Variation génétique et plasmatique des microARNs : impact sur les paramètres biologiques de l'hémostase

Florian Thibord

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Spécialité : Bioinformatique & Épidémiologie Génétique

École doctorale nº393: École Doctorale Pierre Louis de Santé Publique à Paris - Épidémiologie et Sciences de l'Information Biomédicale

réalisée

au Laboratoire de Génomique et Physiopathologie des Maladies Cardiovasculaires (UMRS 1166)

sous la direction de David-Alexandre TRÉGOUËT

présentée par

Florian THIBORD

pour obtenir le grade de :

DOCTEUR DE SORBONNE UNIVERSITÉ

Sujet de la thèse :

Variation génétique et plasmatique des microARNs - Impact sur les paramètres biologiques de l'hémostase

soutenue le 11 décembre 2019

devant le jury composé de :

M.	David-Alexandre TRÉGOUËT	Directeur de thèse
Mme	Macha NIKOLSKI	Rapporteur
M.	Hervé SEITZ	Rapporteur
Mme	Alessandra CARBONE	Examinateur
M.	Daniel GAUTHERET	Examinateur

Remerciements

Je tiens d'abord à remercier mon directeur de thèse, David-Alexandre Trégouët, pour son soutient constant et pour la confiance qu'il a placé en moi depuis qu'il m'a recruté en stage de master, et pour m'avoir proposé de poursuivre l'aventure avec ce projet de thèse. Sa motivation pour la science et sa bonne humeur on été très contagieuses et ont fortement contribué à rendre ces trois années de thèse agréables et émoustillantes. Il a su stimuler ma curiosité pour la bioinformatique, la biologie, et la recherche médicale, et j'ai appris énormément sous sa direction. Je lui en suis très reconnaissant.

Ce projet de thèse n'aurait pas été possible sans un financement, et pour cela je dois remercier le LabEx GenMed et ses directeurs, Jean-François Deleuze et François Cambien, qui ont accepté de prendre en charge mon doctorat, et qui m'ont également invité à présenter mon projet lors du workshop annuel du LabEx, qui fut ma première expérience de communication orale. Je les remercie également pour m'avoir permis d'accéder à la plateforme de calcul de france-génomique, qui m'aura été extrêmement utile tout au long de ma thèse (merci cobalt).

Je remercie également le professeur Pierre-Emmanuel Morange, d'abord parce que ce projet est basé sur les données extraites de la cohorte MARTHA, dont il est responsable, et ensuite pour son aide et ses conseils sur mes différents projets. Je remercie également les membres de son équipe Marseillaise: Pierre, Manal, Louisa et Noémie, qui ont également souvent contribués à mes travaux.

Je tiens également à remercier les rapporteurs de ce manuscrit de thèse, Macha Nikolski et Hervé Seitz, ainsi que les examinateurs, Alessandra Carbone et Daniel Gautheret, d'avoir accepté de participer à mon jury de thèse, d'avoir consacré leur temps à lire ma thèse, et surtout pour leurs retours qui m'auront permis d'améliorer ce travail.

Je remercie particulièrement les membres actifs et anciens du projet IMPROVEMENT. D'abord Claire, qui m'a beaucoup aidé dans l'analyse des données, et pour ses explications sur le protocole de séquençage. Ensuite Marine et Maguelonne, qui ont non seulement accepté de me passer le flambeau sur ce projet, mais qui m'ont également guidé pour appréhender ces données, et elles ont toutes ma gratitude pour les nombreuses discussions qui m'ont permis de mieux cerner et surmonter les difficultés de ce projet.

Du côté de Bordeaux, je remercie Gaëlle Munsch, pour son aide importante sur le dernier papier, et j'espère que son expérience de la thèse avec David sera aussi plaisante que la mienne. Et je remercie Dylan, qui m'a aidé à rendre optimiR accessible au plus grand nombre.

De retour à Paris, je remercie Lise et Wilfried, pour avoir consacré du temps et de l'énergie sur le projet PLXDC2. Je remercie également Waqas, mon ancien voisin de bureau avec qui j'ai passé deux années très agréables et instructives, et qui m'a invité a collaborer sur son projet MACARON avec Varma. Et Anne-Sophie, avec qui j'ai eu le plaisir de partager l'expérience de la thèse sous la direction du chef, et que je dois remercier pour son aide sur mes projets, mais surtout pour avoir contribué à rendre ces trois années très sympathiques. Je remercie aussi les autres membres de l'équipe 1 avec qui j'ai passé de bon moments, notamment lors des congrès, et les anciens bioinfos: Yasmine, Béa, Romain, Véronica, Émeline, Tété, pour les

discussions et les bons moments. Je tiens également à remercier l'équipe administrative pour leur soutient dans mes démarches au cours de ma thèse et pour leur sympathie.

À l'international, je remercie les chercheurs qui m'ont fourni des données pour pouvoir répliquer mes résultats sur mes différents projets, et je remercie les membres de la communauté mirtop pour leur invitation et pour avoir résolu de grandes questions que je me posais sur les isomiRs.

Enfin, je tiens à remercier mon entourage, pour leur soutient dans mon projet et pour avoir rendu ces trois années très agréables. D'abord Vincent, Raphi et Alex, avec qui j'ai eu le privilège de passer de très chouettes moments, musicalement riches, et souvent remplis d'aventures. Ensuite le reste du kiwicrew, que j'ai souvent retrouvé après une longue journée de travail pour décompresser. Puis mes colocs, Pierre et Riccardo, qui m'ont supporté jusque dans la dernière ligne droite, et mon ancienne coloc, Adeline, qui m'a beaucoup influencé dans mon choix de faire une thèse, ce pour quoi je lui suis très reconnaissant. Pour terminer, je tiens à remercier ma famille et particulièrement mes parents, Thierry et Corinne, pour leurs encouragements depuis toujours.

Table des matières

Liste d'abréviations

ADN Acide désoxyribonucléique
ARN Acide ribonucléique
miARN microARN
ARNm ARN messager
lncARN long ARN non codant
3'UTR 3' untranslated region
nt nucléotide
NGS Next Generation Sequencing
qRT-PCR quantitative Reverse Transcriptase Polymérase Chain Reaction
RISC RNA induced silencing complex
TNTase Terminal nucleotydil transferase
NTA Non templated addition
TDMD Target directed microRNA degradation
SNP Single nucleotide polymorphism
eQTL expression Quantitative Trait Locus
LD Déséquilibre de liaison
KO Knock Out
VTE Thrombose veineuse
PE Embolie pulmonaire
TVP Thrombose veineuse profonde
aPTT temps de céphaline activée
PT temps de prothrombine
TGP potentiel de génération de thrombine
PAI-1 Plasminogen activator inhibitor 1
TFPI Tissue factor pathway inhibitor

Introduction

Les microARNs sont des petites molécules d'ARN possédant une taille moyenne de 22 nucléotides. Ils sont produits par de nombreux organismes eukaryotes, et plus de 2000 microARNs ont été identifiés chez l'homme. Leur principale fonction est de réguler l'expression des gènes, en inhibant la traduction d'ARN messagers en protéines dans le cytoplasme. La régulation d'un gène par un microARN est rendue possible grâce à l'association par complémentarité de séquence d'un microARN avec un ARN messager. La majorité des gènes exprimés chez l'homme sont susceptibles d'être ciblés par des microARNs, qui sont ainsi potentiellement impliqués dans de nombreuses voies biologiques.

Depuis leur découverte il y a une vingtaine d'années, les microARNs ont fait l'objet de nombreux travaux visant à identifier les mécanismes impliqués dans leur synthèse et dans la régulation de la traduction des ARN messagers. Plusieurs études ont également identifié des microARNs dont l'expression anormale est associée au développement de phénotypes délétères. Ces petites molécules sont ainsi devenues un sujet de recherche biologique et médicale majeur, afin de découvrir leur implication dans le développement de nombreuses pathologies, dont le cancer, les maladies neurodégénératives ou les maladies cardiovasculaires. Les microARNs sont également des molécules très stables, et ils peuvent être relocalisés en dehors des cellules. On les retrouve notamment dans des biofluides facilement accessibles tels que le plasma, la salive ou l'urine. Ces microARNs circulants pourraient ainsi avoir une utilité en tant que biomarqueurs dans un cadre clinique, par exemple pour prédire le développement d'une maladie, ou pour évaluer la réponse à un traitement.

Ce projet de thèse a été développé afin d'étudier le profil plasmatique des microARNs dans une cohorte de patients ayant développé une ou plusieurs thromboses veineuses. Cette cohorte a été constituée dans le cadre d'un axe de recherche sur les déterminants moléculaires et génétiques de la thrombose veineuse, dirigé par David-Alexandre Trégouët et Pierre-Emmanuel Morange. Pour mener ces recherches, plusieurs milliers de patients atteints de thrombose veineuse ont été recrutés au sein de l'hôpital de La Timone (Marseille, France) entre janvier 1995 et octobre 2012, pour former la cohorte de patients MARTHA (MARseille THrombose Association). Pour chaque patient, des échantillons sanguins ont été collectés au moment de leur entrée dans l'étude, afin d'établir le profil génétique des participants et de mesurer différentes variables liées à l'hémostase.

Pour 435 patients de la cohorte MARTHA, des échantillons de plasma sanguin ont été extraits afin de réaliser un séquençage des microARNs. Une analyse bioinformatique est ensuite requise pour détecter et quantifier les microARNs à partir de ces données de séquençage, afin d'obtenir le profil plasmatique des microARNs chez ces patients. Toutefois, la détection et la quantification des microARNs est une tâche complexe, notamment à cause de leur petite taille, de la similarité de séquence pouvant exister entre différents microARNs, et des modifications de leur séquence qu'ils peuvent subir à cause de variants génétiques ou de modification post-

transcriptionnelles.

Le premier objectif de mon projet de thèse a consisté à développer un outil bioinformatique permettant la détection et la quantification précise de microARNs à partir de données de séquençage. Cet outil devait implémenter une stratégie adéquate pour distinguer les microARNs avec des séquences similaires, et prendre en compte les potentielles variations de séquences des microARNs, notamment en intégrant l'information génétique disponible pour chaque échantillon afin d'améliorer la détection des microARNs avec des variations génétiques.

Le second objectif a consisté à identifier des associations entre les niveaux plasmatiques de microARNs obtenus et un ensemble de variables biologiques et cliniques. Premièrement, j'ai pu identifier des variants génétiques pouvant influencer les niveaux plasmatiques des microARNs grâce à des études d'associations pangénomiques. Ensuite, j'ai recherché les microARNs dont les niveaux plasmatiques sont associés à la récurrence d'épisodes thrombotiques, afin de découvrir des microARNs pouvant servir de biomarqueurs pour prédire le risque de récidive de thrombose. Finalement, j'ai effectué des analyses de corrélation entre les niveaux plasmatiques de microARNs et les variables biologiques de l'hémostase disponibles pour cette cohorte. Collectivement, ces analyses m'ont permis d'identifier des microARNs d'intérêt pour la recherche sur la thrombose veineuse et sur l'hémostase.

Au cours de ce travail, j'ai eu l'occasion de m'intéresser aux travaux de recherches sur les microARNs réalisés depuis leur découverte. J'ai synthétisé le fruit de cette recherche bibliographique au sein du premier chapitre de cette thèse, qui offre un aperçu détaillé des mécanismes de biogénèse des microARNs, de la régulation des ARN messagers, et des mécanismes régulant l'expression et la fonction des microARNs. Une annexe détaillant les fondamentaux de la biologie moléculaire est également disponible, pour couvrir certains éléments qui ne sont pas détaillés dans le premier chapitre. Le second chapitre présente ensuite les méthodes disponibles pour quantifier les microARNs, ainsi que le détail du protocole suivi dans cette étude pour le séquençage des microARNs. Le troisième chapitre détaillera la stratégie mise en oeuvre pour détecter et quantifier les microARNs à partir de données de séquençage. Enfin, le quatrième chapitre présente les mécanismes de l'hémostase et la maladie thromboembolique veineuse, et détaille les études d'associations effectuées à partir des niveaux plasmatiques de microARNs. J'ai également eu l'opportunité de participer à différents projets de recherches au cours de ma thèse, ces travaux annexes seront présentés dans le dernier chapitre.

Chapitre I : Les microARNs chez l'homme

Introduction : du microARN au miRnome

La découverte des microARNs

Le premier microARN a été découvert au début des années 1990 grâce à l'effort combiné de deux groupes de recherche dirigés respectivement par Gary Ruvkun et Victor Ambros. Ces deux groupes se sont initialement intéressés à un mécanisme impliquant deux gènes, *lin-4* et *lin-14*, qui contrôlent la temporalité du développement larvaire chez le nématode caenorhabditis elegans (C. elegans).

Dans un premier temps, le groupe de Ruvkun fit la découverte d'une délétion englobant une partie de la région 3'UTR du gène *lin-14*, qui provoque un phénotype anormal similaire à celui d'un nématode mutant dans lequel *lin-4* n'est pas exprimé [287]. Les auteurs ont alors émis l'hypothèse qu'une interaction entre les produits de ces deux gènes, via la partie 3'UTR de *lin-14*, était nécessaire pour observer un développement normal du nématode. Le groupe d'Ambros a ensuite découvert que le produit du gène *lin-4* n'est pas une protéine, mais qu'il est transcrit en deux petits ARNs non codants. Le premier possède une taille de 61 nucléotides (nt) pouvant se replier sur lui même pour obtenir une structure en forme d'épingle. Le second ARN est plus petit, avec une taille de 22 nt, et possède la même séquence que celle observée à l'extrémité 5' du premier ARN (Figure I.1.A), et est donc potentiellement dérivé de celui-ci par un mécanisme post-transcriptionnel qui était alors inconnu [168].

De plus, ces deux groupes ont remarqué que le petit ARN issu de *lin-4* possèdent une séquence partiellement complémentaire à plusieurs sites localisés dans la région 3'UTR de *lin-14* (Figure I.1.B). Plusieurs sites complémentaires sont d'ailleurs contenus dans la région impactée par la délétion précédemment raportée [287] qui provoque un développement anormal de C. elegans.
Finalement, le groupe de Ruvkun démontra l'importance de cette complémentarité de séquence entre les ARNs issus de *lin-4* et les sites localisés en 3'UTR de *lin-14*. Pour cela, il mit en évidence une diminution des niveau de la protéine LIN-14 lorsque cette complémentarité est effective et que *lin-4* est exprimé [288]. Les auteurs déduisent ainsi qu'une interaction lin-4:lin-14 en 3'UTR de *lin-14* est une régulation post-transcriptionnelle singulière, permettant d'inhiber la traduction de cet ARN en protéine.

Quelques années plus tard, le groupe d'Ambros découvre un second gène, nommé *lin-28*, soumis à la même régulation post-transcriptionnelle par le petit ARN issus de *lin-4* [204]. Un site de fixation par complémentarité est identifié, également en 3'UTR de ce gène. Dès lors, ce mécanisme de régulation post-transcriptionnelle par *lin-4* n'est plus restreint à un seul gène,

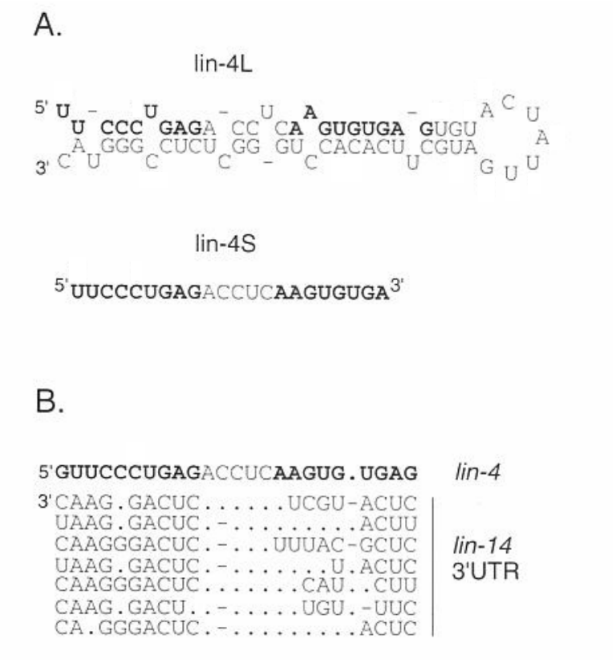


Fig. I.1 – (A) Séquence et structure secondaire des petits ARNs lin-4L (L pour *long*) et lin-4S (S pour *short*). (B) Alignments complémentaires entre *lin-4* et 7 sites localisés dans la partie 3'UTR de *lin-14*. *Image adaptée de [168]*

et les auteurs suggèrent que ce mécanisme singulier est peut-être plus commun qu'ils ne le pensaient: d'autres gènes sont peut-être ciblés par *lin-4*, et surtout, il existe probablement d'autres petits ARNs non codants possédant une fonction similaires à *lin-4*.

Une nouvelle classe d'ARNs

Il faudra attendre le début des années 2000 pour qu'un nouveau petit ARN non codant avec un mécanisme similaire à *lin-4* soit découvert, encore une fois chez C. elegans: l'ARN *let-7* [235, 255]. Et contrairement à *lin-4*, qui n'est présent que chez certains nématodes, des homologues [1] de *let-7* sont présents dans de nombreuses espèces, y compris chez l'homme [223]. Cette découverte va générer un nouvel intérêt pour ces petits ARNs et, rapidement, de nouveaux petits ARNs ayant une taille moyenne de 22 nts sont identifiés chez la drosophile, la souris et chez l'homme, et le terme microARN (miARN) apparaît pour désigner la classe d'ARN les regroupant [155, 164, 167, 205, 156, 5, 112, 175, 143].
Un système d'annotation des miARNs regroupant les gènes orthologues et paralogues en familles est mis en place [4], et une base de données indexant l'ensemble des miARNs identifiés et leurs annotations est mise en ligne en 2004 sous le nom de *microRNA registry* [90]. Cette

1. Des séquences homologues correspondent à des segment d'ADN qui possèdent une origine évolutive commune, et partagent ainsi des similitudes dans leurs séquences. Deux séquences ayant une origine commune à la suite d'un évènement de spéciation sont dites orthologues, tandis que celles ayant une origine comune à la suite d'une duplication sont dites paralogues.

base de données référence alors 506 loci produisant des miARNs issus de six organismes, dont 56 chez l'homme. Cette base de données sera renommée *miRBase* par la suite, et va évoluer de façon continue jusqu'à sa dernière version en 2018 (v22.1), qui indexe désormais 38 589 loci produisant des miARNs dans 271 espèces, dont 1917 chez l'homme.

miRnome et recherche médicale

Devant le nombre croissant de miARNs découverts, il devint clair que ces petits ARNs, autrefois restreints à un mécanisme singulier du nématode, peuvent potentiellement réguler des milliers de gènes dans des centaines d'espèces. Ils sont alors suspectés d'avoir un rôle important dans des voies majeures du développement, et d'être impliqués dans de multiples pathologies. Ainsi, depuis le début des années 2000, les travaux sur les miARNs n'ont cessés de croître, si bien qu'aujourd'hui ils représentent environ 1% du volume d'articles indexés dans pubmed [2] chaque année (Figure I.2).

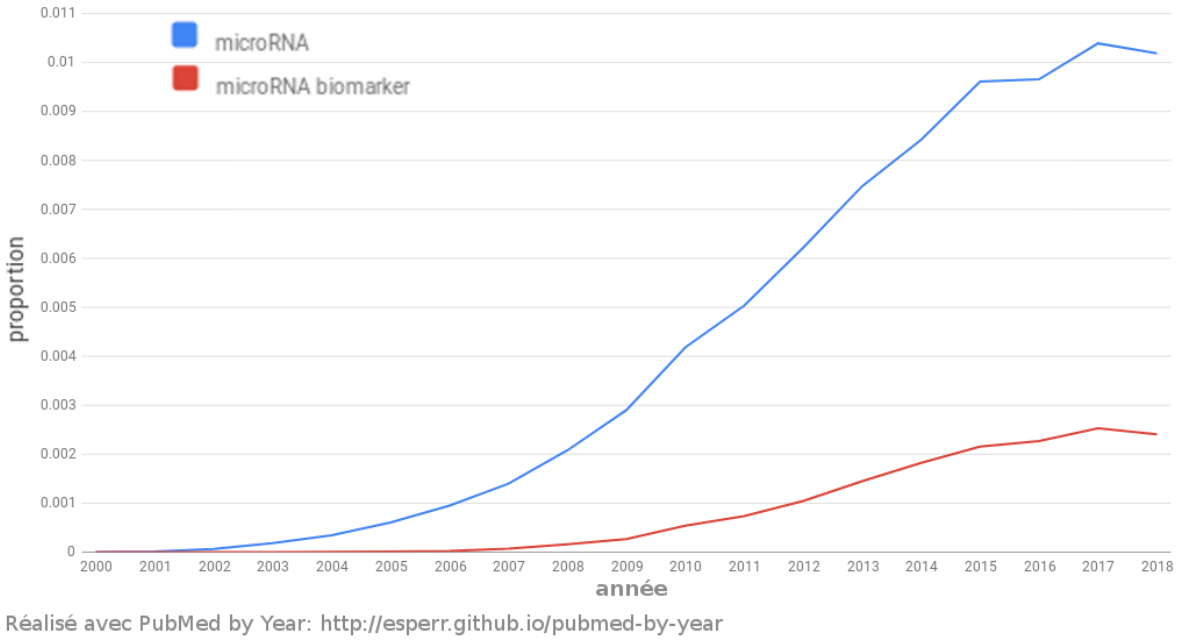


Fig. I.2 – Evolution annuelle de la proportion des articles traitant des microARNs et des microARNs en tant que biomarqueurs indexés dans PubMed

Dans la même période, les technologies de séquençage haut-débit (ou NGS en anglais pour *Next Generation Sequencing*) se sont améliorées, et sont devenues assez abordables pour devenir une alternative aux méthodes de quantification d'ARN classiques telles que les puces de séquençage et la qRT-PCR (en anglais pour *quantitative Reverse-Transcriptase Polymerase Chain Reaction*). Grâce aux méthodes NGS, il est devenu possible d'étudier l'expression des miARNs, de découvrir de nouveaux miARNs, et de détécter les variations de leurs séquences, et ainsi quantifier ce que l'on appelle le *miRnome* (ensemble des miARNs exprimés au sein d'un tissu ou échantillon). Les miARNs ont ainsi été impliqués dans de nombreux mécanismes

2. Base de données indexant les articles publiés dans des revues scientifiques médicales et biologiques

biologiques, et de nombreux travaux "miRnomique" ayant pour cadre des pathologies complexes telles que le cancer, les maladies cardiovasculaires ou neurologiques, ont ainsi permis de découvrir des variations d'expression de miARNs selon le statut des patients, et ce dans différents tissus ou biofluides [3]. Si les miARNs différentiellement exprimés n'ont pas forcément de rôle direct dans le déclenchement ou la progression de la maladie, ils peuvent toutefois servir comme biomarqueurs permettant d'établir un diagnostic, de prévoir le développement d'une pathologie, ou d'évaluer la réponse à un traitement. Pour de nombreuses pathologies, la découverte de nouveaux biomarqueurs fiables pourrait améliorer de façon considérable la prise en charge des patients. Ainsi, la recherche sur l'utilisation de miARNs en tant que biomarqueurs est en constante progression depuis une dizaine d'années (Figure I.2).

Les recherches menées ces 3 dernières décennies ont permis de définir précisément les voies de synthèse endogène des miARNs, les mécanismes qui les impliquent dans la répression de la traduction des ARNs messagers, et les régulateurs de leur expression et de leur fonction. Cette capacité à contrôler l'expression de gènes par complémentarité de séquence a eu un impact majeur sur la recherche en biologie et en médecine, avec notamment le développement de techniques d'inhibition de la traduction par petit ARN interférant (siARN) [72], qui permet de diminuer l'expression de gènes spécifiques. Ce chapitre présente l'état actuel de ces connaissances, qui sont valables chez l'homme, mais dont beaucoup de mécanismes sont partagés par l'ensemble des organismes métazoaires.

1 Transcription et maturation des microARNs

La maturation des miARNs s'effectue en deux étapes: la première dans le noyau et la seconde dans le cytoplasme. Les mécanismes responsables de la transcription et de la maturation du miARN sont présentés dans cette section. On verra ensuite que certains miARNs suivent des voies de maturation différentes, dite non-canoniques.

1.1 Transcription et maturation nucléaire par le microprocesseur

a) Transcription

La transcription des loci produisant des miARNs est généralement effectuée par l'enzyme [4] ARN polymérase II avec l'aide de facteurs de transcription, de la même façon que les ARNs messagers [5] [171, 268]. Ainsi, comme pour tout transcrit produit par l'ARN polymérase II, une coiffe [6] est ajoutée en 5' et une queue poly(A) est ajoutée en 3', protégeant ainsi ses extrémités contre la dégradation par les ribonucléases. Le transcrit est appelé miARN primaire (pri-miARN) et possède une taille de plusieurs centaines, voire plusieurs milliers de nucléotides. Le brin d'ARN peut alors se replier localement sur lui-même, selon les appariements des nucléotides qui le composent. Les appariements observés sont généralement A avec U [7] et C avec G, mais

3. Fluide biologique (salive, plasma, sérum, urine, etc).
4. Protéine impliquée dans le métabolisme facilitant les réactions chimiques.
5. Les détails de ce mécanisme sont donnés dans l'Appendice A (section I. Transcription de l'ADN en ARN).
6. La coiffe est une guanosine méthylée anti-sens (voir Appendice A).
7. U correspond à l'Uracile, qui remplace les Thymines dans les brins d'ARN (voir Appendice A).

on peut aussi observer des appariements de type *wobble*[8]. La section du transcrit correspondant à un miARN possède un repliement local typique en hélice, qui ressemble de façon schématique à une épingle à cheveux, composée d'une tige et d'une tête, avec occasionnellement quelques "bourgeons" dans la tige qui correspondent à des mésappariements. Certains pri-miARNs peuvent contenir plusieurs miARNs, et donc plusieurs repliements en épingles, et sont appelés pri-miARNs polycistroniques (Figure I.3).

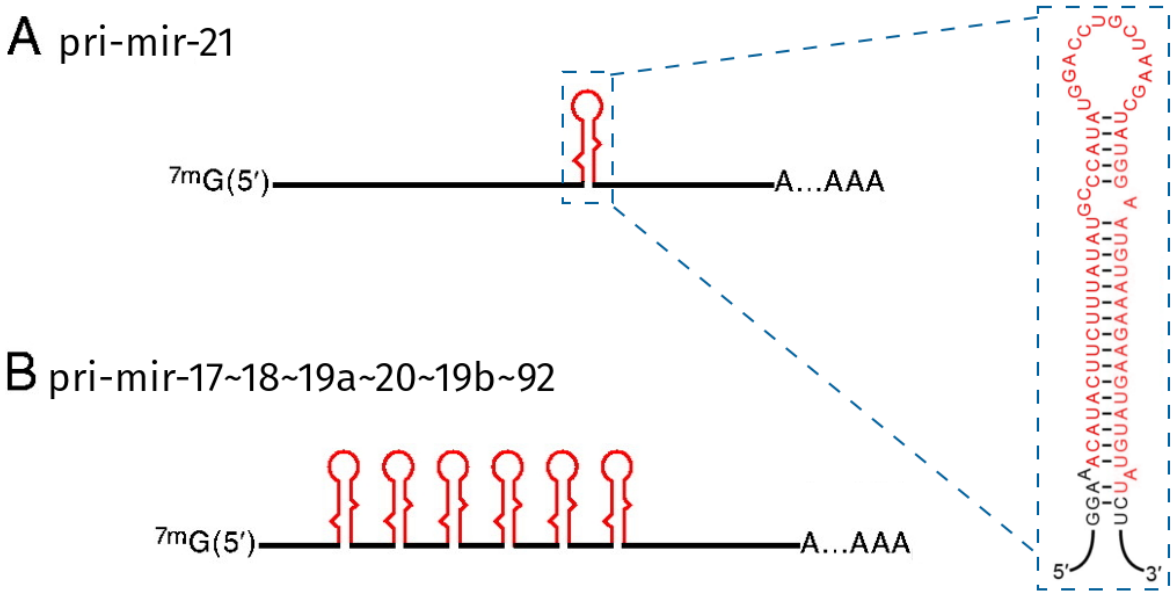


Fig. I.3 – (A) Représentation schématique du pri-mir-21, avec un zoom au niveau du repliement en épingle correspondant au site du miARN. (B) Représentation d'un pri-miARN polycistronique contenant 6 miARNs.

b) Maturation par le microprocesseur

La maturation du pri-miARN implique un complexe enzymatique appelé microprocesseur, composé de la ribonucléase[9] Drosha, et d'un couple de protéines DGCR8 [170, 160, 89]. Cette étape de maturation peut s'effectuer soit à la suite de la transcription, soit en parallèle, auquel cas la queue poly(A) n'est pas ajoutée au transcrit [18]. La ribonucléase Drosha contient un domaine de fixation aux ARNs double brins (dbARNs) lui permettant de s'associer à la tige des pri-miARNs, ainsi que deux domaines RNase III (RIIID) qui lui permettent de cliver les deux bras de l'épingle, notés 5p et 3p[10] [211]. Drosha se fixe à la base de la tige du pri-miARN tandis que le couple de DGCR8 va se lier au niveau de la tête, appelée section apicale (Figure I.4).

8. Appariement non canonique observé principalement dans les ARNs: G avec U, I avec U, I avec A, ou bien I avec C (I correspond à une Inosine, qui est une Adénosine désaminée par l'action des adénosine désaminases ADAR).

9. Enzyme responsable de la modification ou de la dégradation des ARNs. Les endoribonucléase (comme Drosha ou Dicer) découpent les brins d'ARN, tandis que les exoribonucléases retirent des nucléotides aux extrémités des brins d'ARN.

10. La tige de l'épingle est constituée de deux bras liés par la tête, qui sont notés 5p et 3p, correspondant aux côtés 5' et 3' de l'ARN, respectivement.

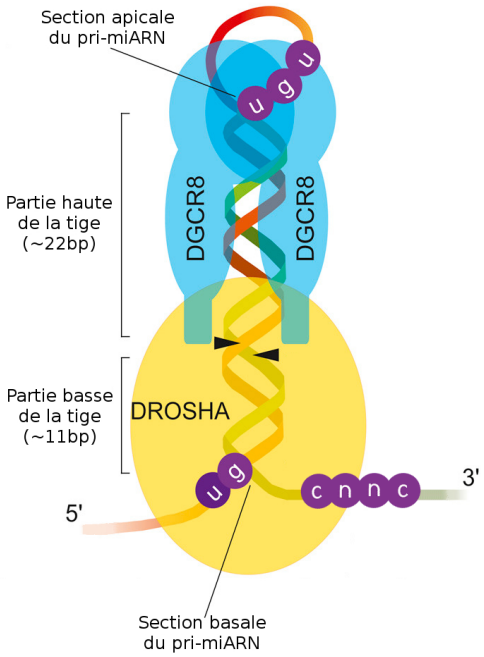


Fig. I.4 – Positionnement du microprocesseur au niveau du repliement en épingle du pri-miARN. *Image adaptée de [211]*

Certains éléments facilitent et améliorent la fonction de découpe du microprocesseur. En particulier, des motifs de séquence spécifiques permettent au microprocesseur de se positionner de façon précise au niveau du site de découpe du pri-miARN: un motif UG à la base de la tige en 5', un motif CNNC [11] à la base de la tige côté 3', et un motif UGU au niveau de la section apicale [71]. Au moins un de ces trois motifs est présents dans 79% des miARNs humains. En particulier, le motif CNNC est un motif reconnu par les protéines SRSF3 et DDX17 qui peuvent contribuer à la découpe réalisée par le microprocesseur [10, 202]. De plus, le microprocesseur est généralement apparié avec un hème [12], qui lui permet de détecter la section apicale de l'épingle et d'effectuer une découpe précise [221]. Dans certains cas, le microprocesseur recrute également d'autres cofacteurs qui facilitent son action [95, 293, 269].

Typiquement, le microprocesseur effectue la découpe à une distance d'environ 11 paires de bases (bp) de la section basale de la tige, ce qui correspond à un tour d'hélice d'ARN double brin, et à environ 22 bp de la section apicale. Suite à la découpe du pri-miARN par le microprocesseur, l'épingle est libérée et conserve sa structure en double brins. Cette épingle est appelée miARN précurseur (pre-miARN).

Parfois, le microprocesseur peut également agir en coopération avec le complexe d'épissage des introns (appelé aussi spliceosome, dont le mécanisme est présenté dans l'Appendice A). Ce mécanisme concerne une partie des miARNs localisés dans des régions intragéniques, c'est à dire au sein d'un gène dont le transcrit correspond à un ARN messager ou à un long ARN non codant (lncARN), qui est appelé gène hôte. Comme les miARNs intergéniques, certains miARNs intragéniques possèdent un promoteur indépendant et sont transcrit indépendamment du gène hôte [187, 199], tandis que le reste des miARNs intragéniques sont transcrits en même

11. D'après la nomenclature IUPAC, N représente n'importe quel nucléotide (A,C,G ou U).
12. Cofacteur contenant un atome de métal, nécessaire à l'activité biologique de certaines protéines.

temps que le gène hôte [238]. Dans ce cas, le microprocesseur peut agir en coopération avec le spliceosome, et produire le pre-miARN en parallèle à l'épissage de l'ARN hôte sans affecter son expression [130, 145, 190].

c) Transport vers le cytoplasme

La découpe par Drosha est légèrement décalée d'un bras à l'autre, et le bras 3p du pre-miARN possède 2 ou 3 nucléotides de plus que le bras 5p, ce qui permet au complexe de transport composé d'Exportin 5 (Xpo5) et RanGTP de reconnaître le pre-miARN [301, 29, 214]. Xpo5-RanGTP est chargé de transporter les pre-miARNs en dehors du noyau vers le cytoplasme, et protège les pre-miARNs d'une dégradation par des enzymes. Cependant, Lorsque Xpo5 n'est pas exprimée, on observe une réduction modeste de l'abondance de miARNs matures sans accumulation de pre-miARNs dans le noyau, ce qui suggère qu'Xpo5 n'est pas la seule protéine capable de transporter les pre-miARNs hors du noyau [144].

1.2 Maturation cytoplasmique et formation du RISC

a) Maturation par Dicer

Une fois exporté dans le cytoplasme, le pre-miARN est reconnu par la ribonucléase Dicer qui va compléter la maturation du miARN [92, 122]. Comme Drosha, Dicer possède un domaine de fixation dbARN et deux domaines RIIID, qui lui permettent de se fixer à un ARN double brins et d'effectuer une découpe. Les extrémités 3' et 5' du miARN sont fixées par Dicer, et différenciées grâce aux nucléotides supplémentaires présents en 3'. Dicer effectue la découpe des bras 5p et 3p à environ 22 nucléotides des extrémités 5' et 3', au niveau de la tête de l'épingle, pour produire un duplex de deux miARNs matures [184, 218] (Figure I.5). Tout comme Drosha, la découpe est légèrement décalée d'un bras à l'autre, ce qui laisse 2 ou 3 nucléotides en plus en 3' du bras 5p. La précision de Dicer peut être influencée par sa protéine partenaire, qui est généralement TRBP et plus rarement PACT [166].

b) Chargement du microARN dans une protéine AGO

Une fois la tête de l'épingle détachée, il reste un duplex de miARNs matures, composé des bras 5p et 3p, d'environ 22 nucléotides chacun. La dernière étape de maturation du miARN consiste à charger un des deux brins de 22 nucléotides dans une protéine AGO [13] à l'aide de protéines chaperones HSC70 et HSP90 [124], et ainsi former un complexe appelé *RNA Induced Silencing Complex* (noté RISC).

Il existe deux hypothèses concernant le mécanisme de formation du RISC. La première, appelée hypothèse "hélicase", est que le duplex de miARNs est d'abord dissocié par une protéine hélicase [14] puis l'un des deux miARNs matures est chargé dans la protéine AGO. La seconde hypothèse est que le duplex est directement chargé dans la protéine AGO. Avant de recevoir le duplex, AGO est activée grâce à l'apport d'ATP [15] qui engendre une modification de la conformation d'AGO lui permettant de charger le duplex de miARNs. Après chargement, un repliement de la structure d'AGO provoque l'expulsion de l'un des deux brins du duplex

13. Chez l'homme, il existe 4 protéines AGO homologues, notées de AGO1 à AGO4. Ces protéines font partie de la famille des Argonautes, qui comprend également les protéines PIWI chez l'homme.
14. Protéine dont le rôle est de dissocier deux brins d'ARN ou d'ADN.
15. l'ATP ou Adénosine triphosphate est un fournisseur d'énergie intervenant dans de nombreuses voies biologiques.

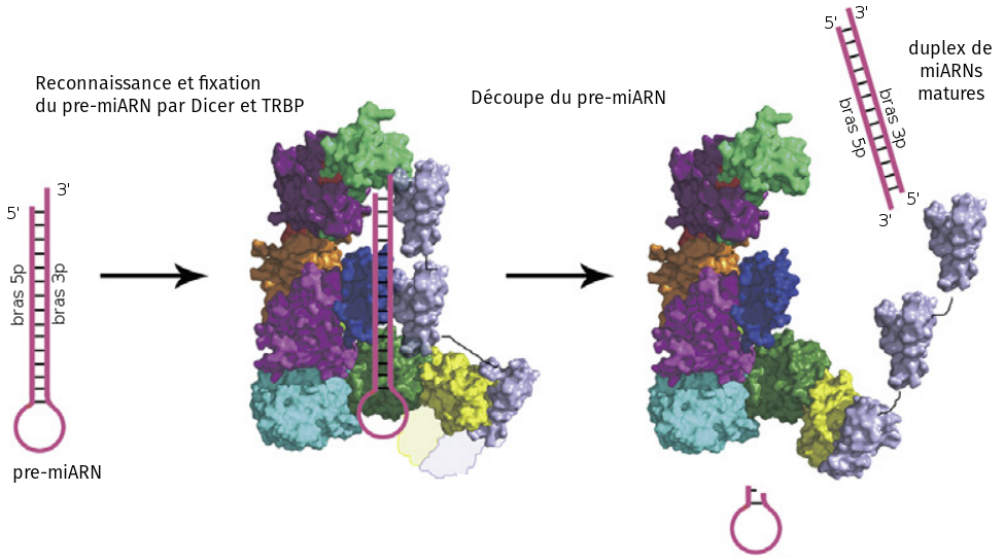


Fig. I.5 – Maturation cytoplasmique du miARN par Dicer et TRBP. *Image adaptée de [179]*

pour ne conserver qu'un brin dans la structure finale (Figure I.6). Il y a plus de preuves en faveur de la seconde hypothèse qui est aujourd'hui favorisée pour expliquer la formation du RISC [134, 208].
Le chargement est facilité lorsque quelques mésappariements de nucléotides ou des appariement peu stables de type wobble sont présents au sein du duplex, qui facilitent la séparation des deux brins [302].

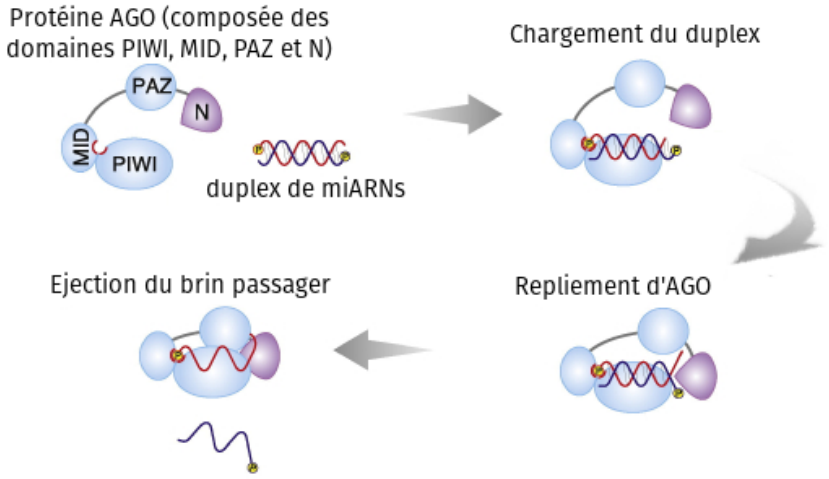


Fig. I.6 – Chargement du miARN mature dans AGO pour former le complexe RISC. *Image adaptée de [208]*

Le brin du duplex chargé dans AGO pour former le complexe RISC est alors mature et est appelé brin guide, tandis que l'autre brin est appelé brin passager (noté parfois miARN*).

Le choix du brin guide dépend principalement de l'orientation du duplex lors du chargement dans AGO. Cette orientation semble être déterminée par la composition de l'extrémité 5' des deux brins, qui est le premier élément reconnu par AGO, et qui va se placer dans une poche interne d'AGO. Cette poche favoriserait soit le brin avec une extrémité 5' possédant un A ou un U [76, 260], soit le brin avec l'extrémité 5' la moins stablement appariée au brin complémentaire [136, 248]. Toutefois, on observe que pour certains pre-miARNs, les deux brins peuvent produire des miARNs fonctionnels [300], et le choix du brin guide peut également dépendre du tissu [182]. Une fois le brin chargé dans la protéine AGO, le complexe RISC est alors effectif et peut cibler des ARNs messagers pour réguler leur traduction. Le brin passager, n'ayant aucun complexe protéique le protégeant des ribonucléases, est alors rapidement dégradé et n'est donc probablement pas fonctionnel.
Le résumé de cette voie de maturation, appelée canonique, est présentée en Figure I.7.

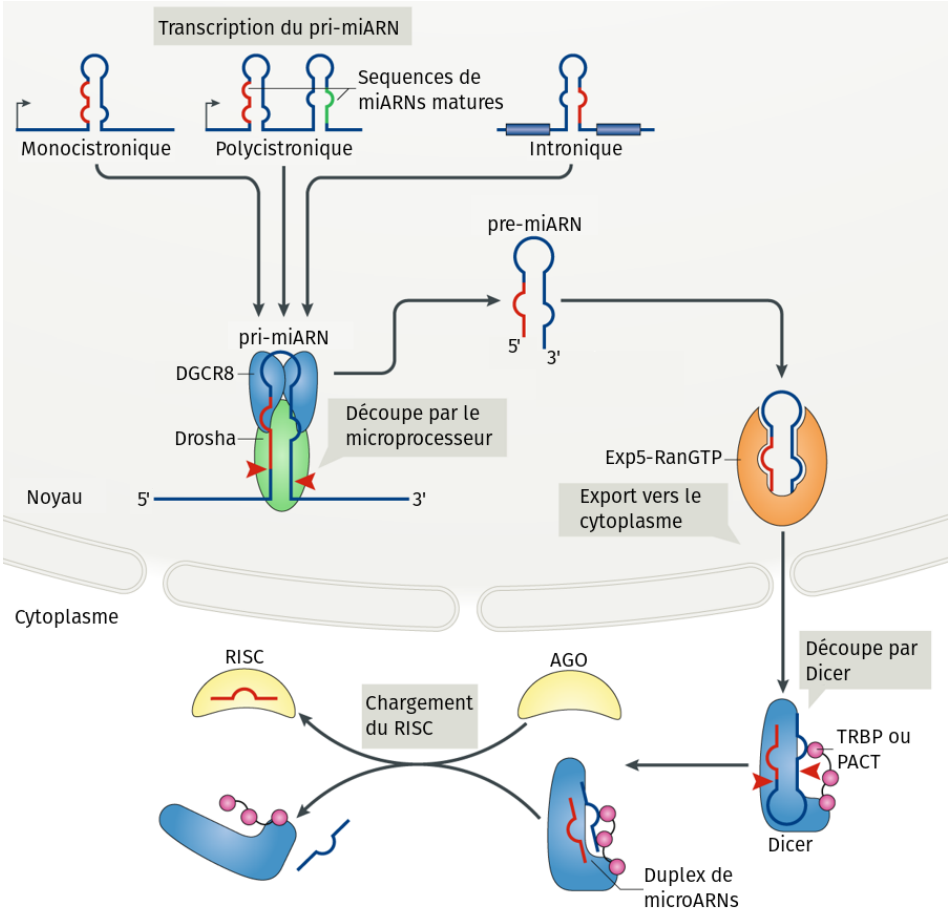


Fig. I.7 – Voie de maturation canonique des miARNs. *Image adaptée de [270]*

1.3 Voies de maturation non canoniques

La biogenèse des miARNs est réalisée en plusieurs étapes, chacune impliquant un ensemble de protéines cruciales à leurs maturations: le microprocesseur pour la maturation du pri-miARN, et Dicer pour la maturation du pre-miARN. Cependant, depuis la découverte de cette

voie canonique, il a été observé que certains miARNs n'ont pas besoin de Drosha ou de Dicer pour leur maturation, et suivent une autre voie de maturation, dite non canonique.

a) Les mirtrons

Les miARNs non canoniques les plus fréquents appartiennent à une classe de miARNs introniques, qui n'ont pas besoin du microprocesseur pour générer le miARN précurseur. Ces miARNs sont localisés dans des introns de taille réduite et l'action du spliceosome permet de générer des pre-miARNs qui peuvent ensuite être exportés dans le cytoplasme et poursuivre leur maturation [241]. Ces ARNs sont à la fois des pre-miARNs et des introns, et sont ainsi appelés mirtrons. Après épissage de l'intron correspondant au mirtron et le débranchement du Lasso [16] correspondant, l'ARN se replie localement en épingle, souvent avec une queue d'ARN adjacente en 3' ou plus fréquemment en 5' [285]. La queue d'ARN est ensuite retirée par une ribonucléase, autre que Drosha mais encore non identifiée, avant de poursuivre la maturation du miARN [14, 215] (Figure I.8). D'après la base de données mirtronDB [56], il y aurait 585 mirtrons précurseurs correspondant à 1136 miARNs matures recensés chez l'homme.

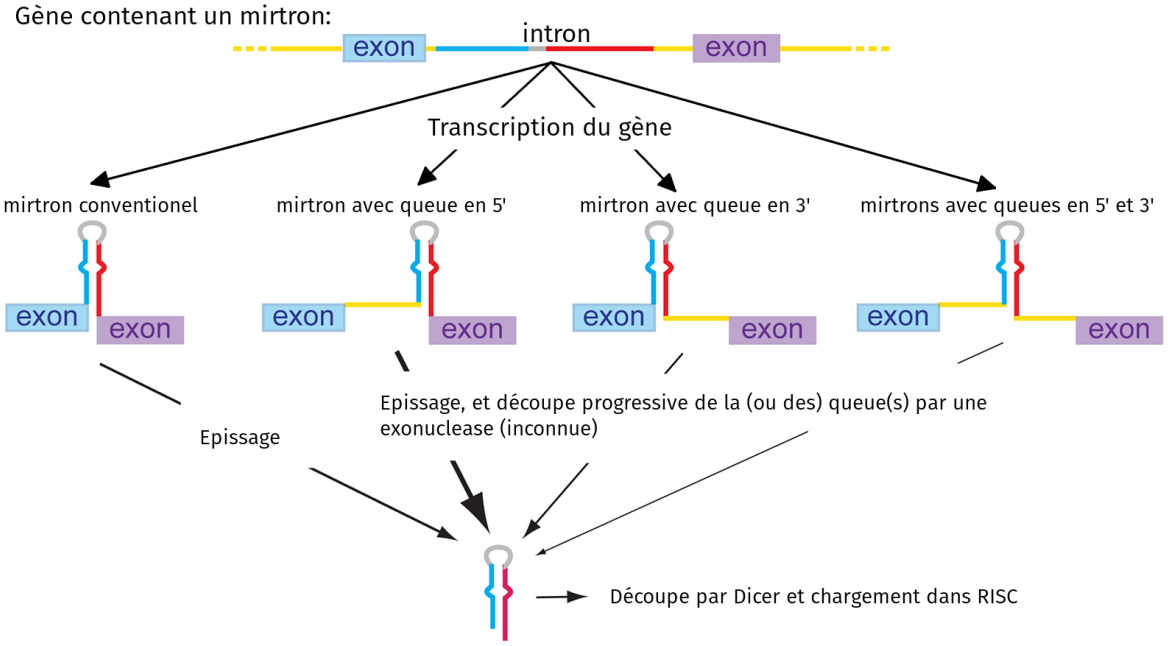


Fig. I.8 – Voie de maturation non canonique des mirtrons. L'épaisseur de la flèche lors de l'épissage reflète la fréquence de chaque type de mirtron. *Image adaptée de [285]*

b) microARNs précurseurs avec coiffe en 5'

Certains miARNs transcrits par ARN polymérase II sont directement exportés dans le cytoplasme par Exportin 1 (Xpo1) sans découpe préalable par le microprocesseur [297]. Ces

16. Lors de l'épissage d'un intron, ce dernier possède un repliement caractéristique appelé "Lasso", où l'extrémité 3' de l'intron est connectée à une Adénine au centre de l'intron.

transcrits conservent ainsi la coiffe en 5' caractéristique des transcrits de la polymérase II. Après découpe par Dicer dans le cytoplasme, c'est le brin 3p qui est systématiquement choisi comme brin fonctionnel, car la coiffe en 5' empêche le brin 5p d'être chargé dans AGO. Parmi les miARNs précurseurs avec coiffe, on trouve notamment les membres de la famille mir-320.

c) Le mir-451

Le miARN non canonique le mieux caractérisé est le mir-451. Ce miARN est très abondant dans les érythrocytes, et une carence en mir-451 entraîne un défaut de maturation des érythroblastes [224, 232]. À l'inverse des mirtrons, la maturation de ce miARN est dépendante du microprocesseur mais pas de Dicer [46]. Après la maturation du pri-miARN par le microprocesseur, le pre-miARN est trop petit pour Dicer, et est directement chargé dans une protéine AGO2 qui va elle même découper la tête de l'épingle, expulser le brin 3p, et ne conserver que le brin 5p. La ribonucléase PARN raffine ensuite la découpe du miARN mature en 3' du brin 5p. Le mir-451 est le seul miARn identifié qui suit cette voie de maturation.

d) Maturation dépendante d'une mono-uridylation

Après la découpe du microprocesseur, la plupart des pre-miARNs possèdent 2 nucléotides supplémentaires en 3', qui sont reconnus par Xpo5 et Dicer pour compléter le cycle de maturation. Cependant, certains pre-miARNs, notamment certains membres de la famille let-7, peuvent n'avoir qu'un seul nucléotide supplémentaire en 3', et nécessitent l'intervention d'une protéine de type Terminal Nucleotydil Transferase (TNTase) pour ajouter, par un mécanisme appelé mono-uridylation, un unique nucléotide Uracile supplémentaire en 3' du pre-miARN [109]. On appelle ce mécanisme *tailing* (en anglais, pour "ajouter à la queue"), et 3 TNTase ont été identifiées comme responsables de cette mono-uridylation: TUT2, TUT4 et TUT7. L'addition d'un nucléotide en 3' permet ainsi de restaurer l'extrémité 3' canonique de 2 nucléotides supplémentaires, qui est reconnue par Xpo5 et Dicer. À noter que ces TNTases peuvent effectuer plusieurs mono-uridylations d'affilée sur un même miARN, notamment en coopération avec la protéine Lin28 [110], ce qui peut entraîner la dégradation du miARN [138].

e) Une majorité de microARNs non-canoniques

Avec la croissances des analyses NGS de miARNs, de plus en plus de miARNs non canoniques sont découverts, et leur nombre dépasserait le nombre de miARNs canoniques [285]. Les miARNs non canoniques sont généralement moins bien conservés que les canoniques et sont souvent moins bien exprimés. Par exemple, les mirtrons sont rapidement dégradés chez la drosophile, après un mécanisme de poly-uridylation effectué par la (TNTase) Tailor [31]. Cela suggère qu'ils ont une fonction de régulation moins importante, ou que leurs surexpression pourrait avoir des conséquences délétères. Il existe cependant des exceptions, tels que les miARNs de la famille mir-320 et le mir-451 qui sont hautement exprimés et qui ont des fonctions importantes identifiées. Le résumé des voies de maturations des miARNs non-canoniques est représenté sur la Figure I.9.

Que les miARNs suivent une voie canonique ou non, ils sont au final tous incorporés dans une protéine AGO pour former un RISC, ce qui souligne la capacité de la protéine AGO à s'adapter aux différents types de miARNs, qui peuvent avoir une strucure singulière, comme le mir-451.

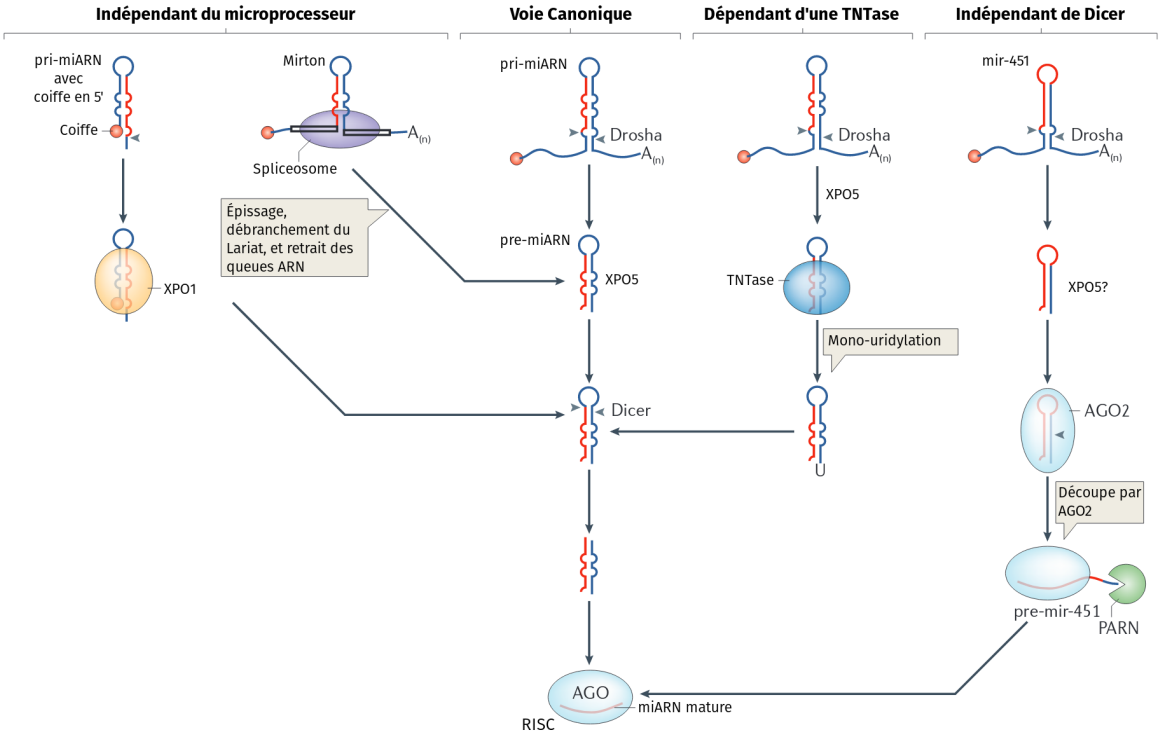


Fig. I.9 – Voie de maturation canonique et voies non canoniques. *Image adaptée de [96]*

f) Pseudo-microARNs pouvant être chargés dans AGO

En plus des duplex de miARNs, les protéines AGO peuvent aussi charger des ARNs simple brin [44]. Ces protéines sont ainsi fréquemment associées à des petits fragments d'ARN d'origine diverse (ARNt, ARNr, vARN, snoARN, snARN, yARN, ARNm, etc) [17] de taille similaire aux miARNs [39, 267], et en particulier les tRFs qui semblent impliqués dans des mécanismes de régulation similaires au miARNs [154]. Ces fragments d'ARN pourraient donc avoir une fonction au delà de celle de l'ARN dont ils sont originaire, contrairement aux miARNs qui n'ont qu'une seule forme fonctionnelle.

On peut aussi mentionner les agotrons, des petits ARNs non codants issus de petits introns, comme les mirtrons. Ces agotrons sont plus grand que les miARNs (environ 100nt) mais pourraient être impliqués dans un mécanisme de régulation des ARN messagers similaire, car ils sont également chargés dans une protéine AGO, sans avoir besoin de Drosha et Dicer pour leur maturation [105].

Les petits fragment d'ARNs sont parfois confondus avec les miARNs, comme le mir-1246 dont la séquence est identique à celle d'un fragment du snARN U2-1 [298]. Si ces pseudo-miARNs partagent une taille similaire au miARNs, leur fonction n'est pas bien établie, et leur association avec AGO n'implique pas qu'ils partagent les même rôles biologiques et mécanismes de régulation que les miARNs, et devraient donc être analysés indépendamment.

17. tARN: ARN de transfert, dont les fragments sont notés tRFs (tRNA Fragments) ; rARN: ARN ribosomal ; vARN: ARN vault ; snoARN: small nucleolar ARN ; snARN: small nuclear ARN ; yARN: ARN Y

1.4 Annotation des microARNs

Depuis leur découverte il y a 20 ans chez l'homme, de nouveaux miARNs sont régulièrement identifiés. Le véritable nombre de miARNs chez l'homme est donc encore inconnu. La base de données miRBase [90] est la principale source d'annotation des miARNs et compte chez l'homme 2588 miARNs matures correspondants à 1881 pre-miARNs dans sa version 21 [18].

Une nomenclature a été créée pour indexer ces miARNs. Chaque identifiant comprend l'espèce auquel appartient le miARN (ex: *hsa* pour l'homme ou *mmu* pour la souris), la distinction pre-miARN ou miARN mature (*mir* ou *miR*, respectivement), un numéro identifiant le miARN (attribué chronologiquement dans l'ordre de leur découverte), et pour les miARNs matures le bras 3p ou 5p du pre-miARN dont il est issu. Par example, hsa-mir-381 correspond au pre-miARN humain 381, et hsa-miR-381-3p correspondau miARN mature issu du bras 3p du précurseur hsa-mir-381.

Pour certains miARNs, des règles spécifiques sont également respéctées. Notamment les miARNs de la famille let-7, qui conservent l'identifiant d'origine (ex: hsa-let-7a-5p). Ensuite, les miARNs paralogues, correspondant à des séquences homologues qui sont apparus à la suite de duplications [25, 111], ont un nom similaire: les miARNs qui partagent une séquence identique reçoivent le même identifiant suivi par un suffixe numérique (ex: hsa-mir-7-1 et hsa-mir-7-2 sont des pre-miARNs paralogues dont la séquence est identique), et les séquences pratiquement identiques reçoivent un suffixe littéral (ex: hsa-mir-4433a et hsa-mir-4433b sont des pre-miARNs paralogues avec quelques nucléotides différents). Pour les miARNs orthologues, qui correspondent aux miARNs conservés à la suite d'un événement de spéciation, les identifiants sont unifiés d'une espèce à l'autre (ex: mmu-miR-16-5p et hsa-miR-16-5p sont des miARNs matures orthologues).

Les séquences référencées dans la miRBase proviennent généralement de travaux réalisés à partir de données NGS, mais certains miARNs référencés sont probablement des pseudo-miARNs, classifiés comme miARNs de façon erronée [50, 35, 104]. Différents critères permettant d'évaluer la fiabilité d'un miARN à partir des données de séquençage peuvent être pris en compte, notamment:

1. Les fragments séquencés, alignés sur un locus, possèdent une extrémité 5' constante (avec une tolérance de 1 ou 2 nucléotides, à cause des inconsistances occasionnelles de découpe de Dicer et Drosha)
2. Les fragments séquencés doivent s'aligner sur les bras 5' et 3' du précurseur, avec 2 nucléotides supplémentaires à chaque extrémités lorsque le précurseur est replié en épingle, typique des découpes de Dicer et Drosha (à l'exception des miARN ayant besoin d'une mono-uridylation, comme certains membres de la famille let-7)
3. Le miARN possède des orthologues, signe d'une conservation phylogénétique
4. Le précurseur possède les motifs reconnus par le microprocesseur: GU, GUG ou CNNC [19]

En se basant sur les deux premiers critères, la base de données MirGeneDB [80] a pu identifier 558 précurseurs fiables, dont la majorité possède au moins un motif reconnu par le microprocesseur, et tous sont conservés au moins chez les mammifères placentaires. Si certains véritables

18. La miRBase a récemment été mise à jour et compte désormais 2654 miARNs correspondant à 1917 pre-miARNs dans sa version 22.1 . Les analyses réalisées dans le cadre de cette thèse ont été effectuées à partir de la version 21.

19. Ce critère n'est pas nécessaire pour les mirtrons, qui complètent leur maturation sans action du microprocesseur.

miARNs ne sont probablement pas identifiés par ces critères stringents, cette base de données offre néanmoins un critère de fiabilité robuste, et indique qu'il faut être vigilant sur la nature des séquences observées dans les données NGS, ainsi que celles qui sont référencées dans la miRBase.

2 Régulation de la traduction par les microARNs

Le complexe RISC a pour principale fonction de transporter le miARN vers un ARN messager (ARNm) pour établir une interaction entre le miARN et un site complémentaire dans la région 3'UTR du transcrit. Une fois l'interaction effective, plusieurs complexes protéiques sont recrutés par RISC pour inhiber la traduction de l'ARN messager.

2.1 Interaction RISC:ARNm

Le modèle de ces interactions a d'abord été déterminé grâce à des analyses bioinformatiques, avant d'être validé par des analyses structurales du RISC.

a) Prédiction des sites de fixation par alignements de séquences

Découverte et définition de la seed Au début des années 2000, plusieurs centaines de miARNs sont découverts chez les vertébrés, mais la façon dont le complexe RISC reconnaît les sites de fixation en 3'UTR du messager n'est pas encore connue, ce qui rend difficile l'identification des ARN messagers ciblés. Déjà en 1993, les groupes d'Ambros et Ruvkun avaient montré avec des alignements de séquences entre *lin-4* et la partie 3'UTR de *lin-14*, que plusieurs sites de la partie 3'UTR étaient partiellement complémentaires avec le miARN, en particulier du côté 5' du miARN (Figure I.1). En 2002, des alignements similaires sont observés chez la drosophile, entre des séquences de miARNs et des sites atypiques dans des régions 3'UTR, car de précédents travaux avaient identifiés que ces sites en 3'UTR étaient associés à une régulation du gène [157]. Généralement, ces analyses montrent qu'un groupe de 7 ou 8 nucléotides, côté 5' du miARN, est parfaitement aligné avec un site localisé dans une région 3'UTR d'un gène régulé par le miARN.

En 2003, le groupe de Bartel et Burge reprend ce type d'analyse en étudiant les séquences de 3'UTR d'ARN messagers (n=$\sim$ 15,000) et de miARNs (n=55) conservées chez l'homme, la souris et le rat (Lewis 2003). Grâce à une analyse systématique de la complémentarité entre les régions 3'UTR, et 7 nucléotides successifs (7mer) graduellement décalés le long du miARN (allant des nucléotides 1..7 à (-7)..(-1) [20]), ils confirment que les séquences de 7mer côté 5' du miARN ont le plus de sites complémentaires dans les régions 3'UTR, et cette complémentarité est conservée entre les trois espèces (Figure I.10).

Ce résultat montre que la région côté 5' du miARN, qu'ils nomment *seed*, et en particulier le 7mer en positions 2..8, est le déterminant majeur des interactions miARN:ARNm, et plusieurs de ces interactions sont validées expérimentalement. Le premier outil de prédiction d'interactions miARN:ARNm chez les vertébrés, nommé TargetScan, est d'ailleurs basé en partie sur ces résultats [172].

20. En numérotant 1 la position du premier nucléotide côté 5', et -1 celle du premier nucléotide côté 3'.

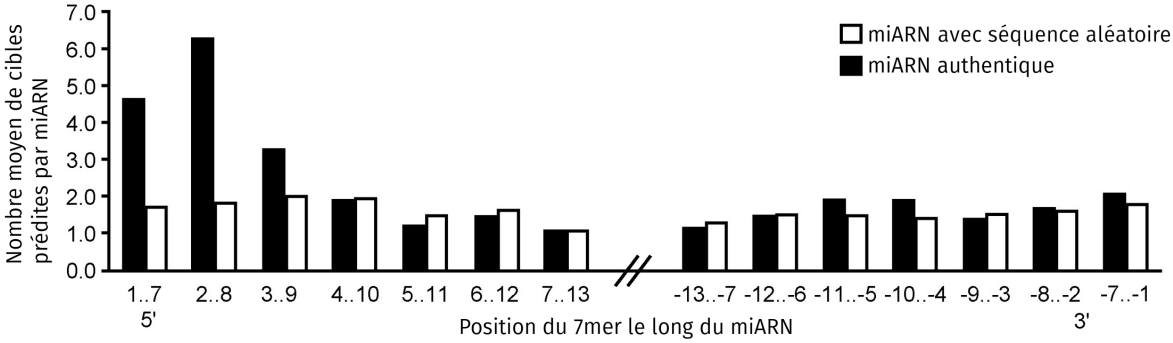


Fig. I.10 – Nombre moyen de cibles prédites par miARN selon la position du 7mer le long du miARN. Les séquences utilisées pour les miARNs et régions 3'UTR des gènes cibles sont conservées chez l'homme le rat et la souris (barres noires). Un second ensemble avec des séquences de miARNs aléatoires a également été utilisé (barres blanches) en tant que contrôle, et on observe que le nombre de sites complémentaires avec ces séquences aléatoires varie très peu selon la position du 7mer. Le nombre moyen de cibles prédites est significativement plus élevé avec les 7mer en positions 1..7 et 2..8 chez les miARNs authentiques par rapport aux contrôles. *Image adaptée de [172]*

Des travaux plus récents ont permis de raffiner ces résultats, et montrent que la seed correspond à un 7mer +/- 1 nucléotide, allant des positions 2..7 à 2..8, qui est généralement suffisant pour déterminer le spectre d'ARN messagers ciblés par un miARN [21, 79]. De plus, les interactions miARN:ARNm nécessitent généralement une complémentarité de type Watson-Crick (appariements C-G et A-T), sans appariements non canoniques tels que les wobbles [282].

Appariements supplémentaires En plus des appariements au niveau de la seed, les alignements de séquences prédisent souvent des appariements complémentaires supplémentaires côté 3' du miARN, particulièrement au niveau des nucléotides 13 à 16 [91]. Les appariements supplémentaires n'auraient généralement que peu d'influence sur l'efficacité de la répression de traduction du messager par le RISC [27, 243, 282]. Cependant, ils pourraient parfois avoir un rôle pour dicter la spécificité des interactions miARN:ARNm, notamment pour les miARNs partageant une seed identique, qui sont regroupés en familles de miARNs.

Le fait que ces miARNs partagent la même seed suggère qu'ils partagent aussi les même cibles, mais quelques exceptions ont été observées, par exemple les miR-25-3p et miR-92-3p font partie de la même famille et sont tous les deux exprimés dans les cardiomyocytes, mais seulement le miR-25-3p est capable d'impacter la traduction du gène *SERCA2a* [275]. De plus, la sous-expression de certains miARNs peut entraîner une réponse phénotypique, malgré l'expression continue d'autres membres de la famille du miARN [22]. Dans ces cas, des appariements supplémentaires en 3' seraient nécessaires pour expliquer le manque de spécificité de la seed, et expliquerait également pourquoi les microARNs d'une même famille ne sont pas forcément interchangeables [34, 200].

On peut donc postuler que les appariements supplémentaires en 3' permettent des interactions plus robustes, et sont essentiels pour certains miARNs dans certaines conditions, mais ils ne sont généralement pas nécessaires.

b) Structure de l'interaction RISC:ARNm

La conformation du miARN au sein de la protéine AGO, obtenue grâce à des analyses structurales et biochimiques, explique l'importance de la seed du miARN pour reconnaître des ARNm cibles.

Tout d'abord, les deux nucléotides aux extrémités 5' et 3' du miARN sont protégés et restent fixés dans deux poches de la protéine, localisées dans les domaines MID et PAZ, respectivement [279] (Figure I.11). Le miARN est majoritairement protégé par AGO et seuls les nucléotides 2 à 6 du miARN sont exposés [45, 243]. Ce sous ensemble de la seed, et plus particulièrement les nucléotides 2 à 4, est utilisé dans un premier temps pour reconnaître un site sur un ARNm ciblé avec des nucléotides complémentaire [45].

Après initiation de l'interaction miARN:ARNm via le début de la seed, la conformation d'AGO est légèrement modifiée et permet la propagation de l'appariement aux nucléotides suivants de la seed tout en exposant les nucléotides 13 à 16 du microARN qui permettent des appariements supplémentaires éventuels en 3' [245] (Figure I.12). Le nucléotide de l'ARNm opposé à celui de l'extrémité 5' du miARN est également fixé dans une poche d'AGO, avec une préférence marquée pour une adénine [245] (Figure I.11.A). La structure finale de l'interaction est représentée en Figure I.11.B.

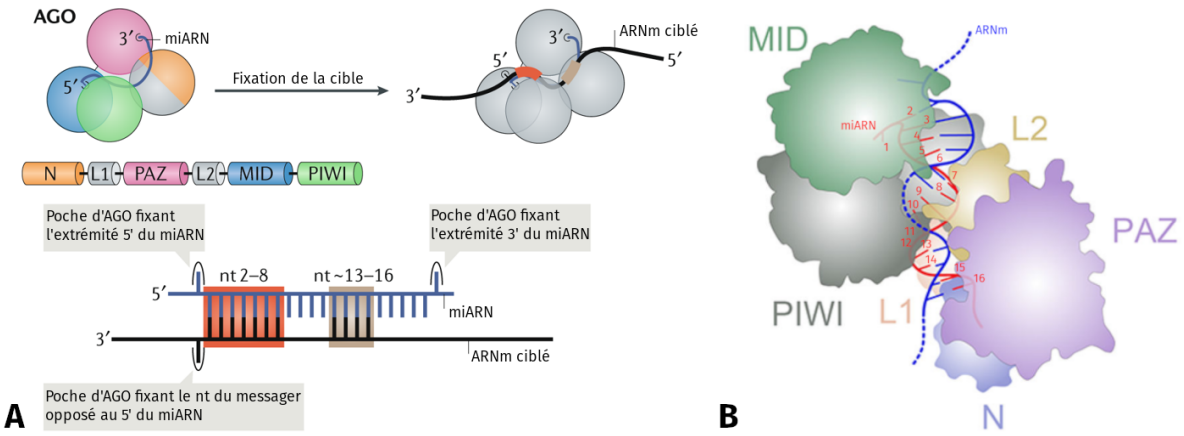


Fig. I.11 – (A) Interaction du complexe RISC par complémentarité de séquence du miARN avec un ARN messager. *Image adaptée de [85]* (B) Représentation détaillée de l'interaction RISC:ARNm. *Image adaptée de [245]*

c) Recherche de cibles par RISC

Pendant le parcours de RISC le long d'un ARNm [45, 149], des interactions faibles avec les nucléotides 2 à 4 de la sous-seed retiennent RISC, qui va tenter de propager la complémentarité de la seed. S'il échoue par manque de complémentarité, alors RISC se désengage rapidement et continue son parcours le long de l'ARNm. Si la propagation est réussie et que le nombre de nucléotides appariés de la seed est au moins égal à 7, alors l'interaction est robuste et RISC reste significativement plus longtemps apparié sur le site complémentaire [45]. On peut supposer que ce temps d'interaction est déterminant pour permettre au RISC de recruter des complexes protéiques nécessaires à l'inhibition de la traduction du messager, et éventuellement

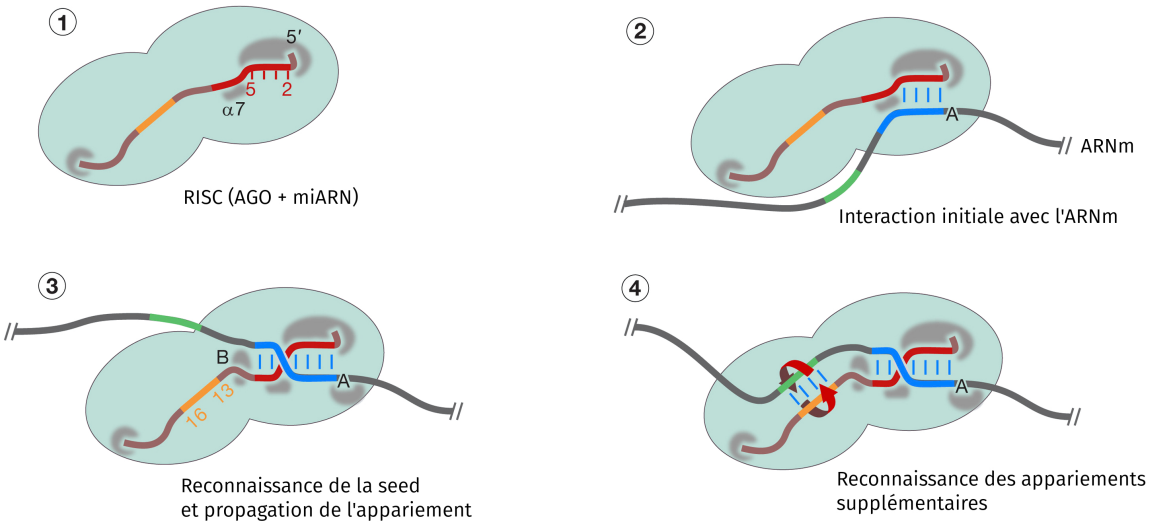


Fig. I.12 – Reconnaissance du messager cible par RISC via la seed du miARN, puis propagation de l'appariement de la seed, avec des appariements supplémentaires éventuels en 3' du miARN. *Image adaptée de [22].*

pour renforcer l'interaction en appariant des nucléotides supplémentaires en 3'.

2.2 Mécanismes de régulation de la traduction par les microARNs

Un miARN peut cibler de multiples ARN messagers, et à l'inverse, un ARN messager peut être ciblé par de multiples miARNs, et même posséder plusieurs sites de fixation pour un même miARN. Le groupe de Bartel a estimé qu'une large majorité des ARN messagers ont des sites de fixation conservés pour les miARNs [79]. Ces interactions permettent la mise en oeuvre de différents mécanismes de régulation de l'expression des gènes, observés principalement dans le cytoplasme.

a) Inhibition de la traduction et déstabilisation de l'ARN messager

Une fois l'appariement miARN:ARNm effectif et stabilisé, la protéine TNRC6 [21] est recrutée par le RISC et interagit avec les protéines PABPC qui sont associées à la queue poly-(A) du messager [127, 233]. Des complexes de déadénylases (CCR4-NOT ou PAN2-PAN3) sont ensuites recrutés par TNRC6, pour raccourcir la queue poly-(A), causant (dans les cellules post-embryonnaire uniquement) une déstabilisation du messager et le retrait de la coiffe 5' du messager par le complexe de décoiffage DCP1-DCP2 [47, 127]. Les messagers sans coiffe 5' ni queue poly-(A) en 3' sont ensuite rapidement dégradés, notamment depuis l'extrémité 5' par l'exoribonucléase XRN1 [17].

De plus, un second mécanisme, qui n'est pas encore bien compris, est mis en place pour inhiber la traduction avant ou pendant la déadénylation du messager. D'après les observations actuelles, il semblerait que l'initiation de la traduction soit perturbée grâce au recrutement par CCR4-NOT de DDX6 et du transporteur eIF4E (noté 4E-T). Il semblerait que la protéine 4E-T tente de se lier avec la protéine 4E à la place de 4G [129], perturbant ainsi l'initiation

21. TNRC6, également appelée GW182, possède 3 homologues chez l'homme (TNRC6A, B et C)

de la traduction, et inhibant la synthèse de protéines par les ribosomes. Quant à la protéine DDX6, il semblerait qu'elle pourrait se lier au complexe de décoiffage, ce qui entraînerait une inhibition de la traduction [52, 127]. Enfin, d'autres observations suggèrent que les facteurs d'initiation de la traduction 4A-1 et 4A-2 (eIF4A-1 et eIF4A-2) peuvent être dissociés de l'ARN messager lorsqu'une interaction miARN:ARNm est effective [127, 193, 81], et ce possiblement sans recrutement préalable de TNRC6 par le RISC [82].
Une représentation de ces mécanismes est synthétisée sur la Figure I.13.

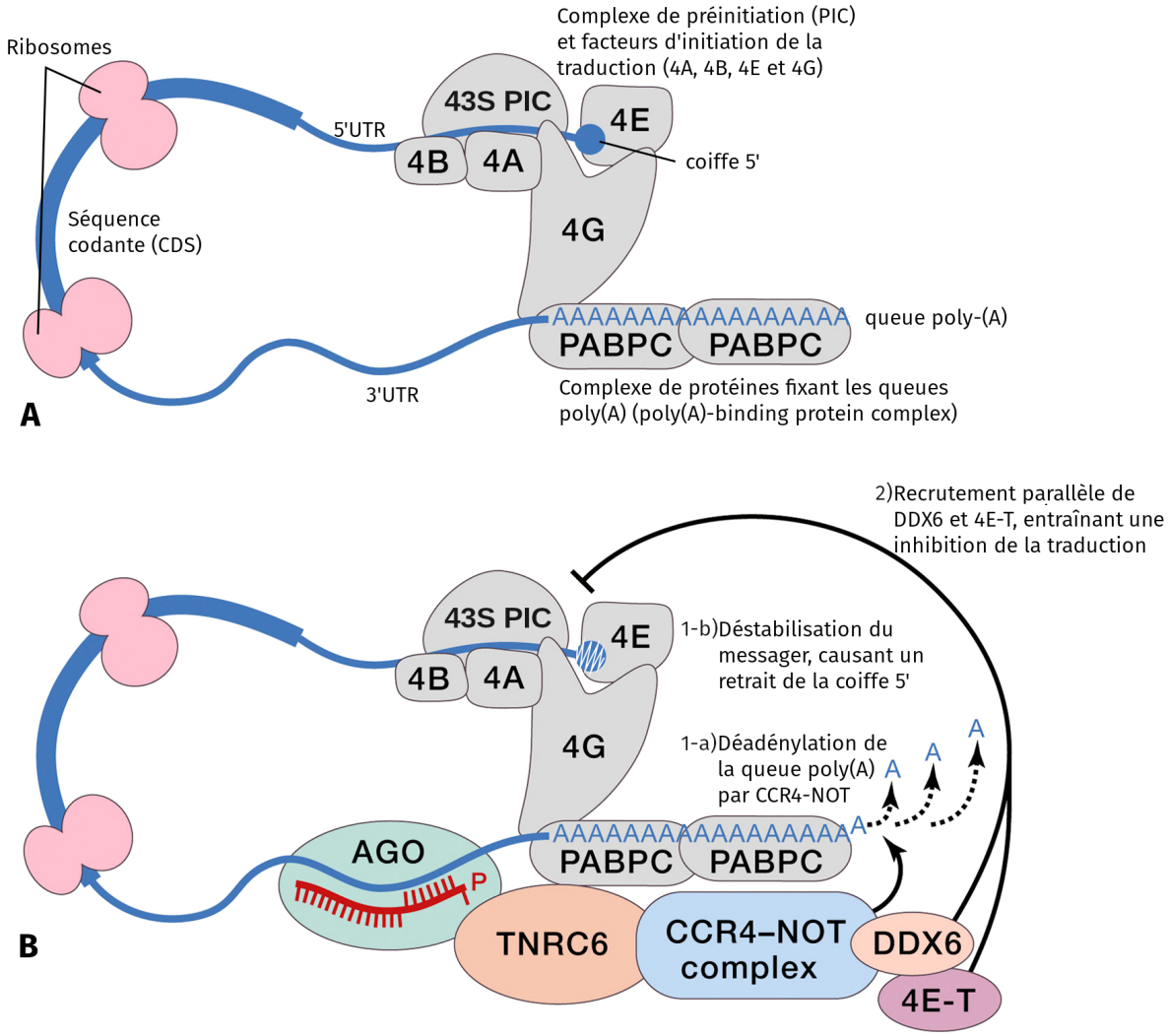


Fig. I.13 – A) ARNm dans sa conformation semi-circulaire, associé aux complexes protéiques nécessaires lors de sa traduction. **B)** Mécanismes de répression de la traduction par RISC, via le recrutement de TNRC6, qui recrute à son tour divers complexes pour déstabiliser le messager, afin d'entraîner sa dégradation (1-a et 1-b) et inhibe la traduction en parallèle (2). *Image adaptée de [22]*

L'interaction RISC:ARNm résulte le plus souvent (dans plus de 65% des cas) avec la dégradation du messager, quel que soit le miARN ou le type de cellule analysé [40, 68]. L'examination de la dynamique de cette répression, réalisée grâce à des relevés sur plusieurs

heures, montre qu'une diminution de l'activité de traduction est observée dans la première heure, avec une diminution des niveaux de protéines produites et une accumulation des ARN messagers, rapidement suivie par une dégradation importante du nombre de messagers [40]. Cette observation suggère que le mécanisme d'inhibition de la traduction est effectif avant ou pendant la mise en place du mécanisme de déstabilisation menant à la dégradation du messager.

b) Découpe de l'ARN messager par AGO2

Chez l'homme, 4 protéines AGO ont été identifiées. Cependant AGO2 est particulière car elle est la seule [22] qui peut catalyser elle-même la découpe d'un brin d'ARN par l'intermédiaire de son domaine PIWI [178, 194]. C'est en outre la seule AGO qui peut compléter la maturation du mir-451 en effectuant une découpe en 5' du bras 3p, comme mentionné précédemment [46].

Lorsqu'un miARN chargé dans AGO2 possède une complémentarité de séquence étendue entre la seed et les sites supplémentaires en 3' avec un ARNm cible, la protéine AGO2 peut alors catalyser la découpe du messager [178, 194]. La découpe du messager est alors effectuée entre les nucléotides appariés aux nucléotides 10 et 11 du miARN [24] et ces appariements 10-11, en plus de la seed et des supplémentaires en 3', sont suffisants pour catalyser la réaction. Ce mode de répression, fréquent chez les plantes et similaire au mécanisme de répression par les siARNs [23] [100], est relativement rare chez l'homme et n'a été observé que pour une vingtaine d'interactions RISC:ARNm [253].

c) Interactions non canoniques

Plusieurs études visant à établir un atlas des interactions miARN:ARNm [24] ont révélé que de nombreuses interactions sont observées avec une complémentarité de la seed limitée à quelques nucléotides, voire inexistante [34, 49, 94, 97, 108, 180]. Parmi ces études, plusieurs montrent également que des interactions sont fréquentes dans les régions codante (CDS) et 5'UTR du messager. Par la suite, les interactions avec une complémentarité imparfaite de la seed sont appelées non-canoniques de liaison, tandis que les interactions liées à une région différente d'un 3'UTR seront appelées non canonique de région.

Interactions non canoniques de liaison Concernant les interactions non canoniques de liaison, l'accumulation des observations suggère que le phénomène est relativement fréquent. Mais l'impact de ces interactions sur les niveaux de traduction est généralement très faible ou inexistante, notamment lorsque la seed ne partage aucun nucléotide complémentaire [1]. Comme le complexe RISC peut parcourir le 3'UTR et tester différentes interactions avec les premiers nucléotides de la seed, il est possible que ces appariements "tests" expliquent la majorité des interactions non canoniques de liaison observées.

Cependant, une méta-analyse rassemblant les données de ces études a permis de définir quelques interactions non canoniques fonctionnelles lorsque quelques mésappariements sont présents dans la seed. Cependant l'efficacité de ces interactions sur la répression de la traduction est réduite [141]. Une nouvelle étude récente montre, en modifiant successivement les nucléotides

22. Une étude récente indique qu'AGO3 possèderait également la possibilité de catalyser la découpe d'un ARNm, mais ne serait effective qu'avec un sous ensemble de miARNs [220].

23. Petits ARNs interférants, notés siRNA en anglais pour *small interfering RNA*.

24. En utilisant les méthodes HITS-CLIP et CLASH développées dans le but de séquencer les régions d'ARNs impliquées uniquement dans des interactions protéine:ARN, et adaptées spécifiquement pour les interactions RISC:ARN.

de la séquence d'un miARN ou d'un site de fixation connu du même miARN, que quelques mésappariements dans la seed peuvent être tolérés si une complémentarité supplémentaire est présente en 3' [24].

De façon générale, les interactions les plus stables, et donc celles qui ont le plus de chance d'être fonctionnelles, reposent majoritairement sur une seed parfaitement appariée. Cependant, il y a potentiellement une utilité fonctionnelle à des interactions non canoniques de liaison peu stables, via un modèle coopératif de plusieurs RISC interagissant avec le même ARNm sur une petite distance (moins de 35nt) [73, 261]. Si une interaction non canonique n'a pas de conséquence fonctionnelle à elle seule, sa coopération avec d'autres RISC à proximité permettrait de renforcer l'interaction avec le messager, et recruter plus facilement les complexes protéiques nécessaires à la déstabilisation de l'ARNm, notamment TNRC6 et CCR4-NOT.

Interactions non canoniques de région L'impact fonctionnel des interactions non canoniques de régions n'est pas encore clairement établi. Récemment, il a été observé que les appariements dans les régions CDS pourraient faire obstacle aux ribosomes et ainsi avoir un effet sur les niveaux de traduction du messager, sans provoquer sa dégradation [304]. D'après cette étude, ces interactions ont fréquemment des appariements supplémentaires étendus en 3', qui correspond au côté de la confrontation avec les ribosomes, mais il n'est pas clairement établi que ces appariements étendus en 3' puissent entraver l'action de traduction du ribosome. Lorsque ces appariements sont effectifs, une diminution de la production de protéines est observée, et ce même en l'absence de TNRC6, ce qui suggère que cette répression est mise en place par un mécanisme encore inconnu.

Toutefois, on ne peut pas exclure que certaines interactions dans les régions CDS ou 5'UTR soient capables de recruter les complexes protéiques nécessaires à la répression de la traduction, comme observé dans les régions 3'UTR.

d) Spécificités des protéines AGO

Les protéines AGO sont très abondantes et très stables. Elles peuvent atteindre 1.5×10^5 protéines par cellule [276] et possèdent une demie vie de plusieurs jours voire plusieurs semaines [58]. Toutefois, la stabilité d'AGO semble être assurée uniquement lorsqu'elle est chargée avec un miARN, car les protéines non chargées peuvent être dégradées [150].

Certaines modifications post-transcriptionnelles d'AGO permettent de réguler son activité [85]. Par exemple, AGO peut relâcher le messager lorsque certains acides aminés sont phosphorylés par CSNK1A1 [88], permettant au RISC de se rediriger vers une nouvelle cible.

Il y a 4 protéines AGO chez l'homme, et elles semblent s'associer sans préférence particulière de séquence avec les miARNs [11, 66, 258], à l'exception notable d'AGO2 avec le mir-451, qui est nécessaire pour compléter la maturation du miARN. Les quatre AGO peuvent charger des duplex de miARNs avec une préférence pour les duplex possédant un mismatch au niveau des nucléotides 8..11 [302]. Toutefois, certaines caractéristiques propres à chaque AGO pourraient avoir une incidence sur l'efficacité de la répression, et sur le spectre de cibles d'un miARN chargé. Par exemple, des motifs structuraux spécifiques à AGO1 et AGO3 peuvent interagir avec la séquence centrale des miARNs de façon différente d'AGO2 et AGO4 [219]. Cela suggère que les différentes AGO chargées avec le même miARN peuvent avoir des différences d'affinités avec une même cible.

