A plusieurs, on est meilleur : du rôle des duplications dans l'adaptation aux insecticides chez les moustiques.

Jean Loup Imbert Claret

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Ecologie, Evolution, Ressources Génétiques, Paléobiologie

Ecole doctorale GAIA

Unité de recherche ISEM

Titre de la thèse

A plusieurs, on est meilleur : du rôle des duplications dans l'adaptation aux insecticides chez les moustiques.

Présentée par Jean-Loup CLARET
Le 21 Décembre 2023

Sous la direction de Pierrick LABBE

Devant le jury composé de

Carole SMADJA, DR, ISEM, CNRS-Montpellier	Présidente du jury
Emmanuelle LERAT, CR, LBBE, CNRS-Lyon	Examinatrice
Claire MEROT, CR, ECOBIO, CNRS-Rennes	Examinatrice
Pierrick LABBE, PR, ISEM, Université de Montpellier	Directeur de thèse
Jean-Philippe DAVID, DR, LECA, CNRS-Grenoble	Rapporteur
Julien VARALDI, MC, LBBE, Université de Lyon	Rapporteur

UNIVERSITÉ DE MONTPELLIER

Thèse pour obtenir le grade de Docteur

de l'Université de Montpellier

Ecole doctorale GAIA

A plusieurs, on est meilleur : du rôle des duplications dans l'adaptation aux insecticides chez les moustiques.

Présentée par Jean-Loup CLARET
Sous la direction de Pierrick LABBE et Pascal Milesi

Remerciements

Né dans le milieu des années 90, j'ai eu le plaisir de grandir dans une culture où fleurissait une pléthore d'univers fantastiques et de science-fiction. Je m'y suis plongé avec bonheur, n'y appréciant dans un premier temps que la sensation d'émerveillement et de découverte qu'ils me procuraient, pour ne réaliser que plus tard leur caractère formateur et le message qu'ils recèlent. Vous vous en serez rendu compte, la grande majorité de ces œuvres suivent une trame commune, rapidement identifiable. Ce sont des récits initiatiques : ils détaillent les tribulations d'un héros ou d'une héroïne aux milles visages, d'une même figure, toujours naïve et souvent bien mal préparée, lancée dans une grande quête qui la marquera immanquablement. Ces aventures lui permettent de mûrir, de s'affirmer, et finalement de se muer en une version plus réaliste et plus humaine de ce qu'elle aspirait à devenir au début de son périple. Ce sont de belles métaphores de développement personnel, du passage à l'âge adulte, couvertes du verni vibrant des mondes fantastiques. Quand je repense à ces trois dernières années, mais aussi par extension à toutes celles qui me séparent maintenant de mon entrée à l'Université, je ne peux m'empêcher d'y voir une restitution moderne de ces récits que j'ai dévoré enfant (mais évidemment bien plus tranquille et moins dangereuse) ! Finalement, cette thèse m'aura à mon tour permis de m'embarquer dans ma propre quête, mon voyage initiatique vers un but lointain et ardemment désiré : l'obtention d'un doctorat.

Si ce préambule de remerciements parle autant de moi, c'est bien parce qu'il existe un autre point commun à tous les récits auxquels je compare mon expérience : le ou la protagoniste n'est jamais seul-e. C'est toujours grâce à l'aide et au soutien de son entourage, de quelques ami-e-s fidèles et compagnons de route, que sa quête aboutit et que le récit trouve sa fin heureuse. Je veux donc remercier tous les membres de ma Communauté de l'Anneau, les sœurs et frères d'armes de mon Alliance Rebelle, qui ont rendu la rédaction de ce manuscrit et l'aboutissement de ces études possibles.

J'ai longtemps cherché comment organiser ces remerciements, par qui commencer et finir, quel chemin suivre pour m'adresser à chacun et chacune et être bien certain de n'oublier personne. Sauf que voilà, il m'a vite fallu me rendre à l'évidence, la tâche est immense et je suis bien trop désorganisé… Je ferai tout de même de mon mieux, et vous recevrez une suite de "merci" qui peut sembler sans queue ni tête, mais qui soyez-en assuré-e-s, vient tout droit du cœur. Aux éventuels-les oublié-e-s et immanquables anonymes à qui je dois pourtant beaucoup, j'adresse ces quelques mots empruntés à un homme petit par la taille, mais grand par le courage : "Je ne remercierai pas la moitié d'entre vous aussi bien que je le voudrais, et je remercie moins de la moitié d'entre vous aussi bien que vous le mériteriez".

Je commencerai par ma famille : ma mère, ma tante et mon oncle. Je vous suis tellement reconnaissant de m'avoir instillé l'envie de comprendre, d'avoir encouragé ma curiosité. Plus encore, vous avez su me faire confiance : en partant de la ferme de l'Hort, jusqu'aux chantiers du village d'Argelliers, pour enfin finir sur les bancs de l'Université ; vous m'avez laissé faire mes choix et m'avez soutenu dans tous mes projets. C'est bien grâce à cette confiance que j'ai pu m'épanouir et me construire. Merci !

J'en profite aussi pour remercier ceux qui m'ont donné du travail avant que je ne me lance dans les études. Simon, François, Jean-Marc et Romain : chacun à votre tour, vous avez tous su me transmettre un peu de la passion que vous avez pour votre métier, tout en faisant

preuve d'une incroyable gentillesse. Je chéris chacune de ces expériences et les souvenirs des moments passés ensemble. En attendant de vous revoir autour d'un bon repas : la bise !

Un grand merci à l'Université des Sciences de Montpellier, dans laquelle je rôde depuis presque dix ans (bonjour le coup de vieux). Je lui dois, entre autres, mes plus belles rencontres. D'abord de formidables enseignants (on y reviendra d'ailleurs), mais aussi et surtout des personnes sans lesquelles j'ai du mal à m'imaginer le quotidien.

Honorine, Maxime et Grégoire, je vous dédicace ce cliché poncif, mais dont vous avez su me montrer tout le sens : il y a la famille dans laquelle on naît et celle que l'on se choisit. Merci d'avoir été présents tout au long de ces années et de continuer à l'être quelle que soit la distance qui nous sépare.

A celle qui me supporte (dans tous les sens du terme) depuis déjà plus de cinq ans. Marion, merci d'être la bouffée d'air frais et le rayon de soleil dont j'ai tant besoin (et notamment lors de ces derniers mois de thèse, pour le moins… compliqués). On remet la même pour cinq ans de plus, qu'est-ce que tu en dis ?

Viennent ensuite les Darwins : pour tous les cours que nous avons suivis ensemble, maintenant devenus des congrès ; les fiches de révisions partagées qui sont à présent des scripts et des articles ; les soirées de jeux qui sont… eh bien, toujours les soirées de jeux, et c'est très bien comme ça ! Je n'aurais pû rêver d'un meilleur groupe avec lequel me plonger dans les méandres de la biologie évolutive. Merci pour tout, les copains, mais surtout, à tout bientôt.

Pour l'ISEM, un labo où il a fait bon travailler. Merci à toutes les personnes qui aident à en faire un endroit où l'on est heureux d'arriver le matin. Aux doctorants avec qui j'ai pu boire (un peu) et me plaindre (beaucoup). A Arthur pour tous les regards entendus qui suivaient le traditionnel "Comment va ?" ; à Marie pour les kilomètres parcourus lors de nos balades sur le campus, son oreille attentive et ses jolies photos d'oiseaux…

Enfin à mon équipe au grand complet, pour les apiculteurs, les fans de zombies et les aficionados de course et de foot. Merci pour les petits mots qui remontent le moral quand vous me voyiez me ratatiner devant mon ordinateur. Merci à Sandra et Patrick pour toute leur aide en manips. Merci à Camille qui m'a sauvé la mise plus d'une fois quand un script s'avérait bien trop compliqué pour moi, et toujours avec le sourire… Merci aussi à Alice pour les innombrables coups de main, mais surtout pour ta bienveillance et les petits "Ca va, tu t'en sors ?" de ces derniers mois.

I would also like to take this opportunity to give my deepest thanks to Lindy McBride and Yuki Haba. You are both amazing researchers and people, and without you, my PhD would just not have been the same. Yuki, I won't forget your teachings: おまえまじ やばい, ね!

Au tour des grands chefs !

Mylène, toi qui m'impressionnait déjà tant avant la thèse, je t'ai vu affronter des épreuves avec une force et un courage que j'admire profondément. Malgré tout, tu as continué à te soucier des autres, tu m'as guidé et conseillé. Merci pour tout ce que tu fais pour l'équipe, c'est en grande partie grâce à toi qu'on s'y sent si bien. Des bisous !

Pascal, merci pour ton enthousiasme contagieux (si, si, je te jure, j'étais enthousiaste aussi !). Merci de ta patience quand tu as réalisé que, si je comprenais vite, il fallait quand

même m'expliquer longtemps. Tu as sû jouer ton rôle à la perfection, alliant gentillesse, boutades et *pep talks* de main de maître, un véritable Pascal le grand frère (pardon). J'ai du mal à exprimer toute la gratitude que je ressens quand je repense à ton accueil lors de ce (trop) court séjour en Suède. Mais bon, heureusement pour moi, il semblerait que tu n'apprennes pas de tes erreurs et que tu aies décidé de supporter mes bêtises encore un moment. J'ai hâte de voir ce que la suite nous réserve !

Enfin, le mot de la fin sera pour Pierrick. Comment dire simplement tout ce que je te dois ? Tu es la personne qui m'a donné envie de me lancer dans la recherche (même si un deuxième Bourguignon -lui aussi capillairement minimaliste d'ailleurs- n'est pas étranger à cette décision). Tu m'as aussi motivé à enseigner, en me montrant qu'en y mettant un peu de *showmanship*, tout est tellement plus intéressant (même les modèles de dynamique des populations, c'est dire !). C'est encore toi qui m'as donné l'opportunité de réaliser toutes ces envies en me prenant en thèse, et tu m'as aidé à chaque pas de ce long chemin. Je garde en tête toutes les discussions dont je ressortais remotivé et les conseils que j'ai essayé tant bien que mal d'appliquer… Grâce à toi, j'ai eu le plaisir de passer trois années auprès d'un directeur de thèse avec qui parler de science, bien sûr, mais aussi d'histoire, de religion, de sociologie, de politique, de sport (merci de m'avoir refilé la fièvre rugby !) et même de généalogie des rois du Second Âge du Silmarillion (si, si !). Merci d'avoir été, largement plus qu'un directeur de thèse, un véritable ami.

Introduction

I. Adaptation et variations du nombre de copies de locus

Le processus d'adaptation peut être observé simplement quand l'altération d'un milieu entraîne des modifications au sein des populations[1] qui l'occupent. Ces modifications sont dues au tri de la diversité génétique par la sélection naturelle, et s'observent donc au fil des générations. Les variants les plus avantageux augmentent progressivement en fréquence, jusqu'à envahir complètement la population (Darwin, 1859). Devant cette apparente simplicité, un lecteur non-informé[2] pourrait se demander pourquoi l'étude de l'adaptation déchaîne les passions et motive toute une communauté scientifique à faire couler tant d'encre (et de larmes). Donner une réponse satisfaisante à cette question mériterait bien plus qu'une thèse et n'est de toute manière pas l'objectif de la mienne. Je vais donc m'en tenir à expliciter quelques notions générales sur l'adaptation et la sélection naturelle qui serviront mon propos tout au long de ce manuscrit, ce qui devrait d'ailleurs suffire pour laisser entrevoir la complexité réelle de ces processus.

Variants structuraux et diversité génétique.

La diversité génétique est générée par des mutations. Elles peuvent prendre plusieurs formes, depuis l'altération d'une unique base d'ADN (Single Nucleotide Polymorphisms -*SNPs*), jusqu'au remaniement de larges séquences génomiques (Structural Variants -*SVs*). Parmi ce dernier type de mutations, on distingue les modifications de la position génomique d'un segment d'ADN ou le changement de son orientation (translocations et inversions respectivement), et celles qui affectent le nombre de copies de locus (Copy Number Variations -*CNVs*), qui comprennent les délétions (suppression de locus) et les duplications (doublement de séquences)[3]. On rencontre aussi l'appellation d'*indels* (contraction d'insertion-délétion) pour des *CNVs* de petite taille.

Quel que soit le type de mutation, la diversité génétique qui en résulte est capitale, surtout parce qu'elle sert de réservoir pour l'adaptation, et est donc au cœur des sciences de l'évolution. Les chercheur-euse-s dans ce domaine se sont longtemps concentré-e-s sur l'étude des *SNPs*, principalement du fait de la difficulté inhérente à la détection des *SVs* (sujet

[1] J'utilise à dessein le terme de population plutôt que celui d'espèce, puisqu'il rend mieux compte de l'échelle à laquelle l'adaptation joue principalement : c'est un processus local, qui affecte un groupe d'individus interféconds soumis aux mêmes conditions environnementales.

[2] Ce que vous n'êtes bien sûr pas, très cher-e lecteur-rice ! Mais on ne sait jamais entre quelles mains pourrait tomber cette thèse...

[3] On n'abordera pas ici le cas de la polyploïdie, qui résulte de duplications du génome complet ; c'est (très) intéressant mais trop éloigné de ce que j'ai étudié.

sur lequel je reviendrai plus tard, en longueur, croyez-moi!). L'existence des *SVs* a pourtant été documentée relativement tôt dans l'histoire de la discipline, notamment avec les travaux pionniers de Sturtevant sur les réarrangements chromosomiques chez la drosophile : il a étudié une mutation du gène *bar* qui entraîne une modification de la forme des yeux en une barre verticale (d'où le nom), modification qui s'aggrave encore lorsque l'allèle muté est dupliqué (Sturtevant, 1925 ; voir aussi Wolfner & Miller, 2016 pour un résumé des travaux de Sturtevant sur le gène *bar*). Sturtevant avait noté la rapidité étonnante avec laquelle les changements du nombre de copies (duplications et délétions de réversion) du gène *bar* s'opéraient, estimant la fréquence de ces mutations à 10^{-3} par génération (Sturtevant, 1925). Ces découvertes seront suivies par celle des éléments transposables (*TE - Transposable Elements*) par Barbara McClintock en 1931[4] : ces séquences d'ADN ubiquistes ne sont pas à proprement parler des *SVs*, mais elles se transposent (de manière autonome ou en utilisant la machinerie cellulaire du génome hôte) à différents points du génome, créant ainsi des séquences qui favorisent l'apparition de *SVs* par recombinaison homologue. Ainsi l'activité des *TEs* induit la formation de *SVs* potentiellements adaptatifs[5] (voir Schrader & Schmitz, 2019 pour revue) et serait à l'origine d'une part importante des variations de taille et de complexité des génomes à plusieurs échelles taxonomiques (Tenaillon *et al.* 2011).

Pendant les années qui suivirent, l'étude des *SVs* fut toutefois reléguée au second plan, l'essentiel des travaux se concentrant sur les *SNPs*. Il faudra attendre les années 2000, pour que l'on assiste à un regain d'intérêt pour les *SVs,* largement motivé par le développement de nouvelles techniques de séquençage (*new generation sequencing - NGS*). Ce bond technologique permet l'analyse de génomes entiers, et rend également possible l'étude de populations naturelles d'espèces non-modèles. Ainsi, de nouvelles preuves viennent étayer les observations de Sturtevant et confirment la fréquence élevée d'apparition des *SVs* (Sebat *et al.*, 2004 ; Iafrate *et al.*, 2004 ; Emerson *et al.*, 2008). En parallèle, des études d'accumulation de mutations mesurent directement les taux d'apparition des *SVs* (Lynch et al., 2008 ; Lipinski et al., 2011 ; Schrider et al., 2013). Ce type d'analyse est libéré de certaines contraintes liées à l'étude de populations naturelles (*e.g.*, elles permettent de limiter l'impact de la sélection naturelle, de la dérive génétique et du biais d'échantillonnage), et elles ont permis d'apporter la preuve que les *SVs* ont une fréquence d'apparition notablement plus élevée que celle des *SNPs* : les évènements de duplication notamment seraient en effet

[4] Ce qui lui vaudra un prix Nobel, bien des années plus tard et après de nombreuses péripéties, en 1983.
[5] L'essentiel de leur impact semble délétère, ils peuvent par exemple altérer voire complètement empêcher l'expression des gènes en s'insérant dans leur séquences codantes (Kidwell & Lisch, 2000 ; 2001)

d'un ordre de grandeur plus fréquents que les *SNPs* (Schrider et al., 2013 ; Katju & Bergthorsson, 2013). Une part plus importante qu'initialement supposée de la diversité génétique est ainsi composée de duplications (voir plusieurs exemples chez les primates : Hansaker *et al.,* 2015 ; Lauer & Gresham 2019 pour une revue).

Sélection : les cas "simples"…

Nous avons vu de quelle manière est générée la diversité génétique, et j'ai déjà précisé que l'adaptation résulte de son tri par la sélection, mais comment s'effectue-t-il ? L'effet de la sélection sur chaque mutation est fonction de son l'impact sur l'ensemble des caractéristiques observables (le phénotype), ou plus concrètement sur la valeur sélective[6] (*fitness*) qui y est associée. Avant de traiter du devenir d'une mutation sous l'effet de la sélection, il faut d'abord s'intéresser à l'efficacité de cette dernière. Un paramètre populationnel influe justement sur cette efficacité : c'est la taille efficace (N_e) de la population. N_e est un concept faisant référence au nombre d'individus d'une population théorique dont les caractéristiques génétiques seraient similaires à celles de la population observée. De manière plus pratique, N_e peut être résumée au nombre d'individus reproducteurs génétiquement distincts composant une population. Elle permet donc de calculer le taux de changement de fréquences alléliques causés par des processus stochastiques (que l'on regroupe sous l'appellation de dérive génétique) dans une population idéale telle que décrite dans le modèle de Wright-Fisher[7] (Wright, 1931 ; mais voir aussi Charlesworth, 2009 et les articles qui y sont cités). La sélection aura un impact majeur sur le changement des fréquences alléliques dans une population de grande taille et génétiquement diverse, quand à l'inverse l'évolution d'une population de faible effectif ou composée d'individus apparentés (donc génétiquement peu distincts) sera plus sujette à la dérive génétique. Par conséquent, une réduction drastique de N_e (on parle de goulot d'étranglement ou *bottleneck ;* voir Fig. Int.1.a) peut entraîner l'élévation à de fortes fréquences, voire la fixation par dérive, de certains allèles y compris délétères (*e.g.* fréquence élevée de la rétinite pigmentaire sur l'île de Tristan de Cûnha par effet de fondation, Thompson, 1978).

Considérons maintenant une mutation arrivant dans une population suffisamment grande et diverse : son devenir sera alors essentiellement dicté par son impact phénotypique. Même dans ces conditions, la dérive pourra toutefois jouer un rôle déterminant et entraîner la

[6] La valeur sélective d'un variant est une mesure relative de son succès reproducteur moyen par rapport à celui du reste de la population, dans un environnement donné et sur plusieurs générations.

[7] Soit une population qui n'est pas affectée par la sélection, la migration et la mutation, panmictique, de taille constante, et aux générations non chevauchantes.

disparition précoce du variant, notamment juste après son apparition, quand la mutation est rare. Une seconde condition importante est que cette mutation soit héritable (transmise aux descendants), sans quoi elle sera irrémédiablement perdue à la mort de son porteur.

Jusqu'à présent, j'ai résumé de façon très binaire l'impact d'une mutation en la qualifiant de délétère ou d'avantageuse. En réalité, l'impact d'une mutation sur la valeur sélective n'est pas fixe, mais dépend d'un équilibre complexe entre sélection et dérive (théorie quasi-neutraliste de Ohta, 1973). Prenons dans un premier temps l'exemple d'une mutation ponctuelle modifiant une séquence codante d'ADN : si elle modifie la chaîne d'acide aminés (on la qualifie alors de mutation non-synonyme), elle peut donc altérer la fonction initiale de la protéine codée. Ce changement est le plus souvent délétère (~90% de *SNPs* non-synonymes délétères d'après l'expérience d'accumulation de mutations de Schrider *et al*., 2013). C'est notamment le cas quand la mutation entraîne un mauvais repliement de la protéine la rendant inactive, voire toxique comme dans certaines formes de la maladie d'Alzheimer, de variants de Creutzfeldt–Jakob ou d'amyloses familiales (Dobson, 2003). Dans ces cas, la sélection (alors qualifiée de purifiante) entraînera la diminution en fréquence de la mutation jusqu'à sa disparition à plus ou moins long terme. Des allèles délétères peuvent cependant être maintenus longtemps dans les populations, notamment s'ils sont récessifs (*i.e.* ils ne s'expriment qu'à l'état homozygote) et se trouvent "cachés" à la sélection dans les génotypes hétérozygotes. Si au contraire, une mutation affecte le phénotype de manière positive (*i.e.* sa valeur sélective s'en trouve augmentée par rapport à la moyenne de la population), sa fréquence augmentera jusqu'à ce qu'elle envahisse la population, on parle alors de sélection positive. Notez que dans le cas de gènes pléiotropes, c'est-à-dire s'exprimant à travers plusieurs traits, la sélection d'une mutation est dépendante de ses impacts dans l'ensemble des différents traits sur lesquels elle agit (*i.e.* la mutation peut modifier favorablement un trait, mais entraîner des effets néfastes sur d'autres)[8]. Au-delà de ces compromis évolutifs (*trade-off*) entre traits, la sélection d'une mutation dépend aussi de l'environnement dans lequel elle ségrège : une mutation avantageuse dans un milieu donné peut ainsi se révéler délétère dans un autre. La sélection est donc le résultat d'une balance complexe à différentes échelles (spatiales, temporelles) et à différents niveaux d'intégrations (génome, organisme, population, *etc*…).

[8] Pour citer le dessinateur Boulet dans sa planche intitulée *Pléiotropie* : "un gène ce n'est pas un programme dédié à un seul truc. Le génome c'est un gros bordel de plein de corrélations et d'interconnexions. C'est comme un gros Rubik's Cube : déplacer un carré peut changer les six faces. "

Au cours de ma thèse, je me suis surtout intéressé à l'échelle génomique. En effet, les variants génétiques de différents locus présents sur un même chromosome sont transmis ensemble à la descendance, on parle de liaison génétique entre variants. En échangeant des séquences d'ADN entre chromosomes homologues, la recombinaison "casse" ces liaisons : plus deux sites sont physiquement éloignés sur un chromosome, plus la probabilité qu'un événement de recombinaison les sépare est élevée (et inversement). En affectant la fréquence d'un allèle, l'impact de la sélection ne se limite donc pas seulement à la position du variant sélectionné, mais peut s'étendre à de large zones génomiques et entraîner tous les variants proches. On parle de balayage sélectif (*selective sweep* ; Fig. Int.1.b) pour désigner le fait que la sélection positive d'un allèle avantageux à un locus donné provoque l'augmentation en fréquence de variants présents sur d'autres locus liés; à l'inverse, la perte de variants liés à une mutation délétère est appelée sélection d'arrière-plan (*background selection* ; Fig. Int.1.c). Comme pour les *bottlenecks* dont je parlais plus tôt, les *selective sweeps* et le processus de *background selection* laissent des empreintes sur le génome (Fig. Int.1), qui sont autant d'indices précieux que les biologistes de l'évolution peuvent utiliser pour reconstruire l'histoire évolutive et démographique des espèces. Toutefois, les balayages sélectifs sont des phénomènes transitoires ; de la même manière, les environnements ne sont pas stables et les conditions démographiques fluctuent. En conséquence, la diversité génétique originale est susceptible d'être rétablie après un événement de sélection, et les signaux de son activité s'érodent avec le temps (Fig.Int.1, court/long terme). La vitesse de cette érosion dépend des taux de mutation et de recombinaison, de l'intensité de la dérive génétique et du flux de gènes (migration), autant de facteurs qui affectent les fréquences alléliques dans une population.

L'impact de la sélection naturelle ne se limite pas à une réduction de la diversité génétique, elle peut également participer activement au maintien du polymorphisme : c'est une forme de sélection dite "balancée" (Fig. Int.1.d), que l'on sous-divise encore en plusieurs types en fonction des mécanismes impliqués. La sélection fréquence-dépendante est le processus de sélection balancée le plus courant, où un variant n'est avantageux que lorsqu'il est rare dans la population. Les différentes stratégies reproductives des lézards à flanc maculés (*Uta stansburiana*) illustrent bien ce type de sélection : trois phénotypes associés à des types de comportements distincts sont maintenus chez les mâles de cette espèce grâce à un système de pierre-feuille-ciseaux (Sinervo & Lively, 1996). Les mâles à gorge bleue établissent de petits territoires et gardent activement leurs femelles ; les mâles à gorge orange envahissent le territoire des autres mâles à la recherche de femelles ; les mâles à gorge jaune s'introduisent sur le territoire des autres mâles en se faisant passer pour des femelles.

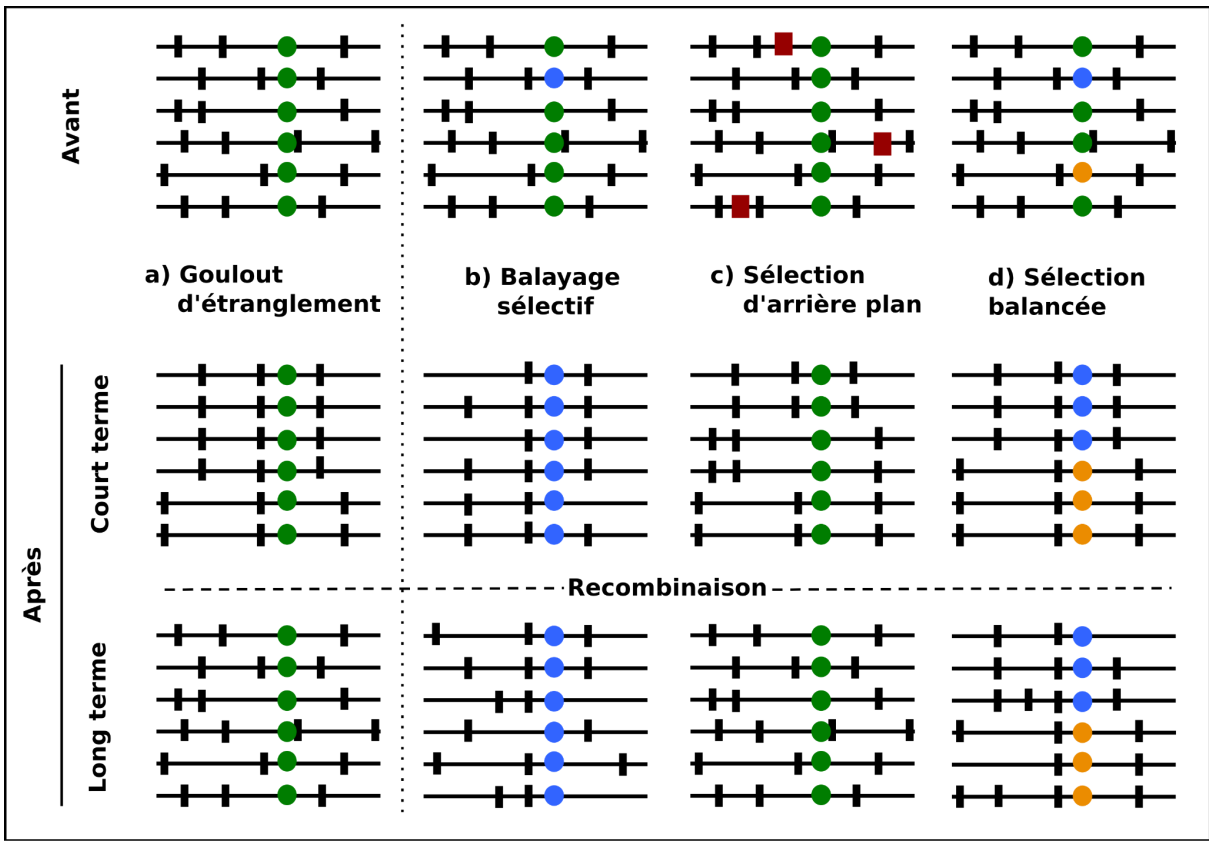


Figure Int.1. Effets démo-génétiques sur la variation neutre. Six chromosomes sont échantillonnés avant et après la sélection ou le goulot d'étranglement, avec l'effet de la recombinaison, immédiat (à court terme) et à plus long terme. Les différentes mutations sont représentées par des lignes verticales noires (mutations neutres), des cercles verts (allèle ancestral du gène focal), des cercles bleus et oranges (allèles avantageux du gène focal) et des cercles rouges (mutations délétères hors du gène focal).

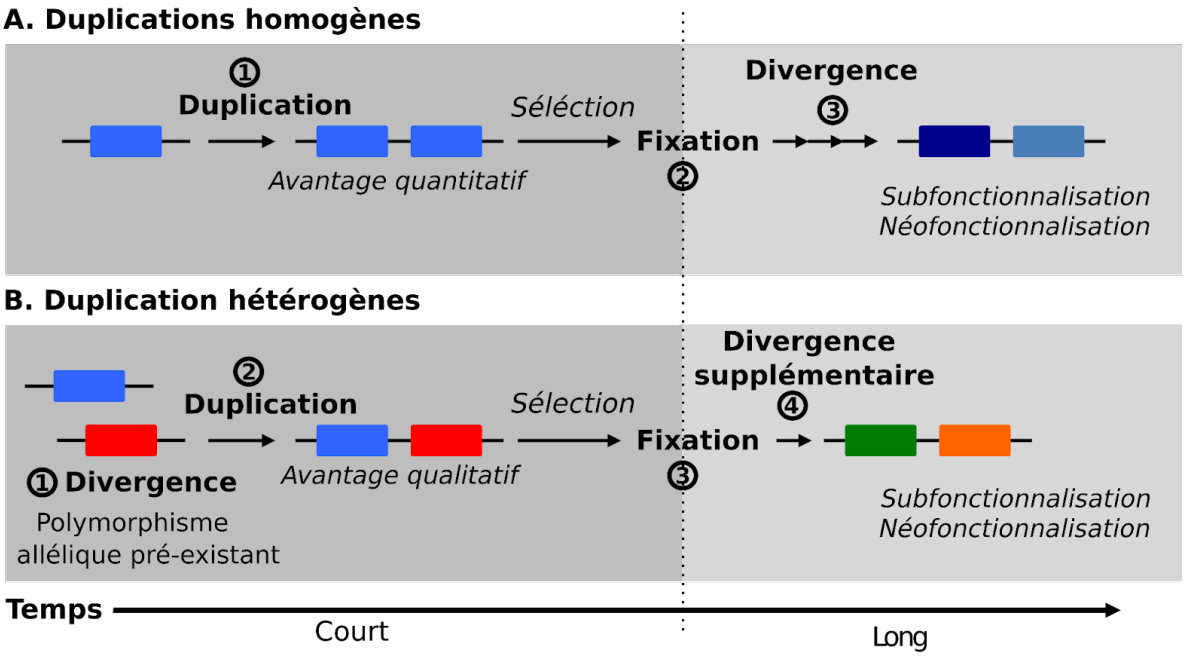
La stratégie “agressive” des mâles oranges est ainsi défaite par la stratégie furtive des mâles jaunes, qui est à son tour vaincue par la stratégie de “garde” des mâles bleus, qui sont en retour dépassés par les mâles oranges agressifs... Le succès reproductif de chacun de ces phénotypes est donc directement dépendant de sa fréquence relative, l’abondance d’une stratégie dans la population la rendant plus susceptible à l’envahissement par une autre. Un second type de sélection balancée s’observe quand le génotype le plus avantageux est l’hétérozygote : on parle alors de superdominance ou hétérosis. Ce génotype hétérozygote augmente alors en fréquence dans la population, mais comme il nécessite par essence l’existence d’au moins deux variants, leur fréquences sont maintenues stables dans la population (Fig. Int.2). Comme dans le cas des balayages sélectifs, la sélection balancée laisse une signature génomique reconnaissable : on la détecte grâce à un nombre excessif[9] de variants en fréquence intermédiaire près du site sélectionné.

Sélection et rôle évolutif des *SVs*.

Jusque-là nous avons vu les différentes formes de sélection qui peuvent agir sur une mutation ponctuelle. Elles peuvent affecter de la même façon un *SV*. Cependant, du fait de la taille du segment d’ADN qu’ils altèrent, les *SVs* engendrent des changements phénotypiques multiples dont les effets sur la valeur sélective sont en moyenne plus importants que ceux dûs aux *SNPs*, et qui peuvent en outre être antagonistes pour les différents locus impactés. Par exemple, la perte complète par délétion ou le doublement de gènes par duplication entraîne la modification de leur expression, induisant une rupture dans leur équilibre de dose (*gene dosage*)[10]. Ces mutations sont ainsi à l’origine d’une pléthore de maladies, et les *SVs* sont très largement étudiés chez l’humain (Feuk *et al*., 2006 ; Buchanan & Scherer, 2008 ; Collins *et al.,* 2020). Par exemple des modifications du gène PMP22 (Peripheral Myelin Protein 22) sont responsables de neuropathies du système nerveux périphérique : l’Hypersensibilité à la Pression (HNPP : Hereditary Neuropathy with liability to Pressure Palsy) lorsqu’il est délété, ou CMT1A (Charcot-Marie-Tooth type 1A) lorsqu’il est dupliqué (et CMT1E ou HNPP lors de mutations ponctuelles ; Jouaud, 2016). S’il n’est ainsi pas étonnant qu’une très large majorité des *SVs* soient délétères, il existe pourtant de forts indices de leur rôle crucial dans

[9] Le “nombre excessif” de variants dont je parle ici, tout comme “la diminution de la diversité génomique” résultant de l’activité de la sélection que je mentionnais plus haut, sont deux appréciations qui nécessitent chacune une valeur de référence. Je n’en parlerai pas ici, mais il existe plusieurs outils pour comparer la diversité génétique observée à un attendu neutre, *i.e.* sans sélection (voir Vitti *et al.,* 2013 pour une revue complète).

[10] Équilibre relatif de chaque gène qui peut être perturbé par l’expression d’une copie supplémentaire ou manquante dont l’apparition résulterait d’un *SV*.



Le fardeau de ségrégation associé au phénotype hétérozygote est absent dans le cas d'une duplication hétérogène.

