Santé des Plantes
Interactions Biotiques
Outils scientifiques Collectifs
BioInformatique et Génomique

Journée Annotation structurale et fonctionnelle des génomes eukaryotes : Annotation des smallRNAs

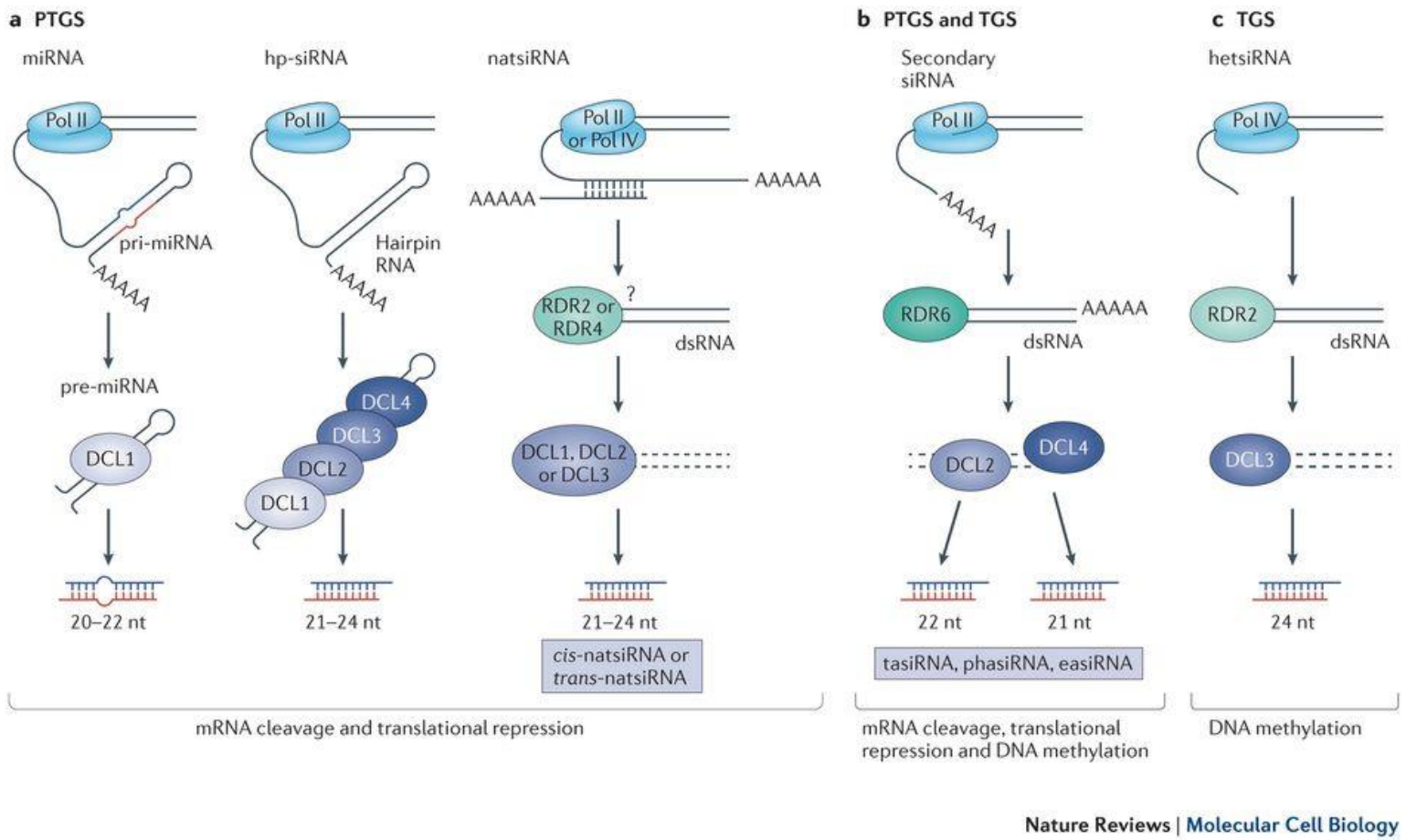
Martine Da Rocha

SCIENCE & IMPACT

Membre de UNIVERSITÉ CÔTE D'AZUR

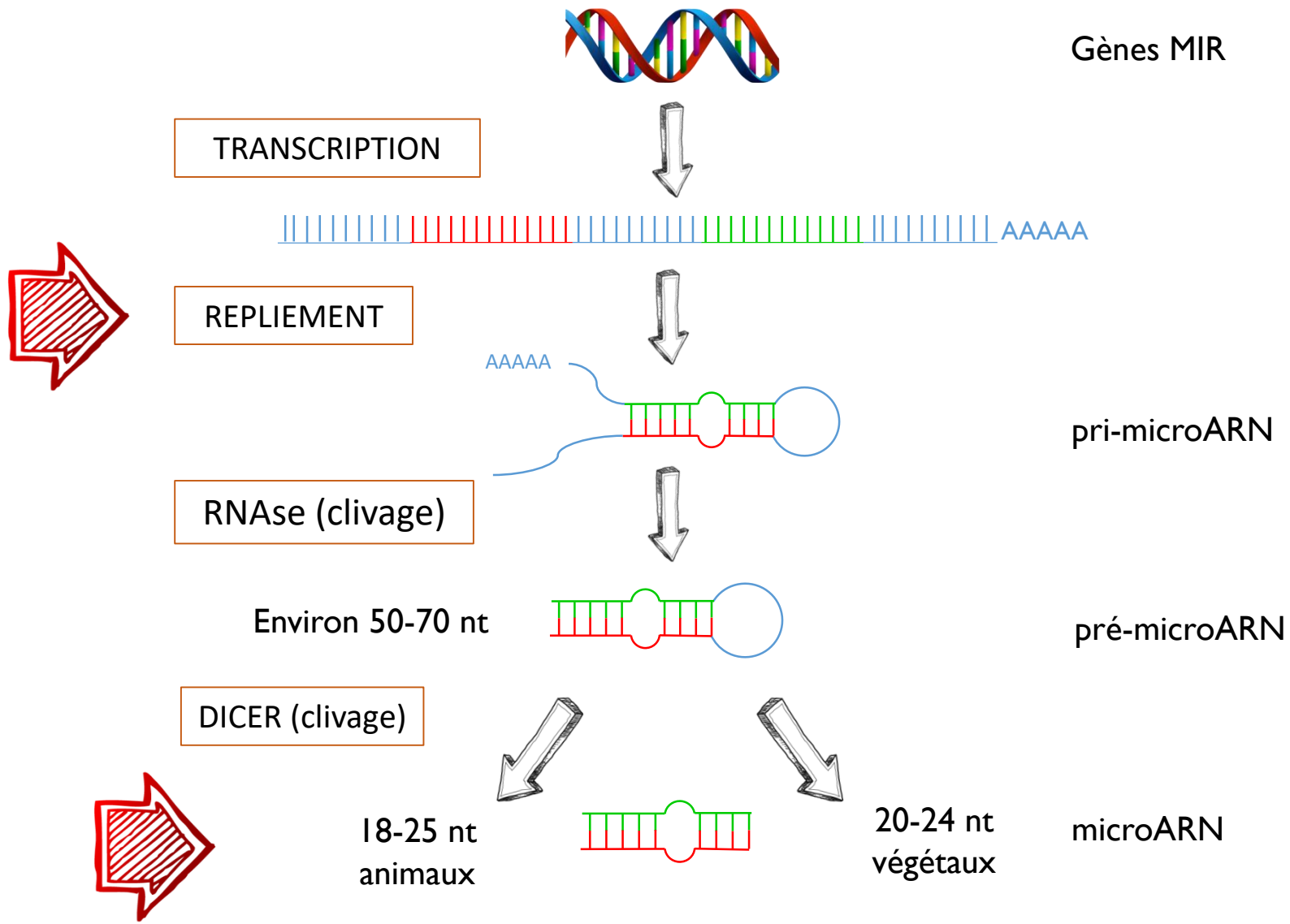
SANTE DES PLANTES - ENVIRONNEMENT

SmallRNA Word



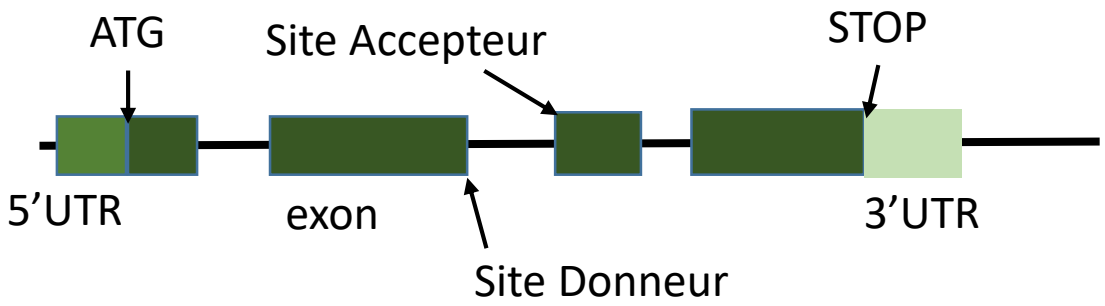
Les SmallRnas -> Repliement -> maturation -> Structure action biologique

Biosynthèse des microARN chez les eucaryotes



Challenge de la prédiction des smallRNAS

Gènes codants pour des protéines



