UNIVERSITÉ DE STRASBOURG

École Doctorale
des Sciences de la Vie
et de la Santé
STRASBOURG

THÈSE

présentée par : Anaïs Vittu

soutenue le : 1 décembre 2015

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline / Spécialité : Sciences de la Vie et de la Santé

Bioinformatique et biologie des systèmes

Outils bioinformatiques pour l'analyse génétique de la résistance du moustique *Anopheles gambiae* vis-à-vis des parasites du paludisme

THÈSE dirigée par :

BLANDIN Stéphanie — Chargé de recherche, Université de Strasbourg

RAPPORTEURS EXTERNES :

COLINGE Jacques — Professeur, Institut de Recherche en Cancérologie de Montpellier

SEITZ Hervé — Chargé de recherche, Institut de Génétique Humaine de Montpellier

EXAMINATEUR INTERNE :

IMLER Jean-Luc — Professeur, Université de Strasbourg

Remerciements

Je tiens à remercier tout d'abord Stéphanie Blandin, ma directrice de thèse, pour m'avoir encadrée durant ces quatre années. Nos discussions m'ont beaucoup apporté, professionnellement et personnellement.

Je tiens également à remercier les membres de mon jury, Jacques Colinge, Jean-Luc Imler et Hervé Seitz pour avoir accepter de juger mon travail de thèse.

Je souhaite aussi remercier Jean-Marc Reichhart, le directeur de l'équipe RIDI, pour m'avoir accueillie au sein de son laboratoire pour mon stage et ma thèse.

Je remercie vivement Laurent Troxler pour tous les échanges bioinformatiques que nous ayons eu. Ils ont été très constructifs et essentiels. Merci pour tes encouragements et ta présence.

Je remercie aussi l'équipe de l'HPC, David Brusson, Romaric David et Michel Ringenbach pour leur soutien, toujours rapide et efficace, dans l'utilisation du cluster.

Je souhaite aussi remercier Marie-Céline, Laurence, Émilie et Laurent pour toutes les discussions à la pause de midi et pour la bonne ambiance qu'il y avait au labo grâce à vous.

Je remercie toute ma famille pour m'avoir soutenue et encouragée pendant ces années.
Merci à ma mère et mon père pour tout et rien.
Merci à ma marraine d'être toujours là.
Merci à mes trois petites sœurs qui font de moi une grande sœur fière.

Merci à Vincent qui m'a supporté tout ce temps avec mes hauts et mes bas.
Merci à ma fille, Hanaé, d'être là et de tout embellir.

Abréviations

ADN	Acide DésoxyriboNucléique
ADNc	ADN complémentaire
ADNg	ADN génomique
ARN	Acide RiboNucléique
b	base
BAC	Bacterial Artificial Clone
CDC	Centers for Disease Control and prevention
dsRNA	double stranded RNA (ARN double brin)
FP	Faux Positif
GP	Golden Path
indels	insertion/délétion
kb	kilobase
Mb	Mégabase
MO	Micro-Organisme
Mo	Mégaoctets
NGS	Next Generation Sequencing
NTS	Nouvelles Technologies de Séquençage
OMS	Organisation Mondiale de la Santé
pb	paire de bases
PE	Paired-End
PEST	Pink Eye STandard
piRNA	piwi RNA (petits ARNs associés à piwi)
QTL	Quantitative Trait Loci
rasRNAi	reciprocal allele-specific RNA interference
siRNA	small interfering RNA (petits ARNs interférents)
SNP	Single Nucleotide Polymorphism
TEP	ThioEster-containing Protein
To	Téraoctets
VIH	Virus de l'Immunodéficience Humaine
VP	Vrai Positif
WGS	Whole Genome Sequencing

Sommaire

Introduction

Introduction

Sommaire

1. Le paludisme, un problème de santé mondial

D'après l'Organisation Mondiale de la Santé (OMS), le paludisme, appelé aussi malaria, est un fléau qui sévit dans 97 pays. Son contrôle et son élimination font parti des grandes préoccupations de la santé mondiale. En effet, 3,2 milliards de personnes vivent ou voyagent dans des pays endémiques. En 2013, 198 millions de cas de paludisme et 584 000 décès associés ont été estimés. Toutefois, ces données sont probablement sous-estimées en raison de l'incertitude des données récoltées sur le terrain (Fig. 1).

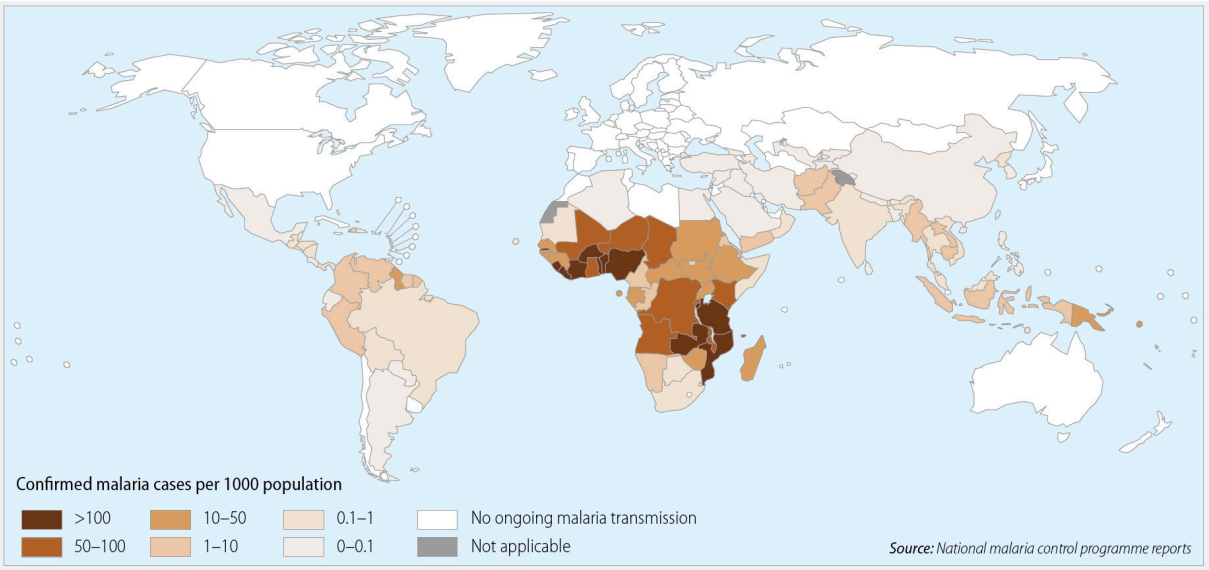


Figure 1. Nombre de cas de paludisme déclarés pour 1 000 personnes. Source : Rapport de l'OMS 2014, p.2

La zone la plus meurtrière se situe en Afrique où 90% des décès ont lieu dont 78% sont des enfants de moins de cinq ans.

1.1. Trois acteurs en jeu

Le responsable du paludisme est un parasite protozoaire du genre Plasmodium. Il existe cinq espèces de Plasmodium liées à la maladie chez l'homme : *Plasmodium falciparum*, *Plasmodium knowlesi*, *Plasmodium vivax*, *Plasmodium malariae* et *Plasmodium ovale*.
P. falciparum et *P. knowlesi* sont responsables des formes les plus sévères de la maladie.
Sa transmission à l'homme s'effectue par les femelles moustiques du genre Anopheles. Seules les femelles prennent des repas sanguins pour assurer la ponte des œufs. Parmi les 430 espèces d'anophèles présentes dans le monde entier, entre 30 à 40 sont vectrices du paludisme.

1.1.1. Le parasite

Le cycle de vie du parasite comporte deux phases : une phase sexuée chez un moustique appelé vecteur et une phase asexuée chez un humain appelé hôte (Fig. 2.)

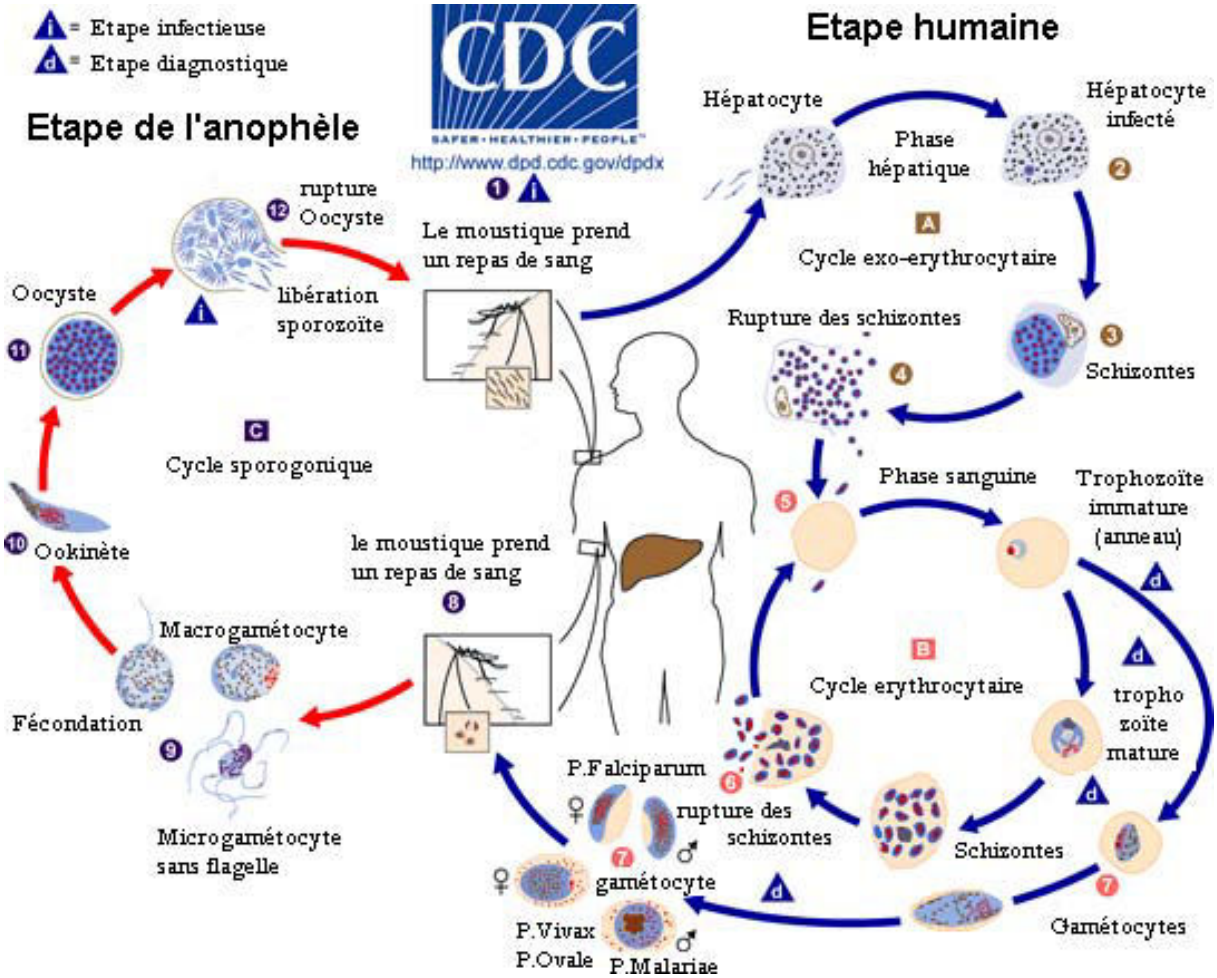


Figure 2. Cycle de vie des parasites responsables du paludisme. A droite, l'étape humaine est composée de deux cycles : le premier exo-érythrocytaire (A) et le deuxième érythrocytaire (B). Au cours de son repas sanguin, la femelle anophèle infectée inocule les parasites à l'homme en injectant de la salive (1). Ces formes injectées, les sporozoïtes, sont transportés via la circulation sanguine vers le foie où ils envahissent les hépatocytes (2). Ils se multiplient et forment des schizontes qui vont faire éclater les cellules hépatiques (4). Les mérozoïtes libérés entrent dans la circulation sanguine et intègrent les globules rouges (5). Une multiplication des parasites s'ensuit alors dans les érythrocytes par séries répétitives d'invasion, de croissance et de division (6). Cette multiplication peut entraîner l'infection de 10^{12} ou plus d'hématies, conduisant à la maladie et aux complications du paludisme. Au cours de ce parasitisme sanguin, qui peut durer des mois s'il n'est pas traité, certains stades érythrocytaires vont quitter les cycles de multiplication asexuée et se développer sur une période d'environ 2 semaines en gamétocytes mâles et femelles matures (7). Les gamétocytes sont les seules formes infectieuses pour le moustique. Ils peuvent être ingérés par un moustique femelle lors d'un repas sanguin (8).
La phase de développement du parasite chez l'anophèle est dite sporogonique (C). La fécondation des gamétocytes a lieu dans l'intestin et produit des zygotes (9). Ces derniers se différencient en ookinètes capables de traverser la paroi intestinale et de s'y loger sous forme d'oocystes (10, 11). La croissance et la division de chaque oocyste produit des milliers de formes haploïdes appelées sporozoïtes. Ces derniers sont libérés 8-15 jours après le repas sanguin selon les espèces et vont envahir les glandes salivaires (12). Le cycle de vie du parasite est terminé lorsque ces sporozoïtes seront transmis de nouveau à l'homme lors d'une piqûre du moustique femelle infectée (1). La transmission peut ainsi continuer. Source : www.cdc.gov

1.1.2. Le moustique vecteur

Les moustiques sont des insectes de l'ordre des Diptères et de la famille des Culicidés.
Ils sont caractérisés par une tête, un thorax et un abdomen, une paire d'ailes, 3 paires de pattes, une trompe en forme de seringue et une paire d'antennes en forme de V. On reconnaît un mâle à ses antennes plumeuses et touffues (Fig. 3).

Même si les moustiques présentent des différences d'habitat, de comportement et des périodes de développement distinctes, ils ont tous en commun un cycle de développement similaire (Fig. 3) avec 2 périodes :

- une phase aquatique, qui est celle du développement larvaire et nymphal
- une phase aérienne, pendant laquelle l'adulte se reproduit

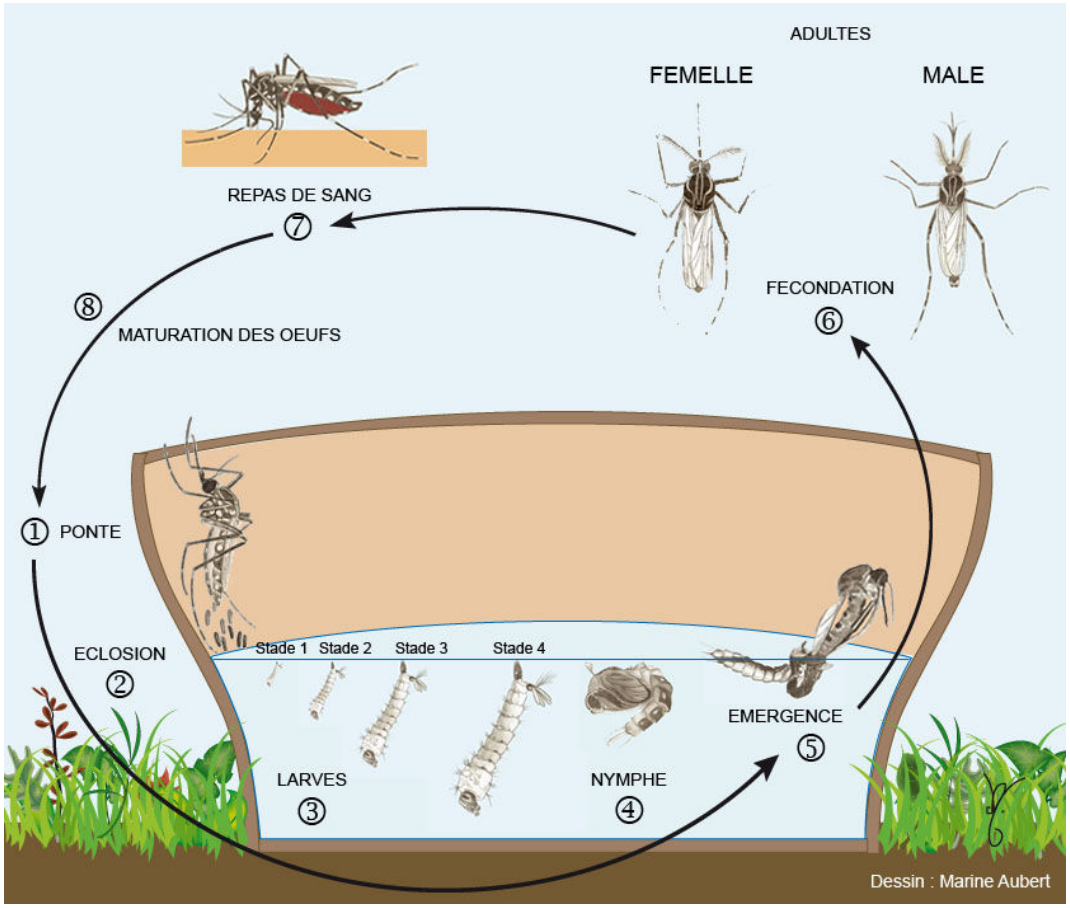


Figure 3. Cycle de vie du vecteur du paludisme : le moustique. La femelle pond entre 50 et 400 œufs un par un sur une surface liquide stagnante (1). Elle peut pondre entre 200 à 1 000 œufs durant sa vie. Si les conditions climatiques le permettent, les œufs éclosent (2)(3). Les stades larvaires sont les seuls stades de croissance au cours desquels le moustique passe de 1-2 mm à 8-12 mm en muant. La larve respire en se collant à la surface de l'eau et s'alimente de microplanctons, d'algues microscopiques et de particules en suspension. Après plusieurs jours, toujours si les conditions climatiques le permettent, la quatrième mue donne naissance à la nymphe (4). Ce stade éphémère dure entre 24 à 48h. La nymphe ne s'alimente plus, la transformation physique commence. L'émergence est le moment où le moustique adulte sort de la mue nymphale (5). Source : Institut Pasteur de Nouvelle-Calédonie (http://www.institutpasteur.nc/les-moustiques-et-la-dengue/)

Les moustiques mâles et femelles sont nectarivores, c'est-à-dire qu'ils s'alimentent de nectar et de jus sucrés des fleurs et des fruits pour couvrir leurs besoins énergétiques. En fait, seule la femelle a besoin d'un repas sanguin pour produire ses œufs. Les femelles anophèles piquent préférentiellement entre le crépuscule et l'aube. Elles sont dites exophiles ou endophiles lorsqu'elles se nourrissent à l'extérieur ou à l'intérieur des maisons.
La durée de vie du moustique va de quelques jours à 1 mois et son rayon d'action est, en moyenne, de 300m mais peut s'étendre jusqu'à 15 km pour certaines espèces.

Parmi les 3 500 espèces de moustiques existantes, celles qui se nourrissent de sang humain sont dites anthropophiles. Parmi elles, on retrouve *Anopheles gambiae*, *A. arabiensis*, *A. funestus*, *A. nili* et *A. moucheti*, les principaux vecteurs du paludisme en Afrique (Fig. 4).

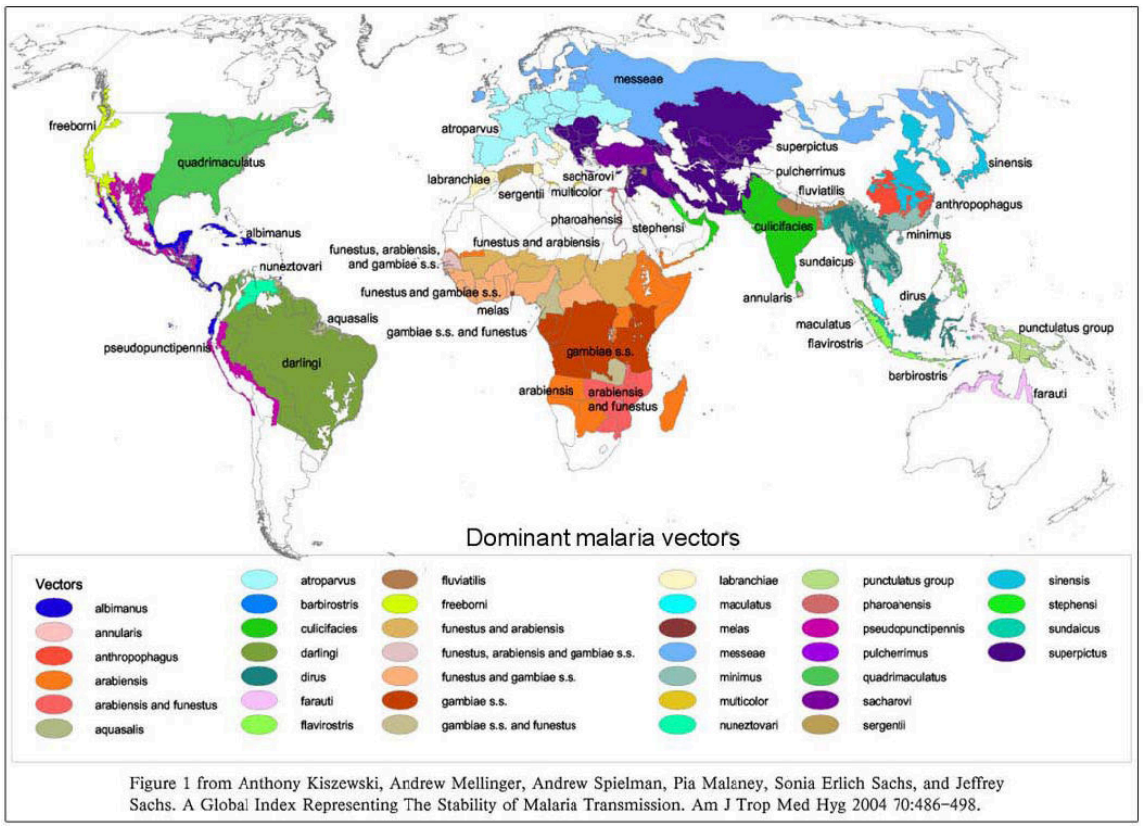


Figure 1 from Anthony Kiszewski, Andrew Mellinger, Andrew Spielman, Pia Malaney, Sonia Erlich Sachs, and Jeffrey Sachs. A Global Index Representing The Stability of Malaria Transmission. Am J Trop Med Hyg 2004 70:486–498.

Figure 4. Carte de la distribution des principaux vecteurs du paludisme dans le monde. Source : OMS

En Afrique, ce sont les espèces du complexe *Anopheles gambiae* qui occupent le plus grand territoire et qui ont la plus forte capacité de transmission (Fig. 4). On parle ici de complexe puisque le complexe *Anopheles gambiae* regroupe plusieurs espèces identiques d'un point de vue morphologique mais différentes sur le plan génétique et biologique. Les divergences observées concernent les milieux fréquentés, la distribution, la reproduction, les préférences trophiques et par conséquent les différences d'efficacité vectorielle (Coluzzi et al., 1979). A l'origine, la distinction des espèces de ce complexe s'est faite par l'étude cytogénétique des chromosomes polytènes. Il est aussi possible de les distinguer par différentes techniques (PCR, marqueurs moléculaires, etc.).
Le moustique compte deux paires de chromosomes autosomes numérotés 2 et 3 et une paire de chromosomes sexuels nommé X (Fig. 5). Les bras de part et d'autre du centromère sont désignés par les lettres R et L pour respectivement right et left.

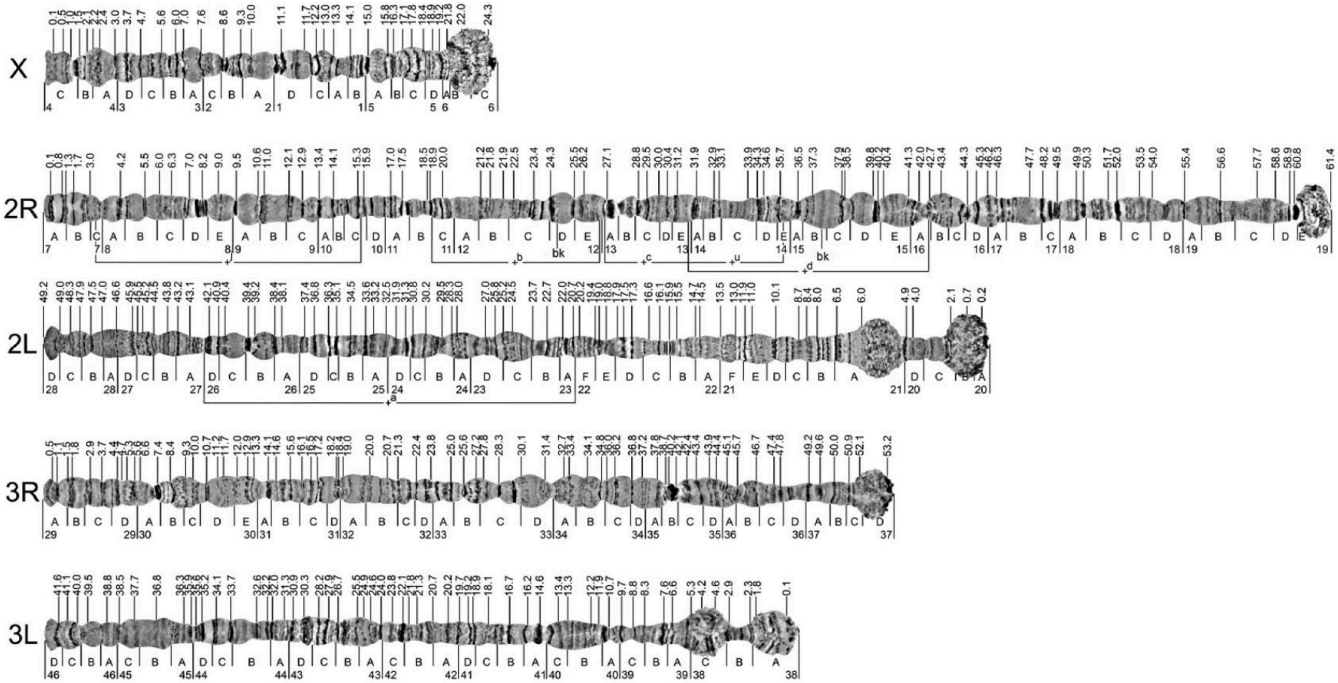


Figure 5. Cartographie cytogénétique d'*Anopheles gambiae*. Les coordonnées sont spécifiées au-dessus des bras des chromosomes et les divisions et sous-divisions se situent sous les bras des chromosomes. La localisation des plus fréquentes inversions chromosomiques est indiquée sous les divisions (il y a 5 inversions paracentriques sur le bras 2R : 2R*j*, *b*, *c*, *d* et *u* et une inversion sur le bras 2L : 2L*a*. Source : (George et al., 2010)

Les espèces du complexe peuvent être distinguées les unes des autres par l'identification d'inversions chromosomiques (rotation à 180° de portions de chromosome). Les inversions paracentriques (situées de part et d'autre du centromère) sont désignées par des lettres minuscules. L'arrangement standard est noté « + ». Par exemple : 2R+ désigne l'arrangement standard du bras droit du chromosome 2 ; 2L*a* désigne l'inversion *a* du bras gauche du chromosome 2.

La combinaison de ces inversions fixées permet de distinguer 6 espèces différentes :

- *Anopheles gambiae sensu stricto* (s.s.)
- *Anopheles bwambae*
- *Anopheles arabiensis*
- *Anopheles melas*
- *Anopheles merus*
- *Anopheles quadriannulatus*

Par la suite, *Anopheles gambiae s.s.* a donné deux formes moléculaires distinctes M et S (Wondji et al., 2002) différentiables par l'arrangement de séquences d'ADN ribosomiques situées sur le chromosome X (Gentile et al., 2001; Torre et al., 2001). Récemment, les deux formes ont été élevées au rang d'espèces et ont été nommées *Anopheles gambiae* et *Anopheles coluzzii,* respectivement (Coetzee et al., 2013).

1.1.3. L'hôte humain

Environ la moitié de la population mondiale est exposée au paludisme. La majorité des cas et des décès surviennent en Afrique subsaharienne. Cependant, l'Asie, l'Amérique latine, le Moyen-Orient et certaines parties de l'Europe sont également affectés.

Les groupes de population les plus spécifiquement à risque sont :

- Les jeunes enfants vivant dans des zones de transmission stable qui n'ont pas encore développé une immunité les protégeant contre les formes les plus sévères de la maladie.
- Les femmes enceintes chez qui le paludisme entraîne des taux élevés de fausses couches, des faibles poids de naissance et peut provoquer des décès maternels, en particulier lors des première et seconde grossesses.
- Les personnes vivant avec le VIH/sida.
- Les voyageurs internationaux en provenance de régions exemptes de paludisme car ils ne sont pas immunisés.
- Les immigrants venus de régions d'endémie et leurs enfants qui vivent dans des zones exemptes de paludisme et qui retournent dans leur pays d'origine. L'immunité anti-paludique nécessite une exposition continue pour être maintenue.

Le paludisme est une maladie caractérisée par des épisodes fébriles aigus. Les symptômes apparaissent généralement 10 à 15 jours après la piqûre de moustique infectante. Les premiers symptômes – fièvre, maux de tête, frissons et vomissements – peuvent être modérés et difficiles à attribuer au paludisme. Les enfants fortement atteints développent fréquemment un ou plusieurs des symptômes suivants : anémie sévère, détresse respiratoire consécutive à une acidose métabolique ou paludisme cérébral. Chez l'adulte, on observe aussi fréquemment une atteinte de tous les organes.
S'il n'est pas traité dans les 24 heures, le paludisme à *P. falciparum* peut évoluer vers une affection sévère souvent mortelle.
Pour les paludismes à *P. vivax* et à *P. ovale*, des rechutes cliniques peuvent se produire des semaines ou des mois après la première infection. Certaines personnes peuvent vivre avec les parasites sans en développer aucun symptôme. Néanmoins, ils peuvent toujours transmettre les parasites lors de piqûres par les femelles moustiques.

1.2. La transmission

Plusieurs facteurs entrent en jeu dans la transmission du paludisme : (1) le parasite, (2) le vecteur, (3) l'hôte et (4) l'environnement.

(1) Le parasite : selon le type de parasite, la durée de certaines phases de maturation diffère. Par exemple, le cycle érythrocytaire, chez l'homme, varie de 1 jour pour *P. knowlesi*, à 2 jours pour *P. falciparum, P. vivax* et *P. ovale* ou 3 jours pour *P. malariae*. Le cycle sporogonique, chez le moustique, dure de 7 à 10 jours. Cela rend donc le moustique plus vite infectieux ou accélère l'apparition et le degré de virulence des symptômes chez l'homme. On sait que la majorité des décès est due à *P. falciparum* et que *P. vivax* et *P. ovale* ont une phase hépatique dormante et constitue donc un réservoir potentiel infectieux.

(2) Le vecteur : les espèces d'anophèles les plus dangereuses sont celles qui ont un degré d'anthropophilie fort, et un cycle de vie assez long pour que le parasite puisse se développer et arriver aux glandes salivaires. De plus, un moustique vivant plus longtemps transmettra la maladie à un plus grand nombre de personnes. Le cycle de vie du moustique varie de 7 à 20 jours selon les espèces et l'environnement.

(3) L'hôte : les facteurs génétiques rendent plus ou moins sensibles les individus. L'immunité développée par la personne entre aussi en compte.

(4) L'environnement : les localités où les populations n'ont pas ou peu accès à des centres de santé, et donc aux messages de prévention, aux diagnostics et aux traitements, sont souvent celles à forte transmission. Les conditions climatiques telles que le taux des précipitations, l'humidité et la température influent sur l'abondance et la survie des moustiques vecteurs. Les premières étapes de la vie des moustiques (œufs, larves, pupes) sont aquatiques. C'est pour cela qu'il y a souvent une recrudescence des cas de paludisme lors de la saison des pluies. La température est aussi très importante pour le parasite et le moustique puisqu'elle affecte directement la durée de leur cycle de vie.

1.3. Les stratégies de contrôle et d'élimination

Depuis plusieurs décennies, l'OMS a mis en place des stratégies de lutte anti-vectorielle sur le terrain afin de réduire la transmission du paludisme. Parmi elles, les insecticides et les vaccins antipaludiques ont été les plus efficaces.

Les deux principales méthodes de prévention sont l'utilisation de moustiquaires imprégnées d'insecticide longue durée et la pulvérisation d'insecticide sur les murs des maisons.
En Afrique, les 41 pays appartenant à la région OMS avaient adopté la politique qui consiste en l'approvisionnement gratuit de moustiquaires traitées à l'insecticide pour les enfants et les femmes enceintes. Alors qu'en 2006, seuls 16 de ces pays envisageaient la possibilité d'étendre cette mesure à l'ensemble de la population à risque, ils sont maintenant 38.
Aujourd'hui, près de la moitié de la population a accès aux moustiquaires imprégnées et un peu plus de 40% dorment dessous. Cette amélioration considérable a fait chuter l'incidence du paludisme dans certains pays.
Concernant les pulvérisations intra-domiciliaires, seuls 15 pays d'Afrique bénéficient de cette intervention pour la prévention et le contrôle de l'épidémie malgré le fait qu'elles soient recommandées dans 42 pays d'Afrique. L'OMS conseille de changer annuellement d'insecticide pour préserver leur efficacité mais cette mesure est difficile à mettre en pratique.

La proportion de la population en Afrique sub-saharienne protégée par au moins une méthode de contrôle du vecteur a considérablement augmenté depuis les années 2000 (Fig. 6).

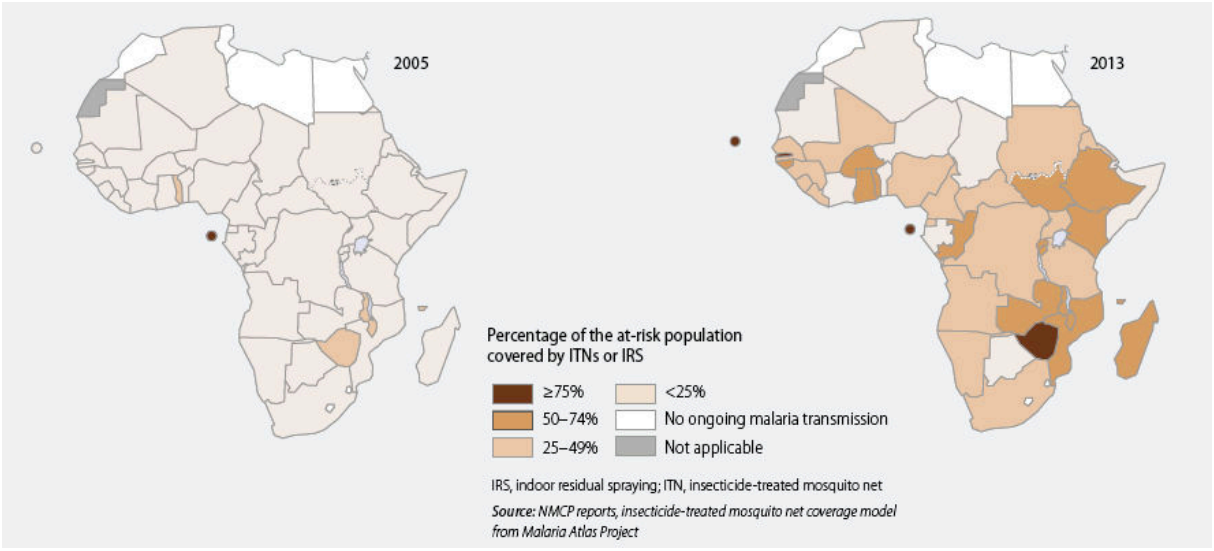


Figure 6. Pourcentage de la population à risque protégée par des moustiquaires imprégnées d'insecticides ou par des pulvérisations d'insecticides intra-domiciliaires, 2005 et 2013. Source : Rapport de l'OMS 2014, p.15

D'autres moyens de lutte sont aussi utilisés seuls ou combinés :

- répulsifs pour la peau,
- vêtements imprégnés,
- réductions des gîtes larvaires,
- lutte anti-larvaire (larvicides).

Pour l'homme, l'OMS recommande :

- l'administration de traitements antipaludiques combinés chez les personnes dont les tests de laboratoire ont révélé la présence des parasites,
- l'utilisation obligatoire de thérapies de combinaison de médicaments antipaludiques efficaces à un composé dérivé de l'artémisinine, surtout dans le cas avéré d'infection à *P. falciparum,*
- l'utilisation de chloroquine pour les cas lié à *P. vivax*, et dans le cas d'une résistance, l'utilisation des thérapies combinées,
- l'utilisation d'une dose unique de primaquine lors d'une suspicion de résistance à l'artémisinine,
- les femmes enceintes doivent recevoir durant leur grossesse un traitement antipaludique préventif et intermittent.

Ces dernières années, l'OMS a noté une forte augmentation de l'utilisation des traitements antipaludiques combinés, dû en majorité à l'augmentation de la production et de l'accès à ces traitements. Malgré cela, la proportion d'enfants de moins de 5 ans recevant un traitement reste plutôt faible (< 20%).

Même s'il y a encore beaucoup de chemin à parcourir, le déploiement d'interventions de lutte anti-vectorielle a permis une importante réduction de la morbidité et de la mortalité dues au paludisme en Afrique. L'OMS a enregistré en 2015 une diminution de 60% de la mortalité au niveau mondial par rapport à 2000 et de 54% dans la Région africaine de l'OMS. En Afrique, le taux de mortalité des enfants a diminué de 58% par rapport à 2000.

1.4. Les résistances

En dépit de cet important recul de la mortalité, de plus en plus de pays signalent des résistances : (1) celle des parasites aux antipaludiques et (2) celle des moustiques aux insecticides.

(1) En 2004, la combinaison thérapeutique à base d'artémisinine (ACT) a été mise en place après l'émergence d'une pharmacorésistance aux traitements chimiques tels que la chloroquine et la sulfadoxine-pyrimethamine (Yeung et al., 2004). Cependant, les parasites du paludisme montrent une grande aptitude à développer une résistance aux médicaments et ces dernières années, certains pays d'Amérique du sud et d'Asie du sud-est dont principalement le Cambodge, ont rapporté des résistances à l'artémisinine (World Malaria Report 2014).

(2) Cela concerne plusieurs espèces de moustiques et toutes les classes d'insecticides. De récents rapports indiquent que certains pays d'Afrique montrent une forte hausse de la résistance aux pyréthroïdes tels que la deltaméthrine. Or, il n'existe à ce jour que peu d'insecticides alternatifs à la fois efficaces et peu coûteux. En 2012, l'OMS publie son Plan mondial pour la gestion de la résistance aux insecticides. Elle invite les pays concernés ainsi que les parties prenantes (milieux universitaires et industriels) à prendre des mesures immédiates pour préserver l'efficacité des méthodes actuelles et à développer de nouveaux insecticides dans les meilleurs délais possibles. Il a été aussi observé la modification des habitudes alimentaires de *A. funestus* au Bénin : face à l'introduction des insecticides, *A. funestus* qui piquait préférentiellement la nuit dans les maisons a changé de comportement pour piquer aux premières lueurs du jour à l'extérieur des habitations (Moiroux et al., 2012).

Si les ACT et les moustiquaires imprégnées restent à ce jour les meilleurs outils de lutte contre ce fléau, l'émergence et l'augmentation de ces résistances ont poussé la communauté scientifique à développer de nouvelles méthodes d'éradication.

En plus de la recherche de nouveaux antipaludiques et de nouveaux insecticides, on trouve parmi les nouveaux moyens de lutte (1) le développement de vaccins, (2) l'utilisation des micro-organismes (MO) contre les moustiques et les parasites et (3) la sélection ou la modification génétique de moustiques.

(1) Le développement de vaccins antipaludiques n'aboutit toujours pas en dépit de nombreuses décennies de recherches intensives. La complexité des plasmodies ralentit considérablement la recherche d'un vaccin efficace. Plus de vingt projets de vaccins sont actuellement en cours d'évaluation dans des essais cliniques ou sont en phase de développement préclinique avancé. Actuellement, les meilleurs montrent une diminution de la prévalence des formes sévères du paludisme. Dans le cas où un vaccin serait mis sur le marché, il sera utilisé en supplémentation et non en remplacement des mesures préventives et thérapeutiques existantes. En effet, le vaccin a lui seul ne peut pas éradiquer le paludisme.

(2) Dans la nature, un ensemble d'ennemis naturels contribuent à la régulation des populations de moustiques. Certains de ces organismes sont produits et utilisés à grande échelle dans plusieurs pays pour limiter le nombre de moustiques : les bactéries *Bacillus thuringiensis* var. *israelensis* et *Bacillus sphaericus,* le nématode entomopathogène, *Romanomermis iyengari*, et le poisson larvivore *Gambusia affinis* (Abagli et al., 2014). De plus, certaines bactéries présentes dans l'organisme du moustique ont une action négative sur le développement des parasites du paludisme permettant de limiter ainsi la transmission des parasites à l'homme (Hughes et al., 2011).

(3) La dissémination de moustiques modifiés génétiquement ou non génétiquement dans l'environnement est en passe d'ouvrir un nouveau front dans la guerre contre les maladies à transmission vectorielle telles que le paludisme. Il y a deux types de méthodes à ne pas confondre :
a. La première est la suppression de la population selon la technique de l'insecte stérile. Le but est d'introduire des insectes irradiés en laboratoire qui engendreront une progéniture non viable. Plusieurs études sur l'irradiation de moustiques ont été réalisés en laboratoire (Helinski et al., 2006, 2008; Oliva et al., 2012) et elles pointent souvent du doigt le problème de la compétition qu'il y aura dans la nature entre des moustiques mâles normaux et des moustiques mâles irradiés pour se reproduire avec des femelles (Baker et al., 1979; Reisen et al., 1980). Les moustiques de laboratoire, de surcroît irradiés, montrent des signes de faiblesse physique (Huho et al., 2007).
b. La deuxième est la méthode de remplacement de la population. Plus complexe techniquement, cette dernière consiste à modifier l'insecte pour qu'il ne puisse plus transmettre le parasite. Elle suppose l'accouplement d'insectes génétiquement modifiés avec la population sauvage et la modification permanente de son patrimoine génétique. Il est alors possible de remplacer les moustiques sensibles aux parasites du paludisme par des moustiques résistants.

2. Facteurs non génétiques du moustique Anopheles gambiae contrôlant le développement du parasite du paludisme

Les facteurs environnementaux comme la composition du microbiote intestinal d'un moustique peuvent modifier sa capacité à transmettre les parasites du paludisme. Le terme microbiote définie les communautés microbiennes qui vivent en contact direct avec l'épithélium de l'organisme hôte. Elles sont composées de bactéries, virus, champignons et protistes.
Le microbiote intestinal du moustique a récemment émergé comme étant un facteur important de la résistance des moustiques aux pathogènes. En particulier, les bactéries intestinales ont montré qu'elles avaient un impact négatif conséquent sur la charge parasitaire paludique à travers des mécanismes de colonisation impliquant soit des interactions directes bactéries-parasites du paludisme, soit via l'activation des réponses immunitaires par les bactéries ciblant aussi les parasites.

Ces interactions se font de différentes manières :

- Les bactéries commensales intestinales stimulent la réponse immunitaire antibactérienne ce qui contrôle la prolifération des bactéries aussi bien que celle des populations des stades ookinètes et oocystes du parasite *Plasmodium* (Dong et al., 2009; Frolet et al., 2006).

- Les molécules sécrétées par les bactéries, telles que les intermédiaires d'oxygène réactif ou les métabolites secondaires, peuvent bloquer ou limiter le développement des parasites dans les intestins (Cirimotich et al., 2011a).
- La charge bactérienne augmente fortement après l'ingestion d'un repas sanguin limitant l'accès des parasites à l'épithélium intestinal (Cirimotich et al., 2011b).
- La présence et l'abondance de certaines bactéries sont significativement liés à la charge parasitaire intestinale (Boissière et al., 2012).

L'utilisation des bactéries pour contrôler le développement du parasite dans le moustique et donc limiter sa transmission à l'homme est un concept relativement nouveau et qui mérite d'être approfondi. La bactérie *Wolbachia* a déjà montré ses preuves ; certaines lignées réduisent la durée de vie du moustique et empêchent le pathogène de terminer son cycle, d'autres sont capables d'inhiber les parasites de *P. falciparum* directement dans l'intestin (Hughes et al., 2011; McMeniman et al., 2009).
Par contre, le problème de la recherche de bactéries utilisables comme stratégie de contrôle du paludisme, c'est que la composition en bactéries des moustiques dépend fortement de son lieu de vie et varie en fonction des différents stades de sa vie (Boissière et al., 2012; Gimonneau et al., 2014).

Bien que les bactéries soient les plus abondantes au sein du microbiote intestinal, il est nécessaire de prendre en compte les virus, champignons et protistes dans la recherche de l'origine non génétique de la résistance d'un moustique aux parasites.

3. Facteurs génétiques du moustique Anopheles gambiae contrôlant le développement du parasite du paludisme

Étant donné le manque de vaccins efficaces et les augmentations des résistances du parasite aux médicaments et du moustique aux insecticides, le développement de nouvelles stratégies de contrôle est crucial pour réduire considérablement la transmission du paludisme (Cirimotich et al., 2011c). Les modifications génétiques constituent alors un nouvel espoir dans la lutte anti-vectorielle. Le moustique, vecteur, devient alors un outil dont on se doit d'approfondir les connaissances génétiques pour en maîtriser la manipulation.

3.1. Moustiques résistants et moustiques sensibles

Les moustiques *Anopheles gambiae* ne sont pas tous sensibles à une infection aux parasites du paludisme. Chez certains moustiques, la réponse antiparasitaire est particulièrement efficace. Elle bloque complètement le développement du parasite à un stade précoce de l'infection, ie dans l'intestin juste après l'ingestion des parasites (Fig. 7 et 8). Ces moustiques ne transmettent donc pas la maladie et sont nommés résistants (Blandin et al., 2009). Il a été observé que les moustiques résistants présentaient deux moyens de destruction des parasites morts : la mélanisation et la lyse. Leurs homologues dont la réponse immunitaire n'est pas apte à la destruction totale des parasites sont nommés sensibles. Des lignées résistantes ou sensibles peuvent être sélectionnées et isolées en laboratoire suggérant qu'il existe des facteurs génétiques qui contrôle l'efficacité de la réponse antiparasitaire du moustique (Collins et al., 1986; Vernick et al., 1995).

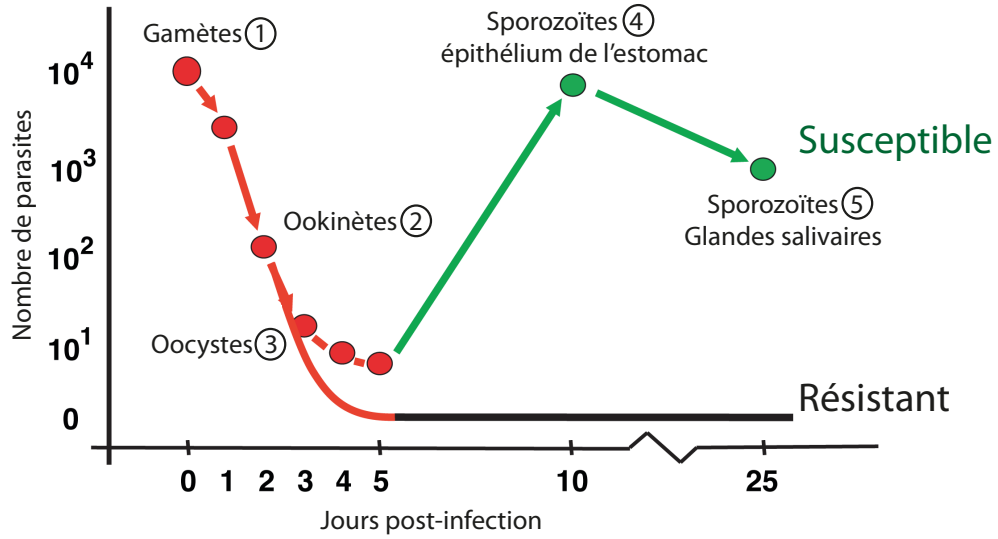


Figure 7. Évolution du nombre de parasites dans les lignées résistantes et sensibles après l'ingestion des gamètes. Les lignées résistantes sont capables d'éliminer les stades (2) et (3) des parasites directement dans l'intestin. La réponse immunitaire des lignées sensibles n'élimine pas tous les parasites et certains oocystes libèrent alors leurs sporozoïtes (4). La seconde perte de parasites est principalement due aux réponses immunitaires dans les autres parties du moustique, lorsque les sporozoïtes se dirigent vers les glandes salivaires et y restent (5). Source : S. Blandin, modifié.

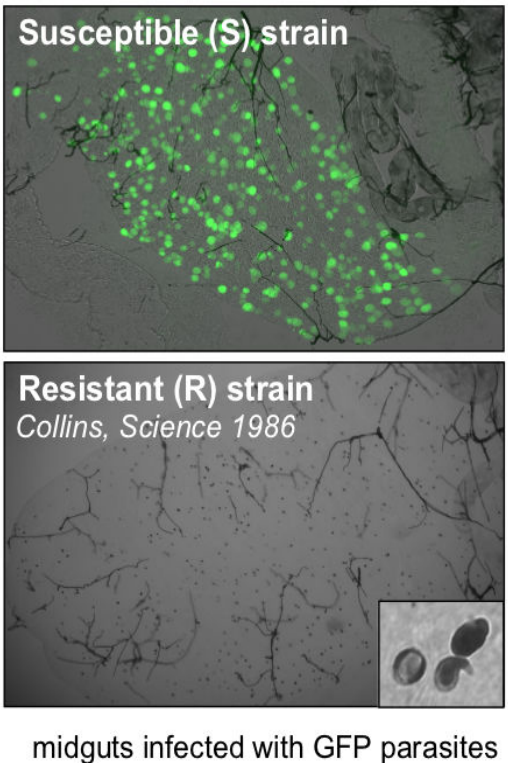


Figure 8. Intestins de moustiques résistants et sensibles. Les parasites vivants expriment la GFP ce qui les rend vert fluorescent et donc visibles au microscope à fluorescence. En dessous, les points noirs représentent les parasites morts et qui ont été mélanisé. Source : (Blandin et al., 2009)

3.2. Identification des facteurs génétiques

Notre laboratoire s'est donc lancé dans l'investigation des bases génétiques qui sous-tendent la résistance des moustiques au parasite du paludisme. L'objectif est d'identifier les régions génomiques et les polymorphismes qui expliquent la destruction complète des parasites.
Pour ce faire, ils ont croisé des moustiques résistants avec des moustiques sensibles et ont obtenu une seconde génération comprenant divers degrés de résistance (Fig. 9).

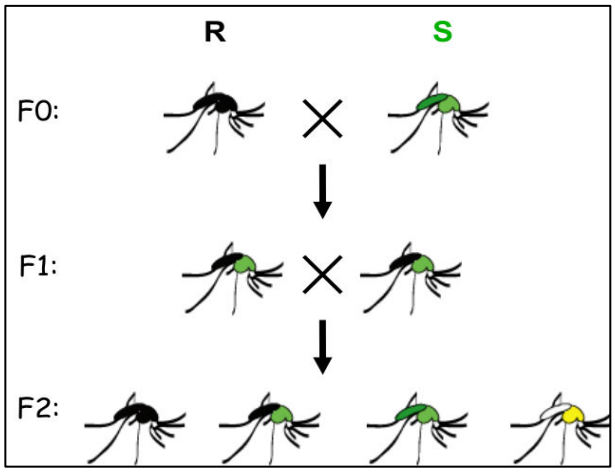


Figure 9. Croisement de lignées résistantes et lignées sensibles et générations produites. Les couleurs représentent les multiples phénotypes observés. Source : (Blandin et al., 2009)

Ces divers degrés de résistance dans la génération F2 indique que la destruction des parasites et leur mélanisation sont des caractères complexes qui sont la conséquence de l'interaction de plusieurs gènes.

3.2.1 QTL mapping

Afin d'identifier les régions associées à la résistance aux parasites, ils ont utilisé une stratégie classique en génétique : la cartographie de QTL. Elle a été réalisée à partir des phénotypes extrêmes de la génération F2 infecté par le parasite murin *Plasmodium berghei* (Fig. 10).

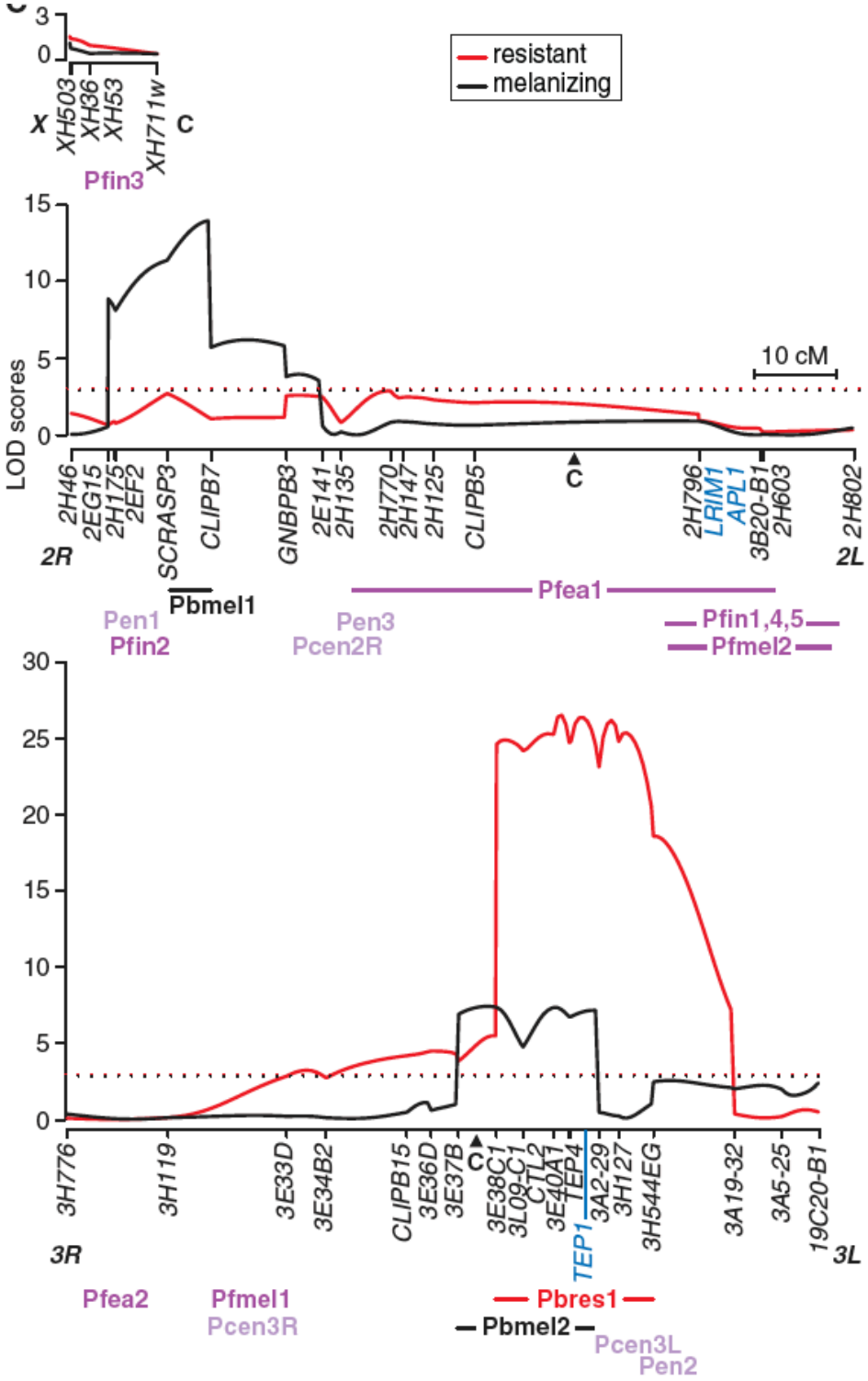


Figure 10. Cartographie de QTL pour les phénotypes de résistance (en rouge) et de mélanisation (en noir). Les seuils de LOD sont indiqués en pointillés horizontaux (ils sont à 3 et 2.88, respectivement). La position du centromère (**C**), les marqueurs génétiques, les bras des chromosomes et les gènes LRIM1, APL1 et TEP1 (en bleu) sont spécifiés sous l'axe des x. Les QTLs précédemment identifiés sont ensuite indiqués en violet clair pour la résistance au parasite simien et en violet foncé pour la résistance au parasite *P. falciparum*. Source : (Blandin et al., 2009).

L'analyse des QTLs liés à la résistance a révélé une région de 19Mb sur le chromosome 3L associée à la résistance à *P. berghei* (Fig. 10). Ce locus a été nommé *Pbres1* pour Plasmodium berghei resistance locus 1.
Dans cette large région qui contient environ 975 gènes, nous en repérons 35 ayant un domaine jouant un rôle dans la fonction immunitaire. Parmi eux se trouve un gène connu (Blandin et al., 2004) codant pour une protéine de type complément : TEP1 (ThioEster-containing Protein 1). C'est une protéine homologue au facteur du complément C3 (protéine faisant partie du système immunitaire inné chez l'homme). Cette molécule sécrétée dans l'hémolymphe a une activité antiparasitaire clé : elle se lie aux parasites et promeut leur destruction. De plus, en étudiant ce gène dans les différentes lignées d'*A. gambiae*, il est apparu qu'il possède plusieurs variants alléliques (Fig. 11). Son rôle antiparasitaire et son polymorphisme font de ce gène un candidat idéal à l'explication des divers degrés de résistance observés.

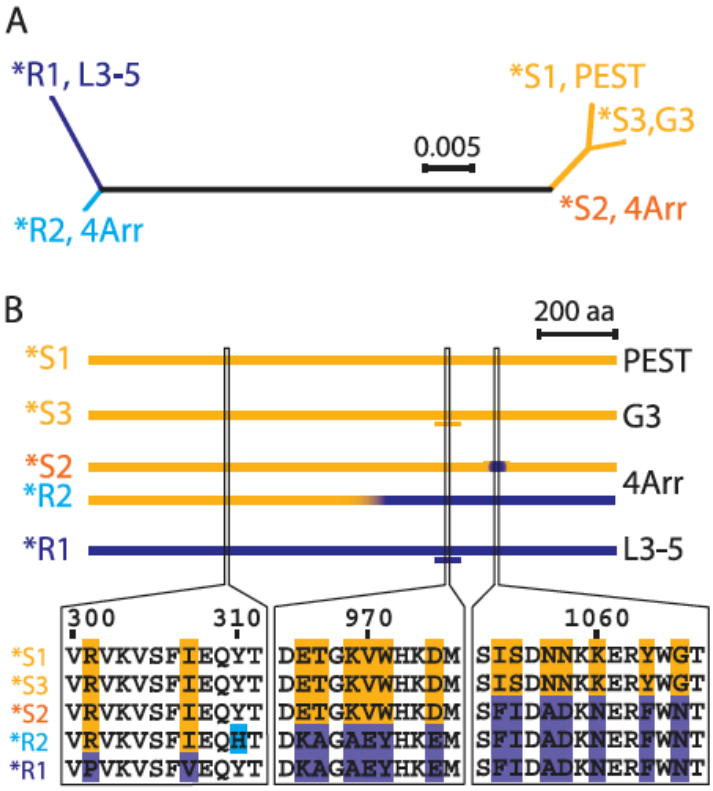


Figure 11. Étude du polymorphisme du gène TEP1. (A) Arbre phylogénétique créé à partir de l'alignement global des séquences d'acides aminés complètes des allèles de TEP1 provenant des lignées L3-5 (*R1), 4Arr (*R2 et *S2), PEST (*S1) et G3 (*S3). L'échelle indique les substitutions d'acides aminés estimés par site. (B) Représentation schématique des séquences d'acides aminés des allèles de TEP1. Les séquences de *S1 et *S3 sont représentées par des barres orange, et *R1 par une barre bleue. Les allèles *S2 et *R2 sont des combinaisons des allèles *S1/*S3 et *R1. Les petites barres horizontales sous *S3 et *R1 indiquent la région ciblé par les dsRNAs dsRa et dsSa. Source : (Blandin et al., 2009)

Cependant, l'étude de la contribution des différents allèles de TEP1 au phénotype n'était pas possible puisque ces allèles sont dans des lignées différentes et que l'efficacité d'un allèle peut dépendre du fond génétique dans lequel il se trouve. Il est alors nécessaire de comparer les allèles dans un même fond génétique, donc dans un même moustique.

3.2.2. Reciprocal allele-specific RNA interference

Pour résoudre ce problème, l'équipe a mis en place la stratégie suivante : (1) ils ont effectué des croisements entre des moustiques résistants (L3-5) et des moustiques sensibles (G3) pour obtenir une lignée contenant un allèle de chaque parent, (2) puis ils ont inhibé spécifiquement chaque allèle par l'injection de sondes dsRNAs allèle-spécifiques.
Il ont conçu, pour réaliser le point (2), une nouvelle approche appelée « reciprocal allele-specific RNAi » (rasRNAi) basé sur le principe de l'ARN interférence (Encadré 1) et qui permet d'évaluer la contribution des différents allèles d'un même gène pour un caractère donné dans un même fond génétique (Fig. 12). Cette technique repose sur l'utilisation d'ARNs double brin (dsRNAs) longs spécifiques de chaque allèle et donc qui ne doivent pas contenir des fragments > 19 paires de bases identiques entre les deux allèles. Cette dernière condition pose alors problème dans le cas des gènes peu polymorphiques.

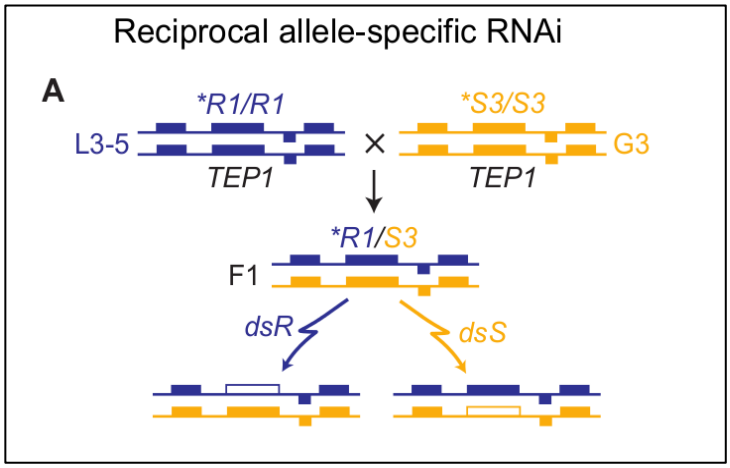


Figure 12. Principe de la méthode « reciprocal allele-specific RNAi » ou rasRNAi. Le croisement d'une lignée L3-5 résistante avec une lignée G3 sensible (chacune homozygote pour le gène TEP1) permet d'obtenir une lignée contenant les deux allèles : résistant *R et sensible *S. L'injection d'un dsRNA ciblant l'un des allèles permet d'étudier l'action de l'allèle non ciblé sur la destruction des parasites. Source : (Blandin et al., 2009)

Encadré 1. L'ARN interférence (ARNi ou RNAi)

L'ARN interférence (« RNA interference ») ou ARNi est un processus naturel que les cellules utilisent pour bloquer l'expression d'un gène. La première découverte de ce phénomène fut en 1990 par des scientifiques qui souhaitaient intensifier la couleur de pétunias. Mais étonnamment, en introduisant un gène responsable de la couleur, ils supprimèrent l'expression de ce gène et obtinrent ainsi des zones ou des pétunias entiers sans pigmentation (van der Krol et al., 1990a, 1990b). Quelques années plus tard, une étude sur le nématode *Caenorhabditis elegans* a montré que l'élément clé qui provoque ce phénomène de suppression est l'ARN double brin (« dsRNA ») (Fire et al., 1998). L'ARN interférence est donc nommée et leurs travaux ont été récompensés par le Prix Nobel en 2006.
Dans le cytoplasme, les dsRNAs longs sont découpés par une enzyme de la famille des RNase III, Dicer, en petites séquences de 20-25 paires de bases (pb) appelées petits ARN interférents (« small interfering RNA or siRNA ») (Fig.1i). Ces siRNAs sont caractérisés par des extrémités 5' phosphorylées et par l'existence aux extrémités 3' de 2 bases non appariées. Une fois le siRNA incorporé dans le complexe RISC (« RNA-induced silencing complex »), un des brins devient le brin guide et s'ancre et alors que l'autre brin, le passager, va être clivé et expulsé du complexe. Ce sont les stabilités thermodynamiques relatives des régions 5' de chaque brin qui vont déterminer le brin guide du brin passager (Khvorova et al., 2003; Schwarz et al., 2003). Une faible stabilité en 5' promeut la sélection de ce brin en tant que guide par le complexe RISC. Un faible pourcentage en GC du duplex facilite la sélection par le complexe et le largage du passager. Le brin guide, s'il est complémentaire de l'ARN messager (ARNm), va permettre la fixation du complexe Ago2-RISC sur l'ARNm. Cette action entraine ensuite la destruction de l'ARNm et donc la suppression de l'expression du gène. L'origine des dsRNAs peut être endogène (expression de transgènes, produits de l'ARN polymérase ARN-dépendent, etc.) ou exogène (injection directe de dsRNAs, infection virales, etc.).
Il a été découvert, dans un second temps, une autre voie de régulation de l'expression des gènes via des microARNs (« miRNA ») endogènes (Fig. 1ii). Les précurseurs des miRNAs, les pri-miRNAs, ont une structure en forme d'épingle à cheveux. Les duplex miRNAs vont être départagés entre le complexe Ago1-RISC et le complexe Ago2-RISC de la voie des siRNAs en se basant sur leur complémentarité partielle (Ago1-RISC) ou parfaite (Ago2-RISC) (Fig. 1). La voie des miRNAs est essentielle puisqu'elle intervient dans de nombreux processus biologiques, chez les plantes et les animaux, qui vont des fonctions de bases vitales pour la cellule aux réponses au stress environnemental.
L'ARNi est un mécanisme conservé présent dans les trois règnes eucaryotes : Fungi, Plantae et Animalia. En plus du maintien de l'intégrité du génome, c'est un mécanisme de défense antiviral chez les plantes et les animaux.
Les fortes spécificité et efficacité de l'ARNi font de lui un outil idéal et, par conséquent, il est utilisé dans la génomique fonctionnelle (analyse des phénotypes perte-de-fonction après l'injection de dsRNA, utilisable sur une large variété d'organismes) et dans le développement de thérapies (traitement des infections virales, des désordres dominants, des désordres neurologiques et de certains types de cancers).

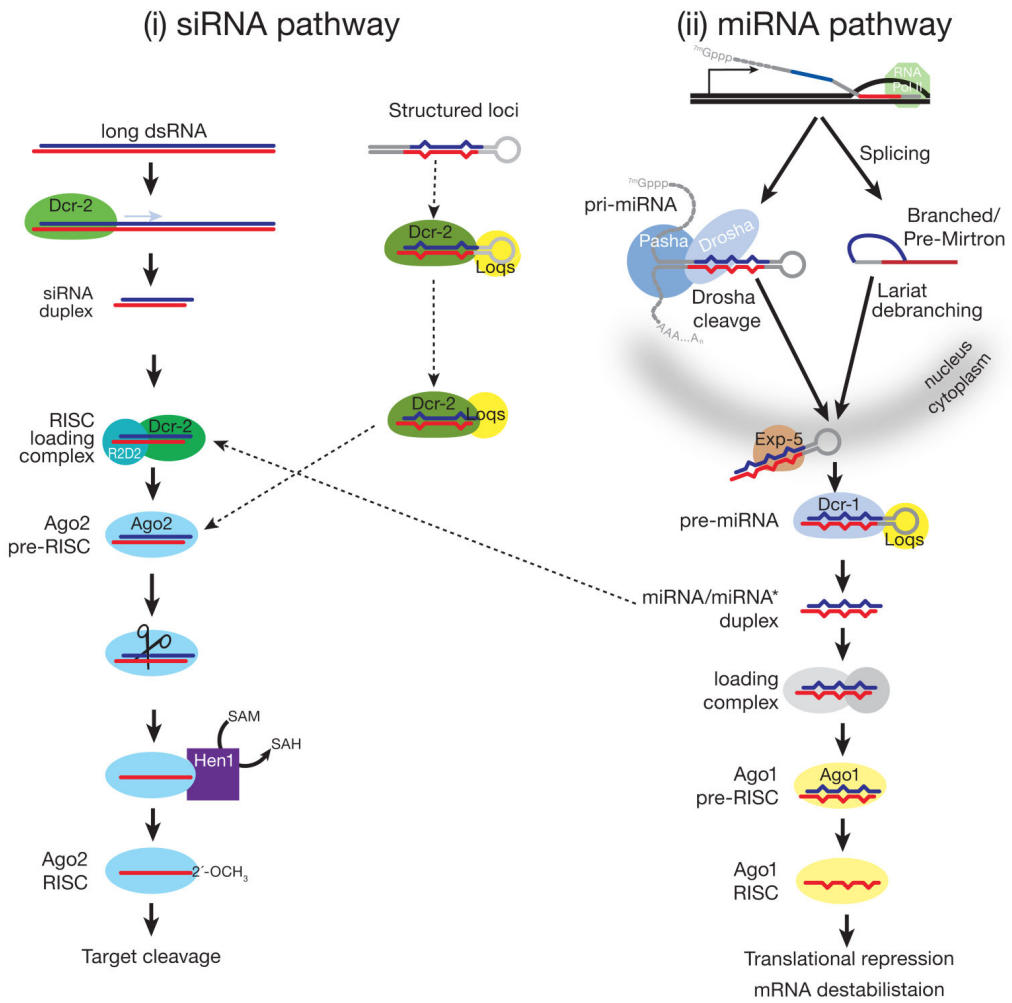


Figure 1. Voies des siRNAs et des miRNAs expliquant le phénomène de l'ARN interférence. Source : (Ghildiyal and Zamore, 2009)

Dans notre cas, deux dsRNAs ciblant l'allèle TEP1*R et l'allèle TEP1*S nous ont permis d'étudier leur action sur les parasites (Fig. 12 et 13). Un dsRNA (dsLacZ) ciblant un gène absent a servi de contrôle négatif et un dsRNA (dsTEP1) ciblant les deux allèles de TEP1 a servi de contrôle positif.

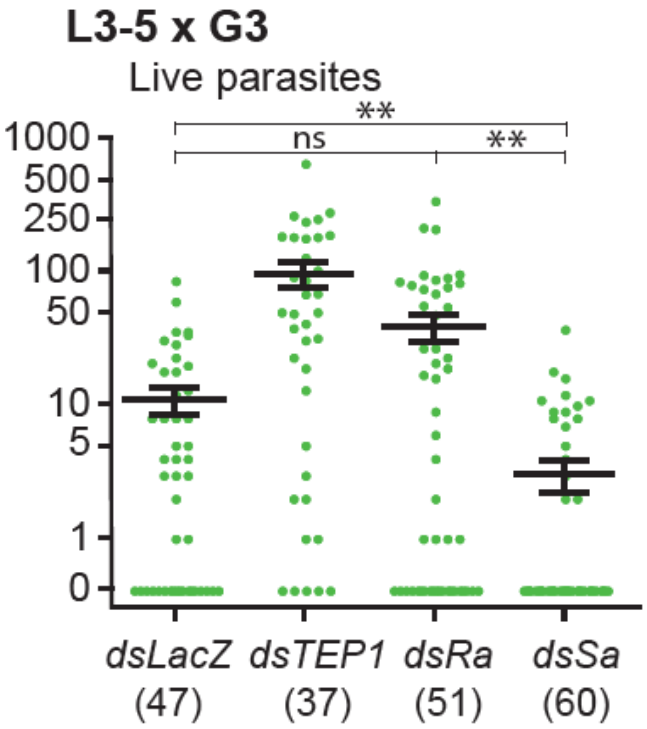


Figure 13. Nombre de parasites vivants comptabilisés par moustique (représenté par un point vert). dsLacZ est le contrôle négatif et dsTEP1 est le contrôle positif. dsRa cible l'allèle résistant TEP1*R et dsSa l'allèle sensible TEP1*S. Le nombre de moustiques disséqués pour chaque type de dsRNA injecté est placé sous le nom du dsRNA. Source : (Blandin et al., 2009)

L'inhibition indépendante de l'expression des allèles résistant et sensible de TEP1 nous a montré que le polymorphisme de TEP1 contrôlait l'efficacité de la réponse immunitaire au parasite *P. berghei* (Fig. 13). Quand l'allèle résistant (dsRa) était ciblé, le nombre de parasites a augmenté, ce qui signifie que la réponse immunitaire fut moins efficace. De plus, quand l'allèle sensible (dsSa) était ciblé, le nombre de parasites diminuait. L'allèle résistant est donc plus efficace pour détruire les parasites si son homologue sensible est absent. L'allèle *S est bien responsable d'une défaillance du système immunitaire contre les parasites *P. berghei*.

L'analyse des trois génération, F0, F1 et F2, du croisement L3-5 / G3 a révélé que les moustiques ayant les mêmes allèles de TEP1 entre la génération F1 et la F2 montraient un phénotype de résistance différent (Fig. 14). 50 à 70% des moustiques F2 avec les allèles *S2,*R1 sont résistants alors que seulement 7% des F1 avec les mêmes allèles le sont.

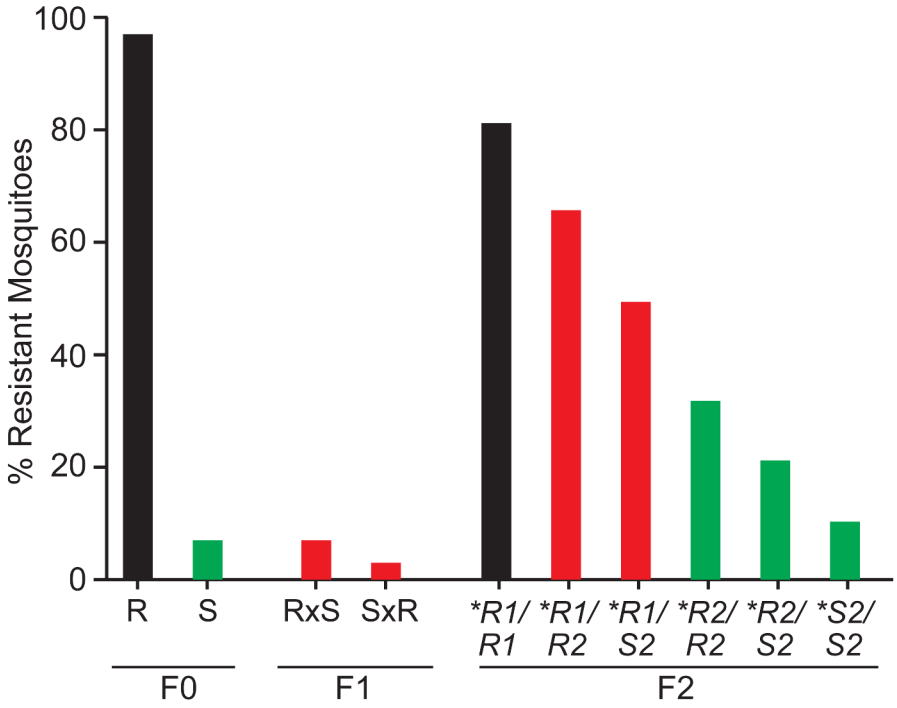


Figure 14. Croisement d'une lignée résistante L3-5 (TEP*R) avec une lignée sensible G3 (TEP1*S). Les couleurs représentent les différents phénotypes de résistance observés : de noir pour très résistant à vert pour très sensible. Source : S. Blandin

L'équipe est donc arrivée à la conclusion que d'autres facteurs que le gène TEP1 jouaient un rôle dans la diversité des phénotypes de résistance.

3.2.3. Limites actuelles et mise en place d'une nouvelle stratégie

Notre objectif est d'identifier ces facteurs génétiques qui contribuent à la résistance du moustique aux parasites du paludisme. Il y a deux méthodes d'identification possibles : (1) en sélectionnant des gènes antiparasitaires et polymorphiques connus (dont ceux présents dans la région *Pbres1*), mais, il faut pour cela avoir des lignées de moustiques avec différents allèles de ces gènes et connaître le polymorphisme de ces mêmes gènes, et (2) en croisant de manière similaire des lignées de moustiques où TEP1 est invariant afin de trouver les facteurs qui interagissent avec lui.

Cependant, certaines limites actuelles réduisent notre champ d'investigation : (1) le manque d'information sur le polymorphisme des lignées de moustiques du laboratoire, (2) la mauvaise qualité du génome d'*Anopheles gambiae* et (3) le peu de marqueurs pour le génotypage.

Nous avons donc mis en place une nouvelle stratégie basée sur de nouvelles lignées de moustiques isolées au laboratoire pour identifier (1) les facteurs qui interagissent avec le polymorphisme de TEP1 et (2) les facteurs génétiques indépendants de TEP1.

Lignée résistante	L3-5	VkR	S1_low
Lignée sensible	G3	VkS	S1_high
Phénotype	Mélanisation / Lyse	Lyse /Mélanisation	Lyse
Génotype	TEP1*R ou *S 2L*a*+/+	TEP1*R ou *S 2L*a*+/+	TEP1*S 2L*a*+/+

Tableau 7. Lignées résistantes et sensibles de moustiques utilisées au laboratoire. Le phénotype indiqué concerne le mode de suppression des parasites morts et le génotype indique l'allèle de TEP1 et le type d'inversion chromosomique (rotation à 180° de portions du chromosome) : 2L*a*+/+ désigne la forme standard de l'inversion *a* sur le bras gauche du chromosome 2.

Les nouvelles lignées utilisées au laboratoire ont été sélectionnées à partir de lignées parentales Vkper et NGousso. Vkper est une lignée provenant du Burkina-Faso et contenant les allèles TEP1*R et TEP1*S. La sélection parmi leur progéniture des moustiques homozygotes pour TEP1*R a donné la lignée VkR et ceux homozygotes pour TEP1*S a donné la lignée VkS. NGousso est une lignée provenant du Cameroun et contenant les allèles TEP1*S1 et TEP1*S2. Des moustiques homozygotes TEP1*S1 ont révélé deux phénotypes extrêmes : l'un résistant et l'autre sensible, ce qui a donné, respectivement, les lignées S1_low et S1_high (Fig. 15).

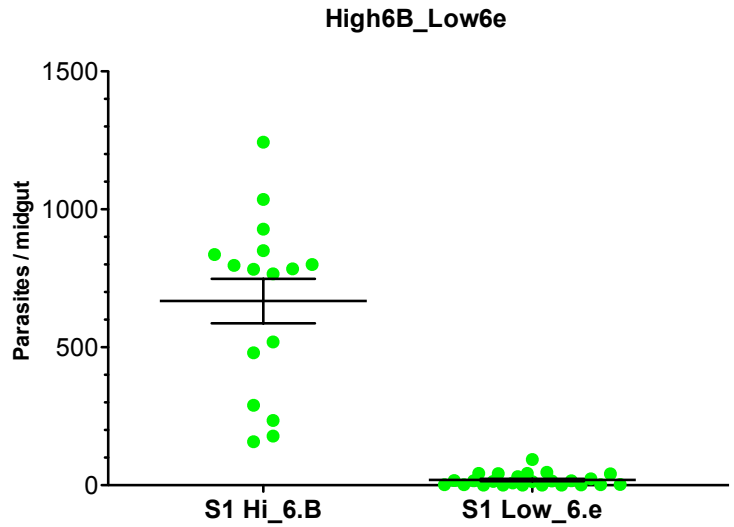


Figure 15. Nombre de parasites vivants par intestin pour la lignée S1_high (S1 Hi_6.B) et la lignée S1_low (S1 Low_6.e). Chaque point vert représente un moustique disséqué.

Notre stratégie est de :

(1) génotyper nos lignées résistantes et sensibles par séquençage ce qui va nous permettre de lister tous les polymorphismes. Ces derniers serviront, par la suite, de marqueurs génétiques pour des cartographies de QTL et dans les régions génomiques d'intérêt, i.e. liées à la résistance, de lister les gènes polymorphiques.

(2) séquencer les transcriptomes pour lister les gènes différentiellement régulés entre lignées résistantes et sensibles. Parmi eux, il sera possible de retenir les gènes situés dans des loci liés à la résistance.

(3) séquencer et d'analyser les petits ARNs.

4. Rôle du séquençage et de la bioinformatique dans l'identification des facteurs génétiques et non génétiques du moustique Anopheles gambiae contrôlant le développement du parasite du paludisme

En génétique, le séquençage permet de déterminer l'ordre d'enchaînement des nucléotides pour une séquence d'ADN donnée. Les deux premières techniques de séquençage apparaissent dans les années 1970. Elles sont développées indépendamment, l'une par l'équipe de Walter Gilbert (Maxam and Gilbert, 1977) aux États-Unis et l'autre par celle de Frederick Sanger (Sanger and Coulson, 1975) en Grande-Bretagne. Les deux hommes ont reçu le Prix Nobel de Chimie en 1980.

Il y a 10 ans sortait la première plateforme de séquençage de 2nde génération : 454 Life Sciences (Margulies et al., 2005). Ont suivi Illumina/Solexa en 2006, puis SOLiD en 2007 et IonTorrent en 2010. Les technologies de 2nde génération sont aussi appelées NTS (Nouvelles Technologies de Séquençage), NGS (Next Generation Sequencing) ou HTS (High-Throughput Sequencing).

Trois étapes principales composent le séquençage à très haut débit :

- la préparation des librairies : fragmentation aléatoire des brins d'ADN, ligature des adaptateurs aux brins, sélection de la taille de brins choisie, amplification par PCR,
- la génération des clusters : répliquer environ 1000 fois la même séquence d'ADN,
- le séquençage : ajout des primers, de la polymérase et des déoxynucléotidestriphosphates (dNTPs) fluorescentes puis lecture à chaque incorporation d'une base complémentaire à la séquence à déterminer.

Actuellement, Illumina est la plateforme majoritairement utilisée en raison de son faible coût, de sa précision et de sa forte production de données (Liu et al., 2012; Mavromatis et al., 2012).

Les technologies de séquençage ont permis d'obtenir le génome de 1 131 archées, 47 654 bactéries et 10 570 eucaryotes selon la Genome OnLine Database (GOLD, (Reddy et al., 2014)).

Le séquençage à très haut débit et les analyses bioinformatiques des données de séquençage ont énormément apporté à la science. Les applications de ces deux domaines combinés sont variées et quasiment illimitées :

- Assemblages de génomes entiers
- Étude d'expression des gènes
- Identification des polymorphismes
- Identification des mutations
- Étude des interactions ADN-protéine
- Découverte et annotation de gènes
- Étude des petits ARNs (microARNs, petits ARN interférents, etc.)
- Étude de structures
- Classification des espèces
- Etc.

L'apport de ces technologies a clairement été important en médecine où la connaissance des maladies génétiques humaines s'est fortement développée (identification des gènes liés aux maladies génétiques rares, étude du génome dans les cas de cancers, thérapies ciblées selon la mutation découverte, etc., (Koboldt et al., 2013)).

Dans le cas des recherches en lien avec le paludisme, le séquençage et la bioinformatique ont eu un fort impact les faisant ainsi rapidement progresser.

4.1. Publication des génomes du parasite du paludisme, de son vecteur et de son hôte naturel

Le lancement du projet Génome humain a débuté au début des années 1990 par la réunion de plusieurs grands centres de séquençage en un consortium international. L'objectif était d'obtenir la séquence complète du génome humain, soit 3,2 milliards de nucléotides. Le consortium a réparti les 24 chromosomes dans les différents centres pour le séquençage. Il y a deux stratégies principales pour réaliser le séquençage d'un génome entier complexe. La méthode choisie pour le projet Génome humain est la méthode « hierarchical shotgun sequencing ». Dans cette approche, l'ADN génomique est découpé en fragments d'environ 150 Mégabases (Mb) et inséré dans des vecteurs BAC (« bacterial artificial clone »), puis transformés dans la bactérie *E. coli* où ils ont été répliqués et stockés. Les inserts BAC ont été ensuite isolés et cartographiés pour déterminer l'ordre de chaque fragment de 150 Mb cloné. Cet enchainement a été répertorié en tant que « Golden Tiling Path ». Chaque fragment BAC de cet enchainement a été divisé de façon aléatoire en plus petites pièces qui ont été, chacune, clonées dans un plasmide et séquencées par les deux brins. Les séquences obtenues ont été alignées entre elles afin de former de plus grandes séquences par assemblage des séquences se chevauchant. Chaque brin avait été séquencé 4 fois pour obtenir une couverture de 8X.
La première ébauche de la séquence du génome humain a été célébrée en juin 2000 à la Maison blanche. Après la publication du génome en 2001 (Venter et al., 2001), le travail de finition s'est achevé en avril 2003. Une version complète et précise à 99,99% de la séquence du génome humain est aujourd'hui librement accessible en ligne, à la disposition des chercheurs du monde entier.

En 1996, un effort international vise à séquencer le parasite *Plasmodium falciparum*, responsable de centaines de millions de cas de paludisme et du décès de plus d'un million d'enfants africains par an. La lignée choisie est composée d'environ 23 Mb réparties sur 14 chromosomes. La méthode sélectionnée pour le séquençage est celle du « whole chromosome shotgun sequencing ». Dans cette approche, des petits fragments de chaque chromosome séparé sont séquencés aléatoirement. Les séquences sont ensuite alignées entre elles pour détecter les superpositions qui serviront pour les assembler en grandes séquences.
La séquence complète de cette lignée fut publiée en 2002 (Gardner et al., 2002).

