Une approche réseau pour l'inférence du rôle des microARN dans la corégulation des processus biologiques

Ricky Bhajun

UNIVERSITÉ GRENOBLE ALPES

THÈSE

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES
Spécialité : **Modèles, Méthodes et Algorithmes en Biologies, Santé et Environnement.**
Arrêté ministériel : 7 août 2006

Présentée par
Ricky Bhajun

Thèse dirigée par **Xavier Gidrol**,
codirigée par **Christian Lajaunie** et
coencadrée par **Laurent Guyon**

préparée au sein du **Laboratoire BIOMICS**
dans **l'École Doctorale Ingénierie pour la Santé, la Cognition et l'Environnement**

Une approche réseau pour l'inférence du rôle des microARN dans la corégulation des processus biologiques

Thèse soutenue publiquement le **8 octobre 2015**,
devant le jury composé de :

M. Denis THIEFFRY
Professeur à l'Ecole Normale Supérieure, Paris, Président
M. Pascal BARBRY
Directeur de Recherche CNRS, Nice, Rapporteur
M. Daniel, GAUTHERET
Professeur à l'Université Paris-Sud, Orsay, Rapporteur
M. Xavier GIDROL
Docteur, CEA, Grenoble, Directeur de thèse
M. Christian LAJAUNIE
Docteur, CBIO, Paris, Co-directeur de thèse
M. Laurent GUYON
Docteur, CEA, Co-encadrant de thèse

Remerciements

Après ces trois années de thèse, je qualifierais le doctorat de grande aventure. Loin des Indiana Johns, Otto Lidenbrock et autres Nathan Drake, la thèse est une aventure scientifique et intellectuelle dans un premier temps, avec tout ce que cela implique comme causes et conséquences : joies et déceptions, découvertes et déconvenues, appréhension et soulagement… Mais la thèse est également une aventure humaine, où nous – étudiants – avons l'occasion de rencontrer de nombreuses personnalités qui forgeront notre habilité à nous *challenger* et à surmonter les nombreux problèmes qui se présenteront à nous durant et après la thèse. Alors que serait une thèse sans sa page de remerciement envers toutes ces personnes ? Probablement une pièce montée dans toute sa splendeur mais sans ces petites touches artistiques faites uniquement de sucre et que la plupart des invités laisseront de toute manière de côté *in fine*.

Viens donc le moment des remerciements et la liste est longue… Je commencerai donc par chaudement remercier mes directeurs de thèse Dr Xavier « Chef » Gidrol et Dr Christian Lajaunie pour leur disponibilité, leurs conseils et aussi simplement le fait de m'avoir fait confiance durant ces trois années afin de mener à bien mon projet. Ca aura été un honneur pour moi d'être encadré par vous deux. Et j'ose espérer, Xavier, que d'ici 22 ans, tu auras le même physique que le grand père faisant de la calisthénie ! Même à cet âge, je ne doute pas que tu continueras à fournir tes conseils et ta formidable expertise (et à écouter du rap !) à tes futurs étudiants comme tu as su le faire avec moi tout au long de ces trois années. Pour moi, tu as été un mentor, un chef d'orchestre, un ami mais surtout et avant tout un véritable 先生. Je remercie également chacun des membres de mon jury de thèse pour le temps qu'ils ont accordé à la lecture de mon manuscrit et pour leur retour très informatif.

Mes seconds remerciement vont bien évidemment à mon encadrant de thèse : le Dr Laurent Guyon qui a su m'aiguiller tout au long de ces travaux, m'a remis dans le droit chemin lorsque je dérivais et m'a soutenu au mieux pendant les moments moins joyeux. J'ai grandement apprécié chacune de nos discussions, toujours plus enrichissantes les unes que les autres et nos fou-rires sur nos termes biologiques inventés ! Je maintiens que certains devraient être adoptés d'ailleurs. Tu m'auras fait connaitre énormément de choses et ça a été un très grand plaisir pour moi de travailler et partager avec toi.

Je tiens également à remercier le Dr Arsia Amir-Aslani pour m'avoir offert la possibilité de suivre la double formation en management durant cette thèse. Cela n'aura pas été facile mais cette formation m'aura profondément changé. Je vous remercie pour tous vos conseils et votre bienveillance. Vous n'êtes pas souvent disponible mais vous restez extrêmement efficace lorsque vous l'êtes ! Sans oublier…*Science **is** a business.*

Mes remerciements vont également à tous les membres du laboratoire BIOMICS : Eric, les deux Patricia, Delphine, Amandine, les deux Stéphanie, Sophie, Ruth, Vincent, Max, Nath… tous autant que vous êtes, vous avez participé de près ou de loin à ma formation et/ou à mes projets et c'est avec plaisir que j'ai pu vous servir et partager avec chacun de vous. Une pensé très particulière pour la descente de la dent de croles de nuit et son (absence de) couché de soleil au sommet… Je m'en souviendrai à tout jamais. En bien, évidemment, Eric !

Viennent ensuite ceux ayant été présents bien avant BIOMICS mais ayant tout de même eu un grand rôle à jouer dans mon orientation : le Dr Yannis François en premier lieux, qui aura su me donner goût à la recherche et la science après ma reprise. Tu auras toujours cru en moi et tu m'auras poussé à aller toujours plus loin. Si j'en suis où j'en suis aujourd'hui, tu en es grandement responsable ! Vient derrière le Dr Didier Rognan et son (ancienne) équipe (Dr Esther Kellenberger, Dr Jamel Meslamani et Dr Jeremy Desaphy). Nous n'avons pas passé beaucoup de temps ensemble mais là aussi, vous tous avez grandement orienté mon parcours.

J'exprime toute ma gratitude envers mes meilleurs amis, qui m'encouragent au mieux lorsque nous nous voyons (trop peu souvent ces dernières années) et me permettent en ces rares occasions d'oublier complétement le stress et les tracas du quotidien. Et puis, combien de personnes peuvent se targuer d'avoir des amis prêts à faire 500 bornes pour aider à un déménagement. Au final vous aviez tous raison… J'aurais effectivement fait une thèse, je vous dois un mac, je crois ! Ces personnes à qui je dois mon caractère et sans aucun doute ma folie sont (sans ordre particulier) : le Dr Omid « patachon » Taghavi – futur néphrologue de haut vol ; M. Nicolas « l'autre taré » Bellot – kinésithérapeute pour qui « le trapèze, c'est pas l'épaule donc pas ma spécialité, mais attend je te la casse comme ça je peux t'aider » ; M. Frédéric « tronc » Bellot – futur chirurgien-dentiste (si, si, chirurgien) et drogué de café ; Mlle Majda « vieille folle » Koubati – future Dr en Pharmacie et ayant de sérieux soucis avec tout objet technologique. Même si nous nous voyons beaucoup moins et que nous nous sommes dispersés, je garde toujours une pensée pour chacun de vous bande de goulaf.

Je tiens également à remercier trois personnes que je considère aujourd'hui comme des membres de ma propre famille et mes propres frères : M. Yacine « vieux fou » Oleemahomed, M. Nawez « jeune fou » Oleemahomed et M. Hussayn « huss » Janally. Il est vrai que chacun d'entre nous est parti de son côté et qu'il est plus que rare que nous nous rencontrions mais c'est toujours un plaisir de vous revoir. Vous me manquez tous les trois, j'espère que nous nous retrouverons tous ensemble, un jour ou l'autre...

Je remercie également mes parents, M. Basdeo Bhajun et Mme Ramawatee Bhajun pour m'avoir offert tout ce qui était dans leur moyen pour ma réussite. Vous n'avez pas eu une vie facile, bien peu de gens s'en rendent réellement compte, mais vous avez su tenir bon jusqu'au bout pour m'orienter sur le droit chemin. Chacun à votre façon, vous m'avez soutenu quel qu'étaient mes choix. J'espère aujourd'hui vous avoir rendu fier, sachez en tout cas, que je suis fier de ce que vous avez accompli avec le peu que vous aviez et que je suis également très fier d'être votre fils. Mes remerciements vont également à ma famille Strasbourgeoise : toute la famille Gunputh qui a également toujours été présente dans les moments difficiles. Tout n'est pas toujours rose et même si je ne le montre pas forcément, sachez que je tiens sincèrement à chacun de vous quatre.

Enfin, je consacrerai ces dernières lignes de remerciements à ma compagne, Mme Noushita Bhojoo, qui me supporte depuis maintenant sept ans. Je sais avoir trouvé en toi ma moitié au sens du poète Aristophane : tu m'écoutes, me soutiens, me conseilles… Je me sens simplement complet auprès de toi. Je te remercie pour ton amour, ta confiance, ta patience et toute l'attention que tu m'accordes dans les bons moments comme les mauvais... मैं तुमसे प्यार करता हूँ.

Au moment où j'écris ces mots, je ne sais pas ce que l'avenir me réserve mais sachez que vous avez tous été une source d'inspiration pour moi et que cette inspiration me restera jusqu'au bout.

Contribution

Au cours de cette thèse, j'ai pu participer de près ou de loin à différents travaux scientifiques, au sein du laboratoire BIOMICS mais également d'autres travaux moins scientifiques pour l'école supérieure de commerce de Grenoble. Voici mes différentes contributions pour ces articles :

Bhajun, R., Guyon, L., Pitaval, A., Sulpice, E., Combe, S., Obeid, P., Haguet, V., Ghorbel, I., Lajaunie, C., and Gidrol, X. (2015). A statistically inferred microRNA network identifies breast cancer target miR-940 as an actin cytoskeleton regulator. Scientific Reports.

Bhajun, R., Guyon L., and Gidrol, X. *Degeneracy and pluripotentiality in miRNA networks confer robustness to gene expression and canalization of phenotypic outcomes*. Soumis à CMLS

Bhajun, R., Guyon, L., Freida, D., Gerbaud, S., Lajaunie, C., and Gidrol, X. A miRNA community of stemness exit. En écriture.

Guyon, L., Lajaunie, C., Fer F., **Bhajun, R.**, Sulpice, E., Pinna, G., Campalans, A., Pablo Radicella, J., Rouillier, P., Mary, M., Combe, S., Obeid, P., Vert, J-P., and Gidrol, X. (2015). *Φ-score : A cell-to-cell phenotypic scoring method for sensitive and sleective hit discovery in cell-based assays*. Scientific Reports.

Amir-Aslani. A., **Bhajun, R**., and Sainte-Foie, N., (2015). Biotech et Pharma : même galère, Biofutur (mai).

Amir-Aslani. A., and **Bhajun, R**., (2014). Le réveil des biotechs françaises, Biofutur (aout).

Bhajun, R., and Amir-Aslani, A., (2015). Les 7 « v » pilliers des mégadonnées, Spectra Biologie (novembre).

Bhajun, R., and Amir-Aslani, A., (2015). Les mégadonnées et la génétique moléculaire, Spectra Biologie (novembre).

Amir-Aslani, A., and **Bhajun, R.**, Mégadonnées et santé (titre provisoire). Non soumis.

Picollet-D'hahan, N., Gerbaud, S., Kermarrec, F., Alcaraz, J-P., Obeid, P., **Bhajun, R**., Guyon, L., Sulpice, E., Cinquin, P., Dolega, M. E., Wagh, J., Gidrol X., and Martin, D. K. (2014). *The modulation of attachment, growth and morphology of cancerous prostate cells by polyelectrolye nanofilms.* Biomaterials.

Lexique

ADN	Acide Désoxyribonucléique
ADNc	ADN complémentaire
AGO	Membres de la famille de protéines Argonautes
ARN	Acide Ribonucléique
Assortativité	Propension des nœuds très connectés d'un réseau à se lier à d'autres nœuds très connectés
BH	Procédure de corrections pour tests multiples introduite par Benjamini et Hochberg
Centralité	Mesures de l'importance des nœuds dans un réseau
CLASH	*Crosslinking ligation and sequencing of hybrids*
CLIP	*Crosslinking and immunoprecipitation*
Cluster	*Groupe d'entités (nœuds dans un réseau p.ex.)*
Déadénylation	Processus de suppression des Adénosine en 3' des ARNm mature
Decapping	Décoiffage - processus de suppression de la coiffe en 5' des ARNm mature
Dégénérescence	Capacités d'éléments structurellement différents à se substituer les uns aux autres
DGCR8	*DiGeorge Syndrom Critical Region Gene 8*
Disassortativité	Propension des nœuds à faible degré à se lier à des nœuds à fort degré. Issu du terme anglais « *disassortativity* »
eIF	*Eukaryotic Translation-initiation Factor - facteur eucaryote de l'initiation de la traduction*
Exon	Portions transcrites d'un gène et restant dans l'ARNm mature après épissage
GEO	Gene Expression Omnibus - base de données regroupant des données d'expression
GO	Gene Ontology - base de données regroupant des informations sur les gènes en termes de processus biologique (BP), compartiment cellulaire (CC) et fonction moléculaire (MF)
GTP	Guanosine triphosphate
GTPases	Enzymes qui lient et hydrolyse la GTP
Intron	Région située entre deux exons d'un ARNm non mature
IRES	*Internal Ribosome Entry Site*
LNA	*Locked-nucleic acid*
MLCII	*Myosin light chain 2*
MRE	*miRNA recognition element*
omique	Analyse d'ensembles
ORF	*Open Reading Frame*
PABPC	*Poly(A)-binding protein*

Petit-monde	Structure particulière des réseaux ou quelques nœuds permettent de relier l'ensemble du réseau tel que le plus court chemin entre n'importe quelle paire de nœuds est faible
Pluripotentialité	Capacités d'un élément à effectuer plusieurs fonctions différentes
PPI	*Protein-protein interaction*
Réseau clairsemé	Réseau à faible densité
Rich clubs	Communauté(s) de hubs dans un réseau
RISC	*RNA-induced Silencing Complex*
ROCK	*Rho-associated protein kinase*
RPE1	Retinal pigmented epithelial cells - lignée cellulaire épithélial de rétine
Région *seed*	Région très conservés en 5' du miARN mature
SG	*Stress Granules*
TNRC6	*Tri-nucleoide repeat containing protein*
TRAP	*Target RNA affinity purification*
TRBP	*The Human Immunodeficiency Virus Transactivating Response RNA-binding Protein*
UTR	*Untranslated Region*

Résumé

L'interférence par l'ARN est un processus selon lequel un petit ARN non codant se lie à un ARN messager cible dans la cellule pour moduler son expression. Ce mécanisme a été conservé au cours de l'évolution : il est retrouvé aussi bien chez les animaux que chez les végétaux. Nous savons aujourd'hui que le rôle de l'interférence par l'ARN est fondamental, dans le développement embryonnaire comme dans la progression tumorale. Les microARN (miARN) sont des ARN non codant endogènes dont l'une des particularités est leur capacité à réguler tout un ensemble de gènes par interférence avec les ARN messagers. Il est ainsi prédit qu'un seul miARN serait capable de réguler plusieurs centaines de gènes différents. La thèse a consisté en l'analyse de la corégulation médiée par les miARN grâce à l'inférence de réseau basée sur le partage de gènes cibles. La corégulation est un phénomène où plusieurs miARN différents interviennent sur les mêmes familles de gènes et donc sur les mêmes processus biologiques. Le travail a plus spécifiquement consisté en la mise en place d'un réseau de miARN, en son analyse topologique mais également en son interprétation biologique. Le but final était de proposer de nouvelles hypothèses biologiques à tester afin de mieux comprendre la corégulation des processus biologiques par les miARN. Au travers de ces travaux, deux groupes de miARN ont pu être mis en évidence, dont l'un impliqué dans la régulation de la signalisation par les petites GTPases – hypothèse par la suite validée par plusieurs expériences *in vitro*. Dans un second temps, une communauté de miARN impliquée dans le maintien de la pluripotence des cellules souches a pu également être mise en évidence. Pour compléter ces analyses, une étude systémique de la topologie des réseaux de miARN a été menée afin de mieux comprendre leur intégration dans les réseaux biologiques et leur rôle dans le devenir cellulaire.

Abstract

RNA interference is a process in which a small non-coding RNA will bind to a specific messenger RNA and regulate its expression. This evolutionary conserved mechanism is found in all superior eukaryotes from plants to mammals. Nowadays, we know that RNA interference is a major regulatory process involved in developmental biology and tumor progression. MicroRNAs (miRNAs) are endogenous (coded in and produced by the cell) non-coding RNAs which are able to regulate a whole set of genes, typically hundreds of genes. This doctoral thesis consisted in the analysis of the miRNA mediated coregulation through a network approach based on target sharing. Coregulation is the process where many different miRNAs will regulate the same set of genes and thus the same biological process. In particular, the work consisted in the inference of a miRNA network, in its topological analysis and also its biological interpretation. Indeed, the final aim of the work was to generate new biological hypothesis. As such, two different groups of miRNAs were first retrieved. One of them was predicted to be involved in the small GTPase signaling and was further validated *in vitro*. Moreover, a miRNA community involved in the maintenance of stem cells pluripotency was also discovered. Finally, a systemic analysis of the target-based miRNAs network was conducted to better understand their integration with biologic networks and their role in cell fate.

Table des matières

Introduction

A. Les microARNs

Présents chez tous les eucaryotes supérieurs, de la plante à l'animal, les microARNs (miARN) sont de petits ARN non codants d'approximativement 20 nucléotides. Ces ARN endogènes sont des régulateurs post-transcriptionels de l'expression des gènes. Bien que produisant des effets généralement peu marqués sur le niveau d'expression des protéines (Baek et al., 2008), nous savons aujourd'hui que les miARN ont un rôle prépondérant dans un grand nombre de processus physiologiques tels que le développement embryonnaire, l'apoptose ou encore la progression tumorale (Alvarez-Garcia and Miska, 2005; Kloosterman and Plasterk, 2006). Dans cette première partie de la thèse, nous nous intéresserons aux particularités des miARN, de leur découverte à la recherche de leurs cibles en passant par leur(s) mécanisme(s) d'action.

1. Une histoire de miARN

Le premier miARN fut découvert chez le nématode *Caenorhabditis elegans* en 1993, conjointement dans les laboratoires de Garry Ruvkun et Victor Ambros (Lee et al., 1993; Wightman et al., 1993). A l'époque, les deux équipes s'intéressaient aux gènes impliqués dans le développement des différents stades larvaires (gènes hétérochroniques) du petit organisme modèle. Une des souches avec mutation nulle pour un gène hétérochronique nommé *lin-4* montrait alors de graves défauts de développement chez l'animal. En ce temps, *lin-4* était déjà connu comme étant nécessaire au bon développement larvaire, notamment pour la transition entre les deux premières phases de ce stade (Chalfie et al., 1981). Quelques années plus tard, des chercheurs démontrèrent qu'une mutation introduite dans le gène *lin-14* (un autre gène hétérochronique) était capable d'inverser le phénotype du mutant nul pour *lin-4* (Ferguson et al., 1987), indiquant donc d'une part un potentiel de régulation négative de *lin-14* par *lin-4* et d'autre part, que les défauts de développement observés plus tôt pouvaient être dus à une accumulation du gène *lin-14* dans les cellules du ver. Durant cette même période, la découverte fortuite d'un autre mutant pour le gène *lin-14* dans le laboratoire d'Ambros et

montrant les mêmes défauts que ceux observés pour la souche *lin-4-null* vint renforcer cette idée de régulation négative par un gène (Lee et al., 2004a).

Afin de mieux comprendre ces différents phénomènes, Ambros et Ruvkun travaillèrent ensemble jusqu'en 1989 pour cloner le gène *lin-14,* mais les deux chercheurs s'orientèrent rapidement vers des travaux indépendants, le premier se focalisant sur *lin-4* et le second sur *lin-14*. De son côté, l'équipe d'Ambros découvrit la présence d'un fragment d'ADN génomique chez *C. elegans* pouvant contenir le gène *lin-4*, la particularité de cette portion génique étant l'absence de codons *start* et *stop*. En outre, l'introduction de mutations dans le cadre de lecture ouvert supposé (*Open Reading Frame* – ORF) du gène ne montrant aucun phénotype particulier amena les auteurs à la conclusion que le gène ne codait pas pour une protéine. Ils mirent également en évidence la présence de deux petits transcrits de 61 nucléotides et 22 nucléotides correspondant respectivement au miARN primaire et mature et qui semblaient correspondre au gène *lin-4* (Lee et al., 1993). L'équipe de Ruvkun, quant à elle, découvrit que la fameuse mutation dans le gène *lin-14* liée aux défauts de développement touchait en fait la région non traduite en 3' du gène (3' *Untranslated Region* – 3'UTR). Ils découvrirent également que la protéine LIN-14 était régulée de manière post-transcriptionel pendant le développement du ver (Wightman et al., 1993). Cependant, ce fut uniquement grâce au partage de leurs résultats en juin 1992 que les deux équipes se rendirent compte d'un fait capital : les transcrits de *lin-4* étaient anti-complémentaires à une séquence en 3'UTR du gène *lin-14*. L'ensemble de leurs résultats corrélaient, une régulation négative gène-gène était donc parfaitement envisageable dans ces conditions !

C'est donc un an plus tard, en décembre 1993 que les deux équipes publièrent leurs résultats de manière indépendante dans le même numéro du journal Cell, décrivant un nouveau système de régulation génique où la somme de leurs deux études montrait que le petit ARN non codant lin-4 était capable de réguler l'expression protéique du gène *lin-14* au travers de sa région 3'UTR.

De manière intéressante, ces conclusions ne marquèrent pas la découverte de l'ARN interférence telle que nous la connaissons aujourd'hui, ni d'ailleurs celle de la régulation génique par ARN endogène. En effet, les premiers signes de ce type de régulation étaient observés avant 1988, essentiellement chez les procaryotes (Simons, 1988). C'est en 1990, soit trois années avant les publications d'Ambros et Ruvkun, qu'une équipe démontra l'existence de ce type de système chez les eucaryotes, notamment chez *Petunia hybrida* (Napoli et al., 1990). En 1992, une étude chez *Dictyostelium discoideum* (Hildebrandt and Nellen, 1992) publiée dans Cell, mit en évidence les premiers signes d'une régulation génique par ARN non codant endogène. Cette étude servit d'ailleurs de support aux premières conclusions d'Ambros et Ruvkun (Lee et al., 2004a). La découverte du mécanisme d'ARN interférence *en soi* fut d'ailleurs attribué à Craig C. Mellow et Andrew Z. Fire (Fire et al., 1998), ce qui leur a d'ailleurs permis d'obtenir le prix Nobel de Physiologie et de Médecine en 2006.

La découverte du deuxième miARN n'arriva que 7 ans après celle de lin-4. Reinhart *et al.* démontrèrent qu'un petit ARN de 21 nucléotides nommé *let-7* (encore un autre gène hétérochronique chez *C. elegans*) contrôlait, quant à lui, les derniers stades de développement du ver (dernier stade larvaire vers premier stade adulte). Alors que l'augmentation de l'activité du gène entrainait une expression précoce des stades adultes du ver, sa suppression pouvait entrainer une inversion des stades adultes vers les stades larvaires. Tout comme pour l'expérience entre *lin-4* et *lin-14*, les auteurs de cette étude découvrirent que *let-7* était capable d'inverser un phénotype nul pour un autre gène (*lin-41)*, que sa séquence était complémentaire à celles de plusieurs gènes dans le génome du ver (notamment *lin-14*, *lin-28*, *lin-41*, *lin-42* et *daf-12*) et enfin que *let-7* était capable de réguler l'expression de *lin-41* spécifiquement au travers de sa 3'UTR (Reinhart et al., 2000). Fait plus intéressant encore, cette équipe montra la même année qu'il existait une très forte conservation de séquence de *let-7* au sein de différentes espèces animales, de l'éponge à l'être humain mais également que ce gène était exprimé dans la majorité des tissus humains (Pasquinelli et al., 2000). Cette seconde étude marqua très probablement le début de la révolution de la recherche sur les petits ARN non

codants endogènes avec une évolution qui restera quasi-exponentielle jusqu'en 2010 (Figure 1).

Les miARN ne furent clairement établis comme une classe indépendante d'ARN non codant régulatrice de l'expression des gènes qu'en 2001, ajoutant ainsi un nouveau niveau de complexité au dogme central de la biologie moléculaire où les ARN sont non seulement des porteurs d'information mais également des régulateurs (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). MiRBase v1, la toute première version de la base de données de référence sur les miARN (« *the miRNA registry* » à l'époque) fût mise en ligne fin 2002. Elle comptait alors 218 entrées (Griffiths-Jones, 2004). Aujourd'hui, la base de données comptabilise 28645 entrées différentes, toutes espèces confondues, dont 2588 miARN matures humains (miRBase v21, Juin 2014). Ce chiffre tend cependant à se stabiliser. La base de données regroupe tout un ensemble d'informations sur les miARN, notamment leur nom

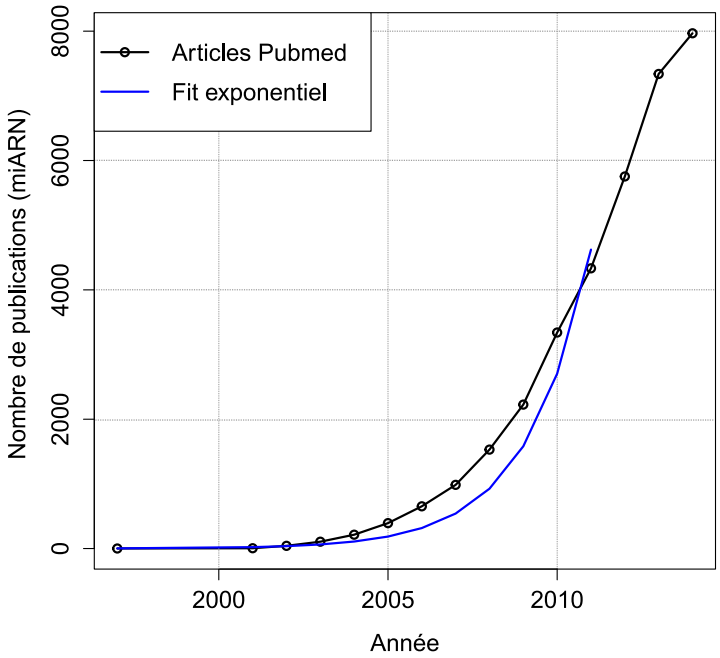


Figure 1. Evolution du nombre de publications associées aux miARN. Recherche dans pubmed depuis 1997 jusqu'à 2014.

de référence, leur emplacement génomique, les séquences, des références de détection des miARN, des liens vers différentes bases de données de prédiction de cibles, etc.

2. Synthèse des miARN

Les miARN sont codés dans le génome comme de longs transcrits primaires, souvent sous la dépendance de leur propre promoteur. Ces derniers peuvent cependant être retrouvés en cluster sur le génome et donc transcrits sous forme de polycistron (Lee et al., 2002). Ils peuvent être classés en quatre groupes distincts suivant leur emplacement génomique (Figure 2) :

- Les miARN exoniques, situés dans des régions non codantes (gènes d'ARN non codants)
- Les miARN exoniques, situés dans des régions codantes.
- Les miARN introniques, situés dans des régions non codantes.
- Les miRNAs introniques, situés dans des régions codantes.

Il existe différentes voies de biogénèse des miARN chez les animaux, généralement communes pour l'ensemble des quatre types. Dans la voie canonique, les miARN sont premièrement transcrits dans le noyau des cellules, clivés une première fois puis exportés vers le cytoplasme où ils seront clivés une seconde fois en miARN mature (Figure 3 A). La maturation des miARN est donc un système compartimentalisé entre le noyau et le cytoplasme. Les autres voies de maturation des miARN empruntent, quant à elles, certains éléments de la voie canonique tout en en contournant d'autres.

a) Synthèse des miARN par la voie canonique et dégradation

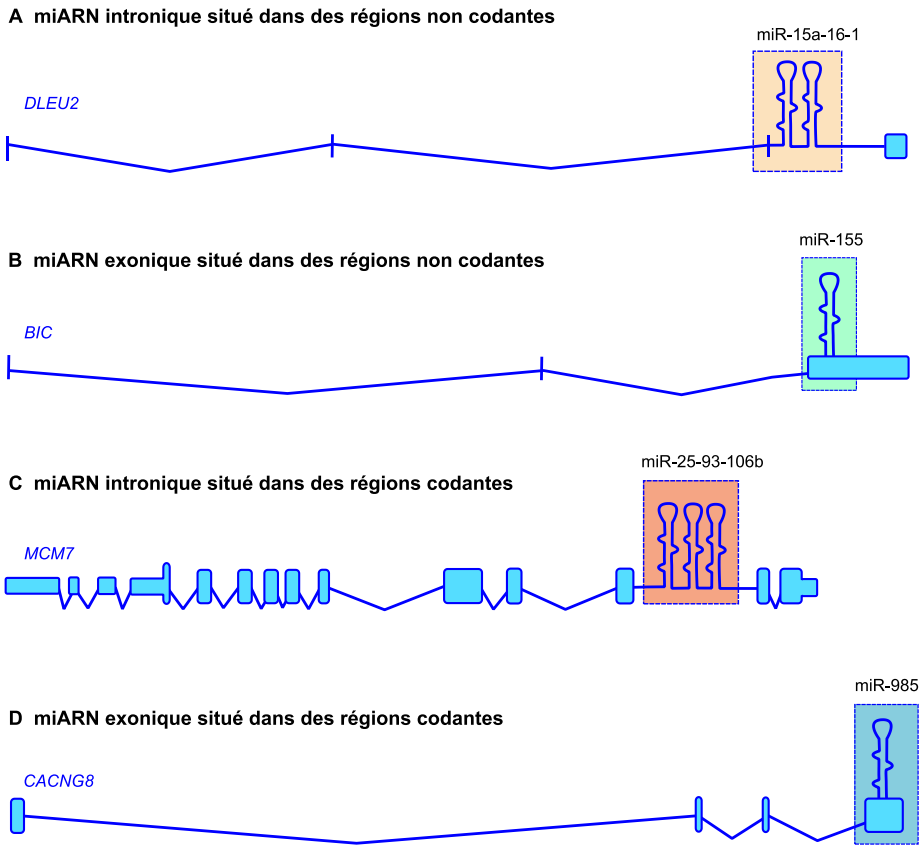


Figure 2. Catégorisation du type de miARN en fonction de leur localisation génomique. A | miARN intronique dans un transcrit non codante comme le pour cluster miR-15a/16-1 se trouvant dans le gène DLEU2. B | miARN exonique situé dans un transcrit non codant comme le miR-155 dans le gène non codant *BIC*. C | miARN intronique situé dans transcrit codant pour une protéine comme le cluster miR-25/93/106b dans le gène MCM7. D | miARN exonique situé dans une région codante comme le miR-985 dans le dernier exon de CACNG8. Les boites bleues indiquent les régions codantes des protéines. Adaptée d'après (Kim et al., 2009).

Le processus de synthèse des miARN commence par leur transcription par l'ARN polymérase II dans le noyau des cellules (Lee et al., 2004b). Tout comme pour l'expression des gènes, cette synthèse est sous la dépendance de différents facteurs de transcription. Les longs transcrits primaires ainsi créés sont appelés pri-miARN (pour *primary-miRNAs)* et – tout comme les ARN messagers (ARNm) – sont coiffés en 5' et polyadénylés en 3' (Cai et al., 2004). Ils possèdent également une (ou plusieurs) structure(s) locale(s) en tige-boucle (Figure 2, Figure 3).

Les pri-miARN et plus spécifiquement les structures en tige boucle sont alors pris en charge par l'endonucléase nucléaire (RNase III) Drosha, dont le rôle est de découper le

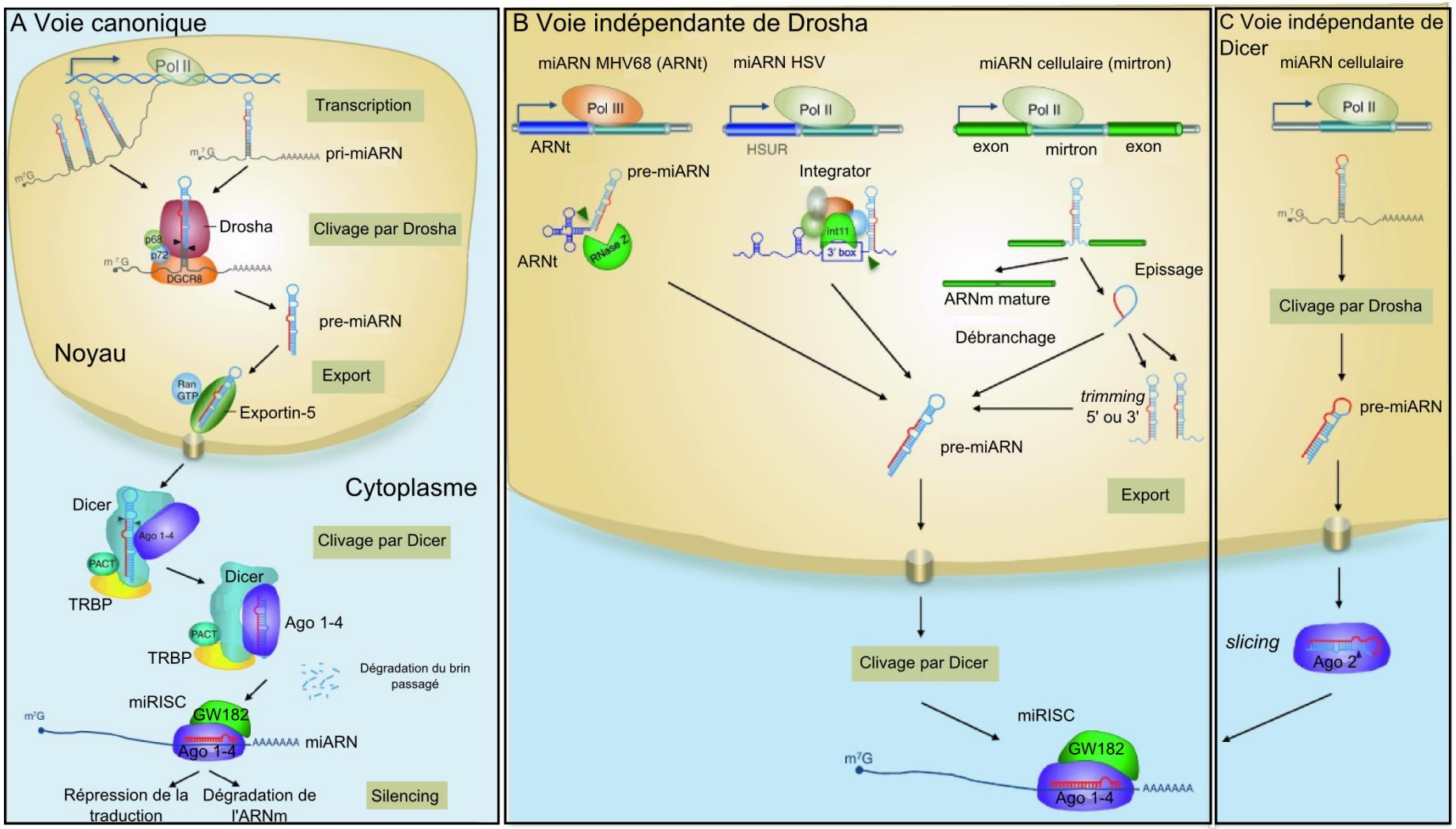


Figure 3. Différentes voies de biogénèse des miARN. A | Voie canonique de la biogénèse des miARN. La première partie se déroule dans le noyau où le miARN est transcrit par la polymérase II, clivé par Drosha et DGCR8 et exporté par l'exportin-5. Dans le cytoplasme, le pre-miARN est clivé par Dicer pour être ensuite chargé dans miRISC. B | Biogénèse indépendante de Drosha. Différentes voies sont présentées : la voie la plus à gauche présente une méthode de biogénèse à partir d'un ARN de transfert initié par conséquent par la polymérase III. Au milieu, une miARN couplé à un ARNm (viral dans ce cas-ci) nécessitant le complexe « *Integrator* » pour la formation du pre-miARN. *Integrator* est un complexe nucléaire de 12 protéines qui s'associe à la polymérase II. Enfin la voie la plus à droite est la voie mirtron où le pre-miARN est formé par épissage. C | Voie indépendante de Dicer où la structure du pri-miARN permet un clivage par Drosha. HSV : Saimiri herpesvirus. Adaptée d'après (Libri et al., 2013).

pri-miARN afin de libérer les tiges boucles pour former des pre-miARN (pour *precursor-miRNA*), d'environ 80 nucléotides. Drosha est une protéine conservée d'environ 160 kDa (Figure 4) (Fortin et al., 2002). Pour assurer sa fonction chez l'être humain, elle forme un complexe appelé « *microprocessor* » d'environ 650 kDa en s'appariant avec son co-facteur DGCR8 (*DiGeorge Syndrom Critical Region Gene 8*) (Gregory et al., 2004). Le clivage par Drosha laisse un certain nombre de nucléotides libres (généralement 2) en 3' de l'ARN (*3' overhang*) (Lee et al., 2003). Cette particularité est une marque essentielle au mécanisme d'ARN interférence puisqu'il ne se limite pas qu'aux miARN mais se retrouve aussi pour d'autres types d'ARN interférents (p.ex. *small hairpin* ARN, *small interfering* ARN, etc.). Le système de reconnaissance entre Drosha et les pri-miARN restent encore assez peu connu, bien que la structure tridimensionnelle du pri-miARN semble être capitale pour l'appariement entre la protéine et l'ARN.

Les pre-miARN sont ensuite transférés du noyau vers le cytoplasme à l'aide de l'Exportin-5 et de son partenaire Ran-GTP. Cet export s'effectue au travers des pores nucléaires, des complexes protéiques attachés à la membrane nucléaire (Figure 3 A). Il semblerait également que le rôle de l'Exportin-5 ne se limite pas à l'export des pre-miARN uniquement mais aussi à la stabilité générale de ces derniers (Bohnsack et al., 2004).

Une fois dans le cytoplasme, les pre-miARN sont clivés par la nucléase Dicer-1 (Figure 4) pour former des duplexes de miARN double brin (Bernstein et al., 2001; Ketting et al., 2001). Dicer-1 est une protéine d'environ 200 kDa très conservée dans l'évolution. Si la protéine seule est capable de cliver les pre-miARN *in vitro*, elle s'associe *in vivo* à différentes protéines comme certains membres de la famille Argonaute (Ago2) ou TRBP (*the human immunodeficiency virus transactivating response RNA-binding protein*) afin d'assurer sa fonction (Carmell et al., 2002; Chendrimada et al., 2005). Ces différents complexes permettent non seulement le clivage des pre-miARN mais servent également de protecteurs et de plateformes pour un assemblage avec le complexe RISC (*RNA-induced silencing complex*).

Les duplexes sont alors séparés par une hélicase encore non identifiée pour former deux miARN matures d'environ 20 nucléotides. Plusieurs hélicases ont été retrouvées associées au complexe RISC, notamment Gemin3 (Mourelatos et al., 2002), DDX5 (Salzman et al., 2007) et MOV10 (Meister et al., 2004). Le rôle précis de ces hélicases n'a toutefois pas encore été découvert. L'un ou l'autre des brins peut alors être incorporé dans le complexe protéique miRISC (pour « *miRNA RISC* », terme spécifique au complexe RISC chargé d'un miARN) ; complexe principalement formé par différents membres de la famille des protéines Argonaute (AGOs) (Figure 3 et Figure 4). Les détails de sélection entre les deux brins sont encore assez incompris. D'après certaines expériences basées sur les siARNs, on pense que le brin majoritairement recruté par le complexe miRISC est le brin dont la stabilité thermodynamique des paires de bases en 5' est la plus faible (Schwarz et al., 2003), l'autre brin se retrouvant dégradé assez rapidement. Historiquement, un astérisque était accolé au nom du brin supposé dégradé pour pouvoir les différentier (p.ex. hsa-miR-1* contre hsa-miR-1). En réalité, le ratio d'expression entre les deux brins pour un même miARN est très variable d'un tissu à un autre. Il existe même des cas où les deux espèces sont retrouvées. Cette notation a donc été abandonnée au profit des termes plus génériques : « miARN-3p » et « miARN-5p », faisant référence respectivement au brin en 3' et 5' du pre-miARN.

Les miARN complexés à RISC sont particulièrement stables mais, tout comme les ARNm, ces derniers sont dégradés *in vivo* au bout d'un certain temps. Cette stabilité varie en fonction des miARN. Par exemple, miR-382 – un miARN impliqué dans le mécanisme associé au virus de l'immunodéficience (VIH) – est assez instable et généralement très rapidement dégradé. Les 7 derniers nucléotides en 3' semblent déterminants dans la stabilité du miARN puisque des mutations dans cette région l'augmentent considérablement (Bail et al., 2010). Dans cette étude, les auteurs montrent que miR-382 est dégradé principalement par l'exoribonucléase 3'-5' exosome mais également par l'exonucléase 5'-3' XRN1 dans une moindre mesure. Une autre étude montre également l'implication de l'exonucléase 3'-5'

Figure 4. Structures schématiques des protéines Dicer, Drosha et AGO2 chez l'humain et la drosophile. Des domaines RNAse III (RIIIa, RIIIb) sont retrouvés dans Dicer et Drosha dans la partie C-terminal ainsi qu'un domaine hélicase en N-terminal. Le domaine PAZ est conservé entre Dicer et AGO. Les protéines AGO portent en plus un domaine PIWI. DUF : *domain of unknown function*. D'après (Meister and Tuschl, 2004)

PNPT1 dans la dégradation de miR-221 (Das et al., 2010). En revanche, aucune endonucléase n'a été identifiées dans ce rôle, pour le moment. D'autres facteurs caractéristiques peuvent aussi avoir une influence sur le *turnover* des miARN ; c'est notamment le cas du cycle cellulaire, de l'abondance des facteurs de croissance ou encore du type cellulaire (Rüegger and Großhans, 2012).

b) Les voies indépendantes de Drosha/DGCR8

Comme leur nom l'indique, ces voies se passent de la protéine Drosha pour la première étape de clivage des miARN. En particulier, la voie « *mirtron* » est probablement la voie alternative la plus établie à l'heure actuelle. Les mirtrons sont des miRNAs intronique en tige-boucle dont les extrémités coïncident avec des sites d'épissage (une extrémité acceptrice et l'autre donneuse) (Ruby et al., 2007; Melamed et al., 2013). Cette particularité permet ainsi un clivage des mirtrons dans le noyau sans passer par la protéine Drosha mais en utilisant le mécanisme d'épissage. Les pre-miARN ainsi créés suivent par la suite la voie canonique de

maturation des miARN (Figure 3 B). Ce mécanisme de création de pre-miARN par épissage a été retrouvé au sein de plusieurs espèces animales (Berezikov et al., 2007).

Quelques auteurs ont pu mettre en évidence d'autres voies indépendantes de Drosha/DGCR8 où les pre-miARN pouvaient être synthétisés à partir de snoARN (*small nucleolar RNA*) (Ender et al., 2008), d'endo-shRNA (*endogenous small haipin RNA*) (Babiarz et al., 2008) ou encore d'ARNt (ARN de transfert, Figure 3 B) (Haussecker et al., 2010). Chaque catégorie possède ses propres caractéristiques de biogénèse mais pour la plupart, ces dernières ne sont pas encore entièrement identifiées (Yang and Lai, 2011).

c) La voie indépendante de Dicer

La voie indépendante de Dicer a été mise en évidence spécifiquement pour le miR-451, un miARN parfaitement conservé chez les vertébrés, du poisson à l'être humain. Dans ce cas, le pri-miRNA est bien clivé par Drosha/DGCR8 pour former un pre-miARN. Ce dernier est cependant légèrement plus petit que la normale : la tige boucle ne mesure qu'environ 18 paires de bases au lieu de 60-70 des pre-miARN classiques. Cette particularité empêche toute interaction avec Dicer mais permet tout de même un recrutement par les protéines de la famille Argonaute. En l'occurrence, c'est la protéine Ago2 qui se charge du clivage en 3' d'une portion du pre-miARN, formant ainsi un ARN de 30 nucléotides. Afin d'obtenir un miARN mature et fonctionnel, une nucléase encore non identifiée clive les derniers nucléotides restants (Figure 3 C) (Cheloufi et al., 2010; Cifuentes et al., 2010; Yang et al., 2010a).

3. Mécanisme d'action des miARN

Comme exposé ci-dessus, les miARN matures seront recrutés par le complexe RISC afin de réguler l'expression de leur(s) gène(s) cible(s) de manière post-transcriptionelle. On considère aujourd'hui que les miARN sont capables de réguler au moins la moitié des gènes du génome chez l'être humain (Friedman et al., 2009a) et que chacun possède au moins une centaine de cibles différentes (Lewis et al., 2005; Lim et al., 2005). A l'inverse et de manière complémentaire, chaque gène est également susceptible d'être régulé par différents miARN

(Wu et al., 2010). Il a également été montré que lorsqu'un gène possédait plusieurs sites de liaison aux miARN et que plusieurs de ces sites étaient occupés, des effets synergiques sur son expression et/ou celle de sa protéine pouvaient être observés (Saetrom et al., 2007; Wu et al., 2010; Fang and Rajewsky, 2011; Ding et al., 2014). Lorsque cette occupation de sites se fait par des miARN différents, on parle de modules de miARN (Ding et al., 2014).

Le recrutement d'un gène par le complexe miRISC s'effectue principalement par appariement de séquence entre le miARN chargé et un ARNm cible. On considère aujourd'hui qu'il existe principalement deux mécanismes d'action pour l'ARN interférence. Le premier nécessite un appariement parfait ou quasi-parfait entre la séquence du miARN et son ARNm cible, ce qui entrainera le clivage et la dégradation de l'ARNm. C'est un mécanisme principalement retrouvé chez les plantes. Le deuxième mécanisme – majoritaire chez les animaux – implique une liaison imparfaite entre miARN et ARNm. Dans ce cas, l'ARNm n'est pas forcément dégradé : on parle donc de « répression de la traduction ». Nous pouvons cependant noter que les deux mécanismes ont été observés autant chez les plantes que chez les animaux : ils ne sont pas limités à l'un ou l'autre des deux règnes.

a) Le(s) complexe(s) RISC

Le complexe RISC est le centre du mécanisme d'ARN interférence. C'est un complexe ribonucléoprotéique dont le cœur est composé d'une protéine argonaute et d'une protéine riche en glycine et tryptophane de 182 kDa chez la drosophile (GW182). Il se charge de recruter tout autant les miARN (miRISC) que d'autres types d'ARN interférants (p. ex. siRISC pour le recrutement des siARNs).

Les protéines argonautes – sur lesquelles sont directement chargés les ARN interférents – peuvent être classées en deux catégories : la sous-famille Ago et la sous-famille Piwi (Carmell et al., 2002). Si les protéines AGO sont retrouvées dans toutes les lignées somatiques, les protéines PIWI semblent n'être retrouvées que dans les lignées germinales. Il existe 4 protéines Argonautes différentes chez les mammifères : AGO1, 2, 3 et 4. Ces

dernières sont constituées de quatre domaines conservés formant une protéine bi-lobaire : le premier lobe est formé par les domaines N-terminal et Piwi-Argonaute-Zwilli (PAZ), alors que les domaines MID et PIWI forment le second (Figure 5). Le domaine PAZ reconnait spécifiquement la partie sortante 3'OH des ARN (Ma et al., 2004) et est également retrouvé chez Dicer (Figure 4). Le domaine MID se lie, quant à lui, au 5' phosphate des ARN et permet donc un ancrage stable des ARN sur les protéines AGO (Ma et al., 2005; Boland et al., 2010). Ce dernier semblerait également être impliqué dans des interactions protéine-protéine et permettrait donc le recrutement de cofacteurs (Till et al., 2007). Enfin, le domaine PIWI montre une conformation similaire aux RNase H, un enzyme dont le rôle est de cliver les fragments d'ARN dans les duplexes ADN-ARN. Cette similarité indique que PIWI serait en fait l'élément permettant le clivage des ARNm chargés dans RISC (on parle d'activité « *slicer* ») (Parker et

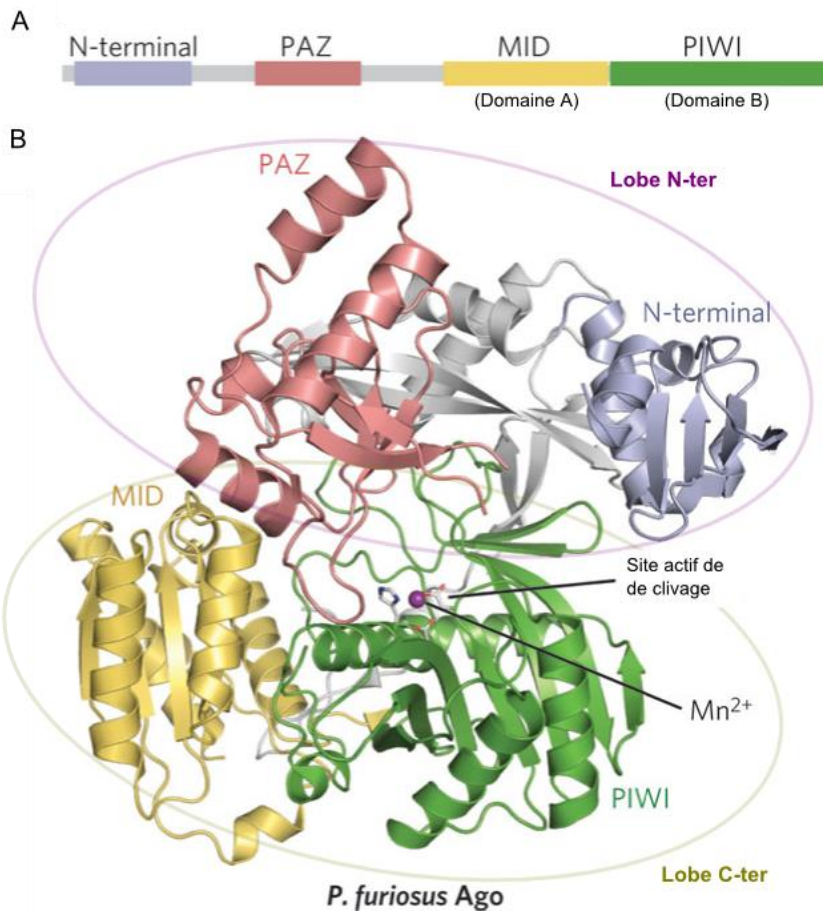


Figure 5. Structure de la protéine AGO. A | Représentation schématique d'une protéine AGO avec ses domaines N-terminal, PAZ, MID et PIWI. B | Structure cristallographique de la protéine AGO chez *P. furiosus* avec les deux lobes représentés. Adaptée d'après (Jinek and Doudna, 2009).

al., 2005; Yuan et al., 2005). Chez les mammifères, seule Ago2 possède l' activité de « *slicing* » (Liu et al., 2004).

Le second constituant du cœur de RISC, GW182, existe sous trois formes paralogues chez les mammifères : TNRC6A, B et C (pour *tri-nucleotide repeat containing protein*). Le domaine N-terminal de ces protéines est très riche en motifs GW, WG ou GWG qui servent de site de liaison aux protéines AGO (Yao et al., 2011). Cette région est suivie d'un domaine putatif d'association à l'ubiquitine (UBA) et d'un domaine riche en glutamine (Q-rich). Enfin, la partie C-terminal est également appelée *silencing domain* à cause de sa faculté à bloquer la traduction. Ces protéines sont également les composants de foci cytoplasmiques dénommés « corps-P » (ou corps-GW) dont le rôle est principalement lié au catabolisme des ARNm et notamment leur déadénylation, *decapping* et dégradation (Eulalio et al., 2007a; Parker and Sheth, 2007).

Ces deux constituants principaux sont également accompagnés d'autres protéines pour le bon déroulement du mécanisme d'ARN interférence. Ces dernières diffèrent en fonction des espèces mais aussi en fonction du type d'ARN interférence considéré (siARN, miARN, shARN, etc.). Il n'existe donc pas un complexe RISC mais différents complexes RISC d'environ 100 kDa pour les plus petits (*low molecular weight* RISC – LMW-RISC) à plus de 2 MDa pour les plus gros (*high molecular weight RISC* – HMW-RISC), chacun avec leurs propres caractéristiques, mais un cœur tout de même sensiblement identique (Höck et al., 2007; Landthaler et al., 2008).

b) Mécanisme chez les végétaux

Dans le règne végétal, le mécanisme de dégradation d'ARNm semble dominer. Ce mode de régulation est clairement établi depuis plusieurs années : les miARN et les ARNm se reconnaissent par anti-complémentarité parfaite ou quasi-parfaite et l'ARNm est clivé directement (Llave et al., 2002; Rhoades et al., 2002). Le clivage s'effectue sur l'ARNm entre les nucléotides 10 et 11 du miARN (clivage endonucléotidique) par les AGOs au travers de

leur domaine PIWI (Figure 6 A) (Souret et al., 2004; Jinek and Doudna, 2009). Les deux fragments d'ARNm sont alors dégradés au niveau des extrémités créées par le clivage : de 3' en 5' par l'Exosome et de 5' en 3' par l'exonucléase XRN4 (German et al., 2008).

A contrario, le mécanisme de répression de la traduction chez les plantes n'avait été démontré jusqu'à récemment que pour quelques gènes uniquement (p. ex. APETALA2 et miR-172, SPL3 et miR-156/157) (Aukerman and Sakai, 2003; Chen, 2004; Gandikota et al., 2007). De récentes études montrent néanmoins que ce mode de régulation ne serait pas si rare qu'on le pensait (Brodersen et al., 2008; Todesco et al., 2010). Le mécanisme n'ayant pas encore été assez étudié, nous n'en connaissons pour le moment que peu de détails mais nous pouvons tout de même noter qu'aucun rapport ne montre que cette répression soit due à une anti-complémentarité imparfaite entre le miARN et l'ARNm comme chez les métazoaires.

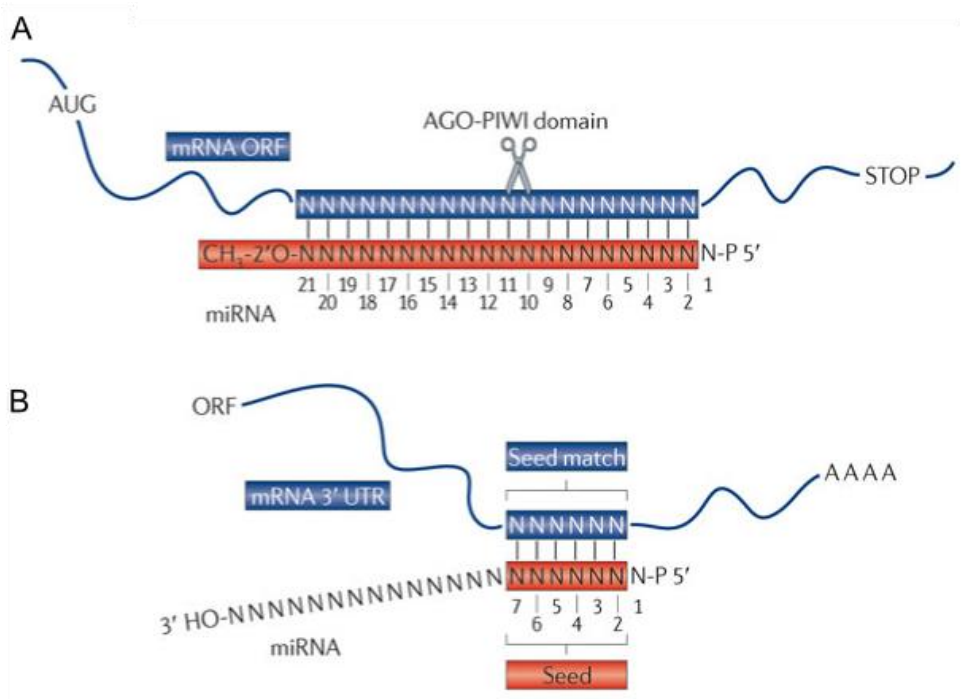


Figure 6. Reconnaissance entre miARN et ARNm par appariements Watson-Crick. A | Reconnaissance chez les plantes avec appariement parfait entre le miARN et son ARNm. Le premier nucléotide du miARN est enfoui dans la protéine AGO. Ce type de reconnaissance parfaite (ou presque parfaite) entraine le clivage par le domaine PIWI d'AGO. B | Reconnaissance chez les animaux avec appariement partiel entre la séquence « seed » du miARN (position 2 à 7) et la 3'UTR de l'ARNm. D'après (Huntzinger and Izaurralde, 2011).

c) Mécanisme chez les métazoaires

Chez les animaux, on pensait jusqu'à présent que les miARN ne reconnaissaient que partiellement leur(s) site(s) de liaison (Figure 6 B) et que ces sites se trouvaient plus généralement dans la partie 3'UTR des gènes – comme ont pu le démontrer Ambros et Ruvkun dans leurs premières expériences. En fait, les liaisons semblent bien plus variées que cela (Doxakis, 2013) et l'existence de sites dans les parties 5'UTR et dans les régions codantes des gènes a également pu être établie (Lytle et al., 2007; Zhou et al., 2009; Hausser et al., 2013). Helwak et coll. ont montré par exemple, et contrairement à ce que l'on pensait, qu'il y aurait en fait une prévalence des liaisons dans les régions codantes des gènes plutôt qu'en 3'UTR mais que ces liaisons en 3'UTR seraient généralement plus stables que les autres (Helwak and Tollervey, 2014). Ces conclusions ne reposent bien évidemment que sur une technique bien particulière et il serait malaisé d'en faire une généralité. En conséquence, nous voyons qu'il est encore aujourd'hui difficile de déterminer *in vivo* les proportions relatives de ces différents sites de liaison.

Un des éléments déterminant de la reconnaissance miARN/ARNm est la séquence « *seed* » : une très courte séquence conservée d'environ 6 nucléotides en 5' du miARN aux positions 2 à 7 (Figure 6 B). Cette séquence bien spécifique est capable à elle seule de mettre en place le *silencing* et est souvent la seule partie du miARN s'appariant effectivement à l'ARNm. Il existe tout de même des exceptions où la partie en 3' du miARN intervient également dans la sélection de cible et, plus souvent, dans la stabilisation de liaison miARN/ARNm notamment aux travers des nucléotides 13 à 16. Dans d'autres cas encore, la *seed* n'intervient pas dans l'interaction (Brennecke et al., 2005; Grimson et al., 2007). Concernant le reste de la séquence du miARN, les positions 9 à 12 font généralement saillie (possiblement pour permettre un clivage de l'ARNm cible par Ago2) alors que le nucléotide en position 1 est enfoui dans la protéine AGO. Ce denier constat est également vrai chez les plantes.

La reconnaissance d'un ARNm par un miARN et son chargement dans le complexe miRISC bloque toutes possibilités de traduction (Figure 7). La traduction est un mécanisme se décomposant en trois phases : l'initiation, l'élongation et la terminaison. Deux éléments essentiels à l'initiation traductionnelle chez les eucaryotes supérieurs sont la coiffe m^7G en 5' et la queue poly-adénylée (poly-A) en 3' des ARNm. En effet, la queue poly-A est reconnue par la PABPC (*Poly(A)-binding protein*), une protéine qui s'associe à eIF4G (*eukaryotic translation-initiation factor 4G*), elle-même associée à la coiffe de l'ARNm via eIF4E ; ces deux protéines associées à eIF4A forment une entité également appelées eIF4F. Dans cette configuration, l'ARNm forme une sorte de boucle permettant l'initiation de sa traduction mais également sa protection contre toute dégradation (Wells et al., 1998). L'intérêt de ces deux protéines ici est leur implication essentielle pour le mécanisme de répression de la traduction.

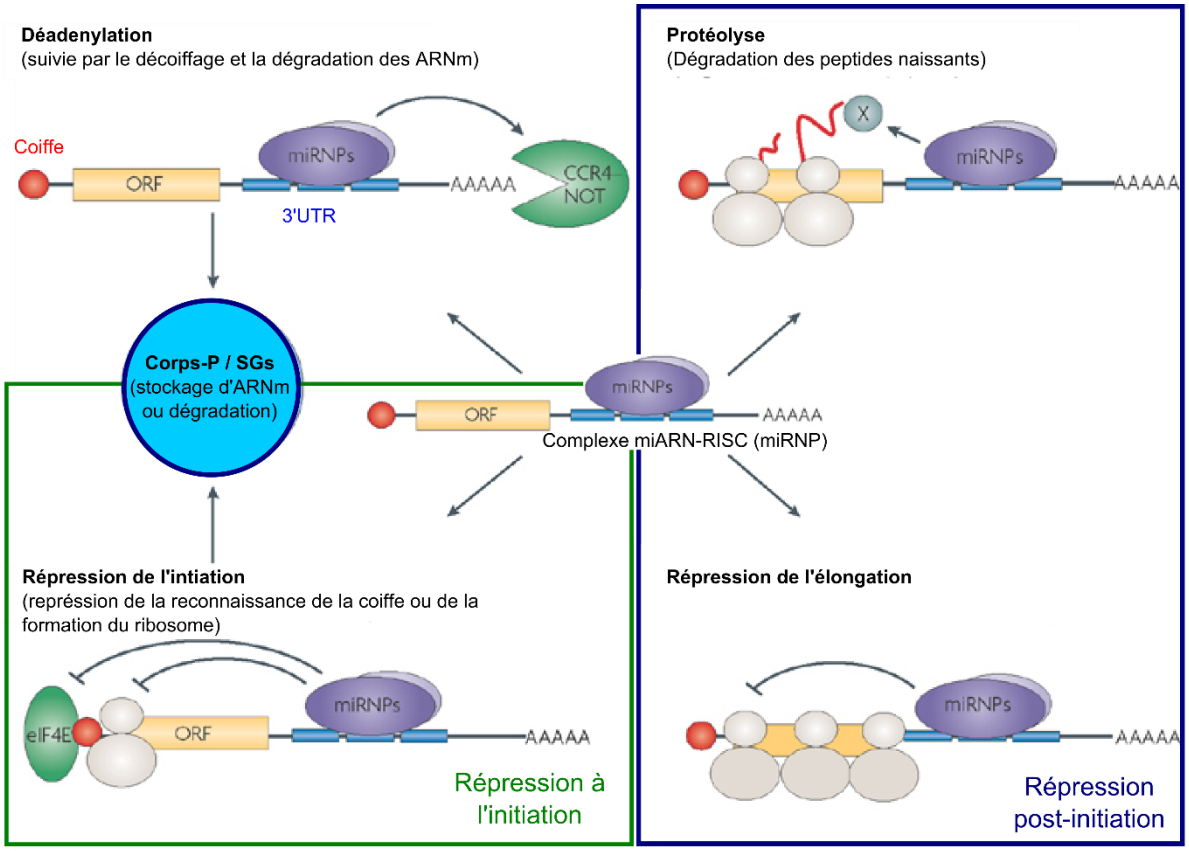


Figure 7. Devenir des ARNm ciblés par les miARN. Après recrutement de l'ARNm par le complexe RISC et son miARN (complexe miRNP – *micro-ribo-nucleo-protein*), l'ARNm est soit déadénylé et dégradé (en haut à gauche). Le complexe miRNP peut également réprimer l'initiation de la traduction par des mécanismes impliquant la reconnaissance de la coiffe ou du ribosome (bas gauche). Dans les deux cas, les ARNm sont envoyés vers les corps-P (ou les granules de stress – SGs). La répression des ARNm peut également se faire à des niveaux post-initiation soit par la protéolyse des peptides naissants (partie haut droite), soit par la répression de l'élongation par dissociation prématurée des ribosomes (partie bas droite). Adaptée d'après (Filipowicz et al., 2008)

En effet, plusieurs auteurs ont pu démontrer des interactions capitales entre la machinerie d'ARN interférence et ces deux dernières. Le mécanisme de répression en lui-même reste tout de même encore controversé puisque des études montrent tout autant des répressions au niveau de l'initiation de la traduction qu'à des stades post-initiation.

(1) Répression à l'initiation

Le principal mécanisme de répression au niveau de l'initiation de la traduction passe par la reconnaissance de la coiffe des ARNm et l'inhibition de la reconnaissance de cette dernière par eIF4F (Figure 7). Paradoxalement, des auteurs ont démontrés que des miARN étaient capables d'inhiber l'expression d'ARNm coiffés mais ne le faisaient pas avec les ARNm avec une coiffe non fonctionnelle, ni en la présence d'IRES (Humphreys et al., 2005; Pillai et al., 2005). Egalement en contradiction avec les deux études citées ci-dessus sur lin-4 et let-7, Ding et Grosshans ont montré que la répression de cibles de let-7 et lin-4 chez *C. elegans* – notamment lin-14 et lin-28 – s'accompagnait d'une diminution de l'association des ARNm aux polysomes (Ding and Grosshans, 2009). Par ailleurs, l'augmentation artificielle de la protéine eIF4F (le complexe se liant à la coiffe) dans des cellules tumorales d'ascite de souris (Krebs-2) permet de supprimer les effets de *silencing* (Mathonnet et al., 2007). D'autres études encore tendent également à confirmer cette hypothèse (Kiriakidou et al., 2007; Thermann and Hentze, 2007; Wakiyama et al., 2007; Zdanowicz et al., 2009), ce qui en fait le mécanisme dominant de la répression pour le moment.

Un mécanisme alternatif a également été proposé dans cette catégorie par Chendrimada et collaborateurs. Ce mécanisme repose sur le blocage du recrutement de la sous-unité 60S du ribosome grâce au recrutement d'eIF6 par miRISC (Chendrimada et al., 2007). EIF6 est une protéine impliquée dans la biogénèse de la grande sous-unité 60S et son transport vers le cytoplasme mais elle empêche surtout toute interaction précoce avec la petite sous unité 40S pour former le ribosome (Figure 7). De par le rôle d'eIF6 dans la maturation de la sous-unité 60S, l'interprétation de leurs résultats reste complexe et le mécanisme est encore controversé.

Néanmoins, une baisse de la densité de ribosome devrait être observée dans les deux cas puisque les deux hypothèses empêcheraient globalement la fixation des ribosomes sur l'ARNm.

(2) Répression post-initiation

La répression post-initiation est historiquement la première hypothèse ayant été formulée pour expliquer l'inhibition de la traduction des ARNm. Lors des premières expériences sur lin-4 chez *C. elegans*, il avait été observé que l'expression des protéines LIN-14 et LIN-28 était réprimée sans impact sur le niveau de leur ARNm respectif. Ces ARNm avaient par ailleurs été repérés dans les polysomes (unité fonctionnelle de plusieurs ribosomes attachés sur un même ARN), indiquant donc que la répression se mettait en place après l'étape d'initiation de la traduction, c'est-à-dire, après fixation des ribosomes sur l'ARNm (Olsen and Ambros, 1999; Seggerson et al., 2002). Il a également été montré que des miARN étaient capables de réprimer des ARNm indépendamment de la coiffe et au travers d'IRES (*internal ribosome entry site*), ajoutant ainsi du crédit à ce type de régulation post-initiation (Petersen et al., 2006; Lytle et al., 2007).

Principalement deux hypothèses de répression post initiation sont aujourd'hui proposées (Figure 7) : i) une dégradation des protéines conjointement à la traduction au travers de protéases non identifiées et recrutées par RISC (Nottrott et al., 2006) et ii) une dissociation prématurée des ribosomes pendant l'élongation à cause de la présence de miARN et de miRISC (Petersen et al., 2006). Nous pouvons noter que ces deux hypothèses sous-tendent des conséquences sur la densité de ribosomes « attachés » à un ARNm : dans le premier cas, la densité de ribosome ne devrait pas changer mais rester constante alors que dans le deuxième cas, la densité de ribosomes devrait diminuer.

(3) Dégradation des ARNm

A l'instar des plantes, il existe quelques cas où une complémentarité parfaite de séquence entre ARNm et miARN entraine clivage et dégradation chez les animaux (Yekta et al., 2004). Ces cas de complémentarité parfaite restent cependant rares dans ce règne bien

que la dégradation des ARNm *en soi* semble ne pas l'être. En effet, plusieurs études combinant des approches de protéomique et de transcriptomique à haut débit menées ces dernières années montrent que la dégradation des ARNm par le mécanisme d'ARN interférence n'est pas l'exception à la règle mais au contraire, le mécanisme dominant malgré des complémentarités de séquences partielles (Baek et al., 2008; Selbach et al., 2008; Hendrickson et al., 2009; Guo et al., 2010). L'ensemble de ces études montrent en fait que l'impact sur l'expression des protéines serait essentiellement dû à la déstabilisation des ARNm. Certains auteurs montrent que la déstabilisation compterait ainsi pour 66 à 90% des cas, toutes situations confondues (Guo et al., 2010; Eichhorn et al., 2014).

Il est également intéressant de noter que deux de ces études tendent à confirmer l'hypothèse de répression à l'initiation plutôt qu'à des stades post-initiation. En effet, au travers d'approche permettant une localisation précise des ribosomes sur l'ARNm, ces auteurs montrent que les miARN ont un effet négatif sur la densité de ribosomes attachés aux ARNm cibles mais également que la diminution de la densité de ribosomes se fait tout au long du messager et non à des endroits spécifiques. Ces résultats semblent donc infirmer les deux hypothèses de répression post-initiation (Hendrickson et al., 2009; Guo et al., 2010).

La déstabilisation des ARNm réprimés par les miARN est un mécanisme encore peu compris. L'hypothèse la plus probable est que cette dernière passe d'abord par la déadénylation de la queux poly-A des ARNm grâce au complexe de déadénylation CAF1-CCR4-NOT1 (Figure 7) – recruté par la protéine GW182 – puis par la suppression de la coiffe par DCP2 (Behm-Ansmant et al., 2006; Eulalio et al., 2007b). Une fois déadénylé et décoiffé, l'ARNm est dégradé de 5' en 3' par l'exonucléase XRN1 (Rehwinkel et al., 2005; Fabian et al., 2009; Piao et al., 2010; Zekri et al., 2013). Il n'y a donc *a priori* pas de clivage direct suivi d'une dégradation des deux parties de l'ARN comme chez les végétaux mais simplement un passage par une des voies de dégradation classique chez les eucaryotes.

En règle générale et quel que soit le mécanisme considéré, les ARNm non traduits s'accumulent dans des foci granulaires cytoplasmiques appelés corps-P (*Processing-bodies*,

P-bodies ou encore *GW-bodies*). En cas de stress ou de répression général de l'initiation de la traduction pour la cellule, cette accumulation peut également se faire dans des agrégats appelés « granule de stress » (*stress granules* ou SGs) (Mazroui et al., 2006; Balagopal and Parker, 2009) (Figure 7). Les corps-P sont des structures subcellulaires impliqués dans la dégradation des ARNm (Eulalio et al., 2007a) mais sont également des sites de stockage temporaires pour les ARNm réprimés (Brengues et al., 2005). De nombreuses expériences prouvent que ces deux types de granules sont enrichis à la fois en miARN, en protéines AGO et en ARNm réprimés par les miARN (Jakymiw et al., 2005; Liu et al., 2005a, 2005b; Leung et al., 2006). Par ailleurs et à l'inverse des corps-P, l'accumulation de protéines AGO dans les SGs semble être dépendante des miARN (Leung et al., 2006). Il est donc vraisemblable que les corps-P et les SGs soient responsables du devenir des ARNm réprimés par les miARN, bien qu'aucune expérience ne le prouve directement à ce jour.

4. Système d'annotation des miARN

Les premiers miARN ont été découverts et analysés principalement par le clonage d'ADNc (ADN complémentaire) (Ambros et al., 2003). Aujourd'hui, la découverte de nouveaux miARN passe essentiellement par le séquençage haut débit et l'analyse bioinformatique (Motameny et al., 2010; Kozomara and Griffiths-Jones, 2011; Sun et al., 2014). Chaque nouveau miARN est premièrement annoté selon un système qui remonte à l'article d'Ambros, Ruvkun et Tuschl (Ambros et al., 2003), puis ajouté à la base de données miRBase (Griffiths-Jones, 2004).

Dans la base de données, chaque nouveau miARN (quel que soit l'espèce) est incrémenté par un identifiant unique. Il en existe deux types : le MI (pour miARN pré-mature) et le MIMAT (pour miARN mature) composés tous deux d'une succession de 7 ou 6 chiffres (MI0000001, pour un miARN non mature et MIMAT000001, pour un miARN mature p.ex.). Ces deux identifiants permettent essentiellement d'éviter la présence de doublon dans la base de données.

Cependant, les miARN sont plus largement reconnus par un autre type de notation prenant en compte l'espèce où a été retrouvé le miARN et un numéro unique représentant approximativement le moment où le miARN a été découvert et/ou sa proximité évolutionnaire avec d'autres miARN connus dans d'autres espèces. Ainsi, les trois premières lettres du nom du miARN font référence à l'espèce dans laquelle ce dernier a été retrouvé (hsa pour *Homo sapiens* p.ex.). Ces trois lettres sont suivies des termes miR pour les miARN mature (hsa-miR) et mir pour les gènes (hsa-mir). Un numéro est ajouté à la suite de ces deux informations de manière généralement séquentielle, sauf lorsqu'il existe une identité de séquence dans une autre espèce où le même numéro est alors utilisé (hsa-miR-121 ou mmu-mir-121 p.ex.). En présence de précurseurs différents ou de loci génomiques exprimant le même miARN mature, un chiffre est ajouté à la suite du nom (hsa-mir-121-1 et hsa-mir-121-2 p.ex.). En cas de très grande proximité de séquence mature, une lettre est directement accolée au numéro du miARN (hsa-miR-121a et hsa-miR-121b p.ex.). Enfin et comme déjà exposé, si un gène de miARN donne naissance à deux miARN matures pouvant tous deux être chargés dans miRISC, les termes 5p (issus du brin 5' du pre-miARN) et 3p (issus du brin 3') sont accolés au miARN (hsa-miR-142-5p et hsa-miR-142-3p p.ex.)

Il existe quelques exceptions à ces règles, notamment pour les miARN let-7 et lin-4, à cause de leur caractère historique mais également pour certains miARN issus d'autres organismes (plantes ou virus p.ex.).

5. Découverte de cibles

Suite à la découverte des premiers miARN, la recherche de cibles de ces régulateurs a très rapidement été un sujet de recherche bouillonnant. Les premières méthodes *in silico* sont apparues dès 2003, notamment pour la drosophile (Enright et al., 2003; Stark et al., 2003; Rajewsky and Socci, 2004) et les mammifères (Lewis et al., 2003; Kiriakidou et al., 2004). Ces études, ainsi que celles qui les ont suivies, ont établi les règles de liaison entre les deux entités afin de prédire les couples potentiels de miARN et d'ARNm. C'est principalement à cause de

l'absence de méthodes haut-débit de caractérisation des cibles des miARN que de nombreux algorithmes différents voient le jour à cette époque. Avec l'avènement des méthodes de séquençage de dernière génération et de la protéomique, ces limites semblent pourtant s'estomper et quelques études montrent enfin des caractérisations d'ensemble du *targetome* (l'ensemble des cibles) des miARN. Cependant, ce type d'expérience reste encore assez cher et lourd à mettre en œuvre rendant encore aujourd'hui l'utilisation des algorithmes de prédiction indispensable.

a) Découverte de cibles *In silico*

Il existe plus d'une dizaine d'algorithmes de prédiction différents pour l'être humain (Tableau 1). En effet, si une recherche de cibles par complémentarité parfaite entre ARNm et miARN chez la plante donne de très bons résultats, ce n'est pas le cas chez les mammifères (Rhoades et al., 2002) – si bien qu'il a fallu très rapidement rajouter un niveau de complexité aux algorithmes de prédiction dans ce règne. Chaque algorithme cherche ainsi à modéliser les interactions miARN/ARNm en se basant sur un ensemble de règles plus ou moins complexes. Ces règles peuvent être classées globalement en deux catégories : i) les règles communes, qui sont des éléments repris par tout un ensemble d'algorithmes et ii) les règles spécifiques à certains algorithmes.

Sans surprise, la première des règles communes est la complémentarité de séquences entre le miARN et l'ARNm, et plus particulièrement pour ce qui concerne la séquence *seed* du miARN. Cette complémentarité antiparallèle est aujourd'hui essentiellement envisagée par des liaisons Watson-Crick canoniques (Adénosine-Uracile et Guanine-Cytosine). La définition de la *seed* peut varier quelque peu en fonction de l'algorithme : si la plupart des études la considèrent comme un hexamère (6mer : position 2 à 7 du miARN), d'autres prennent également en compte des heptamères (7mer : position 2 à 8 ou 1 à 7) ou encore des octomères (8-mer : position 1 à 8 ou 2 à 9) (Lewis et al., 2003; Brennecke et al., 2005; Krek et al., 2005).

Tableau 1. Résumé d'outils de prédiction de cibles. D'après (Peterson et al., 2014)

Nom	Site Web	Utilisation en ligne	Disponibilité du code source	Ajustable par l'utilisateur	Nécessité pour l'utilisateur de fournir des données	Niveau d'utilisation
miRanda	http://www.microrna.org/		X	X	Séquences	Avancé
miRanda-mirSVR	http://www.microrna.org/	X				Tous
TargetScan	http://www.targetscan.org	X				Tous
DIANA-microT-CDS	http://www.microrna.gr/microT-CDS	X				Tous
MirTarget2	http://mirdb.org	X		X		Tous
RNA22-GUI	https://cm.jefferson.edu/rna22v1.0/	X				Intermédiaire
TargetMiner	http://www.isical.ac.in/~bioinfo_miu/targetminer20.htm	X	X		Fichier d'input	Intermédiaire
SVMicrO	http://compgenomics.utsa.edu/svmicro.html	X	X		Séquences	Expert
PITA	http://genie.weizmann.ac.il/pubs/mir07/	X	X	X		Tous
RNAhybrid	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	X	X	X	Séquences	Avancé

Une autre règle commune à plusieurs algorithmes est la conservation de séquences au sein de différentes espèces. Cette conservation (généralement « parfaite ») peut porter aussi bien sur les sites de liaison des ARNm que sur les miARN eux-mêmes. Lewis et al. ont par ailleurs montré que la séquence *seed* était la partie la plus conservée des miARN dans l'évolution (Lewis et al., 2003), etc. La troisième règle commune est l'énergie thermodynamique de liaison (énergie de Gibbs) entre l'ARNm et le miARN puisque plus une liaison est stable plus elle a de chance d'exister (Maragkakis et al., 2009a; Yue et al., 2009). Cette énergie de liaison est généralement calculée *in silico* grâce au *package* de repliement Vienna (Hofacker et al., 1994). Enfin, l'accessibilité des sites de liaison constitue la dernière règle commune. Cette dernière considère les structures secondaires prédites des ARNm, qui pourraient potentiellement empêcher toutes interactions avec les miARN (Mahen et al., 2010).

Les règles spécifiques sont bien plus nombreuses que les règles communes et sont généralement limitées à quelques algorithmes uniquement. Elles servent usuellement de critères secondaires afin de réduire le nombre de faux positifs ou afin de donner un score aux interactions. Quelques exemples, non exhaustifs, de règles non communes sont :

- Les appariements *wobble* (littéralement « appariements bancals ») : autorise une liaison entre une guanine et une uracile dans la région seed (Doench and Sharp, 2004)
- Les appariements non canoniques basés sur la prédiction de structures tridimensionnelles : prend en compte la plus grande flexibilité de liaison ARN/ARN (Leontis, 2002) et ainsi tout autre appariement que A-U et G-C (Gan and Gunsalus, 2013)
- L'abondance de sites cibles : correspond au nombre de sites potentiels de liaison pour un même ARNm (Garcia et al., 2011)
- La composition locale en AU : correspond au pourcentage d'adénines et d'uraciles de part et d'autre de la séquence *seed* (Friedman et al., 2009b; Betel et al., 2010)
- Les sites compensatoires en 3' : l'appariement dans la région 3' des miARN permettant une plus grande stabilité de liaison, exposée plus haut (Friedman et al., 2009b)
- L'énergie de liaison de la *seed* uniquement (Garcia et al., 2011)
- La contribution de position : considère la position du site de liaison sur l'ARNm (Grimson et al., 2007)
- Les approches d'apprentissage : concerne l'utilisation de méthodes d'apprentissage supervisées ou non. Certaines méthodes se basent uniquement sur des règles de liaison ARNm/miARN (Saetrom et al., 2005; Kim et al., 2006; Yousef et al., 2007; Sturm et al., 2010), alors que d'autres utilisent par exemple des données d'expression (Wang and El Naqa, 2008; Bandyopadhyay and Mitra, 2009)
- La recherche dans d'autres régions que les 3'UTR : notamment en 5'UTR et dans les régions codantes des ARNm (Reczko et al., 2012)

Enfin, certaines méthodes cherchent à combiner l'information d'autres algorithmes soit en considérant simplement des intersections et des unions de tables de prédiction (Shirdel et

al., 2011), soit en combinant certaines étapes clés de chaque algorithme et en y ajoutant d'autres règles et/ou informations. Coronnello *et al.* reprennent par exemple des éléments de miRanda (Enright et al., 2003), PITA (Kertesz et al., 2007) et TargetScan (Lewis et al., 2003) tout en ajoutant une information basée sur des données d'expression de miARN chez *D. melanogaster* (Coronnello and Benos, 2013). Zhou et al. reprennent, quant à eux, les deux algorithmes miRanda et TargetScan et les appliquent aux 5'UTR ainsi qu'aux régions codantes pour étendre les prédictions initiales des deux algorithmes (Zhou et al., 2009). Dans la suite de cette section, trois algorithmes seront analysés plus en détails, notamment DIANA-microT version 3 (Maragkakis et al., 2009a) et TargetScan version 6 (Lewis et al., 2005) – qui ont tous deux été utilisés tout au long des travaux de cette thèse. Nous aborderons également l'algorithme miRANDA, essentiellement à cause de son caractère pionniers dans la prédiction de cibles.

(1) TargetScan

TargetScan est probablement un des algorithmes de prédiction de cibles les plus utilisés à l'heure actuelle. Développé à la même époque que miRanda (Enright et al., 2003), cet algorithme n'a pourtant jamais cessé d'être amélioré année après année (Lewis et al., 2003, 2005; Grimson et al., 2007; Friedman et al., 2009a; Garcia et al., 2011). La version en ligne contient des prédictions de cibles chez l'être humain mais également la souris, le ver *C elegans*, la drosophile et le poisson. L'algorithme se base aujourd'hui sur quelques règles communes et plusieurs règles spécifiques afin non seulement de prédire les cibles mais également de leur donner un score.

La première étape de l'algorithme consiste à trouver des couples potentiels miARN-ARNm en recherchant : 1) des appariements entre la séquence *seed* des miARN et la 3'UTR des gènes (8mer et 7mer), 2) des conservations de site de liaison des gènes cibles dans différentes espèces (humain, chien, poulet, rat et souris), 3) le pourcentage de GC pour les zones appariées et 4) des conservations de la séquence seed dans différentes espèces. Il y a donc deux types de « conservations » différentes : la conservation des miARN (par la *seed*)

et la conservation de gènes (par les sites de liaison). TargetScan introduit par ailleurs deux types de 7mer : le 7mer-A1 (appariement parfait de la position 2 à 7 plus une adénine en position 1 du miARN) et le 7mer-m8 (appariement parfait de la position 2 à 8).

L'algorithme se base sur trois niveaux de conservation pour les familles de miARN (conservation des miARN) : « globalement conservé » (présent chez les vertébrés), « conservé » (présent chez les mammifères) et « non conservé » (les autres cas). Dans ce cas, l'algorithme considère uniquement des alignements parfaits de séquences *seed* (séquences identiques).

Pour la conservation de sites de liaison en revanche, il n'y a que deux niveaux de conservation : « conservé » et « non conservé ». Pour définir ces deux niveaux, l'algorithme examine une longueur de branche d'un arbre phylogénique construit à partir d'alignements des sites de liaison d'un même gène chez différentes espèces. Un seuillage sur la longueur des branches et dépendant du type de sites permet la classification (8-mer : seuil de 0.8, 7-mer-m8 : seuil de 1.3 et 7mer-1A : seuil de 1.6).

En se basant sur ces différents états de conservation et le nombre de sites pour un même gène, l'algorithme attribut un premier score représentant la probabilité de ciblage conservé (P_{CT}) entre un miARN et une cible. Nous pouvons noter que ce calcul est effectué uniquement pour les familles de miARN conservés.

Pour le calcul de $P_{CT,}$ un ratio signal sur bruit (S/B) est premièrement calculé pour chaque site de liaison et à chaque seuil de longueur de branche, à partir d'un ensemble négatif de sites comme établi dans (Friedman et al., 2009a). Le score par site est alors défini comme l'estimation Bayésienne de la probabilité qu'un site de liaison soit conservé en raison de la pression de sélection du ciblage par miARN, plutôt que par chance (ou tout autre raison n'impliquant pas les miARN) :

$$P_{CTs} \approx \frac{\frac{S}{B} - 1}{\frac{S}{B}}$$

Un P_{CT} par cible en fonction du nombre de sites de liaison pour un même gène est alors calculé de la manière suivante :

$$P_{CT} = 1 - [(1 - P_{CTs})_{site1} \times (1 - P_{CTs})_{site2} \times (1 - P_{CTs})_{site3} \times, etc.]$$

où un P_{CT} de 1 montre une très forte probabilité de ciblage conservé et 0, une très faible probabilité.

Après cette première étape, un deuxième score – appelé *context+ score* – est calculé à partir de critères secondaires comme le type d'appariements observés, l'enrichissement local en AU, la complémentarité en dehors de la région *seed* (3'UTR) et la distance du site de liaison par rapport à la fin de la 3'UTR (Grimson et al., 2007; Marson et al., 2008). L'abondance en sites de liaison (TA) et la stabilité de liaison de la *seed (*SPS) sont les deux derniers critères ajoutés (Garcia et al., 2011). L'étape de calcul pour ce second score est dérivée de données expérimentales après expression ectopique de différents miARN dans des cellules HeLa. Pour chacun des six critères, un score unique est estimé de façon indépendante par régression linéaire à partir des données citées ci-dessus. Ce score représente la « contribution » de chaque critère. Le *context+ score* est alors défini comme la somme des contributions des différents éléments. Il représente en quelque sorte la probabilité pour une cible donnée d'être réellement une cible, attendu qu'il existe une anticorrélation entre l'expression d'un miARN et ses gènes cibles : ainsi plus le score est faible, plus cette probabilité est favorable.

(2) DIANA-microT

DIANA-microT est un algorithme légèrement plus récent puisque sa version publiée n'est apparue qu'en 2009 (Maragkakis et al., 2009a). Il reste cependant – dans ces toutes premières versions – un des premiers à avoir proposé des cibles chez l'être humain. Également mis à jours régulièrement, l'avantage de cet algorithme réside essentiellement dans le site web qui l'accompagne. En effet, Diana Tools est une référence quant à l'intégration d'outils liés à l'analyse des miARN (Maragkakis et al., 2009b). Le site regroupe ainsi des données de prédictions de cibles (microT, TargetScan, PITA, etc.), des données de validation

de cibles (TarBase), des méthodes d'analyses systémiques (mirPath), des pipelines d'analyses automatiques d'expériences, etc.

Dans sa version 3, l'algorithme débute par la recherche d'éléments de reconnaissance des miARN (MRE pour *miRNA Recongition* Element) dans les 3'UTR des gènes à partir de la seed (9-mer, position 1 à 9 du miARN). Les alignements sont alors classés en deux catégories en fonction du nombre d'appariements observés : en présence de moins de 7 nucléotides appariés ou de paires wobble, les liaisons MRE:*seed* sont qualifiées de faibles alors qu'on les considère comme fortes dans le cas contraire.

Les liaisons fortes passent directement à la suite de l'algorithme. Dans le cas de liaisons faibles en revanche, une MRE n'est pas directement rejetée mais une énergie de liaison de l'hétéroduplexe est d'abord calculée. La MRE est considérée comme potentielle uniquement lorsque cette énergie est inférieure à un seuil prédéfini.

Dans la suite de l'algorithme, un score est alors attribué à chaque MRE en fonction du type d'appariement observé pour un même gène (6-mer, 7-mer, 8-mer, 9-mer, etc.), mais également de la conservation du MRE chez différentes espèces. Dans ce cas, la conservation est définie comme une identité parfaite de MRE dans plusieurs espèces (séquences identiques) et le score de conservation est le nombre de fois que la MRE est retrouvée conservée (par exemple 3, si retrouvé dans trois espèces).

Le score MRE à proprement parlé est calculé pour un miARN r, une catégorie de liaison b et un score de conservation supérieur ou égal à c par la formule suivante :

$$R_{r,b}(c) = 60 \times \frac{N_{r,b}(c)}{\sum_{m=1}^{60} M_{r,m,b}(c)}$$

où $N_{r,b}(c)$ correspond au nombre de MRE pour un vrai miARN et $M_{r,m,b}(c)$ au nombre de MRE pour un ensemble de miARN aléatoires, m étant l'index du miARN aléatoire. Les miARN aléatoires sont des séquences artificiellement créées pour l'occasion et ayant à peu près le même nombre de MREs correspondant au vrai miARN.

La somme pondérée de toutes les MREs pour un même gène forme alors le score final appelé score « miTG ». Ce score représente donc en quelque sorte la probabilité qu'un gène soit effectivement une cible potentielle d'un miARN : plus ce score est élevé, plus la probabilité est forte. Un score supérieur à 7 est considéré comme un bon score.

Pour le calcul des poids de la somme formant le score miTG, un premier ratio sensiblement identique au précédent et basé sur 75 vrais miARN conservés chez l'être humain, le chimpanzé, la souris, le rat, le chien et le poulet est calculé pour chaque catégorie b et score de conservation c, d'après la formule suivante :

$$R'_b(c) = 5 \times \frac{\sum_{r=0}^{r=75} N_{r,b}(c)}{\sum_{r=1}^{75} \sum_{m=1}^{5} M_{r,m,b}(c)}$$

On considère ici 5 faux miARN pour chaque vrai miARN, c'est-à-dire 375 faux miARN en tout. Les poids pour chaque catégorie de sites sont alors évalués d'après la pente d'un ajustement linéaire estimé par la méthode des moindres carrés entre les ratios R'$_b$ et leur score de conservation c correspondant (Figure 8).

La dernière version de DIANA-microT (DIANA-microT-CDS ou v5) (Reczko et al., 2012) considère également les régions codantes en plus des 3'UTR mais intègre surtout des approches d'apprentissage supervisées (modèle linéaire généralisé ici) basées sur des données de PAR-CLIP (*photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation*) – une méthodologie de découverte de cibles de miARN qui sera abordée dans la partie I.B.5.b) (page 35). En se basant sur des données d'expression, la dernière version intègre également un apprentissage sur la contribution de sites multiples et ajoute ainsi

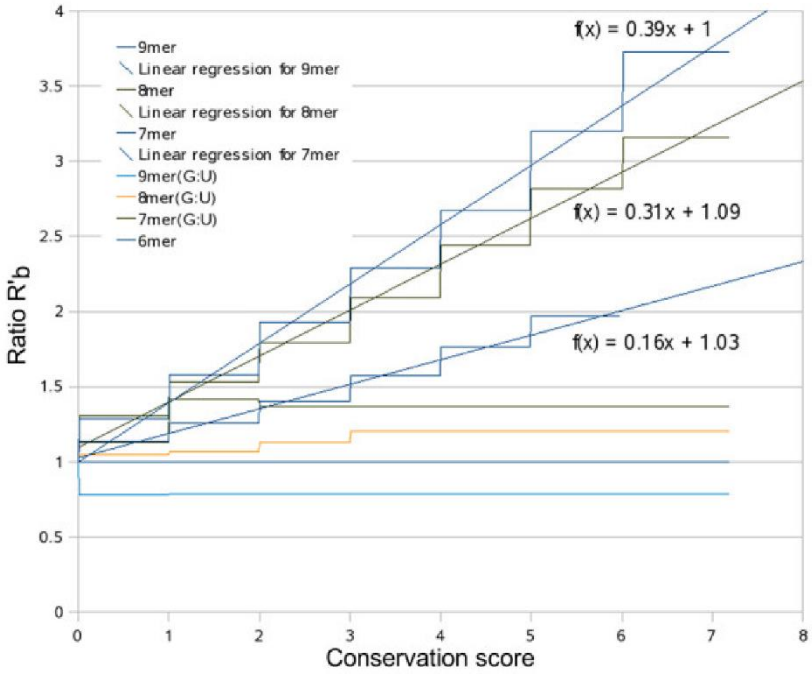


Figure 8. Calcul du poids des catégories de sites dans DIANA-microT.
Tirée de (Maragkakis et al., 2009a).

différentes règles non communes qui n'étaient pas présentes initialement: composition en AU, accessibilité de la 3'UTR, distance à la fin de l'UTR de la région codante etc. Malheureusement, cette version n'ayant été mise en ligne que très tardivement malgré sa date de publication, l'ensemble des travaux de thèse a été réalisé sur la première version publiée de l'algorithme.

(3) miRANDA

Développé en premier lieu pour la drosophile mais rapidement étendu à l'être humain, miRANDA est un des premiers algorithmes à avoir vu le jour (Enright et al., 2003; John et al., 2004). L'algorithme se base sur trois étapes séquentielles : 1) un *scan* des 3'UTR des ARNm pour trouver des sites potentiels de liaison contre un miARN (appariements WC pour l'ensemble du miARN). 2) Le calcul d'une énergie de liaison entre les couples ARNm/miARN avec un nombre de nucléotides appariés supérieurs à un seuil prédéfini. 3) La recherche de conservation de séquences pour les couples dont l'énergie de liaison est inférieure à un seuil prédéfini. Un score est enfin attribué à chaque couple en se basant sur le nombre de nucléotides appariés. L'algorithme prend en compte le nombre de sites de liaison, un score élevé peut donc faire référence aussi bien à un appariement unique et parfait ou plusieurs

appariements imparfaits. La simplicité de ces trois étapes successives et leur pertinence fait qu'ils sont très souvent repris dans la littérature pour la prédiction de cibles des miARN.

En 2010, une version améliorée de l'algorithme combinant deux approches a été mise en ligne. Appelée miRanda-mirSVR (Betel et al., 2010), cette combinaison utilise tout d'abord miRanda pour trouver des couples potentiels miARN/ARNm et, dans un second temps, mirSVR pour donner un score à ces prédictions. MirSVR est un algorithme d'apprentissage basé sur la SVR (*support vector regression*) et entrainé sur des données d'expression après transfection de miARN dans des cellules HeLa. Le rôle de mirSVR est donc de prédire un score représentant l'effet d'un miARN sur l'expression des ARNm en se basant sur la conservation de séquence, l'appariement de *seed*, l'accessibilité du site de liaison, la taille de la 3'UTR, etc.

En règle générale, chaque algorithme possède son propre site web et ses tables de prédictions. Ces dernières peuvent généralement soit être consultées, soit être directement téléchargées sous différents formats depuis le site web en question. Par ailleurs, certains sites proposent en plus des tables de prédiction, les codes sources des algorithmes ayant servi à identifier les cibles potentielles. Enfin, il existe quelques sites qui cherchent à regrouper toutes ces informations en un seul lieu, c'est le cas notamment de miRWalk (Dweep et al., 2011). Nous pouvons noter que ce dernier site ne se limite pas à regrouper des informations mais introduit également un nouvel algorithme que nous n'aborderons pas ici. Bien que ce type de démarches soit intéressant pour la communauté scientifique dans son ensemble, elles affichent toutefois du retard – parfois très conséquent – par rapport aux différentes mises à jour des algorithmes (par exemple, miRWalk n'intègre à l'heure actuelle que la version 5.1 de TargetScan).

Il est estimé aujourd'hui que la précision des algorithmes de prédiction les plus utilisés tourne aux alentours de 50-70%, avec DIANA-microT et TargetScan souvent en tête (Selbach et al., 2008; Alexiou et al., 2009). Le recouvrement entre les algorithmes étant généralement assez faible, certains auteurs déconseillent d'utiliser les intersections de différents algorithmes

pour limiter le nombre de faux positif (Witkos et al., 2011) et bien que ces algorithmes aient fait beaucoup de progrès ces dernières années, il est tout de même nécessaire de valider (ou d'invalider) ces prédictions *in vitro* (voir même *in vivo*).

b) Découverte de cibles *In vitro*

Les méthodes de découvertes de cibles *in vitro* ont ainsi pour but de découvrir les vraies cibles biologiques des miARN. Il existe différentes approches qui peuvent être globalement décomposées en 4 catégories : les constructions reportrices, les approches biochimiques, les méthodes –omiques (transcriptomique et protéomique) et enfin toutes les approches moléculaires. Il existe plusieurs bases de données dédiées répertoriant ce type d'information : nous pouvons par exemple citer DIANA-TarBase, mise à jour en ce début d'année (Vlachos et al., 2015) ou encore miRTarBase (Hsu et al., 2011).

(1) Constructions reportrices

La méthode la plus commune de vérification de cibles est le clonage d'un site de fixation potentiel en aval d'un gène rapporteur (généralement la luciférase). Après transfection de ces constructions, les cellules exprimeront de manière constitutive le gène rapporteur. Une baisse de l'activité du gène indiquera qu'un miARN s'est lié au site de liaison (Figure 9).

Plusieurs approches peuvent être considérées : i) seule la construction site de fixation/gène rapporteur est transfectée, dans ce cas on cherche essentiellement à savoir si le site de liaison est correct ; ii) la construction et un inhibiteur de miARN (LNA – *locked-nucleic-acid,* miARN éponge ou antagomirs p.ex.) sont co-transfectés, c'est alors un lien direct entre site de liaison et miARN qui est recherché et iii) la construction et un agoniste de miARN (miARN *mimic* ou construction avec le pre-miARN p.ex.) sont co-transfectés : dans ce dernier cas, c'est également un lien direct entre site de liaison et miARN qui est recherché, avec l'avantage toutefois d'avoir une réponse dose-dépendante. Idéalement, la même construction mutée pour le site potentiel de fixation du miARN sera utilisée comme contrôle. Il est également possible d'envisager un clonage d'une UTR complète ou d'un ensemble de sites

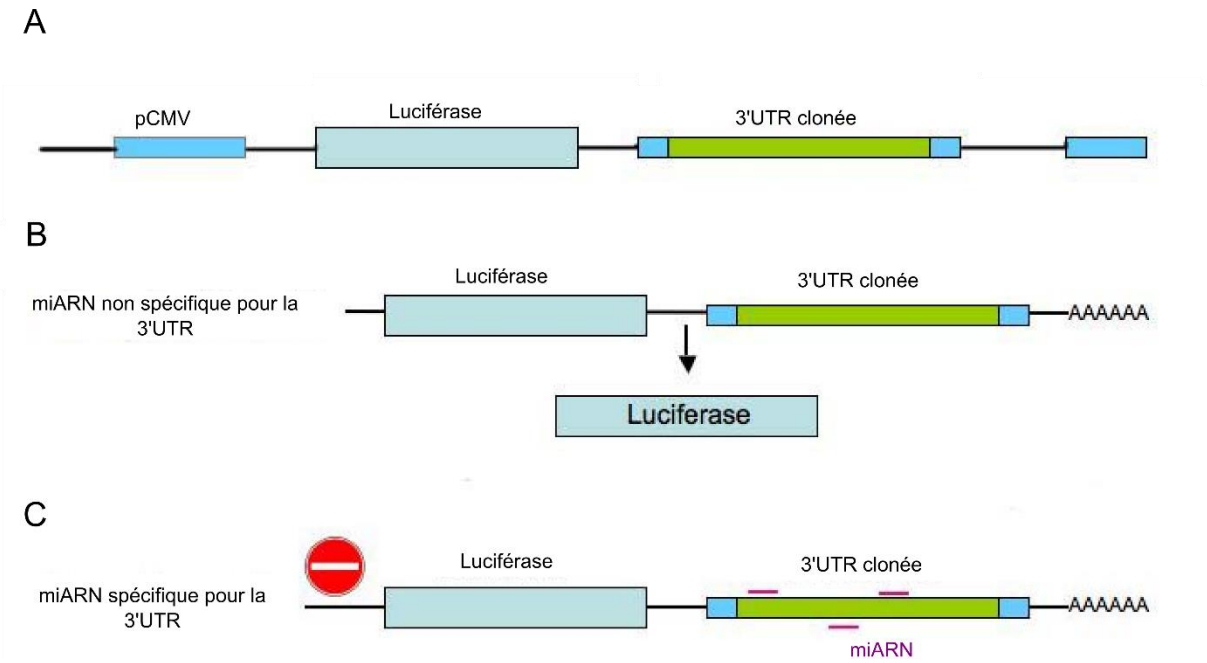


Figure 9. Expérience de gène rapporteur. A | Création du gène chimérique avec un promoteur constitutif (ici celui du cytomégalovirus – CMV), la luciférase et d'une région 3'UTR. B | En présence d'un miARN non spécifique pour la 3'UTR, le gène est transcrit. C | En présence d'un miARN spécifique pour la 3'UTR, la traduction du gène est inhibée et l'activité luciférase est réprimée. Adaptée d'après (Georgantas et al., 2007).

de liaison pour vérifier des effets synergiques ou comprendre des motifs de régulation complexe.

Il est bien connu que beaucoup de facteurs influencent la régulation par les miARN, c'est notamment le cas du type cellulaire ou encore de l'état de différentiation des cellules. Quel que soit le cas, il est donc nécessaire de définir correctement le type cellulaire envisagé lors des expériences initiales. De façon très intéressante, Hwang et al. ont également montré qu'il existait une corrélation entre l'activité ARN interférence liée aux miARN et la densité cellulaire : plus la densité est élevée, plus l'activé des miARN est forte (Hwang et al., 2009). Ce résultat fait resurgir une difficulté supplémentaire lors de l'examen des résultats sur cellules cultivées puisque bien souvent, la densité cellulaire n'est pas prise en compte lors de l'analyse de l'activité du gène rapporteur.

(2) Approches biochimiques

Les approches biochimiques sont des techniques de précipitation généralement haut débits et basées sur le *pull-down* des protéines impliquées dans l'ARN interférence, ou autrement dit, les protéines membres de miRISC comme les protéines AGO principalement.

L'objectif est de retenir le complexe miRISC avec l'ARNm et le miARN afin de déterminer les couples miARN/ARNm.

La technique HITS-CLIP (*high-throughput sequencing to crosslinking immunoprecipitation*) par exemple est une méthode où un lien (réticulation ou *cross-linking*) entre les protéines AGO et le duplex miARN/ARNm est créé par irradiation aux ultra-violets. Après isolation (p.ex. par l'utilisation d'anticorps anti-AGO) et séparation du complexe lié, le séquençage par des techniques de dernière génération ainsi qu'une analyse bioinformatique permet l'identification de la cibles des miARN (Figure 10) (Licatalosi et al., 2008; Chi et al., 2009). Cette méthode a par exemple permis de mettre en évidence des sites de liaison dans les régions codantes à hauteur de 25% et dans les régions introniques à 12% (Chi et al., 2009).