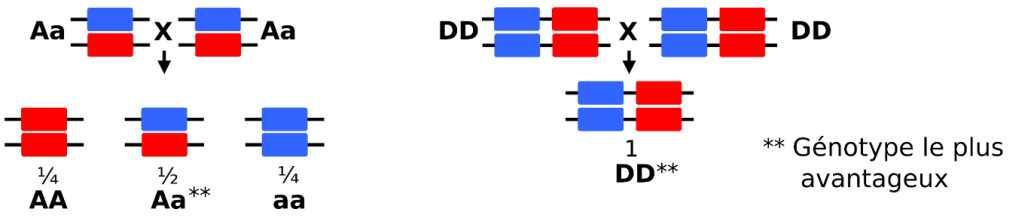


Figure Int.2. Origines et destins adaptatifs des différents types de duplications génétiques (modifié de Milesi *et al.*, 2017). Chaque rectangle représente une copie d'un locus; les couleurs indiquent que les deux copies/allèles ne sont pas identiques (*i.e.* correspondent à des haplotypes différents) et sont associées à des fonctions distinctes. L'échelle de temps est indiquée par la flèche noire et les couleurs d'arrière-plan séparent les processus à court et à long termes. (A) Les duplications homogènes associent deux haplotypes identiques d'un gène (1). Ils peuvent être sélectionnés, par exemple, pour des avantages quantitatifs dus à un rendement accru en protéines, et fixés (2). À long terme, ils peuvent diverger et acquérir de nouvelles fonctions (3). (B) Dans le cas de duplications hétérogènes, la divergence allélique (1) précède la duplication (2). Ces duplications peuvent être sélectionnées en cas de superdominance car elles permettent la production de protéines différentes, et arriver à fixation puisqu'elles ne sont pas affectées par le fardeau de ségrégation associé aux hétérozygotes standards (les descendances des deux croisements et leurs proportions sont indiqués en dessous) (3). Les deux copies remplissent des fonctions différentes avant la fixation, mais peuvent ensuite diverger davantage et se spécialiser encore (4).

les processus évolutifs, résultant de la sélection positive de certains d'entre eux (Wellenreuther *et al.,* 2019 ; voir également revue dans Mérot *et al*., 2020).

Tous les types de *SVs* n'ont pas les mêmes conséquences évolutives. Par exemple, les inversions pourraient être sélectionnées parce qu'elles empêchent la recombinaison chez les hétérozygotes (*i.e.* individus chez lesquels un chromosome porte l'inversion, mais pas le chromosome complémentaire ; ou hétérocaryotypes) : elles entraîneraient ainsi la transmission groupée d'un ensemble d'allèles adaptatifs (le *SV* est alors transmis comme un seul locus, ou supergène). Plusieurs cas d'adaptation et de structuration de populations via des inversions ont ainsi été largement documentés (Wellenreuther & Bernatchez, 2018 ; Faria *et al.,* 2019 ; Mérot *et al.,* 2021 ; Jay *et al.*, 2021 ; Schaal *et al.,* 2022).

Les duplications constituent l'autre grand type de *SVs* qui a tout particulièrement attiré l'attention des biologistes de l'évolution. Il a été proposé que ces mutations soient une source de matériel génétique propre à l'établissement de nouvelles fonctions, les nombreuses familles multigéniques rencontrées dans le vivant résultant de la répétition de ce processus au cours l'évolution (Haldane, 1932 ; Ohno, 1970). Le modèle classique (Ohno, 1970 ; Lynch & Conery, 2000), propose que dans la majorité des cas de duplications, une des copies serait perdue par accumulation de mutations, c'est ce que l'on appelle la *pseudogénisation*. Toutefois dans le cas où les deux copies seraient conservées suffisamment longtemps (*i.e.* sous l'hypothèse que le *SV* ne soit pas perdu par dérive, ou contre-sélectionné), il a été proposé que la pression de sélection impactant chacune des copies serait relâchée, une mutation normalement délétère sur une seule des deux impactant moins la valeur sélective de son porteur puisque la seconde peut toujours assumer la fonction du locus. Il en résulterait deux issues possibles : i) la *néofonctionnalisation* où une des deux copies retient la fonction originelle du gène tandis que l'autre en acquiert une nouvelle ; et ii) la *subfonctionnalisation* où une partie de la fonction du gène est conservée par chaque copie, ce qui permet leur maintien en tandem et leur optimisation par sélection, ou spécialisation (Hughes, 1994 ; Force, 1999; Lynch & Force, 2000). Un exemple largement étudié de néofonctionnalisation est retrouvé dans les gènes de la famille des opsines chez les primates (Hunt *et al.,* 1998 ; Dulai *et al.,* 1999 ; Dominy & Lucas, 2001 ; Surridge *et al.,* 2003). Une duplication du gène codant pour les opsines vertes chez les *Catarrhini* (singes de l'Ancien Monde) aurait mené à la néofonctionnalisation d'une des copies résultant en l'apparition des opsines rouges et de la vision trichromatique chez les espèces de ce clade (Dominy & Lucas, 2001 ; Vallender, 2017). Vous remarquerez que les prédictions sur le devenir des gènes dupliqués, et les exemples que j'en ai donné, nécessitent de longues périodes évolutives pour se mettre en

place : il faut en effet qu'une duplication soit suffisamment fréquente dans la population pour envisager ces évolutions secondaires.

Un moyen rapide de fixer un variant est la sélection naturelle. Mais *quid* de la sélection des duplications de gène à l'échelle micro-évolutive (de l'ordre de centaines de générations) ? On a vu qu'elles sont délétères dans la très grande majorité des cas (Schrider *et al.,* 2013). Cependant, il existe de nombreux exemples de duplications immédiatement adaptatives. Les duplications homogènes (ou amplifications géniques; Fig Int.2.A) peuvent être sélectionnées pour l'avantage quantitatif qu'elles entraînent en engendrant des allèles où un même haplotype est retrouvé en plusieurs copies, et donc une surproduction des protéines codées. On connaît de nombreux exemples de ce phénomène : dans plusieurs espèces d'insectes des duplications entraînent la surexpression d'enzymes de détoxification qui leur permettent de survivre à l'exposition à un xénobiotique (voir revue dans Bass & Field, 2011). Des amplifications du gène de l'amylase permettent aux humains et à leurs plus fidèles compagnons à quatre pattes (*Canis lupus domesticus*) de digérer une alimentation riche en céréales (Perry *et al.,* 2007 ; Axelsson *et al.,* 2013). De la même manière, des duplications homogènes sont impliquées dans l'adaptation à la chaleur (Riehle *et al*., 2001) et aux métaux lourds chez des bactéries (von Rozycki & Nies, 2009) ; ou même encore dans la résistance aux traitements médicamenteux chez des leishmanies (Mary *et al.,* 2010 ; voir Kondrashov, 2012 pour revue détaillée des duplications dans l'adaptation à court terme). D'autres types de duplications associent différents allèles d'un locus sur un même brin d'ADN : on parle alors de duplications hétérogènes, chaque copie correspondant à un haplotype différent. Elles permettent par exemple de fixer un phénotype hétérozygote avantageux en cas de superdominance (Haldane, 1954 ; Milesi *et al.,* 2017). On notera que les deux copies sont différentes dès l'événement de duplication, ce qui les différencie des cas de néofonctionnalisation. Si les patrons génomiques de ces duplications hétérogènes sont indiscernables de ceux générés par néo- ou sub-fonctionnalisation, leur mécanisme de fixation est bien identifié (Innan et Kondrashov, 2010 ; Labbé *et al.,* 2007 ; Milesi *et al.,* 2017 ; Fig.Int.2). On a peu d'exemples de ce type de duplications, certainement du fait de la difficulté associée à leur identification. Le plus documenté est celui des duplications hétérogènes du gène *ace-1* observées chez plusieurs espèces de moustiques, associant sur un même brin d'ADN un allèle associé à la résistance aux insecticides et un second allèle sensible[11].

[11] On remarquera un brusque rétrécissement du champ des références dans une tentative éhontée de recentrer le discours sur le sujet de ma thèse.

II. Résistance aux insecticides chez les moustiques.

J'ai déjà dit quelques mots sur la complexité inhérente aux études de l'adaptation : il est difficile d'identifier une pression de sélection, de comprendre la façon dont elle affecte les populations et comment la myriade de facteurs confondants (démographiques, stochastiques, génétiques…) peuvent brouiller la détection des signatures de la sélection naturelle (Rausher, 1992). Néanmoins, dans le cas de la résistance aux xénobiotiques, la pression de sélection est clairement identifiée, et souvent même quantifiée, puisqu'elle est d'origine anthropique. Les effets biochimiques de la pression de sélection (les xénobiotiques) sont connus, à la fois sur l'organisme et sur leur cible moléculaire : par exemple empêcher la fermeture des canaux sodium-dépendants dans le cas des insecticides pyréthroïdes, et entraîner ainsi la mort du moustique. Enfin, les traitements visent généralement des organismes avec des temps de génération courts (évolution rapide) et de larges effectifs (impact limité de la dérive). Toutes ces caractéristiques font de la résistance aux xénobiotiques un cadre privilégié pour l'étude d'adaptation, notamment à une échelle micro-évolutive. Pendant ma thèse, j'ai justement étudié plusieurs cas d'adaptation aux xénobiotiques chez les moustiques, et je vais maintenant présenter celui qui aura occupé la plus grande partie de mon temps ces trois dernières années : les mutations du gène *ace-1*.

Résistance par modification de l'acétylcholinestérase.

Les années 1950 ont vu l'essor de méthodes de lutte chimique contre des vecteurs de maladies ou ravageurs de cultures. Les moustiques, vecteurs de maladies dont certaines mortelles et parmi les plus prégnantes pour l'humain (paludisme, dengue, fièvre jaune, etc.), ont très tôt été la cible de traitements insecticides utilisant diverses molécules avec des modes d'action différents (Mulla, 1994). Plusieurs cas de résistances à des traitements auparavant efficaces ont toutefois été rapidement signalés (Brown, 1958 ; Ayad & Georghiou, 1975 ; voir Forgash, 1984 pour revue). Par exemple, l'utilisation d'insecticides de type organophosphates et carbamates (OP et CX) a induit l'augmentation en fréquence d'allèles de résistance du gène *ace-1* encodant l'acétylcholinestérase (AChE1). L'AChE1 est une enzyme responsable de l'arrêt de l'influx nerveux par la dégradation du neurotransmetteur acétylcholine (ACh) dans les synapses neuronales, et elle est également impliquée dans le développement du système nerveux chez les invertébrés et les vertébrés (Grisaru *et al.,* 1999 ; Cousin *et al.,* 2005). En se fixant sur l'AChE1, les insecticides OP/CX engendrent une accumulation d'ACh dans les synapses, ce qui entraîne la mort par paralysie (Massoulié *et al.,* 1993). Le premier allèle de résistance identifié au locus *ace-1* était dû à une mutation ponctuelle (notée

G119S[12], j'y référerai par la suite sous l'appellation d'allèle R) acquise indépendamment chez plusieurs espèces de moustiques appartenant aux complexes d'espèces *Anopheles gambiae s.l.* et *Culex pipiens s.l.* (Weill *et al.*, 2003 ; Weill *et al.,* 2004).

Au niveau moléculaire, les impacts de cette mutation sont similaires dans les deux complexes d'espèces (Alout *et al.,* 2008) : la modification d'une glycine en sérine sur le site actif de l'enzyme engendre une diminution de son affinité pour les insecticides OP et CX et permet par conséquent la résistance, mais elle limite aussi l'affinité pour l'ACh, réduisant l'activité enzymatique de plus de 60% par rapport à la protéine sensible (Bourguet *et al.,* 1996 ; Alout *et al.,* 2008). Cette mutation ponctuelle illustre les effets pléiotropes du gène *ace-1* puisque, bien qu'elle permette de survivre à l'exposition aux insecticides, cette modification de l'activité de l'AChE1 a des impacts essentiellement délétères qui se révèlent particulièrement en absence d'insecticides sur de nombreux traits : temps de développement (Raymond *et al.,* 2001), résistance aux infections et à la prédation (Berticat *et al.,* 2004 ; Duron *et al.,* 2006), survie (Gazave *et al.,* 2001) et succès reproductif (Berticat et al., 2002). On a donc affaire à un *trade-off* irréductible entre activité protéique et résistance (Labbé et al. 2007 ; Assogba *et al.,* 2015).

Du fait de son intérêt sanitaire et parfois économique, la dynamique de cette mutation a été suivie de près, et ce depuis plusieurs décades, dans les deux complexes *An. gambiae s.l.* et *Cx. pipiens s.l.*, notamment grâce à l'élaboration de tests moléculaires permettant de différencier les différents phénotypes au gène *ace-1* : [R] (résistant), [RS] (hétérozygote) et [S] (sensible ; Bourguet *et al.,* 1996).

Les *SVs* d'*ace-1* chez *Culex pipiens s.l.*

Les premiers indices de l'existence de *SVs* au locus *ace-1* ont été découverts dans une souche issue d'une population de Martinique : après sélection aux CX (carbamates) répétée sur plusieurs générations, le phénotype hétérozygote [RS] était arrivé à fixation dans la souche. La fixation d'un phénotype hétérozygote est normalement impossible avec des allèles simple copie du fait du fardeau de ségrégation qui l'affecte (Fig. Int.2). L'hypothèse que la souche était porteuse d'une duplication hétérogène associant une copie résistance, R, et une copie sensible, S, du gène *ace-1* a alors été proposée (allèles D, Bourguet *et al.,* 1996). Un peu plus tard, un fort excès d'hétérozygotes par rapport à l'attendu sous panmixie[13] a été

[12] Elle est parfois appelée G280S chez *An. gambiae s.l.*, en fonction de la référence utilisée pour l'annotation de la protéine AChE.

[13] Lorsque l'appariement des gamètes est aléatoire, on peut facilement prédire les fréquences alléliques attendues dans une population de très grande taille ; on parle d'équilibre de Hardy-Weinberg.

découvert dans une population de la région de Montpellier, qui a été interprété comme dû à la présence d'une duplication hétérogène dans cette population (Lenormand *et al.,* 1998). Ces soupçons ont été confirmés formellement par les travaux de Labbé et *al.* en 2007, et les duplications de Martinique et de Montpellier ont été fixées dans des souches de laboratoire. Par ailleurs,ce n'était pas un mais deux allèles D différents (nommés D_2 et D_3) qui ségrégeaient dans les populations montpelliéraines. La fixation de ces allèles dans des souches à fond génétique commun a permis de caractériser leurs effets phénotypiques, et de comparer leur *fitness*. Si les allèles montpelliérains, D_2 et D_3, sont sublétaux à l'état homozygote (phénotype Homozygote Sublétal HS), D_1, l'allèle martiniquais, ne l'est pas (Labbé *et al.,* 2007). L'AChE1 codée par les allèles D_2 et D_3 est fonctionnelle (Labbé *et al.,* 2007 ; Labbé *et al.,* 2014), et il a donc été proposé que l'origine du phénotype HS soit liée à l'architecture génomique des duplications : des mutations liées aux zones de cassure (*breakpoints*) engendrées par la duplication, des mutations dans les gènes embarqués, ou une rupture de l'équilibre du dosage génique pourraient s'exprimer chez les homozygotes tout en ayant des effets limités chez les hétérozygotes si récessives. Cette hypothèse était renforcée par le fait que les allèles D_2 et D_3 complémentent, c'est-à-dire que la *fitness* d'un hétérozygote portant ces deux duplications (D_2/D_3) est similaire à celle d'un homozygote D_1/D_1. Il a alors été proposé que la sublétalité des allèles montpelliérains expliquait leur co-ségrégation. Ceci a été confirmé lors d'une étude plus systématique à l'échelle mondiale ou, au total, 27 allèles D différents ont été identifiés, témoignant de grande diversité des duplications hétérogènes chez *Culex*. Parmi ces allèles D, sur les sept dont la *fitness* a été mesurée, six sont sublétaux et tous les allèles sublétaux complémentent (Milesi *et al.,* 2018).

Les SVs d'*ace-1* chez *Anopheles gambiae s.l.*

En parallèle, une duplication hétérogène similaire a été découverte en 2008 chez *An. gambiae s.l.* (Ag-D_1[14]), par les travaux de Djogbénou et collaborateurs, dont la distribution semblait assez large en Afrique de l'Ouest (Djogbenou *et al.*, 2009). Il a été démontré par la suite qu'elle avait un impact similaire sur la fitness de ses porteurs que celle observée chez *C. pipiens* en Martinique (Cp-D_1 ; Assogba *et al.,* 2015). Un peu plus tard, on a également découvert que tous les allèles R retrouvés chez *An. gambiae s.l.* étaient en fait des duplications homogènes associant plusieurs copies du même haplotype et dont le nombre peut varier entre 2 et 10 (allèles R^x ; Djogbénou *et al.,* 2015 ; Weetman *et al.,* 2015 ; Assogba

[14] pour différencier les allèles des deux complexes d'espèces *An. gambiae s.l.* et *Cx. pipiens s.l.*, on ajoutera si nécessaire Ag et Cx devant le nom de l'allèle, resp.

et al., 2016). Pour ces allèles R^x, il y a une relation quantitative entre le nombre de copies R, la capacité de résistance, mais aussi l'intensité des effets délétères associés. Par exemple, un allèle a trois copies, R^3, est moins résistant et moins coûteux qu'un allèle à cinq copies, R^5. Chaque duplication du locus *ace-1*, qu'elle soit hétérogène ou homogène, est donc associée à un *trade-off* resistance-coût qui lui est propre (Assogba *et al.,* 2016), ce qui représente autant de solutions pour s'adapter à différents régimes de traitement insecticides. Plus étonnant encore, toutes les duplications *ace-1* partagent une architecture commune (allèles Ag-D et Ag-R^x ; Assogba *et al.,* 2016) : tous ces allèles sont composés d'amplicons de 203 Kb en tandem, comportant le gène *ace-1* et 11 autres gènes embarqués (Assogba *et al.,* 2016). Enfin, Assogba et collaborateurs ont mis en évidence l'existence d'une délétion interne dans certains allèles R^x, dont il a été proposée qu'elle soit adaptative puisque 1) elle supprime tous les gènes embarqués à l'exception d'*ace-1*, rétablissant ainsi le dosage initial de ces gènes, et 2) l'allèle R^x délété augmente en fréquence dans des populations naturelles et a introgressé entre espèces du complexe *An. gambiae s.l.* au Togo (Assogba *et al.,* 2018).

III. Et ma thèse dans tout ça ?

L'objectif initial de ma thèse était donc de caractériser l'architecture génomique des allèles D fixés dans des souches chez *Cx. pipiens s.l.* et dont la *fitness* était caractérisée, afin d'identifier les éventuelles différences structurelles qui pourraient expliquer pourquoi certains de ces allèles sont sublétaux et d'autres non. Le second objectif était d'étudier l'impact de ces allèles sur le polymorphisme génétique neutre lors des balayages sélectifs liés à leur succès adaptatif. Nous disposons en effet d'un modèle avec plusieurs *SVs* récents ayant envahi des populations naturelles, idéal pour l'étude de l'émergence des duplications adaptatives, et plus généralement pour tester des modèles mêlant démographie et sélection. Toutefois, n'échappant pas à ce qui est vécu dans la grande majorité des thèses, ma route s'est révélée *nettement* plus sinueuse que prévue[15]…

Dans le premier chapitre de cette thèse, je traiterai des difficultés que j'ai rencontrées dans l'analyse de ces allèles dupliqués, et plus généralement, dans l'analyse génomique des *SVs* dans un organisme non-modèle. Je montrerai notamment comment la nature même de ces allèles rend leur étude particulièrement complexe et facilement sujette à de nombreux biais si l'on n'y prend pas garde. Je présenterai dans ce chapitre le premier article de ma thèse,

[15] "Lourd est le parpaing de la réalité sur la tartelette aux fraises de nos illusions." Boulet, Notes, Tome 3 : *La viande, c'est la force*. Cet auteur, qui m'était inconnu au début de la rédaction, aura finalement largement étoffé le contenu des notes de cette introduction, pour le meilleur et pour le pire.

portant sur l'identification de nouveaux allèles D dans des populations d'*Anopheles* de Côte d'Ivoire. J'y détaillerai comment j'ai pu mettre au point une méthode d'analyse automatisée pour identifier et caractériser les variants structuraux ségrégeant dans ces deux complexes d'espèces.

Dans le deuxième chapitre, j'introduirai en particulier les analyses que j'ai réalisées sur le *PipPop Project*, un jeu de données récemment établi par des collaborateurs américains comptabilisant plus de 800 génomes du complexe *Cx. pipiens s.l.* provenant de 47 pays différents, et qui m'a permis d'y découvrir une grande diversité structurelle pour les allèles D. J'y présenterai aussi l'évolution parallèle observée au locus *ace-1* dans l'adaptation aux insecticides OP/CX chez les moustiques *An. gambiae s.l.* et *Cx. pipiens s.l.,* en commençant par mon second article, dans lequel j'ai participé à l'identification d'un nouvel allèle dupliqué R^x chez *Cx. pipiens s.l.*

Dans le troisième chapitre, j'introduirai les résultats obtenus suite à l'analyse de plusieurs autres gènes impliqués dans la résistance aux insecticides sur ce même jeu de données *PipPop Project.* J'y montrerai en particulier que *ace-1* n'est pas le seul locus où les *SVs* jouent un rôle déterminant dans cette adaptation.

Enfin je discuterai de ce que m'ont appris ces travaux sur l'évolution adaptative des *SVs*, et des perspectives qui s'ouvrent à leur suite pour étudier l'impact de ces *SVs* sur la diversité génétique en lien avec la démographie, notamment grâce au jeu de données que j'ai établi lors de cette thèse, avec des échantillons de plusieurs populations couvrant plus de 40 années d'évolution de la résistance dans la région de Montpellier.

Glossaire des duplications.

Les duplications que j'ai étudiées prennent des formes *très* diverses. Elles présentent des variations à différentes échelles, qui peuvent pourtant avoir une importance capitale pour comprendre leur histoire évolutive et leur impact adaptatif.
Il m'aura fallu attendre la fin de ma thèse pour comprendre à quel point la nomenclature mise en place par mes prédécesseurs est importante, puisqu'elle permet justement de rendre compte de la finesse des variations qui séparent les variants dupliqués[1]. Vous trouverez ici une explication de cette nomenclature, qui devrait vous permettre d'y voir un peu mieux dans le brouillard de technicalités que je m'apprête à vous faire traverser.

- **Duplication** : partie du génome retrouvée en plus de 2 exemplaires dans une espèce diploïde.

- **Amplicon** : un exemplaire de la partie du génome qui est dupliquée.
  NB : La profondeur de couverture (*Depth of Coverage* - *DoC*) observée pour un génome avec duplication aligné sur un génome de référence peut indiquer le nombre d'amplicons.

- **Structure d'une duplication** : caractérisation de la zone du génome affectée par une duplication, en termes de taille, de nombre d'amplicons, d'homogénéité d'amplification au sein de la zone (ex. délétions ou duplications internes à la duplication étudiée). Synonyme dans ma thèse : **architecture**.

Ces dénominations sont générales et concernent l'ensemble de la duplication. Cependant, je m'intéresse dans ma thèse aux duplications qui affectent un gène cible (principalement *ace-1*) pour leur caractère adaptatif. Si on se focalise sur ce gène cible, d'autres éléments de nomenclature deviennent rapidement indispensables...

- **Copie** : réfère au nombre de fois où le gène cible est retrouvé dans le génome (normalement, le nombre de copies est égal au nombre d'amplicons…).
  NB : Par commodité, on pourra considérer en bloc le nombre de copies porteuses de la mutation G119S, ou copies R, pour le distinguer du nombre de celles qui portent la version sensible, ou copies S, sans s'intéresser du tout à la diversité du reste de la séquence entre copies S, entre copies R, ou entre copies R et S (quand on s'intéressera à cette diversité, on parlera d'**haplotypes**, ex. une duplication composée de 4 **amplicons**, 2 **copies** S et 2 copies R, pourra être composée de deux **haplotypes** S mais d'un unique haplotype R en deux exemplaires…)

- **Types de duplication** : on distinguera deux grands types de duplications: i) les **duplications homogènes**, ou **allèles $\mathbf{R^x}$** et S^x, qui ne sont composées que de copies R ou S respectivement, ii) les **duplications hétérogènes**, ou **allèles D**, composés à la fois de copies R et de copies S.

- **Haplotype** : réfère à la séquence particulière d'un exemplaire spécifique du gène cible, et concerne l'ensemble de la séquence connue.
  NB : On regroupera parfois plusieurs haplotypes différents (*i.e.* avec des mutations entre eux sur les autres bases) en catégories basées sur le nucléotide trouvé à la position 119: on parlera alors d'haplotypes R quand les séquences portent la mutation G119S, et d'haplotypes S dans le cas contraire (*NB* : environ équivalents aux copies R et aux copies S, resp., sauf que l'utilisation du terme "haplotype" suggère en plus une diversité de séquences hors mutation G119S).

- **Allèles** : on appellera allèle tout variant affectant le gène cible et/ou la zone dupliquée. Des duplications ayant i) des structures différentes seront donc des allèles, ou des duplications dont ii) le nombre de copies du gène cible diffèrent (copies R, S, ou les deux), ou iii) composées d'haplotypes différents (identité ou association des haplotypes R, S, ou des deux). On distinguera donc i) les ***structural-alleles***, ii) les ***copy-number-alleles*** et iii) les ***sequence-alleles***, qui correspondront respectivement aux trois types de variations décrites.
  NB: On pourra aussi regrouper les allèles par **type de duplication**, allèles $\mathbf{R^x}$, S^x ou **D** (associations de copies, R, S, ou des deux, resp.), éventuellement contrastés avec les allèles simple-copie **R** ou **S**.

[1] C'est malheureusement à vous, cher(es) lecteur(trice)s de payer le prix de la précision de mes propos. J'ai essayé (et mes directeurs plus encore) de rendre la lecture de cette thèse aussi agréable que possible. Mais il vous faudra tout de même endurer des phrases pleines d'informations et de notions diverses, pénibles, parfois alambiquées, et qui ne sont en rien simplifiées par ma fâcheuse tendance à économiser les points au profit des virgules.

Chapitre I. Caractérisation génomique des duplications *Anopheles* et *Culex* : des affres de la bioinformatique.

Le premier objectif de ma thèse était la caractérisation de l'architecture génomique des duplications hétérogènes (allèles D) chez *Cx. pipiens s.l.*, dans le but d'en comprendre l'origine et de tenter d'expliquer leurs impacts différents sur la *fitness*. Ce genre d'analyse génomique avait déjà permis à notre équipe de caractériser l'architecture d'une duplication chez *An. gambiae s.l.* (Assogba *et al.,* 2016). Le protocole que je devais suivre était donc établi, et les données de séquences sur lesquelles je devais travailler étaient déjà disponibles au laboratoire, puisque les génomes des souches porteuses d'allèles D avaient été séquencées en 2016 (Illumina, *pair-ends short-reads*). Tout paraissait donc devoir se dérouler sans accroc… Hélas, la transposition à *Culex* des analyses réalisées chez *Anopheles* a été bien plus complexe que ce que nous ne l'avions espéré, et nous nous sommes rapidement heurtés à un premier écueil[1] que je détaillerai ci-après. Cela m'a obligé à affiner le protocole de caractérisation, et à revoir nos ambitions...

Au final, j'ai dû mettre au point et tester plusieurs approches pour arriver à un protocole d'analyse fiable, transposable et automatisé (Box I.1 pour une présentation de l'outil d'identification que j'ai développé, *ArDu*[2]). A cause des difficultés avec *Culex*, j'ai décidé de me faire la main sur un jeu de données plus "facile" contenant de nouvelles duplications hétérogènes identifiées dans deux populations Ivoiriennes d'*An. coluzzii*, que j'avais commencé à étudier lors de mon stage de deuxième année de master. Par conséquent, j'ai travaillé dans deux cadres distincts: le premier -les données *Anopheles*- que je qualifierai d'idéal pour des raisons que je détaillerai plus bas ; et le second, plus expérimental et complexe, avec les données *Culex.* J'ai alterné entre ces deux cadres de travail, testant d'abord les méthodes et outils sur les données *Anopheles* avant de les confronter aux données *Culex*. Dans ce chapitre, je présenterai la caractérisation de l'architecture génomique de plusieurs duplications dans chaque genre, en détaillant les difficultés qui ont émaillé mes analyses : elles constituent encore aujourd'hui un obstacle à la compréhension de mon sujet de recherche, et soulèvent des questions d'intérêt général sur le fonctionnement, la portée et la confiance que l'on peut avoir dans les analyses de génomique comparative lorsqu'elles impliquent des structures génomiques complexes.

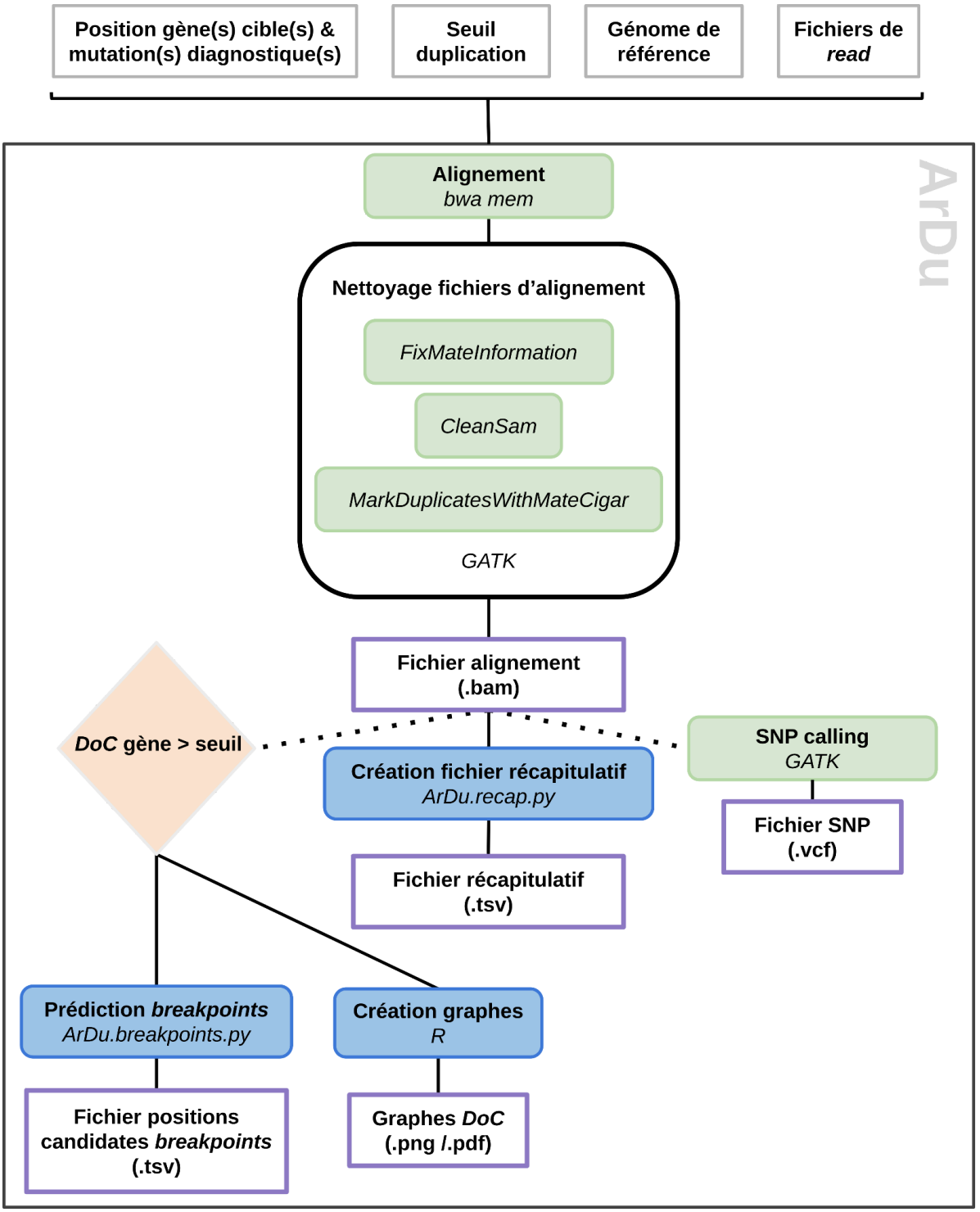
[1] … et puis un autre, et encore un autre.

[2]Avez-vous noté la propension des chercheurs en bioinformatique à faire des jeux de mots (plus ou moins réussis) pour nommer leurs algorithmes ? J'ai voulu m'y essayer aussi… verdict ?

I. Méthode d'analyse des duplications : c'est donc *ArDu*…

Il existe différentes méthodes pour identifier des duplications, qui ont chacune leurs avantages et inconvénients; je ne les détaillerai pas ici mais une revue très complète a été écrite par Lallemand *et al.* (2020). Pour une grande part, ces méthodes sont destinées à des analyses plus "générales" : des scans globaux du génome et des approches qui ne cherchent pas à étudier un locus en particulier. L'approche et la détection systématique se fait au détriment d'une caractérisation fine des duplications (*e.g.* larges intervalles de confiance sur les *breakpoints*, ambiguïtés sur le nombre de copies, et problèmes d'échelle lorsque des réarrangements structuraux sont inclus dans des réarrangements plus grands). Dans mon cas, je m'intéressais uniquement au gène *ace-1* (au début en tout cas, *c.f.* Chapitre III) ; l'objet de mon analyse était donc clairement identifié. Par ailleurs, je disposais d'importantes ressources génomiques, et des travaux précurseurs réalisés par Assogba et collaborateurs (2016) sur la caractérisation des duplications homogènes, R^x, et hétérogènes, *Ag*-D, chez *Anopheles*. J'ai donc suivi l'approche globale proposée par Assogba *et al.* (2016) tout en l'adaptant pour développer un *pipeline* de détection et d'analyse automatique de l'*Ar*chitecture des *Du*plications (*ArDu*, Box I.1). Cet outil repose sur une idée simple : puisqu'un fragment d'ADN dupliqué est en multiples copies, son séquençage produit une plus grande quantité de *reads* par rapport aux autres parties du génome. En alignant les *reads* issus du séquençage d'un individu porteur d'une duplication sur un génome de référence, *i.e.* dépourvu de la duplication en question, on observe une augmentation de la profondeur de couverture (*Depth of Coverage,* DoC) sur la portion de génome qui est en plusieurs copies (zone dupliquée) par rapport aux régions non-dupliquées. Ce changement local de couverture permet d'identifier la position, l'étendue et le nombre de copies de cette duplication. En plus d'un changement de DoC, la présence de duplications va laisser une signature spécifique au niveau des *reads* chevauchant les bornes de la duplication. L'étude fine de cette signature permet la détection précise des positions génomiques de l'insertion de la zone dupliquée, ou points de cassure (*breakpoints,* voir le schéma Fig.I.1). Hélas, c'est là qu'est l'os[3] : ce type d'analyse est *très* dépendant de la qualité des ressources génomiques utilisées, à la fois celle du génome de référence, et celle du séquençage des différents individus. Particulièrement dans un cas où comme ici l'objet de l'étude génomique est la relation structure-fonction : la précision des prédictions sur la taille des amplicons, la position des *breakpoints* et le contenu en gènes embarqués est en effet primordiale pour la compréhension du phénotype HS.