En 2001, un consortium international se constitue dans le but de séquencer le génome complet du moustique *Anopheles gambiae*. Cette espèce étant la principale vectrice du paludisme en Afrique, une meilleure connaissance de son génome permettrait de mieux la combattre. C'est la stratégie du « whole genome shotgun » qui a été choisie et où des petits fragments du génome sont séquencés au hasard. Les superpositions entre les séquences alignées ont permis de les assembler en plus grandes séquences.
Une première ébauche du génome fut publiée en 2002 (Holt et al., 2002a). Depuis, ce génome est en constante amélioration et la toute dernière version a été rendue publique en 2014.

L'apport des séquençages du génome humain, du parasite du paludisme et du vecteur à la communauté scientifique a fortement contribué à accroitre nos connaissances sur :

(1) les interactions entre ce trinôme,
(2) l'origine des résistances des parasites aux traitements,
(3) l'origine des résistances des moustiques aux insecticides,
(4) l'origine des résistances des moustiques aux parasites.

Cela permet également d'envisager de nouvelles stratégies de lutte contre le paludisme comme le remplacement de la population de moustiques sensibles au paludisme par une population de moustiques résistants en identifiant les facteurs génétiques et non génétiques qui sont responsables de cette résistance.

4.2. Utilisation des NGS et de la bioinformatique pour l'identification des facteurs génétiques et non génétiques

4.2.1 Identification des polymorphismes

Un de nos objectifs est de lister les polymorphismes dans les lignées de moustiques du laboratoire qui serviront de marqueurs génétiques et qui permettront de lister les gènes polymorphiques.
Pour ce faire, il est envisagé de réaliser un séquençage paired-end (PE) d'ADN génomique (ADNg) en utilisant la technologie de séquençage d'Illumina (Fig 16).

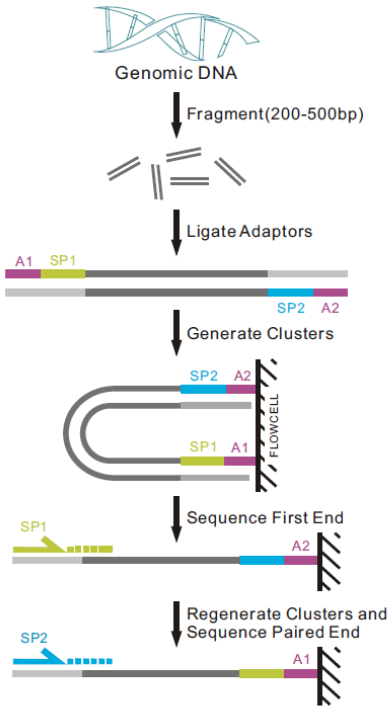


Figure 16. Processus de séquençage paired-end d'ADN génomique Illumina. Source : Illumina.com

Après le séquençage massivement parallèle, les paires de séquences d'ADN, séparées par un nombre de bases déterminé à l'avance, qu'on obtient sont appelées « reads paired-end». En français, « read » se traduit par « lecture » mais j'utiliserai le nom de read pour les qualifier.

Pour obtenir le génome de la lignée séquencé, les reads PE vont être assemblés via l'étape de l'assemblage *de novo*.
Un assemblage est par définition un ensemble de plus grandes séquences provenant de la superposition des reads et se rapprochant, par exemple, le plus possible de la séquence d'un génome. Ces séquences sont appelées contigs si elles ne contiennent pas de base N et scaffolds lorsqu'elles en contiennent. En fait, un scaffold se compose de contigs positionnés les uns par rapport aux autres grâce à l'information de distance entre des reads appariés. Pour les relier entre eux, les « trous » sont comblés par des N.
L'assemblage *de novo* est réalisé quand un génome n'a jamais été séquencé et assemblé. Cette méthode pose plusieurs problèmes que l'on peut retrouver pour certains, par analogie, dans un jeu de puzzle :

- nous avons des millions ou des milliards de pièces/séquences
- nous n'avons pas d'image/génome de référence
- nous ne savons pas quel est l'endroit (brin sens) et l'envers (brin anti-sens)
- certaines pièces/séquences sont défectueuses et ne s'assemblent pas dû aux erreurs de séquençages
- certaines pièces/séquences sont absentes.

Les brins sens (forward) et anti-sens (reverse) sont déterminés d'après le sens de transcription.

Les génomes publiés sont donc souvent incomplet et peuvent être source d'erreurs. Mais l'évolution constante des techniques de séquençage et d'assemblage tend à combler et à corriger les versions assemblées des génomes.

Il existe plusieurs outils d'assemblage qui se basent sur trois types de techniques : l'overlap-layout-consensus (OLC), les greedy assembleurs et le graphe de de Bruijn (Fig 17). Cette dernière est la base de nombreux outils : Velvet, SOAPdenovo, IDBA, etc. (Luo et al., 2012; Peng et al., 2010; Zerbino and Birney, 2008).

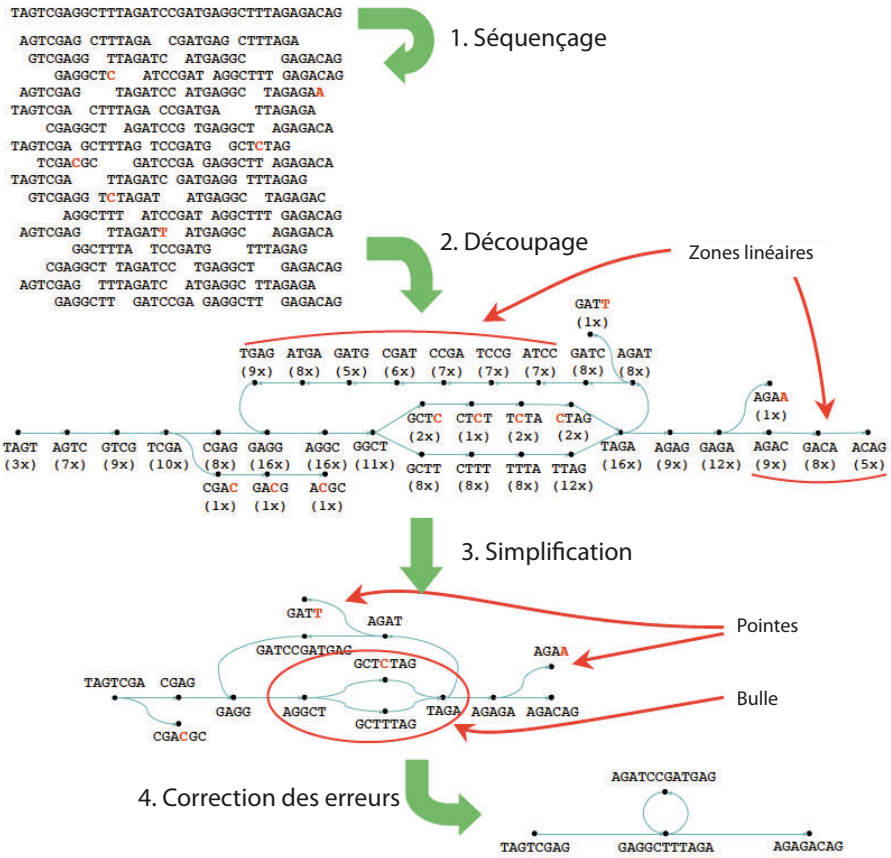


Figure 17. Principe de l'assemblage par le graphe de de Bruijn. Source : (Flicek and Birney, 2009), modifié

Dans les graphes de de Bruijn, le paramètre le plus important est le k qui détermine la taille des k-mers. Ce dernier joue énormément sur la qualité de l'assemblage fini : un tout petit k produira de grands contigs mais avec beaucoup d'erreurs, un trop grand k sortira des contigs plus justes mais trop petits. Il est donc nécessaire de trouver un compromis entre les deux.
Mais comment estimer la meilleure taille de k-mers puisqu'elle diffère à chaque scénario d'assemblage de données. La première possibilité, si on a les ressources informatiques et du temps de disponible, est de lancer plusieurs assemblages avec des tailles de k-mers différentes. Sinon, une autre possibilité est d'utiliser un outil de recherche d'un k optimal pour des données sélectionnées.

Une fois le génome pseudo-reconstitué, les polymorphismes intra-lignées peuvent être identifiés en alignant les reads d'une lignée sur le génome de la même lignée et les polymorphismes inter-lignées en alignant les reads d'une lignée sur le génome d'une autre lignée. Les alignements peuvent être réalisés par des outils comme Bowtie ou Bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2009).
Ensuite, le résultat des alignements doit passer par des outils de prédiction de variants tels que GATK ou SAMtools (Li, 2011; McKenna et al., 2010) pour obtenir une liste de variants probables.

Le séquençage et l'assemblage d'un génome peuvent permettre aussi de clarifier des portions du génome par rapport au génome sélectionné comme référence publique.

4.2.2. Mesure de l'expression des gènes

Pour lister les gènes différentiellement régulés entre lignées résistantes et sensibles, il faut séquencer les transcriptomes de nos lignées.

Pour ce faire, il faut réaliser un séquençage Illumina RNA-Seq, i.e. le séquençage des ARNs via le séquençage de librairies d'ADN complémentaire (ADNc).
Les reads ainsi obtenus peuvent ensuite être alignés par des outils spécifiques prenant en compte les jonctions d'épissage des gènes comme TopHat (Kim et al., 2013; Trapnell et al., 2009). Les alignements serviront de bases aux outils d'étude des gènes différentiellement exprimés comme DESeq (Anders and Huber, 2010).

4.2.3. Étude des petits ARNs

Afin d'étudier les petits ARNs d'un organisme, il est aussi possible de les séquencer grâce à la technologie d'Illumina.
L'alignement des reads sur le génome ou sur des séquences de référence permet d'étudier le profil des petits ARNs produits dans des conditions expérimentales déterminées. Des outils comme Bowtie peuvent être utilisés.

4.2.4. Identification des micro-organismes

Un de nos projets est l'identification des micro-organismes (MO) chez nos lignées de moustiques résistantes et sensibles afin de les comparer.

Il est possible de réaliser un séquençage d'ADNg à partir des moustiques. Pour l'analyse des reads obtenus, des outils spécifiques de l'analyse de métagénomes sont disponibles.
Pour en identifier la provenance, les reads peuvent être (1) alignés sur des bases de données contenant des génomes de référence avec Bowtie, Kraken, etc. (Langmead et al., 2009; Wood and Salzberg, 2014) ou (2) assembler pour ensuite aligner les contigs sur des bases de données avec une combinaison d'outils d'assemblage et d'alignements (metaVelvet, RayMeta, BLAST, etc. (Altschul et al., 1990; Boisvert et al., 2012; Namiki et al., 2012)). On trouve également des outils proposant de faire plusieurs analyses des données métagénomiques : MetAMOS, MOCAT, etc. (Kultima et al., 2012; Treangen et al., 2013).

Dans le cas de l'identification de MO, les bases de données comporteront des génomes de bactéries, de virus, de champignons et de protistes. Ces bases de données sont regroupées dans des centres de ressources biologiques tel que le National Center for Biotechnology Information (NCBI).

Présentation des chapitres

Notre objectif est d'identifier des facteurs génétiques et non génétiques contrôlant la résistance chez le moustique *Anopheles gambiae* aux parasites du paludisme *Plasmodium berghei*. Les récentes technologies de séquençage à très haut débit et la technique de rasRNAi offrent de nouvelles possibilités pour identifier et tester ces facteurs. Mon travail a été le développement et la mise en place de nouvelles méthodes ou outils utilisant ces technologies.

Dans le premier chapitre, notre but est d'identifier les polymorphismes de nos lignées résistantes et sensibles afin de sélectionner des marqueurs génétiques et de lister les gènes polymorphiques. Dans un premier temps, nous avons réalisé le séquençage d'une lignée résistante du laboratoire. L'alignement des séquences obtenues sur le génome de référence d'*A. gambiae* nous a montré qu'il n'est pas possible de déterminer l'ensemble des polymorphismes de la lignée. Nous avons donc assemblé le génome de notre lignée pour lister les polymorphismes. Finalement, nous avons mis en place une stratégie pour identifier les différences génétiques entre moustiques résistants et moustiques sensibles.

Dans le second chapitre, j'ai contribué à l'amélioration de la méthode du « reciprocal allele-specific RNA interference » (rasRNAi) afin de tester les gènes polymorphiques, sélectionnés grâce aux résultats du Chapitre I, entre les lignées résistantes et sensibles. Ceci permettra d'identifier les gènes responsables de la résistance du moustique *Anopheles gambiae* aux parasites du paludisme. Cette méthode permet d'évaluer la contribution des différents allèles d'un même gène pour un caractère donné dans un même fond génétique en injectant un ARN double brin (dsRNA) spécifique de l'allèle à tester et qui ne doit pas contenir des fragments > 19 paires de bases identiques entre les allèles. Cette méthode n'est pas réalisable dans le cas des gènes faiblement polymorphiques. Pour créer une nouvelle sonde dsRNA allèle-spécifique dans ces cas-là, nous avons étudié le processus de découpage d'un dsRNA injecté dans les cellules du moustique par le séquençage et l'analyse des petits ARNs. Forts des observations de ce processus, nous avons élaboré de nouvelles sondes dsRNAs allèle-spécifique nommées siRNA Carrier qui sont capables d'inhiber un allèle spécifique dans les cas de gènes peu polymorphiques.

Dans le troisième chapitre, nous cherchons à identifier la composition du microbiote de nos lignées de moustiques. Les microorganismes jouant un rôle dans sa capacité à transmettre les parasites, nous voulons identifier et comparer les micro-organismes présents chez les moustiques résistants et sensibles. Après la tentative infructueuse d'un outil pour analyser nos données de séquençage d'ADNg d'une de nos lignées, nous avons décidé d'élaborer notre propre outil de classification de données métagénomiques sur les bases de données des bactéries, des virus, des archées et des champignons.

Chapitre I

Détermination des polymorphismes des lignées résistantes et sensibles du moustique *Anopheles gambiae* par l'utilisation des techniques du séquençage et de l'assemblage

Sommaire

I. Contexte

C'est en 2001 dans un contexte où la résistance du parasite aux médicaments anti-paludéens et la résistance du moustique aux insecticides s'étendent fortement que la communauté scientifique lance le programme de séquençage du moustique *Anopheles gambiae*. Les scientifiques espèrent avec le génome du principal vecteur du paludisme en Afrique identifier les gènes impliqués dans la résistance aux insecticides, ceux contrôlant les récepteurs olfactifs et gustatifs responsables de la localisation du repas sanguin et ceux impliqués dans la réponse immunitaire vis-à-vis du parasite *Plasmodium falciparum*. Une meilleure connaissance de la biologie du moustique et de son interaction avec le parasite pourrait permettre de découvrir de nouvelles stratégies de lutte contre le paludisme. Les interactions entre l'homme et les parasites sont aussi ardemment étudiées à l'époque mais la réalisation d'un vaccin se révèle être une tâche plus compliquée que prévue, à cause, entre autres, des variations des antigènes exprimés à la surface du parasite.
La première ébauche du génome d'*Anopheles gambiae* est rendue disponible en mars 2002 après un an de collaborations entre plusieurs laboratoires et centres de séquençage internationaux. En octobre 2002 paraît la publication dans la revue Science (Holt et al., 2002b).
C'est la première fois que la communauté scientifique possède les informations génomiques d'un organisme infectieux (*Plasmodium falciparum*, (Gardner et al., 2002)), de son hôte naturel (l'Homme, (Venter et al., 2001)) et de son vecteur (*Anopheles gambiae*).

Du moustique à l'assemblage

La lignée PEST d'*Anopheles gambiae* a été choisie pour ce projet car une banque de grands fragments génomiques clonés dans des chromosomes artificiels bactériens (BAC) était déjà disponible (via l'équipe de Franck Collins à l'université de Notre-Dame, Etats-Unis). Cette lignée est le fruit d'un croisement entre une lignée venant directement du Kenya et une lignée de laboratoire originaire du Nigeria et ayant la particularité d'avoir une mutation génétique facilement identifiable. Cette mutation liée à l'X donne aux yeux des moustiques une couleur rose et permet ainsi de détecter toute contamination entre colonies. C'est aussi cette mutation qui est à l'origine du nom de la lignée : PEST pour Pink Eye STandard.
C'est la stratégie du « séquençage aléatoire global » (whole genome shotgun ou WGS, Fig.1) où l'on séquence au hasard des petits fragments du génome qui a été choisie. Deux types de banques ont été construites : des librairies BAC à partir de mâles et de femelles mélangés, et des librairies de petites séquences d'ADN génomique avec des tailles différentes d'inserts (2.5, 10 et 50kb) pour chaque sexe. La méthode de séquençage utilisée est celle de Sanger, elle produit des tailles de séquences entre 500 et 700 paires de bases (pb). L'ensemble des données a permis d'obtenir une couverture de 10,2 fois le génome, sachant que le génome du moustique est de 278Mb.

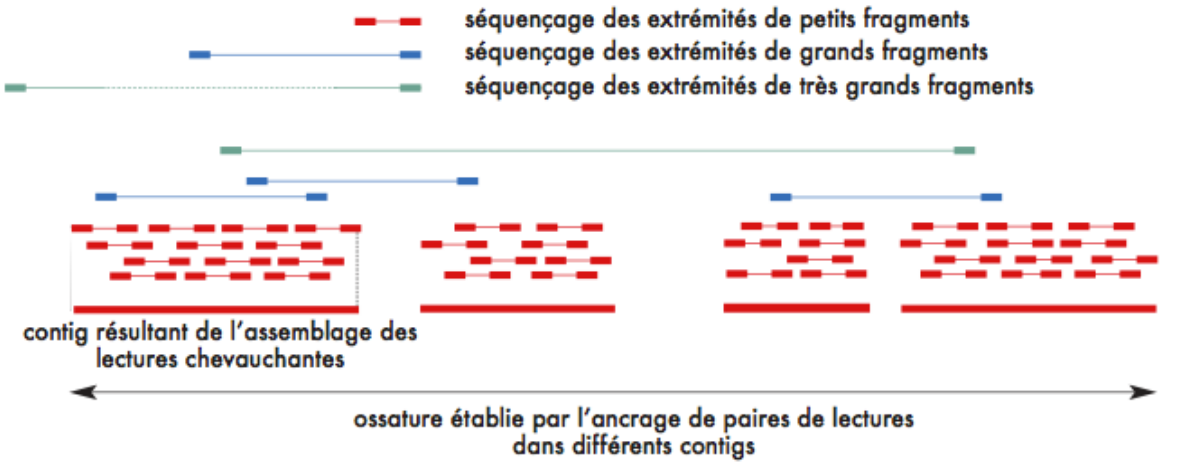


Figure 1. Stratégie du « séquençage aléatoire global » ou WGS. L'assemblage des petits fragments amène à la formation de contigs et l'information de distance entre séquences appairées permet de positionner et d'orienter les contigs les uns par rapport aux autres. (www.genoscope.cnrs.fr)

L'assemblage des séquences lues a été effectué par l'assembleur Celera (du même nom que la firme américaine) qui avait auparavant servi pour les assemblages des génomes de la drosophile, de la souris et de l'humain. Grâce à l'information d'orientation et de distance fournie par les séquences appariées et les BACs, l'assemblage a résulté en 8 987 scaffolds permettant de couvrir 278,3Mb. Parmi eux, 303 ont une taille supérieure à 30kb et recouvre 91% du génome. Chaque scaffold a été ensuite positionné et orienté sur les trois chromosomes du moustique grâce à une carte physique préalablement réalisée. Sur les 278 Mb, les chercheurs ont estimé qu'environ 21Mb avait été artificiellement dupliquées lors de l'assemblage. Ces régions dupliquées, appelées haplotypes, ont plusieurs sources: (1) la présence de multiples génomiques, (2) la possibilité de réelles duplications et (3) le polymorphisme très fort de la lignée PEST. En effet, la diversité nucléotidique, i.e. le nombre de substitutions par site, calculée pour *A. gambiae* est de 0,029 tandis que pour l'homme et la drosophile, elles sont de 0,0012 et 0,017 respectivement. L'origine de ce fort polymorphisme s'explique, entre autres, par le fait que la lignée PEST est le fruit d'un croisement entre deux lignées. C'est ce qui a poussé les chercheurs à garder les haplotypes puisque, avec leur région correspondante dans la séquence de référence, ils sont des représentations des deux lignées originaires.
La première ébauche du génome de *A. gambiae* est connue sous le nom de MOZ1.

Environ 14 000 gènes putatifs ont été annotés automatiquement en utilisant deux suites d'outils d'analyse de génome, celle de Celera et celle d'Ensembl. Ces outils annotent les gènes à partir de séquences d'ADN complémentaire (ADNc) connues, par homologie avec les gènes de la drosophile ou par prédiction *ab initio* à l'aide d'algorithmes qui utilisent les caractères connus des gènes (les codons de départ et de fin, les jonctions d'épissage, etc.) pour identifier de nouveaux gènes. Le séquençage du génome du moustique a permis de dresser la liste de gènes potentiellement impliqués dans les réponses immunitaires du moustique (Christophides et al., 2002) ou dans les mécanismes de résistance aux insecticides (Ranson et al., 2002) et a complètement changé les perspectives de recherche dans le domaine.

Perfectionnement du génome

La seconde version, MOZ2, parut en octobre 2003 (Mongin et al., 2004). Des ambiguïtés dans l'orientation et la localisation de scaffolds furent résolues grâce aux corrections manuelles de différents laboratoires. La taille totale était toujours de 278Mb.

C'est en 2006 que de majeures améliorations ont été réalisées (Sharakhova et al., 2007). Des alignements physiques supplémentaires, une analyse in silico détaillée des scaffolds et des alignements de séquences provenant de clones BAC ont conduit à une nouvelle version du génome : AgamP3. Ce dernier possède un total de 7 gros scaffolds référencés comme le Golden Path et couvrant 273Mb. Les séquences toujours non localisées ont été introduites dans un chromosome UNKN. Les scaffolds contenant des marqueurs spécifiques du chromosome Y ont été regroupés dans un chromosome Y_unplaced. Un total de 15 Mb a été ajouté au génome dont 8Mb ont servi à la reconstruction de régions péricentromériques. Deux Mb ont été reclassé en tant que génomes bactériens et environ 8Mb ont été classés comme assemblages alternatifs probables (Fig.2). Pour chacune de ces régions hyper polymorphiques, l'un des haplotype a été conservé dans le Golden Path tandis que l'autre a été présenté en tant qu'haplotype alternatif. Elles sont au nombre de 166 et sont visualisables et téléchargeables comme des scaffolds simples (Fig.3 et 4). Les divergences entre le Golden Path et la séquence haplotype se situent entre 1 à 4,5%.

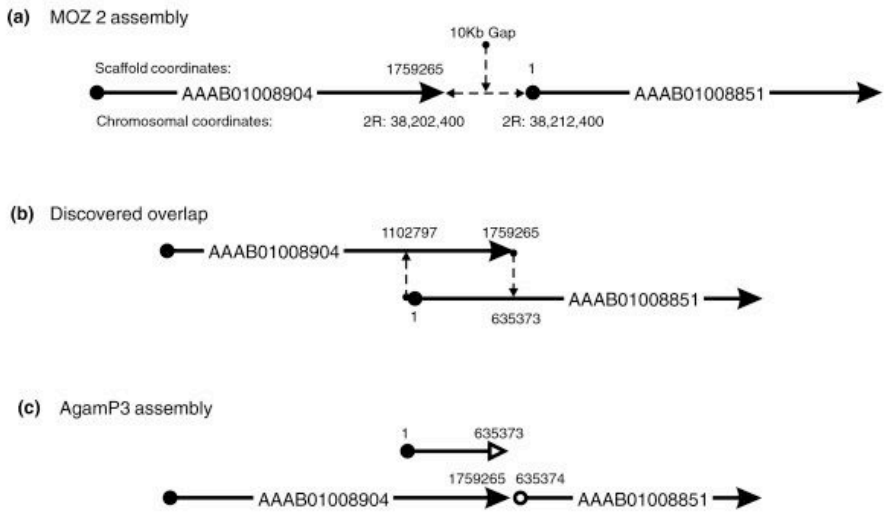


Figure 2. Exemple de scaffolds où les terminaisons adjacentes sont en fait des assemblages alternatifs de la même région. (a) Dans la version MOZ2, les scaffolds AAAB01008904 et AAAB01008851 ont été placés avec une brèche arbitraire de 10kb. (b) Après un nouvel alignement, il apparut qu'une région de 64kb était commune aux deux scaffolds. Cette région présentait cependant des divergences importantes qui ont forcé l'assembleur à les considérer comme des régions différentes du génome. (c) Aussi, la portion appartenant au début du scaffold AAAB01008851 est devenue un haplotype et la région correspondante dans le scaffold AAAB01008904, de meilleure qualité dans son ensemble, a été utilisée dans le Golden Path. Extrait de (Sharakhova et al., 2007)

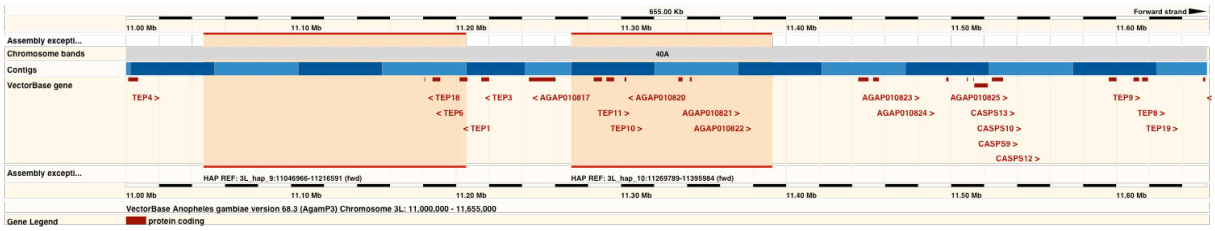


Figure 3. Visualisation de deux haplotypes (9 et 10) positionnés sur le chromosome 3L. La séquence de référence, i.e. le Golden Path est représenté par le trait bleu. Les haplotypes sont représentés par les traits fins rouges. (www.vectorbase.org)

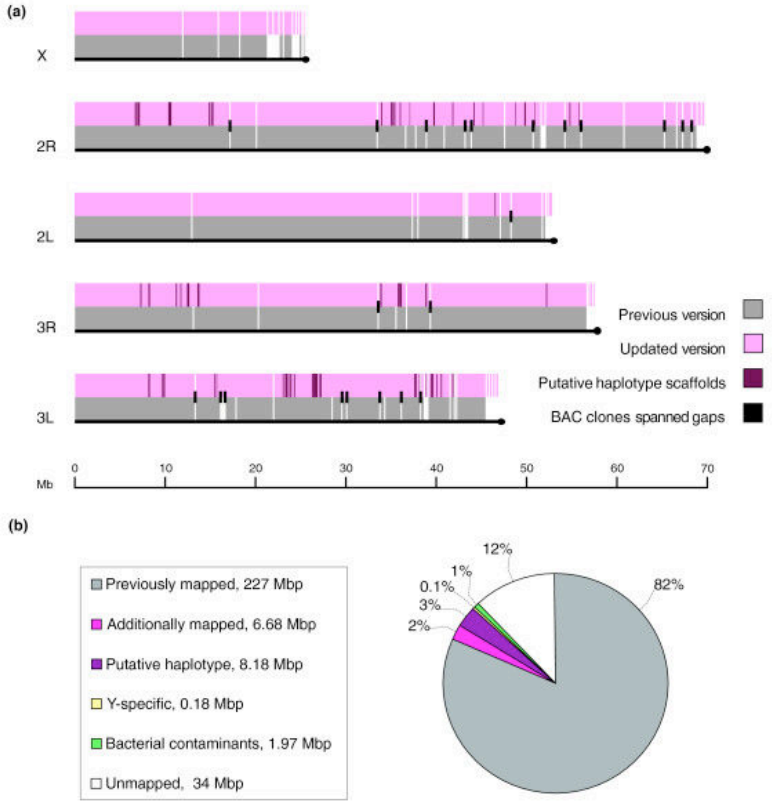


Figure 4. Comparaison entre la version MOZ2 et la nouvelle version AgamP3 de l'assemblage du génome d'*Anopheles gambiae*. (a) Les scaffolds de la précédente et de la nouvelle version sont représentés respectivement par la couleur grise et la couleur rose. Les barres violettes représentent les séquences haplotypes alternatives dont la taille est supérieure à 50Kb. Les barres noires correspondent à une brèche dont les scaffolds de part et d'autre ont été liés par l'information extraite de clones BAC. (b) Le statut mis à jour du projet du génome d'*Anopheles gambiae*. Extrait de (Sharakhova et al., 2007)

Résolution des haplotypes : assemblage des génomes des formes M et S

En 2005, le National Human Genome Research Institute a lancé un programme de séquençage de plusieurs espèces ou sous-espèces, dont les formes M et S d'*Anopheles gambiae* (http://www.genome.gov/15014493) afin de comprendre leurs divergences et de construire deux génomes de référence pour *A. gambiae*.

La colonie Pimperena représentant la forme S et la colonie Mali-NIH représentant la forme M d'*A. gambiae* ont été établie par la collecte en novembre 2005 de femelles adultes après un repas sanguin dans le village de Pimperena au Mali et dans les maisons du village de Niono au Mali, respectivement. AgamS1 et AgamM1 furent les premiers assemblages de novo issus de séquençages aléatoires globaux (WGS) (Lawniczak et al., 2010)., L'assemblage de ces génomes avec Celera a résulté en 13 042 scaffolds couvrant 236Mb pour AgamS1, et en 10 521 scaffolds pour un total de 224Mb pour AgamM1.

En 2013, au vu des différences importantes entre les deux formes moléculaires de *A. gambiae* et de la quasi absence de moustiques hybrides sur le terrain, les deux formes ont été promue au rang d'espèces appartenant au complexe d'espèce d'*A. gambiae* (Coetzee et al., 2013). La forme M est nommée *Anopheles coluzzii* (AcolM1 pour son assemblage) tandis que la forme S conserve le nom *Anopheles gambiae*.

En 2014, une nouvelle version AgamP4 a été publiée. Elle diffère de la précédente par l'ajout d'un génome mitochondrial. La disparition des haplotypes s'explique par la présence des assemblages AgamS1 et AcolM1 et à l'assignation des haplotypes à l'un ou l'autre des génomes.

Comparaison des génomes des formes M et S

L'étude des deux génomes M et S a permis de découvrir plus de 2 millions de SNPs par forme et plus de 150 000 différences fixées entre les deux. Les divergences ne se réservant pas aux régions péricentriques comme observé dans les études de microarray basée sur les gènes (Carneiro et al., 2009; Turner et al., 2005; White et al., 2010) mais s'étalant plutôt sur tout le génome (Lawniczak et al., 2010; Neafsey et al., 2010). Par ailleurs, la nature très étendue de la divergence génomique des formes M et S est liée à leur localisation géographique et la topologie de la divergence entre M et S n'est pas uniforme (Reidenbach et al., 2012).
Ces divergences vont donc jouer un rôle dans les différents degrés de résistance des moustiques aux parasites du paludisme. Par exemple, le locus contenant les trois gènes APL1 codant pour des composants essentiels de la réponse antiparasitaire du moustique s'est avéré fortement polymorphique chez *A. gambiae*, alors qu'une forte réduction de variabilité génétique a été détectée chez *A. coluzzii*, suggérant que ce locus avait été l'objet d'une forte sélection dans cette espèce (Rottschaefer et al., 2011). Fait intéressant, le gène TEP1 montre lui aussi une forte sélection dans des échantillons d'*A. coluzzii* collectés au Mali (White et al., 2011).

Le paludisme humain n'est transmis que par des moustiques du genre *Anopheles*. Cependant, au sein de ce genre, il existe environ 400 espèces dont seules 60 sont vectrices. Plusieurs facteurs peuvent l'expliquer, notamment leur tendance à prendre leur repas sanguin sur des hommes ou sur des animaux, leur lieu de vie (plus ou moins en contact avec l'habitat humain), leur espérance de vie, mais aussi l'efficacité de leur défense antiparasitaire. Afin de chercher des pistes pour expliquer ces différences phénotypiques importantes, qui déterminent la capacité vectorielle du moustique, les génomes de 16 espèces d'*Anopheles* ont récemment été séquencés et comparés.

Séquençages et assemblages d'autres Anopheles

En 2012, les génomes d'*Anopheles darlingui* et d'*Anopheles stephensi* furent séquencés et assemblés. *A. darlingui* est considéré comme le vecteur le plus actif dans les régions néotropicales (Mexique, Amérique du sud et du centre). AdarC3 compte aujourd'hui 2 221 scaffolds pour une taille totale de 136Mb. *A. stephensi* est le vecteur majeur sur des zones très étendues allant du Golfe Persique à la région Thaïlande/Laos/Chine, tout en parcourant toute l'Inde. Astel2 compte 23 371 scaffolds pour une taille totale de 221Mb.

En 2015, une communauté de chercheurs a publié la séquence et l'analyse comparative de 16 génomes d'*Anopheles* (Neafsey et al., 2015). Ces 16 espèces ont été choisies parce qu'elles couvrent de nombreuses branches du genre *Anopheles*, mais aussi en raison de leur localisation géographique et de leurs niches écologiques variées, et de la variabilité de leur capacité vectorielle avec des espèces vectrices du paludisme humain, et d'autres non vectrices (Fig. 5). Elles proviennent d'Afrique, d'Asie, d'Europe et d'Amérique latine.

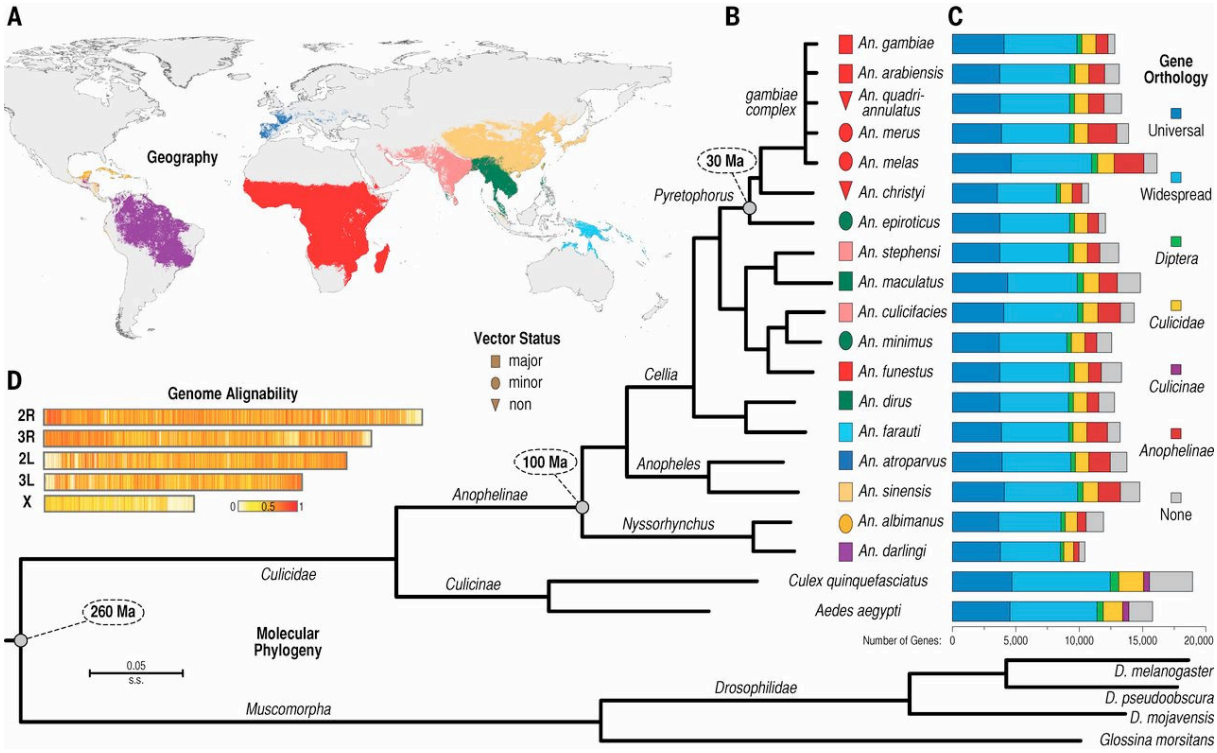


Figure 5. Géographie, état du vecteur, phylogénie moléculaire, orthologie des gènes et alignabilité des génomes des 16 nouvelles espèces d'Anopheles et de quelques diptères sélectionnés. Extrait de (Neafsey et al., 2015)

Plusieurs séquençages Illumina ont été effectués à partir d'ADN génomique et d'ARN provenant du corps entier. Les assemblages ont été réalisés avec ALLPATHS-LG (Gnerre et al., 2010) (Tableau 1).

Species	Assembly Version	GenBank Assembly	Assembly Size (Mb)	Gaps (Kb)	Scaffold N25‡ (Kb)	Scaffold N50‡ (Kb)	Scaffold N75‡ (Kb)	Number of Scaffolds	%GC
A. albimanus	AalbS1	GCA_000349125.1	170.5	6,961	24,066	18,068	8,610	204	49.21
A. arabiensis	AaraD1	GCA_000349185.1	246.6	35,125	10,035	5,604	2,699	1,214	44.68
A. atroparvus	AatrE1	GCA_000473505.1	224.3	35,338	12,071	9,207	4,026	1,371	46.35
A. christyi	AchrA1	GCA_000349165.1	172.7	2,671	18	9	4	30,369	42.74
A. coluzzii	AcolM1	GCA_000150765.1	224.5	14,926	7,194	4,437	1,263	10,521	44.38
A. culicifacies	AculA1	GCA_000473375.1	203.0	15,840	42	22	11	16,162	42.68
A. darlingi	AdarC2	-	134.7	27	272	115	54	2,160	48.39
A. dirus	AdirW1	GCA_000349145.1	216.3	18,566	13,137	6,906	3,666	1,266	46.18
A. epiroticus	AepiE1	GCA_000349105.1	223.5	20,855	743	367	162	2,673	43.95
A. farauti	AfarF1	GCA_000473445.1	181.0	5,030	1,777	1,197	541	550	44.69
A. farauti	AfarF2	GCA_000473445.2	183.1	7,280	22,739	12,895	6,025	310	44.69
A. funestus	AfunF1	GCA_000349085.1	225.2	35,208	1,127	672	380	1,392	41.59
A. gambiae	AgamP3	GCA_000005575.1	273.1	20,655	53,201	49,364	42,390	7†	44.27
A. gambiae	AgamS1	GCA_000150785.1	236.4	8,363	6,249	3,801	1,822	13,042	44.33
A. maculatus	AmacM1	GCA_000473185.1	141.9	10,063	7	4	2	47,797	44.21
A. melas	AmelC1	GCA_000473525.1	227.4	20,678	31	18	10	20,281	44.94
A. melas	AmelC2	GCA_000473525.2	224.2	20,733	31	18	10	20,229	44.84
A. merus	AmerM1	GCA_000473845.1	251.8	33,614	658	342	134	2,753	44.64
A. merus	AmerM2	GCA_000473845.2	288.0	70,729	2,833	1,490	538	2,027	44.64
A. minimus	AminM1	GCA_000349025.1	201.8	15,387	21,278	10,313	5,773	678	42.70
A. quadriannulatus	AquaS1	GCA_000349065.1	283.8	74,862	2,616	1,641	622	2,823	44.76
A. sinensis	AsinS1	GCA_000472065.1	241.4	49,229	151	81	35	11,270	43.93
A. sinensis	AsinS2	GCA_000472065.2	375.8	185,483	1,104	579	208	10,448	43.95
A. stephensi	Astel2	GCA_000300775.2	221.3	11,843	2,767	1,591	597	23,371	44.80
A. stephensi	AsteS1	GCA_000349045.1	225.4	29,185	1,402	837	450	1,110	45.02
Aedes aegypti	AaegL1	GCA_000004015.1	1384.0	73,881	2,717	1,547	742	4,758	38.27
Culex quinquefasciatus	CpipJ1	GCA_000209185.1	579.0	39,083	949	487	221	3,171	37.42
Drosophila melanogaster	BDGP5	GCA_000001215.2†	168.7	67,764	27,905	23,012	22,423	14†	41.74

‡ N25/N50/N75 is the scaffold length, *x*, such that 25%/50%/75% of the genome is assembled on scaffolds of length *x* or longer.
† The AgamP3 and BDGP5 assemblies are mapped to chromosomes, and the BDGP5 assembly from FlyBase (dmel-r5.57) additionally contains heterochromatin.

Tableau 1. Statistiques des assemblages (Neafsey et al., 2015)

Les statistiques des assemblages produits varient considérablement entre les anophèles. Les tailles totales vont de 134Mb à 375Mb avec plus ou moins de brèches et de scaffolds. Le plus grand génome est aussi celui qui a le plus de brèches (185 kb) mais pas celui qui a le plus grand nombre de scaffolds. Par exemple, les deux sous-espèces d'*A. stephensi*, Astel2 et AsteS1, ont quasiment la même taille totale mais un nombre de scaffolds complètement différents, 1 000 pr AsteS1 et 23 000 pour Astel2. Cela pourrait s'expliquer par un polymorphisme important chez Astel2 qui complique fortement l'assemblage et conduit à la création de régions dupliquées artificiellement. C'est surement aussi le cas pour *A. maculatus* qui totalise presque 48 000 scaffolds.

Ces grandes variations entre espèces et sous-espèces ont été l'objet d'un projet visant à séquencer plusieurs lignées naturelles isolées pour identifier leurs polymorphismes. Les moustiques ont été fournis par le Malaria Research and Reference Reagent Resource Center (www.mr4.org). Il y a 6 *A. arabiensis* dont 3 de la lignée SENN et 3 de la lignée KGB, et 15 *A. gambiae* dont 3 L3-5, 3 4ARR, 3 Akron, 5 Kisumu et 1 G3. Toutes les informations et les données séquencées sont disponibles via VectorBase sous le numéro de projet VBP0000002.
Ces données serviront de ressources pour les études d'association du génome entier (GWAS) ou tout autre étude génomique dont le but est d'identifier des loci génétiques responsables de tel ou tel phénotype, tel que la résistance aux insecticides ou aux parasites.

Le génome AgamP3 et notre projet

Notre but est d'identifier les facteurs génétiques impliqués dans la résistance du moustique aux parasites du paludisme murin. Pour cela, il nous faut trouver de nouveaux marqueurs génétiques spécifiques à chaque lignée sensible ou résistante sélectionnée qui serviront au génotypage de croisements. Ensuite, il nous faudra la séquence précise des régions d'intérêts, i.e. identifiées comme porteuses de facteurs génétiques de résistance.
C'est la version AgamP3 qui était disponible lors du début du projet et qui sera utilisée comme référence tout au long de ce travail.

Dans une étude préalable en collaboration avec le Sanger Center, notre laboratoire a séquencé deux BACS de la lignée PEST, de 115kb et de 120kb, couvrant partiellement une région génomique de 650kb qui comprend 10 gènes TEP (Fig.6). En effet, bien que riche en gènes et peu répétitive, le Golden Path était de relative mauvaise qualité, avec notamment de nombreuses interruptions de séquences (NNNs). Ces 2 BACs ont donc été séquences afin d'obtenir une information précise sur le nombre et la séquence exacte des TEPs de ce cluster

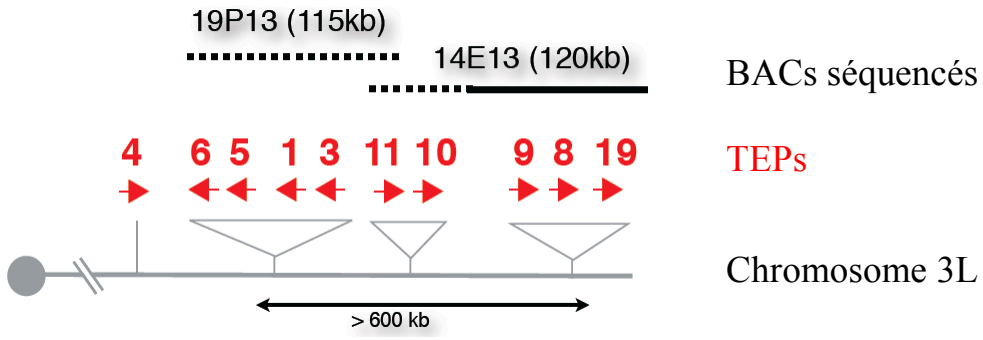


Figure 6. Localisation des régions séquencées sur le chromosome 3L.

Ensuite, le laboratoire a réalisé un alignement avec LAGAN (Brudno et al., 2003) entre le génome AgamP3 et ces deux séquences (Figures 7, 8).

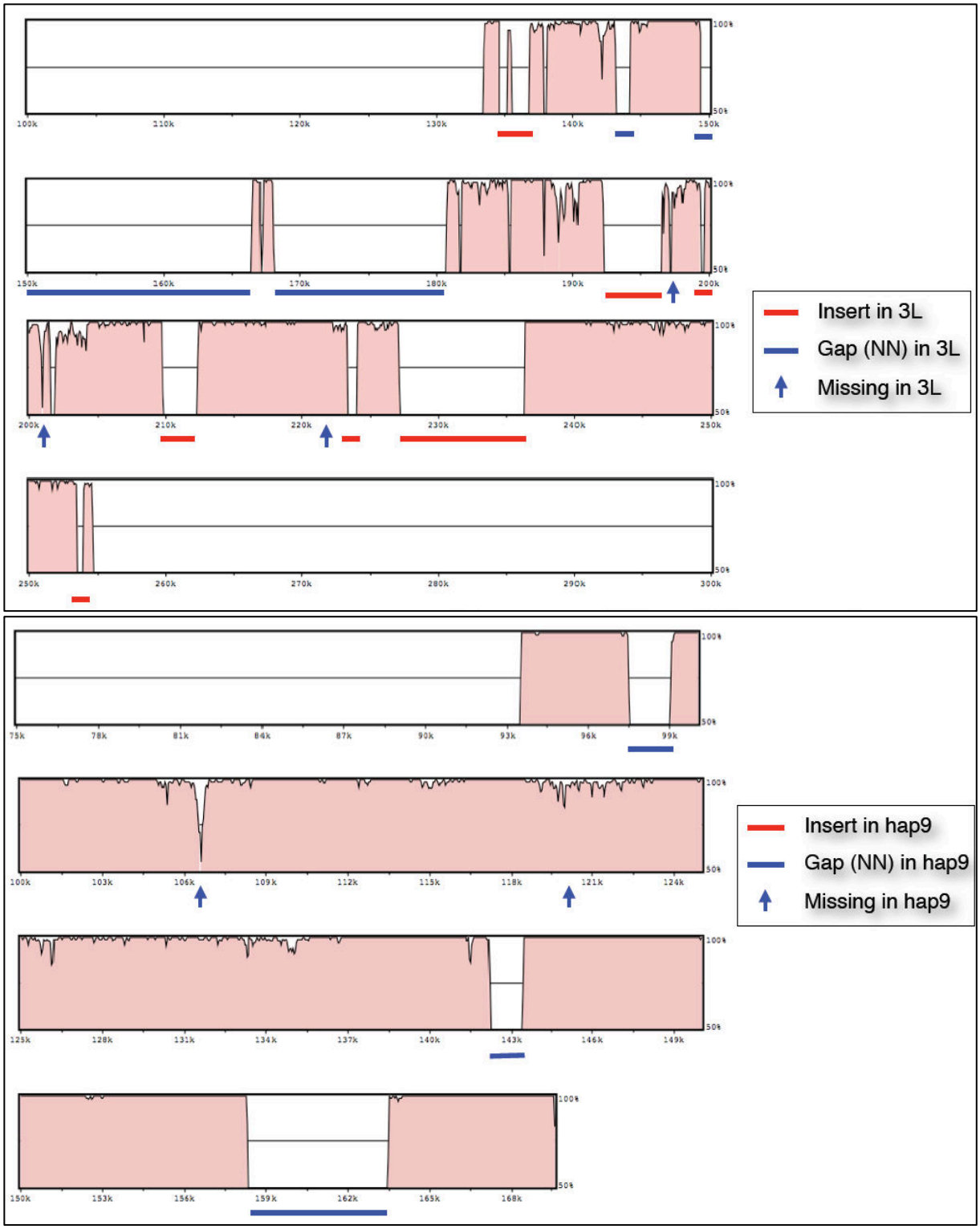


Figure 7. Alignement du BAC 19P13 sur le chromosome 3L (Golden Path) et alignement du même BAC 19P13 sur l'haplotype 9 correspondant à la même portion du chromosome 3L. Les zones rosées correspondent au pourcentage d'alignement. La première région séquencée ressemble plus à la séquence de l'haplotype plutôt qu'au Golden Path.

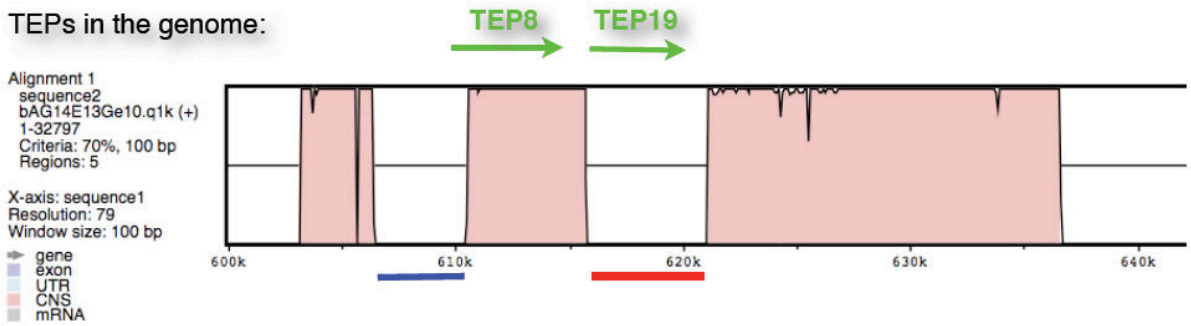


Figure 8. Alignement du BAC 14E13 sur la région du chromosome 3L comprenant les gènes TEP8 et TEP19. La zone contenant le gène TEP19 n'existe pas réellement dans la lignée PEST. Nous pouvons facilement supposer que ce gène est un artefact de l'assemblage puisqu'en fait il s'agit d'une autre variante ou allèle du gène TEP8. Pour cette région, il n'y a pas d'haplotype correspondant.

Pour conclure, l'existence des gènes TEP6, 5, 1, 3, 9 et 8 a été confirmée via ces alignements (Figures 7 et 8) et cet exemple a démontré les faiblesses de AgamP3 : (1) la séquence du Golden Path contient des portions qui ne représentent qu'une partie du polymorphisme de l'espèce, (2) certaines régions, comme celle que nous avons séquencée, présentent une extrême variabilité au sein de l'espèce, (3) malgré les améliorations qui ont conduit à cette version, certaines duplications persistent (exemple de TEP8/19) et (4) le Golden Path présente de nombreuses brèches, insertions ou délétions.

Ces constatations nous amènent à la conclusion que nous ne pouvons nous limiter à l'alignement simple de reads issus de nos lignées séquencées sur le Golden Path qui risque d'être fortement perturbé par le fort polymorphisme de certaines régions, notamment celles couvertes par des haplotypes, et/ou la mauvaise qualité du Golden Path. Par ailleurs, nous avons vu dans l'introduction que le génome d'*A. gambiae* possède de nombreuses inversions qui, du fait de la quasi absence de recombinaison homologue au niveau de l'inversion, divergent fortement entre les régions ancestrales et réarrangées. Là encore, il nous faudra être attentifs aux alignements dans ces régions, en particulier au niveau des points de rupture. Il est donc important pour nous de nous assurer que (1) la plupart des inversions (et notamment les plus grandes) soient consistantes entre les lignées résistantes et les lignées sensibles au sein d'une même paire afin de ne pas supprimer la recombinaison pendant les croisements, et (2) les régions d'intérêt situées au niveau de ces inversions soient correctement assemblées.

Notre projet est de lister tous les polymorphismes entre les différentes lignées de moustiques et de les utiliser pour réaliser des études génétiques sur la résistance du moustique vis-à-vis du parasite. Nous pourrons ensuite répertorier les gènes responsables de cette résistance.
Dans ce chapitre, j'alignerai des reads paired-end issus d'un séquençage d'ADN génomique d'une de nos lignées sur le génome de référence *Anopheles gambiae* et je confirmerai nos premières observations. Puis je démontrerai la difficulté d'assembler le génome de cette lignée, pourtant relativement moins polymorphique pour cette espèce, et a fortiori, de toutes les lignées que nous avons sélectionnées. Enfin, je proposerai une nouvelle stratégie afin (1) de produire une liste de marqueurs génétiques qui pourront être utilisés pour les analyses génétiques (QTL mapping, GWAS) et (2) d'obtenir une séquence précise des régions d'intérêt détectées dans les analyses génétiques.

II. Matériels et méthodes

- Matériel biologique et Séquençage

Les moustiques séquencés proviennent de la lignée R1iso2 isolée à partir de la lignée résistante L3-5 par 2 sélections isofemelles successives. A chaque sélection, la progéniture d'une seule femelle infectée présentant un nombre important de parasites mélanisés est conservée et amplifiée jusqu'à la prochaine sélection. La librairie d'ADN génomique a été préparée à partir de 6 moustiques afin de trouver un bon équilibre entre (1) avoir suffisamment de matériel génétique pour préparer la bibliothèque de séquençage, et (2) limiter le nombre d'individus pour limiter le polymorphisme qui nuit à l'assemblage et pouvoir identifier les SNPs des erreurs de séquençage. Une fois l'ADN génomique prêt, il a été fracturé mécaniquement par une méthode élaborée par la firme Covaris. La librairie a été préparé à partir des fragments obtenus et selon le protocole Illumina pour la préparation des librairies paired-end. Le séquençage a été effectué sur la plateforme Illumina HiSeq 2000 à l'EMBL de Heidelberg en Allemagne. Les reads ont été générés en paired-end, ou reads appariés, dont nous connaissons la distance les séparant, environ 200 bases.
Toutes les étapes précédant celle du séquençage ont été réalisées par mes collègues.

- Matériel informatique

Pour toutes les étapes sauf la visualisation des données, j'ai eu recours à l'HPC qui fournissait suffisament de ressources en mémoire vive (24 Téraoctets) et en nombre de processeurs (380) nécessaires aux analyses.

- Pré-analyse

Il est impératif de vérifier la qualité des données brutes issues du séquenceur avant de commencer toute analyse. FastQC est un outil très complet fournissant un rapport qui nous permet d'évaluer la qualité des données à travers des informations statistiques et visuelles. Les formats d'input acceptés sont les SAM, BAM et FastQ. Le lancement du programme peut se faire soit en ligne de commande soit via une interface graphique. Il a été développé par le groupe de bioinformatique de l'Institut Babraham (Cambridge, Angleterre).

- Alignement

Bowtie est le logiciel d'alignement très populaire que j'ai utilisé pour toutes mes études. C'est un outil permettant l'alignement de larges sets de petits reads sur une ou plusieurs grandes séquences telles que les génomes. Il existe deux versions : Bowtie sorti en 2009 (Langmead et al., 2009) et Bowtie2 sorti en 2012 (Langmead and Salzberg, 2012). Les deux créent un index du ou des séquence(s) de référence basé sur la transformation de Burrows-Wheeler et l'index FM. Bowtie est ensuite plus qualifié pour les tout petits reads jusqu'à 50 bases. Bowtie2 gère les reads de 50 à 1 000 bases et supporte les alignements avec brèches, paired-end et locaux.
Ils peuvent tourner sur plusieurs processeurs et augmenter ainsi la vitesse d'alignement.
Le format de sortie est le format SAM. Pour manipuler les fichiers SAM, j'exploiterai les diverses fonctionnalités de SAMtools (Li et al., 2009).
J'ai eu aussi recours à MUMmer qui est un logiciel d'alignement très rapide de génomes entiers ou de contig/scaffolds entre eux (Delcher et al., 1999, 2002; Kurtz et al., 2004).

- Assemblage

Pour choisir mon outil d'assemblage je me suis basée sur les dernières publications de génomes séquencés, sur la popularité des outils et sur l'étude faite par l'Assemblathon. Il a été démontré par les équipes des Assemblathon 1 (Earl et al., 2011) et 2 (Bradnam et al., 2013) qu'un assembleur doit être choisi en fonction des besoins et du temps et des ressources disponibles. Un assembleur peut donner de plus grands scaffolds alors qu'un autre saura mieux gérer les erreurs, ou un autre sera plus tolérant des répétitions. Ils mettent aussi en avant que le meilleur assembleur d'aujourd'hui ne sera pas forcément celui de demain.

J'ai donc sélectionné les assembleurs suivants :

Assembleurs	Algorithme d'assemblage	Technologie de séquençage	Version utilisée	Référence
ABySS	De Bruijn	Illumina, SOLiD, 454, Sanger	1.5.2	(Simpson et al., 2009)
ALLPATHS-LG	De Bruijn	Illumina, Pacific Biosciences	-	(Gnerre et al., 2010)
Ray	De Bruijn	Illumina, 454, Sanger	v2.2.0	(Boisvert et al., 2010)
SOAPdenovo	De Bruijn	Illumina	SOAPdenovo2-src-r240	(Luo et al., 2012)
Velvet	De Bruijn	Illumina, SOLiD, 454, Sanger	1.2.10	(Zerbino and Birney, 2008)

Tableau 2. Tableau de présentation des assembleurs

ALLPATHS-LG ne sera pas utilisé dans cette étude puisqu'il requière en input deux types de librairies, une avec une taille d'insert d'environ 200pb et une autre avec une taille d'insert d'environ 3 000pb, et que pour cette partie nous n'avions qu'une librairie de reads avec petits inserts.

Ray peut combiner en entrée des reads venant de différentes technologies de séquençage.

Les quatre assembleurs sont capables de travailler sur des clusters. Ray et ABySS utilisent la technologie MPI, ils échangent les données via des messages entre les processeurs, Velvet et SOAPdenovo utilisent la librairie openMP, ils utilisent plusieurs processeurs qui partagent la même mémoire.

Concernant le choix du k-mer, j'ai fait appel à KmerGenie (Chikhi and Medvedev, 2013).
La stratégie de KmerGenie est d'utiliser le plus grand k-mer possible (pour éliminer les répétitions) de telle façon que le génome soit assez couvert par les k-mers. Pour estimer le meilleur k, KmerGenie calcule et construit des histogrammes d'abondance de chaque k-mer. L'axe des X illustre l'abondance et l'axe des Y le nombre de k-mers dont on a observé cette abondance (ou observé x fois).

J'ai par exemple deux reads : ACTCA et GTCA. Si on choisi un k=3, on aura 5 3-mers possibles : ACT, CTC, TCA, GTC et TCA. TCA est présent 2 fois, les trois autres 1 fois (Tableau 3). L'histogramme suivant est donc créé (Fig. 9) :

Abondance	Nombre de k-mers
1	3
2	1
3	0
4	0

Tableau 3. Abondance des k-mers pour notre exemple

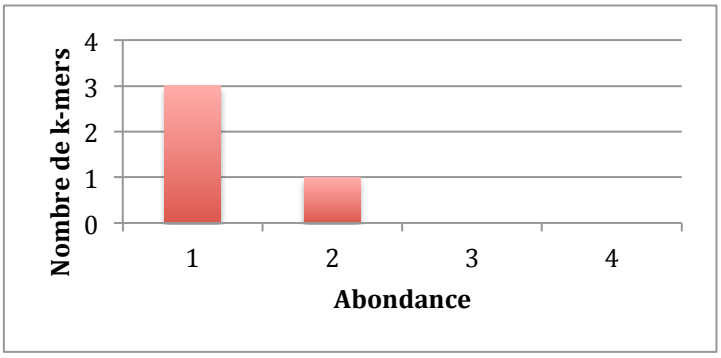


Figure 9. Histogramme d'abondance des k-mers pour notre exemple

L'histogramme diffère selon le k choisi et ressemblera à ces types de graphe issus du papier présentant KmerGenie.

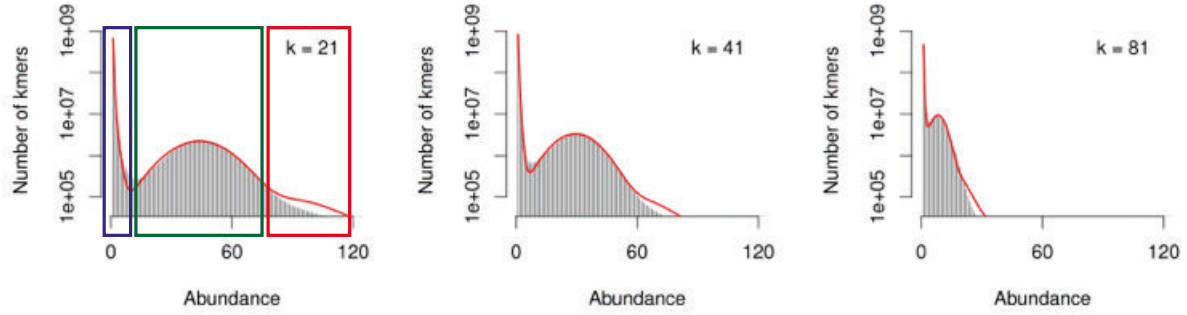


Figure 10. Histogrammes d'abondance du chr14 pour des valeurs de k de 21, 41 et 81. Chaque courbe rouge correspond au modèle statistique optimisé. (Chikhi and Medvedev, 2013)

Un histogramme est divisible en trois parties : la zone bleue contient les k-mers erronés, ils sont présent très peu de fois ou 1 fois, la zone verte contient les k-mers génomiques, c'est une aire génomique représentant le nombre de k-mer distincts permettant de couvrir le génome, la zone rouge contient les k-mers fortement répétés, principalement des artefacts de séquençage.
Un bon k-mer est alors celui qui couvre le plus possible le génome grâce aux k-mers génomiques.

Il ne faut pas oublier que KmerGenie propose un k optimal pour un set de données et qu'il n'est valable que pour celui-ci. De plus, rien n'empêche de faire quand même plusieurs assemblages avec un k-mer différent à chaque fois.
KmerGenie a aussi ses limites : il peut être dans l'incapacité de trouver un k optimal, il est pour le moment inefficace pour des données transcriptomiques et métagénomiques et des données dont la couverture n'est pas uniforme.

- Visualisation de l'alignement et des assemblages

Pour visualiser les alignements et les assemblages produits, j'ai utilisé trois outils : R, Circos et IGV.

R est un logiciel permettant des analyses statistiques et des visualisations graphiques de données.

Circos est un logiciel permettant la visualisation de données sous forme circulaire. Il est idéal pour avoir une vision globale d'alignements sur un génome.

IGV, pour Integrative Genome Viewer (Thorvaldsdóttir et al., 2013), permet quant à lui d'observer les détails des alignements : SNPs, informations paired-end. Le zoom va de quelques bases à un maximum de 99kb.

Circos ne lit pas les fichiers en format SAM générés par les logiciels d'alignement tel que Bowtie, il est donc nécessaire de modifier le format de ces données. J'utilise pour cela les différentes commandes fournies par la suite BEDTools : *genomecov* reporte la couverture du génome en nombre de reads pour chaque position, *makewindows* découpe le génome en fenêtres de x bases, *coverage* calcule la couverture pour chaque position et par fenêtre de x bases. J'ai écrit un script Perl qui permet de créer ensuite le fichier de configuration nécessaire à Circos.

Nous avons donc deux types de couverture : la couverture par base (Fig. 11) et la couverture par fenêtre de 10kb (Fig. 12).

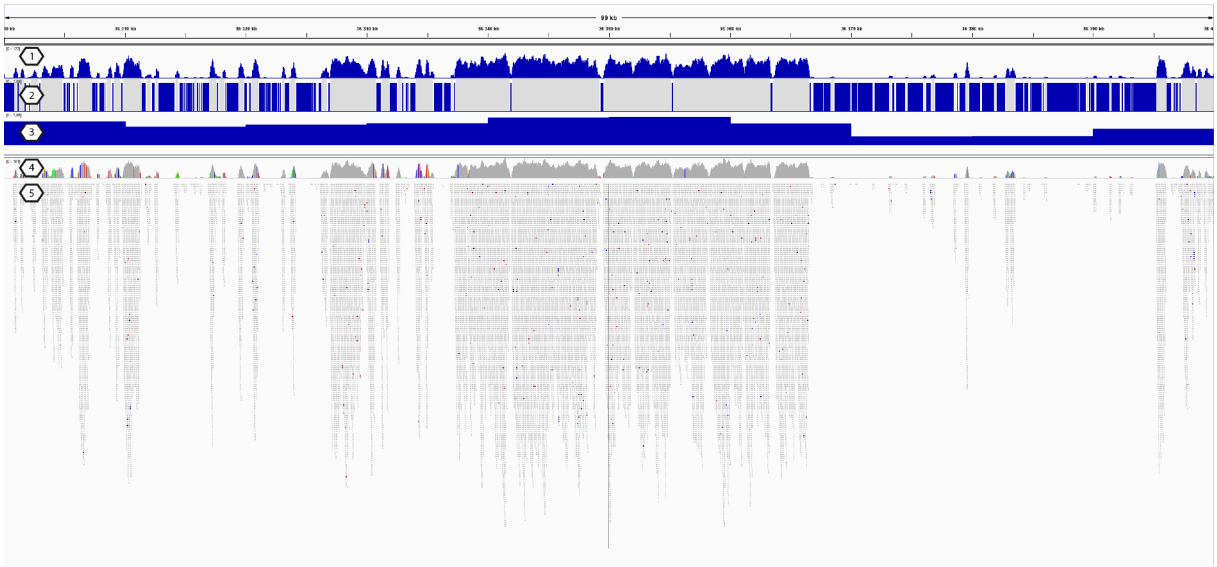


Figure 11. Visualisation par le logiciel IGV des reads alignés et de la couverture par base sur une partie du génome. (1) La couverture par base correspond au nombre de reads qui recouvrent la base. (2) Les zones bleues sont les zones avec une couverture de 0. (3) Les blocs bleus représentent la couverture exprimée en pourcentage par fenêtre de 10kb. (4) Couverture par base avec la visualisation possible des SNPs en couleur. (5) Reads alignés.

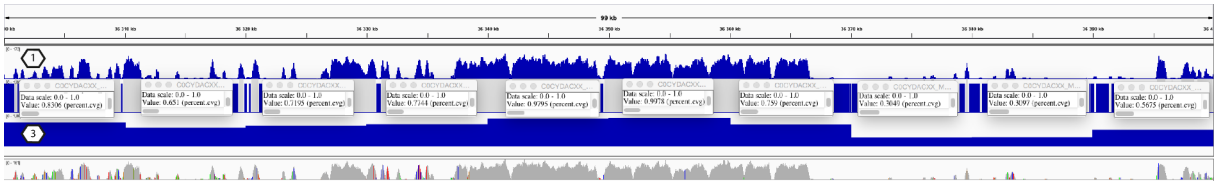


Figure 12. Visualisation par le logiciel IGV des reads alignés et de la couverture par fenêtre de 10kb sur une partie du génome. (1) La couverture par base correspond au nombre de reads qui recouvrent la base. (3) Couverture exprimée en pourcentage par fenêtre de 10kb. Entre les rangées (1) et (3), chaque pourcentage de couverture correspondant à une fenêtre de 10kb est représenté.

- Vérification des assemblages produits

Plusieurs critères sont importants afin de juger et de comparer la qualité des différents assemblages :

- le nombre de scaffolds
- la taille totale de l'assemblage
- la taille moyenne des scaffolds
- la taille du plus grand contig par scaffold
- le nombre de contigs par scaffolds
- le N50 : c'est la taille des séquences telle que toutes les séquences plus grandes couvrent 50 % de l'assemblage
- le pourcentage de bases N
- la consistance interne : c'est le pourcentage de reads correctement ré-alignés sur l'assemblage
- etc.

Certains logiciels tels que QUAST (Gurevich et al., 2013) permettent d'extraire ces différents critères à partir de l'assemblage et de produire les graphes correspondants. L'Assemblathon met à disposition un script Perl qui calcule les différentes métriques d'un assemblage.

- Détection des SNPs et insertions/délétions

Pour détecter les variants (SNPs, indels (insertions/délétions)) de notre assemblage, j'ai mis en place un pipeline alliant plusieurs outils.

Dans un premier temps, les reads sont alignés sur l'assemblage avec Bowtie2. Le prétraitement des données, en utilisant une combinaison des outils SAMtools et GATK (McKenna et al., 2010), facilite la détection des variants par la suite. Les troisième et quatrième étapes sont la détection et la filtration des variants candidats par GATK.

Les variants contenus dans les fichiers issus du pipeline sont visualisés avec les reads dans IGV.

A noter, le conseil actuel pour la détection de variants est de combiner plusieurs outils d'alignement et de détection. En effet, chaque aligneur est plus ou moins tolérant avec les SNPs et gère différemment les insertions/délétions lors de l'alignement. D'autres aligneurs tels que GSNAP (Wu and Nacu, 2010), Novoalign (Novocraft, 2010) et BWA (Li and Durbin, 2009) sont disponibles. Chaque détecteur a un seuil de détection plus ou moins sensible pour dissocier erreurs de séquençage et SNPs réels. Ces détecteurs sont réglables, par exemple en fonction de la ploïdie de l'organisme. SAMtools fournit aussi un détecteur de variants.

III. Résultats

Le séquençage Illumina de 6 moustiques de la lignée résistante R1iso2 a produit 2 fichiers de 36 Go. Ils contiennent chacun 140 millions de reads de 104pb, ce qui nous donne une couverture estimée de 102X pour notre génome de 273Mb. Le premier et le deuxième fichier se composent respectivement du premier et du deuxième read d'une paire.
Dans un premier temps, nous avons vérifié la qualité des reads de chaque fichier. Le rapport établi par FastQC n'a reporté aucun biais ou problème spécifique. La majorité des séquences (134 millions) ont un score supérieur à 30, i.e. que la précision d'identification d'une base est supérieure à 99,9%. Et 111 millions ont un score supérieur à 36, ce qui veut dire que ces bases sont encore plus fiables. Les scores de qualité de chaque base révèlent une bonne homogénéisation de la fiabilité des bases sur l'ensemble des séquences des reads (Fig. 13).

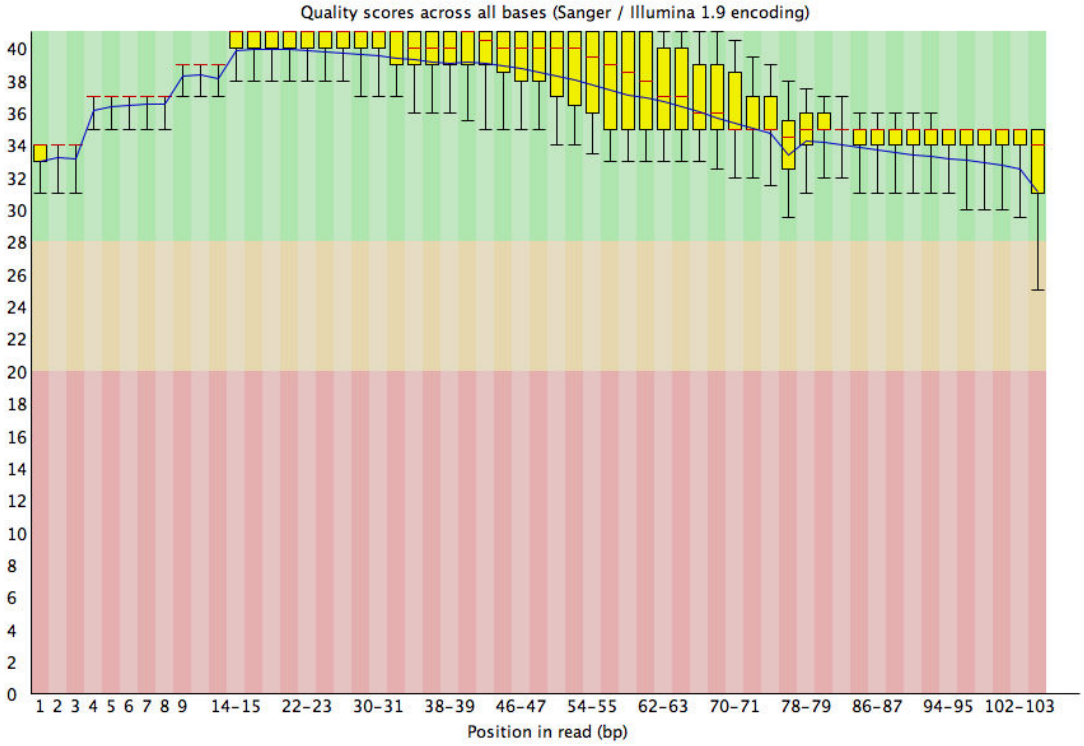


Figure 13. Graphe de qualité par base issu du rapport de FastQC pour le premier fichier. Le profil du second est similaire. Pour chaque base n de l'ensemble des reads, une boite à moustache est représentée avec les valeurs inter-quartiles (25%-75%), la médiane et les valeurs à 10% et 90%. La ligne bleue correspond à la moyenne. Les différentes parties rouge, orange et vert se rapportent respectivement aux qualités très mauvaises, moyennes et très bonnes.

Dans les séquençages Illumina, la qualité a tendance à diminuer vers la fin de la séquence. Cela est en partie dû aux difficultés vers la fin du processus de séquençage d'enlever correctement le fluorophore de la base incorporée précédente, ce qui joue ensuite sur la lecture du fluorophore de la nouvelle base incorporée.

La qualité de nos séquençages a été jugée satisfaisante au vu des statistiques mesurées sur les deux fichiers par FastQC.

1. *Alignement des reads sur le génome de référence d'Anopheles gambiae et ses haplotypes*

1.1. Résultats des alignements avec Bowtie2

Bowtie2 a été sélectionné par rapport à Bowtie pour notre alignement car (1) il gère mieux les reads ayant une taille supérieure à 50pb, (2) il prend en compte les insertions/délétions (indels) et les brèches au sein des reads lors de l'alignement, (3) il autorise les alignements sur les caractères ambigus comme les bases N.
J'ai effectué deux alignements des reads paired-end (PE) sur le génome de référence AgamP3 d'*Anopheles gambiae* accompagné de ses 166 haplotypes (Tableau 4). Pour le premier, j'ai laissé les options d'alignement par défaut, i.e. un alignement global sensitif. Pour le second, j'ai changé la taille du fragment autorisé. Un fragment comporte la taille des deux reads appariés et la distance entre eux. Par défaut, la taille maximum du fragment aligné est à 500b, je l'ai augmenté à 20kb afin d'observer si certaines paires se retrouvent de part et d'autre d'une grande région N par exemple. Cela voudrait alors dire que cette région contenant un grand nombre n de N n'en contiendrait en fait qu'environ 200 qui est la taille moyenne de la distance entre deux reads appariés dans notre échantillon. Une autre règle importante pour les deux alignements est la règle FR, i.e. que le read 1 doit s'aligner d'abord dans le sens Forward puis que le read 2 doit s'aligner après dans le sens Reverse. En fait, l'orientation des reads de la paire doit être inversé dans l'alignement comme ils ont été séquencés. Dans le cas d'une inversion, les reads apparaitront dans le même sens après l'alignement.

1.1.1. Proportion de reads appartenant au génome

	Mapping_1	**Mapping_2**
Options	- Taille maximum de fragment = 500b - FR	- Taille maximum de fragment = 20kb - FR
Taux d'alignement total	90,48%	90,49%
Alignés en paires conformément aux options*	87,65%	87,76%
Alignés en paires non conformes aux options*	0,11%	0,07%
Alignés seuls	2,73%	2,66%

Tableau 4. Statistiques des alignements de Bowtie 2 sur le génome AgamP3. * Les reads qui se sont alignés en paires sont classés dans deux catégories par Bowtie2. La première catégorie contient les reads PE qui, en s'alignant, ont respecté la règle de la taille maximale du fragment et la règle du FR. La seconde catégorie classe les reads PE qui n'ont pas réussi à s'aligner selon la règle du FR (le read 2 s'est aligné avant le read 1) et/ou selon la règle de la taille maximale du fragment (les reads appariés se sont alignés avec une taille de fragment supérieure à celle de la règle).

En augmentant la taille maximale du fragment autorisée, 0,07% des reads PE s'alignant seuls avant s'alignent maintenant en paires non conformes et 0,09% des reads PE alignés qui étaient non conformes sont maintenant conformes. Nous avons encore 0,07% de reads PE alignés et non conformes qui seront intéressants à étudier pour la recherche de fragments dont la taille est supérieure à 20kb et de variants structuraux dans le cas des reads appariés inversés.

Nous avons obtenu 90,49% des reads s'alignant sur le génome. Parmi les 9,5% de reads non alignés, 0,68% se sont alignés sur les transcrits d'*Anopheles gambiae* dont 0,67% se sont alignés en solitaire. En fait, pour ces cas-là, les transcrits sont mieux assemblés que les gènes dans le Golden Path.
Il nous restait encore un peu moins de 9% de reads PE inconnus. Plusieurs explications sont possibles : ils correspondent (1) a des régions hyper-polymorphiques ou (2) a des brèches dans le génome, ou des régions qui n'ont pas été séquencées, (3) ce sont des erreurs de séquençage, ou (4) ils appartiennent a d'autres espèces que le moustique, et notamment au microbiote du moustique. Nous reviendrons sur ce dernier point dans le chapitre 3 qui traite de la détection des micro-organismes.
Il est possible de vérifier le point (1) en alignant ces reads sur les génomes de la forme S d'*A. gambiae* et d'*A. coluzzii*, l'ancienne forme M. Dans les 9%, Bowtie2 a aligné 52,19% sur la forme S et 35,96% sur la forme M. Ce qui confirme le fort taux de polymorphisme partagé entre ces 3 génomes d'Anopheles.

1.1.2. Vérification de la taille des fragments séquencés

La distribution du nombre de bases situées entre deux read d'une même paire a été calculée par Picard et représenté par l'histogramme suivant (Fig. 14). Cela permet de vérifier la concordance entre la taille du fragment choisie pour le séquençage et la taille du fragment réellement séquencée.

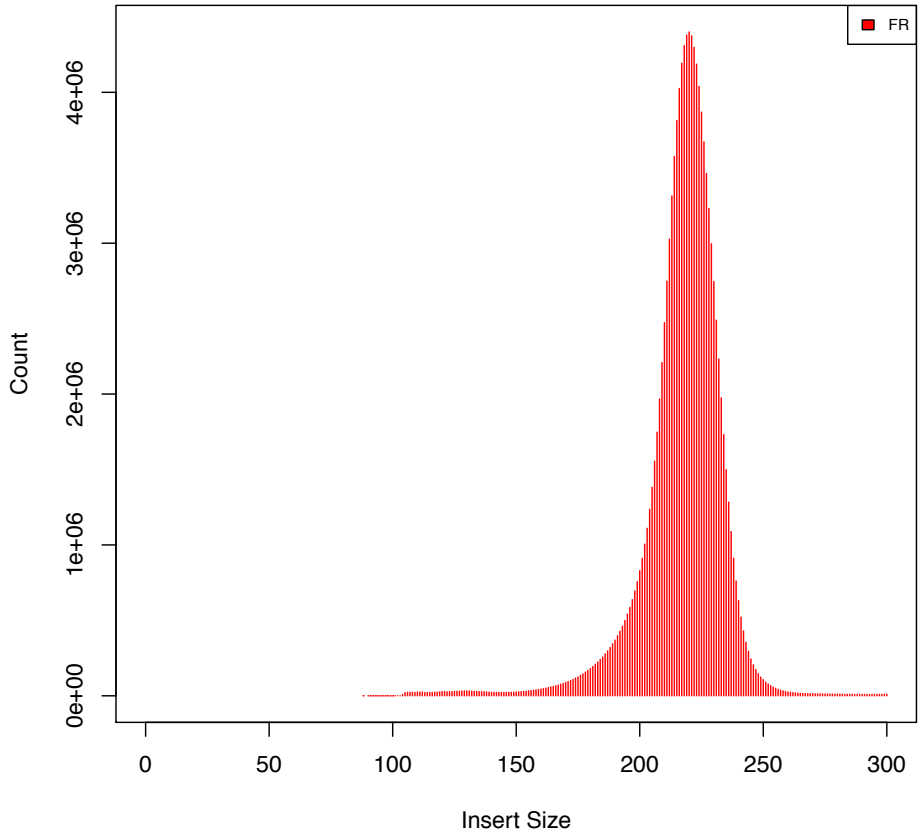


Figure 14. Distribution de la taille des inserts des reads appariés alignés. Normalement, l'insert contient les deux reads de la paire et la séquence contenue entre les deux mais ici le programme défini l'insert comme la partie située entre les deux reads de la paire.

Après l'alignement, on observe des distances entre deux reads appariés conformes comprises entre 88 et 59 842 814 bases avec une médiane à 220 bases +/- 8 bases. La majorité se situant autour de 220b, cela confirme la taille de nos fragments séquencés qui étaient de 400bp. Les plus grandes distances observées sont en fait celles des reads appariés qui se sont alignés chacun de leur côté, souvent l'un sur un chromosome et l'autre sur un autre chromosome. Quand on enlève les reads alignés en solitaire et qu'on ré-effectue les mesures, la distance maximum entre deux reads est de 20 018b, ce qui correspond, environ, au maximum que j'avais autorisé lors du second alignement.

1.2. Couverture des bases des chromosomes

D'après les fichiers d'alignement, il a été mesuré le nombre de bases couvertes par minimum un read (Tableau 5). Chaque chromosome présente une couverture sur plus de 90% des bases quelle soit A, T, C, G ou N. Les bases N sont en fait des bases indéterminées lors de l'assemblage du génome.

Chr.	Nombre de bases total	Pourcentage de bases couvertes (ATCGN)	Nombre total de bases N	Nombre de bases ATCG	Pourcentage de bases couvertes (ATCG)
2R	61 545 105	92,35 %	1 412 652 2,3 %	60 132 453	94,52 %
2L	49 364 325	96,58 %	838 578 1,7 %	48 525 747	98,05 %
3R	53 200 684	94,17 %	974 116 1,8 %	52 226 568	95,92 %
3L	41 963 435	90,39 %	1 204 962 2,9 %	40 758 473	93,06 %
X	24 393 108	94,97 %	1 007 759 4,1 %	23 385 349	99,05 %

Tableau 5. Statistiques du nombre de bases de chaque chromosome couvertes par au minimum un read lors de l'alignement.

93,7% de l'ensemble du génome a été recouvert par au minimum un read et sur ces 93,7% de bases, il a été mesuré une profondeur médiane de couverture de 101X. Ce qui corrobore notre estimation de 102X calculée d'après nos 280 millions de reads. Par ailleurs, il reste 6,3% des bases qui n'ont aucun read aligné. Nous savons que 2,5% des bases du génome sont des bases N. Certaines ont des reads alignés (Fig. 15) et cela représente 0,011% des 273Mb du génome, soit 31 052b.

Figure 15. Visualisation par IGV d'une zone de transition entre une séquence et une zone de N. Quelques bases de la fin de plusieurs reads s'alignant sur la séquence d'avant continuent de s'aligner sur les bases N.

En ne comptabilisant pas les bases N dans chacun des chromosomes, nous obtenons une moyenne de 96,12% de bases ATCG couvertes par au minimum un read.
Il nous reste donc environ 3,88% de bases sans alignement de reads. Pour comprendre la raison de ces zones de non-alignements et étudier la répartition des reads, nous avons réalisé des graphes circulaires à l'aide de Circos nous permettant de visualiser la couverture sur chaque chromosome et ses haplotypes.

1.2.1. Visualisation

Le prochain graphe représente la visualisation circulaire produite par Circos pour le chromosome 2R. Les graphes Circos des chromosomes 2L, 3R, 3L et X sont les annexes I, II, III et IV.