La méthode PAR-CLIP (*Photo-activatable-Ribonucleoside-Enhanced Crosslinking and* Immunoprecipitation) est une version améliorée du HITS-CLIP où le lien covalent est photo-activable, ce qui permet de réduire les réticulations non-spécifiques. Cette technique a été utilisée plus récemment par Hafner et al. dans des cellules HEK293 (Hafner et al., 2010).

Afin de rendre les étapes d'analyse plus simples, la méthode CLASH (*cross-linking ligation and sequencing of hybrids*), quant à elle, reprend les mêmes étapes initiales que la technique HITS-CLIP citée ci-dessus (Kudla et al., 2011; Helwak and Tollervey, 2014). En revanche, la purification s'effectue en tandem, ce qui élimine premièrement le besoin d'anticorps de très haute affinité et permet également une meilleure purification. Cette étape est suivie d'une ligation *in vitro* des duplex miARN-ARNm et de leur séquençage. La ligation permet de s'affranchir de certaines étapes bioinformatiques de déconvolution puisque le miARN et l'ARNm interactants ne forment qu'une seule entité (Figure 10).

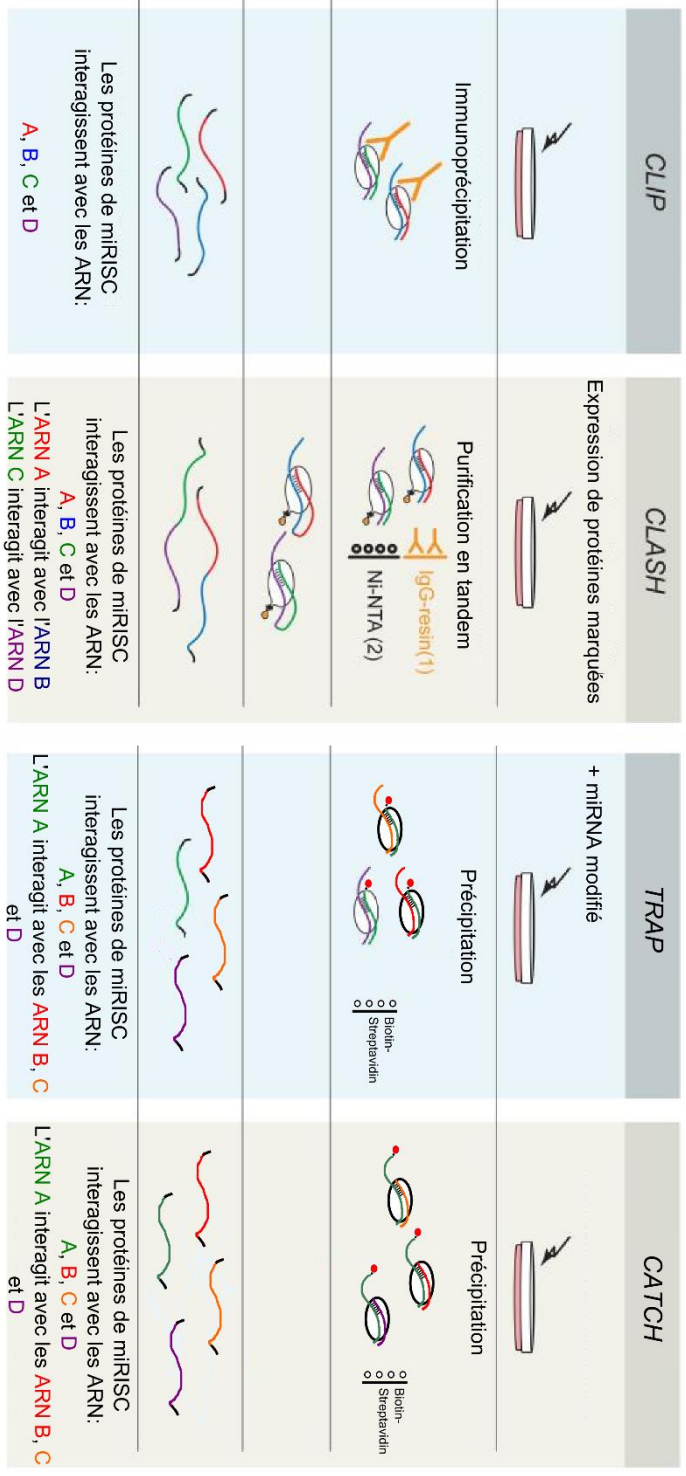


Figure 10. Méthodes biochimiques de découverte de cibles des miARN. A gauche, la méthode CLIP (*crosslinking immunoprecipitation*) repose sur la création de liens réversibles par exposition aux ultra-violets, l'immunoprécipitation des complexes miRNP et le séquençage des ARN. Seule une analyse bioinformatique permet l'identification des cibles des miARN. La méthode CLASH à sa droite introduit une purification en tandem et la création d'une liaison des miARN et ARNm en contact. Le séquençage permet donc de détecter les cibles directes des miARN. La méthode TRAP repose sur une précipitation plus simple basée sur le complexe Biotine-Streptavidine et la PCR des fragments d'ARN. Dans ce cas, ce sont toutes les cibles d'un miARN en particulier qui sont étudiées. Enfin la méthode CATCH est très identique à TRAP dans son ensemble mais se place du point de vue d'un ARNm et étudie l'ensemble des miARN qui le régule. Adaptée d'après (Helwak and Tollervey, 2014).

Ces trois techniques restent cependant lourdes à mettre en place autant

financièrement qu'humainement puisqu'elles nécessitent plusieurs étapes différentes et très spécifiques (immunoprécipitation par anticorps, purification, séquençage, bioinformatique, etc.). MiR-TRAP (*miRNA target RNA affinity purification*) est une méthode cherchant à pallier ces problèmes en utilisant des miARN modifiés par l'ajout de psoralène dans la *seed* et d'une biotine en 3' (Baigude et al., 2012). Le premier composé permet la création du lien entre miRISC, miARN et ARNm par irradiation UV comme pour la méthode HITS-CLIP. Cette irradiation s'effectue toutefois à une longueur d'onde de 360 nm – moins délétère pour les cellules – plutôt qu'à 254. Le deuxième composé, quant à lui, est une molécule permettant l'isolation et la purification du complexe par l'utilisation de billes de streptavidine. Le séquençage est alors effectué simplement par RT-qPCR (*reverse transcriptase quantitative polymerase chain reaction*) après réversion du *cross-linking*.

Un élément redondant pour cet ensemble de méthodes est la focalisation sur les cibles de quelques miARN : elles permettent pour un (ou quelques) miARN(s) donné(s) de déterminer un ensemble de cibles. *A contrario*, la méthode miR-CATCH est une technologie à contre-courant qui ne se focalise pas sur les miARN mais sur les cibles. Son but est donc d'identifier tous les miARN qui vont cibler un gène en particulier. Tout comme MiR-TRAP, cette technique repose sur le *cross-linking* réversible de miRISC, d'un miARN et d'un ARNm et l'isolation/purification du complexe par *pull-down* biotine-streptavidine (Vencken et al., 2015).

(3) Approches –omiques

Les approches -omiques se basent sur les dernières avancées en analyse de transcriptome (séquençage de dernière génération p.ex.) et protéome (spectrométrie de masse p.ex.) pour déterminer l'effet des miARN sur l'expression des ARNm ou des protéines. Ceci se fait essentiellement par comparaison d'une condition avec un miARN induit ou réprimé avec une condition contrôle avec un taux d'expression basal de ce miARN. Ces méthodes ont la particularité d'être quantitatives. En effet, une majorité des techniques présentées dans cette partie donnent non seulement une information sur le nombre de cibles potentielles des

miARN mais également sur l'effet (relatif ou absolu) qu'ont les miARN sur l'expression de leurs ARNm et/ou protéines cibles.

(a) Au niveau des transcrits

Les trois méthodes de choix en transcriptomique sont les puces à ADN (*microarrays*), le séquençage et la PCR (*polymerase chain reaction*). Les deux premières sont des méthodologies à haut-débit alors que la dernière est généralement une méthode à plus faible débit. Les puces à ADN analysent une expression relative du niveau d'expression des ARNm cibles (Lim et al., 2005; Wang and Wang, 2006) alors que le séquençage donne une information absolue sur ces cibles (Xu et al., 2010). Il existe bien évidemment différentes techniques dans les deux cas (p.ex. plus d'une dizaine pour les puces à ADN), les plus utilisées conjointement à l'étude des miARN sont les technologies des sociétés Agilent et Affymetrix pour les puces à ADN et Illumina pour le séquençage. Dans le cas de la PCR, l'objectif est d'amplifier un signal d'ADN à partir de *primers*, il existe différentes versions de la technologie notamment la qPCR qui possède l'avantage d'être quantitative.

Brièvement, dans le cas des puces à ADN, des séquences nucléotidiques complémentaires aux ARNm et fixées sur une lame (les sondes) permettent l'hybridation des ARNm préalablement marqués et extraits de cellules après lysat. Suite à l'hybridation, la puce est lue par un scanner afin de détecter le niveau d'expression de ces ARNm (Figure 11 A). Ce calcul se base généralement sur une activité luminescente (chimique ou biologique), soit par la comparaison d'intensités sur deux canaux (i.e. Cy3 et Cy5, on parle de puces à ADN à deux couleurs, Figure 11 A), soit par l'intensité observée sur un seul canal (puces à ADN à une couleur).

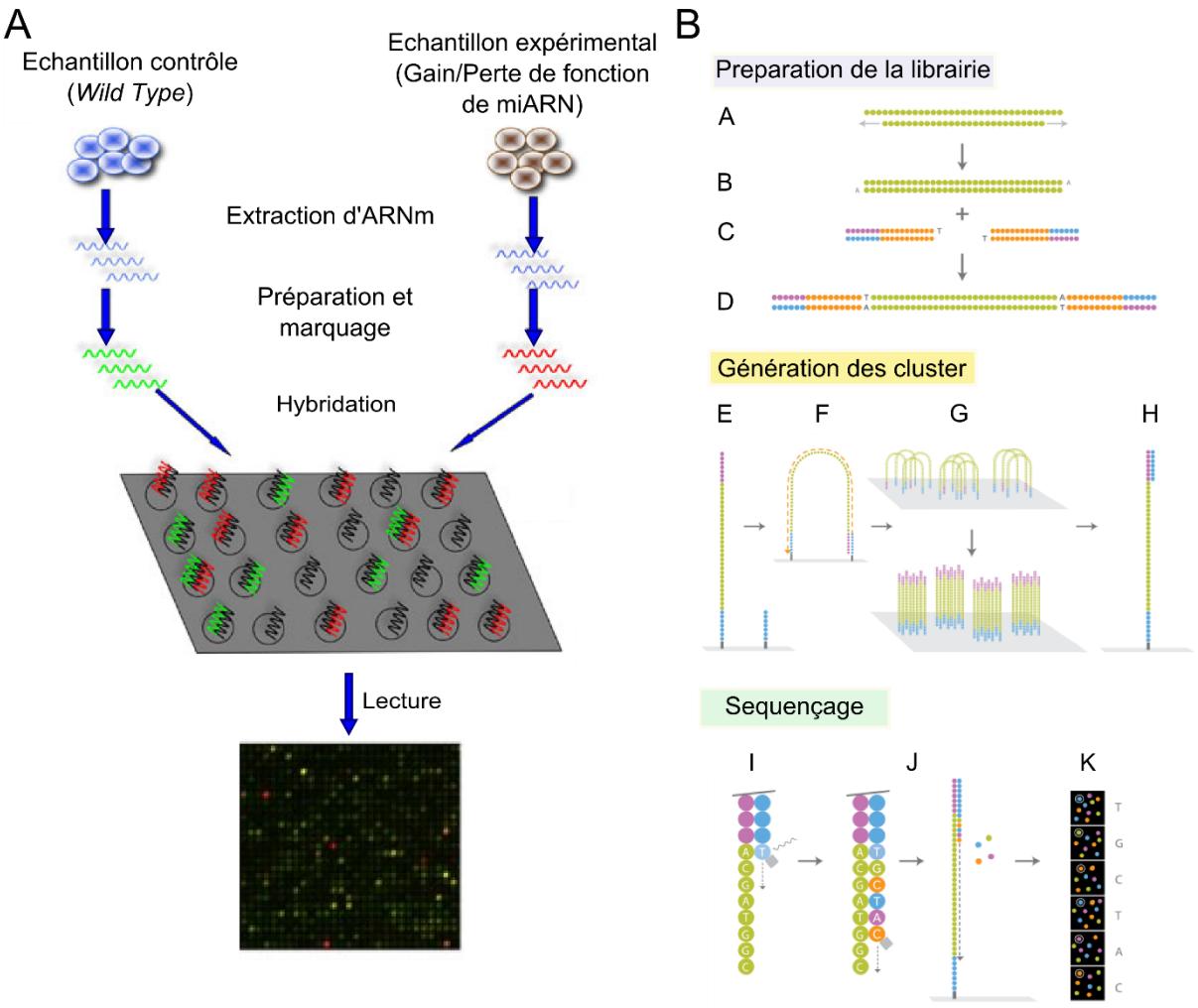


Figure 11. Deux méthodes transcriptomiques d'analyse d'ARNm. A | Expérience de puce à ADN à deux couleurs. L'ensemble tient en quatre étapes : 1) l'extraction des ARNm, 2) leur préparation et marquage, 3) l'hybridation des ARNm à la puce et 4) la lecture de la puce. B | Expérience de séquençage par la technologie Solexa (illumina). La méthode se déroule en trois étapes : 1) la préparation des ARNm, 2) la génération des clusters par pont et 3) le séquençage par luminescence. Adaptée d'après (Yevgeniy, 2011) sur http://bitesizebio.com/7206/introduction-to-dna-microarrays/ et (Michael, 2011) sur http://www.rsc.org/.

Pour le séquençage, après préparation et extraction, les séquences d'ARNm sont attachées à une surface et amplifiées. La répétition de ce processus forme des clusters doubles-brins sur toute la surface qui seront par la suite dénaturés. Le séquençage se fait alors par cycle, après l'ajout itératif de nucléotides terminaux réversibles marqués (c'est à dire qu'ils empêchent la synthèse d'ADN une fois incorporés mais que ce processus est réversible) et de polymérase. Le premier cycle nécessite également l'ajout de *primer* pour démarrer la rétro-synthèse. La surface est excitée par un laser et chaque cluster s'allumera d'une couleur en fonction du nucléotide apparié à la zone à séquencer. Après identification de ce nucléotide, le blocage de la transcription est levé et le cycle reprend jusqu'à ce que l'ensemble des nucléotides ait été identifié (Figure 11 B). Un traitement informatique est par la suite nécessaire

afin « d'aligner les *reads* » sur le génome pour l'identification des ARN séquencés. Dans les deux cas cependant, il existe de nombreux cas de faux positifs et faux négatifs, principalement parce qu'il est très difficile lors de l'utilisation de ces deux méthodes de séparer les cibles directes des cibles indirectes.

Enfin, la PCR est une technique de biologie moléculaire d'amplification génique d'un fragment d'ADN à partir d'amorces prédéfinies. Il en existe différentes variantes dont la RT-qPCR (*Quantitative reverse transcription PCR*), une méthode qui utilise la rétro-transcription (ARN vers ADN) dans un premier temps pour former un ADN circulaire qui sera par la suite amplifié par PCR. L'avantage de cette variante est l'aspect quantitatif : elle permet en effet d'obtenir une information sur l'expression du fragment initialement analysé.

(b) Au niveau des protéines

En analyse protéomique, la technique des puces à protéine est sensiblement identique aux puces à ADN dans son ensemble, à la différence près que ce sont des protéines complexes (anticorps ou protéines « sondes ») qui sont fixés sur les plaques pour lier les protéines cibles afin de déterminer leur expression. Alternativement, la spectrométrie de masse quantitative permet également de mesurer l'expression des protéines. Cette technologie a pour but de caractériser les protéines en estimant des ratios poids moléculaire sur charge (m/z). Pour la découverte de cibles assistée par spectrométrie de masse, l'approche la plus utilisée à l'heure actuelle est probablement la méthode SILAC (*Stable-isotope Labeling by Amino Acids in Cell Culture*), une technique très commune en protéomique quantitative (Mann, 2006). Dans cette approche, des cellules sont cultivées soit sur le milieu lourd (dans lequel des isotopes lourds d'acides aminés ont été ajoutés), soit sur un milieu léger (un milieu classique, sans isotopes lourds). De ce fait, les protéines synthétisées dans les cellules sur milieu lourd auront incorporées les isotopes lourds contrairement aux protéines du deuxième lot. Après quelques temps, les cellules sont mises en commun, digérées et analysées par spectrométrie de masse (Figure 12) (Gruhler and Kratchmarova, 2008). C'est alors le ratio d'intensité des pics peptidiques dans les deux conditions (milieu lourd et léger)

qui définit l'abondance protéique. Afin d'adapter la méthode à la recherche de cibles de miARN, un agoniste de miARN est également utilisé en plus du marquage isotopique (Figure 12), comme pour l'étude de Vinther *et al.* sur miR-1 et pour lequel ils ont découvert une douzaine de cibles dans des cellules HeLa (Vinther et al., 2006). D'autres études avec un débit supérieur ont bien évidemment suivi cette première preuve de concept (Baek et al., 2008; Yang et al., 2010b; Korpal et al., 2011; Yan et al., 2011).

Introduite en 2008 par Selbach et collaborateurs, pSILAC (*pulsed-SILAC*) est une amélioration de la méthode SILAC. Pour cette dernière, les cellules sont premièrement cultivées sur un milieu léger puis transférées vers deux milieux contenant chacun un isotope lourd différent et enfin remises en commun. Les protéines auront donc incorporé uniquement un seul des deux isotopes, définissant ainsi le « pulse » de la méthode. Le ratio entre les deux types de protéines donne alors une estimation de la différence de niveau de traduction (Figure 12). Sachant que seules les protéines nouvellement synthétisées sont considérées ici (c'est-à-dire synthétisées après le pulse), c'est une méthode de choix pour l'analyse de cibles directes des miARN (Selbach et al., 2008).

Il existe également d'autres techniques utilisant la protéomique pour identifier les cibles des miARN. Par exemple, c'est le cas de la méthode iTRAQ utilisée pour découvrir les cibles du miR-21 sur des cellules cancéreuses (Yang et al., 2009) ou encore de la méthode 2D-DIGE (*2D-difference gel electrophoresis*) (Unlü et al., 1997) utilisée pour la découverte des cibles de miR-29a (Figure 12) (Muniyappa et al., 2009). Ces méthodes restent cependant des méthodes bas-débits et sont tout autant lourdes (voir plus) à mettre en place que les deux autres techniques citées ci-dessus (Elliott et al., 2009).

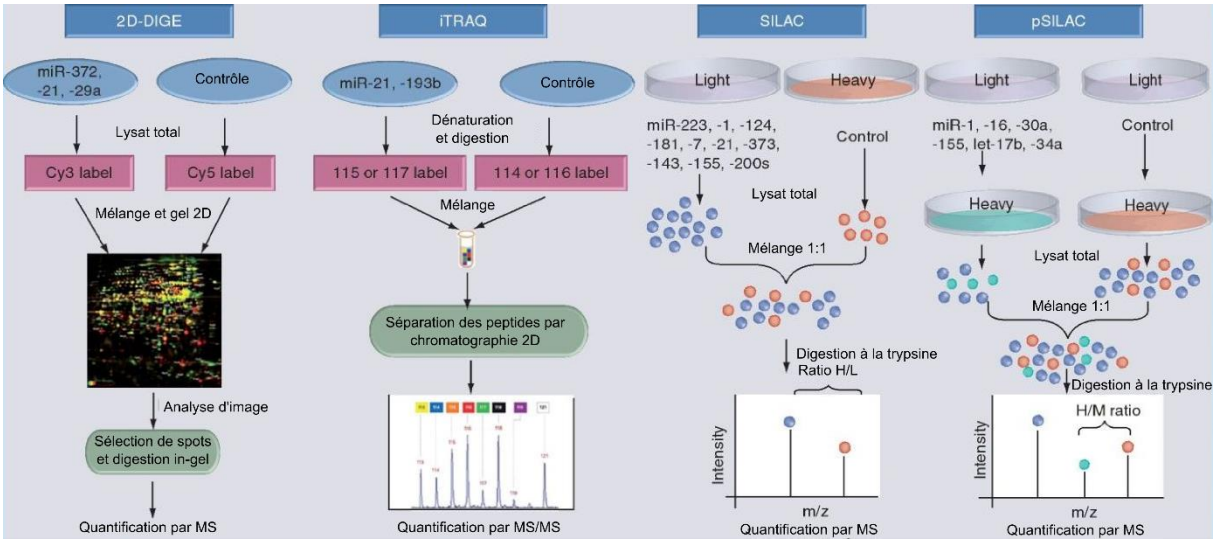


Figure 12. Différentes méthodes d'études des cibles de miARN assistées par spectrométrie de masse. Dans la méthode 2D-DIGE, les deux conditions sont marquées différemment, réunies puis séparées sur gel 2D. Une sélection des spots d'intérêts, la digestion de ces spots puis l'analyse MS permet l'identification des cibles. La méthode iTRAQ suit le même principe que 2D-DIGE mais la séparation se fait par chromatographie. Les méthodes SILAC et pSILAC reposent sur l'utilisation de peptides marqués et une analyse différentielle des spectres MS entre les deux conditions. L'avantage du pSILAC tient dans le pulse qui permet de ne considérer que les nouvelles protéines synthétisées et donc uniquement les cibles directes. Heavy : milieu lourd. Light : milieu léger. MS : spectrométrie de masse. MS/MS : spectrométrie de masse en tandem. D'après (Li et al., 2012)

A quelques exceptions près, un consensus ressortant de l'ensemble de ces méthodes –omiques est l'effet généralement modeste qu'ont les miARN sur l'expression des gènes et des protéines. De fait, l'expression des protéines ou des gènes est souvent assez peu affectée par les miARN avec des changements généralement inférieurs à un facteur 4 (log_2 *Fold-Change* < 2), même s'il existe des cas où ces changements sont bien plus conséquents. Ce constat semble encore plus vrai *in vivo* lorsque les miARN sont utilisés individuellement (Vidigal and Ventura, 2014).

En dernier lieu, les approches moléculaires restent des méthodes bas-débits très utilisées dans la découverte de cibles des miARN. Ces dernière concernent toutes les méthodes couramment utilisées en biologie moléculaire comme le western blot, le northern blot, les méthodes d'hybridation, etc.

Chacune des méthodes exposées ici possède évidemment ses propres avantages et inconvénients (Mestdagh et al., 2011). De plus en plus d'études cherchent donc à combiner plusieurs méthodes afin de remédier aux inconvénients de chacune tout en conservant leurs avantages respectifs. Un exemple particulièrement intéressant est celui de Kaller et

collaborateurs qui utilisent à la fois les algorithmes de prédictions TargetScan et Pictar, la méthode pSILAC, des *microarrays* et des expériences de gènes rapporteurs afin d'étudier les cibles directes et indirectes de miR-34a dans une lignée cellulaire de cancer colorectal (Kaller et al., 2011). Ce genre d'études très intégrées est bien évidemment encore assez rare pour le moment mais elles vont probablement devenir plus fréquentes avec l'amélioration des technologies et la baisse des coûts.

6. Le rôle des miARN dans la régulation de l'expression

Nous avons pu le constater tout au long de ce chapitre, la régulation par le miARN est assez complexe. Non seulement ce sont de fins régulateurs de l'expression des gènes mais ils sont eux-mêmes régulés par des protéines (*p.ex.* les facteurs de transcriptions). Par ce dernier constat, il n'est donc pas étonnant de les voir impliqués dans des boucles de régulation où les deux entités interagissent les unes avec les autres, de manière négative ou positive (Tsang et al., 2007; Bracken et al., 2008; Yu et al., 2008). Ces différents éléments placent en conséquence les miARN au centre de la cohérence des systèmes biologiques (Ebert and Sharp, 2012). Leur habilité à moduler de manière extrêmement fine l'expression des gènes apporte une certaine robustesse à ces systèmes ; c'est-à-dire qu'ils leur permettent de continuer à fonctionner malgré des perturbations internes ou externes (Kitano, 2004). De plus, cette robustesse émerge également du concept de dégénérescence. La dégénérescence peut être définie comme la capacité d'éléments structurellement différents à avoir des fonctions identiques ou à produire des résultats identiques (Edelman and Gally, 2001), une définition qui colle parfaitement aux miARN. Dans le cas d'une perturbation d'un miARN en particulier par exemple, d'autres miARN visant les mêmes gènes seraient tout à fait capables de prendre la place du miARN perturbé pour réajuster le niveau d'expression du gène en question.

Pour être plus précis, il est aujourd'hui considéré que la fonction des miARN dans la régulation de l'expression génique est double : 1) les miARN permettent premièrement de réajuster l'expression des gènes autour d'une moyenne (fonction de « *tuning* ») et 2) ils

permettent deuxièmement de réduire la variance d'expression des gènes (fonction de « *buffering* » ou rhéostats) (Wu et al., 2009; Vidigal and Ventura, 2014). Ces deux fonctions complémentaires peuvent être considérées autour des boucles de régulation. Dans le premier cas, le miARN agit de façon cohérente (seul ou avec une autre entité comme un facteur de transcription) sur l'expression du gène : il(s) réprime(nt) globalement son expression et ajuste(nt) donc son expression moyenne. Dans le second en revanche, le miARN agit de façon incohérente avec son partenaire : l'un réprime le gène alors que l'autre l'induit. Ce sont ces effets contraires qui permettent de réduire l'écart à la moyenne de l'expression du gène.

La régulation par les miARN ne s'arrête pourtant pas à ce niveau puisque plusieurs études récentes viennent complexifier ce modèle en décrivant de nouvelles fonctions potentielles par les miARN. Nous pouvons citer par exemple la régulation positive de l'expression des gènes, les miARN « éponges » ou encore plus récemment le rôle des miARN dans la régulation des pri-miARN.

7. De nouveaux rôles pour les miARN

a) Des miARN activateurs?

Les miARN sont connus essentiellement pour jouer un rôle négatif sur l'expression des gènes et nous avons pu comprendre comment se déroulent les différentes étapes de cette régulation. Pourtant, il semblerait que certains miARN soient potentiellement capables de réguler positivement l'expression de gènes ou autrement dit, d'en augmenter l'expression. Le terme faisant référence à ce phénomène est « ARN activation » (ARNa).

Le premier exemple de ce type de phénomène pour les miARN a été publié fin 2007 (Place et al., 2008), bien qu'il ait été découvert plus tôt pour un autre type d'ARN interférent (Li et al., 2006). Dans cette étude sur cellules PC-3, Place et collaborateurs ont montré que l'augmentation artificielle du niveau de miR-373 (ainsi que du pre-miR-373) augmentait l'expression génique d'E-cadhérine et de CSDC2 (*cold-shock domain containing protein C2*) – deux cibles potentielles du miARN – mais que cet effet était aboli lors de l'introduction de

mutations dans la séquence du miARN. De façon intéressante, les sites de liaison du miARN sur les portions d'ADN des deux protéines se trouvent dans leur région promotrice respective. Ce fait a conduit les auteurs à considérer un lien éventuel entre ARN activateurs (ARNa) et régions promotrices, même si d'autres cibles potentielles avec site de liaison dans le promoteur ne montraient aucun effet similaire.

Cette même année, Vasudevan et collaborateur ont montré que miR-369-3 était également capable d'activer la traduction de TNFα (*tumor necrosis factor*-α) dans des cellules HEK293 (cellules embryonnaires de reins). Dans ce cas en revanche, ce n'est pas au travers de la région promotrice que cette activation se met en place mais grâce à des éléments riches en AU (AREs) situés en 3'UTR de l'ARNm. Les auteurs de cette étude ont également montré que let-7 était tout autant capable d'activer la traduction, mais uniquement lorsque le cycle cellulaire était à l'arrêt : let-7 reprenait en effet son rôle d'inhibiteur dans les cellules prolifératives (Vasudevan et al., 2007). Une partie de ces résultats a de surcroit été confirmée par Mortensen et al. sur des ovocytes de *Xenopus laevis* (Mortensen et al., 2011)

b) Des miRNA appâts ?

L'activité de miARN appâts constitue un autre mécanisme particulièrement intéressant. Pour ce dernier, un miARN est susceptible d'agir comme appât afin de bloquer des interactions protéines/ARNm. Dans ce cas, le miARN ne se lie pas au messager mais directement à la protéine (George and Tenenbaum, 2006). C'est précisément ce qu'ont prouvé Eiring *et al.* (Eiring et al., 2010) pour le miR-328 et la protéine inhibitrice hnRNP E2. Ces auteurs ont en effet montré (*in vivo* et *in vitro*) que le miARN pouvait restaurer la traduction de la protéine CEBPA (*CCAAT/enhancer-binding protein alpha*) en entrant en compétition avec le messager de CEPBA pour la liaison à hnRNP E2 : une fois le miARN connecté à hnRNP E2, CEBPA peut être traduite.

Il existe également des cas où ce sont les ARNm qui jouent le rôle d'appât pour les miARN : on parle alors d'ARN compétitifs endogènes (ARNce). Ces ARNce sont en fait des

« éponges » à miARN (Ebert et al., 2007) qui vont altérer leur(s) fonction(s) en s'y fixant, sans pour autant affecter leur biogénèse (Haga and Phinney, 2012). Et enfin, il existe un dernier cas où des protéines se lient à la 3'UTR supposée être ciblée par des miARN afin de prévenir leur fonctionnement. C'est notamment le cas de la protéine Dnd1 (*Dead end 1*) (Kedde et al., 2007)

c) Des réserves de miARN ?

La régulation de la biogénèse des miARN est sous la dépendance des facteurs de transcription. La stabilité cellulaire des miARN, quant à elle, varie énormément, mais elle est généralement augmentée lorsque les miARN sont associés à certaines protéines (Drosha, RISC, etc.). Une récente étude montre qu'une de ces interactions en particulier permettrait non seulement de stabiliser les miARN mais servirait en fait de réserve de miARN (La Rocca et al., 2015). Les auteurs de cette étude ont montré que, dans les cellules adultes, la plupart des miARN étaient liés à des complexes LMW-RISC, non engagés dans la répression d'ARNm et formant donc une réserve « prête à l'action », alors que dans les cellules prolifératives (et les lignées cellulaires) les miARN étaient liés aux ARNm et à des HMW-RISC. Selon les auteurs, cette différence permettrait à la cellule de faire face à des transitions physiologiques ou à des cas de stress en modulant rapidement les niveaux d'expression des gènes, mais elle pourrait également expliquer pourquoi les cellules animales ont « développé » la voie des miARN tout en ayant conservé la voie des siARNs.

d) Autorégulation de la biogénèse par les miARN?

Comme nous l'avons vu, let-7 fut un des premiers miARN à avoir été découvert. Chez *C. elegans*, ce miARN (entre autres) a permis de caractériser en grande partie le mécanisme de biogénèse des miARN chez les eucaryotes. Pourtant, une récente étude chez ce ver prouve qu'il reste probablement encore à découvrir sur cet organisme et ses miARN, puisque les auteurs de l'étude ont montré que let-7 était capable de réguler sa propre biogénèse en se liant directement au miARN primaire pri-let-7 (Zisoulis et al., 2012).

Dans l'organisme modèle, pri-let-7 possède un site de liaison à la protéine ALG-1 (*Argonaute Like protein-1*), une protéine intervenant dans la maturation du miARN primaire. Le bon déroulement de cette maturation nucléaire n'est toutefois permis qu'en présence de la séquence mature de let-7, qui se lie à un site de liaison conservé en 3' du pri-miARN. Le miARN let-7 se lie donc à son propre miARN primaire pour entrainer le recrutement de la protéine ALG-1 afin de promouvoir sa propre synthèse. Ces résultats montrent également que certaines protéines AGO ainsi que certains miARN matures peuvent se retrouver dans le noyau.

Les interactions entre miARN mature et primaire ne semblent pas non plus se limiter à la régulation positive. En effet, dans une étude publiée la même année, Tang et collaborateurs ont montré que miR-709 était capable de réguler la maturation du pri-mir-15a/16-1 en se liant directement à une séquence complémentaire de 19 nucléotides sur le miARN primaire dans le noyau. Dans ce cas-là, l'interaction empêche la biogénèse des pre-miARN et par conséquent, celle des miARN matures (Tang et al., 2012).

e) Synthèse protéique à partir de pri-miARN

Très récemment, une équipe de chercheurs à Toulouse a pu mettre en évidence un autre rôle pour les miARN : leur traduction en peptide fonctionnel (Lauressergues et al., 2015).

Dans leur étude, les auteurs ont analysé des ORF potentielles dans la séquence des pri-miARN miR-171b et miR-165a chez *Arabidopsis thaliana* et *Medicago truncatula* et ont découvert des petites séquences portant des codons start et stop et pouvant être traduites. Ils ont démontré que ces séquences étaient effectivement traduites en petits peptides fonctionnels – qu'ils ont nommé miPEP – capables d'améliorer l'accumulation de leur miARN mature respectif. Ces miPEP semblent en fait jouer sur l'activation de la transcription des miARN, sans influencer la stabilité de ces derniers.

Pour le moment, cette découverte est limitée au monde végétal mais il serait probablement intéressant de voir si ce phénomène existe aussi chez les animaux. Sa

découverte étant également très récente, peu d'informations sont disponibles sur son fonctionnement exact.

En conclusion, bien que de nombreuses avancées aient été faites pour comprendre le rôle des miARN, il semblerait bien que nous soyons encore loin d'avoir tout découvert. Une question qui reste encore assez peu étudiée au final est le rôle systémique des miARN dans nos cellules et notamment leur manière d'interagir ou de co-agir pour contrôler le destin d'une cellule. Afin de pouvoir répondre à ce genre de question, des approches systémiques et intégrées sont donc nécessaires. Un exemple de telles approches est l'analyse de réseau de régulation de l'expression génique. Ce type d'analyse emprunte les fondamentaux de la théorie des graphes et de l'analyse des réseaux sociaux pour les appliquer aux réseaux biologiques afin de formuler de nouvelles hypothèses ou d'apporter des compléments d'information sur le système en question. La suite de cette introduction décrit ces approches et leurs utilisations, notamment dans le monde des miARN.

B. Les réseaux

L'objectif principal de l'analyse des réseaux, qu'ils soient sociaux, biologiques ou de tout autre type, est la compréhension et/ou la prédiction du comportement de ces systèmes, on parle alors d'études systémiques. L'analyse de réseaux au sens large n'est pas récente : la plupart des auteurs s'accordent à dire qu'elle remonte au problème des ponts de Königsberg et sa résolution par Euler en 1735 - 1736. À l'époque se pose la question de savoir si les sept ponts de la ville de Königsberg peuvent être traversés en un seul voyage et sans repasser deux fois par le même (Figure 13), ce à quoi Euler a répondu par la négative. Depuis, de

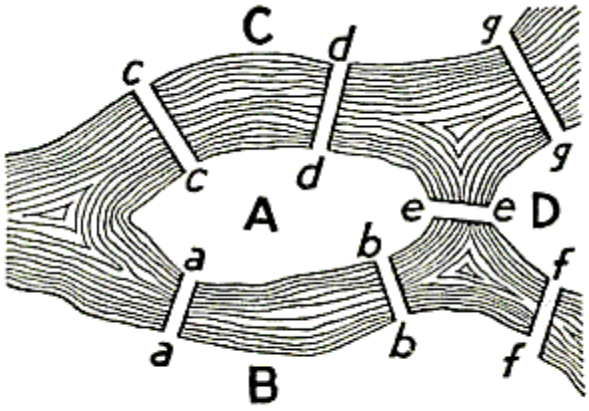


FIGURE 98. *Geographic Map: The Königsberg Bridges.*

Figure 13. Problème des ponts du Königsberg (tiré de mathworld.walfram.com (Weisstein)**)**

nombreuses théories mathématiques ont vu le jour et l'analyse des réseaux tend aujourd'hui vers l'étude des réseaux à grande échelle, où l'on ne se concentre plus sur les propriétés individuelles de quelques éléments mais plutôt sur les propriétés statistiques globales (Onnela et al., 2007; Chen et al., 2011). Ces nouvelles approches ne sont bien évidemment permises que grâce à l'amélioration des technologies informatiques, des méthodes statistiques mais également grâce à un accès plus simple aux données à grande échelle.

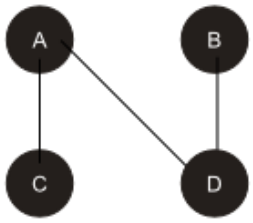
1. Qu'est-ce qu'un réseau ?

Au sens mathématique du terme, un réseau – également appelé « graphe » – est constitué d'un ensemble de nœuds V (pour *Vertices*) reliés par des liens E (pour *Edges*) et dénoté par la formule : $G = (V, E)$.

Un graphe peut être représenté par différents moyens, le plus simple étant la représentation graphique, aisément interprétable à l'œil pour de petites dimensions. Ils peuvent également être symbolisés sous d'autres formats comme des fichiers textes (des fichiers de type csv – *comma separated value* par exemple) ou encore des matrices d'adjacence. Dans le cas des fichiers textes, chaque ligne représente deux nœuds ainsi que le lien qui existe entre ces nœuds ; chacune de ces informations étant séparées par un symbole. Pour les matrices d'adjacence, les colonnes et les lignes représentent les différents

nœuds alors que les intersections des lignes et des colonnes indiquent les liens. Une matrice binaire donnera donc une information sur la présence ou l'absence de liens entre les nœuds (Figure 14).

A



B

A, C, 1
A, D, 1
D, B, 1

C

	A	B	C	D
A	0			
B	0	0		
C	1	0	0	
D	1	1	0	0

Figure 14. Différentes représentations de graphe. A | Un exemple de réseau à 4 nœuds et 3 liens. B | Le même réseau sous format texte. Chaque ligne représente un lien. C | Représentation du graphe A sous forme de matrice d'adjacence.

Dans le cas d'un réseau simple, les nœuds sont simplement reliés entre eux par les liens, sans boucles (une boucle étant un nœud relié à lui-même) ni liens multiples (deux nœuds reliés ensemble par plusieurs liens) (Figure 15 A). Les réseaux peuvent cependant être bien plus complexes que cette définition sommaire. Par exemple, il peut y avoir différents types de nœuds au sein d'un même réseau (protéines et miARN) ou différents types de liens (répression entre un miARN et un gène ou interaction protéine-protéine) (Figure 15 B). Lorsque les nœuds forment deux ensembles distincts l'un de l'autre, on parle de graphe bipartite. Les nœuds ou les liens peuvent posséder certaines propriétés comme des noms ou des valeurs : un nœud par exemple peut être une protéine de type kinase exprimée à un certain niveau dans un certain type cellulaire alors qu'un lien peut donner une constante de dissociation entre deux protéines qui interagissent. Si des valeurs numériques sont associées aux nœuds et/ou aux liens, on parle de graphes pondérés (Figure 15 C). Dans le cas de graphes orientés, les liens possèdent un sens (une protéine qui inhibe une autre protéine mais pas l'inverse) (Figure 15 D). Les graphes peuvent aussi être cycliques (possibilité de partir d'un point A et revenir à ce point en suivant un chemin) ou acycliques. Les graphes acycliques sont d'ailleurs très souvent représentés par des arbres (Figure 15 E). Et enfin, les liens peuvent relier plus que deux nœuds, on parle alors d'hyperliens et d'hypergraphes.

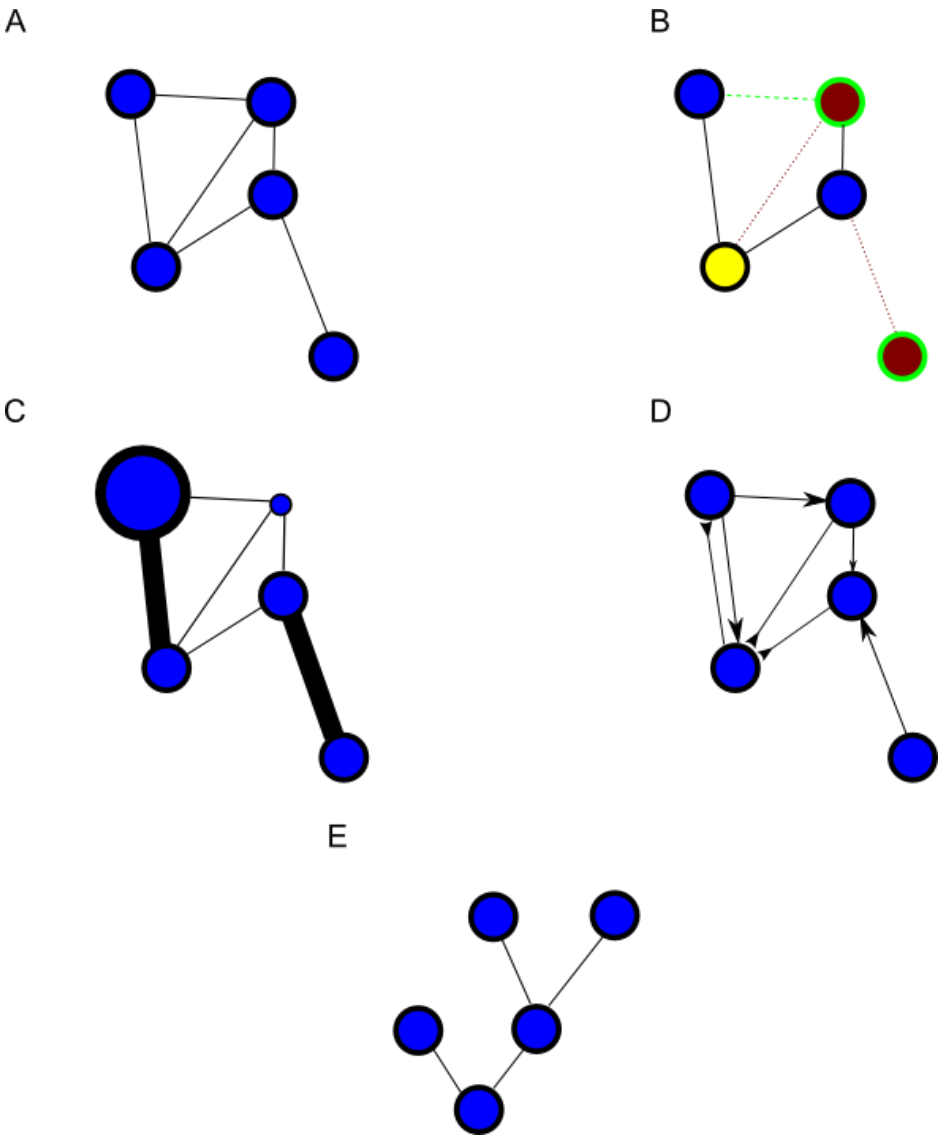


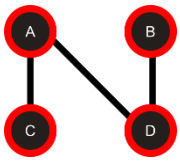
Figure 15. Différents types de graphes. A | Un graphe simple non orienté. B | Un réseau avec différents types de nœuds et de liens. C | Un graph pondéré où chaque nœud et chaque lien possède une certaine propriété associée les faisant paraitre plus ou moins épais. D | Un graphe orienté avec deux types d'orientation différente. E | Un réseau acyclique représenté sous forme d'arbre (dichotomique ici). Le nœud le plus bas porte également le nom de « racine ».

2. Métriques et réseaux : l'analyse des réseaux

L'analyse des réseaux passe par l'observation des propriétés de ces derniers. Ces propriétés ont été définies par différentes métriques comme le degré moyen, le chemin moyen, le diamètre ou encore le coefficient de *clustering* ou les coefficients de centralité, qui permettent de comprendre non seulement comment transite le flux d'information au sein du réseau mais aussi l'importance de la structure du réseau (Freeman, 1978; Albert and Barabási, 2002; Newman, 2003). Cette partie se charge d'exposer certaines de ces propriétés afin que

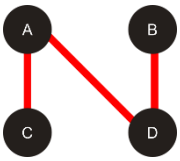
le lecteur puisse mieux appréhender la suite du document. Notons « n » le nombre de nœuds et « m », le nombre de liens, « G » un graphe simple non orienté et son ensemble de nœuds V et de liens E. Nous prendrons les graphes des figures 14 et 15 A comme exemple pour la suite.

Nœud : unité de base de tous types de réseau, c'est un terme qui dérive des sciences computationnelles. Cet élément porte également le nom de vertex (vertices au pluriel), site (physique), acteur (sociologie) ou sommet.



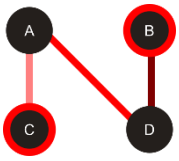
Ex : La Figure 14 est composé de n = 4 nœuds : A, B, C et D.

Lien : la ligne qui connecte deux nœuds. Ce terme porte également le nom d'arête, d'arc ou *edge* en anglais. Les arcs font généralement références à des liens orientés, c'est-à-dire qu'ils possèdent un sens (unique) lorsqu'ils relient des nœuds entre eux. On dit de deux nœuds qu'ils sont adjacents ou voisins, s'ils sont reliés par un lien.



Ex : Sur la Figure 14, les nœuds A et C sont reliés par un lien tout comme les nœuds A, D et D, B mais pas les nœuds A, B ni C, B ; C, D et A, B. En tout, le graphe comporte m = 3 liens.

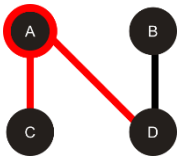
Chemin : ensemble de liens permettant d'aller d'un nœud à un autre, en traversant d'autres nœuds du réseau. La longueur du chemin représente le nombre de liens empruntés pour aller du nœud d'origine au nœud terminal. On parle également de liaison à « n » degré.



Ex : Le chemin reliant C et B sur la Figure 14 est le chemin passant par les nœuds C, A, D et B et empruntant les liens C → A, A → D et D → B. Il n'en existe qu'un seul dans ce cas. Ce chemin est dénoté de la manière suivante : ((C,A),(A,D),(D,B)) et a une longueur de 3. On dit également que les nœuds C et B sont des voisins au troisième degré.

Degré : le nombre de liens k connectés à un nœud. Dans un graphe simple, le degré est égal au nombre de nœuds adjacents à celui considéré. Le degré d'un nœud peut également être appelé connectivité. Dans le cas de graphe orienté, il existe deux types de degré : degré entrant et degré sortant, faisant respectivement référence aux liens arrivant vers le nœud et aux liens sortant du nœud. Les nœuds à plus fort degré dans un réseau sont appelés *hub*.

Ex : Sur la Figure 14, le degré des nœuds A est de 2. De la même manière le degré de D est 2 alors que celui des nœuds C et B est de 1.



Densité : rapport entre le nombre de liens observés sur le nombre de liens possibles. La densité est calculée par la formule suivante :

$$D = \frac{2m}{n.(n-1)}$$

où lorsque la densité est égale à 0, tous les nœuds sont isolés. Si tous les nœuds sont connectés, on parle de graphe complet et la densité vaut 1. La formule est légèrement adaptée dans le cas des graphes orientés.

Ex : Sur la Figure 14, la densité du réseau est de 0.5 puisque seule la moitié des liens entre les quatre nœuds sont présents.

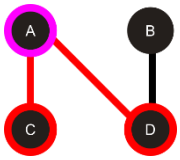
Degré moyen : mesure proche de la densité qui compare le nombre total de nœuds au nombre total de liens du réseau. Il est calculé par la formule suivante :

$$< k > = 2.m/n$$

Ex : Sur la Figure 14, le degré moyen est de 1.5.

Géodésique / Distance : plus court chemin « d » pour aller d'un nœud à un autre. Il peut y avoir plusieurs géodésiques entre deux nœuds. La distance entre un nœud et lui-même est 0. Celle entre deux nœuds isolés est ∞.

Ex : Le géodésique entre les nœuds C et D sur la Figure 14 est le chemin ((C,A),(A,D)) avec une longueur de 2. Dans le cas où un autre nœud était connecté à C et D comme A, il y aurait alors deux géodésiques.



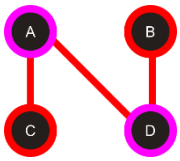
Chemin moyen : moyenne des plus courts chemins entre toutes les paires de nœuds dans le graphe. Le chemin moyen porte également le nom de chemin caractéristique et est calculé par la formule suivante, si d(i,j) représente le plus court chemin entre les nœuds i et j:

$$l_g = \frac{\sum_{i \neq j} d(i,j)}{n(n-1)}$$

Ex : Sur la Figure 14, le chemin moyen est de 1,67

Diamètre : la longueur du plus long géodésique entre n'importe quel couple de nœuds du réseau. C'est en fait le plus long des plus courts chemins du réseau.

Ex : Le diamètre du réseau sur la Figure 14 est de 3. Il est représenté par le chemin ((C,A),(A,D),(D,B)) indépendamment de la direction.

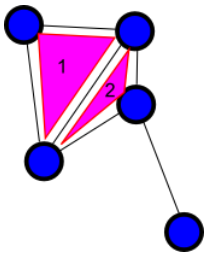


Distribution de degré : l'histogramme du degré des nœuds d'un graphe où p_k représente le nombre de nœuds avec un degré k (cf. Figure 17).

Coefficient de *clustering* : calcul du nombre de triangles (trois nœuds interconnectés) dans le graphe. Le coefficient de *clustering* – également appelé transitivité – fait en quelque sorte référence à l'adage « les amis de mes amis sont mes amis ». Il donne la probabilité moyenne que deux nœuds voisins d'un même nœud soient eux-mêmes connectés. Il est calculé par la formule suivante :

$$C = \frac{3 \times nombre\ de\ triangles\ dans\ le\ réseau}{nombre\ de\ triplets\ connectés}$$

Les triplets sont ici définis comme des nœuds reliés à deux autres, quel que soit l'ordre de ces liens. C est compris entre 0 et 1.

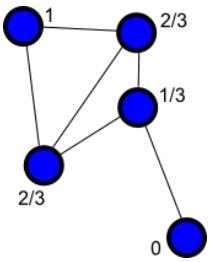


Ex : Le coefficient de clustering du graphe de la Figure 14 est de 0 alors que celui de la Figure 15 A est de 0.6.

Ce coefficient peut également être défini comme une propriété locale des nœuds (Watts and Strogatz, 1998), dans ce cas-là, le calcul se fait de la manière suivante :

$$C_i = \frac{nombre\ de\ triangles\ connectés\ à\ un\ noeud\ i}{nombre\ de\ triplets\ centrés\ sur\ le\ noeud\ i}$$

La moyenne de ces propriétés locales donne un autre coefficient de *clustering* global. Lorsque le nombre de triplets d'un nœud est nul, $C_i = 0$. Les deux coefficients peuvent différer considérablement.



Ex : Sur la Figure 14, tous les nœuds ont un coefficient de clustering local de 0 alors que sur la Figure 15 A, les nœuds ont des coefficients de 1, 2/3, 2/3, 1/3 et 0. Le coefficient de clustering moyen du graphe de la Figure 14 est donc de 0, alors que celui de la Figure 15 A est de 0.53. Il diffère de la définition précédente.

Communauté : un groupe de nœuds plus fortement connectés entre eux qu'avec les autres nœuds du réseau. Ce terme porte également le nom de *cluster* ou de module. Ce sont en fait des sous graphes $G' = (V', E')$ de G, à forte densité. Lorsque l'ensemble des nœuds d'une communauté sont interconnectés, cette dernière porte le nom de « clique ».

Détection de structure en communauté : technique de mise en évidence des structures de graphes à grande échelle (Fortunato, 2010). On parle alors d'algorithme de détection de communautés. Elles supposent qu'un réseau est naturellement composé de *clusters* plus petits et que le rôle du chercheur est de les retrouver. La taille des communautés n'est pas connue à l'avance pour ces méthodes.

Partitionnement : subdivision d'un graphe en plus petits composants avec certaines propriétés. Très utilisé dans le domaine du calcul parallèle, où le principe est de séparer les tâches sur plusieurs processeurs, de telle sorte qu'il y ait le moins de communication inter-processeurs possible (Ulrich Elsner, 1997). La taille des partitions peut être connue à l'avance. Les termes partitionnement et détection de communauté peuvent être confondus mais elles forment bien en réalité deux lignes de recherche différentes (Newman, 2006).

Modularité : différence entre le nombre de liens internes à des communautés (d'après une structure de communautés donnée) et le nombre attendu de liens si ces derniers étaient disposés au hasard (Newman, 2006). La modularité est donc calculée d'après un modèle nul, généralement avec une distribution des degrés des nœuds équivalente au vrai graphe. Elle donne en quelque sorte la qualité d'une séparation d'un graphe en communautés et est un paramètre que les algorithmes de détection de communautés cherchent à maximiser.

Considérons un graphe de n nœuds et m liens pouvant être séparé en deux communautés s. Soit A la matrice d'adjacence du graphe et $k_i k_j / 2m$, le nombre de liens attendus entre les nœuds i et j si les liens étaient placés au hasard. k_i, k_j sont les degrés des nœuds et m, le nombre de liens dans le graphe. Si un nœud i appartient à la communauté 1, $s_i = 1$ et si i appartient à la communauté 2, $s_i = -1$. Le nombre de liens entre i et j est donné par la matrice d'adjacence :

$$A_{ij} = \begin{cases} 1 \; s'il \; existe \; un \; lien \; (i,j) \\ 0 \; sinon \end{cases}$$

Pour toutes les paires de nœuds i et j d'un même groupe (pour chaque groupe), la modularité est alors définie par :

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \frac{(s_i s_j + 1)}{2}$$

Quelques exemples sont fournis sur la Figure 16 où seule la première partition montre une modularité positive.

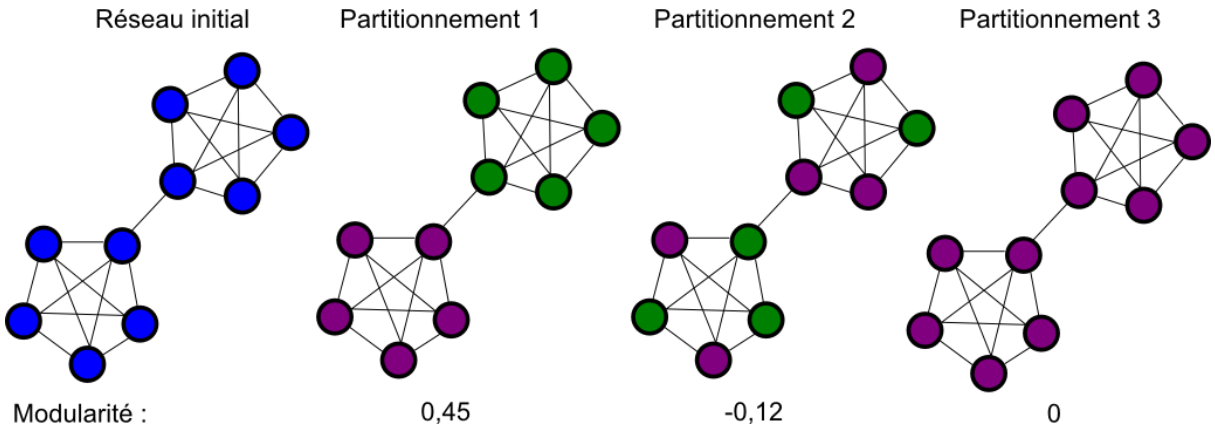


Figure 16. Exemples de mesures de modularité en fonction de différentes partitions. A gauche est représenté le réseau initial où l'on peut distinguer assez facilement deux communautés reliées par un lien. Le partitionnement 1 retrouve exactement les deux communautés. Le partitionnement 2 mixe les deux communautés ensembles et tient de l'aléatoire. Enfin, le partitionnement 3 considère l'ensemble du réseau comme une seule communauté.

Dans le cas de partitionnement de plus de deux communautés, la formule peut être généralisée pour « c » communautés :

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\sum_{l}\delta(c_i, l)\delta(c_j, l)$$

où, δ est le delta de Kronecker :

$$\delta_{kl} = \begin{cases} 1 \; si \; k = l \\ 0 \; si \; k \neq l \end{cases}$$

La modularité prend ses valeurs entre -1 et 1. Une modularité positive montre un partitionnement avec un nombre de liens intra-classes supérieur au nombre de liens attendus au hasard (présence de communautés). Lorsque la modularité est égale à 0, le partitionnement tient de l'aléatoire. Une valeur supérieure à 0,3 indique un bon partitionnement (Clauset et al., 2004)

Centralité : mesure « l'importance » des nœuds dans un réseau quant au potentiel de communication entre ces derniers. Comme l'importance peut être définie de plusieurs manières, il existe différents types de centralité : centralité de degré, centralité *betweenness* et centralité *closeness* (Freeman, 1978), etc. La centralité peut être considérée à un niveau

local (chaque nœud) mais également à un niveau global (un réseau), on parle parfois de mesure de centralisation dans ce dernier cas. Elles sont normalisées pour donner une mesure comprise entre 0 (non central) et 1 (central). Les définitions suivantes sont données pour un nœud (v).

Centralité de degré : degré normalisé d'un nœud.

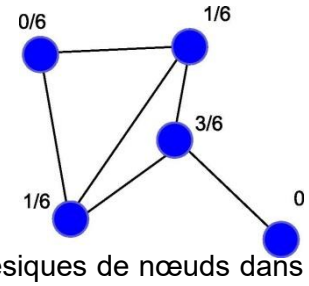
$$C_D(v) = \frac{\sum_{i=1}^{n} A_{iv}}{n-1}$$

Centralité *betweenness* : mesure basée sur la fréquence d'un nœud à être traversé par le géodésique de paires de nœuds dans le réseau. Les nœuds à forte centralité *betweenness* sont des points de contrôle particuliers dans un graphe : ce sont en quelque sorte des « ponts pour les liens » reliant différentes parties du réseau (à ce titre, il existe également une centralité *betweenness* pour les liens) et permettant le transit de l'information plus rapidement entre deux points quelconques du réseau :

$$C_B(v) = \sum_{i}^{n} \sum_{j \neq i \neq v}^{n} \frac{g_{ij}(v)}{g_{ij}}$$

où $g_{ij}(v)$ représente tous les géodésiques reliant les nœuds i et j contenant le nœud v et g_{ij} tous les géodésiques reliant les nœuds i et j avec $i \neq j \neq v$. La mesure est cependant dépendante de la taille du réseau, une généralisation de cette dernière permet de restreindre la mesure entre 0 et 1 et permet de fait la comparaison entre réseaux (normalisation par le nombre de couples de nœuds) :

$$C'_B(v) = \frac{1}{\frac{(n-1)(n-2)}{2}} C_B(v)$$

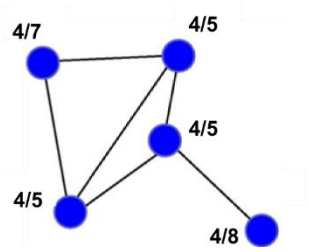


Centralité *closeness* : mesure également basée sur les géodésiques de nœuds dans le réseau. Contrairement à la centralité betweenness, les nœuds à forte centralité closeness sont des nœuds capables d'éviter le contrôle par d'autres nœuds. Ces nœuds ne sont donc pas dépendants des autres pour recevoir l'information qui transite dans le réseau : c'est une

forme d'importance via l'indépendance. Elle est le reflet d'une distance faible à tous les autres nœuds du réseau : plus le nœud est central, plus sa distance totale aux autres nœuds est faible.

$$C_C(v) = \frac{1}{\sum_{i \neq v} d(i,v)}$$

Cette mesure est généralement normalisée pour obtenir une valeur entre 0 et 1 :



$$C'_C(v) = C_C(v)(n-1)$$

Centralisation : propriété de centralité globale qui donne d'une certaine manière l'état de compaction du graphe. La mesure est calculée d'après la formule suivante pour chaque type de centralité locale évoqué :

$$C(G) = \frac{\sum_{i=1}^{n}(max_v\ C_v - C_i)}{\max\left(\sum_{i=1}^{n}(max_v\ C_v - C_i)\right)}$$

avec $\max(\sum_{i=1}^{n}(max_v\ C_v - C_i))$, le maximum théorique (qui dépend du type de graphe considéré) de la somme des différences des centralités locales des nœuds d'un graphe à n nœuds.

L'étude générale de toutes ces propriétés a montré que les réseaux réels (internet, réseaux sociaux ou réseaux biologiques) diffèrent grandement des réseaux dits « aléatoires », introduits par Erdös et Rényi au milieu du 20ème siècle (Erdős and Rényi, 1959). Dans un réseau aléatoire – probablement le modèle de réseau le plus étudié – la connectivité des nœuds suit une loi binomiale et chaque nœud est connecté de façon aléatoire avec une probabilité p (Figure 17 A). Le diamètre de ces réseaux, quant à lui, est proportionnel au logarithme du nombre de nœuds ($log(n)$). *A contrario*, l'étude de la structure des réseaux réels montre des distributions de degré des nœuds particulières (Figure 17), notamment suivant des lois de puissance (Figure 17 B) ($P_k \approx k^{-\gamma}$) et ont été nommés « réseaux sans-échelles »

(Barabási and Albert, 1999). Ce type de réseau est principalement dominé par quelques *hubs* reliant entre eux les nœuds peu connectés et sur lesquelles les nouveaux nœuds du réseau ont tendance à s'apparier avec une plus grande probabilité. Cette tendance des *hubs* à éviter de se connecter entre eux est appelée disassortativité, c'est d'ailleurs une des principales différences observées entre les réseaux sociaux – où les *hubs* ont plutôt tendance à se connecter à d'autres *hubs* – et les réseaux biologiques. Le diamètre de ce modèle de réseau

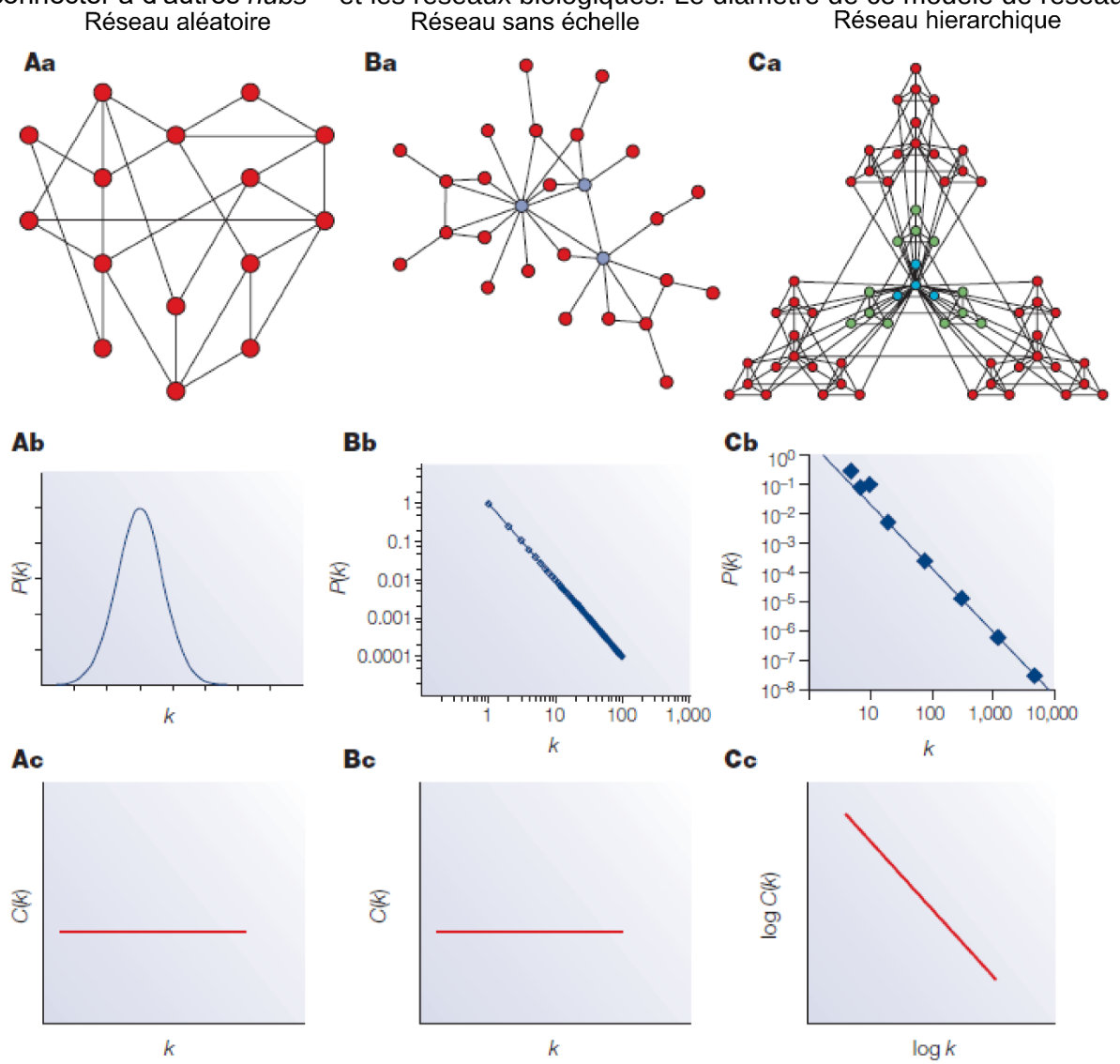


Figure 17. Modèles de réseau. A | a, Un réseau aléatoire où chaque nœud a une certaine probabilité d'être connecté à d'autres nœuds et un nombre de liens moyen égal à p((n(n-1))/2) ; b, La distribution des degrés des nœuds suit une loi de Poisson centré sur <k>, chaque nœuds a donc à peu près le même degré ; c, Le coefficient de *clustering* ne varie pas en fonction du degré des nœuds. B | a, Un réseau sans échelle avec trois *hubs* indiqués en bleu ; b, La distribution des degrés des nœuds est clairement différente et suit une loin de puissance où les *hubs* sont rares. En revanche, il existe beaucoup de nœuds peu connectés de par la forte probabilité de connexion des nouveaux nœuds aux *hubs* ; c, Tout comme pour les réseaux aléatoires, le coefficient de *clustering* ne varie pas en fonction du degré des nœuds. C | a, Un réseau hiérarchique (non abordé dans la thèse), où les clusters s'organisent de manière itérative. La construction du réseau est basée sur un plus petit module (en bleu) qui est alors répété trois fois et dont les nœuds sont reliés au nœud central du premier module ; b, La distribution suit également une loi de puissance ; c A cause du caractère itératif de la construction, le coefficient de *clustering* et le degré des nœuds sont anticorrélés. Tirée de (Barabási and Oltvai, 2004)

est proportionnel à $log(log(n))$ dénotant donc leur caractère très compact. Les réseaux hiérarchiques forment une autre catégorie de structure (Figure 17 C) mais que nous n'aborderons pas dans cette thèse.

La plupart des réseaux biologiques possèdent une forme sans échelle (Albert, 2005) et ils adoptent aussi la propriété de « petit-monde » (voir, ultra-petit-monde), où le géodésique entre la plupart des paires de nœuds est très faible (Watts and Strogatz, 1998). En biologie, ces structures spécifiques semblent être particulièrement liées à l'évolution, où de nouvelles fonctions (ou des modifications d'anciennes fonctions) viennent s'ajouter à celle déjà existantes tout en apportant un avantage à l'espèce dans son adaptation à l'environnement (Hartwell et al., 1999). Nous pouvons noter que cet exemple typique d'adaptation via l'acquisition ou la modification de fonction(s) fait également intervenir les notions de robustesse, de résilience et de flexibilité des systèmes. La robustesse est une notion qui fait référence à la capacité d'un système à subsister et fournir un résultat attendu malgré des perturbations internes ou externes (Kitano, 2004, 2007). La résilience d'un système est une notion proche de la robustesse, mais qui diffère dans le sens où un système résiliant est capable de retrouver son fonctionnement normal malgré les perturbations (Krantz et al., 2009). La résilience est très souvent confondue avec la robustesse bien que les deux concepts soient assez différents l'un de l'autre. De façon très imagée, cette subtile différence réside dans la capacité du système à absorber les dégâts (l'attaque d'un nœud p.ex.) : un système robuste pourrait endurer beaucoup d'attaques sans influence sur sa fonction, alors qu'un système résilient reviendrait à son état d'origine après chaque attaque. Enfin, la flexibilité fait également référence à la capacité d'un système à s'adapter en modifiant (ou en acquérant) de nouvelles fonctions sans pour autant impacter négativement sur les fonctions initiales (Hunter, 2009). En termes de réseaux, tous ces concepts renvoient à l'étude de la stabilité du transit de l'information sous la perturbation des nœuds (ou des liens) du réseau. Une des constantes des réseaux biologiques est qu'ils sont souvent à la fois robustes, résilients et flexibles (Meir

et al., 2002; Csete and Doyle, 2004; Kitano, 2004; Dartnell et al., 2005; Prill et al., 2005; Ciliberti et al., 2007; Viana et al., 2009; Gáspár and Csermely, 2012).

3. Réseaux biologiques

Les réseaux biologiques se construisent à partir de données expérimentales (ou prédictives). L'exemple typique de réseau biologique est le réseau des voies métaboliques, qui cherche à représenter les réactions chimiques dans les cellules. Dans les réseaux métaboliques, les nœuds sont des substrats ou des produits, reliés par des liens orientés représentant les réactions chimiques (Figure 18 A). Un autre exemple très commun est le réseau des interactions protéine-protéine (*protein-protein interaction network – PPI*). Dans ce cas, les graphes sont souvent simples et non-orientés ; chaque nœud représente une protéine et chaque lien, une interaction physique entre deux protéines (un complexe protéique p.ex., Figure 18 B). Les réseaux neuronaux sont un autre exemple de réseau biologique où les nœuds sont des neurones et les liens des liaisons synaptiques permettant le transfert de signaux (Figure 18 C). Nous pouvons également citer d'autres exemples, moins communs, comme les chaines alimentaires, où chaque nœud est un animal/végétal et un lien orienté représente le lien prédateur/proie (Figure 18 D) ; les réseaux vasculaires, où chaque nœud désigne une séparation vasculaire et les liens sont les vaisseaux sanguins eux-mêmes ; ou encore les arbres phylogéniques, pour lesquelles les nœuds sont des espèces et les arcs des liens de parenté entre espèces. Enfin, les réseaux géniques déterminent les effets des gènes les uns sur les autres (induction/répression) et montrent les boucles de rétroaction (Figure 18 E).

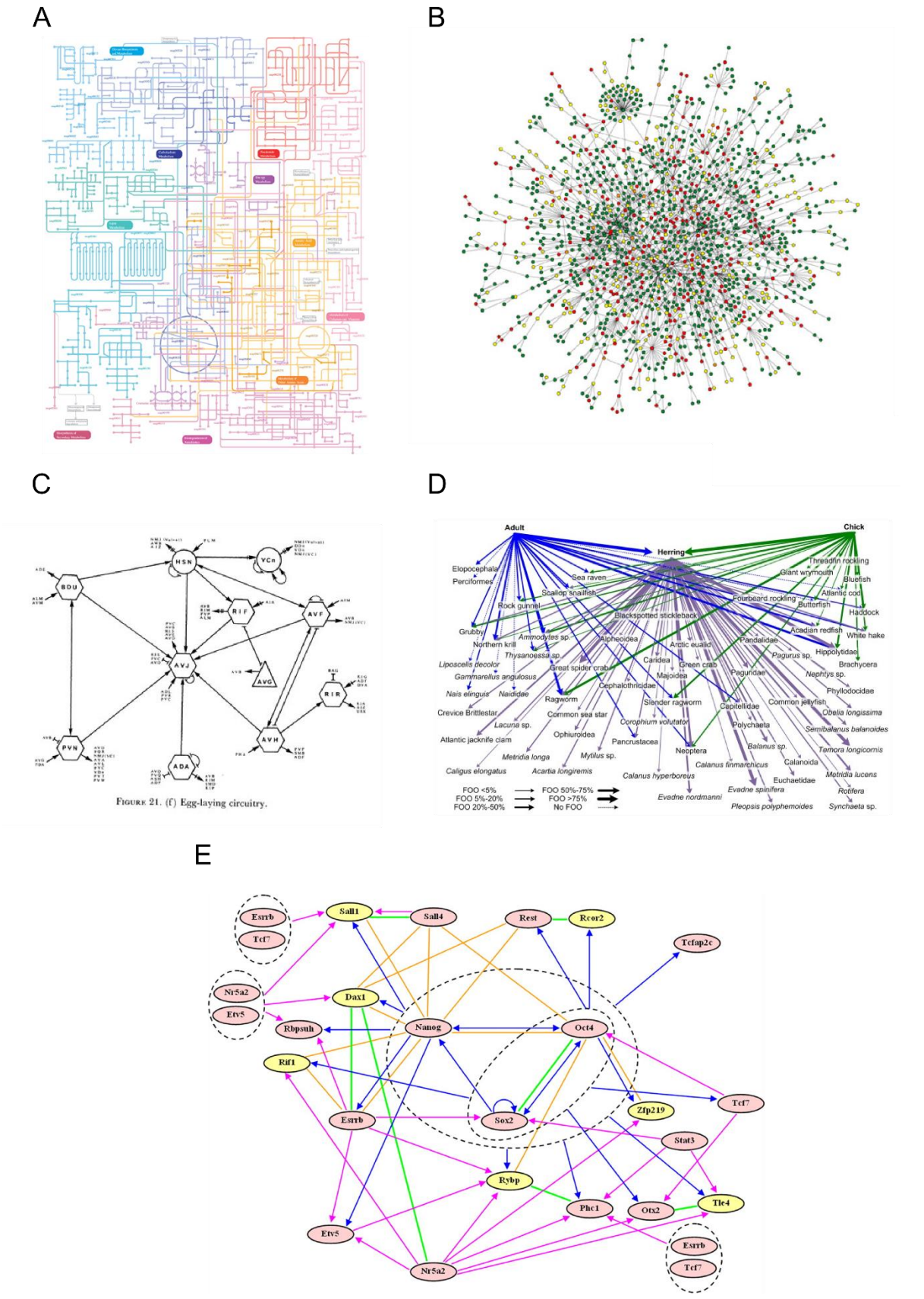


Figure 18. Différents types de réseaux biologiques. A | Le réseau métabolique de KEGG (*Kyoto encyclopedia of genes and genomes*) d'après (Kanehisa et al., 2007). B | Réseau d'interaction protéine-protéine chez la levure d'après (Jeong et al., 2001; Barabási and Oltvai, 2004). C| Un réseau d'interaction neuronal chez *C. elegans* d'après (White et al., 1986) . D | Réseau représentant la chaine alimentaire de F. *arctica* et *C. harengus* d'après (Bowser et al., 2013). E | Réseau de régulation génique dans les cellules souches embryonnaires de souris d'après (Zhou et al., 2007).

D'un point de vue cellulaire, les réseaux biologiques forment non pas un ensemble de réseaux séparés et indépendants mais plutôt une superstructure de réseaux intriqués (un réseau de réseaux) responsable du comportement des cellules (Barabási and Oltvai, 2004; Csete and Doyle, 2004). Ainsi, les réseaux d'interactions PPI sont entrelacés avec ceux des réactions métaboliques et de la signalisation mais également avec ceux de la régulation génique, que ces derniers soient liés aux miARN, aux facteurs de transcription ou à tout autre régulateur. Cette superstructure est par ailleurs dynamique : elle évolue non seulement dans le temps mais répond aussi aux stimuli de l'environnement afin de maintenir l'état de la cellule, quelles que soient les situations qu'elle rencontre. Toute la robustesse et la résilience de ce super-réseau tient notamment dans ses boucles de rétroaction (Kwon and Cho, 2008) et les plus petits modules qui la composent (Prill et al., 2005). Cette vision intégrée est cependant encore peu étudiée mais de nombreuses études se focalisent sur la structure et les propriétés de chacun de ses réseaux indépendamment, notamment sur les aspects de communautés et de boucles de régulation.

Il est alors intéressant de chercher des relations entre la structure physique de ces (sous) réseaux et leur fonction, leur robustesse et leur évolvabilité. Par exemple, les réseaux métaboliques et les réseaux de la transcription/traduction montrent tous les deux des structures en forme de nœud papillon où tout un ensemble de composés (formant une partie du nœud papillon) sera transformé en un autre ensemble de composés (la deuxième partie) au travers d'un petit nombre d'effecteurs (le nœud) (Figure 19). Pour les réseaux métaboliques, le premier éventail est représenté par les nutriments, présents dans l'environnement et qui seront transformés en quelques briques essentielles (sucres, des acides aminées ou encore des acides gras), ce qui permettra de fournir de l'énergie à la cellule mais également de générer des composés plus complexes pour sa survie. Une partie de ces briques (acides aminés et acides nucléiques notamment) sera également utilisée pour la transcription de tout un ensemble de gènes, qui seront eux même pris en charge par un faible nombre de complexes ribonucléoprotéiques afin de créer toutes les protéines nécessaires à

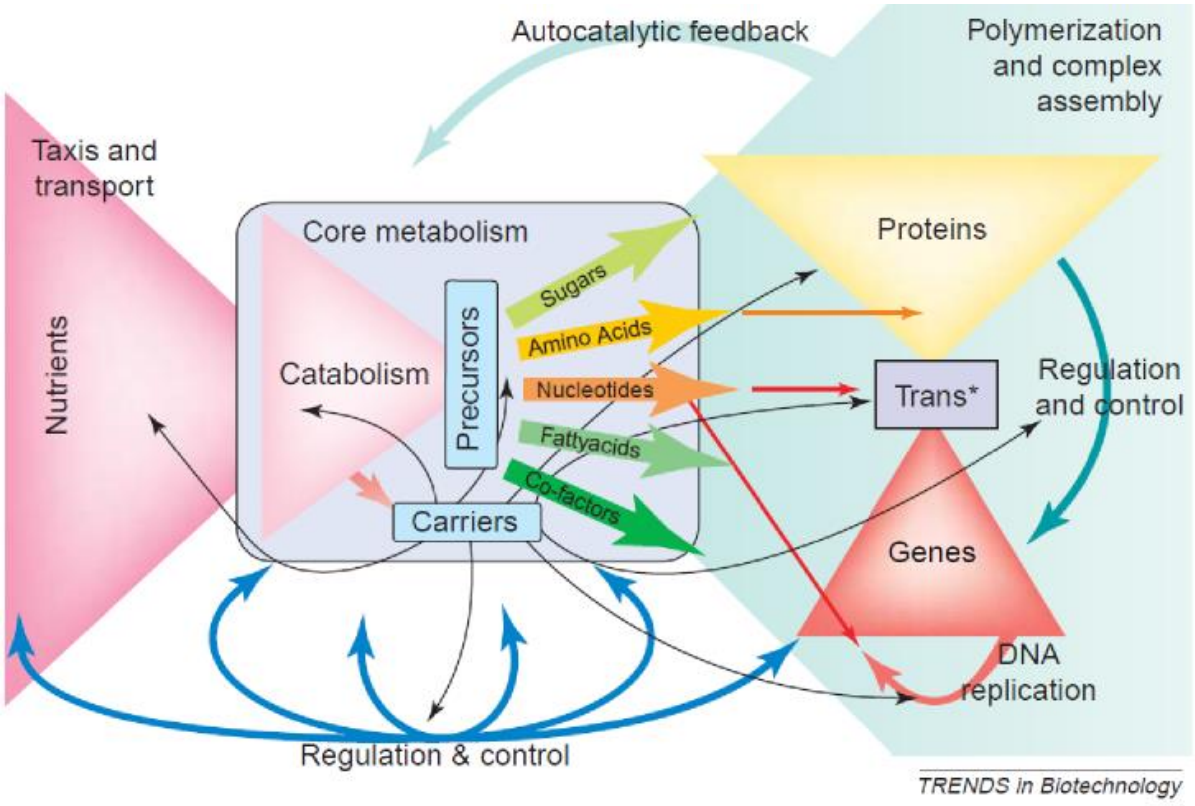


Figure 19. Réseaux intriqués en nœud papillon. D'un côté, le réseau métabolique transforme un éventail de nutriments (rose) en quelques précurseurs qui serviront d'énergie et de composés à la cellule (bleu). Au sein de ces composés, certains seront utilisés au cours de la réplication, la transcription et la traduction, qui forment également un réseau en nœud papillon et dont le nœud est formé cette fois-ci par des ribosomes et autres complexe nécessaire à la traduction. Tirée de (Csete and Doyle, 2004).

la cellule. Les deux réseaux sont donc interconnectés et l'ensemble est bien évidemment contrôlé de manière très fine par des boucles de régulation (flèches bleus de la Figure 19). L'avantage d'avoir un très petit ensemble d'effecteurs est qu'il permet « *la facilitation du contrôle, l'accommodation aux perturbations et aux fluctuations, à plusieurs échelles temporelles et spatiales* » (Csete and Doyle, 2004). Les effecteurs sont en fait particulièrement transposables (utilisables dans d'autres situations et environnements) et permettent donc l'adaptation à de nouveaux environnements, sans besoin de grandes modifications. Ce dernier constat démontre le caractère évolutif de la structure en nœud papillon et c'est au travers de ce caractère qu'en découle toute sa robustesse.

4. Réseaux et microARNs

Concernant les réseaux de miARN, bien que les premiers algorithmes de prédiction de cibles soient apparus dès 2003, les études systémiques sur ces régulateurs n'ont été réalisées

que quelques années plus tard. De prime abord, ces dernières consistaient essentiellement en la recherche de boucles de régulation entre gènes, facteurs de transcription et miARN. Elles ont toutefois très vite basculées vers la recherche des liens entre miARN et processus biologiques ou miARN et pathologies. L'objectif de ces études était (et reste) en fait la compréhension de la place des miARN dans le réseau de régulation génique.

Dans cette sous-partie, nous étudierons quelques travaux ayant eu pour but la construction et l'analyse de réseaux de miARN et ayant servi de base de réflexion pour les travaux de cette thèse. Nous verrons tout autant des réseaux miARN-miARN que des réseaux bipartites miARN-gène, miARN-facteur de transcription ou même miARN-pathologie.

En 2006, Tsang et collaborateurs ont mis au point une méthodologie bioinformatique afin de mettre en évidence des boucles de régulation dans le génome de l'être humain et de la souris sans utiliser d'algorithmes de prédiction (Tsang et al., 2007). Afin d'estimer le niveau d'expression des miARN, ces auteurs se sont basés sur les miARN introniques situés dans les régions codantes de protéines, dont l'expression est supposée être corrélée à celle des protéines. De plus, pour s'affranchir de la différence d'expression qui peut exister au sein d'un même tissu, ils n'ont étudié qu'une population de cellules neuronales homogène. La détermination de l'implication d'un miARN dans des boucles de régulation de type I (ou de type II) a été effectuée en se basant sur la corrélation (ou anticorrélation) d'expression entre un miARN et tous les gènes testés, mais également en cherchant à déterminer si la conservation des sites de liaisons dans les 3'UTR d'ensembles de gènes (région *seed* d'un miARN) n'était pas liée au hasard. La question à laquelle cherchait à répondre les auteurs était donc la suivante : est-ce que certains miARN montrent un biais envers l'un ou l'autre type de boucles de régulation ? Ce à quoi ils ont répondu par l'affirmative : les miARN ont plus tendance à être impliqués dans des boucles de type I. D'autres études se sont focalisées sur les boucles de régulation impliquées dans un sous type cellulaire comme par exemple sur le glioblastome (Sun et al., 2012), le cancer (Hsieh et al., 2015) ou encore les maladies cardiaques (Lin et al., 2015).

Plus proche des travaux de cette thèse, Shalgi *et al.* ont construit et analysé en 2007 plusieurs types de réseaux (miARN-miARN et miARN-facteur de transcription) afin de mieux comprendre le rôle conjoint entre ces deux régulateurs géniques (Shalgi et al., 2007). L'étude de leur réseau miARN-miARN, basé sur le partage de cibles prédites (dont la construction est sensiblement équivalente à celle de la Figure 20 A), a montré des interactions combinatoires entre les miARN qui partageaient des cibles communes. La structure de ce réseau montrait par ailleurs des caractéristiques d'un réseau sans échelle et « petit-monde », dénotant l'importance de quelques *hubs* interagissant avec d'autres miARN moins connectés. Certains gènes partagés semblaient par ailleurs être bien plus ciblés que d'autres. Ils ont donc nommé ces gènes « cibles *hubs* ». Beaucoup de ces « gènes *hubs* », potentiellement corégulés, étaient en fait des régulateurs de la transcription et par conséquent, associés à des processus de développement. Ce constat est particulièrement intéressant puisque les gènes impliqués dans des processus biologiques liés au développement semblent eux même fortement régulés (Borneman et al., 2006). Le réseau constitué de miARN et de facteurs de transcription, basé sur le partage de cibles entre ces deux entités, a montré quant à lui la présence de beaucoup de motifs de régulation. Ils ont également montré que l'expression entre ces couples de miARN et les facteurs de transcription était sensiblement corrélée (ou anti-corrélée) : des résultats qui sont expliqués par différents autres types de boucles de régulation. Cette étude reste une des premières à avoir adopté un point de vue mathématique et statistique pour l'analyse de réseaux de miARN.

Sur un principe assez proche de l'étude précédente, Jiang et collaborateurs ont construit un réseau à deux niveaux (Jiang et al., 2010) : un réseau miARN-miARN basé sur le partage de cibles dans un premier temps (Figure 20 A) et un réseau de pathologies dans un second temps, les deux réseaux étant également reliés par des liens d'interaction miARN-maladie (Figure 20 B), tirés de base de données de référence comme miR2Disease (Jiang et al., 2009) et HMDD (Lu et al., 2008). L'objectif de cette étude était d'identifier de nouveaux miARN impliqués dans certaines maladies en se basant sur le partage de cibles avec d'autres miARN connus pour être impliqués dans ces maladies. L'algorithme recherchait des communautés de miARN partageant les mêmes cibles et donnait un score à ces communautés en fonction du nombre de miARN impliqués dans une certaine pathologie. Ainsi pour chaque maladie, les auteurs donnaient une liste de miARN à « prioriser » pour cette dernière. Les

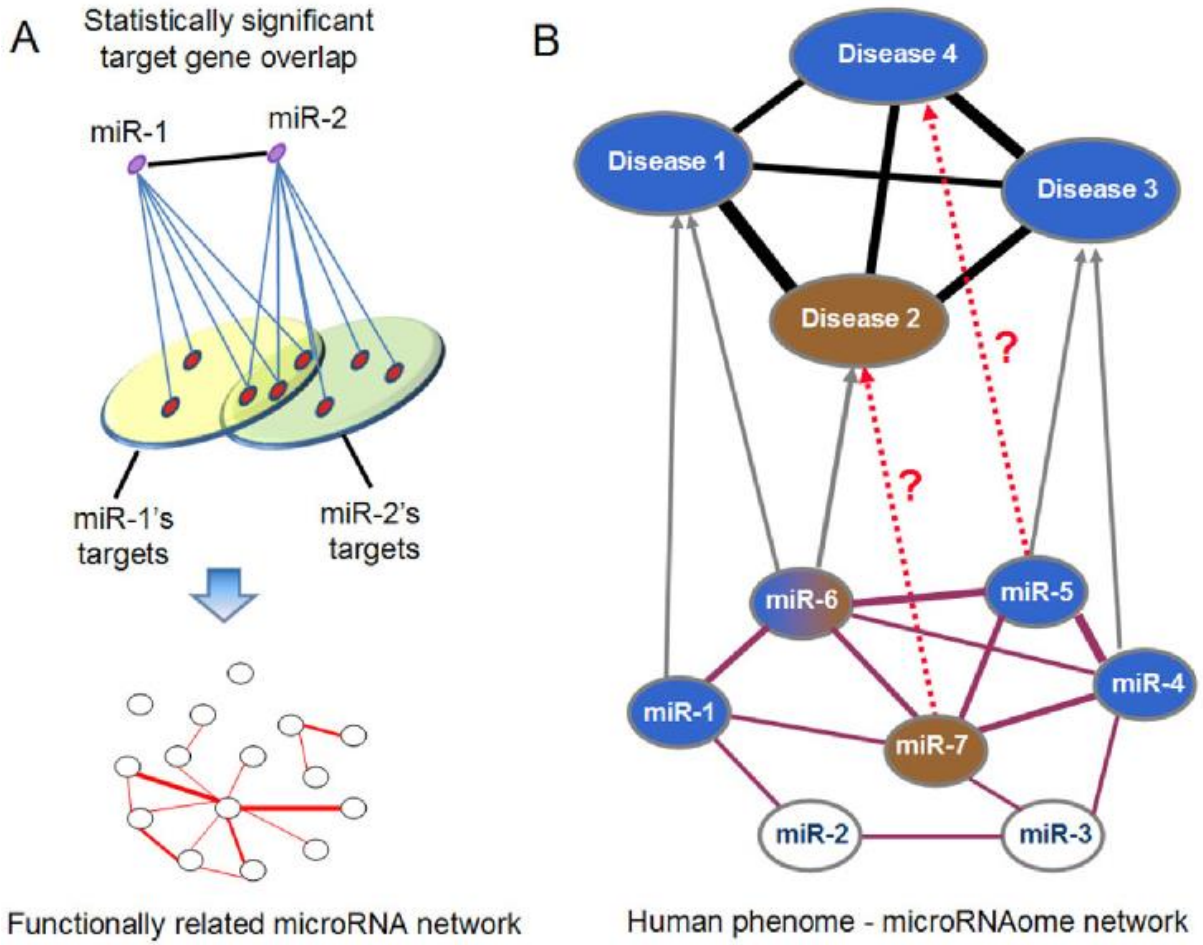


Figure 20. Construction de réseaux d'après (Jiang et al., 2010)**.** A | Construction du réseau miARN-miARN où un lien est créé entre deux miARN en fonction du nombre de cibles partagées. B | Construction du second réseau de pathologies et liaison avec le premier réseau de miARN. Les liens rouges montrent les miARN potentiellement impliqués dans les maladies de par leurs connections à d'autres miARN qui le sont réellement. Tirée de (Jiang et al., 2010)

auteurs ont également cherché à connaître la robustesse des algorithmes de prédiction en reproduisant l'étude avec deux algorithmes différents (TargetScan et PITA). Malgré les très grandes différences de prédiction entre les algorithmes, les performances de leur modèle montraient des résultats très comparables. Afin de valider en partie leur approche, Jiang *et al.* ont effectué une courte revue de littérature sur le cancer du sein.

Poussant la construction et l'analyse de réseaux un peu plus loin, Xu et collaborateurs ont intégré des informations de prédictions de cibles des miARN, d'analyse d'ontologies et d'interactions protéine-protéine afin de construire un réseau miARN-miARN synergétique (Xu et al., 2011a). Pour la construction du réseau, les auteurs ont dans un premier temps extrait les cibles partagées entre des paires de miARN. Ils ont alors conduit une analyse d'enrichissement d'ontologies sur ces cibles communes afin de trouver des « modules fonctionnels ». Une analyse d'enrichissement d'ontologie cherche à découvrir quels processus biologiques sont surreprésentés dans un set de gènes donnés, d'après un test hypergéométrique. A chaque processus est alors associée une P-valeur ainsi que l'ensemble de gènes impliqués dans ce processus. C'est en fait cet ensemble de gènes que les auteurs ont appelé « modules fonctionnels ». Dans leur étude, un module n'a été considéré que lorsque les gènes le constituant interagissaient (distances et chemin caractéristique faibles entre les gènes) dans un réseau d'interaction protéine-protéine (tiré de la base de données HPRD (Keshava Prasad et al., 2009)). Les auteurs ont placé un lien entre deux miARN si et seulement s'ils co-régulaient au moins un module fonctionnel. Le réseau de miARN ainsi édifié montrait des caractéristiques de petit-monde, sans échelle et était constitué de modules. Ce constat a permis aux auteurs de conclure que, de par ces caractéristiques, les miARN pouvaient potentiellement répondre à des perturbations de manière très rapide mais également que les miARN agissaient de manière synergique pour répondre à ces perturbations. Ils ont également mis en évidence un lien entre la proximité physique et la topologie de groupes de miARN et leur implication dans différentes pathologies. De la même manière que les précédentes études, des réseaux basés sur différents algorithmes de

prédiction ont également été construits afin de déterminer la robustesse de leur approche. Il est intéressant de noter qu'une fois encore, si les prédictions diffèrent énormément en fonction des algorithmes, la construction des réseaux et leur analyse semblent bien plus robuste.

Un des exemples les plus frappants d'analyse miARN-pathologie reste l'étude de Volinia et collaborateurs qui ont probablement mis au point une des plus grosses études intégratrices sur les miARN et leur implication dans les cancers (Volinia et al., 2010). En se basant sur l'expression d'environ 100 types de tissus différents sains et cancéreux représentant quelques 4000 échantillons humains, ces auteurs ont construit un ensemble de réseaux bayésiens miARN-miARN où chaque lien représentait en quelque sorte la probabilité que les deux miARN soient coexprimés dans un ensemble de tissus (soit sains, soit cancéreux). Le réseau construit d'après les tissus sains montrait un graphe complet où chaque miARN était connecté à au moins un autre miARN. A l'inverse, les « réseaux cancéreux » montraient des sous-réseaux isolés les uns des autres, avec des différences topologiques notables entre cancers solides et leucémies. Ainsi les miARN impliqués dans ces pathologies ne semblaient plus être sous un contrôle coordonné comme dans les cas normaux. L'analyse des gènes cibles de ces miARN particuliers montrait aussi de clairs enrichissements pour les processus cancéreux. Pour résumer, ces auteurs ont pu montrer, au travers de leur étude, à quel point le réseau de régulation de miARN était perturbé dans les cancers.

D'autres auteurs se sont également intéressés au lien entre les miARN et les cancers d'un point de vue réseau (à plus ou moins grande échelle). C'est le cas par exemple des études d'Aguda et collaborateurs sur la régulation dynamique et réciproque entre les miARN du cluster miR-17-92 et les facteurs de transcription E2F et Myc, tous deux impliqués dans le contrôle de la prolifération cellulaire et l'apoptose (Aguda et al., 2008). Bandyopadhyay *et al.* ont utilisé, quant à eux, une approche d'exploration de données (*data mining*) de liens expérimentaux entre miARN et cancers tirés de la littérature (Bandyopadhyay et al., 2010). Construisant un réseau miARN-cancers, ces auteurs ont pu montrer que les modules de miARN retrouvés impliqués dans les mêmes cancers pouvaient avoir des effets combinatoires,

étant donné que ces miARN partageaient souvent des oncogènes ou des tumeurs suppresseurs. Ils ont également mis en évidence certains miARN *hubs*, impliqués dans un grand nombre de cancers. Un de ces miARN était d'ailleurs miR-21, un des miARN le plus étudié chez l'être humain. Enfin, ils ont montré que les motifs de dérégulation des miARN étaient dépendants du type de tissu considéré et qu'il existait des modules de miARN impliqués dans plusieurs cancers. Toujours sur ce lien miARN-cancer, Plaisier et coll. ont mis au point une méthodologie appelée FIRM (*framework for inference of regulation by miRNAs*) afin de trouver des miARN potentiellement régulateurs de processus cancéreux (Plaisier et al., 2012). Le but de cette méthodologie était de créer des réseaux miARN-cancer basés sur la découverte de cibles potentielles des miARN à partir de motifs de liaison dans les 3'UTR et la coexpression de ces cibles dans des données d'expression. Les auteurs ont ainsi identifié et validé 13 miARN impliqués dans ces processus.

Il existe bien évidemment beaucoup d'autres méthodologies pour inférer des réseaux de miARN, quel que soit leur type. L'algorithme LeMoNe (Bonnet et al., 2010) par exemple est également une méthode d'inférence se basant sur des données d'expression, comme la simple utilisation d'un coefficient de corrélation (Xu et al., 2011b). Une méthodologie qui semble également assez commune dans la construction de graphes à partir de données d'expression est la régression lasso, une forme de régression linéaire (Lu et al., 2011; Alshalalfa et al., 2012; Qabaja et al., 2013). Dans l'étude de Volinia et al., c'est un algorithme appelé Banjo qui avait été utilisé mais il existe également des méthodes bayésiennes permettant d'obtenir d'autres types de réseaux et pouvant porter une information légèrement différente (Zacher et al., 2012). L'objectif principal de toutes ces études reste cependant toujours le même : la compréhension de la place des miARN dans la régulation de l'expression des gènes, dans le devenir de nos cellules et leur implication dans diverses pathologies.

La conclusion que nous pouvons tirer de cet ensemble d'analyses est que les miARN sont des régulateurs fins de l'expression des gènes et qu'ils semblent agir comme des rhéostats et des tampons qui diminuent la variation d'expression génique. Ces derniers

apportent de la robustesse au réseau de la régulation génique en modulant des cibles *hubs* ou particulièrement centraux au réseau. Cette régulation est souvent faite sous forme de modules et de manière combinatoire, voire synergique.

C. miARN en santé humaine

Outre la compréhension du rôle des miARN dans nos cellules, l'intérêt de l'étude des miARN, à un niveau systémique ou non, réside également dans des aspects diagnostiques et thérapeutiques. En effet, de nombreuses études montrent que les miARN sont dérégulés dans une multitude de pathologies. La mesure objective et quantitative de cette dérégulation peut dans une certaine mesure être utilisée comme marqueur de l'état cellulaire (sain/pathologique) : on parle dans ce cas de miARN biomarqueurs. Les biomarqueurs peuvent avoir une visée pronostique (déterminer l'évolution et l'aboutissement d'une maladie, sert à stratifier la maladie) ou diagnostique (déterminer la présence d'une maladie). L'utilisation des miARN n'est cependant pas limitée à celle de biomarqueurs puisqu'ils pourraient aussi être utilisés comme agents thérapeutiques via l'augmentation ou la diminution artificielle de leur concentration cellulaire. Pour aller encore un peu plus loin, les miARN pourraient même être envisagés comme composés théranostiques. La théranostique est la contraction des termes « thérapie » et « diagnostique » c'est en fait le couplage entre un outil diagnostique permettant le choix de la thérapie adaptée. Dans le cas des miARN, le diagnostic et la thérapie pourraient être le miARN en lui-même (test d'expression et restauration/répression).

1. Les miARN comme biomarqueurs

C'est dès 2005 que le potentiel diagnostique des miARN a été perçu. Dans une étude cherchant à mesurer l'expression de 217 miARN sur 334 échantillons dont des cancers, Lu et collaborateurs ont montré qu'ils étaient capables de différencier des tissus cancéreux de tissus sains, mais également de classifier (avec plus ou moins de facilité) les différents types de cancers qu'ils considéraient grâce aux miARN (Lu et al., 2005).

Lorsqu'on examine l'expression des miARN pour définir leur potentiel diagnostique, il semble en fait plus facile de considérer une signature plutôt qu'un miARN isolé. Ceci est probablement dû aux effets souvent limités d'un miARN seul sur le devenir des cellules (Vidigal and Ventura, 2014). Une signature est en fait un ensemble de miARN induits et/ou réprimés dans les tissus malades par rapport aux tissus sains. Un exemple de ce type de signature – découverte par Jung et collaborateurs – est l'ensemble de 5 miARN induits et de 6 miARN réprimés dans des cellules cancéreuses rénales et qui permet de séparer de façon très robuste les tissus cancéreux des tissus sains (Jung et al., 2009). En réalité, deux des miARN (miR-141, réprimé dans les tissus cancéreux et miR-155, induit dans les tissus cancéreux) suffisent à classifier presque parfaitement les deux échantillons que les auteurs ont testés. De la même manière, de nombreuses signatures pour d'autres cancers ont été étudiés : leucémie (Calin et al., 2005), pancréas (Lee et al., 2007), colon (Schepeler et al., 2008), prostate (Schaefer et al., 2010; Martens-Uzunova et al., 2012), poumon (Yanaihara et al., 2006), sein (Mattie et al., 2006), etc. Il existe en fait une pléthore de travaux de ce genre, cherchant à discriminer la pathologie des tissus sains. La plupart des signatures permettent de distinguer les tissus cancéreux des tissus sains dans 80 à 90% des cas et le recouvrement entre miARN dérégulés entre différentes études peut être assez faible, bien que certains semblent plus souvent retrouvés que d'autres (p.ex. miR-21). Un des problèmes constant en revanche est la baisse de ces chiffres de distinction cancéreux/sains – parfois de façon très significative – lorsqu'une signature donnée est utilisée sur d'autres études indépendantes. Malgré ce problème, un test de laboratoire a tout de même été créé et commercialisé par la société Asuragen sous le nom de miRInform® Pancreas. Ce test permet de diagnostiquer les adénocarcinomes canalaires pancréatiques en se basant sur l'expression différentielle entre miR-196a et miR-217 (Szafranska-Schwarzbach et al., 2011). L'utilisation de cette technologie, en plus des tests utilisés à l'heure actuelle pour détecter ce type de cancer, permet l'amélioration significative des diagnostics posés par les praticiens.

Pour le caractère pronostique des miARN en revanche, ce sont très souvent des miARN isolés qui sont étudiés et qui permettent de stratifier l'état de la maladie. Par exemple, l'expression de miR-196a est plus forte dans les adénocarcinomes œsophagiens, le syndrome de Barrett (métaplasie des cellules de la partie basse de l'œsophage) et les lésions dysplasiques, comparés aux muqueuses normales. Son expression est également plus forte dans les dysplasies de haut grade par rapport au syndrome de Barrett et aux dysplasies de faible grade, indiquant donc une corrélation entre l'état d'avancement des adénocarcinomes œsophagiens et l'expression du miARN, ce qui permet d'envisager son potentiel pronostique (Maru et al., 2009). Dans le cas de la prostate, c'est l'expression du miR-221 qui montre une anticorrélation avec la progression tumorale (Spahn et al., 2010; Kneitz et al., 2014) alors que pour les carcinomes hépatiques, c'est la perte de miR-122 qui est associée à la progression de la maladie (Coulouarn et al., 2009). L'expression de miR-21, que nous avons déjà évoqué, semble significativement corrélée avec l'état d'avancement du cancer du sein (Yan et al., 2008) alors que l'expression de miR-155 semble être associé, quant à elle, à une meilleure survie des patients atteints du cancer des poumons (Raponi et al., 2009; Donnem et al., 2011).

Si les premières études sur l'utilité des miARN en tant que biomarqueurs ont été poursuivies essentiellement sur des échantillons de biopsies humaines, on voit aujourd'hui de plus en plus d'études cherchant à identifier les miARN circulants. En effet, les miARN ont été isolés dans de nombreux fluides corporels, que ce soit le sérum, le plasma, les urines, la salives, le lait maternel ou encore le liquide lacrymal (Cortez et al., 2011). L'avantage de ces miARN circulants est leur grande stabilité – due à leur protection par des vésicules lipidiques comme les exosomes (Cheng et al., 2014) ou les corps apoptotiques (Zernecke et al., 2009) ou encore par des lipoprotéines à haute densité (HDL) (Vickers et al., 2011). Il est donc assez facile de récupérer ces miARN (prise de sang p.ex.) et mesurer leur niveau d'expression (qPCR p.ex.).