[3] …


Position gène(s) cible(s) & mutation(s) diagnostique(s)
Seuil duplication
Génome de référence
Fichiers de *read*
ArDu
Alignement
bwa mem
Nettoyage fichiers d'alignement
FixMateInformation
CleanSam
MarkDuplicatesWithMateCigar
GATK
Fichier alignement (.bam)
***DoC* gène > seuil**
Création fichier récapitulatif
ArDu.recap.py
SNP calling
GATK
Fichier SNP (.vcf)
Fichier récapitulatif (.tsv)
Prédiction ***breakpoints***
ArDu.breakpoints.py
Création graphes
R
Fichier positions candidates ***breakpoints*** (.tsv)
Graphes ***DoC*** (.png /.pdf)

BOX I.1. Représentation graphique du *Pipeline* Ar(chitecture)Du(plication). Les fichiers d'entrée nécessaires au *pipeline* sont indiqués par des rectangles gris, les fichiers de sorties sont en mauve, les outils dépendants sont en vert, tandis que les scripts d'analyses développés par mes soins sont indiqués en bleu. Les liaisons en pointillés indiquent les étapes optionnelles ou soumises à une condition logique, représentée par l'étrange losange orange. Les positions génomiques du (ou des) gène(s) cible(s) et des éventuelles mutations diagnostiques (*e.g.* G119S pour *ace-1*) sont renseignées dans un fichier texte, tout comme les paramètres graphiques choisis pour les graphes de profondeur de couverture (taille de la zone à représenter, couleur choisie pour représenter l'emplacement du gène et sa position). L'alignement des fichiers de *reads* sur le génome de référence est réalisé à l'aide de l'outil *bwa mem* (Li & Durbin, 2009). Les fichiers d'alignement produits sont ensuite nettoyés avec différents outils de la suite GATK (outils de la suite PICARD, appelés par GATK; Van der Auwera & O'Connor, 2020). Le fichier d'alignement résultant est analysé avec un script de ma conception (ArDu.recap.py) qui permet d'obtenir la *DoC* normalisée du gène cible. La *DoC* normalisée correspond à la *DoC* moyenne du gène rapportée à la *DoC* médiane du chromosome ou à la *DoC* moyenne d'un gène de référence connu pour être en une seule copie (un fonctionnement un peu similaire à une PCR quantitative). Ainsi, un individu porteur d'un gène en une seule copie aura une *DoC* normalisée proche de 1, tandis que dans le cas d'un homozygote dupliqué, la valeur sera proche de 2 (ou 1.5 pour un hétérozygote dupliqué/mono-copie). Le script *python* ArDu.recap.py fournit également d'autres informations utiles, par exemple la présence d'une mutation diagnostique si sa position est renseignée dans les fichiers d'entrée, ou encore la *DoC* moyenne et médiane des chromosomes. Enfin, la *DoC* du gène cible est comparée au seuil fourni en entrée: si elle est trouvée supérieure, un graphe de *DoC* est produit et un second script (ArDu.breakpoints.py) extrait les positions candidates pour les *breakpoints* de la duplication (ces positions sont détectées grâce aux informations de *soft-clipped reads* et de tailles d'*insert* anormales, voir Fig. I.1).

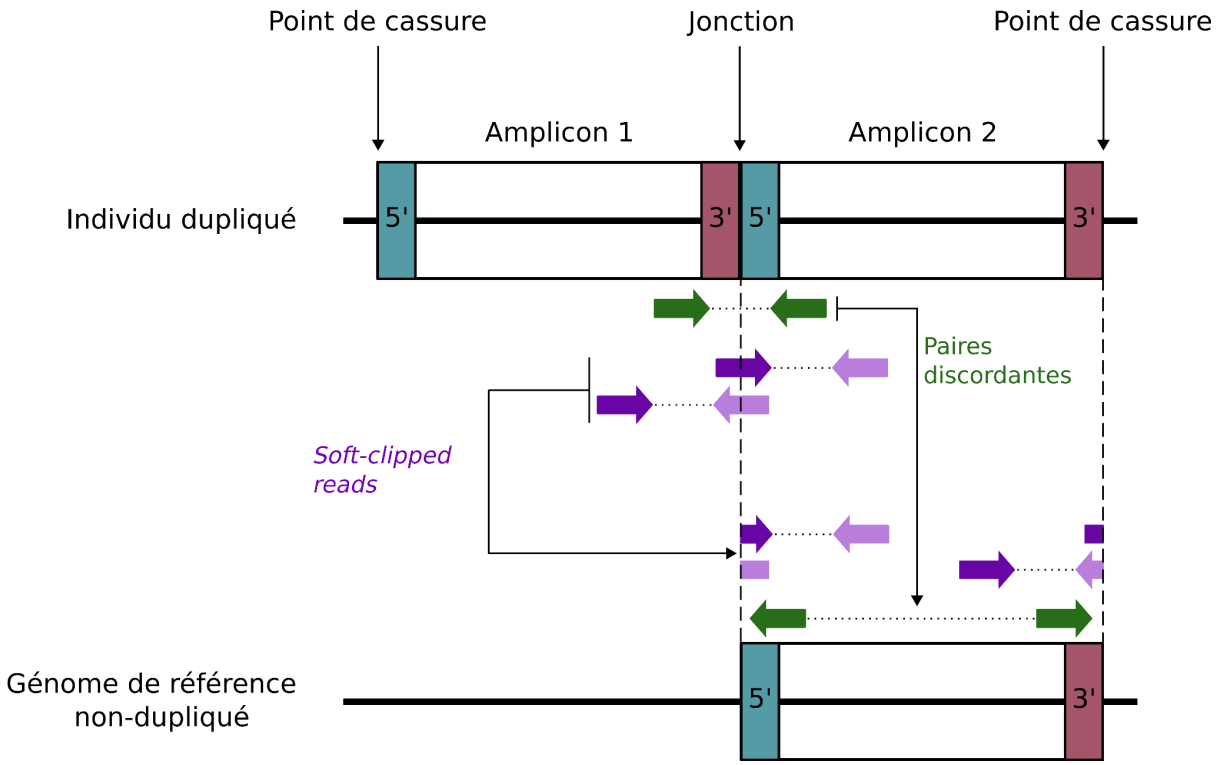


Figure I.1. Protocole de détection des points de cassure des duplications. La caractérisation des bornes exactes de la duplication est rendue possible par l'information de *reads* dont l'alignement est anormal. Les paires discordantes sont représentées en vert, la taille de l'*insert* les séparant de leur *read* apparié donne une estimation de la taille de la zone dupliquée et de l'emplacement des cassures. Les *soft-clipped reads*, en mauve, ne s'alignent que sur une partie de leur séquence: ils sont en fait à cheval sur la jonction des deux copies de la duplication et vont donc s'aligner au début et à la fin de la zone dupliquée sur le génome de référence. Il est ainsi possible d'affiner la position d'un point de cassure en récupérant la position génomique sur laquelle leur alignement s'interrompt.

I.1. Le génome de référence.

Le génome de référence est le premier obstacle de taille que j'ai rencontré dans mes analyses. Je mentionne de manière très désinvolte l'utilisation d'un "génome de référence non dupliqué" dans le protocole d'identification d'une duplication ci-dessus. Ma première année de thèse m'aura permis de comprendre à quel point cette référence et sa qualité sont vitales à la bonne conduite de l'étude que j'entreprenais. Et bien que la diminution des coûts de séquençage ait permis une nette augmentation du nombre de génomes de référence disponibles, leurs qualités sont *très* variables, notamment pour les espèces non-modèles.

Qualité de la référence.

J'ai en effet étudié des espèces pour lesquelles l'état des génomes de référence disponibles étaient (et restent) très différents. *An. gambiae* et *An. coluzzii*, principaux vecteurs du paludisme en Afrique et étudiés à ce titre par de très nombreuses équipes internationales à travers le monde, disposent d'une référence d'excellente qualité (*AgamP4.12*, établie à partir de la souche de laboratoire *An. gambiae PEST* ; Holt *et al.,* 2002 ; Mongin *et al.,* 2004). L'annotation de cet assemblage est le fruit du travail de nombreuses équipes et continue de recevoir des mises à jour régulières. A l'inverse, *Cx. quinquefasciatus* et *Cx. pipiens* sont des vecteurs d'arboviroses plus généralement bénignes (pas toujours, et pas que, mais bon…), et ils sont donc peu étudiés (seules quelques équipes s'y intéressent[4]). Le seul assemblage disponible pour ces espèces au début de ma thèse (*Cx. quinquefasciatus*, CpipJ2 souche *Johannesburg*) était donc composé d'un grand nombre de *contigs* (séquences d'ADN chevauchantes résultant de l'assemblage de *reads*), et *très* loin de la résolution chromosomique du génome d'*An. gambiae*. De même, l'annotation de cet assemblage était essentiellement prédictive (basée sur des modèles théorique de gènes et annotée via un algorithme) ; elle n'avait pas été vérifiée, par exemple par des analyses du transcriptome. Fort heureusement pour moi, un nouvel assemblage mieux résolu à l'échelle du chromosome a rapidement été rendu public après le début de ma thèse, ce qui a *considérablement* facilité mes analyses ; néanmoins, son annotation n'a pas franchement progressé. Les différences entre les génomes de référence disponibles pour les moustiques *Anopheles* et *Culex* ne s'expliquent pas que par leur importance en tant que vecteurs, résultant en un investissement asymétrique dans l'étude de leurs génomes, mais aussi par les tailles et complexités

[4] Ce qui a en revanche l'avantage de m'avoir permis de travailler avec quelques-unes parmi les plus importantes d'entre elles en toute confiance, loin de l'ambiance de compétition que j'ai pu observer pour la communauté étudiant les *Anopheles*.

respectives desdits génomes. Avec 573 Mb contre 281Mb, le génome de *Cx quinquefasciatus* est en effet deux fois plus long que celui d'*An. gambiae.* Cette différence de taille est, pour une petite part, liée à une légère extension du nombre de gènes chez *Cx. quinquefasciatus*, en particulier ceux jouant un rôle dans les réponses immunitaires et l'oxydoréduction (Arensburger *et al.,* 2010). Cette multiplication pourrait s'expliquer par la tendance qu'a ce moustique à coloniser les eaux insalubres et polluées (excès de matières organiques, résidus de biocides et autres composées exogènes…) dans lesquelles très peu d'autres espèces survivent[5]. Cependant, la raison principale du doublement de la taille du génome tient surtout à une activité nettement plus importante des éléments transposables (ET). Le génome de *Cx. quinquefasciatus* est en effet considérablement plus riche en ET : on passe de 10-12% pour *An. gambiae*, à 40-45% pour *Cx. quinquefasciatus* (de Melo & Wallau, 2020 ; H. Alout *pers. com.*)[6]. Ceci rend son assemblage (et fatalement son étude, j'y reviendrai plus tard) *nettement* plus complexe, notamment à petite échelle.

Identité du génome de référence.

Vous aurez peut-être remarqué que je ne mentionne qu'un assemblage de référence pour chaque genre, *An. gambiae* pour le genre *Anopheles*, et *Cx. quinquefasciatus* pour *Culex.* Il est en effet fréquent d'utiliser le génome d'une espèce proche comme référence pour une étude; l'assemblage d'un génome est un travail à part entière, long et fastidieux, et l'utilisation d'une référence proche est souvent suffisante (ou en tout cas considérée comme telle…) pour les analyses se basant sur la recherche d'orthologies entre espèces. Toutefois, suivant l'objectif et l'échelle à laquelle est menée l'analyse, cette pratique peut rapidement devenir un facteur limitant dans l'obtention de résultats sûrs et facilement interprétables, comme cela a été le cas ici : en m'intéressant à la structure de duplications retrouvées dans des espèces différentes, je devais en effet d'une part mener une analyse à une échelle génomique très réduite, qui par conséquent requiert une grande précision d'assemblage local, et d'autre part composer avec les divergences entres espèces-sœurs au sein de chaque genre. Mes sujets d'études appartiennent en effet à des groupes d'espèces aux relations taxonomiques complexes et échangent souvent des gènes entre elles (promis, je vous en parlerai en détail avant le chapitre suivant, c'est passionnant !). Si l'analyse de duplications provenant de populations *An. coluzzii* sur une référence *An. gambiae* n'a pas posé de

[5] Je vous laisse imaginer les joies que représente le terrain quand on veut récolter des échantillons de larves.
[6] Ca pourrait être pire notez, j'aurais pu travailler sur des *Aedes* : par exemple, près de 50% du génome est composé d'ET chez *A. aegypti* (Tu & Coates,2004 ; Nene *et al.,* 2007 ; Arensburger *et al.,* 2011).

problèmes, la réciproque n'est pas vraie chez *Culex*, où l'analyse des souches d'origine *Cx. pipiens* s'est rapidement révélée plus problématique que celles provenant de *Cx. quinquefasciatus*.

I.2. Les données génomiques brutes.

Trois souches porteuses de duplications hétérogènes d'origine diverses (Tab. I.1) (D_1, D_2, D_3) ont été créées à l'ISEM de sorte qu'elles partagent un même fond génétique, celui de la souche de référence sensible (SLAB ; voir Labbé *et al.,* 2007). Elles ont été séquencées en 2016, et re-séquencées en 2021 (*Illumina, pair-end, short-reads* dans les deux cas), en ajoutant en plus une nouvelle souche, SRQ, elle aussi introgressée sur le fond génétique de SLAB mais porteuse d'une duplication homogène, R^x, provenant de Martinique (Tab. I.1).

Concernant les données *Anopheles*, nous disposions d'une vaste banque de données sous la forme du *Anopheles gambiae 1000 genomes project* phase 2 AR1 (2017), un projet de séquençage de moustiques de populations naturelles appartenant au complexe *Anopheles gambiae s.l.* (*An. gambiae s.s.* et *An. coluzzii*). Ces données en accès libre comprennent plus de 1100 génomes de moustiques provenant de 13 pays d'Afrique Sub-Saharienne, et permettent donc une approche populationnelle de la question de la résistance aux insecticides. En plus de ces ressources, nous disposions de données génomiques (*Illumina* et *Nanopore*) provenant de populations naturelles d'*An. coluzzii* de Côte d'Ivoire.

II. Identification des duplications *ace-1*.

II.1. *Anopheles s. l.*

Article 1 : "Despite structural identity, *ace-1* heterogenous duplication resistance alleles are quite diverse in *Anopheles* mosquitoes." Jean-Loup Claret, Marion Di-Liegro, Alice Namias, Benoit Assogba, Patrick Makoundou, Alphonsine Koffi, Cédric Pennetier, Mylène Weill, Pascal Milesi et Pierrick Labbé. *in press in Heredity*.

Fin 2020, on ne connaissait qu'une seule duplication hétérogène *ace-1* dans le genre *Anopheles* (Ag-D_1, Assogba *et al.,* 2016). Au début de ma thèse, j'ai démontré qu'une diversité d'allèles D étonnante se maintiennent dans deux populations naturelles d'*An. coluzzii* de Côte d'Ivoire. Les premiers soupçons de l'existence de nouveaux allèles D sont nés suite à l'analyse des fréquences génotypiques *ace-1* des populations de Yamoussoukro et Yopougon. Nous savions déjà qu'un allèle D, très probablement Ag-D_1, ségrège dans ces

populations, car nous détections de façon récurrente un excès significatif de phénotypes hétérozygotes apparents par rapport aux fréquences génotypiques attendues sous panmixie. Pour le confirmer, nous avons utilisé un test PCR spécifique de l'allèle Ag-D_1 développé par Assogba *et al.,* (2016), et nous avons effectivement trouvé de nombreux individus porteurs de cet allèle. Notre étude a montré néanmoins qu'au moins un allèle non-dupliqué sortait également positif à ce test qui porte sur la séquence de la copie sensible de la duplication ; toutefois les individus qui en étaient porteurs restaient suffisamment rares pour que nos estimations de fréquences soient robustes.

A notre grand étonnement, la fréquence de l'allèle Ag-D_1 ne permettait cependant pas d'expliquer complètement les excès d'hétérozygotes observés : une fois la fréquence d'Ag-D_1 prise en compte, des excès d'hétérozygotes persistaient de manière significative dans trois des six points géographiques et temporels étudiés (voir **Claret et al., in press ;** Tab.I.2). Une origine purement technique (PCR) à ces discordances ayant été rapidement écartée, nous avons décidé d'étudier l'existence potentielle d'allèles dupliqués différents de Ag-D_1. Nous avons pour cela suivi un protocole pour le moins complexe et laborieux (voir les détails dans **Claret et al., in press**) permettant 1) d'identifier les porteurs de duplications hétérogènes différentes de Ag-D_1, et 2) de les séquencer pour évaluer la diversité de ces allèles, à la fois en termes de diversité au locus *ace-1* et en termes de structure génomique. Nous avons notamment testé une nouvelle approche développée par Namias *et al.* (2023) pour obtenir directement les haplotypes du gène *ace-1* à partir d'un séquençage Nanopore de produits PCR, une méthode nettement plus rapide et moins laborieuse que l'approche traditionnelle par clonage et séquençage Sanger. Pour la partie structure génomique, c'est sur ces données que je me suis exercé pour développer le *pipeline* détaillé dans la Box I.1.

Avec ces données, nous avons pu effectivement mettre en évidence la co-ségrégation de plusieurs allèles Ag-D dans les populations de Yamoussoukro et Yopougon, avec un polymorphisme maintenu depuis au moins 2012 (Fig. 2 dans **Claret et al., in press)**. Ces nouveaux allèles Ag-D partagent tous avec Ag-D_1 le même haplotype pour la copie résistante (qui reste l'unique décrit à ce jour, retrouvé également dans les allèles Ag-R^x). En revanche ils sont tous porteurs d'haplotypes différents pour leur copie S, et la totalité des mutations discriminant les copies sont synonymes ou se situent dans les introns (**Claret *et al., in press***). A l'aide du pipeline *ArDu*, nous avons aussi mis en évidence que tous ces allèles Ag-D étaient similaires (pour ne pas dire identiques) en termes de structure génomique, y compris avec les allèles Ag-D_1 et Ag-R^x (Fig. 3 dans **Claret et al., *in press***). Ceci a été confirmé ensuite par un test PCR spécifique de la jonction entre les amplicons développé pour les

Table I.1. Origine des souches "*ace-1*" et qualité des alignements. Le tableau présente les noms, les allèles *ace-1* et l'origine taxonomique de la souche associée, ainsi que les statistiques d'alignement sur le génome de référence en fonction des années de séquençage : % réf. couverte (pourcentage du génome de référence couvert lors de l'alignement), BASEQ (moyenne qualité de base) et MAPQ (moyenne de qualité d'alignement).

Souche	Origine	Référence
SLAB (référence sensible)	*Cx. quinquefasciatus* (Etat-Unis)	Georghiou *et al.,* 1966
Ducos / Cp-D_1	*Cx. quinquefasciatus* (Martinique)	Labbé *et al.,* 2007
Maurin / Cp-D_2	*Cx. pipiens* (Montpellier)	Labbé *et al.,* 2007
Biface / Cp-D_3	*Cx. pipiens* (Montpellier)	Labbé *et al.,* 2007
SRQ (R^x)	*Cx. quinquefasciatus* (Martinique)	Milesi *et al.,* 2022

Espèce	Allèle	Année	Duplication			Génome		
			% réf. couverte	BASEQ	MAPQ	% réf. couverte	BASEQ	MAPQ
Cx. quinquefasciatus	Cp-D1	2016	77.7	35.3	53.1	71.5	35.6	51.8
Cx. p. pipiens	Cp-D2		50.7	36.1	49.3	74.3	36.1	52.4
Cx. p. pipiens	Cp-D3		80.7	36.3	51.1	74.8	36.3	52.1
Cx. quinquefasciatus	Slab		84.4	35.4	52.9	75.2	35.2	52.6
Cx. quinquefasciatus	Cp-D1	2021	74	35.6	52.7	75.1	35.3	52.4
Cx. p. pipiens	Cp-D2		47.7	35.6	48.7	71.8	35.5	52.1
Cx. p. pipiens	Cp-D3		48.5	35.7	48.9	71	35.7	51
Cx. quinquefasciatus	Cp-R2x		71.3	35.5	53.2	71.8	35.5	51.8
Cx. quinquefasciatus	Slab		79.3	35.6	51.3	72.3	35.5	51.8

allèles Ag-R^x et Ag-D_1 (Assogba et al. 2016): tous les nouveaux allèles sont positifs pour ce test, ce qui indique que leurs amplicons sont en tandem et que les points de cassures sont identiques, ou ne diffèrent que de quelques paires de bases.

Je discuterai de ce que nous suggèrent cette étude sur l'origine et les processus permettant le maintien d'un tel polymorphisme dans le chapitre II ; dans le cadre de ce chapitre, ce sont surtout les conséquences en termes pratiques d'étude des *SV* que je souhaite aborder. En effet, cette étude très fine de la diversité de séquences et de structures chez *An. gambiae s.l.* n'a été rendue possible que par la qualité des ressources disponibles pour cette espèce. J'ai pu grâce à cela affiner mon *pipeline*, et surtout tester différentes approches de mises en évidence de ces duplications. Par exemple, j'ai montré que se baser uniquement sur la DoC de la mutation ponctuelle conférant la résistance (G119S) pour évaluer le nombre de copies R et S d'un allèle dupliqué est assez hasardeux. Une approche plus robuste est de combiner cette information avec celles de la couverture du gène *ace-1* rapportée à celle du chromosome et d'un gène de référence mono-copie. En effet, même avec une grande profondeur de séquençage et un génome de référence d'excellente qualité, la couverture locale peut s'écarter fortement de la couverture moyenne le long du génome et amener à des conclusions erronées. Cette rigueur dans la multiplication des mesures bioinformatiques pour obtenir des résultats robustes, bien qu'évidente à énoncer, me semble cependant être trop souvent négligée dans nombre d'études génomiques.

II.2. *Culex s.l.*

De l'importance de la qualité des données de séquençage.

Au début de ma thèse, je disposais donc des données brutes de séquençage des génomes complets de quatre souches, Slab (la souche sensible de référence), et trois souches porteuses d'allèles dupliqués différents (Cp-D_1, Cp-D_2, Cp-D_3) introgressés sur Slab. La même technologie de séquençage avait été utilisée pour les quatre souches : *Illumina* 125pb *paired-end reads*, 500pb *insert*, 30X de profondeur de couverture (DoC). Malgré des métriques de qualité de *reads* tout à fait satisfaisantes (distribution de taille, contenu en séquences adaptatrices, taux de GC), il s'est néanmoins rapidement avéré que ces données ne permettaient pas d'étudier la zone dupliquée convenablement. En effet, en suivant le même *pipeline* que celui développé pour *An. gambiae s.l.*, la caractérisation des duplications *Culex* à partir des données de séquençage de 2016 s'est soldée par un échec pour deux des trois souches dupliquées...

L'allèle Cp-D_1, originaire d'une population *Cx. quinquefasciatus*[7], est la seule souche dupliquée pour laquelle nous ayons pu observer une nette augmentation de la DoC (Fig. I.2). Pour Cp-D_2 et Cp-D_3, s'il nous semblait effectivement observer une légère augmentation de DoC cohérente avec la présence d'une duplication, il fallait une certaine dose de conviction pour la repérer, et il était impossible d'en délimiter les bornes, notamment du fait de fortes variations de *DoC* sur l'ensemble du chromosome 3 (Fig I.2). Pour tenter de comprendre pourquoi nous avions de telles incohérences (ça a été l'objet principal de mes travaux pendant les 12 premiers mois de ma thèse…), nous nous sommes d'abord penchés plus en profondeur sur les statistiques d'alignement: nous n'avons cependant pas pu identifier de différence marquante entre les données de séquençage des souches porteuses de Cp-D_2 et Cp-D_3 par rapport à Cp-D_1, si ce n'est que l'ensemble des souches présentaient des distributions anormales de tailles d'*insert* (distance entre les *paired-mate*), avec de fortes déviations par rapport au mode attendu de 300 pb pour Cp-D_3, ou des distributions décalées vers de larges tailles d'insert pour le reste des souches (Fig I.3). Ne semblant pas pouvoir incriminer la qualité de nos données de séquençage (ou alors de façon très marginale), nous avons alors suspecté que la divergence entre *Cx. quinquefasciatus* et *Cx. pipiens* pouvait être la cause de notre incapacité à observer les duplications Cp-D_2 et Cp-D_3 via un alignement sur un génome de référence *Cx. quinquefasciatus.* Pour tester cette hypothèse, nous avons tenté de limiter les séquences pouvant bruiter notre analyse. Nous avons donc réduit notre analyse des variations de *DoC* aux seuls exons; nous avons été rassurés d'observer une *DoC* normalisée (voir Box I.1) moyenne de 2.63 et 2.54 pour les exons *ace-1* de Cp-D_2 et Cp-D_3 respectivement, puisque ces valeurs, bien qu'élevées, s'approchent de celle attendue chez un homozygote dupliqué (*i.e.* proche de 2). En étudiant la *DoC* moyenne des gènes entourant *ace-1*, nous détections des valeurs proches de 2 pour au moins 5 gènes situés en aval sur le chromosome, ce qui était également compatible avec l'hypothèse de larges duplications embarquant plusieurs gènes, et avec ce qui était clairement observé chez Cp-D_1. Malheureusement, du fait de larges segments génomiques peu couverts, l'ajout des séquences introniques dans l'analyse résultait en une diminution de la *DoC* des gènes (1.71 et 1.58 pour *ace-1* chez Cp-D_2 et Cp-D_3 resp.) et mes espoirs d'arriver à identifier la structure de ces souches suivaient la même tendance.

[7] Pour rappel, Cp-D_1 est originaire d'une population appartenant à la même espèce que le génome de référence, *Cx. quinquefasciatus*, à l'inverse de Cp-D_2 et Cp-D_3 qui sont d'origine *Cx. pipiens.*

Table I.2. *Breakpoints* et contenu génique des duplications *Culex s.l.* Les *breakpoints* sont indiqués en gras pour chaque souche. Sont également indiqués le nom de chaque gène identifié et leur fonction (quand elle est connue).

Type	Début	Fin	Identité / gène	Fonction
Breakpoint 5'	**71634406**		**Cp-D_3**	-
Breakpoint 5'	**71641523**		**Cp-D_2**	-
cds	71641790	71643628	*ß-catenin like protein 1*	?
Breakpoint 5'	**71643959**		**Cp-D_1 / SRQ (R^x)**	-
cds	71658081	71773312	*ace-1*	Dégrade le neurotransmetteur acétylcholine dans les synapses neuronales.
cds	71783093	71798736	*Haemolymph juvenile hormone binding protein*	Chez *An. gambiae*, régule l'embryogénèse, le développement larvaire, et la maturation reproductive des adultes (Zalewska *et al.,* 2009)
cds	71812974	71900425	*Fibroblast growth factor receptor 3*	Interagit avec les facteurs de croissance des fibroblastes, déclenchant une cascade de signaux qui influencent la mitogenèse et la différenciation cellulaire. Régule la ramification neuronale chez *Caenorhabditis elegans* (Schutzman *et al.,* 2001)
cds	71836182	71836950	*BTB/POZ domain zinc finger*	Constitue un des types les plus communs de domaine de liaison de l'ADN (voir Collins *et al.,* 2001 pour revue)
cds	71906012	71962027	*Heparan-sulfate 6-O-sulfotransferase 1*	Chez *C. elegans,* joue un rôle dans la ramification neuronale *in vivo* (Tornberg *et al.,* 2011)
Breakpoint 3'	**~71967000**		**Cp-D_3**	-
cds	71974633	71983795	Protein of unknown function (DUF1777)	?
exon 1	71986306	71987430	SPRY domain-containing protein 7	Trouvé dans de nombreuses protéines eucaryotes avec un large éventail de fonctions. Impliqué dans d'importantes voies de signalisation telles que le traitement de l'ARN, la régulation de la méthylation de l'histone H3, l'immunité innée ou le développement embryonnaire (D'Cruz *et al.,* 2013; Diaz-Granados *et al.,* 2016)
Breakpoint 3'	**~71988000**		**Cp-D_2**	
exon 2	71988220	71988506	*SPRY domain-containing protein 7*	
cds	72000861	72038556	*protein couch potato*	Régule des transcripts spécifiques au système nerveux. La perte de fonction crée des adultes hypoactifs chez *Drosophila* (Bellen *et al.,* 1992)
Breakpoint 3'	**72055228**		**Cp-D_1 / SRQ (R^x)**	-

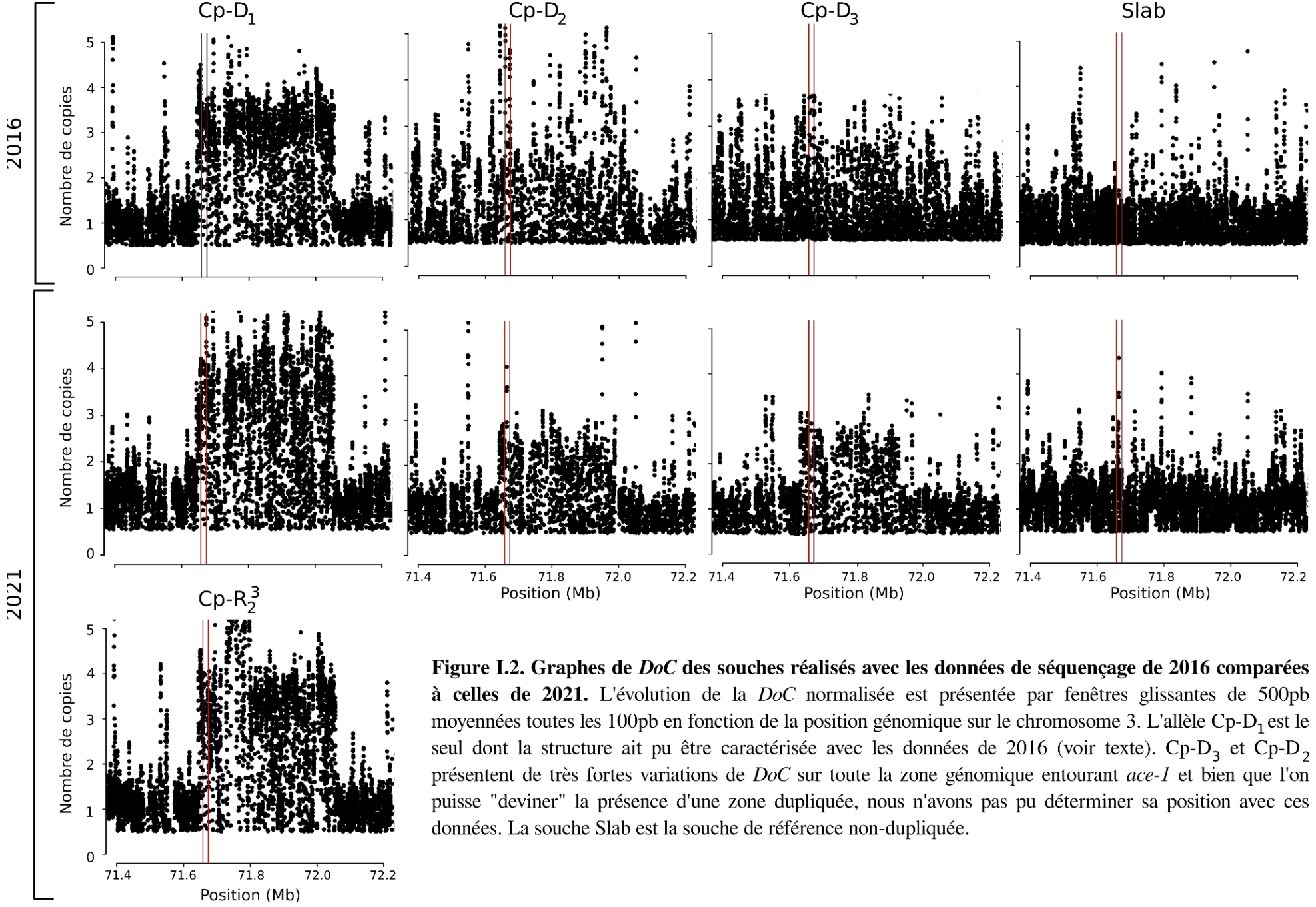


Figure I.2. Graphes de *DoC* des souches réalisés avec les données de séquençage de 2016 comparées à celles de 2021. L'évolution de la *DoC* normalisée est présentée par fenêtres glissantes de 500pb moyennées toutes les 100pb en fonction de la position génomique sur le chromosome 3. L'allèle Cp-D_1 est le seul dont la structure ait pu être caractérisée avec les données de 2016 (voir texte). Cp-D_3 et Cp-D_2 présentent de très fortes variations de *DoC* sur toute la zone génomique entourant *ace-1* et bien que l'on puisse "deviner" la présence d'une zone dupliquée, nous n'avons pas pu déterminer sa position avec ces données. La souche Slab est la souche de référence non-dupliquée.

