Université de Montréal

Modélisation de réseaux d'interactions des microARN et analyse et validation expérimentale de leurs boucles minimales avec des facteurs de transcription

par
Véronique Lisi

Institut de recherche en immunologie et cancer
Faculté de médecine

Thèse présentée à la Faculté de médecine
en vue de l'obtention du grade de Ph.D.
en biologie moléculaire

Décembre 2012

Résumé

Les microARN (miARN) sont de petits ARN non-codants qui répriment la traduction de leurs gènes cibles par hybridation à leur ARN messager (ARNm). L'identification de cibles biologiquement actives de miARN est cruciale afin de mieux comprendre leurs rôles. Ce problème est cependant difficile parce que leurs sites ne sont définis que par sept nucléotides. Dans cette thèse je montre qu'il est possible de modéliser certains aspects des miARN afin d'identifier leurs cibles biologiquement actives à travers deux modélisations d'un aspect des miARN. La première modélisation s'intéresse aux aspects de la régulation des miARN par l'identification de boucles de régulation entre des miARN et des facteurs de transcription (FT). Cette modélisation a permis, notamment, d'identifier plus de 700 boucles de régulation miARN/FT, conservées entre l'humain et la souris. Les résultats de cette modélisation ont permis, en particulier, d'identifier deux boucles d'auto-régulation entre LMO2 et les miARN miR-223 et miR-363. Des expériences de transplantation de cellules souches hématopoïétiques et de progéniteurs hématopoïétiques ont ensuite permis d'assigner à ces deux miARN un rôle dans la détermination du destin cellulaire hématopoïétique.

La deuxième modélisation s'intéresse directement aux interactions des miARN avec les ARNm afin de déterminer les cibles des miARN. Ces travaux ont permis la mise au point d'une méthode simple de prédiction de cibles de miARN dont les performances sont meilleures que les outils courant. Cette modélisation a aussi permis de mettre en lumière certaines conséquences insoupçonnées de l'effet des miARN, telle que la spécificité des cibles de miARN au contexte cellulaire et l'effet de saturation de certains ARNm par les miARN. Cette méthode peut également être utilisée pour identifier des ARNm dont la surexpression fait augmenter un autre ARNm par l'entremise de miARN partagés et dont les effets sur les ARNm non ciblés seraient minimaux.

Mots-clés : microARN, facteurs de transcriptions, modélisation, leucémie, miR-223, miR-363, LMO2, prédiction de cibles de miARN.

Abstract

microRNAs (miRNAs) are small non coding RNAs that repress the translation of their target genes by pairing to their messenger RNA (mRNA). The identification of miRNAs' biologically active targets is a difficult problem because their binding sites are defined by only seven nucleotides. In this thesis, I show that it is possible to model specific aspects of miRNAs to identify their biologically active targets through two modeling of each one aspect of miRNAs. The first modeling considers the miRNAs regulations through the identification of regulatory loops between miRNAs and transcription factors (TFs). Through this modeling, we identified over 700 miRNA/TF regulatory loops conserved between human and mouse. With the results of this modeling, we were able to identify, in particular, two regulatory loops between LMO2 and the miRNAs miR-223 and miR-363. Using hematopoietic stem cells and progenitor cells transplantation experiment we showed that miR-223 and miR-363 are involved in hematopoietic cell fate determination.

The second modeling focuses directly on the interaction between miARN and messenger RNA (mRNA) to determine the miRNA targets. With this work, we developed a simple method for predicting miRNA targets that outperforms the current state of the art tool. This modeling also highlighted some unsuspected consequences of miRNA effects such as the cell context specificity and the saturation of mRNA targets by miRNA. This method can also be used to identify mRNAs whose overexpression increases the expression level of another mRNA through their shared miRNA and whose global effects on other genes are minimal.

Keywords : microRNA, transcription factors, modelling, leukemia, miR-223, miR-363, LMO2, miRNA target prediction.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ADN	acide désoxyribonucléique
ALL	acute lymphoblastic leukemia
ARN	acide ribonucléique
ARNm	ARN messager
ARNnc	ARN non codant
ARNr	ARN ribosomal
ARNt	ARN de transfert
ChIP	Immonuprécipitation de la chromatine
CLP	common lymphoid progenitor
DN	cellules double négative CD4-CD8-
DP	cellules double positives CD4+CD8+
FT	facteur de transcription
HITS-CLIP	high-throughput sequencing of RNAs isolated by crosslinking immunoprecipitation
HSC	hematopoietic stem cell
kb	kilobase
KSL	cellules Kit+Sca+Lin-
LMO2	Lim Only Domain 2
miARN	micro ARN
MRE	miRNA recognition element
nt	nucléotide
PCR	polymerase chain reaction
POLII	polymérase à ARN II
pre-miARN	précurseur de miARN
pri-miARN	transcrit primaire de miARN
pSILAC	pulsed stable isotope labelling with amino acids in cell culture
RefSeq	reference sequence
région amorce	nucléotides 2 à 8 d'un miARN mature
triloop	structure de l'ARN formée de 5 nucléotides où le premier et le dernier sont appariés

UTR	untranslated region

À ma mère qui me connait mieux que moi-même

Remerciements

Entreprendre et réussir des études graduées nécessitent de la persévérance, de la discipline et une grande éthique du travail. Je tiens tout d'abord à remercier mes parents de m'avoir donné ces trois grandes qualités qui m'ont permis de mener à bien mes études. Merci aussi à vous, à ma chère sœur Geneviève et à mon amoureux Michel pour votre support inconditionnel et vos encouragements durant les moments difficiles et pour avoir partagé mes succès et mes joies. Il y a un peu de vous dans cette thèse.

Un grand merci à mon directeur de recherche, Dr François Major, pour la très grande liberté de recherche. Merci pour ton soutien et tes encouragements dans les moments difficiles et de m'avoir poussée à toujours croire en moi-même. Merci pour tes idées de fou qui ne devraient pas fonctionnées mais qui finissent généralement par être plus proches de l'intuition du génie. Dans ton laboratoire et sous ta supervision, j'ai appris énormément non seulement sur la science mais également sur ce qui fait une bonne scientifique et un bon chef de laboratoire. Je doute qu'il y ait beaucoup d'endroits où j'aurais pu apprendre autant.

Merci aussi à ma co-directrice non officielle, Dr Trang Hoang, tout d'abord pour m'avoir accueillie dans son laboratoire puis pour m'avoir guidée dans le labyrinthe qu'étaient pour moi l'hématopoïèse et les méthodes expérimentales permettant de l'étudier. Votre support et vos idées ont été très appréciés.

Merci aux membres présents et passés des laboratoires Major et Hoang pour les discussions, scientifiques ou non, l'aide et le support. Un merci particulier à Marie-Claude, d'abord mon mentor au laboratoire puis une grande amie. Avec toi, j'ai tout appris de la biologie moléculaire expérimentale et ce fut très agréable. Merci à Shanti, André et Véronique pour l'aide avec les souris, si ce n'étaient de vous, je n'y serais pas arrivé. Merci à Romain pour les discussions lorsqu'on était si peu au laboratoire, à Karine dont le projet m'a permis une étude au laboratoire de la structure d'ARN, à Paul pour ses scripts *bash* de la mort, à Marc-Frederick pour l'optimisation de code

(j'oublies pas notre pari) et à Nathanaël pour la belle collaboration. Merci à Julie R. qui m'a donné un bon coup de main au laboratoire et à Jean pour l'aide, les conseils et les discussions. Merci aussi au membre d'adoption du laboratoire Major et mon amie du tout début, Marieke, qui m'a permis de faire mes toutes premières armes au laboratoire. Sans toi, le début du doctorat n'aurait pas été aussi agréable.

Finalement, merci aux organismes subventionnaires qui m'ont octroyés une bourse me permettant de poursuivre mes études: Instituts de Recherche en Santé du Canada (IRSC), Programme de Biologie Moléculaire de l'Université de Montréal et la fondation Cole. Je désire également remercier les membres du jury qui ont lu et évalué cette thèse.

1 Introduction

1.1 Tâche à accomplir

Le dogme central de la biologie moléculaire propose que l'acide désoxyribonucléique (ADN) d'une cellule fournisse les instructions aux molécules actives, les protéines, via les ARN messagers (ARNm) définissant ainsi les ARN codant, *i.e.* ceux qui "codent" pour des protéines. L'identification récente d'ARN non codant (ARNnc), qui ne "codent" pas pour des protéines mais sont biologiquement actifs modifie notre compréhension du fonctionnement cellulaire. Les ARN ribosomaux (ARNr) et les ARN de transfert (ARNt) sont des ARNnc connus depuis longtemps. Leur rôle important est centralisé à la traduction des ARNm en protéines. À l'opposé, les ARNnc récemment identifiés sont impliqués dans la majorité des processus cellulaires. Par leur grand champ d'action, ils ajoutent un niveau de régulation supplémentaire aux mécanismes cellulaires. Parmi les ARNnc identifiés, les microARN (miARN) forment une classe de petits ARNnc qui agissent sur la régulation des niveaux d'expression protéique de leurs cibles. Les fonctions des miARN sont très variées; ils sont requis pour le développement embryonnaire normal (Lee, Feinbaum et al. 1993, Lee, Kim et al. 2004), et le cycle cellulaire (Carleton, Cleary et al. 2007), ils influencent la différenciation cellulaire (Song and Tuan 2006), sont impliqués dans l'immunité (Pedersen and David 2008) et ont été liés à diverses maladies telle que le cancer (Dalmay and Edwards 2006, Kent and Mendell 2006, Medina and Slack 2008), les maladies neurodégénératives (Enciu, Popescu et al. 2012) et les maladies cardio-vasculaires (Small and Olson 2011), pour ne nommer que celles-ci. Cette diversité de fonctions semble en opposition avec le fait que l'effet d'un miARN sur une cible spécifique est généralement plutôt faible, de l'ordre d'une diminution de 50% du niveau d'expression (Baek, Villén et al. 2008, Selbach, Schwanhäusser et al. 2008), une variation qui existe naturellement dans la population pour la majorité des transcrits et ne mène à aucune conséquence négative (Cheung, Conlin et al. 2003).

La recherche présentée ici vient d'un besoin d'améliorer notre compréhension du rôle des miARN dans la régulation de l'expression des gènes. Pour ce faire, il est important de caractériser les cibles réelles et directes des miARN. Les évidences expérimentales des dernières années ont permis de brosser un portrait grossier des fonctions de certains miARN, entre autres les plus fortement exprimés, dans une variété de contextes. Cependant, l'identification de nouveaux miARN est beaucoup plus rapide que la validation de leurs fonctions. Cet état de fait est notamment dû au grand nombre de cibles potentielles pour chaque miARN et à l'inexistence de techniques à grande échelle permettant de vérifier rapidement toutes les cibles directes d'un miARN. Il est donc primordial de trouver des approches permettant d'identifier avec grande confiance les cibles potentielles d'un miARN. Ceci, d'une part afin de les valider plus facilement et d'autre part afin de permettre ainsi d'annoter plus rapidement les fonctions des miARN.

1.2 Thèse

1.2.1 La thèse

La modélisation de certains aspects de la régulation par des miARN basée sur des données expérimentales permet non seulement d'identifier des cibles potentielles mais surtout de prédire les cibles effectives de miARN. Deux aspects spécifiques des miARN sont modélisés pour démontrer la thèse, chacun s'intéressant à une facette différente des miARN.

Les miARN sont des régulateurs de l'expression des gènes dont le niveau d'expression est lui-même régulé par d'autres facteurs. Plusieurs études ont montré que la plupart des miARN exhibe un patron d'expression spécifique à un contexte cellulaire (cf. section 2.2). Cette spécificité d'expression pourrait être déterminée par les cibles que les miARN régulent. Si tel est le cas, les cibles fonctionnelles d'un miARN pourraient être déterminées par l'identification de leurs co-régulations avec des protéines de régulation de l'expression telles que les FT. Ceci pourrait permettre,

d'une part, d'identifier les cibles biologiquement actives d'un miARN parmi un ensemble de cibles potentielles et d'autre part d'annoter un miARN avec la fonction de cette protéine.

Basée sur cette idée, la première approche de modélisation utilisée identifie un sous-ensemble des boucles de co-régulation possibles entre un miARN et une protéine régulatrice de l'expression en se concentrant sur les facteurs de transcriptions (FT) (voir Figure 1-1). Ceux-ci possèdent quatre caractéristiques qui, d'une part, simplifient l'identification de boucles de co-régulation avec les miARN et, d'autre part, permettent de démontrer l'utilité d'identifier les cibles des miARN à travers les boucles de régulation. Ces caractéristiques sont qu'ils régulent directement la transcription de leur cible, qu'ils peuvent réguler l'expression de miARN, que la description des caractéristiques de leur cible est généralement bien établie comparativement aux cibles d'autres régulateurs d'expression et que leur fonction est généralement bien documentée.

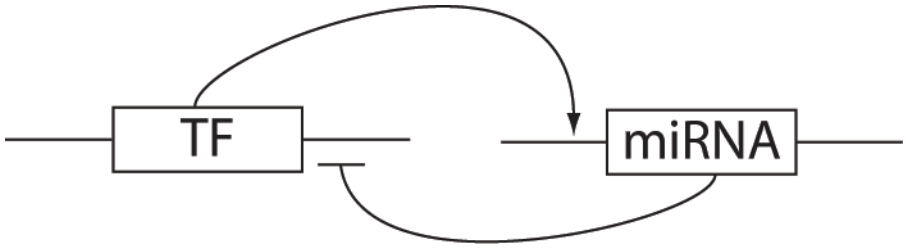


Figure 1-1 Représentation graphique de la première modélisation.
Cette modélisation consiste à identifier les cibles d'un miARN parmi les cibles potentielles en identifiant celles qui sont des facteurs de transcription et qui peuvent réguler le miARN.

Plusieurs approches ont été développées afin d'identifier les cibles d'un miARN en utilisant une panoplie de caractéristiques (pour une liste exhaustive, voir la section 2.5.1). Durant les dernières années, la complexification importante des modèles de prédiction de cibles de miARN n'a pas menée à une augmentation significative des performances des outils de prédictions. Cet état de fait suggère qu'une caractéristique clé permettant d'identifier les cibles des miARN reste inutilisée.

C'est dans ce contexte et afin d'identifier des caractéristiques jusqu'alors inconnues du mécanisme d'action des miARN qu'a été faite la prédiction de cibles de miARN par

modélisation des possibilités d'association des miARN aux ARNm (voir Figure 1-2). Cette approche de modélisation identifie les cibles biologiquement appropriées d'un miARN parmi un ensemble de cibles potentielles en considérant, en plus de l'affinité d'un miARN pour un site sur un ARNm, le contexte cellulaire à l'étude, *i.e.* les quantités du miARN à l'étude, des autres miARN et des cibles potentielles.

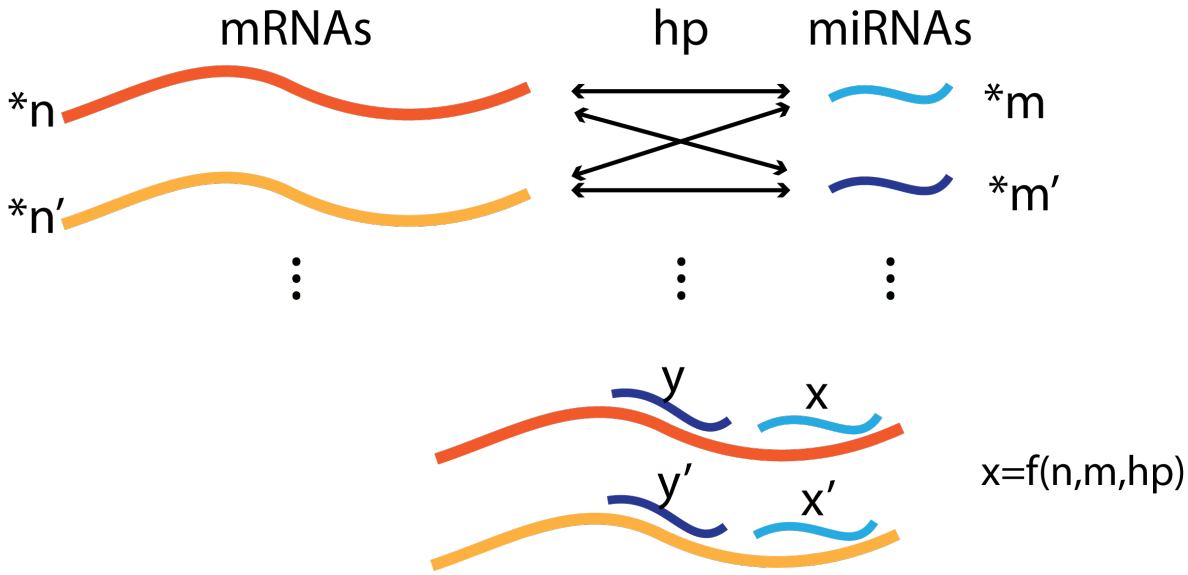


Figure 1-2 Représentation graphique de la deuxième modélisation.
La modélisation prend en considération les quantités de chaque ARNm et miARN (*n, *m) et les probabilités d'hybridation (hp) de chaque miARN à chacune des positions de chacun des ARNm (*i.e.* le contexte cellulaire) pour produire une assignation des miARN sur les ARNm. Le nombre de miARN assigné (x) à une position donnée est déterminé par une fonction qui prend en considération le nombre de miARN (m), le nombre d'ARNm (m) et la probabilité d'hybridation (hp). Ce modèle implique que deux contextes cellulaires différents produisent deux assignations différentes.

1.2.2 L'antithèse

La modélisation de certains aspects de la régulation par des miARN présente trois défis de taille. Le premier est que les données expérimentales sont bruitées de façon inhérente à cause de la faible précision expérimentale mais aussi et surtout par la complexité du réseau d'interactions engendré par la régulation par les miARN. Ces deux caractéristiques combinées à l'effet subtil causé par les miARN font en sorte qu'il est difficile d'identifier, hors de tout doute, les cibles directes des miARN. Le deuxième défi est qu'il est impraticable de modéliser l'intégrité de nos connaissances sur les

miARN dans un seul outil puisque plusieurs parmi celles-ci sont en opposition directe. On peut penser notamment au fait que l'utilisation de la conservation entre espèces diminue le bruit dans les prédictions qui est en opposition avec le fait que certaines cibles sont spécifiques à une espèce donnée. Le troisième défi est que malgré l'étendue de nos connaissances sur les miARN, le réseau d'interaction des miARN avec les autres composants cellulaires est probablement beaucoup plus étendu que ce qui est présentement connu. Il a par exemple été montré très récemment que les miARN peuvent réguler le métabolisme de certains lipides présents dans la cellule, une possibilité qui n'était, jusqu'à récemment, pas envisagée (Rayner, Fernandez-Hernando et al. 2012). Il semble donc utopique de tenter de modéliser quelque chose dont les caractéristiques sont d'être à la fois bruité, complexe et partiellement inconnu.

1.2.3 La synthèse

Nous proposons ici que cette modélisation est possible en faisant quelques simplifications clés et en ne modélisant qu'une facette des miARN à la fois. Ces simplifications permettent de conserver assez d'information sur les cibles des miARN pour pouvoir les identifier. Grâce à ces simplifications qui permettent la modélisation, celle-ci mène à mieux comprendre le rôle des miARN, facilite leur annotation et met en lumière des conséquences inconnues de leur effet.

1.3 Résultats principaux

Cette thèse développe l'idée de modéliser certains aspects des miARN afin d'identifier leurs cibles biologiquement actives à travers deux manuscrits correspondant chacun à la modélisation d'un aspect des miARN. Le premier manuscrit présente la modélisation des régulations des miARN par l'identification de boucles de régulation entre des miARN et des FT. Cette approche a permis, notamment, d'identifier plus de 700 boucles de régulation miARN/FT conservées entre l'humain et la souris. La modélisation a, en particulier, permis d'identifier deux

boucles d'auto-régulation entre LMO2 et les miARN miR-223 et miR-363. Des expériences de transplantation de cellules souches hématopoïétiques et de progéniteurs hématopoïétiques ont ensuite permis d'assigner à ces deux miARN un rôle dans la détermination du destin cellulaire hématopoïétique.

Le deuxième manuscrit présente la modélisation des interactions des miARN avec les ARNm afin de déterminer les cibles des miARN. Ces travaux ont permis la mise au point d'une méthode simple de prédiction de cibles de miARN dont les performances sont meilleures que les outils courants. Cette modélisation a aussi permis de mettre en lumière certaines conséquences insoupçonnées de l'effet des miARN, telle que la spécificité liée au contexte cellulaire des cibles de miARN et l'effet de saturation de certains ARNm par les miARN. Cette méthode peut également être utilisée pour identifier des ARNm dont la surexpression fait augmenter un autre ARNm par l'entremise de miARN partagées et dont les effets sur les ARNm non ciblés seraient minimaux.

1.4 Contributions

Les contributions des résultats des travaux de recherche présentés dans cette thèse se situent principalement au niveau de la compréhension de la fonction des miARN. Certains résultats obtenus lors de la validation de la première modélisation ont des applications en recherche sur l'hématopoïèse et la leucémie. Finalement, des travaux de recherche connexes ont des applications dans le domaine de recherche sur la structure de l'ARN.

1.4.1 Contributions dans la communauté des miARN

La première modélisation a permis d'identifier plus de 700 boucles de régulation entre des miARN et des FT. Ces boucles représentent une première étape importante vers l'annotation complète de ces miARN. La méthode par laquelle ces boucles ont été identifiées permet d'obtenir une liste de cibles potentielles qui ont une haute probabilité de véracité, un résultat en soi très utile. De plus, à travers le rôle

connu des FT, ces boucles fournissent un contexte dans lequel les miARN ont vraisemblablement un rôle biologique. Par exemple, les boucles impliquant LMO2, un facteur de transcription requis pour l'hématopoïèse normale, permettent d'émettre l'hypothèse d'un rôle hématopoïétique des miRNA de ces boucles: miR-223 et miR-363. La prédiction de l'existence de ces boucles oriente donc la recherche d'autres cibles d'un miARN (tel que miR-223 et miR-363), vers celles qui ont une fonction connue dans le contexte dans lequel le miARN est impliqué (tel que l'hématopoïèse pour miR-223 et miR-363).

La deuxième modélisation a permis la mise au point d'une méthode simple et précise de prédiction des cibles de miARN en fonction d'une identité cellulaire précise définie par l'abondance absolue des divers miARN et ARNm. La considération de l'identité cellulaire fait en sorte qu'il est possible d'utiliser cette méthode afin de prédire l'effet dû directement aux miARN d'une modification de l'abondance de miARN ou d'ARNm sur tous les ARNm exprimés dans le contexte cellulaire à l'étude. Ces travaux de recherche, tout comme certains travaux antérieurs, mettent en évidence la variété de cibles des miARN. Ils permettent aussi d'expliquer plusieurs observations faites sur les cibles des miARN qui ne peuvent pas être expliquées par les autres outils de prédictions de cibles telles que la spécificité liée au contexte cellulaire et la saturation d'ARNm par des miARN.

1.4.2 Contributions à la biologie moléculaire en dehors de la communauté des miARN

Les travaux de validation de la modélisation des boucles de régulation des miARN ont nécessité, entre autres, la quantification précise des niveaux protéiques de LMO2, un facteur de transcription, à partir d'échantillons de petite taille. Pour ce faire, j'ai mis au point une technique de détection par cytométrie en flux qui m'a permis de quantifier les niveaux de LMO2 dans diverses populations rares de cellules primaires hématopoïétique. Une telle quantification n'avait jamais été réalisée auparavant. LMO2 est un des nombreux acteurs de l'hématopoïèse normale et son activation

aberrante mène à des leucémies lymphoblastiques des cellules T. Une connaissance plus précise et une quantification juste et adaptée à des échantillons de petite taille des niveaux de LMO2 pourront mener à une meilleure compréhension des rôles de LMO2 dans l'hématopoïèse normale et aberrante.

1.4.3 Contributions à la communauté de la structure de l'ARN

Les travaux de recherche présentés ici ont été réalisés dans un laboratoire de recherche dont le centre d'intérêt primaire est la structure de l'ARN. En plus des travaux de modélisation de certains aspects des miARN, j'ai également contribué à l'amélioration de la modélisation de la structure tridimensionnelle des ARN formant une structure en tige boucle se terminant par une boucle de trois nucléotides non appariés (triloop). Pour ce faire, j'ai bâti un catalogue des structures connues d'ARN formant une triloop et j'ai ensuite développé une méthode de prédiction de la structure tridimensionnelle basée sur le type d'appariement présent à la base de la triloop. Ce catalogue permet de raffiner la modélisation tridimensionnelle de la structure d'ARN.

J'ai également contribué à des travaux de recherche visant à modéliser la relation structure/activité d'un ARN. Les validations expérimentales que j'ai réalisées suite aux prédictions informatiques permettent de mieux comprendre les déterminants structuraux de la fonction d'un ARN bien spécifique, le motif sarcin/ricin de la sous-unité de taille 23S du ribosome de *E. coli*. De plus, les résultats expérimentaux valident la méthode de prédiction informatique et donc celle-ci pourra maintenant être appliquée à d'autres ARN dont l'activité biologique est intimement liée à la structure.

1.5 Résumé des chapitres

1.5.1 Chapitre 2: Travaux précédents

Ce chapitre présente une revue de la littérature des connaissances actuelles dans le domaine des miARN en se limitant à ce qui est nécessaire à la compréhension des travaux de recherche présentés ici. On y retrouve une description de la biogénèse des miARN ainsi que de leur mécanisme d'expression. On y traite ensuite de ce qui est connu de la régulation de l'expression des miARN et des effets des miARN, notamment au niveau des boucles d'auto-régulation. Ces informations mettent en lumière la nouveauté de notre modélisation pour la prédiction de ces boucles de régulation entre miARN et FT. Finalement, on y décrit les principaux outils de prédiction de cibles de miARN pour introduire la modélisation permettant la prédiction de cibles de miARN effectuée dans le cadre de cette thèse.

1.5.2 Chapitre 3: Modélisation des boucles de régulations miARN/FT

Ce chapitre présente l'article traitant de l'identification de cibles fonctionnelles de miARN par modélisation de la régulation des miARN en fonction de leurs interactions avec des facteurs de transcription. Les détails de la méthode informatique permettant d'identifier les boucles d'autorégulation y sont présentés. Les deux boucles impliquant LMO2 et les miARN miR-223 et miR-363 sont ensuite vérifiées expérimentalement par diverses stratégies de quantifications des niveaux des miARN et de LMO2 en conditions normales et en surexpression des miARN ou de LMO2. Une fonction de miR-223 et de miR-363 et des boucles impliquant LMO2 dans l'hématopoïèse est proposée basée sur des expériences de transplantation de cellules souche hématopoïétiques et de progéniteurs multipotents surexprimant l'un ou l'autre de ces miARN.

1.5.3 Chapitre 4: Modélisation des possibilités d'association de miARN et d'ARNm

Ce chapitre présente l'article traitant de la prédiction de cibles de miRNA par modélisation des possibilités d'association entre les miARN et les ARNm. Les détails de la méthode informatique élaborée pour prédire les cibles des miARN sont décrits de même que les résultats expérimentaux publiés reproduits par la méthode. Les conséquences du modèle sont ensuite décrites par l'entremise de trois comportements observés lors des diverses expériences virtuelles de surexpression de miRNA ou d'ARNm et de comparaisons de cibles dans divers contextes cellulaires. Ces phénomènes sont la saturation d'un ARNm, les effets de cascade et de spécificité. La saturation se produit lorsque l'augmentation de miARN ne produit aucun effet parce que tous les sites possibles de liaison de miARN sont occupés. Les effets de cascades se produisent lorsque l'augmentation d'un miARN fait diminuer un ARNm par déplacement d'un autre miARN sur un autre ARNm. Finalement, les effets de spécificité de contexte sont identifiés lorsqu'un ARNm est la cible d'un miRNA dans un contexte mais pas dans un autre.

1.5.4 Chapitre 5: Travaux de recherche connexes

Dans le cadre de la réalisation de mes travaux de thèse, j'ai participé à deux autres modélisations qui, quoique n'étant pas liées à l'identification de cibles de miARN, décrivent des approches intéressantes de modélisation qui permettent de mieux comprendre des phénomènes biologiques. Ce chapitre présente les deux articles correspondant à ces deux autres modélisations

1.5.4.1 Catalogue des structures de "triloop"

Cette section présente l'article décrivant une modélisation de la structure tridimensionnelle fine d'un ARN replié en tige boucle par l'utilisation d'un catalogue de structures connues. Les détails de la construction du catalogue basée sur une classification à partir des éléments informatifs sont donnés. Suivent ensuite quelques

cas intéressants de variations de structure dues au dynamisme de la structure de l'ARN ou à la présence de molécules externes. Une méthode de modélisation à partir d'un tel catalogue est ensuite détaillée.

1.5.4.2 Analyse quantitative de la relation structure/activité

Cette section présente un article traitant de la modélisation de la relation entre la structure et la fonction biologique du motif sarcin-ricin de la boucle E du ribosome de *E. coli*. Cet article présente une méthode informatique de prédiction de l'activité biologique d'une séquence en se basant sur sa structure prédite. Suit une vérification des prédictions informatiques par des essais de croissance de bactéries ayant une variété de mutations de séquences dans ce motif et dont la structure correctement repliée permet la résistance à certains antibiotiques.

1.5.5 Chapitre 6: Discussion et conclusion

Le chapitre 0 présente la conclusion des travaux présentés et est divisé en deux sections. La première résume et discute des résultats principaux présentés dans cette thèse et leur implication dans un contexte plus global. La deuxième section discute de directions de recherche futures qui découlent des travaux présentés dans cette thèse.

2 Travaux précédents sur les miARN

Tel que mentionné précédemment, le rôle de divers ARN dans la cellule est connu depuis longtemps. Que ce soit les ARN de transfert (ARNt), les ARN messagers (ARNm) ou les ARN ribosomaux (ARNr), l'importance de ces molécules ne fait aucun doute. Plus récemment, un nouveau groupe d'ARN a été identifié, les ARN non-codants (ARNnc). Ces ARN sont transcrits par la cellule mais ne codent pas pour des protéines. Ils ne sont généralement pas traduit mais des évidences expérimentales suggèrent que certains ARNnc puissent l'être (Ingolia, Lareau et al. 2011). Plusieurs types de ces ARNnc ont été identifiés et associés à une variété de fonctions. Ils sont généralement classés en deux groupes distincts selon leur taille : les longs et les petits ARNnc.

Parmi les longs ARNnc, on retrouve, entre autres, les longs ARNnc intergéniques (long intergenic non-coding RNA, lincRNA) qui ont commencé à être caractérisés en 2009 et sont impliqués dans une variété de fonctions biologiques (Guttman, Amit et al. 2009, Khalil, Guttman et al. 2009). Un autre long ARNnc récemment identifé est l'ARN circulaire (circular RNA, circRNA) dont la fonction semble être d'attirer les miARN pour les empêcher de lier leur cible naturelle (Hansen, Wiklund et al. 2011, Hansen, Jensen et al. 2013, Memczak, Jens et al. 2013).

Dans le cas des petits ARNnc, on retrouve les ARN associés aux protéines PIWI (piwi associated RNAs, piRNAs) identifiés en 2006 (Aravin, Gaidatzis et al. 2006, Girard, Sachidanandam et al. 2006, Grivna, Beyret et al. 2006, Watanabe, Takeda et al. 2006) et requis pour le contrôle d'éléments génomiques mobiles tels que les rétrotransposons, en particulier dans les cellules germinales (Siomi, Sato et al. 2011). Deux ARNnc localisés au noyau sont connus depuis plus longtemps : les petits ARN nucléolaire (small nucleolar RNA, snoRNA) et les petits ARN nucléaires (small nuclear RNA, snRNA). Les premiers guident les modifications chimiques de d'autres ARNnc tels que les ARNr et les ARNt (Bachellerie, Cavaille et al. 2002) alors que les second sont impliqués dans la maturation des ARNm (Matera, Terns et al. 2007).

On retrouve également parmi ce groupe les microARN (miARN), les petits ARNnc les mieux caractérisé autant du point de vue de leur expression et maturation que du point de vue de leur mécanisme d'actions, leurs cibles et leurs fonctions. Les miARN sont au cœur de cette thèse et donc les connaissances actuelles sur les miARN, notamment leur biogénèse, leur rôle, leur mécanisme d'expression et certaines de leurs régulations connues sont décrites ici.

2.1 Biogénèse des microARN

Le premier miARN (lin-4) a été identifié en 1993 dans l'organisme *C. elegans* par son effet sur la protéine Lin-14 durant le développement embryonnaire. Le mécanisme d'action proposé pour la régulation de Lin-14 par lin-4 impliquait une interaction anti sens entre lin-4 et une région du 3'UTR de Lin-14 complémentaire à lin-4 (Lee, Feinbaum et al. 1993). Ce mécanisme a été démontré peu après par l'utilisation d'un gène rapporteur contenant la séquence complémentaire à lin-4 (Wightman, Ha et al. 1993). Un autre miARN, let-7, a été identifié en 2000, toujours chez *C. elegans* (Reinhart, Slack et al. 2000). Peu après, ce miARN (mais non lin-4) a été détecté dans plusieurs autres espèces incluant *D. melanogaster* et *H. sapiens* (Pasquinelli, Reinhart et al. 2000). En 2001, toujours dans *C. elegans*, un ensemble de miARN ont été identifiés simultanément par divers groupes (Lagos-Quintana, Rauhut et al. 2001, Lau, Lim et al. 2001, Lee and Ambros 2001). Ces résultats et l'identification d'un miARN chez les vertébrés ont fait perdre aux miARN leur étiquette de curiosité spécifique aux nématodes. Ces diverses évidences ont permis d'émettre l'hypothèse que plusieurs autres miARN existent et restent à être découverts dans la majorité des espèces. Une conséquence de cette hypothèse étant qu'ils constituent un réseau de régulations post-transcriptionnelles complètement inconnues jusqu'alors. L'importance des implications de ces découvertes a mené à plusieurs études des mécanismes de transcription, maturation et régulation des miARN et de leurs rôles

biologiques. Ceci est bien reflété par la croissance importante du nombre de publications scientifiques sur les miARN depuis leur identification (Figure 2-1).

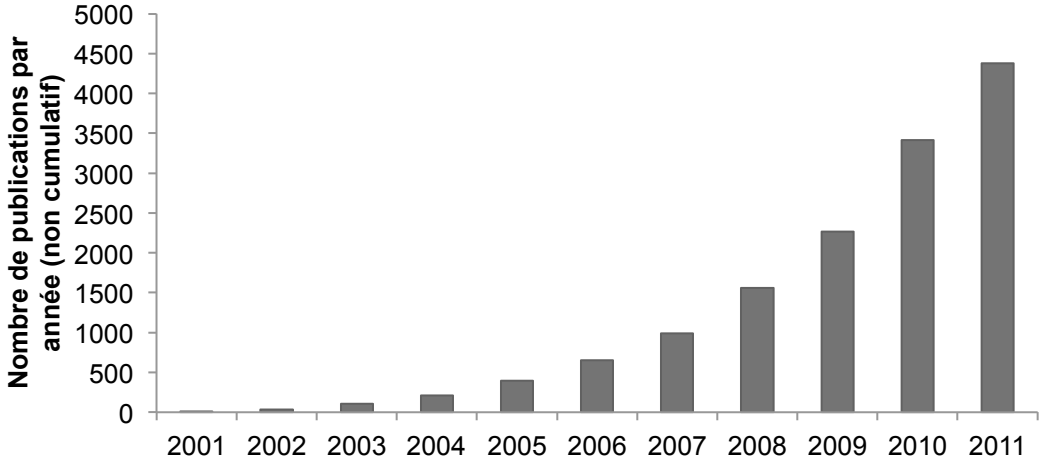


Figure 2-1 Croissance des publications sur les miARN.
Nombre de publications répertoriées dans Pubmed, par année, correspondant au mot-clé "microRNA" depuis leur identification en 2001 jusqu'à la fin de l'année 2011.

Il existe plusieurs ressemblances entre la biogénèse des miARN et celle des ARNm. Le modèle actuel de production des miARN propose que la polymérase à ARN II (POLII) soit responsable de la transcription du miARN; comme c'est le cas pour la majorité des transcrits d'ARN messagers (ARNm) (Cai, Hagedorn et al. 2004, Lee, Kim et al. 2004) (voir Figure 2-2). Le produit de cette transcription est nommé le transcrit primaire de miARN (pri-miARN). Ce transcrit contient une coiffe et queue de polyadénylation (Cai, Hagedorn et al. 2004), comme c'est le cas pour les ARNm. Ce long transcrit primaire dont la taille peut atteindre six kilo bases (kb) mais est généralement retrouvée entre trois et quatre kb (Saini, Griffiths-Jones et al. 2007) est ensuite clivé par un complexe protéique nommé le microprocesseur. Celui-ci est composé, entre autres, de la ribonucléase DROSHA et de son co-facteur DGCR8 (Lee, Ahn et al. 2003). Plusieurs autres protéines sont retrouvées dans ce complexe dont les hélicases de type DEAD-box p68 et p72 et des ribonucléoprotéines (Fukuda, Yamagata et al. 2007, Guil and Caceres 2007). La protéine DGCR8 est responsable du positionnement du microprocesseur sur le pri-miARN alors que Drosha effectue le clivage (Lee, Ahn et al.

2003, Han, Lee et al. 2006). Ce clivage donne naissance au précurseur de miARN (pre-miARN), une tige-boucle dont la taille est d'environ 70 nucléotides (nt) (Liu, Fortin et al. 2008).

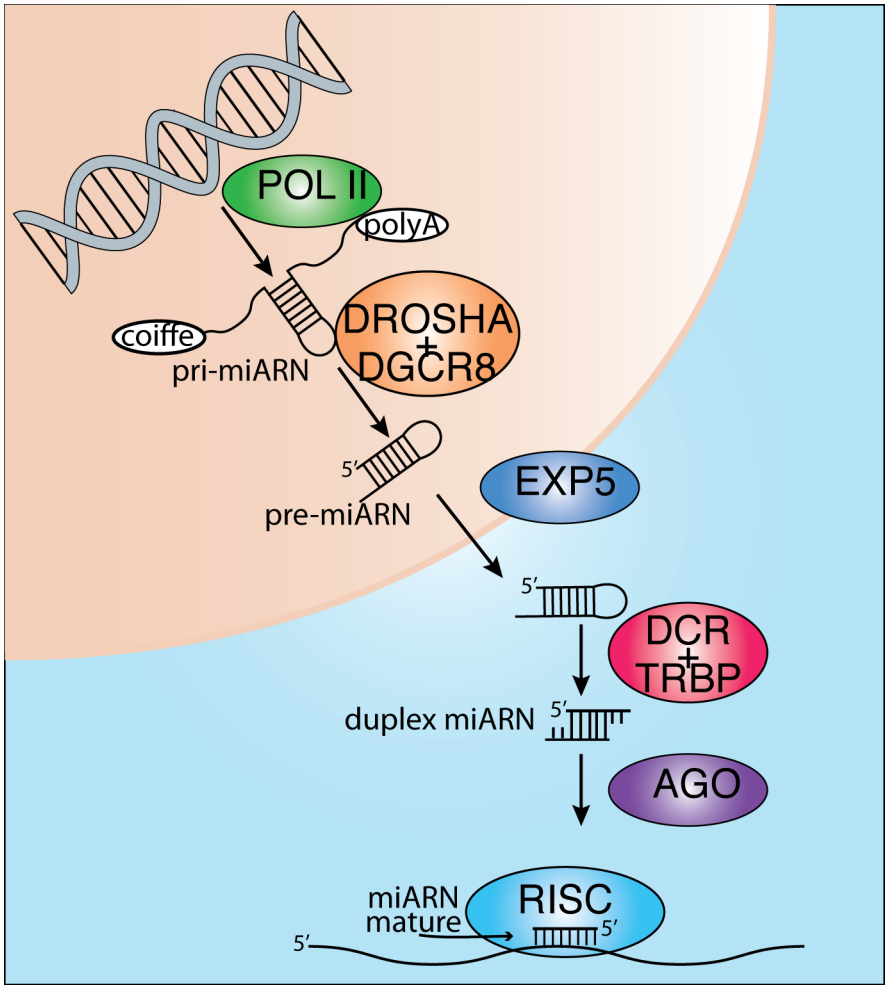


Figure 2-2 Biogénèse des miARN
Le gène de miARN est transcrit par POLII pour donner naissance au pri-miARN. Celui-ci est ensuite clivé par DROSHA en une tige boucle d'environ 70 nt qui est exportée au cytoplasme par la protéine Exportin5 (EXP5). La tige boucle est ensuite clivée par DICER en un double brin d'ARN dont un des brins est incorporé au complexe de silencement par la protéine AGO2.

Un mécanisme alternatif de génération du pre-miARN est celui des mirtrons, utilisé pour produire certains miARN introniques. Lors de l'épissage alternatif de l'ARNm, l'intron contenant le mirtron forme une tige-boucle qui sert directement de pre-miARN sans nécessiter d'autres formes de maturation. Ce mécanisme ne requiert donc

pas le complexe protéique contenant DROSHA et DGCR8. Ce mécanisme a tout d'abord été identifié chez la drosophile (Okamura, Hagen et al. 2007) puis chez l'humain où, notamment, miR-877 et miR-1224 sont produits ainsi (Berezikov, Chung et al. 2007).

Le pre-miARN obtenu de l'un ou l'autre mécanisme est ensuite exporté vers le cytoplasme par la protéine exportin-5 (EXP-5) (Yi, Qin et al. 2003, Bohnsack, Czaplinski et al. 2004, Lund, Guttinger et al. 2004) où il est clivé à nouveau par la protéine Dicer (DCR), une ribonucléase III (Bernstein, Caudy et al. 2001, Ketting, Fischer et al. 2001, Knight and Bass 2001). Le domaine PAZ de DCR lui permet de reconnaître le surplomb simple brin à l'extrémité 3' du pre-miARN (Song, Liu et al. 2003, Ma, Ye et al. 2004) alors que ses domaines RNAse III permettent le clivage à environ 21 nt du domaine PAZ. Ce clivage génère un autre surplomb simple brin et donc le résultat est un double brin d'ARN dont la taille est celle du miARN mature additionnée d'un surplomb de deux nt à chaque extrémité 3' du double brin (Bartel 2004). Dépendamment du contexte cellulaire (Ro, Park et al. 2007) ou de déterminants de séquences ou de structures (Khvorova, Reynolds et al. 2003, Schwarz, Hutvagner et al. 2003, Griffiths-Jones, Hui et al. 2011, Yang, Phillips et al. 2011), l'un ou l'autre (ou les deux) de ces deux brins est (sont) ensuite incorporé(s) à un complexe protéique effectif, nommé RISC (pour Rna Induced Silencing Complexe), contenant entre autres, une protéine AGO.

L'hybridation du complexe RISC à sa cible sur l'ARNm mène à une diminution d'expression de la protéine. Cette diminution d'expression peut être due à une déstabilisation de l'ARNm suite à la déadénylation de la queue polyA. La traduction de l'ARNm peut elle même être affectée par un blocage de l'initiation ou de l'élongation de la traduction ou encore par la protéolyse du peptide nouvellement formé (Filipowicz, Bhattacharyya et al. 2008) Plus rarement chez les vertébrés, il est possible que le miARN mène au clivage de l'ARNm (Liu, Carmell et al. 2004).

Les régions caractérisées comme cibles de miARN se retrouvent principalement dans les régions 3'UTR des ARNm. Cependant, la caractérisation à grande échelle de toutes

les cibles de miARN a montré que les régions codantes sont presque autant ciblées que les régions 3'UTR. Cette étude a aussi montrée que les autres régions génomiques, dont font partie entre autre les régions 5'UTR, les régions intergéniques et les introns sont ciblées par les miARN (Chi, Zang et al. 2009). Une multitude d'évidences expérimentales indiquent que la région amorce ("seed") du miARN, soit les nucléotides en position deux à huit à l'extrémité 5' du miARN, doit être parfaitement ou presque parfaitement complémentaire à la cible pour mener à un effet du miARN sur celle-ci (Doench and Sharp 2004, Kloosterman, Wienholds et al. 2004, Brennecke, Stark et al. 2005). Cette spécificité de seulement sept nucléotides signifie qu'en moyenne, tous les 16kb du génome contiennent une cible potentielle de chaque miARN possible (4^7=16384).

2.2 Expression des miARN

Jusqu'à maintenant, un peu plus de 2000 miARN différents ont été recensés chez l'humain (Kozomara and Griffiths-Jones 2011). La croissance du nombre de miARN recensé est en déclin depuis avril 2011 ce qui porte à croire que la majorité des miARN exprimés chez l'humain ont été répertoriés (voir Figure 2-3). Il semblerait donc que sur les plus de 1 billion (4^{20} = 1e+12) miARN possible de séquences différentes, seule une infime portion de ceux-ci existent. Des 2000 miARN connus chez l'humain, la majorité de ceux-ci ont également été recensés chez la souris et d'autres espèces, indiquant que les séquences sélectionnées comme miARN ne sont pas aléatoires mais probablement liées à l'évolution.

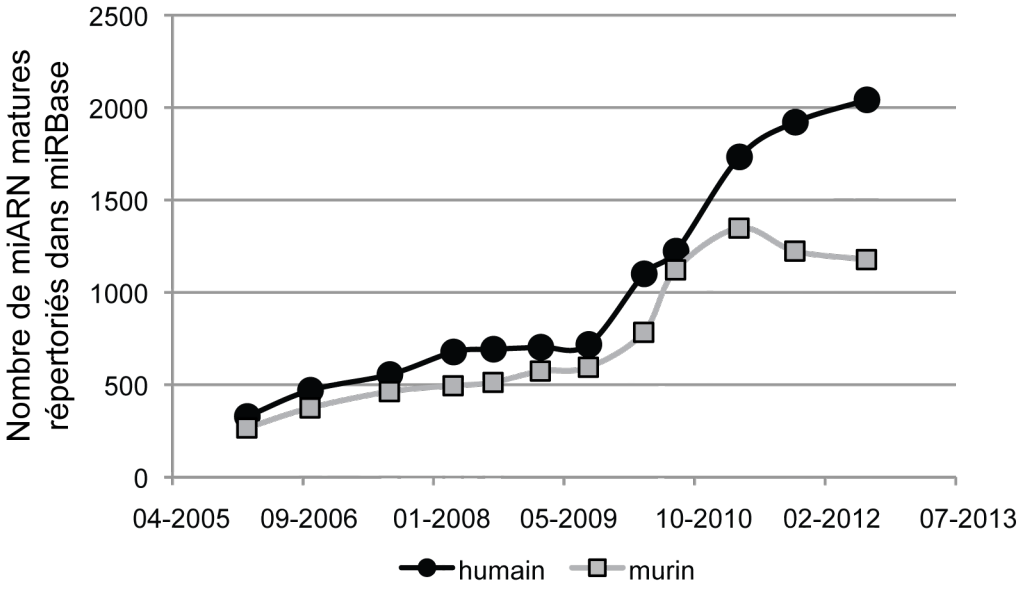


Figure 2-3 Croissance du nombre de miARN connus chez l'humain et la souris.
Nombre de miARN mature recensés chez l'humain et la souris dans miRBase à différents temps correspondant chacun à une version différente de miRBase (versions 8 à 19). Depuis la version 17 (avril 2011) la croissance du nombre de miARN humain connus diminue alors que le nombre de miARN murin connus décroit. Cet état de fait suggère que la majorité des miARN existants dans ces deux espèces ont déjà été répertoriés.

Globalement, les niveaux d'expression des miARN sont relativement élevés; le miARN le plus exprimé dans une cellule étant présent à au moins 25 000 et jusqu'à 50 000 copies par cellules (Bissels, Wild et al. 2009, Shin, Nam et al. 2010, Janas, Wang et al. 2012). À titre de comparaison, près de 32 000 ARNm sont recensés dans RefSeq et chacun de ceux-ci est exprimé à moins de 10 000 copies par cellule (Velculescu, Madden et al. 1999). Tout comme les ARNm, plusieurs miARN ont un patron d'expression spécifique à un tissus ou à un type cellulaire. C'est le cas par exemple de miR-142, miR-181 et miR-223 qui sont spécifiquement exprimés dans le système hématopoïétique (Chen, Li et al. 2004) et de miR-208 qui lui est exprimé dans le cœur (van Rooij, Sutherland et al. 2007). Certains autres miARN connus semblent être ubiquitaires tel que miR-16, miR-23a et miR-29a mais dans ces cas, leur niveau d'expression varie grandement d'un tissus/type cellulaire à l'autre conférant donc une certaine spécificité d'expression (Landgraf, Rusu et al. 2007).

Cette diversité et spécificité des niveaux d'expression a été mise à profit afin de classer des cancers en leurs divers types, et ce, de façon plus efficace qu'en utilisant les profils d'expression des ARNm (Lu, Getz et al. 2005). Ces travaux ont servis de base à d'autres qui ont menés à l'identification de miARN qui puissent servir d'indicateur de diagnostic et de pronostique dans diverses maladies telles que la leucémie lymphocytique chronique (Calin, Ferracin et al. 2005), le cancer du poumon (Yanaihara, Caplen et al. 2006) et le cancer du sein (Iorio, Casalini et al. 2008, Lowery, Miller et al. 2008). L'existence de tels patrons d'expression des miARN, qui rappellent ceux des ARNm suggère que l'expression des miARN est régulée par des FT, comme c'est le cas pour les ARNm.

2.3 Fonction des miARN

La fonction d'un miARN est directement liée aux fonctions des gènes que ce miARN cible. Des études à grande échelle des cibles des miARN ont montré qu'un miARN peut cibler jusqu'à plusieurs centaines d'ARNm différents (Selbach, Schwanhäusser et al. 2008). Un aussi grand nombre de cibles suggère que celles-ci sont impliquées dans plusieurs fonctions biologiques différentes et donc que les miARN, en tant qu'ensemble, forment une classe de molécules dont les rôles sont très diversifiés et ne peuvent être liés à une seule fonction cellulaire spécifique. En fait, la majorité des ARNm représentent des cibles potentielles de miARN (Friedman, Farh et al. 2009) et donc ceux-ci seraient impliqués dans la régulation de la majorité des phénomènes cellulaires.

Par l'étendue et la diversité des gènes qu'ils ciblent, les miARN forment un réseau complexe de régulation de l'expression des protéines précédemment insoupçonné. Les frontières de ce réseau de régulation par les miARN sont encore mal définies parce que, d'une part, toutes les cibles de chaque miARN ne sont pas connues et d'autre part, parce que tous les miARN pouvant être transcrits n'ont pas encore été identifiés. Plusieurs expériences ont été menées afin de tester la nécessité des miARN

au développement normal d'un organisme par délétion des diverses protéines requises à leur maturation. Chez la souris, la perte de Dicer1 (DCR1) est létale tôt dans le développement et mène à des embryons dépourvus de cellules souches embryonnaires (Bernstein, Kim et al. 2003). La délétion spécifique aux cellules T de DCR1 mène à une différenciation aberrante de celles-ci (Muljo, Ansel et al. 2005). La perte de DGCR8 quand à elle mène à une augmentation de l'auto-renouvellement des cellules souches embryonnaires, au détriment de leur différenciation (Wang, Medvid et al. 2007). Ces évidences expérimentales suggèrent que les miARN ont un rôle important dans diverses fonctions biologiques, notamment la différenciation cellulaire. De plus, dans certains processus cellulaires, notamment ceux qui impliquent une importante modification des niveaux d'expressions géniques, le rôle de plusieurs miARN a été démontré et caractérisé.

Chez les mammifères, l'un des premiers systèmes dans lequel le rôle des miARN a été bien étudié est l'hématopoïèse, le processus successif qui permet de différencier des cellules souches hématopoïétiques en une variété de progéniteurs multipotents et ensuite en cellules différenciés du système sanguin (voir Figure 2-4). Une première étude publiée en 2004 a montrée que certains miARN murins (miR-142, miR-181 et miR-223) ont un patron d'expression spécifique aux cellules hématopoïétiques, posant ainsi l'hypothèse que ces trois miARN ont une fonction spécifique à ce système. Cette hypothèse a été supportée par l'observation que la modification des niveaux d'expression de ces miARN affecte le destin cellulaire (Chen, Li et al. 2004). Suite à ces travaux, l'implication d'au moins un miARN dans la majorité des étapes de l'hématopoïèse normale a été démontré. Le rôle de certains des miARN associés à l'hématopoïèse est brièvement décrit ici afin d'illustrer le propos sans alourdir inutilement le texte; des revues plus détaillées des rôles connus des miARN dans l'hématopoïèse normale (Garzon and Croce 2008, Havelange and Garzon 2010, O'Connell, Rao et al. 2010) et aberrante (Kluiver, Kroesen et al. 2006, Garzon, Calin et al. 2009) sont disponibles.

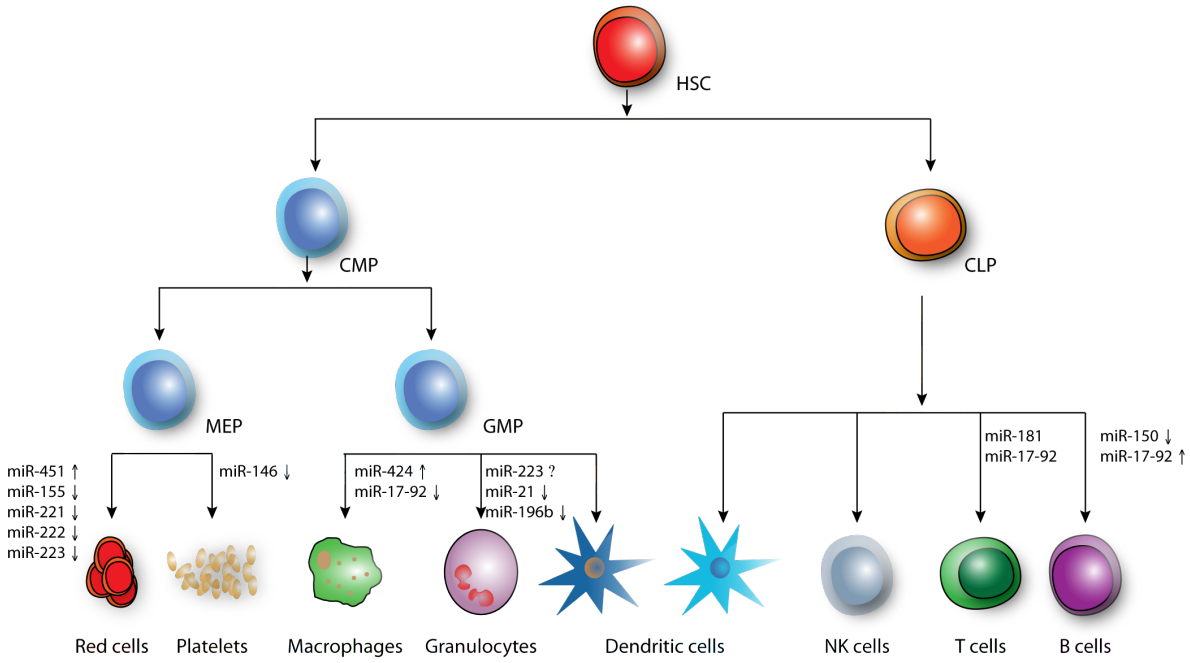


Figure 2-4 Représentation de l'hématopoïèse et des miARN impliqués.
D'une cellule souche hématopoïétique (HSC) dérivent les progéniteurs communs myéloïdes (CMP) et les progéniteurs lymphoïdes (CLP). Des premiers dérivent ensuite les progéniteurs érythrocytes et mégacaryocytes (MEP) et les progéniteurs granulocyte et monocytes (GMP). Divers miARN sont requis (↑) ou doivent être absents (↓) pour permettre la différenciation terminale des cellules hématopoïétiques. Dans le cas de miR-223, des expériences complémentaires supportent des hypothèses contradictoires sur son rôle dans la différenciation en granulocytes (indiqué par un ?). Des études préliminaires suggèrent que miR-181 et miR-17-92 ont un rôle à jouer dans la différentiation des cellules T mais il n'est présentement pas connu s'ils sont requis ou s'ils doivent être absents.

Plusieurs miARN ont été identifiés comme des joueurs clés de la différenciation myéloïde. Des études *in vitro* de gain et de perte de fonctions ont montré que miR-223 active la différenciation en granulocytes, un type cellulaire différencié provenant de la lignée myéloïde (Fazi, Rosa et al. 2005, Fazi, Racanicchi et al. 2007). Par contre, un modèle de souris n'exprimant pas miR-223 a aussi révélé une augmentation de la prolifération des granulocytes (Johnnidis, Harris et al. 2008); mettant en évidence les différences dans les fonctions des miARN pouvant exister entre des lignées cellulaires et des cellules primaires. Le rôle de plusieurs autres miARN dans la différenciation myéloïde a également été démontré. Par exemple, la surexpression de miR-21 et miR-196b dans des cellules de moelle osseuse de souris a montré que ces deux miARN coopèrent pour contrôler la différenciation en granulocytes (Velu, Baktula et al. 2009).

La différenciation des monocytes et des macrophages est promue par miR-424 (Rosa, Ballarino et al. 2007). Le groupe de miARN miR-17-92 quant à lui inhibe la différenciation des monocytes et leur inhibition accélère la différenciation (Forrest, Kanamori-Katayama et al. 2010). La différenciation mégakaryocytique serait inhibée par miR-146 (Labbaye, Spinello et al. 2008). Finalement, dans le cas de l'érythropoïèse l'expression élevée de miR-451 ainsi qu'une diminution de l'expression de miR-155, miR-221, miR-222 et miR-223 semblent être liés à la prolifération normale des érythroblastes (Felli, Fontana et al. 2005, Masaki, Ohtsuka et al. 2007, Felli, Pedini et al. 2009).

Dans la lignée lymphoïde, le groupe de miARN miR-17-92 promeut la différenciation en cellule B (Ventura, Young et al. 2008) alors que la surexpression de miR-150 altère la formation des cellules B matures (Xiao, Calado et al. 2007, Zhou, Wang et al. 2007). Pour ce qui a trait à la différenciation en cellules T matures, miR-181 (Li, Chau et al. 2007) et le groupe de miARN miR-17-92 semblent nécessaire à certains stades de la formation des cellules T matures (Xiao, Srinivasan et al. 2008).

Outre leur rôle dans l'hématopoïèse, les miARN ont aussi un rôle important dans divers autres processus de différenciation cellulaire (Song and Tuan 2006), le cycle cellulaire (Carleton, Cleary et al. 2007), la réponse au stress (Mendell and Olson 2012) et diverses maladies telles que le cancer (Dalmay and Edwards 2006), les problèmes cardio-vasculaires (Small and Olson 2011) et les maladies neurodégénératives (Enciu, Popescu et al. 2012).

Le large éventail de fonctions des miARN en tant que classe de molécules et la possibilité pour les miARN de cibler les ARNm de certaines protéines non atteignables par les approches thérapeutiques standard en font aussi des molécules pharmacologiques intéressantes. Plusieurs miARN, mimiques de miARN ou inhibiteurs de miARN sont présentement à l'étude pour le traitement de divers cancer tel que le cancer du poumon (Trang, Medina et al. 2010) et du foie (Kota, Chivukula et

al. 2009) et certaines maladies cardio-vasculaires (Montgomery, Hullinger et al. 2011).

2.4 Régulation de l'expression des miARN

La majorité des miARN découverts en 2001 sont localisés dans des régions intergéniques ou en orientation anti sens de gènes (Lagos-Quintana, Rauhut et al. 2001, Lau, Lim et al. 2001, Lee and Ambros 2001). Ce positionnement génomique suggère que leur transcription est indépendante de toute autre transcription et utilise un promoteur propre au transcrit de miARN. Par contre, des analyses plus récentes de la distribution génomique des miARN (incluant ceux récemment identifiés) semblent montrer que ceux-ci se retrouvent principalement dans des introns d'ARNm (Rodriguez, Griffiths-Jones et al. 2004). Dans ce cas, leur transcription utiliserait soit le même promoteur que celui de l'ARNm dans lequel il se trouve soit un promoteur propre au miARN (Baskerville and Bartel 2005).

Indifféremment de la localisation génomique d'un miARN, il semble qu'un FT lié à une région promotrice soit nécessaire à leur transcription. Une meilleure compréhension de la fonction des miARN requiert une meilleure compréhension de leurs mécanismes d'expression. Ceci nécessite en particulier une analyse des facteurs de transcriptions régissant la production de ces transcrits. Le fait que les miARN puissent cibler les FT, ceux là même qui régulent leur expression, mène naturellement à s'interroger sur l'existence possible de boucles de régulation entre des miARN et des FT. L'intérêt de ces boucles est d'autant plus grand que certains types de boucles de régulation mènent à l'apparition de comportements émergents (voir section 2.4.2). Divers travaux de recherche ont permis d'identifier expérimentalement quelques boucles de régulation miARN/FT.