La première étude ayant proposé la possibilité d'utiliser les miARN circulant comme biomarqueur est l'étude de Mitchell et coll. qui ont montré que miR-141 était significativement

augmenté dans le sérum des patients atteints de cancer avancé de la prostate (Mitchell et al., 2008). Depuis, de nombreuses études ont dévoilé ce potentiel pour divers cancers, autant pour le diagnostic (Lodes et al., 2009; Bryant et al., 2012; Chiam et al., 2015; Erbes et al., 2015; Hou et al., 2015; Shin et al., 2015) que le pronostic (Brase et al., 2011; Moltzahn et al., 2011; Antolín et al., 2015; Wozniak et al., 2015). Les miARN circulants ne sont évidemment pas étudiés que dans le cas des cancers mais aussi pour les malformations congénitales comme celles touchant le cœur (Zhu et al., 2013) ou plus singulièrement, celles affectant le statut des spermatozoïdes (Wu et al., 2012). Malheureusement, les miARN circulant n'améliorent pas vraiment la sensibilité et la spécificité à distinguer les cellules saines des autres cellules par rapport aux miARN présents dans les tissus, puisque les chiffres observés sont sensiblement identiques aux analyses sur tissus. En revanche, leur étude est bien plus simple. Les deux méthodologies gardent tout de même un fort potentiel puisqu'elles montrent toutes deux des chiffres bien supérieurs à ce que l'expression des gènes seuls permet d'obtenir. Des études à bien plus grande échelle et intégrées permettraient sans doute d'obtenir de meilleures statistiques afin de déterminer plus aisément des signatures diagnostiques et/ou pronostiques spécifiques.

2. L'ARNi thérapeutique

Les miARN peuvent être utilisés de plusieurs manières pour des aspects thérapeutiques : soit en augmentant artificiellement leur expression cellulaire, que ce soit par l'utilisation de mimic (Trang et al., 2011), de pre-miARN ou tout autre composé ; soit en diminuant leur expression ou en empêchant leur fonctionnement par l'utilisation de LNA, d'antagomiR ou d'éponges à miARN par exemple. De nombreuses études montrent que les miARN ont un intérêt thérapeutique d'un point de vue préclinique, mais nous ne nous focaliserons ici que sur les exemples les plus marquants, notamment ceux ayant passés les phases précliniques.

Le premier miARN que nous évoquerons est miR-122, que nous avions déjà évoqué pour des aspects diagnostiques dans les carcinomes hépatiques. Ce miARN, qui semble assez spécifique des cellules hépatiques, est associé à la réplication du virus de l'hépatite C (Jopling et al., 2005; Chang et al., 2008; Jangra et al., 2010). Ainsi, l'utilisation de constructions anti-sens cherchant à bloquer le miR-122 permettait de réduire significativement la réplication virale (Jopling et al., 2005). Depuis, plusieurs inhibiteurs ont montrés un effet positif sur la répression du virus. Un de ces inhibiteurs est Miravirsen, un inhibiteur du miR-122 (LNA) ayant terminé sa phase 2a et développé initialement par Santaris Pharma (racheté récemment par le groupe pharmaceutique suisse Roche) (de Jong and Jacobson, 2014). Ce composé est en fait le tout premier inhibiteur de miARN à être entré en phase clinique. Aujourd'hui, ce n'est pourtant plus l'unique inhibiteur de miR-122 puisque la compagnie Regulus Therapeutics possède également son propre inhibiteur, actuellement en phase 2. D'autres composés en phase préclinique sont également à l'étude.

miR-34 est un miARN particulièrement intéressant puisqu'il semble impliqué dans au moins une quinzaine de cancers différents (Bader, 2012). Dans ces différents cancers, il a principalement un rôle de tumeur suppresseur, c'est-à-dire qu'il inhibe la progression tumorale lorsqu'il est induit. Il n'est donc pas étonnant de voir un mimic de ce miARN en phase clinique afin de traiter le cancer du foie. Ce composé, appelé MRX34, est le premier mimic à être entré en phase clinique ; il est produit par la société miRNA Therapeutics. Il est actuellement en phase 1, phase qui devrait se terminer d'ici fin 2015. Plus récemment, l'ADRI (*Australia's Asbestos Diseases Research Institute*) cherchait à recruter pour une étude de phase 1 concernant un mimic de miR-16 contre le mésothéliome (Reid et al., 2013), une forme de cancer mésothélial.

Ces trois produits sont pour le moment les seuls à être entrés en phase clinique même si de nombreux autres exemples peuvent être trouvés en phase préclinique. Nous pouvons par exemple citer miR-155 pour les maladies inflammatoires (Worm et al., 2009), miR-208 pour les maladies cardiaques (Montgomery et al., 2011), miR-21 pour les fibroses (Liu et al.,

2010), miR-33 pour l'athérosclérose et l'artériosclérose (Najafi-Shoushtari et al., 2010; Rayner et al., 2010) et miR-92a pour les maladies vasculaires (Bonauer et al., 2009). En outre, les sociétés miRNA therapeutics et Regulus possèdent également toutes deux des produits en phase préclinique comme let-7 (cancers) et miR-16 (cancers) pour la première et miR-10b (cancer du sein), miR-21 (gliome et fibrose), miR-103/107 (diabète), miR-182 (métastase du foie), miR-33a/b (athérosclérose) et miR-380-5p (neuroblastome) pour la seconde. Miragen est une quatrième compagnie avec quatre produits en phase préclinique : miR-15a (infarctus du myocarde), miR-195 (infarctus du myocarde), miR-208 (arrêt cardiaque) et miR-451 (hématopoïèse).

En conclusion, même si nous pouvons dire que les miARN ont bel et bien un fort potentiel curatif, cette voie thérapeutique est aujourd'hui restreinte par la délivrance des composés. En effet, le verrou principal de cette technologie reste à ce jour la méthode de délivrance des ARN, permettant non seulement leur transport vers les organes concernés mais également leur stabilisation *in vivo*. La plupart des méthodes de délivrance reposent aujourd'hui sur l'encapsulation des drogues (mimic, LNA, etc.) par des vecteurs de deux types : les vecteurs viraux (lentivirus, adénovirus, etc.) et non viraux (vésicules nanolipidiques, modification chimique des ARN, etc.). Si ces vecteurs permettent d'augmenter considérablement la stabilité des composés, leurs effets secondaires sont non négligeables : activation du système immunitaire, réaction locale lors d'injection intraveineuse, délivrance non systémique mais limitée à certaines organes (le foie en particulier) ou encore baisse du potentiel curatif des composés due à l'encapsulation (Zhang et al., 2013; Chen et al., 2015). Malgré ces défauts, les miARN tendent tout de même à se présenter aujourd'hui comme une troisième classe thérapeutique à mi-chemin entre les deux autres classes : les petites entités chimiques (*new molecular entity* – NME) et les molécules biologiques (*biological entity* – BLE*).*

D. Conclusions

Les miARN jouent un rôle capital dans le devenir des cellules et nous pouvons donc penser qu'en comprenant parfaitement les réseaux de régulation cellulaire, il soit possible *in fine* d'améliorer sensiblement la santé des patients. Même si de nombreux progrès restent encore à faire afin d'arriver à ce niveau de compréhension, chaque pierre apportée à l'édifice permet tout de même de s'en rapprocher. Les travaux de cette thèse suivent cet objectif puisqu'ils ont mené à la découverte de miARN oncogènes potentiels (oncomiR) et miARN suppresseurs de tumeurs potentiels ainsi qu'à l'identification d'une communauté de miARN impliquées dans le maintien de l'état indifférencié des cellules souches.

Ce manuscrit est construit autour de trois chapitres principaux et un chapitre d'ouverture exposant la majeure partie des travaux ayant été réalisés pendant la thèse. Nous commencerons donc dans le chapitre 1 par présenter la méthodologie de construction d'un type de réseaux miARN-miARN, son analyse mathématique et statistique. Nous poursuivrons dans le chapitre 2 par l'étude de groupes de *hubs* composant un des réseaux et la validation de leur implication dans les processus biologiques prédits. Nous poursuivrons dans le chapitre 3 par l'extraction et la caractérisation d'une communauté de miARN potentiellement importante pour le renouvèlement des cellules souches. Enfin, une analyse théorique du réseau de miARN chez l'homme sera proposée dans le chapitre 4, faisant aussi office de chapitre « d'ouverture ».

Matériels et méthodes

La plupart des calculs et des analyses ont été menés dans le logiciel libre de statistiques et représentation graphique R (Ihaka and Gentleman, 1996; R Core Team, 2012). Similaire au langage S – développé au laboratoire Bell par John Chambers et collaborateurs (Becker et al., 1988; Chambers, 1998) – R est un langage très utilisé aujourd'hui dans le domaine de la bioinformatique et des statistiques en général. Cette popularité vient non seulement de sa puissance mais également des nombreux *packages* d'analyse libres pouvant être récupérés facilement au travers de divers répertoires (p.ex. CRAN). Pour les aspects bioinformatiques, c'est principalement au travers du répertoire Bioconductor que les *packages* d'analyse sont déposés et récupérés. BiomaRt, par exemple, est un *package* particulièrement intéressant puisqu'il permet d'obtenir relativement aisément tout un ensemble d'informations biologiques sur les gènes et/ou les miARNs comme leur(s) emplacement(s) génomique(s), leurs identifiants dans différentes bases de données (p.ex. NCBI ou Ensembl) ou encore leur(s) fonction(s) (Durinck et al., 2009).

A. Bases de données de prédiction

Principalement deux bases de données de prédiction ont été utilisées tout au long des travaux de cette thèse : DIANA-microT version 3 (Juillet 2009), téléchargée depuis le site web http://diana.cslab.ece.ntua.gr/microT/ et TargetScan version 6.2 (Juin 2012), téléchargée sur http://www.targetscan.org/vert_61/.

1. DIANA-microT

Seules les prédictions chez l'humain ont été prises en compte, quel que soit leur score miTG (miTG > 0). Selon ces critères, les 555 miARNs de la base de données ciblent un ensemble de 18986 gènes différents (identifiant « Ensembl ») pour un total d'environ 2 millions d'interactions miARN-gène. Dans ce cas, 60% des miARNs possèdent entre 2 000 et 5 000 sites de liaison et les quatre miARNs hsa-miR-495, -548c-3p, -590-3p et -603 montrent une

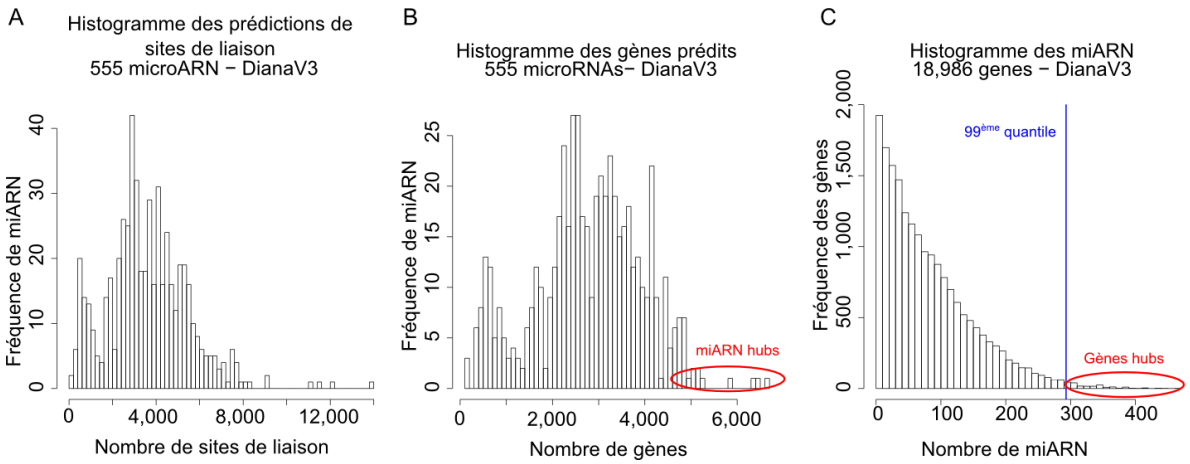


Figure 21. Caractéristiques de la base de données de prédiction DIANA-microT. A | Histogramme des prédictions de sites de liaison (un gène peut avoir plusieurs sites de liaison). B | Histogramme du nombre de gènes cibles prédits pour chaque miARN. C | Histogramme du nombre de miARN ciblant un gène (pour l'ensemble des gènes).

affinité pour plus de 10 000 sites de liaison – ce qui représente environ 6 000 cibles différentes. Enfin, quelques 190 gènes sont ciblés par au moins 293 miARN différents (99ème percentile de la distribution) (Figure 21). Ces gènes, extrêmement ciblés par les miARNs, seront dénommés par la suite gènes (ou protéines) hubs en suivant la formulation de (Shalgi et al., 2007).

2. TargetScan

Dans le cas de TargetScan, c'est toujours l'union des versions conservée et non conservée qui a été considérée. De la même manière qu'avec DIANA-microT, aucun score n'a été pris en compte pour limiter le nombre de prédictions, et ceci afin de garder le plus petit nombre de faux négatifs possible malgré une augmentation probable du nombre de faux

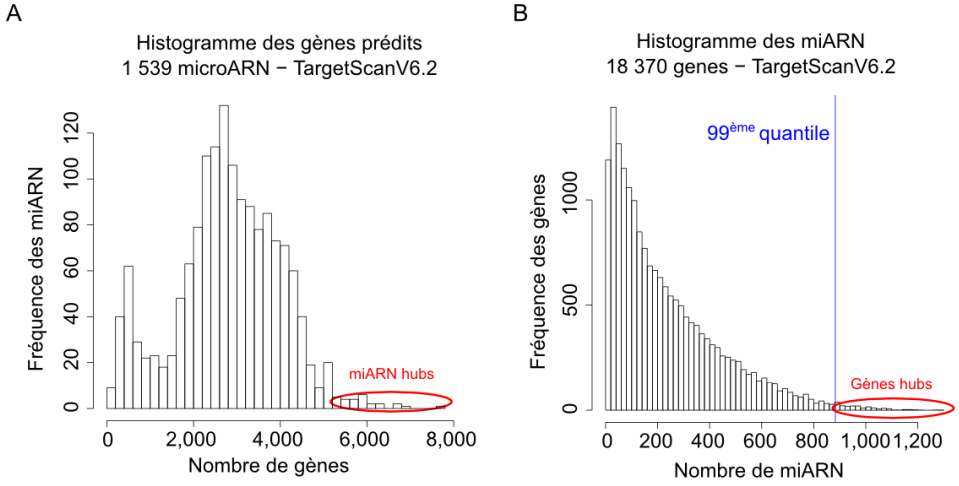


Figure 22. Caractéristiques de la base de données de prédiction TargetScan (humain). A | Histogramme des prédictions de sites de liaison (un gène peut avoir plusieurs sites de liaison). B | Histogramme des prédictions des gènes. C | Histogramme de la régulation des miARNs.

positifs. Chez l'humain, l'union des deux fichiers donne un ensemble de 1 539 miARNs ciblant 18 370 gènes (identifiants « Entrez »). Environ 60% des miARNs ciblent entre 1 500 et 4 500 gènes (Figure 22) et 185 gènes hubs sont prédits comme ciblés par plus de 883 miARNs.

Des prédictions chez d'autres espèces ont également été étudiées indépendamment les unes des autres : notamment le chien, le taureau, le chimpanzé, le rat, la souris, le cheval, le macaque, l'ornithorynque, etc. (Tableau 2). Ces comparaisons seront abordées dans la dernière partie du chapitre 1.

Tableau 2. Caractéristiques générales des prédictions pour différentes espèces avec TargetScan. Min (max) représente le nombre minimum (maximum) de gènes ciblés par les miARN et 25, 50 et 75% représentent respectivement le premier, le second et le troisième quartile du nombre de gènes ciblés par les miARN pour les différentes espèces.

Espèces	# miARN	# gènes	# interactions	Min	25%	50%	75%	Max
Homo sapiens	1 539	18 370	4 305 160	89	2 143	2 796	3 634	7 624
Mus musculus	779	16 961	1 666 429	55	1 660	2 172	2 750	4 839
Bos taurus	634	17 136	1 515 723	141	1 850	2 387	3 030	7 638
Pan troglodites	524	17 888	1 416 354	89	2 094	2 686	3 556	7 545
Macaca mulatta	485	17 532	1 292 877	139	2 106	2 640	3 341	6 228
Galus galus	467	7 480	287 454	31	366	580	818	2 246
Rattus norvegicus	438	16 137	908 602	79	1 638	2 116	2 652	3 925
Equus caballus	360	17 343	870 822	153	1 959	2 426	3 103	5 470
Ornithorhynchus anatinus	306	11 274	317 930	102	752	995	1 309	2 913
Canis familiaris	288	17 202	708 282	245	1 982	2 411	3 088	5 302
Xenopus tropicalis	156	6 167	68 867	30	265	416	612	1 215
Monodelphis domestica	144	12 709	233 428	129	1 271	1 623	2 055	3 820

3. Différences entre DIANA-microT et TargetScan

A cause des différences entre les deux algorithmes, les deux bases prédisent des cibles différentes pour les miARNs. En analysant le recouvrement de prédiction pour chaque miARN présent dans les deux bases de données par l'indice *meet/min* ($meet/min_{miARNx} = A \cap B / \min(\#A, \#B)$, avec A et B représentant les cibles prédites d'un même miARN dans chacune des bases), nous pouvons constater que ce dernier n'est environ que de 60% (Figure 23). La distribution du recouvrement sur la Figure 23 est d'ailleurs bimodale : une première partie centrée aux alentours de 55% et une autre à 76%. Ces deux parties montrent que pour certains miARNs, les prédictions sont très proches entre les deux algorithmes alors qu'elles diffèrent bien plus pour d'autres. Un des miARNs dont les cibles sont particulièrement recouvertes est hsa-miR-513a-3p avec une valeur de 84%. Aucun miARN ne montre un recouvrement de 100%.

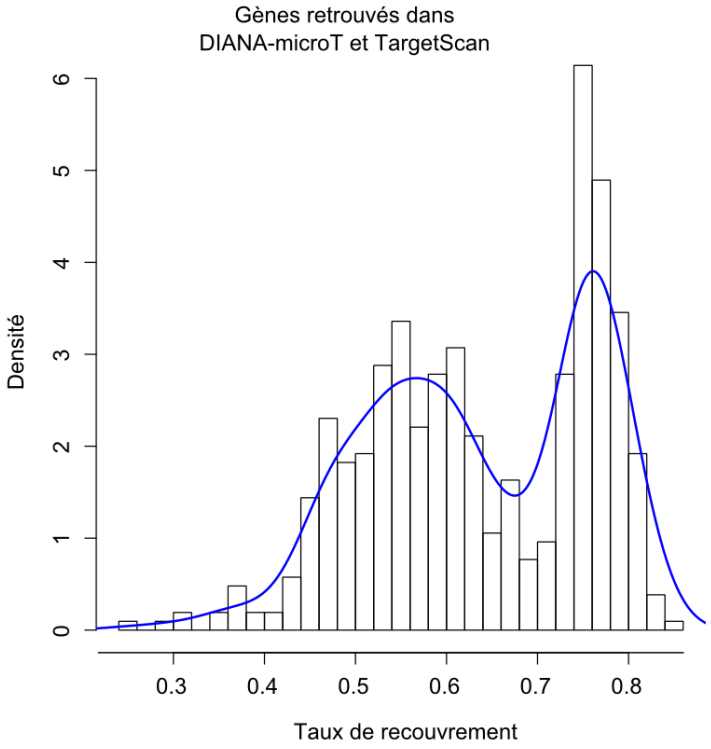


Figure 23. Recouvrement entre les prédictions de TargetScan v6.2 et DIANA-microT v3. Pour chaque miARN présent à la fois dans DIANA-microT et TargetScan, le recouvrement de cibles est calculé par l'indice meet/min.

B. Construction de réseaux

La plupart des travaux présentés dans cette thèse se basent sur l'algorithme DIANA-microT sauf lorsque précisé autrement. Chaque réseau de miARN a été construit de façon indépendante en fonction de la base de prédiction et de l'espèce mais toujours selon la même méthodologie (Figure 24). Cette méthode se base sur l'hypothèse qu'en partageant des cibles communes, les miARNs seraient capables de coréguler les mêmes processus biologiques. Dans notre cas, un réseau est ainsi représenté par un ensemble de miARNs pour les nœuds et les liens représentent le pourcentage de cibles partagées entre deux miARN. Il existe différents indices pour définir le partage de cibles (Goldberg and Roth, 2003), celle utilisée dans ces travaux est l'indice de Simpson (ou indice *meet/min*) :

$$meet/min_{(A,B)} = \frac{A \cap B}{\min(\#A, \#B)}$$

où A ∩ B est le nombre de gènes en commun régulés par les miARN A et B et #A et #B représente le nombre de cibles régulées par les miARN A et B respectivement.

En se basant sur cet indice, les réseaux construits sont des graphes complets et pondérés : tous les miARNs sont interconnectés deux à deux, avec un certain pourcentage compris entre 0 (ne partagent aucune cible) et 100 (partagent exactement les mêmes cibles). Ces réseaux étant des réseaux extrêmement denses, leur analyse s'en retrouve grandement complexifiée autant d'un point de vue computationnel que visuel. Afin de pouvoir analyser plus aisément ces réseaux, un seuil meet/min a été appliqué sur les liens : un lien n'est gardé que si le pourcentage meet/min dépasse ce seuil. Après cette transformation, on se retrouve avec un réseau binaire : un lien est présent entre deux miARNs uniquement si ces derniers partagent « assez » de cibles.

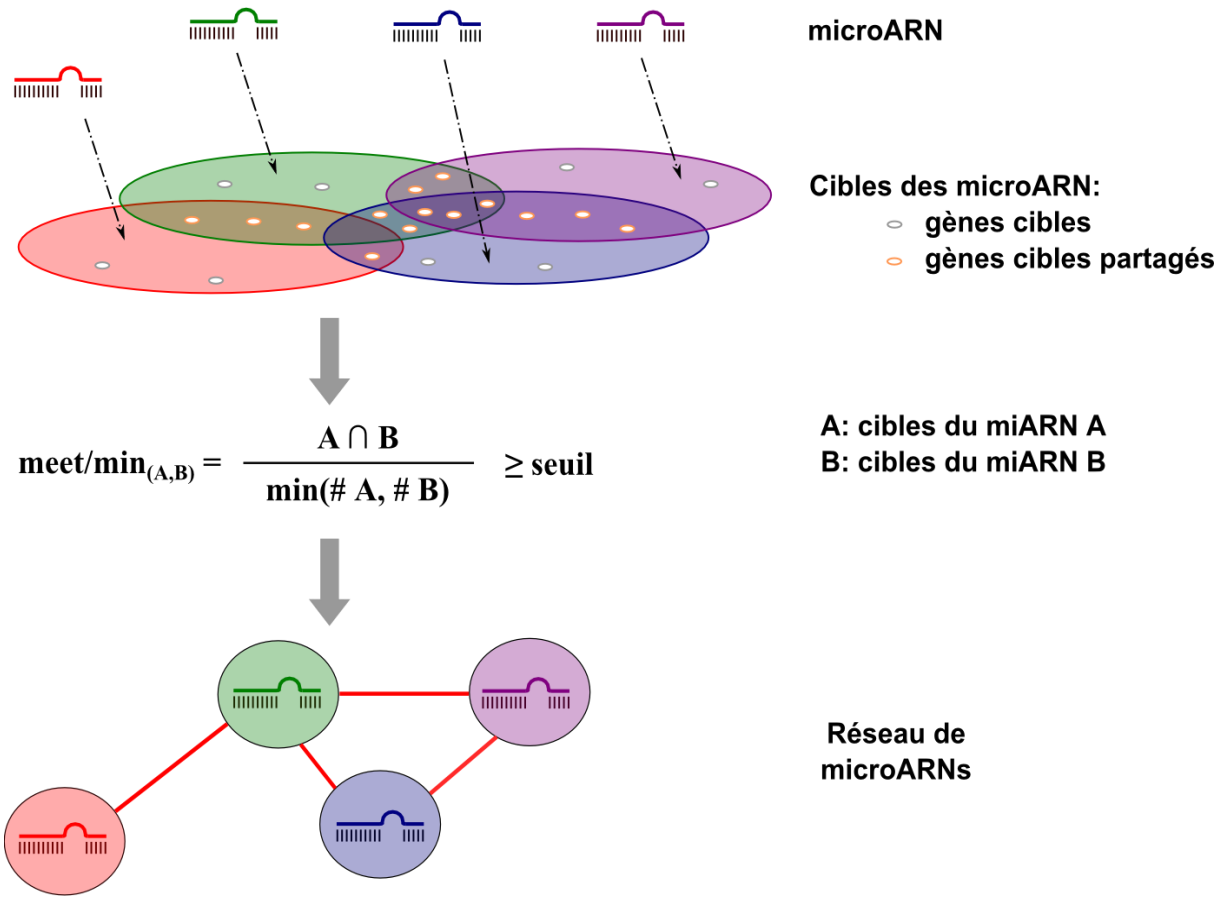


Figure 24. Construction de réseaux de miARNs basé sur le partage de cibles. Dans un premier temps, une analyse du recouvrement de cibles entre toutes les paires de miARNs est faite puis un seuil est mis en place sur ces recouvrements afin de construire le réseau binaire de miARN.

En pratique, pour un algorithme et une espèce donnée, plusieurs réseaux ont été construits en fonction de différents seuils *meet/min* (typiquement de 0 à 100 par incrément de 1, soit 101 réseaux). Pour chacun de ces 101 réseaux, les propriétés évoquées dans l'introduction ont été calculées (densité, coefficient de clustering, chemin moyen, centralité, etc.). Ces propriétés ont alors été comparées entre les différents réseaux pour déterminer un seuil optimal : c'est ce que l'on appelle communément une approche par seuil multiple (*multiple-threshold-approach*) – couramment utilisée dans l'étude des réseaux neuronaux (van Wijk et al., 2010; Langer et al., 2013). Le seuil meet/min a été défini =automatiquement comme le niveau pour lequel était observé un maximum de nœuds connectés au réseau (minimum de nœuds isolés) pour un minimum de liens. Ces deux critères induisent une densité de réseau faible (réseau « *sparse* »). La construction et l'analyse des réseaux forment l'ensemble du chapitre 1 de ce document. Des comparaisons de réseaux de miARNs à

différents seuils avec un réseau aléatoire de 555 nœuds et 2911 liens (Erdős and Rényi, 1959), un réseau sans échelle de 555 nœuds (Barabási and Albert, 1999) et un réseau d'interaction protéine-protéine tiré d'expériences de double hybride chez la levure (Y2H) (Bu et al., 2003) ont également été menées. Pour les réseaux aléatoires, la construction se fait itérativement à partir d'un nombre de nœuds et d'un nombre de liens donnés. Chaque lien a la même probabilité d'apparaitre pendant la construction. Pour les réseaux sans échelle, seul le nombre de nœuds est nécessaire. Les liens sont ajoutés de telle sorte que le réseau possède *in fine* les propriétés classiques des réseaux sans échelle.

Ces différentes étapes ont été menées au travers du package « igraph » de R (Csardi and Nepusz, 2006). La visualisation a été faite essentiellement à l'aide du logiciel libre Cytoscape (Shannon et al., 2003) et de l'algorithme « *Unweighted Spring Embedded* ». Cet algorithme considère les nœuds comme des boules et les liens entre les nœuds, comme des ressorts. Dans la représentation graphique, les communautés très connectées auront donc tendance à se regrouper sous la « pression élastique » alors que les nœuds peu connectés seront éloignés les uns des autres.

C. Détection de communautés

La découverte des « clubs assortis », dont l'analyse sera évoquée dans le chapitre 2, est basée sur le principe des « rich clubs » (Colizza et al., 2006; Boulet et al., 2011). Ces clubs riches sont en fait ces communautés particulières formées de hubs très interconnectés. Ils ont généralement une plus grande importance puisqu'ils ont tendance à dominer et influencer le réseau. Dans le cadre de ces travaux, nous avons nommé ces communautés « clubs assortis », à cause de la tendance assortative de ces sous-graphes (c'est-à-dire, la tendance qu'ont les hubs à également se lier à d'autres hubs). Le processus de découverte de ces clubs repose sur le classement des nœuds en fonction de leur degré et l'analyse des propriétés des sous-réseaux formés par les *n* premiers hubs rangés.

Pour la détection de communautés (Fortunato, 2010) à proprement parler, trois algorithmes ont été utilisés : *edge betweenness* (Newman and Girvan, 2004), *fast greedy* (Clauset et al., 2004) et *multilevel* (Blondel et al., 2008), l'ensemble de ces algorithmes étant implémenté dans le package R igraph (Csardi and Nepusz, 2006).

L'algorithme *multilevel* est similaire aux algorithmes de classification hiérarchique ascendante. C'est un processus itératif où chaque nœud est d'abord considéré comme une communauté isolée. A chaque étape, les nœuds sont fusionnés au sein de la communauté pour laquelle la fusion maximise la modularité (cf. page 52). Lorsqu'aucun nœud ne peut être réassigné, chaque communauté est considérée comme un unique nœud et le processus se poursuit jusqu'à ce que chaque nœud du réseau forme une communauté unique ou alors lorsque la modularité ne peut plus être augmentée (Figure 25).

L'algorithme *fast greedy* repose sur un principe similaire à *multilevel* à la différence près que l'algorithme *multilevel* cherche à équilibrer la taille des communautés à fusionner. Cette différence, pouvant sembler mineure *a priori*, peut entrainer de gros changements dans les communautés retrouvées.

Enfin, l'algorithme *edge betweenness* se base également sur une classification hiérarchique. Dans ce cas en revanche, la classification est descendante et l'algorithme se base sur une analyse des liens et non pas des nœuds. Pour cette méthodologie, les liens du réseau sont retirés dans l'ordre décroissant de leur score de centralité betweenness. De façon plus précise, le score de centralité de chaque lien est premièrement calculé dans le réseau. Le lien avec le plus fort score est alors retiré et les scores de centralité sont recalculés afin que le cycle puisse se poursuivre. On obtient en fin de compte un dendrogramme complet du graphe où les feuilles représentent les nœuds individuels et la racine, le graphe complet. Afin de déterminer le seuil où couper dans le dendrogramme pour définir les communautés, l'algorithme se base également sur la modularité, calculée à chaque niveau du dendrogramme (Figure 25).

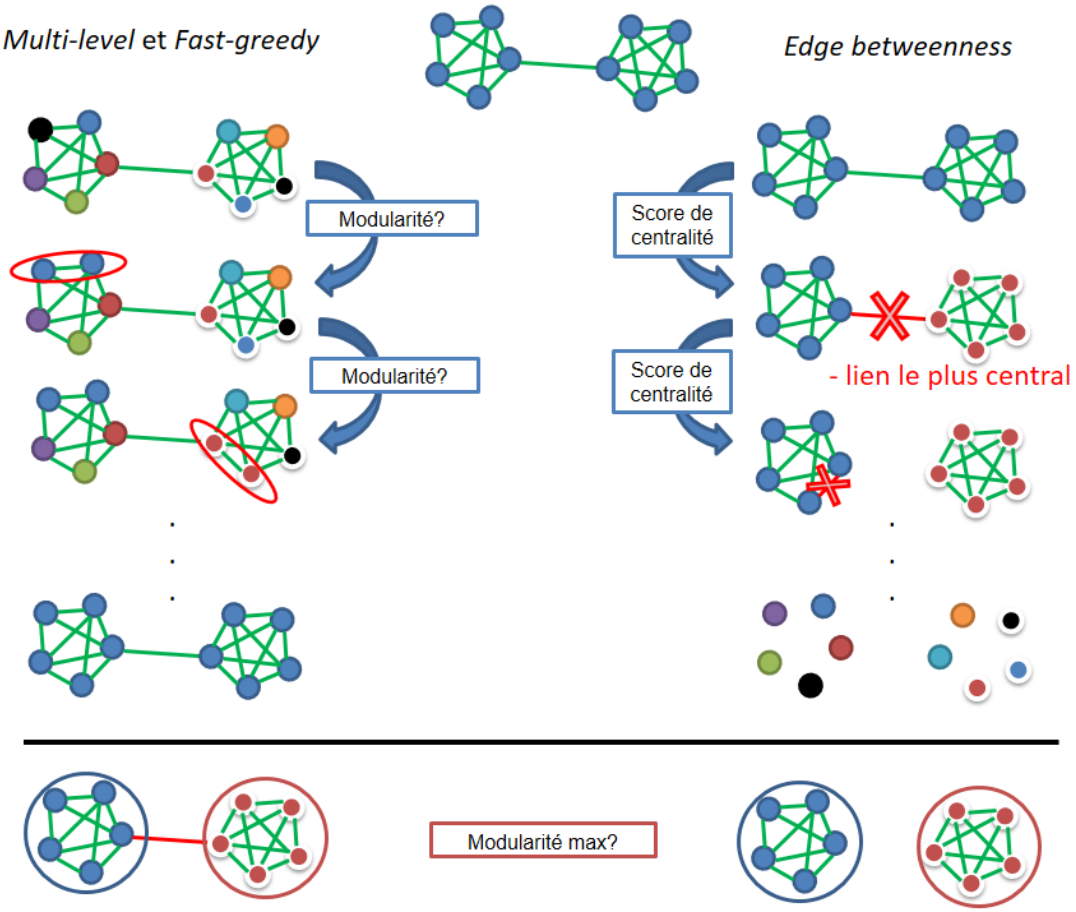


Figure 25. Exemple des algorithmes de recherche de communautés. A gauche, un algorithme de type classification ascendante où chaque nœud est initialement une communauté à part (*multilevel* et *fast greedy*) et l'algorithme cherche à regrouper des nœuds ensemble en maximisant la modularité. A droite, un algorithme de recherche de communauté de type classification descendante où les liens sont retirés de manière itérative jusqu'à ce qu'il n'y ait plus de liens (*edge betweenness*). A chaque étape les scores de centralité sont recalculés.

Ces différents algorithmes ont été utilisés principalement dans le cadre du chapitre 3 pour découvrir une communauté de miARNs impliqués dans le maintien de la totipotence des cellules souches embryonnaires. Pour identifier cette communauté, les trois algorithmes ont en l'occurrence été utilisés sur les réseaux binaires de TargetScan et DIANA-microT indépendamment. Ceci a mené à la création de six partitions différentes (trois pour chaque réseau) composés de différents clusters.

D. Enrichissement d'ontologie

1. Gene Ontology

L'ontologie des gènes (*Gene Ontology* – GO) est une forme d'annotation unifiée et contrôlée de la fonction et la localisation des gènes à toutes les espèces et dont le but est la structuration de la description des gènes et des protéines (Ashburner et al., 2000). Elle est basée sur trois niveaux : les processus biologiques (BP), les fonctions moléculaires (MF) et les compartiments cellulaires (CC). Ces trois niveaux sont en fait les racines de trois arbres, c'est à dire trois graphes acycliques dirigés, où chaque nœud est un terme d'ontologie (p.ex. « régulation des petites GTPase » ou « régulation de la transcription »). Dans ces graphes, chaque annotation peut avoir un ou plusieurs fils et un ou plusieurs parents (Figure 26). Le parent correspond souvent à une définition ontologique plus générique que celle du fils (p.ex. « régulation de la transcription » contre « régulation négative de la transcription »). Les liens entre parents et fils sont de plusieurs natures : « *is_a* », « *part_of_a* », « *regulates* », « *negatively regulates* » et « *positively regulates* ».

Plusieurs gènes pouvant être associés à un même terme, il est courant de chercher les termes statistiquement surreprésentés dans un ensemble de gènes donnés. On parle alors

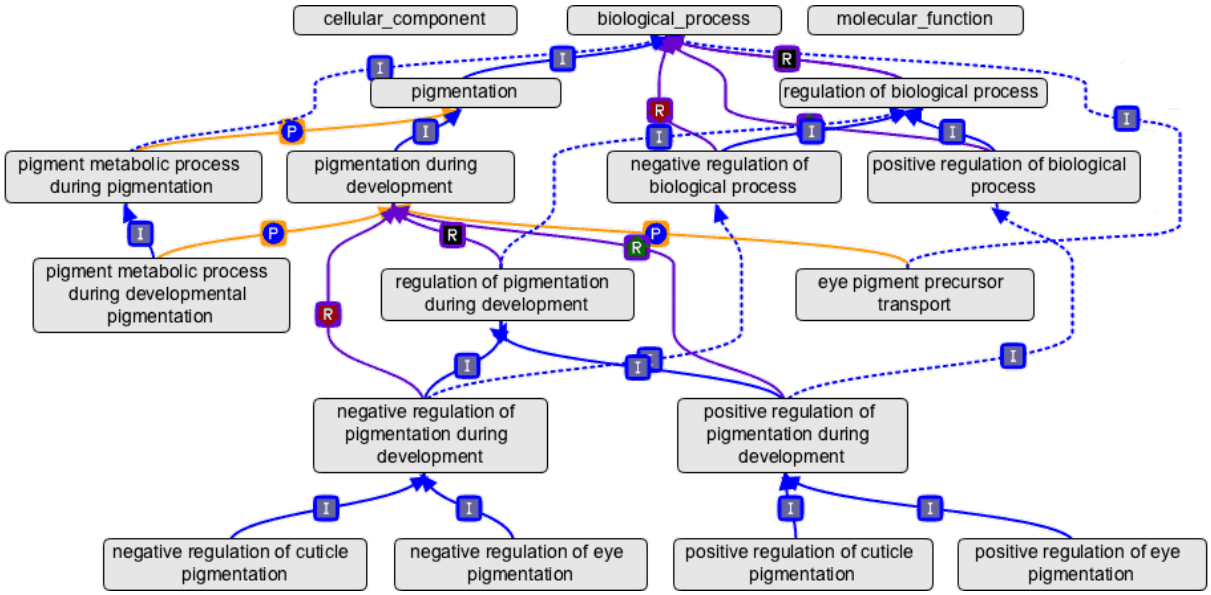


Figure 26. Structure d'une partie de l'arbre « processus biologique » de GO. Source : http://geneontology.org/page/ontology-structure. I : is_a ; R : part_of_a.

de l'analyse d'enrichissement d'ontologie des gènes (*gene ontology enrichment analysis*). Dans notre cas, chaque miARN peut cibler un ensemble de gènes, dont une partie peut intervenir dans des processus identiques. En analysant par exemple l'enrichissement d'ontologie BP de ces gènes, on cherche indirectement à connaitre les processus biologiques régulés – avec la plus grande probabilité – par les miARNs (ou les processus corégulés, si l'on considère les gènes ciblés par plusieurs miARNs comme dans ces travaux).

2. Analyse d'enrichissement d'ontologie

L'approche classique pour ce type d'analyse est le test exact de Fisher, un test basé sur des tables de contingence (matrice 2 x 2) et qui permet le calcul exact de probabilités. Soit la matrice de contingence suivante :

	Catégorie 1 (# gènes cibles par différents miARN)	**| Catégorie 1** (# gènes non ciblés par différents miARN)	**Total**
Catégorie 2 (# gènes associés à un terme GO)	a	b	**a + b**
| Catégorie 2 (# gènes non associés à un terme GO)	c	d	**c + d**
Total	**a + c**	**b + d**	**N (a + b + c + d)**

Dans un test exact de Fisher, la probabilité associée à cette table est calculée de la manière suivante :

$$p(a) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}$$

où $\binom{i}{j}$ est le coefficient binomial et « | » désigne la négation d'ensemble (*i.e.* ne faisant pas partie de).

La p-valeur est ensuite calculée comme une somme de probabilités :

$$\mathrm{p-valeur} = \sum_{x \geq a} p(x)$$

Une p-valeur inférieure à un seuil α donné induit le rejet de l'hypothèse nulle de l'indépendance entre les deux catégories. Dans notre cas, nous dirons que les communautés de miARN sont enrichies (au travers des gènes cibles) pour un (des) terme(s) donné(s). Les gènes considérés sont les gènes partagés par différents miARNs d'une même communauté : dans le cas des clubs assortis et des communautés, c'est principalement les gènes partagés par au moins 50% des membres des clubs qui ont été considérés alors que pour des zones plus vastes du réseau, ce sont les gènes partagés par au moins 25% afin d'éviter de n'avoir que des gènes hubs. Dans la plupart des cas cependant, des gammes complètes de pourcentages de gènes partagés ont été analysées.

Les enrichissements ont été calculés pour l'ensemble des termes ontologiques de la base de données GO (BP, MF, CC). Pour calculer ces enrichissements, le package R TopGO (Alexa et al., 2006) a été utilisé, principalement pour sa rapidité mais également à cause de son module permettant de ne pas considérer les termes trop génériques (option « elim »). Les termes très génériques (associés à plus de 5000 gènes) et trop spécifiques (associés à moins de 10 gènes) n'ont de toute manière pas été considérés. Si les termes trop génériques ne devraient en théorie pas apparaitre dans les analyses d'ontologie, certaines erreurs d'annotations dans les arbres GO (notamment dus aux processus d'annotation automatisés ou la redondance des identifiants géniques) peuvent tout de même les faire ressortir significativement.

Une information capitale lors de ce type d'analyse est le fond (*background*) de gènes utilisé (N du tableau ci-dessus) : dans notre cas, ce *background* dépendait de l'algorithme et il correspondait presque toujours à l'ensemble des gènes prédits par l'algorithme en question (sauf lorsque précisé autrement) (*i.e.* les 18986 gènes pour DIANA-microT et les 18370 gènes pour TargetScan chez l'être humain). Dans quelques cas toutefois, ce n'est pas l'ensemble des gènes qui a été pris en compte mais uniquement une portion de ces gènes (p.ex. uniquement les gènes réprimés dans un type cellulaire par rapport à un autre – toujours pour un algorithme donné). La définition de gènes réprimés (ou induits) dans un type cellulaire sera

abordée un peu plus loin dans cette partie et dans le chapitre 3 pour les enrichissements des gènes réprimés dans les cellules souches embryonnaires et partagés par des miARNs.

Dans les tableaux d'enrichissement présentés dans ce document, les termes « annotés, retrouvés, classicFisher et BH.pVal » font respectivement référence au nombre de gènes annotés pour le terme GO en question (total), le nombre de gènes testés et annotés pour le terme GO, à la p-valeur classique du test de Fisher et à la p-valeur corrigée par la procédure de Benjamini et Hochberg (voir paragraphe suivant).

3. Problème des tests multiples

Dans ce type d'étude d'enrichissement et dans bien d'autres cas, des milliers d'hypothèses sont testées en parallèle. Ce genre de comparaisons multiples entraine un biais dans les statistiques puisque le risque de se tromper augmente significativement avec le nombre de tests (Figure 24 A) – on parle d'augmentation du risque de première espèce (erreur de type I) ou du taux de faux positifs. Ainsi, lorsque l'on effectue un test statistique à un seuil α de 5%, le risque de rejeter à tort l'hypothèse nulle est de 5% ; mais lorsque l'on effectue 10 tests indépendants, le risque de se tromper au moins une fois sur l'ensemble des tests est non

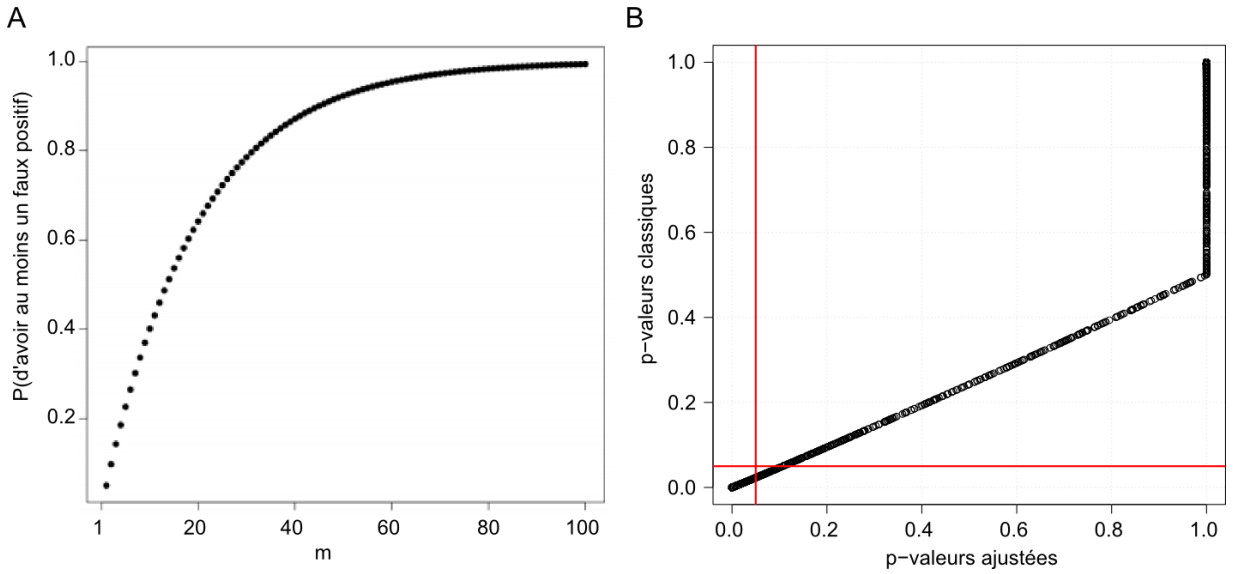


Figure 27. Problème des tests multiples et correction. A | Probabilité d'obtenir un faux positif en fonction du nombre de tests. Plus le nombre de tests augmente, plus la probabilité de se tromper augmente. B | Différence entre p-valeurs et p-valeurs corrigées par la méthode de Benjamini et Hochberg. Exemple trivial sur les p-valeurs associées à une distribution normale $\boldsymbol{N(0,20)}$ de 10 000 valeurs. Les deux barres rouges représentent des seuils arbitraires à 5%.

plus de 5% mais de 40% ! Avec une centaine de tests, on se retrouve avec une probabilité de se tromper de quasiment 100% sur au moins un des tests effectués (Figure 27 A) :

$$P(faire\ au\ moins\ une\ erreur\ dans\ m\ tests) = 1 - (1 - \alpha)^m$$

De nombreuses méthodes ont vu le jour pour pallier ce problème, notamment les méthodes de Benjamini et Hochberg (Benjamini and Hochberg, 1995) ou de Bonferroni (Bonferroni, 1936). Pour ces travaux, seule la correction de Benjamini et Hochberg a été utilisée, parce que la méthode est moins drastique que d'autres corrections - notamment celle de Bonferroni, qui consiste à diviser les p-valeurs par le nombre d'hypothèses testées.

Considérons un ensemble $H_1, \ldots . H_m$ d'hypothèses nulles et leur probabilité respective $P_{(1)}, etc. P_{(m)}$. La procédure de Benjamini et Hochberg cherche tout d'abord à ordonner les p-valeurs dans l'ordre croissant puis, pour un seuil α donné, cherche le plus grand k tel que $P_{(k)} \leq \frac{k}{m}\alpha$. La procédure rejette alors toutes les hypothèses $H_{(i)}$ avec $i = 1, \ldots, k$. Un exemple de la transformation pour des p-valeurs associées à une distribution normale est donné dans la Figure 27 B, où l'on peut constater que la méthode restreint bien le nombre d'hypothèses acceptées à un seuil α de 5%.

Toutefois, cette correction n'est valable que lorsque les m hypothèses sont indépendantes. Ce dernier constat pose un problème dans le cas de l'analyse d'enrichissement ontologique puisque les termes des arbres ne sont pas indépendants. Les différentes corrections sont donc inadéquates dans le cas d'analyse d'enrichissement GO mais restent particulièrement utilisées dans le domaine – bien que le problème soit reconnu par différents auteurs (Blüthgen et al., 2005; Huang et al., 2009) – raison pour laquelle nous avons tout de même calculé et gardé l'information corrigée.

Une autre particularité de la correction pour tests multiples est le concept de la recherche exploratoire contre la recherche confirmatoire (Goeman and Solari, 2011). La distinction entre ces deux domaines est précisément le taux de fausses découvertes. Lors d'une expérience en recherche confirmatoire, l'objectif est d'avoir le plus faible taux de faux

positifs possible. Par exemple, un objectif serait de diagnostiquer le cancer de la prostate et d'éviter à tout prix les faux diagnostiques (test positif à la maladie mais absence de maladie), très dommageables pour l'individu (particulièrement psychologiquement), en plus d'être parfaitement inutile. *A contrario*, la recherche exploratoire n'a pas ce genre de contrainte puisque l'objectif est de mettre en évidence de nouvelles hypothèses à tester et à confirmer par la suite. Ce type de recherche peut donc se permettre plus de souplesse sur le risque d'erreur de type I. Dans notre situation, nous sommes typiquement dans le cas de la recherche exploratoire : nous cherchons à trouver de nouvelles hypothèses à tester et confirmer (ou infirmer) pour les miARNs. Il s'est donc agi de trouver un compromis entre stringence des méthodes de corrections multiples et souplesse de l'analyse vis-à-vis de la significativité statistique des conclusions. Durant ces travaux, un seuil α de 5% a donc été considéré sur les p-valeurs corrigées pour définir le seuil de significativité des enrichissements, même si ce seuil n'a pas été considéré comme une limite stricte.

E. Validation de l'implication des miR-661, -612 et -940 dans la voie des petites GTPases.

La découverte de ces trois miARNs formant un des deux clubs assortis découverts sur DIANA-microT et la prédiction du (des) processus biologique(s) qu'ils co-régulent forment le chapitre 2 de ce document. Seront exposés ici les protocoles des différentes expériences pour valider les prédictions, notamment le *western blotting*, l'immunofluorescence, les tests « *transwells* » et les tests de blessures. L'ensemble de ces expériences ont été réalisées par des collègues expérimentateurs au laboratoire et je me suis chargé d'analyser les données, à l'exception des analyses d'images de blessures.

1. Culture cellulaire et transfection

L'ensemble des validations a été faite sur des cellules épithéliales de rétine immortalisées (*Human telomerase-immortalised retinal pigmented epithelial cells* – hTERT-RPE1 ou RPE1 plus simplement). Ces cellules ont été mises en culture à 37°C et 5% de CO_2

dans du milieu de culture (DMEM/F12, Invitrogen, Carlsbad, Californie) auquel a été ajouté 10% de sérum bovin, 2 mM de glutamine, 100 U/ml de streptomycine. La transfection a été faite en utilisant la lipofectamine RNAiMax (Invitrogen, Carlsbad, Californie) pendant 48 heures. La concentration finale des mimics de miARN et des siARNs était de 20 nM. Les miARNs mimics ont été achetés chez Thermo Scientific (Waltham, Massachusetts) et Dharmacon (Lafayette, Colorado). Le siARN AllStars (Qiagen, Venio, Hollande), conçu pour ne cibler aucun gène, a été utilisé comme contrôle négatif et l'inhibiteur de ROCK, Y27632 (refY0503 : Sigma-Aldrich, St. Louis, Missouri) comme contrôle positif. Ce dernier a été ajouté à 10 µM pendant les 24 dernières heures.

2. Lyse des cellules, extraction protéique et western blot.

Le lysat protéique a été préparé dans du tampon réfrigéré RIPA (Thermo Scientific) avec un cocktail d'inhibiteurs de protéases (*complete mini* ; Roche, Bâle, Suisse), 1mM de PMSF, 2 mM Na_3VO_4 et de glycérophosphate. Les homogénats ont été séparés par centrifugation à 15 000 g pendant 15 min à 4°C. Les protéines (10 µg/colonne) ont été déposées sur un gel de polyacrylamide-SDS, puis transférées sur des membranes de nitrocellulose et bloqués avec 3% de BSA (dans du TBST) pour 1 heure puis incubées toute la nuit à 4°C avec des anticorps primaires polyclonaux de lapin contre la chaine légère de la phosphomyosine 2 (1 :1000, Cell Signaling Technology, Denver, Massachusetts) dans 3% de BSA. La visualisation a été faite par des anticorps conjugués-PRH (peroxydase de raifort, anti-lapin ; Santa Cruz Biotechnology, Dallas, Texas). Pour le contrôle de chargement, des anticorps polyclonaux GAPDH de lapin ont été utilisés.

3. Création des lamelles de micropatron

Des lamelles de verre ont été enduites par centrifugation à 3 000 tours/min pendant 30 s avec un promoteur d'adhésion (TI Prime ; MicroChemicals, Madhya Pradesh, Inde) puis avec 0.5% de polystyrène dilué dans du toluène. La couche de polystyrène a également été traitée au plasma d'oxygène (FEMTO ; Diener Electronics, Allemagne) pour 10 s à 30 W et incubée

avec 0.1 mg/ml de polylysine polyéthylène-glycol (JenKem Technology, Beijing, China) dans 10 nM d'Hepès à pH 7,4 et à température ambiante pendant 1 heure. Les lamelles ont ensuite été séchées. Les lamelles recouvertes de polyéthylène-glycol ont alors été placées en contact avec un masque optique portant le micropatron transparent (Topan Photomask, Round Rock, Texas) en utilisant une chambre sous vide puis exposées aux UV pendant 5 min (UVO Cleaner ; Jelight Company, Irvine, Californie). Les lames de micropatron ont ensuite été rincées une première fois dans du PBS et incubées pendant 30 min dans une solution de 50 µg/ml de solution de fibronectine bovine (Sigma-Aldrich, St. Louis, Missouri) et 5 µg/ml de fibrinogène Alexa Fluor 646 – ou d'Alexa Fluor 542 (Invitrogen, Carlsbad, Californie). Avant de déposer les cellules sur les micropatrons, ces derniers ont été lavés trois fois avec du PBS stérilisé. Les petits micropatrons de fibronectine utilisés avaient une taille de 500 µm², alors que les larges avaient une taille de 1000 µm².

L'objectif des analyses sur patron est de réduire la variabilité morphologique des cellules (taille, forme, etc.). Comme ces dernières ne se fixent qu'au niveau du patron et que ce dernier contraint la forme cellulaire (des cercles dans ce cas, mais il existe également des patrons avec des formes différentes), l'analyse et la comparaison cellule-à-cellule est grandement facilité.

4. Marquage par immunofluorescence

De la même manière qu'exposée précédemment, les cellules RPE1 ont été transfectées par le siARN AllStars ou les mimics de miR-612, -661 et -940 à 20 nM pendant 48 heures. Le contrôle positif (Y27632) a été ajouté aux cellules transfectées par le siARN AllStars à 10 µM pour les dernières 24 heures. Les cellules ont ensuite été déposées sur les micropatrons. Après adhésion sur ces derniers (2 heures environ), les cellules ont été pré-perméabilisées pendant 15 s avec 0.1% de Triton X-100 dans du tampon de cytosquelette à pH 6.1 et fixées dans 4% de paraformaldéhyde pendant 15 min à température ambiante. Elles ont ensuite été rincées deux fois avec du PBS et incubées dans 0.1 M de chlorure d'ammonium

pendant 10 min. Les cellules ont ensuite été bloquées avec 3% de BSA dans du $PBS^{Ca2+, Mg2+}$ pendant 30 min. Les images ont été prises par un microscope droit (BX61 ; Olympus, Tokyo, Japon). Les images des marquages de myosine et d'actine présentées sont des projections maximales de différentes cellules alignées (grâce aux patrons) acquises par un objectif au 100X. Les images ont été traitées et analysées par le logiciel ImageJ (Abramoff et al., 2004). Afin de mesurer l'intensité de la fluorescence d'actine et de myosine, les cellules ont d'abord été segmentées et la densité intégrée de la fluorescence a été calculée uniquement au sein de ces segmentations. Pour cette dernière, un filtre médian de 15 pixels de rayon a d'abord été appliqué sur le canal FITC (marquage phalloïdine) puis une méthode d'auto-seuillage par la méthode « Li » a été utilisée (Schneider et al., 2012).

Pour la culture classique, les cellules ont été transfectées dans des plaques 8 puits avec les mimics et le siARN AllStars. Des marquages pour visualiser la vinculine et l'actine ont été utilisés. Dans ce cas, les images ont été prises à l'AxioImager® (ZEISS, Oberkochen, Allemagne).

5. Expérience *transwell*

Pour cette expérience, les cellules RPE1 ont été déposées dans des plaques de microtitration 6 puits, cultivées pendant un jour et transfectées par les mimics des trois miARNs ou du contrôle négatif siARN AllStars à une concentration finale de 20 nM. Les cellules ont été trypsinées 24h après la transfection et re-suspendues dans un milieu de culture sans sérum bovin (0% FBS) et comptées par Scepter™ 2.0 (Merck, Millipore, Billerica, Massachusetts). Un nombre similaire de cellules a ensuite été déposé sur les inserts Transwell® (Corning, membrane en polycarbonate, taille des pores : 5.0 µm). La migration cellulaire au travers des pores a été induite par la présence d'un milieu de culture complet avec sérum (10% FBS) dans le compartiment inférieur des puits. Après dix-huit heures, la migration cellulaire a été stoppée puis les cellules ont été rincées ($PBS^{Ca2+, Mg2+}$), fixées (PFA 4%) et perméabilisées (100% de méthanol). Pour le comptage des cellules, les noyaux ont été marqués à l'Hoechst 33342. Les

cellules situées sur le côté du dépôt (coté supérieur des inserts) ont été retirées avec un coton-tige. Des images du coté inférieur ont ensuite été prises (10 champs par insert) avec un microscope à épifluorescence (Imager Z1 de ZEISS, Oberkochen, Allemagne) en utilisant le logiciel AxioVision avec un objectif 10X (Plan-Neofluar 10X/0.30). La quantification du nombre de cellules ayant traversées vers le coté inférieur a été faite semi-automatiquement en utilisant le logiciel ImageJ ou manuellement.

Quatre expériences indépendantes avec un nombre variable de cellules déposées (afin de mieux appréhender l'augmentation ou la baisse du nombre de cellules passant par les pores) ont été conduites. Chaque condition (miR-612, -661, -940 et siARN AllStars) a été reproduite trois fois pour un total de trente observations par expérience et par condition. Pour fusionner l'ensemble des expériences, chaque condition a été normalisée par le nombre médian de cellules dans la condition avec le siARN AllStars ($\text{mediane}_{\text{SiRNA-AllStarsx}}$) :

$$\widetilde{N}_{x,r,\tau} = log10\ (\frac{N_{x,r,\tau}}{mediane_{siRNA-AllStars_x}})$$

où N est le nombre de cellules comptées et $\widetilde{N}$, le nombre de cellules normalisé pour chaque condition, réplique et expérience. x est l'expérience (1 à 4), r, la réplique pour chaque expérience (1 à 3) et τ, les différentes conditions (siARN AllStars ou les mimics). $\widetilde{N}$ est donc normalisé à 0 pour le siARN AllStars. Les p-valeurs ont été calculées par le test non paramétrique de Mann-Whitney.

6. Test de blessure

Les cellules RPE1 ont été déposées sur des plaques de microtitration transparentes de 96 puits avec du milieu DMEM en présence de 10% de sérum fœtal de veau et sans antibiotique. Après 24 heures, les cellules ont été transfectées en triplicat toujours dans les mêmes conditions qu'évoquées précédemment. Après 48 heures, de larges blessures de 500 µm ont été faites sur chaque tapis cellulaire en utilisant un peigne de 96 pointes (« *wound replicator* », V&P Scientific, San Diego, Californie). Les blessures ont été immédiatement

imagées à 5 heures, 7,5 heures et 10 heures après la blessure. Pour laisser le temps aux cellules de s'adapter aux stresses mécaniques et thermiques induits par la blessure, aucune image n'a été prise à 2,5 heures. A chaque temps, les images ont été acquises à huit temps d'exposition en utilisant un microscope sans lentille parallélisé (96 senseur d'image CMOS – STMicroelectronics, Grenoble, France) placé sous la plaque de microtitration (Ghorbel et al., 2014a). L'imagerie holographique repose sur l'enregistrement digital de figures de diffractions produites par les cellules sous illumination cohérente. Les huit images prises à différents temps d'exposition ont été combinées pour produire une image contrastée et non saturée en utilisant une approche *High Dynamic Range* (HDR). Les bords des blessures ont été automatiquement détectés sur ces dernières en utilisant un procédé appelé *k-means/Markov random field* et un *parallel double snake* (Ghorbel et al., 2014a, 2014b). Les résultats ont par la suite été validés à l'œil. Enfin, l'évolution de la taille moyenne des blessures a été normalisée par rapport à l'état initial de la taille de la blessure et mise en commun pour chaque condition de transfection.

F. Données d'expression

Les données d'expression utilisées dans cette étude sont intégralement issues de la base de données GEO (*Gene Expression Omnibus*) (Edgar, 2002). Cette base de données réunit tout un ensemble de données d'expression, quelle que soit la méthode expérimentale. Ces informations se présentent sous la forme de tableaux à deux dimensions avec les lignes représentant les sondes et les colonnes représentant différentes informations comme le nom des gènes, la valeur d'expression, l'emplacement génomique, etc.

1. Ensembles de données

Dans le cadre du chapitre 1, sept séries de données sur l'expression de miARNs ont été téléchargées depuis GEO : GSE19505 (cortex préfrontal et foie), GSE23527 (muscle squelettique), GSE24205 (sang), GSE25508 (poumons), GSE31309 (sein), GSE34933 (prostate) et GSE38389 (muqueuse colorectale). Ce sont essentiellement des tissus sains qui ont été analysés pour cet ensemble. Trois autres ensembles de données sur l'expression des

miARNs dans le sein ont été utilisés pour le chapitre 2 : GSE31309, GSE38867 et GSE44124. Dans ce dernier cas, c'est une comparaison entre tissus sains et cancéreux qui a été menée. Enfin, deux autres ensembles ont été ajoutés pour l'analyse de l'expression différentielle des miARNs (et des gènes) dans les cellules souches embryonnaires et les cellules somatiques : GSE14473 et GSE42446. Ces deux dernières séries ont été utilisées dans le chapitre 3 de ce manuscrit. La composition plus précise et les publications associées à ces différents ensembles de données sont exposées dans le Tableau 3.

Tableau 3. Composition en échantillon des données d'expression de miARN. Xn : nombre d'échantillon (X17 : 17 échantillons du même type tissulaire)

Identifiant	Technology	Composition	Publication
GSE1950	Agilent	Cerebellum sain X3	
		Cortex préfrontal sain X3	
		Foie sain X3	
GSE23527	LC	Muscle squelettique jeune sain X 19	(Drummond et al., 2011)
		Muscle squelettique vieux sain X 17	
GSE24205	Agilent	Sang sain X21	(Mattila et al., 2011)
		Sang cancéreux X20	
GSE25508	Agilent	Poumons sains exposés à l'amiante X13 (différentes parties)	(Nymark et al., 2011)
		Poumons sains non-exposés à l'amiante X13 (différentes parties)	
		Poumons cancéreux exposés à l'amiante X13 (différentes parties)	
		Poumons cancéreux non-exposés à l'amiante X13 (différentes parties)	
		Autres X8 (commerciaux, hamartomés, tuberculosés, etc.)	
GSE31309	Febit	Sein sain X57	(Schrauder et al., 2012)
		Sein cancéreux X48 (stade précoce)	
GSE34933	Agilent	Prostate saine X12 (différents types)	(Chen et al., 2012; He et al., 2012, 2013)
		Prostate cancéreuse X12 (différents types)	
GSE38389	Exiqon	Muqueuse colorectale saine X71	(Gaedcke et al., 2012)
		Muqueuse colorectale cancéreuse X69	
GSE38867	Agilent	Sein sain X7	
		Sein cancéreux X7 (ductal)	

		Sein cancéreux X7 (invasif)	
		Sein cancéreux X7 (métastatique)	
GSE44124	Agilent	Sein sain X3 (Pool de 10 échantillons)	(Feliciano et al., 2013)
		Sein cancéreux X50	
GSE14473	Agilent	Lignées cellulaires souches non différentiées X9 (9 lignées)	(Stadler et al., 2010)
		Lignées cellulaires souches différentiées X9 (9 lignées)	
		N-tera non-différentiée	
		N-tera différentiée	
		Placenta X3	
GSE42446	Agilent	Lignée cellulaire souche non différentiée X10	(Koyanagi-Aoi et al., 2013)
		Cellule souche induite X49	
		Lignée cellulaire somatique X6	
		Lignées cellulaire cancéreuse X5	
		Lignée cellulaire cancéreuse souche X2	

2. Traitement des données brutes

Les données brutes (identifiant GSE) de ces différentes séries de données ont été téléchargées et extraites de GEO. En pratique, chaque fichier de données brutes dans GEO est associé à un échantillon bien spécifique (identifiant GSM). Pour une même expérience, « n » fichiers doivent donc être lus et transformés en matrice à deux dimensions, où chaque colonne représente un échantillon et chaque ligne, une sonde. Lorsqu'au moins une valeur négative était présente dans les données, ces dernières ont été ajustées en ajoutant une constante telle que la valeur minimale observée sur l'ensemble soit égale à 1 (Reich et al., 2006) :

$$si\ \min(x) \leq 0\ alors\ \tilde{x}_i = x_i + abs(\min(x)) + 1$$

où x_i est l'expression brute de la sonde i et $\tilde{x}_i$ la nouvelle valeur d'expression estimée de la sonde. Après ce premier ajustement, les données ont été transformées en log2. Enfin, une normalisation au quantile a été utilisée sur les données transformées. Dans certains cas, plusieurs sondes correspondant à un même miARN (ou gène) sont sur les plaques. Dans cette

situation et afin d'obtenir une valeur d'expression unique par miARN (ou gène), c'est la médiane d'expression des sondes qui a été utilisée.

Le principe de la normalisation au quantile est de faire en sorte que plusieurs distributions aient les mêmes propriétés statistiques, notamment en termes de quantiles (médiane, 1er et 3ème quartile, etc.). La procédure commence par ordonner chaque colonne dans l'ordre croissant tout en conservant l'information des rangs pour chaque individu. Pour chaque ligne, la moyenne est ensuite calculée et est affectée en lieu et place des valeurs initiales. Cette nouvelle matrice est alors réordonnée selon les rangs initiaux pour former la matrice normalisée :

Matrice initiale				
A	6	4	9	3
B	9	8	2	2
C	1	2	5	7

Rangs des données				
A	ii	ii	iii	ii
B	iii	iii	i	i
C	i	i	ii	iii

Matrice ordonnée				
A	1	2	2	2
B	6	4	5	3
C	9	8	9	7

Moyenne des lignes

1.75 = i
4.5 = ii
8.25 = iii

Matrice normalisée				
A	4.5	4.5	8.25	4.5
B	8.25	8.25	1.75	1.75
C	1.75	1.75	4.5	8.25

Après cette transformation, les colonnes de la matrice ont la même distribution et peuvent être aisément comparées. Elle a été implémentée avec la fonction *quantile.normalize* des *packages* R affyPLM (Bolstad, 2004), AgiMicroRna (López-Romero et al., 2010) ou preprocessCore (Bolstad, 2013) en fonction du type de données analysées. Par ailleurs, certains de ces packages permet également la lecture et l'extraction des données de manière automatique avec des fonctions comme *read.image*. La force de la méthode de normalisation quantile réside dans sa simplicité mais également dans sa grande sensibilité et sa grande spécificité (Pradervand et al., 2009).

Pour les données d'expression utilisées au chapitre 2, une étape supplémentaire d'analyse a été menée au travers de l'utilisation du *package* Limma (Smyth, 2005) pour déterminer la significativité statistique sur l'expression différentielle du miR-940 sur les trois ensembles de données. Limma introduit une relation linéaire entre les valeurs d'expression observées et les conditions expérimentales et permet le calcul de p-valeurs sur la significativité statistique de l'expression différentielle. Elle est considérée comme l'une des méthodes les plus robustes dans ce genre de cas (Bolstad et al., 2003).

3. Analyse différentielle

L'analyse différentielle d'expression est une méthodologie de comparaison de l'expression d'une entité entre deux conditions (ou plus). Cette analyse se fait généralement en comparant les expressions en log2 entre les conditions – on parle alors d'analyse de *log Fold Change* (*logFC*). Par exemple, si A et B sont deux conditions distinctes (sain et cancéreux), l'expression différentielle d'une entité est définie comme :

$$logFC = log2\left(\frac{A}{B}\right) = log2(A) - log2(B)$$

Comme les expressions peuvent grandement différer d'une méthode à l'autre mais aussi d'un ensemble de données à un autre, il n'y a pas de réel consensus pour définir le seuil pour lequel on considère avoir une expression différentielle. Ainsi, de nombreuses études reposent sur des tests statistiques plutôt que des seuils sur le *logFC*. C'est notamment ce que cherche à faire le *package* Limma, mais nous pouvons également citer d'autres méthodes plus directes comme le test-t, le test Wilcoxon ou encore l'ANOVA (Cui and Churchill, 2003).

L'utilisation du *Fold Change* ne permet pas d'obtenir de niveau de confidence sur les conclusions mais reste facilement interprétable : un *logFC* de 2 (en base 2) indique par exemple une expression quatre fois supérieure dans la première condition par rapport à la deuxième. L'intérêt de cette mesure dans notre cas réside précisément dans cette facilité

d'interprétation puisque le logFC peut très bien être apposé sur le réseau comme attribut des nœuds afin d'apporter une information supplémentaire aux réseaux de miARNs.

Les logFC seront utilisés essentiellement dans le chapitre 3 afin de déterminer l'expression différentielle des miARNs (et des gènes) entre les cellules souches embryonnaires et les cellules somatiques après la normalisation de leur expression par la méthode des quantiles exposée plus haut (cf. page 100). Pour GSE14473, le logFC a été calculé pour toutes les lignées de l'ensemble par paire (p. ex. cellules H1 non différenciées contre cellules H1 différenciées). Pour GSE42446, un logFC entre chaque lignée souche et chaque lignée somatique a été calculé, soit 60 logFC différents (Tableau 2 : 10 lignées souches pour 6 ignées somatique). La moyenne de ces mesures sur les deux ensembles de données a été utilisée pour être affichée sur les réseaux de miARNs, nous considérons donc un niveau moyen d'expression différentielle des miARNs.

Des seuils de +1,2 et +2,3 pour GSE14473 et GSE42446 respectivement ont été définis pour considérer les miARNs les plus différentiellement exprimés. Ces seuils correspondent en fait aux trente miARNs les plus surexprimés dans les cellules souches.

Comme un miARN est supposé réduire l'expression des gènes et des protéines, une étude conjointe entre les miARNs induits et les gènes réprimés dans les cellules souches a également été menée. Pour l'expression différentielle des gènes sur GSE42446, la même procédure a été conduite mais deux seuils différents ont été définis pour l'étude : un premier à -1.5, définissant les gènes moyennement réprimés (environ 2 200 gènes) et un deuxième à -3 (environ 800 gènes), définissant les gènes les plus réprimés dans les cellules souches.

4. Enrichissement en hits

Après avoir exposé les logFC (ou tout autre score) sur le réseau (c'est-à-dire, comme attribut des nœuds) et trouvé des communautés, une analyse particulièrement intéressante est de chercher les communautés montrant des enrichissements en hits (p.ex. beaucoup de

forts logFC ou de nombreux scores de crible particulièrement élevés au sein d'une même communauté).

Pour ce faire, une étude classique de « randomisation » par permutation a été utilisée : l'objectif de cette méthode est de comparer la valeur moyenne (ou médiane) des scores des membres de la communauté analysée à une distribution aléatoire de scores moyens (ou médians). Dans le cas d'une communauté à *n* nœuds, *x* sélections aléatoires (typiquement 10 000) de *n* nœuds dans le réseau sont effectuées. Le score moyen (ou médian) des scores associés aux nœuds est alors calculé à chaque étape afin de construire une distribution aléatoire (ou modèle nul). La p-valeur est enfin définie comme la fraction d'évènements – dans la distribution aléatoire – supérieure à la valeur moyenne des scores des membres de la communauté réelle. La p-valeur maximale que l'on puisse obtenir dans cette situation dépend du nombre de répétitions : cette limite est égale à 1/*n*.

G. miARNs : séquences, emplacements génomiques, alignements

L'information relative aux miARNs, notamment leur séquence mature, leur emplacement génomique et leurs différents identifiants ont été récupérés principalement depuis le site web de miRBase. Le packages BiomaRt a fourni l'emplacement chromosomique (Durinck et al., 2005).

Les séquences des miARNs (et notamment les séquences seed 8-mer 2 à 9) ont été alignés grâce au logiciel Jalview (Clamp et al., 2004) et à l'algorithme d'alignement multiple ClustalW implémenté dans Jalview en utilisant les paramètres définis par défaut (Thompson et al., 1994). Un arbre de distances moyennes a ensuite été calculé sur les alignements. Cet arbre a permis la découverte de douze clusters de *seed* de miARNs après seuillage (coupure dans à un certain niveau de l'arbre correspondant aux branches les plus longues). Toujours avec ce même logiciel, une séquence consensus pour chacun des clusters a été calculée.

H. Cribles ARNi

Dans le cadre du chapitre 3, les données d'un crible ARN interférence ont également été utilisées. Dans ce crible *genome-wide*, des LNA (943 LNA différents visant 943 miARNs de la version 17 de miRBase et 38 miARNs propriétaires miRPlus^{TM}) ont été utilisés afin de trouver les miARNs impliqués dans le maintien de la totipotence ou, au contraire, ceux induisant la différenciation. L'objectif était de bloquer l'action des miARNs dans des cellules souches et d'observer les conséquences sur l'expression de la protéine POU5F1 (ou OCT4) (au travers d'une construction reportrice GFP) – utilisée ici comme marqueur de la totipotence des cellules sur des cellules souches embryonnaires H1 (Figure 28).

1. Protocol expérimental

Dans un premier temps, les puits de 9 plaques 384 puits (Grenier) ont été revêtues de 10 µl de matrigel à concentration finale de 0,65 mg/ml (dilué dans du milieu DMEM/F12 sans enzyme) pour 30 min à 37°C et passées à la centrifugeuse à 1400 RPM puis laissées dans un incubateur pendant 2 heures à 37°C. Après ce temps, le surnageant a été retiré et les plaques ont été gardées à 4°C. Les LNA (Exiqon) et siARN contrôles Oct-4, Nanog, GFP et AllStars (Dharmacon) ont ensuite été déposés dans les puits à une concentration finale de 50 nM (4 répliques pour les LNA). Pour la transfection, un mix de 0,05 µl de Dharmafect1 (Dharmacon) et de 4,95 d'OptiMEM (Invitrogen) a été ajouté dans chaque puits et les plaques ont été mises à incuber pour 20 min.

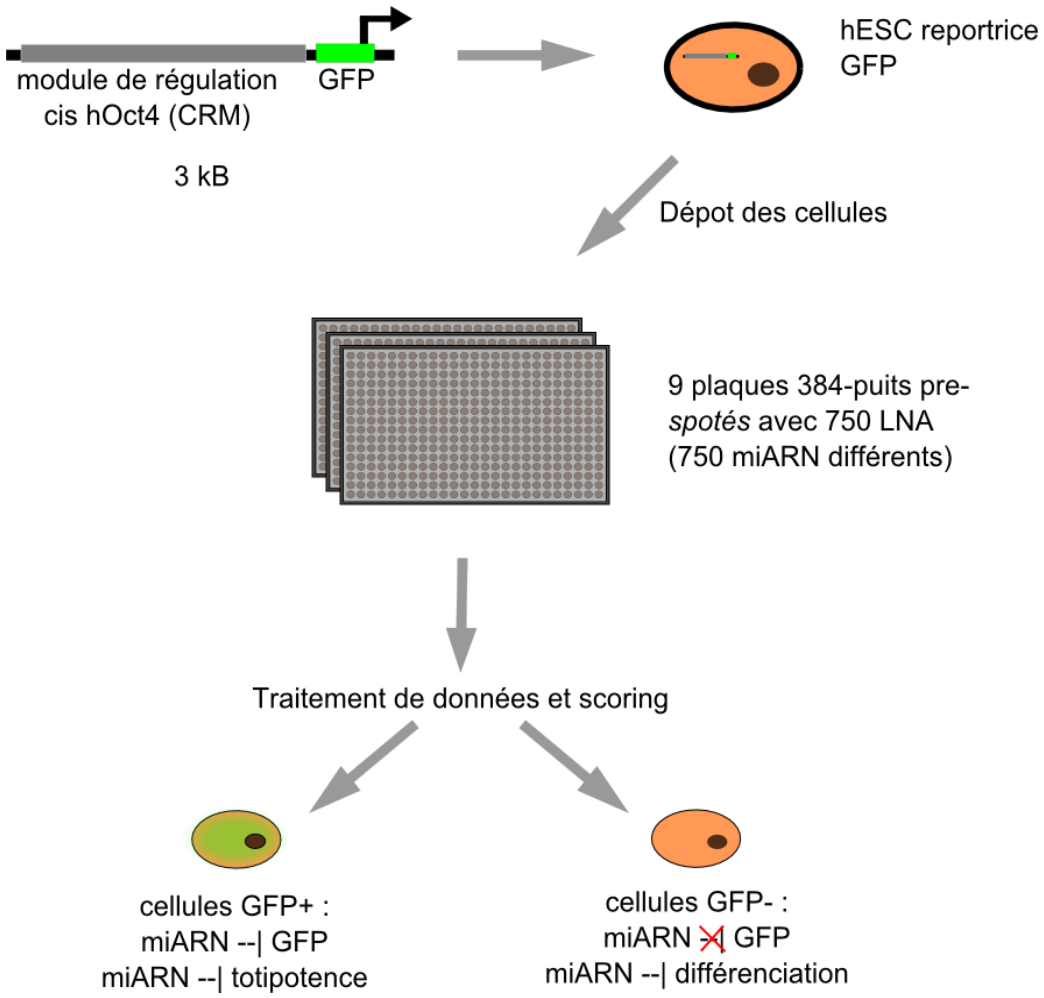


Figure 28. Schéma du crible ARNi. Environ 750 inhibiteurs de miARN différents **(**LNA) visant des miARNs ont été utilisés pour déterminer les miARNs influençant le devenir des cellules souches. Une cellule fluorescente correspond dans ce cas à une cellule totipotente, c'est-à-dire ayant gardé son potentiel de différenciation alors qu'une cellule sans fluorescence correspond à une cellule différenciée.

Pour cette étude, des cellules H1 (Thomson et al., 1998) modifiées ont été utilisées. Une région de 3 064 paires de base de la région promotrice du gène humain POUF5F1 (OCT4) a été clonée en amont d'un gène rapporteur GFP, remplaçant le promoteur de Cytomégalovirus du plasmide N-EGFP avec un marqueur de sélection à la Généticine (Gibco – présentement Life Technology, Carlsbad, Californie). Cette construction POU5F1-GFP (2 µg) a été transfectée dans les cellules H1 en utilisant du FuGENE (Roche, Bâle, Suisse). Les lignées résistances sont apparues 2 semaines après la sélection chimique.

Le lignée cellulaire H1 POU5F1-GFP (Chia et al., 2010) a alors été déposée – après passage à la trypsine – dans les puits pour un total d'environ 3 000 cellules par puits dans 40 µl de milieu de culture et 10 µM d'inhibiteur Rock (Calbiochem – présentement Merck KGaA, Darmstadt, Allemagne). La transfection s'est faite pendant 72 heures avec changement de milieu tous les 24 heures. 12 heures après le dépôt des cellules, un changement de milieu a permis l'évacuation du mix de transfection.

Après fixation des cellules, la lecture des plaques a été faite à l'ArrayScan Cellomics® (Life Technologies, Carlsbad, Californie) et les résultats sur l'intensité du signal GFP (cellule à cellule) ont été exportés en fichier csv. L'analyse de ces fichiers a été faite avec le logiciel R. Un fichier de données standard comporte des milliers voir des millions de lignes, représentant chacune une cellule unique avec ses différentes caractéristiques formant les colonnes du fichier (intensité de fluorescence, taille, circularité, traitement associé, emplacement, etc.).

2. Analyse des données

Brièvement, l'analyse de cribles ARNi (ou chimique) passe principalement par quatre étapes (Boutros and Ahringer, 2008) :

- Lecture et analyse des données brutes (recherche d'*outliers*, seuillage, etc.)
- Analyse de qualité (Z' factor, étude des contrôles, etc.)
- Normalisation et *scoring* (Zscore, Bscore, etc.)

- Sélection de hits

L'analyse des données consiste à détecter – et corriger le cas échéant – différents biais pouvant influencer les analyses. Ces biais peuvent être de nature aléatoire ou systématique (Caraus et al., 2015). Les biais aléatoires sont des biais liés aux différences intrinsèques qui existent d'une expérience à l'autre (p.ex. des données aberrantes locales à un ou quelques puits). Ces derniers sont essentiellement corrigés par la mise en place de répliques techniques et biologiques. Les effets spatiaux, quant à eux, sont des biais systématiques (mais pas les seuls). Ce sont essentiellement des problèmes techniques et environnementaux pouvant significativement influencer le nombre de faux positifs lors de la sélection de hits. Un exemple de biais systématiques serait un séchage non uniforme entrainant une augmentation du nombre de hits en bord de plaque (« *border effect* ») ou encore une erreur robotique entrainant des « *patterns* » remarquables (ligne à ligne ou colonne à colonne). Il existe différentes méthodes pour circonvenir à ces problèmes (correction *lowess*, Bscore, SPAWN, etc.) mais que nous n'aborderons pas dans ce manuscrit. L'objectif de ces méthodes est de transformer les données de telle sorte que les effets spatiaux soient corrigés.

L'étude de la qualité des plaques forme un prérequis à la poursuite de l'étude du crible puisqu'elle définit l'exploitabilité des données. Le Z' factor (Zhang et al., 1999) est une mesure simple de la qualité des plaques qui se base sur les contrôles internes aux plaques (à ne pas confondre avec le Zscore) :

$$Z'factor = 1 - \frac{3(\sigma_p + \sigma_n)}{|\mu_p - \mu_n|}$$

où σ_p et μ_p représentent respectivement l'écart type et la moyenne associée à un contrôle positif (ayant un effet marqué, p.ex. un siARN dirigé vers Oct4-GFP dans notre cas) et σ_n et μ_n, ces mêmes mesures associées à un contrôle négatif (sans effet attendu). Pour les cribles chimiques, un Z' factor de 1 équivaut à un crible parfait où la démarcation entre les contrôles positif et négatif est grande. Lorsque cette mesure est inférieure à 0.5, le crible est considéré

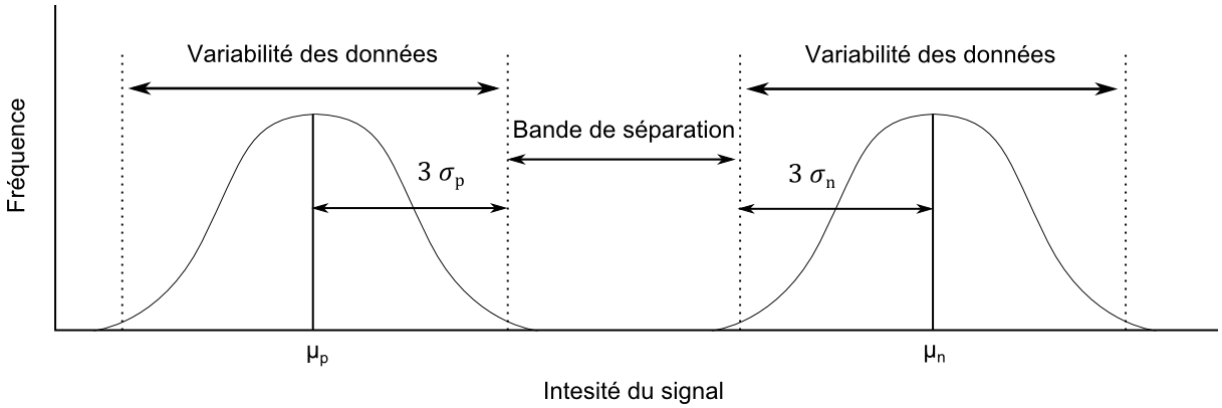


Figure 29. Illustration de la variation et la séparation des données pour les contrôles (positif « p » et négatif « n ». Un Z'factor élevé indique une forte séparation (grande bande de séparation) entre les deux distributions. A 0 (ou moins), il n'y a aucune bande de séparation et les deux distributions se chevauchent. D'après (Zhang et al., 1999)

comme peu exploitable. Entre 1 et 0.5, le crible est très bon et peut être analysé sans problème (Figure 29). A cause de l'absence de très bons contrôles négatifs et positifs pour les cribles ARN interférence, ces chiffres ne sont pas parfaitement adaptés à ce type d'études. Il n'est donc pas rare de voir des cribles ARN interférence avec des Z' factor inférieur à 0,5 sans pour autant être non-exploitables.

Dans le cas de données cellule-à-cellule, le Z'factor peut être calculé sur les données brutes ou sur les données agrégées (par champs, puits, conditions ou des combinaisons de ces trois dernières). Les données brutes montrant généralement beaucoup de variabilité, il est souvent plus intéressant d'agréger les données dans un premier temps (par champs, puis par puits) pour calculer le score. C'est la solution qui a été choisie dans cette analyse. Pour agréger les données cellules-à-cellules, une première médiane a été calculée pour chaque champ de vue sur l'intensité de GFP. Ces médianes ont ensuite été utilisées pour calculer une médiane de l'intensité de GFP par puits (Figure 30).

Après l'étape de vérification des données, vient la normalisation et le *scoring*. La méthode de normalisation Zscore (ou score standard) est probablement la méthode la plus utilisée pour le criblage haut-débit :

$$Zscore_i = \frac{x_i - \mu_n}{\sigma_n}$$

Cellules	Puits	Champs	Traitements	GFP
Cellule1	PuitsA	Champ1	siARN1	150
Cellule2	PuitsA	Champ1	siARN1	120
Cellule3	PuitsA	Champ1	siARN1	124
Cellule4	PuitsA	Champ2	siARN1	160
Cellule5	PuitsA	Champ2	siARN1	167
Cellule6	PuitsA	Champ2	siARN1	154
Cellule7	PuitsA	Champ2	siARN1	128
Cellule8	PuitsA	Champ3	siARN1	103
Cellule9	PuitsA	Champ3	siARN1	189
Cellule10	PuitsA	Champ3	siARN1	150
Cellule11	PuitsB	Champ1	siARN2	70
Cellule12	PuitsB	Champ1	siARN2	75
Cellule13	PuitsB	Champ1	siARN2	62
Cellule14	PuitsB	Champ2	siARN2	54
Cellule15	PuitsB	Champ2	siARN2	79
Cellule16	PuitsB	Champ2	siARN2	92
Cellule17	PuitsB	Champ2	siARN2	160
Cellule18	PuitsB	Champ3	siARN2	95
Cellule19	PuitsB	Champ3	siARN2	76
Cellule20	PuitsB	Champ3	siARN2	72

Agrégat par champs

Champs	Puits	Trts	GFP
Champ1	PuitsA	siARN1	124
Champ2	PuitsA	siARN1	157
Champ3	PuitsA	siARN1	150
Champ1	PuitsB	siARN2	70
Champ2	PuitsB	siARN2	85.5
Champ3	PuitsB	siARN2	76

Agrégat par Puits

Puits	Trts	GFP
PuitsA	siARN1	150
PuitsB	siARN2	76

Figure 30. Exemple d'agrégation des données par champs puis par puits. A chaque étape, l'intensité de GFP est agrégée en calculant une médiane.

où i représente une condition ; x, la valeur associée à cette condition et σ_n et μ_n représentent respectivement la déviation standard et la moyenne des contrôles négatifs (ou de la plaque entière). Une variation plus robuste de ce score repose sur l'utilisation de la médiane et la déviation absolue de la médiane (*Median Absolute Deviation* – MAD) en lieu et place de la moyenne et de la déviation standard – on parle dans ce cas de Zscore robuste.

Cette méthode de normalisation fait également office de méthode de *scoring* puisqu'elle possède les mêmes caractéristiques que les distributions normales centrées réduites : le Zscore (classique) représente la déviation des valeurs associées aux conditions étudiées par rapport à une moyenne. Un Zscore supérieur à 3 (déviation supérieure à 3σ par rapport au centre de la distribution) représente environ 0.1% des observations (soit une probabilité d'être observé au hasard de 0.0027) : une observation déviant assez de la moyenne pour être significativement considérée comme un hit. La distribution, supposée

Gaussienne (sous l'hypothèse nulle) étant symétrique, un score inférieur à -3 représente aussi 0.1% des observation et correspond à une baisse de la fluorescence. Dans notre cas, un Zscore négatif équivaut à une baisse de fluorescence de la GFP. Sachant que les LNA inhibent les miARN, nous pouvons dire que le miARN visé est impliqué dans la différenciation. A l'inverse, un Zscore positif montre une augmentation de fluorescence et ainsi une implication des miARNs dans la maintenance de la totipotence. Tout comme le Z'factor, le Zscore a préférablement été calculé sur les données agrégées (par champs puis par puits).

De la même manière qu'avec les logFC, ces scores ont été appliqués sur le réseau afin de découvrir des communautés potentiellement enrichies en hit, en suivant les mêmes étapes qu'exposées précédemment et notamment l'approche par *randomisation*. Dans ce cas en revanche, c'est la valeur absolue des scores qui a été étudiée pour prendre en compte tous les LNA ayant une activité sur le devenir des cellules souches.

Chapitre 1 :
Construction et analyse topologique de réseaux de microARNs

A. Introduction

L'hypothèse de base du travail présenté dans cette thèse est la possibilité pour les miARN de réguler des processus biologiques similaires lorsque ces derniers partagent des cibles en commun. Comme nous l'avons déjà évoqué, l'idée de corégulation n'est pas nouvelle et semble par ailleurs prendre de plus en plus de poids dans notre compréhension du rôle biologique des miARN (Vidigal and Ventura, 2014). Cette corégulation peut être de plusieurs natures : coopérative, additive, synergique ou encore dégénérée.

Dans le cas de la coopérativité, les miARN sont capables d'interagir ensemble lors de la régulation génique, cette coopération peut ou non amplifier l'effet des miARN séparément. L'additivité et le synergisme sont des formes particulières de coopération. Pour l'additivité, les miARN agissent de concert sur des gènes et l'effet de l'un et l'autre s'additionne pour réguler de manière plus forte l'expression génique. Le synergisme, quant à lui, est un effet qui va au-delà de la simple additivité et permet donc une régulation encore plus forte par deux entités conjointes. Dans le cas de la dégénérescence, les miARN n'agissent pas forcément de concert mais peuvent se substituer les uns aux autres dans diverses situations pour maintenir l'état cellulaire (robustesse du système) – on parle aussi souvent de « redondance fonctionnelle ».

Pour répondre à ces différentes hypothèses et trouver des familles de miARN pouvant interagir, le principe a été dans un premier temps de construire des réseaux de miARN basés sur le partage de cibles et dans un second temps, d'analyser la structure et les propriétés de ces différents réseaux. Nous nous intéresserons en premier lieu aux prédictions chez l'être humain dans DIANA-microT puis dans TargetScan. Nous analyserons et comparerons alors en fin de chapitre des topologies de réseaux chez différentes espèces.

B. DIANA-microT : construction et analyse d'un réseau de miARN

La méthode de construction des réseaux repose sur un principe simple : le calcul du pourcentage de cibles (prédites) partagées entre tous les miARN pris deux à deux par l'indice meet/min ($meet/\min = {A \cap B}/{\min(A,B)}$, Figure 31) et la définition d'un seuil optimal pour créer des réseaux binaires. Une fois ce seuil défini, il s'agit d'analyser la structure et les propriétés des réseaux. La plupart des résultats et des validations présentés dans cette partie et le chapitre suivant sont basées sur l'algorithme de prédiction de cibles de miARN DIANA-microT v3, contenant les prédictions pour 555 miARN couramment étudiés. Ces 555 miARN sont ceux qui étaient présents dans la base de données miRBase lors des prédictions par la 3ème version de DIANA-microT courant 2009. En conséquence, ces miARN ont statistiquement plus de chance d'avoir été étudiés par une équipe et l'information disponible pour ces derniers est plus grande que celle concernant des miARN plus récents.

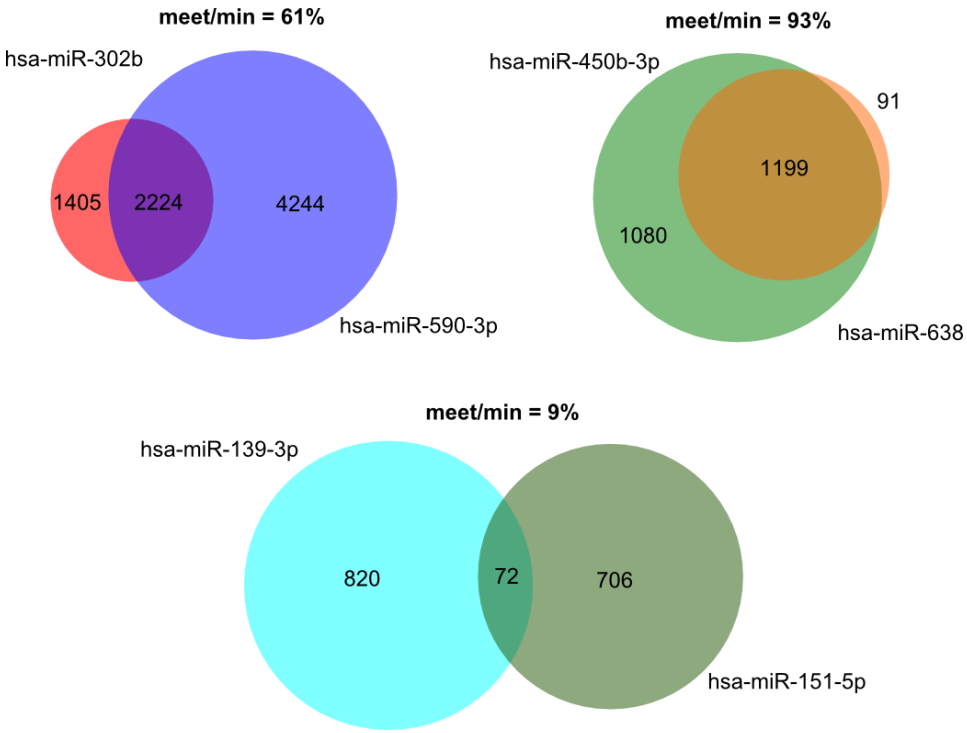


Figure 31. Exemple de recouvrement de cibles entre miARN et indice meet/min.

Les réseaux non-seuillés sont des réseaux pondérés avec une « force » pour chaque lien. L'objectif du seuillage est de réduire le nombre de liens afin de diminuer la densité des réseaux. En effet, même si de nombreux algorithmes ont été implémentés en biologie des systèmes pour analyser les réseaux pondérés, la plupart de ces algorithmes ont été

développés pour des réseaux clairsemés (faible densité) – c'est-à-dire pour lesquels la matrice d'adjacence est creuse. Ainsi, même si ces algorithmes sont NP-difficiles (***N****on deterministic* ***P****olynomial-time hard)*, cela ne pose généralement aucun problème en biologie des systèmes classique à cause de l'espace faible à parcourir dans les matrices d'adjacence. Comme notre réseau non seuillé est un réseau ultra-dense, ces méthodes n'y sont clairement pas adaptées. Par conséquent et malgré la perte d'information, nous avons choisi de simplifier le réseau pondéré en réseaux binaires par le seuillage de l'indice meet/min. Le choix du seuil étant une étape souvent délicate à cause de la subjectivité induite par ce choix, nous avons utilisé une approche par « seuil multiple » pour pallier ce problème. Ces approches consistent en fait à comparer les réseaux à différents seuils afin de déterminer un seuil optimal en fonction des comparaisons (van Wijk et al., 2010; Langer et al., 2013).

1. Construction et choix de seuil optimal

La Figure 32 montre différents réseaux binaires créés à partir de la base de données DIANA-microT pour différents seuils meet/min. A 40%, la plupart des miARN sont encore interconnectés et l'analyse à l'œil est impossible. C'est à partir de 50% de cibles communes que le réseau prend une forme facilement interprétable : le réseau semble être formé de deux parties extrêmes reliées par une plus petite partie centrale. A partir de 60%, le réseau global est de plus en plus petit et de petites communautés isolées apparaissent au fur et à mesure que le seuil augmente. Au-delà de 80%, seules les familles de miARN sont encore observées. Ces familles sont essentiellement des miARN ayant presque exactement les mêmes séquences (la famille let-7 p.ex. avec let-7a, let-7b, let-7c, etc.). Si cette information est intéressante, elle n'est cependant pas nouvelle puisque les familles de miARN basées sur les séquences *seed* sont déjà connues. Il s'agissait en effet de trouver des familles plus larges (en termes de partage de cibles) que celles uniquement basées sur les séquences.

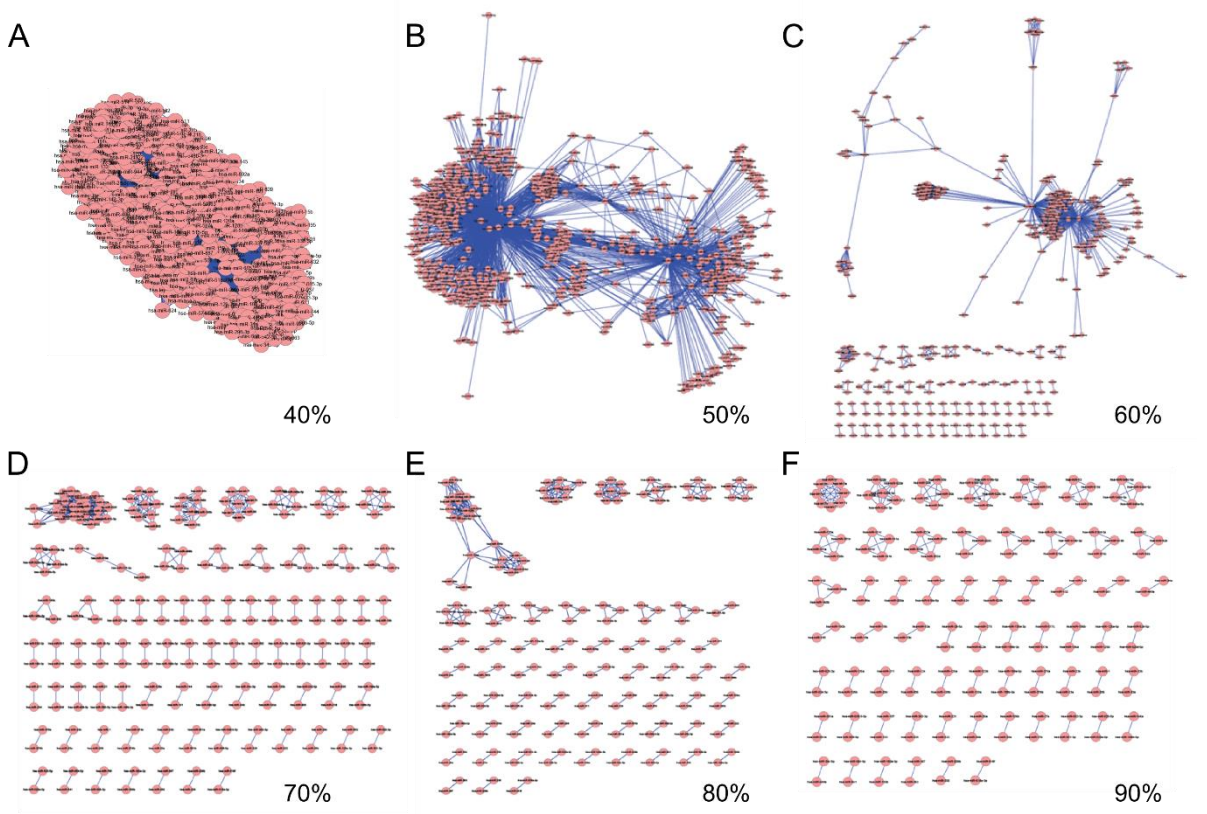


Figure 32. Réseaux de partage de cibles de miARN pour différents seuils meet/min. A | Seuil = 40% (meet/min). Seuls les miARN avec au moins 40% de cibles communes possèdent un lien dans le réseau. Presque tous les miARN sont interconnectés, la densité de réseau est grande. B | Seuil = 50%. Un réseau à plus faible densité, plus facilement analysable. C | Seuil = 60%. Le réseau montre beaucoup moins de nœuds interconnectés mais quelques communautés isolées du reste du réseau. D | Seuil = 70%. A partir de ce niveau-là, il n'y a pas plus que des petites communautés faisant référence aux familles de miARN (let-7 p.ex.). E | Seuil = 80%. F | Seuil = 90%. Seules les familles avec des séquences quasiment identiques sont retrouvées.

Différentes propriétés de réseaux en fonction de seuils meet/min sont représentées sur la Figure 33. Tous les nœuds sont connectés à au moins un voisin jusqu'à meet/min 48%, seuil pour lequel il existe un faible nombre de liens dans le réseau. Au seuil 40%, la densité du réseau est proche de 0, ce qui indique un réseau clairsemé (*sparse*) avec beaucoup de nœuds mais peu de liens. Le coefficient de *clustering* atteint son minimum aux alentours d'une valeur meet/min de 50%. Seuil pour lequel les deux mesures de centralité sont proches de leur maximum respectif. C'est à ce niveau également qu'un maximum de nœuds connectés pour un minium de liens est observé : seuls 4 nœuds sont isolés du réseau pour 2911 liens. Le seuil meet/min pour ces prédictions a donc été fixé à 50%. On garde ainsi un maximum de nœuds connectés dans le réseau pour un minimum de liens.

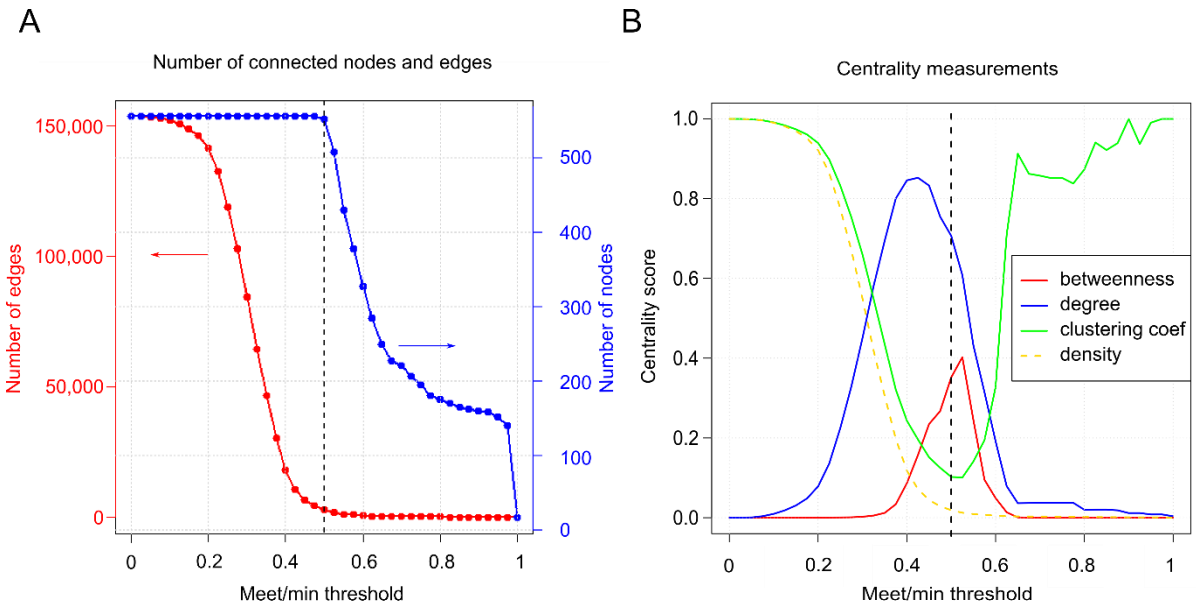


Figure 33. Caractéristiques des réseaux DIANA-microT pour différents seuils meet/min. A | Nombre de liens et nombre de nœuds non isolés en fonction de seuils meet/min. B | Centralisation *betweenness* et degré, coefficient de *clustering* (global) et densité des réseaux en fonction de seuils meet/min. Le seuil meet/min s'étend de 0 à 100% avec incrément de 1.