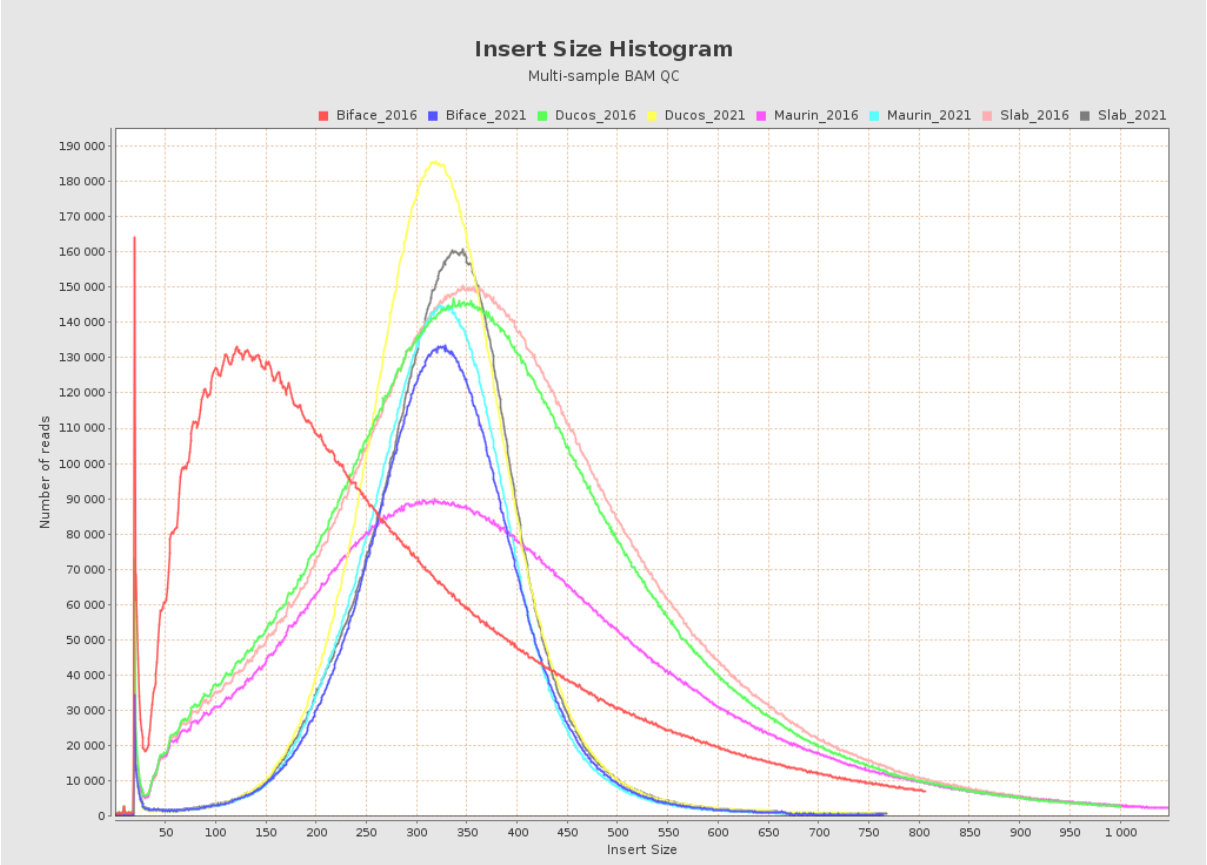


Figure I.3. Distribution de la taille d'*insert* des différentes souches pour les séquençages de 2016 et 2021. Figure obtenue à l'aide de l'outil *qualimap* (Okonechnikov, 2015). La souche Biface (Cp-D_3) présente une distribution très différente de l'attendu avec un mode proche de 125pb, pour un attendu de 300pb ; pour les autres souches (Maurin, Cp-D_2 ; Ducos, Cp-D_1 ; Slab, S), le mode de la distribution est plus proche de 300pb mais présente une variance beaucoup plus importante. Ces différences n'ont pas été retrouvées dans les données de séquençage de 2021, qui sont nettement moins variables.

Heureusement[8], dans le cadre d'une autre étude (**Milesi, Claret *et al.* 2022**), nous avons eu besoin de séquencer une nouvelle souche, SRQ, porteuse de l'allèle Cp-R_2^x (dont je parlerai plus en détail dans le prochain chapitre), aussi nous en avons profité pour séquencer de nouveau les souches porteuses de Cp-D_1, Cp-D_2 et Cp-D_3 (*Illumina* 150 pb *paired-end reads*, 300 pb *insert*, 30X *DoC*). En 2021, nous avons donc reçu les nouvelles données de séquençage pour les trois souches D, avec SRQ en plus (Tab. I.1). A notre grande surprise, l'essentiel des problèmes que nous avions rencontrés jusqu'à présent se sont résolus d'eux-mêmes : sans rien changer au *pipeline* d'analyse, nous avons été capables de produire des graphes de *DoC* montrant plus clairement les zones dupliquées, pour les allèles Cp-D (duplications hétérogènes) y comprit ceux de *Cx. pipiens*, comme pour la duplication homogène Cp-R_2^x (Fig. I.2). En plus de confirmer les résultats préliminaires obtenus pour Cp-D_1, nous avons ainsi été en mesure d'affiner la position des *breakpoints* pour tous les allèles, ce qui s'était avéré impossible avec les précédentes données. Un point à noter cependant est que, pour les données de 2016 comme pour celles de 2021, nous constatons toujours un effet de l'origine taxonomique des allèles dupliqués sur l'alignement des données génomiques dans cette région. En effet, les allèles S (Slab), Cp-D_1 ou Cp-R_2^x, tous issus de l'espèce *Cx. quinquefasciatus*, présentent des taux de couverture du génome de référence compris entre 70 et 80% (pourcentage moyen couvert: 77,34 ; écart-type 4.51), quand ceux de Cp-D_2 ou Cp-D_3 issus de l'espèce *Cx. pipiens*, sont plus proches des 50% (48,98 ; écart-type 1,27). Ces différences liées à l'origine des allèles ne se retrouvent pas en dehors de la zone dupliquée (*t*.test, *p*-value = 0.8), apportant la preuve directe de l'efficacité du protocole de *backcrossing* utilisé pour introgresser indépendamment chacun des allèles de résistance dans un fond génétique commun (Labbé *et al.,* 2007 ; **Milesi, Claret *et al.*, 2022**) .

Identification des points de cassure et diversité des tailles des duplications.

Pour toutes les souches dupliquées, prédire la position approximative des *breakpoints* était donc relativement aisé avec les données de 2021. J'ai cherché à les définir précisément en utilisant, comme pour *An. gambiae*, l'information des tailles d'*insert* et la position des *soft-clipped reads.* Le signal s'est avéré très clair pour la borne 5' des amplicons, mais les bornes 3' de Cp-D_2 et Cp-D_3 n'ont pas pu être identifiées aussi précisément. Je n'ai pas pu trouver de *soft-clipped reads* permettant de déterminer la position à la base près, aussi il persiste une incertitude d'1 kb sur la position des bornes 3' de ces deux souches (estimée

[8] … pour ma santé mentale et celle de tous-tes ceux et celles qui devaient subir mes lamentations, geignements incessants et doléances diverses.

Table I.3. *Breakpoints* et contenu génique des duplications *Culex s.l.* Nombres de copies estimées par qPCR. La position des amorces sur le chromosome 3 est indiquée pour chaque gène cible. Les valeurs de nombre de copies et d'erreurs correspondent à la moyenne des mesures de trois réplicats techniques sur trois réplicats biologiques (trois individus pour chaque souche à l'exception de Cp-D_2 pour lequel nous ne disposons que d'un individu homozygote).

Souche	Gène cible	Position	Nombre de Copies	Erreur
Cp-D_1	*Haemolymph juvenile hormone biding protein HJHBP*	exon 3 71783350 - 71788512	3.01	0.123
Cp-D_2			1.93	0.010
Cp-D_3			1.77	0.084
Slab			1.02	0.044
Cp-R_2[3]			5.23	0.268
Cp-D_1	*SPRY domain-containing protein 7*	exon 2 71986306 - 71987430	2.62	0.08.24E-02
Cp-D_2			1.77	1.35E-01
Cp-D_3			0.91	9.25E-02
Slab			0.86	2.64E-02
Cp-R_2[3]			2.70	1.02E-01
Cp-D_1	*Protein couch potato*	exon 7 72038425 - 72038556	2.93	1.74E-01
Cp-D_2			0.93	6.60E-02
Cp-D_3			0.96	6.30E-02
Slab			1.01	3.97E-02
Cp-R_2[3]			2.98	1.77E-01

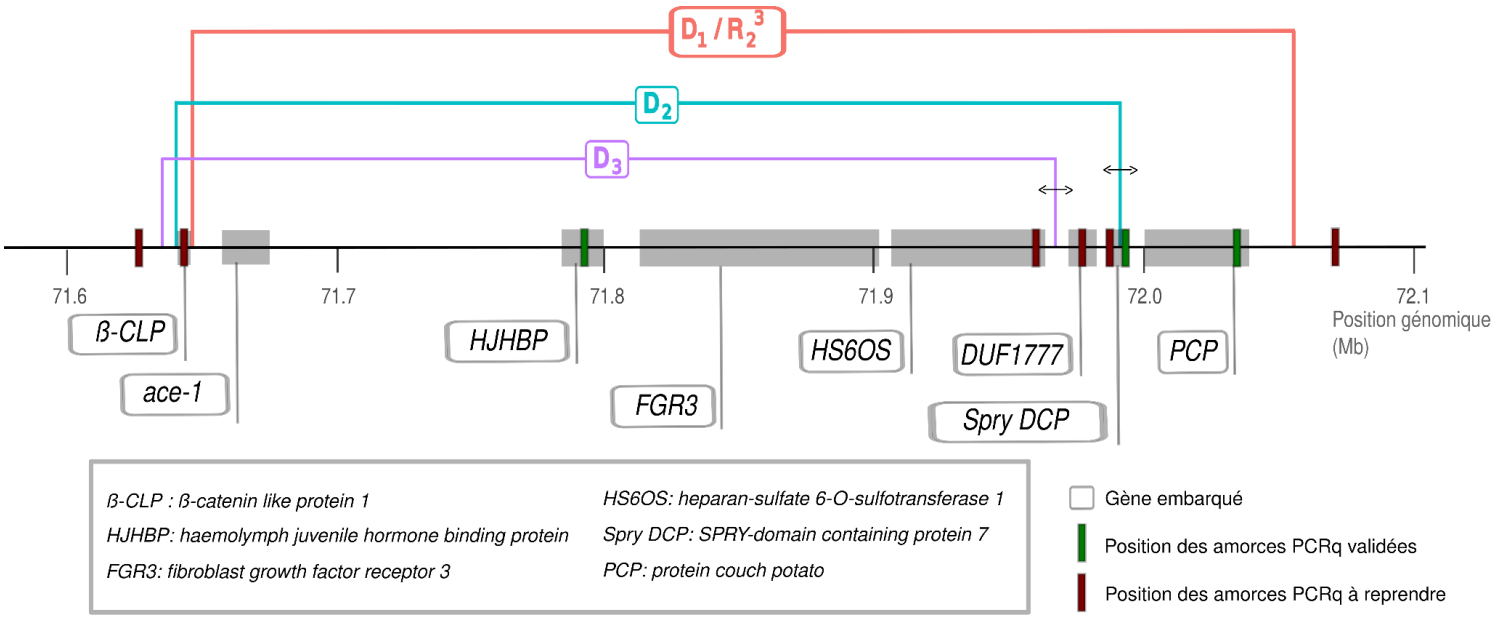


Figure I.4. *Breakpoints* et gènes embarqués des souches dupliquées *Cx. pipiens s.l.* Les positions génomiques des *breakpoints*, des gènes et des amorces utilisées pour la vérification par PCR quantitative sont indiqués le long du chromosome 3. Les amorces colorées en vert ont permis d'obtenir des mesures du nombre de copies par qPCR (voir texte et Tab. I.3), celles en rouge ont échoué.

entre 71,967 Mb et 71,968 Mb pour Cp-D_2, et entre 71,988Mb et 71,989 Mb pour Cp-D_3 ; Tab. I.2).

Néanmoins, ayant à ce stade acquis une certaine (et saine ?) méfiance vis-à-vis des résultats des analyses bioinformatiques, j'ai décidé de suivre la pratique habituelle du laboratoire[9], et de confirmer ces observations par des tests plus directs basés sur des PCR. J'ai ainsi mis au point des amorces spécifiques dans des zones conservées, ciblant des exons au sein et autour des duplications afin de mesurer le nombre de copies de ces zones par PCR quantitative. En distribuant ces amorces le long des duplications, j'ai ainsi pu confirmer les DoC prédites par la bioinformatique, ainsi que la taille des différents allèles Cp-D (Fig I-4 et Tab I-2). Malheureusement, les amorces élaborées pour le gène codant pour la *ß-catenin like protein 1*, situé entre la borne 5' de Cp-D_3 et de Cp-D_1/Cp-R_2^x, et pour le gène DUF1777 (Fig. I.4) n'ont pas fonctionné, certainement du fait d'un manque de spécificité. Il me faudra donc en sélectionner de nouvelles pour finir ces analyses et confirmer convenablement les estimations bio-informatiques[10]. Les bornes de Cp-D_1/Cp-R_2^x ayant pu être identifiées précisément, nous avons également conçu des amorces s'alignant de part et d'autre de la jonction entre deux amplicons, à l'image de la PCR jonction élaborée par Assogba *et al.* (2016) pour les duplications *An. gambiae s.l.*. Cette PCR a permis de confirmer les prédictions sur la position des *breakpoints* de ces deux duplications, mais aussi de confirmer que les amplicons qui les composent (deux pour Cp-D_1 et trois pour Cp-R_2^x) s'organisent en tandem.

Le premier résultat marquant de ces analyses est que, contrairement à ce qui avait été observé pour *An. gambiae*, les duplications décrites chez *C. pipiens s.l.* ne sont pas toutes identiques: les allèles Cp-D_1, Cp-D_2 et Cp-D_3 présentent tous des bornes différentes; en revanche, Cp-R_2^x (duplication homogène) et Cp-D_1 (duplication hétérogène), tous deux originaires de Martinique, partagent les mêmes bornes (Tab. I.2). Si les bornes 5' semblent plus conservées que les 3' entre duplications, leurs positions varient tout de même de plusieurs kb. Les duplications *Cx. pipiens s.l.*, se révèlent aussi bien plus larges que celles rencontrées chez *An. gambiae s.l.*, de 330 kb pour Cp-D_3 à 350 kb pour Cp-D_2, et même 411 kb pour l'allèle Cp-D_1 (et Cp-R_2^x), une variabilité qui contraste avec l'unique structure retrouvée pour l'ensemble des duplications *An. gambiae s.l.*.

Un second résultat marquant, là-encore en contraste avec les deux copies seulement (1 R, 1S) retrouvées pour tous les allèles Ag-D de Côte d'Ivoire, est que nous avons trouvé une

[9] Merci Mylène !
[10] A un moment, après le rendu de ce manuscrit.

hétérogénéité du nombre de copies pour les allèles Cp-D : si Cp-D_2 et Cp-D_3 présentent également deux copies (1R, 1S ; confirmant Labbé *et al.,* 2007), Cp-D_1 en comporte trois (2R, 1S). Il est également intéressant de constater que Cp-R_2^x présente lui aussi trois copies R identiques (confirmant **Milesi, Claret *et al.,* 2022**) : étant donné que les haplotypes *ace-1* R et S retrouvés au sein de Cp-D_1 ne diffèrent que pour la mutation responsable de la résistance (G119S), on peut émettre l'hypothèse que cet allèle serait issu de la réversion d'une des copies R de Cp-R_2^x en copie S plutôt que d'une duplication *de novo* (ce qui avait déjà été proposé, voir Labbé *et al.,* 2007).

Gènes embarqués : comprendre les différences de valeur sélective des allèles Cp-D ?

Pour rappel, les allèles Cp-D_2 et Cp-D_3 sont sublétaux à l'état homozygote, ce qui n'est pas le cas de Cp-D_1 ; nous savons par ailleurs grâce à des études antérieures que cela est probablement dû à des mutations embarquées dans la duplication et affectant d'autres gènes qu'*ace-1* (Labbé *et al.,* 2007 ; 2014 et Milesi *et al.,* 2018). J'ai donc étudié le contenu en gène embarqués de chaque duplication pour tenter de relier leur structure à leur impact sur la valeur sélective. Pour ce faire, j'ai d'abord utilisé l'annotation du génome *Cx. quinquefasciatus* disponible sur NCBI. Cependant, j'ai vite réalisé que cette annotation faite de manière automatique était loin d'être optimale: j'ai notamment rencontré des inadéquations entre les coordonnées fournies pour les gènes et les résultats d'alignement effectués pour *ace-1*. J'ai donc préféré refaire un alignement de la librairie de gènes NCBI (National Center for Biotechnology Information, 1988 ; en utilisant *Blast ;* Altschul *et al.,* 1990) de plusieurs moustiques et diptères (*Anopheles, Aedes, Culex, Drosophila*) sur la zone dupliquée, et le comparer à l'annotation (Tab. I.2). Cette analyse a surtout permis de mettre en évidence que le contenu en gènes des duplications varie en fonction des allèles D, les allèles les plus longs (Cp-D_1 et Cp-R_2^x) embarquant sept autres gènes en plus de *ace-1*, Cp-D_2 six (le *breakpoint* 3' se situant au coeur du 6ème gène), et Cp-D_3 seulement quatre (Tab. I-2). Parmi ces gènes embarqués, les fonctions de plusieurs sont connues, et j'ai pu en identifier qui semblent jouer un rôle majeur dans des processus métaboliques et développementaux : l'*Haemolymph juvenile hormone binding protein*, le *fibroblast growth factor receptor* ou la *protein couch potato* (Tab. I.2).

Ces variations dans le nombre et la nature des gènes embarqués dans les duplications sont fortement susceptibles d'impliquer des effets différents en termes de valeur sélective. En effet, un changement de dosage génique est généralement associé à une altération de quantité de la protéine qui en est le produit, qui peut perturber des équilibres biochimiques, par

exemple en rompant le dosage relatif entre différents gènes. Des variations dans le dosage de gènes codant pour des protéines ayant un rôle métabolique important pourraient donc altérer le métabolisme et affecter la valeur sélective (Veita, 2004). A ce stade, il est toutefois difficile de relier les différences de contenus en gènes embarqués aux phénotypes de sublétalité à l'état homozygote. En effet, il est par exemple surprenant que la duplication comportant le plus de gènes embarqués ($Cp\text{-}D_1$) soit la seule qui ne présente pas ce phénotype. Il aurait semblé intuitif qu'à l'inverse plus de gènes embarqués entraînent des effets sur la valeur sélective plus importants, d'autant plus que cet allèle possède trois copies. Le fait que le *breakpoint* 3' de l'allèle $Cp\text{-}D_2$ se trouve dans la séquence codante d'un gène (certainement sur le deuxième exon du gène *SPRY domain-containing protein 7,* voir Fig I.4 et Tab. I.2 et I.3) est une piste intéressante pour expliquer son phénotype homozygote sublétal, car ce n'est pas le cas de $Cp\text{-}D_3$ qui présente pourtant aussi ce phénotype : comme ces deux allèles complémentent, on cherche des altérations différentes. La nature des données de séquençage (*short reads*) ne permet toutefois pas de reconstituer avec certitude les haplotypes des différentes copies de ces gènes embarqués, et nous ne pouvons donc pas savoir si les protéines encodées sont conservées ou altérées. Il apparaît donc crucial à ce stade de compléter ces analyses par des approches de séquençage *long-reads*, mais aussi par des approches plus fonctionnelles (transcription, expression), afin d'explorer le lien entre phénotype et duplication. Le séquençage ARN des souches dupliquées à différents stades de développement est prévu prochainement pour répondre à ces questions (dans le cadre de la thèse de Maxime Prat) et pourrait donc apporter quelques lumières sur ce lien.

III. Conclusion sur les analyses bioinformatiques

Si j'ai choisi de présenter en détail les problèmes que nous avons rencontrés avec le premier jeu de données de séquençage et avec le génome de référence et les espèces, c'est principalement pour l'intérêt que je porte aux questions qu'ils nous ont fait nous poser sur l'aspect "boîte noire" que peuvent parfois prendre les analyses de génomique[11]. Toutes les métriques disponibles de qualité de *reads* indiquent que les données à notre disposition étaient de bonne qualité, mais nous nous sommes malgré tout retrouvés dans l'impossibilité de détecter ce qui s'est avéré plus tard être de larges duplications génomiques. Il est possible que la qualité de l'ADN extrait pour ces séquençages soit en cause ici (je n'ai pas retrouvé d'information sur la qualité de l'ADN reçu par l'entreprise qui s'est occupé du séquençage de

[11] En toute honnêteté, c'est aussi parce qu'il fallait que ça sorte. Identifier ces structures aura été au cœur de mes préoccupations pendant près d'un an !

2016), ou tout simplement que l'amélioration des technologies de séquençage puisse expliquer les différences de résultats que j'ai observé entre les deux jeux de séquences.

Un mot sur la confiance à accorder à mes analyses bio-informatiques.

Avec du recul, et considérant tous les points que j'ai détaillés plus haut, je suis assez étonné de l'adéquation entre les prédictions faites *via* des analyses de bio-informatique et les résultats des premiers tests PCR (incomplets, certes, mais c'est prometteur!). Je trouve également intéressant que cet exemple illustre si bien les défauts que peuvent avoir les analyses se basant uniquement sur une analyse bio-informatique et les difficultés liées à l'analyse fine de structures dans les espèces-soeurs.

Title: Despite structural identity, *ace-1* heterogenous duplication resistance alleles are quite diverse in *Anopheles* mosquitoes

Running title: *ace-1* duplications in *Anopheles* mosquitoes.

Authors: Jean-Loup Claret[1], Marion Di-Liegro[1], Alice Namias[1], Benoit Assogba[2], Patrick Makoundou[1], Alphonsine Koffi[3], Cédric Pennetier[4], Mylène Weill[1], Pascal Milesi[5,6], Pierrick Labbé[1,7]*

1. ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France

2. Medical Research Council, Unit The Gambia at London School of Hygiene and Tropical Medicine

3. National Institute of Public Health/Pierre Richet Institute, Bouake, Côte d'Ivoire

4.Institute of Research for Development (IRD), Marseille, France

5. Plant Ecology and Evolution, Department of Ecology and Genetics, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden

6. Science for Life Laboratory (SciLifeLab), Uppsala University, 75237 Uppsala, Sweden

7. Institut Universitaire de France (IUF), Paris, France

***** corresponding author: pierrick.labbe@umontpellier.fr

Word count: 8,336 words in main text

Abstract (206 words)

Anopheles gambiae s.l. has been the target of intense insecticide treatment since the mid-XXth century to try and control malaria. A substitution in the *ace-1* locus has been rapidly selected for, allowing resistance to organophosphate and carbamate insecticides. Since then, two types of duplication of the *ace-1* locus have been found in *An. gambiae s.l*. populations: homogeneous duplications that are composed of several resistance copies, or heterogeneous duplications that contain both resistance and susceptible copies. Heterogeneous duplications confer a permanent intermediate trade-off between the resistance induced by the mutation in the presence of insecticides and the disadvantages it incurs in their absence. So far, a single heterogeneous duplication has been described in *An. gambiae s.l*. populations (in contrast with the multiple duplicated alleles found in *Culex pipiens* mosquitoes). We used an innovative approach, combining long and short read sequencing with Sanger sequencing to precisely identify and describe at least nine different heterogeneous duplications, in two populations of *An. gambiae s.l*.. We show that these alleles share the same structure as the previously identified heterogeneous and homogeneous duplications, namely 203-kb tandem amplifications with conserved breakpoints. Our study sheds new light on the origin and maintenance of these alleles in *An. gambiae s.l*. populations, and their role in mosquito adaptation.

Introduction

Human activities have huge impacts on the environment, these diverse anthropogenic modifications can lead to spectacular adaptations in species subjected to them (see for example Otto 2018 or Hendry *et al.,* 2017). Among these, resistance to biocides (e.g. antibiotic resistance in bacteria or resistance to pesticides in crop pests and disease vectors) are probably the most studied and the best understood, because of their crucial impacts on economy and public health. From an evolutionary biology point of view, these are also major models to decipher the genetics of the adaptation (e.g. polygenic *vs* mono- or oligo-genic) or to understand how evolutionary processes shape these dynamics (e.g. spatial variation in selective pressure intensity, fluctuating selection overtime; Guillemaud *et al.,* 1998, David *et al.,* 2010, Milesi *et al.,* 2016). Notably, studying resistance revealed the complexity and the diversity of the genomic structural rearrangements underlying adaptations, well beyond the role of single nucleotide polymorphisms (SNPs). For example, cases of xenobiotic resistance linked to gene duplications are plenty (Devonshire & Sawicki, 1979, Leister, 2004, Labbé *et al.,* 2007, Kwon *et al.,* 2010, Patterson *et al.,* 2018). In the present study, we focused on one of these well-known models, the case of insecticide resistance in the malaria-vector mosquito *Anopheles gambiae s.l.*, and on the genomic nature and diversity of resistance alleles at the *ace-1* locus.

An. gambiae s.l. has been the target of intense insecticide treatment since the mid-XXth century, particularly on the African continent, to control malaria. While pyrethroids (PYR) are the most used insecticides, organophosphates (OP) and carbamates (CX) have also been utilised since the 1950's. They target acetylcholinesterase (AChE), an enzyme that regulates the activity of the synaptic neurotransmitter acetylcholine (Weill *et al.,* 2003). A unique substitution in the AChE-encoding gene *ace-1,* resulting in a glycine to serine substitution at the 280 codon of the protein (G280S). The substitution enables resistance to OP/CX by hindering their binding to AChE (R allele, Weill *et al.,* 2004), often referred to as the G119S mutation according to its position in the homolog gene of *Torpedo californica,* where AChE structure was first elucidated (Schumacher

et al., 1986). A nomenclature we will adopt here. This mutation has been independently selected for in multiple mosquito species (Weill *et al.,* 2003; Huchard *et al.,* 2006). However, the enzymatic activity of the protein encoded by the R allele is 60% lower than its wild-type susceptible counterpart (S allele; Bourguet *et al.,* 1997, Alout *et al.,* 2008, Labbé *et al.,* 2014). As a result, the R alleles are selected against in the absence of OP/CX insecticides. So far in *An. gambiae s.l.*, G119S is the only SNP mutation that has been found responsible for OP/CX insecticide resistance, introgressed from *An. coluzzii* to *An. gambiae s.s.* (Weill *et al.,* 2003, Djogbenou *et al.,* 2008, Assogba *et al.,* 2016, Grau-Bové *et al.,* 2021; a couple of other mutations have however been found in other mosquito species, Alout *et al.,* 2007a, 2007b).

Other types of mutations, *i.e.* structural variants (SVs), have also been selected in response to the use of these insecticides. Two types of duplication of the *ace-1* locus have been found in *An. gambiae s.l.* (Fig. 1A): i) homogeneous duplications, *i.e.* composed of several R copies, *i.e.* (R^x alleles; Assogba *et al.,* 2016, Grau-Bové *et al.,* 2021), or ii) heterogeneous duplications, containing both R and S copies (D alleles; Assogba *et al.,* 2015, 2016, Grau-Bové *et al.,* 2021). R^x alleles confer higher resistance levels and are favoured in highly-treated areas. Yet the heterogeneous duplications (D alleles) enable the fixation of the heterozygous phenotype (Labbé *et al.,* 2014, Assogba *et al.,* 2015, Milesi *et al.,* 2017), *i.e.* an intermediate trade-off between resistance in the presence of insecticides and disadvantage in their absence (or "selective cost", but see Lenormand *et al.,* 2018). While R^x and S remain respectively the fittest alleles in highly-treated and non-treated areas, D alleles are the fittest in populations exposed to intermediate selective pressures, in those areas exposed to reduced concentrations of treatments *per se*, or to temporal or geographical variations in treatment intensity (Labbé *et al.,* 2014, Milesi *et al.,* 2017).

Diversity in duplicated alleles (and more generally in copy-number variations or CNV) can result from two types of variation: i) variation in the DNA sequence of the amplicons, in particular the *ace-1* haplotypes, and/or ii) copy-number variations (for example R^x alleles carry different copy-numbers of the same haplotype). In the present study, we focussed on haplotype variations:

92 we will refer to alleles carrying different *ace-1* haplotypes as “sequence-alleles”, and those
93 differing in number of copies as “copy-number-alleles”. In *An. gambiae s.l.*, a single R sequence-
94 allele, but with multiple copy-number-alleles or R^x alleles, was found in several African countries
95 (Assogba *et al.,* 2016). Similarly, only one D sequence-allele, named Ag-D_1 (thereafter D_1 for
96 simplicity), has been formally described, using direct sequencing of cloned fragments of the *ace-*
97 *1* gene (real haplotypes, Djogbénou *et al.,* 2008). D_1 carries two *ace-1* copies, one R copy (the
98 haplotype is identical to that found in R^x alleles) and one S copy, in 203-kb tandem amplicons
99 (Assogba *et al.,* 2016). D_1 is found all over West Africa (Assogba *et al.,* 2018), which is in sharp
100 contrast with *Cx. pipiens*, where several different D sequence-alleles are often found segregating
101 in the same population (Milesi *et al.,* 2018). Grau-Bové *et al.,* (2021) recently analysed a large
102 dataset of Illumina paired-end genomes of *An. gambiae s.l.* from all over Africa (The *Anopheles*
103 *gambiae* 1000 Genomes Consortium (2021): Ag1000G phase 3). Based on the variations in depth
104 of coverage of the alternative bases at position 119 (*i.e.* of R or S sequences), they suggested
105 that several copy-number-alleles, differing in their numbers of R and S copies, might actually be
106 segregating in Africa, while the nature of their data (*i.e.* short-read sequencing) was unreliable to
107 assess whether multiple D sequence-alleles were segregating in the African populations.
108 Evidence for the existence of diversity in D alleles could help us understand the origin of
109 these mutations, and more generally how SVs are selected for short-term adaptation. So, we
110 adopted a comprehensive approach to assess the *ace-1* haplotype diversity and the number of R
111 and S copies in heterogeneous duplications and used information from various sources to
112 understand the origin of this diversity. As previous studies have shown that the frequency of D
113 alleles in *An. gambiae s.l.* was particularly high in Ivory Coast (Assogba *et al.,* 2018) and as Grau-
114 Bové *et al.* (2021) suggested that several D copy-number-alleles could segregate there, we
115 analysed the structures and diversity of the heterogeneous duplications present in two natural
116 populations of Ivory Coast, Yamoussoukro and Yopougon. By screening samples from Assogba
117 *et al.* (2018), we found the presence of the D_1 allele alone could not explain the observed
118 frequencies of D alleles. By cloning and sequencing a large part of the *ace-1* locus for several

individuals, we obtained the various haplotypes of each of their D(R) and D(S) copies. We also tested a more recently developed, and logistically easier approach, based on long-read sequences of PCR products (Namias *et al.,* 2023). We revealed that at least nine different *ace-1* D alleles segregate in these two populations. Using whole genome sequencing, we showed that at least five of these alleles share the exact same structure, two 203-kb tandem amplicons, with the exact same breakpoints. Finally, we discuss what these findings suggest in terms of duplication origin, but also the role of SVs in the adaptation process.

Material & Methods

Sampling and identification

We focused on two localities from Ivory Coast where the presence of the D_1 heterogeneous duplication has already been documented (Assogba *et al.,* 2018, Grau-Bové *et al.,* 2021), and where resistance has been monitored since 2012. We first used DNA previously extracted from adult mosquito samples, collected in 2012, 2015, 2016, and preserved at -80°C in the lab (Assogba *et al.,* 2018). To assess the frequencies at the time of the study, we collected new fourth instar (L_4) larvae samples in the same sites, in 2019. They were identified as *An. coluzzii* through a multiple PCR protocol: the first PCR was able to discriminate *An. arabiensis* from *An. gambiae s.s.* and *An. coluzzii* (Supporting information Tab. 1, “Species”; Scott *et al.,* 1993), and a second one distinguishing *An. gambiae s.s.* from *An. coluzzii* (Supp. Info. Tab. 1, “Form”; Favia *et al.,*1997).

DNA extraction and PCR conditions.

We extracted DNA from individual L_4 following a protocol modified from Collins *et al.* (1987). Briefly, each larva was ground in 200µL CTAB buffer (100mM Tris HCL, pH8.0, 10mM EDTA,1.4M NaCl, 2% CTAB), then incubated for 15 minutes, at 60°C. 200µL of chloroform with

4% of isoamyl alcohol were added and the solution was centrifuged for 10 minutes at 8000 rotations/min. The supernatant was transferred to a new tube with 200µL of isopropanol, to precipitate DNA at room temperature. DNA was washed with 400µL of 70% ethanol after 10 minutes of centrifugation (10000 rotations/min), dried, and then rehydrated in 50µL H_2O.

The PCR tests described below were performed using the Promega PCR kit (Madison, Winsconsin, USA) with *ca*. 50 ng of genomic DNA into 40µL of PCR-mix and using the following: 94°C for 30s, annealing temperature for 30s, and 72°C for 1 to 2 min for a total of 30 cycles (primers and annealing temperatures are listed in Supp. Info. Tab. 1).

Heterogeneous duplication detection and frequency estimation

***ace-1 phenotyping*.** We performed the "*ace-1* phenotype" PCR-RFLP test described in Djogbénou *et al.,* (2008): it amplifies an 817bp sequence of the *ace-1* locus encompassing the resistance-diagnostic G119S mutation. This mutation generates an *Alu I* restriction site and enables the distinction between three phenotypes: resistant homozygous [RR], susceptible homozygous [SS], and heterozygous [RS] (Fig. 1B-1). However, it does not enable the differentiation between standard heterozygous individuals for single-copy alleles (RS) and individuals carrying a heterogeneous duplicated allele (D), as D alleles associate both susceptible D(S) and resistant D(R) haplotypes of the *ace-1* locus (Assogba *et al.,* 2015). Alleles with multiple identical copies (e.g. R^x) also cannot be distinguished from alleles carrying the same haplotype but as single-copy.

***Estimation of D allele frequencies*.** D allele frequencies can nevertheless be inferred from the phenotypic frequencies (Tab. 1), as their presence in a population causes an excess of [RS] relative to what is expected under Hardy-Weinberg equilibrium (Lenormand *et al.,* 1998). We took advantage of this observation and used the same approach as in Assogba *et al.* (2018) to compute the D allele frequencies, independently for each year and location, implementing the

maximum-likelihood approach developed by Lenormand *et al.* (1998). We calculated the log-likelihood, *L*, of observing all the data as follow:

$$\square = \sum_{\square} \square_{\square\square\square} \square\square(\square_{\square\square\square})$$

with n_{ijt} and f_{ijt}, the observed number and the predicted frequency of individuals with phenotype *i* in population *j* at time *t*, respectively. (L_{max}) was maximised for each sample using a simulated annealing algorithm (Labbé *et al.,* 2009; Lenormand, *et al.,* 1998). The support limits (SL, equivalent to 95% confidence intervals) were defined as the minimum and maximum allele frequencies that did not significantly decrease the likelihood. Recursions and likelihood maximization algorithms were written and compiled with Lazarus v1.0.10 (http://www.lazarus.freepascal.org/).