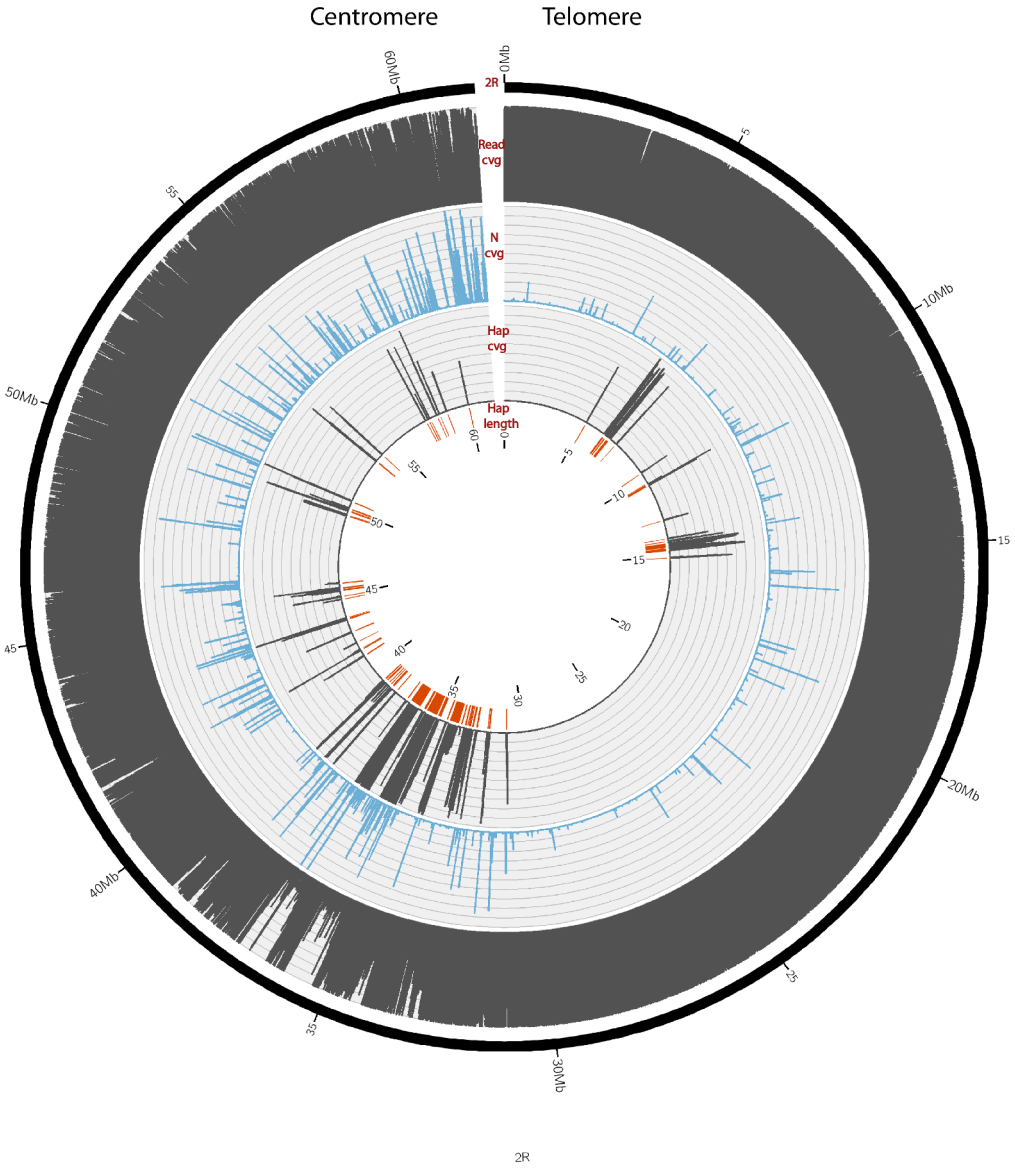


Figure 16. Visualisation circulaire par le logiciel Circos de la répartition des reads alignés sur le chromosome 2R et ses haplotypes. La lecture se fait par couche concentrique de l'extérieur vers l'intérieur : (1) la séquence du chromosome de référence, (2) la couverture en nombre de reads alignés sur le Golden Path par fenêtre de 10kb, (3) la proportion de bases indéterminées N par fenêtre de 10kb, (4) la couverture en nombre de reads alignés sur les haplotypes par fenêtre de 10kb et (5) la position et la taille des haplotypes. Pour chaque couche concentrique, l'échelle est comprise entre 0 et 100% et chaque graduation correspond à une augmentation de 10%.

1.2.2. Statistiques

Les statistiques présentées ci-après sont la base des graphes Circos.

Chr.	Couverture moyenne (%)	Couverture minimale (%) (# fenêtres de 10kb)	Couverture maximale (%)	Couverture moyenne région centromérique (5Mb) (%)	Couverture moyenne région télomérique (5Mb) (%)
2R	94,52	0 (4)	100	87,12	98,72
2L	98,05	0 (8)	100	90,34	98,82
3R	95,92	0 (9)	100	86,80	98,60
3L	93,06	0 (13)	100	85,45	87,38
X	99,05	0 (3)	100		

Tableau 6. Statistiques de couverture des reads sur les chromosomes du génome de référence. Les bases N ne sont pas prises en compte.

Chr.	Nombre total de bases N	Pourcentage en nombre de bases du chromosome	Nombre de régions contenant des N	Taille moyenne (en base)	Taille minimale (en base)	Taille maximale (en base)
2R	1 412 652	2,3%	1 658	851	19	36 426
2L	838 578	1,7%	957	875	19	28 883
3R	974 116	1,8%	1 128	862	19	24 291
3L	1 204 962	2,9%	1 272	946	19	31 062
X	1 007 759	4,1%	1 287	782	19	21 131

Tableau 7. Statistiques des bases N des chromosomes du génome de référence.

Chr.	Nombre d'haplotypes	Taille totale	Couverture moyenne (%)	Nombre de bases N	Couverture moyenne (sans les N, %)
2R	70	5 061 155	55,32 %	981 872	75,45
2L	8	290 959	32,89 %	92 732	59,32
3R	34	2 441 208	48,88 %	494 986	67,40
3L	54	3 996 177	57,11 %	872 997	80,58

Tableau 8. Statistiques de couverture des reads sur les haplotypes.

La première constatation observée sur les graphes Circos des chromosomes 2 et 3 (Fig. 16 et Annexes I à IV), c'est l'augmentation du nombre de N lorsqu'on s'approche du centromère, ce qui impacte le nombre de reads alignés. Pour l'ensemble des 4 régions centromériques, nous avons une moyenne de 87,42% de bases couvertes alors qu'elle est de 93,33% pour les deux chromosomes entiers (Tableau 6). Ces régions, riches en séquences répétées sont généralement très difficiles à séquencer et à assembler.

Concernant le chromosome 2R, sa couverture est très bonne avec une moyenne de 98,39% sur les 30 premières Mb, puis elle diminue à 86,61% sur le reste. Sur le graphe (Fig. 16), cette baisse du nombre de bases couvertes s'explique par un nombre de bases N plus conséquent et la présence de plusieurs haplotypes, 14 dans la première partie et 56 dans la seconde. C'est d'ailleurs le bras 2R qui cumulent le plus d'haplotypes avec une taille totale à 5Mb. De plus, les haplotypes concentrés dans la région 35-39,2Mb, profitent d'une meilleure couverture que le Golden Path, soit respectivement 75,31% pour 62,76% (Fig. 16). Cet exemple illustre une nouvelle fois que le Golden Path (GP) ne représente qu'une « version » du génome dans les régions hyper-polymorphiques, et que les haplotypes sont, dans certains cas, plus proches de la séquence de notre lignée que le GP. Le chromosome 2R contient aussi le plus grand nombre, 1 658, et les plus grandes régions N. La taille maximale d'une succession de N est de 36kb (Tableau 7), ce qui explique la moins bonne couverture générale (92,36%) que les autres.
C'est le chromosome 2L qui montre la meilleure couverture avec 96,38% (Tableau 6). Il y a plus de bases couvertes sur le 2L que les autres car c'est celui qui possède le moins d'haplotypes, seulement 8, et le moins de bases N, soit 1,7%.
Par contre, le chromosome 3L présente la moins bonne couverture à seulement 90,4% et elle est plutôt homogène sur l'ensemble (Tableau 6, Annexe III). Il possède une forte proportion de N, 2,9% mais ce n'est pas lui qui a la plus forte proportion (Tableau 7). Cependant, il a 54 haplotypes qui se partagent en 4 régions condensées (Tableau 8, Annexe III) et certains haplotypes sont plus couverts que le GP comme nous avons vu sur le chr. 2R.
Pour le chromosome 3R, nous avons une couverture moyenne de 94,18%, une faible proportion de N, 1,8%, et 34 haplotypes rangés en 4 régions (Annexe II, Tableaux 6, 7, 8). Même si on ne le voit pas sur le graphe Circos, la couverture est parfois meilleure sur les haplotypes que sur le GP.
Le chromosome X présente une couverture assez moyenne de 94,97% dû majoritairement aux 4,1% de bases N (Tableau 7, Annexe IV). Ce chromosome n'a pas d'haplotypes.

1.3. Origines d'une mauvaise couverture

Nous avons observé deux raisons principales qui expliquent les diminutions dans la couverture du génome et dans les 37 fenêtres de 10kb qui n'ont aucun read aligné. Premièrement, on a remarqué que le plus souvent, une diminution de la couverture et une augmentation des Ns étaient parallèles (Fig. 17). Deuxièmement, on a observé des diminutions lors de la présence d'haplotype où les reads s'alignent préférentiellement (Fig. 18).

Figure 17. Visualisation par le logiciel IGV des reads alignés sur le chromosome 2L entre les bases 10 000 à 70 000. Nous avons 6 fenêtres de 10 kb chacune, représentées par les flèches rouges. Les pourcentages de bases couvertes et de bases N dans chaque fenêtre sont indiqués au-dessus des flèches.

Les fenêtres 2, 3 et 4 situées dans la figure 17 sont un bon exemple de la cause de la baisse de la couverture. Pour la n°2, 56% des bases sont des N, les 44% restants, des bases ATCG, ont été couvertes par des reads. Quand la somme de ces deux pourcentages n'arrivent pas à 100%, c'est qu'il y a des bases ATCG non couvertes.

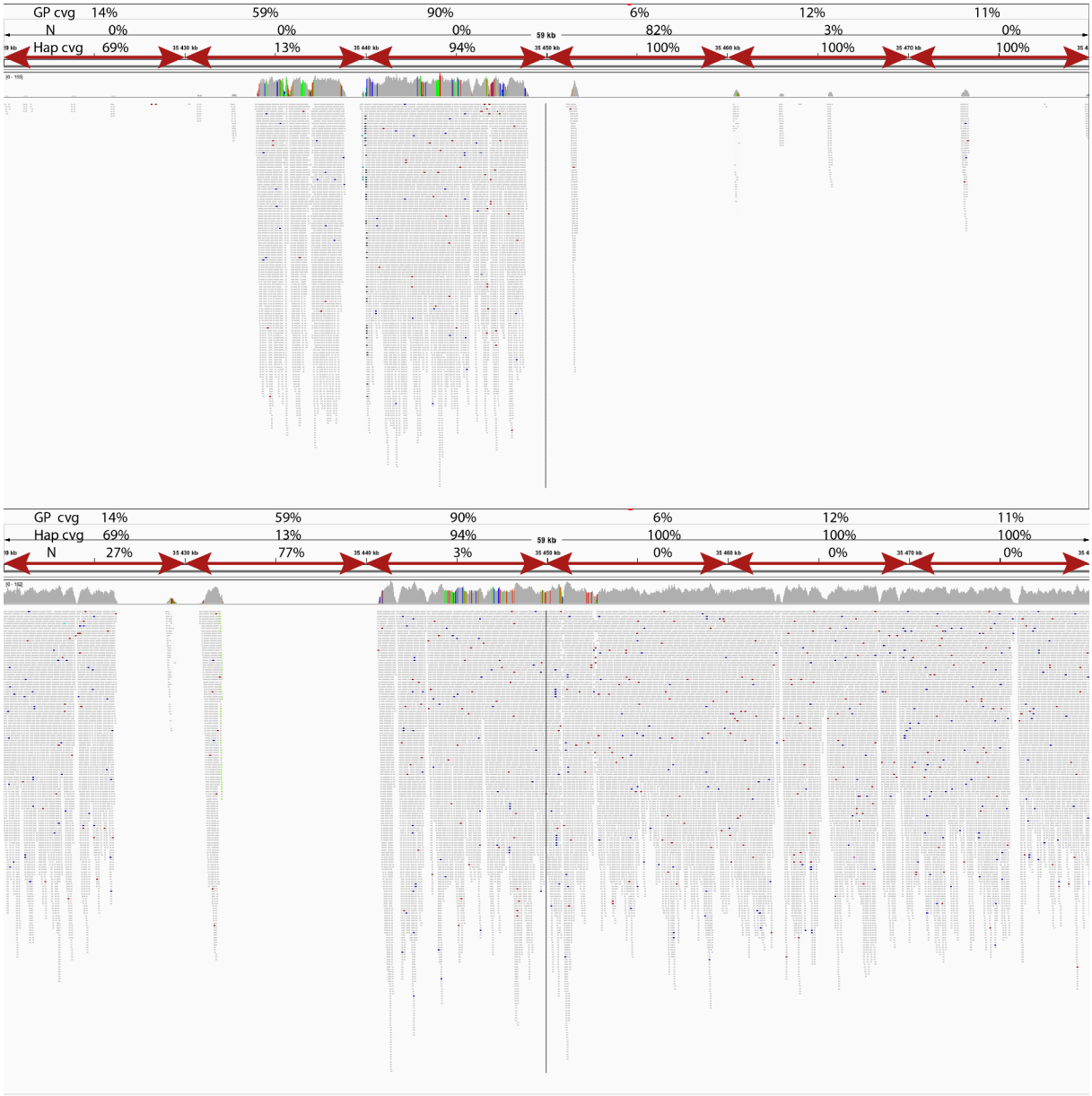


Figure 18. Visualisation par le logiciel IGV des reads alignés sur le chromosome 2R et sur son haplotype 2R_hap_32 entre les bases 35 420 000 à 35 480 000. Nous avons 6 fenêtres de 10 kb chacune, représentées par les flèches rouges. Les pourcentages de bases couvertes sur le GP, de bases N et de bases couvertes sur l'haplotype dans chaque fenêtre sont indiqués au-dessus des flèches.

Dans cet exemple (Fig. 18), la couverture est meilleure sur l'haplotype que sur le GP pour les fenêtres 1, 3, 4, 5 et 6 de 10kb. La chute de la couverture sur l'haplotype de la fenêtre 2 est due à la présence de N. La fenêtre 3 présente une bonne couverture que ce soit sur le GP ou sur l'haplotype, cependant la séquence du GP n'est pas la même que celle de l'haplotype. D'ailleurs, les reads alignés montrent un fort taux de SNPs colorés que ce soit sur le GP ou l'haplotype.

Cependant, plusieurs régions de 10kb ont montré des couvertures faibles qui ne s'expliquent ni par la présence de N, ni par celle d'un haplotype : 361 régions de 10kb présentent un taux de bases couvertes par soit une base ATCG, soit une base N qui n'excède pas les 70% et 137 présentent un taux inférieur à 50%.
Pour comprendre pourquoi très peu de bases sont couvertes, nous avons examiné les régions dont le pourcentage des bases couvertes et des bases N n'excédait pas 20%. Nous avons reporté ci-après les différents cas observés.

- Exemple 1

Cette première zone couverte sur 10,27% des bases est entourée de nombreuses portions de bases N (Fig. 19). Dans les 30kb de part et d'autre, le nombre de bases couvertes et le nombre de bases N se compensent. Ce premier exemple montre bien la mauvaise qualité de cet assemblage AgamP3 sur cette portion.

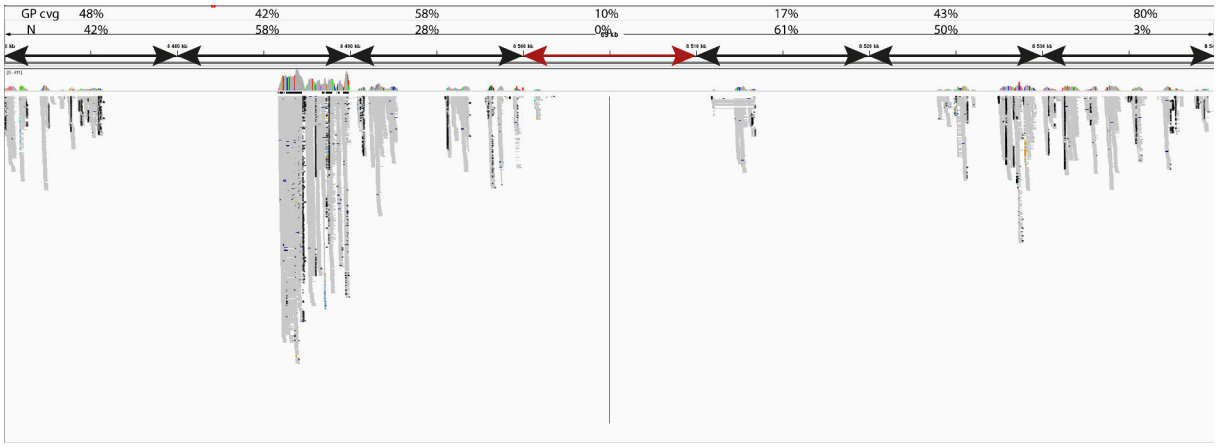


Figure 19. Visualisation par le logiciel IGV d'une zone de 10kb (signalée par la flèche rouge au centre) faiblement couverte. La zone est située sur le chromosome 2L entre 8,5 et 8,51Mb. Elle est entourée de 30kb de part et d'autre.

- Exemple 2

La deuxième zone a très peu de bases couvertes (6,73%). La couverture autour, qui est très bonne, chute considérablement aux abords de la zone alors qu'il n'y a aucune base N (Fig. 20). Les quelques pics présents dans cette région fantôme de 19,7kb comprennent en majorité des mauvais alignements : soit les reads sont seuls, soit ils sont alignés mais présentent beaucoup de SNPs et de petits indels. Le fait qu'il n'y ai aucun read alignés parfaitement dans cette région peut indiquer deux cas : cette région n'existe pas dans notre lignée ou c'est un artéfact de l'assemblage.

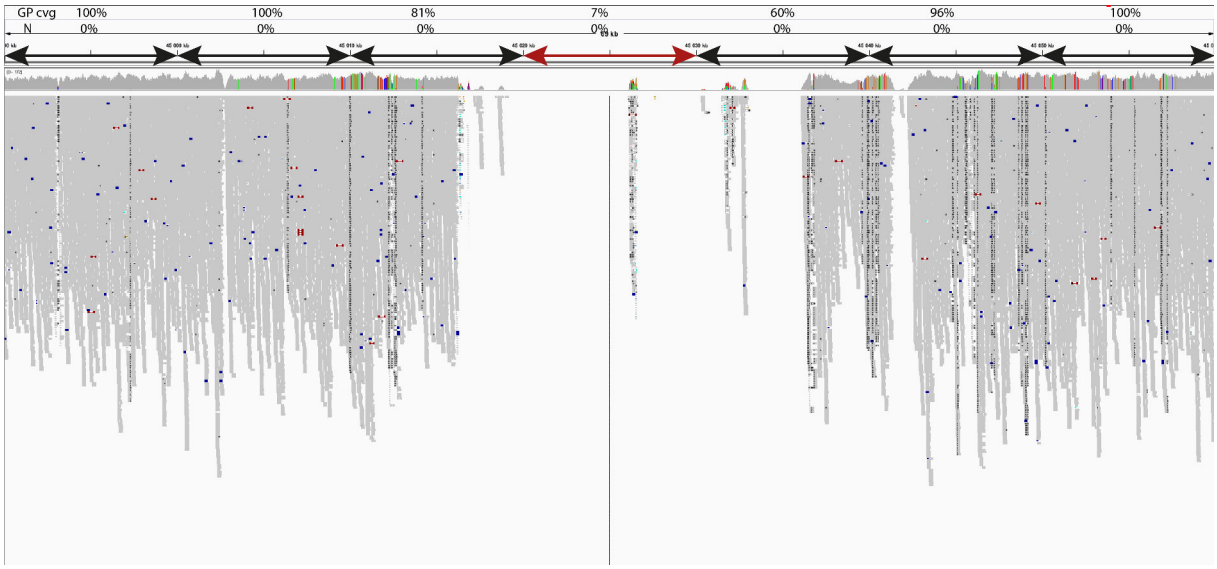


Figure 20. Visualisation par le logiciel IGV d'une zone de 10kb (signalée par la flèche rouge au centre) faiblement couverte. La zone est située sur le chromosome 2L entre 45,02 et 45,03Mb. Elle est entourée de 30kb de part et d'autre.

- Exemple 3

Une autre zone de 10kb, cette fois-ci sur le chromosome 2R, nous montre une forte concentration de petits indels et de SNPs autour (Fig. 21). En fait, cette région très polymorphique compte 2 haplotypes de part et d'autre de notre zone de 10kb. Dans ce cas, sur les 30 premiers kb qui correspondent à la fin d'un haplotype, la couverture et l'alignement sont meilleurs sur l'haplotype. La fin de l'haplotype qui ressemble donc plus à notre lignée signifie la fin d'une bonne couverture et de bons alignements. Il y a ensuite des alignements mais qui sont très imparfaits jusqu'à un trou de 1kb pourtant sans N. En fait, les reads de ces 1kb ont tous aligné sur le kb correspondant dans l'haplotype. Nous pouvons donc supposer que notre zone de 10kb comme ses alentours ne reflète pas le génome de notre lignée.

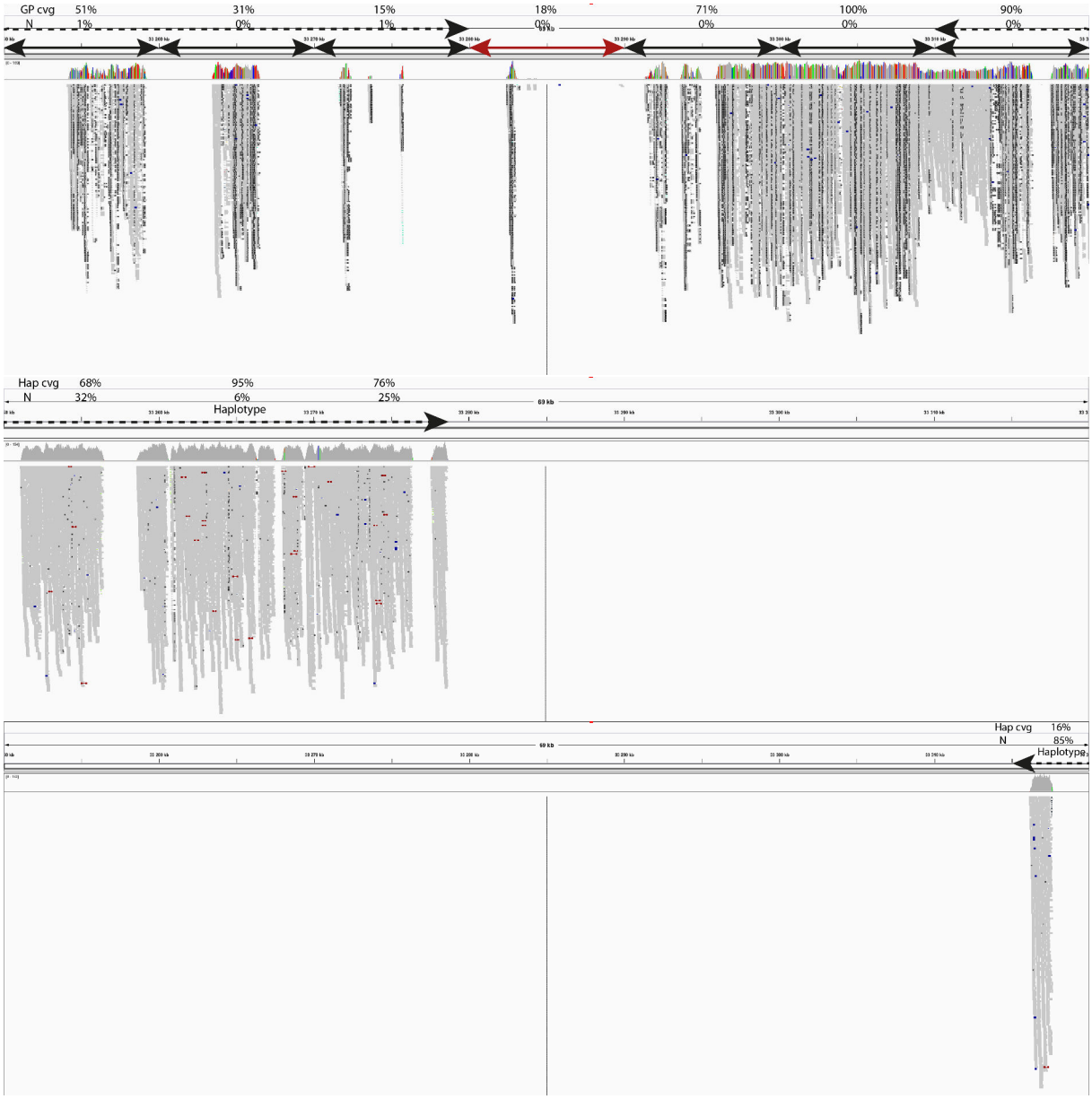


Figure 21. Visualisation par le logiciel IGV d'une zone de 10kb (signalée par la flèche rouge au centre) faiblement couverte. Les flèches en pointillés représentent la position des haplotypes par rapport au GP. La zone est située sur le chromosome 2R entre 33,28 et 33,29Mb. Elle est entourée de 30kb de part et d'autre. La deuxième et la troisième partie de l'image montre l'alignement des reads sur les haplotypes qui entoure notre zone de 10kb.

- Exemple 4

Dans le prochain cas, nous avons aussi une fin d'haplotype où la couverture était très bonne (Fig. 22). Cependant, les 25kb situés après ne sont quasiment pas couverts malgré un faible taux de N dans cette région. L'avant dernière fenêtre de 10kb possède 6 000 bases N qui peuvent signifier une jointure entre deux contigs assemblés. Nous pouvons conclure que le contig qui a été inséré dans le Golden Path ne concorde pas du tout avec notre lignée mais plutôt le contig qui a été inséré dans l'haplotype.

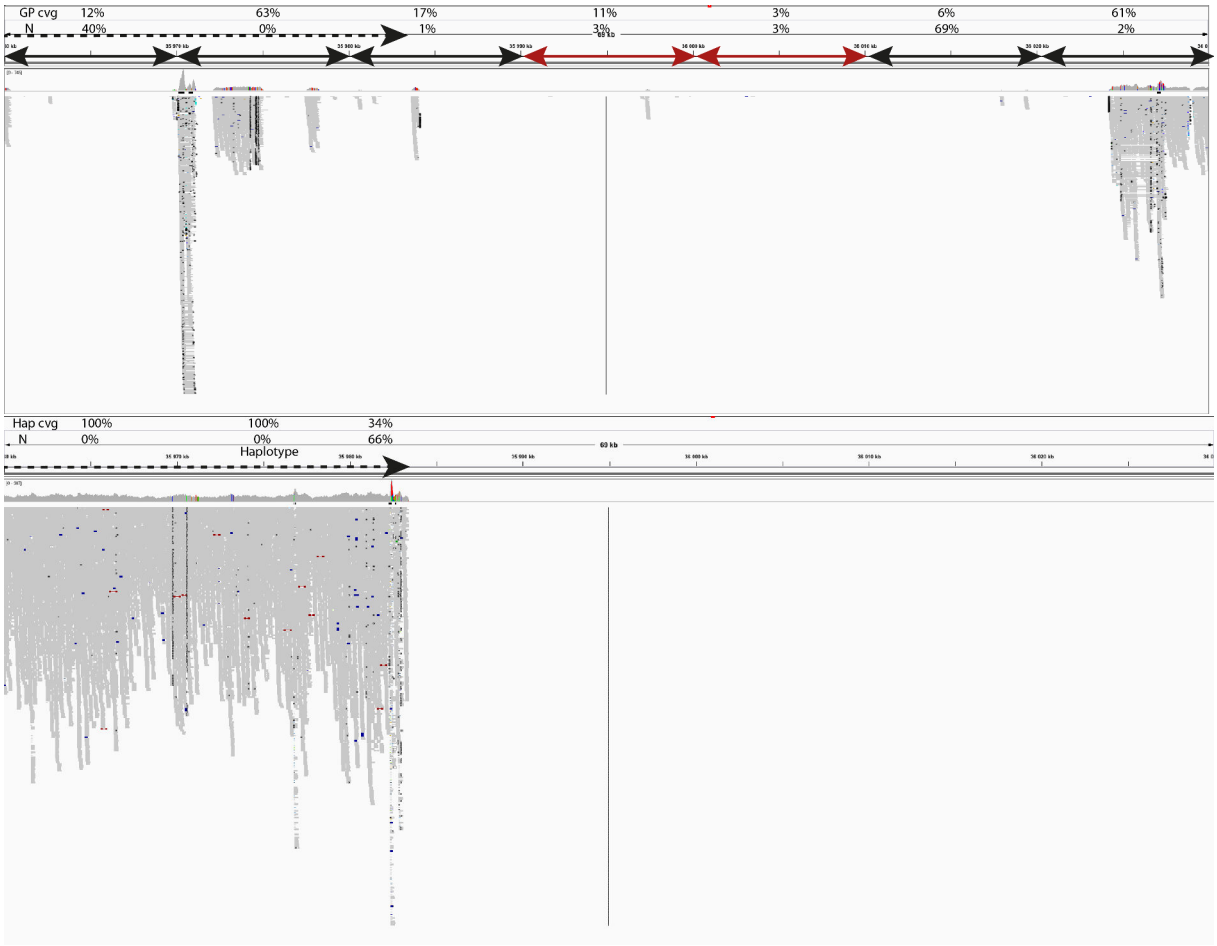


Figure 22. Visualisation par le logiciel IGV de 2 zones de 10kb concomitantes (signalées par les flèches rouges) faiblement couvertes. La flèche en pointillés représente la position de l'haplotype sur le GP. Les zones sont situées sur le chromosome 2R entre 35,99 et 36,1Mb. Elles sont entourées de 30kb à gauche et de 20kb à droite. La deuxième partie de l'image montre l'alignement des reads sur l'haplotype.

- Exemple 5

Nous avons aussi observé des zones où les reads PE étaient séparés par une trop grande distance (Fig. 23). Dans ce cas, cela signifie qu'il y a eu une délétion dans notre lignée. Dans cet exemple, il n'y a pas du tout d'haplotypes. Par contre, on observe beaucoup de petites insertions/délétions de part et d'autre de la délétion de 30kb. Il y a aussi une zone ambiguë à droite qui présente des reads PE séparés par 3kb qui sont en fait très bien couverts. Dans ce cas, il existe probablement 2 formes différentes dans la lignée R1iso2 ; une avec la délétion et une identique au GP.

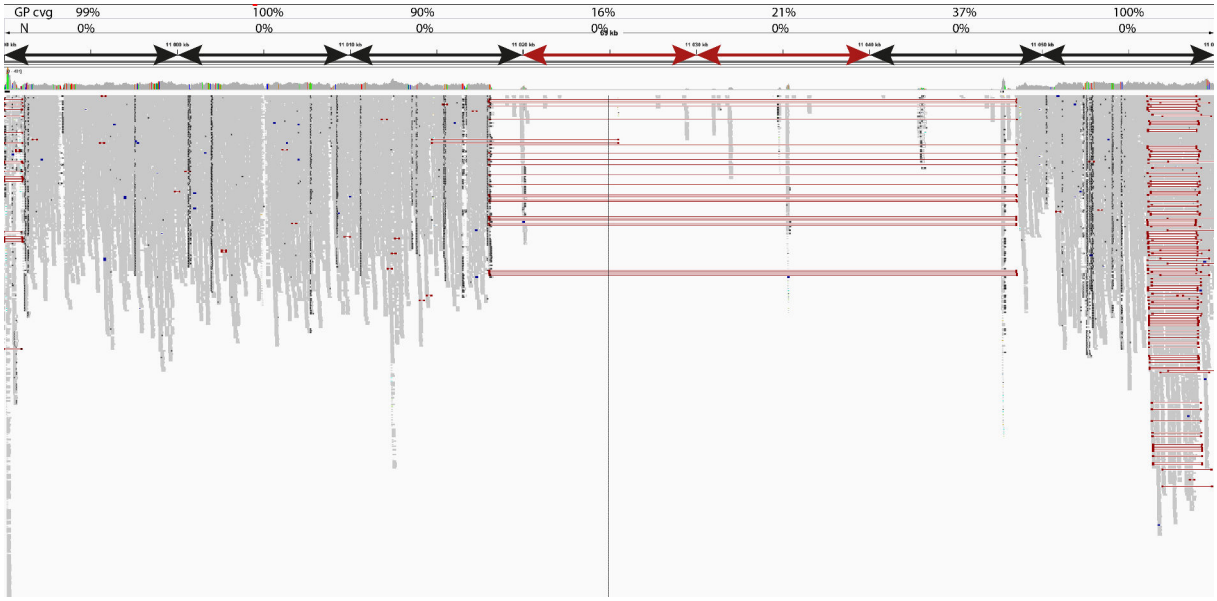


Figure 23. Visualisation par le logiciel IGV de 2 zones de 10kb concomitantes (signalées par les flèches rouges) faiblement couvertes. Les zones sont situées sur le chromosome 3R entre 11,02 et 11,04Mb. Elles sont entourées de 30kb à gauche et de 20kb à droite.

- Exemple 6

Une autre zone de 10kb faiblement couverte est ici entourée de régions apparaissant comme très polymorphiques, avec beaucoup de N, quelques indels, quelques insertions (Fig. 24). Cette zone montre à elle seule la mauvaise qualité du génome et la non représentativité de notre lignée.

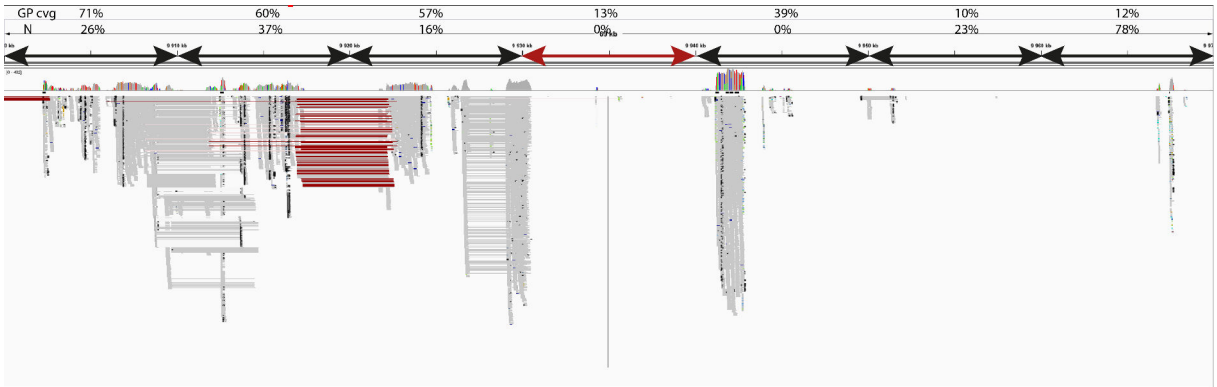


Figure 24. Visualisation par le logiciel IGV d'une zone de 10kb (signalées par la flèche rouge) faiblement couverte. La zone est située sur le chromosome 3L entre 9,93 et 9,94Mb. Elle est entourée de 30kb de part et d'autre.

De plus, en détaillant le pool de reads alignés juste après notre zone faiblement couverte au centre, je me suis rendu compte que quelques reads de début et de fin de cette région s'alignaient seul. Il s'est avéré que leurs paires s'alignaient soit sur le chromosome 2R (noir), soit sur le chromosome 3R (orange) (Fig. 25). En fait, cette région de 1 700b qui est localisée sur le chromosome 3L pourrait (1) s'intégrer en s'inversant dans le chromosome 2R quelque part entre les bases 1 555 800 et 1 555 900 mais pourrait aussi (2) s'intégrer tel quel dans le chromosome 3R quelque part entre les bases 18 482 800 et 18 482 900 environ (Fig. 25). De plus, en récupérant la séquence de 1700b et en la BLASTant sur le génome d'*A. gambiae*, il s'est avéré qu'elle s'alignait avec des pourcentage de similarité allant de 88,8% à 95,7% sur des portions d'environ 1700b sur les chromosomes : 3R (6 régions différentes), 3L (5 autres positions), 2R (4 régions différentes), X (1 région) et UNKN (3 régions différentes). Il s'agit donc d'une région répétée.

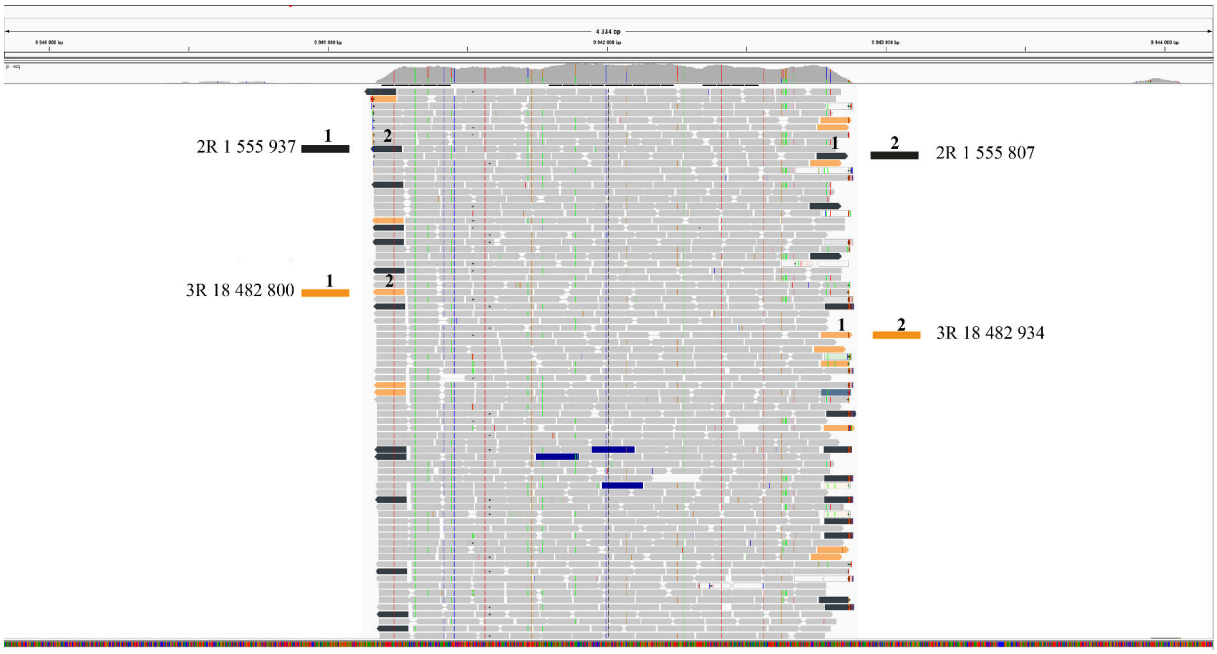


Figure 25. Zoom IGV sur le pool de reads alignés à droite dans la Figure 24. Les reads colorés en bleu ont une taille d'insert trop petite, les autres couleurs symbolisent le chromosome sur lequel l'autre read de la paire s'est aligné : les reads orangés ont leur read correspondant sur le chromosome 3R, les reads noirs sur le chromosome 2R. Les informations de part et d'autre du pool de reads sont les exemples montrant le premier ou le deuxième read d'une paire situé sur le chromosome spécifié et à la position indiquée.

1.4. Conclusion

Nous avons d'abord vu que la couverture dépendait fortement de la présence de N ou des haplotypes sur lesquels nos reads s'alignaient mieux parfois. Puis, nous venons de voir avec les zones de 10kb qui ne contenaient que très peu de bases N (< 20%), aucun haplotype et pourtant très peu de reads alignés (< 20%) qu'il y a plusieurs explications au faible alignement des reads : (1) des mauvais assemblages qui provoquent des duplications, inversions, insertions, délétions, (2) un fort taux de polymorphisme, (3) un réarrangement chromosomique différent dans le génome de notre lignée. Ces exemples pris représentent des cas extrêmes de faible couverture, cependant on retrouve les mêmes causes qui engendrent la faible couverture sur des zones un peu plus couvertes. En comptabilisant les fenêtres de 10kb qui montre une couverture en reads inférieure à 70% qui n'est pas compensée par une couverture en bases N ou un haplotype, on en trouve 719, soit environ 3% des chromosomes 2, 3 et X. Toutes ces régions ATCG non couvertes ou très peu/mal couvertes en plus des régions N sont non utilisables pour la recherche de polymorphismes.

Il est donc possible de trouver des marqueurs génétiques à partir de l'alignement des reads issus de notre lignée R1iso2 sur le génome de référence, que ce soit sur le Golden Path ou sur les haplotypes, dans les régions non problématiques. Cependant, il existe de nombreuses régions mal couvertes dont certaines des régions ayant un lien avec la résistance du moustique aux parasites (exemple de la région des gènes TEP sur le chromosome 3L) où il sera impossible de spécifier des marqueurs génétiques et encore moins de lister les polymorphismes.
Cela nous a démontré la nécessité de séquencer et d'assembler le génome de nos propres lignées pour mener à bien les génotypages des moustiques sensibles et résistants et nos recherches de régions liées à la résistance au parasite du paludisme.

2. Assemblage de novo du génome de la lignée R1iso2

2.1. Sélection du k-mer et assemblages

La première étape était la détermination de la taille du k-mer. Ce dernier, ayant un rôle décisif dans la formation des scaffolds, peut être estimé par KmerGenie. Les graphes issus de KmerGenie montrent l'abondance des différents k-mers formés selon la taille k choisie (Fig. 26).

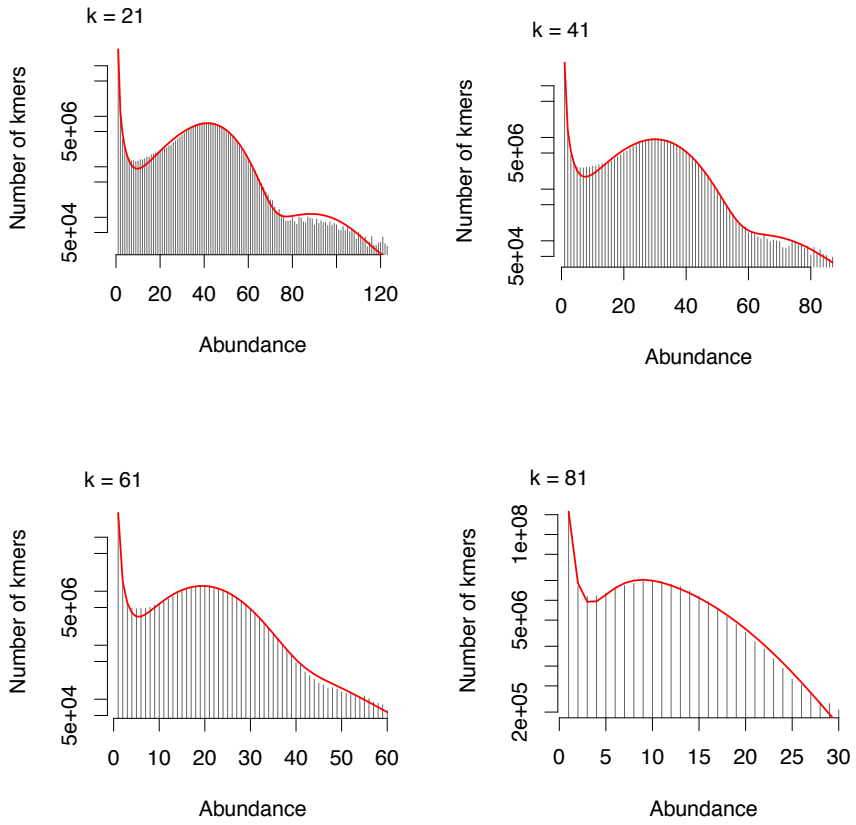


Figure 26. Histogrammes d'abondance pour des valeurs de k de 21, 41, 61 et 81 pour nos données.

KmerGenie a déterminé que le k optimal pour nos données est de 61 puisqu'il présente la plus grande abondance de k-mers génomiques différents disponibles pour assembler les données brutes.
Les premiers assemblages avec les logiciels ABySS, Ray, SOAPdenovo et Velvet ont été faits avec des 61-mers.

	ABySS	Ray	SOAPdenovo	Velvet
Temps d'exécution	3h11	1j16h26	2h10	3h08

Tableau 9. Temps d'exécution des différents assembleurs utilisés

Chaque assemblage a été réalisé sur l'HPC sur 1 nœud à 16 cœurs avec une mémoire disponible de 256Go. C'est SOAPdenovo le plus rapide, suivent Velvet, ABySS et Ray (Tableau 9).
Au niveau de la mémoire utilisée, Velvet, ABySS et SOAPdenovo ont demandé une taille de mémoire supérieure à 180 Go. Ray fonctionne avec très peu de mémoire disponible comparé aux trois autres.

2.2. Évaluation des assemblages produits

Pour estimer la qualité des assemblages produits, on peut se baser sur plusieurs critères. Les plus importants ont été regroupés dans le tableau suivant (Tableau 10). Un bon assemblage est un assemblage avec le moins de scaffolds possible tout en étant de grande taille.

Critères	ABySS	Ray	SOAPdenovo	Velvet
Nombre de scaffolds	361 840	363 401	306 246	81 825
Taille totale des scaffolds	277 273 369	310 301 799	277 593 582	224 647 142
Taille totale des scaffolds en pourcentage de la taille estimée de référence	101.6%	113.7%	101.7%	82.3%
Plus grand scaffold	264 640	139 279	394 341	457 125
Plus petit scaffold	61	100	100	121
Nombre de scaffolds > 500 nt	34 763 (9.6%)	68 007 (18.7%)	53 256 (17.4%)	29 358 (35.9%)
Moyenne des tailles des scaffolds	766	854	906	2 745
Pourcentage d'assemblage des contigs mis en scaffold	20.7%	22.9%	35.5%	64.9%
Pourcentage d'assemblage des contigs non scaffoldés	79.3%	77.1%	64.5%	35.1%
Nombre de contigs	368 305	373 477	324 937	101 322
Taille totale des contigs	276 915 658	305 546 363	276 691 538	222 641 962
Plus grand contig	263 788	131 462	279 721	112 256
Plus petit contig	61	100	2	51
Nombre de contigs > 500 nt	38 880 (10.6%)	78 083 (20.9%)	57 177 (17.6%)	44254 (43.7%)
Moyenne des tailles des contigs	752	818	852	2 197

Tableau 10. Statistiques des 4 assemblages générés selon les critères de qualité les plus importants. Les chiffres en rouge montrent le moins bon assemblage dans le critère sélectionné, inversement pour les chiffres en vert qui montrent le meilleur assemblage.

La première chose que l'on remarque dans ce tableau, c'est la grande différence qu'il existe entre les assemblages produits. Même si les assembleurs utilisent la même base, i.e. le graphe de de Bruijn, chaque algorithme varie et forme des enchainements différents de séquences.
C'est pour cela qu'il est conseillé de réaliser des assemblages provenant de plusieurs assembleurs pour optimiser l'assemblage final produit.
Concernant nos résultats, Ray créé le plus grand nombre de scaffolds avec une taille totale de 310Mb, ce qui dépasse la taille estimée du génome de 273Mb. Nous pouvons alors supposer qu'on aura plusieurs séquences pour une portion commune et que cela est dû au polymorphisme important.

Malgré le fait que la taille totale des scaffolds produits est la plus petite, Velvet a construit le moins de scaffolds et leur taille est largement supérieure aux autres. La taille moyenne se situe vers 2700 bases pour Velvet et entre 750 à 900 pour les trois autres. De plus, Velvet a le mieux réussi l'assemblage des contigs en scaffolds : 64,9% de l'assemblage produit sont des scaffolds.
ABySS et SOAPdenovo se situent entre les deux. Cependant ABySS a construit plus de petits contigs ou scaffolds et il a eu plus de mal à intégrer les contigs en scaffolds.

La consistance interne est aussi un critère d'évaluation très important. Elle correspond au pourcentage de reads utilisés pour l'assemblage qui sont correctement réalignés sur les scaffolds. L'alignement a été effectué par Bowtie2.

	ABySS	Ray	SOAPdenovo	Velvet
Taux d'alignement total	95,70 %	99,30 %	97,04 %	94,58 %
Reads PE correctement alignés	85,36 %	97,46 %	79,9 %	85,52 %
Reads PE non correctement alignés	1,11 %	0,45 %	1,42 %	1,06 %
Reads alignés seul	9,23 %	1,39 %	15,72 %	8 %

Tableau 11. Réalignement des reads qui ont servi à l'assemblage sur les scaffolds issus de l'assemblage. Les chiffres en vert signalent les meilleures statistiques de chaque critère.

C'est Ray qui montre la meilleure consistance interne puisqu'il a le meilleur taux de reads appariés correctement réalignés soit 97,46% (Tableau 11). Suivent Velvet, ABySS et SOAPdenovo en terme de reads PE correctement réalignés. Il a aussi le moins de reads PE réalignés mais qui ne suivent pas les règles du PE et le moins de reads réalignés seuls, i.e. que l'autre read appairé est soit sur un autre chromosome soit ne s'est pas aligné du tout. Cela montre un très bon assemblage à partir des données brutes.

Comme nous avions un génome de référence disponible, nous avons aligné les scaffolds de chaque assembleur sur le Golden Path d'*A. gambiae* avec MUMmer (Tableau 12) pour estimer le taux de contigs correctement assemblés. Il faut cependant noter que le génome actuel n'étant pas une très bonne référence, ce critère n'est pas le plus important.

	ABySS	Ray	SOAPdenovo	Velvet
% bases couvertes (# bases du GP : 273 093 681)	73,7 %	60,2 %	62 %	74,5 %
# contigs mal assemblés	14 787	20 403	8 983	9 991
# contigs redondants	44 755	112 391	82 834	18 583
# contigs non alignés (dont # contigs alignés partiellement)	224 133 (6 608)	188 833 (10 007)	127 922 (12 396)	15 580 (6 217)

Tableau 12. Statistiques des alignements des scaffolds sur le Golden Path d'*A. gambiae* AgamP3.

Ray qui a produit le plus de scaffolds surpassant la taille du génome n'a réussi qu'à recouvrir 60,2% des bases. En fait, il a construit énormément de contigs redondants. Il a aussi beaucoup de contigs non alignés mais cela ne vaut pas forcément dire que ce sont des chimères vu que AgamP3 contient beaucoup de zones N et que les haplotypes non pas été comptabilisés. En tout cas, Velvet qui avait construit le moins de contigs mais de plus grandes tailles a recouvert le plus de bases du GP. Il a aussi le moins de contigs redondants (18 583) et le moins de contigs non alignés (15 580). C'est ABySS qui a le plus de contigs mal assemblés et non alignés. Le fait qu'il ait utilisé le moins de reads PE correctement réalignés explique ces mauvais assemblages.

Pour conclure sur ces quatre premiers assemblages réalisés, nous avons vu qu'il est très difficile de reconstruire notre génome même avec une couverture de 102X et une lignée relativement peu polymorphique par rapport aux autres. Le fait qu'*Anopheles gambiae* présente beaucoup de polymorphismes complique fortement la tâche des assembleurs et amène à la formation de plusieurs contigs pour une même région (voir l'assemblage par Ray). De plus, le plus grand scaffold construit est de 450 000 bases ce qui, à l'échelle du génome, est faible (0,17% du génome). L'incapacité à produire de plus grands scaffolds provient de l'utilisation d'une seule librairie de reads avec des petits inserts. L'association de cette librairie à une librairie avec des grands inserts ou à un séquençage de 3ème génération produisant de longues séquences devrait considérablement améliorer l'assemblage final.

3. Détermination des polymorphismes

Pour identifier les variants intra- et inter-lignées et désigner des marqueurs génétiques, on a recours à un alignement des reads provenant d'une lignée sur un séquence de référence choisie puis à l'utilisation de programmes détecteurs de variants sur les fichiers d'alignements.

Nous avons effectué l'alignement par Bowtie2 des reads de notre lignée R1iso2 sur le génome de référence d'*Anopheles gambiae* AgamP3 et ses haplotypes pour identifier les polymorphismes inter-lignées. Nous avons ensuite utilisé des scripts de préparation des données proposés par BCFtools et GATK qui enlève les reads dupliqués, puis détecte et réaligne les petites zones contenant des indels pour distinguer les vrais variants. Puis nous avons utilisé deux scripts, un de GATK et un de SAMtools pour lister les variants.

	GATK	SAMtools
# total de variants identifiés	2 635 772	2 785 325
# SNPs	2 161 488	2 274 122
# locus multi-SNPs	3 005	0
Ratio Ti/Tv	1,22	1,23
# insertions	224 368	90 884
# délétions	240 299	49 170
# variants complexes	6 896	371 149
# variants mixtes	2 721	0
# variants structuraux	724	0

Tableau 13. Listes et nombres des types de variants identifiés par GATK et SAMtools. Ti : transition, Tv : tranversion. Les variants sont considérés comme mixtes si le même site contient à la fois des SNPs et des indels et comme structuraux quand la taille est égale ou supérieure à 50bp.

Entre les deux outils, SAMtools a tendance a être un peu plus laxiste dans son identification des SNPs. Il liste des variants dans des régions d'alignements médiocres ou suspicieuses par rapport à de régions mieux couvertes (Fig. 27).

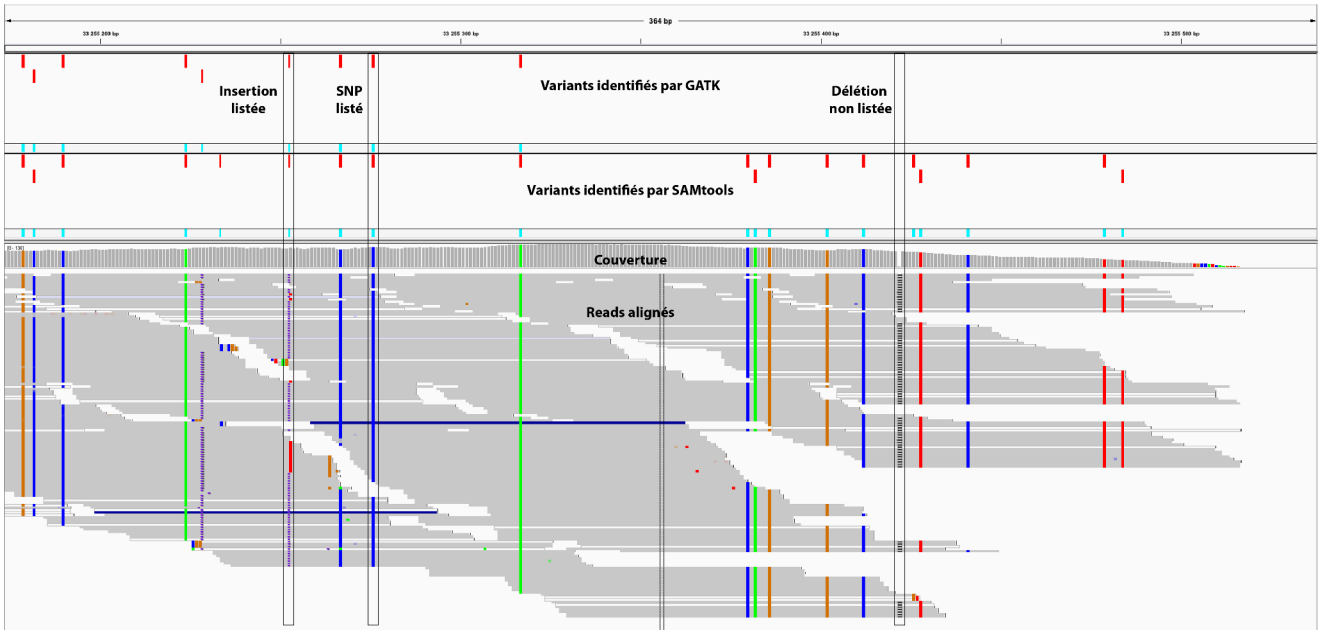


Figure 27. Visualisation par le logiciel IGV permettant de voir les variants listés par GATK en premier, par SAMtools en deuxième et les reads alignés en troisième. Les différentes couleurs représente les substitutions dans les reads : vert pour un A, bleu pour un C, rouge pour un T et marron pour un G. A gauche, il y a deux exemples d'une insertion et d'un SNP simultanément listés. A droite, les substitutions résultant de l'alignement n'ont pas été listées par GATK à cause du faible nombre de reads superposés. Par contre, la délétion n'a été listée par aucun des deux.

GATK et SAMtools ont détecté 2 027 798 sites contenant des variants communs. 304 031 et 304 431 sites sont uniques à GATK et SAMtools respectivement, et 319 096 sites se chevauchent mais ils n'ont pas été classés dans la même catégorie (Fig. 28). Le cas encadré en rouge au centre a été identifié comme SNP par GATK (Fig. 28, encadré à droite) et comme

INDEL par SAMtools (Fig. 28, encadré à gauche). En fait, sur les 99 reads, il y a 19 « bonnes » bases, dont 10 présentent une insertion ensuite, contre 80 substitutions de T > C.

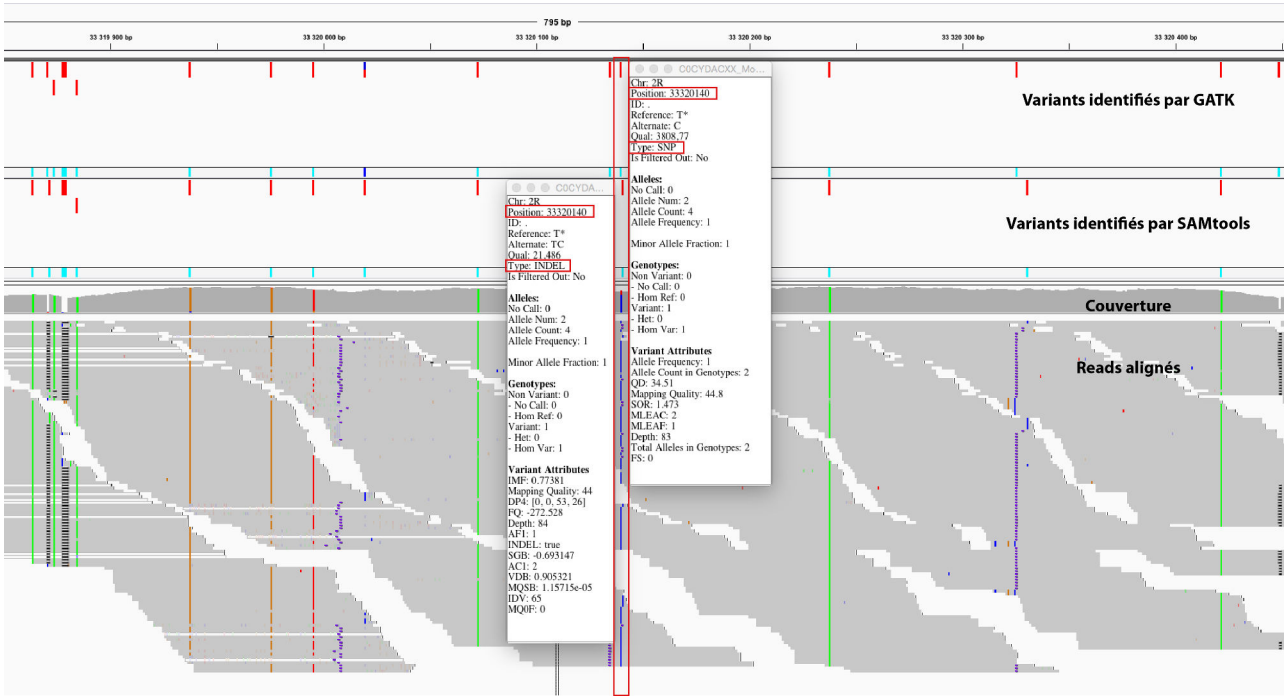


Figure 28. Visualisation par le logiciel IGV permettant de voir les variants listés par GATK en premier, par SAMtools en deuxième, la couverture en troisième et les reads alignés en quatrième. Les différentes couleurs représente les substitutions dans les reads : vert pour un A, bleu pour un C, rouge pour un T et marron pour un G.

GATK semble meilleur pour lister des variants qui ont le plus de chances d'en être de réels ; les variants identifiés le sont dans des régions de meilleure qualité et l'appellation du variant est plus logique. Toutefois, pour obtenir une liste de variants sûrs pour le génotypage et l'évaluation du polymorphisme, le mieux est de garder les variants communs aux deux outils. Les variants identifiés ici sont ceux de notre lignée par rapport au génome de référence. Les variants intralignée seront aussi à déterminer et ce sera plus informatif sur le polymorphisme de notre lignée R1iso2.

IV. Conclusions et perspectives

Nous avons vu, en alignant nos reads sur le génome et les haplotypes d'*Anopheles gambiae* version P3, que (1) ce génome contenait encore beaucoup de portions inconnues et identifiées par des N, (2) les haplotypes sont le résultat d'un fort taux de polymorphisme et (3) il y a beaucoup de problèmes d'assemblage, duplications, inversions, insertions, délétions.
Tout cela nous montre les limites actuelles du génome d'*Anopheles gambiae* disponible sur VectorBase. En fait, seulement 61% du génome est utilisable pour réaliser les alignements des reads. Les relatives mauvaise qualité et représentativité de nos lignées nous ont donc démontré la nécessité de séquencer et d'assembler nos propres lignées pour mener à bien les génotypages des moustiques sensibles et résistants et nos recherches de régions liées à la résistance au parasite du paludisme.

Cependant, nous avons obtenu des assemblages bien différents tant au niveau de la qualité que de la quantité. Et il très difficile de déterminer quel est le meilleur. Sur tous les critères évalués, ABySS et SOAPdenovo ressortent moins bons que Ray et Velvet. Entre les deux, Velvet produit les plus grands scaffolds mais en petite quantité alors que Ray produit un grand nombre de scaffolds permettant de surpasser la taille du génome de référence. Ce qui donne des régions redondantes. C'est Ray qui présente la meilleure consistance interne et Velvet qui a réussi à recouvrir le plus de bases du GP. Concernant les temps d'exécution, Ray est clairement le plus long mais il ne prend que très peu de mémoire par rapport à ABySS, SOAPdenovo et Velvet. Les problèmes rencontrés lors de l'assemblage sont dus en majorité au polymorphisme de la lignée séquencée et à l'utilisation d'une seule librairie de petits inserts.

Ces premières esquisses d'assemblage *de novo* de notre génome nous ont permis d'en tirer les leçons suivantes :

(1) Séquencer des individus seuls pour limiter le polymorphisme
Nous avions utilisé 6 moustiques pour générer ce set de données représentatif du polymorphisme de la lignée R1iso2 que nous considérions faible. Le mieux aurait été de faire plusieurs librairies d'individus isolés pour limiter au maximum le polymorphisme de chaque librairie.

(2) Estimer une couverture suffisante
(Desai et al., 2013) ont montré avec des génomes de bactéries et de levure qu'une couverture de 30-50X produisait des assemblages corrects. Illumina a aussi montré avec le génome de E. coli qu'à partir de 30X une augmentation de la couverture n'améliorait pas l'assemblage produit. Par contre, une augmentation de la couverture veut dire plus de reads donc cela entraine une augmentation de la consommation de mémoire lors de l'assemblage. Ce qui peut être un facteur limitant lorsque l'on a que des machines avec peu de mémoire. Dans le cas de plus gros génomes, celui du panda géant a été assemblé avec une couverture de 56X (Li et al., 2010) et, plus récemment, ceux du projet des 16 génomes de moustiques (Neafsey et al., 2013) avec une couverture de 100X. La couverture doit être choisie en fonction du motif du séquençage, du type de séquençage et de la complexité du génome.

(3) Vérifier la qualité des reads
La qualité des reads est un critère très important. Il est obligatoire de juger de la qualité de ses reads via un logiciel ou tout autre script avant de commencer tout assemblage. Pour certaines données, il est possible et même conseillé d'éliminer certains reads jugés anormaux ou de mauvaise qualité, de rogner le bout 3' de reads ayant un très faible score de qualité (ce qui arrive chez Illumina). Des logiciels sont disponibles tels que FASTX-Toolkit (Hannon Lab, hannonlab.cshl.edu), Trimmomatic (Bolger et al., 2014), etc.
Quelques logiciels sont aussi capables de détecter et de corriger les erreurs de séquençage dans les reads dont Quake, (Kelley et al., 2010).

(4) Sélectionner un ou plusieurs k-mer
Le choix du paramètre k lors d'un assemblage joue énormément sur la qualité du résultat. Nous savons qu'il faut choisir un k à chaque nouveau set de reads. Comme il est difficile d'en choisir un et qu'il n'est pas conseillé de le choisir arbitrairement, nous recommandons d'utiliser un logiciel permettant une estimation du meilleur k ou d'en sélectionner plusieurs lors de l'assemblage.

(5) Combiner les technologies de séquençage
Il est aussi possible de combiner plusieurs technologies de séquençage afin de cumuler les avantages et de peut-être de minimiser les inconvénients. Cela s'appelle un assemblage hybride. Le but ici n'est pas d'assembler chaque set de reads séparément pour les réunir ensuite, c'est plutôt de les assembler simultanément.
Les génomes des bactéries *Acinetobacter baylyi et Pseudomonas syringae assemblés avec des reads Roche/454 et Illumina/Solexa sont meilleurs que ceux issus d'une seule technologie (Aury et al., 2008; Reinhardt et al., 2009).*
Mais cela demande alors un assembleur capable de gérer un assemblage hybride, c'est à dire de prendre en compte des reads issus de deux technologies différentes.
Ray est capable de mixer 454 et Illumina et il a permis de diminuer les erreurs dans les contigs produits pour trois génomes de bactéries (Boisvert et al., 2010). ALLPATHS-LG peut associer des reads Illumina avec des reads PacBio et a produit les meilleures statistiques d'assemblage pour quatre génomes de bactéries (Utturkar et al., 2014). Le but ici est de surtout associer de petits reads PE avec de très longs reads issus de la 3ème génération de séquençage afin d'aider lors de la création des scaffolds.

(6) Combiner plusieurs assembleurs
Il est possible de réaliser différents assemblages avec des assembleurs différents et on refait un assemblage à partir des assemblages précédents. Bambus (Pop et al., 2004) est un outil qui accepte les outputs de plusieurs assembleurs et qui propose de choisir les options de scaffolding.

(7) Séquencer et assembler des données transcriptomiques
Pour améliorer un assemblage de séquences d'ADN génomique, il est conseillé de séquencer et d'assembler des séquences d'ARN. Plusieurs assembleurs de novo de données RNA-Seq sont disponibles dont Trinity (Grabherr et al., 2011) et Bridger (Chang et al., 2015).

(8) Prévoir des ressources informatiques adaptées
Dans notre cas, nous avions à reconstituer un génome de 273Mb en ayant une couverture de 100X. Nous avions 280 millions de reads répartis en deux fichiers formant un total de 70Gb. Suivant l'assembleur utilisé, il nous a fallu avoir une mémoire disponible supérieure à 200Gb.

L'assemblage *de novo* d'un génome aussi grand est une tâche considérable en terme de ressources humaines et informatiques. Le temps nécessaire est aussi à prendre en compte. Ce qui n'est pas forcément adapté pour notre projet.

Nous avons donc redéfini notre stratégie d'étude pour la suite.
Notre but est de déterminer les différences génétiques entre moustiques résistants et moustiques sensibles. Pour ce faire, nous avons défini la méthode suivante :

1- nous réaliserons le séquençage Illumina d'ADN et d'ARN issus de moustiques individuels pour obtenir de petits reads.
2- nous alignerons ces reads sur le génome de référence qui sera sélectionné en fonction du type de polymorphisme voulu : intra- ou inter-lignées.
3- nous listerons tous les variants à partir des alignements précédents, éventuellement nous identifierons les variations du nombre de copies.
4- les polymorphismes inter-lignées identifiés seront ensuite utilisés comme marqueurs pour cartographier les QTLs et les polymorphismes intra-lignées pour l'étude d'association sur génome entier (GWAS).
5- les régions d'intérêts montrant un lien fort avec le degré de résistance et dont la séquence du GP est de mauvaise qualité ou ne correspond pas aux séquences de nos lignées seront réassemblées à partir des petits reads par une combinaison d'assemblage *de novo* par Velvet et Ray et de réalignements automatique et manuel. La consistance interne permettra de vérifier l'exactitude des régions assemblées. La connaissance précise de la séquence de ces régions nous apportera une liste exacte des polymorphismes présents et donc des gènes polymorphiques.
6- l'étude d'expression des gènes grâce aux données RNA-Seq nous dressera une liste des gènes différentiellement exprimés entre moustiques résistants et sensibles.
7- les étapes 5 et 6 nous fournissent, à travers les listes de gènes polymorphiques et de gènes différentiellement exprimés, une liste de gènes candidats à tester.

Cette stratégie sera réalisée pour les prochains séquençages des génomes et des transcriptomes de 6 lignées sélectionnées de moustiques résistants et sensibles.

Chapitre II

Optimisation de la méthode du « reciprocal allele-specific RNA interference » (rasRNAi) afin d'identifier les gènes responsables de la résistance du moustique *Anopheles gambiae* aux parasites du paludisme

Sommaire

I. Contexte

Les moustiques de l'espèce *Anopheles gambiae* ne sont pas tous sensibles à une infection aux parasites du paludisme. Dans certaines lignées de moustiques, la réponse antiparasitaire est particulièrement efficace. Elle bloque complètement le développement du parasite à un stade précoce de l'infection, ie dans l'intestin juste après l'ingestion des parasites. Un des acteurs majeurs de cette réponse est un gène codant pour une protéine de type complément : le gène TEP1 (ThioEster-containing Protein 1) (Blandin et al., 2004). C'est une protéine homologue au facteur du complément C3 (protéine faisant partie du système immunitaire innée chez l'homme). Cette molécule sécrétée dans l'hémolymphe a une activité antiparasitaire clé : elle se lie aux parasites et promeut leur destruction. De plus, ce gène possède plusieurs variants alléliques. Grâce à la technique du « reciprocal allele-specific RNA interference » (rasRNAi), les allèles de TEP1 ont été reliés aux divers degrés de résistance observés. Cependant, à lui seul TEP1 n'est pas responsable du phénomène de résistance du moustique. Notre objectif est d'identifier d'autres facteurs génétiques entrant en compte dans l'extinction des parasites *P. berghei* chez nos lignées de moustiques.

TEP1 fait partie d'une famille multigénique dont certains ont un rôle dans la réponse antiparasitaire. De plus, 9 d'entre eux sont localisés parmi les 975 gènes de la région *Pbres1* du chromosome 3L (Fig. 1).

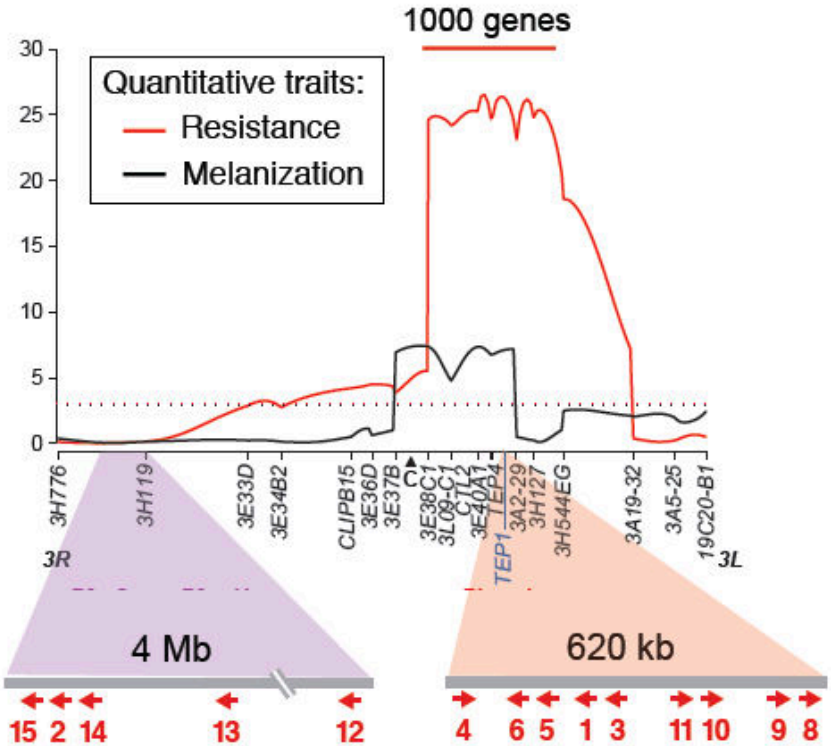


Figure 1. Cartographie génétique du chromosome 3 liant génotype et phénotype. Le tracé rouge représente le phénotype de résistance, le tracé noir correspond à la mélanisation des parasites. Les marqueurs génétiques sont spécifiés sous l'axe des X avec l'inscription de quelques gènes impliqués dans la résistance (CTL2, TEP4 et TEP1). La zone des 1000 gènes correspond au locus Pbres1. Les zones colorées permettent de situer et d'orienter les différents TEPs signalés en rouge. Source : S. Blandin

7 gènes de cette famille multigénique ont été testés par l'injection d'un dsRNA spécifique dans des moustiques G3, ie sensibles sur leur rôle dans la résistance du moustique aux parasites du paludisme murin. Pour chaque gène TEP ciblé, le nombre de parasites vivants dans l'intestin des moustiques a été comptabilisé et comparé au nombre de parasites après l'injection d'un dsRNA ciblant LacZ (gène exogène au génome du moustique) (Fig. 2).

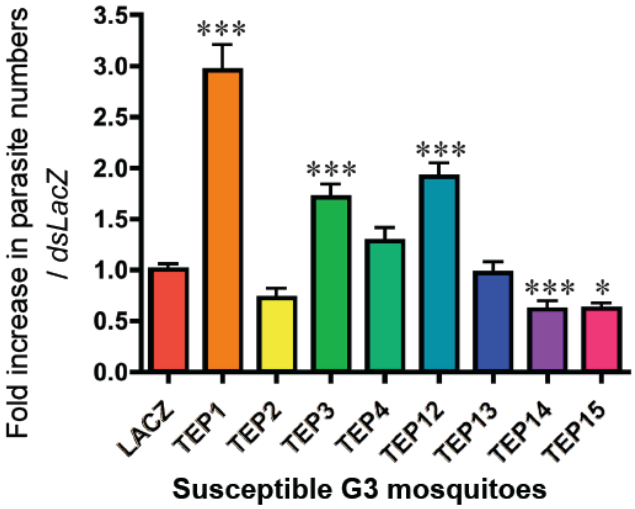


Figure 2. Impact de l'inhibition des 7 TEPs sélectionnés sur le nombre de parasites vivants chez des moustiques sensibles. LacZ représente le contrôle négatif et TEP1 le contrôle positif. Source : S. Blandin

L'injection du dsRNA ciblant TEP3 et TEP12 provoque une augmentation de 1,5 à 2 fois le nombre de parasites chez les moustiques G3. Ils ont donc un rôle dans la destruction des parasites. TEP2, TEP4 et TEP13 ne montrent pas de changement significatif. Quant à TEP14 et TEP15, leur inhibition a eu une répercussion négative sur les parasites. Cela pourrait signifier qu'ils ont soit une action positive sur le développement des parasites, soit une action négative sur la réponse immunitaire mise en place contre les parasites.
TEP3, TEP12, TEP14 et TEP15 sont donc des gènes anti/pro-parasitaires.

Pour être un candidat idéal à l'explication de la résistance du moustique aux parasites *P. berghei*, les gènes doivent présenter des polymorphismes en plus du fait d'avoir une action antiparasitaire.
TEP3 est un gène situé sur le chromosome 3L à côté du gène TEP1. Il compte 12 exons codants qui totalisent 4 074 bases. TEP12 est situé sur le bras R du chromosome 3. Il compte 4 exons codants pour 2 550 bases.
Après séquençage de ces deux gènes dans nos propres lignées de moustiques résistants et sensibles, il est apparu plusieurs 3 allèles différents pour TEP3 (Fig. 3) et 4 pour TEP12 (Fig. 4). Les lignées L3-5 et G3M possédant un seul allèle, ce sont ces derniers qui ont été clonés et utilisés pour la séquence des dsRNAs.

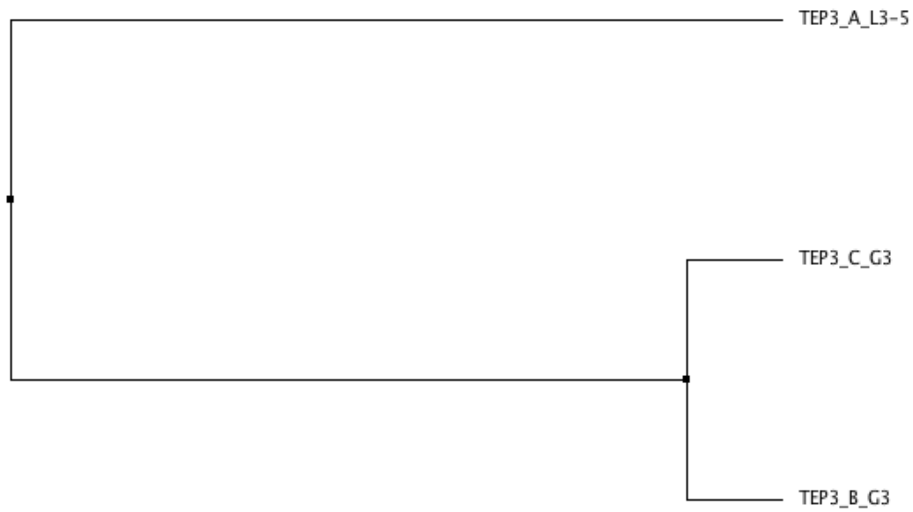


Figure 3. Arbre phylogénétique des allèles du gène TEP3 identifiés chez la lignée résistante L3-5 (TEP3_A) et la lignée sensible G3 (TEP3_B et _C).

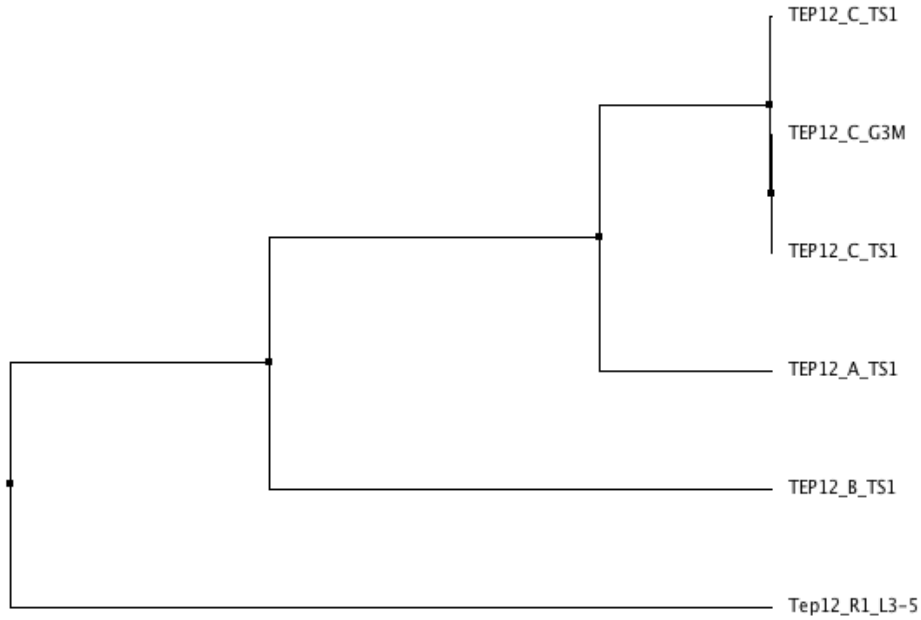


Figure 4. Arbre phylogénétique des allèles du gène TEP12 identifiés chez la lignée résistante L3-5 (TEP12_R1) et les lignées sensibles TS1 (TEP12_A, _B et _C) et G3M (TEP12_C).

Dans les cas de TEP3 et TEP12, en raison de leur faible polymorphisme, il est impossible de choisir une séquence aussi grande et qui inhibera seulement l'allèle choisi (Fig. 5 et 6). TEP1 étant très polymorphique (Fig. 7), la séquence du dsRNA injecté de 75 paires de bases (pb) était bien spécifique de l'allèle ciblé puisqu'elle ne contenait pas de fragments >= 19 paires de bases identiques.

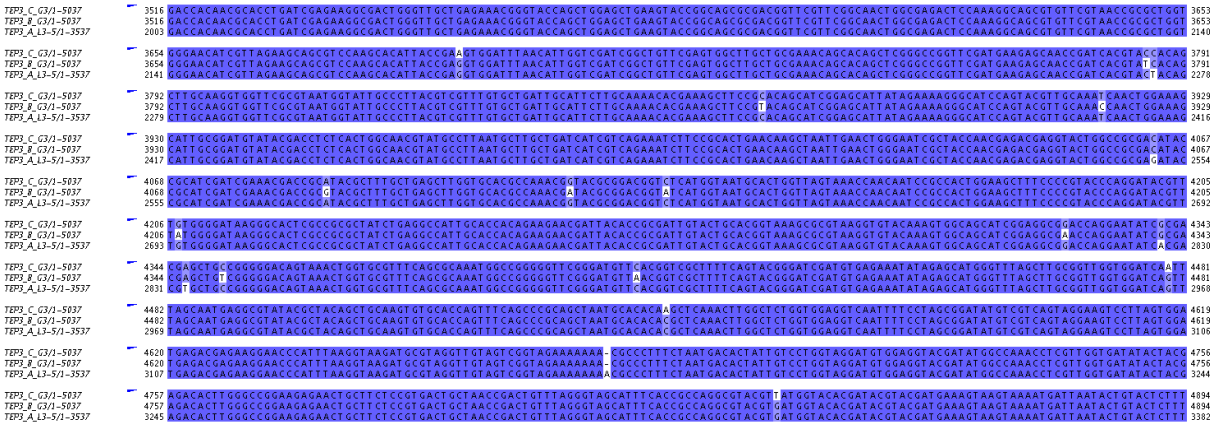


Figure 5. Alignement des 3 allèles de TEP3 : TEP3_A_L3-5, TEP3_B_G3 et TEP3_C_G3. Plus la conservation entre les allèles est forte, plus la coloration bleue s'intensifie.

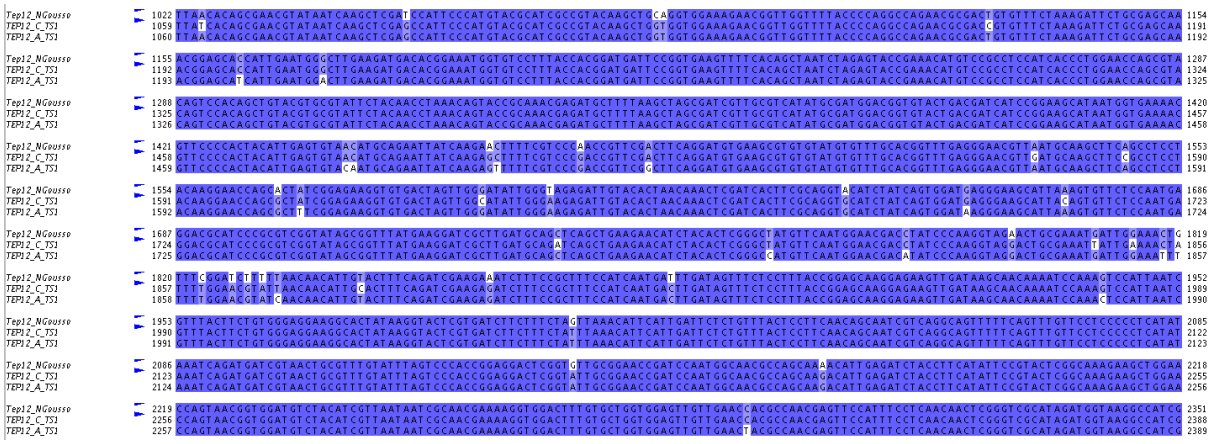


Figure 6. Alignement de 3 allèles de TEP12 : TEP12_NGousso, TEP12_A_TS1 et TEP12_C_TS1. Plus la conservation entre les allèles est forte, plus la coloration bleue s'intensifie.

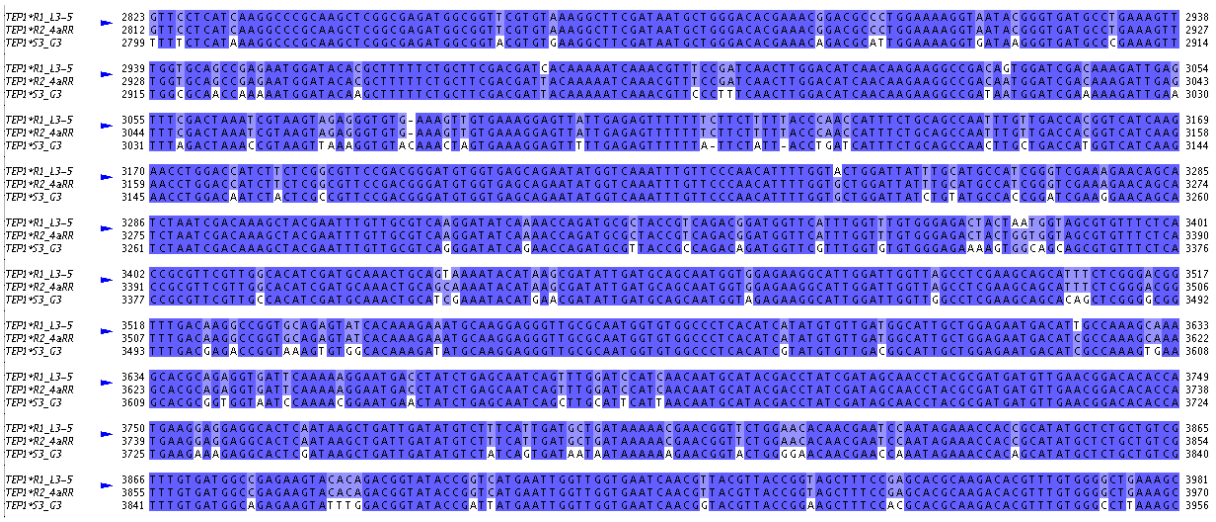


Figure 7. Alignement de 3 allèles de TEP1 : TEP1*R1, *R2 et *S3. Plus la conservation entre les allèles est forte, plus la coloration bleue s'intensifie.

Afin d'inhiber spécifiquement chaque allèle de ces deux gènes, TEP3 et TEP12, le laboratoire a testé l'injection directe de siRNAs dans des moustiques adultes. L'expérience a été réalisée sur le gène TEP1 afin de comparer l'efficacité d'inhibition d'un siRNA par rapport à celle d'un dsRNA.
Ils ont injecté 2 dsRNAs ciblant le gène exogène LacZ et le gène TEP1 et 3 siRNAs : un siScrambled qui est un enchainement de 21 bases qui n'existe pas dans *A. gambiae* et deux siRNAs ciblant TEP1 : siTEP1 et siTEP1m qui a la même séquence que siTEP1 mais avec des bases LNA® (Locked Nucleic Acid) (Fig. 8).