Pour mieux appréhender les caractéristiques de ces différents réseaux, une comparaison entre trois réseaux de miARN (meet/min 25%, 50% et 75%) et trois autres réseaux a également été menée (Tableau 4). Les trois autres réseaux étaient un réseau aléatoire (Erdős and Rényi, 1959), un réseau sans échelle (Barabási and Albert, 1999) et un réseau d'interaction protéine-protéine (Bu et al., 2003).

Tableau 4. Comparaison entre trois réseaux miARN à différents seuils meet/min, un graphe aléatoire, un graphe sans échelle et un réseau réel d'interaction protéine-protéine

	meet/min 50%	Interactome humain	Graphe sans échelle	Graphe aléatoire	meet/min 25%	meet/min 75%
Nœuds connectées	**551**	2361	551	551	555	195
Liens	**2911**	7182	2740	2911	119160	389
Degré moyen	**10.49**	6.08	9.95	10.57	429.41	1.4
Densité	**0.019**	0.0026	0.01	0.02	0.78	0.0025
Diamètre	**5**	16	8.07	5.01	2	3
Chemin caractéristique	**2.51**	4.65	2.26	2.92	1.22	1.26
Coef. de *clustering*	**0.1**	0.1	0.05	0.02	0.83	0.85
Centralisation *betweenness*	**0.35**	0.01	0	0.01	0.0005	0.0001
Centralisation de degré.	**0.71**	0.01	0.2	0.02	0.22	0.037

La densité du réseau d'interaction protéine-protéine est bien plus faible que celles des autres graphes, mis à part le réseau miARN à meet/min 75% où seules les familles de miARN sont retrouvées. En revanche, les valeurs de centralisation (centralité globale) du graphe à meet/min 50% sont en général bien supérieures à celles des autres graphes – tout comme le coefficient de *clustering* – dénotant donc une organisation sous-jacente du réseau reposant d'une part sur des hubs et, d'autre part, sur des groupes de miARN interconnectés. Etant données les grandes différences de centralisation entre le réseau à meet/min 50% et le réseau aléatoire, nous pouvons supposer que la forme du réseau n'est clairement pas liée au hasard. Avec un diamètre de 5 et un chemin moyen de 2,5, le réseau est également un graphe plutôt compact (Figure 32 B).

2. Analyse de la topologie du réseau à meet/min 50%

Sous cette condition et mise à part les caractéristiques de centralisation, le graphe montre un comportement proche des réseaux sans échelle ($R^2 = 0,64$) : il comprend beaucoup de nœuds à faible degré (points verts sur la Figure 34) et peu de nœuds à fort degré indiqués en bleus et rouges sur la Figure 34. Ces nœuds à fort degré sont les hubs du réseau, nous reviendrons sur cette notion un peu plus tard dans le texte (cf. page 122). Il est formé de modules et est disassortatif. En effet, la distribution de degré suit à peu près une loi de puissance (Figure 34 A). Le faible coefficient de détermination peut être expliqué par l'influence des quatre nœuds isolés que l'on peut observer dans la partie gauche de la Figure 34 A. Par ailleurs, sur cette même figure, nous pouvons noter la présence d'un groupe de hubs dans la partie basse droite (onze points les plus bas). Des anticorrélations entre le nombre de voisins des nœuds du réseau et le coefficient de *clustering* moyen (Figure 34 B) ainsi que le voisinage moyen de ces nœuds (Figure 34 C) sont également observés. La première information donne une indication sur la connectivité des voisins d'un nœud : les voisins des nœuds faiblement connectés sont souvent reliés (ils forment des clusters) contrairement aux voisins des hubs du réseau. Ces nœuds, à la fois faiblement connectés et *clusterisés*, constituent en fait une partie des familles que l'on retrouve à des seuils meet/min supérieurs : dans ce cas, le réseau est

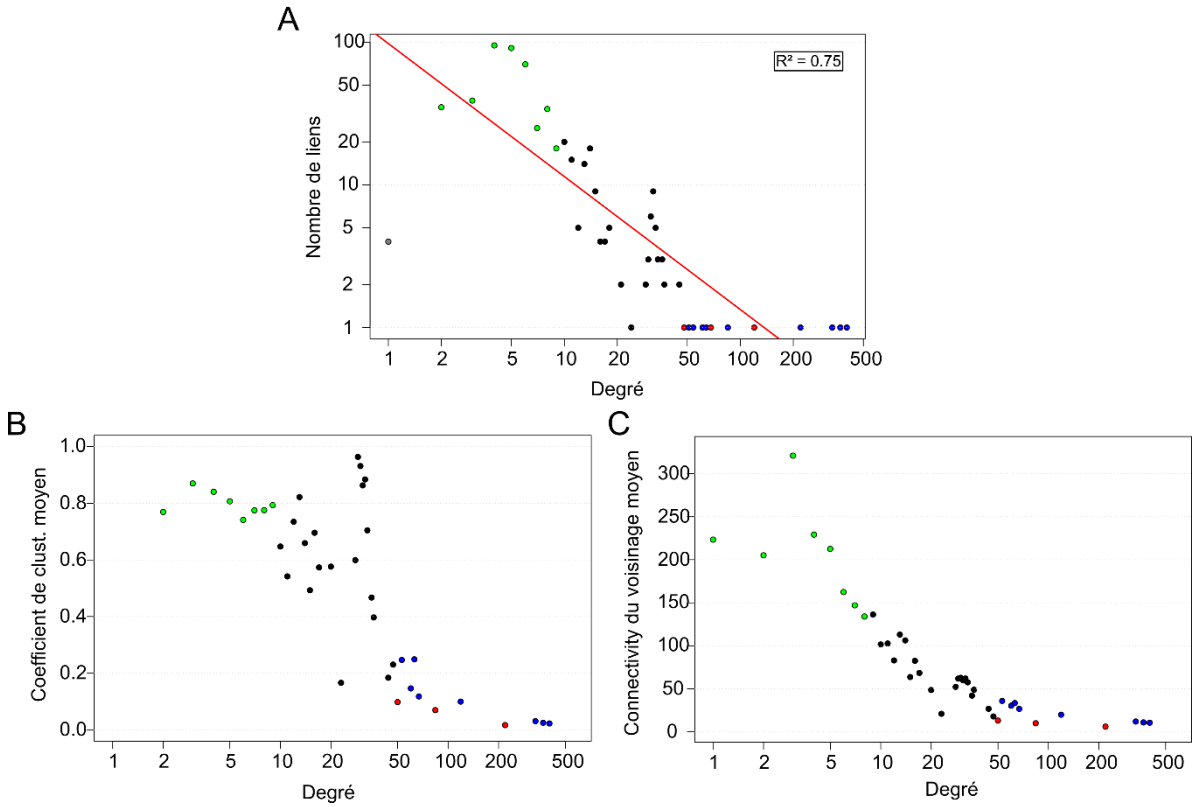


Figure 34. Caractéristiques du réseau DIANA-microT meet/min 50%. A | Distribution de degré. Une légère correspondance avec un graphe sans échelle est observée (R^2 = 0,75, calculé sans les quatre nœuds isolés indiqués en gris sur la figure). La plupart des nœuds du graphe sont des nœuds avec de faibles degrés. La ligne rouge désigne un ajustement linéaire sur le logarithme des données. B | Distribution de coefficient de clustering moyen par miARN. Donne une indication sur la connectivité des voisins d'un nœud. C | Distribution de la connectivité voisine moyenne. Représente la connectivité moyenne de tous les voisins d'un nœud pour un degré donné. En bleu et rouge sont indiqués les hubs du réseau et en vert les nœuds peu connectés.

qualifié de modulaire. La connectivité moyenne du voisinage, quant à elle, donne une information sur la propension des hubs à se connecter de préférence aux nœuds à faible degré – phénomène appelé « disassortativité ».

Toutes ces caractéristiques (densité faible, modularité, disassortativité et mesures de centralisation élevées) font du réseau de miARN de DIANA-microT un « petit monde » où l'information transite facilement d'un nœud vers un autre, avec quelques hubs centraux pouvant coordonner cette information. Dans notre cas, l'information transitant (même si l'on ne peut pas exactement parler de transit ici) correspondrait à la corégulation des processus biologiques. C'est à ce seuil meet/min bien spécifique de 50% que des caractéristiques très proches de ce que l'on connait aujourd'hui des réseaux biologiques sont observées : des réseaux peu denses, disassortatifs, modulaires et dominés par quelques hubs (Albert, 2005).

3. Les clubs assortis

Lorsque l'on analyse la densité des sous-graphes induits par les hubs en fonction de leur degré, nous pouvons constater que le degré des nœuds baisse très rapidement pour atteindre un premier palier entre dix et cinquante nœuds (Figure 35). Une faible rupture peut être observée au niveau du cinquantième nœud, seuil au-delà duquel le degré décroit alors stablement. Au vu de la Figure 35, il est difficile de considérer les nœuds au-delà de la quinzième position comme de véritables hubs puisqu'on se situe alors dans la zone de décroissance stable. Les onze premiers en revanche forment deux clusters : le premier indiqué en bleu sur la figure, avec une densité proche de 0.8 et un second indiqué en rouge, avec une

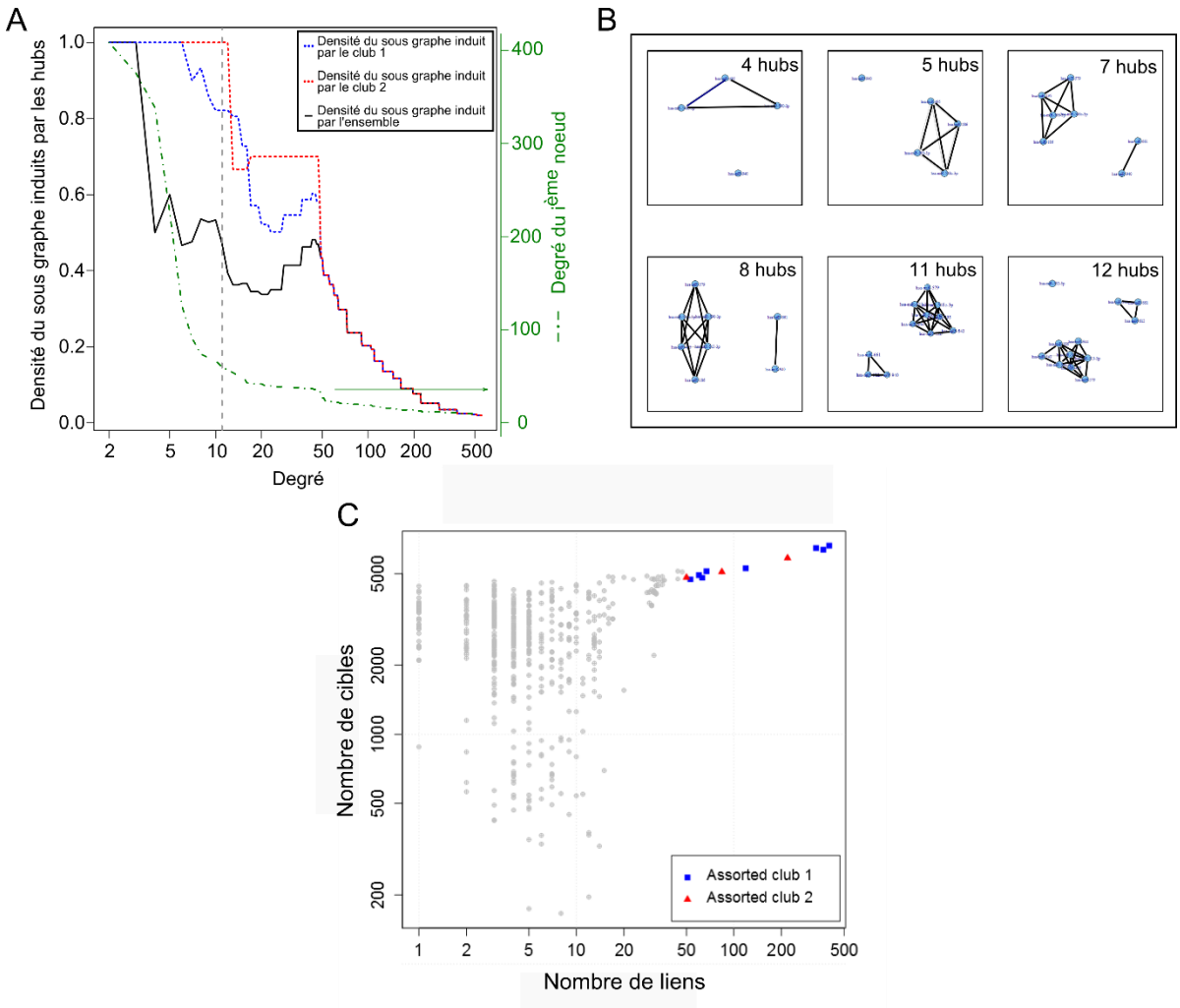


Figure 35. Détection des « clubs assortis ». A | Les nœuds sont classés du plus haut au plus faible degré. En abscisse sont représentés les sous-graphes induits par les *n* premiers hubs (p.ex. à *n* = 5, le sous-graphe induit par les 5 premiers hubs du réseau). B | Sous-graphes induits pour différent nombre de hubs classés par ordre décroissant. A 11 hubs, deux clusters séparés sont retrouvés avec des densités respectives de 0.8 et 1. C | Nombre de cibles prédites contre nombre de liens pour chaque nœuds du réseau. En bleu sont signalés les 8 membres du club assorti 1 et en rouge ceux du club assorti 2.

densité de 1. Le premier groupe est constitué de huit miARN et a été nommé « club assorti 1 » alors que le deuxième n'est formé que de trois miARN et a été nommé « club assorti 2 » (Figure 35 B). Ces densités très élevées montrent les fortes connectivités qui existent entre les miARN et, de façon plus intéressante, le fort nombre de cibles qu'ils partagent. Par ailleurs, ces connections très proches basées sur le partage de cibles renforcent grandement l'idée de régulation potentielle de processus biologiques en commun.

Le reste des nœuds se connectent alors à l'un ou l'autre des deux clubs créant aussi les deux pôles que nous pouvions déjà observer plus tôt dans le réseau global. Le quarante-huitième nœud relie les deux parties du réseau afin de ne former plus qu'un réseau unique (Figure 35 A). Etant donné la baisse brutale de densité pour les deux clubs assortis après l'ajout du douzième hub et sachant que ce dernier n'est connecté à aucun des deux clusters, les clubs assortis associés à cette base de prédictions ont été définis comme les onze premiers hubs. Les miARN formant ces deux clusters sont (dans l'ordre décroissant de degré) hsa-miR-495, -548c-3p, 590-3p, -940, -186, -661, -579, -513-3p, -543, -944 et -612. Hsa-miR-940, -661 et -612 sont par ailleurs les membres du club assorti 2.

Les membres des clubs assortis sont non seulement des hubs dans le réseau mais sont aussi les miARN avec le plus de cibles (Figure 35 C). En effet, malgré la méthode de normalisation imposée par l'indice meet/min, il existe tout de même une corrélation entre le nombre de liens des nœuds et le nombre de cibles prédites pour chaque miARN (surtout visible pour les hubs). La plupart de ces hubs possède également de très fortes valeurs de centralité *betweenness* (de 0,4 à 0.67). De fait, parmi les onze hubs formant les clubs assortis, sept sont retrouvés parmi les treize nœuds avec la plus forte centralité *betweenness* dans le graphe. Ces sept miARN (les trois du club assorti 2 et quatre, du club assorti 1) sont également placés à des positions très centrales dans le réseau et définissent deux pôles (Figure 36, en rouge et bleu). C'est particulièrement le cas des miARN hsa-miR-548c-3p, -590-3p et -495 pour le club assorti 1 et des trois miARN du club assorti 2. Les quatre autres miARN des clubs

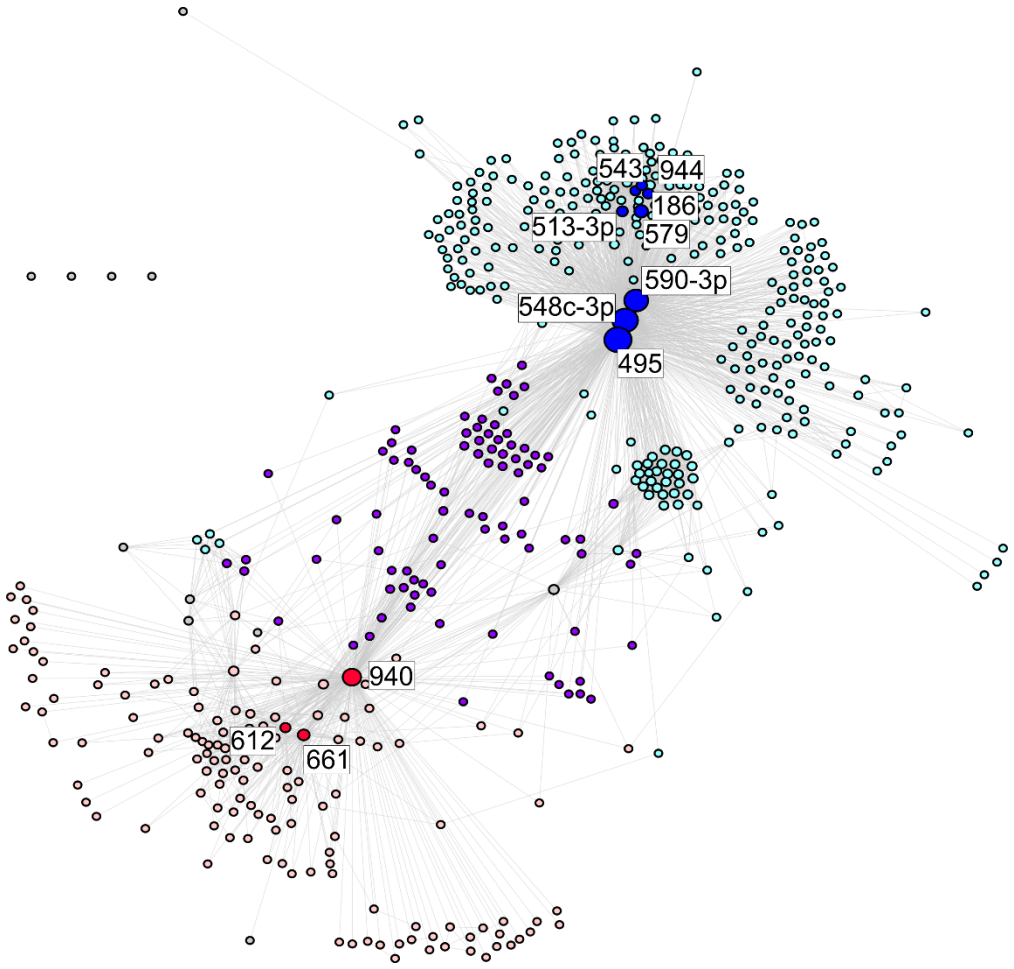


Figure 36. Le réseau de miARN de DIANA-microT. En bleu sont représentés les huit membres du club assorti 1 et en cyan tous les miARN connectés à au moins un de ces huit nœuds. En rouge sont représentés les trois membres du club assorti 2 et en rose, chaque miARN relié à au moins un de ces trois nœuds. Enfin en violet sont représentés les nœuds connectés à la fois au moins un des deux membres des clubs assortis et en gris ceux qui ne sont connectés à aucun des deux. La taille de chaque nœud est proportionnelle à leur degré.

assortis sont moins « centrés » dans le réseau ce qui explique leur valeur de centralité *betweenness* plus faible.

4. Zones d'influence des clubs assortis

La Figure 36 montre le réseau de miARN basé sur DIANA-microT à meet/min 50% coloré avec un code faisant référence à l'organisation sous-jacente liée aux clubs assortis. Dans ce réseau binaire, chaque nœud représente un miARN unique et un lien entre deux miARN indique que ces deux derniers partagent au moins 50% de cibles en commun. Les nœuds cyans représentent tous les miARN connectés à au moins un membre du club assorti 1, alors que les nœuds rose sont ceux connectés à au moins un membre du club assorti 2. En

violet sont représentés les nœuds connectés à la fois aux deux clubs et enfin, le gris représente ceux n'étant connectés à aucun des deux clusters.

Nous pouvons clairement constater que le réseau forme bien deux pôles (en rose et en cyan) qui sont typiquement organisés autour des deux clubs assortis représentés en rouge et en bleu. Ce code permet donc de clairement visualiser le rôle central et influençant des hubs sur le reste des miARN du réseau et notamment sur les deux pôles. A cause de ce rôle d'influence, les deux parties en rose et cyan ont été nommées « zone d'influence » des clubs assortis. Ces deux zones sont séparées par la zone violette, qui a été nommée « zone intermédiaire » à cause son caractère limitrophe entre les zones d'influence. La zone d'influence du club 1 est constituée de 315 miARN, celle du club 2 de 129 miARN et enfin, la zone intermédiaire est quant à elle formée par 89 miARN différents – soit 11 miARN non reliés à ces différents groupes. Cette organisation bien spécifique explique parfaitement les fortes valeurs de centralisation que l'on observe pour le réseau (Tableau 4).

5. Expression des miARN dans différents tissus

Afin d'évaluer le pouvoir de regulation du réseau de miARN, nous avons par la suite coloré le réseau en fonction de l'expression des miARN dans quatre tissus normaux (le sein, la muqueuse colorectale, les poumons, la prostate). Nous pouvons premièrement constater que les expressions sont assez similaires sur les différents tissus analysés avec des variations généralement mineures des miARN d'un tissu à l'autre (Figure 37). Ainsi, nous pouvons repérer quelques miARN formant des communautés coexprimées dans les différents tissus (encerclées en rouge et violet par exemple). Au contraire, d'autres groupes semblent plutôt différentiellement exprimés entre les tissus. Ces miARN seraient potentiellement des miARN permettant la différenciation entre les tissus considérés.

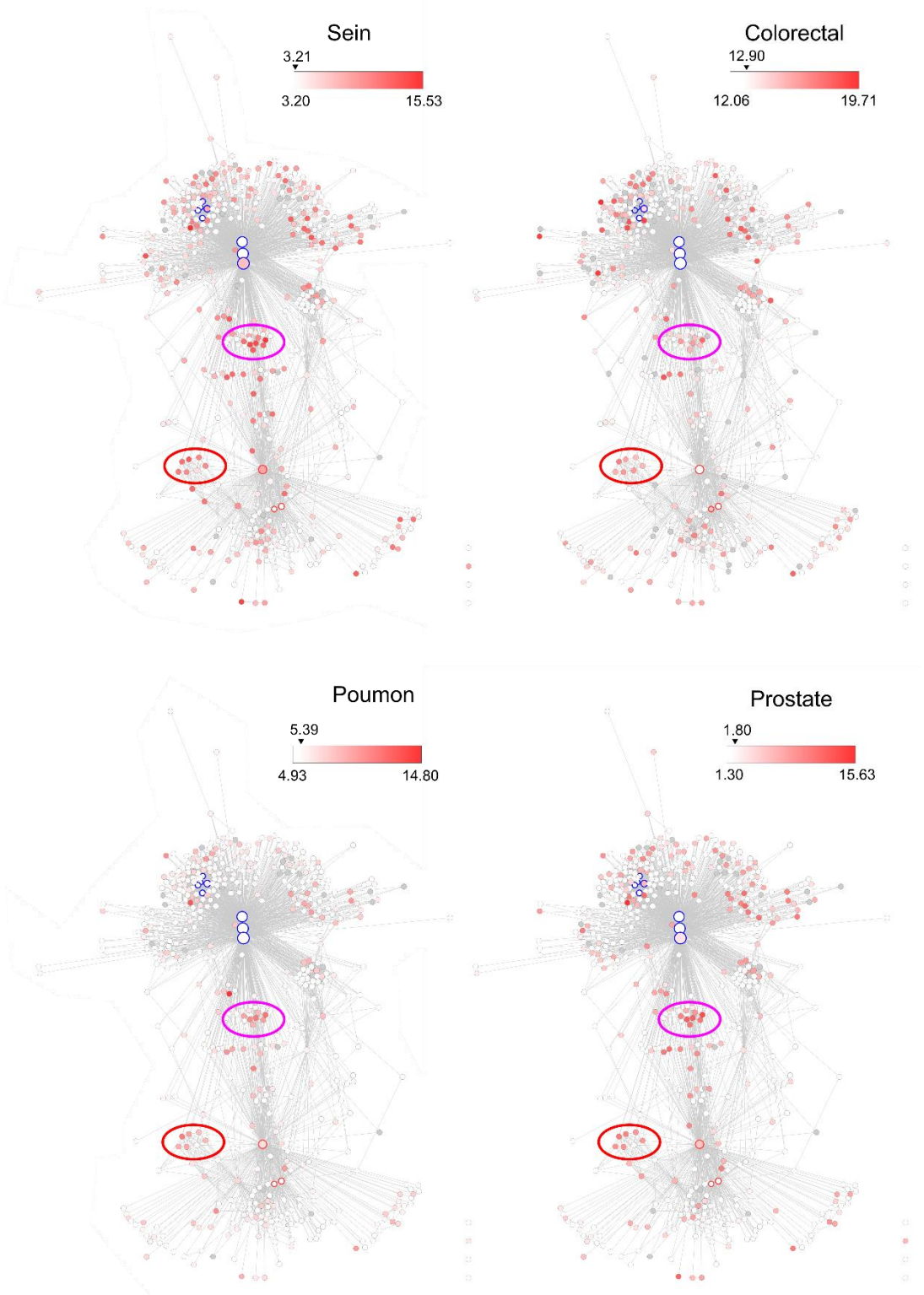


Figure 37. Expression des miARN dans différents tissus. Sont représentés des tissus sains de sein (GSE45666), de la muqueuse colorectale (GSE38389), des poumons (GSE25508) et de la prostate (GSE34933). Une coloration linéaire du blanc vers le rouge est utilisée pour visualiser l'expression des miARN. Cette échelle débute à la médiane d'expression des miARN (flèches) dans chaque ensemble de données et s'étend jusqu'au maximum observé. La boite complète s'étend de la valeur d'expression minimal observé jusqu'à la valeur maximale. Les miARN non testés sont représentés en gris. Les miARN fortement exprimés sont donc indiqué en rouge, ceux faiblement exprimés sont blanc.

La classification hiérarchique des tissus cités ci-dessus et d'autres tissus en se basant sur l'expression des miARN permet de mettre en évidence une claire séparation entre les tissus épithélioglandulaires et les autres tissus (cérébraux et musculaires) (Figure 38). Les échantillons de sang du sein et de la prostate se classifient également avec les tissus épithélioglandulaires mais se retrouvent en fait légèrement à l'écart de ces derniers. Les tissus rectaux semblent bien plus dissimilaires des autres tissus glandulaires même s'ils se retrouvent globalement dans le même sous-groupe. De façon similaire, l'expression des miARN permet de classifier les tissus cérébraux ensemble d'une part et d'autre part les tissus

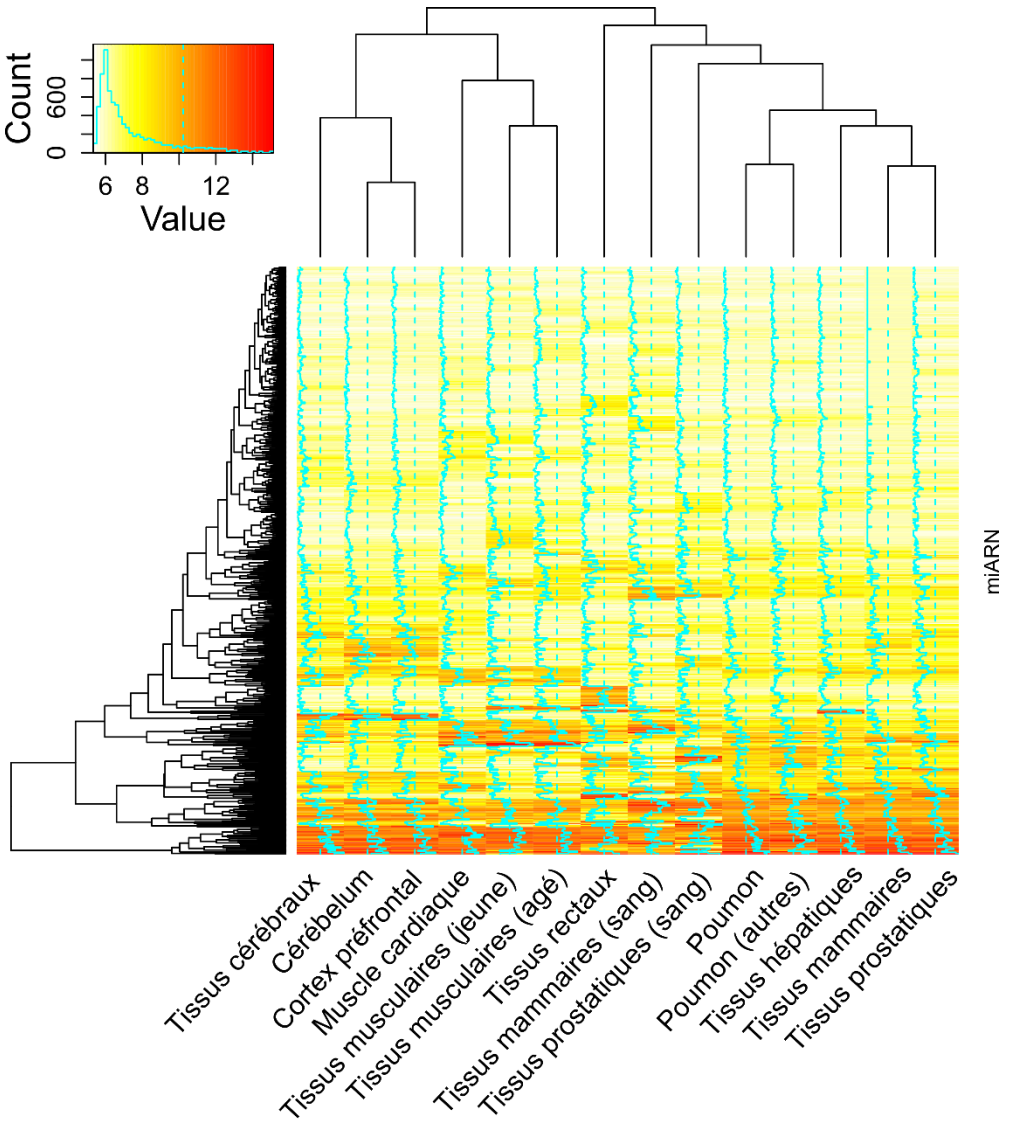


Figure 38. Classification hiérarchique de différents tissus sains en fonction de l'expression des miARN. Afin de pouvoir comparer les tissus et les expériences entre elles, une normalisation au quantile a été utilisée entre les différents tissus. En rouge sont signalés les miARN très exprimés et en blanc ceux très peu exprimés. Le jaune donne une indication sur les miARN modérément exprimés.

musculaires (cardiaques ou striés squelettiques) (Figure 38). Il semblerait que seuls 50 à 100 miARN par tissus soient très fortement exprimés.

C. TargetScan : robustesse de la construction et de l'analyse

Comme nous l'avons déjà vu, la qualité et le recouvrement peuvent différer sensiblement entre différents algorithmes de prédiction. Nous avons ainsi déjà pu constater que seules 60% de cibles prédites en moyenne sont communes pour les miARN entre TargetScan et DIANA-microT. Une question très importante dans la construction des réseaux était donc la robustesse de la construction vis-à-vis des algorithmes de prédiction.

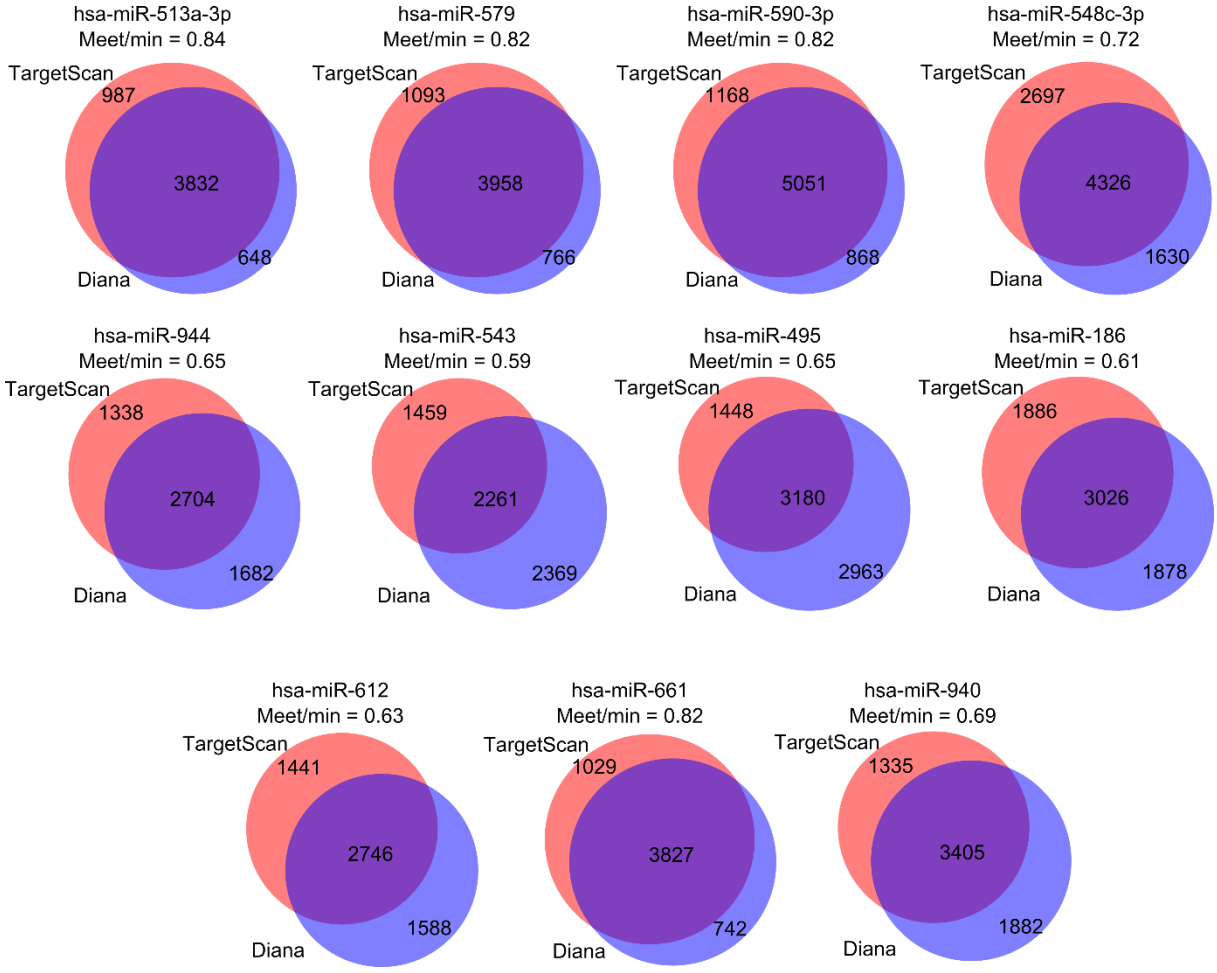


Figure 39. Diagrammes de Venn des recouvrements entre les cibles des membres des deux clubs assortis de DIANA-microT. En bleu sont représentées les prédictions de DIANA-microT et en rouge, celles de TargetScan. Pour évaluer le recouvrement, c'est également l'indice meet/min qui a été utilisée. En moyenne, le recouvrement est de 71%.

1. Recouvrement de cibles prédites pour les clubs assortis entre DIANA-microT et TargetScan

Le recouvrement de cibles prédites entre les deux algorithmes est assez faible (aux alentours de 60%, cf. page 84). L'analyse du recouvrement des membres des clubs assortis montre un taux légèrement supérieur mais toujours inférieur à 85%. Le miARN montrant le plus de recouvrement est le miR-513a-3p avec une valeur de 84% alors que celui montrant le moins de recouvrement est miR-543 avec une valeur de 59% (Figure 39).

2. Comparaison des propriétés et de la topologie des réseaux

Malgré les fortes différences de cibles prédites observées entre les deux algorithmes, autant en termes de prédiction que de nombre de miARN, les propriétés globales des réseaux construits selon le même processus montrent de flagrantes similarités (Figure 40). Ainsi, le maximum de différences entre le minimum de nœuds isolés et le minimum de liens se situe également aux alentours de 50% pour TargetScan. Pour être précis, ce dernier se trouve ici à 54%. A ce seuil, le réseau montre également des caractéristiques de petit monde et est modulaire et disassortatif – les propriétés (centralité, coefficient de *clustering* etc.) des réseaux TargetScan et DIANA microT étant quasiment superposables (Figure 40 A et B contre C).

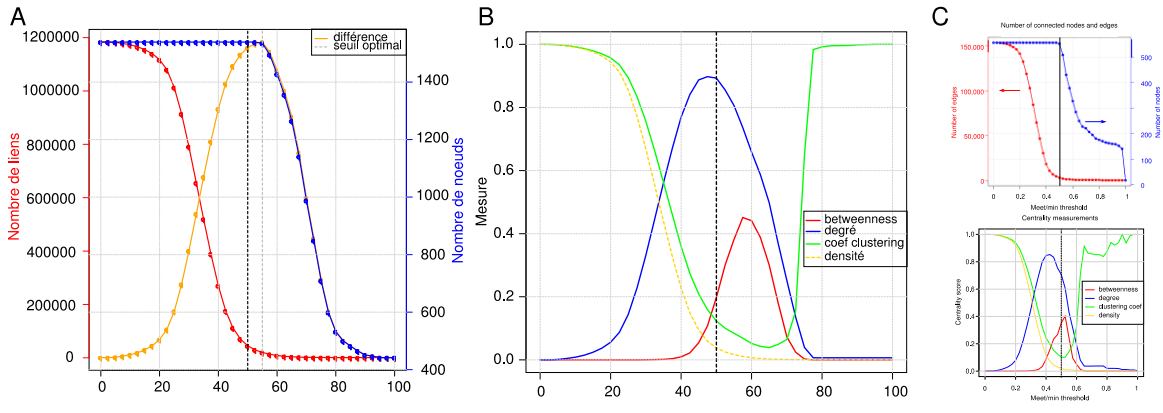


Figure 40. Caractéristiques des réseaux TargetScan pour différents seuils meet/min. A | Nombre de liens et nombre de nœuds non isolés en fonction de seuils meet/min. B | Centralité *betweenness* et degré, coefficient de *clustering* et densité des réseaux en fonction de seuils meet/min. C | Propriétés de réseaux pour DIANA-microT. Le seuil meet/min s'étend de 0 à 100% avec incrément de 1.

Le réseau prend également une forme bipolaire, organisé autour de quelques hubs centraux aux deux pôles (Figure 41), dont certains membres des deux clubs assortis. Une

différence majeure réside tout de même dans les nœuds centraux retrouvés entre les deux algorithmes et notamment les hubs qui ne sont plus exactement les mêmes : les clubs assortis de DIANA-microT ne sont en effet plus vraiment des clubs assortis à proprement parler sur ce nouveau réseau. Si la plupart des hubs diffèrent bien d'un algorithme à l'autre, les nœuds à fortes centralité *betweenness* restent très similaires puisque cinq nœuds sur sept avec la plus forte centralité *betweenness* découverts sur DIANA-microT sont également repérés parmi les trente nœuds à plus forte centralité *betweenness* sur TargetScan. Ces derniers sont en fait des nœuds bien spécifiques formant des liens entre les deux parties du réseau et sont composés de : miR-548c-3p, -590-3p, -661, -186 et -940, par ordre décroissant de valeur de centralité. Sur les onze hubs de DIANA-microT formant les deux clubs assortis, sept sont retrouvés parmi les quarante nœuds à plus fort degré sur TargetScan malgré un nombre de miARN trois fois supérieur dans cette dernière.

3. Comparaison des hubs, des nœuds centraux et des liens entre miARN

De plus, en reportant le code couleur de la Figure 36 sur le réseau TargetScan à un seuil meet/min 54, une structure assez similaire au réseau DIANA-microT peut être observée (Figure 41). Les clubs assortis retrouvés sur DIANA-microT restent de fait encore très centraux au réseau et définissent globalement les deux pôles, même si ce ne sont plus les plus gros hubs. Les zones d'influence des deux clusters sont également sensiblement similaires avec les zones cyan et rose globalement bien séparées. En revanche, la zone intermédiaire est bien moins visible comme zone de séparation des deux pôles, bien qu'une partie mineure de ces miARN soit tout de même placée entre les deux zones d'influence.

Les onze hubs identifiés avec DIANA-microT ont également été comparés entre les deux algorithmes. De façon intéressante, les deux clubs sont connectés avec des profils de densité très similaires entre les deux réseaux (Figure 42 A et B). Dans les deux cas, la densité des sous-graphes induits par les membres des clubs assortis de DIANA-microT décroît aux alentours de meet/min 50 – 60. Ces deux constats prouvent que les connections entre les miARN restent identiques malgré le changement d'algorithme. La différence majeure dans ce cas est la connexion entre les deux clusters au travers du miARN miR-548c-3p (Figure 42 C) formant un bloc uni, contrairement au réseau DIANA-microT.

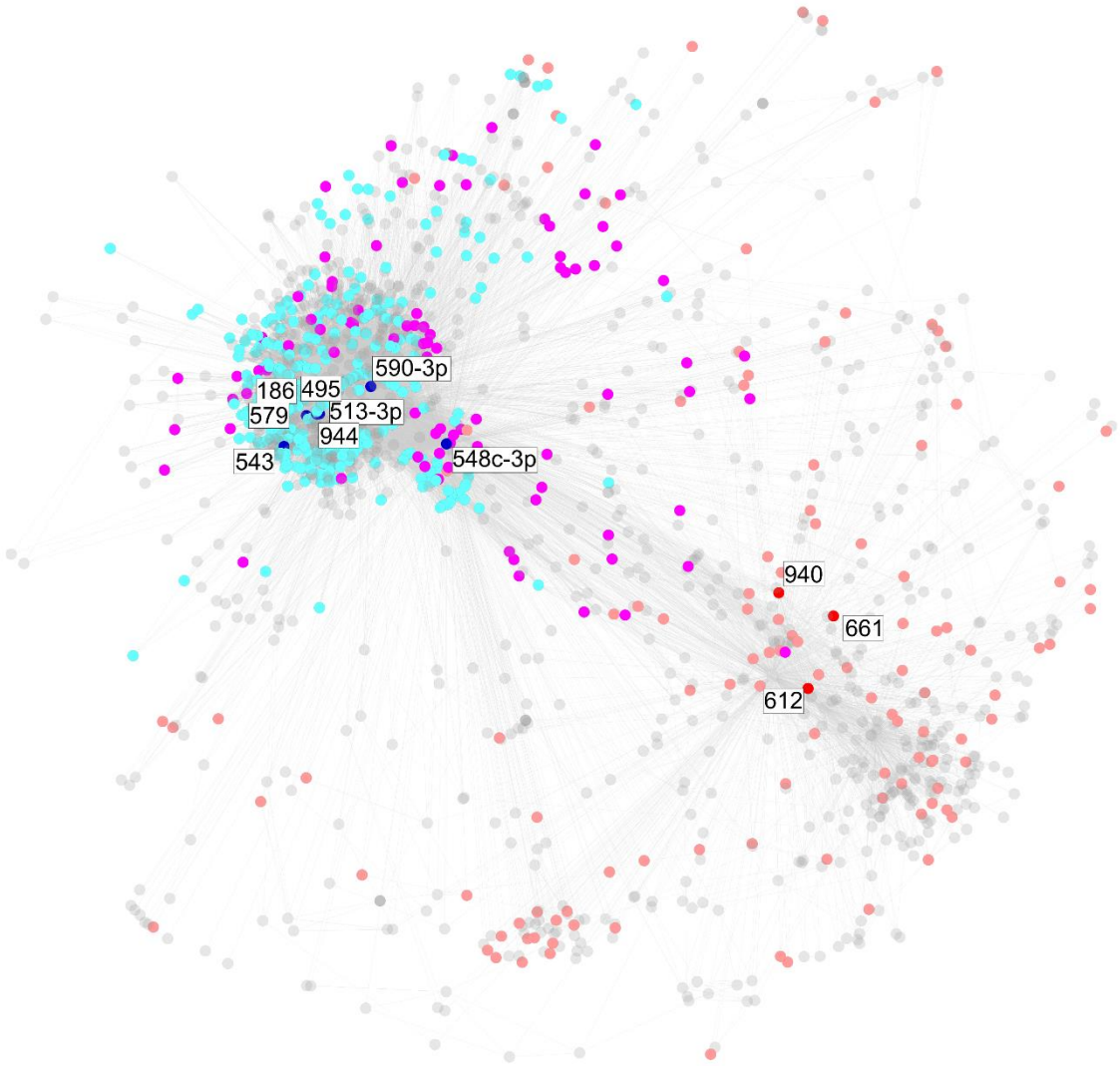


Figure 41. Le réseau de miARN de TargetScan. Le code couleur représenté ici est celui de la Figure 36 : en bleu sont représentés les membres du club assorti 1, en rouge ceux du club assorti 2. En cyan, rose et violet sont respectivement représentées les zones d'influence des clubs 1 et 2 et la zone intermédiaire. Les nœuds non présents dans DIANA-microT sont colorés en gris.

En conclusion de cette sous-partie dédiée à la robustesse du processus de construction et d'analyse, nous pouvons dire que malgré toutes les différences observées pour ce changement d'algorithme de prédiction de cibles, les relations en termes de partage de cibles et donc de connexions entre les miARN restent très similaires d'un algorithme à l'autre, tout comme la structure des réseaux. La construction basée sur le partage de cibles et l'analyse qui s'en suit montrent donc une certaine robustesse vis-à-vis des algorithmes de

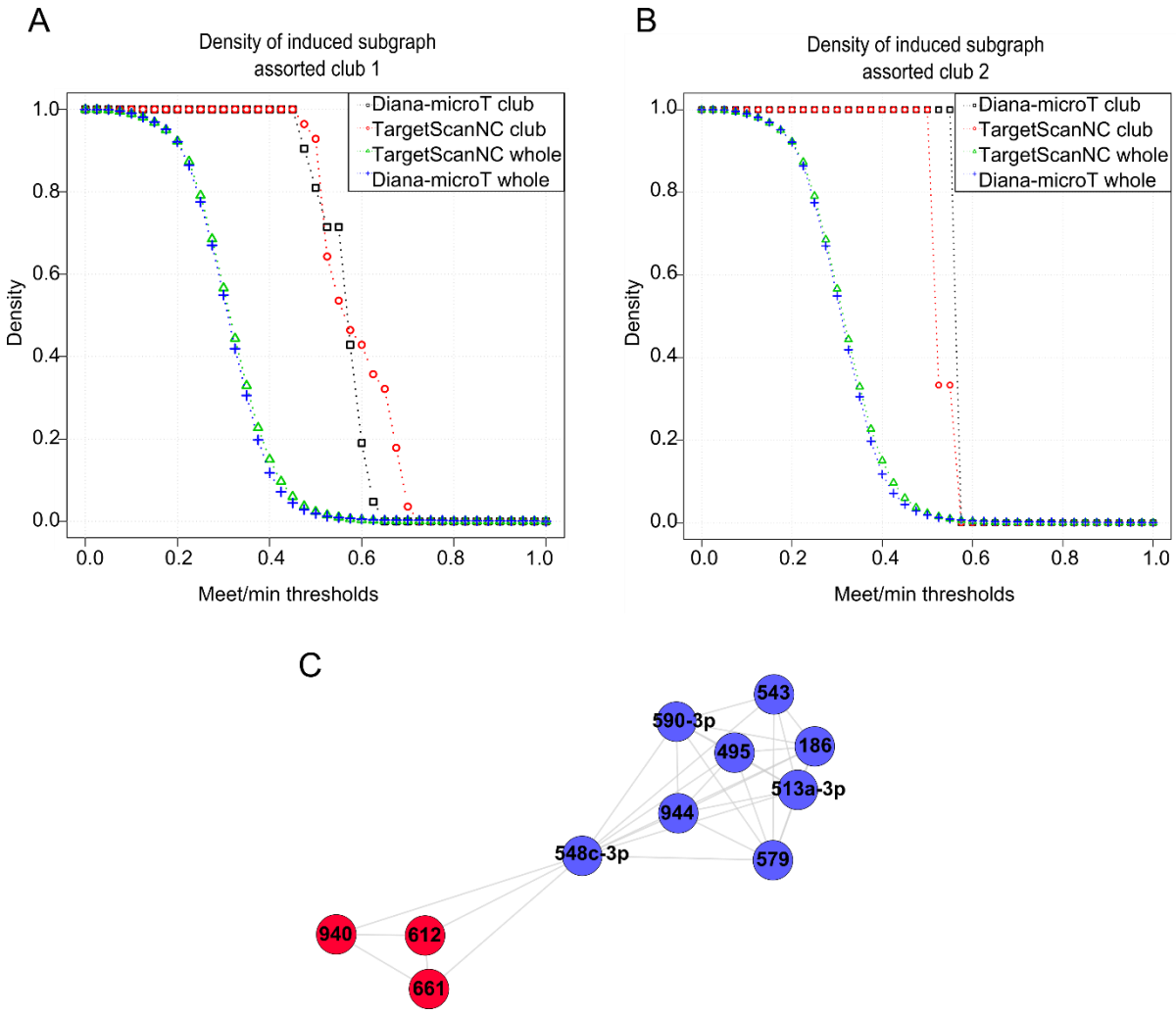


Figure 42. Comparaison des sous-graphes induits et des profils de densité des deux clubs assortis de DIANA-microT entre les deux algorithmes. A | Densité du club assorti 1 en fonction de différents seuils meet/min. B | Densité du club assorti 2 en fonction de différents seuils meet/min. C | Sous-graphe induit par les clubs assortis 1 et 2 de DIANA-microT sur le réseau TargetScan à un seuil meet/min 54. La densité des sous-graphes formés par les clubs assortis 1 et 2 ont une densité de 0.93 et 1 respectivement dans le réseau TargetScan.

prédiction.

D. Réseaux de miARN et évolution ?

Afin de bénéficier des prédictions dans TargetScan pour d'autres espèces et de comparer des réseaux de miARN entre différentes espèces (le taureau, le chien, le cheval, la poule, l'homme, le macaque, l'opossum, la souris, l'ornithorynque, le chimpanzé, le rat et enfin une espèce de crapaud), des réseaux indépendants pour chaque espèce présente dans TargetScan ont été construits de la même manière qu'avec les réseaux de DIANA-microT et de TargetScan chez l'être humain.

En termes de nombre de miARN et de prédictions (cf. Tableau 2), TargetScan montre 1539 miARN prédits pour cibler 18 370 gènes avec 4 305 160 interactions miARN-gène chez l'homme, suivi par la souris avec 779 miARN et 16 961 gènes pour 1 666 429 interactions. Chez le chimpanzé, il n'existe que 524 miARN et 17888 gènes pour 1 418 354 interactions et l'opossum ne possède que 144 miARN, 12 709 gènes qui seraient des cibles prédites et 233 428 interactions et se classe donc comme l'espèce avec le plus faible nombre de miARN et de prédictions.

Il est difficile en réalité de savoir si ces différents nombres sont réellement associés à des différences évolutionnaires entre les espèces ou simplement dus à un manque d'information et/ou un biais sur les espèces les plus étudiées. En effet, nous pourrions supposer que les nombres entre l'homme et le singe soient plus proches qu'entre la souris et l'homme par exemple, à cause de leur rapprochement phylogénique, ce qui n'est pas le cas ici. Malgré ce doute, nous pouvons remarquer que le nombre d'interactions observées est linéairement corrélé avec le nombre de miARN d'une espèce (coefficient de corrélation = 0,97 et $R^2 = 0.94$: Figure 43). La médiane du nombre de gènes prédits régulés par un miARN est en moyenne de 1937 pour les douze espèces. Chez l'homme, les miARN régulent en moyenne 2796 gènes prédits et le crapaud est l'espèce où les miARN régulent généralement le moins de gènes avec une médiane de 265.

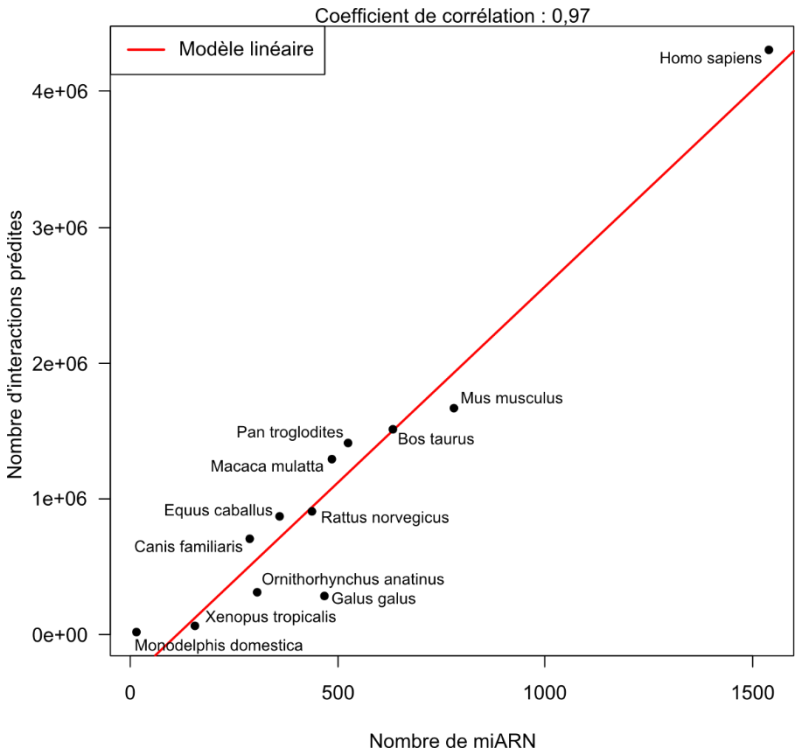


Figure 43. Nombre d'interactions contre nombre de miARN pour différentes espèces dans TargetScan v6.2.

La Figure 44 montre la première étape de ces constructions notamment l'étape d'analyse de seuil optimal pour chacune des douze espèces analysées. Le seuil optimal varie bien en fonction des espèces mais ce dernier semble plus ou moins corrélé à la phylogénie : p.ex. l'homme et le chimpanzé montrent des propriétés particulièrement similaires en fonction des seuils meet/min, tout comme la souris, le rat et l'opossum. Par ailleurs, mis à part le macaque et l'opossum, les seuils optimaux se situent généralement 40 et 60%. La forme des courbes des nœuds isolés est également particulièrement intéressante puisque chez l'être humain et le taureau, les courbes ne forment pas de palier aux alentours 80%, contrairement aux autres espèces. Ce dernier constat est probablement dû à la présence de grands clusters de miARN interconnectés et partageant un grand nombre de cibles. C'est notamment le cas chez l'opossum. Dans tous les cas, quelques miARN différents partageant exactement les mêmes cibles sont retrouvés (meet/min 100%). Ces groupes ne sont pas exclusivement des familles de miARN (let-7 p.ex.) même si ces dernières forment une grande partie des clusters les plus fortement connectés. En effet, l'indice meet/min prenant en compte le minimum du nombre de gènes régulés comme dénominateur, les miARN avec peu de gènes prédits

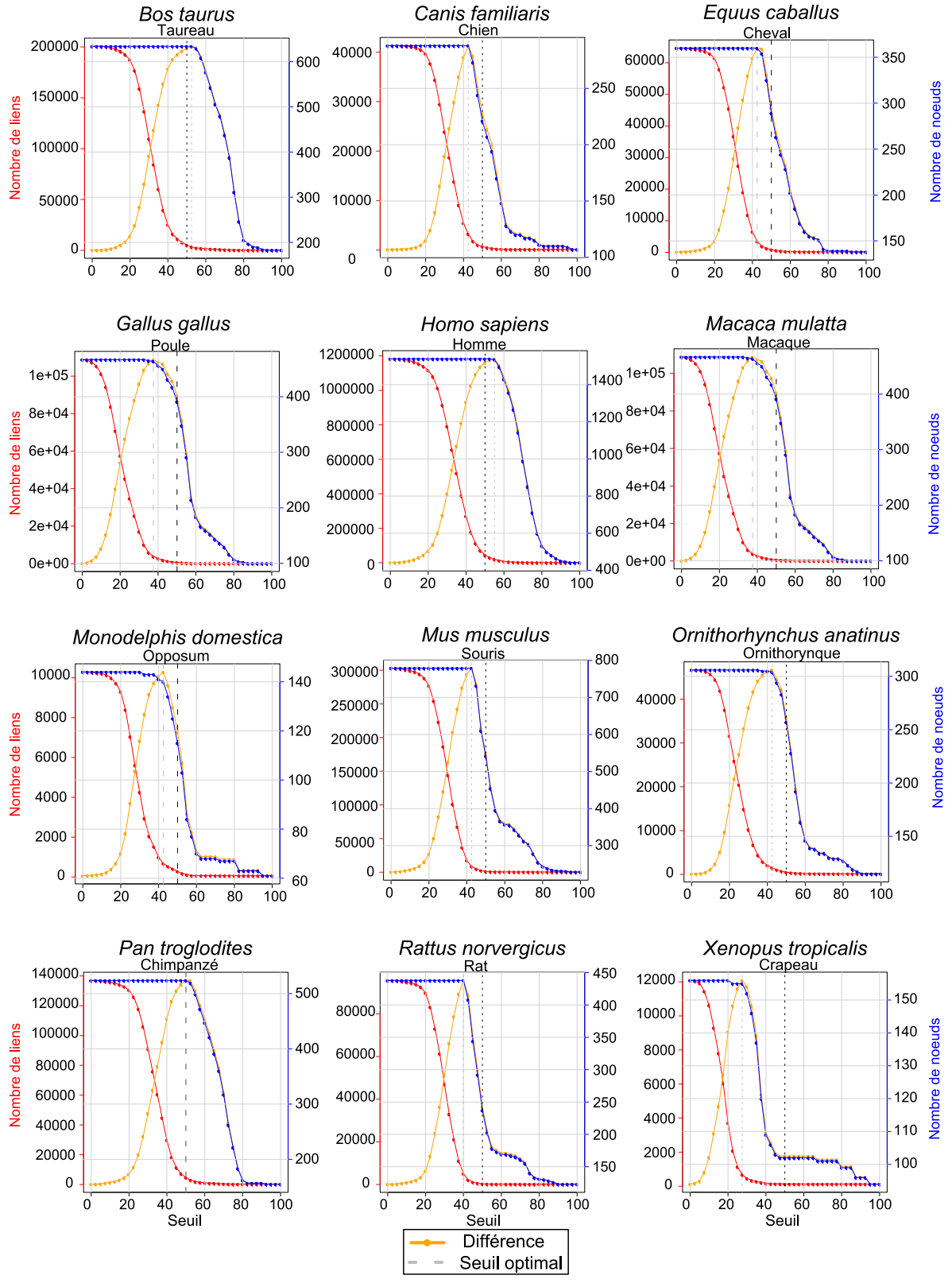


Figure 44. Nombre de liens et nombre de nœuds connectés au réseau pour différentes espèces. Prédictions tirées de TargetScan v6.2. Les courbes rouges représentent le nombre de liens. Les courbes bleues représentent le nombre de nœuds non isolés. Les courbes jaunes montrent la différence entre le nombre de liens et le nombre de nœuds connectés (normalisés). Lorsque la différence est grande, nous sommes en présence d'un graphe avec beaucoup de miARN toujours connectés entre eux mais peu de liens. La barre grise représente le seuil optimal (le maximum de différence entre nombre de nœuds connectés et nombre de liens) et la barre noire un seuil de 50%. Seuil meet/min de 0 à 100 avec incrément de 1.

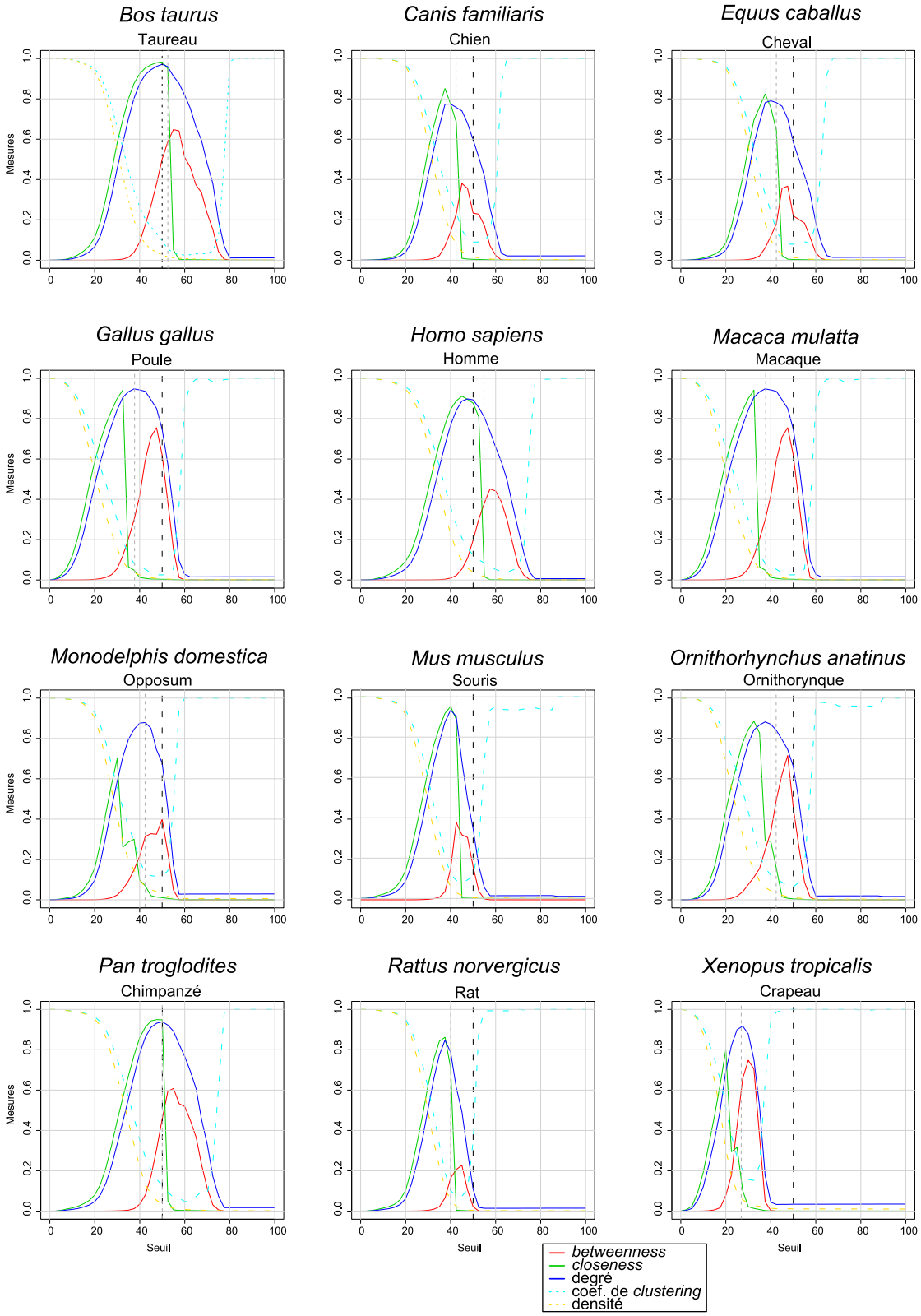


Figure 45. Propriétés de réseaux pour différentes espèces et différents seuils meet/min. Prédictions tirées de TargetScan v6.2. Les paramètres représentés sont les centralités *betweenness*, *closeness* et degré respectivement en rouge, vert et bleu. Sont également représentés le coefficient de *clustering* ainsi que la densité des graphes en fonction de seuil meet/min (de 0 à 100 avec incrément de 1).

peuvent influencer les groupes qui sont toujours présents à des forts meet/min et se lier notamment (avec plus de probabilité) aux miARN avec beaucoup de gènes prédits.

Les propriétés des réseaux à différents seuils meet/min sont représentées sur la Figure 45 où l'on peut globalement tirer des conclusions similaires aux précédentes. En effet, l'ensemble des propriétés montrent de fortes similarités lorsque l'on considère la phylogénie des espèces. Cinq groupes peuvent grossièrement être distingués en fonction des propriétés : i) le chien, le cheval et le macaque ; ii) l'opossum, l'ornithorynque et la poule ; iii) le rat et la souris ; iv) l'homme, le chimpanzé et le taureau et enfin v) le crapaud qui est isolé des autres espèces surtout à cause des patterns bien spécifiques. Ces patterns sont probablement dus à un nombre de miARN, de gènes et de prédictions bien plus faibles que chez les autres espèces. Dans tous les cas, les formes des différentes courbes sont fortement similaires d'une espèce à l'autre. Nous pouvons donc supposer que tous ces réseaux (à leur seuil optimal respectif) sont – tout comme ce qui a déjà été évoqué jusqu'à présent dans la thèse – structurés autour de quelques nœuds très centraux dominant le reste des miARN.

E. Conclusions et discussion

Les miARN jouent des rôles cruciaux dans nos cellules en régulant divers processus biologiques de façon coordonnée et en ciblant différents gènes qui peuvent appartenir à des voies biologiques différentes ou similaires. La biologie des systèmes semble donc être un moyen de choix pour analyser et comprendre ce rôle complémentaire. Grâce à la méthodologie d'inférence basée sur le partage de cible ainsi qu'à l'analyse de réseau présentées dans ce chapitre, nous avons pu mettre en évidence des structures de réseaux spécifiques, non liées au hasard et pouvant potentiellement être associées à des rôles biologiques.

La notion de co-ciblage n'est pourtant pas peu étudiée puisque Tsang et collaborateurs ont par exemple déjà évoqué la prévalence de ce type de régulation en s'intéressant à certains aspects de familles de miARN (Tsang et al., 2010). Ils ont ainsi émis l'hypothèse que si

différents miARN visent les mêmes cibles, un plus grand potentiel de régulation pourrait émerger. En réalité, les processus biologiques étant représentés non pas par un seul et unique gène mais par tout un réseau intriqué de gènes, la corégulation des processus biologiques peut également passer par la régulation de différents gènes appartenant au même processus. Une question que nous adressons partiellement au travers de notre méthodologie mais sans spécifiquement s'y attarder. Xu et collaborateur se sont spécialement intéressés à ce problème en basant leur inférence non pas uniquement sur le partage de cibles mais également sur la similarité des termes GO des gènes (Xu et al., 2011a). En revanche, dans leur cas, l'ontologie est plutôt utilisée comme méthode de validation des clusters alors que nous cherchons à l'utiliser comme méthode d'inférence des fonctions cellulaires potentiellement corégulées.

A cause du pourcentage de faux positifs des prédictions *in silico* – évalué à environ 66% (Maragkakis et al., 2009b) – certains auteurs se basent uniquement sur des prédictions validées (Alshalalfa et al., 2013). Même si ces prédictions sont généralement plus robustes (certaines montrent tout de même un certain taux de faux positifs et de faux négatifs pour les cibles directes des miARNs), l'information est encore limitée. Par exemple, au moment de l'écriture du premier manuscrit (Bhajun et al., 2015), seule une cible était validée pour miR-612 dans TarBase (Sethupathy et al., 2006) et aucune n'était retrouvée dans miRecords (Xiao et al., 2009). Cette limitation rend encore aujourd'hui l'utilisation des données de prédiction incontournable pour des approches systémiques. Un autre point crucial dans la méthodologie d'inférence est le choix du seuil qui est très souvent subjectif. Dans notre cas, nous avons compensé ce choix arbitraire en analysant statistiquement et de manière exploratoire l'ensemble des graphes et des informations disponibles. Ainsi, s'il existe bien de fortes différences à un niveau individuel des miARN, les propriétés d'ensemble des réseaux sont robustes.

Après cette première étape de construction et d'analyse, nous aborderons l'interprétation biologique des clubs assortis, le rôle co-régulateur potentiel des onze miARN composant les deux clusters ainsi que leur rôle d'influence sur le reste du réseau.

Chapitre 2 :
Analyse des clubs assortis

A. Introduction

Les clubs assortis sont définis comme des groupes de hubs ayant une forte influence sur le reste des membres du réseau. Dans notre cas, nous avons mis en évidence deux clubs assortis formés respectivement par huit et trois miARN (Tableau 5). Etant donné la forte connectivité qui existe au sein de ces deux clubs, nous pouvons supposer qu'il existe une certaine probabilité que les miARN composant ces deux groupes co-régulent certaines fonctions cellulaires. L'objectif de ce chapitre est d'exposer les prédictions de fonctions des gènes ciblés pour chacun des deux clubs et également d'aborder les validations biologiques ayant permis de valider nos prédictions pour le club assorti 2.

Tableau 5. Composition des miARN des clubs assortis et information du nombre de cibles.

Club assorti 1				Club assorti 2			
Nom	**Degré**	**Rang**	**Cibles**	**Nom**	**Degré**	**Rang**	**Cibles**
miR-495	402	1	6 626	miR-940	219	4	5 848
miR-548c-3p	369	2	6 364	miR-661	84	6	5 094
miR-590-3p	332	3	6 468	miR-612	50	11	4 819
miR-186	119	5	5 285				
miR-579	67	7	5 129				
miR-513a-3p	63	8	4 795				
miR-543	60	9	4 934				
miR-944	53	10	4,722				

Comme exposé dans le chapitre précédent, les deux clubs induisent la topologie bipolaire du réseau de miARN. Une autre question que nous aborderons donc dans cette partie est l'influence en termes de fonctions cellulaires qu'ont les deux clubs sur le reste du réseau ainsi que le biais introduit par les gènes hubs – des gènes beaucoup ciblés par les miARN. Dans un dernier temps, nous verrons les différences sur les prédictions de fonctions cellulaires induites par le changement d'algorithme autant sur les clubs assortis que sur leur zone d'influence et la robustesse de l'analyse d'ontologie face aux faux positifs et aux faux négatifs.

B. Club assorti 1

1. Description du groupe

Le club assorti 1 est constitué de huit miARN (Tableau 5), miR-495 étant le plus gros hub du réseau. Ce dernier est connecté à 72% des 555 miARN composant le réseau DIANA-microT et est prédit comme ciblant 6 626 cibles différentes (13 900 sites de liaison). En moyenne, les membres de ce cluster sont prédits comme ciblant approximativement 5 000 gènes. 5 276 gènes sont ciblés par au moins quatre des huit miARN du cluster (*i.e.* partagé par au moins 50% des miARN du club). Seuls 540 gènes sont ciblés par l'ensemble des huit miARN. MiR-944, quant à lui, est le plus petit des hubs et est connecté à « seulement » 10% des 555 miARNs.

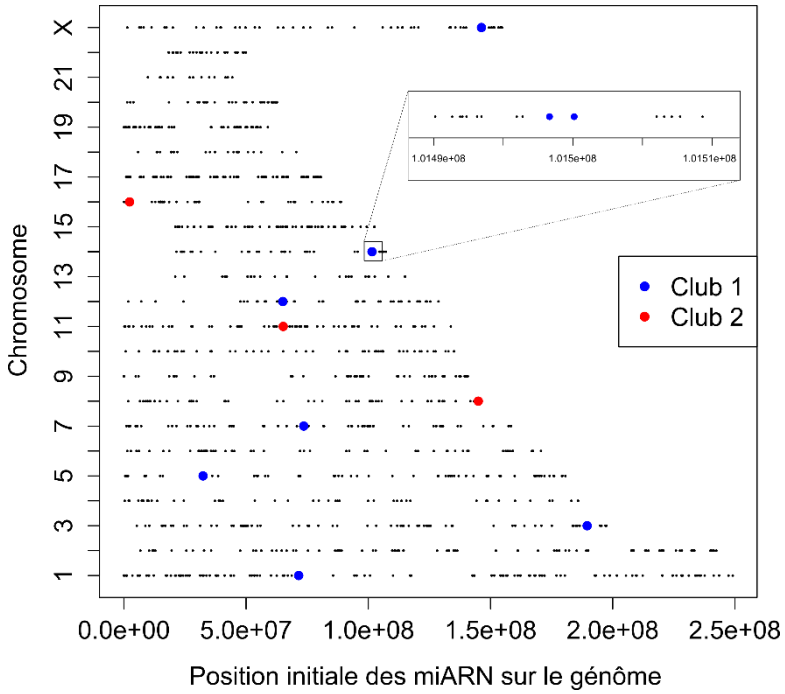


Figure 46. Emplacement génomique des miARN dans le génome humain. Chaque point désigne l'emplacement chromosomique en paires de bases d'un miARN chez l'être humain (est indiqué en fait la position la plus en 5' de l'ensemble des miARN de miRBase). En bleu sont signalés les huit miARN du club assorti 1 et en rouge, les trois miARN du club assorti 2. Seuls deux miARNs sont retrouvés en cluster sur le génome

Une particularité de la transcription des miARN, évoquée dans l'introduction, est leur propension à être transcrits ensemble sous forme de polycistron. Dans notre cas, seuls les deux miR-495 et miR-543 sont retrouvés proches dans le génome. Ils sont tous les deux

localisés sur le chromosome 14 et séparés par 1 500 paires de bases (Figure 46). Nous pouvons constater que le reste des miARN est plutôt bien réparti sur les autres chromosomes.

Les miARN des clubs assortis ne sont probablement pas transcrits sous forme de polycistron. En revanche, le statut de club assorti pourrait potentiellement être associé à une co-transcription si les miARN se retrouvaient proches les uns des autres dans un même territoire au sein du noyau. Les données Hi-C (extension de la technologie 3C pour *chromosome conformation capture* (van Berkum et al., 2010)) sont des données permettant d'obtenir ce type d'information, grâce à l'utilisation de conditions permettant la formation de liaisons covalentes (*cross-links*) entre des portions d'ADN et des protéines. La Figure 47 montre les interactions entre des portions de 20 000 paires de bases portant l'ensemble des miARNs présent dans la version 21 de miRBase. Les données montrent peu d'interactions interchromosomiques entre les miARN (nombre d'interaction généralement inférieur à 10). Il

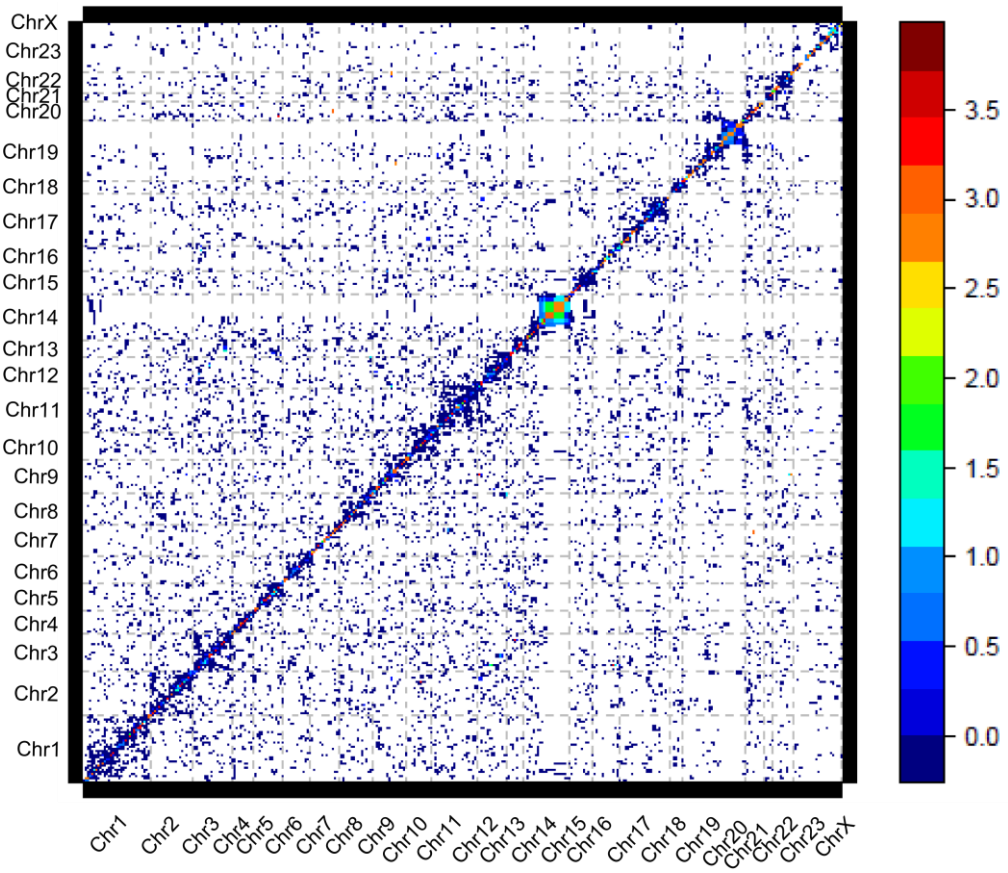


Figure 47. Données d'interaction chromosomique pour l'ensemble des miARNs de miRBase sur les 23 paires de chromosomes. Sont représentés les log10 du nombre d'interaction entre les portions d'ADN portant les miARN. D'après l'ensemble de données GSE35156 avec un *binning* de 20 000 (Dixon et al., 2012).

existe en revanche beaucoup d'interactions intrachromosomiques, c'est typiquement le cas du chromosome 14 où l'on peut observer de très nombreuses interactions. Une étude plus spécifique des membres des clubs assortis ne montre toutefois aucune interaction significative entre leur portion respective.

Comme ce club est constitué des hubs avec les plus forts degrés, une faible spécificité de régulation de fonctions pourrait être attendue, à cause de la corrélation entre le nombre de cibles et le nombre de liens. En effet, nous pouvons supposer que plus un miARN (ou un groupe de miARN) régule de cibles, moins ce miARN a de spécificité. Il agirait en fait non pas sur une fonction particulière mais sur plusieurs fonctions de manière aspécifique. C'est typiquement le cas des miARN hubs qui visent et partagent beaucoup de cibles prédites. Chaque autre groupe – relié à ces clusters – amènerait donc de la spécificité mais également de la redondance fonctionnelle. De fait, sur les 540 gènes cibles partagées par les huit membres du club, 17% sont en fait des gènes hubs – les gènes les plus ciblés par les miARNs. Ce pourcentage représente presque la moitié des gènes hubs de l'ensemble des données (Figure 48 A - p-valeur de Fisher = 10^{-90}).

2. Prédiction de processus biologiques

Pour prédire quelles fonctions biologiques pourraient être corégulées par les miARN de ce cluster, une étude d'enrichissement d'ontologie GO (Ashburner et al., 2000) a été menée sur les gènes partagés par le club 1. Afin de ne pas être trop restrictif, ce sont tous les gènes ciblés par au moins 50% des membres du groupes qui ont été utilisés (c'est-à-dire, tous les gènes retrouvés ciblés par au moins quatre miARN différents dans le club assorti 1).

Le Tableau 6 récapitule les résultats de cette analyse pour les trois branches principales de GO (BP – processus biologique, MF – fonction moléculaire et CC – compartimentation cellulaire). Un clair enrichissement pour la régulation des ARNm, la transcription et l'expression génique est retrouvé dans le tableau avec des p-valeurs entre 10^{-5} et 10^{-8}. En s'intéressant aux annotations moins significatives, des termes associés à la

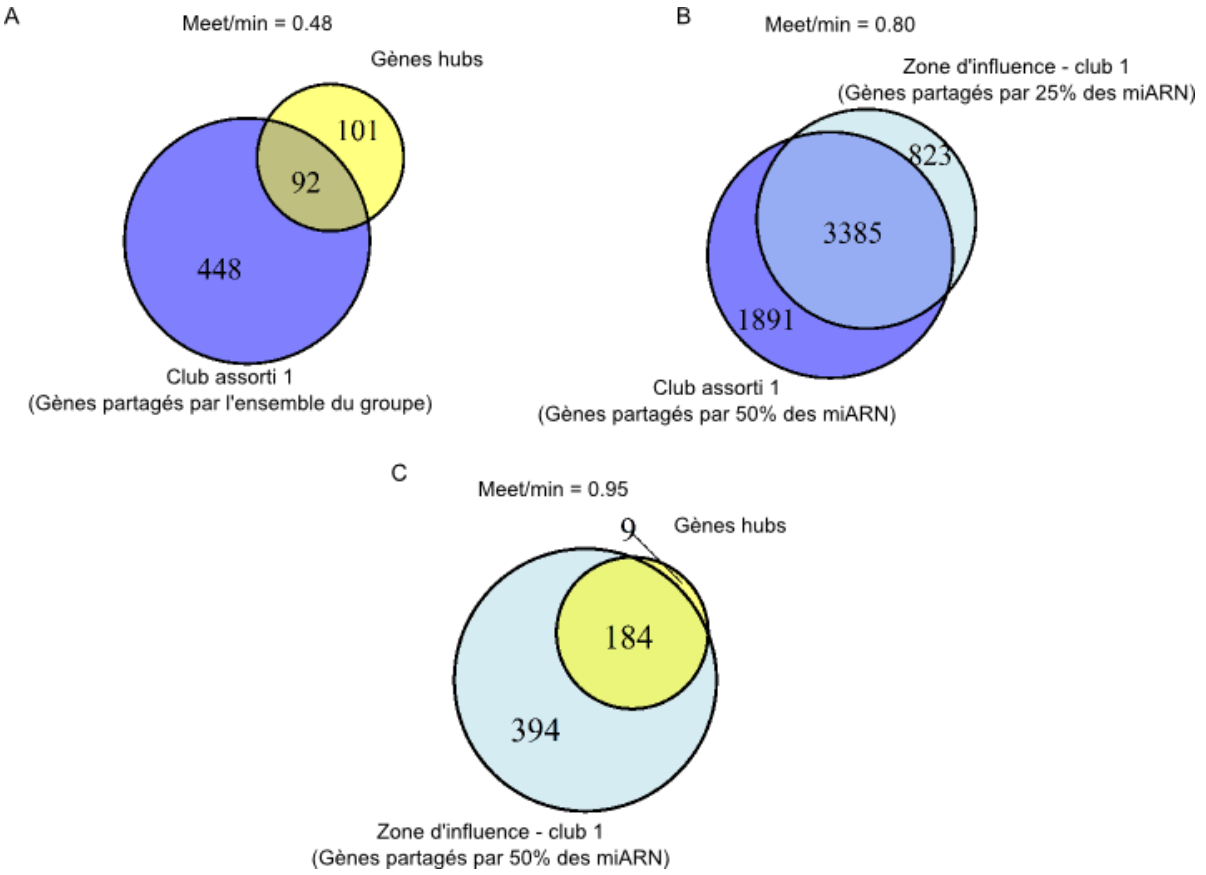


Figure 48. Diagramme de Venn de recouvrement de cibles entre les gènes hubs, le club assorti 1 et sa zone d'influence. A | Recouvrement de cibles entre les gènes hubs et les gènes partagés par l'ensemble des miARN du club assorti 1. B | Recouvrement de cibles entre les gènes partagés par 50% des membres du club assorti 1 (4/8) et 25% des membres de sa zone d'influence (79/315). C | Recouvrement de cibles entre les gènes partagés par 50% des membres de la zone d'influence du club assorti 1 (158/315) et les gènes hubs.

modification de protéines (p-valeur corrigée = 3 x 10^{-4}), au transport endosomal (p-valeur corrigée = 5 x 10^{-4}) et à la régulation de locomotion (p-valeur corrigée = 10^{-2}) peuvent être retrouvés. De manière corrélée, beaucoup de protéines ciblées par les miARN du club sont nucléaires (p-valeur corrigée = 1,6 x 10^{-7}). Enfin, au niveau MF, ce sont des termes associés à la liaison aux ions métalliques que nous pouvons apercevoir (p-valeur corrigée = 6,3 x 10^{-10}).

Tableau 6. Enrichissement GO pour les gènes ciblés par 50% des membres du club assortis 1 (BP, MF et CC). Soit 5 276 gènes, ciblés par au moins 4 miARNs sur les 8 du club. Seules les dix premières annotations sont présentées.

BP					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:2000112	regulation of cellular macromolecule biosynthetic process	2817	989	7.50E-12	3.08E-08
GO:0019219	regulation of nucleobase-containing compound metabolic process	3046	1058	2.10E-11	4.31E-08
GO:0031326	regulation of cellular biosynthetic process	2998	1040	5.00E-11	6.85E-08
GO:0010556	regulation of macromolecule biosynthetic process	2877	1000	8.50E-11	8.73E-08
GO:0051252	regulation of RNA metabolic process	2674	935	1.10E-10	9.04E-08
GO:0009889	regulation of biosynthetic process	3026	1045	1.50E-10	1.03E-07
GO:0051171	regulation of nitrogen compound metabolic process	3121	1074	1.90E-10	1.12E-07
GO:0006355	regulation of transcription, DNA-templated	2602	909	2.80E-10	1.44E-07

GO:0010468	regulation of gene expression	3004	1035	3.60E-10	1.64E-07
GO:2001141	regulation of RNA biosynthetic process	2618	912	5.10E-10	2.10E-07

MF

GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0046872	metal ion binding	3551	1222	1.70E-13	6.34E-10
GO:0008270	zinc ion binding	1780	657	5.10E-13	9.52E-10
GO:0043169	cation binding	3593	1224	4.70E-12	5.85E-09
GO:0043167	ion binding	3603	1226	6.50E-12	6.06E-09
GO:0046914	transition metal ion binding	2039	732	1.30E-11	9.70E-09
GO:0019787	small conjugating protein ligase activity	257	118	1.70E-08	1.06E-05
GO:0004842	ubiquitin-protein transferase activity	242	112	2.30E-08	1.23E-05
GO:0005488	binding	10826	3307	1.90E-07	8.86E-05
GO:0016881	acid-amino acid ligase activity	285	121	2.10E-06	0.000871
GO:0003676	nucleic acid binding	2959	975	3.40E-06	0.00127

CC

GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0005622	intracellular	11345	3491	2.80E-11	3.58E-08
GO:0005634	nucleus	5415	1759	3.40E-10	1.63E-07
GO:0044424	intracellular part	11054	3399	5.20E-10	1.63E-07
GO:0044464	cell part	12849	3894	5.90E-10	1.63E-07
GO:0005623	cell	12850	3894	6.40E-10	1.63E-07
GO:0043231	intracellular membrane-bounded organelle	8665	2697	3.40E-08	6.93E-06
GO:0043227	membrane-bounded organelle	8674	2699	3.80E-08	6.93E-06
GO:0043229	intracellular organelle	9583	2953	2.30E-07	3.67E-05
GO:0043226	organelle	9594	2955	2.90E-07	4.11E-05
GO:0000139	Golgi membrane	494	195	7.40E-07	9.45E-05

Bien que certaines de ces annotations soient très générales, les résultats montrent un fort enrichissement en termes ontologiques liés aux liaisons ARN/ADN et à la régulation des ARNm dénotant donc une implication probable du club dans la corégulation transcriptionnelle.

3. Revue de la littérature sur la corégulation potentielle par les miARN du club

Malheureusement, peu d'information sur le rôle coopératif des huit miARN peut être retrouvée dans la littérature et encore moins sur une corégulation de régulateurs transcriptionels. Néanmoins, miR-186 et miR-543 sont tous les deux cités dans différentes études sur le vieillissement cellulaire (Kim et al., 2012; Nidadavolu et al., 2013) démontrant la possibilité de leur coaction. miR-495 et miR-543 ont également été identifiés comme actant sur la transition épithélio-mésenchymateuse (Haga and Phinney, 2012). Par ailleurs, miR-495 est également connu pour son rôle dans la différenciation cellulaire et la prolifération (Simion et al., 2010; Chen et al., 2013; Prévot et al., 2013) et également pour son rôle de tumeur suppresseur (Jiang et al., 2012). miR-186 semblerait avoir un effet sur les récepteurs

purinergiques pro-apoptotique (Zhou et al., 2008) mais aucun rôle direct sur l'apoptose ne lui est reconnu. miR-590-3p semble impliqué dans la mort des neurones (Villa et al., 2011) alors que miR-513a-3p est impliqué dans la réponse immunitaire médiée par l'interféron gamma (IFN-γ) (Gong et al., 2009). Enfin, miR-548c-3p est impliqué dans les processus de réparation de l'ADN en agissant sur la traduction de la protéine TOP2A (Srikantan et al., 2011). En se basant sur ces informations, il est assez difficile de tirer des conclusions sur une éventuelle complémentarité régulationnelle entre les membres du club 1.

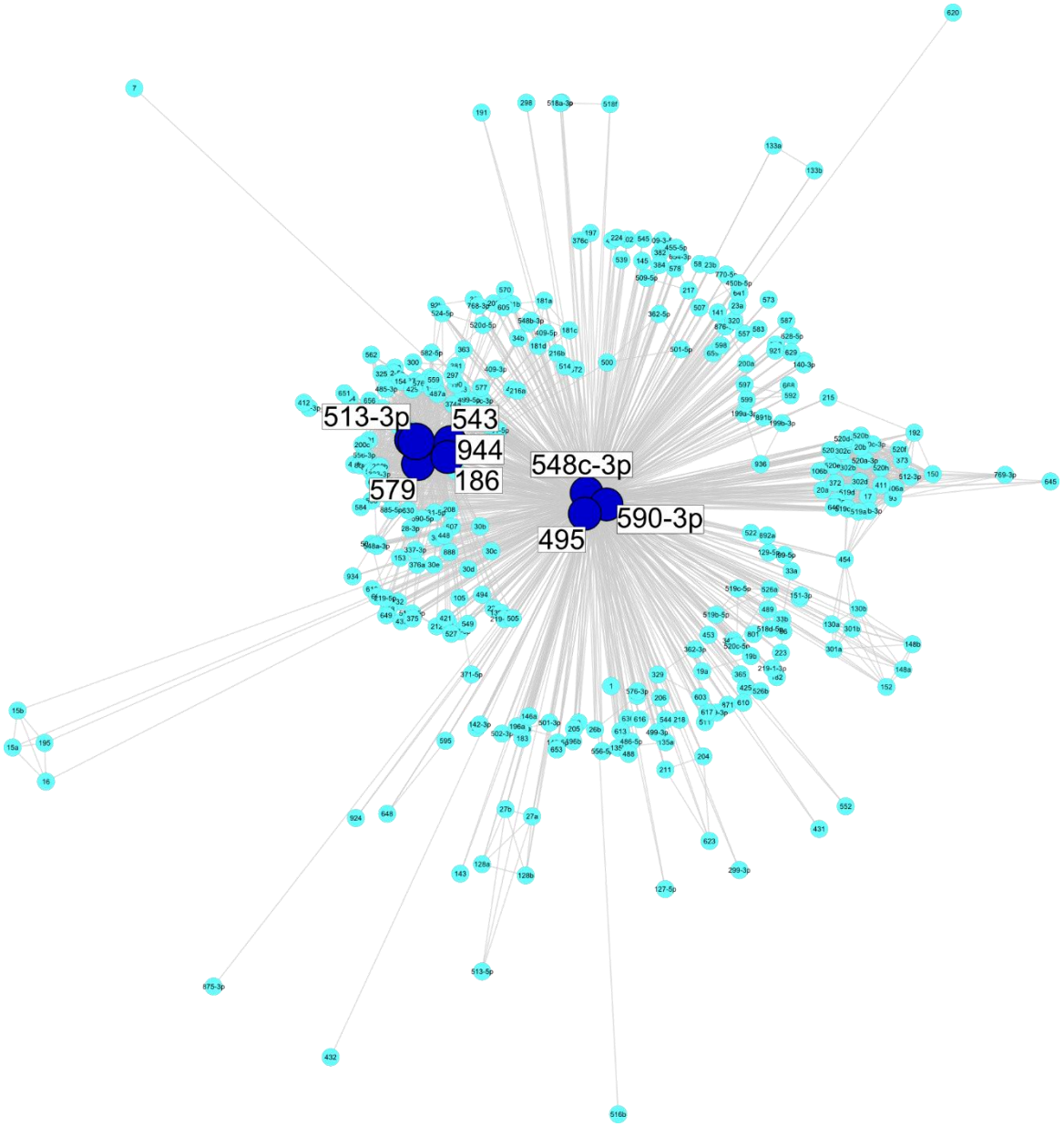


Figure 49. Le club assorti 1 et sa zone d'influence. Les nœuds ont été extraits du réseau complet de DIANA-microT à un seuil meet/min 50%.

4. Zone d'influence du club 1

La zone d'influence du club assorti 1 est composé de 315 miARN, elle forme la plus grande des trois zones déjà évoquées (Figure 49). Les trois miARN miR-548c-3p, -590-3p et -495 sont particulièrement centraux dans la zone. Pour analyser d'un point de vue biologique cette zone, c'est en priorité les gènes ciblés par au moins 25% des miARN de la zone qui ont été étudiés – soit les gènes retrouvés ciblés par au moins 79 miARN sur 315. Comme une bonne partie des gènes sont partagés entre la zone d'influence et le club assorti en soi (Figure 48 B), nous pouvons supposer *a priori* une corrélation des enrichissements GO entre les deux ensembles.

Effectivement, la zone d'influence semble également impliquée dans la régulation de la transcription mais également dans le développement et la différentiation (Tableau 7 : p-valeurs corrigées comprises entre 10^{-4} et 10^{-7}). Ces résultats confirment bien l'influence des miARN hubs sur les miARN qui les entourent. En restreignant l'analyse sur l'ensemble des gènes partagés par 50% des 315 miARN de la zone (au lieu de 25%), c'est essentiellement des enrichissements associés au développement du système nerveux qui ressortent, c'est-à-dire des enrichissements d'ontologies liés aux gènes hubs (voir partie « protéines hubs » page 148).

Tableau 7. Enrichissement GO pour les gènes partagés par 25% des membres de la zone d'influence du club assortis 1 (BP, MF et CC). Soit 4 208 gènes, ciblés par au moins 79 miARNs sur les 315 du cluster. Seules les dix premières annotations sont présentées.

BP					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0007399	nervous system development	1539	472	8.10E-11	3.33E-07
GO:2000112	regulation of cellular macromolecule biosynthetic process	2817	790	7.70E-09	1.58E-05
GO:0010556	regulation of macromolecule biosynthetic process	2877	799	4.20E-08	4.81E-05
GO:0019219	regulation of nucleobase-containing compound metabolic process	3046	840	6.50E-08	4.81E-05
GO:0031323	regulation of cellular metabolic process	3967	1069	7.80E-08	4.81E-05
GO:0048869	cellular developmental process	2496	700	8.10E-08	4.81E-05
GO:0031326	regulation of cellular biosynthetic process	2998	827	8.20E-08	4.81E-05
GO:0009889	regulation of biosynthetic process	3026	832	1.40E-07	6.99E-05
GO:0023052	signaling	4226	1130	1.60E-07	6.99E-05
GO:0035556	intracellular signal transduction	1716	497	1.70E-07	6.99E-05

MF					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0046872	metal ion binding	3551	986	1.70E-11	6.34E-08
GO:0008270	zinc ion binding	1780	529	1.50E-10	1.59E-07

GO:0043169	cation binding	3593	989	1.50E-10	1.59E-07
GO:0043167	ion binding	3603	991	1.70E-10	1.59E-07
GO:0046914	transition metal ion binding	2039	588	2.50E-09	1.87E-06
GO:0005488	binding	10826	2650	1.60E-06	0.000995
GO:0008092	cytoskeletal protein binding	579	185	2.00E-06	0.00107
GO:0005515	protein binding	6665	1689	2.60E-06	0.00121
GO:0004842	ubiquitin-protein transferase activity	242	87	9.60E-06	0.00398
GO:0019904	protein domain specific binding	513	162	1.80E-05	0.00672

CC

GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0030054	cell junction	682	235	2.50E-11	3.19E-08
GO:0044464	cell part	12849	3123	3.70E-09	1.70E-06
GO:0005623	cell	12850	3123	4.00E-09	1.70E-06
GO:0005622	intracellular	11345	2775	5.70E-07	1.63E-04
GO:0042995	cell projection	1053	314	6.40E-07	1.63E-04
GO:0005911	cell-cell junction	266	97	1.20E-06	2.55E-04
GO:0016020	membrane	6622	1677	2.00E-06	3.65E-04
GO:0045202	synapse	429	142	3.10E-06	4.95E-04
GO:0005886	plasma membrane	3563	937	4.90E-06	6.50E-04
GO:0043005	neuron projection	531	169	5.40E-06	6.50E-04

En résumé, les enrichissements de la zone d'influence du club assorti 1 suivent sensiblement ceux du club en lui-même. La zone d'influence semble plutôt orientée vers la régulation des régulateurs de la transcription.

C. Les gènes/protéines hubs

Les gènes (ou protéines) hubs ont été définis dans la publication de Shalgi et collaborateurs comme des gènes beaucoup plus ciblés par les miARN que les autres (Shalgi et al., 2007). Dans leur étude, les auteurs expliquent que ce ciblage préférentiel peut partiellement être expliqué par des régions 3'UTR géniques généralement plus longues, mais que ce constat seul ne suffit pas en soi.

En reprenant la même définition pour les gènes hubs que ces auteurs, nous avons dans notre cas 193 gènes hubs ciblés par au moins 293 miARN (cf. Figure 21). L'analyse d'enrichissement GO dans ce cas montre essentiellement des termes liés au système nerveux et à son développement (p-valeurs corrigées = 0,011). Pour les fonctions moléculaires, ce sont les termes liés aux activités « kinase » qui sont retrouvées et enfin, pour les compartiments cellulaires, des termes rattachés à la cavéole et aux neurones. Dans les deux cas en revanche,

les p-valeurs sont plus élevées (Tableau 8 : p-valeurs corrigées > 0,05 ; les p-valeurs non corrigées sont comprises entre 10^{-3} et 10^{-5}).

Tableau 8. Enrichissement GO pour les 193 gènes hubs (BP, MF et CC). Seules les dix premières annotations sont présentées.

BP

GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0007417	central nervous system development	590	23	4.50E-06	0.0109
GO:0007399	nervous system development	1539	42	5.30E-06	0.0109
GO:0007420	brain development	425	18	1.80E-05	0.0247
GO:0040011	locomotion	1118	32	3.60E-05	0.0312
GO:0048869	cellular developmental process	2496	56	3.80E-05	0.0312
GO:0030154	cell differentiation	2348	53	5.70E-05	0.0358
GO:0031016	pancreas development	74	7	6.10E-05	0.0358
GO:0022008	neurogenesis	1010	28	2.00E-04	0.0913
GO:0032502	developmental process	3958	76	2.00E-04	0.0913
GO:0048856	anatomical structure development	3476	68	0.00033	0.106

MF

GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0019199	transmembrane receptor protein kinase activity	80	7	7.70E-05	0.287
GO:0004713	protein tyrosine kinase activity	138	8	0.00042	0.457
GO:0005275	amine transmembrane transporter activity	76	6	0.00044	0.457
GO:0016362	activin receptor activity, type II	3	2	0.00049	0.457
GO:0003700	sequence-specific DNA binding transcription factor activity	916	24	0.00072	0.46
GO:0001071	nucleic acid binding transcription factor activity	918	24	0.00074	0.46
GO:0005030	neurotrophin receptor activity	4	2	0.00098	0.475
GO:0015171	amino acid transmembrane transporter activity	62	5	0.00121	0.475
GO:0004672	protein kinase activity	574	17	0.00124	0.475
GO:0016773	phosphotransferase activity, alcohol group as acceptor	683	19	0.00134	0.475

CC

GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0005901	caveola	56	6	5.70E-05	0.0728
GO:0031252	cell leading edge	231	11	0.00012	0.0766
GO:0033267	axon part	109	7	0.00036	0.105
GO:0042995	cell projection	1053	26	4.00E-04	0.105
GO:0032444	activin responsive factor complex	3	2	0.00043	0.105
GO:0044304	main axon	32	4	0.00057	0.105
GO:0030054	cell junction	682	19	0.00062	0.105
GO:0000932	cytoplasmic mRNA processing body	34	4	0.00073	0.105
GO:0043005	neuron projection	531	16	0.00074	0.105
GO:0071141	SMAD protein complex	4	2	0.00086	0.11

Ces résultats sont très proches de ceux observés pour les gènes partagés par 50% de membres de la zone d'influence du club 1. De plus, presque tous les gènes hubs sont retrouvés parmi ces gènes partagés (Figure 48 C). Nous pouvons donc supposer que les enrichissements observés pour la zone d'influence sont biaisés par les gènes hubs – enrichis pour le développement neuronal.