2.3 Autres mécanismes associés aux microARNs

La répression de la traduction dans le cytoplasme est le mécanisme principal des miARNs, mais ils pourraient également être impliqués dans d'autres mécanismes. En particulier, des observations suggèrent que les miARNs pourraient être impliqués dans un mécanisme permettant d'améliorer l'activité de traduction de certains ARN messagers. Ils semblerait également qu'ils aient une fonction dans le noyau des cellules, en permettant de contrôler l'activation et la répression de la transcription. Les observations de ces mécanismes sont encore limitées, impliquent à l'heure actuelle un petit nombre de miARNs, et méritent de plus amples investigations.

a) Surexpression post-transcriptionnelle par les microARNs

À l'inverse du mécanisme habituellement observé, certaines observations suggèrent que des interactions RISC:ARNm auraient comme conséquence d'améliorer l'activité de traduction de certains ARN messagers dans des conditions cellulaires spécifiques. Par exemple, dans les cellules à l'état de repos (noté G0), où la traduction des messagers est affaiblie par la voie PI3K/mTOR, la traduction peut être réactivée par l'intermédiaire d'un RISC avec l'aide des protéines FXR1a, PARN et p97 [36, 273, 274]. Un autre exemple concerne les ARNm avec un motif 5'TOP [25], dont la traduction est limitée dans certains états cellulaires excepté lorsque le miARN miR-10a vient se fixer directement après le motif 5'TOP [307]. La présence du miARN empêcherait la liaison d'autres complexes inhibiteurs sur le motif 5'TOP, permettant ainsi au messager d'être traduit.

b) Activités nucléaires des miARNs

Alors que le mécanisme principalement étudié des miARNs a lieu dans le cytoplasme, de nombreux miARNs matures chargés dans un RISC sont également observés dans le noyau, et leur profil semble dépendre du type de cellules [177]. Le transport de RISC du cytoplasme vers le noyau est effectué par Importin8 [283], et pour certains miARNs tel que le miR-29b, un motif de 6 nucléotides spécifiques en 3' est nécessaire et suffisant pour être importé dans le noyau [123]. Les différents travaux sur le sujet rapportent principalement 2 mécanismes liés aux miARNs: la dégradation de transcrits, ou une régulation de la transcription par un mécanisme épigénétique.

Dégradation de transcrits nucléaires De façon similaire au mécanisme observé dans le cytoplasme, une dégradation des transcrits est possible dans le noyau, grâce à la présence d'autres complexes liés à la voie de répression cytoplasmique des ARN messagers, tels que TNRC6 et le complexe CCR4-NOT [83, 244, 246]. Par exemple, dans le noyau de cellules souches, il a été montré qu'un RISC peut cibler des ARNs par complémentarité de la seed dans les régions 3'UTR, mais aussi plus largement dans les régions codantes et les introns, et provoquer la dégradation d'un transcrit par la voie TNRC6 / CCR4-NOT [244].

miARNs et mécanismes épigénétiques De façon plus surprenante, les miARNs pourraient être impliqués dans un mécanisme épigénétique permettant la régulation transcriptionnelle de gènes ciblés. En effet, certains travaux suggèrent qu'un RISC pourrait interagir avec

25. À l'inverse de la plupart des ARN messagers avec coiffe 5', qui débutent avec une adénine, les ARN messagers 5'TOP débutent avec une cytosine suivie par un segment de 4 à 14 pyrimidines (C et U).

des régions ciblées du génome, par complémentarité de séquence avec le miARN chargé, et entraîner la modification du profil épigénétique localement, soit grâce à une modification des queues d'histones (principalement H3K4, H3K9 et H3K27), soit par une modification des niveaux de méthylation de l'ADN.

Ce mécanisme a été initialement observé dans le cadre de recherches utilisant des siARNs [101, 140], mais des observations suggèrent qu'un RISC chargé avec un miARN aurait un impact similaire. Plusieurs modèles ont été élaborés pour expliquer ces interactions, et proposent qu'un RISC pourrait interagir soit directement avec l'ADN au niveau des régions promotrices de la transcription, soit avec les transcrits naissants, synthétisés par l'ARN polymérase [177, 237, 284]. Une fois le RISC lié à un site, il peut recruter des complexes protéiques, et modifier les niveaux d'expression de transcrits localement, en modifiant le profil épigénétique. Ces interactions permettraient ainsi soit d'activer la transcription [185, 188, 229], soit de la désactiver [63, 139, 262, 303]. Cependant, les mécanismes et les complexes protéiques responsables de ces modifications ne sont pas encore clairement identifiés [174, 177].

L'activité et le rôle des miARNs nucléaires sont encore peu étudiés par rapport à l'activité cytoplasmique des miARNs, et méritent de plus amples investigations pour valider et mieux comprendre les mécanismes observés.

3 Régulation de l'expression et de la fonction des MicroARNs

Si les mécanismes responsables de la synthèse et de la fonction cytoplasmique des miARNs sont bien caractérisés, les mécanismes qui influencent leur stabilité et leur abondance au sein d'une cellule sont moins bien connus. La section qui suit donne un aperçu dynamique de l'activité et du cycle de renouvellement des miARNs. Ensuite, les mécanismes impliqués dans la dégradation, les modifications post-transcriptionnelles, et le contrôle transcriptionnel des miARNs seront présentés, afin d'avoir un aperçu de la régulation complexe de l'expression et de la fonction des miARNs.

3.1 Abondance et stabilité des miARNs

Des travaux récents sur des fibroblastes embryonnaires de souris ont permis d'élucider la dynamique temporelle de la production à la dégradation de 176 miARNs, donnant une idée des rythmes de production et de la stabilité des miARNs chez les mammifères [146]. Ils ont ainsi observé que quelques minutes suffisent pour compléter la maturation d'un miARN, et le miARN mature peut-être chargé dans un RISC et être effectif dans les 30 minutes suivantes [26]. Toujours d'après cette étude, les niveaux de production des miARNs sont généralement corrélés avec leur abondance, et peuvent être très élevés pour les miARNs les plus abondants, une cellule pouvant produire jusqu'à $\sim$ 100 copies d'un microARN mature chaque minute. Ce taux est bien plus important que ceux des ARN messagers les plus rapidement produits rapportés (moins de 10 molécules par minute par cellule [247]).

26. D'après les résultats d'une étude similaire chez la drosophile [234], mais avec un protocole différent, le temps de chargement est plus long et requiert en moyenne 1 heure, temps pendant lequel 40% des miARNs sont dégradés, ce qui implique une compétition entre les miARNs pour être chargés dans un RISC. Il y a donc soit une distinction entre les deux espèces, soit une mesure plus précise dans l'un des deux protocoles, et des analyses complémentaires sont nécessaires pour comprendre cette distinction.

La stabilité peut varier significativement selon l'identité du miARN, avec d'un côté des brins extrêmement stables possédant une demi vie pouvant atteindre une semaine [239], et à l'inverse des miARNs qui sont dégradés en quelques heures voire quelques minutes, en particulier dans les cellules neuronales [236]. En moyenne, les demi-vies observées des miARNs dans les fibroblastes murins sont supérieures à 24h, et sont ainsi environ 10 fois supérieures à celles des ARN messagers, inférieures à 3h. Ces durées de vies élevées permettent ainsi à un miARN d'avoir à tout moment une concentration supérieure à 1000 molécules par cellule, ce qui représente une abondance bien supérieure à celle typiquement observée pour les ARN messagers (environ 17 molécules par cellule) [62, 247]. La différence de stabilité entre différents miARNs semble être majoritairement dictée par leur séquence, qui est le seul élément permettant de les distinguer, mais aucun motif de séquence lié à la stabilité des miARNs n'a encore été clairement identifié.

3.2 Dégradation active des miARNs

Les mécanismes qui entraînent une dégradation rapide des miARNs sont peu connus. Les deux principaux mécanismes identifiés sont la dégradation par la protéine Tudor-SN (mécanisme appelé *TumiD* en anglais pour *Tudor-SN directed miRNA Degradation*), et la dégradation orchestrée par la cible (noté *TDMD* en anglais pour *Target-Directed miRNA Degradation*).

La protéine Tudor-SN a été récemment identifiée comme la première endoribonucléase responsable de la dégradation de certains miARNs [69]. Cette enzyme peut découper les miARNs lorsqu'ils sont libres ou lorsqu'ils sont chargés dans un RISC. La découpe est effectuée principalement au niveau des dinucléotides CA ou UA, lorsqu'ils sont localisés dans une région centrale du miARN (à au moins 5 nucléotides des extrémités). Les morceaux du miARN résultants de cette découpe sont ensuite dégradés par des exoribonucléases.

Le second mécanisme, TDMD, se déroule au niveau d'une interaction RISC:ARNm, et requiert une haute complémentarité de séquence entre le miARN et le messager. Lorsque les nucléotides supplémentaires en 3' sont suffisamment appariés avec la cible, l'extrémité 3' du miARN peut être relâchée de sa poche protectrice dans le domaine PAZ d'AGO [251]. Cette extrémité est alors exposée, et peut être modifiée par des TNTases et exoribonucléases, qui vont, respectivement, soit ajouter des nucléotides par un processus de *tailing* (et ajouter principalement des Adénines ou des Uraciles), soit en retirer par un processus appelé *trimming*. Éventuellement, le miARN est entièrement dégradé, mais les étapes menant à cette dégradation ne sont pas encore clairement établies. Il a été proposé que lors d'un TDMD, les miARNs sont d'abord sujets à un tailing, jusqu'à ce que l'extrémité 3' soit accessible à des exoribonucléases, qui vont ensuite effectuer un nombre suffisant de trimming pour éventuellement dissocier le miARN du RISC, et finalement dégrader entièrement le miARN [6].

L'exemple le mieux caractérisé de TDMD implique le miR-7-5p et sa cible le lncARN *Cyrano*, qui entraîne une réduction des niveaux de ce miARN de 97% [148]. Les observations actuelles de TDMD indiquent que le mécanisme est efficace principalement dans les cellules neuronales [59], ce qui est cohérent avec la tendance dynamique du système nerveux, nécéssitant des mécanismes actifs de régulation. Cependant, ce mécanisme est peut-être plus répandu, car une étude visant à découvrir des interactions résultant en TDMD dans des fibroblastes murins suggère que des centaines d'interactions peuvent résulter en TDMD, et décrivent en particulier une interaction entre le 3'UTR de *SERPINE1* et les miR-30b/c-5p qui résulte en TDMD dans ce tissu [87].

De plus, il a été rapporté qu'un miARN peut être dissocié du RISC lorsqu'il partage une haute complémentarité avec sa cible, laissant le miARN vulnérable à des ribonucléases [58]. On peut postuler que cette dissociation est causée par un TDMD, ou bien par une modification post-transcriptionnelle d'AGO (comme mentionné dans la section I.2.2.d).

Le nombre de miARNs affectés par ces mécanismes n'est pas connu, mais ils pourraient être principalement responsables de la régulation des miARNs avec une faible stabilité.

3.3 Séquestration des miARNs et ARNs compétiteurs

L'activité de miARNs spécifiques peut également être régulée sans dégradation, en les séquestrant par complémentarité de séquence, grâce à des transcrits non codants tels que les lncARNs ou les ARNs circulaires. Ces transcrits sont appelés "éponges", car ils possèdent généralement un nombre important de sites de fixation, leur permettant de séquestrer de nombreux miARNs. Par exemple, les ARNs circulaires *Syr* permet d'éponger le miR-138-5p, et le *CDR1as* peut éponger le mir-7-5p [103, 266]. Ces éponges permettraient ainsi de réguler les niveaux de miARNs spécifiques, et donc potentiellement d'augmenter l'expression d'ARN messagers normalement ciblés par ces miARNs séquestrés. Ce principe est souvent utilisé en laboratoire pour limiter l'expression de miARNs spécifiques en introduisant un nombre conséquent de séquences synthétisées complémentaires au miARN ciblé (nommés antagomiRs).

Ce principe de séquestration des miARNs par des transcrits pour permettre la sur-expression d'autres transcrits ciblés par ces miARNs, est à la base d'un modèle de régulation connu sous le nom d'hypothèse des ARN compétiteurs endogènes (notés ceARN) [242]. Selon cette hypothèse, un changement de l'expression d'un ARN ciblés par un miARN pourrait se répercuter sur les niveaux d'autres ARN messagers ciblés par le même miARN, et ainsi former un vaste réseau de régulation entre les ARN messagers par l'intermédiaire des miARNs.

Pour comprendre les conséquences de ce modèle, on peut considérer l'exemple de l'augmentation des niveaux d'expression d'un ARNm ciblé par un miARN, qui va alors augmenter le nombre de sites de fixations disponibles pour ce miARN, ce qui va entraîner une compétition de cet ARNm titrateur avec les autres ARN messagers ciblés par ce même miARN, et donc impacter la régulation et les niveaux d'expression de ces autres ARN messagers compétiteurs.

La validité de cette hypothèse dépend principalement de la quantité d'un miARN considéré, de la quantité de l'ARNm titrateur par rapport aux autres messagers ciblés par le miARN en compétition, et de la quantité de sites de fixations pour le miARN. Et malgré les hautes abondances de miARNs, les sites de fixations sont potentiellement bien plus nombreux, grâce à leur petite taille. En moyenne, pour les 90 miARNs les plus conservés, il existe plusieurs centaines de sites complémentaires chez l'homme [79], et bien plus si on considère également les sites non canoniques (détaillés dans la section I.3.2.b). Ainsi, il est peu probable que les sites de fixations soient saturés par des miARNs. Ce résultat a plusieurs implications [22]:

1) Une augmentation des niveaux d'un miARN, même légère, pourra avoir un effet sur les niveaux de répression des ARNs ciblés. En effet, si le nombre de sites de fixation est plus important que le nombre de miARNs disponibles, alors une augmentation du nombre de miARNs permettra de couvrir plus de sites de fixations et permettra donc d'accentuer la répression ;
2) Les miARNs faiblement exprimés ont probablement un effet négligeable sur la répression. Car devant le nombre important de sites de fixation possibles, un miARN doit être

suffisament exprimé pour cibler continuellement le même spectre d'ARN messagers et avoir un effet de régulation significatif ;

3) Enfin, en lien avec l'hypothèse ceARN, une augmentation des niveaux d'expression d'un ARN titrateur, et donc du nombre de sites de fixation, doit être importante pour avoir un effet sur la quantité disponible du miARN impliqué dans sa répression. Cependant, l'augmentation nécessaire de l'ARN titrateur pour déclencher cette compétition semble devoir être trop importante pour être observée naturellement *in vivo*. En effet, dans une étude visant à estimer l'augmentation nécessaire d'un ARN titrateur permettant d'observer une dérégulation des autres ARN messagers ciblés par un miARN [61], les auteurs ont observé que l'augmentation de l'ARN titrateur nécessaire était très élevée, et au mieux très rarement atteignable dans un contexte cellulaire normal.

3.4 Modifications de la séquence des microARNs : les isomiRs

Chaque miARN possède une séquence de référence, d'environ 22 nt, qui correspond au fragment d'ARN obtenu après transcription et maturation par le microprocesseur et Dicer. Ces séquences de références sont associées à chaque miARN dans les bases de données qui les référencent, telle que la miRBase. Cependant, l'application des technologies de séquençage haut débit sur les petits ARNs a permis de découvrir que la séquence des miARNs peut être variable, et que pour de nombreux miARNs, les séquences avec des variations sont plus abondantes que la séquence de référence [191]. Ces variations de séquence sont dues principalement à:

— des inconsistances lors du processus de maturation du miARN
— des modifications post-transcriptionnelles
— des variants génétiques

Les variations de séquences sont particulièrement fréquentes sur les extrémités du miARN, et correspondent à l'ajout de nucléotides, par *tailing*, ou au retrait de nucléotides, par *trimming* (ces mécanismes ont été mentionnés précédemment dans les sections I.1.3.d et I.3.2).

Les miARNs avec une variation de séquence par rapport à la séquence de référence sont ainsi des isoformes de miARNs, et sont donc appelés isomiRs [203], ou séquences non canoniques.

a) Imprécisions de Dicer et Drosha

Lors de la maturation du miARN, les RNAses Drosha et Dicer effectuent deux découpes successives qui déterminent les extrémités des miARNs. L'évaluation *in vitro* des sites de découpe de Drosha a permis de révéler des sites alternatifs pour de nombreux miARNs [137]. Les miARNs qui ont des sites alternatifs sont principalement ceux dont le pri-miARN n'a pas de motifs UGU dans la section apicale ou UG dans la section basale, dont la présence permet d'améliorer la précision de découpe du microprocesseur.

Le site de découpe de Dicer peut également être modifié. En particulier, selon la protéine partenaire de Dicer, TRBP ou PACT, la position du site de découpe peut changer d'un ou deux nucléotides [290].

La conséquence majeure de ces découpe alternatives est de décaler la seed du miARN, qui débute généralement au niveau du second nucléotide du miARN, et donc de modifier le spectre d'ARN messagers ciblés par le miARN. La découpe de Drosha définit la seed du bras 5p du pre-miARN, tandis que la découpe de Dicer définit celle du bras 3p (Figure I.14).

Pour certains miARNs paralogues, qui possèdent des séquences de références identiques, quelques variations nucléotidiques dans la séquence du précurseur, mais pas dans celle du

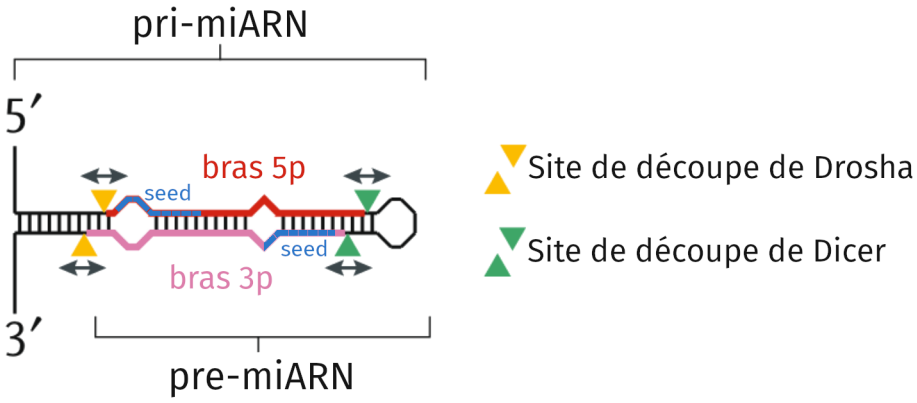


Fig. I.14 – Découpes alternatives de Drosha et Dicer. *Image adaptée de [85]*

miARN mature, peuvent également influencer les découpe de Dicer et Drosha. Ces différences génèrent ainsi de façon inattendue des miARNs paralogues avec des séquences différentes pouvant avoir des fonctions différentes [28].

Par rapport à une séquence de référence, ces découpes alternatives entraînent le trimming ou le tailing d'un ou plusieurs nucléotides aux extrémités. Dans le cas de tailing, les nucléotides ajoutés correspondent à ceux présents dans la séquence d'origine, au sein du pri-miARN. On parle alors de *templated tailing*, qui est différent du tailing provoqué par des TNTases, appelé *non-templated tailing*, ou plus fréquemment *non-templated addition* (noté NTA).

b) Modifications post-transcriptionnelles

Non-templated addition en 3' Le NTA est effectué par les protéines TNTases qui ajoutent principalement des Adénines (par la TNTase nommée GLD2) et des Thymines (par les TNTases nommées TUT1, TUT4, et TUT7) à l'extrémité 3' du miARN [295]. Ces additions sont observées à la fois sur les miARNs matures, et sur les pre-miARNs, notamment pour les pre-miARNs de la famille let-7 qui ont besoin de ce mécanisme pour compléter leurs maturations (voir section I.1.3.d). Les miARNs matures ont leur extrémité 3' généralement protégée par AGO, mais dans certaines conditions, comme lors d'un TDMD (voir section I.4.2), l'extrémité 3' est exposée et vulnérable à des modifications par les TNTases.

Cette modification, qui n'a pas d'impact sur la seed, permettrait selon certaines observations d'améliorer la stabilité de certains miARNs, en particulier le miR-122 pour lequel les NTA sont fréquents [131, 38]. Elles pourraient également améliorer l'affinité de certaines interactions miARN:ARNm [299] par extension des appariements en 3' du miARN, et ainsi améliorer l'efficacité de la répression de la traduction [252]. Elle permettrait également de promouvoir le TDMD, qui requiert une forte complémentarité en 3'.

Trimming par exoribonucléase Le trimming en 3' est la modification la plus fréquemment observée avec le NTA [191], mais c'est également la modification la moins bien caractérisée chez l'homme. Il semble que la majorité des évènement de trimming sont produits lorsque le miARN est chargé dans AGO et que l'extrémité 3' n'est plus protégée dans le domaine PAZ. En effet, en utilisant une AGO mutée qui ne possède plus de poche dans son domaine PAZ

pour protéger l'extrémité 3' du miARN, il a été récemment observé que les isomiRs avec du trimming étaient bien plus abondants qu'avec l'utilisation d'une AGO non mutée [251]. Or la seule condition connue où l'extrémité 3' est libérée de la poche PAZ d'AGO naturellement est lors d'un TDMD. Pour expliquer la haute fréquence des évènements de trimming, on peut donc postuler que le TDMD est très répandu. Ou alors, on peut émettre l'hypothèse qu'après une certaine limite de temps, l'extrémité 3' du miARN n'est plus fixée dans la poche d'AGO, ce qui la rend vulnérable aux exoribonucléases et TNTases. Cette dernière hypothèse est d'ailleurs consistante avec les observations récentes du groupe de Bartel, qui indique que l'abondance des isomiRs avec trimming ou tailing est corrélée avec l'âge des miARNs [146].

On peut également mentionner le mécanisme de trimming par exoribonucléases spécifique aux mirtrons, qui entraîne la dégradation des queues d'ARN qui sont associées en 3' ou 5' du mirtron [14, 215] (voir section I.1.3.a). Cette dégradation peut alors résulter avec des extrémités alternatives, avec du trimming si l'extrémité est plus courte que sur la séquence de référence, ou du templated tailing si elle est plus longue.

On peut également postuler que les extrémités du pre-miARN, après découpe par Drosha, peuvent être exposées aux exonucléases, comme elles le sont aux TNTases, mais aucun mécanisme n'a été identifié allant dans ce sens.

Edition d'ARN A>I par ADAR Des modifications post-transcriptionnelles internes à la séquence peuvent également être observées, à cause du mécanisme appelé "édition ARN". L'édition ARN la plus fréquente est la désamination d'une Adénine qui résulte en Inosines (I), notée "A>I". Cette modification est effectuée par la protéine Adenosine deaminase acting on RNA (ADAR) [133], qui cible les ARNs doubles brins tels que les épingles formées par les pri-miARNs ou les pre-miARNs. Ces modifications peuvent ainsi modifier la seed et donc réguler la fonction d'un miARN.

La fréquence d'édition A>I dépendrait du tissu, et ce mécanisme a été observé principalement dans les tissus cérébraux [158]. Comme l'Inosine se lie par complémentarité au C au lieu du T, cette modification peut créer des bourgeonnements, qui modifient la structure du pri- ou pre-miARN, et peut donc aussi gêner la découpe de Dicer ou Drosha et perturber le processus de maturation du miARN. Cette modification est souvent notée A>G à la place de A>I, car les Inosines se lient par complémentarité aux Cytosines, et lors de la préparation des librairies pour le séquençage, les Inosines sont remplacées par des Guanines, qui sont complémentaires aux Cytosines.

Un mécanisme plus rare d'édition "C>U" a également été observé dans des cellules T-regs, dans lesquelles ce mécanisme modifie la séquence du miR-100-5p [209]. Mais le mécanisme responsable de cette modification n'est pas encore identifié.

c) Variations génétiques

Des variants génétiques peuvent être localisés dans la séquence d'un miARN. Ils modifient donc la séquence du miARN, et potentiellement la seed, et peuvent ainsi changer le spectre d'ARN messagers ciblés par le miARN. Ces variants génétiques peuvent également influencer les niveaux de production d'un miARN, soit en perturbant le mécanisme de transcription, soit en déstabilisant la structure du pri- ou du pre-miARN, perturbant ainsi les mécanismes de maturation du miARN [42].

D'après une analyse entre les miARNs référencés dans la miRBase (version 18) et les variants indexés dans la base de donnée dbSNP (version 137), environ 24% des miARNs

matures possèdent au moins un SNP [27] dans leur séquence, et 9% des miARNs possèdent un SNP dans leur seed [102]. Toutefois, la majorité de ces variants sont rares, en particulier ceux qui sont localisés dans la seed des miARNs, qui est la région la mieux conservée des miARNs. Certains variants localisés dans des miARNs peuvent entraîner des phénotypes importants tels que la formation d'un kératocône impactant la cornée [118] ou une surdité [195].
Par la suite, on appellera *polymiR* un isomiR contenant un variant génétique.

d) Résumé

Un résumé des différents isomiRs et des mécanismes responsables de leur formation est représenté sur la Figure I.15.

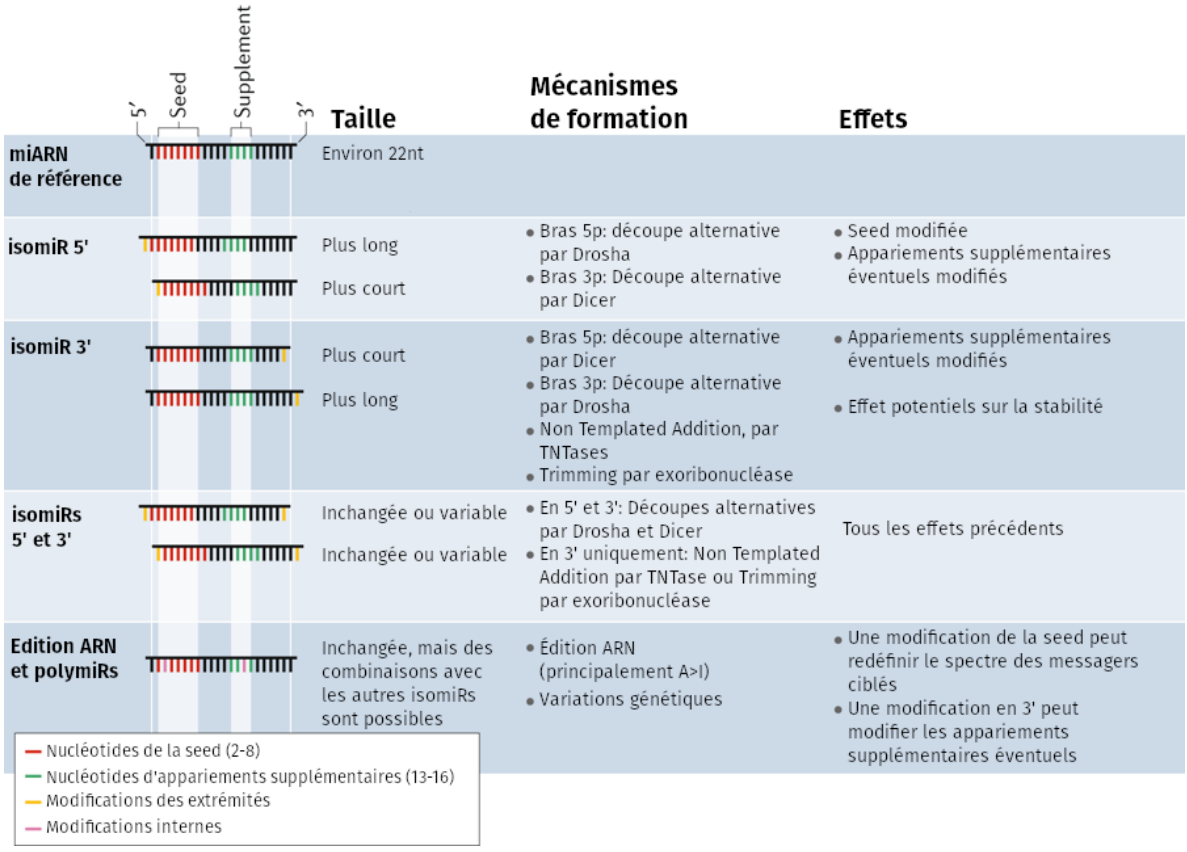


Fig. I.15 – Résumé des différents isomiRs, des conséquences qui leur sont associées, et des mécanismes responsables de leur formation. La taille de l'isomiR est également comparée à la séquence de référence

Les modifications en 5' du miARN peuvent être générées soit par une découpe alternative de Drosha ou Dicer, soit par un variant génétique, ou bien par une édition ARN. Ces modifications ont un impact plus important que celles en 3', car l'extrémité 5' contient la seed aux positions 2..8, qui peut être modifiée à cause de ces variations de séquences. Et une modification de la seed du miARN va impacter sa fonction en redéfinissant le spectre d'ARN messagers ciblés.

27. en anglais pour *Single Nucleotide Polymorphisme*, voir l'Appendice A pour plus d'information sur les variants génétiques

Cependant, les modifications en 3' du miARN sont beaucoup plus fréquentes, car cette extrémité est vulnérable aux mêmes mécanismes que l'extrémité 5', mais aussi au tailing par TNTases et au trimming par exoribonucléases. Les variations en 3' ont un impact modéré sur la fonction du miARN. Cependant des appariement supplémentaires en 3' sont parfois nécessaires, et des modifications sur cette extrémité pourraient ainsi avoir un effet sur certaines interactions.

La fonction des isomiRs n'est pas encore clairement établie, mais certains isomiRs peuvent avoir un effet sur la régulation par les miARNs dans un contexte particulier. Par exemple, le miR-223-3p possède deux isoformes qui régulent l'expression de cibles distinctes dans les neutrophiles murins [50]. De plus, les isomiRs sont potentiellement contrôlés génétiquement, car différents travaux rapportent que certains isomiRs peuvent être différentiellement exprimés selon le sexe, la population étudiée, ou la pathologie analysée [263, 181].

3.5 Régulation génétique et épigénétique

Les miARNs sont également assujettis à des mécanismes de régulation génétiques et épigénétiques [28], comme toutes les classes d'ARNs. Dans des cas particuliers, leur expression peut être totalement supprimée, par exemple à cause de mutations rares, comme la délétion d'un locus contenant un miARN. Plus fréquemment, leurs expression peut être modifiée, plus ou moins sévèrement, par des variations génétiques ou épigénétiques.

Afin d'identifier des variants génétiques pouvant moduler l'expression d'un miARN, on peut réaliser des études d'association entre le génotype d'un variant et les niveaux d'expression d'un miARN. Ces analyses reposent généralement sur un modèle linéaire additif, permettant d'associer le nombre d'allèles portés par un individu avec le niveau d'expression du miARN transcrit (comme sur l'exemple représenté sur la Figure I.16), en prenant en compte des variables d'ajustement souvent associées aux niveaux de miARNs telles que l'âge et le sexe [113, 240].

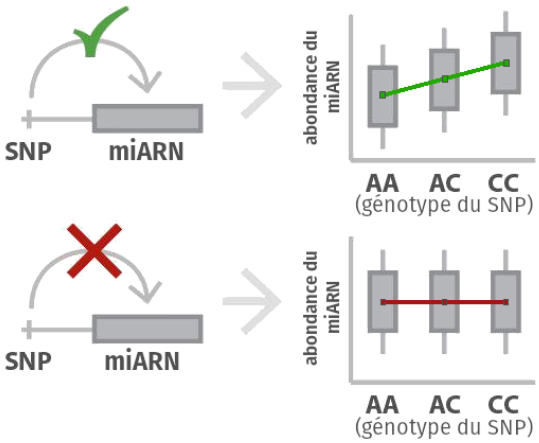


Fig. I.16 – Deux exemples de SNPs régulant (en haut) ou non (en bas) les niveaux d'expression d'un miARN. Les boxplots, à droite, représentent les niveaux d'expression du miARN selon le génotype du variant.

28. Les mécanismes de régulation génétique et épigénétiques sont présentés dans l'Appendice A.

Lorsqu'on mène ces analyses sur l'ensemble des variants du génome, elles sont appelées GWAS (en anglais pour *Genome Wide Association Study*) et les variants associés significativement à des variations d'expression sont appelés eQTLs (en anglais pour *expression Quantitative Trait Loci*). Des GWAS permettant d'identifier des eQTLs pour les miARNs ont ainsi été réalisées dans plusieurs tissus [30, 53, 84, 115, 116, 163, 222, 231, 254], permettant d'identifier collectivement des eQTLs pour une centaine de miARNs. La vaste majorité de ces eQTLs sont localisés à proximité du locus génomique du miARN, et sont appelés cis-eQTLs, en opposition aux trans-eQTLs, qui interagissent à distance. Ces cis-eQTLs sont généralement localisés dans des régions régulatrices (ex: ilôts CpG, promoteurs, enhancers, ou sites de fixation pour les facteurs de transcription), ou dans la séquence du miARN.

En particulier, les variants présents dans les ilôts CpG peuvent impacter localement la méthylation de l'ADN, l'un des principaux mécanismes épigénétique permettant le contrôle de la transcription. D'ailleurs, environ 10% des miARNs sont directement colocalisés avec des ilôts CpG pouvant être méthylés [201], ce qui les rends particulièrement susceptibles à un contrôle épigénétique de leur expression. Par exemple, un ilôt CpG méthylé à proximité du miARN polycistronique let-7a-1/f-1/d peut désactiver totalement son expression [151]. Récemment, une étude d'association épigénome-entier (EWAS) entre les niveaux de méthylation de sites CpG et des niveaux d'expression des miARNs dans le sang total a mis en évidence 40 miARNs dont l'expression est contrôlée par les niveaux de méthylation d'ilôts CpG [114], dont 11 sont localisés dans la région 14q32 du génome, déjà connue comme étant sous contrôle épigénétique [249]. Les modification post transcriptionnelles des queues d'histones (HPTM) sont également associées à un contrôle épigénétique de l'expression de nombreux miARNs, et la base de données EpimiR [57] a été créée afin de recenser les miARNs dont l'expression est contrôlée par la méthylation de l'ADN ou les HPTM.

3.6 Régulation par re-localisation en dehors du cytoplasme

L'activité principale des miARNs est observée dans le cytoplasme, elle peut donc être régulée en relocalisant les miARNs dans le noyau (déjà mentionné dans la section I.3.3.b), ou bien en les expulsant en dehors de la cellule. Les miARNs extracellulaires, appelés également miARNs circulants, sont présents dans de nombreux biofluides, tels que le sérum, le plasma, l'urine, le liquide cérébro-spinal, le liquide séminal, le lait maternel, les larmes, ou la salive [51, 281]. Ils sont transportés au sein de vésicules extracellulaires (EV, aussi appelés exosomes), de lipoprotéines, ou simplement chargés dans un RISC [8, 272] et sont ainsi protégés d'une dégradation par les ribonucléases.

Les mécanismes cellulaires qui entraînent l'évacuation de ces miARNs ne sont pas encore bien déterminés. Ils sont en partie expulsés passivement lors de l'apoptose, mais un mécanisme actif semble également évacuer les miARNs par l'intermédiaire d'EVs. Les hypothèses principales expliquants leur localisation extracellulaire sont [271]:

— un rejet résultant d'une régulation purement intracellulaire ;
— un rejet permettant une communication intercellulaire.

La régulation intracellulaire permettrait à une cellule d'évacuer des miARNs pour réguler leur abondance dans le cytoplasme. Tandis que la communication intercellulaire permettrait à un groupe de cellules d'envoyer des protéines et ARNs, dont des miARNs, à d'autres cellules situées soit à proximité, soit joignables par circulation dans le biofluide. Un exemple d'une telle communication a récemment été observé, avec un miARN exprimé par le tissu adipeux d'une souris, évacué par EV, puis intégré par des cellules hépatiques pour réguler un ARNm exprimé dans le foie [265]. Cependant, les quantité de miARNs observées dans les EVs sont

très limitées, et semblent être insuffisantes pour impacter les niveaux de traduction dans une cellule réceptrice. La part majoritaire de miARNs circulants correspond à ceux qui sont simplement chargés dans un RISC [8, 272]. Des travaux supplémentaires sont nécessaires pour comprendre les mécanismes mis en oeuvres lors de l'évacuation des miARNs et lors de la récéption des miARNs circulants. Cette piste est un nouvel axe de recherche actif [225] et ouvre des perspectives intéressantes sur les possibilités de régulation intercellulaire par les miARNs.

4 Intérêt de l'étude des microARNs en recherche biomédicale

Dans les parties précédentes, les mécanismes responsables de la synthèse, de la fonction et de la régulation des miARNs ont été présentés. Les miARNs y sont décrits comme des molécules remarquablement stables et abondantes (section I.4.1), dont le rôle est de réguler la majeure partie du transcriptome (section I.3.2). Ce sont donc des acteurs omniprésents qui sont actifs dans de nombreuses voies biologiques. Ils sont aussi eux-mêmes vulnérables à de nombreux mécanismes pouvant réguler leurs expressions et leurs fonctions. Dans certains contextes particuliers, génétiques ou environnementaux, une dérégulation importante de l'expression ou de la fonction des miARNs peut avoir lieu et entraîner le développement de phénotypes anormaux. Ils sont ainsi potentiellement impliqués dans diverses maladies complexes ou rares, et sont donc le sujet de nombreux travaux de recherches médicales. Cette dernière partie décrit le rôle et l'ampleur de l'impact des miARNs dans la physiologie cellulaire, avant de présenter les différents axes de recherches médicales sur les miARNs.

4.1 Rôle physiologique des microARNs

Les miARNs sont considérés comme les sculpteurs du transcriptome [22], qui raffinent les niveaux de production des protéines. Le niveau d'expression d'un gène donné dépend de l'abondance de l'ARNm correspondant, mais aussi de l'abondance et de la stabilité des miARNs qui le ciblent. Ces paramètres varient significativement selon le tissu et le type cellulaire dans lequel l'ARNm et le miARN sont exprimés [98], et de l'état du cycle cellulaire ou du développement de l'organisme [210]. L'activité des miARNs a donc pour conséquence d'obtenir une régulation spécifique à un contexte cellulaire, que l'on peut supposer correspondre à un niveau optimal de production de protéines pour le bon fonctionnement cellulaire.

Pour que la régulation d'un gène par un miARN soit effective, il faut évidemment que le miARN et sa cible soient exprimés dans la même cellule au même moment, mais surtout que le miARN soit suffisamment exprimé pour avoir un effet de régulation significatif sur l'expression de sa cible. Car certains miARNs ne sont pas exprimés dans certains contextes cellulaires, ou alors trop peu exprimés pour avoir un effet sur l'expression des ARN messagers ciblés. En effet, à cause du nombre important de sites de fixation disponibles pour un miARN donné, il est estimé qu'environ 1000 copies du miARN sont nécessaires pour avoir un effet [146]. Par exemple, lorsqu'un certain seuil de production d'un miARN est atteint, il peut déclencher une transition dans le développement cellulaire et biologique, comme découvert par les équipes de Ruvkun et Ambros chez le nématode c. elegans, où la régulation de *lin-14* par *lin-4* permet au ver de passer d'un stade de développement au suivant. Ainsi, seul un sous-ensemble de miARNs est fonctionnel dans un contexte cellulaire spécifique.

Lorsque les miARNs exprimés sont suffisament abondants de manière continue, ils peuvent alors assurer la stabilité nécessaire au fonctionnement cellulaire. Concernant les quelques miARNs avec une stabilité faible ou variable, ils ont vraisemblablement un rôle plus dynamique, permettant une réponse à un contexte cellulaire ou environnemental spécifique. Par exemple, lorsque des cellules neuronales de la rétine sont exposées à la lumière, l'abondance de certains miARNs naturellement peu stables est réduite en moins de 90 minutes, grâce à une diminution de leur production [153]. Dans certaines conditions spécifiques, comme lors d'une infection virale [43, 169, 186] ou pendant certaines périodes du cycle cellulaire [236], une dégradation rapide de certains miARNs peut être observée, qui résulte potentiellement d'une régulation active de leur expression ou de leur fonction.

4.2 Ampleur de l'impact des microARNs sur l'expression des gènes

Afin de déterminer le rôle et l'impact d'un miARN spécifique sur le transcriptome et le protéome [29], et plus largement sur l'organisme, de nombreuses expérimentations de surexpression ou de suppression de l'expression par *knock-out* (KO en anglais pour délétion) de miARNs individuels ont été réalisées dans des lignées cellulaires humaines et dans des modèles murins [22].

Il est ainsi possible de déterminer les transcrits ciblés par un miARN, en comparant les niveaux d'expression des transcrits entre la condition où le miARN est exprimé normalement, et la condition où il est exprimé de façon anormale (surexprimé ou KO), et en même temps mesurer l'importance de la répression effectuée par un miARN sur ces différents transcrits. Il est alors également possible d'observer les phénotypes développés dans la condition où le miARN est exprimé de façon anormale. De plus, pour les miARNs qui ont une fonction redondante, notamment lorsqu'ils sont membres d'une même famille, il faut parfois modifier l'expression de l'ensemble des membres de la famille pour voir apparaître un phénotype important [30].

En comparant les niveaux d'expression des transcrits et des protéines entre les deux conditions (avec miARN et condition KO), on observe de façon surprenante que l'impact de la régulation d'un miARN est généralement modeste. En effet, le changement de l'expression des transcrits et des protéines peut parfois atteindre 50% mais il est généralement inférieur à 20% [16, 250]. On peut toutefois noter que la coopération de plusieurs miARNs différents, fixés sur plusieurs sites proches sur un ARNm, peuvent parfois accentuer l'efficacité de la répression (comme mentionné précédemment dans la section I.3.2.b) [73, 261], mais cette condition n'a pas été prise en compte par ces expérimentations.

Cependant, malgré l'impact modeste sur les niveaux de production de protéines, des phénotypes souvent sévères sont observés presque systématiquement lors d'un KO d'un miARN conservé dans un modèle murin: 69 des 90 familles de miARNs conservées depuis le poisson sont associées à un phénotype anormal lors d'un KO de la famille du miARN, et parmi les 21 familles restantes, plusieurs n'ont pas encore été testées [22]. Parmi les phénotypes observés, des développements anormaux sont fréquents (au niveau du squelette, cerveau, muscles, coeur, poumons, reins, foie, lignées hématopoïétiques, etc), et accompagnés de dérèglements cellulaires (au niveau de la formation des synapses, la polyploïdisation, la genèse des cils, etc), ou physiologiques (en dérégulant la fonction cardiaque, pression sanguine, métabolisme des lipides

29. Le protéome désigne l'ensemble des protéines qui résultent de la traduction des gènes
30. En particulier pour les expérimentations de type KO, où il faut parfois supprimer l'expression de tous les membres d'une famille de miARNs.

et du cholestérol, mobilisation des glycogènes, fibrose, développement embryonnaire, etc) et impactent régulièrement la viabilité de l'embryon ou de la souris.

Il est ainsi apparent que les miARNs conservés ont un rôle indispensable pour obtenir des conditions de développement saines. Mais l'impact des miARNs moins conservés, qui ont émergés récemment chez l'homme, est plus difficile à étudier, faute d'avoir un modèle animal dans lequel mener des expérimentations.

Des KO hétérozygotes, permettant de désactiver un seul des deux allèles exprimant un miARN, ont également été effectués, permettant de réduire de moitié l'expression d'un miARN. Dans ces expérimentations, il n'y pas généralement pas de réponse phénotypique [22], révélant ainsi que la majorité des ARN messagers ciblés sont insensibles à un changement majeur de l'expression d'un miARN, à l'exception notable du mir-96 et du polycistronique mir-17∼92, qui engendrent des haploinsuffisances causant une surdité et une anomalie du squelette, respectivement [195, 60]. Cela suggère que les gènes haploinsuffisants ciblés par un miARN sont davantage vulnérables à un changement de l'expression du miARN [228].

4.3 Recherche biomédicale

Depuis la premier cas d'implication d'un miARN dans le développement d'une maladie humaine en 2002 [41], où une mutation responsable de la délétion du cluster de miARN mir-15∼mir-16 a été associée au développement d'une leucémie, de nombreuses études ont été lancées pour déterminer l'implication des miARNs dans diverses pathologies. Le développement de phénotypes anormaux est possible lorsque l'expression d'un miARN est altérée, ou lorsqu'une interaction miARN:ARNm est perturbée. Ces dérégulations dépendent alors d'un contexte particulier, génétique ou environnemental. De plus, s'ils ne sont pas directement impliqués dans le développement d'un phénotype anormal, certains miARNs peuvent avoir leur expression associée au développement de la pathologie, et être ainsi utilisables comme biomarqueurs.

a) Recherche de microARNs impliqués dans le développement d'une maladie

Recherche de miARNs dérégulés Afin d'identifier des miARNs impliqués dans le développement d'une maladie, on commence généralement par effectuer une comparaison entre le miRnome d'individus malades et celui d'individus sains. Le miRnome doit être mesuré dans un tissu adapté à l'étude du phénotype délétère, ou à partir de biofluides facilement accessibles lorsque le tissu d'intérêt est indéterminé, ou si il requiert une procédure d'extraction trop invasive. Cette approche permet d'identifier les miARNs différentiellements exprimés entre les deux groupes, qui sont ainsi associés au phénotype délétère, et donc potentiellement responsables du développement de la maladie.

La base de donnée HDMM (*human miRNA disease database* en anglais) recense les associations identifiées entre des miARNs et des maladies [117], et référence dans sa version actuelle (v3.2) 35,547 associations impliquants 1206 miARNs et 893 pathologies. Cependant, certaines études référencées reposent sur des cohortes de petites tailles, avec une faible puissance statistique, et l'implication de miARNs dans le développement de certains phénotypes délétères est peut être parfois surestimée. Ainsi, des réplications de ces études d'association sont nécessaires pour valider les résultats observés, et réduire l'ensemble de miARNs suspects à un sous-ensemble robuste et consistant entre les études.

Dans le cas de maladies rares, il est possible que la dérégulation d'un miARN à elle seule soit responsable du développement du phénotype, auquel cas cette approche est particulièrement

efficace. Cependant, dans le cas de maladies complexes répandues, plusieurs facteurs collaborent dans le développement du phénotype délétère. La dérégulation d'un miARN peut parfois être l'un de ces facteurs, même si la dérégulation est modeste et n'entraîne pas à elle seule le développement d'un phénotype délétère. Ainsi l'identification de ces miARNs modestement dérégulés n'est rendue possible qu'en intégrant un nombre élevé de patients dans l'étude, afin d'obtenir une puissance statistique permettant de détecter une association significative avec le développement de la maladie.

La dérégulation du miARN suspecté peut également être investiguée. En particulier, le contexte génétique et épigénétique peut influencer de façon importante les niveaux d'expression d'un miARN, et des études de type GWAS ou EWAS (mentionnées dans la partie I.4.5) permettent d'établir les déterminants génétiques ou épigénétiques responsables d'une dérégulation du miARN, et donc potentiellement responsables du développement de la maladie.

Identification des interactions miARN:ARNm Afin de comprendre le phénotype engendré par la dérégulation d'un miARN, il faut aussi identifier le ou les gènes dont l'expression est impactée par la dérégulation du miARN, et donc *in fine* identifier l'interaction miARN:ARNm qui est dérégulée.

Afin de déterminer ces interactions fonctionnelles entre un miARN et un ARNm, plusieurs approches existent. D'abord les approches directes à grande échelle, basée sur des techniques d'immunoprécipitation visant les protéines AGO, permettent d'établir un atlas des interactions miARN:ARNm. Mais comme cela a été mentionné dans la partie I.2.2.c cette technique manque de fiabilité et ne permet pas d'établir définitivement les interactions fonctionnelles, c'est à dire celles qui ont un impact sur l'expression du gène ciblé par le miARN.

Des approches indirectes peuvent aussi être employées. On les appelle indirectes car elles ne permettent pas d'établir si il y a une interaction directe entre le miARN et l'ARNm. Elles reposent sur des analyses de l'expression du transcriptome et du protéome. On peut par exemple effectuer des analyses de corrélations entre l'expression de miARNs et l'expression de gènes. Une corrélation négative entre un miARN et un gène peut suggérer une régulation du gène par le miARN. On peut également faire une analyse plus précise, en désactivant l'expression d'un miARN (par exemple en utilisant un antagomiR ou par KO CRISPR). En comparant l'expression des transcrits et protéines lorsque le miARN est exprimé ou désactivé, on peut alors établir quels sont les gènes potentiellement régulés par le miARN. Cependant, les approches indirectes ne permettent pas d'affirmer que le miARN est directement responsable de la régulation, ou si un intermédiaire est impliqué pour modifier les niveaux d'expression du gène.

Parmi les approches directes à petite échelle, les expérimentation basées sur les gènes rapporteur comme la luciférase permettent d'établir l'impact de la régulation d'un miARN sur un gène particulier. Le principe repose sur la combinaison de la séquence 3'UTR d'un gène d'intérêt et la partie codante d'un gène rapporteur tel que la luciférase, une protéine bioluminescente. On peut alors mesurer l'impact d'un miARN sur l'expression d'un gène dont il cible le 3'UTR, en comparant le niveau de fluorescence émise lorsque le miARN est exprimé ou désactivé.

Enfin, une méthode permettant de déterminer de façon précise une interaction est de perturber un site de fixation suspecté (par exemple en utilisant une méthode de type CRISPR), et d'observer la réponse phénotypique de cette perturbation. Grâce à cette méthode, plusieurs études ont permis d'associer le développement de phénotypes sévères avec la perturbation d'un site de fixation miARN:ARNm unique [65, 67, 192]. Cependant, cette méthode ne prend

pas en compte une coopération potentielle de plusieurs miARNs, qui peuvent accentuer la répression collaborativement, mais qui n'ont pas d'effets observables lorsqu'il sont seuls.

Le choix de la stratégie dépend donc des informations dont on dispose. En particulier, si des miARNs et des gènes spécifiques sont suspectés, les méthodes directes à petite échelle donneront des résultats plus précis. Une fois l'interaction clairement établie, il faudra recourir à des expérimentation plus avancées, telles qu'un KO du miARN, par exemple dans un modèle animal compatible, pour découvrir son implication dans un phénotype délétère. Il faut également noter que les conditions de laboratoire ne permettent pas toujours d'obtenir un effet observable. Par exemple, certains phénotypes sont observés dans des contextes cellulaires spécifiques (tissu, type de cellules, cycle cellulaire, stress, prédispositions génétiques, etc).

b) Utilisation de microARNs comme biomarqueurs

Les biomarqueurs sont d'origine moléculaire, histologique, physiologique, ou radiographique, et peuvent servir de témoins d'une activité biologique. Ils peuvent être utilisés notamment pour établir un diagnostic, prédire le développement d'une pathologie, ou évaluer la réponse à un traitement. Des biomarqueurs fiables pourraient ainsi faciliter la prise en charge de patients, en termes de temps, de fiabilité du diagnostic, de traitement et donc dans l'absolu en termes de coûts, grâce à une simple mesure. La stabilité et la facilité d'extraction des miARNs circulants à partir de biofluides en font des candidats idéaux dans la recherche de biomarqueurs pour diverses pathologies [198].

Ce domaine de recherche est très actif, notamment dans la recherche sur le cancer. La découverte d'un profil d'expression de miARNs unique à différents types de cancers permettrait d'obtenir des signatures distinctives pouvant faciliter leurs diagnostics [198]. Les données accumulées ces dernières années sont prometteuses, par exemple pour discriminer les patients atteints d'une forme agressive du cancer de la prostate, qui requiert un traitement immédiat, des patients atteints du forme moins aggressive. Une équipe a ainsi identifié 14 miARNs présents uniquement dans le sérum des patients atteints de la forme peu aggressive [197], et l'utilisation clinique de cette signature de miARNs permettrait ainsi d'éviter un traitement lourd à des patients qui n'en ont pas besoin.

Dans le domaine des maladies cardiovasculaires, les biomarqueurs constituent également un outil essentiel pour les cliniciens. Ils sont notamment utilisés pour le diagnostic de l'infarctus du myocarde (par mesure des troponines cardiaques), de l'hypertension (par mesure de la pression sanguine), ou pour prédire le risque de rejet d'une transplantation cardiaque (avec Allomap, un test basé sur l'expression génique) [99]. Cependant, beaucoup de biomarqueurs manquent de fiabilité, et beaucoup de pathologies cardiovasculaires pourraient être mieux diagnostiquées et traitées avec la découverte de nouveaux biomarqueurs robustes, tels que les miARNs.

Cependant, plusieurs challenges doivent être surmontés pour identifier une signature de miARNs robuste comme biomarqueurs. D'abord il faut prendre en compte les facteurs techniques, tels que la méthode de prélèvement et leur conservation, ou les méthodes de profilage, qui peuvent influencer la variabilité des miARNs. Il est possible de réduire cette variabilité au sein d'une étude, si le protocole est strictement identique d'un patient à l'autre. Ainsi, si un biais d'expression est introduit à un certain point, il sera identique pour tous les échantillons étudiés, permettant ainsi de les comparer entre eux. Ensuite, il faut considérer la variabilité introduite par des facteurs biologiques, comme l'âge ou le sexe, ou le contexte génétique. Afin de trouver la variabilité associée uniquement au phénotype, ces paramètres biologiques doivent être pris en compte lors des tests d'association. Enfin, les biomarqueurs

identifiés doivent suivre un strict processus de validation, avec notamment une réplication dans des cohortes indépendantes. Afin d'être considéré dans un cadre clinique, il faut également déterminer un protocole permettant d'établir le risque associé aux biomarqueurs avec une analyse quantitative absolue. Par exemple, grâce à une mesure par PCR (technique présentée dans le chapitre suivant) en ayant identifié au préalable les seuils critiques associés à un risque.

Chapitre II : Mesure des microARN

1 Méthodes de quantification

Plusieurs méthodes permettent de quantifier les miARNs exprimés dans un échantillon, les principales étant les puces d'expression (*microarray*), la qRT-PCR (pour *quantitative Reverse Transcriptase Polymerase Chain Reaction* en anglais), et le séquençage haut débit (NGS) des petits ARNs.

1.1 qRT-PCR

La qRT-PCR, aussi appelée PCR en temps réel, permet de quantifier une séquence particulière d'ARN ou d'ADN présente dans un échantillon. La mesure est effectuée grâce à un marqueur fluorescent, et elle effectuée en continu sur plusieurs cycles d'amplification PCR. La cinétique complète de la réaction de polymérisation permet d'obtenir une quantification absolue de la quantité initiale de la séquence d'ARN ou ADN ciblée.

Dans le cas d'une mesure d'ARN, la séquence ciblée est au préalable convertie en ADN complémentaire (noté ADNc). Le principe de cette technique repose sur des amorces (*primers* en anglais) qui peuvent venir se fixer à la séquence ciblée par complémentarité de séquence, ainsi que des protéines Reverse-Transcriptases qui permettent une transcription inverse de la séquence en ADNc, en partant du primer. Ensuite, pour chaque cycle PCR, des primers spécifiques à la séquence ciblée ainsi que des ADN polymérase sont introduites. Une fois les primers fixés sur la séquence ciblée, l'ADN polymerase synthétise le brin complémentaire à la séquence. Des sondes fluorescentes pouvant se fixer sur les ADN double brin sont également introduites, et émettent une fluorescence une fois fixées. Cette technique, répétée sur plusieurs cycles, permet de mesurer en temps réel l'accumulation de la séquence ciblée. Chaque cycle permet de doubler la quantité de la séquence présente dans l'échantillon, et le nombre de cycles d'amplification nécessaires pour obtenir un signal significatif reflète la quantité initiale de la séquence dans l'échantillon. Un faible nombre de cycles nécessaires reflète donc une quantité initialement élevée de la séquence, tandis qu'un nombre élevé de cycles nécessaires reflète une séquence initialement peu abondante.

Cette technique est parallélisable, avec l'utilisation de plaques contenant plusieurs puits, chaque puit contenant des primers complémentaires à une séquence spécifique.

1.2 Microarrays

Les microarray ont été développés pour mesurer l'expression d'un grand nombre de séquences en parallèle. Le principe repose sur l'hybridation des fragments d'ADN ou d'ARN contenus dans un échantillon à des sondes. Comme pour la PCR, si les fragments ciblés sont des fragments d'ARN, ils sont au préalable convertis en ADNc. Chaque sonde possède une séquence complémentaire à un fragment ciblé, et une fluorescence est émise lors de l'hybridation de la

séquence à la sonde. Chaque sonde est disposée à un endroit pré-déterminé et fixe sur la puce, ce qui permet de déterminer la quantité d'un fragment présent dans l'échantillon grâce à la localisation et à l'intensité de la fluorescence sur la puce.

1.3 Séquençage haut-débit

Le séquençage haut-débit (ou NGS) est la dernière méthode en date développée pour quantifier les miARNs. Cette méthode permet d'obtenir directement la séquence des ARNs présents dans un échantillon. Une pré-sélection sur la taille des fragments séquencés permet de concentrer l'effort de séquençage sur les fragments ayant la taille typique des miARNs.

Le NGS est la méthode offrant donc le maximum d'information, et permet notamment d'analyser la séquence des miARNs, de quantifier des miARN non référencés, et de quantifier les isomiRs. Ces analyses ne sont pas possibles avec les méthodes qRT-PCR et microarray, qui reposent sur des techniques d'hybridation, et ne permettent donc que de mesurer des séquences fixées à l'avance (et proposent généralement les séquences de références indexées par la base de données miRBase). Cependant, une analyse bioinformatique conséquente est nécessaire pour trier la quantité massive d'information obtenue par NGS.

1.4 Choix de la méthode

Chaque méthode possède des avantages uniques, et le choix de la méthode va dépendre des objectifs de l'étude. Si aucun miARN particulier n'est visé par l'étude, par exemple dans le cadre d'une phase initiale de découverte, une approche globale comme les microarray ou le séquençage NGS permettant une mesure de nombreux miARNs sera requise. Mais si un ou quelques miARNs spécifiques sont visés par l'étude, la qRT-PCR est plus appropriée. Pour les études à grande échelle échelle, avec de nombreux échantillons, les microarray ont l'avantage d'un prix très abordable. Mais la baisse du prix du séquençage NGS, ainsi que sa haute sensitivité et les possibilités supplémentaires qu'il offre, notamment l'analyse de miARN non référencés, d'isomiRs, et d'autres fragments d'ARN, en font une méthode beaucoup plus puissante. Dans le cadre de cette thèse, c'est la méthode NGS qui a été choisie, dont le protocole détaillé est présenté dans la partie suivante.

2 Protocole détaillé du séquençage de petits ARNs

Le séquençage NGS requiert la préparation d'une librairie, contenant uniquement les ARNs que l'on souhaite séquencer. Dans le cadre de ce projet, un protocole expérimental a donc été suivi pour obtenir des librairies de petits ARNs, avec une taille proche des miARNs, à partir d'échantillons de plasma. Les étapes principales de ce protocole sont:

— la purification des ARNs
— la ligation d'adaptateurs
— la sélection de la taille des ARNs

Ce protocole a été effectué par Mme Claire Perret (UMRS 1166, Paris) en suivant les indications fournies par les fabricants des kits de purification et de préparation utilisés. Le séquençage en lui-même a été réalisé par la plateforme de séquençage de l'Institut du Cerveau et de la Moelle épinière (ICM, Hôpital de la Pitié Salpêtrière, Paris).

2.1 Purification et préparation des librairies

Purification des échantillons La première étape du protocole consiste à purifier les échantillons, c'est à dire d'isoler les ARNs des autres éléments contenus dans le plasma. La purification a été réalisée avec le kit miRNeasy Serum/Plasma de Qiagen (www.qiagen.com), à partir de 400 μL de plasma conservé dans des tubes contenant du citrate de sodium (permettant d'empêcher la coagulation des échantillons). Dans un premier temps, le réactif *QIAzol Lysis* est introduit pour faciliter la lyse des membranes lipidiques des vésicules et dénaturer les complexes protéiques, tout en désactivant les RNAses. Les miARNs présents dans les vésicules et dans les RISC circulants sont ainsi libérés sans être dégradés. L'ajout de chloroforme permet la séparation en 2 phases: une phase aqueuse contenant les ARNs et un peu d'ADN, et une phase phénolique contenant quelques protéines. Une interface entre les deux phases contient la plupart des protéines et de l'ADN. La phase contenant les ARNs peut alors être extraite pour obtenir un échantillon purifié après élution (Figure II.1).

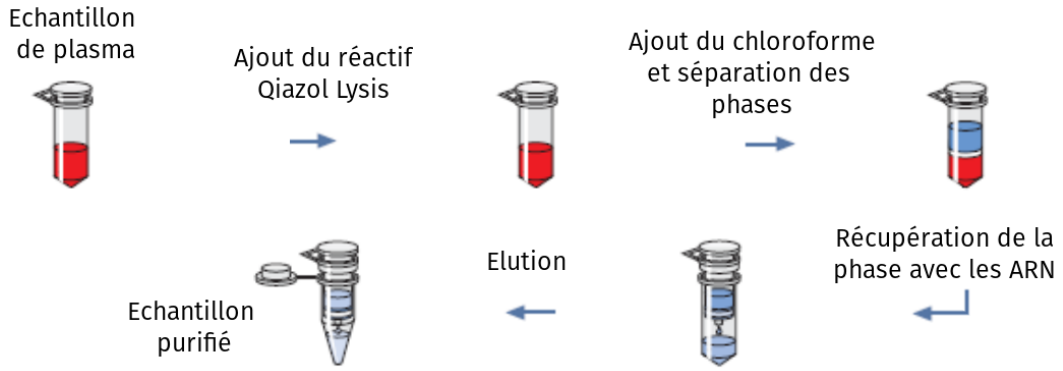


Fig. II.1 – Purification des librairies avec le kit miRNeasy de QIAGEN

Ligation des adaptateurs La seconde étape a pour but d'ajouter les adaptateurs aux extrémités des brins d'ARN. Ces adaptateurs sont des séquences de nucléotides permettant d'amorcer le séquençage des brins d'ARN. Cette étape est effectuée avec le kit NEBNext Multiplex Small RNA Library Prep Set pour Illumina (www.neb.com), à partir de 6 μL d'ARN extraits lors de l'étape précédente. Dans un premier temps, les adaptateurs 5' et 3' sont ligaturés aux extrémités correspondantes des ARNs, et un *primer* complémentaire à l'adaptateur 3' est introduit. Le *primer* permet d'initier la transcription inverse d'un brin d'ADN complémentaire (ADNc) au brin d'ARN, qui est rendue possible grâce à l'introduction d'une Reverse-Transcriptase. Les brins d'ARN sont alors éliminés pour ne conserver que les brins d'ADNc. Les séquences des adaptateurs contiennent également les séquences P5 et P7, qui sont les séquences universelle d'Illumina, cruciales lors du séquençage. Une séquence "code barre", de 6 nucléotides, est également présente entre l'adaptateur 3' et la séquence P7. Cette séquence est appelée code barre car elle permet d'identifier l'échantillon séquencé. Enfin, 15 cycles de *Polymerase Chain Reaction* (PCR) permettent de produire les duplex d'ADNc, et d'amplifier le nombre de duplex. Le protocole de ligation des adaptateurs est représenté sur la Figure II.2. Les fragments d'ARN sont donc convertis en ADNc, et sont encapsulés entre deux adaptateurs, comme représenté sur la Figure II.3.

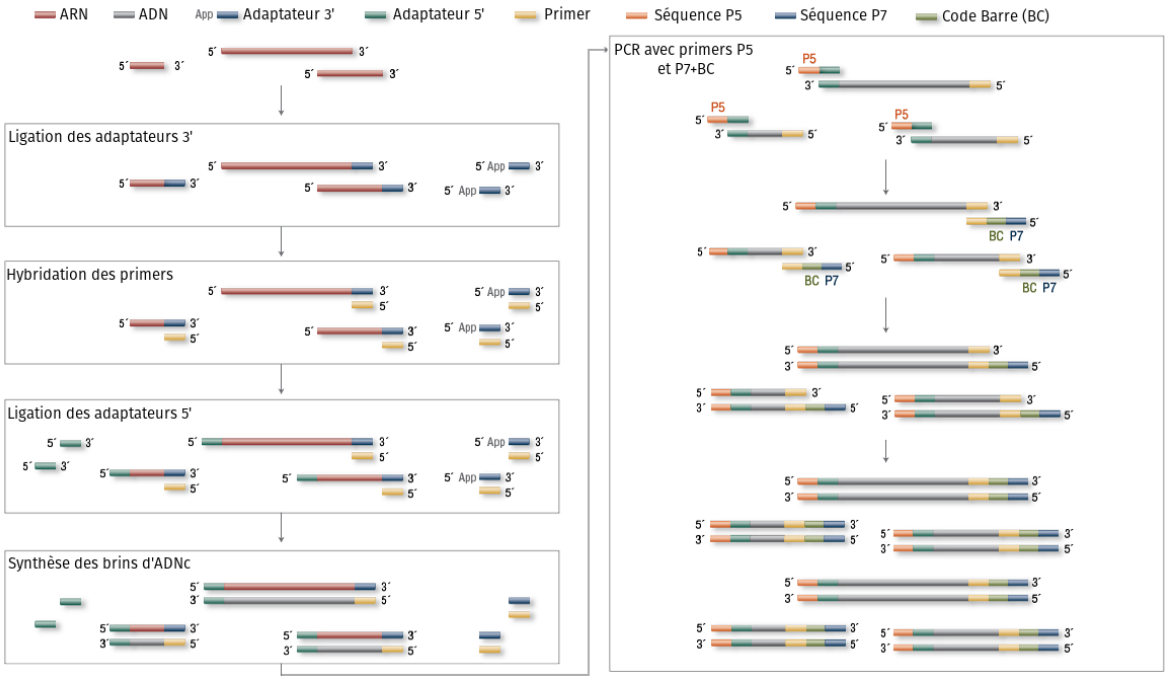


Fig. II.2 – Ligation des adaptateurs avec le protocole de NEBnext.

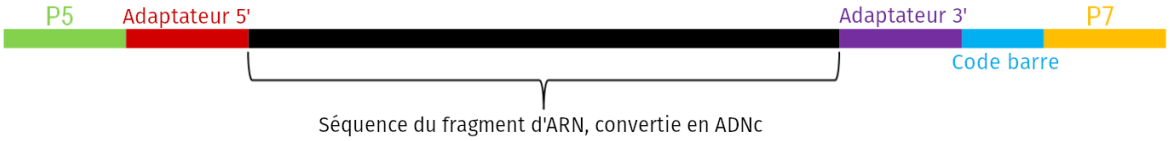


Fig. II.3 – Un fragment d'ARN prêt à être séquencé.

Sélection des petits ARNs La troisième et dernière étape de la préparation des librairies consiste à sélectionner les brins selon leur taille. Les miARNs ont une taille d'environ 22 nucléotides, à laquelle il faut ajouter les adaptateurs ligaturés (y compris le code barre et les séquences P5/P7), pour obtenir la taille des duplex, soit environ 143 paires de bases (bp). Un protocole fourni par NEB, basé sur l'extraction par billes magnétiques XP AMPure, a été optimisé afin d'extraire les brins d'ARN entre 100 et 160bp en 3 étapes. D'abord les petits fragments sont éliminés grâce à des billes purifiantes 1.5X permettant de capturer les fragments de taille supérieure à 100bp. A partir des fragments capturés, des billes 1.3X permettent de capturer les fragments de taille supérieure à 160bp. Ces derniers sont alors éliminés pour ne conserver que les fragments avec une taille entre 100 et 160bp.

Multiplexage Grâce à l'utilisation des codes barres, permettant d'identifier l'échantillon d'origine auquel appartient un fragment, les librairies peuvent être multiplexée, c'est à dire mélangées entre elles pour être séquencées en une seule opération. Le nombre d'échantillons multiplexés va dépendre du nombre de fragments que l'on veut séquencer par individu, mais aussi du séquenceur et du protocole de séquençage envisagé. Typiquement, plusieurs centaines

de millions de fragments peuvent être séquencés en une opération, donc en multiplexant 24 échantillons en une librairie, comme effectué dans le cadre de ce protocole, on peut obtenir la séquence de plus de 10 millions de fragments pour chaque échantillon.

Dans le cadre de ce projet, 435 échantillons plasmatiques ont été sélectionnés, et des réplicats ont été préparés pour estimer la qualité du séquençage. Deux types de réplicats ont ainsi été préparés: des réplicats biologiques, qui ont suivi le protocole de préparation indépendamment, et des réplicats techniques, qui ont été préparés à partir des librairies finalisées. Au total, 45 réplicats ont été préparés dont 7 biologiques et 38 techniques. Ainsi, 480 échantillons ont été préparés au total, multiplexés en 20 librairies.

2.2 Séquençage haut débit

Les librairies multiplexées ont été séquencées sur une machine Illumina NextSeq500 (www.illumina.com), qui peut séquencer environ 400 millions de fragments simultanément en une dizaine d'heures. Le protocole est réalisé automatiquement, et requiert, en plus de la librairie de fragments à séquencer, les réactifs nécessaires au séquençage ainsi qu'un *flow cell*, le support sur lequel sont déposés les fragments pour être séquencés. La méthode développée par Illumina est le séquençage par synthèse massivement parallélisé, qui comprend deux étapes principales: l'amplification des fragments en *clusters*[1], et le séquençage par synthèse des brins complémentaires.

Amplification des fragments en clusters Les fragments d'ADNc sont d'abord répartis sur le flow cell, sur lequel sont déjà fixés plusieurs milliards d'oligos[2] de deux types: l'un complémentaire à la séquence P5 de l'adaptateur et l'autre complémentaire à la séquence P7. Ainsi, les fragments déposés sur le flow cell sont dénaturés et viennent se lier par les 2 extrémités aux oligos par complémentarité. Des polymérases sont ensuite introduites pour synthétiser le brin complémentaire de l'ADNc en partant de l'oligo (cette étape est appelée phase d'extension). Le duplex formé est ensuite dénaturé pour libérer le brin original, qui est éliminé, afin de conserver uniquement les nouveaux brins synthétisés, qui sont prolongés par les oligos rattachés au flow cell.

Pendant la phase d'amplification, les fragments vont se replier pour apparier l'autre extrémité du fragment à un oligo secondaire libre, et des polymérases sont à nouveau introduites pour synthétiser le brin complémentaire à partir de l'oligo secondaire. Ensuite les duplex sont à nouveau dénaturés, mais comme les deux brins sont désormais prolongés par les oligos, ils sont tous les deux conservés. Le cycle est répété une trentaine de fois, permettant de générer des clusters contenant chacun des milliers de copies identiques du fragment original et de son complémentaire.

Le flow cell contient alors des centaines de millions de clusters, chacun composé de milliers de copies d'un fragment, à la fois l'original rattaché par la séquence P5 et le complémentaire rattaché par la séquence P7. Les fragments originaux attachés par l'extension P5 au flow cell sont éliminés, pour conserver uniquement les fragments complémentaire, permettant de séquencer les fragments dans un sens unique. Cette phase d'amplification est représentée sur la Figure II.4.

1. un cluster de fragments correspond à un nombre important de fragments localisés dans un perimètre réduit
2. Séquence de nucléotides synthétique

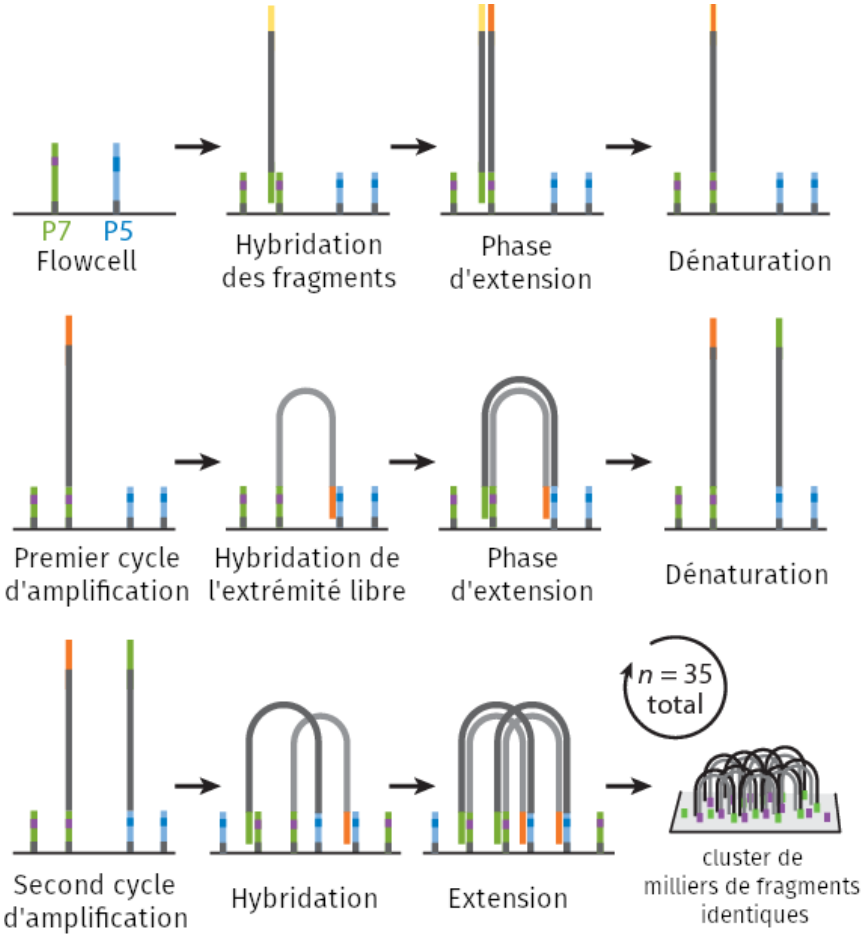


Fig. II.4 – Phase d'amplication permettant d'obtenir des clusters de fragments identiques.

Séquençage par synthèse Lors de la phase de séquençage, un brin complémentaire à chaque fragment présent est synthétisé sur plusieurs cycles. Les nucléotides incorporés pour cette synthèse sont liés à des molécules qui émettent une fluorescence distincte à chaque nucléotide lorsqu'ils sont appariés à un brin [3]. À chaque cycle, un nucléotide est donc apparié sur chaque fragment, et émet une fluorescence qui est amplifiée grâce aux cluster de fragments identiques. Cette fluorescence est alors capturée par des caméras pour enregistrer le nucléotide séquencé.

Le principe de capture du NextSeq500 repose sur deux caméras, chacune avec un filtre permettant de capturer des longueurs d'ondes spécifiques, permettant à la première caméra de capturer les fluorescence vertes et à la seconde de capturer les rouges. Les fluorescence émises par chaque nucléotide sont les suivantes: la cytosine émet du rouge, la thymine émet du vert, l'adénine émet à la fois du rouge et du vert, et la guanine n'émet pas de fluorescence. Ainsi, la lecture des images capturées à chaque cycle permet de déterminer la séquence du fragment synthétisé. La Figure II.5 représente ce processus.

3. La fluorescence est émise par chaque nucléotide hybridé est rendue possible grâce à un laser qui excite les molécule liées à chaque nucléotide

Le nombre de cycles de la phase du séquençage détermine la taille des fragment séquencés. Ce nombre a été fixé à 76 dans le cadre de ce projet.

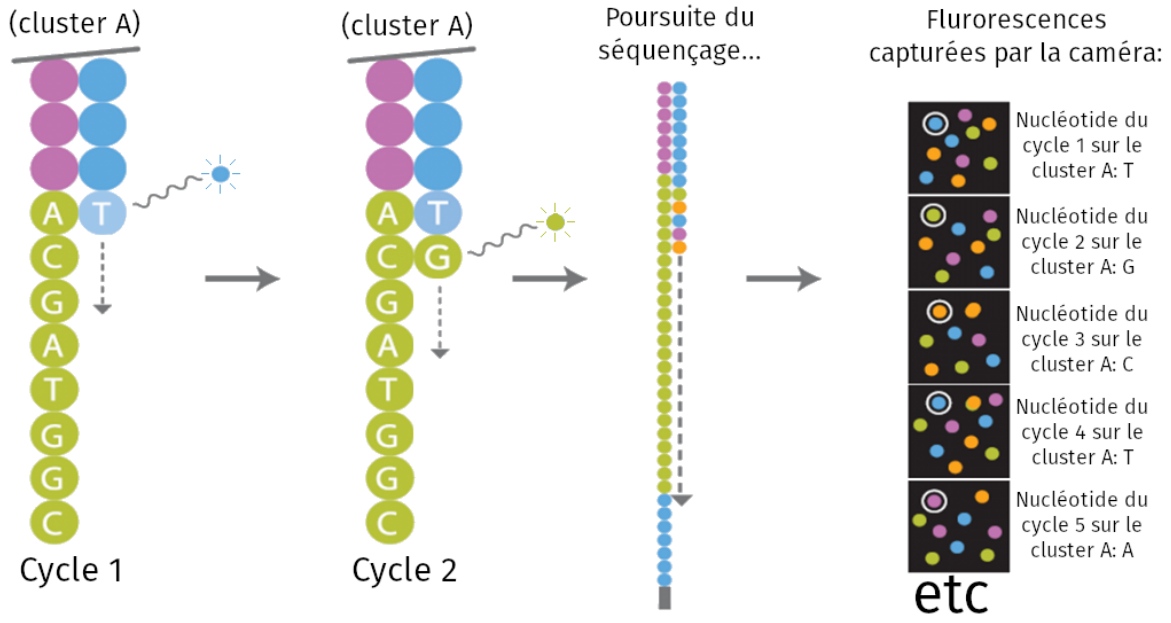


Fig. II.5 – Séquençage par synthèse d'Illumina.

2.3 Qualité du séquençage et fichier fastq

Un indice reflétant la qualité de lecture est associé à chaque nucléotide séquencé. Cet indice est un score (appelé BQ-score pour *Base Quality score* en anglais) de type phred, qui indique la probabilité d'une erreur pour chaque nucléotide interprété:

- BQ - score = 10 indique une probabilité d'erreur de 0.1
- BQ - score = 20 indique une probabilité d'erreur de 0.01
- BQ - score = 30 indique une probabilité d'erreur de 0.001
- BQ - score = 40 (score maximum) indique une probabilité d'erreur de 0.0001

Plus le score est élevé, plus la lecture est fiable. Ce score est ensuite encodé en caractères ASCII permettant d'assigner un symbole unique aux numéros 0 à 40 (par exemple " !" correspond au score 0, ou "A" correspondant au score 32).

A la suite du séquençage, la machine Illumina produit un fichier *fastq* par échantillon (démultiplexé), qui contient l'ensemble des lectures des fragments séquencés, appelés *reads*, et les BQ-score associés à chaque base. Chaque read est précédé d'un en-tête avec notamment un identifiant unique, les coordonnées du cluster correspondant sur le flow cell, et un code illumina indiquant les reads de mauvaise qualité (qui sont en général automatiquement filtrés). Chaque fichier contient plusieurs millions de reads, et une analyse bioinformatique est nécessaire pour les annoter, c'est à dire pour identifier le petit ARN auquel il correspond.

2.4 Bruit de fond de séquençage

Idéalement, les reads issus du séquençage représentent parfaitement les miARNs présents dans l'échantillon séquencé. Cependant, les technologies NGS actuelles ne permettent pas encore d'obtenir un séquençage parfait, notamment à causes d'erreurs parfois induites lors du séquençage ou de la préparation des librairies, voire parfois à cause d'une contamination externe de l'échantillon. Ces perturbations génèrent ce qu'on appelle le bruit de fond de séquençage, qui complexifie l'analyse des données de séquençage. En particulier, l'étape de préparation des librairies peut être une source importante de bruit de fond de séquençage.

En amont du séquençage, la préparation des librairies est réalisée grâce à des kits spécifiques, dont les plus populaires sont TruSeq (fabriqué par Illumina), NEBnext (fabriqué par NEB), et plus récemment NEXTflex (fabriqué par Bioo). La préparation requiert plusieurs étapes qui ont été détaillées précédemment: la ligation des adaptateurs, la synthèse des brins d'ADN complémentaires et l'amplification PCR. Chaque étape, en particulier la ligation des adaptateurs, peut introduire un biais, favorisant certaines séquences au détriments d'autres, et impactent ainsi la précision du séquençage [126, 230, 291, 142]. Si le même protocole est suivi pour tous les échantillons au sein d'une étude, ces biais n'auront pas de conséquence sur la comparaison d'un miARN entre différents échantillons, mais ils affecteront les résultats lorsqu'il s'agira de comparer des miARNs entre eux. De plus, ces biais peuvent également impacter la reproductibilité de résultats obtenus par différentes études utilisant des protocoles différents.

De plus, cette étape de préparation peut également altérer la séquence des miARNs. Dans une étude récente [292] ayant pour objectif de quantifier les biais propres aux kits de préparation les plus utilisés pour séquencer les petits ARNs, des quantités équimolaires de 962 miARNs synthétiques (avec une séquence fixe) ont été séquencées avec chacun des 3 kits les plus populaires. Et de façon surprenante, des séquences avec des variations ressemblant aux isomiRs, appelées pseudo-isomiRs, ont été détectées en grande quantité dans les données issues du séquençage (Figure II.6).

Le kit NEXTflex, dont la particularité est d'utiliser des séquences aléatoires dans les adaptateurs, est celui qui génère le moins de pseudo-isomiRs (n=2221), suivi des kits Truseq (n=5021) et NEBnext (n=9074). L'expression de chaque pseudo-isomiR est également très importante, en particulier pour le kit NEBnext avec une expression moyenne des pseudo-isomiRs qui est 60% plus importante que pour le kit NEXTflex. Les pseudo-isomiRs observés sont principalement des modifications en 3' de la séquence du miARN. Ce bruit introduit lors de la préparation des échantillons risque donc d'entraver considérablement l'étude des isomiRs à partir des données NGS, et des progrès pour limiter ce bruit sont nécessaires.

Les kits intégrants des adaptateurs aléatoires semblent avoir de meilleurs résultats, comme le confirme une nouvelle méthode baptisée AQ-seq [142] qui a récemment été développée dans le but de réduire significativement ces biais, et intègre des adaptateurs aléatoires ainsi que des optimisations spécifiques au séquençage des petits ARNs.

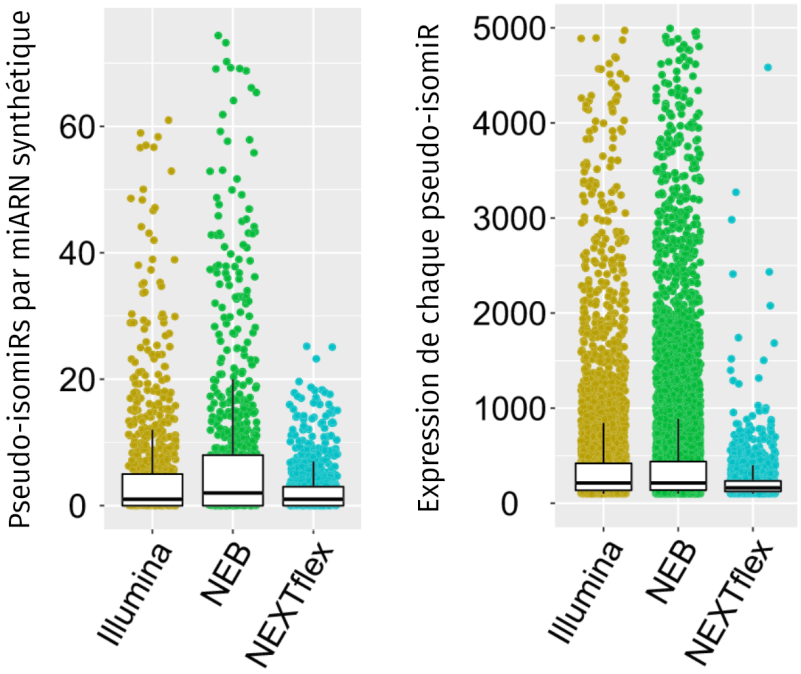


Fig. II.6 – Les pseudo-isomiRs ont été quantifiés pour les trois kits de séquençage utilisés. Sur le panel de droite, le nombre de pseudo-isomiRs est représenté pour chacun des 962 miARNs synthétiques, tandis que l'expression de chaque pseudo-isomiR est indiquée sur le panel de droite. *Image adaptée de [292]*

Chapitre III : Détection et quantification de microARNs avec optimiR

1 Introduction: principes, challenges et travaux préliminaires

La première étape de l'analyse bioinformatique de données de séquençage NGS est similaire pour tous les types de données séquencées (ADN, ARNs, ou petits ARNs): elle débute par une procédure d'alignement. Cette procédure consiste à trouver l'alignement le plus pertinent d'un read contre un ensemble de séquences de référence, appelé librairie de référence. Lors de cette procédure, une comparaison est effectuée entre la séquence introduite (le read) et celles de la librairie de référence, afin d'identifier la séquence de la librairie de référence la plus proche de celle du read. L'alignement permet ainsi d'identifier chaque read, par exemple grâce aux coordonnées génomiques du locus sur lequel il s'aligne [1], ou grâce au nom du transcrit sur lequel il s'aligne, suivant le choix de la librairie de référence utilisée. Comme les miARNs sont des molécules de petite taille (environ 22nt), ils sont entièrement séquencés au sein d'un seul read [2], on peut donc quantifier l'abondance d'un miARN en comptant les reads alignés sur un même locus ou un même transcrit (Figure III.1).

1.1 Challenges de l'alignement de microARNs

Afin de quantifier de façon précise les miARNs séquencés dans un échantillon, il faut limiter le nombre de faux négatifs et de faux positifs lors de l'alignement. Les faux négatifs correspondent à:

— un fragment séquencé correspondant à un miARN qui n'a pas été aligné (ex: isomiR ou polymiR).

Et les faux positifs correspondent à:

— l'alignement d'un fragment séquencé correspondant à un miARN sur le mauvais miARN ;
— l'alignement d'un fragment séquencé ne correspondant pas à un miARN sur un miARN.

Il n'y a actuellement pas de standard définissant les règles d'alignement des miARNs pour obtenir un résultat optimal. Ceci est dû à la nature complexe des miARNs. D'un côté à cause de leur petite taille et des nombreux miARNs paralogues qui induisent une homogénéité de

1. On parle également de *mapping* lorsque l'alignement permet de localiser le locus d'origine d'un read
2. Comme le séquençage est réalisé sur 76 cycles, les reads ont tous une taille de 76nt, et contiennent donc à la fois la séquence du miARN, et le début de la séquence des adaptateurs. Cette dernière est retirée des reads avant l'alignement.

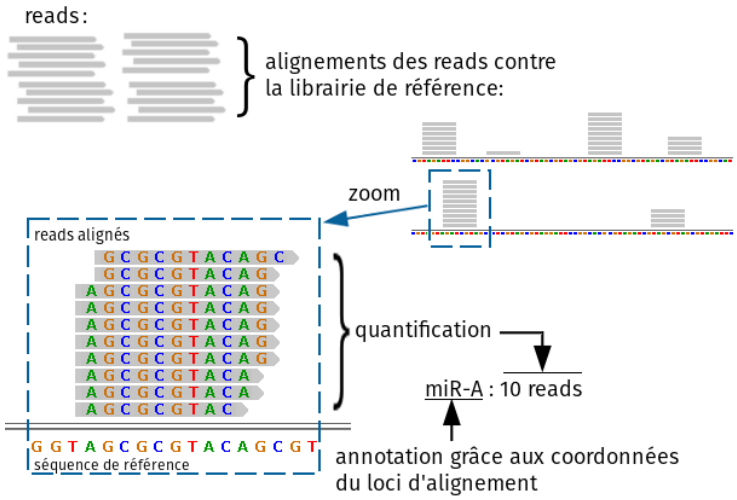


Fig. III.1 – Exemple d'identification et de quantification par alignement d'un miARN fictif séquencé (dont la taille est volontairement diminuée, dans un soucis de clarté).

séquence inter-miARN (Figure III.2.A). De l'autre côté à cause des polymiRs et des isomiRs, qui introduisent une hétérogénéité de séquence intra-miARN (Figure III.2.B).