Discriminating new D alleles from D_1 allele and standard heterozygotes. Before this study a single D allele had been characterised in *An. gambiae s.l.*, referred to as D_1 (Assogba *et al.,* 2018, Grau Bové *et al.,* 2021). For each population and year, we tested whether this allele alone could explain the estimated frequency of D alleles, or if more alleles (hereafter D_i) could segregate in the populations. To do so, we used a PCR-RFLP test specific to the D_1 susceptible *ace-1* copy ("D_1" PCR-RFLP-test, Supp. Info. Tab. 1; Assogba *et al.,* 2015) on all individuals with an [RS] phenotype (those that could harbour a D allele): the same 817bp fragment of the *ace-1* gene is PCR-amplified and an *AvaI* restriction site specific to the D_1(S) copy distinguishes between D_1 carriers ([D_1+] phenotype) and individuals that do not carry D_1 ([D_1-] phenotype; Fig. 1B-2). We then compared a model considering only three alleles (R, S, D_1) with models considering either four (R, S, S_{D1}, D_1) or five alleles (R, S, S_{D1}, D_1, D_i), using likelihood ratio tests (Labbé *et al.*, 2009; Milesi *et al.*, 2016); we took into consideration the possibility of occurrence of single-copy susceptible alleles S_{D1}, carrying the same *AvaI*-diagnostic mutation as the D_1(S) copy (Tab. 2). Our goal was then to characterize the *ace-1* haplotypes present in these potentially new D_i alleles. We first sequenced (Sanger sequencing ABI 3500 xL, Applied Biosystems by Thermo Fisher

196 Scientific) the 817bp *ace-1* PCR product for all the [RS D_1-] (Fig. 1B-3), *i.e.* individuals that could
197 carry a D but not D_1 (except for three controls). If, as expected, these individuals harbour at least
198 one D_i allele (genotypes D_iR, D_iS or D_iD_j), then up to three *ace-1* haplotypes should be present in
199 the PCR product (D_i(R), D_i(S) and another one), which would result in SNPs with multiple peaks
200 in the Sanger sequence (e.g. two peaks for an heterozygote). Providing the different haplotypes
201 carry different SNPs, one could expect diagnostic “triple peaks” (positions at which three different
202 SNPs can be found) in this mix sequence. This enables discrimination of D_i carriers from standard
203 RS heterozygotes. Although powerful to detect new D alleles, this approach can lead to an
204 underestimation of the new D_i allele frequencies: when the *ace-1* haplotypes are similar between
205 D alleles and single-copy resistance or susceptible alleles, triple peaks would not be detected.
206
207 **Heterogeneous duplication diversity.**
208 To identify the different haplotypes present in the *ace-1* PCR products of D_i carriers (*i.e.* “triple-
209 peak” individuals), we used two approaches: i) classic Sanger sequencing, which requires a
210 preliminary TA cloning step to provide individual haplotypes from the mixed products, and ii) an
211 approach initially developed to assess the diversity of *Wolbachia cid* genes multigenic family,
212 which is less tedious and more sensitive than TA cloning (Namias *et al.*, 2023). In this approach,
213 the PCR product is directly sequenced using Nanopore long-reads: each read then corresponds
214 to one individual haplotype.
215
216 ***TA cloning/ Sanger sequencing*.** We purified the *ace-1* PCR products of 22 “triple-peak”
217 individuals, using the BS664-250 Preps EZ-10 Spin Column PCR Purification kit (New England
218 BioLabs, Evry France). The purified products were then cloned (TOPO TA Cloning Kit pCR 2.1-
219 TOPO Vector and TOP10F' invitrogen bacteria). For each individual, we genotyped 24 clones
220 (Supp. Info. Tab. 2) using the *AluI* RFLP test (to discriminate R and S copies). All R haplotypes
221 recovered so far in *An. gambiae s.l.* were strictly identical on this 817bp fragment: we Sanger-

sequenced one resistance clone (*i.e.* D(R) or R), and at least 11 susceptible clones (*i.e.* D(S) or S; ABI 3500 xL, Applied Biosystems by Thermo Fisher Scientific).

***Nanopore sequencing of ace-1 PCR products*.** We directly sequenced the purified ace-1 PCR product of 12 “triple-peak” individuals (with a mean ≈1000X coverage for each individual) using Nanopore long-reads technology to capture, in a single read, each full 817 bp amplicon. Six of these individuals, which had been previously analysed with the TA cloning/Sanger sequencing approach, served as controls to assess the reliability of the Nanopore sequencing-based approach (Supp. Info. Tab. 2). We then adapted the bioinformatic pipeline developed by Namias *et al.* (2023): reads were mapped on a reference file containing two reference haplotypes, one R and one S, using *minimap2 v.2.24* (Li, 2018) with the options *map-ont* and without secondary alignments. SNPs were then called using *bcftools 1.15*, with the *config-ont* option, and a minimum mapping quality of 10. Finally, haplotype phasing was performed using *WhatsHap 1.4* (Martin *et al.*, 2023) with the default options. As mentioned in Namias *et al.* (2023), some heterozygous SNPs on the S haplotypes were not called: although they were supported by a high number of reads, the read distribution did not fit with a diploid framework (with only two S copies). We used the script provided by Namias *et al.* (2023) to recover those SNPs.

***ace-1 haplotype trees*.** To assess the diversity of *ace-1* haplotypes obtained through the two approaches, we aligned and compared them using a Neighbour-Joining phylogram compiled by *MEGA* software (MEGA11: Molecular Evolutionary Genetics Analysis version 11; Tamura *et al.*, 2021). We added several known *ace-1* haplotypes to this phylogram: an R reference haplotype and the reference D_1(S) copy haplotype. We used the information from molecular tests, whole genome sequencing, *ace-1* PCR product long-read sequencing and haplotype frequency to discriminate the haplotypes corresponding to a D(S) copy from those corresponding to a single-copy S allele, in each individual.

To infer the geographical origin of the newly characterized D alleles, we then added publicly-available *ace-1* S haplotypes to our alignment, from samples collected in the same time period as or samples and in Ivory Coast or nearby countries (Ghana, sequences from Weetman *et al.,* 2015; and Benin, Burkina Faso and Ivory Coast from Assogba *et al.,* 2018; see Supp. Info. Tab. 3). We then computed a maximum likelihood phylogram using a Tamura 3 parameters model with gamma distributed invariant site (G+I) mutation rates (best model fit determined with MEGA11). The phylogram was then plotted using the *ggplot2* (Wickham H. 2016) and *ggtree* packages (Xu *et al.*, 2022) using the R software (v.4.2.2, R Core Team, 2022; https://www.R-project.org/).

Duplication architecture

***Genomic characterisation*.** To determine the structural architecture of the newly characterised D_i alleles, we followed the protocol developed by Assogba *et al.* (2016) for D_1. The whole genomes of twelve "triple-peak" individuals were sequenced using Illumina paired-end sequencing (WGS, 150bppb reads, 350 bppb insert size; Supp. Info. Tab. 2). Reads from each individual were mapped to the *An. gambiae* PEST reference genome assembly (AgamP4.13; https://www.vectorbase.org) using the *bwa* (*-mem*) algorithm (Li & Durbin, 2009). The per-base depth of coverage (*pb*DoC) between positions 3,436,000 and 3,639,000 of the 2R chromosome (*ace-1* lies between positions 3,484,107 and 3,495,790) was obtained using the *samtools* suite (Danecek *et al.,*2021). We then standardized them ($pbDoC_{std}$), dividing each by the average μDoC calculated over the whole 2R chromosome ($pbDoC_{std}$ = *pb*DoC/μDoC) and plotting the $pbDoC_{std}$ along this chromosome, using R software. It allowed a fine scale observation of the structure of duplications (location, size, gene copy number, etc.). To determine the precise position of the duplication breakpoints, we isolated reads mapping at ±1 kb from the putative breakpoints determined from the $pbDoC_{std}$ graphs and analysed the insert-size among discordant paired-reads (*i.e.* paired-reads from each side of the junction between amplicons would map on each extremity of the amplicons, with an apparent insert-size equal to the amplicon size) and the

frequency of soft-clipped reads (*i.e.* the reads encompassing the junction and mapping partially on each extremity of the amplicons; see Assogba *et al.,* 2016, Fig. S1).

***Molecular validation of structural homologies*.** Assogba *et al.* (2016) developed a diagnostic PCR test for R^x and D_1 duplications, which amplifies a 460bp sequence overlapping the junction between the amplicons (Supp. Info. Tab. 1, "Junction"). We used this PCR to further assess whether the newly identified D_i alleles also shared the junction and the breakpoints of both D_1 and R^x alleles, using susceptible [SS] individuals as controls.

Results

Ivory Coast populations are highly polymorphic for D alleles.

We first analysed samples collected in two populations of Ivory Coast (Yamoussoukro and Yopougon) for which resistance was monitored between 2012 and 2016 (Assogba *et al.,* 2018) (Tab. 1). We built the first model (Model A) to analyse the "*ace-1* phenotype" data and found that the two populations showed a significant excess of heterozygotes compared to panmixia, suggesting the presence of D alleles at relatively high frequencies (>0.15; Tab. 2A). Using the specific "D_1" molecular test based on a single diagnostic mutation (Assogba *et al.,* 2015), we built a second model (Model B) taking this information into account (Fig. 1B-3), including the rare S allele, subsequently named S_{D1} and identified from [SS] individuals positive for this "D_1" test (Tab. 1). It showed that D_1 alone did not explain the observations as well as the third model (Model C) which considered several D alleles (model C fitted significantly better than model B for three out of the six populations, Yopougon 2015, and Yamoussoukro 2015 and 2016, p <0.05, Tab. 2C), with more coherent R and overall D frequencies (Tab. 2A and C *vs* B). This strongly supports the presence of at least one other D allele segregating in these populations. Interestingly, D_1 was not always the most frequent D allele (e.g. Yamoussoukro 2015 and 2016, Tab. 2C).

To identify the alleles present in these populations, we first Sanger-sequenced a fragment of the *ace-1* locus for RR individuals. Only one single resistance haplotype was found (already known from previous studies; Weill *et al.,* 2003; Djogbenou *et al.,* 2008; Assogba *et al.,* 2016; Grau-Bové *et al.,* 2021), *i.e.* no other R sequence-allele (although there were probably different copy-numbers-alleles; Assogba *et al.,* 2016; Djogbénou *et al.,* 2015).

The only other known resistance allele was D_1. We amplified the same PCR fragment in [RS, D_1-] individuals, (individuals that could carry a D but not D_1; *N* = 97, Tab. 1), plus three D_1 carriers as controls. The PCR product mixes several haplotypes, both R and S, which results in multiple peaks for SNPs in the Sanger sequence (Fig. 1A, see "Discriminating new D alleles from D_1 allele and standard heterozygotes", Materials and Methods). As expected, some individuals (N = 22) displayed several positions with triple peaks (henceforth "triple-peak" individuals), confirming the existence of at least three haplotypes (R and S) in the mix and the presence of other D alleles (D_i) segregating in these *An. gambiae s.l.* populations. We found six more "triple-peak" individuals in new samples collected in 2019 in Yopougon and Yamoussoukro (among 27 [RS] individuals analysed for each population), consequently, a total of 28 "triple-peak" individuals from 2012 to 2019 were identified.

To describe the diversity of the D_i resistance alleles, we used two approaches. First, Sanger sequencing of PCR products, which required a preliminary TA cloning step to get individual R and S haplotypes from the mix. As this protocol is tedious and difficult to apply to large numbers of individuals, we also tested another approach, using Nanopore long-reads sequencing of PCR products to directly access the various haplotypes carried by each individual (one read corresponds to one haplotype; Namias *et al.,* 2023). Over the 28 individuals "triple-peak", 22 were cloned and 12 were Nanopore sequenced (Supp. Info. Tab. 2). Six individuals among the cloned 22 served as controls for the Nanopore approach (Namias *et al.,* 2023): adapting the previously described pipeline, we were able to recover the exact same haplotypes with both approaches (the ≈ 300X high coverage of each haplotype allows easily correcting the

PCR and/or sequencing errors). This demonstrates the robustness of Namias *et al.*'s (2023) approach, which could be used to process much larger samples in the future.

As expected, for each "triple-peak" individual (carrying at least one D allele), three different *ace-1* haplotypes were identified, one R (119S), and two S (119G) with different SNPs. The R haplotypes of all individuals were identical to D_1(R) (the R copy carried by D_1) and to the haplotype recovered from RR individuals: as a result, a unique sequence carrying the G119S mutation is present in all resistance alleles, whether R or D. The different D resistance alleles differed only in the S copies they carried (or D(S)): we found 26 different S haplotypes, which could be a D(S) or a S single-copy allele (see Fig. 1 B "combined PCR"). They differed by only a few mutations, mostly found in introns, resulting in a relatively low divergence (d = 0.012; d_{exons}= 0.008 *vs.* $d_{introns}$=0.027). To discriminate the D(S) copies from the S alleles, we compiled a neighbour-joining phylogram with S haplotypes recovered from the 28 "triple-peak" individuals.

We then assigned the various haplotypes as follow:

i) We expected the D(S) haplotypes to be more frequent as they are directly selected for in the presence of insecticides (they provide resistance) and our protocol selected specifically for D-carriers; therefore, clusters of identical S haplotypes were more likely to correspond to D(S) copies than to S alleles (Supp. Info. Fig. 1A). Moreover, when several individuals had a first S haplotype in a cluster, their second S haplotype would often be different and attributed to single-copy S alleles.

ii) This approach was first validated by the observation of an expected cluster corresponding to D_1(S) (purple, Fig. 2). The second haplotypes found in the same individuals were different, as expected. For example, the individuals Yam19-11 and Yam19-24 (Fig. 2, Supp. Info. Tab. 4): two haplotypes (Yam19-11-S1 and Yam19-14-S1) were similar to D_1(S), but the two others (Yam19-11-S2 and Yam19-14-S2) were different; they were identified as D_1S individuals.

iii) Similarly, a second large cluster was found (pink, Fig. 2), which could unambiguously be assigned to a second D allele, henceforth, D_2.

iv) A third and a fourth smaller clusters were found, with respectively three and two individuals sharing one S haplotype. We tentatively named them D_3(S) and D_4(S) (dark green and orange, resp., Fig. 2).

v) Rather than inflating the number of potential D alleles, we chose to conservatively consider individuals carrying one haplotype similar to D_1(S) or D_2(S) as DS heterozygotes, even if the second S haplotype was found in another cluster, so they could actually be D_iD_j heterozygotes (e.g. Yam19-39 and Yam19-11 carried either D_1(S) or D_2(S), but their second S haplotypes clustered together; Fig. 2, Supp. Info. Tab. 4). This parsimonious approach was further supported by genomic analyses (see below) that confirmed three such individuals were DS heterozygotes (Yop16-41, Yop16-6, Yop16-16; Supp. Info. Fig. 2). Following this principle, we identified the D(S) copy of Yop12-45 as Yop12-45-S2, tentatively named D_5(S) (light blue, Fig. 2), because Yop12-45-S1 was identical to Yop12-54-S1, while Yop12-54-S2 was identical to D_1(S) (Fig. 2, Supp. Info. Tab. 4).

vi) For some individuals, the total number of *ace-1* copies they carried was independently known from genomic analyses (WGS, see Material and methods): Yop16-60 carried four copies (Fig. 3), but only three different haplotypes, one R and two S. Yop16-60 carried two different D alleles, it was a D_2D_3 heterozygote (Fig. 2): Yop16-60-S1 was identical to D_2(S), and Yop16-60-S2 belonged to the tentative D_3(S) cluster, which incidentally confirmed the existence of the D_3 allele.

vii) Conversely, genomics showed that Yop16-50 carried three *ace-1* copies (Supp. Info. Fig. 2) and was a DS heterozygote. As Yop16-50-S1 was identical to D_2(S), it strongly suggested that Yop16-50-S2 was the single-copy S allele. Yop15-49-S1, which was identical to Yop16-50-S2 (Fig. 2, Supp. Info. Tab. 4), was most probably an S allele too. Consequently, as Yop16-49 carried a duplicated allele (one R and two S haplotypes), it was the second S haplotype, Yop15-49-S2, that was the D(S) copy (despite being isolated in the tree; Fig. 2); this new allele has been named D_6.

viii) For the last three D-carriers, Yam15-41, Yam16-42 and Yop15-3, both S haplotypes have single occurrences in the tree (Fig. 2, Supp. Info. Tab. 4). Although each carried at least one new

D allele, hence D_7, D_8 and D_9, which S haplotypes corresponded to their D(S) copies remained undetermined.

From the 28 analysed triple-peak individuals, we were able to conservatively infer that at least nine different D sequence-alleles (including D_1) were segregating in Yamoussoukro and Yopougon. From the same 28 individuals, we also recovered 17 susceptible S alleles (Fig 2 and Supp. Info. Tab. 4). For comparison, only one other resistance haplotype, R, was found in our samples, the same as that encountered in across West Africa. Apart from D_1, D_2 and D_3, their D(S) haplotype identification remained however unsure. Different D alleles (up to four) were recovered in each population and year (Supp. Info. Tab. 4), indicating that the two populations remained polymorphic for these resistance alleles from 2012 to 2019 (> 70 generations). D_1, D_2 and D_3 in particular, were found over the whole period, while the other D alleles were found only once. However, it does not mean that these alleles have appeared and disappeared rapidly: our study protocol was designed to assess the overall diversity of D alleles, but not to infer their frequency, or even their presence over time (the sampling size was too limited, one to eight individuals per year and population, and limited to those displaying the triple-peak signal).

Of course, nine D sequence-alleles is a minimum estimate of the real diversity of the duplicated resistance alleles (our approach was not exhaustive). This relatively high diversity, segregating in only two Ivory Coast populations, immediately begs the question of their molecular origins: i) could a single original duplication event followed by secondary rearrangements (e.g. recombinations, deletions) have generated this diversity? or ii) are multiple independent events of duplication required, one for each allele? To try and answer this question, we used two approaches, the first one based on phylogeography, the second using genomics.

ace-1 haplotypes do not show structure at the geographical or species level.

We first tried to assess whether the different D alleles could be associated with a particular geographical origin in West Africa. We added around 30 *ace-1* S haplotypes from neighbouring countries (Benin, Burkina Faso, Ghana) of both *An. gambiae* and *An. coluzzii* and computed a

phylogenetic model (Fig. 4; Tamura 3 parameters G+I, see Materials). This revealed no clustering pertaining to geographical origin or species for the different S copies (whether S or D(S)). Despite testing different models of evolution, we found that close or identical S copies can be found in all countries, and that none of the different D(S) copies are associated with a particular country. In fact, the only highly supported nodes are those of the D(S) clusters (Fig. 4). The diversity is relatively low and mostly concentrated in the introns. Translating the exonic part of the *ace-1* haplotypes carried by D alleles showed that all mutations were synonymous, except G119S for the D(R) haplotypes.

Overall, we have no evidence if the different D alleles identified in the present study have originated in the populations where we found them or elsewhere, and cannot rule out the possibility of a unique origin for all.

All D alleles share a common genomic architecture.

We then took advantage of the bioinformatic approach developed by Assogba *et al.* (2016) to analyse the genomic architecture of these alleles (copy number, amplicon size, breaking points, etc.): we already knew that D_1 and R^x alleles share the same breaking points, but what about these new D alleles? Different genomic structures would support independent origins.

***Copy number and amplicon size**:* Twelve individuals carrying different D alleles (D_2, D_3, D_4, D_7, D_8, D_9) were sequenced using Illumina paired-end sequencing (Supp. Info. Tab. 2 and 4; we could not get enough DNA for individuals carrying D_5 and D_6). The reads were first mapped onto the reference *A. gambiae* PEST genome (Vector-Base; AgamP4.13). We also mapped short-reads from the susceptible reference strain Kisumu as a non-duplicated reference. For each individual, we calculated a standardised depth of coverage ($pb\text{DoC}_{std}$) for each base in a region surrounding *ace-1* (see Materials and Methods). As expected, Kisumu's $pb\text{DoC}_{std}$ remained close to 1 over the whole region (Fig. 3A). By contrast, all D-carriers displayed a consistent $pb\text{DoC}_{std}$ increase over a 203 kb region encompassing the *ace-1* locus, similar to that seen for D_1 (Fig. 3B and Supp. Info. Fig. 2). For 11 individuals we observed a 1.5-fold $pb\text{DoC}_{std}$ increase (1.5±0.14 for

the duplicated region *vs.* 1±0.16 for the flanking non-duplicated regions), which is consistent with a DS genotype (Supp. Info. Fig. 2). Yop16-60, displayed a 2-fold *pb*DoC$_{std}$ increase in the same area (Fig. 3B), suggesting a DD genotype (actually D_2D_3, see D(S) copies identification above).

***Breakpoint positions**:* We further combined information from insert-size among discordant read pairs, *i.e.* pairs overlapping the amplicon junction, and local enrichment in soft-clipped reads, *i.e.* reads overlapping the amplicon junction, to precisely map the breakpoints of the duplication. For all the analysed D alleles, i) the insert-size of the discordant read pairs showed the duplication to be 203 kb, and ii) a significant increase of soft-clipped reads was found on positions similar to the 5' and 3' breakpoint positions previously identified for D_1 (position 3,436,927 and position 3,639,836; resp.; Fig. 3B; Assogba *et al.,* 2016). We finally submitted these individuals to the specific PCR test designed by Assogba *et al.* (2016) that amplifies a 460bp pb fragment overlapping the amplicon junction in D_1 and R^x alleles. The fragment was amplified in all 12 individuals.

Together, this PCR test and the genomic analysis indicate that all the D alleles share the exact same genomic architecture, *i.e.* two amplicons only, with the same boundaries and sizes (without any internal deletion as seen in some R^x alleles, Assogba *et al.,* 2016, 2018).

Discussion

An unsuspected duplicated allele diversity: beyond the spotlight effect

So far, in *An. gambiae s.l.*, the only *ace-1* SNP that has been linked to OP/CX insecticide resistance is G119S (Weill *et al.,* 2003; Djogbenou *et al.,* 2008; Assogba *et al.,* 2016; Grau-Bové *et al.,* 2021; a couple of other substitutions have been found in other mosquito species e.g. Alout *et al.,*2007a, 2007b). Moreover, it appears that this mutation occurred only once in *An. gambiae s.l.* (introgressing from *An. coluzzii* to *An. gambiae s.s.*; Djogbenou *et al.,* 2008), so that a single R haplotype has rapidly spread over all West Africa (Weill *et al.,* 2003; Djogbenou *et al.,* 2008;

457 Assogba *et al.,* 2016, this study). Interestingly, no single-copy R allele has been found, only
458 homogeneous duplications, with various copy-number-alleles of repeated identical R haplotype
459 (R^x, Fig. 1A; Assogba *et al.,* 2016, Grau-Bové *et al.,* 2021).

460 Despite indications of more variations (Grau-Bové *et al.,* 2021), the only resistance allele
461 known in *An. gambiae s.l.* was one heterogeneous duplication, D_1, also found across all West
462 Africa (Djogbenou *et al.,* 2008; 2009; Assogba *et al.,* 2016): this allele carries one R and one S
463 copy, resulting in a [RS] phenotype using the usual molecular test (Fig. 1). Our study showed that
464 significant excess of apparent heterozygotes in two populations of Ivory Coast could not be
465 explained by the presence of D_1 alone (Tab. 2). Through a multi-approach genotyping and
466 sequencing protocol, we have further evidenced high diversity of *ace-1* resistance alleles in West
467 African *An. gambiae s.l.*, with eight new D sequence-alleles, found in only two Ivory Coast
468 populations. All D alleles carry one identical R haplotype, but carry a unique S haplotype, *i.e.* their
469 D(S) copy is different (similarly to most of the 27 different D alleles described so far in *Cx. pipiens*;
470 Milesi *et al.,* 2018). They also share the same genomic architecture as D_1, *only* two *ace-1* copies,
471 the same amplicon size and breakpoints. The D allele diversity is high compared to a unique R
472 sequence-allele, but it is only half that of single-copy S alleles, segregating in the same
473 populations (17 different S haplotypes, Supp. Info. Tab. 4).

474 The fact that only D_1 had been described since the seminal work of Djogbénou *et al.* (2008),
475 despite regular surveys (Djogbenou *et al.*, 2008; Assogba *et al.*, 2016, 2018; Grau-Bové *et al.*,
476 2021, Kouamé *et al.*,2023), highlights a bias in the way these variants are studied: the “classical”
477 approach, based on field-caught individuals, crossed in the laboratory with a reference
478 susceptible strain (Labbé *et al., 2007*; Assogba *et al.,* 2016; Milesi *et al.,* 2018), will only retain
479 genotypes frequent enough to be sampled, and more importantly, individuals fit enough to survive
480 and reproduce in the laboratory. It’s a major problem when studying duplicated alleles that are
481 often plagued with strong deleterious effects (Innan & Kondrashov, 2010; Schrider & Hahn, 2010;
482 Schrider *et al.,* 2013; Milesi *et al.,* 2018). On the other hand, surveys relying on specific molecular
483 tests for a few diagnostic mutations are prone to a strong “spotlight effect”, *i.e.* they can only find

what they are looking for, especially when these mutations are not directly causal for the resistance phenotype.

The last decade has seen a giant leap in sequencing and bioinformatics analyses based on NGS data, which are now affordable for extensive surveys of natural populations. However, there are also limitations when it comes to precisely describing (at the sequence level) structural variants in natural populations, as is the case in our study. For example, in Ivory Coast, Grau-Bové *et al.* (2021) suggested variation in the number of S copies in D alleles (*i.e.* copy-number-alleles) of which we found no evidence: all the D alleles identified in the present study carried only two copies, one S and one R haplotype. While we cannot exclude the existence of D copy-number-alleles, our analyses suggest that this discrepancy may come from how the number of copies were determined in these two studies. Identifying S/R copy number ratio only from the ratio of allelic coverage at a single diagnostic position (here the G119S point mutation, see Grau-Bové *et al.,* 2021) can lead to inaccurate copy-number estimations, especially with low depth of coverage (see simulations in Karunarathne *et al.*, 2023). In our study, using the average depth of coverage across the whole *ace-1* gene to assess the number of copies and to deduce the genotype, proved to be more reliable (Supp. Info Tab. 5). Similarly, despite indications that suggested the potential existence of D sequence-alleles, Grau-Bové *et al.*'s study did not allow their specific identification, because haplotype reconstruction is particularly difficult from short-read data when several copies are present. We demonstrated the potential of long-read sequencing to overcome this issue. We described the same haplotypes through direct long-read sequencing of the PCR mix and bioinformatics analyses and with the logistically heavy but reliable, TA cloning/Sanger sequencing approach. As long-read sequences become more accessible and reliable, some limitations may reduce, especially for structural variant detection and study (Mantere *et al.*, 2019; De Coster *et al.*, 2023, Namias *et al.*, 2023, although these methods are still limited for the amplicon size we are using; see Hook and Timp, 2023 for a review).

The molecular origins of D alleles remain to be confirmed, although secondary recombinations are likely

The surprising diversity of D resistance alleles found in two populations poses questions on their origin(s) in *Anopheles* mosquitoes.

Deletions/duplications are frequent for multi-copy genes (e.g. esterases, Milesi *et al.,* 2016), due to unequal recombination, so that the existence of different R^x copy-number-alleles is expected (note that all the copies have the same size exactly, although a secondary internal deletion has been described; Assogba et *al.*, 2016, 2018). The existence of several D alleles carrying one identical R haplotype and different S haplotypes could be explained by two different scenarios, as proposed for D alleles in *Cx. pipiens* (Labbé *et al.*, 2007, Milesi *et al.*, 2018). The first one requires multiple independent unequal recombination events in RS heterozygotes (scenario 1, Fig. 5). The second scenario (scenario 2, Fig. 5) only requires one unequal recombination event, followed by secondary recombination event between the D(S) copy bound in the duplication (or one R copy of a R^2 allele) and a single-copy S allele in heterozygous DS (or R^2S) individual. These secondary recombination events could be limited to the *ace-1* sequence or much larger (up to 203kb, *i.e.*, encompassing all the genes embedded in the duplication).

In *Cx. pipiens*, both scenarios are probably at play: both D(R) and D(S) copies differ between some D alleles, which is expected in scenario 1; however, many share the same D(R) copy, and the D(S) copies are found as single-copy S alleles in the same populations, a pattern expected in scenario 2 (Milesi *et al.*, 2018). In *An. gambiae*, all D alleles (as well as R^x) display a strict structural homology, *i.e.*, the exact same boundaries and breakpoints, which would require frequent and precisely localized *de novo* unequal recombination under scenario 1. While Assogba *et al.* (2016) did find a *harbinger* transposable element on the 3' end of the amplicon, it is not evident that it fits such recurrent recombination events in the same genomic area. Therefore, scenario 1 does not appear to be the most parsimonious. The diversity of D alleles we observed was more likely to be generated by secondary recombination events between S copies. Single-copy S alleles with close haplotype sequences are found in the same populations (Fig. 2), further

supporting the hypothesis that the diversity of D alleles was likely due to secondary recombinations, but the lack of geographic structure tends to weaken it (Fig. 4). To confidently evaluate those hypotheses, complete haplotypes of the duplicated alleles are required. This could soon be possible, with the improvement of long-read sequencing (but for the amplicon size, see above).

Nonetheless, both scenarios imply a high recombination rate in this genomic region, as these alleles are fairly recent in terms of evolutionary time: OP and CX insecticides have only been used for 50-60 years to control *An. gambiae s.l.* population; resistance was first reported in the late 1990's-early 2000's (Elissa *et al.,* 1994; N'Guessan *et al.,* 2003); the first reports of a D allele are even more recent (Djogbenou *et al.*, 2008).

How is the *ace-1* resistance alleles diversity maintained in the populations?

Several studies in *An. gambiae* or *Cx. pipiens* have shown that various resistance alleles at the *ace-1* locus conferred different fitness (e.g. Assogba et *al.*, 2015, 2016 in *An. gambiae*, Labbé et *al.*, 2007, 2014, Milesi et *al.*, 2018, 2022 in *Cx. pipiens*). For the homogenous duplicated alleles R^x, a larger number of copies (e.g. 5 *vs* 3) confers both a higher resistance level and higher selective disadvantages, although it does not follow a linear pattern (Assogba *et al.*, 2016 in *An. gambiae*, Milesi *et al.*, 2022 in *Cx. pipiens*). These alleles appear to be selected in areas exposed to intense selective pressure, the copy-number-allele diversity is potentially being maintained by small insecticide treatment fluctuations (Assogba et al., 2018). On the other hand, D_1, the only heterogenous duplicated allele known before the present study in *An. gambiae*, has been shown to confer a phenotype similar to standard heterozygotes RS: it provides more resistance than S but less than R^x alleles, and less selective disadvantages than R^x but more than S (Assogba et al., 2015, 2016), as observed for D alleles in *Cx. pipiens* (Labbé et al., 2014). These alleles are selected for in areas where insecticide treatments are moderate or fluctuating (Assogba 2015, 2016, Lenormand *et al.*, 1998, Labbé *et al.*, 2007, 2014, Milesi *et al.*, 2018, 2017).

In the present study, in all samples except those from Yopougon in 2012, the resistant heterozygous phenotype was more frequent than the susceptible phenotype, and the frequency of [RR] individuals was very low (maximum one individual per sample, Supp. Info. Tab. 4). The overall D frequency remained globally stable from 2012 to 2019, despite a slight increase in resistance (binomial test, p = 5.5x10^{-11} and p = 2.4x10^{-4} for Yopougon and Yamoussoukro, respectively; Supp. Info. Fig. 3). As the response to change in selective pressure is usually fast for resistance alleles (even seasonal; Lenormand *et al.*, 1999, Milesi *et al.*, 2016, Milesi *et al.,* 2017), it suggests that the insecticide treatments did not change much over the period of the study. The higher frequency of D over R^{x} among resistance alleles and the persistence of S alleles at relatively high frequencies, further suggests that these populations were exposed to moderate (or fluctuating) treatment intensities.

We found nine D sequence-alleles that differ only by their D(S) copy and up to four of them have been co-segregating in the same population over at least 70 generations (Supp. Info. Tab. 4). At first glance, these alleles are not expected to confer different resistance levels. Previous studies in *Cx. pipiens* suggests that the intermediate resistance level displayed by D alleles depends entirely on the association of an R and an S copy (*i.e.* one carrying the G119S mutation and the other not), but not on the sequence of the R or S alleles captured in the duplication (Labbé *et al.,* 2007, 2014, Milesi *et al.,* 2018). While D alleles are clearly selected for over the S and R^{x} alleles in conditions of intermediate selective pressures, the D sequence-alleles found here are expected to be neutral in terms of selective advantage. A first parsimonious explanation would be that the polymorphism observed in D alleles reflects the rate at which new D alleles are generated, *i.e.* the duplication rate - genetic drift equilibrium considering scenario 1, or the recombination rate - genetic drift equilibrium in scenario 2. In such cases, the diversity of D alleles should be similar to that of the S alleles, and their frequency spectrum should also be similar. However, the diversity of the D alleles is lower than the neutral expectancy provided by the diversity of S alleles found in the exact same individuals: in the 28 “triple-peak” individuals (56 alleles in total), we identified nine different D haplotypes over 29 D sequences, but significantly more S haplotypes (17 different

haplotypes over 27 S sequences; binomial test, χ^2 = 4.5, *df* = 1, *p* = 0.03). Similarly, some D alleles segregate at a much higher frequency than the S alleles (Tab. S4 and Supp. Info. Fig. 1B): for instance, D_2 and D_1 were found 11 and 8 times, respectively (relative frequency over all D alleles = 0.38 and 0.28, resp.; Supp. Info. Fig. 1B), while one S allele was found five times and the second most frequent S three times (relative frequency over all S alleles = 0.18 and 0.11, resp.; Supp. Info. Fig. 1B). Our limited sample size (1 to 8 D carriers per sample) and the fact that we captured a D_2D_3 heterozygote (under panmixia $f_{D2D3}= 2f_{D2}f_{D3}$), further suggests that the D alleles captured in this study must segregate at a higher frequency. It is unlikely that the D(S) copies would be a random sample of the S diversity, considering that our discovery approach was meant to maximise the diversity in D alleles while not affecting S allele diversity.