D. Club assorti 2

1. Description du groupe

Le club assorti 2 est composé des trois miARN miR-940, -661 et -612 (Tableau 5). En moyenne, les miARN de ce petit club sont prédits pour cibler 5 254 gènes. Parmi eux, 4 596 gènes seraient régulés par au moins deux des trois miARN (i.e. régulés par au moins 50% des miARN du club). Les trois miARN sont situés sur trois chromosomes différents : miR-940 est situé sur le chromosome 16, miR-661 sur le chromosome 8 et enfin miR-612 sur le chromosome 11 (Figure 46, en rouge). Par ailleurs, ces derniers ne semblent pas proches les uns des autres en trois dimensions au sein du noyau (Figure 47). Comme ce cluster de miARN

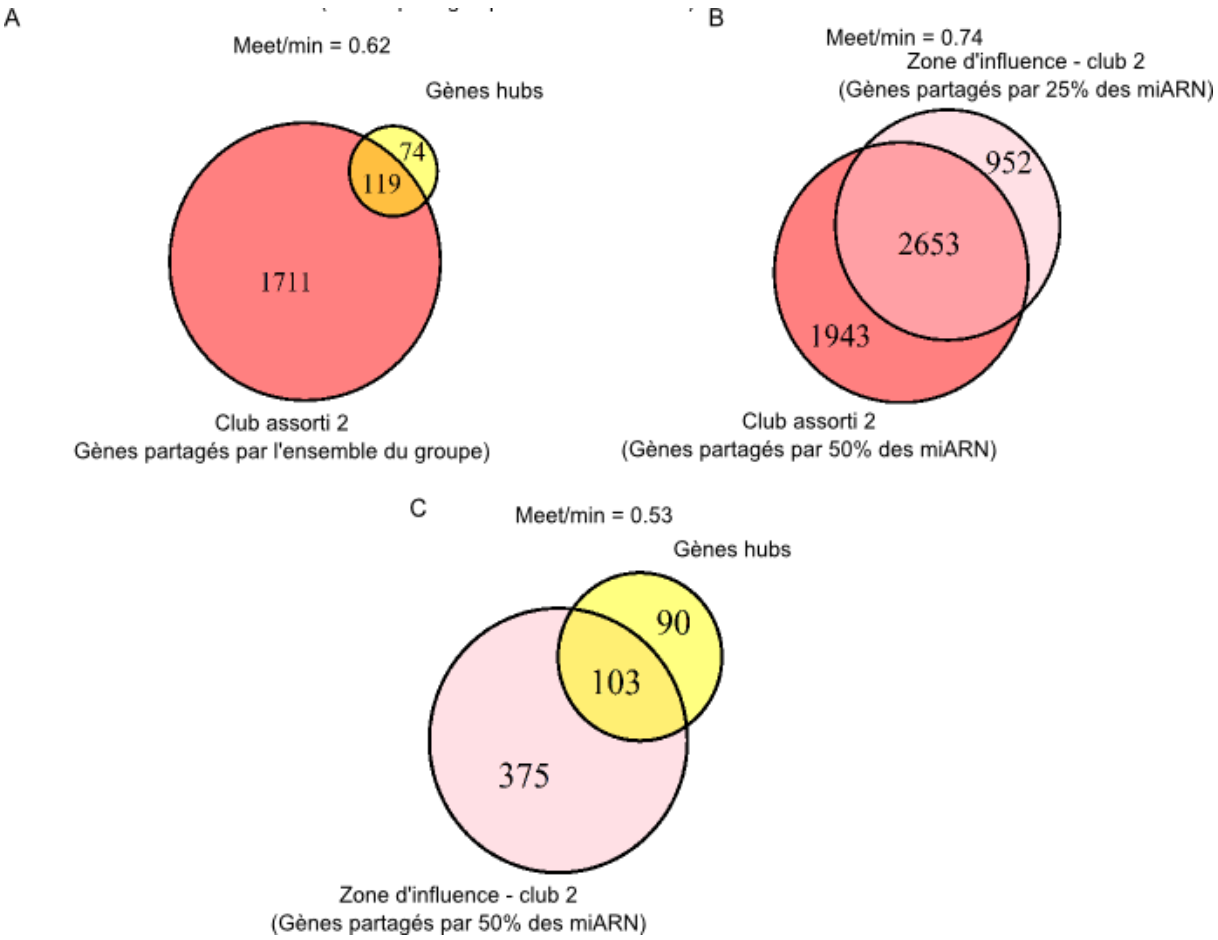


Figure 50. Diagramme de Venn de recouvrement de cibles entre les gènes hubs, le club assorti 2 et sa zone d'influence. En rouge foncé sont représentées les gènes cibles partagées par les membres du club assorti 2. En rouge clair, les gènes cibles partagées par les membres de la zone d'influence du club assorti 2 et en jaune, les gènes hubs dans les prédictions. A | Recouvrement de cibles entre les gènes hubs et les gènes partagés par l'ensemble des miARN du club assorti 2. B | Recouvrement de cibles entre les gènes partagés par 50% des membres du club assorti 2 (2/3) et 25% des membres de sa zone d'influence (33/129). C | Recouvrement de cibles entre les gènes partagés par 50% des membres de la zone d'influence du club assorti 2 (65/129) et les gènes hubs.

montre moins de cibles partagées que le club assorti 1, il pourrait montrer plus de spécificité que le premier – tout en gardant en mémoire qu'il reste tout de même un club assorti et que chacun des miARN sont des hubs dans le réseau. Sur les 1 830 gènes partagés par l'ensemble des trois miARN, 6,5% sont des gènes hubs. Ce pourcentage représente 62% des 193 gènes hubs de DIANA-microT (Figure 50 – p-valeur de Fisher = 10^{-71}). 52 gènes hubs sont partagés entre les clubs assortis 1 et 2.

2. Prédiction de processus biologiques

De la même manière qu'avec le club assorti 1, le Tableau 9 montre les « enrichissements GO » pour les gènes partagés par 50% des miARN du club. Le terme « transduction du signal médiée par les petites GTPase » est le terme le plus de significatif (p-valeur corrigée de 5 x 10^{-3}). Ce dernier est suivi par d'autres termes associés à la communication cellulaire et la signalisation (p-valeurs corrigées < 0,05). Un nombre significatif de gènes partagés par ce cluster est associé à une localisation membranaire (membrane de la cellule, des organelles ou encore du plasma) et aux jonctions cellulaires (p-valeurs corrigées < 10^{-5}). Enfin, parmi les fonctions moléculaires enrichies, nous retrouvons les liaisons aux phospholipides (p-valeur corrigée = 0,036). D'autres termes liés à la régulation et la liaison aux petites GTPases peuvent être retrouvés dans la liste avec des p-valeurs plus élevées (Tableau 9).

Tableau 9. Enrichissement GO pour les gènes partagés par 50% des membres du club assortis 2 (BP, MF et CC). Soit 4 596 gènes, ciblés par au moins 2 miARNs sur les 3 du cluster. Seules les dix premières annotations sont présentées.

BP					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0007264	small GTPase mediated signal transduction	575	186	2.40E-06	0.00567
GO:0007154	cell communication	4338	1148	3.00E-06	0.00567
GO:0023051	regulation of signaling	1778	503	5.00E-06	0.00567
GO:0009966	regulation of signal transduction	1553	444	6.20E-06	0.00567
GO:0007399	nervous system development	1539	440	6.90E-06	0.00567
GO:0035556	intracellular signal transduction	1716	484	1.10E-05	0.00668
GO:0023052	signaling	4226	1113	1.30E-05	0.00668
GO:0048583	regulation of response to stimulus	2027	563	1.30E-05	0.00668
GO:0007265	Ras protein signal transduction	353	118	3.40E-05	0.0155
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	581	180	5.70E-05	0.0234

MF

GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0005543	phospholipid binding	469	154	9.70E-06	0.0362
GO:0008289	lipid binding	675	207	4.70E-05	0.0634
GO:0005085	guanyl-nucleotide exchange factor activity	150	58	5.10E-05	0.0634
GO:0005083	small GTPase regulator activity	257	88	0.00016	0.124
GO:0030695	GTPase regulator activity	396	127	0.00018	0.124
GO:0008092	cytoskeletal protein binding	579	177	2.00E-04	0.124
GO:0046872	metal ion binding	3551	935	0.00026	0.131
GO:0051020	GTPase binding	145	54	0.00028	0.131
GO:0060589	nucleoside-triphosphatase regulator activity	408	129	0.00032	0.133
GO:0001228	RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	46	22	4.00E-04	0.149

CC

GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0016020	Membrane	6622	1761	6.00E-12	7.66E-09
GO:0071944	cell periphery	3635	998	8.20E-09	5.24E-06
GO:0044425	membrane part	5179	1378	1.50E-08	5.43E-06
GO:0005886	plasma membrane	3563	977	1.70E-08	5.43E-06
GO:0030054	cell junction	682	220	2.70E-07	6.90E-05
GO:0005912	adherens junction	183	72	2.30E-06	0.00049
GO:0070161	anchoring junction	200	76	5.60E-06	0.00102
GO:0016021	integral component of membrane	4323	1134	1.60E-05	0.0023
GO:0005925	focal adhesion	118	49	1.70E-05	0.0023
GO:0031090	organelle membrane	2159	594	1.80E-05	0.0023

En conclusion, bien que les p-valeurs ne soient pas exceptionnellement faibles, nous voyons tout de même une bonne cohérence de termes GO associés aux petites GTPase pour ce cluster.

3. Validations biologiques

Les petites GTPase sont une famille d'hydrolases impliquées dans différents processus cellulaires comme la prolifération, la morphologie ou encore le transport (Somlyo and Somlyo, 2000). Pour confirmer nos hypothèses, plusieurs expériences *in vitro* sur des cellules épithéliales de rétine de l'œil (RPE1) ont été menées, en s'intéressant notamment à la distribution spatiale et au niveau de phosphorylation de la chaine légère de la myosine II (MCLII ou phosphomyosine) après surexpression des trois miARN du club. MLCII est un substrat des protéines ROCK (*Rho-associated protein kinases*) et un produit terminal de la signalisation des petites GTPase (Somlyo and Somlyo, 2000), ce qui en fait un témoin idéal pour suivre cette voie de signalisation. La phosphorylation de MLCII permet en fait l'activation de la myosine et la création, par conséquent, des forces contractiles au travers du

cytosquelette d'actomyosine par effet de « coulissage » des filaments d'actine sur ceux de myosine (Charras and Paluch, 2008).

Afin de normaliser la forme des cellules et l'architecture du cytosquelette d'actine, les cellules ont été déposées sur des micropatrons de fibronectine (cf. page 96). La Figure 51 montre l'architecture des cellules RPE1 avec un traitement contrôle (siARN AllStars) ainsi qu'un inhibiteur chimique de ROCK (Y27632 – qui bloque la phosphorylation) sur des patrons circulaires de 500 µm². Comme attendu, une baisse globale de la phosphorylation de MLCII pouvait effectivement être observée avec Y27632. En revanche, ces patrons étant très petits par rapport aux cellules RPE1, les cellules semblaient trop comprimées. Comme cette compression pouvait influencer la contraction cellulaire, les dépôts suivants ont été effectués sur des patrons de 1000 µm². Ce changement permettait non seulement d'obtenir des cellules généralement moins contraintes mais également d'observer des niveaux de phosphorylation plus grands.

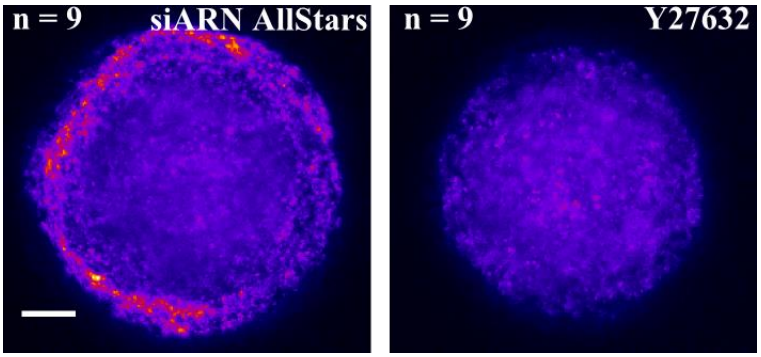


Figure 51. Immunomarquage de la phosphomyosine II des cellules RPE1 déposées sur patron circulaire de 500 µm². Neuf images différentes pour chaque condition ont été prises, alignées et projetées sur une seule image en utilisant les valeurs médianes des pixels (projection médiane sous ImageJ). Couleur de la table « *fire* » de ImageJ. Barre, 5 µm.

La quantification des niveaux de phosphorylation de MLCII par Western Blot de cellules non restreintes sur patron démontrait une baisse générale de la phosphorylation lorsque chacun des trois miARN étaient artificiellement surexprimés dans les cellules (Figure 52). Cette baisse générale prouve que ces trois miARN sont impliqués dans la voie de signalisation des petites GTPase bien que nous soyons incapables de dire à quel niveau. Les cellules surexprimant miR-612 et miR-940 semblaient relâchées et sans contraintes, des phénotypes

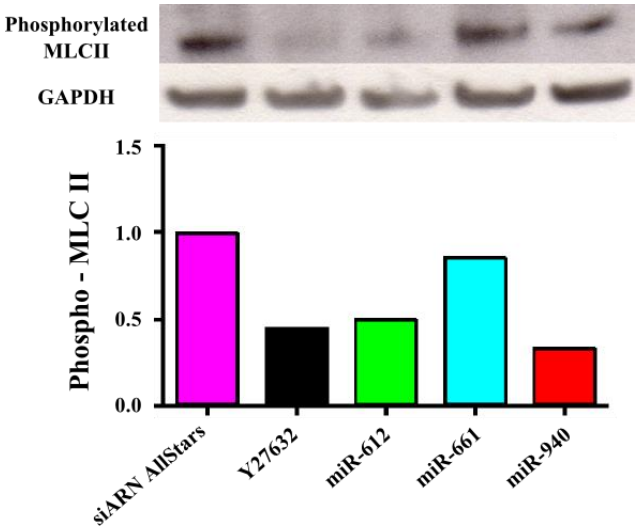


Figure 52. Western blot de la phosphomyosine II. La GAPDH tient lieu de contrôle de dépôt. Le diagramme montre les niveaux normalisés par rapport à la condition avec le contrôle : siRNA AllStars.

proches de celles traitées avec Y27632. Leurs filaments d'actine étaient globalement désorganisés avec une absence de fibres de stress et d'arcs transverses en comparaison aux cellules traités avec le contrôle négatif. Au contraire, les cellules avec une surexpression ectopique de miR-661 montraient un grand nombre de fibres de stress décorées de myosine II et des cellules très contractées avec des fibres très denses (Figure 53 A). En plus de la baisse du niveau de phosphorylation avec ce miARN, une réorganisation spatiale de la MLCII du bord des cellules vers l'intégralité de la surface cellulaire était aussi observée. Par ailleurs, une quantification plus fine du niveau de phosphorylation des cellules pour l'ensemble des conditions montrait que cette baisse était très marquée pour miR-940 (p-valeur = $1,2 \times 10^{-5}$) mais plus modérée pour les miR-612 et miR-661 (p-valeurs de 0,04 et 0,058 respectivement) (Figure 53 B). Enfin, une augmentation de l'intensité de fluorescence des filaments d'actine pouvait être observée sur les cellules traitées avec le miR-661 (p-valeur = 0,012). A l'inverse, les cellules traitées avec les deux autres miARN montraient une diminution de l'intensité de fluorescence (p-valeurs = 0,031 et $1,2 \times 10^{-5}$ pour miR-612 et -940 respectivement) (Figure 53 C). Cinq images typiques de cellules déposées sur patron circulaire et pour chaque condition utilisée sont représentées sur la Figure 54. Ces images récapitulent les conclusions précédentes mais au niveau de chaque cellule individuelle.

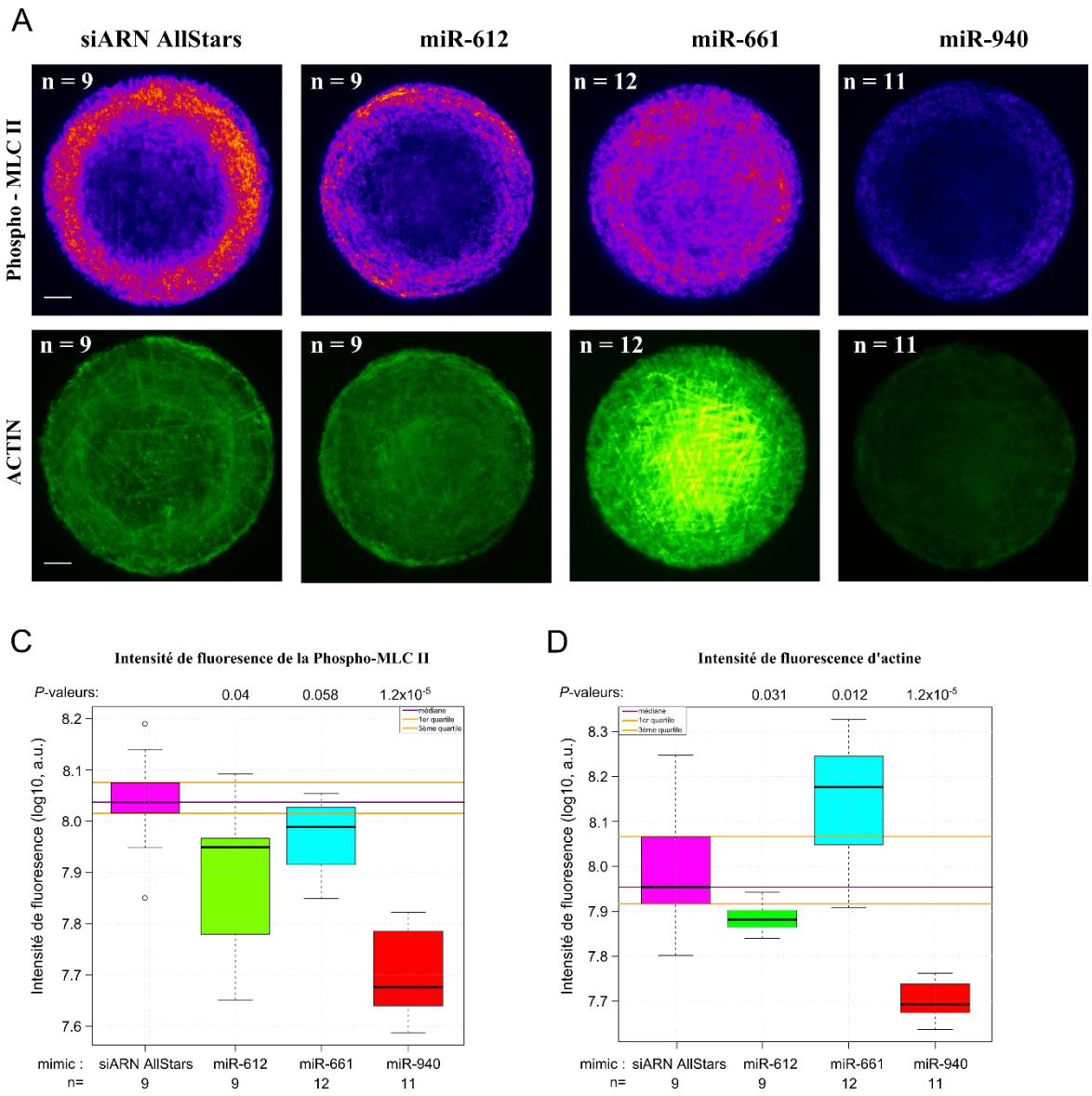


Figure 53. Implication des miR-661, -612 et -940 dans la voie de signalisation des petites GTPase. A | Immunomarquage de la phosphomyosine II et des filaments d'actine avec des tables de couleurs « *fire* » et « *green* » respectivement. Les trois miARNs ont été transfectés dans les cellules afin de mimer une surexpression des miARN. Pattern circulaire de 1000 µm². Procédure d'analyse et de comparaison d'image similaire à A. Barre, 5 µm. B | Log10 des niveaux de phosphorylation calculés sur les images présentées en C après segmentation automatique des cellules dans ImageJ. C | Log10 de l'intensité de la fluorescence des filaments d'actine. a.u. : unité arbitraire. p-valeurs issus du test bilatéral de Mann-Whitney.

En résumé de ces premières expériences, les trois miARN semblent bel et bien avoir un impact sur la voie de signalisation des petites GTPases puisqu'une baisse de la phosphorylation de MLCII est observable dans chaque cas. En revanche, le phénotype des cellules n'est pas le même après la surexpression artificielle des miARN : les cellules traitées avec les miR-940 et miR-612 montrent plutôt des formes relâchées alors que celles traitées avec miR-661 montrent une sur-contraction cellulaire par rapport à la condition avec le contrôle négatif. Ces phénomènes pouvant grandement influencer le comportement cellulaire en

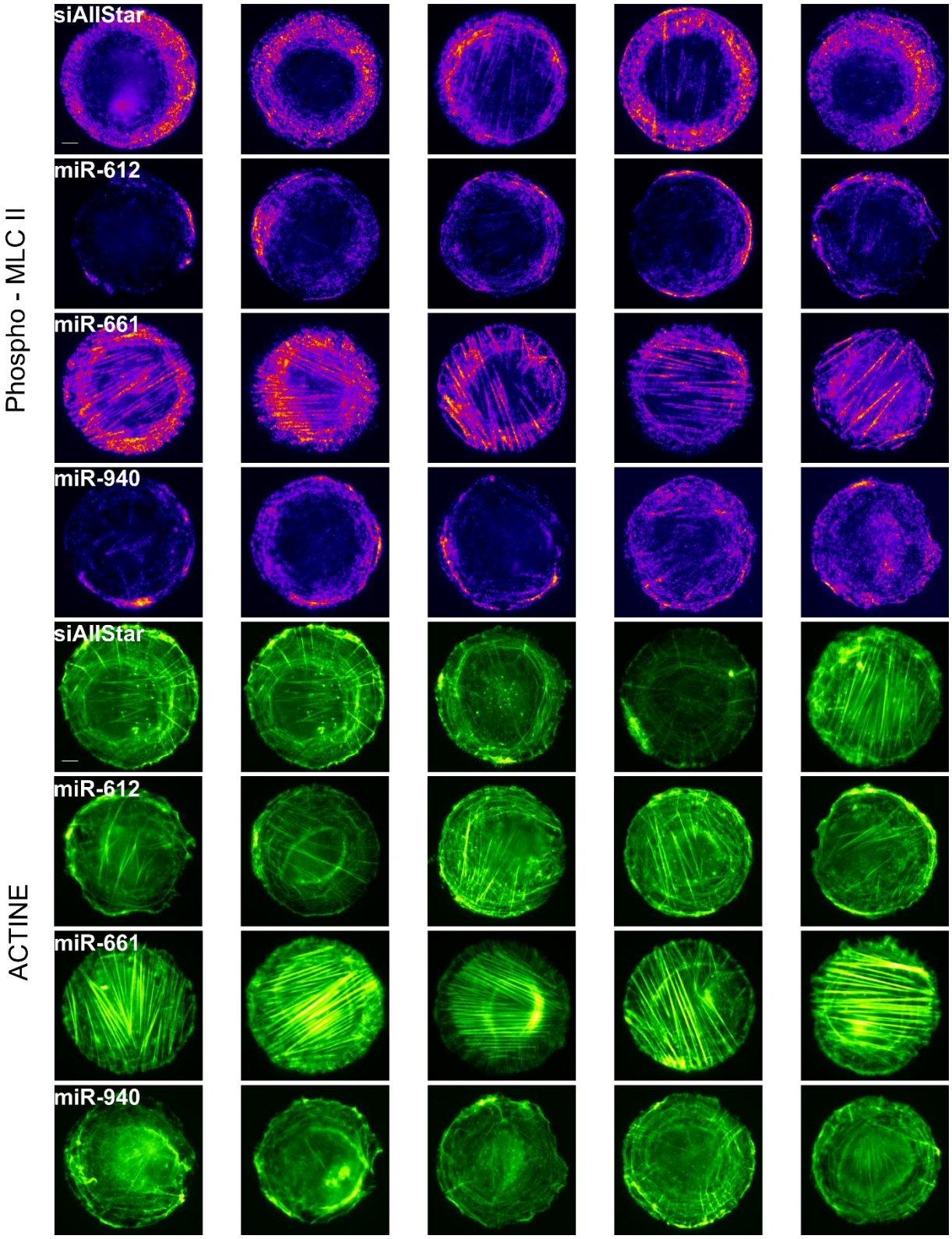


Figure 54. Immunomarquage d'image individuelle des filaments d'actine et de la phosphomyosine II sur des patrons circulaires de fibronectine de 1000 µm². Sur chaque patron, seule une cellule est capable d'adhérer. La cellule prend la forme du patron, en l'occurrence un cercle. Cette normalisation structurale permet donc des comparaisons et des superpositions plus aisées.

termes de migration et de prolifération, deux autres expériences ont été menées afin d'observer le comportement des cellules et leur habilité à se déformer et à migrer en présence

d'un fort niveau d'expression de chacun des trois miARN.

Ces deux expériences sont le test *transwell* et le test de blessure. Dans le cas du *transwell*, l'objectif est de comprendre la dynamique du cytosquelette et la propension des cellules à passer au travers de petits pores dans une membrane. Les pores étant bien plus petit que le diamètre cellulaire, les cellules doivent obligatoirement se déformer pour pouvoir traverser la membrane. Pour le test de blessure, il s'agit d'évaluer la capacité des cellules à se diviser et/ou à migrer pour assurer la fermeture d'une blessure au milieu d'un tapis cellulaire.

Dans le test *transwell*, miR-661 produisait une claire augmentation du nombre de cellules passant au travers des pores par rapport au contrôle négatif (Figure 55, p-valeur = 1,3 x 10^{-8}). De manière corrélée, le miARN entrainait une fermeture bien plus rapide de la blessure par rapport au contrôle (Figure 56). Par ailleurs, les cellules traitées avec miR-661 étaient bien plus confluentes avec des marquages bien plus forts que pour les autres conditions (Figure

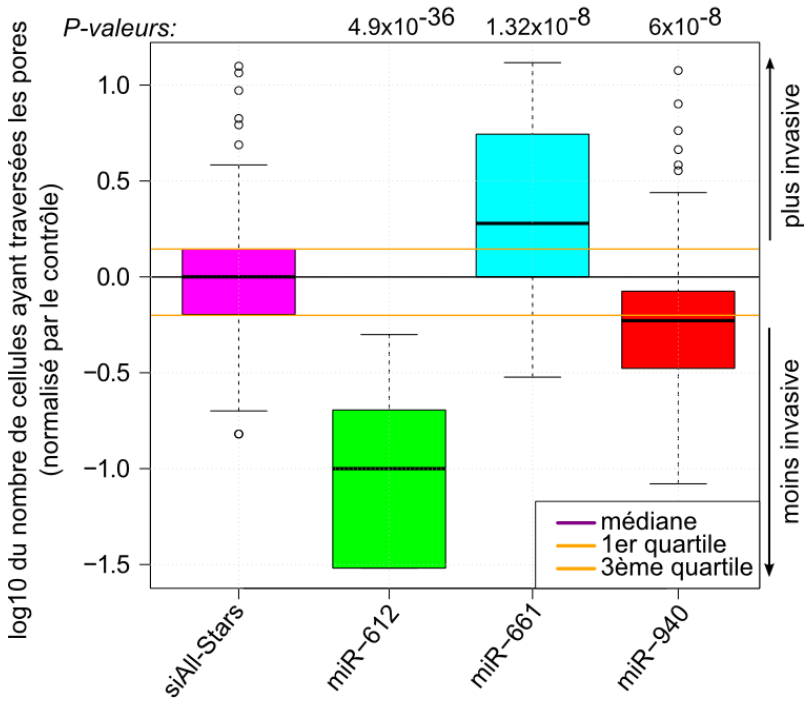


Figure 55. *Transwell* et effet de l'expression ectopique des miR-612, -661 et -940 sur la migration des cellules RPE1. Graphe de motilité : nombre de cellules normalisé passant par les pores de 5 µm. Le comptage des cellules s'est fait 18 heures après dépôt des cellules sur les membranes poreuses. Quatre expériences différentes avec un nombre initial variable de cellules déposées ont été menées. Le nombre de cellules pour chaque condition a été normalisé par rapport au contrôle négatif siRNA-AllStars et les quatre expériences ont été assemblées en une seule.

57). Ces résultats suggèrent que miR-661 augmente fortement la motilité cellulaire ainsi que la division cellulaire lorsqu'il est surexprimé. A l'inverse, miR-612 produisait une baisse d'un facteur 10 du nombre de cellules traversant les pores dans le *transwell* (p-valeur = 4,9 x 10^{-36}) et bloquait complétement la fermeture de la blessure. De plus, les cellules transfectées avec le miARN montraient des interactions cellulaires et des formes très différentes : les cellules étaient premièrement beaucoup moins confluentes et formaient des réseaux creux très proches des surfaces épithéliales (Figure 57). De la même manière, les cellules transfectées avec le miR-940 montraient une baisse générale du nombre de cellules pouvant traverser les pores (p-valeur = 6 x 10^{-8}). Cette baisse était toutefois moins importante qu'avec le miR-612. Par ailleurs, le miARN n'induisait pas de changements sur la fermeture des blessures par

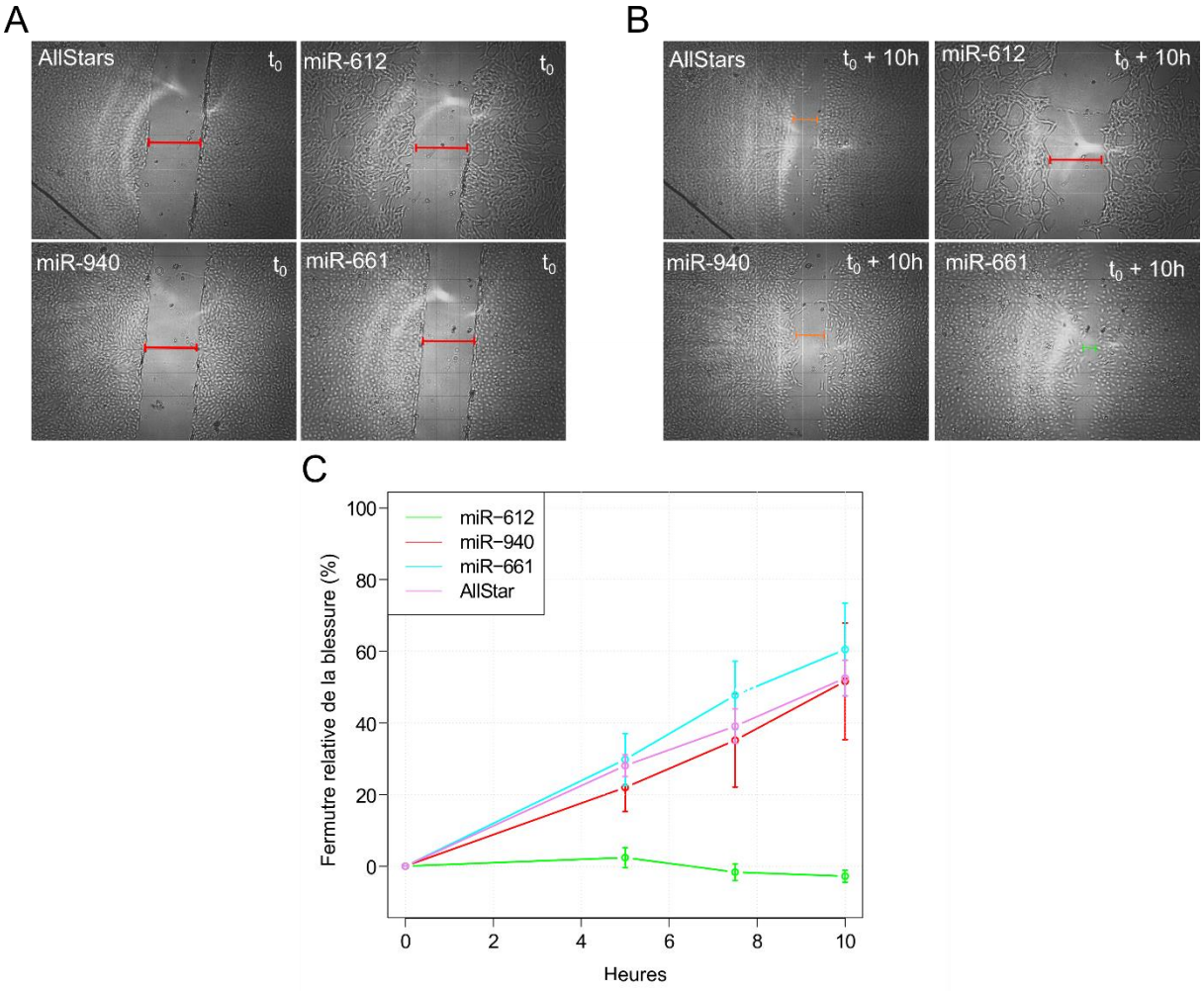


Figure 56. Test de blessure et effet de l'expression ectopique des miR-612, -661 et -940 sur la migration et la prolifération des cellules RPE1. A | Images de blessure à t_0 pour les différentes conditions. B | Images de blessure à t_0+10h. Vert : blessure très fermée, orange : blessure moyennement fermée, rouge : blessure similaire à t_0. C | Les blessures ont été faites sur des cellules à confluence après transfection des trois miARNs. L'expérience a été menée pendant 10 heures avec des images prises à t_0, t_0+5h, t_0+7.5h et t_0+10h. Chaque condition était présente en triplicat.

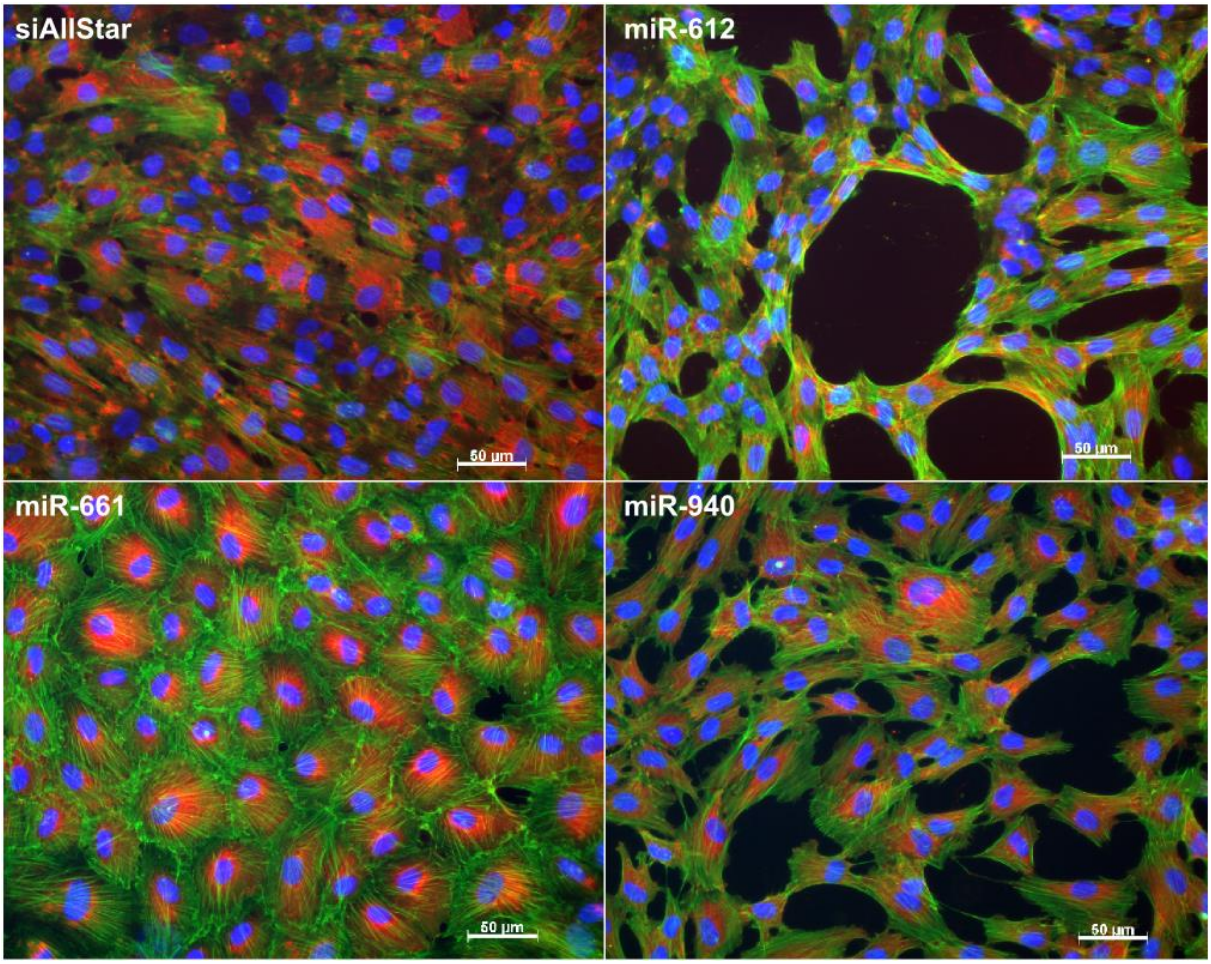


Figure 57. Images des cellules RPE1 après surexpression des miR-612, -661 et -940. Marquage phalloïdine (actine en vert), vinculine (rouge) et noyaux (Hoechst) des cellules transfectées avec les trois miARNs. Les images ont été prises avec un objectif 20x.

rapport au contrôle (Figure 56). L'impact du miARN sur la forme des cellules restait malgré tout très remarquable et les cellules montraient un phénotype similaire au miR-612 – avec, toutefois encore, un phénotype moins marqué (Figure 57).

En conclusion, l'ensemble de ces résultats confirment les prédictions que nous avions faites sur la fonction cellulaire des gènes ciblés par les miARN du club assorti 2. Les trois miARN interviennent effectivement sur la voie de signalisation des petites GTPases mais à ce stade nous n'avons pas encore caractérisé le type de corégulation qu'ont les trois miARN sur la fonction, ni à quel niveau de la cascade ils interviennent.

4. Expression de miR-940 dans les cellules du sein

Sachant que le rôle de miR-940 n'était pas connu au moment de sa découverte par l'approche systémique proposée dans cette thèse, nous nous sommes également intéressés

à son expression – ainsi qu'à celle des miR-612 et -661 – dans différents tissus et notamment le sein.

Dans les trois ensembles de données étudiés (Schrauder et al., 2012; Feliciano et al., 2013), seul miR-940 a été retrouvé différentiellement exprimé. La Figure 58 montre les données d'expression du miARN entre des tissus sains de sein et des tumeurs du sein, obtenues à partir de trois études totalement indépendantes. Dans les trois études, nous pouvons constater que le miARN est réprimé dans les cancers du sein par rapport aux tissus sains (logFC de -0,26, pour GSE44124. -0,53 pour GSE31309 et -0,57 pour GSE38867). Dans deux de ces cas les p-valeurs sont très faibles. Pour le troisième cas, la p-valeur élevée est

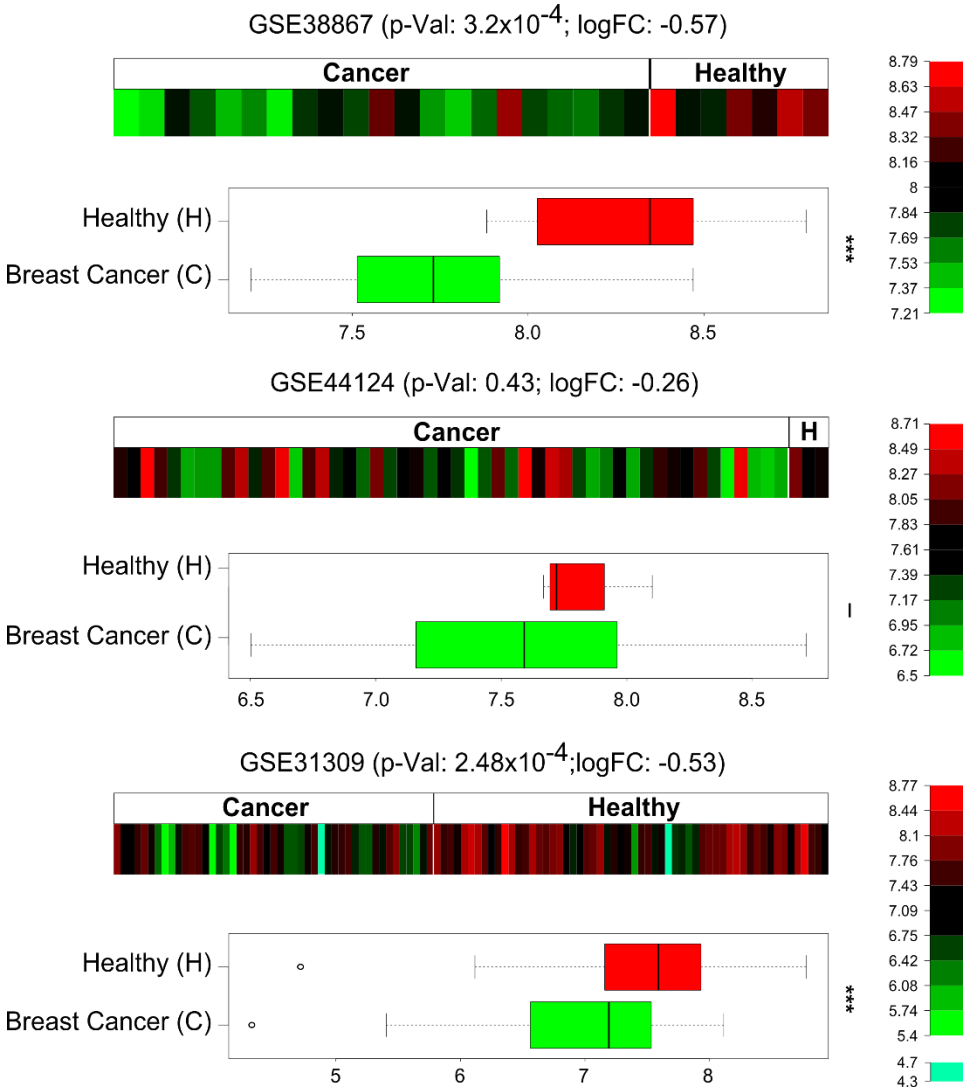


Figure 58. Expression relative du miR-940 dans le sein. L'expression du miR-940 sur trois ensembles de données sur le cancer du sein tirés de GEO (GSE38867, GSE44124 et GS31309) (Schrauder et al., 2012; Feliciano et al., 2013). p-valeurs issues d'une analyse *limma*. LogFC est *log Fold Change*, c'est-à-dire le logarithme du ratio cancéreux et sain.

probablement due au nombre très faible de tissus sains (3 tissus). Ces résultats sont en accord avec ceux que nous avons déjà abordés précédemment et confirme le rôle de miR-940 dans la dynamique cellulaire tout en lui faisant apparaitre un nouveau rôle potentiel dans la progression tumorale, notamment dans le cancer du sein.

5. Zone d'influence

Tout comme le club assorti 1, les membres du club assorti 2 définissent leur propre zone d'influence, composée dans ce cas de 129 miARN (Figure 59). Les gènes partagés par 25% des membres de la zone (au moins 33 miARN sur 129) montrent, dans ce cas, des enrichissements généraux pour la signalisation et la communication cellulaire mais également

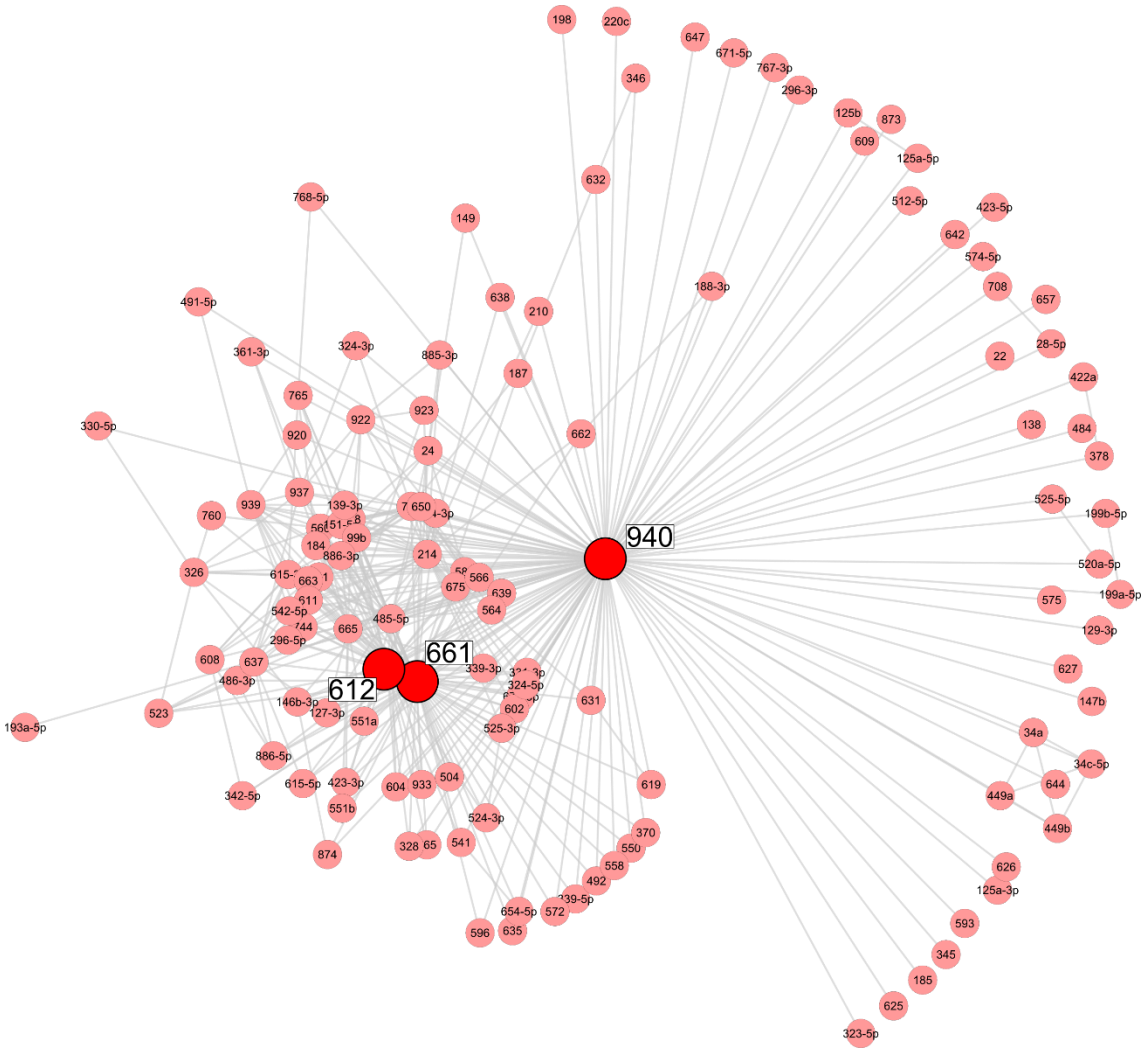


Figure 59. Le club assorti 2 et sa zone d'influence. Les nœuds ont été extraits du réseau complet de DIANA-microT à un seuil meet/min 50%.

le développement du système nerveux (Tableau 10) avec des p-valeurs corrigées entre 10^{-6} et 10^{-14}. A cause du fort nombre de gènes recouverts entre ceux partagés par les membres de la zone et ceux partagés par le club assorti 2 (Figure 50 B), il existe encore une fois une corrélation des enrichissements entre le club assorti et sa zone d'influence. Il est assez probable que les enrichissements en termes GO associés au système nerveux soient à nouveau en partie dus aux gènes hubs. En limitant l'analyse aux gènes partagés par 50% des 129 miARN, un biais encore plus conséquent pour le système nerveux pouvait être observé. Le recouvrement entre les gènes hubs et ces gènes étant proche de 50% (Figure 50 C), cette significativité accrue n'est pas étonnante.

Tableau 10. Enrichissement GO pour les gènes ciblés par 25% des membres de la zone d'influence du club assortis 2 (BP, MF et CC). Soit 3605 gènes, ciblés par au moins 33 miARNs sur les 129 du cluster. Seules les dix premières annotations sont présentées.

BP					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0007154	cell communication	4338	1020	1.40E-18	5.75E-15
GO:0023052	Signaling	4226	986	1.30E-16	2.67E-13
GO:0007399	nervous system development	1539	406	1.10E-13	1.51E-10
GO:0023051	regulation of signaling	1778	455	5.30E-13	5.44E-10
GO:0007268	synaptic transmission	574	176	1.10E-11	9.04E-09
GO:0035637	multicellular organismal signaling	652	193	3.30E-11	2.26E-08
GO:0007165	signal transduction	3789	860	7.50E-11	4.40E-08
GO:0007264	small GTPase mediated signal transduction	575	173	9.00E-11	4.57E-08
GO:0009966	regulation of signal transduction	1553	394	1.00E-10	4.57E-08
GO:0019226	transmission of nerve impulse	644	189	1.20E-10	4.93E-08

MF					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0005083	small GTPase regulator activity	257	97	2.30E-12	8.58E-09
GO:0030695	GTPase regulator activity	396	129	1.10E-10	2.05E-07
GO:0008092	cytoskeletal protein binding	579	172	3.80E-10	4.73E-07
GO:0060589	nucleoside-triphosphatase regulator activity	408	129	9.40E-10	8.77E-07
GO:0005543	phospholipid binding	469	141	6.50E-09	4.85E-06
GO:0005085	guanyl-nucleotide exchange factor activity	150	54	9.80E-07	0.00061
GO:0008289	lipid binding	675	176	4.90E-06	0.00261
GO:0005096	GTPase activator activity	232	72	9.90E-06	0.00415
GO:0016773	phosphotransferase activity, alcohol group as acceptor	683	176	1.00E-05	0.00415
GO:0003779	actin binding	324	93	2.00E-05	0.00653

CC					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0016020	membrane	6622	1462	1.30E-18	1.66E-15
GO:0071944	cell periphery	3635	866	1.20E-17	7.66E-15
GO:0005886	plasma membrane	3563	847	6.80E-17	2.89E-14
GO:0030054	cell junction	682	211	8.80E-15	2.81E-12
GO:0044425	membrane part	5179	1146	3.40E-13	8.68E-11
GO:0045202	synapse	429	138	2.20E-11	4.68E-09
GO:0043005	neuron projection	531	155	3.90E-09	7.11E-07
GO:0016021	integral component of membrane	4323	944	5.90E-09	9.42E-07

GO:0030136	clathrin-coated vesicle	177	66	7.10E-09	9.58E-07
GO:0031224	intrinsic component of membrane	4424	963	7.50E-09	9.58E-07

En conclusion, la zone d'influence du club assorti 2 suit les enrichissements observés pour le club assorti 2 : les miARN de cette zone semble plutôt être orientés dans la régulation de la signalisation cellulaire.

E. Zone intermédiaire

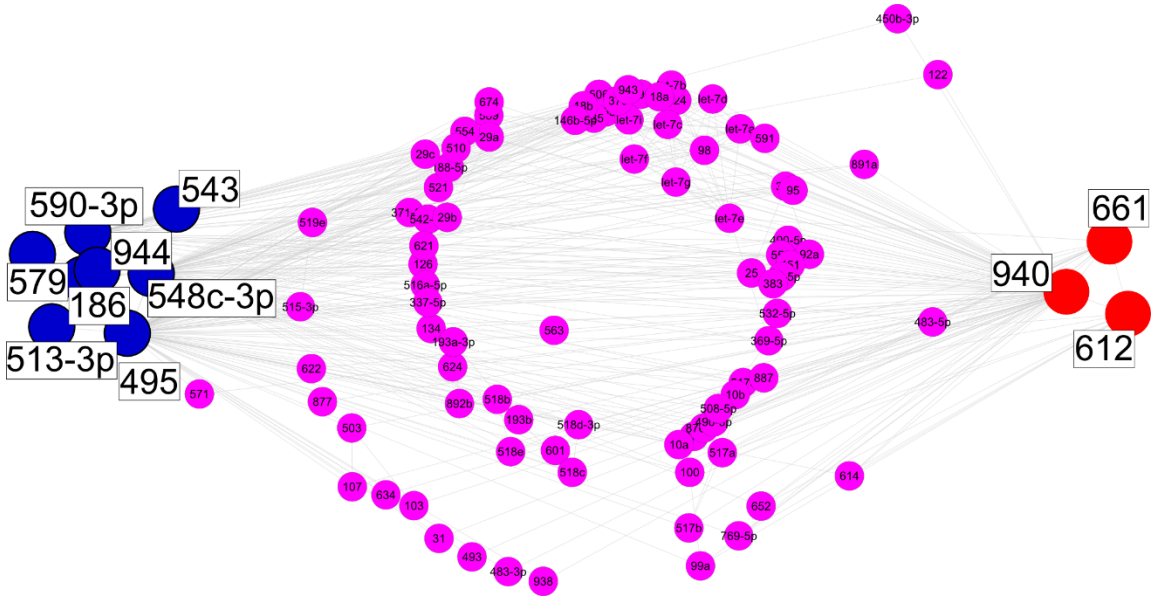


Figure 60. Les clubs assortis 1 et 2 et la zone transitoire. Les nœuds ont été extraits du réseau complet de DIANA-microT à un seuil meet/min de 50%.

Entre les deux zones d'influence des clubs assortis 1 et 2 se situe la zone intermédiaire – la plus petite des trois zones – dont les 89 miARN sont reliés à la fois à au moins un des membres de chacun des deux clubs (Figure 60). Contrairement aux deux zones d'influence, la zone intermédiaire ne montre aucun enrichissement particulier. Etant donné la liaison entre les deux clubs, ce résultat peut sembler logique : aucun des deux groupes de fonctions retrouvées pour les deux clubs ne domine l'autre. En analysant les 3 011 gènes partagés par 25% des miARN de la zone intermédiaire (ciblés par au moins 23 miARN sur 89), nous observons de très faibles enrichissements orientés vers le transport avec des p-valeurs corrigées entre 10^{-1} et 10^{-4} (les neuf premières annotations des processus biologiques). Nous pouvons également noter la présence d'un terme associé au système nerveux. Des fonctions

moléculaires associées aux jonctions cellulaires et aux membranes sont tout de même récupérés avec des p-valeurs corrigées entre 10^{-2} et 10^{-6}. En revanche, pour les gènes ciblés par au moins 50% des miARN de cette zone, aucun enrichissement n'est retrouvé.

Tableau 11. Enrichissement GO pour les gènes partagés par 25% des membres de la zone intermédiaire (BP, MF et CC). Soit 3 011 gènes, ciblés par au moins 23 miARNs sur les 89 du cluster. Seules les dix premières annotations sont présentées.

BP					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0007399	nervous system development	1539	255	1.50E-08	6.16E-05
GO:0051179	localization	3782	538	8.00E-07	0.00164
GO:0035329	hippo signaling	29	13	1.10E-05	0.0151
GO:0006865	amino acid transport	97	27	1.90E-05	0.0195
GO:0055085	transmembrane transport	772	130	3.30E-05	0.0271
GO:0007417	central nervous system development	590	103	5.30E-05	0.0363
GO:0048856	anatomical structure development	3476	482	7.80E-05	0.0406
GO:0030030	cell projection organization	822	135	7.90E-05	0.0406
GO:0006810	transport	3040	426	9.70E-05	0.0443
GO:0022008	neurogenesis	1010	160	0.00011	0.0448

MF					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0008092	cytoskeletal protein binding	579	104	1.10E-05	0.0317
GO:0005543	phospholipid binding	469	87	1.70E-05	0.0317
GO:0008270	zinc ion binding	1780	266	2.90E-05	0.0361
GO:0046914	transition metal ion binding	2039	298	5.10E-05	0.0476
GO:0046872	metal ion binding	3551	485	0.00018	0.134
GO:0008013	beta-catenin binding	57	17	0.00025	0.156
GO:0043169	cation binding	3593	488	3.00E-04	0.16
GO:0043167	ion binding	3603	488	0.00039	0.17
GO:0017002	activin-activated receptor activity	7	5	0.00041	0.17
GO:0019904	protein domain specific binding	513	86	0.00067	0.231

CC					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0030054	cell junction	682	131	6.80E-09	8.68E-06
GO:0016020	Membrane	6622	887	4.30E-08	2.75E-05
GO:0031252	cell leading edge	231	53	1.20E-06	0.000511
GO:0071944	cell periphery	3635	505	5.40E-06	0.00172
GO:0005886	plasma membrane	3563	494	9.20E-06	0.00235
GO:0044425	membrane part	5179	691	1.30E-05	0.00277
GO:0045202	Synapse	429	80	1.90E-05	0.00347
GO:0030027	Lamellipodium	109	28	4.70E-05	0.0075
GO:0042995	cell projection	1053	165	6.00E-05	0.00851
GO:0044463	cell projection part	553	95	9.30E-05	0.0119

En résumé, la zone de transition montre peu d'enrichissement particulier. Les miARN de cette zone étant partagés entre les deux clubs, les fonctions de l'un et l'autre des clubs se retrouvent probablement partagées dans la zone et aucun ne semble prendre l'ascendant sur l'autre.

F. Robustesse des enrichissements

1. Robustesse au changement d'algorithme

Les cibles potentielles des miARN des deux clubs étant partiellement différentes d'un algorithme à l'autre, nous pouvons nous demander quels sont les effets de ces différences sur les enrichissements. En effet, nous avons déjà vu que le changement d'algorithme entraine effectivement des variations en termes de degré des miARN mais que les réseaux construits indépendamment sont sensiblement identiques, tout comme les liens reliant les miARN des clubs assortis.

Le Tableau 12 montre les enrichissements aux trois niveaux BP, MF et CC pour les gènes partagés par les huit membres du club assorti 1 et prédits par TargetScan. Dans un premier temps, les résultats dans l'ensemble des trois tableaux sont sensiblement identiques à ceux obtenus avec les prédictions de DIANA-microT avec un coefficient de corrélation entre les deux séries d'enrichissement de 0,89 (Figure 61 A). Dans ce cas, la significativité des termes est légèrement plus faible.

Tableau 12. Enrichissement GO pour les gènes partagés par 50% des membres du club assortis 1 (BP, MF et CC). D'après les prédictions de TargetScan.

BP					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:2000112	regulation of cellular macromolecule biosynthetic process	2955	961	1.60E-10	6.70E-07
GO:0010556	regulation of macromolecule biosynthetic process	3019	977	3.70E-10	7.75E-07
GO:0031326	regulation of cellular biosynthetic process	3150	1012	9.60E-10	1.03E-06
GO:0009889	regulation of biosynthetic process	3181	1021	9.80E-10	1.03E-06
GO:0006464	cellular protein modification process	2399	779	2.40E-08	1.30E-05
GO:0036211	protein modification process	2399	779	2.40E-08	1.30E-05
GO:0010468	regulation of gene expression	3159	1002	2.60E-08	1.30E-05
GO:0060255	regulation of macromolecule metabolic process	3946	1230	2.70E-08	1.30E-05
GO:0019219	regulation of nucleobase-containing compound metabolic process	3191	1011	2.80E-08	1.30E-05
GO:0006355	regulation of transcription, DNA-templated	2732	876	3.30E-08	1.38E-05

MF					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0046872	metal ion binding	3796	1217	6.60E-13	2.50E-09
GO:0043169	cation binding	3842	1224	4.30E-12	6.95E-09
GO:0043167	ion binding	3852	1226	5.50E-12	6.95E-09
GO:0008270	zinc ion binding	1907	649	2.00E-11	1.90E-08
GO:0046914	transition metal ion binding	2179	726	9.80E-11	7.43E-08
GO:0005488	binding	11411	3262	4.50E-08	2.84E-05
GO:0019787	small conjugating protein ligase activity	260	110	1.90E-07	0.000103
GO:0004842	ubiquitin-protein transferase activity	245	104	3.40E-07	0.000161

GO:0016881	acid-amino acid ligase activity	288	117	9.90E-07	0.000417
GO:0016879	ligase activity, forming carbon-nitrogen bonds	320	125	4.70E-06	0.00178

CC					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0044464	cell part	13624	3861	8.60E-15	6.12E-12
GO:0005623	cell	13625	3861	9.50E-15	6.12E-12
GO:0005622	intracellular	11960	3429	9.60E-13	4.12E-10
GO:0044424	intracellular part	11637	3338	8.60E-12	2.77E-09
GO:0043227	membrane-bounded organelle	9097	2650	4.40E-10	1.07E-07
GO:0043231	intracellular membrane-bounded organelle	9088	2647	5.00E-10	1.07E-07
GO:0043226	organelle	10085	2898	1.70E-08	3.13E-06
GO:0043229	intracellular organelle	10072	2893	2.40E-08	3.87E-06
GO:0005634	nucleus	5682	1689	1.20E-07	1.72E-05
GO:0005794	Golgi apparatus	1095	366	1.90E-06	0.000245

Le Tableau 13, quant à lui, montre les enrichissements pour le club assorti 2 avec les prédictions de TargetScan. Ici encore, il existe une forte similarité entre les résultats obtenus avec DIANA-microT : le club assorti 2 montre également une implication potentielle dans la régulation de la signalisation par les petites GTPases avec un coefficient de corrélation de 0,73 dans ce cas (Figure 61 B). La différence pour ce cluster est que la significativité des résultats est supérieure à ceux obtenus avec DIANA-microT.

Tableau 13. Enrichissement GO pour les gènes partagés par 50% des membres du club assortis 2 (BP, MF et CC). D'après les prédictions de TargetScan.

BP					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0023051	regulation of signaling	1853	498	2.00E-07	0.000838
GO:0007265	Ras protein signal transduction	365	122	4.10E-07	0.000859
GO:0007268	synaptic transmission	599	183	7.90E-07	0.0011
GO:0035637	multicellular organismal signaling	679	203	1.10E-06	0.00115
GO:0016192	vesicle-mediated transport	894	257	1.40E-06	0.00117
GO:0051056	regulation of small GTPase mediated signal transduction	349	115	1.90E-06	0.00133
GO:0019226	transmission of nerve impulse	671	199	2.50E-06	0.0015
GO:0009966	regulation of signal transduction	1615	431	3.50E-06	0.00168
GO:0007264	small GTPase mediated signal transduction	593	178	3.80E-06	0.00168
GO:0051234	establishment of localization	3238	813	4.00E-06	0.00168

MF					
GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0005083	small GTPase regulator activity	272	101	1.40E-08	5.31E-05
GO:0005543	phospholipid binding	483	155	2.60E-07	0.000493
GO:0046872	metal ion binding	3796	953	6.10E-07	0.000771
GO:0030695	GTPase regulator activity	417	135	9.00E-07	0.000853
GO:0043167	ion binding	3852	963	1.20E-06	0.000885
GO:0043169	cation binding	3842	960	1.40E-06	0.000885
GO:0060589	nucleoside-triphosphatase regulator activity	429	136	2.90E-06	0.00157
GO:0022857	transmembrane transporter activity	887	250	1.20E-05	0.00548
GO:0008289	lipid binding	705	204	1.30E-05	0.00548
GO:0005099	Ras GTPase activator activity	109	44	1.60E-05	0.00607

CC

GO.ID	Terme	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0016020	membrane	7088	1805	4.90E-21	6.32E-18
GO:0044425	membrane part	5594	1434	6.20E-16	4.00E-13
GO:0005886	plasma membrane	3800	1000	3.40E-13	1.46E-10
GO:0071944	cell periphery	3871	1014	7.80E-13	2.51E-10
GO:0031224	intrinsic component of membrane	4809	1219	2.50E-11	6.45E-09
GO:0016021	integral component of membrane	4704	1194	3.30E-11	7.09E-09
GO:0030054	cell junction	698	216	1.40E-08	2.58E-06
GO:0044459	plasma membrane part	1987	533	3.80E-08	6.12E-06
GO:0031090	organelle membrane	2249	582	1.60E-06	0.000229
GO:0045202	synapse	445	140	1.90E-06	0.000245

Au final, s'il existait déjà une robustesse dans la construction et l'analyse des réseaux ainsi que dans les liaisons entre les membres des clubs assortis, nous observons également une certaine robustesse dans l'enrichissement de certains termes d'ontologie, ou autrement dit, une robustesse de la prédiction de l'implication des miARN dans certaines fonctions biologiques. En effet, dans les deux cas, les enrichissements sont très similaires autant au niveau des termes retrouvés que des p-valeurs (Figure 61).

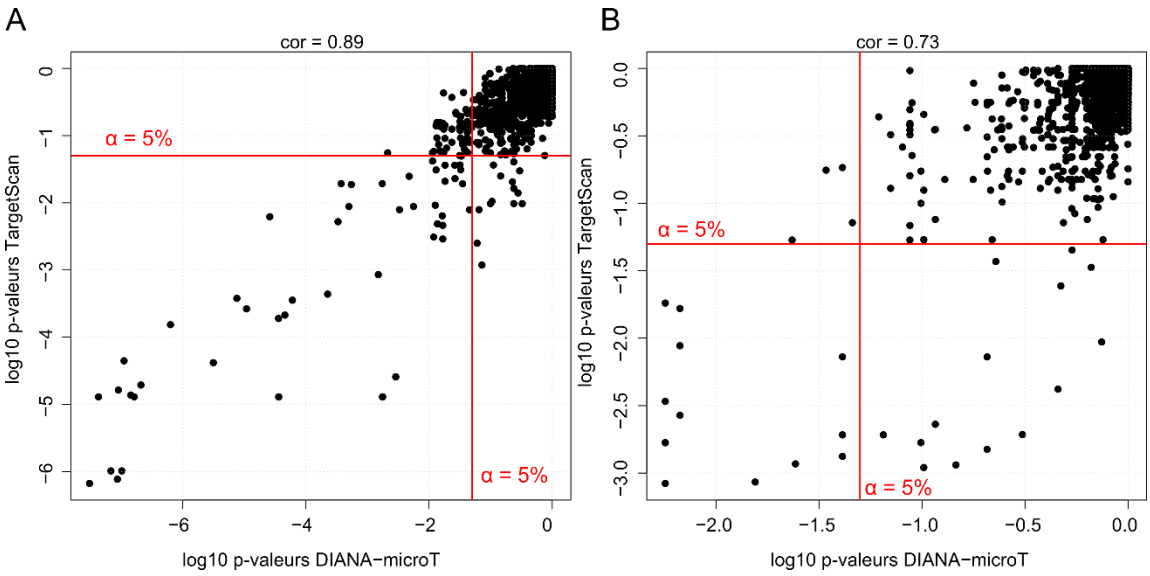


Figure 61. Corrélation entre les p-valeurs corrigées des annotations avec TargetScan contre celles de DIANA-microT. A | Club assorti 1. B | Club assorti 2.

2. Robustesse des ontologies face aux faux positifs et faux négatifs

Du fait de la spécificité et de la sensibilité imparfaites des algorithmes de prédictions de cibles des miARN, une dernière question concernant nos prédictions réside dans leur robustesse face aux « fausses » prédictions ; c'est-à-dire face aux faux positifs et aux faux négatifs. Nous avons donc étudié cette dernière pour le club assorti 2 en faisant varier aléatoirement les prédictions de façon indépendante et à quatre niveaux différents pour

chaque catégorie : 5%, 10%, 20% et 30%. Par exemple, pour étudier l'effet qu'aurait 5% supplémentaire de faux positifs sur les prédictions, nous avons ajouté 5% de cibles (choisi au hasard parmi l'ensemble des gènes connus et non déjà prédits pour le(s) miARN(s) en question) puis recalculé les enrichissements. Pour les faux négatifs, nous avons retiré un certain pourcentage de gènes prédits pour être réguler par les miARN. En répétant l'opération un grand nombre de fois, on obtient un histogramme.

La Figure 62 montre cette étude principalement pour les dix meilleures annotations du club assorti 2. Sur le panel A, nous pouvons observer l'évolution de la p-valeur pour le meilleur terme du club assorti 2 (Tableau 9) face au retrait de gènes au hasard dans l'ensemble de gènes partagés par les trois miARN. Cette dernière reste assez stable jusqu'à 20% de faux négatifs où une baisse assez significative peut alors être observée (trois ordres de grandeur). Concernant le taux de faux positifs, le schéma reste globalement le même avec des effets plus marqués à partir de 20% (Figure 62 B). Comme les p-valeurs dépendent également du nombre de gènes cibles, il est également intéressant de considérer les rangs dans les annotations plutôt que leur significativité. Les panels C et D montre l'évolution du rang des cinq meilleures annotations pour le club assorti 2 sous les effets des taux de faux négatifs (C) et de faux positifs (D). En général, les différentes annotations gardent des rangs sensiblement identiques jusqu'environ 20%, niveau à partir duquel une grande variabilité commence à apparaitre. C'est également le cas pour les cinq meilleures annotations suivantes (6 à 10) visualisables sur les panels E et F. Nous pouvons toutefois constater que même à 30%, les dix meilleures annotations sont tout de même souvent retrouvées parmi les vingt meilleures annotations. En revanche, les cinq meilleures annotations ont plutôt tendance à se retrouver plus haut dans la liste lorsque le taux de fausses prédictions augmente.

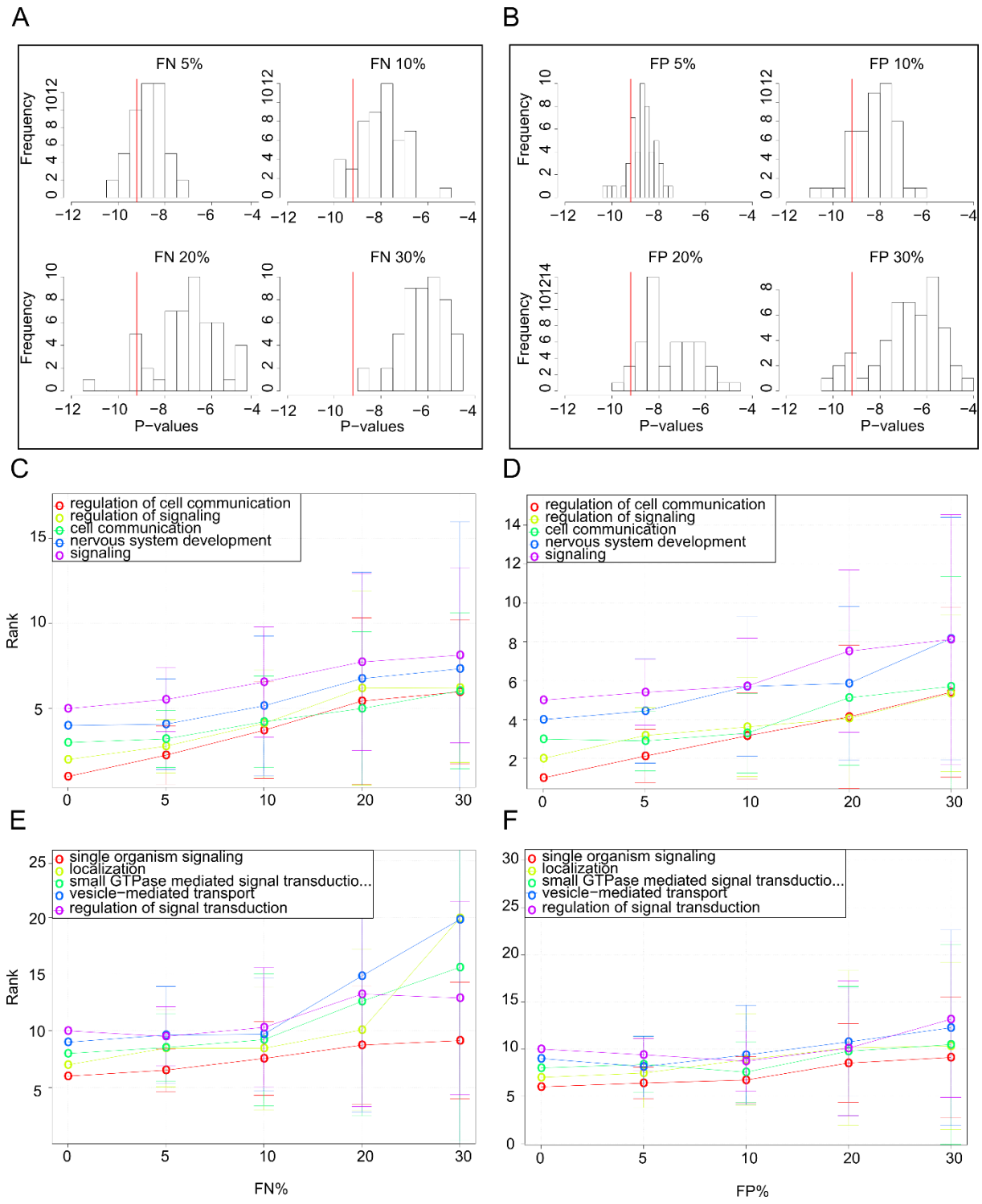


Figure 62. Robustesse des ontologies face aux fausses prédictions. A | p-valeurs du terme le plus enrichi pour le club assorti 2 après retrait de 5, 10, 20 et 30% de cibles (faux négatifs). B | p-valeurs du terme le plus enrichi pour le club assorti 2 après ajout de 5, 10, 20 et 30% de cibles (faux positifs). Barre rouge : p-valeur sans changement | Rang des cinq meilleures annotations pour le club assorti 2 et suivi de ces rang en fonction des faux négatifs. D | Rang des cinq meilleures annotations pour le club assorti 2 et suivi de ces rangs en fonction des faux positifs. E | Suite des meilleures annotations (6 à 10) pour le club assorti 2 et suivi de ces rangs en fonction des faux négatifs. F | Suite des meilleures annotations (6 à 10) pour le club assorti 2 et suivi de ces rang en fonction des faux positifs. 250 changements aléatoires pour chaque point d'étude. Sur C, D, E et F sont représentés la moyenne des 250 permutations et la déviation standard.

Ces résultats indiquent donc une certaine robustesse des analyses d'ontologie face

aux prédictions des cibles. De façon très intéressante, il semble assez rare d'augmenter artificiellement la p-valeurs même lorsque beaucoup de gènes sont rajoutés dans les tests (30% de faux positifs typiquement). Les enrichissements que nous observons pour le club assorti 2 sont donc probablement réellement dus aux gènes annotés pour les fonctions biologiques et non pas simplement un effet du nombre de gènes testés.

G. Conclusions et discussion

Nous avons pu mettre en évidence deux clubs assortis et prédire leur(s) implication(s) biologique(s) ainsi que leur rôle d'influence sur les miARN les entourant. Cette analyse globale permet notamment de donner une information sur les fonctions biologiques les plus probablement corégulées par les clubs assortis et également les autres miARN du réseau. En revanche, l'analyse ne peut pas donner en soi d'informations sur les fonctions très spécifiques de chacun des miARN. Nos prédictions s'arrêtent donc forcément à une vision systémique de la régulation par les miARN. Nous voyons ainsi qu'il existe trois grandes communautés de miARN : ceux plutôt impliqués dans la régulation de la signalisation, ceux impliqués dans la régulation transcriptionelle et ceux entre les deux ne montrant pas d'enrichissement particulier. Par ailleurs nous avons pu tester expérimentalement certaines de nos prédictions sur un des deux clubs.

Peu d'informations étaient disponibles au moment de l'écriture du premier manuscrit pour le deuxième club, composé de miR-612, -661 et -940,. L'analyse GO nous a permis de prédire une implication des trois miARN dans la transduction des signaux par les petites GTPases. Cette hypothèse impliquait une possibilité pour les trois miARN d'influencer le cytosquelette et la motilité cellulaire. Après validation fonctionnelle *in vitro*, nous avons pu confirmer ces différentes hypothèses en montrant que les miARN agissent sur le cytosquelette au travers de la phosphorylation de MLCII, un élément clé dans le contrôle du cytosquelette. De façon plus étonnante, la surexpression ectopiques des miARN montrait des phénotypes différents sur les cellules RPE1 puisque miR-661 entrainait une modification de la distribution

spatiale de la phosphorylation contrairement aux miR-612 et miR-940 qui baissait les niveaux de phosphorylation sans influencer fortement leur distribution. Ce phénomène antagoniste a également pu être confirmé par des expériences d'invasion permettant donc de confirmer l'implication du club assorti 2 dans la voie de signalisation des petites GTPases, la régulation du cytosquelette d'actine, la motilité cellulaire ainsi que l'invasion. Les mécanismes moléculaires impliqués dans ces comportements antagonistes devraient être plus finement caractérisés dans le futur.

En corrélation avec nos résultats, Tao et collaborateurs ont montré que miR-612 avait un effet inhibiteur dans les carcinomes hépatiques autant sur la prolifération, la migration, l'invasion et la métastase. De plus, le miARN semble avoir un effet sur les étapes initiales et finales de la cascade métastatique en réprimant des invasions locales et les colonisations distales (Tao et al., 2013). De façon similaire, miR-661 a été montré comme impliqué dans l'invasion des cellules cancéreuses mammaires en ciblant spécifiquement les gènes Nectin-1 et StarD10 (Vetter et al., 2010). Au moment des premières expériences sur le club assorti, aucune information n'était reportée pour miR-940. Depuis, plusieurs auteurs ont pu montrer pour miR-940 des rôles similaires à ceux que nous avons montrés. Par exemple, Rajendiran et collaborateurs ont prouvé que miR-940 est capable de supprimer la migration et l'invasion des cellules cancéreuses de prostates en contrôlant l'expression de MIEN1 (Rajendiran et al., 2014). Le miARN est par ailleurs surexprimé dans les tissus normaux et sous-exprimé dans les tissus tumoraux, plaçant donc miR-940 comme un outil potentiel de diagnostic et de pronostic pour le cancer de la prostate. De la même manière, Ma et collaborateur ont mis en évidence des résultats similaires pour le carcinome nasopharyngé (Ma et al., 2014). MiR-940 semble par ailleurs également impliqué dans les adénocarcinomes pancréatiques (Song et al., 2015) et hépatiques (Yuan et al., 2015) toujours selon le même schéma, c'est à dire en réprimant la progression tumorale. Nous restons cependant les premiers à avoir montré un rôle potentiel du miARN dans le cancer du sein.

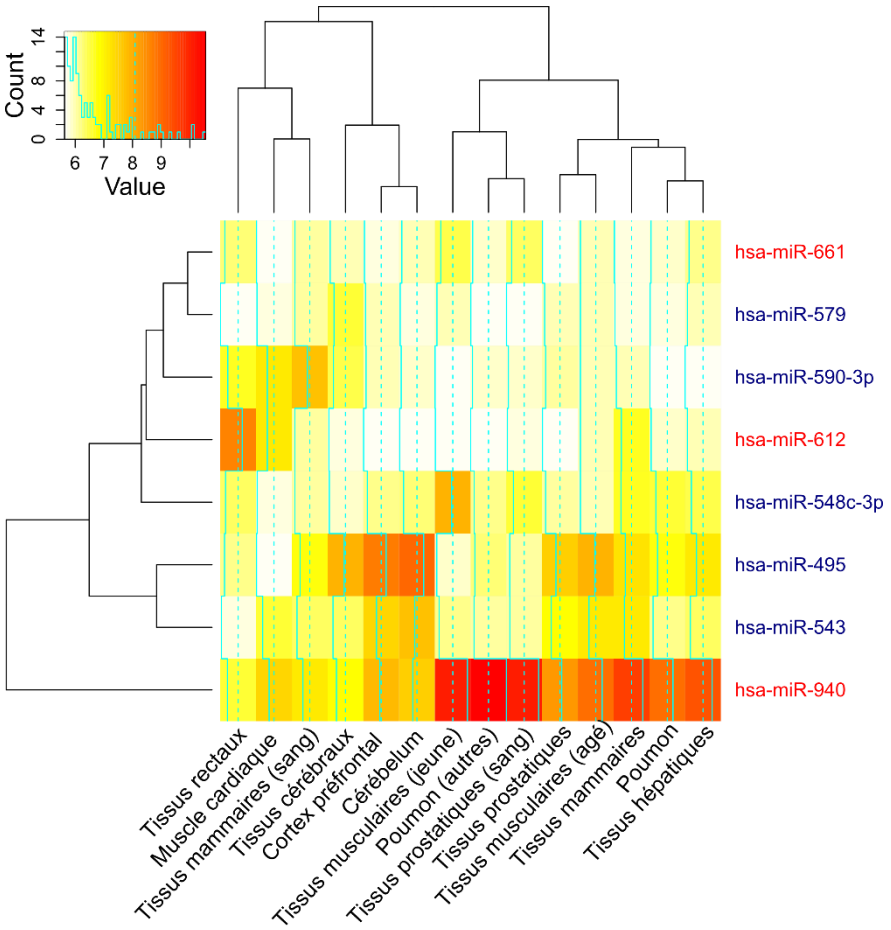


Figure 63. Expression des membres des clubs assortis dans différents tissus (cf. page 100). Seuls les miARN dont l'expression était retrouvée sur l'ensemble des tissus analysés sont présentés. Soit 8 miARN sur les 11 formant les clubs assortis. En bleu sont représentés les membres du club assorti 1 et en rouge, ceux du club assorti 2.

Concernant l'expression des membres des clubs assortis, nous pouvons constater que les miARN ne sont, en règle générale, pas coexprimés – il semblerait d'ailleurs que pour la plupart des tissus, seul un membre des deux clubs soit fortement exprimé à la fois (Figure 63). MiR-940 semble particulièrement exprimé dans les épithéliums glandulaires alors que miR-495, quant à lui, est plutôt exprimé dans différentes régions du cerveau. Enfin, dans les tissus rectaux, c'est miR-612 qui semble prendre le relai. En revanche, et contrairement aux résultats sur l'ensemble des miARN, la classification tissus glandulaire/autres tissus semble moins présente avec ces onze miARN uniquement.

Les clubs assortis définissent deux sphères séparées par une zone intermédiaire représentant les trois types de miARN déjà évoqués. Une grande corrélation existant entre les enrichissements des clubs assortis et leur sphère respective, nous avons donc nommé les

deux sphères « zone d'influence des clubs assortis ». Une idée importante derrière toutes ces analyses est la notion d'exploration globale se basant sur les gènes partagés. En effet, bien que les deux sphères soient impliquées l'une et l'autre dans différentes fonctions cellulaires – prédites à partir des gènes partagés, certains miARN des zones pourraient tout de même ne pas avoir d'implication dans ces fonctions. Il existe en fait simplement une plus grande probabilité que ces miARN soient impliqués dans ces processus. En gardant donc à l'esprit que notre analyse se limite à des aspects globaux de la régulation biologique par les miARN, nous pouvons noter un certain lien entre la forme du réseau (les deux clubs assortis et les différentes zones du réseau) et les fonctions biologiques partagées par les miARNs (signalisation et transcription). Par ailleurs, nous avons également vu que la robustesse ne se limite pas à la construction des réseaux et aux liens entre les miARN mais s'étend également à l'analyse d'ontologie puisque peu importe l'algorithme, les prédictions restent identiques. Nous pouvons également noter qu'en ajoutant des fausses prédictions aux ensembles de données, les résultats obtenus par les analyses d'ontologie restent plutôt stables.

.

Chapitre 3 :
Analyse d'une communauté de miARN impliquée dans la pluripotence

A. Introduction

Les cellules souches sont capables soit de donner naissance à de nombreux types cellulaires ou de s'auto-renouveler. Elles peuvent être pluripotentes et se différencier dans tous les types cellulaires d'un tissu (p.ex. les cellules souches hématopoïétiques) ou totipotentes (p.ex. les cellules souches embryonnaires) et recréer un organisme entier, avec toutes ses composantes cellulaires.

Les cellules souches embryonnaires (cellules totipotentes issues de la masse cellulaire interne du blastocyste) et pluripotentes induites (cellules pluripotentes dédifférenciées à partir de cellules somatiques) (Takahashi and Yamanaka, 2006) sont aujourd'hui des cellules très étudiées pour leur potentiel thérapeutique (Kehat et al., 2001; Wobus and Boheler, 2005; Ludwig et al., 2006; Tabar and Studer, 2014). Par exemple, une récente étude singulièrement frappante est celle de Chong et collaborateurs qui ont réussi à repeupler un cœur de primate à partir de cellules souches embryonnaires humaines, montrant par la même occasion tout le potentiel de ces cellules dans les thérapies cellulaires (Chong et al., 2014). Ces dernières possèdent donc un fort potentiel thérapeutique clinique : leur étude et la compréhension de leur régulation relève d'un fort intérêt scientifique, économique et social.

Comme nous l'avons vu dans l'introduction, les miARN ont été premièrement découverts pour leur rôle dans le développement chez *C. elegans* (Lee et al., 1993; Wightman et al., 1993; Pasquinelli et al., 2000). Il n'est donc pas étonnant que ces derniers aient rapidement été identifiés comme des régulateurs des cellules souches (Houbaviy et al., 2003; Laurent et al., 2008; Marson et al., 2008; Melton et al., 2010; Stadler et al., 2010; Yi and Fuchs, 2011; Li and He, 2012). Ainsi, plusieurs familles de miARN sont souvent caractérisées pour leur implication dans la reprogrammation cellulaire : c'est particulièrement le cas de la famille miR-302 et du cluster miR-17-92 (Subramanyam et al., 2011; Parchem et al., 2014). En particulier, l'étude de Anokye-Danso et collaborateurs a prouvé que l'utilisation de miARN par gain de fonction – notamment du cluster miR-302/367 – en lieu et place du fameux cocktail de dédifférenciation formé par les facteurs de transcription Oct4, Sox2, Klf4 et c-Myc et mis au

point par l'équipe de Yamanaka en 2006, permettait un bien meilleur rendement de cellules dédifférenciées (Anokye-Danso et al., 2011).

L'objectif de ce chapitre est de présenter une communauté de miARN impliqués dans la régulation des cellules souches, communauté que nous avons identifiée à partir de l'inférence des réseaux de miARN basés sur le partage de cibles. Afin de garder un maximum d'informations dans l'inférence de cette communauté, c'est l'union des résultats sur les deux algorithmes DIANA-microT et TargetScan qui a été utilisée. Plus précisément, différents algorithmes de détection des communautés ont été appliqués sur chaque réseau, permettant d'obtenir différentes partitions des réseaux. Dans chacun de ces partitionnements, une communauté en particulier contenait toujours miR-17 et les miARN de la famille miR-302 : c'est la combinaison des sous-graphes contenant ces miARN qui a formé la « communauté souche » identifiée ici. Dans ce chapitre, nous verrons donc dans un premier temps le processus de découverte de la communauté, puis nous étudierons l'expression des miARN de la communauté dans des cellules souches pour enfin nous attarder sur l'influence de ces derniers sur le potentiel de différenciation (ou, au contraire, le potentiel de maintien du caractère souche) des cellules souches.

B. Identification de la communauté souche

Afin d'identifier une communauté de miARN impliquée dans les cellules souches, nous nous sommes intéressés aux deux réseaux inférés à partir de TargetScan et de DIANA-microT, avec les seuils déjà évoqués jusqu'à présent : meet/min 50% pour le réseau DIANA-microT et 54% pour le réseau TargetScan. Trois algorithmes de détection de communautés

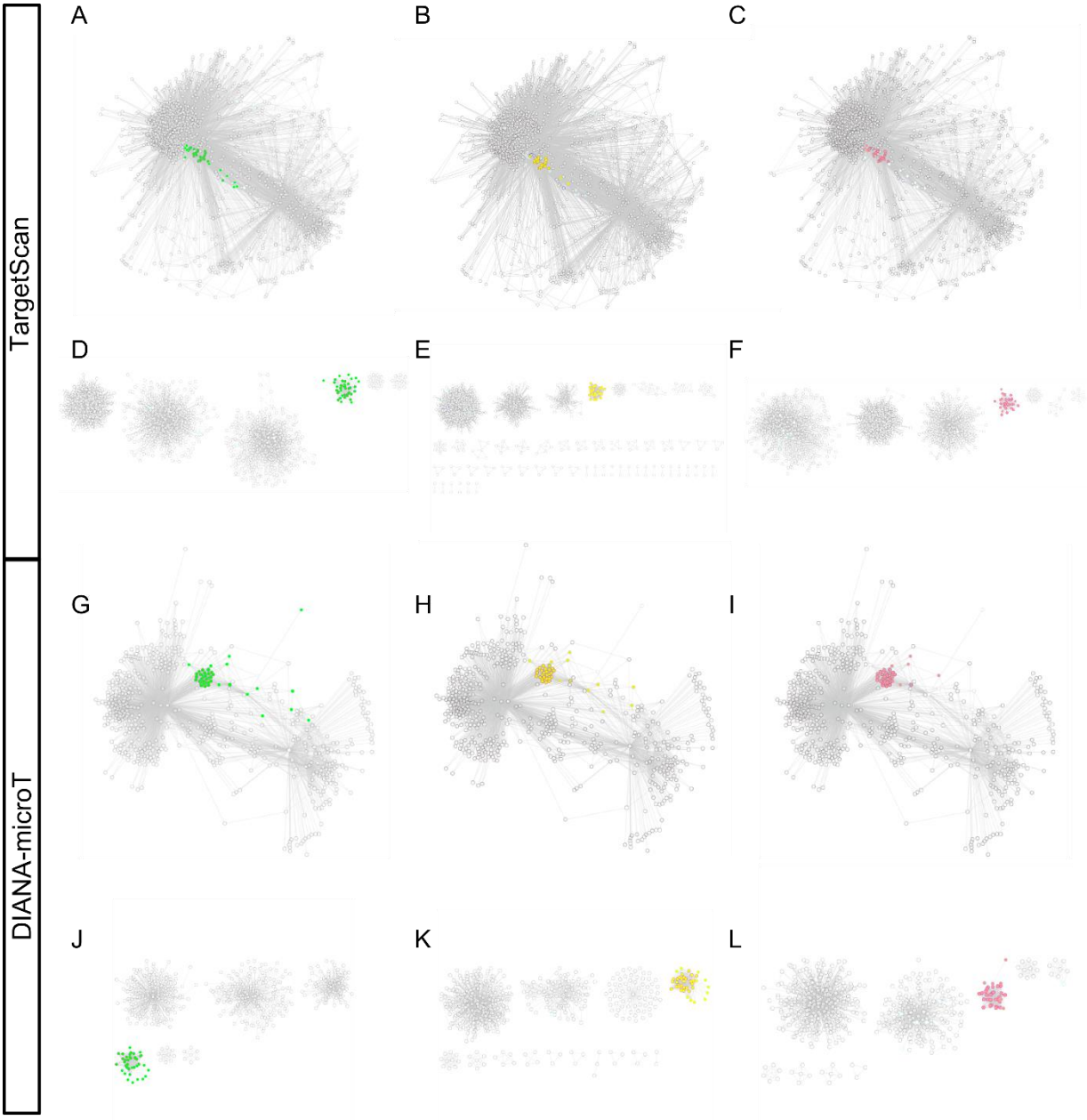


Figure 64. Différents partitionnement pour les deux réseaux DIANA-microT et TargetScan. Les deux premières lignes (A à F) montrent les partitionnements sur le réseau TargetScan et les deux dernières (G à L), ceux sur le réseau DIANA-microT. Les lignes D, E, F et J, K, L montrent les partitions obtenues avec les trois algorithmes pour TargetScan et DIANA-microT respectivement. La première et la troisième ligne montrent la communauté souche dans le réseau entier. Les nœuds verts sont associés à la communauté souche retrouvée avec l'algorithme *multilevel*, ceux en jaune à l'algorithme *fast greedy* et, enfin, ceux en rose à l'algorithme *edge betweenness*. Ces communautés sont définies comme celles contenant les membres de la famille miR-302 et miR-17.

différents ont alors été utilisés sur les deux réseaux indépendamment, menant donc à six partitionnements différents : trois pour chaque réseau (Figure 64).

Nous pouvons constater que les algorithmes détectent généralement trois grandes communautés constituées de plus de 100 miARN (les trois communautés visibles sur la gauche des panels D, E, F, J, K et L). Ces trois communautés correspondent en fait aux trois zones évoquées dans les précédents chapitres (zone d'influence des clubs assortis et zone intermédiaire). Les algorithmes détectent ensuite d'autres communautés plus petites en nombre variable et chacune composée de 2 à 40 miARN en fonction des algorithmes de

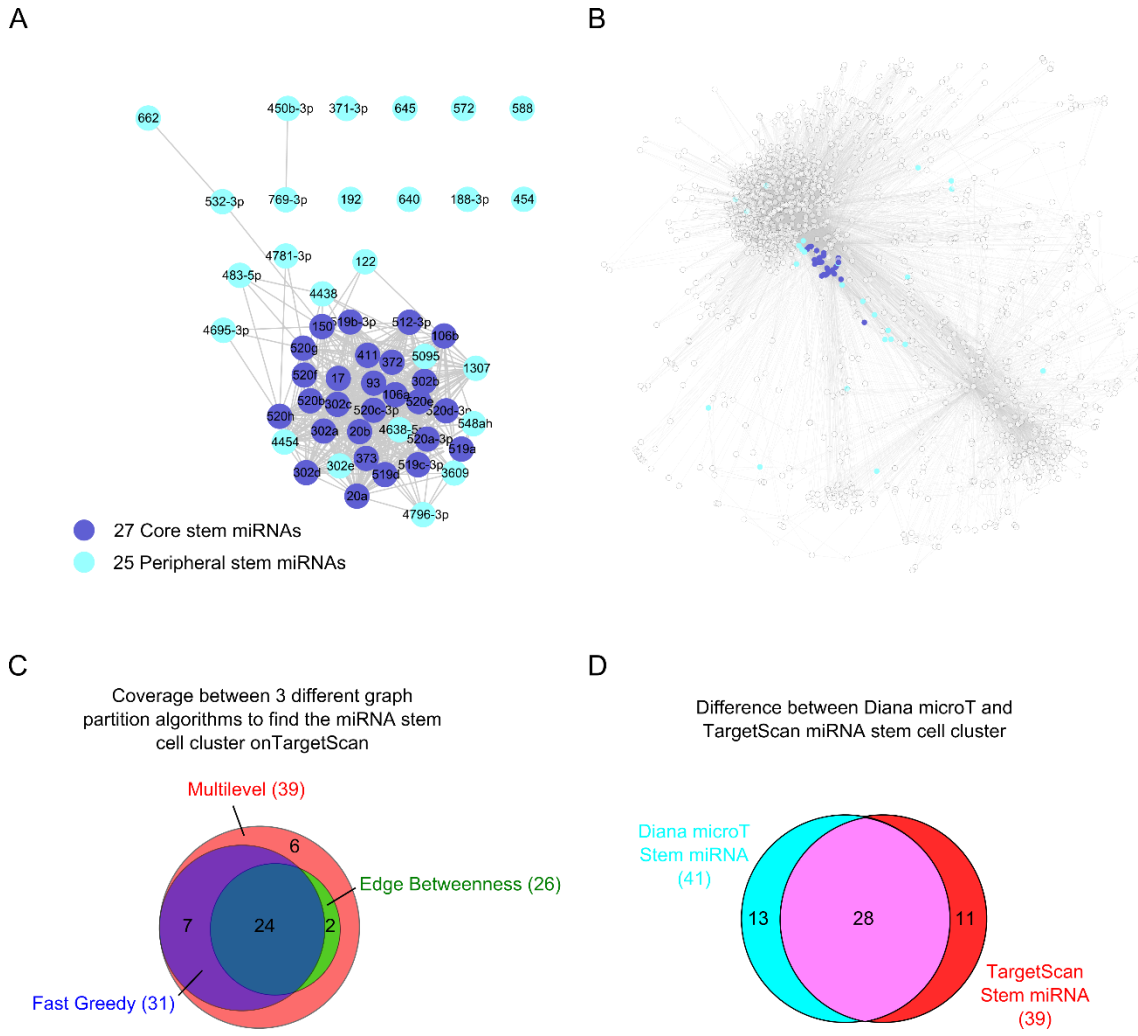


Figure 65. La communauté souche inférée à partir de TargetScan. A | Union de l'ensemble des miARN retrouvés dans les six communautés contenant les miR-302 et miR-17. Sous-graphe induit par ces miARN depuis TargetScan. Les nœuds non connectés sont des miARN qui ne sont pas retrouvés dans les partitions de TargetScan mais uniquement celles de DIANA-microT. Le code couleur indique le nombre de fois qu'un miARN est retrouvé parmi les six communautés (4 fois ou plus pour le « *core* »). B | Réseau TargetScan avec un seuil meet/min de 54% et mise en évidence de la communauté souche complète. C | Recouvrement de miARN entre les trois communautés du réseau TargetScan issues des trois algorithmes *multilevel*, *fast greedy* et *edge betweenness*. D | Recouvrement entre les miARN retrouvés dans les trois communautés de TargetScan et ceux retrouvés dans DIANA-microT.

partitionnement. L'objectif était alors de chercher dans ces partitions les communautés contenant les miR-302a/b/c/d et miR-17 (Suh et al., 2004; Laurent et al., 2008; Stadler et al., 2010). Ces partitions sont indiquées en couleur sur la Figure 64 et possèdent des densités comprises entre 0,53 et 0,71. Nous avons alors nommé « communauté souche » l'union des miARN composant ces six partitions.

Pour mieux appréhender la communauté souche et sa stabilité, nous avons par la suite compté le nombre de fois qu'un miARN était présent dans les communautés des six partitions.

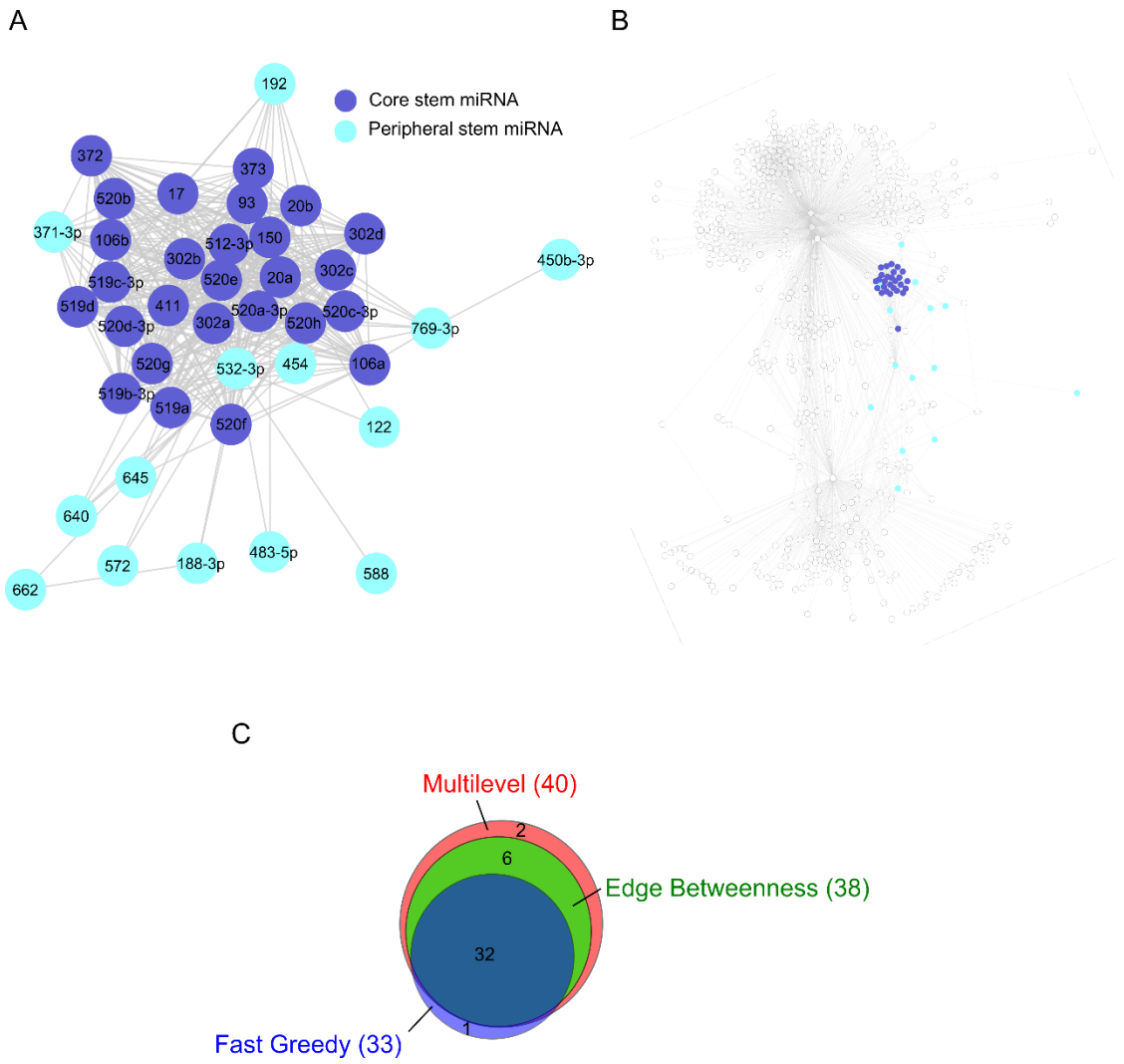


Figure 66. La communauté souche inférée à partir de DIANA-microT. A | Union de l'ensemble des miARN retrouvés dans les six communautés contenant les miR-302 et miR-17. Sous-graphe induit par ces miARN depuis DIANA-microT. Les nœuds non connectés sont des miARN qui ne sont pas retrouvés dans les partitions de DIANA-microT mais uniquement celles de TargetScan. Le code couleur indique le nombre de fois qu'un miARN est retrouvé parmi les six communautés (4 fois ou plus pour le « *core* »). B | Réseau DIANA-microT avec un seuil meet/min de 50% et mise en évidence de la communauté souche complète. C | Recouvrement de miARN entre les trois communautés du réseau TargetScan issues des trois algorithmes *multilevel*, *fast greedy* et *edge betweenness*.

Les Figure 65 et Figure 66 montrent l'inférence de cette communauté à partir de TargetScan ou de DIANA-microT respectivement. Vingt-sept miARN ont été retrouvés dans quatre ou plus de partitions différentes. Etant donné leur grande stabilité face aux algorithmes de détection de communautés, ces miARN ont été nommés « miARN souches cœurs » et sont indiqués en bleu foncé sur les deux figures. Le sous-graphe induit par ces miARN possède une densité de 1 que ce soit sur TargetScan ou DIANA-microT. Vingt-cinq miARN ont été retrouvés dans moins de trois partitions, ces miARN ont donc été appelés « miARN souches périphériques » (Figure 65 A et B et Figure 66 A et B). Parmi ces vingt-cinq miARN, onze miARN sont limités à TargetScan, notamment miR-1307, miR-302e, ou encore miR-3609 et tous ceux ayant un identifiant supérieur à ce dernier (p.ex. -4438, -4454, etc.)

Les recouvrements entre les miARN récupérés par les différents algorithmes sur TargetScan et DIANA-microT sont représentés sur les panels C des deux figures. L'algorithme *multilevel* englobe généralement les résultats des deux autres algorithmes : les miARN retrouvés par les algorithmes *fast greedy* et *edge betweenness* sont aussi retrouvés par *multilevel*. Un seul miARN (miR-450b-3p) est spécifique à l'algorithme *fast greedy* sur sa partition de DIANA-microT (Figure 66 C). Vingt-huit miARN sont retrouvés en commun entre les partitions de TargetScan et DIANA-microT. Onze miARN sont limités à TargetScan et treize, à DIANA-microT (Figure 65 D) sachant que 989 miARN de TargetScan ne sont pas présents dans la base de données DIANA-microT.