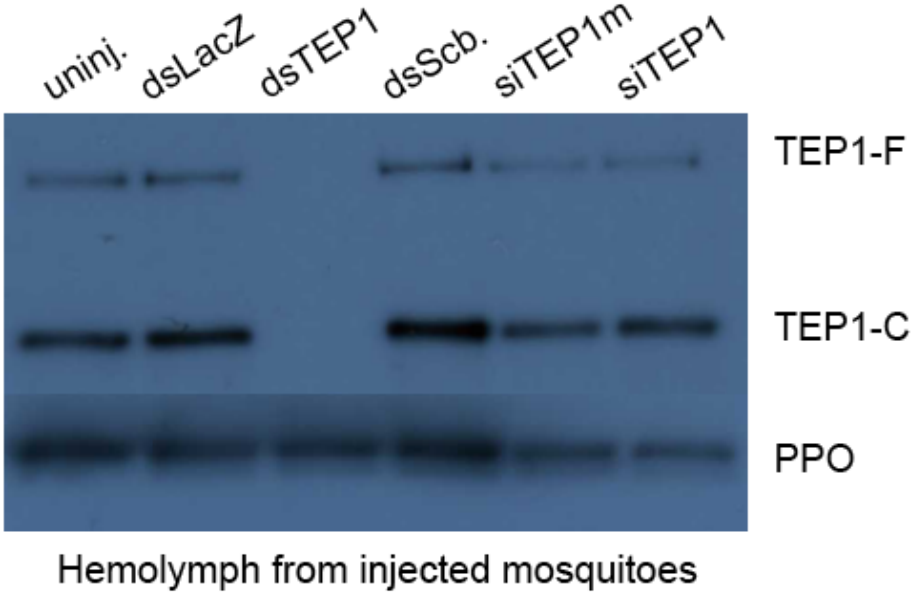


Figure 8. Western blot montrant les résultats des injections de différents dsRNAs (dsLacZ, dsTep1) et de différents siRNAs (siScb., siTep1m et siTep1). Les quatre premiers, la non-injection, l'injection de dsLacZ, de dsTep1et du siRNA siScb servent de contrôle. siTep1m et siTep1 ont la même séquence de 21 bases mais le premier est un siRNA modifié. Sa séquence est constitué de bases LNA® (Locked Nucleic Acid). Les deux bandes représentent les deux formes de Tep1 : Tep1-F, la forme protéique complète, et Tep1-C, la forme clivée. Les bandes de charge du bas montre l'expression d'un gène domestique, PPO, et servent de contrôle. Source : S. Blandin

Les injections des siRNAs siTep1 et siTep1m n'ont pas généré d'inhibition de la production des formes protéiques F et C du gène TEP1 par rapport à l'injection du dsRNA dsTep1 qui est efficace (Fig. 8). La taille de la séquence injectée est donc un critère essentiel dans la réussite du processus d'inhibition d'un gène.
Chez la drosophile, le « soaking » de siRNAs n'est pas efficace pour inhiber l'expression d'un gène du fait que les cellules n'internalisent pas les siRNAs (Saleh et al., 2006). La sélection se fait au niveau de récepteurs externes des cellules et les siRNAs n'ont pas les critères de taille requis pour entrer. Par contre, la transfection des siRNAs fonctionne parfaitement.

L'injection directe de siRNAs ne fonctionnant pas, il a fallu trouver un moyen de les transporter dans les cellules puisqu'il est essentiel d'avoir de petites séquences. Dans les cas de faible polymorphisme, les petites séquences sont les seules qui peuvent être allèle-spécifiques. Il faut toutefois prendre en compte que ce transporteur devra avoir une taille suffisante (>80pb) pour être efficace.
Dans ce chapitre, notre objectif est de créer une sonde dsRNA allèle-spécifique pour inhiber efficacement les gènes faiblement polymorphiques. Pour savoir comment construire cette sonde, il faut savoir comment les cellules traitent les dsRNAs injectés. Nous avons donc séquencé et étudié les petits ARNs issu de dsRNAs injectés dans des moustiques adultes.

II. Matériels et méthodes

Note : toutes les techniques de biologie utilisées dans ce chapitre (PCR quantitative, western blot, synthèse et injection des dsRNAs, préparation des libraires pour le séquençage, etc.) ont été réalisées par mes collègues.

Plusieurs lignées de moustiques présentes au laboratoire ont été utilisées dans ce projet : une lignée résistante : L3-5 et trois lignées sensibles : G3, G3M et TS1. G3M est une lignée isogénique descendant de la lignée G3. TS1 a été sélectionnée à partir de la lignée T4 (Bernardini et al., 2014) qui présente trois allèles de TEP1 : *R2, *S1 et *S2. La lignée TS1 possède seulement l'allèle *S1.

Plusieurs dsRNAs ont été utilisés tout au long du projet :

(1) dsLacZ qui provient d'un gène exogène à *A. gambiae*. C'est notre contrôle négatif pour toutes les expériences.
(2) dsTep1 qui provient du gène TEP1. dsTep1*R et dsTep1*S proviennent des allèles des lignées L3-5 et G3, respectivement.
(3) dsTep3 regroupe 2 dsRNAs injectés ensemble et provenant de deux régions différentes du gène TEP3. dsTep3-R1 et dsTep3-G3 proviennent des allèles des lignées L3-5 et G3M.
(4) dsTep12 regroupe 3 dsRNAs injectés ensemble et se suivant dans le gène TEP12. dsTep12-R1 et dsTep12-G3 proviennent des allèles des lignées L3-5 et G3M.
(5) ds151-329 et ds329-151 sont les nouvelles sondes que nous avons mises au point.

Tous les séquençages réalisés sont des smallRNA-Seq et ont été effectués sur la plateforme HiSeq2000 d'Illumina. Les librairies ont été préparées selon le protocole NEBNext® Multiplex Small RNA Library Prep Set for Illumina®. Nous avons réalisé 5 séquençages comprenant de 2 à 9 échantillons identifiés par un index unique (Tableau 1). L'ensemble des séquençages a produit de 4 à 60 millions de reads de 3 à 52 bases.

Identifiant du séquençage	Nom du dsRNA injecté	Lignée de moustiques	Identifiant	Informations supp.
D0YEFACXX-JRC7	dsLacZ	TS1	dsLacZ_1	Contrôle - Réplicat 1
	dsTep1	TS1	dsTep1_1	Réplicat 1
D0YEFACXX-JRC8	dsLacZ	TS1	dsLacZ_2	Contrôle - Réplicat 2
	dsTep1	TS1	dsTep1_2	Réplicat 2
C1HW5ACXX-JRC9	dsTep1	TS1	dsTep1-1.6h	6h - Réplicat 1
	dsTep1	TS1	dsTep1-1.24h	24h - Réplicat 1
	dsTep1	TS1	dsTep1-1.48h	48h - Réplicat 1
	dsTep1	TS1	dsTep1-1.72h	72h - Réplicat 1
	dsTep1	TS1	dsTep1-2.6h	6h - Réplicat 2
	dsTep1	TS1	dsTep1-2.24h	24h - Réplicat 2
	dsTep1	TS1	dsTep1-2.48h	48h - Réplicat 2
	dsTep1	TS1	dsTep1-2.72h	72h - Réplicat 2
	dsTep1	TS1	dsTep1-2.6d	6 jours
D1UJHACXX-JRC11	dsTep3	TS1	dsTep3mix	
	dsTep12	TS1	dsTep12mix	
	ds151-329	TS1	ds151-329	
	ds329-151	TS1	ds329-151	
C2LVTACXX-JRC12	dsTep3-G3	TS1	dsTep3-G3.TS1	
	dsTep12-G3	TS1	dsTep12-G3.TS1	
	dsTep3-R1	TS1	dsTep3-R1.TS1	
	dsTep12-R1	TS1	dsTep12-R1.TS1	
C3PEGACXX-JRC13	dsTep1*R	L3-5	dsTep1*R-L35.1	Réplicat 1
	dsTep1*S	L3-5	dsTep1*S-L35.1	Réplicat 1
	dsTep1*R	L3-5	dsTep1*R-L35.2	Réplicat 2
	dsTep1*S	L3-5	dsTep1*S-L35.2	Réplicat 2
	dsTep1*R	G3	dsTep1*R-G3.1	Réplicat 1
	dsTep1*S	G3	dsTep1*S-G3.1	Réplicat 1
	dsTep1*R	G3	dsTep1*R-G3.2	Réplicat 2
	dsTep1*S	G3	dsTep1*S-G3.2	Réplicat 2

Tableau 1. Détails des séquençages réalisés. (a) Identifiants des 5 séquençages et identifiants (index) de chaque échantillon pour un séquençage. (b) Les données regroupent pour chaque échantillon séquencé son identifiant de séquençage, le nom du gène ou de l'allèle ciblé, les lignées dans lesquelles ont été injecté les dsRNAs, l'identifiant de l'échantillon et des informations supplémentaires.

Analyses et outils bioinformatiques

Tout d'abord, il a fallu retirer les séquences des adaptateurs de chaque read. L'opération a été achevée par Cutadapt (version 1.0 à 1.3, (Martin, 2011)).
Ensuite, j'ai vérifié la qualité des séquençages produits à l'aide de FastQC.

Pour les analyses, j'ai utilisé plusieurs outils disponibles (Tableaux 2 et 3) et j'ai écrits mes propres scripts en langage Perl.

Bowtie a servi à l'alignement de chaque pool de reads séquencés sur le dsRNA injecté correspondant. Excel a été utilisé pour manipuler les données analysées et créer des graphes pour la visualisation. WebLogo est un logiciel de création de logos issus de l'alignement de séquences multiples d'acides aminés ou d'acides nucléiques. Cela permet de visualiser si il existe une séquence consensus aux multiples séquences.

Outil	Version	Publication
Bowtie	1.0.0	(Langmead et al., 2009)
Excel		
WebLogo	3.3	(Crooks et al., 2004; Schneider and Stephens, 1990)

Tableau 2. Détails des outils utilisés pour analyser les données de ce chapitre

siRNA design tools	Criteria	Designer lab team or company
Block-iT™ RNAi Designer	ESM = Default / Tuschl, %GC	Life Technologies
DSIR	ST	(Vert et al., 2006)
E-RNAi	ST, ESM = Weighted / Rational	(Horn and Boutros, 2010)
OligoWalk	/	(Lu and Mathews, 2008)
OptiRNA	/	(Ladunga, 2007)
siDesignCenter	%GC	Dharmacon
siDirect	ESM = Ui-Tei / Reynolds / Amarzguioui, %GC	(Naito et al., 2004, 2009)
siRNA Design Service	GC	eurofins Genomics
siWizard	GC	InvivoGen

Tableau 3. Détails des outils de prédiction des siRNAs efficaces à partir d'une séquence à inhiber. Au niveau des règles qui définissent l'efficacité d'un siRNA, il est parfois possible de sélectionner la méthode de calcul de l'efficacité (ESM = efficiency scoring method) ou alors de déterminer quelques critères tels que le pourcentage en GC autorisé, le seuil minimum d'efficacité à afficher (ST) ou alors certains outils ne laisse pas du tout le choix comme OligoWalk.

III. Résultats

1. Profils des données de séquençage

Rapport de qualité
Les statistiques calculées ont révélé une très bonne qualité de l'ensemble des séquençages réalisés (Fig. 10). Tous les séquençages ont produit un profil quasiment similaire.

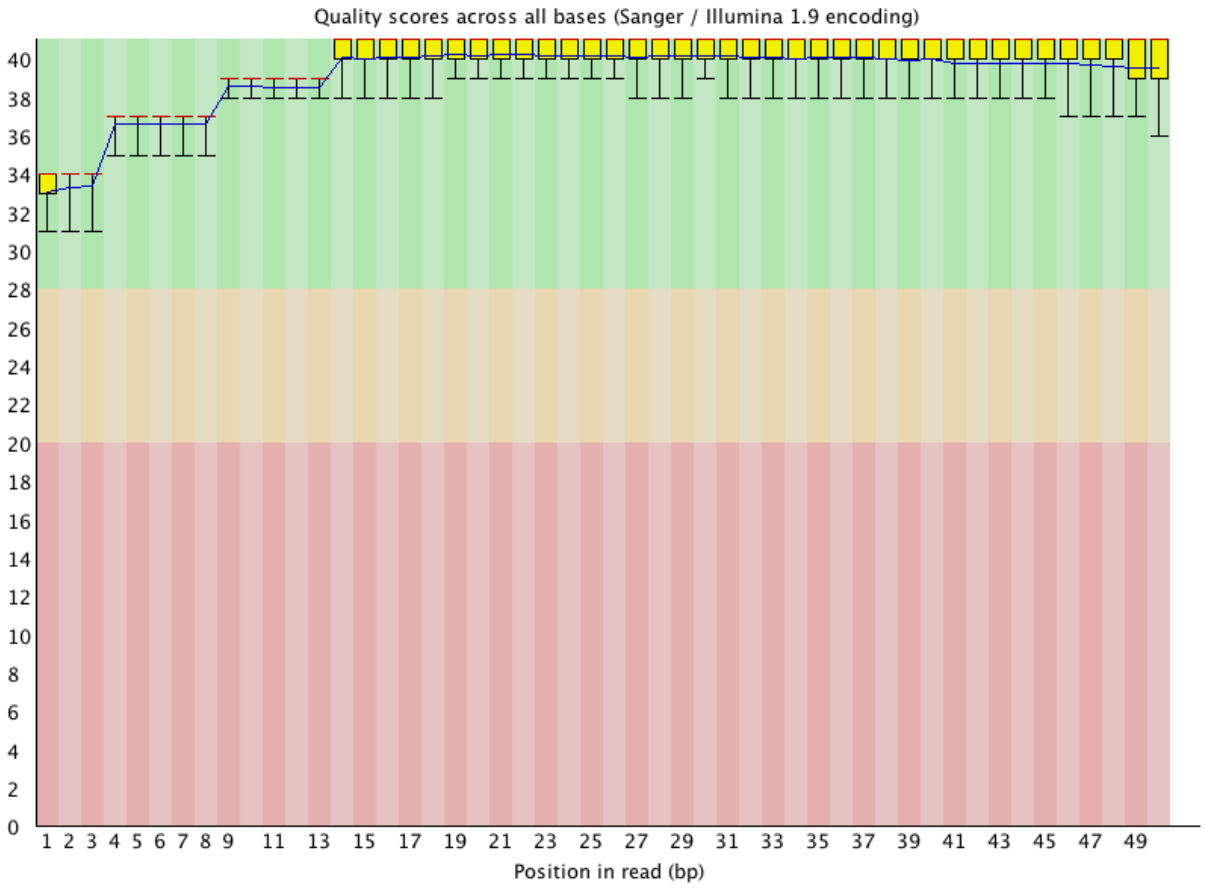


Figure 9. Graphe de qualité par base de toutes les séquences d'un échantillon (dsTep1).

Sélection des tailles à garder
Les tailles des reads de nos librairies sont toutes comprises entre 3 et 52 bases. Les séquences qui nous intéressent, les siRNAs, mesurent 21pb. Pour déterminer la taille minimum à garder, nous avons aligné les reads de 4 librairies en autorisant seulement les alignements parfaits sur une séquence de 720 bases d'un gène non présent dans l'organisme du moustique, le gène LacZ. Il est donc impossible que nous retrouvions des reads appartenant à LacZ dans nos échantillons. C'est à partir de 15 bases que les reads ne s'alignent plus sur la séquence de LacZ. Nous avons donc choisi le minimum à 15b.
Pour toutes les analyses qui ont été faites, nous n'avons gardé que les reads ayant une taille comprise entre 15 et 40b inclus.

Distribution du nombre total de reads séquencés par échantillon

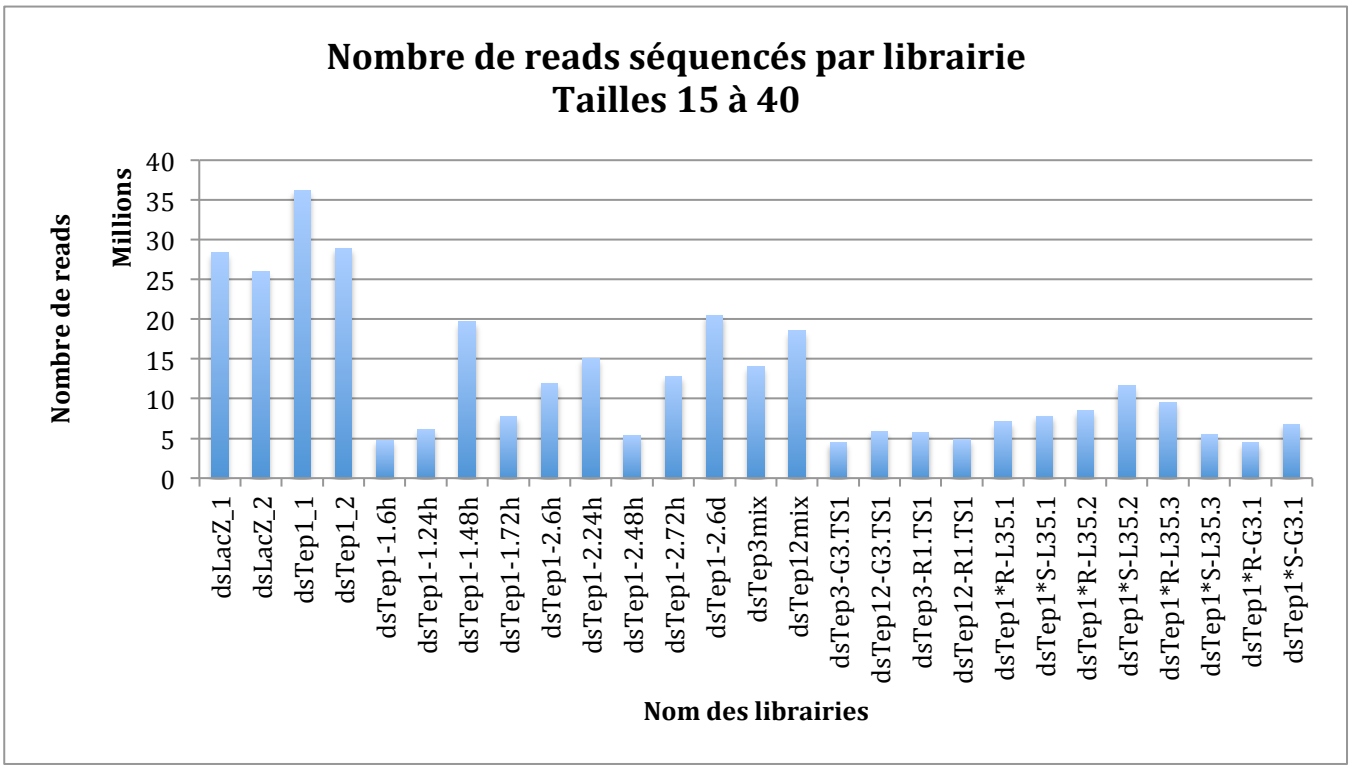


Figure 10. Nombres de reads de 15 à 40b séquencés pour chaque librairie réalisée.

Les variations du nombre de reads séquencés sont assez fortes (Fig. 10). Quelques librairies ne totalisent que 5-6 millions de reads alors que d'autres dépassent les 25 millions de reads séquencés. L'explication de ces variations se situe peut-être au niveau des préparations des librairies, entre l'injection et l'extraction des petits ARNs.

Profil des petits ARNs séquencés

Les petits ARNs de 21b à 29b représentent plus de la moitié des reads de 15 à 40b séquencés (Fig. 11). Parmi ces petits ARNs, on retrouve les microARNs (miRNAs) qui sont de petites séquences comprises entre 20 et 24pb, ensuite, de 25 à 29pb, on peut assimiler ces petites séquences aux piwi-ARNs (piRNAs). Le pic de 22-mers doit être principalement constitué de miRNAs appartenant au moustique.

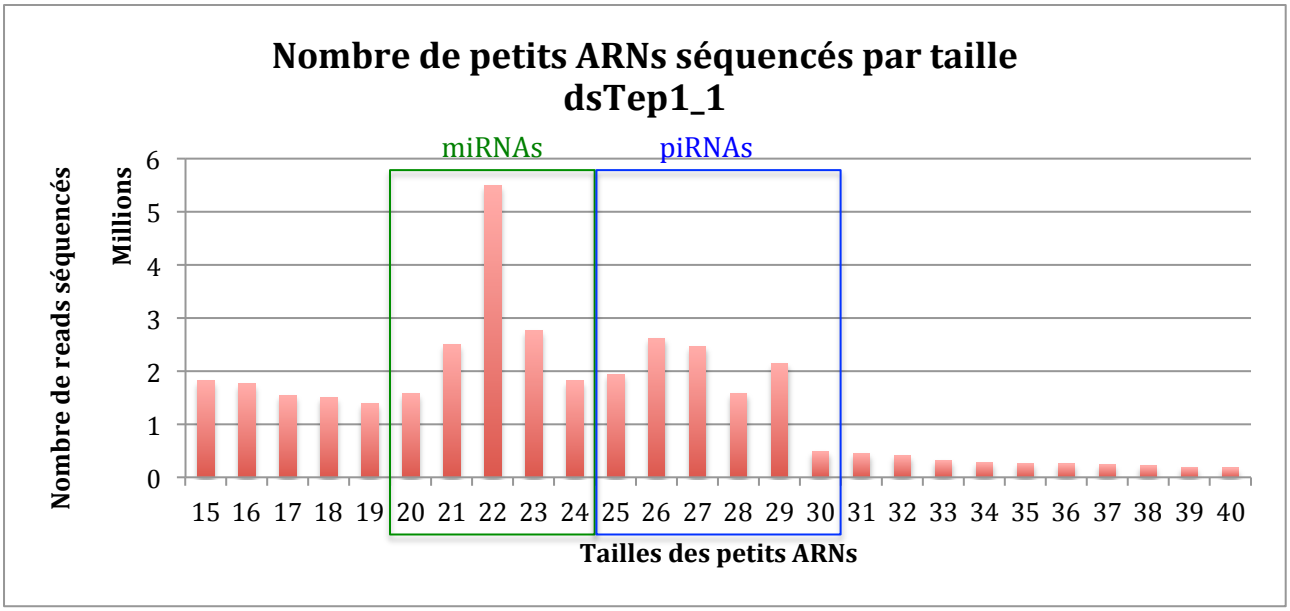


Figure 11. Distribution des petits ARNs séquencés par taille pour l'échantillon dsTep1_1. Les miRNAs sont encadrés en vert et les piRNAs en bleu.

En étudiant les profils de chacune des librairies, il est apparu deux types de profils différents. Le premier profil montre une majorité écrasante de reads de 28b séquencés par rapport aux autres tailles (Fig. 12) et le deuxième présente plusieurs pics de tailles séquencées : 22, 26, 27 et 29 bases (Fig. 13).

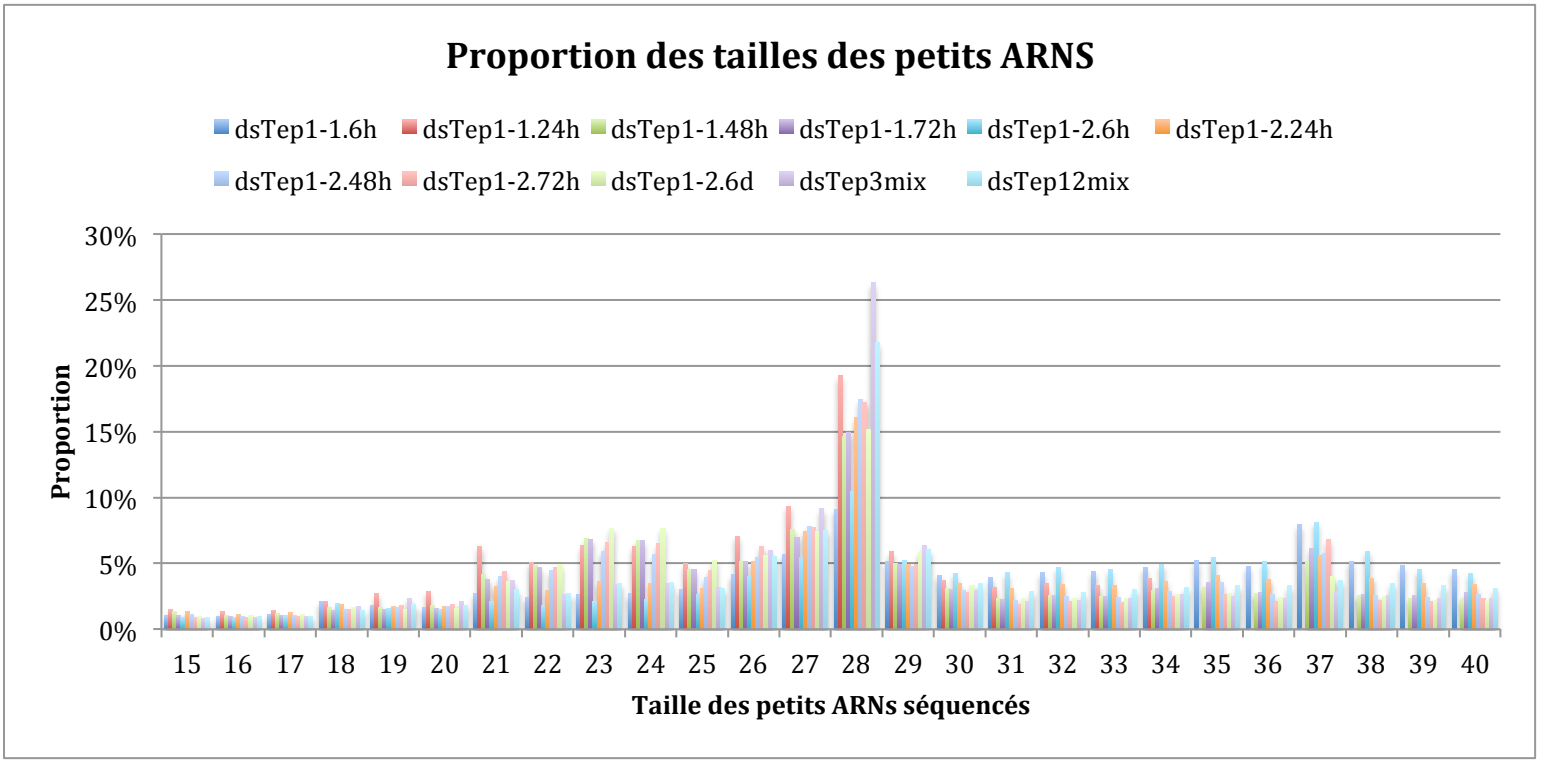


Figure 12. Distribution des tailles des petits ARN séquencés dans chaque échantillon présentant un pic de reads à 28b. La proportion a été calculée sur l'ensemble des reads d'une taille de 15 à 40b.

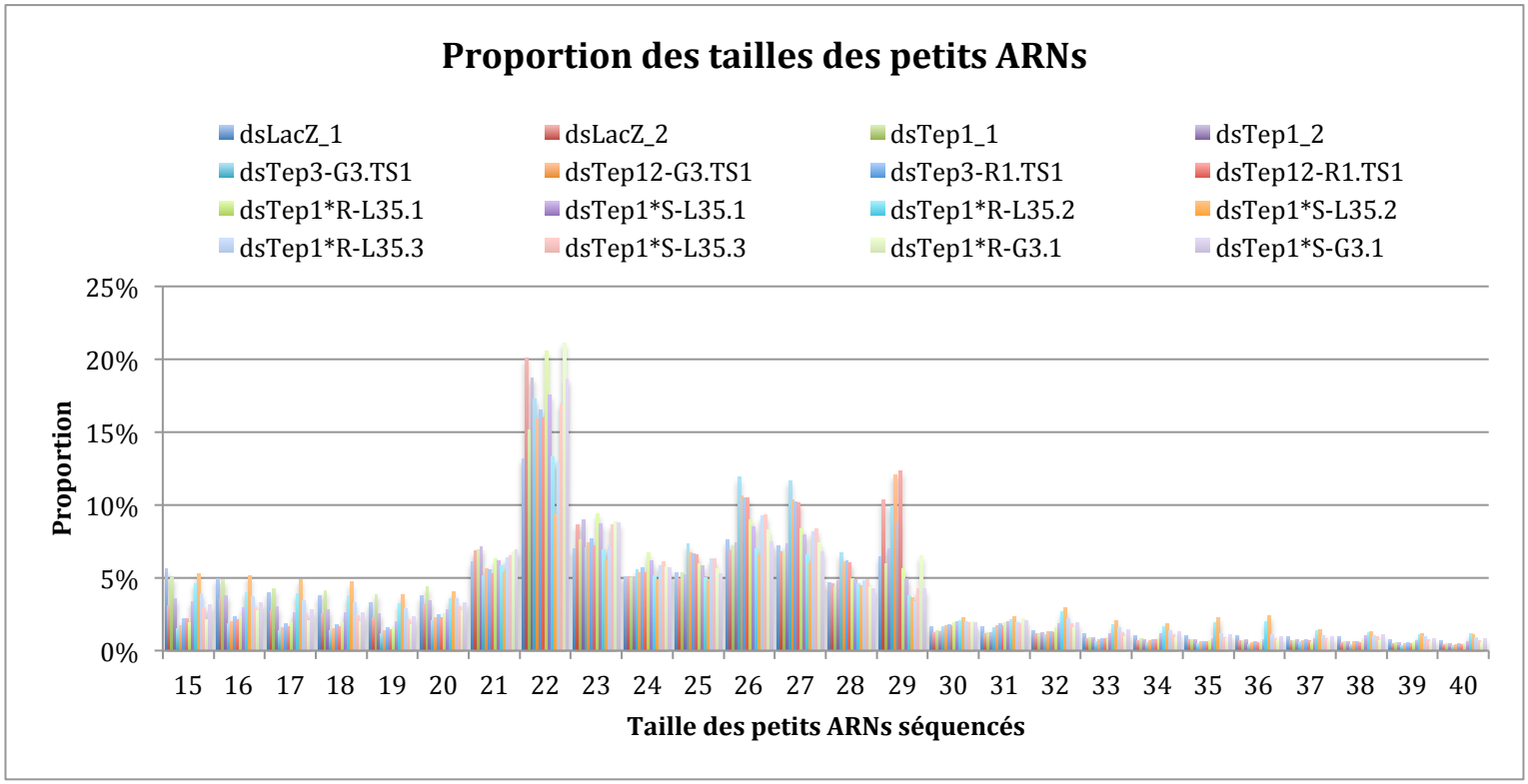


Figure 13. Distribution des tailles des petits ARN séquencés dans chaque échantillon présentant des pics de reads à 22, 26, 27 et 29b. La proportion a été calculée sur l'ensemble des reads d'une taille de 15 à 40b.

Nous ne remarquons pas de corrélation entre ces deux profils et l'allèle ou le gène ciblé. Le temps entre l'injection et l'extraction n'est pas mis en cause puisque tous les échantillons ciblant Tep1 et modifiant ce temps sont tous regroupés dans le premier profil (Fig. 12). De même, la lignée de moustique utilisée pour l'injection n'engendre pas un profil déterminé. Par contre, le premier profil regroupe les librairies C1HW5ACXX (dsTep1 au fil du temps) et D1UJHACXX (dsTep3/12mix) (Tableau 1) alors que le second profil regroupe les trois autres librairies D0YEFACXX, C2LVTACXX et C3PEGACXX (Tableau 1). Nous pouvons émettre l'hypothèse que ce soit la préparation de la librairie qui influe sur le type de petits ARNs produits, peut-être lors de la liaison des adaptateurs aux petits ARNs ou lors de la sélection des bandes de charges correspondant aux tailles de fragments d'environ 140 pb (miRNAs) et d'environ 150 pb (piRNA) (NEBNext® Multiplex Small RNA Library Prep Set for Illumina®).

Quel que soit le type de profil, les miRNAs et les piRNAs doivent principalement provenir du génome du moustique.

Profil des petits ARNs provenant du génome d'*Anopheles gambiae*

En alignant avec Bowtie les reads de chaque échantillon sur le génome d'*Anopheles gambiae* AgamP3, on obtient en moyenne 85% d'alignements dans les différentes librairies. Le profil de la distribution des tailles alignées est similaire pour tous les échantillons (Fig. 14).

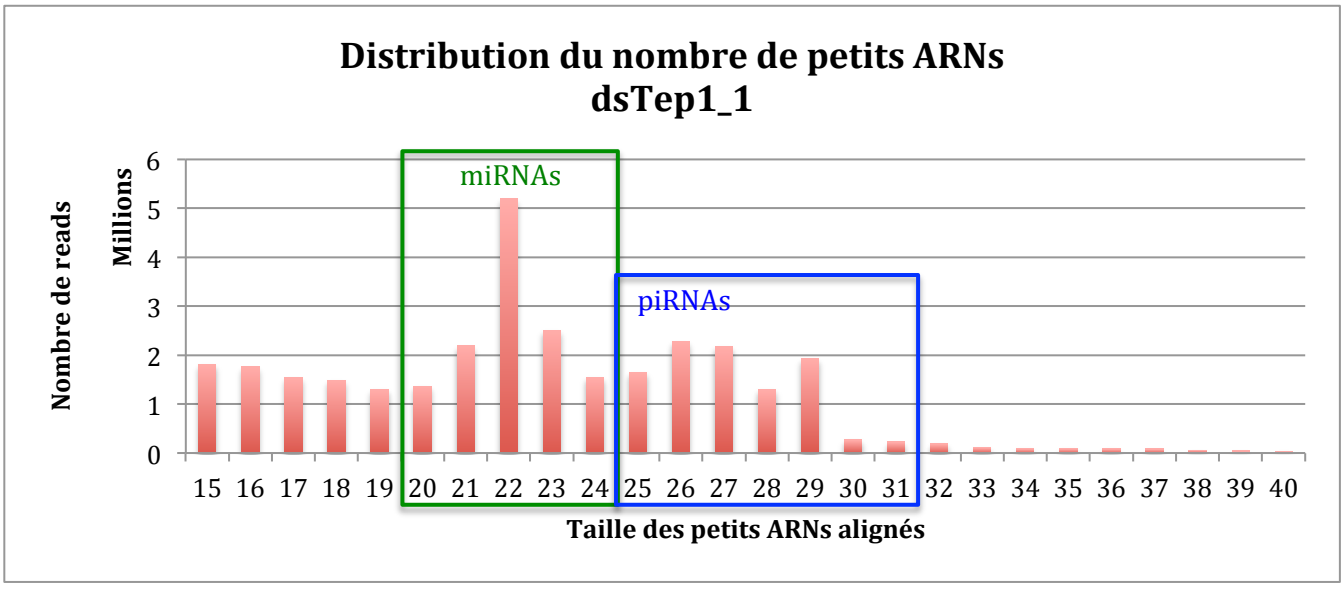


Figure 14. Distribution du nombre de petits ARNs de l'échantillon dsTep1_1 alignés sur le génome d'*Anopheles gambiae* en fonction de leur taille. Les tailles correspondant au microARNs sont encadrées en vert et celles des piwiARNs en bleu.

La majorité des petits ARNs de taille 21 à 29b appartiennent au génome du moustique. C'est le type de profil attendu puisque la catégorie des petits ARNs est composée des siRNAs (21pb), des miRNAs (20-24pb) et des piwiRNAs (25-30pb).
Il y a aussi plus de petites séquences de taille inférieure à 20b que de séquences de taille supérieure à 29b. Nous pouvons supposer que ce sont principalement des produits de dégradation du génome.

2. Mise en évidence du processus de découpage d'un dsRNA injecté

Afin de comprendre comment les cellules du moustique découpent un dsRNA, nous avons aligné les petits ARNs séquencés de 15 à 40b sur la séquence du dsRNA injecté correspondant.

2.1. Profil des petits ARNs provenant de la séquence du dsRNA

2.1.1. Disparités dans la proportion des petits ARNs alignés

Le nombre total de petits ARNs s'alignant parfaitement sur la séquence du dsRNA varie considérablement d'un échantillon à l'autre (Fig.15). Nous avons un minimum de 0,38% à un maximum de 22,29% de petits ARNs alignés sur l'ensemble des petits ARNS séquencés.

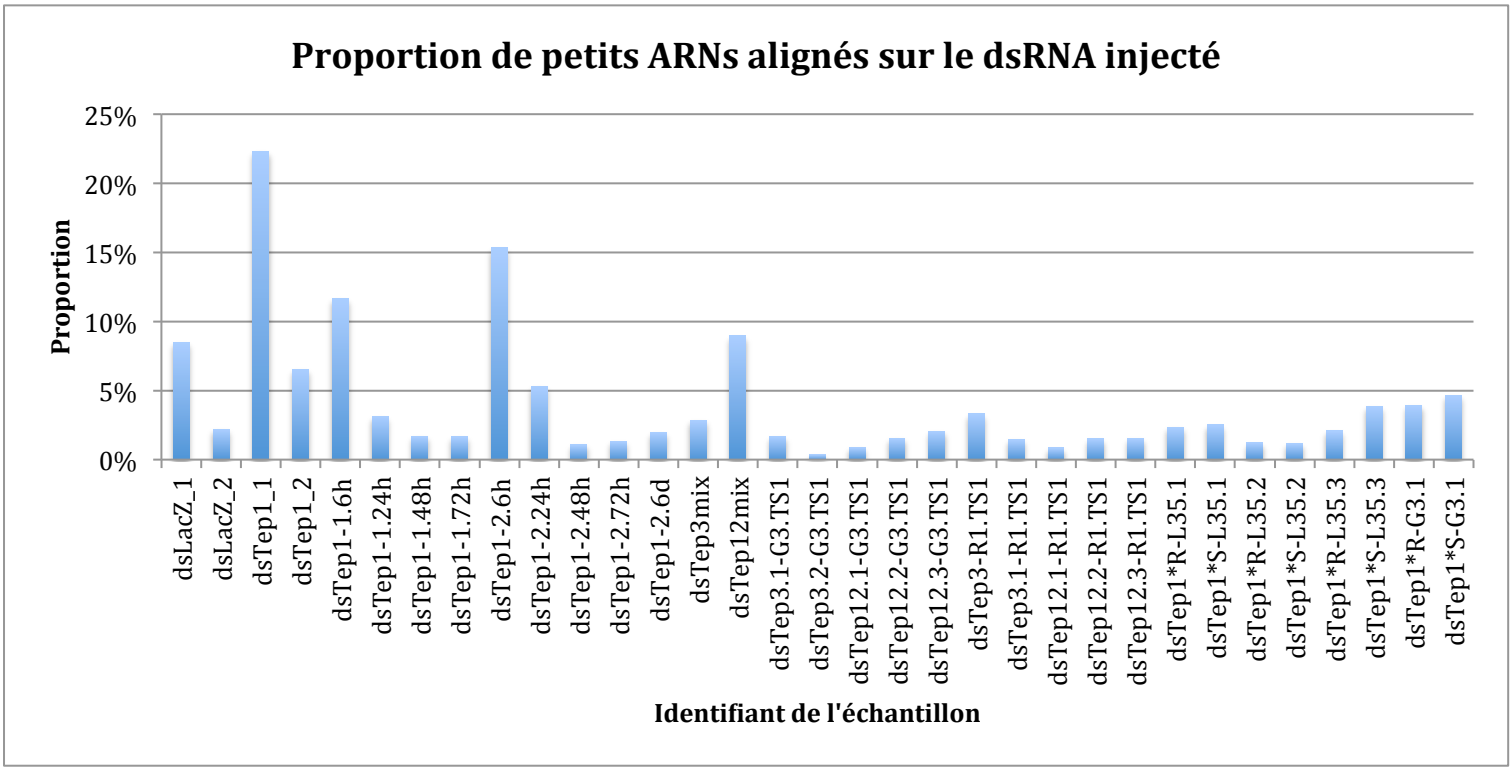


Figure 15. Proportion de petits ARNs de tous les échantillons s'alignant sur la séquence du dsRNA injecté correspondant.

Comme ces gènes sont endogènes (sauf LacZ), il est fort probable qu'un certain nombre de petits ARNs proviennent du gène et non du dsRNA injecté. Pour vérifier, nous avons aligné les petits ARNs provenant de l'échantillon dsLacZ_1 et _2 dont le dsRNA est dsLacZ sur le gène Tep1. En fait, seulement 0,1% des petits ARNs alignés appartiennent au gène Tep1, ce qui très faible et ne risque pas de gêner nos analyses. On peut supposer que ces 0,1% proviennent d'erreurs lors du démultiplexage de librairies. Les reads s'alignant sur les dsRNA proviennent bien du dsRNA injecté.

Il y a plusieurs explications possibles à ce phénomène dont une que nous pouvons vérifier : le temps passé entre l'injection et l'extraction des ARNs. Nous avons à notre disposition 5 échantillons où les ARNs ont été extraits 6 heures (h), 24 h, 48 h, 72 h et 6 jours après l'injection d'un dsRNA. Les quatre premières expériences ont des réplicats biologiques.
Cependant, nous venons de voir une grande variabilité dans le nombre total de petits ARNs s'alignant sur le dsRNA. Pour pouvoir comparer efficacement deux ou plusieurs échantillons, nous avons normalisé les données.

2.1.2. Normalisation du nombre de petits ARNs

Nous avons divisé le nombre de petits ARNs alignés de chaque taille par le nombre total de petits ARNs alignés sur le dsRNA. Puis, nous avons multiplié ces divisions par un million pour obtenir des parties par million ou ppm (Fig. 16).

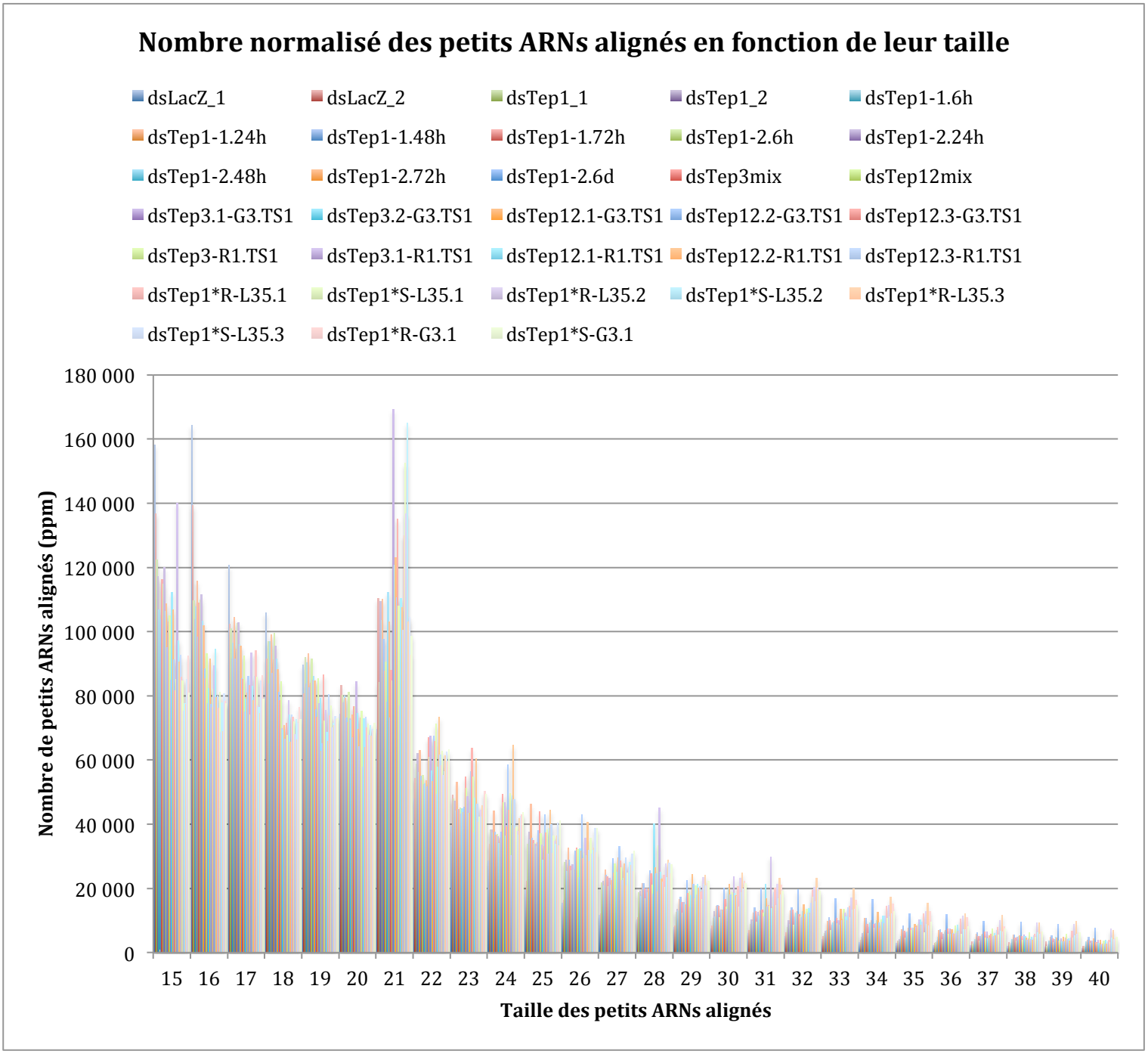


Figure 16. Normalisation du nombre de petits ARNs selon le nombre total de petits ARNs alignés sur le dsRNA.

Les dsRNAs ont majoritairement produit des petits ARNs de 21b, ce qui correspond à la taille des siRNAs attendus. Ce découpage a été observé la première fois sur un modèle in vitro de *Drosophila melanogaster* (Zamore et al., 2000). Cependant, malgré la normalisation, il existe encore une assez grande variation dans le nombre de 21-mers issus du dsRNA injecté.
On peut noter une valeur aberrante en taille 15 : 771 ppm pour l'échantillon dsTep1-1.24h. Le nombre de petits ARNs de taille inférieure ou égale à 15 chute, en effet, brutalement dans cet échantillon. Ce qui n'est pas le cas dans le réplicat biologique, dsTep1-2.24h. Je suspecte un problème lors du multiplexage ou du démultiplexage pour ces tailles.

Les deux échantillons qui sortent de l'ensemble par un nombre plus important de séquences de taille inférieure à 19b et un nombre plus faible de séquences de taille supérieure à 23b sont les échantillons dsLacZ_1 et dsLacZ_2 (Fig. 16). Les séquences de 20 à 22b étant dans la moyenne par rapport aux autres échantillons. Le fait que ce soit une séquence exogène joue peut-être sur son processus de découpage et le profil des petits ARNs produits, sauf pour la production des siRNAs à plus ou moins une base près.

Nous pouvons maintenant comparer nos échantillons entre eux et vérifier si le temps entre l'injection du dsRNA et celui de l'extraction des ARNs modifie le profil des petits ARNs issus du dsRNA.

2.1.3. Évolution de l'abondance des petits ARNs au cours du temps post-injection

Un même dsRNA ciblant Tep1 a été injecté dans des moustiques à un temps 0. L'extraction des petits ARNs a été effectuée à +6h, +24h, +48h, +72h avec des réplicats et +6j. L'alignement des petits ARNs sur le dsRNA dsTep1 montre que le nombre de petits ARNs issus du dsRNA chute d'environ 10% en 48h puis remonte légèrement (Fig. 17).

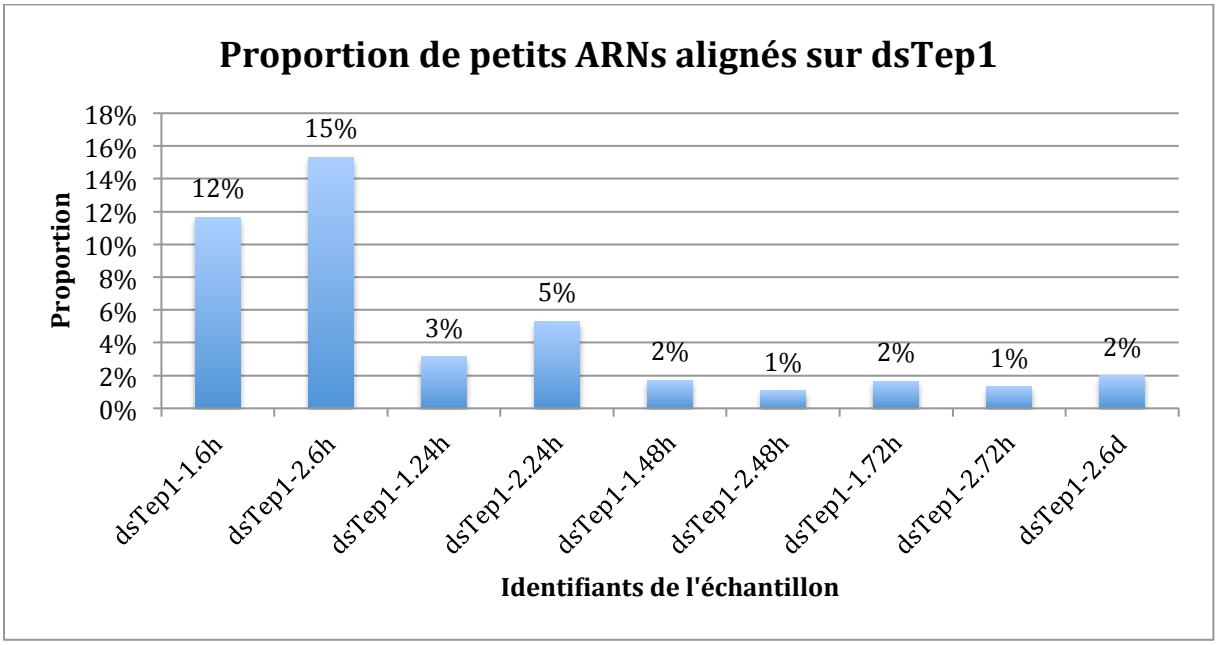


Figure 17. Proportion de petits ARNs alignés sur le dsRNA dsTep1 injecté.

Concernant la distribution des tailles des petits ARNs, on observe le même pic à 21b quel que soit la durée post-injection (Fig. 18 et 19). Par contre, l'évolution du nombre de 21-mers n'est pas du tout la même dans les deux réplicats (Fig. 18 et 19). On note de nouveau la présence de la valeur aberrante des 15-mers dans le premier réplicat (Fig. 18).

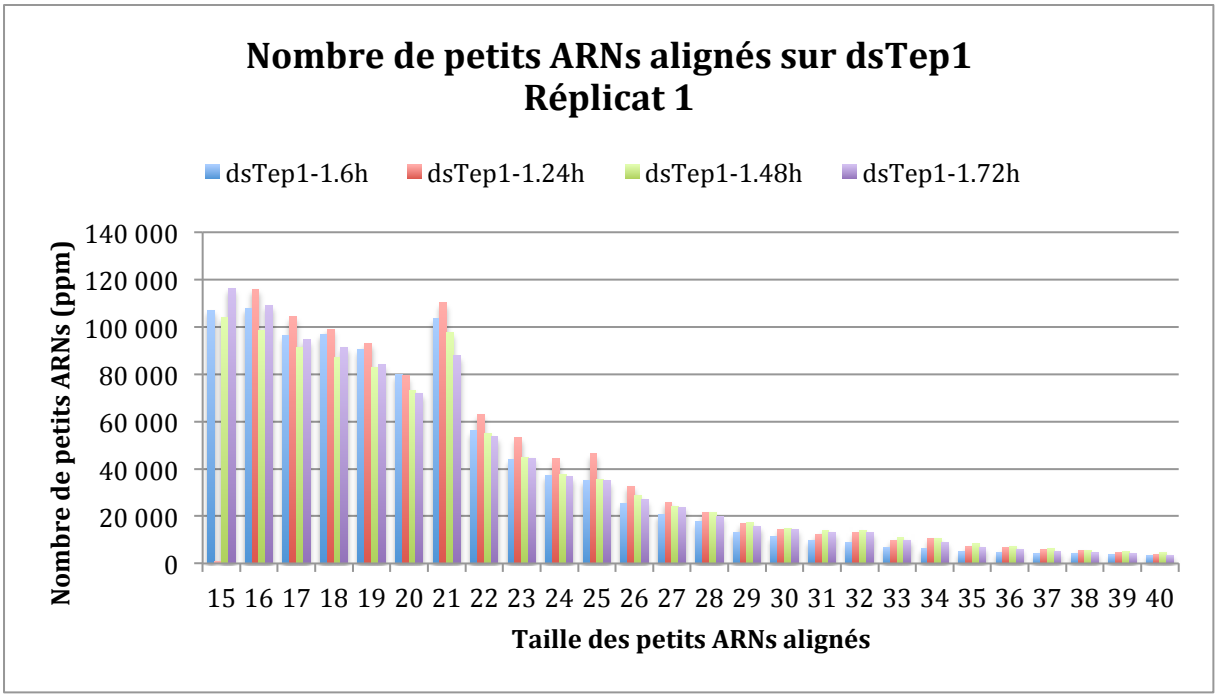


Figure 18. Nombre de petits ARNs de différentes tailles alignés sur dsTep1 en fonction du temps post-injection. Ces 4 échantillons font partie du réplicat 1.

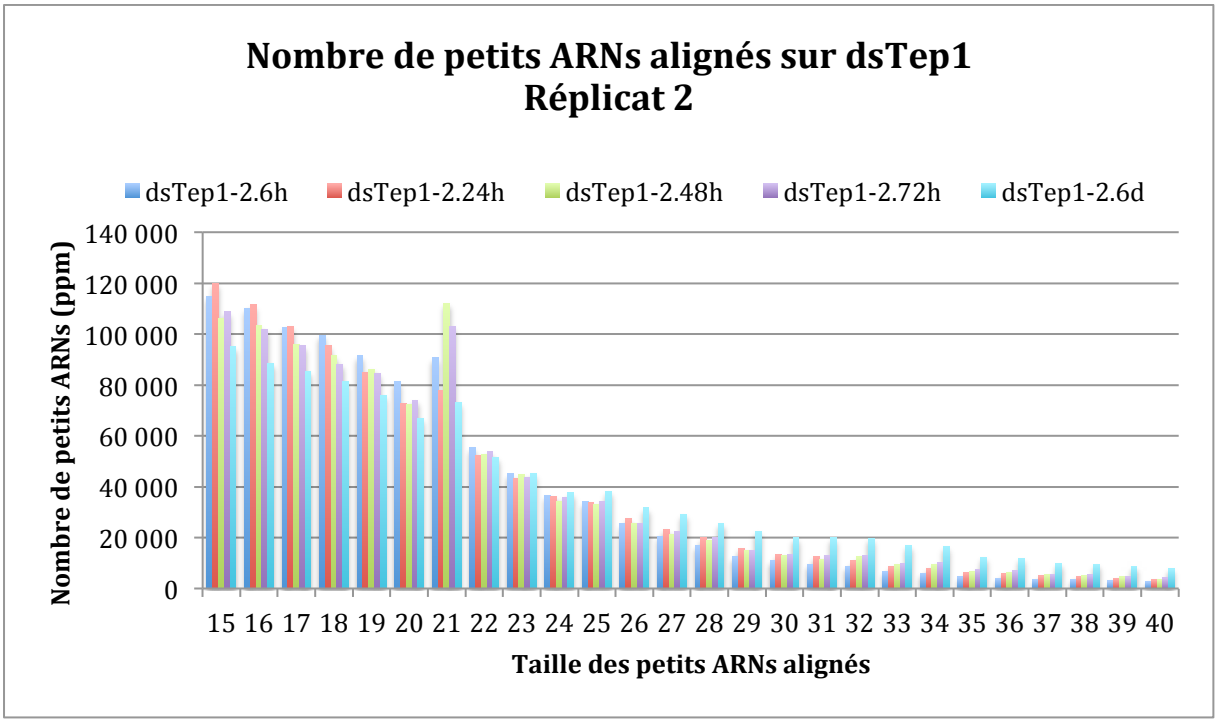


Figure 19. Nombre de petits ARNs de différentes tailles alignés sur dsTep1 en fonction du temps post-injection. Ces 5 échantillons font partie du réplicat 2.

Dans la première expérience, le nombre de 21-mers augmente dans les 24h post-injection puis diminue à +48h et +72h (Fig. 18). Alors que dans la seconde expérience, la quantité de 21-mers diminue dans les premières 24h, puis atteint son maximum à +48h pour réduire de nouveau (Fig. 19). Dans les deux cas, la quantité de 21-mers est plus faible à la dernière mesure enregistrée (+72h pour dsTep1-1 et +6j pour dsTep1-2) par rapport à la première mesure (+6h). il serait intéressant de réaliser un nouveau réplicat pour connaître la réelle évolution du nombre de 21-mers après l'injection d'un dsRNA.
D'autant plus qu'en étudiant l'évolution de toutes les tailles au fil du temps, c'est la seule qui présente autant de variations dans le temps (Fig. 20), dans le cas où l'on ne prend pas en compte la valeur aberrante des 15-mers de dsTep1-1.24h.

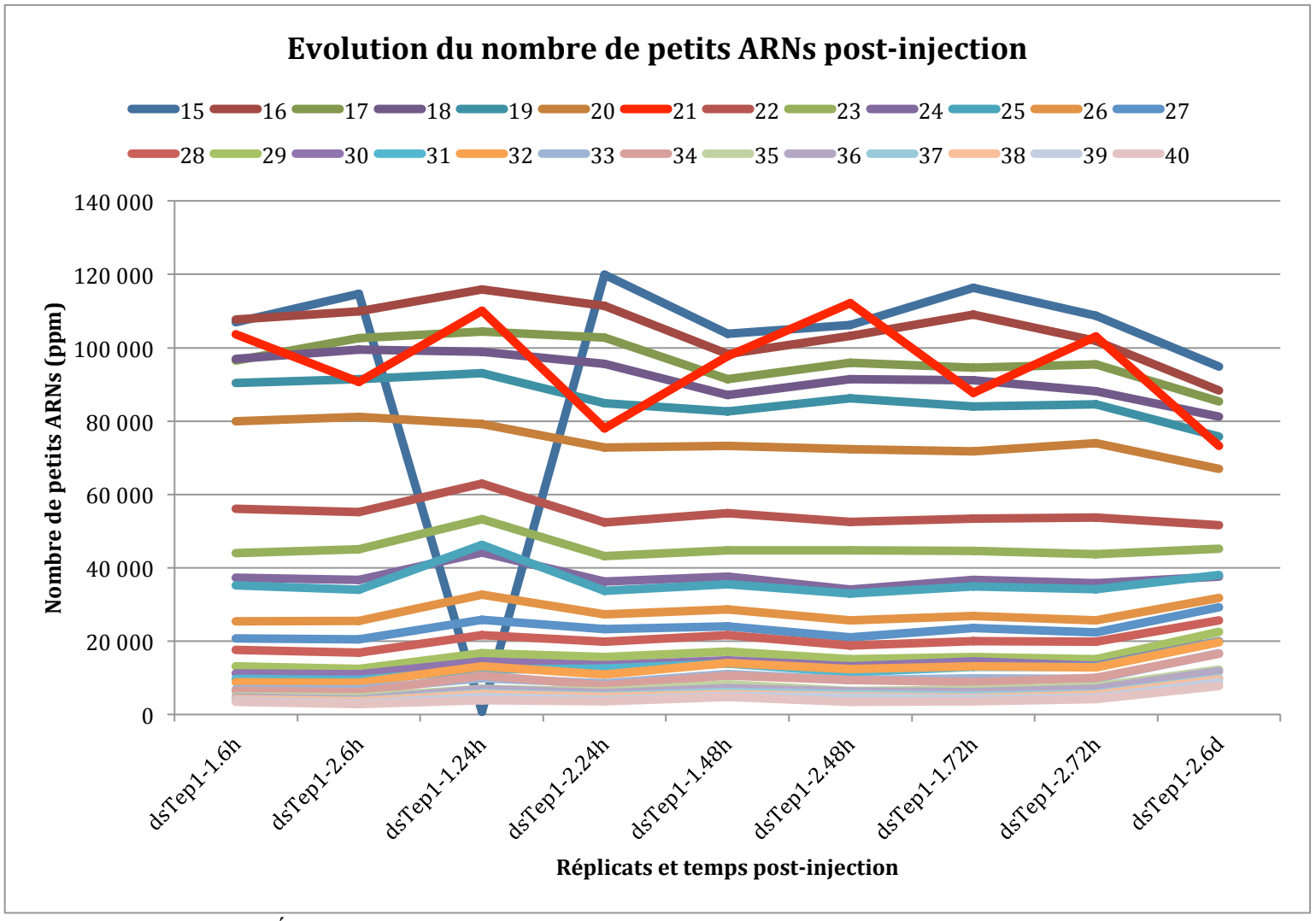


Figure 20. Évolution du nombre de petits ARNs post-injection de dsTep1 dans les deux réplicats. L'évolution du nombre de 21-mers est indiquée en rouge.

Pour les analyses suivantes tendant à étudier comment est traité le dsRNA injecté par les cellules, nous avons choisi de se focaliser sur les reads dont la taille se situe entre 19 à 24 bases puisque cet intervalle contient nos siRNAs de 21b.

2.2. Répartition hétérogène des petits ARNs sur la séquence du dsRNA injecté

Grâce au fichier d'alignement SAM fourni par Bowtie, j'ai pu placé la première position de chaque read qui s'est aligné sur la séquence du dsRNA correspondant injecté (Fig. 21).

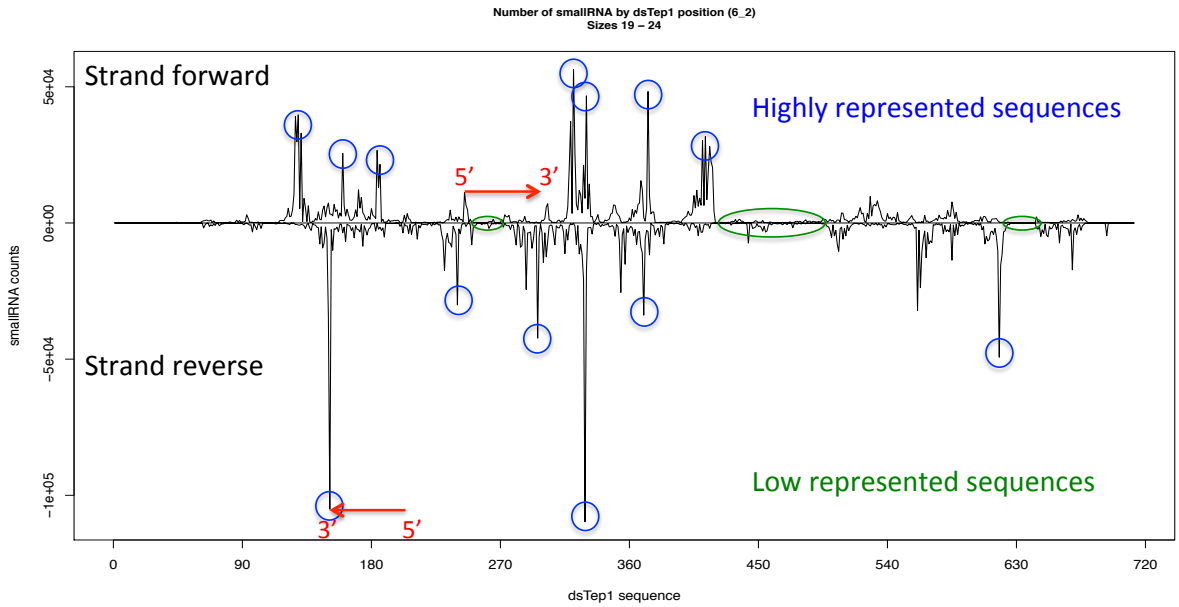


Figure 21. Répartition des petits ARNs des tailles 19 à 24b sur la séquence du dsRNA dsTep1 (échantillon dsTep1_1). En positif sont représentés les reads s'alignant sur le brin sens du dsRNA, et en négatif les reads s'alignant sur le brin anti-sens. L'orientation des reads placés est signalée par les indicateurs rouges. Les cercles bleus indiquent les pics et donc les reads surreprésentés. Les cercles verts indiquent des zones ayant un faible nombre de reads alignés.

La répartition des reads sur la séquence du dsRNA a été réalisée pour chaque échantillon. L'étude de chaque graphe a conduit aux mêmes observations et conclusions. Pour l'exemple, j'ai pris l'échantillon dsTep1_1 (Fig. 21). La première observation est l'alignement hétérogène des petits ARNs sur la séquence du dsRNA. Alors que certaines régions ont été nommées zones « déserts » en raison du peu ou du non alignement des petits ARNs, on remarque que certains petits ARNs sont surreprésentés par rapport aux autres et forment des pics facilement identifiables. De plus, le profil des petits ARNs issus du dsRNA est consistant entre les réplicats (Fig. 22) et au fil du temps (Fig. 23).

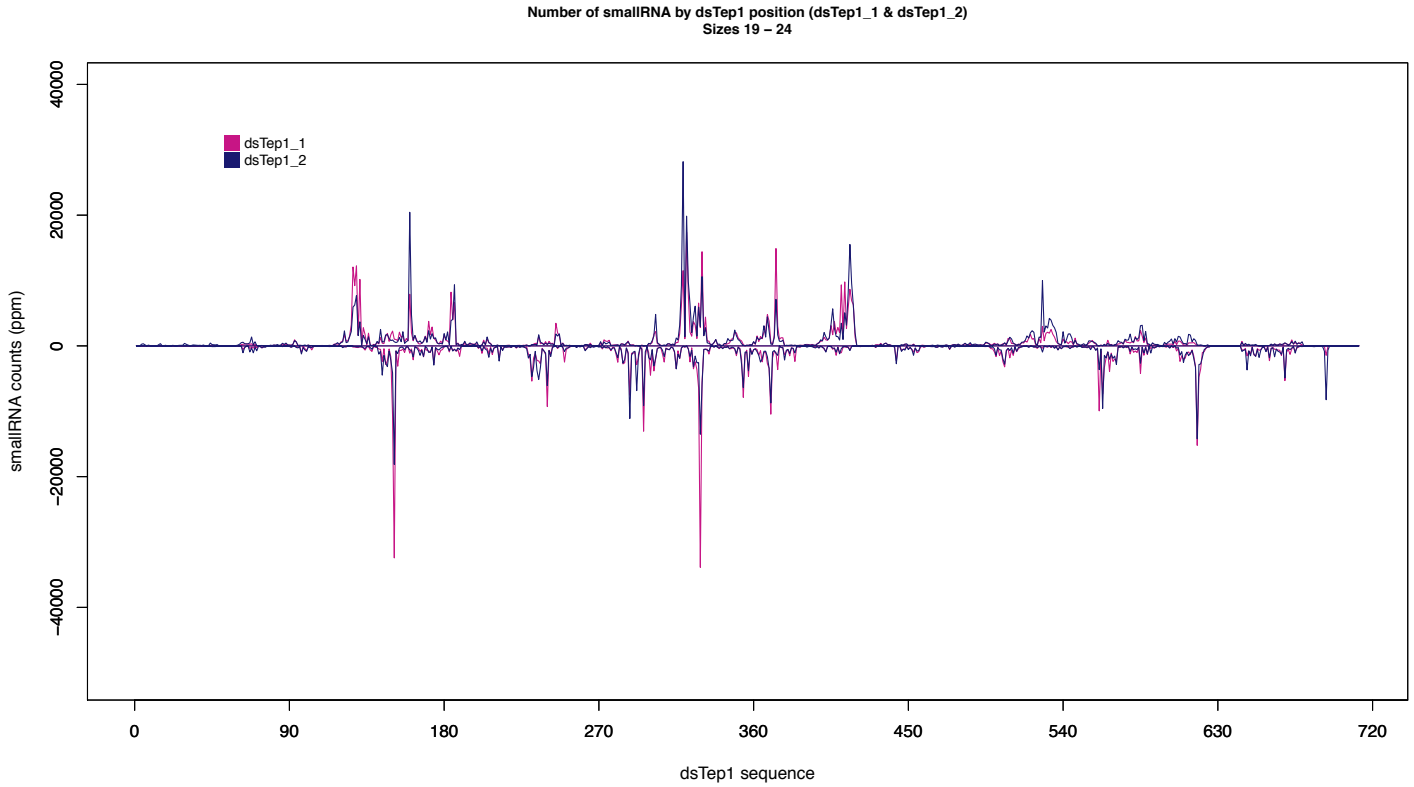


Figure 22. Répartition des petits ARNs de tailles 19 à 24b sur la séquence de dsTep1. Les deux réplicats sont représentés : dsTep1_1 et dsTep1_2.

Dans tous les cas ou nous avons réalisé des réplicats biologiques, les profils des petits ARNs obtenus sont identiques au niveau des séquences surreprésentées, les pics, ainsi qu'au niveau des régions « déserts » (Fig.22). Pour les mêmes petits ARNs surreprésentés, seule la quantité varie d'un réplicat à l'autre.

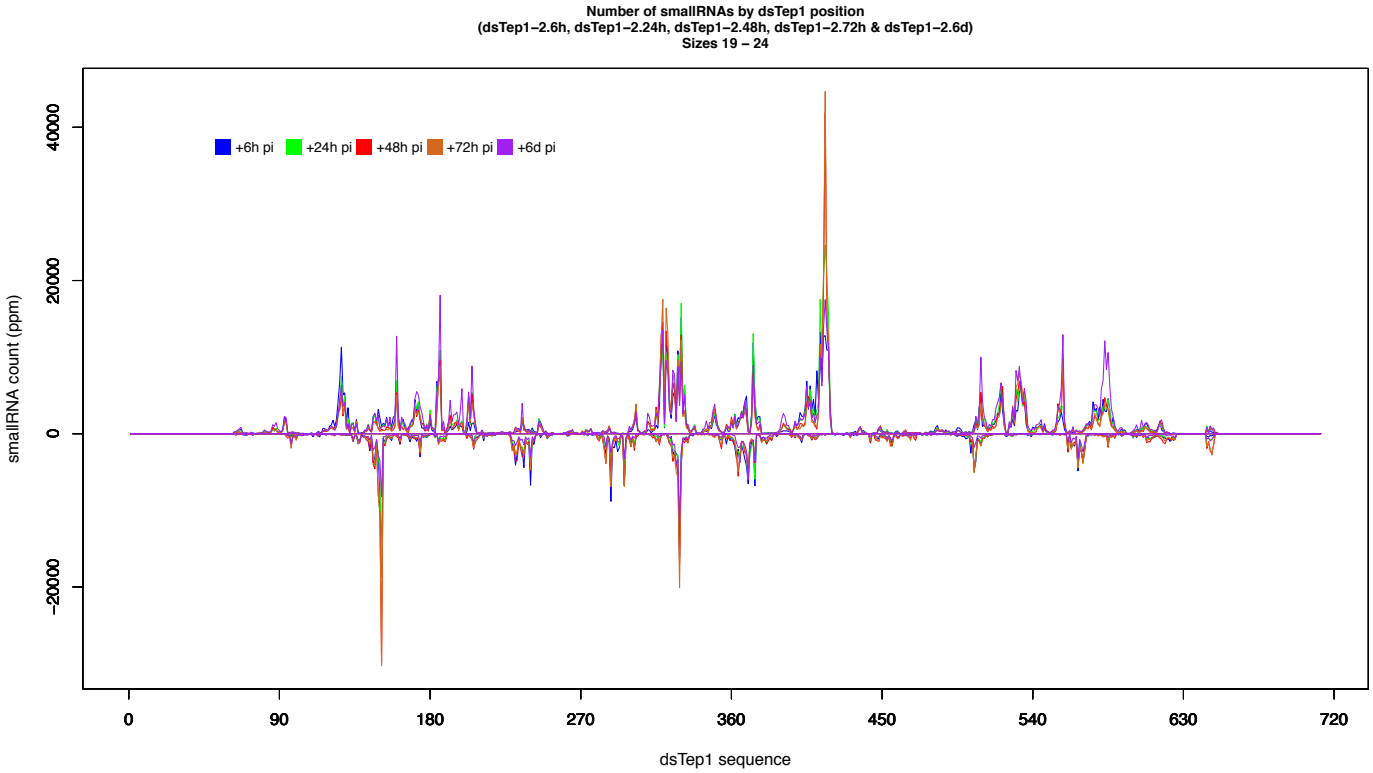


Figure 23. Répartition des petits ARNs de tailles 19 à 24b sur la séquence de dsTep1 au cours de différents temps post-injection : +6h, +24h, +48h, +72h et +6j.

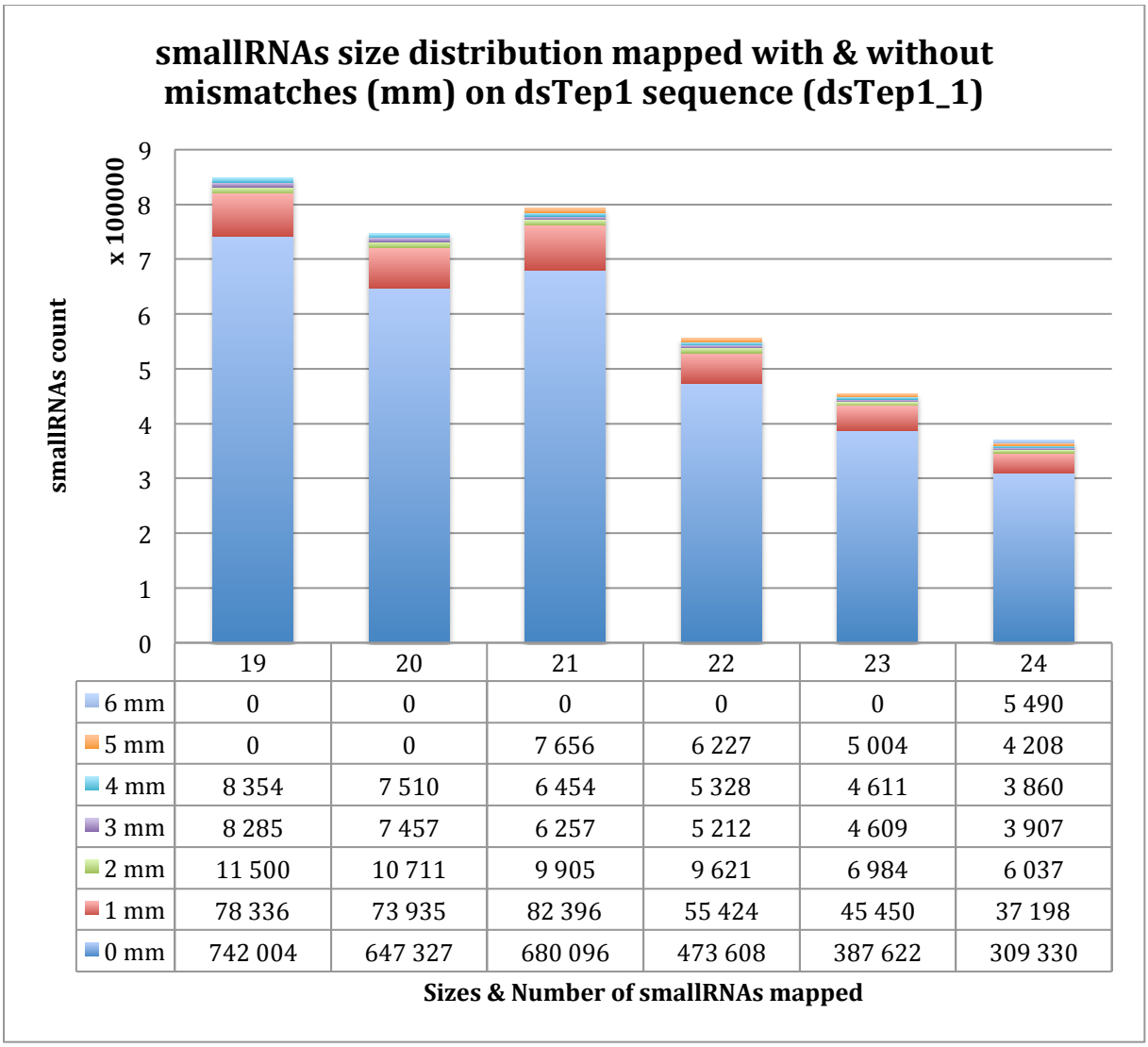
De même, les petits ARNs surreprésentés et les régions « déserts » restent identiques quel que soit le nombre d'heures passées après l'injection du dsRNA (+6h, +24h, +48h, +72h et +6j, Fig. 23). Un petit ARN surreprésenté dès le début est conservé tout au long des 6jours post-injection. On peut émettre l'hypothèse que ces petits ARNs sont protégés, peut-être parce qu'ils sont chargés dans le complexe RISC-Ago2.

Ce profil des petits ARNs alignés sur chaque dsRNA injecté a été formé à partir d'alignements parfaits. Je n'ai pas autorisé d'erreurs entre les reads et la séquence du dsRNA pour être sûre que les reads s'alignant provenaient bien du dsRNA injecté.
Nous avons donc cherché à savoir ensuite si les reads avec erreurs modifiaient le profil de la distribution des reads sur le dsRNA.

2.3. Augmentation du nombre de substitutions sur l'extrémité 3' des petits ARNs

Les reads ont donc été réalignés sur chaque dsRNA par Bowtie en autorisant des erreurs lors de l'alignement. J'ai décidé d'autoriser le plus de mésalignements possibles entre la totalité de la séquence du read et la référence sachant que Bowtie me propose le meilleur alignement avec les limites imposées. En effet, Bowtie ne propose que le meilleur alignement pour chaque read après avoir testé 125 positions d'alignement différentes et que la somme des valeurs de qualité des mésalignements ne doit pas dépasser la valeur de 70. Ces valeurs de qualité sont attribuées pour chaque base par le séquenceur et vont de 0 à 41. Plus le score est élevé, plus la probabilité qu'il s'agisse d'une mauvaise lecture de la base diminue.

Dans le cas de dsTep1, Bowtie a aligné de 14 à 19% de petits ARNs, de taille 19 à 24, en plus par rapport au nombre total de petits ARNs s'alignant parfaitement (Fig. 24).



	19	20	21	22	23	24
6 mm	0	0	0	0	0	5 490
5 mm	0	0	7 656	6 227	5 004	4 208
4 mm	8 354	7 510	6 454	5 328	4 611	3 860
3 mm	8 285	7 457	6 257	5 212	4 609	3 907
2 mm	11 500	10 711	9 905	9 621	6 984	6 037
1 mm	78 336	73 935	82 396	55 424	45 450	37 198
0 mm	742 004	647 327	680 096	473 608	387 622	309 330

Figure 24. Nombre de petits ARNs de tailles 19 à 24 s'alignant sur la séquence dsTep1 avec et sans substitutions pour l'échantillon dsTep1_1. Les petits ARNs s'alignant ont été partagés en 7 groupes comprenant, du premier en bas au dernier en haut, de 0 à 6 erreurs (ou mismatch, mm).

La répartition de ce nouvel ensemble de petits ARNs sur la séquence dsTep1 montre qu'il n'y a pas de nouveau pic mais plutôt des contributions aux pics existants (Fig. 25). Ce qui signifie que ce sont les mêmes séquences issues du découpage du dsRNA mais qu'elles ont subi des substitutions.

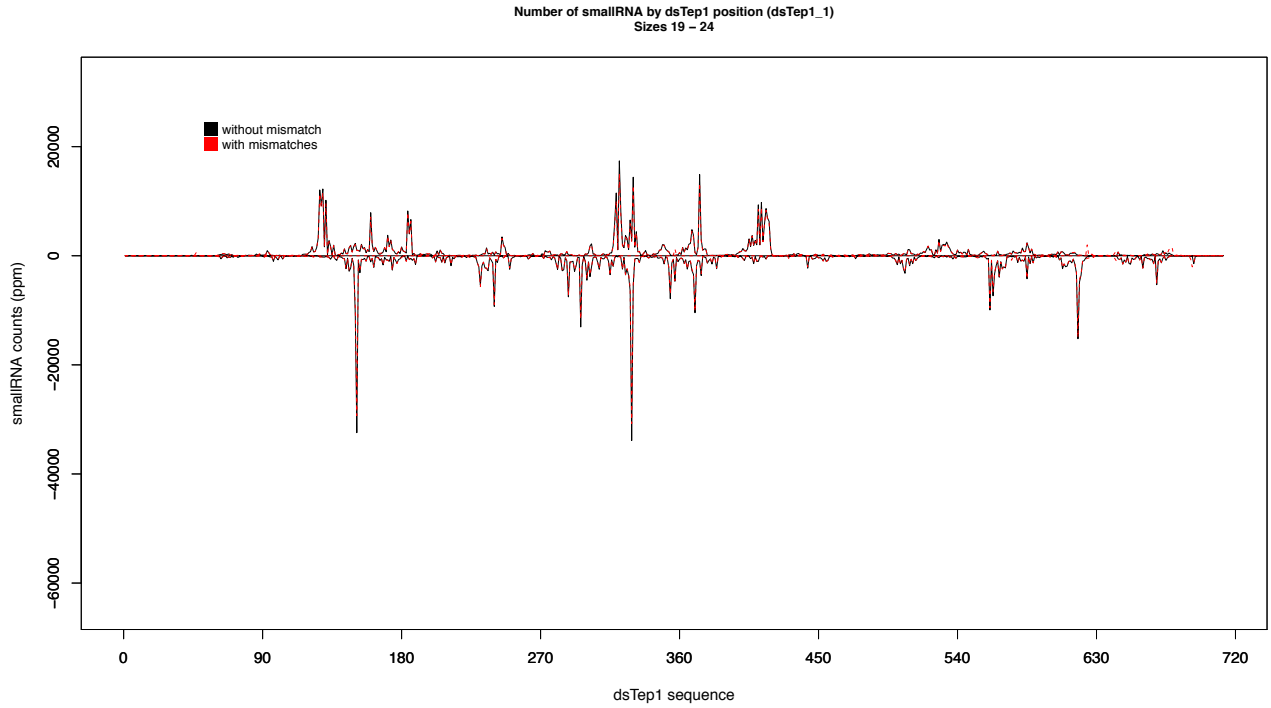


Figure 25. Répartition des petits ARNs s'alignant avec (en rouge) et sans erreurs (en noir) sur la séquence dsTep1. Les petits ARNs proviennent de l'échantillon dsTep1_1.

Ces mêmes observations ont été faites sur l'ensemble des échantillons dsLacZ, dsTep3 et dsTep12.

Nous avons ensuite regardé où étaient localisées les substitutions dans les petits ARNs (Fig. 26).

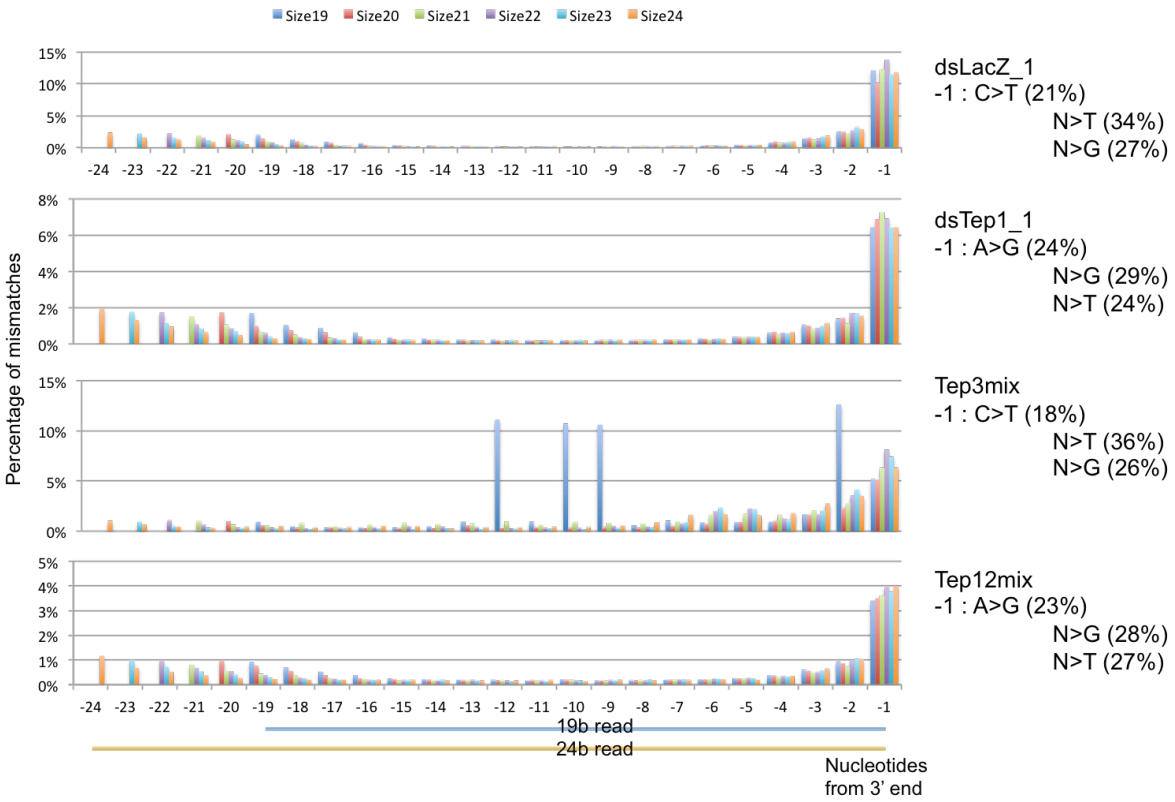


Figure 26. Localisation des substitutions sur les séquences de 19 à 24b pour les dsRNAs dsLacZ, dsTep1, dsTep3 et dsTep12. Pour chaque échantillon, les bases sont comptabilisées à partir de la première base de l'extrémité 3' de la séquence. Les reads de 19b se termineront alors avant les reads de 24b, comme on peut le voir en bas de l'image. L'axe est négatif pour indiquer le sens 3'-5' de droite à gauche. A droite, sont représentées les substitutions majeures qui se produisent à la première base de l'extrémité 3' pour chaque échantillon.