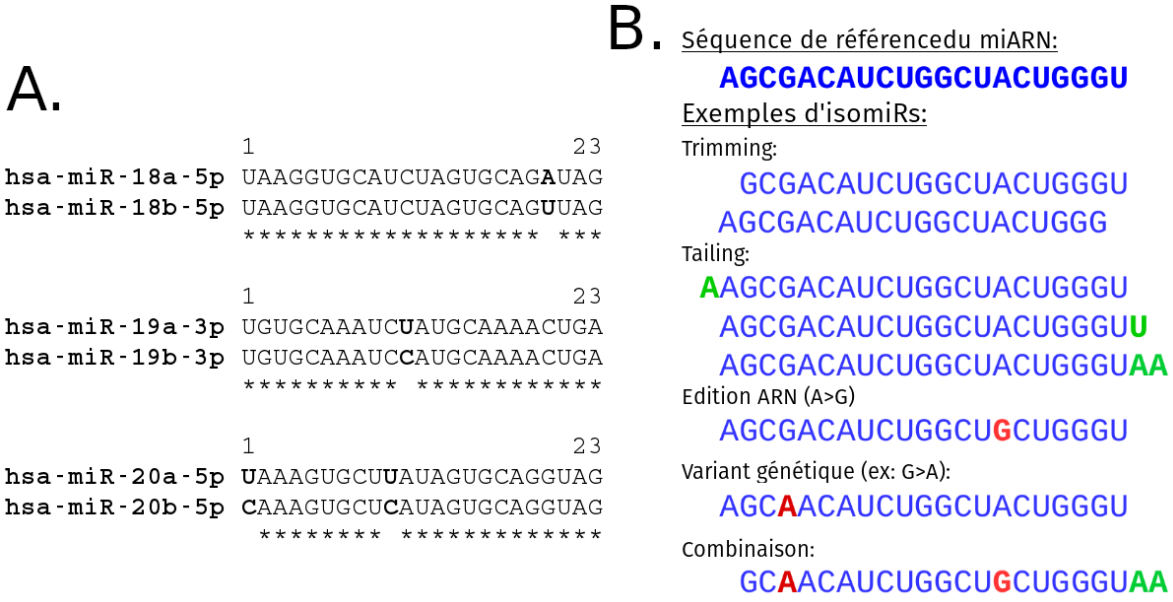


Fig. III.2 – (A) Exemples de miARNs paralogues. (B) Exemples d'isomiRs.

L'alignement des reads doit donc être assez stringent pour distinguer les miARNs paralogues, sans générer de faux négatifs. Et l'alignement être assez permissif pour aligner les polymiRs et isomiRs, sans introduire de faux positifs. Si un alignement stringent génère uniquement des alignements parfaits, un alignement permissif requiert l'autorisation d'alignements imparfaits,

c'est à dire avec des mésappariements, aussi appelés *mismatchs*[3].

L'autorisation de mismatchs lors de l'alignement de reads est largement adoptée pour les protocoles analysant de longs fragments, afin d'aligner des fragments contenant des variants génétiques. Cependant, pour l'analyse de long fragments, les mismatchs représentent une petite partie de la séquence (ex: pour des fragments de 250 nucléotides, 2 mismatchs représentent moins de 1% de la séquence) mais pour l'analyse des miARNs, d'une taille d'environ 20 nucléotides, 2 mismatchs représentent déjà 10% de la séquence. Cependant, certains isomiRs ont des nucléotides modifiés par rapport à la séquence de référence: les isomiRs avec édition ARN, les polymiRs, et les isomiRs avec tailing par TNTase. Ces isomiRs requièrent donc l'utilisation de mismatchs pour être alignés.

De plus, à cause de la petite taille des reads et de l'homologie entre certains miARNs, les reads peuvent parfois s'aligner de façon multiple[4], c'est à dire sur plusieurs loci, et sont appelés alignements ambigus. Le dernier challenge est donc de déterminer quel est l'alignement le plus probable pour un read qui s'aligne de façon ambiguë, c'est à dire déterminer le miARN dont il est originaire.

1.2 Choix de la librairie de référence pour l'alignement

Le choix de la référence dépend d'abord des objectifs de l'étude. Si un de ces objectifs est d'identifier des miARNs *de novo*, c'est à dire des miARNs qui n'ont pas encore été identifiés, alors l'alignement contre le génome entier est actuellement la meilleure façon d'atteindre cet objectif. Par exemple, de nouveaux miARNs humains potentiels ont encore récemment été identifiés [3, 15]. Cette approche est particulièrement préférable si l'étude est basée sur des organismes dont les miARNs ont été peu, ou pas étudiés. Chez l'homme, il est raisonnable de présumer que les données accumulées pendant les vingt dernières années ont permis d'identifier la grande majorité des miARNs, ou du moins les miARNs les plus conservés et les plus abondants, et donc ceux qui ont vraissemblablement un rôle biologique.

En pratique, on peut utiliser une référence plus restreinte que le génome complet pour l'alignement des miARNs. D'abord pour des raisons pratiques, car l'alignement de l'ensemble des reads sur le génome peut prendre du temps, notamment lorsqu'il faut aligner les lectures de centaines d'échantillons. Ensuite car les fragments séquencés sont de petite taille, les reads ont donc une grande probabilité de s'aligner de façon ambiguë sur différents loci homologues du génome, ce qui rend l'identification du fragment séquencé plus complexe. Si l'analyse est centrée sur l'étude de transcrits connus, tels que les miARNs référencés dans la miRBase, alors restreindre la taille de la librairie de référence permet de réduire à la fois le temps d'alignement et la complexité de l'analyse. Pour cette raison, l'alignement des miARNs est souvent effectué sur les séquences des miARNs précurseurs ou matures fournies par la miRBase.

Suivant la stratégie d'alignement implémentée, il peut être préférable d'utiliser les séquences de précurseurs, en particulier pour aligner les miARNs avec un ou plusieurs évènements de tailing. Car les reads contenant du tailing ont une taille supérieure à celle de la séquence de référence du miARN mature, et en général, les méthodes d'alignements classiques ne permettent pas d'aligner un read sur une séquence de taille plus petite. De plus la séquence du précurseur

3. Un mismatch (en anglais pour mésappariement) lors d'un alignement correspond à une différence d'un nucléotide entre la séquence du read et celle de la référence.

4. Le terme "alignement multiple" peut avoir différentes significations, et est souvent utilisé pour désigner la comparaison d'au moins 3 séquences (généralement pour inférer leur homologie). Dans ce chapitre, ce terme désignera les alignements ambigus, c'est à dire les reads étant alignés sur plusieurs loci distincts (aussi appelé *cross-mapping reads* en anglais).

permet de vérifier si les nucléotides ajoutés par tailing sont identiques à ceux présents dans la séquence du précurseur, permettant ainsi de discriminer le *templated tailing* (produit par une imprécision de découpe par Dicer ou par le microprocesseur) du *non templated tailing* (ou NTA, produit par les TNTases). On pourra toutefois noter qu'une TNTase pourrait ajouter des nucléotides identiques à la séquence du précurseur, faisant ainsi passer un *non templated tailing* pour un *templated tailing*, en particulier si les nucléotides ajoutés par tailing sont des A ou des T.

1.3 Travaux préliminaires

Afin de déterminer la stratégie d'alignement à utiliser pour quantifier les miARNs, j'ai effectué quelques travaux préliminaires. Parmi ces travaux, j'en ai sélectionné deux qui sont présentés ici. Le premier permet d'avoir un aperçu de l'impact d'un alignement permissif sur le nombre d'alignements ambigus résultants. Et le second permet d'évaluer la complexité de la détection des évènements d'édition ARN, en prenant en compte le bruit de fond de séquençage. J'ai également eu l'occasion de tester de nombreux pipelines d'alignements de miARNs dont les stratégies implémentées seront présentées.

a) Effets d'un alignement permissif

Afin de comparer les résultats entre un alignement stringent (sans mismatch) et un alignement permissif (avec mismatchs), j'ai effectué une comparaison de l'alignement de plusieurs jeux de données, sur plusieurs librairies de référence, en incrémentant le nombre de mismatchs autorisés (0, 1 et 2). Ce travail permet en particulier d'illustrer l'effet du nombre de mismatchs sur le nombre d'alignements ambigus résultants.

Les 3 librairies de références utilisées sont a) le génome humain (référence hg38) ; b) les séquences des pré-miARNs indexées dans la miRBase 21 (n=1881) ; et c) celles des miARNs matures correspondants (n=2588). Les 4 jeux de données utilisés sont i) le jeu *miRBase*, identique à la librairie c), correspondant aux 2588 séquences des miARNs matures indexés dans la miRBase 21 ; ii) le jeu *shuffled*, correspondant à un ensemble de 321 séquences aléatoires de pseudo-miARNs (générées par l'étude de Lewis et al 2003 [172]) ; iii) deux échantillons *ÉchantillonA* et *ÉchantillonB*, tirés aléatoirement parmi les 435 échantillons plasmatiques séquencés pour les petits ARNs dans le cadre de cette thèse, composés d'environ 15 millions et 12 millions de reads, respectivement. L'alignement a été réalisé avec le programme *bowtie* [162], paramétré de façon à reporter tous les alignements qui respectent le nombre de mismatchs autorisés (option *-a*). Chaque jeu de données a été aligné sur chacune des 3 librairies de références, dans un premier temps sans aucun mismatch autorisé, puis avec 1 mismatch autorisé, et enfin avec 2 mismatchs (option *-v i*, i étant un entier correspondant au nombre de mismatchs autorisés). À noter que les alignements sur le génome peuvent être de type complémentaire inversés, afin de détecter les miARNs transcrits à partir du brin opposé, mais ces alignements sont désactivés sur les librairies de référence des miARNs précurseur et matures (grâce à l'option *–norc*).

Cette stratégie est naïve, car elle n'est pas optimisée pour rapporter les alignements les plus probables. Elle permet néanmoins de comparer la proportion des reads qui s'alignent de façon unique ou de façon ambiguë, suivant différentes librairies de référence, et suivant le niveau de stringence matérialisé par l'autorisation de mismatchs. Les résultats de ces alignements sont donnés dans la Table III.1.

		miRBase (n=2588)			Shuffled (n=316)			ÉchantillonA (n=14,905,552)			ÉchantillonB (n=12,013,535)		
	Nombre de mismatches:	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**
Genome	Temps	2"	14"	22"	1"	1"	15"	15'33	81'53	762'20	11'35	69'40	753'02
	% de reads alignés	100.00%	100.00%	100.00%	0.95%	8.86%	53.16%	35.31%	51.04%	59.50%	27.15%	43.14%	54.37%
	% de reads alignés de façon unique	82.70%	60.28%	22.06%	0.00%	67.86%	39.88%	22.24%	14.40%	5.30%	24.41%	15.21%	6.17%
	% de reads alignés de façon ambiguë	17.30%	39.72%	77.94%	100.00%	32.14%	60.12%	77.76%	85.60%	94.70%	75.59%	84.79%	93.83%
	Nombre total d'alignements	136,396	860,102	3,026,001	13	293	4,310	120,739,124	609,347,611	4,712,694,271	71,825,616	398,632,469	3,911,673,631
pre-miARNs	Temps	1"	1"	1"	1"	1"	1"	1'07	1'17	3'52	46"	58"	3'34
	% de reads alignés	100.00%	100.00%	100.00%	0.00%	0.00%	0.00%	6.88%	9.03%	9.91%	5.89%	7.58%	8.55%
	% de reads alignés de façon unique	91.34%	87.06%	81.41%	x	x	x	89.22%	80.81%	64.69%	89.44%	78.96%	60.04%
	% de reads alignés de façon ambiguë	8.66%	12.94%	18.59%	x	x	x	10.78%	19.19%	35.31%	10.56%	21.04%	39.96%
	Nombre total d'alignements	2948	3494	4581	x	x	x	1,166,806	1,769,089	2,522,098	799,466	1,200,036	1,796,646
miARNs	Temps	1"	1"	1"	1"	1"	1"	58'	1'13	3'21	46"	55"	2'42
	% de reads alignés	100.00%	100.00%	100.00%	0.00%	0.00%	0.00%	6.23%	7.05%	7.26%	5.31%	5.95%	6.14%
	% de reads alignés de façon unique	97.95%	94.13%	89.84%	x	x	x	99.53%	89.08%	68.95%	99.62%	87.66%	66.21%
	% de reads alignés de façon ambiguë	2.05%	5.87%	10.16%	x	x	x	0.47%	10.92%	31.05%	0.38%	12.34%	33.79%
	Nombre total d'alignements	2685	2942	3340	x	x	x	934,172	1,217,312	1,648,966	641,055	831,314	1,135,021

Table III.1 – Comparaison d'alignements stringents et permissifs

miRBase Si on considère l'alignement des 2588 séquences de miARNs matures de la miRBase, elles sont logiquement toutes alignées, quelle que soit la librairie de référence utilisée. Quand elles sont alignées contre elles-mêmes, on observe que même en l'absence de mismatchs, 2.05% des séquences sont alignées de façon ambiguë, car certaines séquences de miARNs sont parfaitement identiques, ou incluses dans la séquence d'un autre miARN. Ce résultat augmente jusqu'à 10.16% lorsque l'on autorise 2 mismatchs, à cause de l'homologie de certains miARNs. Sur le génome, l'autorisation de 2 mismatchs résulte avec plus de 3 millions d'alignements au total, ce qui indique qu'il y a beaucoup de loci sur le génome avec des séquences homologues à celles des miARNs.

Shuffled Le jeu de données *shuffled*, composé de séquences aléatoires d'environ 22 nucléotides, n'a naturellement aucun alignement reporté sur les séquences des précurseurs et des matures. Seulement 3 séquences (0.95% du total) sont alignées sur le génome lorsqu'aucun mismatch n'est autorisé, mais plus de la moitié des reads trouvent un loci où s'aligner lorsque 2 mismatchs sont autorisés. Ainsi, à cause de la petite taille des miARNs, on peut trouver des séquences homologues même avec des séquences aléatoires, lorsque la librairie de référence est assez grande. Ce résultat suggère donc que certaines petites séquences d'origine exogène peuvent éventuellement être alignées sur une librairie de référence humaine, notamment lorsque l'alignement est permissif et que la librairie est assez grande.

Échantillons réels Concernant les échantillons A et B, il y a remarquablement peu de reads alignés sur les séquences de la miRBase, avec un maximum de 7.26% des reads alignés lorsque 2 mismatchs sont autorisés, ce qui indique un faible nombre de miARNs présents dans ces échantillons. Sur le génome, on observe jusqu'à 60% de reads alignés lorsque 2 mismatchs sont autorisés, ce qui suggère que la majorité des reads présents dans les échantillons séquencés correspondent à des fragments d'ARN qui ne sont pas des miARNs. Pour les 40% de reads non alignés, il y a plusieurs explications possibles:

— Les séquences des fragments non alignées ont été soumises à trop de modifications, soit post-transcriptionnelles, soit lors de la préparation des librairies ou lors du séquençage, pour être identifiables avec seulement 2 mismatchs autorisés ;

— Il peut s'agir de séquences d'origine exogène, par exemple à cause d'une contamination lors de la préparation [165], ou à cause d'autres petits ARNs d'origine virale tels que les cytomégalovirus [226, 196] qui peuvent être présents dans le plasma. Pour vérifier cette hypothèse, les fragments non alignés contre le génome humain les plus importants

(présents avec plus de 10.000 copies) ont été alignés contre la base de donnée *nr* avec l'outil *Blastn*. Pour chaque fragment, de nombreux alignements contre divers génomes ont été retournés, en particulier des génomes de fungi, de rongeurs et de microalgues. Ces fragments ne semblent donc pas avoir une origine unique, mais diverse et complexe. Enfin, le nombre d'alignements ambigus sur les séquences de la miRBase est pratiquement nul lorsqu'aucun mismatch n'est autorisé, ce qui suggère que les miARNs présents dans ces échantillons ne font principalement pas partie de ceux qui possèdent une séquence identique ou incluse dans celle d'un autre miARN. Mais ce nombre augmente considérablement avec le nombre de mismatchs autorisés: plus de 30% des reads sont alignés de façon ambiguë lorsque 2 mismatchs sont autorisés.

Dans l'ensemble, on observe que la taille de la librairie de référence et le nombre de mismatchs autorisés sont corrélés avec le temps d'analyse et la proportion de reads alignés de façon ambiguë. Pour l'alignement des petits ARNs, il faut donc trouver une stratégie d'alignement permettant de limiter le nombre d'alignement ambigus, et / ou de résoudre l'ambiguïté de ces alignements multiples, afin de trouver l'alignement le plus probable.

b) Détection des évènements d'édition ARN

Certains sites ARN sont préférentiellement édités par les protéines ADAR. D'après la base de données REDIportal [227], qui recense les sites d'édition ARN A>G, 130 sites d'éditions sont localisés dans la séquence de miARNs matures. J'ai donc effectué une analyse à partir des échantillons ÉchantillonA et ÉchantillonB, utilisés précédemment, pour déterminer sur chacun de ces 130 sites la proportion de reads comportant une substitution A>G. Pour cela, j'ai effectué un alignement de ces deux échantillons sur les séquences de miARNs matures, en autorisant 1 mismatch. En cas d'alignement ambigus, les alignements sans mismatchs sont rapportés en priorité.

Pour ÉchantillonA, seulement 26 miARNs contenant un des 130 sites d'édition sont exprimés, contre 24 pour ÉchantillonB. Sur les 26 miARNs de ÉchantillonA, seuls 6 ont des reads comportant une substitution A>G sur les sites référencés par REDIportal, et 5 pour ÉchantillonB, dont 3 sont communs aux deux échantillons. La proportion de reads contenant une substitution A>G par rapport au nombre de reads alignés sur le miARN varie entre 0.01% et 0.49% . Il y a donc un très faible nombre de reads qui supportent ce phénomène d'édition sur ces sites.

Afin de déterminer si ces sites peuvent être identifiés de manière fiable comme des site d'édition A>G, et non comme des substitution induites par du bruit de fond lors de la préparation des librairies ou lors du séquençage, j'ai effectué une estimation du taux moyen de substitutions observé par base alignée sur l'ensemble des reads alignés sur les séquences de miARNs matures. Pour que le résultat ne soit pas biaisé par les variants génétiques et les évènements de NTA, j'ai exclu les positions avec des variants génétiques et je n'ai considéré que les bases internes des reads alignés (en excluant les 2 bases situées aux extrémités 3' et 5').

Pour l'échantillon ÉchantillonA, le taux de substitution moyen est de 0.39%, contre 0.42% pour ÉchantillonB. On peut donc considérer que pour chaque position, environ 0.40% des nucléotides alignés sont assujettis à du bruit de fond induit lors de la préparation des librairies et du séquençage. La substitution majoritairement observée est T>C (28% des substitutions en moyenne sur les 2 échantillons), suivie de A>G (19%) et T>A (12%), le reste des substitutions s'étalant entre 2 et 10%.

Ainsi, les fréquences des sites d'édition observées pour ces deux échantillons (inférieures

à 0.5%) sont difficiles à distinguer du bruit de fond (environ 0.4%). Les sites d'édition ARN sont donc particulièrement complexes à détecter dans nos données. Il est aussi possible que les miARNs présents dans nos échantillons soient très faiblement édités, car ce phénomène a principalement été identifié dans des tissus cérébraux. J'ai donc fait le choix de ne pas considérer les évènements d'édition ARN pour établir la stratégie d'alignement.

c) Description des pipelines d'alignement et stratégies existantes

Plusieurs traitements peuvent être effectués avant et après l'alignement. En amont de l'alignement on effectue généralement une étape de contrôle qualité du séquençage, et on réduit la taille des reads en retirant les séquences en 3' qui correspondent aux adaptateurs séquencés à la suite du fragment. Après alignement, on peut par exemple raffiner les alignements et quantifier les reads alignés sur chaque locus, pour calculer les abondances de chaque miARN.

Afin d'automatiser et de paralléliser ces traitements, ces étapes sont implémentées dans un *pipeline* bioinformatique, qui permet d'effectuer ces tâches les unes à la suite des autres pour obtenir un traitement identique et reproductible pour l'ensemble des échantillons analysés. Un pipeline correspond donc à une suite ordonnée d'opérations qui vont successivement traiter et transformer les données brutes, pour finalement générer un ou plusieurs documents résumant l'information extraite, dans un format compréhensible et/ou exploitable pour des analyses supplémentaires.

De nombreux pipelines sont disponibles pour aligner et quantifier les miARNs séquencés, mais comme il n'existe pas encore de consensus concernant la stratégie d'alignement des miARNs, ils diffèrent par rapport à la librairie de référence, l'outil d'alignement et les paramètres utilisés, ainsi que le format d'annotation des alignements. Quand j'ai débuté ce travail de recherche en 2016, il existait déjà une quarantaine de pipelines d'alignements pour les miARNs faisant l'objet d'une publication. Certains se distinguent parce qu'ils offrent des analyses particulières, telles que:

— l'identification de miARNs de novo: miRDeep2 [78] ;
— l'identification d'autres petits ARNs non codants: sRNAnalyzer [294] ou sRNAbench [23] ;
— l'analyse simultanée de plusieurs échantillons: miRge [20].

Ces pipelines implémentent donc des stratégies d'alignement différentes, permettant généralement de détecter les isomiRs en minimisant les alignements ambigus. Les trois principales sont les suivantes.

Faire confiance au programme d'alignement La stratégie la plus simple est de faire confiance au programme d'alignement utilisé (ex: *bowtie*) pour miniser les alignements ambigus en rapportant l'alignement le plus probable. Cela est rendu possible en utilisant certains paramètres avancés des outils d'alignement, qui permettent d'attribuer un score à chaque alignement, reflétant notamment le nombre de mismatchs, et de rapporter l'alignement avec le meilleur score dans le cas d'alignements ambigus. Cependant, les scores des outils d'alignement utilisés ont généralement été développés pour l'alignement de longs fragments d'ARN ou d'ADN, et ces scores ne se transposent pas forcément à l'alignement de miARNs. Car l'alignement avec le meilleur score est celui qui possède le moins de mismatchs, mais pour l'alignement des miARNs on peut trouver des exemples où ce n'est pas le cas:

— Si un read s'aligne sur un premier locus avec un mismatch interne à la séquence, et sur un second locus avec deux mismatchs sur l'extrémité 3', alors on peut justifier que le

second alignement est plus probable, car on observe plus fréquemment des isomiRs avec du tailing en 3' que des polymiRs ou des évènements d'édition ARN qui pourraient induire un mismatch interne à la séquence.

— Si un read s'aligne sans aucun mismatch sur deux références, mais lorsque l'on compare la séquence du read avec celle des deux miARNs matures sur lesquelles il s'aligne, il apparaît avec un évènement de trimming en 5' par rapport à la première référence, tandis que sa séquence est identique à la seconde référence. Dans un tel cas, les outils d'alignement classiques considèrent les deux alignements équivalents et ne font pas de choix, alors que l'on peut justifier que le second alignement est plus probable, car le trimming en 5' est peu fréquent.

Stratégie du score optimisé pour l'alignement de miARNs Afin de ne pas se reposer sur l'outil d'alignement pour rapporter l'alignement le plus probable dans le cas d'alignements ambigus, il est possible de définir un score optimisé pour les miARNs. Cette stratégie consiste à aligner les reads sans demander à l'outil de rapporter l'alignement le plus probable, puis d'assigner un score optimisé pour les miARNs à chaque alignement, pour finalement sélectionner l'alignement qui possède le meilleur score. Par exemple, le pipeline SeqBuster [217] attribue un ordre de probabilité suivant l'isomiR observé en comparant le read à la séquence de référence du miARN sur lequel il est aligné: si le read est identique à la séquence du miARN sur lequel il est aligné, alors l'alignement reçoit un score de niveau 1, si il apparaît avec un évènement de trimming, il reçoit un score de niveau 2, si il apparaît avec un variant génétique il reçoit un score de niveau 3, etc. Et l'alignement le plus probable est celui avec le score de niveau le moins élevé.

Stratégie de la librairie sur-mesure Une autre approche notable est l'utilisation d'un alignement stringent avec une librairie de référence personnalisée, contenant les séquences des miARNs, ainsi que les séquences des isomiRs pouvant être produits à partir de chaque miARN. Cette méthode permet d'aligner les reads sans mismatchs, ce qui diminue considérablement le nombre d'alignements ambigus. Le pipeline isomiRage [206] implémente ce principe pour identifier les isomiRs avec du trimming et du tailing. Les séquences des miARNs matures dans la librairie de référence sont dupliquées et possède des dizaines de versions avec du tailing de 1, 2 ou 3 nucléotides, soit en 3', soit en 5'. Cependant, les variations internes dues aux variants génétiques et aux évènements d'édition ARN ne sont pas considérées par IsomiRage, voire même la combinaison de plusieurs isoformes, comme celle d'un événement de trimming en 3' suivie d'un événement de tailing en 3'. Le pipeline miRge utilise également ce principe pour détecter une vingtaine de variants génétiques présents dans les miARNs, et contient donc dans sa librairie de référence deux séquences pour chacun de ces polymiRs: une séquence contenant l'allèle de référence et une séquence contenant l'allèle alternatif.

Pondération des alignements ambigus Les stratégies précédemment présentées permettent de réduire le nombre d'alignement multiples, et parfois de résoudre l'ambiguïté en déterminant l'alignement le plus probable. Cependant, il n'est pas toujours possible de prendre une décision. Par exemple si deux alignements ambigus ont le même score d'alignement. Dans ce cas, une étape de normalisation peut être appliquée, pour qu'un read aligné de façon multiple n'ai pas plus de poids qu'un read qui s'aligne de façon unique. Ainsi si un read s'aligne sur n loci, les n alignements sont conservés, mais un poids de $1/n$ est appliqué à chaque alignement.

d) Résumé

Ces travaux préliminaires m'ont d'abord permis d'évaluer l'effet d'un alignement stringent (sans autoriser de mismatchs) ou permissif (en autorisant des mismatchs) sur l'alignement des miARNs. La table III.2 résume les différents isomiRs qui peuvent être détectés avec un alignement permissif ou stringent, et l'effet sur le nombre de faux positifs et faux négatifs obtenus. Trois types d'isomiRs possèdent des différences par rapport à la séquence de référence et ne sont pas détectés par un alignement stringent: les isomiRs avec NTA, les isomiRs avec édition ARN, et les polymiRs. Ce défaut entraîne un nombre de faux négatifs élevé avec un alignements stringent, mais il permet d'obtenir un faible nombre de faux positifs. A l'inverse, l'alignement permissif permet de détecter n'importe quel isomiR, si il est assez permissif, c'est à dire en autorisant assez de mismatchs, et donc d'obtenir un nombre de faux négatif pratiquement nul. Cependant, l'alignement permissif entraîne un nombre très élevé de faux positifs, notamment à cause du nombre important d'alignements ambigus, et dans une moindre mesure à cause de l'alignement de séquences ne correspondant pas à des miARNs pouvant potentiellement s'aligner sur des miARNs.

Ensuite, l'analyse des stratégies d'alignement de miARNs existantes m'a permis de trouver des moyens de réduire le nombre d'alignements ambigus, afin de réduire le nombre de faux positifs.

	Alignement Permissif	**Alignement Stringent**
miARNs de référence	Aligné	Aligné
Modifications sur les extrémités:		
· isomiRs (trimming)	Aligné	Aligné
· isomiRs (tailing templated)	Aligné [(a)]	Aligné [(a)]
· isomiRs (tailing NTA)	Aligné [(a)]	Non aligné
Modifications sur la séquence centrale:		
· isomiRs (édition ARN)	Aligné	Non aligné
· polymiRs	Aligné	Non aligné
Combinaison de modifications	Aligné	Parfois aligné [(b)]
Faux Positifs (*)	Très élevé	Faible
Faux Négatifs (**)	Très faible ou Nul[(c)]	Élevé

TABLE III.2 – Effet d'un alignement permissif ou stringent sur la détection des miARNs et sur le nombre de faux négatifs et faux positifs.
(*) Les faux positifs sont principalement dus aux alignements ambigus. (**) Les faux négatifs correspondent au rejet d'isomiRs. [(a)] Si la librairie de référence ou l'outil d'alignement permet d'aligner une séquence plus longue que la séquence de référence du miARN. [(b)] Dépend des combinaisons. [(c)] Suivant le degré de permissivité.

2 Méthodes: La stratégie d'optimiR

Le pipeline que j'ai développé dans le cadre de ce projet de thèse, nommé optimiR, a fait l'objet d'une publication dans la revue *RNA* sous le titre: " *OPTIMIR, a novel algorithm for integrating available genome-wide genotype data into miRNA sequence alignment analysis*" . La stratégie d'alignement implémentée dans ce pipeline repose sur les principes suivants:

— L'intégration de l'information génétique pour l'alignement des polymiRs ;
— L'utilisation d'un algorithme d'alignement local, permettant d'obtenir à la fois un alignement stringent sur la partie centrale de la séquence, et un alignement permissif

sur les extrémités ;
— Une résolution des alignements ambigus avec un score optimisé pour les miARNs, et une pondération des alignements ambigus non résolus.

Tous les types d'isomiRs peuvent être détectés par cette stratégie, à l'exception des évènements d'édition ARN, qui sont difficiles à distinguer du bruit de séquençage dans nos échantillons. Cette stratégie permet ainsi d'obtenir un faible nombre de faux négatifs, et la combinaison d'un alignement local stringent avec la résolution d'alignements ambigus permet d'obtenir également un faible nombre de faux positifs.

2.1 Intégration de l'information génétique: détection des polymiRs

La principale originalité d'optimiR repose sur l'intégration de l'information génétique, qui est de plus en plus souvent disponible, grâce à la réduction du coût des puces de séquençage ADN et au développement du séquençage complet du génome. J'ai ainsi développé une méthode d'alignement des miARNs qui prend en compte cette information génétique lors de l'alignement pour détecter les polymiRs.

Une liste de variants génétiques et les génotypes associés pour chaque échantillon analysé peut être fournie dans un fichier au format vcf [5]. Dans ce cas, une nouvelle librairie de référence est créée à partir des séquences de miARNs matures fournies par la miRBase: les séquences de miARNs contenant un des variants génétiques fournis sont dupliquées, la première séquence étant celle d'origine, contenant l'allèle de référence du variant, tandis que la seconde séquence contient l'allèle alternatif. Je me suis ainsi inspiré de la "stratégie de la librairie sur mesure", mais avec optimiR, la création de la nouvelle librairie est dynamique, c'est à dire qu'elle est générée à chaque utilisation du pipeline, et dépend de la liste de variants fournis.

Pour chaque polymiR, le nombre de reads alignés sur chacune des séquences, celle contenant l'allèle de référence ou celle contenant l'allèle alternatif, est ensuite comparé avec le génotype pour vérifier la consistance de l'alignement. Pour les génotypes hétérozygote, l'expression de chaque allèle peut être quantifiée, afin d'estimer l'impact d'un variant sur l'expression du miARN.

Cette méthode permet donc un alignement personnalisé pour chaque échantillon à partir des variants et des génotypes fournis.

2.2 Alignement local: détection des évènements de tailing

La seconde originalité est l'utilisation d'un algorithme d'alignement local, qui permet de forcer un nombre nul de mismatch sur une sous partie du read, appelée *seed*. Cet algorithme permet ainsi un alignement permissif hors de la seed, c'est à dire sur les extrémités du read, et un alignement stringent au niveau de la seed, c'est à dire au centre de la séquence. L'alignement local permet également d'aligner des reads qui sont plus longs que la séquence de référence, et permet ainsi d'aligner des isomiRs avec tailing sur les séquences de miARNs matures.

Plusieurs outils implémentent cet algorithme, qui a été développé à l'origine pour aligner des reads contenant de grande insertions ou délétions, ou pour aligner des fragments d'ARN contenant la séquence limitrophe entre deux exons, voire même pour aligner des reads dont la

5. Le format vcf est largement adopté pour lister un ensemble de variants génétiques, auxquels peuvent être associés des annotations et les génotypes d'une liste d'échantillons

séquence de l'adaptateur n'a pas été retirée. Dans tous ces cas, une ou deux des extrémités du read ne correspondra pas à la séquence de référence. Donc seule une sous-partie du read, la seed, est utilisée pour l'alignement contre la référence. Lorsqu'un alignement avec la seed est trouvé, il est propagé vers chaque extrémité du read [6]. Lorsque les nucléotides aux extrémités ne correspondent plus à la référence, ou parce qu'ils dépassent la référence, ils sont alors *soft-clippés*. Les nucléotides soft-clippés ne sont pas pris en compte par l'outil pour qualifier la qualité de l'alignement, contrairement aux mismatchs. De plus, contrairement aux alignements permissifs basé sur les mismatchs, il n'y a pas besoin de donner une limite aux nombre de bases mésappariées, qui sont ici soft-clippées.

Pour optimiR, j'ai utilisé l'outil *bowtie2* [161] et paramétré l'alignement local de façon à définir une seed de 17 nucléotides. Comme la seed doit être parfaitement alignée et qu'elle est de taille suffisamment grande, cette stratégie permet de distinguer la plupart des miARNs paralogues, et les nucléotides soft-clippés correspondent aux événements de tailing sur les extrémités (Figure III.3).

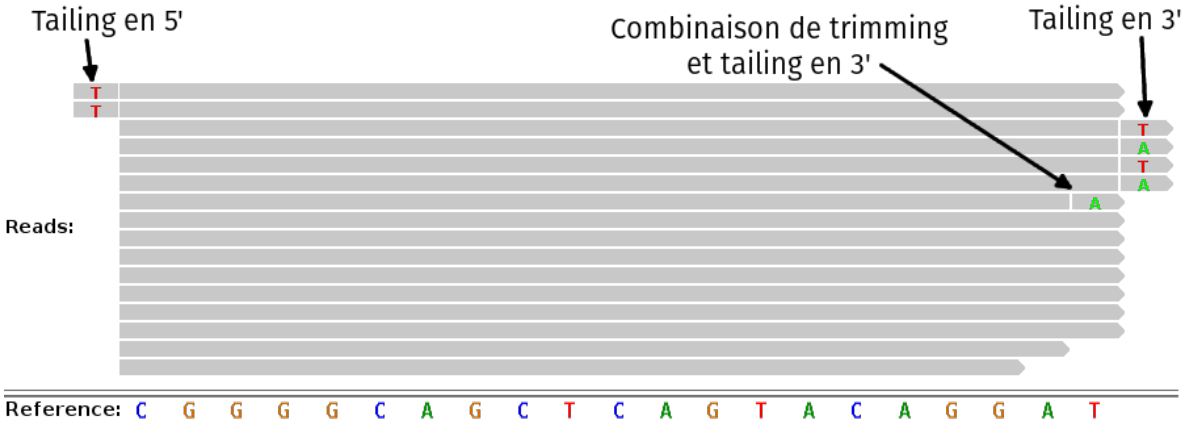


Fig. III.3 – Exemple d'alignement local. Les nucléotides qui diffèrent de la séquence de référence ou qui la dépassent sont soft-clippés.

2.3 Résolution de l'ambiguïté des alignement multiples

Malgré une stratégie d'alignement relativement stringente, des alignements multiples peuvent subsister. Afin de résoudre l'ambiguïté de ces alignements, j'ai développé un score très simple qui permet de sélectionner en priorité l'alignement qui a subit le moins d'évènement de trimming ou de tailing. Chaque événement observé sur l'extrémité 3' pénalise l'alignement d'un point, tandis que les événements en 5', plus rares, pénalisent l'alignement de 4 points (Figure III.4). L'alignement avec le score le plus faible est conservé. Dans le cas où n alignements reçoivent un score minimal, ils sont alors conservés, chacun avec une pondération de 1/n.

En appliquant OptimiR aux jeux de données utilisés dans la section III.1.3.a, on observe que le pourcentage de reads alignés pour les échantillons A et B sont plus élevés qu'avec la

6. On peut d'ailleurs noter que cette stratégie est similaire à celle que les miARNs emploient pour trouver une interaction avec un ARNm.

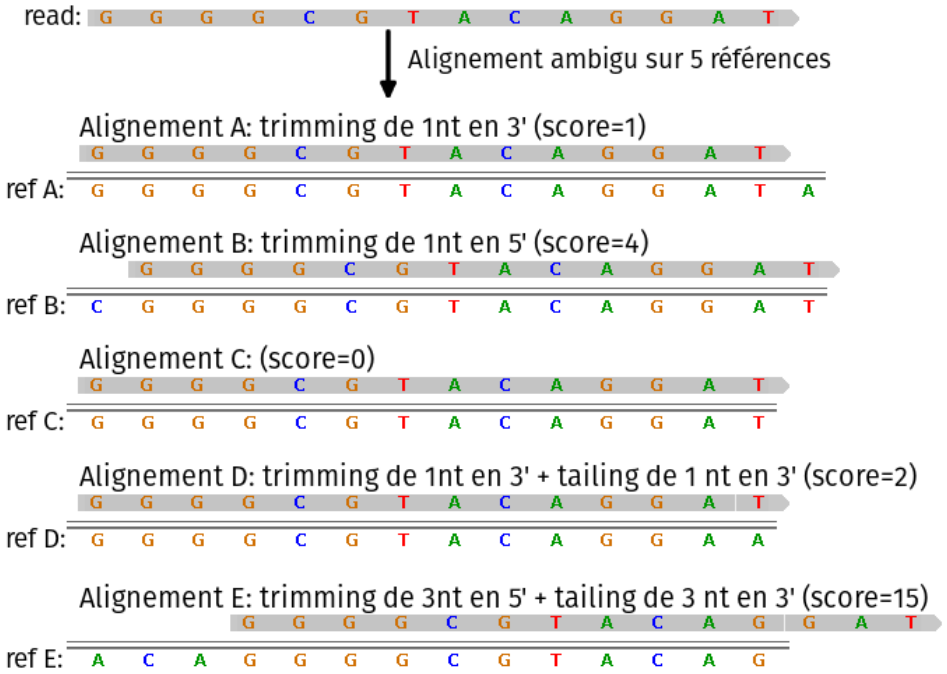


Fig. III.4 – Exemple de résolution d'un alignment multiple. L'alignement C possède le score le plus faible, il est donc conservé, tandis que les autres sont supprimés.

stratégie d'alignement naïve, avec un pourcentage d'alignement ambigus comparable à celui observé lorsqu'aucun mismatch n'est autorisé (voir Table III.3).

Le pourcentage d'alignements multiples avec le jeu miRBase (qui est donc aligné contre lui-même) est égal à 1.51%, et reflète le nombre de séquences de miARNs parfaitement identiques car la résolution par score n'a pas permis de résoudre ces ambiguïtés. Ainsi les séquences identiques sont alignées à la fois sur elles-même, mais aussi sur les autres séquences identiques. Dans ce cas de figure, l'utilisation de la pondération permet de ramener le compte de chaque miARN à 1 read aligné par séquence. Car si n reads sont identiques, et s'alignent sur n séquences identiques, alors chacun de ces alignements reçoit un poids de $1/n$, donc chaque référence possède $n * 1/n$ reads alignés, soit 1 read aligné. Ainsi, en alignant les séquences de la miRBase contre elles-mêmes, chaque séquence de la librairie possède exactement 1 read aligné avec optimiR.

	miRBase	Shuffled	ÉchantillonA	ÉchantillonB
Temps	2”	1”	31”	24”
% de reads alignés	100.00%	0.00%	8.72%	7.31%
% de reads alignés de façon unique	98.49%	x	99.71%	99.73%
% de reads alignés de façon ambiguë	1.51%	x	0.29%	0.27%
Nombre total d'alignements	2663	x	1,304,070	860,647

TABLE III.3 – Application d'optimiR aux jeux de données utilisés dans la section III.1.3.a

3 Article: *OPTIMIR, a novel algorithm for integrating available genome-wide genotype data into miRNA sequence alignment analysis*

BIOINFORMATICS

OPTIMIR, a novel algorithm for integrating available genome-wide genotype data into miRNA sequence alignment analysis

FLORIAN THIBORD,[1,2,3] CLAIRE PERRET,[1,2] MAGUELONNE ROUX,[1,2] PIERRE SUCHON,[4,5] MARINE GERMAIN,[1,2,3] JEAN-FRANÇOIS DELEUZE,[6,7] PIERRE-EMMANUEL MORANGE,[4,5] and DAVID-ALEXANDRE TRÉGOUËT,[1,2,3] on behalf of the GENMED CONSORTIUM

[1]Sorbonne Universités, Université Pierre et Marie Curie (UPMC Univ Paris 06), Institut National pour la Santé et la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1166, Team Genomics and Pathophysiology of Cardiovascular Diseases, 75013 Paris, France

[2]Institute for Cardiometabolism and Nutrition (ICAN), 75013 Paris, France

[3]INSERM UMR_S 1219, Bordeaux Population Health Research Center, University of Bordeaux, 33076 Bordeaux, France

[4]Laboratory of Haematology, La Timone Hospital, 13885 Marseille, France

[5]Institut National pour la Santé et la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1062, Nutrition Obesity and Risk of Thrombosis, Center for CardioVascular and Nutrition Research (C2VN), Aix-Marseille University, 13885 Marseille, France

[6]Centre National de Recherche en Génomique Humaine, Direction de la Recherche Fondamentale, CEA, 91057 Evry, France

[7]CEPH, Fondation Jean Dausset, 75011 Paris, France

ABSTRACT

Next-generation sequencing is an increasingly popular and efficient approach to characterize the full set of microRNAs (miRNAs) present in human biosamples. MiRNAs' detection and quantification still remain a challenge as they can undergo different posttranscriptional modifications and might harbor genetic variations (polymiRs) that may impact on the alignment step. We present a novel algorithm, OPTIMIR, that incorporates biological knowledge on miRNA editing and genome-wide genotype data available in the processed samples to improve alignment accuracy. OPTIMIR was applied to 391 human plasma samples that had been typed with genome-wide genotyping arrays. OPTIMIR was able to detect genotyping errors, suggested the existence of novel miRNAs and highlighted the allelic imbalance expression of polymiRs in heterozygous carriers. OPTIMIR is written in python, and freely available on the GENMED website (http://www.genmed.fr/index.php/fr/) and on Github (github.com/FlorianThibord/OptimiR).

Keywords: microRNA; next-generation sequencing; alignment; genetic variations; isomiRs

INTRODUCTION

With an average length of 22 nucleotides (nt), microRNAs (miRNAs) belong to a class of small noncoding RNAs known to regulate gene expression by binding messenger RNAs (mRNAs) and interfering with the translational machinery (Bartel 2004; Filipowicz et al. 2008). MiRNAs are transcribed from primary miRNA sequences (pri-miRNAs) and fold into a hairpin-like structure, which is sequentially processed by two ribonucleases, DROSHA and DICER. The former cleaves the pri-miRNA into a pre-miRNA and the latter completes the miRNA's maturation by cleaving the pre-miRNA near its loop to produce a miRNA duplex composed of two mature strands (Kim et al. 2009). Exceptionally, some miRNAs follow a slightly different pathway where only one ribonuclease is needed to complete the maturation (Kim et al. 2016). In any case, only one of the two mature strands is loaded in an effective protein complex called RISC, while the other is degraded (Kawamata and Tomari 2010). This selection seems mostly driven by the thermodynamic stability of both ends forming the duplex (Meijer et al. 2014).

There is emerging interest in performing miRNA profiling in body fluids or tissues in order to identify novel molecular determinants of human diseases (Mitchell et al. 2008; Pulcrano-Nicolas et al. 2018). Such miRNA profiling can

Corresponding author: david-alexandre.tregouet@inserm.fr
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.069708.118. Freely available online through the *RNA* Open Access option.

be achieved using hybridization (microarray), next-generation sequencing (NGS), or real time-quantitative polymerase chain reaction (RT-qPCR) techniques. With 2588 known mature miRNAs in humans according to miRBase version 21 (Kozomara and Griffiths-Jones 2014), RT-qPCR would be cumbersome on a genomic scale, but is widely recognized as a gold standard for the validation of few miRNAs. The NGS technology is becoming more popular than microarrays because of its greater detection sensitivity, and higher accuracy in differential expression analysis (Git et al. 2010; Tam et al. 2014). NGS applied to small RNAs revealed a great diversity in the sequences of mature miRNAs originating from the same hairpin. This diversity is mostly attributable to the deletion and addition of nucleotides at the miRNAs' extremities (also known as trimming and tailing events, respectively), due to the activity of terminal nucleotidyl transferases, exoribonucleases, or imprecise cleavage by DROSHA and DICER (Wyman et al. 2011; Neilsen et al. 2012; Ameres and Zamore 2013). To a lesser extent, the ADAR protein acting on double stranded RNAs and responsible for A-to-I editing is also known to target miRNAs (Nishikura 2016). These post-transcriptional editing mechanisms have been shown to affect miRNAs' function and stability (Kawahara et al. 2007; Burroughs et al. 2010; Chiang et al. 2010; Katoh et al. 2015). Lastly, genetic variations have also been shown to contribute to the sequence diversity of miRNAs, and to affect their function and expression (Mencía et al. 2009; Gong et al. 2012; Han and Zheng 2013; Cammaerts et al. 2015). MiRNAs subject to post-transcriptional events and/or genetic variations are generally referred to as isomiRs. In the following, we will use the expression "polymiR" to refer to the subclass of isomiRs harboring genetic polymorphisms in their miRNA sequence.

The first step in the bioinformatics analysis of miRNA sequencing (miRSeq) data consists in aligning sequenced reads to a reference library of mature miRNAs. This step may be challenging because (i) the aforementioned variability of isomiRs could lead to imperfect alignments to the reference library; (ii) sequenced reads may correspond to (fragments of) other molecules (e.g., other small noncoding RNAs like piRNA, tRNA, yRNA…), captured during the preparation of the libraries, that might share a high similarity with miRNAs because of their small length and thus might be confused with miRNAs (Chen and Heard 2013; Heintz-Buschart et al. 2018); (iii) some miRNAs have homologous sequences that are identical or very similar, thus a single read might align ambiguously to multiple reference sequences. In this work, we investigate the impact of the presence of polymorphisms in the sequence of mature miRNAs on their alignment and their expression in the context of miRSeq profiling applied to samples that have also been typed for genome-wide genotype data. This is a situation we anticipate to become rather common with the rise of increasingly affordable genome-wide association studies (GWAS) and the decreasing cost of next-generation exome/genome sequencing techniques. In that context, we developed an original bioinformatics workflow called OPTIMIR, for pOlymorPhism inTegratIon for MIRna data alignment, that integrates genetic information from genotyping arrays or DNA sequencing into the miRSeq data alignment process with the aim of improving the accuracy of polymiRs alignment, while accommodating for other isomiRs detection and ambiguously aligned reads. In addition, OPTIMIR allows to assess the association of genotypes on polymiRs with corresponding polymiRs' expression. OPTIMIR was evaluated in the plasma samples of 391 individuals, part of the MARTHA study (Oudot-Mellakh et al. 2012).

RESULTS

OPTIMIR is composed of three main steps (Materials and Methods). First, miRSeq data are aligned to a reference library upgraded with sequences integrating alternative alleles of genetic variations. A correction is then applied for ambiguous and unreliable alignments via a scoring approach. Finally, polymiR alignments are evaluated to only retain those that are consistent with input genotypes in case these have been provided by the user. The general workflow was summarized in Figure 1.

MiRNA alignment

OPTIMIR was evaluated on 391 miRNA sequencing data files totaling 7,390,947,662 sequencing reads. After preprocessing the sequenced reads, that included adapters' removal and selecting only reads with size ranging between 15 and 27 nt, 2,922,446,965 reads (39.54% of total reads) remained for alignment. 562,040,494 of these reads (19.23%) were then mapped to mature miRNA reference sequences, of which 10,937,479 (1.95% of mapped reads) aligned ambiguously to two sequences or more. The application of the OPTIMIR scoring algorithm for alignment disambiguation resulted in a unique solution for 91.6% of these cross-mapping reads. OPTIMIR computes a score based on the number of editing events that a sequenced read could be compatible with and keeps alignments with the lowest scores (i.e., the lowest number of editing events). For example, if a given read aligns to two reference sequences, perfectly on the first one (score of zero), but with a missing base on the 3′ end (score of one) on the second, then only the first alignment is kept. A score of zero corresponds to a perfect match, and each editing event adds penalties to the score. Modifications in the 5′ end are more penalizing, as they are less frequently observed. If a cross-mapping read receives an equal score on different alignments, then its weight is divided accordingly to the number of equivalent alignments (see Materials and Methods). For 89.5% of reads with multiple alignments,

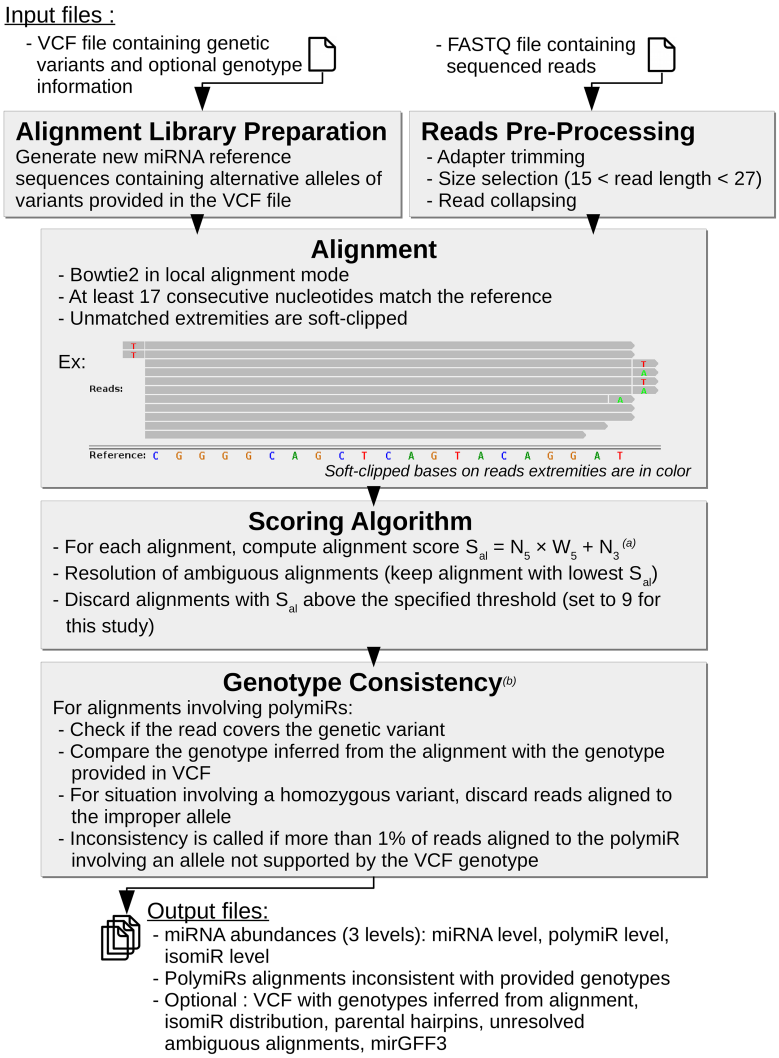


FIGURE 1. OPTIMIR workflow. From an optional VCF file and a FASTQ file, OPTIMIR performs the alignment of miRSeq data and results in the generation of abundances files containing the expressions of sequenced miRNAs. (*a*) The score S_{al} is based on the number of editing events observed on the 5′ end and 3′ end of a read (N_5 and N_3, respectively), with a weight W_5 applied to the 5′ end events (set to four in this study). (*b*) The genotype consistency is checked only if the user provided genotypes for the variants in the VCF file.

the difference between the two lowest scores was greater than two (Fig. 2A) which would correspond to alignments that differ from each other by at least two modifications in the 3′ end.

After alignment disambiguation, scores ranged from zero to 76 with 98.0% of alignment scores lower or equal to nine. Beyond this threshold, the quantile distribution curve rapidly increased indicating that scores with higher values are very sparse and suggesting that such alignments with a very high number of editing events are likely improper alignments (Fig. 2B). As a consequence, for the following, we decided to discard any alignment with a score greater than nine. Among the 550,946,055 remaining reads, more than 40% were perfectly aligned or involved templated additions, which were not penalized as these nucleotides are present in the parental hairpin sequence. An additional 40% of alignments involved reads with a single event in the 3′ end (see Fig. 2C).

IsomiRs distribution

A total of 197,808,779 (35.9%) reads perfectly aligned to mature miRNAs; such reads being generally referred to as canonical miRNAs, and received an alignment score of zero. This confirms previous observations suggesting that a substantial amount of miRNAs are mainly represented by alternative isomiRs (Wallaert et al. 2017; Wu et al. 2018).

The most common observed editing events were on miRNAs' 3′ end, with ∼34% of trimming, ∼17% tailing with nontemplated nucleotides, ∼5% tailing with templated nucleotides, and a similar proportion of trimming events followed by tailing. The latter modification could also be interpreted as nucleotide variation due to genetic variants, or other less frequent post-transcriptional editing events such as A-to-I editing. It should also be noted that library preparation and sequencing could also contribute to a significant amount of fragment modifications that are falsely detected as isomiRs (Wright et al. 2019).

The 5′ end was much less frequently edited, with 94% of reads having no editing on this extremity. Nevertheless, the most frequently observed modification on 5′ ends was trimming that affected 4.4% of all reads. This may be of biological relevance since such trimming could shift the miRNA binding seed that is crucial for the miRNA to bind to its mRNA targets.

The distribution of 3′ and 5′ ends modifications on mapped reads observed over the 391 samples processed by OPTIMIR is shown in Figure 3.

Alignment on polymiRs

Over all samples, 220,156 reads mapped to 46 polymiRs for a total of 1786 distinct alignments. As detailed in Table 1, 19 polymiRs have reads that aligned to an

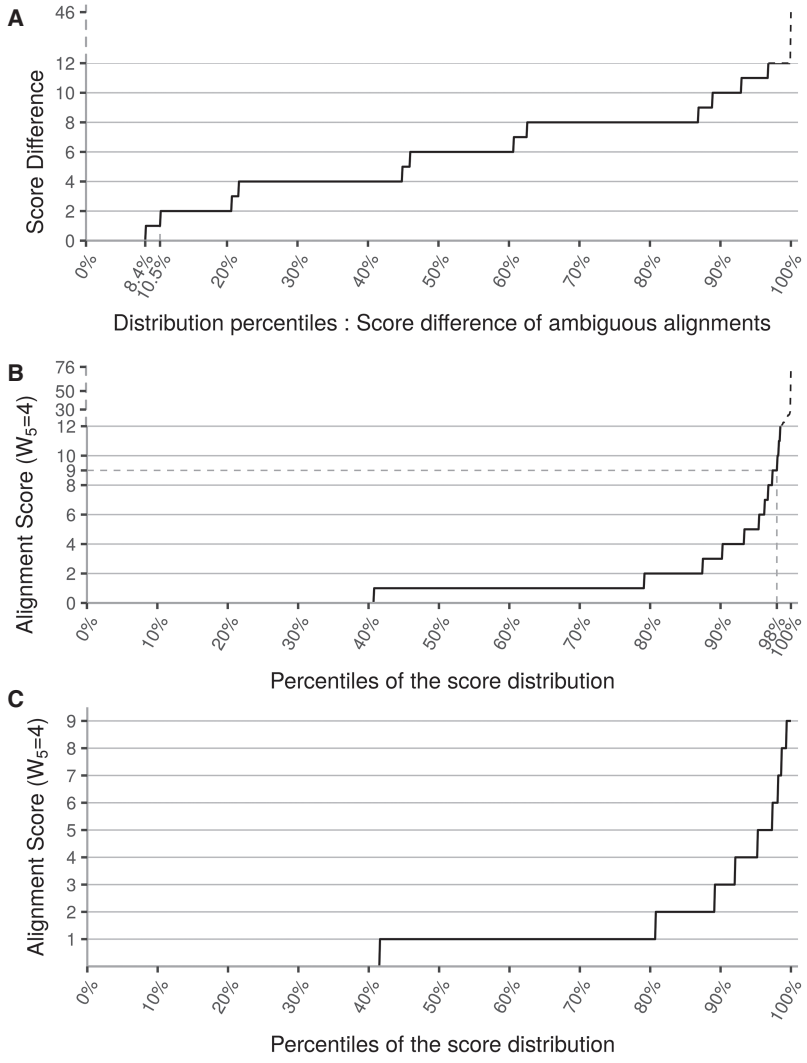


FIGURE 2. Influence of the scoring algorithm. (*A*) This figure represents the percentile distribution of the score differences between the two best alignments of each cross-mapping read. A total of 8.4% of cross-mapping reads have a score difference of zero, which means they aligned on different sequences with the same score. These alignments could not be disambiguated. W5 indicates the penalty weight for events in the 5′ end, fixed to four in this study. (*B*) The alignment score percentile distribution after cross-mapping reads disambiguation represented in percentiles. A total of 98% of alignments have a score lower or equal to nine. The remaining 2% of alignments have a score ranging from 10 to 76, which were categorized as unreliable isomiRs because of the unlikely number of editing events they would have undergone. (*C*) The same distribution as in *B* but after removing alignments with a score higher than nine.

alternative sequence that was introduced in the upgraded reference library. Two polymiRs (hsa-miR-6796-3p and hsa-miR-1269b) harbor two SNPs, and for both of them only the reference sequence with both common alleles were found to be expressed. Among the remaining 44 polymiRs that harbor only one SNP, 15 were expressed with both alleles. It is important to mention that the allele present in the miRBase reference sequence may not be the most common one (e.g., the rs2155248 and hsa-miR-1304-3p) which may lead to improperly discarding of reads if stringent alignment (with no mismatch allowed) to the original miRBase library is applied.

In total, 155 alignments involving 721 reads (0.33% of reads mapped to polymiRs) were found inconsistent with the individual genotypes, among which 145 alignments were supported by less than five reads. These alignments were discarded and not further discussed as they most probably correspond to sporadic mechanisms that mimic genetic variations such as modifications induced during library preparation or sequencing (Wright et al. 2019), or uncommon posttranscriptional editing events (see cMaterials and Methods). The remaining 10 alignments, involving 507 reads spread over 10 samples and five polymiRs, are detailed in Table 2 and further investigated in the next section.

Investigation of inconsistent genotypes

Situations where a substantial number of reads aligned inconsistently on a polymiR were reported by OPTIMIR to allow for further investigations.

The first case of inconsistent genotype concerns individual PVP28 imputed to be homozygous for the rs12473206-G allele while showing numerous reads mapping to both versions of the hsa-miR-4433b-3p polymiR. Sanger resequencing revealed that this individual is in fact heterozygous for this variant which is much more compatible with the number of observed reads at this locus. A rather similar inconsistent genotype was observed for individual SSP20 but lack of available DNA did not allow us to perform Sanger validation. Lack of DNA also prevented us from investigating deeper the inconsistent genotypes observed for rs72631820/hsa-miR-339-3p and rs6771809/hsa-miR-6826-5p.

The inconsistent genotype observed for individual AJA9 at rs6841938 and hsa-miR-1255-5p was challenging. Sanger resequencing confirmed that this individual was indeed homozygous for the rs6841938-G allele although it has numerous reads mapping to both versions of hsa-miR-1255-5p. However, the use of BLAST web-service

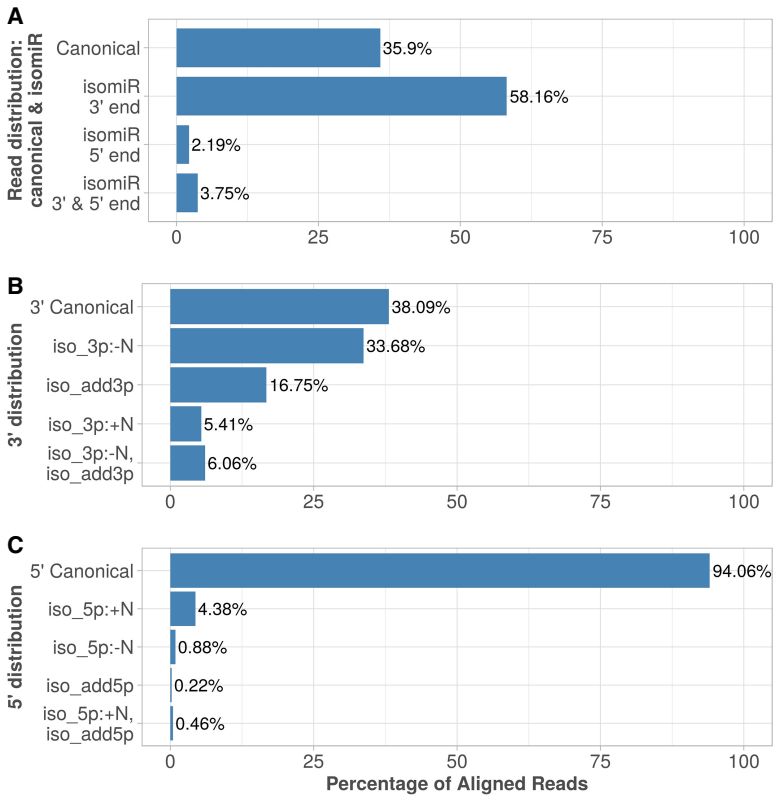


FIGURE 3. IsomiRs profiling. (*A*) Reads aligned by OPTIMIR distributed between: canonical miRNA (identical to the reference), isomiR 3′ end/5′ end (modifications due to trimming or tailing observed on the 3′ or 5′ end, respectively), and isomiR with both ends edited. (*B*) Distribution of variations observed on the 3′ end of aligned miRNAs: 3′Canonical (no modification observed); iso_3p: −N (trimming); iso_3p:+N (templated tailing); iso_add3p (nontemplated tailing); iso_3p: −N,iso_add3p (combination of trimming and nontemplated tailing). (*C*) Distribution of variations observed on the 5′ end of aligned miRNAs: 5′Canonical (no modification observed); iso_5p:+N (trimming); iso_5p:−N (templated tailing); iso_add5p (nontemplated tailing); iso_5p:+N,iso_add5p (combination of trimming and nontemplated tailing).

from NCBI (Altschul et al. 1990) revealed that reads aligned on the polymiR sequence containing the rs6841938-A allele perfectly match to the *chr1:167,998,699-167, 998,720* region but on the opposite strand where hsa-miR-1255-3p should be located. This observation could be compatible with the presence on this opposite strand of an unreported miRNA locus as has already been observed for other miRNAs (e.g., hsa-mir-4433a and hsa-mir-4433b). Going back to the discarded alignments revealed that 18 other samples homozygous for rs6841938-G had one or two reads that mapped to this opposite strand. These alignments had been discarded since they were supported by less than five reads (see *Genotype consistency analysis* section in the Materials and Methods).

The last inconsistent genotypes were observed at rs3817551/hsa-miR-7107-3p for five independent individuals. For four of them with available DNA, Sanger resequencing confirmed the initial homozygous genotype for rs3817551-G allele while these individuals were found to express only the hsa-miR-7107-3p version with the T allele. On average, these inconsistent alignments received an alignment score of 1.41, and none of the reads involved mapped to other sequences. Concerning the samples that had reads aligned on this same reference with a consistent genotype, the average alignment score was very close to 1.47. This score indicates that reads share a high sequence similarity with the hsa-miR-7107-3p, and there is no significant difference between the group with an inconsistent genotype, and the group with a consistent one. BLAST analysis did not enable us to identify any homologous sequence that could explain these observations that still remain to be further investigated in order to be sure that associated reads do originate from hsa-miR-7107-3p.

Analysis of polymiRs allele-specific expression

As shown in Table 1, 29 polymiRs with a SNP in their mature sequence were found to be expressed in individuals heterozygotes for this SNP. While one could anticipate that, in heterozygous individuals for such SNP, the polymiRs could have balanced expression (as measured by read counts) of both the reference and alternative sequences, this was hardly observed. Indeed, we observed a strong preference for either the reference or the alternative version of the polymiR in heterozygotes.

Figure 4 shows the alignments of 4 polymiRs involving heterozygous variants according to the vcf genotype. The dotted line represents a theoretical balanced expression between reference and alternative sequences. We can see that, for hsa-miR-1255b-5p/rs6841938 and hsa-miR-5189-3p/rs35613341, dots are close to the y-axis, which are situations where only the reference allele is expressed. The polymiR with the expression closest to allelic balance for many samples is hsa-miR-4433b-3p but, even then, the average rate of alternative reads across all 147 samples is of 0.8 (see Supplemental Table 4 for details on polymiRs involved in heterozygous situations). As the last case, 65 MARTHA individuals were heterozygous for the rs2925980 variant associated with polymiR hsa-miR-7854-3p. These genotypes could be considered as reliable as this SNP was directly typed on the array. Among these

TABLE 1. Summary information at detected polymiRs

polymiR	rsID	Number of samples expressing the polymiR with genotype[a]			Number of reads aligned on polymiR sequence integrating the allele		Number of inconsistent reads removed on the sequence integrating the allele		Inconsistent alignments to investigate[b]
		0/0	0/1	1/1	Reference	Alternative	Reference	Alternative	
hsa-miR-4433b-3p	rs12473206 (C/G)	223 (/230)	147 (/147)	14 (/14)	63,762	96,915	4	475	2
hsa-miR-1255b-5p	rs6841938 (G/A)	244 (/323)	48 (/66)	2 (/2)	29,642	66	0	52	1
hsa-miR-1304-3p	rs2155248 (G/T)	0 (/0)	5 (/7)	256 (/384)	146	11,928	4	0	0
hsa-miR-339-3p	rs72631820 (T/C)	213 (/386)	3 (/5)	0 (/0)	6293	0	0	21	1
hsa-miR-7854-3p	rs2925980 (A/G)	22 (/159)	65 (/167)	35 (/65)	392	2015	1	4	0
hsa-miR-4745-5p	rs10422347 (C/T)	86 (/300)	14 (/81)	0 (/10)	2330	15	0	0	0
hsa-miR-5189-3p	rs35613341 (C/G)	42 (/196)	20 (/161)	0 (/34)	1315	0	2	1	0
hsa-miR-7107-3p	rs3817551 (T/G)	20 (/157)	33 (/177)	0 (/57)	1104	1	122	0	5
hsa-miR-4741	rs7227168 (C/T)	40 (/293)	4 (/93)	0 (/5)	695	0	0	0	0
hsa-miR-548l	rs13447640 (G/A)	30 (/373)	0 (/18)	0 (/0)	613	0	0	0	0
hsa-miR-3620-5p	rs2070960 (C/T)	47 (/348)	4 (/42)	0 (/1)	459	0	0	0	0
hsa-miR-6826-5p	rs6771809 (T/C)	11 (/286)	11 (/92)	2 (/13)	199	243	0	21	1
hsa-miR-4638-5p	rs146528803 (G/A)	20 (/369)	1 (/22)	0 (/0)	313	0	0	0	0
hsa-miR-3622a-5p	rs66683138 (G/A)	6 (/231)	4 (/138)	0 (/22)	241	24	0	0	0
hsa-miR-4707-3p	rs2273626 (C/A)	4 (/86)	6 (/195)	0 (/110)	162	0	0	0	0
hsa-miR-1269a	rs73239138 (G/A)	5 (/236)	2 (/131)	0 (/24)	107	31.5[d]	0	0	0
hsa-miR-4781-3p	rs74085143 (G/A)	8 (/366)	0 (/25)	0 (/0)	135	0	0	0	0
hsa-miR-6763-3p	rs3751304 (C/T)	1 (/39)	2 (/153)	2 (/199)	86	29	0	0	0
hsa-miR-4804-5p	rs266435 (C/G)	0 (/7)	3 (/91)	7 (/293)	0	114	1	0	0
hsa-miR-4999-5p	rs72996752 (A/G)	5 (/189)	1 (/163)	0 (/39)	86	26	0	1	0
hsa-miR-6839-5p	rs7804972 (G/A)	1 (/153)	2 (/196)	0 (/42)	56	0	1	0	0
hsa-miR-4459	rs73112689 (C/T)	3 (/245)	1 (/133)	0 (/13)	51	0	0	0	0
hsa-miR-585-3p	rs62376935 (C/T)	0 (/330)	2 (/52)	0 (/9)	2	48	0	0	0
hsa-miR-3928-5p	rs5997893 (A/G)	1 (/38)	1 (/157)	1 (/196)	47	9	0	0	0
hsa-miR-548al	rs515924 (A/G)	0 (/307)	1 (/80)	1 (/4)	0	37	0	0	0
hsa-miR-3117-3p	rs12402181 (G/A)	1 (/288)	0 (/97)	0 (/6)	35	0	0	0	0
hsa-miR-3130-3p	rs2241347 (C/T)	2 (/254)	3 (/119)	0 (/18)	35	0	0	0	0
hsa-miR-4532	rs73177830 (G/A)	2 (/344)	0 (/44)	0 (/3)	34	0	0	0	0
hsa-miR-302c-3p	rs199971565 (del)	1 (/371)	0 (/20)	0 (/0)	30	0	0	0	0
hsa-miR-4772-5p	rs62154973 (C/T)	3 (/339)	0 (/51)	0 (/1)	25	0	0	0	0
hsa-miR-1343-5p	rs2986407 (T/C)	0 (/20)	0 (/111)	2 (/260)	0	24	0	0	0
hsa-miR-548ap-5p	rs4414449 (G/A)	2 (/47)	11 (/180)	1 (/164)	16	4	4	0	0
hsa-miR-6885-5p	rs78293125 (A/G)	1 (/341)	0 (/47)	0 (/3)	15	0	0	0	0
hsa-miR-548h-3p	rs73235381 (C/T)	3 (/379)	0 (/11)	0 (/1)	14	0	0	0	0
hsa-miR-4433a-5p	rs12473206 (C/G)	1 (/230)	2 (/147)	0 (/14)	12	0	0	0	0
hsa-miR-6887-5p	rs1688017 (G/A)	1 (/152)	0 (/193)	0 (/46)	10	0	0	0	0
hsa-miR-4482-5p	rs45596840 (G/A)	1 (/167)	0 (/179)	0 (/45)	10	0	0	0	0
hsa-miR-4520-5p	rs8078913 (C/T)	0 (/87)	0 (/198)	1 (/106)	0	9	0	0	0
hsa-miR-1229-5p	rs2291418 (G/A)	2 (/363)	0 (/27)	0 (/1)	7	0	0	0	0
hsa-miR-6805-3p	rs56312243 (C/T)	1 (/337)	0 (/50)	0 (/4)	5	0	0	0	0
hsa-miR-4302	rs11048315 (G/A)	1 (/278)	2 (/102)	0 (/11)	1	2	0	6	0
hsa-miR-4520-3p	rs8078913 (C/T)	0 (/87)	1 (/198)	0 (/106)	1	0	0	0	0
hsa-miR-548h-5p	rs9913045 (G/A)	0 (/143)	0 (/197)	0 (/51)	0	0	1	0	0
hsa-miR-6863	rs12708966 (G/A)	1 (/385)	0 (/6)	0 (/0)	1	0	0	0	0
hsa-miR-6796-3p	rs3745199 (C/G) rs3745198 (C/G)	3 (/131) 3 (/131)	0 (/193) 0 (/190)	0 (/67) 0 (/70)	56	0[c]	0	0	0

Continued

TABLE 1. *Continued*

polymiR	rsID	Number of samples expressing the polymiR with genotype[a]			Number of reads aligned on polymiR sequence integrating the allele		Number of inconsistent reads removed on the sequence integrating the allele		Inconsistent alignments to investigate[b]
		0/0	0/1	1/1	Reference	Alternative	Reference	Alternative	
hsa-miR-1269b	rs12451747 (A/C) rs7210937 (G/C)	0 (/82) 1 (/326)	1 (/171) 0 (/60)	0 (/138) 0 (/5)	31.5 (d)	0[c]	0	0	0
	TOTAL	1062	400	324	108574.5	111540.5	140	581	10

[a]Between parentheses is the total number of samples carrying the variant with homozygous reference allele (0/0), heterozygous (0/1), or homozygous alternative allele (1/1).
[b]Alignments with a high number of inconsistent reads (higher than 1% of the sum of reads gathered by the polymiR sequences), which is incompatible with sporadic events replicating a genetic variant.
[c]For multivariant polymiRs, reads did not map to any of the sequence integrating alternative haplotypes.
[d]The sequence of hsa-miR-1269a integrating the rs73239138-A allele is identical to the hsa-miR-1269b sequence. Hence, each of the 63 reads that cross-mapped on both sequences received a count of ½ on each sequence.

65 individuals, 54 expressed only the alternative version of polymiR hsa-miR-7854-3p, and the remaining 11 expressed both polymiR versions with a mean ratio of 0.96 in favor of the alternative allele.

Finally, we used the RNAfold program (Lorenz et al. 2011) to predict the secondary structure of the 29 pri-miRNA (i.e., hairpins) induced by the presence of a SNP in the polymiR sequence (Supplemental Fig. S1). Most genetic variations create either a new bulge, or a wobble pairing, or have no impact on the secondary structure. A notable exception relates to rs35613341 located on the hsa-miR-5189-3p where the G allele completely changed the secondary structure of the hairpin (see Supplemental Fig. S1[u]) making it difficult to access for the DICER and DROSHA machinery. In MARTHA samples, 161 individuals were heterozygous and 34 homozygous for the rs35613341-G allele. None of these individuals were found to express the alternative sequence of polymiR hsa-miR-5189-3p which could support the hypothesis that the rs35613341-G allele impacts the maturation of this miRNA.

Lastly, by the completion of the OPTIMIR pipeline (with a scoring threshold of nine as mentioned above and keeping only genotype consistent reads) on MARTHA samples, 7.45% of sequenced reads were aligned. This value had to be compared with 7.68% and 8.24% obtained by two other recent pipelines for miRSeq data, sRNAnalyzer (Wu et al. 2017) and miRge (Baras et al. 2015), respectively, executed using default parameters. These discrepancies could be explained by the different alignment strategies implemented by these tools, which by default allow up to two mismatches for read alignment, while OPTIMIR does not allow any mismatch in aligning sequenced reads and relies on local alignment to capture isomiRs with modified extremities (see OPTIMIR description). When we ran

TABLE 2. Inconsistent genotypes unlikely due to sequencing errors

polymiR	rsID	Alleles: reference, alternative	MAF	RSQ	SAMPLE	Genotype from vcf	Counts reference	Counts alternative	SANGER validation
hsa-miR-4433b-3p	rs12473206	C/G	0.23	0.99	PVP28	C/C	84	*333*	G/C
hsa-miR-4433b-3p	rs12473206	C/G	0.23	0.99	SSP20	C/C	135	*8*	N.A.
hsa-miR-339-3p	rs72631820	T/C	0.009	0.98	CED24	T/T	139	*7*	N.A.
hsa-miR-6826-5p	rs6771809	T/C	0.14	0.98	DTD19	T/T	1	*20*	N.A.
hsa-miR-1255b-5p	rs6841938	G/A	0.09	0.96	AJA9	G/G	92	*28*	G/G
hsa-miR-7107-3p	rs3817551	T/G	0.36	0.96	MSM20	G/G	27	0	N.A.
hsa-miR-7107-3p	rs3817551	T/G	0.36	0.96	MCC20	G/G	15	0	G/G
hsa-miR-7107-3p	rs3817551	T/G	0.36	0.96	NFN22	G/G	23	0	G/G
hsa-miR-7107-3p	rs3817551	T/G	0.36	0.96	NHN19	G/G	25	0	G/G
hsa-miR-7107-3p	rs3817551	T/G	0.36	0.96	GLG20	G/G	17	0	G/G

MAF, minor allele frequency. RSQ, imputation quality criteria (r2).

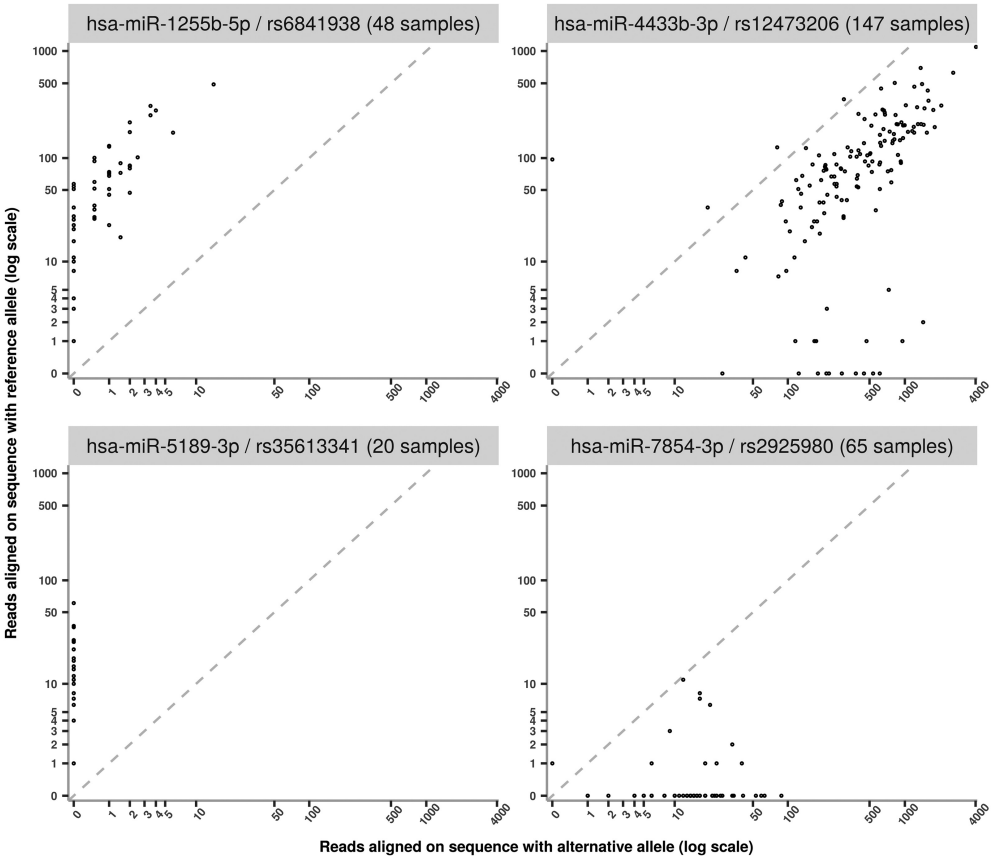


FIGURE 4. Examples of allele imbalanced expressions observed at four polymiRs. On the *x*-axis is shown the expression of the alternative allele, while the *y*-axis represents the expression of the reference allele. Expressions were $\log_{10}$-transformed.

sRNAnalyzer with only one mismatch allowed, the percentage of aligned reads dropped to 7.38% (compared to 7.68% with two mismatches). Of note, it was not possible to modify the number of authorized mismatches in miRge.

The files generated by OPTIMIR include (i) global abundances of miRNAs (counts of isomiRs and polymiRs are merged with the reference mature sequences' counts); (ii) specific abundances for each polymiR sequence; (iii) specific abundances for each isomiR sequence; (iv) alignments that are inconsistent with provided genotypes; (v) two annotation files containing details on templated nucleotides and alignments that could not be disambiguated. In addition, OPTIMIR allows to generate its results into the recently developed mirGFF3 format (Desvignes et al. 2018) aimed at unifying results of any miRSeq data analysis.

DISCUSSION

In this work, we propose a novel algorithm, OPTIMIR, for aligning miRNA sequences obtained from NGS and we applied it to plasma samples of 391 individuals from the MARTHA study. Borrowing some ideas from other alignment pipelines such as the addition of new sequences to the reference library corresponding to allelic versions of polymiRs (Baras et al. 2015; Russell et al. 2018) and a scoring strategy for handling cross-mapping reads (Urgese et al. 2016) or for discarding unlikely isomiRs (Bofill-De Ros et al. 2018), OPTIMIR has two features that make it unique. First, OPTIMIR is based on a scoring strategy that incorporates biological knowledge on miRNA editing to identify the most likely alignment in the presence of cross-mapping reads. Second, OPTIMIR allows the user to provide genotype information, in particular data obtained from genome-wide genotyping arrays, to improve alignment accuracy. This option revealed several interesting observations when OPTIMIR was applied to MARTHA plasma samples.

First, it allowed the identification of improperly imputed genotypes despite overall good imputation quality. Second, it suggested the existence of a new miRNA not indexed in the miRBase v21 database that would be located on the opposite strand of the hsa-miR-1255b-5p. Thirdly, it suggested that reads aligned to hsa-miR-7107-3p are likely false alignments and would more likely come from other noncoding fragments that share sequence similarity with

hsa-miR-7107-3p. These last two hypotheses would need to be further validated but this is out of the scope of the bioinformatics workflow described in the current work. Even more interesting was the study of polymiRs' expressions in heterozygous individuals for SNPs on these polymiRs. OPTIMIR clearly showed that plasma allelic expression of polymiRs is unbalanced for most polymiRs. One allelic version of a polymiR is much more expressed than the other and this is not necessarily the one carrying the most common allele. This observation is consistent with previous works showing that SNPs in (pri-) miRNA sequences can influence miRNA expression through their impact on the DROSHA/DICER RNases' machinery (Duan et al. 2007; Cammaerts et al. 2015). Epigenetic mechanisms could also explain the allelic imbalance expression of miRNAs (Morales et al. 2017).

Several limitations shall however be acknowledged. OPTIMIR requires to fix two parameters, a weight W_5 for penalizing 5′ end editing event (set to four in the current application) and a score threshold (set to nine here) to discard unreliable alignments. The former tends to have little impact on the general findings (see Supplemental Table S2), while the second may be study-specific and may depend on the study sample-size and the kind of tissue analyzed. These parameters can be easily modified by the user. There is so far no gold standard program for miRNA alignment analysis but our preliminary study suggests that OPTIMIR aligns slightly less reads than two other recently proposed software, sRNAnalyzer (Wu et al. 2017) and miRge (Baras et al. 2015). This is likely due to the higher number of mismatches allowed by the latter two for aligning reads while OPTIMIR tends to be more stringent. Without extensive investigations including experimental validations, it is not possible to really appreciate which alignments are the correct ones.

Finally, several improvements could be considered such as (i) the integration of A-to-I editing events in the definition of our reference library and of our scoring strategy, even if we anticipate that it might be difficult to distinguish these rare events from sequencing errors and (ii) the extension of the OPTIMIR workflow to analyze other small coding RNAs (e.g., piRNA and tRNA, rRNA, snRNA, or yRNA derived fragments) that are generally sequenced together with miRNAs in a miRseq profiling. Applications to other tissues from the samples processed for miRSeq data deserve to be conducted to generalize the findings observed in the current plasma samples.

MATERIALS AND METHODS

The MARTHA data set

The MARseille Thrombosis Assocation study is a collection of patients with venous thrombosis (VT) recruited at the La Timone Hospital (Marseille, France) between 1994 and 2005 and aimed at identifying novel molecular determinants for VT and its associated endophenotypes (Oudot-Mellakh et al. 2012; Dick et al. 2014).

For the present study, 391 MARTHA participants with available plasma samples were processed for plasma miRNA profiling through miRSeq. These individuals had been previously typed for genome-wide genotyping arrays and imputed for single nucleotide polymorphisms (SNPs) available in the 1000G reference database (Germain et al. 2015).

MiRNA extraction and preparation followed the same protocol as the one previously described in Roux et al. (2018). Briefly, from 400 µL of plasma, total RNA was first extracted using the miRNeasy serum/plasma kit for Qiagen. MiRNA libraries were then prepared using the NEBNext Multiplex Small RNA Library Preparation Set for adapter ligation and PCR, with adapter sequences GATCGGAAGAGCACACGTCTGAACTCCAGTCAC (3′ adapter) and CGACAGGTTCAGAGTTCTACAGTCCGACGATC (5′ adapter) followed by a size selection using AMPure XP beads. Pools of equal quantity of 24 purified libraries were constructed and tagged with different indexes. Pools were then sequenced using a 75 bp single-end strategy on an Illumina NextSeq500 instrument.

The OPTIMIR workflow

Alignment

Prealignment data processing. 3′ adapters were removed using cutadapt (Martin 2011) with a base quality filter set to 28. Remaining reads with sequence length between 15 and 27, which generally correspond to miRNA sequences, were then kept for alignment. Note that identical reads were collapsed together to decrease the computational burden associated with processing *n* times *n* identical reads.

Definition of an alignment reference library. Read alignment generally starts by the selection of a reference library to which reads shall be aligned. The miRBase 21 database (Kozomara and Griffiths-Jones 2014) containing known human mature miRNAs is usually adopted for miRSeq data. We first upgrade this reference library by adding new sequences corresponding to the alternate forms of polymiRs showing genetic polymorphisms in their mature sequence as previously proposed (Baras et al. 2015; Russell et al. 2018). In case a polymiR contains more than one polymorphism, new sequences corresponding to all possible haplotypes are generated. These variants, that are provided by the OPTIMIR's user in a vcf format file (Danecek et al. 2011), are mapped to miRNAs using miRBase miRNA coordinates file (i.e., positions of miRNAs on the human reference genome). The generation of new sequences is automated via a standalone python script provided with the OPTIMIR pipeline.

For the current application to MARTHA GWAS data, we identified 88 single nucleotide polymorphisms (SNPs) for which we have a reliable genotype data, defined as SNPs with imputation $r^2 > 0.8$. Note that some SNPs may map to two distinct miRNAs if the latter are transcribed from opposite strands. Some miRNAs may also have more than one SNP in their sequence. In our application, five SNPs mapped to two distinct miRNAs and three miRNAs contained more than one SNP (see Supplemental

Figure S2 for examples). As a result, the reference library was upgraded with 96 new alternative sequences corresponding to all possible haplotypes derived from the 90 identified polymiRs.

Read alignment process. For read alignment, we opted for the bowtie2 software (Langmead and Salzberg 2012) that can handle trimming and tailing events at the reads' extremities via its local alignment mode which has been shown to be efficient for miRSeq data alignment (Ziemann et al. 2016). Only reads with a sequence of at least 17 consecutive nucleotides (defined as the alignment seed) that perfectly match with the reference library are kept in the analysis (see Supplemental Table S1 for details concerning the choice of the seed value and its consequences on isomiR detection). In summary, the OPTIMIR pipeline does not authorize any mismatch in the central sequence of a read but allows variations at its extremities to address post-transcriptional editing. The handling of miRNAs with genetic variations is addressed by the use of the upgraded reference library described above. For miRNAs that underwent tailing events, or trimming events followed by tailing events, additional bases exceeding or differing from the reference are soft-clipped and do not participate in the alignment. With a limited read length of 27 nt, the maximum number of bases that can be soft-clipped was set to 10. Reads were allowed to align to multiple reference sequences in order to take into account the different mature miRNAs with similar sequences from which they could originate. Finally, we did not allow reverse complement alignment as small RNAs were first ligated with different 5′ and 3′ adapters before single-end sequencing, which implies that RNA strands were sequenced in only one direction.

Resolution of ambiguous alignments

To handle multiple ambiguous alignments, OPTIMIR integrated a scoring algorithm aimed at identifying the most plausible alignment(s) while discarding likely erroneous ones. Of note, beforehand, for reads mapping to a mature miRNA that can be produced by two different pri-miRNAs (e.g., hsa-miR-1255b-5p can originate from hsa-mir-1255b-1 or hsa-mir-1255b-2 located on chromosome 4 and 1, respectively), we used the information on templated tailed nucleotides (i.e., nucleotides in the pri-miRNA sequences (also available in the miRBase 21) that surround the mature miRNA sequence) to deduce from which locus these reads might come from. This information, that has no impact on the alignment per se, is stored in an output file (named *expressed_hairpins.annot*).

Each alignment was assigned a score based on the number of trimming and tailing events that could make a given read perfectly match with a mature miRNA sequence. Since trimming and tailing are frequently observed in the 3′ end of miRNAs but rare in the 5′ end (Neilsen et al. 2012; Wu et al. 2018), a more penalizing weight was applied on events observed on the 5′ end. Alignments with a lower score would be considered as more reliable as they would correspond to a miRNA with less editing events. The alignment score is calculated as follows:

$$\text{Alignment score} = N_5 \times W_5 + N_3,$$

where N_5 and N_3 represent the number of editing events observed on the 5′ and 3′ extremities of the read, respectively. W_5 is the weight applied to the events observed on the 5′ end. Several W_5 values were tested and their impact on the alignment results are shown in Supplemental Table S2. Finally, for our application, W_5 was set to four in order to resolve as many ambiguous alignments as possible without penalizing too many 5′ events compared to 3′ events as they represent ∼60% of aligned reads. Templated tailed nucleotides do not count as editing events as they tend to validate the parentage of a read to its reference. These templated nucleotides are most likely the result of imprecise cleavage by DROSHA and DICER. However, they might occasionally result from the action of the terminal nucleotidyl transferase that adds the same nucleotides as those surrounding the original sequence, in which case they cannot be distinguished.