Tableau 14. miARN de la communauté souche. En rouge sont indiqués les miARN permettant de définir les communautés souches ; en bleu, ceux retrouvés uniquement avec DIANA-microT ; en vert, ceux retrouvés uniquement avec TargetScan et en noir, ceux retrouvés dans les deux bases de données.

miARN cœurs			**miARN périphériques**		
MIMAT	**miARN**	**Nombre d'apparition**	**MIMAT**	**miARN**	**Nombre d'apparition**
MIMAT0000103	hsa-miR-106a	6	MIMAT0000421	hsa-miR-122	3
MIMAT0000680	hsa-miR-106b	6	MIMAT0005951	hsa-miR-1307	2
MIMAT0000451	hsa-miR-150	4	MIMAT0004613	hsa-miR-188-3p	2
MIMAT0000070	**hsa-miR-17**	6	MIMAT0000222	hsa-miR-192	3
MIMAT0000075	hsa-miR-20a	6	**MIMAT0005931**	**hsa-miR-302e**	3
MIMAT0001413	hsa-miR-20b	6	MIMAT0017986	hsa-miR-3609	2
MIMAT0000684	**hsa-miR-302a**	6	MIMAT0000723	hsa-miR-371-3p	3
MIMAT0000715	**hsa-miR-302b**	6	MIMAT0018956	hsa-miR-4438	1
MIMAT0000717	**hsa-miR-302c**	6	MIMAT0018976	hsa-miR-4454	2
MIMAT0000718	**hsa-miR-302d**	6	MIMAT0004910	hsa-miR-450b-3p	1
MIMAT0000724	hsa-miR-372	6	MIMAT0003885	hsa-miR-454	2
MIMAT0000726	hsa-miR-373	6	MIMAT0019695	hsa-miR-4638-5p	1

MIMAT0003329	hsa-miR-411	6		MIMAT0019789	hsa-miR-4695-3p	2
MIMAT0002823	hsa-miR-512-3p	6		MIMAT0019943	hsa-miR-4781-3p	1
MIMAT0002869	hsa-miR-519a	6		MIMAT0019971	hsa-miR-4796-3p	2
MIMAT0002837	hsa-miR-519b-3p	5		MIMAT0004761	hsa-miR-483-5p	2
MIMAT0002832	hsa-miR-519c-3p	5		MIMAT0020600	hsa-miR-5095	1
MIMAT0002853	hsa-miR-519d	6		MIMAT0004780	hsa-miR-532-3p	2
MIMAT0002834	hsa-miR-520a-3p	6		MIMAT0018972	hsa-miR-548ah	2
MIMAT0002843	hsa-miR-520b	6		MIMAT0003237	hsa-miR-572	2
MIMAT0002846	hsa-miR-520c-3p	6		MIMAT0003255	hsa-miR-588	1
MIMAT0002856	hsa-miR-520d-3p	6		MIMAT0003310	hsa-miR-640	3
MIMAT0002825	hsa-miR-520e	6		MIMAT0003315	hsa-miR-645	3
MIMAT0002830	hsa-miR-520f	5		MIMAT0003325	hsa-miR-662	1
MIMAT0002858	hsa-miR-520g	6		MIMAT0003887	hsa-miR-769-3p	3
MIMAT0002867	hsa-miR-520h	6				
MIMAT0000093	hsa-miR-93	6				

Au final, cinquante-deux miARN composent la communauté souche que nous avons identifiée avec deux niveaux de stabilité différents : vingt-sept miARN cœurs et vingt-cinq miARN périphériques (Tableau 14). Ces miARN représentent l'union des communautés contenant les miARN miR-302 et miR-17 issues des six partitionnements différents. Par ailleurs, les membres de ces deux familles font partie du cœur de la communauté exception faite de miR-302e.

C. Alignement de *seed*

Sachant que chacun des deux algorithmes, TargetScan et DIANA-microT, prédit les cibles des miARN en se basant sur leur séquence *seed*, une question particulièrement importante est l'effet de la séquence *seed* sur l'identification de la communauté souche. En effet, à cause de cet effet de similarité de séquence, les groupements dans les réseaux peuvent se faire selon deux manières différentes : sur des cibles identiques à cause de séquences identiques (famille let-7) ou alors sur des cibles communes mais des séquences différentes. Ainsi, l'alignement des séquences *seed* (octamer, position 2 à 9 des séquences matures de miARN dans ce cas-là) des cinquante-deux miARN permet d'appréhender cet effet *seed* sur les regroupements dans les réseaux.

La Figure 67 montre cette analyse réalisée avec le programme Jalview (Clamp et al., 2004) et l'algorithme d'alignement multiple ClustalW (Thompson et al., 1994). En se basant sur les alignements de séquences *seed*, huit clusters différents peuvent être extraits pour quatre miARN isolés (différentes couleurs sur la Figure 67). De façon intéressante, un certain

enrichissement de séquence peut être observé notamment avec la séquence consensus AAGUGC, regroupant une vingtaine de miARN (groupe 9 en violet). Cette séquence consensus était déjà décrite par Laurent et collaborateurs dans leur étude en 2008 (Laurent

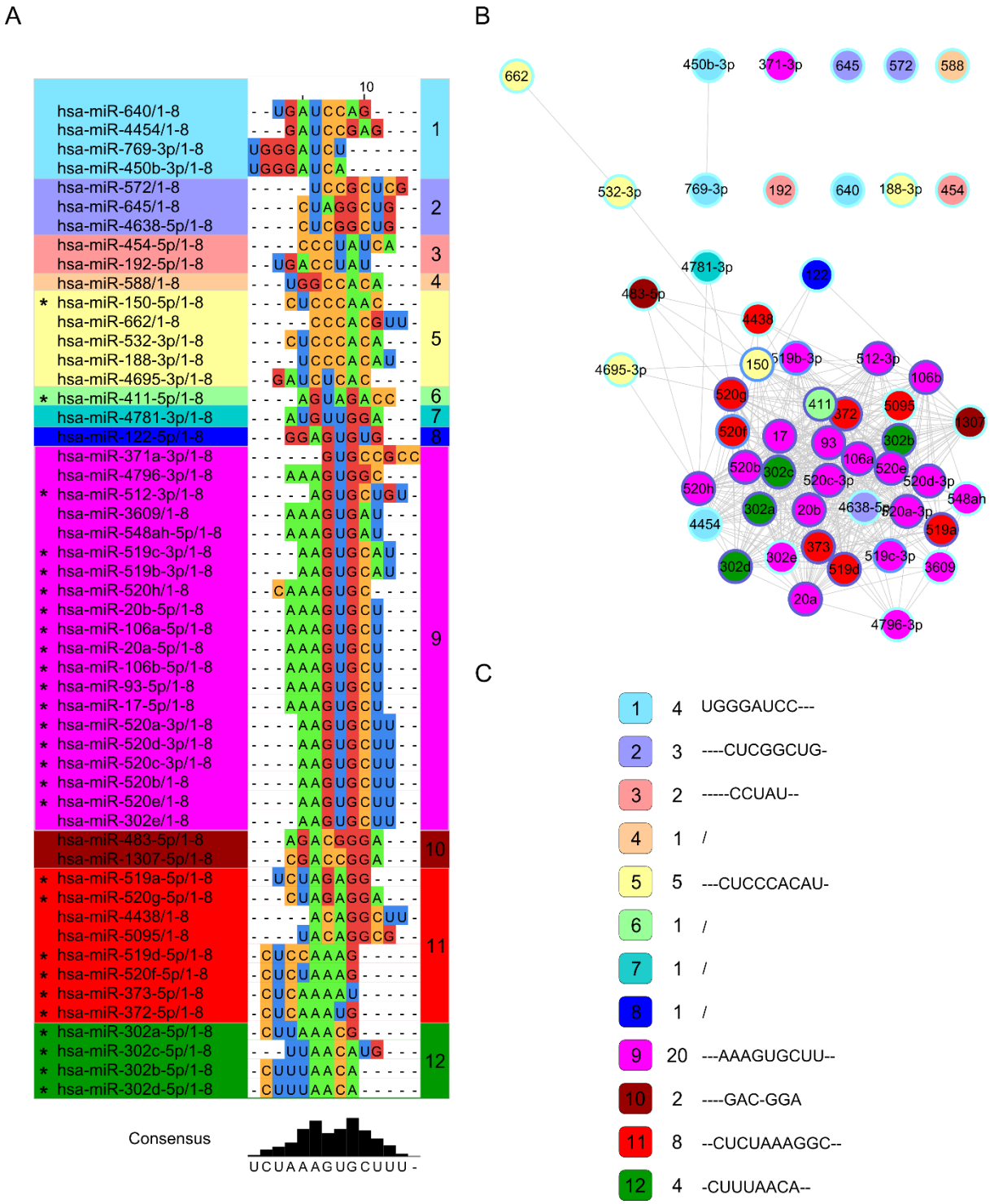


Figure 67. Alignement des séquences *seed* des miARN de la communauté souche. A | Alignement multiple des octamers *seed* (position 2 à 9) par le programme ClustalW. Les différents clusters ont été extraits à partir d'un arbre de similarité construit sur l'alignement. B | Communauté souche du sous-graphe de TargetScan. Le code couleur correspond aux clusters du panel A Le pourtour des nœuds indiquent leur appartenance aux deux groupes de miARN (cœur et périphérique). C | Séquence consensus des différents clusters. * = miARN du *core*.

et al., 2008). Le groupe est composé de membres du cluster 17 (Mendell, 2008) et de la famille miR-520. Le second plus gros cluster (groupe 11 en rouge) sur la figure comporte huit miARN dont miR-372, -373 et une autre partie de la famille miR-520 et possède la séquence consensus suivante : CUCUAAA. Quatre membres de la famille miR-302 (groupe 12 en vert) sont regroupés ensemble avec une séquence plus proche du deuxième groupe. Ces trois clusters forment globalement la partie cœur de la communauté souche (repérés par des « * » sur la Figure 67 A).

Le troisième cluster en taille est composé de cinq miARN (miR-150, -662, -532-3p, -188-3p et -4695-3p ; groupe 5 en jaune) et montre une séquence *seed* consensus proche des groupes 2 (bleu clair : miR-572, -645 et-4638-5p), 3 (rouge claire : miR-454, -192) et au miARN isolé miR-588. Les deux derniers clusters sont composés respectivement de miR-640, -4454, -769-3p et -450b-3p (groupe 1 en cyan) et miR-483 et -1307 (groupe 10 en brun). La plupart de ces miARN fait partie de la périphérie de la communauté souche.

En conclusion, si une partie des miARN est bien groupée à cause des séquences *seed*, nous pouvons constater que d'autres groupes sont également regroupés uniquement par le partage de cibles sans similarité évidente de séquence.

D. Expression des miARN de la communauté souche

Afin d'observer l'expression des miARN de la communauté dans des cellules souches (embryonnaires dans un premier temps et induites par la suite), deux ensembles de données d'expression (GSE14473 et GSE42446) ont été analysés (Stadler et al., 2010; Koyanagi-Aoi et al., 2013). Dans les deux cas, nous nous sommes intéressés au ratio d'expression entre cellules non différenciées et cellules différenciées. Pour l'étude GSE14473, les comparaisons ont été faites par paires entre les mêmes types de cellules (p.ex. cellules H1 dans un état différencié contre cellules H1 dans un état non différenciées). Dans le cas de l'étude GSE42446, les comparaisons ont été faites entre tous les types cellulaires non-différenciés et

tous les types cellulaires différenciés. Une étude par permutation pour calculer l'enrichissement en hits positifs a également été menée pour chacun des ensembles de données.

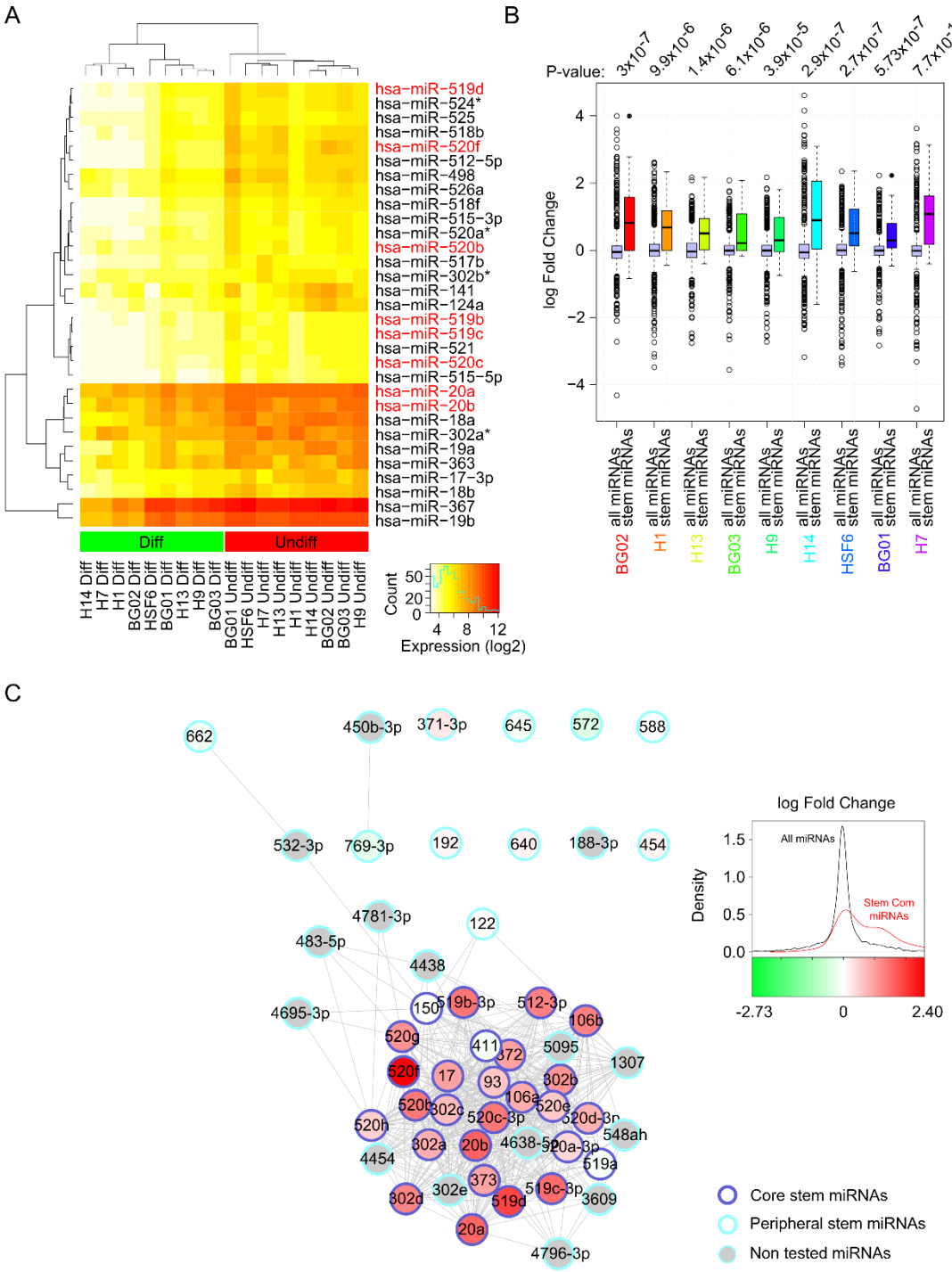


Figure 68. Expression de miARN dans différentes cellules souches (GSE14473). A | Expression relative des miARN les plus surexprimés dans l'état non différencié. En rouge sont mis en évidence les miARN de la communauté souche. B | Boxplot des expressions différentielles (logFC) sur les neuf types cellulaires souches étudiés. En gris est représenté l'expression de l'ensemble des miARN et en couleurs, celles de tous les miARN de la communauté souche. C | LogFC des miARN dans la communauté souche sur le sous-graphe de TargetScan.

1. GSE14473 : différents modèles de cellules souches embryonnaires

La Figure 68 A montre l'expression relative des miARN les plus surexprimés dans les cellules non différenciées par rapport aux cellules différenciées sur l'ensemble de données GSE14473. Parmi les trente miARN les plus différentiellement exprimés entre cellules souches et cellules différenciées sur l'ensemble de données, huit sont des membres de la communauté

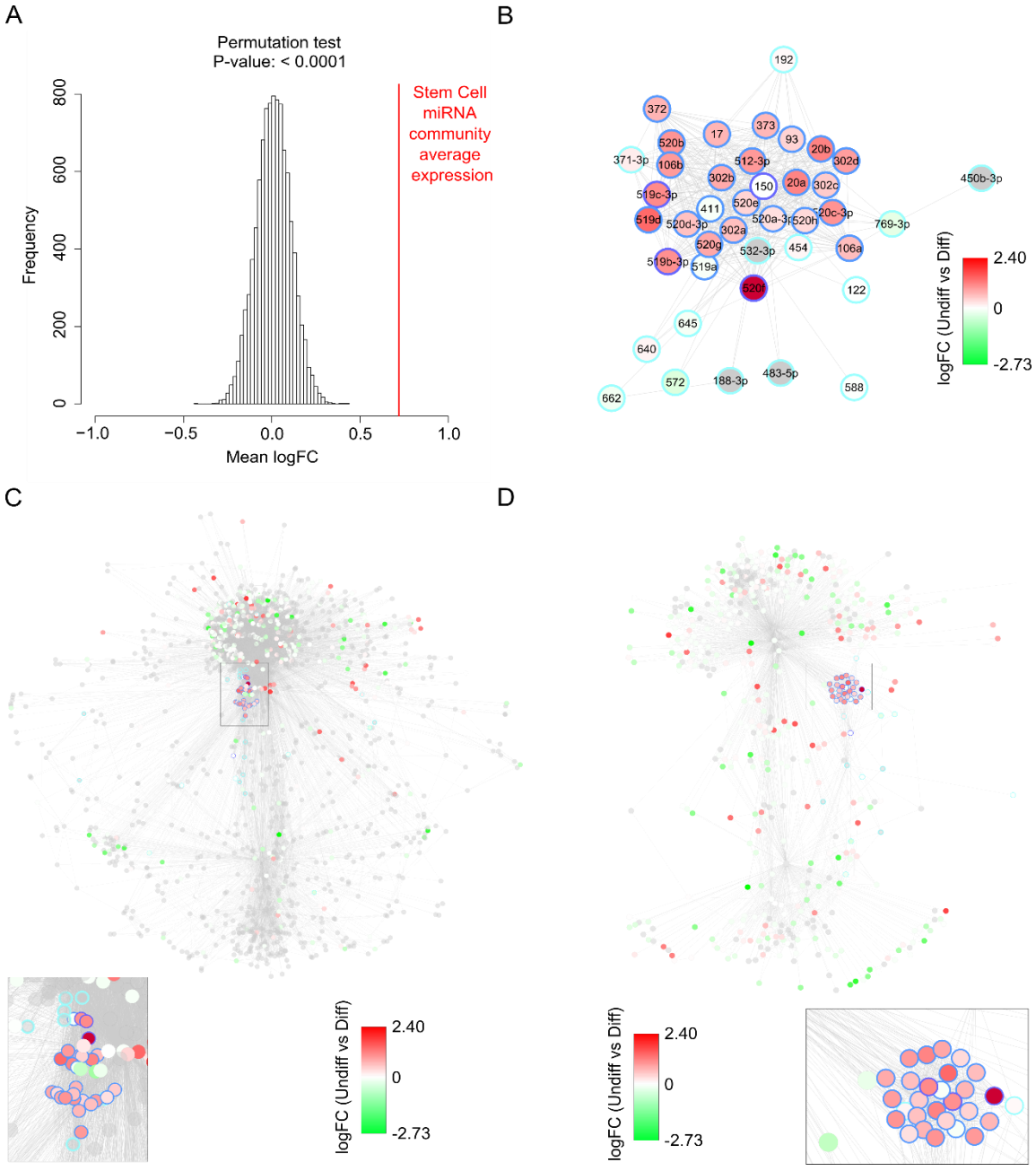


Figure 69. Enrichissement en hits positifs dans la communauté souche dans différentes cellules souches (GSE14473). A | Histogramme des valeurs moyennes de logFC. **B |** LogFC des miARN dans la communauté souche sur le sous-graphe de DIANA-microT. C et D | LogFC des miARN testés sur les réseaux de TargetScan et DIANA-microT respectivement. Les nœuds gris sont les nœuds n'ayant pas été testés dans l'ensemble de données.

souche que nous avons identifiée (noms surlignés en rouge sur la figure). En comparant les log *Fold Change* ($logFC_{miARN_i} = log2\left(ExpressionTissu1_{miARN_i}/ExpressionTissu2_{miARN_i}\right)$) pour chaque type cellulaire souche entre l'ensemble des miARN et ceux de la communauté souche uniquement, nous pouvons observer un clair enrichissement en hits positifs (déplacement de la distribution) dans la communauté souche (Figure 68 B). Par ailleurs, les p-valeurs associées à ces différences sont toutes situées entre 10^{-5} et 10^{-10}, confirmant ainsi que la plupart des miARN de la communauté souche sont globalement surexprimés dans les cellules pluripotentes. La valeur moyenne des logFC pour la communauté est de 0,72 avec une p-valeur de permutation de 10^{-04} (Figure 69, cf. procédure page 105).

2. GSE42446 : modèles de cellules souches embryonnaires et induites.

Des conclusions similaires peuvent être tirées à partir de l'ensemble de données GSE42446 (cellules souches embryonnaires et induites) avec encore une fois un enrichissement en miARN plus exprimés dans l'état indifférencié (Figure 70). Dans ce cas, neuf miARN parmi les plus différentiellement surexprimés font partie de la communauté souche (Figure 70 A), ces derniers ne sont cependant pas exactement les mêmes que précédemment. En l'absence de paires de types de cellules (p.ex. un type de cellule souche dans l'état différencié et le même type dans un état non-différencié), nous avons mené des comparaisons entre tous les types cellulaires souches et tous les types cellulaires somatiques. Sur le panel B de la Figure 70 est représenté une de ces comparaisons entre tous les types souches testés dans l'ensemble de données et des cellules DP74 (mélanocytes). La Figure 81 en annexe montre toutes les autres comparaisons effectuées et d'autres types de cellules somatiques. Dans toutes les configurations, les résultats sont similaires : les miARN de la communauté sont significativement surexprimés dans les cellules souches.

Les miARN de la communauté sont significativement surexprimés par rapport à l'ensemble des miARN testés : les p-valeurs sont effectivement très faibles.

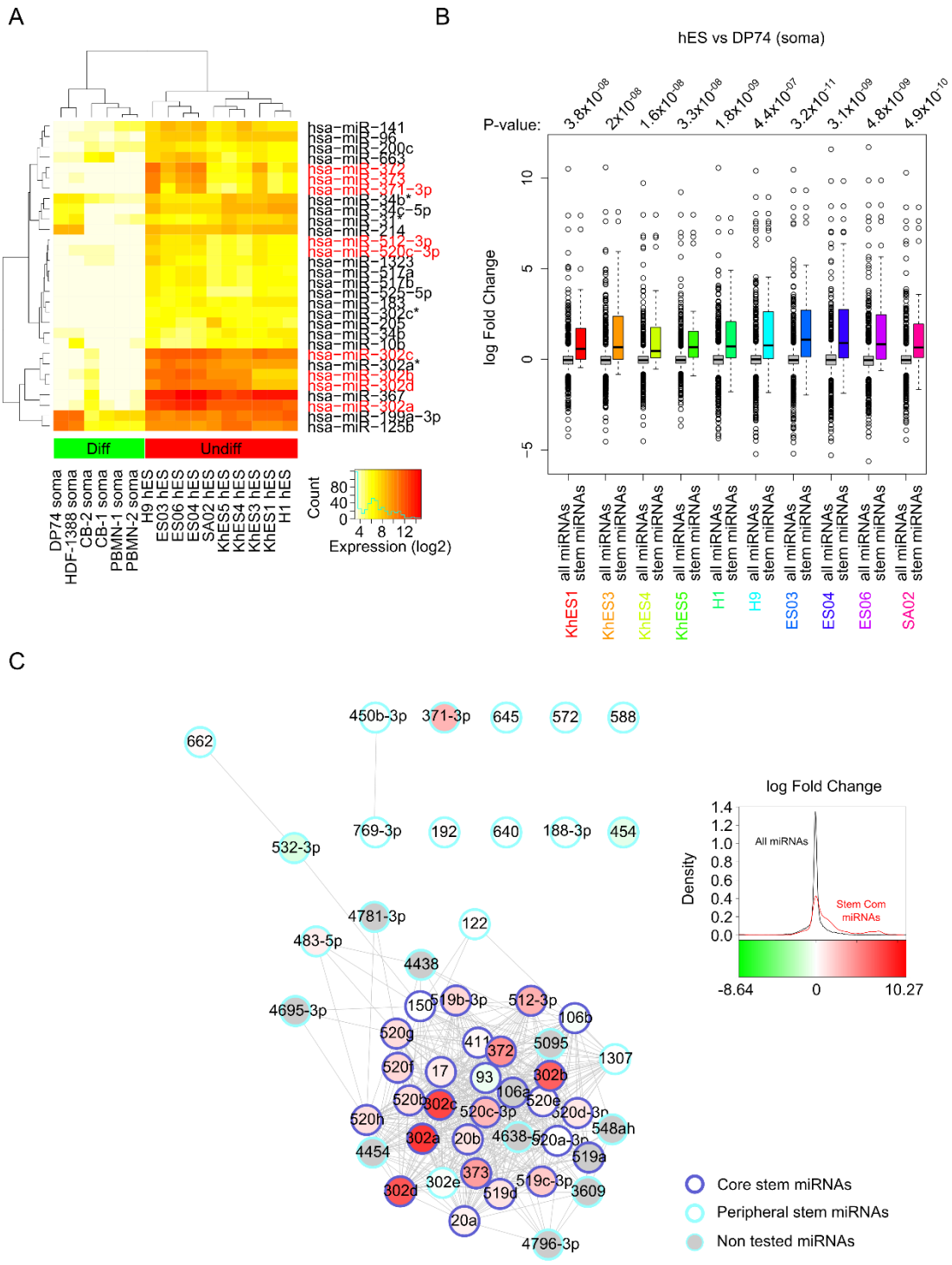


Figure 70. Expression de miARN dans différentes lignées souches (GSE42446). A | Expression relative des miARN les plus surexprimés dans l'état non différencié. En rouge sont mis en évidence les miARN de la communauté souche. B | Boxplot des expressions différentielles (logFC) sur dix types de cellules souches contre des cellules somatique DP74. En gris est représenté l'expression de l'ensemble des miARN et en couleurs, celle de tous les miARN de la communauté souche. C | LogFC des miARN dans la communauté souche sur le sous-graphe de TargetScan.

Remarquablement, bien que l'ensemble des miARN soient généralement surexprimés (Figure 70 C et Figure 71 B, C et D), nous pouvons constater que quatre miARN de la famille miR-302 dominent complètement en termes d'expression. La moyenne des logFC des miARN de la

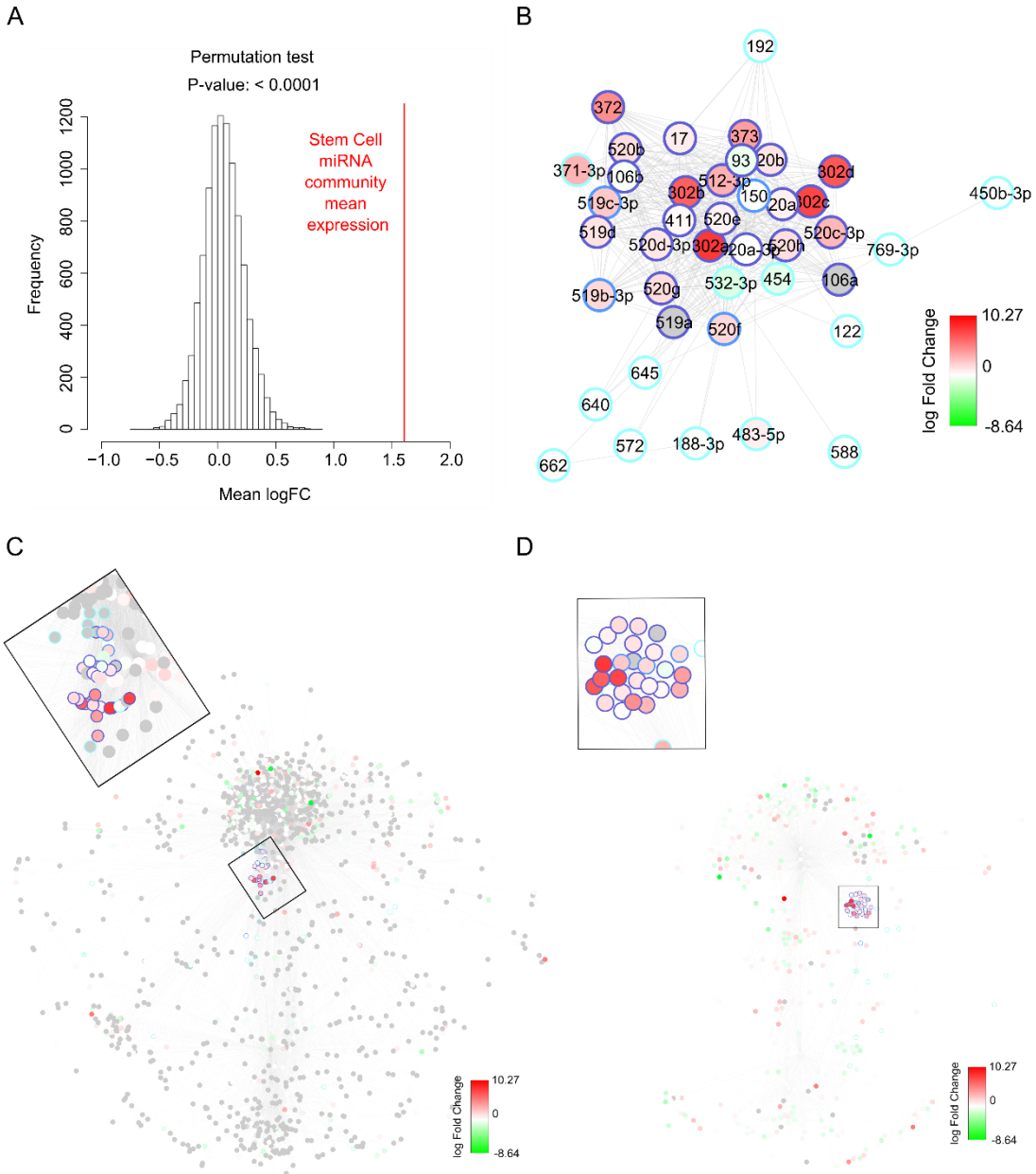


Figure 71. Enrichissement en hits positifs dans la communauté souche pour différentes lignées souches (GSE42446). A | Histogramme des valeurs moyennes en logFC après 10000 permutations. Une observation est la moyenne des logFC d'un groupe de 52 miARN pris au hasard. B **|** LogFC des miARN dans la communauté souche sur le sous-graphe de DIANA-microT. C et D | LogFC des miARN testés sur les réseaux de TargetScan et DIANA-microT respectivement. Les nœuds gris sont les nœuds n'ayant pas été testés dans l'ensemble de données.

communauté est cette fois-ci de 1,6 avec une p-valeur de permutation inférieure à 10^{-04} (Figure 71 A).

Enfin, sur ce même ensemble de données, nous nous sommes également intéressés à l'expression différentielle des miARN entre des cellules souches humaines induites et des cellules somatiques (Figure 72). Les résultats sur les cellules souches embryonnaires et les cellules souches induites sont quasiment similaires (Figure 72 B). En effet, en considérant un

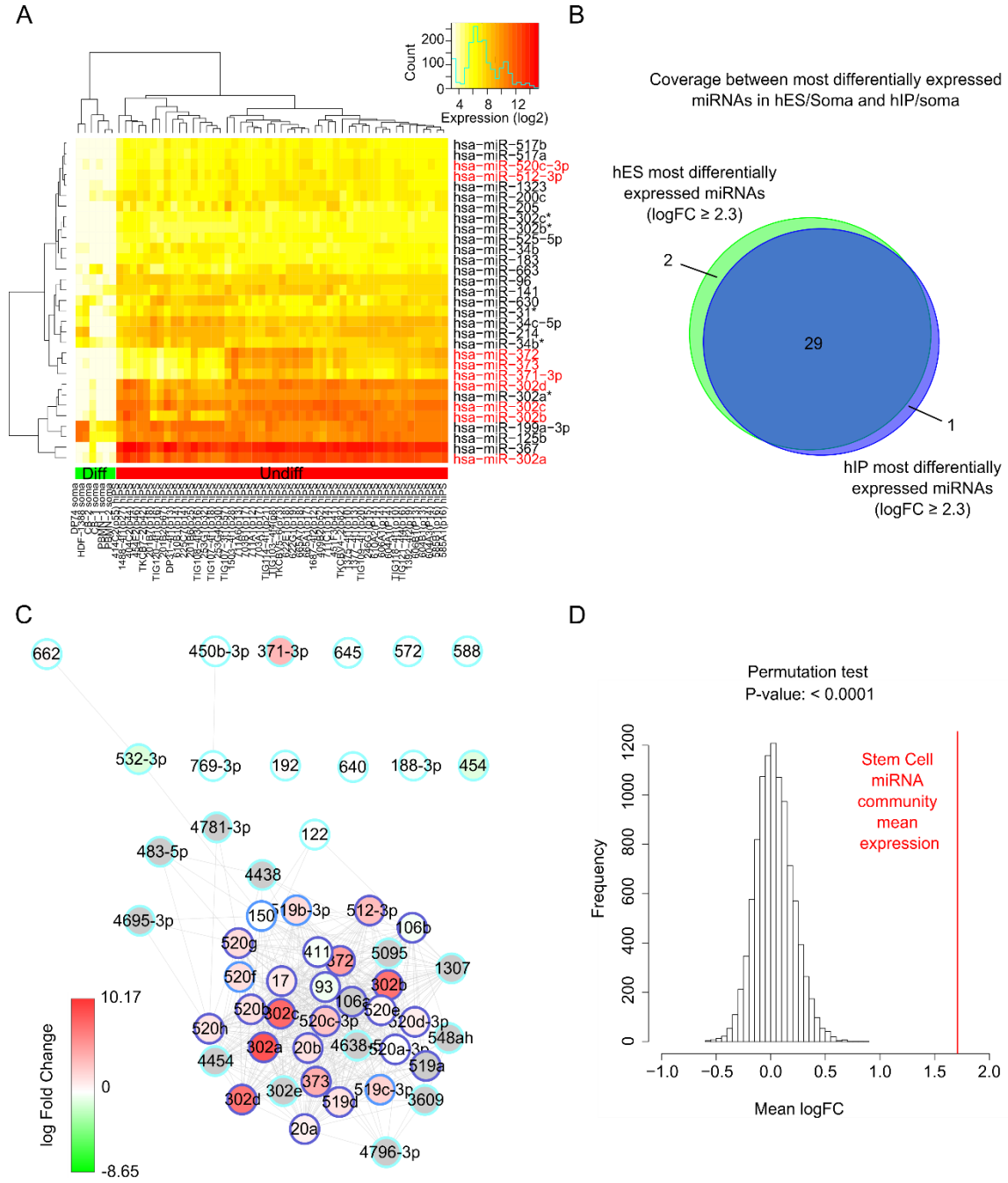


Figure 72. Expression de miARN dans différentes cellules souches induites (GSE42446). A | Expression relative des miARN les plus différentiellement exprimés entre état différencié et non-différencié (cellules souches induites). En rouge sont mis en évidence les miARN de la communauté souche. B | Comparaison entre les miARN les plus différentiellement exprimés (logFC ≥ 2,3) sur les cellules souches embryonnaires et les cellules souches induites. C | LogFC des miARN dans la communauté souche sur le sous-graphe de TargetScan. D | Histogramme des valeurs moyennes en logFC après 10000 permutations. Une observation est la moyenne des logFC d'un groupe de 52 miARN pris au hasard.

logFC supérieur à 2,3 (seuil correspondant à environ 30 « hits »), vingt-neuf des miARN les plus différentiellement exprimés sont retrouvés en commun entre les deux groupes. Deux seulement sont uniques aux cellules souches embryonnaires (miR-630 et -302b*) et un seul est limité aux cellules souches induites (miR-10b). La valeur moyenne des logFC de la

communauté souche est d'environ 1,9 avec une p-valeur de permutation inférieure à 10^{-04} (Figure 72 D).

En conclusion, les miARN de la communauté souche que nous avons identifiée sont bien plus exprimés dans les cellules souches (autant embryonnaires qu'induites) dénotant donc une forte implication de la communauté dans sa quasi-entièreté dans le maintien de la pluripotence des cellules souches. La communauté est par ailleurs significativement enrichie en miARN surexprimés, ce qui démontre bien la valeur de notre inférence basée sur le partage de cibles.

E. Fonctions biologiques et corégulation

Etant donné l'emplacement de la communauté souche dans les deux réseaux (i*.e.* dans la sphère d'influence du club assorti 1), un enrichissement d'ontologie (cf. page 147) envers la régulation transcriptionelle pourrait être attendu à partir des gènes partagés par les membres de la communauté. Par la suite, seule les prédictions de TargetScan seront étudiées. En effet, TargetScan possède l'ensemble des 52 miARN contrairement à DIANA-microT et, comme nous avons pu l'observer, les différences d'enrichissement d'ontologie entre les différentes régions des graphes basés sur les deux algorithmes sont minimes.

Le Tableau 15 confirme bien l'hypothèse formulée ci-dessus : les enrichissements semblent tous associés à la régulation de la transcription avec des p-valeurs corrigées par la procédure de Benjamini et Hochberg (Benjamini and Hochberg, 1995) aux alentours de 10^{-06} pour les processus biologiques et un peu moins faibles pour la localisation cellulaires et les fonctions moléculaires des cibles des miARN. Les miARN de la communauté souche ciblent donc des gènes liés à la régulation de la transcription et pourraient donc potentiellement coréguler cette fonction cellulaire.

Afin d'étudier plus précisément le potentiel de corégulation des miARN de la communauté, nous avons par la suite analysé l'effet du nombre de miARN sur les enrichissements. L'objectif était de suivre l'évolution du nombre de gènes participant à la «

régulation de la transcription » (GO:0006355) en fonction du nombre de miARN ciblant des gènes. Par exemple, à onze miARN, nous avons calculé les enrichissements GO pour l'ensemble des gènes ciblés par au moins onze des miARN de la communauté souche. Ce calcul a été étendu de un à cinquante-deux miARN.

Tableau 15. Enrichissement GO pour les gènes partagés par 50% des 52 miARN de la communauté souche (BP, MF et CC). Soit 2 230 gènes, partagés par au moins 26 miARNs sur les 52 du cluster. Seules les dix premières annotations sont présentées.

BP					
GO.ID	Termes	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0019219	regulation of nucleobase-containing comp.	4 137	628	1.30E-09	1.70E-06
GO:0010556	regulation of macromolecule biosynthetic.	3 658	564	1.30E-09	1.70E-06
GO:2000112	regulation of cellular macromolecule bio.	3 548	549	1.30E-09	1.70E-06
GO:0051252	regulation of RNA metabolic process	3 373	525	1.50E-09	1.70E-06
GO:0009059	macromolecule biosynthetic process	4 543	681	1.60E-09	1.70E-06
GO:0010468	regulation of gene expression	3 930	599	2.00E-09	1.70E-06
GO:0034645	cellular macromolecule biosynthetic proc.	4 407	662	2.20E-09	1.70E-06
GO:0031326	regulation of cellular biosynthetic proc.	3 806	581	3.20E-09	1.93E-06
GO:0009889	regulation of biosynthetic process	3 851	587	3.20E-09	1.93E-06
GO:0051171	regulation of nitrogen compound metaboli.	4 231	637	3.80E-09	2.06E-06

MF					
GO.ID	Termes	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0046872	metal ion binding	3 949	594	2.70E-08	2.67E-05
GO:0043169	cation binding	4 018	601	5.20E-08	2.67E-05
GO:0016740	transferase activity	2 007	322	4.10E-07	1.405E-04
GO:0003677	DNA binding	2 326	363	1.10E-06	2.673E-04
GO:0016773	phosphotransferase activity, alcohol gro.	690	129	1.30E-06	2.673E-04
GO:0016301	kinase activity	752	135	6.90E-06	1.182E-03
GO:0016772	transferase activity, transferring phosp.	893	152	3.50E-05	5.140E-03
GO:0003676	nucleic acid binding	3 764	539	8.30E-05	1.067E-02
GO:0004672	protein kinase activity	585	103	1.80E-04	2.056E-02
GO:0004674	protein serine/threonine kinase activity	437	80	2.70E-04	2.776E-02

CC					
GO.ID	Termes	Annotés	Retrouvés	classicFisher	BH.pVal
GO:0031252	cell leading edge	308	67	1.50E-06	9.57E-04
GO:0098588	bounding membrane of organelle	2 071	315	9.00E-06	2.87E-03
GO:0031090	organelle membrane	2 667	388	4.90E-05	1.04E-02
GO:0000139	Golgi membrane	569	99	1.60E-04	2.02E-02
GO:0030027	lamellipodium	142	33	1.80E-04	2.02E-02
GO:0005901	caveola	65	19	1.90E-04	2.02E-02
GO:0044431	Golgi apparatus part	763	125	3.30E-04	3.01E-02
GO:0005794	Golgi apparatus	1 263	193	4.50E-04	3.59E-02
GO:0044459	plasma membrane part	2 159	310	7.90E-04	5.60E-02
GO:0045121	membrane raft	218	43	9.60E-04	6.12E-02

Ces résultats sont exposés sur la Figure 73 où nous pouvons constater que la p-valeur du terme est minimale entre 4 et 24 miARN, seuil au-delà duquel les p-valeurs ainsi que la différence entre les gènes retrouvés et les gènes attendus augmentent rapidement. La p-valeur minimale est retrouvée pour 21 miARN (2 063 gènes partagés), bien que le maximum

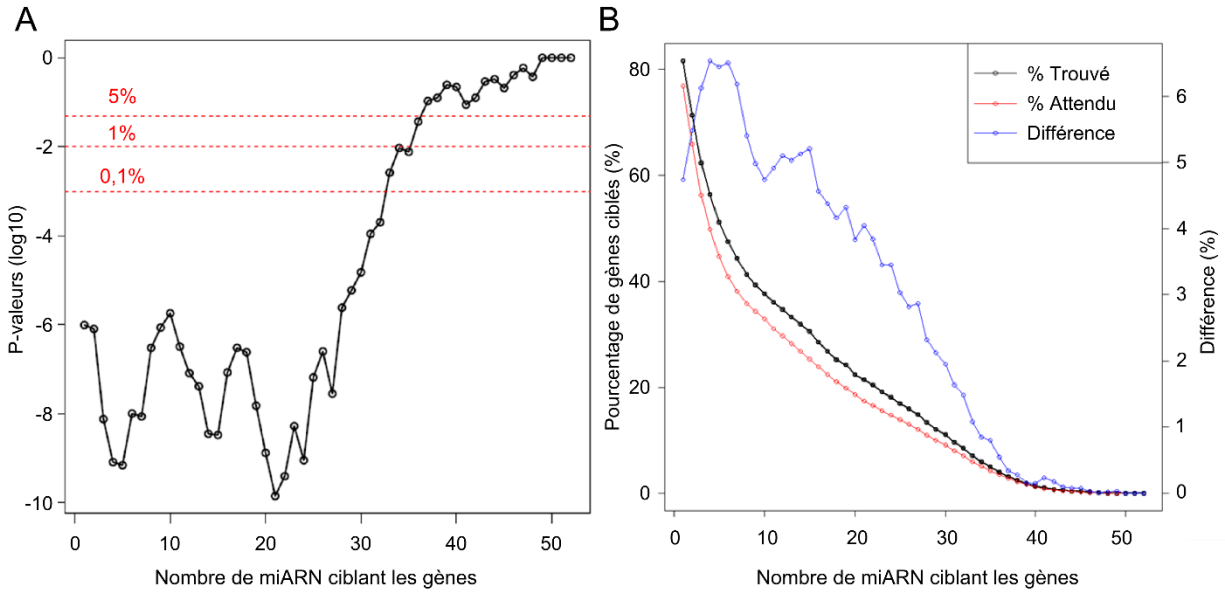


Figure 73. Evolution de l'enrichissement du terme GO « régulation de la transcription » en fonction du nombre de miARN. A | Evolution de la p-valeur du terme « régulation de la transcription » (GO:0006355). B | Evolution de la différence entre pourcentage de gènes retrouvés et pourcentage de gènes attendus.
$\%\ \mathbf{Trouvé} = (\#\ \boldsymbol{gènes\ retrouvés}/\#\ \boldsymbol{gènes\ partagés\ par\ x\ miARN}) \times \mathbf{100}$.
$\%\ \mathbf{Attendu} = (\#\ \boldsymbol{gènes\ attendu\ au\ hasard}/\#\ \boldsymbol{gènes\ partagés\ par\ x\ miARN}) \times \mathbf{100}$.
$\boldsymbol{Différence} = \%\ \boldsymbol{Trouvé} - \%\boldsymbol{Attendu}$.
L'analyse porte sur les gènes co-ciblés par les miARN : p.ex. pour $\boldsymbol{x}$ = 10, ce sont tous les gènes ciblés par au moins 10 miARN qui sont utilisés pour calculer les enrichissements GO.

de différences entre le pourcentage de gènes retrouvés et le pourcentage de gènes attendus soit à son maximum à 4 miARN. D'après cette significativité accrue entre 4 et 24 miARN, nous pouvons supposer que les miARN pourraient agir de concert dans les processus de régulation transcriptionnelle.

Si ces informations restent intéressantes d'un point de vue global, elles ne nous permettent pas de convenablement caractériser la communauté souche de manière plus spécifique puisque ces fonctions restent très génériques. L'objectif suivant a donc été d'étudier les gènes prédits ciblés par les miARN de la communauté mais moins exprimés dans les cellules souches. Autrement dit, nous avons cherché à analyser la portion de gènes la plus probablement régulée par les miARN dans les cellules souches (*i.e.* la partie de gènes déstabilisés). Pour cela, nous avons également utilisé le module « elim » du *package* TopGO afin de ne garder que les termes spécifiques des tableaux d'enrichissement.

En considérant uniquement les gènes les plus sous-exprimés dans les cellules souches par rapport aux cellules somatiques dans l'ensemble de données GSE42446 (logFC

≤ -3, environ 900 gènes), un enrichissement assez clair envers le système immunitaire (réponse immunitaire, réponse inflammatoire, défense face aux virus…) peut être observé (p-valeur corrigées entre 10^{-09} et 10^{-23}) – phénomène déjà évoqué par Lukk et coll. (Lukk et al., 2010). Parmi ces 900 gènes, environ 500 sont prédits pour être ciblés par au moins un miARN de la communauté souche. L'analyse de ce sous-ensemble de gènes montre également un enrichissement pour le système immunitaire, avec néanmoins des p-valeurs moins élevées (10^{-09} à 10^{-18}). Cette baisse de significativité peut toutefois s'expliquer par le nombre total de gènes testés, bien inférieur dans le deuxième cas. Le Tableau 16 récapitule les vingt gènes sous-exprimés les plus ciblés par les miRNA de la communauté souche. Nous pouvons remarquer qu'une bonne partie de ces gènes sont des kinases mais il semble n'y avoir aucun enrichissement particulier dans ce petit ensemble de gènes. En limitant encore plus l'ensemble de gènes sous-exprimés analysés à ceux ciblés par au moins vingt-six miARN différents (50% des membres de la communauté souche), les mêmes effets que précédemment sont observés : un enrichissement pour le système immunitaire accompagné d'une baisse de la significativité des termes (p-valeurs > 10^{-06}).

Tableau 16. Gènes les moins exprimés dans les cellules souches et les plus ciblés par les miARN de la communauté.

Id Entrez	Symbol HGNC	# miARN	Type et fonction
8837	CFLAR	49	Caspase
83544	DNAL1	49	Dynéine
5865	RAB3B	48	Protéine de transport (Superfamille Ras)
2395	FXN	47	Protéine mitochondriale (Phosphorylation oxydative)
7170	TPM3	47	Myosine
7468	WHSC1	47	Histone méthyl-transférase
29767	TMOD2	47	Tropomoduline (impliqué dans la polymérisation d'actine)
55275	VPS53	47	Transport
57404	CYP20A1	47	Oxydoreductase
284716	RIMKLA	47	Glutathione synthase
646851	FAM227A	47	?
659	BMPR2	46	Kinase
3140	MR1	46	Associé aux antigènes
4043	LRPAP1	46	LDL
4915	NTRK2	46	Kinase
9422	ZNF264	46	Facteur de transcription putatif
11122	PTPRT	46	Phosphatase
22848	AAK1	46	Kinase
23112	TNRC6B	46	Impliqué dans l'interférence par ARN
25778	DSTYK	46	Kinase

Par la suite, nous nous sommes intéressés aux liens entre les gènes sous-exprimés dans l'état pluripotent pour différents seuils de logFC (-1,5, -2, -2,5 et -3) et leur corégulation potentielle par les miARN pour différents seuils de nombre de miARN (1 à 52 parmi la communauté souche). Nous avons, pour ce faire, calculé une p-valeur basée sur le test de Fisher afin de déterminer la probabilité d'association entre les gènes sous-exprimés et leur ciblage par les miARN (comme mis en exemple dans le Tableau 17). De façon très simplifiée, l'objectif de cette analyse était de savoir si les gènes sous-exprimés étaient plus ciblés par les miARN de la communauté souche qu'ils le seraient au hasard.

Tableau 17. Exemple d'analyse de probabilité de co-ciblage des gènes sous-exprimés (logFC ≤ -2) dans l'état souche par au moins 10 miARN de la communauté souche.

	Nombre de gènes sous-exprimé dans l'état souche (logFC ≤ -2)	**Nombre de gènes non sous-exprimé dans l'état souche (logFC ≤ -2)**	**Total**	
Nombre de gènes corégulé par au moins 10 miARN	423	4 246	4 669	
Nombre de gènes non corégulé par au moins 10 miARN	1108	12 543	13 651	
Total	**1 531**	**16 789**	**18 320**	**p-valeur = 0,025**

Dans un premier temps, nous constatons que les gènes les plus perturbés (logFC -2.5 et -3) ne sont pas statistiquement plus ciblés par les miARN de la communauté souche que ce qui serait attendu par hasard (Figure 74). En effet, les p-valeurs à ces deux niveaux de logFC sont toujours élevées et ce, quel que soit le nombre de miARN considérés (Figure 74 A). Le nombre de gènes ciblés par les miARN est ainsi proche, voire inférieur à ce qui est attendu (Figure 74 B). Au contraire, pour les deux seuils de logFC -2 et -1,5, une forte baisse de p-valeurs est observable entre 7 et 24 miARN avec un minimum à 14 miARN dans les deux cas. Au-delà d'une vingtaine de miARN en revanche, les p-valeurs ont tendance à augmenter de nouveau.

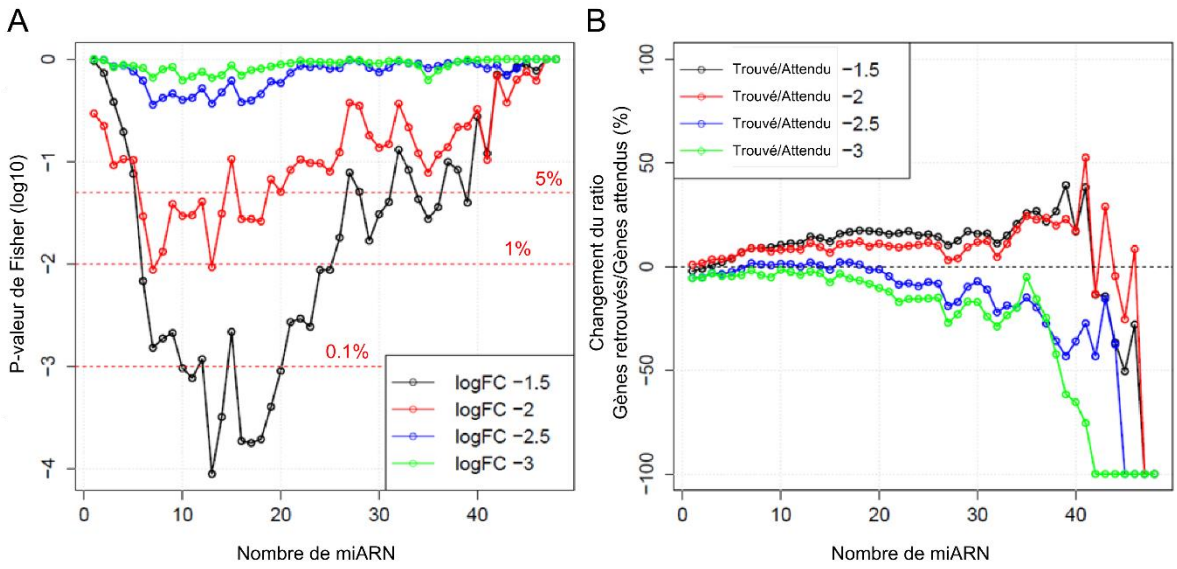


Figure 74. Evolution de la p-valeur de Fisher entre l'association gènes sous-exprimés dans les cellules souches et nombre de miARN prédits comme cibles des gènes. A | Evolution de la p-valeur de Fisher. B | Evolution du ratio gènes trouvés sur gènes attendus normalisé. Par exemple, les proportions sont équivalentes lorsque le ratio est égal à 0 alors qu'à 100, il y a deux fois plus (en pourcentage) de gènes trouvés par rapport aux gènes attendus au hasard. Différents seuils de logFC permettent de définir les gènes sous-exprimés dans les cellules souches.

En conclusion, il semblerait que les gènes les plus négativement perturbés dans les cellules souches (logFC de -2,5 et -3) ne le soient pas (essentiellement) à cause des miARN de la communauté souche. Ces derniers ne semblent en effet pas plus ciblés par les miARN de la communauté souche qu'au hasard. Les gènes moyennement et faiblement perturbés, quant à eux, semblent en revanche bien plus ciblés par les miARN que ce que l'on pourrait attendre au hasard. Par ailleurs, les significativités de ces résultats étant bien supérieures en considérant 4 à 24 miARN, une corégulation potentiellement synergique de ces gènes semblerait bien plus probable. Ces résultats sont en accord avec les propriétés récemment démontrées pour les miARN, selon lesquelles les miARN n'agiraient pas comme des interrupteurs génétiques, mais plutôt en synergie comme des rhéostats pour réguler finement l'expression de centaines de gènes afin d'établir et de confirmer un devenir cellulaire initié par d'autres mécanismes (Ebert 2012, Bartel & Chen 2004, Hornstein & Shomton 2006).

F. Crible d'inhibiteur de miARN

L'objectif d'un crible par inhibiteurs est de bloquer l'effet d'un miARN et d'observer les conséquences phénotypiques de cette inhibition, et ce pour un ensemble de miARN. Pour

cette étude nous avons utilisé des *LNA,* pour *locked nucleic acid*. Le rôle des *LNA* est donc de se lier au miARN, un peu comme les ARN éponges. Dans notre cas, l'expression des miARN n'est donc pas directement modifiée mais les inhibiteurs agissent sur la capacité de ces miARN à agir sur leur(s) gène(s) cible(s).

Environ 750 inhibiteurs ont été utilisés contre les miARN humains. Le modèle cellulaire était des cellules souches H1 modifiées par l'introduction d'une construction *Oct4-GFP* (Chia et al., 2010). Oct4 étant un facteur de transcription surtout exprimé dans l'état pluripotent des cellules souches (Anokye-Danso et al., 2011), ce dernier peut aussi être utilisé comme marqueur de l'état cellulaire (pluripotent ou non). De par la construction introduite dans les cellules, la protéine GFP est traduite en même temps que Oct4 (protéine qui n'est exprimée que dans les cellules pluripotentes), ce qui entraine une augmentation de fluorescence dans les cellules souches. En fonction du niveau d'expression de la GFP, nous pouvons conclure sur le niveau de différenciation des cellules. Ainsi un miARN dont l'inhibition augmenterait fortement l'expression de la GFP, nous indiquerait que ce miARN est plutôt impliqué dans la différenciation.

Les données de la plaque 1 (sur les 9 du crible) sont représentées sur la Figure 75. Nous pouvons premièrement constater que la transfection a bien fonctionnée étant donnée la baisse significative du niveau d'expression de la GFP avec le siARN GFP (en vert sur les panels A et B et montrant 17% de cellules GFP positives en moyenne sur l'ensemble des plaques). Une diminution de cette intensité est également visible pour le siARN dirigé contre Oct4 même si cette dernière est moins remarquable (brun, 41% de cellules GFP positives en moyenne). En revanche, le siARN dirigé contre Nanog (un autre facteur de transcription exprimé dans les cellules souches) montre généralement une différence minime par rapport au siARN AllStars (53% de cellules GFP positives en moyenne contre 58% pour le siARN AllStars). Le panel C de la Figure 75 permet de visualiser l'intensité moyenne des cellules pour les différents puits de la plaque. Nous pouvons noter qu'il n'y a aucun effet spatial particulier sur cette plaque (tout comme sur les autres plaques du crible). L'analyse des données brutes

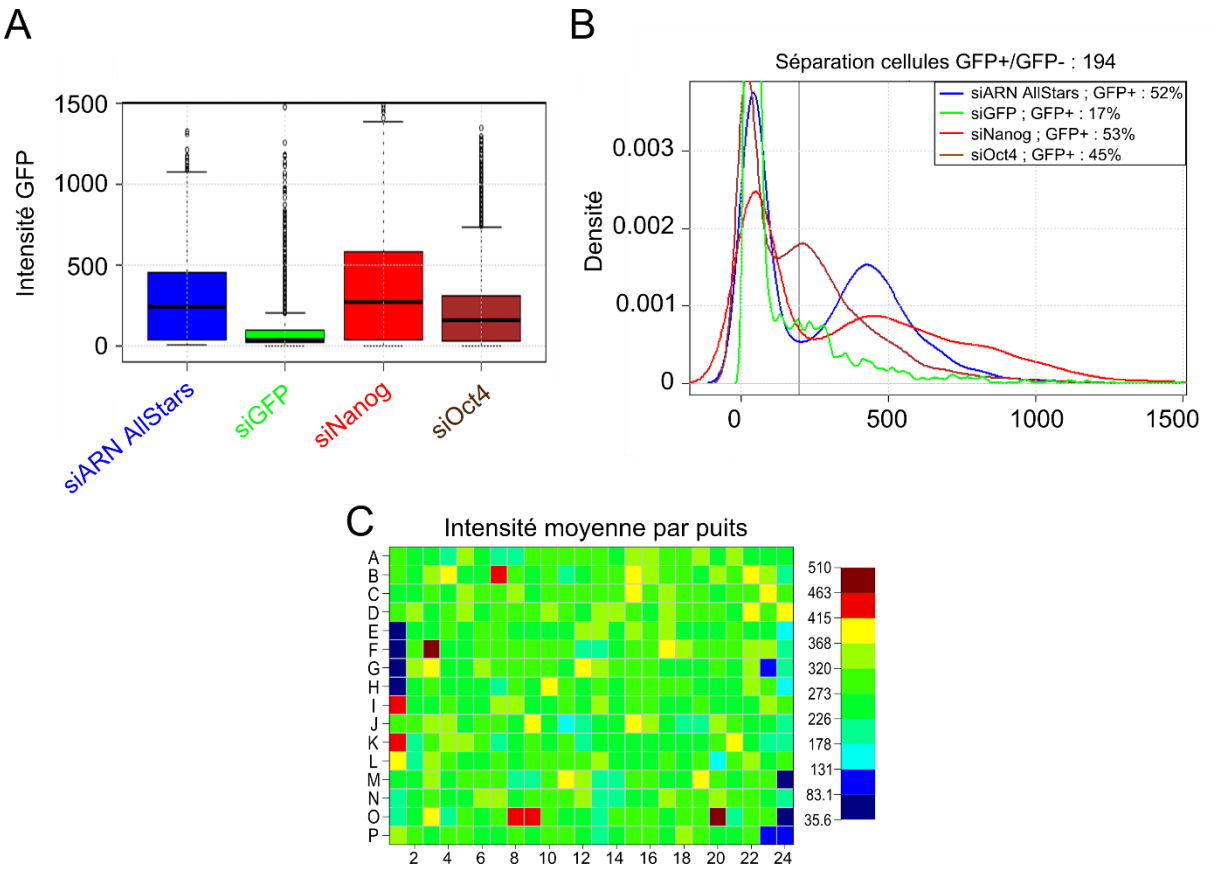


Figure 75. Données brutes des contrôles de la plaque 1 (sur 9) du crible. A | Intensité des siARN contrôles. B | Densité d'intensité des siARN contrôles : en bleu est représenté le siARN AllStars, contrôle négatif. En vert, le siARN contre la GFP ; en rouge, siARN contre Nanog (un autre facteur de transcription des cellules souches) et en brun, le siARN contre Oct4 ; ces trois siARN sont les contrôles positifs. C | Intensité moyenne par puits sur la plaque. Les contrôles sont situés sur les bords.

montrent donc des transfections réussies et l'absence de biais spatiaux (cf. Figure 82 en annexe).

La Figure 76 montre l'ensemble des résultats du crible. Un premier constat est l'effet assez peu marqué des inhibiteurs sur l'expression de la GFP autant positivement que négativement. En effet, les Zscores de l'intensité moyenne de GFP par puits (cf. page 109) s'étendent uniquement entre -1,3 et 2,6, valeurs assez faibles mais pas contradictoires avec les changements d'expression observés dans la plupart des études. Malgré tout, un certain nombre de miARN ressortent comme hits dans le crible. Les miARN membres de la communauté souche, indiqués en rouge sur le panel A, sont répartis sur l'ensemble de la distribution ordonnée des Zscores. Quelques-uns sont ainsi des hits négatifs forts (p.ex. miR-1307) et d'autres des hits positifs forts (p.ex. miR-17) (Tableau 18), d'autres ne montrent aucun phénotype particulier. Un tel constat pourrait être associé à une distribution aléatoire des

scores. Ainsi, afin de savoir si la communauté souche montrait plus de hits forts qu'il serait attendu au hasard, nous avons également étudié l'enrichissement de la communauté souche en hits en caractérisant l'hypothèse nulle H_0 par permutation, et ceci en étudiant la moyenne des scores absolus des miARN de la communauté (sans distinction entre hits positifs et hits négatifs et ceci afin de détecter tous les miARN capables d'affecter la « souchitude »). La p-valeur associée à cette analyse est de 0,015 (Figure 75 D) dénotant donc un enrichissement faible mais significatif en scores élevés dans la communauté : cette dernière regroupe bel et bien plus de miARN avec de forts scores qu'attendu par le simple jeu du hasard.

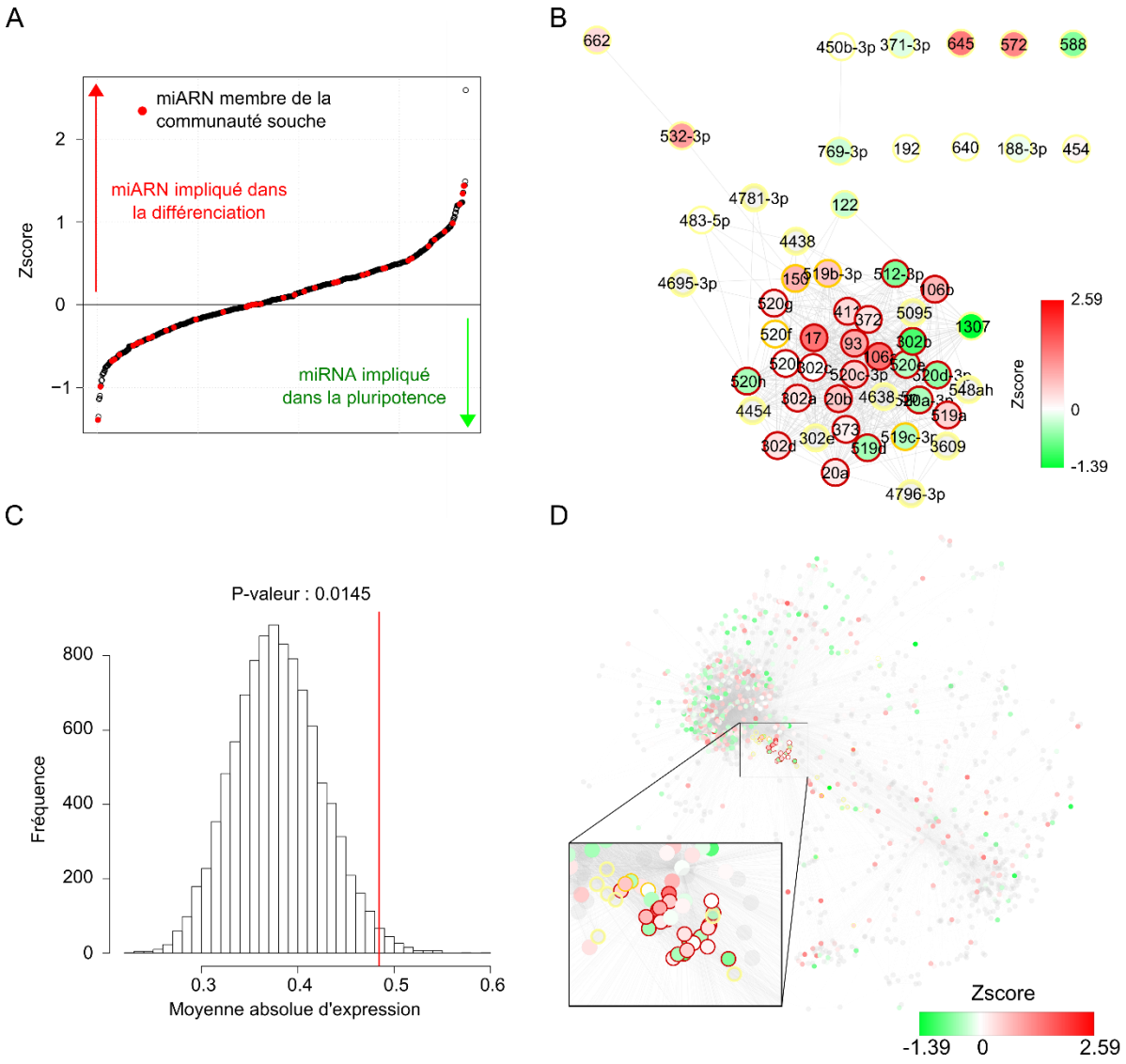


Figure 76. Scores sur le crible par inhibiteurs. A | Zscores ordonnés de l'ensemble des miARN testés. B | Communauté souche et Zscores. C | Histogramme de la moyenne des valeurs absolues des scores de la communauté après 10000 permutations. Une observation est la moyenne des logFC d'un groupe de 52 miARN pris au hasard. D | Zscores sur le réseau entier. Le gris représente des miARN non testés dans le crible.

Il est en outre intéressant de constater que les miARN les mieux caractérisés pour leur association avec la « souchitude », notamment miR-17, miR-106 et la plupart des membres de la famille miR-302, sont retrouvés comme hits positif. En effet, leur inhibition entraîne une augmentation de l'expression de la GFP, et souligne donc leur rôle dans l'initiation de la différenciation (Wu et al., 2012a). Ceci est d'autant plus intéressant qu'une augmentation de GFP est particulièrement difficile à observer car les cellules sont déjà totipotentes et expriment la GFP sous le contrôle d'OCT4.

Tableau 18. Les vingt « meilleurs » hits dans le crible. En rouge sont indiqués les miARN membres de la communauté souche.

Nom de miARN	Z	mimat
hsa-miR-1307	-1.39	MIMAT0005951
hsa-miR-346	-1.35	MIMAT0000773
hsa-miR-361-3p	-1.141	MIMAT0004682
hsa-miR-615-5p	-1.13	MIMAT0004804
hsa-miR-130a*	-1.1	MIMAT0004593
hsa-miR-302b	-0.99	MIMAT0000715
hsa-miR-212	-0.98	MIMAT0000269
hsa-miR-25	-0.91	MIMAT0000081
hsa-miR-512-5p	-0.91	MIMAT0002822
hsa-let-7e	-0.84	MIMAT0000066
	…	
hsa-miR-663	1.23	MIMAT0003326
hsa-miR-498	1.23	MIMAT0002824
hsa-miR-34b*	1.24	MIMAT0000685
hsa-miR-645	1.34	MIMAT0003315
hsa-miR-1296	1.35	MIMAT0005794
hsa-miR-106a	1.44	MIMAT0000103
hsa-miR-17	1.44	MIMAT0000070
hsa-miR-181d	1.44	MIMAT0002821
hsa-miR-148b	1.49	MIMAT0000759
hsa-miR-142-5p	2.59	MIMAT0000433

Nous pouvons noter également que contrairement aux données d'expression, les hits de ce crible sont retrouvés à la fois dans les parties positive et négative des scores, c'est-à-dire que l'inhibition des miARN de la communauté souche entraîne aussi bien une augmentation du potentiel pluripotent qu'une répression (même si ces effets sont très faibles). Cette double potentialité n'est pourtant pas antinomique avec ce que nous avons déjà vu

jusqu'à présent, de par la complexité de régulation des phénotypes cellulaires faisant intervenir gènes codants, non codants et facteurs épigénétiques. Par exemple, miR-302b – un membre de la famille miR-302 – montre un score négatif, c'est-à-dire affectant plutôt la pluripotence des cellules, alors qu'on s'attendrait à ce qu'il agisse plutôt sur la différenciation au vu de nos précédentes conclusions. Il faudrait en revanche des études *in vivo* plus poussées pour mieux comprendre le rôle individuel de chaque miARN. De plus, si nos hypothèses sur la corégulation sont correctes, l'inhibition d'un miARN parmi tout un ensemble pourrait aisément être compensée par d'autres miARN, ce qui expliquerait aussi les faibles scores observés d'une manière générale, et l'action synergique de plusieurs acteurs de cette communauté sur le maintien de la pluripotence ou au contraire la sortie de ce stade et l'initiation de la différenciation. La différenciation terminale étant elle, régulée par d'autres acteurs.

G. Discussion et conclusions

Comme proposé par Chia et collaborateurs, l'identification des molécules nécessaires au maintien du caractère pluripotent des cellules (ou alors celles nécessaires au contraire à la sortie de ce statut et à l'initiation de la différenciation des cellules souches vers certains types cellulaires particuliers), est fondamentale dans la compréhension des mécanismes qui gouvernent le renouvellement des cellules ou leur différenciation (Chia et al., 2010). Cette identification se déroule à différents niveaux : les auteurs de cette étude se sont par exemple focalisés sur les gènes codants et ont identifié PRDM14 comme régulateur critique des cellules souches embryonnaires. Mais nous pouvons également citer d'autres types de régulateurs comme les facteurs épigénétiques (Wang et al., 2015), les longs ARN non codants (Ng et al., 2012) ou encore – et c'est le sujet de cette étude – les miARN.

Au travers de notre approche intégrée, nous avons mis en évidence une communauté de miARN potentiellement impliquée dans la régulation des cellules souches. Composée de cinquante-deux miARN dont vingt-sept miARN « *core* » (beaucoup plus connectés entre eux et retrouvés par différents algorithmes de partitionnement), les membres de cette communauté

montrent premièrement et généralement une expression significativement accrue dans les cellules souches (embryonnaire et induites) comparativement aux cellules somatiques. Si une partie de ces régulateurs (la famille miR-302 et le cluster miR-17 notamment) était déjà connu pour leur implication dans ces processus cellulaires (Suh et al., 2004; Laurent et al., 2008; Gangaraju and Lin, 2009), nous avons inféré de nouveaux miARN comme, par exemple, miR-1307 – miARN qui est également ressorti comme le hit négatif (impliqué dans la pluripotence et l'auto-renouvèlement) le plus efficace du crible par inhibiteurs. C'est d'ailleurs la première fois que ce miARN est prédit impliqué dans la régulation des cellules souches. En ne se focalisant que sur l'expression des miARN et non pas le partage de cibles, ce miARN n'aurait certainement pas pu être identifié. Une caractérisation plus fine de ses propriétés *in vivo* reste toutefois nécessaire pour réellement comprendre son rôle.

Bien entendu, d'autres miARN en dehors de la communauté sont également importants dans le renouvellement ou la différenciation des cellules souches, c'est par exemple le cas de la famille let-7, qui se retrouve isolée de cette communauté. Cette famille a cependant plutôt tendance à être moins exprimée dans les cellules souches et bloque le potentiel de renouvellement des cellules (Ibarra et al., 2007; Melton et al., 2010).

Par des approches statistiques, nous avons également pu mettre en évidence dans ce chapitre des potentialités de corégulation par les miARN. Pour la communauté souche, ce potentiel semble par ailleurs être plus probable premièrement entre 4 et 24 miARN et deuxièmement pour les gènes dont l'expression est subtilement modifiée. En effet, d'après nos résultats, la différence d'expression des gènes les plus perturbés ne peut pas être expliquée uniquement par la (co)régulation par les miARN. Ce constat est aussi en adéquation avec différentes études sur l'effet modeste des miARN sur l'expression de leurs ARNm cibles (Baek et al., 2008; Muniyappa et al., 2009). D'un point de vue global, les miARN de la communauté souche montrent une plus forte tendance à la régulation potentielle des processus liés à la transcription (plutôt que la signalisation). Plus spécifiquement, ce sont des

gènes liés au système immunitaire qui sont réprimés dans les cellules souches par rapport aux cellules somatiques, phénomène très peu reporté dans la littérature

Dans la suite du document, nous nous intéresserons au lien entre la structure du réseau et les processus biologiques que nous venons d'évoquer. Un modèle théorique où le réseau de miARN s'intégrerait avec les autres réseaux biologiques connus sera ainsi proposé comme ouverture aux travaux de cette thèse. L'objectif de ce modèle est de mieux comprendre comment s'intègre la régulation complexe des gènes par les miARN dans ces autres réseaux biologiques.

Chapitre 4 :
Analyse systémique des réseaux de miARN

A. Introduction

Comme nous l'avons vu dans l'introduction générale, les miARN n'agiraient pas vraiment comme des interrupteurs génétiques mais plutôt comme des rhéostats, réajustant de façon synergique et très fine l'expression de gènes codants pour des protéines afin de renforcer le devenir cellulaire déclenché par d'autres mécanismes (Chen et al., 2004; Hornstein and Shomron, 2006; Ebert and Sharp, 2012).

La robustesse est définie comme la capacité d'un système biologique à maintenir une fonction donnée malgré des perturbations internes ou externes (Kitano, 2004). Les propriétés des miARN évoquées ci-dessus suggèrent leur importance pour accroître la robustesse des systèmes biologiques. Comme exposé par Ebert et Sharp (Ebert and Sharp, 2012), ceci est également démontré par les éléments suivants : (1) les gènes dont l'expression est spécifique à certains tissus tendent à avoir une plus grande région 3'UTR avec plus de sites de liaison aux miARN (Stark et al., 2005) ; (2) l'expression des miARN augmente et se diversifie au cours du développement embryonnaire (Thomson et al., 2006) tout comme les 3'UTR qui ont tendance à se rallonger via des choix de sites de polyadénylation alternatifs (Ji et al., 2009) ; et enfin (3) la diversité du répertoire de miARN dans le génome animal augmente avec la complexité de l'organisme (Lee et al., 2007a; Heimberg et al., 2008). Plusieurs propriétés, comme la modularité, l'architecture en nœud papillon et la dégénérescence, ont déjà été décrites dans différents modèles comme apportant de la robustesse à des systèmes biologiques complexes (Tieri et al., 2010; Whitaker et al., 2012).

Afin d'avoir une vision plus globale du rôle régulateur des miARN dans l'expression du génome, nous essaierons de caractériser, dans ce chapitre et d'une manière plus théorique, les propriétés de nos réseaux notamment leur modularité, leur architecture et leur dégénérescence.

B. Les réseaux de miARN montrent-ils une structure en nœud papillon ?

Une version simplifiée de l'organisation sous forme de nœud papillon du métabolisme microbien est représenté sur la (Figure 77) (Csete and Doyle, 2004). Dans cette structure, tout un ensemble de nutriments sont décomposés via la partie catabolique à gauche pour produire quelques précurseurs et briques de construction (p.ex. des acides aminés, des nucléotides, des acides gras, des sucres). Ces briques qui constituent le nœud – cœur de la structure – se déploient ensuite vers la biosynthèse des macromolécules dans la partie droite du nœud papillon.

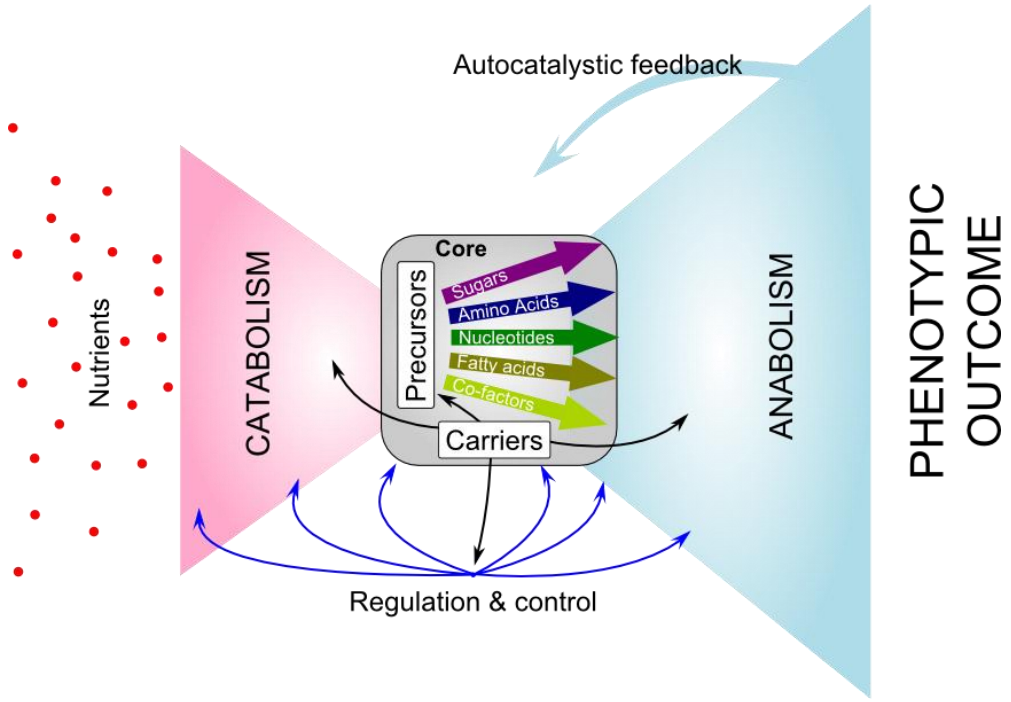


Figure 77. Réseau métabolique en forme de nœud papillon. A gauche, les nutriments sont décomposés en quelques précurseurs qui serviront d'énergie et de briques à la cellule pour l'anabolisme. L'ensemble est finement régulé et contrôlé au travers de différents systèmes. Adapté d'après (Csete and Doyle, 2004).

Nous avions déjà évoqué la présence de trois « modules » ou « sphères d'influence et zone intermédiaire » dans le réseau de miARN (cf. page 124). Les deux sphères – retrouvées aussi bien sur DIANA-microT que sur TargetScan – suggèrent une forme des réseaux en nœud papillon (Figure 78). En menant les mêmes analyses avec les algorithmes miRanda et PITA, la même structure est d'ailleurs approximativement retrouvée (Figure 78). De manière plus ou moins évidente en fonction des algorithmes, nous pouvons noter la conservation

globale des deux zones d'influence. La partie rouge de la Figure 78 est associée, via les gènes cibles partagés, à la régulation de la signalisation cellulaire et plus particulièrement la signalisation par les petites GTPases. Dans le cadre d'une structure en nœud papillon, cette partie pourrait donc potentiellement participer à la régulation de l'information venant de l'environnement cellulaire. L'autre grande partie du réseau (les nœuds en bleu) cible, elle, plutôt des gènes impliqués dans la régulation transcriptionnelle. Cette dernière pourrait participer à l'orientation vers un phénotype cellulaire spécifique en réponse aux signaux entrants (via la partie rouge). La partie centrale du réseau (en violet – zone intermédiaire), en revanche, n'a montré aucun enrichissement particulier et se retrouve connectée aux deux

Figure 78. Réseau de miARN avec différents algorithmes de prédiction et structure en nœud papillon. Les couleurs indiquent les différents groupes retrouvés avec l'organisation sous-jacente retrouvée avec DIANA-microT (deux sphères d'influence et une zone intermédiaire). Le rouge montre la zone d'influence « signalisation par les petites GTPases » (sphère d'influence 2) et la bleu, celle de la « régulation transcriptionelle » (sphère d'influence 1). La partie violette fait référence à la zone intermédiaire.

autres parties du réseau. Elle est également plus petite et concentrée que les deux précédentes.

Selon Csete et Doyle (Csete and Doyle, 2004), le nœud papillon peut être considéré comme une combinaison de deux systèmes dégénérés couplés par une partie centrale (un cœur) constituée de quelques éléments clés. D'après les auteurs, la modularité, ainsi que les règles et les interfaces par lesquelles les modules interagissent, permettent le recyclage des éléments clés du réseau. En comparant l'architecture du réseau DIANA-microT (Figure 78 : DIANA-microT) à ces caractéristiques, il semble illégitime d'évoquer une véritable structure en nœud papillon pour le réseau de miARN. En effet, si les parties bleu et rouge conviennent bien à la définition de Csete et Doyle, la partie violette en revanche n'est pas constituée d'éléments clés agissant comme briques de bases ou monnaie d'échanges entre les deux modules mais sont simplement d'autres miARN. Par conséquent, bien que ce réseau de miARN possède trois modules, nous pensons que ces derniers ne correspondent pas à la définition d'une architecture en nœud papillon. En revanche, nous suggérons que ces réseaux font partie intégrante d'un réseau d'expression de gènes bien plus large.

C. Le réseau de miARN humains fait partie d'un plus large réseau d'expression de gènes et de signalisation possédant une structure en lavallière.

Le réseau de régulation de l'expression des gènes présente aussi une structure en nœud de papillon (Figure 79 : partie droite). Les gènes forment alors la partie entrante du nœud papillon et les protéines forment la partie sortante. Au cœur de cette structure, un nombre réduit d'enzymes très conservées, notamment les polymérases, et les autres composantes des machineries de transcription et de traduction forment le nœud du réseau et permette le recyclage de briques de base que sont les ribonucléotides et les acides aminés.

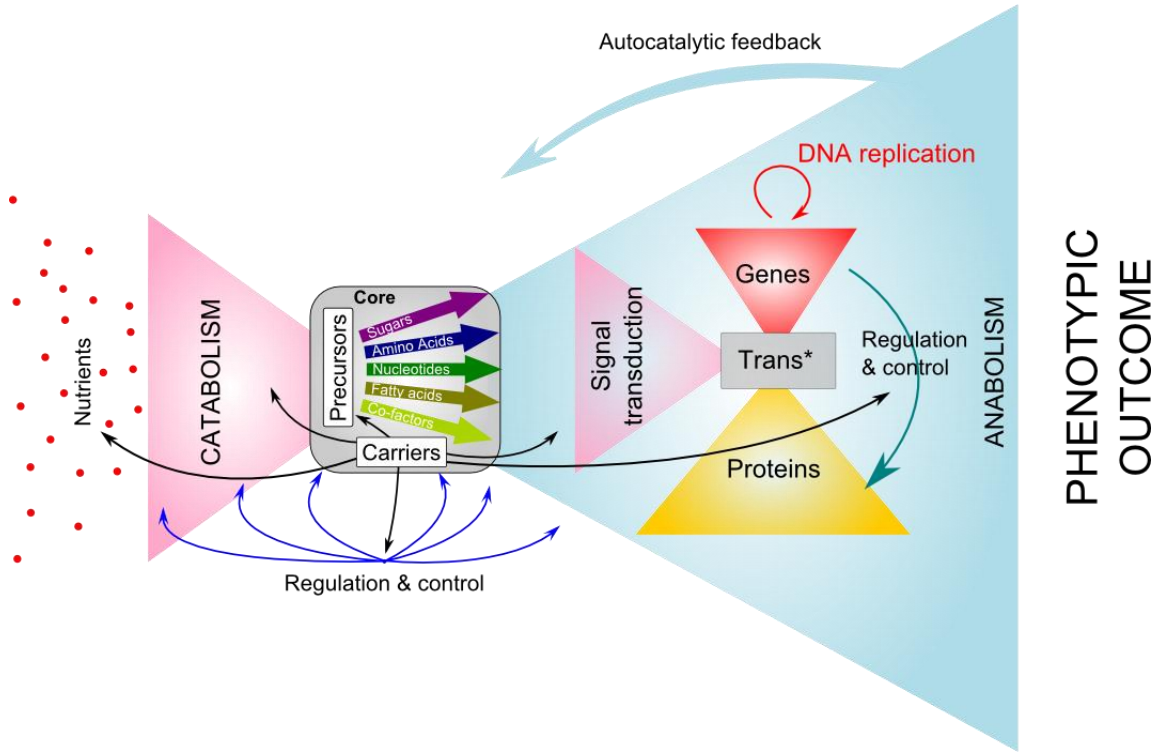


Figure 79. Réseau métabolique et d'expression de gènes en forme de nœud papillon. Le principe est le même que sur la Figure 77 avec une nouvelle dimension pour l'anabolisme au travers du nœud papillon formé par le bloc « *trans** » (transcription/traduction). Cette brique est composée de polymérase et de composantes universelles permettant la transcription des gènes (codants et non codants) et la traduction des gènes (codants). Ce nœud papillon est sous la dépendance de signaux extérieurs ou intérieurs à la cellule. Modifiée d'après (Csete and Doyle, 2004).

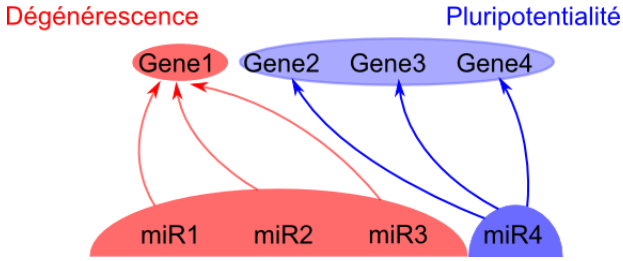
Ce cœur est représenté par le terme générique « *trans** » (pour *transcription* et *translation*) dans la Figure 79).

L'organisation en nœud papillon permet non seulement de gérer le flux de matériaux et d'énergie, mais également de contrôler le flux d'information au sein de la cellule. Les réseaux de signalisation par exemple condensent l'information venant de l'environnement cellulaire au travers de récepteurs, de cascades de signalisation et des messagers secondaires (molécules permettant la transduction d'un signal externe vers l'intérieur de la cellule) afin de modifier l'expression de gènes, au niveau de *Trans**. Cette modification permet en fait à la cellule de s'adapter à ces nouvelles conditions et d'ajuster son phénotypes.

Nous proposons ici une nouvelle structure de réseau, imbriquant le réseau de régulation de l'expression de gènes et le réseau de signalisation pour former un réseau encore plus large, possédant une architecture particulière, en lavallière (figure ci-contre), et clé pour la biosynthèse des macromolécules (Figure 79). Comme pour la structure en nœud papillon, la structure en lavallière possède en son cœur un module constitué d'éléments peu nombreux et très conservés, en l'occurrence le module *trans**. Nous pensons que le réseau de miARN s'intègre également à ce large réseau de régulation génique en agissant à la fois sur la machinerie de transcription et de traduction.

D. Comment le réseau de miARN s'intègre-t-il à ce plus large réseau ?

La dégénérescence fait référence à la capacité d'éléments structurellement différents à effectuer la même fonction dans un système complexe (Edelman and Gally, 2001; Friston and Price, 2003; Whitaker et al., 2012). Souvent confondu avec la redondance, la dégénérescence correspond au concept « plusieurs structures-une fonction » alors que la redondance correspond plutôt au concept « une structure-une fonction » répété. A l'instar du code génétique (Edelman and Gally, 2001), les miARN sont un exemple parfait de dégénérescence puisqu'un même gène peut potentiellement être régulé par des miARN possédant des séquences nucléotidiques différentes (exemple ci-contre).



Dans un système complexe, la redondance de fonction permet d'apporter de la robustesse au système ou, autrement dit, de permettre au système de s'adapter aux changements imprévus avec un impact minimal sur la fonctionnalité du système. La dégénérescence apporte plus de robustesse que la simple redondance : en cas de défaillance d'un élément, cette dernière permet de compenser la perte de l'élément défaillant (p.ex. en cas de défaillance du GPS par défaut d'alimentation, une carte IGN en papier est toujours la

bienvenue). Un autre avantage de la dégénérescence par rapport à la redondance dans le cadre des systèmes biologiques est l'évolvabilité (Edelman and Gally, 2001; Whitacre and Bender, 2010). En effet, la flexibilité des systèmes dégénérés leur permet aussi de développer de nouvelles fonctionnalités avec un avantage sélectif potentiel. A cause de leur caractère dégénéré, les miARN offrent un meilleur potentiel d'adaptation de l'expression génique en réponse directe aux changements de l'environnement.

La pluripotentialité s'oppose, elle, à la dégénérescence : elle correspond au paradigme « une structure-plusieurs fonctions » – tout comme les *moonlighting proteins,* des protéines ayant chacune plusieurs fonctions différentes dans la cellule (Jeffery, 2003) ou encore certaines kinases qui phosphorylent des dizaines de protéines substrats assurant diverses fonctions cellulaires. Sachant qu'un miARN peut reconnaître une centaine de cibles différentes possédant des fonctions différentes, les miARN montrent également des propriétés de pluripotentialité. Par exemple, let-7e-5p semble cibler 2 100 gènes différents validés sur TarBase v7.0.

Nous avons donc analysé les propriétés de dégénérescence et de pluripotentialité de chacune des sphères d'influence du réseau de miARN avec DIANA-microT en analysant certaines de leurs caractéristiques (Table 1).

Table 1. Caractéristiques des sphères d'influences. Etudes des cibles prédites sur DIANA-microT v3. Le nombre de cibles uniques désigne les cibles visées par au moins un miARN de la zone. Les cibles non-redondantes font références à des cibles prédites pour un seul miARN de la zone (c'est-à-dire des cibles uniques à un miARN parmi l'ensemble des membres de la zone). Les cibles génériques sont les cibles visées par l'ensemble des miARN de chaque zone. Le nombre de gènes hubs désigne les cibles génériques étant également des gènes hubs (Shalgi et al., 2007), c'est-à-dire des gènes beaucoup plus ciblés par l'ensemble des miARN du miRnome.

Caractéristiques	Sphère d'influence 1	Sphère d'influence 2
Fonction biologique prédite	Régulation transcriptionnelle	Transduction du signal
Nombre de miARN	323	132
Nombre de liens	1 938	407
Nombre moyen de cibles	3 873	3 508
Nombre de cibles uniques	18 762	17 780
Nombre de cibles non-redondantes	362	946
Nombre de cibles génériques	231	270

Nombre de gènes hubs	142	77
Ratio A (# cibles non-redondantes / # miARN)	1,12	7,17
Ratio B (# cibles unique / # miARN)	58,09	134,7
Ratio C (# cibles génériques / # miARN)	0,72	2,05

Les ratios A, B et C donnent une indication des propriétés de pluripotentialité et de dégénérescence des différentes sphères du réseau. Plus les ratios sont élevés, plus les miARN visent de cibles et ainsi plus leur pluripotentialité est marquée. Au contraire, plus les ratios sont faibles, moins les miARN visent de cibles différentes : ils possèdent donc une dégénérescence plus importante. Nous pouvons constater que la sphère d'influence 2 impliquée dans la transduction du signal montre des ratios plus élevés et donc une certaine pluripotentialité. Ces différentes caractéristiques montrent que les miARN de cette zone ciblent un nombre plus grand de gènes codants. De la même manière, les gènes hubs (Shalgi et al., 2007), sont moins ciblés par les membres de cette zone. Ainsi, cette sphère d'influence agit principalement sur la transduction du signal et permettrait, au travers de la pluripotentialité, une régulation divergente (« *fan out* ») : chacun pouvant réguler une myriade de signaux différents venant de l'environnement extérieur. Au contraire, la sphère d'influence 1, agissant principalement sur la régulation transcriptionnelle, montre plus de dégénérescence avec des ratios plus faibles. Cette partie du réseau régulerait donc de manière convergente (« *fan in* ») les signaux vers des facteurs de transcription clés au cœur du module *trans**.

E. Le réseau de miARN participerait à la canalisation du devenir phénotypique

Il y a de ça 60 ans, Waddington posait le concept de canalisation des programmes développementaux qui – comme l'eau qui s'écoule au milieu d'une vallée, canalisée par les collines environnantes – décrivait comment l'acquisition d'un phénotype au cours du développement est un phénomène stéréotypé, robuste et canalisé (Waddington, 1959). Comme exposé plus tôt, l'étude des miARN a montré que ces derniers avaient une importance fonctionnelle capitale mais tout de même dispensable (Vidigal and Ventura, 2014). Pour

expliquer ce paradoxe, Wu et coll. ont récemment proposé un nouveau rôle non conventionnel – en plus de la régulation génique – pour les miARN : la canalisation phénotypique (Cohen et al., 2006; Choi et al., 2007; Varghese and Cohen, 2007; Martinez et al., 2008; Wu et al., 2009).

Le rôle des miARN est double : ajuster la moyenne des niveaux d'expression des cibles (*tuning*) et réduire la variance des expressions des cibles (*buffering*) (Wu et al., 2009). Bien que ces fonctions soient indépendantes, elles ne sont pas mutuellement exclusives. Dans le réseau de régulation transcriptionnelle, les facteurs de transcription et les miARN pourraient agir de manière complémentaire. En termes d'expression, les facteurs de transcription joueraient alors un rôle dominant dans l'ajustement de la moyenne d'expression des gènes. Le rôle de *buffering* serait ainsi surtout tenu par les miARN afin de pallier tout élément « perturbateur » interne ou externe.

Nous savons aujourd'hui, suite à de nombreuses études, que les miARN et les facteurs de transcription forment effectivement des boucles de rétroaction (Tsang et al., 2007) et agissent de concert. En ce qui concerne le réseau de miARN, nous pensons que la dégénérescence de la sphère d'influence bleue (régulation de la transcription) et la pluripotentialité de la sphère d'influence rouge (régulation de la signalisation) pourrait agir de concert pour maintenir le niveau d'expression d'un gène X mais aussi pour gérer finement la variance de cette expression (Figure 80). En effet, quelle que soit la variation spatio-temporelle de l'expression de ces miARN au cours du développement, en fonction de la localisation spécifique dans un organe ou en réponse aux stress, les miARN pourraient agir comme des amortisseurs pour permettre l'ajustement robuste de l'expression des gènes cibles. Cette hypothèse pourrait être vérifiée en déterminant dans un certain nombre de conditions, si l'expression des miARN est plus variable que celle de leurs gènes cibles. Plus précisément, nous posons l'hypothèse que la dégénérescence participerait principalement au *tuning* de l'expression génique alors que la pluripotentialité participerait plutôt au *buffering* de cette expression. Nous rappelons toutefois que ces fonctions ne sont pas mutuellement exclusives. Ces deux phénomènes, en agissant de concert au sein de structures modulaires en nœud

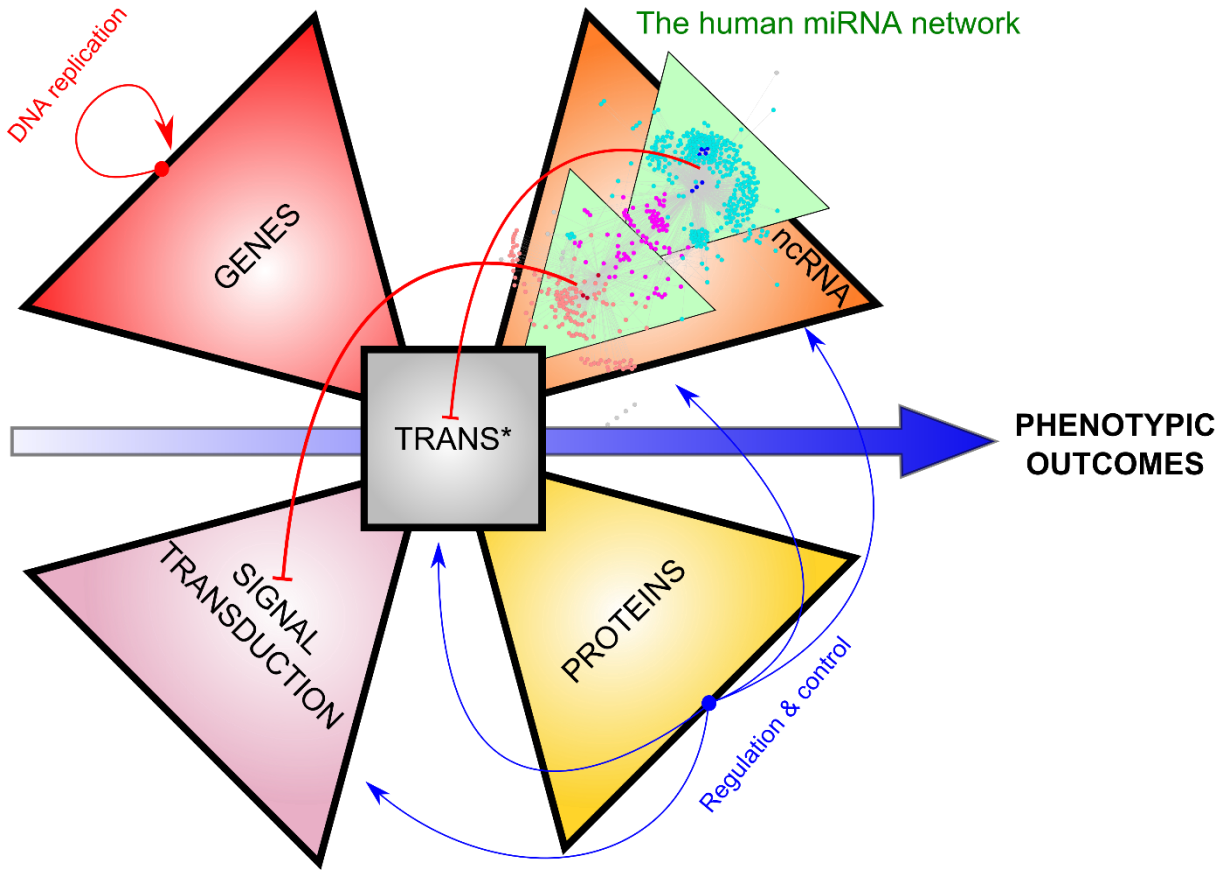


Figure 80. Modèle d'intégration du réseau de miARN dans le réseau plus large de l'expression des gènes. Le réseau de miARN s'intègre à la lavallière de la régulation des gènes en agissant à la fois au niveau de la transduction du signal et au niveau du réseau de régulation des gènes, au sein de boucles de rétroaction (flèches rouges). Trans* (machinerie transcriptionelle et traductionnelle).

papillon ou lavallière apporteraient de la robustesse aux systèmes biologiques complexes. A nouveau, cette hypothèse serait vérifiable dans le futur, en analysant les propriétés de dégénérescence et de pluripotentialité des miARN participant plutôt au processus de développement, par rapport à ceux plutôt impliqués dans la réponse au stress des tissus adultes ou différenciés. A ce titre, il est intéressant de rappeler que les gènes dont l'expression est spécifique à un tissu tendent à présenter de plus grandes régions 3'UTR avec plus de sites de liaison aux miARN (Stark et al., 2005) et que l'expression des miARN augmente et se diversifie au cours du développement embryonnaire (Thomson et al., 2006).

Pour revenir sur le concept de canalisation, certains gènes permettraient en fait d'orienter le phénotype vers la bonne vallée phénotypique alors que d'autres canaliseraient plutôt le phénotype le rendant ainsi moins sensible aux perturbations extérieures (Wu et al., 2009). Les facteurs de transcriptions et les miARN coopéreraient pour réguler les deux

processus mais nous proposons que les facteurs de transcription et la dégénérescence des miARN seraient déterminant dans le choix de la vallée alors que la pluripotentialité des miARN serait plus efficace dans le processus de canalisation afin d'éviter tout débordement dans d'autres vallées : par exemple, les vallées correspondant à un phénotype cancéreux.

Conclusions et perspectives générales

Les miARN sont de fins régulateurs de l'expression génique. Ils permettent un ajustement particulièrement précis de l'expression des gènes en agissant à la fois comme des rhéostats et des tuners (Wu et al., 2009; Vidigal and Ventura, 2014). Ces deux fonctions complémentaires coexistent en partie parce que chaque miARN est susceptible de réguler plusieurs gènes mais aussi parce qu'un même gène peut potentiellement être régulé par différents miARN. L'ensemble des travaux de cette thèse repose sur une hypothèse simple et basée sur ces deux derniers constats : les miARN partageant des cibles communes seraient-ils capables de coréguler les mêmes processus biologiques en agissant de manière conjointe sur les mêmes groupes de gènes ?

Pour répondre à cette question et proposer de nouvelles hypothèses biologiques vis-à-vis de la corégulation des processus biologiques par les miARN, des réseaux de miARN basés sur le partage de cibles prédites ont tout d'abord été construits. Cette construction s'est appuyée sur une métrique – déjà utilisée par d'autres auteurs dans leurs travaux sur les réseaux de miARN (Shalgi et al., 2007; Jiang et al., 2010) – prenant en compte l'intersection des gènes partagés, divisée par le minimum de gènes régulés par l'un ou l'autre des miARN. En imposant alors des seuils sur l'ensemble des paires de miARN, des réseaux binaires de miARN ont pu être créés. Pour des seuils faibles (entre 0 et 40%), les réseaux étaient très denses et difficilement analysables. Au contraire, lorsque ces seuils étaient très élevés (entre 70 et 100%), seules les familles de miARN, dont les séquences sont pratiquement identiques, étaient retrouvées – une information déjà connue. Ce n'est qu'à des seuils intermédiaires (typiquement 50%) que les réseaux avaient un potentiel de découverte intéressant.

Une perspective de cette partie du travail consisterait à ajouter d'autres informations dans le processus d'inférence du réseau. Le premier exemple serait les niveaux d'expression des miARN et de leurs cibles pour différents tissus, comme déjà initiés par différents auteurs (Alshalalfa et al., 2012, 2013; Qabaja et al., 2013). Etant données les fortes variations d'expression entre différents tissus, il conviendrait soit de construire un modèle probabiliste où seules les expressions différentielles très conservées entre miARN et gènes seraient gardées,

soit de construire différents réseaux en fonction des tissus. Les deux types d'informations sont complémentaires : le premier cas permettrait de repérer les miARN agissant de concert dans tout un ensemble de tissus alors que le deuxième permettrait de déterminer les miARN spécifiques à certains types cellulaires uniquement. Les interactions protéine-protéine (Hsu et al., 2008) et la régulation des miARN par les facteurs de transcription (Wang et al., 2010; Ye et al., 2012) seraient également des informations particulièrement pertinentes dans l'inférence de réseaux pour la compréhension de la régulation génique et protéique.