Pour les quatre dsRNA injectés, les substitutions se produisent principalement sur la première base de l'extrémité 3' dont le taux avoisine les 8% contre 1% en moyenne sur la totalité. Le taux de substitutions par position baisse pour être inférieur à 1% de la 4ème à la 17ème bases incluse, pour ensuite remonter légèrement sur l'extrémité 5'. Pour dsTep3, on a observé un très fort taux de substitutions sur les positions -2, -9, -10 et -12. Je n'ai pas d'explications sur ce phénomène.
Au niveau du type de substitutions, on a observé deux cas. Le premier regroupe dsLacZ et dsTep3 et présente une substitution en faveur d'un U à 34 et 36% respectivement. Le second groupe compte dsTep1 et dsTep12 et présente une substitution en faveur d'un G, à 29 et 28% respectivement. Les substitutions principales sont en fait des transitions d'un C>T pour le premier groupe et d'un A>G pour le second. Les transitions sont des changements entre soit deux purines, soit deux pyrimidines et les transversions sont des changements entre une purine et une pyrimidine. En général, les transitions sont deux fois plus courantes que les transversions (Vol'kenshtein, 1976).

Nous nous sommes attachés à comprendre pourquoi les dernières bases et, en particulier la dernière, présentaient un taux de substitutions assez fort ?

Nous savons que les dernières bases sont intégrées dans le domaine PAZ de Ago2 (Fig. 27). Après le largage du brin passager, cette extrémité 3' est 2'-O-méthylée par la protéine méthyltransférase Hen1 (Hua Enhancer1) (Yang et al., 2006). Cette méthylation améliore la stabilité du guide dans le complexe et empêche toute modification de cette extrémité.

Figure 27. Positionnement du guide dans le complexe RISC et lien avec l'ARNm cible. La première base de l'extrémité 5' s'ancre dans le domaine MID et l'extrémité 3' dans le domaine PAZ. Cette extrémité 3' est 2'-O-méthylée pour la protéger de toute modification. De plus, elle améliore la stabilité du guide dans le complexe.

En fait, les quatre dernières bases d'un guide ne contribuent que très peu lors de l'appariement des bases du guide à l'ARNm complémentaire et ne jouent pas sur l'efficacité de l'inhibition (Hong et al., 2014; Wee et al., 2012). Par contre, cette extrémité 3' est délogée de son domaine lors de l'appariement à cause de la structure en hélice d'Ago2 et se retrouve donc libre (Wang et al., 2009).

Plusieurs explications sont donc possibles à ce fort taux de substitutions en 3' :

(1) Dans les cas où la méthyl transférase Hen1 ne joue pas son rôle, les siRNAs qui ne sont pas méthylés donc pas protégés peuvent être sujets à des modifications de l'extrémité 3' lors du délogement du domaine PAZ. Dans ce cas, il a été montré chez *Chlamydomonas reinhardtii* et *Arabidopsis* que la nucléotidyl transférase Hen1 Suppressor1 (Heso1) ajoute principalement des uridyl à cette extrémité 3' et que cette uridylation amenait ensuite à la dégradation du siRNA (Ibrahim et al., 2010; Zhao et al., 2012). Une autre voie de dégradation des siRNAs non méthylés est le découpage par une exonucléase de l'extrémité 3' vers la 5'.

(2) Dans le cas où la voie des siRNAs est surchargée, les siRNAs peuvent être pris en charge par la voie des miRNAs et liés à Ago1. Dans cette voie, les miRNAs ne sont pas protégés et sont plus souvent modifiés. Ils peuvent subir une uridylation et une activité exonucléase 3'-5' (Ameres et al., 2010).

Ces substitutions peuvent avoir plusieurs conséquences qui ont été observées par (De et al., 2013) chez les mammifères. Cette étude a montré que les erreurs en 3' :

- stabilisent les interactions entre Ago2 et le siRNA,
- facilite le largage du brin d'ARNm coupé,
- améliorent l'inhibition de l'expression de cibles fortement exprimées.

Nous venons de voir que le découpage du dsRNA produisait un petit nombre de séquences avec substitutions qui contribuent aux pics déjà existants. Il n'y a donc pas de modification du profil des reads issus du découpage d'un dsRNA. De plus, ces substitutions, en majorité des transitions, sont principalement situées à l'extrémité 3' de ces séquences. Leur rôle chez le moustique n'a pas été déterminé.
Nous avons ensuite cherché si la présence d'un ou plusieurs SNPs dans le dsRNA injecté pouvait modifier le profil des petits ARNs produits.

2.4. Évaluation de la présence de SNPs dans la séquence du dsRNA sur le profil des petits ARNs produits

Cette partie a été menée dans le but de répondre à deux questions :

(1) Est-ce que l'environnement génétique du moustique modifie le profil des petits ARNs issus du dsRNA injecté ? En d'autres termes, la présence de l'allèle ciblé influence-t'elle le processus de découpage d'un dsRNA ?
(2) Est-ce qu'un ou plusieurs SNP(s) modifie le profil des petits ARNs issu du dsRNA injecté ? Dans ce cas, nous aurions des petits ARNs surreprésentés et des régions « déserts » différents.

Nous avons donc (1) comparer le découpage du même dsRNA dans un environnement génétique différent, i.e. le découpage du dsRNA *S, par exemple, dans des moustiques sensibles qui possèdent l'allèle ciblé et dans des moustiques résistants qui n'ont pas cet allèle, et (2) comparer le découpage de deux dsRNAs différents dans un même environnement génétique, i.e. le découpage des dsRNAs *S et *R dans les moustiques sensibles par exemple.

Pour ce faire, nous avons sélectionné le gène TEP1 pour son fort polymorphisme et les gènes TEP3 et TEP12 pour leur faible polymorphisme. Puis nous avons réalisé les protocoles suivants :

(1) pour TEP1 : un dsRNA ciblant l'allèle TEP1*S a été injecté dans des moustiques sensibles G3 et résistants L3-5 et un dsRNA ciblant l'allèle TEP1*R a été injecté dans les mêmes lignées (Fig. 28),
(2) pour les gènes TEP3 et TEP12 : les injections des dsRNAs ciblant les allèles *S et *R ont été réalisées dans la lignée sensible TS1 (Fig. 28).

Il a été injecté 2 dsRNAs différents ciblant le gène TEP3 (dsTep3.1, dsTep3.2) et 3 dsRNAs différents ciblant le gène TEP12 (dsTep12.1, dsTep12.2 et dsTep12.3).

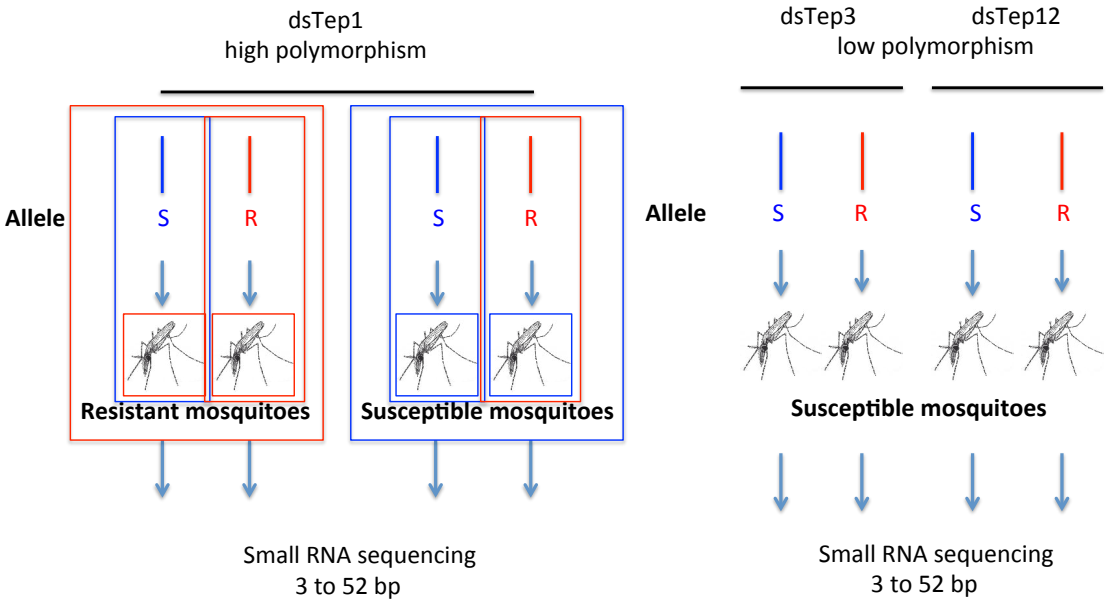


Figure 28. Protocole des injections des dsRNAs ciblant les allèles sensibles *S et résistants *R des gènes TEP1, TEP3 et TEP12 dans des lignées de moustiques résistantes (L3-5) et sensibles (G3 pour TEP1 et TS1 pour TEP3 et TEP12). Le phénotype résistant est associé à la couleur rouge et le sensible à la couleur bleu. Le type de séquençage réalisé est aussi indiqué.

Parmi les échantillons séquencés pour cette partie, nous avons 2 réplicats pour dsTep1*R/*S injecté dans des moustiques L3-5 et G3 (dsTep1*R-L35.1, dsTep1*S-L35.1, dsTep1*R-L35.2, dsTep1*S-L35.2, dsTep1*R-G3.1, dsTep1*S-G3.1, dsTep1*R-G3.2, dsTep1*S-G3.2) et une seule expérience pour les autres : dsTep3-G3.TS1, dsTep12-G3.TS1, dsTep3-R1.TS1 et dsTep12-R1.TS1.

Pour répondre à la première question (Est-ce que l'environnement génétique du moustique modifie le profil des petits ARNs issus du dsRNA injecté ?), nous avons aligné les petits ARNs de 21b, qui correspondent aux siRNAs, sur les séquences des dsRNAs injectés correspondants. Nous avons donc 2 graphes qui représentent :

(1) le dsRNA dsTep1*R injecté dans des moustiques résistants et des moustiques sensibles (Fig. 29),
(2) le dsRNA dsTep1*S injecté dans des moustiques résistants et des moustiques sensibles (Fig. 30),

Chacune des injections compte un réplicat biologique.

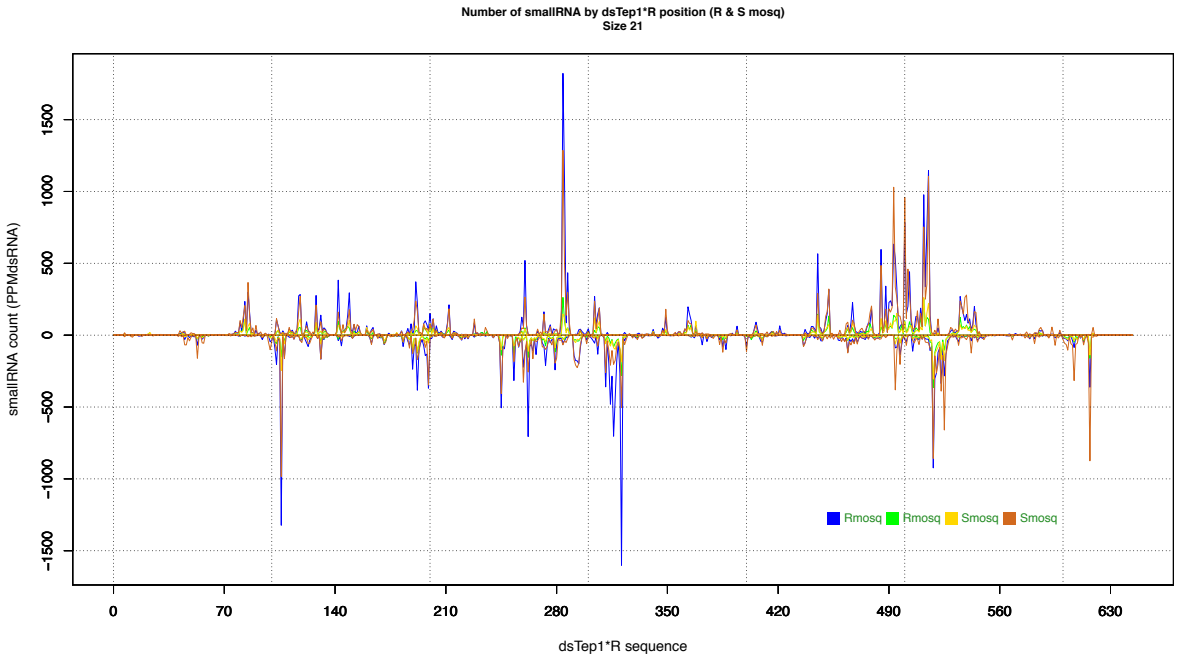


Figure 29. Répartition des petits ARNs de 21b issus du découpage de dsTep1*R injecté dans des moustiques résistants (en bleu et en vert, le réplicat) et dans des moustiques sensibles (en jaune et en marron, le réplicat).

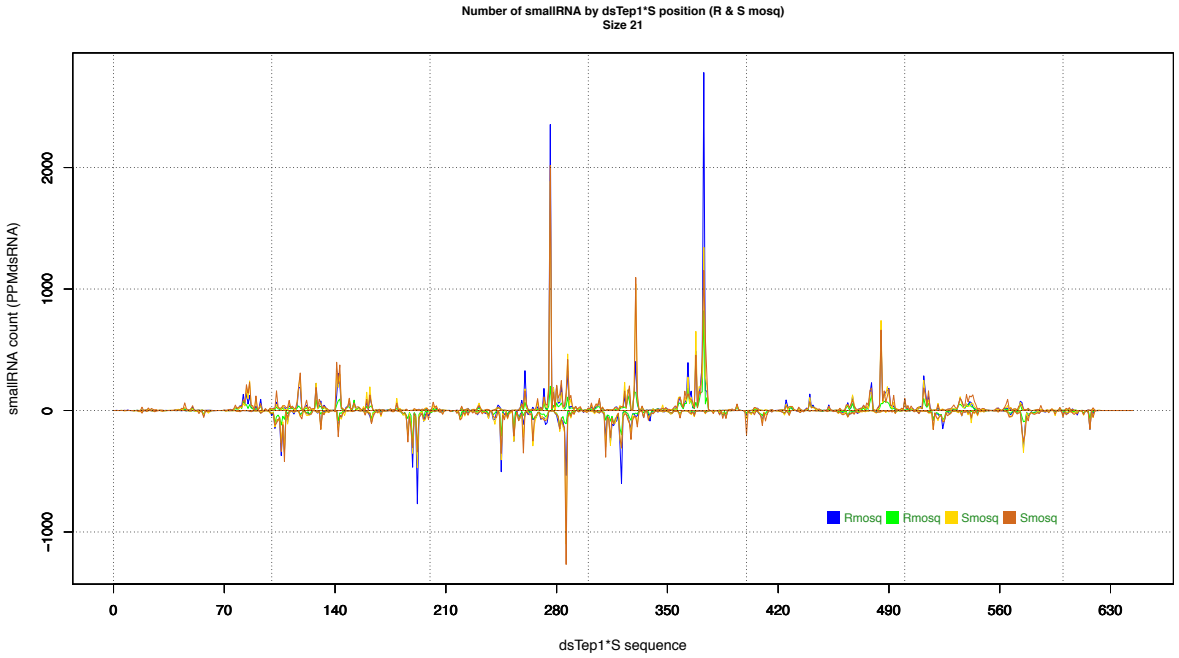


Figure 30. Répartition des petits ARNs de 21b issus du découpage de dsTep1*S injecté dans des moustiques résistants (en bleu et en vert, le réplicat) et dans des moustiques sensibles (en jaune et en marron, le réplicat).

Les profils des petits ARNs, i.e. les pics et les déserts, provenant du même dsRNA sont similaires quelque soit l'environnement génétique, résistant ou sensible, du moustique. La présence de l'allèle ciblé dans l'environnement génétique du moustique ne modifie pas le profil des petits ARNs issus du découpage du dsRNA.

Pour répondre à la deuxième question (est-ce qu'un ou plusieurs SNP(s) modifie le profil des petits ARNs issu du dsRNA injecté ?), nous avons aligné les petits ARNs de 21b, qui correspondent aux siRNAs, sur les séquences des dsRNAs injectés correspondants. Nous avons donc 4 graphes qui représentent :

(1) les dsRNAs dsTep1*R et *S injectés dans des moustiques résistants (Fig. 31),
(2) les dsRNAs dsTep1*R et *S injectés dans des moustiques sensibles (Fig. 32),
(3) les dsRNAs dsTep3*R et *S injectés dans des moustiques sensibles (Fig. 33) et
(4) les dsRNAs dsTep12*R et *S injectés dans des moustiques sensibles (Fig. 34).

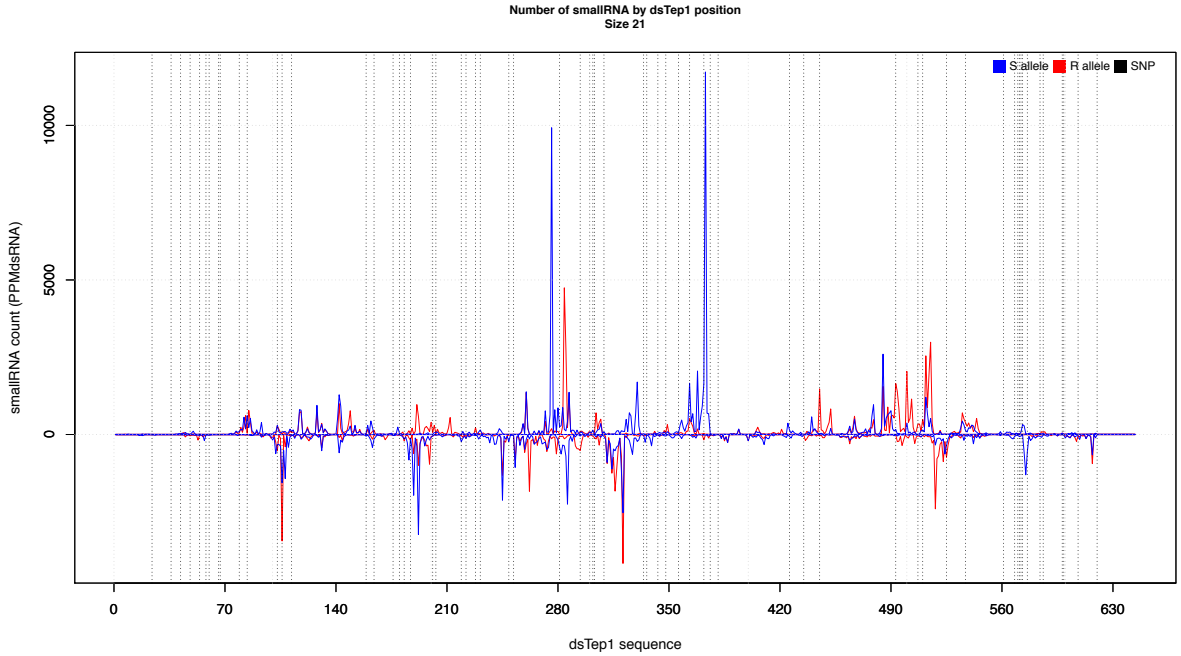


Figure 31. Répartition des petits ARNs de 21b issus des découpages de dsTep1*S (en bleu) et dsTep1*R (en rouge) injectés dans les moustiques résistants. Les traits noirs verticaux représentent la position des SNPs.

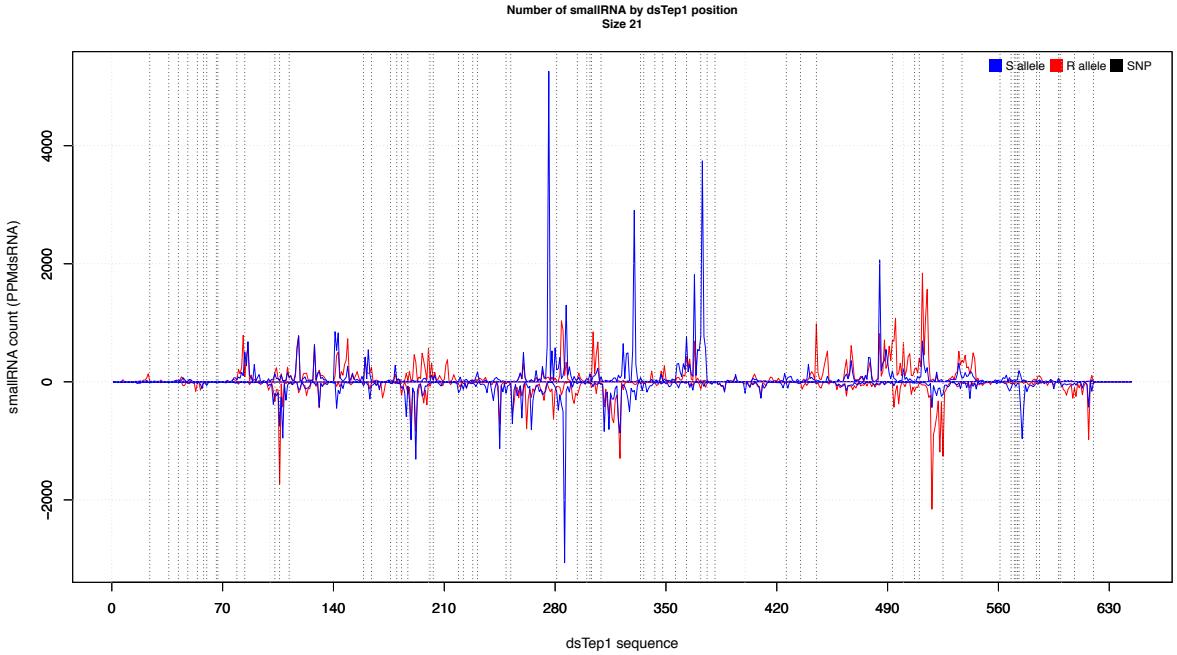


Figure 32. Répartition des petits ARNs de 21b issus des découpages de dsTep1*S (en bleu) et dsTep1*R (en rouge) injectés dans les moustiques sensibles. Les traits noirs verticaux représentent la position des SNPs.

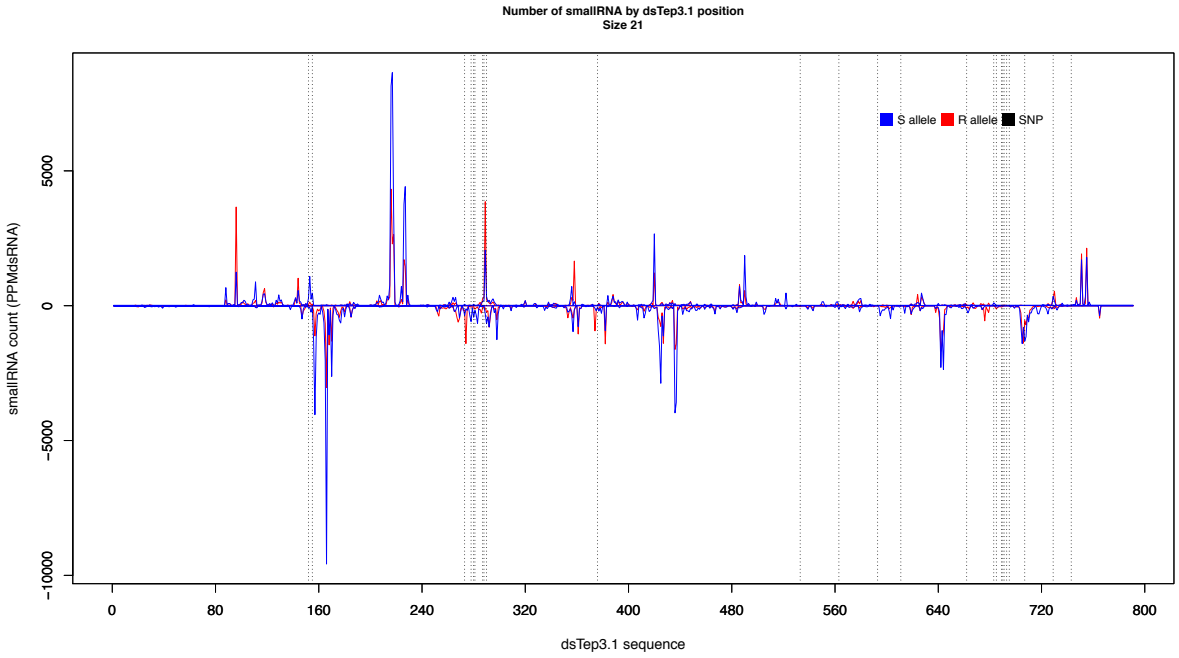


Figure 33. Répartition des petits ARNs de 21b issus de dsTep3-G3 et de dsTep3-R1 alignés sur la séquence de dsTep3.1. L'allèle G3, sensible, est en bleu et l'allèle R1, résistant, est en rouge. Les traits noirs verticaux représentent la position des SNPs.

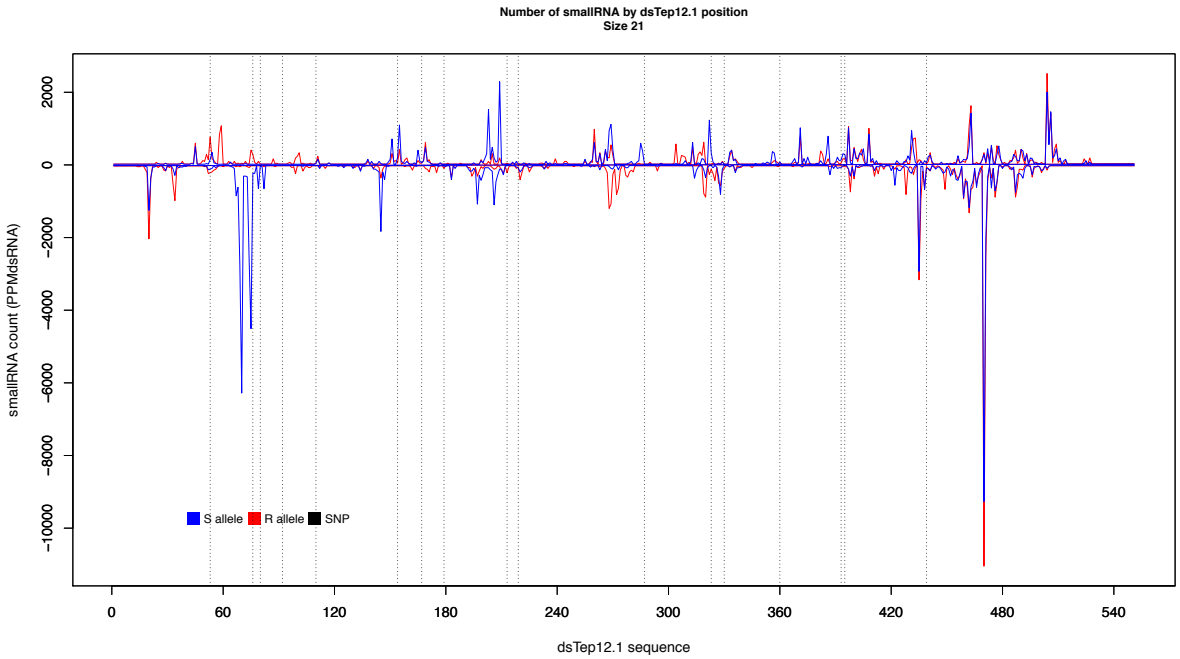


Figure 34. Répartition des petits ARNs de 21b issus de dsTep12-G3 et de dsTep12-R1 alignés sur la séquence de dsTep12.1. L'allèle G3, sensible, est en bleu et l'allèle R1, résistant, est en rouge. Les traits noirs verticaux représentent la position des SNPs.

Dans les quatre graphes précédents et pour les dsRNAs dsTep3.2, dsTep12.2 et dsTep12.3, le profil des petits ARNs issus du découpage du dsRNA diffère en fonction de la séquence injectée. Il y a des pics et des « déserts » communs mais aussi des zones propres à chacun des dsRNAs injectés, qu'il cible l'allèle résistant ou l'allèle sensible.

Pour évaluer la corrélation entre la présence d'un SNP et la modification du profil des petits ARNs produits, nous avons mesuré, pour chaque position, le ratio du nombre normalisé de séquences de 21b issues du dsRNA ciblant l'allèle *S sur le nombre normalisé de séquences de 21b issues du dsRNA ciblant l'allèle *R. Nous avons ensuite transformé les données par un log10. Les calculs ont été effectués sur chacun des brins sens et antisens.

Nous avons sélectionné les positions des séquences surreprésentées (les pics) des dsRNAs *S et *R et avons regardé si leur nombre variaient significativement entre le dsRNA*R et le dsRNA*S. Une séquence a été considérée comme surreprésentée si le nombre de séquences qui la compose est supérieur à la moyenne plus une fois l'écart-type. Pour déterminer s'il y a bien une différence significative dans le nombre de séquences issues des dsRNAs *S et *R, nous avons établi une valeur seuil de la moyenne plus deux fois l'écart-type. Puis nous avons regardé si ces positions comportaient au minimum un SNP.

Par exemple, pour l'échantillon dsTep3.1, il y a 72 positions qui correspondent à un pic dont 24 qui comportent au minimum un SNP. Parmi les 72 pics, 16 ont subi une variation significative du nombre de séquences dans le dsRNA opposé et parmi les 24 pics, 9 ont aussi subi une variation significative (Tableau 4).

	dsTep1		**dsTep3.1**		**dsTep3.2**		**dsTep12.1**		**dsTep12.2**		**dsTep12.3**	
	Tous	>= 1 SNP	Tous	>= 1 SNP	Tous	>= 1 SNP	Tous	>= 1 SNP	Tous	>= 1 SNP	Tous	>= 1 SNP
Nombre de très grand pics issus des dsRNAs *R et *S	146	129	72	24	79	23	79	43	91	17	91	35
Ayant une variation significative de séquences	40	40	16	9	20	15	26	25	13	10	21	21
Test du χ^2 pour l'indépendance	(**)		(*)		(****)		(****)		(****)		(****)	

Tableau 4. Corrélation entre la présence d'au minimum un SNP et la variation dans le nombre de 21-mers pour les séquences surreprésentées dans les gènes TEP1, TEP3 et TEP12. dsTep1 regroupe les deux expériences des figures 28 et 29, i.e. les dsRNAs*R et *S injectés dans la lignée résistante et la lignée sensible, respectivement.

Pour dsTep1, il existe une corrélation entre la présence d'au minimum un SNP et la variation du nombre de 21-mers entre les dsRNAs *R et *S, mais elle est plutôt intermédiaire. En fait, cela est dû au fort polymorphisme du gène TEP1, comme on peut le voir sur les figures 28 et 29 où les nombreux SNPs sont présents tout au long de la séquence.
Pour dsTep3.2, dsTep12.1, dsTep12.2 et dsTep12.3, la corrélation est très forte entre la présence d'au minimum un SNP et la variation du nombre de 21-mers entre les dsRNAs *R et *S. Parmi les séquences surreprésentées, lorsqu'il y a une variation significative du nombre de 21-mers entre les dsRNAs *S et *R, c'est à cause de la présence d'au moins un SNP dans la séquence. dsTep3.1 montre une assez faible corrélation que nous suspections déjà lors de l'analyse de la répartition des petits ARNs (Fig. 33).

La présence d'au minimum un SNP est donc responsable de l'abondance, faible ou forte, d'une séquence issue du découpage d'un dsRNA injecté.

Nous avons voulu savoir si il y avait une position préférentielle du SNP dans la séquence de 21b qui engendrait ces variations. Il est apparu que les SNPs étaient présents de manière homogène dans les séquences ayant des différences significatives d'amplitude (Fig. 35).

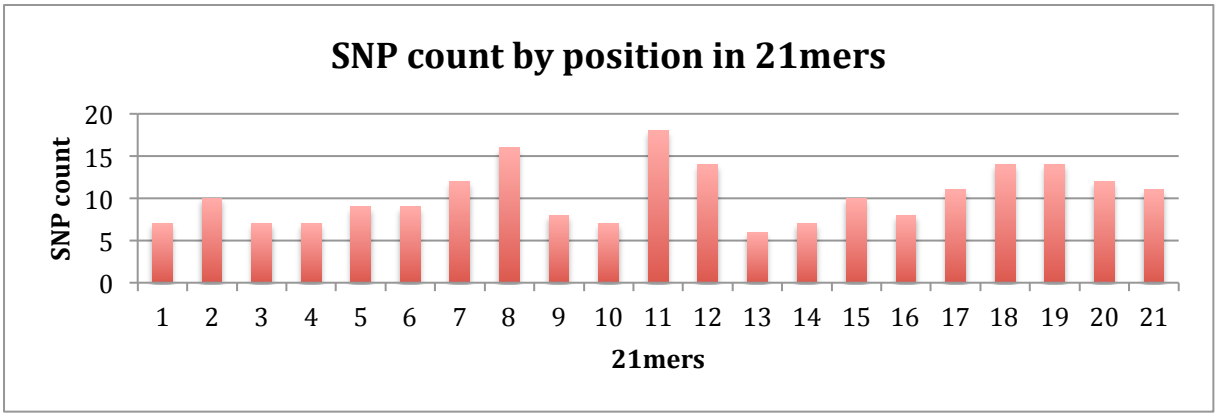


Figure 35. Compilation de toutes les positions des SNPs dans toutes les séquences de 21b issues des dsRNAs injectés et ayant montré une variation significative entre les dsRNAs *R et *S. Tous les échantillons ont été comptabilisés.

En regardant la nature de ces SNPs, il est apparu qu'il y avait une forte proportion de transitions (Fig. 36).

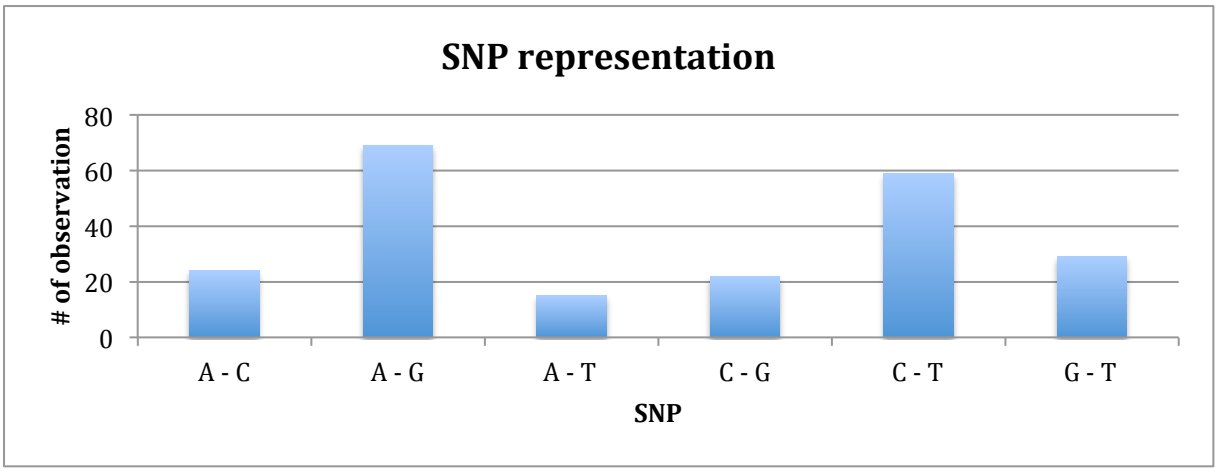


Figure 36. Type et fréquence des substitutions observées des SNPs engendrant une variation significative du nombre de séquences de 21b. Tous les échantillons ont été comptabilisés.

Pour le futur, il serait intéressant de regarder si un SNP précis favorise l'un ou l'autre des allèles.

Nous nous sommes penchés ensuite sur les séquences surreprésentées du profil, ie les pics. Notre hypothèse est que leur « chargement » dans le complexe RISC les protège de la dégradation et que probablement, ces petits ARNs dont les siRNAs sont les plus efficaces pour inhiber l'expression du gène cible.

2.5. Caractéristiques des petits ARNs surreprésentés

Le premier point que l'on a vérifié était le partage ou non d'un motif commun aux séquences surreprésentées. Pour ce faire, les séquences qui possédaient un nombre de reads alignés supérieur à la moyenne plus deux fois l'écart-type ont été définies comme surreprésentées par rapport aux autres (Fig. 37). Cette valeur seuil a été choisie après la comparaison de plusieurs valeurs seuil (en couleur sur la fig. 37).

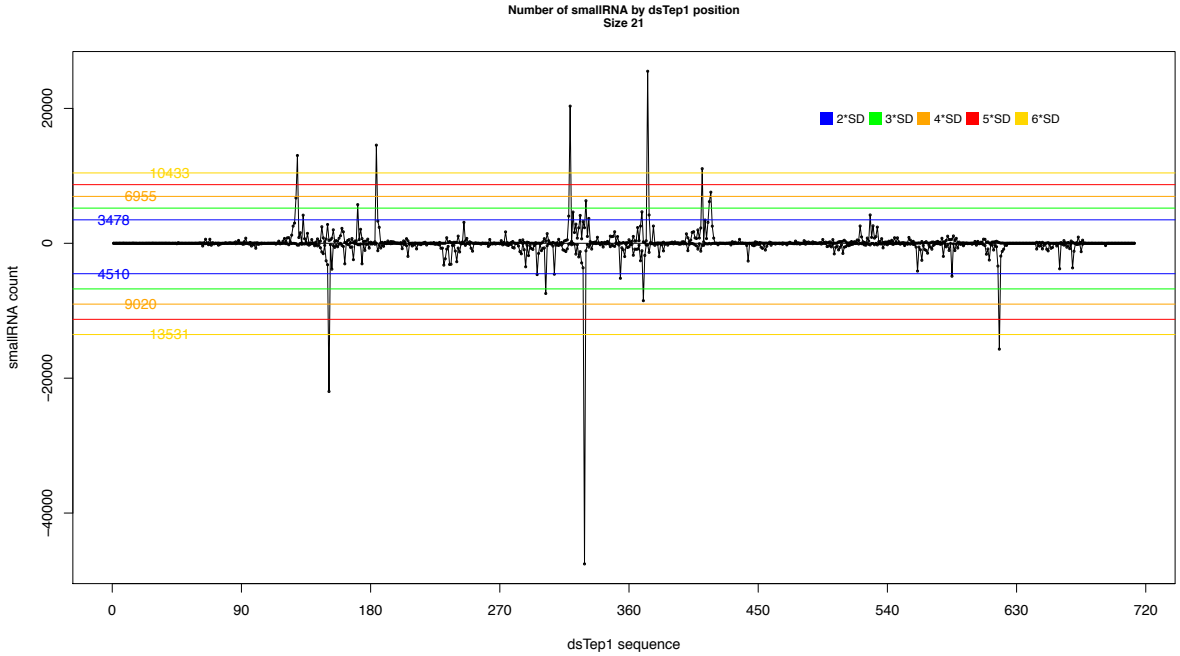


Figure 37. Répartition des petits ARNs de 21b sur dsTep1 (échantillon dsTep1_1) et sélection des séquences surreprésentées. Chaque trait de couleur correspond à la valeur seuil déterminée par la moyenne plus deux fois l'écart-type (bleu), plus trois fois l'écart-type (vert), plus quatre fois l'écart-type (orange), plus cinq fois l'écart-type (rouge) et plus 6 fois l'écart-type (jaune). Les seuils des brins sens et antisens diffèrent.

Une fois les seuils de chaque brin calculés, on a compilé les séquences définies comme surreprésentées du brin sens et du brin antisens dans un même fichier. WebLogo se charge d'aligner ces séquences entre elles et de déterminer les motifs communs. Chaque taille de 19 à 24b a été analysée indépendamment. Nous avons donc obtenu 6 WebLogo différents pour les reads de 19 à 24b issus des dsRNAs dsLacZ, dsTep1, dsTep3 et dsTep12 (Fig. 38). Dans le cas de dsLacZ et dsTep1, les réplicats ont été regroupés. dsTep3 et dsTep12 ne comportaient pas de réplicats.

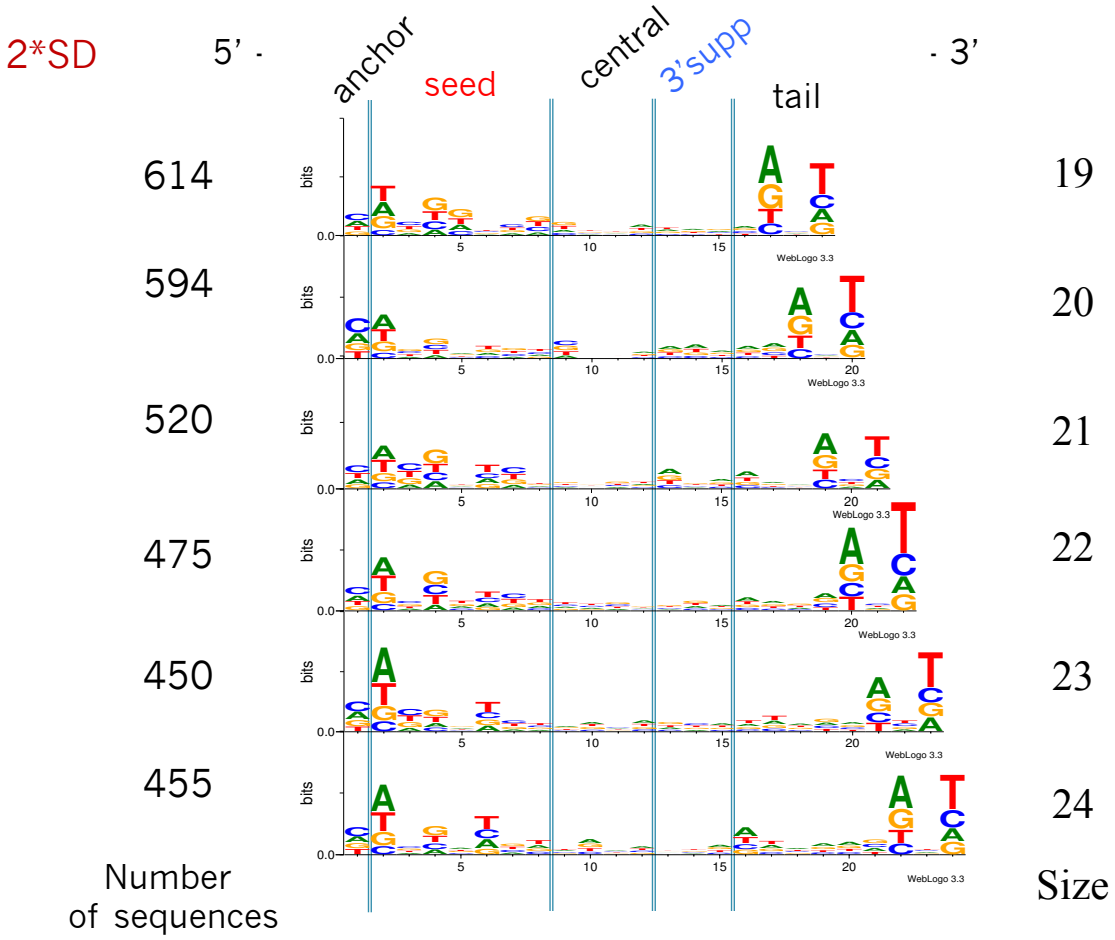


Figure 38. WebLogo des séquences surreprésentées de 19 (en haut) à 24b (en bas) pour dsLacZ_1 & _2, dsTep1_1 & _2, dsTep3mix et dsTep12mix. Pour chaque position, le logo informe sur la conservation de la séquence par la hauteur totale de la pile et sur la fréquence relative de chaque base par la hauteur de la base. A gauche est affiché le nombre de séquences qui a servi à construire le WebLogo pour chaque taille. En haut, on retrouve les 5 parties d'un siRNA. Dans cette figure, les séquences sont montrées telles que séquencées. En fait, comme ce sont des petits ARNs, la thymine devrait être remplacée par l'uracile (U).

Quel que soit la taille du read, la séquence a tendance à commencer par la cytosine (C) plutôt qu'une autre. La seconde base qui est un peu plus conservée favorise l'adénine (A). Le reste de la graine (« seed »), de la partie centrale (« central ») et de la partie 3' supplémentaire (« 3' supp ») sont très faiblement conservée. Par contre, les bases des positions -1 et -3 à l'extrémité 3' sont beaucoup plus conservées et, pour ces positions, l'adénine et la thymine (l'uracile (U) dans les petits ARNs) sont plus fréquentes, respectivement.

Dans la séquence d'un siRNA, le positionnement et l'enchainement des bases détermine si la séquence servira de guide pour inhiber l'ARN messager (ARNm) cible. Nous avons donc regardé les règles qui régissent la sélection d'un guide pour les comparer à nos observations précédentes.

En premier lieu, un siRNA est composé de deux brins contenant 2 bases non duplexées aux régions 3' (Fig. 39). Quand le duplex est incorporé par le complexe Ago2, un des brins devient le brin guide et s'ancre et alors que l'autre brin, le passager, va être clivé et expulsé du complexe. Ce sont les stabilités thermodynamiques relatives des régions 5' de chaque brin qui vont déterminer le brin guide du brin passager (Khvorova et al., 2003; Schwarz et al., 2003). Une faible stabilité en 5' promeut la sélection de ce brin en tant que guide par le complexe RISC (Fig. 39). Un faible pourcentage en GC du duplex facilite la sélection par le complexe et le largage du passager.

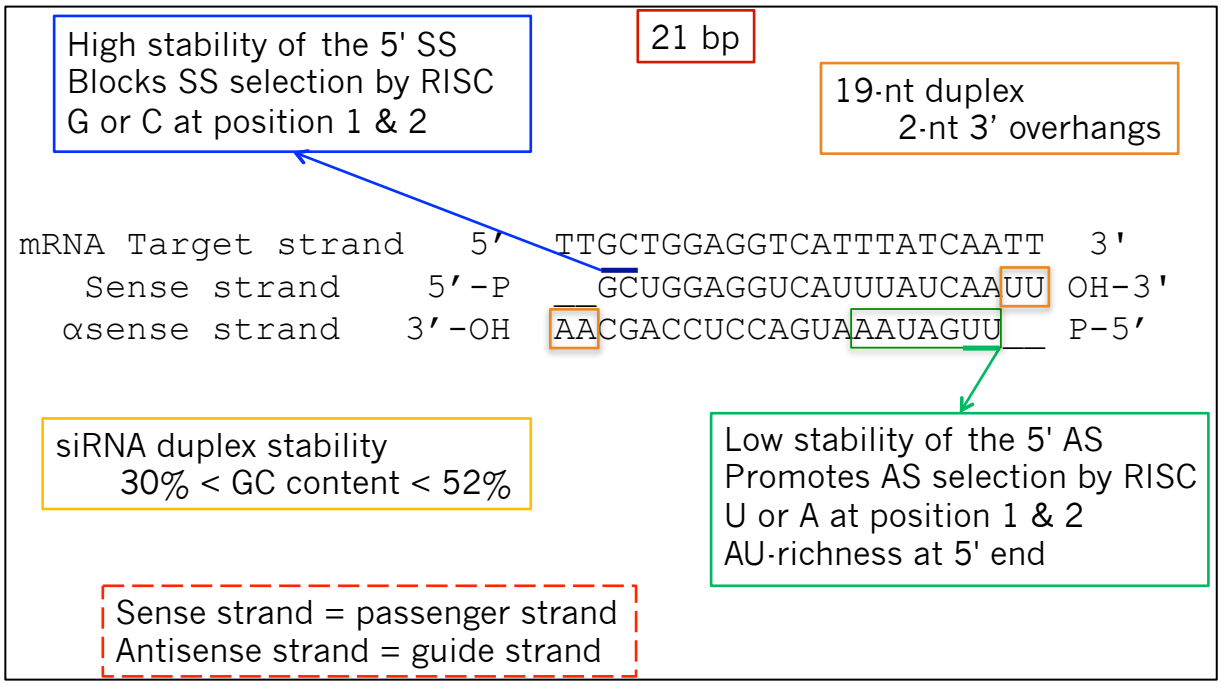


Figure 39. Définition d'un siRNA et critères de sélection des brins guide et passager lors du chargement du siRNA duplex dans Ago2. Un siRNA est une séquence de 21pb dont les 2 bases aux extrémités 3' ne sont pas appariées (encadrés rouge et orange). Un brin destiné à devenir guide doit avoir (1) des liaisons faibles telles que les liaisons A-U à l'extrémité 5' (encadré vert), (2) un pourcentage de GC compris entre 30 et 52% (encadré jaune), (3) des liaisons fortes telles que les liaisons G-C à l'extrémité 3' (encadré bleu). Seul le brin anti-sens sélectionné en tant que guide peut cibler l'ARN messager complémentaire et inhiber l'expression du gène (encadré pointillés rouge).

Une fois la séquence guide sélectionnée, plusieurs critères vont déterminer son efficacité à inhiber le gène cible (Reynolds et al., 2004; Shah et al., 2007). Ces critères ont été définis chez les mammifères et la drosophile (Ameres et al., 2011; Wee et al., 2012).
Le guide se découpe alors en 5 parties : l'ancre, la graine, la partie centrale, la partie 3' supplémentaire et la queue (Fig. 40). L'ancre et la queue se fixent dans deux domaines dans les domaines MID et PAZ (Fig. 40). La partie graine, si elle est complémentaire du brin d'ARN messager à cibler, permettra ensuite la fixation du complexe RISC sur ce brin d'ARNm ciblé. Seul le brin anti-sens du siRNA peut se lier à l'ARNm et promouvoir son clivage.

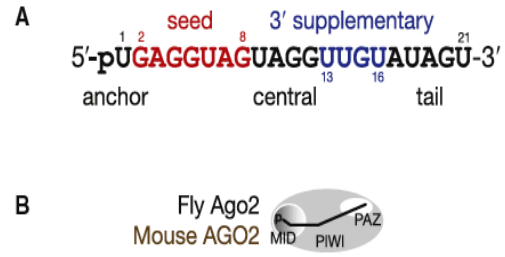


Figure 40. Caractéristiques d'un siRNA dans le complexe Ago2. (A) Les cinq parties d'un siRNA lors de son chargement dans le complexe Ago2. (B) Position du siRNA dans le complexe Ago2 chez la drosophile et la souris. L'ancre se fixe dans le domaine MID et la queue se retrouve dans le domaine PAZ. Source : (Wee et al., 2012)

Par rapport à nos observations, un C pour la première base de l'extrémité 5', les séquences guides chez les mammifères commencent plutôt par un U ou un A (Fig. 40). Cependant, il a été montré que chez les drosophiles, les siRNAs endogènes étaient départagés entre AGO2 et AGO1 par leur taille et la nature de la première base de l'extrémité 5' (Fig. 41, (Ameres et al., 2011)). Les siRNAs endogènes qui commencent par un C ou un U s'associent à AGO2, ce qui confirmerait notre observation de la surreprésentation du C chez le moustique.

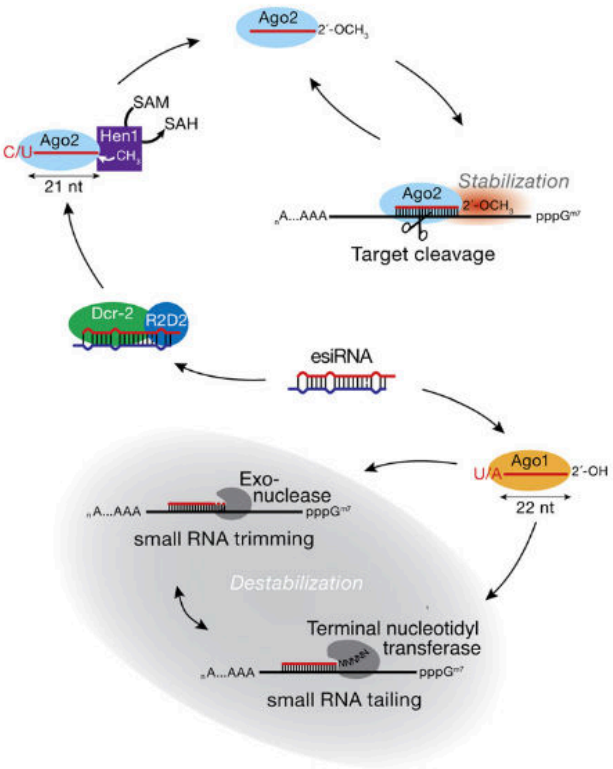


Figure 41. Modèle de partage des siRNAs endogènes entre AGO1 et AGO2 chez *Drosophila melanogaster*. Les siRNAs commençant par un A ou un U s'associent préférentiellement à AGO1 alors que ceux commençant par un C ou un U s'associent avec AGO2. Source : (Ameres et al., 2011).

La deuxième base de notre séquence WebLogo qui montre une préférence pour un A ou un U peut compenser le C de la première base car une extrémité 3' moins stable grâce aux liaisons AT promeut la sélection du brin antisens par le complexe RISC.

Dans le cas des dernières bases surreprésentées, la fin du duplex de 19b d'une séquence guide a plutôt tendance a avoir de fortes liaisons GC ce qui s'oppose à notre observation d'un A en position -3. Il n'y a pas d'observations d'un rôle pour les deux dernières bases non appariés de l'extrémité 3' d'une séquence guide.

Nous avons ensuite regardé si les trois séquences des trois pics les plus grands chez dsTep1 montraient les caractéristiques observées. C'est la séquence du guide qui nous intéresse et le guide est forcément sur le brin antisens pour pouvoir se lier au brin sens de l'ARNm (Fig. 39). Les positions de ces trois pics sont 618, 151 et 329 par ordre croissant du nombre de reads (Fig. 42).

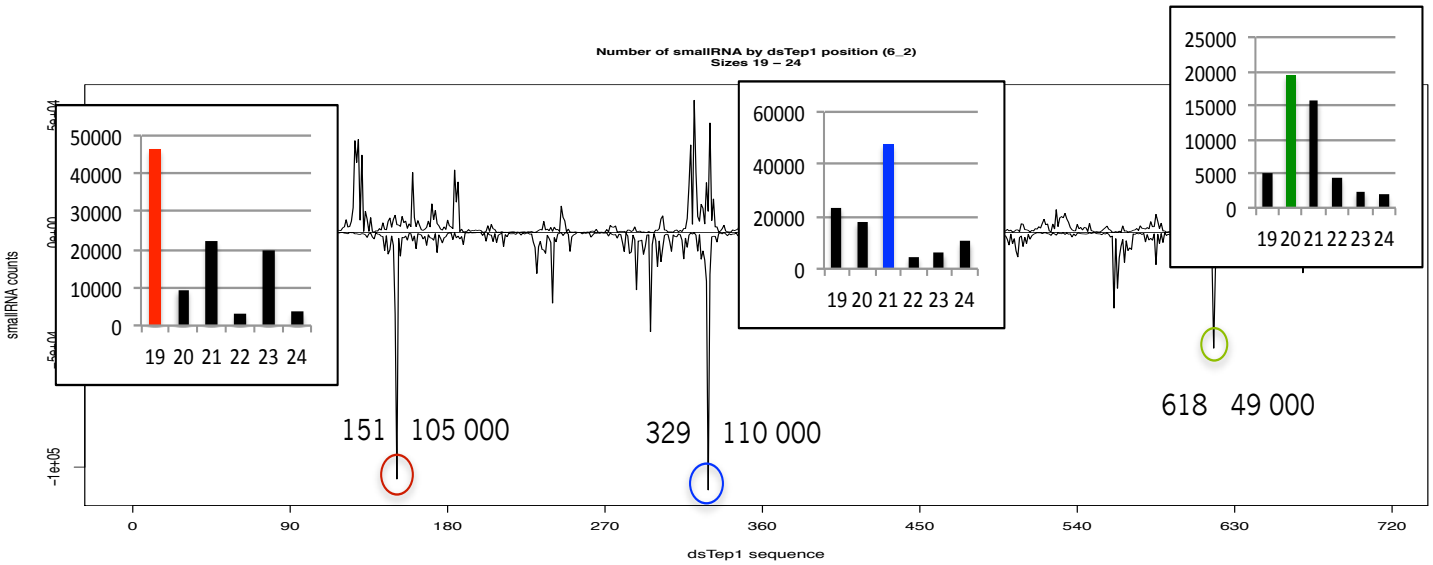


Figure 42. Distribution des petits ARNs alignés de 19 à 24b sur dsTep1 (échantillon dsTep1_1) et sélection des trois plus grands pics sur le brin anti-sens. Chaque pic est entouré de sa position à gauche et du nombre brut de séquences à droite. Les histogrammes montrent la distribution des tailles qui forment le pic.

Pour le pic 151, il y a un plus grand nombre de reads de 19b, pour le pic 329, ce sont les 21b et pour le pic 618, ce sont les 20b. Les 21b arrivant en seconde position pour les pics 151 et 618. Pour les trois séquences principales de 19b (pic 151), de 20b (pic 618) et de 21b (pic 329) des trois pics, le ratio entre la séquence guide et la séquence passager est de, respectivement, 30 pour 1, 365 pour 1 et 47 pour 1. Cette observation était attendue puisque les brins passagers sont détruits une fois le siRNA duplexé incorporé dans le complexe RISC.

Nous avons récupéré la séquence de 21b correspondant à ces trois positions (Fig. 43).

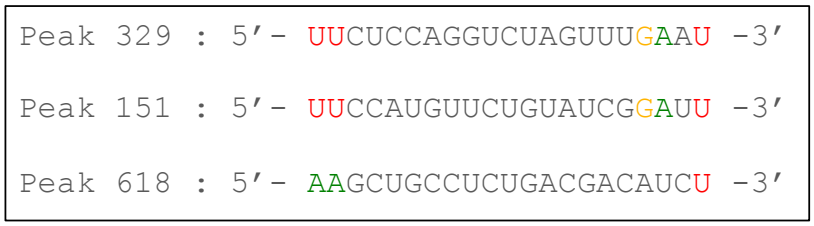


Figure 43. Séquences de 21b des trois pics par ordre décroissant du nombre de reads de 21b. Les bases colorées selon le type de base indiquent les positions importantes pour la désignation d'un guide efficace.

Les trois séquences présentent toutes le T à l'extrémité 3'. Les deux premières présentent aussi le A en positon -3. Seule la dernière comporte un A en seconde position et aucune ne commence par un C.
Par contre, on retrouve le U ou le A en début de séquence qui promeut la sélection de ce brin par AGO2 chez les mammifères et un G en position -4 qui augmente la stabilité de l'extrémité 3'.

Mais les différentes règles qui déterminent l'efficacité d'un siRNA ne sont peut-être pas toutes applicables chez le moustique.

Nous avons donc cherché à comparer une liste de siRNA prédits comme efficaces d'après les règles d'efficacité établies sur les mammifères et la drosophile à nos petits ARNs surreprésentés.

2.6. Décalage entre les petits ARNs surreprésentés et les siRNAs prédits comme efficace

Plusieurs outils de prédiction de siRNAs sont disponibles et se basent sur des règles établies pour calculer un score d'efficacité et prédire les meilleurs à partir de la séquence d'un gène à inhiber (Horn and Boutros, 2010, 2013; Shah et al., 2007; Ui-Tei et al., 2004).

Il est aussi possible d'évaluer des siRNAs candidats, ce que nous avons fait pour nos trois pics. Ces derniers ont montré un faible pourcentage d'efficacité de l'ordre de 66,1% pour le pic 329, de 51,9% pour le pic 151 et de 51,2% pour le pic 618. Ces pourcentages ont été calculés d'après l'algorithme de score établi par (Shah et al., 2007).

Sur les 21 outils, seuls 9 ont été utilisés notamment parce qu'ils ne restreignaient pas la liste des génomes de référence.
Le nombre de siRNAs prédits par chaque outil varie fortement. Certains outils se limite à un nombre précis alors que d'autres rapportent tous les siRNAs trouvés. Nous avons eu entre 10 à 285 siRNAs prédits pour la séquence de dsTep1. En sélectionnant les 10 meilleurs de chaque outil (sauf pour les outils siDirect et siWizard où il a été gardé 16 et 13 séquences), il est apparu que seulement 12 siRNAs sur les 99 étaient communs à deux outils.
Ceci montre que les outils de prédiction, même s'ils utilisent pour certains d'entre eux les mêmes règles, ne s'accordent pas sur les séquences des siRNAs prédits comme les plus efficaces dans une séquence donnée.

Nous avons positionné les siRNAs prédits comme étant les plus efficaces de chaque outil sur la séquence de dsTep1 pour visualiser leur distribution par rapport à nos séquences surreprésentées chez dsTep1 (Fig. 44).

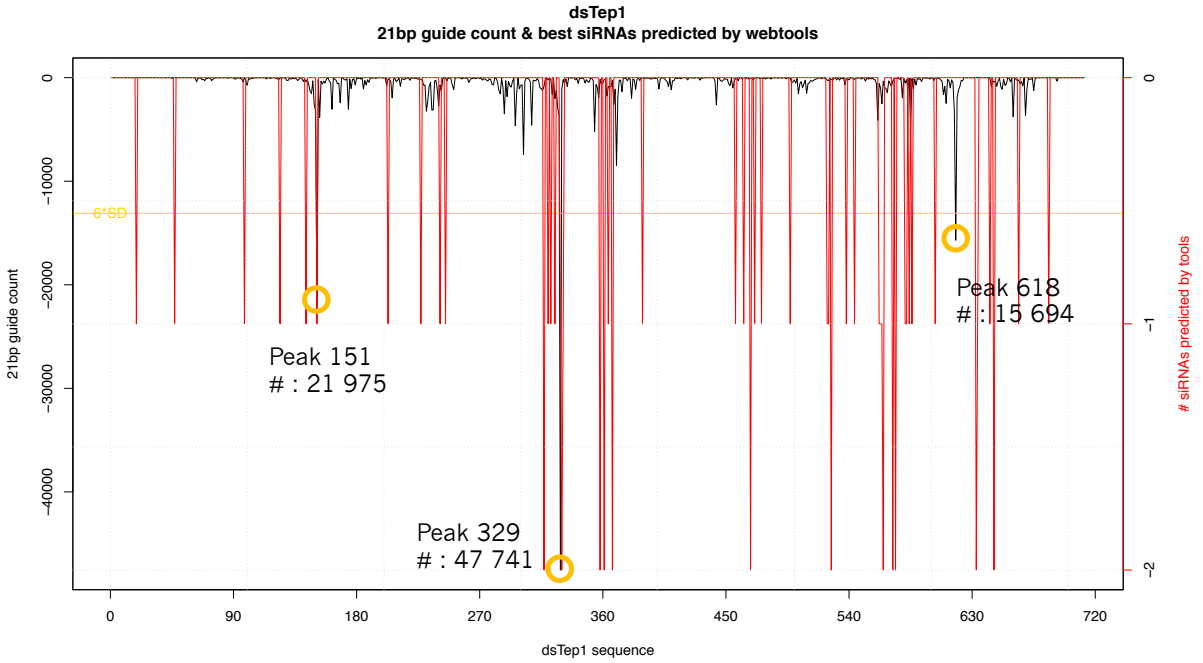


Figure 44. Répartition des petits ARNs de 21b en noir et des siRNAs prédits en rouge sur la séquence antisens du dsRNA dsTep1. Nos trois plus grands pics sont entourés et le nombre de reads de 21b associé est indiqué. L'échelle du nombre de petits ARNs est à gauche et l'échelle du nombre de siRNAs placés par position à droite.

Tout d'abord, on retrouve un maximum de 2 siRNAs prédits sur les 12 positions. La majorité des siRNAs prédits se situent dans des zones « déserts » sur la séquence de dsTep1.
Cependant, parmi les 10 meilleurs siRNAs, OptiRNA avait prédit la séquence du pic 151 et siRNA Design Service celle du pic 329. De plus, on retrouve 6 positions autour du pic 329 où il a été prédit des siRNAs : positions 317, 320, 322, 325, 330 et 331.

Nous avons également sélectionné dans les échantillons dsLacZ, dsTep3mix, et dsTep12mix les meilleurs pics des séquences de 21b qui correspond aux siRNAs (Tableau 5). Dans les cas de Tep3 et dsTep12 où il y avait plusieurs dsRNAs injectés, nous avons regroupé les dsRNAs sur la séquence entière du gène pour déterminer lesquels produisaient les plus grands pics par rapport aux autres. C'est dsTep3.1 et dsTep12.3 qui produisent les plus grands pics.

Gènes	Position des pics 21-mers	Nombre de séquences
LacZ	27	13 932
	161	13 766
TEP1	329	47 741
	151	21 975
	618	15 694
TEP3	623	1 603
	624	1 601
	417	1 347
	418	1 192
	147	1 081
TEP12	325	26 924

Tableau 5. Position et nombre de séquences des pics pour les gènes LacZ, TEP1, TEP3 et TEP12.

Nous avons ensuite regardé si des siRNAs avaient été prédits sur et +/- 5b autour de ces 11 pics (Fig. 45).

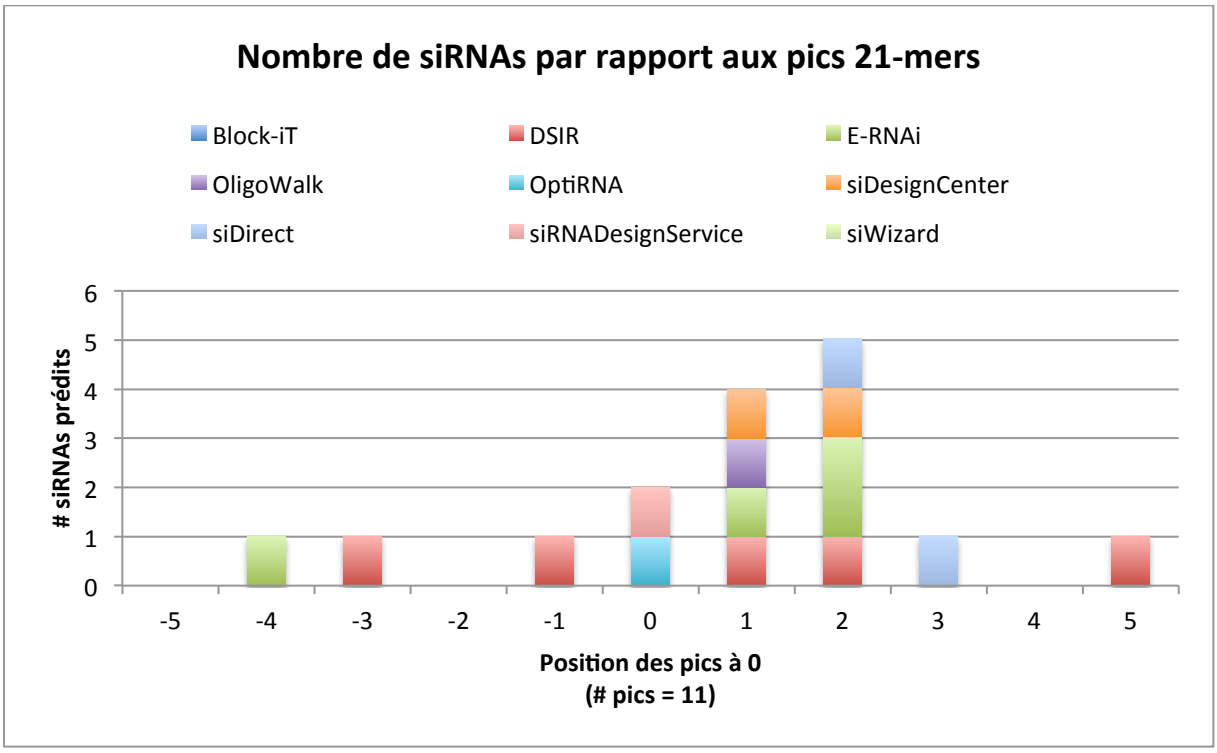


Figure 45. Nombre de siRNAs prédits par les 9 outils et se situant entre -5 et + 5 bases autour des 11 pics de 21b des gènes LacZ, TEP1, TEP3 et TEP12.

Sur les 401 siRNAs prédits pour les 7 séquences dsLacZ, dsTep1, dsTep3.1, dsTep3.2, dsTep12.1, dsTep12.2 et dsTep12.3, seuls 16 se trouvent dans les 5b autour d'un pic 21-mers et ils sont prédits par les 9 outils différents (Fig. 45). DSIR est le seul outil qui propose 5 siRNAs dans cette sélection.
On retrouve les 2 siRNAs des pics 329 et 151 de dsTep1. Les outils n'ont pas prédit de siRNAs correspondant à nos pics dans les autres dsRNAs. Par contre, 9 des siRNAs prédits sont placés à +1 et +2b de nos pics.

Nous avons donc décidé de vérifier la capacité des deux siRNAs les plus représentés, ie les pics 329 et 151, à inhiber le gène Tep1.

3. Élaboration de nouvelles sondes dsRNAs allèle-spécifique

3.1. Design d'un siRNA carrier

Cependant, nous savons qu'il est impossible d'inhiber un gène en injectant directement des siRNAs dans les cellules de moustiques. Il est nécessaire d'avoir une séquence d'au minimum 80b de long pour une inhibition efficace.
Nous avons alors pensé à associer nos siRNAs à une séquence « porteuse ». Pour déterminer la nature de cette séquence, nous avons choisi de sélectionner une séquence qui ne produisait que très peu de petits ARNs dans la séquence du dsRNA dsLacZ (Fig. 46). Ainsi même les quelques petits ARNs quand même produits n'interfèreront pas dans l'efficacité de la sonde.

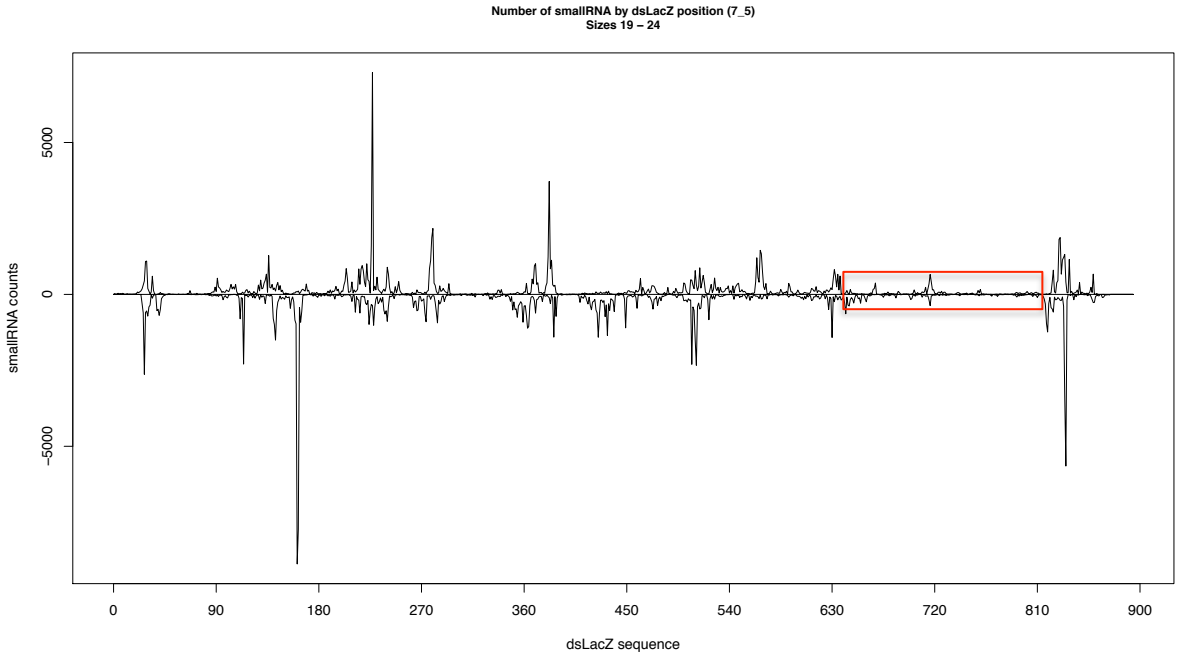


Figure 46. Distribution des petits ARNs de 19 à 24b sur la séquence dsLacZ (dsLacZ_1). La portion encadrée correspond à la séquence qui va servir de transporteur pour les siRNAs.

Nous avons ainsi créé une nouvelle sonde dsRNA qui contient à ses extrémités la séquence de 24b d'un siRNA sélectionné et entre les deux, la séquence « porteuse » de 176 pb issue de dsLacZ. Cette nouvelle sonde a été nommée siRNA Carrier (Fig. 47).

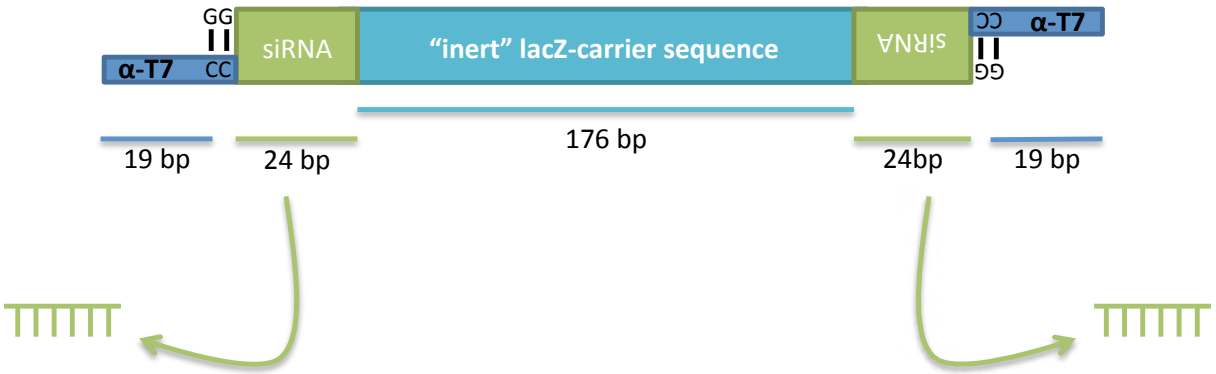


Figure 47. Détails d'un siRNA Carrier. Les séquences de 24pb d'un siRNA sont attachées aux extrémités d'une séquence de 176pb issue de LacZ. La séquence du promoteur T7 conduit à l'addition de di-mers G-C à l'extrémité des siRNAs. Source : N. Jelly

Pour tester nos siRNAs 329 et 151, il a donc été créé deux sondes, la première ds151-329 contenant le siRNA du pic 151 à gauche et le pic 329 à droite et la seconde ds329-151 contenant le pic 329 à gauche et le pic 151 à droite (Fig. 48A).
Cette expérience nous permettra de (1) vérifier la production des siRNAs testés et (2) vérifier l'efficacité du siRNA Carrier à inhiber efficacement le gène ciblé.

Après l'injection de ces deux sondes dans des moustiques, il a été réalisé un séquençage et un alignement des petits ARNs (Fig. 48BC) et une mesure de l'expression relative de Tep1 par PCR quantitative (Fig. 49 et 50).

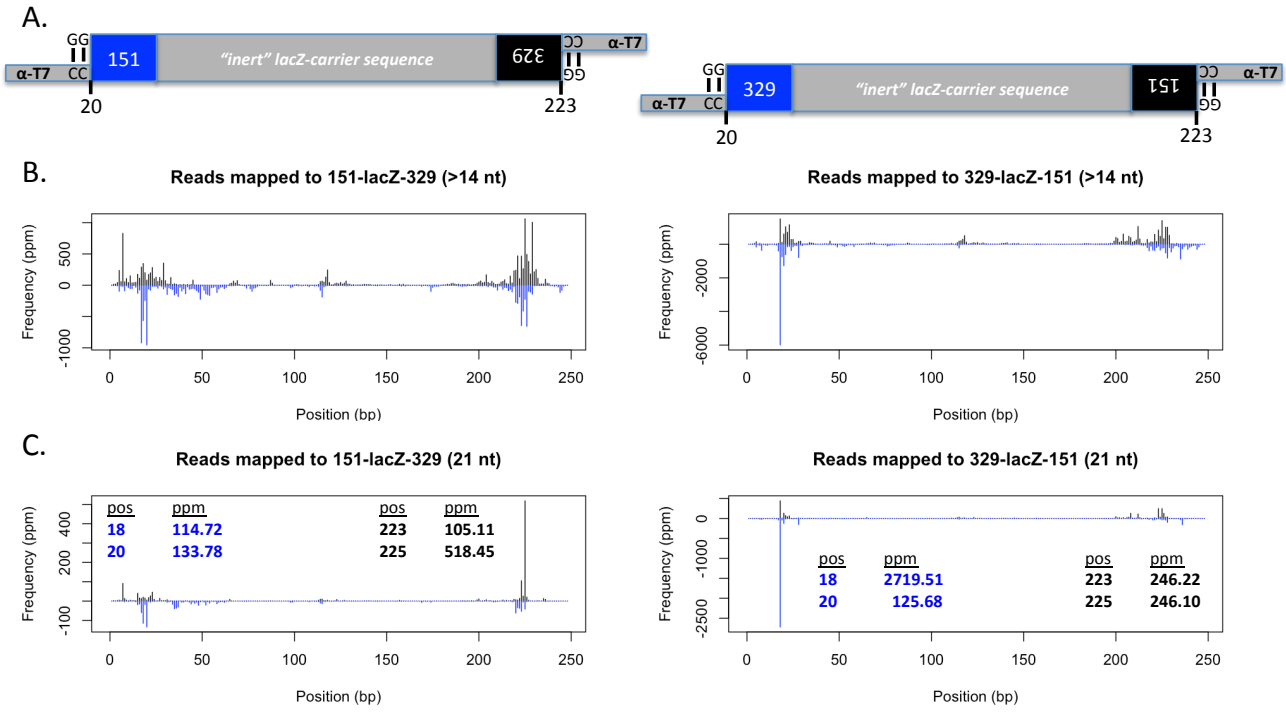


Figure 48. (A) Structure des sondes siRNA Carrier injectées pour tester les pics 329 et 151. (B) Distribution des reads de taille supérieure à 14b sur la sonde correspondante. (C) Distribution des reads de 21b sur la sonde correspondante. Source : J. R. Clayton

Il y a bien une très faible production de petits ARNs au niveau de la partie centrale qui correspond en fait à la zone « désert » de dsLacZ. Cela confirme le fait que notre porteur est bien neutre.
Ensuite, la majorité des reads proviennent des séquences de 24b correspondant aux pics 329 et 151. Dans les quatre cas, toutes les tailles de reads ou seulement les 21b, les pics majeurs sont aux positions 18 et 20 au début de la sonde et aux positions 223 et 225 à la fin de la sonde. En fait, la position 20 et la position 223 se rapportent à la première base de la séquence de 24b que nous voulions tester, ie soit le pic 151 soit le pic 329. Les positions 20-2 et 223+2 contiennent les CC de la fin du promoteur T7.

Cette expérience nous révèle que nos sondes siRNA Carrier produisent des siRNAs fortement représentés provenant seulement de notre séquence de 24b sélectionnée. Mais cette information n'est pas intéressante si elle n'est pas couplée aux résultats de la PCR quantitative (PCRq) qui mesure l'expression de Tep1 pour vérifier l'efficacité de la sonde.

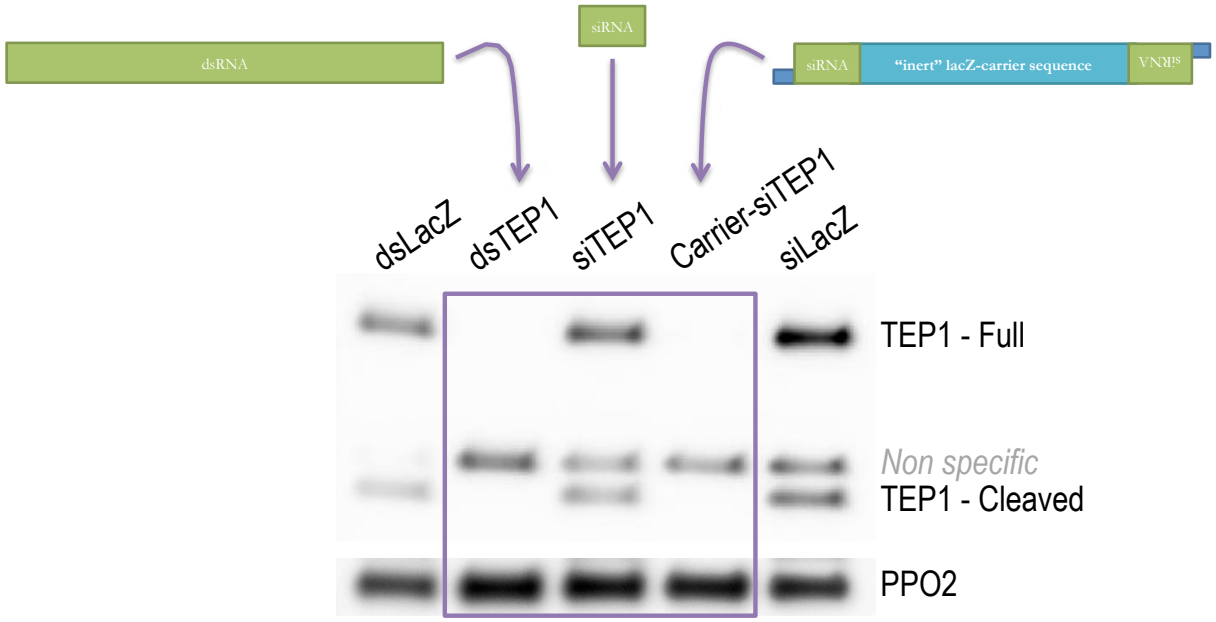


Figure 49. Western blot du gène Tep1 pour tester son inhibition par l'injection de différentes sondes. dsLacZ et siLacZ sont les contrôles négatifs. PPO2 est le gène domestique de contrôle. dsTep1 est un dsRNA long, siTep1 est un petit siRNA de 21b et Carrier-siTep1 est notre nouvelle sonde dsRNA qui transporte deux siRNAs spécifiques de Tep1. Source : N. Jelly.

L'injection de notre siRNA Carrier Carrier-siTEP1 est aussi efficace que l'injection du dsRNA dsTep1 pour inhiber l'expression du gène TEP1 (Fig. 49).

Nous avons ensuite testé les séquences de 24pb des trois plus grands pics de dsTep1, dsTep3 et dsTep12 indépendamment les uns des autres ainsi que des séquences de 24pb provenant de régions « déserts » dans ces mêmes dsRNAs (Tableau 6). Il a été fait 2 réplicats.

Gènes	Positions des pics	Positions des déserts
TEP1	329 - 151 - 618	631
TEP3	623 - 417 - 147	526 – 528
TEP12	396 - 438 - 254	129 - 239

Tableau 6. Positions des pics et des « déserts » sélectionnés pour intégrer un siRNA Carrier.

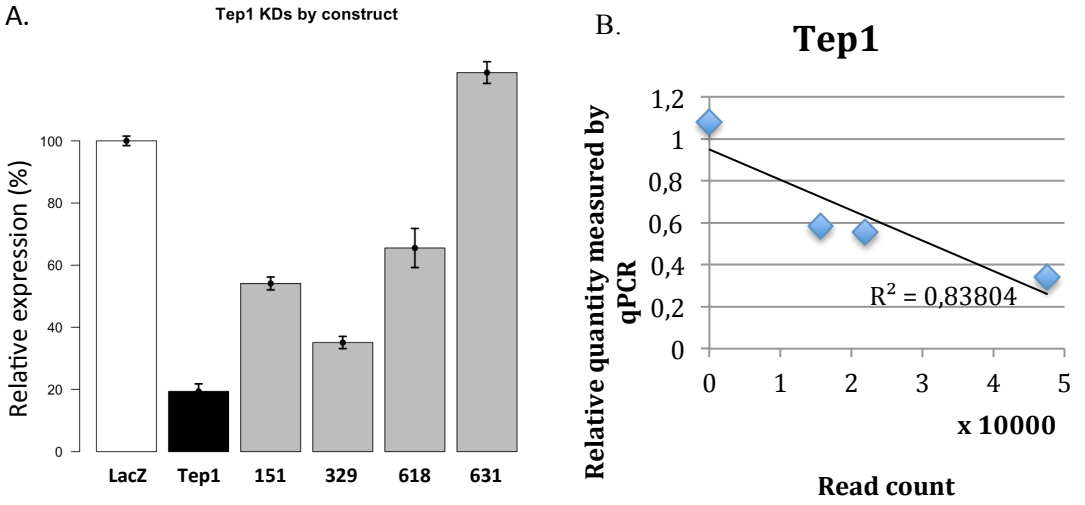


Figure 50. (A) Mesure de l'expression relative du gène TEP1 après injection d'un dsLacZ pour le contrôle négatif, d'un dsTep1 pour le contrôle positif et des sondes contenant deux séquences 151, deux 329, deux 618 et une zone « désert » dans dsTep1, 631 (Source : J. R. Clayton). (B) Corrélation entre l'expression relative du gène Tep1 et le nombre de siRNAs de 21b de chaque pic ou désert sélectionné.

L'expression relative du gène TEP1 est fortement corrélée au nombre de siRNAs produits par le dsRNA (Fig. 50). Plus le pic est grand donc plus il y a de siRNAs présents, plus l'expression du gène est inhibée.
Cependant, cette corrélation n'est pas conservée pour les deux autres gènes testés : TEP3 et TEP12 (Fig. 51).