It appears that the observed diversity and frequencies of D alleles are probably not the results of neutral processes only. How could selection contribute to explaining the persistence of D resistance allele polymorphism at the population scale? In *Cx. pipiens* several populations are polymorphic for D alleles (Milesi *et al,.* 2018). These D alleles are associated with various deleterious effects expressed only in homozygotes and independent from *ace-1* (Labbé *et al.*, 2014) and they complement each other (each D allele compensates for each other's flaws, so that a D_iD_j heterozygote is fine; Labbé *et al.*, 2007, Milesi *et al.,* 2018). Simulations have shown that such alleles can be maintained in the same population by frequency-dependent selection (Milesi *et al.,* 2018). Gene duplications are extensive structural rearrangements that are known to be largely detrimental (Schrider *et al.*, 2013; Katju and Berthogsson, 2013), whether for structural reasons (e.g., gene disruption, deleterious mutation hitchhiking) or in relation to gene-dosage imbalance. It would not be surprising that similarly detrimental D alleles, that could complement each other, might be found in *An. gambiae s.l.* too. However, D_1 was not sublethal when homozygous in lab experiments (although less fit than S in absence of insecticides; Assogba *et al.,* 2015) and this frequency-dependent selection among D alleles should last only until S and R^x alleles are eliminated. On the other hand, S allele frequency remained relatively high and globally stable over the whole period of the study (Tab. 2, Supp. Info. Fig. 4), arguing for a complex

618 balancing selection situation, as was observed in *Cx. pipiens* (Milesi *et al.,* 2018). A definitive
619 approach to assess the primacy of selection over neutral processes in the observed D diversity
620 would require measuring the fitness of these new D alleles, either by establishing a mosquito line
621 for each of these alleles, introgressed on a unique genomic background (e.g., Labbé *et al.,* 2007,
622 Assogba *et al.,* 2015, 2016, Milesi *et al.,* 2018) or by monitoring their dynamics in the populations
623 over many years (e.g., Labbe *et al.*, 2009, Milesi *et al.*, 2016). Both require extensive long-lasting
624 effort. By revealing the existence of this unexpectedly high polymorphism of D allele in *An.*
625 *gambiae s.l., our study* represents the first step in that direction.

626 **Conclusion**

627 Altogether our findings highlight the relatively high local diversity and frequency of *ace-1*
628 heterogeneous duplications implicated in the adaptation to OP/CX insecticides in *Anopheles*
629 mosquitoes, particularly when compared to the regional uniqueness of the other resistance allele
630 (R). This diversity likely results from frequent secondary recombination events between single-
631 copy and duplicated alleles, not only restricted to *ace-1* but potentially involving a large portion of
632 the 203kb amplicon. Our study supports a role for selection in the maintenance of D allele
633 polymorphism in the populations, but further investigations are required to better assess the
634 relative roles of neutral processes and selection.

635 More generally, our study highlights the challenges that must be overcome when analysing
636 large-scale structural variants (SV). The last 15 years have seen an increase of interest in the
637 role of SVs in the adaptation process (e.g. Wellenreuther *et al.,* 2019). Gene copy-number
638 variations are a perfect example of genomic mutations that are particularly arduous to investigate
639 for technical reasons (they are mostly just more of the same sequence, all mapping together on
640 reference genomes), but also because of their often multi-allelic nature. For large-scale segmental
641 duplications, our study of *An. gambiae s.l.* clearly demonstrates that both the number of copies
642 (R^x alleles) and the specific sequence carried by the copies (D alleles) are relevant to understand
643 their evolution. It also showed that the study of copy-number variations (CNVs) is prone to

misinterpretations with rushed approaches (as both the number and the nature of the copies can affect the phenotype).

These large genomic mutations are frequent and ubiquitous (Emerson *et al.,* 2008; Itsara *et al.,* 2009; Reams *et al.,* 2010; Langley *et al.,* 2012; Katju & Bergthorsson, 2013; Schrider *et al.,* 2013, Remnant et al., 2013, Mérot *et al.*, 2020), they played a decisive role in the evolution of living organisms and are still determinant in the adaptation process, even at the micro-evolutionary scale (e.g. Kondrashov, 2012 and references therein); therefore they are worth the painstaking endeavour to study them.

Acknowledgement. We are very grateful to Sarah A. Kelly for the revision of the manuscript. The sequencing was performed at Polo d'Innovazione di Genomica, Genetica e Biologia (Italy) as part of the InfraVec 2 EU program (website: https://infravec2.eu/), and by the GenSeq platform at ISEM (Clone Sanger sequences). The computations were enabled by resources in project [NAISS 2023/23-5] provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at UPPMAX, funded by the Swedish Research Council through grant agreement no. 2022-06725.

Funding. This work was funded by the French ANR programme (project "ArchR", grant number ANR-20-CE34-0007). This study is a contribution of the Institut des Sciences de l'Evolution de Montpellier (UMR 5554, CNRS-UM-IRD-EPHE)

Author's contribution. M.W. and P.L. designed the research; B.A., C.P., A.K., P.M. and M.D-L produced the data; J-L.C., M.D-L., A.N., P.M. and P.L. analyzed data; J-L.C., P.M. and P.L. wrote the first draft; P.L., P.M and M.W. reviewed the manuscript.

Competing financial interests. The authors have no competing financial interests to declare.

References

- Alout, H., Berthomieu, A., & C. Berticat, C. (2007a). Different amino-acid substitutions confer insecticide resistance through acetylcholinesterase 1 insensitivity in *Culex vishn*ui and *Culex tritaeniorhynchus* (Diptera: Culicidae) mosquitoes from China. Journal of Medical Entomology 44:463–469. https://doi.org/10.1603/0022-2585(2007)44[463:dascir]2.0.co;2
- Alout, H., Berthomieu, A., Hadjivassilis, A., & Weill M. (2007b). A new amino-acid substitution in acetylcholinesterase 1 confers insecticide resistance to *Culex pipiens* mosquitoes from Cyprus. Insect Biochem Mol Biol 37:41–47. https://doi.org/10.1016/j.ibmb.2006.10.001
- Alout, H., Djogbénou, L., Berticat, C., Chandre, F., & Weill, M. (2008). Comparison of *Anopheles gambiae* and *Culex pipiens* acetycholinesterase 1 biochemical properties. Comparative

Biochemistry and Physiology Part B: Biochemistry and Molecular Biology, 150(3), 271-277. https://doi.org/10.1016/j.cbpb.2008.03.008

- Assogba, B. S., Djogbénou, L. S., Milesi, P., Berthomieu, A., Perez, J., Ayala, D., Chandre, F., Makoutodé, M., Labbé, P., & Weill, M. (2015). An *ace-1* gene duplication resorbs the fitness cost associated with resistance in *Anopheles gambi*ae, the main malaria mosquito. Scientific Reports, 5(1), 14529. https://doi.org/10.1038/srep14529
- Assogba, B. S., Milesi, P., Djogbénou, L. S., Berthomieu, A., Makoundou, P., Baba-Moussa, L. S., Fiston-Lavier, A.-S., Belkhir, K., Labbé, P., & Weill, M. (2016). The *ace-1* locus is amplified in all resistant *Anopheles gambiae* mosquitoes : fitness consequences of homogeneous and heterogeneous duplications. PLoS Biology, 14(12), e2000618. https://doi.org/10.1371/journal.pbio.2000618
- Assogba, B. S., Alout, H., Koffi, A., Penetier, C., Djogbénou, L. S., Makoundou, P., Weill, M., & Labbé, P. (2018). Adaptive deletion in resistance gene duplications in the malaria vector *Anopheles gambiae*. Evolutionary Applications, 11(8), 1245-1256. https://doi.org/10.1111/eva.12619
- Bourguet, D., Roig, A., Toutant, J.-P., & Arpagaus, M. (1997). Aanalysis of molecular forms and pharmacological properties of acetylcholinesterase in several mosquito species. Neurochemistry International, 31(1), 65-72. https://doi.org/10.1016/S0197-0186(96)00118-0
- Collins, F. H., Mendez, M. A., Rasmussen, M. O., Mehaffey, P. C., Besansky, N. J., & Finnerty, V. (1987). A ribosomal RNA gene probe differentiates member species of the *Anopheles gambiae* complex. *The American journal of tropical medicine and hygiene*, *37*(1), 37-41. https://doi.org/10.4269/ajtmh.1987.37.37
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... & Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience, 10(2), giab008. https://doi.org/10.1093/gigascience/giab008
- David, J.-P., Coissac, E., Melodelima, C., Poupardin, R., Riaz, M. A., Chandor-Proust, A., & Reynaud, S. (2010). Transcriptome response to pollutants and insecticides in the dengue vector Aedes aegypti using next-generation sequencing technology. BMC Genomics, https://doi.org/10.1186/1471-2164-11-216

- De Coster, W., Weissensteiner, M. H., & Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. Nature Reviews Genetics, 22(9), 572-587. https://doi.org/10.1038/s41576-021-00367-3
- Devonshire, A. and Sawicki, R. (1979) Insecticide-resistant *Myzus persicae* as an example of evolution by gene duplication. Nature 280, 140–141. https://doi.org/10.1038/280140a0
- Djogbénou, L., F. Chandre, A. Berthomieu, R. K. Dabiré, A. Koffi, H. Alout, & M. Weill. (2008) Evidence of introgression of the *ace-1*^{R} mutation and of the *ace-1* duplication in west African *Anopheles gambiae s. s.* PLoS ONE, 3:e2172, 1–7. https://doi.org/10.1371%2Fjournal.pone.0002172
- Djogbénou, L.S., Assogba, B., Essandoh, J., Constant, E. A. V., Makoutodé, M., Akogbéto, M., Donnelly, M. J., & Weetman, D. (2015). Estimation of allele-specific *Ace-1* duplication in insecticide-resistant *Anopheles* mosquitoes from West Africa. Malaria Journal, 14, 507. https://doi.org/10.1186/s12936-015-1026-3
- Elissa, N., Mouchet, J., Rivière, F., Meunier, J.Y. & Yao, K. (1994) Sensibilité d'*Anopheles gambiae* aux insecticides en Côte d'Ivoire. Cahiers Santé, 4, 95–99.
- Emerson, J. J., Cardoso-Moreira, M., Borevitz, J. O., & Long, M. (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. Science, 320(5883), 1629-1631. https://doi.org/10.1126/science.1158078
- Favia, G., Della Torre, A., Bagayoko, M., Lanfrancotti, A., Sagnon, N., Touré, Y. T., & Coluzzi, M. (1997). Molecular identification of sympatric chromosomal forms of *Anopheles gambiae* and further evidence of their reproductive isolation. Insect Molecular Biology, 6(4), 377-383. https://doi.org/10.1046/j.1365-2583.1997.00189.x
- Grau-Bové, X., Lucas, E., Pipini, D., Rippon, E., van 't Hof, A. E., Constant, E., Dadzie, S., Egyir-Yawson, A., Essandoh, J., Chabi, J., Djogbénou, L., Harding, N. J., Miles, A., Kwiatkowski, D., Donnelly, M. J., Weetman, D., & The *Anopheles gambiae* 1000 Genomes Consortium. (2021). Resistance to pirimiphos-methyl in West African Anopheles is spreading via duplication and introgression of the *Ace1* locus. PLoS Genetics, 17(1), e1009253. https://doi.org/10.1371/journal.pgen.1009253

- Guillemaud, T., Lenormand, T., Bourguet, D., Chevillon, C., Pasteur, N., & Raymond, M. (1998). Evolution of resistance in *Culex pipiens*: allele replacement and changing environment. Evolution, 52(2), 443-453. https://doi.org/10.1111/j.1558-5646.1998.tb01644.x
- Hendry, A. P., Gotanda, K. M., & Svensson, E. I. (2017). Human influences on evolution, and the ecological and societal consequences. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1712), 20160028. https://doi.org/10.1098/rstb.2016.0028
- Hook, P. W., & Timp, W. (2023). Beyond assembly: the increasing flexibility of single-molecule sequencing technology. Nature Review Genetics, 24, pages 627–641. doi: 10.1038/s41576-023-00600-1.
- Huchard, E., Martinez, M., Alout, H., Douzery, E. J. P., Lutfalla, G., Berthomieu, A., Berticat, C., Raymond, M., & Weill, M. (2006). Acetylcholinesterase genes within the Diptera : Takeover and loss in true flies. Proceedings of the Royal Society B: Biological Sciences, 273(1601), 2595-2604. https://doi.org/10.1098/rspb.2006.3621
- Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nature reviews. Genetics 11:97–108. https://doi.org/10.1038/nrg2689
- Itsara, A., Cooper, G. M., Baker, C., Girirajan, S., Li, J., Absher, D., ... & Eichler, E. E. (2009). Population analysis of large copy number variants and hotspots of human genetic disease. The American Journal of Human Genetics, 84(2), 148-161. https://doi.org/10.1016%2Fj.ajhg.2008.12.014
- Karunarathne, P., Zhou, Q., Schliep, K., & Milesi, P. (2023). A comprehensive framework for detecting copy number variants from single nucleotide polymorphism data: 'rCNV', a versatile r package for paralogue and CNV detection. Molecular Ecology Resources 23:1772–1789. https://doi.org/10.1111/1755-0998.13843
- Katju, V., & Bergthorsson, U. (2013). Copy-number changes in evolution: rates, fitness effects and adaptive significance. Frontiers in genetics, 4, 273. https://doi.org/10.3389/fgene.2013.00273
- Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. Proceedings of the Royal Society B: Biological Sciences, 279(1749), 5048-5057. https://doi.org/10.1098/rspb.2012.1108
- Kouamé, R. M. A., Lynd, A., Kouamé, J. K. I., Vavassori, L., Abo, K., Donnelly, M. J., Edi, C. & Lucas, E.. 2023. Widespread occurrence of copy number variants and fixation of pyrethroid target site

763 resistance in *Anopheles gambiae* (*s.l.*) from southern Côte d'Ivoire. Current Research in
764 Parasitology & Vector-Borne Diseases 3:100117.
765 • Kwon, D. H., J. M. Clark, S. H. Lee, J. M. Clarkt, & S. H. Lee. 2010. Extensive gene duplication of
766 acetylcholinesterase associated with organophosphate resistance in the two-spotted spider mite.
767 Insect Molecular Biology 19:195–204. https://doi.org/10.1111/j.1365-2583.2009.00958.x
768 • Labbé, P., Berthomieu, A., Berticat, C., Alout, H., Raymond, M., Lenormand, T., & Weill, M. (2007).
769 Independent duplications of the acetylcholinesterase gene conferring insecticide resistance in the
770 mosquito *Culex pipiens*. Molecular Biology and Evolution, 24, 1056–1067.
771 https://doi.org/10.1093/molbev/msm025
772 • Labbé, P., Sidos, N., Raymond, M., & Lenormand, T. (2009). Resistance gene replacement in the
773 mosquito *Culex pipiens* : fitness estimation from long-term cline series. Genetics, 182(1), 303-312.
774 https://doi.org/10.1534/genetics.109.101444
775 • Labbé, P., Milesi, P., Yébakima, A., Pasteur, N., Weill, M., & Lenormand, T. (2014). Gene-dosage
776 effects on fitness in recent adaptive duplications : *ace-1* in the mosquito *Culex pipiens*. Evolution,
777 68(7), 2092-2101. https://doi.org/10.1111/evo.12372
778 • Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C. G., Schrider, D. R., Pool, J. E., ... & Begun, D. J.
779 (2012). Genomic variation in natural populations of Drosophila melanogaster. Genetics, 192(2),
780 533-598. https://doi.org/10.1534%2Fgenetics.112.142018
781 • Leister, D. (2004) Tandem and segmental gene duplication and recombination in the evolution of plant
782 disease resistance genes. Trends in genetics, 3, 116-122. https://doi.org/10.1016/j.tig.2004.01.007
783 • Lenormand, T., N. Harmand, and R. Gallet. 2018. Cost of resistance: an unreasonably expensive
784 concept. Rethinking Ecology 3:51-70 https://doi.org/10.3897/rethinkingecology.3.31992.
785 • Lenormand, T., Guillemaud, T., Bourguet, D. , & Raymond, M. (1998). Appearance and sweep of a
786 gene duplication: adaptive response and potential for new functions in the mosquito *Culex pipiens*.
787 Evolution 52:1705. https://doi.org/10.1111/j.1558-5646.1998.tb02250.x
788 • Lenormand, T., D. Bourguet, T. Guillemaud, and M. Raymond. (1999). Tracking the evolution of
789 insecticide resistance in the mosquito *Culex pipiens*. Nature 400, 861-864.
790 https://doi.org/10.1038/23685

- Li H, Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. Bioinformatics, 25(14), 1754-60. https://doi.org/10.1093/bioinformatics/btp324
- Li, H. (2018). Minimap2 : Pairwise alignment for nucleotide sequences. Bioinformatics, 34(18), 3094-3100. https://doi.org/10.1093/bioinformatics/bty191
- Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-read sequencing emerging in medical genetics. Frontiers in genetics, 10, 426. https://doi.org/10.3389/fgene.2019.00426
- Martin, M., Ebert, P., Marschall, T. (2023). Read-Based Phasing and Analysis of Phased Variants with WhatsHap. In: Peters, B.A., Drmanac, R. (eds) Haplotyping. Methods in Molecular Biology, vol 2590. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-2819-5_8
- Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. Trends in Ecology & Evolution, 35(7), 561-572. https://doi.org/10.1016/j.tree.2020.03.002
- Milesi, P., Lenormand, T., Lagneau, C., Weill, M., & Labbé, P. (2016). Relating fitness to long-term environmental variations *in natura*. Molecular Ecology, 25(21), 5483-5499. https://doi.org/10.1111/mec.13855
- Milesi, P., Weill, M., Lenormand, T., & Labbé, P. (2017). Heterogeneous gene duplications can be adaptive because they permanently associate overdominant alleles. Evolution Letters, 1(3), 169-180. https://doi.org/10.1002/evl3.17
- Milesi, P., Assogba, B. S., Atyame, C. M., Pocquet, N., Berthomieu, A., Unal, S., Makoundou, P., Weill, M., & Labbé, P. (2018). The evolutionary fate of heterogeneous gene duplications : A precarious overdominant equilibrium between environment, sublethality and complementation. Molecular Ecology, 27(2), 493-507. https://doi.org/10.1111/mec.14463
- Milesi, P., Claret, J.-L., Unal, S., Weill, M., & Labbé, P. (2022). Evolutionary trade-offs associated with copy number variations in resistance alleles in *Culex pipiens* mosquitoes. Parasites Vectors 15:484. https://doi.org/10.1186/s13071-022-05599-8
- Namias, A., Sahlin, K., Makoundou, P., Bonnici, I., Sicard, M., Belkhir, K., & Weill, M. (2023). Nanopore sequencing enables multigenic family reconstruction. Computational and Structural Biotechnology Journal 21:3656–3664. https://doi.org/10.1016/j.csbj.2023.07.012

- N'Guessan, R., Darriet, F., Guillet, P., Carnevale, P., Traore-Lamizana, M., Corbel, V., Koffi, A. A., & Chandre, F. (2003). Resistance to carbosulfan in *Anopheles gambiae* from Ivory Coast, based on reduced sensitivity of acetylcholinesterase. Medical and Veterinary Entomology, 17(1), 19-25. https://doi.org/10.1046/j.1365-2915.2003.00406.x
- Otto, S. P. (2018). Adaptation, speciation and extinction in the Anthropocene. Proceedings of the Royal Society B, 285, 20182047. https://doi.org/10.1098/rspb.2018.2047
- Patterson, E. L., Pettinga D. J., Ravet K., Neve P., Gaines T. A. (2018) Glyphosate resistance and EPSPS gene duplication: convergent evolution in multiple plant species, Journal of Heredity, 109, 117–125, https://doi.org/10.1093/jhered/esx087
- Reams, A. B., Kofoid, E., Savageau, M., & Roth, J. R. (2010). Duplication frequency in a population of *Salmonella enterica* rapidly approaches steady state with or without recombination. Genetics, 184(4), 1077-1094. https://doi.org/10.1534/genetics.109.111963
- Remnant, E. J., Good, R. T., Schmidt, J. M., Lumb, C., Robin, C., Daborn, P. J., & Batterham, P. (2013). Gene duplication in the major insecticide target site, *Rdl*, in *Drosophila melanogaster*. Proceedings of the National Academy of Sciences of the United States of America 110:14705–10. https://doi.org/10.1073/pnas.1311341110
- Schrider, D. R., & Hahn, M. W. (2010). Gene copy-number polymorphism in nature. Proceedings of the Royal Society B: Biological Sciences, 277(1698), 3213-3221. https://doi.org/10.1098/rspb.2010.1180
- Schrider, D. R., Houle, D., Lynch, M., & Hahn, M. W. (2013). Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics, 194(4), 937-954. https://doi.org/10.1534/genetics.113.151670
- Schumacher M., Camp S., Maulet Y., Newton M., MacPhee-Quigley K., Taylor S. S., Friedmann T. & Taylor P. (1986) Primary structure of *Torpedo californica* acetylcholinesterase deduced from its cDNA sequence. Nature 319, 407-409.
- Scott, J. A., Brogdon, W. G., & Collins, F. H. (1993). Identification of single specimens of the *Anopheles gambiae* complex by the polymerase chain reaction. The American journal of tropical medicine and hygiene, 49(4), 520-529. https://doi.org/10.4269/ajtmh.1993.49.520

- Tamura, K., Stecher, G., & Sudhir Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11, Molecular Biology and Evolution, 38, 3022–3027, https://doi.org/10.1093/molbev/msab120
- Weetman, D., Mitchell, S. N., Wilding, C. S., Birks, D. P., Yawson, A. E., Essandoh, J., Mawejje, H. D., Djogbenou, L. S., Steen, K., Rippon, E. J., Clarkson, C. S., Field, S. G., Rigden, D. J., & Donnelly, M. J. (2015). Contemporary evolution of resistance at the major insecticide target site gene *Ace-1* by mutation and copy number variation in the malaria mosquito *Anopheles gambiae*. Molecular Ecology, 24(11), 2656-2672. https://doi.org/10.1111/mec.13197
- Weill, M., Lutfalla, G., Mogensen, K., Chandre, F., Berthomieu, A., Berticat, C., Pasteur. N. Philips, A., Fort. P. & Raymond, M. (2003). Insecticide resistance in mosquito vectors. Nature, 423(6936), 136-137. https://doi.org/10.1038/423136b
- Weill, M., C. Malcolm, F. Chandre, K. Mogensen, A. Berthomieu, M. Marquine, & M. Raymond. 2004. The unique mutation in *ace-1* giving high insecticide resistance is easily detectable in mosquito vectors. Insect molecular biology 13:1–7. https://doi.org/10.1111/j.1365-2583.2004.00452.x
- Wellenreuther, M., Mérot, C., Berdan, E., & Bernatchez, L. (2019). Going beyond SNPs : The role of structural genomic variants in adaptive evolution and species diversification. Molecular Ecology, 28(6), 1203-1209. https://doi.org/10.1111/mec.15066
- Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics. Version, 2(1), 1-189.
- Xu, S., Li, L., Luo, X., Chen, M., Tang, W., Zhan, L., ... & Yu, G. (2022). Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. iMeta, 1(4), e56.

FIGURES

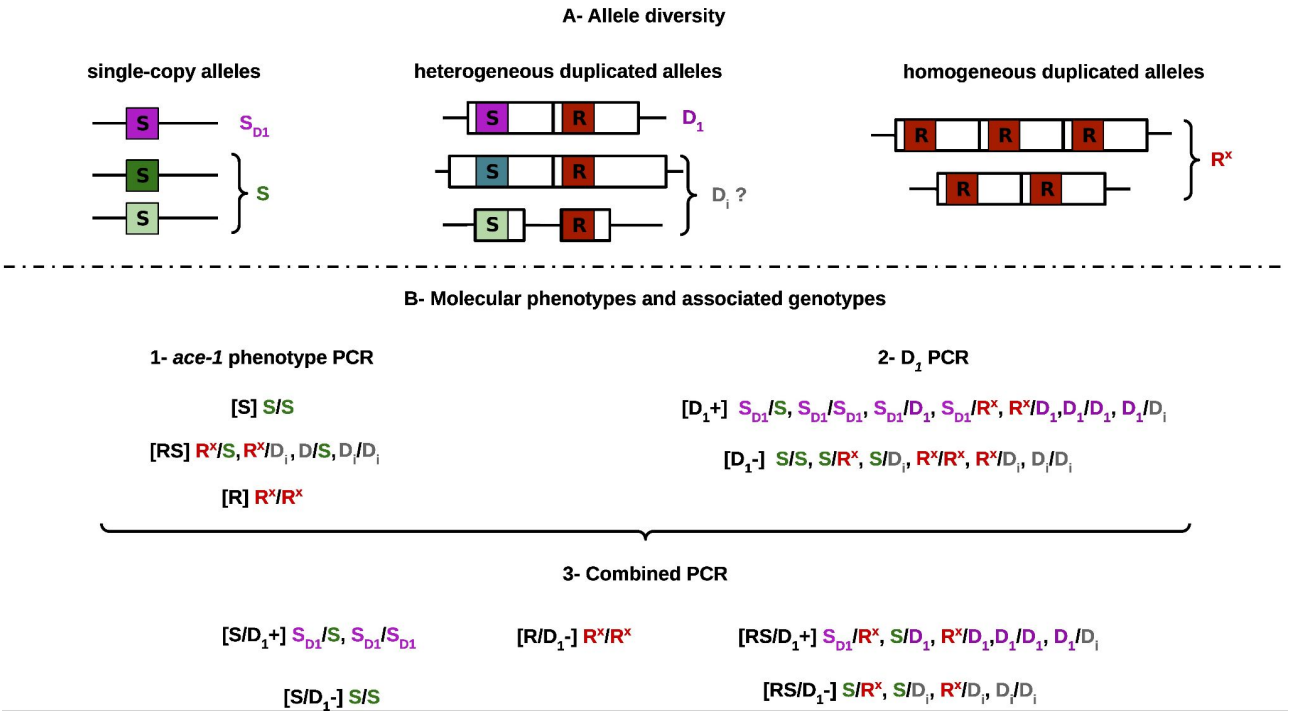


Figure 1. Diversity of *ace-1* alleles, and molecular phenotyping. (Modified from Assogba *et al.,* 2018).

(A) The various alleles found at the *ace-1* locus: different single-copy S alleles[a] on the left (green) and homogeneous duplicated alleles R^x on the right (here with 2 or 3 R copies, R^3 and R^2 resp., in red). The central part illustrates the known D_1 heterogenous allele, with its D(S) copy (in pink) and its D(R) copy (in red), as well other heterogeneous D_i alleles (with different architectures depending on the size of the amplified region). NB: the single-copy S_{D1} allele has the same sequence as the D(S) copy of *ace-1* (hence the same color). (B) The two PCR used to identify the genotypes of triple peaks samples. The combined information of the « *ace-1* phenotype » PCR (1) and the «D_1 PCR» (2) allow the partial discrimination of 5 phenotypes with 13 possible genotypes (3).

[a] multi-copy S alleles or S^x could also exist, but they would not be distinguished here from single-copy alleles.

Locality	Year	Resistance phenotype			
		[RR]	[SS] ([D_1+])	[RS] ([D_1+])	N
Yopougon	2012	0	40 (1)	**15 (3)**	55
	2015	1	13 (0)	**45 (25)**	59
	2016	1	20 (2)	**37 (18)**	58
Yamoussoukro	2012	0	21 (1)	**25 (14)**	46
	2015	1	12 (2)	**35 (15)**	48
	2016	0	9 (1)	**30 (15)**	39

Table 1: *ace-1* phenotype diversity in Yamoussoukro and Yopougon. Samples (identified by year and locality) originate from Assogba et al.'s (2018) study. [RR], [RS] and [SS] phenotypes were obtained through the "*ace-1* phenotype" PCR (see Material and Methods, Fig. 1B). The number of [D_1+] individuals are indicated in brackets ("D_1" PCR, Fig. 1B).

Model A- 3 alleles model

Locality	Year	*R*	*S*	*D*	LRT (χ^2)	*p-value*
Yopougon	**2012**	0.00 [0.00:0.18]	0.85	0.15 [0.00:0.22]	2.38	0.12^{NS}
	2015	0.13 [0.03:0.27]	0.47	0.40 [0.23:0.55]	24.42	$7.73x10^{-7***}$
	2016	0.13 [0.03:0.27]	0.61	0.27 [0.11:0.41]	11.94	$5.50x10^{-4***}$
Yamoussoukro	**2012**	0.00 [0.00:0.20]	0.68	0.32 [0.10:0.43]	9.27	$2.33x10^{-3**}$
	2015	0.14 [0.03:0.29]	0.53	0.33 [0.16:0.49]	14.84	$1.17x10^{-4***}$
	2016	0.00 [0.00:0.22]	0.50	0.50 [0.26:0.63]	19.27	$1.13x10^{-4***}$

Model B- 4 alleles model

Locality	Year	*R*	*S*	S_{D1}	D_1
Yopougon	**2012**	0.11 [0.06:0.18]	0.85	0.01 [0.00:0.05]	0.03 [0.01:0.07]
	2015	0.25 [0.16:0.34]	0.51	0.00 [0.00:0.03]	0.24 [0.16:0.33]
	2016	0.21 [0.14:0.30]	0.60	0.03 [0.00:0.08]	0.16 [0.09:0.24]
Yamoussoukro	**2012**	0.14 [0.08:0.23]	0.68	0.02 [0.00:0.07]	0.16 [0.09:0.25]
	2015	0.27 [0.18:0.38]	0.54	0.04 [0.01:0.11]	0.15 [0.08:0.24]
	2016	0.24 [0.15:0.36]	0.53	0.02 [0.00:0.10]	0.20 [0.11:0.31]

Model C- 5 alleles model

Locality	Year	*R*	*S*	S_{D1}	D_1	D_i	LRT (χ^2)	*p-value*
Yopougon	**2012**	0.00 [0.00:0.17]	0.84	0.01 [0.00:0.05]	0.03 [0.00:0.07]	0.12 [0.00:0.20]	1.60	0.21^{NS}
	2015	0.13 [0.03:0.27]	0.47	0.00 [0.00:0.04]	0.24 [0.16:0.33]	0.16 [0.01:0.31]	4.61	0.03*
	2016	0.13 [0.03:0.27]	0.58	0.03 [0.00:0.09]	0.16 [0.09:0.24]	0.11 [0.00:0.25]	2.37	0.12^{NS}
Yamoussoukro	**2012**	0.00 [0.00:0.20]	0.67	0.02 [0.00:0.07]	0.16 [0.09:0.25]	0.16 [0.00:0.26]	2.30	0.13^{NS}
	2015	0.14 [0.03:0.29]	0.49	0.04 [0.01:0.12]	0.15 [0.08:0.24]	0.18 [0.02:0.35]	4.99	0.03*
	2016	0.00 [0.00:0.22]	0.47	0.03 [0.00:0.11]	0.20 [0.11:0.31]	0.30 [0.07:0.44]	7.07	0.01**

Table 2: Estimated allele frequencies under different models.

For each locality and sampling year, the frequencies of the different alleles (along with their support limits, *i.e.* roughly equivalent to 95% confidence intervals, brackets) have been estimated using a maximum likelihood approach under different assumptions (see text).

Model A - The first model considers only the phenotypes resulting from the "*ace-1* phenotype" test, thus 3 alleles, R, S and D, for 3 phenotypes and 6 genotypes (Fig. 1B). LRT corresponds to the likelihood ratio test between this model and another one without any D allele (chi-square distribution with 1 degree of freedom). Model B **-** The second model considers the phenotypes resulting from the combination of the "D_1" and "resistance phenotype" tests, *i.e.* adding the specific information on D_1 frequency. Four alleles are thus

considered, R, S, S_{D1} (as some [SS] are also [D_1+]) and D_1, for 5 phenotypes and 9 genotypes (Fig. 1B-3, without genotypes including the D_i allele). Model C- The third model analyses the same data, but considers 5 alleles, R, S, S_{D1}, D_1 and D_i (*i.e.* at least another D allele), for 5 phenotypes and 13 genotypes (Fig. 1B-3, all genotypes). The *p*-value of the LRT comparing models B and C[a] (chi-square distribution with 1 degree of freedom) is indicated for each population (if significant, model C better fits the data than model B).

[a]models A and B cannot be compared using LRT as they are not fitted to the same dataset (3 *vs* 5 phenotypes).

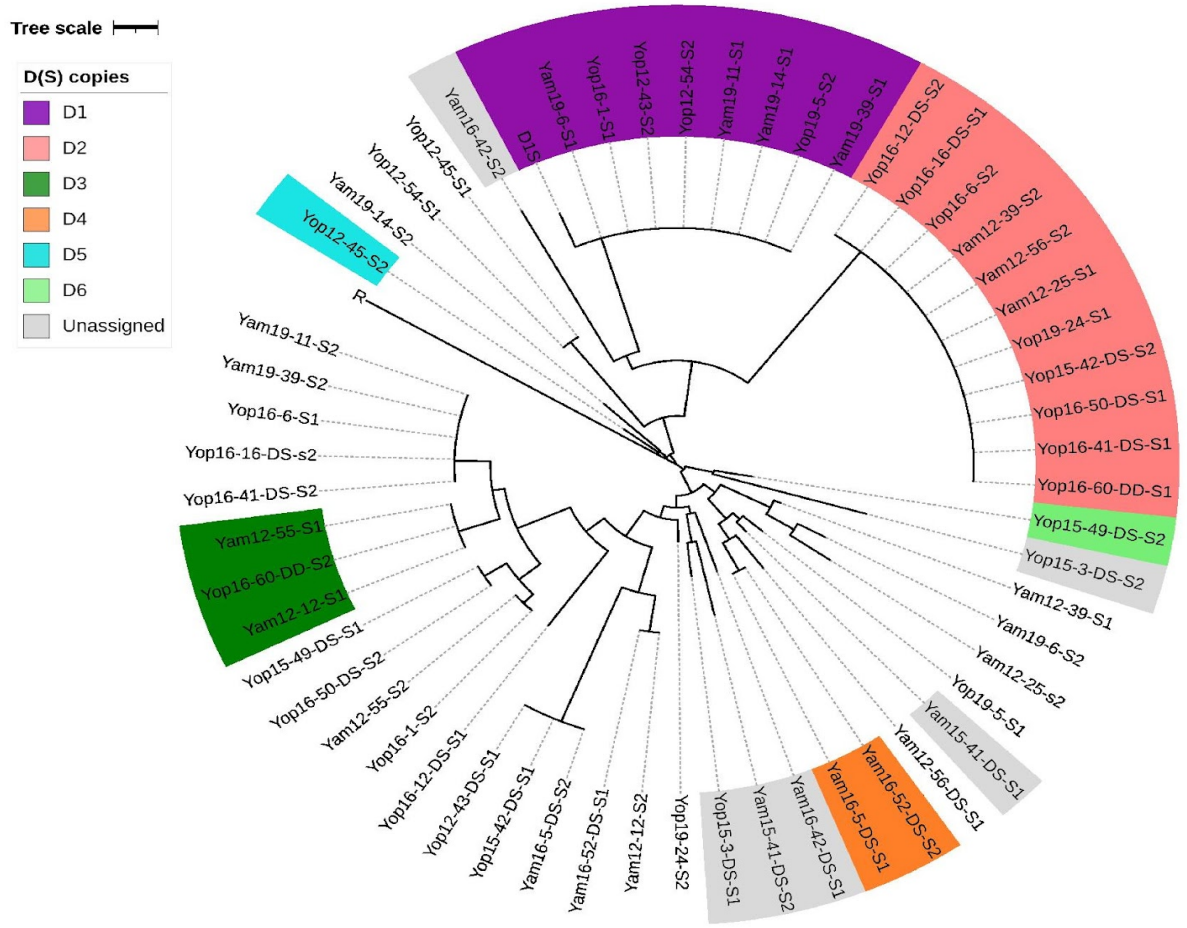


Figure 2 : Diversity of the *ace-1* S and D(S) sequences in individuals displaying triple peaks.