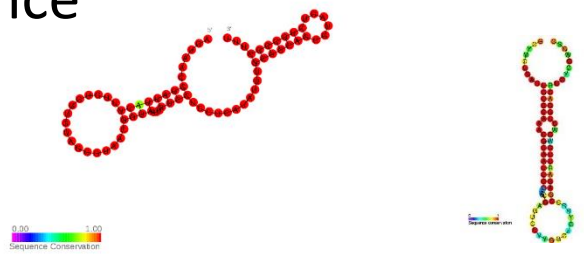
Gènes codants pour des SmallRNA

Schéma de structure lisible dans la séquence

- cadres de lecture ouverts
- sites d'épissage
- fonction liée à la séquence

Absence de schéma de structure lisible dans la séquence

- fonction lié au repliement de la séquence



Challenge de la prédiction des smallRNAS

- Critère de prédiction :
  - A partir de données de séquençage.
  - En fonction de la nature des petits ARNs : structure en hairpin
  - Pour les miRNAs à partir des miRNA connues (Conservation des séquences)

problème beaucoup de faux positif

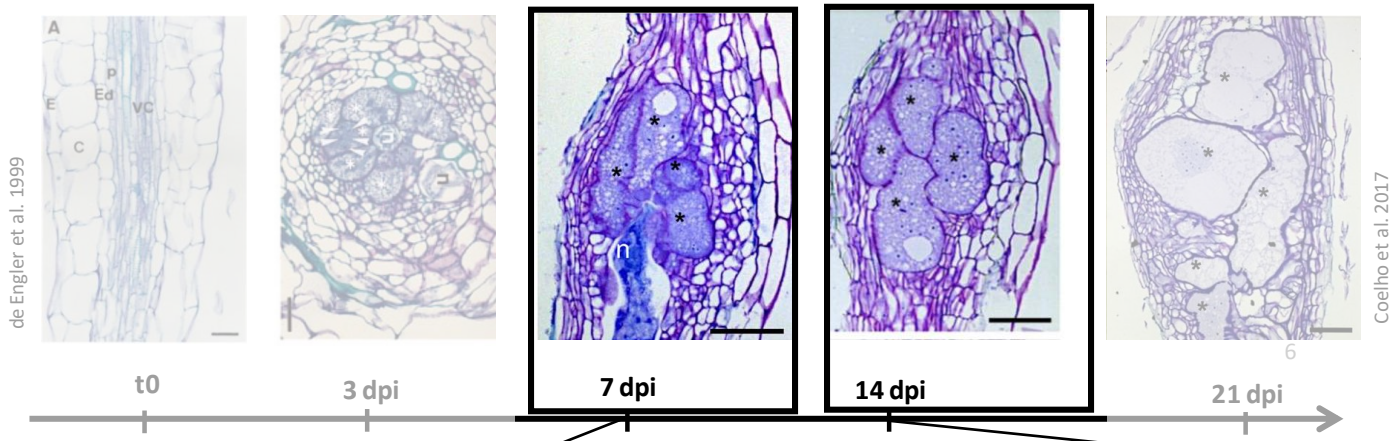
Stratégies choisies :