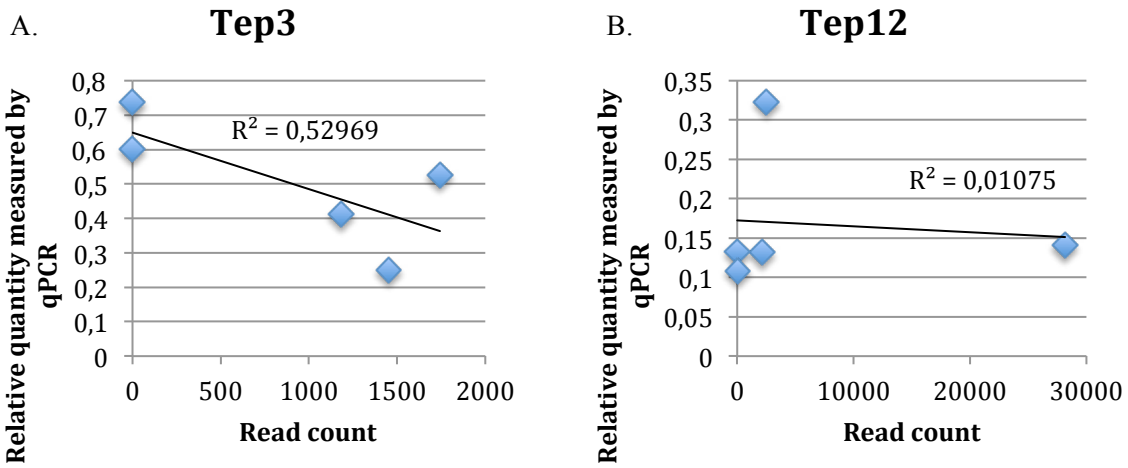


Figure 51. Corrélation entre l'expression relative du gène TEP3 (A) et du gène TEP12 (B) et le nombre de siRNAs de 21b de chaque pic ou désert sélectionné.

Pour ces deux gènes, la capacité à inhiber efficacement l'expression du gène n'est pas corrélée à la fréquence des siRNAs. Pour le gène TEP3, le plus grand pic, donc celui qui a le plus de séquences, ne donne pas la meilleure inhibition de l'expression du gène. Pour TEP12, les deux zones « désert » inhibent plus l'expression du gène que les pics.

Il y a plusieurs explications possibles au fait qu'il n'y a pas de corrélation pour ces deux gènes par rapport à TEP1 :

- l'injection a été réalisée dans la lignée G3M, or, à ce moment-là nous ne connaissions pas les séquences exactes et le nombre d'allèles des gènes TEP3 et TEP12 dans cette lignée,
- nous ne savons pas si la présence des CC à l'extrémité 3' des siRNAs influence le découpage ou le chargement des siRNAs dans le complexe.

Pour conclure, il a été démontré qu'en injectant des siRNAs avec une séquence porteuse, il était possible d'inhiber l'expression du gène. Cependant, il n'y a pas de forte corrélation entre l'abondance d'un siRNA et sa capacité à inhiber le gène correspondant.

3.2. Évaluation d'un siRNA carrier allèle-spécifique

Notre objectif est de créer des siRNA Carriers allèle-spécifique et jusque lors, nous n'avions pas intégré de siRNAs contenant des SNPs dans nos siRNA carriers. Nous avons donc testé pour les gènes TEP3 et TEP12 qui sont faiblement polymorphiques, l'injection de siRNA Carriers contenant des SNPs (Fig. 52).
Dans un premier temps, nous avons choisi deux siRNAs différents présentant 7 SNPs entre l'allèle sensible G3 et l'allèle résistant L3-5. Puis nous avons réduit le nombre de SNPs à 3 avec deux nouveaux siRNAs.
Les injections ont été réalisées dans la lignée G3M dont on sait qu'elle possède l'allèle sensible G3 de chaque gène. Il y a eu 3 réplicats.

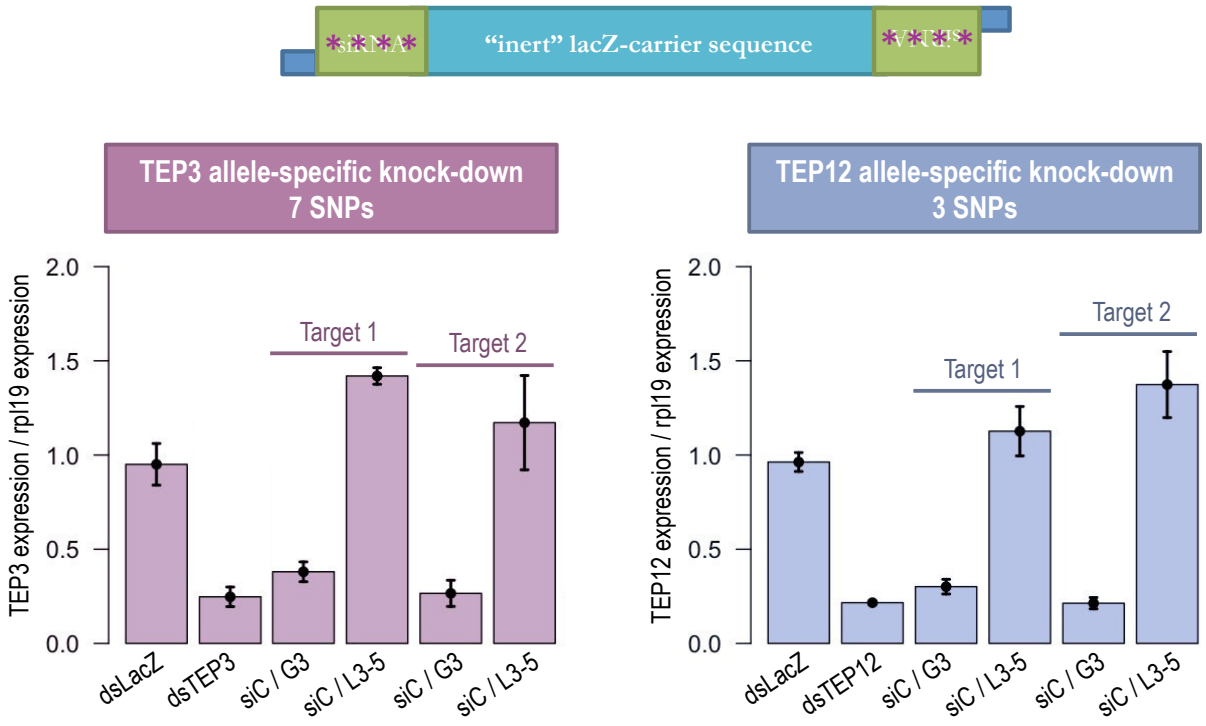


Figure 52. Expression relative des allèles des gènes TEP3 (à gauche) et TEP12 (à droite) après l'injection de siRNA Carriers allèle-spécifique dans la lignée G3M. dsLacZ est le contrôle négatif, dsTep3 et dsTep12 sont les contrôles positifs. Target 1 et 2 sont deux siRNAs différents ciblant l'allèle sensible G3 l'allèle résistant L3-5. Source : N. Jelly

Les siRNAs Carrier provenant de l'allèle G3 ont bien inhibé l'expression des gènes TEP3 et TEP12 alors que ceux provenant de l'allèle L3-5 n'ont pas inhibé les gènes puisque la séquence des siRNAs n'était pas complémentaire de l'allèle présent chez cette lignée : l'allèle G3 (Fig. 52). La présence des SNPs dans nos siRNA Carrier permet donc de cibler spécifiquement l'allèle choisi.
Nous avons aussi testé si le siRNA Carrier était toujours allèle-spécifique dans le cas où le siRNA ne contenait qu'un seul SNP (Fig. 53 et 54).

Target 1	**L3-5**	GCATTAAGGCCTCCAT**T**ATGACGGG
	G3M	GCATTAAGGCCTCCAT**C**ATGACGGG
Target 2	**L3-5**	TCGCATTAAGGCCTCCAT**T**ATGACG
	G3M	TCGCATTAAGGCCTCCAT**C**ATGACG

Figure 53. Position du SNP des deux siRNAs « Target » dans les allèles sensible G3 et résistant L3-5.

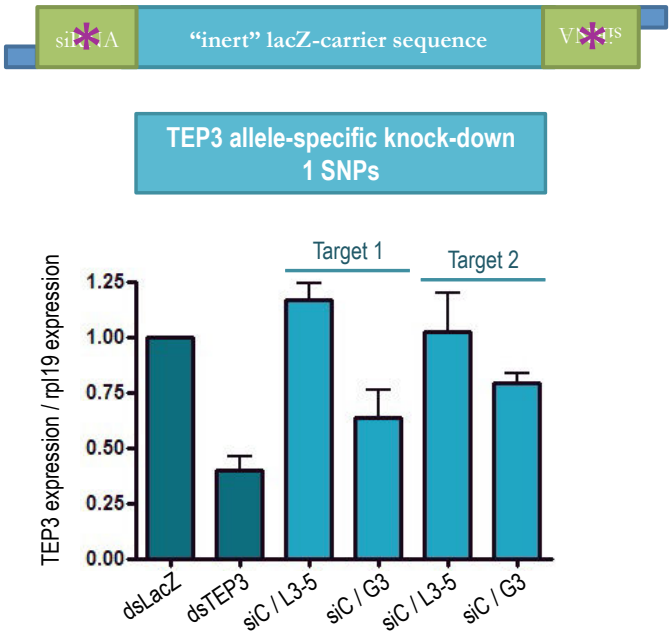


Figure 54. Expression relative des allèles du gène TEP3 après l'injection d'un siRNA Carrier allèle-spécifique dans la lignée G3M. dsLacZ est le contrôle négatif, dsTep3 est le contrôle positif. Target 1 et 2 sont deux siRNAs différents ciblant l'allèle sensible G3 l'allèle résistant L3-5. Source : N. Jelly

Le siRNA Carrier contenant qu'un seul SNP est toujours allèle-spécifique puisqu'il inhibe toujours le gène TEP3 (allèle G3) dans le cas où le siRNA qui ciblait l'allèle G3 a été injecté (Fig. 54).
Toutefois, la présence du SNP dans les siRNAs Carrier a eu un impact limité sur l'inhibition du gène TEP3. Il serait intéressant de sélectionner d'autres siRNAs en modifiant la position du SNP dans la séquence.

IV. Conclusions et Perspectives

Notre objectif était d'optimiser la méthode de rasRNAi qui permet d'inhiber seulement l'expression d'un des allèles d'un gène choisi pour tester sa contribution au phénotype de résistance observé. La partie à développer était la méthode à utiliser dans le cas de gènes à cibler faiblement polymorphiques et où il était impossible d'injecter de longs dsRNAs sans cibler les deux allèles. Ce qui est le cas pour les gènes TEP3 et TEP12 dont nous voulons tester leur contribution allélique au phénotype de résistance.

Pour arriver à créer de nouvelles sondes dsRNAs allèle-spécifiques, nous avons étudié le phénomène de découpage de plusieurs dsRNAs injectés longs de 500 à 900b ciblant différents gènes endogènes, TEP1, TEP3 et TEP12 et un gène exogène, LacZ, pour le contrôle. Après l'analyse des petits ARNs séquencés, nous avons pu faire plusieurs observations :

(1) les petits ARNs issus du dsRNA injecté ne sont pas répartis de manière homogène sur la séquence du dsRNA injecté : certaines séquences sont très abondantes par rapport à d'autres (repérables par la présence d'un pic) alors que d'autres régions du dsRNA ne sont pas présentes dans le pool de reads séquencés (zones que l'on a qualifié de « déserts »).

(2) le profil de distribution (répartition des pics et des « déserts ») des petits ARNs issus du dsRNA injecté est dans une certaine mesure indépendant de la séquence environnante et stable au cours du temps post-injection.

(3) les siRNAs de 19 à 24b montrent un taux de substitutions, surtout des transitions, à l'extrémité 3'. Ceci suggère que soit les siRNAs n'ont pas été protégés par la méthylation de leur extrémité 3', soit qu'ils ont été incorporés dans la voie des miRNAs où il n'y a pas de protection en 3'.

Ces premières injections ont été réalisées pour cibler le gène quelque soit ses allèles. Nous avons donc ensuite injecté des dsRNAs spécifiques de deux allèles d'un même gène dans deux lignées de moustiques, porteuses ou non de l'allèle ciblé. Les allèles sensible et résistant de chaque gène (TEP1, TEP3 et TEP12) ont été ciblés dans des moustiques sensibles et résistants. Nous avons pu conclure de cette expérience que :

(4) la présence de SNPs dans la séquence d'un dsRNA injecté affecte son découpage et produit un profil de petits ARNs (pics et « déserts ») différent.

(5) un même dsRNA est découpé de la même façon quelque soit l'environnement génétique du moustique.

Toutes ces observations nous ont permis de créer une nouvelle sonde dsRNA appelé « siRNA Carrier ». L'analyse des petits ARNs séquencés combinée à la mesure de l'efficacité du knockdown par Western blot et/ou par qPCR nous a permis de démontrer, dans certains cas, l'efficacité des ces siRNA Carriers dont l'homologie avec la cible est limitée puisqu'il y a seulement 48b (20% du siRNA Carrier) qui proviennent de l'allèle à cibler. Cependant, nous n'avons pas observé de lien direct entre l'abondance d'une séquence et son efficacité lorsqu'elle est placée dans le siRNA Carrier, même s'il semble que les séquences correspondant aux pics sont globalement plus efficaces que les déserts.
De plus, nous avons montré que ces siRNA Carriers sont allèle-spécifiques, avec toutefois une action limitée lorsqu'un seul SNP distingue les deux allèles. Il faudrait réaliser d'autres injections de siRNAs contenant qu'un seul SNP pour évaluer leur réelle efficacité d'inhibition du gène cible.

Chapitre III

Élaboration d'un pipeline d'analyse de données métagénomiques pour l'identification des micro-organismes chez les lignées résistantes et sensibles du moustique *Anopheles gambiae*

Sommaire

I. Contexte

La capacité d'un moustique à transmettre les parasites du paludisme est déterminée par différents facteurs. En plus des facteurs génétiques du moustique qui contrôlent la résistance, deux autres entrent en jeu : les facteurs de virulence des parasites (Lambrechts et al., 2005; Molina-Cruz et al., 2013) et les facteurs environnementaux tel que la composition du microbiote intestinal.
En effet, outre les parasites du paludisme dont il est le vecteur, les moustiques hébergent d'autres microorganismes avec lesquels ils établissent des relations symbiotiques allant du mutualisme au parasitisme. C'est le terme microbiote qui regroupe l'ensemble de ces communautés microbiennes (virus, bactéries, levures et protistes) colonisant soit les parties externes de l'hôte (l'épithélium, les lumens, etc.) soit les parties internes. Les bactéries intracellulaires sont connues pour jouer un rôle dans les fonctions vitales des insectes telles que la nutrition, la digestion, la reproduction, le développement ou la protection contre les pathogènes (Douglas, 2011; Minard et al., 2013). Les virus, les champignons et les bactéries commensales montrent aussi des aptitudes à modifier le comportement de leur insecte hôte (Lewis and Lizé, 2015). Ces modifications que l'on observe sont appelées phénotype étendu.
Les bactéries intracellulaires et commensales sont capables de réduire les infections virales et parasitiques en activant les réponses immunitaires de leur hôte ou en inhibant directement le développement du pathogène (Fig. 1).

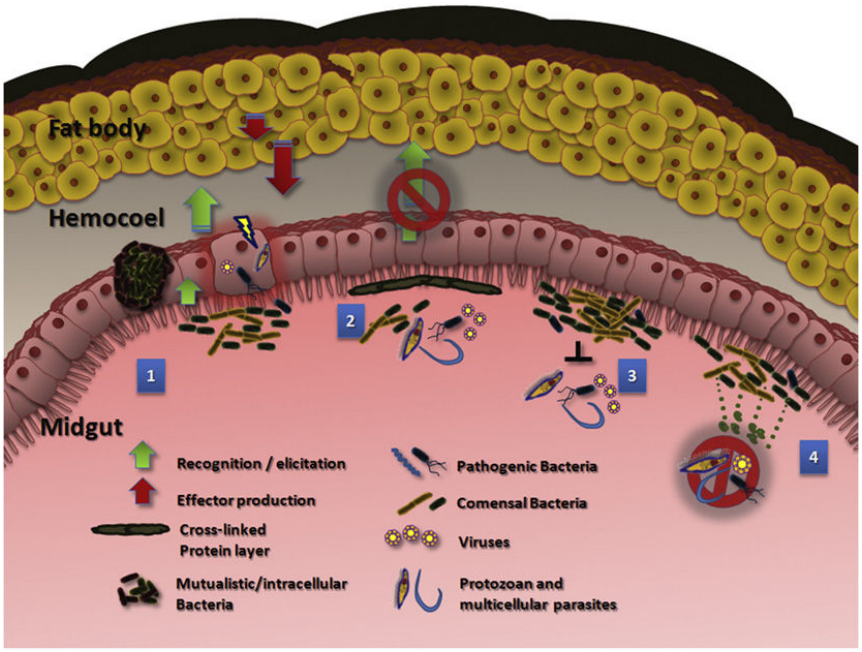


Figure 1. Les microorganismes du système digestif ont un impact sur la capacité des pathogènes à infecter leur vecteur. (1) Les bactéries commensales et symbiotiques associées au système digestif stimulent une réponse immunitaire antibactérienne qui a aussi un impact négatif sur le développement des pathogènes. (2) Une protéine sécrétée par les cellules épithéliales qui inhibe la sur-activation du système immunitaire a aussi un rôle de protection des pathogènes dans les intestins (Kumar et al., 2010). (3) Les populations bactériennes se répandent rapidement dans les intestins après l'ingestion de sang et peuvent former une sorte de barrière physique entre les pathogènes et la barrière épithéliale. (4) Les molécules sécrétées par les bactéries, telles que les intermédiaires d'oxygène réactif ou les métabolites secondaires, peuvent tuer ou interférer avec les parasites dans les intestins avant que l'infection se produise (Cirimotich et al., 2011a). Source : (Cirimotich et al., 2011b).

Chez les moustiques, c'est la flore intestinale microbienne qui est en contact direct avec les parasites fraichement ingérés. C'est aussi au niveau de l'intestin que se produit la plus grande destruction de parasites. Pour tester l'impact de cette flore microbienne sur le taux d'infection par *P. falciparum*, des chercheurs ont d'abord débarrassé les moustiques de leur flore par des traitements antibiotiques puis ils les ont infecté avec *P. falciparum*. Les moustiques aseptiques étaient significativement plus sensibles à l'infection que les moustiques avec leur flore naturelle (Dong et al., 2009). Il a aussi été démontré que plusieurs gènes du moustique sont surexprimés sous l'influence de cette flore intestinale dont plusieurs facteurs anti-Plasmodium (Dong et al., 2006). Plus récemment, le pyroséquençage de plusieurs moustiques issus de sites différents a permis de caractériser la population microbienne intestinale de chaque moustique et de révéler ainsi une grande variation dans la composition bactérienne liée à l'environnement (Boissière et al., 2012). De plus, l'abondance des entérobactériacées a clairement été liée au taux d'infection par le parasite *P. falciparum* (Boissière et al., 2012). Il y a donc bien une corrélation entre la présence de certaines souches de bactérie et la diminution du nombre de parasites vivants au sein du moustique (Cirimotich et al., 2011a).
Par ailleurs, la variation génétique des moustiques peut déterminer la composition en microorganismes des moustiques individuels, ce qui par la suite influence la réponse immunitaire anti-Plasmodium. Le polymorphisme de certains gènes de la réponse immunitaire, dont plus particulièrement des fibronectines de type III, a été lié à l'abondance de *S. marcescens* et d'autres entérobactériacées chez *A. gambiae* (Stathopoulos et al., 2014). Il faut aussi prendre en compte la variation génétique des bactéries, surtout depuis qu'il a été démontré que la variation intra-spécifique de S. marcescens détermine l'impact des bactéries sur la résistance aux parasites du paludisme P. berghei (Bando et al., 2013).

Au laboratoire, un groupe de moustiques sélectionnés génétiquement pour leur forte sensibilité aux parasites contenait encore, après une infection à *P. berghei*, des moustiques complètement résistants, sans aucun parasite vivant. Ceci suggère que d'autres facteurs, non génétiques, contribuent à contrôler le développement des parasites chez ces moustiques.

Connaissant le rôle des microorganismes sur le taux d'infection par les parasites du paludisme au sein d'un moustique, nous avons cherché à identifier la composition en microorganismes de toutes nos lignées de moustiques afin de comparer (1) nos moustiques résistants et sensibles et (2) nos moustiques au sein d'une même lignée. Cela nous permettra peut-être d'identifier des espèces uniques à certaines populations et ainsi d'établir un lien entre leur présence et le degré de résistance.

En microbiologie, la méthode classique de découverte ou de caractérisation d'une flore microbienne est la culture in vitro en boite de Pétri. Cependant, la principale limite de cette technique est l'impossibilité de recréer l'environnement complexe comme le corps d'un moustique. La plupart des bactéries qui nécessitent cette complexité pour croitre ne seront pas identifiées. Afin de déjouer cette limite, l'identification des espèces bactériennes se fait maintenant par une approche métagénomique.
La métagénomique est l'étude de l'ensemble des organismes vivants dans un même milieu complexe tel que l'océan, le sol, les intestins, etc. Elle permet grâce au séquençage direct des génomes d'un échantillon de déterminer la présence et l'abondance des organismes. Avec les banques de données contenant les génomes déjà séquencés, il est possible d'identifier jusqu'à la souche d'une espèce. Dans les cas de génomes non séquencés ou non connus, il est possible de les regrouper à leurs plus proches voisins dans un arbre phylogénétique.

Pour donner un exemple de l'utilisation de la métagénomique, je vais citer le grand projet de TARA Océans (http://oceans.taraexpeditions.org). Tara est une goëlette qui a parcouru les mers et océans du Globe entre 2009 et 2013 pour prélever des échantillons de planctons en 209 lieux différents. Ces échantillons contiennent chacun un ensemble de virus, bactéries, protozoaires, petits métazoaires comme les copépodes, d'organismes gélatineux et de larves de poissons dont les espèces et leur quantité diffèrent selon la localité d'où l'échantillon provient. Grâce à l'utilisation seule ou combinée de techniques d'analyses microscopique, morphologique et métagénomique (Brum et al., 2015; Lima-Mendez et al., 2015; Sunagawa et al., 2015; Vargas et al., 2015), nous avons donc une meilleure connaissance des écosystèmes planctoniques, ce qui nous permet de savoir comment les différents organismes interagissent les uns avec les autres et avec leur environnement physico-chimique.

Les premiers projets de métagénomique n'ont pas pu utiliser les protocoles standards d'analyses génomiques : séquencer le plus possible, assembler les reads en séquences plus grandes pour finalement les annoter. En effet, les jeux de données produits ainsi présentent plusieurs difficultés :

- Leur assemblage de novo est complexe voire parfois impossible du fait de la diversité des génomes et de l'abondance variable des populations. Les algorithmes des assembleurs classiques sont conçus pour assembler les reads en une seule séquence consensus. Les génomes les plus gros et les plus présents seront peut être assemblés, mais présenteront vraisemblablement de nombreuses erreurs dues à des reads provenant d'autres espèces et présentant une certaine similarité. Les espèces très faiblement représentées dans l'échantillon n'auront que très peu de reads et ont toutes les chances d'être considérées comme des artéfacts et ainsi ignorées.
- L'alignement des reads pose aussi problème. Les microorganismes telles que les bactéries partagent souvent des séquences communes et lorsqu'elles sont supérieures à la taille des reads, ces derniers peuvent donc s'aligner sur plusieurs génomes et il est donc impossible de savoir d'où provient réellement la séquence. Il est important de prendre en compte certains reads s'alignant avec moins de similarité dans les cas où les génomes ne seraient pas disponibles dans les banques de données mais qui contiendrait leurs « proches ». Mais comment savoir où poser la limite ?

Il est important d'avoir connaissance de ces difficultés avant de commencer toute analyse. Il faut aussi garder en tête qu'il est très difficile voir impossible d'avoir la liste exacte et précise de la composition d'un échantillon.

La majorité des projets en métagénomique (identifications bactériennes) est clairement dominée par le séquençage des unités 16S des ARN ribosomaux (Kuczynski et al., 2010) mais ces derniers montrent quelques limites (Liu et al., 2011; Manichanh et al., 2008). En effet, l'estimation des populations peut être biaisée par les larges différences du nombre de copies d'ARNr entre les espèces assez proches et par l'utilisation des primers de PCR qui peuvent ne pas amplifier tous les ARNr. Un séquençage complet des séquences d'ADN présentes dans l'échantillon est plutôt conseillé car moins cher et plus rapide (Manichanh et al., 2008).

La métagénomique est une discipline très récente qui génère une masse importante de données. Avec la baisse constante du prix des séquençages, de plus en plus d'analyses métagénomiques sont lancées. Ce n'est donc plus la génération des données qui est coûteuse mais plutôt leurs analyses. Aussi seuls quelques logiciels parmi les plus connus en génomique sont utilisables en pratique (SOAPdenovo, Velvet, Bowtie, BLAST).

Cependant la métagénomique et la génomique ne répondent pas aux mêmes questions et ces dernières années, de nouveaux logiciels adaptés à la métagénomique ont été créés.

Deux types de logiciel ont émergés :

- ⇨ Des outils affectés à une tâche unique (Tableau 1)
- ⇨ Des pipelines permettant une analyse complète de données de métagénomique (Tableau 2)

Outils	Site	Plateformes	Mises à jour régulières
Genometa	http://genomics1.mh-hannover.de/genometa/	Testé sur Linux et Windows	Non
Kraken	http://ccb.jhu.edu/software/kraken/	Linux	Oui
Krona	http://sourceforge.net/p/krona/home/krona/	Linux, Mac	Oui
MEGAN	http://ab.inf.uni-tuebingen.de/software/megan5/	Linux, Mac, Windows	Oui
MetaVelvet	http://metavelvet.dna.bio.keio.ac.jp/	Linux	Pas depuis 2012
RayMeta	http://denovoassembler.sourceforge.net/	Testé sur Linux	Oui

Tableau 1. Outils de métagénomique dédiés à une analyse unique

Parmi eux, les quatre outils suivants ont été initialement développés spécifiquement pour la métagénomique :

- Genometa est un outil faisant partie de IGB (Integrative Genome Browser) et permettant à l'aide de Bowtie d'aligner et de classer les reads sur une base de données de bactéries et d'archées (Davenport et al., 2012).
- Kraken attribue des étiquettes taxonomiques grâce à l'alignement exact de k-mers entre les reads et une base de données microbienne (Wood and Salzberg, 2014).
- Krona est un outil interactif de visualisation des reads alignés sur des bases de données microbiennes (Ondov et al., 2011).
- MEGAN (MEta Genome ANalyzer) est le premier à voir le jour en 2007 (Huson et al., 2007). Le but de MEGAN est d'analyser les résultats obtenus à partir d'un BLAST entre des reads et une base de données de séquences de références du type NCBI. Il effectue une classification taxonomique et fonctionnelle utilisant la classification de KEGG (Kyoto Encyclopedia for Genes and Genomes) et/ou de SEED (Overbeek et al., 2005). La fenêtre graphique permet l'exploration des résultats. Des comparaisons de plusieurs sets de données sont possibles. MEGAN est en constante évolution, la version actuelle étant la cinquième.

D'autres, par contre, sont une amélioration d'outils existants en vue de leur utilisation pour la métagénomique. C'est ainsi que les assembleurs Velvet et Ray ont sorti MetaVelvet (Namiki et al., 2012) et Ray Meta (Boisvert et al., 2012).

Chacun de ces six outils nécessitent l'utilisation combinée d'autres outils pour obtenir une analyse complète des données de métagénomiques. De nouveaux outils ou pipelines, répondant à l'exigence d'une analyse complète, ont donc vu le jour en parallèle.

Outils	Site	Plateformes	Mises à jour régulières
metaBEETL	http://beetl.github.io/BEETL/	Linux, Mac	Oui
MetAMOS	https://github.com/marbl/metAMOS	Linux, Mac	Oui
MG-RAST	http://metagenomics.anl.gov/	Serveur disponible via Internet	Oui
MOCAT	http://vm-lux.embl.de/~kultima/MOCAT/index.html	Machines UNIX	Oui

Tableau 2. Outils de métagénomique permettant une analyse complète des données

Ces pipelines prennent tous en entrée le fichier de reads issus du séquenceur et fournissent la liste des espèces identifiées. Certains effectuent des analyses supplémentaires :
metaBEETL classe les reads de manière taxonomique par leur indexation et celle des séquences de référence (Ander et al., 2013). Il utilise le même algorithme d'indexation BWT (Burrows-Wheeler Transform) que Bowtie, seulement il n'a pas de limite de nombre ou de la taille des séquences de référence.
MetAMOS est un pipeline d'analyse automatique et modulable dont le code source est libre (Treangen et al., 2013). Il est possible de générer à partir de données métagénomiques des assemblages, des annotations taxonomiques et fonctionnelles, des listes de cadres de lecture ouverts, des motifs de variants, des graphes Krona.
Le serveur MG-RAST est une plateforme d'analyse automatique des données métagénomiques (Meyer et al., 2008). Le pipeline assigne les reads grâce aux bases de données nucléiques et protéiques. Des résumés d'ordre phylogénétique et fonctionnel sont générés à partir des données et des outils pour la comparaison de métagénomes sont disponibles. L'accès aux données est privatisé mais il est possible de les rendre accessibles pour permettre la collaboration et l'échange entre laboratoires.
MOCAT est un pipeline décrit comme modulable et configurable permettant l'analyse rapide des données de séquençage Illumina (Kultima et al., 2012). MOCAT utilise des outils existants (FastX, SOAP, Prodigal, etc) pour le contrôle de la qualité, l'alignement, l'assemblage et la prédiction de gènes.

Notre objectif est de lister les microorganismes présents chez chacune de nos lignées de moustiques afin d'identifier des espèces qui pourraient avoir un rôle dans la résistance de leur moustique hôte aux parasites du paludisme.
Dans cette partie, je démontrerai tout d'abord la nécessité de développer notre propre outil de détection des microorganismes par rapport aux outils existants ne répondant pas à notre problématique, puis je présenterai notre nouveau pipeline et analyserai avec un set de données métagénomique issu d'une nos lignées de moustiques.

II. Matériels et méthodes

- Matériel biologique

Pour tester l'efficacité des outils de métagénomique sélectionnés sur des données réelles, j'ai téléchargé deux sets de données du bioprojet 48475 du Human Microbiome Project (HMP) dont on connaît la composition (http://hmpdacc.org/HMMC/ et http://www.ncbi.nlm.nih.gov/bioproject/48475).
Le premier set nommé MockE est composé d'une quantité égale de bactéries et le deuxième nommé MockS contient des quantités variables de bactéries (Tableau 3).

	MockE		MockS	
Organisme	Copies 16S	Masse d'ADNg	Copies 16S	Masse d'ADNg
Acinetobacter baumannii ATCC 17978	100 000	1.60E-10	10 000	1.60E-11
Bacillus cereus ATCC 10987	100 000	3.73E-11	100 000	3.73E-11
Bacteroides vulgatus ATCC 8482	100 000	1.52E-10	1 000	1.52E-12
Clostridium beijerinckii ATCC 51743	100 000	3.81E-11	100 000	3.81E-11
Deinococcus radiodurans DSM 20539	100 000	1.76E-09	1 000	1.76E-11
Enterococcus faecalis ATCC 47077	100 000	2.22E-11	1 000	2.22E-13
Escherichia coli ATCC 700926	100 000	2.71E-11	1 000 000	2.71E-10
Helicobacter pylori ATCC 700392	100 000	4.50E-11	10 000	4.50E-12
Lactobacillus gasseri DSM 20243	100 000	1.53E-11	10 000	1.53E-12
Listeria monocytogenes ATCC BAA-679	100 000	3.98E-11	10 000	3.98E-12
Methanobrevibacter smithii ATCC 35061	100 000	9.50E-11	1 000 000	9.50E-10
Neisseria meningitidis ATCC BAA-335	100 000	6.87E-11	10 000	6.87E-12
Propionibacterium acnes DSM16379	100 000	1.39E-10	10 000	1.39E-11
Pseudomonas aeruginosa ATCC 47085	100 000	1.80E-10	100 000	1.80E-10
Rhodobacter sphaeroides ATCC 17023	100 000	1.30E-10	100 000	6.97E-11
Staphylococcus aureus ATCC BAA-1718	100 000	6.97E-11	1 000 000	1.31E-09
Staphylococcus epidermidis ATCC 12228	100 000	1.31E-10	100 000	1.83E-11
Streptococcus agalactiae ATCC BAA-611	100 000	1.83E-11	1 000 000	4.70E-10
Streptococcus mutans ATCC 700610	100 000	4.70E-11	1 000	8.11E-13
Streptococcus pneumoniae ATCC BAA-334	100 000	8.11E-11	100 000	6.97E-11

Tableau 3. Contenance des deux sets MockE et MockS. La masse d'ADN génomique (ADNg) est en grammes. Cette masse est calculée à partir de la masse d'ADNg d'une copie 16S multiplié par le nombre de copies 16S intégrées dans l'échantillon pour chaque bactérie.

Afin de tester l'efficacité des outils sélectionnés et d'évaluer la méthode de détection des microorganismes de notre outil, j'ai simulé un ensemble de séquences supposé proche de ce qui pourrait être obtenu à partir d'un échantillon de moustique entier et dont la composition exacte, en terme de composition et d'abondance est dans ce cas connue. Ce jeu de séquences comprend, outre le moustique, du parasite, de la levure, des bactéries et des virus.

ART est un outil de simulation de reads issus de différentes plateformes de séquençage (Huang et al., 2012). La version utilisée est Vanilla Ice Cream. ART permet de créer un fichier contenant, pour notre projet, des reads Illumina avec des modèles d'erreurs empiriques et des profils de qualité issus de données de séquençage Illumina. Dans le fichier de sortie, nous avons pour chaque génome ou groupe de génomes présents dans le fichier initial, un nombre de reads simulés :

- *Anopheles gambiae* : 33 496 498 reads
- *Plasmodium berghei* : 1 225 830 reads
- *Saccharomyces cerevisiae* : 482 823 reads
- 6 génomes de virus : 96 302 reads
- 450 bactéries et archées (215 genres) : 92 755 983 reads

Le fichier comptabilise un total de 128 057 436 reads de 75 bases pour un poids de 29,06 Go.

Dans le cadre du projet d'identification des facteurs génétiques contrôlant la résistance des moustiques aux parasites (Chapitre I), nous disposons de données de séquençage d'ADN génomique de moustiques. Cependant, par rapport à notre objectif, il faut prendre en compte que ces ADNg ont été préparés à partir de moustiques entiers et que les échantillons ne contiennent pas seulement les microorganismes intestinaux mais aussi les intracellulaires et ceux des parties externes des organes du moustique.

- Analyses métagénomiques

Parmi tous les outils disponibles pour effectuer une analyse de données métagénomiques, certains ne seront pas utilisables pour notre projet car non adaptés soit à notre problématique, soit à nos données. Le but de MOCAT est d'assembler les reads puis de prédire les gènes codant des protéines au sein des contigs. Le serveur MG-RAST débute par une étape pour enlever les reads d'organismes modèle tels que la drosophile, la souris, l'homme. Mais cette liste est limitée et ne contient pas le génome du moustique. MEGAN requière en input un fichier d'alignements BLAST. Or, faire un BLAST de notre masse de reads demande énormément de temps et des ressources informatiques adaptées. De plus un BLAST ne prend pas en compte la qualité des reads, ni l'information paired-end (PE). metaBEETL classe 31 millions de reads en 42h, nous en avons 280 millions. Ce qui pose la question si on peut trouver un outil plus rapide.

Trois outils ont manifesté des signes de dysfonctionnement lors de l'installation ou lors des essais. Genometa ne se lance pas en raison d'un problème avec ma version de Java alors que IGB fonctionne très bien. Vu que Genometa date de 2011 et qu'il n'est plus suivi, je n'utiliserai pas cet outil pour mes analyses. MetaVelvet bloque à l'étape du scaffolding quelque soit la machine (Linux ou Mac) sur laquelle il tourne. J'ai trouvé sur différents blogs plusieurs utilisateurs ayant la même erreur, malheureusement aucune solution n'était proposée. J'ai contacté les développeurs qui m'ont dit qu'ils allaient voir d'où venait le problème, mais je n'ai plus eu de réponse et aucune mise à jour du programme n'a été faite. Malgré le fait que la publication de MetaVelvet témoigne des avantages qu'il a par rapport à Velvet dans le cas de métagénomes, une autre publication démontre qu'il n'y a aucune différence significative entre les deux programmes (Vázquez-Castellanos et al., 2014). RayMeta classait toujours les 128 millions de reads du set simulé 10 jours après le début de la classification. Le programme s'est arrêté à 10j puisque c'est le temps de fonctionnement maximum autorisé d'un programme sur l'HPC. Il aurait été possible de prolonger ce temps exceptionnellement mais il est clair qu'il vaut mieux trouver un outil beaucoup plus rapide.

Il nous reste donc un pipeline permettant une analyse complète : MetAMOS, un outil de classification : Kraken, et un outil de visualisation : Krona. Ces trois outils seront donc utilisés dans ce chapitre et combinés à d'autres outils pour analyser nos données métagénomiques (Tableau 4).

Outils	Version	Publication
BLAST	2.2.30+	(Altschul et al., 1990; 2008)
Bowtie	1.1.1	(Langmead et al., 2009)
FastQC	0.10.1	-
Kraken	0.10.4	(Wood and Salzberg, 2014)
Krona	2.4	(Ondov et al., 2011)
MetAMOS	1.5rc3	(Treangen et al., 2013)
SolexaQA++	3.1.3	(Cox et al., 2010)
Velvet	1.2.10	(Zerbino and Birney, 2008)

Tableau 4. Outils sélectionnés pour l'identification des microorganismes dans un set de données métagénomiques

MetAMOS sera utilisé sur le set de données simulées avec les paramètres par défaut pour évaluer sa capacité à détecter les microorganismes présents.
FastQC et SolexaQA++ sont des outils de vérification de la qualité des reads. Ils seront intégrés dans notre nouvel outil de détection.
BLAST, Bowtie et Kraken alignent et assignent les reads à une espèce donnée présente dans la base de données utilisée. Bowtie et Kraken, ayant une fonction similaire, seront comparés sur leur capacité à détecter les microorganismes présents dans des sets de données identiques. Les résultats détermineront leur place au sein du nouvel outil. BLAST sera utilisé dans le nouvel outil pour aligner des contigs issus de l'assemblage par Velvet (vu dans le premier chapitre) sur des bases de données choisies.
Krona est un outil de visualisation sous forme circulaire du nombre de reads assignés à chaque espèce. Il sera intégré au nouvel outil afin de créer un fichier HTML interactif à partir du fichier issu de Kraken.

Pour les alignements avec Bowtie et BLAST, les génomes de référence proviennent des bases de données RefSeq du National Center for Biotechnology Information (NCBI). J'ai téléchargé les génomes des bactéries, des virus, des phages et des champignons à la date du 13 octobre 2014. Dans le cas de Kraken, leur site fournit une base de données contenant les génomes des virus, des bactéries et des archées provenant de la base RefSeq du NCBI en date du 8 décembre 2014.

- Matériel informatique
Pour la création de notre outil, j'ai utilisé mon MacBook Pro et toutes les analyses ont été réalisées sur l'HPC, mieux équipé en mémoire vive (24 Téraoctets (To) pour l'HPC contre 8 Gigaoctets (Go) pour le MacBook Pro) et en nombre de processeurs (380 processeurs pour l'HPC contre 1 pour le MacBook Pro).

III. Résultats

1. Application de l'outil MetAMOS à nos données simulées

1.1. Présentation

MetAMOS est un condensé d'outils bioinformatiques permettant de réaliser des analyses automatiques et reproductibles de données génomiques et métagénomiques. Il se veut le plus polyvalent possible en proposant le plus d'outils possibles: alignement, assemblage, analyse de variants, recherche d'ORFs, classification taxonomique, annotation fonctionnelle, etc. Ce programme regroupe donc un panel impressionnant d'outils : FastQC, Bowtie, SOAP, Velvet, BLAST, Krona, pour n'en citer que quelques uns. Le fait de limiter le nombre de lignes de codes à écrire et de rassembler autant d'outils fait de MetAMOS un outil très attractif autant pour les biologistes que pour les bioinformaticiens.

1.2. Installation

L'installation se fait automatiquement par le lancement d'un script python avec des paramètres modifiables. Par défaut, MetAMOS installe environ 350 Go de bases de données, en plus des outils. Pour les équipes n'ayant besoin que d'une base de données contenant les bactéries, virus et champignons, comme c'est le cas pour nous, une mini base peut être téléchargée et installée indépendamment mais elle est à demander par mail. Il faut ensuite spécifier lors du lancement du script qu'il ne faut pas qu'il installe les bases par défaut. L'installation est assez longue et un rapport sur les outils installés ou non est fourni à la fin. Pour les outils dont on aurait besoin et qui n'ont pas été installés, la recherche de la cause de la non-installation est compliquée. Les fichiers logs sont difficiles d'accès et éparpillés. De plus, les exigences de librairies ou packages pour le bon fonctionnement de chaque outil peuvent demander des mises à jour au niveau informatique. Mises à jour qui peuvent perturber le fonctionnement d'autres programmes du système.

1.3. Lancement des analyses

MetAMOS est composé de deux principales commandes : initPipeline et runPipeline. Le premier sert à créer les nouveaux projets et à initialiser les fichiers contenant les données à traiter. Le deuxième lance les différentes étapes d'analyse sur le projet sélectionné. Il y a 13 étapes par défaut. Le set de données utilisé pour tester MetAMOS est notre fichier de données simulées de 30 Go. MetAMOS a été lancé avec les paramètres par défaut.

Le programme s'est arrêté une première fois lors de l'étape de l'assemblage par SOAP. En cause, le fichier soapconfig.txt introuvable et qui est normalement créé lors de la première commande. En fait, un fichier config.txt était bien créé et il contenait les informations nécessaires à SOAP pour l'assemblage. Le fait de le renommer manuellement et de relancer la deuxième commande a réglé le problème. La cause du deuxième arrêt n'était pas due au fonctionnement des outils mais à l'atteinte du maximum de mon quota d'espace disque autorisé sur le serveur de l'HPC. En fait, notre petit fichier de données simulées de 30 Go a généré un peu plus de 1,3 To de données, bien que MetAMOS ne soit pas arrivé jusqu'à la fin de ses étapes. Après explications par un des développeurs, il s'avère que MetAMOS garde tous les fichiers de toutes les analyses faites par chaque outil. Ce sont les répertoires de l'assemblage et de la validation des assemblages qui sont les plus gros, 1 To à eux deux.

1.4. Conclusion

MetAMOS nous était apparu comme un couteau suisse fort pratique où tous les outils nécessaires à nos analyses étaient regroupés. Cependant c'est un outil très lourd qui demande beaucoup de mémoire vive, beaucoup d'espace de stockage et de la main d'œuvre adaptée pour fonctionner. Malgré le nombre important de publications en métagénomique et la multiplicité des analyses permises par MetAMOS, aucune n'utilise cet outil, ce qui suggère que d'autres groupes ont rencontré les mêmes problèmes que nous. J'ai déjà contacté l'équipe de développement sur certains points d'amélioration possibles.

2. Mise en place d'une nouvelle méthode d'analyse de nos données métagénomiques

2.1. Méthode initiale

Cet outil (Fig. 2) a vu le jour en 2011 lors de mon stage de Master. Une des tâches qui m'avait été confiées était la détection de virus et de la microsporidie Tubulinosema ratisbonensis (organisme non séquencé) dans des données de séquençage RNA-Seq de la mouche du vinaigre, Drosophila melanogaster. On voulait aussi étudier l'expression des gènes de la drosophile pour voir si l'interaction avec la microsporidie produisait des changements.

Dans un premier temps (étape I), les reads sont alignés avec TopHat (Trapnell et al., 2009) et Bowtie sur un ou plusieurs génomes de référence sélectionnés. Le fichier contenant les reads alignés peut être utilisé pour réaliser des analyses supplémentaires comme, par exemple, l'étude des gènes de l'organisme choisi. Les reads non alignés sont assemblés avec Trinity (Grabherr et al., 2011) pour former des contigs (étape II). Ces contigs sont BLASTés dans une base de données choisie du NCBI pour identifier l'organisme de provenance (étape III). Les résultats sont filtrés pour en sortir les plus pertinents (étape IV).
Dans le cadre du projet, les reads de Drosophila melanogaster ont d'abord été retiré du fichier initial grâce à l'alignement. Les reads restants ont formé plusieurs contigs qui ont permis d'identifier la présence de virus dans notre échantillon. Par contre, à ce stade, nous n'avions aucune trace de microsporidie. J'ai alors eu l'idée de réitérer mes étapes (symbolisé par la flèche rouge sur la figure 2) mais en enlevant cette fois-ci les reads des virus identifiés lors de la première itération. Cela a, en effet, permis de trouver des concordances dans la famille des microsporidies à la fin du second passage.

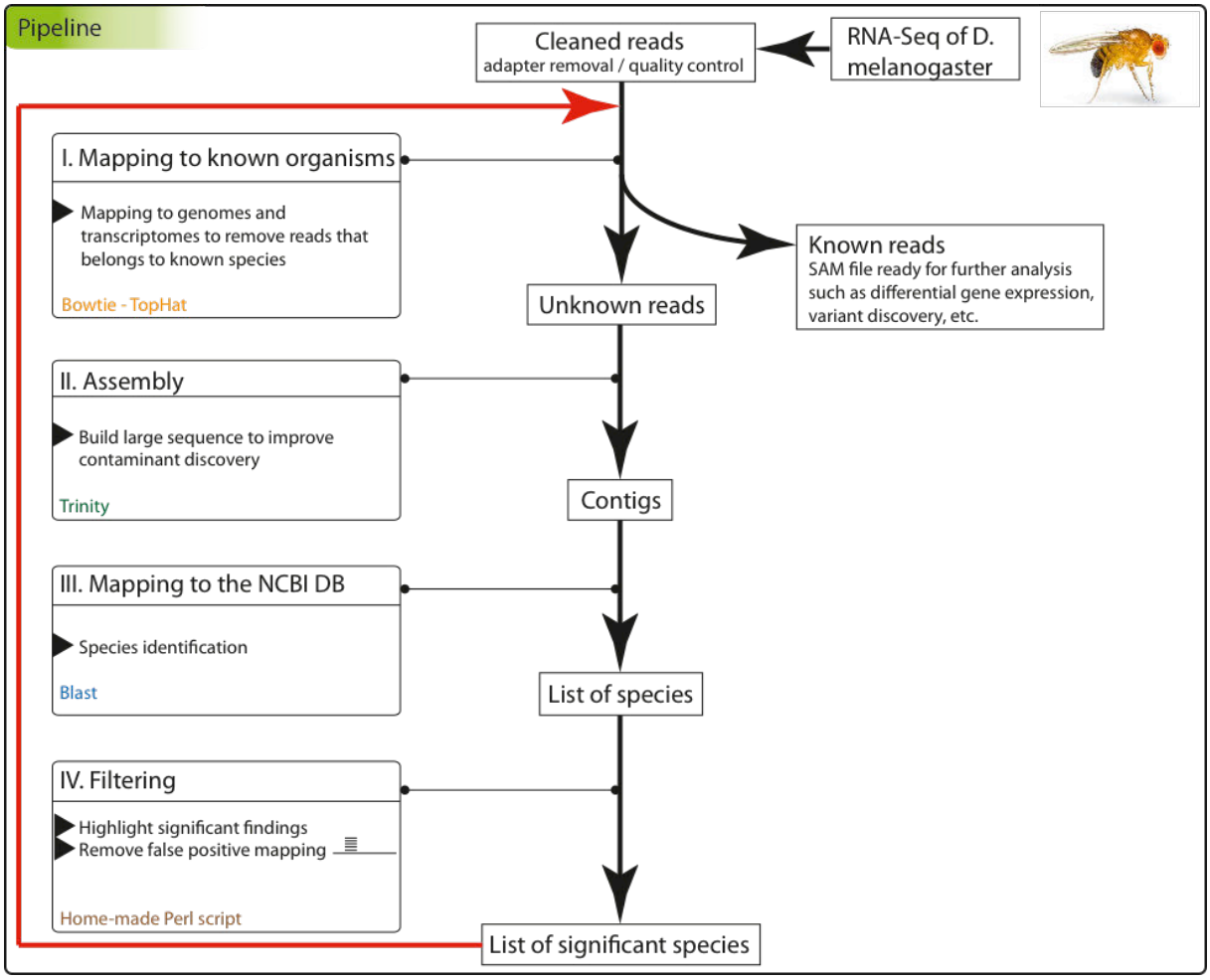


Figure 2. Données, méthodes et outils du pipeline initial.

2.2. Adaptations

Ce pipeline n'était cependant pas adapté à notre problématique, à savoir la découverte des microorganismes présents dans nos données de séquençage de moustiques. En effet, (1) nos données sont des reads d'ADNg et non des données transcriptomiques, il a donc fallu changer les outils puisque TopHat et Trinity sont spécialisés dans le traitement de données RNA-Seq ; (2) mon pipeline initial identifie les microorganismes par assemblage et BLAST, ce qui est efficace sur les espèces largement représentées, mais ne permet pas d'identifier des espèces moyennement ou faiblement représentées dans les données de séquençage. Dans ce cas, il est possible d'utiliser Bowtie pour aligner les reads sur des bases de données du NCBI téléchargées.
Avec l'avènement de la métagénomique, plusieurs logiciels, dont Kraken, ont été publiés et sont a priori plus adaptés aux données métagénomiques que Bowtie. Nous avons donc comparé Kraken et Bowtie afin de déterminer quel était le meilleur outil pour l'identification et la quantification des espèces présentes.

2.2.1. Évaluation des outils Bowtie et Kraken

Pour comparer les deux outils l'un par rapport à l'autre, j'ai sélectionné deux échantillons dont la composition en ADN génomique de chaque bactérie est connue, mockE et mockS, et un échantillon simulé (Tableau 5).

Set de données	MockE	MockS	Données simulées
Nombre de genres	17	17	215
Nombre d'espèces	20	20	330
Nombre de reads	6 562 065	7 932 819	92 755 983

Tableau 5. Caractéristiques des sets de données utilisés pour la comparaison des outils Bowtie et Kraken.

Le set mockE, pour mockEven, est composé d'une quantité égale de bactéries en terme de copies 16S mesurées, soit 100 000. Malgré cette équivalence, la masse d'ADN génomique de chaque bactérie varie entre 1,53E-11 et 1,76E-9 grammes. Cette masse est calculée à partir de la masse d'ADNg d'une copie 16S multiplié par le nombre de copies 16S intégrées dans l'échantillon pour chaque bactérie. Pour le set mockS (mockStaggered), le nombre de copies 16S varie entre mille et un million et la masse d'ADNg entre 2,22E-13 et 1,31E-9 grammes.
Dans le cas du set de données simulées, cet échantillon ne contient que des bactéries et le nombre de reads de chaque espèce qui le compose est connu.

Bowtie et Kraken ont tous les deux fonctionné sur l'HPC avec 16 processeurs. Bowtie a aligné les reads sur une base de données contenant des bactéries et des archées avec les options d'alignement par défaut. Kraken a aligné les reads sur sa mini base de données fournie sur leur site internet et qui contient les bactéries, les archées et les virus.
Au niveau du temps d'exécution des outils sur les trois sets de données, Bowtie est 300 fois plus rapide que Kraken pour les sets MockE et MockS et 28 fois plus rapide pour les données simulées (Tableau 6).

	MockE	MockS	Données simulées
Bowtie	1 min 5 s	1 min 39 s	9 min 33 s
Kraken	307 min	377 min	237 min

Tableau 6. Temps d'exécution des outils Bowtie et Kraken sur les trois sets de données.

La différence des bases de données de référence peut expliquer en partie la rapidité de Bowtie mais c'est plutôt la méthode de classification des reads qui est responsable du temps d'exécution des deux outils.
La complexité du dernier set se ressent chez Bowtie qui multiplie par 9 son temps d'exécution (Tableau 6). Alors qu'on s'y attend aussi chez Kraken, celui-ci passe une à deux heures en moins pour la classification de ce set. Alors que Kraken aligne deux fois moins de reads que Bowtie sur ce même set (Tableau 7), ce qui fait que l'écart entre les deux outil est plus faible, cela n'explique pas le fait qu'il passe moins de temps sur ce set comparé aux deux autres qui ont un nombre bien plus faible de reads (Tableau 5, ~11 fois moins).

	MockE		MockS		Données simulées	
	Bowtie	Kraken	Bowtie	Kraken	Bowtie	Kraken
Reads classés (%)	79,67%	75,40%	68,17%	67,07%	95,63 %	43,29 %

Tableau 7. Pourcentage de reads classés par Bowtie et Kraken pour les trois sets de données.

Concernant le nombre de reads classés pour les trois sets, c'est Bowtie qui en classe le plus qu'ils soient bien ou mal classés (Tableau 3). Pour les sets MockE et MockS, les deux outils ont classé presque le même nombre de reads. Par contre, il y a une nette différence sur le dernier set, puisque Kraken a classé un peu moins de la moitié des reads contrairement à Bowtie qui en classe 95% (Tableau 3). Il sera intéressant de voir par la suite si, malgré ce faible taux de classement, Kraken est capable de détecter toutes nos espèces.

Nous avons ensuite évalué chaque outil sur (1) sa sensibilité, ie leur capacité à détecter une espèce réellement présente, (2) sa spécificité, ie leur capacité à ne pas détecter une espèce non présente et (3) son estimation quantitative des espèces identifiées. Les évaluations se feront selon les niveaux taxonomiques du genre et de l'espèce. Dans notre projet, le but est d'identifier d'abord les genres présents dans nos échantillons puis si possible les espèces.

2.2.1.1. Sensibilité et spécificité

Notre premier critère d'évaluation se base sur la capacité des deux outils à identifier une bactérie qui est réellement présente dans le set de données de départ. Si la bactérie a été identifiée, c'est que Bowtie ou Kraken lui a attribué au minimum un read. Ce critère qui est la sensibilité se traduit par la proportion de vrais positifs (bactéries identifiées et réellement présentes) sur le nombre total de bactéries réellement présentes dans le set de départ. Plus cette proportion se rapproche de 100%, plus l'outil est sensible puisqu'il est apte à détecter une bactérie présente.

Notre second critère d'évaluation est la spécificité des 2 outils, i.e. leur capacité à ne pas identifier une bactérie qui n'est pas présente dans le set de départ. Dans ce cas, nous calculons la proportion de faux positifs (bactéries identifiées mais réellement absentes) par rapport au nombre total de bactéries identifiées par l'outil. Plus notre proportion est proche de 0%, plus l'outil est spécifique.

Nous allons commencer par analyser les genres puis les espèces identifiées dans les trois sets par Bowtie et Kraken.

2.2.1.1.1. Genres identifiés

Genre	MockE		MockS		Données simulées	
Critères	Bowtie	Kraken	Bowtie	Kraken	Bowtie	Kraken
# bactéries détectées	399	223	479	231	650	641
# moyen de reads alignés	13 103	21 965	11 290	22 212	136 469	63 179
# minimum / maximum de reads alignés	1 / 2 016 786	1 / 1 891 496	1 / 2 301 076	1 / 2 651 216	2 / 3 697 244	1 / 1 746 594
Vrais positifs (VP)	17	17	17	17	215	215
Faux positifs (FP)	382	206	462	214	435	426
Faux négatifs (FN)	0	0	0	0	0	0

Tableau 8. Bactéries identifiées au niveau du genre par Bowtie et Kraken sur les trois sets de données : MockE, MockS et les données simulées. Pour chaque set et l'outil utilisé, sont indiqués : le nombre de bactéries identifiées, le nombre moyen de reads alignés par genre, le nombre minimum et maximum de reads alignés par genre et, parmi les bactéries identifiées, le nombre de vrais positifs (VP), de faux positifs (FP) et de faux négatifs (FN).

Sur les trois sets, Kraken identifie toujours moins de bactéries même si l'écart est moins marqué sur le dernier set (Tableau 8). Au niveau du nombre moyen de reads par genre, Kraken assigne deux fois plus de reads que Bowtie sur les petits sets MockE et MockS. Pour les données simulées, c'est l'inverse qui se produit (Tableau 8, 63 000 reads pour Kraken contre 136 000 reads pour Bowtie). Nous avions vu que Kraken avait classé un peu moins de la moitié des reads présents dans ce set de départ alors que Bowtie en avait classé 95% (Tableau 7). Cela peut donc expliquer cette inversion.
En ce qui concerne les vrais positifs (VP), Bowtie et Kraken trouvent les mêmes (Tableau 8). Par contre, Kraken détecte moins de faux positifs (FP) dans le cas des trois sets de données.

Pour tenter d'éliminer ces FP, nous avons examiné le nombre de reads alignés pour chaque espèce trouvée. Ce nombre varie considérablement, allant de 1 read à environ 2 millions de reads. On remarque d'ailleurs que les VP possèdent un plus grand nombre de reads que les FP. Cependant, comme il n'y a pas de rupture brute entre ces deux classes, nous avons cherché à déterminer une valeur seuil qui permettrait de séparer les VP et les FP ou tout du moins d'éliminer le plus de FP possible. Afin de déterminer la valeur qui servira de filtre, j'ai recalculé la proportion de VP et de FP après l'application de différents filtres basés sur le nombre de reads assignés par genre.

Pour les trois sets, Kraken présente la même quantité ou beaucoup moins de FP après l'application de n'importe quel filtre sur les résultats (Fig. 3, 4 et 5).
Pour le set MockE et le set de données simulées, l'application du filtre à 1 000 reads minimum par genre sur les résultats de Kraken permet d'éliminer tous les FP et de garder tous les VP (Fig. 3 et 5).

Pour le set MockS, la valeur filtre de 10 000 reads minimum chez Kraken permet d'éliminer tous les FP mais n'identifie que 11 genres sur les 17 présents. Les 6 genres non identifiés sont en fait les 6 genres les plus faiblement représentés. Le fait de descendre la valeur à 1 000 reads minimum augmente le nombre de VP à 15 mais aussi celui des FP (12% des genres identifiés après la filtration, Fig. 4).

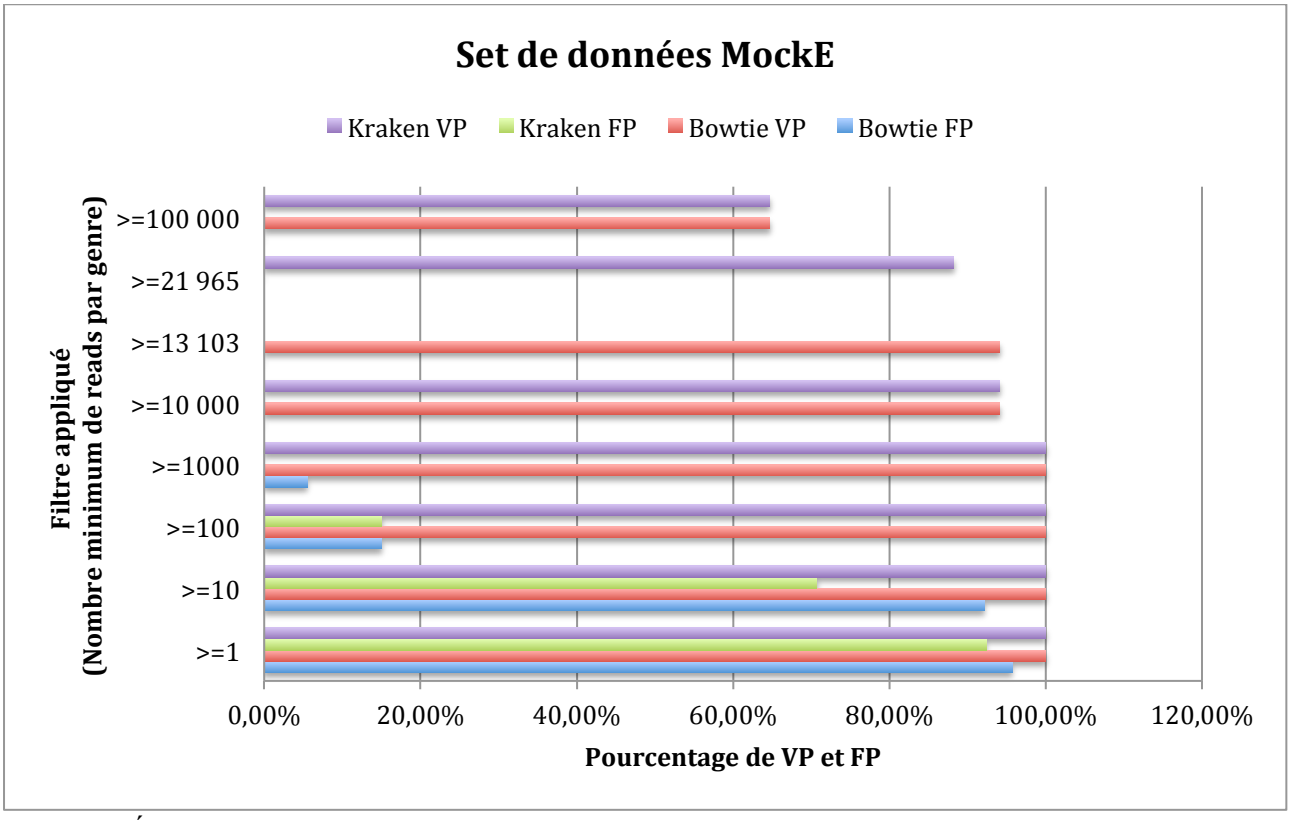


Figure 3. Évolution du pourcentage de genres VP et FP identifiés par Bowtie et Kraken sur le set MockE en fonction du nombre minimal de reads choisi comme filtre. Le pourcentage de VP est calculé par le nombre de VP après le filtre sur le nombre total de genres à identifier. Le pourcentage de FP est calculé par le nombre de FP après le filtre sur le nombre total de genres identifiés après le filtre. Les filtres 13 103 pour Bowtie et 21 965 pour Kraken correspondent à la moyenne du nombre de reads alignés par genre.

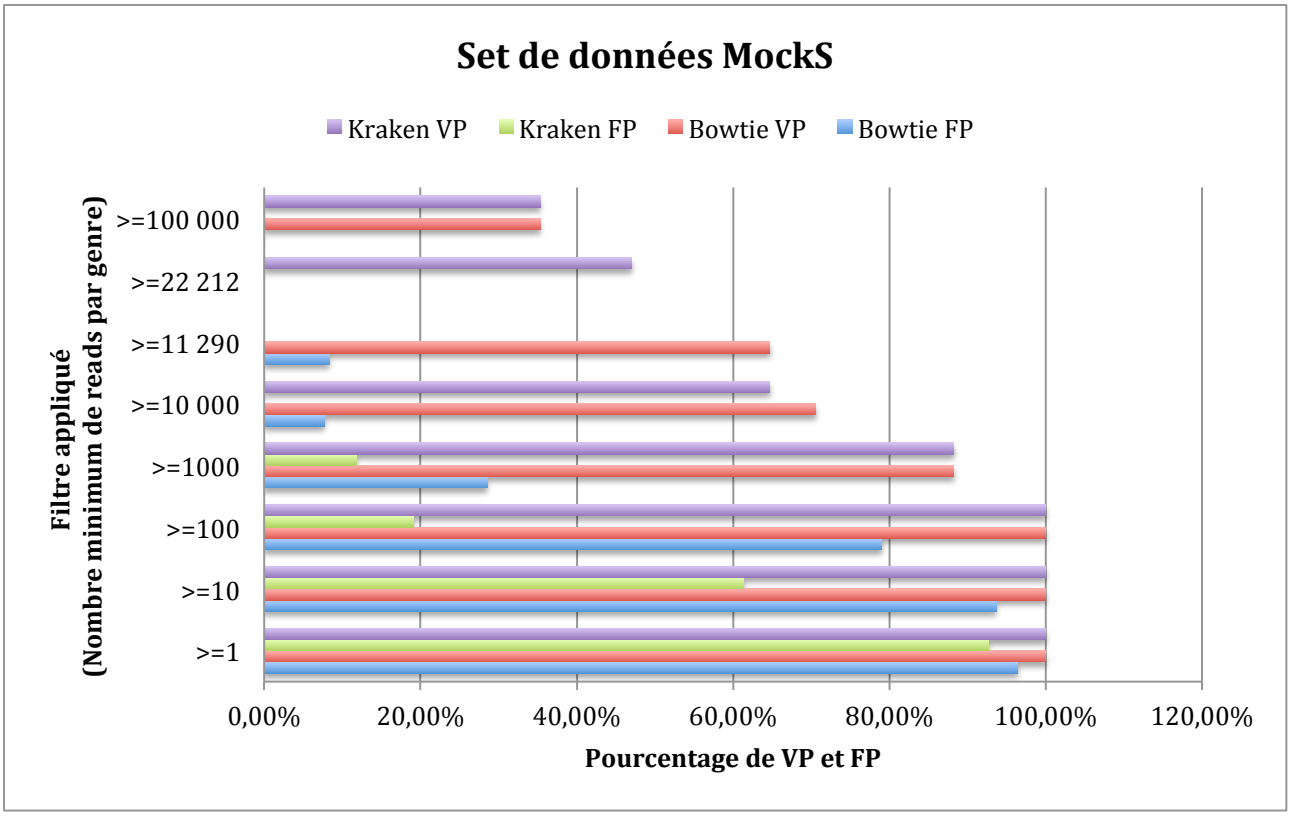


Figure 4. Évolution du pourcentage de genres VP et FP identifiés par Bowtie et Kraken sur le set MockS en fonction du nombre minimal de reads choisi comme filtre. Le pourcentage de VP est calculé par le nombre de VP trouvé après le filtre sur le nombre total de genres à identifier. Le pourcentage de FP est calculé par le nombre de FP après le filtre sur le nombre de genres identifiés en totalité après le filtre. Les filtres 11 290 pour Bowtie et 22 212 pour Kraken correspondent à la moyenne du nombre de reads alignés par genre.

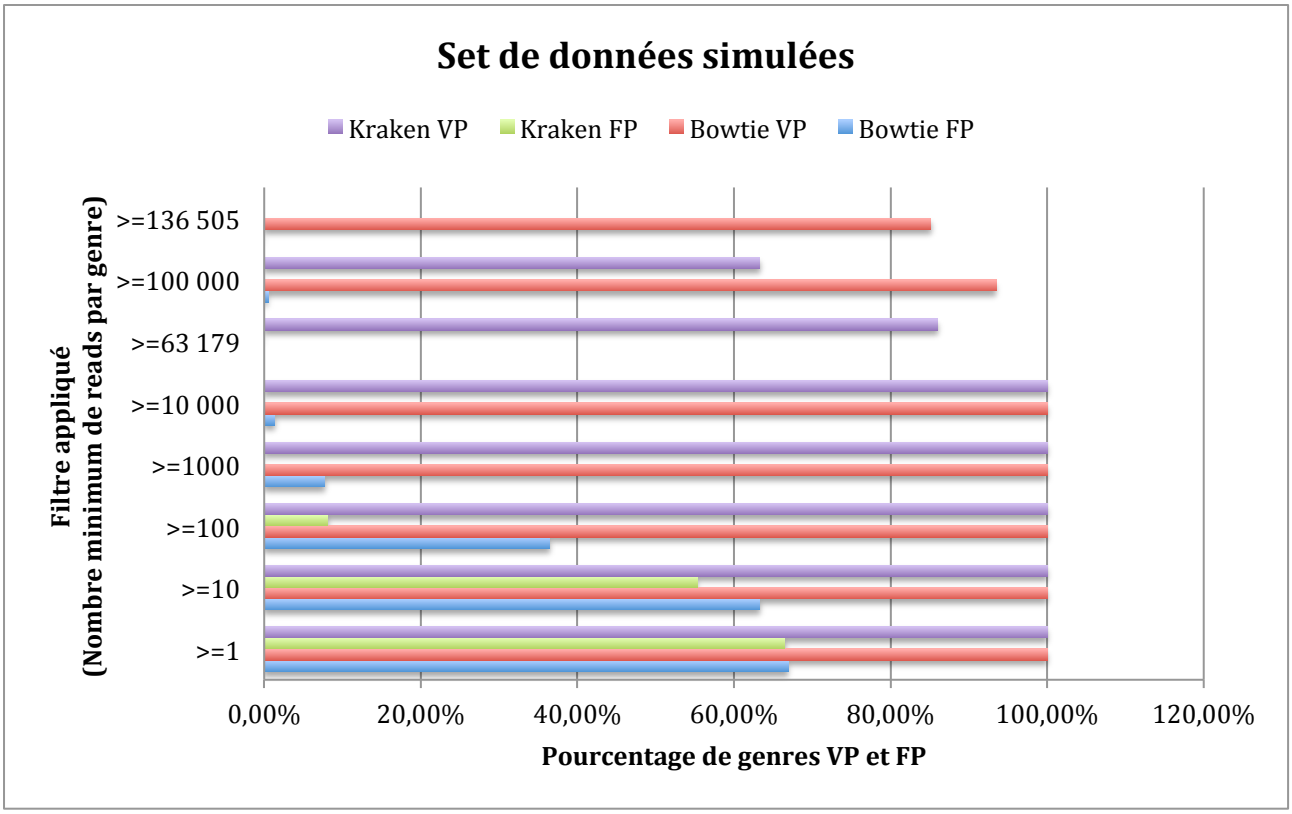


Figure 5. Évolution du pourcentage de genres VP et FP identifiés par Bowtie et Kraken sur le set de données simulées en fonction du nombre minimal de reads choisi comme filtre. Le pourcentage de VP est calculé par le nombre de VP trouvé après le filtre sur le nombre total de genres à identifier. Le pourcentage de FP est calculé par le nombre de FP après le filtre sur le nombre de genres identifiés en totalité après le filtre. Les filtres 136 505 pour Bowtie et 63 179 pour Kraken correspondent à la moyenne du nombre de reads alignés par genre.

2.2.1.1.2. Espèces identifiées

Nous avons ensuite analysé les résultats des bactéries identifiées par Bowtie et Kraken en se basant directement sur les espèces identifiées.

Espèces	MockE		MockS		Données simulées	
Critères	Bowtie	Kraken	Bowtie	Kraken	Bowtie	Kraken
# bactéries détectées	823	411	969	392	1 447	1 423
# moyen de reads alignés	6 352	11 176	5 580	12 422	61 324	25 854
# minimum / maximum de reads alignés	1 / 2 008 560	1 / 1 876 372	1 / 1 573 214	1 / 1 346 885	2 / 2 128 983	1 / 877 220
Vrais positifs (VP)	20	20	20	20	330	330
Faux positifs (FP)	803	391	949	372	1 117	1 093
Faux négatifs (FN)	0	0	0	0	0	0

Tableau 9. Bactéries identifiées au niveau de l'espèce par Bowtie et Kraken sur les trois sets de données : MockE, MockS et les données simulées. Pour chaque set et l'outil utilisé, il est indiqué le nombre de bactéries identifiées, le nombre moyen de reads alignés par espèce, le nombre minimum et maximum de reads alignés par espèce et, parmi les bactéries identifiées, le nombre de vrais positifs (VP), de faux positifs (FP) et de faux négatifs (FN).

Comme pour les genres, Kraken identifie moins de bactéries même si l'écart est plus faible sur le dernier set (Tab. 9). De même au niveau du nombre moyen de reads par espèce, Kraken assigne deux fois plus de reads que Bowtie sur les petits sets MockE et MockS. Pour les données simulées, c'est l'inverse qui se produit (Tab. 9, 25 000 reads pour 61 000 reads). Cette inversion peut être expliquée par le fait que Kraken avait classé un peu moins de la moitié des reads présents dans ce set de départ alors que Bowtie en avait classé 95% (Tab. 7). Comme pour les genres, Bowtie et Kraken trouvent les mêmes VP et Kraken détecte moins de FP (Tab. 9).

Afin de réduire le nombre de ces FP dans nos résultats, nous avons testé les mêmes filtres que précédemment en faisant varier le nombre de reads minimal, et compté combien de VP et FP étaient sélectionnés.

L'analyse des résultats obtenus après l'application des différents filtres sur les trois sets révèle les mêmes observations et déductions que la précédente analyse sur les genres identifiés pour les sets MockE et MockS.
Sur les deux petits échantillons, Kraken présente le même nombre ou beaucoup moins de FP que Bowtie sur un filtre équivalent (Fig. 6 et 7). La valeur filtre permettant d'éliminer tous les FP est de 1 000 pour MockE et de 10 000 pour MockS (Fig. 6 et 7). Dans le cas de MockS, il y a aussi une perte de VP, seulement 12 espèces sont identifiées sur les 20 présentes (Fig. 7).

Pour le set de données simulées, il a été rajouté une valeur filtre à 200 000 reads minimum pour enfin obtenir une liste de résultats sans FP par Bowtie (Fig. 8). Toutefois, seule la moitié des espèces présentes ont été identifiées.
En fait, chez Kraken, certaines espèces dont leur genre a été identifié et a passé le filtre ne passent pas le filtre des 1 000 reads minimum (3 espèces soit 0,91% de VP perdues, Fig. 8). Parmi ces trois espèces, deux font partie du genre *Hydrogenobaculum* qui contient 95 000 reads ce qui lui fait passer le filtre mais elles présentent seulement 24 et 32 reads. En fait, il est fortement probable que les reads soit situés sur une portion commune aux deux espèces ce qui fait que Kraken ne sachant pas d'où ils proviennent réellement préfère les attribuer au rang taxonomique supérieur. Alors que Bowtie va les assigner soit à l'un soit à l'autre. En alignant avec BLAST les génomes de ces deux espèces, H. sp. HO et H. sp. SN, il est apparu qu'ils ont 99% de bases communes. Pour la dernière espèce qui présente 418 reads attribués, son genre, *Chloroflexus*, présente 158 964 reads. Kraken doit aussi hésiter entre les génomes des espèces de *Chloroflexus* disponibles dans la base de données.

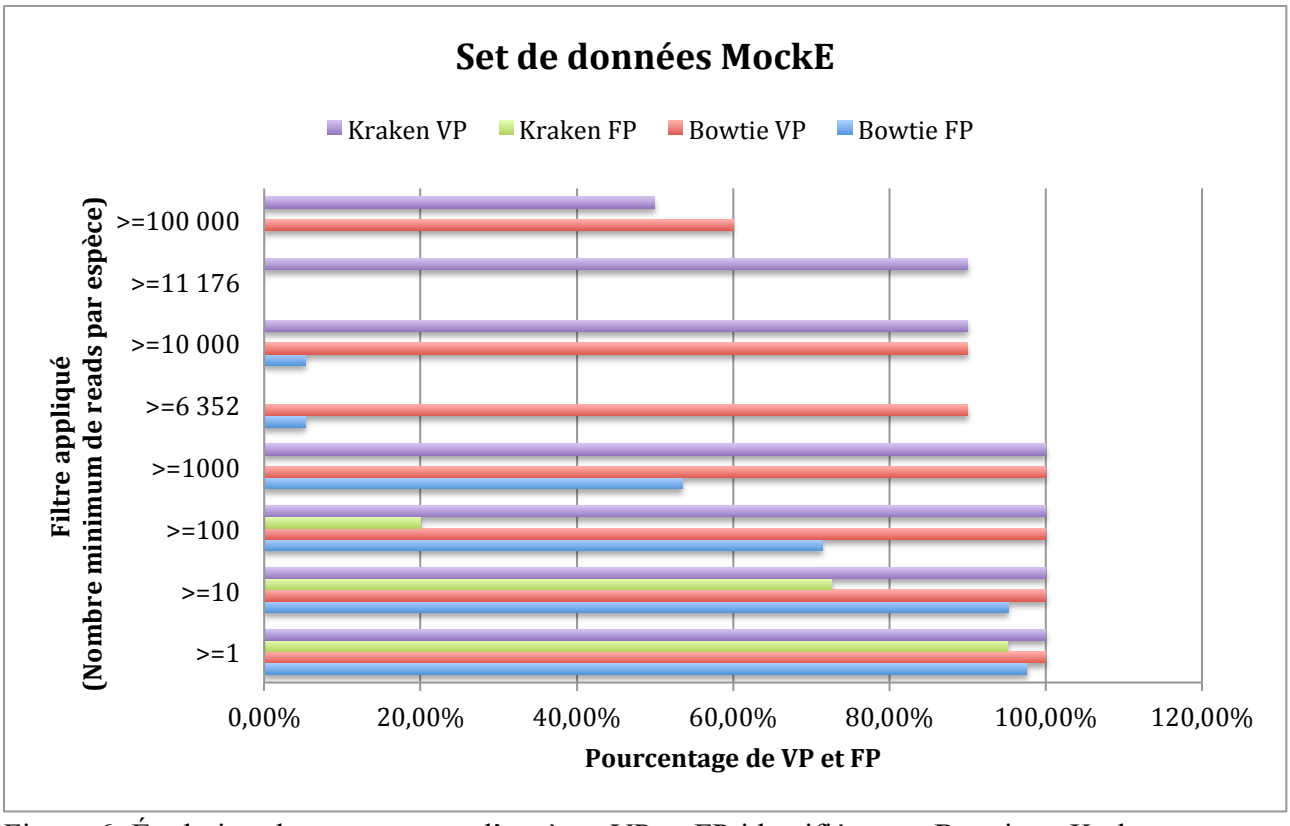


Figure 6. Évolution du pourcentage d'espèces VP et FP identifiées par Bowtie et Kraken sur le set MockE en fonction du nombre minimal de reads choisi comme filtre. Le pourcentage de VP est calculé par le nombre de VP trouvé après le filtre sur le nombre total d'espèces à identifier. Le pourcentage de FP est calculé par le nombre de FP après le filtre sur le nombre d'espèces identifiées en totalité après le filtre. Les filtres 6 352 pour Bowtie et 11 176 pour Kraken correspondent à la moyenne du nombre de reads alignés par genre.

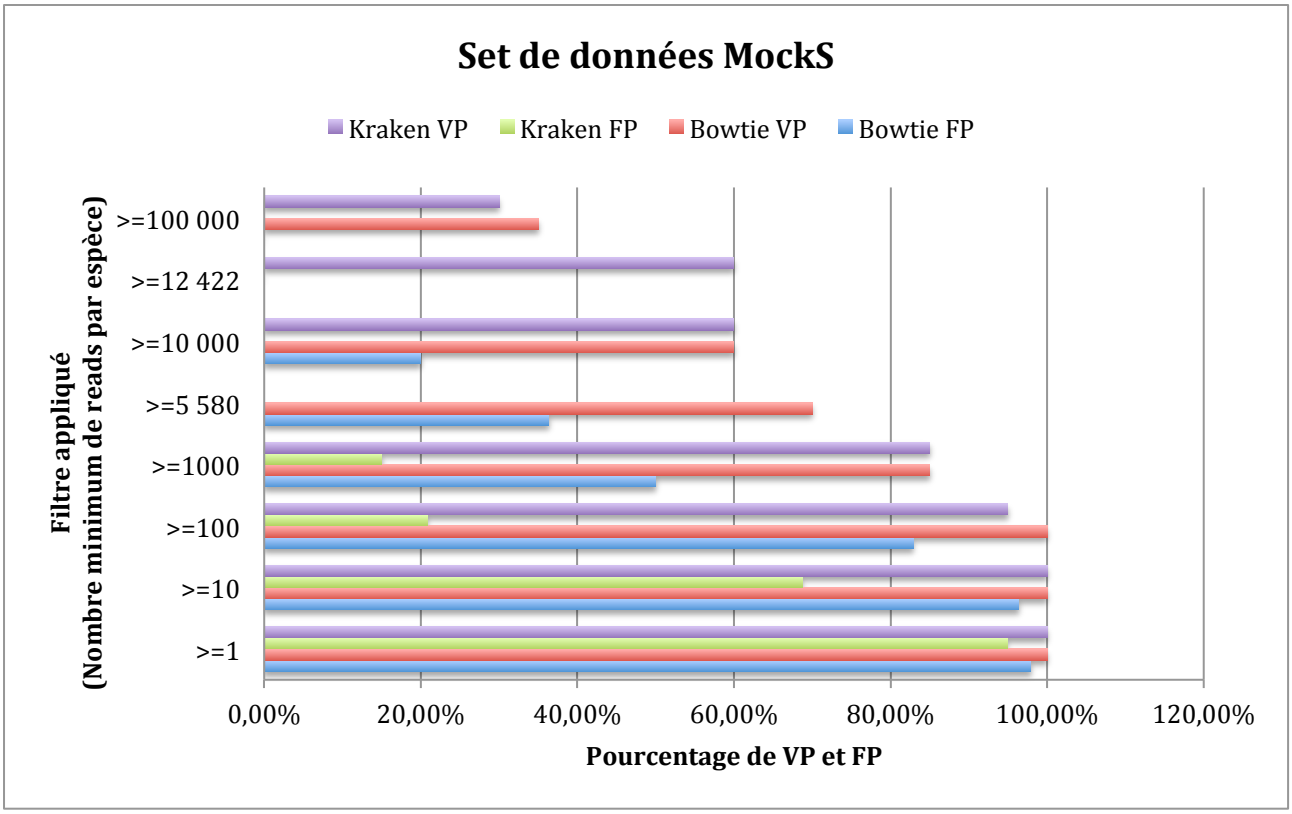


Figure 7. Évolution du pourcentage d'espèces VP et FP identifiées par Bowtie et Kraken sur le set MockS en fonction du nombre minimal de reads choisi comme filtre. Le pourcentage de VP est calculé par le nombre de VP trouvé après le filtre sur le nombre total d'espèces à identifier. Le pourcentage de FP est calculé par le nombre de FP après le filtre sur le nombre d'espèces identifiées en totalité après le filtre. Les filtres 5 580 pour Bowtie et 12 422 pour Kraken correspondent à la moyenne du nombre de reads alignés par genre.

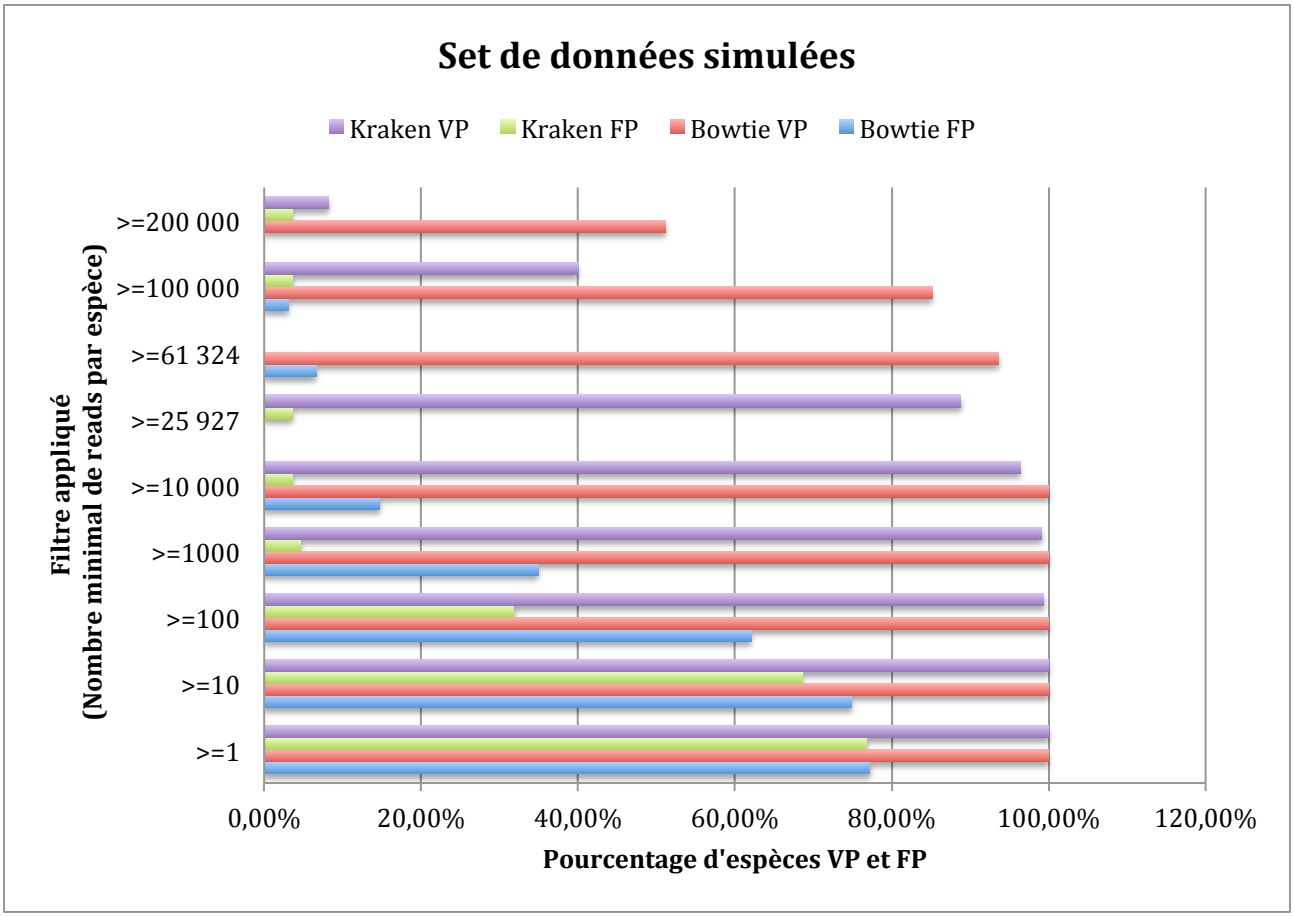


Figure 8. Évolution du pourcentage d'espèces VP et FP identifiées par Bowtie et Kraken sur le set de données simulées en fonction du nombre minimal de reads choisi comme filtre. Le pourcentage de VP est calculé par le nombre de VP trouvé après le filtre sur le nombre total d'espèces à identifier. Le pourcentage de FP est calculé par le nombre de FP après le filtre sur le nombre d'espèces identifiées en totalité après le filtre. Les filtres 61 324 pour Bowtie et 25 927 pour Kraken correspondent à la moyenne du nombre de reads alignés par genre.