This phylogram represents the diversity of the S copies retrieved (TA cloning/Sanger sequencing and/or Nanopore sequencing, see Supp. Info. Tab. 2) from individuals displaying triple peaks in the mix sequence of the "*ace-1* resistance phenotype" PCR product (see Materials). Samples are coded as follows: locality (Yam for Yamoussoukro and Yop for Yopougon)/sampling year and individual number (-x). The 12 samples whose genotype was obtained through short-read sequencing are further coded with DS (duplicated heterozygote) or DD (duplicated homozygote). For each individual, the two S copies are indicated as copy S1 and copy S2 (assigned randomly). Sequences identified as single-copy S alleles are not highlighted. Copies identified as probable D(S) copies (see text) are highlighted according to the corresponding putative D allele (legend). Unassigned sequences, *i.e.* S sequences that are found in an individual carrying a D allele, but that cannot yet be assigned to S or D(S) for lack of data, are highlighted in grey.

NB: the haplotypes assigned to $D_1(S)$, including the D_1 controls, differed by one mutation from the canonical sequence (GenBank: KM875635.1).

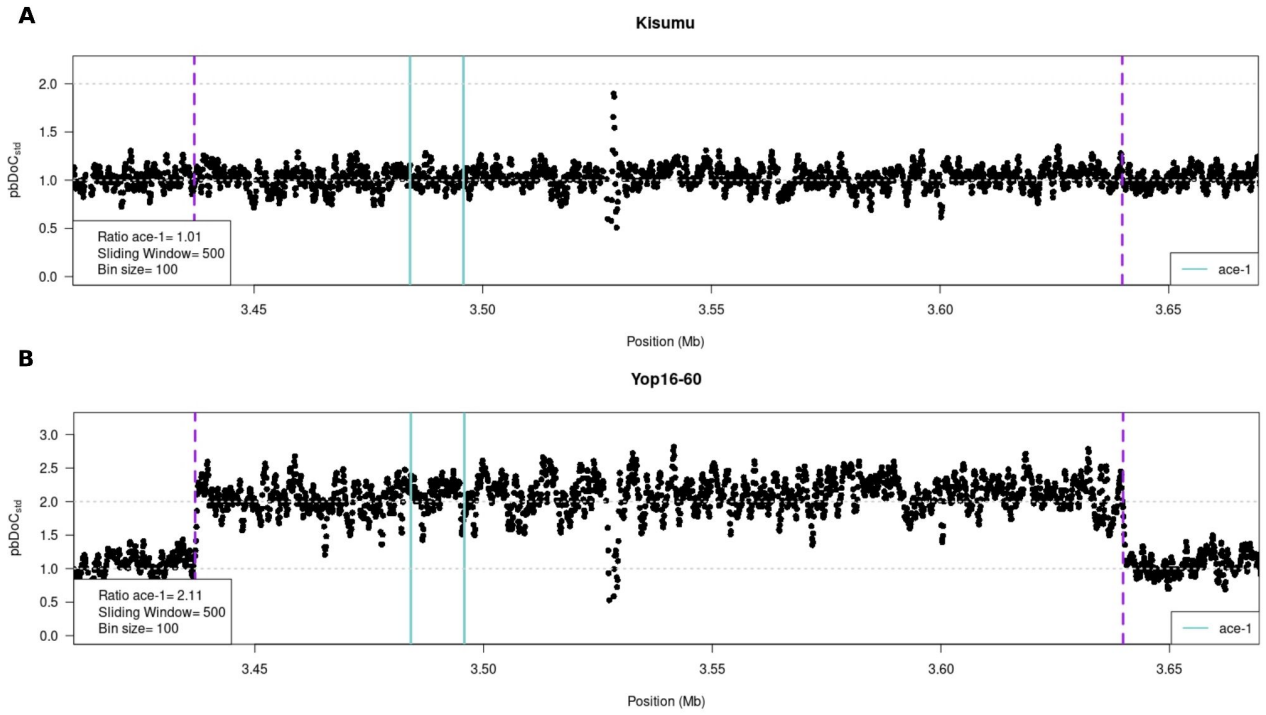


Figure 3 : Genomic architecture of the *ace-1* D alleles.

In each graph, we presented the variation of the standardized per-base depth of coverage (*pb*DoC$_{std}$, with 1 being the mean *pb*DoC calculated over the whole chromosome) along the region of interest, from 3.4 to 3.7 MB of chromosome 2R. Each dot is the mean *pb*DoC$_{std}$ calculated every 100 bases (bin size) over 500-base sliding windows. The purple dashed lines represent the amplicon limits of the D_1 and R^x alleles (Assogba et al. 2018); the cyan lines represent the *ace-1* gene location.

A- The susceptible strain Kisumu is the single-copy S allele reference, with no particular variation of *pb*DoC$_{std}$ (mean = 1). B- The second graph represents the *pb*DoC$_{std}$ variation for the individual Yop16-60 (as a representative example of the D alleles analyzed in the present study; similar graphs for the other individuals analyzed are shown in Supp. Info. Fig. 1). A 2-fold increase reveals the amplicon size and location, similar to D_1; it is consistent with a D_i/D_j genotype (two S and two R copies). All D_i alleles share the same breakpoints as D_1; however, the other individuals display only a 1.5 increase, as expected for genotypes DS (two S and 1R copies; Supp. Info. Fig. 1).

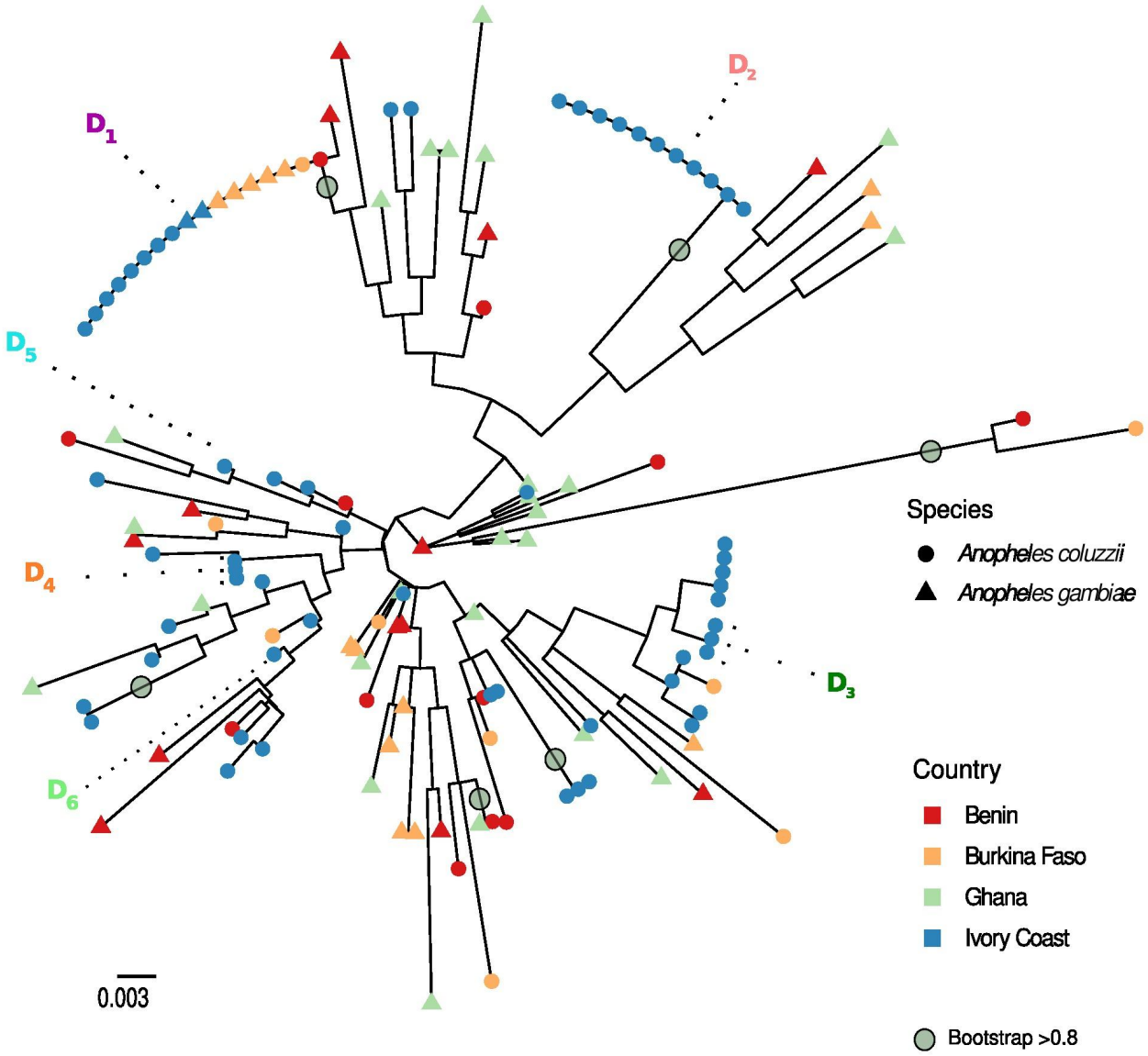


Figure 4. *ace-1* S diversity in West Africa. Maximum likelihood tree of *A. coluzzii* and *A. gambiae* S sequences from West Africa. For each sequence, species (triangle for *A. gambiae* and circle for *A. coluzzii)*, and geographical origin (Benin in red, Burkina Faso in orange, Ghana in green, and Ivory Coast in blue, as in the inlet map) are indicated. Accession numbers can be found in Supp. Info. Tab. 3. The D(S) copies identified in the present study are also indicated, as well as the bootstraps confidence ≥0.8 with grey circles.

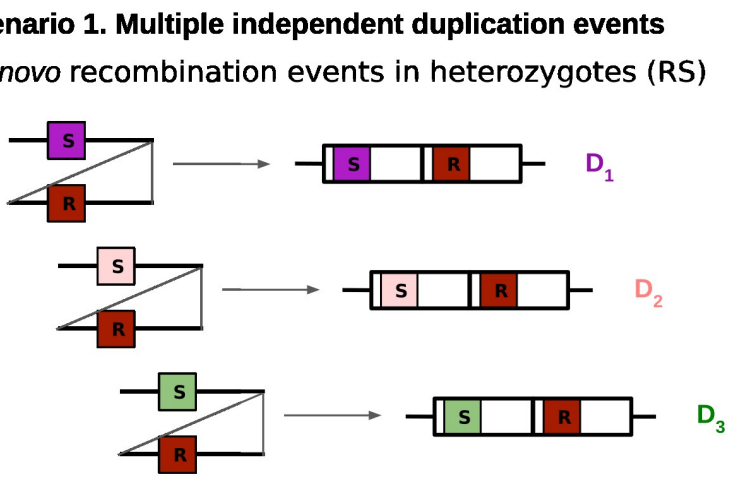


Figure 5. Possible scenarios for the origin of Ivory Coast *ace-1* duplications.

Scenario 1 requires several independent duplication events on the same breakpoints, whereas scenario 2 considers a first duplication event followed by secondary recombinations occurring in the amplicon that bears the S copy (either the whole amplicon, or only part of the the *ace-1* locus, bottom, or any size between). *NB: the alleles presented here are for illustration only, as the present study did not allow firmly distinguishing the two scenarios, or any secondary recombination span. Similarly, the oblique lines are used to illustrate the recombination events, but not a particular molecular mechanism.*

Supporting information Table 1. Molecular tests.

Various PCR were used for identifying the species (Species) and the molecular form (Form) of the analysed individuals, as well as their *ace-1* phenotype, whether they carry the D_1 allele (D_1), and whether they carry a duplicated allele sharing the same junction between amplicons as D_1 and R^x alleles (Junction). The primers used and the reaction characteristics are indicated.

PCR	Primers	Annealing temperature
Species	**-** ***Anopheles gambiae*** **AG** 5' - CTG GTT TGG TCG GCA CGT TT - 3' **-** ***Anopheles arabiensis*** **AA** - 5' - AAG TGT CCT TCT CCA TCC TA - 3' **-** ***Anopheles melas*** **AM** - 5' - AAG TGT CCT TCT CCA TCC TA - 3' **- Universal UN** - 5' - CTG TGC CCC TTC CTC GAT GT - 3'	56°C
Form	**- 6.1a** 5' - TCGCCTTAGACCTTGCGTTA - 3'. **- 6.1b** 5 - CGCTTCAAGAATTCGAGATAC - 3'	52°C
ace-1 **phenotype**	**- AgEx3univdir** 5' - GAT CGT GGA CAC CGT GTT CG - 3' **- AgEx3univrev** 5' - AGG ATG GCC CGC TGG AAC AG - 3'	57°C
D1	**- AgEx3univdir** 5' - GAT CGT GGA CAC CGT GTT CG - 3' **- AgEx4rev** 5' - TCG CTG CAT CTG CTG TCC GCC CT -3'	57°C
Junction	**- Agduplispedir2** 5' - CTC TTA AGG TGG CGT TGT TCC - 3' **- Agduplisperev1** 5' - TTC CGC ACA AAA GGT TGG GCA - 3'	60°C

Supporting information Table 2. Summary of the different protocols applied to each sample of our study.

For each individual, we summarised whether its *ace-1* haplotypes were determined through TA-cloning/Sanger sequencing and/or long-read sequencing (Nanopore). We also indicated those whose whole genome was sequenced (Illumina) (see Materials).

Sample	TA-cloning/ Sanger	Nanopore	Whole genome Illumina
Yam12_12	Yes	No	No
Yam12_25	Yes	No	No
Yam12_39	Yes	No	No
Yam12_55	Yes	No	No
Yam12_56	Yes	No	No
Yam15_41	No	Yes	Yes
Yam16_42	Yes	Yes	Yes
Yam16_52	No	Yes	Yes
Yam16_5	Yes	Yes	Yes
Yam19_11	Yes	No	No
Yam19_14	Yes	No	No
Yam19_39	Yes	No	No
Yam19_6	Yes	No	No
Yop12_43	Yes	No	No
Yop12_45	Yes	No	No
Yop12_54	Yes	No	No
Yop15_3	No	Yes	Yes
Yop15_42	Yes	Yes	Yes
Yop15_49	No	Yes	No
Yop16_12	Yes	Yes	Yes
Yop16_16	Yes	No	Yes
Yop16_1	Yes	Yes	No
Yop16_41	No	Yes	Yes
Yop16_50	No	Yes	Yes
Yop16_60	No	Yes	Yes
Yop16_6	Yes	No	Yes
Yop19_24	Yes	No	No
Yop19_5	Yes	No	No

Supporting Information Table 3. Sequences dataset from previously published studies.
The different sequences used in this study are indicated with their SRA accession number when available.

name	SRA_accession	name	SRA_accession
Yam12_12	PRJNA971118	GhaaF4b_wt	KP165384.1
Yam12_25	PRJNA971118	GhaaF4a_wt	KP165383.1
Yam12_39	PRJNA971118	GhaaF1b_wt	KP165382.1
Yam12_55	PRJNA971118	GhaaF1a_wt	KP165381.1
Yam12_56	PRJNA971118	GhaaE1b_wt	KP165380.1
Yam15_41	PRJNA971118	GhaaE1a_wt	KP165379.1
Yam16_42	PRJNA971118	GhaaD1b_wt	KP165378.1
Yam16_52	PRJNA971118	GhaaD1a_wt	KP165377.1
Yam16_5	PRJNA971118	GhaaC1b_wt	KP165376.1
Yam19_11	PRJNA971118	GhaaC1a_wt	KP165375.1
Yam19_14	PRJNA971118	GhaaB2b_wt	KP165374.1
Yam19_39	PRJNA971118	GhaaB2a_wt	KP165373.1
Yam19_6	PRJNA971118	GhaaM_F2b	KP165362.1
Yop12_43	PRJNA971118	GhaaM_F2a	KP165361.1
Yop12_45	PRJNA971118	GhanaD3b_wt	KP165342.1
Yop12_54	PRJNA971118	GhanaD3a_wt	KP165341.1
Yop15_3	PRJNA971118	GhanaC3b_wt	KP165340.1
Yop15_42	PRJNA971118	GhanaC3a_wt	KP165339.1
Yop15_49	PRJNA971118	GhanaB3b_wt	KP165338.1
Yop16_12	PRJNA971118	GhanaB3a_wt	KP165337.1
Yop16_16	PRJNA971118	GhanaA3b_wt	KP165336.1
Yop16_1	PRJNA971118	GhanaA3c_wt	KP165335.1
Yop16_41	PRJNA971118	GhanaA3a_wt	KP165334.1
Yop16_50	PRJNA971118	GhanaA1b_wt	KP165333.1
Yop16_60	PRJNA971118	GhanaA1a_wt	KP165332.1
Yop16_6	PRJNA971118		
Yop19_24	PRJNA971118		
Yop19_5	PRJNA971118		

Supporting information Table 4. Probable genotypes of D-carrying individuals.
The probable genotype of the 28 individuals identified as carrying at least one D allele, *i.e.* the triple-peak individuals, has been inferred from the phylogram (Fig. 2) and the genomic analyses, as described in Results. Each individual carried two S copies (S1 and S2, randomly assigned), one being D_i(S) and the other one being either a single-copy S allele, or another D_j(S) copy ("?" is used when undetermined). Each D allele is coloured as in Fig. 2. The different single-copy S alleles are identified as S_i, where *i* (*i*= A to Q) indicates the haplotype. Whole-genome-Illumina-sequenced individuals are indicated with a "*".

Individual	Copy S1	Copy S2
Yam12-12	D_3(S)	S_A
Yam12-25	D_2(S)	S_B
Yam12-39	S_C	D_2(S)
Yam12-55	D_3(S)	S_D
Yam12-56	S_E	D_2(S)
Yam15-41*	D_7(S)? (S_F)	D_7(S)? (S_F)
Yam16-42*	D_8(S)? (S_G)	D_8(S)? (S_G)
Yam16-5*	D_4(S)	S_H
Yam16-52*	S_A	D_4(S)
Yam19-11	D_1(S)	S_I
Yam19-14	D_1(S)	S_J
Yam19-39	D_1(S)	S_I
Yam19-6	D_1(S)	S_K
Yop12-43	S_H	D_1(S)
Yop12-45	S_L	D_5(S)
Yop12-54	S_L	D_1(S)
Yop15-3*	D_9(S)? (S_M)	D_9(S)? (S_M)
Yop15-42*	D_2(S)	S_H
Yop15-49	S_N	D_6(S)
Yop16-1	D_1(S)	S_D
Yop16-12*	S_O	D_2(S)
Yop16-16*	D_2(S)	S_I
Yop16-41*	D_2(S)	S_I
Yop16-50*	D_2(S)	S_N
Yop16-6*	S_I	D_2(S)
Yop16-60*	D_2(S)	D_3(S)
Yop19-5	S_P	D_1(S)
Yop19-24	D_2(S)	S_Q

Supporting information Table 5. Inferring total, R and S copy numbers from genomic data for triple-pic individuals.

For each individual, the depth of coverage (DOC) was analysed first for the ratio between *ace-1* mean DOC and the mean DOC over the whole chromosome 2R (mean *ace-1*/mean chrom), and the ratio between *ace-1* mean DOC and the mean DOC of the single-copy reference gene *ace-2* (mean *ace-1/*mean *ace-2*) were computed. The expected ratios depend on the total number of *ace-1* copies: 1, 1.5, 2 and 2.5 respectively for 2, 3, 4 and 5 copies; the deduced number of *ace-1* copies is thus indicated.

We then analysed the number of reads *N* for R and S haplotypes at the position diagnostic (one base only) for resistance (S haplotypes carry a G, R haplotypes carry a A) found in the Illumina whole-genome sequencing, giving the frequency of the R haplotypes among all the reads covering this position (%R). This frequency is expected to be 0.5, 0.33 and 0.25 for frequencies corresponding respectively to 1R:1S (RS or DD), 1R:2S (DS) and 1R:3S (e.g. D alleles with multiple S copies).

The copy number for each haplotype (R or S) was deduced considering both the *ace-1* total copy number and the frequency of R haplotypes. When considering only the diagnostic base, some ratios are not compatible with the total copy number (bolded; NB: these individuals actually carry DS genotypes, *i.e*. 1R:2S; see text).

	ace-1			**diagnostic base**			
individual	**mean *ace-1*/ mean chrom**	**mean *ace-1*/ mean *ace-2***	**deduced *ace-1* total copy number**	***N* R reads**	***N* S reads**	**%R**	**deduced number of haplotypes**
Yam15-41	1.58	1.49	3	15	29	0.34	1R:2S
Yam16-42	1.56	1.38	3	13	26	0.33	1R:2S
Yam16-5	1.71	1.47	3	14	25	0.36	1R:2S
Yam16-52	1.54	1.47	3	12	36	**0.25**	not compatible
Yop15-3	1.47	1.46	3	14	26	0.35	1R:2S
Yop15-42	1.55	1.41	3	10	28	**0.26**	not compatible
Yop16-12	1.65	1.81	3	18	31	0.37	1R:2S
Yop16-16	1.56	1.31	3	16	34	0.32	1R:2S
Yop16-41	1.58	1.54	3	16	31	0.34	1R:2S
Yop16-50	1.54	1.55	3	16	23	**0.41**	not compatible
Yop16-6	1.61	1.55	3	12	25	0.32	1R:2S
Yop16-60	2.11	02.09	4	27	28	0.49	2R:2S

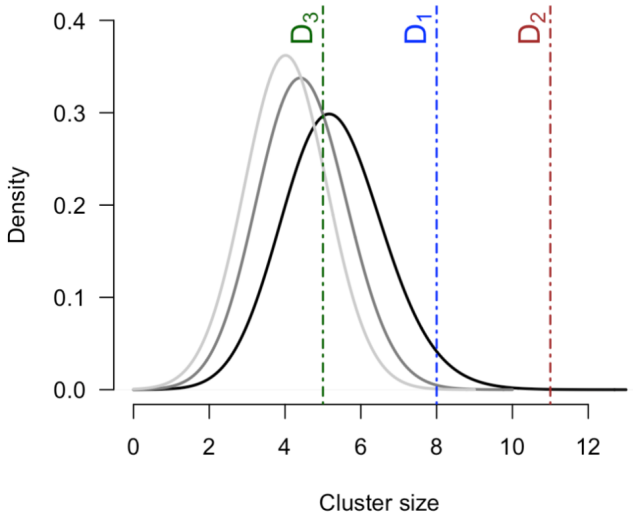
Supporting information Figure 1. A) Expected cluster sizes in a random draw, and B) observed occurrence distribution of the S and D haplotypes.

A) To test our assumption that D(S) sequences could be recognized because they would be part of larger cluster than single-copy S alleles, we computed the distribution of the expected number of identical sequences in a random draw of 56 sequences (2 per 28 diploid individuals) out of the 26 different S sequences that composed our dataset (9 D(S) and 17 single-copy S), over 100,000 iterations. The probability of observing a given number of sequences in a cluster (n_{obs}) is given by 1-(the corresponding quantile in the simulated distribution).

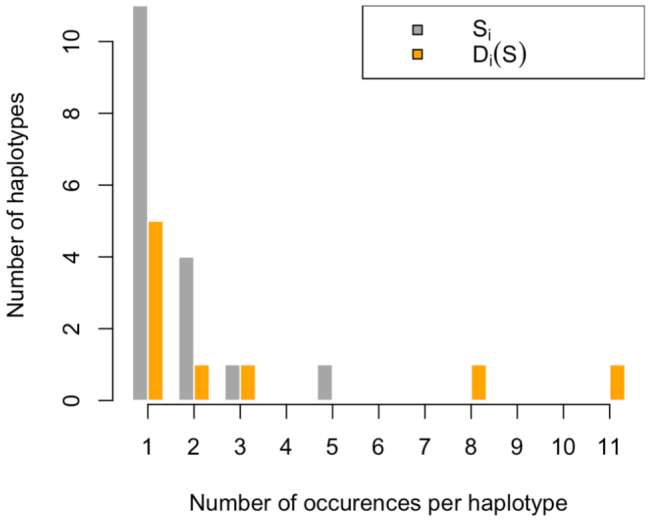
Solid lines represent the expected number of identical sequences for the first, second and third largest clusters (black, dark grey, light grey, resp.) over 100,000 iterations (see text). The dashed-dotted lines are the observed cluster sizes for D_1, D_2 and D_3 (blue, brown and green respectively). Both D_1 (n_{obs} = 8, $p < 0.001$) and D_2 (n_{obs} = 11, $p < 0.001$) clusters were significantly larger than expected, whereas three-sequence clusters are expected in random draws ($p = 0.89$). Thus, apart from the independently-confirmed D_3 allele (see text), D_4(S), D_5(S) and D_6(S) identification remains tentative.

B) Distribution of number of occurrences per haplotype in S and D alleles recovered from the 28 individuals.

A)

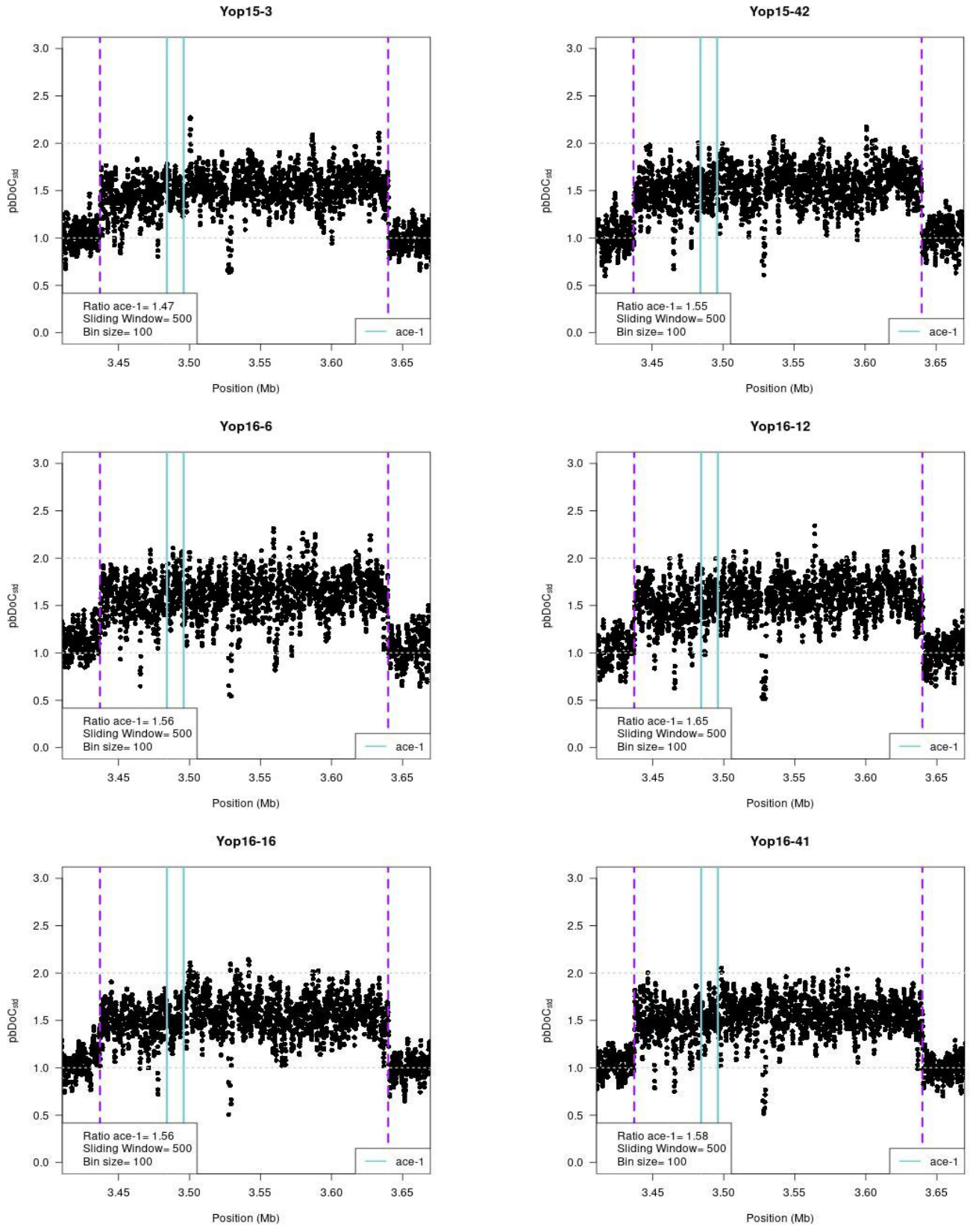


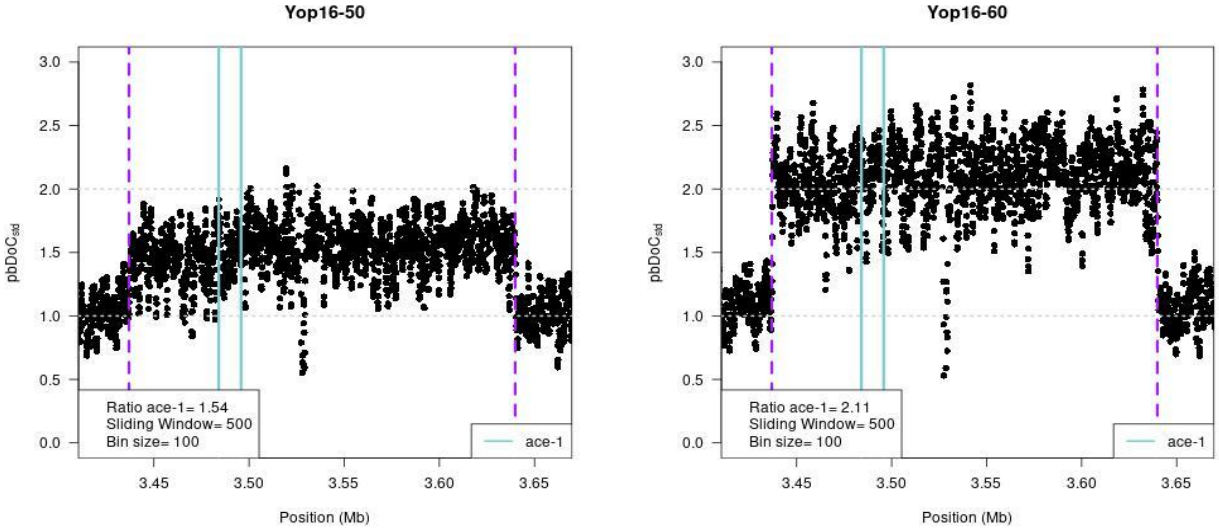
B)



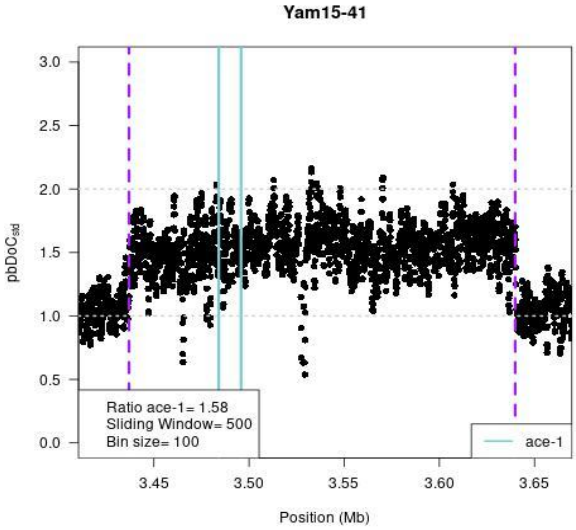
Supporting information Figure 2. Heterogeneous allele structure.

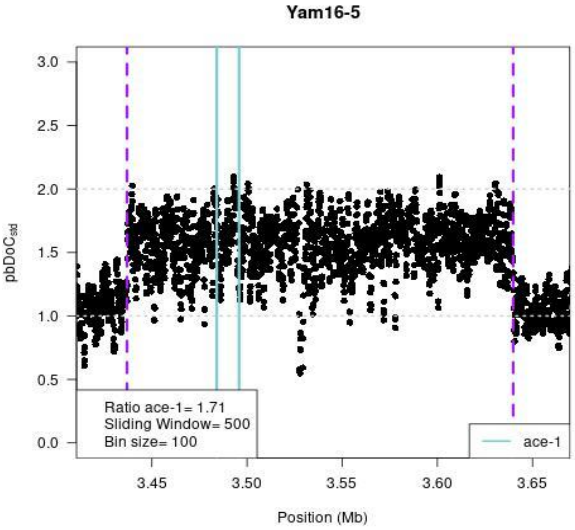
In each graph, we present the variation of the standardized per-base depth of coverage (*pb*DoC$_{std}$, with 1 being the mean *pb*DoC calculated over the whole chromosome) along the chromosomal region of interest (absiss, from 3.4 to 3.7 MB along the chromosome 2R). Each dot is the mean *pb*DoC$_{std}$ calculated every 100 bases (bin size) over 500-base sliding windows. The purple dashed lines represent the amplicon limits of the D_1 and R^x alleles (Assogba et al. 2018); the cyan lines represent the *ace-1* gene location.