2.4.1 Boucles connues miARN/FT

Les premières boucles d'auto-régulation entre un miARN et un facteur de transcription ont été identifiées en 2005. La toute première de ces boucles, identifiée avant même qu'il ne soit clair que les FT pouvaient réguler l'expression des miARN implique le facteur de transcription c-MYC qui régule à la fois E2F1 et miR-20a. Ce dernier régule également E2F1, formant ainsi une boucle de régulation composée de 2 FT et d'un miARN (O'Donnell, Wentzel et al. 2005) (voir Figure 2-5C). Cette boucle permettrait de contrôler étroitement les signaux prolifératifs de la cellule. Suite à l'identification de cette boucle de nombreuses autres boucles de tous types topologiques ont été identifiées par diverses méthodes et dans divers contextes physiologiques. C'est le cas de la boucle impliquant les FT C/EBPα et NFI-A et le miARN miR-223 qui, dans des cellules humaines en culture, régulerait la granulopoïèse (Fazi, Rosa et al. 2005), ou de celle impliquant DAF-12 et let-7 chez *C. elegans* qui permettrait d'intégrer les signaux environnementaux et le minutage développemental (Hammell, Karp et al. 2009). En fait, des boucles de régulation miARN/FT ont été identifiées dans la majorité des organismes modèles communément utilisés et sont impliquées dans une variété de fonction (voir Tableau 2-I). Il ne fait donc aucun doute que les boucles de régulation miARN/FT sont largement répandues et présentes dans tous les processus biologiques. Les phénomènes observés dans ces contextes ne sont pas simplement dus à la cible du miARN mais bien à la présence de la boucle de régulation avec le FT.

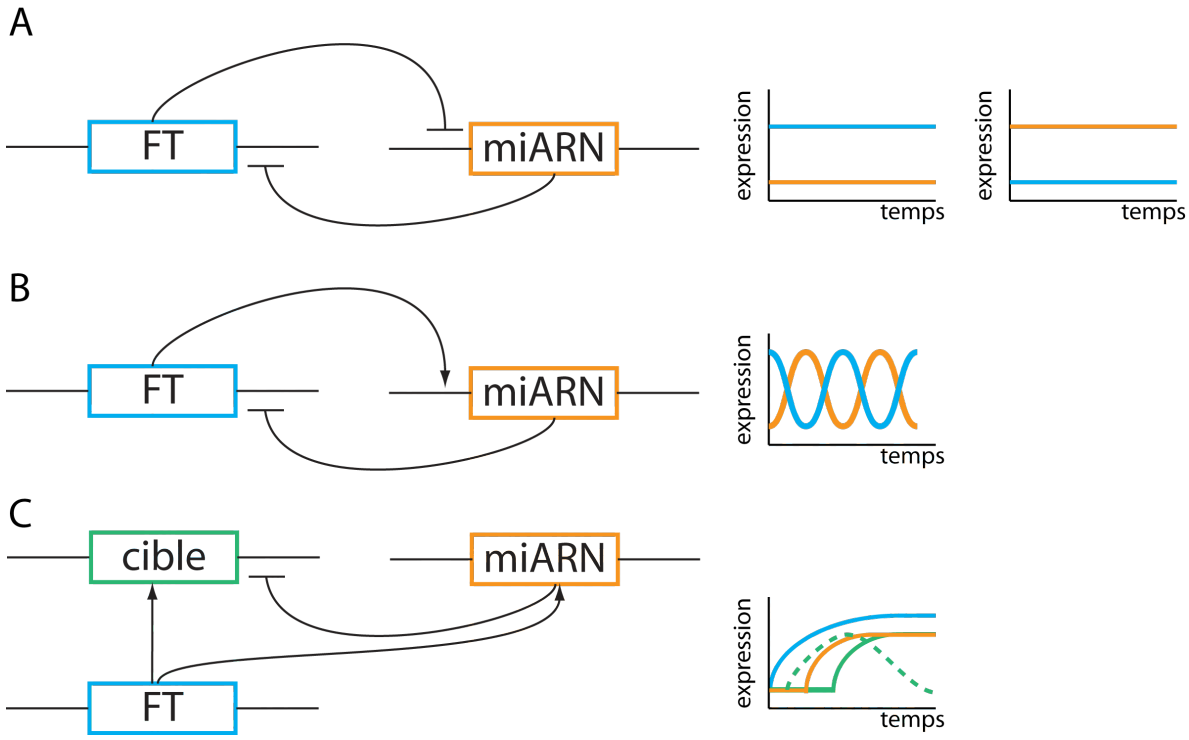


Figure 2-5 Exemples de comportements d'expression causés pas des boucles de régulation.
A Lorsque deux éléments forment une boucle double négative, l'un ou l'autre des deux éléments est exprimé, jamais les deux. Cette situation est représentée à droite où soit le FT (bleu) ou le miARN (orange) est exprimé. Le passage d'un état à l'autre est le résultat d'une modification extérieure des niveaux du miARN ou du FT. Ce type de boucle est notamment observé entre NFIA et miR-223 et également entre LMO2 et miR-223. **B** Lorsque deux éléments forment une boucle positive/négative, l'expression de ceux-ci oscille dans le temps. L'augmentation du miARN fait diminuer les niveaux du FT jusqu'à ce que ceci mène à une diminution des niveaux du miARN. Cette diminution du miARN fait augmenter le FT ce qui augmente le miARN à son tour et la boucle recommence. Ce comportement est représenté à droite. **C** Un réseau de régulation où un FT cible à la fois un miARN et sa cible accélère la réponse cellulaire. Ceci est représenté à droite par l'expression du FT et de sa cible en présence (courbe pointillée) et en absence (courbe verte pleine) d'effet du miARN sur cette cible. La vitesse d'augmentation du miARN (courbe orange) et de la cible en absence d'effet du miARN (courbe pleine verte) sur celle-ci dépend d'autres facteurs que de la topologie du réseau.

Un intérêt particulier de ces boucles réside dans le fait que les diverses topologies peuvent mener à des effets drastiquement différents. Une boucle double négative entre un FT et un miARN peut permettre l'apparition d'un comportement de type "interrupteur" où soit le FT ou le miARN sont exprimés mais pas les deux (voir Figure 2-5A) (Ferrell 2002). C'est le cas de la boucle entre NFI-A et miR-223 dans des cellules non différenciées (Fazi, Rosa et al. 2005). Une boucle positive/négative quant

à elle permet un comportement d'oscillation où le FT et le miARN sont exprimés en alternance (voir Figure 2-5B) (Alon 2007). Finalement, une boucle où un miARN régule négativement une cible et où une autre protéine (tel qu'un facteur de transcription) régule positivement à la fois le miARN et la cible permet d'accélérer la réponse cellulaire à un changement (Rosenfeld, Elowitz et al. 2002, Herranz and Cohen 2010). C'est le cas de la boucle où le FT MYC régule à la fois E2F1-3 et le groupe de miARN miR-17/20 qui lui-même régule E2F1-1 (voir Figure 2-5C) (O'Donnell, Wentzel et al. 2005, Sylvestre, De Guire et al. 2007). Ce même type de boucles, lorsqu'elle implique plusieurs cibles partagées par le miARN et le FT permet de tamponner le bruit associé aux programmes d'expression génétique (Osella, Bosia et al. 2011). Toute modification des niveaux d'expression de miARN impliqués dans ces types de boucles se répercute à travers le réseau de régulation et peut donc avoir des effets plus importants que ceux liés à sa cible directe. Il est donc primordial de bien comprendre les topologies de réseaux de régulation dans lesquelles se retrouvent les miARN afin de comprendre leurs rôles.

Tableau 2-I Boucles de régulation miARN/FT identifiées expérimentalement

miARN	FT	Organisme modèle	Contexte	Références
lsy-6, miR-273	COG-1, DIE-1	*C. elegans*	Choix du destin cellulaire neuronal	(Johnston, Chang et al. 2005)
miR-7	YAN	*D. melanogaster*	Développement de l'œil	(Li and Carthew 2005)
miR-223	NFI-A	Lignées cellulaires et cellules primaires en culture	Granulopoïèse	(Fazi, Rosa et al. 2005)
groupe miR-106a	AML1	Lignées cellulaires et cellules primaires en culture	Différenciation monocytique	(Fontana, Pelosi et al. 2007)
miR-133	PITX3	*M. musculus*	Différenciation neuronale	(Kim, Inoue et al. 2007)
groupe miR-17-92	C-MYC, E2F	Lignées cellulaires	Apoptose et prolifération	(O'Donnell, Wentzel et al. 2005, Sylvestre, De Guire et al. 2007)
miR-17-5p, miR-20	CYCLIN D1	Tissus et lignées cellulaires humaines	Cancer du sein	(Yu, Wang et al. 2008)
miR-27a	RUNX1	Lignées cellulaires	Différenciation mégakaryocytique	(Ben-Ami, Pencovich et al. 2009)
miR-15a	C-MYB	Lignées cellulaires et cellules primaires en culture	Différenciation érythroïde	(Zhao, Kalota et al. 2009)
let-7	DAF-12	*C elegans*	Minutage du développement	(Hammell, Karp et al. 2009)

2.4.2 Prédiction de boucles d'autorégulation miARN/FT

Afin de mieux cerner le rôle des miARN et compte tenu de l'importance biologique des boucles d'autorégulation déjà identifiées entre des miARN et des FT nous avons décidé de modéliser les interactions directes entre les miARN et les facteurs de transcription dans des réseaux de régulation simples. Cette modélisation permet d'identifier le rôle des miARN impliqués dans ces boucles. Pour ce faire, nous avons développé un outil de prédiction de boucles d'auto-régulation entre les miARN et les FT, conservées entre la souris et l'humain. Des plus de 700 boucles prédites, deux impliquent le facteur de transcription LMO2, une protéine essentielle à l'hématopoïèse primitive normale et un oncogène dont l'expression aberrante mène à des leucémies des cellules T (Nam and Rabbitts 2006). Ceci nous a permis d'émettre l'hypothèse que les deux miARN impliqués dans une boucle de régulation avec LMO2,

miR-223 et miR-363, ont un rôle dans l'hématopoïèse. Puisque l'importance du rôle des miARN dans l'hématopoïèse est largement établie (voir chapitre 2.3), de même que la fonction de LMO2, nous avons cherché à mieux caractériser le rôle de ces boucles et de ces deux miARN durant l'hématopoïèse. Des tests fonctionnels nous ont permis de cerner le rôle de ces boucles plus spécifiquement dans le maintien des cellules souches hématopoïétiques et dans la sélection du destin cellulaire entre les lignées myéloïdes et lymphoïdes.

La prédiction de boucles d'auto-régulation, la validation expérimentale de certaines de celle-ci ainsi que l'identification du rôle hématopoïétique de miR-223 et de miR-363 font l'objet du chapitre 3 de cette thèse.

2.5 Détermination de la fonction des miARN

La fonction d'un miARN est définie par les gènes qu'il cible. Ainsi, un miARN est dit avoir une fonction d'inhibition de l'hématopoïèse s'il cible des gènes qui contribuent à l'hématopoïèse. L'identification formelle des cibles d'un miARN est faite lorsqu'il est montré qu'un miARN se lie sur un ARNm ou diminue l'expression d'un ARNm ou de la protéine pour lequel il code. Diverses approches expérimentales ont été développées afin d'identifier les cibles de miARN. Les principales techniques utilisées sont détaillées dans ce chapitre et résumées dans le Tableau 2-II.

La première technique développée pour l'identification de cibles de miARN est l'utilisation de rapporteurs, en luciférase ou en fluorescence contenant la cible potentielle d'un miARN. La co-expression du rapporteur et du miARN diminue les niveaux du premier par rapport à la co-expression du rapporteur et d'un autre miARN. Cette technique est directe et lorsqu'utilisée en combinaison avec la mutation du site de liaison potentiel du miARN, elle permet d'identifier une cible avec une grande confiance. Le temps requis à la construction des rapporteurs fait en sorte qu'il est difficile, voire impossible d'utiliser cette technique afin d'identifier toutes les cibles d'un miARN; elle se prête mieux à l'étude de cibles potentielles identifiées par une

autre approche. Un autre inconvénient de cette technique est que l'interaction directe détectée ne peut être mesurée qu'après surexpression ou inhibition d'un miARN. Il est difficile d'évaluer l'importance biologique d'une interaction observée dans un système où la cible et le miARN ne sont pas exprimés à leurs niveaux naturels dans ce contexte cellulaire.

Une variante de cette technique consiste à quantifier les niveaux de transcrits (par transcription inverse suivie d'un PCR quantitatif) ou de protéine (par immunobuvardage) d'une cible potentielle après surexpression ou inhibition du miARN. Dans ce cas, un effet détecté n'est pas nécessairement le résultat d'une interaction directe. Cependant, la quantification des transcrits, permet de tester rapidement un nombre important de cibles potentielles qui peuvent ensuite être validées par une technique directe.

Certaines approches à grande échelle ont été développées afin d'identifier toutes les cibles d'un miARN sans avoir recours *a priori* à une liste de candidats potentiels (obtenue par exemple par un outil de prédictions de cibles). C'est le cas du pSILAC (pulsed stable isotope labelling with amino acids in cell culture) qui permet d'identifier, par protéomique, les changements dans la production protéique suite à la surexpression d'un miARN. Une autre approche à grande échelle utilise le séquençage à haut débit ou les micro puces à ADN afin d'identifier les transcrits qui varient après surexpression ou inhibition d'un miARN. Ces deux approches sont indirectes puisqu'elles identifient tous les changements suite à la surexpression/diminution d'un miARN, que ceux ci soient directement dus au miARN ou à un effet secondaire à la modification du miARN.

Tableau 2-II Comparaison des méthodes expérimentales permettant l'identification de cibles de miARN.

Méthode	Type	Modification(s) cellulaire(s) requise(s)	Avantage(s)	Inconvénient(s)
Rapporteur (en luciférase ou fluorescence	Directe	Augmentation d'un miARN et expression du rapporteur contenant la cible.	Faible coût, possibilité de tester plusieurs miARN différents sur une même cible potentielle, mesure l'effet sur l'expression protéique.	Approche par candidat, difficile de tester plusieurs cibles et miARN simultanément.
Quantification des ARNm suite à l'augmentation ou à la diminution d'un miARN.	Indirecte	Inhibition ou augmentation d'un miARN.	Permet de tester rapidement plusieurs cibles potentielles.	Approche par candidat, ne permet pas de détecter des changements dans l'efficacité de traduction qui ne sont pas dus au clivage du miARN.
Immunobuvardage suite à l'augmentation ou à la diminution d'un miARN.	Indirecte	Inhibition ou augmentation d'un miARN.	Mesure l'effet sur l'expression protéique.	Approche par candidat, nombre limité de cibles interrogeables.
pSilac	Indirecte	Inhibition ou augmentation d'un miARN.	Mesure l'effet sur l'expression protéique.	Nombre de protéines quantifiables limité, coût.
Micropuce ou séquençage après modification d'un miARN	Indirecte	Inhibition ou augmentation d'un miARN.	Permet de mesurer tous les transcrits.	Coût, ne permet pas de détecter des changements dans l'efficacité de traduction qui ne sont pas dus au clivage du miARN.
HITS-CLIP	Semi-directe	Aucune	Identifie toutes les interactions miARN/ARNm sans modification cellulaire.	Complexité et coût.

L'identification de cibles de miARN sans aucune modification des niveaux de miARN et d'ARNm n'est présentement possible qu'en utilisant la technique du HITS-CLIP (high-throughput sequencing of RNAs isolated by crosslinking

immunoprecipitation) (Chi, Zang et al. 2009). Cette technique relativement complexe nécessite d'abord de lier les ARN et les protéines entre eux, généralement par exposition à des rayons UV. Une immunoprécipitation utilisant un anticorps spécifique contre AGO, une protéine présente à faible distance du miARN et de l'ARNm lors de l'hybridation de ceux-ci, permet ensuite d'isoler tous les ARN (entre autres les ARNm et miARN) liés à cette protéine. Les liaisons sont ensuite renversées, les ARN isolés et amplifiés pour être ensuite séquencés à haut débit. On obtient alors d'une part, un ensemble d'ARNm liés à la protéine immunoprécipitée et d'autre part un ensemble de miARN également liés à cette protéine. Il est possible de réconcilier les deux par une recherche de motifs de six à huit nucléotides dans les transcrits immunoprécipités et en comparant les motifs identifiés aux régions amorces des 20 familles de miARN immunoprécipités les plus exprimées (deux miARN sont dans la même famille si leur région amorce ont la même séquence). Le désavantage principal de cette méthode, outre le fait qu'elle soit complexe, est qu'elle ne permet pas d'associer hors de tout doute un miARN et un ARNm ce qui en fait une méthode semi-directe. La recherche de motifs de six à huit nucléotides présents dans les régions amorces des miARN et dans les ARNm immunoprécipités est imprécise et n'est pas totalement efficace. Par exemple, dans l'étude originale, 27% des régions présentant un signal de séquençage robuste ne peuvent être associées à aucune des 20 familles de miARN les plus hautement exprimées.

Il n'existe donc présentement aucune méthode expérimentale permettant d'identifier rapidement et précisément toutes les cibles d'un miARN donné. Une meilleure compréhension du rôle des miARN nécessite plus d'études directes des cibles des miARN. Ces études nécessitent des outils de prédiction de cibles de miARN précis et exacts.

2.5.1 Outils de prédictions de cibles de miARN

La découverte de plusieurs miARN en 2001 a été suivie par la mise au point d'une variété d'outils informatiques de prédiction de cibles de miARNs. Un des premiers outils appliqués aux prédictions dans les génomes mammifères, et probablement le plus connu et utilisé présentement est l'outil TargetScan dont la première version a été publiée en 2003 (Lewis, Shih et al. 2003). Cette première version de l'outil identifie, pour chaque miARN conservé dans les génomes de l'humain, de la souris et du rat, les sites conservés dans les 3'UTRs de ces trois espèces qui possèdent une complémentarité avec un miARN en considérant l'ensemble du miARN, pas seulement la région amorce. Cet outil a été modifié à plusieurs reprises pour incorporer les nouvelles connaissances sur les mécanismes d'action des miARN tel que l'importance de la région amorce du miARN (Lewis, Burge et al. 2005), l'importance du contexte de séquence du site prédit (Grimson, Farh et al. 2007) et le raffinement du calcul de conservation de sites potentiels de liaison de miARN (Friedman, Farh et al. 2009). Depuis la première version de cet outil, plusieurs autres méthodes ont également été développées pour identifier les cibles des miARN. La très grande majorité des approches utilisent de diverses façons la complémentarité de la région amorce du miARN et la conservation à divers degrés. C'est le cas, notamment, de Diana-microT (Maragkakis, Alexiou et al. 2009), miRanda (John, Enright et al. 2004) (qui tient également compte de la stabilité thermodynamique du complexe miARN/ARNm) et de Pictar (qui utilise un modèle de Markov caché pour estimer la probabilité qu'une séquence représente un site de liaison d'un miARN par rapport à la probabilité qu'elle représente le bruit de fond) (Krek, Grun et al. 2005). Certaines approches utilisent des critères complètement différents tel que PITA qui est basé sur un modèle purement thermodynamique de l'interaction miARN/ARNm (Kertesz, Iovino et al. 2007) et RNA22 qui lui est basé sur une recherche de patrons récurrents dans les ARNm (Miranda, Huynh et al. 2006).

Une évaluation indépendante de neuf algorithmes de prédiction de cibles de miARN largement utilisés illustre bien la divergence qui existe entre les prédictions et l'identification expérimentale des cibles de miARN. En utilisant les paramètres par défaut de chacun de ces algorithmes et en comparant les prédictions de chacun d'eux à des résultats obtenus par pSILAC (Selbach, Schwanhäusser et al. 2008), la précision (nombre de prédictions correctes / nombre de prédictions totales) varie entre 24% et 58% (Alexiou, Maragkakis et al. 2009) pour des sensibilités inférieures à 50%. En supposant que la méthode de pSILAC soit capable d'identifier la majorité des cibles des miARN, ce résultat illustre que la prédiction correcte des cibles de miARN reste un problème non résolu et que des approches radicalement différentes doivent être développées afin d'améliorer la précision des méthodes existantes. L'identification expérimentale des cibles de miARN par des méthodes directes étant largement dépendante de l'identification informatique de cibles potentielles, une augmentation de la précision des outils de prédiction de cibles de miARN se transformerait directement en une réduction du nombre de sites à tester expérimentalement.

2.5.2 Modélisation et prédiction du microtargetome

Afin d'améliorer la précision de la prédiction de cibles de miARN, nous avons décidé d'utiliser une approche radicalement différente de celles utilisées jusqu'à maintenant. Les évidences expérimentales à la base des hypothèses qui nous ont mené à cette approche sont détaillées dans la section 4.2 de cette thèse. Brièvement, nous modélisons les interactions miARN/ARNm en utilisant un modèle qui soit le plus simple possible tout en respectant certaines caractéristiques physiques de l'hybridation miARN/ARNm telles que l'affinité d'un miARN donné pour un ARNm donné et la quantité de chaque entité. Afin d'être en mesure de reproduire le plus grand nombre d'effets conséquents aux miARN, le modèle a été mis au point en se basant sur des données expérimentales illustrant la complexité des mécanismes d'action des miARN. Il a ensuite été éprouvé sur quatre jeux de données de surexpression de miARN suivi d'analyse à grande échelle des niveaux des divers

ARNm. Ce modèle ne nécessite pas l'utilisation de filtres tels que la conservation entre les espèces, le contexte du site ou l'appariement de la région non-amorce du miARN afin de limiter le nombre de cibles potentielles. L'utilisation des quantités de chaque entité combinées à un algorithme d'appariement simple permet de reproduire la vaste majorité des résultats expérimentaux testés et met en lumière de nouveaux comportements observés du microtargetome. Les détails de cette approche font l'objet du chapitre 4 de cette thèse.

3 Article 1: Genome-Wide Identification of microRNAs and Transcription Factors Regulatory Loops Highlights the Role of the LMO2/miR-223 -363 Loops in Hematopoiesis Cell Fate Determination

en préparation pour soumission

3.1 Contribution des co-auteurs

VL: design expérimental informatique et de biologie moléculaire, réalisation des expériences informatiques et de biologie moléculaire, analyse des résultats, rédaction du manuscrit.

MCS: design expérimental de biologie moléculaire, analyse des résultats de biologie moléculaire.

TH: design expérimental de biologie moléculaire, analyse des résultats de biologie moléculaire.

FM: design expérimental informatique, analyse des résultats informatiques, rédaction du manuscrit.

3.2 Mise en situation

Tel que mentionné précédemment, quelques boucles de régulation entre des miARN et des FT ont été identifiées expérimentalement. L'identification de ces boucles nous a conduit à poser deux hypothèses. La première, qu'il est possible de modéliser de telles boucles de régulation. La deuxième que cette modélisation permettrait d'identifier rapidement toutes les boucles de régulation existante. Une conséquence de l'identification de ces boucles de régulation est qu'il devient alors possible d'annoter les miARN impliqués dans ces boucles avec la fonction des FT auxquels ils sont associés.

L'article présenté ici décrit le modèle qui a été développé pour tester cette hypothèse. Afin de prédire les sites de liaison des FT sur les promoteurs de miARN, cette modélisation utilise d'une part les données génomiques disponibles publiquement sur l'humain et la souris et d'autre part les bases de données répertoriant les caractéristiques des sites de liaison des FT. La prédiction des cibles de miARN est faite en combinant les prédictions de deux algorithmes considérés comme les plus performants. Cette modélisation a permis de prédire plus de 700 boucles de

régulation présentes à la fois chez l'humain et la souris. De celles-ci, cinq impliquant des FT possédant un rôle dans l'hématopoïèse ont été testées au laboratoire et se sont révélées réelles. De ces cinq boucles, deux ont ensuite été sélectionnées afin d'effectuer des tests fonctionnels pour déterminer leur rôle. Ces deux boucles impliquent le FT LMO2 et les miARN miR-223 et miR-363.

La sélection des boucles impliquant LMO2 pour des études fonctionnelles n'est pas le fruit du hasard. LMO2 est un FT essentiel à l'hématopoïèse normale (Warren, Colledge et al. 1994). Celui-ci est la composante limitante d'un complexe multiprotéique de régulation de la transcription composé des protéines LDB1, SCL, E2A et GATA1, ces trois dernières protéines étant les seules à liée l'ADN directement (Wadman, Osada et al. 1997). L'expression aberrante de LMO2, notamment suite à des translocations chromosomiques, provoque des leucémies lymphoblastiques aigues des cellules T (Nam and Rabbitts 2006). De plus, l'expression de miR-223 a été bien charactérisée dans plusieurs sous-populations hématopoïétiques. Ces deux éléments fournissent une base solide à partir de laquelle tester nos prédictions dans un système dynamique.

Genome-Wide Identification of microRNAs and Transcription Factors Regulatory Loops Highlights the Role of the LMO2/miR-223 -363 Loops in Hematopoiesis Cell Fate Determination

Véronique Lisi[1,2], Marie-Claude Sincennes[1,3], Trang Hoang[1,4], François Major[1,5]

[1]Institute for Research in Immunology and Cancer, [2]Department of Molecular Biology, [3]Department of Biochemisty, [4]Department of Pharmacy, [5]Department of Computer Science and Operations Research, Université de Montréal, Montréal, Québec H3C 3J7, Canada

Keywords: hematopoietic stem cell, auto-regulatory loop, microRNA, transcription factor, LMO2, miR-223, miR-363

3.3 Abstract

LMO2 is a transcription factor (TF) required for hematopoiesis. Its expression level must be tightly regulated since overexpression of LMO2 can lead to T cell acute lymphoblastic leukemia. However, the mechanisms by which this regulation occurs is not fully understood. It was reported that part of LMO2's regulation occurs through miR-223. Here, we show that LMO2 and miR-223 are co-regulated through a feedback loop. We first computationally predicted direct feedback loops between several TFs and miRNAs, and we identified two loops involving LMO2: one with miR-223 and another with miR-363. We then validated these predictions experimentally, and characterized the implications of these loops on c-Kit+ Sca-1+ Lin- (KSL) and common lymphoid progenitor (CLP) primary bone marrow cell populations after overexpression of either miRNA. These results underline a role for miR-223 and miR-363 in the differentiation of hematopoietic progenitors. Our approach revealed the widespread existence of auto-regulatory loops between miRNAs and TFs, highlighting their overall importance in gene regulation.

3.4 Introduction

The Lim domain only 2 gene (LMO2) is a small transcription factor (TF) required for normal hematopoiesis (Warren, Colledge et al. 1994, Yamada, Warren et al. 1998). Its improper expression following chromosomal rearrangements yields to T cell acute lymphoblastic leukemia (Nam and Rabbitts 2006). LMO2 does not bind DNA directly. Rather its two Lim domain allow for protein-protein interactions. Thus, its transcriptional activity is mediated through the formation of a large DNA binding multi-protein complex composed of SCL, LDB1, E2A and GATA1, all of which, except LDB1, bind DNA directly (Wadman, Osada et al. 1997). Once formed, this complex can activate or repress the transcription of its target genes (Herblot, Steff et al. 2000, Tremblay, Herblot et al. 2003, Chang, Draheim et al. 2006, Hanawalt 2007). LMO2 levels are tightly regulated during hematopoiesis (Lecuyer, Lariviere et al. 2007, Landry, Bonadies et al. 2009) but the mechanisms by which these are regulated are not fully known. At least part of this regulation seems to be through the effect of microRNAs (miRNAs) since LMO2 was shown to be a direct target of miR-223 (Fazi, Rosa et al. 2005, Yuan, Wang et al. 2008, Felli, Pedini et al. 2009, Malumbres, Sarosiek et al. 2009, Zhang, Jima et al. 2009, Sun, Shen et al. 2010).

MiRNAs are short endogenous ribonucleic acids (RNA), which form a class of post-transcriptional RNA down-regulators of most genes by destabilization or degradation of their target mRNAs. The scope of this mechanism in cell regulation is not yet fully understood, but accumulating evidences suggest that it is central to genetic expression (Hall and Russell 2005) and crucial in the elucidation of the oncogenic process (Kent and Mendell 2006). MiRNAs were shown to be generally involved in both positive and negative transcription regulatory networks (Tsang, Zhu et al. 2007). It was shown that miR-20a, a member of the miR-17-92 miRNA cluster, and the E2F TFs participate in an auto-regulatory feedback loop: i.e., the E2Fs bind upstream of miR-20a and activate its transcription, whereas mir-20a directly targets the E2Fs' mRNAs (Sylvestre, De Guire et al. 2007). Similar regulation loops were

observed by various groups in a diversity of cell lines (O'Donnell, Wentzel et al. 2005, Ben-Ami, Pencovich et al. 2009), cell types (Fazi, Rosa et al. 2005, Fontana, Pelosi et al. 2007, Yu, Wang et al. 2008) and animal models such as worm (Johnston, Chang et al. 2005, Hammell, Karp et al. 2009), drosophila (Li and Carthew 2005) and mouse (Kim, Inoue et al. 2007, Zhao, Kalota et al. 2009).

Another approach determined that TF and miRNAs are significantly more functionally linked than any other gene class (Croft, Szklarczyk et al. 2012). Embedding miRNAs in regulation networks gives rise to specific behaviors. Double-negative loops lead to instability in the system. Ultimately, they act as a switch mechanism, where the system expresses either the miRNA or the TF, but not both (Sneppen, Krishna et al. 2010). On the other hand, positive-negative loops (known as negative feedback loops) have an oscillating behavior that tends to maintain both the miRNA and TF at stable levels (Becskei and Serrano 2000). Given the extent of regulatory loops between miRNAs and TFs, we developed a computational approach, whose results predict, among others, two loops with LMO2: one with miR-223 and another with miR-363. Given the importance of LMO2 in hematopoiesis and the fact that LMO2 was shown to be a target of miR-223, we studied the role of these loops in hematopoiesis. We characterized their action during hematopoietic stem cell (HSC) differentiation, and in particular in the lymphoid cell fate decision.

3.5 Results

3.5.1 Auto-regulatory loops prediction

We devised a computer program to identify auto-regulatory loops between miRNAs and TFs based on the information available from various public resources (see Experimental procedures and Figure 3-1). We applied the program to the human and mouse genomes and kept only the loops present in both. We identified 779 loops involving 130 different miRNAs and 182 different TFs (Table A1-I). Our program does not rely on any *a priori* knowledge of the role of the miRNAs or the TF. This explains why it more than triples the number of loops, when compared to the 243 that were identified using manual curation of the literature (Wang, Lu et al. 2010).

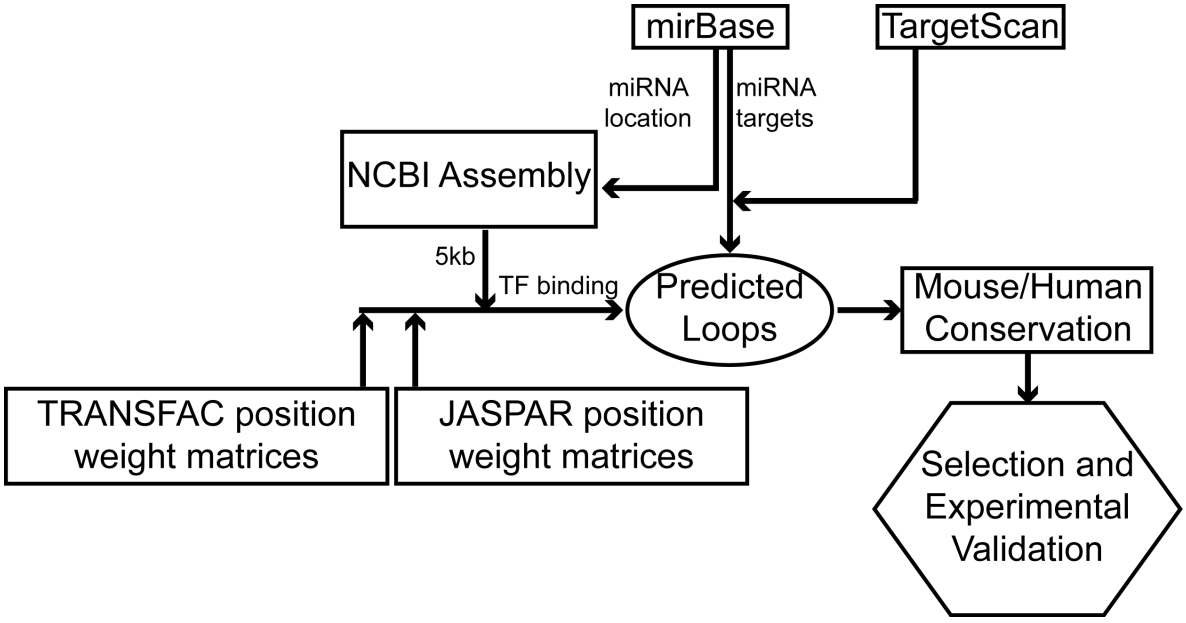


Figure 3-1 Schematic representation of the computational approach used to predict miRNA/TF auto-regulatory loops.
miRBase provides the genomic positions for each miRNAs. We extract from the NCBI human and mouse genome assembly the promoter of each miRNA and apply to it a the TF binding site identification algorithm which takes as input a set of position weight matrices defining each TF binding site sequence characteristics. The miRNAs targets are extracted from both mirBase (now known as microcosm) and TargetScan.

Interestingly, among the predicted loops we find two involving LMO2: one with miR-223 and another with miR-363. We thus investigated the expression levels of miR-223 and miR-363 to assess whether or not it is coherent with these predictions.

3.5.2 LMO2 levels inversely correlate with those of miR-223 in hematopoietic progenitors and erythroid sub-populations

As a first step towards the identification of the possible cooperation between miR-223 and LMO2, we characterized the levels of LMO2 in an array of sorted populations of primary murine cells. We assessed the expression of the LMO2 protein by intracellular staining of LMO2 followed by flow cytometry, allowing for a more accurate quantification of the protein levels of LMO2 in these small samples, which are not amenable to Western blotting. Coherent with the known expression of LMO2, we observed that, in the progenitors populations, the expression of LMO2 was the highest in the LT-HSC, ST-HSC, KSL, LMPP and MEP populations ($LMO2_{high}$ populations) and the lowest in the CMP and GMP populations ($LMO2_{low}$ populations) (Figure 3-2a and A1-1). The flow cytometry results were transformed to a single value for each population tested as described in Methods (Figure 3-2b). This allowed for an easier comparison with the levels of miR-223 quantified by qRT-PCR in these same populations. We observed that in the $LMO2_{high}$ populations, expression of miR-223 is relatively low (compared to the levels in the GMP population for example), whereas in the $LMO2_{low}$ populations, the expression of miR-223 is higher (Figure 3-2b - miR-223 was quantified only in the ST-HSC, LMPP and MEP populations).

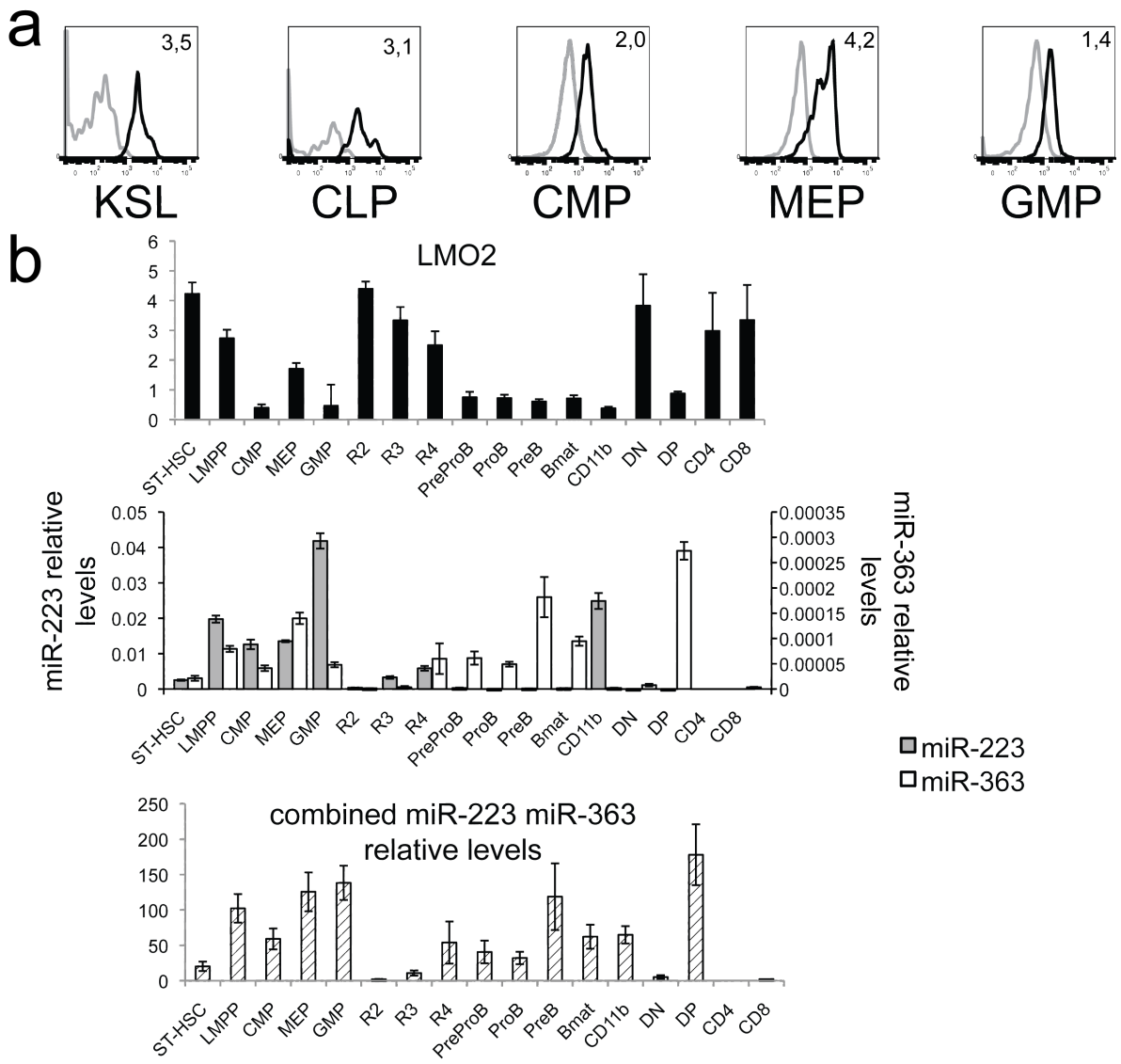


Figure 3-2 LMO2 and miR-223/miR-363 mutually exclusive expression.
(**a**) LMO2 levels in 5 hematopoietic progenitor populations highlighting the regulation of LMO2 as differentiation progresses. LMO2 levels were assayed by intracellular staining followed by flow cytometry. The number in the upper right of each histogram is a quantification of the signal compared to the background (see Method). The levels of LMO2 are the highest in the KSL and MEP populations. (KSL: Kit+Sca+Lin-, CLP: common lymphoid progenitors, CMP: common myeloid progenitors, MEP: megakaryocytic, erythroid progenitors, GMP: granulocyte, macrophage progenitors). (**b**) Quantification (see Method) of the results of the detection of LMO2 by flow cytometry in various populations of the hematopoietic system. The relative LMO2 levels are compared to those of miR-223 and miR-363 as assayed by RT-qPCR. In the populations where LMO2's level are high, neither miR-223 nor miR-363 is highly expressed whereas in the populations where LMO2's level are low, either miR-223 or miR-363 or both are highly expressed. Values represent the mean and error bars the standard deviation of 5 independent experiments. (ST-HSC: short term HSC, LMPP: lymphoid-primed multipotent-progenitors, DP: double positive CD4+CD8+, DN: double negative).

Cells undergoing erythroid differentiation can be followed using Ter119 and CD71 surface markers and classified in four groups (R1 to R4) as differentiation progresses (Socolovsky, Nam et al. 2001). In these populations the levels of LMO2 are decreasing as differentiation progresses from R1 to R4, whereas the levels of miR-223 increases during this process (Figure 3-2b). These results are coherent with those expected with double negative feedback loop in which either the miRNA or the TF is expressed, but not both. In the B cell populations, the levels of LMO2 are overall low. However, the levels of miR-223 are also relatively low. This suggested that other regulators contribute to the expression of LMO2. We hypothesized that other miRNAs could contribute to this regulation.

In the progenitor and erythroid populations, miR-363 globally follow a similar expression patterns as that of miR-223. One exception to this is in the GMP population. In this population, the miR-223 levels are the highest among all the populations we tested which is not the case for miR-363 whose levels are similar to those in the LMPP and CMP populations. Another set of populations in which the levels of miR-223 and miR-363 differ is in the B cell populations. In these, the levels of miR-363 are relatively high compared to those of miR-223, which are relatively low.

In the thymus, the levels of LMO2 were the lowest in the $CD4^{+}CD8^{+}$ population (double positive - DP) and the levels of miR-363 were the highest. In the other cell populations of the thymus (double negative (DN), CD4- and CD8-), the levels of LMO2 are high and the levels of both miARNs are low. In the CD11b+ population, the levels of LMO2 were low and the levels of miR-223 were high. Globally, in each population tested, whenever LMO2 levels were low, miR-223, miR-363 or both miRNAs expressions were high. Inversely, when the levels of LMO2 were high, neither miRNA was highly expressed. This is more easily observable when the sum of the relative levels of miR-223 and of miR-363 is reported (Figure 3-2b, third panel). This inverse expression of LMO2 and the two miRNAs in a wide array of hematopoietic cell populations, together with the prediction of regulatory loops between LMO2 and the

miRNAs support the hypothesis that the miRNAs regulate the levels of LMO2 through two double-negative feedback loops.

3.5.3 LMO2 binds the promoter of both miR-223 and miR-363

To assess whether the inverse correlation in expression observed between LMO2 and miR-223/-363 is the result of a two double negative feedback loops, we first verified if LMO2 binds the promoter of both miRNAs. Our algorithm predicted that LMO2 binds two sites on the promoter of each of miR-223 and miR-363: one E-box and one GATA-site (see Figure 3-3a). By chromatin immuno-precipitation (ChIP) of LMO2, we showed that in addition to the classical *Gypa* promoter LMO2 binding site, three of the predicted sites are indeed bound by LMO2, the two GATA-sites and the E-box on the miR-223 promoter, when compared to the negative control, the *Hprt* promoter. This suggests a regulatory role of LMO2 on these two miRNAs. Combined with evidences from the literature that LMO2 is a target of miR-223, these results bring support to the presence and role of these two regulatory loops in hematopoietic cells.

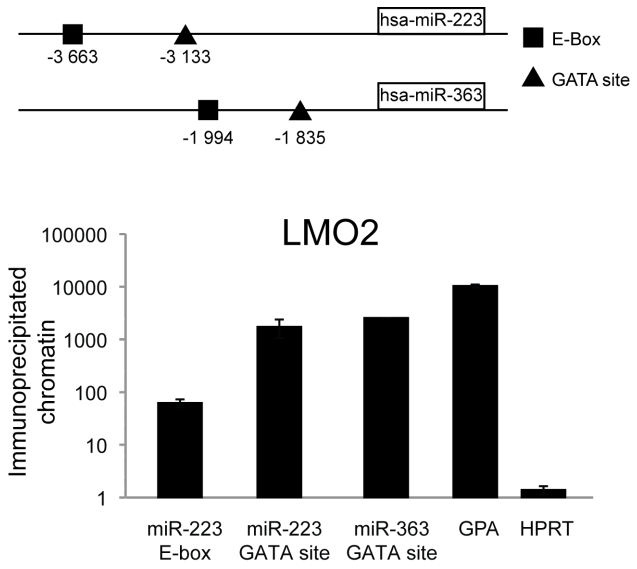


Figure 3-3 LMO2 binds the miRNAs promoters.
The genomic contexts of the predicted binding sites of LMO2 on the miR-223/-363 promoters are graphically represented. The ChIPs performed in TF1 cells against LMO2 shows the TF binds the promoter of the miRNAs when compared to the GYPA promoter, LMO2's classical binding site and to the negative control HPRT.

3.5.4 LMO2 negatively regulates miR-223/-363

We next verified if this binding of LMO2 on the promoter of miR-223 and miR-363 yields a biological regulation. We overexpressed LMO2 by retroviral infection in the Jurkat cell line, expressing LMO2 at low levels, and observed its effects on the expression of the miRNA. The LMO2 overexpression increased by three orders of magnitude the levels of LMO2 and resulted in a 16- and 8-fold decrease on the levels of respectively miR-223 and miR-363 (Figure 3-4a). Together with the binding of LMO2 on the miRNAs promoter, this result confirms the negative regulatory role of LMO2 on the two miRNAs.

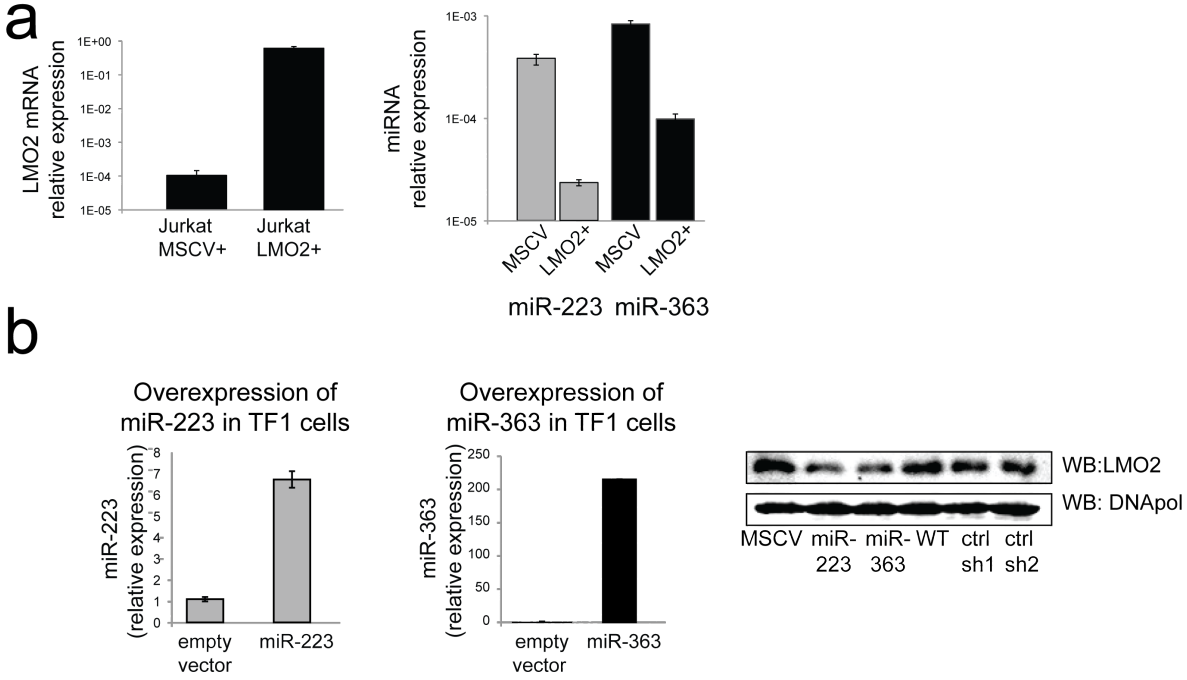


Figure 3-4 LMO2 and miR-223/-363 are involved in double negative regulation loops.
(**a**) Jurkat cells overexpressing LMO2 show a lower expression of both miR-223 and miR-363 compared to cells overexpressing the empty vector. (**b**) TF1 cells overexpression either miRNA show a decreased expression of LMO2 as assayed by western blotting of LMO2. Two control shRNAs do not impact the levels of LMO2.

3.5.5 miR-223 and miR-363 negatively regulate LMO2

Similarly, we examined the effect of overexpressing either miRNAs by retroviral infection of TF1 cells, which express LMO2 at high levels, on the levels of LMO2. Cells overexpressing either miRNAs have decreased expressions of LMO2 (Figure 3-4b). The overexpression of two shRNAs designed to target ETO2 did not decrease the levels of LMO2. This suggests the two miRNAs regulate the protein levels of LMO2, further supporting the presence of these two predicted auto-regulatory loops.

3.5.6 Effect of the miR-223 and miR-363 loops with LMO2 in a dynamic context

We next investigated the role of these two loops. Because of LMO2's established role in hematopoiesis, we hypothesized that the loops could modulate hematopoietic differentiation. We tested this hypothesis *in vivo* through bone marrow transplantation assays. We overexpressed either miRNAs by retroviral infections in fetal liver cells depleted of terminally differentiated cells. These cells were then transplanted into irradiated hosts. We measured two outputs,six months post transplantation: hematopoietic stem and progenitor cell maintenance through the reconstitution of the KSL population and lymphoid differentiation through the CLP population reconstitution *in vivo*, (Figure 3-5a). To ensure that any observed effect was due to the LMO2/miR-223-363 loops, we also transplanted animals with cells expressing a shRNA against LMO2. We confirmed that the overexpression of each miRNA in primary cells lead to a decrease of the levels of LMO2, assayed by flow cytometry following intracellular staining (Figure 3-5b). We observed that the reconstitution of the KSL population is significantly altered by each of the miRNA overexpression (Figure 3-5c). The median number of KSL per animal in the control animals is 4085 cells, whereas mice transplanted with cells overexpressing miR-223 or miR-363 have a median number of KSL cells of 1264 and 2317 respectively. This is also observed with the shRNA against LMO2, where the number of KSL cell decreases 2-fold in the animals transplanted with the shRNA agains LMO2 (Figure 3-5c). Since the phenotype is observed with both miRNAs and with the shRNA against LMO2, the effect is likely due to a decrease in LMO2 expression suggesting the miRNAs have a biological impact on LMO2 expression in hematopoiesis. We thus evaluated the impact of decreased LMO2 levels on the cell fate determination of the transplanted cells. We observed that in the control animals, the ratio of the number of CLP cells over the number of KSL cells is 1.085, meaning that KSL are differentiating almost equally as myeloid and lymphoid progenitors (Figure 3-5d). This ratio is significantly increased when the animals are transplanted with cells overexpressing either miRNA (ratio of

1.61 and 1.63 for the animals transplanted with cells overexpressing miR-223 and miR-363 respectively) (Figure 3-5d). The significant increase in the number of CLP is not matched to a significant decrease in the ratio of any of the myeloid progenitors (Figure 3-5e). This suggests that the miRNAs favor the lymphoid differentiation without affecting the myeloid differentiation.

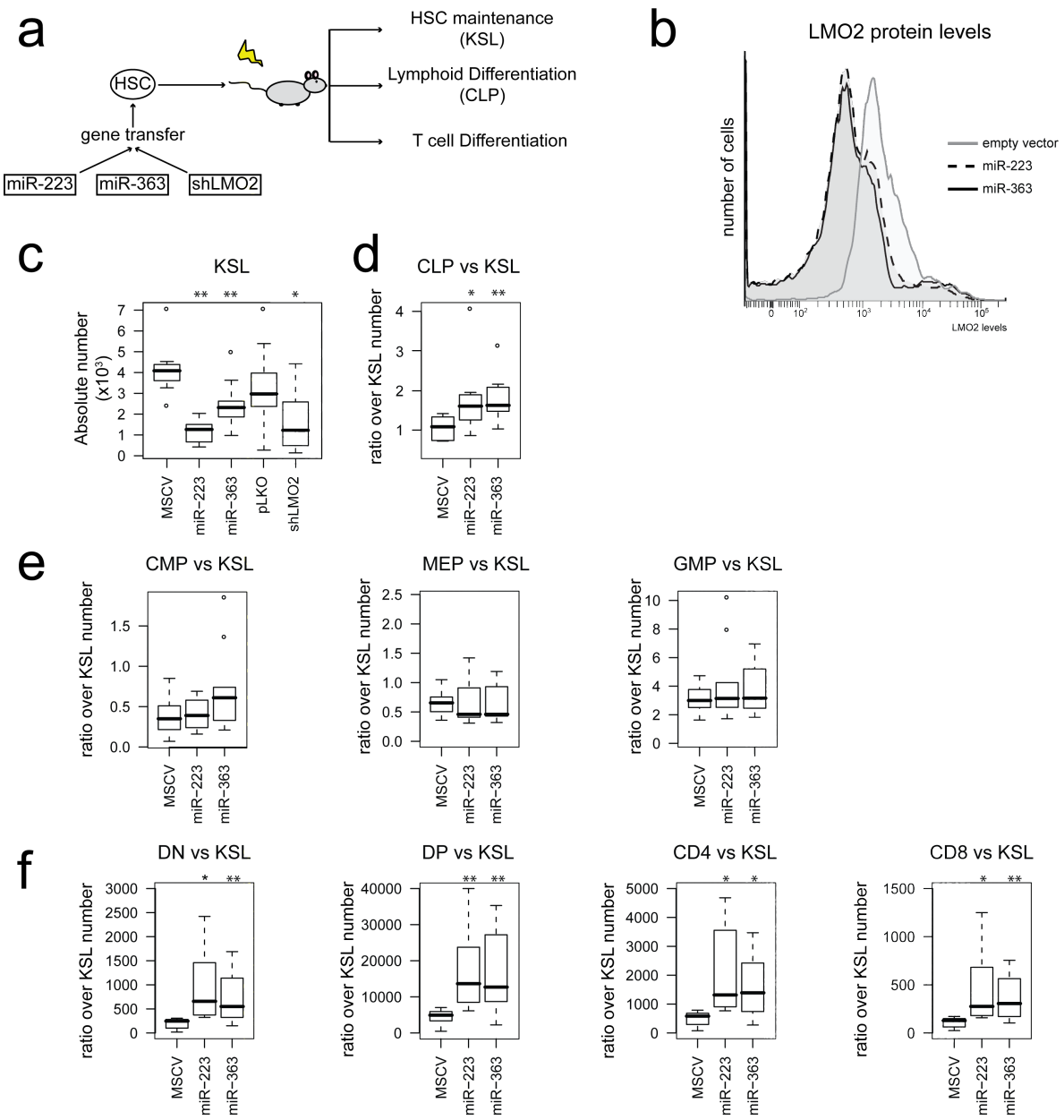


Figure 3-5 Transplantation assay.

(**a**) Fetal liver cells are harvested and depleted of terminally differentiated cells before infection with miR-223, miR-363, or an shRNA against LMO2. These cells are then transplanted in irradiated host that will undergo hematopoiesis with the cells overexpressing either miRNA or the shRNA. Two parameters are studied: hematopoietic stem cell (HSC) maintenance and lymphoid differentiation. (**b**) The infection of primary cells with either miR-223 or miR-363 decreases the levels of LMO2 as assayed by intracellular staining followed by flow cytometry. (**c**) The infection with either miRNA or the shRNA against LMO2 decreases the number of KSL cells per animal. (**d**) miR-223 and miR-363 increases the ratio of the number of CLP cells over the number of KSL cells. (**e**) The increase in the CLP/KSL ratio is not matched with a decrease of any given myeloid progenitor. (**f**) The two miRNAs increase the ratio of each of the number of T cell in each population over the number of KSL. (*: student t-test p-value < 0.05, **: student t-test p-value < 0.01, n≥8).

We next evaluated if the observed bias towards lymphoid differentiation has an impact on T cell maturation. We observed that both miRNAs increase the ratio of the number of differentiated T cell over the number of KSL for all four T cell populations (Figure 3-5f). Further this increase is not solely due to the increase in CLP ratio since the ratio of the number of differentiated T cell over the number of CLP cell is also increased, although the increase is not as important (data not shown). Finally, the number of B cell and their ratio over the number of KSL or CLP are unaffected by either of the miRNAs (data not shown).

Taken together, these results suggest that miR-223 and miR-363 favor the lymphoid differentiation without affecting significantly the myeloid differentiation. From the CLP population, the two miRNAs favor the T cell differentiation without a significant impact on the B cell populations. These two effects combined result in an important increase in the number of differentiated T cells. Since undifferentiated T cells characterize the leukemia originating from an aberrant expression of LMO2, the results obtained here suggest that miR-223 and miR-363 could be used to restore LMO2 levels and favor T cell differentiation.

3.6 Discussion

3.6.1 Unprecedented and systematic prediction of miRNA-TF auto-regulatory loops

MiRNAs are powerful regulators of gene expression because of the large set of genes they each regulate. In that sense, they are reminiscent of TFs, which also each regulate a large array of genes. The co-regulations between miRNAs and TFs yield specific behaviors that are propagated throughout the network of interaction. Identifying these regulatory loops and understanding their behavior is thus necessary to our comprehension of the roles of miRNAs. Our unique method utilizes the state of the art in target prediction of miRNAs and TFs to detect direct co-regulatory loops between them with few false positives. The broadness of high confidence predictions indicates that these loops are rather common.

3.6.2 New understanding of the role of LMO2 during hematopoiesis

The role of the LMO2 transcription factor has been well characterized in erythropoiesis, in part through the identification of a handful of target genes regulated by the LMO2 containing complex (reviewed in (Lecuyer and Hoang 2004)). Its causal role in leukemia is also well documented. We expand our knowledge on the roles of LMO2 by identifying two miRNAs that are direct targets of LMO2 in early hematopoiesis. Since these can also regulate other genes, our work increases the set of potential LMO2 targets to include the genes regulated by LMO2 through the two miRNAs. We further identified a role for LMO2 through the miRNA loops in KSL maintenance, as well as in lymphoid cell fate determination. This is to our knowledge the first time a decrease in the levels of LMO2 (rather than a knockout) has been used to highlight its role in both early hematopoiesis and T-cell differentiation.

3.6.3 Annotating miRNAs

Dozens of miRNAs have not yet been annotated. Our loop predictor can be used to remedy, in part, to this situation. When a loop is predicted between a miRNA and a TF, we can hypothesize that the miRNA is implicated in at least some of the known roles of the TF. For example, the predictions of the LMO2-miR-223/-363 loops led us to test the hypothesis that these miRNAs have an effect during hematopoiesis and to uncover their roles in lymphoid cell fate decision through the reconstitution of the CLP population and the maintenance of the KSL population; a hypothesis that we would not have made otherwise. Since there are so many miRNAs with unknown or limited associated function, and new miRNAs are constantly being identified, it is important to be able to rapidly annotate these miRNAs. The identification of loops using our method provides a mean to do this annotation using the known roles of TFs.

3.6.4 Studying miRNA in context

The reported widespread effect of miRNAs has been shown to be subtle (Baek, Villén et al. 2008, Selbach, Schwanhäusser et al. 2008) on the vast majority of genes, with some drastic exceptions such as the effect of lin-4 and let-7 on, respectively, lin-14 and lin-41. This apparent inconsistency can be partly explained by the presence of regulation loops between miRNAs and TFs and by the overall behavior resulting from these loops. It is thus important when addressing the roles of specific miRNAs to consider not only the regulation of the miRNA on a given target, but also the global effects of this regulation on the complete regulation network. Our work provides the first step in the identification of these large regulatory networks by identifying many of the direct regulation loops.

It was shown that drosophila miR-9a represses dLMO (Bejarano, Smibert et al. 2010), the LMO2 ortholog in the fly. Loss of this regulation via modification of dLMO 3'UTR or by deletion of miR-9a causes loss of adult wing tissue. However, complete inhibition of the miRNA pathway by loss of two critical proteins, Pasha and Dcr-1

yielded a weaker phenotype than that observed with the loss of miR-9a. This further highlights the importance of studying miRNAs in their contexts.

3.6.5 Linking the ENCODE project

Recent reports published as a result of the ENCODE project highlighted the fact that many TFs target more non-coding RNAs, including miRNAs than protein-coding genes (Gerstein, Kundaje et al. 2012). Another conclusion of this study was that miRNAs tend to target sets of TFs that physically interact with each other to shut down the entire functional unit. In this context, the existence of a direct negative feedback loop between a miRNA and a TF may highlight the limiting factors of a functional unit. This is the case for LMO2, which is usually the limiting factor in the assembly of the transcriptional regulation protein complex.

3.7 Materials and Methods

3.7.1 Identification of known miRNAs, their genomic location, promoter region and putative targets

The 940 known human miRNAs and 590 known mouse miRNAs are extracted from release 15 of mirBase (Griffiths-Jones 2004, Griffiths-Jones, Grocock et al. 2006) providing names and genomic locations of all identified miRNAs in the Ensembl release 46 of the genomes. This is used to extract the 5kb region upstream of the miRNA transcription start site. This region is defined as the miRNA promoter. For miRNAs embedded in genes (in exons, introns or UTR regions), the 1.5kb upstream of each transcript it is included in (i-e: each transcript's promoter) is also considered as the miRNA promoter region. Therefore, for each miRNA, we consider at least its direct promoter but if it applies, we also consider one or many upstream regions as the promoter. Each of these regions is treated individually. The set of targets for each miRNA is the union of the predicted targets from MicroCosm version 5 (formely known as miRBase::Targets)(Enright, John et al. 2003) and TargetScan release 5.1 (Friedman, Farh et al. 2009).