2.2.1.1.3. Conclusions

Après avoir évaluer Bowtie et Kraken sur leur sensibilité et leur spécificité, nous en avons déduit : (1) Bowtie et Kraken sont aussi sensibles l'un que l'autre puisqu'ils ont identifié toutes les bactéries réellement présentes, (2) Kraken est plus performant que Bowtie en terme de spécificité puisqu'après l'application d'un filtre, il permet d'éliminer tous les FP des résultats avec aucune ou une faible perte de bactéries réellement présentes, (3) il est préférable de chercher à identifier directement les genres puis d'examiner les espèces des genres identifiés.

Pour le choix de la valeur filtre, elle est à déterminer selon la quantité de reads assignés en totalité et par genre. Malgré la différence du nombre de reads assignés en moyenne à un genre, nous avons vu qu'elle pouvait être identique (la valeur filtre du set MockE et du set de données simulées est de 1 000 reads minmum). Elle dépend aussi du but de l'analyse. La perte de VP lors de l'application du filtre peut être préférée à la présence de FP résiduels.
Dans ce cas où nous souhaitons être sûr des bactéries présentes quitte à en perdre quelques unes, la valeur filtre à appliquer sera plus haute.

2.2.1.2. Estimation quantitative des espèces

Nous avons ensuite cherché à évaluer la capacité des 2 outils à estimer l'abondance des espèces présentes dans les échantillons initiaux : MockE, MockS et les données simulées.
Les tableaux contenant les bactéries présentes, leur masse d'ADNg (sets MockE et MockS) ou leur nombre de reads initial (set de données simulées) ainsi que le nombre de reads que leur a attribué Bowtie et Kraken sont disponibles en Annexes V, VI et VII.

Pour les sets MockE et MockS, Bowtie et Kraken sont relativement proches en nombre de reads attribués (Annexe V et VI). Seulement 2 espèces, *Escherichia coli* et *Enterococcus faecalis*, présentent un nombre de reads assignés par Bowtie 2,5 voire 4 fois supérieur à celui de Kraken (Annexe V). En fait, si on remonte aux genres, ce n'est plus le cas pour *Enterococcus* qui totalise 70 000 reads attribués par Bowtie et 68 000 reads par Kraken, mais toujours pour *Escherichia* qui présente encore une différence de 4 fois plus de reads attribués par Bowtie (20 000 pour 80 0000 reads). En fait, 80 000 reads sont bien attribués par Kraken dans la famille des *Enterobacteriaceae* mais seulement 20 000 vont au genre *Escherichia*. Nous pouvons donc en déduire que les 60 000 reads restants ont des portions communes à plusieurs genres/espèces d'*Enterobacteriaceae* ce qui fait que Kraken les attribue à la famille et non à un genre/espèce précis selon sa méthode de classification. A ces deux espèces s'ajoutent *Streptococcus pneumoniae* et *Neisseria meningitidis* dans le set MockS qui affichent deux fois moins de reads chez Kraken que Bowtie pour la même raison (Annexe VI).

Par contre, dans les deux sets, certaines espèces sont sous-représentées en nombre de reads, que ce soit par Bowtie ou Kraken : *Pseudomonas aeruginosa* et *Methanobrevibacter smithii* (Annexe V). Pour ces deux espèces, il n'y a pas de dispersion de reads sur des espèces proches par Bowtie ou de reads assignés au rang taxonomique supérieur par Kraken. Par contre, dans le cas de *Methanobrevibacter smithii*, la masse d'ADNg a été calculée selon la masse d'ADNg théorique d'une copie 16S alors que pour toutes les autres bactéries, la masse d'ADNg d'une copie 16S de chaque bactérie avait pu être mesurée par PCR quantitative. D'après le nombre de reads attribués par les deux outils, on pourrait supposer avoir une masse d'ADNg totale d'environ 3E-11 grammes pour *Methanobrevibacter smithii*, ce qui ferait une masse d'ADNg d'une copie 16S à environ 3E-16 grammes au lieu des 9,5E-16 grammes théorique. Cela a été vérifié dans le set MockS par une nouvelle sous-estimation par les deux outils. Je n'ai pas d'explication concernant la sous-estimation de *Pseudomonas aeruginosa* dans le set MockE qui n'est pas retrouvée dans le set MockS.

La corrélation entre le nombre de reads alignés et la masse d'ADN génomique de chaque espèce présente dans l'échantillon est plutôt bonne dans le set MockE (Fig. 9, $R^2 \sim 0,9$ pour Bowtie et Kraken) malgré la présence de valeurs « outliers » (une sur-estimation et une sous-estimation sur les 20 espèces présentes). Cependant, elle est plutôt mauvaise dans le set MockS (Fig. 10, $R^2 \sim 0,5$ pour Bowtie et Kraken). Ce faible R^2, la cohérence entre Bowtie et Kraken dans ce set et plusieurs incohérences par rapport au set MockE (une espèce qui a la même masse d'ADNg dans les deux sets ne présente pas du tout la même quantité de reads attribués dans chacun des sets) nous ont poussé à douter de la véracité des masses d'ADNg de quelques espèces du set MockS. Il s'est avéré qu'en téléchargeant de nouveau le tableau décrivant le contenu des sets MockE et MockS, celui-ci ne présentait pas les mêmes masses d'ADNg pour les dernières espèces du set MockS. Il reste donc encore un doute fortement probable quant à la véracité de certaines masses d'ADNg de ce set. Il ne sera donc pas possible de prendre en compte l'abondance estimée par Bowtie et Kraken sur ce set dans notre évaluation.

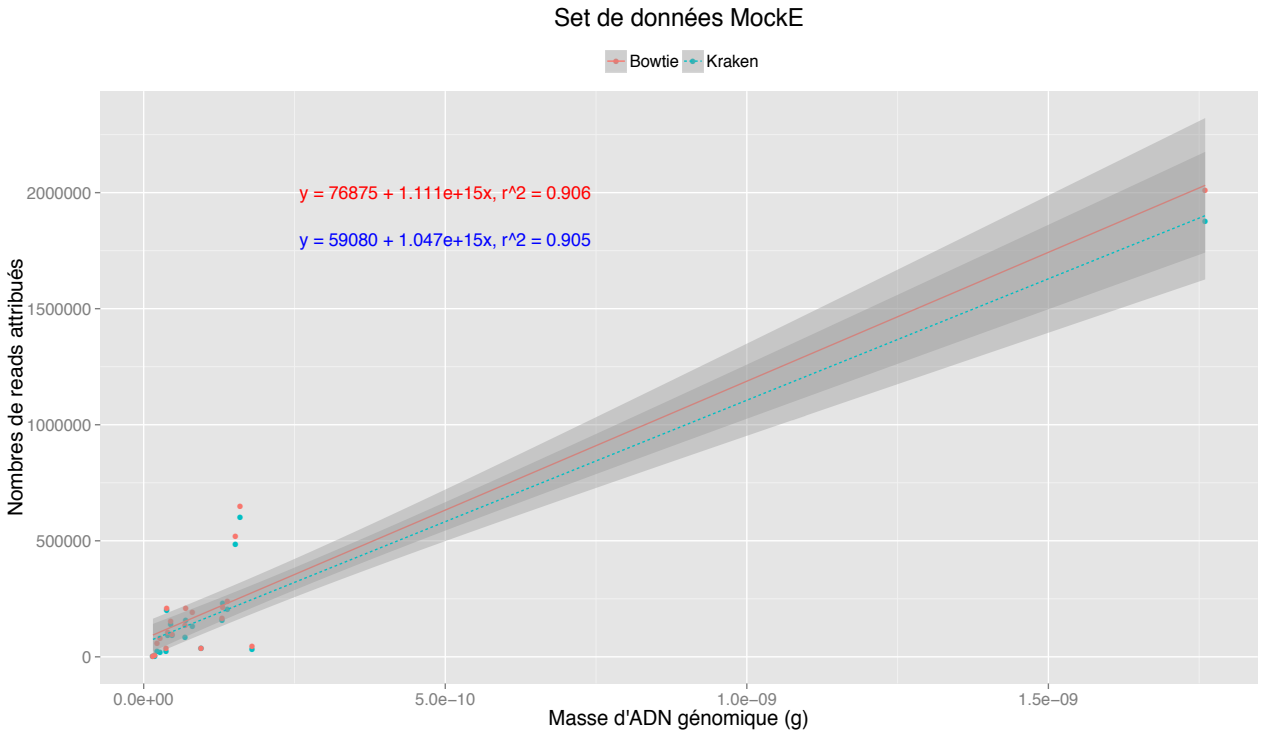


Figure 9. Corrélation linéaire simple entre la masse d'ADN génomique d'une espèce du set MockE et le nombre de reads assigné par Bowtie et Kraken à cette même espèce.

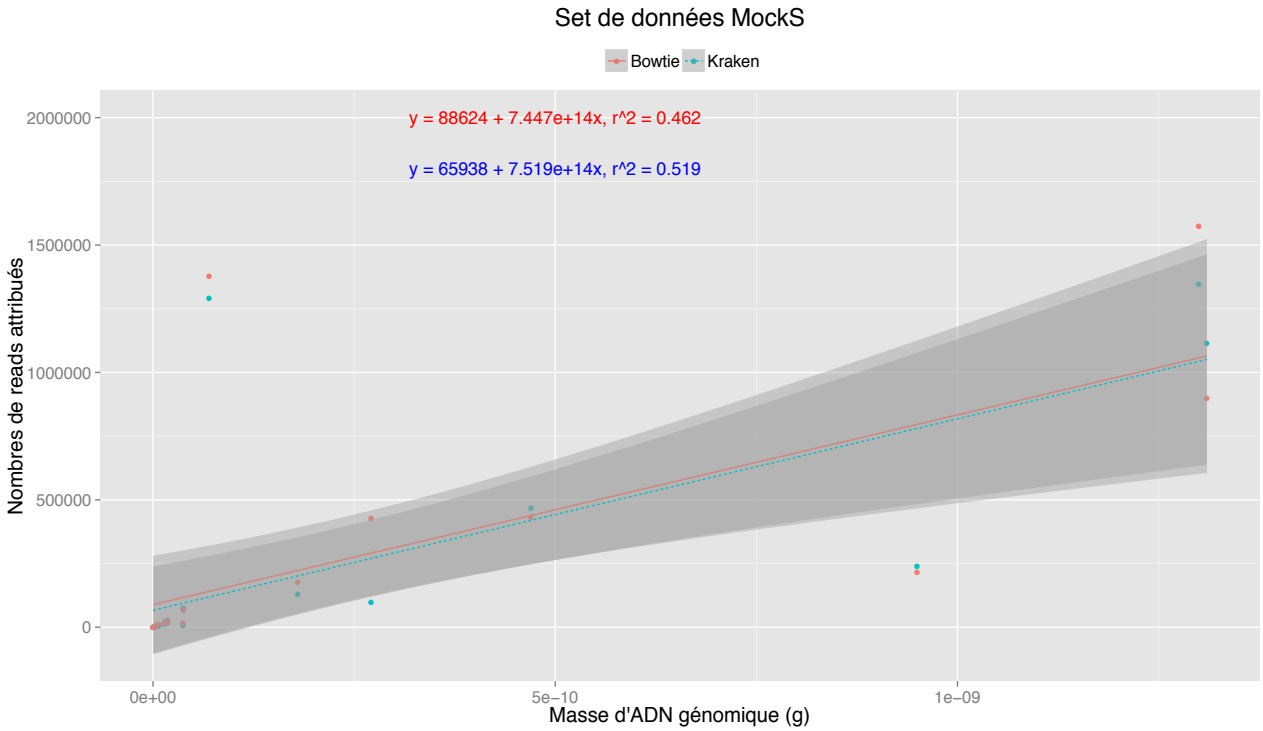


Figure 10. Corrélation linéaire simple entre la masse d'ADN génomique d'une espèce du set MockS et le nombre de reads assigné par Bowtie et Kraken à cette même espèce.

J'ai également réalisé une régression linéaire robuste sur le set de données MockE ce qui a permis d'améliorer les r^2 de Bowtie et Kraken à 0,974 et 0,977. En fait, une régression linéaire robuste ne prend pas en compte les valeurs qu'il considère comme valeurs « outliers » donc la corrélation sera plus robuste. Ces dernières sont les valeurs avant la dernière dans ce set de données : une sous-estimation et deux sur-estimations dont une des deux n'en été pas une puisque la masse d'ADNg de cette espèce n'est pas correcte. Dans le cas de cette régression, Kraken se montre plus robuste que Bowtie.
De plus, pour l'ensemble des données, les deux outils sont concordants sur leur nombre de reads attribués.

Dans le cas des données simulées, nous avions vu que Kraken avait classé seulement la moitié (43%) des reads du set alors que Bowtie en avait classé 95% (Tab 7) et que le nombre de reads assigné au genre ou à l'espèce identifié(e) par Bowtie est bien plus important que celui assigné par Kraken (Tableaux 8 et 9). On peut alors se demander si cela est observé sur quelques espèces ou sur une majorité. Dans le cas où seulement quelques espèces seraient touchées, cela provoquerait un biais dans l'estimation de l'abondance des bactéries.
Connaissant le nombre exact de reads appartenant à chaque genre ou espèce, j'ai exprimé le nombre de reads attribués par Bowtie et Kraken en proportion du total de reads du genre ou de l'espèce. Par exemple, pour *Acetobacter pasteurianus*, le nombre de reads dans l'échantillon de contrôle est de 187 910, Bowtie a aligné 179 373 reads sur cette espèce, soit 95% et Kraken en a aligné 84 337, soit 45%.

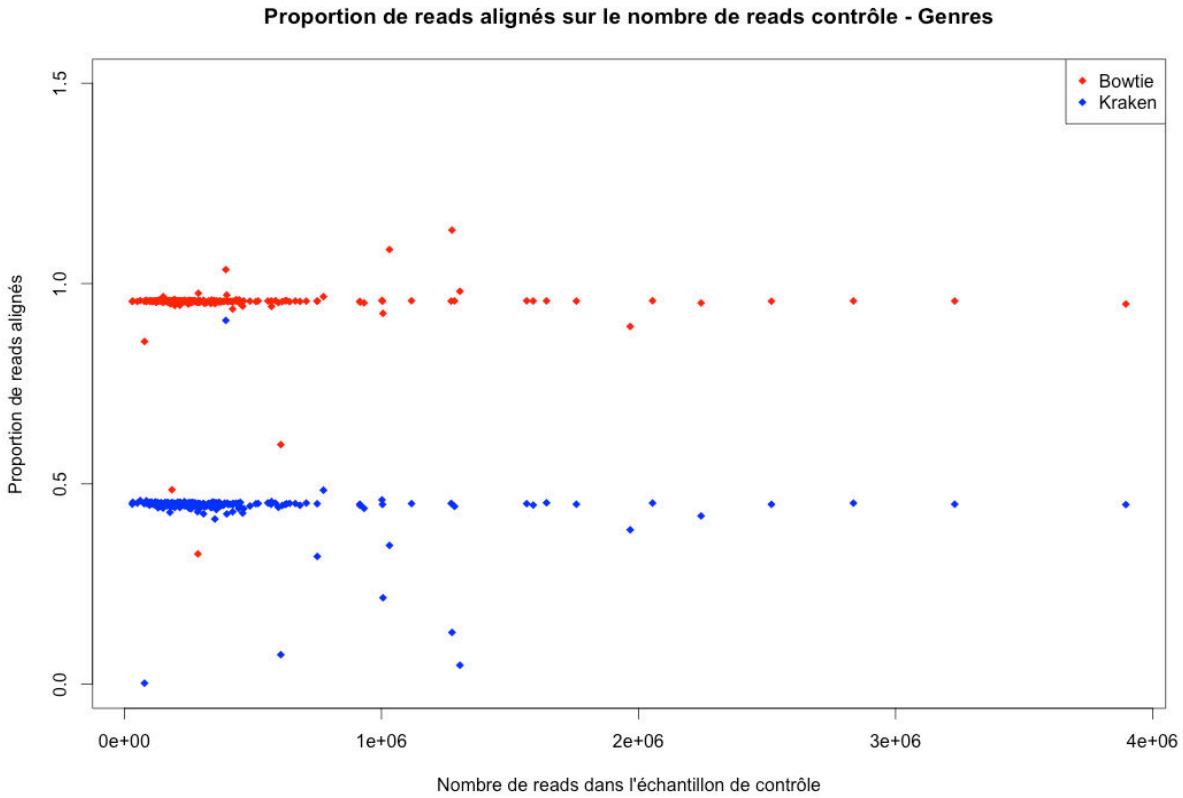


Figure 11. Proportion de reads assignés par Bowtie et Kraken aux genres identifiés par rapport au nombre de reads contrôle. Chaque point représente un genre.

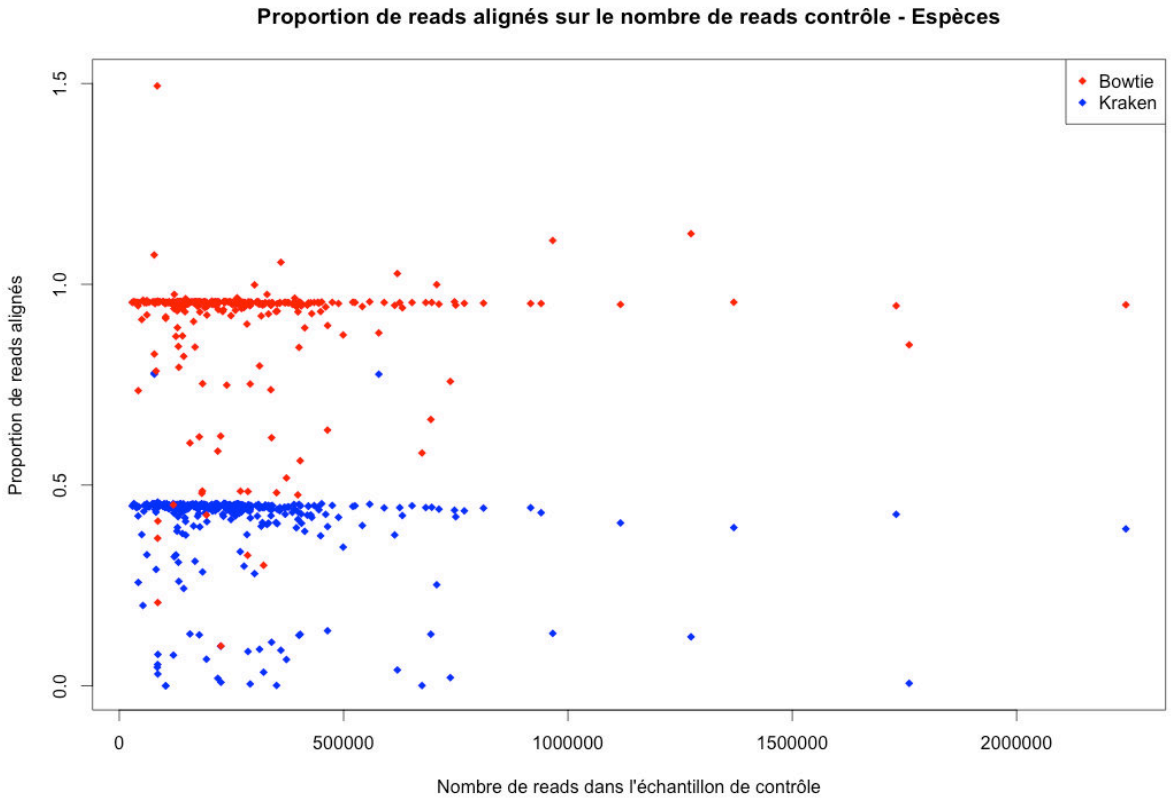


Figure 12. Proportion de reads assignés par Bowtie et Kraken aux espèces qu'ils ont chacun identifié par rapport au nombre de reads contrôle. Chaque point représente une espèce.

En comparant les deux outils, je me suis rendue compte que Kraken (en bleu sur les Figures 11 et 12) a comme une sorte de limite à 46% avec seulement 3 points en dehors (Fig. 11 et 12). Les trois étant à 77%. Pour Bowtie, en rouge sur les mêmes figures, le seuil serait plutôt à 96% avec 10 points au dessus dont un à 149% (Fig. 11 et 12).
En fait, sur un profil similaire, ie mêmes genres ou espèces identifié(e)s, la différence est dans le fait que Kraken assigne deux fois moins de reads à chaque genre ou espèce (Fig. 11 et 12). Cependant, il y a quelques genres et espèces qui ne suivent pas cette tendance : 7 genres sur les 215 présents et 49 espèces sur les 330 présentes affichent un nombre de reads assignés par Bowtie supérieur à deux fois et demi le nombre de reads assignés par Kraken. En fait, le nombre plus faible de reads attribués à chacun de ces cas par Kraken s'explique par sa méthode de classification qui assigne au rang taxonomique supérieur un read attribuable à plusieurs bactéries. Cela va donc forcément se répercuter sur l'estimation de l'abondance des bactéries. Plus on se rapproche de la lignée dans la taxonomie des bactéries à identifier, plus la méthode de classification de Kraken entraine une fausse sous-estimation de l'abondance.
Cela nous a été confirmé par les figures suivantes (Fig. 13 et 14) qui présentent, pour Kraken, un R^2 de 0,66 pour les espèces alors qu'il est à 0,95 pour les genres. Bowtie est largement au-dessus avec un R^2 de 0,97 pour les espèces et 0,99 pour les genres (Fig. 13 et 14).
Le fait que Kraken assigne deux fois moins de reads ne modifie pas l'abondance des bactéries entre elles.

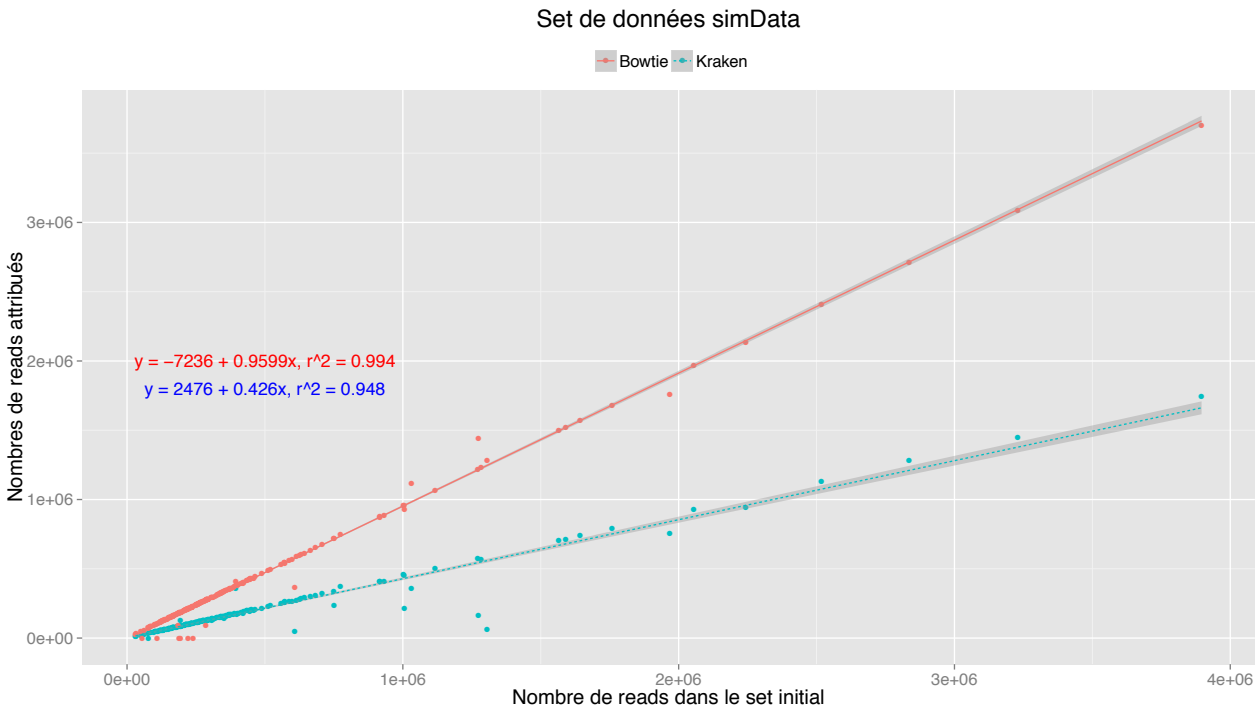


Figure 13. Corrélation entre le nombre de reads par genre présents dans le set initial et le nombre de reads attribués au même genre par Bowtie et Kraken

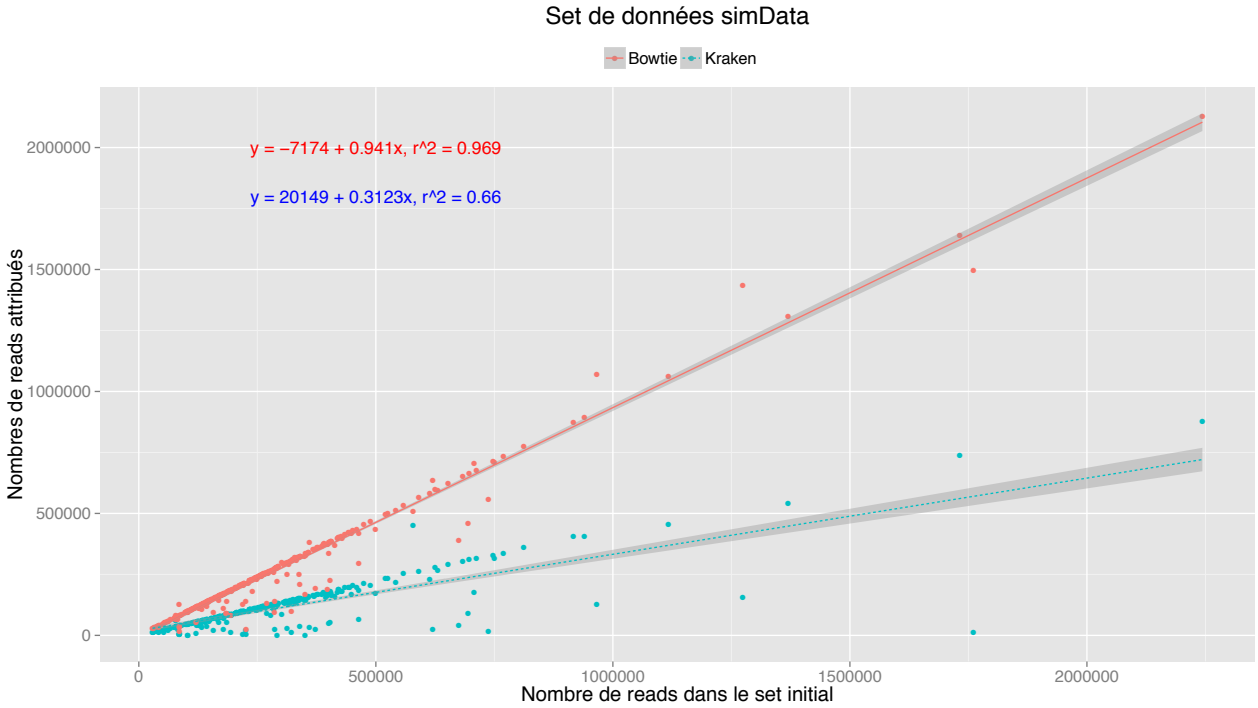


Figure 14. Corrélation entre le nombre de reads par espèce présentes dans le set initial et le nombre de reads attribués à la même espèce par Bowtie et Kraken

J'ai également réalisé une régression linéaire robuste sur les genres identifiés ce qui a fournit des r^2 de 1 pour Bowtie et Kraken. Ils sont donc aussi robuste l'un que l'autre après avoir retirer les valeurs « outliers ».

2.2.1.3. Conclusion

Par les tests précédents, nous avons vu que

(1) Bowtie et Kraken sont aussi sensibles l'un que l'autre pour détecter des bactéries réellement présentes.

(2) cependant, Kraken est plus spécifique que Bowtie en identifiant moins de bactéries non présentes dans l'échantillon initial. Dans le cas d'espèces proches à l'espèce présente, Bowtie aligne les reads des portions communes sur l'ensemble de ces espèces, augmentant ainsi le nombre de FP. L'application d'un nombre de reads minimal à garder permet de diminuer fortement ce nombre de bactéries FP chez Kraken et plus faiblement chez Bowtie.

(3) Bowtie est plus apte que Kraken à estimer l'abondance des bactéries identifiées et réellement présentes. L'exactitude de l'estimation diminue avec l'augmentation de la précision du rang taxonomique de la bactérie pour les deux outils.

Finalement, nous avons donc choisi plutôt Kraken que Bowtie pour l'étape de la classification des reads pour plusieurs raisons :

- sa méthode de classification est plus solide, surtout lorsque l'on se rapproche des espèces ou lignées,
- il attribue les reads lorsqu'il est sûr de leur provenance, sinon il les attribue au rang taxonomique supérieur,
- il identifie moins de bactéries FP,
- en se basant sur le nombre de reads assigné, les bactéries VP sont facilement identifiables par leur nombre plus important de reads attribués,
- les bactéries FP peuvent être en grande partie éliminées par l'application d'une valeur seuil tel que le nombre de reads minimal attribué à chaque bactérie,
- il fournit une estimation acceptable de l'abondance des bactéries dans un échantillon, sachant que cette estimation baisse avec l'augmentation de la précision du rang taxonomique.

Malgré le fait que Bowtie se soit montré très rapide sur les trois sets, les analyses effectuées par Kraken sont tout à fait réalisables en quelques heures.
De plus, Kraken fournit les profils taxonomiques jusqu'aux lignées dans le fichier de résultats. Un autre avantage de Kraken est la disponibilité de scripts dans le package permettant de rendre le fichier de résultats plus lisible, de le filtrer, de le convertir en un autre format, tout ceci afin de faciliter l'interprétation des résultats. Pour Bowtie, j'ai moi-même écrit les scripts pour interpréter et filtrer les résultats à partir du fichier SAM.
Concernant les bases de données, Kraken a sa propre mini base téléchargeable sur son site et contenant les bactéries, virus et archées. Elle est facilement mise à jour. Il est aussi possible de créer sa propre base de données. Pour Bowtie, c'est plus compliqué. Il faut d'abord télécharger sur NCBI RefSeq les génomes des organismes souhaités, les regrouper en un fichier fasta et en créer l'index pour Bowtie. C'est considérablement plus long à faire. Et il faut savoir que les bases de données NCBI sont mises à jour régulièrement.

Nous avons donc décidé d'intégrer Kraken en outil principal de classification des reads dans notre nouveau pipeline. Pour lister les bactéries ou tout autre microorganisme réellement présent, nous sélectionnerons les genres qui auront un nombre minimal déterminé de reads assignés. Cette valeur seuil sera à établir selon la quantité de reads assignés à chaque genre identifié. Il n'est pas possible de désigner une valeur fixe.

Cependant, en raison du nombre important de reads non alignés par Kraken et du cas où les espèces ne seraient pas dans la base de données, il a été décidé d'effectuer deux étapes parallèles avec le fichier de sortie contenant les reads non classés : une avec assemblage et une avec alignement. Le nouveau pipeline est présenté dans le paragraphe suivant.

2.3. Nouveau pipeline de détection des micro-organismes

Mon pipeline regroupe plusieurs outils qui sont gérés automatiquement par un script Perl et un fichier de configuration. Son nom est ICoMiO pour Identification of Contaminant MicroOrganisms (Fig. 15). Il comprend plusieurs étapes détaillées ci-dessous.

2.3.1. Description

La première étape est la vérification de la qualité des données de séquençage. Deux logiciels sont disponible : FastQC et SolexaQA. Ils donnent tous les deux des informations statistiques et graphiques sur la qualité des reads. FastQC est par défault. Il est aussi possible de passer cette étape, il suffit de le signaler dans le fichier de configuration.

La deuxième étape (optionnelle) est la suppression des reads appartenant à une ou plusieurs espèces présentes dans l'échantillon. Cette étape est importante, en particulier dans le cas d'échantillons avec une large proportion de génomes connus. Le fait de les retirer réduit la taille du fichier à analyser donc le temps d'analyse du pipeline. De plus, cela facilite les étapes de la classification et de l'assemblage. Pour ôter les reads d'une ou plusieurs espèces connues, on créé un fichier contenant le nom de ces espèces. Le nom de ce fichier sera ensuite mis dans le fichier de configuration pour être pris en compte. Pour enlever les reads, j'utilise Bowtie. Il est rapide et efficace. Le nombre d'erreurs autorisées peut être modifié via le fichier de configuration et une de ses options permet la création d'un fichier contenant les reads non alignés. A l'issue de cette deuxième étape, nous obtenons aussi des fichiers d'alignement SAM sur les organismes qui ont été retirés. Cela nous permet d'effectuer si besoin des analyses complémentaires sur les génomes du moustique et du parasite telles que des études d'expression des gènes. Cette étape peut être passée si l'échantillon ne contient qu'une communauté microbienne ou si l'on ne souhaite pas retirer les reads.

La troisième étape est celle de la classification des reads. Kraken est donc l'outil sélectionné. Il n'y a pas d'options à définir, Kraken gère seul la classification puis mon pipeline fait tourner les scripts de post-classification proposés par Kraken. Les fichiers produits sont un fichier .output contenant les résultats, un fichier .report contenant les résultats sous forme plus compréhensible avec pourcentage et nombre de reads alignés, noms et rangs taxonomiques, un fichier .labels avec pour chaque identifiant de read classé le nom et la taxonomie de l'espèce sur laquelle il s'est aligné, et un fichier .mpa.report (format MeatPhlAn) pour réaliser ensuite un graphe Krona. Les reads non alignés sont aussi sortis dans un fichier .fq qui sert de base pour les étapes suivantes.

Les étapes 4/5 et 6 permettent de compléter les résultats de Kraken dans les cas où des génomes sont absents de sa bases de données.

La quatrième étape est l'assemblage des reads non classés par Kraken. Pour ce faire, j'utilise Velvet. Le choix de la taille du k-mer est à faire dans le fichier de configuration ainsi que le style de reads (petits, longs, appairés, etc.). Un dossier nommé Velvet_Assembly_k(k-mer) contiendra tous les fichiers sortis de Velvet dont les contigs.

La cinquième étape est le BLAST des contigs sur les bases de données RefSeq téléchargées du NCBI. La recherche se fait en megablast. 4 fichiers tabulés commençant par Contigs_ contiennent les résultats de BLAST pour les bactéries, les virus, les phages et les champignons. Des scripts que j'ai écrit en Perl vont filtrer et synthétiser les résultats par espèce et par lignée.

La sixième étape est l'alignement par Bowtie des reads non classés par Kraken. Le taux d'erreurs de l'alignement est définissable dans le fichier de configuration. Les bases de données sont les quatre mêmes que pour le BLAST. Les fichiers de sortie sont sous format SAM. J'ai aussi écrit des scripts Perl pour récapituler les résultats au niveau du genre, de l'espèce et de la lignée. Trois fichiers tabulés seront produits pour chaque fichier SAM.

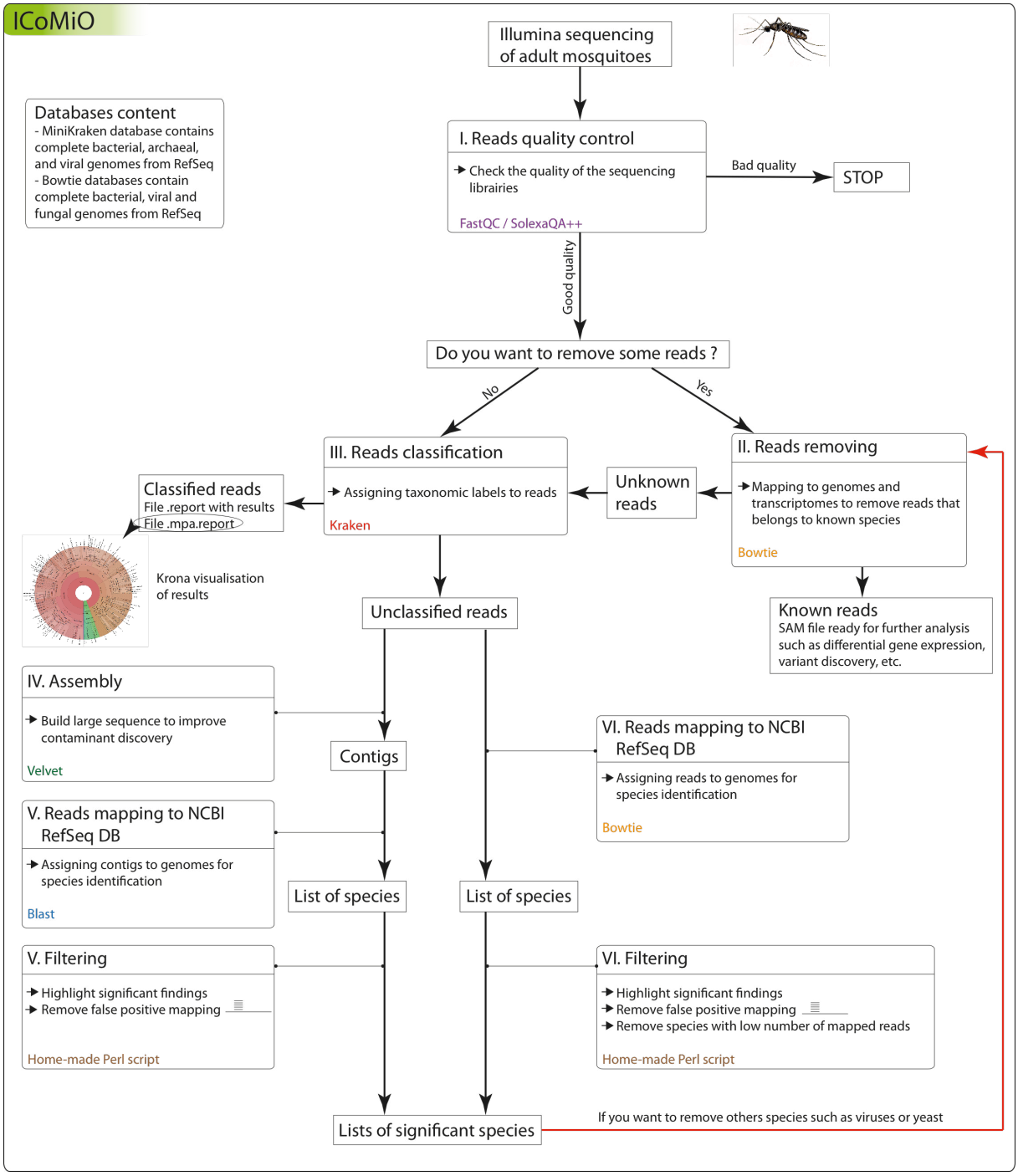


Figure 15. Données, méthodes et outils du nouveau pipeline ICoMiO

Le fichier de configuration rassemble toutes les informations nécessaires au bon déroulement du pipeline. Pour les étapes qui sont optionnelles, le choix de les effectuer ou non est à définir dans ce fichier.
Les champs à remplir sont :

- Informations générales
    - Type de séquençage : gDNA ou mRNA
    - Nom du fichier
    - Nom du second fichier dans le cas de reads paired-end
    - Nombre de cœurs à utiliser pour faire tourner les outils
    - Chemin complet de la racine jusqu'au dossier contenant les génomes de référence
    - les reads sont en paired-end : yes / no

- Vérification de la qualité
    - Contrôle de la qualité des reads : yes / no
    - Choisir l'outil pour le contrôle : fastqc / solexaqa

- Création des index des bases de données
    - Construction des index : yes / no
    - Chemin complet de la racine jusqu'au dossier qui contiendra les index des génomes de référence

- Alignement pour ôter les reads des organismes connus
    - Retirer les reads des organismes connus : yes / no
    - Nom du fichier contenant les noms des organismes à retirer
    - Choix des options d'alignement : -n 2 -l 28 -e 70 / -v 0 / -v 1 / -v 2 / -v 3 (il est possible de sélectionner ses propres valeurs de n, l et e, voir le manuel de Bowtie pour en savoir plus)

- Assemblage
    - Catégorie des reads : -short / -shortPaired / -short2 / -shortPaired2 / -long / -longPaired
    - Taille des k-mers : soit on rentre le k choisi, soit on sélectionne un minimum, un maximum et un saut comme ceci : m,M,s

- Alignement pour la classification des reads
    - Choix des options d'alignement : -n 2 -l 28 -e 70 / -v 0 / -v 1 / -v 2 / -v 3 (il est possible de sélectionner ses propres valeurs de n, l et e, voir le manuel de Bowtie pour en savoir plus)

Tout en haut du fichier de configuration, j'ai ajouté une incrémentation qui démarre à 1. En fait, tous les fichiers issus de ce lancement de ICoMiO seront regroupés dans le dossier run_1. Si l'on souhaite relancer un nouvel ICoMiO pour le même set de données mais en changeant des paramètres tels que les options d'alignements ou la taille du k-mer, c'est possible en créant un nouveau fichier de configuration avec les nouvelles options et avec l'incrémentation à 2.

Par ailleurs, en plus des fichiers de résultats, un dossier nommé Logs est créé durant le fonctionnement de ICoMiO. Il regroupe les fichiers de sortie écran de chaque étape. Il est très utile en cas de problème avec un outil.

Pour une utilisation optimale, il est conseillé lors de l'installation des outils nécessaires à son fonctionnement d'ajouter dans le fichier bash profile le chemin complet de la racine aux exécutables de chaque outil.

2.3.2. Fonctionnement de ICoMiO sur le set de données simulées

Nous avons testé ICoMiO sur un fichier nommé simData.fq contenant des reads simulés d'*Anopheles gambiae*, de *Plasmodium berghei*, de *Saccharomyces cerevisiae*, de virus et de bactéries.
Quelques préparations sont nécessaires avant de lancer le script. Tout d'abord, il faut créer le fichier de configuration qui regroupe quelques critères du fichier d'input, le choix des étapes à réaliser et les différentes options des outils utilisés (Fig. 16).

```
##################################
##      ICoMiO config file      ##
##################################

iteration                   : 1 [increment each time you run ICoMiO for the same dataset]

##############################
##      Data infos          ##
##############################
sample_type                 : gDNA [gDNA, smallRNA]
sample_file_name_1          : simData.fq
sample_file_name_2          : none [put 'none' in case of single file]
nb_threads                  : 16
path_to_DB                  : /home/ibmc/vittu/Genomes_Indexes
paired_end                  : no [yes/no]


##############################
##      Quality             ##
##############################
check_quality               : yes [yes/no]
quality                     : fastqc [fastqc, solexaqa]


##############################
##      Indexes             ##
##############################
build_indexes               : no [yes/no]
indexes_directory           : /home/ibmc/vittu/Genomes_Indexes


##############################
##      Removing            ##
##############################
orgaToRemove_file           : listOrgaToRemove.txt
removing_action             : yes [yes/no]
bowtie_removing_settings    : -n 2 -l 28 -e 70 [-n 2 -l 28 -e 70, -v 0, -v 1, -v 2, -v 3]


##############################
##      Velvet              ##
##############################
read_category               : -short [-short, -shortPaired, -short2, -shortPaired2, -long, -longPaired]
kmer                        : 61 [m,M,s] #k-mer lengths such that m ≤ k < M with a step of s

##############################
##      Bowtie              ##
##############################
bowtie_mapping_settings     : -n 2 -l 28 -e 70 [-n 2 -l 28 -e 70, -v 0, -v 1, -v 2, -v 3]
```

Figure 16. Fichier de configuration ICoMiO_simData.txt pour le fichier simData.fq

Ce fichier doit être dans le même dossier que le fichier contenant les reads, ici simData.fq.

Ensuite, comme nous voulons retirer les reads du moustique et du parasite, nous créons un fichier listOrgaToRemove.txt contenant le nom des index des organismes à enlever. Le notre contient « AgamP3-Haplotypes » et « Pberghei ». Ce fichier doit aussi être avec le fichier simData.fq.

Voici donc le contenu de notre dossier :

- ICoMiO_simData.cfg
- listOrgaToRemove.txt
- simData.fq

Nous pouvons maintenant lancer notre script dans un terminal avec la commande suivante :
$ ICoMiO.pl ICoMiO_simData.cfg

Le terminal affiche le déroulement des étapes tel que présenté sur l'image suivante (Fig. 17).

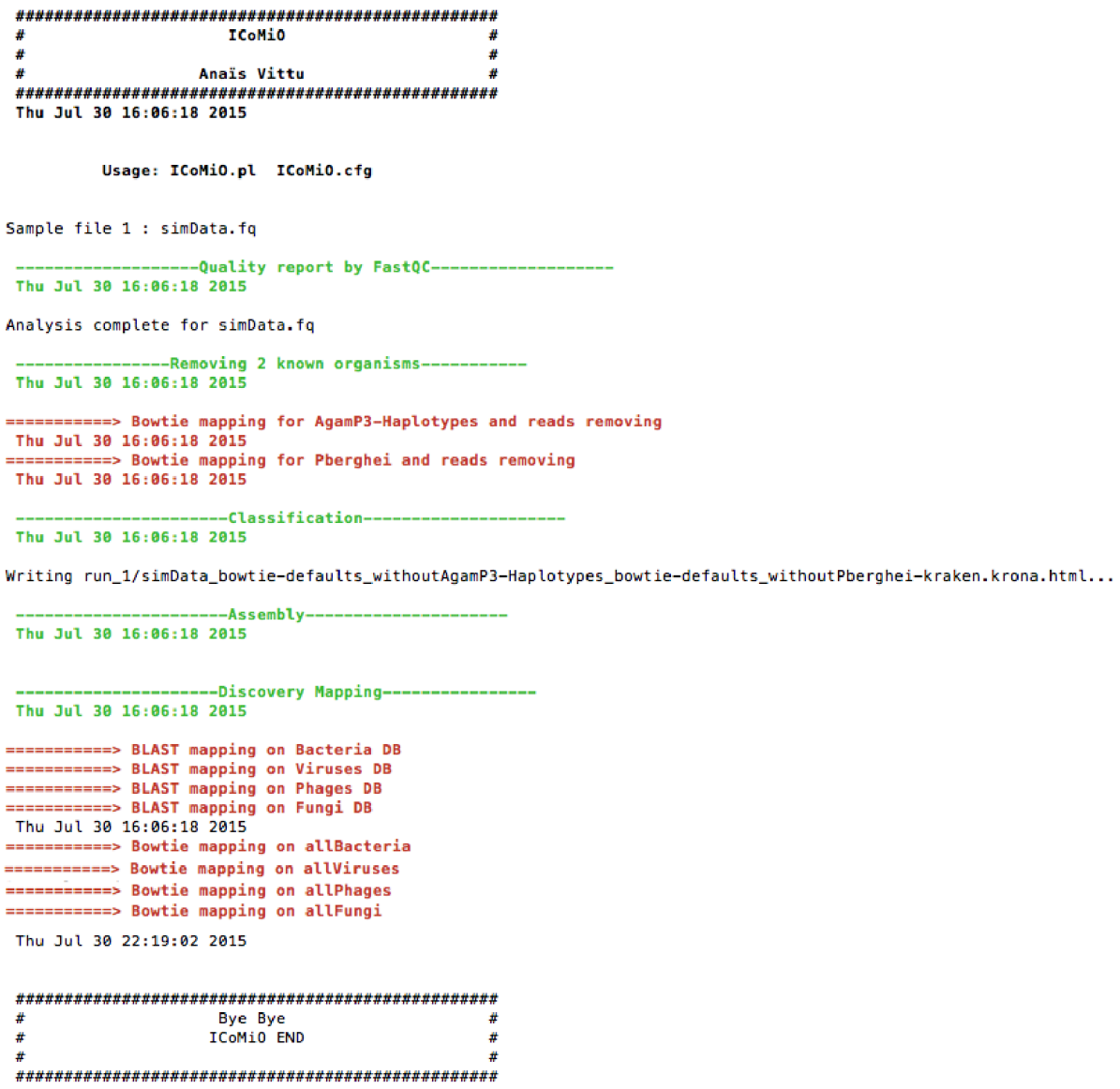
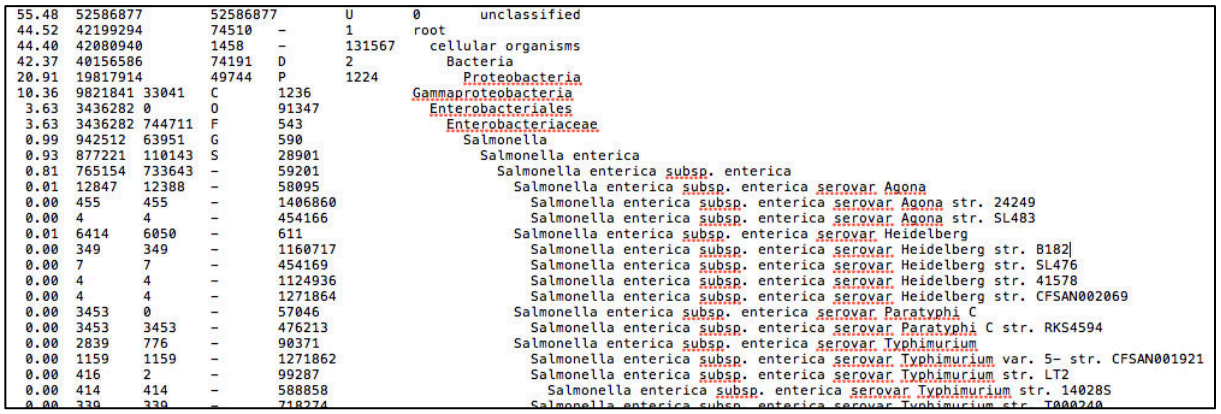


Figure 17. Déroulement des étapes d'ICoMiO affiché dans le terminal

Tous les fichiers de résultats commencent par le nom du fichier initial (Annexe VIII). On retrouve donc le rapport de FastQC, les résultats de la classification par Kraken, ceux du BLAST des contigs et ceux de l'alignement de Bowtie, les fichiers SAM des alignements de Bowtie sur les organismes enlevés et les fichiers .fq contenant les reads restants après avoir ôté les organismes.

1) étape de la classification par Kraken

Kraken a classé 44,52% des 94 786 171 reads qu'il restait après avoir retiré ceux d'*A. gambiae* et ceux de *P. berghei*.
Ce dernier fournit un fichier .report contenant les résultats sous forme de texte (Fig. 18).



```
 55.48  52586877        52586877        U       0       unclassified
 44.52  42199294        74510   -       1       root
 44.40  42080940        1458    -       131567    cellular organisms
 42.37  40156586        74191   D       2           Bacteria
 20.91  19817914        49744   P       1224          Proteobacteria
 10.36  9821841 33041   C       1236              Gammaproteobacteria
  3.63  3436282 0       O       91347               Enterobacteriales
  3.63  3436282 744711  F       543                   Enterobacteriaceae
  0.99  942512  63951   G       590                     Salmonella
  0.93  877221  110143  S       28901                     Salmonella enterica
  0.81  765154  733643  -       59201                       Salmonella enterica subsp. enterica
  0.01  12847   12388   -       58095                         Salmonella enterica subsp. enterica serovar Agona
  0.00  455     455     -       1406860                         Salmonella enterica subsp. enterica serovar Agona str. 24249
  0.00  4       4       -       454166                          Salmonella enterica subsp. enterica serovar Agona str. SL483
  0.01  6414    6050    -       611                           Salmonella enterica subsp. enterica serovar Heidelberg
  0.00  349     349     -       1160717                         Salmonella enterica subsp. enterica serovar Heidelberg str. B182
  0.00  7       7       -       454169                          Salmonella enterica subsp. enterica serovar Heidelberg str. SL476
  0.00  4       4       -       1124936                         Salmonella enterica subsp. enterica serovar Heidelberg str. 41578
  0.00  4       4       -       1271864                         Salmonella enterica subsp. enterica serovar Heidelberg str. CFSAN002069
  0.00  3453    0       -       57046                         Salmonella enterica subsp. enterica serovar Paratyphi C
  0.00  3453    3453    -       476213                          Salmonella enterica subsp. enterica serovar Paratyphi C str. RKS4594
  0.00  2839    776     -       90371                         Salmonella enterica subsp. enterica serovar Typhimurium
  0.00  1159    1159    -       1271862                         Salmonella enterica subsp. enterica serovar Typhimurium var. 5- str. CFSAN001921
  0.00  416     2       -       99287                           Salmonella enterica subsp. enterica serovar Typhimurium str. LT2
  0.00  414     414     -       588858                            Salmonella enterica subsp. enterica serovar Typhimurium str. 14028S
  0.00  339     339     -       718274                          Salmonella enterica subsp. enterica serovar Typhimurium str. T000240
```

Figure 18. Début du fichier .report de Kraken. Il y a six colonnes représentant (1) le pourcentage de reads couverts par le clade pour ce taxon, (2) le nombre de reads couverts par le clade pour ce taxon, (3) le nombre de reads assignés directement au taxon, (4) un code indiquant (U)nclassified, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, ou (S)pecies, les autres rang sont indiqués par '-', (5) l'identifiant taxonomique du NCBI et (6) le nom scientifique du rang indenté.

A partir des fichiers de résultats de Kraken, il est également fourni un graphe interactif visualisable dans un navigateur (Fig. 19). On peut aisément voir les organismes majoritaires trouvés et parcourir jusqu'au rang de l'espèce, sélectionner seulement les virus, par exemple, prendre une photo svg du graphe ou rechercher un nom précis. Plusieurs paramètres sont modifiables en haut à gauche de la fenêtre: la taille du graphe, la taille de caractère, la profondeur taxonomique à représenter.

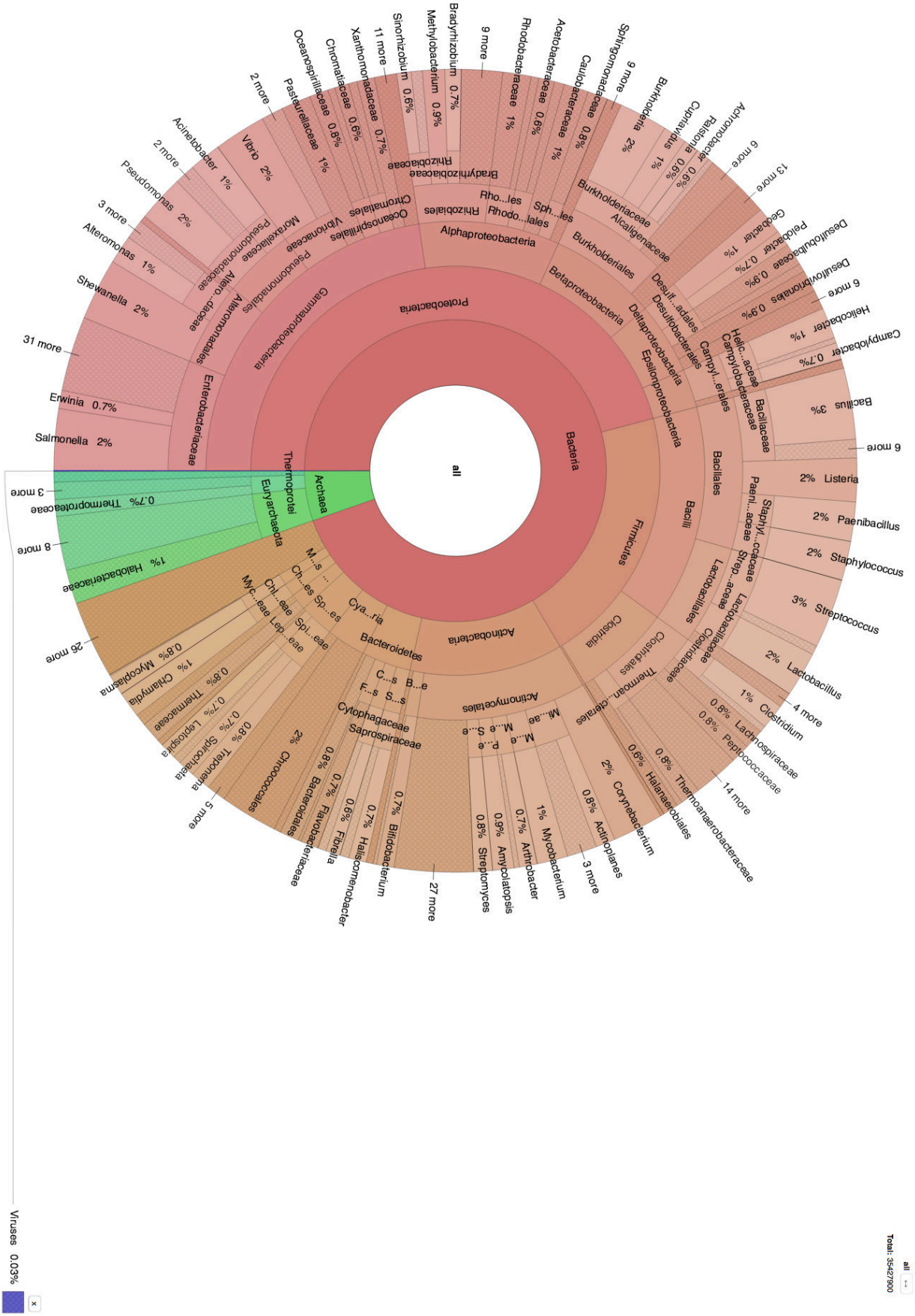


Figure 19. Graphe Krona de la composition de notre échantillon simData.

Pour les bactéries et les archées, Kraken a assigné des reads sur 675 genres. Pour déterminer la liste des bactéries réellement présentes, nous avons examiné les nombres de reads attribués aux genres identifiés. Pour ce faire, le contenu du fichier .report de Kraken a été copié dans un fichier Excel. Les genres identifiés ont au minimum 1 read et au maximum 1 746 594 reads avec une moyenne de 60 324 reads. De plus, en classant le nombre de reads par genre par ordre croissant, il a été observé un gap assez conséquent de 12 000 reads entre *Citrobacter* à 994 reads et *Candidatus Sulcia* à 12 982 reads. Il a donc été choisi que la valeur filtre soit à 10 000 reads minimum pour éliminer le plus possible de FP. Nous avons obtenu 215 genres, qui, après vérification sont tous des VP. L'application de ce filtre a permis d'identifier toutes nos bactéries réellement présentes tout en éliminant tous les FP.
Sur l'ensemble des 675 genres initialement identifiés, il a été trouvé 1 430 espèces de bactéries avec un nombre de reads attribués allant de 1 à 877 221 et une moyenne de 24 766. En gardant uniquement les espèces provenant des 215 genres identifiés qui ont passé le filtre, nous en avons 832. Elles ont au minimum 1 read à maximum 877 221 reads ave une moyenne de 42 510. Parmi elles, nous retrouvons nos 330 espèces de bactéries réellement présentes. Il n'y a pas de rupture brute du nombre de reads assignés à chaque espèce. Il est donc difficile d'établir une valeur filtre pour éliminer quelques FP, surtout lorsque certaines espèces VP, 4 exactement, ont eu un nombre de reads assignés par Kraken inférieur à 1 000 (24, 32, 418 et 678 reads). Si nous avions décidé arbitrairement de poser le filtre à 1 000 reads minimum, nous aurions garder 330 espèces dont 4 FP et perdu les 4 VP précédents. La moyenne des 502 espèces éliminées après ce filtre ont une moyenne de 90 reads assignés contre une moyenne de 107 040 reads assignés pour les 330 espèces passant le filtre des 1 000 reads minimum. Le fait de calculer la moyenne des reads attribués aux espèces en-dessous et au-dessus d'une valeur filtre peut aider au choix de cette valeur.

Pour les virus, Kraken a identifié 35 genres et 129 espèces. Dans le fichier .report de Kraken, le nombre de reads affecté à chaque genre est assez bas par rapport aux bactéries. Il varie de 1 read minimum à 37 789 reads maximum avec une moyenne à 1 153. Il y a une brutale rupture entre le premier et le deuxième genre, passant de 37 789 reads à 1 562 reads. Il n'est clairement pas possible d'établir une valeur seuil à 10 000, nous garderions que le premier résultat et nous savons qu'il y a 6 virus présents. En plaçant la valeur filtre à 100, nous éliminons 30 genres avec une moyenne de 7 reads et en gardons 5 pour une moyenne de 589 reads. Si on met la valeur filtre à 10 reads minimum, nous éliminons 26 genres avec une moyenne de 2 reads et en gardons 8 avec une moyenne de 314 reads. Je pense qu'ici le choix de la valeur seuil va dépendre du choix entre (1) sacrifier quelques VP pour avoir une liste dont on peut être sûr ou (2) essayer d'avoir tous les VP et prendre le risque de garder des FP. Dans notre cas, nous avons choisi la première option et avons décidé d'appliquer le filtre à 100 reads minimum. Nous avons donc obtenu 5 genres. Cependant, il manque un VP puisque notre set compte 6 genres de virus. En fait, ce genre ne possède que 90 reads dans le set simulé ce qui fait que Kraken ne lui a assigné que 48 reads. Les 32 espèces découlant des 5 genres identifiés possèdent de 1 à 9 811 reads avec une moyenne à 356 reads. Nous avons placé notre valeur filtre à 100 reads minimum pour la même raison que précédemment et avons obtenu 5 espèces qui sont toutes des VP.

Pour conclure sur cette première étape, nous avons bien réussi à identifier les bactéries et virus effectivement présents dans le set simulé initial. Cependant, la présence de *Saccharomyces cerevisiae* n'a pas été détectée. Cela s'explique par son absence dans la base de données de Kraken. Nous allons voir si les étapes suivantes ont été capables de compléter ces premiers résultats.

2) étape de l'assemblage par Velvet

Les 55,48% des reads non classés par Kraken ont été assemblés par Velvet. Nous avons obtenu 2 contigs de 122 et 149 paires de bases (pb). Après un BLAST, ils proviennent de deux bactéries, respectivement, *Coxiella burnetii* et *Burkholderia pseudomallei* qui sont bien présentes dans l'échantillon de départ. Elles avaient déjà été identifiées par Kraken.
Cette étape n'a pas révélé la présence de *S. cerevisiae*.

3) étape du deuxième alignement par Bowtie

L'alignement des reads non classés par Kraken a identifié :
- pour les bactéries : 624 genres et 1 392 espèces,
- pour les virus : 11 genres et 15 espèces et
- pour les champignons : 19 genres et 23 espèces.

Pour les bactéries, l'application du filtre à 10 000 reads minimum a permis de garder 218 genres (dont les 215 effectivement présents) et celui à 1 000 reads minimum de garder 458 espèces (dont les 330 bien présentes).
Pour les virus, le filtre à 100 reads minimum permet de garder 5 espèces avec 1 FP et 1 FN.
Concernant les phages, Bowtie en détecte 252 génomes. Un filtre à 1 000 reads minimum permet de réduire cette liste à 23 génomes qui contiennent l'unique VP. Tous les autres sont des phages de bactéries présentes dans l'échantillon. En fait, les reads qui s'alignent sur ces phages s'alignent aussi sur les génomes des bactéries correspondantes, surement parce que ces portions sont en fait intégrées dans le génome de la bactérie.
Enfin, Bowtie identifie 23 espèces de champignons dont 19 genres différents. Le VP qui correspond à *S. cerevisiae*, se démarque des autres par un nombre de reads à 461 000 alors que les FP sont entre 2 et 84 reads. Il est donc aisé de l'identifier parmi tous les résultats.

Conclusion

ICoMiO a donc identifié tous nos organismes que nous avions dans notre fichier de données simulées de 30 Go en 6 heures 12 minutes. Il a utilisé entre 180 et 256 Go de mémoire vive.
Dans ce cas, Velvet n'a pas révélé d'informations supplémentaires mais Bowtie a permis d'identifier *S. cerevisiae*, non identifié par Kraken. Il serait alors nécessaire d'ajouter dans la base de données de Kraken les champignons. L'application des filtres sur les résultats de Kraken et Bowtie a permis de confirmer leur utilisation nécessaire à la diminution de FP.

De plus, ICoMiO nous a fournit tous les fichiers de résultats nécessaires plus des fichiers temporaires comme les fichiers SAM issus des alignements pour retirer les reads d'organismes connus qui pourront servir dans les cas où l'on souhaite étudier les produits de séquençage de ces organismes.

3. Application de ICoMiO sur des données de séquençage

ICoMiO a été utilisé sur les données de séquençage d'ADNg de la lignée R1iso2 du chapitre de l'assemblage pour identifier les micro-organismes qu'il contient. Les données sont des reads PE de 104 pb répartis en deux fichiers de 36 Go. ICoMiO a mis 11h 39 min et 21s pour analyser ces données. C'est donc un temps relativement raisonnable pour obtenir une analyse métagénomique complète.

Tout d'abord, ICoMiO a enlevé 78 447 815 reads, soit 56,26% de l'échantillon, provenant du génome d'*Anopheles gambiae* et aucun read provenant du parasite *P. berghei*, ce qui est normal, ces moustiques n'étaient pas infectés.
Cela nous laisse un total de 60 984 783 reads inconnus.

1) Kraken

Kraken a ensuite classé seulement 0,52% des 60 984 783 reads (Fig. 20).

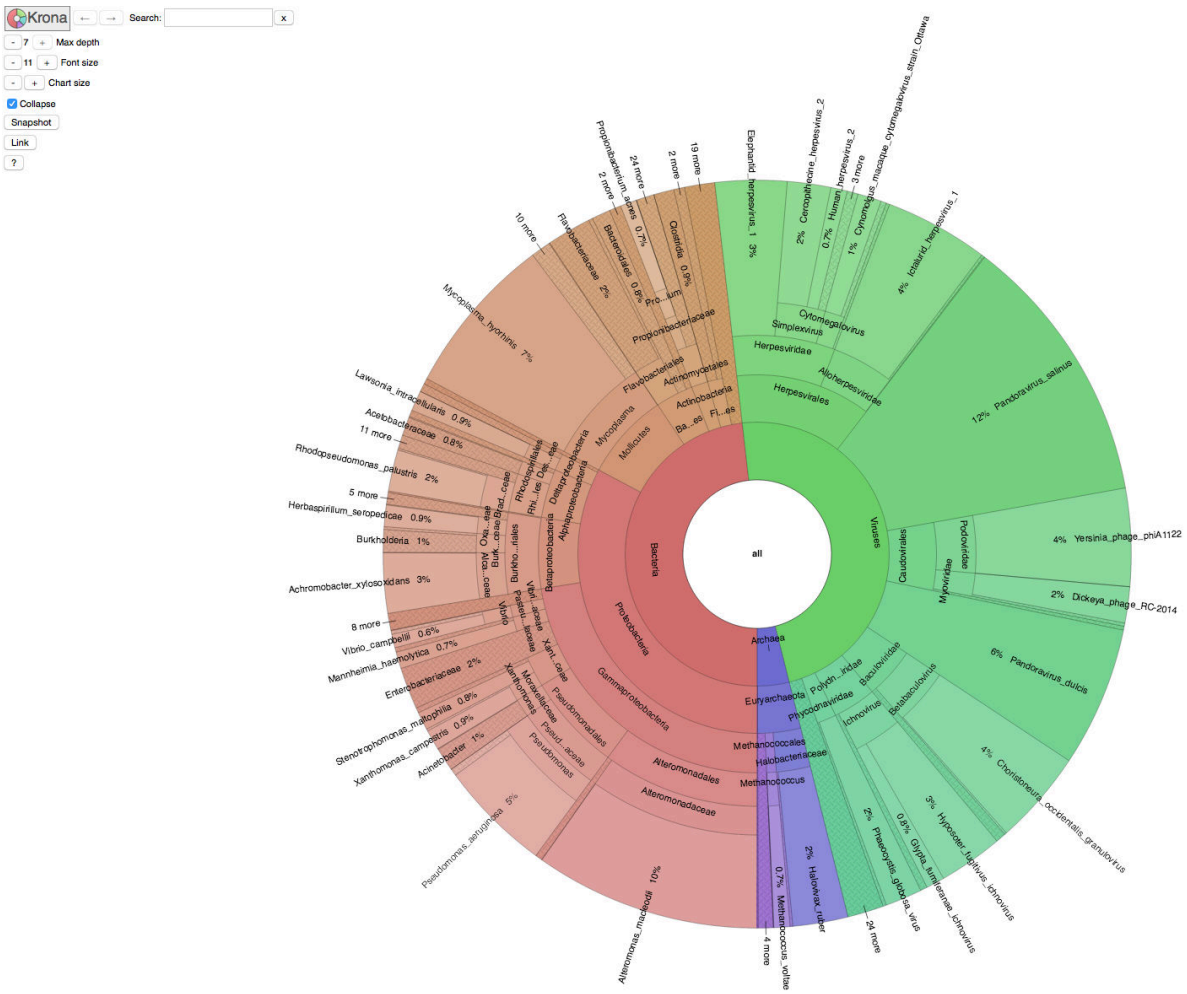


Figure 20. Graphe Krona des organismes identifiés par Kraken sur les données PE d'ADN génomique. Les bactéries sont de couleurs orange/rouge, les virus en vert et les archées en bleu.

Sur les 316 142 reads classés, il y a 48% de bactéries, 48% de virus et 4% d'archées.

a. Bactéries identifiées par Kraken

Dans le fichier .report de Kraken, nous avons 5 phylums qui sont présents à plus de 1% en nombre de reads assignés (Fig. 21). Il y a une majorité de Protéobactéries (33%), puis des Mollicutes (8%), des Bactéroïdes (3%), des Actinobactéries (1%) et des Firmicutes (1%). A part les Mollicutes, les quatre autres sont les phylum les plus représentés chez les moustiques (Boissière et al., 2012).

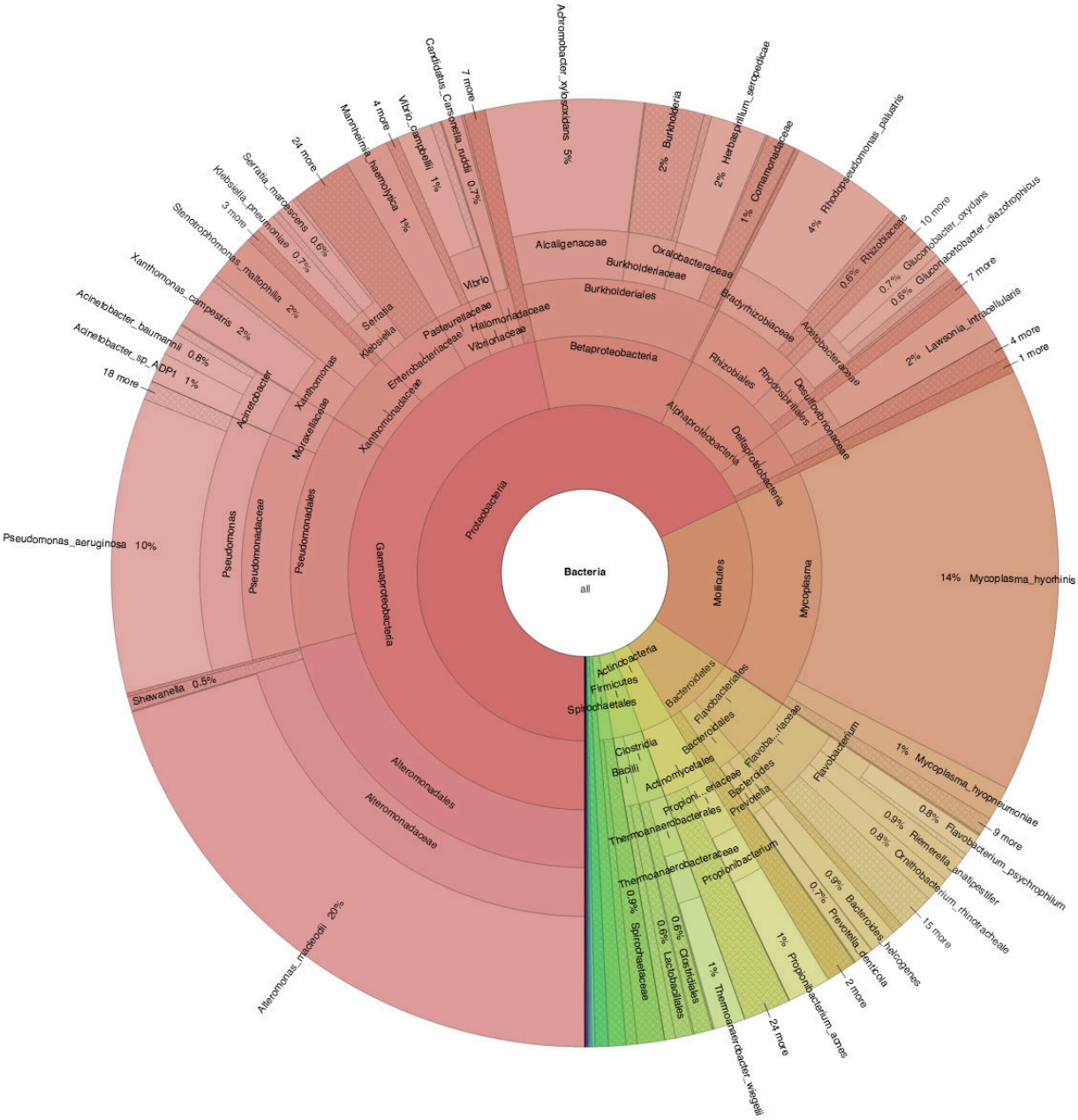


Figure 21. Graphe Krona des bactéries identifiées par Kraken dans le séquençage d'ADNg d'une lignée de moustique.

Kraken a identifié 341 genres. Ils ont au minimum 1 read, au maximum 11 507 reads et une moyenne de 172 reads. Il n'y a pas de rupture dans le nombre de reads assignés. En plaçant un filtre arbitraire à 100 reads minimum, nous gardons 60 genres ayant de 104 à 11 507 reads et une moyenne de 908 reads. En choisissant le filtre à 1 000 reads minimum, nous gardons 10 genres ayant de 1 036 à 11 507 reads et une moyenne de 3 923 reads. Nous avons décidé de garder les 10 genres à plus de 1 000 reads pour établir une liste de bactéries dont la probabilité qu'elles soient réellement présentes est la plus forte possible.
Ces 10 genres ont donné 67 espèces avec un minimum et un maximum identiques aux genres et une moyenne à 561 reads (Tableau 10).

Nombre de reads assignés au genre	Genre	Nombre de reads assignés à l'espèce majoritaire	Espèce	Gram
11 507	*Alteromonas*	11 507	*Alteromonas macleodii*	Négatif
9 751	*Mycoplasma*	8 230	*Mycoplasma hyorhinis*	Négatif
6 333	*Pseudomonas*	5 721	*Pseudomonas aeruginosa*	Négatif
3 060	*Achromobacter*	3 060	*Achromobacter xylosoxidans*	Négatif
2 107	*Rhodopseudomonas*	2 107	*Rhodopseudomonas palustris*	Négatif
1 784	*Burkholderia*	354	*Burkholderia cepacia complex*	Négatif
1 415	*Acinetobacter*	442	*Acinetobacter baumannii*	Négatif
1 147	*Xanthomonas*	1 027	*Xanthomonas campestris*	Négatif
1 094	*Herbaspirillum*	1 094	*Herbaspirillum seropedicae*	Négatif
1 036	*Lawsonia*	1 036	*Lawsonia intracellularis*	Négatif

Tableau 10. Bactéries identifiées par Kraken et ayant plus de 1 000 reads assignés au genre. Les espèces notées sont celles qui ont le plus de reads dans les 10 genres identifiés. Les bactéries sont classées par ordre décroissant du nombre de reads par genre. La dernière colonne présente le résultat de la coloration de Gram pour les espèces indiquées.

La majorité des bactéries identifiées appartiennent à la classe des Gamma-protéobactéries. L'espèce *Alteromonas macleodii* représente 20% de l'ensemble des bactéries identifiées dans nos moustiques. Cependant, elle n'a jamais été identifiée chez d'autres espèces ou même au sein d'*Anopheles gambiae*. Le genre *Pseudomonas* est, par contre, un habitué des lignées de moustiques (Boissière et al., 2012).
Les genres *Achromobacter*, *Rhodopseudomonas*, *Burkholderia*, *Acinetobacter, Xanthomonas* et *Herbaspirillum,* qu'ils soient plus ou moins représentés dans notre lignée, ont déjà été retrouvé chez d'autres lignées ou espèces de moustiques (Boissière et al., 2012; Cirimotich et al., 2011a; Minard et al., 2013; Osei-Poku et al., 2012; Rani et al., 2009; Valiente Moro et al., 2013). De plus, certains genres de bactéries trouvés dans ces mêmes études ont été identifiés par Kraken mais n'ont pas passé le filtre des 1 000 reads. Il y en a 22 sur les 50 genres qui ont quand même plus de 100 reads.

Les genres *Mycoplasma* et *Lawsonia* sont plutôt des bactéries colonisant les hommes, les bovins ou les porcins. Leur présence ici est surprenante mais peut-être expliquée par l'utilisation des souris pour nourrir les femelles moustiques. Ces bactéries pouvant les infecter aussi. Leur réelle existence devrait être vérifiée chez les moustiques et les souris utilisées.

Alors que Boissière et al. (2012) avait identifié chez leurs moustiques de laboratoire une Flavobactérie, *Elizabethkingia spp*., qui représentait 95% de leur échantillon, nous n'en avons aucune trace dans notre lignée. Cette espèce avait aussi été retrouvée précédemment chez d'autres moustiques de différents laboratoires (Dong et al., 2009; Kämpfer et al., 2011). Toutefois, les moustiques issus des milieux naturels ne présentent qu'un faible pourcentage de cette même espèce (Boissière et al., 2012).

Il se trouve que la majorité de nos bactéries, dont les plus présentes, sont des Gram négatif (Tableau 10). Le moustique est peut-être un environnement favorable à leur développement. De plus, il a été observé l'inhibition de la formation des oocystes par les bactéries Gram négatif (Cirimotich et al., 2011a). Nos moustiques séquencés venant d'une lignée résistante, il serait maintenant intéressant de voir si c'est le cas chez nos lignées sensibles et sur différents échantillons dans le temps.

b. Archées identifiées par Kraken

Kraken a attribué 14 224 reads aux archées. Il y a 25 genres qui ont de 1 à 2 811 reads avec une moyenne de 186 reads. Il y a une rupture assez forte de 1 900 reads entre le premier genre, *Halovivax* (2811 reads) et le deuxième genre, *Methanococcus* (910 reads). Nous avons donc gardé uniquement le premier résultat. Les 2 811 reads du genre ont tous été attribués à l'espèce *Halovivax ruber XH-70*. Le fait qu'elle soit gram négatif confirme que l'environnement du moustique est un terrain favorable à leur développement.
Les Halobacteriaceae sont des bactéries halophiles extrême, ie qu'elles nécessitent un environnement hypersalé pour vivre. Une espèce de cette famille a déjà été identifiée chez *Culex quinquefasciatus* (Reegan et al., 2013).

c. Virus identifiés par Kraken

Kraken a identifié 45 genres ayant au minimum 1 read et au maximum 21 122 reads et avec une moyenne de 1 200 reads. En examinant le nombre de reads assignés, nous avons remarqué une rupture entre deux genres : *Alphabaculovirus* à 527 reads et *Cytomegalovirus* à 1 487 reads. Nous avons décidé de garder les genres avec un nombre de reads supérieur ou égale à 1 000. Il reste donc 9 genres, ce qui permet d'obtenir 25 espèces. Elles ont entre 1 et 5 287 reads avec une petite rupture entre 940 et 1 206 reads. Cela nous a donc donné 11 espèces à plus de 1000 reads (Tableau 11).

Nombre de reads	Genre	Nom complet	Hôtes connus	Transmission
13 869	Pandoravirus	Pandoravirus salinus	Amibes	Via l'eau
7 139	Pandoravirus	Pandoravirus dulcis	Amibes	Via l'eau
5 287	Ictalurivirus	Ictalurid herpesvirus 1	Poissons	Via l'eau
4 982	T7likevirus	Yersinia phage phiA1122	Bactéries	Via le milieu
4 808	Betabaculovirus	Choristoneura occidentalis granulovirus	Arthropodes	Voie orale-fécale
3 651	Betaherpesvirinae	Elephantid herpesvirus 1	Vertébrés	Via des fluides corporels
3 271	Ichnovirus	Hyposoter fugitivus ichnovirus	Guêpes qui sont elles-même parasites des Lepidoptères	Via la descendance
2 212	Alphaherpesvirinae	Cercopithecine herpesvirus 2	Vertébrés	Via des fluides corporels
1 876	Prymnesiovirus	Phaeocystis globosa virus	Algues	Via l'eau
1 835	Non classé (Famille des Myoviridae)	Dickeya phage RC-2014	Bactéries	Via le milieu
1 206	Betaherpesvirinae	Cynomolgus macaque cytomegalovirus strain Ottawa	Vertébrés	Via des fluides corporels

Tableau 11. Virus identifiés par Kraken et ayant plus de 1 000 reads assignés à l'espèce. Les espèces notées sont celles dont le genre à passer le filtre. Les espèces sont classées par ordre décroissant du nombre de reads. Les deux dernières colonnes présentent leurs hôtes connus et leur moyen de transmission.

Tous les virus identifiés ont un ADN double brin, ce qui est assez logique puisque le séquençage a été effectué sur une préparation d'ADN génomique. Les virus à ARN ne sont pas détectables.
Parmi les virus identifiés, 3 genres dont 2 sont les plus représentés (*Pandoravirus*, *Ictalurivirus* et *Prymnesiovirus*) se transmettent dans les milieux aquatiques. Leur présence peut s'expliquer par le fait que les premiers stades du développement du moustique se passent dans l'eau. Nous avons aussi 2 genres parasitant des bactéries reflétant la possibilité de trouver ces bactéries dans notre échantillon. Après vérification, les bactéries *Yersinia* et *Dickeya* ont été identifiées par Kraken mais elles ont seulement 59 et 38 reads, respectivement.
Les deux virus colonisant les vertébrés sont en fait des virus simiens et il est très étonnant de les trouver ici. De même pour ceux colonisant les arthropodes. Il se pourrait qu'ils soient de faux positifs.

2) Velvet et BLAST

Avec les reads restants, Velvet a construit 170 402 contigs ayant un spectre de taille de 121 à 38 450 bases. Leur taille moyenne se situe à 797 bases et nous avons 23% des contigs qui ont une taille supérieure à 1 000 bases.
Dans notre cas, le BLAST des contigs n'a pas permis d'identifier de nouvelles espèces de bactéries ou de virus après Kraken. Cependant, nous avons pu identifier avec certitude la présence de *Saccharomyces cerevisiae* avec 7 768 contigs de taille moyenne de 1 318 bases. La présence de cette levure est probable dans notre échantillon puisque les premiers stades larvaires de ces moustiques avaient été nourri avec de la levure et qu'il y a pu avoir une colonisation à ce moment-là.

3) Bowtie

L'alignement avec Bowtie des reads non utilisés par Kraken n'a pas apporté de nouvelles informations à part confirmer la présence de *S. cerevisiae*.

IV. Conclusions et perspectives

Nous voulions connaître la composition en microorganismes de nos lignées de moustiques afin d'identifier des organismes uniques entre plusieurs lignées de moustiques. Après la recherche et l'essai infructueux d'un outil permettant l'analyse métagénomique d'un échantillon séquencé, nous avons développé un outil listant les espèces présentes dans un échantillon après séquençage. L'outil créé, ICoMiO, est en fait un pipeline regroupant plusieurs outils d'analyse de données de séquençage et des scripts Perl personnels permettant de filtrer et résumer les résultats du BLAST et de Bowtie.

ICoMiO peut être utilisé sur un échantillon uniquement microbien ou sur un échantillon contenant l'organisme hôte et ses micro-organismes (MO).
Il permet :

- l'étude des organismes hôte : des analyses telles que l'étude de l'expression des gènes ou la détection de SNPs sont possibles après l'alignement
- l'étude du ou des pathogènes présents : expression des gènes, etc.
- une estimation de la composition d'un échantillon
- la détection de bactéries, de virus, de phages, d'archées et de champignons présents dans les bases de données sélectionnées
- la détection de nouvelles espèces par assemblage (espèces non séquencées)
- l'alignement de reads sur des espèces proches dans le cas d'espèces non séquencées
- la détection d'espèces faiblement représentées

Après des tests concluants sur des données simulées et réelles, ICoMiO a permis de détecter chez notre lignée résistante R1iso2 un ensemble de microorganismes. Nous avons identifié plusieurs virus qu'il conviendra de tester s'ils sont réellement présent ou s'il s'agit d'une probable contamination. Nous avons également identifié plusieurs espèces de bactéries dont la majorité a déjà été retrouvée chez des moustiques. Nous avons aussi établi la présence d'une levure, *S. cerevisiae*, dans notre lignée.

ICoMiO sera utilisé par la suite pour déterminer la composition des microbiotes des futurs séquençages de lignées résistantes et sensibles afin de détecter d'une part des différences de microbiotes entre les lignées résistantes et sensibles, et d'autre part, au sein d'une lignée, de chercher des corrélations entre la composition du microbiote des moustiques et leur niveau d'infection.