- Utilisation de réplicas biologique => même prédiction / dans les réplicas
- Si possible : recouper les résultats de plusieurs méthodes
- Attention au prédiction liée aux homologies : risque de propagés des erreurs

Contexte de l'études : Identification des sRNAs exprimés dans la racine de tomate lors de l'infection par un nématode

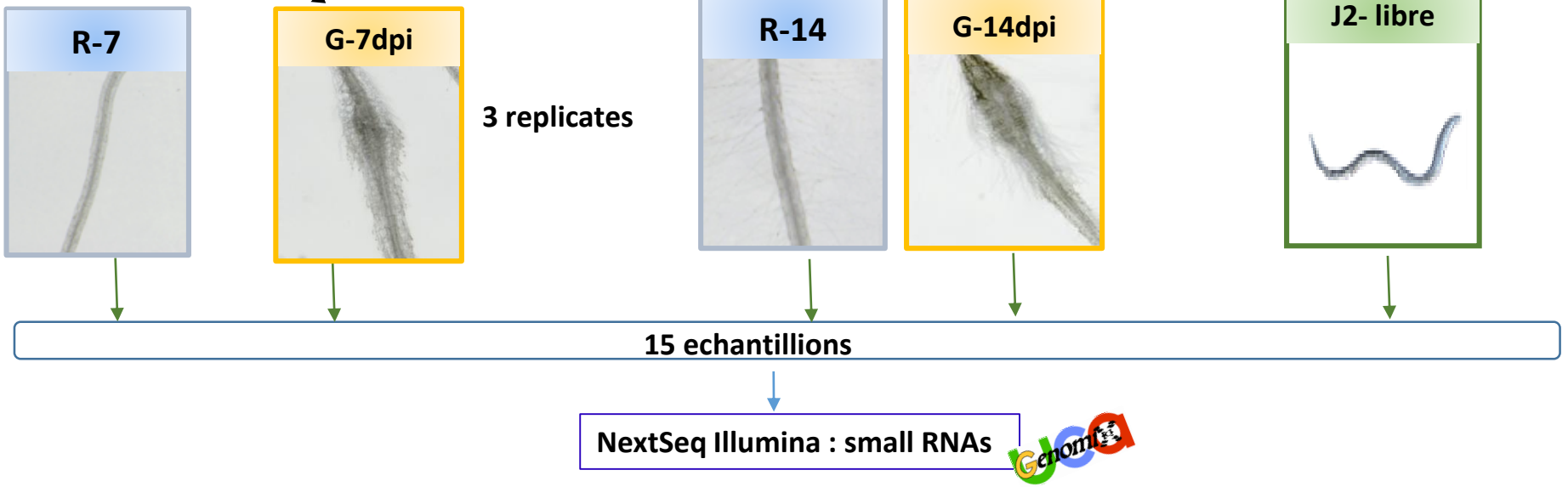
M. incognita

Solanum lycopersicum

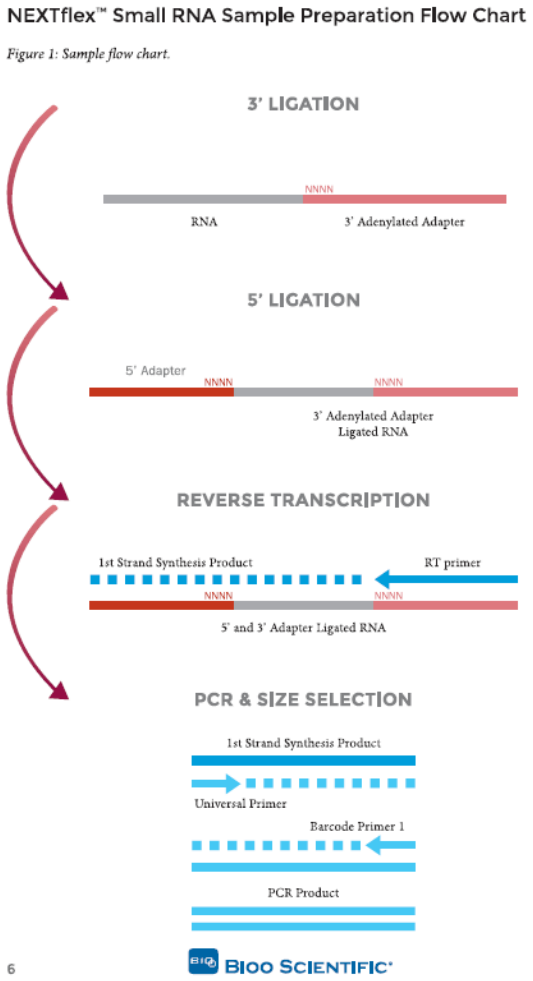


Galles (G) : Racines infectées

Racines saine (R)