By the end of the scoring algorithm, the alignment with the lowest score was retained. In case of an ambiguous read with *n* possible alignments having the same score, all alignments are kept and assigned to a weight of *1/n*, and corresponding alignments are listed in an output file (named *remaining_ambiguous.annot*).

Of note, bowtie2 also integrates an alignment score. However, this scoring is general and does not integrate the biological knowledge on editing events specific to miRNAs. The OPTIMIR scoring algorithm also differs from the one recently proposed in IsomiR-SEA pipeline (Urgese et al. 2016) which is based on the number of observed mismatches and the difference in size between a given read and the reference mature miRNAs.

Genotype consistency analysis

In case users provide genetic information for individuals that have been miRSeq profiled, the last step of the OPTIMIR workflow is to provide a comparison analysis of the genotype data provided by the user (in a standard vcf format) and the genotype data that could be inferred from the sequenced reads aligned to polymiRs. For an individual whose reads aligned onto a polymiR sequence that harbors the alternate allele of a SNP, consistency will be called if this individual is either heterozygous or homozygous for this allele in the provided vcf genotype file. Inconsistent alignments are discarded but saved in an output file (named *inconsistents.sam*).

Indeed, it may occur that some reads align to a polymiR sequence that harbors a given allele in an individual that is not expected to carry it. This could be due to modifications induced during library preparation or sequencing that mimic genetic variations (Wright et al. 2019). However, for a given individual and a given polymiR, if this event is observed for a large number of reads, another explanation must be looked for. For instance, this could occur when reads originate from sequenced fragments of other small noncoding RNAs that share a high similarity with a polymiR. Such situations are detailed in a separate output (named *consistency_table.annot*).

To detect these situations, we set up a threshold based on the number of reads aligned to each allele of the polymiR: if the presence of an alternative allele is supported by more than 1% of the total reads (i.e., reads with the reference or alternative allele) that aligned to its associated polymiR, then this allele is considered as plausible. For example, given a polymiR with 980 reads aligned to the sequence integrating the reference allele and 20 reads aligned to the sequence integrating the alternative allele, then the percentage of reads supporting the presence of the alternative allele is 2%. Such a situation would then be considered inconsistent for an individual genotyped as homozygous for the reference allele and would deserve to be further investigated.

Note that, when polymiRs have less than 500 aligned reads in total, a threshold of five supporting reads instead of a percentage of 1% was used. These parameters can be modified by the users.

Care is needed to call inconsistency when a polymiR may have homologous mature sequences and one of them is polymorphic. For example, the mature miRNA hsa-miR-1255b-5p can originate either from the pri-miRNA hsa-miR-1255b-1 located on chromosome 4 or from hsa-mir-1255b-2 located on chromosome 1. However, only the chromosome 4 copy contains a variant. If reads with the alternate allele can easily be deduced to originate from hsa-miR-1255b-1, reads with the reference allele can come from both chromosome 1 and 4 copies. As a consequence, homozygous carriers of the alternate allele can still have reads mapping to chromosome 1 copies and such reads shall not be considered as inconsistent. We have listed in Supplemental Table S3 all mature miRNAs that have multiple pri-miRNAs sequences and tagged those that are polymorphic.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

F.T. and M.R. were financially supported by the GENMED Laboratory of Excellence on Medical Genomics, Agence Nationale de la Recherche (ANR-10-LABX-0013). D-A.T. was financially supported by the "EPIDEMIOM-VTE" Senior Chair from the Initiative of Excellence of the University of Bordeaux. MiRNA sequencing in the MARTHA study was performed on the iGenSeq platform (ICM Institute, Paris) and supported by a grant from the European Society of Cardiology for Medical Research Innovation.

Received November 26, 2018; accepted February 21, 2019.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410. doi:10.1016/S0022-2836(05)80360-2

Ameres SL, Zamore PD. 2013. Diversifying microRNA sequence and function. *Nat Rev Mol Cell Biol* **14:** 475–488. doi:10.1038/nrm3611

Baras AS, Mitchell CJ, Myers JR, Gupta S, Weng L-C, Ashton JM, Cornish TC, Pandey A, Halushka MK. 2015. miRge—a multiplexed method of processing small RNA-seq data to determine microRNA entropy. *PLoS One* **10:** e0143066. doi:10.1371/journal.pone.0143066

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297. doi:10.1016/S0092-8674(04)00045-5

Bofill-De Ros X, Chen K, Chen S, Tesic N, Dusan R, Skundric N, Nessic S, Varjacic V, Williams EH, Malhotra R, et al. 2018. QuagmiR: a cloud-based application for isomiR big data analytics. *Bioinformatics* doi:10.1093/bioinformatics/bty843

Burroughs AM, Ando Y, de Hoon MJL, Tomaru Y, Nishibu T, Ukekawa R, Funakoshi T, Kurokawa T, Suzuki H, Hayashizaki Y, et al. 2010. A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome Res* **20:** 1398–1410. doi:10.1101/gr.106054.110

Cammaerts S, Strazisar M, De Rijk P, Del Favero J. 2015. Genetic variants in microRNA genes: impact on microRNA expression, function, and disease. *Front Genet* **6:** 186. doi:10.3389/fgene.2015.00186

Chen CJ, Heard E. 2013. Small RNAs derived from structural non-coding RNAs. *Methods* **63:** 76–84. doi:10.1016/j.ymeth.2013.05.001

Chiang HR, Schoenfeld LW, Ruby JG, Auyeung VC, Spies N, Baek D, Johnston WK, Russ C, Luo S, Babiarz JE, et al. 2010. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev* **24:** 992–1009. doi:10.1101/gad.1884710

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27:** 2156–2158. doi:10.1093/bioinformatics/btr330

Desvignes T, Loher P, Eilbeck K, Ma J, Urgese G, Fromm B, Sydes J, Aparicio-Puerta E, Barrera V, Espin R, et al. 2018. Unification of miRNA and isomiR research: the mirGFF3 format and the mirtop API. *bioRxiv* doi:10.1101/505222

Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aïssi D, Wahl S, Meduri E, Morange P-E, Gagnon F, Grallert H, et al. 2014. DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383:** 1990–1998. doi:10.1016/S0140-6736(13)62674-4

Duan R, Pak C, Jin P. 2007. Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. *Hum Mol Genet* **16:** 1124–1131. doi:10.1093/hmg/ddm062

Filipowicz W, Bhattacharyya SN, Sonenberg N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* **9:** 102–114. doi:10.1038/nrg2290

Germain M, Chasman DI, de Haan H, Tang W, Lindström S, Weng L-C, de Andrade M, de Visser MCH, Wiggins KL, Suchon P, et al. 2015. Meta-analysis of 65,734 individuals identifies *TSPAN15* and *SLC44A2* as two susceptibility loci for venous thromboembolism. *Am J Hum Genet* **96:** 532–542. doi:10.1016/j.ajhg.2015.01.019

Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, Bertone P, Caldas C. 2010. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA* **16:** 991–1006. doi:10.1261/rna.1947110

Gong J, Tong Y, Zhang H-M, Wang K, Hu T, Shan G, Sun J, Guo A-Y. 2012. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. *Hum Mutat* **33:** 254–263. doi:10.1002/humu.21641

Han M, Zheng Y. 2013. Comprehensive analysis of single nucleotide polymorphisms in human microRNAs. *PLoS One* **8:** e78028. doi:10.1371/journal.pone.0078028

Heintz-Buschart A, Yusuf D, Kaysen A, Etheridge A, Fritz JV, May P, de Beaufort C, Upadhyaya BB, Ghosal A, Galas DJ, et al. 2018. Small RNA profiling of low biomass samples: identification and removal of contaminants. *BMC Biol* **16:** 52. doi:10.1186/s12915-018-0522-7

Katoh T, Hojo H, Suzuki T. 2015. Destabilization of microRNAs in human cells by 3′ deadenylation mediated by PARN and CUGBP1. *Nucleic Acids Res* **43:** 7521–7534. doi:10.1093/nar/gkv669

Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K. 2007. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* **315:** 1137–1140. doi:10.1126/science.1138050

Kawamata T, Tomari Y. 2010. Making RISC. *Trends Biochem Sci* **35:** 368–376. doi:10.1016/j.tibs.2010.03.009

Kim VN, Han J, Siomi MC. 2009. Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* **10:** 126–139. doi:10.1038/nrm2632

Kim Y-K, Kim B, Kim VN. 2016. Re-evaluation of the roles of *DROSHA*, *Exportin 5*, and *DICER* in microRNA biogenesis. *Proc Natl Acad Sci* **113:** E1881–E1889. doi:10.1073/pnas.1602532113

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42:** D68–D73. doi:10.1093/nar/gkt1181

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357. doi:10.1038/nmeth.1923

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6:** 26. doi:10.1186/1748-7188-6-26

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17:** 10–12. doi:10.14806/ej.17.1.200

Meijer HA, Smith EM, Bushell M. 2014. Regulation of miRNA strand selection: follow the leader? *Biochem Soc Trans* **42:** 1135–1140. doi:10.1042/BST20140142

Mencía A, Modamio-Høybjør S, Redshaw N, Morín M, Mayo-Merino F, Olavarrieta L, Aguirre LA, del Castillo I, Steel KP, Dalmay T, et al. 2009. Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat Genet* **41:** 609–613. doi:10.1038/ng.355

Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, Peterson A, Noteboom J, O'Briant KC, Allen A, et al. 2008. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci* **105:** 10513–10518. doi:10.1073/pnas.0804549105

Morales S, Monzo M, Navarro A. 2017. Epigenetic regulation mechanisms of microRNA expression. *Biomol Concepts* **8:** 203–212. doi:10.1515/bmc-2017-0024

Neilsen CT, Goodall GJ, Bracken CP. 2012. IsomiRs—the overlooked repertoire in the dynamic microRNAome. *Trends Genet* **28:** 544–549. doi:10.1016/j.tig.2012.07.005

Nishikura K. 2016. A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol* **17:** 83–96. doi:10.1038/nrm.2015.4

Oudot-Mellakh T, Cohen W, Germain M, Saut N, Kallel C, Zelenika D, Lathrop M, Trégouët D-A, Morange P-E. 2012. Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. *Br J Haematol* **157:** 230–239. doi:10.1111/j.1365-2141.2011.09025.x

Pulcrano-Nicolas A-S, Proust C, Clarençon F, Jacquens A, Perret C, Roux M, Shotar E, Thibord F, Puybasset L, Garnier S, et al. 2018. Whole-blood miRNA sequencing profiling for vasospasm in patients with aneurysmal subarachnoid hemorrhage. *Stroke* **49:** 2220–2223. doi:10.1161/STROKEAHA.118.021101

Roux M, Perret C, Feigerlova E, Mohand Oumoussa B, Saulnier P-J, Proust C, Trégouët D-A, Hadjadj S. 2018. Plasma levels of hsa-miR-152-3p are associated with diabetic nephropathy in patients with type 2 diabetes. *Nephrol Dial Transplant* **33:** 2201–2207. doi:10.1093/ndt/gfx367

Russell PH, Vestal B, Shi W, Rudra PD, Dowell R, Radcliffe R, Saba L, Kechris K. 2018. miR-MaGiC improves quantification accuracy for small RNA-seq. *BMC Res Notes* **11:** 296. doi:10.1186/s13104-018-3418-2

Tam S, de Borja R, Tsao M-S, McPherson JD. 2014. Robust global microRNA expression profiling using next-generation sequencing technologies. *Lab Invest* **94:** 350–358. doi:10.1038/labinvest.2013.157

Urgese G, Paciello G, Acquaviva A, Ficarra E. 2016. isomiR-SEA: an RNA-seq analysis tool for miRNAs/isomiRs expression level profiling and miRNA-mRNA interaction sites evaluation. *BMC Bioinformatics* **17:** 148. doi:10.1186/s12859-016-0958-0

Wallaert A, Van Loocke W, Hernandez L, Taghon T, Speleman F, Van Vlierberghe P. 2017. Comprehensive miRNA expression profiling in human T-cell acute lymphoblastic leukemia by small RNA-sequencing. *Sci Rep* **7:** 7901. doi:10.1038/s41598-017-08148-x

Wright C, Rajpurohit A, Burke EE, Williams C, Collado-Torres L, Kimos M, Brandon NJ, Cross AJ, Jaffe AE, Weinberger DR, et al. 2019. Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods. *bioRxiv* doi: 10.1101/445437

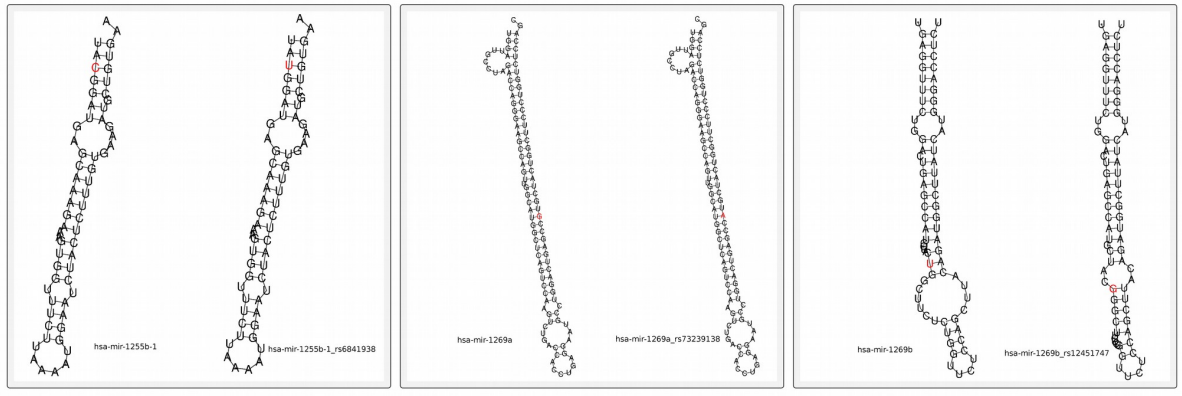
Wu X, Kim T-K, Baxter D, Scherler K, Gordon A, Fong O, Etheridge A, Galas DJ, Wang K. 2017. sRNAnalyzer—a flexible and customizable small RNA sequencing data analysis pipeline. *Nucleic Acids Res* **45:** 12140–12151. doi:10.1093/nar/gkx999

Wu CW, Evans JM, Huang S, Mahoney DW, Dukek BA, Taylor WR, Yab TC, Smyrk TC, Jen J, Kisiel JB, et al. 2018. A Comprehensive Approach to Sequence-oriented IsomiR annotation (CASMIR): demonstration with IsomiR profiling in colorectal neoplasia. *BMC Genomics* **19:** 401. doi:10.1186/s12864-018-4794-7

Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, Krouse MA, Webster PJ, Tewari M. 2011. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res* **21:** 1450–1461. doi:10.1101/gr.118059.110

Ziemann M, Kaspi A, El-Osta A. 2016. Evaluation of microRNA alignment techniques. *RNA* **22:** 1120–1138. doi:10.1261/rna.055509.115

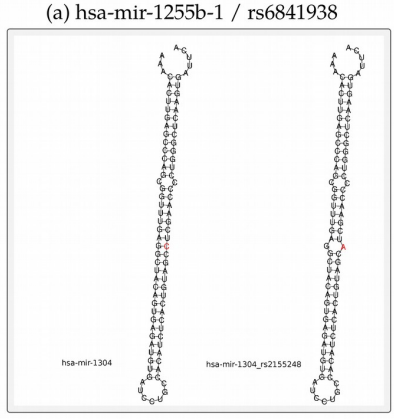
Supplemental Figure S1 (1/4): RNAfold predictions of polymiRs secondary structures that are expressed by heterozygous samples



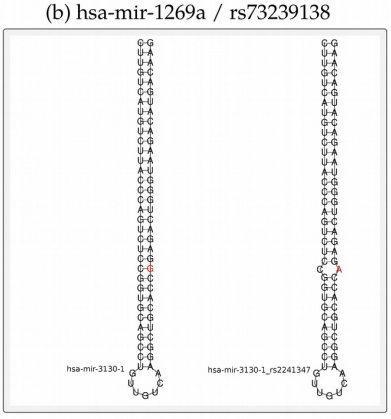
(a) hsa-mir-1255b-1 / rs6841938

(b) hsa-mir-1269a / rs73239138

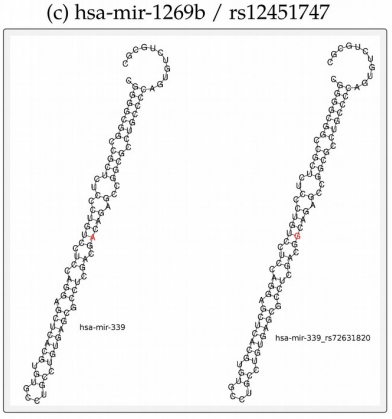
(c) hsa-mir-1269b / rs12451747



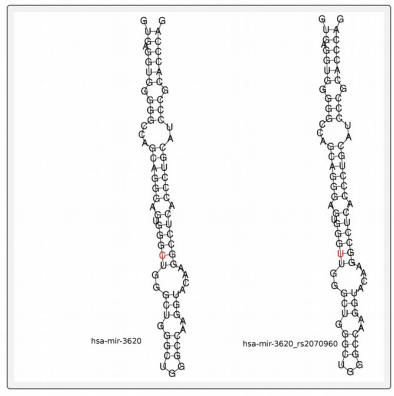
(d) hsa-mir-1304 / rs2155248



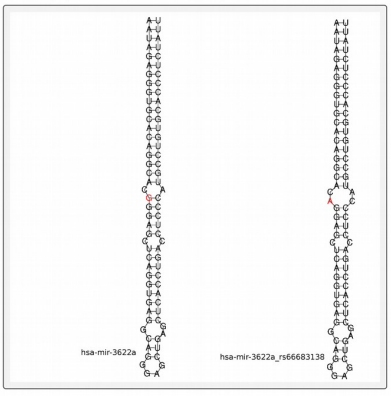
(e) hsa-mir-3130-1 / rs2241347



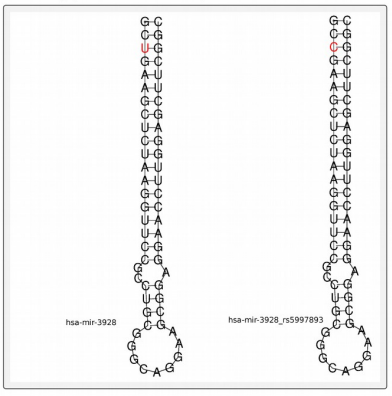
(f) hsa-mir-339 / rs72631820



(g) hsa-mir-3620 / rs2070960

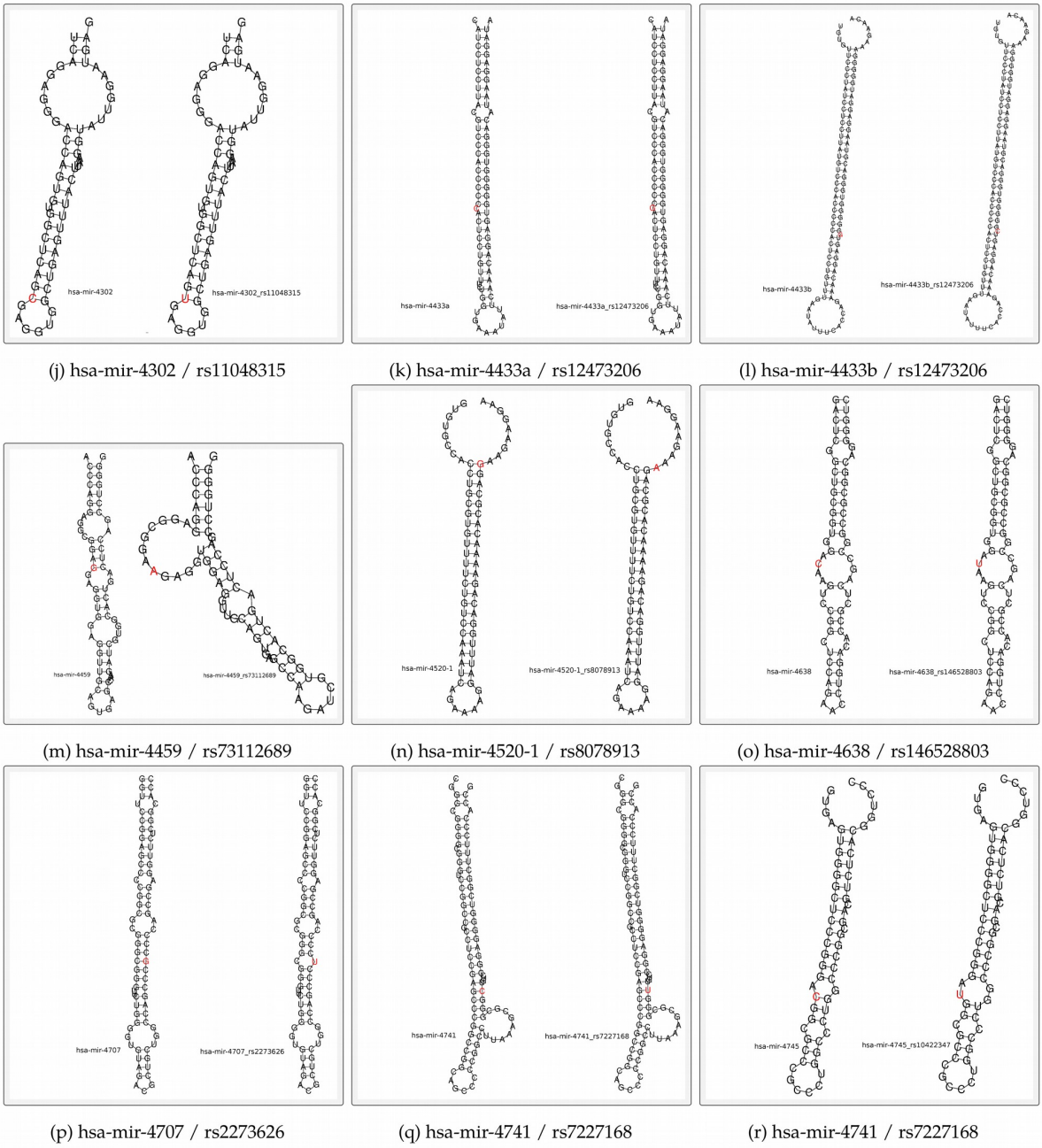


(h) hsa-mir-3622a / rs66683138



(i) hsa-mir-3928 / rs5997893

Supplemental Figure S1 (2/4): RNAfold predictions of polymiRs secondary structures that are expressed by heterozygous samples

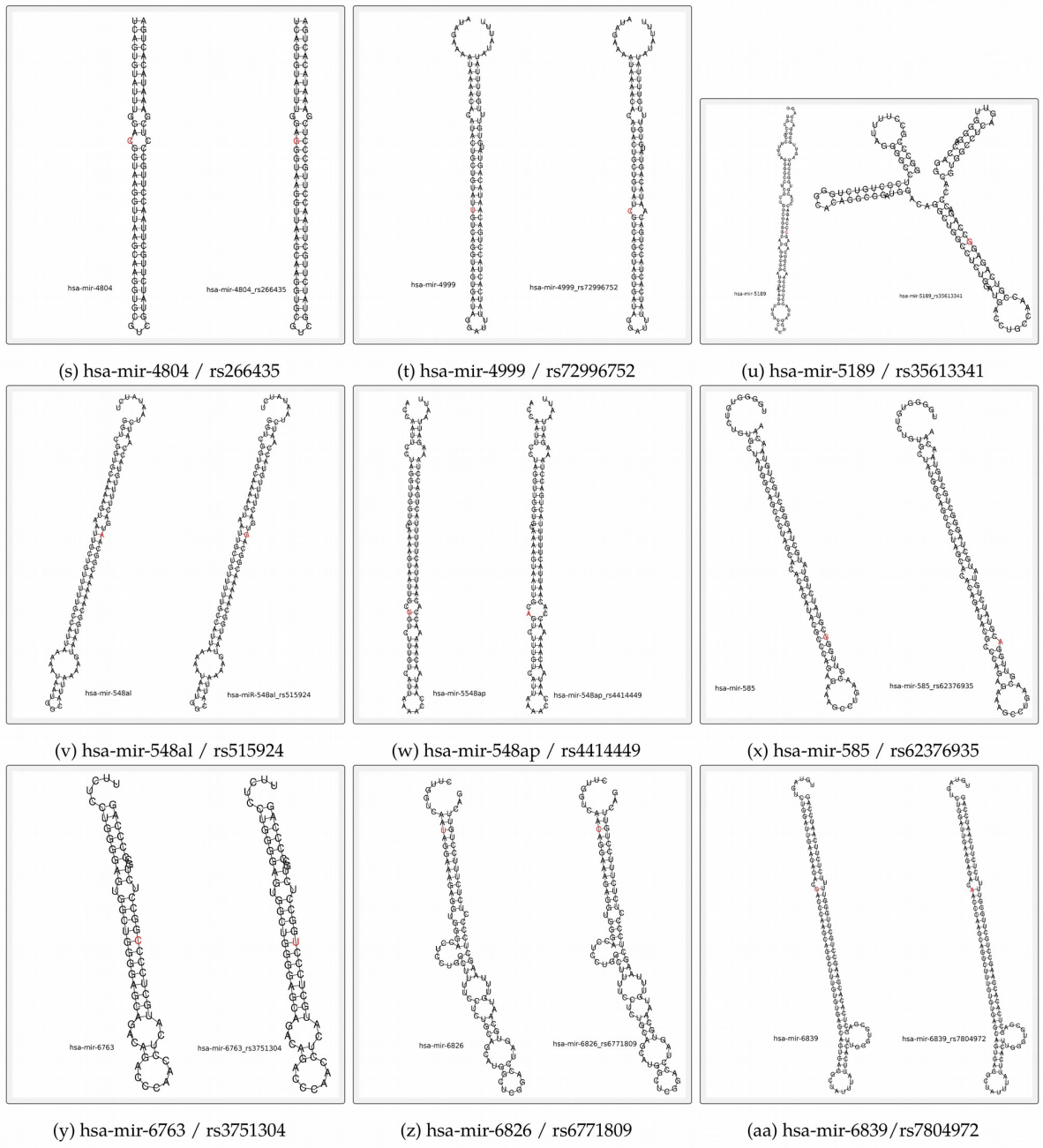


(j) hsa-mir-4302 / rs11048315 (k) hsa-mir-4433a / rs12473206 (l) hsa-mir-4433b / rs12473206

(m) hsa-mir-4459 / rs73112689 (n) hsa-mir-4520-1 / rs8078913 (o) hsa-mir-4638 / rs146528803

(p) hsa-mir-4707 / rs2273626 (q) hsa-mir-4741 / rs7227168 (r) hsa-mir-4741 / rs7227168

Supplemental Figure S1 (3/4): RNAfold predictions of polymiRs secondary structures that are expressed by heterozygous samples

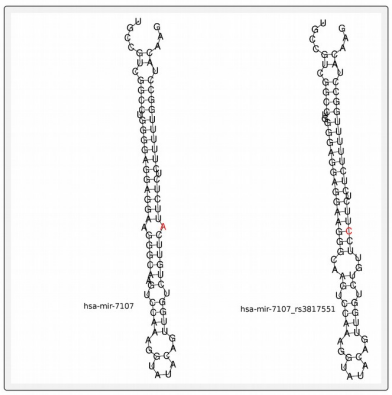


(s) hsa-mir-4804 / rs266435 (t) hsa-mir-4999 / rs72996752 (u) hsa-mir-5189 / rs35613341

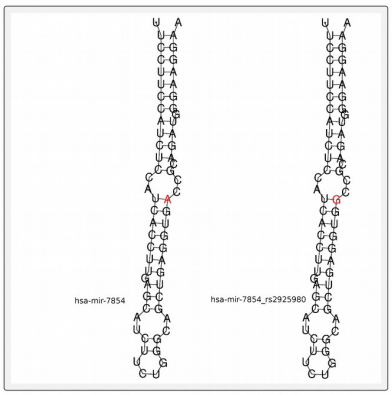
(v) hsa-mir-548al / rs515924 (w) hsa-mir-548ap / rs4414449 (x) hsa-mir-585 / rs62376935

(y) hsa-mir-6763 / rs3751304 (z) hsa-mir-6826 / rs6771809 (aa) hsa-mir-6839/rs7804972

Supplemental Figure S1 (4/4): RNAfold predictions of polymiRs secondary structures that are expressed by heterozygous samples

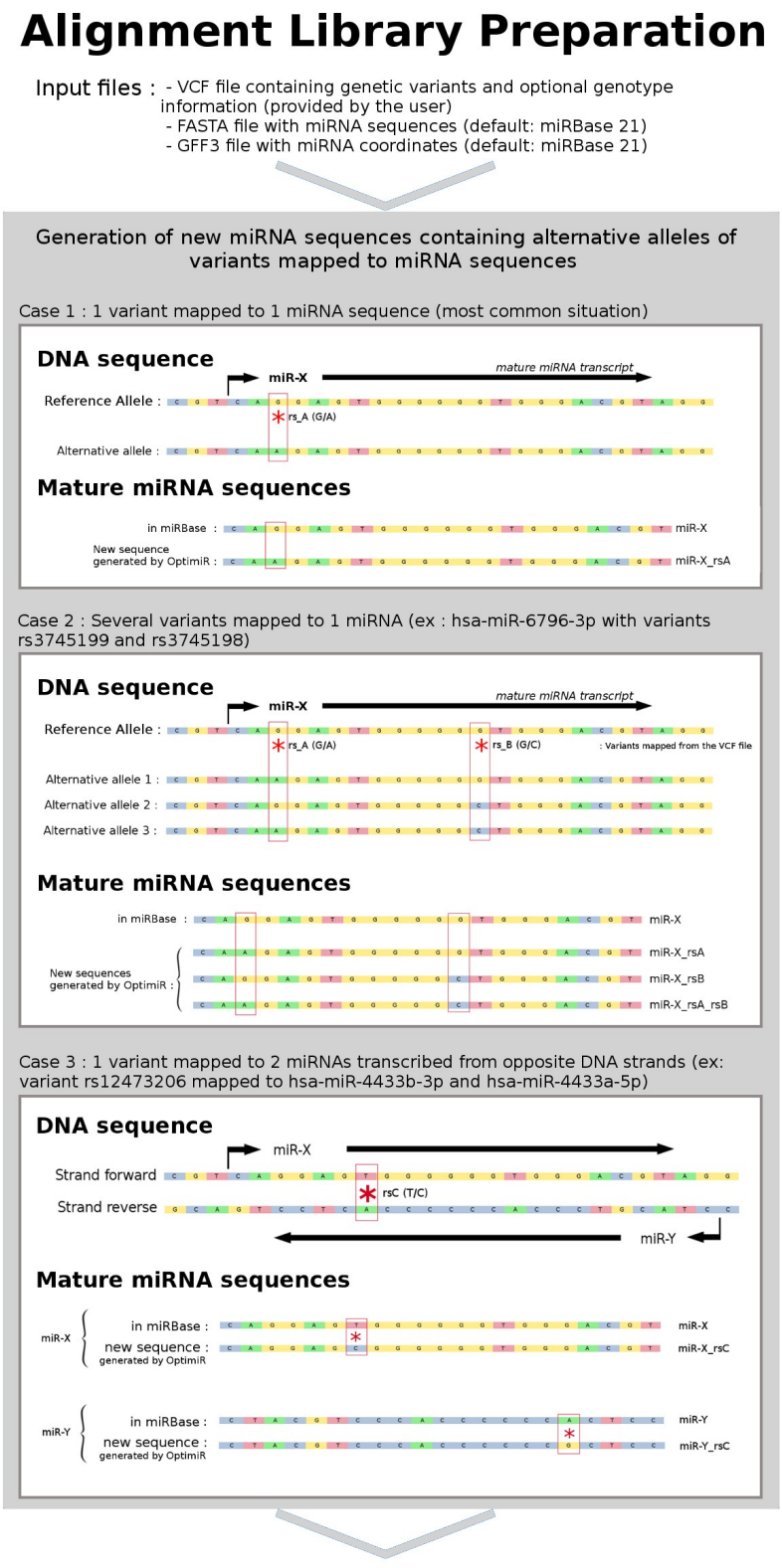


(ab) hsa-mir-7107 / rs3817551



(ac) hsa-mir-7854 / rs2925980

Supplemental Figure S2 : Library preparation workflow



Supplemental Table 1: Bowtie2 alignment of miRSeq data with different alignment seed length

Bowtie2 Seed Length	After Alignment							
	TOTAL reads	TOTAL ambigous reads	miRs: Sequences expressed (total: 2588)	Number canonical miRs	Nb Isoformes	Nb Reliable Isoforms (*)	Nb isoform per miR (average)	Nb reliable isoform per miR (average)
12	595,023,218	68,575,414	2,588	915	1,302,378	20,424	503.24	7.89
13	584,166,507	63,720,478	2,579	915	850,849	17,078	329.91	6.62
14	576,260,068	60,707,314	2,482	915	541,446	14,959	218.15	6.03
15	572,838,687	59,470,291	2,134	915	390,611	13,625	183.04	6.38
16	568,210,216	55,599,076	1,819	915	289,447	12,370	159.12	6.80
17	561,926,285	10,936,664	1,678	915	207,555	10,740	123.69	6.40
18	556,992,220	8,965,700	1,621	915	150,358	9,549	92.76	5.89
19	550,679,360	3,248,902	1,587	915	104,165	8,177	65.64	5.15
20	541,809,265	2,779,129	1,559	915	68,827	6,761	44.15	4.34
21	526,199,215	1,043,802	1,540	915	45,386	5,468	29.47	3.55
22	505,017,606	987,609	1,512	915	30,144	4,222	19.94	2.79

The default seed length in bowtie2 very-sensitive-local mode is 20. However, diminushing the seed length can lead to an increase in sensitivity, and thus in isomiR detection.
Seed length = 17 is a good compromise as it allows for a better sensitivity than longer seeds & a great reduction of ambiguous alignments compared to shorter seed lengths.
(*) Reliable isoforms are expressed with at least 5x in at least 10 samples

Supplemental Table 2: OptimiR resolution of ambiguous alignment with different W5 values (Bowtie2 seed = 17)

W_5	After Cross-Mapping Reads Resolution (Seed length = 17)							
	TOTAL reads	TOTAL ambigous reads remaining	miRs: Sequences expressed (total 2588)	Number canonical miRs	Nb Isoformes	Nb Reliable Isoforms (*)	Nb isoform per miR (average)	Nb reliable isoform per miR (average)
1	561,926,285	929,198	1,667	915	192,523	10,220	115.49	6.13
2	561,926,285	929,193	1,664	915	192,435	10,221	115.65	6.14
3	561,926,285	927,518	1,664	915	192,394	10,220	115.62	6.14
4	561,926,285	927,284	1,664	915	192,374	10,220	115.61	6.14
5	561,926,285	927,976	1,664	915	192,393	10,220	115.62	6.14
6	561,926,285	927,389	1,664	915	192,378	10,220	115.61	6.14

Application of OptimiR with different W5 values.
Although results are close between different weights, W5 = 4 showed the best results concerning the resolution of ambiguous alignments.

Supplemental Table 3: List of mature miRNAs that can originate from different hairpins

EXTRAIT de la table: Le reste de la table est disponible en ligne à l'adresse: https://rnajournal.cshlp.org/content/25/6/657/suppl/DC1

MIRNA	HAIRPINS						VARIANT:
	Name	Chrom	Start	End	Sens	ID	
hsa-miR-1972							
	hsa-mir-1972-1	chr16	15010321	15010397	-	MI0009982	
	hsa-mir-1972-2	chr16	70030346	70030422	+	MI0015977	
hsa-miR-3199							
	hsa-mir-3199-1	chr22	27920525	27920612	-	MI0014247	**rs75321888**
	hsa-mir-3199-2	chr22	27920526	27920611	+	MI0014248	
hsa-miR-4444							
	hsa-mir-4444-1	chr2	177212726	177212799	+	MI0016787	
	hsa-mir-4444-2	chr3	75214476	75214549	+	MI0019111	
hsa-miR-5701							
	hsa-mir-5701-1	chr15	20940252	20940333	+	MI0019308	
	hsa-mir-5701-2	chr15	21513959	21514040	+	MI0019593	
	hsa-mir-5701-3	chr15	21951242	21951323	+	MI0031522	
hsa-miR-521							
	hsa-mir-521-2	chr19	53716594	53716680	+	MI0003163	
	hsa-mir-521-1	chr19	53748636	53748722	+	MI0003176	
hsa-miR-450a-5p							
	hsa-mir-450a-1	chrX	134540341	134540431	-	MI0001652	
	hsa-mir-450a-2	chrX	134540508	134540607	-	MI0003187	
hsa-miR-3158-5p							
	hsa-mir-3158-1	chr10	101601417	101601497	+	MI0014186	
	hsa-mir-3158-2	chr10	101601417	101601497	-	MI0014187	

Supplemental Table 4: PolymiRs expressed by heterozygous carriers

polymiR	rsID	Number of Samples	Counts Reference	Counts Alternative	Mean Rate	Median Rate	Min Rate	Max Rate
hsa-miR-4433b-3p	rs12473206	147	19012	85651	0.81	0.83	0.00	1.00
hsa-miR-7854-3p	rs2925980	65	42	1198	0.96	1.00	0.00	1.00
hsa-miR-1255b-5p	rs6841938	48	4034.5	64.5	0.01	0.01	0.00	0.08
hsa-miR-7107-3p	rs3817551	33	574	1	0.00	0.00	0.00	0.06
hsa-miR-5189-3p	rs35613341	20	385	0	0.00	0.00	0.00	0.00
hsa-miR-4745-5p	rs10422347	14	300	15	0.07	0.00	0.00	1.00
hsa-miR-6826-5p	rs6771809	11	24	209	0.91	1.00	0.00	1.00
hsa-miR-548ap-5p	rs4414449	11	14	3	0.09	0.00	0.00	0.50
hsa-miR-4707-3p	rs2273626	6	107	0	0.00	0.00	0.00	0.00
hsa-miR-1304-3p	rs2155248	5	146	73	0.20	0.02	0.00	0.54
hsa-miR-3622a-5p	rs66683138	4	61	24	0.25	0.00	0.00	1.00
hsa-miR-3620-5p	rs2070960	4	15	0	0.00	0.00	0.00	0.00
hsa-miR-4741	rs7227168	4	67	0	0.00	0.00	0.00	0.00
hsa-miR-3130-3p	rs2241347	3	19	0	0.00	0.00	0.00	0.00
hsa-miR-4804-5p	rs266435	3	0	38	1.00	1.00	1.00	1.00
hsa-miR-339-3p	rs72631820	3	49	0	0.00	0.00	0.00	0.00
hsa-miR-6763-3p	rs3751304	2	66	0	0.00	0.00	0.00	0.00
hsa-miR-585-3p	rs62376935	2	2	48	0.98	0.98	0.95	1.00
hsa-miR-4433a-5p	rs12473206	2	6	0	0.00	0.00	0.00	0.00
hsa-miR-6839-5p	rs7804972	2	8	0	0.00	0.00	0.00	0.00
hsa-miR-4302	rs11048315	2	0	2	1.00	1.00	1.00	1.00
hsa-miR-1269a	rs73239138	2	5	31.5	0.50	0.50	0.00	1.00
hsa-miR-4999-5p	rs72996752	1	0	26	1.00	1.00	1.00	1.00
hsa-miR-4638-5p	rs146528803	1	15	0	0.00	0.00	0.00	0.00
hsa-miR-1269b	rs12451747	1	31.5	0	0.00	0.00	0.00	0.00
hsa-miR-4520-3p	rs8078913	1	0.5	0.5	0.50	0.50	0.50	0.50
hsa-miR-548al	rs515924	1	0	4	1.00	1.00	1.00	1.00
hsa-miR-4459	rs73112689	1	11	0	0.00	0.00	0.00	0.00
hsa-miR-3928-5p	rs5997893	1	0.5	0.5	0.50	0.50	0.50	0.50

4 Discussion

Apports de la stratégie optimiR La méthode de détection et de quantification des miARNs implémentée dans le pipeline optimiR se distingue des stratégies existante en intégrant l'information génétique, permettant de quantifier avec précision les polymiRs. La combinaison de cette approche avec la stratégie d'alignement local et de résolution d'alignements ambigus permet également de limiter le nombre de faux négatifs et de faux positifs. Toutefois, il reste un certain nombre de faux négatifs inévitables avec notre approche, à cause de la contrainte d'un alignement parfait de la séquence centrale des reads. Cette contrainte permet de diminuer significativement le nombre d'alignements ambigus, mais les reads dont la séquence centrale est altérée lors du séquençage ou lors de la préparation des librairies ne pourront pas être alignés. Et pour la même raison, les isomiRs avec édition ARN ne seront pas non plus détectés.

Perspectives d'amélioration La diminution de faux négatifs et la détection des isomiRs avec édition ARN permettrait d'améliorer la spécificité de l'outil. Par exemple, une seconde étape d'alignement moins stringente à partir des reads non alignés permettrait de récupérer une grande partie de faux négatifs. Cependant, en réduisant la stringence de l'alignement on risque également d'introduire des faux positifs et d'impacter la sensibilité de l'outil. Cette perspective demande donc une réfléction approfondie pour conserver une haute sensibilité et spécificité.

Parmi les autres perspectives d'amélioration, on pourrait également étendre la librairie de référence avec les séquences d'autres petits ARNs non codants séquencés en même temps que les miARNs. De nombreux fragments d'ARNs, issus par exemple d'ARNt, d'ARNr, de snoARN, de snARN ou d'ARN Y, peuvent ainsi être séquencés en même temps que les miARNs, et leur quantification peut présenter un intérêt pour la recherche bio-médicale, par exemple pour identifier de nouveaux biomarqueurs, ou pour découvrir de nouveaux mécanismes biologiques.

Comparaison avec les méthodes existantes Pendant les 3 années de ce projet de thèse, au moins 10 nouveaux pipelines permettant la quantification de miARNs ont fait l'objet d'une publication. Le nombre important de méthodes disponibles reflète l'importance de ce domaine de recherche, et le besoin d'une méthode de quantification précise des miARNs, avec une haute spécificité et sensibilité. Cependant, il est particulièrement difficile de comparer ces méthodes entre elles, pour évaluer leurs spécifité et sensibilité. La principale raison est que chaque outil dispose de son propre format d'annotation des résultats, et très souvent, seules les abondances finales des miARNs détectés sont générées par ces outils, sans information sur le nombre d'alignements ambigus, ou le détail des isomiRs identifiés.

La communauté miRtop Pour faciliter la comparaison d'outils, la communauté miRtop, rassemblant des chercheurs travaillant sur la quantification de miARNs et d'isomiRs, a récemment développé un nouveau format d'annotation, nommé miRGFF3. Ce format permet en particulier d'indiquer le détail des abondances pour chaque isomiR, de recenser les reads qui n'ont pas été alignés, et d'indiquer les alignements ambigus. Après la publication d'optimiR, j'ai été invité à rejoindre ce groupe, et j'ai ainsi eu l'opportunité de participer au développement du format miRGFF3, qui a fait l'objet d'une publication dans la revue *Bioinformatics* (incluse en Appendice B). Grâce à ce format unique, qui a été intégré à optimiR, il sera donc possible de comparer les résultats obtenus par des outils différents, afin d'identifier les forces et faiblesses des différents pipelines d'alignement.

Chapitre IV : MicroARNs et paramètres de l'hémostase

1 Introduction: Hémostase, maladie thromboembolique veineuse, et implication des microARNs

1.1 Hémostase

L'hémostase correspond à un ensemble de mécanismes naturels permettant au sang de se maintenir dans les vaisseaux sanguins dans un état fluide. En cas de dommage vasculaire entraînant une brèche dans un vaisseau, une vasoconstriction du vaisseau permet une diminution du flux sanguin, et une accumulation de plaquettes sanguines (phase de l'hémostase primaire) et de fibrine générée par la cascade de la coagulation (phase de l'hémostase secondaire) permettent la formation d'un caillot sanguin, appelé thrombus. La fibrine permet ainsi d'agréger les plaquettes, mais aussi différents éléments du sang tels que les érythrocytes [1] et ainsi former un thrombus assez large pour bloquer la paroi endomagée du vaisseau, et stopper l'hémorragie.

Hémostase primaire Lors d'un dommage de l'endothélium, le collagène, normalement isolé, est exposé aux plaquettes en circulation qui viennent s'y lier, et l'adhésion est renforcée par le facteur von Willebrand (vWF), secrété par les plaquettes et les cellules endothéliales, résultant en un amas qui bloque le saignement. Les plaquettes sont alors dans leur forme activée, qui entraîne une modification leur structure pour faciliter leur adhésion. L'activation déclenche aussi la production de microparticules [2] contenant divers ARNs et protéines, notamment le vWF et le thromboxane A2, qui active des plaquettes additionnelles. L'agrégat de plaquettes est ensuite renforcé par des fibres de fibrines, produites par la cascade de la coagulation.

Hémostase secondaire La cascade de la coagulation correspond à une série de réactions impliquant différentes molécules dont des enzymes, qui sont tour à tour activées pour catalyser la réaction suivante, permettant de générer la fibrine. Ces protéines sont appelées facteurs de coagulation, qui sont généralement identifiés par des chiffres romains, avec un "a" minuscule pour indiquer la forme activée. La liste des facteurs et de leurs fonctions est donnée dans la Table IV.1. La cascade de la coagulation est divisée en deux voies : la voie extrinsèque (activée par le facteur tissulaire relâché dans le sang lors d'un dommage de l'endothélium) et la voie intrinsèque (activée par contact avec les structures sous endothéliales), qui activent toutes les

1. Aussi appelés globules rouges, des cellules anucléées dont la fonction est le transport d'O_2 et de CO_2. Ce sont les cellules sanguines les plus nombreuses.
2. Des vésicules extracellulaires de petite taille.

Facteur	Nom	Fonction
I	Fibrinogène	Renforce le thrombus dans sa forme activée (fibrine)
II	Prothrombine	Active I, V, VIII, XI, XIII, protéine C (via thrombomoduline), plaquettes
III	Facteur Tissulaire (TF)	Cofacteur du VIIa
V	Proaccélérine	Cofacteur du X
VII	Proconvertine	Active IX et X
VIII	Anti-hémophile A	Cofacteur du IX, lié au vWF
IX	Anti-hémophile B	Active X à l'aide du VIIIa
X	Stuart-Prower	Active II à l'aide du Va
XI	Rosenthal	Active IX, et prékallikréine
XII	Hageman	Active XI
XIII	Laki-Lorand	Stabilise les dépots de Fibrine

TABLE IV.1 – Facteurs de la coagulation.
L'ancien facteur VI correspond à la forme activée du facteur V, tandis que le facteur IV correspond aux ions calcium Ca^{2+} nécessaires à l'activation de nombreux facteurs.

deux la voie commune et finale du facteur X, de la thrombine, et du fibrinogène qui, dans sa forme activée, correspond à la fibrine (Figure IV.1).

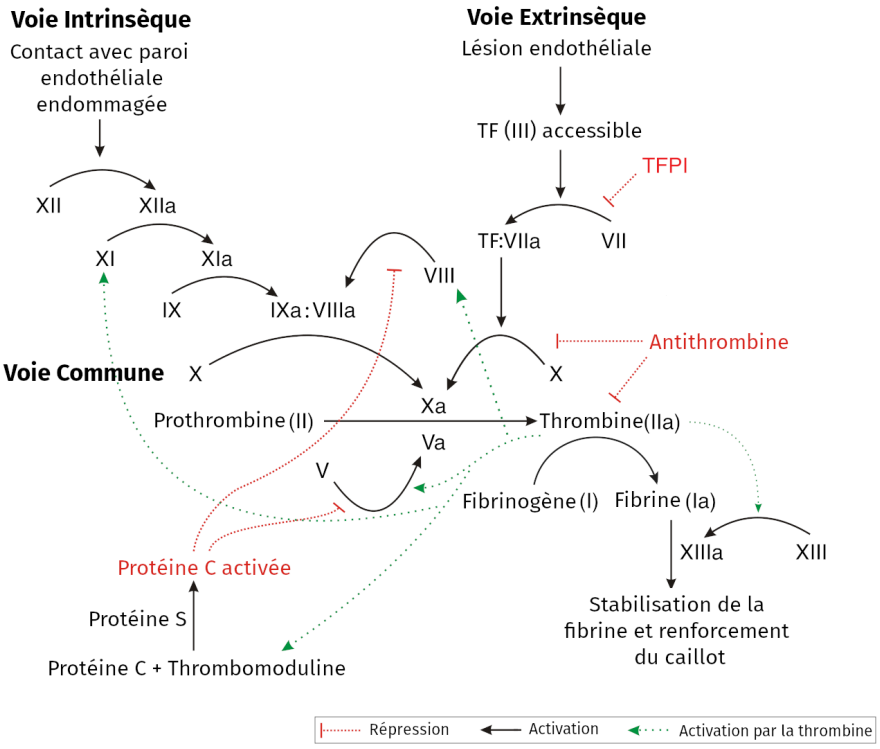


Fig. IV.1 – Cascade de la coagulation

La thrombine, initialement produite en petite quantité, active à son tour de nombreux facteurs permettant d'amplifier sa propre production, et donc la production de fibrine, dans une boucle de rétroaction positive. En parallèle, la présence de thrombine permet également l'activation de la Protéine C, produite à partir de thrombomoduline et de la Protéine S, dont la fonction est de stopper la production de FVa. C'est donc un facteur anticoagulant, tout comme le TFPI (en anglais pour *Tissue Factor Pathway Inhibitor*) et l'Antithrombine, qui

régulent la cascade de coagulation et évitent une généralisation de l'état thrombotique.

Fibrinolyse À la fin du processus thrombotique, et après réparation de la lésion endothéliale, le caillot sanguin est progressivement dissout par le mécanisme de fibrinolyse. Ce mécanisme résulte de l'action de la plasmine, qui dégrade la fibrine. La plasmine est la version active du plasminogène, qui est activé par tPA ou uPA, les activateurs tissulaires ou urokinase du plasminogène, respectivement. Ces activateurs peuvent avoir leur action inhibée par le PAI-1 (en anglais pour *Plasminogen Activator Inhibitor 1*), qui inhibe ainsi la fibrinolyse.

L'hémostase est ainsi profondément régulée et dépendante de l'équilibre entre de nombreux facteurs pro- et anti-coagulants. Ainsi, une dérégulation de cet équilibre peut avoir des effets délétères importants, tels qu'une hémorragie ou une thrombose.

1.2 Maladie thromboembolique veineuse

a) Description et épidémiologie

La thrombose veineuse (VTE) est la troisième maladie cardiovasculaire la plus fréquente, avec une incidence annuelle estimée à 1.5 cas pour 1000 personnes [107]. Elle se manifeste sous deux formes: la thrombose veineuse profonde (TVP) et l'embolie pulmonaire (EP). La TVP, ou phlébite profonde, correspond à la formation d'un thrombus sur la paroi d'une veine, généralement au niveau d'une valvule [3] dans les membres inférieurs, pouvant occasionner une réduction du débit sanguin. Lorsque le thrombus se détache de la paroi veineuse on parle alors d'embole, pouvant circuler dans le réseau sanguin jusqu'à une artère pulmonaire et provoquer une EP (Figure IV.2), caractérisée par une obstruction de la circulation sanguine vers les poumons, qui peut être fatale.

La VTE est très rare avant l'adolescence et augmente de façon exponentielle avec l'âge, pour atteindre 1 cas pour 100 personnes chaque année chez les plus de 55 ans [207], et touche plus fréquemment les hommes que les femmes. L'EP est fatale dans environ 10% des cas, et les survivants ont un mauvais pronostic, avec un taux de mortalité d'environ 20% la première année [74, 207], et jusqu'à 50% des patients développent des douleurs, un œdème ou des ulcères, qui constituent le syndrome post-thrombotique, réduisant significativement leur qualité de vie [128]. De plus, environ 30% des patients vont développer une nouvelle VTE dans les 10 ans suivant le premier épisode [107].

La formation de thrombus est favorisée par l'association de trois facteurs, appelée triade de Virchow (du nom de Rudolf Virchow):

— une altération du flux sanguin, provoquée par exemple par une immobilisation de longue durée ;
— un dysfonctionnement ou une altération de l'endothélium, résultant par exemple d'un traumatisme local ou d'une chirurgie ;
— une hypercoagulabilité, qui peut être influencée par des facteurs externes comme le tabac et l'âge, mais aussi par des facteurs génétiques entraînant une déficience d'anticoagulant naturels.

3. Les veines contiennent des valvules qui imposent un sens de circulation au sang. Une valvule est représentée sur la Figure IV.2.

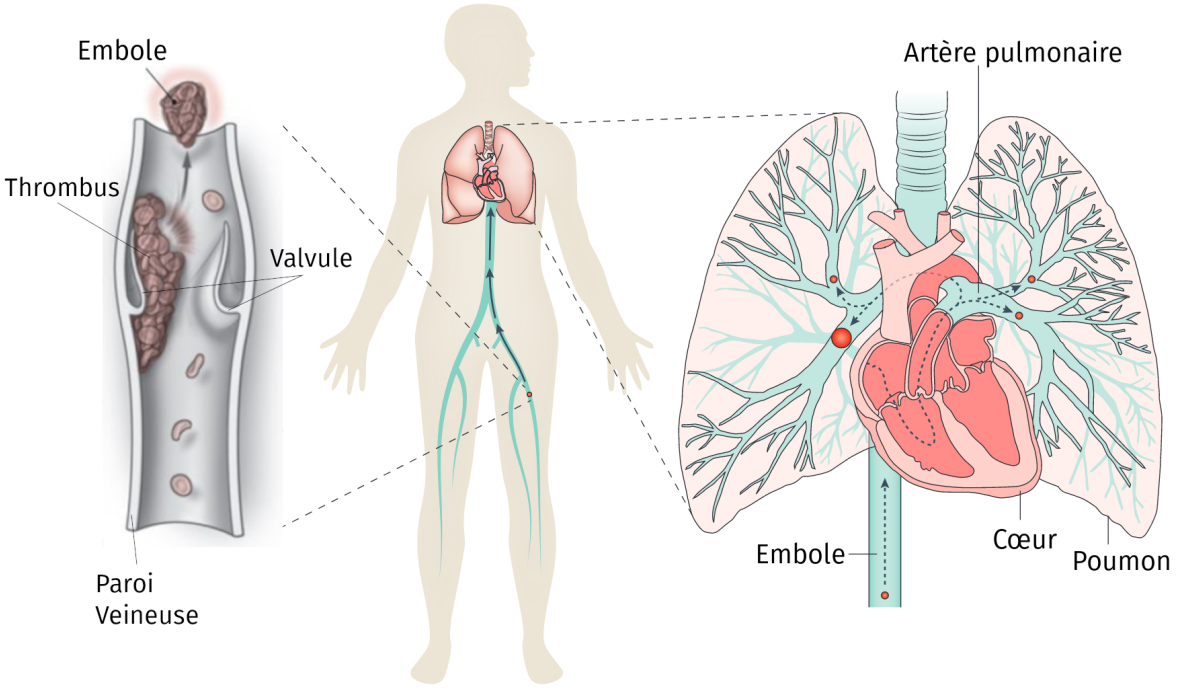


Fig. IV.2 – Thrombus formé au niveau d'une valvule dans une veine d'un membre inférieur. Une partie du thrombus est détachée, et l'embole résultante circule vers les artères pulmonaires. L'embole obstrue éventuellement la circulation pulmonaire, à cause de la réduction du diamètre des artères. *Image adaptée de [120]*

Environ la moitiée des cas de VTE sont provoqués par des facteurs externes, tels qu'une chirurgie, un trauma, une immobilisation, ou un cancer. Parmi les cas de VTE non provoquées, une part significative est attribuée à des facteurs génétiques héritables [306], pouvant causer une dérégulation de l'hémostase. Les principaux facteurs de risques héritables identifiés pour la VTE sont des déficiences des inhibiteurs de la coagulation (antithrombine, protéine C et protéine S), qui entraînent naturellement un état d'hypercoagulation. Toutefois, les mutations génétiques responsables de ces anomalies sont rares, avec une fréquence inférieure à 1% . D'autres variants plus fréquents, mais avec un risque moins élevé, ont été identifiés: le variant du F5 Leiden (rs6025), le F2 G20210A (rs1799963), et le rs2066865 du FGG (un gène entrant dans la composition du fibrinogène avec FGA et FGB). Une vingtaine de variants additionnels, plus fréquents mais avec un risque moins élevé, ont récemment été identifiés [176, 86]. De plus, les individus avec un groupe sanguin différent de O ont des niveaux de vWF plus élevés [75], et sont également à risque.

b) Prévention et traitement

Malgré sa prévalence importante, la VTE manque encore de biomarqueurs robustes pour prédire le développement de la maladie ou le risque de récidive. La mesure des D-dimères, un produit de la dégradation de la fibrine, est l'un des rares biomarqueurs utilisé cliniquement pour effectuer un diagnostic de la VTE. Ce test possède une très haute sensibilité (∼98%) mais manque de spécificité [119] car les D-dimères peuvent avoir des niveaux plus élevés avec l'âge, ou en réponse à une infection, voire même à cause d'une inflammation chronique, et

augmentent ainsi le risque de faux positifs. Il est donc difficile d'émettre un diagnostic à partir de résultats avec de hauts niveaux de D-dimères, mais les bas niveaux sont fiables pour conclure à une absence de risques immédiats pour la VTE. Si le taux de D-dimères est élevé, le diagnostic doit être confirmé par imagerie, via une échographie Doppler pour la TVP, ou un angioscanner pour l'EP.

Différentes mesures sont également effectuées afin de mesurer l'efficacité de la coagulation, notamment le temps de céphaline activée (ou aPTT en anglais pour *activated partial thromboplastine time*), le temps de prothrombine (TP), et le potentiel de génération de thrombine (TGP, en anglais pour *Thrombin Generation Potential*). Le TP permet de déterminer l'efficacité de la voie de coagulation extrinsèque, tandis que l'aPTT permet de mesurer celle de la voie intrinsèque. Enfin, le test TGP permet de mesurer en temps réel la quantité de thrombine générée après activation de la coagulation, et permet d'établir les caractéristique *lag time* (le temps nécessaire au début de la génération de thrombine), *thrombin peak* (la quantité maximale de thrombine générée) et l'*ETP* (ou *Endogenous Thrombin Potential* qui correspond à la quantité totale de thrombine générée).

Pour déterminer si un élément singulier de la coagulation est déficient, d'autres paramètres de l'hémostase sont également fréquemment mesurés, comme la quantification des plaquettes, du vWF, ou des différents acteurs de la coagulation et de la fibrinolyse. Un dépistage des facteurs génétiques à risques peut également être effectué, notamment chez les cas de VTE non provoquée.

Afin de réduire le risque de développement de thrombus, des traitements à base d'anticoagulants sont généralement prescrits tels que l'héparine [4], les anti-vitamines K [5], ou les inhibiteurs du FXa et du FIIa. Cette approche est efficace pour réduire le risque de VTE, mais peuvent augmenter le risque d'hémorragie. La recherche de nouveaux biomarqueurs robustes, d'anticoagulants diminuant le risque de saignements, et de nouvelles cibles moléculaires permettant un traitement moins risqué, font partie des objectifs principaux de la recherche sur la VTE.

1.3 microARNs et hémostase

Plusieurs miARNs impliqués dans les mécanismes de l'hémostase ont été identifiés [7]. Ils seraient impliqués dans la régulation de l'activité des plaquettes, et dans la régulation de l'expression de plusieurs éléments de la cascade de la coagulation. Ils pourraient donc avoir un rôle dans le développement de pathologies associées aux troubles de l'hémostase, comme une hémorragie ou une thrombose, et ils pourraient avoir une utilité en tant que biomarqueurs pour ces pathologies, par exemple pour prédire le risque de développement de la VTE ou le risque de récidive.

microARNs plaquettaires Les premiers miARNs impliqués dans les mécanismes de l'hémostase sont issus des plaquettes [159]. Les plaquettes sont des cellules anucléées issues du cytoplasme des mégakaryocytes, et contiennent ainsi des miARNs produits par les mégakaryocytes. Différents travaux suggèrent un rôle pour ces miARNs dans la régulation de l'activité plaquettaire et l'hémostase.

4. L'héparine amplifie l'activité de l'Antithrombine
5. Les vitamines K sont des molécules requises pour l'activation de plusieurs facteurs de la coagulation (II, IX, VII et X).

Afin d'étudier l'impact des miARNs sur l'activité plaquettaire, un KO de Dicer a été réalisé dans les mégakaryocytes d'un modèle murin [159]. Plusieurs conséquences ont été observées, parmi lesquelles une augmentation de recepteurs plaquettaires impliqués dans la formation de thrombus, une augmentation de la réactivité [6] des plaquettes, ainsi que des temps de saignement plus courts.

Plusieurs miARNs plaquettaires ont par la suite été investigués afin de déterminer leurs rôles dans l'hémostase primaire. Par exemple, les miR-223-3p et miR-126-3p, qui sont présents dans des quantités particulièrement élevées dans les plaquettes, ont été identifiés comme régulateurs de l'activité plaquettaire: un KO du miR-223-3p dans un modèle murin entraîne une réponse plaquettaire amplifiée en présence de thrombine ou de collagène [70], tandis qu'un KO du miR-126-3p entraîne une réduction de l'aggrégation de plaquettes en modifiant l'expression de plusieurs récepteurs plaquettaires [132].

De plus, les plaquettes génèrent également des microparticules (MP) lors de leur activation, qui participent à la formation du thrombus, notamment grâce à des protéines d'adhésion celullaire (ex: Fibrinogène, Fibronectine, vWF, Thrombospondine, Vitronectine, Sélectine P) et des facteurs de coagulation (ex: Facteurs V, XI, XIII, PAI-1, TFPI, Antithrombine, Plasminogène, ou Protéine S), mais qui peuvent également contenir des miARNs. Si un nombre important de MPs relâchent leur contenu dans le plasma, plusieurs travaux *in vitro* suggèrent qu'une partie des MPs pourrait être absorbée par des cellules distantes, en particulier les cellules endothéliales [296], permettant ainsi à des miARNs contenus dans ces MPs de réguler des messagers ciblés dans ces cellules [259]. D'ailleurs, la protéine VAMP8, qui participe à la production de ces MPs, serait elle même directement régulée par le miR-96-5p [152]. Les plaquettes seraient d'ailleurs des contributeurs importants de miARNs circulants dans le plasma [289], en particulier les miARNs miR-223-3p, miR-191-5p, miR-126-3p et miR-150-5p qui sont très abondants dans les plaquettes. Ces miARNs circulants pourraient ainsi correspondre à des biomarqueurs de l'activité plaquettaires.

Interactions de miARNs avec les gènes impliqués dans l'hémostase L'impact des miARNs dans l'hémostase est encore peu connu, mais de nombreuses interactions miARN:ARNm impliquant des gènes responsables de l'hémostase ont été identifiées. Les recherches d'interactions à grande échelle ont permis d'identifier plus d'un millier d'interactions de miARNs avec des gènes de l'hémostase. Cependant ces résultats manquent de fiabilité et ne permettent pas de déterminer les interactions fonctionnelles (comme mentionné dans les sections I.2.2.c et I.4.2.a), c'est-à-dire celles qui résultent en une diminution de l'expression de gènes ciblés. De nombreux travaux à plus petite échelle, à partir de tests par Luciférase, centrés sur les gènes impliqués dans l'hémostase, ont permis d'identifier une cinquantaine d'interactions fonctionnelles *in vitro* dont la liste est présentée dans la table IV.2.

Récemment, une étude à moyenne échelle a été menée afin d'identifier les interactions fonctionnelles entre des miARNs et de nombreux gènes impliqués dans l'hémostase [213]. Grâce à la combinaison de tests par Luciférase et d'une technique appelée miTRAP [7] [32], ce travail a permis d'identifier une cinquantaine de nouvelles interactions fonctionnelles avec des gènes de l'hémostase, dont un extrait est présenté dans la table IV.3.

Toutefois, si des cibles de miARNs ont été identifiées parmi les gènes impliqués dans l'hémostase, l'impact de cette régulation sur l'hémostase, notamment en termes de thrombose

6. Propention des plaquettes à être activées

7. Cette technique permet la co-purification d'un ARNm spécifique et des miARNs qui interagissent avec ce dernier. Les miARNs peuvent ensuite être identifiés par séquençage.

ARNm (Protéine)	miARN
F3 (FIII)	miR-106b, miR-126, miR-19a/b, miR-20a/b, miR-223, miR-520g, miR-93
F11 (FXI)	miR-181a, miR-145, miR-544
FGA (Fibrinogène chaîne A)	miR-759, miR-29c
FGB (Fibrinogène chaîne B)	miR-409-3p
PLAT (Activateur du Plasminogène, Tissulaire)	miR-133a, miR-144, miR-21-5p, miR-340
PLAU (Activateur du Plasminogène, Urokinase)	miR-181c, miR-193a/b-3p, miR-198, miR-23b-3p
PROS (Protéine S)	miR-494
SERPINE1 (PAI-1)	miR-10a, miR-143-3p, miR-145-5p, miR-148a-3p, miR-192-5p, miR-301a-3p, miR-30b/c-5p, miR-421, miR-99a-5p
TFPI (TFPI)	miR-27a/b-3p, miR-494, miR-500
THBS1 (Thrombospondine-1)	let-7g-5p, miR-126-5p, miR-182-5p, miR-194, miR-200a, miR-27b-3p, miR-467, miR-487b, miR-let-7f
VWF (vWF)	miR-24

Table IV.2 – 49 interactions miARN:ARNm liées à l'hémostase identifiées par test Luciférase dans différente études (le bras effectif du miARN n'est pas toujours communiqué)

ou d'hémorragie, est encore peu connu.

miARNs et risque de VTE Deux études basées sur des cohortes de patients atteints de VTE ont récemment identifiées des miARNs dont les niveaux d'expression dans le plasma sont associés au risque de développer une thrombose, et au risque de récidive.

La première, menée par l'équipe de Hansen [257] a évalué la différence d'expression de 97 miARNs plasmatiques, mesurés par RT-qPCR, entre un groupe de 20 patients ayant développé une VTE (non provoquée) dans les 5 années précédant l'étude, et un groupe de 20 contrôles n'ayant jamais développé de VTE. Cette étude a identifié 9 miARNs avec une différence d'expression entre les deux groupes ($p < 0.05$ sans correction pour le nombre de tests effectués):

miARNs sur-exprimés chez les cas miR-10b-5p, miR-320a, miR-320b, miR-424-5p, miR-423-5p

miARNs sous-exprimé chez les cas miR-103a-3p, miR-191-5p, miR-301a-3p, miR-199b-3p

La seconde étude, menée par le groupe de Zöller [278] a récemment identifié des miARNs associés au risque de récidive de la VTE. Ils ont évalué la différence d'expression de 110 miARNs plasmatiques, mesurés par RT-qPCR, entre un groupe de 39 patients ayant eu plusieurs épisodes thrombotiques, et un groupe de 39 patients n'ayant eu qu'un seul épisode.

ARNm (Protéine)	**miARN**
F11 (FXI)	miR-103a-3p, miR-1255a, miR-148b-3p, miR-151a-3p, miR-15b-5p, miR-181b-5p, miR-24-3p, miR-30a/d-3p, miR-96-5p
F7 (FVII)	miR-19a/b-3p
F8 (FVIII)	miR-18a-5p, miR-30e-3p, miR-34a-5p, miR-454-3p, miR-532-5p, miR-7-5p, miR-874-3p
FGA (Fibrinogène chaîne A)	miR-193b-3p, miR-194-5p
FGG (Fibrinogène chaîne B)	miR-151a-5p, miR-193a-5p, miR-452-5p, miR-99b-3p
SERPINC1 (Antithrombine)	miR-186-5p, miR-19b-3p

TABLE IV.3 – 27 interactions miARN:ARNm identifiées dans une étude ayant utilisé une combinaison de miTRAP et de tests par Luciférase

Ils ont ainsi identifié 12 miARNs associés au risque de récidive (en utilisant une correction FDR[8] de 25%):

miARNs sur-exprimés chez les cas miR-15b-5p, miR-222-3p, miR-26b-5p, miR-532-5p, miR-21-5p, and miR-30c-5p

miARNs sous-exprimé chez les cas miR-106a-5p, miR-197-3p, miR-652-3p, miR-361-5p, miR-27b-3p,miR-103a-3p

Ces résultats sont encourageants pour la découverte de biomarqueurs liés au risque de VTE et de récidive. Cependant, les cohortes impliquées dans ces études sont de petites tailles, ce qui entraîne un manque de puissance pour obtenir des associations significatives robustes. Le seuil de significativité de ces études est par conséquence relativement laxiste, et des réplications dans des cohortes indépendantes sont nécessaires pour valider ces résultats.

2 Matériel et méthodes: Associations des niveaux plasmatiques des microARNs avec des variables cliniques et biologiques

Pour ce projet de thèse, j'ai mené des analyses à partir des niveaux d'expression des miARNs circulants quantifiés avec optimiR. Dans un premier temps, j'ai effectué une recherche des variants génétiques pouvant influencer ces niveaux d'expression. Par la suite, j'ai mis en oeuvre une analyse permettant d'identifier les miARNs associés au risque de récidive de VTE, et à différents paramètres impliqués dans l'hémostase. Les résultats de ces analyses ont fait l'objet de l'article *Bayesian network analysis of plasma microRNA sequencing data in patients with venous thrombosis* accepté dans la revue *European Heart Journal Supplement*. Cette section présente brièvement le matériel et les méthodes utilisés lors de cette étude, et les résultats seront discutés dans la dernière section.

8. Le False Discovery Rate, ou correction de Benjamini-Hochberg pour les tests multiples, permet de limiter le pourcentage de faux positifs parmi les résultats significatifs (avec $p < 0.05$).

2.1 Matériel: La cohorte MARTHA

La cohorte MARTHA est composée de patients non apparentés d'origine européenne ayant développé une ou plusieurs VTE. Cette cohorte a été créée notamment pour identifier de nouveaux déterminants moléculaires et génétiques de la maladie, afin d'améliorer les stratégies de diagnostic et de traitement, et pour découvrir des biomarqueurs fiables permettant d'évaluer le risque de survenue et de récurrence de la maladie.

Pour chaque patient de la cohorte MARTHA, un historique des incidents thrombotiques a été réalisé, et des échantillons sanguins ont été prélevés. Aucun des participants ne possède les facteurs de risques génétiques les plus importants, incluant les déficience en antithrombine, protéine C et protéine S, ou de variant homozygote pour le F5 Leiden ou le F2 G20210A.

Dans le cadre de ce projet, les échantillons plasmatiques de 435 participants ont été sélectionnés pour le séquençage de petits ARNs.

Données génétiques Le profil génétique d'environ 1500 participants de MARTHA a été obtenu avec des puces à ADN, permettant de génotyper environ 600,000 variants génétiques. Une imputation [9] des variants manquants a été effectuée par Marine Germain, ingénieure de recherche au sein de l'équipe menée par David-Alexandre Trégouët, à partir des données de 1000 génomes [264], pour inférer le génotype d'environ 6,000,000 de variants supplémentaires.

Suivi de la récidive d'incident thrombotique Un sous ensemble des participants de l'étude MARTHA fait l'objet d'un suivi clinique, permettant ainsi de recenser les récidives d'incidents thrombotiques depuis l'entrée des patients dans la cohorte. De plus, les échantillons sanguins ont été collectés au moment de l'entrée du patient dans la cohorte, à la suite d'un évènement thrombotique. On pourra donc estimer la valeur prédictive des niveaux plasmatiques de miARNs extraits dans ces échantillons par rapport au risque de récidive.

Mesures des paramètres biologiques de l'hémostase Différentes mesures ont été réalisées chez une majorité des participants à l'étude. En particulier les tests d'aPPT, de TP et de TGT ont été effectués, et plusieurs facteurs impliqués dans l'hémostase ont été mesurés:

- les facteurs de coagulation FI, FV, FVIII et FXI
- les facteurs anti-coagulation Protéine C, Protéine S, Antithrombine et TFPI
- les facteurs régulants la fibrinolyse PAI-1 et TAFI
- le vWF

Les D-dimères ont également été mesurés, ainsi que les quantités moyennes des différents composants du sang (érythrocytes, plaquettes, et leucocytes).

2.2 Méthodes: Recherche d'associations avec les niveaux plasmatiques des microARNs

a) Normalization des données de miARNs quantifiées

Parmi les 435 échantillons plasmatiques séquencés pour les petits ARNs, 391 ont également des données génétiques disponibles. Le pipeline optimiR a donc été utilisé pour quantifier les

9. Une imputation des variants génétiques permet d'inférer le génotype de variants par analyse haplotypique. Car les variants sont souvent associés de façon préférentielle avec d'autres variants, on dit qu'ils sont alors en déséquilibre de liaison (LD). Ils forment ainsi des structures haplotypiques, et le génotype d'un ou plusieurs variants au sein d'un haplotype premet d'inférer celui des autres variants de l'haplotype en LD.

miARNs de ce sous ensemble de 391 individus. Certains échantillons ont été retirés de l'étude, soit parce qu'ils présentaient des signes d'hémolyse ($n = 34$), soit parce qu'ils possèdent un nombre particulièrement faible de reads séquencés alignés sur des miARNs ($n = 3$ échantillons possèdent moins de 100,000 reads alignés).

Les abondances de miARNs obtenues avec optimiR sont appelées données brutes, et possèdent une grande variabilité entre les échantillons, principalement à cause du protocole d'extraction et de séquençage. Afin de pouvoir comparer les échantillons entre eux, une étape de normalisation est donc nécessaire. Cette étape a été réalisée avec la méthode `rlog` de la librairie `DESeq2` [183]. Seuls les miARNs exprimés avec au moins 5 CPM [10] dans 75% des échantillons restants ont été considérés pour la normalisation. Enfin, après normalisation, 10 échantillons *outliers* [11] ont été détectés par PCA (en anglais pour *Principal Component Analysis*) et ont été retirés de l'étude.

Au final, les niveaux normalisés d'expression de 162 miARNs ont été obtenus pour 344 échantillons. Ces données seront utilisées pour les études d'association pangénomiques et les études d'association avec les variables biologiques de l'hémostase.

b) Associations génétiques

Pour chacun des 162 miARNs, une GWAS a été effectuée afin de trouver des variants susceptibles de modifier leurs niveaux d'expression. Les variables d'ajustement suivantes ont été prises en compte pour ces analyses: âge, sexe, prise d'anticoagulant, et une combinaison de 3 miARNs mesurés par PCR (let-7d/g/i-5p) qui reflètent la quantité totale de miARNs dans l'échantillon [48]. Pour chaque étude GWAS, plus de 6,000,000 de variants sont testés, le seuil de significativité utilisé est donc généralement de 5×10^{-8}. Comme 162 miARNs sont testés, une correction de type Bonferonni a été appliquée pour baisser ce seuil à 3.2×10^{-10}. Toutefois, les associations au seuil "classique" ont également été étudiées.

Réplication et méta-analyse Récemment, une analyse similaire a été menée par le groupe canadien de Nikpay [212], pour 143 miARNs plasmatiques chez 710 individus sains, et les résultats de ces analyses ont été rendus publiques sur la plateforme `zenodo.org`. Nous avons donc récupéré ces résultats, qui ont dans un premier temps été utilisés pour répliquer les associations identifiées dans MARTHA, puis de façon réciproque pour répliquer dans MARTHA les associations identifiées par cette étude. Ensuite, les résultats des deux études ont été combinés dans une méta-analyse avec effets aléatoires, afin d'augmenter la puissance de l'analyse et découvrir des variants significatifs additionnels. Pour cette analyse, nous émettons donc l'hypothèse que les effets génétiques sur les niveaux plasmatiques des miARNs observés dans MARTHA sont similaires chez des individus sains.

c) Associations avec la récidive de thrombose veineuse et les paramètres de l'hémostase

Analyse de survie pour l'association de miARNs avec le risque de récidive Parmi les 344 échantillons sélectionnés pour les analyses d'association, 228 proviennent de patients qui ont acceptés de participer à un suivi clinique, permettant de recenser la survenue

10. *Counts per Millions* en anglais, correpondant à: $miRNA_{al}$ / ($total\_miRNAs_{al}$ / 1,000,000), avec $miARN_{al}$ le nombre de reads alignés sur un miARN et $total\_miARNs_{al}$ le nombre total de reads alignés

11. Échantillons dont les niveaux d'expression de miARNs sont significativement différents des autres. Ces outliers ont donc un profil de miARNs particulier, et doivent d'être étudiés indépendemment.

d'un nouvel épisode thrombotique. Pendant ce suivi, 41 patients parmi les 228 ont subit une récidive. Afin d'établir l'association entre le niveaux d'expression des miARNs et le risque de récidive chez ces 228 individus suivis, une analyse de survie basée sur un modèle de Cox a été mise en oeuvre par Gaëlle Munsch, lors de son stage de M2 sous la direction de David-Alexandre Trégouët. Ce modèle permet de prendre en compte l'âge de l'individu lors de la collecte de l'échantillon, ainsi que son âge lors de la récidive. Les variables d'ajustement âge, sexe, indice de masse corporelle et le statut fumeur / non fumeur ont été prises en compte dans ce modèle.

Corrélations entre les niveaux plasmatiques des microARNs et les variables biologiques de l'hémostase Les données dont nous disposons ne permettent pas d'établir des interactions directes entre les miARNs et les gènes impliquées dans l'hémostase. On peut néammoins déterminer les corrélations entre les niveaux d'expression des miARNs plasmatiques et les variables de l'hémostase disponibles pour cette étude. Ces analyses de corrélation permettent de suggérer des relations pouvant exister entre les miARNs circulants et les variables de l'hémostase. En particulier, une corrélation entre les niveaux d'expression d'un miARN et les niveaux d'une protéine circulante (comme les facteurs pro et anti-coagulation) peut suggérer l'implication du miARN, directe ou indirecte, dans la régulation de cette protéine. Tandis qu'une corrélation avec les différentes mesures déterminant l'efficacité de coagulation (aPPT, PT et TGT) permettent de suggérer l'implication d'un miARN dans le processus de la coagulation.

Réduction du nombre de tests par réseaux bayésien Lors des analyses d'associations entre les niveaux d'expression des 162 miARNs et la récidive, ou les paramètres de l'hémostase, une correction de la p-value doit être effectuée par rapport au nombre de tests réalisés, afin de diminuer le nombre de potentiels faux positifs, qui correspondent aux associations significatives obtenues par chance. Avec une correction de type Bonferonni, le seuil de la p-value est réduit à 3.1×10^{-4} pour 162 tests. Cependant, en considérant la taille modeste de la cohorte de cette étude, un manque de puissance statistique risque d'entraver les chances de trouver des associations significatives à ce seuil.

Pour cette étude, nous proposons une stratégie originale permettant de réduire le nombre de miARNs testés, qui repose sur la construction de réseaux bayésiens. Cette stratégie a été mise en oeuvre par Gaëlle Munsch. La construction de réseaux bayésiens permet de représenter graphiquement les relations d'influences entre plusieurs variables, au sein d'un graphe orienté acyclique, dont les noeuds correspondent aux variables étudiées, dans notre cas les niveaux d'expression des miARNs. Les relations d'influences entre les noeuds sont modélisées par des arcs orientés, dont l'orientation peut être forcée à l'aide d'informations dites *à priori*. Comme il n'existe pas de relation connue entre les miARNs, aucun *à priori* n'est fourni pour la construction du réseau. Les noeuds terminaux, qui sont influencés par d'autres noeuds mais qui n'influencent aucun noeuds, intègrent donc l'effet cumulé des noeuds parents.

Le réseau bayesien construit à partir des niveaux d'expressions de miARNs est ainsi structuré en plusieurs sous-réseaux, au sein desquels les niveaux de miARNs possèdent des liens d'influence, et dont le miARN terminal cumule l'information du sous-réseau. Les miARNs terminaux sont donc les plus intéressants pour une étude d'association avec une variable d'intérêt.

Cette approche non supervisée, décrite plus en détails dans la publication, nous a permis d'identifier 15 miARNs terminaux, qui ont été utilisés pour les études d'association avec la récidive et les variables de l'hémostase. Cette stratégie originale permet de réduire le nombre le

tests effectués et d'augmenter les chances d'identifier des associations significatives. Cependant, nous avons tout de même effectué les analyses d'association avec l'ensemble des miARNs. D'abord parce que cette approche de réduction est expérimentale et mérite d'être reproduite dans une cohorte indépendante pour vérifier que les miARNs terminaux sont consistants. Ensuite pour pouvoir répliquer les résultats observés dans d'autres études, en particulier les résultats de l'étude menée par le groupe de Zöller sur la récidive. Et enfin pour générer une ressource combinant l'ensemble des analyses d'associations, pouvant être utile à de futur travaux sur l'implication des miARNs dans l'hémostase.

3 Article: *Bayesian network analysis of plasma microRNA sequencing data in patients with venous thrombosis*

Cet article a été accepté pour une publication dans la revue *European Heart Journal Supplements* (Oxford University Press).

Bayesian network analysis of plasma microRNA sequencing data in patients with venous thrombosis

Florian Thibord[1,2], Gaëlle Munsch[1], Claire Perret[3], Pierre Suchon[4], Maguelonne Roux[3], Manal Ibrahim-Kosta[4,7], Louisa Goumidi[7], Jean-François Deleuze[5,6], Pierre-Emmanuel Morange[4,7,*], David-Alexandre Trégouët[1,*], on behalf of the GENMED consortium[1]

[1]Bordeaux Population Health Research Center, UMR_S 1219, INSERM, University of Bordeaux, 33076 Bordeaux, France
[2]Sorbonne-Université, Pierre Louis Doctoral School of Public Health, Paris, France.
[3]Sorbonne Universités, Université Pierre et Marie Curie (UPMC Univ Paris 06), INSERM UMR_S 1166, 75013 Paris,France
[4]Laboratory of Haematology, La Timone Hospital, Marseille, France.
[5]Centre National de Recherche en Génomique Humaine, Direction de la Recherche Fondamentale, CEA, 91057 Evry, France
[6]CEPH, Fondation Jean Dausset, Paris, France
[7]INSERM UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, Center for CardioVascular and Nutrition research (C2VN), Aix-Marseille University, Marseille, France
[*]Contributed equally to this work

Corresponding author: David-Alexandre Trégouët ; phone: +33 5 47 30 42 54 ; mail: david-alexandre.tregouet@u-bordeaux.fr; david-alexandre.tregouet@inserm.fr

Abstract

MicroRNAs (miRNAs) are small regulatory RNAs participating to several biological processes and known to be involved in various pathologies. Measurable in body fluids, miRNAs have been proposed to serve as efficient biomarkers for diseases and/or associated traits. We here performed a next-generation-sequencing based profiling of plasma miRNAs in 344 patients with venous thrombosis (VT) and assessed the association of plasma miRNA levels with several haemostatic traits and the risk of VT recurrence. Among the most significant findings, we detected an association between hsa-miR-199b-3p and hematocrit levels (p=0.0016), these two markers having both been independently reported to associate with VT risk. We also observed suggestive evidence for association of hsa-miR-370-3p (p=0.019), hsa-miR-27b-3p (p=0.016) and hsa-miR-222-3p (p=0.049) with VT recurrence, the observations at the latter two miRNAs confirming the recent findings of Wang et al. (Clin Epigenetics 2019). Besides, by conducting Genome Wide Association Studies on miRNA levels and meta-analyzing our results with some publicly available, we identified 21 new associations of SNP with plasma miRNA levels at the statistical significance threshold of $p<5\times10^{-8}$, some of these associations pertaining to thrombosis associated mechanisms.
In conclusion, this study provides novel data about the impact of miRNAs' variability in haemostasis and new arguments supporting the association of few miRNAs with the risk of recurrence in patients with venous thrombosis.

1 Introduction

Venous thrombosis (VT), including deep vein thrombosis (DVT) and pulmonary embolism (PE), affects about 1,200,000 individuals each year in Europe and is thus the third most common cardiovascular disease after coronary artery disease and stroke.[1] It is a severe disorder that leaves many patients (25 to 50%) with a debilitating post-thrombotic syndrome[2] and whose PE manifestation kills many of them (6% acute, 20% after one year).[3]

About 50% of VT are unprovoked, i.e., they occur without clear external factors like surgery, trauma, immobilization, hormone use or cancer. The annual recurrent rate is 6% and about 25% of patients with unprovoked VT will face a recurrent event after a six-month course of anticoagulant treatment.[4] Thus, the secondary prevention of VT in this specific population group of patients with a first unprovoked VT is a major health issue.

There is an urgent need to better understand the pathophysiological mechanisms leading to VT in order to develop targeted therapeutic and preventative strategies to save lives, improve quality of life and reduce health care costs. Effective preventative options are available in the form of anticoagulant treatments, but these are associated with major bleeding complications. There are unmet needs to develop predictive biomarkers with high sensitivity and specificity for accurate identification of patients who will develop a recurrence, to avoid unacceptably high risk of bleeding complications in patients at low risk of recurrence. Indeed, preventing thrombosis without inducing bleeding is the holy grail of anticoagulant therapy. Currently, there are no commercially available anticoagulants that achieve this.

Predicting the risk of recurrence as well as discriminating between fatal (PE) and non fatal (DVT) events in unprovoked VT patients remain challenging. There is so far no established biomarkers that serve these aims, even if D-dimers measurement has been proposed[5] but lacks specificity. We here propose a comprehensive microRNA profiling from plasma samples of VT patients aimed at discovering microRNA derived biomarkers discriminating between PE and DVT, and associated with VT recurrence. MicroRNAs (miRNAs) represent a class of small ($\sim$22 nucleotides) noncoding RNAs that participate in genes post-transcriptional regulation.[6] It is now well established that miRNAs are involved in the development of human diseases, in particular cardiovascular ones.[7] Several genes participating to thrombosis associated mechanisms have already been suspected to be subject to miRNA regulation.[8–11] So far, epidemiological studies looking for association of plasma miRNAs with VT outcomes are still sparse. Using plasma samples of 20 VT cases and 20 healthy individuals, Starikova et al. assessed the association of 97 miRNAs with VT risk among which 9 were found significantly ($p < 0.05$) associated with the outcome.[12] As for Wang et al.,[13] by looking for the association of 110 miRNAs with the risk of VT recurrence in plasma samples of 39 cases and 39 controls, twelve miRNAs were identified. None of these observations, that were obtained on miRNA data profiled using RT-qPCR techniques, have yet been replicated.

Briefly, we here performed plasma miRNA profiling in 391 VT patients using a next-generation sequencing technology and assessed the association of identified miRNAs with several haemostatic traits and VT associated clinical outcomes. Association analyses were conducted using an original Bayesian Network inference strategy aimed at identifying miRNAs with the highest abilities to serve as relevant biomarkers. In addition, we integrated genome wide genotype data with miRNA expression levels in order to identify miRNAs that are under a strong genetic control.