L'utilisation des MO d'un insecte pour le développement de stratégies de contrôle de maladies est un concept relativement nouveau et qui mérite d'être approfondi. L'utilisation des microbes a déjà été employée dans les biopesticides. La bactérie *Bacillus thuringiensis* et le champignon *Beauveria bassiana* ont été évalués sur le terrain pour le contrôle des larves de moustiques (Bukhari et al., 2011; Fillinger et al., 2003; Nartey et al., 2013). L'application la plus récente des MO dans le contrôle de la transmission d'une maladie humaine est le projet de recherche « Eliminate Dengue » (http://www.eliminatedengue.com/program), qui exploite les moustiques *Aedes aegypti* infectés par la bactérie *Wolbachia* pour contrôler la transmission du virus (Hoffmann et al., 2011; Rasgon, 2011; Walker et al., 2011). En plus d'avoir l'avantage d'être écologiquement sain, *Wolbachia* est capable de bloquer le développement d'un grand nombre de pathogènes dont les virus de la Dengue, du chikungunya et de la fièvre jaune et les parasites du genre *Plasmodium* (Cirimotich et al., 2011b; Hoffmann et al., 2011; van den Hurk et al., 2012; Iturbe-Ormaetxe et al., 2011; Kambris et al., 2010; Moreira et al., 2009; Rasgon, 2011; Walker et al., 2011).

Quelques lignées de *Wolbachia* réduisent la durée de vie des insectes et donc limite les chances du pathogène de terminer leur cycle à l'intérieur de l'insecte (McMeniman et al., 2009). Dans le cas du moustique *A. gambiae*, deux lignées de *Wolbachia* ont été capables d'inhiber les parasites de *P. falciparum* au sein de l'intestin (Hughes et al., 2011). Récemment, une toute nouvelle lignée de *Wolbachia* a été découverte en Afrique dans des populations de moustiques *A. gambiae* (Baldini et al., 2014), ce qui ouvre le champ de recherche des interactions entre elle, *A. gambiae* et le parasite *Plasmodium falciparum*.

L'étude des relations entre l'organisme hôte (le moustique), le pathogène (les parasites du paludisme) et les microorganismes peuvent donc conduire à l'utilisation de ces derniers pour contrôler le pathogène au sein du moustique et enrayer sa transmission à l'homme.

Plusieurs améliorations ou de nouveaux développements de mon outil pourront être réalisées :
(1) seule l'analyse de données de séquençage d'ADNg a été intégrée dans le pipeline, l'analyse de données RNA-Seq sera à rajouter,
(2) il serait intéressant d'ajouter un petit script Perl permettant de résumer et de filtrer le fichier .report issu de Kraken,
(3) et de créer notre propre base de données de Kraken en y ajoutant à l'actuelle les génomes du règne Fungi,
(4) les scripts Perl qui filtrent et résument les fichiers de sortie de BLAST et de Bowtie peuvent être améliorés au niveau des filtres à appliquer.

Conclusions et perspectives

Sur le terrain, certains moustiques sont naturellement capables d'éliminer les parasites et ne transmettent donc pas la maladie à l'homme. La capacité du moustique à transmettre les parasites dépend de plusieurs facteurs : la résistance du moustique au parasite, la virulence du parasite vis-à-vis du moustique et l'environnement.
L'objectif du laboratoire est d'identifier des facteurs génétiques et non génétiques contrôlant la résistance chez le moustique *Anopheles gambiae* aux parasites du paludisme *Plasmodium berghei*. Les récentes technologies de séquençage à très haut débit et les outils bioinformatiques offrent de nouvelles possibilités pour identifier ces facteurs. De plus, la récente méthode du « reciprocal allele-specific RNA interference » (rasRNAi) permet de tester la contribution des facteurs génétiques à la résistance. Mon travail a été le développement et la mise en place de nouvelles méthodes ou outils utilisant ces technologies.

Un de mes projets a été la mise en place d'une stratégie d'identification des polymorphismes dans les lignées résistantes et sensibles du laboratoire. Nous avons réalisé qu'il n'était pas possible d'obtenir l'ensemble des polymorphismes ni en alignant des reads issus du séquençage d'une lignée de moustique sur le génome de référence d'*A. gambiae*, ni en assemblant le génome de notre lignée.
Nous avons donc établi la méthode suivante :

8- nous réaliserons le séquençage Illumina d'ADN et d'ARN issus de moustiques individuels pour obtenir de petits reads.
9- nous alignerons ces reads sur le génome de référence qui sera sélectionné en fonction du type de polymorphismes voulu : intra- ou inter-lignées.
10- nous listerons tous les variants à partir des alignements précédents, éventuellement nous identifierons les variations du nombre de copies.
11- les polymorphismes inter-lignées identifiés seront ensuite utilisés comme marqueurs pour cartographier les QTLs et les polymorphismes intra-lignées pour l'étude d'association sur génome entier (GWAS).
12- les régions d'intérêts montrant un lien fort avec le degré de résistance et dont la séquence du GP est de mauvaise qualité ou ne correspond pas aux séquences de nos lignées seront réassemblées à partir des petits reads par une combinaison d'assemblage *de novo* par Velvet et Ray et de réalignements automatique et manuel. La consistance interne permettra de vérifier l'exactitude des régions assemblées. La connaissance précise de la séquence de ces régions nous apportera une liste exacte des polymorphismes présents et donc des gènes polymorphiques.
13- l'étude d'expression des gènes grâce aux données RNA-Seq nous dressera une liste des gènes différentiellement exprimés entre moustiques résistants et sensibles.
14- les étapes 4 et 5 nous fournissent, à travers les listes de gènes polymorphiques et de gènes différentiellement exprimés, une liste de gènes candidats à tester.

Cette méthode permettra d'identifier les polymorphismes de 6 lignées de moustiques résistants et sensibles du laboratoire afin de (1) sélectionner des marqueurs génétiques nécessaires à la cartographie de QTL ou à l'étude d'association sur le génome entier (GWAS), (2) lister les gènes polymorphiques et (3) obtenir la séquence précise de régions d'intérêts comme les loci de résistance.

En parallèle, j'ai contribué à l'élaboration de nouvelles sondes ARN double-brin (dsRNAs) allèle-spécifique pour la méthode du rasRNAi en identifiant le processus de découpage du dsRNA injecté chez des moustiques par l'analyse des petits ARNs séquencés issus du dsRNA injecté. Cette méthode permet de comparer la contribution de 2 allèles d'un même gène à un phénotype donné dans un même environnement génétique. Dans les cas où le gène dont nous souhaitions évaluer la contribution de chacun de ces allèles à la résistance est faiblement polymorphique, il n'était pas possible d'injecter un dsRNA long ordinaire car il aurait inhibé l'expression des deux allèles. J'ai réalisé les analyses des petits ARNs séquencés et en étudiant ceux issus de la séquence du dsRNA injecté, j'ai observé les faits suivants :

(6) les petits ARNs issus du dsRNA injecté ne sont pas répartis de manière homogène sur la séquence du dsRNA injecté : certaines séquences sont très abondantes par rapport à d'autres (repérables par la présence d'un pic) alors que d'autres régions du dsRNA ne sont pas présentes dans le pool de reads séquencés (zones que l'on a qualifié de « déserts »).
(7) le profil de distribution (répartition des pics et des « déserts ») des petits ARNs issus du dsRNA injecté est dans une certaine mesure indépendant de la séquence environnante et stable au cours du temps post-injection.
(8) les siRNAs de 19 à 24b montrent un taux de substitutions, surtout des transitions, à l'extrémité 3'. Ceci suggère que soit les siRNAs n'ont pas été protégés par la méthylation de leur extrémité 3', soit qu'ils ont été incorporés dans la voie des miRNAs où il n'y a pas de protection en 3'.
(9) la présence de SNPs dans la séquence d'un dsRNA injecté affecte son découpage et produit un profil de petits ARNs (pics et « déserts ») différent.
(10) un même dsRNA est découpé de la même façon quelque soit l'environnement génétique du moustique.

Forts de ces informations, nous avons établi une nouvelle sonde dsRNA qui contient à ses extrémités la séquence de 24b d'un siRNA allèle-spécifique et entre les deux, une séquence « porteuse » de 176 pb issue d'un gène exogène au moustique, LacZ. Cette nouvelle sonde a été nommée siRNA Carrier. L'analyse des petits ARNs séquencés combinée à la mesure de l'efficacité du knockdown par Western blot et/ou par qPCR nous a permis de démontrer, dans certains cas, l'efficacité des ces siRNA Carriers dont l'homologie avec la cible est limitée puisqu'il y a seulement 48b (20% du siRNA Carrier) qui proviennent de l'allèle à cibler. Cependant, nous n'avons pas observé de lien direct entre l'abondance d'une séquence et son efficacité lorsqu'elle est placée dans le siRNA Carrier, même s'il semble que les séquences correspondant aux pics sont globalement plus efficaces que les déserts.
De plus, nous avons montré que ces siRNA Carriers sont allèle-spécifiques, avec toutefois une action limitée lorsqu'un seul SNP distingue les deux allèles.

Il est donc maintenant possible de tester la contribution d'un allèle au phénotype de résistance dans les cas de gènes faiblement polymorphiques.
Cependant, afin d'établir si le siRNA Carrier allèle-spécifique est réellement efficace pour inhiber un allèle précis d'un gène peu polymorphique, il serait nécessaire de réaliser d'autres injections de siRNAs ne contenant qu'un seul SNP à différentes positions pour ce même gène déjà testé et également de tester d'autres gènes avec un ou plusieurs SNP(s).

En troisième partie de ma thèse, j'ai constitué un pipeline pour identifier la composition en micro-organismes (MO) d'une de nos lignées de laboratoire que nous avons séquencé. L'outil créé, ICoMiO pour Identification of Contaminant MicroOrganisms, est en fait un pipeline regroupant plusieurs outils d'analyse de données de séquençage et des scripts Perl personnels permettant de filtrer et résumer les résultats du BLAST et de Bowtie.

ICoMiO peut être utilisé sur un échantillon uniquement microbien ou sur un échantillon contenant l'organisme hôte et ses micro-organismes.
Il permet :

- l'étude des organismes hôte : des analyses telles que l'étude de l'expression des gènes ou la détection de SNPs sont possibles après l'alignement,
- l'étude du ou des pathogènes présents : expression des gènes, etc.,
- une estimation de la composition d'un échantillon,
- la détection de bactéries, de virus, de phages, d'archées et de champignons présents dans les bases de données sélectionnées,
- la détection de nouvelles espèces par assemblage (espèces non séquencées),
- l'alignement de reads sur des espèces proches dans le cas d'espèces non séquencées,
- la détection d'espèces faiblement représentées.

ICoMiO a permis de détecter chez notre lignée résistante R1iso2 un ensemble de microorganismes. Nous avons identifié plusieurs virus qu'il conviendra de tester s'ils sont réellement présent ou s'il s'agit d'une probable contamination. Nous avons également identifié plusieurs espèces de bactéries dont la majorité a déjà été retrouvée chez des moustiques. Nous avons aussi établi la présence d'une levure, *S. cerevisiae*, dans notre lignée.

ICoMiO sera utilisé par la suite pour déterminer la composition des microbiotes des futurs séquençages de lignées résistantes et sensibles afin de détecter d'une part des différences de microbiotes entre les lignées résistantes et sensibles, et d'autre part, au sein d'une lignée, de chercher des corrélations entre la composition du microbiote des moustiques et leur niveau d'infection.

L'étude des relations entre l'organisme hôte (le moustique), le pathogène (les parasites du paludisme) et les micro-organismes peuvent donc conduire à l'utilisation de ces derniers pour contrôler le pathogène au sein du moustique et enrayer sa transmission à l'homme.

Plusieurs améliorations ou de nouveaux développements de ICoMiO pourront être réalisées :
(1) seule l'analyse de données de séquençage d'ADNg a été intégrée dans le pipeline, l'analyse de données RNA-Seq sera à rajouter,
(2) il serait intéressant d'ajouter un petit script Perl permettant de résumer et de filtrer le fichier .report issu de Kraken,
(3) et de créer notre propre base de données de Kraken en y ajoutant à l'actuelle les génomes du règne Fungi,
(4) les scripts Perl qui filtrent et résument les fichiers de sortie de BLAST et de Bowtie peuvent être améliorés au niveau des filtres à appliquer.

L’ensemble de mes travaux a permis de développer et/ou d’améliorer des outils ou des méthodes liés à l’analyse bioinformatique de données NGS et à l’analyse génétique telle que la méthode du rasRNAi dans le but d’étendre les possibilités d’investigation de la recherche des facteurs de résistance des moustiques *A. gambiae* aux parasites du paludisme. Connaître les facteurs génétiques et non génétiques qui sont responsables de la résistance nous permettra de les manipuler afin de réduire ou d’arrêter la transmission de la maladie dans le milieu naturel. Ces méthodes ou outils ne sont pas spécifiques à nos données et à notre but biologique, ils peuvent donc être appliqués à d’autres thèmes de recherche.

Bibliographie

Abagli, A.Z., Alavo, T.B.C., and Brodeur, J. (2014). Microorganismes entomopathogènes, prédateurs et parasites des moustiques: Perspectives pour la lutte raisonnée contre les vecteurs du paludisme en Afrique sub-saharienne. Int. J. Biol. Chem. Sci. *8*, 340–354.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Ameres, S.L., Horwich, M.D., Hung, J.-H., Xu, J., Ghildiyal, M., Weng, Z., and Zamore, P.D. (2010). Target RNA-directed trimming and tailing of small silencing RNAs. Science *328*, 1534–1539.

Ameres, S.L., Hung, J.-H., Xu, J., Weng, Z., and Zamore, P.D. (2011). Target RNA-directed tailing and trimming purifies the sorting of endo-siRNAs between the two Drosophila Argonaute proteins. RNA *17*, 54–63.

Ander, C., Schulz-Trieglaff, O.B., Stoye, J., and Cox, A.J. (2013). metaBEETL: high-throughput analysis of heterogeneous microbial populations from shotgun DNA sequences. BMC Bioinformatics *14*, S2.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Aury, J.-M., Cruaud, C., Barbe, V., Rogier, O., Mangenot, S., Samson, G., Poulain, J., Anthouard, V., Scarpelli, C., Artiguenave, F., et al. (2008). High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics *9*, 603.

Baker, R.H., Reisen, W.K., Sakai, R.K., Hayes, C.G., Aslamkhan, M., Saifuddin, U.T., Mahmood, F., Perveen, A., and Javed, S. (1979). Field Assessment of Mating Competitiveness of Male Culex tritaeniorhynchus Carrying a Complex Chromosomal Aberration. Ann. Entomol. Soc. Am. *72*, 751–758.

Baldini, F., Segata, N., Pompon, J., Marcenac, P., Robert Shaw, W., Dabiré, R.K., Diabaté, A., Levashina, E.A., and Catteruccia, F. (2014). Evidence of natural Wolbachia infections in field populations of Anopheles gambiae. Nat. Commun. *5*.

Bando, H., Okado, K., Guelbeogo, W.M., Badolo, A., Aonuma, H., Nelson, B., Fukumoto, S., Xuan, X., Sagnon, N. 'fale, and Kanuka, H. (2013). Intra-specific diversity of Serratia marcescens in Anopheles mosquito midgut defines Plasmodium transmission capacity. Sci. Rep. *3*, 1641.

Bernardini, F., Galizi, R., Menichelli, M., Papathanos, P.-A., Dritsou, V., Marois, E., Crisanti, A., and Windbichler, N. (2014). Site-specific genetic engineering of the Anopheles gambiae Y chromosome. Proc. Natl. Acad. Sci. *111*, 7600–7605.

Blandin, S., Shiao, S.-H., Moita, L.F., Janse, C.J., Waters, A.P., Kafatos, F.C., and Levashina, E.A. (2004). Complement-Like Protein TEP1 Is a Determinant of Vectorial Capacity in the Malaria Vector Anopheles gambiae. Cell *116*, 661–670.

Blandin, S.A., Wang-Sattler, R., Lamacchia, M., Gagneur, J., Lycett, G., Ning, Y., Levashina, E.A., and Steinmetz, L.M. (2009). Dissecting the Genetic Basis of Resistance to Malaria Parasites in Anopheles gambiae. Science *326*, 147–150.

Boissière, A., Tchioffo, M.T., Bachar, D., Abate, L., Marie, A., Nsango, S.E., Shahbazkia, H.R., Awono-Ambene, P.H., Levashina, E.A., Christen, R., et al. (2012). Midgut Microbiota of the Malaria Mosquito Vector Anopheles gambiae and Interactions with Plasmodium falciparum Infection. PLoS Pathog *8*, e1002742.

Boisvert, S., Laviolette, F., and Corbeil, J. (2010). Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. J. Comput. Biol. *17*, 1519–1533.

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. *13*, R122.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., et al. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience *2*, 10.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., NISC Comparative Sequencing Program, Green, E.D., Sidow, A., and Batzoglou, S. (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. *13*, 721–731.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., Vargas, C. de, Gasol, J.M., et al. (2015). Patterns and ecological drivers of ocean viral communities. Science *348*, 1261498.

Bukhari, T., Takken, W., and Koenraadt, C.J. (2011). Development of Metarhizium anisopliae and Beauveria bassiana formulations for control of malaria mosquito larvae. Parasit. Vectors *4*, 23.

Carneiro, M., Ferrand, N., and Nachman, M.W. (2009). Recombination and Speciation: Loci Near Centromeres Are More Differentiated Than Loci Near Telomeres Between Subspecies of the European Rabbit (Oryctolagus cuniculus). Genetics *181*, 593–606.

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., Cramer, C.L., and Huang, X. (2015). Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. *16*, 30.

Chikhi, R., and Medvedev, P. (2013). Informed and Automated k-Mer Size Selection for Genome Assembly. Bioinformatics.

Christophides, G.K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P.T., Collins, F.H., Danielli, A., Dimopoulos, G., et al. (2002). Immunity-Related Genes and Gene Families in Anopheles gambiae. Science *298*, 159–165.

Cirimotich, C.M., Dong, Y., Clayton, A.M., Sandiford, S.L., Souza-Neto, J.A., Mulenga, M., and Dimopoulos, G. (2011a). Natural microbe-mediated refractoriness to Plasmodium infection in Anopheles gambiae. Science *332*, 855–858.

Cirimotich, C.M., Ramirez, J.L., and Dimopoulos, G. (2011b). Native Microbiota Shape Insect Vector Competence for Human Pathogens. Cell Host Microbe *10*, 307–310.

Cirimotich, C.M., Clayton, A.M., and Dimopoulos, G. (2011c). Low- and High-Tech Approaches to Control Plasmodium Parasite Transmission by Anopheles Mosquitoes. J. Trop. Med. *2011*.

Coetzee, M., Hunt, R.H., Wilkerson, R., Della Torre, A., Coulibaly, M.B., and Besansky, N.J. (2013). Anopheles coluzzii and Anopheles amharicus, new members of the Anopheles gambiae complex. Zootaxa *3619*, 246–274.

Collins, F.H., Sakai, R.K., Vernick, K.D., Paskewitz, S., Seeley, D.C., Miller, L.H., Collins, W.E., Campbell, C.C., and Gwadz, R.W. (1986). Genetic selection of a Plasmodium-refractory strain of the malaria vector Anopheles gambiae. Science *234*, 607–610.

Coluzzi, M., Sabatini, A., Petrarca, V., and Di Deco, M.A. (1979). Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. Trans. R. Soc. Trop. Med. Hyg. *73*, 483–497.

Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics *11*, 485.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: A Sequence Logo Generator. Genome Res. *14*, 1188–1190.

Davenport, C.F., Neugebauer, J., Beckmann, N., Friedrich, B., Kameri, B., Kokott, S., Paetow, M., Siekmann, B., Wieding-Drewes, M., Wienhöfer, M., et al. (2012). Genometa - A Fast and Accurate Classifier for Short Metagenomic Shotgun Reads. PLoS ONE *7*, e41224.

De, N., Young, L., Lau, P.-W., Meisner, N.-C., Morrissey, D.V., and MacRae, I.J. (2013). Highly Complementary Target RNAs Promote Release of Guide RNAs from Human Argonaute2. Mol. Cell *50*, 344–355.

Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L. (1999). Alignment of whole genomes. Nucleic Acids Res. *27*, 2369–2376.

Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. (2002). Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. *30*, 2478–2483.

Desai, A., Marwah, V.S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., Kulkarni, V., and Jere, A. (2013). Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data. PLoS ONE *8*, e60204.

Dong, Y., Aguilar, R., Xi, Z., Warr, E., Mongin, E., and Dimopoulos, G. (2006). Anopheles gambiae Immune Responses to Human and Rodent Plasmodium Parasite Species. PLoS Pathog *2*, e52.

Dong, Y., Manfredini, F., and Dimopoulos, G. (2009). Implication of the Mosquito Midgut Microbiota in the Defense against Malaria Parasites. PLoS Pathog. *5*.

Douglas, A.E. (2011). Lessons from Studying Insect Symbioses. Cell Host Microbe *10*, 359–367.

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., Yu, H.O.K., Buffalo, V., Zerbino, D.R., Diekhans, M., et al. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. Genome Res. *21*, 2224–2241.

Fillinger, U., Knols, B.G.J., and Becker, N. (2003). Efficacy and efficiency of new Bacillus thuringiensis var. israelensis and Bacillus sphaericus formulations against Afrotropical anophelines in Western Kenya. Trop. Med. Int. Health *8*, 37–47.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806–811.

Flicek, P., and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. Nat. Methods *6*, S6–S12.

Frolet, C., Thoma, M., Blandin, S., Hoffmann, J.A., and Levashina, E.A. (2006). Boosting NF-κB-Dependent Basal Immunity of Anopheles gambiae Aborts Development of Plasmodium berghei. Immunity *25*, 677–685.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., et al. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. Nature *419*, 498–511.

Gentile, G., Slotman, M., Ketmaier, V., Powell, J.R., and Caccone, A. (2001). Attempts to molecularly distinguish cryptic taxa in Anopheles gambiae s.s. Insect Mol. Biol. *10*, 25–32.

George, P., Sharakhova, M.V., and Sharakhov, I.V. (2010). High-resolution cytogenetic map for the African malaria vector Anopheles gambiae. Insect Mol. Biol. *19*, 675–682.

Ghildiyal, M., and Zamore, P.D. (2009). Small silencing RNAs: an expanding universe. Nat. Rev. Genet. *10*, 94–108.

Gimonneau, G., Tchioffo, M.T., Abate, L., Boissière, A., Awono-Ambéné, P.H., Nsango, S.E., Christen, R., and Morlais, I. (2014). Composition of Anopheles coluzzii and Anopheles gambiae microbiota from larval to adult stages. Infect. Genet. Evol. *28*, 715–724.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2010). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality Assessment Tool for Genome Assemblies. Bioinformatics.

Helinski, M.E., Parker, A.G., and Knols, B.G. (2006). Radiation-induced sterility for pupal and adult stages of the malaria mosquito Anopheles arabiensis. Malar. J. *5*, 41.

Helinski, M.E., Hassan, M.M., El-Motasim, W.M., Malcolm, C.A., Knols, B.G., and El-Sayed, B. (2008). Towards a sterile insect technique field release of Anopheles arabiensis mosquitoes in Sudan: Irradiation, transportation, and field cage experimentation. Malar. J. *7*, 65.

Hoffmann, A.A., Montgomery, B.L., Popovici, J., Iturbe-Ormaetxe, I., Johnson, P.H., Muzzi, F., Greenfield, M., Durkan, M., Leong, Y.S., Dong, Y., et al. (2011). Successful establishment of Wolbachia in Aedes populations to suppress dengue transmission. Nature *476*, 454–457.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.C., Wides, R., et al. (2002a). The Genome Sequence of the Malaria Mosquito Anopheles gambiae. Science *298*, 129–149.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.C., Wides, R., et al. (2002b). The Genome Sequence of the Malaria Mosquito Anopheles gambiae. Science *298*, 129–149.

Hong, S.W., Park, J.H., Yun, S., Lee, C.H., Shin, C., and Lee, D. (2014). Effect of the guide strand 3′-end structure on the gene-silencing potency of asymmetric siRNA. Biochem. J. *461*, 427–434.

Horn, T., and Boutros, M. (2010). E-RNAi: a web application for the multi-species design of RNAi reagents--2010 update. Nucleic Acids Res. *38*, W332–W339.

Horn, T., and Boutros, M. (2013). Design of RNAi Reagents for Invertebrate Model Organisms and Human Disease Vectors. In siRNA Design, D.J. Taxman, ed. (Totowa, NJ: Humana Press), pp. 315–346.

Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics *28*, 593–594.

Hughes, G.L., Koga, R., Xue, P., Fukatsu, T., and Rasgon, J.L. (2011). Wolbachia infections are virulent and inhibit the human malaria parasite Plasmodium falciparum in Anopheles gambiae. PLoS Pathog. *7*, e1002043.

Huho, B.J., Ng'habi, K.R., Killeen, G.F., Nkwengulila, G., Knols, B.G.J., and Ferguson, H.M. (2007). Nature beats nurture: a case study of the physiological fitness of free-living and laboratory-reared male Anopheles gambiae s.l. J. Exp. Biol. *210*, 2939–2947.

van den Hurk, A.F., Hall-Mendelin, S., Pyke, A.T., Frentiu, F.D., McElroy, K., Day, A., Higgs, S., and O'Neill, S.L. (2012). Impact of Wolbachia on Infection with Chikungunya and Yellow Fever Viruses in the Mosquito Vector Aedes aegypti. PLoS Negl Trop Dis *6*, e1892.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. *17*, 377–386.

Ibrahim, F., Rymarquis, L.A., Kim, E.-J., Becker, J., Balassa, E., Green, P.J., and Cerutti, H. (2010). Uridylation of mature miRNAs and siRNAs by the MUT68 nucleotidyltransferase

promotes their degradation in Chlamydomonas. Proc. Natl. Acad. Sci. U. S. A. *107*, 3906–3911.

Iturbe-Ormaetxe, I., Walker, T., and O' Neill, S.L. (2011). Wolbachia and the biological control of mosquito-borne disease. EMBO Rep. *12*, 508–518.

Kambris, Z., Blagborough, A.M., Pinto, S.B., Blagrove, M.S.C., Godfray, H.C.J., Sinden, R.E., and Sinkins, S.P. (2010). Wolbachia Stimulates Immune Gene Expression and Inhibits Plasmodium Development in Anopheles gambiae. PLoS Pathog *6*, e1001143.

Kämpfer, P., Matthews, H., Glaeser, S.P., Martin, K., Lodders, N., and Faye, I. (2011). Elizabethkingia anophelis sp. nov., isolated from the midgut of the mosquito Anopheles gambiae. Int. J. Syst. Evol. Microbiol. *61*, 2670–2675.

Kelley, D.R., Schatz, M.C., and Salzberg, S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol. *11*, R116.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs Exhibit Strand Bias. Cell *115*, 209–216.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K., and Mardis, E.R. (2013). The Next-Generation Sequencing Revolution and Its Impact on Genomics. Cell *155*, 27–38.

van der Krol, A.R., Mur, L.A., Beld, M., Mol, J.N., and Stuitje, A.R. (1990a). Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. Plant Cell *2*, 291–299.

van der Krol, A.R., Mur, L.A., de Lange, P., Mol, J.N., and Stuitje, A.R. (1990b). Inhibition of flower pigmentation by antisense CHS genes: promoter and minimal sequence requirements for the antisense effect. Plant Mol. Biol. *14*, 457–466.

Kuczynski, J., Costello, E.K., Nemergut, D.R., Zaneveld, J., Lauber, C.L., Knights, D., Koren, O., Fierer, N., Kelley, S.T., Ley, R.E., et al. (2010). Direct sequencing of the human microbiome readily reveals community differences. Genome Biol. *11*, 210.

Kultima, J.R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D.R., Arumugam, M., Pan, Q., Liu, B., Qin, J., et al. (2012). MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. PLoS ONE *7*, e47656.

Kumar, S., Molina-Cruz, A., Gupta, L., Rodrigues, J., and Barillas-Mury, C. (2010). A Peroxidase/Dual Oxidase System Modulates Midgut Epithelial Immunity in Anopheles gambiae. Science *327*, 1644–1648.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol. *5*, R12.

Ladunga, I. (2007). More complete gene silencing by fewer siRNAs: transparent optimized design and biophysical signature. Nucleic Acids Res. *35*, 433–440.

Lambrechts, L., Halbert, J., Durand, P., Gouagna, L.C., and Koella, J.C. (2005). Host genotype by parasite genotype interactions underlying the resistance of anopheline mosquitoes to Plasmodium falciparum. Malar. J. *4*, 3.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Lawniczak, M.K.N., Emrich, S.J., Holloway, A.K., Regier, A.P., Olson, M., White, B., Redmond, S., Fulton, L., Appelbaum, E., Godfrey, J., et al. (2010). Widespread Divergence Between Incipient Anopheles gambiae Species Revealed by Whole Genome Sequences. Science *330*, 512–514.

Lewis, Z., and Lizé, A. (2015). Insect behaviour and the microbiome. Curr. Opin. Insect Sci.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinforma. Oxf. Engl. *27*, 2987–2993.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., et al. (2010). The sequence and de novo assembly of the giant panda genome. Nature *463*, 311–317.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., Roux, S., Vincent, F., et al. (2015). Determinants of community structure in the global plankton interactome. Science *348*, 1262073.

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., and Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics *12*, S4.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of Next-Generation Sequencing Systems. J. Biomed. Biotechnol. *2012*, 1–11.

Lu, Z.J., and Mathews, D.H. (2008). Efficient siRNA selection using hybridization thermodynamics. Nucleic Acids Res. *36*, 640–647.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience *1*, 18.

Manichanh, C., Chapple, C.E., Frangeul, L., Gloux, K., Guigo, R., and Dore, J. (2008). A comparison of random sequence reads versus 16S rDNA sequences for estimating the biodiversity of a metagenomic library. Nucleic Acids Res. *36*, 5180–5188.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. Nature *437*, 376–380.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, pp. 10–12.

Mavromatis, K., Land, M.L., Brettin, T.S., Quest, D.J., Copeland, A., Clum, A., Goodwin, L., Woyke, T., Lapidus, A., Klenk, H.P., et al. (2012). The Fast Changing Landscape of Sequencing Technologies and Their Impact on Microbial Genome Assemblies and Annotation. PLoS ONE *7*, e48837.

Maxam, A.M., and Gilbert, W. (1977). A new method for sequencing DNA. Proc. Natl. Acad. Sci. U. S. A. *74*, 560–564.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

McMeniman, C.J., Lane, R.V., Cass, B.N., Fong, A.W.C., Sidhu, M., Wang, Y.-F., and O'Neill, S.L. (2009). Stable Introduction of a Life-Shortening Wolbachia Infection into the Mosquito Aedes aegypti. Science *323*, 141–144.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., et al. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics *9*, 386.

Minard, G., Mavingui, P., and Moro, C.V. (2013). Diversity and function of bacterial microbiota in the mosquito holobiont. Parasit. Vectors *6*, 146.

Moiroux, N., Gomez, M.B., Pennetier, C., Elanga, E., Djènontin, A., Chandre, F., Djègbé, I., Guis, H., and Corbel, V. (2012). Changes in Anopheles funestus Biting Behavior Following Universal Coverage of Long-Lasting Insecticidal Nets in Benin. J. Infect. Dis. *206*, 1622–1629.

Molina-Cruz, A., Garver, L.S., Alabaster, A., Bangiolo, L., Haile, A., Winikor, J., Ortega, C., van Schaijk, B.C.L., Sauerwein, R.W., Taylor-Salmon, E., et al. (2013). The Human Malaria Parasite Pfs47 Gene Mediates Evasion of the Mosquito Immune System. Science *340*.

Mongin, E., Louis, C., Holt, R.A., Birney, E., and Collins, F.H. (2004). The Anopheles gambiae genome: an update. Trends Parasitol. *20*, 49–52.

Moreira, L.A., Iturbe-Ormaetxe, I., Jeffery, J.A., Lu, G., Pyke, A.T., Hedges, L.M., Rocha, B.C., Hall-Mendelin, S., Day, A., Riegler, M., et al. (2009). A Wolbachia Symbiont in Aedes aegypti Limits Infection with Dengue, Chikungunya, and Plasmodium. Cell *139*, 1268–1278.

Naito, Y., Yamada, T., Ui-Tei, K., Morishita, S., and Saigo, K. (2004). siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference. Nucleic Acids Res. *32*, W124–W129.

Naito, Y., Yoshimura, J., Morishita, S., and Ui-Tei, K. (2009). siDirect 2.0: updated software for designing functional siRNA with reduced seed-dependent off-target effect. BMC Bioinformatics *10*, 392.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. gks678.

Nartey, R., Owusu-Dabo, E., Kruppa, T., Baffour-Awuah, S., Annan, A., Oppong, S., Becker, N., and Obiri-Danso, K. (2013). Use of Bacillus thuringiensis var israelensis as a viable option in an Integrated Malaria Vector Control Programme in the Kumasi Metropolis, Ghana. Parasit. Vectors *6*, 116.

Neafsey, D.E., Lawniczak, M.K.N., Park, D.J., Redmond, S.N., Coulibaly, M.B., Traore, S.F., Sagnon, N., Costantini, C., Johnson, C., Wiegand, R.C., et al. (2010). SNP Genotyping Defines Complex Gene-Flow Boundaries Among African Malaria Vector Mosquitoes. Science *330*, 514–517.

Neafsey, D.E., Christophides, G.K., Collins, F.H., Emrich, S.J., Fontaine, M.C., Gelbart, W., Hahn, M.W., Howell, P.I., Kafatos, F.C., Lawson, D., et al. (2013). The Evolution of the Anopheles 16 Genomes Project. G3 GenesGenomesGenetics.

Neafsey, D.E., Waterhouse, R.M., Abai, M.R., Aganezov, S.S., Alekseyev, M.A., Allen, J.E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., et al. (2014). Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science 1258522.

Neafsey, D.E., Waterhouse, R.M., Abai, M.R., Aganezov, S.S., Alekseyev, M.A., Allen, J.E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., et al. (2015). Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science *347*, 1258522.

Oliva, C.F., Jacquet, M., Gilles, J., Lemperiere, G., Maquart, P.-O., Quilici, S., Schooneman, F., Vreysen, M.J.B., and Boyer, S. (2012). The Sterile Insect Technique for Controlling Populations of Aedes albopictus (Diptera: Culicidae) on Reunion Island: Mating Vigour of Sterilized Males. PLoS ONE *7*, e49414.

Ondov, B.D., Bergman, N.H., and Phillippy, A.M. (2011). Interactive metagenomic visualization in a Web browser. BMC Bioinformatics *12*, 385.

Osei-Poku, J., Mbogo, C.M., Palmer, W.J., and Jiggins, F.M. (2012). Deep sequencing reveals extensive variation in the gut microbiota of wild mosquitoes from Kenya. Mol. Ecol. *21*, 5138–5150.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.-Y., Cohoon, M., Crécy-Lagard, V. de, Diaz, N., Disz, T., Edwards, R., et al. (2005). The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. Nucleic Acids Res. *33*, 5691–5702.

Peng, Y., Leung, H.C., Yiu, S.-M., and Chin, F.Y. (2010). IDBA–a practical iterative de Bruijn graph de novo assembler. In Research in Computational Molecular Biology, (Springer), pp. 426–440.

Pop, M., Kosack, D.S., and Salzberg, S.L. (2004). Hierarchical Scaffolding With Bambus. Genome Res. *14*, 149–159.

Rani, A., Sharma, A., and Rajagopal, R. (2009). Bacterial diversity analysis of larvae and adult midgut microflora using culture-dependent and culture-independent methods in lab-reared and field-collected Anopheles stephensi-an Asian malarial vector.

Ranson, H., Claudianos, C., Ortelli, F., Abgrall, C., Hemingway, J., Sharakhova, M.V., Unger, M.F., Collins, F.H., and Feyereisen, R. (2002). Evolution of Supergene Families Associated with Insecticide Resistance. Science *298*, 179–181.

Rasgon, J.L. (2011). Dengue fever: Mosquitoes attacked from within. Nature *476*, 407–408.

Reddy, T.B.K., Thomas, A.D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E.A., and Kyrpides, N.C. (2014). The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. gku950.

Reegan, A.D., Paulraj, M.G., and Ignacimuthu, S. (2013). Isolation and characterization of halotolerant bacteria associated with the midgut of Culex quinquefasciatus Say (Diptera: Culicidae). Pak. J. Biol. Sci. PJBS *16*, 1311–1317.

Reidenbach, K.R., Neafsey, D.E., Costantini, C., Sagnon, N., Simard, F., Ragland, G.J., Egan, S.P., Feder, J.L., Muskavitch, M.A.T., and Besansky, N.J. (2012). Patterns of Genomic Differentiation between Ecologically Differentiated M and S Forms of Anopheles gambiae in West and Central Africa. Genome Biol. Evol. *4*, 1202–1212.

Reinhardt, J.A., Baltrus, D.A., Nishimura, M.T., Jeck, W.R., Jones, C.D., and Dangl, J.L. (2009). De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae. Genome Res. *19*, 294–305.

Reisen, W.K., Sakai, R.K., Baker, R.H., Rathor, H.R., Raana, K., Azra, K., and Niaz, S. (1980). Field Competitiveness of Culex tritaeniorhynchus Giles Males Carrying a Complex Chromosomal Aberration: a Second Experiment. Ann. Entomol. Soc. Am. *73*, 479–484.

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S., and Khvorova, A. (2004). Rational siRNA design for RNA interference. Nat. Biotechnol. *22*, 326–330.

Rottschaefer, S.M., Riehle, M.M., Coulibaly, B., Sacko, M., Niaré, O., Morlais, I., Traoré, S.F., Vernick, K.D., and Lazzaro, B.P. (2011). Exceptional Diversity, Maintenance of Polymorphism, and Recent Directional Selection on the APL1 Malaria Resistance Genes of Anopheles gambiae. PLoS Biol *9*, e1000600.

Saleh, M.-C., van Rij, R.P., Hekele, A., Gillis, A., Foley, E., O'Farrell, P.H., and Andino, R. (2006). The endocytic pathway mediates cell entry of dsRNA to induce RNAi silencing. Nat. Cell Biol. *8*, 793–802.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. *94*, 441–448.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.

Schwarz, D.S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the Assembly of the RNAi Enzyme Complex. Cell *115*, 199–208.

Shah, J.K., Garner, H.R., White, M.A., Shames, D.S., and Minna, J.D. (2007). sIR: siRNA Information Resource, a web-based tool for siRNA sequence design and analysis and an open access siRNA database. BMC Bioinformatics *8*, 178.

Sharakhova, M.V., Hammond, M.P., Lobo, N.F., Krzywinski, J., Unger, M.F., Hillenmeyer, M.E., Bruggner, R.V., Birney, E., and Collins, F.H. (2007). Update of the Anopheles gambiae PEST genome assembly. Genome Biol. *8*, R5.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., and Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. Genome Res. *19*, 1117–1123.

Stathopoulos, S., Neafsey, D.E., Lawniczak, M.K.N., Muskavitch, M.A.T., and Christophides, G.K. (2014). Genetic Dissection of Anopheles gambiae Gut Epithelial Responses to Serratia marcescens. PLoS Pathog *10*, e1003897.

Sunagawa, S., Coelho, L.P., Chaffron, S., and Kultima, J.R. (2015). Structure and function of the global ocean microbiome.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Torre, A. della, Fanello, C., Akogbeto, M., Dossou-yovo, J., Favia, G., Petrarca, V., and Coluzzi, M. (2001). Molecular evidence of incipient speciation within Anopheles gambiae s.s. in West Africa. Insect Mol. Biol. *10*, 9–18.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Treangen, T.J., Koren, S., Sommer, D.D., Liu, B., Astrovskaya, I., Ondov, B., Darling, A.E., Phillippy, A.M., and Pop, M. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol. *14*, R2.

Turner, T.L., Hahn, M.W., and Nuzhdin, S.V. (2005). Genomic Islands of Speciation in Anopheles gambiae. PLoS Biol *3*, e285.

Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., and Saigo, K. (2004). Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. Nucleic Acids Res. *32*, 936–948.

Utturkar, S.M., Klingeman, D.M., Land, M.L., Schadt, C.W., Doktycz, M.J., Pelletier, D.A., and Brown, S.D. (2014). Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics *30*, 2709–2716.

Valiente Moro, C., Tran, F.H., Nantenaina Raharimalala, F., Ravelonandro, P., and Mavingui, P. (2013). Diversity of culturable bacteria including Pantoea in wild mosquito Aedes albopictus. BMC Microbiol. *13*, 70.

Vargas, C. de, Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Bescot, N.L., Probert, I., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. Science *348*, 1261605.

Vázquez-Castellanos, J.F., García-López, R., Pérez-Brocal, V., Pignatelli, M., and Moya, A. (2014). Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. BMC Genomics *15*, 37.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. Science *291*, 1304–1351.

Vernick, K.D., Fujioka, H., Seeley, D.C., Tandler, B., Aikawa, M., and Miller, L.H. (1995). Plasmodium gallinaceum: a refractory mechanism of ookinete killing in the mosquito, Anopheles gambiae. Exp. Parasitol. *80*, 583–595.

Vert, J.-P., Foveau, N., Lajaunie, C., and Vandenbrouck, Y. (2006). An accurate and interpretable model for siRNA efficacy prediction. BMC Bioinformatics *7*, 520.

Vol'kenshtein, M.V. (1976). Probabilities of transversions and transitions. Mol. Biol. (Mosk.) *10*, 605–608.

Walker, T., Johnson, P.H., Moreira, L.A., Iturbe-Ormaetxe, I., Frentiu, F.D., McMeniman, C.J., Leong, Y.S., Dong, Y., Axford, J., Kriesner, P., et al. (2011). The wMel Wolbachia strain blocks dengue and invades caged Aedes aegypti populations. Nature *476*, 450–453.

Wang, Y., Juranek, S., Li, H., Sheng, G., Wardle, G.S., Tuschl, T., and Patel, D.J. (2009). Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. Nature *461*, 754–761.

Wee, L.M., Flores-Jasso, C.F., Salomon, W.E., and Zamore, P.D. (2012). Argonaute Divides Its RNA Guide into Domains with Distinct Functions and RNA-Binding Properties. Cell *151*, 1055–1067.

White, B.J., Cheng, C., Simard, F., Costantini, C., and Besansky, N.J. (2010). Genetic association of physically unlinked islands of genomic divergence in incipient species of Anopheles gambiae. Mol. Ecol. *19*, 925–939.

White, B.J., Lawniczak, M.K.N., Cheng, C., Coulibaly, M.B., Wilson, M.D., Sagnon, N., Costantini, C., Simard, F., Christophides, G.K., and Besansky, N.J. (2011). Adaptive divergence between incipient species of Anopheles gambiae increases resistance to Plasmodium. Proc. Natl. Acad. Sci. *108*, 244–249.

Wondji, C., Simard, F., and Fontenille, D. (2002). Evidence for genetic differentiation between the molecular forms M and S within the Forest chromosomal form of Anopheles gambiae in an area of sympatry. Insect Mol. Biol. *11*, 11–19.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

Wu, T.D., and Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873–881.

Yang, Z., Ebright, Y.W., Yu, B., and Chen, X. (2006). HEN1 recognizes 21–24 nt small RNA duplexes and deposits a methyl group onto the 2′ OH of the 3′ terminal nucleotide. Nucleic Acids Res. *34*, 667–675.

Yeung, S., Pongtavornpinyo, W., Hastings, I.M., Mills, A.J., and White, N.J. (2004). Antimalarial Drug Resistance, Artemisinin-Based Combination Therapy, and the Contribution of Modeling to Elucidating Policy Choices. Am. J. Trop. Med. Hyg. *71*, 179–186.

Zamore, P.D., Tuschl, T., Sharp, P.A., and Bartel, D.P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell *101*, 25–33.

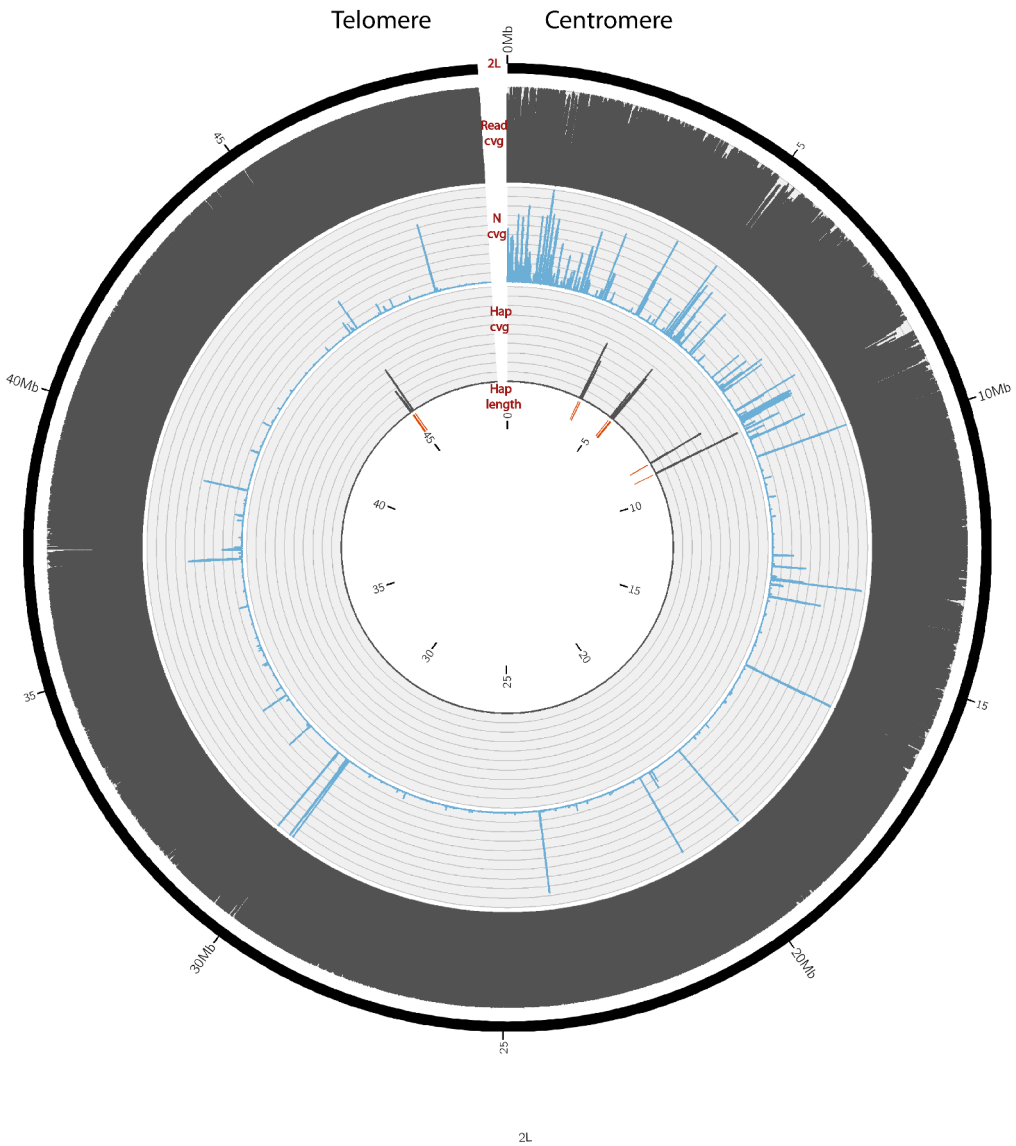
Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhao, Y., Yu, Y., Zhai, J., Ramachandran, V., Dinh, T.T., Meyers, B.C., Mo, B., and Chen, X. (2012). HESO1, a nucleotidyl transferase in Arabidopsis, uridylates unmethylated miRNAs and siRNAs to trigger their degradation. Curr. Biol. CB *22*, 689–694.

(2008). BLAST® Command Line Applications User Manual (National Center for Biotechnology Information (US)).

Annexe I

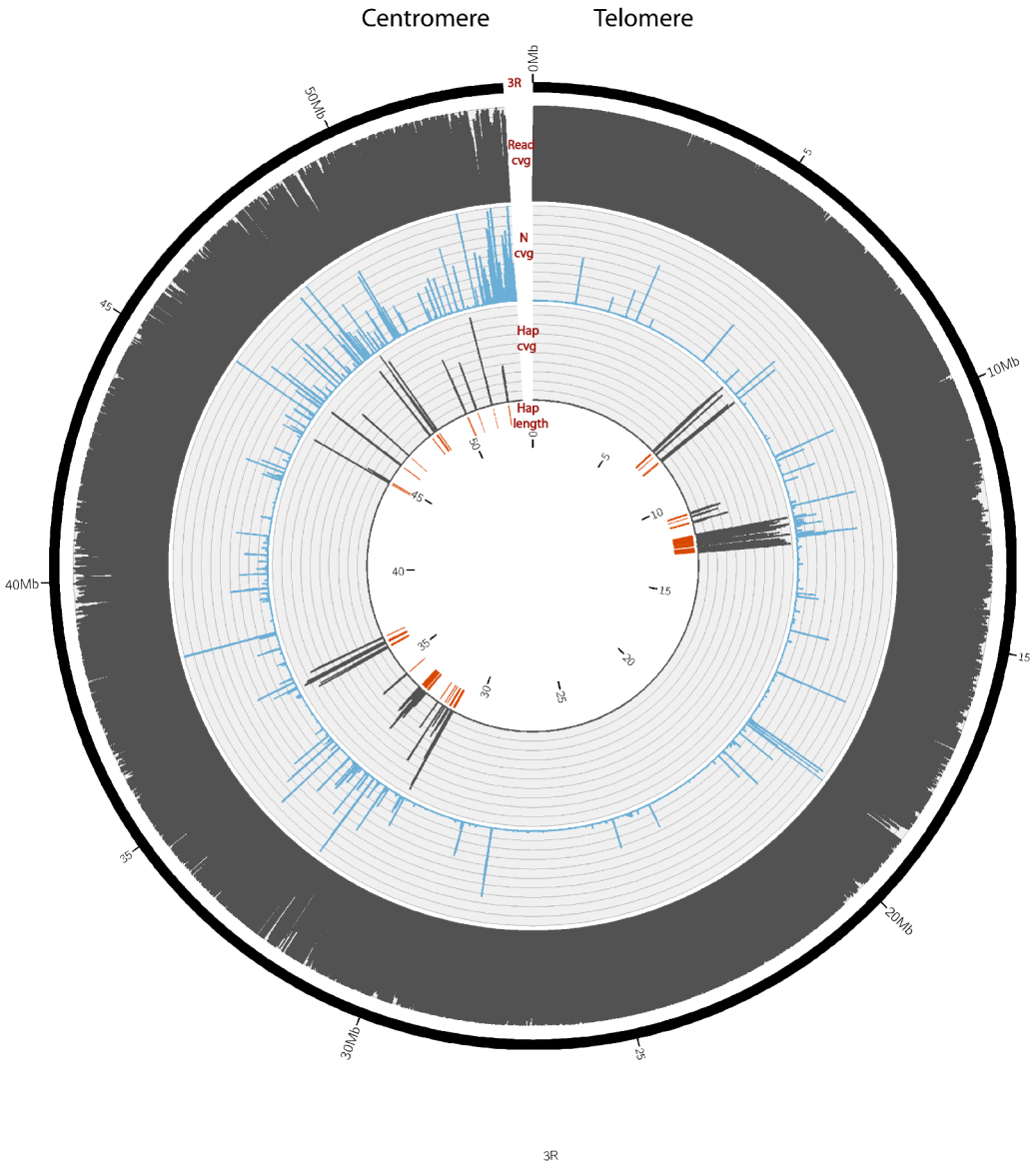
Visualisation Circos des reads alignés sur le chromosome 2L et ses haplotypes



Répartition des reads alignés sur le chromosome 2L et ses haplotypes. La lecture se fait de l'extérieur vers l'intérieur : (1) la séquence du chromosome de référence, (2) la couverture en reads alignés sur le Golden Path par fenêtre de 10kb, (3) la proportion de N par fenêtre de 10kb, (4) la couverture en reads alignés sur les haplotypes par fenêtre de 10kb et (5) la position et la taille des haplotypes. Pour chaque rangée, l'échelle est comprise entre 0 et 100% et les graduations correspondent à une augmentation de 10%.

Annexe II

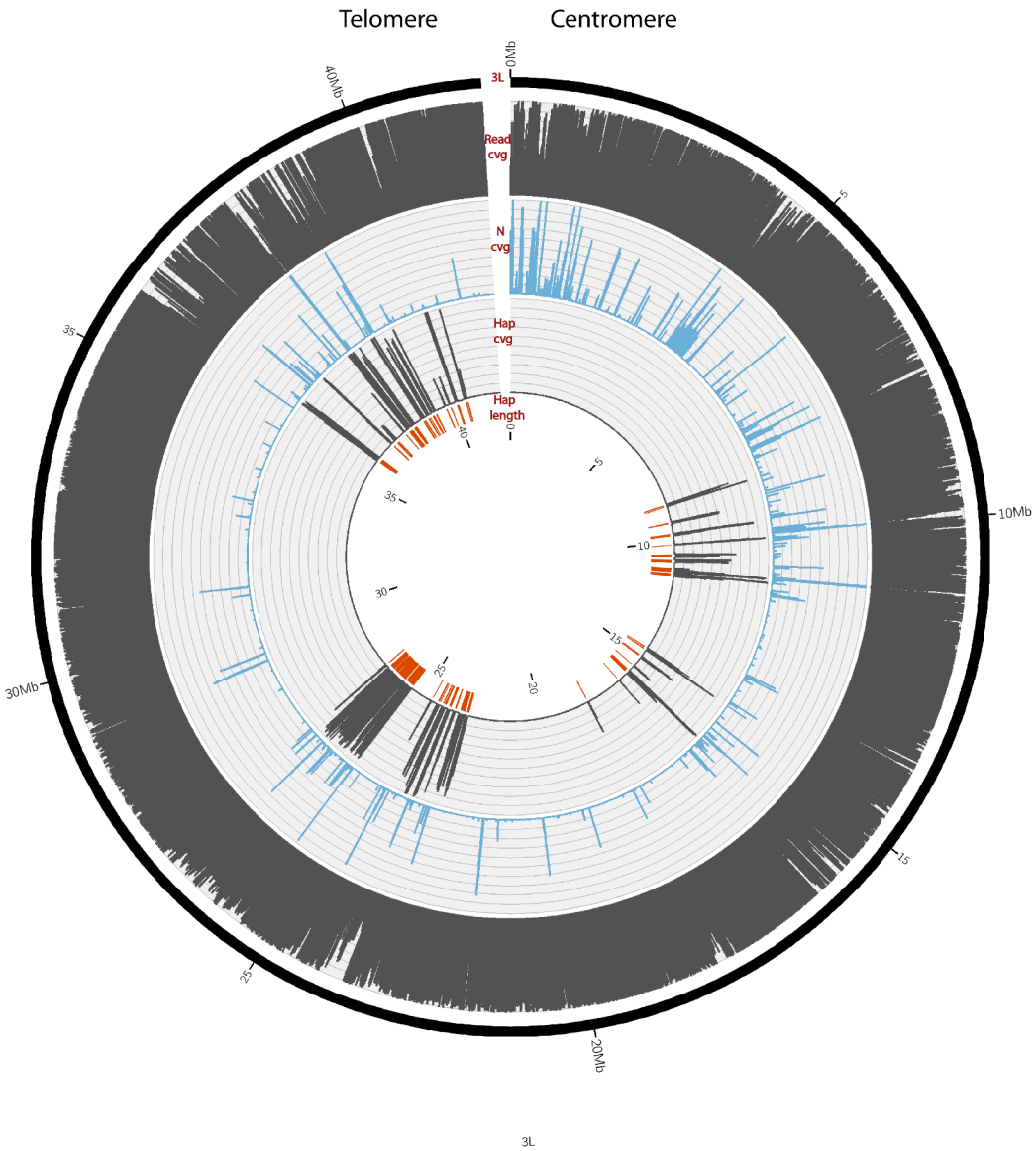
Visualisation Circos des reads alignés sur le chromosome 3R et ses haplotypes



Répartition des reads alignés sur le chromosome 3R et ses haplotypes. La lecture se fait de l'extérieur vers l'intérieur : (1) la séquence du chromosome de référence, (2) la couverture en reads alignés sur le Golden Path par fenêtre de 10kb, (3) la proportion de N par fenêtre de 10kb, (4) la couverture en reads alignés sur les haplotypes par fenêtre de 10kb et (5) la position et la taille des haplotypes. Pour chaque rangée, l'échelle est comprise entre 0 et 100% et les graduations correspondent à une augmentation de 10%.

Annexe III

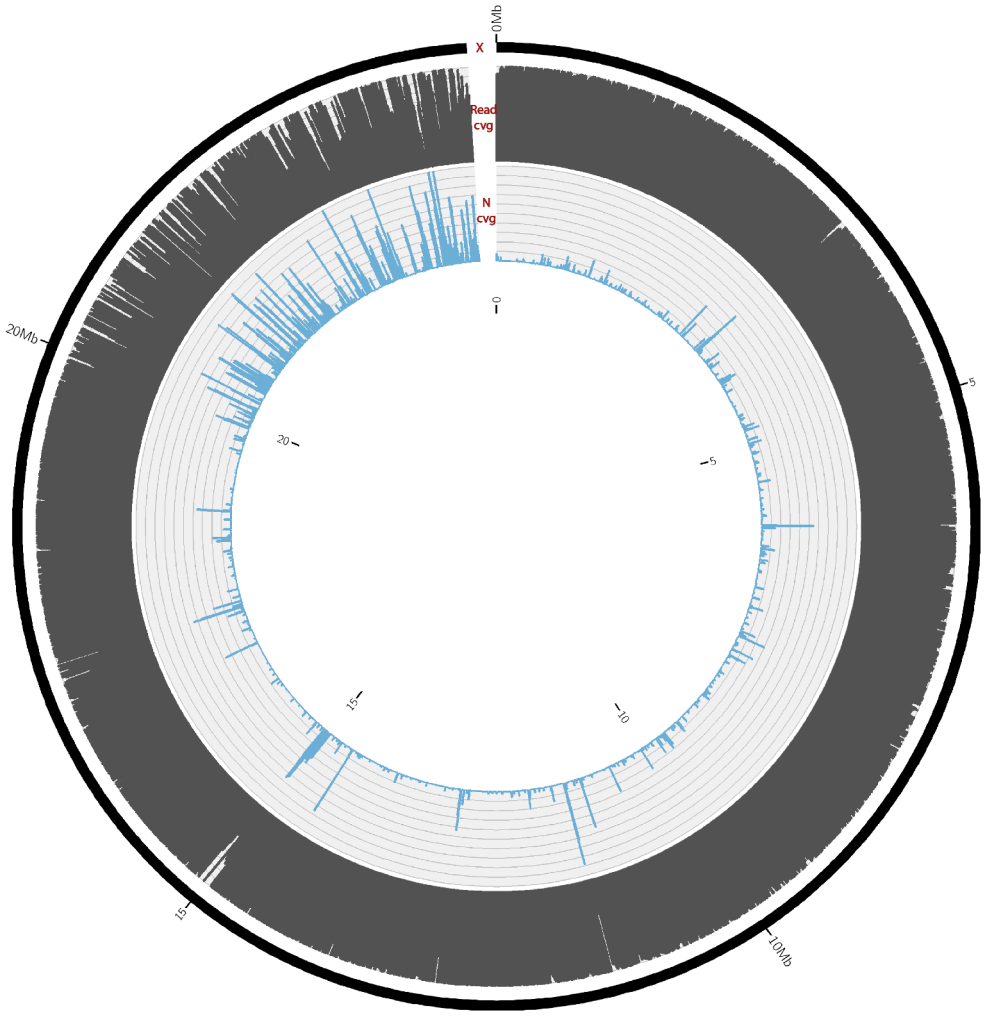
Visualisation Circos des reads alignés sur le chromosome 3L et ses haplotypes



Répartition des reads alignés sur le chromosome 3L et ses haplotypes. La lecture se fait de l'extérieur vers l'intérieur : (1) la séquence du chromosome de référence, (2) la couverture en reads alignés sur le Golden Path par fenêtre de 10kb, (3) la proportion de N par fenêtre de 10kb, (4) la couverture en reads alignés sur les haplotypes par fenêtre de 10kb et (5) la position et la taille des haplotypes. Pour chaque rangée, l'échelle est comprise entre 0 et 100% et les graduations correspondent à une augmentation de 10%.

Annexe IV

Visualisation Circos des reads alignés sur le chromosome X



X

Répartition des reads alignés sur le chromosome X. La lecture se fait de l'extérieur vers l'intérieur : (1) la séquence du chromosome de référence, (2) la couverture en reads alignés sur le Golden Path par fenêtre de 10kb, (3) la proportion de N par fenêtre de 10kb. Pour chaque rangée, l'échelle est comprise entre 0 et 100% et les graduations correspondent à une augmentation de 10%.

Annexe V

Nombre de reads attribués par Bowtie et Kraken sur les espèces présentes dans le set de données MockE

Espèces	Masse d'ADNg (g)	Kraken	Bowtie
Lactobacillus gasseri	1,53E-11	1 343	1 447
Streptococcus agalactiae	1,83E-11	4 209	4 550
Enterococcus faecalis	2,22E-11	23 459	59 525
Escherichia coli	2,71E-11	18 504	79 752
Bacillus cereus	3,73E-11	22 183	37 678
Clostridium beijerinckii	3,81E-11	200 350	207 809
Listeria monocytogenes	3,98E-11	90 906	104 612
Helicobacter pylori	4,50E-11	138 902	151 738
Streptococcus mutans	4,70E-11	91 535	96 056
Neisseria meningitidis	6,87E-11	83 126	137 315
Staphylococcus aureus	6,97E-11	158 295	206 882
Streptococcus pneumoniae	8,11E-11	131 703	192 096
Methanobrevibacter smithii	9,50E-11	36 692	37 848
Rhodobacter sphaeroides	1,30E-10	157 344	166 497
Staphylococcus epidermidis	1,31E-10	230 329	211 290
Propionibacterium acnes	1,39E-10	205 340	238 954
Bacteroides vulgatus	1,52E-10	485 331	520 260
Acinetobacter baumannii	1,60E-10	599 533	648 079
Pseudomonas aeruginosa	1,80E-10	34 205	43 497
Deinococcus radiodurans	1,76E-09	1 876 372	2 008 560

Le tableau est classé par la masse d'ADN génomique (ADNg) croissante.

Annexe VI

Nombre de reads attribués par Bowtie et Kraken sur les espèces présentes dans le set de données MockS

Espèces	Masse d'ADNg (g)	Kraken	Bowtie
Enterococcus faecalis	2,22E-13	87	271
Streptococcus pneumoniae	8,11E-13	317	833
Bacteroides vulgatus	1,52E-12	1 340	1 322
Lactobacillus gasseri	1,53E-12	339	344
Listeria monocytogenes	3,98E-12	3 863	4 117
Helicobacter pylori	4,50E-12	5 465	5 211
Neisseria meningitidis	6,87E-12	4 537	9 841
Propionibacterium acnes	1,39E-11	12 703	15 089
Acinetobacter baumannii	1,60E-11	24 430	23 449
Deinococcus radiodurans	1,76E-11	14 561	16 942
Streptococcus agalactiae	1,83E-11	26 320	24 525
Bacillus cereus	3,73E-11	8 728	13 558
Clostridium beijerinckii	3,81E-11	74 686	67 864
Staphylococcus aureus	6,97E-11	1 288 905	1 378 203
Pseudomonas aeruginosa	1,80E-10	127 853	174 454
Escherichia coli	2,71E-10	98 966	427 280
Streptococcus mutans	4,70E-10	467 179	432 200
Methanobrevibacter smithii	9,50E-10	238 679	214 216
Rhodobacter sphaeroides	1,30E-09	1 346 885	1 573 214
Staphylococcus epidermidis	1,31E-09	1 115 437	898 112

Le tableau est classé par la masse d'ADN génomique (ADNg) croissante.

Annexe VII

Nombre de reads attribués par Bowtie et Kraken sur les espèces présentes dans le set de données simulées

Espèces	Masse d'ADNg (g)	Kraken	Bowtie
Candidatus Sulcia	28975	12978	27673
Syncytium symbiont	30625	13766	29256
Nanoarchaeum equitans	32725	14861	31299
Candidatus Carsonella	32900	14774	31472
Candidatus Riesia	38290	17195	36559
Mycoplasma genitalium	38630	17138	36909
Blattabacterium sp. (Blaberus	41940	17746	39697
Blattabacterium sp. (Blatta	42295	10900	31097
Mycoplasma wenyonii	43345	19353	41387
Ureaplasma parvum	50110	18881	45716
Candidatus Blochmannia	52745	10566	50659
Candidatus Kinetoplastibacterium	54805	23825	52488
Mycoplasma conjunctivae	56314	25375	53721
Mycoplasma haemocanis	61330	27333	58504
Mycoplasma hyopneumoniae	61405	27860	58921
Borrelia burgdorferi	61520	20095	56830
Mycoplasma fermentans	66665	29701	63682
Candidatus Azobacteroides	74280	33752	71043
Halyomorpha halys	76735	34591	73364
Chlamydia psittaci	77875	60659	83558
Chlamydophila psittaci	78145	60659	64572
Candidatus Midichloria	78915	35801	75439
Ehrlichia muris	79780	35042	75934
Candidatus Liberibacter	81820	35401	78289
Rickettsia akari	82070	23803	64338
Rickettsia rickettsii	84670	3928	126497
Rickettsia japonica	85535	4568	31451
Buchnera aphidicola	85595	39197	82023
Rickettsia philipii	85845	2545	17825
Rickettsia rhipicephali	86020	6747	35296
Ignicoccus hospitalis	86500	39094	82636
Ehrlichia canis	87665	39045	83737
Mycoplasma penetrans	90575	41106	86705
Thermoplasmatales archaeon	97405	43517	93247
Candidatus Endolissoclinum	98745	44587	94439
Thermocrinis albus	100035	45404	95585
Bartonella clarridgeiae	101515	45193	96787
Campylobacter lari	101695	45770	96995

Taylorella asinigenitalis	102674	45621	97706
Hydrogenobaculum sp. SN	103525	24	94736
Hydrogenobaculum sp. HO	103525	32	95144
Methanococcus aeolicus	104630	47028	99923
Thermoplasma volcanium	105650	47974	100912
Candidatus Nitrosopumilus	109330	48307	104317
Candidatus Methanomethylophilus	111115	50004	106239
Gardnerella vaginalis	111155	50022	106332
Nautilia profundicola	111760	50231	106792
Mycoplasma hyorhinis	111767	49282	106696
Pyrococcus yayanosii	114450	51303	109162
Pyrococcus horikoshii	115900	51983	110566
Helicobacter bizzozeronii	116424	52185	110974
Erysipelothrix rhusiopathiae	119195	53434	114037
Leuconostoc citreum	119750	52858	113637
Methanocorpusculum labreanum	120330	54625	115012
Thermotoga naphthophila	120650	9252	54395
Pyrobaculum islandicum	121760	52415	114370
Streptococcus pyogenes	122570	39421	119490
Thermoproteus tenax	122765	54735	117026
Pyrolobus fumarii	122880	55320	117418
Dictyoglomus turgidum	123700	55302	118103
Anaerococcus prevotii	125535	56708	119854
Lactobacillus sakei	125640	55923	119596
Melissococcus plutonius	126065	56304	120373
Leuconostoc gelidum	126230	41178	109866
Leuconostoc mesenteroides	126435	55714	120456
Streptococcus thermophilus	128660	49568	120180
Methanococcus voltae	129090	58356	123356
Bifidobacterium animalis	129235	58077	123747
Francisella cf.	129685	51168	115687
Lactobacillus johnsonii	131085	55242	122707
Chlorobium phaeovibrioides	131120	57657	124327
Thermovirga lienii	131180	58893	125539
Streptococcus lutetiensis	131700	40543	111332
Streptococcus constellatus	132740	34549	105319
Coxiella burnetii	133920	60151	128179
Pyrobaculum calidifontis	133950	60646	127848
Thermococcus barophilus	134005	60053	128227
Mycoplasma gallisepticum	134985	59176	128521
Methanobacterium sp. MB1	135315	60882	129319
Campylobacter concisus	136800	62099	130548
Zymomonas mobilis	137090	61461	131263
Helicobacter cinaedi	138555	62474	132283
Bifidobacterium adolescentis	139300	60130	131193
Candidatus Hamiltonella	140673	61950	134409

Lactobacillus kefiranofaciens	140865	53440	122761
Ammonifex degensii	141945	64290	135742
Streptococcus iniae	143320	62432	135502
Neisseria gonorrhoeae	143590	34852	117829
Calditerrivibrio nitroreducens	143855	65119	137621
Riemerella anatipestifer	144425	64079	138225
Streptococcus equi	144480	63699	138076
Streptococcus gordonii	146440	61305	136436
Streptococcus salivarius	147810	60524	142533
Rickettsia prowazekii	148230	55670	140649
Pasteurella multocida	150495	66158	143373
Haemophilus parasuis	151275	66903	143866
Methylacidiphilum infernorum	152475	68807	145971
Aggregatibacter actinomycetemcomitans	153901	67929	146379
Halanaerobium praevalens	153950	69211	147208
Cycloclasticus sp. P1	157545	20376	95287
Vulcanisaeta distributa	158275	71341	151230
Prochlorococcus marinus	160719	72349	153715
Thiomicrospira crunogena	161845	73285	154696
Deinococcus geothermalis	164480	74389	157154
Acetohalobium arabaticum	164635	74500	157412
Actinobacillus suis	165643	65972	150305
Lactobacillus buchneri	166700	74346	159096
Rothia dentocariosa	167065	75251	159391
Caldicellulosiruptor obsidiansis	168820	52400	142443
Xylella fastidiosa	169045	75707	161776
Methanoregula boonei	169525	76633	162182
Sulfurovum sp. NBC37-1	170815	77390	163231
Staphylococcus carnosus	171090	76152	162792
Sulfuricurvum kujiense	171650	76534	163627
Methanobacterium sp. AL-21	172250	77776	164704
Staphylococcus lugdunensis	173046	76575	164762
Halanaerobium hydrogeniformans	174205	78274	166696
Bartonella tribocorum	174600	73770	165013
Mannheimia haemolytica	177220	76034	169516
Nitrosomonas eutropha	177400	78469	168666
Caldicellulosiruptor lactoaceticus	178320	22664	110591
Gallibacterium anatis	179155	79618	170117
Sulfolobus islandicus	179490	71077	167115
Sulfolobus tokodaii	179650	81019	171948
Treponema succinifaciens	182120	82229	173373
Tepidanaerobacter sp. Re1	183990	83470	88240
Flavobacteriaceae bacterium	184540	83180	176431
Ketogulonicigenium vulgare	185070	82994	89820
Thermoanaerobacter wiegelii	185670	52692	139741
Acetobacter pasteurianus	187910	84337	179373

Halogeometricum borinquense	188035	84547	179702
Candidatus Nitrososphaera	188920	85452	180661
Kangiella koreensis	190135	85296	181773
Brachyspira pilosicoli	190549	84302	181853
Coprococcus sp. ART55/1	190669	82152	179522
Desulfovibrio desulfuricans	191560	86546	183309
Enterococcus sp. 7L76	194287	12925	82686
Methanosphaerula palustris	194860	87733	186160
Desulfurispirillum indicum	195225	87563	186221
Listeria ivanovii	195255	79875	180294
Thermacetogenium phaeum	195935	86425	185198
Haloferax mediterranei	196590	86689	187081
Lactobacillus plantarum	202975	89552	193419
Psychromonas sp. CNPT3	203490	91202	193848
Treponema brennaborense	203705	91520	194529
Deinococcus radiodurans	203998	91510	194717
Bacillus coagulans	204868	91557	195426
Meiothermus ruber	206495	92687	197869
Desulfurivibrio alkaliphilus	206515	93352	197445
Halorhabdus utahensis	207785	90465	197348
Nitrosospira multiformis	212280	95887	203057
Ruegeria sp. TM1040	213395	95511	203726
Candidatus Arthromitus	213755	93198	204120
Denitrovibrio acetiphilus	214805	96959	205458
Rubrobacter xylanophilus	215045	97619	205603
Faecalibacterium prausnitzii	215757	95172	204085
Meiothermus silvanus	216625	97155	206691
Rhodothermus marinus	217440	98703	207997
Spirochaeta africana	219055	99032	209399
Brucella melitensis	219624	4160	128380
Melioribacter roseus	220025	99353	210458
Corynebacterium glutamicum	220625	97410	210371
Azospirillum sp. B510	220755	98060	206866
Spirochaeta sp. Buddy	221095	99816	211238
Cyanobium gracile	222820	99998	212886
Sphingopyxis alaskensis	223010	100009	213012
Acidiphilium cryptum	225945	22359	140535
Brucella pinnipedialis	226615	2042	22565
Coprococcus catus	230502	98563	214895
Nitrosococcus oceani	232110	97994	217435
Legionella pneumophila	233329	106040	223444
Synechococcus sp. PCC	234015	106368	223858
Aequorivita sublithincola	234710	105953	224419
Desulfotalea psychrophila	234890	106491	224432
Allochromatium vinosum	235125	104583	224415
Sulfobacillus acidophilus	236745	107302	226627

Synechocystis sp. PCC	238000	107722	227761
Clostridium cf.	239574	104869	179426
Desulfovibrio gigas	243551	109449	232853
Pelobacter carbinolicus	244390	110355	233719
Hyphomonas neptunium	247000	111510	236318
Geobacter sulfurreducens	247615	111330	236559
Alistipes finegoldii	248945	103133	229509
Rhodobacter capsulatus	249260	109867	238068
Thermus thermophilus	250118	107855	238035
Dinoroseobacter shibae	252635	113347	241358
Natrinema pellirubrum	252695	109526	239860
Fibrobacter succinogenes	256200	115507	245077
Desulfobulbus propionicus	256790	115373	245670
Treponema azotonutricium	257040	116190	245584
Bacillus pseudofirmus	257265	115635	245142
Arthrobacter arilaitensis	257280	114964	245006
Erwinia tasmaniensis	258895	109540	242122
Haloterrigena turkmenica	259265	113519	246874
Asticcacaulis excentricus	260270	117541	248770
Thalassolituus oleivorans	261355	118303	249751
Vibrio cholerae	262585	116373	253768
Desulfomicrobium baculatum	262840	119099	251229
Gluconacetobacter diazotrophicus	262940	117112	250858
Desulfococcus oleovorans	262940	119060	251257
Bdellovibrio bacteriovorus	265905	120573	254342
Clostridium botulinum	266172	120087	254798
Phenylobacterium zucineum	266415	119200	254326
Thioalkalivibrio nitratireducens	266820	120121	254944
Pelobacter propionicus	267200	120103	254927
Caulobacter crescentus	267795	115141	253535
Bacillus subtilis	269580	90101	255672
Vibrio anguillarum	270113	118985	130975
Leadbetterella byssophila	270640	122164	259110
Bacillus cytotoxicus	272465	118041	255999
Cyanobacterium aponinum	274270	123840	262034
Halobacillus halophilus	276705	124359	264296
Streptococcus pneumoniae	278124	82971	261979
Sphingobium sp. SYK-6	279955	125403	267254
Vibrio fischeri	281850	124513	267738
Cellulomonas fimi	284420	125379	270723
Cronobacter sakazakii	284575	107228	256407
Peptoclostridium difficile	286015	125259	93035
Shigella dysenteriae	286830	24578	138835
Granulicella tundricola	287275	128980	274544
Eubacterium limosum	287780	129853	275046
Desulfitobacterium dehalogenans	288115	126043	273787

Saprospira grandis	289680	130340	276918
Pantoea sp. At-9b	291245	126737	275713
Azoarcus sp. BH72	291735	128371	278002
Mycobacterium bovis	291780	1431	219358
Tsukamurella paurometabola	291990	130159	278616
Rhizobium etli	292105	122134	276923
Oscillibacter valericigenes	294000	131972	280966
Marinobacter adhaerens	294790	131710	281176
Neisseria meningitidis	301515	84304	301085
Cellvibrio japonicus	305100	137143	291440
Lactococcus lactis	307730	130114	294563
Lysinibacillus sphaericus	309320	139067	294621
Spirochaeta smaragdinae	310260	140284	296787
Porphyromonas gingivalis	312213	138104	296783
Yersinia pseudotuberculosis	312625	28539	249120
Arthrobacter sp. FB24	313260	139051	298288
Geobacter sp. M21	316385	125779	291477
Ochrobactrum anthropi	318875	141664	303603
Bifidobacterium longum	319030	129766	302567
Shigella sonnei	321680	10978	96625
Shewanella frigidimarina	323015	143350	306836
Desulfosporosinus meridiei	324900	145220	309788
Bacillus infantis	325645	146790	310560
Xanthomonas oryzae	329345	132388	321069
Campylobacter jejuni	332565	134823	308012
Syntrophobacter fumaroxidans	332680	149560	317112
Vibrio vulnificus	333845	147432	318271
Rubrivivax gelatinosus	336215	148290	320289
Glaciecola sp. 4H-3-7+YE-5	336820	149647	319822
Mycobacterium abscessus	337810	151217	249119
Pseudomonas mendocina	338185	143321	320478
Bordetella bronchiseptica	339450	36919	209784
Erwinia billingiae	340010	149256	323058
Marinomonas sp. MWYL1	340020	152664	324299
Propionibacterium acnes	341365	153285	325751
Serratia liquefaciens	349240	142103	325594
Chloroflexus aurantiacus	350565	418	168662
Geobacter sp. M18	351825	154763	334844
Enterobacter aerogenes	352015	142216	328538
Azotobacter vinelandii	357685	155968	340758
Azorhizobium caulinodans	357980	158987	341816
Mycobacterium intracellulare	360160	32003	379910
Herbaspirillum seropedicae	367590	163507	350977
Mycobacterium gilvum	369845	158095	352884
Verminephrobacter eiseniae	371115	164833	354341
Mycobacterium indicus	372600	24479	192925

Thermomonospora curvata	375930	167605	358808
Gordonia polyisoprenivorans	377985	168351	360663
Salinispora arenicola	385755	166736	366554
Microcystis aeruginosa	389515	176060	372269
Paenibacillus polymyxa	390965	172095	377554
Cupriavidus taiwanensis	394615	155377	372645
Vibrio harveyi	397950	172692	189283
Klebsiella oxytoca	398267	165238	370933
Planctomyces brasiliensis	400440	181097	382837
Lactobacillus casei	400585	50475	337664
Mycobacterium sp. JLS	403225	51835	226047
Nakamurella multipartita	404015	181156	385803
Paenibacillus terrae	405555	173677	383604
Pseudomonas syringae	406245	164499	386474
Mesorhizobium australicum	413365	159156	368550
Pseudomonas resinovorans	419055	178182	397180
Photobacterium profundum	421451	188598	401798
Vibrio nigripulchritudo	421470	188689	401560
Pseudomonas aeruginosa	426840	181797	407292
Streptococcus agalactiae	428985	179924	397657
Rhodococcus erythropolis	434420	195064	414802
Nostoc sp. PCC	442335	196691	422453
Verrucosispora maris	444930	195030	424068
Burkholderia thailandensis	448750	167766	418326
Fibrella aestuarina	451132	204717	431303
Achromobacter xylosoxidans	460071	196518	434092
Micromonospora sp. L5	464165	63775	295651
Ralstonia eutropha	464300	184226	416668
Paenibacillus sp. Y412MC10	474775	213393	453481
Sinorhizobium meliloti	488610	205080	465244
Burkholderia ambifaria	498990	172478	436030
Leptospira borgpetersenii	520525	232714	497350
Clostridium acetobutylicum	525405	235613	501629
Ralstonia pickettii	541715	216344	511738
Haliscomenobacter hydrossis	558110	252552	533840
Burkholderia sp. 383	578410	448926	508435
Lactobacillus rhamnosus	590465	261754	563759
Bradyrhizobium japonicum	613825	230752	581585
Yersinia pestis	619820	24603	636387
Actinoplanes friuliensis	625070	277527	596600
Streptomyces davawensis	630980	267682	594049
Corynebacterium diphtheriae	652545	292759	622984
Xanthomonas axonopodis	674720	678	391309
Amycolatopsis mediterranei	683125	303371	652118
Bacillus cereus	694505	89382	460734
Chlamydia trachomatis	696474	309815	664970

Cyanothece sp. PCC	707495	178231	706988
Streptococcus suis	712425	313606	677256
Burkholderia mallei	737855	15266	559415
Helicobacter pylori	747120	326888	714555
Methylobacterium extorquens	749930	315799	710969
Corynebacterium pseudotuberculosis	769040	335284	732628
Bacillus amyloliquefaciens	811805	359460	773639
Alteromonas macleodii	916840	406818	873041
Staphylococcus aureus	940236	405578	895267
Burkholderia pseudomallei	966140	126410	1071536
Acinetobacter baumannii	1117118	453326	1061030
Escherichia coli	1274363	155814	1434892
Listeria monocytogenes	1369713	540035	1308560
Shewanella baltica	1731640	739831	1639025
Mycobacterium tuberculosis	1760555	11524	1495156
Salmonella enterica	2243541	877220	2128983

Le tableau est classé par la masse d'ADN génomique (ADNg) croissante.

Annexe VIII

Contenu simplifié du dossier après le fonctionnement de ICoMiO sur les données simulées

Les fichiers issus d'une même étape sont surlignés d'une même couleur : FastQC, Bowtie pour retirer les reads connus, Kraken, Velvet, BLAST et Bowtie pour la classification des reads. Les caractères en italiques sont à remplacer par les options du fichier de configuration :

- ICoMiO_simData.cfg
- listOrgaToRemove.txt
- simData.fq
- run_1/
    - Logs/
        - bowtie_mapping*DBname*.out
        - fastqc.out
        - kraken.out
        - velvet_*k*-mer.out
    - Velvet_Assembly_k*k*/
        - Graph2
        - LastGraph
        - Log
        - PreGraph
        - UnusedReads.fa
        - contigs.fa
        - stats.txt
    - Contigs_BLAST-*DBname*.tab
    - Contigs_BLAST-*DBname*.stats.tab
    - Contigs_BLAST-*DBname*.filtered.tab
    - simData_fastqc.html
    - simData_fastqc.zip
    - simData_bowtie-*options*-AgamP3-Haplotypes.sam
    - simData_bowtie-*options*-withoutAgamP3-Haplotypes.fq
    - simData_bowtie-*options*- withoutAgamP3-Haplotypes_bowtie-*options*-Pberghei.sam
    - simData_bowtie-*options*- withoutAgamP3-Haplotypes_bowtie-*options*-withoutPberghei.fq
    - simData_bowtie-*options*-withoutAgamP3-Haplotypes_bowtie-*options*-withoutPberghei-kraken.output
    - simData_bowtie-*options*-withoutAgamP3-Haplotypes_bowtie-*options*-withoutPberghei-kraken.report
    - simData_bowtie-*options*-withoutAgamP3-Haplotypes_bowtie-*options*-withoutPberghei-kraken.mpa-report
    - simData_bowtie-*options*-withoutAgamP3-Haplotypes_bowtie-*options*-withoutPberghei-kraken.krona
    - simData_bowtie-*options*-withoutAgamP3-Haplotypes_bowtie-*options*-withoutPberghei-kraken.krona.html

- simData_bowtie-*options*-withoutAgamP3-Haplotypes_bowtie-*options*-withoutPberghei-kraken-unclassified-out.fq
- simData_bowtie-*options*-withoutAgamP3-Haplotypes_bowtie-*options*-withoutPberghei-kraken-unclassified-out_bowtie-options-*DBname*.sam
- simData_bowtie-*options*-withoutAgamP3-Haplotypes_bowtie-*options*-withoutPberghei-kraken-unclassified-out_bowtie-options-*DBname*.stats.tab
- simData_bowtie-*options*-withoutAgamP3-Haplotypes_bowtie-*options*-withoutPberghei-kraken-unclassified-out_bowtie-options-*DBname*.filtered.tab

UNIVERSITÉ DE STRASBOURG

Anaïs Vittu

Inserm
Institut national
de la santé et de la recherche médicale

Outils bioinformatiques pour l'analyse génétique de la résistance du moustique *Anopheles gambiae* vis-à-vis des parasites du paludisme

Résumé

Au cours de ma thèse, j'ai développé et mis en place de nouvelles méthodes utilisant les récentes technologies du séquençage à très haut débit, des outils bioinformatiques et du « reciprocal allele-specifique RNA interference » (rasRNAi) dans l'objectif d'identifier les facteurs génétiques et non génétiques responsables de la résistance du moustique *Anopheles gambiae* aux parasites du paludisme murin *Plasmodium berghei*.

J'ai mis en place une stratégie d'identification des polymorphismes dans les lignées résistantes et sensibles afin de sélectionner des marqueurs génétiques pour de futures analyses génétiques et lister les gènes polymorphiques.

J'ai contribué à l'élaboration de nouvelles sondes ARN double-brin (dsRNAs) allèle-spécifique pour la méthode du rasRNAi en identifiant le processus de découpage du dsRNA injecté chez des moustiques par l'analyse des petits ARNs séquencés issus du dsRNA injecté.

J'ai élaboré un pipeline pour identifier la composition du microbiote des lignées sensibles et résistantes dans le but de les comparer.

Mots-clés : *Anopheles gambiae*, paludisme, facteurs de résistance, séquençage à très haut débit, bioinformatique

Summary

During my PhD, I developed and implemented new methods and tools using the latest technologies of the Next Generation Sequencing, bioinformatics tools and the « reciprocal allele-specific RNA interference » (rasRNAi) method with the aim of identifying genetic and non-genetic factors responsible for the resistance of the mosquito *Anopheles gambiae* to the mouse malaria parasites *Plasmodium berghei*.

I have implemented a strategy for identifying polymorphisms in the resistant and susceptible lines to (1) select genetic markers for future genetic analysis and (2) list the polymorphic genes.

I contributed to the development of a new allele-specific dsRNA probe for the rasRNAi method by identifying how mosquitoes process the injected dsRNA by the analysis of sequenced small RNAs from the injected dsRNA.

I developed a pipeline to identify the microbiota composition in susceptible and resistant lines in order to compare them.

Keywords: *Anopheles gambiae*, malaria, resistance factors, next generation sequencing, bioinformatics