3.7.2 Transcription factors

The set of transcription factors contains all the genes with GO term GO:0030528 (transcription regulator activity) of GO:0003677 (DNA binding). These were obtained from the Ensembl human genome assembly 46.

3.7.3 Transcription factors binding sites predictions

For each target of a specific miRNA that is a TF, we compare its position weight matrix (PWM) as described by either TRANSFAC's version 7.0 (Matys, Kel-Margoulis et al. 2006) or Jaspar (Stormo 2000) to the promoter region of the miRNA to find similarities. To do this comparison, we use a modified version of the MatInspector algorithm (Quandt, Frech et al. 1995). For each of the tested regions, we built a

background matrix of identical length to the PWM of the transcription factor under study where each identical column is the relative proportion of each nucleotide in the promoter region. The ratio of the score obtained by the MatInspector algorithm with the PWM and the one obtained with the background matrix is then computed for each region of the promoter of size equal to the size of the TFBS; the highest the ratio, the more similar to the TFBS the region is. We apply this procedure independently for the mouse and human genomics information. On each dataset, the threshold on the ratio score is set at 4 standard deviation of the mean of the normal distribution. When a loop has a score above the threshold for the human and for the mouse data, the loop is considered predicted and is then attributed a score. The final score is the sum of the score returned by our algorithm (values ranging from 0 to 2), a discretization of the p-value returned by the miRanda algorithm which is used in the microcosm database of miRNA targets (p_value < 1e-10, score = 1, 1e-10 <= p_value <1e-5, score = 0.5, 1e-5 <= p_value < 0, score = 0.25) and of the percentile returned by the TargetScan algorithm divided by 100. The higher the score, the more confidence we have in the existence of the loop.

3.7.4 Accuracy of the method

We evaluate the false positive predictions in our approach by looking at the number of auto-regulatory loops generated by this method in a random setting (i.e. when no loop should be predicted). We built a random graph based on the same properties as the actual graph. In this graph 200 miRNAs each have 1500 out-going edges to one of the 35 000 genes and the other 275 miRNAs have 3500 out going edge, based on the targets distributions of mirBase:Targets. Out of the 35000 gene nodes, 1900 are TF and they each have 200 out going edges to miRNAs. This is likely an over-estimation of the actual number of miRNA each TF targets, however, the lower the actual number is, the better the accuracy of the methods. In this bipartite graph, we calculate the number of auto-regulatory loops, i-e the number of miRNAs that have, for a TF, both an out-going and an in-coming edge. In this random setting, 9217 loops

(on average in 50 different random generation) are predicted. Our filter on the scores greater than four standard deviations applied to 9217 random loops returns 885 loops. The conservation criteria between human and mouse returns only 15% of the loops. In the case of the random generation, only 132 loops would be returned. Thus, we expect that out of the 779 loops predicted by our algorithm, 647 are true positive loops, the remaining ones being false positive resulting from the property of the data.

3.7.5 Comparison with Transmir

To evaluate the number of false negative predictions our approach makes, we compare our results, obtained using miRNA and TFs target predictions with those obtained from a literature survey and available as the Transmir database (Wang, Lu et al. 2010). Of the 245 loops found in this database, only 30 involve TFs for which at least one position weight matrix (PWM) is available either in the Jaspar or Transfac database. Out of these 30 loops with which we can compare, only 8 are predicted by our approach using our strict acceptance threshold. This shows that, as designed, our approach is making far more false negative predictions than false positives. This provides confidence in the approach and allows the testing of the predictions and their use as hypothesis generators to annotate the miRNAs.

3.7.6 Experimental validation of the model

We selected three additionnal predicted loops involving TFs with known roles in hematopoiesis to test the binding of the transcription factor to the predicted binding site on the promoter of each of the miRNA. Of these loops, two involve C/EBPβ, and one involves GATA1.

3.7.6.1 C/EBPβ binds the promoter of both miR-155 and miR-212

Using our algorithm, we predicted a regulation loop between miR-155 and C/EPBβ and between this TF and miR-212. CEPBβ is a bZIP transcription factor that binds DNA as a homodimer or heterodimer with other members of the C/EBP family.

It is required for the generation of B cells in the bone marrow (Chen, Liu et al. 1997). The oncogenic role of miR-155 is well established (Tam and Dahlberg 2006, Lawrie 2007) and it is expressed in various lymphomas (Eis, Tam et al. 2005, Kluiver, Poppema et al. 2005, Kluiver, Haralambieva et al. 2006, Kluiver, van den Berg et al. 2006, Ramkissoon, Mainwaring et al. 2006), in particular Burkitt's lymphomas (Calin, Liu et al. 2004, Calin, Sevignani et al. 2004, Metzler, Wilda et al. 2004, He, Thomson et al. 2005, Ramkissoon, Mainwaring et al. 2006) and in some solid tumors (Iorio, Ferracin et al. 2005, Roldo, Missiaglia et al. 2006, Volinia, Calin et al. 2006, Lee, Gusev et al. 2007) making it an interesting candidate to test our predictions. miR-212 is part of a potential miRNA cluster with miR-132. It is highly expressed in most terminally differentiated cells of the hematopoietic system and at low levels in progenitor cells (Petriv, Kuchenbauer et al. 2010). It is also known to be expressed and play a role in the brain (Olsen, Klausen et al. 2009) and it is downregulated in gastric cancer (Wada, Akiyama et al. 2010).

We predicted two binding sites on the promoter of, respectively, miR-155 and miR-212 for the transcription factor C/EBPβ (Figure A1-2a). The ChIP using an anti-C/EBPβ antibody showed that C/EPBβ binds these 4 predicted sites and not the *c-kit* promoter, our negative control, compared to the positive control, the promoter of *IL1β* (Tsukada, Saito et al. 1994). This confirmed the prediction that C/EPBβ binds the promoters of miR-155 and miR-212 and supports the existence of the predicted regulation loops. Furthermore, it was recently shown that the overexpression of miR-155 leads to a decrease in C/EBPβ levels (Selbach, Schwanhausser et al. 2008) thus strongly supporting this predicted regulation loop.

3.7.6.2 Gata1 binds the promoter of miR-489

MiR-489 is located within an intron of the gene *CALCR*, 3kb downstream of the start site of the shortest isoform of *CALCR* (in human, 41kb in mouse) and 90kb (42kb in mouse) downstream of its longest isoform. It thus has two potential promoter regions, its immediate promoter (5kb directly upstream of the miRNA start site) and

the CALCR promoter. In each of these regions, there are three potential binding sites for GATA1 (Figure A1-2b). The ChIP using an anti-GATA1 antibody showed that GATA1 binds the six predicted sites, compared to the negative control, the c-KIT promoter.

3.7.7 Cell culture

TF1 cells were cultured as described previously (Krosl, He et al. 1998). K562, HL-60, Kg1a, U937, and NFS-60 cells were maintained in IMDM (GIBCO Invitrogen Corporation, Burlington, Ontario, Canada) supplemented with 10% FCS (GIBCO Invitrogen Corporation, Burlington, Ontario, Canada). The cells were passaged every second day at $1x10^5$/ml. NFS-60 cells were also supplemented with 25μl/ml of Wehi3 supernatant. 32D cells were maintained in IMDM supplemented with 10% FCS and 10ng/ml of IL-3. They were passaged every second day at 2.5×10^4/ml. The Mel cells were maintained in DMEM (GIBCO Invitrogen Corporation, Burlington, Ontario, Canada) supplemented with 10% FCS and passaged every second day at 5×10^4/ml.

3.7.8 RNA analysis, chromatin immuno-precipitation and real-time PCR

Total RNA was extracted using miRNeasy RNA extraction kit (Qiagen) following manufacturer's protocol. mRNA reverse transcription reaction was primed with random oligonucleotides and dT oligonucleotides. miRNA reverse transcription was performed with Exiqon's Universal cDNA synthesis kit. Real-time PCR was performed using PerfeCTa SYBR Green FastMix from Quanta on a StepOnePlus apparatus (Applied Biosystems) using the primers provided as supplemental information. ChIP assays were performed as previously described (Lécuyer, Herblot et al. 2002) and the real-time PCRs were performed as described using the primers provided as supplemental information.

3.7.9 Infections and Western Blotting

The LMO2 expression vector was described previously (Lécuyer, Herblot et al. 2002). The pre-hsa-miR-223 and pre-hsa-miR-363 cDNA were cloned into the murine-stem-cell-virus (MSCV)-GFP vector using XhoI and EcoRI. TF1, Jurkat and GP+E cells were infected as previously described (Lahlil, Lecuyer et al. 2004) for 48 hours and GFP sorted on a BD FacsAria flow cytometry apparatus 7 days post infection. Fetal liver cells were harvested and infected as previously described (Lecuyer, Lariviere et al. 2007).

LMO2 antibody was purchased from R&D systems (cat# AF2726). Lamin antibody was purchased from Santa Cruz (cat# sc-6217).

3.7.10 LMO2 Intracellular staining

Cells were fixed and washed with BD Cytofix/Cytoperm and Perm/Wash buffer according to manufacturer's instruction and stained with LMO2 antibody (goat - previously described) in staining buffer (phosphate-buffered saline (PBS) with 2% FCS) followed with secondary staining with anti-goat-Cy3 antibody in staining buffer. The gating strategy used to isolate each population is schematically described in Figure A1-3. The quantification of LMO2 based on the flow cytometry results is done as follows:

$$\frac{\left(median(LMO2)_{pop}\right)-\left(median(LMO2)_{pop,unstainedMice}\right)}{\left(median(LMO2)_{allCells}\right)-\left(median(LMO2)_{allCells,unstainedMice}\right)}$$

4 Article 2: Predicting and modelling the microRNA targetome

en préparation pour publication

4.1 Contribution des co-auteurs

VL: design expérimental, réalisation des expériences, analyse des résultats, rédaction du manuscrit.

NW: design expérimental, réalisation des expériences, analyse des résultats, rédaction du manuscrit.

FM: design expérimental, analyse des résultats, rédaction du manuscrit.

4.2 Mise en situation

La première des caractéristiques identifiées comme étant un déterminant d'une cible d'un miARN est la complémentarité de base du miARN avec sa cible potentielle. Cette caractéristique a ensuite été modifiée pour n'inclure que les nucléotides de la région amorce du miARN soit les nucléotides des positions deux à huit du miARN. Les outils de prédiction de cibles de miARN ont été mis au point en utilisant cette caractéristique. D'autres caractéristiques tel que la conservation de la région à travers un nombre d'espèce variable, le contexte de la séquence (le positionnement de la cible potentielle dans le 3'UTR de l'ARNm), la composition en AU de la région, ont été ajoutés afin de limiter le nombre de prédictions. Cependant, malgré ces filtres des prédictions, les outils génèrent encore trop de cibles potentielles pour chaque miARN pour qu'elles puissent toute être testées facilement. De plus les prédictions des divers outils ne sont pas concordantes. Il semble également que plusieurs de ces cibles potentielles ne puissent être confirmées expérimentalement. De plus, plusieurs cibles connues et validées expérimentalement ne sont pas prédites par ces outils. Cet état de fait nous a mené à émettre l'hypothèse qu'au moins un des facteurs permettant d'identifier une cible réelle d'un miARN d'une non-cible était manquant.

Nous avons tout d'abord émis l'hypothèse que la structure locale de l'ARNm pouvait empêcher ou favoriser l'appariement du miARN sur une cible potentielle. Cependant,

aucun de nos résultats expérimentaux ne supportaient cette hypothèse. Pendant ce temps, une étude faite par apprentissage machine arrive à la conclusion que la structure est le facteur le moins important des nombreux facteurs que les auteurs ont testés (Reczko, Maragkakis et al. 2011). Nous avons donc éliminé cette hypothèse.

La deuxième hypothèse testée était que la structure formée lors de l'appariement du miARN sur sa cible potentielle devait avoir certaines caractéristiques spécifiques pour mener à un effet observable. Cette hypothèse était également supportée par les travaux d'un autre groupe qui avait observé un effet de la structure sur l'efficacité d'un miARN (Long, Lee et al. 2007). Cependant, cette hypothèse n'a pu être vérifiée sur des jeux de données plus exhaustifs et a donc été également éliminée.

La troisième hypothèse que nous avons testée est celle qui est présentée dans le manuscrit qui suit. Nous avons émis l'hypothèse qu'il était possible de modéliser les interactions entre les miARN et les ARNm et que cette modélisation permettrait de prédire les cibles des miARN. Pour modéliser ces interactions, nous avons considéré tous les appariements possibles des régions amorces de chaque miARN sur chacun des ARNm. Nous avons également considéré les quantités de chaque espèce d'ARN. Nous avons utilisé un algorithme d'appariement simple pour modéliser l'interaction d'un miARN sur un ARNm dans un contexte donné. Le détail de la méthode informatique utilisée pour le développement et la validation de l'algorithme ainsi que les résultats obtenus sont présentés dans ce chapitre.

Predicting and modelling the microTargetome

Véronique Lisi, Nathanael Weill, François Major

Institute for Research in Immunology and Cancer, Department of Computer Science and Operations Research, Université de Montréal, Montréal, Québec H3C 3J7, Canada

Keywords: microRNA, target prediction, formal model, modeling

4.3 Abstract

Regenerative and personalised medicine promise to replace or fix damaged cells and tissues in patients. This requires pluripotent cells and recent studies indicated that we can restore the original stemness of mammalian cells by regulating their genes. However, we partially know how cells regulate gene expression, which limits our ability to rationally intervene in cellular programs in pathological conditions. Here, we introduce miRBooking, an algorithm to determine *a priori* post-transcriptional events involving microRNAs. MiRBooking establishes the cross hybridisation of microRNAs with messenger RNAs (the microtargetome) in specific cellular contexts described by their RNA stoichiometry, which we realised was the missing link for accurate predictions. We show that the microtargetome provides adequate information to predict gene levels in various cellular identities, including small RNA off-targets. Applications of miRBooking include identifying yet unsuspected genes involved in cellular functions, as well as designing RNA-based strategies to induce specific cellular programs.

4.4 Introduction

The development of multicellular organisms moves forward from embryonic pluripotent stem cells to terminally differentiated cells, such as blood cells, neurones, fibroblasts, etc. Cellular diversity is achieved, at least in part, through alterable gene expression systems that use microRNAs (miRNAs) (Ambros 2004, Bartel 2004, He and Hannon 2004). Although human differentiated cells do not largely exert tissue generation post embryogenesis, their plasticity has been established using various mixtures of protein or miRNA factors. Importantly, human cells have been reprogrammed to a pluripotent state (Takahashi and Yamanaka 2006, Anokye-Danso, Trivedi et al. 2011), and programmed to a specialised state from either already terminally differentiated cells (Yoo, Sun et al. 2011, Liu, Li et al. 2012) or pluripotent states (Wichterle, Lieberam et al. 2002).

The design of RNA-based interventions to control cellular identities requires a profound understanding of gene regulation mechanisms. In particular, knowing which miRNAs target which genes and how they mediate translational repression are crucial. However, miRNA target identification is experimentally difficult and currently depends on computational approaches that have low predictive accuracies (Alexiou, Maragkakis et al. 2009). One of the problems comes from the short sequence complementarity between miRNA seeds (nucleotides 2 to 8) and their messenger RNA (mRNA) target sites (Bartel 2009), thereby creating a large number of potential targets. Despite many efforts to identify factors that could be used to filter these potential targets, how miRNAs and their targets find and bind to each other, and how such interactions interfere with mRNA translation remain central questions.

Recently, it was shown that the abundance of a target mRNA is a key factor in determining the activity of miRNAs and other small RNAs (Ebert, Neilson et al. 2007, Arvey, Larsson et al. 2010, Mukherji, Ebert et al. 2011, Tay, Kats et al. 2011). Consistent with this view, among a series of factors we tested, we found miRNA abundance to be the best to identify experimentally validated miRNAs targeting *PTEN*

among all miRNAs expressed in the DU145 cell line (Tay, Kats et al. 2011) (AUC > 0.9) (see Methods). With miRNA and mRNA abundance playing such an important role in targeting, we considered incorporating this parameter along with others to determine the microtargetome of a cell.

A biological consequence of this is that there is no unique microtargetome that fits all cellular identities, but rather as many as there are cellular identities. The miRNAs that target a given set of genes in a given cellular context can target a different set in a different one, solely based on differential individual mRNA and miRNA expression levels. Thus, microtargetomes cannot be computed using a simple miRNA to mRNA target relationship. We developed miRBooking, a matching algorithm that simulates miRNA-mRNA hybridisation by taking into account absolute RNA abundance and seed hybridisation probabilities. MiRBooking establishes a snapshot of which miRNAs occupy which mRNAs for a given absolute RNA quantification. As RNA quantities in the cell are dynamic, the microtargetome changes accordingly. MiRBooking needs to be invoked each time a change in absolute quantities occurs. We will show below how microtargetome comparative analysis can be made using miRBooking to predict the outcomes of routinely used experiments that modify the microtargetome, such as gene knockout and small RNA knockdown assays, and reporter genes.

4.5 Results

4.5.1 Modelling the microtargetome

Computing microtargetomes requires miRNA-mRNA seed hybridisation probabilities and a cellular context described by its RNA content. We matched each miRNA seed described in miRBase (Griffiths-Jones, Saini et al. 2008) to every positions in the mRNAs described in the RefSeq repository (Pruitt, Tatusova et al. 2009), and kept those with at most one mismatch, defining over 300,000,000 MiRNA Recognition Elements (MREs) (see Methods). We obtained absolute RNA abundance of cell lines from real-time PCR and microarray expression data (Yun, Heisler et al. 2006, Bissels, Wild et al. 2009) (see Methods).

For a given cell line, we considered the MREs for which both the miRNA and mRNA were expressed. Each considered MRE defines a number of instances that is equal to the absolute abundance of its corresponding mRNA. The purpose of miRBooking is to simulate the hybridisation of miRNAs with mRNAs, i.e. to "book" miRNAs into MRE instances.

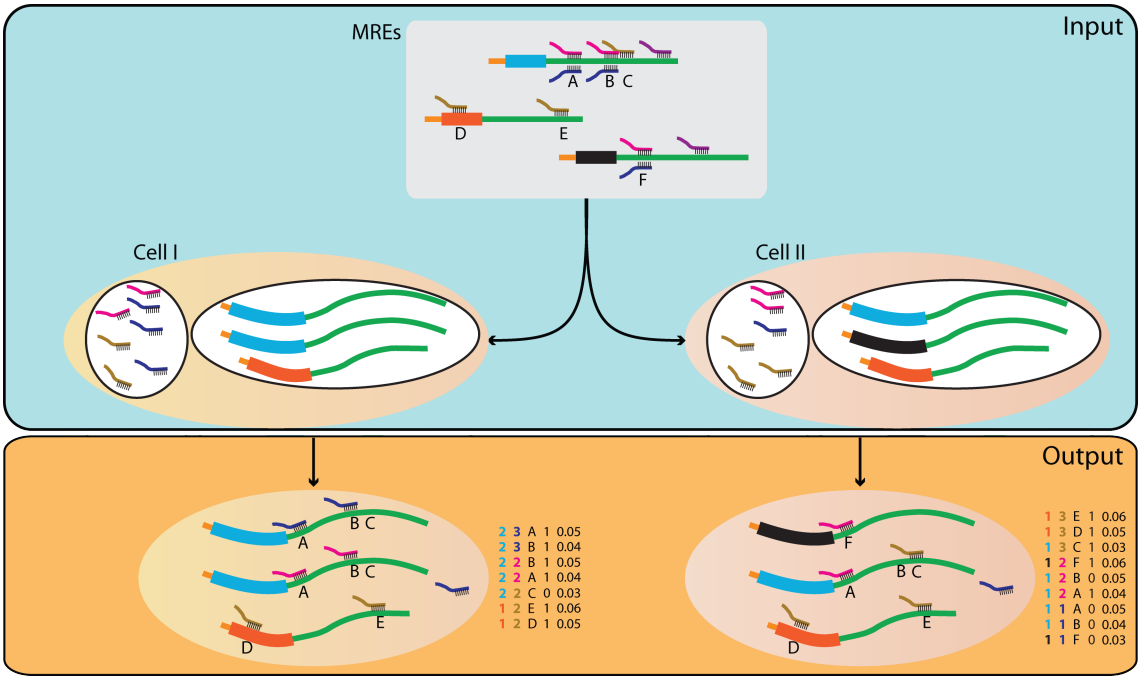
MiRNAs act like enzymes and affect the expression of genes proportionally to their quantities (Arvey, Larsson et al. 2010), i.e. not all matching resolves the miRNA targeting problem. Rather, we must optimise the matches proportionally to RNA abundance and hybridisation probabilities. This is accomplished by assigning the MREs in sorted order of mRNA quantities, miRNA quantities, and hybridisation probabilities. At the beginning of the algorithm, all MRE instances are free. Then, for each MRE in the sorted list, we book a number of free MRE instances proportionally to:

$$\mathrm{q}(mRNA) \times \log_{\alpha}\left(\mathrm{q}(miRNA)\right) \times hp \ ,$$

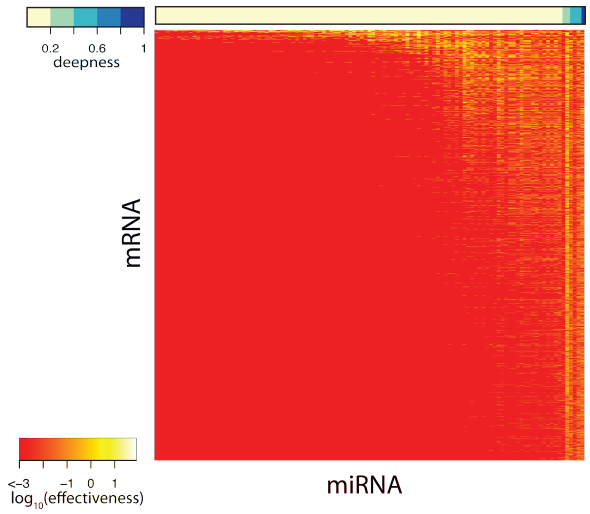
where *q(mRNA)* is the initial mRNA quantity, *q(miRNA)* is the current miRNA quantity, and *hp* is the hybridisation probability between miRNA and mRNA (see Methods).

Figure 4-1 schematises how miRBooking computes the microtargetomes of two different cells that contain different absolute RNA abundance. This quantification of the microtargetome defines the deepness and effectiveness of miRNAs. The deepness is the percentile of the lowest-abundant mRNA occupied by a miRNA. The effectiveness is the number of occupied MRE instances by a miRNA on an mRNA weighted by their hybridisation probabilities (i.e. affinities of the occupied MRE instances). The typical output of miRBooking can be visualised using the deepness and effectiveness values (Figure 4-1b). Although almost all mRNAs are occupied by miRNAs (indicated by the red colour), only a small fraction are effectively targeted (indicated by the yellow colour). Not surprisingly, the miRNAs targeting the deepest mRNAs are the most effective.

a



b



c

```
    hp=0.06
   GAUGGACGUGACAUUCGUGAAAC-5' miR-17-5p
                  |||||||
5'-GGAUUAAUAAAGAUGGCACUUUC PTEN

    hp=0.02
   UCGGAUAGGACCUAAUGAACUU-5' miR-26a-5p
                 .||||||
5'-AUGGAUUCGACUUAGACUUGAC PTEN

    hp=0.01
   AGUCAAAACGUAUCUAAACGUGU-5' miR-19a-3p
                  |||||.|
5'-UGUCACUGCUUGUUGUUUGCGCA PTEN

    hp=0.01
   GAUGGACGUGCUUGUCGUGAAAC-5' miR-93-5p
                  |||||.|
5'-UUAUAAUGGGCUUUUGCACUGUU PTEN
```

Figure 4-1 MiRBooking algorithm.

a, MirBooking applied to two different cells leads to two different microtargetomes. The input is constituted of the RefSeq MRE hybridisation probabilities and the RNA abundance of the cell. Cell I is defined by two copies of mRNA-cyan and one copy of an mRNA-orange, and two copies of miR-pink, three copies of miR-dblue, and two copies of miR-brown; cell II by one copy of mRNA-cyan, one copy of mRNA-orange and one copy of mRNA-black, and two copies of miR-pink, one copy of a miR-dblue, and three copies of miR-brown. The MREs are sorted by mRNA and miRNA absolute abundance, and hybridisation probabilities. The MRE sorted lists for each cell are shown to the right of the output microtargetomes. Each MRE is defined by a line of numbers (from left to right): the absolute abundance of mRNA, the absolute number of miRNA, a MRE identifier, the number of miRNA assigned in the microtargetome, and the MRE hybridisation probability. Consider the first line of cell I: 2 3 A 1 0.05. There are 2 copies of mRNA-cyan and three copies of miR-dblue, and these are the most abundant RNAs. The hybridisation probability of MRE A is 0.05, the highest probability for miR-dblue on mRNA-cyan. Therefore, one copy of miR-dblue is assigned to MRE A. Compared to cell II, the first assigned MRE is E, because in cell II miR-brown is the most abundant. The miRNAs are booked following the order of the MRE sorted list proportionally to RNA absolute abundance and hybridisation probabilities. In the microtargetome of cell I, one copy of mRNA-cyan was assigned two copies of miR-dblue and the other copy was assigned two copies of miR-pink, and the mRNA-orange was assigned two miR-brown, including one in the CDS region. **b**, DU145 microtargetome at a glance. Log10(effectiveness) of the top 20% effective miRNAs on the top 20% targeted mRNAs in DU145 cells (red to white). Deepness of each of the top 20% effective miRNAs (white to blue). **c**, Examples of a 8mer and three 6mer seed matching sites of a few miRNAs predicted to hybridise at various locations of PTEN when overexpressed in DU145 cells. Each of these miRNAs hybridises to PTEN at multiple sites, whose hybridisation probabilities vary from 0.06 to 0.01 for miR-17-5p; 0.04 to 0.02 for miR-26a-5p; 0.05 to 0.01 for miR-19a-3p; and, 0.06 to 0.01 for miR-93-5p. The nucleotides involved in the seed region of the duplex are shown in bold. The Watson-Crick base pairs are shown with vertical lines. The mismatches are shown with dots.

We computed the microtargetomes of the DU145 and PC3 human prostate cancer cell lines. In these cell lines, the microtargetomes were defined by an average of about 100,000 occupied MRE instances, and we observed increased fractions of more than 5% of MREs in the 3'UTR regions compared to the RefSeq site distribution (Figure A2-1b). This bias was also observed in the results of Ago1 HITS-CLIP experiments performed in P13 mouse brain cells (Chi, Zang et al. 2009).

The microtargetome of DU145 cells included miRNAs covering a wide range of hybridisation probabilities. The best seed matching probabilities correspond to 8mer sites (Bartel's nomenclature (Bartel 2009)), whereas the lowest seed probabilities correspond to variations of 6mer sites that contain one mismatch (Figure 4-1c). RNA duplexes often include mismatches which contribute stabilising energies by forming various hydrogen bonding patterns (Leontis and Westhof 2001, Lerman, Kennedy et al. 2011). The seed hybridisation probabilities we used in miRBooking were derived from the MC-Fold RNA secondary structure prediction software, and thus account for these energetic contributions (Parisien and Major 2008).

4.5.2 Simulating overexpression experiments

MiRNAs and proteins are critical in cell differentiation and diseases through modulations of their expression levels (Lee, Feinbaum et al. 1993, Lu, Getz et al. 2005). Cell biologists study the role of genes by altering their expression levels and observing the effects of such on a specific function. We can simulate these experiments and analyse their effects on the microtargetome using miRBooking. First, we compute the microtargetome in a reference cellular context. Second, we change the abundance of the studied miRNA or mRNA to that of the modified context. Finally, we compute the modified microtargetome, which we can compare with the reference.

Gene downregulation by the overexpression of miRNAs has been extensively studied (Sethupathy, Corda et al. 2006, Baek, Villén et al. 2008, Selbach, Schwanhäusser et al. 2008). However, it was recently shown that the modulation of a gene expression was caused by the variable abundance of competitive endogenous RNAs (ceRNAs), i.e. mRNAs and long intergenic non-coding RNAs (lincRNAs) sharing the same miRNAs (Cesana, Cacchiarelli et al. 2011, Song, Carracedo et al. 2011, Tay, Kats et al. 2011). These two phenomena relate to the microtargetome, and their outcome is predicted by miRBooking.

Pandolfi and co-workers identified eight miRNAs directly targeting a transfected *PTEN* reporter in DU145 cells by RNA Immunoprecipitation followed by RT-qPCR: miR-17-5p, miR-19a, miR-19b, miR-20a, miR-26a, miR-93, miR-106a and miR-106b, and they used miR-191 as a true negative (Tay, Kats et al. 2011). They also identified three ceRNAs which overexpressions increased PTEN levels as measured by Western Blot quantification: *CNOT6L*, *VAPA*, *SERINC1* (Tay, Kats et al. 2011). We used these data to calibrate miRBooking (see Methods).

Next, we wished to determine if miRBooking could predict changes in gene expression levels directly from microtargetome comparative analysis. If we assume the contribution of a MRE to a gene's repression is proportional to its hybridisation probability (affinity), then we can estimate its expression level after an RNA

abundance fluctuation by the ratio of the number of occupied MREs in the reference divided by that in the modified context, each weighted by their hybridisation probabilities:

$$foldchange_{reference \to modified}(gene) = \frac{\sum_{MREs \in reference} \text{occupancy}(gene) \times hp}{\sum_{MREs \in modified} \text{occupancy}(gene) \times hp},$$

where occupancy(*gene*) is the total number of booked MRE instances on *gene*.

We compared the predicted sums of occupancies (cooperative effectiveness) in DU145 cells in which we incorporated two synthetic reporters with respectively 0 or 1 MRE for miR-20a-5p. This virtual experiment mimics that performed by Mukherji and co-workers (Mukherji, Ebert et al. 2011). They created two reporters: mCherry with a 3'UTR containing 0 or 1 binding site for miR-20a-5p; and, eYFP. They used quantitative fluorescence microscopy to compare the effect of miR-20-5p on both reporters and observed a gene expression threshold. We observe a similar threshold (Figure A2-2).

We overexpressed in turn *CNOT6L*, *VAPA*, *SERINC1* and compared the miRBooking computed fold change of PTEN with the experimental relative expression (Figure 4-2a). The upregulation of PTEN for all three overexpressed ceRNAs correlate. The relative expression measured experimentally and the computed fold change indicate that miRBooking captures PTEN's behaviour with sufficient accuracy to mimic the microtargetome logic.

As we expected a complex relation between miRNA occupancies and protein output, we were surprised to also observe a high correlation between the predicted fold change and a series of Western blotting measurements of the PTEN and VAPA proteins after overexpressions of eight miRNAs shown by Pandolfi and co-workers to target both (Tay, Kats et al. 2011) (R^2 over 0.7) (Figure 4-2b). This observation corroborates with some previous attempts to link miRNA occupancy with mRNA repression and protein output (Baek, Villén et al. 2008, Guo, Ingolia et al. 2010,

Mukherji, Ebert et al. 2011). Moreover, this formulation of the microtargetome indicates that any context can provide a buffering activity for some genes, those that have low affinities for the miRNAs expressed in that context.

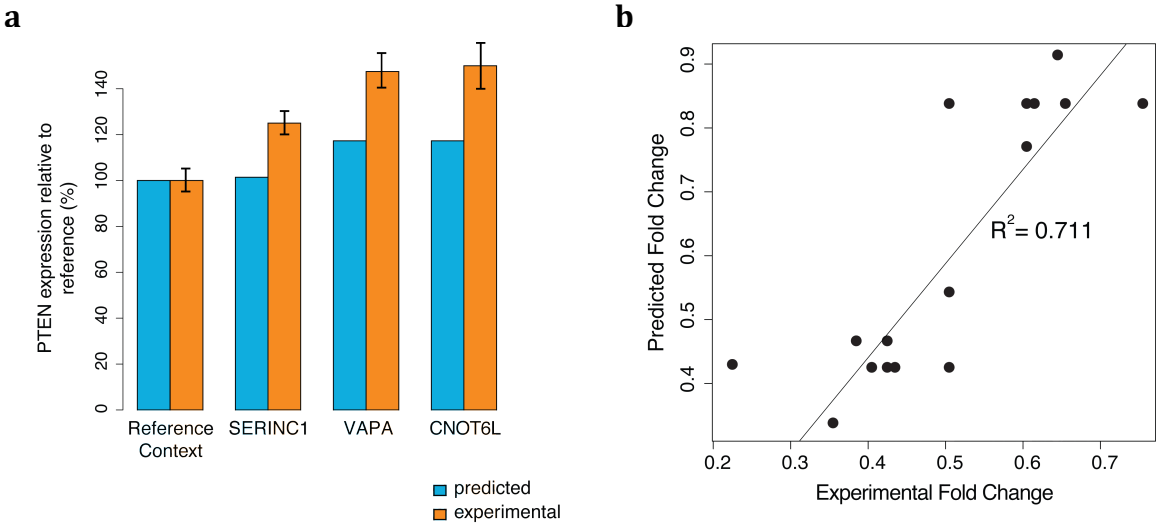


Figure 4-2 MiRBooking calibration and validation.
a, The PTEN protein expression in a reference context and in response to the overexpressions of three ceRNAs as measured by Western Blot quantification and reported by Pandolfi and co-workers (orange) vs. miRBooking computed fold change (cyan). For VAPA and CNOT6L, we averaged the two reported relative expressions as two different constructs were used to express each a portion of ceRNA 3'UTRs. **b**, PTEN and VAPA protein expressions as measured by Western Blot quantification and reported by Pandolfi and co-workers in response to overexpressions of eight miRNAs they confirmed to target PTEN when overexpressed in DU145 cells vs. miRBooking computed fold change.

4.5.3 Estimating disturbance

Off-target effects of designed small RNAs hinders their use in functional studies and as therapeutics (Jackson and Linsley 2010). Off-target effects come from targeting non-intended genes. It is almost impossible to get rid of such because imperfect complementarity is inherent to RNA hybridisation, and it is recognised that designed small RNAs function like miRNAs (Doench, Petersen et al. 2003). Undesired small RNA effects can also come from miRNA dilution, i.e. if the targeted gene is not the most abundant, then the small RNA targets other more abundant genes. It has also been shown that another problem with designed small RNAs is that they are inefficient to repress some genes at some perfect complementary sites. Possible explanations

include a lack of mRNA accessibility due to local structure and the presence of mRNA-binding proteins (Vella, Choi et al. 2004, Vella, Reinert et al. 2004). MiRBooking results rather point to sites that yield weak fold change (due to small Δoccupancies from a reference to a modified context), even though the small RNAs reach their targets.

Comparative analysis of microtargetomes allows us to evaluate the fold change of all genes, and thus to include all effects, at once. The global effects resulting from any RNA abundance modification can be evaluated by comparing the fold change of all mRNAs. To assess the range of such effects, we evaluated the fold change of all mRNAs after the overexpression of each miRNA expressed in DU145 cells (Figure 4-3a). Because gene downregulations corresponding to fold changes over 0.5 ($\log_2$ ratio > -1) may not be significant to a cell, we may be interested only in those equal or below some ratio, e.g. 0.5 ($\log_2$ ratio <= -1).

We defined the ***down-disturbance*(*x*)** resulting from an RNA abundance change by the number of genes for which a fold change of *x* or less is predicted after the change; or similarly the ***up-disturbance(x)*** as the number of genes for which a fold change of *x* or more is predicted after the change. For instance, seven miRNAs shown by Pandolfi and co-workers to target *PTEN* in DU145 are predicted to disturb around 1000 genes, down-disturbance(0.5) ≅ 1000 genes, whereas miR-17 overexpression disturbs over 3000 genes (Figure 4-3a). The range of down-disturbance(0.5) in the DU145 cells goes from 0 (let-7e) to over 4000 genes (miR-1310). Half of the miRNAs expressed in DU145 cells in overexpression simulations disturb the expression of less than 1000 genes. Computing the down-disturbance(0.5) of a small RNA used in a gene-knockdown experiment represents a good approximation of the global effects it can cause to the cells under study.

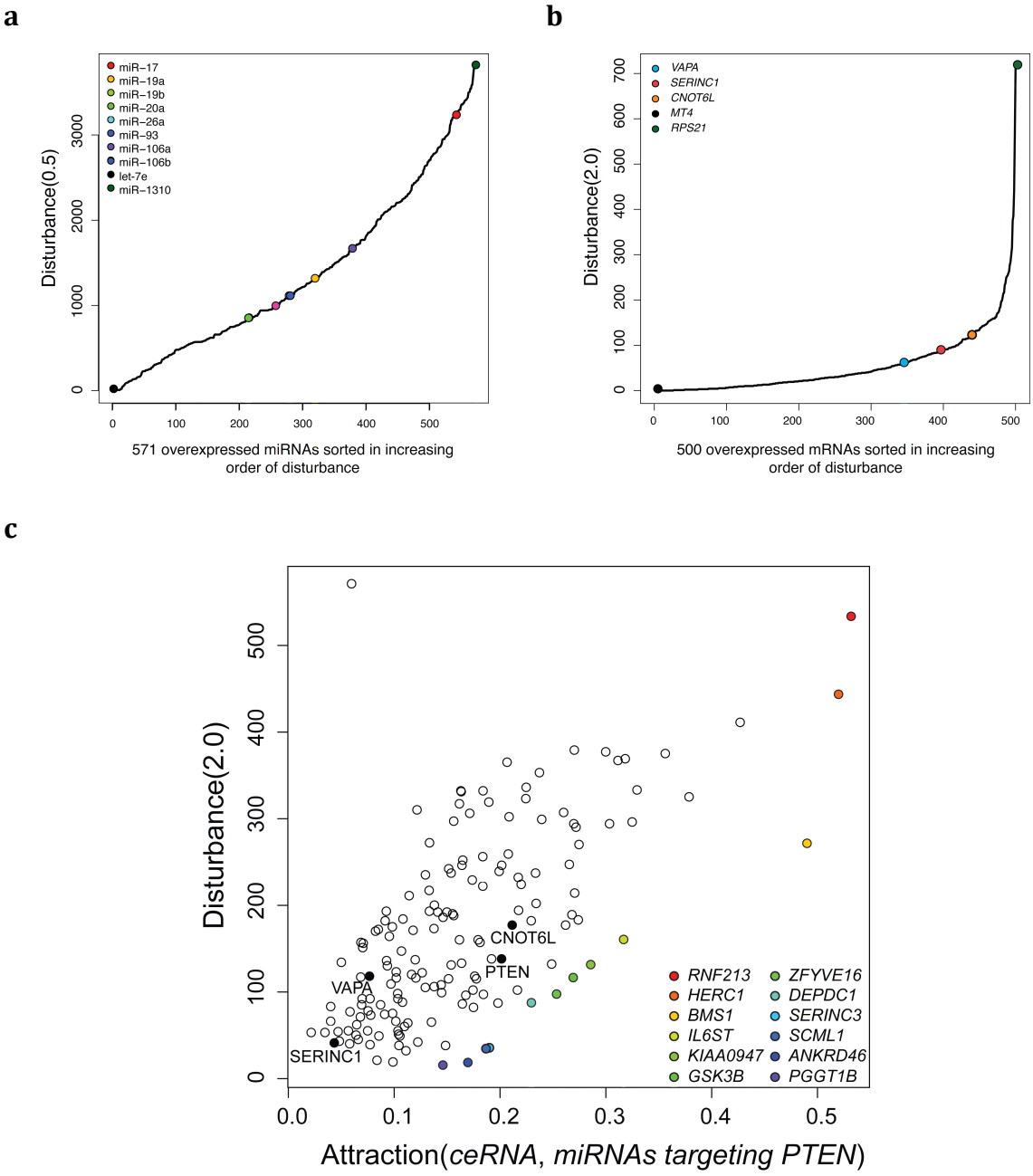


Figure 4-3 Disturbance created by miRNA and mRNA overexpression.

a, The disturbance(0.5) of all genes expressed in DU145 cells when each miRNA was overexpressed. MiR-1310 disturbs the most (over 4000 genes) and let-7e the least (0 gene). The miRNAs in coloured circles correspond to those that were shown experimentally to target *PTEN* when overexpressed in DU145 cells. **b**, The disturbance(2) of all genes when each of 500 mRNAs are overexpressed in DU145 cells. *Rps21* disturbs the most and *Mt4* the least the genes expressed in DU145 cells. The mRNAs shown in coloured circles correspond to those that were shown to compete endogenously with *PTEN*. **c**, Disturbance(2.0) for all miRBooking predicted *PTEN* ceRNAs vs. their attraction for nine miRNAs targeting *PTEN*. The mRNAs identified by coloured circles are optimal in disturbance for their level of attraction. The three *PTEN* ceRNAs *VAPA*, *CNOT6L*, and *SERINC1,* which were experimentally studied by Pandolfi and co-workers, are labelled and shown in black circles.

We determined the up-disturbance of the three ceRNAs of *PTEN* in DU145 cells shown by Pandolfi and co-workers to increase PTEN's levels (Tay, Kats et al. 2011). Similarly to our interest in down-disturbance(0.5), we were interested in significant upregulation, which corresponds to up-disturbance(2.0), or fold change over 2.0 ($\log_2$ ratio > 1). The three ceRNAs of *PTEN* lead to up-disturbance(2.0) of around 100 genes. The range of up-disturbance(2.0) for the mRNAs expressed in the DU145 cells goes from 0 (*MT4*) to over 700 genes (*RPS21*) (Figure 4-3b). Up to 75% of the mRNAs expressed in DU145 cells disturb less than 100 genes, indicating that mRNA overexpression disturbs much less a cell than miRNA overexpression.

4.5.4 Optimising the selection of miRNAs and ceRNAs

The influence of a given miRNA on a gene's miRNA-induced repression can be estimated by:

$$\text{influence}(gene, miRNA) = \frac{\sum_{\text{MREs occupied by miRNA}} occupancy(gene) \times hp}{\sum_{\text{All MREs}} occupancy(gene) \times hp}$$

The ***influence*** of a miRNA on the repression of a gene is the ratio of its occupancy weighted by its hybridisation probabilities over the overall occupancies of all miRNAs on this gene weighted by their hybridisation probabilities. In a simulation of an overexpression of *PTEN* in DU145 cells, 300 copies of *PTEN* were found occupied by 37 miRNAs distributed in 643 MREs, and so *PTEN*'s total occupancy is 643. Among the miRNAs, miR-494 was the one with the most influence, with a little more than 10% of the total influence.

To reduce the expression of a target gene, choosing miRNAs (but also siRNAs and shRNAs) that have a great influence on its repression while having low down-disturbance(0.5) is sensible. Similarly, if we want to increase *PTEN* levels, it would be beneficial to increase the expression of ceRNAs that have great attraction for the miRNAs targeting *PTEN* while having low up-disturbance(2.0). The ***attraction*** of a mRNA for a miRNA is the miRNA occupancy on the mRNA over the sum of its occupancies on all mRNAs:

$$attraction(gene, miRNA) = \frac{occupancy_{miRNA}(gene)}{\sum_{\text{All genes}} occupancy_{miRNA}(gene)}$$

To identify the ceRNAs, we first fix the targeted mRNA to the desired absolute abundance (e.g. *PTEN* increased to the fifteen percentile), and ran miRBooking to identify the miRNAs that target it. The other mRNAs predicted to be targeted by the same miRNAs define the ceRNAs (see Methods). Figure 4-3c shows the attraction of 160 *PTEN* ceRNAs for nine miRNAs predicted by miRBooking to target *PTEN* when expressed in the fifteen percentile, as well as their up-disturbance(2.0) in ceRNA overexpression contexts. The *Pten* ceRNAs that were experimentally studied by Pandolfi and co-workers are *Vapa, CNOT6L,* and *SERINC1,* of which the first two were identified by miRBooking. We nevertheless included *Serinc1*, and found it has one of the lowest disturbance (less than 50 genes), but also a very low attraction for the *PTEN*'s miRNAs identified by miRBooking (less than 5%) (Figure 4-3c). Among other ceRNAs we considered in the simulations, *RNF213* had the maximum attraction for *PTEN*'s miRNAs (over 50%), but is predicted to generate high disturbance (more than 500 genes) (Figure 4-3c). *SERINC3* is predicted to be a good compromise as it has the most attraction for *PTEN*'s miRNAs among the mRNAs that disturb less than 100 genes in the DU145 cells (near 20%).

Selecting ceRNAs that simultaneously maximise the attraction for the miRNAs of a target mRNA and minimise gene disturbance requires the analysis of the RNA content of the cells used. In particular, how many miRNAs will remain on a target mRNA after

the overexpression of selected ceRNAs depends on the attraction of the mRNAs that are more expressed than the target as its level gradually increases. An essential condition to induce an increase of the target gene is that at least one of the shared miRNAs must possess a deepness that reaches the target gene at its endogenous level. The sum of attractions represents a reasonable approximation of this phenomenon.

4.5.5 Identifying functional RNAs

A recent study proposed miR-132 acting as a tumour suppressor of prostate cancer (Formosa, Lena et al. 2012). MiR-132 was shown to be silenced in prostate cancer by DNA methylation. The study showed that increased levels of miR-132 considerably reduced the phenotypes associated to prostate cancer proliferation. The genes *HBEGF* and *TLN2* were suspected to play the role of oncogenes as they were predicted to be direct targets of miR-132. However, neither sole nor a combination of these two gene knockdown reproduced the phenotype observed when miR-132 levels were increased (Formosa, Lena et al. 2012).

Using miRBooking, we analysed the common targets of miR-132 in both the PC3 and DU145 cell lines (see Methods). We found *EEF1A1* among a short list of five candidate genes through which miR-132 could mediate a tumour suppressor activity: *EEF1A1*, *GAPDH*, *PGAM1*, *PPIA*, and *PKM2*. Interestingly, *EEF1A1* is a gene involved in proliferation, invasion, and migration (Zhu, Yan et al. 2009), and its upregulation was recently determined as a hallmark of prostate cancer (Scaggiante, Dapas et al. 2012). We also found PKM2 and PGAM1 in the short list identified by miRBooking, which are two genes that could synergistically work with EEF1A1 to sustain the high rate of glycolysis required in highly proliferating cells (Vander Heiden, Locasale et al. 2010).

4.5.6 Linking microtargetome, microarray, and biological pathway data

Next, we asked to which degree changes in gene expression can be predicted from microtargetome analysis. We wondered how much this information would

contribute to the identification of biological pathways related to the deregulation of specific genes found in modified contexts. It was recently shown using genome-wide gene expression analysis and luciferase reporter assays that the oncogene PNP is directly regulated by two tumour suppressor miRNAs, miR-1 and miR-133a, in the DU145 and PC3 cell lines (Kojima, Chiyomaru et al. 2012).

To mimic the transfections of miR-1 and miR-133a, we predicted the sets of genes that would be downregulated in the DU145 and PC3 cell lines by increasing the absolute abundance of these miRNAs by one order of magnitude compared to the average miRNA abundance. As in the microarray data, *Pnp* and all other genes observed downregulated by both miRNA overexpressions (that were present in our quantifications) were predicted to be downregulated by miRBooking, i.e. eight over the fourteen identified in the microarrays: *TAGLN2*, *ZAK*, *PVRL2*, *C4ORF34*, *PNP*, LASS2, *AHNAK*, and *MMP14*. Furthermore, miRBooking predicted all miR-1 and miR-133a targets, whereas six were not predicted by TargetScan 5.1 (Friedman, Farh et al. 2009): miR-1 and miR-133a targets *ZAK* and *AHNAK*, the miR-1 target *MMP14*, and the miR-133a target *PVRL2*.

This indicates, as suspected, that including RNA abundance and simulating miRNA-mRNA hybridisation increased greatly the predictive accuracy of miRNA target identification (see Methods). The miRBooking algorithm predicts more true positives and less false positives than any other method independently of the value chosen to discriminate between mRNAs that are either affected or non-affected by miRNA overexpressions (Figure A2-2).

We then hypothesised that microtargetome predictions by miRBooking had a better ability to link miRNAs and cellular functions than microarray data. To test this notion, we compared the biological annotations of the genes predicted by miRBooking with those identified in the microarray experiments, and evaluated if these sets belong to specific biological pathways (see Methods). Using the predicted genes, the most significant pathway category was "*cancer*" with Z-scores between 6 and 11,

whereas the Z-scores of the genes identified by microarray data were between 1 and 8 (for the "*cancer*" category; negative controls produced Z-scores between -0.84 and 3.8). The best Z-scores obtained in the case of miR-1 in DU145 cells using the microarray data are over 6 for the pathway category "*replication and repair*". This interpretation of the KEGG pathways suggests that the miRBooking predictions corroborate better with the cancer annotation of miR-1 and miR-133a in the DU145 and PC3 cell lines than the microarray data (Kojima, Chiyomaru et al. 2012) (Table. S1 and S2).

4.6 Discussion

MiRBooking uses a very simple hybridisation system based on seed pairing. A key factor that significantly improved accuracy resides in using a matching algorithm and considering the RNA stoichiometry. The high rates of false positive of most methods reflect a restrained and static view of the microtargetome. This view was probably perpetuated by attempting to learn targeting rules from overexpression assays, which reflect only a single abundance of the RNAs studied, that of their maximum.

It was recently suggested that miRNAs form a large postranscriptionnal regulation program containing about 250,000 interactions (Sumazin, Yang et al. 2011). The miRBooking model suggests that less than half of those, 100,000, suffice to maintain each cellular identity, i.e. less than 1% of the potential network defined by seed complementarity. It would be interesting to interpret this result in the context of the ENCODE project (Dunham, Kundaje et al. 2012).

The measures of disturbance represent good approximations of the overall effects of RNA content modifications. Our model corroborates with accumulating evidences that miRNAs contribute to the robustness of the cell (Ebert and Sharp 2012). A majority of miRNAs in overexpression disturbs a small fraction of all genes. The overexpression of mRNAs induces even smaller disturbance on the microtargetome. MiRNAs ensure that sudden RNA increases, or invasions, are rapidly taken care of by the microtargetome while it maintains the basic gene regulatory function. MiRBooking is one of the first tool that determines *a priori* post-transcriptional effects involving miRNAs, a critical step in the design of RNA-based rational interventions. This approach will contribute to the advancement of strategies used in regenerative medicine.

4.7 Methods

The Methods are fully described in the Supplemental Information. The miRNA target predictions produced by miRBooking in two different cell lines with various miRNA and mRNA overexpressions are available online at www.major.iric.ca/Web/microtargetomes.

4.8 Acknowledgment

We thank Cédric Saule for his help with computing the heptamer energy tables. Stephen Leong for creating the microtargetome repository site. All members of the RNA engineering lab for useful discussions. Victor Ambros, Etienne Gagnon, Gerardo Ferbeyre, and Damien D'Amours for their careful readings of this manuscript and relevant comments. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada (Discovery program), the Canadian Institutes of Health Research (CIHR) (grant number MOP-93679), and the National Institutes of Health (grant number 1R01GM088813-01A1). VL is a recipient of PhD scholarships from the CIHR and the Cole Foundation. NW is a CIHR postdoctoral fellow.

5 Travaux de recherche connexes

5.1 Article 3: A comparative analysis of the triloops in all high-resolution RNA structures reveals sequence structure relationships.

Publié dans RNA vol 13: 1537-1545

2007

5.1.1 Contribution des co-auteurs

VL: Design expérimental, réalisation des expériences, analyse des résultats, rédaction du manuscrit

FM: Design expérimental, analyse des résultats, rédaction du manuscrit

5.1.2 Mise en situation

L'ARN est une molécule très souple et dynamique qui peut adopter une variété de structures dans l'espace. La structure d'un ARN est généralement décomposée en sa structure primaire (sa séquence), sa structure secondaire (l'ensemble des paires de bases) et sa structure tridimensionnelle. La structure tridimensionnelle adoptée par un ARN peut être déterminée expérimentalement par diverses méthodes, les plus usuelles étant la résonnance magnétique nucléaire et la cristallographie et diffraction de rayon-X. Il existe également diverses techniques permettant de modéliser (prédire) les structures possibles d'un ARN, notamment par l'utilisation de contraintes (Major, Turcotte et al. 1991). L'identification de contraintes raisonnables est la principale difficulté inhérente à la modélisation de structure d'ARN par utilisation de contraintes.

5.1.2.1 Éléments structuraux fréquemment adoptés par l'ARN

Les propriétés de l'ARN, contrairement à l'ADN, lui permettent d'adopter une multitude de structures différentes. Ces structures sont habituellement composées de divers éléments structuraux récurrents: double hélice, boucle, renflement ("bulge"). Une double hélice d'ARN est formée lorsqu'un minimum de trois résidus consécutifs s’apparient à un même nombre de résidus consécutifs, sur le même brin d'ARN ou sur un autre brin. Ces appariements impliquent le plus souvent les faces Watson-Crick des résidus mais tout type d'appariement peut permettre la formation d'une double

hélice. Lorsqu'une double hélice est formée d'un même brin d'ARN replié sur lui-même, les nucléotides non appariés qui permettent le changement d'orientation nécessaire à la formation de la double hélice forment alors une boucle. La combinaison de la double hélice et de la boucle est habituellement nommée tige-boucle. Le nombre minimal de nucléotides non appariés pouvant permettre une double hélice terminée par un appariement Watson-Crick est de deux. Lorsqu'un nucléotide dont les voisins font partie d'une double hélice ne s'apparie pas dans cette hélice, ce nucléotide forme alors un renflement.

5.1.2.2 Relation séquence structure de l'ARN

Le nombre de séquences d'ARN biologiquement actives ne cesse d'augmenter, de même que le nombre de structures tridimensionnelles disponible. Cependant, ce dernier augmente à un rythme beaucoup plus lent puisque les méthodes de détermination de la structure n'ont pas le même rendement que, notamment, les séquenceurs à haut débit. Il est donc important de pouvoir prédire rapidement et correctement la structure adoptée par un ARN en ne se basant que sur sa séquence.

Tel que mentionné précédemment, la prédiction de structure de l'ARN par contraintes n'est possible qu'en ayant un grand ensemble de contraintes précises. La prédiction de structure secondaire de l'ARN permet d'obtenir plusieurs de ces contraintes sur les bases formant des tiges. Cependant, aucune contrainte n'est typiquement donnée par la structure secondaire sur les bases formant des boucles autre que la liaison covalente, par l'entremise du squelette, de ces bases. Par ailleurs, les régions boucles de l'ARN sont, par définition, des régions possédant peu d'interactions intramoléculaires en comparaison avec les régions de tige-boucles. Les bases de ces régions sont donc disponibles aux interactions intermoléculaires, que ce soit avec d'autres ARN, de l'ADN ou des protéines. Cependant, tous les nucléotides des régions boucles ne sont pas disponibles pour des interactions intermoléculaires puisque certains de ceux-ci forment des interactions intramoléculaires qui ne sont pas

capturées par les prédictions de structure secondaires. Il est donc primordial de développer des méthodes qui permettent d'identifier ces interactions.

Nous avons émis l'hypothèse qu'en analysant les structures tridimensionnelles déterminées expérimentalement et en se concentrant sur un élément de structure récurrent, il serait possible de développer une méthode permettant de prédire les interactions intra-moléculaires de cet élément de structure. Nous avons développé la méthode en utilisant le motif de "triloop", une boucle formée de cinq nucléotides dont le premier et le dernier s'apparient à l'origine de la boucle. Ce motif a été sélectionné parce qu'il représente le motif le plus simple de boucle. Cette analyse nous a permis de développer une méthode permettant la prédiction de la présence et absence d'interactions intra-moléculaires dans une séquence formant une tige boucle à trois nucléotides. Cette méthode peut être utilisée comme source de contraintes supplémentaires lors de la modélisation tridimensionnelle de structures inconnues.

A comparative analysis of the triloops in all high-resolution RNA structures reveals sequence-structure relationships

Véronique Lisi and François Major

Institute for Research in Immunology and Cancer, Department of Computer Science and Operations Research, Université de Montréal, Montréal, Québec H3C 3J7, Canada

Keywords: RNA, triloop, motif, structure, comparative analysis, three-dimensional modeling

5.1.3 Abstract

Despite an increasing number of experimentally determined RNA structures, the gap between the number of structures and that of RNA families is still growing. To overcome this limitation, efficient and reliable RNA modeling methodologies must be developed. In order to reach this goal, here, we show how triloop sequence–structure relationships have been inferred through a systematic analysis of all triloops found in available high-resolution structures. The structural annotation of all triloops allowed us to define discrete states of the triloop's conformational space, and therefore an explicit sequence-to-structure relation. The sequence–structure relationships inferred from this explicit relation are presented in a convenient modeling table that provides a limited set of possible three-dimensional structures given any triloop sequence. The table is indexed by the two nucleotides that form the triloop's flanking base pair, since they are shown to provide the most information about the triloop three-dimensional structures. We also report the observations in the X-ray crystallographic structures of important conformational variations, which we believe might be the result of RNA dynamic.

5.1.4 Introduction

Single-stranded RNAs fold in three-dimensional (3D) space and adopt a diversity of conformations conferring various biological functions. A recent accumulation of high-resolution X-ray crystallographic structures (Berman, Westbrook et al. 2000) offers an opportunity to study further RNA architecture. In particular, the various ribosomal RNA (rRNA) structures (Wimberly, Varani et al. 1993, Ban, Nissen et al. 2000, Harms, Schluenzen et al. 2001, Nissen, Ippolito et al. 2001, Brodersen, Clemons et al. 2002) show many repeated structural motifs in the context of their hosts. Generally acknowledged and studied motifs include the loop E (Varani, Wimberly et al. 1989, Wimberly, Varani et al. 1993, Szewczak and Moore 1995), the GNRA tetraloop (Jucker and Pardi 1995, Jucker, Heus et al. 1996), pseudoknots (Giedroc, Theimer et al. 2000) , the kink-turn (Klein, Schmeing et al. 2001), the A-minor motif (Nissen, Ippolito et al. 2001), the T-loop (Nagaswamy and Fox 2002), the UNR U-turn (Gutell, Cannone et al. 2000), and the lonepair triloop (Lee, Cannone et al. 2003).

In the meantime, in order to study RNA structures systematically, our laboratory developed a series of computational tools that are compliant with the RNA ontology (Leontis, Altman et al. 2006). Given a set of 3D structures, *MC-Annotate* interprets and labels the RNA base pairing and stacking interactions (Gendron, Lemieux et al. 2001, Lemieux and Major 2002), whereas *MC-Search* determines the locations of user-defined structural motifs (Hoffmann, Mitchell et al. 2003, Olivier, Poirier et al. 2005). Given the large and increasing amount of high-resolution RNA structural data, it is now difficult to design sound and complete motif studies without such systematic tools. For instance, unforeseen examples of the so-called GNRA tetraloop motif do not adhere to the G-N-R-A consensus sequence (Huang, Nagaswamy et al. 2005), or were found in interior loops rather than tetraloops. However, these dissident examples fulfill the same role of stabilizing tertiary interactions between two adjacent stems (Lemieux and Major 2006).

The latter observation was facilitated by a systematic root-mean-square deviation (RMSD) classification of cyclic motifs in the 23S rRNA of *Haloarcula marismortui* (Ban, Nissen et al. 2000). The cyclic motifs were shown to preserve the same base pairing and stacking interactions (Lemieux and Major 2006). Consequently, classifying RNA motifs according to their base interactions or using RMSD is equivalent, and defines a discrete structure space that enables sequence-structure mapping, i.e. sequences can be linked with a number of distinct structures.

Here, taking advantage of the recent X-ray crystallographic structures and of our computational tools, we study the triloop motif in different RNAs, species, and contexts. First, we systematically search for all triloops. Second, we classify them according to the types and positions of their base interactions. Then, we build a discrete and explicit sequence-to-structure relation, which we use to link triloop sequences with base pairing and stacking constraints. Such constraints are useful to narrow down the size of triloops' conformational space in the context of structure prediction and 3D modeling (Major, Turcotte et al. 1991, Massire and Westhof 1998, Jossinet and Westhof 2005, Major and Thibault 2007).

The triloop motif is particularly interesting due to its major role in a variety of organisms and pathways, for instance in the promotion of some virus replication (Huang, Alexandrov et al. 2001, Olsthoorn and Bol 2002), viral synthesis (Haasnoot, Bol et al. 2003), and iron response (McCallum and Pardi 2003). Furthermore, a previous triloop study revealed an important structural diversity (Lee, Cannone et al. 2003), making this motif an ideal candidate for a base interaction characterization. This previous work identified triloops where the only base pair is the closing one, whereas here all triloops are identified regardless of the presence of intra-loop base pairs.

5.1.5 Results

5.1.5.1 Triloop motif

We define the triloop motif as a sequence of five adjacent nucleotides, where the first and last form a base pair in the 3D structure Figure 5-1A. This definition corresponds to the classical view of a triloop: three free nucleotides closed by a base pair. Our choice to accept all possible base-pairing types for the flanking (indicated by the circled 'P' in Figure 5-1A) comes from recent observations showing near 50% of large rRNA base pairs are non-Watson-Crick (Lemieux and Major 2002), and also from the results of a previous study indicating that only a small fraction of lonepair triloops are flanked by canonical Watson-Crick base pairs (Lee, Cannone et al. 2003). The dotted lines in Figure 5-1A indicate that any intra-loop interactions are allowed. This flexibility is necessary to account for potentially different structures for any triloop sequence.