2 Materials & Methods

2.1 The MARTHA miRNA sequencing study

The MARseille THrombosis Association project refers to a collection of VT patients recruited at the La Timone Hospital in Marseille, France, initially between 1994 and 2005 and further extended over the 2010-2012 period. Detailed description of this collection has already been previously provided.[14]

The present study relies on a subsample of 391 VT patients that had been previously genotyped for genome-wide polymorphisms using dedicated genotyping array[15,16] and with available plasma samples. For each sample, total RNA was extracted from 400 μL citrate plasma sample using miRNeasy Serum/Plasma kit from Qiagen. From 6 μL of total RNA, plasma miRNA libraries were then prepared with NEBNext Multiplex Small RNA Library Prep Set for Illumina. The manufacturer's protocol was followed, with an optimized size selection method via Ampure XP beads, a specific dilution of adapters to 1/10, and 15 cycles of PCR amplification, using adapter sequences GATCGGAAGAGCACACGTCTGAACTCCAGTCAC and CGACAGGTTCAGAGTTCTACAGTCCGACGATC for 3' and 5' ends respectively. Detailed characteristics of the experimental protocol for libraries preparation and sequencing have already been described.[17]

2.2 miRNA alignment and quantification processes

Sequenced data were processed with the bioinformatic OptimiR pipeline[17] in order to detect and quantify miRNAs. Briefly, OptimiR aligned miRNAs to a library composed of mature miRNA references sequences from miRBase 21.[18] For miRNA integrating genetic variants in their sequence (called polymiRs), the reference library was upgraded by OptimiR with sequences integrating alternative alleles. Ambiguous alignments were resolved using a scoring algorithm that keeps only the most likely

alignment while considering the frequent post transcriptional modifications that miRNAs can undergo.[19] Reads aligned on polymiRs were kept if they were consistent with the sample's genotype, otherwise they were discarded.[17]

From the resulting miRNA abundances, we performed several quality assessments in order to discard unreliable data. First, samples that were poorly sequenced, i.e with less than 100,000 reads aligned, were discarded ($n = 3$) as well as samples identified to be hemolyzed ($n = 34$). The degree of hemolysis was determined based on the optical density at 414nm, and values exceeding 0.2 were defined as hemolyzed samples.[20] Finally, in order to retain only highly expressed miRNAs, we kept only those with at least 5 counts in at least 75% of the remaining samples.

Abundances were then normalized using the rlog method from the DESeq2 R library.[21] This normalization process takes into account differences in library sizes due to library preparation and sequencing protocols, and stabilize variance across miRNAs and samples to respect homoscedasticity constraints for further analysis. Principal component analysis (PCA) was applied to normalized abundances in order to identify individuals with outliers miRNA profiles. Individuals deviating by 3 standard deviation from the centers of the first four PCAs ($n = 10$) were further excluded from downstream analyzes, leaving 344 individuals for Bayesian network and association analyses.

2.3 Bayesian Network analysis

A Bayesian Network (BN) is a probabilistic directed acyclic graphical model that represents relationships among a large number of variables (here mainly miRNAs) with the aim of modeling the dependencies/interactions and conditional independencies between variables.[22,23] Generally, any BN is defined by a directed acyclic graph structure $G = (V,E)$ where V is the set of variables and E the set of edges representing the directional relationships between variables and P a joint probability distribution of the variables in the network. Three types of nodes can be identified in a given BN: the root nodes that are variables found to influence several other variables but are not themselves influenced by any other variables, the internal nodes that are both influenced by and modulate other variables, and finally terminal nodes that are variables that are not identified as influencing others (see Figure 1). Any variable influencing another variable in the network is referred to as a parental node for this later variable.

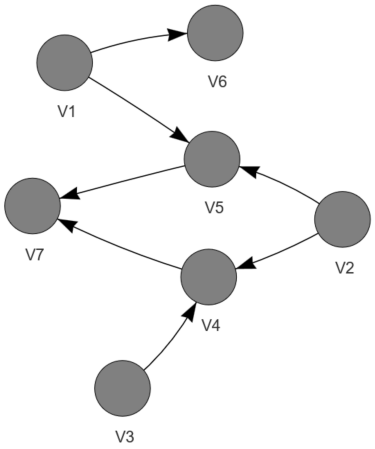


Figure 1 – A bayesian network example

In this illustrative BN example, variables V1, V2 and V3 are root nodes, V4 and V5 are internal nodes and V6 and V7 are terminal nodes. V3 is also a parental node for V4 which is itself a parental node for V7.

In the following, we will mainly focus on terminal nodes assuming that such nodes, as integrating the cumulative upstream effects of other variables, would serve as more relevant and powerful endophenotypes to be tested in relation to some outcomes of interest. In that context, BN analysis can also be viewed as a data reduction technique since, instead of testing the association of all initial variables with a given outcome, only the terminal nodes will be tested for association, reducing then the multiple testing burden. In this work, BNs will be constructed with the «bnlearn» package[24] that implements the relatively fast tabu search algorithm handling both discrete and continuous variables. In the current application, BNs will be created from all expressed miRNAs but also with the age and sex variables. These two latter variables have been shown to have strong influence on circulating miRNA levels[25,26] and their integration in the BN analysis can then add information to more efficiently model the dependencies and conditional independence between some miRNAs.

Because tabu search is a greedy search algorithm, it may end up into a local optimum. To overcome such situation and to assess the stability of the BN analysis in identifying robust terminal nodes, we generated 2,000 bootstrapped datasets composed of 95% of the initial samples and for each boostrapped datasets, we randomly shuffled the way the input variables were ordered in the initial dataset. For each shuffled bootstrapped dataset, a BN was constructed and the terminal nodes identified. After 2,000 bootstrap, we calculated the number of times a given variable was identified as terminal node.

In order to assess whether the observed distribution of the number of terminal node's occurrences deviates from

the null hypothesis of no correlation structure between miRNAs, a permutation strategy was adopted. For each permutation, we randomly selected at least 40 variables whose values were permuted between individuals in order to break down the original data correlation structure. We generated 2,000 of such permuted datasets and constructed a BN on each of them. From these permuted BNs, we counted the maximum number of times a given variable (that could be any miRNA, age or sex) was identified as a terminal node and used this maximum value as a cut off to identify robust terminal miRNAs in the unpermuted analysis above.

2.4 Association analysis with haemostatic traits and clinical outcomes

Identified terminal miRNAs were tested for association with several haemostatic traits available in MARTHA participants (see Table 1). Association analyses were performed using linear regression model and adjusted for age, sex, anticoagulant therapy and combined plasma levels of hsa-let-7d-5p, hsa-let-7g-5p and let-7i-5p measured by qPCR, which serve as a control reference of miRNA levels.[27] Individuals under anticoagulant therapy at the time of blood sampling were excluded for the analysis on protein C, protein S and prothrombin time. For association testing, log-transformation was applied to the following variables: Activated Thrombin Generation Potential biomarkers (Endogenous Thrombin Potential, Lagtime), Partial Thromboplastin Time, Factor VIII, Homocystein, Plasminogen Activator Inhibitor-1, Tissue Factor Principal Inhibitor and von Willebrand Factor.
Terminal miRNAs were also tested for association with the DVT vs PE outcome using a logistic regression model while a Cox model was used to assess their association with VT recurrence whose information was available in 228 patients only. For the latter analysis, we applied the Cox survival model with left truncature[28] and adjusted for age, sex, body mass index and smoking. To address the multiple testing issue associated with the number of terminal miRNAs that will be tested for association with the phenotypes, we applied a Bonferroni correction based on the effective number of independent variables.[29]

2.5 Genome Wide miR-eQTL analysis

As MARTHA participants have been typed for high-density genotyping arrays and imputed for common polymorphisms available in the 1000G reference panel, we performed genome-wide association study (GWAS) on each expressed miRNA for identifying miRNA expression quantitative trait loci (miR-eQTL) using the mach2QTL program.[30] Analyses were performed under the assumption of additive genetic effects and adjusting for the following covariates: sex, age of blood collection, anticoagulant prescription, RT-qPCR measured hsa-let-7 combination,[27] and the 4 first principal genetic components retrieved from PCA analysis as previously described.[15,16] GWAS results were filtered out for variants with minor allele frequency lower than 0.05 and with imputation criterion r^2 below 0.4. Finally, we combined the results of our miR-eQTL analysis with those previously described by Nikpay et al.[31] and available at zenodo.org/record/2560974 in order to identify additional SNP $\times$ miRNA associations. For this, a random-effect model based meta-analysis was adopted as implemented in the GWAMA software.[32] SNP $\times$ miRNA associations were considered as cis effects when the SNP maps $\pm$ 1 Mb from the mature miRNA position. Otherwise, they were considered as trans. Any association with *p-value* $< 3.2 \times 10^{-10}$ corresponding to the Bonferroni threshold corrected for the number of tested SNP $\times$ miRNA associations was considered as genome-wide significant. We also used a miRNA-wide threshold of $p < 5 \times 10^{-8}$, the standard statistical threshold generally advocated in the context of a single GWAS, to identify additional suggestive associations.

3 Results

3.1 The MARTHA miRNA cohort

Detailed description of the clinical and biological characteristics of the 344 participants is shown in Table 1. Of note, 228 patients have been followed for the risk of recurrence for a mean time period of 11.4 $\pm$ 4.3 years. During this period, 41 patients experienced a new VT event.

After the application of the OptimiR workflow, 162 miRNAs were found expressed in the 344 MARTHA participants. Full miRNA data are provided in Supplementary Table 1. The most expressed miRNA was the hsa-miR-122-5p (Supplementary Figure 1), a miRNA known to be mainly expressed in liver and that was previously shown to be amongst the most abundant plasma miRNAs.[33] Additional highly expressed miRNAs were hsa-miR-486-5p, hsa-miR-92a-3p and hsa-miR-451a (Supplementary Figure 1). Of note, the 25 most expressed miRNAs accounted for more than 90% of all sequenced reads that were aligned to miRNA mature sequences.

3.2 BN analysis of miRNA data

Under the null hypothesis of no specific structure in the miRNA data, all miRNAs were identified as a terminal

Table 1 – Characteristics of the MARTHA miRNA cohort

Variables	N	Mean±SD (1)
Gender (Male / Female)	344	144 / 200
Age (years)	344	52.1 ± 14.5
Smoking (Yes/No)	343	94 / 249
BMI (kg/m^2)	331	25.86 ± 4.62
Deep Vein Thrombosis / Pulmonary Embolism	344	259 / 85
Anticoagulant therapy (Yes/No)	344	122 / 222
Antithrombin (IU/ml)	313	102.41 ± 11.59
Activated Partial Thromboplastin Time (sec)	341	33.42 ± 6.02
Ddimers (μg/mL)	184	0.39 ± 0.33 (2)
FV (IU/ml)	150	109.21 ± 22.26
FVIII (IU/dl)	294	135.07 ± 48.31
FXI (IU/ml)	336	130.78 ± 31.99
Fibrinogen (g/L)	342	3.42 ± 0.66
Hematocrit (L/L)	343	0.42 ± 0.03
Homocysteine (μmol/L)	304	12.26 ± 5.65
Platelet count (G/L)	344	254.62 ± 64.91
Mean platelet volume (fL)	344	7.90 ± 0.77
Hemoglobin (g/dL)	344	140.42 ± 13.19
PAI-1 (UI/ml)	272	12.25 ± 13.44
Protein C (IU/ml)	318	99.55 ± 40.56
Protein S (IU/ml)	322	81.3 ± 27.49
TAFI (μg/mL)	336	15.27 ± 4.72
TFPI (ng/ml)	336	14.17 ± 6.84
vWF (IU/dl)	308	154.34 ± 67.74
Prothrombin Time (%)	344	87.63 ± 27.95
Thrombin Generation	193	
• Endogeneous Thrombin Potential (nM.min)		1761.44 ± 280.31
• Peak (nM)		340.35 ± 57.51
• Lagtime (min)		3.34 ± 1.17
VT recurrence during follow-up (Yes/No)	228	41 / 187

(1) Count data are shown for categorical variables, other reported values were mean ± standard deviation.
(2) In about 50% participants, D-dimers values were below the detection limit (0.22) and thus discarded. Mean and SD were then computed over all Ddimer values > 0.22.

node at least once and, on average, a miRNA was found as a terminal node in 6.3% ± 3.5 of the permuted BNs, with a maximum of 18.3%. Using the latter threshold, the bootstrap BN analysis identified 15 terminal miRNAs and the number of times each of them was found as a terminal node in boostrapped BNs is shown in Figure 2.

3.3 Association of miRNAs' levels with VT associated biological and clinical traits

The application of the Li and Ji multiple testing procedure[29] estimated the number of effective independent terminal miRNAs as 14, leading to an adapted Bonferroni threshold of 3.6×10^{-3}. At this statistical level, only one association between terminal miRNAs and haemostatic traits was detected. Plasma levels of hsa-miR-199b-3p was negatively correlated ($\rho = -0.17$, $p = 0.0016$) with hematocrit levels. Interestingly, this miRNA has recently been reported to associate with VT risk[12] whose association with hematocrit levels have already been described.[34,35] The full results of the scan for association between miRNAs and haemostatic traits are given in Supplementary Table 2.

Of note, the strongest association of terminal miRNAs with recurrence risk was observed for hsa-miR-370-3p (HR $= 1.77[1.09 - 2.88], p = 0.019$), this miRNA being also the terminal miRNA that discriminated the most between DVT and PE (OR for PE $= 0.72[0.49 - 1.05], p = 0.090$) (Table 2). Of interest, one of our terminal miRNAs, hsa-miR-197-3, was reported to associate with VT recurrence in Wang et al.[13] However, we did not observe here such trend for association (HR $= 0.78[0.35 - 1.76], p = 0.55$). Nevertheless, among the 9 additional miRNAs reported in Wang et al. and also exressed in MARTHA, we found two with a suggestive association with VT recurrence: hsa-miR-27b-3p (HR $= 0.4[0.2 - 0.79], p = 0.016$) and hsa-miR-222-3p (HR $= 1.76[1.01 - 3.08], p = 0.049$) (Supplementary table 3).

3.4 miR-eQTL analyses

At the pre-specified genome-wide statistical level of 3.2×10^{-10}, 3 SNP × miRNA associations, all cis, were identified in the MARTHA study (Table 3). These were observed for rs12473206 with hsa-miR-4433b-3p ($p = 8.12 \times 10^{-35}$), rs2127870 with hsa-miR-625-3p ($p = 9.57 \times 10^{-26}$) and rs140930133 with hsa-miR-941 ($p = 5.07 \times 10^{-15}$). The latter two have already been observed in whole blood[36] and adipose tissue.[37] Using a more liberal miRNA-wide threshold of $p = 5 \times 10^{-8}$, 10 additional suggestive associations, 1 in cis and 9 in trans, were observed (Table 3). Regional association plots and boxplot summarizing the genotype × miRNA

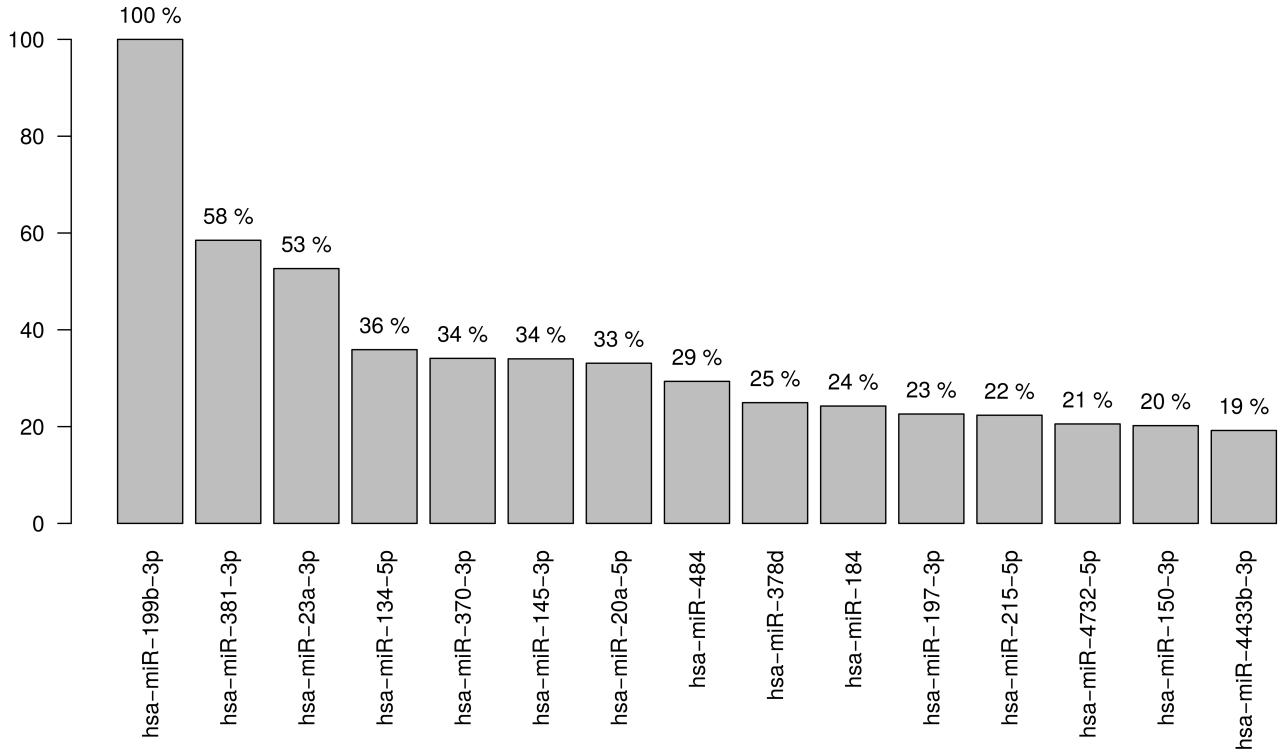


Figure 2 – Percentage of significant terminal miRNAs found in 2000 bootstrapped bayesian networks
The bootstrap BN analysis identified 15 terminal miRNAs with an occurrence percentage over the significance threshold (18.3%) determined by the permutation analysis.

Table 2 – Association of terminal miRNAs with VT outcomes in the MARTHA miRNA study

miRNA	VT recurrence		Pulmonary Embolism vs Deep Vein Thrombosis	
	HR [95%CI]	*p* (1)	OR [95%CI]	*p* (2)
hsa-miR-370-3p	1.77 [1.09 – 2.88]	0.019	0.72 [0.49 - 1.05]	0.090
hsa-miR-184	0.53 [0.30 – 0.95]	0.024	1.23 [0.92 - 1.66]	0.153
hsa-miR-4732-5p	0.41 [0.18 – 0.92]	0.024	0.70 [0.39 - 1.22]	0.218
hsa-miR-4433b-3p	1.54 [1.04 – 2.29]	0.033	1.01 [0.75 - 1.36]	0.930
hsa-miR-215-5p	0.63 [0.37 – 1.09]	0.091	1.11 [0.73 - 1.67]	0.633
hsa-miR-134-5p	1.58 [0.85 – 2.91]	0.142	0.89 [0.57 - 1.39]	0.601
hsa-miR-381-3p	1.45 [0.83 – 2.56]	0.194	0.81 [0.53 - 1.23]	0.327
hsa-miR-145-3p	0.51 [0.15 – 1.76]	0.278	0.62 [0.24 - 1.56]	0.311
hsa-miR-23a-3p	0.67 [0.26 – 1.70]	0.393	1.00 [0.51 - 1.93]	0.999
hsa-miR-197-3p	0.78 [0.35 – 1.76]	0.555	1.41 [0.79 - 2.56]	0.251
hsa-miR-150-3p	1.23 [0.53 – 2.83]	0.629	0.90 [0.49 - 1.66]	0.743
hsa-miR-484	1.20 [0.56 – 2.59]	0.637	1.27 [0.69 - 2.38]	0.447
hsa-miR-199a-3p	0.80 [0.22 – 2.86]	0.726	1.17 [0.46 - 2.97]	0.746
hsa-miR-378d	0.81 [0.15 – 4.56]	0.812	0.41 [0.10 - 1.46]	0.184
hsa-miR-20a-5p	1.09 [0.40 – 2.95]	0.863	0.74 [0.36 - 1.52]	0.411

HR = Hazard Ratio; OR = Odds Ratio; CI = Confidence Interval
(1) *p*-values were obtained from the Likelihood Ratio Test statistic associated with a Cox survival model adjusted for age, sex, BMI and smoking.
(2) *p*-values obtained from a logistic model ajusted for age, sex, BMI and smoking

associations at these 13 main candidates are shown in supplementary materials.

Of note, the most significant association was observed between hsa-miR-4433b-3p and rs12473206, a variant located within the mature miRNA sequence. It can be speculated that this variant impacts the maturation process of the miRNA or its target spectrum, and thus influences its plasma expression levels. In addition, two SNPs with cis effects on miRNA levels (thereafter referred to as cis miSNPs) have been previously found to associate with levels of the protein encoded by the miRNA host gene. In whole blood, the miSNP rs2127870 was reported to influence FUT8 levels,[38] FUT8 being the host gene for hsa-miR-625-3p. Similarly, the DNAJC5 rs2427555 that is in very strong linkage disequilibrium with the miSNP rs140930133 we here found associated with plasma hsa-miR-941 levels, has been reported to influence the expression of DNAJC5 in lymphoblastoid cells.[39] These observations are supportive elements for the observed miSNP associations and would suggest a joint regulation of hsa-miR-625-3p and hsa-miR-941 expressions with those of their host genes as already documented for several miRNAs.[40]
One trans-eQTL located in the long non-coding RNA

Table 3 – Significant associations at the 5×10^{-8} statistical level between SNPs and plasma miRNA levels in the MARTHA miRNA study

miRNA	miRNA host gene	Top SNP Associated	MAF	R^2	Chr	Distance to 5' miRNA / Position (1)	Effect (SD)	*p*	SNP Genomic Context
Cis associations									
hsa-miR-4433b-3p	intergenic	rs12473206	0.23	0.99	2	-13	0.979 (0.080)	8.12×10^{-35}	exonic_ncRNA (hsa-miR-4433b)
hsa-miR-625-3p	FUT8	rs2127870	0.27	0.99	14	141025	0.533 (0.051)	9.57×10^{-26}	intergenic
hsa-miR-941	DNAJC5	rs140930133	0.19	0.97	20	8822	-0.349 (0.045)	5.07×10^{-15}	Intronic (DNAJC5)
hsa-miR-432-5p	RTL1	rs201969986	0.29	0.95	14	177423	-0.346 (0.063)	3.31×10^{-8}	intergenic
Trans associations									
hsa-miR-184		rs144867605	0.07	0.82	11	75957983	0.804 (0.134)	2.02×10^{-9}	intergenic
hsa-miR-654-5p		rs11109171	0.44	0.99	12	98098091	-0.246 (0.042)	3.28×10^{-9}	intergenic
hsa-miR-320c		rs10151482	0.06	0.93	14	41934917	0.427 (0.074)	6.47×10^{-9}	intergenic
hsa-miR-184		rs143007764	0.06	0.65	3	142899139	0.916 (0.161)	1.14×10^{-8}	intergenic
hsa-miR-1-3p		rs73245753	0.12	0.79	4	26292392	0.589 (0.105)	2.31×10^{-8}	intergenic
hsa-miR-330-3p		rs1554362	0.45	0.82	2	101221457	-0.227 (0.041)	2.81×10^{-8}	intronic (LINC01849)
hsa-miR-582-3p		rs4522365	0.13	0.83	15	29964742	0.314 (0.057)	2.91×10^{-8}	intergenic
hsa-miR-4446-3p		chr12:95274192:I	0.09	0.61	12	95274192	-0.492 (0.089)	3.07×10^{-8}	intergenic
hsa-miR-320d		rs12800249	0.05	0.63	11	21240436	0.481 (0.088)	4.33×10^{-8}	Intronic (NELL1)

MAF = minor allele frequency; R^2 = imputation quality criterion
(1) For cis associations, the distance to 5' miRNA is reported, while for trans associations, the hg19 position is reported.

(lncRNA) LINC01849 was associated with hsa-miR-330-3p. The identified trans miSNP, rs1554362, is also an eQTL for the PDCL3 transcript levels in different tissues according to the GTeX database.[41] Another intronic miSNP located in the NELL1 gene was associated with hsa-miR-320d levels. The seven other trans eQTL are located in intergenic regions. We sought to in silico replicate these miSNP associations using the results from Nikpay et al.[31] who scanned for genetic polymorphisms associated with miRNA levels in 710 plasma samples. Unfortunately, as the Nikpay et al. study relied on a genotyping array focusing mainly on coding regions and used a very stringent imputation quality criterion ($r^2 > 0.9$), it was not possible to assess all our candidate associations. Only 4 were testable (hsa-miR-941 $\times$ rs140930133, hsa-miR-432-5p $\times$ rs201969986, hsa-miR-654-5p $\times$ rs11109171, hsa-miR-320c $\times$ rs10151482) among which only the association of rs140930133 with hsa-miR-941 levels replicated ($p = 6.3 \times 10^{-11}$).

Conversely, we looked into the MARTHA results to replicate the 223 miSNP associations that were significantly ($p < 5 \times 10^{-8}$) detected in the Nipkay et al. study. We were able to test 92 of them among which 37 replicated at the nominal level of $p = 0.05$ in MARTHA (Table 4). These involved 29 cis and 8 trans miSNP associations.

Among these 8 trans miSNP associations, three deserve to be highlighted. First, plasma levels of hsa-miR-143-3p were influenced by the intronic ZFPM2 rs4734879, ZFPM2 being a locus reported to associate with venous thrombosis risk42 and platelet function.[42] In MARTHA, plasma levels of hsa-miR-143-3p were negatively significantly correlated with BMI ($\rho = -0.24, p = 3.6 \times 10^{-4}$) and borderline significant with PAI-1 activity levels ($\rho = -0.21, p = 5.3 \times 10^{-3}$) (Supplementary Table 2). Second, hsa-miR-126-3p plasma levels were associated with the rs600038 located in the promoter region of the ABO gene. This polymorphism is in strong linkage disequilibrium (LD) with several other ABO polymorphisms that are known to associate with VT risk, including the rs579459 ($r^2 = 0.99$) tagging for the A1 ABO blood group. In MARTHA, plasma levels of hsa-miR-126-3p were strongly and positively correlated ($\rho \sim 0.20$) with red cells ($p = 1.73 \times 10^{-5}$), lymphocytes ($p = 2.5 \times 10^{-4}$), platelets ($p = 5.9 \times 10^{-4}$) and polynuclear ($p = 6.0 \times 10^{-4}$) (Supplementary Table 2). Third, polymorphisms (rs970280, rs11070216) in the promoter region of the THBS1 gene were found associated with plasma levels of hsa-miR-222-3p. This miRNA has been previously reported to associate with the risk of VT recurrence[13] and has a suggestive association ($p = 0.049$) in our study (Supplementary Table 3), where it positively correlated with antithrombin levels ($\rho = 0.21, p = 8.8 \times 10^{-4}$) (Supplementary Table 2). THBS1 encodes Thrombospondin-1 and is known to be involved in angiogenesis and platelet aggregation.[43,44]

Finally, we performed a random effect meta-analysis of both datasets in order to discover additional miSNPs. At the 5×10^{-8} statistical threshold, we identified 7 new cis and 5 new trans miSNP associations (Table 5). None of these miSNP associations appeared to involve loci with documented link with thrombosis related traits.

Table 4 – Association of SNPs with plasma miRNA levels identified in Nikpay et al (Cardiovasc Res 2019) that nominally replicated (p <0.05) in MARTHA miRNA study

					NIKPAY (N=710)				MARTHA (N=344)				
miRNA	**SNP**	**Chr**	**Position(bp)**	**EA**	**EAF**	**Effect**	**SE**	***p***	**EAF**	**R^2**	**Effect**	**SE**	***p*** (1)
Cis associations													
miR-941	rs2427550	20	62547575	A	0.23	-0.157	0.023	3.96×10^{-11}	0.19	0.99	-0.339	0.044	5.76×10^{-15}
miR-584-5p	rs17795259	5	148416952	C	0.15	0.268	0.018	1.35×10^{-45}	0.15	0.99	0.213	0.043	4.82×10^{-7}
miR-4433b-5p	rs2059631	2	64574682	A	0.43	0.289	0.017	1.57×10^{-56}	0.45	1	0.129	0.029	4.96×10^{-6}
miR-139-3p	rs4944563	11	72316881	C	0.17	0.169	0.026	1.18×10^{-10}	0.14	1	0.182	0.042	6.82×10^{-6}
miR-181a-5p	rs74746864	1	199023240	G	0.11	0.175	0.025	4.12×10^{-12}	0.13	0.95	0.221	0.066	4.27×10^{-4}
miR-425-5p	rs7623513	3	142100428	C	0.15	-0.044	0.007	7.48×10^{-10}	0.12	0.95	-0.166	0.054	1.04×10^{-3}
let-7e-5p	rs2198171	19	52174483	G	0.27	-0.089	0.014	3.10×10^{-10}	0.25	0.97	-0.124	0.043	1.83×10^{-3}
miR-197-3p	rs7355073	1	110129740	T	0.16	-0.078	0.011	1.23×10^{-12}	0.19	1	-0.118	0.041	2.10×10^{-3}
miR-26b-5p	rs12623740	2	219665715	A	0.49	-0.060	0.007	3.37×10^{-18}	0.51	0.99	-0.138	0.051	3.24×10^{-3}
miR-152-3p	rs9910516	17	46183160	A	0.23	0.093	0.016	1.52×10^{-8}	0.27	0.95	0.089	0.033	3.44×10^{-3}
miR-27b-3p	rs10993381	9	97639463	T	0.07	0.170	0.016	2.00×10^{-24}	0.06	0.99	0.148	0.055	3.86×10^{-3}
miR-182-5p	rs2693738	7	129431977	G	0.32	0.115	0.02	2.36×10^{-8}	0.37	0.82	0.166	0.063	4.30×10^{-3}
miR-181a-3p	rs1434282	1	199010721	C	0.27	0.211	0.022	9.03×10^{-21}	0.26	0.98	0.122	0.048	5.57×10^{-3}
miR-181a-5p	rs12125200	1	198992043	A	0.27	0.340	0.013	1.13×10^{-111}	0.24	0.96	0.124	0.049	5.79×10^{-3}
miR-584-5p	rs4147470	5	148528107	T	0.49	-0.131	0.014	7.71×10^{-20}	0.51	1	-0.081	0.032	6.15×10^{-3}
miR-26b-5p	rs833083	2	219336959	T	0.41	-0.076	0.006	3.96×10^{-30}	0.43	0.81	-0.137	0.057	7.96×10^{-3}
miR-181a-5p	rs878254	1	199257141	A	0.48	-0.122	0.015	3.54×10^{-15}	0.49	0.9	-0.104	0.045	0.010
miR-181a-5p	rs2360961	1	199000277	C	0.4	-0.151	0.016	4.39×10^{-20}	0.40	0.94	-0.095	0.043	0.014
miR-30d-5p	rs13282464	8	135707922	T	0.15	0.092	0.007	2.02×10^{-33}	0.17	1	0.047	0.023	0.020
miR-4433b-5p	rs6740438	2	64528086	C	0.13	0.163	0.029	1.78×10^{-8}	0.15	0.98	0.083	0.041	0.022
miR-30d-5p	rs13268530	8	135727196	T	0.15	0.095	0.007	1.68×10^{-35}	0.17	0.99	0.045	0.023	0.024
miR-21-5p	rs2665392	17	57809453	A	0.16	0.059	0.011	3.59×10^{-8}	0.16	0.88	0.078	0.041	0.027
miR-4433b-5p	rs35503140	2	64539015	C	0.21	-0.130	0.022	9.86×10^{-9}	0.19	0.95	-0.071	0.037	0.029
miR-584-5p	rs9325124	5	148248818	A	0.39	-0.085	0.015	7.62×10^{-9}	0.45	1	-0.056	0.031	0.036
miR-181a-5p	rs3861924	1	199121330	A	0.18	0.137	0.02	2.06×10^{-11}	0.20	0.96	0.097	0.054	0.037
miR-1908-5p	rs174561	11	61582708	C	0.3	0.151	0.012	4.76×10^{-31}	0.26	1	0.052	0.030	0.040
miR-151a-3p	rs11167012	8	141968408	A	0.42	0.059	0.006	3.79×10^{-24}	0.40	1	0.061	0.036	0.045
miR-139-3p	rs10898849	11	72269302	T	0.25	0.124	0.022	3.30×10^{-8}	0.27	1	0.054	0.032	0.046
let-7i-5p	rs6581454	12	62934442	G	0.47	0.039	0.006	3.04×10^{-11}	0.44	0.99	0.034	0.021	0.049
Trans associations													
miR-222-3p	rs11070216	15	39817245	T	0.19	-0.067	0.012	4.87×10^{-8}	0.19	0.97	-0.198	0.051	5.06×10^{-5}
miR-222-3p	rs970280	15	39864403	G	0.32	-0.064	0.010	8.79×10^{-10}	0.32	0.94	-0.113	0.042	3.57×10^{-3}
miR-143-3p	rs4734879	8	106583124	G	0.28	0.239	0.031	2.88×10^{-14}	0.24	0.96	0.098	0.038	5.60×10^{-3}
miR-1-3p	rs11906462	20	61158952	T	0.20	0.310	0.033	6.28×10^{-20}	0.23	0.42	0.262	0.116	0.012
miR-320a	rs1443651	2	68569316	G	0.45	-0.036	0.006	7.12×10^{-10}	0.44	1	-0.053	0.028	0.029
miR-16-5p	rs137214	22	35288857	T	0.28	0.041	0.007	1.76×10^{-8}	0.29	0.97	0.088	0.050	0.040
miR-126-3p	rs600038	9	136151806	C	0.21	0.055	0.009	5.95×10^{-9}	0.34	1	0.041	0.024	0.041
miR-320c	rs1443651	2	68569316	G	0.45	-0.031	0.005	2.77×10^{-10}	0.44	1	-0.066	0.039	0.045

R^2 = imputation quality criterion; EA = Effect Allele; EAF = Effect Allele Frequency
(1) One sided test *p*-value

Table 5 – Significant ($p < 5\times10^{-8}$) associations of miSNP with miRNA plasma levels derived from the MARTHA miRNA and Nipkay et al. (Cardiovasc Res 2019) meta-analysis

					MARTHA					Nikpay				Combined (N=1054)			
miRNA	**Chr**	**Position (bp)**	**SNP**	**EA**	**EAF**	**R^2**	**Effect**	**SE**	***p***	**EAF**	**Effect**	**SE**	***p***	p^{het} (1)	**Effect**	**SE**	***p*** (2)
Cis associations																	
miR-181b-5p	1	199257141	rs878254	A	0.485	0.90	-0.054	0.032	0.0916	0.480	-0.071	0.013	1.64×10^{-7}	0.61	-0.069	0.012	3.18×10^{-8}
miR-148a-3p	7	25991977	rs9639523	T	0.375	0.87	-0.081	0.034	0.0191	0.344	-0.072	0.013	2.03×10^{-7}	0.80	-0.073	0.013	8.41×10^{-9}
let-7a-5p	9	96916230	rs10512230	T	0.287	1	0.040	0.031	0.1934	0.315	0.026	0.004	6.49×10^{-8}	0.67	0.027	0.005	2.19×10^{-8}
let-7d-5p	9	97229465	rs4497033	T	0.492	0.99	-0.061	0.036	0.0895	0.463	-0.028	0.005	1.50×10^{-7}	0.36	-0.029	0.005	3.85×10^{-8}
miR-2110	10	115933905	rs17091403	T	0.091	1	-0.141	0.043	1.13×10^{-3}	0.074	-0.103	0.023	9.90×10^{-6}	0.44	-0.112	0.020	4.34×10^{-8}
miR-342-3p	14	100256449	rs8011282	C	0.474	0.99	0.095	0.030	1.39×10^{-3}	0.487	0.067	0.014	5.65×10^{-6}	0.41	0.073	0.013	3.68×10^{-8}
miR-99b-5p	19	52160843	rs11084100	C	0.392	1	-0.067	0.024	5.17×10^{-3}	0.419	-0.065	0.012	1.12×10^{-7}	0.94	-0.066	0.011	1.50×10^{-9}
Trans associations																	
miR-215-5p	2	171402733	rs724806	C	0.252	0.97	0.091	0.057	0.1123	0.326	0.143	0.027	1.44×10^{-7}	0.40	0.134	0.024	4.09×10^{-8}
miR-10b-5p	7	13236107	rs6948643	G	0.264	1	-0.071	0.040	0.0766	0.285	-0.090	0.017	2.84×10^{-7}	0.66	-0.087	0.016	4.62×10^{-8}
let-7d-3p	11	2611449	rs1024164	A	0.133	0.87	-0.083	0.034	0.0147	0.092	-0.065	0.013	7.78×10^{-7}	0.63	-0.068	0.012	3.18×10^{-8}
miR-378a-3p	11	133763476	rs10894759	A	0.317	0.99	0.066	0.028	0.0206	0.296	0.059	0.011	7.86×10^{-7}	0.82	0.060	0.011	3.58×10^{-8}
miR-7-5p	15	41614621	rs7163989	G	0.293	0.99	-0.112	0.041	6.68×10^{-3}	0.278	-0.089	0.016	1.48×10^{-7}	0.61	-0.093	0.016	2.70×10^{-9}

R^2 = Imputation quality criterion; EA = Effect Allele; EAF = Estimated Allele Frequency
(1) *p*-value of the test for heterogeneity between the MARTHA and Nipkay studies
(2) *p*-value of the combined effect obtained through a random effect meta-analysis of the results of both studies

4 Discussion

In this study, we reported the largest investigation to date of miRNA plasma profiling in a cohort of VT patients. Capitalizing on the application of a next generation sequencing technology, known to be more efficient and sensitive to detect and quantify miRNAs compared to microarray or RT-qPCR techniques, we were able to detect 162 highly expressed miRNAs. These miRNAs were then tested for association with several VT related phenotypes including 38 haematological traits and VT recurrence. In order to deal with the correlation between miRNA levels and reduce the multiple testing burden associated with the number of tested miRNAs, we deployed an original Bayesian Network analysis aimed at identifying miRNAs that could serve as more powerful biomarkers for the investigated traits. In addition, as our studied VT patients had been previously typed for genome-wide genotypes, we were able to perform GWAS on each of the 162 miRNAs, and combined our results with some previously obtained in disease-free individuals in order to identify novel associations of common SNPs with plasma miRNA levels.

Several conclusions could be derived from this work. First, we did not identify any miRNA that significantly associated with the risk of VT recurrence. In our study, the miRNA that discriminated the most between patients with or without recurrence, but also between DVT vs PE patients, was the hsa-miR-370-3p. Several works have already reported the involvement of has-miR-370-3p in lipids metabolism[45–48] and one of the most robust target gene for hsa-miR-370-3p is CPT1A[49] whose role in lipid metabolism is also very documented.[50–52] Hsa-miR-370-3p is also predicted to target drug-metabolism genes such CYP2D6 and VKORC1L150 that are related to the warfarin anticoagulant pharmacotherapy. Aside this miRNA, we observed a trend of association with VT recurrence for the hsa-mir-27b-3p and hsa-miR-222-3p that had been previously identified in Wang et al.[13] but these associations ($p = 0.016$ and $p = 0.0495$, respectively) did not survive any multiple testing correction (Supplementary Table 3).

Larger studies would be mandatory to confirm these observations and increase our chance to identify other miRNAs associated with the risk of recurrence in VT patients. Second, we observed several significant associations of miRNAs with haematological traits that deserve further replication in independent studies. One can highlight the significant correlation between hematocrit levels and plasma levels of hsa-miR-199b-3p, a miRNA that has been reported to be associated with VT risk.[12] Third, our miR-QTL study identified about 25 significant ($p < 5 \times 10^{-8}$) associations of SNPs with plasma miRNA levels, of which, to the best of our knowledge, 21 have never been reported, including a dozen of trans associations. These associations could help deciphering the genomic architecture of complex diseases where miRNAs are involved. For example, plasma levels of hsa-miR-143-3p were found to be associated with the rs4734879 mapping to ZFPM2, a gene known to associate with platelet function[42] and VT risk.[53] We also observed a strong association of rs12473206 with plasma levels of hsa-miR-4433b-3p, a miRNA whose serum levels have recently shown to be associated with stroke.[54] The impact of this SNP on stroke risk deserves to be further and deeply investigated. The results of our GWAS on miRNA levels were combined with those obtained by Nipkay et al.[31] and freely available at https://zenodo.org/. However, only SNPs with imputation quality greater than 0.90 are available at this resource, which has hampered our ability to replicate some of the main associations observed in the MARTHA miRNA study. To facilitate future studies aimed at disentangling the genetic regulation of miRNAs, the results of the 162 GWAS performed on miRNA levels in MARTHA will be available for download at https://zenodo.org/.

Altogether, this study produced a rich source of information relating plasma miRNAs and biological/clinical traits associated with VT that could be of great use to generate and/or validate new hypothesis.

Acknowledgments

F.T, G.M and M.G were financially supported by the GENMED Laboratory of Excellence on Medical Genomics (ANR-10-LABX-0013). DA.T was financially supported by the «EPIDEMIOM-VTE» Senior Chair from the Initiative of Excellence of the University of Bordeaux. MiRNA sequencing in the MARTHA study was performed on the iGenSeq platform (ICM Institute, Paris) and supported by a grant from the European Society of Cardiology for Medical Research Innovation. Bioinformatics and statistical analyses benefit from the CBiB computing centre of the University of Bordeaux.

Supplementary data

Supplementary materials, which consist of 3 supplementary tables, are available online. They can also be provided by the corresponding author (contact informations are provided on the first page).

References

[1] Samuel Z. Goldhaber. Venous thromboembolism: epidemiology and magnitude of the problem. *Best Practice & Research.*

Clinical Haematology, 25(3):235–242, September 2012.

[2] Jean-Philippe Galanaud, Manuel Monreal, and Susan R. Kahn. Epidemiology of the post-thrombotic syndrome. *Thrombosis Research*, 164:100–109, April 2018.

[3] Richard H. White. The epidemiology of venous thromboembolism. *Circulation*, 107(23 Suppl 1):I4–8, June 2003.

[4] Paolo Prandoni, Enrico Bernardi, Antonio Marchiori, Anthonie W. A. Lensing, Martin H. Prins, Sabina Villalta, Paola Bagatella, Donatella Sartor, Andrea Piccioli, Paolo Simioni, Antonio Pagnan, and Antonio Girolami. The long term clinical course of acute deep vein thrombosis of the arm: prospective cohort study. *BMJ (Clinical research ed.)*, 329(7464):484–485, August 2004.

[5] Clive Kearon, Sameer Parpia, Frederick A. Spencer, Sam Schulman, Scott M. Stevens, Vinay Shah, Kenneth A. Bauer, James D. Douketis, Steven R. Lentz, Craig M. Kessler, Jean M. Connors, Jeffrey S. Ginsberg, Luciana Spadafora, and Jim A. Julian. Long-term risk of recurrence in patients with a first unprovoked venous thromboembolism managed according to d-dimer results; A cohort study. *Journal of thrombosis and haemostasis: JTH*, April 2019.

[6] David P. Bartel. Metazoan MicroRNAs. *Cell*, 173(1):20–51, March 2018.

[7] David D. McManus and Jane E. Freedman. MicroRNAs in platelet function and cardiovascular disease. *Nature Reviews. Cardiology*, 12(12):711–717, December 2015.

[8] Alexandre Marchand, Carole Proust, Pierre-Emmanuel Morange, Anne-Marie Lompré, and David-Alexandre Trégouët. miR-421 and miR-30c inhibit SERPINE 1 gene expression in human endothelial cells. *PloS One*, 7(8):e44532, 2012.

[9] Ana B. Arroyo, Ascensión M. de Los Reyes-García, Raúl Teruel-Montoya, Vicente Vicente, Rocío González-Conejero, and Constantino Martínez. microRNAs in the haemostatic system: More than witnesses of thromboembolic diseases? *Thrombosis Research*, 166:1–9, 2018.

[10] Carla Y. Vossen, Astrid van Hylckama Vlieg, Raúl Teruel-Montoya, Salam Salloum-Asfar, Hugoline de Haan, Javier Corral, Pieter Reitsma, Bobby P. C. Koeleman, and Constantino Martínez. Identification of coagulation gene 3'UTR variants that are potentially regulated by microRNAs. *British Journal of Haematology*, 177(5):782–790, 2017.

[11] Bengt Sennblad, Saonli Basu, Johanna Mazur, Pierre Suchon, Angel Martinez-Perez, Astrid van Hylckama Vlieg, Vinh Truong, Yuhuang Li, Jesper R. Gådin, Weihong Tang, Vera Grossman, Hugoline G. de Haan, Niklas Handin, Angela Silveira, Juan Carlos Souto, Anders Franco-Cereceda, Pierre-Emmanuel Morange, France Gagnon, Jose Manuel Soria, Per Eriksson, Anders Hamsten, Lars Maegdefessel, Frits R. Rosendaal, Philipp Wild, Aaron R. Folsom, David-Alexandre Trégouët, and Maria Sabater-Lleal. Genome-wide association study with additional genetic and post-transcriptional analyses reveals novel regulators of plasma factor XI levels. *Human Molecular Genetics*, 26(3):637–649, 2017.

[12] Irina Starikova, Simin Jamaly, Antonio Sorrentino, Thorarinn Blondal, Nadezhda Latysheva, Mikhail Sovershaev, and John-Bjarne Hansen. Differential expression of plasma miRNAs in patients with unprovoked venous thromboembolism and healthy control individuals. *Thrombosis Research*, 136(3):566–572, September 2015.

[13] Xiao Wang, Kristina Sundquist, Peter J. Svensson, Hamideh Rastkhani, Karolina Palmér, Ashfaque A. Memon, Jan Sundquist, and Bengt Zöller. Association of recurrent venous thromboembolism and circulating microRNAs. *Clinical Epigenetics*, 11(1):28, 2019.

[14] Tiphaine Oudot-Mellakh, William Cohen, Marine Germain, Noémie Saut, Choumous Kallel, Diana Zelenika, Mark Lathrop, David-Alexandre Trégouët, and Pierre-Emmanuel Morange. Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. *British Journal of Haematology*, 157(2):230–239, April 2012.

[15] Marine Germain, Noémie Saut, Nicolas Greliche, Christian Dina, Jean-Charles Lambert, Claire Perret, William Cohen, Tiphaine Oudot-Mellakh, Guillemette Antoni, Marie-Christine Alessi, Diana Zelenika, François Cambien, Laurence Tiret, Marion Bertrand, Anne-Marie Dupuy, Luc Letenneur, Mark Lathrop, Joseph Emmerich, Philippe Amouyel, David-Alexandre Trégouët, and Pierre-Emmanuel Morange. Genetics of venous thrombosis: insights from a new genome wide association study. *PloS One*, 6(9):e25581, 2011.

[16] Marine Germain, Daniel I. Chasman, Hugoline de Haan, Weihong Tang, Sara Lindström, Lu-Chen Weng, Mariza de Andrade, Marieke C. H. de Visser, Kerri L. Wiggins, Pierre Suchon, Noémie Saut, David M. Smadja, Grégoire Le Gal, Astrid van Hylckama Vlieg, Antonio Di Narzo, Ke Hao, Christopher P. Nelson, Ares Rocanin-Arjo, Lasse Folkersen, Ramin Monajemi, Lynda M. Rose, Jennifer A. Brody, Eline Slagboom, Dylan Aïssi, France Gagnon, Jean-Francois Deleuze, Panos Deloukas, Christophe Tzourio, Jean-Francois Dartigues, Claudine Berr, Kent D. Taylor, Mete Civelek, Per Eriksson, Cardiogenics Consortium, Bruce M. Psaty, Jeanine Houwing-Duitermaat, Alison H. Goodall, François Cambien, Peter Kraft, Philippe Amouyel, Nilesh J. Samani, Saonli Basu, Paul M. Ridker, Frits R. Rosendaal, Christopher Kabrhel, Aaron R. Folsom, John Heit, Pieter H. Reitsma, David-Alexandre Trégouët, Nicholas L. Smith, and Pierre-Emmanuel Morange. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44a2 as two susceptibility loci for venous thromboembolism. *American Journal of Human Genetics*, 96(4):532–542, April 2015.

[17] Florian Thibord, Claire Perret, Maguelonne Roux, Pierre Suchon, Marine Germain, Jean-François Deleuze, Pierre-Emmanuel Morange, David-Alexandre Trégouët, and GENMED Consortium. OPTIMIR, a novel algorithm for integrating available genome-wide genotype data into miRNA sequence alignment analysis. *RNA (New York, N.Y.)*, 25(6):657–668, 2019.

[18] Ana Kozomara and Sam Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1):D68–D73, January 2014.

[19] Stefan L. Ameres and Phillip D. Zamore. Diversifying microRNA sequence and function. *Nature Reviews. Molecular Cell Biology*, 14(8):475–488, August 2013.

[20] Michaela B. Kirschner, J. James B. Edelman, Steven Chuan-Hao Kao, Michael P. Vallely, Nico Van Zandwijk, and Glen

Reid. The Impact of Hemolysis on Cell-Free microRNA Biomarkers. *Frontiers in Genetics*, 4, 2013.

[21] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[22] Parameswaran Ramachandran, Daniel Sánchez-Taltavull, and Theodore J. Perkins. Uncovering robust patterns of microRNA co-expression across cancers using Bayesian Relevance Networks. *PloS One*, 12(8):e0183103, 2017.

[23] Katrin Töpner, Guilherme J. M. Rosa, Daniel Gianola, and Chris-Carolin Schön. Bayesian Networks Illustrate Genomic and Residual Trait Connections in Maize (Zea mays L.). *G3: Genes|Genomes|Genetics*, 7(8):2779–2789, June 2017.

[24] Marco Scutari. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(1):1–22, July 2010.

[25] Barend W. Florijn, Roel Bijkerk, Eric P. van der Veer, and Anton Jan van Zonneveld. Gender and cardiovascular disease: are sex-biased microRNA networks a driving force behind heart failure with preserved ejection fraction in women? *Cardiovascular Research*, 114(2):210–225, 2018.

[26] Tianxiao Huan, George Chen, Chunyu Liu, Anindya Bhattacharya, Jian Rong, Brian H. Chen, Sudha Seshadri, Kahraman Tanriverdi, Jane E. Freedman, Martin G. Larson, Joanne M. Murabito, and Daniel Levy. Age-associated microRNA expression in human peripheral blood is associated with all-cause mortality and age-related traits. *Aging Cell*, 17(1), February 2018.

[27] Xi Chen, Hongwei Liang, Danping Guan, Cheng Wang, Xiaoyun Hu, Lin Cui, Sidi Chen, Chunni Zhang, Junfeng Zhang, Ke Zen, and Chen-Yu Zhang. A combination of Let-7d, Let-7g and Let-7i serves as a stable reference for normalization of serum microRNAs. *PloS One*, 8(11):e79652, 2013.

[28] Wei-Yann Tsai, Nicholas P. Jewell, and Mei-Cheng Wang. A Note on the Product-Limit Estimator Under Right Censoring and Left Truncation. *Biometrika*, 74(4):883–886, 1987.

[29] J. Li and L. Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221–227, September 2005.

[30] Yun Li, Cristen J. Willer, Jun Ding, Paul Scheet, and Gonçalo R. Abecasis. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834, December 2010.

[31] Majid Nikpay, Kaitlyn Beehler, Armand Valsesia, Jorg Hager, Mary-Ellen Harper, Robert Dent, and Ruth McPherson. Genome-wide identification of circulating-miRNA expression quantitative trait loci reveals the role of several miRNAs in the regulation of Cardiometabolic phenotypes. *Cardiovascular Research*, January 2019.

[32] Reedik Mägi and Andrew P. Morris. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics*, 11(1):288, May 2010.

[33] Mercedes Rubio, Mariona Bustamante, Carles Hernandez-Ferrer, Dietmar Fernandez-Orth, Lorena Pantano, Yaris Sarria, Maria Piqué-Borras, Kilian Vellve, Silvia Agramunt, Ramon Carreras, Xavier Estivill, Juan R. Gonzalez, and Alfredo Mayor. Circulating miRNAs, isomiRs and small RNA clusters in human plasma and breast milk. *PloS One*, 13(3):e0193527, 2018.

[34] Sigrid K. Braekkan, Ellisiv B. Mathiesen, Inger Njølstad, Tom Wilsgaard, and John-Bjarne Hansen. Hematocrit and risk of venous thromboembolism in a general population. The Tromso study. *Haematologica*, 95(2):270–275, February 2010.

[35] Suely Meireles Rezende, Willem M. Lijfering, Frits R. Rosendaal, and Suzanne C. Cannegieter. Hematologic variables and venous thrombosis: red cell distribution width and blood monocyte count are associated with an increased risk. *Haematologica*, 99(1):194–200, January 2014.

[36] Tianxiao Huan, Jian Rong, Chunyu Liu, Xiaoling Zhang, Kahraman Tanriverdi, Roby Joehanes, Brian H. Chen, Joanne M. Murabito, Chen Yao, Paul Courchesne, Peter J. Munson, Christopher J. O'Donnell, Nancy Cox, Andrew D. Johnson, Martin G. Larson, Daniel Levy, and Jane E. Freedman. Genome-wide identification of microRNA expression quantitative trait loci. *Nature Communications*, 6:6601, March 2015.

[37] Mete Civelek, Raffi Hagopian, Calvin Pan, Nam Che, Wenpin Yang, Paul S. Kayne, Niyas K. Saleem, Henna Cederberg, Johanna Kuusisto, Peter S. Gargalovic, Todd G. Kirchgessner, Markku Laakso, and Aldons J. Lusis. Genetic regulation of human adipose microRNA expression and its consequences for metabolic traits. *Human Molecular Genetics*, 22(15):3023–3037, August 2013.

[38] Benjamin B. Sun, Joseph C. Maranville, James E. Peters, David Stacey, James R. Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, Clare Oliver-Williams, Mihir A. Kamat, Bram P. Prins, Sheri K. Wilcox, Erik S. Zimmerman, An Chi, Narinder Bansal, Sarah L. Spain, Angela M. Wood, Nicholas W. Morrell, John R. Bradley, Nebojsa Janjic, David J. Roberts, Willem H. Ouwehand, John A. Todd, Nicole Soranzo, Karsten Suhre, Dirk S. Paul, Caroline S. Fox, Robert M. Plenge, John Danesh, Heiko Runz, and Adam S. Butterworth. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, 2018.

[39] Barbara E. Stranger, Alexandra C. Nica, Matthew S. Forrest, Antigone Dimas, Christine P. Bird, Claude Beazley, Catherine E. Ingle, Mark Dunning, Paul Flicek, Daphne Koller, Stephen Montgomery, Simon Tavaré, Panos Deloukas, and Emmanouil T. Dermitzakis. Population genomics of human gene expression. *Nature Genetics*, 39(10):1217–1224, October 2007.

[40] Yu-Ping Wang and Kuo-Bin Li. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. *BMC genomics*, 10:218, May 2009.

[41] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, June 2013.

[42] William J. Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L. Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A. Kostadima, John J. Lambourne, Suthesh Sivapalaratnam, Kate Downes, Kousik Kundu, Lorenzo Bomba, Kim Berentsen, John R. Bradley, Louise C. Daugherty, Olivier Delaneau, Kathleen Freson, Stephen F. Garner, Luigi Grassi, Jose Guerrero, Matthias Haimel, Eva M. Janssen-Megens, Anita Kaan, Mihir Kamat, Bowon Kim, Amit Mandoli, Jonathan Marchini, Joost H. A. Martens, Stuart Meacham, Karyn Megy, Jared O'Connell, Romina Petersen, Nilofar Sharifi, Simon M. Sheard, James R. Staley, Salih Tuna, Martijn van der Ent, Klaudia Walter, Shuang-Yin Wang, Eleanor Wheeler, Steven P. Wilder, Valentina

Iotchkova, Carmel Moore, Jennifer Sambrook, Hendrik G. Stunnenberg, Emanuele Di Angelantonio, Stephen Kaptoge, Taco W. Kuijpers, Enrique Carrillo-de Santa-Pau, David Juan, Daniel Rico, Alfonso Valencia, Lu Chen, Bing Ge, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yang, Roderic Guigo, Stephan Beck, Dirk S. Paul, Tomi Pastinen, David Bujold, Guillaume Bourque, Mattia Frontini, John Danesh, David J. Roberts, Willem H. Ouwehand, Adam S. Butterworth, and Nicole Soranzo. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, 167(5):1415–1429.e19, 2016.

[43] Patrick R. Lawler and Jack Lawler. Molecular basis for the regulation of angiogenesis by thrombospondin-1 and -2. *Cold Spring Harbor Perspectives in Medicine*, 2(5):a006627, May 2012.

[44] C. Trumel, M. Plantavid, S. Lévy-Tolédano, A. Ragab, J. P. Caen, E. Aguado, B. Malissen, and B. Payrastre. Platelet aggregation induced by the C-terminal peptide of thrombospondin-1 requires the docking protein LAT but is largely independent of alphaIIb/beta3. *Journal of thrombosis and haemostasis: JTH*, 1(2):320–329, February 2003.

[45] Dimitrios Iliopoulos, Konstantinos Drosatos, Yaeko Hiyama, Ira J. Goldberg, and Vassilis I. Zannis. MicroRNA-370 controls the expression of microRNA-122 and Cpt1alpha and affects lipid metabolism. *Journal of Lipid Research*, 51(6):1513–1523, June 2010.

[46] Wei Gao, Hui-Wei He, Ze-Mu Wang, Huan Zhao, Xiao-Qing Lian, Yong-Sheng Wang, Jun Zhu, Jian-Jun Yan, Ding-Guo Zhang, Zhi-Jian Yang, and Lian-Sheng Wang. Plasma levels of lipometabolism-related miR-122 and miR-370 are increased in patients with hyperlipidemia and associated with coronary artery disease. *Lipids in Health and Disease*, 11:55, May 2012.

[47] R. O. Benatti, A. M. Melo, F. O. Borges, L. M. Ignacio-Souza, L. a. P. Simino, M. Milanski, L. A. Velloso, M. A. Torsoni, and A. S. Torsoni. Maternal high-fat diet consumption modulates hepatic lipid metabolism and microRNA-122 (miR-122) and microRNA-370 (miR-370) expression in offspring. *The British Journal of Nutrition*, 111(12):2112–2122, June 2014.

[48] Dan Tian, Yin Sha, Jing-Min Lu, and Xian-Jin Du. MiR-370 inhibits vascular inflammation and oxidative stress triggered by oxidized low-density lipoprotein through targeting TLR4. *Journal of Cellular Biochemistry*, 119(7):6231–6237, 2018.

[49] Chih-Hung Chou, Sirjana Shrestha, Chi-Dung Yang, Nai-Wen Chang, Yu-Ling Lin, Kuang-Wen Liao, Wei-Chi Huang, Ting-Hsuan Sun, Siang-Jyun Tu, Wei-Hsiang Lee, Men-Yee Chiew, Chun-San Tai, Ting-Yen Wei, Tzi-Ren Tsai, Hsin-Tzu Huang, Chung-Yu Wang, Hsin-Yi Wu, Shu-Yi Ho, Pin-Rong Chen, Cheng-Hsun Chuang, Pei-Jung Hsieh, Yi-Shin Wu, Wen-Liang Chen, Meng-Ju Li, Yu-Chun Wu, Xin-Yi Huang, Fung Ling Ng, Waradee Buddhakosai, Pei-Chun Huang, Kuan-Chun Lan, Chia-Yen Huang, Shun-Long Weng, Yeong-Nan Cheng, Chao Liang, Wen-Lian Hsu, and Hsien-Da Huang. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research*, 46(D1):D296–D302, January 2018.

[50] France Gagnon, Dylan Aïssi, Alain Carrié, Pierre-Emmanuel Morange, and David-Alexandre Trégouët. Robust validation of methylation levels association at CPT1a locus with lipid plasma levels1. *Journal of Lipid Research*, 55(7):1189–1191, July 2014.

[51] Alexis C. Frazier-Wood, Stella Aslibekyan, Devin M. Absher, Paul N. Hopkins, Jin Sha, Michael Y. Tsai, Hemant K. Tiwari, Lindsay L. Waite, Degui Zhi, and Donna K. Arnett. Methylation at CPT1a locus is associated with lipoprotein subfraction profiles. *Journal of Lipid Research*, 55(7):1324–1330, 2014.

[52] Marguerite R. Irvin, Degui Zhi, Roby Joehanes, Michael Mendelson, Stella Aslibekyan, Steven A. Claas, Krista S. Thibeault, Nikita Patel, Kenneth Day, Lindsay Waite Jones, Liming Liang, Brian H. Chen, Chen Yao, Hemant K. Tiwari, Jose M. Ordovas, Daniel Levy, Devin Absher, and Donna K. Arnett. Epigenome-wide association study of fasting blood lipids in the Genetics of Lipid-lowering Drugs and Diet Network study. *Circulation*, 130(7):565–572, August 2014.

[53] Derek Klarin, Connor A. Emdin, Pradeep Natarajan, Mark F. Conrad, and Sekar Kathiresan. Genetic Analysis of Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates Obesity as a Causal Risk Factor. *Circulation. Cardiovascular genetics*, 10(2), April 2017.

[54] Takumi Sonoda, Juntaro Matsuzaki, Yusuke Yamamoto, Takashi Sakurai, Yoshiaki Aoki, Satoko Takizawa, Shumpei Niida, and Takahiro Ochiya. Serum MicroRNA-Based Risk Prediction for Stroke. *Stroke*, 50(6):1510–1518, June 2019.

Bayesian Network Analysis of plasma microRNA sequencing data in patients with venous thrombosis: Supplementary Materials

Supplementary tables descriptions

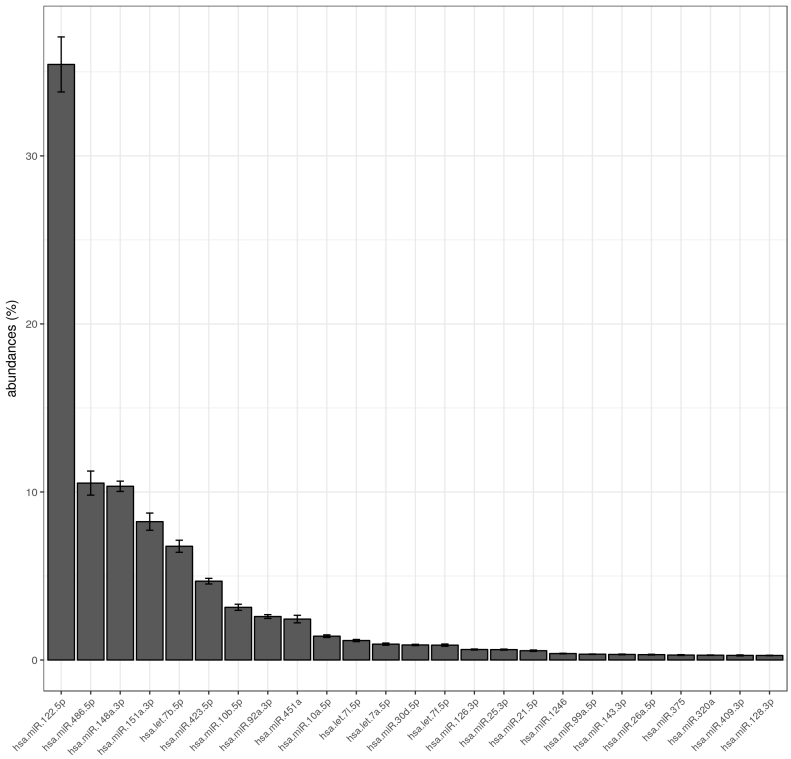
Supplementary table 1: Raw abundances of plasma microRNA sequencing data

Supplementary Table 2: Correlation between plasma miRNA levels and haematological traits

Supplementary Table 3: Association of plasma miRNA levels with the risk of VT recurrence in the MARTHA miRNA study

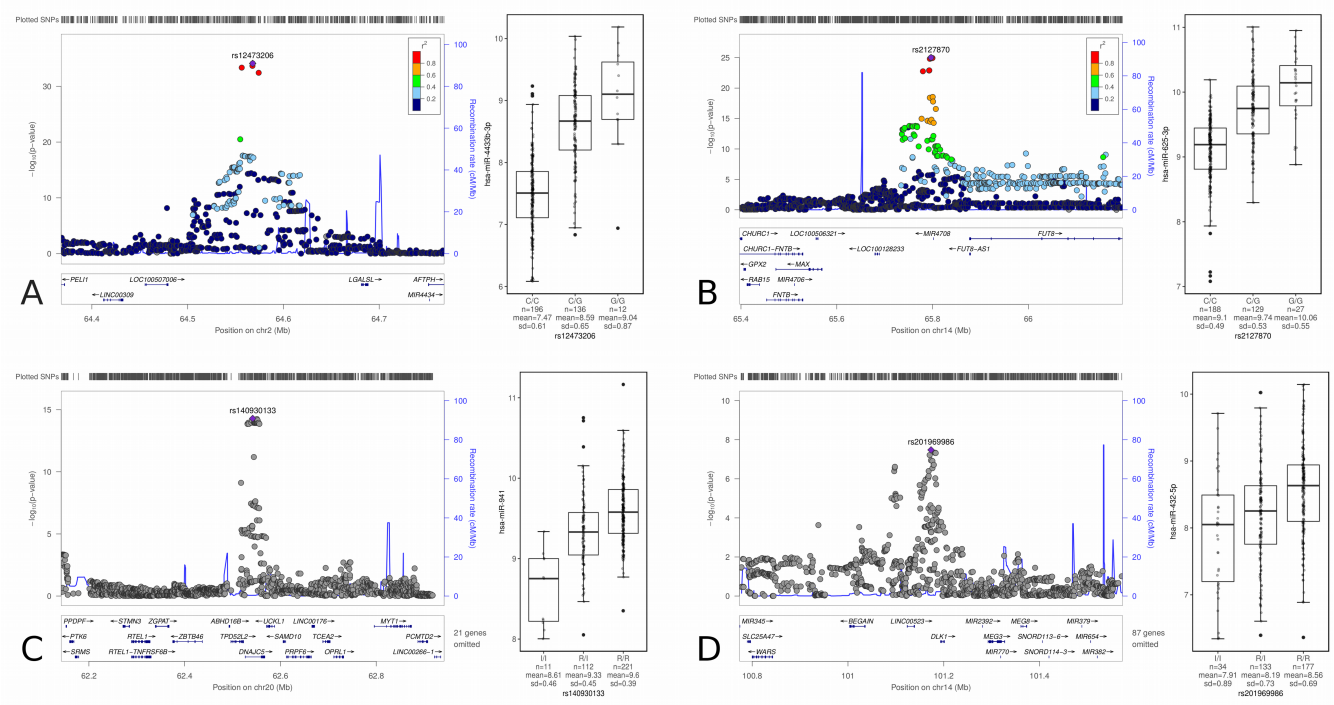
Supplementary Figures

Supplementary Figure 1 : Raw abundances of the 25 most expressed microRNAs



Abundances are shown for each microRNA as a percentage of total microRNAs quantified in the 344 MARTHA patients

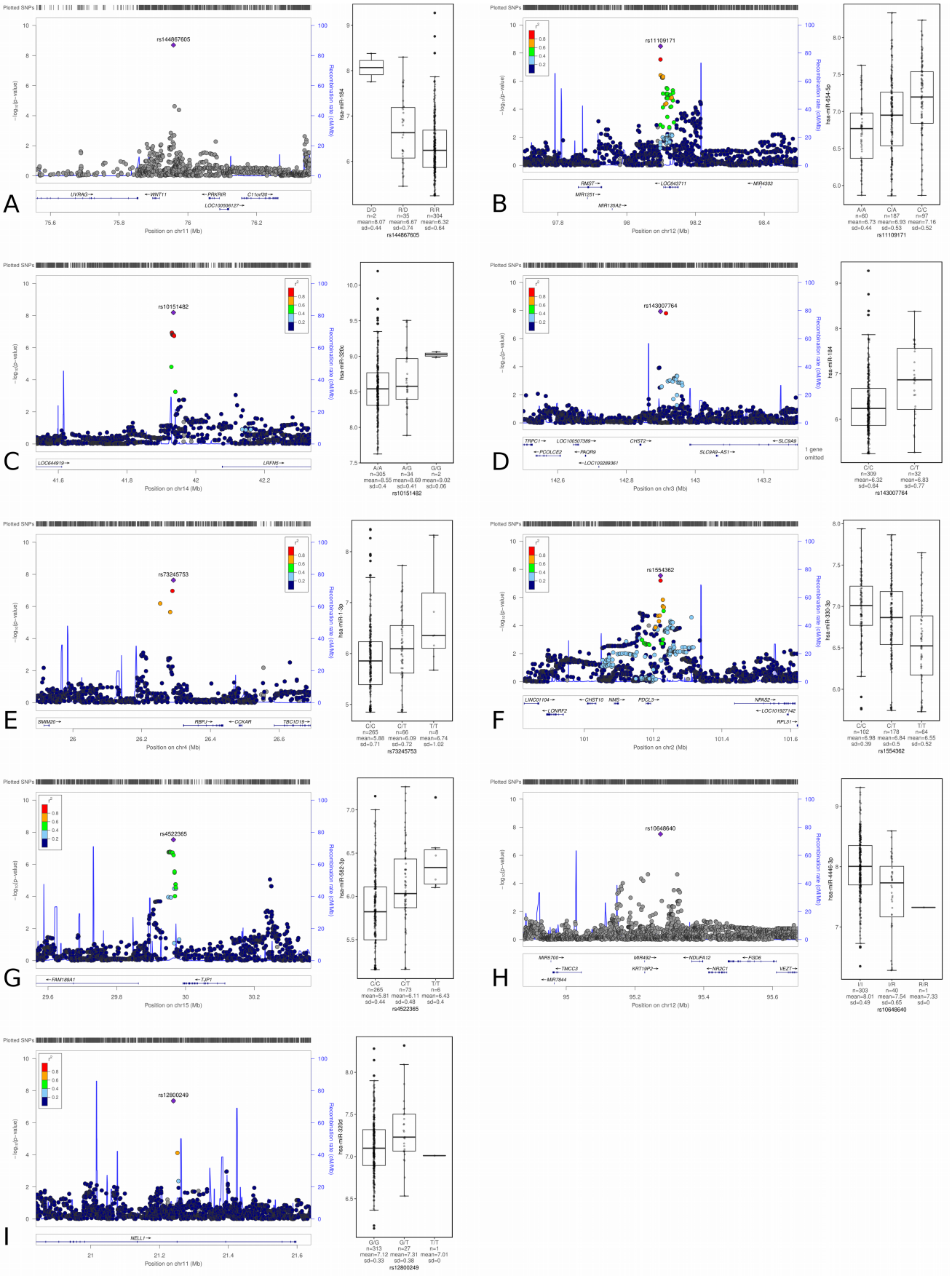
Supplementary Figure 2 : Regional association plots and boxplots of significant cis associations



For each figure, a regional association plot (on the left) shows the associations of SNPs in the region centered on the peak SNP, and a boxplot (on the right) shows the distribution of miRNA expression for each genotype of the peak SNP. (A) hsa-miR-4433b-3p and rs12473206. (B) hsa-miR-625-3p and rs2127870. (C) hsa-miR-941 and rs1409330133. (D) hsa-miR-432-5p and rs201969986.

Supplementary Figure 3 (on the next page): Regional association plots and boxplots of significant trans associations

(A) hsa-miR-184 and rs144867605. (B) hsa-miR-654-3p and rs11109171. (C) hsa-miR-320c and rs10151482. (D) hsa-miR-184 and rs143007764. (E) hsa-miR-1-3p and rs73245753. (F) hsa-miR-330-3p and rs1554362. (G) hsa-miR-582-3p and rs4522365. (H) hsa-miR-4446-3p and rs10648640. (I) hsa-miR-320d and rs12800249.


A
rs144867605
Position on chr11 (Mb)
rs144867605
B
rs11109171
Position on chr12 (Mb)
rs11109171
C
rs10151482
Position on chr14 (Mb)
rs10151482
D
rs143007764
Position on chr3 (Mb)
rs143007764
E
rs73245753
Position on chr4 (Mb)
rs73245753
F
rs1554362
Position on chr2 (Mb)
rs1554362
G
rs4522365
Position on chr15 (Mb)
rs4522365
H
rs10648640
Position on chr12 (Mb)
rs10648640
I
rs12800249
Position on chr11 (Mb)
rs12800249

Supplementary Table 3: Association of plasma miRNA levels with the risk of VT recurrence in the MARTHA miRNA study
Extrait de la table: la table complète et le reste des supplementary materials
peuvent être fournis par le corresponding author.

miRNA	HR [95%CI]	p
hsa-miR-27a-3p	0.19 [0.07 - 0.52]	0.001
hsa-miR-941	3.34 [1.53 - 7.33]	0.0027
hsa-miR-330-3p	2.99 [1.36 - 6.55]	0.0054
hsa-miR-191-5p	3.01 [1.29 - 7.07]	0.0101
hsa-miR-451a	0.57 [0.36 - 0.89]	0.0105
hsa-miR-30e-3p	3.79 [1.34 - 10.77]	0.0107
hsa-miR-320a	0.21 [0.06 - 0.73]	0.0108
hsa-miR-744-5p	2.43 [1.22 - 4.83]	0.011
hsa-miR-320d	0.26 [0.09 - 0.78]	0.013
hsa-miR-4446-3p	2.47 [1.2 - 5.09]	0.0137
hsa-miR-139-3p	3.07 [1.25 - 7.54]	0.0149
hsa-miR-27b-3p	0.4 [0.2 - 0.79]	0.0161
hsa-miR-370-3p	1.77 [1.09 - 2.88]	0.0194
hsa-miR-184	0.53 [0.3 - 0.95]	0.0235
hsa-miR-409-3p	1.68 [1.06 - 2.66]	0.0244
hsa-miR-4732-5p	0.41 [0.18 - 0.92]	0.0244
hsa-miR-320b	0.23 [0.06 - 0.91]	0.0262
hsa-miR-6842-3p	2.08 [1.06 - 4.08]	0.0314
hsa-miR-4433b-3p	1.54 [1.04 - 2.29]	0.0332
hsa-miR-99b-3p	2.57 [1.08 - 6.08]	0.0343
hsa-miR-1180-3p	0.45 [0.21 - 0.97]	0.0357
hsa-miR-16-2-3p	0.52 [0.28 - 0.96]	0.0382
hsa-miR-485-5p	1.89 [1.02 - 3.48]	0.0416
hsa-miR-654-5p	1.91 [1.01 - 3.6]	0.0448
hsa-miR-222-3p	1.76 [1.01 - 3.08]	0.0495
...		

Association were tested using a Cox model adjusted for age, sex , smoking and BMI
Reported p-values were obtained from the Likelihood Ratio Test (LRT) statistic

4 Discussions et conclusions

4.1 Associations génétiques

Les GWAS réalisées à partir des 162 miARNs ont permis d'identifier seulement 3 cis-eQTLs significatifs en utilisant la correction de Bonferroni pour les tests multiples, dont deux ont déjà été identifiés dans des études similaires [115, 53]. L'association du miR-4433b-3p avec le variant rs12473206, localisé dans la séquence miR-4433b-3p, est la seule qui n'a pas été découverte précédemment. Cette association n'est donc détéctable que si l'étape de quantification des miARNs prend en compte la détection de polymiRs, comme le permet optimiR.

En utilisant le seuil "classique" ($p < 5 \times 10^{-8}$) de significativité pour les GWAS, on trouve 10 associations supplémentaires inédites, dont 1 cis-eQTL, et 9 trans-eQTLs. En comparaison, l'étude menée par Nikpay avec 710 échantillons a détécté 223 eQTLs au seuil classique de significativité. Le manque d'association significatives dans MARTHA est donc probablement lié à un manque de puissance, lié à la taille modeste de la cohorte. Malgré ce manque de puissance, notre étude a permis de répliquer 37 associations identifiées par Nikpay, permettant ainsi de confirmer ces associations. De plus, la combinaison des résultats obtenus par les deux études dans une méta-analyse a permis d'identifier 12 associations supplémentaires, dont 10 qui n'ont pas été découvertes par de précédentes études. Au total cette étude a donc permis de répliquer une quarantaine d'associations précédemment identifiées, et de découvrir 21 nouvelles associations. Ces résultats démontrent qu'il est possible d'identifier des eQTLs pour des miARNs circulants.

Associations d'intérêt pour l'étude des troubles de l'hémostase Parmi les trans-eQTLs répliqués à partir des données de Nikpay, l'association entre le miR-222-3p et les variants rs11070216 et rs970280 sont les plus significatives ($p = 5.06 \times 10^{-5}$ et $p = 3.57 \times 10^{-3}$, respectivement). Ces deux variants sont localisés en amont du gène *THBS1*, et possèdent des variants en LD précédemment rapportés comme influançant les niveaux d'expression de *THBS1* [286]. De plus, certains de ces variants en LD sont localisés dans les introns de *THBS1*. Cela suggère qu'une modulation de l'expression de *THBS1* par ces variants pourrait influencer l'expression du miR-222-3p. Une interaction directe a d'ailleurs récemment été identifiée entre ce miARN et ce gène dans un modèle porcin [305]. Le gène *THBS1* correspond à la protéine d'adhésion cellulaire thrombospondine-1 pouvant former des complexes avec plusieurs facteurs impliqués dans l'hémostase, tels que le fibrinogène ou le plasminogène, et est exprimé notamment par les plaquettes lors de leur activation. On peut également mentionner deux autres associations avec un effet trans, également découvertes par Nikpay et répliquées dans MARTHA: la première concerne le miR-143-3p et le variant rs4734879 intronique du gène *ZFPM2*, et la seconde concerne le miR-126-3p et le variant rs600038 localisé dans le promoteur d'*ABO*. Ces deux associations peuvent avoir un intérêt pour l'étude des troubles de l'hémostase, car le variant rs4734879 du gène *ZFPM2* est en LD ($r^2 = 0.70$) avec le rs4602861 précédemment rapporté comme associé au risque de VTE [147], et ce variant est également en LD avec le rs6993770 ($r^2 = 0.97$) associé à l'abondance de plaquettes [9]. Le variant rs600038 du locus *ABO* est quand à lui en LD ($r^2 = 0.99$) avec le rs495828, lui aussi précédemment rapporté comme associé avec le risque de VTE [106].

Mécanismes de régulation de l'expression des miARNs On peut suggérer deux mécanismes par lesquels les cis-eQTLs peuvent réguler les niveaux d'expression des miARNs.

Le premier est celui d'un variant localisé dans la séquence du pri- ou du pre-miARN, comme c'est le cas pour le miR-4433b-3p et du rs12473206, qui peut interférer avec le processus de maturation du miARN, et donc influencer les niveaux de production du miARN. Le second correspond à des variants localisés au niveau de régions régulatrices de l'expression, tels que des enhancer, promoteurs, ou ilôts CpG. Par exemple, un cluster de miARNs localisés sur le locus génomique 14q32, auquel appartient le miR-432-5p qui est régulé par le cis-eQTL rs201969986 dans MARTHA, est connu comme étant régulé épigénétiquement. Dans notre étude sur les données de MARTHA, plusieurs miARNs appartenant à ce cluster sont également régulés suggestivement par ce variant rs201969986: le miR-409-3p ($p = 2.95 \times 10^{-7}$), le miR-381-3p ($p = 2.4 \times 10^{-6}$), le miR-654-5p ($p = 9.6 \times 10^{-6}$), le miR-127-3p ($p = 2.3 \times 10^{-5}$), et le miR-379-5p ($p = 2.9 \times 10^{-5}$). Grâce à une précédente étude sur l'associations de variants génétiques avec les niveaux de méthylation de l'ADN dans MARTHA [12], nous avons constaté que le variant rs201969986 est également fortement associé aux niveaux de méthylation d'un site CpG à moins de 3000 nucléotides du variant (le cg18089426, $p = 4 \times 10^{-44}$). Cela suggère que ce variant génétique, ou un variant en LD avec ce dernier, pourrait influencer localement les niveaux de méthylation, et ainsi influencer les niveaux de transcription des gènes et du cluster de miARNs présents sur ce locus. D'après le catalogue GWAS [12] [37], ce locus est associé à plusieurs maladies complexes telles que la sclérose latérale amyotrophique, le diabète de type 1 et 2, ou la scoliose, mais aussi avec l'abondance de plaquettes [9], ce qui en fait un locus intéressant pour la recherche sur les troubles de l'hémostase. Si les miARNs présents dans cette région ne sont pas directement impliqués dans ces pathologies, ils pourraient néammoins servir de biomarqueurs reflètant les niveaux de régulation de cette région.

En revanche, il est difficile d'identifier le mécanisme pouvant mener à une régulation en trans d'un miARN. Certaines associations en trans semblent impliquer un miARN et un ARNm ciblé dans le cytoplasme, comme le miR-222-3p et le gène *THBS1*. Cependant, comme mentionné dans la section I.4.3, à cause de l'abondance du nombre de cibles potentielles pour un miARN, il est peu probable que la modification de l'expression d'un ARNm ciblé puisse influencer l'abondance d'un miARN. Toutefois, si le gène est transcrit en un ARN connu comme régulant l'expression du miARN par un mécanisme tel que le TDMD, alors une modification de l'expression de ce gène pourrait se répercuter par une modification de l'abondance du miARN. Par exemple, une association trans entre le miR-7-5p et un variant influençant les niveaux d'expression du lncARN *Cyrano*, un ARN connu comme régulant les niveaux du miR-7-5p par TDMD, a été identifié par la méta-analyse.

4.2 Associations avec le risque de récidive de la VTE et avec les variables biologiques de l'hémostase

Malgré l'utilisation de la technique de réduction par réseaux bayésiens, cette étude n'a pas permis de découvrir d'association significative entre les 15 miARNs terminaux et la récidive ou les variables de l'hémostase. Cela suggère que l'hypothèse utilisée pour cette réduction, qui stipule que les miARNs terminaux sont les plus à même de faire office de biomarqueurs, en tant qu'intégrateurs de l'effet cumulé de miARNs parents dans le même sous réseau, n'est pas valide pour les traits analysés. Toutefois, ce manque de résultats significatifs peut aussi s'expliquer par la taille modeste de la cohorte, qui ne permet pas d'obtenir une puissance statistique nécessaire pour observer des associations significatives. Cependant, plusieurs associations suggestives ($p < 0.05$) sont intéressantes parmi les 162 miARNs testés.

12. Ressource bioinformatique qui recense les résultats de GWAS dont les résultats ont été publiés.

Par exemple, 25 miARNs possèdent une association suggestive avec le risque de récidive, dont 15 sont sur-exprimés chez les patients ayant subit une récidive, et 10 sous-exprimés. Parmi ces associations suggestives, les miARNs miR-27-3p (sous-exprimé) et miR-222-3p (sur-exprimé) sont particulièrement intéressants car ils ont été précédemment associés au risque de récidive dans l'étude du groupe de Zöller, avec la même tendance d'expression. De plus le miARN le plus significativement associé au risque de récidive est le miR-27a-3p ($p = 0.001$, sous-exprimé), qui appartient à la même famille que le miR-27b-3p, et qui partage donc potentiellement les mêmes cibles. Ces deux miARNs ont précédemment été identifiés comme régulant l'expression du TFPI [2, 13], mais il n'y a aucune corrélation entre ces deux miARNs et les niveaux plasmatiques du TFPI dans nos données (miR-27a-3p: $\rho = -0.02, p = 0.5$ et miR-27b-3p: $\rho = 0.09, p = 0.2$). Le miR-27b-3p a également été identifié comme régulant directement l'expression de THBS1 [277], ce qui est un point commun avec le miR-222-3p. Parmi les associations suggestives avec les variables de l'hémostase dans MARTHA, le miR-222-3p est positivement corrélé aux taux d'antithrombine ($\rho = 0.21, p = 8.8 \times 10^{-4}$), et le miR-27b-3p est légèrement corrélé avec les taux de PAI-1 ($\rho = 0.14, p = 3.8 \times 10^{-2}$).

La corrélation la plus significative observée dans MARTHA concerne le miR-193b-5p et les niveaux plasmatiques de PAI-1 ($\rho = 0.33, p = 2.6 \times 10^{-7}$), un inhibiteur de la fibrinolyse régulant les activateurs du plasminogène. De plus, son paralogue miR-193a-5p est également associé à PAI-1 ($\rho = 0.3, p = 2.9 \times 10^{-5}$). Comme la corrélation entre ces deux miARNs et PAI-1 est positive, il n'y a probablement pas de régulation directe entre ces deux miARNs et le PAI-1, mais il est possible que ces miARNs régulent un gène interagissant avec PAI-1, pouvant influencer ses niveaux plasmatiques. Par exemple, ces deux miARNs on été prédemment identifiés comme régulant directement l'expression de PLAU, un activateur du plasminogène dont l'action est inhibée par par PAI-1.

Globalement, les corrélations observées entre les niveaux plasmatiques des 162 miARNs et les traits de l'hémostase sont modestes, avec des valeurs comprises entre -0.25 et 0.33. Ces corrélations ne permettent pas d'affirmer l'implication directe d'un miARN dans la régulation d'un trait de l'hémostase, mais elles peuvent néammoins servir pour supporter de précédentes ou futures observations. Par exemple le miR-10a-5p précédemment rapporté comme régulant directement la production de PAI-1 [173], possède une corrélation négative avec les taux plasmatiques de PAI-1 dans nos données ($\rho = -0.14, p = 2.1 \times 10^{-3}$), supportant ainsi le rôle du miR-10a-5p dans la régulation directe de la production de la protéine PAI-1.

4.3 Conclusions

Cette étude n'a pas permis d'établir définitivement l'implication de miARNs dans la régulation des paramètres de l'hémostase ou de découvrir de nouveaux biomarqueurs robustes pour la récidive de la VTE. Cependant, plusieurs associations suggestives intéressantes ont été découvertes, en particulier les associations des miR-222-3p et miR-27b-3p avec la récidive de la VTE, qui avaient précédemment été rapportées par le groupe de Zöller. Cette étude a également permis d'identifier de nouvelles associations génétiques avec les niveaux plasmatiques de plusieurs miARNs. Ces associations ont permis d'établir les déterminants génétiques responsables du contrôle de l'expression de plusieurs miARNs, et pourraient avoir un intérêt pour les travaux de recherche bio-médicale ciblant ces miARNs.

Toutes les données générées au cours de cette étude sont publiées en tant que matériel supplémentaire avec l'article publié dans *European Heart Journal Supplement*, et les résultats des associations génétiques sont disponibles sur la plateforme `zenodo.org`. Ces données

pourraient ainsi être une ressource majeure pour de futures études entre les miARNs et les paramètres de l'hémostase.

Perspectives Plusieurs pistes sont envisageables pour poursuivre cette étude. Premièrement, ont pourrait étendre cette analyse aux autres petits ARNs non codants ou fragments d'ARN séquencés en même temps que les miARNs, mentionnés dans la section I.1.3.f. En particuliers, les fragments d'ARN Y sont fréquemment observés dans le plasma, et ils ont récemment été associés avec l'activité plaquettaire [132]. Ainsi ces autres petits ARNs non codants pourraient servir de biomarqueurs reflètant l'activité plaquettaire ou une dérégulation de l'hémostase. Ensuite, on pourrait également effectuer une recherche de miARNs de novo dans nos échantillons, afin de potentiellement découvrir des miARNs non indexés dans la miRBase. Enfin, les données produites dans cette étude pourraient également être réutilisées dans le cadre d'une méta-analyse avec une étude similaire dans une nouvelle cohorte, afin d'augmenter la puissance statistique nécessaire à l'observation d'associations significatives.

Travaux Annexes

En parallèle de mon projet de thèse, j'ai eu l'opportunité d'être impliqué dans plusieurs projets de recherches, qui m'ont permis de développer des compétences supplémentaires en bioinformatique et biostatistique, et de partager mes connaissances sur les microARNs.