Séquençage des smallRNAs

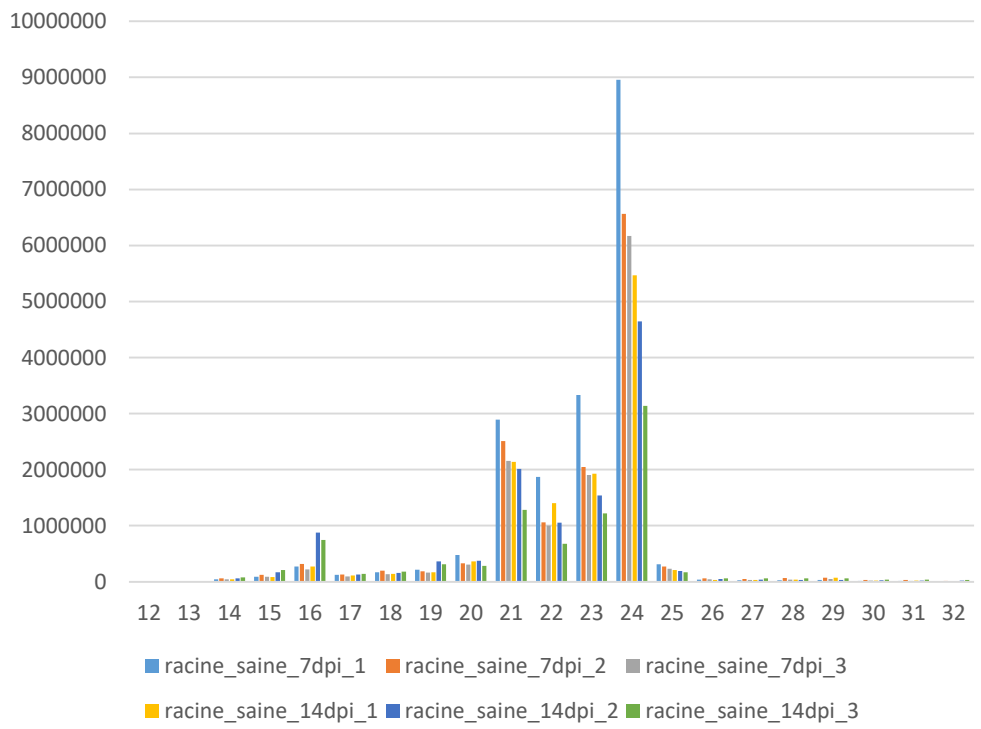


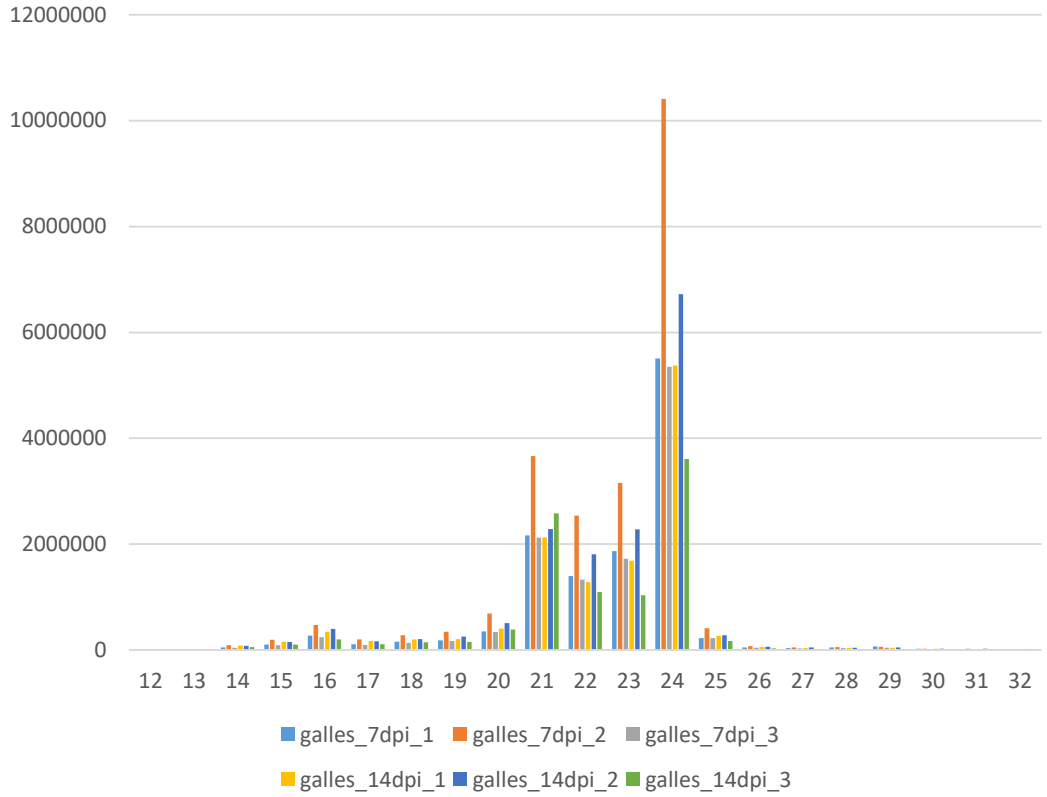
Purification basé sur la :

- Taille
- Adaptateur liaisons 5'P libres (ARNm ont une coiffe liaison impossible)

	Tissu végétal	Tissu animal	Nombre de reads
Référence Nématodes		J2_libres_1	38'251'985
		J2_libres_2	40'119'751
		J2_libres_3	34'527'534
Référence - Racines de 14 jours	Racines_saines_14dpi_1		13'971'658
	Racines_saines_14dpi_2		13'457'914
	Racines_saines_14dpi_3		10'334'113
Référence - Racines de 7 jours	Racines_saines_7dpi_1		20'995'892
	Racines_saines_7dpi_2		15'823'763
	Racines_saines_7dpi_2		14'167'136
Expérience - 14 jours après infection	Galles_14dpi_1		14'342'492
	Galles_14dpi_2		17'499'887
	Galles_14dpi_3		10'991'707
Expérience - 7 jours après infection	Galles_7dpi_1		14'083'472
	Galles_7dpi_2		25'164'984
	Galles_7dpi_3		13'419'506

Tailles de fragments obtenues





Pics à 24 et 21 nucléotides

Outils bioinformatique utilisés

- Tous les smallRNA : ShortStack

RNA. 2013 Jun;19(6):740-51. doi: 10.1261/rna.035279.112. Epub 2013 Apr 22.

ShortStack: comprehensive annotation and quantification of small RNA genes.

Axtell MJ[1].