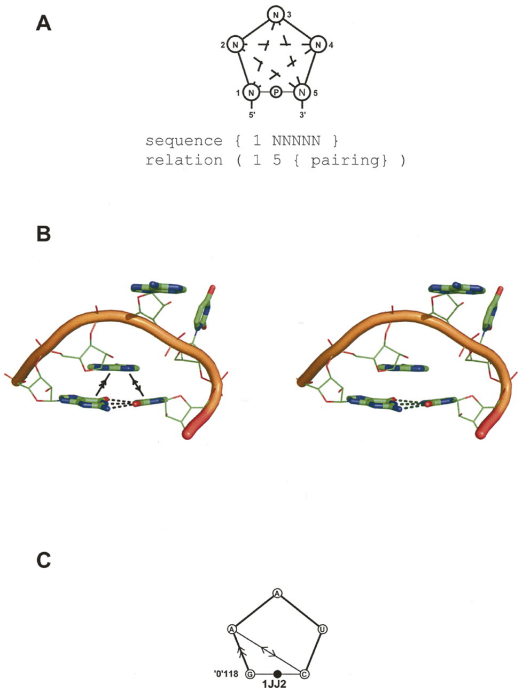


Figure 5-1. Triloop motif.
(*A*) Interaction graph (*above*) and MC-Search input descriptor (*below*). The nucleotides are numbered from 1 to 5 and from 5′ to 3′. The lines indicate nucleotide interactions. The bold lines indicate the presence of phosphodiester linkages (backbone). The dotted lines indicate possible intraloop interactions. Any nucleotide type is accepted, represented by the letter "N" (IUPAC code) at each node. Any flanking base pairing type is tolerated, indicated by the circled letter "P" between nucleotides 1 and 5. (*B*) Three-dimensional structure of a typical triloop in the 23S rRNA of *H. marismortui* (PDB code 1JJ2). This triloop is flanked by a Watson–Crick *cis* base pair. The H-bonds in the base pair are shown by dotted lines. Nucleotide A2 stacks with both nucleotides involved in the flanking base pair, indicated by the arrows. Nucleotides A3 and U4 bulge out of the triloop structure. (*C*) Interaction graph of the triloop instance shown in *B*. The flanking GC Watson–Crick *cis* base pair is represented by the black dot. The arrows indicate base stacking as in *B*. Nucleotide G1 corresponds to the residue identifier 118 in chain "0" of the PDB file 1JJ2.

5.1.5.2 Triloop sequence-structure relationships

Applying the MC-Search program with our definition of the triloop and the high−resolution RNA structures available at the PDB (Berman, Westbrook et al. 2000) results in 922 triloops that were found in 86 different PDB files (for the complete list, see Table A3-I in the on-line Supplementary Materials). Triloops were found in a wide variety of RNA families, including the 5S, 16S and 23S rRNAs with and without antibiotics, transfer RNAs (tRNA), ribonuclease P RNAs, many viruses, riboswitches, group I introns, as well as several RNA aptamers bound to proteins. The

indistinguishable triloops in sequence and *structure* (see Materials & Methods) were grouped together, yielding 104 different *specimens* (or groups). Figure 5-1B shows the 3D structure of a typical example of a 23S rRNA triloop specimen. Figure 5-1C shows the base interaction graph of the typical triloop specimen shown in Figure 5-1B. 55 different triloop structures define 55 such distinct and discrete interaction graphs, in contrast to the continuous RNA 3D space. 60 different triloop sequences fold in these 55 distinct *structures*. The set of triloop interaction graphs can also be seen as a partition of the triloop conformational space, which is convenient to define a straightforward sequence-structure relation, as shown in Figure 5-2.

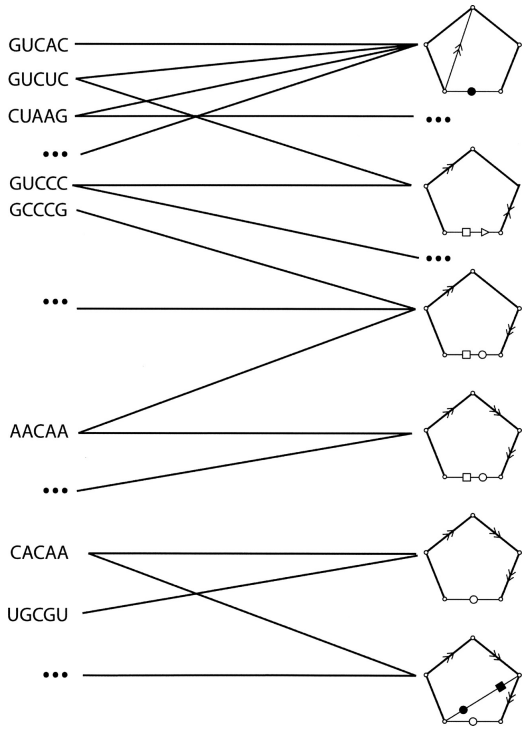


Figure 5-2. Sequence–structure relation.
Only subsets of triloop sequences and structures are shown. The sequences are linked to the structures in which they are found. Three different symbols represent the base-pairing types accordingly to the Leontis and Westhof nomenclature (see Materials and Methods). The base-stacking types are shown using arrowtips, accordingly to the Major and Thibault nomenclature (see Materials and Methods).

5.1.5.3 Triloop dynamic

This explicit sequence–structure relation allows us to study the sequence–structure relation in both ways. First, we can take a fixed sequence and study the different structures it has fold into. Second, although not explored in this work, we can pick a given structure and compare the sequences that were threaded in it. In the context of our studies, the sequence–structure relation reveals the many-to-many relationships between some triloop sequences and structures, i.e., some sequences fold in different structures, and some structures accommodate multiple sequences.

5.1.5.3.1 Fixing the sequence

Interestingly, when we take a close look at given sequences, we observe small structural changes that occur in different crystals of the same RNAs and sites. Three

types of such small structural variations were noticed: (1) stack movement along the backbone, (2) base stacking, and (3) base-pairing formation/disappearance. These three types have been observed in the 32 available X-ray crystal structures of the 23S rRNA of *H. marismortui*.

The first type was detected in a CUAAG triloop found at position 1186, where four different triloop structures are observed (see Figure 5-3A). The base stacking between nucleotides 3 and 4 is present in all 32 X-ray structures, whereas another stack interaction moves along the phosphodiester linkages. It appears between nucleotides 4 and 5 in the PDB 1VQN (Figure 5-3A, second from left), between nucleotides 1 and 2 in 1YJW and 1VQ8 (Figure 5-3A, third and fourth from left). In the latter X-ray crystallographic structure, another base-stacking interaction is observed between nucleotides 4 and 5 (Figure 5-3A, right), also present in 1VQN (Figure 5-3A, second from left).

The second type of triloop dynamic is intraloop base-stacking formation. For instance, three different structures are observed in a CGCGA triloop found at position 218 (Figure 5-3B). No intraloop base stacking is present in 1YIT (Figure 5-3B, left). However, base stacking between nucleotides 3 and 5 appears in 1FFK (Figure 5-3A, middle), and then a second base-stacking interaction between nucleotides 1 and 3 is observed in 1VQ4 (Figure 5-3B, right).

Finally, intraloop base-pairing formation is observed in a triloop at position 138 (Figure 5-3C). The triloop in 1VQ8 and five other 23S rRNAs do not show any intraloop base pairing (Figure 5-3A, left), whereas a Hoogsteen U1-G3 *cis* base pair that participates in the formation of a base triple, involving the flanking base pair, is observed in 1VQM (Figure 5-3C, right) and another 23S rRNA. This site in the 24 other X-ray crystal structures of the 23S rRNA of *H. marismortui* shows pentaloops, not triloops.

This base-pairing formation inside a triloop site of the rRNA is not common, since only four triloop structures among the 55 have this feature. In addition to the

example shown above, a UUAAG triloop is found at position 1966 of an X-ray crystallographic structure of the 23S rRNA of *H. marismortui* (data not shown). In this case, all 32 X-ray crystallographic structures maintain a *cis* Watson–Crick/Hoogsteen U2–A4 base pair. On the other hand, a CACAA triloop is found at position 934 of the X-ray crystallographic structure of the 16S rRNA of *Thermus thermophilus*. In 2J00, a *cis* Watson–Crick/Hoogsteen C1–A4 base pair is observed, whereas it disappears in 1FJG (data not shown). This latter case falls into our third type of triloop dynamic. Finally, the last case is a CCCGG triloop found at position 1028 in the 16S rRNA of *T. thermophilus*, where a *cis* Watson–Crick C2–G5 base pair participates in the formation of a base triple with the flanking (data not shown).

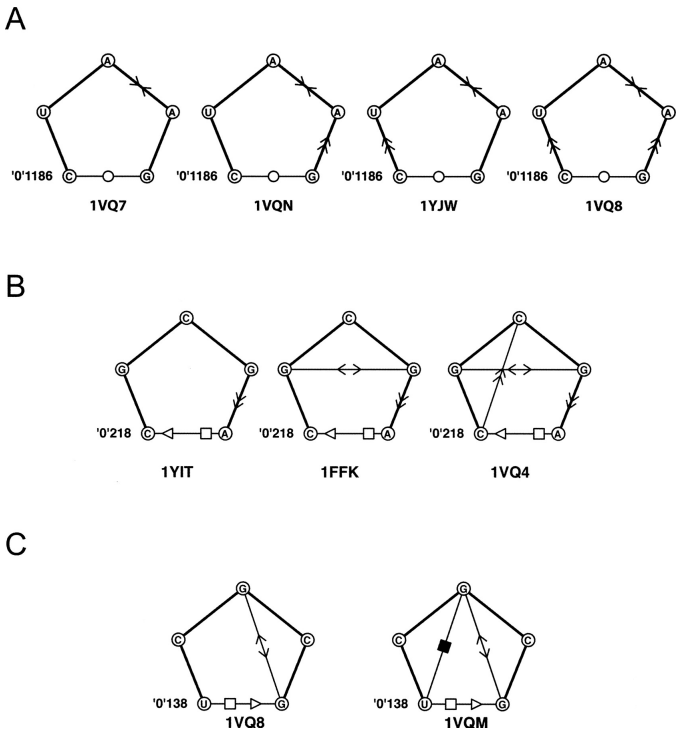


Figure 5-3 Interaction dynamics.
(*A*) Base-stacking interactions along the backbone. Four structures for the same 23S rRNA site ("0"1186) differ by the positions of base-stacking interactions. (*B*) Intraloop stacking interactions. Three structures for the same 23S rRNA site ("0"218) differ by intraloop base-stacking interactions. The leftmost structure has no intraloop interaction. Then, an intraloop base stacking appears between nucleotides 2 and 4. Finally, a second intraloop base-stacking interaction appears between nucleotides 1 and 3. (*C*) Intraloop base-pairing interactions. Two structures for the same 23S rRNA site ("0"138) differ by the appearance of an intraloop Hoogsteen *cis* base-pairing interaction between nucleotides 1 and 3.

5.1.5.3.2 Ligand influcence

The sequence–structure relation can also serve to measure the effect of ligand binding on triloop structures. We studied an interaction among two triloops and a GNRA tetraloop in the X-ray crystal structure of the 16S rRNA of *T. thermophilus* (see Figure 5-4A). In 2J00, the rRNA is bound to paromomycin antibiotics. Noticeably, the second and third nucleotides of, respectively, the triloop capping helix 10 (H10) and helix 17 (H17) (nucleotides 202 and 461) bulge out of the triloop, toward each other's triloop. This junction is significantly different in 1FKA, where the 16S rRNA is not bound to antibiotics (Figure 5-4B). In this case, both nucleotides fold back toward

their respective loops. The triloop capping H17 in 1FKA is not a longer triloop, but a rather large-size loop, in which even the last Watson–Crick base pair in the stem is broken, as shown in the interaction graph of Figure 5-4C. Furthermore, the triloop capping H10 loses the stack between nucleotides 1 and 3. The GNRA tetraloop of helix 15 (H15) switches to a triloop.

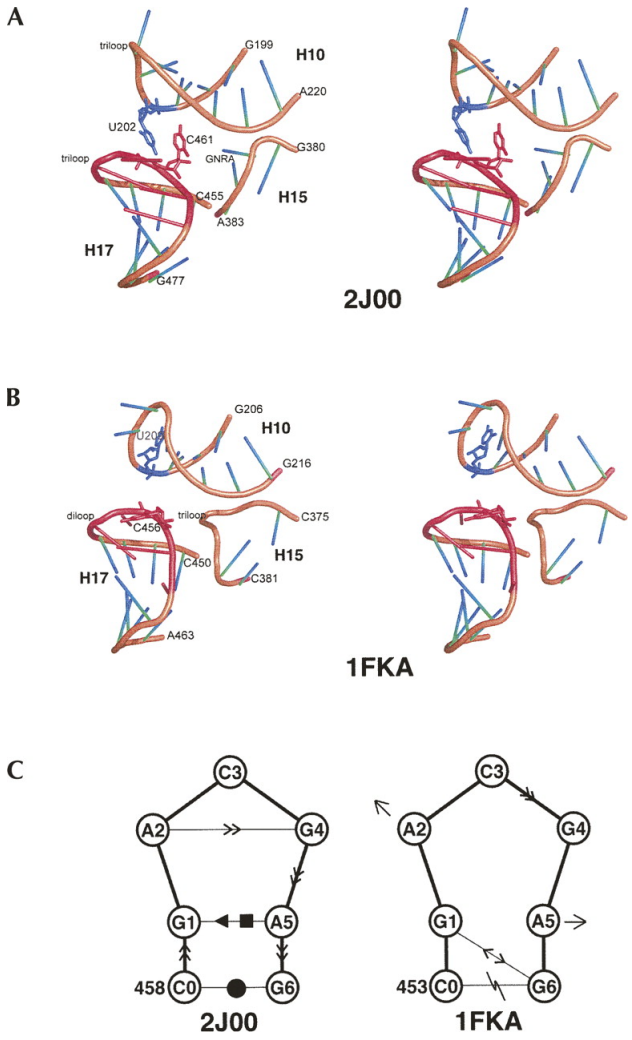


Figure 5-4. Antibiotics and triloop structure.
Two views of a triloop found in the 16S rRNA of *T. thermophilus*. The triloop is represented in a red cartoon, where the third nucleotide of the triloop is shown using sticks. A nucleotide of an adjacent loop is shown in blue, also using sticks. (*A*) Stereoview of the triloop found in PDB 2J00 at position 459 with two neighbor loops. The third nucleotide of the triloop and a nearby one occupy the cavity created by the three RNA loops. (*B*) Stereoview of the triloop found in PDB 1FKA at position 454 (resolution of 3.3 Å). The third nucleotide of the triloop and the nearby one moved toward their respective loops and outside the cavity. (*C*) Structural graphs of the triloop observed in 2J00 (*left*) and in 1FKA (*right*). The triloop and base pair next to the flanking are disrupted by the conformational changes, possibly induced by the presence of antibiotics.

Interestingly, an antibiotic bound X-ray crystal structure of the 23S rRNA of *T. thermophilus* (2J01) exhibits two overlapping triloops at positions 475 and 476 (see Figure 5-5). We were curious about the implication of the antibiotics in this peculiar triloop arrangement. The same overlapping triloops are found at the same site in the

X-ray crystallographic structure of *Escherichia coli* (2AW4) (Figure 5-5B, middle). However, this structure is not bound to antibiotics. We thus inspected other *E. coli* X-ray crystallographic structures and found one bound to kasugamyin antibiotics, 1VS6 (Figure 5-5B left). This one, surprisingly, shows only the triloop at position 476; the triloop at position 475 is lost.

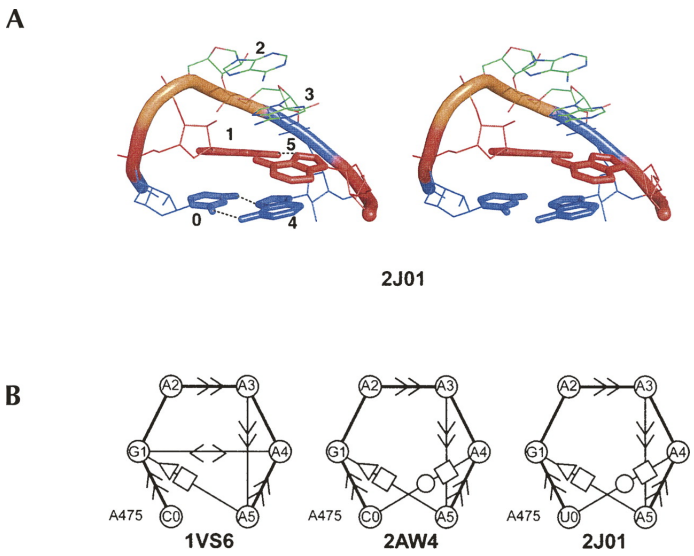


Figure 5-5. Ribosomal overlapping triloops.
(*A*) Stereoview of the six nucleotides forming the overlapping triloops, as found in PDB file 2J01 (*T. thermophilus* with antibiotics). Nucleotides 4 and 5 cross one from each other to achieve the overlapping triloops. (*B*) Structural graphs of the overlapping triloops shown in *A* (*right*), the single triloop of PDB file 1VS6 (*E. coli* with antibiotics) (*left*), and the overlapping triloops in PDB file 2AW4 (*E. coli* without antibiotics) (*middle*).

5.1.5.4 Sequence distributions

Out of the 60 sequences found, 20 are not specific to a single structure, confirming that no trivial, or direct, triloop sequence–structure relationships exist. However, we noticed that the three nucleotides of the loop, nucleotides 2, 3, and 4, display a greater variation than the complete sequence, and most triplets are found in many different triloop structures (data not shown).

We further confirmed this by using an information theory approach (see Materials and Methods). Out of near 2 bits of mutual sequence–structure information available in our data, >1 bit (in fact, 1.13 bits/1.88 bits = >60% of the information) is provided by the flanking base-pair nucleotides (nucleotides 1 and 5). In other words, the flanking base pair is the most informative sequence element of the triloop 3D structure.

We therefore focused our interest on the flanking base pair, and indexed the structural information by it (see Table A3-II), yielding to a modeling table (see Table A3-III). The table returns, from the knowledge of the flanking base-pair sequence of a triloop, the possible types of the flanking base pair and, for each, the possible interactions among the nucleotides of the triloop. Figure 5-6 shows the example of all triloops flanked by a CG base pair (CNNNG). Here, 10 different alternatives are proposed. In general, small numbers of structures (less than 20) are proposed, instead of almost 20,000 theoretical types: three types of interaction for each of the nine possible intraloop nucleotide interactions = 3^9, or 19,683 for any given triloop when no a priori information is available.

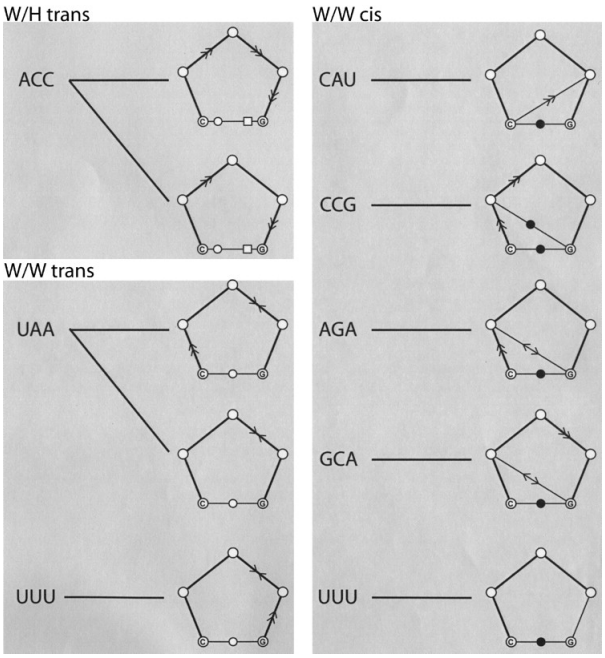


Figure 5-6. Section of the modeling table.
The triloop specimens with a CG flanking base pair are shown using interaction graphs. For each graph, the observed triplet (nucleotides 2–4) sequences are given. Three groups are defined by the flanking base-pair types: W/H *trans*, W/W *trans*, and W/W *cis*.

5.1.6 Discussion

5.1.6.1 Systematic search

The large number of triloops found in this study indicates the increasing importance of using a systematic tertiary structure approach to search for RNA motifs. As an example, here, MC-Search and MC-Annotate helped us find nine additional triloops in the 23S rRNA subunit of *H. Marismortui*, taking to 22 the total number. This represents an increase of about ~40% in comparison to the previous study (Lee, Cannone et al. 2003) (see Figure A3-1 in the online Supplemental Materials).

5.1.6.2 Nucleotide dynamic

Fourteen triloop sites fold in more than one structure. A comparative analysis of these triloop sites revealed three types of nucleotide interaction interplay: (1) a movement of base-stacking and base-pairing interactions along the backbone path, (2) the appearance/disappearance of intraloop base-stacking and base-pairing interactions, and (3) changes in interaction types. Among these, more stability is observed in the base-pairing types than in any other nucleotide interplay. Interestingly, these observations correspond to regions in the X-ray crystallographic structures that, if the RNA was static rather than dynamic, should have folded similarly. They can indeed be the effect of a competition between different folds, of different folding pathways that depend on the environment and conditions, or of an oscillation among several possible conformations, possibly due to nucleotide dynamics, which would occur until a late folding step or when a ligand approaches.

5.1.6.3 Three-dimensional modeling

The analysis of the counterpart of the 16S rRNA triloop at position A459 in *T. thermophilus* (PDB file 2J00 versus 1FKA) allowed us to observe a degenerated triloop conformation (see Figure 5-4C). The triloop structure observed in 2J00 was resolved

bound to the antibiotic paromomycin (Figure 5-4A,C, left), whereas the degenerated one was resolved in a transcription-activated state, free of any antibiotics. In the structure bound to the antibiotic, the two triloops (in H10 and H17) expose a nucleotide inside the cavity created by the junction of the three loops. In the second structure, the two exposed nucleotides have moved toward their respective loop, leaving the cavity empty. The reason why the cavity is left empty in the activated state is unknown. Nevertheless, it shows that the two triloops were affected in a similar fashion by the experimental conditions.

The interactions found in the triloop of the X-ray crystallographic structure with the antibiotics can be obtained from the degenerated one by simple 3D modeling manipulations. First, flipping out C3 would allow for the stacking between G4 and A2. Second, bringing A5 inside the loop would allow for a pairing with G1. Finally, C0 and G6 can be slightly altered to retrieve the last Watson–Crick *cis* base pair of the adjacent stem. The degenerated structure observed in 1FKA can possibly be a preliminary state of the triloop, or the result of conformational changes induced by the activated state of the ribosome. Besides, the investigators of this X-ray crystallographic structure mentioned that long-range conformational changes could be transmitted along what they call a long structural pillar that includes H17 (Schluenzen, Tocilj et al. 2000)

Recently, RNA dynamics have also been reported in the literature in two interior loops. The X-ray crystallographic structures of the 23S rRNA of *Deinococcus radiodurans* and *E. coli* contain an interior loop in helix 40. Turner and coworkers have solved this interior loop by NMR and found a lower ground state than that of the X-ray crystals (Shankar, Kennedy et al. 2006). An important difference between the two states is the disappearance of a noncanonical A–A base pair in the X-ray crystal structures. Similarly, the cytoplasmic A site in the small rRNA of *Homo sapiens* has been solved by X-ray crystallography, showing two distinct structures (Kondo, Urzhumtsev et al. 2006). The two structures are referred to as the OFF and ON states,

which correspond, to a free A site and to a loaded one, respectively, where a tRNA brings a new amino acid to a nascent protein. The ON state shows two bulging out nucleotides that pair inside the helix in the OFF state. The two examples above show clearly the appearance/disappearance of intraloop base-stacking and base-pairing interactions, such as those observed in the triloops.

The comparative analysis also revealed the presence of two overlapping triloops in two different species, *E. coli* and *T. thermophilus*, which share four of their five nucleotides (Figure 5-5). The effect created by the presence of antibiotics differs in each species, and could thus be species specific. The *T. thermophilus* structure with antibiotics (2J01) shows the presence of two triloops. The *E. coli* structure with antibiotics (1VS6) only shows one triloop. Finally, the *E. coli* structure without antibiotics (2AW4) restores the second triloop, a structure identical to that of the *T. thermophilus* with the antibiotics. Worth mentioning, this triloop site is also located in H17, but in *E. coli*. These results show that ligand binding has an extremely important effect on RNA structure, which can be structure specific, and cannot be accounted for in any context-free RNA structure predictor.

5.1.6.4 Three-dimensional modeling

Our first approach to capture sequence–structure relationships in triloops focused on a structural classification based on the 5-nucleotide (nt) sequence. However, the 60 different sequences spanning 102 specimens were insufficient to derive valid statistics. We then focused our attention on the middle triplets (nucleotides 2–4), based on the argument that isosteric base-pair substitutions can introduce noise in the statistics, since the partners can be almost any nucleotide with little influence on the geometry of the triloop (by isostericity definition) (Lescoute, Leontis et al. 2005). Unfortunately, most triplets are found across more than one class, and thus are not invariants of particular structures, as we wished. We then focused on the last alternative to find an invariant in triloop structures: the flanking base-pairing

type, which was previously linked to tetraloop structures (Cheong, Varani et al. 1990, Woese, Winker et al. 1990, Antao, Lai et al. 1991, Heus and Pardi 1991).

As a matter of fact, the flanking bases bring more than half (1.13/1.88 bits) the mutual information between triloop sequence and structure. As an example, consider the following diverse sequence and structure triplets: AGU, AGA, CAU, and AAC. These triplets are found in one, three, six, and two different structures, respectively. However, when these triplets are combined with a flanking GA base pair, they all fold in a single structure, and the base-pairing type is either sugar/Hoogsteen or sugar/Watson–Crick. Furthermore, in general, the structural class of one specific triplet changes in function of its combined flanking base-pairing type. Consider, in particular, GCG, which folds in a structure flanked by a sugar/Hoogsteen *trans* C–A base pair, in a different one when flanked by a Watson–Crick *trans* U–U base pair, in another one when it is flanked by a sugar/Hoogsteen *trans* G–A base pair, and finally in yet another one when flanked by a sugar/Bifurcated-sugar *cis* U–A base pair.

The correlation between the flanking base-pairing type, sequence, and triloop structure applies to all sequences but five, which preserve the flanking base-pairing type while changing structures. However, we find that such a small number of "dynamic" sequences suggests that the nucleotide interplay is not a general phenomenon unless we do not have enough structural data to observe it further. Or perhaps, only a limited number of sequences exhibit the interaction dynamics, and all other structures can be predicted precisely from sequence. The former hypothesis is supported by NMR data. Consider the PDB file 1NBR, which contains 15 NMR models of an iron responsive element (IRE). The 15 models classify into three structural graphs, two of which show nucleotide interplay: a base-stacking movement along the backbone path and the appearance of an intraloop base-stacking interaction (data not shown). Corroborating with our modeling hypothesis, the flanking base-pairing type does not change among the 15 NMR models. We find that the IRE solution exhibits a sampling of different triloop conformations.

In terms of 3D modeling, the few possible flanking base-pairing types for each partner pair, in general, implies the consideration of multiple alternative structures for any given triloop sequence (see Table A3-III in the online Supplemental Materials). Applied to the IRE triloop sequence, CAGUG, for instance, we find 11 possible subclasses with two different flanking base-pairing types, Watson–Crick/Hoogsteen and Watson–Crick. Among the triplets in the CG entry, we do not find the IRE sequence. However, one is very close, CAGAG, which would be a good first guess. In fact, it actually points to the structure that includes the right flanking base-pairing type and two of the three intraloop interactions of the NMR models. In a real modeling application, however, all alternatives need to be considered.

The importance of the flanking base pair in the sequence–structure relationships can be explained by the identity and positioning of the flanking bases, which strongly direct the backbone conformation and, by ricochet, influence the position of the other bases, which in turn, defines their interactions.

The work presented here highlights the structural diversity of the RNA triloop motif in different contexts. Despite this diversity, we were able to extract sequence–structure relationships. The flanking base-pair sequence provides useful RNA modeling information. The systematic classification approach developed in this work, based on discrete folding states defined by base interactions, is easily scalable to any RNA motif. It would provide valuable information about any RNA motif and modeling approach. Our intention is to build similar classifications for other RNA motifs.

5.1.7 Materials and methods

5.1.7.1 Triloop database

All RNA X-ray crystallographic structures of resolutions higher or equal to 3 Å, which were available in the PDB (Berman, Westbrook et al. 2000) in October 2006, were considered. Classical RNA triloops are composed of five contiguous nucleotides, where the first and last form a flanking base pair, and the three others form the loop (see Figure 5-1). This information was input to MC-Search to find all triloop sites and to produce the triloop database. MC-Search takes an RNA motif description (such as in Figure 5-1A, bottom) and a database of 3D structures, and returns the sites that match the motif in the database by applying a classical graph isomorphism algorithm.

5.1.7.2 Triloop annotation

The triloops were analyzed by the MC-Annotate computer program(Gendron, Lemieux et al. 2001), which labels nucleotide interactions from atomic coordinates. MC-Annotate applies a base-pairing classifier (Lemieux and Major 2002), which returns its type, for each base pair found, using the Leontis and Westhof nomenclature (Leontis and Westhof 2001) (see below). Base stacks are found using the Gabb method (Gabb, Sanghani et al. 1996).

5.1.7.3 Nomenclature

We consider three kinds of nucleotide interactions: base pairing, base stacking, and nucleotide linkage. The Leontis and Westhof nomenclature is used for labeling the base-pairing types (Leontis and Westhof 2001) . We use three different symbols to represent the three edges of a base: the Watson–Crick (W; • *cis*; ○ *trans*), Hoogsteen (H; ■ *cis*; □ *trans*) and sugar (S; ◄ *cis*; ◅ *trans*). The *cis*/*trans* indicates the relative orientation of the backbone across the median of the plane formed by the two partners. For instance, a sheared sugar/Hoogsteen *trans* G–A base pair is written G◅□A, or S/H. We use a single symbol when the groups involved in H bonds in the two

bases are on the same edge. For instance, we write X□Y instead of X□□Y. Bifurcated base-pairing types are indicated by Bs and Bh, where Bs points to the bifurcating group between the sugar and Watson–Crick edges, and Bh between the Watson–Crick and Hoogsteen edges (Lemieux and Major 2002).

Base-stacking types are shown using arrows. The tip of the arrows indicates the normal to the plane of the base, defined so that any base in a classical A-RNA type double helix has its normal vector oriented toward the 3′-strand endpoint (Major and Thibault 2006). In pyrimidines, we use a right-handed axis system to define a normal by the rotational vector around atoms N1 to N6. In purines, the normal is reversed to that of pyrimidines, since the atoms of their pyrimidine rings are numbered in the reversed order. Two arrows pointing in the same direction indicate A-RNA double-helix type: up (>>) or down (<<), depending which base is named first (i.e., B1>>B2 means B2 is stacked upward of B1, or B1 is stacked downward of B2). Two other types are possible, but less frequent in RNAs inward (B1> <B2; B1 or B2 is stacked inward of B2 or B1, respectively) and outward, respectively (B1< >B2; B1 or B2 is stacked outward of B2 or B1, respectively).

In the structural graphs, thick lines represent the presence of phosphodiester linkages, whereas a thin line, or the absence of a line, indicates 2 nt that are not linked covalently. The crystallographic structure numbering system is used throughout the article.

5.1.7.4 Mutual sequence-structure information

The mutual sequence–structure information has been obtained from building the first layer of an ID3 decision tree from an information table (Genesereth and Nilsson 1987), in which we assigned one of the eight structural triloop topology to each sequence (see Table A3-III in the online Supplemental Materials). The mutual information is calculated by:

$$I(data) = \sum_{i=1}^{n} p_i * \log_2 p_i$$

where p_i is the probability of class *i*, that is:

p_i = (nb of occurrences of class *i* / total number of occurrences)

$I(data) = 1.88$ bits

We then calculate the gain in information of the flanking base pair nucleotides by first calculating the information (as above) for this criterion, $I(data_v)$, where *i* takes all the values in the criterion:

criterion = [AA,AC,AG,AU,CA,CC,…,UU]

Then, the gain in information of this criterion (flanking base pair) in our data is calculated as:

$$Gain(data,criterion) = I(data) - \sum_{v \in criterion} \frac{|data_v|}{|data|} I(data_v)$$

where $|data_v|$ is the number of occurrences of a particular flanking base pair.

$Gain(data,criterion) = 1.88 - 0.75 = 1.13$ bits

5.1.8 Supplemental Data

All Supplemental Materials are available at www.major.iric.ca and in A3.

5.1.9 Acknowledgments

This work was supported by a grant from the Canadian Institutes of Health Research (CIHR) (MT-14604) to F.M. F.M. is a CIHR investigator and a member of the Centre Robert-Cedergren of the Université de Montréal. V.L., at the time of this work, held a CIHR scholarship to encourage higher education in bioinformatics (Université de Montréal, programme biT). The authors thank Dr. Robin Gutell for providing them with the image of the secondary structure of the 23S rRNA of *H. marismortui*.

5.2 Article 4: RNA Sequence Design Using a Three-Dimensional Quantitative Structure-Activity Relationships Approach

en préparation pour publication

5.2.1 Contribution des co-auteurs

KSO: Conception et implantation de la méthode, analyse et discussion des résultats, rédaction du manuscript.

VL: Réalisation et interprétation de la validation expérimentale.

SH: Lecture et critique du manuscript.

FM: Conception et supervision du dévelopement de la méthode, analyse et discussion des résultats, rédaction du manuscript.

5.2.2 Mise en situation

Durant mes études doctorales, j'ai collaboré à une autre modélisation qui reflète bien la puissance de la combinaison de la modélisation informatique et de la biologie moléculaire afin de mieux comprendre certains phénomènes biologiques. Les travaux de recherche d'une collègue, Karine St-Onge, l'ont amené à modéliser la relation structure/activité d'un motif d'ARN, le domaine sarcin/ricin de la sous-unité ribosomale de taille 23S de *E. coli* en utilisant une méthode d'analyse quantitative de cette relation. Ce domaine, composé de 29 nt, tire son nom du fait qu'il est le lieu d'attaque de deux toxines: la ricine, issus du ricin et α-sarcin. Par deux modifications différentes du ribosome, ces toxines inhibent l'interaction de celui-ci avec les facteurs d'élongation EF-Tu et EF-G (Hausner, Atmadja et al. 1987, Moazed, Robertson et al. 1988, Brigotti, Rambelli et al. 1989). La délétion de la région de la boucle du domaine sarcin/ricin (nucléotides 2653 à 2667, voir Figure 5-7) confère, *in vivo*, un phénotype létal dominant (Lancaster, Lambert et al. 2008), ce qui est cohérent avec la très grande conservation de cette région à travers une variété d'espèces (Gutell, Schnare et al. 1992). Diverses études de mutagénèse dirigée ont montrées que la structure de cette boucle est importante à l'établissement d'un ribosome fonctionnel (Macbeth and Wool 1999, Chan, Sitikov et al. 2000, Chan, Dresios et al. 2006, Chan and Wool 2008) ce qui en fait un domaine de choix pour l'analyse de la relation structure/activité.

Les travaux de recherche sur la modélisation de la relation structure/activité faits au laboratoire ont menés à la prédiction d'un ensemble de séquences de la boucle du domaine dont la structure serait équivalente à celle de la séquence sauvage. Ces séquences mèneraient à la formation d'un ribosome fonctionnel et donc à la survie des bactéries. Un autre ensemble de structures seraient suffisamment différentes de la structure de la séquence sauvage pour mener à un ribosome non fonctionnel qui donc ne permettrait pas la survie des bactéries. Afin de vérifier l'exactitude de ces prédictions, nous avons mis au point une technique expérimentale permettant de détecter les mutations menant à un ribosome fonctionnel de celles menant à un ribosome non fonctionnel. Cette méthode expérimentale, combinée à l'approche de modélisation informatique, permet d'étudier indirectement la structure de ces variantes de séquence.

5.2.2.1 Méthodes expérimentales

5.2.2.1.1 Prédictions de séquences viables et léthales

La modélisation de la relation structure/activité utilise comme référence la structure cristallographique publiée (Correll, Beneken et al. 2003). Chacune des mutations associées à un phénotype (viable ou létal) est ensuite modélisée en trois dimensions en utilisant l'outil MC-Sym (Major, Turcotte et al. 1991, Parisien and Major 2008) et en lui fournissant comme contraintes les appariements présents dans la structure de référence. Ces modélisations ont été faites sur la région de la boucle du motif de sarcin/ricin, la région identifiée précédemment comme étant clé de l'activité de ce motif. À cette région a été ajoutée la première paire de base de la double hélice à la base de cette boucle (voir Figure 5-7). Les diverses structures sont ensuite comparées entre elles et à la structure de référence. Les déterminants de l'activité d'une séquence sont identifiés par apprentissage machine en considérant l'accessibilité et la charge de chacun des atomes des structures de chacune des séquences, toutes modélisées en appliquant les contraintes présentes dans la

structure sauvage. Ces déterminants sont ensuite utilisés afin de prédire le phénotype de nouvelles séquences.

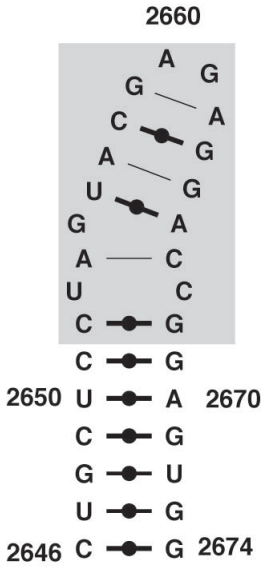


Figure 5-7 Structure secondaire du domaine sarcin/ricin basée sur la structure expérimentale. La structure secondaire du domaine sarcin/ricin de l'ARNr de *E. coli* est présentée. La région en gris est celle qui a été utilisée durant la modélisation de la relation structure/activité. La numérotation est celle de la structure expérimentale

5.2.2.1.2 Vérification des prédictions

Le système expérimental choisi pour tester le phénotype associé aux nouvelles séquences est basé sur celui utilisé dans plusieurs autres analyses de l'activité de cette région (Macbeth and Wool 1999, Chan, Sitikov et al. 2000, Chan, Dresios et al. 2006, Chan and Wool 2008). Dans ce système, le plasmide pLK45 contient d'une part l'ARNr de *E. coli* sous le contrôle du promoteur λP_L. Dans ce plasmide, l'ARNr de la sous-unité 23S contient une mutation (A2058G) conférant la résistance à l'antibiotique érythromycine. Ce plasmide contient également un gène de résistance à l'ampicilline. Les bactéries DH1/cI expriment le répresseur thermolabile λcI du promoteur λP_L ainsi qu'un gène de résistance à la kanamycine. Ces bactéries ont été transformées avec le plasmide pLK45 et sélectionnées à basse température (30°C) par la kanamycine, pour s'assurer de l'expression du répresseur, et par l'ampicilline pour

s'assurer de la présence du plasmide pLK45. À cette température, le répresseur λcI est actif et donc l'ARNr du plasmide pLK45 n'est pas transcrit et n'affecte pas la prolifération des bactéries. À température élevée (42°C), le répresseur est inactivé et donc l'ARNr du plasmide pLK45 est exprimé. Dans ces conditions, environ 30% des ribosomes exprimés proviennent du plasmide (Powers and Noller 1990, Macbeth and Wool 1999). Lorsque l'érythromycine est combinée à l'incubation à haute température, seul l'ARNr provenant du plasmide est exprimé et il est alors possible de déterminer si la séquence présente dans ce plasmide génère un ribosome fonctionnel en étudiant la prolifération des bactéries. Les diverses mutations de séquence du motif sarcin/ricin ont été effectuées en utilisant la trousse de mutagénèse dirigée QuickChange II XL de Agilent Technologies en suivant les instructions du manufacturier. Les mutations ont été vérifiées par séquençage.

Pour les essais de prolifération, les bactéries DH1/cI contenant le plasmide pLK45 codant pour l'ARNr sauvage ou un de ses mutants ont été cultivées dans un milieu LB contenant 50μg/mL d'ampicilline et 30μg/mL de kanamycine à 30°C jusqu'à ce que la culture atteigne une absorbance de 0.6 à 650nm. Ces cultures ont ensuite été diluées (de 10^{-1} à 10^{-4}) et déposée en gouttes de 7ul sur des pétris d'agar contenant soit 50μg/mL d'ampicilline et 30μg/mL de kanamycine (A+K) ou 50μg/mL d'ampicilline, 50μg/mL de kanamycine et 50μg/mL d'érythromycine (A+K+E). Les pétris ont été incubés à 30°C (A+K) ou à 42°C (A+K+E) pendant 16-20 heures. Les expériences ont été réalisées à trois reprises en duplicata et des résultats représentatifs sont présentés.

5.2.2.2 Résultats

Des 30 séquences pour lesquelles des prédictions d'activité ont été faites, nous avons identifié huit séquences à tester expérimentalement (voir Tableau 5-I). Chacune de celles-ci possède des caractéristiques intéressantes nous permettant de mieux comprendre la relation entre la structure et la séquence. Les séquences SK12 et SK13 présentent des mutations complémentaires à la base de la région modélisée. Ces deux

séquences sont conservées dans le domaine sarcin/ricin de d'autres espèces, suggérant que celles-ci soient viables également dans le contexte de l'ARNr de *E. coli*. La séquence SK19 est également supportée par les alignements de séquences. Cette mutation pourrait entraîner un léger changement de structure pour permettre un appariement entre G2655 et C2665, exposant le A2654 plutôt que le G2655. Les autres séquences ne sont pas supportées par les alignements et constituent donc de véritables nouvelles prédictions. SK26 formerait une paire non-canonique G2658-G2663 ou causerait un important réarrangement du registre. Les séquences SK28 et SK30 perturbent également la même région de la structure. Finalement, les séquences SK36 et SK40 présentent plusieurs modifications différentes qui devraient affecter toute la structure.

Tableau 5-I Séquences du motif sarcin/ricin testées expérimentalement.
Les mutations effectuées sont présentées en caractères gras et en lettres minuscules.

Identifiant	Mutation	Séquence
WT	-	CUAGUACGAGAGGACCG
SK12	C2652U	**u**UAGUACGAGAGGACCG
SK13	G2668A	CUAGUACGAGAGGACC**a**
SK19	A2665C	CUAGUACGAGAGG**c**CCG
SK26	C2658G	CUAGUA**g**GAGAGGACCG
SK28	C2658U/G2663C	CUAGUA**u**GAGA**c**GACCG
SK30	C2658U/G2663U	CUAGUA**u**GAGA**u**GACCG
SK36	C2652A/U2653G/G2659C/G2661C/A2662G /G2664A/A2665G/C2666A	**ag**AGUAC**c**U**cg**G**aga**CG
SK40	C2652U/U2653C/A2654G/U2656G/G2664U G2668A	**ucg**G**g**ACGAGAG**u**ACC**a**

Les bactéries transformées avec les divers mutants ont été cultivées tel que décrit précédemment. Lorsque cultivées à 30°C, les bactéries transformées avec chaque variante de séquence croissent de manière similaire aux bactéries ayant la séquence sauvage (voir Figure 5-8), tel qu'attendu du fait qu'à cette température, le répresseur est activé et donc l'ARNr du plasmide n'est pas exprimé. Cependant, puisque celles-ci croissent sur un milieu contenant de l'ampicilline, le plasmide contenant une variante de séquences d'ARNr (pLK45) est présent. Lorsque les bactéries sont incubées à 42°C, le répresseur est inactivé et donc l'ARNr du plasmide est exprimé. En présence d'érythromycine, seule la sous-unité 23S du ribosome provenant du plasmide pLK45 est fonctionnelle et donc la prolifération des bactéries

est entièrement dépendante de la fonctionnalité du ribosome possédant la variante de séquence. Les tests de croissance à 42°C montrent que les bactéries exprimant les séquences KS26, KS36 et KS40 ne sont pas ou très peu viables. Ces séquences se replieraient donc en une structure différente qui ne permettrait pas au ribosome d'être fonctionnel alors que toutes les autres séquences peuvent adopter la même structure que la séquence sauvage. Ces résultats concordent avec les prédictions basées sur la modélisation de la relation structure/activité et supportent l'exactitude de la méthode.

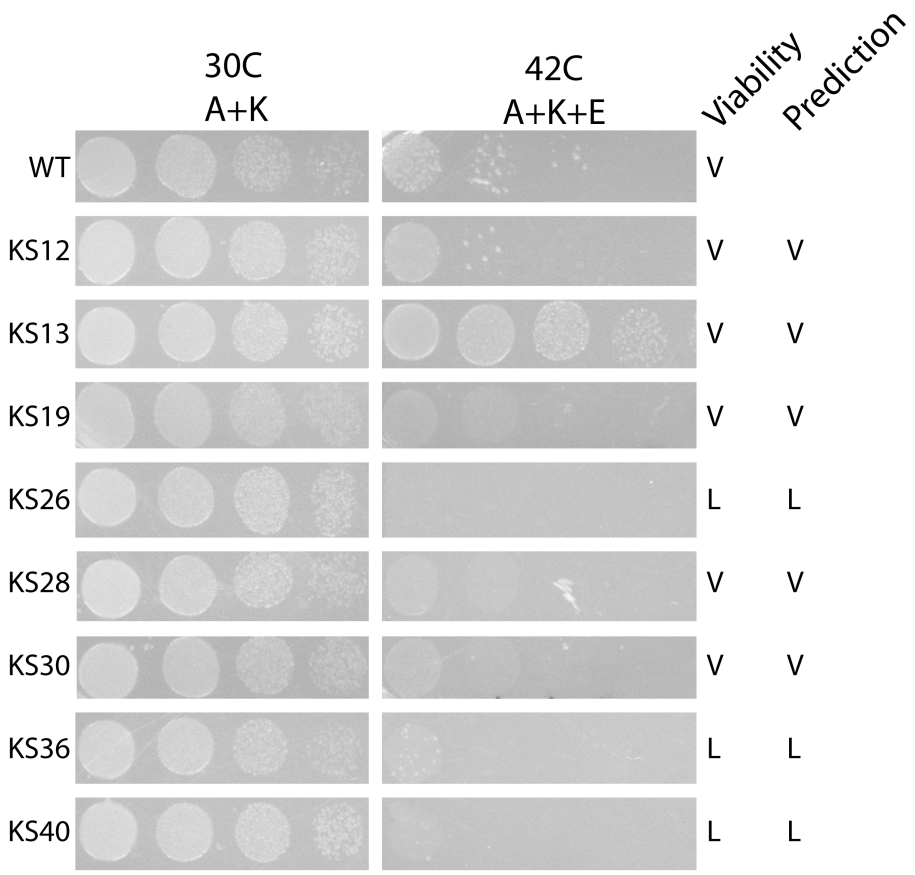


Figure 5-8 Test de croissance des diverses variantes de séquence du motif sarcin/ricin.
Toutes les variantes de séquence ont une croissance similaire lorsque déposées sur un milieu contenant de l'ampicilline et de la kanamycine et incubées à 30°C. À la température supérieure et en présence d'érythromycine, les bactéries exprimant les mutations de séquence de l'ARNr KS26, KS36 et KS40 ne prolifèrent pas (séquences létales (L)) alors que les autres prolifèrent (séquences viables (V)). Toutes les séquences testées sont cohérentes avec les prédictions issues de la modélisation.

5.2.2.3 Discussion

La méthode expérimentale choisie repose sur un test fonctionnel lié à la structure de la séquence d'intérêt plutôt que sur une analyse de la structure directement. Ce choix a été motivé d'une part par la modélisation à la base des prédictions. Cette modélisation évalue d'abord la similitude d'une structure prédite pour une nouvelle séquence à la structure de la séquence sauvage. Cette similitude est ensuite transformée en une prédiction d'activité, dans le cas du domaine sarcin/ricin, la fonctionnalité du ribosome par l'analyse de la viabilité des bactéries. Puisque la

modélisation permet de prédire l'activité, il est naturel de tester expérimentalement cette même activité.

D'autre part, une analyse directe de la structure peut difficilement avoir une précision de résolution équivalente à ce que la cellule peu distinguer. La machinerie cellulaire, particulièrement dans le cas de la traduction par le ribosome est optimisée pour reconnaître avec précision le positionnement de certains atomes ou groupements chimiques. Il n'existe présentement aucune méthode expérimentale permettant de déterminer avec cette même précision la structure d'un ARN replié dans un contexte identique au contexte cellulaire.

Lors de la modélisation de la structure, nous avons fait abstraction des autres parties du ribosome et des interactions du domaine sarcin/ricin avec ces autres parties du ribosome et avec les protéines ribosomales. Ceci est possible sous l'hypothèse que le bon positionnement des atomes et des groupes chimiques permet de conserver ces interactions. Dans le contexte d'une analyse directe de la structure, il est très difficile de conserver toutes ces interactions dont plusieurs sont inconnues. Il est donc difficile de s'assurer que la structure obtenue est en tout point identique à celle de la séquence sauvage.

L'analyse de résultats de tests fonctionnels peut parfois être malaisée parce que plusieurs hypothèses peuvent expliquer ces résultats. Dans le cas des tests fonctionnels présentés ici, d'autres hypothèses que celle avancée peuvent être émises pour expliquer le résultat. En effet, on peut imaginer que le changement de séquence ou de structure rendrait le ribosome sensible à l'érythromycine sans pour autant le rendre non fonctionnel dans un contexte sans cet antibiotique. La structure pourrait aussi être la même mais le changement de séquence pourrait faire en sorte que la structure n'est plus reconnue par les facteurs d'élongation ou d'autres protéines. Cependant, ces tests fonctionnels combinés aux prédictions issues de la modélisation de la relation structure/activité renforcent l'hypothèse que les séquences viables conservent la même structure que la structure de la séquence sauvage et que les

séquences létales abolissent cette structure. Ceci met en lumière l'utilité de la modélisation informatique dans un contexte de compréhension de phénomènes biologiques.

RNA Sequence Design Using a Three-Dimensional Quantitative Structure-Activity Relationships Approach

Karine St-Onge[1], Véronique Lisi[2], Sylvie Hamel[1] and François Major[1,2]

[1]Department of Computer Science and Operations Research, [2]Institute for Research in Immunology and Cancer, Université de Montréal, PO Box 6128, Downtown Station, Montréal, Québec, H3C 3J7, CANADA

Keywords: RNA, structure, prediction, QSAR

5.2.3 Abstract

Ribonucleic acids are first-choice molecules to intervene in cellular programs since they both carry and control genetic information. Tens of RNA therapeutics are currently under clinical trials, and RNA-synthetic components that react to prefixed conditions and release therapeutics RNAs have been conceived and prototyped. However, engineering RNAs with predetermined structure and function requires profound understanding of RNA structure-activity relationships. Here, we introduce a principal component analysis approach, *MC-QSAR*, to discriminate RNA variants that preserve a given function. *MC-QSAR* builds the 3D structural profile of active sequences from a set of known active and inactive sequences. After the supervised learning step, the 3D structure of new sequences is matched to the profile: those fitting the profile are predicted to be active, while those that do not are predicted to be inactive. To exemplify *MC-QSAR*, we built a training set of twelve 23S ribosomal sarcin-ricin loop sequences, four known to be viable and eight to be lethal. Information about the key nucleotides of this loop was obtained by selecting the best parameters using leave-one-out cross validation. These nucleotides are either involved directly with the loop interactions with elongation factors or to maintain the necessary structural features in place for thereof. Besides, *MC-QSAR* was successful in predicting the outcomes of 23 out of 24 sarcin-ricin loops in transgenic bacteria.

5.2.4 Introduction

The unmatched flexibility of ribonucleic acids (RNAs) makes them molecules of choice for therapeutics (Burnett and Rossi 2012). RNA sequences fold in three-dimensional (3D) structures that confer precise biological function. The surface of RNA 3D structures offers a variety of chemical groups including several hydrogen donors and acceptors, the reactive 2'OH, and an electronegative phosphodiester chain. In tandem with proteins to carry and control genetic information, RNAs and proteins form numerous complexes.

RNA structures are dynamics (Bailor, Sun et al. 2010). The many possible conformations available to an RNA are determined by its sequence of nucleotides. These conformations determine function. Therefore, any change in the sequence can alter its available structures and function. However, predicting to which extent any given mutation in the RNA sequence has on its function is difficult. To address this problem, we are taking a quantitative structure-activity relationships (QSAR) approach.

QSAR approaches are classically used to optimize the chemistry of ligands bound to receptors in the context of rational drug design. QSAR methods have been used to optimize protein primary (Caballero, Fernandez et al. 2006) and topological structures (Perez Gonzalez, Gonzalez Diaz et al. 2003, Cabrera, Gonzalez et al. 2006), as well as topological descriptors of chemical structures (Xiao, Xiao et al. 2002). QSAR methods have been applied to RNAs as well. They have been used successfully to probe anticancer activity (Helguera, Rodriguez-Borges et al. 2007), to predict the local binding affinity constants between a specific nucleotide and an antibiotic (Marrero Ponce, Castillo Garit et al. 2005) or to quickly identify and predict miRNAs (Gonzalez-Diaz, Vilar et al. 2007). The predictor variables used to model these RNA QSAR approaches range from 2D descriptors like chemical topology (planar representation of atoms and their chemical bonds) (Helguera, Rodriguez-Borges et al. 2007), to vectors of numerical values representing nucleotide properties (experimental molar

absorption coefficients, single excitation energies, oscillation strength values, etc.) (Marrero Ponce, Castillo Garit et al. 2005), and finally to thermodynamic molecular descriptors such as free energies, entropies and melting temperatures (Gonzalez-Diaz, Vilar et al. 2007). All those models were calibrated using linear discriminant analysis techniques and validated using either cross-validation techniques (Helguera, Rodriguez-Borges et al. 2007), leave-one-out jack-knife experiments (Marrero Ponce, Castillo Garit et al. 2005) or ROC curve analysis (Gonzalez-Diaz, Vilar et al. 2007).

Here, we introduce a QSAR method, *MC-QSAR*, to determine if an RNA sequence exhibits the same function as that measured experimentally on a training set of sequences. The 3D structures accessible to the training set sequences are predicted using high-resolution experimental structures, if any, and RNA 3D structure prediction. In comparison to traditional QSAR approaches, the goal here is not to optimize any profile, but rather to distinguish among new sequences those that would be active (exhibit the function of the training sequences) and inactive (do not exhibit the function).

The training set can be composed of active and inactive sequences. We use principal component analysis (Pearson 1901) (PCA) to build the electrostatic profile of active structures based on the commonly exposed charges of the 3D structures of the active sequences that are absent in the inactive sequences. The predicted structures of a new sequence for which we want to determine functionality are then mapped to this electrostatic profile. If the structure of the new sequence fits the profile, then it is determined to be active or inactive otherwise.

To illustrate and validate *MC-QSAR*, we use the sarcin-ricin loop (SRL), located in one of the longest conserved RNA sequences in large ribosomal subunits (Cannone, Subramanian et al. 2002) (see Figure 5-9A). The SRL contains the classical GNRA tetraloop (Jucker and Pardi 1995). The name SRL is attributed to the fact that two protein toxins, α-sarcin and ricin, bind to it and inactivate protein synthesis by blocking the ribosome's interactions with elongation factors (EFs) (Moazed,

Robertson et al. 1988, Szewczak and Moore 1995). The α-sarcin catalyzes the hydrolysis of the phosphodiester linkage between the R and A of the GNRA tetraloop (Endo and Wool 1982), while the ricin favours the depurination of the N in the GNRA (Endo, Mitsui et al. 1987). Since the SRL is highly conserved across ribosomes, it makes it an interesting case to study its structure-activity relationships.

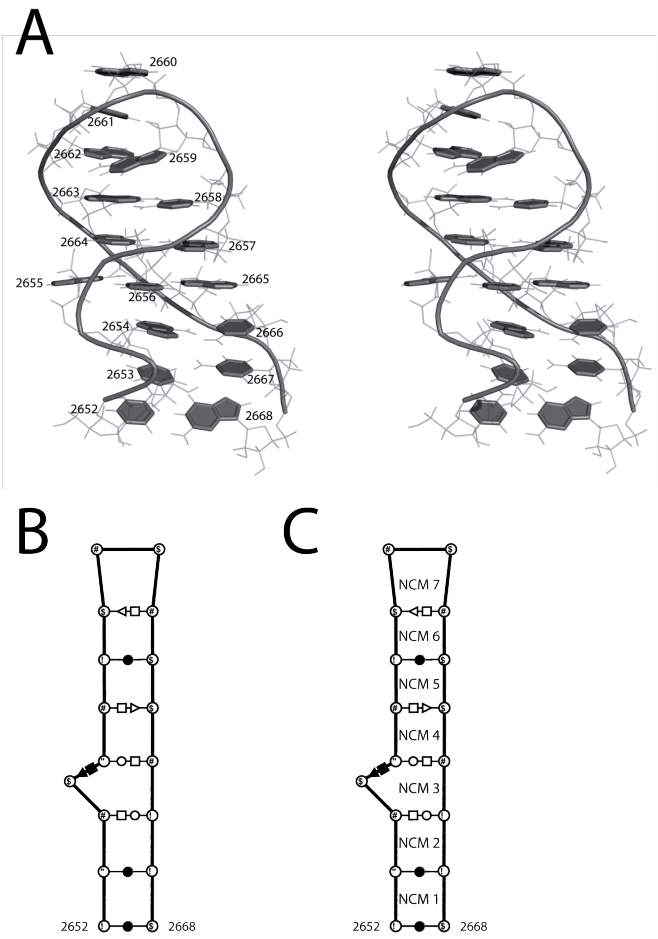


Figure 5-9 SRL.
A Stereo view of the SRL. Nucleotides are illustrated by planar forms and phosphodiester links by cylinders. **B** Tertiary structure of the SRL in the 23S rRNA of *E. coli* (B2652-B2668) using *MC-Annotate* (Gendron, Lemieux, & Major, 2001). Canonical base pairs are represented with a black circle according to Leontis-Westhof notation (Leontis & Westhof, 2001), sugar edge is represented with a triangle and hoogsteen edge with a square. Filled symbol indicates that the base pair is in cis orientation and blank symbol in trans. Dark line represents phosphodiester link. **C** NCMs are identified into the tertiary structure of the SRL. Same symbols are used as that in B).

To enhance the validation of *MC-QSAR*, we use two more examples: domain II tested for erythromycin resistance and P loop tested for cell growth. The domain II

mutations that cause resistance to erythromycin are located in a hairpin structure between nucleotides 1198 and 1247 (see Figure 5-10B). This is close to a short open reading frame in the 23S rRNA that encodes a pentapeptide whose expression in vivo renders cells resistant to erythromycin. Therefore, a possible mechanism of resistance caused by domain II mutations may be related to an increased expression of the pentapeptide (Dam, Douthwaite et al. 1996). The P loop mutations that cause lethal mutants are located between nucleotides 2249 and 2254 (see Figure 5-11B). Evidence is present for the participation of the P loop of 23 S rRNA in establishing the tertiary structure of the peptidyl transferase center. Nucleotide substitutions were introduced into the P loop, which participates in peptide bond formation through direct interaction with the CCA end of P site-bound tRNA (Gregory and Dahlberg 1999).