Les réseaux construits suivant notre méthodologie possèdent tous la même particularité : les miARN reliés dans le réseau sont des miARN partageant beaucoup de cibles en commun (au moins 50% selon les seuils définis dans les études présentées dans ce document). Lorsque des groupes de miARN, plus ou moins tous reliés les uns aux autres, sont retrouvés, nous avons affaire à des communautés de miARN. L'identification de ces communautés dans le réseau permet ainsi de trouver des groupes de miARN pouvant potentiellement coréguler des processus biologiques similaires ; ces processus étant identifiés grâce aux cibles partagées par les miARN des communautés. C'est au travers de ce processus d'analyse de graphe que nous avons pu mettre en évidence plusieurs groupes de miARN corégulant des processus biologiques similaires. De la même manière que pour l'inférence des réseaux, l'analyse des processus biologiques pourraient se faire en fonction de données d'expression transcriptomique ou protéique pour restreindre les cibles de façon tissus-spécifiques. Dans notre cas en l'occurrence, nous prédisons effectivement les processus les plus probablement corégulés mais sans donner d'information sur la spécificité tissulaire ou temporelle de cette corégulation ni d'ailleurs si la corégulation a effectivement lieu. L'application de score d'expression aux réseaux comme exposé dans les différents chapitres permet tout de même de partiellement répondre à ce problème. Par cette approche, nous avons pu constater que les miARN des deux clubs assortis n'étaient pas souvent coexprimés dans les différents tissus analysés et nous pouvons supposer que la même conclusion peut être tirée pour les gènes cibles de ces miARN. Pour autant, les validations

biologiques montrent tout de même l'intérêt de notre approche et il reste important de garder à l'esprit l'aspect global de nos analyses : malgré nos prédictions, chaque miARN du réseau pris séparément peut parfaitement agir sur d'autres processus que ceux que nous avons prédits.

Dans le cadre de l'analyse des communautés du réseau, nous avons premièrement mis en évidence deux clubs assortis de miARN dominant fortement le réseau et lui imposant sa structure. La particularité de ces deux groupes était le caractère à la fois *hub* (possédant beaucoup de voisins dans le réseau et, par conséquent, partageant beaucoup de cibles en commun avec d'autres miARN du réseau) et interconnectés. La prédiction des processus biologiques corégulés a montré une implication, pour le plus petit des deux clubs, dans la régulation de la signalisation par les petites GTPase. En validant ces prédictions *in vitro*, nous avons aussi montré le potentiel tumeur-suppresseur de miR-612 et miR-940 ainsi que le potentiel oncogénique de miR-661. En effet, si les trois miARN montraient bien tous une implication dans la voie des petites GTPase, les phénotypes cellulaires observés lors de leur expression ectopique étaient complètement différents. miR-612 et miR-940 entraînaient une forte baisse de la prolifération cellulaire en agissant de manière négative sur la flexibilité du cytosquelette et la prolifération des cellules. Au contraire, miR-661 montrait une tendance à augmenter la flexibilité et la prolifération cellulaire. Par ailleurs, nous avons également mis en évidence pour la première fois une tendance de miR-940 à être réprimé dans les cancers du sein par rapport aux cellules mammaires saines.

Pour la poursuite de ces travaux, il s'agirait d'identifier dans un premier temps et avec plus de précision les cibles impliquées dans les phénotypes observés et, dans un second temps, de confirmer les effets des miARN *in vitro* (sur d'autres types cellulaires par exemple) et *in vivo,* que ce soit dans un cadre thérapeutique ou diagnostique. Une autre question se pose aussi sur les effets conjoints des miARN. En effet et comme déjà évoqué, dans le cas des trois miARN du club assorti 2, les données d'expression n'ont pas montré de coexpression des miARN. Tout au contraire, pour les huit tissus analysés, seul un des trois miARN du club

2 était modérément à fortement exprimé dans un même tissu (variant en fonction des tissus), les autres n'étant quasiment pas exprimés. Il serait ainsi également pertinent de tester l'effet conjoint des miARN *in vitro* ou *in vivo* que ce soit par l'activation (ou l'inhibition) complémentaire de deux miARN – une stratégie récemment explorée dans le traitement de certains cancers (Jayawardena et al., 2012; Kasinski et al., 2015) – ou au contraire en jouant sur des effets opposés (un inhibiteur du miR-661 et un activateur du miR-940 par exemple). La combinatoire reste cependant compliquée à appliquer dans le cadre de validations biologiques fines.

Dans une seconde partie d'analyse des réseaux, nous avons identifié une communauté de miARN impliquée dans la régulation du caractère totipotent des cellules souches. Composée de cinquante-deux miARN, dont vingt-sept miARN très connectés, cette communauté montrait une expression moyenne nettement plus élevée dans l'état souche par rapport à l'état différenciée des cellules. L'approche réseau nous a permis de découvrir de nouveaux miARN – en plus de ceux déjà connu dans la littérature et dont les *seed* sont similaires (Laurent et al., 2008). Par ailleurs, nous avons également montré que les miARN de la communauté étaient susceptibles d'agir de concert sur un ensemble de gènes impliqués dans la régulation transcriptionelle et en particulier sur les gènes faiblement réprimés dans les cellules souches. Enfin, nous avons pu démontrer au travers d'un crible au niveau du génome que les miARN de cette communauté étaient particulièrement impliqués dans les aspects de « souchitude » des cellules mais que leur répression par des LNA entraînait tout autant des effets négatifs que des effets positifs forts sur ces aspects. En effet, la communauté montrait un enrichissement en « hits » aussi bien positifs que négatifs.

Tout comme pour les clubs assortis, il s'agirait dans la suite de valider plus en profondeur *in vitro* et *in vivo* les miARN de cette communauté, autant sur les aspects cibles (communes ou spécifiques) que sur le rôle exact de chacun des miARN dans la « souchitude ». Il s'agirait également de mener une analyse sur la complémentarité des miARN – bien que la complexité de combinatoire soit encore plus élevée dans ce cas – et

aussi de mieux comprendre comment les miARN et les facteurs de transcription du cocktail de dé-différentiation (Oct4, Sox2, Nanog, etc.) agissent ensemble pour donner un certain phénotype cellulaire. Toutes ces analyses s'orienteraient donc vers une meilleure compréhension du devenir des cellules souches. Les réseaux que nous avons caractérisés dans cette thèse gardent tout de même d'autres potentialités de découverte puisque différentes autres communautés peuvent être retrouvées visuellement et/ou en utilisant des algorithmes de détection. Une partie que nous n'avons pas abordée dans ce document mais qui fera l'objet de recherches futures au travers de leur intégration dans un site web consultable publiquement. L'aspect principal de ce site sera l'intégration de réseaux à différents seuils et la possibilité de visualiser des données d'expression (ou tout autre score) sur ces réseaux directement en ligne, afin de mettre en évidence des communautés ou des parties de réseaux enrichis. A plus long terme, l'intégration d'autres algorithmes de prédiction pourraient également être envisagée, tout comme la mise en ligne de réseaux de miARN plus « complexes ».

En conclusions, les travaux de cette thèse auront permis d'apporter de nouvelles connaissances à la (co)régulation des processus biologiques médiés par les miARN mais ont également posé les fondations pour des études systémiques plus poussées.

Bibliographie

Abramoff, M.D., Magalhães, P.J., and Ram, S.J. (2004). Image processing with ImageJ. Biophotonics Int. *11*, 36–42.

Aguda, B.D., Kim, Y., Piper-Hunter, M.G., Friedman, A., and Marsh, C.B. (2008). MicroRNA regulation of a cancer network: consequences of the feedback loops involving miR-17-92, E2F, and Myc. Proc. Natl. Acad. Sci. U. S. A. *105*, 19678–19683.

Albert, R. (2005). Scale-free networks in cell biology. J. Cell Sci. *118*, 4947–4957.

Albert, R., and Barabási, A.-L. (2002). Statistical mechanics of complex networks. Rev. Mod. Phys. *74*, 47–97.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics *22*, 1600–1607.

Alexiou, P., Maragkakis, M., Papadopoulos, G.L., Reczko, M., and Hatzigeorgiou, A.G. (2009). Lost in translation: an assessment and perspective for computational microRNA target identification. Bioinformatics *25*, 3049–3055.

Alshalalfa, M., Bader, G.D., Goldenberg, A., Morris, Q., and Alhajj, R. (2012). Detecting microRNAs of high influence on protein functional interaction networks: a prostate cancer case study. BMC Syst. Biol. *6*, 112.

Alshalalfa, M., D Bader, G., Bismar, T.A., and Alhajj, R. (2013). Coordinate MicroRNA-Mediated Regulation of Protein Complexes in Prostate Cancer. PLoS One *8*, e84261.

Alvarez-Garcia, I., and Miska, E.A. (2005). MicroRNA functions in animal development and human disease. Development *132*, 4653–4662.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., et al. (2003). A uniform system for microRNA annotation. RNA *9*, 277–279.

Anokye-Danso, F., Trivedi, C.M., Juhr, D., Gupta, M., Cui, Z., Tian, Y., Zhang, Y., Yang, W., Gruber, P.J., Epstein, J.A., et al. (2011). Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. Cell Stem Cell *8*, 376–388.

Antolín, S., Calvo, L., Blanco-Calvo, M., Santiago, M.P., Lorenzo-Patiño, M.J., Haz-Conde, M., Santamarina, I., Figueroa, A., Antón-Aparicio, L.M., and Valladares-Ayerbes, M. (2015). Circulating miR-200c and miR-141 and outcomes in patients with breast cancer. BMC Cancer *15*, 297.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Aukerman, M.J., and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. Plant Cell *15*, 2730–2741.

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

Bader, A.G. (2012). miR-34 - a microRNA replacement therapy is headed to the clinic. Front. Genet. *3*, 120.

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

Baigude, H., Ahsanullah, Li, Z., Zhou, Y., and Rana, T.M. (2012). miR-TRAP: a benchtop chemical biology strategy to identify microRNA targets. Angew. Chem. Int. Ed. Engl. *51*, 5880–5883.

Bail, S., Swerdel, M., Liu, H., Jiao, X., Goff, L.A., Hart, R.P., and Kiledjian, M. (2010). Differential regulation of microRNA stability. RNA *16*, 1032–1039.

Balagopal, V., and Parker, R. (2009). Polysomes, P bodies and stress granules: states and fates of eukaryotic mRNAs. Curr. Opin. Cell Biol. *21*, 403–408.

Bandyopadhyay, S., Mitra, R., Maulik, U., and Zhang, M.Q. (2010). Development of the human cancer microRNA network. Silence *1*, 6.

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

Barabási, A.-L., and Albert, R. (1999). Emergence of Scaling in Random Networks. Science (80-.). *286*, 509–512.

Barabási, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. *5*, 101–113.

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988). The new S language. Pacific Grove.

Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., and Izaurralde, E. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes Dev. *20*, 1885–1898.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B *57*, 289–300.

Berezikov, E., Chung, W.-J., Willis, J., Cuppen, E., and Lai, E.C. (2007). Mammalian mirtron genes. Mol. Cell *28*, 328–336.

Van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. J. Vis. Exp.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature *409*, 363–366.

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

Bhajun, R., Guyon, L., Pitaval, A., Sulpice, E., Combe, S., Obeid, P., Haguet, V., Ghorbel, I., Lajaunie, C., and Gidrol, X. (2015). A statistically inferred microRNA network identifies breast cancer target miR-940 as an actin cytoskeleton regulator. Sci. Rep. *5*, 8336.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. *2008*, P10008.

Blüthgen, N., Brand, K., Cajavec, B., Swat, M., Herzel, H., and Beule, D. (2005). Biological profiling of gene groups utilizing Gene Ontology. Genome Inform. *16*, 106–115.

Bohnsack, M.T., Czaplinski, K., and Gorlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA *10*, 185–191.

Boland, A., Tritschler, F., Heimstädt, S., Izaurralde, E., and Weichenrieder, O. (2010). Crystal structure and ligand binding of the MID domain of a eukaryotic Argonaute protein. EMBO Rep. *11*, 522–527.

Bolstad, B.M. (2004). Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization.

Bolstad, B.M. (2013). preprocessCore: A collection of pre-processing functions. R Packag. Version *1*.

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

Bonauer, A., Carmona, G., Iwasaki, M., Mione, M., Koyanagi, M., Fischer, A., Burchfield, J., Fox, H., Doebele, C., Ohtani, K., et al. (2009). MicroRNA-92a controls angiogenesis and functional recovery of ischemic tissues in mice. Science *324*, 1710–1713.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilita.

Bonnet, E., Tatari, M., Joshi, A., Michoel, T., Marchal, K., Berx, G., and Van de Peer, Y. (2010). Module network inference from a cancer gene expression data set identifies microRNA regulated modules. PLoS One *5*, e10162.

Borneman, A.R., Leigh-Bell, J.A., Yu, H., Bertone, P., Gerstein, M., and Snyder, M. (2006). Target hub proteins serve as master regulators of development in yeast. Genes Dev. *20*, 435–448.

Boulet, R., Mazzega, P., and Bourcier, D. (2011). A network approach to the French system of legal codes—part I: analysis of a dense network. Artif. Intell. Law *19*, 333–355.

Boutros, M., and Ahringer, J. (2008). The art and design of genetic screens: RNA interference. Nat. Rev. Genet. *9*, 554–566.

Bowser, A.K., Diamond, A.W., and Addison, J.A. (2013). From puffins to plankton: a DNA-based analysis of a seabird food chain in the northern Gulf of Maine. PLoS One *8*, e83152.

Bracken, C.P., Gregory, P.A., Kolesnikoff, N., Bert, A.G., Wang, J., Shannon, M.F., and Goodall, G.J. (2008). A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition. Cancer Res. *68*, 7846–7854.

Brase, J.C., Johannes, M., Schlomm, T., Fälth, M., Haese, A., Steuber, T., Beissbarth, T., Kuner, R., and Sültmann, H. (2011). Circulating miRNAs are correlated with tumor progression in prostate cancer. Int. J. Cancer *128*, 608–616.

Brengues, M., Teixeira, D., and Parker, R. (2005). Movement of eukaryotic mRNAs between polysomes and cytoplasmic processing bodies. Science *310*, 486–489.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA-target recognition. PLoS Biol. *3*, e85.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. Science *320*, 1185–1190.

Bryant, R.J., Pawlowski, T., Catto, J.W.F., Marsden, G., Vessella, R.L., Rhees, B., Kuslich, C., Visakorpi, T., and Hamdy, F.C. (2012). Changes in circulating microRNA levels associated with prostate cancer. Br. J. Cancer *106*, 768–774.

Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., et al. (2003). Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic Acids Res. *31*, 2443–2450.

Cai, X., Hagedorn, C.H., and Cullen, B.R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA *10*, 1957–1966.

Calin, G.A., Ferracin, M., Cimmino, A., Di Leva, G., Shimizu, M., Wojcik, S.E., Iorio, M. V, Visone, R., Sever, N.I., Fabbri, M., et al. (2005). A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N. Engl. J. Med. *353*, 1793–1801.

Caraus, I., Alsuwailem, A.A., Nadon, R., and Makarenkov, V. (2015). Detecting and overcoming systematic bias in high-throughput screening technologies: a comprehensive review of practical issues and methodological solutions. Brief. Bioinform.

Carmell, M.A., Xuan, Z., Zhang, M.Q., and Hannon, G.J. (2002). The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes Dev. *16*, 2733–2742.

Chalfie, M., Horvitz, H.R., and Sulston, J.E. (1981). Mutations that lead to reiterations in the cell lineages of C. elegans. Cell *24*, 59–69.

Chambers, J.M. (1998). Programming with Data: A Guide to the S Language.

Chang, J., Guo, J.-T., Jiang, D., Guo, H., Taylor, J.M., and Block, T.M. (2008). Liver-specific microRNA miR-122 enhances the replication of hepatitis C virus in nonhepatic cells. J. Virol. *82*, 8215–8223.

Charras, G., and Paluch, E. (2008). Blebs lead the way: how to migrate without lamellipodia. Nat. Rev. Mol. Cell Biol. *9*, 730–736.

Cheloufi, S., Dos Santos, C.O., Chong, M.M.W., and Hannon, G.J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature *465*, 584–589.

Chen, C.-Z., Li, L., Lodish, H.F., and Bartel, D.P. (2004). MicroRNAs modulate hematopoietic lineage differentiation. Science *303*, 83–86.

Chen, G., Ward, B.D., Xie, C., Li, W., Wu, Z., Jones, J.L., Franczak, M., Antuono, P., and Li, S.-J. (2011). Classification of Alzheimer Disease, Mild Cognitive Impairment, and Normal Cognitive Status with Large-Scale Network Analysis Based on Resting-State Functional MR Imaging. Radiology.

Chen, J., He, H., Jiang, F., Militar, J., Ran, P., Qin, G., Cai, C., Chen, X.-B., Zhao, J., Mo, Z., et al. (2012). Analysis of the specific pathways and networks of prostate cancer for gene expression profiles in the Chinese population. Med. Oncol. *29*, 1972–1984.

Chen, S.-M., Chen, H.-C., Chen, S.-J., Huang, C.-Y., Chen, P.-Y., Wu, T.-W.E., Feng, L.-Y., Tsai, H.-C., Lui, T.-N., Hsueh, C., et al. (2013). MicroRNA-495 inhibits proliferation of glioblastoma multiforme cells by downregulating cyclin-dependent kinase 6. World J. Surg. Oncol. *11*, 87.

Chen, X. (2004). A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. Science *303*, 2022–2025.

Chen, Y., Gao, D.-Y., and Huang, L. (2015). In vivo delivery of miRNAs for cancer therapy: Challenges and strategies. Adv. Drug Deliv. Rev. *81C*, 128–141.

Chendrimada, T.P., Gregory, R.I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., and Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature *436*, 740–744.

Chendrimada, T.P., Finn, K.J., Ji, X., Baillat, D., Gregory, R.I., Liebhaber, S.A., Pasquinelli, A.E., and Shiekhattar, R. (2007). MicroRNA silencing through RISC recruitment of eIF6. Nature *447*, 823–828.

Cheng, L., Sharples, R.A., Scicluna, B.J., and Hill, A.F. (2014). Exosomes provide a protective and enriched source of miRNA for biomarker profiling compared to intracellular and cell-free blood. J. Extracell. Vesicles *3*.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479–486.

Chia, N.-Y., Chan, Y.-S., Feng, B., Lu, X., Orlov, Y.L., Moreau, D., Kumar, P., Yang, L., Jiang, J., Lau, M.-S., et al. (2010). A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature *468*, 316–320.

Chiam, K., Wang, T., Watson, D.I., Mayne, G.C., Irvine, T.S., Bright, T., Smith, L., White, I.A., Bowen, J.M., Keefe, D., et al. (2015). Circulating Serum Exosomal miRNAs As Potential Biomarkers for Esophageal Adenocarcinoma. J. Gastrointest. Surg.

Choi, W.-Y., Giraldez, A.J., and Schier, A.F. (2007). Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. Science *318*, 271–274.

Chong, J.J.H., Yang, X., Don, C.W., Minami, E., Liu, Y.-W., Weyers, J.J., Mahoney, W.M., Van Biber, B., Cook, S.M., Palpant, N.J., et al. (2014). Human embryonic-stem-cell-derived cardiomyocytes regenerate non-human primate hearts. Nature *510*, 273–277.

Cifuentes, D., Xue, H., Taylor, D.W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G.J., Lawson, N.D., et al. (2010). A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. Science *328*, 1694–1698.

Ciliberti, S., Martin, O.C., and Wagner, A. (2007). Robustness can evolve gradually in complex regulatory gene networks with varying topology. PLoS Comput. Biol. *3*, e15.

Clamp, M., Cuff, J., Searle, S.M., and Barton, G.J. (2004). The Jalview Java alignment editor. Bioinformatics *20*, 426–427.

Clauset, A., Newman, M., and Moore, C. (2004). Finding community structure in very large networks. Phys. Rev. E *70*, 066111.

Cohen, S.M., Brennecke, J., and Stark, A. (2006). Denoising feedback loops by thresholding--a new role for microRNAs. Genes Dev. *20*, 2769–2772.

Colizza, V., Flammini, A., Serrano, M.A., and Vespignani, A. (2006). Detecting rich-club ordering in complex networks. Nat. Phys. *2*, 110–115.

Coronnello, C., and Benos, P. V (2013). ComiR: Combinatorial microRNA target prediction tool. Nucleic Acids Res. *41*, W159–W164.

Cortez, M.A., Bueso-Ramos, C., Ferdin, J., Lopez-Berestein, G., Sood, A.K., and Calin, G.A. (2011). MicroRNAs in body fluids--the mix of hormones and biomarkers. Nat. Rev. Clin. Oncol. *8*, 467–477.

Coulouarn, C., Factor, V.M., Andersen, J.B., Durkin, M.E., and Thorgeirsson, S.S. (2009). Loss of miR-122 expression in liver cancer correlates with suppression of the hepatic phenotype and gain of metastatic properties. Oncogene *28*, 3526–3536.

Csardi, G., and Nepusz, T. (2006). The igraph Software Package for Complex Network Research. InterJournal Complex Sy *1695*.

Csete, M., and Doyle, J. (2004). Bow ties, metabolism and disease. Trends Biotechnol. *22*, 446–450.

Cui, X., and Churchill, G. (2003). Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. *4*, 210.

Dartnell, L., Simeonidis, E., Hubank, M., Tsoka, S., Bogle, I.D.L., and Papageorgiou, L.G. (2005). Robustness of the p53 network and biological hackers. FEBS Lett. *579*, 3037–3042.

Das, S.K., Sokhi, U.K., Bhutia, S.K., Azab, B., Su, Z.-Z., Sarkar, D., and Fisher, P.B. (2010). Human polynucleotide phosphorylase selectively and preferentially degrades microRNA-221 in human melanoma cells. Proc. Natl. Acad. Sci. U. S. A. *107*, 11948–11953.

Ding, J., Li, X., and Hu, H. (2014). MicroRNA modules prefer to bind weak and unconventional target sites. Bioinformatics *31*, 1366–1374.

Ding, X.C., and Grosshans, H. (2009). Repression of C. elegans microRNA targets at the initiation level of translation requires GW182 proteins. EMBO J. *28*, 213–222.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Doench, J.G., and Sharp, P.A. (2004). Specificity of microRNA target selection in translational repression. Genes Dev. *18*, 504–511.

Donnem, T., Eklo, K., Berg, T., Sorbye, S.W., Lonvik, K., Al-Saad, S., Al-Shibli, K., Andersen, S., Stenvold, H., Bremnes, R.M., et al. (2011). Prognostic impact of MiR-155 in non-small cell lung cancer evaluated by in situ hybridization. J. Transl. Med. *9*, 6.

Doxakis, E. (2013). Principles of miRNA-target regulation in metazoan models. Int. J. Mol. Sci. *14*, 16280–16302.

Drummond, M.J., McCarthy, J.J., Sinha, M., Spratt, H.M., Volpi, E., Esser, K.A., and Rasmussen, B.B. (2011). Aging and microRNA expression in human skeletal muscle: a microarray and bioinformatics analysis. Physiol. Genomics *43*, 595–603.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., Moor, B. De, Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Dweep, H., Sticht, C., Pandey, P., and Gretz, N. (2011). miRWalk--database: prediction of possible miRNA binding sites by “walking” the genes of three genomes. J. Biomed. Inform. *44*, 839–847.

Ebert, M.S., Neilson, J.R., and Sharp, P.A. (2007). MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat. Methods *4*, 721–726.

Ebert, M.S.S., and Sharp, P.A.A. (2012). Roles for microRNAs in conferring robustness to biological processes. Cell *149*, 515–524.

Edelman, G.M., and Gally, J.A. (2001). Degeneracy and complexity in biological systems. Proc. Natl. Acad. Sci. U. S. A. *98*, 13763–13768.

Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

Eichhorn, S.W., Guo, H., McGeary, S.E., Rodriguez-Mias, R.A., Shin, C., Baek, D., Hsu, S.-H., Ghoshal, K., Villén, J., and Bartel, D.P. (2014). mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. Mol. Cell *56*, 104–115.

Eiring, A.M., Harb, J.G., Neviani, P., Garton, C., Oaks, J.J., Spizzo, R., Liu, S., Schwind, S., Santhanam, R., Hickey, C.J., et al. (2010). miR-328 functions as an RNA decoy to modulate hnRNP E2 regulation of mRNA translation in leukemic blasts. Cell *140*, 652–665.

Elliott, M.H., Smith, D.S., Parker, C.E., and Borchers, C. (2009). Current trends in quantitative proteomics. J. Mass Spectrom. *44*, 1637–1660.

Ender, C., Krek, A., Friedländer, M.R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N., and Meister, G. (2008). A human snoRNA with microRNA-like functions. Mol. Cell *32*, 519–528.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003). MicroRNA targets in Drosophila. Genome Biol. *5*, R1.

Erbes, T., Hirschfeld, M., Rücker, G., Jaeger, M., Boas, J., Iborra, S., Mayer, S., Gitsch, G., and Stickeler, E. (2015). Feasibility of urinary microRNA detection in breast cancer patients and its potential as an innovative non-invasive biomarker. BMC Cancer *15*, 193.

Erdős, P., and Rényi, A. (1959). On random graphs. Publ. Math. Debrecen *6*, 290–297.

Eulalio, A., Behm-Ansmant, I., and Izaurralde, E. (2007a). P bodies: at the crossroads of post-transcriptional pathways. Nat. Rev. Mol. Cell Biol. *8*, 9–22.

Eulalio, A., Rehwinkel, J., Stricker, M., Huntzinger, E., Yang, S.-F., Doerks, T., Dorner, S., Bork, P., Boutros, M., and Izaurralde, E. (2007b). Target-specific requirements for enhancers of decapping in miRNA-mediated gene silencing. Genes Dev. *21*, 2558–2570.

Fabian, M.R., Mathonnet, G., Sundermeier, T., Mathys, H., Zipprich, J.T., Svitkin, Y. V, Rivas, F., Jinek, M., Wohlschlegel, J., Doudna, J.A., et al. (2009). Mammalian miRNA RISC recruits CAF1 and PABP to affect PABP-dependent deadenylation. Mol. Cell *35*, 868–880.

Fang, Z., and Rajewsky, N. (2011). The impact of miRNA target sites in coding sequences and in 3'UTRs. PLoS One *6*, e18067.

Feliciano, A., Castellvi, J., Artero-Castro, A., Leal, J.A., Romagosa, C., Hernández-Losa, J., Peg, V., Fabra, A., Vidal, F., Kondoh, H., et al. (2013). miR-125b acts as a tumor suppressor in breast tumorigenesis via its novel direct targets ENPEP, CK2-α, CCNJ, and MEGF9. PLoS One *8*, e76247.

Ferguson, E.L., Sternberg, P.W., and Horvitz, H.R. (1987). A genetic pathway for the specification of the vulval cell lineages of Caenorhabditis elegans. Nature *326*, 259–267.

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

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

Fortin, K., Nicholson, R., and Nicholson, A. (2002). Mouse ribonuclease III. cDNA structure, expression analysis, and chromosomal location. BMC Genomics *3*, 26.

Fortunato, S. (2010). Community detection in graphs. Phys. Rep. *486*, 75–174.

Freeman, L.C. (1978). Centrality in social networks conceptual clarification. Soc. Networks *1*, 215–239.

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

Friston, K.J., and Price, C.J. (2003). Degeneracy and redundancy in cognitive anatomy. Trends Cogn. Sci. *7*, 151–152.

Gaedcke, J., Grade, M., Camps, J., Søkilde, R., Kaczkowski, B., Schetter, A.J., Difilippantonio, M.J., Harris, C.C., Ghadimi, B.M., Møller, S., et al. (2012). The rectal cancer microRNAome--microRNA expression in rectal cancer and matched normal mucosa. Clin. Cancer Res. *18*, 4919–4930.

Gan, H.H., and Gunsalus, K.C. (2013). Tertiary structure-based analysis of microRNA-target interactions. RNA *19*, 539–551.

Gandikota, M., Birkenbihl, R.P., Höhmann, S., Cardon, G.H., Saedler, H., and Huijser, P. (2007). The miRNA156/157 recognition element in the 3' UTR of the Arabidopsis SBP box gene SPL3 prevents early flowering by translational inhibition in seedlings. Plant J. *49*, 683–693.

Gangaraju, V.K., and Lin, H. (2009). MicroRNAs: key regulators of stem cells. Nat. Rev. Mol. Cell Biol. *10*, 116–125.

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

Gáspár, M.E., and Csermely, P. (2012). Rigidity and flexibility of biological networks. Brief. Funct. Genomics *11*, 443–456.

Georgantas, R.W., Hildreth, R., Morisot, S., Alder, J., Liu, C., Heimfeld, S., Calin, G.A., Croce, C.M., and Civin, C.I. (2007). CD34+ hematopoietic stem-progenitor cell microRNA expression and function: a circuit diagram of differentiation control. Proc. Natl. Acad. Sci. U. S. A. *104*, 2750–2755.

George, A.D., and Tenenbaum, S.A. (2006). MicroRNA modulation of RNA-binding protein regulatory elements. RNA Biol. *3*, 57–59.

German, M.A., Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L.A., Nobuta, K., German, R., et al. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat. Biotechnol. *26*, 941–946.

Ghorbel, I., Bertacchi, N., Gidrol, X., and Haguet, V. (2014a). Parallelized contact imaging and automated analysis of cell migration dynamics. Proc. 37th Int. Meet. Ger. Soc. Cell Biol. 71.

Ghorbel, I., Rossant, F., Bloch, I., and Paques, M. (2014b). Modeling a parallelism constraint in active contours. Application to the segmentation of eye vessels and retinal layers. Proc. 18th IEEE Int. Conf. Image Process. 445–448.

Goeman, J.J., and Solari, A. (2011). Multiple Testing for Exploratory Research. Stat. Sci. *26*, 584–597.

Goldberg, D.S., and Roth, F.P. (2003). Assessing experimentally derived interactions in a small world. Proc. Natl. Acad. Sci. *100*, 4372–4376.

Gong, A.-Y., Zhou, R., Hu, G., Li, X., Splinter, P.L., O'Hara, S.P., LaRusso, N.F., Soukup, G.A., Dong, H., and Chen, X.-M. (2009). MicroRNA-513 regulates B7-H1 translation and is involved in IFN-gamma-induced B7-H1 expression in cholangiocytes. J. Immunol. (Baltimore, Md. 1950) *182*, 1325–1333.

Gregory, R.I., Yan, K.-P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. Nature *432*, 235–240.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Res. *32*, D109–D111.

Grimson, A., Farh, K.K.-H., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell *27*, 91–105.

Gruhler, S., and Kratchmarova, I. (2008). Stable isotope labeling by amino acids in cell culture (SILAC). Methods Mol. Biol. *424*, 101–111.

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

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., et al. (2010). PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. J. Vis. Exp.

Haga, C.L., and Phinney, D.G. (2012). MicroRNAs in the Imprinted DLK1-DIO3 Region Repress the Epithelial-to-Mesenchymal Transition by Targeting the TWIST1 Protein Signaling Network. J. Biol. Chem. *287*, 42695–42707.

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature *402*, C47–C52.

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

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

He, H., Zhu, J., Chen, X., Chen, S., Han, Z., Dai, Q., Ling, X., Fu, X., Lin, Z., Deng, Y., et al. (2012). MicroRNA-23b downregulates peroxiredoxin III in human prostate cancer. FEBS Lett. *586*, 2451–2458.

He, H., Han, Z., Dai, Q., Ling, X., Fu, X., Lin, Z., Deng, Y., Qin, G., Cai, C., Chen, J., et al. (2013). Global analysis of the differentially expressed miRNAs of prostate cancer in Chinese patients. BMC Genomics *14*, 757.

Heimberg, A.M., Sempere, L.F., Moy, V.N., Donoghue, P.C.J., and Peterson, K.J. (2008). MicroRNAs and the advent of vertebrate morphological complexity. Proc. Natl. Acad. Sci. U. S. A. *105*, 2946–2950.

Helwak, A., and Tollervey, D. (2014). Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). Nat. Protoc. *9*, 711–728.

Hendrickson, D.G., Hogan, D.J., McCullough, H.L., Myers, J.W., Herschlag, D., Ferrell, J.E., and Brown, P.O. (2009). Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. PLoS Biol. *7*, e1000238.

Hildebrandt, M., and Nellen, W. (1992). Differential antisense transcription from the Dictyostelium EB4 gene locus: implications on antisense-mediated regulation of mRNA stability. Cell *69*, 197–204.

Höck, J., Weinmann, L., Ender, C., Rüdel, S., Kremmer, E., Raabe, M., Urlaub, H., and Meister, G. (2007). Proteomic and functional analysis of Argonaute-containing mRNA-protein complexes in human cells. EMBO Rep. *8*, 1052–1060.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatshefte F&#65533;r Chemie Chem. Mon. *125*, 167–188.

Hornstein, E., and Shomron, N. (2006). Canalization of development by microRNAs.

Hou, B., Ishinaga, H., Midorikawa, K., Shah, S.A., Nakamura, S., Hiraku, Y., Oikawa, S., Murata, M., and Takeuchi, K. (2015). Circulating microRNAs as novel prognosis biomarkers for head and neck squamous cell carcinoma. Cancer Biol. Ther.

Houbaviy, H.B., Murray, M.F., and Sharp, P.A. (2003). Embryonic Stem Cell-Specific MicroRNAs. Dev. Cell *5*, 351–358.

Hsieh, W.-T., Tzeng, K.-R., Ciou, J.-S., Tsai, J.J., Kurubanjerdjit, N., Huang, C.-H., and Ng, K.-L. (2015). Transcription factor and microRNA-regulated network motifs for cancer and signal transduction networks. BMC Syst. Biol. *9 Suppl 1*, S5.

Hsu, C.-W., Juan, H.-F., and Huang, H.-C. (2008). Characterization of microRNA-regulated protein-protein interaction network. Proteomics *8*, 1975–1979.

Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J., Chiu, C.-M., et al. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. *39*, D163–D169.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Humphreys, D.T., Westman, B.J., Martin, D.I.K., and Preiss, T. (2005). MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. Proc. Natl. Acad. Sci. U. S. A. *102*, 16961–16966.

Hunter, P. (2009). Robust yet flexible. In biological systems, resistance to change and innovation in the light of it go hand in hand. EMBO Rep. *10*, 949–952.

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

Hwang, H.-W., Wentzel, E.A., and Mendell, J.T. (2009). Cell-cell contact globally activates microRNA biogenesis. Proc. Natl. Acad. Sci. U. S. A. *106*, 7016–7021.

Ibarra, I., Erlich, Y., Muthuswamy, S.K., Sachidanandam, R., and Hannon, G.J. (2007). A role for microRNAs in maintenance of mouse mammary epithelial progenitor cells. Genes Dev. *21*, 3238–3243.

Ihaka, R., and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. J. Comput. Graph. Stat. *5*, 299–314.

Jakymiw, A., Lian, S., Eystathioy, T., Li, S., Satoh, M., Hamel, J.C., Fritzler, M.J., and Chan, E.K.L. (2005). Disruption of GW bodies impairs mammalian RNA interference. Nat. Cell Biol. *7*, 1267–1274.

Jangra, R.K., Yi, M., and Lemon, S.M. (2010). Regulation of hepatitis C virus translation and infectious virus production by the microRNA miR-122. J. Virol. *84*, 6615–6625.

Jayawardena, T.M., Egemnazarov, B., Finch, E.A., Zhang, L., Payne, J.A., Pandya, K., Zhang, Z., Rosenberg, P., Mirotsou, M., and Dzau, V.J. (2012). MicroRNA-mediated in vitro and in vivo direct reprogramming of cardiac fibroblasts to cardiomyocytes. Circ. Res. *110*, 1465–1473.

Jeffery, C.J. (2003). Moonlighting proteins: old proteins learning new tricks. Trends Genet. *19*, 415–417.

Jeong, H., Mason, S.P., Barabási, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41–42.

Ji, Z., Lee, J.Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc. Natl. Acad. Sci. U. S. A. *106*, 7028–7033.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. *37*, D98–D104.

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., Liu, Y., and Wang, Y. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. BMC Syst. Biol. *4*, S2.

Jiang, X., Huang, H., Li, Z., He, C., Li, Y., Chen, P., Gurbuxani, S., Arnovitz, S., Hong, G.-M., Price, C., et al. (2012). miR-495 is a tumor-suppressor microRNA down-regulated in MLL-rearranged leukemia. Proc. Natl. Acad. Sci. U. S. A. *109*, 19397–19402.

Jinek, M., and Doudna, J.A. (2009). A three-dimensional view of the molecular machinery of RNA interference. Nature *457*, 405–412.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. (2004). Human MicroRNA targets. PLoS Biol. *2*, e363.

De Jong, Y.P., and Jacobson, I.M. (2014). Antisense therapy for hepatitis C virus infection. J. Hepatol. *60*, 227–228.

Jopling, C.L., Yi, M., Lancaster, A.M., Lemon, S.M., and Sarnow, P. (2005). Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. Science *309*, 1577–1581.

Jung, M., Mollenkopf, H.-J., Grimm, C., Wagner, I., Albrecht, M., Waller, T., Pilarsky, C., Johannsen, M., Stephan, C., Lehrach, H., et al. (2009). MicroRNA profiling of clear cell renal cell cancer identifies a robust signature to define renal malignancy. J. Cell. Mol. Med. *13*, 3918–3928.

Kaller, M., Liffers, S.-T., Oeljeklaus, S., Kuhlmann, K., Röh, S., Hoffmann, R., Warscheid, B., and Hermeking, H. (2011). Genome-wide characterization of miR-34a induced changes in protein and mRNA expression by a combined pulsed SILAC and microarray analysis. Mol. Cell. Proteomics *10*, M111.010462.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., et al. (2007). KEGG for linking genomes to life and the environment. Nucleic Acids Res. *36*, D480–D484.

Kasinski, A.L., Kelnar, K., Stahlhut, C., Orellana, E., Zhao, J., Shimer, E., Dysart, S., Chen, X., Bader, A.G., and Slack, F.J. (2015). A combinatorial microRNA therapeutics approach to suppressing non-small cell lung cancer. Oncogene *34*, 3547–3555.

Kedde, M., Strasser, M.J., Boldajipour, B., Oude Vrielink, J.A.F., Slanchev, K., le Sage, C., Nagel, R., Voorhoeve, P.M., van Duijse, J., Ørom, U.A., et al. (2007). RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. Cell *131*, 1273–1286.

Kehat, I., Kenyagin-Karsenti, D., Snir, M., Segev, H., Amit, M., Gepstein, A., Livne, E., Binah, O., Itskovitz-Eldor, J., and Gepstein, L. (2001). Human embryonic stem cells can differentiate into myocytes with structural and functional properties of cardiomyocytes. J. Clin. Invest. *108*, 407–414.

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

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res. *37*, D767–D772.

Ketting, R.F., Fischer, S.E., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. Genes Dev. *15*, 2654–2659.

Kim, S.-K., Nam, J.-W., Rhee, J.-K., Lee, W.-J., and Zhang, B.-T. (2006). miTarget: microRNA target gene prediction using a support vector machine. BMC Bioinformatics *7*, 411.

Kim, S.Y., Lee, Y.-H., and Bae, Y.-S. (2012). MiR-186, miR-216b, miR-337-3p, and miR-760 cooperatively induce cellular senescence by targeting α subunit of protein kinase CKII in human colorectal cancer cells. Biochem. Biophys. Res. Commun. *429*, 173–179.

Kim, V.N., Han, J., and Siomi, M.C. (2009). Biogenesis of small RNAs in animals. Nat. Rev. Mol. Cell Biol. *10*, 126–139.

Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. Genes Dev. *18*, 1165–1178.

Kiriakidou, M., Tan, G.S., Lamprinaki, S., De Planell-Saguer, M., Nelson, P.T., and Mourelatos, Z. (2007). An mRNA m7G Cap Binding-like Motif within Human Ago2 Represses Translation. Cell *129*, 1141–1151.

Kitano, H. (2004). Biological robustness. Nat. Rev. Genet. *5*, 826–837.

Kitano, H. (2007). Towards a theory of biological robustness. Mol. Syst. Biol. *3*, 137.

Kloosterman, W.P., and Plasterk, R.H.A. (2006). The Diverse Functions of MicroRNAs in Animal Development and Disease. Dev. Cell *11*, 441–450.

Kneitz, B., Krebs, M., Kalogirou, C., Schubert, M., Joniau, S., van Poppel, H., Lerut, E., Kneitz, S., Scholz, C.J., Ströbel, P., et al. (2014). Survival in patients with high-risk prostate cancer is predicted by miR-221, which regulates proliferation, apoptosis, and invasion of prostate cancer cells by inhibiting IRF2 and SOCS3. Cancer Res. *74*, 2591–2603.

Korpal, M., Ell, B.J., Buffa, F.M., Ibrahim, T., Blanco, M.A., Celià-Terrassa, T., Mercatali, L., Khan, Z., Goodarzi, H., Hua, Y., et al. (2011). Direct targeting of Sec23a by miR-200s influences cancer cell secretome and promotes metastatic colonization. Nat. Med. *17*, 1101–1108.

Koyanagi-Aoi, M., Ohnuki, M., Takahashi, K., Okita, K., Noma, H., Sawamura, Y., Teramoto, I., Narita, M., Sato, Y., Ichisaka, T., et al. (2013). Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. Proc. Natl. Acad. Sci. U. S. A. *110*, 20569–20574.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. *39*, D152–D157.

Krantz, M., Ahmadpour, D., Ottosson, L.-G., Warringer, J., Waltermann, C., Nordlander, B., Klipp, E., Blomberg, A., Hohmann, S., and Kitano, H. (2009). Robustness and fragility in the yeast high osmolarity glycerol (HOG) signal-transduction pathway. Mol. Syst. Biol. *5*, 281.

Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., et al. (2005). Combinatorial microRNA target predictions. Nat. Genet. *37*, 495–500.

Kudla, G., Granneman, S., Hahn, D., Beggs, J.D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. Proc. Natl. Acad. Sci. U. S. A. *108*, 10010–10015.

Kwon, Y.-K., and Cho, K.-H. (2008). Quantitative analysis of robustness and fragility in biological networks based on feedback dynamics. Bioinformatics *24*, 987–994.

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

Landthaler, M., Gaidatzis, D., Rothballer, A., Chen, P.Y., Soll, S.J., Dinic, L., Ojo, T., Hafner, M., Zavolan, M., and Tuschl, T. (2008). Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. RNA *14*, 2580–2596.

Langer, N., Pedroni, A., and Jäncke, L. (2013). The Problem of Thresholding in Small-World Network Analysis. PLoS One *8*, e53199.

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

Laurent, L.C., Chen, J., Ulitsky, I., Mueller, F.-J., Lu, C., Shamir, R., Fan, J.-B., and Loring, J.F. (2008). Comprehensive MicroRNA Profiling Reveals a Unique Human Embryonic Stem Cell Signature Dominated by a Single Seed Sequence. Stem Cells *26*, 1506–1516.

Lauressergues, D., Couzigou, J.-M., Clemente, H.S., Martinez, Y., Dunand, C., Bécard, G., and Combier, J.-P. (2015). Primary transcripts of microRNAs encode regulatory peptides. Nature *520*, 90–93.

Lee, C.-T., Risom, T., and Strauss, W.M. (2007a). Evolutionary conservation of microRNA regulatory circuits: an examination of microRNA gene complexity and conserved microRNA-target interactions through metazoan phylogeny. DNA Cell Biol. *26*, 209–218.

Lee, E.J., Gusev, Y., Jiang, J., Nuovo, G.J., Lerner, M.R., Frankel, W.L., Morgan, D.L., Postier, R.G., Brackett, D.J., and Schmittgen, T.D. (2007b). Expression profiling identifies microRNA signature in pancreatic cancer. Int. J. Cancer *120*, 1046–1054.

Lee, R., Feinbaum, R., and Ambros, V. (2004a). A short history of a short RNA. Cell *116*, S89–S92.

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

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

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

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., et al. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415–419.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S.H., and Kim, V.N. (2004b). MicroRNA genes are transcribed by RNA polymerase II. EMBO J. *23*, 4051–4060.

Leontis, N.B. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. Nucleic Acids Res. *30*, 3497–3531.

Leung, A.K.L., Calabrese, J.M., and Sharp, P.A. (2006). Quantitative analysis of Argonaute protein reveals microRNA-dependent localization to stress granules. Proc. Natl. Acad. Sci. U. S. A. *103*, 18125–18130.

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

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.

Li, C., Xiong, Q., Zhang, J., Ge, F., and Bi, L.-J. (2012). Quantitative proteomic strategies for the identification of microRNA targets. Expert Rev. Proteomics *9*, 549–559.

Li, L.-C., Okino, S.T., Zhao, H., Pookot, D., Place, R.F., Urakami, S., Enokida, H., and Dahiya, R. (2006). Small dsRNAs induce transcriptional activation in human cells. Proc. Natl. Acad. Sci. *103*, 17337–17342.

Li, M.A., and He, L. (2012). microRNAs as novel regulators of stem cell pluripotency and somatic cell reprogramming. Bioessays *34*, 670–680.

Libri, V., Miesen, P., van Rij, R.P., and Buck, A.H. (2013). Regulation of microRNA biogenesis and turnover by animals and their viruses. Cell. Mol. Life Sci. *70*, 3525–3544.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769–773.

Lin, Y., Sibanda, V.L., Zhang, H.-M., Hu, H., Liu, H., and Guo, A.-Y. (2015). MiRNA and TF co-regulatory network analysis for the pathology and recurrence of myocardial infarction. Sci. Rep. *5*, 9653.

Liu, G., Friggeri, A., Yang, Y., Milosevic, J., Ding, Q., Thannickal, V.J., Kaminski, N., and Abraham, E. (2010). miR-21 mediates fibrogenic activation of pulmonary fibroblasts and lung fibrosis. J. Exp. Med. *207*, 1589–1597.

Liu, J., Carmell, M.A., Rivas, F. V, Marsden, C.G., Thomson, J.M., Song, J.-J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science *305*, 1437–1441.

Liu, J., Valencia-Sanchez, M.A., Hannon, G.J., and Parker, R. (2005a). MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. Nat. Cell Biol. *7*, 719–723.

Liu, J., Rivas, F. V, Wohlschlegel, J., Yates, J.R., Parker, R., and Hannon, G.J. (2005b). A role for the P-body component GW182 in microRNA function. Nat. Cell Biol. *7*, 1261–1266.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science *297*, 2053–2056.

Lodes, M.J., Caraballo, M., Suciu, D., Munro, S., Kumar, A., and Anderson, B. (2009). Detection of cancer with serum miRNAs on an oligonucleotide microarray. PLoS One *4*, e6229.

López-Romero, P., González, M.A., Callejas, S., Dopazo, A., and Irizarry, R.A. (2010). Processing of Agilent microRNA array data. BMC Res. Notes *3*, 18.

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., et al. (2005). MicroRNA expression profiles classify human cancers. Nature *435*, 834–838.

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., and Cui, Q. (2008). An Analysis of Human MicroRNA and Disease Associations. PLoS One *3*, e3420.

Lu, Y., Zhou, Y., Qu, W., Deng, M., and Zhang, C. (2011). A Lasso regression model for the construction of microRNA-target regulatory networks. Bioinformatics *27*, 2406–2413.

Ludwig, T.E., Levenstein, M.E., Jones, J.M., Berggren, W.T., Mitchen, E.R., Frane, J.L., Crandall, L.J., Daigh, C.A., Conard, K.R., Piekarczyk, M.S., et al. (2006). Derivation of human embryonic stem cells in defined conditions. Nat. Biotechnol. *24*, 185–187.

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., Huber, W., Ukkonen, E., and Brazma, A. (2010). A global map of human gene expression. Nat. Biotechnol. *28*, 322–324.

Lytle, J.R., Yario, T.A., and Steitz, J.A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. Proc. Natl. Acad. Sci. U. S. A. *104*, 9667–9672.

Ma, J., Sun, F., Li, C., Zhang, Y., Xiao, W., Li, Z., Pan, Q., Zeng, H., Xiao, G., Yao, K., et al. (2014). Depletion of intermediate filament protein Nestin, a target of microRNA-940, suppresses tumorigenesis by inducing spontaneous DNA damage accumulation in human nasopharyngeal carcinoma. Cell Death Dis. *5*, e1377.

Ma, J.-B., Ye, K., and Patel, D.J. (2004). Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. Nature *429*, 318–322.

Ma, J.-B., Yuan, Y.-R., Meister, G., Pei, Y., Tuschl, T., and Patel, D.J. (2005). Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature *434*, 666–670.

Mahen, E.M., Watson, P.Y., Cottrell, J.W., and Fedor, M.J. (2010). mRNA secondary structures fold sequentially but exchange rapidly in vivo. PLoS Biol. *8*, e1000307.

Mann, M. (2006). Functional and quantitative proteomics using SILAC. Nat. Rev. Mol. Cell Biol. *7*, 952–958.

Maragkakis, M., Alexiou, P., Papadopoulos, G.L., Reczko, M., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Simossis, V.A., et al. (2009a). Accurate microRNA target prediction correlates with protein repression levels. BMC Bioinformatics *10*, 295.

Maragkakis, M., Reczko, M., Simossis, V.A., Alexiou, P., Papadopoulos, G.L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., et al. (2009b). DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Res. *37*, W273–W276.

Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M., Brambrink, T., Johnstone, S., Guenther, M.G., Johnston, W.K., Wernig, M., Newman, J., et al. (2008). Connecting microRNA Genes to the Core Transcriptional Regulatory Circuitry of Embryonic Stem Cells. Cell *134*, 521–533.

Martens-Uzunova, E.S., Jalava, S.E., Dits, N.F., van Leenders, G.J.L.H., Møller, S., Trapman, J., Bangma, C.H., Litman, T., Visakorpi, T., and Jenster, G. (2012). Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. Oncogene *31*, 978–991.

Martinez, N.J., Ow, M.C., Barrasa, M.I., Hammell, M., Sequerra, R., Doucette-Stamm, L., Roth, F.P., Ambros, V.R., and Walhout, A.J.M. (2008). A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity. Genes Dev. *22*, 2535–2549.

Maru, D.M., Singh, R.R., Hannah, C., Albarracin, C.T., Li, Y.X., Abraham, R., Romans, A.M., Yao, H., Luthra, M.G., Anandasabapathy, S., et al. (2009). MicroRNA-196a is a potential marker of progression during Barrett's metaplasia-dysplasia-invasive adenocarcinoma sequence in esophagus. Am. J. Pathol. *174*, 1940–1948.

Mathonnet, G., Fabian, M.R., Svitkin, Y. V, Parsyan, A., Huck, L., Murata, T., Biffo, S., Merrick, W.C., Darzynkiewicz, E., Pillai, R.S., et al. (2007). MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. Science *317*, 1764–1767.

Mattie, M.D., Benz, C.C., Bowers, J., Sensinger, K., Wong, L., Scott, G.K., Fedele, V., Ginzinger, D., Getts, R., and Haqq, C. (2006). Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies. Mol. Cancer *5*, 24.

Mattila, H., Schindler, M., Isotalo, J., Ikonen, T., Vihinen, M., Oja, H., Tammela, T.L.J., Wahlfors, T., and Schleutker, J. (2011). NMD and microRNA expression profiling of the HPCX1 locus reveal MAGEC1 as a candidate prostate cancer predisposition gene. BMC Cancer *11*, 327.

Mazroui, R., Sukarieh, R., Bordeleau, M.-E., Kaufman, R.J., Northcote, P., Tanaka, J., Gallouzi, I., and Pelletier, J. (2006). Inhibition of ribosome recruitment induces stress granule formation independently of eukaryotic initiation factor 2alpha phosphorylation. Mol. Biol. Cell *17*, 4212–4219.

Meir, E., von Dassow, G., Munro, E., and Odell, G.M. (2002). Robustness, Flexibility, and the Role of Lateral Inhibition in the Neurogenic Network. Curr. Biol. *12*, 778–786.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol. Cell *15*, 185–197.

Meister, G., and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. Nature *431*, 343–349.

Melamed, Z., Levy, A., Ashwal-Fluss, R., Lev-Maor, G., Mekahel, K., Atias, N., Gilad, S., Sharan, R., Levy, C., Kadener, S., et al. (2013). Alternative splicing regulates biogenesis of miRNAs located across exon-intron junctions. Mol. Cell *50*, 869–881.

Melton, C., Judson, R.L., and Blelloch, R. (2010). Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. Nature *463*, 621–626.

Mendell, J.T. (2008). miRiad roles for the miR-17-92 cluster in development and disease. Cell *133*, 217–222.

Mestdagh, P., Lefever, S., Pattyn, F., Ridzon, D., Fredlund, E., Fieuw, A., Ongenaert, M., Vermeulen, J., Paepe, A. De, Wong, L., et al. (2011). The microRNA body map: dissecting microRNA function through integrative genomics. Nucleic Acids Res. *39*, e136–e136.

Michael, G. (2011). Single molecule sequencing.

Mitchell, P.S., Parkin, R.K., Kroh, E.M., Fritz, B.R., Wyman, S.K., Pogosova-Agadjanyan, E.L., Peterson, A., Noteboom, J., O'Briant, K.C., Allen, A., et al. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. Proc. Natl. Acad. Sci. U. S. A. *105*, 10513–10518.

Moltzahn, F., Olshen, A.B., Baehner, L., Peek, A., Fong, L., Stöppler, H., Simko, J., Hilton, J.F., Carroll, P., and Blelloch, R. (2011). Microfluidic-based multiplex qRT-PCR identifies diagnostic and prognostic microRNA signatures in the sera of prostate cancer patients. Cancer Res. *71*, 550–560.

Montgomery, R.L., Hullinger, T.G., Semus, H.M., Dickinson, B.A., Seto, A.G., Lynch, J.M., Stack, C., Latimer, P.A., Olson, E.N., and van Rooij, E. (2011). Therapeutic inhibition of miR-208a improves cardiac function and survival during heart failure. Circulation *124*, 1537–1547.

Mortensen, R.D., Serra, M., Steitz, J.A., and Vasudevan, S. (2011). Posttranscriptional activation of gene expression in Xenopus laevis oocytes by microRNA-protein complexes (microRNPs). Proc. Natl. Acad. Sci. U. S. A. *108*, 8281–8286.

Motameny, S., Wolters, S., Nürnberg, P., and Schumacher, B. (2010). Next Generation Sequencing of miRNAs - Strategies, Resources and Methods. Genes (Basel). *1*, 70–84.

Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. Genes Dev. *16*, 720–728.

Muniyappa, M.K., Dowling, P., Henry, M., Meleady, P., Doolan, P., Gammell, P., Clynes, M., and Barron, N. (2009). MiRNA-29a regulates the expression of numerous proteins and reduces the invasiveness and proliferation of human carcinoma cell lines. Eur. J. Cancer *45*, 3104–3118.

Najafi-Shoushtari, S.H., Kristo, F., Li, Y., Shioda, T., Cohen, D.E., Gerszten, R.E., and Näär, A.M. (2010). MicroRNA-33 and the SREBP host genes cooperate to control cholesterol homeostasis. Science *328*, 1566–1569.

Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. Plant Cell *2*, 279–289.

Newman, M., and Girvan, M. (2004). Finding and evaluating community structure in networks. Phys. Rev. E *69*, 026113.

Newman, M.E.J. (2003). The structure and function of complex networks. arXiv:cond-mat/0303516.

Newman, M.E.J. (2006). Modularity and community structure in networks. Proc. Natl. Acad. Sci. U. S. A. *103*, 8577–8582.

Ng, S.-Y., Johnson, R., and Stanton, L.W. (2012). Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. EMBO J. *31*, 522–533.

Nidadavolu, L.S., Niedernhofer, L.J., and Khan, S.A. (2013). Identification of microRNAs dysregulated in cellular senescence driven by endogenous genotoxic stress. Aging (Albany. NY). *5*, 460–473.

Nottrott, S., Simard, M.J., and Richter, J.D. (2006). Human let-7a miRNA blocks protein production on actively translating polyribosomes. Nat. Struct. Mol. Biol. *13*, 1108–1114.

Nymark, P., Guled, M., Borze, I., Faisal, A., Lahti, L., Salmenkivi, K., Kettunen, E., Anttila, S., and Knuutila, S. (2011). Integrative analysis of microRNA, mRNA and aCGH data reveals asbestos- and histology-related changes in lung cancer. Genes. Chromosomes Cancer *50*, 585–597.

Olsen, P.H., and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev. Biol. *216*, 671–680.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Menezes, M.A. de, Kaski, K., Barabási, A.-L., and Kertész, J. (2007). Analysis of a large-scale weighted network of one-to-one human communication. New J. Phys. *9*, 179–179.

Parchem, R.J., Ye, J., Judson, R.L., LaRussa, M.F., Krishnakumar, R., Blelloch, A., Oldham, M.C., and Blelloch, R. (2014). Two miRNA clusters reveal alternative paths in late-stage reprogramming. Cell Stem Cell *14*, 617–631.

Parker, J.S., Roe, S.M., and Barford, D. (2005). Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. Nature *434*, 663–666.

Parker, R., and Sheth, U. (2007). P bodies and the control of mRNA translation and degradation. Mol. Cell *25*, 635–646.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Müller, P., et al. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature *408*, 86–89.

Petersen, C.P., Bordeleau, M.-E., Pelletier, J., and Sharp, P.A. (2006). Short RNAs repress translation after initiation in mammalian cells. Mol. Cell *21*, 533–542.

Peterson, S.M., Thompson, J.A., Ufkin, M.L., Sathyanarayana, P., Liaw, L., and Congdon, C.B. (2014). Common features of microRNA target prediction tools. Front. Genet. *5*, 23.

Piao, X., Zhang, X., Wu, L., and Belasco, J.G. (2010). CCR4-NOT deadenylates mRNA associated with RNA-induced silencing complexes in human cells. Mol. Cell. Biol. *30*, 1486–1494.

Pillai, R.S., Bhattacharyya, S.N., Artus, C.G., Zoller, T., Cougot, N., Basyuk, E., Bertrand, E., and Filipowicz, W. (2005). Inhibition of translational initiation by Let-7 MicroRNA in human cells. Science *309*, 1573–1576.

Place, R.F., Li, L.-C., Pookot, D., Noonan, E.J., and Dahiya, R. (2008). MicroRNA-373 induces expression of genes with complementary promoter sequences. Proc. Natl. Acad. Sci. U. S. A. *105*, 1608–1613.

Plaisier, C.L., Pan, M., and Baliga, N.S. (2012). A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. Genome Res. *22*, 2302–2314.

Pradervand, S., Weber, J., Thomas, J., Bueno, M., Wirapati, P., Lefort, K., Dotto, G.P., and Harshman, K. (2009). Impact of normalization on miRNA microarray expression profiling. RNA *15*, 493–501.

Prévot, P.-P., Augereau, C., Simion, A., Van den Steen, G., Dauguet, N., Lemaigre, F.P., and Jacquemin, P. (2013). Let-7b and miR-495 Stimulate Differentiation and Prevent Metaplasia of Pancreatic Acinar Cells by Repressing HNF6. Gastroenterology *145*, 668–678.e3.

Prill, R.J., Iglesias, P.A., and Levchenko, A. (2005). Dynamic properties of network motifs contribute to biological network organization. PLoS Biol. *3*, e343.

Qabaja, A., Alshalalfa, M., Bismar, T.A., and Alhajj, R. (2013). Protein network-based Lasso regression model for the construction of disease-miRNA functional interactions. EURASIP J. Bioinform. Syst. Biol. *2013*, 3.

R Core Team (2012). R: A Language and Environment for Statistical Computing.

Rajendiran, S., Parwani, A. V, Hare, R.J., Dasgupta, S., Roby, R.K., and Vishwanatha, J.K. (2014). MicroRNA-940 suppresses prostate cancer migration and invasion by regulating MIEN1. Mol. Cancer *13*, 250.

Rajewsky, N., and Socci, N.D. (2004). Computational identification of microRNA targets. Dev. Biol. *267*, 529–535.

Raponi, M., Dossey, L., Jatkoe, T., Wu, X., Chen, G., Fan, H., and Beer, D.G. (2009). MicroRNA classifiers for predicting prognosis of squamous cell lung cancer. Cancer Res. *69*, 5776–5783.

Rayner, K.J., Suárez, Y., Dávalos, A., Parathath, S., Fitzgerald, M.L., Tamehiro, N., Fisher, E.A., Moore, K.J., and Fernández-Hernando, C. (2010). MiR-33 contributes to the regulation of cholesterol homeostasis. Science *328*, 1570–1573.

Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., and Hatzigeorgiou, A.G. (2012). Functional microRNA targets in protein coding sequences. Bioinformatics *28*, 771–776.

Rehwinkel, J., Behm-Ansmant, I., Gatfield, D., and Izaurralde, E. (2005). A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. RNA *11*, 1640–1647.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. Nat. Genet. *38*, 500–501.

Reid, G., Pel, M.E., Kirschner, M.B., Cheng, Y.Y., Mugridge, N., Weiss, J., Williams, M., Wright, C., Edelman, J.J.B., Vallely, M.P., et al. (2013). Restoring expression of miR-16: a novel approach to therapy for malignant pleural mesothelioma. Ann. Oncol. *24*, 3128–3135.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901–906.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. Cell *110*, 513–520.

La Rocca, G., Olejniczak, S.H., González, A.J., Briskin, D., Vidigal, J.A., Spraggon, L., DeMatteo, R.G., Radler, M.R., Lindsten, T., Ventura, A., et al. (2015). In vivo, Argonaute-bound microRNAs exist predominantly in a reservoir of low molecular weight complexes not associated with mRNA. Proc. Natl. Acad. Sci. *112*, 201424217.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83–86.

Rüegger, S., and Großhans, H. (2012). MicroRNA turnover: when, how, and why. Trends Biochem. Sci. *37*, 436–446.

Saetrom, O., Snøve, O., and Saetrom, P. (2005). Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. RNA *11*, 995–1003.

Saetrom, P., Heale, B.S.E., Snøve, O., Aagaard, L., Alluin, J., and Rossi, J.J. (2007). Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res. *35*, 2333–2342.

Salzman, D.W., Shubert-Coleman, J., and Furneaux, H. (2007). P68 RNA helicase unwinds the human let-7 microRNA precursor duplex and is required for let-7-directed silencing of gene expression. J. Biol. Chem. *282*, 32773–32779.

Schaefer, A., Jung, M., Mollenkopf, H.-J., Wagner, I., Stephan, C., Jentzmik, F., Miller, K., Lein, M., Kristiansen, G., and Jung, K. (2010). Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma. Int. J. Cancer *126*, 1166–1176.

Schepeler, T., Reinert, J.T., Ostenfeld, M.S., Christensen, L.L., Silahtaroglu, A.N., Dyrskjøt, L., Wiuf, C., Sørensen, F.J., Kruhøffer, M., Laurberg, S., et al. (2008). Diagnostic and prognostic microRNAs in stage II colon cancer. Cancer Res. *68*, 6416–6424.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675.

Schrauder, M.G., Strick, R., Schulz-Wendtland, R., Strissel, P.L., Kahmann, L., Loehberg, C.R., Lux, M.P., Jud, S.M., Hartmann, A., Hein, A., et al. (2012). Circulating micro-RNAs as potential blood-based markers for early stage breast cancer detection. PLoS One *7*, e29770.

Schwarz, D.S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the Assembly of the RNAi Enzyme Complex. Cell *115*, 199–208.

Seggerson, K., Tang, L., and Moss, E.G. (2002). Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation. Dev. Biol. *243*, 215–225.

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58–63.

Sethupathy, P., Corda, B., and Hatzigeorgiou, A.G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA *12*, 192–197.

Shalgi, R., Lieber, D., Oren, M., and Pilpel, Y. (2007). Global and Local Architecture of the Mammalian microRNA–Transcription Factor Regulatory Network. PLoS Comput Biol *3*, e131.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. *13*, 2498–2504.

Shin, V.Y., Siu, J.M., Cheuk, I., Ng, E.K.O., and Kwong, A. (2015). Circulating cell-free miRNAs as biomarker for triple-negative breast cancer. Br. J. Cancer.

Shirdel, E.A., Xie, W., Mak, T.W., and Jurisica, I. (2011). NAViGaTing the Micronome – Using Multiple MicroRNA Prediction Databases to Identify Signalling Pathway-Associated MicroRNAs. PLoS One *6*, e17429.

Simion, A., Laudadio, I., Prévot, P.-P., Raynaud, P., Lemaigre, F.P., and Jacquemin, P. (2010). MiR-495 and miR-218 regulate the expression of the Onecut transcription factors HNF-6 and OC-2. Biochem. Biophys. Res. Commun. *391*, 293–298.

Simons, R.W. (1988). Naturally occurring antisense RNA control--a brief review. Gene *72*, 35–44.

Smyth, G.K. (2005). Limma: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using {R} and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, eds. (New York: Springer), pp. 397–420.

Somlyo, A.P., and Somlyo, A. V. (2000). Signal transduction by G-proteins, Rho-kinase and protein phosphatase to smooth muscle and non-muscle myosin II. J. Physiol. *522*, 177–185.

Song, B., Zhang, C., Li, G., Jin, G., and Liu, C. (2015). MiR-940 inhibited pancreatic ductal adenocarcinoma growth by targeting MyD88. Cell. Physiol. Biochem. *35*, 1167–1177.

Souret, F.F., Kastenmayer, J.P., and Green, P.J. (2004). AtXRN4 degrades mRNA in Arabidopsis and its substrates include selected miRNA targets. Mol. Cell *15*, 173–183.

Spahn, M., Kneitz, S., Scholz, C.-J., Stenger, N., Rüdiger, T., Ströbel, P., Riedmiller, H., and Kneitz, B. (2010). Expression of microRNA-221 is progressively reduced in aggressive prostate cancer and metastasis and predicts clinical recurrence. Int. J. Cancer *127*, 394–403.

Srikantan, S., Abdelmohsen, K., Lee, E.K., Tominaga, K., Subaran, S.S., Kuwano, Y., Kulshrestha, R., Panchakshari, R., Kim, H.H., Yang, X., et al. (2011). Translational Control of TOP2A Influences Doxorubicin Efficacy▿. Mol. Cell. Biol. *31*, 3790–3801.

Stadler, B., Ivanovska, I., Mehta, K., Song, S., Nelson, A., Tan, Y., Mathieu, J., Darby, C., Blau, C.A., Ware, C., et al. (2010). Characterization of microRNAs involved in embryonic stem cell states. Stem Cells Dev. *19*, 935–950.

Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. (2003). Identification of Drosophila MicroRNA targets. PLoS Biol. *1*, E60.

Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell *123*, 1133–1146.

Sturm, M., Hackenberg, M., Langenberger, D., and Frishman, D. (2010). TargetSpy: a supervised machine learning approach for microRNA target prediction. BMC Bioinformatics *11*, 292.

Subramanyam, D., Lamouille, S., Judson, R.L., Liu, J.Y., Bucay, N., Derynck, R., and Blelloch, R. (2011). Multiple targets of miR-302 and miR-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells. Nat. Biotechnol. *29*, 443–448.

Suh, M.-R., Lee, Y., Kim, J.Y., Kim, S.-K., Moon, S.-H.S.Y., Lee, J.Y., Cha, K.-Y., Chung, H.M., Yoon, H.S., Moon, S.-H.S.Y., et al. (2004). Human embryonic stem cells express a unique set of microRNAs. Dev. Biol. *270*, 488–498.

Sun, J., Gong, X., Purow, B., and Zhao, Z. (2012). Uncovering MicroRNA and Transcription Factor Mediated Regulatory Networks in Glioblastoma. PLoS Comput. Biol. *8*, e1002488.

Sun, Z., Evans, J., Bhagwate, A., Middha, S., Bockol, M., Yan, H., and Kocher, J.-P. (2014). CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. BMC Genomics *15*, 423.

Szafranska-Schwarzbach, A.E., Adai, A.T., Lee, L.S., Conwell, D.L., and Andruss, B.F. (2011). Development of a miRNA-based diagnostic assay for pancreatic ductal adenocarcinoma. Expert Rev. Mol. Diagn. *11*, 249–257.

Tabar, V., and Studer, L. (2014). Pluripotent stem cells in regenerative medicine: challenges and recent progress. Nat. Rev. Genet. *15*, 82–92.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Tang, R., Li, L., Zhu, D., Hou, D., Cao, T., Gu, H., Zhang, J., Chen, J., Zhang, C.-Y., and Zen, K. (2012). Mouse miRNA-709 directly regulates miRNA-15a/16-1 biogenesis at the posttranscriptional level in the nucleus: evidence for a microRNA hierarchy system. Cell Res. *22*, 504–515.

Tao, Z.-H., Wan, J.-L., Zeng, L.-Y., Xie, L., Sun, H.-C., Qin, L.-X., Wang, L., Zhou, J., Ren, Z.-G., Li, Y.-X., et al. (2013). miR-612 suppresses the invasive-metastatic cascade in hepatocellular carcinoma. J. Exp. Med. *210*, 789–803.

Thermann, R., and Hentze, M.W. (2007). Drosophila miR2 induces pseudo-polysomes and inhibits translation initiation. Nature *447*, 875–878.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. *22*, 4673–4680.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. Science *282*, 1145–1147.

Thomson, J.M., Newman, M., Parker, J.S., Morin-Kensicki, E.M., Wright, T., and Hammond, S.M. (2006). Extensive post-transcriptional regulation of microRNAs and its implications for cancer. Genes Dev. *20*, 2202–2207.

Tieri, P., Grignolio, A., Zaikin, A., Mishto, M., Remondini, D., Castellani, G.C., and Franceschi, C. (2010). Network, degeneracy and bow tie. Integrating paradigms and architectures to grasp the complexity of the immune system. Theor. Biol. Med. Model. *7*, 32.

Till, S., Lejeune, E., Thermann, R., Bortfeld, M., Hothorn, M., Enderle, D., Heinrich, C., Hentze, M.W., and Ladurner, A.G. (2007). A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. Nat. Struct. Mol. Biol. *14*, 897–903.

Todesco, M., Rubio-Somoza, I., Paz-Ares, J., and Weigel, D. (2010). A collection of target mimics for comprehensive analysis of microRNA function in Arabidopsis thaliana. PLoS Genet. *6*, e1001031.

Trang, P., Wiggins, J.F., Daige, C.L., Cho, C., Omotola, M., Brown, D., Weidhaas, J.B., Bader, A.G., and Slack, F.J. (2011). Systemic delivery of tumor suppressor microRNA mimics using a neutral lipid emulsion inhibits lung tumors in mice. Mol. Ther. *19*, 1116–1122.

Tsang, J., Zhu, J., and van Oudenaarden, A. (2007). MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Mol. Cell *26*, 753–767.

Tsang, J.S., Ebert, M.S., and van Oudenaarden, A. (2010). Genome-wide Dissection of MicroRNA Functions and Cotargeting Networks Using Gene Set Signatures. Mol. Cell *38*, 140–153.

Ulrich Elsner, A.S. (1997). Graph Partitioning - A Survey. Chemnitz, Ger. Tech. Univ. Chemnitz 97–27.

Unlü, M., Morgan, M.E., and Minden, J.S. (1997). Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. Electrophoresis *18*, 2071–2077.

Varghese, J., and Cohen, S.M. (2007). microRNA miR-14 acts to modulate a positive autoregulatory loop controlling steroid hormone signaling in Drosophila. Genes Dev. *21*, 2277–2282.

Vasudevan, S., Tong, Y., and Steitz, J.A. (2007). r. Science *318*, 1931–1934.

Vencken, S., Hassan, T., McElvaney, N.G., Smith, S.G.J., and Greene, C.M. (2015). miR-CATCH: microRNA capture affinity technology. Methods Mol. Biol. *1218*, 365–373.

Vetter, G., Saumet, A., Michele, M., Vallar, L., Le Bechec, A., Laurini, C., Sabbah, M., Arar, K., Theillet, C., Lecellier, C.-H., et al. (2010). miR-661 expression in SNAI1-induced epithelial to mesenchymal transition contributes to breast cancer cell invasion by targeting Nectin-1 and StarD10 messengers. Oncogene *29*, 4436–4448.

Viana, M.P., Tanck, E., Beletti, M.E., and Costa, L. da F. (2009). Modularity and robustness of bone networks. Mol. Biosyst. *5*, 255–261.

Vickers, K.C., Palmisano, B.T., Shoucri, B.M., Shamburek, R.D., and Remaley, A.T. (2011). MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. Nat. Cell Biol. *13*, 423–433.

Vidigal, J.A., and Ventura, A. (2014). The biological functions of miRNAs: lessons from in vivo studies. Trends Cell Biol. *25*, 137–147.

Villa, C., Fenoglio, C., De Riz, M., Clerici, F., Marcone, A., Benussi, L., Ghidoni, R., Gallone, S., Cortini, F., Serpente, M., et al. (2011). Role of *hnRNP-A1* and miR-590-3p in Neuronal Death: Genetics and Expression Analysis in Patients with Alzheimer Disease and Frontotemporal Lobar Degeneration. Rejuvenation Res. *14*, 275–281.

Vinther, J., Hedegaard, M.M., Gardner, P.P., Andersen, J.S., and Arctander, P. (2006). Identification of miRNA targets with stable isotope labeling by amino acids in cell culture. Nucleic Acids Res. *34*, e107.

Vlachos, I.S., Paraskevopoulou, M.D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I.-L., Maniou, S., Karathanou, K., Kalfakakou, D., et al. (2015). DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Res. *43*, D153–D159.

Volinia, S., Galasso, M., Costinean, S., Tagliavini, L., Gamberoni, G., Drusco, A., Marchesini, J., Mascellani, N., Sana, M.E., Jarour, R.A., et al. (2010). Reprogramming of miRNA networks in cancer and leukemia. Genome Res. *20*, 589–599.

Waddington, C.H. (1959). Canalization of development and genetic assimilation of acquired characters. Nature *183*, 1654–1655.

Wakiyama, M., Takimoto, K., Ohara, O., and Yokoyama, S. (2007). Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. Genes Dev. *21*, 1857–1862.

Wang, A., Yue, F., Li, Y., Xie, R., Harper, T., Patel, N.A., Muth, K., Palmer, J., Qiu, Y., Wang, J., et al. (2015). Epigenetic Priming of Enhancers Predicts Developmental Competence of hESC-Derived Endodermal Lineage Intermediates. Cell Stem Cell *16*, 386–399.

Wang, J., Lu, M., Qiu, C., and Cui, Q. (2010). TransmiR: a transcription factor-microRNA regulation database. Nucleic Acids Res. *38*, D119–D122.

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

Wang, X., and Wang, X. (2006). Systematic identification of microRNA functions by combining target prediction and expression profiling. Nucleic Acids Res. *34*, 1646–1652.

Watts, D.J., and Strogatz, S.H. (1998). Collective dynamics of “small-world” networks. Nature *393*, 440–442.

Weisstein, E.W. Königsberg Bridge Problem.

Wells, S.E., Hillner, P.E., Vale, R.D., and Sachs, A.B. (1998). Circularization of mRNA by Eukaryotic Translation Initiation Factors. Mol. Cell *2*, 135–140.

Whitacre, J.M., and Bender, A. (2010). Networked buffering: a basic mechanism for distributed robustness in complex adaptive systems. Theor. Biol. Med. Model. *7*, 20.

Whitaker, W.R., Davis, S.A., Arkin, A.P., and Dueber, J.E. (2012). Engineering robust control of two-component system phosphotransfer using modular scaffolds. Proc. Natl. Acad. Sci. U. S. A. *109*, 18090–18095.

White, J.G., Southgate, E., Thomson, J.N., and Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *314*, 1–340.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell *75*, 855–862.

Van Wijk, B.C.M., Stam, C.J., and Daffertshofer, A. (2010). Comparing Brain Networks of Different Size and Connectivity Density Using Graph Theory. PLoS One *5*, e13701.

Witkos, T.M., Koscianska, E., and Krzyzosiak, W.J. (2011). Practical Aspects of microRNA Target Prediction. Curr. Mol. Med. *11*, 93–109.

Wobus, A.M., and Boheler, K.R. (2005). Embryonic stem cells: prospects for developmental biology and cell therapy. Physiol. Rev. *85*, 635–678.

Worm, J., Stenvang, J., Petri, A., Frederiksen, K.S., Obad, S., Elmén, J., Hedtjärn, M., Straarup, E.M., Hansen, J.B., and Kauppinen, S. (2009). Silencing of microRNA-155 in mice during acute inflammatory response leads to derepression of c/ebp Beta and down-regulation of G-CSF. Nucleic Acids Res. *37*, 5784–5792.

Wozniak, M.B., Scelo, G., Muller, D.C., Mukeria, A., Zaridze, D., and Brennan, P. (2015). Circulating MicroRNAs as Non-Invasive Biomarkers for Early Detection of Non-Small-Cell Lung Cancer. PLoS One *10*, e0125026.

Wu, C.-I., Shen, Y., and Tang, T. (2009). Evolution under canalization and the dual roles of microRNAs: a hypothesis. Genome Res. *19*, 734–743.

Wu, N., Sulpice, E., Obeid, P., Benzina, S., Kermarrec, F., Combe, S., and Gidrol, X. (2012a). The miR-17 family links p63 protein to MAPK signaling to promote the onset of human keratinocyte differentiation. PLoS One *7*, e45761.

Wu, S., Huang, S., Ding, J., Zhao, Y., Liang, L., Liu, T., Zhan, R., and He, X. (2010). Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3' untranslated region. Oncogene *29*, 2302–2308.

Wu, W., Hu, Z., Qin, Y., Dong, J., Dai, J., Lu, C., Zhang, W., Shen, H., Xia, Y., and Wang, X. (2012b). Seminal plasma microRNAs: potential biomarkers for spermatogenesis status. Mol. Hum. Reprod. *18*, 489–497.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. *37*, D105–D110.

Xu, G., Fewell, C., Taylor, C., Deng, N., Hedges, D., Wang, X., Zhang, K., Lacey, M., Zhang, H., Yin, Q., et al. (2010). Transcriptome and targetome analysis in MIR155 expressing cells using RNA-seq. RNA *16*, 1610–1622.

Xu, J., Li, C.-X., Li, Y.-S., Lv, J.-Y., Ma, Y., Shao, T.-T., Xu, L.-D., Wang, Y.-Y., Du, L., Zhang, Y.-P., et al. (2011a). MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. Nucleic Acids Res. *39*, 825–836.

Xu, J., Li, C.-X., Lv, J.-Y., Li, Y.-S., Xiao, Y., Shao, T.-T., Huo, X., Li, X., Zou, Y., Han, Q.-L., et al. (2011b). Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. Mol. Cancer Ther. *10*, 1857–1866.

Yan, G.-R., Xu, S.-H., Tan, Z.-L., Liu, L., and He, Q.-Y. (2011). Global identification of miR-373-regulated genes in breast cancer by quantitative proteomics. Proteomics *11*, 912–920.

Yan, L.-X., Huang, X.-F., Shao, Q., Huang, M.-Y., Deng, L., Wu, Q.-L., Zeng, Y.-X., and Shao, J.-Y. (2008). MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. RNA *14*, 2348–2360.

Yanaihara, N., Caplen, N., Bowman, E., Seike, M., Kumamoto, K., Yi, M., Stephens, R.M., Okamoto, A., Yokota, J., Tanaka, T., et al. (2006). Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell *9*, 189–198.

Yang, J.-S., Maurin, T., Robine, N., Rasmussen, K.D., Jeffrey, K.L., Chandwani, R., Papapetrou, E.P., Sadelain, M., O'Carroll, D., and Lai, E.C. (2010a). Conserved vertebrate mir-451 provides a platform for Dicer-independent, Ago2-mediated microRNA biogenesis. Proc. Natl. Acad. Sci. U. S. A. *107*, 15163–15168.

Yang, J.-S., and Lai, E.C. (2011). Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. Mol. Cell *43*, 892–903.

Yang, Y., Chaerkady, R., Beer, M.A., Mendell, J.T., and Pandey, A. (2009). Identification of miR-21 targets in breast cancer cells using a quantitative proteomic approach. Proteomics *9*, 1374–1384.

Yang, Y., Chaerkady, R., Kandasamy, K., Huang, T.-C., Selvan, L.D.N., Dwivedi, S.B., Kent, O.A., Mendell, J.T., and Pandey, A. (2010b). Identifying targets of miR-143 using a SILAC-based proteomic approach. Mol. Biosyst. *6*, 1873–1882.

Yao, B., Li, S., Lian, S.L., Fritzler, M.J., and Chan, E.K.L. (2011). Mapping of Ago2-GW182 functional interactions. Methods Mol. Biol. *725*, 45–62.

Ye, H., Liu, X., Lv, M., Wu, Y., Kuang, S., Gong, J., Yuan, P., Zhong, Z., Li, Q., Jia, H., et al. (2012). MicroRNA and transcription factor co-regulatory network analysis reveals miR-19 inhibits CYLD in T-cell acute lymphoblastic leukemia. Nucleic Acids Res. *40*, 5201–5214.

Yekta, S., Shih, I.-H., and Bartel, D.P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. Science *304*, 594–596.

Yevgeniy, G. (2011). Introduction to DNA Microarrays.

Yi, R., and Fuchs, E. (2011). MicroRNAs and their roles in mammalian stem cells. J. Cell Sci. *124*, 1775–1783.

Yousef, M., Jung, S., Kossenkov, A. V, Showe, L.C., and Showe, M.K. (2007). Naïve Bayes for microRNA target predictions--machine learning for microRNA targets. Bioinformatics *23*, 2987–2992.

Yu, Z., Wang, C., Wang, M., Li, Z., Casimiro, M.C., Liu, M., Wu, K., Whittle, J., Ju, X., Hyslop, T., et al. (2008). A cyclin D1/microRNA 17/20 regulatory feedback loop in control of breast cancer cell proliferation. J. Cell Biol. *182*, 509–517.

Yuan, B., Liang, Y., Wang, D., and Luo, F. (2015). MiR-940 inhibits hepatocellular carcinoma growth and correlates with prognosis of hepatocellular carcinoma patients. Cancer Sci.

Yuan, Y.-R., Pei, Y., Ma, J.-B., Kuryavyi, V., Zhadina, M., Meister, G., Chen, H.-Y., Dauter, Z., Tuschl, T., and Patel, D.J. (2005). Crystal structure of A. aeolicus argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. Mol. Cell *19*, 405–419.

Yue, D., Liu, H., and Huang, Y. (2009). Survey of Computational Algorithms for MicroRNA Target Prediction. Curr. Genomics *10*, 478–492.

Zacher, B., Abnaof, K., Gade, S., Younesi, E., Tresch, A., and Fröhlich, H. (2012). Joint Bayesian inference of condition-specific miRNA and transcription factor activities from combined gene and microRNA expression data. Bioinformatics *28*, 1714–1720.

Zdanowicz, A., Thermann, R., Kowalska, J., Jemielity, J., Duncan, K., Preiss, T., Darzynkiewicz, E., and Hentze, M.W. (2009). Drosophila miR2 primarily targets the m7GpppN cap structure for translational repression. Mol. Cell *35*, 881–888.

Zekri, L., Kuzuoğlu-Öztürk, D., and Izaurralde, E. (2013). GW182 proteins cause PABP dissociation from silenced miRNA targets in the absence of deadenylation. EMBO J. *32*, 1052–1065.

Zernecke, A., Bidzhekov, K., Noels, H., Shagdarsuren, E., Gan, L., Denecke, B., Hristov, M., Köppel, T., Jahantigh, M.N., Lutgens, E., et al. (2009). Delivery of microRNA-126 by apoptotic bodies induces CXCL12-dependent vascular protection. Sci. Signal. *2*, ra81.

Zhang, J., Chung, T., and Oldenburg, K. (1999). A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. J. Biomol. Screen. *4*, 67–73.

Zhang, Y., Wang, Z., and Gemeinhart, R.A. (2013). Progress in microRNA delivery. J. Control. Release *172*, 962–974.

Zhou, L., Qi, X., Potashkin, J.A., Abdul-Karim, F.W., and Gorodeski, G.I. (2008). MicroRNAs miR-186 and miR-150 Down-regulate Expression of the Pro-apoptotic Purinergic P2X7 Receptor by Activation of Instability Sites at the 3′-Untranslated Region of the Gene That Decrease Steady-state Levels of the Transcript. J. Biol. Chem. *283*, 28274–28286.

Zhou, Q., Chipperfield, H., Melton, D.A., and Wong, W.H. (2007). A gene regulatory network in mouse embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *104*, 16438–16443.

Zhou, X., Duan, X., Qian, J., and Li, F. (2009). Abundant conserved microRNA target sites in the 5'-untranslated region and coding sequence. Genetica *137*, 159–164.

Zhu, S., Cao, L., Zhu, J., Kong, L., Jin, J., Qian, L., Zhu, C., Hu, X., Li, M., Guo, X., et al. (2013). Identification of maternal serum microRNAs as novel non-invasive biomarkers for prenatal detection of fetal congenital heart defects. Clin. Chim. Acta. *424*, 66–72.

Zisoulis, D.G., Kai, Z.S., Chang, R.K., and Pasquinelli, A.E. (2012). Autoregulation of microRNA biogenesis by let-7 and Argonaute. Nature *486*, 541–544.

Annexes

A. Figures supplémentaires

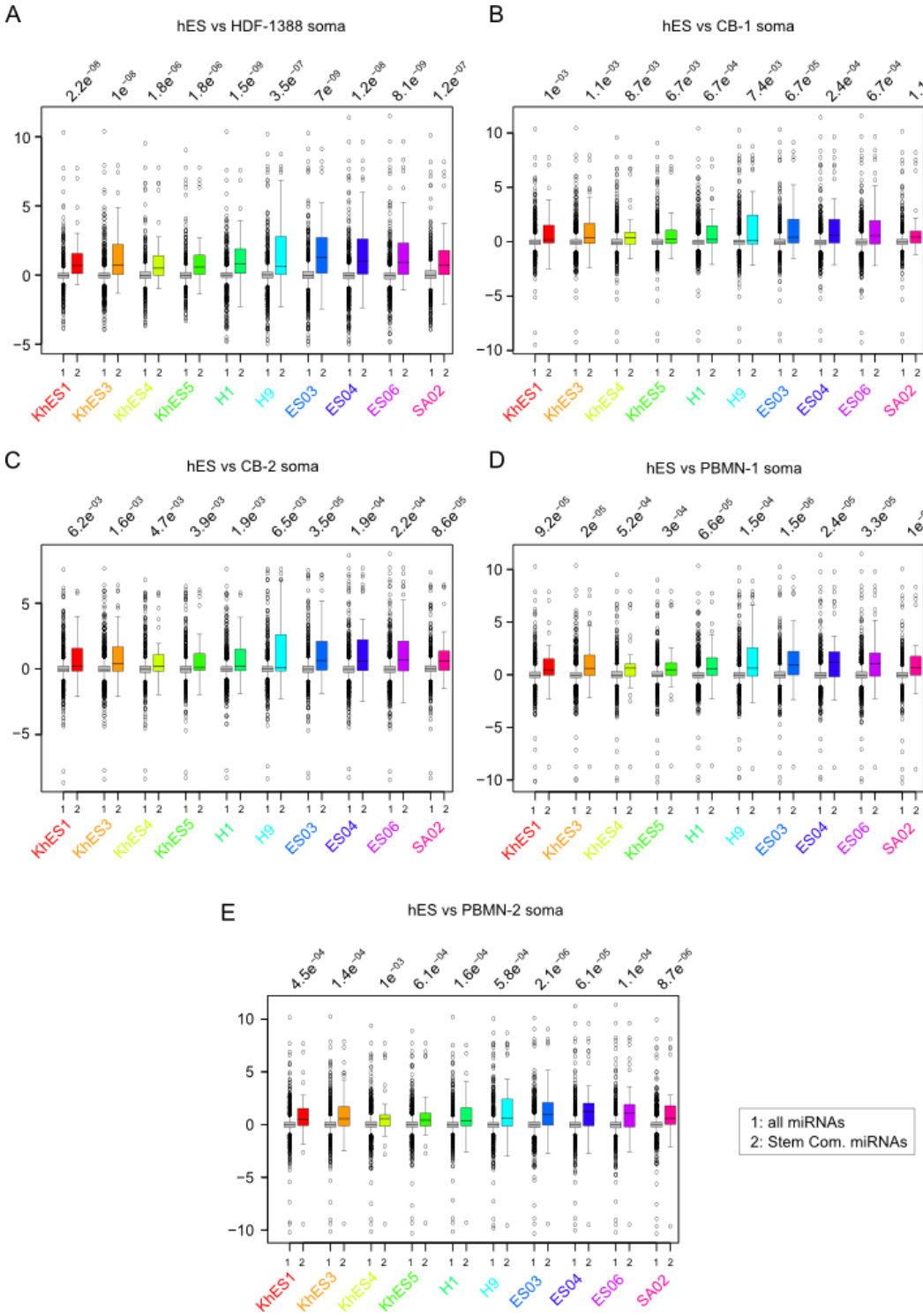


Figure 81. Boxplot des expressions différentielles (logFC) sur dix types de cellules souches contre différents types de cellules somatiques. En gris est représenté l'ensemble des miARN de l'ensemble de données et en couleurs tous les miARN de la communauté souche.

.

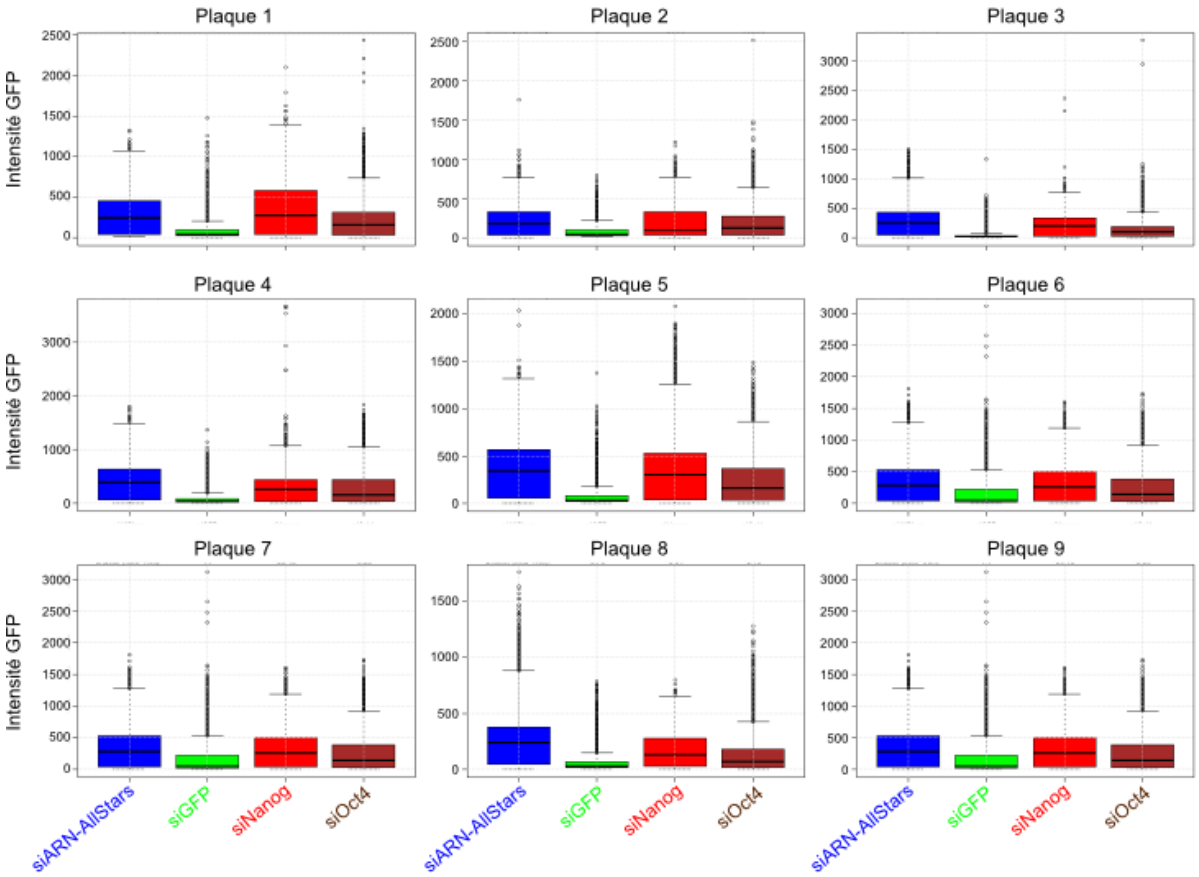


Figure 82. Données brutes des contrôles des neufs plaques du crible. Sont représentés ici uniquement les siARN contrôles. : En bleu est représenté le siARN-AllStars, contrôle négatif. En vert, le siARN contre la GFP ; en rouge, siARN contre Nanog (un autre facteur de transcription des cellules souches) et en brun, le siARN contre Oct4 ; ces trois siARN sont les contrôles positifs.

B. Articles parus dans la presse

INTERVIEW

Arsia AMIR-ASLANI *

Mieux se préparer en complétant sa thèse en sciences par un mastère professionnel

L'insertion professionnelle d'un futur docteur es sciences se prépare dès la première année de thèse affirme Arsia Amir-Aslani, Directeur de Programme de l'*Advanced Master's* « *Biotechnology and Pharmaceutical Management* » et du parcours Mastère/Doctorat à Grenoble Ecole de Management. Les étudiants comme Ricky Bhajun, actuellement en deuxième année de ce double parcours, se réjouissent des opportunités offertes.

Spectra Analyse : Pourquoi avoir créé le mastère avancé « *Biotechnology and Pharmaceutical Management* » dans le parcours Mastère/ Doctorat à Grenoble École de Management ?

Arsia Amir-Aslani : Les difficultés d'insertion auxquelles les docteurs sont confrontés hors du secteur public reflètent surtout l'inadéquation de la formation doctorale et le projet professionnel des doctorants de l'université d'une part et les attentes du marché de l'emploi d'autre part. Or, dans un contexte de multiplication des connaissances interdisciplinaires et de développement de nouvelles technologies par le secteur des biotechnologies, il devient clairement impératif de former les scientifiques à des métiers de managers d'affaires afin qu'ils soient en mesure de gérer des projets R&D en travaillant dans un cadre international et pluridisciplinaire. En effet, la double compétence est devenue une exigence incontournable sur ce marché très compétitif.

Spectra Analyse : La spécialisation doctorale est souvent perçue comme un handicap inhibant la capacité des docteurs es sciences à se tourner vers de nouveaux champs de connaissance. Qu'en pensez-vous ?

AAA : Contrairement aux idées préconçues, c'est précisément cette phase active de leur apprentissage, le « *Learn by doing* », qui va leur permettre de mieux s'adapter et leur donner les atouts pour évoluer avec plus de facilité dans un contexte industriel en modification continuelle. En effet, pendant leurs trois années de thèse, les docteurs es sciences démontrent leurs capacités à identifier, valider, traiter un problème et au final, à prendre une décision tout en étant autonome. Ainsi, ils sont également à même de pouvoir initier une progression de leur apprentissage et de leur profil de formation. Grace à leur capacité de questionner, leur esprit critique et leur aptitude à conduire le changement, ces docteurs sont en mesure de répondre aux demandes des entreprises alors que ces dernières et les intéressés eux-mêmes ignorent le plus souvent ces possibilités. Cette spécialisation permet donc de pallier cette méconnaissance réciproque. Elle rectifie une perception erronée selon laquelle les docteurs seraient uniquement capables de faire de la recherche et non comme des collaborateurs potentiels possédant des profils de compétences adaptables et valorisables à tous les niveaux de l'entreprise.

Spectra Analyse : Comment comptez-vous enseigner le management aux docteurs es sciences ?

AAA : Pour les adultes en formation, l'une des exigences actuelle est d'apprendre plus vite et mieux. Il faut par conséquent que les contenus soient en lien direct avec les attentes et les objectifs des entreprises et de leurs futurs cadres. D'où l'adoption d'une démarche inversée, car il faut aborder la problématique en s'interrogeant sur ces attentes et ces objectifs. La question étant de savoir ce que les docteurs es sciences doivent savoir faire à l'issue d'un tel enseignement.

Spectra Analyse : quelles sont ces compétences à acquérir ?

AAA : Il a notamment été jugé essentiel qu'avec l'acquisition des connaissances spécifiques nécessaires pour bien comprendre le secteur des biotechnologies, les docteurs bénéficient d'une approche pédagogique innovante permettant d'assurer le lien entre la recherche et la vie économique. En effet, dans un contexte industriel, ils doivent pouvoir se « *reposer sur les deux pieds* » que sont la compétence technique et la maîtrise opérationnelle. C'est pour cette raison nous insistons sur une approche par les processus d'innovation visant à comprendre la problématique du point de vue des porteurs d'innovation et selon l'ensemble de ses dimensions : humaines et relationnelles, financières, économiques et sociales.

Spectra Analyse : En quoi cette approche diffère-t-elle de l'enseignement doctoral classique ?

AAA : Les initiatives telles que les *Doctoriales* existant au sein des écoles doctorales des

* Contact : E-mail : arsia.amir-aslani@grenoble-em.com

Opportunités

INTERVIEW

différentes universités constituent surtout un éveil intellectuel. Celui-ci est trop isolé du contexte et s'avère insuffisant à rendre les doctorants opérationnels. À Grenoble Ecole de Management, et dans le cadre de notre participation au sein du campus GIANT, le cursus proposé permet aux doctorants de suivre l'*Advanced Master's « Biotechnology & Pharmaceutical Management »*. Ce programme leur permet d'acquérir les compétences managériales spécifiques du secteur des sciences de la vie. Il se déroule sur une période de trois ans avec un volume horaire compris entre 350 et 400 heures. Les futurs docteurs es sciences ont l'obligation de compléter le mastère en 3 ans, en choisissant les différents modules à compléter. Ce qui se passe d'une manière parfaitement flexible, de sorte à ne pas léser leurs travaux de thèse. Par ailleurs, la formation est centrée sur le renforcement des liens entre recherche et industrie, entre nouvelles technologies et marchés. Cet aspect, incite les doctorants à regarder leurs propres travaux de recherche dans une perspective industrielle.

Spectra Analyse : Les étudiants doctorants sont-ils aidés dans ce parcours ?

AAA : Il faut savoir que GRAL finance chaque année un nouveau doctorant pendant trois ans. Cette année, afin de rendre cette formation accessible à un plus grand nombre de candidats, Grenoble École de Management, a décidé en accord avec son engagement au sein du campus GIANT, de proposer deux bourses d'études de 5000 €/an/étudiant pendant trois ans. Elles permettent à certains des futurs docteurs es sciences de suivre cette formation dans de meilleures conditions.

Témoignage

Entretien avec Ricky Bhajun, doctorant et en deuxième année du parcours Mastère/Doctorat

Spectra Analyse : Quel a été votre parcours universitaire ?

RB : Mon parcours a été classique. Après une 1ère année commune de licence en biologie, je me suis orienté vers la bio-informatique, notamment pour la double compétence que cette discipline assez récente permettait d'acquérir. Après ma licence, j'ai poursuivi par un master en bio-informatique et en biologie structurale que j'ai obtenu avec les honneurs. C'est donc tout naturellement que je me suis orienté vers un doctorat en bio-informatique. À l'heure actuelle, j'effectue ma thèse au CEA de Grenoble où je travaille notamment sur les cribles ARN interférent (ARNi), un domaine en pleine expansion dans le domaine des biotechnologies.

Spectra Analyse : Quelles étaient vos motivations pour acquérir la double compétence science/management?

RB : L'une de mes premières motivations a été la curiosité. Dans le cadre du Labex GRAL (laboratoires d'excellence : *Alliance Grenobloise pour la Biologie Structurale et Cellulaire Intégrées*), un appel d'offre a été lancé pour l'obtention d'une bourse permettant à un doctorant en 1ère année de suivre – en même temps que sa thèse – la formation en management des industries pharmaceutiques et biotechnologiques. J'ai été l'heureux élu cette année-là. Aujourd'hui je pense que c'est à l'occasion de mes premiers cours en management que je me suis découvert une réelle passion pour cet univers qui me plaçait à

Ricky Bhajun effectue sa thèse au CEA à Grenoble sur les cribles ARN interférent tout en suivant le mastère de Grenoble école de management.

Mieux se préparer en complétant sa thèse en sciences par un mastère professionnel

l'interface des sciences et du management. Bien entendu, les avantages de cette position se sont ensuite révélés au fil du temps et leur mise à profit est devenue ma principale source de motivation : j'allais bénéficier d'une triple compétence en biologie/informatique/management avec de meilleures opportunités professionnelles à la sortie même de la thèse.

Spectra Analyse : Comment avez-vous pu concilier cette thèse de management avec la formation MS Bio ?

RB : Cela ce fait essentiellement grâce à une bonne gestion du temps car j'effectue le mastère spécialisé sur trois années consécutives, ce qui permet de répartir la charge de travail annuelle qu'aurait un étudiant classique sur ces trois années. Les différents cours que je suis à l'école de commerce sont également validés par l'école doctorale (ED) et ce dans le cadre des unités d'enseignement obligatoires en vue d'une insertion professionnelle postdoctorale. Je n'ai donc que très peu de cours à suivre auprès de l'ED. Un dernier point est la complémentarité de ma double formation biologie et informatique et certains aspects du management, je pense en l'occurrence à la prédiction sur séries temporelles, une compétence statistique qui m'a permis d'impressionner quelques membres de la promotion 2012-2013 !

Spectra Analyse : Quels sont les bons points de cette formation pour un thésard?

RB : J'y ai appris certains rouages des industries pharmaceutiques et biotechnologiques qui m'étaient jusqu'alors complètement inconnus. J'ai notamment pu aborder la complexité de la tâche consistant à mener une simple idée en science jusqu'à sa concrétisation, que ce soit une technologie ou un nouveau médicament. J'ai également compris comment présenter son produit afin d'apporter de la valeur à ses clients et de les captiver. Clairement, cette formation m'a permis de gagner en assurance dans mes présentations tout en me donnant une vision plus globale de la science et de son potentiel en matière de business. Ma formation est cependant loin d'être terminée et je ferai très probablement de nouvelles découvertes durant ces deux prochaines années.

Spectra analyse : Cette formation vous permet-elle de porter un nouveau regard sur vos travaux de thèse ?

RB : Oui clairement. Je dirais avoir grâce à elle une meilleure vision d'ensemble de mon projet. Je comprends mieux la façon de l'insérer dans la thématique du laboratoire et les autres travaux de recherche, et en particulier, la façon dont ils pourraient apporter de la valeur en dehors du laboratoire. Bien sûr, ce nouveau regard ne se limite pas à mes travaux : au département des sciences du vivant du CEA, j'ai la chance d'être dans un univers très orienté « nouvelles technologies » et où les start-ups en biotechnologie ne sont pas rares. Aujourd'hui et de par cette formation, je suis en mesure de comprendre les choix de ces *start-ups* pour s'installer dans un marché et une économie difficile, mais également capable d'avoir un regard critique sur certains de ces choix.

Spectra Analyse : Cette formation aura-t-elle un impact sur votre choix de carrière ?

RB : Ma première année de formation a déjà modifié mon choix de carrière. Il serait dommage de ne pas valoriser cette double compétence scientifique et managériale d'ici la fin de ma thèse car c'est une opportunité en or. À la fin de ma thèse, je m'orienterai donc très probablement vers l'industrie privée. Cependant, quant à savoir si je limiterai à terme mes activités au management de la science ou à la science uniquement, c'est une autre histoire.