Description de Shorstack

Script perl qui nécessite :

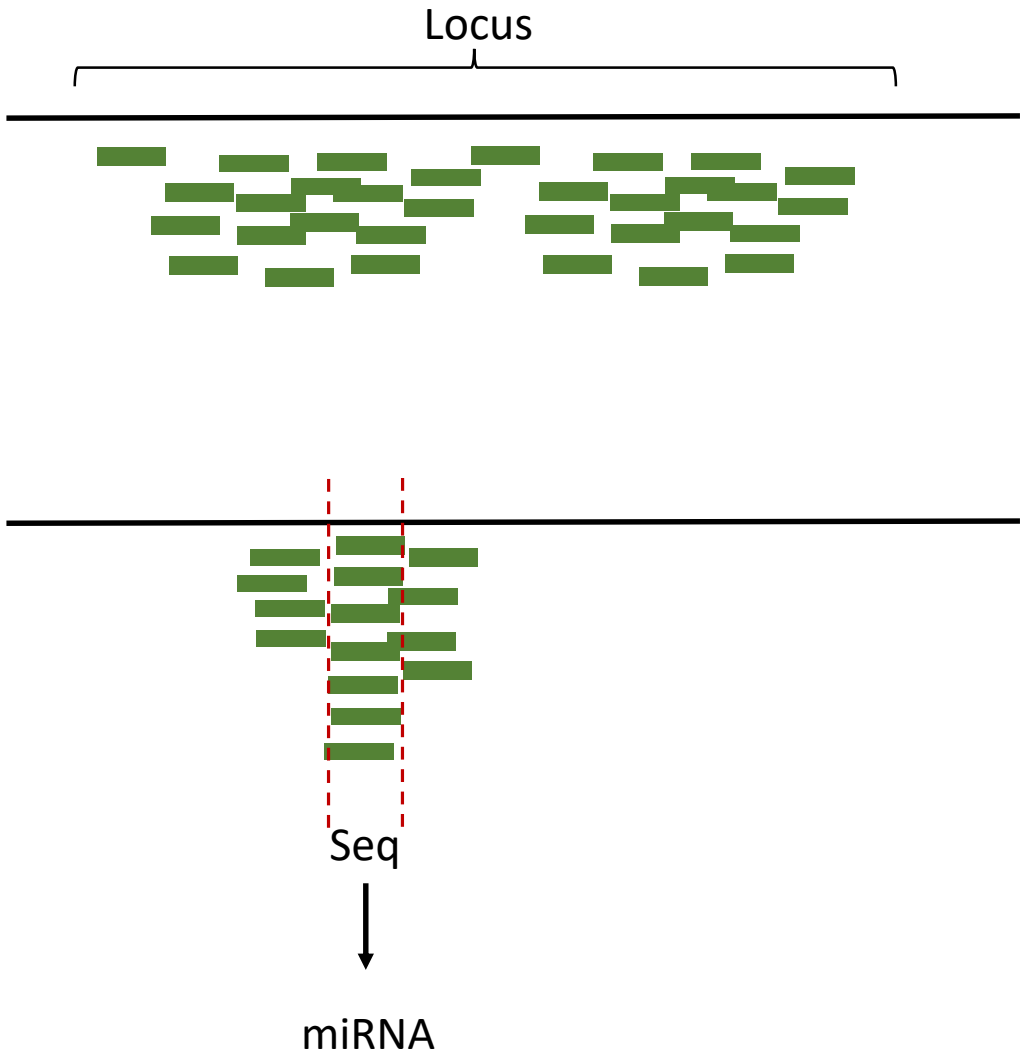
- Samtools
- RNALfold (Vienna RNA Package)
- RNAeval (Vienna RNA Package)

Utilisation Bowtie pour le mapping (peut être utilisé directement dans Shortstack) :

- Optimiser pour les reads <50 nt
- Supporte les données Solid.

Ce défini comme étant très stringent pour la détermination des miRNAs

Fonctionnement de Shorstack



Identification de « ilot « -> Accumulation des reads

- Niveau de profondeur défini
- Extension de ces ilots aux ilots voisins (variation des smallRNAs le long des précurseurs)

Création de Locus

- Comptage aux niveaux des locus de la zone la plus couverte : définition de la séquence la plus représenté Seq .

- Si 20 <Seq< 24 (plante) ou 25 (animal)

Analyse par Rnafold -> Hairpin => miRNA

Utilisation de Shorstack

1/Analyse de chaques réplicas indépendamment.

2/Elimination de la redondance entre les banques / conservation seulement des locus présent dans 2 réplicas

3/Relance Shortstack mais en lui donnant la liste des locus à déterminer

Fichier de sortie de Shorstack

Fichier de résultats générale :

#Locus Name Length Reads RPM UniqueReads FracTop Strand
MajorRNA MajorRNAReads Complexity DicerCall
MIRNA PhaseScore Short Long 20 21 22 23 24

Fichier de résultats pour chaque MiRNA :

Cluster_613 Original Location: Minc3s00023:218025-218100 Displayed Location: Minc3s00023:218006-218119 Strand: +

UACCAAAAGUUUUGUUUAUCAGGAUUGGUAUACAUAGGUAAUUUUGUCUUCUUUAAAAAAAAUUUUUUUUUAAUUACCAUGUGAACCCAUUCUGAUUUUAUAAGACUUUUAAUUA

)))..((((((((((..(((((((((.(((.((((((.((((((...........(((((((....))))))))))))))))))).))).)))))))...))))))))))))(((((

.....UUACCAUGUGAACCCAUUCUGA..... miRNA l=22 a=1978

.....AGGAUUGGUAUACAUAGGUAAUU..... miRNA-star l=23 a=63

Résultats de Shorstack

Solanum lycopersicum

Mapping :
~ 90% de lectures mappés
~ 30% mapper de façon multiple

locus prédit : 183632
miRNA : 111

M. incognita

Mapping :
~ 90% de lectures mappés
~ 70% mapper de façon multiple

locus prédit : 20210
miRNA : 91

Cohérents avec les résultats des analyses réalisées par Axtel, beaucoup plus de prédiction sur les plantes.

Outils bioinformatique utilisés et dédiés au miRNA

Nucleic Acids Res. 2012 Jan;40(1):37-52. doi: 10.1093/nar/gkr688. Epub 2011 Sep 12.

miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.