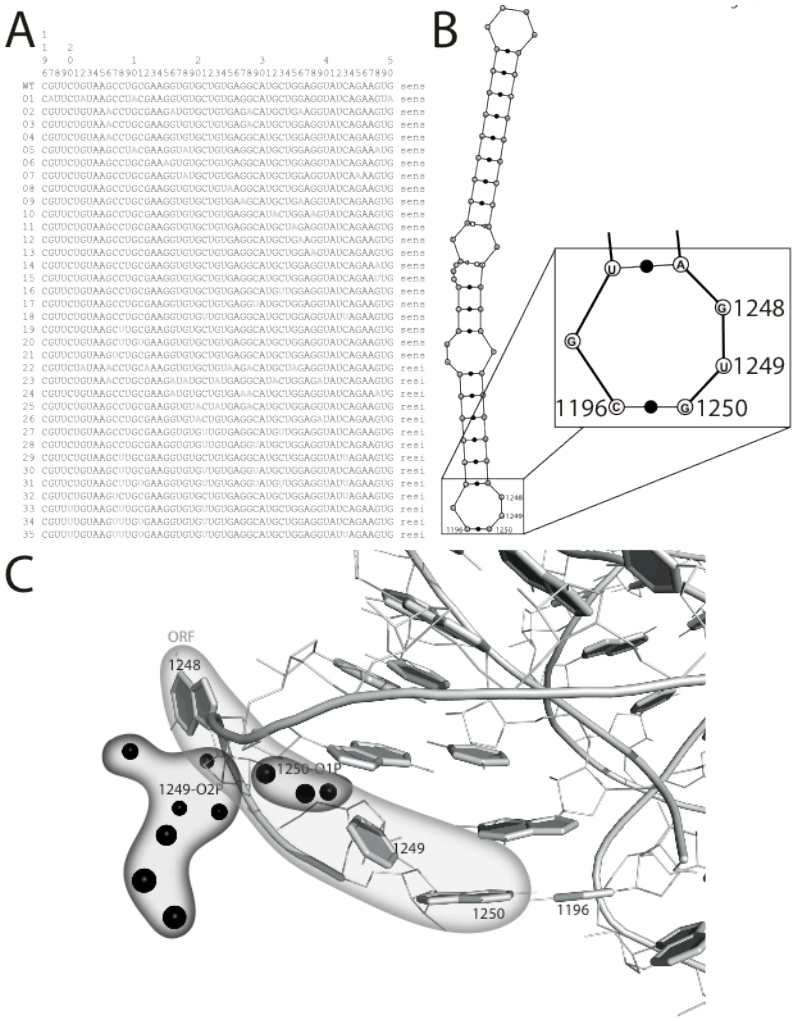


Figure 5-10 Domain II tested for erythromycin resistance.
A The thirty-six sequences from nucleotides 1196 to 1250 of the 23S *E. coli* that were tested for erythromycin resistance (Douthwaite, Powers et al. 1989, Aagaard and Douthwaite 1994, Leviev, Levieva et al. 1995, Dam, Douthwaite et al. 1996) mutations are indicated by bold/gray nucleotides. The numbering and wild-type (WT) sequence comes from the 23S *E. coli*. Notation sens/resi indicates that the sequence is sensitive or resistant to erythromycin. **B** The tertiary structure using *MC-Annotate* (Gendron, Lemieux et al. 2001). A close-up of the most significant NCM from the PCA analysis is shown. The numbering is the same as in A). Canonical base pairs are represented with a black circle according to Leontis-Westhof notation (Leontis and Westhof 2001), sugar edges are represented with a triangle and hoogsteen edges with a square. Filled symbol indicates that the base pair is in cis orientation and blank symbol in trans. Dark line represents phosphodiester link. **C** The 3D structure where the domain II is represented. Nucleotides are illustrated by planar forms and phosphodiester links by cylinders. The black areas represent significant atoms (indicated by spheres, where a bigger sphere is more significant) for discriminate sequences according to erythromycin resistance. The gray area indicates the short open reading frame (ORF) between nucleotides 1248 and 1250.

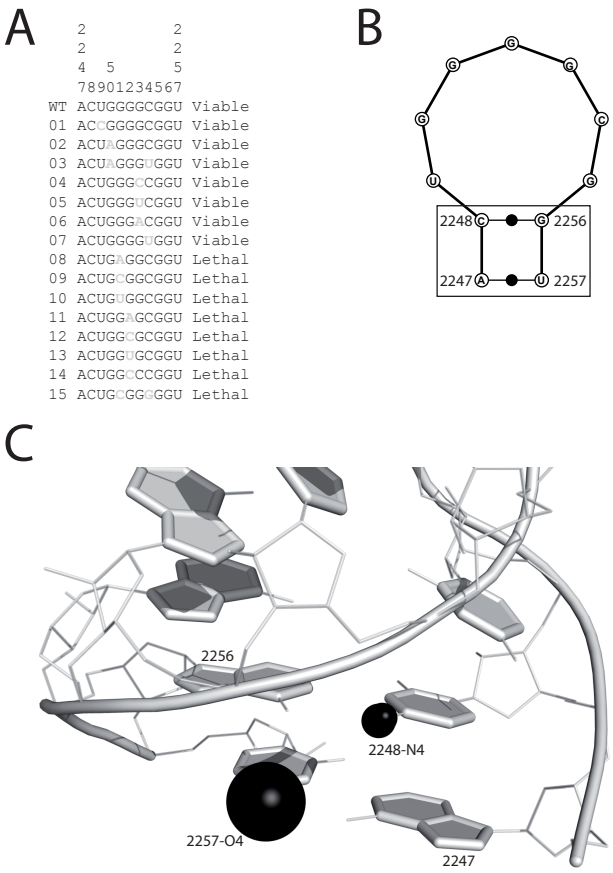


Figure 5-11 P loop tested for cell growth.
A The 16 sequences from nucleotides 2247 to 2257 of the 23S *E. coli* that were tested for cell growth (Gregory, Lieberman et al. 1994, Lieberman and Dahlberg 1994, O'Connor, Brunelli et al. 1995, Samaha, Green et al. 1995, Porse, Thi-Ngoc et al. 1996, Spahn, Remme et al. 1996, Green, Samaha et al. 1997, Bocchetta, Xiong et al. 1998, Gregory and Dahlberg 1999), mutations are indicated by bold/gray nucleotides. The numbering and wild-type (WT) sequence comes from the 23S *E. coli*. Notation viable/lethal indicates that the sequence is viable or lethal (cell growth). **B** The tertiary structure using *MC-Annotate* (Gendron, Lemieux et al. 2001). The most significant NCM from the PCA analysis is shown with a box. The numbering is the same as in A). Canonical base pairs are represented with black circle according to Leontis-Westhof notation (Leontis and Westhof 2001), filled symbol indicates that the base pair is in cis orientation. Dark line represents phosphodiester link. **C** The 3D structure where the hairpin is represented. Nucleotides are illustrated by planar forms and phosphodiester links by cylinders. Significant atoms are represented with spheres, where a bigger sphere is more significant for discriminating sequences according to cell growth.

5.2.5 Results

5.2.5.1 SRL predictions

Data set. Table 5-II shows the *MC-QSAR* activity predictions of the training set sequences. The four viable sequences and seven of the eight lethal sequences were correctly predicted. So we accurately predicted activity of 94% of the sequences from the training set. Moreover, we predict the activity of thirty new sequences (see Table 5-III) eight sequences that were in the alignment of the bacterial 23S rRNA subunit; eleven variants of the 2658-2663 base pair; and eleven randomly generated sequences that conserve dinucleotides composition. From the eight sequences that were in the alignment, we predict five sequences as viable and three as lethal. Then from the eleven variants of the 2658-2663 base pair, we predict three sequences as viable and eight as lethal. Finally from the eleven randomly generated sequences, we predict three sequences as viable and five as lethal. Note that no model is produced for three of the random sequences so the prediction is not possible.

Table 5-II Predictions of the training set.
Sequences are identified with the number (or WT) to their left in the first column. WT identifier is for the wild-type sequence, from the 23S rRNA of *E. coli* (B2652-B2668). The type of model for each sequence is indicated in the second column; MFE for the minimum free energy model and RMSD for the closest model in RMSD of the seed structure. The activities (viable/lethal for growth of cells (Macbeth and Wool 1999, Chan, Sitikov et al. 2000, Chan, Dresios et al. 2006, Chan and Wool 2008) and the activity predictions (viable/lethal) from LOOCV are shown for each sequence in third and fourth column.

ID	Model	Activity	Prediction
WT	MFE	Viable	Viable
WT	RMSD	Viable	Viable
01	MFE	Lethal	Lethal
02	MFE	Viable	Viable
02	RMSD	Viable	Viable
03	MFE	Lethal	Lethal
04	MFE	Lethal	Lethal
05	MFE	Lethal	Lethal
06	MFE	Viable	Viable
06	RMSD	Viable	Viable
07	MFE	Lethal	Lethal
08	MFE	Lethal	Viable
09	MFE	Lethal	Lethal
10	MFE	Lethal	Lethal
11	MFE	Viable	Viable
11	RMSD	Viable	Viable

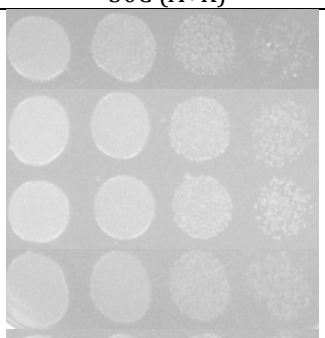
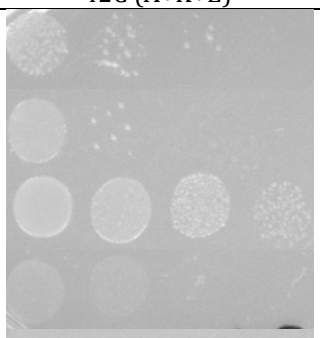
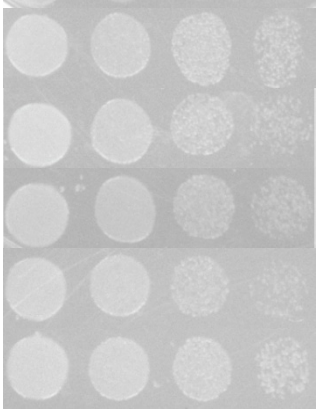
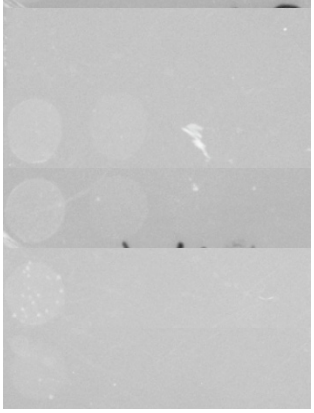
Table 5-III The data set.
Sequences are identified with the number (or WT) to their left in the first column. WT identifier is for the wild-type sequence, from the 23S rRNA of *E. coli* (B2652-B2668). Mutations in sequences are shown in gray in the second column. Crosses are used to identify sequences that are in the alignment of bacterial 23S rRNA sequences in the third column. The words "viable" or "lethal" are used to identify sequences that are tested experimentally (growth cells) in 4 papers (Macbeth and Wool 1999, Chan, Sitikov et al. 2000, Chan, Dresios et al. 2006, Chan and Wool 2008) in the fourth to seventh column. The activity predictions are indicated in the eighth column. Note that no model is produced for three of the random sequences so the prediction is not possible.

IDs	Sequence	Alignment	Chan *et al.*, JMB, 2000	Macbeth and Wool, JMB, 1999	Chan et al., JMB, 2006	Chan and Wool, JMB, 2008	Activity prediction
WT	CUAGUACGAGAGGACCG	×	viable	viable	viable	viable	viable
01	CUAGUAGGAGACGACCG		lethal			lethal	lethal
02	CUAGUAUGAGAAGACCG	×	viable				viable
03	CUAGUAAGAGAUGACCG		lethal				lethal
04	CUAGUACGAGACGACCG		lethal			lethal	lethal
05	CUACUACGAGAGGACCG	×		lethal		lethal	lethal
06	CUAAUACGAGAGGACCG			viable		viable	viable
07	CUAUUACGAGAGGACCG			lethal		lethal	lethal
08	CUAGUACGAGCGGACCG				lethal		viable
09	CUAGUACGAGGGGACCG				lethal		lethal
10	CUAGUACGAGUGGACCG				lethal		lethal
11	CUAGUACGUGAGGACCG					viable	viable
12	UUAGUACGAGAGGACCG	×					viable
13	CUAGUACGAGAGGACCA	×					viable
14	AUAGUACGAGAGGACCU	×					lethal
15	UUAGUACGCAAGGACCG	×					viable
16	CUUGUACGAGAGGACCG	×					viable
17	UUUGUACGAGAGGACCA	×					lethal
18	UUAGUACGAGAGGAUUU	×					lethal
19	CUAGUACGAGAGGCCCG	×					viable
20	CUAGUAAGAGAAGACCG						lethal
21	CUAGUAAGAGACGACCG						lethal
22	CUAGUAAGAGAGGACCG						lethal
23	CUAGUACGAGAAGACCG						lethal
24	CUAGUACGAGAUGACCG						lethal
25	CUAGUAGGAGAAGACCG						lethal
26	CUAGUAGGAGAGGACCG						lethal
27	CUAGUAGGAGAUGACCG						lethal
28	CUAGUAUGAGACGACCG						viable
29	CUAGUAUGAGAGGACCG						viable
30	CUAGUAUGAGAUGACCG						viable
31	UCAGUAGACCGGAGACG						viable
32	CUAGGACGAGAGUCACG						-
33	AGAGUCGACUAGGGACC						viable
34	CGAGUCACUAGGGAGAC						-
35	GGAGUAGACCACUCGGA						lethal
36	AGAGUACCUCGGAGACG						lethal
37	GGAGUACCGAGGACUCA						viable
38	CGAGUAGACGGGACUCA						-
39	GGUACUCGAGAGGACCA						lethal

40	UCGGGACGAGAGUACCA	lethal
41	AGGACUCGAGAGGUACC	lethal

Prediction set. From the thirty sequences from the data set, we selected eight sequences to be experimentally tested (cell growth): three sequences that were in the alignment of the bacterial 23S rRNA subunit; three variants of the 2658-2663 base pair; and two randomly generated sequences that conserve dinucleotides composition. The activity predictions are shown inTable 5-IV, where all of the eight tested sequences were correctly predicted.

Table 5-IV New sequences prediction.
Sequences are identified with numbers (or WT) in the first column. WT identifier is for the wild-type sequence, from the 23S rRNA of *E. coli* (B2652-B2668). Mutations in sequences are shown in gray in the second column. The third and fourth columns show the effect on the growth of *E. coli* cells of mutations in a 23 S rRNA gene in a plasmid-encoded rrnB operon. (30C A+K and 42C A+K+E). The experimental activities and the activity predictions from *MC-QSAR* (viable/lethal) are shown for each sequence in the fifth and sixth column. Act.: activity, Pred.: prediction

ID	Sequence	Experimental		Act.	Pred.
		30C (A+K)	42C (A+K+E)		
WT	CUAGUACGAGAGGA CCG			V	V
12	UUAGUACGAGAGGA CCG			V	V
13	CUAGUACGAGAGGA CCA			V	V
19	CUAGUACGAGAGGC CCG			V	V
26	CUAGUAGGAGAGGA CCG			L	L
28	CUAGUAUGAGACGA CCG			V	V
30	CUAGUAUGAGAUGA CCG			V	V
36	AGAGUACCUCGGAG ACG			L	L
40	UCGGGACGAGAGUA CCA			L	L

PCA analysis. The most influent nucleotide cyclic motif (NCM) in SRL according to the LOOCV analysis is the NCM5 (see Figure 5-9C) composed of nucleotides 2657-2658, 2663-2664. From that NCM, we observe that the O6 atom from nucleotide 2664, N3 (2663), N4 (2658) and O2P (2657 and 2658) (Figure 5-12) play a role in cell growth. However, the most significant atom is O6 from nucleotide 2664, means that a mutation that modifies the electrostatic area near this atom has more chance to change the activity of the sequence.

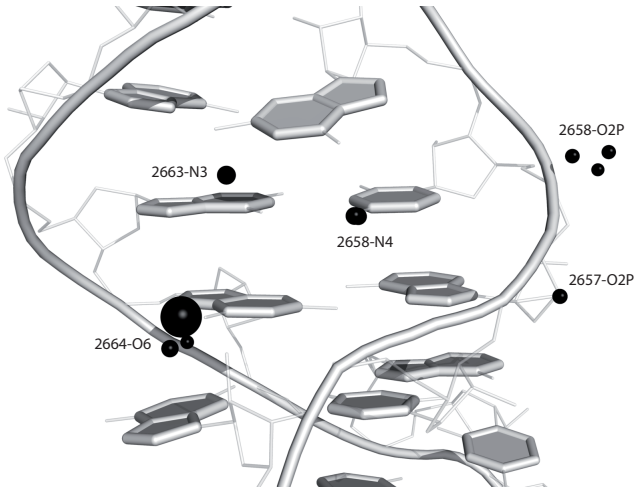


Figure 5-12 PCA analysis.
The 3D representation of the SRL centered on the NCM 5 composed by nucleotides 2657, 2658, 2663 and 2664. Nucleotides are illustrated by planar forms and phosphodiester links by cylinders. The atoms that are the most important are represents with black spheres.

5.2.5.2 Other example predictions

Domain II tested for erythromycin resistance. The LOOCV analysis reveals that the NCM containing nucleotides 1196-1198 and 1248-1250 is the most significant to discriminate sequences according to erythromycin resistance. That analysis shows that the best number of clusters to use is twenty-seven. Finally, the LOOCV allows us to correctly predict thirty-three of the forty-four sequences from the training set. Then, the PCA analysis indicates two significant atoms (O2P from nucleotide 1249 and O1P from 1250) that play a role in erythromycin resistance.

P loop tested for cell growth. The LOOCV analysis reveals that the NCM containing nucleotides 2247, 2248, 2256 and 2257 is the most significant to discriminating sequences according to cell growth. This analysis reveals that the best number of clusters is thirteen. In addition, the LOOCV enables us to accurately predict twelve of the sixteen sequences from the training set. Next, the PCA analysis indicates two significant atoms (N4 from nucleotide 2248 and O4 from 2257) that play a role in cell growth.

5.2.6 Discussion

5.2.6.1 SRL predictions

Using the *MC-Sym* software (Major, Turcotte et al. 1991), we can now predict the 3D structure of RNA sequences more easily than ever before. Accurate predictions are essential to study the electrostatic features of RNAs. Using the PCA analysis, we can now study simultaneously the effect of a large number of features and capture those that explain a phenomenon (here growth of cells bearing the RNA sequences). The combined advances in RNA structure predictions and PCA analysis allow us to develop an RNA 3D QSAR method, *MC-QSAR*.

MC-QSAR allows us to identify the electrostatic profile of active structures based on the commonly exposed charges of the 3D structures of the active sequences that are absent in the inactive sequences, but also to distinguish between among new sequences those that would be active and inactive. We model new SRL sequences (see Table A4-I) and we determine functionality by mapping them to the electrostatic profile. We correctly predict twenty-three of the twenty-four experimentally-tested sequences, therefore 96% of predictions are correct.

Beyond the activity predictions, we analyze the PCA analysis and we observe that the O6 atom from nucleotide 2664 plays a major role in cell growth (see Figure 5-12). According to that result, Spackova and Sponer identified an important ion-binding site, interconnecting the N7 atom from nucleotide 2263 and O6 from 2664 (Spackova and Sponer 2006). Moreover, the 2664 is within the site of ribosome-inactivating proteins (Yassin and Mankin 2007).

Our PCA analysis also reveals that O2P atom from nucleotide 2657 among others play a role in cell growth (see Figure 5-12). Consistent with this result, Uchiumi and co-workers recognized that the 2657 nucleotide is strongly protected by the binding of L3 and L6 proteins with the RNA (Uchiumi, Sato et al. 1999). Moreover Spackova and Sponer found a hydration site in this area between N2 atom from

nucleotide 2664 and O2P from 2657. In the crystal structure these atoms are in close contact, but during the equilibration period the distance between N2 (2664) and O2P (2657) increases and a hydration site is formed (Spackova and Sponer 2006).

In another vein, for the clusters that participate the most in the PCA analysis (data not shown), we observe that two of them are positioned near to the N7 atom of the 2663 nucleotide and the C5 atom of the 2658 nucleotide as identified by Correll and co-workers. These functional groups, which are identical in the wild-type and the viable mutation but are different in the lethal mutation, are part of the putative EF binding surface (Correll, Beneken et al. 2003).

5.2.6.2 Other example predictions

Domain II tested for erythromycin resistance. The PCA analysis indicates two significant atoms (O2P from nucleotide 1249 and O1P from 1250) that play a role in erythromycin resistance (see Figure 5-10). In agreement with this result, Nteo mentions that translation of a pentapeptide (E-peptide) in cis, encoded in the rRNA has been reported to mediate erythromycin resistance in *Escherichia coli*. This E-peptide is encoded in a short open reading frame (ORF) between nucleotides 1248 and 1265 at the junctions of domain II and III. Mutations that affect translation initiation signals of the E-peptide mini gene (Shine-Dalgarno region and initiator codon GUG) abolish erythromycin resistance, suggesting that the size of the peptide and its amino acid sequence are essential for its functions (Nteo 2006).

P loop tested for cell growth. The PCA analysis indicates two significant atoms (N4 from nucleotide 2248 and O4 from 2257) that play a role in cell growth (see Figure 5-11). In correspondence with this result, Moazed and Noller among others identified that nucleotides 2256 and 2257 are located within the P-loop of domain V of the 23 S rRNA and are protected by the 3' terminus of the P-site-bound peptidyl tRNA (Steiner, Kuechler et al. 1988, Moazed and Noller 1989, Moazed and Noller 1991, Green, Switzer et al. 1998).

5.2.7 Methods

5.2.7.1 SRL seed structure

The SRL tertiary structure of the *E. coli* 23S rRNA (PDB 2AWB B2652-B2668) is a typical SRL example, which we define as the seed structure (see Figure 5-9A). This structure is used as the template for a bacterial alignment of sequences in eight hundred and six species [Personal communication] (see Table 5-V). The sequences from the alignment that we use in this work have tested mutations (Macbeth and Wool 1999, Chan, Sitikov et al. 2000, Chan, Dresios et al. 2006, Chan and Wool 2008) between nucleotides 2655 to 2663. From this region of mutations, we choose to include three additional base pairs to this sequence to complete the seed structure: C2652●G2668, U2653●C2667, and A2654□○C2666. With these additional base pairs, the structure will be more stable for the 3D structure prediction.

Table 5-V The alignment sequences.
Sequences are identified with the number (or WT) in the first column. WT identifier is for the wild-type sequence, from the 23S rRNA of *E. coli* (B2652-B2668). The alignment of bacterial 23S rRNA sequences from residue 2652 to 2668 represents the SRL in the second column. Mutations in sequences are shown in gray in the second column. The frequency of each sequence among the eight hundred and six sequences of the alignment is indicated in the third column.

ID	Sequence	Frequency
WT	CUAGUACGAGAGGACCG	457
12	UUAGUACGAGAGGACCG	261
13	CUAGUACGAGAGGACCA	72
14	AUAGUACGAGAGGACCU	6
02	CUAGUAUGAGAAGACCG	2
15	UUAGUACGCAAGGACCG	1
16	CUUGUACGAGAGGACCG	1
17	UUUGUACGAGAGGACCA	1
18	UUAGUACGAGAGGAUUU	1
05	CUACUACGAGAGGACCG	1
19	CUAGUACGAGAGGCCCG	1
-	CUAGUACGANAGGACCG	1
-	CUAKUACGAGAGGACCG	1

The seed structure includes a base triple: G2655◀■U2656○□A2665 (see Figure 5-9) and is made of seven NCMs, numbered 1 to 7 in Figure 5-9B. NCMs were

shown to be building blocks of RNA structures (Lemieux and Major 2006, Parisien and Major 2008), and thus we use it to predict the 3D structure of SRL sequences.

5.2.7.2 Data set

The data set is made of forty-two sequences (see Table 5-III and Table A4-I). Twelve were taken from the literature (training set): the seed sequence (WT) and eleven mutants (01, 02, ..., 11) that were tested in bacteria cell growth(Macbeth and Wool 1999, Chan, Sitikov et al. 2000, Chan, Dresios et al. 2006, Chan and Wool 2008). Eight sequences (12, 13, ..., 19) were obtained from the bacterial alignment of 23S rRNAs. Eleven variants (20, 21, ..., 30) of the 2658-2663 base pair (Chan, Sitikov et al. 2000). Finally, eleven random sequences (31, 32, ..., 41) that conserve the dinucleotide composition from the seed sequence.

5.2.7.3 Prediction set

For the prediction set, we choose sequences 12, 13 and 19 from the alignment: sequences 12 and 13 have many occurrences (two hundred and seventy-one and seventy-two occurrences) while sequence 19 has only one occurrence. We also choose sequences 26, 28 and 30; they have mutations localized in the same base pair as sequences 01, 02, 03 and 04 (Chan, Sitikov et al. 2000). We include sequences 36 and 40 that are random sequences that conserve the dinucleotide composition from the seed sequence.

5.2.7.4 QSAR method

The Figure 5-13 illustrates how we determine the essential features of the biological function, here the viability and the growth of cells, from a set of sequences. To do this, we build a set of 3D models using *MC-Sym* for each sequence from the training set (see *Structure prediction* section). From this set of models, we divide each model into a set of NCMs (see *Split into NCMs* section). We clusterize atoms from each NCM to obtain a set of atom's clusters (see *Atoms clustering* section). Then we analyze the electrostatic features of each cluster considering accessible area and partial

charge of atoms within those clusters (see *Features computation* section) and we encode this information in vectors. We use the Principal Component Analysis (PCA) (Pearson 1901) to convert the electrostatic feature's vectors of each model into uncorrelated variables called principal components (see *Principal Component Analysis* section). Those principal components are used to build the electrostatic profile of the training set. Finally, to determine the activity of a new sequence (see *Activity prediction* section), we analyze the electrostatic profile of this new sequence relative to that from training set sequences.

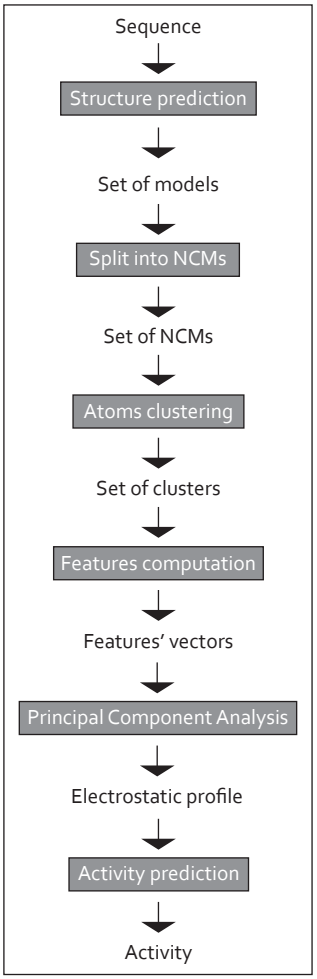


Figure 5-13 QSAR method.
The QSAR method determines the essential features of the biological function from a set of sequences. To do this, we build a set of 3D models using a structure prediction program, *MC-Sym* (Major, Turcotte et al. 1991) for each sequence. From this set of models, we split each model into a set of NCMs. We clusterize atoms from each NCM to obtain a set of the atom's clusters. Then we analyze the electrostatic features of each cluster and we encode this information in vectors. We use the PCA (Pearson 1901) to convert the electrostatic features vectors of each model into principal components and we build the electrostatic profile of the training set. Finally, to determine the activity of a new sequence, we analyze the electrostatic profile of this new sequence relative to that from training set sequences.

Structure prediction. For each sequence from the data set, we use *MC-Sym* to generate a set of 3D models that conserve base pair types from the seed structure (see Table A4-I). We refine each 3D models using energy minimization (TINKER limited memory L-BFGS) and keep the minima. For the viable sequences from the training set (WT, 02, 06, and 11), we also keep the closest model in RMSD to the seed structure

(see Figure A4-1 Training set models.), which gives us a set of eight viable and eight lethal models from the training set.

Split into NCMs. We align models together using *MC-RMSD* (Gendron, Lemieux et al. 2001) on all atoms of the structures. We divide the whole structure analysis into NCM analyses as the divide and conquer strategy. Models are split in NCMs (see Figure 5-9C) where the first NCM (called NCM 1) is composed by nucleotides 2652-2653,2667-2668; the second NCM (called NCM 2) is composed by nucleotides 2653-2654,2666-2667; etc. At this step, we obtain a set of seven NCMs, where each NCM is composed of an alignment from the data set models.

Atoms clustering. For each NCM, we use the k-means algorithm (Steinhaus, 1956) on atoms into the align models to cluster them (see Figure A4-2A). The idea of the k-means algorithm is to assign each atom into k clusters, where an atom is assigned to the cluster that minimizes the distance with the cluster centroid (center of mass, average of all atoms). In this study, we use k=13 because it is the k that maximizes the LOOCV analysis (see *Leave-One-Out analysis* section). At this step, atoms (from data set's models) of each NCM are combined into a set of thirteen clusters.

Features computation. We build a feature's vector for each model from the data set which contains the atom's electrostatic contributions of the thirteen clusters within a model. To do this for the cluster i within the model m for example, we sum the partial charge area (see **Eq. 2**) of each atom within the cluster i for the model m, see **Eq. 1**.

Eq. 1

$$PartialCharge_{i,m} = \sum_{j=1}^{NbAtoms_{i,m}} PartialChargeArea_j$$

where $PartialCharge_{i,m}$ is the partial charge of the cluster i within the model m, $NbAtoms_{i,m}$ is the number of atoms that are members of the cluster i within the model m, $PartialChargeArea_j$ is the partial charge area of atom j (see **Eq. 2**).

To calculate the partial charge area of an atom j (see **Eq. 2**), we multiply its accessible area (see example in Figure A4-2B) obtained with the *pymol* program (Delano 2002) (for a probe of water where its radius is 1.4A) with its partial charge (atomic partial charge parameters from Amber(Wang, Cieplak et al. 2000), see Table A4-II).

Eq. 2

$$PartialChargeArea_j = PartialCharge_j \times Area_j$$

where $PartialChargeArea_j$ is the partial charge area of atom j, $PartialCharge_j$ is the partial charge of atom j, $Area_j$ is the accessible area of atom j.

At this step, we obtain a feature's vector of length thirteen (from the LOOCV analysis) for each model from the data set.

Principal Component Analysis. For performing the machine learning step, we use feature's vectors from the models part of the training set. We build the electrostatic profile of active structures based on the commonly exposed charges of the active sequences that are absent in the inactive sequences. To do this, we use PCA (Pearson 1901) to convert the feature's vectors of each model into uncorrelated variables called principal components (see Figure A4-3). The first principal component is the one that represents the data more accurately (the highest variance), the second principal component is the second best representative of the data (the second highest variance), and so on. Models are now vectors of values called scores. In order to get an electrostatic profile as accurate as possible, we compute as many components as electrostatic features (here, resulting in thirteen principal components).

At this step, we obtain the electrostatic profile of models from the training set by performing the PCA on feature's vectors thereof.

Activity prediction. To determine the activity of a new sequence, we analyze the electrostatic profile of this new sequence relative to that from training set sequences. To do this, we compute the distance, using weighted Euclidean distance as shown in

Eq. 3, between the electrostatic profile of the model from the new sequence and that of models from the training set.

Eq. 3

$$Distance_{i,j} = \sum_{k=1}^{NbComp} V_k (Score_{i,k} - Score_{j,k})^2$$

where $Distance_{i,j}$ is the weighted Euclidean distance between the model i and the model j, *NbComp* is the number of components in the PCA analysis (here, *NbComp*=13), V_k is the variance of component k, and $Score_{i,k}$ is the PCA score of the model i for the component k. We determine the activity of a new sequence as the same as the activity of the closest model from the training set of this model's new sequence, using the nearest neighbor algorithm.

5.2.7.5 Leave-One-Out analysis

To select the best set of parameters (NCM analysis and number of clusters used in the k-means algorithm) for our method *MC-QSAR*, we perform the Leave-One-Out Cross-Validation (LOOCV) technique (Plutowski, Sakata et al. 1994) on electrostatic profiles from the training set. For each NCM analysis (see *Split into NCMs* section), we execute different LOOCV using one to fifteen clusters (see *Atoms clustering* section) for each NCM analysis. Among the one hundred and five LOOCV (result of the combination of seven NCMs and fifteen cluster sizes), we keep one that minimizes the prediction errors (see *Activity prediction* section). Here, one analysis produces one prediction error (seeTable 5-II), the NCM analysis for NCM 5 (see Figure 5-9C) with thirteen clusters.

5.2.7.6 Other example predictions

Domain II tested for erythromycin resistance. We applied *MC-QSAR* on thirty-six sequences from nucleotides 1196 to 1250 of the 23S *E. coli* that were tested for erythromycin resistance (Douthwaite, Powers et al. 1989, Aagaard and Douthwaite

1994, Leviev, Levieva et al. 1995, Dam, Douthwaite et al. 1996). Then, we applied LOOCV to identify the best NCM and the number of clusters to use.

P loop tested for cell growth. We applied *MC-QSAR* on sixteen sequences from nucleotides 2247 to 2257 of the 23S *E. coli* that were tested for cell growth (Gregory, Lieberman et al. 1994, Lieberman and Dahlberg 1994, O'Connor, Brunelli et al. 1995, Samaha, Green et al. 1995, Porse, Thi-Ngoc et al. 1996, Spahn, Remme et al. 1996, Green, Samaha et al. 1997, Bocchetta, Xiong et al. 1998, Gregory and Dahlberg 1999). Then, we apply LOOCV to identify the best NCM and the number of clusters to use.

5.2.7.7 Experimental methods.

Bacterial stains, plasmids and mutagenesis. *E. Coli* DH1 strain containing the thermolabile λcI repressor (referred to as DH1/cI) and the pLK45 plasmid were generously provided by the Wool lab. The pLK45 plasmid contains the *rrn*B operon under the control of the λP_L promoter, an ampicillin selection marker and a 23S mutation conferring erythromycin resistance (A2058G). An increase in temperature to 42°C induces expression of the pLK45 plasmid encoded rRNA. Mutagenesis of the 23S ribosome were performed using the QuickChange II XL mutagenesis kit from Agilents technologies according to the manufacturer's instructions.

Growth assay. DHI/cI cells with wild type or mutant 23S rRNA were grown in LB containing 50μg/mL ampicillin and 30μg/mL kanamycin at 30°C to an absorbance of 0.6 at 650nm, diluted (10^{-1} to 10^{-5}) and applied in 7μl drops on agar plates containing either 50μg/mL ampicillin and 30μg/mL kanamycin or 50μg/mL ampicillin, 50μg/mL kanamycin and 50μg/mL erythromycin. The plates were incubated at 30°C (amicillin and kanamycin) or 42°C (ampicillin and kanamycin – ampicillin, kanamycin and erythromycin) for 16-20 hours. Experiments were performed twice in duplicates and representative results are shown.

5.2.8 Acknowledgments

We would like to thank Paul Dallaire and Marc-Frédérick Blanchette for constructive discussions. We thank Eric Westhof for providing the alignment. This work was supported by grants from the Canadian Institutes of Health Research (CIHR) (MT-14604) to F.M., and from the Natural Sciences and Engineering Research Council of Canada (NSERC) (170165-01) to F.M. and (262965-2011) to S.H. S.H. holds a NSERC University Faculty Award. K.S. is supported by a scholarship from the Fonds Québécois de la Recherche sur la Nature et les Technologies.

6 Conclusion

Ce chapitre résume les principaux résultats de recherche présentés dans cette thèse et souligne leurs importances individuelles et collectives à la compréhension de la fonction des miARN. J'y discute également de l'apport de ces modélisations au domaine de la biologie moléculaire en général. Diverses avenues de recherche intéressantes constituant la suite logique des présents travaux sont ensuite décrites.

6.1 Résumé des principaux résultats de recherche

Tel que mentionné précédemment, l'objectif de cette thèse était d'identifier les cibles effectives des miARN en modélisant certains aspects de leur fonction. Deux approches complémentaires ont été utilisées et les résultats de chacune d'elles accroissent notre compréhension du rôle des miARN.

6.1.1 Modélisation des boucles de régulation miARN.

La première approche de modélisation utilisée se rapporte aux boucles de régulation entre des miARN et des FT. Cette modélisation a permis d'identifier plus de 700 de ces boucles de co-régulation. Celles-ci jettent un nouvel éclairage sur l'étendue de la régulation par les miARN pour plusieurs raisons. Tout d'abord, le nombre de boules identifiées représente une augmentation significative par rapport à ce qui est rapporté dans la littérature. De plus, on observe une grande variété dans les FT identifiés formant ces boucles. Une partie importante de ceux-ci ont été caractérisés comme des FT clés dans divers systèmes. C'est le cas, notamment, de LMO2 qui possède un rôle important dans l'hématopoïèse.

Les résultats de la validation expérimentale de ces boucles ont permis, tout d'abord de supporter la précision de la méthode informatique utilisée. Ces résultats ont aussi permis d'assigner un rôle dans le destin cellulaire hématopoïétique à deux

de celles-ci impliquant le FT LMO2 et les miARN miR-223 et miR-363. Ces travaux de recherche ont également permis d'étendre le rôle hématopoïétique de LMO2 au maintient des cellules souches hématopoïétiques et à la détermination du destin cellulaire durant les premières étapes de l'hématopoïèse. Deux aspects du rôle de LMO2 qui étaient jusqu'alors inconnus.

Cette modélisation et la validation expérimentale qui l'a suivie constituent donc une preuve de concept de la possibilité d'assigner un rôle à un miARN à partir du rôle du FT avec lequel il forme une boucle de régulation. Dans le cas présent, l'hypothèse d'un rôle hématopoïétique pour miR-223 et miR-363 corrobore avec des évidences expérimentales obtenues par d'autres équipes de recherche. La modélisation apporte donc un élément additionnel supportant l'hypothèse de leur fonction hématopoïétique.

Cette modélisation permet donc de générer des hypothèses raisonnables sur la fonction de miARN, similairement à ce qu'un criblage des miARN exprimés spécifiquement dans certains tissus permettrait de faire. Cependant, la modélisation présente plusieurs avantages par rapport à un criblage. Tout d'abord, elle peut être faite sur tous les miARN et tous les FT connus à la fois. Ceci signifie que la quantité d'information obtenue est largement supérieure à ce qu'un criblage des miARN spécifiquement exprimés permettrait d'obtenir. Un autre avantage d'une telle modélisation est que l'avancement des connaissances, lorsqu'il est inclut dans la modélisation, permet de l'améliorer, alors qu'à l'inverse, un criblage devient rapidement désuet dès que notre compréhension d'un phénomène biologique est modifiée par de nouvelles données expérimentales. Étant donné ces caractéristiques, la modélisation est un outil de choix pour mieux comprendre la biologie moléculaire.

Dans un contexte où de plus en plus d'études sont faites pour utiliser le mécanisme d'action des miARN dans un but thérapeutique, il est primordial de bien le comprendre et de cerner son étendue. L'identification de boucles de régulations

miARN/FT est un pas de plus dans la compréhension de ce mécanisme. Cette compréhension est nécessaire afin d'utiliser ce mécanisme à son plein potentiel.

L'élaboration d'une thérapie qui utiliserait, par exemple, miR-223 afin de cible un gène X pourrait tirer bénéfice de l'existence de la boucle double négative entre miR-223 et LMO2. Selon cette boucle, l'augmentation de miR-223 par un apport extérieur mène entre autres à la diminution de LMO2, un régulateur négatif de l'expression de miR-223. Ceci mène subséquemment à l'augmentation des niveaux endogènes de miR-223, ce qui pourrait contribuer à rendre une thérapie utilisant ce miARN plus efficace. Par contre, cette même boucle, dans un contexte où l'expression de LMO2 est fortement positivement régulée pourrait mener à la diminution de miR-223 et annuler son effet thérapeutique potentiel.

La présence de boucles de régulation entre miARN et FT nous porte à croire que la cellule utilise tout ce qui est à sa disposition et recycle des éléments de régulation afin de produire une variété de comportements dans divers système en n'utilisant qu'un nombre limité de composantes. Par exemple, une boucle de régulation double négative, telles que celles identifiées lors de la première modélisation, permet de n'utiliser qu'un seul signal dans la gestion de l'expression inverse d'un miARN et d'un FT. Sans ce type de boucle de régulation, la cellule devrait probablement utiliser un régulateur pour le miARN, un autre pour le FT et un senseur pour ajuster les niveaux du miARN à ceux du FT (et vice-versa). L'utilisation d'une boucle de régulation semble une solution beaucoup plus efficace. De plus, ces boucles de régulations sont, selon toute vraisemblance, elle même dans des réseaux de régulation plus complexes. Ces réseaux de régulation sont hautement connectés; en particulier lorsque l'on considère les connections indirectes (par exemple, SCL régule LMO2 qui régule miR-223 et donc SCL régule indirectement miR-223). Le réel réseau de régulation utilisé par la cellule est donc probablement plus près d'une clique (au sens mathématique du terme) que d'un arbre de régulation. Hors, beaucoup de travail a été fait dans l'esprit des relations un à un, en voyant les réseaux d'interaction comme des graphs faiblement connectés et dont les signaux sont directionnels et ne

mènent qu'à peu ou pas de rétroaction plutôt que comme des cliques. Mes travaux de recherche m'amènent à croire qu'il faut maintenant penser en terme de relations plusieurs à plusieurs. Ceci nous mènera à revoir beaucoup de méthodes expérimentales utilisées et à remettre en question plusieurs des conclusions tirées. Il faut garder en tête que lorsque plusieurs règles sont nécessaires à l'explication de divers phénomènes reliés, il est possible qu'une règle plus complexe existe pour expliquer en même temps tous ces phénomènes. Cette règle est probablement celle à mettre en application. Mes travaux de recherche m'amènent à croire que la biologie se doit d'aller vers des hypothèses englobantes plutôt que de fractionner les phénomènes et de proposer des hypothèses et des explications différentes pour chaque sous phénomène.

Récemment, diverses études ont montrées qu'il est possible de reprogrammer des cellules différenciées en cellules pluripotentes en utilisant divers FT ou miARN. La connaissance de l'existence de boucles de régulation entres des FT et des miARN jouant un rôle dans cette reprogrammation pourrait être utilisée pour la faciliter. La connaissance et l'utilisation de ces boucles pourraient permettre l'utilisation de moins de facteurs (miARN ou FT) que ce qui a été utilisé. Ultimement, il serait possible de n'utiliser qu'un seul facteur clé qui, par l'entremise de ses boucles de régulation, permettrait d'augmenter ou de diminuer un ensemble d'autres facteurs. Ceux-ci, à leur tour, modifieraient un autre ensemble de facteurs. Ultimement, la toute première modification mènerait à une cascade dirigée d'autres modifications qui utiliseraient les systèmes de régulation de la cellule pour mener à la reprogrammation souhaitée. Ce genre de manipulation de la cellule n'est possible que si l'on connaît finement les divers réseaux de régulation dont la cellule disposent et utilisés pour la reprogrammation.

Dans le cadre de cette première modélisation, nous n'avons considéré que les boucles de régulations prédites en utilisant les données génomiques de deux espèces: *Homo sapiens* et *Mus musculus*. Ce choix a été fait afin de limiter le nombre de faux positifs de la méthode. Il est connu que la prédiction de sites de liaison de FT mène à

de haut taux de faux positifs. Il en va de même de la prédiction de cibles de miARN. En ne conservant que les boucles de régulation présentes dans les deux espèces, les prédictions sont enrichies en vrais positifs par rapport aux faux positifs. Par contre, cette approche ne permet pas d'identifier des boucles qui seraient spécifiques à une ou à l'autre des espèces, en particulier celles utilisant des miARN propres à l'espèce.

Cette première modélisation a été faite en utilisant les meilleures prédictions de cibles de miARN disponibles au moment d'effectuer les travaux, soit les prédictions de TargetScan. Depuis, la modélisation des interactions miARN/ARNm menant à la prédiction de cibles de miARN a été faite et a mis en évidence les limitations des méthodes précédentes. On pourrait croire que cette deuxième modélisation invalide les travaux qui ont été faits dans la première modélisation mais ce n'est pas le cas. Au contraire, les résultats de la deuxième modélisation peuvent être utilisés afin d'améliorer la première. Les prédictions obtenues alors ne seront que meilleures puisqu'utilisant des bases plus solides. Il en est de même pour l'identification des cibles des FT. De meilleures données telles que celles tout récemment obtenues par le projet ENCODE amélioreront nécessairement les résultats de la modélisation de boucles de régulation miARN/FT.

6.1.2 Modélisation des interactions miARN/ARNm

Dans le cadre de la deuxième approche de modélisation, l'intérêt était centré sur les interactions entre les miARN et les ARNm. Ces travaux ont permis d'élaborer un modèle prédisant quels ARNm sont touchés par l'augmentation ou la diminution d'un miARN. Similairement, le modèle proposé permet de prédire les effets, à travers le réseau de miARN, du changement d'un ARNm sur tous les autres ARNm. Nous avons également pu émettre l'hypothèse que les miARNs peuvent modifier un groupe fonctionnel de gènes. Le modèle développé lors de ces travaux est unique et ses prédictions ne peuvent être faites par aucun autre outil existant.

Cette amélioration dans la précision des prédictions est due à un important changement de paradigme. Les résultats des premiers travaux de recherche de cibles

de miARN semblaient montrer que lorsqu'un miARN cible un ARNm donné dans un contexte donné, alors il le cible dans tous les contextes et chez toutes les espèces. Des travaux plus récents indiquent par contre que ceci n'est pas toujours le cas. Le nouveau paradigme qui a été développé pour la deuxième modélisation est basé sur l'hypothèse qu'un miARN peut cibler des ensembles différents d'ARNm dans des contextes différents. Ce changement de paradigme a mené à la considération des quantités de miARN et d'ARNm dans la prédiction des cibles de miARN. Le fait que la précision des prédictions augmente avec ce changement de paradigme et des évidences expérimentales récemment publiées (Khan, Betel et al. 2009, Arvey, Larsson et al. 2010, Mukherji, Ebert et al. 2011) supportent l'hypothèse à la base de celui-ci. Les conséquences de ces travaux de recherche sont importantes.

Cette modélisation met tout d'abord en évidence la dépendance des cibles d'un miARN au contexte cellulaire définit par la quantité de chacun des miARN et des ARNm. Ce phénomène a également été observé par d'autres équipes (Arvey, Larsson et al. 2010, Mukherji, Ebert et al. 2011, Bossel Ben-Moshe, Avraham et al. 2012). En d'autres mots, ces résultats impliquent qu'une cible de miARN dans un contexte ne l'est pas nécessairement dans un autre contexte. Il convient donc d'être circonspect lors de l'extrapolation de résultats expérimentaux réalisés dans un contexte vers un autre contexte. Cette dépendance au contexte cellulaire implique également qu'une observation faite dans une lignée cellulaire peut très bien ne pas être reproductible dans des cellules primaires puisque celles-ci représentent des contextes très différents d'une lignée cellulaire. Dans le même ordre d'idée, une observation faite chez la souris ne peut être facilement extrapolée à l'être humain, entre autres parce que les régions 3'UTR, des régions hautement ciblées par les miARN, ne sont généralement pas bien conservées entre les espèces. Dans un contexte de développement de thérapies basées sur le mécanisme d'action des miARN, il semble qu'il soit difficile de définir un bon système dans lequel réaliser les tests préliminaires.

Cette dépendance implique également qu'une modification du contexte cellulaire, qu'elle soit naturelle (comme dans le cas de différenciation cellulaire) ou artificielle

(par modifications des quantités de miARN ou d'ARNm dans le cadre d'expériences) mène à une réorganisation parfois très importante de l'association miARN/ARNm. Cette réorganisation signifie que l'analyse et la compréhension des résultats expérimentaux après modification du contexte cellulaire peut être ardue, voire impossible sans la connaissance du réseau d'interactions présent avant et après la modification. Il convient également de noter que cette réorganisation mène souvent à des effets secondaires sur certains ARNm. C'est le cas, par exemple, lorsque l'augmentation d'un miARN fait en sorte que celui-ci cible un ARNm à un endroit qui serait, sinon, ciblé par un autre miARN. Cet autre miARN cible alors un ARNm qui ne serait pas ciblé si l'augmentation du premier miARN n'avait été faite. L'observateur ne perçoit alors que la diminution du deuxième ARNm et conclut donc, à tort, que celui-ci est une cible du premier miARN alors que la véritable cible de ce miARN, à la base de cet effet secondaire, ne diminuant pas n'est donc pas observée. Il est cependant possible de séparer les effets directs des effets secondaires lorsque l'analyse des résultats expérimentaux est faite en combinaison de l'analyse des résultats de la même expérience faite virtuellement.

Le changement de paradigme nécessaire à la modélisation des interactions miARN/ARNm nécessite une quantification de chacun des miARN et ARNm. Pour obtenir ces quantités, nous avons utilisé les données publiques de micro-puces. Plusieurs problèmes connus sont associés à cette technologie. Tout d'abord, la quantification relative des divers ARNm est biaisée par la longueur de ceux-ci. Ce problème n'est pas présent dans la quantification des miARN puisque ceux-ci présentent tous sensiblement la même longueur. Ensuite, avec cette technologie, il n'est possible de quantifier que ce qui est connu. Ce problème est particulièrement criant dans le cas des micro-puces de miARN puisque de nouveaux miARN sont régulièrement isolés. Une solution à ce problème est d'utilisé des données de séquençage à haut débit qui peuvent identifier, dans un échantillon, des ARN encore inconnus. Par contre, la petite taille des miARN pose encore quelques difficultés techniques lors du traitement informatique des données de séquençage, entre autres

parce qu'il est difficile de départager un miARN inconnu faiblement exprimé de divers produits de dégradation de l'ARN. De plus, peu de jeux de données sont présentement disponibles où un même type cellulaire a été séquencé autant pour les ARNm que pour les miARN, idéalement sur la même plateforme.

Un autre problème, présent autant dans les résultats obtenus par micro-puce que par séquençage à haut débit est que ces expériences ne donnent que des données relatives d'un ARN par rapport aux autres. Dans notre modélisation, il serait possible de n'utiliser que des données relatives. Cependant, les deux techniques de quantification ne peuvent être faites simultanément pour les ARNm et les miARN. Les ARNm sont typiquement isolés par sélection des transcrits polyadénylés, ce qui permet d'éliminer les ARNr et les ARNt qui sont tellement abondants qu'il ne serait pas possible d'obtenir de bons résultats de quantification des ARNm sans cette étape. Cependant, les miARN ne possèdent pas cette caractéristique et doivent donc être isolé par une sélection basée sur la taille de ceux-ci. Au final, deux expériences différentes doivent donc être faites pour quantifier les ARNm et les miARN ce qui donne deux ensembles de quantifications relatives. Pour notre modélisation, ces quantifications relatives doivent être ramenées dans un même espace comparable et doivent donc être transformées (par exemple en quantifications absolues soit en nombre de copies par cellules) en considérant, entre autre, la quantité d'ARN utilisé, l'efficacité de la réaction de transcription inverse et la quantité d'ARN dans une cellule du type cellulaire à l'étude.

La modélisation faite ici prédit ce qui se passerait si une cellule ayant exactement les quantités d'ARN tel que défini dans les données d'entrées. Cependant, toutes les cellules ne sont pas absolument identiques. Pour représenter cette diversité, il serait intéressant de faire varier les quantités de chacun des transcrits (de ±10% par exemple) et de moyenner les effets de plusieurs dizaines de modélisation. Ceci permettrait d'obtenir des prédictions plus robustes.

L'un des problèmes les plus importants de la prédiction de cibles de miARN est qu'il est présentement difficile d'établir l'étendue du nombre de cibles effectivement utilisées par la cellule. Ce problème sera peut être contré par le développement de méthodes expérimentales permettant de déterminer de façon directe, précise et rapide toutes les cibles d'un miARN. Une variante de la technique du PAR-CLIP (qui ne permet pas d'assigner un miARN spécifique à un site de liaison de miARN identifié sur un ARNm) qui utiliserait des anticorps reconnaissant spécifiquement un miARN pourrait remédier à ce problème et fournir des jeux de données complets avec lesquels tester et améliorer les outils de prédictions de cibles de miARN.

6.1.3 La modélisation en tant qu'outil d'étude en biologie moléculaire

Les résultats obtenus par les deux approches de modélisation présentées à travers cette thèse ont permis de mieux comprendre le rôle et les mécanismes d'action des miARN. Ils ont aussi montré les bénéfices de l'utilisation de la modélisation à la compréhension du fonctionnement des miARN, un des domaines de recherche de la biologie moléculaire. L'utilisation de la modélisation en biologie moléculaire représente un outil supplémentaire dans la boîte à outil à laquelle le biologiste moléculaire a accès. Cette approche est relativement peu utilisée, particulièrement lorsque l'organisme d'intérêt est l'humain. Les résultats présentés ici supportent l'hypothèse que l'utilisation de la modélisation en biologie moléculaire est une approche qui puisse être utile.

6.2 Travaux futurs

La modélisation des boucles de régulation entre miARN et FT a permis d'identifier un grand nombre de boucles potentielles. Seules deux de ces boucles ont été soumises à une variété de tests fonctionnels afin d'identifier leur rôle. Plusieurs des autres boucles identifiées impliquent des FT dont au moins une partie du rôle est connu. Il serait dont possible d'identifier des rôles pour plusieurs des miARN présents

dans les prédictions de boucles de régulation. De ceux-ci, nombreux sont ceux pour lesquels il n'existe présentement aucun rôle connu, d'où l'intérêt d'utiliser la modélisation comme base à la génération d'hypothèse sur le rôle de ces miARN.

Cette première modélisation a été faite en utilisant les données génomiques d'*Homo sapiens* et de *Mus musculus* dû à un intérêt quant aux résultats obtenus pour ces deux espèces. Il serait intéressant de reproduire la modélisation en utilisant les données génomiques des diverses espèces de *Caenorhabditis* pour lesquels des données génomiques, en particulier de miARN sont disponibles, soit *C. briggsae*, *C. elegans* et *C. remanei*. La comparaison entre trois espèces très proches l'une de l'autre permettrait d'étudier la relation entre des changements biologiques (tel que la reproduction hermaphrodite de *C. briggsae* et de *C. elegans* et sexuée de *C. remanei*) et la présence ou l'absence de boucles de régulation entre miARNs et FT. De plus cette comparaison entre trois espèces très proches l'une de l'autre par rapport à une ou deux espèces éloignées (telles que *M. musculus* et *H. sapiens*) permettrait d'évaluer la rapidité d'acquisition et de perte de ces boucles de régulations par rapport à la vitesse d'acquisition et de perte de FT et de miARN. Similairement, il serait possible de comparer les boucles présentes dans les génomes de divers rongeurs (souris, rat, hamster) par rapport à celles présentes dans les génomes de divers hominidés (humain, gorille, orang-outan, chimpanzé). Ces données génomiques et l'annotation des miARN dans celles-ci ont récemment été publiées et ces analyses sont maintenant possibles.

La modélisation des interactions entre miARN et ARNm ne permet pas, pour l'instant, de prédire exactement l'effet d'un miARN sur un ARNm. Pour ce faire, il faudrait un modèle basé sur des résultats expérimentaux qui puisse prendre en considération non seulement le nombre de miARN sur un ARNm par rapport à une référence mais également l'effet de la variation de cet ARNm sur tout le microtargetome. On peut en effet penser qu'une augmentation de miARN sur un ARNm le ferait graduellement diminuer et donc augmenterait le nombre de miARN disponibles pour cibler d'autres ARNm. On peut émettre l'hypothèse que dans ce

contexte, un ou un petit ensemble d'ARNm diminuerait grandement et qu'un grand nombre d'ARNm diminuerait légèrement. Un modèle pouvant prédire ce genre de comportement permettrait de meilleures analyses des effets des miARN.

Cette modélisation considère présentement tous les miARN et tous les ARNm connus, incluant les variantes de transcrits de ces ARNm. Par contre, il existe d'autres ARNnc répertoriés qui pourraient influencer la modélisation des interactions. L'ajout de ces ARNnc ajouterait de la puissance au modèle. Pour ce faire, il faudrait notamment quantifier ces divers ARNnc et déterminer lesquels peuvent se lier à des ARNm et empêcher la liaison de miARN, lesquels peuvent avoir un effet par eux même sur les ARNm et lesquels peuvent directement affecter la fonction des miARN. De façon similaire, plusieurs protéines peuvent également être liées aux ARNm et empêcher les miARN de se lier. Certaines sont connues et pourraient être ajoutées au modèle pour améliorer la précision des prédictions.

La suite logique des deux modélisations présentées dans cette thèse est de les combiner dans un modèle dynamique. Brièvement, dans un premier temps, l'assignation des interactions entre miARN et ARNm serait faite tel que dans la deuxième modélisation de cette thèse. Ensuite, la connaissance des boucles de régulation entre miARN et FT serait utilisée en combinaison avec l'assignation miARN/ARNm pour modifier les quantités des diverses espèces d'ARN. Une nouvelle assignation miARN/ARNm pourrait alors être faite basée sur les nouvelles quantités, et ainsi de suite. Il serait intéressant de voir si ce système dynamique convergerait vers une suite stable d'états ou pas. Il pourrait aussi être possible d'évaluer la similarité des divers états en évaluant l'effet des diverses modifications d'interactions de miARN sur les divers ARNm. Il est possible qu'il existe une suite stable d'états du système qui soit tous équivalents dans le sens ou l'effet net sur chaque ARNm est nul.

L'effet des miARN sur une cible donnée est souvent subtil, de l'ordre d'une réduction de 50% des niveaux d'expression. Cet effet relativement modeste complique l'identification de cibles de miARN. Nous avons montré que de petits changements

dans les niveaux d'expression des miARN ou des ARNm peuvent mener à des changements importants du microtargetome. Dans ce contexte, il ne semble pas raisonnable de tenter d'identifier des cibles de miARN en utilisant des populations de cellules qui ne soient pas parfaitement homogène. C'est le cas notamment lorsque des cellules sont transfectées pour exprimer un ARNm rapporteur. Mes travaux de recherche m'amènent à croire que pour mieux comprendre le rôle des miARN, nous devrons analyser leurs effets dans des conditions qui soient le plus proche possible du contexte biologique et que les expériences devront se faire une cellule à la fois afin de mesurer chaque effet indépendamment et non pas de mesurer un effet moyenné sur toute une population.

En vingt ans, les miARN sont passés d'une curiosité observée chez *C. elegans* et étudiée par quelques laboratoires à une classe d'ARNnc étudiés dans des centaines de laboratoires pour leur rôle dans toutes les fonctions biologiques et les dérèglement de celles-ci, de *C. elegans* à l'humain en passant par la drosophile et la souris. Ceci reflète, selon moi, le rôle central qu'occupent les miARN dans les processus biologiques. Il ne viendrait jamais à l'esprit de quiconque de ne pas considérer le rôle potentiel des protéines impliquées dans un système biologique à l'étude. De la même manière, je crois qu'il est maintenant impossible de bien comprendre un système biologique en ne considérant que le rôle des protéines et pas celui des divers ARN, codants et non codants, possiblement impliqués.

7 Bibliographie

Aagaard, C. and S. Douthwaite (1994). "Requirement for a conserved, tertiary interaction in the core of 23S ribosomal RNA." Proc Natl Acad Sci U S A **91**(8): 2989-2993.

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

Alon, U. (2007). "Network motifs: theory and experimental approaches." Nat Rev Genet **8**(6): 450-461.

Ambros, V. (2004). "The functions of animal microRNAs." Nature **431**(7006): 350-355.

Anokye-Danso, F., C. M. Trivedi, D. Juhr, M. Gupta, Z. Cui, Y. Tian, Y. Zhang, W. Yang, P. J. Gruber, J. A. Epstein and E. E. Morrisey (2011). "Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency." Cell Stem Cell **8**(4): 376-388.

Antao, V. P., S. Y. Lai and I. Tinoco, Jr. (1991). "A thermodynamic study of unusually stable RNA and DNA hairpins." Nucleic Acids Res **19**(21): 5901-5905.

Aravin, A., D. Gaidatzis, S. Pfeffer, M. Lagos-Quintana, P. Landgraf, N. Iovino, P. Morris, M. J. Brownstein, S. Kuramochi-Miyagawa, T. Nakano, M. Chien, J. J. Russo, J. Ju, R. Sheridan, C. Sander, M. Zavolan and T. Tuschl (2006). "A novel class of small RNAs bind to MILI protein in mouse testes." Nature **442**(7099): 203-207.

Arvey, A., E. Larsson, C. Sander, C. S. Leslie and D. S. Marks (2010). "Target mRNA abundance dilutes microRNA and siRNA activity." Mol Syst Biol **6**: 363.

Bachellerie, J. P., J. Cavaille and A. Huttenhofer (2002). "The expanding snoRNA world." Biochimie **84**(8): 775-790.

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

Bailor, M. H., X. Sun and H. M. Al-Hashimi (2010). "Topology links RNA secondary structure with global conformation, dynamics, and adaptation." Science **327**(5962): 202-206.

Ban, N., P. Nissen, J. Hansen, P. B. Moore and T. A. Steitz (2000). "The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 A Resolution." Science **289**(5481): 905-920.

Barrett, T., D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov and A. Soboleva (2011). "NCBI GEO: archive for functional genomics data sets--10 years on." Nucleic Acids Res **39**(Database issue): D1005-1010.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-297.

Bartel, D. P. (2009). "MicroRNAs: target recognition and regulatory functions." Cell **136**(2): 215-233.

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

Becskei, A. and L. Serrano (2000). "Engineering stability in gene networks by autoregulation." Nature **405**(6786): 590-593.

Bejarano, F., P. Smibert and E. C. Lai (2010). "miR-9a prevents apoptosis during wing development by repressing Drosophila LIM-only." Dev Biol **338**(1): 63-73.

Ben-Ami, O., N. Pencovich, J. Lotem, D. Levanon and Y. Groner (2009). "A regulatory interplay between miR-27a and Runx1 during megakaryopoiesis." Proc Natl Acad Sci U S A **106**(1): 238-243.