1 Associations génétiques avec les niveaux plasmatiques du Facteur V

Avant de mettre en application des études de type GWAS pour trouver les déterminants génétiques des niveaux d'expression plasmatiques des miARNs, je me suis formé à ce type d'analyse sur les niveaux plasmatiques du Facteur V. Ce trait a été mesuré chez des participants de l'étude MARTHA, et cette analyse est particulièrement intéressante car il n'y a que très peu de connaissances sur les variants génétiques pouvant influencer les niveaux plasmatiques du Facteur V.

Le FV, décrit brièvement dans la section IV.1.1, est un facteur de la cascade de la coagulation qui intervient au niveau de la voie commune en tant que co-facteur du FX, pour activer le FII en thrombine. J'ai ainsi effectué une GWAS sur les niveaux plasmatiques du FV, qui ont été mesurés chez 510 individus. Cette analyse a permis de mettre en évidence une association significative avec le variant rs6027, localisé dans un exon du gène *F5*, qui avait précédemment été identifié comme pouvant influencer de façon significative les niveaux de production du gène *F5*. Deux associations suggestives ont également été découvertes, la première avec le variant rs27218, intronique au gène *MAST4*, et la seconde avec le variant rs927826, intronique au gène *PLXDC2*. Ces deux associations ont été analysées dans la cohorte indépendante MARTHA12, constituée de 1156 individus, et seule l'association avec le variant de *PLXDC2* a été répliquée. Des travaux supplémentaires de biologie moléculaire ont par la suite permis de confirmer l'influence du gène *PLXDC2* sur l'expression du *F5*, ainsi que d'autres gènes de la cascade de la coagulation. Les résultats de ces travaux ont fait l'objet d'une publication dans la revue *Journal of Thrombosis and Haemostasis*, incluse dans la partie suivante.

Les résultats de certaines analyses que j'ai effectué sur le *F5* m'ont par la suite ammené à contribuer à une autre étude, visant à déterminer la contribution de deux variants exoniques du *F5*, le Leiden rs6025 et le rs4524, sur le risque de VTE. Cette étude a fait l'objet d'une publication dans la revue *Scientific Reports*, disponible en Appendice C.

2 Article: *A Genome Wide Association Study on plasma FV levels identified PLXDC2 as a new modifier of the coagulation process*

Received: 6 March 2019 | Accepted: 1 July 2019

DOI: 10.1111/jth.14562

ORIGINAL ARTICLE

A Genome Wide Association Study on plasma FV levels identified *PLXDC2* as a new modifier of the coagulation process

Florian Thibord[1,2,3] | Lise Hardy[3,4] | Manal Ibrahim-Kosta[5,6] | Noémie Saut[5,6] | Anne-Sophie Pulcrano-Nicolas[1,3,4] | Louisa Goumidi[6] | Mete Civelek[7] | Per Eriksson[8,9] | Jean-François Deleuze[10,11] | Wilfried Le Goff[3,4] | David-Alexandre Trégouët[2,3] | Pierre-Emmanuel Morange[5,6,12]

[1]Pierre Louis Doctoral School of Public Health, Sorbonne-Université, Paris, France

[2]Institut National pour la Santé et la Recherche Médicale (INSERM) Unité Mixte de Recherche en Santé (UMR_S) 1219, Bordeaux Population Health Research Center, University of Bordeaux, Bordeaux, France

[3]INSERM UMR_S 1166, Université Pierre et Marie Curie (UPMC Univ Paris 06), Sorbonne Université, Paris, France

[4]ICAN Institute of Cardiometabolism and Nutrition, Paris, France

[5]Laboratory of Haematology, La Timone Hospital, Marseille, France

[6]C2VN, Aix Marseille Univ, INSERM, INRA, Marseille, France

[7]Department of Biomedical Engineering, Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia

[8]Department of Medicine, Cardiovascular Medicine Unit, BioClinicum, Karolinska Institutet, Stockholm, Sweden

[9]Karolinska University Hospital, Solna, Sweden

[10]Centre National de Recherche en Génomique Humaine, Direction de la Recherche Fondamentale, CEA, Evry, France

[11]CEPH, Fondation Jean Dausset, Paris, France

[12]CRB Assistance Publique - Hôpitaux de Marseille, HemoVasc (CRB AP-HM HemoVasc), Marseille, France

Correspondence
Pierre-Emmanuel Morange, Department of Haematology, CHU Timone, 264 rue Saint-Pierre, 13385 Marseille Cedex 05, France.
Email: pierre.morange@ap-hm.fr

Funding information
EPIDEMIOM-VTE Senior Chair from the Initiative of Excellence of the University of Bordeaux; GENMED Laboratory of Excellence on Medical Genomics, Grant/Award Number: ANR-10-LABX-0013

Abstract

Background: Factor V (FV) is a circulating protein primarily synthesized in the liver, and mainly present in plasma. It is a major component of the coagulation process.

Objective: To detect novel genetic loci participating to the regulation of FV plasma levels.

Methods: We conducted the first Genome Wide Association Study on FV plasma levels in a sample of 510 individuals and replicated the main findings in an independent sample of 1156 individuals.

Results: In addition to genetic variations at the *F5* locus, we identified novel associations at the *PLXDC2* locus, with the lead *PLXDC2* rs927826 polymorphism explaining ~3.7% ($P = 7.5 \times 10^{-15}$ in the combined discovery and replication samples) of the variability of FV plasma levels. In silico transcriptomic analyses in various cell types confirmed that *PLXDC2* expression is positively correlated to *F5* expression. SiRNA experiments in human hepatocellular carcinoma cell line confirmed the role of

Manuscript handled by: John-Bjarne Hansen

Final decision: John-Bjarne Hansen, 1 July 2019

David-Alexandre Trégouët and Pierre-Emmanuel Morange contributed equally to this study.

PLXDC2 in modulating factor *F5* gene expression, and revealed further influences on *F2* and *F10* expressions.

Conclusion: Our study identified *PLXDC2* as a new molecular player of the coagulation process.

KEYWORDS
biomarkers, coagulation factor, computational biology, factor V, genetics

Essentials

- Little is known about the regulation of Factor V plasma levels.
- A Genome Wide Association Study (GWAS) was performed on Factor V plasma levels.
- Genetic variations at the PLXDC2 gene are associated with plasma levels of Factor V.
- This study identifies a novel player of the coagulation process.

1 | INTRODUCTION

Coagulation factor V (FV) plays an important and dual role in the regulation of blood coagulation by exhibiting both pro- and anticoagulant functions (reviewed in[1]). In plasma, single-chain FV expresses anticoagulant activity as a cofactor of both TFPI and activated protein C. In situations where the coagulation system is triggered, FV is converted to a highly effective procoagulant cofactor to activated factor X (FXa), which activates prothrombin to thrombin, which in turn catalyzes fibrin deposition and activates platelets. Although FV deficiency is known to associate with bleeding tendency, it has recently been proposed that FV plasma levels were associated with the risk of venous thrombosis (VT) in an ambiguous pattern. Indeed, in 2377 VT patients and 2943 controls of the MEGA study, individuals with either high (>1.22 IU/mL) or low (<0.57 IU/mL) FV levels were at higher risk of disease.[2]

Until now, only one genetic factor, the HR2 haplotype located in the *F5* gene, has been robustly found associated with FV plasma levels[3] even though it has been hypothesized that FV Leiden genotype could also modulate FV plasma levels.[2,4]

We here report the results of the first genome wide association study (GWAS) aiming to identify new genetic determinants of FV plasma levels.

2 | MATERIALS AND METHODS

2.1 | Study description

This work builds on two independent samples of unrelated VT patients of European ancestry recruited at the Thrombophilia center of La Timone Hospital (Marseille, France), the MARseille THrombosis Association (MARTHA) and MARTHA12 cohorts.[5] MARTHA patients were used for the discovery GWAS, whereas MARTHA12 individuals were considered for the replication step. All participants provided written informed consent, and the protocol was approved by the ethics committee of the participating institution.

The MARTHA project has already been extensively described.[6,7] It is composed of unrelated subjects of European origin, with the majority being of French ancestry, consecutively recruited at the Thrombophilia center of La Timone hospital (Marseille, France) between January 1994 and October 2012. All patients had a documented history of VT and are free of well-characterized genetic risk factors including AT (Antithrombin), PC (Protein C), or PS (Protein S) deficiency, homozygosity for FV Leiden or FII G20210A, and lupus anticoagulant. The MARTHA12 study is an independent sample of 1245 VT patients recruited between 2010 and 2012, according to the same criteria as the MARTHA patients.[5,8]

2.2 | Hemostatic traits measurements

FV activity plasma levels (FV:C) were measured using human FV deficient plasma on automated coagulometers (STA® analysers from Diagnostica Stago). Activated partial thromboplastin time (aPTT) and prothrombin time (PT) were measured in plasma with the use of automated coagulometers. Tests were conducted on the same day as blood collection or within a few weeks (after freezing). PT was recorded as a ratio of the patient's PT to the mean normal PT, and aPTT was recorded in seconds (s). FV:C plasma levels were available in 510 participants of the discovery phase and 1156 participants of the replication. FV antigen plasma levels (FV:Ag) were measured by ELISA on freezing samples using the ZYMUTEST Factor V kit (Hyphen Biomed) in a random sample of 60 patients from the MARTHA study (n = 20 per rs927826 genotypes). Factor II (FII) and Factor X (FX) activity levels in plasma were measured in MARTHA and MATHA12 cohorts by using human FII- or FX-deficient plasma on automated coagulometers in a total of 738 patients with no anti-vitamin K antagonists (VKA) at the time of blood sampling.

2.3 | Genotyping

DNA samples of the MARTHA participants were typed with high-density genotyping arrays and imputed for single nucleotide

TABLE 1 Description of the MARTHA and MARTHA12 cohorts

	Discovery MARTHA N = 510	Replication MARTHA12 N = 1156
Age	47.3 ± 16.0	50.4 ± 15.4
Gender (M/F)	176/334	527/629
Current smoker (%)	24.1	20.2
FV Leiden carriers (%)	15.5	11.0
BMI, kg/m^2	25.34 ± 4.96	26.26 ± 5.06
Vitamin K antagonist users (%)	3.7	28
FV:C plasma levels (IU/mL)	1.068 ± 0.235	1.089 ± 0.217
Activated partial thromboplastin time, s	31.68 ± 3.71	33.61 ± 6.95
Prothrombin time (%)	99.11 ± 14.72	83.64 ± 30.03

Note: Mean ± standard deviation.

polymorphisms (SNPs) available in the 1000G reference as previously described.[5] In MARTHA12, wet-lab genotyping was performed using Taqman assay. Genotyping was done using Taqman 5′ nuclease assay (ThermoFischer Scientific, ref. C___2386120_10 (rs27218, *MAST4*), C___8879235_10 (rs1409338, *PLXDC2*)), following the supplier instructions, on an LC480® Real time PCR Instrument (Roche Life sciences), using dedicated Dual Color Hydrolysis Probe program. Reactions were made in a 10 µL final volume in a 1 × final concentration of Genotyping Master Mix (ThermoFischer Scientific ref. 4371355), with 25 ng of DNA, and 0,1 µL of 20× Genotyping Assay containing probes, nucleotides, and DNA Taq Polymerase.

2.4 | Genetic association analyses

For the GWAS discovery phase, association analyses of imputed SNPs with plasma FV levels were performed using linear regression analysis adjusted for age, sex, and the four first principal components derived from the GWAS data as implemented in the Mach2QTL program.[9] Only SNPs with acceptable imputation quality (r^2 > .5) and minor allele frequency >0.05 were considered. At the replication stage, associations of SNPs with plasma FV levels were assessed using standard linear regression analysis adjusted for age and sex. Haplotype association analyses were conducted using THESIAS software.[10] Results from the discovery and replication studies were meta-analyzed via a fixed-effects model based on the inverse-variance weighting method. A replicated SNP was further tested for association with additional hemostatic traits including PT, aPTT, FII, and FX circulating levels. For these analyses, patients with VKA at the time of blood sampling were excluded.

2.5 | RNA interference mediated PLXDC2 silencing experiments

Small interfering RNA (siRNA) *PLXDC2* gene silencing was conducted in human hepatocytes, a key cell type for *F5* regulation, to validate in vitro the association between *PLXDC2* and *F5* gene expressions. Additional expression of coagulation factors *F2*, *F7*, and *F10* were also measured to assess the specificity of the *PLXDC2* association with *F5* expression.

The human hepatocellular carcinoma cell line Hep3B (American Type Culture Collection) was grown at 37°C in 5% CO2 in Dulbecco's Modified Eagle's Medium containing 10% fetal calf serum, 2 mmol/L glutamine and 100 U/mL penicillin/streptomycin. Cells were seeded on 12-well plates at 100 000 cells per well and were transfected 72 hours later with 50 nmol/L control siRNA or an siRNA targeting the *PLXDC2* gene, s39607 (sequence ctacagaagatgataccaa from Thermofisher) using lipofectamine RNAiMax (Life Technologies) according to the manufacturer's instructions. Twenty-four hours following transfection, total RNA was extracted (NucleoSpin RNA II kit; Macherey-Nagel) and reverse transcribed (high-capacity cDNA reverse transcription kit, Life Technologies) and gene expression was analyzed by real-time qPCR using a LightCycler LC480 (Roche). Primers used for quantification of *PLXDC2* and *F5*, *F2*, *F7*, and *F10* are provided as Supporting Information. Expression of mRNA levels was normalized to human non-POU domain containing, octamer-binding housekeeping gene (NONO), human alanin (ALA) and human heat shock protein 90 kDa alpha (cytosolic) class B member 1 (HSP90AB1). Data (Table S1) were expressed as a fold change in mRNA expression relative to control cells.

The impact of *PLXDC2* knock-down on gene expression was tested via two-way analysis of variance.

3 | RESULTS

FV activity plasma levels (FV:C) were measured using human FV deficient plasma on automated coagulometers. The discovery GWAS was composed of 510 patients with venous thromboembolism (VTE) assessed for 6 264 382 SNPs and main significant findings were tested for replication in an independent sample of 1156 patients (see Materials & Methods). Main clinical and biological characteristics of the studied samples are shown in Table 1.

The Manhattan and Quantile-Quantile plots summarizing the GWAS results are shown in Figures S1-S4. About 70 SNPs achieved genome-wide significance ($P < 5 \times 10^{-8}$), with all SNPs mapping to the *F5* locus on chromosome 1q24.2. The lead SNP, rs72708008, with association *P* value = 8.43×10^{-12}, was in high linkage disequilibrium (LD) (r^2 > .80) with several SNPs including the rs6027 (p.Asp2222Gly, $P = 1.27 \times 10^{-11}$) known to tag the *F5* HR2 haplotype whose association with FV:C plasma levels has already been established.[11] After conditioning on rs6027, no additional signal reached genome-wide significance (Figure S3 and Table S2), the smallest *P* value being now $P = 3.05 \times 10^{-7}$ and observed at the *PLXDC2* locus. Because two other

TABLE 2 Association with FV activity plasma levels of lead SNPs at two loci that reached suggestive evidence for association in the discovery GWAS

	MAST4 rs27218[a]				*PLXDC2* rs927826[a,b]			
	CC	CT	TT	MAF	TT	TG	GG	MAF
Discovery								
Mean FV plasma levels ± SD	1.03 ± 0.22 N = 271	1.09 ± 0.23 N = 199	1.23 ± 0.29 N = 40	0.274	1.11 ± 0.22 N = 249	1.04 ± 0.24 N = 224	0.95 ± 0.19 N = 37	0.293
Additive allele effect[c]		+0.075 ± 0.015 $P = 9.02 \times 10^{-7}$				−0.076 ± 0.016 $P = 1.10 \times 10^{-6}$		
Replication								
Mean FV plasma levels ± SD	1.102 ± 0.21 N = 605	1.07 ± 0.22 N = 434	1.101 ± 0.21 N = 89	0.271	1.13 ± 0.22 N = 537	1.05 ± 0.21 N = 461	1.02 ± 0.20 N = 87	0.293
Additive allele effect		−0.019 ± 0.010 $P = .054$				−0.062 ± 0.010 $P = 2.89 \times 10^{-10}$		

Abbreviations: MAF, minor allele frequency; SD, standard deviation.
[a]Imputation quality criterion was .984 and .996 for *MAST4* rs27218 and *PLXDC2* rs927826, respectively, in the discovery GWAS.
[b]In the replication study, rs927826 was substituted by rs1409338 that served as a proxy ($r^2 = .97$) because of some technical issues in wet-lab genotyping the rs927826.
[c]Association testing was adjusted for age, sex, and the four main genetics components in the discovery cohort and on age and sex in the replication study.

SNPs at the *F5* locus, rs6025 (p.Arg534Gln) and rs4524 (p.Lys858Arg), have been extensively studied in relation to VT risk,[5] we further analyzed their joint association with rs6027 with respect to FV:C plasma levels in both the discovery and replication cohorts using haplotype analysis. This analysis revealed that in addition to the strong decreasing effect of the rs6027-C allele ($\beta = -.151 \pm .02$, $P = 1.93 \times 10^{-12}$), the rs4524-C allele was associated with a slight decrease ($\beta = -.033 \pm .01$, $P = .003$) in FV:C plasma levels (Table S3). Altogether, these two *F5* SNPs explained ~7% of the variability of FV:C plasma levels.

Aside the aforementioned *F5* association signal, two other loci with several SNPs in LD exhibited suggestive statistical evidence for association with FV:C plasma levels in the discovery cohort, with *P*-values ranging between ~10^{-6} and 10^{-5} (Figure S1, Table S2). These loci were *MAST4* and *PLXDC2* where lead SNPs, rs27218 (intronic; $P = 9.02 \times 10^{-7}$) and rs927826 (intronic; $P = 1.10 \times 10^{-6}$), respectively, were looked for replication. Although the association of *MAST4* rs27218 did not replicate (Table 2), the association of *PLXDC2* rs927826 with plasma FV:C levels was confirmed ($P = 2.89 \times 10^{-10}$). In both the discovery and replication samples, the rs927826-G allele was associated with decreased FV:C plasma levels, $\beta = -.076 \pm .016$ ($P = 1.10 \times 10^{-6}$) and $\beta = -.062 \pm .010$ ($P = 2.89 \times 10^{-10}$), respectively. When both samples were combined, the overall statistical evidence for association of rs927826 with FV:C levels reached $P = 7.54 \times 10^{-15}$, with no evidence for heterogeneity across the two samples ($P = .458$), and with the rs927826 explaining ~3.5% of FV:C plasma variability. As a reminder, the *PLXDC2* locus was the top locus in the GWAS analysis conditioned on *F5* rs6027 (Table S2) and the rs927826 ranked 4th ($P = 3.13 \times 10^{-7}$) in this conditional GWAS.

To assess whether *PLXDC2* could be a true determinant of FV levels and not a determinant of the clotting assay used to measure FV:C levels, we measured FV antigen (FV:Ag) levels by ELISA in a random sample of 60 patients from the MARTHA study. The ELISA data confirmed the specificity of the association of PLXDC2 genotypes with FV plasma levels as the rs927826-G allele was associated with significant ($P = .028$) decrease of FV:Ag levels (Table 3). Note, in this subsample, the correlation between the two FV measurements was 0.76.

We further examined the correlation between *PLXDC2* and *F5* gene expression in several genome-wide gene expression data from multiple cell lines and tissues.[12–17] In all investigated resources except macrophages, we observed positive correlations between *PLXDC2* and *F5* expressions, the strongest correlation ($r = .45$, $P = 3.94 \times 10^{-14}$) being observed in liver (Figure S5) where FV synthesis mainly occurs (Table 4). We also assessed the correlation of liver *PLXDC2* expression with that of other coagulation genes expressed in the liver *(F2*, *F7* and *F10)*. We observed a significant positive, but less strong, correlation with *F2* expression ($r = .17$, $P = .007$) and a very low correlation, if any, with *F7* and *F10* expressions (Table 4).

To strengthen the perspective of a specific effect of the rs927826 variant on FV levels, we tested its association with FII and FX plasma levels in 738 participants. Even though FII and FX were moderately correlated with FV plasma levels ($r = .35$, $P = 4.8 \times 10^{-17}$ and $r = .32$, $P = 4.40 \times 10^{-14}$ for FII and FX, respectively), we did not find any association between rs927826 and both FII ($\beta = -.002 \pm .01$, $P = .83$) and FX ($\beta = .003 \pm .01$, $P = .81$) plasma levels (Table S4).

To follow up on these genetic epidemiological findings, we conducted preliminary in vitro study to assess whether *PLXDC2* gene expression could associate with *F5* gene expression in human liver cells (see Materials and Methods). As shown in Figure 1, knock-down expression of *PLXDC2* using siRNA was associated with a significant decrease liver expression of *F5* ($P = .020$). Concurrently, expressions of other coagulation genes (*F2*, *F7*, and *F10*) were also measured, and we also observed a significant decrease of *F2* ($P = .004$) and *F10* ($P = .0003$) expressions when *PLXDC2* is silenced. *F7* expression remained unchanged ($P = .1$).

TABLE 3 Association of rs927826 genotypes with plasma FV antigen (FV:Ag) and activity (FV:C) in a sample of 60 patients from the MARTHA cohort

rs927826	FV:Ag (IU/mL)	FV:C (IU/mL)
TT (n = 20)	0.647 ± 0.163	1.182 ± 0.271
TG (n = 20)	0.621 ± 0.133	1.204 ± 0.265
GG (n = 20)	0.545 ± 0.140	1.006 ± 0.200
P value[a]	*P* = 0.0286	*P* = 0.0263

Note: Mean ± standard deviation.
[a]Associations were tested using a linear regression model adjusted for age and sex, under the assumption of an additive model.

4 | DISCUSSION

Altogether, these results strongly support the role of *PLXDC2* in the regulation of *F5* gene, and more generally in the coagulation cascade. Interestingly, the *PLXDC2* rs927826 has recently been found associated with activated partial thromboplastin time in a Japanese population, but with no formal replication of the statistical findings or experimental validation.[18] We did not observe such association in our population but did observe an association of rs927826 with prothrombin time (Table S4). This polymorphism is in very strong LD (r^2 > .80) with other *PLXDC2* SNPs (Table S5), all located in intronic regions subject to epigenetic regulation,[19] but only the rs927826 is so far predicted to affect transcription factors' binding.[20]

Little is known about the PLXDC2 gene/protein. It is recognized as a cell surface transmembrane receptor expressed in various tissues (see, e.g., GTEx Portal[21]) without so far clear arguments that it could be directly involved in mRNA regulation. We showed that *PLXDC2* expression correlated with *F5* expression in different cell lines and that *F5* mRNA expression was significantly reduced in the liver in the *PLXDC2* knock-down experiment. Interestingly, the siRNA experiment revealed that the *PLXDC2* gene expression could also modulate the liver expression of *F2* and *F10* genes that might indicate a more generalized effect on coagulation. However, we did not find any association between the rs927826 variant and either FII or FX plasma levels.

Of interest, in a recent plasma proteomic profiling study,[22] the *F5* rs6027 and *PLXDC2* rs927826 variants we found here associated

TABLE 4 Correlations between *PLXDC2* and *F2*/*F5*/*F7*/*F10* gene expressions in six different tissues

	Macrophages[a] N = 684	Whole blood[b] N = 78	Monocytes[c] N = 849	Endothelial cells[d] N = 157	Adipose tissue[e] N = 200	Liver[f] N = 253
F2	NA	−0.27 *P* = 0.02	NA	0.16 *P* = 0.05	NA	0.17 *P* = 0.007
F5	0.03 *P* = .45	0.39 *P* = 0.001	0.17 $P = 5.83 \times 10^{-6}$	0.33 $P = 4.87 \times 10^{-5}$	0.31 $P = 6.32 \times 10^{-6}$	0.45 $P = 3.94 \times 10^{-14}$
F7	−0.02 *P* = 0.56	−0.33 *P* = 0.003	−0.05 *P* = 0.18	−0.08 *P* = 0.31	−0.06 *P* = 0.37	0.03 *P* = 0.61
F10	NA	NA	NA	−0.06 *P* = 0.49	−0.08 *P* = 0.26	0.08 *P* = 0.21

NA: Expressions were not available
Corresponding publication for each study: a[12]; b[13]; c[14]; d[15]; e[16]; f[17]

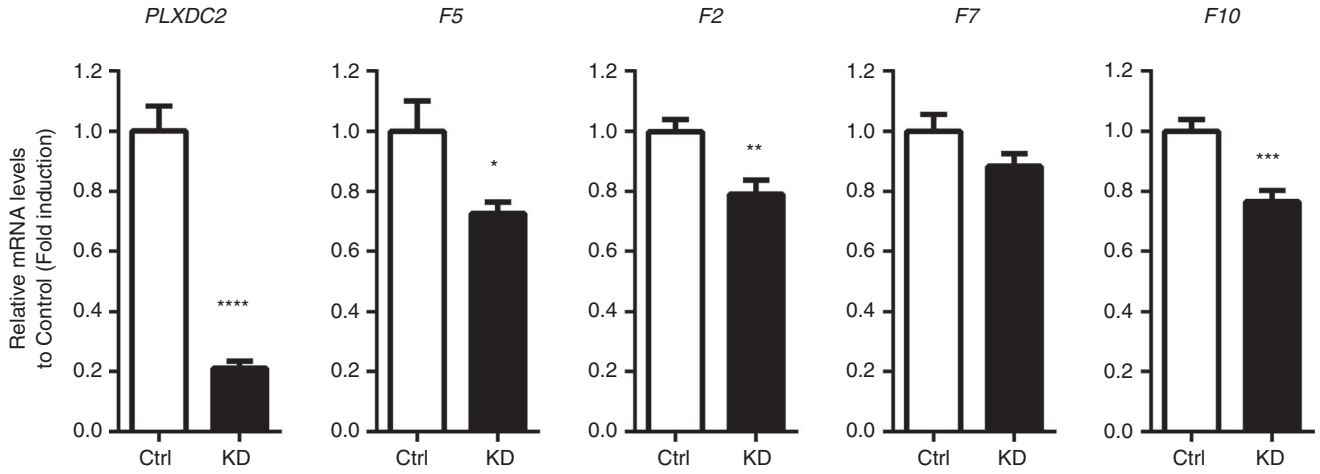


FIGURE 1 Impact of *PLXDC2* gene silencing on *F5* gene expression in liver. Abbreviation: KD = siRNA KnockDown. Quantification of *PLXDC2* and *F5*, *F2*, *F7*, and *F10* mRNA levels in Hep3B cells after 24 hours of treatment with 50 nM siRNA directed to *PLXDC2* gene (n = 9 for both controls and siRNA KD). A 79% decrease ($P < 10^{-8}$) in *PLXDC2* gene expression was followed by a significant decrease in *F5* (−27%, *P* = .02), *F2* (−21%, *P* = .004), and *F10* (−23%, *P* = .0003) gene expression, whereas *F7* expression was not significantly decreased (−12%, *P* = .11)

with plasma FV levels were both found influencing the plasma levels of the same four proteins (CD3E, CDH7, SLC22A16, and TOR1AIP1), such observations adding support for *PLXDC2* and *F5* belonging to a same regulatory pathway.

The strength of our study lies in its novelty. Indeed, there is limited information currently available regarding the regulation of plasma levels of FV. We performed the first GWAS on FV plasma levels and replicated the findings in an independent sample of subjects that limited the risk of spurious findings. Moreover, we increased the specificity of the finding by replicating the observed association between *PLXDC2* rs927826 genotypes and FV activity plasma levels by using an ELISA measuring FV antigen levels in plasma. Several limitations must be acknowledged. First, this study was performed in VT patients, and validation of the observed genetic association deserves to be investigated in healthy individuals to assess whether the observed effect size also holds in non-diseased individuals. Second, the size of our discovery cohort was relatively modest compared with what is currently done in a GWAS context, which has likely hampered our chance to detect additional loci participating to FV regulation. Additional efforts would be needed to measure FV plasma levels in independent cohorts with both available plasma and GWAS data to better characterize the genetic regulation underlying FV plasma variability. With large samples, it would also be interesting to assess the contribution of rare variants which was not possible in the current GWAS study mainly focusing on common polymorphisms. Finally, PLXDC2 is recognized as a cell-surface transmembrane receptor and it is then unclear how it could be involved in the regulation of *F5* gene expression. Moreover, the correlation between *PLXDC2* and *F2* expression in the liver and the significant decrease of *F2* and *F10* gene expressions in the *PLXDC2* knock-down experiment might indicate a more generalized effect on coagulation. Further experimental studies are mandatory to decipher the underlying mechanism.

In conclusion, all these observations point out the existence of a new player in the coagulation cascade whose exact molecular contribution needs to be extensively investigated. Its impact on FV-mediated coagulation related disorders also warrants further investigations as whole blood *PLXDC2* expression levels have been reported to be associated with stroke.[23]

ACKNOWLEDGEMENTS

F.T. was financially supported by the GENMED Laboratory of Excellence on Medical Genomics (ANR-10-LABX-0013). MARTHA genetic investigations were also supported by the GENMED laboratory of Excellence. D-A.T. was financially supported by the «EPIDEMIOM-VTE» Senior Chair from the Initiative of Excellence of the University of Bordeaux.

CONFLICT OF INTEREST

A patent deposit process has been initiated by J-F.D., D-A.T., and P-E.M. The other authors state that they have no conflict of interest.

AUTHOR CONTRIBUTIONS

All statistical analyses were performed by F. Thibord under the supervision of D-A. Trégouët who designed the study together with P.-E. Morange. Wet-lab genotyping was performed by N. Saut, FV:C measurements by M. Ibrahim-Kosta, FV:Ag measurements by L. Goumidi, and experimental works by L. Hardy under the joint supervision of D-A. Trégouët and W. Le Goff. A-S. Pulcrano-Nicolas performed additional bioinformatics analyses. M. Civelek and P. Eriksson provided supplemental datasets for replication. Genetic data acquisition was coordinated by D-A. Trégouët, J-F. Deleuze, and P-E. Morange. Manuscript was drafted by F. Thibord and D-A. Trégouët and further reviewed by all coauthors.

ORCID

Florian Thibord https://orcid.org/0000-0003-2229-8322

REFERENCES

1. Dahlbäck B. Novel insights into the regulation of coagulation by factor V isoforms, tissue factor pathway inhibitorα, and protein S. J Thromb Haemost. 2017;15:1241–50.
2. Rietveld IM, Bos MHA, Lijfering WM, Li-Gao R, Rosendaal FR, Reitsma PH, et al. Factor V levels and risk of venous thrombosis: the MEGA case-control study. Res Pract Thromb Haemost. 2018;2:320–6.
3. Vos HL. Inherited defects of coagulation Factor V: the thrombotic side. J Thromb Haemost. 2006;4:35–40.
4. Kamphuisen PW, Rosendaal FR, Eikenboom JC, Bos R, Bertina RM. Factor V antigen levels and venous thrombosis: risk profile, interaction with factor V leiden, and relation with factor VIII antigen levels. Arterioscler Thromb Vasc Biol. 2000;20:1382–6.
5. Germain M, Chasman DI, de Haan H, Tang W, Lindström S, Weng L-C, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. Am J Hum Genet. 2015;96:532–42.
6. Antoni G, Oudot-Mellakh T, Dimitromanolakis A, Germain M, Cohen W, Wells P, et al. Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. BMC Med Genet. 2011;12:102.
7. Oudot-Mellakh T, Cohen W, Germain M, Saut N, Kallel C, Zelenika D, et al. Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. Br J Haematol. 2012;157:230–9.
8. Suchon P, Germain M, Delluc A, Smadja D, Jouven X, Gyorgy B, et al. Protein S Heerlen mutation heterozygosity is associated with venous thrombosis risk. Sci Rep. 2017;7:45507.
9. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34:816–34.
10. Tregouet DA, Garelle V. A new JAVA interface implementation of THESIAS: testing haplotype effects in association studies. Bioinformatics. 2007;23:1038–9.
11. Lunghi B, Iacoviello L, Gemmati D, Dilasio MG, Castoldi E, Pinotti M, et al. Detection of new polymorphic markers in the factor V gene: association with factor V levels in plasma. Thromb Haemost. 1996;75:45–8.
12. Codoni V, Blum Y, Civelek M, Proust C, Franzén O, Björkegren JLM, et al. Preservation analysis of macrophage gene coexpression

between human and mouse identifies PARK2 as a genetically controlled master regulator of oxidative phosphorylation in humans. G3 (Bethesda). 2016;6:3361–71.
13. Pulcrano-Nicolas A-S, Proust C, Clarençon F, Jacquens A, Perret C, Roux M, et al. Whole-blood miRNA sequencing profiling for vasospasm in patients with aneurysmal subarachnoid hemorrhage. Stroke. 2018;49:2220–3.
14. Rotival M, Zeller T, Wild PS, Maouche S, Szymczak S, Schillert A, et al. Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS Genet. 2011;7.
15. Erbilgin A, Civelek M, Romanoski CE, Pan C, Hagopian R, Berliner JA, et al. Identification of CAD candidate genes in GWAS loci and their expression in vascular cells. J Lipid Res. 2013;54:1894–905.
16. Stancáková A, Civelek M, Saleem NK, Soininen P, Kangas AJ, Cederberg H, et al. Hyperglycemia and a common variant of GCKR are associated with the levels of eight amino acids in 9,369 Finnish men. Diabetes. 2012;61:1895–902.
17. Folkersen L, van't HF, Chernogubova E, Agardh HE, Hansson GK, Hedin U, et al. Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease. Circ Cardiovasc Genet 2010;3:365–73.
18. Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat Genet. 2018;50:390–400.
19. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40:D930–4.
20. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22:1790–7.
21. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648–60.
22. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558:73.
23. O'Connell GC, Petrone AB, Treadway MB, Tennant CS, Lucke-Wold N, Chantler PD, et al. Machine-learning approach identifies a pattern of gene expression in peripheral blood that can accurately detect ischaemic stroke. NPJ Genom Med. 2016;1:16038.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Thibord F, Hardy L, Ibrahim-Kosta M, et al. A Genome Wide Association Study on plasma FV levels identified *PLXDC2* as a new modifier of the coagulation process. *J Thromb Haemost*. 2019;00:1–7. https://doi.org/10.1111/jth.14562

3 Participations à d'autres travaux de recherche

Grâce à mes compétences dans le développement d'outils bioinformatiques, j'ai eu l'opportunité d'apporter mon aide sur un projet d'outil bioinformatique développé par Waqasuddin Khan, qui a réalisé un post-doctorat sous la direction de David-Alexandre Trégouët. Cet outil a pour but d'identifier des MNVs (en anglais pour *Multi Nucleotides Variants*), qui correspondent à une combinaison de deux variants ou plus, localisés au sein d'un même codon. Chaque codon, ou triplet de nucléotides, étant traduit en acide aminé selon le code génétique, la présence d'un variant génétique dans un codon peut engendrer la production d'un acide aminé différent, et donc impacter la fonction de la protéine produite. Les outils d'annotation de variants usuels permettent d'identifier les acides aminés impactés par la présence d'un variant, mais ne prennent généralement pas en compte la combinaison de plusieurs variants au sein d'un même codon: les MNVs. L'outil développé par Waqasuddin Khan, nommé MACARON, permet ainsi d'identifier ces MNVs et de prédire les acides aminés engendrés par ces MNVs. J'ai été chargé de tester cet outil sur différents jeux de données afin d'identifier certains points à améliorer, et j'ai apporté mon aide pour mettre en oeuvre ces modifications. Cet outil a fait l'objet d'une publication dans la revue *Bioinformatics*, disponible en Appendice D.

J'ai également eu l'occasion de contribuer à un projet de recherche sur les miARNs en apportant mon expertise sur le sujet pour un projet de recherche visant à identifier des miARNs pouvant servir de biomarqueurs pour le risque de développer un vasospasme chez des patients ayant subit une hémorragie sous-arachnoïdienne. Cette étude a été menée par Anne-Sophie Pulcrano-Nicolas lors de sa thèse sous la direction de David-Alexandre Trégouët, et les résultats ont fait l'objet d'une publication dans la revue *Stroke*, disponible en Appendice E.

Enfin, j'ai été invité au cours de ma thèse à participer à différents projets développés par la communauté miRtop. Le premier projet de cette communauté a aboutit avec la création du format miRGFF3, mentionné dans la section III.4, permettant d'unifier l'annotation des miARNs et isomiRs lors de la quantification par alignements des données de séquençage de petits ARNs. Un outil éponyme, miRtop, a également été créé afin d'effectuer des analyses comparatives à partir de données ayant le format miRGFF3. Un dernier projet, toujours en cours, vise a étudier l'impact du choix du kit de préparation de la librairie utilisé lors du séquençage sur la précision et la qualité de l'analyse des miARNs et des isomiRs.

Annexe A : Bases de la biologie moléculaire

1 De l'ADN aux protéines

Au début du vingtième siècle, on connaissait déjà la notion des gènes, qui définissent des caractères biologiques héritables, notamment grâce aux travaux de Gregor Mendel. Cependant, on ne savait pas comment ces caractères étaient transmis, ni par quel support. A partir des années 1950, des découvertes majeures en biologie moléculaire ont permis d'élucider ces questions, et ainsi de révolutionner la recherche en biologie et en médecine. La première étape fondamentale fût d'établir la structure de l'acide désoxyribonucléique (ADN) et son rôle en tant que support héritable de l'information génétique [77, 280]. Francis Crick propose alors que l'information génétique est encodée par une succession des quatre types de bases nucléotidiques qui composent l'ADN: adénine (A), cytosine (C), guanine (G) et thymine (T). Les couples de bases C-G ou A-T peuvent se former grâce à des liaisons hydrogènes, permettant ainsi à l'ADN d'avoir une structure en deux brins complémentaires antisens (Figure A.1). Cette structure en double brins complémentaires facilite sa réplication lors de la division cellulaire, et assure de cette façon que les cellules issues de la division héritent de la même information génétique. L'ADN est le composant principal des chromosomes, et chez l'homme, plus de 3 milliards de nucléotides sont répartis sur 23 paires de chromosomes [1] localisées dans le noyau cellulaire.

Les cellules contiennent également des macromolécules appelées protéines, qui assurent une multitude de fonctions essentielles au sein de l'organisme. Une protéine est constituée d'une séquence d'acides aminés, aussi appelée chaîne polypeptidique. Le repliement de cette chaîne dépend des affinités des acides aminés entre eux et avec le milieu, et il détermine aussi sa fonction. Il existe une vingtaine d'acides aminés [2] distincts, et leurs agencement le long d'une chaîne polypeptidique est dicté par les gènes, qui sont encodés par des segments de nucléotides le long de l'ADN. Le mécanisme permettant de traduire le flux d'information encodé dans les gènes en protéines, repose sur deux principes découverts peut après la résolution de la structure de l'ADN: le code génétique [3] [55, 189, 135] et les ARNs messagers [33, 93, 125].

Le code génétique permet d'assigner à chaque triplet de nucléotides un acide aminé, et ainsi de traduire un gène en une séquence d'acides aminés pour former une protéine. Ces triplets de nucléotides sont appelés codons, et les 64 combinaisons possibles [4] de codons permettent d'encoder les 20 acides aminés connus avec redondance (Figure A.2). Certains codons ont une

1. 22 paires d'autosomes et 1 paire de chromosomes sexuels, sans compter l'ADN mitochondrial
2. 20 acides aminés standards et 2 non standards (selenocysteine + pyrrolysine)
3. L'expression code génétique correspond généralement à la matrice permettant de traduire une suite de nucléotide en acides aminés (présentée dans la figure AI.2).
4. Il y a 64 combinaisons de triplets parmi 4 nucléotides possibles ($4^3 = 64$)

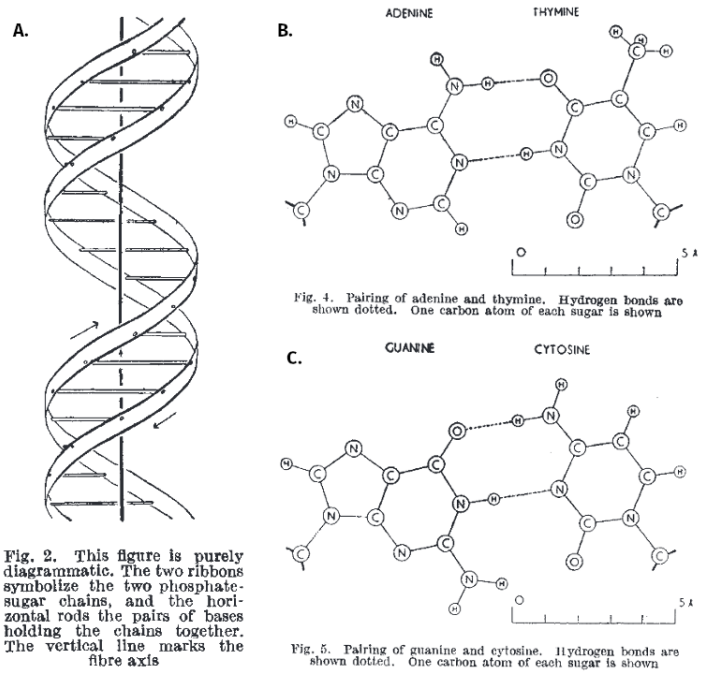


Fig. 2. This figure is purely diagrammatic. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Fig. 4. Pairing of adenine and thymine. Hydrogen bonds are shown dotted. One carbon atom of each sugar is shown

Fig. 5. Pairing of guanine and cytosine. Hydrogen bonds are shown dotted. One carbon atom of each sugar is shown

Fig. A.1 – a) Représentation schématique de la structure en double hélice de l'ADN. b) Appariement d'une Adénine et d'une Thymine. c) Appariement d'une Guanine et d'une Cytosine (après 1953, il sera découvert qu'il y a généralement 3 liaisons hydrogènes entre une guanine et une cytosine, et non deux, comme représenté ici). *Figure adaptée de l'article [280].*

fonction spécifique: le codon AUG correspond à l'acide aminé méthionine et aussi au codon "START" qui signale l'initiation de la traduction, tandis que les codons UAA, UAG et UGA correspondent à des codons STOP, qui signalent la fin de la traduction.

L'ARN (acide ribonucléique) messager (ARNm) est une copie de la séquence d'un gène réalisée par la protéine ARN polymérase [5], et contient donc la même information que celle au niveau de l'ADN. Le brin d'ARN est ainsi identique à la séquence d'ADN dupliquée, excepté les thymines qui sont substituées par des uraciles (U) dans l'ARN. Ce messager est traduit en protéine par l'intermédiaire des ribosomes et des ARNs de transferts (ARNt), qui vont permettre de construire la chaîne d'acides aminés selon la suite de codons identifiés (Figure A.3).

Le mécanisme permettant la copie d'une séquence ADN en ARN est appelé transcription, et celui permettant d'assembler une protéine à partir d'un ARN messager est appelé traduction (Figure A.4).

5. L'ARN polymérase a été découverte en 1960 Hurwitz, Stevens et Loe [121]. Il existe plusieurs protéines ARN polymérases chez l'homme, notées de I à III. La transcription des messagers est effectuée par l'ARN polymérase II

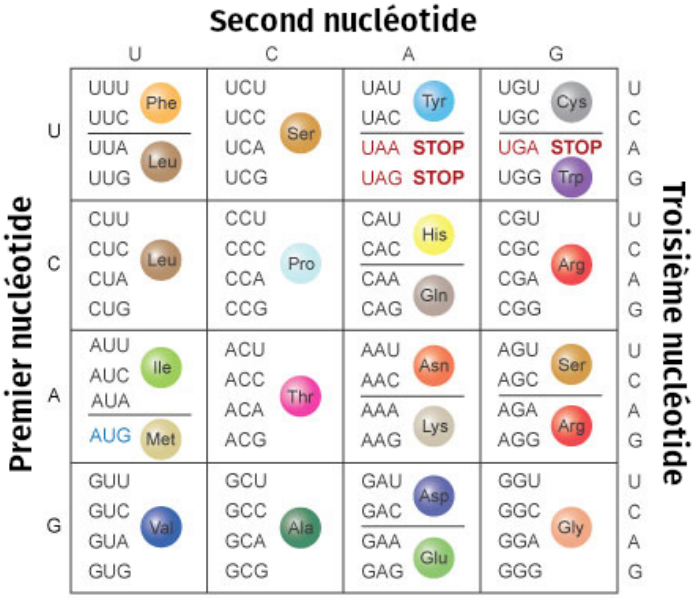


Fig. A.2 – Le code génétique. *Image de © Nature Education*

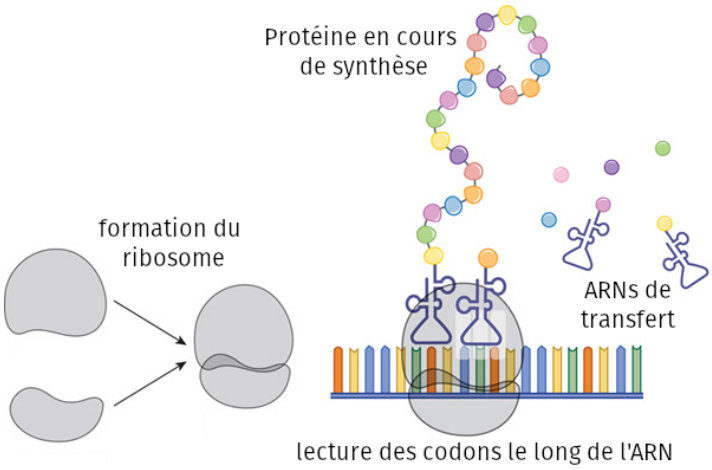


Fig. A.3 – Synthèse d'une protéine. *Image de © Nature Education*

2 Transcription et maturation du messager

Les gènes peuvent ainsi être définis comme les segments d'ADN qui sont transcrits en ARN. Lors de la transcription d'un gène, un ensemble de protéines vient se fixer en amont du segment d'ADN, sur un site spécifique appelé promoteur. Parmi ce complexe de protéines, on trouve des facteurs de transcription essentiels à l'initiation de la transcription, et l'ARN polymérase qui va générer le transcrit. Le produit de la transcription possède une structure spécifique, et un sens. Les extrémités sont notées 5' et 3', et le transcrit est toujours généré du 5' vers le 3'. La polymérase va simultanément lire la séquence ADN est produire un brin

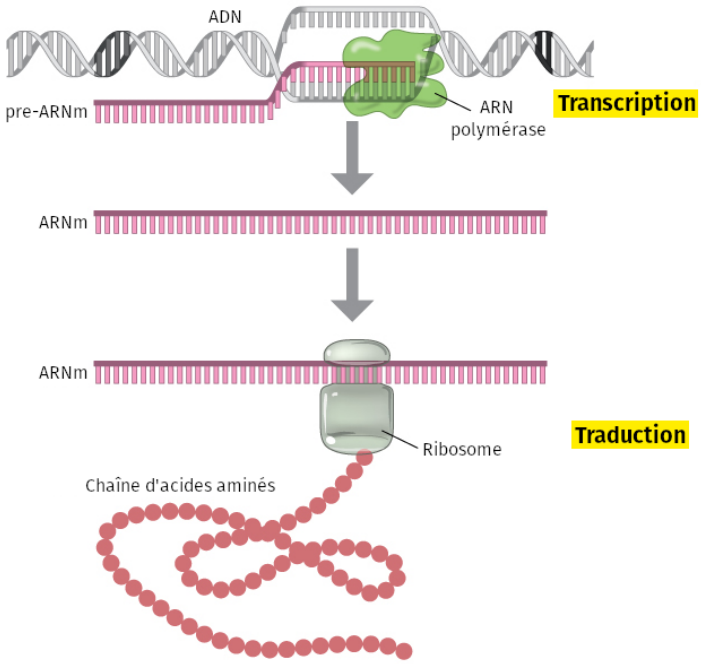


Fig. A.4 – De l'ADN à la protéine: mécanismes de transcription et de traduction. *Image de © Nature Education*

d'ARN complémentaire et antiparallèle (Figure A.5).

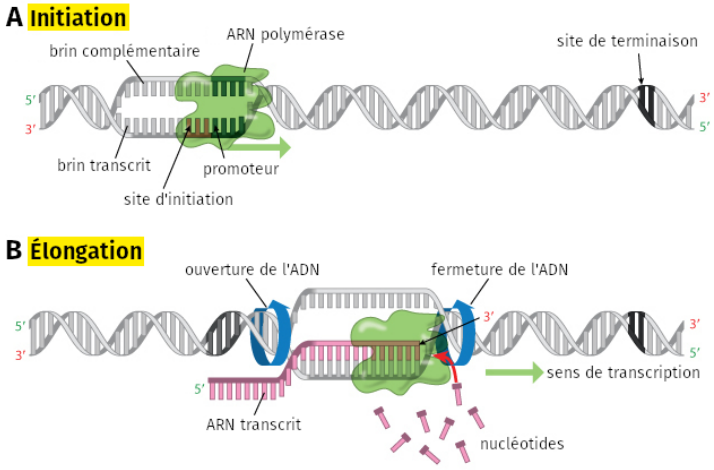


Fig. A.5 – Transcription de l'ADN en ARN. *Image de © Nature Education*

Les régions situées aux extrémités 5' et 3' du transcrit sont notées UTR (*Untranslated Regions*[6]), et ne sont pas traduites. Comme les régions UTR, certains segments de l'ARN ne sont pas traduits, mais contrairement aux UTR, ils sont dissociés et exclus du transcrit. Les régions exclues sont appelées introns, et celles retenues sont appelées exons. Le processus

6. Région non traduite, en français

d'exclusion des introns est appelé épissage.

Un ARN messager va ainsi subir plusieurs étapes de maturation avant d'être traduit en protéine. Les 3 étapes de la maturation des ARNs messagers sont: la pose d'une coiffe en 5', [7] la polyadénylation en 3', [8] et l'épissage des introns. [9] Dès le début de la transcription par l'ARN polymérase, une coiffe composée d'un guanosine méthylée triphosphate antisens (notée 7mG) est ajoutée à l'extrémité 5' naissante du transcrit. Cette coiffe permet à la fois de protéger le transcrit d'une dégradation, et également d'être reconnue par les ribosomes dans le cytoplasme pour initier la traduction. Une fois la transcription de l'ARN achevée, un complexe protéique effectue une découpe du transcrit au niveau de l'extrémité 3', puis ajoute une queue d'environ 250 adénines (notée poly-A). La queue poly-A permet, comme la coiffe, de protéger le transcrit d'une dégradation. Elle facilite également l'export du transcrit dans le cytoplasme et sa traduction. Le site d'ajout de la queue poly-A est appelé site de polyadénylation, et une majorité de gènes possèdent plusieurs sites de polyadénylation pouvant générer des transcrits alternatifs. Enfin, l'épissage des introns est effectué soit pendant ou immédiatement après la transcription par un complexe protéique appelé spliceosome, constitué de protéines et de petits ARNs. Certains exons sont parfois éliminés en même temps que les introns, générant des transcrits alternatifs, pouvant générer des protéines avec des structures et fonctions distinctes (Figure A.6). Il est estimé que 95% des transcripts avec de multiples exons subissent des épissages alternatifs [216].

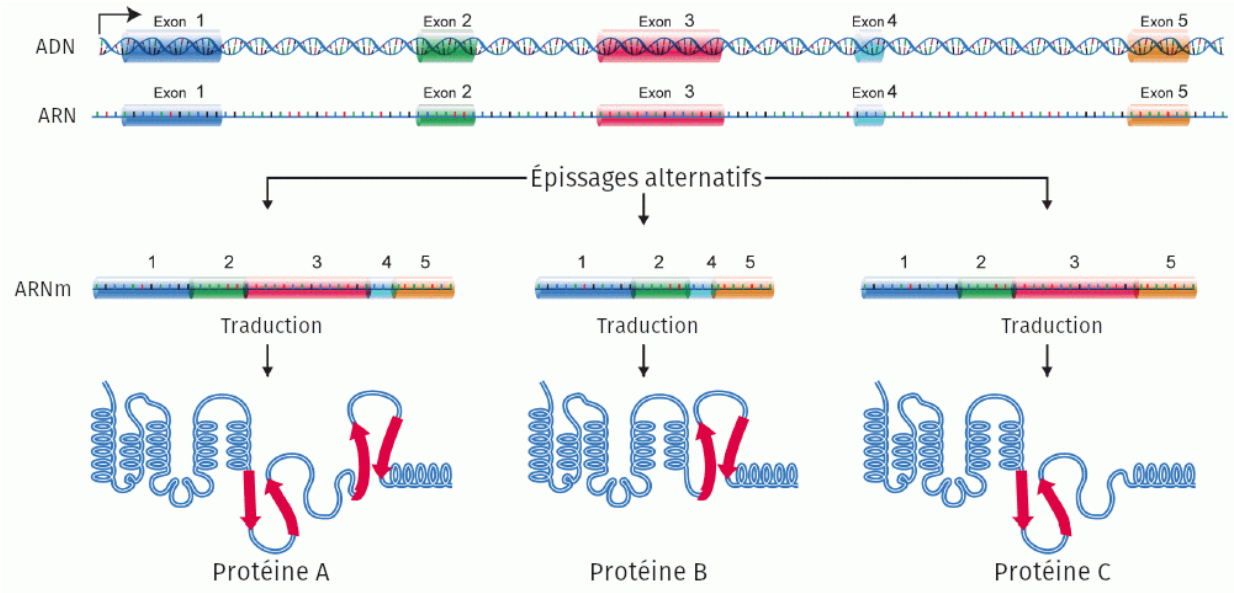


Fig. A.6 – Epissages alternatifs. *Image du NHGRI*

Ces trois modifications post-transcriptionnelles permettent ainsi d'obtenir un messager mature (Figure A.7) qui pourra être traduit en protéine une fois exporté dans le cytoplasme.

7. Découvert en 1974 indépendamment par les équipes de Reddy et Busch [19]
8. Découvert en 1971 par les équipes de Darnell, Edmond et Lee [54]
9. Découvert par les laboratoires dirigés par Sharp et Roberts en 1977 [26].

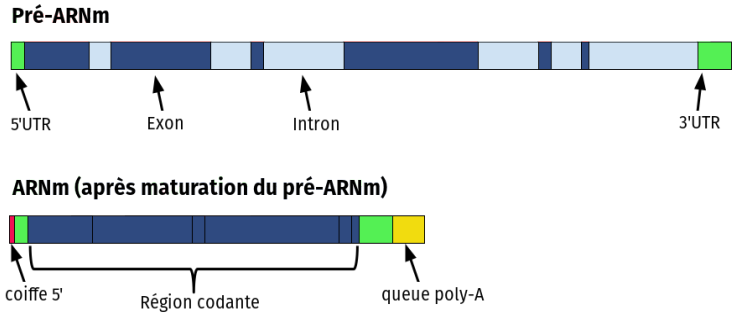


Fig. A.7 – Maturation du messager

Chez l'homme, les gènes traduits en protéines représentent moins de 3% du génome (soit approximativement 20 000 gènes), alors que plus de 83% du génome est transcrit en ARN [10]. Ce phénomène est appelé transcription pervasive, et implique qu'une très large majorité d'ARNs n'est pas traduite en protéines. Parmis ces ARNs on trouve notamment l'ARN ribosomal qui est un composant principal des ribosomes, les ARNs de transfert qui permettent de faire correspondre les codons aux acides aminés, ou les petits ARNs faisant partie du spliceosome. Les gènes correspondants à ces ARNs sont dits non-codants (ARNnc), car leurs séquence ne code pas pour une protéine. Il existe plusieurs classes d'ARNnc, et on distingue généralement les petits ARNs non codants (taille inférieure à 200bp) des long non codants. Parmi les petits non codants, on trouve en particulier les microARNs, qui régulent la traduction des messagers en protéines, en se fixant à la partie 3'UTR de l'ARN. Ils ont ainsi une fonction de régulation post-transcriptionnelle des gènes.

3 Régulation de l'expression des gènes

L'ADN s'enroule autour de protéines appelées histones, et forment ensemble la chromatine qui compose les chromosomes. Selon la densité d'histones présentes dans une région de l'ADN, la chromatine est plus ou moins accessible par les complexe protéiques permettant la transcription des gènes. Lors de la transcription, les histones sont fréquemment ré-organisées le long de la chromatine afin de libérer des régions d'ADN pour permettre la transcription. Généralement, la transcription est activée lorsque la chromatine est repliée localement et que le promoteur est en contact avec une région plus ou moins distante en amont appelée enhancer. Le contact entre l'enhancer et le promoteur est alors facilité par l'intermédiaire du complexe médiateur, aussi appelé complexe de Kornberg. La transcription est activée lorsque tous ces éléments sont en place (Figure A.8).

L'expression d'un gène dépend ainsi en premier lieu de la disponibilité de l'ensemble des acteurs de la transcription, et de l'état de la chromatine. Certains gènes peuvent également être régulé par la fixation d'un répresseur au niveau du promoteur, qui empêche la machinerie de transcription d'effectuer sa tâche. C'est notamment le cas de l'opéron [11] lactose.

10. Estimations faites en 2012 par le projet ENCODE [64]
11. Un opéron est une région de l'ADN contenant plusieurs gènes transcrits simultanément

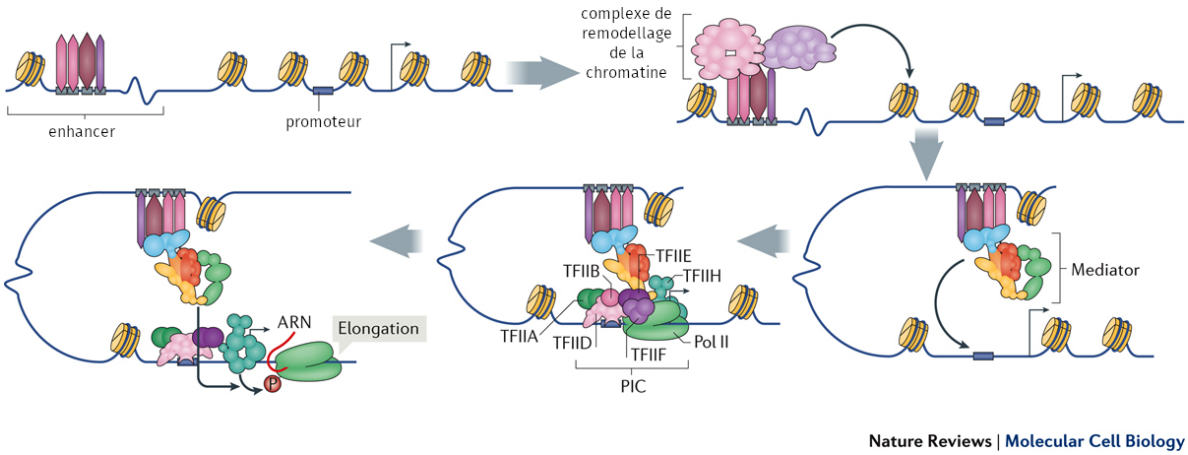


Fig. A.8 – Initiation de la transcription. *Image adaptée de [256]*

Régulation épigénétique L'accessibilité de la chromatine est principalement dictée par les histones, le complexe protéique autour duquel s'enroule l'ADN pour constituer un nucléosome, qui possède également des queues d'acides aminés accessibles par des protéines tierces. Certains acides aminés spécifiques de ces queues d'histones subissent régulièrement des modifications chimiques qui impactent la structure de la chromatine et par la même occasion le niveau d'expression des gènes à proximité. Plus d'une centaine de modifications chimiques des queues d'histones ont été recensées, les plus fréquentes étant la méthylation et l'acétylation.

L'accessibilité de la chromatine par la machinerie de transcription peut être également entravée par des modifications chimiques au niveau de l'ADN. Ces modifications concernent principalement la méthylation des cytosines. Une région fortement méthylée (hyperméthylation) est généralement associée à une inhibition de la transcription de la région. Les cytosines méthylées sont généralement suivies d'une guanine, et sont souvent densément regroupées dans des régions appelées îlots CpG généralement localisées en amont d'un promoteur.

Les modifications des queues d'histones et la méthylation de l'ADN constituent le profil épigénétiques, qui correspond à l'ensemble des modifications biochimiques qui impactent l'expression des gènes sans modifier la séquence de l'ADN, et qui sont héritables au cours de la division cellulaire.

Régulation génétique La différence d'expression génique entre individus s'explique d'une part par les conditions environnementales, et d'autre part par les variations génétiques, qui sont des différences au niveau de la séquence ADN duent à des mutations. Ces variants sont principalement le résultats d'erreurs lors de la division cellulaire et de la réplication de l'ADN, et sont aussi causées par des dommages de l'ADN (par exemple lors de l'exposition à des radiations) qui n'ont pas été correctement réparés. On peut également observer des insertions et des délétions provoquées par des éléments mobiles (transposons ADN) qui peuvent se déplacer d'un locus [12] à un autre ou se recopier le long de l'ADN.

Le génome d'un individu, composé d'environ trois milliards de nucléotides, peut être obtenu grâce aux technologies de séquençage de l'ADN. Le Human Genome Project (projet

12. Location génomique

génome humain) lancé dans les années 90 a permis de séquencer entièrement le premier génome humain, et par la même occasion de développer de nouvelles technologies de séquençage, dites NGS (Next Generation Sequencing ou séquençage de nouvelle génération). Si le Human Genome Project a mis une dizaine d'années à aboutir, la technologie NGS a permis par la suite de séquencer rapidement de nouveaux génomes, et ainsi d'établir la diversité génétique inter-individuelle. Les projets HapMap, 1000 genomes, HRC ou encore TopMed ont permis d'établir des catalogues de variants génétiques observés chez l'homme ainsi que leurs fréquence selon la population étudiée.

On distingue deux types de variants génétiques suivant leur taille. Parmi les variations génétiques à grande échelles, on peut observer des délétions, insertions, duplications, inversions et substitution de larges segments d'ADN, voire même des translocations, qui correspondent à l'échange de deux segments d'ADN sur deux chromosomes différents (Figure A.9). Ces mutations, appelées aussi réarrangements chromosomiques, sont rares excepté dans les cellules cancéreuses où elles sont fréquentes et ont une forte probabilité d'impacter la viabilité de la cellule.

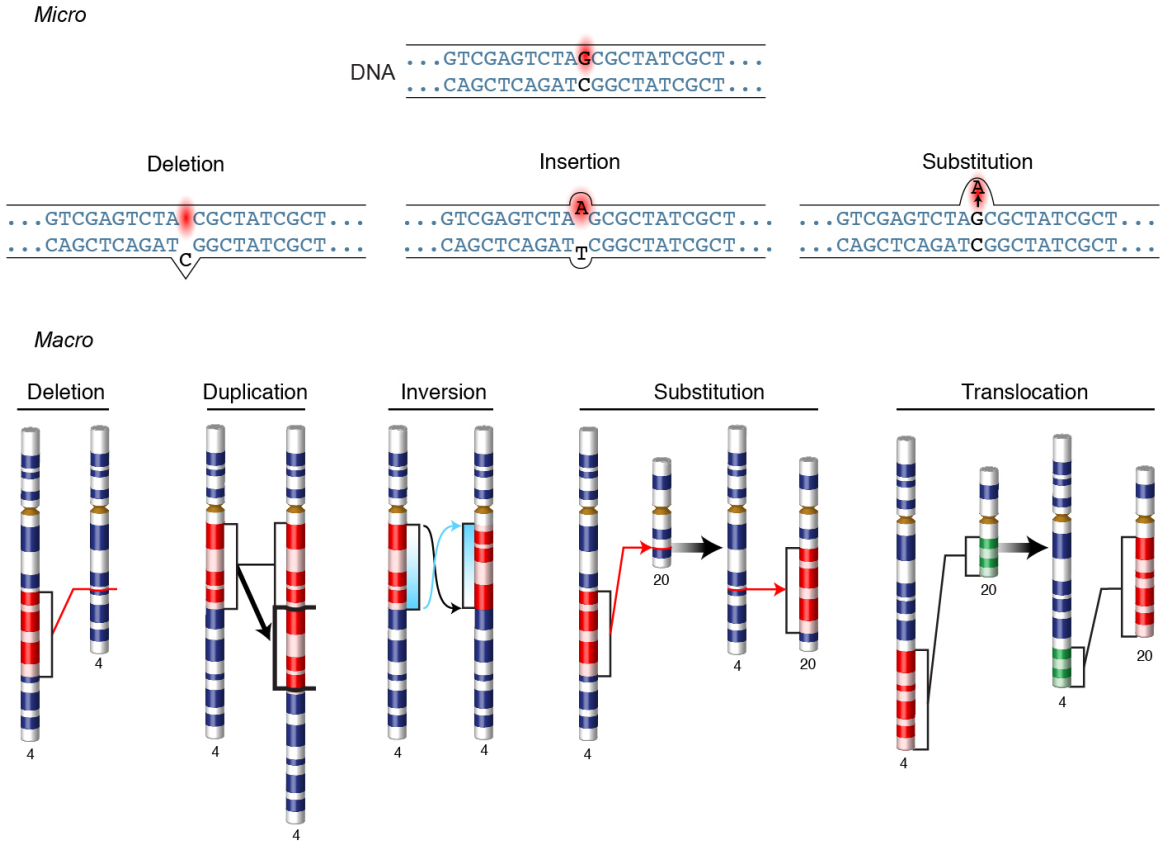


Fig. A.9 – Variations génétiques. *Images fournie par le NHGRI*

Les variations génétiques à petites échelle, qui impactent un ou quelques nucléotides sont beaucoup plus fréquentes. La substitution d'un nucléotide par un autre est la plus commune, mais on observe également fréquemment une insertion ou une délétion d'un ou plusieurs nucléotides. Ces variants sont appelée Single Nucleotide Variant (SNV), ou Single Nucleotide

Polymorphisme (SNP) lorsque le variant est fréquemment observée (fréquence $> 1\%$).

Ces variants apparaissent tout le long du génome, mais sont plus fréquemment observés dans les régions non transcrites du génome. Les variants dans les régions transcrites peuvent avoir un effet délétère, comme impacter l'expression ou la fonction du gène correspondant, et ont moins de chance d'être propagées dans une population qu'une variation neutre ou avec un effet bénéfique. Ce phénomène s'appelle la pression de sélection. Ainsi on observe davantage de variants dans les régions intergénique (entre deux gènes) qu'intragénique (au sein d'un gène). Similairement, les variants intragéniques sont plus fréquemment observés dans les introns que dans les exons.

Il faut toutefois noter que des variants dans des régions non codantes peuvent être délétères, par exemple si ils sont localisés sur des sites régulant l'expression des gènes comme les promoteurs ou enhancers. À l'inverse, les variants exoniques ne sont pas forcément délétères, car on observe souvent des substitutions avec un effet neutre, dits synonymes, qui n'impactent pas la séquence d'acide aminés de la protéine traduite grâce à la redondance du code génétique. De plus, les variants délétères peuvent subsister au sein d'une population si ils n'affectent pas la viabilité de l'organisme ou ses capacités de reproduction. Certains variants sont effectifs uniquement si les deux copies alléliques sont concernées, c'est à dire que le même variant est présent sur les deux éléments de la paire de chromosomes. Enfin, de nombreux variants ont un faible impact négatif, mais l'accumulation de variants avec un faible impact au sein d'un organisme peut avoir des effets importants.

Annexe B : *Unification of miRNA and isomiR research: the mirGFF3 format and the mirtop API*

Bioinformatics, 2019, 1–6
doi: 10.1093/bioinformatics/btz675
Advance Access Publication Date: 29 August 2019
Original Paper

OXFORD

Genome analysis

Unification of miRNA and isomiR research: the mirGFF3 format and the mirtop API

Thomas Desvignes [1,*], Phillipe Loher[2], Karen Eilbeck[3], Jeffery Ma[2], Gianvito Urgese[4], Bastian Fromm [5], Jason Sydes [1], Ernesto Aparicio-Puerta[6], Victor Barrera[7], Roderic Espín[8], Florian Thibord[9,10], Xavier Bofill-De Ros[11], Eric Londin[2], Aristeidis G. Telonis[2], Elisa Ficarra[4], Marc R. Friedländer[5], John H. Postlethwait [1], Isidore Rigoutsos[2], Michael Hackenberg[6], Ioannis S. Vlachos[12], Marc K. Halushka [13] and Lorena Pantano [14,*]

[1]Institute of Neuroscience, University of Oregon, Eugene, OR 97403, USA, [2]Computational Medicine Center, Thomas Jefferson University, Philadelphia, PA 19144, USA, [3]University of Utah, Biomedical Informatics, Salt Lake City, UT 84108, USA, [4]Department of Control and Computer Engineering, Politecnico di Torino, Torino 10129, Italy, [5]Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Stockholm 114 18, Sweden, [6]Computational Epigenomics Laboratory, Genetics Department and Biotechnology Institute and Biosanitary Institute, University of Granada, Granada 18002, Spain, [7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA, [8]Universitat Oberta de Catalunya, Barcelona 08018, Spain, [9]Sorbonne Université, Pierre Louis Doctoral School of Public Health, Paris 75006, France, [10]Institut National pour la Santé et la Recherche Médicale (INSERM) Unité Mixte de Recherche en Santé (UMR_S), University of Bordeaux, Bordeaux 33076, France, [11]RNA Biology Laboratory, Center for Cancer Research, National Cancer Institute, Frederick, MD 21702, USA, [12]Non-coding Research Lab, Department of Pathology, Cancer Research Institute, Harvard Medical School Initiative for RNA Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02115, USA, [13]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA and [14]Bioinformatics Core, The Picower Institute for Learning and Memory, Cambridge, MA 02139, USA

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty

Received on March 19, 2019; revised on July 17, 2019; editorial decision on August 24, 2019; accepted on August 28, 2019

Abstract

Motivation: MicroRNAs (miRNAs) are small RNA molecules (~22 nucleotide long) involved in post-transcriptional gene regulation. Advances in high-throughput sequencing technologies led to the discovery of isomiRs, which are miRNA sequence variants. While many miRNA-seq analysis tools exist, the diversity of output formats hinders accurate comparisons between tools and precludes data sharing and the development of common downstream analysis methods.

Results: To overcome this situation, we present here a community-based project, miRNA Transcriptomic Open Project (miRTOP) working towards the optimization of miRNA analyses. The aim of miRTOP is to promote the development of downstream isomiR analysis tools that are compatible with existing detection and quantification tools. Based on the existing GFF3 format, we first created a new standard format, mirGFF3, for the output of miRNA/isomiR detection and quantification results from small RNA-seq data. Additionally, we developed a command line Python tool, mirtop, to create and manage the mirGFF3 format. Currently, mirtop can convert into mirGFF3 the outputs of commonly used pipelines, such as seqbuster, isomiR-SEA, sRNAbench, *Prost!* as well as BAM files. Some tools have also incorporated the mirGFF3 format directly into their code, such as, miRge2.0, IsoMIRmap and OptimiR. Its open architecture enables any tool or pipeline to output or convert results into mirGFF3. Collectively, this isomiR categorization system, along with the accompanying mirGFF3 and *mirtop* API, provide a comprehensive solution for the standardization of miRNA and isomiR annotation, enabling data sharing, reporting, comparative analyses and benchmarking, while promoting the development of common miRNA methods focusing on downstream steps of miRNA detection, annotation and quantification.

Availability and implementation: https://github.com/miRTop/mirGFF3/ and https://github.com/miRTop/mirtop.
Contact: desvignes@uoneuro.uoregon.edu or lpantano@iscb.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

microRNAs (miRNAs) are the best known class of small RNAs and were discovered in the nematode worm *Caenorhabditis elegans* (Bartel, 2004, 2018). It was first reported that the gene *lin-4* generated a 22 nucleotide (nt) long RNA molecule that bound to the 3'-UTR of the *lin-14* gene transcript, thereby regulating its expression during larval development (Lee *et al.*, 1993). miRNA genes are transcribed into a primary RNA (pri-miRNA) that is processed into a hairpin-like miRNA precursor (pre-miRNA) after cutting off the 5' and 3'-tails by Drosha and DGCR8 proteins (Denli *et al.*, 2004). The pre-miRNA hairpin is then exported to the cytoplasm and processed by Dicer, which cleaves off the hairpin loop and releases a miRNA duplex about 22 nt long (Perron and Provost, 2008). Originally, it was believed that only one strand of the duplex is retained and incorporated into the RNA-induced silencing complex thereby mediating gene silencing by imperfect base pairing between the miRNA and the 3'-UTR of target messenger RNAs (mRNA) (Vella *et al.*, 2004). It was later shown that both arms of a miRNA can produce mature miRNAs, either simultaneously (Yang *et al.*, 2011), or in a tissue specific manner (Londin *et al.*, 2015; Telonis *et al.*, 2015). In addition, it was shown experimentally that both amino acid-coding sequences (Tay *et al.*, 2008) and 5'-UTRs (Zhou and Rigoutsos, 2014) can also be targeted by miRNAs. miRNAs are essential to virtually all biological processes including, but not limited to, cell differentiation, cell proliferation, cell death, fat metabolism and neuronal cell fate (Bartel, 2004, 2018). Moreover, the deficit or excess of miRNAs have been associated with several human diseases, such as myocardial infarction and different types of cancer (Ardekani and Naeini, 2010). miRNAs reside not only inside cells, but also in a variety of biofluids (Zhang *et al.*, 2018; Zhou *et al.*, 2018), which led to the suggestion that they could be used as non-invasive disease biomarkers or even therapies (Liu *et al.*, 2014; Pan *et al.*, 2018; Telonis *et al.*, 2015; Zhang *et al.*, 2018, 2019).

IsomiRs are sequence variants from annotated miRNAs (Desvignes *et al.*, 2015; Fromm *et al.*, 2015; Kim *et al.*, 2019). IsomiRs were first described by Morin *et al.* (Morin *et al.*, 2008) in human stem cell lines using next generation sequencing technologies. Sequence variations can affect different parts of the mature miRNA sequence as consequences of different biochemical processes (Pantano *et al.*, 2010). Variations at the 5' and 3'-ends can be due to imprecision of the Drosha/Dicer cutting machinery(Bofill-De Ros *et al.*, 2019b; Gu *et al.*, 2012). Moreover, it has already been shown that, in humans, these endpoint variations may differ in both healthy individuals and patients (Loher *et al.*, 2014; Telonis *et al.*, 2015; Telonis and Rigoutsos, 2018). In fact, isomiRs are likely genetically controlled because they depend on a person's sex, population origin and ethnicity (Loher *et al.*, 2014; Telonis *et al.*, 2015), as well as on tissue, tissue state and disease subtype (Magee *et al.*, 2018; Telonis *et al.*, 2017). Non-templated nucleotide additions at the 3'-end can be due to terminal uridylyl transferases that generally add adenine or uridine nucleotides (Menezes *et al.*, 2018). Variations and non-templated additions at the 3'-end can assume a new function considering that the tail end of a miRNA can contain a cell-compartment localization signal (Hwang *et al.*, 2007) or change target-specificity (Yang *et al.*, 2019). Finally, post-transcriptional processing of miRNAs can generate nucleotide changes at any position of the sequence by RNA processing enzymes, such as A-to-I editing by RNA-specific adenosine deaminase (ADAR) enzymes (Kawahara *et al.*, 2007). The specific function of isomiRs is still not well understood, but multiple studies have suggested a context-specific effect of isomiRs on gene regulation (Garate *et al.*, 2018; Menezes *et al.*, 2018; Telonis *et al.*, 2017; Trontti *et al.*, 2018). This conclusion is further supported by the fact that different isomiRs from the same mature miRNA can target virtually non-overlapping sets of transcripts (Engkvist *et al.*, 2017; Kume *et al.*, 2014; Tan *et al.*, 2014; Telonis *et al.*, 2015; Yang *et al.*, 2019).

Several tools have been developed to analyze miRNAs and their respective isomiRs (Lukasik *et al.*, 2016) (Supplementary Table S1). These tools differ in their alignment strategies, ways to handle cross-mapping events, abundance cutoffs, or isomiR annotation methods. Many tools operate by mapping sequenced reads on a curated database such as MirGeneDB (https://doi.org/10.1101/258749), miRBase, miRCarta, or RNACentral (Backes *et al.*, 2018; Fromm *et al.*, 2015; Kozomara and Griffiths-Jones, 2014; Sweeney *et al.*, 2019). Some tools allow users to provide custom database of interest: e.g. species-specific annotation, all members of the *let-7* family, miRNA precursors or shRNA products. The benefit of such an approach is its speedy execution due to the small size of the search space. Databases like isomiR Bank, re-analyze public datasets and share the annotation through a web-page (Zhang *et al.*, 2016). Each of these tools report isomiRs in a different format and with different levels of complexity (Supplementary File S1).

Seqbuster integrates its own aligner to maximize the number of isomiRs analyzed but only retains isomiRs with a maximum of one nucleotide change within the miRNA (missing the cases with more changes) and three different nucleotides at each end (Pantano *et al.*, 2010). Seqbuster outputs a tabular delimited file with a column for each isomiR type. It works with the isomiR Bioconductor package to detect expression data and isomiR differences (https://doi.org/doi:10.18129/B9.bioc.isomiRs).

isomiR-SEA (Urgese *et al.*, 2016) implements a miRNA-specific alignment procedure for comparing each read of the sample to all the miRNA sequences from miRBase and MirGeneDB, collecting uniquely and multi-mapped sequences (Fromm *et al.*, 2015; Kozomara and Griffiths-Jones, 2014). This tool annotates the positions of the variations (mismatches and indels) enabling fine categorization of each aligned read that can be classified as canonical miRNA or one of the isomiRs described in Supplementary Tables S2 and S3. isomiR-SEA then outputs a detailed isomiR expression quantification table, with a focus on the conserved miRNA-mRNA interaction sites. isomiR-SEA is implemented in C++ using functions collected in the SeqAn bioinformatics library (Reinert *et al.*, 2017).

sRNAbench (Aparicio-Puerta *et al.*, 2019) applies a bowtie seed alignment option (Langmead *et al.*, 2009), either to the genome (genome mode) or to miRNA reference sequences (library mode), to score only the first L nucleotides (by default $L = 19$ allowing one mismatch) and therefore does not take into account mismatches at the 3'-end of the read caused by any post-transcriptionally added nucleotides. sRNAbench clusters all reads that map to the reference precursor within a window of the canonical mature miRNA sequence (3 nt upstream of the start coordinate and 5 nt downstream of the end coordinate) and applies a hierarchical isomiR classification scheme. The sRNAbench tool has several tab-separated output files for isomiR analysis.

miRge2.0 maximizes isomiR discovery by iteratively mapping reads to user-defined miRNA and non-miRNA libraries using bowtie with a final step of loose alignment to the miRNA reads of any unaligned sequences (Lu *et al.*, 2018). miRge2.0, has a threshold option to remove called miRNAs whose reads are predominantly isomiR, rather than canonical, based on a user-specified threshold to correct for false positive miRNAs.

Prost! (PRocessing Of Small Transcripts) quantifies and annotates miRNA expression (Desvignes *et al.*, 2018). *Prost!* uses the global aligner BBMap (https://sourceforge.net/projects/bbmap/) to align transcripts to a user-specifiable genome assembly allowing for the identification of post-transcriptional modifications (e.g. non-templated additions, editing, alternative cutting) as well as identifying whether an isomiR can equally likely originate from one or more

genomic loci. *Prost!* then groups transcripts based on genomic location(s) and each group of sequences is annotated with user-defined databases of mature miRNAs, miRNA precursors and other types of RNAs. Genomic location groups with identical annotations are further combined and can be used for downstream differential expression analyses.

IsoMiRmap maps and quantifies isomiRs by considering both a miRNA library and the genome (ignoring miRNAs region). The IsoMiRmap tool (Loher and Rigoutsos—Personal Communication), currently in development, considers the entire genome when mapping while having modest computational requirements. Considering the entire genome has the advantage of being able to flag whether or not an isomiR is exclusive to the miRNA library or if it could have been transcribed from a gene different from that of the canonical miRNA sequence. The IsoMiRmap tool outputs in various formats, including HTML, tab separated files and mirGFF3.

OptimiR produces miRNA and isomiR expression abundances, and optionally integrates genetic information to retain or discard alignments depending on their consistency with the genotype of the sample (Thibord *et al.*, 2019). If the user provides a vcf file with genetic variants located on mature miRNAs, the miRBase reference library is automatically updated with new sequences that integrate the variants. The alignment procedure relies on bowtie2 local alignment mode, without any mismatch allowed in the central sequence (Langmead and Salzberg, 2012). A customizable score is then computed for each alignment, which resolves cross-mapping events and discards unreliable alignments.

Although the analysis of miRNAs and their isomiRs has dramatically changed over the past several years, a lack of consensus persists among bioinformatic tools used to describe and study the isomiR landscape. Tools generate different output file types with different structures and isomiR notations. This lack of homogeneity, which has an advantage in representing a diversity of ways of approaching isomiRs, however, prevents the evaluation of each tool to case-specific situations and precludes data sharing and the development of common downstream analyses that would be independent of the tool used for detection and quantification.

To overcome this situation, we present here mirGFF3, a standardized output format for the analysis of miRNAs and their isomiRs based on transcriptomic sequencing data. mirGFF3 was created to fit all research fields and as many tools as possible with the idea of democratization and standardization of data analysis. This new file format allows the storage of relevant miRNA/isomiR information and was developed based on the existing GFF3 (General Feature Format) format (https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md), commonly used in genome annotation and mRNA analyses. Importantly, mirGFF3 uses an ontological naming system to relate identified sequence features to the sequence ontology project (Eilbeck *et al.*, 2005). Moreover, we developed a Python API (mirtop) that supports general file operations as well as importing miRNA tool output files and converting and exporting them into the new mirGFF3 format to promote the development of downstream tools usable by all in a collaborative environment.

2 Results

To communicate ideas, define standards, and to develop successful formats and tools useful to the majority of researchers in the miRNA community, we created the miRNA Transcriptomic Open Project (miRTOP), an entirely open source and community-based project. The project is open for participation to any member of the miRNA community, regardless of level of seniority and status. miRTOP serves as an incubator of ideas that helps improve miRNA analysis standards and boost collaboration. All updates and progress reports are and will continue to be publicly available as they happen, and discussion summaries have already been released through GitHub. The project owns its own GitHub organization which articulates so far four different repositories: (1) the main web page, (2) the mirGFF3 format, (3) the mirtop API and (4) the incubator (https://github.com/miRTop/incubator/issues), where new ideas take form. Additional repositories will be added as projects develop. The miRTOP group uses the GitHub project web pages to organize different analyses in a transparent, communicative and inclusive manner to promote collaboration and equality among all members of the miRNA community. Crowd-supported projects have recently started to emerge in bioinformatics research (Lesurf *et al.*, 2015). As one of them, miRTOP encourages communication, collaboration and community-driver problem solving as well as decision making. The research problems are selected by the miRNA community and commonly addressed.

2.1 The mirGFF3 file format: definition and explanation

The mirGFF3 format was developed based on the original GFF3 format, taking advantage of the coordinate system information that GFF3 can handle and the possibility to store attributes in column 9 (Supplementary File S2). The GFF3 format is commonly used for the annotation of genomic coordinates and is a popular data exchange format, particularly within the Generic Model Organism Database (O'Connor *et al.*, 2008) and genome browsing applications such as Ensembl or IGV (Thorvaldsdottir *et al.*, 2013; Zerbino *et al.*, 2018). The mirGFF3 format definition and corresponding descriptions are maintained on the mirGFF3 specific GitHub page, and have been deposited in the FAIRsharing (Sansone *et al.*, 2019) and EDAM databases (Ison *et al.*, 2013). Other file formats, such as BED, BAM, or VCF files, were considered but their customization to miRNA data would have necessitated more complicated alterations to be unbiased compared to the adaptation of the original GFF3 file format. For instance, many extra columns would have been necessary to adapt the BED file format to define miRNA attribute information, and for BAM and VCF files, several different isomiR attribute tags would have to be implemented in addition to the already mandatory ones. In contrast, the original GFF3 file format already provides a structure fulfilling all the miRNA and isomiR requirements without the need to create a totally new file format or extension.

In the mirGFF3 format, the columns 'seqid', 'source', 'type', 'start', 'end' and 'strand', are used as defined in the original GFF3 format (Supplementary Table S2). The column 'type' accepts the terms 'ref_miRNA' or 'isomiR' which are part of the sequence ontology project for miRNA definition (http://www.sequenceontology.org/browser/current_release/term/SO: 0000276) as SO: 0002166 and SO: 0002167, respectively. The column 'score' is available for each tool to use freely if additional information needs to be added or specified. The column 'phase' is ignored in the mirGFF3 format given that it refers specifically to reading frame in protein coding sequences. Finally, column 9, 'attribute', was adapted to contain all the relevant information concerning the metadata that characterize each specific isomiR (Supplementary Table S3). In the mirGFF3 definition, attributes starting with a capitalized letter are reserved to the attributes listed in Supplementary Table S3, but custom attributes can be added by adding their descriptor in lower case.

mirGFF3 format accepts headers that include sample origin, names and other custom information used to parse the data by the API framework. All header lines should start with the string '##'. Four header lines are mandatory: (1) the mirGFF3 format version, (2) the database used for annotation, (3) the sample names and (4) the tool used for annotation and quantification (Fig. 1a). The database line can point to any of the already published resources and their version: miRBase (Kozomara and Griffiths-Jones, 2014), miRCarta (Backes *et al.*, 2018b), miRGeneDB (Fromm *et al.*, 2015), or a custom database. For existing databases, the version should be provided. For custom databases, an optional link to download the coordinates or precursor sequences should be provided. Sample names should be given after the character string 'COLDATA' and should contain the sample names, each separated by a ',' (comma) character (if more than one sample). The header string, 'TOOLS', is used to inform the tool used to detect miRNAs and isomiRs from the transcriptomic sequencing data. If the attribute 'Filter' is used, a line starting with the character string 'FILTER', which explains the possible values this attribute refers to, should be added for the user to filter the file content based on these criteria (Fig. 1a). In addition, we encourage users to add any header line that could provide additional useful

Fig. 1. The mirGFF3 file format and the *mirtop* API. (**a**) Example of input file header and the required lines: file format version, database used and samples included. (**b**) Examples of sequence compression to uniquely identify each sequence. (**c**) Examples of isomiRs with changes at the 5'- and 3'-ends and their respective variant attributes. The first sequence represents a portion of the miRNA precursor (i.e. pre-miRNA) and the second sequence is the reference isomiR defined by the user or the database used. Bold orange letters indicate templated additions. Bold red letters indicate non-templated additions at the 3'-end of the isomiR. Orange strikethrough letters indicate non-templated nucleotide additions at the 3'-end of the isomiR. (**d**) Example of nucleotide changes at different positions/regions of the isomiR and their respective naming. (**e**) The *mirtop* API workflow shows the main formats accepted as input files, the functions the Python API have implemented and the output file formats

information concerning the original small RNA-seq analysis. The most common information could be, but is not limited to, the command line/parameters used to generate the mirGFF3 file, the date of generation, or the description of any custom adjustments done during the previous annotation and quantification analysis.

The column 'attribute' (column 9) of the original GFF3 format was adapted to contain all the metadata relevant to miRNA/isomiR analyses [e.g. the exact sequence variation(s) that an isomiR displays, its expression in different samples, its mapping location] with the possibility to filter and classify isomiRs by mature miRNA(s) and/or pre-miRNAs(s) (Supplementary Table S3). The unambiguous identification of isomiRs between studies is still an open problem because different tools utilize different nomenclature and categorization systems. To this end, we adopted a unique identifier (UID), or 'IsomiR license plate', inspired by the MINTmap approach for tRNA fragments (Loher *et al.*, 2017; Pliatsika *et al.*, 2016) and acting as a sequence-dependent unique ID that is independent of genome assembly or species and does not require an arbitrary naming mechanism. Any isomiR sequence can be translated into a UID and any UID can be converted back to the isomiR sequence it represents (Fig. 1b). In the mirGFF3 format, a UID name, e.g. 'iso-NN-C{N}', concatenates three pieces of information, each separated by a '-' (dash). First, the prefix 'iso-' specifies that this sequence corresponds to an isomiR; second, the length (NN) of the sequence is provided and third, the suffix is the encoded nucleotide sequence. The nucleotide sequence conversion method follows the rules of MINTmap in which each nucleotide triplet is encoded into a single character therefore reducing the string length by one third.