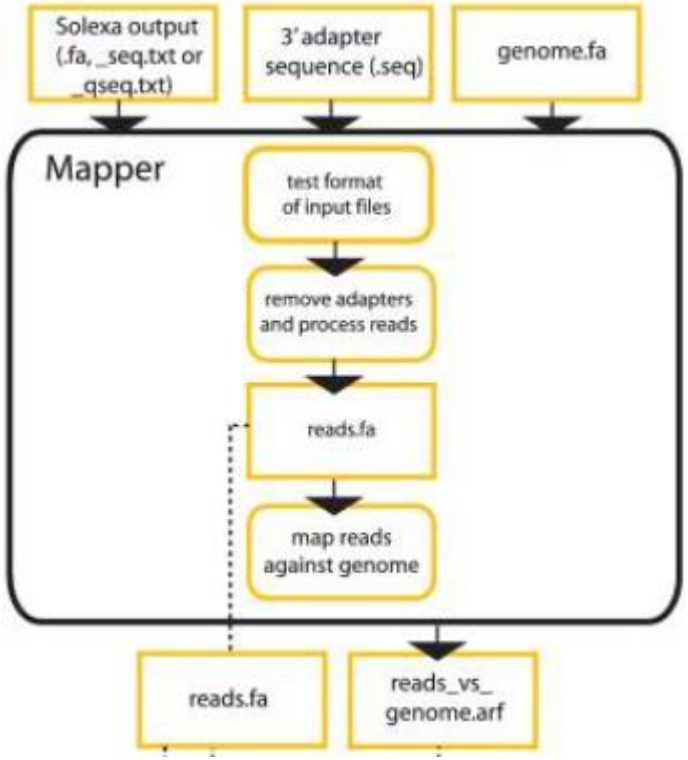
Friedländer MR[1], Mackowiak SD, Li N, Chen W, Rajewsky N.

Bioinformatics. 2017 Aug 15;33(16):2446-2454. doi: 10.1093/bioinformatics/btx210.

miRCat2: accurate prediction of plant and animal microRNAs from next-generation sequencing datasets.

Paicu C[1,2], Mohorianu I[2,3], Stocks M[2], Xu P[3], Coince A[3], Billmeier M[3], Dalmay T[3], Moulton V[2], Moxon S[3].

Fonctionnement de Mirdeep2 : Module mapper



- Suppression des adaptateurs si nécessaire
- Les reads avec une séquence identique sont réduites pour supprimer la redondance. (Un nombre dans les nouvelles entêtes des fasta indique combien de fois la séquence est présente)
- Mapping avec par défaut 0 mismatch autorisé sur une graine de 18 nucléotides
- Fichier de sortie .arf et reads.fa

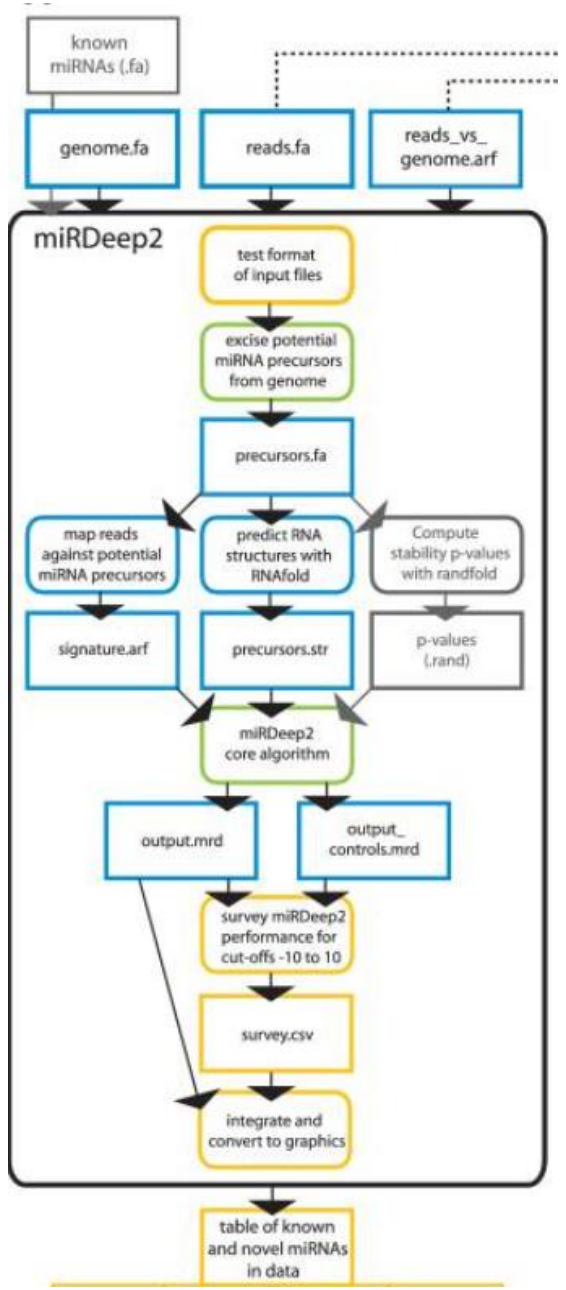
Fonctionnement de Mirdeep2

Données d'entrées :

- Fichier de sortie .arf et reads.fa
- Génome
- miRNA connu

Etapes :

- Si les reads sont positionnés plus de 5 dans ce cas rejeté.
- Ne conserve que les alignements n'ayant aucun mismatch sur 18 nucléotides.
- Les zones contenant des reads mappés sont considérés comme précurseur potentiel et analysé avec RNAfold
- Recherche si les séquences de miRNAprécurseurs connues se retrouvent sur le génome.
- Les miRNAs matures connus sont aussi positionnés sur le génome
- Statistique de performance de l'annotation est déterminé à partir des annotations déjà connus



Adaptation mirDeep2 au plante

BMC Genomics. 2011 Feb 16;12:108. doi: 10.1186/1471-2164-12-108.

Characterization of statistical features for plant microRNA prediction.

Thakur V[1], Wanchana S, Xu M, Bruskiewich R, Quick WP, Mosig A, Zhu XG.