Berezikov, E., W. J. Chung, J. Willis, E. Cuppen and E. C. Lai (2007). "Mammalian mirtron genes." Mol Cell **28**(2): 328-336.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). "The Protein Data Bank." Nucleic Acids Res. **28**(1): 235-242.

Bernstein, E., A. A. Caudy, S. M. Hammond and G. J. Hannon (2001). "Role for a bidentate ribonuclease in the initiation step of RNA interference." Nature **409**(6818): 363-366.

Bernstein, E., S. Y. Kim, M. A. Carmell, E. P. Murchison, H. Alcorn, M. Z. Li, A. A. Mills, S. J. Elledge, K. V. Anderson and G. J. Hannon (2003). "Dicer is essential for mouse development." Nat Genet **35**(3): 215-217.

Bissels, U., S. Wild, S. Tomiuk, A. Holste, M. Hafner, T. Tuschl and A. Bosio (2009). "Absolute quantification of microRNAs by using a universal reference." RNA **15**(12): 2375-2384.

Bocchetta, M., L. Xiong and A. S. Mankin (1998). "23S rRNA positions essential for tRNA binding in ribosomal functional sites." Proc Natl Acad Sci U S A **95**(7): 3525-3530.

Bohnsack, M. T., K. Czaplinski and D. Gorlich (2004). "Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs." Rna **10**(2): 185-191.

Bossel Ben-Moshe, N., R. Avraham, M. Kedmi, A. Zeisel, A. Yitzhaky, Y. Yarden and E. Domany (2012). "Context-specific microRNA analysis: identification of functional microRNAs and their mRNA targets." Nucleic Acids Research.

Brennecke, J., A. Stark, R. B. Russell and S. M. Cohen (2005). "Principles of microRNA-target recognition." PLoS Biol **3**(3): e85.

Brigotti, M., F. Rambelli, M. Zamboni, L. Montanaro and S. Sperti (1989). "Effect of alpha-sarcin and ribosome-inactivating proteins on the interaction of elongation factors with ribosomes." Biochem J **257**(3): 723-727.

Brodersen, D. E., W. M. Clemons, Jr., A. P. Carter, B. T. Wimberly and V. Ramakrishnan (2002). "Crystal structure of the 30 S ribosomal subunit from Thermus thermophilus: structure of the proteins and their interactions with 16 S RNA." J Mol Biol **316**(3): 725-768.

Burnett, J. C. and J. J. Rossi (2012). "RNA-based therapeutics: current progress and future prospects." Chem Biol **19**(1): 60-71.

Caballero, J., L. Fernandez, J. I. Abreu and M. Fernandez (2006). "Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of human lysozyme mutants." J Chem Inf Model **46**(3): 1255-1268.

Cabrera, M. A., I. Gonzalez, C. Fernandez, C. Navarro and M. Bermejo (2006). "A topological substructural approach for the prediction of P-glycoprotein substrates." J Pharm Sci **95**(3): 589-606.

Cai, X., C. H. Hagedorn and B. R. Cullen (2004). "Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs." Rna **10**(12): 1957-1966.

Calin, G. A., M. Ferracin, A. Cimmino, G. Di Leva, M. Shimizu, S. E. Wojcik, M. V. Iorio, R. Visone, N. I. Sever, M. Fabbri, R. Iuliano, T. Palumbo, F. Pichiorri, C. Roldo, R. Garzon, C. Sevignani, L. Rassenti, H. Alder, S. Volinia, C. G. Liu, T. J. Kipps, M. Negrini and C. M. Croce (2005). "A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia." N Engl J Med **353**(17): 1793-1801.

Calin, G. A., C. G. Liu, C. Sevignani, M. Ferracin, N. Felli, C. D. Dumitru, M. Shimizu, A. Cimmino, S. Zupo, M. Dono, M. L. Dell'Aquila, H. Alder, L. Rassenti, T. J. Kipps, F. Bullrich, M. Negrini and C. M. Croce (2004). "MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias." Proc Natl Acad Sci U S A **101**(32): 11755-11760. Epub 12004 Jul 11729.

Calin, G. A., C. Sevignani, C. D. Dumitru, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizu, S. Rattan, F. Bullrich, M. Negrini and C. M. Croce (2004). "Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers." Proc Natl Acad Sci U S A **101**(9): 2999-3004. Epub 2004 Feb 2918.

Cannone, J. J., S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Muller, N. Pande, Z. Shang, N. Yu and R. R. Gutell (2002). "The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs." BMC Bioinformatics **3**: 2.

Carleton, M., M. A. Cleary and P. S. Linsley (2007). "MicroRNAs and cell cycle regulation." Cell Cycle **6**(17): 2127-2132.

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

Chan, Y.-L., J. Dresios and I. G. Wool (2006). "A pathway for the transmission of allosteric signals in the ribosome through a network of RNA tertiary interactions." Journal of Molecular Biology **355**(5): 1014-1025.

Chan, Y.-L. and I. G. Wool (2008). "The integrity of the sarcin/ricin domain of 23 S ribosomal RNA is not required for elongation factor-independent peptide synthesis." Journal of Molecular Biology **378**(1): 12-19.

Chan, Y. L., A. S. Sitikov and I. G. Wool (2000). "The phenotype of mutations of the base-pair C2658.G2663 that closes the tetraloop in the sarcin/ricin domain of Escherichia coli 23 S ribosomal RNA." Journal of Molecular Biology **298**(5): 795-805.

Chang, P. Y., K. Draheim, M. A. Kelliher and S. Miyamoto (2006). "NFKB1 is a direct target of the TAL1 oncoprotein in human T leukemia cells." Cancer Res **66**(12): 6008-6013.

Chen, C. Z., L. Li, H. F. Lodish and D. P. Bartel (2004). "MicroRNAs modulate hematopoietic lineage differentiation." Science **303**(5654): 83-86.

Chen, X., W. Liu, C. Ambrosino, M. R. Ruocco, V. Poli, L. Romani, I. Quinto, S. Barbieri, K. L. Holmes, S. Venuta and G. Scala (1997). "Impaired generation of bone marrow B lymphocytes in mice deficient in C/EBPbeta." Blood **90**(1): 156-164.

Cheong, C., G. Varani and I. Tinoco, Jr. (1990). "Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC." Nature **346**(6285): 680-682.

Cheung, V. G., L. K. Conlin, T. M. Weber, M. Arcaro, K. Y. Jen, M. Morley and R. S. Spielman (2003). "Natural variation in human gene expression assessed in lymphoblastoid cells." Nat Genet **33**(3): 422-425.

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

Correll, C. C., J. Beneken, M. J. Plantinga, M. Lubbers and Y. L. Chan (2003). "The common and the distinctive features of the bulged-G motif based on a 1.04 A resolution RNA structure." Nucleic Acids Res **31**(23): 6806-6818.

Croft, L., D. Szklarczyk, L. J. Jensen and J. Gorodkin (2012). "Multiple independent analyses reveal only transcription factors as an enriched functional class associated with microRNAs." BMC Syst Biol **6**: 90.

Dalmay, T. and D. R. Edwards (2006). "MicroRNAs and the hallmarks of cancer." Oncogene **25**(46): 6170-6175.

Dam, M., S. Douthwaite, T. Tenson and A. S. Mankin (1996). "Mutations in domain II of 23 S rRNA facilitate translation of a 23 S rRNA-encoded pentapeptide conferring erythromycin resistance." J Mol Biol **259**(1): 1-6.

Delano, W. L. (2002). "The PyMOL Molecular Graphics System." DeLano Scientific.

Doench, J. G., C. P. Petersen and P. A. Sharp (2003). "siRNAs can function as miRNAs." Genes Dev **17**(4): 438-442.

Doench, J. G. and P. A. Sharp (2004). "Specificity of microRNA target selection in translational repression." Genes Dev **18**(5): 504-511.

Douthwaite, S., T. Powers, J. Y. Lee and H. F. Noller (1989). "Defining the structural requirements for a helix in 23 S ribosomal RNA that confers erythromycin resistance." J Mol Biol **209**(4): 655-665.

Dunham, I., A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kelllis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassman, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, M. Snyder, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R.

Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elinitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigo, R. C. Hardison, T. J. Hubbard, M. Kellis, W. J. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Starnatoyannopoulos, S. A. Tennebaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A. Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. J. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B. K. Lee, A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D. Lieb, G. E. Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan, Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J.

Vielmetter, E. C. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter, J. S. Newberry, S. E. Levy, D. M. Absher, A. Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H. Margulies, S. C. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khurana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, A. Tanzer, E. Tapanan, M. L. Tress, M. J. van Baren, N. Walters, S. Washieti, L. Wilming, A. Zadissa, Z. Zhengdong, M. Brent, D. Haussler, M. Kellis, A. Valencia, M. Gerstein, A. Raymond, R. Guigo, J. Harrow, T. J. Hubbard, S. G. Landt, S. Frietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyenger, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Larnarre-Vincent, J. Leng, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, B. Pei, D. Raha, L. Ramirez, B. Reed, J. Rozowsky, A. Sboner, M. Shi, C. Sisu, T. Slifer, H. Witt, L. Wu, X. Xu, K. K. Yan, X. Yang, K. Y. Yip, Z. Zhang, K. Struhl, S. M. Weissman, M. Gerstein, P. J. Farnham, M. Snyder, S. A. Tenebaum, L. O. Penalva, F. Doyle, S. Karmakar, S. G. Landt, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, D. Patacsil, T. Slifer, A. Victorsen, X. Yang, M. Snyder, K. P. White, T. Auer, L. Centarin, M. Eichenlaub, F. Gruhl, S. Heerman, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, Z. Weng, T. W. Whitfield, J. Wang, P. J. Collins, S. F. Aldred, N. D. Trinklein, E. C. Partridge, R. M. Myers,

J. Dekker, G. Jain, B. R. Lajoie, A. Sanyal, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. S. Hansen, L. Boatman, E. Haugen, R. Humbert, G. Jain, A. K. Johnson, E. M. Johnson, T. M. Kutyavin, B. R. Lajoie, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, P. Sabo, M. E. Sanchez, R. S. Sandstrom, A. Sanyal, A. O. Shafer, A. B. Stergachis, S. Thomas, R. E. Thurman, B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. A. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, R. Kaul, J. Dekker, J. A. Stamatoyannopoulos, I. Dunham, K. Beal, A. Brazma, P. Flicek, J. Herrero, N. Johnson, D. Keefe, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. P. Wilder, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, A. Kundaje, R. C. Hardison, W. Miller, B. Giardine, R. S. Harris, W. Wu, P. J. Bickel, B. Banfai, N. P. Boley, J. B. Brown, H. Huang, Q. Li, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, M. M. Hoffman, A. O. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, Z. Weng, S. Iyer, X. Dong, M. Greven, X. Lin, J. Wang, H. S. Xi, J. Zhuang, M. Gerstein, R. P. Alexander, S. Balasubramanian, C. Cheng, A. Harmanci, L. Lochovsky, R. Min, X. J. Mu, J. Rozowsky, K. K. Yan, K. Y. Yip and E. Birney (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

Ebert, M. S., J. R. Neilson and P. A. Sharp (2007). "MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells." Nat Methods **4**(9): 721-726.

Ebert, M. S. and P. A. Sharp (2012). "Roles for microRNAs in conferring robustness to biological processes." Cell **149**(3): 515-524.

Eis, P. S., W. Tam, L. Sun, A. Chadburn, Z. Li, M. F. Gomez, E. Lund and J. E. Dahlberg (2005). "Accumulation of miR-155 and BIC RNA in human B cell lymphomas." Proc Natl Acad Sci U S A **102**(10): 3627-3632.

Enciu, A. M., B. O. Popescu and A. Gheorghisan-Galateanu (2012). "MicroRNAs in brain development and degeneration." Mol Biol Rep **39**(3): 2243-2252.

Endo, Y., K. Mitsui, M. Motizuki and K. Tsurugi (1987). "The mechanism of action of ricin and related toxic lectins on eukaryotic ribosomes. The site and the

characteristics of the modification in 28 S ribosomal RNA caused by the toxins." J Biol Chem **262**(12): 5908-5912.

Endo, Y. and I. G. Wool (1982). "The site of action of alpha-sarcin on eukaryotic ribosomes. The sequence at the alpha-sarcin cleavage site in 28 S ribosomal ribonucleic acid." J Biol Chem **257**(15): 9054-9060.

Enright, A. J., B. John, U. Gaul, T. Tuschl, C. Sander and D. S. Marks (2003). "MicroRNA targets in Drosophila." Genome Biol **5**(1): R1.

Fazi, F., S. Racanicchi, G. Zardo, L. M. Starnes, M. Mancini, L. Travaglini, D. Diverio, E. Ammatuna, G. Cimino, F. Lo-Coco, F. Grignani and C. Nervi (2007). "Epigenetic silencing of the myelopoiesis regulator microRNA-223 by the AML1/ETO oncoprotein." Cancer Cell **12**(5): 457-466.

Fazi, F., A. Rosa, A. Fatica, V. Gelmetti, M. De Marchis, C. Nervi and I. Bozzoni (2005). "A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis." Cell **123**(5): 819-831.

Fazi, F., A. Rosa, A. Fatica, V. Gelmetti, M. L. De Marchis, C. Nervi and I. Bozzoni (2005). "A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis." Cell **123**(5): 819-831.

Felli, N., L. Fontana, E. Pelosi, R. Botta, D. Bonci, F. Facchiano, F. Liuzzi, V. Lulli, O. Morsilli, S. Santoro, M. Valtieri, G. A. Calin, C. G. Liu, A. Sorrentino, C. M. Croce and C. Peschle (2005). "MicroRNAs 221 and 222 inhibit normal erythropoiesis and erythroleukemic cell growth via kit receptor down-modulation." Proc Natl Acad Sci U S A **102**(50): 18081-18086.

Felli, N., F. Pedini, P. Romania, M. Biffoni, O. Morsilli, G. Castelli, S. Santoro, S. Chicarella, A. Sorrentino, C. Peschle and G. Marziali (2009). "MicroRNA 223-dependent expression of LMO2 regulates normal erythropoiesis." Haematologica **94**(4): 479-486.

Ferrell, J. E., Jr. (2002). "Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability." Curr Opin Cell Biol **14**(2): 140-148.

Filipowicz, W., S. N. Bhattacharyya and N. Sonenberg (2008). "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?" Nat Rev Genet **9**(2): 102-114.

Fontana, L., E. Pelosi, P. Greco, S. Racanicchi, U. Testa, F. Liuzzi, C. M. Croce, E. Brunetti, F. Grignani and C. Peschle (2007). "MicroRNAs 17-5p-20a-106a control monocytopoiesis through AML1 targeting and M-CSF receptor upregulation." Nat Cell Biol **9**(7): 775-787.

Formosa, A., A. M. Lena, E. K. Markert, S. Cortelli, R. Miano, A. Mauriello, N. Croce, J. Vandesompele, P. Mestdagh, E. Finazzi-Agro, A. J. Levine, G. Melino, S. Bernardini and E. Candi (2012). "DNA methylation silences miR-132 in prostate cancer." Oncogene.

Forrest, A. R., M. Kanamori-Katayama, Y. Tomaru, T. Lassmann, N. Ninomiya, Y. Takahashi, M. J. de Hoon, A. Kubosaki, A. Kaiho, M. Suzuki, J. Yasuda, J. Kawai, Y. Hayashizaki, D. A. Hume and H. Suzuki (2010). "Induction of microRNAs, mir-155, mir-222, mir-424 and mir-503, promotes monocytic differentiation through combinatorial regulation." Leukemia **24**(2): 460-466.

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

Fukuda, T., K. Yamagata, S. Fujiyama, T. Matsumoto, I. Koshida, K. Yoshimura, M. Mihara, M. Naitou, H. Endoh, T. Nakamura, C. Akimoto, Y. Yamamoto, T. Katagiri, C. Foulds, S. Takezawa, H. Kitagawa, K. Takeyama, B. W. O'Malley and S. Kato (2007). "DEAD-box RNA helicase subunits of the Drosha complex are required for processing of rRNA and a subset of microRNAs." Nat Cell Biol **9**(5): 604-611.

Gabb, H. A., S. R. Sanghani, C. H. Robert and C. Prevost (1996). "Finding and visualizing nucleic acid base stacking." J Mol Graph **14**(1): 6-11, 23-14.

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

Garzon, R. and C. M. Croce (2008). "MicroRNAs in normal and malignant hematopoiesis." Curr Opin Hematol **15**(4): 352-358.

Gendron, P., S. Lemieux and F. Major (2001). "Quantitative analysis of nucleic acid three-dimensional structures." Journal of Molecular Biology **308**(5): 919.

Genesereth, M. R. and N. J. Nilsson (1987). Logical Foundations of Artificial Intelligence. Los Altos, Morgan Kaufmann Publishers.

Gerstein, M. B., A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman and M. Snyder (2012). "Architecture of the human regulatory network derived from ENCODE data." Nature **489**(7414): 91-100.

Giedroc, D. P., C. A. Theimer and P. L. Nixon (2000). "Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting." Journal of Molecular Biology **298**(2): 167.

Girard, A., R. Sachidanandam, G. J. Hannon and M. A. Carmell (2006). "A germline-specific class of small RNAs binds mammalian Piwi proteins." Nature **442**(7099): 199-202.

Gonzalez-Diaz, H., S. Vilar, L. Santana, G. Podda and E. Uriarte (2007). "On the applicability of QSAR for recognition of miRNA bioorganic structures at early stages of organism and cell development: embryo and stem cells." Bioorg Med Chem **15**(7): 2544-2550.

Green, R., R. R. Samaha and H. F. Noller (1997). "Mutations at nucleotides G2251 and U2585 of 23 S rRNA perturb the peptidyl transferase center of the ribosome." J Mol Biol **266**(1): 40-50.

Green, R., C. Switzer and H. F. Noller (1998). "Ribosome-catalyzed peptide-bond formation with an A-site substrate covalently linked to 23S ribosomal RNA." Science **280**(5361): 286-289.

Gregory, S. T. and A. E. Dahlberg (1999). "Mutations in the conserved P loop perturb the conformation of two structural elements in the peptidyl transferase center of 23 S ribosomal RNA." J Mol Biol **285**(4): 1475-1483.

Gregory, S. T., K. R. Lieberman and A. E. Dahlberg (1994). "Mutations in the peptidyl transferase region of E. coli 23S rRNA affecting translational accuracy." Nucleic Acids Res **22**(3): 279-284.

Griffiths-Jones, S. (2004). "The microRNA Registry." Nucleic Acids Res **32**(Database issue): D109-111.

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

Griffiths-Jones, S., J. H. Hui, A. Marco and M. Ronshaugen (2011). "MicroRNA evolution by arm switching." EMBO Rep **12**(2): 172-177.

Griffiths-Jones, S., H. K. Saini, S. van Dongen and A. J. Enright (2008). "miRBase: tools for microRNA genomics." Nucleic Acids Res **36**(Database issue): D154-158.

Grimson, A., K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim and D. P. David P. Bartel (2007). "MicroRNA targeting specificity in mammals: determinants beyond seed pairing." Mol Cell **27**(1): 91-105.

Grivna, S. T., E. Beyret, Z. Wang and H. Lin (2006). "A novel class of small RNAs in mouse spermatogenic cells." Genes Dev **20**(13): 1709-1714.

Guil, S. and J. F. Caceres (2007). "The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a." Nat Struct Mol Biol **14**(7): 591-596.

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

Gutell, R. R., J. J. Cannone, D. Konings and D. Gautheret (2000). "Predicting U-turns in Ribosomal RNA with Comparative Sequence Analysis." J. Mol. Biol. **300**(4): 791-803.

Gutell, R. R., M. N. Schnare and M. W. Gray (1992). "A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures." Nucleic Acids Res **20 Suppl**: 2095-2109.

Guttman, M., I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn and E. S. Lander (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." Nature **458**(7235): 223-227.

Haasnoot, P. C. J., J. F. Bol and R. C. L. Olsthoorn (2003). "A plant virus replication system to assay the formation of RNA pseudotriloop motifs in RNA-protein interactions." PNAS **100**(22): 12596-12600.

Hall, P. and S. Russell (2005). "New perspectives on neoplasia and the RNA world." Hematological Oncology **23**(2): 49-53.

Hammell, C. M., X. Karp and V. Ambros (2009). "A feedback circuit involving let-7-family miRNAs and DAF-12 integrates environmental signals and developmental timing in Caenorhabditis elegans." Proc Natl Acad Sci U S A **106**(44): 18668-18673.

Han, J., Y. Lee, K. H. Yeom, J. W. Nam, I. Heo, J. K. Rhee, S. Y. Sohn, Y. Cho, B. T. Zhang and V. N. Kim (2006). "Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex." Cell **125**(5): 887-901.

Hanawalt, P. C. (2007). "Paradigms for the three rs: DNA replication, recombination, and repair." Mol Cell **28**(5): 702-707.

Hansen, T. B., T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard and J. Kjems (2013). "Natural RNA circles function as efficient microRNA sponges." Nature **495**(7441): 384-388.

Hansen, T. B., E. D. Wiklund, J. B. Bramsen, S. B. Villadsen, A. L. Statham, S. J. Clark and J. Kjems (2011). "miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA." EMBO J **30**(21): 4414-4422.

Harms, J., F. Schluenzen, R. Zarivach, A. Bashan, S. Gat, I. Agmon, H. Bartels, F. Franceschi and A. Yonath (2001). "High resolution structure of the large ribosomal subunit from a mesophilic eubacterium." Cell **107**(5): 679-688.

Hausner, T. P., J. Atmadja and K. H. Nierhaus (1987). "Evidence that the G2661 region of 23S rRNA is located at the ribosomal binding sites of both elongation factors." Biochimie **69**(9): 911-923.

Havelange, V. and R. Garzon (2010). "MicroRNAs: emerging key regulators of hematopoiesis." Am J Hematol **85**(12): 935-942.

He, L. and G. J. Hannon (2004). "MicroRNAs: small RNAs with a big role in gene regulation." Nat Rev Genet **5**(7): 522-531.

He, L., J. M. Thomson, M. T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S. W. Lowe, G. J. Hannon and S. M. Hammond (2005). "A microRNA polycistron as a potential human oncogene." Nature **435**(7043): 828-833.

Helguera, A. M., J. E. Rodriguez-Borges, X. Garcia-Mera, F. Fernandez and M. N. Cordeiro (2007). "Probing the anticancer activity of nucleoside analogues: a QSAR model approach using an internally consistent training set." J Med Chem **50**(7): 1537-1545.

Herblot, S., A. M. Steff, P. Hugo, P. D. Aplan and T. Hoang (2000). "SCL and LMO1 alter thymocyte differentiation: inhibition of E2A-HEB function and pre-T alpha chain expression." Nat Immunol **1**(2): 138-144.

Herranz, H. and S. M. Cohen (2010). "MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems." Genes Dev **24**(13): 1339-1344.

Heus, H. A. and A. Pardi (1991). "Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops." Science **253**(5016): 191-194.

Hoffmann, B., G. T. Mitchell, P. Gendron, F. Major, A. A. Andersen, R. A. Collins and P. Legault (2003). "NMR structure of the active conformation of the Varkud satellite ribozyme cleavage site." Proc. Natl Acad. Sci. U.S.A. **100**(12): 7003-7008.

Huang da, W., B. T. Sherman and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc **4**(1): 44-57.

Huang, H., A. Alexandrov, X. Chen, T. W. B. III, H. Zhang, K. Dutta and S. M. Pascal (2001). "Structure of an RNA Hairpin from HRV-14." Biochemistry **40**: 8055-8064.

Huang, H.-C., U. M. A. Nagaswamy and G. E. Fox (2005). "The application of cluster analysis in the intercomparison of loop structures in RNA." RNA **11**(4): 412-423.

Ingolia, N. T., L. F. Lareau and J. S. Weissman (2011). "Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes." Cell **147**(4): 789-802.

Iorio, M. V., P. Casalini, E. Tagliabue, S. Menard and C. M. Croce (2008). "MicroRNA profiling as a tool to understand prognosis, therapy response and resistance in breast cancer." Eur J Cancer **44**(18): 2753-2759.

Iorio, M. V., M. Ferracin, C. G. Liu, A. Veronese, R. Spizzo, S. Sabbioni, E. Magri, M. Pedriali, M. Fabbri, M. Campiglio, S. Menard, J. P. Palazzo, A. Rosenberg, P. Musiani, S. Volinia, I. Nenci, G. A. Calin, P. Querzoli, M. Negrini and C. M. Croce (2005). "MicroRNA gene expression deregulation in human breast cancer." Cancer Res **65**(16): 7065-7070.

Jackson, A. L. and P. S. Linsley (2010). "Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application." Nat Rev Drug Discov **9**(1): 57-67.

Janas, M. M., B. Wang, A. S. Harris, M. Aguiar, J. M. Shaffer, Y. V. Subrahmanyam, M. A. Behlke, K. W. Wucherpfennig, S. P. Gygi, E. Gagnon and C. D. Novina (2012). "Alternative RISC assembly: Binding and repression of microRNA-mRNA duplexes by human Ago proteins." RNA **18**(11): 2041-2055.

John, B., A. J. Enright, A. Aravin, T. Tuschl, C. Sander and D. S. Marks (2004). "Human MicroRNA targets." PLoS Biol **2**(11): e363.

Johnnidis, J. B., M. H. Harris, R. T. Wheeler, S. Stehling-Sun, M. H. Lam, O. Kirak, T. R. Brummelkamp, M. D. Fleming and F. D. Camargo (2008). "Regulation of progenitor cell proliferation and granulocyte function by microRNA-223." Nature **451**(7182): 1125-1129.

Johnston, R. J., Jr., S. Chang, J. F. Etchberger, C. O. Ortiz and O. Hobert (2005). "MicroRNAs acting in a double-negative feedback loop to control a neuronal cell fate decision." Proc Natl Acad Sci U S A **102**(35): 12449-12454.

Jossinet, F. and E. Westhof (2005). "Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure." Bioinformatics **21**(15): 3320-3321.

Jucker, F. M., H. A. Heus, P. F. Yip, E. H. Moors and A. Pardi (1996). "A network of heterogeneous hydrogen bonds in GNRA tetraloops." J Mol Biol **264**(5): 968-980.

Jucker, F. M. and A. Pardi (1995). "GNRA tetraloops make a U-turn." RNA **1**(2): 219-222.

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

Kent, O. A. and J. T. Mendell (2006). "A small piece in the cancer puzzle: microRNAs as tumor suppressors and oncogenes." Oncogene **25**(46): 6188-6196.

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

Ketting, R. F., S. E. Fischer, E. Bernstein, T. Sijen, G. J. Hannon and R. H. Plasterk (2001). "Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans." Genes Dev **15**(20): 2654-2659.

Khalil, A. M., M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander and J. L. Rinn (2009). "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression." Proc Natl Acad Sci U S A **106**(28): 11667-11672.

Khan, A. A., D. Betel, M. L. Miller, C. Sander, C. S. Leslie and D. S. Marks (2009). "Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs." Nature Biotechnology **27**(6): 549-555.

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

Kim, J., K. Inoue, J. Ishii, W. B. Vanti, S. V. Voronov, E. Murchison, G. Hannon and A. Abeliovich (2007). "A MicroRNA feedback circuit in midbrain dopamine neurons." Science **317**(5842): 1220-1224.

Klein, D. J., T. M. Schmeing, P. B. Moore and T. A. Steitz (2001). "The kink-turn: a new RNA secondary structure motif." EMBO J. **20**: 4212-4221.

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

Kluiver, J., E. Haralambieva, D. de Jong, T. Blokzijl, S. Jacobs, B. J. Kroesen, S. Poppema and A. van den Berg (2006). "Lack of BIC and microRNA miR-155 expression in primary cases of Burkitt lymphoma." Genes Chromosomes Cancer **45**(2): 147-153.

Kluiver, J., B. J. Kroesen, S. Poppema and A. van den Berg (2006). "The role of microRNAs in normal hematopoiesis and hematopoietic malignancies." Leukemia **20**(11): 1931-1936.

Kluiver, J., S. Poppema, D. de Jong, T. Blokzijl, G. Harms, S. Jacobs, B. J. Kroesen and A. van den Berg (2005). "BIC and miR-155 are highly expressed in Hodgkin, primary mediastinal and diffuse large B cell lymphomas." J Pathol **207**(2): 243-249.

Kluiver, J., A. van den Berg, D. de Jong, T. Blokzijl, G. Harms, E. Bouwman, S. Jacobs, S. Poppema and B. J. Kroesen (2006). "Regulation of pri-microRNA BIC transcription and processing in Burkitt lymphoma." Oncogene.

Knight, S. W. and B. L. Bass (2001). "A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans." Science **293**(5538): 2269-2271.

Kojima, S., T. Chiyomaru, K. Kawakami, H. Yoshino, H. Enokida, N. Nohata, M. Fuse, T. Ichikawa, Y. Naya, M. Nakagawa and N. Seki (2012). "Tumour suppressors miR-1 and miR-133a target the oncogenic function of purine nucleoside phosphorylase (PNP) in prostate cancer." Br J Cancer **106**(2): 405-413.

Kondo, J., A. Urzhumtsev and E. Westhof (2006). "Two conformational states in the crystal structure of the Homo sapiens cytoplasmic ribosomal decoding A site." Nucleic Acids Res **34**(2): 676-685.

Kota, J., R. R. Chivukula, K. A. O'Donnell, E. A. Wentzel, C. L. Montgomery, H. W. Hwang, T. C. Chang, P. Vivekanandan, M. Torbenson, K. R. Clark, J. R. Mendell and J. T. Mendell (2009). "Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model." Cell **137**(6): 1005-1017.

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

Krek, A., D. Grun, M. Poy, R. Wolf, L. Rosenberg, E. Epstein, P. MacMenamin, I. da Piedade, K. Gunsalus, M. Stoffel and N. Rajewsky (2005). "Combinatorial microRNA target predictions." Nature Genetics **37**(5): 495-500.

Krosl, G., G. He, M. Lefrancois, F. Charron, P. H. Romeo, P. Jolicoeur, I. R. Kirsch, M. Nemer and T. Hoang (1998). "Transcription factor SCL is required for c-kit expression and c-Kit function in hemopoietic cells." J Exp Med **188**(3): 439-450.

Labbaye, C., I. Spinello, M. T. Quaranta, E. Pelosi, L. Pasquini, E. Petrucci, M. Biffoni, E. R. Nuzzolo, M. Billi, R. Foa, E. Brunetti, F. Grignani, U. Testa and C. Peschle (2008). "A three-step pathway comprising PLZF/miR-146a/CXCR4 controls megakaryopoiesis." Nat Cell Biol **10**(7): 788-801.

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

Lahlil, R., E. Lecuyer, S. Herblot and T. Hoang (2004). "SCL assembles a multifactorial complex that determines glycophorin A expression." Mol Cell Biol **24**(4): 1439-1452.

Lancaster, L., N. J. Lambert, E. J. Maklan, L. H. Horan and H. F. Noller (2008). "The sarcin-ricin loop of 23S rRNA is essential for assembly of the functional core of the 50S ribosomal subunit." RNA **14**(10): 1999-2012.

Landgraf, P., M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foa, J.

Schliwka, U. Fuchs, A. Novosel, R. U. Muller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. De Vita, D. Frezzetti, H. I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan and T. Tuschl (2007). "A mammalian microRNA expression atlas based on small RNA library sequencing." Cell **129**(7): 1401-1414.

Landry, J. R., N. Bonadies, S. Kinston, K. Knezevic, N. K. Wilson, S. H. Oram, M. Janes, S. Piltz, M. Hammett, J. Carter, T. Hamilton, I. J. Donaldson, G. Lacaud, J. Frampton, G. Follows, V. Kouskoff and B. Gottgens (2009). "Expression of the leukemia oncogene Lmo2 is controlled by an array of tissue-specific elements dispersed over 100 kb and bound by Tal1/Lmo2, Ets, and Gata factors." Blood **113**(23): 5783-5792.

Lau, N., L. Lim, E. Weinstein and D. Bartel (2001). "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans." Science **294**(5543): 858-862.

Lawrie, C. H. (2007). "MicroRNAs and haematology: small molecules, big function." Br J Haematol **137**(6): 503-512.

Lécuyer, E., S. Herblot, M. Saint-Denis, R. Martin, C. G. Begley, C. Porcher, S. H. Orkin and T. Hoang (2002). "The SCL complex regulates c-kit expression in hematopoietic cells through functional interaction with Sp1." Blood **100**(7): 2430-2440.

Lecuyer, E. and T. Hoang (2004). "SCL: from the origin of hematopoiesis to stem cells and leukemia." Exp Hematol **32**(1): 11-24.

Lecuyer, E., S. Lariviere, M. C. Sincennes, A. Haman, R. Lahlil, M. Todorova, M. Tremblay, B. C. Wilkes and T. Hoang (2007). "Protein stability and transcription factor complex assembly determined by the SCL-LMO2 interaction." J Biol Chem **282**(46): 33649-33658.

Lee, E. J., Y. Gusev, J. Jiang, G. J. Nuovo, M. R. Lerner, W. L. Frankel, D. L. Morgan, R. G. Postier, D. J. Brackett and T. D. Schmittgen (2007). "Expression profiling identifies microRNA signature in pancreatic cancer." Int J Cancer **120**(5): 1046-1054.

Lee, J. C., J. J. Cannone and R. R. Gutell (2003). "The Lonepair Triloop: A New Motif in RNA Structure." Journal of Molecular Biology **325**(1): 65.

Lee, R. and V. Ambros (2001). "An extensive class of small RNAs in Caenorhabditis elegans." Science **294**(5543): 862-864.

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

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim and V. N. Kim (2003). "The nuclear RNase III Drosha initiates microRNA processing." Nature **425**(6956): 415-419.

Lee, Y., M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek and V. N. Kim (2004). "MicroRNA genes are transcribed by RNA polymerase II." Embo J **23**(20): 4051-4060.

Lemieux, S. and F. Major (2002). "RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire." Nucl. Acids Res. **30**(19): 4250-4263.

Lemieux, S. and F. Major (2006). "Automated extraction and classification of RNA tertiary structure cyclic motifs." Nucleic Acids Res **34**(8): 2340-2346.

Leontis, N. B., R. B. Altman, H. M. Berman, S. E. Brenner, J. W. Brown, D. R. Engelke, S. C. Harvey, S. R. Holbrook, F. Jossinet, S. E. Lewis, F. Major, D. H. Mathews, J. S. Richardson, J. R. Williamson and E. Westhof (2006). "The RNA Ontology Consortium: An open invitation to the RNA community." RNA **12**: 533-541.

Leontis, N. B. and E. Westhof (2001). "Geometric nomenclature and classification of RNA base pairs." RNA **7**(4): 499-512.

Lerman, Y. V., S. D. Kennedy, N. Shankar, M. Parisien, F. Major and D. H. Turner (2011). "NMR structure of a 4 x 4 nucleotide RNA internal loop from an R2 retrotransposon: identification of a three purine-purine sheared pair motif and comparison to MC-SYM predictions." RNA **17**(9): 1664-1677.

Lescoute, A., N. B. Leontis, C. Massire and E. Westhof (2005). "Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments." Nucl. Acids Res. **33**(8): 2395-2409.

Leviev, I., S. Levieva and R. A. Garrett (1995). "Role for the highly conserved region of domain IV of 23S-like rRNA in subunit-subunit interactions at the peptidyl transferase centre." Nucleic Acids Res **23**(9): 1512-1517.

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

Lewis, B. P., I.-h. Shih, M. W. Jones-Rhoades, D. P. David P. Bartel and C. B. Burge (2003). "Prediction of mammalian microRNA targets." Cell **115**(7): 787-798.

Li, Q. J., J. Chau, P. J. Ebert, G. Sylvester, H. Min, G. Liu, R. Braich, M. Manoharan, J. Soutschek, P. Skare, L. O. Klein, M. M. Davis and C. Z. Chen (2007). "miR-181a is an intrinsic modulator of T cell sensitivity and selection." Cell **129**(1): 147-161.

Li, X. and R. W. Carthew (2005). "A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the Drosophila eye." Cell **123**(7): 1267-1277.

Lieberman, K. R. and A. E. Dahlberg (1994). "The importance of conserved nucleotides of 23 S ribosomal RNA and transfer RNA in ribosome catalyzed peptide bond formation." J Biol Chem **269**(23): 16163-16169.

Liu, J., M. A. Carmell, F. V. Rivas, C. G. Marsden, J. M. Thomson, J. J. Song, S. M. Hammond, L. Joshua-Tor and G. J. Hannon (2004). "Argonaute2 is the catalytic engine of mammalian RNAi." Science **305**(5689): 1437-1441.

Liu, X., K. Fortin and Z. Mourelatos (2008). "MicroRNAs: biogenesis and molecular functions." Brain Pathol **18**(1): 113-121.

Liu, X., F. Li, E. A. Stubblefield, B. Blanchard, T. L. Richards, G. A. Larson, Y. He, Q. Huang, A. C. Tan, D. Zhang, T. A. Benke, J. R. Sladek, N. R. Zahniser and C. Y. Li (2012). "Direct reprogramming of human fibroblasts into dopaminergic neuron-like cells." Cell Res **22**(2): 321-332.

Long, D., R. Lee, P. Williams, C. Chan, V. Ambros and Y. Ding (2007). "Potent effect of target structure on microRNA function." Nat Struct Mol Biol **14**(4): 287-294.

Lowery, A. J., N. Miller, R. E. McNeill and M. J. Kerin (2008). "MicroRNAs as prognostic indicators and therapeutic targets: potential effect on breast cancer management." Clin Cancer Res **14**(2): 360-365.

Lu, J., G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub (2005). "MicroRNA expression profiles classify human cancers." Nature **435**(7043): 834-838.

Lund, E., S. Guttinger, A. Calado, J. E. Dahlberg and U. Kutay (2004). "Nuclear export of microRNA precursors." Science **303**(5654): 95-98.

Ma, J. B., K. Ye and D. J. Patel (2004). "Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain." Nature **429**(6989): 318-322.

Macbeth, M. R. and I. G. Wool (1999). "The phenotype of mutations of G2655 in the sarcin/ricin domain of 23 S ribosomal RNA." Journal of Molecular Biology **285**(3): 965-975.

Major, F. and P. Thibault (2007). RNA Tertiary Structure Prediction. Bioinformatics: From Genomes to Therapies. T. Lengauer. Weinheim, Germany, Wiley-VCH. **I:** 491-539.

Major, F., M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion and R. Cedergren (1991). "The combination of symbolic and numerical computation for three-dimensional modeling of RNA." Science. **253**(5025): 1255-1260.

Malumbres, R., K. A. Sarosiek, E. Cubedo, J. W. Ruiz, X. Jiang, R. D. Gascoyne, R. Tibshirani and I. S. Lossos (2009). "Differentiation stage-specific expression of microRNAs in B lymphocytes and diffuse large B-cell lymphomas." Blood **113**(16): 3754-3764.

Maragkakis, M., P. Alexiou, G. L. Papadopoulos, M. Reczko, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, V. A. Simossis, P. Sethupathy, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas and A. G. Hatzigeorgiou (2009). "Accurate

microRNA target prediction correlates with protein repression levels." BMC Bioinformatics **10**: 295.

Marrero Ponce, Y., J. A. Castillo Garit and D. Nodarse (2005). "Linear indices of the 'macromolecular graph's nucleotides adjacency matrix' as a promising approach for bioinformatics studies. Part 1: prediction of paromomycin's affinity constant with HIV-1 psi-RNA packaging region." Bioorg Med Chem **13**(10): 3397-3404.

Masaki, S., R. Ohtsuka, Y. Abe, K. Muta and T. Umemura (2007). "Expression patterns of microRNAs 155 and 451 during normal human erythropoiesis." Biochem Biophys Res Commun **364**(3): 509-514.

Massire, C. and E. Westhof (1998). "MANIP: an interactive tool for modelling RNA." J. Mol. Graphics Modell. **16**(4-6): 197-205, 255-257.

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

Matys, V., O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel and E. Wingender (2006). "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes." Nucleic Acids Res **34**(Database issue): D108-110.

McCallum, S. A. and A. Pardi (2003). "Refined solution structure of the iron-responsive element RNA using residual dipolar couplings." J Mol Biol. **326**(4): 1037-1050.

Medina, P. P. and F. J. Slack (2008). "microRNAs and cancer: an overview." Cell Cycle **7**(16): 2485-2492.

Memczak, S., M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble and N. Rajewsky (2013). "Circular RNAs are a large class of animal RNAs with regulatory potency." Nature **495**(7441): 333-338.

Mendell, J. T. and E. N. Olson (2012). "MicroRNAs in stress signaling and human disease." Cell **148**(6): 1172-1187.

Metzler, M., M. Wilda, K. Busch, S. Viehmann and A. Borkhardt (2004). "High expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma." Genes Chromosomes Cancer **39**(2): 167-169.

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

Moazed, D. and H. F. Noller (1989). "Interaction of tRNA with 23S rRNA in the ribosomal A, P, and E sites." Cell **57**(4): 585-597.

Moazed, D. and H. F. Noller (1991). "Sites of interaction of the CCA end of peptidyl-tRNA with 23S rRNA." Proc Natl Acad Sci U S A **88**(9): 3725-3728.

Moazed, D., J. M. Robertson and H. F. Noller (1988). "Interaction of elongation factors EF-G and EF-Tu with a conserved loop in 23S RNA." Nature **334**(6180): 362-364.

Montgomery, R. L., T. G. Hullinger, H. M. Semus, B. A. Dickinson, A. G. Seto, J. M. Lynch, C. Stack, P. A. Latimer, E. N. Olson and E. van Rooij (2011). "Therapeutic inhibition of miR-208a improves cardiac function and survival during heart failure." Circulation **124**(14): 1537-1547.

Mukherji, S., M. S. Ebert, G. X. Zheng, J. S. Tsang, P. A. Sharp and A. van Oudenaarden (2011). "MicroRNAs can generate thresholds in target gene expression." Nat Genet **43**(9): 854-859.

Muljo, S. A., K. M. Ansel, C. Kanellopoulou, D. M. Livingston, A. Rao and K. Rajewsky (2005). "Aberrant T cell differentiation in the absence of Dicer." J Exp Med **202**(2): 261-269.

Nagaswamy, U. and G. E. Fox (2002). "Frequent occurrence of the T-loop RNA folding motif in ribosomal RNAs." RNA **8**(9): 1112-1119.

Nam, C. H. and T. H. Rabbitts (2006). "The role of LMO2 in development and in T cell leukemia after chromosomal translocation or retroviral insertion." Mol Ther **13**(1): 15-25.

Nissen, P., J. A. Ippolito, N. Ban, P. B. Moore and T. A. Steitz (2001). "RNA tertiary interactions in the large ribosomal subunit: The A-minor motif." PNAS **98**(9): 4899-4903.

Nteo, M. (2006). Investigation of Novel Erythromycin Resistance Mechanisms Arising from Heterologous Expression of Gram Positive DNA in Escherichia Coli. Johannesburg, South Africa, University of Johannesburg.

O'Connell, R. M., D. S. Rao, A. A. Chaudhuri and D. Baltimore (2010). "Physiological and pathological roles for microRNAs in the immune system." Nat Rev Immunol **10**(2): 111-122.

O'Connor, M., C. A. Brunelli, M. A. Firpo, S. T. Gregory, K. R. Lieberman, J. S. Lodmell, H. Moine, D. I. Van Ryk and A. E. Dahlberg (1995). "Genetic probes of ribosomal RNA function." Biochem Cell Biol **73**(11-12): 859-868.

O'Donnell, K., E. Wentzel, K. Zeller, C. Dang and J. Mendell (2005). "c-Myc-regulated microRNAs modulate E2F1 expression." Nature **435**(7043): 839-843.

O'Donnell, K. A., E. A. Wentzel, K. I. Zeller, C. V. Dang and J. T. Mendell (2005). "c-Myc-regulated microRNAs modulate E2F1 expression." Nature **435**(7043): 839-843.

Okamura, K., J. W. Hagen, H. Duan, D. M. Tyler and E. C. Lai (2007). "The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila." Cell **130**(1): 89-100.

Olivier, C., G. Poirier, P. Gendron, A. Boisgontier, F. Major and P. Chartrand (2005). "Identification of a Conserved RNA Motif Essential for She2p Recognition and mRNA Localization to the Yeast Bud." Mol. Cell Biol. **25**(11): 4752-4766.

Olsen, L., M. Klausen, L. Helboe, F. C. Nielsen and T. Werge (2009). "MicroRNAs Show Mutually Exclusive Expression Patterns in the Brain of Adult Male Rats." PLoS ONE **4**(10): e7225.

Olsthoorn, R. C. L. and J. F. Bol (2002). "Role of an Essential Triloop Hairpin and Flanking Structures in the 3' Untranslated Region of Alfalfa Mosaic Virus RNA in In Vitro Transcription." J. Virol. **76**(17): 8747-8756.

Osella, M., C. Bosia, D. Corá and M. Caselle (2011). "The role of incoherent microRNA-mediated feedforward loops in noise buffering." PLoS Computational Biology **7**(3): e1001101.

Parisien, M. and F. Major (2008). "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data." Nature **452**(7183): 51-55.

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

Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space." Philos. Mag. **6**(2): 559-572.

Pedersen, I. and M. David (2008). "MicroRNAs in the immune response." Cytokine **43**(3): 391-394.

Perez Gonzalez, M., H. Gonzalez Diaz, R. Molina Ruiz, M. A. Cabrera and R. Ramos de Armas (2003). "TOPS-MODE based QSARs derived from heterogeneous series of compounds. Applications to the design of new herbicides." J Chem Inf Comput Sci **43**(4): 1192-1199.

Petriv, O. I., F. Kuchenbauer, A. D. Delaney, V. Lecault, A. White, D. Kent, L. Marmolejo, M. Heuser, T. Berg, M. Copley, J. Ruschmann, S. Sekulovic, C. Benz, E. Kuroda, V. Ho, F. Antignano, T. Halim, V. Giambra, G. Krystal, C. J. F. Takei, A. P. Weng, J. Piret, C. Eaves, M. A. Marra, R. K. Humphries and C. L. Hansen (2010). "Comprehensive microRNA expression profiling of the hematopoietic hierarchy." Proceedings of the National Academy of Sciences of the United States of America.

Plutowski, M., S. Sakata and H. White (1994). Cross-validation estimates IMSE. San Mateo, CA, Morgan Kaufman.

Porse, B. T., H. P. Thi-Ngoc and R. A. Garrett (1996). "The donor substrate site within the peptidyl transferase loop of 23 S rRNA and its putative interactions with the CCA-end of N-blocked aminoacyl-tRNA(Phe)." J Mol Biol **264**(3): 472-483.

Powers, T. and H. F. Noller (1990). "Dominant lethal mutations in a conserved loop in 16S rRNA." Proceedings of the National Academy of Sciences of the United States of America **87**(3): 1042-1046.

Pruitt, K. D., T. Tatusova, W. Klimke and D. R. Maglott (2009). "NCBI Reference Sequences: current status, policy and new initiatives." Nucleic Acids Res **37**(Database issue): D32-36.

Quandt, K., K. Frech, H. Karas, E. Wingender and T. Werner (1995). "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data." Nucleic Acids Res **23**(23): 4878-4884.

Ramkissoon, S. H., L. A. Mainwaring, Y. Ogasawara, K. Keyvanfar, J. P. McCoy, Jr., E. M. Sloand, S. Kajigaya and N. S. Young (2006). "Hematopoietic-specific microRNA expression in human cells." Leuk Res **30**(5): 643-647.

Rayner, K. J., C. Fernandez-Hernando and K. J. Moore (2012). "MicroRNAs regulating lipid metabolism in atherogenesis." Thromb Haemost **107**(4): 642-647.

Reczko, M., M. Maragkakis, P. Alexiou, G. L. Papadopoulos and A. G. Hatzigeorgiou (2011). "Accurate microRNA Target Prediction Using Detailed Binding Site Accessibility and Machine Learning on Proteomics Data." Front Genet **2**: 103.

Reinhart, B., F. Slack, M. Basson, A. Pasquinelli, J. Bettinger, A. Rougvie, H. Horvitz and G. Ruvkun (2000). "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans." Nature **403**(6772): 901-906.

Ro, S., C. Park, D. Young, K. M. Sanders and W. Yan (2007). "Tissue-dependent paired expression of miRNAs." Nucleic Acids Res **35**(17): 5944-5953.

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

Roldo, C., E. Missiaglia, J. P. Hagan, M. Falconi, P. Capelli, S. Bersani, G. A. Calin, S. Volinia, C. G. Liu, A. Scarpa and C. M. Croce (2006). "MicroRNA expression abnormalities in pancreatic endocrine and acinar tumors are associated with distinctive pathologic features and clinical behavior." J Clin Oncol **24**(29): 4677-4684.

Rosa, A., M. Ballarino, A. Sorrentino, O. Sthandier, F. G. De Angelis, M. Marchioni, B. Masella, A. Guarini, A. Fatica, C. Peschle and I. Bozzoni (2007). "The interplay between the master transcription factor PU.1 and miR-424 regulates human monocyte/macrophage differentiation." Proc Natl Acad Sci U S A **104**(50): 19849-19854.

Rosenfeld, N., M. B. Elowitz and U. Alon (2002). "Negative autoregulation speeds the response times of transcription networks." J Mol Biol **323**(5): 785-793.

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

Samaha, R. R., R. Green and H. F. Noller (1995). "A base pair between tRNA and 23S rRNA in the peptidyl transferase centre of the ribosome." Nature **377**(6547): 309-314.

Scaggiante, B., B. Dapas, S. Bonin, M. Grassi, C. Zennaro, R. Farra, L. Cristiano, S. Siracusano, F. Zanconati, C. Giansante and G. Grassi (2012). "Dissecting the expression of EEF1A1/2 genes in human prostate cancer cells: the potential of EEF1A2 as a hallmark for prostate transformation and progression." Br J Cancer **106**(1): 166-173.

Schluenzen, F., A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi and A. Yonath (2000). "Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution." Cell **102**(5): 615.

Schwarz, D. S., G. Hutvagner, T. Du, Z. Xu, N. Aronin and P. D. Zamore (2003). "Asymmetry in the assembly of the RNAi enzyme complex." Cell **115**(2): 199-208.

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

Selbach, M., B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin and N. Rajewsky (2008). "Widespread changes in protein synthesis induced by microRNAs." Nature **455**(7209): 58-63.

Sethupathy, P., B. Corda and A. Hatzigeorgiou (2006). "TarBase: A comprehensive database of experimentally supported animal microRNA targets." RNA **12**(2): 192-197.

Shankar, N., S. D. Kennedy, G. Chen, T. R. Krugh and D. H. Turner (2006). "The NMR structure of an internal loop from 23S ribosomal RNA differs from its structure in crystals of 50s ribosomal subunits." Biochemistry **45**(39): 11776-11789.

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

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

Small, E. M. and E. N. Olson (2011). "Pervasive roles of microRNAs in cardiovascular biology." Nature **469**(7330): 336-342.

Sneppen, K., S. Krishna and S. Semsey (2010). "Simplified models of biological networks." Annu Rev Biophys **39**: 43-59.

Socolovsky, M., H. Nam, M. D. Fleming, V. H. Haase, C. Brugnara and H. F. Lodish (2001). "Ineffective erythropoiesis in Stat5a(-/-)5b(-/-) mice due to decreased survival of early erythroblasts." Blood **98**(12): 3261-3273.

Song, J. J., J. Liu, N. H. Tolia, J. Schneiderman, S. K. Smith, R. A. Martienssen, G. J. Hannon and L. Joshua-Tor (2003). "The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes." Nat Struct Biol **10**(12): 1026-1032.

Song, L. and R. S. Tuan (2006). "MicroRNAs and cell differentiation in mammalian development." Birth Defects Res C Embryo Today **78**(2): 140-149.

Song, M. S., A. Carracedo, L. Salmena, S. J. Song, A. Egia, M. Malumbres and P. P. Pandolfi (2011). "Nuclear PTEN regulates the APC-CDH1 tumor-suppressive complex in a phosphatase-independent manner." Cell **144**(2): 187-199.

Spackova, N. and J. Sponer (2006). "Molecular dynamics simulations of sarcin-ricin rRNA motif." Nucleic Acids Res **34**(2): 697-708.

Spahn, C. M., J. Remme, M. A. Schafer and K. H. Nierhaus (1996). "Mutational analysis of two highly conserved UGG sequences of 23 S rRNA from Escherichia coli." J Biol Chem **271**(51): 32849-32856.

Steiner, G., E. Kuechler and A. Barta (1988). "Photo-affinity labelling at the peptidyl transferase centre reveals two different positions for the A- and P-sites in domain V of 23S rRNA." EMBO J **7**(12): 3949-3955.

Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics **16**(1): 16-23.

Sumazin, P., X. Yang, H.-S. Chiu, W.-J. Chung, A. Iyer, D. Llobet-Navas, P. Rajbhandari, M. Bansal, P. Guarnieri, J. Silva and A. Califano (2011). "An Extensive MicroRNA-Mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma." Cell **147**(2): 370-381.

Sun, W., W. Shen, S. Yang, F. Hu, H. Li and T. H. Zhu (2010). "miR-223 and miR-142 attenuate hematopoietic cell proliferation, and miR-223 positively regulates miR-142 through LMO2 isoforms and CEBP-beta." Cell Res **20**(10): 1158-1169.

Sylvestre, Y., V. De Guire, E. Querido, U. Mukhopadhyay, V. Bourdeau, F. Major, G. Ferbeyre and P. Chartrand (2007). "An E2F/miR-20a Autoregulatory Feedback Loop." Journal of Biological Chemistry **282**(4): 2135-2143.

Sylvestre, Y., V. De Guire, E. Querido, U. K. Mukhopadhyay, V. Bourdeau, F. Major, G. Ferbeyre and P. Chartrand (2007). "An E2F/miR-20a Autoregulatory Feedback Loop." J Biol Chem **282**(4): 2135-2143.

Szewczak, A. A. and P. B. Moore (1995). "The sarcin/ricin loop, a modular RNA." J Mol Biol **247**(1): 81-98.

Takahashi, K. and S. Yamanaka (2006). "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." Cell **126**(4): 663-676.

Tam, W. and J. E. Dahlberg (2006). "miR-155/BIC as an oncogenic microRNA." Genes Chromosomes Cancer **45**(2): 211-212.

Tay, Y., L. Kats, L. Salmena, D. Weiss, S. M. Tan, U. Ala, F. Karreth, L. Poliseno, P. Provero, F. Di Cunto, J. Lieberman, I. Rigoutsos and P. P. Pandolfi (2011). "Coding-Independent Regulation of the Tumor Suppressor PTEN by Competing Endogenous mRNAs." Cell **147**(2): 344-357.

Trang, P., P. P. Medina, J. F. Wiggins, L. Ruffino, K. Kelnar, M. Omotola, R. Homer, D. Brown, A. G. Bader, J. B. Weidhaas and F. J. Slack (2010). "Regression of murine lung tumors by the let-7 microRNA." Oncogene **29**(11): 1580-1587.

Tremblay, M., S. Herblot, E. Lecuyer and T. Hoang (2003). "Regulation of pT alpha gene expression by a dosage of E2A, HEB, and SCL." J Biol Chem **278**(15): 12680-12687.

Tsang, J., J. Zhu and A. van Oudenaarden (2007). "MicroRNA-Mediated Feedback and Feedforward Loops Are Recurrent Network Motifs in Mammals." Mol Cell **26**(5): 753-767.

Tsukada, J., K. Saito, W. R. Waterman, A. C. Webb and P. E. Auron (1994). "Transcription factors NF-IL6 and CREB recognize a common essential site in the human prointerleukin 1 beta gene." Mol Cell Biol **14**(11): 7285-7297.

Uchiumi, T., N. Sato, A. Wada and A. Hachimori (1999). "Interaction of the sarcin/ricin domain of 23 S ribosomal RNA with proteins L3 and L6." J Biol Chem **274**(2): 681-686.

van Rooij, E., L. B. Sutherland, X. Qi, J. A. Richardson, J. Hill and E. N. Olson (2007). "Control of stress-dependent cardiac growth and gene expression by a microRNA." Science **316**(5824): 575-579.

Vander Heiden, M. G., J. W. Locasale, K. D. Swanson, H. Sharfi, G. J. Heffron, D. Amador-Noguez, H. R. Christofk, G. Wagner, J. D. Rabinowitz, J. M. Asara and L. C. Cantley (2010). "Evidence for an alternative glycolytic pathway in rapidly proliferating cells." Science **329**(5998): 1492-1499.

Varani, G., B. Wimberly and I. J. Tinoco (1989). "Conformation and dynamics of an RNA internal loop." Biochemistry **28**: 7760-7772.

Velculescu, V. E., S. L. Madden, L. Zhang, A. E. Lash, J. Yu, C. Rago, A. Lal, C. J. Wang, G. A. Beaudry, K. M. Ciriello, B. P. Cook, M. R. Dufault, A. T. Ferguson, Y. Gao, T. C. He, H. Hermeking, S. K. Hiraldo, P. M. Hwang, M. A. Lopez, H. F. Luderer, B. Mathews, J. M. Petroziello, K. Polyak, L. Zawel and K. W. Kinzler (1999). "Analysis of human transcriptomes." Nature Genetics **23**(4): 387-388.

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

Vella, M. C., K. Reinert and F. J. Slack (2004). "Architecture of a validated microRNA::target interaction." Chem Biol **11**(12): 1619-1623.

Velu, C. S., A. M. Baktula and H. L. Grimes (2009). "Gfi1 regulates miR-21 and miR-196b to control myelopoiesis." Blood **113**(19): 4720-4728.

Ventura, A., A. G. Young, M. M. Winslow, L. Lintault, A. Meissner, S. J. Erkeland, J. Newman, R. T. Bronson, D. Crowley, J. R. Stone, R. Jaenisch, P. A. Sharp and T. Jacks (2008). "Targeted deletion reveals essential and overlapping functions of the miR-17 through 92 family of miRNA clusters." Cell **132**(5): 875-886.

Volinia, S., G. A. Calin, C. G. Liu, S. Ambs, A. Cimmino, F. Petrocca, R. Visone, M. Iorio, C. Roldo, M. Ferracin, R. L. Prueitt, N. Yanaihara, G. Lanza, A. Scarpa, A. Vecchione, M. Negrini, C. C. Harris and C. M. Croce (2006). "A microRNA expression signature of human solid tumors defines cancer gene targets." Proc Natl Acad Sci U S A **103**(7): 2257-2261.

Wada, R., Y. Akiyama, Y. Hashimoto, H. Fukamachi and Y. Yuasa (2010). "miR-212 is downregulated and suppresses methyl-CpG-binding protein MeCP2 in human gastric cancer." Int J Cancer **127**(5): 1106-1114.

Wadman, I. A., H. Osada, G. G. Gr|[Uuml]|Tz, A. D. Agulnick, H. Westphal, A. Forster and T. H. Rabbitts (1997). "The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins." The EMBO Journal **16**(11): 3145.

Wadman, I. A., H. Osada, G. G. Grutz, A. D. Agulnick, H. Westphal, A. Forster and T. H. Rabbitts (1997). "The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins." EMBO J **16**(11): 3145-3157.

Wang, J., P. Cieplak and P. A. Kollman (2000). "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?" Journal of Computational Chemistry **21**(12): 1049-1074.

Wang, J., M. Lu, C. Qiu and Q. Cui (2010). "TransmiR: a transcription factor-microRNA regulation database." Nucleic Acids Res **38**(Database issue): D119-122.

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

Warren, A. J., W. H. Colledge, M. B. Carlton, M. J. Evans, A. J. Smith and T. H. Rabbitts (1994). "The oncogenic cysteine-rich LIM domain protein rbtn2 is essential for erythroid development." Cell **78**(1): 45-57.

Watanabe, T., A. Takeda, T. Tsukiyama, K. Mise, T. Okuno, H. Sasaki, N. Minami and H. Imai (2006). "Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes." Genes Dev **20**(13): 1732-1743.

Wichterle, H., I. Lieberam, J. A. Porter and T. M. Jessell (2002). "Directed differentiation of embryonic stem cells into motor neurons." Cell **110**(3): 385-397.

Wightman, B., I. Ha and G. Ruvkun (1993). "Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans." Cell **75**(5): 855-862.

Wimberly, B., G. Varani and I. J. Tinoco (1993). "The conformation of loop E of eukaryotic 5S ribosomal RNA." Biochemistry **32**(4): 1078-1087.

Woese, C. R., S. Winker and R. R. Gutell (1990). "Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops"." Proc Natl Acad Sci U S A **87**(21): 8467-8471.

Xiao, C., D. P. Calado, G. Galler, T.-H. Thai, H. C. Patterson, J. Wang, N. Rajewsky, T. P. Bender and K. Rajewsky (2007). "MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb." Cell **131**(1): 146-159.

Xiao, C., L. Srinivasan, D. P. Calado, H. C. Patterson, B. Zhang, J. Wang, J. M. Henderson, J. L. Kutok and K. Rajewsky (2008). "Lymphoproliferative disease and autoimmunity in mice with increased miR-17-92 expression in lymphocytes." Nat Immunol **9**(4): 405-414.