The 'Variant' attribute constitutes another characteristic of the mirGFF3 format specifically organized to maximize clarity, communication and standardization across the community. The 'Variant' attribute follows an isomiR description and miRNA–mRNA interaction characteristic adapted from the isomiR-SEA format (Urgese *et al.*, 2016). Briefly, isomiR modification characterizations are based on a comparison of the sequence of a given isomiR to its reference miRNA in the chosen database. Changes on the 5'-end of the sequence, related to the start of the miRNA, are described as 'iso_5p' and changes on the 3'-end of the sequence, related to the tail of the miRNA, are described as 'iso_3p'. This annotation prefix is followed by details on the nucleotide changes of this isomiR compared to the provided reference. A '–' (minus sign) is used if the isomiR start or end is upstream compared to the reference extremity. In contrast, a '+' (plus sign) is used if the isomiR start or end is downstream compared to the reference endpoint (Fig. 1c) (Loher *et al.*, 2014; Telonis *et al.*, 2015). For example, an isomiR with both 'iso_5p:–1' and 'iso_3p:+1' as 'Variant' attributes would be two nucleotides longer than its reference: one nucleotide longer at the 5'-end and one nucleotide longer at the 3'-end (Fig. 1c). In both 'iso_5p' and 'iso_3p' cases, the nucleotide additions have to be templated additions, meaning that these nucleotides are encoded in the genome. In cases of 5'-non-templated additions (additions that do not match the reference genomic sequence), isomiRs are described as 'iso_add5p' (Fig. 1c). Similarly, in the case of 3'-non-templated additions, isomiRs are described as 'iso_add3p'.

Finally, isomiRs that present nucleotide changes in their sequence that do not affect their extremities are described as 'iso_snv' (single nucleotide variant). This type of isomiR is further divided into five subtypes (Fig. 1d): (1) 'iso_snv_seed', when the nucleotide variation is located in the seed of the detected isomiR between nucleotides 2 to 7; (2) 'iso_snv_central_offset', when the nucleotide variation is located at the seed offset position, at nucleotide 8, a nucleotide that is relevant to the strength of the miRNA-mRNA interaction; (3) 'iso_snv_central', when the nucleotide variation is located in the central part of the miRNA, between nucleotide 9 to 12, (4) 'iso_snv_central_supp', when the nucleotide variation is located in the supplementary region of the miRNA, between nucleotides 13 to 17 and (5) 'iso_snv', when the nucleotide variation is located in any other position in the miRNA, nucleotides 1, and 18 to the end of the miRNA.

The 'Filter' attribute was adapted from the variant caller format file, where it is used to decide whether a variant passes or not the user-defined filtering options. Each annotation and quantification tool has the possibility to attribute to each isomiR a reliability score that can be any custom value defined in the additional header lines of the mirGFF3 file.

The 'Hits' attribute is used to represent the number of times that the read name/sequence matches the database with different isomiR changes. For example, in the human genome assembly GRCh38 (International Human Genome Sequencing Consortium, 2004), iso-22-DV0Y6O6N3 (with sequence AATGCACCTGGGCAAGGAT TCT) can be attributed to both MIMAT0002871&hsa-miR-500a-3p (ATGCACCTGGGCAAGGATTCTG) and MIMAT0004775 &hsa-miR-502-3p (AATGCACCTGGGCAAGGATTCA) with different but presumably equally likely variations from the two references. In the first assignment case (hsa-miR-500a-3p), the 5'-end and 3'-end differ from the reference by one nucleotide, whereas in the second assignment case (hsa-miR-502-3p), the 3'-end differs from the reference by the insertion of one nucleotide. By setting 'Hits = 2' and representing the sequence in two lines (with 'Parent' attribute being one of the references in each line), both possible origins can be adequately captured. The 'Expression' value for the variant is set to the number of total reads for the sequence, and not a proportion of them, and the 'UID' attribute can be used to parse the file and avoid over-counting. A different example could be the isomiR iso-23-UPVMX5I80O (with sequence TACAGTAGTCTGCACATTGGT TA) that can be attributed to three different loci located on three different human chromosomes: MIMAT0004563&hsa-miR-199b-3p on chromosome 9 and MIMAT0000232&hsa-miR-199a-3p on chromosomes 1 and 19 (all three loci having ACAGTAGTCTGCAC ATTGGTTA as reference sequence). In this situation, one can set 'Hits = 3' and take a similar approach as above. Alternatively,

because this isomiR perfectly matches each genomic location in the exact same way, it could be listed in a single line with the 'Hits' attribute set to '1' and the 'Parent' attribute would be used to reflect the multiple possible origin by having the three reference names separated by a comma character.

We realized that column 9 can be overwhelmed by the number of attributes it contains; for that reason, a mirGFF3 file can be converted into a tabular format facilitating the parsing by other tools or custom scripts. In addition, mirGFF3 can be output as a GTF format changing the separator character used in the 'Attributes' column.

2.2 The mirtop API framework

The API framework '*mirtop*' was developed in Python (v.2.7 and v3.6) and uses other common bioinformatics packages. It operates BAM files (pysam) (Li *et al.*, 2009), Bed files (pybedools, bedtools) (Dale *et al.*, 2011; Quinlan, 2014) and standard IO processes with sequences (Biopython) (Cock *et al.*, 2009). The *mirtop* package is based on a central class that converts each line of the mirGFF3 file into a Python class structure, containing all the information related to each isomiR. A validation step for mirGFF3 rules and restrictions occurs at the creation of the file, avoiding errors that can be difficult to uncover later.

The *mirtop* API framework contains five different operations: importing, converting, managing, compiling and exporting (Fig. 1e). The importers in *mirtop* have so far been coded to import and convert the output files of seqbuster (bcbio-nextgen), miRge2.0, isomirSEA, sRNAbench and *Prost!* into the mirGFF3 format. Furthermore, IsoMiRmap, miRge2.0, OPTIMIR and QuagmiR (Bofill-De Ros *et al.*, 2019a) have already implemented the mirGFF3 format into their outputs. This is an indication of the short adaptation time required thanks to the use of the standard GFF3 format *mirtop* uses. The *mirtop* operator can also manage and compile mirGFF3 files allowing joining, filtering on single or multiple files and transformation of the mirGFF3 information into a count matrix. Finally, *mirtop* exporters create the final mirGFF3 file and can also convert it into other output formats commonly used for downstream analyses. Currently, *mirtop*, in addition to the mirGFF3 format, can export to FASTA, isomiRs (Bioconductor package, https://bioconductor.org/packages/release/bioc/html/isomiRs.html) and VCF formats, which are all used in a diversity of visualization and analysis tools for isomiR characterization and variant calling (http://www.internationalgenome.org/wiki/Analysis/vcf4.0/).

The conversion of several different tool outputs into a common file format will help researchers and developers focus on downstream analyses without being limited to only one quantifying tool and a specific output format. The *mirtop* API will therefore help boost the development of universal downstream analyses, enhancing the reproducibility and quality of miRNA and isomiR biology research.

3 Discussion

Here, we present a community-backed effort to standardize, homogenize and enhance the ways researchers report, share and communicate miRNA results. We have organized a community with a common goal for miRNA/isomiR result standardization, and created mirGFF3, an adapted GFF3 file format. mirGFF3 was specifically designed to contain all relevant information concerning miRNAs and isomiRs identified in small RNA-seq data, regardless of the upstream methods or downstream use-cases. This new format represents the first consensus supported by multiple experts in the field for the report of isomiR variations and abundances in one or more biological samples produced by high throughput sequencing technologies. The mirGFF3 format is complementary to existing bioinformatics tools that support GFF3 files and aligns to the transcriptomic communities that have based their mRNA annotations on GFF3 files. Similar to BAM or VCF file formats, mirGFF3 contains all the information necessary to re-analyze the data in the same way as when the raw output file from any analysis pipeline is available. The API framework, *mirtop*, which enables the conversion of miRNA quantification tool outputs and the processing of general statistics and count matrices, will serve as a catalyst for the use of the mirGFF3 format. The *mirtop* API supports any version of the mirGFF3 format and can convert older files to the latest version if needed.

The mirGFF3 file format and the *mirtop* API tool are the results of an open-membership international miRNA community created to promote open source code sharing in a collaborative and well-supported bioinformatic environment. The mirGFF3 format and associated *mirtop* API will encourage the miRNA community to develop downstream analysis protocols independent of the initial tool that was used for detection and quantification. The mirGFF3 format will provide a common entry point for a variety of applications ranging from the annotation of miRNAs/isomiRs or filtering for technical errors inherent to each library preparation protocol (Giraldez *et al.*, 2018), to visualization, variant calling, differential expression, clustering, or any other sequence analyses.

We are currently using mirGFF3 and *mirtop* to study the accuracy of isomiRs detection across laboratories, protocols and tools by re-analyzing multiple publicly available datasets (Giraldez *et al.*, 2018; Kim *et al.*, 2019; Wright *et al.*, 2019). The current status of this project can be accessed at: https://github.com/miRTop/isomir_accuracy_meta_analysis. The use of the mirGFF3 format and *mirtop* makes comparisons easier, more transparent and reproducible.

The miRTOP group is and will remain open to any researcher interested in small RNA analysis at any level, from experimental scientists to computational biologists. miRTOP was created by members of the miRNA research community for the miRNA research community and offers networking and organization to improve and to promote collaborative research.

Acknowledgements

Authors thank Peter Batzel for suggesting to us adapting the GFF3 format, Rafael Alis for helping in the debugging the mirGFF3 conversion function, Yin Lu for integrating mirGFF3 into miRge2.0 and Shruthi Bandyadka for integrating the tabular exporter operation.

Funding

This work was supported by grant PLR-1543383 and OPP-1543383 of the National Science Foundation (T.D. and J.H.P.), B.F. and M.R.F. acknowledge funding from the Strategic Research Area (SFO) program of the Swedish Research Council (VR) through Stockholm University, M.K.H. was supported by grant 1R01HL137811 of the National Institutes of Health, National Heart Lung Blood Institute, F.T. was financially supported by the GENMED laboratory of excellence on medical genomics (ANR-10-LABX-0013) and I.S.V. was supported by the George and Marie Vergottis Fellowship of Harvard Medical School.

Conflict of Interest: none declared.

References

Aparicio-Puerta,E. *et al.* (2019) sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Res.*, **47**, W530–W535.

Ardekani,A.M. and Naeini,M.M. (2010) The role of microRNAs in human diseases. *Avicenna J. Med. Biotechnol.*, **2**, 161–179.

Backes,C. *et al.* (2018) miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res.*, **46**, D160–D167.

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

Bartel,D.P. (2018) Metazoan microRNAs. *Cell*, **173**, 20–51.

Bofill-De Ros,X. *et al.* (2019a) QuagmiR: a cloud-based application for isomiR big data analytics. *Bioinformatics*, **35**, 1576–1578.

Bofill-De Ros,X. *et al.* (2019b) Structural differences between Pri-miRNA paralogs promote alternative drosha cleavage and expand target repertoires. *Cell Rep.*, **26**, 447–459.e4.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Dale,R.K. *et al.* (2011) Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, **27**, 3423–3424.
Denli,A.M. *et al.* (2004) Processing of primary microRNAs by the microprocessor complex. *Nature*, **432**, 231–235.
Desvignes,T. *et al.* (2015) miRNA nomenclature: a view incorporating genetic origins, biosynthetic pathways, and sequence variants. *Trends Genet.*, **31**, 613–626.
Desvignes,T. *et al.* (2018) miRNA analysis with Prost! Reveals evolutionary conservation of organ-enriched expression and post-transcriptional modifications in three-spined stickleback and zebrafish. *Sci. Rep.*, **9**, 2045–2322.
Eilbeck,K. *et al.* (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
Engkvist,M.E. *et al.* (2017) Analysis of the miR-34 family functions in breast cancer reveals annotation error of miR-34b. *Sci. Rep.*, **7**, 9655.
Fromm,B. *et al.* (2015) A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu. Rev. Genet.*, **49**, 213–242.
Garate,X. *et al.* (2018) Identification of the miRNAome of early mesoderm progenitor cells and cardiomyocytes derived from human pluripotent stem cells. *Sci. Rep.*, **8**, 8072.
Giraldez,M.D. *et al.* (2018) Comprehensive multi-center assessment of small RNA-seq methods for quantitative miRNA profiling. *Nat. Biotechnol.*, **36**, 746–757.
Gu,S. *et al.* (2012) The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing in vivo. *Cell*, **151**, 900–911.
Hwang,H.-W. *et al.* (2007) A hexanucleotide element directs microRNA nuclear import. *Science*, **315**, 97–100.
International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
Ison,J. *et al.* (2013) EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, **29**, 1325–1332.
Kawahara,Y. *et al.* (2007) Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science*, **315**, 1137–1140.
Kim,H. *et al.* (2019) Bias-minimized quantification of microRNA reveals widespread alternative processing and 3' end modification. *Nucleic Acids Res.*, **47**, 2630–2640.
Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**(**Database issue**), D68–D73.
Kume,H. *et al.* (2014) A-to-I editing in the miRNA seed region regulates target mRNA selection and silencing efficiency. *Nucleic Acids Res.*, **42**, 10050–10060.
Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
Lee,R.C. *et al.* (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
Lesurf,R. *et al.* (2015) ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res.*, **44**, D126–D132.
Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
Liu,N. *et al.* (2014) A four-miRNA signature identified from genome-wide serum miRNA profiling predicts survival in patients with nasopharyngeal carcinoma. *Int. J. Cancer J. Int. Du Cancer*, **134**, 1359–1368.
Loher,P. *et al.* (2014) IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies. *Oncotarget*, **5**, 8790–8802.
Loher,P. *et al.* (2017) MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Sci. Rep.*, **7**, 41184.
Londin,E. *et al.* (2015) Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc. Natl. Acad. Sci. USA*, **112**, E1106-15.
Lukasik,A. *et al.* (2016) Tools4miRs—one place to gather all the tools for miRNA analysis. *Bioinformatics*, **32**, 2722–2724.
Lu,Y. *et al.* (2018) miRge 2.0 for comprehensive analysis of microRNA sequencing data. *BMC Bioinformatics*, **19**, 275.
Magee,R.G. *et al.* (2018) Profiles of miRNA isoforms and tRNA fragments in prostate cancer. *Sci. Rep.*, **8**, 5314.
Menezes,M.R. *et al.* (2018) 3' RNA uridylation in epitranscriptomics, gene regulation, and disease. *Front. Mol. Biosci.*, **5**, 61.
Morin,R.D. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.
O'Connor,B.D. *et al.* (2008) GMODWeb: a web framework for the generic model organism database. *Genome Biol.*, **9**, R102.
Pan,J. *et al.* (2018) A two-miRNA signature (miR-33a-5p and miR-128-3p) in whole blood as potential biomarker for early diagnosis of lung cancer. *Sci. Rep.*, **8**, 16699.
Pantano,L. *et al.* (2010) SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res.*, **38**, e34.
Perron,M.P. and Provost,P. (2008) Protein interactions and complexes in human microRNA biogenesis and function. *Front. Biosci. J. Virtual Library*, **13**, 2537–2547.
Pliatsika,V. *et al.* (2016) MINTbase: a framework for the interactive exploration of mitochondrial and nuclear tRNA fragments. *Bioinformatics*, **32**, 2481–2489.
Quinlan,A.R. (2014) BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protocols Bioinformatics*, **47**, 11.12.1–34.
Reinert,K. *et al.* (2017) The SeqAn C++ template library for efficient sequence analysis: a resource for programmers. *J. Biotechnol.*, **261**, 157–168.
Sansone,S.-A. *et al.* (2019) FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.*, **37**, 358.
Sweeney,B.A. *et al.* (2019) RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res.*, **47**, D221–D229.
Tan,G.C. *et al.* (2014) 5' isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res.*, **42**, 9424–9435.
Tay,Y. *et al.* (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, **455**, 1124–1128.
Telonis,A.G. *et al.* (2015) Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. *Nucleic Acids Res.*, **43**, 9158–9175.
Telonis,A.G. *et al.* (2017) Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. *Nucleic Acids Res.*, **45**, 2973–2985.
Telonis,A.G. and Rigoutsos,I. (2018) Race disparities in the contribution of miRNA isoforms and tRNA-derived fragments to triple-negative breast cancer. *Cancer Res.*, **78**, 1140–1154.
Thibord,F. *et al.* (2019) OPTIMIR, a novel algorithm for integrating available genome-wide genotype data into miRNA sequence alignment analysis. *RNA*, **25**, 657–668.
Thorvaldsdottir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinformatics*, **14**, 178–192.
Trontti,K. *et al.* (2018) Strong conservation of inbred mouse strain microRNA loci but broad variation in brain microRNAs due to RNA editing and isomiR expression. *RNA*, **24**, 643–655.
Urgese,G. *et al.* (2016) isomiR-SEA: an RNA-Seq analysis tool for miRNAs/isomiRs expression level profiling and miRNA-mRNA interaction sites evaluation. *BMC Bioinformatics*, **17**, 148.
Vella,M.C. *et al.* (2004) Architecture of a validated microRNA: target interaction. *Chem. Biol.*, **11**, 1619–1623.
Wright,C. *et al.* (2019) Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods. *BMC Genom.*, **20**, 513.
Yang,A. *et al.* (2019) 3' Uridylation Confers miRNAs with non-canonical target repertoires. *Mol. Cel.*, **S1097–2765**, 30386–30387.
Yang,J.-S. *et al.* (2011) Widespread regulatory activity of vertebrate microRNA* species. *RNA*, **17**, 312–326.
Zerbino,D.R. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
Zhang,Y. *et al.* (2016) IsomiR Bank: a research resource for tracking IsomiRs. *Bioinformatics*, **32**, 2069–2071.
Zhang,Z. *et al.* (2018) Circular RNA: new star, new hope in cancer. *BMC Cancer*, **18**, 834.
Zhang,Y. *et al.* (2019) A 5-microRNA signature identified from serum microRNA profiling predicts survival in patients with advanced stage non-small cell lung cancer. *Carcinogenesis*, **40**, 643–650.
Zhou,H. and Rigoutsos,I. (2014) MiR-103a-3p targets the 5' UTR ofGPRC5Ain pancreatic cells. *RNA*, **20**, 1431–1439.
Zhou,X. *et al.* (2018) Plasma miRNAs in diagnosis and prognosis of pancreatic cancer: a miRNA expression analysis. *Gene*, **673**, 181–193.

Annexe C : *Minor allele of the factor V K858R variant protects from venous thrombosis only in non-carriers of factor V Leiden mutation*

SCIENTIFIC REPORTS

OPEN

Minor allele of the factor V K858R variant protects from venous thrombosis only in non-carriers of factor V Leiden mutation

Received: 7 August 2018
Accepted: 6 February 2019
Published online: 06 March 2019

M. Ibrahim-Kosta[1,2], P. Suchon[1,2], F. Couturaud[3], D. Smadja[4,5], R. Olaso[6], M. Germain[7], N. Saut[1,2,8], L. Goumidi[2], C. Derbois[6], F. Thibord[7,9], S. Debette[7,10], P. Amouyel[11], J. F. Deleuze[6,12], P. van Doorn[13], E. Castoldi[13], E. Patin[14,15,16], M. C. Alessi[1,2], D. A. Trégouët[7] & P. E. Morange[1,2,8]

Factor V serves an important role in the regulation of blood coagulation. The rs6025 (R534Q) and rs4524 (K858R) polymorphisms in the *F5* gene, are known to influence the risk of venous thrombosis. While the rare Q534 (factor V Leiden) allele is associated with an increased risk of venous thrombosis, the minor R858 allele is associated with a lower risk of disease. However, no study has deeply examined the cumulative impact of these two variations on venous thrombosis risk. We study the association of these polymorphisms with the risk of venous thrombosis in 4 French case-control populations comprising 3719 patients and 4086 controls. We demonstrate that the Q534 allele has a dominant effect over R858. Besides, we show that in individuals not carrying the Q534 allele, the protective effect of the R858 allele acts in a dominant mode. Thrombin generation-based normalized activated protein C sensitivity ratio was lower in the 858R/R homozygotes than in the 858K/K homozygotes (1.92 ± 1.61 vs 2.81 ± 1.57, $p = 0.025$). We demonstrate that the R858 allele of the *F5* rs4524 variant protects from venous thrombosis only in non-carriers of the Q534 allele of the *F5* rs6025. Its protective effect is mediated by reduced factor VIII levels and reduced activated protein C resistance.

Factor V (FV) serves an important role in the regulation of blood coagulation, having both pro- and anticoagulant properties[1]. FV circulates in blood as a precursor of activated FV (FVa), which serves as a cofactor to FXa in the activation of prothrombin to thrombin. The procoagulant activity of FVa is under strict control by activated protein C (APC), which cleaves multiple peptide bonds in FVa[2]. FV also has two identified anticoagulant activities, as a cofactor to APC in the inactivation of factor VIIIa[3] (FVIIIa) and as a cofactor to the coagulation inhibitor tissue factor pathway inhibitor (TFPIα) in the inhibition of FXa[4,5].

[1]Laboratory of Haematology, La Timone Hospital, Marseille, France. [2]C2VN, Aix Marseille Univ, INSERM, INRA, C2VN, Marseille, France. [3]University Brest, France, Department of Chest Diseases and Internal Medicine, Hôpital de la Cavale Blanche, Brest, France. [4]Service d'hématologie biologique, AP-HP, Hôpital Européen Georges Pompidou, Paris, France. [5]Université Paris Descartes, Sorbonne Paris Cité, France, Inserm, UMR-S1140, Paris, France. [6]Centre National de Recherche en Génomique Humaine, Direction de la Recherche Fondamentale, CEA, Evry, France. [7]INSERM UMR_S 1219, Bordeaux Population Health Research Center, University of Bordeaux, Bordeaux, France. [8]CRB Assistance Publique - Hôpitaux de Marseille, HemoVasc (CRB AP-HM HemoVasc), Marseille, France. [9]Sorbonne-Université, Pierre Louis Doctoral School of Public Health, Paris, France. [10]Department of Neurology, Bordeaux University Hospital, Bordeaux, France. [11]Univ. Lille, INSERM, Centre Hosp. Univ Lille, Institut Pasteur de Lille, LabEx DISTALZ-UMR1167 - RID-AGE - Risk factors and molecular determinants of aging-related diseases, Epidemiology and Public Health Department, F-Lille, France. [12]CEPH, Fondation Jean Dausset, Paris, France. [13]Department of Biochemistry, Cardiovascular Research Institute Maastricht, Maastricht University, Maastricht, The Netherlands. [14]Human Evolutionary Genetics Unit, Department of Genomes & Genetics, Institut Pasteur, Paris, France. [15]CNRS, UMR2000, Paris, France. [16]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, Paris, France. D. A. Trégouët and P. E. Morange contributed equally. Correspondence and requests for materials should be addressed to P.E.M. (email: pierre.morange@ap-hm.fr)

	rs6025 (R534Q)				rs4524 (K858R)			
	R/R	R/Q	Q/Q					
	FVL$^-$/FVL$^-$	FVL$^+$/ FVL$^-$	FVL$^+$/ FVL$^+$	*P* value[a]	K/K	K/R	R/R	*P* value[a]
EDITH								
Controls	1103 (95%)	56 (5%)	1	$8.79\ 10^{-10}$	656 (56%)	445 (38%)	69 (6%)	0.0077
Cases	1030 (88%)	138 (12%)	3		704 (61%)	396 (34%)	51 (5%)	
EOVT								
Controls	1170 (95%)	58 (5%)	0	$1.72\ 10^{-16}$	672 (55%)	477 (39%)	79 (6%)	0.010
Cases	340 (83%)	70 (17%)	1		255 (62%)	136 (33%)	20 (5%)	
FARIVE								
Controls	561 (95%)	27 (5%)	0	$8.43\ 10^{-5}$	314 (54%)	220 (38%)	44 (8%)	0.0031
Cases	532 (89%)	62 (11%)	1		363 (63%)	184 (32%)	31 (5%)	
MARTHA								
Controls	1052 (95%)	58 (5%)	0	$5.26\ 10^{-23}$	586 (53%)	460 (41%)	64 (6%)	$3.19\ 10^{-6}$
Cases	1202 (68%)	340 (22%)	0		973 (63%)	490 (32%)	79 (5%)	
COMBINED								
Controls	3886 (95%)	199 (~5%)	1	$4.37\ 10^{-63}$	2228 (55%)	1602 (39%)	256 (6%)	$2.14\ 10^{-11}$
Cases	3104 (83%)	610 (16%)	5 (1‰)		2295 (62%)	1206 (33%)	181 (5%)	

Table 1. Association of *F5* rs6025 and rs4524 with VT risk in four French case-control studies. [a]Cochran Armitage trend test's *p*-value. FVL$^-$: absence of Factor V Leiden mutation; FVL$^+$: presence of Factor V Leiden mutation.

A large number of missense polymorphisms in the *F5* gene coding for FV has been reported[6]. Among these, two genetic variations are now well established to affect the risk of venous thrombosis (VT): FV Leiden (FVL, rs6025, R534Q) identified by Bertina *et al.*[7] and the Lysine to Arginine substitution at amino acid 858 (rs4524, K858R) identified by Smith *et al.*[8]. The Q534 allele is the major genetic risk factor of VT and has a frequency of ~5% in the general population of European descent. The Q534 allele has been associated with a ~3 fold increased risk of VT through resistance to APC[9]. Conversely, the minor R858 allele of rs4524 has been associated with a protective OR (~0.8) for VT[8]. This association has been replicated in other studies and meta-analyses[10–14]. There is some evidence that the protective effect of the R858 allele is mediated through its influence on plasma APC resistance[15,16].

Although several studies have previously attempted to address the joint influence of the rs6025 and rs4524 polymorphisms on VT risk[10,12,14,15], none of them have properly taken into account the linkage disequilibrium (LD) between the two polymorphisms to accurately estimate their respective influence on disease risk. Indeed, haplotype analysis is not only adapted to detect interactive effects between polymorphisms[17,18] but is also particularly well suited to identify the true contribution on a trait of a polymorphisms from what is due to its LD with other variant(s)[19,20].

In this work, we performed a comprehensive haplotype analysis of the rs6025 and rs4524 in a case-control setting totaling 3716 VT patients and 4086 controls in order to better estimate their true impact on VT risk. In addition, we supplemented our epidemiological observations with experimental data on the functional impact of the rs4524 on the PC anticoagulant pathway.

Results

Association of rs6025 (R534Q) and rs4524 (K858R) variants with VT. Genotype distributions of the two studied *F5* variants in cases and controls are provided in Table 1. As already well documented, the presence of the Q534 allele was about 3-fold more frequent in cases than in controls (0.083 vs 0.025, $p = 4.37\ 10^{-63}$). Conversely, the minor allele (R858) of the rs4524 variant was less frequent in cases than in controls (0.213 vs 0.259, $p = 2.14\ 10^{-11}$). The Q534 allele was associated with an increased risk of VT (OR = 3.61 [3.06–4.24]), while the R858 was associated with a decreased risk of VT (OR = 0.77 [0.72–0.84]) in the combined study samples with no evidence for heterogeneity across studies ($p = 0.778$).

Linkage disequilibrium and haplotype analyses of the rs6025 (R534Q) and rs4524 (K858R) variants. The two variants were in complete negative ($D' = -1$) LD, generating 3 haplotypes: R534/K858 (H1), Q534/K858 (H2) and R534/R858 (H3). Haplotype distributions in cases and controls are shown in Table 2. These distributions were very consistent across the four studies. In agreement with the results of the single variant analyses, the unique haplotype (H2) carrying the Q534 allele was more frequent in cases than in controls, whereas the unique haplotype (H3) carrying the R858 form was less frequent in cases, homogeneously across the four studies.

Association of F5 diplotypes with VT. Because of the strong LD between the two *F5* variants, the three observed haplotypes generated 5 diplotypes, i.e pairs of haplotypes carried by a given individual. Association of these diplotypes with VT risk in the combined studies are shown in Table 3. In the absence of the Q534 allele, carrying one or two copies of the H3 haplotype was homogeneously associated with a protective OR for VT, OR = 0.78 [0.70–0.88] and OR = 0.74 [0.58–0.94], respectively. The test for heterogeneity between these two ORs was not significant ($p = 0.68$) indicating a dominant effect of the H3 haplotype in non carriers of the Q534 allele

		Controls	Cases	OR (95%CI)
EDITH		**(N = 1150)**	**(N = 1140)**	
H1	R534/K858	0.73	0.73	—
H2	Q534/K858	0.02	0.06	2.58 [1.85–3.61]
H3	R534/R858	0.25	0.21	0.85 [0.74–0.98]
EOVT		**(N = 1228)**	**(N = 411)**	
H1	R534/K858	0.72	0.70	—
H2	Q534/K858	0.02	0.09	3.51 [2.38–5.17]
H3	R534/R858	0.26	0.21	0.86 [0.70–1.04]
FARIVE		**(N = 575)**	**(N = 577)**	
H1	R534/K858	0.71	0.73	—
H2	Q534/K858	0.02	0.06	2.26 [1.41–3.61]
H3	R534/R858	0.27	0.21	0.78 [0.65–0.95]
MARTHA		**(N = 1110)**	**(N = 1542)**	
H1	R534/K858	0.71	0.68	—
H2	Q534/K858	0.03	0.11	4.96 [3.67–6.71]
H3	R534/R858	0.26	0.21	0.83 [0.73–0.95]

Table 2. Association of haplotypes derived from *F5* rs6025 (R534Q) and rs4524 (K858R) with VT risk.

Diplotype		Controls	Cases	
H1	H1	2064 (51%)	1812 (49%)	Reference
H1	H2	146 (4%)	474 (13%)	OR = 2.99 [2.36–3.79] $p = 1.58\ 10^{-19}$
H2	H3	53 (1%)	131 (4%)	OR = 2.36 [1.61–3.45] $p = 1.06\ 10^{-5}$
H2	H2	0	5 (0, 1%)	NA
H1	H3	1545 (38%)	1071 (29%)	OR = 0.78 [0.70–0.88] $p = 5.97\ 10^{-5}$
H3	H3	255 (6%)	178 (5%)	OR = 0.74 [0.58–0.94] $p = 0.013$

Table 3. Distribution of diplotypes derived from *F5* rs6025 (R534Q) rs4524 (K858R) variants in the combined cases and control population. H1 haplotype refers to the R534/K858 haplotype. H2 haplotype represents the unique haplotype carrying the FVL mutation (Q534/K858). H3 haplotype represents the unique haplotype carrying the rare R858 allele (R534/R858). Odds Ratios (OR) were adjusted for age, sex and study population.

By contrast, we observed a dominant effect of the H2 haplotype (tagging for the Q534 allele), as H2 carriers were exposed to the same VT risk whether they were carrying the H1H2 or the H2H3 diplotype, OR = 2.99 [2.36–3.79] and OR = 2.36 [1.61–3.45], respectively (p for heterogeneity = 0.30).

Association of F5 diplotypes with quantitative biological phenotypes. Association of *F5* diplotypes with plasma levels of FV and normalized Agkistrodon Contortrix Venom ratio (ACVn) from MARTHA GWAS and MARTHA12 cases are shown in Table 4. Association analyses for ACVn were conducted after excluding individuals on anticoagulant therapy. We observed a significant association between ACVn and the H2 haplotype (tagging for the Q534 allele: $p = 1.79\ 10^{-169}$ for H1H2 and $p = 1.00\ 10^{-59}$ for H2H3). Carrying a diplotype including the H3 haplotype (tagging for R858) in the absence of the Q534 allele had no effect on the quantitative biological phenotypes measured (ACVn and FV plasma levels). Only a trend toward an association between H3 haplotype and lower APC resistance in the ACVn test was observed under a dominant model ($p = 0.06$).

To get more insight into the protective effect of the rs4524 R858 allele, we measured thrombin generation at 10 pM tissue factor (TF) in the absence and presence of APC in 25 homozygous carriers of this allele (858R/R) and in 25 non-carriers (858K/K), all without Q534 allele (Fig. 1). While the area under the thrombin generation curve without APC (endogenous thrombin potential (ETP)$^{-APC}$, Fig. 1A) was similar in the two genotype groups (707.8 ± 158.3 nM.min vs 716.0 ± 139.1 nM.min, $p = 0.876$ after adjustment for age and sex), the ETP plus APC (ETP^{+APC}, Fig. 1B) was lower in the 858R/R homozygotes than in the 858K/K homozygotes (143.4 ± 133.3 nM.min vs 205.2 ± 127.7 nM.min) and the difference was close to significance after correction for age and sex ($p = 0.067$). Accordingly, the normalised APC sensitivity ratio (nAPCsr, Fig. 1C) was also lower in the 858R/R homozygotes than in the 858K/K homozygotes (1.92 ± 1.61 vs 2.81 ± 1.57), with $p = 0.025$. This indicates that the minor allele of the rs4524 polymorphism is associated with reduced plasma APC resistance in the ETP-based assay, in line with its protective effect against VT.

In contrast, no significant difference in APC sensitivity ratio (APCsr) between 858R/R and 858K/K homozygotes was observed in the Immunochrom assay (2.05 ± 0.18 vs 1.97 ± 018, $p = 0.169$ after correction for age and sex) (Fig. 1D), which specifically measures APC resistance arising from poor FVIIIa inactivation. However, FVIII levels (Fig. 1E) were somewhat lower in 858R/R homozygotes than in 858K/K homozygotes (99.7 ± 26.4 IU/dL vs 113.9 ± 19.4 IU/dL, $p = 0.039$).

Diplotype	Log ACVn*			FV plasma levels (IU/mL)		
	n		*p*	n		*p*
H1H1	450	0.051 (0.235)	Ref	480	1.09 (0.236)	Ref
H1H2	87	−0.723 (0.210)	$1.79\ 10^{-169}$	98	1.09 (0.261)	0.42
H2H3	29	−0.700 (0.193)	$1.00\ 10^{-59}$	31	1.13 (0.268)	0.20
H1H3	281	0.072 (0.261)	0.14	293	1.07 (0.210)	0.18
H3H3	39	0.096 (0.244)	0.20	42	1.05 (0.149)	0.53

Table 4. Association of *F5* diplotypes with quantitative biological phenotypes in MARTHA GWAS and MARTHA 12. H1 haplotype refers to the R534/K858 haplotype. H2 haplotype represents the unique haplotype carrying the FV Leiden mutation (Q534/K858). H3 haplotype represents the unique haplotype carrying the rare R858 allele (R534/R858). ACVn: normalized Agkistrodon Contortrix Venom ratio; FV: Factor V. Association analyses were adjusted for age, sex and MARTHA substudy group. *Analyses were conducted after excluding individuals under anticoagulant therapy.

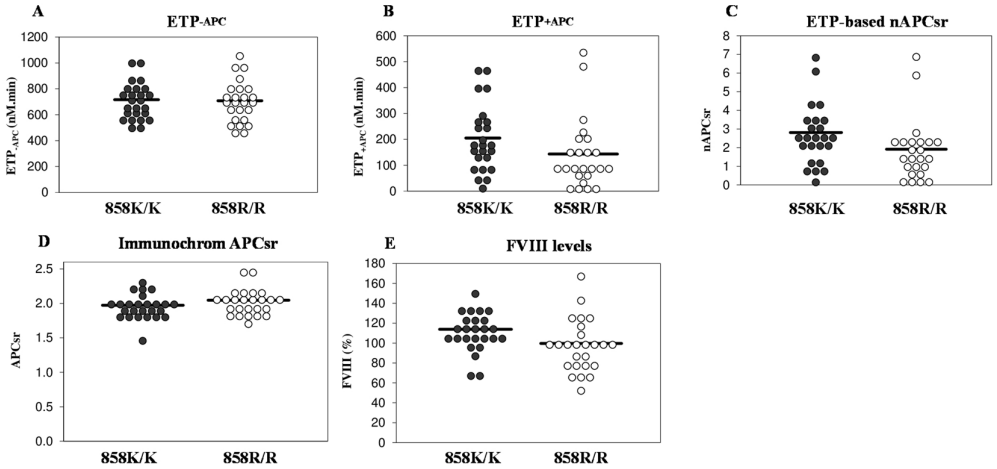


Figure 1. Coagulation parameters in 858K/K and 858R/R homozygotes. Distributions of the ETP^{-APC} (**A**), ETP^{+APC} (**B**) and nAPCsr (**C**) measured with thrombin generation-based assay, the APCsr measured with the Immunochrom assay (**D**), and FVIII levels (**E**) in 25 homozygous carriers of rs4524 (858R/R) and in 25 non-carriers (858K/K). The horizontal lines represent the means of the distributions. Please note that APC resistance increases with increasing ETP-based nAPCsr and with decreasing Immunochrom APCsr.

Discussion

In the present study, we performed the first haplotype/diplotype association analysis of the *F5* rs6025 and rs4524 polymorphisms with respect to VT risk. This analysis that efficiency takes into account the LD between the two polymorphisms for accurately estimating their true impact on VT risk, demonstrated that the R858 allele of the rs4524 is protective against VT but only in individuals not carrying the Q534 allele of the rs6025. Conversely, the Q534 allele is a risk factor for VT whatever the allele present at the rs4524 locus.

Due to a complete negative LD between the 2 variants, the R858 and the Q534 alleles are never on the same haplotype in the French studied populations.

Whereas the mechanism by which the Q534 allele induces a hypercoagulable state is well documented[1], the mechanisms underlying the protection from VT provided by the R858 allele deserve to be clarified. The K858R variant is located in the N-terminal part of the B-domain and is part of the so-called G-allele, a particular *F5* allele which is characterized by having guanines (G) instead of adenines (A) at nucleotide positions 2391, 2663 (rs4524), 2684 and 2863[6]. These variants are always co-inherited in Caucasians and the last three lead to amino acid substitutions. However, not much is known about the functional effects of most of the B-domain variants. Since the B-domain is removed upon activation of FV, an effect on the activity of FVa seems unlikely, but an effect on the anticoagulant functions of FV, when the B-domain is retained, cannot be excluded. In fact, the FV B-domain is known to be important for the APC-cofactor function of FV in the inactivation of FVIIIa[21]. In this respect, Kostka *et al.*[15] have shown that the R858 allele is less frequent than expected in patients with APC resistance in the absence of the Q534 allele. In this study, APC resistance was measured with the Immunochrom® APC response Test Kit, which specifically measures the APC-cofactor activity of FV in FVIIIa inactivation. In addition, Mingozzi *et al.*[16] have observed that the R858 allele is associated with lower APC resistance (measured with the activated partial thromboplastin time aPTT-based assay) in asymptomatic Q534 heterozygotes. In the present study we did not observe any association between the H3 haplotype (tagging the R858 allele) and the ACVn, an aPTT-based APC

resistance assay, nor with FV plasma levels. Moreover, the effect of the H2 haplotype (tagging for the Q534 allele) on APC resistance was similar irrespective of the haplotype (H1 or H3) on the counterpart allele.

To get more insight into the biological mechanisms responsible for the R858 allele, additional functional assays exploring the PC anticoagulant pathway were performed in 25 homozygous carriers of R858 and in 25 non-carriers. While we could not prove that R858/R858 homozygotes are less APC-resistant than K858/K858 homozygotes in the Immunochrom test (probably due to the insufficient number of patients tested in relation to the 'assay window', which is very narrow), we did observe reduced APC resistance in R858/R858 homozygotes using the ETP-based assay, essentially confirming that the R858 variant attenuates APC resistance. In addition, we found that FVIII levels are ~13% lower in R858/R858 homozygotes than in K858/K858 homozygotes, which could contribute to the protective effect of the R858 allele.

Apart from APC resistance, it has been recently reported that the B-domain of FV is also important for the interaction with TFPIα[22] and that it can be alternatively spliced to yield a form of FV (FV-short) with considerably increased affinity for TFPIα[23]. Therefore, it is tempting to speculate that the rs4524 polymorphism might also affect the FV-TFPIα interaction, thereby influencing the TFPIα-cofactor activity of FV in the inhibition of FXa[4,5] and/or the inhibition of FV activation and prothrombinase by TFPIα[22–24]. This could also explain why the effect of K858R on APC resistance is more easily detected with the ETP-based assay (which is triggered with tissue factor and is very sensitive to TFPIα[25] than with aPTT-based assays, which rely on the intrinsic coagulation pathway.

In conclusion, we have provided additional evidence that the common R858 variant protects against VT and that its protective effect is mediated by reduced FVIII levels and reduced APC resistance. Additional functional studies are needed to clarify whether this polymorphism also affects the interaction of FV with TFPIα and whether this interaction contributes to the protective effect of R858 on VT.

Materials and Methods

The present work was based on 4 case-control studies for VT, namely EDITH, EOVT, FARIVE, and the MARTHA Genome-Wide Association Study (GWAS), where VT events (pulmonary embolism and/or deep vein thrombosis) were objectively diagnosed. Detailed descriptions have already been published[13,26] and are summarized in the supplementary text. The association of FV diplotypes with quantitative phenotypes was assessed in two independent collections of VT cases, the MARTHA GWAS and the MARTHA12[13,27] study (described in the supplementary text).

Participants of the case-control EOVT and MARTHA GWAS studies have been typed with high density Illumina DNA arrays and imputed for 1000G reference database as part of previous Genome-Wide Association Studies[13,27]. In both studies, the rs6025 and rs4524 were well imputed (imputation criterion $r^2 > 0.80$). In the FARIVE and EDITH studies, the rs6025 and rs4524 polymorphisms were genotyped by Taqman Technology (Applied Biosystems C___ 11975250_10 and C___8919444_1 respectively). MARTHA12 participants were typed with the Illumina HumanExome BeadChip v1.0 that includes both the rs6025 and rs4524.

All patients have been informed and their consents have been obtained.

The procedures employed were reviewed and approved by the *Assistance Publique des Hopitaux de Marseille* institutional review committee. All methods were performed in accordance with the relevant guidelines and regulations.

Functional assays. In MARTHA GWAS and MARTHA12, fasting blood was drawn and biological parameters measured in platelet-poor plasma. Using those plasma, we performed several experiments in order to assess the functional impact of the rs4524 on the PC anticoagulant pathway by different assays. The ACV test is a global aPTT-based test exploring functional defects in the PC anticoagulant pathway. In MARTHA patients free from any anticoagulant treatment, the ACV test was performed as described by Robert *et al.*[28,29]. Briefly, the results were expressed as ACV ratio (ACVr) which was obtained by dividing the aPTT plus ACV (Protac®) by the aPTT. ACVn was then calculated by dividing the ACVr of the patient by the ACVr of a normal plasma control. FV plasma levels were measured by a clotting-based assay using an automated coagulometer (STA-R, Diagnostica Stago).

Among the cases of MARTHA12 that were not on anticoagulant treatment at the time of blood collection, 25 homozygous carriers of the minor allele of the rs4524 polymorphism (858R/R) and 25 sex-matched non-carriers (858K/K) were selected for functional assays. None of the selected individuals carried FVL (i.e. all were R534/R534).

For these selected patients, thrombin generation at 10 pM TF in the absence and presence of APC was measured using the Calibrated Automated Thrombography (CAT) method[30]. The area under the thrombin generation curve (in nM.min), calculated by the Thrombinoscope software, was used as the main output parameter. The APC concentration (7 nM) was chosen such as to reduce the ETP of normal pooled plasma to ~10% of its value in the absence of APC (%rest = $ETP^{+APC}/ETP^{-APC} = 10\%$). The nAPCsr was calculated by dividing the %rest of each sample plasma by the %rest of normal pooled plasma measured on the same plate. The nAPCsr varies between 0 and 10 and is directly correlated with APC resistance.

The Immunochrom APC Resistance assay, which specifically detects APC resistance arising from poor FVIIIa inactivation, was carried out as described in Brugge *et al.*[31]. This assay is based on the chromogenic measurement of FVIIIa activity before and after a standardised treatment of the (diluted) sample plasma with APC. The assay outcome is expressed as APCsr, which is defined as the ratio between the FVIIIa activity determined in the absence and presence of APC. The Immunochrom APCsr is inversely correlated with APC resistance. All samples were measured in duplicate.

Statistical analysis. The association of *F5* genotypes with VT risk was tested using the Cochran-Armitage trend test after having checked for the genotype distributions consistency with Hardy-Weinberg equilibrium. Analyses were performed separately in each study and results were then meta-analyzed using the Mantel-Haensel

methodology implementing a fixed effect model. Due to the complete negative ($D' = -1$) linkage disequilibrium between the *F5* variants, haplotypes and diplotypes could be manually reconstructed from genotypes. The statistical significance of the linkage disequilibrium D' coefficient was assessed by the maximum likelihood approach implemented in the THESIAS software[32] that relies on the methodology proposed by Thompson *et al*. 1988[33]. Associations of haplotypes/diplotypes with VT risk and quantitative phenotypes were tested using a generalized linear model and were adjusted for age, sex and study group. Heterogeneity across populations was tested using the Cochran's Q statistic.

Coagulation parameters measured in plasma were compared between carriers and non-carriers of the rs4524 polymorphism using linear regression analysis adjusted for age and sex.

References

1. Dahlbäck, B. Pro- and anticoagulant properties of factor V in pathogenesis of thrombosis and bleeding disorders. *Int. J. Lab. Hematol.* **38**(Suppl 1), 4–11 (2016).
2. Segers, K., Dahlbäck, B. & Nicolaes, G. A. F. Coagulation factor V and thrombophilia: background and mechanisms. *Thromb. Haemost.* **98**, 530–542 (2007).
3. Shen, L. & Dahlbäck, B. Factor V and protein S as synergistic cofactors to activated protein C in degradation of factor VIIIa. *J. Biol. Chem.* **269**, 18735–18738 (1994).
4. Peraramelli, S. *et al*. Role of exosite binding modulators in the inhibition of Fxa by TFPI. *Thromb. Haemost.* **115**, 580–590 (2016).
5. Santamaria, S. *et al*. Factor V has an anticoagulant cofactor activity that targets the early phase of coagulation. *J. Biol. Chem.* **292**, 9335–9344 (2017).
6. Vos, H. L. Inherited defects of coagulation Factor V: the thrombotic side. *J. Thromb. Haemost. JTH* **4**, 35–40 (2006).
7. Bertina, R. M. *et al*. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* **369**, 64–67 (1994).
8. Smith, N. L. *et al*. Association of genetic variations with nonfatal venous thrombosis in postmenopausal women. *JAMA* **297**, 489–498 (2007).
9. Morange, P. E. & Tregouet, D. A. Lessons from genome-wide association studies in venous thrombosis. *J. Thromb. Haemost. JTH* **9**(Suppl 1), 258–264 (2011).
10. Bezemer, I. D., Bare, L. A., Arellano, A. R., Reitsma, P. H. & Rosendaal, F. R. Updated analysis of gene variants associated with deep vein thrombosis. *JAMA* **303**, 421–422 (2010).
11. Dahm, A. E. A. *et al*. Candidate gene polymorphisms and the risk for pregnancy-related venous thrombosis. *Br. J. Haematol.* **157**, 753–761 (2012).
12. Smith, N. L. *et al*. Replication of findings on the association of genetic variation in 24 hemostasis genes and risk of incident venous thrombosis. *J. Thromb. Haemost. JTH* **7**, 1743–1746 (2009).
13. Germain, M. *et al*. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am. J. Hum. Genet.* **96**, 532–542 (2015).
14. Gran, O. V. *et al*. Joint effects of cancer and variants in the factor 5 gene on the risk of venous thromboembolism. *Haematologica* **101**, 1046–1053 (2016).
15. Kostka, H. *et al*. Frequency of polymorphisms in the B-domain of factor V gene in APC-resistant patients. *Thromb. Res.* **99**, 539–547 (2000).
16. Mingozzi, F. *et al*. A FV multiallelic marker detects genetic components of APC resistance contributing to venous thromboembolism in FV Leiden carriers. *Thromb. Haemost.* **89**, 983–989 (2003).
17. Pearce, E. *et al*. Haplotype effect of the matrix metalloproteinase-1 gene on risk of myocardial infarction. *Circ. Res.* **97**, 1070–1076 (2005).
18. Tregouet, D. A. *et al*. Specific haplotypes of the P-selectin gene are associated with myocardial infarction. *Hum. Mol. Genet.* **11**, 2015–2023 (2002).
19. Klerkx, A. H. *et al*. Haplotype analysis of the CETP gene: not TaqIB, but the closely linked -629C– > A polymorphism and a novel promoter variant are independently associated with CETP concentration. *Hum. Mol. Genet.* **12**, 1111–1123 (2003).
20. Soubrier, F. *et al*. High-resolution genetic mapping of the ACE-linked QTL influencing circulating ACE activity. *Eur. J. Hum. Genet.* **10**, 553–561 (2002).
21. Thorelli, E., Kaufman, R. J. & Dahlbäck, B. The C-terminal region of the factor V B-domain is crucial for the anticoagulant activity of factor V. *J. Biol. Chem.* **273**, 16140–16145 (1998).
22. Wood, J. P. *et al*. Tissue factor pathway inhibitor-alpha inhibits prothrombinase during the initiation of blood coagulation. *Proc. Natl. Acad. Sci. USA* **110**, 17838–17843 (2013).
23. Vincent, L. M. *et al*. Coagulation factor V(A2440G) causes east Texas bleeding disorder via TFPIα. *J. Clin. Invest.* **123**, 3777–3787 (2013).
24. van Doorn, P., Rosing, J., Wielders, S. J., Hackeng, T. M. & Castoldi, E. The C-terminus of tissue factor pathway inhibitor-α inhibits factor V activation by protecting the Arg1545 cleavage site. *J. Thromb. Haemost. JTH* **15**, 140–149 (2017).
25. de Visser, M. C. H. *et al*. Determinants of the APTT- and ETP-based APC sensitivity tests. *J. Thromb. Haemost. JTH* **3**, 1488–1494 (2005).
26. Suchon, P. *et al*. Protein S Heerlen mutation heterozygosity is associated with venous thrombosis risk. *Sci. Rep.* **7**, 45507 (2017).
27. Germain, M. *et al*. Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. *PloS One* **7**, e38538 (2012).
28. Robert, A., Eschwège, V., Hameg, H., Drouet, L. & Aillaud, M. F. Anticoagulant response to Agkistrodon contortrix venom (ACV test): a new global test to screen for defects in the anticoagulant protein C pathway. *Thromb. Haemost.* **75**, 562–566 (1996).
29. Oudot-Mellakh, T. *et al*. Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. *Br. J. Haematol.* **157**, 230–239 (2012).
30. Hemker, H. C. *et al*. The calibrated automated thrombogram (CAT): a universal routine test for hyper- and hypocoagulability. *Pathophysiol. Haemost. Thromb.* **32**, 249–253 (2002).
31. Brugge, J. M. *et al*. Expression of the normal factor V allele modulates the APC resistance phenotype in heterozygous carriers of the factor V Leiden mutation. *J. Thromb. Haemost. JTH* **3**, 2695–2702 (2005).
32. Tregouet, D. A. & Garelle, V. A new JAVA interface implementation of THESIAS: testing haplotype effects in association studies. *Bioinformatics.* **23**, 1038–1039 (2007).
33. Thompson, E. A. *et al*. The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. *Am. J. Hum. Genet.* **42**, 113–124 (1988).

Acknowledgements
The authors would like to thank three research programs managed by the National Research Agency (ANR) as part of the French Investment for the Future initiative genetic investigations: the GENMED Laboratory of Excellence on Medical Genomics (ANR-10-LABX-0013), the French Clinical Research Infrastructure Network on Venous Thrombo-Embolism (F-CRIN INNOVTE) and the ICAN Institute for Cardiometabolism and Nutrition (ANR-10-IAHU-05) who partially supported performing genetic investigations in the MARTHA GWAS, MARTHA12, EDITH and FARIVE studies. TF was financially supported by a PhD grant from the GENMED Laboratory of Excellence on Medical Genomics (ANR-10-LABX-0013). vDP was supported by grant nr. 2014-1 from the Dutch Thrombosis Foundation. The present project was supported by a grant from CSL Behring.

Author Contributions
I.K.M. did the selection of samples used in the functional study and wrote the paper. S.P. participated to the functional study and statistical data analysis. C.F., S.D., O.R., D.C., D.S., A.P., D.J.F. and A.M.C. designed the participating studies. G.M., P.E. and T.F. participated to the statistical data analysis. S.N. was in charge of the genotyping work. G.L. participated to the selection of samples used in the functional study and statistical data analysis. P.v.D. and C.E. participated to the functional analysis and wrote the paper. T.D.A. and M.P.E. supervised the all project and wrote the paper.

Additional Information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-40172-x.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Annexe D : *MACARON: a python framework to identify and re-annotate multi-base affected codons in whole genome/exome sequence data*

Bioinformatics, 34(19), 2018, 3396–3398
doi: 10.1093/bioinformatics/bty382
Advance Access Publication Date: 3 May 2018
Applications Note

OXFORD

Sequence analysis

MACARON: a python framework to identify and re-annotate multi-base affected codons in whole genome/exome sequence data

Waqasuddin Khan[1,2], Ganapathi Varma Saripella[1,2], Thomas Ludwig[3], Tania Cuppens[3], Florian Thibord[1,2], FREX Consortium, Emmanuelle Génin[3], Jean-Francois Deleuze[4] and David-Alexandre Trégouët[1,2,*] on behalf of the GENMED Consortium

[1]Sorbonne Universités, UPMC Université Paris 06, INSERM UMR_S 1166, F-75013 Paris, France, [2]ICAN Institute for Cardiometabolism and Nutrition, F-75013 Paris, France, [3]INSERM U1078, Génétique, Génomique Fonctionnelle et Biotechnologies, Université de Bretagne Occidentale, CHU Brest, F-29238 Brest, France and [4]Centre National de Recherche en Génomique Humaine (CNRGH), Direction de la Recherche Fondamentale, CEA, Institut de Biologie François Jacob, F-91000 Evry, France

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received and revised on April 16, 2018; editorial decision on May 1, 2018; accepted on May 2, 2018

Abstract

Summary: Predicted deleteriousness of coding variants is a frequently used criterion to filter out variants detected in next-generation sequencing projects and to select candidates impacting on the risk of human diseases. Most available dedicated tools implement a base-to-base annotation approach that could be biased in presence of several variants in the same genetic codon. We here proposed the MACARON program that, from a standard VCF file, identifies, re-annotates and predicts the amino acid change resulting from multiple single nucleotide variants (SNVs) within the same genetic codon. Applied to the whole exome dataset of 573 individuals, MACARON identifies 114 situations where multiple SNVs within a genetic codon induce an amino acid change that is different from those predicted by standard single SNV annotation tool. Such events are not uncommon and deserve to be studied in sequencing projects with inconclusive findings.
Availability and implementation: MACARON is written in python with codes available on the GENMED website (www.genmed.fr).
Contact: david-alexandre.tregouet@inserm.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Variant annotation is a crucial step in whole genome/exome sequencing analyses aimed at identifying putative causal variants, especially in a clinical context (Ding *et al.*, 2014). For example, for a rare inherited disease, one often starts to filter out detected variants according to the anticipated mode of inheritance, the type of variations (e.g. synonymous, non-synonymous, stop gain/loss, splice, etc.), allele frequencies and their predicted deleteriousness. There is a plethora of annotation tools (Cingolani *et al.*, 2012; McLaren *et al.*, 2016; Yang and Wang, 2015) but most of them implement a base-to-base approach to annotate single-nucleotide variants (SNVs). However, the presence of several SNVs at the same locus, in particular within the same genetic codon, may bias annotations. For example, two synonymous SNVs in the same codon can generate a non-synonymous variation that would be missed by standard annotation tools. To our knowledge, there is only one program, MAC

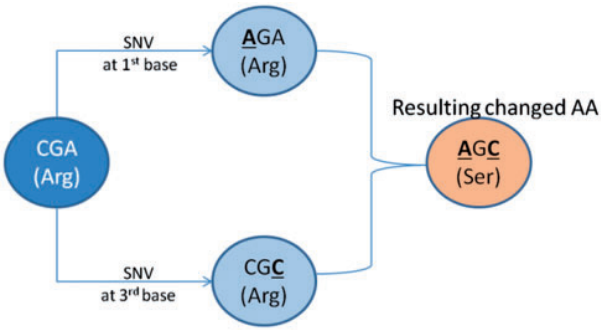


Fig. 1. Illustration of the impact of the presence of two single nucleotide variations within the same genetic codon on the resulting amino acid change

(Wei *et al.*, 2015), that accommodates multiple SNVs simultaneously. However, it is restricted to adjacent SNVs and cannot then properly address the situation when two SNVs affect the first and the third base of a genetic codon. In addition, it does not use the information on genetic code triplet structure. As a consequence, it considers the same way two SNVs affecting the adjacent bases of a genetic codon, and two SNVs affecting the last base of a codon and the first base of the next codon. To fill these gaps, we propose a simple python-based algorithm, MACARON (for **M**ulti-b**A**se **C**odon-**A**ssociated variant **R**e-annotati**ON**) to identify and to more accurately annotate multiple SNVs occurring within the same genetic codon (Fig. 1). We illustrate MACARON's relevance by an application to whole exome sequencing data of 573 subjects.

2 Implementation and application

2.1 Workflow

The overall algorithmic steps of MACARON are given below and illustrated as Supplementary Figure S1. The algorithm of MACARON is written in python language and can run on any LINUX/UNIX-like environment. Two pre-installed software, GATK (McKenna *et al.*, 2010) and SnpEff (Cingolani *et al.*, 2012) should be available for a complete run of MACARON. Briefly, MACARON starts with a VCF file as an input with no restriction on file format specifications. After identifying a list of candidate SNVs that occur within the same genetic codon along with their corrected amino acid changes, a second step consists in reading through the original BAM files to extract reads information and to confirm the presence of multiple SNVs on the same reads.

First, starting with a VCF file, MACARON utilizes GATK's VariationFiltration walker (Van der Auwera *et al.*, 2013) with parameters of –clusterSize 2 and –clusterWindowSize 3 followed by the SelectVariants tool to identify adjacent SNVs and SNVs that are 2 bps apart. Then, coding SNVs are selected based on the SnpEff functional annotation classes: SILENT, MISSENSE and NONSENSE (temp_file1). At the third step, SNVs that cluster within the same genetic codon are kept and new amino acid (AA) changes are written in temp_file2 and temp_file3. Next, clustered SNVs whose resulting AA changes are different from the original ones are stored in temp_file4. In case of a multi-sample VCF file, a scan is then performed on temp_file4 to identify clustered SNVs that are present in at least one individual. Results are stored in a final output text file containing all those SNVs identified within the same genetic codon and for which the allelic status is heterozygous or homozygous compared to the reference. At the final step, in order to confirm that identified clustered SNVs are harbored on the same reads, we used an in-house BASH-shell script (available with MACARON code) to read through the original BAM files that have been used for VCF file generation and to report the number of reads that harbor all variant alleles at the identified clustered SNVs. This script needs a subset of BAM files covering 50 bps over each clustered SNVs.

2.2 Results

MACARON was applied to the whole exome sequencing data of 573 healthy individuals as part of the FREX initiative in which 625 984 exonic SNVs were identified (Genin *et al.*, 2017). MACARON identified 114 multi-base affected codons in 194 participants. All identified affected codons were impacted by two SNVs (these were referred to as paired codon SNVs, pcSNVs) and no codon was identified that was simultaneously affected at all its 3 bases. From the identified pcSNVs, 83 were affecting codon positions 1 and 2, 23 codons were affected at positions 2 and 3 and the remaining 8 were affected at positions 1 and 3. Detailed distribution of the identified pcSNVs according to different criteria including allele frequencies, amino acid changes and predicted deleteriousness is given in Supplementary Table S1. Several observations could be made. For example, of these pcSNVs, 30 involved two rare [i.e. never reported or reported with minor allele frequency <0.01 in the gnomeAD database (Lek *et al.*, 2016)] SNVs, 15 involved one rare and one common SNV and 69 based on two common SNVs. These types of pcSNVs were referred to as 'double-rare', 'single-rare' and 'double-common' pcSNVs, respectively. The number of private (i.e. present in only one individual) pcSNVs were 16 (53%), 11 (~73%) and 3 (~4%) ~ among 'double-rare', 'single-rare' and 'double-common' pcSNVs, respectively. No pcSNV was generated from two synonymous SNVs but 26 were defined from one synonymous and one non-synonymous SNV. For 114 pcSNVs, the resulting amino acid change was different from the two original SNVs. Using the popular functional effect prediction tool SIFT (Ng and Henikoff, 2003), we observed that nine pcSNVs were predicted to be 'damaging' while the two original SNVs were predicted to be 'tolerated'. Conversely, two pcSNVs were predicted to be 'tolerated' or 'neutral' while the two original SNVs were predicted to be 'damaging'. For this application, MACARON took ~1 h on an Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60 GHz processor ×32 cores machine equipped with 64 GB of RAM on UBUNTU 16.04 LTS operating system to screen, re-annotate pcSNVs and validate them from BAM files.

3 Conclusion

MACARON is a new annotation tool for characterizing multiple SNVs within a same codon detected in WGS/WES studies. Its application to real data suggests that the frequency of pcSNVs is underappreciated and that inaccurate annotation of such genetic variations could contribute to explain inconclusive findings in DNA sequencing analyses.

Acknowledgements

Members of the GENMED and FREX consortia are listed in supplements.

Funding

This work was supported by the GENMED Laboratory of Excellence on Medical Genomics [ANR-10-LABX-0013 to WK, GV-S, FT] and the France Genomique National Infrastructure [ANR- 10-INBS-0009 to FREX consortium].

Conflict of Interest: none declared.

References

Cingolani,P. *et al*. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: sNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.

Ding,L. *et al*. (2014) Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet*., **15**, 556–570.

Genin,E. *et al*. (2017) The French Exome (FREX) Project: a population-based panel of exomes to help filter out common local variants. *Genet. Epidemiol*., **41**, 691–691.

Lek,M. *et al*. (2016) Analysis of protein-coding genetic variation in 60, 706 humans. *Nature*, **536**, 285–291.

McKenna,A. *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*., **20**, 1297–1303.

McLaren,W. *et al*. (2016) The ensembl variant effect predictor. *Genome Biol*., **17**, 122.

Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*., **31**, 3812–3814.

Van der Auwera,G.A. *et al*. (2013) From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinf*., **43**, 11.10. 1–11.10.33.

Wei,L. *et al*. (2015) MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC Genomics*, **16**, 569.

Yang,H. and Wang,K. (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc*., **10**, 1556–1566.

Annexe E : *Whole-Blood miRNA Sequencing Profiling for Vasospasm in Patients With Aneurysmal Subarachnoid Hemorrhage*

Brief Report

Whole-Blood miRNA Sequencing Profiling for Vasospasm in Patients With Aneurysmal Subarachnoid Hemorrhage

Anne-Sophie Pulcrano-Nicolas, MSc; Carole Proust; Frédéric Clarençon, MD, PhD; Alice Jacquens; Claire Perret, MSc; Maguelonne Roux; Eimad Shotar, MD; Florian Thibord, MSc; Louis Puybasset, MD, PhD; Sophie Garnier, PhD; Vincent Degos, MD, PhD; David-Alexandre Trégouët, PhD

Background and Purpose—Arterial vasospasm is a well-known delayed complication of aneurysmal subarachnoid hemorrhage (aSAH). However, no validated biomarker exists to help clinicians discriminating patients with aSAH who will develop vasospasm (VSP^+) and identifying those who then deserve aggressive preventive therapy. We hypothesized that whole-blood miRNAs could be a source of candidate biomarkers for vasospasm.

Methods—Using a next-generation sequencing approach, we performed whole-blood miRNA profiling between VSP^+ patients with aSAH and patients who did not develop vasospasm (VSP^-) in a prospective cohort of 32 patients. Profiling was performed on the admission day and 3 days before vasospasm.

Results—Four hundred forty-two miRNAs were highly expressed in whole blood of patients with aSAH. Among them, hsa-miR-3177-3p demonstrated significant ($P=5.9\times10^{-5}$; $P_{\text{Bonferroni corrected}}=0.03$) lower levels in VSP^- compared with VSP^+ patients. Looking for whole-blood mRNA correlates of hsa-miR-3177-3p, we observed some evidence that the decrease in hsa-miR-3177-3p levels after aSAH was associated with an increase in *LDHA* mRNA levels in VSP^- ($P<10^{-3}$) but not in VSP^+ ($P=0.66$) patients.

Conclusions—Whole-blood miRNA levels of hsa-miR-3177-3p could serve as a biomarker for vasospasm.

Clinical Trial Registration—URL: https://www.clinicaltrials.gov. Unique identifier: NCT01779713.

(*Stroke*. 2018;49:2220-2223. DOI: 10.1161/STROKEAHA.118.021101.)

Key Words: biomarkers ■ humans ■ microRNAs ■ prospective studies ■ vasospasm, intracranial

Intracranial aneurysm rupture is most frequently responsible for aneurysmal subarachnoid hemorrhage (aSAH), leading to a true cerebral aggression responsible for neurological insults but also impacting on many other organism's functions.[1] One of the more dreadful aSAH complications is the occurrence of cerebral vasospasm. Cerebral vasospasm consists in a thickening and temporary contraction of an artery vessel occurring in 30% of patients with aSAH, on average between 4 and 12 days after the bleeding. This contraction may lead to hypoxia, which may in turn lead to severe neurological sequela.

Although diagnostic markers have been proposed,[1,2] there are to date no validated biomarkers that can help discriminating patients with aSAH who will develop vasospasm (VSP^+) from those who will not (VSP^-). Any patient admitted in neurointensive care units for an aSAH usually undergoes an aggressive preventive treatment, consisting in an invasive monitoring and administration of a vasodilator drug, the nimodipine,[3] that is associated with severe side effects, such as cerebral and pulmonary edema.[4]

Hypothesizing that whole-blood miRNAs could be a suitable source of candidate biomarkers for vasospasm, we report here the result of the first whole-blood next-generation sequencing miRNA profiling in a cohort of 32 patients with aSAH prospectively followed for cerebral vasospasm.

Materials and Methods

VASOGENE study was registered on ClinicalTrials with the unique identifier NCT01779713. miRNA data described in this work are available in the European Genome-Phenome Archive platform under the acronym access code VASOGENE.

VASOGENE Study

The VASOGENE study was approved by its local ethics committees (Commission Nationale de l'informatique et des Libertés [CNIL] and Comité Consultatif sur le Traitement de l'Information en Matière

Received February 20, 2018; final revision received June 6, 2018; accepted June 27, 2018.

From the INSERM UMR-S 1166 (A.-S.P.-N., C. Proust, C. Perret, M.R., F.T., S.G., D.-A.T.) and Groupe de Recherche Clinique Biosfast (F.C., E.S.), Sorbonne Universités, University Pierre et Marie Curie, Université Paris 06, France; ICAN Institute of Cardiometabolism and Nutrition, Paris, France (A.-S.P.-N., C. Proust, C. Perret, M.R., F.T., S.G., D.-A.T.); Department of Neuroradiology (F.C., E.S.) and Department of Anesthesia and Intensive Care (A.J., L.P., V.D.), Pitié-Salpetrière Hospital, Assistance Publique-Hopitaux de Paris, France; and INSERM UMR 1141, Université Paris 7, France (A.J., L.P., V.D.).

The online-only Data Supplement is available with this article at https://www.ahajournals.org/doi/suppl/10.1161/STROKEAHA.118.021101.

Correspondence to David-Alexandre Trégouët, PhD, INSERM UMR-S 1166, Sorbonne Universités, University Pierre et Marie Curie, Université Paris 06, 91 Blvd de l'Hôpital, 75013 Paris, France. Email david-alexandre.tregouet@inserm.fr

de Recherche dans le Domaine de la Santé [CCTIRS]), and all VASOGENE participants provided informed written consent.

The VASOGENE cohort is composed of 89 patients with aSAH recruited from January 2013 to December 2016 at the neurointensive care unit of Pitié-Salpêtrière Hospital (Paris, France). Participants were patients with aSAH hospitalized in the 48 hours after the aneurysm rupture and treated in the first 96 hours by embolization or surgery. All patients were French individuals, excluding blacks, Hispanics, and Asians, aged ≥18 years. Patients were followed in the neurointensive care unit for at least 12 days. Each day, a transcranial Doppler sonography was performed to diagnose vasospasm. When transcranial Doppler was equivocal or for patients with poor temporal window, a digital subtraction angiography was performed to confirm the suspicion of vasospasm. For all patients with aSAH, a blood sample was collected daily from the entry in the neurointensive care unit till day 12.

mRNA/miRNA Substudy

The present study deals with a subsample of the whole VASOGENE cohort composed of 16 VSP^+ patients retrospectively matched to 16 patients with aSAH who did not develop vasospasm after 12 days (VSP^-), matching being performed as much as possible for age, sex, and hemorrhage severity. For these 16 VSP^+/VSP^- pairs, we analyzed miRNA/mRNA levels on whole-blood samples collected at the admission day (D_0) and 3 days (D_{v3}) before the day VSP^+ patients developed vasospasm (or the corresponding day for their matched VSP^- patients). Detailed description of the genome-wide gene and miRNA expression profiling is given in the online-only Data Supplement. The design of this study is summarized in Figure I in the online-only Data Supplement.

Statistical Association Analyses

Association between miRNA abundance and vasospasm was tested using a linear mixed model adjusted for age and sex (Methods in the online-only Data Supplement). A Bonferroni correction was applied to identify significant associations. miRNAs found significantly associated with the risk of vasospasm in the miRNA sequencing analysis were requantified by reverse transcription-quantitative polymerase chain reaction for technical validation of the results (online-only Data Supplement). Similar linear models were used to identify candidate mRNA correlates of significant miRNAs (Methods in the online-only Data Supplement).

Results

Clinical characteristics of the VASOGENE and of the miRNA substudy populations are shown in the Table.

In total, 1512 known mature miRNAs were detected among which only 442 were considered as expressed and tested for association with vasospasm. Full association results are summarized in the Q-Q plot shown in Figure II in the online-only Data Supplement and listed in Table I in the online-only Data Supplement. One miRNA, hsa-miR-3177-3p, was significantly ($P=5.9\times10^{-5}$; $P_{\text{Bonferroni corrected}}=0.03$) associated with the risk of vasospasm, with higher level in VSP^+ than in VSP^- patients (6.20±0.47 versus 5.62±0.61; Figure 1). Using reverse transcription-quantitative polymerase chain reaction measurements, the significant association of hsa-miR-3177-3p with vasospasm was confirmed ($P=0.03$; Figure III in the online-only Data Supplement). Looking deeply to these results revealed that hsa-miR-3177-3p levels slightly decreased between D_0 and D_{v3} in VSP^- (5.89 versus 5.41; $P=0.037$), whereas no change was observed in VSP^+ patients (6.20 versus 6.18; $P=0.63$; Figure 1).

We then scanned for mRNA expressions that could associate with hsa-miR-3177-3p levels. No single association

Table. VASOGENE Cohort

	Whole Study			Ancillary miRNA Study		
	VSP^+	VSP^-		VSP^+	VSP^-	
	n=32	n=57	*P* Value*	n=16	n=16	*P* Value*
Age, y	49.53 (10.06)	55.33 (12.00)	0.01	49.19 (10.98)	51.62 (12.70)	0.57
Female sex (%)	21 (65.63%)	39 (68.42%)	0.70	11 (68.75%)	11 (68.75%)	1.0
Smoker (%)	22 (68.75%)	28 (49.12%)	0.14	11 (68.75%)	9 (56.25%)	0.72
Fisher grade			0.16			0.11
1	2	9		2	1	
2	6	11		0	5	
3	7	6		4	3	
4	16	27		10	7	
5	1	4		0	0	
WFNS score			0.02			0.13
1	13	23		6	10	
2	14	10		8	3	
3	0	4		0	2	
4	5	13		2	1	
5	0	7		0	0	
GCS>13	26	33	0.08	13	12	0.06

Shown data: mean (SD) for quantitative variable and count (%) for qualitative variable. GCS indicates Glasgow coma scale; VSP+, vasospasm positive; VSP−, vasospasm negative; and WFNS, World Federation of Neurological Surgeons.

*Association test *P* value derived from ANOVA and χ^2 test statistics for quantitative and qualitative data, respectively.

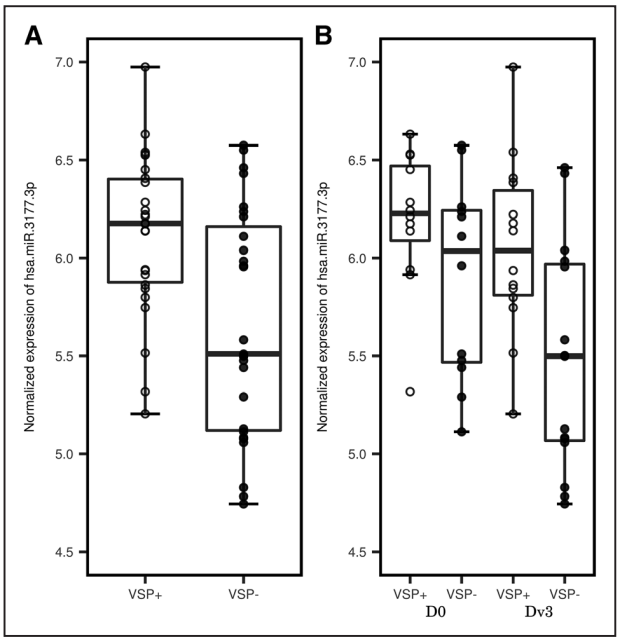


Figure 1. Whole-blood expression of hsa-miR-3177-3p in patients with aneurysmal subarachnoid hemorrhage with vasospasm (VSP$^+$) and without vasospasm (VSP$^-$) in the whole VASOGENE cohort (**A**) and separately at D_0 and D_{v3} (**B**).

reached the Bonferroni threshold of 2.3×10^{-6} (Table II in the online-only Data Supplement). However, among the 3 loci that exhibited suggestive statistical ($P<10^{-4}$) correlation with hsa-miR-3177-3p levels (Methods in the online-only Data Supplement), *LOC100506532* (ρ=0.45; $P=4.15\times10^{-5}$), *Mucin 1* (ρ=0.34; $P=4.76\times10^{-5}$), and *LDHA* (ρ=−0.38; $P=8.7\times10^{-5}$), we observed that the correlation between the mean difference of hsa-miR-3177-3p and the mean difference of *LDHA* mRNA was much stronger in VSP$^-$ (ρ=−0.81; P=0.001) than in VSP$^+$

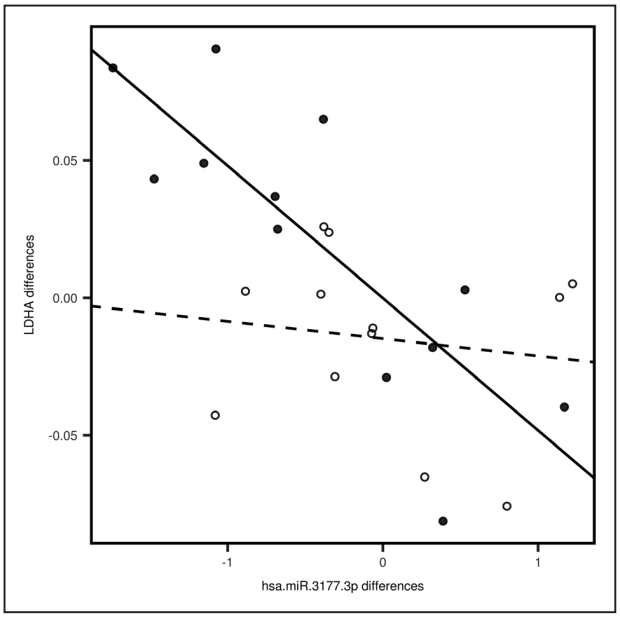


Figure 2. Correlation between changes in hsa-miR-3177-3p and in *LDHA* mRNA over time in whole-blood samples of patients with aneurysmal subarachnoid hemorrhage with vasospasm (VSP$^+$) and without vasospasm (VSP$^-$).

(ρ=−0.14; P=0.657; Figure 2). Following an opposite pattern to that observed for hsa-miR-3177-3p, *LDHA* mRNA levels were rather constant between D_0 and D_{v3} (7.35±0.02 versus 7.36±0.06; P=0.69) in VSP$^+$ but slightly increased over time in VSP$^-$ (7.33±0.04 versus 7.36±0.05; P=0.12; Figure IV in the online-only Data Supplement).

We also sought for miRNAs whose mean expression difference between D_0 and D_{v3} could differ according to the vasospasm status but did not observe any miRNA that achieved statistical significance (Table III in the online-only Data Supplement).

Discussion

We here deployed a next-generation sequencing approach to identify candidate miRNAs associated with vasospasm in whole-blood samples of patients with aSAH followed prospectively for vasospasm. To our knowledge, this is the first study using such integrative approach in the context of cerebral vasospasm and the largest cohort of patients with aSAH prospectively followed for vasospasm and studied for miRNAs and mRNAs.

This study revealed that increased hsa-miR-3177-3p levels were associated with vasospasm risk in patients with aSAH. Little is known about hsa-miR-3177-3p except it is highly expressed in the brain and cerebellum.[5] We also observed that this increase in hsa-miR-3177-3p levels was accompanied with a decrease in *LDHA* gene expression. Several works support the role of *LDHA*, which is also highly expressed in the brain,[6] as a good candidate for vasospasm. *LDHA* mRNA and protein levels have been shown to be modulated after cerebral artery occlusion in rats.[7] *LDHA* expression in brain microvascular endothelial cells was demonstrated to be influenced by hypoxia[8]—a key regulatory mechanism involved in vasospasm.[9] Finally, genetic variations at the *LDHA* locus have been reported to associate[10] with plasma concentrations of acute-phase serum amyloid A—an inflammatory marker known to be associated with cerebral disorders.[11,12]

Despite being supported by strong statistical and biological evidences, our results suffer from some limitations. The size of our cohort, despite the largest involved to date in a miRNA/mRNA study for vasospasm, is still relatively modest and our cohort limited to patients of European ancestry. Even if an association between miRNA and vasospasm reached statistical significance after multiple testing correction, we cannot exclude that we missed additional miRNA associations because of small sample size and power issues. Second, we do not provide replication of our main statistical findings in an independent cohort—a mandatory step to propose elevated hsa-miR-3177-3p levels in whole blood as a biomarker for vasospasm and to validate the association between hsa-miR-3177-3p and *LDHA*. Besides, further experimental works would be needed to investigate whether the observed association between hsa-miR-3177-3p levels and *LDHA* mRNA levels reflects a direct physical interaction between hsa-miR-3177-3p and one of its target genes or whether it involves an additional intermediate partner that remains to be identified. But this is out of the scope of the present epidemiological work.

Summary

We identified elevated hsa-miR-3177-3p levels in whole blood as candidate marker for the risk of vasospasm in patients with aSAH.

Sources of Funding

The VASOGENE study was financially supported by a grant from Comité d'orientation et de suivi des essais cliniques of the Institut National pour la Santé Et la Recherche Médicale and Association pour la recherche clinique et expérimentale en anesthésie réanimation of the La Pitié-Salpêtrière Hospital. miRNA sequencing was performed on the iGenSeq platform (Institut du Cerveau et de la Moelle épinière, Paris); thanks to a support from the European Society of Cardiology Grant for Medical Research Innovation. M. Roux and F. Thibord were financially supported by the Laboratory of Excellence on Medical Genomics (ANR-10-LABX-0013).

Disclosures

Dr Clarençon is a consultant at or reports an advisory relationship with Balt, Medtronic, and Penumbra. The other authors report no conflicts.