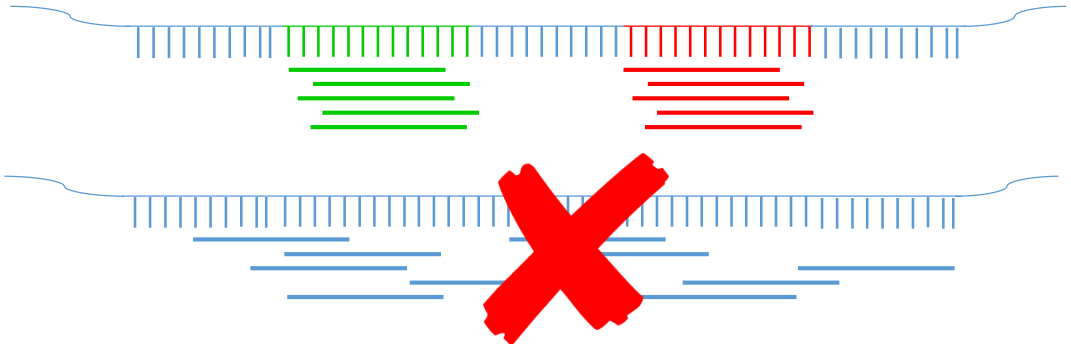
Feature		Original parameters of miRDeep	Plant specific (monocot)	Dicot specific (if any)
MFE	Known	Cumulative Distribution Function: $F(x) = e^{-e(x-location)/scale}$ Location = 32; Scale = 5.5	Log-odds score: $f(x)=a(b+ex*c)$	Log-odds score:$f(x)=a(b+ex*c)$
	Background	Cumulative Distribution Function: $F(x) = e^{-e(x-location)/scale}$ Location = 23; Scale = 4.8	a = 1.339e-12 b = 2.778e-13 c = 45.843	a = 4.46e-4 b = 9.125e-5 c = 26.929
Stability (log-odds)	Stable	1.6	1.37	0.63
	Unstable	-2.2	-3.624	-3.17
Nucleus conservation (log-odds)	Conserved	3	7.63	
	Non-conserved	-0.6	-1.17	
Excision length		140 nt	300 nt	
Paired	Total	≥14	≥15 nt	
Unpaired	Total	NA	≤5 nt	
	Consecutive unpaired	NA	≤3 nt	
Bulge	Total	could be as high as 5 nt #	≤2 nt	
Maximum multiple hits of deep-seq read		5	20	

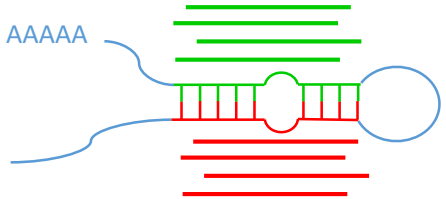
Table 1
List of miRDeep parameters estimated for plant specific miRNAs.

Modification du programme pour que celui corresponde au dicotylédone

miRCat2 : critères d'identification des microARN

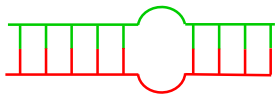
- Sélection de zones candidates (accumulation d'alignements de lectures)





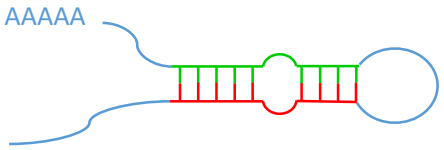
Utilisation de fenêtre coulissante

- Nombre maximal de positions différentes où s'alignent les lectures
- Taille des produits de clivage : *dicer call*
  - *18-25nt pour les animaux*
  - *20-24nt pour les plantes*



Variable en fonction du monde animal ou végétal

- Calcul de l'énergie de repliement



miRCat2 :

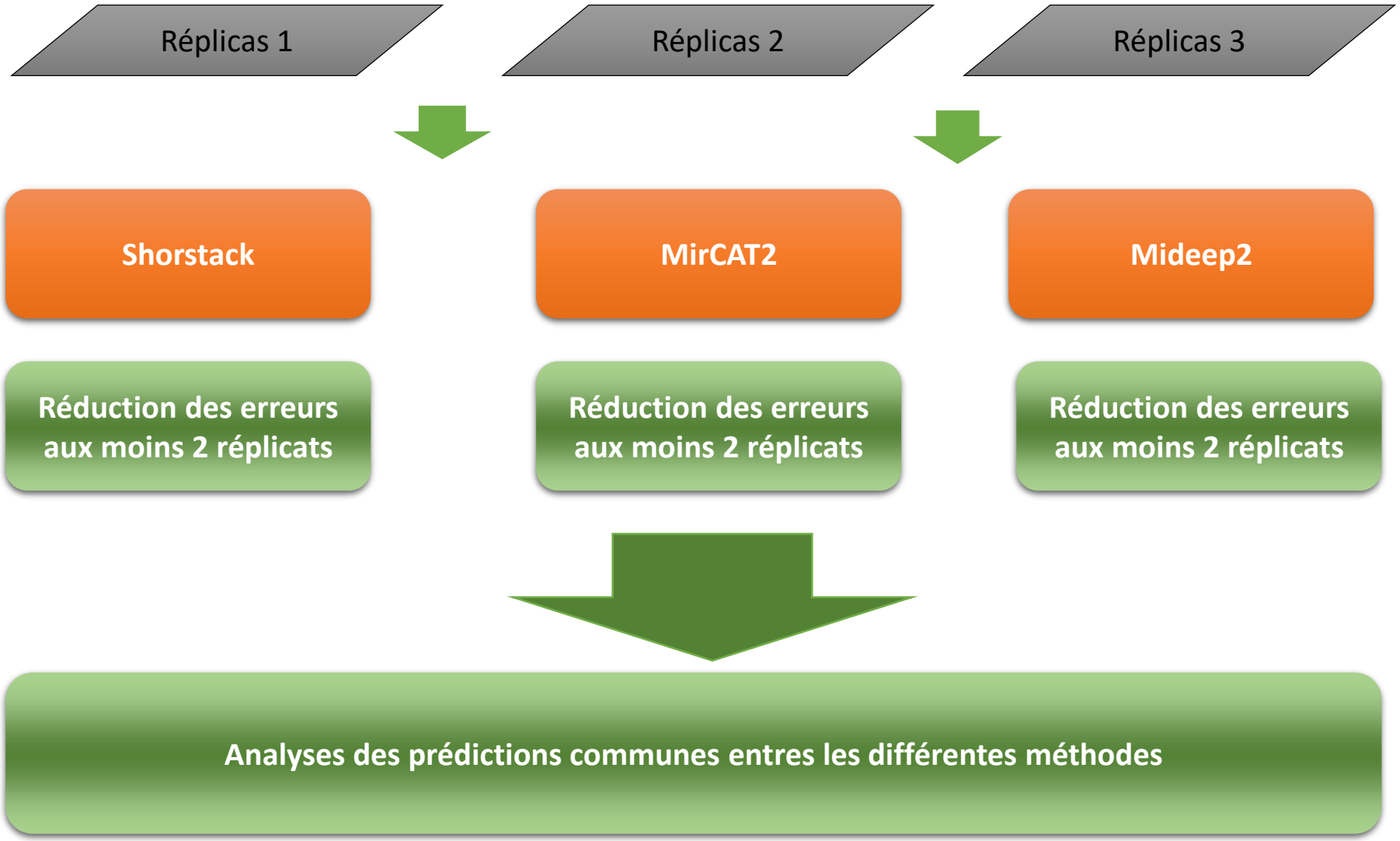
Le UEA sRNA Workbench est disponible ici:

http://srna-workbench.cmp.uea.ac.uk.

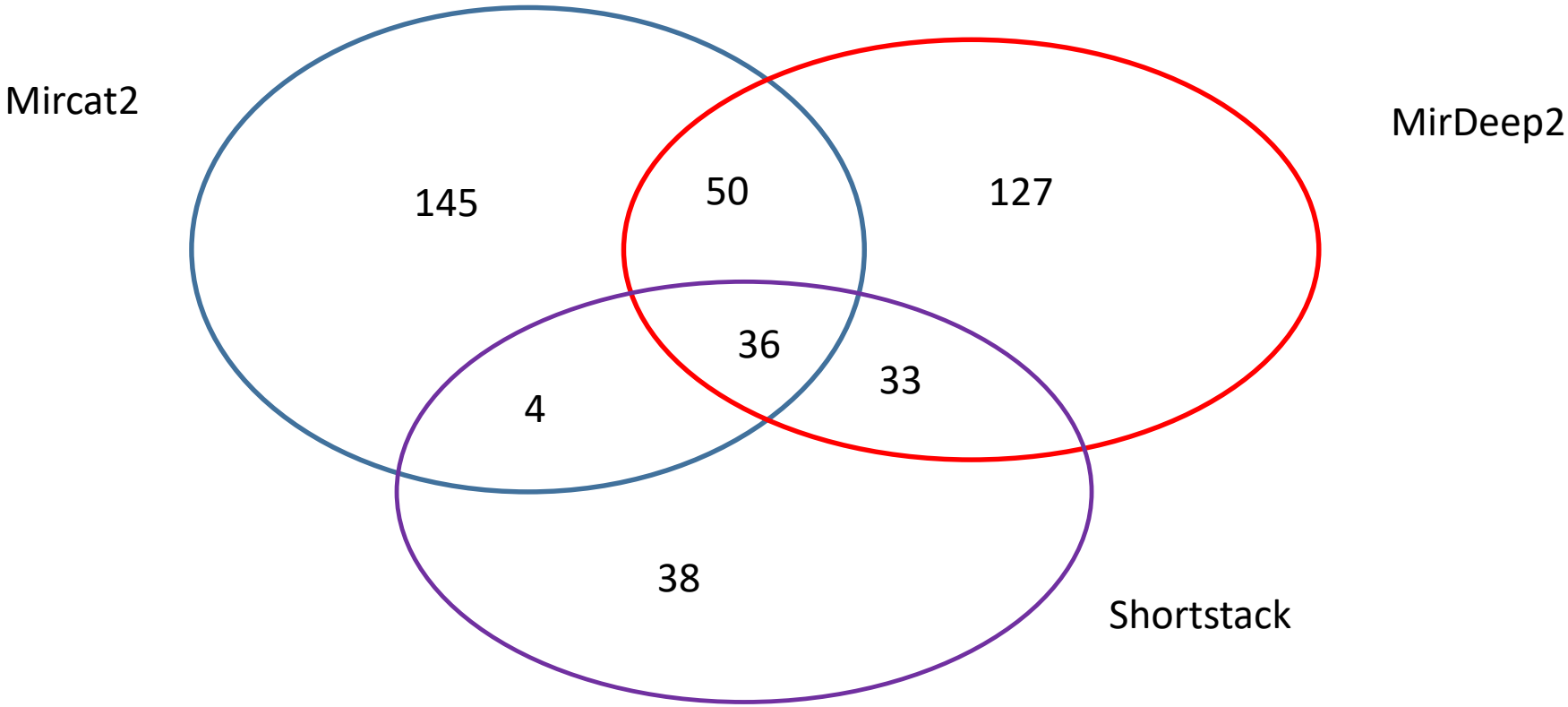
Le code is available at: https://github.com/sRNAworkbenchuea/UEA_sRNA_Workbench.

- Fichier de parametrage en fonction des caractéristiques des plantes /animals (format json)
- Galére d'utilisation via l'interface du Workbench.

Pipeline d’analyse miRNA



Résultats : Croisement des différents prédictions sur la tomate



Algorithme en cours de réécriture car particulier non pris en compte :
Notamment inversion miRNA /Star miRNA

Base de données sur les smallRNA

- www.mirbase.org/ : Base de données dédiés au miRNA pour l'ensemble des espèces -> dernière mise à jour en Mars 2018 (après 4 ans sans mise à jours).
- DASHR – Database of small human noncoding RNAs (www.lisanwanglab.org/DASHR/smdb.php);
- Base de données dédiées au small RNA chez le mais et le riz (htpp://sundarlab.ucdavis.edu/smrnas/)
- Rfam base de données dédié au ARN non spécifique des smallRNAs

Stéphanie Jaubert-Possamai

Anais Vacquier

Virginie MAGNONE
Nicolas Pons
Kevin LEBRIGAND