Xiao, Z., Y. D. Xiao, J. Feng, A. Golbraikh, A. Tropsha and K. H. Lee (2002). "Antitumor agents. 213. Modeling of epipodophyllotoxin derivatives using variable selection k nearest neighbor QSAR method." J Med Chem **45**(11): 2294-2309.

Yamada, Y., A. J. Warren, C. Dobson, A. Forster, R. Pannell and T. H. Rabbitts (1998). "The T cell leukemia LIM protein Lmo2 is necessary for adult mouse hematopoiesis." Proc Natl Acad Sci U S A **95**(7): 3890-3895.

Yanaihara, N., N. Caplen, E. Bowman, M. Seike, K. Kumamoto, M. Yi, R. M. Stephens, A. Okamoto, J. Yokota, T. Tanaka, G. A. Calin, C. G. Liu, C. M. Croce and C. C. Harris (2006). "Unique microRNA molecular profiles in lung cancer diagnosis and prognosis." Cancer Cell **9**(3): 189-198.

Yang, J. S., M. D. Phillips, D. Betel, P. Mu, A. Ventura, A. C. Siepel, K. C. Chen and E. C. Lai (2011). "Widespread regulatory activity of vertebrate microRNA* species." RNA **17**(2): 312-326.

Yassin, A. and A. S. Mankin (2007). "Potential new antibiotic sites in the ribosome revealed by deleterious mutations in RNA of the large ribosomal subunit." J Biol Chem **282**(33): 24329-24342.

Yi, R., Y. Qin, I. G. Macara and B. R. Cullen (2003). "Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs." Genes Dev **17**(24): 3011-3016.

Yoo, A. S., A. X. Sun, L. Li, A. Shcheglovitov, T. Portmann, Y. Li, C. Lee-Messer, R. E. Dolmetsch, R. W. Tsien and G. R. Crabtree (2011). "MicroRNA-mediated conversion of human fibroblasts to neurons." Nature **476**(7359): 228-231.

Yu, Z., C. Wang, M. Wang, Z. Li, M. C. Casimiro, M. Liu, K. Wu, J. Whittle, X. Ju, T. Hyslop, P. McCue and R. G. Pestell (2008). "A cyclin D1/microRNA 17/20 regulatory feedback loop in control of breast cancer cell proliferation." J Cell Biol **182**(3): 509-517.

Yuan, J., F. Wang, J. Yu, G. Yang, X. Liu and J. Zhang (2008). "MicroRNA-223 reversibly regulates erythroid and megakaryocytic differentiation of K562 cells." J Cell Mol Med.

Yun, J. J., L. E. Heisler, Hwang, II, O. Wilkins, S. K. Lau, M. Hyrcza, B. Jayabalasingham, J. Jin, J. McLaurin, M. S. Tsao and S. D. Der (2006). "Genomic DNA functions as a universal external standard in quantitative real-time PCR." Nucleic Acids Res **34**(12): e85.

Zhang, J., D. Jima, C. Jacobs, R. Fischer, E. Gottwein, G. Huang, P. Lugar, A. Lagoo, D. Rizzieri, D. Friedman, J. Weinberg, P. Lipsky and S. Dave (2009). "Patterns of microRNA expression characterize stages of human B cell differentiation." Blood.

Zhao, H., A. Kalota, S. Jin and A. M. Gewirtz (2009). "The c-myb proto-oncogene and microRNA-15a comprise an active autoregulatory feedback loop in human hematopoietic cells." Blood **113**(3): 505-516.

Zhou, B., S. Wang, C. Mayr, D. P. Bartel and H. F. Lodish (2007). "miR-150, a microRNA expressed in mature B and T cells, blocks early B cell development when expressed prematurely." Proc Natl Acad Sci U S A **104**(17): 7080-7085.

Zhu, G., W. Yan, H. C. He, X. C. Bi, Z. D. Han, Q. S. Dai, Y. K. Ye, Y. X. Liang, J. Wang and W. Zhong (2009). "Inhibition of proliferation, invasion, and migration of prostate cancer cells by downregulating elongation factor-1alpha expression." Mol Med **15**(11-12): 363-370.

Annexe 1. Genome-Wide Identification of microRNAs and Transcription Factors Regulatory Loops Highlights the Role of the LMO2/miR-223 -363 Loops in Hematopoiesis Cell Fate Determination (Supplementary Information)

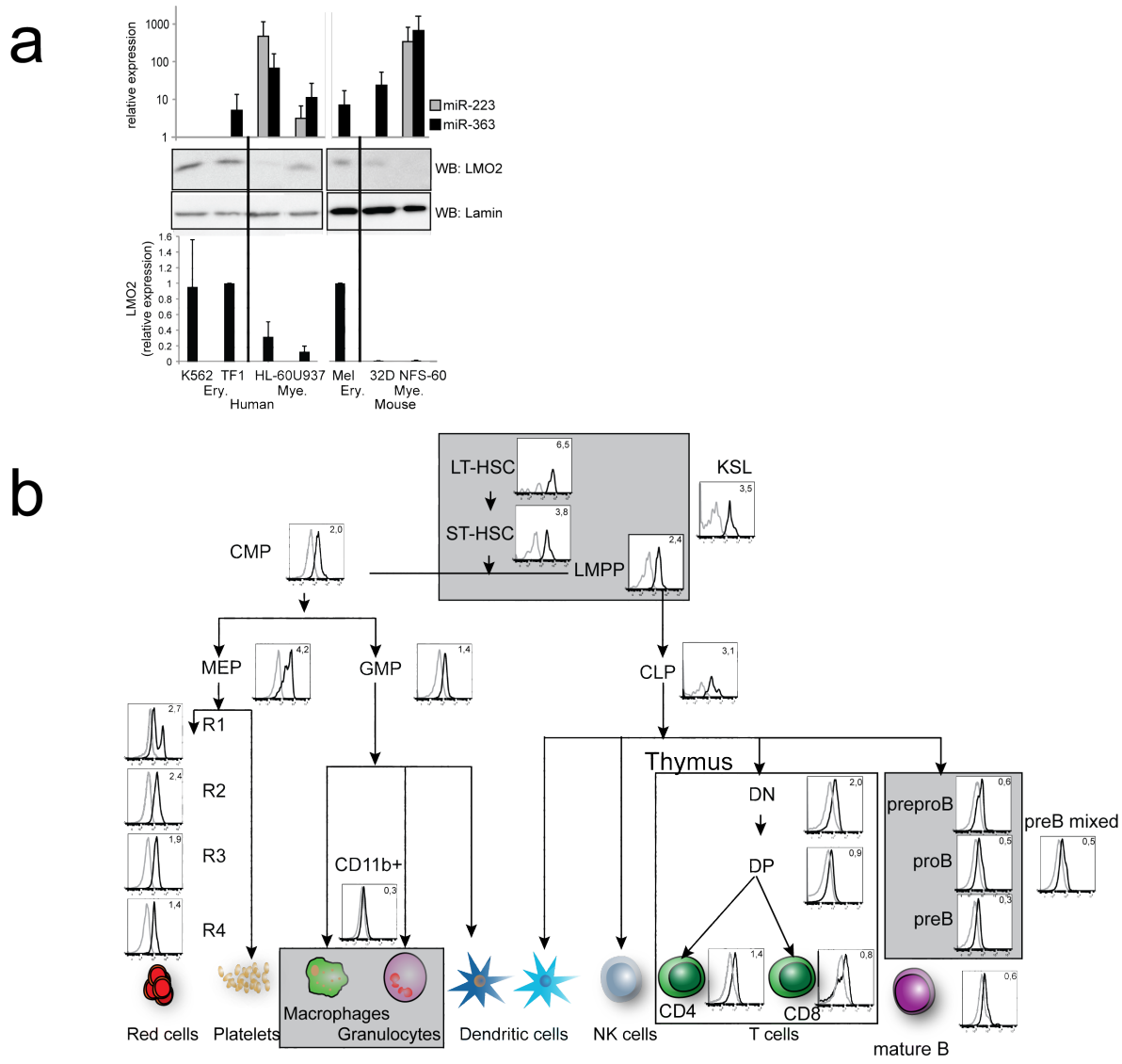


Figure A1-1 LMO2 levels in hematopoietic cell lines and primary cells.
(**a**)LMO2 levels are compared to those of miR-223 and miR-363 in a panel of hematopoietic cell lines. Cell lines with high levels of LMO2 (e.g. K562 and TF1) are low in miR-223 and miR-363. Cell lines with high levels of miR-223 or miR-363 (e.g. 32D and NFS-60) are low in protein and mRNA levels of LMO2. (**b**) LMO2 levels assayed by intracellular staining followed by flow cytometry and presented on a schematic representing the differentiation of hematopoietic stem cell in differentiated cells. The dark line represents the levels of LMO2 and the light line represents the background fluorescence.

Table A1-I List of the 779 predicted auto-regulatory loops conserved between human and mouse.

miRNA	TF	miRNA	TF	miRNA	TF
let-7b	arid3a	mir-20a	lhx6	mir-30d	six4
let-7b	hand1	mir-20a	mtf1	mir-30d	sox12
let-7b	nfat	mir-20a	nfat	mir-30d	sox4
let-7d	arid3a	mir-20a	nkx22	mir-30d	sox9
let-7d	bach1	mir-20a	nkx3-2	mir-30d	sp4
let-7d	e2f	mir-20a	nmyc	mir-30e	bach2
let-7d	gabpa	mir-20a	prrx1	mir-30e	bcl6
let-7d	hand1	mir-20a	sox4	mir-30e	e2f
let-7d	hlf	mir-20a	stat3	mir-30e	foxo3
let-7d	hoxc9	mir-20a	tal1	mir-30e	hoxa1
let-7d	meis2	mir-20a	vsx1	mir-30e	hoxa5
let-7d	myc	mir-20b	hif1a	mir-30e	irf4
let-7d	nmyc	mir-20b	stat3	mir-30e	maf
let-7d	sox13	mir-21	glis2	mir-30e	nr2f2
let-7e	arid3a	mir-21	prrx2	mir-30e	nr4a2
let-7e	gabpa	mir-21	sox2	mir-30e	runx1
let-7e	hand1	mir-21	sox5	mir-30e	six4
let-7e	hlf	mir-21	sox7	mir-30e	sox12
let-7e	hoxa9	mir-211	foxc1	mir-30e	sox4
let-7e	hoxc9	mir-211	sp1	mir-30e	sp4
let-7e	hoxd1	mir-212	arx	mir-31	bach2
let-7e	mef	mir-212	brca1	mir-31	e2f
let-7e	myc	mir-212	foxa1	mir-31	osr2
let-7e	nmyc	mir-212	foxo3	mir-32	sp1
let-7e	pou2f2	mir-212	gata2	mir-325	mzf1
let-7e	sox13	mir-212	hsf1	mir-325	nfat
let-7e	zfp2	mir-212	maf	mir-346	bcl6
let-7g	arid3a	mir-212	meis1	mir-346	klf4
let-7g	e2f	mir-212	mtf1	mir-346	mafb
let-7g	foxa2	mir-212	myc	mir-346	pax2
let-7g	gabpa	mir-212	nr4a2	mir-346	pbx1
let-7g	hand1	mir-212	pdx1	mir-346	pdx1
let-7g	hoxd1	mir-212	prrx2	mir-346	prrx2
let-7g	meis2	mir-212	rfx3	mir-346	smad3
let-7g	nfat	mir-212	six4	mir-346	zic1
let-7g	nmyc	mir-212	sox11	mir-34a	fos
let-7g	sox13	mir-212	sox4	mir-34a	nfe2
let-7i	e2f	mir-212	sox5	mir-34a	nr4a2
mir-100	sox3	mir-212	stat4	mir-34a	sox4
mir-100	sox5	mir-212	tcf7l2	mir-34a	yy1
mir-100	zbtb7b	mir-212	tead1	mir-363	gata2
mir-106a	hif1a	mir-214	esrra	mir-363	gata6
mir-106a	stat3	mir-214	foxo4	mir-363	lmo2
mir-107	arnt	mir-214	hand1	mir-363	rfx1

mir-107	bach2
mir-107	foxj2
mir-107	klf4
mir-107	myb
mir-107	six4
mir-130a	ar
mir-130a	hif1a
mir-130a	hoxa5
mir-130a	myb
mir-130a	nfat
mir-130a	sox5
mir-130a	sp1
mir-130a	zbtb7b
mir-130a	zeb1
mir-130b	ar
mir-132	arx
mir-132	brca1
mir-132	foxa1
mir-132	foxo3
mir-132	gata2
mir-132	hoxb13
mir-132	hsf1
mir-132	maf
mir-132	meis1
mir-132	mtf1
mir-132	nr4a2
mir-132	pdx1
mir-132	prrx2
mir-132	rfx3
mir-132	six4
mir-132	sox11
mir-132	sox4
mir-132	sox5
mir-132	tcf7l2
mir-132	tead1
mir-133b	foxc1
mir-133b	meis1
mir-133b	nfat
mir-133b	sp1
mir-134	nfe2
mir-134	yy1
mir-135b	sp1
mir-135b	stat6
mir-136	elk1
mir-137	e2f
mir-137	glis2
mir-137	klf4
mir-214	sox8
mir-214	sp1
mir-214	zic1
mir-215	egr1
mir-215	runx1
mir-216a	irx3
mir-216a	sp4
mir-216b	cux1
mir-216b	dlx2
mir-216b	e2f
mir-216b	ebf1
mir-216b	hlf
mir-216b	hoxa1
mir-216b	jun
mir-216b	nkx2-4
mir-216b	rfx4
mir-217	foxo4
mir-217	maf
mir-217	sox11
mir-217	zeb1
mir-217	zic1
mir-22	srf
mir-221	arnt
mir-221	e2f
mir-221	emx2
mir-221	er
mir-221	ets1
mir-221	fos
mir-221	foxa2
mir-221	foxj1
mir-221	hoxa7
mir-221	irf2
mir-222	arnt
mir-222	e2f
mir-222	er
mir-222	ets1
mir-222	fos
mir-222	hoxa7
mir-222	irf2
mir-223	elf5
mir-223	foxo3
mir-223	lmo2
mir-223	mafb
mir-223	mybl1
mir-223	pitx1
mir-223	zfx
mir-224	hoxa5
mir-363	sp1
mir-363	yy1
mir-367	e2f
mir-367	gata2
mir-367	gata6
mir-367	hand1
mir-367	hoxc8
mir-367	hoxd10
mir-367	klf4
mir-367	lmo2
mir-367	mtf1
mir-367	rfx1
mir-367	sox11
mir-367	sox4
mir-367	sp1
mir-367	tead1
mir-367	tgif
mir-367	zbtb3
mir-375	gata6
mir-376b	hoxa5
mir-376b	nfat
mir-376b	nfe2
mir-376c	nfe2
mir-376c	nr4a2
mir-376c	sox11
mir-377	arnt
mir-377	egr1
mir-377	en1
mir-377	ets1
mir-377	hoxd10
mir-377	mybl1
mir-377	pou2f3
mir-377	tead1
mir-377	yy1
mir-378	foxq1
mir-378	gfi1
mir-378	hoxb1
mir-378	mtf1
mir-378	prrx2
mir-379	foxf2
mir-379	hsf1
mir-381	ebf1
mir-381	ets1
mir-381	foxo1
mir-381	klf4
mir-381	mtf1
mir-381	sox4

mir-137	maf	mir-224	sox11	mir-381	yy1
mir-137	max	mir-23a	elf5	mir-382	ets1
mir-137	nfat	mir-23a	foxk1	mir-382	maf
mir-137	sox11	mir-23a	foxo4	mir-383	fos
mir-137	sp1	mir-23a	hoxa3	mir-383	nfe2
mir-137	zbtb7b	mir-23a	irf1	mir-383	sox11
mir-137	zic3	mir-23a	maf	mir-410	yy1
mir-141	foxa2	mir-23a	meis1	mir-411	foxo1
mir-141	gata6	mir-23a	mtf1	mir-411	sp1
mir-141	nfe2	mir-23a	nkx3-2	mir-412	stat6
mir-141	sox11	mir-23a	otp	mir-421	bach2
mir-141	sox17	mir-23a	six4	mir-421	dlx3
mir-141	sox5	mir-23a	sox11	mir-421	elk4
mir-141	stat4	mir-23a	zeb1	mir-421	hoxa5
mir-141	stat5a	mir-25	gata2	mir-421	hsf2
mir-147	stat6	mir-26b	arid3a	mir-421	mafb
mir-148b	sp1	mir-26b	e2f	mir-421	meis2
mir-149	fos	mir-26b	hoxa5	mir-421	sox3
mir-149	gsc	mir-26b	hoxa9	mir-421	sox4
mir-154	klf4	mir-26b	hoxc4	mir-421	tbp
mir-154	maf	mir-26b	hoxd13	mir-431	en1
mir-154	tcf7l2	mir-26b	hoxd3	mir-431	meis1
mir-155	bach1	mir-26b	klf4	mir-431	stat4
mir-155	cebpb	mir-26b	mnx1	mir-433	gata3
mir-155	e2f	mir-26b	nfe2	mir-448	bach2
mir-155	ehf	mir-26b	sox17	mir-448	e2f
mir-155	ets1	mir-27a	foxo1	mir-448	ebf1
mir-155	fos	mir-27a	gata2	mir-448	egr3
mir-155	hbp1	mir-27a	gata3	mir-448	foxj3
mir-155	jun	mir-27a	hoxa5	mir-448	foxo3
mir-155	lef1	mir-27a	irf4	mir-448	mtf1
mir-155	mafb	mir-27a	nfat	mir-448	otx2
mir-155	meis1	mir-27a	nfe2	mir-448	pax2
mir-155	myb	mir-27a	sox11	mir-448	sox11
mir-155	rreb1	mir-27a	sp1	mir-448	zbtb7b
mir-155	sox1	mir-27a	tbp	mir-449a	e2f
mir-155	sox10	mir-298	ar	mir-449a	usf1
mir-155	sox11	mir-298	nfat	mir-449a	yy1
mir-155	sp1	mir-29a	arnt	mir-449b	e2f
mir-155	tcf7l2	mir-29a	mafb	mir-449b	nmyc
mir-155	zic3	mir-29a	sp1	mir-449b	usf1
mir-15a	e2f	mir-29a	yy1	mir-449b	yy1
mir-15a	esrra	mir-29c	sp1	mir-452	arnt
mir-15a	hoxa3	mir-29c	yy1	mir-452	klf4
mir-15a	myb	mir-300	nfat	mir-452	zic1
mir-15a	nfat	mir-300	zic1	mir-484	nfat
mir-15a	nfe2	mir-301a	ar	mir-488	sp1

mir-15a	nfic	mir-301a	bach2	mir-489	foxj3
mir-15a	plagl1	mir-301a	e2f	mir-489	lmo2
mir-15a	pou6f1	mir-301a	hbp1	mir-495	foxc1
mir-15a	six4	mir-301a	hoxa3	mir-495	foxo1
mir-15a	smad3	mir-301a	hoxa5	mir-495	nfe2
mir-15a	sox5	mir-301a	id1	mir-496	nfe2
mir-15a	tbp	mir-301a	maf	mir-496	sox11
mir-15b	e2f	mir-301a	mafb	mir-496	srf
mir-15b	hoxa3	mir-301a	myb	mir-497	en2
mir-15b	myb	mir-301a	nfat	mir-497	esrra
mir-15b	nfat	mir-301a	otx2	mir-497	hoxa3
mir-15b	nfe2	mir-301a	sox4	mir-497	hoxd1
mir-15b	nfic	mir-301a	sp1	mir-497	myb
mir-15b	plagl1	mir-301a	zbtb7b	mir-497	mybl1
mir-15b	six4	mir-301a	zeb1	mir-497	nfat
mir-15b	smad3	mir-301b	ar	mir-497	nfe2
mir-17	brca1	mir-301b	myb	mir-497	nfic
mir-17	foxj3	mir-301b	nfat	mir-497	six4
mir-17	hif1a	mir-301b	sox4	mir-497	smad3
mir-17	mtf1	mir-301b	sox5	mir-497	sox5
mir-17	nfat	mir-301b	zeb1	mir-497	tbp
mir-17	nkx3-2	mir-302a	barx2	mir-497	tcf3
mir-17	nmyc	mir-302a	deaf1	mir-500	glis2
mir-17	nr2e3	mir-302a	e2f	mir-503	nfe2
mir-17	stat3	mir-302a	foxj3	mir-504	alx3
mir-17	vsx1	mir-302a	lhx6	mir-504	arnt
mir-181c	ets1	mir-302a	mtf1	mir-504	ebf1
mir-181c	nfat	mir-302a	nmyc	mir-504	foxf2
mir-181c	nkx3-2	mir-302a	nr2f2	mir-504	gata3
mir-181d	ets1	mir-302a	nr4a2	mir-504	nfya
mir-181d	gata6	mir-302a	rora	mir-504	pax2
mir-181d	nkx3-2	mir-302a	runx1	mir-504	pitx3
mir-182	foxo1	mir-302a	tal1	mir-504	pou2f2
mir-182	foxo3	mir-302b	alx4	mir-504	sox13
mir-182	maf	mir-302b	barx2	mir-505	foxj3
mir-182	myb	mir-302b	bcl6	mir-505	mef
mir-182	rfx3	mir-302b	e2f	mir-505	meis1
mir-182	sox5	mir-302b	foxf2	mir-505	nmyc
mir-182	xbp1	mir-302b	foxj3	mir-539	stat4
mir-183	foxo1	mir-302b	lhx6	mir-539	yy1
mir-183	hlf	mir-302b	mtf1	mir-544	maf
mir-183	mef	mir-302b	mybl1	mir-544	nfe2
mir-183	nfe2	mir-302b	nmyc	mir-551b	fos
mir-183	nr4a2	mir-302b	nr2f2	mir-568	e2f
mir-186	bach2	mir-302b	nr4a2	mir-568	ebf1
mir-186	e2f	mir-302b	rora	mir-568	ets1
mir-186	sox11	mir-302b	runx1	mir-568	gabpa

mir-186	sp4	mir-302b	tal1	mir-568	gata3
mir-186	srf	mir-302b	vsx1	mir-568	gata6
mir-186	stat4	mir-302c	e2f	mir-568	lmo2
mir-186	yy1	mir-302c	lhx6	mir-568	mybl1
mir-18a	hif1a	mir-302c	mtf1	mir-568	nmyc
mir-18a	mtf1	mir-302c	rora	mir-568	six4
mir-18a	vax2	mir-302d	barx2	mir-568	yy1
mir-18b	hif1a	mir-302d	deaf1	mir-592	foxo1
mir-191	nfe2	mir-302d	e2f	mir-592	gmeb1
mir-192	sox30	mir-302d	foxj3	mir-592	maf
mir-193b	ets1	mir-302d	lhx6	mir-592	pax2
mir-193b	nfe2	mir-302d	mtf1	mir-592	rfx1
mir-195	en2	mir-302d	nmyc	mir-592	rfx4
mir-195	esrra	mir-302d	nr2f2	mir-652	ddit3
mir-195	hoxa3	mir-302d	nr4a2	mir-652	nfat
mir-195	myb	mir-302d	rora	mir-652	rora
mir-195	nfat	mir-302d	runx1	mir-653	e2f
mir-195	nfe2	mir-302d	tal1	mir-653	fos
mir-195	nfic	mir-30a	arid3a	mir-653	nfat
mir-195	six4	mir-30a	atf1	mir-653	zbtb1
mir-195	smad3	mir-30a	bach2	mir-665	dlx3
mir-195	sox5	mir-30a	bcl6	mir-665	egr3
mir-195	tbp	mir-30a	foxo3	mir-665	irf5
mir-196b	gata6	mir-30a	hoxa5	mir-665	nkx3-2
mir-196b	hoxa5	mir-30a	irf4	mir-665	srf
mir-196b	sox11	mir-30a	irx4	mir-665	stat3
mir-196b	sox12	mir-30a	maf	mir-668	ebf1
mir-196b	tcf7	mir-30a	nr2f2	mir-708	nfat
mir-19a	dlx1	mir-30a	nr3c1	mir-744	hmbox1
mir-19a	dlx3	mir-30a	nr4a2	mir-744	nfat
mir-19a	gsc	mir-30a	pknox2	mir-744	pax2
mir-19a	klf4	mir-30a	runx1	mir-744	rfx1
mir-19a	nmyc	mir-30a	six4	mir-744	six3
mir-19a	pou3f2	mir-30a	sox12	mir-873	elk1
mir-19a	rfx1	mir-30a	sox4	mir-873	max
mir-19a	six4	mir-30a	sox9	mir-873	maz
mir-19a	sox4	mir-30a	sp4	mir-873	nfe2
mir-19a	sox5	mir-30b	arid3a	mir-873	pdx1
mir-19a	tgif	mir-30b	atf1	mir-873	tead1
mir-200c	ets1	mir-30b	bach1	mir-874	esrra
mir-200c	gata2	mir-30b	bach2	mir-874	hsf1
mir-200c	gsc	mir-30b	bcl6	mir-874	nfat
mir-200c	hmbox1	mir-30b	e2f	mir-92b	gata2
mir-200c	klf4	mir-30b	foxo3	mir-92b	gata6
mir-200c	myb	mir-30b	hlf	mir-92b	hand1
mir-200c	sox1	mir-30b	hoxa1	mir-92b	lmo2
mir-200c	sp1	mir-30b	hoxb8	mir-92b	nkx2-4

mir-200c	srf	mir-30b	hsf1	mir-92b	rfx1
mir-200c	tbp	mir-30b	irf4	mir-92b	sox11
mir-200c	zeb1	mir-30b	irx4	mir-92b	sox4
mir-203	dlx5	mir-30b	maf	mir-92b	sp1
mir-203	ebf1	mir-30b	meis2	mir-92b	tead1
mir-203	foxj3	mir-30b	nkx22	mir-92b	tgif
mir-203	foxk1	mir-30b	nr2f2	mir-96	foxo1
mir-203	hbp1	mir-30b	nr4a2	mir-96	foxo3
mir-203	hdx	mir-30b	pknox2	mir-96	foxo4
mir-203	mafb	mir-30b	runx1	mir-96	hoxa5
mir-203	smad3	mir-30b	six4	mir-96	maf
mir-204	ddit3	mir-30b	sox12	mir-96	mtf1
mir-204	er	mir-30b	sox4	mir-96	myb
mir-204	foxc1	mir-30b	sox9	mir-96	nfe2
mir-204	meis1	mir-30b	sp4	mir-96	sox5
mir-204	nr4a2	mir-30d	arid3a	mir-96	tbp
mir-204	sp1	mir-30d	bach2	mir-98	e2f
mir-206	bach2	mir-30d	bcl6	mir-98	gabpa
mir-206	ets1	mir-30d	e2f	mir-98	hand1
mir-206	meis1	mir-30d	foxo3	mir-98	myc
mir-206	nr4a2	mir-30d	hoxb8	mir-98	nmyc
mir-206	plag1	mir-30d	irf4	mir-99a	nfe2
mir-206	sp1	mir-30d	maf	mir-99a	zbtb7b
mir-208a	ets1	mir-30d	nr2f2	mir-99b	cebpb
mir-20a	alx4	mir-30d	nr4a2	mir-99b	hoxa1
mir-20a	foxj3	mir-30d	pitx1	mir-99b	id1
mir-20a	hif1a	mir-30d	pknox2	mir-99b	zbtb7b
mir-20a	irf1	mir-30d	runx1		

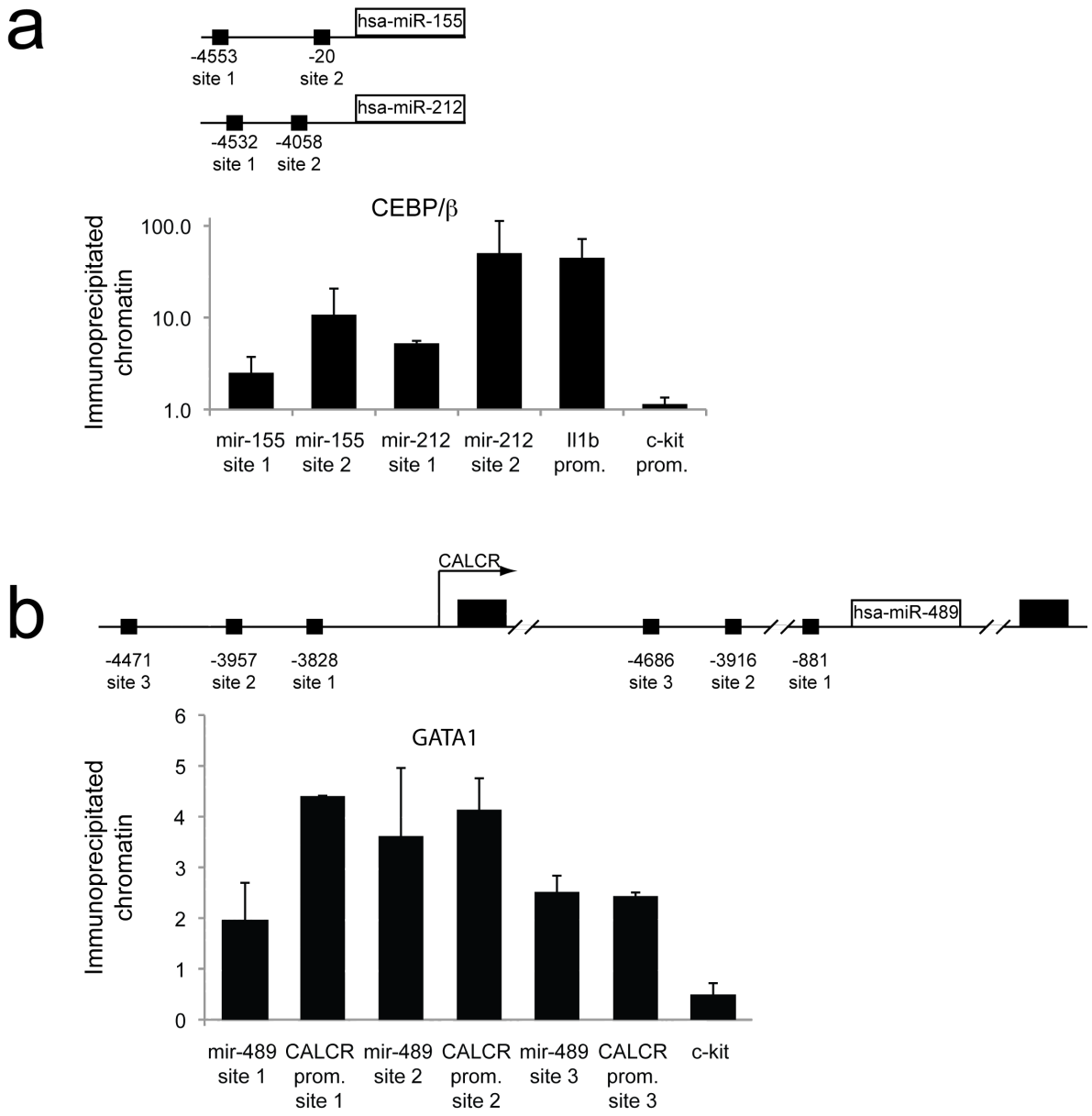


Figure A1-2 The TF binds the promoter of the miRNA in three other predicted loops
ChIP showing the binding of the TF to the promoter of the miRNA for 3 other loops predicted: the one between (a)C/EBPb and miR-155/-212 and the one between (b) GATA1 and miR-489.

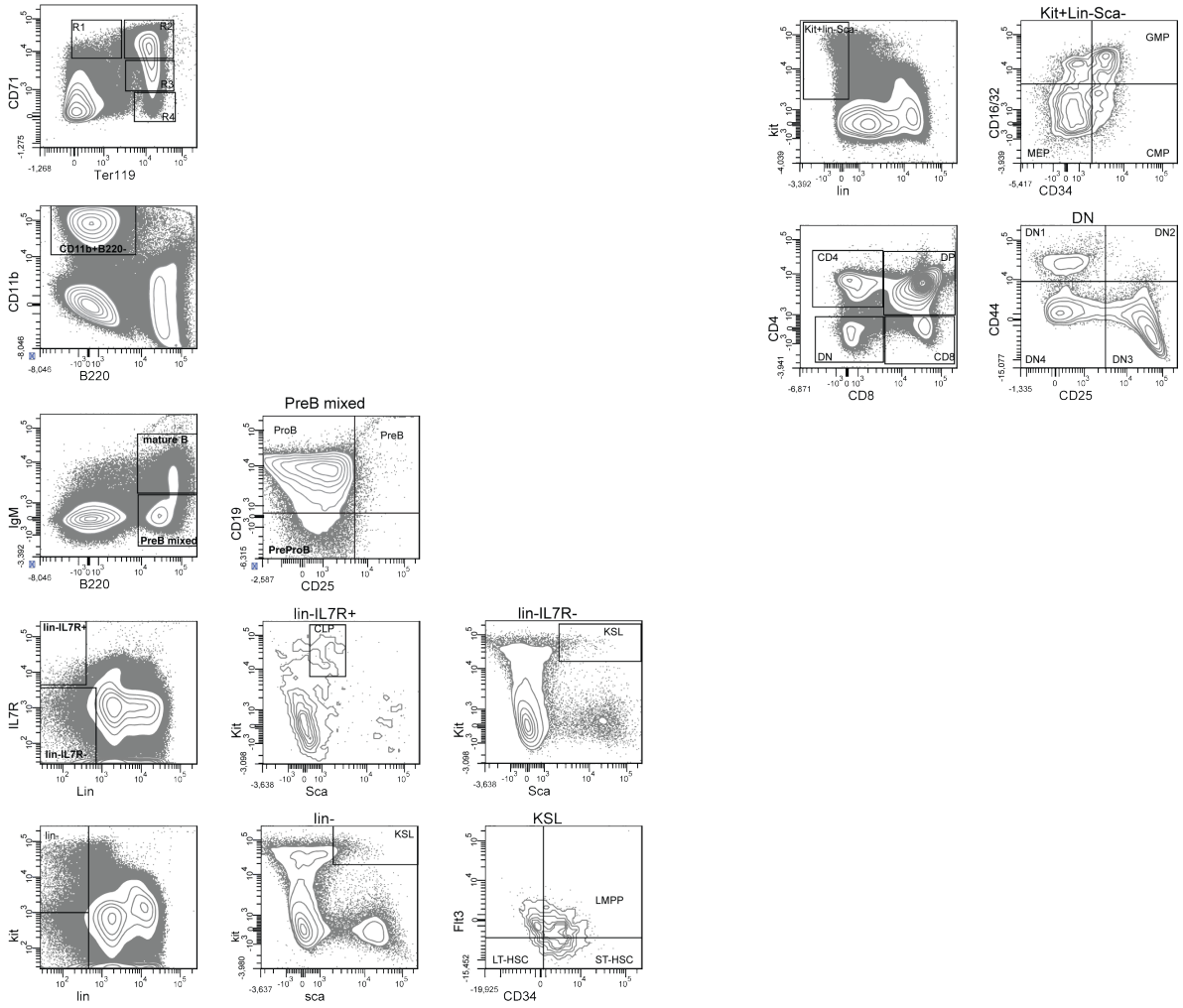


Figure A1-3 Flow cytometry gating strategy for each population assayed.

Annexe 2. Predicting and modeling the microTargetome (Methods and Supplementary Information)

Methods

Assessing factors discriminating miRNAs targeting PTEN

Receiver Operating Characteristic (ROC) curves were used to assess the scores from three factors to identify eight miRNAs (miR-17-5p, miR-19a, miR-19b, miR-20a, miR-26a, miR-93, miR-106a and miR-106b) targeting *PTEN* in DU145 cells (Tay, Kats et al. 2011). We used the area under the ROC curve (AUC) to quantify the classification potential of each factor. We considered three categories of AUC values: values <= 0.5 indicated a random classification; values between 0.5 and 0.75 indicated a poor classification; and values > 0.75 indicated a good classification. The three factors we considered were: miRNA abundance (leading to an AUC of 0.91); TargetScan (Friedman, Farh et al. 2009) conserved scores (AUC = 0.73) and various non-conserved scores (AUC between 0.41 and 0.61); and, seed hybridisation probabilities (see next Section) (AUC = 0.60).

Defining hybridisation probabilities and MREs

The mature miRNA sequences were obtained from miRBase release 18 (November 2011) (Kozomara and Griffiths-Jones 2011). The mRNA sequences were obtained from NCBI RefSeq release 49 (Pruitt, Tatusova et al. 2009). We kept only the non "predicted" mRNA sequences. The seed miRNA sequences correspond to nucleotides 2 to 8 of the mature miRNAs. The seed of all mature miRNAs were mapped to all mRNA positions in the 5'UTR, CDS, and 3'UTR. The free energies of the

seed miRNA-mRNA duplexes, ΔG, were calculated using the MC-Fold software (Parisien and Major 2008), which incorporates the energy contributions of non-canonical base pairs possibly forming at mismatch positions. The energy contribution of each base pair was multiplied by its probability to adopt the Watson-Crick geometry. The assumption being that any other base pairing geometry would interfere with the duplex formation. The duplex energies were normalised assuming a Boltzmann distribution, and converted in hybridisation probabilities using the following: . The hybridisation probability of duplex *d*, between a seed miRNA and a mRNA target site, is given by the ratio of the power of its ΔG divided by kT, where k is the Boltzmann constant and T the temperature of the system, over the sum of the power ΔG divided by kT of all possible seeds. Thus, the sum of all hybridisation probabilities for any given seed equals 1. We kept as miRNA recognition elements (MREs) the sites that defined at most one mismatch, resulting in a list of ~300,000,000 MREs. The MRE hybridisation probability distribution and their mRNA localizations are shown in Figure A2-1. It is noteworthy that the heptamer distribution has a bias toward the 3'UTRs (42%), compared to the sequence fraction in the RefSeq 3'UTRs (36%). The distributions of the MREs predicted by our algorithm (see below) in the three cellular contexts are similar to that of the heptamers.

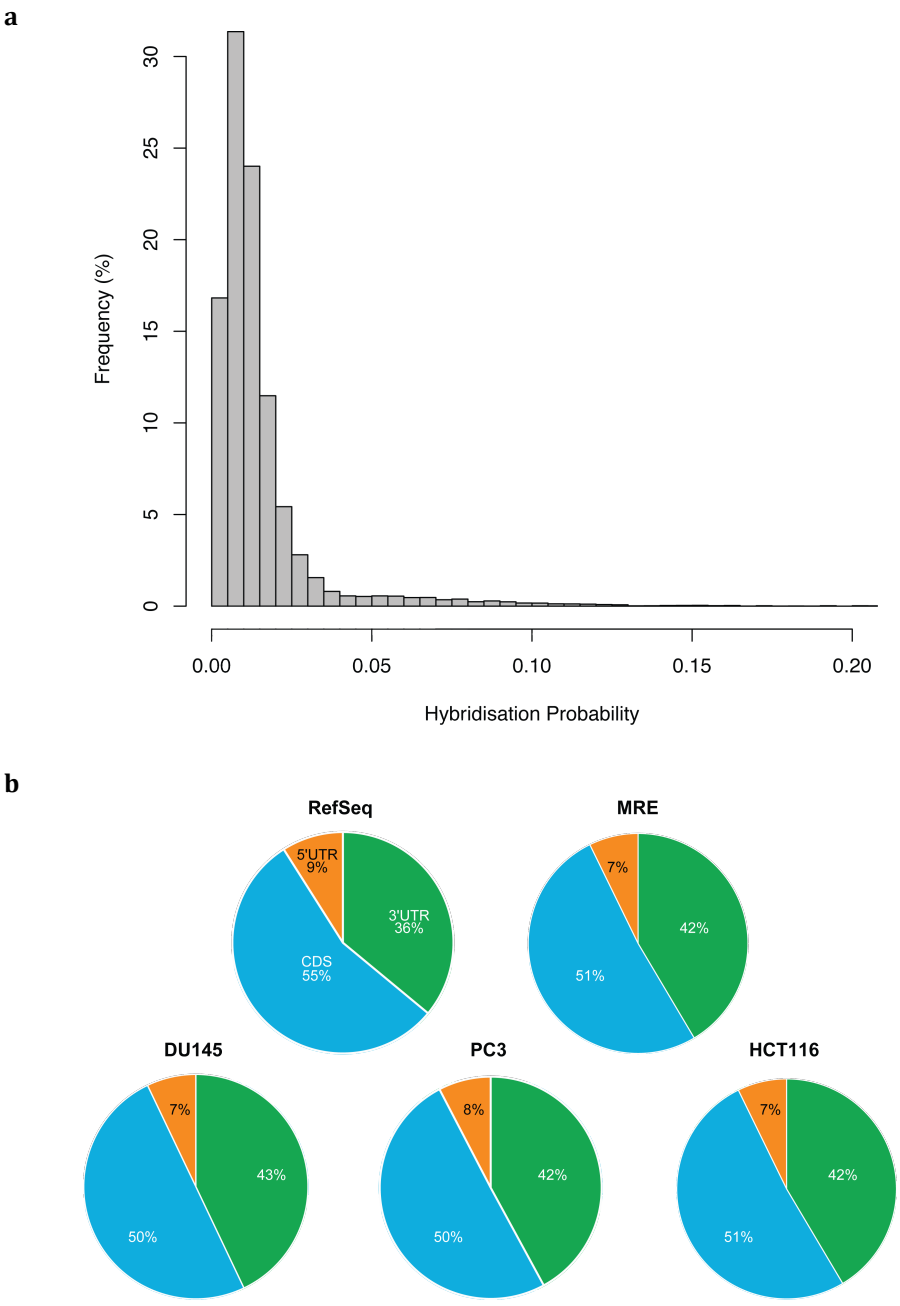


Figure A2-1 MRE probability and mRNA localization distributions.
a, Frequencies of the hybridisation probability values of all MREs with at most one mismatch. **b**, Sequence vs. MRE distributions in RefSeq (top), and predicted occupied MRE distributions in DU145, PC3 and HCT116 cells (bottom).

Determining RNA abundance

mRNA. We used the known absolute quantifications of eight mRNAs in human CD4+ and CD8+ cells (Yun, Heisler et al. 2006). These allowed us to use a linear regression of *log* transformed data ($R^2 = 0.5$) to convert a set of relative gene expression data into absolute quantities (Yun, Heisler et al. 2006). These absolute expression data were used as a reference to convert any other gene expression data into absolute quantities by fitting the mean and the standard deviation of the *log* transformed data of the considered gene expression to the reference.

miRNA. Absolute quantifications of miRNA expression data were determined from liver samples (Bissels, Wild et al. 2009). It was observed that the median and the maximum quantities of miRNAs are respectively 633 and 52,567 copies per cell. These values were used to transform microarray expression data into absolute quantities using a linear fitting of the *log* transformed data to adopt the same median and maximum.

Datasets. We used the GSE4959 samples from the Gene Expression Omibus (Barrett, Troup et al. 2011) (GEO) to quantify the DU145, HCT116, and PC3 cell lines. For each cell line, the hybridisation of cDNA was performed on Affymetrix Human Genome U95 Array, a combination of five microarrays. To avoid a bias among the five datasets, we realigned the expression data using the log transformed expression value of *GAPDH* and *ACTB* genes. For the probes matching more than one mRNA: if the probe matched several transcripts of the same gene, then we considered the longest 3'UTRs of this gene as defined by the RefSeq (Pruitt, Tatusova et al. 2009) annotation; and, we did not consider the probes that bind different genes.

Designing MiRBooking algorithm

Determining miRNA booking quantities. The miRNA assignment equation we used in miRBooking accounts for the power law observed in miRNA activity dilution, where

the log base ⟨ is linked to the kinetic relationship of targeted genes as a function of their abundance (Arvey, Larsson et al. 2010):

$$q\left(miRNA_{booking}\right) = q\left(mRNA\right) \times \log_{\alpha}\left(q\left(miRNA\right)\right) \times hp.$$

The region on a mRNA that is considered occupied when a miRNA is booked to a MRE is 46 nucleotides centred on the miRNA. This 46-nucleotide window is to account for a minimum footprint span of Ago1 binding site on an mRNA (Chi, Zang et al. 2009). Our algorithm establishes the matches in respectively O(n^2) and O(n) time and space complexities, where n is the number of MREs.

Determining a hybridisation probability threshold. We established a threshold hybridisation probability (hp = 0.0118) as that of the highest probability of Pandolfi and co-workers negative control, miR-191, that was shown to be absent on an overexpressed *PTEN*-reporter gene in DU145 cells (Tay, Kats et al. 2011).

Calibrating miRBooking. The *log* base, ⟨, was determined so that it allowed for the observation of a *PTEN* upregulation when a series of *PTEN*'s competitive endogenous mRNAs (ceRNAs) were overexpressed. We simulated the overexpression of an RNA (either mRNA or miRNA) in a given cellular context by increasing its quantity while keeping the original quantities for the others. We increased the quantities of overexpressed ceRNAs one order of magnitude over the average endogenous expression level of mRNAs. For a variety of powers of 2 for ⟨, 2^1 to 2^{15}, we overexpressed in turn three *PTEN* ceRNAs, i.e. whose expressions were shown by Western Blot by Pandolfi and co-workers to influence that of *PTEN*: *CNOT6L*, *VAPA*, and *SERINC1* (Tay, Kats et al. 2011). We identified 2^{14} as the best value for ⟨. This is the only value for which overexpressions of these three ceRNAs decreased *PTEN*'s occupancies (number of miRNAs targeting *PTEN*) and presence of eight miRNAs directly targeting a transfected *PTEN* reporter in DU145 cells. The presence of these miRNAs was shown by Pandolfi and co-workers by RNA Immunoprecipitation

followed by RT-qPCR (Tay, Kats et al. 2011): miR-17-5p, miR-19a, miR-19b, miR-20a, miR-26a, miR-93, miR-106a and miR-106b.

Gene upregulations were determined by a fold change greater than 1, foldchange(PTEN) > 1, where:

$$foldchange_{reference \to \text{modified}}(gene) = \frac{\sum_{MREs \in reference} occupancy(gene) \times hp}{\sum_{MREs \in \text{modified}} occupancy(gene) \times hp}.$$

Similarly, gene downregulations were defined by a fold change smaller than 1, foldchange(*gene*) < 1.

Validating miRBooking. We validated the occupancies computed by miRBooking by comparing our predicted values with those determined experimentally by van Oudenaarden and co-workers (Mukherji, Ebert et al. 2011). As for their experiment, we built a synthetic mRNA reporter with 0 or 1 binding site for miR-20a-5p. We expressed the reporters at various levels and compared the predicted occupancies of all miRNAs (cooperative effectiveness) on each of the reporters (Fig. S2). We observed a similar behaviour between the comparative cooperative effectiveness and the quantitative fluorescence microscopy data obtained from the experiment (see Fig. 1c in reference (Mukherji, Ebert et al. 2011)).

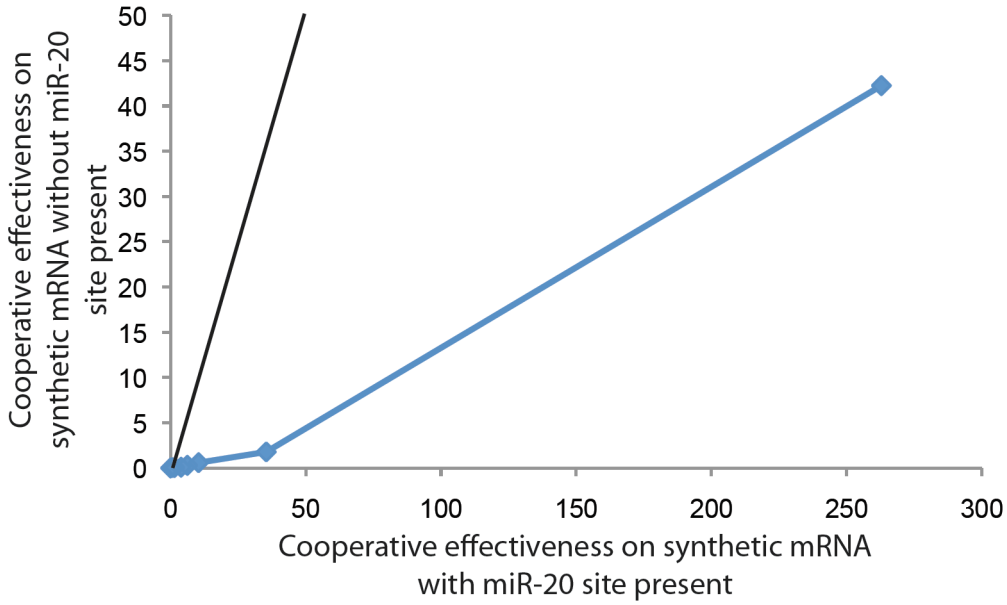


Figure A2-2 Predicted effectiveness of a synthetic reporter with 1 or 0 target site for miR-20a-5p.
Cooperative effectiveness of the miRNAs expressed in DU145 of a reporter containing 0 (Y-axis) or 1 (X-axis) miR-20a-5p binding site. The black line represents the expected occupancies if the two reporters generated the same cooperative effectiveness (diagonal), compared to the blue curve which indicates the effectiveness of the miRNAs on the reporter with 1 site.

Moreover, we compared predicted fold changes with those experimentally obtained for PTEN and VAPA proteins when eight miRNAs targeting both were individually overexpressed (Tay, Kats et al. 2011). The correlation between the predicted fold change and the protein expression measurements was $R^2 = 0.711$ (see main text Fig. 2b).

Selecting ceRNAs (for *PTEN*)

The ceRNAs of a target mRNA (target) are the mRNAs (ceRNAs) in a cell context that share the pool of miRNAs (miRpool). The miRpool are the miRNAs booked on the target at a desired abundance, and can be identified by the microtargetome. We identified the ceRNAs that have the potential to be occupied by

miRpool in endogenous conditions. We used the MREs obtained from hybridisation probabilities as an approximation.

For example, we wanted to increase *PTEN* up to the 15% most expressed mRNAs in DU145 cells. We thus ran miRBooking with *PTEN*'s abundance fixed at the fifteen percentile, and identified the miRNAs having a positive influence on *PTEN*, in this case nine miRNAs: let-7a-5p, let-7c, miR-19b-3p, miR-374a-5p, miR-494, miR-628-3p, miR-668, let-7d-3p, and miR-374-5p. Then, we identified 160 ceRNAs that are occupied by at least the nine miRNAs targeting *PTEN* when expressed in the fifteen percentile.

Identifying miR-132-mediated oncogenes in prostate cancer

MiR-132 levels are low in both the DU145 and PC3 cell lines. Its targets in these cell lines are thus subject to be affected by miR-132 increased levels. We identified fourteen common targets of miR-132 in both cell lines. Five that are not coding for ribosomal proteins: EEF1A1, GAPDH, PGAM1, PPIA, and PKM2; and, ten that do: RPL37A, RPL38, RPLP1, RPL10, RPS17, RPS20, RPS29, RPL28, RPS2, and RPL8. The five that are not coding for ribosomal proteins represent putative oncogenes.

Determining miRBooking target prediction accuracy

Using mirBooking, we simulated the overexpression of miR-1 and miR-133a in DU145 and PC3 cell lines. In each experiment, we compared the microtargetomes before and after overexpression and computed the *repression* of each mRNA's expression by:

$$repression_{\text{reference}\rightarrow\text{modified}}(mRNA) = \sum_{MREs}\left[\#\,MRE_{\text{reference}}(mRNA) - \#\,MRE_{\text{modified}}(mRNA)\right] \times hp_{MRE}$$

We compared our predicted repressions with those measured by microarray (Gene Expression Omnibus GSE26032) (Kojima, Chiyomaru et al. 2012). The platforms used in these experiments and our quantifications differed, and thus we only considered the genes that are mapped in both microarrays. Each sample of the dataset GSE26032

contains the $\log_2$ ratios of the expressed mRNA intensities before and after miRNA overexpressions.

Using microarray experiments to evaluate any miRNA target prediction algorithm provides pros and cons. The main advantage is the large coverage of mRNAs. However, the biological signal is imprecise as the $\log_2$ ratio approaches 0. While considering only the most affected mRNAs (e.g. $\log_2$ ratios < -1) would provide better precision, it might also lead to underestimations of the wide impact of a miRNA overexpression. Our aim here is to evaluate the miRNA target prediction algorithms using the microarray data.

The eight mRNAs found affected by the miRNA overexpressions in the microarrays were predicted by miRBooking. These represent true positives. To evaluate the false positive rate, we computed the AUC. We considered the repression metric to discriminate between affected and non-affected mRNAs. Choosing different $\log_2$ ratio thresholds changes the sets of affected and non-affected mRNAs. For instance, a threshold of -2 defines only five affected mRNAs, whereas a threshold of -0.5 defines 754 affected mRNAs. To decrease the impact of the noise present in the microarrays, we evaluated the AUC at different $\log_2$ ratio thresholds: from 0 to the minimum $\log_2$ ratio in each experiment using incremental steps of 0.01.

Using the same approach, we evaluated and compared the predictive power of TargetScan with and without considering the sequence conservation criterion (Fig. S3). We used TargetScan to identify the mRNAs targeted by miR-1 and miR-133a, indeed independently of the cellular context. We computed the AUC by sorting the mRNAs by the scores produced by TargetScan. For the mRNAs that were not predicted to be targeted, we arbitrarily assigned a score of 10,000.

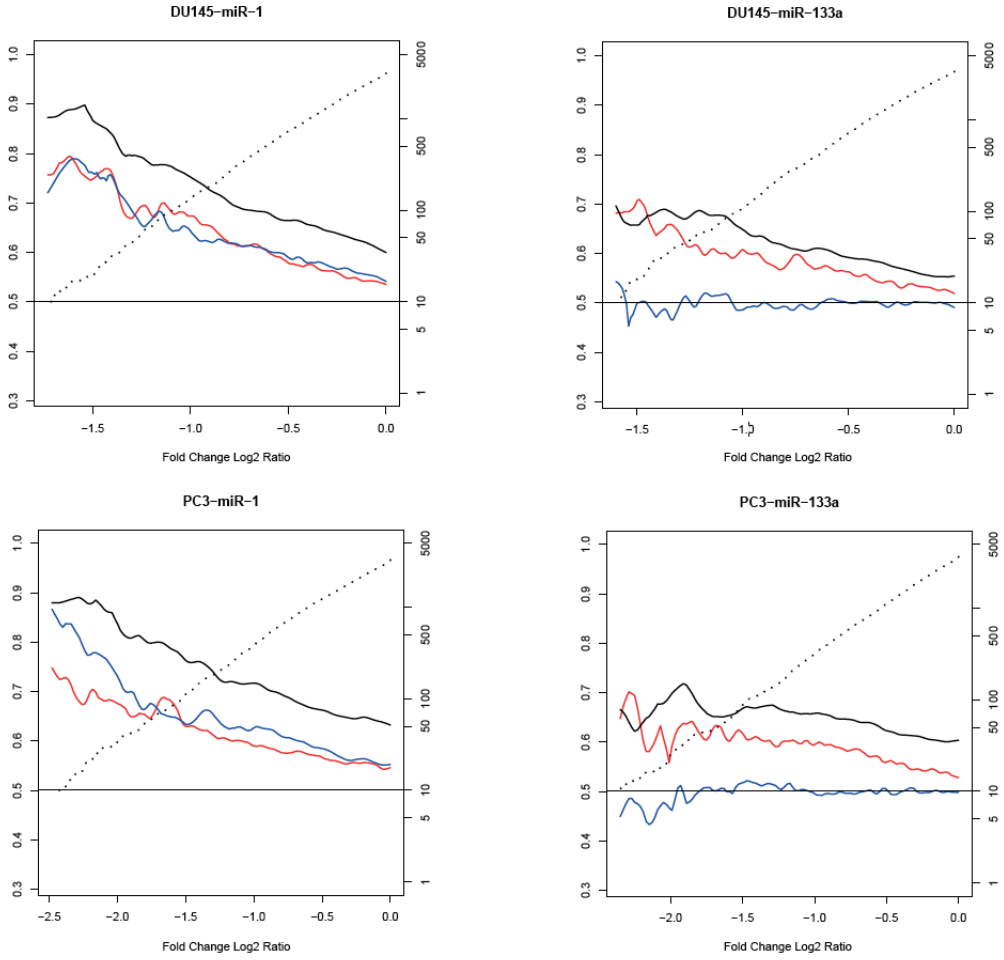


Figure A2-3 Predictive power of miRBooking and TargetScan.
The predictive power of the algorithms was evaluated in four different miRNA overexpression experiments. The AUCs are shown using red (TargetScan), blue (TargetScan without evolutionary conservation), and black (miRBooking) curves. The dashed lines indicate the number of mRNAs experimentally determined as affected at different $\log_2$ ratio thresholds (right y-axis). The horizontal line corresponds to the AUC that would have been obtained by a random predictor (AUC = 0.5).

Linking sets of genes and biological pathways

We used the DAVID bioinformatics platform (Huang da, Sherman et al. 2009) and KEGG (Kanehisa, Goto et al. 2012). Note that miRBooking can only predict the mRNAs for which we know the quantities, which define a specific gene space. Similarly, the gene space of a microarray is defined by the list of genes for which a probe was defined. The DAVID bioinformatics platform defines this gene space as the "background", which can be specified by the user.

The KEGG Pathway Database is a collection of manually drawn pathway maps representing our current knowledge on the molecular interaction and reaction networks for: metabolisms, genetic information processing, environmental information processing, cellular processes, organismal systems, and human diseases. The database is organised as a three level tree, where a leaf corresponds to a single biological pathway defined by a group of genes. For instance, the genes reported to be linked to prostate cancers by any means form the *prostate cancer* leaf (level 3 in the tree) that belongs to the *cancer* category (level 2 in the tree; numbered 6.1 in KEGG), which belongs to *human diseases* (level 1 in the tree; numbered 6 in KEGG).

Knowing the KEGG pathway database and background, DAVID determines *a priori* probabilities of each pathway. For a given group of genes, DAVID computes the associated P-value of each pathway. The link between a miRNA and its biological pathways can be established by the genes it targets and the DAVID P-values associated to the pathways.

We compared the pathways obtained from the genes identified downregulated by the overexpressions of miR-1 and miR-133a in the DU145 and PC3 cell lines, as predicted by miRBooking (Tab. S1), or determined in the microarray data (Kojima, Chiyomaru et al. 2012) (Tab. S2) using their respective backgrounds.

The number of genes directly targeted by miR-1 and miR-133a in the simulations is greater than 1,000 and links to a number of pathways, *N*, of which a fraction belongs to the *cancer* category (labeled 6.1 in the KEGG pathway database). To evaluate if these pathways are enriched in cancer, we first computed the Z-scores of all KEGG pathway categories (level 2 in the KEGG tree) for each simulation:

$Z_{category} = \frac{L - \mu}{\sigma}$, where L is the number of pathways in this category (level 3 in the KEGG tree), μ and σ are the mean and standard deviation of the numbers of pathways belonging to this category as determined by 1000 random selection of *N* pathways. Z-scores > 3 are considered significant.

Negative controls. We simulated the overexpression of three randomly picked miRNAs: miR-27a-3p, miR-30b-5p, and let-7e-3p. None of them gave a significant enrichment in the *cancer* category with Z-scores between -0.69 and 1.5, but for let-7e-3p in the PC3 cell line with a z-score of 3.8.

Table A2-I List of KEGG pathways from predicted genes.
List generated using David from the analysing of the genes that have been predicted to be targeted by miR-1 and miR-133a after their overexpression in the DU145 and PC3 cell lines.