References

1. Weiss N, Sanchez-Peña P, Roche S, Beaudeux JL, Colonne C, Coriat P, et al. Prognosis value of plasma S100B protein levels after subarachnoid aneurysmal hemorrhage. *Anesthesiology*. 2006;104:658–666.
2. Triglia T, Mezzapesa A, Martin JC, Verdier M, Lagier D, Dufour H, et al. Early matrix metalloproteinase-9 concentration in the first 48h after aneurysmal subarachnoid haemorrhage predicts delayed cerebral ischaemia: an observational study. *Eur J Anaesthesiol*. 2016;33:662–669. doi: 10.1097/EJA.0000000000000494
3. Grosset DG, Straiton J, du Trevou M, Bullock R. Prediction of symptomatic vasospasm after subarachnoid hemorrhage by rapidly increasing transcranial Doppler velocity and cerebral blood flow changes. *Stroke*. 1992;23:674–679.
4. Sehba FA, Pluta RM, Zhang JH. Metamorphosis of subarachnoid hemorrhage research: from delayed vasospasm to early brain injury. *Mol Neurobiol*. 2011;43:27–40. doi: 10.1007/s12035-010-8155-z
5. Hinske LC, França GS, Torres HA, Ohara DT, Lopes-Ramos CM, Heyn J, et al. miRIAD–integrating microRNA inter- and intragenic data. *Database (Oxford)*. 2014;2014:bau099. doi: 10.1093/database/bau099
6. Valvona CJ, Fillmore HL, Nunn PB, Pilkington GJ. The regulation and function of lactate dehydrogenase A: therapeutic potential in brain tumor. *Brain Pathol*. 2016;26:3–17. doi: 10.1111/bpa.12299
7. Zeng XLiu N, Zhang J, Wang L, Zhang Z, Zhu J,. Inhibition of miR-143 during ischemia cerebral injury protects neurones through recovery of the hexokinase 2-mediated glucose uptake. *Biosci Rep*. 2017;37:BSR20170216. doi: 10.1042/BSR20170216
8. Shi Q, Liu X, Wang N, Zheng X, Fu J, Zheng J. Nitric oxide from brain microvascular endothelial cells may initiate the compensatory response to mild hypoxia of astrocytes in a hypoxia-inducible factor-1α dependent manner. *Am J Transl Res*. 2016;8:4735–4749.
9. Ciurea AV, Palade C, Voinescu D, Nica DA. Subarachnoid hemorrhage and cerebral vasospasm - literature review. *J Med Life*. 2013;6:120–125.
10. Marzi C, Albrecht E, Hysi PG, Lagou V, Waldenberger M, Tönjes A, et al. Genome-wide association study identifies two novel regions at 11p15.5-p13 and 1p31 with major impact on acute-phase serum amyloid A. *PLoS Genet*. 2010;6:e1001213. doi: 10.1371/journal.pgen.1001213
11. Azurmendi L, Lapierre-Fetaud V, Schneider J, Montaner J, Katan M, Sanchez JC. Proteomic discovery and verification of serum amyloid A as a predictor marker of patients at risk of post-stroke infection: a pilot study. *Clin Proteomics*. 2017;14:27. doi: 10.1186/s12014-017-9162-0
12. Brea D, Sobrino T, Blanco M, Fraga M, Agulla J, Rodríguez-Yáñez M, et al. Usefulness of haptoglobin and serum amyloid A proteins as biomarkers for atherothrombotic ischemic stroke diagnosis confirmation. *Atherosclerosis*. 2009;205:561–567. doi: 10.1016/j.atherosclerosis.2008.12.028

Bibliographie

[1] Vikram Agarwal, George W. Bell, Jin-Wu Nam, and David P. Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4, August 2015. 21

[2] H. O. Ali, A. B. Arroyo, R. González-Conejero, B. Stavik, N. Iversen, P. M. Sandset, C. Martínez, and G. Skretting. The role of microRNA-27a/b and microRNA-494 in estrogen-mediated downregulation of tissue factor pathway inhibitor. *Journal of thrombosis and haemostasis: JTH*, 14(6):1226–1237, 2016. 113

[3] Julia Alles, Tobias Fehlmann, Ulrike Fischer, Christina Backes, Valentina Galata, Marie Minet, Martin Hart, Masood Abu-Halima, Friedrich A. Grässer, Hans-Peter Lenhof, Andreas Keller, and Eckart Meese. An estimate of the total number of true human miRNAs. *Nucleic Acids Research*, 47(7):3353–3364, April 2019. 51

[4] Victor Ambros, Bonnie Bartel, David P. Bartel, Christopher B. Burge, James C. Carrington, Xuemei Chen, Gideon Dreyfuss, Sean R. Eddy, Sam Griffiths-Jones, Mhairi Marshall, Marjori Matzke, Gary Ruvkun, and Thomas Tuschl. A uniform system for microRNA annotation. *RNA*, 9(3):277–279, March 2003. 4

[5] Victor Ambros, Rosalind C. Lee, Ann Lavanway, Peter T. Williams, and David Jewell. MicroRNAs and other tiny endogenous RNAs in C. elegans. *Current biology: CB*, 13(10):807–818, May 2003. 4

[6] Stefan L. Ameres, Michael D. Horwich, Jui-Hung Hung, Jia Xu, Megha Ghildiyal, Zhiping Weng, and Phillip D. Zamore. Target RNA-directed trimming and tailing of small silencing RNAs. *Science (New York, N.Y.)*, 328(5985):1534–1539, June 2010. 25

[7] Ana B. Arroyo, Ascensión M. de Los Reyes-García, Raúl Teruel-Montoya, Vicente Vicente, Rocío González-Conejero, and Constantino Martínez. microRNAs in the haemostatic system: More than witnesses of thromboembolic diseases ? *Thrombosis Research*, 166:1–9, 2018. 87

[8] Jason D. Arroyo, John R. Chevillet, Evan M. Kroh, Ingrid K. Ruf, Colin C. Pritchard, Donald F. Gibson, Patrick S. Mitchell, Christopher F. Bennett, Era L. Pogosova-Agadjanyan, Derek L. Stirewalt, Jonathan F. Tait, and Muneesh Tewari. Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proceedings of the National Academy of Sciences of the United States of America*, 108(12):5003–5008, March 2011. 32, 33

[9] William J. Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L. Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A. Kostadima, John J. Lambourne, Suthesh Sivapalaratnam, Kate Downes, Kousik Kundu, Lorenzo Bomba, Kim Berentsen, John R. Bradley, Louise C. Daugherty, Olivier Delaneau, Kathleen Freson, Stephen F. Garner, Luigi Grassi, Jose Guerrero, Matthias Haimel, Eva M. Janssen-Megens, Anita Kaan, Mihir Kamat, Bowon Kim, Amit Mandoli, Jonathan

Marchini, Joost H. A. Martens, Stuart Meacham, Karyn Megy, Jared O'Connell, Romina Petersen, Nilofar Sharifi, Simon M. Sheard, James R. Staley, Salih Tuna, Martijn van der Ent, Klaudia Walter, Shuang-Yin Wang, Eleanor Wheeler, Steven P. Wilder, Valentina Iotchkova, Carmel Moore, Jennifer Sambrook, Hendrik G. Stunnenberg, Emanuele Di Angelantonio, Stephen Kaptoge, Taco W. Kuijpers, Enrique Carrillo-de Santa-Pau, David Juan, Daniel Rico, Alfonso Valencia, Lu Chen, Bing Ge, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yang, Roderic Guigo, Stephan Beck, Dirk S. Paul, Tomi Pastinen, David Bujold, Guillaume Bourque, Mattia Frontini, John Danesh, David J. Roberts, Willem H. Ouwehand, Adam S. Butterworth, and Nicole Soranzo. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, 167(5):1415–1429.e19, 2016. 111, 112

[10] Vincent C. Auyeung, Igor Ulitsky, Sean E. McGeary, and David P. Bartel. Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell*, 152(4):844–858, February 2013. 8

[11] Asuka Azuma-Mukai, Hideo Oguri, Toutai Mituyama, Zhi Rong Qian, Kiyoshi Asai, Haruhiko Siomi, and Mikiko C. Siomi. Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. *Proceedings of the National Academy of Sciences of the United States of America*, 105(23):7964–7969, June 2008. 22

[12] Dylan Aïssi, Jessica Dennis, Martin Ladouceur, Vinh Truong, Nora Zwingerman, Ares Rocanin-Arjo, Marine Germain, Tara A. Paton, Pierre-Emmanuel Morange, France Gagnon, and David-Alexandre Trégouët. Genome-Wide Investigation of DNA Methylation Marks Associated with FV Leiden Mutation. *PLoS ONE*, 9(9), September 2014. 112

[13] Ana B Arroyo, Salam Salloum-Asfar, Carlos Pérez-Sánchez, Raúl Teruel-Montoya, Silvia Navarro, Nuria García-Barberá, Ginés Luengo-Gil, Vanessa Roldán, John-Bjarne Hansen, Chary López-Pedrera, Vicente Vicente, Rocío González-Conejero, and Constantino Martínez. Regulation of TFPI expression by miR-27a/b-3p in human endothelial cells under normal conditions and in response to androgens. *Scientific Reports*, 7:43500, 2017. 113

[14] Joshua E. Babiarz, J. Graham Ruby, Yangming Wang, David P. Bartel, and Robert Blelloch. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes & Development*, 22(20):2773–2785, October 2008. 12, 29

[15] Christina Backes, Tobias Fehlmann, Fabian Kern, Tim Kehl, Hans-Peter Lenhof, Eckart Meese, and Andreas Keller. miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Research*, 46(D1):D160–D167, January 2018. 51

[16] Daehyun Baek, Judit Villén, Chanseok Shin, Fernando D. Camargo, Steven P. Gygi, and David P. Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, September 2008. 34

[17] Sophie Bail, Mavis Swerdel, Hudan Liu, Xinfu Jiao, Loyal A. Goff, Ronald P. Hart, and Megerditch Kiledjian. Differential regulation of microRNA stability. *RNA (New York, N.Y.)*, 16(5):1032–1039, May 2010. 19

[18] Monica Ballarino, Francesca Pagano, Erika Girardi, Mariangela Morlando, Davide Cacchiarelli, Marcella Marchioni, Nicholas J. Proudfoot, and Irene Bozzoni. Coupled RNA Processing and Transcription of Intergenic Primary MicroRNAs. *Molecular and Cellular Biology*, 29(20):5632–5638, October 2009. 7

[19] A. K. Banerjee. 5'-terminal cap structure in eucaryotic messenger ribonucleic acids. *Microbiological Reviews*, 44(2):175–205, June 1980. 129

[20] Alexander S. Baras, Christopher J. Mitchell, Jason R. Myers, Simone Gupta, Lien-Chun Weng, John M. Ashton, Toby C. Cornish, Akhilesh Pandey, and Marc K. Halushka. miRge - A Multiplexed Method of Processing Small RNA-Seq Data to Determine MicroRNA Entropy. *PloS One*, 10(11):e0143066, 2015. 55

[21] David P. Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, January 2009. 17

[22] David P. Bartel. Metazoan MicroRNAs. *Cell*, 173(1):20–51, March 2018. 17, 19, 20, 26, 33, 34, 35, 187

[23] Guillermo Barturen, Antonio Rueda, Maarten Hamberg, Angel Alganza, Ricardo Lebron, Michalis Kotsyfakis, Bu-Jun Shi, Danijela Koppers-Lalic, and Michael Hackenberg. sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. *Methods in Next Generation Sequencing*, 1(1), January 2014. 55

[24] Winston R. Becker, Benjamin Ober-Reynolds, Karina Jouravleva, Samson M. Jolly, Phillip D. Zamore, and William J. Greenleaf. High-Throughput Analysis Reveals Rules for Target RNA Binding and Cleavage by AGO2. *Molecular Cell*, July 2019. 21, 22

[25] Eugene Berezikov. Evolution of microRNA diversity and regulation in animals. *Nature Reviews. Genetics*, 12(12):846–860, November 2011. 15

[26] Arnold J. Berk. Discovery of RNA splicing and genes in pieces. *Proceedings of the National Academy of Sciences of the United States of America*, 113(4):801–805, January 2016. 129

[27] Grant Bertolet, Natee Kongchan, Rebekah Miller, Ravi K. Patel, Antrix Jain, Jong Min Choi, Alexander B. Saltzman, Amber Christenson, Sung Yun Jung, Anna Malovannaya, Andrew Grimson, and Joel R. Neilson. MiR-146a wild-type 3' sequence identity is dispensable for proper innate immune function in vivo. *Life Science Alliance*, 2(1):e201800249, February 2019. 17

[28] Xavier Bofill-De Ros, Wojciech K. Kasprzak, Yuba Bhandari, Lixin Fan, Quinn Cavanaugh, Minjie Jiang, Lisheng Dai, Acong Yang, Tie-Juan Shao, Bruce A. Shapiro, Yun-Xing Wang, and Shuo Gu. Structural Differences between Pri-miRNA Paralogs Promote Alternative Drosha Cleavage and Expand Target Repertoires. *Cell Reports*, 26(2):447–459.e4, January 2019. 28

[29] Markus T. Bohnsack, Kevin Czaplinski, and Dirk Gorlich. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA (New York, N.Y.)*, 10(2):185–191, February 2004. 9

[30] Christelle Borel, Samuel Deutsch, Audrey Letourneau, Eugenia Migliavacca, Stephen B. Montgomery, Antigone S. Dimas, Charles E. Vejnar, Homa Attar, Maryline Gagnebin, Corinne Gehrig, Emilie Falconnet, Yann Dupré, Emmanouil T. Dermitzakis, and Stylianos E. Antonarakis. Identification of cis- and trans-regulatory variation modulating microRNA expression levels in human fibroblasts. *Genome Research*, 21(1):68–73, January 2011. 32

[31] Diane Bortolamiol-Becet, Fuqu Hu, David Jee, Jiayu Wen, Katsutomo Okamura, Ching-Jung Lin, Stefan L. Ameres, and Eric C. Lai. Selective Suppression of the Splicing-Mediated MicroRNA Pathway by the Terminal Uridyltransferase Tailor. *Molecular Cell*, 59(2):217–228, July 2015. 13

[32] Juliane Braun, Danny Misiak, Bianca Busch, Knut Krohn, and Stefan Hüttelmaier. Rapid identification of regulatory microRNAs by miTRAP (miRNA trapping by RNA in vitro affinity purification). *Nucleic Acids Research*, 42(8):e66, April 2014. 88

[33] S. Brenner, F. Jacob, and M. Meselson. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature*, 190(4776):576, May 1961. 125

[34] James P. Broughton, Michael T. Lovci, Jessica L. Huang, Gene W. Yeo, and Amy E. Pasquinelli. Pairing beyond the Seed Supports MicroRNA Targeting Specificity. *Molecular Cell*, 64(2):320–333, 2016. 17, 21

[35] Miguel Brown, Hemant Suryawanshi, Markus Hafner, Thalia A. Farazi, and Thomas Tuschl. Mammalian miRNA curation through next-generation sequencing. *Frontiers in Genetics*, 4:145, 2013. 15

[36] Syed I. A. Bukhari, Samuel S. Truesdell, Sooncheol Lee, Swapna Kollu, Anthony Classon, Myriam Boukhali, Esha Jain, Richard D. Mortensen, Akiko Yanagiya, Ruslan I. Sadreyev, Wilhelm Haas, and Shobha Vasudevan. A Specialized Mechanism of Translation Mediated by FXR1a-Associated MicroRNP in Cellular Quiescence. *Molecular Cell*, 61(5):760–773, March 2016. 23

[37] Annalisa Buniello, Jacqueline A. L. MacArthur, Maria Cerezo, Laura W. Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L. Whetzel, Ridwan Amode, Jose A. Guillen, Harpreet S. Riat, Stephen J. Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A. Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, January 2019. 112

[38] David Burns, Andrea D'Ambrogio, Stephanie Nottrott, and Joel D. Richter. CPEB and two poly(A) polymerases control miR-122 stability and p53 mRNA translation. *Nature*, 473(7345):105–108, May 2011. 28

[39] Alexander Maxwell Burroughs, Yoshinari Ando, Michiel Jan Laurens de Hoon, Yasuhiro Tomaru, Harukazu Suzuki, Yoshihide Hayashizaki, and Carsten Olivier Daub. Deep-sequencing of human argonaute-associated small RNAs provides insight into miRNA sorting and reveals argonaute association with RNA fragments of diverse origin. *RNA Biology*, 8(1):158–177, 2011. 14

[40] Julien Béthune, Caroline G. Artus-Revel, and Witold Filipowicz. Kinetic analysis reveals successive steps leading to miRNA-mediated silencing in mammalian cells. *EMBO reports*, 13(8):716–723, August 2012. 20, 21

[41] George Adrian Calin, Calin Dan Dumitru, Masayoshi Shimizu, Roberta Bichi, Simona Zupo, Evan Noch, Hansjuerg Aldler, Sashi Rattan, Michael Keating, Kanti Rai, Laura Rassenti, Thomas Kipps, Massimo Negrini, Florencia Bullrich, and Carlo M. Croce. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, 99(24):15524–15529, November 2002. 35

[42] Sophia Cammaerts, Mojca Strazisar, Peter De Rijk, and Jurgen Del Favero. Genetic variants in microRNA genes: impact on microRNA expression, function, and disease. *Frontiers in Genetics*, 6, May 2015. 29

[43] Demián Cazalla, Therese Yario, Joan A. Steitz, and Joan Steitz. Down-regulation of a host microRNA by a Herpesvirus saimiri noncoding RNA. *Science (New York, N.Y.)*, 328(5985):1563–1566, June 2010. 34

[44] Li-Ling Chak and Katsutomo Okamura. Argonaute-dependent small RNAs derived from single-stranded, non-structured precursors. *Frontiers in Genetics*, 5, 2014. 14

[45] Stanley D. Chandradoss, Nicole T. Schirle, Malwina Szczepaniak, Ian J. MacRae, and Chirlmin Joo. A Dynamic Search Process Underlies MicroRNA Targeting. *Cell*, 162(1):96–107, July 2015. 18

[46] Sihem Cheloufi, Camila O. Dos Santos, Mark M. W. Chong, and Gregory J. Hannon. A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*, 465(7298):584–589, June 2010. 13, 21

[47] Chyi-Ying A. Chen and Ann-Bin Shyu. Mechanisms of deadenylation-dependent decay. *Wiley Interdisciplinary Reviews: RNA*, 2(2):167–183, 2011. 19

[48] Xi Chen, Hongwei Liang, Danping Guan, Cheng Wang, Xiaoyun Hu, Lin Cui, Sidi Chen, Chunni Zhang, Junfeng Zhang, Ke Zen, and Chen-Yu Zhang. A combination of Let-7d, Let-7g and Let-7i serves as a stable reference for normalization of serum microRNAs. *PloS One*, 8(11):e79652, 2013. 92

[49] Sung Wook Chi, Julie B. Zang, Aldo Mele, and Robert B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, July 2009. 21

[50] H. Rosaria Chiang, Lori W. Schoenfeld, J. Graham Ruby, Vincent C. Auyeung, Noah Spies, Daehyun Baek, Wendy K. Johnston, Carsten Russ, Shujun Luo, Joshua E. Babiarz, Robert Blelloch, Gary P. Schroth, Chad Nusbaum, and David P. Bartel. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes & Development*, 24(10):992–1009, May 2010. 15, 31

[51] Stephen S. C. Chim, Tristan K. F. Shing, Emily C. W. Hung, Tak-yeung Leung, Tze-kin Lau, Rossa W. K. Chiu, and Y. M. Dennis Lo. Detection and Characterization of Placental MicroRNAs in Maternal Plasma. *Clinical Chemistry*, 54(3):482–490, March 2008. 32

[52] Chia-ying Chu and Tariq M. Rana. Translation Repression in Human Cells by MicroRNA-Induced Gene Silencing Requires RCK/p54. *PLOS Biology*, 4(7):e210, 2006. 20

[53] Mete Civelek, Raffi Hagopian, Calvin Pan, Nam Che, Wen-pin Yang, Paul S. Kayne, Niyas K. Saleem, Henna Cederberg, Johanna Kuusisto, Peter S. Gargalovic, Todd G. Kirchgessner, Markku Laakso, and Aldons J. Lusis. Genetic regulation of human adipose microRNA expression and its consequences for metabolic traits. *Human Molecular Genetics*, 22(15):3023–3037, August 2013. 32, 111

[54] Diana F. Colgan and James L. Manley. Mechanism and regulation of mRNA polyadenylation. *Genes & Development*, 11(21):2755–2766, November 1997. 129

[55] F. H. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192:1227–1232, December 1961. 125

[56] Bruno Henrique Ribeiro Da Fonseca, Douglas Silva Domingues, and Alexandre Rossi Paschoal. mirtronDB: a mirtron knowledge base. *Bioinformatics*, 35(19):3873–3874, October 2019. 12

[57] Enyu Dai, Xuexin Yu, Yan Zhang, Fanlin Meng, Shuyuan Wang, Xinyi Liu, Dianming Liu, Jing Wang, Xia Li, and Wei Jiang. EpimiR: a database of curated mutual regulation

between miRNAs and epigenetic modifications. *Database: The Journal of Biological Databases and Curation*, 2014, March 2014. 32

[58] Nabanita De, Lisa Young, Pick-Wei Lau, Nicole-Claudia Meisner, David V. Morrissey, and Ian J. MacRae. Highly Complementary Target RNAs Promote Release of Guide RNAs from Human Argonaute2. *Molecular Cell*, 50(3):344–355, May 2013. 22, 26

[59] Manuel de la Mata, Dimos Gaidatzis, Mirela Vitanescu, Michael B. Stadler, Corinna Wentzel, Peter Scheiffele, Witold Filipowicz, and Helge Großhans. Potent degradation of neuronal miRNAs induced by highly complementary targets. *EMBO reports*, 16(4):500–511, April 2015. 25

[60] Loïc de Pontual, Evelyn Yao, Patrick Callier, Laurence Faivre, Valérie Drouin, Sandra Cariou, Arie Van Haeringen, David Geneviève, Alice Goldenberg, Myriam Oufadem, Sylvie Manouvrier, Arnold Munnich, Joana Alves Vidigal, Michel Vekemans, Stanislas Lyonnet, Alexandra Henrion-Caude, Andrea Ventura, and Jeanne Amiel. Germline deletion of the miR-17 92 cluster causes skeletal and growth defects in humans. *Nature Genetics*, 43(10):1026–1030, September 2011. 35

[61] Rémy Denzler, Vikram Agarwal, Joanna Stefano, David P. Bartel, and Markus Stoffel. Assessing the ceRNA Hypothesis with Quantitative Measurements of miRNA and Target Abundance. *Molecular Cell*, 54(5):766–776, June 2014. 27

[62] Rémy Denzler, Sean E. McGeary, Alexandra C. Title, Vikram Agarwal, David P. Bartel, and Markus Stoffel. Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression. *Molecular Cell*, 64(3):565–579, November 2016. 25

[63] Vittoria Di Mauro, Silvia Crasto, Federico Simone Colombo, Elisa Di Pasquale, and Daniele Catalucci. Wnt signalling mediates miR-133a nuclear re-localization for the transcriptional control of Dnmt3b in cardiac cells. *Scientific Reports*, 9, June 2019. 24

[64] Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Nadav S. Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J. Luo, Eddie Park, Kimberly Persaud, Jonathan B. Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, and Thomas R. Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, September 2012. 130

[65] Yair Dorsett, Kevin M. McBride, Mila Jankovic, Anna Gazumyan, To-Ha Thai, Davide F. Robbiani, Michela Di Virgilio, Bernardo Reina San-Martin, Gordon Heidkamp, Tanja A.

Schwickert, Thomas Eisenreich, Klaus Rajewsky, and Michel C. Nussenzweig. MicroRNA-155 suppresses activation-induced cytidine deaminase-mediated Myc-Igh translocation. *Immunity*, 28(5):630–638, May 2008. 36

[66] Anne Dueck, Christian Ziegler, Alexander Eichner, Eugene Berezikov, and Gunter Meister. microRNAs associated with the different human Argonaute proteins. *Nucleic Acids Research*, 40(19):9850–9862, October 2012. 22

[67] Matyas Ecsedi, Magdalene Rausch, and Helge Großhans. The let-7 microRNA directs vulval development through a single target. *Developmental Cell*, 32(3):335–344, February 2015. 36

[68] Stephen W. Eichhorn, Huili Guo, Sean E. McGeary, Ricard A. Rodriguez-Mias, Chanseok Shin, Daehyun Baek, Shu-hao Hsu, Kalpana Ghoshal, Judit Villén, and David P. Bartel. mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. *Molecular Cell*, 56(1):104–115, October 2014. 20

[69] Reyad A. Elbarbary, Keita Miyoshi, Jason R. Myers, Peicheng Du, John M. Ashton, Bin Tian, and Lynne E. Maquat. Tudor-SN–mediated endonucleolytic decay of human cell microRNAs promotes G1/S phase transition. *Science*, 356(6340):859–862, May 2017. 25

[70] Amro Elgheznawy and Ingrid Fleming. Platelet-Enriched MicroRNAs and Cardiovascular Homeostasis. *Antioxidants & Redox Signaling*, 29(9):902–921, August 2017. 88

[71] Wenwen Fang and David P. Bartel. The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes. *Molecular Cell*, 60(1):131–145, October 2015. 8

[72] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811, February 1998. 6

[73] Mathieu N. Flamand, Hin Hark Gan, Vinay K. Mayya, Kristin C. Gunsalus, and Thomas F. Duchaine. A non-canonical site reveals the cooperative mechanisms of microRNA-mediated silencing. *Nucleic Acids Research*, 45(12):7212–7225, July 2017. 22, 34

[74] Linda E. Flinterman, Astrid van Hylckama Vlieg, Suzanne C. Cannegieter, and Frits R. Rosendaal. Long-term survival in a large cohort of patients with venous thrombosis: incidence and predictors. *PLoS medicine*, 9(1):e1001155, January 2012. 85

[75] Massimo Franchini, Franco Capra, Giovanni Targher, Martina Montagnana, and Giuseppe Lippi. Relationship between ABO blood group and von Willebrand factor levels: from biology to clinical implications. *Thrombosis Journal*, 5:14, September 2007. 86

[76] Filipp Frank, Nahum Sonenberg, and Bhushan Nagar. Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature*, 465(7299):818–822, June 2010. 11

[77] Rosalind E. Franklin and R. G. Gosling. Molecular Configuration in Sodium Thymonucleate. *Nature*, 171(4356):740, April 1953. 125

[78] M. R. Friedlander, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52, January 2012. 55

[79] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, January 2009. 17, 19, 26

[80] Bastian Fromm, Tyler Billipp, Liam E. Peck, Morten Johansen, James E. Tarver, Benjamin L. King, James M. Newcomb, Lorenzo F. Sempere, Kjersti Flatmark, Eivind Hovig, and Kevin J. Peterson. A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annual Review of Genetics*, 49(1):213–242, 2015. 15

[81] Akira Fukao, Yuichiro Mishima, Naoki Takizawa, Shigenori Oka, Hiroaki Imataka, Jerry Pelletier, Nahum Sonenberg, Christian Thoma, and Toshinobu Fujiwara. MicroRNAs trigger dissociation of eIF4ai and eIF4aii from target mRNAs in humans. *Molecular Cell*, 56(1):79–89, October 2014. 20

[82] Takashi Fukaya, Hiro-Oki Iwakawa, and Yukihide Tomari. MicroRNAs block assembly of eIF4f translation initiation complex in Drosophila. *Molecular Cell*, 56(1):67–78, October 2014. 20

[83] Keith T. Gagnon, Liande Li, Yongjun Chu, Bethany A. Janowski, and David R. Corey. RNAi factors are present and active in human cell nuclei. *Cell Reports*, 6(1):211–221, January 2014. 23

[84] Eric R. Gamazon, Dana Ziliak, Hae Kyung Im, Bonnie LaCroix, Danny S. Park, Nancy J. Cox, and R. Stephanie Huang. Genetic architecture of microRNA expression: implications for the transcriptome and complex traits. *American Journal of Human Genetics*, 90(6):1046–1063, June 2012. 32

[85] Luca F. R. Gebert and Ian J. MacRae. Regulation of microRNA function in animals. *Nature Reviews. Molecular Cell Biology*, 20(1):21–37, January 2019. 18, 22, 28, 187, 188

[86] Marine Germain, Daniel I. Chasman, Hugoline de Haan, Weihong Tang, Sara Lindström, Lu-Chen Weng, Mariza de Andrade, Marieke C. H. de Visser, Kerri L. Wiggins, Pierre Suchon, Noémie Saut, David M. Smadja, Grégoire Le Gal, Astrid van Hylckama Vlieg, Antonio Di Narzo, Ke Hao, Christopher P. Nelson, Ares Rocanin-Arjo, Lasse Folkersen, Ramin Monajemi, Lynda M. Rose, Jennifer A. Brody, Eline Slagboom, Dylan Aïssi, France Gagnon, Jean-Francois Deleuze, Panos Deloukas, Christophe Tzourio, Jean-Francois Dartigues, Claudine Berr, Kent D. Taylor, Mete Civelek, Per Eriksson, Cardiogenics Consortium, Bruce M. Psaty, Jeanine Houwing-Duitermaat, Alison H. Goodall, François Cambien, Peter Kraft, Philippe Amouyel, Nilesh J. Samani, Saonli Basu, Paul M. Ridker, Frits R. Rosendaal, Christopher Kabrhel, Aaron R. Folsom, John Heit, Pieter H. Reitsma, David-Alexandre Trégouët, Nicholas L. Smith, and Pierre-Emmanuel Morange. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44a2 as two susceptibility loci for venous thromboembolism. *American Journal of Human Genetics*, 96(4):532–542, April 2015. 86

[87] Francesco Ghini, Carmela Rubolino, Montserrat Climent, Ines Simeone, Matteo J. Marzi, and Francesco Nicassio. Endogenous transcripts control miRNA levels and activity in mammalian cells by target-directed miRNA degradation. *Nature Communications*, 9(1):3119, 2018. 25

[88] Ryan J. Golden, Beibei Chen, Tuo Li, Juliane Braun, Hema Manjunath, Xiang Chen, Jiaxi Wu, Vanessa Schmid, Tsung-Cheng Chang, Florian Kopp, Andres Ramirez-Martinez, Vincent S. Tagliabracci, Zhijian J. Chen, Yang Xie, and Joshua T. Mendell. An Argonaute phosphorylation cycle promotes microRNA-mediated silencing. *Nature*, 542(7640):197–202, 2017. 22

[89] Richard I. Gregory, Kai-Ping Yan, Govindasamy Amuthan, Thimmaiah Chendrimada, Behzad Doratotaj, Neil Cooch, and Ramin Shiekhattar. The Microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014):235–240, November 2004. 7

[90] Sam Griffiths-Jones. The microRNA Registry. *Nucleic Acids Research*, 32(Database issue):D109–D111, January 2004. 4, 15

[91] Andrew Grimson, Kyle Kai-How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, 27(1):91–105, July 2007. 17

[92] A. Grishok, A. E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D. L. Baillie, A. Fire, G. Ruvkun, and C. C. Mello. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106(1):23–34, July 2001. 9

[93] Francois Gros, H. Hiatt, Walter Gilbert, C. G. Kurland, R. W. Risebrough, and J. D. Watson. Unstable Ribonucleic Acid Revealed by Pulse Labelling of Escherichia Coli. *Nature*, 190(4776):581, May 1961. 125

[94] Stefanie Grosswendt, Andrei Filipchyk, Mark Manzano, Filippos Klironomos, Marcel Schilling, Margareta Herzog, Eva Gottwein, and Nikolaus Rajewsky. Unambiguous Identification of miRNA:Target Site Interactions by Different Types of Ligation Reactions. *Molecular Cell*, 54(6):1042–1054, June 2014. 21

[95] Sonia Guil and Javier F. Cáceres. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nature Structural & Molecular Biology*, 14(7):591–596, July 2007. 8

[96] Minju Ha and V. Narry Kim. Regulation of microRNA biogenesis. *Nature Reviews. Molecular Cell Biology*, 15(8):509–524, August 2014. 14, 187

[97] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141, April 2010. 21

[98] Marc K. Halushka, Bastian Fromm, Kevin J. Peterson, and Matthew N. McCall. Big Strides in Cellular microRNA Expression. *Trends in genetics : TIG*, 34(3):165–167, March 2018. 33

[99] Perry V. Halushka, Andrew J. Goodwin, and Marc K. Halushka. Opportunities for microRNAs in the Crowded Field of Cardiovascular Biomarkers. *Annual Review of Pathology*, 14:211–238, January 2019. 37

[100] A. J. Hamilton and D. C. Baulcombe. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science (New York, N.Y.)*, 286(5441):950–952, October 1999. 21

[101] Jiang Han, Daniel Kim, and Kevin V. Morris. Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 104(30):12422–12427, July 2007. 24

[102] Miao Han and Yun Zheng. Comprehensive analysis of single nucleotide polymorphisms in human microRNAs. *PloS One*, 8(11):e78028, 2013. 30

[103] Thomas B. Hansen, Trine I. Jensen, Bettina H. Clausen, Jesper B. Bramsen, Bente Finsen, Christian K. Damgaard, and Jørgen Kjems. Natural RNA circles function as efficient microRNA sponges. *Nature*, 495(7441):384–388, March 2013. 26

[104] Thomas B. Hansen, Jørgen Kjems, and Jesper B. Bramsen. Enhancing miRNA annotation confidence in miRBase by continuous cross dataset analysis. *RNA Biology*, 8(3):378–383, May 2011. 15

[105] Thomas B. Hansen, Morten T. Venø, Trine I. Jensen, Anne Schaefer, Christian K. Damgaard, and Jørgen Kjems. Argonaute-associated short introns are a novel class of gene regulators. *Nature Communications*, 7:11538, 2016. 14

[106] J. A. Heit, S. M. Armasu, Y. W. Asmann, J. M. Cunningham, M. E. Matsumoto, T. M. Petterson, and M. De Andrade. A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q. *Journal of thrombosis and haemostasis: JTH*, 10(8):1521–1531, August 2012. 111

[107] John A. Heit, Frederick A. Spencer, and Richard H. White. The epidemiology of venous thromboembolism. *Journal of Thrombosis and Thrombolysis*, 41:3–14, 2016. 85

[108] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153(3):654–665, April 2013. 21

[109] Inha Heo, Minju Ha, Jaechul Lim, Mi-Jeong Yoon, Jong-Eun Park, S. Chul Kwon, Hyeshik Chang, and V. Narry Kim. Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. *Cell*, 151(3):521–532, October 2012. 13

[110] Inha Heo, Chirlmin Joo, Jun Cho, Minju Ha, Jinju Han, and V. Narry Kim. Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA. *Molecular Cell*, 32(2):276–284, October 2008. 13

[111] Jana Hertel, Manuela Lindemeyer, Kristin Missal, Claudia Fried, Andrea Tanzer, Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, and Students of Bioinformatics Computer Labs 2004 and 2005. The expansion of the metazoan microRNA repertoire. *BMC genomics*, 7:25, February 2006. 15

[112] Hristo B. Houbaviy, Michael F. Murray, and Phillip A. Sharp. Embryonic stem cell-specific MicroRNAs. *Developmental Cell*, 5(2):351–358, August 2003. 4

[113] Tianxiao Huan, George Chen, Chunyu Liu, Anindya Bhattacharya, Jian Rong, Brian H. Chen, Sudha Seshadri, Kahraman Tanriverdi, Jane E. Freedman, Martin G. Larson, Joanne M. Murabito, and Daniel Levy. Age-associated microRNA expression in human peripheral blood is associated with all-cause mortality and age-related traits. *Aging Cell*, 17(1), February 2018. 31

[114] Tianxiao Huan, Michael Mendelson, Roby Joehanes, Chen Yao, Chunyu Liu, Ci Song, Anindya Bhattacharya, Jian Rong, Kahraman Tanriverdi, Joshua Keefe, Joanne M. Murabito, Paul Courchesne, Martin G. Larson, Jane E. Freedman, and Daniel Levy. Epigenome-wide association study of DNA methylation and microRNA expression highlights novel pathways for human complex traits. *Epigenetics*, 0(0):1–16, July 2019. 32

[115] Tianxiao Huan, Jian Rong, Chunyu Liu, Xiaoling Zhang, Kahraman Tanriverdi, Roby Joehanes, Brian H. Chen, Joanne M. Murabito, Chen Yao, Paul Courchesne, Peter J. Munson, Christopher J. O'Donnell, Nancy Cox, Andrew D. Johnson, Martin G. Larson,

Daniel Levy, and Jane E. Freedman. Genome-wide identification of microRNA expression quantitative trait loci. *Nature Communications*, 6:6601, March 2015. 32, 111

[116] R. Stephanie Huang, Eric R. Gamazon, Dana Ziliak, Yujia Wen, Hae Kyung Im, Wei Zhang, Claudia Wing, Shiwei Duan, Wasim K. Bleibel, Nancy J. Cox, and M. Eileen Dolan. Population differences in microRNA expression and biological implications. *RNA biology*, 8(4):692–701, August 2011. 32

[117] Zhou Huang, Jiangcheng Shi, Yuanxu Gao, Chunmei Cui, Shan Zhang, Jianwei Li, Yuan Zhou, and Qinghua Cui. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Research*, 47(D1):D1013–D1017, January 2019. 35

[118] Anne E. Hughes, Declan T. Bradley, Malcolm Campbell, Judith Lechner, Durga P. Dash, David A. Simpson, and Colin E. Willoughby. Mutation Altering the miR-184 Seed Region Causes Familial Keratoconus with Cataract. *American Journal of Human Genetics*, 89(5):628–633, November 2011. 30

[119] M. V. Huisman and F. A. Klok. Diagnostic management of acute deep vein thrombosis and pulmonary embolism. *Journal of thrombosis and haemostasis: JTH*, 11(3):412–422, March 2013. 86

[120] Menno V. Huisman, Stefano Barco, Suzanne C. Cannegieter, Gregoire Le Gal, Stavros V. Konstantinides, Pieter H. Reitsma, Marc Rodger, Anton Vonk Noordegraaf, and Frederikus A. Klok. Pulmonary embolism. *Nature Reviews Disease Primers*, 4:18028, May 2018. 86, 188

[121] Jerard Hurwitz. The Discovery of RNA Polymerase. *Journal of Biological Chemistry*, 280(52):42477–42485, December 2005. 126

[122] G. Hutvágner, J. McLachlan, A. E. Pasquinelli, E. Bálint, T. Tuschl, and P. D. Zamore. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science (New York, N.Y.)*, 293(5531):834–838, August 2001. 9

[123] Hun-Way Hwang, Erik A. Wentzel, and Joshua T. Mendell. A hexanucleotide element directs microRNA nuclear import. *Science (New York, N.Y.)*, 315(5808):97–100, January 2007. 23

[124] Shintaro Iwasaki, Maki Kobayashi, Mayuko Yoda, Yuriko Sakaguchi, Susumu Katsuma, Tsutomu Suzuki, and Yukihide Tomari. Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. *Molecular Cell*, 39(2):292–299, July 2010. 9

[125] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356, June 1961. 125

[126] Anitha D. Jayaprakash, Omar Jabado, Brian D. Brown, and Ravi Sachidanandam. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Research*, 39(21):e141, November 2011. 46

[127] Stefanie Jonas and Elisa Izaurralde. Towards a molecular understanding of microRNA-mediated gene silencing. *Nature Reviews Genetics*, 16(7):421–433, July 2015. 19, 20

[128] Susan R. Kahn, Anthony J. Comerota, Mary Cushman, Natalie S. Evans, Jeffrey S. Ginsberg, Neil A. Goldenberg, Deepak K. Gupta, Paolo Prandoni, Suresh Vedantham, M. Eileen Walsh, Jeffrey I. Weitz, and American Heart Association Council on Peripheral Vascular Disease, Council on Clinical Cardiology, and Council on Cardiovascular and

Stroke Nursing. The postthrombotic syndrome: evidence-based prevention, diagnosis, and treatment strategies: a scientific statement from the American Heart Association. *Circulation*, 130(18):1636–1661, October 2014. 85

[129] Anastasiia Kamenska, Clare Simpson, Caroline Vindry, Helen Broomhead, Marianne Bénard, Michèle Ernoult-Lange, Benjamin P. Lee, Lorna W. Harries, Dominique Weil, and Nancy Standart. The DDX6–4e-T interaction mediates translational repression and P-body assembly. *Nucleic Acids Research*, 44(13):6318–6334, July 2016. 19

[130] Naoyuki Kataoka, Megumi Fujita, and Mutsuhito Ohno. Functional association of the Microprocessor complex with the spliceosome. *Molecular and Cellular Biology*, 29(12):3243–3254, June 2009. 9

[131] Takayuki Katoh, Yuriko Sakaguchi, Kenjyo Miyauchi, Takeo Suzuki, Shin-Ichi Kashiwabara, Tadashi Baba, and Tsutomu Suzuki. Selective stabilization of mammalian microRNAs by 3' adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. *Genes & Development*, 23(4):433–438, February 2009. 28

[132] Dorothee Kaudewitz, Philipp Skroblin, Lukas H. Bender, Temo Barwari, Peter Willeit, Raimund Pechlaner, Nicholas P. Sunderland, Karin Willeit, Allison C. Morton, Paul C. Armstrong, Melissa V. Chan, Ruifang Lu, Xiaoke Yin, Filipe Gracio, Katarzyna Dudek, Sarah R. Langley, Anna Zampetaki, Emanuele de Rinaldis, Shu Ye, Timothy D. Warner, Alka Saxena, Stefan Kiechl, Robert F. Storey, and Manuel Mayr. Association of MicroRNAs and YRNAs with Platelet Function. *Circulation research*, 118(3):420–432, February 2016. 88, 114

[133] Yukio Kawahara, Molly Megraw, Edward Kreider, Hisashi Iizasa, Louis Valente, Artemis G. Hatzigeorgiou, and Kazuko Nishikura. Frequency and fate of microRNA editing in human brain. *Nucleic Acids Research*, 36(16):5270–5280, September 2008. 29

[134] Tomoko Kawamata and Yukihide Tomari. Making RISC. *Trends in Biochemical Sciences*, 35(7):368–376, July 2010. 10

[135] H. G. Khorana, H. Büuchi, H. Ghosh, N. Gupta, T. M. Jacob, H. Kössel, R. Morgan, S. A. Narang, E. Ohtsuka, and R. D. Wells. Polynucleotide Synthesis and the Genetic Code. *Cold Spring Harbor Symposia on Quantitative Biology*, 31:39–49, January 1966. 125

[136] Anastasia Khvorova, Angela Reynolds, and Sumedha D. Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–216, October 2003. 11

[137] Baekgyu Kim, Kyowon Jeong, and V. Narry Kim. Genome-wide Mapping of DROSHA Cleavage Sites on Primary MicroRNAs and Noncanonical Substrates. *Molecular Cell*, 66(2):258–269.e5, April 2017. 27

[138] Boseon Kim, Minju Ha, Luuk Loeff, Hyeshik Chang, Dhirendra K. Simanshu, Sisi Li, Mohamed Fareh, Dinshaw J. Patel, Chirlmin Joo, and V. Narry Kim. TUT7 controls the fate of precursor microRNAs by using three different uridylation mechanisms. *The EMBO Journal*, 34(13):1801–1815, July 2015. 13

[139] Daniel H. Kim, Pål Saetrom, Ola Snøve, and John J. Rossi. MicroRNA-directed transcriptional gene silencing in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, 105(42):16230–16235, October 2008. 24

[140] Daniel H. Kim, Louisa M. Villeneuve, Kevin V. Morris, and John J. Rossi. Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nature Structural & Molecular Biology*, 13(9):793–797, September 2006. 24

[141] Doyeon Kim, You Me Sung, Jinman Park, Sukjun Kim, Jongkyu Kim, Junhee Park, Haeok Ha, Jung Yoon Bae, SoHui Kim, and Daehyun Baek. General rules for functional microRNA targeting. *Nature Genetics*, 48(12):1517–1526, December 2016. 21

[142] Haedong Kim, Jimi Kim, Kijun Kim, Hyeshik Chang, Kwontae You, and V. Narry Kim. Bias-minimized quantification of microRNA reveals widespread alternative processing and 3' end modification. *Nucleic Acids Research*, 47(5):2630–2640, March 2019. 46

[143] John Kim, Anna Krichevsky, Yonatan Grad, Gabriel D. Hayes, Kenneth S. Kosik, George M. Church, and Gary Ruvkun. Identification of many microRNAs that copurify with polyribosomes in mammalian neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1):360–365, January 2004. 4

[144] Young-Kook Kim, Boseon Kim, and V. Narry Kim. Re-evaluation of the roles of DROSHA, Exportin 5, and DICER in microRNA biogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 113(13):E1881–E1889, March 2016. 9

[145] Young-Kook Kim and V. Narry Kim. Processing of intronic microRNAs. *The EMBO journal*, 26(3):775–783, February 2007. 9

[146] Elena Kingston and David Bartel. Global analyses of the dynamics of mammalian microRNA metabolism. *Genome Research*, September 2019. 24, 29, 33

[147] Derek Klarin, Connor A. Emdin, Pradeep Natarajan, Mark F. Conrad, INVENT Consortium, and Sekar Kathiresan. Genetic Analysis of Venous Thromboembolism in UK Biobank Identifies the ZFPM2 Locus and Implicates Obesity as a Causal Risk Factor. *Circulation. Cardiovascular Genetics*, 10(2), April 2017. 111

[148] Benjamin Kleaveland, Charlie Y. Shi, Joanna Stefano, and David P. Bartel. A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. *Cell*, 174(2):350–362.e17, 2018. 25

[149] Misha Klein, Stanley D. Chandradoss, Martin Depken, and Chirlmin Joo. Why Argonaute is needed to make microRNA target search fast and reliable. *Seminars in Cell & Developmental Biology*, 65:20–28, May 2017. 18

[150] Hotaka Kobayashi, Keisuke Shoji, Kaori Kiyokawa, Lumi Negishi, and Yukihide Tomari. VCP Machinery Mediates Autophagic Degradation of Empty Argonaute. *Cell Reports*, 28(5):1144–1153.e4, July 2019. 22

[151] Martha V. Koerner, Kashyap Chhatbar, Shaun Webb, Justyna Cholewa-Waclaw, Jim Selfridge, Dina De Sousa, Bill Skarnes, Barry Rosen, Mark Thomas, Joanna Bottomley, Ramiro Ramires-Solis, Christopher Lelliott, David J. Adams, and Adrian Bird. An Orphan CpG Island Drives Expression of a let-7 miRNA Precursor with an Important Role in Mouse Development. *Epigenomes*, 3(1), March 2019. 32

[152] A. A. Kondkar, M. S. Bray, S. M. Leal, S. Nagalla, D. J. Liu, Y. Jin, J. F. Dong, Q. Ren, S. W. Whiteheart, C. Shaw, and P. F. Bray. VAMP8/endobrevin is overexpressed in hyperreactive human platelets: suggested role for platelet microRNA. *Journal of Thrombosis and Haemostasis*, 8(2):369–378, 2010. 88

[153] Jacek Krol, Volker Busskamp, Ilona Markiewicz, Michael B. Stadler, Sebastian Ribi, Jens Richter, Jens Duebel, Silvia Bicker, Hans Jörg Fehling, Dirk Schübeler, Thomas G. Oertner, Gerhard Schratt, Miriam Bibel, Botond Roska, and Witold Filipowicz. Characterizing light-regulated retinal microRNAs reveals rapid turnover as a common property of neuronal microRNAs. *Cell*, 141(4):618–631, May 2010. 34

[154] Canan Kuscu, Pankaj Kumar, Manjari Kiran, Zhangli Su, Asrar Malik, and Anindya Dutta. tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA (New York, N.Y.)*, 24(8):1093–1105, 2018. 14

[155] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, 294(5543):853–858, October 2001. 4

[156] Mariana Lagos-Quintana, Reinhard Rauhut, Jutta Meyer, Arndt Borkhardt, and Thomas Tuschl. New microRNAs from mouse and human. *RNA (New York, N.Y.)*, 9(2):175–179, February 2003. 4

[157] Eric C. Lai. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature Genetics*, 30(4):363, April 2002. 16

[158] Pablo Landgraf, Mirabela Rusu, Robert Sheridan, Alain Sewer, Nicola Iovino, Alexei Aravin, Sébastien Pfeffer, Amanda Rice, Alice O. Kamphorst, Markus Landthaler, Carolina Lin, Nicholas D. Socci, Leandro Hermida, Valerio Fulci, Sabina Chiaretti, Robin Foà, Julia Schliwka, Uta Fuchs, Astrid Novosel, Roman-Ulrich Müller, Bernhard Schermer, Ute Bissels, Jason Inman, Quang Phan, Minchen Chien, David B. Weir, Ruchi Choksi, Gabriella De Vita, Daniela Frezzetti, Hans-Ingo Trompeter, Veit Hornung, Grace Teng, Gunther Hartmann, Miklos Palkovits, Roberto Di Lauro, Peter Wernet, Giuseppe Macino, Charles E. Rogler, James W. Nagle, Jingyue Ju, F. Nina Papavasiliou, Thomas Benzing, Peter Lichter, Wayne Tam, Michael J. Brownstein, Andreas Bosio, Arndt Borkhardt, James J. Russo, Chris Sander, Mihaela Zavolan, and Thomas Tuschl. A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell*, 129(7):1401–1414, June 2007. 29

[159] Patricia Landry, Isabelle Plante, Dominique L. Ouellet, Marjorie P. Perron, Guy Rousseau, and Patrick Provost. Existence of a microRNA pathway in anucleate platelets. *Nature Structural & Molecular Biology*, 16(9):961–966, September 2009. 87, 88

[160] Markus Landthaler, Abdullah Yalcin, and Thomas Tuschl. The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. *Current biology: CB*, 14(23):2162–2167, December 2004. 7

[161] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357, April 2012. 59

[162] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, March 2009. 52

[163] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A. C. t Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, The Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E. Antonarakis, Robert Hasler, Ann-Christine Syvanen, Gert-Jan van Ommen, Alvis Brazma,

Thomas Meitinger, Philip Rosenstiel, Roderic Guigo, Ivo G. Gut, Xavier Estivill, and Emmanouil T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, September 2013. 32

[164] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science (New York, N.Y.)*, 294(5543):858–862, October 2001. 4

[165] Martin Laurence, Christos Hatzis, and Douglas E. Brash. Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. *PLOS ONE*, 9(5):e97876, May 2014. 53

[166] Ho Young Lee, Kaihong Zhou, Alison Marie Smith, Cameron L. Noland, and Jennifer A. Doudna. Differential roles of human Dicer-binding proteins TRBP and PACT in small RNA processing. *Nucleic Acids Research*, 41(13):6568–6576, July 2013. 9

[167] R. C. Lee and V. Ambros. An extensive class of small RNAs in Caenorhabditis elegans. *Science (New York, N.Y.)*, 294(5543):862–864, October 2001. 4

[168] R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, December 1993. 3, 4, 187

[169] Sanghyun Lee, Jaewon Song, Sungchul Kim, Jongkyu Kim, Yujin Hong, Youngkyun Kim, Donghyun Kim, Daehyun Baek, and Kwangseog Ahn. Selective degradation of host MicroRNAs by an intergenic HCMV noncoding RNA accelerates virus production. *Cell Host & Microbe*, 13(6):678–690, June 2013. 34

[170] Yoontae Lee, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, Patrick Provost, Olof Rådmark, Sunyoung Kim, and V. Narry Kim. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956):415–419, September 2003. 7

[171] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal*, 23(20):4051–4060, October 2004. 6

[172] Benjamin P. Lewis, I.-hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, December 2003. 16, 17, 52, 187

[173] Chao Li, Hua-Yu Zhu, Wen-Dong Bai, Lin-Lin Su, Jia-Qi Liu, Wei-Xia Cai, Bin Zhao, Jian-Xin Gao, Shi-Chao Han, Jun Li, and Da-Hai Hu. MiR-10a and miR-181c regulate collagen type I generation in hypertrophic scars by targeting PAI-1 and uPA. *FEBS letters*, 589(3):380–389, January 2015. 113

[174] Long-Cheng Li. Chromatin remodeling by the small RNA machinery in mammalian cells. *Epigenetics*, 9(1):45–52, January 2014. 24

[175] Lee P. Lim, Margaret E. Glasner, Soraya Yekta, Christopher B. Burge, and David P. Bartel. Vertebrate MicroRNA Genes. *Science*, 299(5612):1540–1540, March 2003. 4

[176] Sara Lindstrom, Lu Wang, Erin N. Smith, William Gordon, Astrid van Hylckama Vlieg, Mariza de Andrade, Jennifer A. Brody, Jack W. Pattee, Jeffrey Haessler, Ben M. Brumpton, Daniel I. Chasman, Pierre Suchon, Ming-Huei Chen, Constance Turman, Marine Germain, Kerri L. Wiggins, James MacDonald, Sigrid K. Braekkan, Sebastian M. Armasu, Nathan Pankratz, Rebecca D. Jackson, Jonas B. Nielsen, Franco Giulianini, Marja K.

Puurunen, Manal Ibrahim, Susan R. Heckbert, Scott M. Damrauer, Pradeep Natarajan, Derek Klarin, Paul S. de Vries, Maria Sabater-Lleal, Jennifer E. Huffman, Theo K. Bammler, Kelly A. Frazer, Bryan M. McCauley, Kent Taylor, James S. Pankow, Alexander P. Reiner, Maiken E. Gabrielsen, Jean-François Deleuze, Chris J. O'Donnell, Jihye Kim, Barbara McKnight, Peter Kraft, John-Bjarne Hansen, Frits R. Rosendaal, John A. Heit, Bruce M. Psaty, Weihong Tang, Charles Kooperberg, Kristian Hveem, Paul M. Ridker, Pierre-Emmanuel Morange, Andrew D. Johnson, Christopher Kabrhel, David-Alexandre Trégouët, and Nicholas L. Smith. Genomic and Transcriptomic Association Studies Identify 16 Novel Susceptibility Loci for Venous Thromboembolism. *Blood*, August 2019. 86

[177] Hongyu Liu, Cheng Lei, Qin He, Zou Pan, Desheng Xiao, and Yongguang Tao. Nuclear functions of mammalian MicroRNAs in gene regulation, immunity and cancer. *Molecular Cancer*, 17(1):64, February 2018. 23, 24

[178] Jidong Liu, Michelle A. Carmell, Fabiola V. Rivas, Carolyn G. Marsden, J. Michael Thomson, Ji-Joon Song, Scott M. Hammond, Leemor Joshua-Tor, and Gregory J. Hannon. Argonaute2 is the catalytic engine of mammalian RNAi. *Science (New York, N.Y.)*, 305(5689):1437–1441, September 2004. 21

[179] Zhongmin Liu, Jia Wang, Hang Cheng, Xin Ke, Lei Sun, Qiangfeng Cliff Zhang, and Hong-Wei Wang. Cryo-EM Structure of Human Dicer and Its Complexes with a Pre-miRNA Substrate. *Cell*, 173(5):1191–1203.e12, 2018. 10, 187

[180] Gabriel B. Loeb, Aly A. Khan, David Canner, Joseph B. Hiatt, Jay Shendure, Robert B. Darnell, Christina S. Leslie, and Alexander Y. Rudensky. Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Molecular Cell*, 48(5):760–770, December 2012. 21

[181] Phillipe Loher, Eric R. Londin, and Isidore Rigoutsos. IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies. *Oncotarget*, 5(18):8790–8802, September 2014. 31

[182] Eric Londin, Phillipe Loher, Aristeidis G. Telonis, Kevin Quann, Peter Clark, Yi Jing, Eleftheria Hatzimichael, Yohei Kirino, Shozo Honda, Michelle Lally, Bharat Ramratnam, Clay E. S. Comstock, Karen E. Knudsen, Leonard Gomella, George L. Spaeth, Lisa Hark, L. Jay Katz, Agnieszka Witkiewicz, Abdolmohamad Rostami, Sergio A. Jimenez, Michael A. Hollingsworth, Jen Jen Yeh, Chad A. Shaw, Steven E. McKenzie, Paul Bray, Peter T. Nelson, Simona Zupo, Katrien Van Roosbroeck, Michael J. Keating, George A. Calin, Charles Yeo, Masaya Jimbo, Joseph Cozzitorto, Jonathan R. Brody, Kathleen Delgrosso, John S. Mattick, Paolo Fortina, and Isidore Rigoutsos. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 112(10):E1106–E1115, March 2015. 11

[183] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. 92

[184] Ian J. Macrae, Kaihong Zhou, Fei Li, Adrian Repic, Angela N. Brooks, W. Zacheus Cande, Paul D. Adams, and Jennifer A. Doudna. Structural basis for double-stranded RNA processing by Dicer. *Science (New York, N.Y.)*, 311(5758):195–198, January 2006. 9

[185] Shahana Majid, Altaf A. Dar, Sharanjot Saini, Soichiro Yamamura, Hiroshi Hirata, Yuichiro Tanaka, Guoren Deng, and Rajvir Dahiya. MicroRNA-205–directed transcriptional activation of tumor suppressor genes in prostate cancer. *Cancer*, 116(24):5637–5649, 2010. 24

[186] Lisa Marcinowski, Mélanie Tanguy, Astrid Krmpotic, Bernd Rädle, Vanda J. Lisnić, Lee Tuddenham, Béatrice Chane-Woon-Ming, Zsolt Ruzsics, Florian Erhard, Corinna Benkartek, Marina Babic, Ralf Zimmer, Joanne Trgovcich, Ulrich H. Koszinowski, Stipan Jonjic, Sébastien Pfeffer, and Lars Dölken. Degradation of cellular mir-27 by a novel, highly abundant viral transcript is important for efficient virus replication in vivo. *PLoS pathogens*, 8(2):e1002510, February 2012. 34

[187] Annalisa Marsico, Matthew R Huska, Julia Lasserre, Haiyang Hu, Dubravka Vucicevic, Anne Musahl, Ulf Andersson Orom, and Martin Vingron. PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome Biology*, 14(8):R84, 2013. 8

[188] Masayuki Matsui, Yongjun Chu, Huiying Zhang, Keith T. Gagnon, Sarfraz Shaikh, Satya Kuchimanchi, Muthiah Manoharan, David R. Corey, and Bethany A. Janowski. Promoter RNA links transcriptional regulation of inflammatory pathway genes. *Nucleic Acids Research*, 41(22):10086–10109, December 2013. 24

[189] J. H. Matthaei, O. W. Jones, R. G. Martin, and M. W. Nirenberg. Characteristics and composition of RNA coding units. *Proceedings of the National Academy of Sciences of the United States of America*, 48:666–677, April 1962. 125

[190] Chiara Mattioli, Giulia Pianigiani, and Franco Pagani. Cross talk between spliceosome and microprocessor defines the fate of pre-mRNA. *Wiley interdisciplinary reviews. RNA*, 5(5):647–658, October 2014. 9

[191] Matthew N. McCall, Min-Sik Kim, Mohammed Adil, Arun H. Patil, Yin Lu, Christopher J. Mitchell, Pamela Leal-Rojas, Jinchong Xu, Manoj Kumar, Valina L. Dawson, Ted M. Dawson, Alexander S. Baras, Avi Z. Rosenberg, Dan E. Arking, Kathleen H. Burns, Akhilesh Pandey, and Marc K. Halushka. Toward the human cellular microRNAome. *Genome Research*, 27(10):1769–1781, October 2017. 27, 28

[192] Katherine McJunkin and Victor Ambros. A microRNA family exerts maternal control on sex determination in C. elegans. *Genes & Development*, 31(4):422–437, 2017. 36

[193] H. A. Meijer, Y. W. Kong, W. T. Lu, A. Wilczynska, R. V. Spriggs, S. W. Robinson, J. D. Godfrey, A. E. Willis, and M. Bushell. Translational repression and eIF4a2 activity are critical for microRNA-mediated gene regulation. *Science (New York, N.Y.)*, 340(6128):82–85, April 2013. 20

[194] Gunter Meister, Markus Landthaler, Agnieszka Patkaniowska, Yair Dorsett, Grace Teng, and Thomas Tuschl. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular Cell*, 15(2):185–197, July 2004. 21

[195] Angeles Mencía, Silvia Modamio-Høybjør, Nick Redshaw, Matías Morín, Fernando Mayo-Merino, Leticia Olavarrieta, Luis A. Aguirre, Ignacio del Castillo, Karen P. Steel, Tamas Dalmay, Felipe Moreno, and Miguel Angel Moreno-Pelayo. Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nature Genetics*, 41(5):609–613, May 2009. 30, 35

[196] Mesfin K. Meshesha, Isana Veksler-Lublinsky, Ofer Isakov, Irit Reichenstein, Noam Shomron, Klara Kedem, Michal Ziv-Ukelson, Zvi Bentwich, and Yonat Shemer Avni.

The microRNA Transcriptome of Human Cytomegalovirus (HCMV). *The Open Virology Journal*, 6:38–48, 2012. 53

[197] Brittany L. Mihelich, Joseph C. Maranville, Rosalie Nolley, Donna M. Peehl, and Larisa Nonn. Elevated Serum MicroRNA Levels Associate with Absence of High-Grade Prostate Cancer in a Retrospective Cohort. *PLOS ONE*, 10(4):e0124245, April 2015. 37

[198] Patrick S. Mitchell, Rachael K. Parkin, Evan M. Kroh, Brian R. Fritz, Stacia K. Wyman, Era L. Pogosova-Agadjanyan, Amelia Peterson, Jennifer Noteboom, Kathy C. O'Briant, April Allen, Daniel W. Lin, Nicole Urban, Charles W. Drescher, Beatrice S. Knudsen, Derek L. Stirewalt, Robert Gentleman, Robert L. Vessella, Peter S. Nelson, Daniel B. Martin, and Muneesh Tewari. Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences*, 105(30):10513–10518, July 2008. 37

[199] Alex Mas Monteys, Ryan M. Spengler, Ji Wan, Luis Tecedor, Kimberly A. Lennox, Yi Xing, and Beverly L. Davidson. Structure and activity of putative intronic miRNA promoters. *RNA (New York, N.Y.)*, 16(3):495–505, March 2010. 8

[200] Michael J. Moore, Troels K. H. Scheel, Joseph M. Luna, Christopher Y. Park, John J. Fak, Eiko Nishiuchi, Charles M. Rice, and Robert B. Darnell. miRNA-target chimeras reveal miRNA 3'-end pairing as a major determinant of Argonaute target specificity. *Nature Communications*, 6:8864, November 2015. 17

[201] Sara Morales, Mariano Monzo, and Alfons Navarro. Epigenetic regulation mechanisms of microRNA expression. *Biomolecular Concepts*, 8(5-6):203–212, December 2017. 32

[202] Masaki Mori, Robinson Triboulet, Morvarid Mohseni, Karin Schlegelmilch, Kriti Shrestha, Fernando D. Camargo, and Richard I. Gregory. Hippo signaling regulates microprocessor and links cell-density-dependent miRNA biogenesis to cancer. *Cell*, 156(5):893–906, February 2014. 8

[203] Ryan D. Morin, Michael D. O'Connor, Malachi Griffith, Florian Kuchenbauer, Allen Delaney, Anna-Liisa Prabhu, Yongjun Zhao, Helen McDonald, Thomas Zeng, Martin Hirst, Connie J. Eaves, and Marco A. Marra. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research*, 18(4):610–621, April 2008. 27

[204] E. G. Moss, R. C. Lee, and V. Ambros. The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. *Cell*, 88(5):637–646, March 1997. 3

[205] Zissimos Mourelatos, Josée Dostie, Sergey Paushkin, Anup Sharma, Bernard Charroux, Linda Abel, Juri Rappsilber, Matthias Mann, and Gideon Dreyfuss. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes & Development*, 16(6):720–728, March 2002. 4

[206] Heiko Muller, Matteo Jacopo Marzi, and Francesco Nicassio. IsomiRage: From Functional Classification to Differential Expression of miRNA Isoforms. *Frontiers in Bioengineering and Biotechnology*, 2:38, 2014. 56

[207] I. A. Naess, S. C. Christiansen, P. Romundstad, S. C. Cannegieter, F. R. Rosendaal, and J. Hammerstrøm. Incidence and mortality of venous thrombosis: a population-based study. *Journal of thrombosis and haemostasis: JTH*, 5(4):692–699, April 2007. 85

[208] Kotaro Nakanishi. Anatomy of RISC: how do small RNAs and chaperones activate Argonaute proteins? *Wiley Interdisciplinary Reviews. RNA*, 7(5):637–660, 2016. 10, 187

[209] Vinny Negi, Deepanjan Paul, Sudipta Das, Prashant Bajpai, Suchita Singh, Arijit Mukhopadhyay, Anurag Agrawal, and Balaram Ghosh. Altered expression and editing of miRNA-100 regulates iTreg differentiation. *Nucleic Acids Research*, 43(16):8057–8065, September 2015. 29

[210] Joel R. Neilson, Grace X. Y. Zheng, Christopher B. Burge, and Phillip A. Sharp. Dynamic regulation of miRNA expression in ordered stages of cellular development. *Genes & Development*, 21(5):578–589, March 2007. 33

[211] Tuan Anh Nguyen, Myung Hyun Jo, Yeon-Gil Choi, Joha Park, S. Chul Kwon, Sung-chul Hohng, V. Narry Kim, and Jae-Sung Woo. Functional Anatomy of the Human Microprocessor. *Cell*, 161(6):1374–1387, June 2015. 7, 8, 187

[212] Majid Nikpay, Kaitlyn Beehler, Armand Valsesia, Jorg Hager, Mary-Ellen Harper, Robert Dent, and Ruth McPherson. Genome-wide identification of circulating-miRNA expression quantitative trait loci reveals the role of several miRNAs in the regulation of Cardiometabolic phenotypes. *Cardiovascular Research*, January 2019. 92

[213] J. Nourse, J. Braun, K. Lackner, S. Hüttelmaier, and S. Danckwardt. Large-scale identification of functional microRNA targeting reveals cooperative regulation of the hemostatic system. *Journal of Thrombosis and Haemostasis*, 16(11):2233–2245, 2018. 88

[214] Chimari Okada, Eiki Yamashita, Soo Jae Lee, Satoshi Shibata, Jun Katahira, Atsushi Nakagawa, Yoshihiro Yoneda, and Tomitake Tsukihara. A high-resolution structure of the pre-microRNA nuclear export machinery. *Science (New York, N.Y.)*, 326(5957):1275–1279, November 2009. 9

[215] Katsutomo Okamura and Eric C. Lai. Endogenous small interfering RNAs in animals. *Nature Reviews. Molecular Cell Biology*, 9(9):673–678, September 2008. 12, 29

[216] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, December 2008. 129

[217] Lorena Pantano, Xavier Estivill, and Eulàlia Martí. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Research*, 38(5):e34–e34, March 2010. 56

[218] Jong-Eun Park, Inha Heo, Yuan Tian, Dhirendra K. Simanshu, Hyeshik Chang, David Jee, Dinshaw J. Patel, and V. Narry Kim. Dicer recognizes the 5' end of RNA for efficient and accurate processing. *Nature*, 475(7355):201–205, July 2011. 9

[219] Mi Seul Park, Raul Araya-Secchi, James A. Brackbill, Hong-Duc Phan, Audrey C. Kehling, Ekram W. Abd El-Wahab, Daniel M. Dayeh, Marcos Sotomayor, and Kotaro Nakanishi. Multidomain Convergence of Argonaute during RISC Assembly Correlates with the Formation of Internal Water Clusters. *Molecular Cell*, July 2019. 22

[220] Mi Seul Park, Hong-Duc Phan, Florian Busch, Samantha H. Hinckley, James A. Brackbill, Vicki H. Wysocki, and Kotaro Nakanishi. Human Argonaute3 has slicer activity. *Nucleic Acids Research*, 45(20):11867–11877, November 2017. 21

[221] Alexander C. Partin, Byung-Cheon Jeong, Emily Herrell, Gary Hon, Tri D. Ngo, and Yunsun Nam. Heme enables proper positioning of Drosha and DGCR8 on primary microRNAs. *Nature Communications*, 8(1):1737, December 2017. 8

[222] Leopold Parts, Asa K. Hedman, Sarah Keildson, Andrew J. Knights, Cei Abreu-Goodger, Martijn van de Bunt, José Afonso Guerra-Assunçao, Nenad Bartonicek, Stijn van Dongen,

Reedik Magi, James Nisbet, Amy Barrett, Mattias Rantalainen, Alexandra C. Nica, Michael A. Quail, Kerrin S. Small, Daniel Glass, Anton J. Enright, John Winn, Panos Deloukas, Emmanouil T. Dermitzakis, Mark I. McCarthy, Timothy D. Spector, Richard Durbin, and Cecilia M. Lindgren. Extent, Causes, and Consequences of Small RNA Expression Variation in Human Adipose Tissue. *PLoS Genetics*, 8(5), May 2012. 32

[223] A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, J. Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, November 2000. 4

[224] David M. Patrick, Cheng C. Zhang, Ye Tao, Huiyu Yao, Xiaoxia Qi, Robert J. Schwartz, Lily Jun-Shen Huang, and Eric N. Olson. Defective erythroid differentiation in miR-451 mutant mice mediated by 14-3-3zeta. *Genes & Development*, 24(15):1614–1619, August 2010. 13

[225] James G. Patton, Jeffrey L. Franklin, Alissa M. Weaver, Kasey Vickers, Bing Zhang, Robert J. Coffey, K. Mark Ansel, Robert Blelloch, Andrei Goga, Bo Huang, Noelle L'Etoille, Robert L. Raffai, Charles P. Lai, Anna M. Krichevsky, Bogdan Mateescu, Vanille J. Greiner, Craig Hunter, Olivier Voinnet, and Michael T. McManus. Biogenesis, delivery, and function of extracellular RNA. *Journal of Extracellular Vesicles*, 4:27494, 2015. 33

[226] Sébastien Pfeffer, Alain Sewer, Mariana Lagos-Quintana, Robert Sheridan, Chris Sander, Friedrich A. Grässer, Linda F. van Dyk, C. Kiong Ho, Stewart Shuman, Minchen Chien, James J. Russo, Jingyue Ju, Glenn Randall, Brett D. Lindenbach, Charles M. Rice, Viviana Simon, David D. Ho, Mihaela Zavolan, and Thomas Tuschl. Identification of microRNAs of the herpesvirus family. *Nature Methods*, 2(4):269–276, April 2005. 53

[227] Ernesto Picardi, Anna Maria D'Erchia, Claudio Lo Giudice, and Graziano Pesole. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Research*, 45(D1):D750–D757, January 2017. 54

[228] Natalia Pinzón, Blaise Li, Laura Martinez, Anna Sergeeva, Jessy Presumey, Florence Apparailly, and Hervé Seitz. microRNA target prediction programs predict many false positives. *Genome Research*, 27(2):234–245, 2017. 35

[229] Robert F. Place, Long-Cheng Li, Deepa Pookot, Emily J. Noonan, and Rajvir Dahiya. MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 105(5):1608–1613, February 2008. 24

[230] Carsten A. Raabe, Thean-Hock Tang, Juergen Brosius, and Timofey S. Rozhdestvensky. Biases in small RNA deep sequencing data. *Nucleic Acids Research*, 42(3):1414–1426, February 2014. 46

[231] Mattias Rantalainen, Blanca M. Herrera, George Nicholson, Rory Bowden, Quin F. Wills, Josine L. Min, Matt J. Neville, Amy Barrett, Maxine Allen, Nigel W. Rayner, Jan Fleckner, Mark I. McCarthy, Krina T. Zondervan, Fredrik Karpe, Chris C. Holmes, and Cecilia M. Lindgren. MicroRNA expression in abdominal and gluteal adipose tissue is associated with mRNA expression levels and partly genetically driven. *PloS One*, 6(11):e27338, 2011. 32

[232] Kasper D. Rasmussen, Salvatore Simmini, Cei Abreu-Goodger, Nenad Bartonicek, Monica Di Giacomo, Daniel Bilbao-Cortes, Rastislav Horos, Marieke Von Lindern, Anton J. Enright, and Dónal O'Carroll. The miR-144/451 locus is required for erythroid homeostasis. *The Journal of Experimental Medicine*, 207(7):1351–1358, July 2010. 13

[233] Jan Rehwinkel, Isabelle Behm-Ansmant, David Gatfield, and Elisa Izaurralde. A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. *RNA*, 11(11):1640–1647, November 2005. 19

[234] Brian Reichholf, Veronika A. Herzog, Nina Fasching, Raphael A. Manzenreither, Ivica Sowemimo, and Stefan L. Ameres. Time-Resolved Small RNA Sequencing Unravels the Molecular Principles of MicroRNA Homeostasis. *Molecular Cell*, July 2019. 24

[235] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–906, February 2000. 4

[236] Olivia S. Rissland, Sue-Jean Hong, and David P. Bartel. MicroRNA Destabilization Enables Dynamic Regulation of the miR-16 Family in Response to Cell-Cycle Changes. *Molecular Cell*, 43(6):993–1004, September 2011. 25, 34

[237] Thomas C Roberts. The MicroRNA Biology of the Mammalian Nucleus. *Molecular Therapy - Nucleic Acids*, 3:e188, January 2014. 24

[238] Antony Rodriguez, Sam Griffiths-Jones, Jennifer L. Ashurst, and Allan Bradley. Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Research*, 14(10a):1902–1910, October 2004. 9

[239] Eva van Rooij, Lillian B. Sutherland, Xiaoxia Qi, James A. Richardson, Joseph Hill, and Eric N. Olson. Control of Stress-Dependent Cardiac Growth and Gene Expression by a MicroRNA. *Science*, 316(5824):575–579, April 2007. 25

[240] Trine B. Rounge, Sinan U. Umu, Andreas Keller, Eckart Meese, Giske Ursin, Steinar Tretli, Robert Lyle, and Hilde Langseth. Circulating small non-coding RNAs associated with age, sex, smoking, body mass and physical activity. *Scientific Reports*, 8(1):17650, December 2018. 31

[241] J. Graham Ruby, Calvin H. Jan, and David P. Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149):83–86, July 2007. 12

[242] Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language ? *Cell*, 146(3):353–358, August 2011. 26

[243] William E. Salomon, Samson M. Jolly, Melissa J. Moore, Phillip D. Zamore, and Victor Serebrov. Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides. *Cell*, 162(1):84–95, July 2015. 17, 18

[244] Aishe A. Sarshad, Aster H. Juan, Ana Iris Correa Muler, Dimitrios G. Anastasakis, Xiantao Wang, Pavol Genzor, Xuesong Feng, Pei-Fang Tsai, Hong-Wei Sun, Astrid D. Haase, Vittorio Sartorelli, and Markus Hafner. Argonaute-miRNA Complexes Silence Target mRNAs in the Nucleus of Mammalian Stem Cells. *Molecular Cell*, 71(6):1040–1050.e8, September 2018. 23

[245] Nicole T. Schirle, Jessica Sheu-Gruttadauria, and Ian J. MacRae. Structural basis for microRNA targeting. *Science (New York, N.Y.)*, 346(6209):608–613, October 2014. 18, 187

[246] Daniel Schraivogel, Susann G. Schindler, Johannes Danner, Elisabeth Kremmer, Janina Pfaff, Stefan Hannus, Reinhard Depping, and Gunter Meister. Importin-beta facilitates nuclear import of human GW proteins and balances cytoplasmic gene silencing protein levels. *Nucleic Acids Research*, 43(15):7447–7461, September 2015. 23

[247] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011. 24, 25

[248] Dianne S. Schwarz, György Hutvágner, Tingting Du, Zuoshang Xu, Neil Aronin, and Phillip D. Zamore. Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115(2):199–208, October 2003. 11

[249] Hervé Seitz, Hélène Royo, Marie-Line Bortolin, Shau-Ping Lin, Anne C. Ferguson-Smith, and Jérôme Cavaillé. A Large Imprinted microRNA Gene Cluster at the Mouse Dlk1-Gtl2 Domain. *Genome Research*, 14(9):1741–1748, September 2004. 32

[250] Matthias Selbach, Björn Schwanhäusser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, September 2008. 34

[251] Jessica Sheu-Gruttadauria, Paulina Pawlica, Shannon M. Klum, Sonia Wang, Therese A. Yario, Nicole T. Schirle Oakdale, Joan A. Steitz, and Ian J. MacRae. Structural Basis for Target-Directed MicroRNA Degradation. *Molecular Cell*, July 2019. 25, 29

[252] Jessica Sheu-Gruttadauria, Yao Xiao, Luca FR Gebert, and Ian J. MacRae. Beyond the seed: structural basis for supplementary microRNA targeting by human Argonaute2. *The EMBO Journal*, page e101153, April 2019. 28

[253] Chanseok Shin, Jin-Wu Nam, Kyle Kai-How Farh, H. Rosaria Chiang, Alena Shkumatava, and David P. Bartel. Expanding the microRNA targeting code: functional sites with centered pairing. *Molecular Cell*, 38(6):789–802, June 2010. 21

[254] Katherine J. Siddle, Matthieu Deschamps, Ludovic Tailleux, Yohann Nédélec, Julien Pothlichet, Geanncarlo Lugo-Villarino, Valentina Libri, Brigitte Gicquel, Olivier Neyrolles, Guillaume Laval, Etienne Patin, Luis B. Barreiro, and Lluís Quintana-Murci. A genomic portrait of the genetic architecture and regulatory impact of microRNA expression in response to infection. *Genome Research*, 24(5):850–859, May 2014. 32

[255] F. J. Slack, M. Basson, Z. Liu, V. Ambros, H. R. Horvitz, and G. Ruvkun. The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. *Molecular Cell*, 5(4):659–669, April 2000. 4

[256] Julie Soutourina. Transcription regulation by the Mediator complex. *Nature Reviews Molecular Cell Biology*, 19(4):262–274, April 2018. 131, 188

[257] Irina Starikova, Simin Jamaly, Antonio Sorrentino, Thorarinn Blondal, Nadezhda Latysheva, Mikhail Sovershaev, and John-Bjarne Hansen. Differential expression of plasma miRNAs in patients with unprovoked venous thromboembolism and healthy control individuals. *Thrombosis Research*, 136(3):566–572, September 2015. 89

[258] Hong Su, Melanie I. Trombly, Jian Chen, and Xiaozhong Wang. Essential and overlapping functions for mammalian Argonautes in microRNA silencing. *Genes & Development*, 23(3):304–317, February 2009. 22

[259] Sunderland Nicholas, Skroblin Philipp, Barwari Temo, Huntley Rachael P., Lu Ruifang, Joshi Abhishek, Lovering Ruth C., and Mayr Manuel. MicroRNA Biomarkers and Platelet Reactivity. *Circulation Research*, 120(2):418–435, January 2017. 88

[260] Hiroshi I. Suzuki, Akihiro Katsura, Takahiko Yasuda, Toshihide Ueno, Hiroyuki Mano, Koichi Sugimoto, and Kohei Miyazono. Small-RNA asymmetry is directly driven by mammalian Argonautes. *Nature Structural & Molecular Biology*, 22(7):512–521, July 2015. 11

[261] Pål Sætrom, Bret S.E. Heale, Ola Snøve, Lars Aagaard, Jessica Alluin, and John J. Rossi. Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Research*, 35(7):2333–2342, April 2007. 22, 34

[262] Yuliang Tan, Bo Zhang, Tao Wu, Geir Skogerbø, Xiaopeng Zhu, Xiangqian Guo, Shunmin He, and Runsheng Chen. Transcriptional inhibiton of Hoxd4 expression by miRNA-10a in human breast cancer cells. *BMC molecular biology*, 10:12, February 2009. 24

[263] Aristeidis G. Telonis, Phillipe Loher, Yi Jing, Eric Londin, and Isidore Rigoutsos. Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. *Nucleic Acids Research*, 43(19):9158–9175, October 2015. 31

[264] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015. 91

[265] Thomas Thomou, Marcelo A. Mori, Jonathan M. Dreyfuss, Masahiro Konishi, Masaji Sakaguchi, Christian Wolfrum, Tata Nageswara Rao, Jonathon N. Winnay, Ruben Garcia-Martin, Steven K. Grinspoon, Phillip Gorden, and C. Ronald Kahn. Adipose-derived circulating miRNAs regulate gene expression in other tissues. *Nature*, 542(7642):450–455, 2017. 32

[266] Daniel W. Thomson and Marcel E. Dinger. Endogenous microRNA sponges: evidence and controversy. *Nature Reviews Genetics*, 17(5):272–283, May 2016. 26

[267] Daniel W. Thomson, Katherine A. Pillman, Matthew L. Anderson, David M. Lawrence, John Toubia, Gregory J. Goodall, and Cameron P. Bracken. Assessing the gene regulatory properties of Argonaute-bound small RNAs of diverse genomic origin. *Nucleic Acids Research*, 43(1):470–481, January 2015. 14

[268] Zhan Tong, Qinghua Cui, Juan Wang, and Yuan Zhou. TransmiR v2.0: an updated transcription factor-microRNA regulation database. *Nucleic Acids Research*, 47(D1):D253–D258, January 2019. 6

[269] Michele Trabucchi, Paola Briata, Mariaflor Garcia-Mayoral, Astrid D. Haase, Witold Filipowicz, Andres Ramos, Roberto Gherzi, and Michael G. Rosenfeld. The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs. *Nature*, 459(7249):1010–1014, June 2009. 8

[270] Thomas Treiber, Nora Treiber, and Gunter Meister. Regulation of microRNA biogenesis and its crosstalk with other cellular pathways. *Nature Reviews Molecular Cell Biology*, 20(1):5, January 2019. 11, 187

[271] Andrey Turchinovich, Alexander G. Tonevitsky, William C. Cho, and Barbara Burwinkel. Check and mate to exosomal extracellular miRNA: new lesson from a new approach. *Frontiers in Molecular Biosciences*, 2, April 2015. 32

[272] Andrey Turchinovich, Ludmila Weiz, Anne Langheinz, and Barbara Burwinkel. Characterization of extracellular circulating microRNA. *Nucleic Acids Research*, 39(16):7223–7233, September 2011. 32, 33

[273] Shobha Vasudevan. Posttranscriptional Upregulation by MicroRNAs. *Wiley Interdisciplinary Reviews: RNA*, 3(3):311–330, 2012. 23

[274] Shobha Vasudevan, Yingchun Tong, and Joan A. Steitz. Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation. *Science*, 318(5858):1931–1934, December 2007. 23

[275] Christine Wahlquist, Dongtak Jeong, Agustin Rojas-Muñoz, Changwon Kho, Ahyoung Lee, Shinichi Mitsuyama, Alain van Mil, Woo Jin Park, Joost P. G. Sluijter, Pieter A. F. Doevendans, Roger J. Hajjar, and Mark Mercola. Inhibition of *miR-25* improves cardiac contractility in the failing heart. *Nature*, 508(7497):531–535, April 2014. 17

[276] Dongmei Wang, Zhaojie Zhang, Evan O'Loughlin, Thomas Lee, Stephane Houel, Dónal O'Carroll, Alexander Tarakhovsky, Natalie G. Ahn, and Rui Yi. Quantitative functions of Argonaute proteins in mammalian development. *Genes & Development*, 26(7):693–704, April 2012. 22

[277] Jie-Mei Wang, Jun Tao, Dan-Dan Chen, Jing-Jing Cai, Kaikobad Irani, Qinde Wang, Hong Yuan, and Alex F. Chen. MicroRNA miR-27b rescues bone marrow-derived angiogenic cell function and accelerates wound healing in type 2 diabetes mellitus. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 34(1):99–109, January 2014. 113

[278] Xiao Wang, Kristina Sundquist, Peter J. Svensson, Hamideh Rastkhani, Karolina Palmér, Ashfaque A. Memon, Jan Sundquist, and Bengt Zöller. Association of recurrent venous thromboembolism and circulating microRNAs. *Clinical Epigenetics*, 11(1):28, February 2019. 89

[279] Yanli Wang, Gang Sheng, Stefan Juranek, Thomas Tuschl, and Dinshaw J. Patel. Structure of the guide-strand-containing argonaute silencing complex. *Nature*, 456(7219):209–213, November 2008. 18

[280] J. D. Watson and F. H. C. Crick. Genetical Implications of the Structure of Deoxyribonucleic Acid. *Nature*, 171(4361):964, May 1953. 125, 126, 188

[281] Jessica A. Weber, David H. Baxter, Shile Zhang, David Y. Huang, Kuo How Huang, Ming Jen Lee, David J. Galas, and Kai Wang. The microRNA spectrum in 12 body fluids. *Clinical Chemistry*, 56(11):1733–1741, November 2010. 32

[282] Liang Meng Wee, C. Fabián Flores-Jasso, William E. Salomon, and Phillip D. Zamore. Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell*, 151(5):1055–1067, November 2012. 17

[283] Yao Wei, Limin Li, Dong Wang, Chen-Yu Zhang, and Ke Zen. Importin 8 regulates the transport of mature microRNAs into the cell nucleus. *The Journal of Biological Chemistry*, 289(15):10270–10275, April 2014. 23

[284] Marc S. Weinberg and Kevin V. Morris. Transcriptional gene silencing in humans. *Nucleic Acids Research*, 44(14):6505–6517, August 2016. 24

[285] Jiayu Wen, Erik Ladewig, Sol Shenker, Jaaved Mohammed, and Eric C. Lai. Analysis of Nearly One Thousand Mammalian Mirtrons Reveals Novel Features of Dicer Substrates. *PLoS computational biology*, 11(9):e1004441, September 2015. 12, 13, 187

[286] Harm-Jan Westra, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, Benjamin P. Fairfax, Katharina Schramm, Joseph E. Powell, Alexandra Zhernakova, Daria V. Zhernakova, Jan H. Veldink, Leonard H. Van den Berg, Juha Karjalainen, Sebo Withoff, André G. Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter A. C. 't Hoen, Eva Reinmaa, Krista Fischer, Mari Nelis, Lili Milani, David Melzer, Luigi Ferrucci, Andrew B. Singleton, Dena G. Hernandez, Michael A. Nalls, Georg Homuth, Matthias Nauck, Dörte Radke, Uwe Völker,

Markus Perola, Veikko Salomaa, Jennifer Brody, Astrid Suchy-Dicey, Sina A. Gharib, Daniel A. Enquobahrie, Thomas Lumley, Grant W. Montgomery, Seiko Makino, Holger Prokisch, Christian Herder, Michael Roden, Harald Grallert, Thomas Meitinger, Konstantin Strauch, Yang Li, Ritsert C. Jansen, Peter M. Visscher, Julian C. Knight, Bruce M. Psaty, Samuli Ripatti, Alexander Teumer, Timothy M. Frayling, Andres Metspalu, Joyce B. J. van Meurs, and Lude Franke. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, October 2013. 111

[287] B. Wightman, T. R. Bürglin, J. Gatto, P. Arasu, and G. Ruvkun. Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. *Genes & Development*, 5(10):1813–1824, October 1991. 3

[288] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–862, December 1993. 3

[289] Peter Willeit, Anna Zampetaki, Katarzyna Dudek, Dorothee Kaudewitz, Alice King, Nicholas S. Kirkby, Roxanne Crosby-Nwaobi, Marianna Prokopi, Ignat Drozdov, Sarah R. Langley, Sobha Sivaprasad, Hugh S. Markus, Jane A. Mitchell, Timothy D. Warner, Stefan Kiechl, and Manuel Mayr. Circulating microRNAs as novel biomarkers for platelet activation. *Circulation Research*, 112(4):595–600, February 2013. 88

[290] Ross C. Wilson, Akshay Tambe, Mary Anne Kidwell, Cameron L. Noland, Catherine P. Schneider, and Jennifer A. Doudna. Dicer-TRBP complex formation ensures accurate mammalian microRNA biogenesis. *Molecular Cell*, 57(3):397–407, February 2015. 27

[291] Kenneth W. Witwer and Marc K. Halushka. Toward the promise of microRNAs – Enhancing reproducibility and rigor in microRNA research. *RNA Biology*, 13(11):1103–1116, November 2016. 46

[292] Carrie Wright, Anandita Rajpurohit, Emily E. Burke, Courtney Williams, Leonardo Collado-Torres, Martha Kimos, Nicholas J. Brandon, Alan J. Cross, Andrew E. Jaffe, Daniel R. Weinberger, and Joo Heon Shin. Comprehensive assessment of multiple biases in small RNA sequencing reveals significant differences in the performance of widely used methods. *BMC Genomics*, 20(1):513, June 2019. 46, 47, 188

[293] Han Wu, Shuying Sun, Kang Tu, Yuan Gao, Bin Xie, Adrian R. Krainer, and Jun Zhu. A splicing-independent function of SF2/ASF in microRNA processing. *Molecular Cell*, 38(1):67–77, April 2010. 8

[294] Xiaogang Wu, Taek-Kyun Kim, David Baxter, Kelsey Scherler, Aaron Gordon, Olivia Fong, Alton Etheridge, David J. Galas, and Kai Wang. sRNAnalyzer-a flexible and customizable small RNA sequencing data analysis pipeline. *Nucleic Acids Research*, October 2017. 55

[295] Stacia K. Wyman, Emily C. Knouf, Rachael K. Parkin, Brian R. Fritz, Daniel W. Lin, Lucas M. Dennis, Michael A. Krouse, Philippa J. Webster, and Muneesh Tewari. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Research*, 21(9):1450–1461, September 2011. 28

[296] Luoxing Xia, Zhi Zeng, and Wai Ho Tang. The Role of Platelet Microparticle Associated microRNAs in Cellular Crosstalk. *Frontiers in Cardiovascular Medicine*, 5, April 2018. 88

[297] Mingyi Xie, Mingfeng Li, Anna Vilborg, Nara Lee, Mei-Di Shu, Valeria Yartseva, Nenad Šestan, and Joan A. Steitz. Mammalian 5'-capped microRNA precursors that generate a single microRNA. *Cell*, 155(7):1568–1580, December 2013. 12

[298] Yi-Fan Xu, Bethany N. Hannafon, Ujjwol Khatri, Amy Gin, and Wei-Qun Ding. The origin of exosomal miR-1246 in human cancer cells. *RNA biology*, 16(6):770–784, 2019. 14

[299] Acong Yang, Xavier Bofill-De Ros, Tie-Juan Shao, Minjie Jiang, Katherine Li, Patricia Villanueva, Lisheng Dai, and Shuo Gu. 3' Uridylation Confers miRNAs with Non-canonical Target Repertoires. *Molecular Cell*, 0(0), June 2019. 28

[300] Jr-Shiuan Yang, Michael D. Phillips, Doron Betel, Ping Mu, Andrea Ventura, Adam C. Siepel, Kevin C. Chen, and Eric C. Lai. Widespread regulatory activity of vertebrate microRNA* species. *RNA (New York, N.Y.)*, 17(2):312–326, February 2011. 11

[301] Rui Yi, Yi Qin, Ian G. Macara, and Bryan R. Cullen. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & Development*, 17(24):3011–3016, December 2003. 9

[302] Mayuko Yoda, Tomoko Kawamata, Zain Paroo, Xuecheng Ye, Shintaro Iwasaki, Qinghua Liu, and Yukihide Tomari. ATP-dependent human RISC assembly pathways. *Nature Structural & Molecular Biology*, 17(1):17–23, January 2010. 10, 22

[303] Scott T. Younger and David R. Corey. Transcriptional gene silencing in mammalian cells by miRNA mimics that target gene promoters. *Nucleic Acids Research*, 39(13):5682–5691, July 2011. 24

[304] Kai Zhang, Xiaorong Zhang, Zhiqiang Cai, Jie Zhou, Ran Cao, Ya Zhao, Zonggui Chen, Dehe Wang, Wen Ruan, Qian Zhao, Guangqiao Liu, Yuanchao Xue, Yan Qin, Bing Zhou, Ligang Wu, Timothy Nilsen, Yu Zhou, and Xiang-Dong Fu. A Novel Class of MicroRNA Recognition Elements That Function Only in Open Reading Frames. *Nature structural & molecular biology*, 25(11):1019–1027, November 2018. 22

[305] Weihua Zhu, Min Yang, Jinnan Shang, Yiliang Xu, Yuanlang Wang, Qiangqiang Tao, Liang Zhang, Yueyun Ding, Yige Chen, Dongdong Zhao, Chonglong Wang, Mingxing Chu, Zongjun Yin, and Xiaodong Zhang. MiR-222 inhibits apoptosis in porcine follicular granulosa cells by targeting the THBS1 gene. *Animal Science Journal = Nihon Chikusan Gakkaiho*, 90(6):719–727, June 2019. 111

[306] Bengt Zöller, Xinjun Li, Jan Sundquist, and Kristina Sundquist. Familial transmission of venous thromboembolism: a cohort study of 80 214 Swedish adoptees linked to their biological and adoptive parents. *Circulation. Cardiovascular Genetics*, 7(3):296–303, June 2014. 86

[307] Ulf Andersson Ørom, Finn Cilius Nielsen, and Anders H. Lund. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Molecular Cell*, 30(4):460–471, May 2008. 23

Table des figures

Liste des tableaux

Florian THIBORD **11 décembre 2019**

Sujet : Variation génétique et plasmatique des microARNs - Impact sur les paramètres biologiques de l'hémostase

Résumé : Les microARNs (miARNs) sont les membres d'une classe de petits ARNs non codants d'environ 22 nucléotides, dont le mécanisme principal est de réguler l'expression des gènes dans le cytoplasme. Leurs importance est telle qu'il est estimé que la majorité des gènes humains sont régulés par ces petits ARNs, et ils sont ainsi potentiellement impliqués dans le développement de nombreuse pathologies. La séquence des miARNs peut-être soumise à des variations post-transcriptionnelles et des variations génétiques générant alors des séquences isoformes appelées isomiRs. Afin de détécter et quantifier précisemment l'expression des miARNs à partir de données de séquençage, cette hétérogénéité intra-miARN, due aux isomiRs, doit être prise en compte, tout comme l'homogénéité inter-miARN due aux miARNs paralogues. Le pipeline optimiR, développé dans le cadre de cette thèse, permet de surmonter ces challenges grâce notamment à l'intégration de l'information génétique des échantillons analysés, ainsi qu'à une stratégie d'alignement originale, qui permettent de détecter les isomiRs tout en distinguant les miARNs paralogues. Les données analysées lors de cette thèse proviennent de la cohorte MARTHA, composée de patients ayant développé une thrombose veineuse (VTE), parfois avec récidive. L'expression normalisée de 162 miARNs obtenue pour 344 patients a ensuite été utilisée afin d'analyser: 1) les déterminants génétiques de l'expression de ces miARNs; 2) l'association des miARNs avec le risque de récidive pour la VTE; 3) les corrélations avec certains paramètres biologiques de l'hémostase. Collectivement, ces analyses m'ont permis d'identifier des microARNs d'intérêt pour la recherche sur la thrombose veineuse et sur l'hémostase.

Mots clés : microARN; Génétique; Séquençage; Etudes d'Associations; Hémostase; Thrombose Veineuse

Subject : Genetic and plasmatic variations of microRNAs - Impact on haemostatic traits

Abstract: MicroRNAs (miRNA) are small non coding RNAs with an average size of 22 nucleotides, mainly known to regulate gene expression in the cytoplasm. These small RNAs are estimated to regulate the majority of human genes, and are potentially involved in several diseases. MiRNA sequences might contain genetic variants and can undergo post-transcriptional variations, which generate miRNA isoforms called isomiRs. In order to accurately detect and quantify miRNA expression, isomiRs as well as paralogous miRNAs must be accounted for. The optimiR pipeline developed during this project overcome these challenges by integrating genetic information and by implementing an original strategy based on local alignement. Sequencing data were obtained from the MARTHA cohort, which is composed of french unrelated patients who experienced venous thrombosis (VTE). Normalized expression of 162 miRNAs from 334 patients were used to analyze: 1) the genetic determinants of miRNA expression; 2) the association of miRNA expression levels with VTE recurence; 3) the correlations between miRNA expression levels and hemostatic traits. As a whole, these analyses allowed me to identify miRNAs of interest for the study of VTE and hemostasis.

Keywords : microRNA; Genetics; Sequencing; Association Studies; Hemostasis; Veinous Thrombosis