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
RNAI
FUNCTIONAL CLUSTERING
NETWORK TOPOLOGY

Received
8 August 2014

Accepted
14 January 2015

Published
12 February 2015

Correspondence and requests for materials should be addressed to X.G. (xavier.gidrol@cea.fr)

* These two authors equally contributed to the work.

A statistically inferred microRNA network identifies breast cancer target miR-940 as an actin cytoskeleton regulator

Ricky Bhajun[1,2,3]*, Laurent Guyon[1,2,3]*, Amandine Pitaval[1,2,3], Eric Sulpice[1,2,3], Stéphanie Combe[1,2,3], Patricia Obeid[1,2,3], Vincent Haguet[1,2,3], Itebeddine Ghorbel[1,2,3], Christian Lajaunie[4,5,6] & Xavier Gidrol[1,2,3]

[1]Univ. Grenoble Alpes, iRTSV-BGE, F-38000 Grenoble, France, [2]CEA, iRTSV-BGE, F-38000 Grenoble, France, [3]INSERM, BGE, F-38000 Grenoble, France, [4]Center for Computational Biology - CBIO, Mines ParisTech, F-77300 Fontainebleau, France, [5]Institut Curie, F-75248 Paris, France, [6]INSERM, U900, F-75248 Paris, France.

MiRNAs are key regulators of gene expression. By binding to many genes, they create a complex network of gene co-regulation. Here, using a network-based approach, we identified miRNA hub groups by their close connections and common targets. In one cluster containing three miRNAs, miR-612, miR-661 and miR-940, the annotated functions of the co-regulated genes suggested a role in small GTPase signalling. Although the three members of this cluster targeted the same subset of predicted genes, we showed that their overexpression impacted cell fates differently. miR-661 demonstrated enhanced phosphorylation of myosin II and an increase in cell invasion, indicating a possible oncogenic miRNA. On the contrary, miR-612 and miR-940 inhibit phosphorylation of myosin II and cell invasion. Finally, expression profiling in human breast tissues showed that miR-940 was consistently downregulated in breast cancer tissues

MicroRNAs are a class of endogenous, small (19–25 nucleotides), single-stranded non-coding RNAs that regulate gene expression in all eukaryotic organisms. In metazoans, microRNAs most commonly bind to the 3′ untranslated region (3′UTR) of their mRNA target transcript and cause translational repression and/or mRNA degradation. Every microRNA is predicted to regulate from a dozen to thousands of genes, including transcription factors. This fine-tuning of protein expression is known to be involved in many physiological processes, such as development, apoptosis, signal transduction and even cancer progression[1,2]. More than 2,000 mature human microRNAs are listed in the 20th release of miRBase: http://www.mirbase.org (2014) (Date of access:19/08/2013), and some authors hypothesise that the majority of human genes are regulated by microRNAs[3].

Since their discovery in 1993[4], a fair understanding of their role in animal development and in the onset and progression of diseases[2], as well as of their potential use in therapies[5], has been gathered. However, the cooperative behaviour of microRNAs is still under investigation. A growing body of experimental evidence suggests that microRNAs can regulate genes through complementarity, meaning that microRNAs can act together to regulate individual genes or groups of genes involved in similar processes[6]. For example, Hu and co-workers demonstrated that transducing a cocktail of precursor microRNAs (miR-21, miR-24 and miR-221) can result in more effective engraftment of transplanted cardiac progenitor cells[7]. Consistent with these discoveries, Zhu *et al.* demonstrated that miR-21 and miR-221 coregulate 56 gene ontology (GO) processes[8]. In the same study, the authors also showed that cotransfection of miR-1 and miR-21 increases H_2O_2-induced myocardial apoptosis and oxidative stress.

These recent findings support the idea of microRNA-mediated cooperative regulation but also argue for the use of systemic approaches, notably based on graph theory, to decipher individual and complementary roles of microRNAs. Some work has been conducted to use recent high-throughput experiment-derived data sets to infer microRNA synergistic relationships[9–12]. Herein, we present a microRNA network based on target similarities among microRNAs to infer clusters of microRNAs. Clusters are defined as groups of microRNAs sharing a set of common targets, predicted by either DIANA-microT v3[13] or TargetScan v6.2[14]. Some authors have used GO enrichment analysis as a confirmatory tool for their clustering approach[11]. In our case, GO enrichment is not used to infer networks but as a way to estimate the probable metabolic pathway(s) a cluster of microRNAs could co-regulate. Moreover, the novelty of our approach is to consider not only clusters of microRNAs but also

"microRNA hubs", *i.e.,* highly connected microRNAs presenting a crucial role in the network. We further defined these interconnected microRNA hubs as "assorted clubs" of microRNAs.

This target-based microRNA network shows many similarities with known biological networks and is constructed around two microRNA assorted clubs. These two clubs influence the overall shape of the network and thus the microRNAs connected to them. One of the assorted clubs was predicted to play a role in small GTPase signalling. Small GTPase proteins are divided in two subfamilies: members of the Ras subfamily which regulate cell proliferation and differentiation, and members of the Rho subfamily which control cytoskeleton and cell motility but can also act on proliferation[15]. Strikingly, all three microRNAs in the club, miR-612, miR-661, and miR-940, efficiently downregulate small GTPase signalling. However, their cellular function diverges showing that microRNAs acting on similar pathways can lead to opposite outputs. Indeed, Transwell assays and wound healing assays demonstrate that over-expression of miR-661 leads to a dramatic increase in cell motility, while miR-612 and miR-940 reduce this capacity. In addition, miR-940 was found consistently downregulated in breast cancer tissues, indicating a putative role of this microRNA in cancer progression.

Results

A target-based microRNA complementary network. To evaluate how microRNAs could act together on cellular processes, we intended to infer networks based on microRNA target sharing. We used the genome-wide DIANA-microT v3.0 prediction database[13], comprising 555 human microRNAs, 18,986 genes and nearly 2 million interactions. We considered only *Homo sapiens* interactions and did not take into account any score but rather used all available information. In consequence, around 60% microRNAs are predicted to have between 2,000 and 5,000 different target sites and four microRNAs (miR-495, miR-548c-3p, miR-590-3p and miR-603) are predicted to exhibit affinity towards more than 10,000 target sites – which represents around 6000 genes (Supplementary Figures S1a & S1b). Furthermore, 193 genes are targeted by more than 293 microRNAs (Supplementary Figure S1c), which is the 99th quantile of the microRNA-mediated target regulation histogram. These 193 proteins will therefore be referred to as gene (or protein) hubs[9].

Following Shalgi *et al.*[9], we built a target-based microRNA networks based on the idea that if two microRNAs share a common set of genes, they could act on the same pathway(s) and compensate for each other, or act complementarily, on this pathway. Considering the DIANA-microT v3.0 predictions, each node of the network corresponds to a microRNA and each edge between two nodes to the proportion of shared targets between two microRNAs (Figure 1). We used the meet/min metric (or Simpson index) to infer the strength of the edge between two microRNAs[16,17]. In our case, the meet/min index takes into account the number of shared genes between two microRNAs, divided by the minimum number of regulated genes between the microRNAs. This metric takes its value between 0 and 1, where 1 implies the exact same targets, whereas 0 implies no common target. The network thus constructed contained 555 nodes and 153,735 weighted edges with a density of 1, meaning that all nodes in the network are interconnected with different strengths.

Many algorithms have been implemented in systems biology to analyse weighted graphs, notably in protein-protein interaction (PPI) networks. Some algorithms are based on maximal cliques finding and ranking. Although algorithms based on this method are NP-hard, it is not a problem in PPI networks due to their sparse properties[18]. However these methods are not suited to our network which is highly dense. As a consequence, and in spite of information loss, the weighted graph was simplified into a binary graph by defining a meet/min threshold through a "multiple-thresholds-approach"[19,20], which consists of the comparison of different thresholded networks.

To define an appropriate threshold, we analysed changes in the network properties for different meet/min thresholds. Density and clustering coefficients are common properties used in this sense[21]. The density measures the number of edges compared to the number of awaited edges if every node were connected in the network. In addition to the number of edges and connected nodes in the graphs, we also considered different centrality measurements such as betweenness and degree centrality[22]. Our aim was to keep the maximum number of connected nodes in the graph but still be able to analyse it.

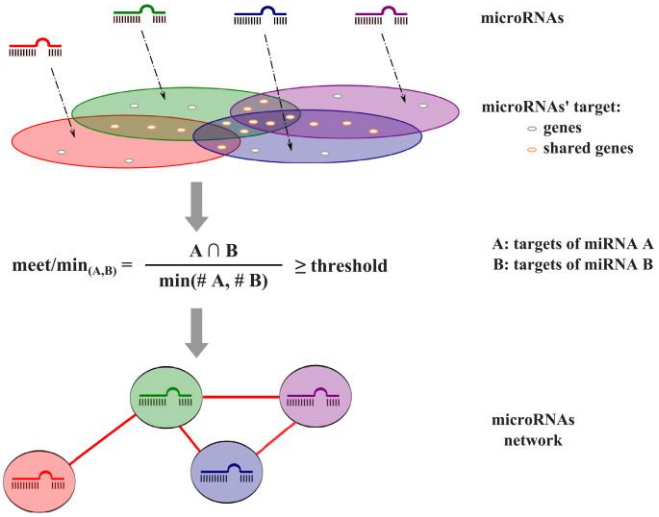


Figure 1 | Construction of the microRNA network. The meet/min metric measures target coverage between microRNAs. A threshold (0.5 is chosen throughout the article) is imposed on the meet/min edges, thus defining a binary network of microRNAs that share common targets. #: number of.

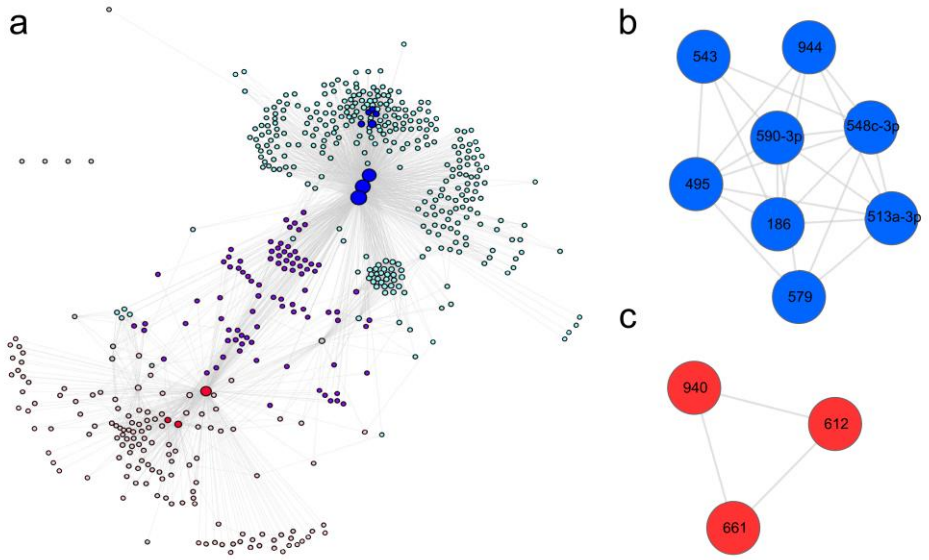


Figure 2 | DIANA-microT network at meet/min 0.5. (a)The network can be divided into two parts (pink and cyan) linked by a few common microRNAs in the middle (purple). In cyan are the nodes that are connected to at least one microRNA of the assorted club 1; in pink are the nodes connected to the assorted club 2; and in purple are the nodes connected to both groups. The nodes not directly connected to the assorted clubs are in grey. Four nodes remained isolated from the entire graph; they are shown in the top left part. Node size is proportional to the node degree. (b) Assorted club 1 has a density of 0.8. It comprises miR-495, miR-548c-3p, miR-590-3p, miR-186, miR-579, miR-513a-3p, miR-543 and miR-944. (c) The assorted club 2 has a density of 1 and is composed of miR-661, miR-612 and miR-940.

Every node is connected to at least one neighbour in the graph until meet/min reaches 0.48, at which point the number of edges decreases sharply (Supplementary Figure S1d). At meet/min 0.4, the density is close to 0 (Supplementary Figure S1e), meaning that the graph is sparse (scattered nodes that are not highly connected to each other), in contrast to the weighted one. The global clustering coefficient gives the degree to which nodes in a graph tend to form clusters (or to put it more simply: the number of "triangles" in the network). The clustering coefficient reaches its minimum near meet/min 0.5 (Supplementary Figure S1e). Finally, centrality measurements evaluate the centrality of each node within the network. Betweenness centrality gives for a given node the number of shortest paths from all pairs of nodes that pass through this node, whereas degree centrality gives information on the degree of this node, that is, the number of edges linked to the node. Node-level centrality measurements can additionally be averaged into single graph-level scores. In our case, the two graph-level centrality measurements are high at this 0.5 meet/min threshold (Supplementary Figure S1e and Supplementary Table I). Furthermore, the network contains 555 nodes and 2,911 edges with a density of 0.02 (Figure 2a). Increasing the threshold further leads to networks formed of isolated modules where only microRNA families can be found (let-7, miR-17/miR-93 cluster, etc.) and whose seed sequences are almost identical in each cluster.

To fully appreciate centralities, we compared 3 graphs for 3 different meet/min thresholds (0.25, 0.5, 0.75) to a random network with 555 nodes and 2,911 edges[23], a scale-free graph iterative construction with 555 nodes[24] and a human PPI based on the yeast two-hybrid system[25] (Supplementary Table I). The density of the human interactome graph is much lower than the other graphs except for the meet/min 0.75 and meet/min 1 graphs, where only microRNA families are found (*e.g.*, the let-7 family). The centrality measurements of the meet/min 0.5 graph are in general much higher than every other graph – as is the clustering coefficient – showing an underlying organization of the meet/min 0.5 network based on hubs and closely linked groups of microRNAs. With a diameter (longest of the shortest paths between any two nodes) of 5 and an average path length (average shortest path between all possible pairs of nodes) of 2.5, the meet/min 0.5 graph is a rather compact graph (Figure 2a and Supplementary Table I). We thus set the meet/min threshold to 0.5 as follows: imposing the condition that two microRNAs are connected in the graph only if they share 50% of their targets. Under this condition, the graph shows a slight scale-free behaviour (R^2 = 0.64), is formed of modules and is dissassortative (Supplementary Figure S2) – a dissassortative network being a network where low degree nodes tend to connect more often to higher degree nodes. It also tends to be a small-world network with high centrality measurements where information is easily transmitted from one node to another, and with central hubs coordinating information. Small-world networks are typical networks where nodes are not all connected to others but are easily reachable through other common nodes. Interestingly, it is at this threshold that we observed these specific characteristics, which are in concordance with our current understanding of biological networks[26]. Although threshold choice is always subjective, we compensated for this arbitrary factor by applying an exploratory statistical analysis to the whole graph.

Deciphering "assorted clubs" of microRNAs. Barabási and Oltvai defined "modules (or clusters)" as highly interconnected groups of nodes[21]. In our model, a cluster comprises interconnected microRNAs that all share a high number of targets.

To test and decipher the underlying organization based on previously described hubs, we followed a "rich-club strategy"[27]. According to the authors, a rich club can be defined as interconnected hubs in a

network with a density of 1 (clique), *i.e.*, a group with a central and influential role. In our case, as the density is high but below 1, we chose to name the groups "assorted clubs" referring to the assortative behaviour of these interconnected and central microRNAs. The analysis of the density formed by the induced subgraph of the *i* first nodes of highest degree (that is, the hubs sorted by their degree) reveals the presence of two assorted clubs (Supplementary Figure S3). At 11 hubs, the first assorted club (assorted club 1) is formed by 8 microRNAs with density of 0.8 and is further emphasised in blue in Figure 2a and b. The second assorted club (assorted club 2) is formed by 3 microRNAs (shown in red in Figure 2a and c) with a density of 1. Knowing that the graph global density is 0.02, the high density values highlight the close connectivity between the different microRNAs and, more interestingly, the high number of shared targets between all of them. Their close connections further reinforce the idea of a common co-regulated biological process between the different microRNAs. As the 12th microRNA is neither connected to the first assorted club nor to the second, we decided to define two clubs with the first 11 degree hubs of the network (Figures 2b and 2c). The two clubs are regrouped into one single network with the 48th microRNA (Supplementary Figure S3).

Despite the normalization imposed by the meet/min formulae, there is still a correlation between the number of potential targets of a microRNA and its number of neighbours in the network (Supplementary Figure S4). Thus, the hubs are not only the microRNAs that are highly connected to other microRNAs but also those with the highest number of predicted targets. Interestingly, most of the hubs also have a high node-level betweenness centrality value (ranging from 0.40 to 0.67). Seven out of the 11 hubs presented here can also be found within the 13 first betweenness centrality sorted nodes. These 7 microRNAs (the 3 microRNAs from the assorted club 2 and 4 from the assorted club 1, namely miR-495, miR-548c-3p, miR-590-3p and miR-186) also seem to be placed at key central positions in the network, defining two separate zones (Figure 2a). The other 4 microRNAs are the remaining members of the assorted club 1. They are, however, more offset on the graph, explaining their lower betweenness centrality values.

To further visualise the structure of the graph organised around the central hubs, we color-coded in cyan the microRNAs linked to at least one of the members of the assorted club 1, in pink the neighbours of at least one of the members of club 2, and in purple the microRNAs connected to at least one member of each cluster (Figure 2a). With this colour scheme, we clearly see that there are three parts structured around the two assorted clubs. The purple part delineates a trench between the two clubs (intermediate zone) that are central to the two extreme zones (cyan and pink). We thus named the two extreme zones as the "sphere of influence" of the assorted clubs. This general organization explains the high graph-level centrality measurements that we observe across the network.

Assorted club 1. The assorted club 1 is composed of 8 microRNAs (Supplementary Table II), including the microRNA with the highest degree in the graph (miR-495). The latter is connected to 72% of the miRNome – hereby defined as the 555 microRNAs of DIANA-microT v3. miR-495 is also predicted to target 6,626 different genes and has 13,900 different target sites. On average, the microRNAs of this group target approximately 5,000 genes. As many as 5,276 genes are shared by at least 4 microRNAs (50%) of the cluster, and 540 genes are shared by all 8 microRNAs. Within this club, only miR-495 and miR-543 are clustered on the genome. They are both localised on chromosome 14 and separated by approximately 1,500 base pairs (Supplementary Figure S5).

As this cluster is composed of hubs of the highest degree in the network and because there is a correlation between the number of targets and the number of edges for a microRNA, one would expect this group to have low specificity. This can be explained by the fact that microRNA hubs may have to regulate a large number of genes at the same time. Indeed, we can suppose that the more genes a group of microRNAs has to regulate, the less specificity it will have – as other groups regulating a part of those genes would bring redundancy. As such, within the 540 genes shared by all 8 microRNAs, 17% of shared genes of this cluster are gene hubs. This represents 47% of all protein hubs (Supplementary Figure S6a - Fisher exact test *P*-value = 10^{-90}).

To further interpret the role of microRNA clusters, we looked at the enrichment of gene ontology (GO)[28] for the coregulated genes of each cluster. A Benjamini-Hochberg (BH) correction was used to account for multiple testing hypothesis correction[29], even though we did not consider the corrected *P*-values as pure decision-making values (see the Methods section for a brief discussion). Only the genes that were shared by at least 50% of all members of the clusters were used in this analysis (4 microRNAs). Using the package TopGO[30] to calculate the GO enrichment, we found enrichment in mRNA processing, transcription and gene expression on the biological process (BP) level of GO (Table 1: assorted club 1), for which BH corrected *P*-values (P_{BH}) ranged from 10^{-5} to 10^{-8}. When considering less generic annotations, we found enrichment for protein modification ($P_{BH} = 3 \times 10^{-4}$), endosomal transport ($P_{BH} = 5 \times 10^{-4}$) and regulation of locomotion ($P_{BH} = 10^{-2}$). Consistent with these findings, "nucleus" was found as the localization of a significantly high number of genes targeted by the microRNAs ($P_{BH} = 1.6 \times 10^{-7}$) (Supplementary Table III. Cellular Component). Accordingly, metal ion binding ($P_{BH} = 6.3 \times 10^{-10}$) was found in the molecular function (MF) category (Supplementary Table III. Molecular Function). Although they are very general annotations, the results correlate with DNA/RNA binding and mRNA processing, showing that the cluster seems to regulate transcription regulators.

Unfortunately, a literature review of the 8 microRNAs reveals little information on their cooperative behaviour and their role in the regulation of transcription factors. However, miR-186 and miR-543 are both cited by different studies in cellular aging[31,32], demonstrating the possibility of their coaction. miR-495 and miR-543 were both identified – with other microRNAs – as actors in the epithelial to mesenchymal transition (EMT)[33]. miR-495 is known to have an effect on cell differentiation and proliferation[34–36] and is also known to be a tumour suppressor[37]. Although miR-186 is known to have an effect on a proapoptotic purinergic receptor[38], it is not known whether this microRNA has a direct role on apoptosis. Similarly, miR-590-3p has a role in neuronal death[39], whereas miR-513a-3p is known to be involved in the immune system response mediated by interferon gamma (IFN-γ)[40]. Finally, miR-548c-3p is involved in the DNA repair process by acting on TOP2A translation[41]. Based on existing knowledge of the biological role of the microRNAs in question, it is difficult to draw conclusions about their complementary behaviour.

Nonetheless, the positions of the members of this assorted club are central to the graph, especially for miR-548c-3p, miR-590-3p, and miR-495 (Figure 2b). By their neighbourhood positioning, they clearly define what we have called a "sphere of influence" represented in cyan in Figure 2a. To understand the biological role of this sphere, we also looked at the GO enrichment using the genes that were shared by 25% of the 315 microRNAs (Table 1: sphere of influence club I and Supplementary Table IV). As most of the coregulated genes of the assorted club and its sphere of influence are shared (Supplementary Figure S6c), one could *a priori* anticipate an enrichment correlation between the two. The sphere appears to be involved not only in transcription regulation – just as the assorted club is itself – but also in development and differentiation (P_{BH} ranging from 10^{-4} to 10^{-7}). This further demonstrates how the hubs influence the other microRNAs around them. By further restricting the genes used for the enrichment calculation to target genes shared by at least 50% of the 315 microRNAs, we saw a clear focus of the ontology on "nervous system development" (Supplementary Table V). This statement can

Table 1 | Gene Ontology enrichment of the assorted club 1 and its sphere of influence (biological process). Only the 20th first terms of each list are represented. In green are represented the terms related to nervous system development, a bias induced by the protein hubs

Assorted club I (5,276 genes shared by 50% of the microRNAs – at least 4 microRNAs/8)

GO.ID	Term	Annotated	Significant	classicFisher	BH.pval
GO:2000112	regulation of cellular macromolecule biosynthetic process	2817	989	7.50E-12	3.08E-08
GO:0019219	regulation of nucleobase-containing compound metabolic process	3046	1058	2.10E-11	4.31E-08
GO:0031326	regulation of cellular biosynthetic process	2998	1040	5.00E-11	6.85E-08
GO:0010556	regulation of macromolecule biosynthetic process	2877	1000	8.50E-11	8.73E-08
GO:0051252	regulation of RNA metabolic process	2674	935	1.10E-10	9.04E-08
GO:0009889	regulation of biosynthetic process	3026	1045	1.50E-10	1.03E-07
GO:0051171	regulation of nitrogen compound metabolic process	3121	1074	1.90E-10	1.12E-07
GO:0006355	regulation of transcription. DNA-dependent	2602	909	2.80E-10	1.44E-07
GO:0010468	regulation of gene expression	3004	1035	3.60E-10	1.64E-07
GO:2001141	regulation of RNA biosynthetic process	2618	912	5.10E-10	2.10E-07
GO:0006351	transcription. DNA-dependent	2847	980	1.70E-09	6.35E-07
GO:0031323	regulation of cellular metabolic process	3967	1322	9.20E-09	3.15E-06
GO:0034645	cellular macromolecule biosynthetic process	3685	1231	2.40E-08	7.59E-06
GO:0009059	macromolecule biosynthetic process	3756	1251	3.70E-08	1.09E-05
GO:0032774	RNA biosynthetic process	2909	984	9.50E-08	2.60E-05
GO:0019222	regulation of metabolic process	4375	1435	1.40E-07	3.60E-05
GO:0060255	regulation of macromolecule metabolic process	3757	1245	1.50E-07	3.63E-05
GO:0080090	regulation of primary metabolic process	3884	1283	2.00E-07	4.57E-05
GO:0070647	protein modification by small protein conjugation or removal	612	240	2.80E-07	6.06E-05
GO:0044249	cellular biosynthetic process	4466	1453	1.10E-06	2.26E-04

Sphere of influence Club I (4,208 genes shared by 25% of the microRNAs – at least 79 microRNAs/315)

GO.ID	Term	Annotated	Significant	classicFisher	BH.pval
GO:0007399	nervous system development	1539	472	8.10E-11	3.33E-07
GO:2000112	regulation of cellular macromolecule biosynthetic process	2817	790	7.70E-09	1.58E-05
GO:0010556	regulation of macromolecule biosynthetic process	2877	799	4.20E-08	4.81E-05
GO:0019219	regulation of nucleobase-containing compound metabolic process	3046	840	6.50E-08	4.81E-05
GO:0031323	regulation of cellular metabolic process	3967	1069	7.80E-08	4.81E-05
GO:0048869	cellular developmental process	2496	700	8.10E-08	4.81E-05
GO:0031326	regulation of cellular biosynthetic process	2998	827	8.20E-08	4.81E-05
GO:0009889	regulation of biosynthetic process	3026	832	1.40E-07	6.99E-05
GO:0023052	signaling	4226	1130	1.60E-07	6.99E-05
GO:0035556	intracellular signal transduction	1716	497	1.70E-07	6.99E-05
GO:0030154	cell differentiation	2348	657	3.40E-07	1.20E-04
GO:0007154	cell communication	4338	1154	3.50E-07	1.20E-04
GO:0000902	cell morphogenesis	843	263	4.20E-07	1.33E-04
GO:0006351	transcription. DNA-dependent	2847	782	4.80E-07	1.36E-04
GO:0006355	regulation of transcription. DNA-dependent	2602	720	5.00E-07	1.36E-04
GO:0051252	regulation of RNA metabolic process	2674	738	5.30E-07	1.36E-04
GO:0032502	developmental process	3958	1058	5.90E-07	1.42E-04
GO:0007167	enzyme linked receptor protein signaling pathway	784	246	6.20E-07	1.42E-04
GO:2001141	regulation of RNA biosynthetic process	2618	723	6.60E-07	1.42E-04
GO:0009653	anatomical structure morphogenesis	1827	521	6.90E-07	1.42E-04

be explained by the protein hubs introduced earlier. The ontology enrichment of the 193 protein hubs from DIANA-microT shows the same focus on "nervous system development" (Supplementary Table VI). As the limitation imposed for the second enrichment calculation on the sphere (genes shared by 50% of the microRNAs) includes many of the protein hubs (Supplementary Figure S6e), it biases the result of the enrichment toward nervous development processes.

Assorted club 2. Assorted club 2 is composed of 3 microRNAs (Supplementary Table IV), namely miR-940, miR-661 and miR-612. On average, each microRNA of this cluster is predicted to target 5,254 genes. A total of 4,596 genes are predicted to be regulated by at least 2 microRNAs of the cluster, defining the consensus set of genes used for the GO enrichment. The three microRNAs are localised on three different chromosomes. miR-940 is located on chromosome 16, miR-661 on chromosome 8, and miR-612 on chromosome 11 (Supplementary Figure S5). This second assorted club should exhibit more specificity than the first, as the microRNAs target fewer mRNAs on average than the assorted club 1 microRNAs. Within the 1,830 genes that are shared by all three microRNAs, approximately 6.5% are protein hubs. This represents more than 61% of the DIANA-microT protein hubs (Supplementary Figure S6b - Fisher exact test P-value = 10^{-71}). Fifty-two protein hubs are shared between the assorted club 1 and this club.

"Small GTPase-mediated signal transduction" was the most enriched GO term at the level of BP ($P_{BH} = 5 \times 10^{-3}$), followed by terms involved in cell communication and signalling ($P_{BH} < 0.05$) (Table 2: assorted club 2). A significant number of proteins targeted by this cluster are localised in the membrane (cell membrane, organelle membrane, plasma membrane, etc.) and cell junction ($P_{BH} < 10^{-5}$) (Supplementary Table VII: Cellular Component). Finally, we found enrichment in phospholipid binding (BH corrected P-value = 0.036) when looking at molecular functions (Supplementary Table VII: Molecular Function). No other term passed our statistical criteria, even though small GTPase regulation and binding ("Ras GTPase binding", "Rho guanyl-nucleotide exchange factor activity")

Table 2 | Gene Ontology enrichment of the assorted club 2 and its sphere of influence (biological process). Only the 20th first terms of each list are represented. In green are represented the terms related to nervous system development, a bias induced by the protein hubs

Assorted Club II (4,596 genes shared by 50% of the microRNAs – at least 2 microRNAs/3)

GO.ID	Term	Annotated	Significant	classicFisher	BH.pval
GO:0007264	small GTPase mediated signal transduction	575	186	2.40E-06	5.67E-03
GO:0007154	cell communication	4338	1148	3.00E-06	5.67E-03
GO:0023051	regulation of signaling	1778	503	5.00E-06	5.67E-03
GO:0009966	regulation of signal transduction	1553	444	6.20E-06	5.67E-03
GO:0007399	nervous system development	1539	440	6.90E-06	5.67E-03
GO:0035556	intracellular signal transduction	1716	484	1.10E-05	6.68E-03
GO:0023052	signaling	4226	1113	1.30E-05	6.68E-03
GO:0048583	regulation of response to stimulus	2027	563	1.30E-05	6.68E-03
GO:0007265	Ras protein signal transduction	353	118	3.40E-05	1.55E-02
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	581	180	5.70E-05	2.34E-02
GO:0016192	vesicle-mediated transport	858	254	6.50E-05	2.43E-02
GO:0007165	signal transduction	3789	994	1.00E-04	3.42E-02
GO:0048011	nerve growth factor receptor signaling pathway	219	77	1.30E-04	4.11E-02
GO:0006897	endocytosis	335	110	1.40E-04	4.11E-02
GO:0010646	regulation of cell communication	1317	371	1.50E-04	4.11E-02
GO:0051056	regulation of small GTPase mediated signal transduction	336	110	1.60E-04	4.11E-02
GO:0035725	sodium ion transmembrane transport	16	11	1.90E-04	4.59E-02
GO:0009653	anatomical structure morphogenesis	1827	499	2.70E-04	6.16E-02
GO:0051179	localization	3782	986	3.00E-04	6.49E-02
GO:0007167	enzyme linked receptor protein signaling pathway	784	229	3.50E-04	7.04E-02

Sphere of Influence Club II (3,605 genes shared by 25% of the microRNAs – at least 33 microRNAs/129)

GO.ID	Term	Annotated	Significant	classicFisher	BH.pval
GO:0007154	cell communication	4338	1020	1.40E-18	5.75E-15
GO:0023052	signaling	4226	986	1.30E-16	2.67E-13
GO:0007399	nervous system development	1539	406	1.10E-13	1.51E-10
GO:0023051	regulation of signaling	1778	455	5.30E-13	5.44E-10
GO:0007268	synaptic transmission	574	176	1.10E-11	9.04E-09
GO:0035637	multicellular organismal signaling	652	193	3.30E-11	2.26E-08
GO:0007165	signal transduction	3789	860	7.50E-11	4.40E-08
GO:0007264	small GTPase mediated signal transduction	575	173	9.00E-11	4.57E-08
GO:0009966	regulation of signal transduction	1553	394	1.00E-10	4.57E-08
GO:0019226	transmission of nerve impulse	644	189	1.20E-10	4.93E-08
GO:0007265	Ras protein signal transduction	353	117	2.10E-10	7.84E-08
GO:0048666	neuron development	714	204	2.80E-10	9.59E-08
GO:0022008	neurogenesis	1010	271	4.10E-10	1.30E-07
GO:0035556	intracellular signal transduction	1716	425	5.40E-10	1.58E-07
GO:0048856	anatomical structure development	3476	789	8.30E-10	2.27E-07
GO:0030030	cell projection organization	822	226	1.40E-09	3.60E-07
GO:0030182	neuron differentiation	876	238	1.60E-09	3.87E-07
GO:0007409	axonogenesis	479	145	2.00E-09	4.57E-07
GO:0048699	generation of neurons	952	254	2.80E-09	5.96E-07
GO:0048731	system development	3014	691	2.90E-09	5.96E-07

were also found in the list at less significant *P*-values, confirming the first enrichment described above. Following this clear enrichment in small GTPase signalling, we also found enrichment in cell organization and cellular development, but with higher corrected and uncorrected *P*-values (< 0.001 and < 0.3, respectively). Overall, there was a consistency in enriched GO terms around GTPase signalling, even though *P*-values were not strikingly significant.

To confirm the prediction regarding this cluster, we looked at the localization and quantified the level of phosphorylated myosin light chain II (MLCII) in Retinal Pigment Epithelial (RPE1) cells using phospho-myosin light chain II antibodies. MLCII is a substrate of Rho-associated protein kinases (ROCK) but is also an ideal readout to monitor small GTPase signalling, as it is the end product of this signalling cascade[42]. Additionally, the cells were plated on micropatterned fibronectin to normalise their shape and actin cytoskeleton architecture, which was also monitored via phalloidin distribution[43]. Figure 3a shows RPE1 cells treated by siRNA-AllStars (negative control) and Y27632, a chemical inhibitor of ROCK used as a positive control, on 500 μm^2 circular fibronectine patterns. A decrease in the global phosphorylation of MLCII was observed with Y27632. However, as the patterns were small compared to the size of RPE1 cells, RPE1 cells were constricted, which involved less cellular contraction. 1000 μm^2 circular fibronectine patterns were used in the following experiments so that cells had fewer restrictions and also so that a generally higher level of phosphorylation could be observed. The quantification of MLCII phosphorylation by Western blot on unrestricted cells showed a decrease in phosphorylation with each human microRNA mimic in comparison to the siAllStars negative control (Figure 3b), which confirmed the involvement of the three microRNAs in small GTPase signalling. Furthermore, Figure 3c shows that upon transfection of miR-612 and miR-940 mimics, the cells were relaxed and exhibited the same behaviour as the cells treated with Y27632. Their actin filaments were disorganised, with an absence of stress fibres and transverse arcs compared to cells treated with siRNA-AllStars (Figure 3c). In contrast, the miR-661 mimic revealed a higher number of myosin-decorated stress fibres

www.nature.com/scientificreports

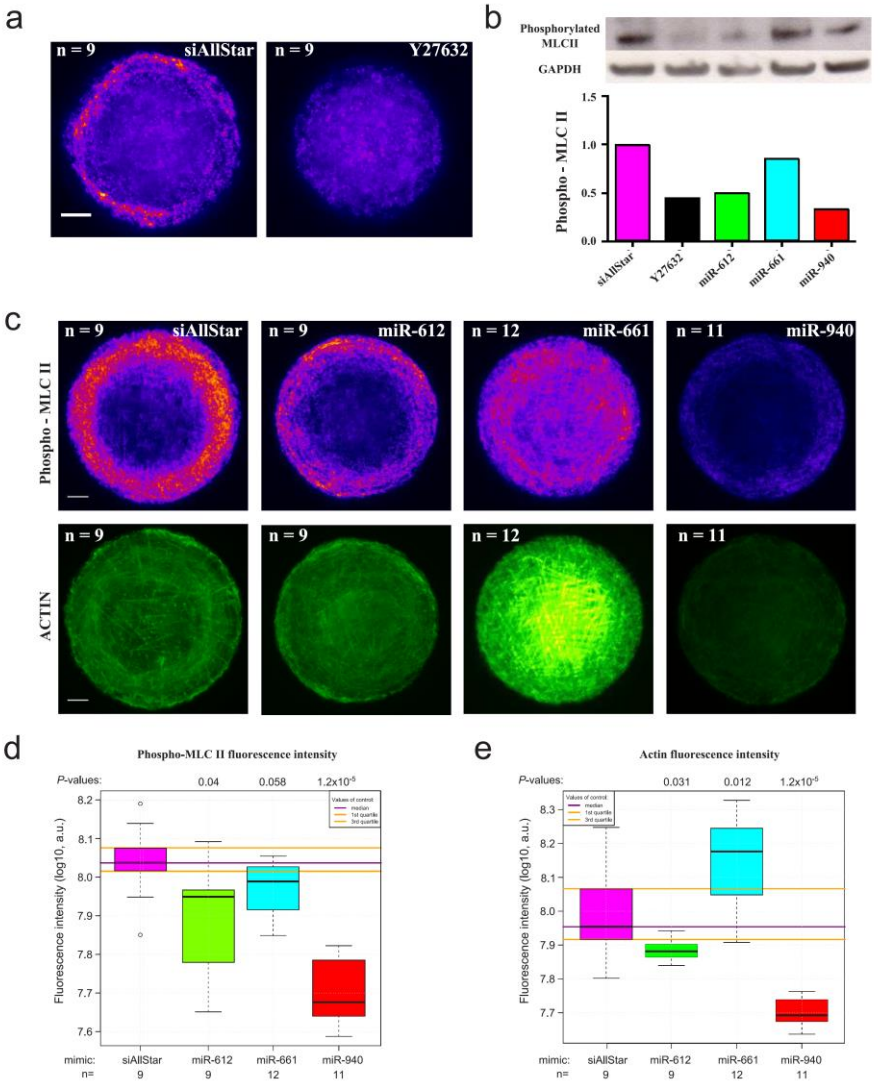


Figure 3 | Involvement of miR-661, miR-612 and miR-940 in small GTPase signalling. (a) Immunostaining of phosphomyosin II and actin filaments of RPE1 cells plated on 500 μm² circular fibronectin patterns. RPE1 cells were transfected with siRNA-AllStars (siAllStars, negative control) and Y27632 (ROCK inhibitor). They were immunolabeled for phosphomyosin II. Nine different images for each condition were taken, aligned and projected into a single image by using the median value of all images for each pixel (Median Z projection of ImageJ). Rescaled with the same conditions, the images were color-coded with the "fire" look-up table to highlight intensity variations. Scale bar, 5 μm. (b) Western blot of phosphomyosin II. RPE1 cells were lysed and supplemented with protease inhibitor. A total of 10 μg of proteins were deposited and hybridised to MLCII antibodies. GAPDH was used as a loading control. The bar plot shows the GAPDH-normalised signal rescaled to siAllStars. (c) Immunostaining of phosphomyosin II and actin filaments. RPE1 cells were transfected with miR-612, miR-661 or miR-940 mimics and immunolabeled for phosphomyosin II and actin fibres on 1000 μm² circular fibronectin patterns. For myosin and actin images, 9 to 12 images were taken, aligned and projected into a single image. They were color-coded with the "fire" and "green hot" look-up table to highlight intensity variations for myosin and actin staining, respectively. Scale bar, 5 μm. (d) Log10 of phosphorylated myosin II fluorescence intensity. The integrated fluorescence intensity of myosin was calculated from single images after cell segmentation for each condition. *P*-values were calculated using the non-parametric two-sided Mann-Whitney test and the number of observations (n) for this calculation. a.u.: arbitrary units. (e) Log10 of actin fluorescence intensity. The integrated fluorescence intensity of actin filaments was calculated from single images after cell segmentation for each condition. *P*-values are calculated using the non-parametric two sided Mann-Whitney test and the number of observations (n) for this calculation. a.u.: arbitrary units.

and highly contracted cells with dense fibres (Figure 3c). However, we clearly observed that miR-661 induced a spatial reorganization of MLCII from the border of the cells to the entire cell surface. With an image-based phosphorylation quantification, we observed that the overall phosphorylation level of MLCII was only significantly reduced with the overexpression of miR-612 and miR-940 (*P*-values = 0.04 and 1.2×10^{-5}, respectively) (Figure 3d). Finally, an increase in actin filament staining was observed in RPE1 cells treated with miR-661 (*P*-value = 0.012, Figure 3e). In contrast, there was a decrease in actin staining following miR-612 and miR-940 treatment (*P*-values = 0.031 and 1.2×10^{-5}, respectively).

A transwell assay and a wound healing assay were carried out to investigate the influence of the three microRNAs on the cytoskeleton dynamics and the ability of the cells to modify their shape and migrate. The aim of the transwell assay was to capture the dynamics of the cell cytoskeleton and the ability of cells to migrate through holes whereas the wound healing assay determined the ability of the cells to divide and migrate. miR-661 clearly produced an increase in the number of cells that went through the transwell membrane compared to the control (Figure 4a, *P*-value = 1.3×10^{-8}, Mann-Whitney test). In the same manner, miR-661 allowed the cells to close the wound faster compared to the control (Figure 4b) but also more constricted cells with increased actin staining (Figure 4c). These two results show that miR-661 greatly enhances cell motility and division when overexpressed. Conversely, miR-612 produced a significant decrease in the number of cells crossing the transwell membrane (*P*-value = 4.9×10^{-36}, Mann-Whitney test) and completely blocked the closure of the wound. The microRNA also induces changes in the general shape of the cells and their interaction with each other. Indeed, the cells are less constricted and form a hollow network resembling epithelium surface (Figure 4c). In the same way, miR-940 also produced a decrease in the number of cells (*P*-value = 6×10^{-8}, Mann-Whitney test) but to a lesser extent, and exhibited no clear difference overall on the wound healing assay compared to the control. However, the impact of the microRNA overexpression on the shape of the cells is still highly visible and follows the same trend as miR-612 with, again, a less marked phenotype (Figure 4c). These results strongly support our ontology prediction with respect to the assorted club 2 and our previous observations from the micropatterned assay where strong effects on cell motility and the cytoskeleton organization was observed.

To assess the regulation power of the network microRNA, we then colour-coded the network according to microRNA expression in 8 different normal tissues (notably breast, colorectal mucosa, lung, prostate, blood, prefrontal cortex, liver and muscle). Interestingly, miR-940 is moderately to highly expressed (pink to red node) in 7 tissues out of 8 whereas miR-661 and miR-612 both show little or no expression (white node) in the same tissues. On the other hand, only miR-612 is expressed in normal colorectal mucosa (Supplementary Figure S7 and S8). The three microRNAs seem rarely co-expressed in the different tested tissues. This is consistent with the phenotypic outputs that we observed after ectopic overexpression of these microRNAs in RPE1 cells.

Differential expressions of the three microRNAs between healthy and cancerous tissues were then carried out on Gene Expression Omnibus (GEO) data. Interestingly, miR-940 was found consistently downregulated in breast cancer (log fold change of −0.26 on GSE44124, -0.53 on GSE31309, and -0.57 on GSE38867). Moreover, two out of the three datasets tested showed a statistically very significant differential expression (Figure 5). These data are consistent with our previous results on the role of miR-940 in cell dynamics and might shed light on a probable role of this microRNA in breast cancer progression.

Just as with assorted club 1, the members of assorted club 2 also define their own sphere of influence. This sphere is composed of 129 microRNAs (pink in Figure 2a). With a limitation to the 4,208 genes shared by 25% of the sphere microRNAs (at least 33 microRNAs), there is enrichment mostly for "signalization" but also for "nervous system development". The corrected *P*-values range from 10^{-6} to 10^{-14} (Table 2: sphere of influence Club 2 and Supplementary Table VIII). The annotations again follow the same trend as the enrichment of the assorted club 2 due to the high number of shared targets (Supplementary Figure S6d). With genes shared by at least 50% of the 129 microRNAs, a higher bias toward nervous system development is observed (Supplementary Table IX). In both cases, the enrichment for brain development can be explained by the protein hubs shared by the microRNAs of the sphere (Supplementary Figure S6f).

The transitory zone of influence (purple on Figure 2) is not highly enriched for any particular process (Supplementary Table X and XI). Indeed, the enrichment of the 3,011 genes shared by at least 25% of the transitory zone (23 microRNAs on 89) is enriched mainly for "transport" with BH corrected *P*-values ranging from 10^{-1} to 10^{-4} (9 first annotations in BP). For genes shared by at least 50% the microRNAs of this zone, no enrichment could be found. So, in general and reassuringly, no clear enrichment is found for the group of microRNAs that are connected to both assorted clubs.

Robustness of the approach. To assess the robustness of our approach, we compared the results obtained using DIANA-microT v3 to those from TargetScan v6.2 (June 2012)[44], a prediction tool also based on seed sequence analysis. We chose to limit this study to the non-conserved version of the TargetScan algorithm, as no real proof indicates that targets which are conserved across species are more accurate than non-conserved targets[45], and also to reduce the number of false negative prediction. Due to the differences in the prediction algorithms, the two databases predict different microRNA targets. These differences in target prediction are further illustrated in Supplementary Figure S9, where we observe that the coverage of microRNA targets between the two databases is – on average – only approximately 60% (meet/min, Supplementary Figure S9). The coverage is slightly higher when considering only the members of the assorted clubs but is still below 70%. The most covered microRNA is miR-513a-3p, with a meet/min value of 84%, whereas the least covered microRNA is miR-543, with a meet/min value of 59% (Supplementary Figure S9a).

Despite the differences in target prediction, the TargetScan network built with the same process described above has a similar two-part structure to that obtained using DIANA-microT (Supplementary Figure S10). The two spheres of influence seen in the DIANA-microT network are also present in the TargetScan network. We also see that the two spheres of influence are likewise organised around a small number of central nodes, comprising the assorted clubs from DIANA-microT. The different properties of the network (clustering coefficient, centrality measurements, number of connected nodes, etc.) are almost identical to those of DIANA-microT network (Supplementary Figure S11). One major difference is the fact that the hubs from the network are different, and as such, the assorted club 2 from the DIANA-microT graph does not come out as an assorted club in TargetScan. However, even though they are no longer hubs, they are still very central to the network and globally still define two spheres of influence (Supplementary Figure S10). The intermediate zone is less obvious, even though the microRNAs from this zone are still mostly located between the two spheres. Despite the fact that we observe three-fold more microRNAs in the TargetScan network, 5 out of the 7 hubs with high betweenness centrality previously described in the DIANA-microT network are found within the 30 first betweenness centrality sorted nodes. The 5 microRNAs are, in decreasing order of centrality: miR-548c-3p, miR-590-3p, miR-661, miR-186 and miR-940. Lastly, 7 hubs from the 11 hubs discovered by DIANA-microT are retrieved within the 40 degree sorted nodes on the TargetScan network (in decreasing order by degree: miR-548c-3p, miR-590-3p,

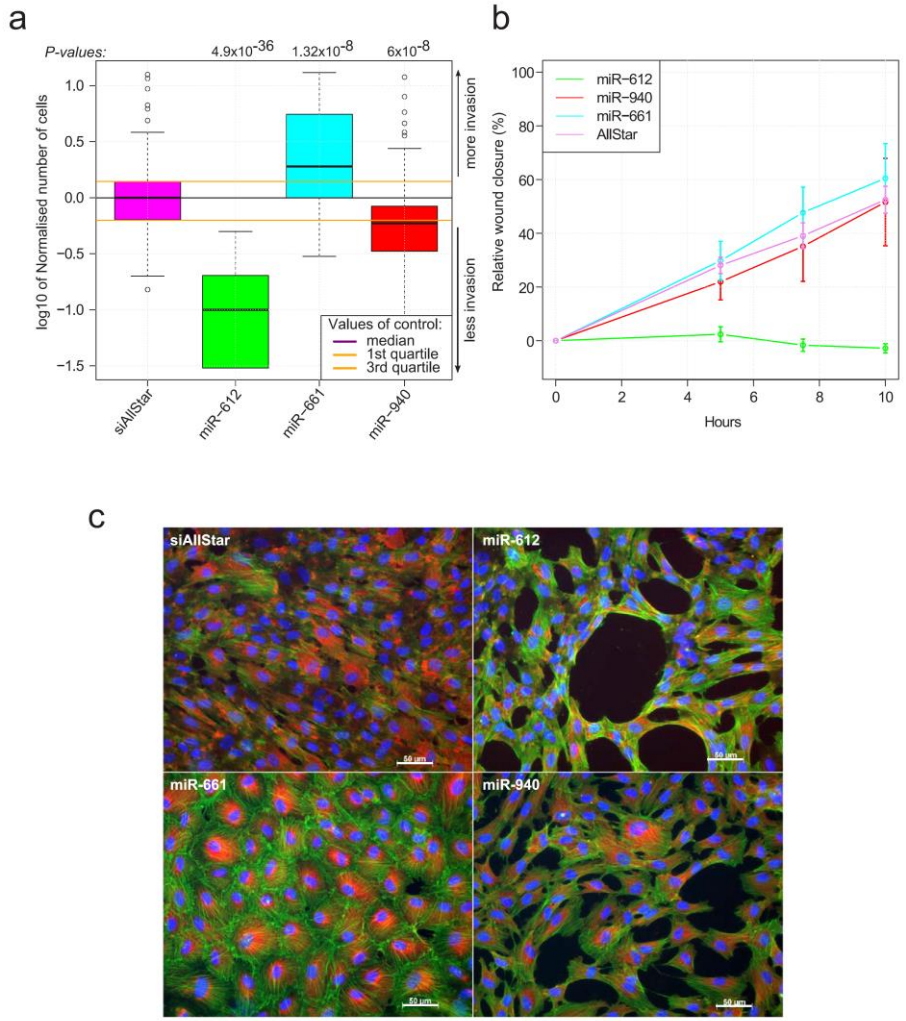


Figure 4 | Effect of miR-612, miR-661 and miR-940 on RPE1 migration and proliferation. (a) Motility graph: Normalised number of cells for the transwell assay. RPE1 cells were independently transfected with mimics of miR-661, miR-612 and miR-940. The number of cells that passed through the 5 μm holes after 18 hours were counted. Four independent experiments were conducted. The cell number was normalised based on the negative control cells, and all four experiments were pooled. *P*-values were calculated using the non-parametric two sided Mann-Whitney test. (b) Wound healing assay. Relative wound closure after a scratch was made in confluent cells transfected by mimics of the three microRNAs. The experiment was conducted on 10 hours with images taken at t_0, t_0+5h, $t_0+7.5h$, and t_0+10h. Each condition was present in triplicates. (c) Vinculin and phalloïdin immunostaining images of the cells transfected by mimics of the three microRNAs. The images were taken at the AxioImager®.

miR-579, miR-186, miR-513a-3p, miR-661, miR-495 and lastly miR-940).

The 11 hubs identified by DIANA-microT were compared with the two prediction algorithms. Interestingly, the two assorted clubs are connected with density profiles that are almost identical in both networks (Supplementary Figure S12a and b). This proves that the connections between the microRNAs are robust across database changes. A major difference is that the two clubs are now connected in the TargetScan network. This connection is made *via* miR-548c-3p (Supplementary Figure S12c). Importantly, when looking at gene ontology enrichment for the two assorted clubs with TargetScan predictions, the results are equivalent to DIANA-microT target ontology enrichment (Supplementary Tables XII and XIII). Surprisingly, even more significant *P*-values are observed for the targets predicted by TargetScan for the assorted club 2 (typically by one order of magnitude), again with a focus on GTPase signalling.

To conclude, even though the hubs are not the same between different target prediction databases and despite all the differences

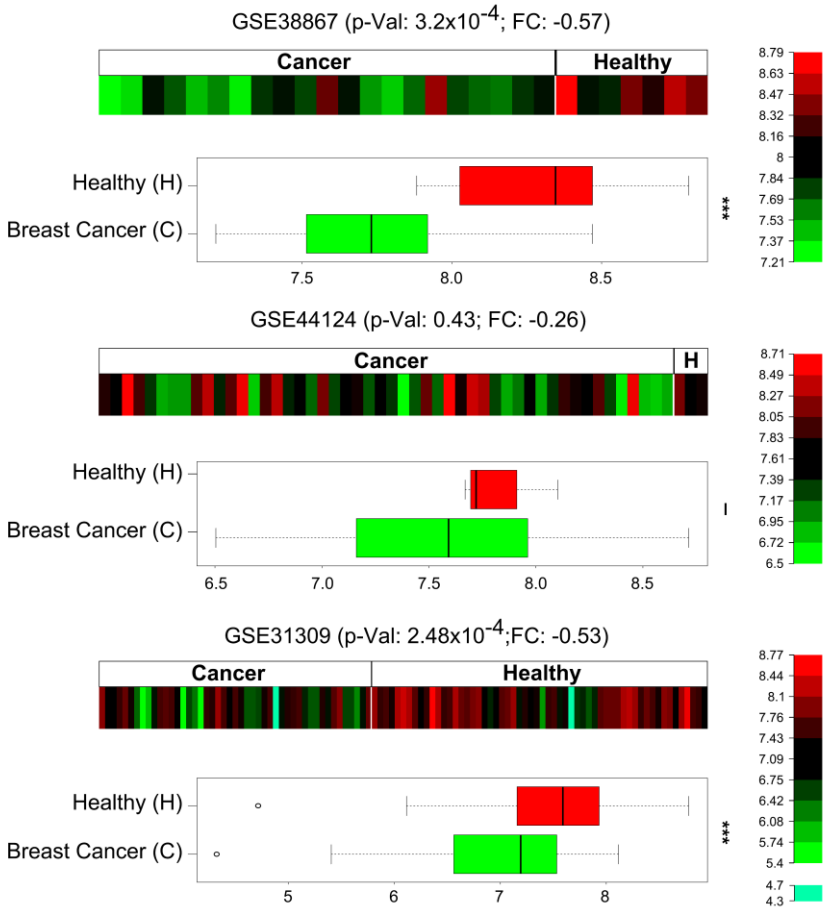


Figure 5 | Relative expression of miR-940 in breast cancer. Expression of miR-940 on three datasets of human breast cancer taken from GEO (GSE38867, GSE44124 and GSE31309). The expression of the microRNA is consistently downregulated in breast cancer tissues on the three experiments. On two out of three microarray sets, miR-940 is differentially expressed with high significance based on limma analysis (p-Value < 0.001).

in target prediction, the relations (target sharing and, therefore, connections) between the different microRNAs remain consistent, as do the ontology prediction and the general shape of the networks.

Discussion

MicroRNAs are crucial entities that regulate diverse biological processes in cells by targeting many different mRNAs. Furthermore, many genes are targeted by at least a few microRNAs. Under these two assumptions, systems biology seems to be the method of choice to characterise the complementary role of microRNAs. Here, we present a way to infer microRNA networks, taking into account target similarities based on the principle that if two microRNAs share similar targets, they may coregulate similar pathways. Tsang *et al.* already suggested that microRNA cotargeting is prevalent in the cell and considered microRNA families in their study[46], where they hypothesised that different microRNAs targeting the same genes would imply a wider range of target-level modulation. In fact, two microRNAs can also be functionally related if they regulate different genes that reside in the same pathways, a question that was addressed by Xu *et al.*[6] when they developed an approach for microRNA networks based not only on target sharing but also on similarities in GO biological processes. In our case, we sought to obtain information on the processes that the group might coregulate rather than to use ontology as a tool to infer networks. More importantly, we focused our analysis on the role of hub microRNAs, i.e., microRNAs that are more connected than others due to their high number of predicted targets.

Although it is known that *in silico* predictions have a high percentage of false positives, we decided to keep every available prediction in our analysis. With the sensitivity of microRNA target prediction algorithms being approximately 66%[47], some studies rely solely on experimentally validated microRNA-gene interactions[48]. However, even though these data may be considered more robust, they are still very limited. For example, only one target is validated for miR-612 in the latest miRTarBase release v4.5[49], and none of those in miRecords v4 are validated[50]. Not only do the predictions have a high false positive rate, but the coverage between different algorithms is

also low (approximately 60% between TargetScan v6.2 and DIANA-microT v3: Supplementary Figure S9). Regardless of the aforementioned issues, the networks built independently with the two algorithms gave almost equivalent results, both in terms of network architecture and in terms of hypotheses for the biological implications of the two groups of interconnected hubs.

The interconnected hubs defined two spheres of microRNAs separated by an intermediate zone. A high correlation between the enrichment of the assorted clubs and their respective spheres could be observed, hence the name "sphere of influence". An important idea in our analysis is the notion of global exploration, meaning that even though the two spheres are mostly involved in their own respective pathways, some microRNAs of the two zones might have little or no involvement in the corresponding pathways. Keeping in mind that we restricted our analysis to a global view of microRNA-mediated biological regulation, we note the link between the structure of the graph (two subnetworks with central hubs) and the biological functions shared by many (but not all) microRNAs in each sphere of influence.

The second central hub group, named "assorted club 2", was composed of miR-612, miR-661 and miR-940. Very little has been reported on the functional role of these microRNAs in human cells. The analysis of GO among gene targets of the 3 microRNAs was enriched in terms related to small GTPase-mediated signal transduction and, as a consequence, may show an involvement of the 3 microRNAs on the cytoskeleton and affect cell motility. We performed functional validation experiments and confirmed that mimics of each of these microRNAs were acting on the cytoskeleton through phosphorylation of myosin II, a key molecular step in cytoskeleton control. However, the introduction of the microRNAs into RPE1 cells induced different phenotypic outcomes. Strikingly, the ectopic expression of miR-661 strongly modified the spatial phosphorylation of myosin II, while in contrast, overexpression of either miR-612 or miR-940 inhibited myosin II phosphorylation (Figure 3). This antagonistic phenotypic outcome was further confirmed by invasion experiments (Figure 4). Together, these experimental validations confirmed the involvement of the assorted club 2 in the regulation of small GTPase signalling, the actin cytoskeleton, cell motility and cell invasion. Because the three microRNAs lead to different phenotypic effects, it might be not surprising that these microRNAs are not expressed at the same time in a tissue. Confirming this statement, the three miRNAs were not found co-expressed in breast, prostate, colorectal mucosa, lung, blood, prefrontal cortex, liver and muscle normal tissues. Therefore, it would have been difficult to infer this network relying on the expression level of microRNAs. These data further emphasise the relevance of the target-based microRNA networks that we have inferred.

To our knowledge, miR-940 had never been reported as differentially express in breast cancer. This absence might be explained by the fact that miR-940 is never one of the most differentially expressed microRNA in these datasets even though its expression trend is consistent. Here, we demonstrated for the first time its capacity to modulate cell cytoskeleton and reduce RPE1 cell migration and invasion and showed that its expression is reduced in breast cancer (Figure 5). Also in agreement with our results, it was recently shown that miR-612 exerts an inhibitory effect on hepatocellular carcinoma, proliferation, migration, invasion, and metastasis. Moreover, miR-612 appears to be involved in both the initial and final steps of the metastatic cascade by suppressing local invasion and distant colonization[51]. Similarly, our results appear to be in agreement with reports from Vetter *et al.*, who have shown that miR-661 contributes to breast cancer cell invasion through the targeting of Nectin-1 and StarD10[52]. Furthermore, we demonstrated here that miR-612 and miR-661 also regulate cell motility via opposite effects on myosin II phosphorylation.

In the near future, we will further investigate the mechanism of action of miR-661 with regard to the p53 status of the cells, as it was recently reported that miR-661 may either suppress or promote cancer aggressiveness, depending on the p53 status[53].

Methods

Target prediction datasets. The flat file of DIANA-microT version 3.0 (July 2009)[13] was downloaded from the web site http://diana.cslab.ece.ntua.gr/microT/. The database consists of genome-wide computationally predicted associations between microRNAs and their predicted targets in ensemble id format. Only *Homo sapiens* gene-microRNA information was considered. Scores and multiple binding sites were not taken into account (miTG score > 0), so that the lowest possible level of false negative prediction was considered. Under this restriction, there are 555 microRNAs that are predicted to regulate 18,986 different genes. The DIANA-microT dataset comprised our main dataset for network building.

To compare the network outcome with that of another prediction algorithm, TargetScan v6.2 non-conserved[44] was downloaded from the web site http://www.targetscan.org/vert_61/ (June 2012) (Date of access: 15/06/2013). Again, only *Homo sapiens* gene–microRNA information were considered, which led to 1,531 microRNAs regulating 18,366 genes. No other criterion of selection was used.

Network building. Based on the idea that microRNAs that share the same targets might act in the same pathways, a microRNA undirected weighted graph $G = (N, E)$ was built, where each node N represents a microRNA and each edge E represents the percentage of shared targets. To define the percentage of shared targets, the meet/min index (or Simpson index) was used. The metric is highlighted in ref. 16 and can be simply defined as:

$$meet/\min_{(A,B)} = \frac{A \cap B}{min(\#A, \#B)} \quad (1)$$

where A and B represent the sets of genes regulated by microRNA A and microRNA B, respectively, and # stands for the number of targets regulated by the corresponding microRNA. The network built upon this metric is a densely connected network with 555 nodes and 153,735 weighted edges.

To analyse this dense weighted network, it was further simplified into a binary network by defining a meet/min threshold under which edges between microRNAs in the network were deleted. To do so, different graph properties were calculated and compared for different meet/min thresholds. We calculated the clustering coefficient, two different centrality measurements (degree and betweenness centrality), the density, and the assortativity coefficient[54]. We defined the meet/min threshold as 0.5. The final network comprised 555 nodes and 2,911 edges.

These microRNA graphs were also compared to 3 different graphs. The first graph was a random graph based on the Erdős-Rényi random graph construction algorithm[23], where each edge has the same probability of appearance during construction. We considered for this construction 551 nodes and 2,911 edges, as for the meet/min 0.5 network. The second graph was a scale-free graph based on the Barabási-Albert model[24] considering 551 nodes; at the end of the construction, the network will have scale-free properties as defined by Barabási and Albert. The final graph is a real network of a human protein-protein interaction map based on yeast two-hybrid[25].

The graphs were built using the package igraph[55] from the R statistical environment[56] and visualised using Cytoscape with the unweighted spring embedded layout[57].

Assorted club deciphering. First, the hubs of the network were sorted out according to the degree of each node. We then looked at the induced subgraph formed by the i^{th} first sorted hubs (beginning with the first two hubs). Ultimately, 11 hubs were considered, defining two "assorted clubs"[27]: the first comprising 8 microRNAs, and the second comprising 3 microRNAs. The 12^{th} hub was isolated from the 11 first hubs.

Ontology enrichment calculation for microRNA clusters. The ontology enrichments were calculated using the R package topGO[30]. All levels of gene ontology (GO) were considered: biological process (BP), molecular function (MF) and cellular component (CC). For BP terms, only those annotated for less than 5,000 genes and more than 10 genes were considered. To calculate the enrichment, Ensembl ids were used with DIANA-microT v3 predictions, and Entrez ids were used with TargetScan v6. All genes predicted by the respective algorithms were used as background genes for the two enrichment analyses. The ontology enrichment in terms of biological processes, molecular function and cellular component categories was built on genes that were shared by at least 50% of the group of microRNAs (*e.g.*, for a cluster of 3 microRNAs, a gene must be predicted to be regulated by at least 2 microRNAs to be considered). For the spheres of influence, a lower threshold of 25% was also used to reduce the restriction on the shared genes.

A classic Fisher's exact test was considered for the enrichment analysis. *P*-values were adjusted using the Benjamini and Hochberg correction[29], and all enriched annotations obtained for each level of GO were used in the multiple testing step (4,109 in BP, 1,277 in CC, and 3,732 in MF with DIANA-microT background genes). The corrected *P*-values were considered significant when < 0.05, although we did not consider this criteria as a strict decision-making threshold. The reader can refer to other publications for a discussion[58,59]. Briefly, multiple testing procedures generally suppose that all tested hypotheses are independent. In the case of GO enrichment, the structure of the tree involves dependences among the different annotations, which

make the BH correction too stringent in this case. We thus also considered the annotations with less significant corrected *P*-values.

Cell culture and transfection. Human telomerase-immortalised Retinal Pigmented Epithelial cells (hTERT-RPE1 or RPE1, Takara Bio Inc., Kyoto, Japan) were grown at 37°C and 5% CO_2 in DMEM/F12 (Invitrogen, Carlsbad, California) supplemented with 10% fetal calf serum, 2 mM glutamine, 100 U/ml penicillin and 100 U/ml streptomycin. Reverse transfection was performed using lipofectamine RNAiMax (Invitrogen, Carlsbad, California) for 48 h. The final concentration of microRNA mimic and siRNA was 20 nM. The microRNA mimics (miR-612, miR-940 and miR-661) were purchased from Thermo Scientific (Waltham, Massachusetts) Dharmacon (miRIDIAN). The siRNA sequence AllStars scramble (Qiagen, Venio, The Netherlands) was used as a negative control, and ROCK inhibitor Y27632 (refY0503: Sigma-Aldrich, St. Louis, Missouri) was used as a positive control. Y27632 was added at 10 μM for the last 24 h.

Cell lysis, protein extraction, and Western blotting. Protein lysates were prepared in ice-cold RIPA (Thermo Scientific), supplemented with protease cocktail inhibitor (complete mini; Roche, Basel, Switzerland), 1 mM PMSF, 2 mM Na_3VO_4, and glycerophosphate. Homogenates were cleared by centrifugation at 15,000 g for 15 min at 4°C. A total of 10 μg of proteins were run on SDS-polyacrylamide gels and blotted onto nitrocellulose membranes. Blots were blocked in 3% BSA (in TBST) for 1 h and then incubated with the primary polyclonal rabbit antibody against phosphomyosin light chain 2 (1:1000; Cell Signaling Technology, Danvers, Massachusetts) in 3% BSA overnight at 4°C. Visualisation was performed using a horseradish peroxidase-conjugated antibody (anti-rabbit; Santa Cruz Biotechnology, Dallas, Texas). For the loading control, a GAPDH rabbit polyclonal antibody was used.

Micropatterning. Glass coverslips were first spin coated at 3,000 rpm for 30 s with an adhesion promoter (TI Prime; MicroChemicals, Madhya Pradesh, India) and then with 0.5% polystyrene dissolved in toluene. The polystyrene layer was further oxidised with an oxygen plasma treatment (FEMTO; Diener Electronics, Germany) for 10 s at 30 W and incubated with 0.1 mg/ml polylysine polyethylene-glycol (JenKem Technology, Beijing, China) in 10 mM Hepes, pH 7.4, at room temperature for 1 h. Coverslips were then dried by spontaneous dewetting. Polyethylene-glycol–coated slides were placed in contact with an optical mask holding the transparent micropatterns (Toppan Photomasks, Round Rock, Texas) using a home-made vacuum chamber and exposed for 5 min to deep UV light (UVO Cleaner; Jelight Company, Irvine, California). Micropatterned slides were washed once in PBS and finally incubated for 30 min with a solution of 50 μg/ml bovine fibronectine solution (Sigma-Aldrich, St. Louis, Missouri) and 5 μg/ml Alexa Fluor 646– or Alexa Fluor 542–labelled fibrinogen (Invitrogen, Carlsbad, California). Before plating cells, patterned coverslips were washed three times with sterilised PBS. The small circular fibronectine micropattern has a size of 500 μm^2, and the large circular fibronectine micropattern is 1000 μm^2.

Immunostaining. RPE1 cells were transfected with AllStars siRNA or mimics of miR-612, miR-661 or miR-940 at 20 nM for 48 h. The positive control (Y27632) was added to the siRNA AllStars-transfected cells at 10 μM for the last 24 h. The transfected cells were then plated on the micropatterns.

After cell adhesion onto the micropatterns (2 h), RPE1 cells were pre-permeabilized for 15 seconds with 0.1% Triton X-100 in cytoskeleton buffer pH 6.1 and fixed in 4% paraformaldehyde in cytoskeleton buffer for 15 min at room temperature. They were then rinsed twice with PBS and incubated in 0.1 M ammonium chloride in PBS for 10 min. Cells were then blocked with 3% BSA in $PBS^{Ca2+\ Mg2+}$ for 30 min.

Images were taken with an upright microscope (BX61; Olympus, Tokyo, Japan). Fluorescent images of myosin and actin staining are maximal projections of different aligned cells acquired with oil immersion objectives at 100× (NA = 1.4) mounted on a piezo ceramic (Physics Instruments, Lederhosen, Germany). Five individual and typically representative images can be found in Supplementary Figure S13. The microscope was controlled with Metamorph software (MDS Analytical Technologies, Toronto, Canada). The images were processed using ImageJ software[60]. To measure the intensity of myosin and actin fluorescence, individual cells were first segmented, and the integrated density of fluorescence was calculated on the segmented cells for both phosphomyosin and actin. The segmentation was performed as follows: a median filter with a radius of 15 pixels was applied on the FITC channel images (phalloidin-stained images), followed by an auto-thresholding using the "Li" method.

For the unrestricted cell images, the cells were transfected in an 8 wells with mimics of the siRNA as previously described. Vinculin and phalloidine staining were used and images were taken at the AxioImager® (ZEISS, Oberkochen).

Transwell assay. RPE1 cells were seeded in six-well microtiter plates, cultured for one day, and then transfected with mimics of miR-612, miR-661, or miR-940 or with a negative control siRNA (siRNA AllStars) at a final concentration of 20 nM. Twenty-four hours after transfection, cells were trypsinized, re-suspended in culture medium without fetal bovine serum (0% FBS) and counted using a Scepter™ 2.0 (Merck Millipore, Billerica, Massachusetts). A similar cell number was then dispensed on the Transwell® inserts (Corning, polycarbonate membrane, 5.0 μm pore size), and cell migration was induced by the presence of complete culture medium containing serum (10% FBS) in the lower compartments. After 18 hours, cell migration was stopped, and cells were washed ($PBS^{Ca2+\ Mg2+}$), fixed (PFA 4%) and permeabilized (100% methanol). For cell counting, cell nuclei were stained with Hoechst 33342, and cells from the deposition side (upper compartment) were removed with a cotton swab. Images of the lower side of the insert were captured (10 different fields of view per insert) with an epifluorescence microscope (Imager Z1 from Zeiss, Oberkochen, Germany) using AxioVision software with a 10-fold objective (Plan-Neofluar 10x/0.30). The quantification of the number of cells that reached the lower side of the insert was performed either semi-automatically using ImageJ software[60] or manually.

Four independent experiments with varying initial numbers of deposited cells (to better emphasise either the increase or decrease in the number of cells passing through the holes) were conducted. Each condition (miR-612, miR-661, miR-940, and siRNA-AllStars) was represented in triplicate, leading to 30 observations per experiment and per condition (Supplementary Figure S14). To pool the four experiments together, each condition of each experiment was normalised to the median number of cells in the siRNA-AllStars condition ($median_{siRNA-AllStars_x}$), considering all replicates:

$$\tilde{N}_{x,y,\tau} = log10\left(\frac{N_{x,r,\tau}}{median_{siRNA-AllStars_x}}\right) \quad (2)$$

where N is the number of counted cells and $\tilde{N}$ the normalised number of cells for each condition, replicate and experiment; x is the experiment (1, 2, 3 or 4), r, the replicate for each experiment (1, 2 or 3), and, the different conditions (siRNA-AllStars or mimics of miR-612, miR-661 or miR-940). $\tilde{N}$ is thus normalised at 0 for siRNA-AllStars. *P*-values were then calculated using the non-parametric two-sided Mann-Whitney test (Wilcoxon test) available in R statistical software.

Wound healing assay. RPE1 cells were deposited into a transparent-bottomed 96-well microtiter plate with DMEM medium in presence of 10% FBS and without any antibiotics. After 24 h, the RPE1 cells were transfected either with siAllStars or the 3 microRNAs at 20 nM in triplicates. 48 hours after transfection, 500 μm-large wounds with low width variability were produced in every confluent cell cultures using a wound replicator equipped with 96 pins (V&P Scientific, San Diego, California). The wounds were immediately imaged as well as 5 h, 7.5 h and 10 h after wound formation. No image was taken 2.5 h after wound formation in order to let the cells recover from mechanical and thermal stress resulting from the process of wound formation.

At every time point, the images of the wounds were acquired at 8 exposure times using parallelised holographic microscopy, *i.e.*, an array of 96 Complementary Metal Oxide Semiconductor (CMOS) image sensors (STMicroelectronics, Grenoble, France) placed under the 96-well microtiter plate[61]. Holographic microscopy relies on the digital recording of the diffraction patterns (*i.e.*, holograms) made by the cells under coherent illumination. The 8 images taken at various exposure times for every time point were combined to produce a contrasted, non-saturated image using a High Dynamic Range (HDR) approach. The edges of the wound were automatically detected on the HDR images by a k-means/Markov random field process and a parallel double snake[62]. Results of automated wound segmentations were validated by eye. Evolution of the average width of the wounds was finally normalised in relation to the initial wound width and pooled for every transfection condition.

Expression data. To assess for microRNA expression in different tissues, samples from 8 different normal tissues were downloaded from GEO. The 8 downloaded datasets comprised GSE19505 (Prefrontal cortex and liver), GSE23527 (Skeletal muscle), GSE24205 (Blood), GSE25508 (Lung), GSE31309 (Breast), GSE34933 (Prostate), and GSE38309 (Colorectal mucosa). The raw data were downloaded and extracted from GEO. Datasets showing negative numbers were adjusted by adding a constant such that the minimum observed value was equal to 1 and then log2 transformed. Finally, a quantile normalization was applied[68]. Cytoscape was used with the DIANA-microT network to visualize microRNA expression levels for each dataset separately. Represented on the networks is the median value of each microRNA expression in control tissues. Min and max observed expression on each dataset were used to define the legend limits with a continuous gradation from white to red. The white limit was set as the median expression on each dataset so that not all nodes would be highlighted. As a consequence, only microRNAs with expression 50% over all miRNA expression(s) in a dataset (moderately to highly expressed on a chip) are coloured. Differences in colour spreading thus reflect the difference in expression score distribution across datasets (Supplementary Figure S15).

For the differential expression analysis, the raw data of the three breast cancer microRNA expression datasets were retrieved on Gene Expression Omnibus (GEO)[63] under the ID GSE31309, GSE38867 and GSE44124. Only datasets with at least 20 samples were considered. As such, GSE31309 comprises 105 samples with 57 healthy controls. GSE38867 is composed of 28 samples with 7 normal tissues, and finally GSE44124 is built of 53 samples with 3 pools of normal tissue. All data were log2 transformed and quantile normalised[64] using either affyPLM package[65] or AgiMicroRna package[66] in R, depending on the data type. After data normalisation, the package Limma[67] was used to calculate *P*-values for differential expression of miR-940 in breast cancer tissues against normal tissues. As we investigated only one microRNA per dataset, no correction for multiple testing was applied.

All calculations in this paper were conducted in the statistical environment R[56]. Box-and-whisker plots show the lower and upper quartiles (25–75%) with a line at the

median. Whiskers extend to 1.5 times the interquartile range (defined as quartile 75% – quartile 25%). The circles show data outside the whiskers ("outliers").

1. Kloosterman, W. P. & Plasterk, R. H. A. The Diverse Functions of MicroRNAs in Animal Development and Disease. *Dev. Cell* **11**, 441–450 (2006).
2. Alvarez-Garcia, I. & Miska, E. A. MicroRNA functions in animal development and human disease. *Development* **132**, 4653–4662 (2005).
3. Yi, R. & Fuchs, E. MicroRNAs and their roles in mammalian stem cells. *J. Cell Sci.* **124**, 1775–83 (2011).
4. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854 (1993).
5. Van Rooij, E. & Olson, E. N. MicroRNA therapeutics for cardiovascular disease: opportunities and obstacles. *Nat. Rev. Drug Discov.* **11**, 860–872 (2012).
6. Xu, J. *et al.* MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features. *Nucleic Acids Res.* **39**, 825–836 (2011).
7. Hu, S. *et al.* Novel microRNA prosurvival cocktail for improving engraftment and function of cardiac progenitor cell transplantation. *Circulation* **124**, S27–34 (2011).
8. Zhu, W. *et al.* Dissection of Protein Interactomics Highlights MicroRNA Synergy. *PLoS One* **8**, e63342 (2013).
9. Shalgi, R., Lieber, D., Oren, M. & Pilpel, Y. Global and Local Architecture of the Mammalian microRNA–Transcription Factor Regulatory Network. *PLoS Comput Biol* **3**, e131 (2007).
10. Jiang, Q. *et al.* Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* **4 Suppl 1**, S2 (2010).
11. An, J., Choi, K. P., Wells, C. A. & Chen, Y.-P. P. Identifying co-regulating microRNA groups. *J. Bioinform. Comput. Biol.* **8**, 99–115 (2010).
12. Vlachos, I. S. *et al.* DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. *Nucleic Acids Res.* **40**, W498–504 (2012).
13. Maragkakis, M. *et al.* Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics* **10**, 295 (2009).
14. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell* **120**, 15–20 (2005).
15. Gómez, J., Martínez-A, C., González, A. & Rebollo, A. Dual role of Ras and Rho proteins: At the cutting edge of life and death. *Immunol. Cell Biol.* **76**, 125–134 (1998).
16. Goldberg, D. S. & Roth, F. P. Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci.* **100**, 4372–4376 (2003).
17. Fuxman Bass, J. I. *et al.* Using networks to measure similarity between genes: association index selection. *Nat. Methods* **10**, 1169–76 (2013).
18. Liu, G., Wong, L. & Chua, H. N. Complex discovery from weighted PPI networks. *Bioinformatics* **25**, 1891–1897 (2009).
19. Van Wijk, B. C. M., Stam, C. J. & Daffertshofer, A. Comparing Brain Networks of Different Size and Connectivity Density Using Graph Theory. *PLoS One* **5**, e13701 (2010).
20. Langer, N., Pedroni, A. & Jäncke, L. The Problem of Thresholding in Small-World Network Analysis. *PLoS One* **8**, e53199 (2013).
21. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–13 (2004).
22. Freeman, L. C. Centrality in social networks conceptual clarification. *Soc. Networks* **1**, 215–239 (1978).
23. Erdős, P. & Rényi, A. On random graphs. *Publ. Math. Debrecen* **6**, 290–297 (1959).
24. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286**, 509–512 (1999).
25. Bu, D. *et al.* Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Res.* **31**, 2443–2450 (2003).
26. Albert, R. Scale-free networks in cell biology. *J. Cell Sci.* **118**, 4947–57 (2005).
27. Colizza, V., Flammini, A., Serrano, M. A. & Vespignani, A. Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110–115 (2006).
28. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
29. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
30. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
31. Kim, S. Y., Lee, Y.-H. & Bae, Y.-S. MiR-186, miR-216b, miR-337-3p, and miR-760 cooperatively induce cellular senescence by targeting α subunit of protein kinase CKII in human colorectal cancer cells. *Biochem. Biophys. Res. Commun.* **429**, 173–179 (2012).
32. Nidadavolu, L. S., Niedernhofer, L. J. & Khan, S. A. Identification of microRNAs dysregulated in cellular senescence driven by endogenous genotoxic stress. *Aging* **5**, 460–473 (2013).
33. Haga, C. L. & Phinney, D. G. MicroRNAs in the Imprinted DLK1-DIO3 Region Repress the Epithelial-to-Mesenchymal Transition by Targeting the TWIST1 Protein Signaling Network. *J. Biol. Chem.* **287**, 42695–42707 (2012).
34. Simion, A. *et al.* MiR-495 and miR-218 regulate the expression of the Onecut transcription factors HNF-6 and OC-2. *Biochem. Biophys. Res. Commun.* **391**, 293–298 (2010).
35. Chen, S.-M. *et al.* MicroRNA-495 inhibits proliferation of glioblastoma multiforme cells by downregulating cyclin-dependent kinase 6. *World J. Surg. Oncol.* **11**, 87 (2013).
36. Prévot, P.-P. *et al.* Let-7b and miR-495 Stimulate Differentiation and Prevent Metaplasia of Pancreatic Acinar Cells by Repressing HNF6. *Gastroenterology* **145**, 668–678.e3 (2013).
37. Jiang, X. *et al.* miR-495 is a tumor-suppressor microRNA down-regulated in MLL-rearranged leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19397–19402 (2012).
38. Zhou, L., Qi, X., Potashkin, J. A., Abdul-Karim, F. W. & Gorodeski, G. I. MicroRNAs miR-186 and miR-150 Down-regulate Expression of the Pro-apoptotic Purinergic P2X7 Receptor by Activation of Instability Sites at the 3′-Untranslated Region of the Gene That Decrease Steady-state Levels of the Transcript. *J. Biol. Chem.* **283**, 28274–28286 (2008).
39. Villa, C. *et al.* Role of *hnRNP-A1* and miR-590-3p in Neuronal Death: Genetics and Expression Analysis in Patients with Alzheimer Disease and Frontotemporal Lobar Degeneration. *Rejuvenation Res.* **14**, 275–281 (2011).
40. Gong, A.-Y. *et al.* MicroRNA-513 regulates B7-H1 translation and is involved in IFN-gamma-induced B7-H1 expression in cholangiocytes. *J. Immunol.* **182**, 1325–1333 (2009).
41. Srikantan, S. *et al.* Translational Control of TOP2A Influences Doxorubicin Efficacy. *Mol. Cell. Biol.* **31**, 3790–3801 (2011).
42. Somlyo, A. P. & Somlyo, A. V. Signal transduction by G-proteins, Rho-kinase and protein phosphatase to smooth muscle and non-muscle myosin II. *J. Physiol.* **522**, 177–185 (2000).
43. Pitaval, A., Christ, A., Curtet, A., Tseng, Q. & Théry, M. Probing ciliogenesis using micropatterned substrates. *Methods Enzymol.* **525**, 109–130 (2013).
44. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell* **120**, 15–20 (2005).
45. Wang, X. & El Naqa, I. M. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* **24**, 325–32 (2008).
46. Tsang, J. S., Ebert, M. S. & van Oudenaarden, A. Genome-wide Dissection of MicroRNA Functions and Cotargeting Networks Using Gene Set Signatures. *Mol. Cell* **38**, 140–153 (2010).
47. Maragkakis, M. *et al.* DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res.* **37**, W273–W276 (2009).
48. Alshalalfa, M., D Bader, G., Bismar, T. A. & Alhajj, R. Coordinate MicroRNA-Mediated Regulation of Protein Complexes in Prostate Cancer. *PLoS One* **8**, e84261 (2013).
49. Hsu, S.-D. *et al.* miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **39**, D163–9 (2011).
50. Xiao, F. *et al.* miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.* **37**, D105–10 (2009).
51. Tao, Z.-H. *et al.* miR-612 suppresses the invasive-metastatic cascade in hepatocellular carcinoma. *J. Exp. Med.* **210**, 789–803 (2013).
52. Vetter, G. *et al.* miR-661 expression in SNAI1-induced epithelial to mesenchymal transition contributes to breast cancer cell invasion by targeting Nectin-1 and StarD10 messengers. *Oncogene* **29**, 4436–4448 (2010).
53. Hoffman, Y., Bublik, D. R., Pilpel, Y. & Oren, M. miR-661 downregulates both Mdm2 and Mdm4 to activate p53. *Cell Death Differ.* **21**, 302–9 (2014).
54. Newman, M. Mixing patterns in networks. *Phys. Rev. E* **67**, 026126 (2003).
55. Csardi, G. & Nepusz, T. The igraph Software Package for Complex Network Research. *InterJournal Complex Syst* **1695**, 1695 (2006).
56. R Core Team. *R: A Language and Environment for Statistical Computing.* (2012). Available at: http://www.r-project.org/ (Accessed: 26th October 2012).
57. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
58. Blüthgen, N. *et al.* Biological profiling of gene groups utilizing Gene Ontology. *Genome Inform.* **16**, 106–115 (2005).
59. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
60. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
61. Ghorbel, I., Bertacchi, N., Gidrol, X. & Haguet, V. Parallelized contact imaging and automated analysis of cell migration dynamics. Paper presented at the *37th Int. Meet. Ger. Soc. Cell Biol.* 71. Regensburg, Germany (2014, March 18–21).
62. Ghorbel, I., Rossant, F., Bloch, I. & Paques, M. Modeling a parallelism constraint in active contours. Application to the segmentation of eye vessels and retinal layers. Paper presented at the *18th IEEE Int. Conf. Image Process.* 445–448. Brussels, Belgium (2011, Sept. 11–14).
63. Edgar, R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
64. Bolstad, B. M., Irizarry, R., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
65. Bolstad, B. M. *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization.* 274 (2004). Available at: http://

bmbolstad.com/Dissertation/Bolstad_2004_Dissertation.pdf (Accessed: 17th June 2014).
66. López-Romero, P., González, M. A., Callejas, S., Dopazo, A. & Irizarry, R. A. Processing of Agilent microRNA array data. *BMC Res. Notes* **3**, 18 (2010).
67. Smyth, G. K. [limma: Linear Models for Microarray Data] *Bioinforma. Comput. Biol. Solut. Using R Bioconductor* [Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. & Huber, W.] [397–420] (Springer, 2005).
68. Pradervand, S. *et al.* Impact of normalization on miRNA microarray expression profiling. *RNA* **15**, 493–501 (2009).

Acknowledgments

This work was funded by grants from the Agence Nationale de la Recherche and from CEA.

Author contributions

L.G. and X.G. conceived and designed the study. R.B., L.G., C.L. built the networks and performed the statistical analysis. A.P. performed the western blotting and immunofluorescence experiments. E.S., S.C. performed the transwell assays. P.O., V.H. and I.G. performed and analysed the wound healing assays. R.B., L.G., X.G. analysed the results. R.B., L.G., X.G. wrote the manuscript with input from all authors.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Bhajun, R. *et al.* A statistically inferred microRNA network identifies breast cancer target miR-940 as an actin cytoskeleton regulator. *Sci. Rep.* **5**, 8336; DOI:10.1038/srep08336 (2015).

Biotechs et pharmas, même galère

Après des décennies de faste, l'industrie pharmaceutique essuie, elle aussi, une crise. Davantage encline au rachat qu'à l'innovation, elle doit désormais revoir les modèles économiques qui ont fait sa grandeur. Dans son sillage, le secteur biotech, longtemps considéré comme la bouée de sauvetage de son aîné historique, subit la même lame de fond(s). Suivra-t-il la même voie pour s'en sortir ? Décryptage.

Avec sa R&D faiblement productive et la perte de brevets sur de nombreux médicaments, l'industrie pharmaceutique fait, depuis quelques années, face à un de ses plus grands défis : maximiser la rentabilité des portefeuilles de produits afin de maintenir ses résultats et d'accéder à des produits innovants grâce à des partenariats, notamment avec les sociétés de biotechnologies. Dans l'environnement actuel, la mise en place de stratégies de management du cycle de vie des produits (« *Product Lifecycle Management* », PLM) ou de stratégies basées sur la fusion et l'acquisition de sociétés, de produits ou de licences, apparaît comme une priorité stratégique pour de nombreuses sociétés pharmaceutiques (1).

Depuis une trentaine d'années, les sociétés de biotechnologies ont réussi à bouleverser, jusqu'à le remettre en cause, le modèle de découverte de médicaments grâce à un éventail d'expertises et d'outils novateurs. Avec une approche plus rationnelle de la conception des médicaments, l'industrie biotech a, en effet, obtenu des résultats concrets et probants en matière de bénéfices commerciaux. C'est le cas, par exemple, de BiogenIdec qui commercialise plusieurs produits d'immunothérapie contre la sclérose en plaques, ou encore de la société biopharmaceutique californienne Amylin dont les développements contre le diabète (tel Byetta®) ont incité le géant américain Bristol-Myers Squibb a l'acquérir (2).

Or à un moment où la demande de l'industrie pharmaceutique pour des technologies et des produits innovants est plus élevée que jamais, le secteur biotechnologique éprouve beaucoup de difficulté à commercialiser ses produits malgré la qualité et l'efficacité de ses portefeuilles de candidats-médicaments et/ou de ses plateformes technologiques. Et pour cause, 30 ans après le début de l'ère biotechnologique et avec comme objectif d'assurer sa pérennité mais également de maximiser son retour sur investissement, le secteur doit faire face aux mêmes enjeux que son aîné le secteur pharmaceutique : une productivité de sa R&D relativement faible – seulement deux produits biologiques ont été approuvés en 2013 contre six en 2011 et en 2012 (3) – au regard de coûts toujours élevés, la perte de brevets sur les médicaments « *bioblockbusters* », un recours aux activités de fusion/acquisition ainsi que la mise en place de stratégies de management du cycle de vie des produits.

les auteurs
Arsia Amir-Aslani, Ricky Bhajun et Nicolas Sainte-Foie
Grenoble École de Management, Grenoble

Que ce soit en Europe ou aux États-Unis, peu d'entreprises de biotechnologie présentent un chiffre d'affaire supérieur à 500 M$. Parmi elles, le « Club des cinq » formé par Amgen, Biogen Idec, Celgen, Gilead Sciences et Regeneron Pharmaceuticals génère à lui-seul 75 % du chiffre d'affaires total du secteur des biotechnologies aux États-Unis. D'après (2).

APRÈS LES *BLOCKBUSTERS*, LES *BIOBLOCKBUSTERS*

L'aspect du *blockbuster* a évolué depuis les années 2000. Principalement basés sur de nouvelles entités chimiques (NEC), et généralement développés autour d'indications prescrites par les médecins généralistes, les premiers médicaments du genre pénétraient le marché par leur utilisation sur une population large et pour traiter de manière plus ou moins sélective une pathologie souvent chronique. Les nouveaux *blockbusters* sont de plus en plus issus de nouvelles entités biologiques (NEB). En 2011, sur les 20 médicaments les plus vendus, 8 étaient ainsi issus des biotechnologies. Ces biomédicaments sont utilisés sur des populations plus restreintes, pour traiter des pathologies plus rares (4).

Les sociétés biotech constituent la nouvelle ligne de développement du modèle des *blockbusters*. Cependant, une fois sur le marché, et malgré leur succès croissant, ces produits, considérés comme plus sélectifs et responsables de moins d'effets indésirables, provoquent tout de même de tels effets, à l'instar de certains médicaments d'origine chimique. De plus, si la recherche de nouveaux médicaments a débouché sur des produits biopharmaceutiques importants, elle n'a pas su réduire le risque ni améliorer la productivité des programmes de R&D du secteur, et ce malgré le vif succès que connaît ce dernier – en plus des 8 médicaments dans le top 20 des ventes, presque 50 % des autorisations de mise sur le marché sont issues des biotechnologies.

LE DÉFI DE LA CRÉATION DE VALEUR

Ce bilan reste relativement mitigé, s'il est placé en perspective de l'ensemble des programmes de R&D menés par les entreprises biotechnologiques. À titre d'exemple, sur la période comprise entre janvier 2006 et décembre 2007, seuls 63 produits issus du secteur biotech – 47 issus des biotechnologies et 16 issus de collaborations biotech-pharma – ont obtenu l'approbation de la Food & Drug Administration (FDA) pour une mise sur le marché, contre 86 échecs cliniques en phase III – 68 issus des biotechnologies et 18 issus des collaborations. Sur les 47 produits 100 % biotech approuvés par l'agence sanitaire américaine, seuls 9 étaient réellement innovants. Les autres s'inspiraient de mécanismes d'action validés par les autorités ou s'inscrivaient dans une recherche d'optimisation du rendement d'un produit via, par exemple, l'extension de son indication. De plus, pour chaque produit innovant approuvé, on constatait en moyenne cinq échecs cliniques en phase III pour des développements équivalents en termes d'innovation. Par ailleurs, sur les 24 produits qui n'ont pas obtenu d'autorisation, 8 appartenaient à la catégorie des produits innovants issus des biotechnologies (5).

En partant du constat qu'un *blockbuster* peut assurer, presque à lui seul, la profitabilité des grands laboratoires pharmaceutiques – comme l'ont démontré l'oméprazol/Prilosec (inhibiteur de la pompe à protons) pour Astra Zeneca, atorvastatine/Lipitor (hypocholestérolémiant) pour Pfizer ou encore fexofénadine/Allegra (antihistaminique) à l'époque d'Aventis –, il est aisé d'appréhender les conséquences dramatiques de cette « bioconcurrence » sur leurs revenus. Dans le secteur biotech, une part considérable des profits peut ne dépendre que de deux, voire d'un seul produit. Après l'expiration du brevet correspondant, la perte de propriété sur l'un de ces *bioblockbusters* peut induire une érosion considérable des volumes de vente au profit de génériques. Si cette érosion, variable selon les pays, reste moins importante que pour les produits d'origine chimique – le coût de développement et de production des biosimilaires[*1] étant beaucoup plus élevé –, elle peut néanmoins affecter notablement la profitabilité des firmes qui les produisent (encadré ci-dessous). Amgen en a notamment fait l'expérience au terme du brevet protégeant la synthèse d'EPO.

L'arrivée des biosimilaires

Depuis une dizaine d'années, nombre de produits issus des biotechnologies se trouvent dans la même situation que les *blockbusters* d'origine chimique : certains ont déjà perdu et d'autres vont perdre prochainement leurs brevets. Et ils doivent, eux aussi, faire face à une concurrence de la part de produits non génériques, d'origine biotechnologique. De nombreux brevets d'anticorps monoclonaux vont ainsi bientôt tomber dans le domaine public et permettre le développement de médicaments dits « biosimilaires[*1] ». Ainsi le Comité des médicaments à usage humain (Committee for Medicinal Products for Human use, CMPH) de l'Agence médicale européenne a-t-il recommandé, le 28 juin 2013, la mise sur le marché du premier produit biosimilaire, Remicade, l'anticorps monoclonal du laboratoire américain MSD destiné, entre autres, à traiter la polyarthrite rhumatoïde. Les brevets d'autres anticorps monoclonaux déjà commercialisés, tels que MabThera (leucémie lymphoïde) ou Herceptin (cancer du sein) du laboratoire suisse Roche, tomberont dans le domaine public dans les cinq prochaines années.

LA CONSOLIDATION PAR LA FUSION/ ACQUISITION

Comme pour le secteur pharmaceutique, les fusions/acquisitions se présentent comme une stratégie de recherche d'innovation pour le secteur biotech : le rapprochement entre deux sociétés est censé favoriser l'acquisition et l'appropriation de nouveaux actifs incorporels (c'est-à-dire des brevets, des technologies ou des projets R&D), mais également le regroupement de ressources financières permettant de développer de nouvelles technologies (6).

Le grand nombre d'entreprises actives dans le secteur entraîne une dispersion des ressources financières et humaines sur des projets voisins et, souvent, très peu différenciés. Lors de l'élaboration de sa stratégie, l'entreprise biotech doit identifier les facteurs de compétitivité, tels que la qualité des produits, la nature de leur technologie, l'existence de partenariats et le volume des ressources financières, et s'efforcer de mieux les maîtriser que ses concurrents. Il faut donc intégrer de nouvelles données relatives à ces facteurs tout au long de la chaîne de valeur, depuis la découverte de la molécule d'intérêt jusqu'au médicament. Cette évolution exige un remaniement fondamental du cadre organisationnel à l'intérieur de ces sociétés, ainsi que des approches et modèles commerciaux. La solution idéale étant une enseigne unique (« *one stop shop* »), où les entreprises sont capables de proposer une gamme de technologies indépendantes et complémentaires. Il est donc désormais plus important, lors d'une fusion, de mettre en commun les technologies et le savoir-faire. Un motif de plus en plus évoqué par les PME biotech pour justifier le recours à des opérations de fusion ou d'acquisition avec d'autres PME du secteur (6). Mais, pour cela, ces sociétés seront obligées de s'engager, à leur tour, dans un processus de consolidation, afin de rechercher une masse critique, celle-ci permettra de développer uniquement les médicaments pour lesquels les entreprises biotech disposent de la plus grande prévisibilité sur le comportement du produit et donc sur la réussite de son développement. En effet, le taux d'échec reste très élevé pendant la phase clinique du fait, principalement, que, dans leur immense majorité, les sociétés biotech sont fondées autour d'une seule technologie.

Selon une étude de 2014, menée par le cabinet américain d'audit financier Ernst & Young, la valeur totale des fusions et acquisitions impliquant des sociétés biotech américaines ou européennes en 2013 a ainsi atteint 47,6 milliards de dollars (Md$), soit une augmentation de 106 % par rapport à 2012. Sans compter la mégafusion Amgen-Onyx Pharmaceuticals, cette même année, les accords biotech-biotech ont augmenté en valeur de 68 % pendant la période pour atteindre 10,6 Md$ (2).

À l'image du secteur pharmaceutique, il s'agit désormais de jouer sur les économies d'échelle et de réduire le risque spécifique au développement d'une molécule ou d'une technologie particulières afin d'améliorer les conditions de financement.

CE QUE VEULENT LES MARCHÉS FINANCIERS

Les résultats de l'industrie mondiale des biotechnologies ont fortement rebondi en 2013, tirés vers le haut par la bonne performance d'entreprises telles que l'américaine Gilead, boostée par le développement et les ventes de son traitement contre l'hépatite C commercialisé sous le nom Sovaldi. Les

[*1] En France, le Code de la santé publique les définit comme des médicaments biologiques de même composition qualitative et quantitative en substance active et de même forme pharmaceutique qu'un médicament biologique de référence mais qui ne remplissent pas les conditions pour être regardés comme une spécialité générique en raison notamment de la variabilité de la matière première et des procédés de fabrication (article L5121-1-15).

entreprises cotées en bourse ont connu une croissance à deux chiffres de leurs revenus, ainsi qu'une forte augmentation des montants des levées de fonds.

Malgré les bons résultats financiers globaux, la plupart des entreprises biotech évoluent dans un environnement aux ressources limitées, contrairement aux sociétés pharmaceutiques qui échappent à cette contrainte grâce à leur forte capacité d'autofinancement (2). Néanmoins, les attentes de la part des investisseurs institutionnels en termes de performances boursières sont les mêmes. D'où une nécessité accrue de justifier une stratégie d'allocation efficace de leur capital dans leurs investissements en R&D. De plus, à l'instar du secteur pharmaceutique, dont les 20 premiers laboratoires réalisent pratiquement 80 % des ventes de médicaments, une grande partie de la croissance observée dans le secteur biotech est le fait d'un groupe relativement restreint de sociétés en phase commerciale, comprenant notamment la suisse Actelion, la néerlandaise Qiagen, la française Eurofins Scientific ou l'irlandaise Shire.

Enfin, comme dans le secteur pharmaceutique, la forte pression exercée par les investisseurs institutionnels sur le potentiel de création de valeur actionnariale des entreprises contraint également les sociétés de biotechnologies à déployer des stratégies de croissance externe.

PRODUCT LIFECYCLE MANAGEMENT (PLM)

Il y a indiscutablement beaucoup trop de sociétés biotech développant des produits très proches, qui nécessitent souvent des moyens financiers importants. Les entreprises du secteur se retrouvent dans l'obligation de mettre en place une politique de portefeuille de produits réaliste mais également de s'inscrire dans des stratégies de PLM, en vue de prolonger la durée de commercialisation de leurs produits. Or, les grandes firmes biopharmaceutiques utilisent le PLM comme un véritable outil stratégique : il leur permet d'accompagner un produit tout au long de son cycle de vie afin d'optimiser son exploitation dans un environnement en constant changement à cause de l'apparition de nouvelles concurrences, indications médicales, formulations du produit sur le marché.

Pour le secteur des biotechnologies, la gestion du cycle de vie d'un biomédicament est d'autant plus cruciale que la durée de la R&D nécessaire en amont de sa commercialisation est de plus en plus longue, laissant de fait très peu de temps à l'exploitation commerciale du produit. C'est pourquoi les stratégies de PLM sont initiées en prévision de la perte d'un brevet. Ce management s'appuie notamment sur la recherche d'extension des indications médicales d'un produit déjà commercialisé. Sur la période comprise entre janvier 2006 et décembre 2007, 103 produits avaient ainsi été approuvés par les autorités réglementaires américaines dont une vingtaine étaient des biomédicaments « *me-too*[*2] » et 18 des extensions. À l'instar de leur rôle dans le secteur pharmaceutique, les stratégies de PLM font aujourd'hui partie intégrante des préoccupations de l'industrie biopharmaceutique. Elles ont pour objectif d'influencer favorablement le retour sur investissement du portefeuille de produits en allongeant la durée de vie d'exploitation commerciale afin de préserver une part de marché conséquente face à la concurrence des produits biosimilaires.

VERS UN MODÈLE BIOTECH INTÉGRÉ

Les sociétés biotechnologiques sont confrontées à autant de défis que d'opportunités, alors que l'ensemble de leur secteur d'activité se transforme. Ces entreprises sont aujourd'hui obligées de se fixer des objectifs en fonction de la configuration de l'environnement économique, financier et réglementaire. Elles doivent ainsi tenir compte non seulement des besoins immédiats en termes d'innovation, mais également des attentes des investisseurs institutionnels et des ressources dont elles disposent.

[*2] Médicament ayant le même mécanisme d'action que le médicament d'origine mais présentant des propriétés pharmacodynamiques différentes. Par exemple, Tadalafil (Cialis) et Vardenafil (Levitra) sont des « *me-too* » du Sildenafil (Viagra).

Objectif FIPCO

Le taux de mortalité élevé des projets R&D développés par les entreprises biotech pourrait être expliqué en partie par leur tentative de maîtriser, comme des entreprises pharmaceutiques, l'ensemble de la chaîne de valeur de découverte d'un médicament, depuis la recherche jusqu'à la phase réglementaire et commerciale avec pour but de devenir une « *Fully Integrated Pharmaceutical Company* » ou FIPCO. Des entreprises telles qu'Amgen, Celgene, Gilead, Shire, etc. y sont parvenues. Ce modèle économique offre une plus grande visibilité opérationnelle et améliore donc la viabilité du projet. Néanmoins, le financement d'équipements, d'unités de production, d'essais cliniques ou encore de l'initiation de la commercialisation sont des investissements potentiellement très importants, qui peuvent mettre en péril la stabilité financière de l'entreprise. Sur les 4 000 sociétés (publiques et privées) du secteur des biotechnologies, seules une petite soixantaine sont arrivées à ce stade. Le modèle FIPCO est risqué mais il a pour mérite de garder en interne toute la valeur créée par un projet R&D. À l'inverse, les petites structures sont dépendantes des partenariats pour financer leurs projets à haut risque. Ces dernières se trouvent, par conséquent, dans l'obligation de partager le rendement financier de leurs travaux avec un partenaire industriel.

Malgré les difficultés rencontrées, la réorganisation économique du secteur vers la biopharmaceutique passe par le modèle dit « FIPCO » (*Fully Integrated Pharmaceutical Company*) (encadré ci-dessus) qui devrait, à terme, révolutionner les marchés pharmaceutiques traditionnels. Pour y parvenir, et en vue d'assurer sa pérennité, le secteur des biotechnologies tente, par le biais des activités de fusion/acquisition, de dégager le maximum de synergies entre deux entreprises grâce à une concentration de leurs efforts de recherche, notamment en tirant parti d'une complémentarité entre classes thérapeutiques de biomédicaments, pour élargir leur portefeuille de produits et de technologies. ■

(1) Sandner P, Ziegelbauer K (2008) *Drug Discov Today* 13, 457-63
(2) Ernst & Young (2014) *Beyond borders Biotechnology Industry Report*, tinyurl.com/EY-biotech2014
(3) HBM Partners (2013) *Trends in US New Drug Approvals*, www.hbmpartners.com/report
(4) Amir-Aslani A *et al.* (2013) *Spectra Biologie* 199, 24-5
(5) Czerepak E, Ryser S (2008) *Nat Rev Drug Discov* 7, 197-8
(6) Buisson R *et al.* (2014) *Biofutur* 350, 36-40