DU145				PC3			
miR-1		miR-133a		miR-1		miR-133a	
Term	P-Value	Term	P-Value	Term	P-Value	Term	P-Value
Focal adhesion	2.80E-04	Cell cycle	5.60E-09	*Prostate cancer*	2.20E-04	Spliceosome	3.10E-07
Phosphatidylinositol signaling system	2.90E-04	Pancreatic cancer	9.50E-08	Focal adhesion	4.30E-04	Endocytosis	4.00E-07
Pathways in cancer	4.00E-04	Chronic myeloid leukemia	1.60E-07	Glioma	4.30E-04	Cell cycle	1.30E-06
Glioma	1.10E-03	Spliceosome	4.70E-07	Pathways in cancer	7.10E-04	Chronic myeloid leukemia	1.60E-06
Insulin signaling pathway	4.00E-03	Aminoacyl-tRNA biosynthesis	1.00E-06	Insulin signaling pathway	1.70E-03	Aminoacyl-tRNA biosynthesis	2.20E-06
Colorectal cancer	4.10E-03	Pyrimidine metabolism	1.30E-05	Pancreatic cancer	2.70E-03	Pancreatic cancer	9.30E-06
Aldosterone-regulated sodium reabsorption	4.20E-03	Endocytosis	1.30E-05	Non-small cell lung cancer	3.50E-03	Pyrimidine metabolism	1.50E-05
Prostate cancer	6.40E-03	Ubiquitin mediated proteolysis	2.30E-05	Aldosterone-regulated sodium reabsorption	4.30E-03	Ubiquitin mediated proteolysis	3.80E-05
Melanoma	7.20E-03	DNA replication	4.30E-05	Renal cell carcinoma	4.50E-03	mTOR signaling pathway	5.70E-05
Dorso-ventral axis formation	1.00E-02	Neurotrophin signaling pathway	7.90E-05	Phosphatidylinositol signaling system	5.20E-03	DNA replication	7.90E-05
ECM-receptor interaction	1.00E-02	mTOR signaling pathway	8.60E-05	Melanoma	5.70E-03	Axon guidance	1.20E-04
Non-small cell lung cancer	1.10E-02	Renal cell carcinoma	9.10E-05	Colorectal cancer	6.30E-03	Renal cell carcinoma	2.00E-04
Renal cell carcinoma	1.40E-02	*Prostate cancer*	1.10E-04	Chronic myeloid leukemia	7.20E-03	Small cell lung cancer	5.20E-04
Neurotrophin signaling pathway	1.40E-02	Lysine degradation	1.20E-04	Dorso-ventral axis formation	1.30E-02	Citrate cycle (TCA cycle)	6.00E-04
Pancreatic cancer	1.60E-02	Lysosome	1.80E-04	Neurotrophin signaling pathway	1.40E-02	N-Glycan biosynthesis	6.40E-04
Acute myeloid leukemia	1.80E-02	Small cell lung cancer	2.30E-04	Axon guidance	1.40E-02	Lysine degradation	7.00E-04
Wnt signaling pathway	2.90E-02	Pathways in cancer	2.70E-04	Acute myeloid leukemia	1.50E-02	Valine, leucine and isoleucine degradation	7.00E-04
Small cell lung cancer	2.90E-02	ErbB signaling pathway	2.80E-04	Small cell lung cancer	1.80E-02	Lysosome	9.30E-04
Chronic myeloid leukemia	3.10E-02	Purine metabolism	4.60E-04	Progesterone-mediated oocyte maturation	3.50E-02	ErbB signaling pathway	1.30E-03
Inositol phosphate metabolism	4.20E-02	Acute myeloid leukemia	4.80E-04	Endometrial cancer	3.60E-02	Neurotrophin signaling pathway	1.60E-03
mTOR signaling pathway	4.70E-02	Non-small cell lung cancer	6.20E-04	ECM-receptor interaction	3.60E-02	Base excision repair	1.70E-03
GnRH signaling pathway	4.90E-02	Endometrial cancer	7.00E-04	Type II diabetes mellitus	4.50E-02	Nucleotide excision repair	1.90E-03
Long-term potentiation	5.10E-02	Progesterone-mediated oocyte maturation	9.60E-04	mTOR signaling pathway	4.50E-02	Insulin signaling pathway	2.00E-03
Jak-STAT signaling pathway	5.60E-02	Colorectal cancer	1.10E-03	Fc gamma R-mediated phagocytosis	4.50E-02	Acute myeloid leukemia	2.20E-03
Progesterone-mediated oocyte maturation	5.60E-02	Nucleotide excision repair	1.10E-03	TGF-beta signaling pathway	5.70E-02	Purine metabolism	2.30E-03
Fc gamma R-mediated phagocytosis	6.10E-02	Valine, leucine and isoleucine degradation	1.10E-03	Wnt signaling pathway	6.00E-02	Pathways in cancer	2.80E-03
Type II diabetes mellitus	6.70E-02	Citrate cycle (TCA cycle)	1.30E-03	Bladder cancer	6.20E-02	Non-small cell lung cancer	2.80E-03
Melanogenesis	6.80E-02	Apoptosis	1.30E-03	Long-term potentiation	6.40E-02	RNA polymerase	4.00E-03
		Tight junction	2.00E-03	Jak-STAT signaling pathway	6.70E-02	Oocyte meiosis	6.10E-03
		Adherens junction	2.50E-03	Riboflavin metabolism	7.00E-02	SNARE interactions in vesicular transport	6.70E-03

	N-Glycan biosynthesis	2.60E-03	GnRH signaling pathway	7.40E-02	Progesterone-mediated oocyte maturation	7.40E-03
	B cell receptor signaling pathway	2.90E-03	Glycerophospholipid metabolism	7.60E-02	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	7.90E-03
	Base excision repair	3.20E-03	Adherens junction	7.90E-02	*Prostate cancer*	8.50E-03
	Inositol phosphate metabolism	3.70E-03			TGF-beta signaling pathway	9.50E-03
	Insulin signaling pathway	4.30E-03			Apoptosis	9.50E-03
	Oocyte meiosis	5.20E-03			Valine, leucine and isoleucine biosynthesis	1.10E-02
	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	5.60E-03			p53 signaling pathway	1.10E-02
	T cell receptor signaling pathway	5.80E-03			Inositol phosphate metabolism	1.30E-02
	Valine, leucine and isoleucine biosynthesis	8.80E-03			Fatty acid metabolism	1.40E-02
	Fatty acid metabolism	9.40E-03			Colorectal cancer	1.50E-02
	Focal adhesion	1.00E-02			Focal adhesion	1.70E-02
	SNARE interactions in vesicular transport	1.10E-02			Glioma	2.10E-02
	p53 signaling pathway	1.30E-02			Phosphatidylinositol signaling system	2.60E-02
	Glioma	1.30E-02			Basal transcription factors	2.90E-02
	Wnt signaling pathway	1.80E-02			Homologous recombination	2.90E-02
	Bladder cancer	1.90E-02			Adherens junction	2.90E-02
	RNA degradation	1.90E-02			RNA degradation	3.00E-02
	Homologous recombination	2.20E-02			Adipocytokine signaling pathway	3.00E-02
	Alanine, aspartate and glutamate metabolism	2.70E-02			Gap junction	4.10E-02
	Steroid biosynthesis	2.70E-02			T cell receptor signaling pathway	4.70E-02
	Phosphatidylinositol signaling system	2.80E-02			Endometrial cancer	5.30E-02
	Thyroid cancer	3.10E-02			Other glycan degradation	6.10E-02
	Butanoate metabolism	3.30E-02			Tight junction	6.50E-02
	Sphingolipid metabolism	3.30E-02			Huntington's disease	6.80E-02
	Huntington's disease	3.40E-02			Wnt signaling pathway	7.30E-02
	Gap junction	4.00E-02			Selenoamino acid metabolism	7.40E-02
	Long-term potentiation	4.10E-02			B cell receptor signaling pathway	8.60E-02
	Synthesis and degradation of ketone bodies	4.30E-02			Thyroid cancer	8.70E-02
	Fc gamma R-mediated phagocytosis	4.50E-02			Steroid biosynthesis	9.10E-02
	Other glycan degradation	5.00E-02			Bladder cancer	9.70E-02
	MAPK signaling pathway	5.00E-02			Propanoate metabolism	1.00E-01
	RNA polymerase	5.00E-02				
	Amino sugar and nucleotide sugar metabolism	6.30E-02				
	Notch signaling pathway	6.90E-02				
	TGF-beta signaling pathway	7.10E-02				
	Regulation of actin cytoskeleton	7.60E-02				
	Propanoate metabolism	7.80E-02				
	Biosynthesis of unsaturated fatty acids	7.90E-02				
	Basal transcription factors	8.70E-02				
	Adipocytokine signaling pathway	9.20E-02				

		Dorso-ventral axis formation	9.20E-02				
		NOD-like receptor signaling pathway	9.80E-02				

Table A2-II List of KEGG pathways from microarray genes.

List generated using David from the analysing of the genes that have been identified by microarray analysis to be targeted by miR-1 and miR-133a after their overexpression in the DU145 and PC3 cell lines.

DU145				PC3			
miR-1		miR-133a		miR-1		miR-133a	
Term	P-Value	Term	P-Value	Term	P-Value	Term	P-Value
Ribosome	3.70E-09	Spliceosome	1.80E-10	Focal adhesion	6.40E-09	Adherens junction	2.30E-06
Spliceosome	2.30E-06	DNA replication	5.50E-08	Pathways in cancer	5.50E-08	Pathways in cancer	2.40E-05
DNA replication	9.40E-06	Cell cycle	2.00E-06	Lysosome	1.30E-06	Fc gamma R-mediated phagocytosis	2.70E-04
Homologous recombination	6.80E-05	Homologous recombination	7.00E-05	Small cell lung cancer	4.00E-05	Pathogenic Escherichia coli infection	3.00E-04
Cell cycle	1.20E-04	N-Glycan biosynthesis	1.30E-04	Pancreatic cancer	5.40E-05	Insulin signaling pathway	5.70E-04
N-Glycan biosynthesis	1.70E-04	Lysosome	1.90E-04	Regulation of actin cytoskeleton	7.80E-05	Neurotrophin signaling pathway	6.80E-04
Mismatch repair	3.60E-04	Mismatch repair	1.30E-03	Fc gamma R-mediated phagocytosis	9.90E-05	Focal adhesion	8.20E-04
Pyrimidine metabolism	3.90E-04	Progesterone-mediated oocyte maturation	1.60E-03	Leukocyte transendothelial migration	1.30E-04	Ribosome	9.90E-04
Graft-versus-host disease	6.90E-04	Pancreatic cancer	2.40E-03	Neurotrophin signaling pathway	1.40E-04	Chronic myeloid leukemia	1.60E-03
RNA degradation	7.30E-04	Oocyte meiosis	3.20E-03	Colorectal cancer	2.40E-04	N-Glycan biosynthesis	1.80E-03
Type I diabetes mellitus	8.20E-04	Nucleotide excision repair	4.70E-03	Adherens junction	2.80E-04	Steroid biosynthesis	1.90E-03
Purine metabolism	3.90E-03	Pyrimidine metabolism	5.00E-03	ECM-receptor interaction	5.30E-04	ErbB signaling pathway	2.00E-03
Lysine degradation	5.90E-03	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	6.40E-03	Axon guidance	8.10E-04	Gap junction	2.10E-03
Aminoacyl-tRNA biosynthesis	8.90E-03	Base excision repair	1.10E-02	*Prostate cancer*	8.80E-04	Colorectal cancer	2.70E-03
Systemic lupus erythematosus	1.00E-02	Oxidative phosphorylation	1.30E-02	Glioma	9.10E-04	Lysosome	2.90E-03
One carbon pool by folate	2.50E-02	Systemic lupus erythematosus	1.50E-02	Cell cycle	1.40E-03	B cell receptor signaling pathway	3.20E-03
Small cell lung cancer	2.70E-02	Other glycan degradation	1.60E-02	DNA replication	2.40E-03	T cell receptor signaling pathway	3.60E-03
Nucleotide excision repair	2.80E-02	p53 signaling pathway	1.70E-02	Chronic myeloid leukemia	5.70E-03	Endocytosis	4.20E-03
Pathways in cancer	4.20E-02	Proteasome	1.70E-02	p53 signaling pathway	7.20E-03	Regulation of actin cytoskeleton	4.90E-03
Pathogenic Escherichia coli infection	4.40E-02	Endocytosis	1.70E-02	Endocytosis	8.10E-03	mTOR signaling pathway	5.70E-03
Base excision repair	4.50E-02	Vibrio cholerae infection	2.00E-02	Bladder cancer	8.60E-03	Dorso-ventral axis formation	5.90E-03
Dorso-ventral axis formation	4.70E-02	Sphingolipid metabolism	3.10E-02	Chemokine signaling pathway	9.80E-03	Tight junction	6.80E-03
ECM-receptor interaction	5.60E-02	Purine metabolism	3.40E-02	Apoptosis	1.20E-02	Spliceosome	1.10E-02
Fc gamma R-mediated phagocytosis	5.70E-02	Alzheimer's disease	4.00E-02	Natural killer cell mediated cytotoxicity	1.20E-02	MAPK signaling pathway	1.10E-02
Allograft rejection	6.10E-02	Adherens junction	4.10E-02	Graft-versus-host disease	1.20E-02	Homologous recombination	1.30E-02
Cysteine and methionine metabolism	6.30E-02	Pathogenic Escherichia coli infection	5.00E-02	Renal cell carcinoma	1.40E-02	*Prostate cancer*	1.30E-02
Nicotinate and nicotinamide metabolism	7.10E-02	Huntington's disease	5.10E-02	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	1.50E-02	Axon guidance	1.40E-02
Limonene and pinene degradation	7.30E-02	beta-Alanine metabolism	6.60E-02	Progesterone-mediated oocyte maturation	1.50E-02	Renal cell carcinoma	1.60E-02
p53 signaling pathway	9.50E-02	Glioma	6.60E-02	Cytokine-cytokine receptor interaction	1.50E-02	Pancreatic cancer	1.60E-02
Thyroid cancer	9.50E-02	Non-homologous end-joining	7.30E-02	Type I diabetes mellitus	1.90E-02	ECM-receptor interaction	1.70E-02
		Small cell lung cancer	7.90E-02	Endometrial cancer	2.20E-02	Small cell lung cancer	1.70E-02
		Fructose and mannose metabolism	8.00E	NOD-like receptor	2.40E	Leukocyte	1.80E-

	-02	signaling pathway	-02	transendothelial migration	02
Regulation of actin cytoskeleton	8.20E-02	Aldosterone-regulated sodium reabsorption	2.70E-02	Vibrio cholerae infection	2.60E-02
mTOR signaling pathway	8.50E-02	Steroid biosynthesis	2.90E-02	Oocyte meiosis	2.80E-02
Steroid biosynthesis	8.70E-02	Homologous recombination	3.00E-02	Sphingolipid metabolism	2.80E-02
Ribosome	9.60E-02	Pathogenic Escherichia coli infection	3.20E-02	Notch signaling pathway	3.20E-02
Parkinson's disease	9.80E-02	RIG-I-like receptor signaling pathway	3.30E-02	Cell cycle	3.30E-02
		T cell receptor signaling pathway	3.90E-02	Glioma	3.60E-02
		Mismatch repair	4.00E-02	Apoptosis	3.70E-02
		ABC transporters	4.00E-02	Non-small cell lung cancer	4.60E-02
		N-Glycan biosynthesis	4.10E-02	Ubiquitin mediated proteolysis	4.60E-02
		Non-small cell lung cancer	4.20E-02	Progesterone-mediated oocyte maturation	4.60E-02
		Phosphatidylinositol signaling system	4.30E-02	Thyroid cancer	4.90E-02
		Jak-STAT signaling pathway	5.10E-02	Melanogenesis	5.50E-02
		TGF-beta signaling pathway	5.20E-02	PPAR signaling pathway	6.10E-02
		MAPK signaling pathway	5.40E-02	Adipocytokine signaling pathway	6.10E-02
		B cell receptor signaling pathway	5.50E-02	Bladder cancer	7.80E-02
		Viral myocarditis	5.50E-02	Acute myeloid leukemia	7.90E-02
		Adipocytokine signaling pathway	5.60E-02	Endometrial cancer	7.90E-02
		Basal cell carcinoma	5.60E-02	Nitrogen metabolism	9.50E-02
		Cytosolic DNA-sensing pathway	5.60E-02	GnRH signaling pathway	9.90E-02
		Folate biosynthesis	6.00E-02		
		Dilated cardiomyopathy	6.30E-02		
		O-Glycan biosynthesis	6.90E-02		
		Tight junction	7.00E-02		
		Calcium signaling pathway	7.10E-02		
		Allograft rejection	7.20E-02		
		mTOR signaling pathway	7.40E-02		
		Gap junction	7.90E-02		
		Oocyte meiosis	8.30E-02		
		Melanoma	8.80E-02		
		Type II diabetes mellitus	9.70E-02		

Annexe 3. A comparative analysis of the triloops in all high-resolution RNA structures reveals sequence-structure relationships (Supplementary Information)

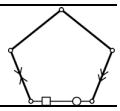
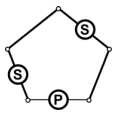
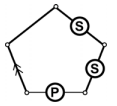
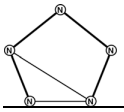
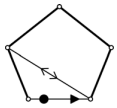
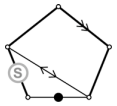
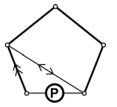
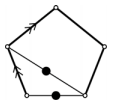
Table A3-I. PDB files.
All PDB files where triloops were found

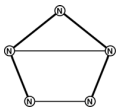
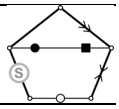
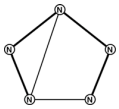
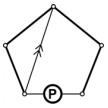
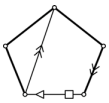
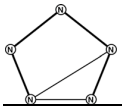
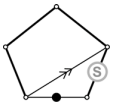
1EFW	1GTR	1K9M	1N8R	1QRT	1U8D	1VQK	1Y39	1ZSE	2CSX	2HOL
1EUY	1GTS	1KD1	1NJI	1QRU	1U9S	1VQL	1YHQ	2A43	2CT8	2HOO
1EXD	1H3E	1L3D	1NTA	1QTQ	1VC6	1VQM	1YI2	2AZX	2CV0	2J00
1FFK	1HC8	1M90	1O0C	1QU2	1VQ4	1VQN	1YIJ	2B2D	2CV1	2J01
1FG0	1HR2	1MMS	1Q81	1QU3	1VQ5	1VQO	1YIT	2B57	2CV2	2J02
1FJG	1I9V	1MZP	1Q82	1QVG	1VQ6	1VQP	1YJ9	2BS0	2DR2	2J03
1G59	1J1U	1N32	1Q86	1S72	1VQ7	1XMQ	1YJN	2BS1	2FK6	437D
1GAX	1JJ2	1N77	1QA6	1SER	1VQ8	1Y26	1YJW	2BTE	2GDI	6MSF
1GID	1K8A	1N78	1QRS	1U0B	1VQ9	1Y27	1ZJW	2CKY	2GDI	

Table A3-II. Triloop instances

The 104 triloop specimens found in RNA-3A. Classes are named using Roman numbers (I to VIII) and shown on the left using the common topology of the structural graphs of its specimens. For each specimen, the structural subclass in indicated by a name that corresponds to its nucleotide interactions: L: phosphodiester Linkage with no combined stack or pair interactions; S: Strack; P: Pair. The four interactions along the backbone path constitute the first part of the name (i.e. I.LLLL means a subclass of class I defined by four phosphodiester Linkages with no other combined interactions). The dashes separate the backbone from the intra-loop interactions. The intra-loop interactions are listed after the first dash and are sorted by the first nucleotide involved in the interacton. The nucleotides are specified using the triloop numbering defined in (Firgue 1a). For example, subclass II.LLSL-S2-5 means a subclass of class II where nucleotides 1 and 2, 2 and 3, and 4 and 5 are simply Linked without any other interactions, nucleotides 3 and 4 are Stacked (and linked), and nucleotides 2 and 5 are involved in intra-loop Stacking. The subclasses with an asterisk after their names were previously reported in the literature. The numbers, underneath the subclass name and separated by a slash, are, respectively, the number of specimens and the number of triloop sites found in RNA-3A in this subclass. Each specimen is defined by a row indicating the number of sites (under Subclass), the structure number (Str) within the subclass, the RNA type (RNA), the PDB file identifier (PDB code), the PDB residue number of nucleotide 1 (A1 location), the sequence, the specific base pairing and stacking interaction types (Interactions), and the base pairing type of the flanking base pair with cis/trans (Flanking) and parallel/antiparallel notations (Para/anti). The letter 'n' in each row indicates that the speciment was previously reported (New). The rows that do not have an RNA, PDB code and A1 location represent extrapolated sequences (see Materials & Methods) of thecurrent structure. Each subclass is also represented by a structural graph (Diagram), where the specific or generalized interaction types are shown. The presence of specific base peiring or stacking types is indicated by corresponding symbols from the nomenclature (see Materials & Methods). However, when more than one base pairing or base stacking types are found, they are indicated, respectively, by black circled P or S. Grey circled S indicate stacking interactions at the limit of the automated annotation. These cases were examined visually and were determined as possibly stacking given small local motion.

Class	Subclass	Str	RNA	PDB code	A1 location	Sequence	Interactions			Flanking	Para/ anti	New	Diagram
I	LLLL*												
	8/54												
	2	1	23S	2J01	A2749	AAGCA				H/W t	a		
	8	2	tRNA	1G59	B515	GCGGU				S/H c	p		
	5	3	5S	1Q86	B3022	GUUGC				H/W c	p		
	2	4	16S	1FJG	A1053	GCAUG				W/H c	a	n	
	1		RNase P	1U9S	A188	GCCCA				W/H t	a	n	
	32		23S	1FFK	0'335	UGACA				W/H t	a		
	2	5	16S	2J00	A201	CUUUG				W/W c	a		
	2		Intron I	1HR2	B235	AUCUU				W/W c	a		
	LLLS*												
	4/63												
	7	1	23S	1YIT	0'218	CGCGA	up			S/H t	a		
	21	2	5S	1YIT	9'22	GUUGC	up			S/W t	p	n	
	4	3	riboswitch	2B57	A47	UUUCU	up			W/H t	a		
	31		23S	1YJN	0'2482	GAUAA	up			W/H t	p	n	
	LLSL*												
	7/21												
	5	1	16S	1FJG	A64	GUGCG	up			W/H c	p	n	
	2		16S	1N32	A1053	GCAUG	up			W/H c	p	n	
	2		tRNA	2DR2	B54	UUCGA	up			W/H t	a	n	
	1		23S	1N8R	A2597	UUAAA	up			W/H t	a	n	
	1	2	16S	1FJG	A352	CAGCA	up			W/S c	p		
	2	3	23S	2J01	A1925	CUAAG	up			W/W c	p		
	8	4	23S	1VQ7	0'1186	CUAAG	in			W/W t	p	n	
	LLSS*												
	6/38												
	1	1	16S	1XMQ	A352	CAGCA	up	up		W/H c	p		
	2		16S	2J00	A618	CUCAA	up	up		W/Bh t	a		
	1	2	16S	1N32	A352	CAGCA	up	up		W/H c	p		
	3		16S	1FJG	A618	CUCAA	up	up		W/H t	a		
	1	3	23S	1VQN	0'1186	CUAAG	in	down		W/W t	p	n	
	30	4	23S	1JJ2	0'1813	UGACU	up	up		W/W c	a		
	LSLS												
	6/39												
	1	1	tRNA	1SER	T9	GCCCG	up	up		H/W t	a		
	1		tRNA	1B23	R9	AACAA	up	up		H/W t	a		
	14	2	23S	1N8R	A2033	GUCCC	up	in		W/Bs c	a		
	2		23S	2J01	A1992	GUCUC	up	in		W/Bs c	a		
	3	3	Virus	1L3D	A8	CACCG	up	up		W/H t	a		
	18	4	23S	1FFK	0'2033	GUCCC	up	in		W/W c	a		
	LSSS*												
	5/8												
	1	1	tRNA	1H3E	B9	UCCCG	up	up	up	H/W t	a		
	1		tRNA	1U0B	A9	AACAA	up	up	up	H/W t	a		
	1	2	Virus	2A43	A7	CACCG	up	up	up	W/H t	a		
	3	3	16S	1FJG	A934	CACAA	up	up	up	W/W t	p	n	
	2		Riboswitch	2GDI	X28	UGCGU	up	up	up	W/W t	p		
	PLLL												
	1/2												
	2	1	Intron I	1GID	A235	AUCUU	S/H t			W/W t	a		
	SLLL*												
	6/50												
	1	1	23S	1FFK	0'326	GAUAC	up			W/H c	p	n	
	2		16S	2J00	A1315	UGCAA	up			W/H t	a		
	1		tRNA	2CT8	C54	UUCGA	up			W/H t	a	n	
	1		tRNA	2J02	W54	UUCGA	in			W/H t	a		
	31	2	23S	1M90	A326	GAUAC	up			W/W c	a	n	
	14		23S	1N8R	A118	GAAUC	up			W/W c	a		

	SLLS												
	1/32												
	32	1	23S	1FFK	0'2784	ACGCA	in	up		H/W t	a		
	SLSL*												
	24/272												
	21	1	tRNA	1EUY	B954	UUCGA	up	up		W/H t	a		
	34		23S	1FFK	0'481	UGCAA	up	up		W/H t	a	n	
	32		23S	1FFK	0'505	CGAAA	up	up		W/H t	a	n	
	12		23S	1JJ2	0'624	UUUGA	up	up		W/H t	a	n	
	36		23S	1FFK	0'1388	UGAGA	up	up		W/H t	a	n	
	32		23S	1FFK	0'2597	UUAAA	up	up		W/H t	a	n	
	32		23S	1FFK	0'313	UGGAA	up	up		W/H t	a	n	
	9		tRNA	1FFY	T54	UUCAA	up	up		W/H t	a		
	2		tRNA	1GAX	C953	UUCAA	up	up		W/H t	a		
	3		16S	1FJG	A1177	GGAAG	up	up		W/H t	a	n	
	1		16S	1FJG	A1315	UGCAA	up	up		W/H t	a	n	
	5		16S	1FJG	A323	UGAGA	up	up		W/H t	a	n	
	5		16S	1FJG	A956	UUAAU	up	up		W/W t	p	n	
	1		RNase P	1U9S	A122	AGAGA	up	up		W/H t	a		
	2		RNase P	1U9S	A170	UGAAA	up	up		W/H t	a		
	8		tRNA	2AZX	C554	UUCGA	up	up		W/H t	a		
	4		riboswitch	2H0J	A39	UGAGA	up	up		W/H t	a		
	2		23S	2J01	A499	UGAAA	up	up		W/H t	a		
	2		23S	2J01	A2562	UUAAA	up	up		W/H t	a		
	2		23S	2J01	A306	UGGGA	up	up		W/H t	a		
	2		23S	2J01	A475	UGAAA	up	up		W/H t	a		
	1	2	tRNA	2J00	W54	UUCGA	in	up		W/H t	a		
	23	3	23S	1YJW	A1186	CUAAG	up	in		W/W t	p	n	
	1	4	23S	1HC8	D332	UUAAA	up	up		W/W c	p		
	SLSS*												
	4/8												
	1	1	23S	1Q81	A2597	UUAAA	up	up	up	W/H c	a	n	
	3	2	23S	1HC8	C132	UUAAA	up	up	in	W/W t	p		
	3		23S	1MMS	C1082	UUAAA	up	up	in	W/W t	p		
	1	3	23S	1VQ8	0'1186	CUAAG	up	in	down	W/W t	p	n	
II	LLLL-S2-5												
	1/11												
	11	1	23S	1FFK	0'842	CAAUA	out			W/S c	a		
						UAAUA	out			W/S c	a		
	LLSL-S2-5*												
	2/34												
	2	1	23S	2J01	A1752	CGCAG	up	out		W/W c	a		
	32		23S	1FFK	0'1808	CGCAG	up	out		W/W c	a	n	
						UGCAG	up	out		W/W c	a		
	SLLL-S2-5												
	3/41												
	21	1	23S	1VQ8	0'842	CAAUA	up	out		W/S c	a		
	2	2	23S	2J01	A319	CAGAG	up	out		W/W c	a		
	18		23S	1VQ8	0'118	GAAUC	up	out		W/W c	a		
						UAAUA	up	out		W/W c	a		
						AGCUG	up	out		W/W c	a		
						CGCUG	up	out		W/W c	a		
						UGCUG	up	out		W/W c	a		
	SSLL-P2-5												
	1/1												
	1	1	16S	1XMQ	A1028	CCCGG	up	up	W/W c	W/W c	a		
						ACCGG	up	up	W/W c	W/W c	a		
						GCCGG	up	up	W/W c	W/W c	a		

III	LLLS-S2-4												
	4/17												
	3	1	16S	1FJG	A459	GACGA	up	up		S/H	c	a	
	2		16S	2J02	A460	GACGA	up	up		S/H	t	a	
	2		23S	2J01	A956	GAUAA	up	up		S/H	t	a	
	10	2	23S	1FFK	0'218	CGCGA	up	out		S/H	t	a	
						GAAGA	up	out		S/H	t	a	
						GAUGA	up	out		S/H	t	a	
	LLSS-P2-4*												
	1/32												
	32	1	23S	1FFK	0'1966	UUAAG	W/H	c up	in	W/W	t	p	n
IV	LLLL-S1-3												
	5/30												
	17	1	23S	1FFK	0'671	AGUAU	up			Bs/H	t	a	
	6	2	Aptamer	2BS0	S9	CUAAG	up			W/W	c	a	
	1		Aptamer	6MSF	R6	GUCAC	up			W/W	c	a	
	3		16S	1FJG	A201	CUUUG	up			W/W	c	a	
	3		16S	1FJG	A838	GUCUC	up			W/W	c	a	
						CAAAG	up			W/W	c	a	
						CGAAG	up			W/W	c	a	
						CAAUG	up			W/W	c	a	
						CGAUG	up			W/W	c	a	
						CUAUG	up			W/W	c	a	
						GAAAC	up			W/W	c	a	
						GCAAC	up			W/W	c	a	
						CUAAC	up			W/W	c	a	
						GCAGC	up			W/W	c	a	
						GUAGC	up			W/W	c	a	
						GAAUC	up			W/W	c	a	
						GCAUC	up			W/W	c	a	
						GUAUC	up			W/W	c	a	
						GACAC	up			W/W	c	a	
						GGCAC	up			W/W	c	a	
						GACUC	up			W/W	c	a	
						GGCUC	up			W/W	c	a	
						GAGAC	up			W/W	c	a	
						GCGAC	up			W/W	c	a	
						GGGAC	up			W/W	c	a	
						GUGAC	up			W/W	c	a	
	LLLS-S1-3												
	1/2												
	2	1	23S	2J01	A247	GGCGA	up	up		S/H	t	a	
						CACGA	up	up		S/H	t	a	
						CGCGA	up	up		S/H	t	a	
						CUCGA	up	up		S/H	t	a	
						GACCA	up	up		S/H	t	a	
						GACGA	up	up		S/H	t	a	
						GAGAA	up	up		S/H	t	a	
						GCGAA	up	up		S/H	t	a	
						GGCCA	up	up		S/H	t	a	
						GGGAA	up	up		S/H	t	a	
						GUCCA	up	up		S/H	t	a	
						GUCGA	up	up		S/H	t	a	
						GUGAA	up	up		S/H	t	a	
V	LLLL-S1-4*												
	1/32												
	32	1	23S	1FFK	0'1651	CCAUG	up			W/W	c	a	n
						ACGGA	up			W/W	c	a	

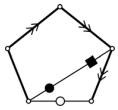
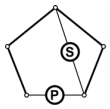
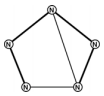
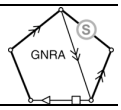
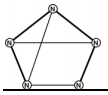
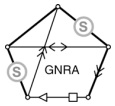
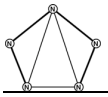
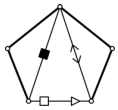
	LSSS-P1-4												
	1/2												
	2	1	16S	2J00	A934	CACAA	up	up	up	W/H c W/W t	p		
						CAACA	up	up	up	W/H c W/W t	p		
						GCCGA	up	up	up	W/H c W/W t	p		
VI	LLLL-S3-5												
	3/14												
	6	1	23S	1VQ8	0'138	UCGCG	out			H/S t	p		
						UCGGG	out			H/S t	p		
						UCGUG	out			H/S t	p		
	2	2	23S	2J01	A1798	UGCGA	out			S/Bs c	a		
						CCAAA	out			S/Bs c	a		
						CCAGA	out			S/Bs c	a		
						CCAUA	out			S/Bs c	a		
						CCGGA	out			S/Bs c	a		
						CCGUA	out			S/Bs c	a		
						UGCAA	out			S/Bs c	a		
	6	3	tRNA	1GTR	B33	UCUGA	up			S/H t	a		
						UCUAA	up			S/H t	a		
	LSLL-S3-5												
	5/84												
	1	1	Virus	1XJR	A22	GAGUA	up	up		S/H t	a		
	32		23S	1FFK	0'494	CAGAA	up	up		S/H t	a		
	30		23S	1KD1	A1707	GCGAA	up	up		S/H c	a		
	19		23S	1FFK	A1276	UCAUA	up	up		S/H c	a		
						GAGCA	up	up		S/H c	a		
						CAGUA	up	up		S/H c	a		
						GAAAA	up	up		S/H c	a		
						GAACA	up	up		S/H c	a		
						GAAGA	up	up		S/H c	a		
						GAAUA	up	up		S/H c	a		
						GAGAA	up	up		S/H c	a		
						GAGCA	up	up		S/H c	a		
						GCAAA	up	up		S/H c	a		
						GCACA	up	up		S/H c	a		
						GCAGA	up	up		S/H c	a		
						GCAUA	up	up		S/H c	a		
						GCGCA	up	up		S/H c	a		
						GCGUA	up	up		S/H c	a		
						UCAAA	up	up		S/H c	a		
						UCACA	up	up		S/H c	a		
						UCAGA	up	up		S/H c	a		
	2	2	23S	2J01	A642	GAACA	up	in		S/W c	p		
						GAAUA	up	in		S/W c	p		
						GCACA	up	in		S/W c	p		
						GCAUA	up	in		S/W c	p		
						GGACA	up	in		S/W c	p		
						GGAUA	up	in		S/W c	p		
	LSLS-S3-5*												
	2/33												
	1	1	23S	2J01	A476	GAAAA	up	down	up	S/H t	a		
	32		23S	1FFK	0'482	GCAAA	up	down	up	S/H t	a	n	
VII	LLLS-S1-3S2-4												
	1/4												
	4	1	23S	1VQ4	0'218	CGCGA	up	up	out	S/H t	a		
VIII	LLLL-P1-3S3-5												
	1/2												
	2	1	23S	1VQM	0'138	UCGCG	H/H c	out		H/S t	p		

Table A3-III. Triloop modeling table.

The 16 possible flanking base partner pairs are represented in a 4X4 matrix (first partner on the left and last partner at the top). For each partner pair, the possible flanking base apiring types are listed together with the observed and extrapolated (underlined) triplets pointing to specific triloop subclasses (class name followed by the subclass name). Three base partner pairs were neither observed nor extrapolated: CC, CU and UC, represented by a X in their corresponding matrix entries.

	A	C	G	U
A	H/W ACA I.LSLS H/W AGC I.LLLL H/W CGC I.SLLS W/H GAG I.SLSL W/W CGG V.LLLL-S1-4		W/W CCG II.SSLL-P2-5 W/W GCU II.SLLL-S2-5	Bs/H GUA IV.LLLL-S1-3 W/W UCU I.LLLL W/W UCU I.PLLL
C	S/Bs CAA CAG CAU CGG VI.LLLL-S3-5 CGU S/H AGU VI.LSLL-S3-5 S/H ACG GCG UCG IV.LLLS-S1-3 S/H AGA VI.LSLL-S3-5 S/H GCG I.LLLS S/H GCG VII.LLLS-S1-3S2-4 S/H GCG III.LLLS-S2-4 WBh AGC UCA I.LLSS W/H GAA I.SLSL W/H UCA I.LLSS W/S AAU II.SLLL-S2-5 W/S AGC I.LLSL W/W AAU II.LLLL-S2-5 W/W AAC ACA V.LSSS-P1-4 W/W ACA I.LSSS	W/W UAA IV.LLLL-S1-3	W/H ACC I.LSLS W/H ACC AAA AAU GAA GAU W/W UAA UAU UUU W/W AGA II.SLLL-S2-5 W/W CAU V.LLLL-S1-4 W/W CCG II.SSLL-P2-5 W/W GCA II.LLSL-S2-5 W/W GCU II.SLLL-S2-5 W/W UAA I.LLSL W/W UUU I.LLLL W/W UAA I.LLSS W/W UAA I.SLSL	
G	S/H AAA AAC AAG AAU VI.LSLL-S3-5 AGA AGC CAA CAC CAG CAU CGA CGC CGU S/H ACG AAG III.LLLS-S2-4 S/H ACC ACG AGA CGA IV.LLLS-S1-3 GCC GCG GGA UCC UCG UGA S/H ACG AUA AUG III.LLLS-S2-4 S/H AGA LLLS-S1-3S2-4 S/H AGU VI.LSLL-S3-5 S/H CAA VI.LSLS-S3-5 S/W AAC AAU CAC CAU VI.LSLL-S3-5 GAC GAU W/H AGC I.LLSS W/H AUA I.LLLS W/H CCC I.LLLL W/W AAU II.SLLL-S2-5 W/W ACA IV.LLLL-S1-3 W/W CCG V.LSSS-P1-4	S/W UUG I.LLLL S/W UUG I.LLLS W/Bs UCC UCU I.LSLS W/H AUA I.SLLL W/W AAA AAU IV.LLLL-S1-3 W/W AAU AUA I.SLLL W/W AAU II.SLLL-S2-5 W/W ACU AGA CAA CAG IV.LLLL-S1-3 CAU CGA GCA GCU GGA UAG UAU UCA UCU UGA W/W UCC I.LSLS	H/W CCC I.LSLS W/H CAU I.LLLL W/H CAU UGC I.LLSL W/H GAA I.SLSL W/W CCG II.SSLL-P2-5	S/H CGG I.LLLL	
U	S/Bs GCG GCA VI.LLLL-S3-5 S/H CAU CAA CAC CAG VI.LSLL-S3-5 S/H CUG CUA VI.LLLL-S3-5 S/H CAA VI.LSLS-S3-5 W/H UAA I.SLSS W/H GAC I.LLLL W/H UAA UCG I.LLSL W/H GCA UCG I.SLLL W/H GAA GAG GCA GGA I.SLSL GGG UAA UCA UCG UUA UUG W/W AAU II.LLLL-S2-5 W/W AAU II.SLLL-S2-5		H/S CGC CGG CGU VIII.LLLL-P1-3S3-5 H/S CGC CGG CGU VI.LLLL-S3-5 H/W CCC I.LSSS W/W GCA II.LLSL-S2-5 W/W GCU II.SLLL-S2-5 W/W UAA III.LLSS-P2-4	W/H UAA I.SLSL W/H UUC I.LLLS W/W GAC I.LLSS W/W GCG I.LSSS	

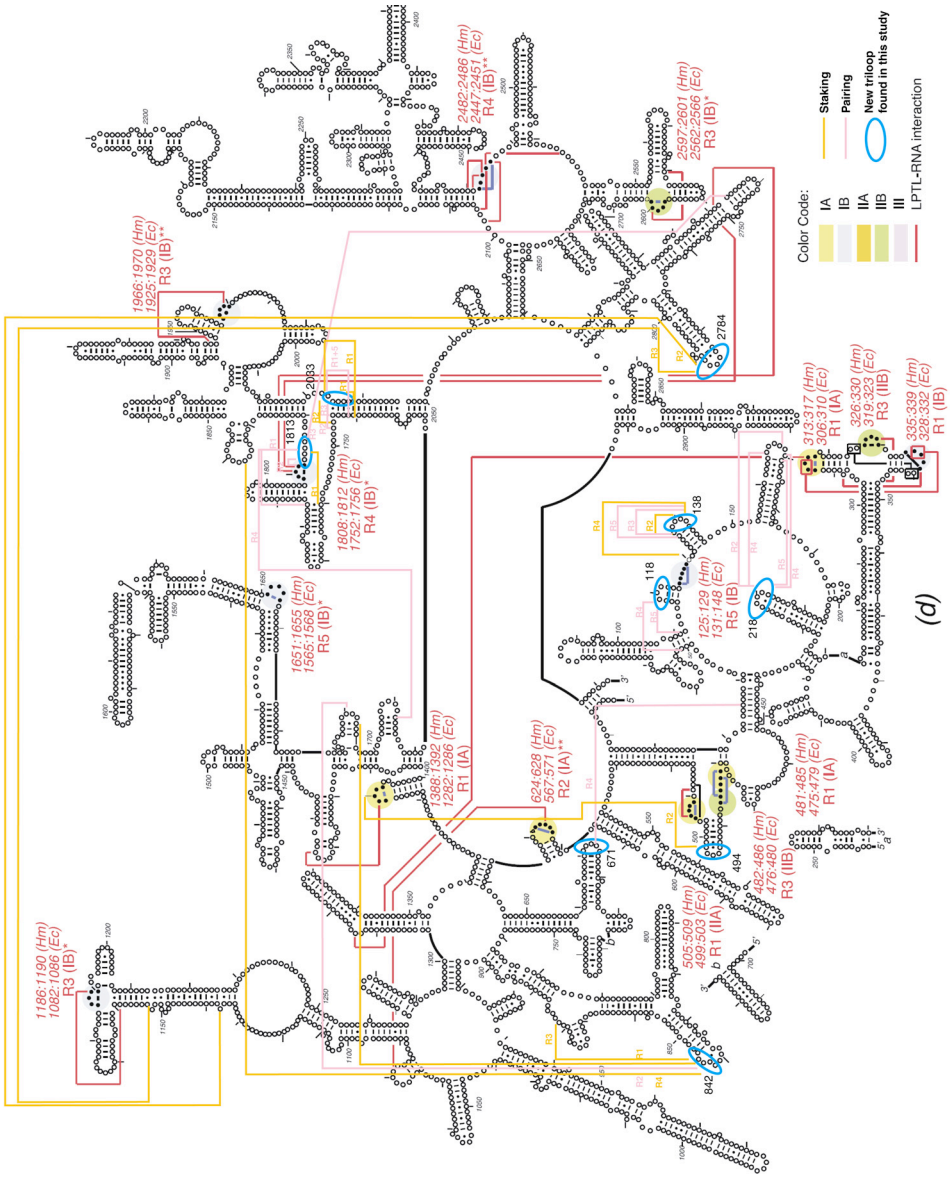


Figure A3-1 Secondary structure of the 23S rRNA subunit of *H. Marismortui*
Secondary structure of the 23S rRNA subunit of *H. Marismortui*, sowing the trillops found by (Lee, Cannone et al. 2003) in various colors. The nine additional triloops we found are shown using cyan ellipses. The tertiary interactions of the new triloops are shown using orange lines for stacking interactions and pink lines for pairing interactions. The triloop nucleotides involved in the tertiary interactions are represented by the letter "R", followed by their nucleotide numbers in the triloop (1 to 5). The original Figure and authorization to update were kindly provided and given by Dr. Robin Gutell.

Annexe 4. RNA Sequence Design Using a Three-Dimensional Quantitative Structure-Activity Relationships Approach (Supplementary Information)

Table A4-I Activity predictions

Sequences are identified with the number (or WT) to their left in the first column. WT identifier is for the wild-type sequence, from the 23S rRNA of *E. coli* (B2652-B2668). Mutations in sequences are shown in gray in the second column. Crosses are used to identify sequences that are in the alignment of bacterial 23S rRNA sequences in the third column. The origin of NCM [27] is mentioned, in the fourth to tenth column, for each NCM and each sequence, where "blank" indicates that we use all occurrences of the NCM in the PDB (Protein Data Bank); "make" indicates that the NCM does not appear in the PDB and we build it using base pairing substitution into a backbone template from the PDB; "mut." indicates that the NCM does not appear in the PDB and it is not possible to construct it as described by the "make" and therefore we mutate the nucleobase identity to the specified sequence into the WT NCM; bold identifiers indicate that the 2655-2656 nucleotides are not paired; italic identifiers indicate that the mutation is made into other NCMs than the WT one; and underline identifiers indicate that base pair types are not the same as the WT one. The number of models for each sequence is shown, the minimum free energy [29] and the minimal distance from the seed structure in the eleventh to thirteenth column. The model used for each sequence is indicated in the fourteenth column; MFE for the minimum free energy model and RMSD for the closest model in RMSD of the seed structure. The activity (viable/lethal for growth of cells from [23][24][25][26] and our experiments) and the activity prediction (viable/lethal) are shown for each sequence in the fifteenth to sixteenth column.

IDs	Sequences	Align.	NCM 1	NCM 2	NCM 3	NCM 4	NCM 5	NCM 6	NCM 7	Nb. models	Min. free energy	Distance from seed	Model	Act	Pred
WT	CUAGUACGAGAGGACCG	×								82	-154.244	0.292780	MFE	V	V
WT	CUAGUACGAGAGGACCG	×								82	-154.244	0.292780	RMSD	V	V
01	CUAGUAGGAGACGACCG									123	-153.915	0.545254	MFE	L	L
02	CUAGUAUGAGAAGACCG	×								38	-141.877	0.887899	MFE	V	V
02	CUAGUAUGAGAAGACCG	×								38	-141.877	0.887899	RMSD	V	V
03	CUAGUAAGAGAUGACCG									297	-148.853	0.399450	MFE	L	L
04	CUAGUACGAGACGACCG									4755	-137.172	1.043702	MFE	L	L
05	CUACUACGAGAGGACCG	×			make					2716	-137.959	0.853049	MFE	L	L
06	CUAAUACGAGAGGACCG				make					687	-147.103	0.720840	MFE	V	V
06	CUAAUACGAGAGGACCG				make					687	-147.103	0.720840	RMSD	V	V
07	CUAUUACGAGAGGACCG				make					1888	-134.788	0.839605	MFE	L	L
08	CUAGUACGAGCGGACCG								mut.	5462	-122.86	0.744745	MFE	L	V
09	CUAGUACGAGGGGACCG								make	1230	-145.995	0.611596	MFE	L	L
10	CUAGUACGAGUGGACCG								make	328	-139.397	0.725250	MFE	L	L
11	CUAGUACGUGAGGACCG									36	-148.779	0.660903	MFE	V	V
11	CUAGUACGUGAGGACCG									36	-148.779	0.660903	RMSD	V	V
12	UUAGUACGAGAGGACCG	×								3296	-138.229	0.705416	MFE	V	V
13	CUAGUACGAGAGGACCA	×								3404	-139.714	1.039440	MFE	V	V
14	AUAGUACGAGAGGACCU	×								3316	-135.676	0.667155	MFE	-	L
15	UUAGUACGCAAGGACCG	×								6807	-142.586	0.756959	MFE	-	V
16	CUUGUACGAGAGGACCG	×			mut.					328	1350.44	4.530229	MFE	-	V
17	UUUGUACGAGAGGACCA	×		make	mut.					904	1618.42	4.692289	MFE	-	L
18	UUAGUACGAGAGGAUUU	×			make					10000	-137.25	1.134765	MFE	-	L
19	CUAGUACGAGAGGCCCG	×			make	Make				10000	790.175	1.181008	MFE	V	V
20	CUAGUAAGAGAAGACCG						mut.	mut.		424	-93.1315	0.921480	MFE	-	L
21	CUAGUAAGAGACGACCG									54	-129.534	0.924141	MFE	-	L
22	CUAGUAAGAGAGGACCG						mut.	make		440	-91.5752	0.683969	MFE	-	L
23	CUAGUACGAGAAGACCG									3567	-138.342	0.799503	MFE	-	L
24	CUAGUACGAGAUGACCG									21	-106.055	1.017566	MFE	-	L
25	CUAGUAGGAGAAGACCG						make	mut.		46	-104.791	0.880023	MFE	-	L
26	CUAGUAGGAGAGGACCG									76	-123.753	0.887077	MFE	L	L
27	CUAGUAGGAGAUGACCG									168	-143.322	0.682374	MFE	-	L
28	CUAGUAUGAGACGACCG									39	-126.435	1.316499	MFE	V	V
29	CUAGUAUGAGAGGACCG									33	-147.811	0.470994	MFE	-	V

30	CUAGUAUGAGAUGACCG						1349	-141.013	1.123861	MFE	V	V
31	UCAGUAGACCGGAGACG		make	Make	*mut.*	*mut.*	*mut.*	586	2086.3	5.368651	MFE	-	V
32	CUAGGACGAGAGUCACG		make	Make				0	-	-	-	-	-
33	AGAGUCGACUAGGGACC		make	mut.	*mut.*	*mut.*	*mut.*	1806	800.374	4.652008	MFE	-	V
34	CGAGUCACUAGGGACC		make	mut.	*mut.*	*mut.*	make	0	-	-	-	-	-
35	GGAGUAGACCACUCGGA		make	Make			make	10302	-52.8482	3.597762	MFE	-	L
36	AGAGUACCUCGGAGACG		make	Make		*mut.*	*mut.*	67321	1.005e+07	4.192982	MFE	L	L
37	GGAGUACCGAGGACUCA		make	Make		*mut.*	*mut.*	1476	-27.0752	5.000734	MFE	-	V
38	CGAGUAGACGGGACUCA		make	Make	*mut.*	mut.	*mut.*	0	-	-	-	-	-
39	GGUACUCGAGAGGACCA	make	mut.	mut.				74650	-98.6177	1.928139	MFE	-	L
40	UCGGGACGAGAGUACCA	make	make	Make				53679	-107.701	3.467435	MFE	L	L
41	AGGACUCGAGAGGUACC		make	mut.				71198	1.000e+08	6.110375	MFE	-	L

Table A4-II Partial charges.

Partial charge [32] for each atom (in row) in each type of nucleotide (in column).

Atoms/Nucleotides	A	C	G	U
C2	0.5875	0.7538	0.7657	0.4687
C4	0.3053	0.8185	0.1222	0.5952
C5	0.0515	0.5215	0.1744	-0.3635
C6	0.7009	0.0053	0.4770	-0.1126
C8	0.2006		0.1374	
N1	-0.7615	-0.0484	-0.4787	0.0418
N2			-0.9672	
N3	-0.6997	-0.7584	-0.6323	-0.3549
N4		-0.9530		
N6	-0.9019			
N7	-0.6073		-0.5709	
N9	-0.0251		0.0492	
O2		-0.6252		-0.5477
O4				-0.5761
O6			-0.5597	
H1			0.3424	
H2	0.04			
H3				0.3164
H5		0.1928		0.1811
H6		0.1928		0.2188
H8	0.1553		0.1640	
1H2			0.4364	
2H2			0.4364	
1H4		0.4234		
2H4		0.4234		
1H6	0.4115			
2H6	0.4115			
C1’	0.0394	0.0066	0.0191	0.0674
C2’	0.0670	0.0670	0.0670	0.0670
C3’	0.2022	0.2022	0.2022	0.2022
C4’	0.1065	0.1065	0.1065	0.1065
C5’	0.0558	0.0558	0.0558	0.0558
O2’	-0.6139	-0.6139	-0.6139	-0.6139
O3’	-0.5246	-0.5246	-0.5246	-0.5246
O4’	-0.3548	-0.3548	-0.3548	-0.3548
O5’	-0.4989	-0.4989	-0.4989	-0.4989
H1’	0.2007	0.2029	0.2006	0.1824
H2’	0.0972	0.0972	0.0972	0.0972
H3’	0.0615	0.0615	0.0615	0.0615
H4’	0.1174	0.1174	0.1174	0.1174
1H5’	0.0679	0.0679	0.0679	0.0679
2H5’	0.0679	0.0679	0.0679	0.0679
HO2’	0.4186	0.4186	0.4186	0.4186
P	1.1662	1.1662	1.1662	1.1662
O1P	-0.7760	-0.7760	-0.7760	-0.7760
O2P	-0.7760	-0.7760	-0.7760	-0.7760

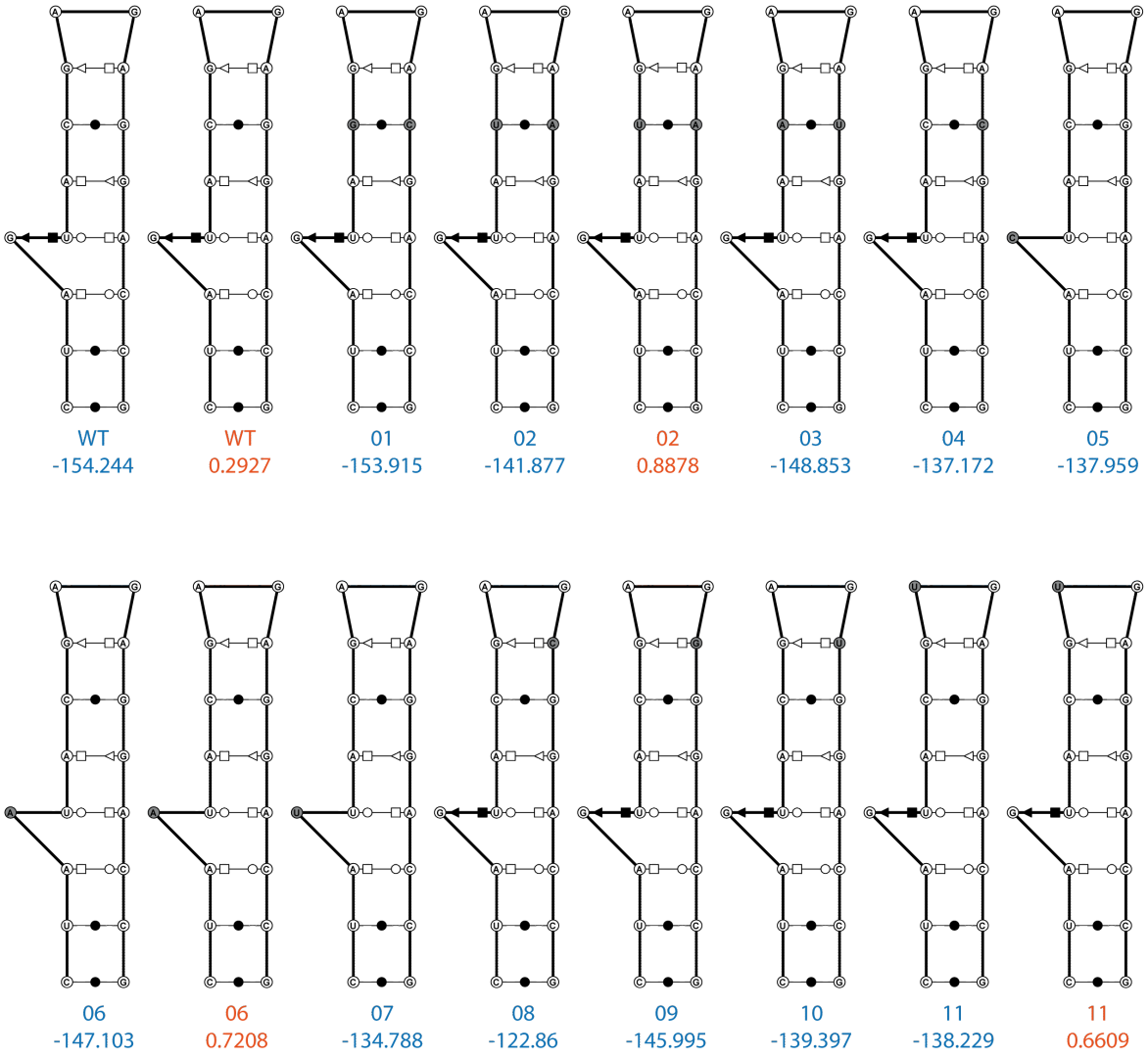


Figure A4-1 Training set models.
The tertiary structure of the models used for the training set. Models are labelled, under tertiary structure, using IDs of sequences (WT, 01, …, 11). Under labels, the free energy in kcal/mole (blue) and the RMSD in Angstrom (orange) of each model are shown. The models are annotated using *MC-Annotate* (Gendron, Lemieux et al. 2001). Canonical base pairs are represented with a black circle according to Leontis-Westhof notation (Leontis and Westhof 2001), sugar edge is represented with a triangle and hoogsteen edge with a square. Filled symbol indicates that the base pair is in cis orientation and blank symbol in trans. Dark line represents phosphodiester link. Nucleotide's gray background indicates mutations from WT sequence.

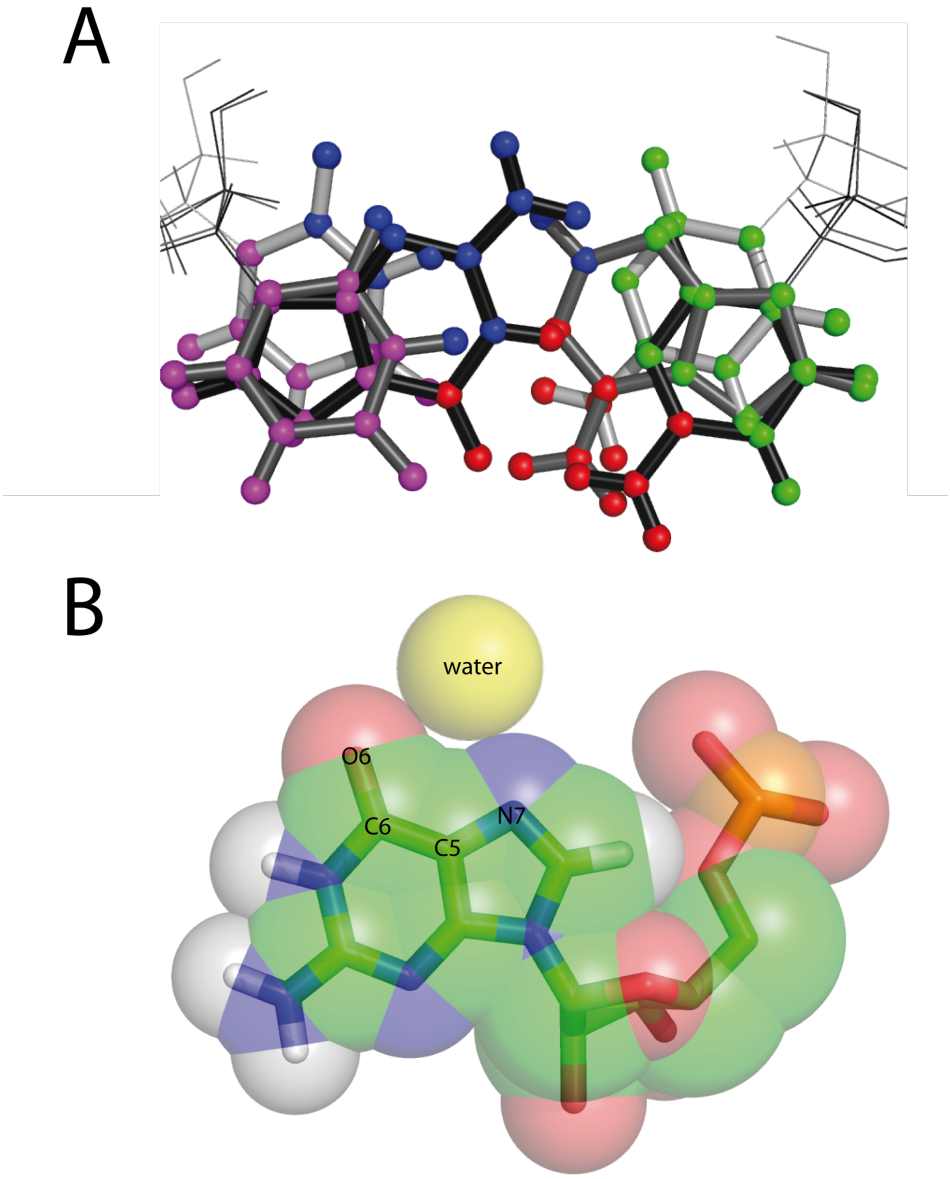


Figure A4-2 Examples.
A) An example of an alignment of three base pairs (GC in black, UA in dark gray and UC in pale gray). Atoms used for the clustering are shown with spheres. The clustering atoms in 4 clusters are represented by each color (purple, blue, red and green). B) The accessible surface is calculated by a probe sphere (here a water molecule in yellow) as it rolls over the RNA (here a guanine (G) nucleotide represented by van der waals radius). In this example, the water molecule has access to the O6 and N7 atoms, but not to the C5 and C6 atoms. Nitrogen atoms are in blue; oxygen in red; carbon in green; phosphate in orange and hydrogen in white.

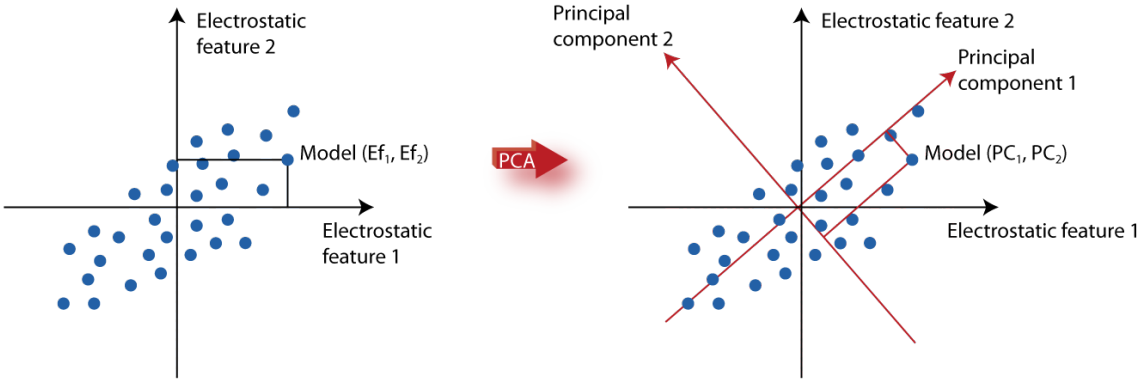


Figure A4-3 Principal component analysis.
An example of PCA with two dimensions. Left) Models (blue dots) are represented in a graph where coordinates of each dot are coupled (Electrostatic feature 1, Electrostatic feature 2). Right) Same models as left. The first principal component represents the axis with the most variance. The second component is the orthogonal axis with the most variance to the first one. A model is the couple (Principal component 1, Principal component 2).