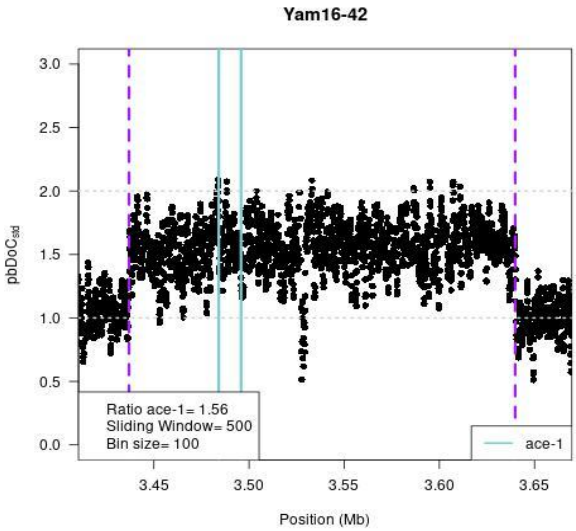


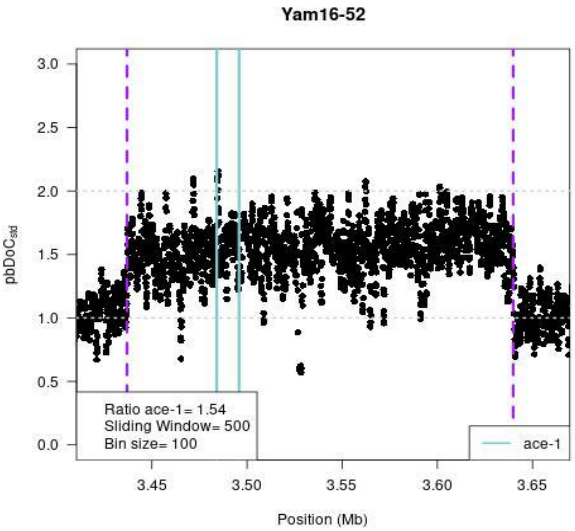

Yop16-50
pbDoC$_{std}$
Ratio ace-1= 1.54
Sliding Window= 500
Bin size= 100
ace-1
Position (Mb)

Yop16-60
pbDoC$_{std}$
Ratio ace-1= 2.11
Sliding Window= 500
Bin size= 100
ace-1
Position (Mb)


Yam15-41
pbDoC$_{std}$
Ratio ace-1= 1.58
Sliding Window= 500
Bin size= 100
ace-1
Position (Mb)

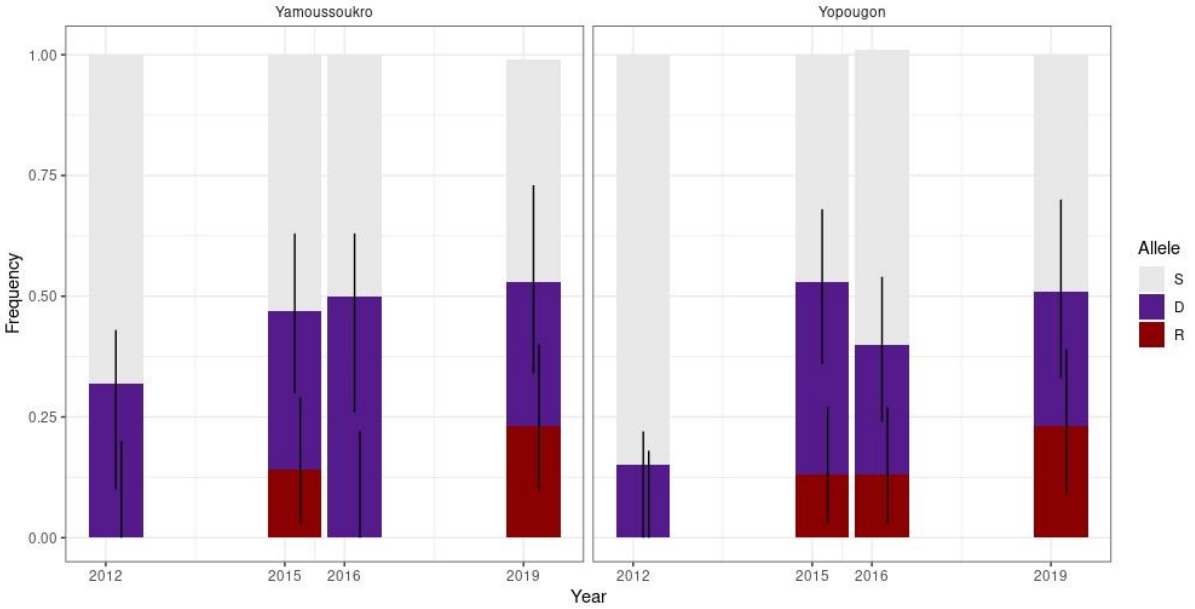

Yam16-5
pbDoC$_{std}$
Ratio ace-1= 1.71
Sliding Window= 500
Bin size= 100
ace-1
Position (Mb)


Yam16-42
pbDoC$_{std}$
Ratio ace-1= 1.56
Sliding Window= 500
Bin size= 100
ace-1
Position (Mb)


Yam16-52
pbDoC$_{std}$
Ratio ace-1= 1.54
Sliding Window= 500
Bin size= 100
ace-1
Position (Mb)

Supporting information Figure 3: R, S and D allelic frequencies estimated in Yamoussoukro and Yopougon.

Data from Assogba *et al.* 2018 (years 2012 to 2016), and from 2019 (this study) were used. We estimated the allelic frequencies considering a 3-allele model (all D alleles are considered together, Tab. 1 Model A) through a maximum likelihood approach (see also *Estimation of duplicated genotype frequencies* in Materials); they are shown in stacked plots with bars corresponding to upper and lower support limits (≈95% confidence intervals).



Aparté taxonomique : de la complexité des complexes.

Jusqu'à présent, j'ai adopté une approche décontractée pour parler des complexes d'espèces *Anopheles gambiae s.l.* et *Culex pipiens s.l.*, comme s'il s'agissait de deux grandes entités populationnelles. Je me suis permis cette simplification dans un premier temps pour faciliter la présentation de la résistance par duplication du gène *ace-1*, sans embrouiller davantage les lecteurs par des considérations taxonomiques. J'ai conservé cette approche tout au long du chapitre I par commodité, puisque je n'y analysais que des données issues de populations homogènes de Côte-d'Ivoire, uniquement *An. coluzzii,* ou des souches croisées sur Slab, *i.e.* des hybrides artificiels. Cependant, le *PipPop Project* (Y. Haba et L. McBride), dont je vais utiliser les données dans le chapitre II, contient des individus appartenant à divers taxons du complexe *Culex pipiens s.l.*. Aussi, il me semble maintenant nécessaire de faire une présentation (rapide et très orientée) des deux complexes d'espèces que j'ai étudiés et des relations phylogénétiques qui les unissent, afin de fournir les informations nécessaires pour apprécier l'intérêt de leur étude parallèle[1].

A.1. Divergence entre *Anopheles gambiae s.l.* et *Culex pipiens s.l.*

Commençons par prendre un peu de recul pour retracer brièvement l'histoire évolutive des moustiques en général. Les moustiques (*Culicidae*) sont une famille de diptères très ancienne, puisque leur émergence daterait de 220 millions d'années (Lorenz *et al.,* 2021), et très diversifiée: en Juillet 2023, on reconnaît 3719 espèces réparties en 111 genres (Mosquito Taxonomic Inventory, Harbach, 2013). On y distingue deux sous-familles, les *Anophelinae* et les *Culicinae*, auxquelles appartiennent respectivement les genres *Anopheles* et *Culex*[2]. Elles auraient divergé il y a environ 197.5 millions d'années[3] (Lorenz *et al.,* 2021). En considérant une moyenne (actuelle, mais extrapolons) de 10 générations par an, ce seraient donc plus d'un milliard de générations qui séparent ces deux genres à l'heure actuelle[4]. Les échanges de gènes entre *Anopheles gambiae s.l.* et *Culex pipiens s.l.* sont donc impossibles par croisement, ces deux complexes d'espèces étant bien trop divergents. Ils sont en revanche

[1] Sans compter que j'avais dit que je le ferai…Chose promise, chose due.
[2] Si jamais on avait des doutes sur les petits préférés de chaque sous-famille.
[3] Pour les personnes qui, comme moi, sont un peu rouillées sur les dates et l'enchaînement des temps géologiques, 190 millions d'années ça nous amène quelque part pendant le Jurassique inférieur. Le dernier ancêtre commun de ces moustiques devait donc être à l'origine des nuits blanches de maintes espèces de dinosaures (mais pas de *T-rex* dont l'émergence est bien plus tardive que le Jurassique, j'ai vérifié…).
[4] Pour aider à réaliser ce que cela représente, c'est nettement plus que le temps écoulé depuis notre ancêtre commun avec les girafes (85-95 Ma), ou même que depuis notre ancêtre commun avec les ornithorynques (160-180 Ma)! On tient de la super anecdote de dîner mondain là, non?

monnaie courante au sein de chaque complexe, et jouent même un rôle clef dans l'histoire des allèles de résistance.

A.2. *Anopheles gambiae s.l.*

Le complexe d'espèces *Anopheles gambiae s.l.* est composé d'un minimum de neuf taxons (O'Loughlin, 2020), aux relations phylogénétiques particulièrement difficiles à élucider. Au sein de ce complexe, seuls les hybrides mâles sont stériles, au contraire des femelles qui profitent de la vigueur hybride (Coluzzi *et al.,* 1979). D''importantes introgressions inter-spécifiques ont été découvertes sur les autosomes des membres de ce complexe, compliquant encore la caractérisation de leurs relations de parenté (Fontaine *et al.,* 2015). L'analyse récente des chromosomes X a toutefois permis d'établir que les deux taxons cibles de ma thèse, *An.* coluzzii et *An. gambiae s.s.*, résultent d'un événement de spéciation très récent (0.5 millions d'années ; Fontaine *et al.,* 2015). Ces deux taxons-frères sont les principaux vecteurs du paludisme (avec le plus distant *An. arabiensis* ; Zoh *et al.,* 2020). Autrefois connus sous les appellations de "formes moléculaires" M (*An. coluzzii*) et S (*An. gambiae s.s.* ; Coetzee *et al.,* 2013), ils étaient principalement discriminés sur la présence de larges *SVs* portés par les chromosomes 2 et X, dont des inversions englobant plusieurs gènes liés à des traits phénotypiques importants tels que la résistance aux insecticides et à la dessiccation (voir Coluzzi *et al.,* 1985 ; Lehmann & Diabaté, 2008). Ces deux taxons sont retrouvés en sympatrie sur une large fraction de leurs aires de répartition respectives, mais occupent des niches écologiques différentes: *An. gambiae* se reproduit principalement dans des habitats aquatiques temporaires engendrés par la saison des pluies, tandis qu'*An. coluzzii* se reproduit toute l'année et est plus facilement retrouvé dans des milieux anthropisés (Fig. A.1.; Lehman & Diabate, 2008; Constantini *et al.,* 2009). Contrairement aux autres croisements dans le complexe d'espèces, les hybrides *An. gambiae s.s.* et *An. coluzzii* mâles et femelles sont entièrement fertiles (Coatzee *et al.,* 2013), aussi l'hybridation *in natura* est fréquente (jusqu'à 20% d'hybrides dans certaines populations, Caputo *et al.,* 2011). Ces hybridations auraient largement favorisé les flux de gènes impliqués dans la capacité vectorielle et dans la résistance aux insecticides (Weill *et al.,* 2003; Djogbenou *et al.,* 2008; Fontaine *et al.,* 2015; Assogba *et al.,* 2016; Grau-Bové *et al.,* 2021).

A.3. *Culex pipiens s.l.*

Comme dans le cas d'*Anopheles*, le complexe d'espèces *Cx. pipiens s.l.* regroupe plusieurs taxons dont le statut phylogénétique aura fait, et continue de faire, couler beaucoup

d'encre. Les taxons sont impossibles à distinguer morphologiquement et les premières classifications de ce complexe se basaient essentiellement sur des critères écologiques. On en distingue généralement cinq: *Cx. australicus* et *Cx. globocoxitus*, endémiques d'Australie, *Cx. pallens*, retrouvé en Chine et Asie du Sud Est, et deux espèces dont l'aire de répartition est la plus large, *Cx. quinquefasciatus*, présent dans toute la zone intertropicale, et *Cx. pipiens s.s.*, cosmopolite dans les zones plus tempérées. Tous ces taxons s'hybrident le long de clines délimitant leurs aires de répartition respectives, notamment pour *Cx. pipiens* et *Cx. quinquefasciatus* en Amérique du Nord, au Moyen Orient et en Australie (Fig.A.1.; Fonseca *et al.,* 2009). Une exception notable s'observe dans les populations d'Afrique du Sud, pour lesquelles on n'observe pas d'hybridation entre *Cx. pipiens* et *Cx. quinquefasciatus* (Cornel *et al.,* 2003). Il a d'ailleurs été proposé que les populations locales, identifiées comme *Cx. quinquefasciatus,* appartiendraient en fait à un troisième taxon différencié (Dumas *et al.,* 2016); l'analyse comparative des séquences *ace-1* menée par Milesi *et al.* (2018) place la séquence d'une souche de cette région en groupe externe et soutient cette hypothèse ("*fun" fact*: ô cruelle ironie, la souche JHB, une souche *Cx. quinquefasciatus* de Johannesburg, est justement celle qui a été choisie pour séquencer le génome de référence de ce complexe d'espèces… Ce n'est pas la dernière fois que je râle à ce sujet, vous êtes prévenus).

On distingue aussi deux écotypes différents au sein de *Cx. pipiens* (parfois élevés au rang d'espèces dans la littérature, mais ce point reste très discuté): i) *Cx. p. pipiens* vit en surface, se reproduit en essaims, et les femelles prennent leurs repas sanguins préférentiellement sur des hôtes aviaires; ii) *Cx. p. molestus*, aussi connu comme "moustique du métro de Londres", vit en gîtes souterrains, ne se reproduit pas en essaim, et les femelles sont plus volontiers anthropophiles. La capacité d'autogénie (capacité des femelles à pondre sans repas sanguin préalable[5]) a longtemps été utilisée pour distinguer ces deux écotypes, mais des études récentes ont montré que ce ne serait en fait pas une caractéristique spécifique à *Cx. p. molestus* (Arich *et al.,* 2022; Haba *et al.,* 2022). *Cx. p. molestus* était autrefois cité comme un cas d'école de spéciation parapatrique: sa découverte dans le métro de Londres par les réfugiés des bombardements de 1945 avait fait naître l'hypothèse qu'il résultait d'une spéciation dans les réseaux souterrains du *Tube*[6]. Cette origine a depuis été réfutée, au profit d'une origine moyen-orientale suivie de colonisations indépendantes *via* des spécialisations

[5] D'ailleurs si elles pouvaient exploiter cette capacité à bien fond et passer le mot, on apprécierait le geste.
[6] J'ai moi-même assisté à des cours en amphithéâtre où un professeur enthousiaste utilisait cet exemple (un grand chauve tatoué, vous remettez ?), et s'en désole aujourd'hui. Ainsi va la Science !

de niche différentes pour chaque écotype, et qui résultent en des différences particulièrement marquées en Europe du Nord (Arich *et al.,* 2022; et voir Haba *et al.,* 2022 pour une revue).

De manière similaire à ce qui s'observe au sein du complexe *An. gambiae s.l.*, l'introgression de larges segments génomiques entre taxons du complexe *Cx. pipiens s.l.* a plusieurs fois été détectée, notamment entre *Cx. pipiens* et *Cx. quinquefasciatus* (Urbanelli *et al.,* 1997; Kothera *et al.,* 2012), dont des gènes de résistance aux insecticides (superlocus *ester*: Labbé *et al.,* 2005; *ace-1:* Fonseca *et al.,* 2009; *vgsc*: Zhou *et al.,* 2009).

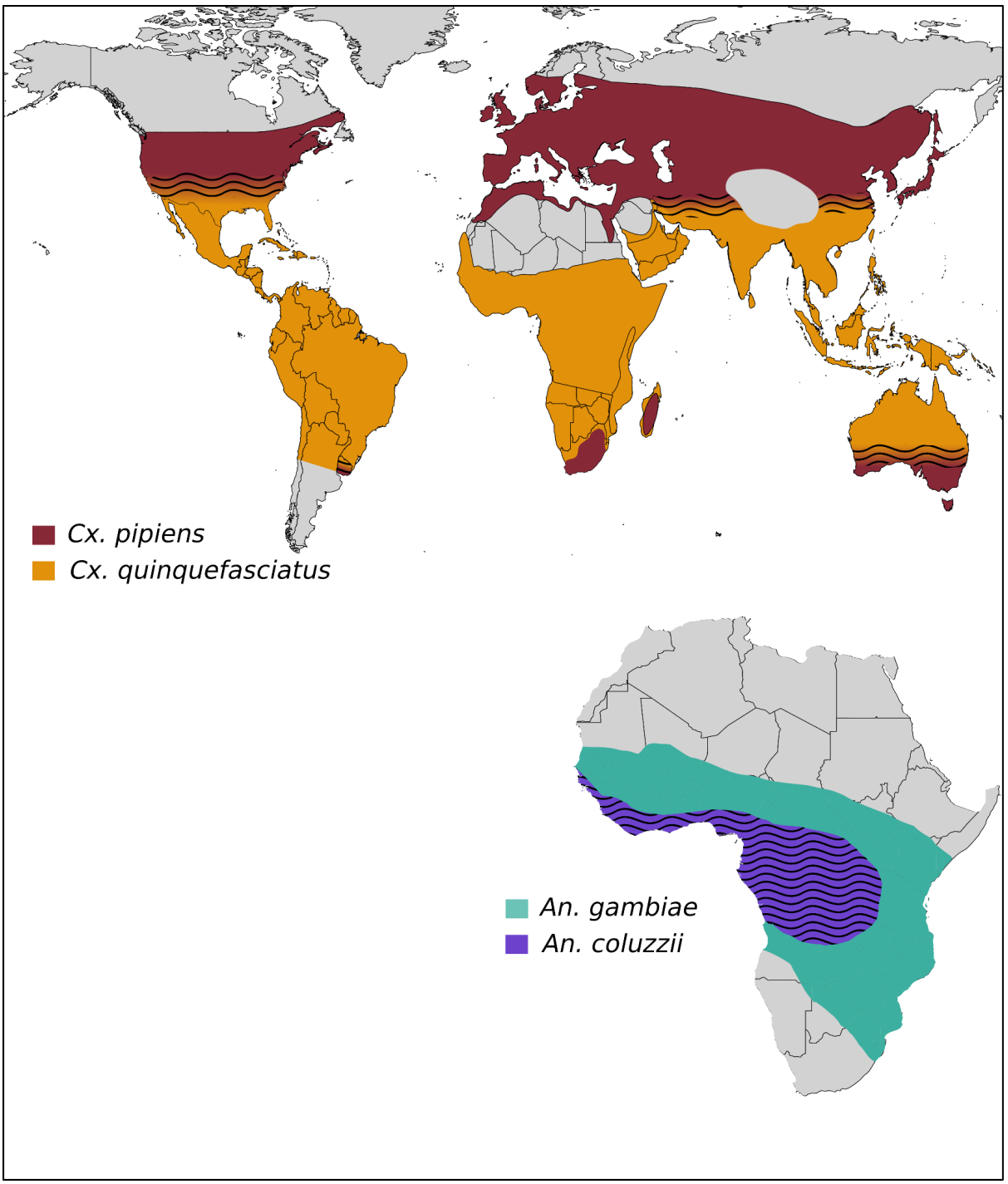


Figure A.1 : Aires de répartition des principales espèces étudiées pendant ma thèse. Aire *Culex* modifiée depuis Haba *et al.* (2022) et aire *Anopheles* modifiée depuis Fontaine *et al.* (2015). Les zones dans lesquelles de l'hybridation entre espèces est attestée sont indiquées par des lignes ondulées.

Chapitre II. Approche populationnelle de la structure des allèles dupliqués et évolution parallèle au locus *ace-1*.

En caractérisant l'architecture des allèles *ace-1* dupliqués fixés dans ces souches disponibles à l'ISEM, j'ai fait le premier pas pour établir l'origine du phénotype homozygote sublétal (HS) observé chez les porteurs de certains allèles D, un des objectifs initiaux de ma thèse. Toutefois, la compréhension de ce phénotype demandera encore beaucoup de travail: une résolution encore plus fine reste en effet nécessaire pour aller plus loin, même si mes travaux ont permis d'apporter des éléments de réponse intéressants et de nouvelles pistes à explorer (ex. les différences de gènes embarqués avec la *ß-catenin like protein 1* ou le gène coupé par la borne de Cp-D_2).

En revanche, un point majeur de ce travail est qu'il m'a permis de mettre au point un outil, *ArDu*, qui me permet de "facilement" identifier et caractériser la structure des duplications sur un gène candidat. Un des avantages d'*ArDu* est qu'il peut être employé sur de grands jeux de données, et l'occasion de l'utiliser s'est d'ailleurs rapidement présentée! Mon équipe travaille depuis plusieurs années en collaboration avec le groupe de recherche de Lindy McBride du département d'écologie et de biologie évolutive de l'Université de Princeton (USA). Elle s'intéresse notamment à l'histoire évolutive du complexe *Cx. pipiens s.l.*. Pour l'explorer, son doctorant Yuki Haba a acquis pendant sa thèse un jeu de données simplement monumental : le *PipPop Project* comprend actuellement les génomes de 830 moustiques du monde entier appartenant à ce complexe ! Non contents de l'excellence de leur travail, ces deux chercheurs sont également des collègues particulièrement généreux, totalement ouverts à la collaboration pour faire avancer la connaissance : ils ont ainsi eu l'extrême gentillesse de partager le fruit de leur travail avec mon équipe, me laissant carte blanche pour y explorer tout ce que je voulais en lien avec la résistance aux insecticides ![7] J'ai ainsi pu me lancer dans l'analyse des structures d'*ace-1* sur des données bien différentes de celles que j'avais jusqu'alors étudiées, délaissant mes assemblages hybrides, qui n'ont de naturel que la région autour d'*ace-1* (le reste provenant de cette *lab frea*k qu'est devenue Slab après plus de 50 ans confinée en insectariums), pour des individus prélevés directement en populations naturelles, avec toutes leurs histoires démographiques et de sélection particulières. Bref, un nouveau défi, mais tellement enthousiasmant !

[7] Les mots me manquent pour exprimer ma gratitude. L'accès à ce jeu de données et les interactions avec ces collègues extra-ordinaires (scientifiquement comme humainement) ont profondément impactés ma thèse et ont élargi le spectre des questions auxquelles j'ai pu, et je continue, de m'intéresser.

I. A la recherche des allèles dupliqués: explorer le *PipPop Project*

Le *PipPop Project* a pour objectif d'établir une base de données génomiques des populations naturelles *Cx. pipiens s.l.* (à l'image de ce qui existe chez *An. gambiae s.l.* avec le *Anopheles gambiae* 1000 Genomes Consortium, 2021). Il est toujours en développement, et le nombre de génomes séquencés continue de grandir. La partie que j'ai analysée comprend 830 génomes (profondeur de séquençage 15X, *short reads paired end, Illumina*), provenant de trois taxons (694 échantillons identifiés par analyse d'admixture, dont 181 *Cx. pipiens,* 259 *Cx. molestus,* 116 *Cx. quinquefasciatus*, et 138 hybrides) réparties en Afrique, Asie, Amérique, Europe et Océanie. La quasi-totalité des 46 pays échantillonnés sont représentés par plusieurs localités (153 au total), dans lesquelles entre cinq et quinze moustiques ont été séquencés.

I.1. Structure des allèles.

Quelques considérations techniques.

J'ai tiré parti du pipeline *ArDu* pour explorer les données du *PipPop Project* à la recherche de duplications du gène *ace-1*. La transposition du protocole d'analyse des données de séquençage des souches, plus couvertes (séquençage 30X), aux données de populations naturelles (15X) a demandé d'abaisser le seuil de nombre de copies séparant les duplications du reste du génome en simple copie (de 1.5 à 1.3) pour accommoder la plus faible couverture relative des génomes et la présence d'hétérozygotes. Cette modification s'est immanquablement accompagnée d'un certain nombre de faux positifs (une trentaine sur les 830 individus du screen) qu'il m'a alors fallu trier "à la main" : j'ai vérifié les graphes de DoC et le nombre de copies des gènes flanquants pour séparer le bon grain de l'ivraie (*i.e.* les dupliqués du reste). Autre point d'intérêt, j'expliquais dans le chapitre I que le nombre de copies du gène pouvait être calculé en utilisant la *DoC* moyenne du chromosome sur lequel se situe le gène (DoC_{CHROM}), ou en utilisant la *DoC* moyenne d'un gène de référence non dupliqué (DoC_{REF}; à la manière d'une qPCR, voir Box. I.1). En analysant les données du *PipPop Project*, j'ai découvert que le nombre de copies établi à partir de la DoC_{REF} était systématiquement plus fiable que celui calculé avec la DoC_{CHROM} (il produisait moins de faux positifs). Les valeurs de nombre de copies dont je parlerai dans la suite de ce chapitre ont donc toutes été obtenues par cette approche, en utilisant le gène *ace-2* comme référence, un paralogue d'*ace-1* fréquemment utilisé pour les qPCR au sein de mon équipe.

Dans l'ensemble, la détection des duplications du gène *ace-1* s'est ensuite déroulée sans accroc, et j'ai pu identifier un total de 76 moustiques porteurs d'une duplication dans un échantillonnage couvrant l'ensemble du globe (Tab. II. 1).

J'ai retrouvé ici les différences de pourcentages de couverture du génome de référence que j'avais identifiées entre souches dupliquées : on observe une nette diminution pour tous les individus classés comme *Cx. p. pipiens* ou *Cx. p. molestus* (moyenne 51.5% couvert, écart type 2.9) par rapport à ceux identifiés comme *Cx. quinquefasciatus* (moyenne 72.39% couvert, écart type 1.9). Si ces différences étaient limitées à la zone entourant *ace-1* dans les souches, elles s'observent ici sur la totalité du génome (ce qui me rassure de nouveau quant à l'efficacité du protocole d'introgression utilisé pour la fixation des souches).

J'ai tout de même rencontré des difficultés dans l'identification des points de cassure (*breakpoints*) de ces duplications, et ce quelque soit leur origine taxonomique. Pour la quasi totalité des individus dupliqués, je n'ai pas réussi à trouver de congruence entre les données de tailles d'*insert* et de positions de *soft-clipped reads* (voir Chap. I). On voit sans doute ici une limite de mon protocole d'analyse, qui semble nécessiter une forte profondeur de séquençage (en tous cas supérieure à 15X) pour permettre une identification formelle des *breakpoints*. Il est aussi probable que des hétérozygotes D_x/D_y existent dans ce jeu de données, dans ce cas certaines des structures que j'observe sont des chimères de celles des duplications qu'ils portent, et les signaux de leur bornes seront d'autant plus difficiles à interpréter.

Des structures c'est bien, mais des allèles ce serait mieux!

Dans ce chapitre, je vais vous présenter des structures d'allèles dupliqués et soigneusement éviter d'évoquer la nature des copies *ace-1* qu'elles portent (*i.e.* les différents haplotypes *ace-1*). En effet, dans le cas qui nous concerne, il me semble hasardeux de prédire le génotype d'un échantillon uniquement depuis des données NGS, et mes griefs contre cette approche sont notamment motivés par les analyses que j'ai menées sur les données de populations naturelles du *PipPop Project,* du *Anopheles gambiae 1000 Genome Project* et des populations naturelles de Côte d'Ivoire (**Claret *et al.*, soumis**). Je m'explique : ici, le génotypage bioinformatique repose sur le ratio des variants d'une unique position de l'assemblage de référence, qui correspond dans la réalité aux données du séquençage de plusieurs zones génomiques différentes (*i.e.* les haplotypes *ace-1* de chaque amplicon d'une duplication). Les analyses que j'ai menées sur la profondeur de couverture m'ont montré que cette dernière varie beaucoup, que ce soit à l'échelle chromosomique, dans les gènes, et

même dans les exons. Il résulte de cette variation que les différents haplotypes *ace-1*, et surtout les différentes positions permettant leur identification comme copie R ou S, ne sont pas couvertes de manière égale. Qui plus est, cette différence de couverture n'est *a priori* pas prédictible, *i.e.* la base sensible n'est pas systématiquement moins couverte que la base résistante, et inversement[8]. Pour résumer, il m'apparaît que la couverture des différentes bases d'une mutation diagnostique n'est pas un bon prédicteur des ratios de copies R/S d'*ace-1* d'un allèle dupliqué (**Claret *et al.,* soumis**), même si cela est régulièrement utilisé dans la littérature (Djogbénou *et al.,* 2015, Grau-Bové *et al.,* 2021). Pour avoir une idée réelle de ce ratio, il semble préférable d'utiliser des qPCR, ou idéalement une combinaison des deux approches. C'est ce que nous avons fait dans le cas de Cp-D_1, ce qui a permis de montrer que cet allèle comporte trois copies *ace-1*, dont deux copies R et une copie S. Cet haplotype S ne diffère d'ailleurs de l'haplotype porté par les copies R que par l'unique mutation G119S sur 700 pb (Labbé et al. 2007, Milesi et al. 2018).

Admettons néanmoins pour un instant que l'on choisisse d'utiliser malgré tout ces estimations du nombre de copies. Se pose alors le problème de leur attribution. Prenons pour exemple un individu pour lequel on observe un ratio de une copie S pour deux copies R : il peut être un hétérozygote R^2/S, mais également un D/R^1. De la même façon, dans les cas d'individus ne portant que des copies R, il est impossible de savoir comment elles se répartissent: un individu avec six copies R peut en effet avoir plusieurs génotypes possibles, R^1/R^5, R^2/R^4 ou R^3/R^3. L'attribution des copies me paraît alors hautement spéculative, raison pour laquelle j'ai préféré m'en abstenir.

Diversité de structures des duplications *ace-1*.

Sur les 76 individus dupliqués, j'ai pu identifier un total de six nouvelles structures, toutes différentes des trois précédemment identifiées à partir des souches (Cp-D_1, Cp-D_2 et Cp-D_3, Fig II.1 α à ε). Au regard de la faible précision des prédictions de *breakpoints*, j'ai écarté trois potentielles structures. Leurs points de cassures étaient à chaque fois situés à moins de 10 kb de ceux de structures déjà identifiées dans le même pays.

Un certain nombre de structures semblent avoir subi des réarrangements secondaires : i) dans le cas de γ et δ, j'ai identifié des duplications partageant des *breakpoints* similaires, mais présentant des nombres de copies différents (variations de trois à cinq copies observées sur tous les gènes embarqués, Tab. II.1). ii) Plus étonnant encore, la structure ε^+ montre une

[8] Mon instinct me murmure d'ailleurs qu'une faible profondeur de séquençage aggraverait encore ces différences, mais là on est dans le sentiment plus que dans l'analyse objective.

Table II.1. Répartition des structures des allèles dupliqués *ace-1 Culex pipiens s.l.* dans le *PipPop Project.* Pour chaque structure sont indiqués le taxon et la population dans lesquelles elle est retrouvée, sa taille et le nombre d'amplicons qu'elle comporte. Des structures communes sont souvent partagées entre *Cx. p. pipiens* et *Cx. p.* molestus, et dans une moindre mesure entre *Cx. quinquefasciatus et Cx. p. molestus*. Seule la structure *α* est partagée entre *Cx. quinquefasciatus* et *Cx. p. pipiens*, mais probablement plutôt du fait d'une apparition indépendante que d'une introgression d'un taxon vers l'autre.

Structure	Taxon	Pays	Région	Nombre de copies	Position des points de cassure et taille
α	*Cx. quinquefasciatus*	Cameroon	Yaoundé	2	71 648 000 - 71 791 000 : 143 kb
	hybride *quinquefasciatus-molestus*	USA	Atlanta	2	
	Cx. quinquefasciatus	USA	Bakersfield	2	
	Cx. quinquefasciatus	USA	Florida	2	
	Cx. quinquefasciatus	USA	Palmdale	2	
	hybride *quinquefasciatus-molestus*	USA	San Jose	2	
	Cx. p. pipiens	Algérie	Oran	2	
β	*Cx. p. pipiens*	Egypte	Ash Sharqiya	2	71 602 000 - 71 908 000 : 306 kb
	Cx. p. molestus	Espagne	Garrapinillos	2	
	hybride *pipiens-molestus*	Espagne	Sevilla	2	
	Cx. p. pipiens	Israël	Jordan Valley	2	
	Cx. p. pipiens	Tunisie	Testour	2	
γ	*Cx. p. pipiens*	Afrique du Sud	Gauteng	3	71 634 000 - 72 012 000 : 378 kb
	Cx. p. pipiens	Egypte	Giza	3	
	Cx. p. pipiens	Egypte	Le Caire	3	
	Cx. p. pipiens	Espagne	Extremadura	3	
	Cx. p. pipiens	Grèce	Athène	3	

	hybride *pipiens-molestus*	Grèce	Chania	3	
	Cx. p. pipiens	Israël	Emek	3	
	hybride *pipiens-molestus*	Israël	Rosh HaAyin	4	
	Cx. p. pipiens	Israël	Yehuda	3	
	Cx. p. pipiens	Italie	Rome	3 à 4	
	Cx. p. molestus	Italie	Veneto	3 à 5	
	Cx. p. molestus	Turquie	Ankara	3	
	Cx. p. pipiens	Turquie	Yalova	3	
δ	*Cx. p. pipiens*	Algérie	Algiers	3	71 622 000 - 71 975 000 : 363 kb
	Cx. p. pipiens	Ethiopie	Addis Ababa	3	
	Cx. p. pipiens	Russie	Moscou	3 à 4	
	Cx. p. pipiens	Russie	Volgograd	3 à 5	
Cp-D_2	*Cx. p. molestus*	France	Montpellier	2	71 641 523 - 71 988 000 : 346 kb
	Cx. quinquefasciatus	Afrique du Sud	Vhembe	2	
	Cx. quinquefasciatus	Arabie Saoudite	Jeddah	2	
	Cx. quinquefasciatus	Malaisie	Melaka	2	
ε+	*Cx. quinquefasciatus*	Qatar	Al Rayyan	3	
	Cx. quinquefasciatus	Qatar	Doha	4	71 648 000 - 720 080 000 : 432 kb
	Cx. quinquefasciatus	Mozambique	Maputo	2	
	Cx. quinquefasciatus	Taiwan	Taipei	2	
ε+*	*Cx. quinquefasciatus*	Soudan	Khartoum	4	

augmentation de *DoC* de la zone autour du gène de l'HJHBP, la même que celle déjà observée chez Cp-R_2[3]: les bornes de cette duplication interne étant partagées, il est possible que les duplications *ace-1* aient embarqué une autre duplication pré-existante de ce gène, expliquant ainsi la structure complexe retrouvée aujourd'hui (Fig. II.1.iii) Enfin, j'ai été agréablement surpris[9] de découvrir sur la structure ε^{+}* une diminution de la *DoC* sur 210 kb, semblant débuter sur la borne 3' de la duplication interne HJHBP et s'arrêtant à 70 kb de la fin de l'amplicon : elle semble correspondre à une délétion secondaire qui retirerait presque tous les gènes embarqués à l'exception d'*ace-1* et de deux autres locus (HJHBP et un gène non identifié, Fig. II.1), similaire à celle observée dans les allèles R^x chez *An. gambiae s.l.* (Assogba *et al.,* 2016).

J'ai annoté les éléments transposables retrouvées dans la zone dupliquée (500 kb recouvrant toutes les structures identifiées ; *RepeatMasker,* Smit *et al.,* 2015 ; Insecta *library* Dfam 3.1, RepBase-20181026). J'en ai trouvé beaucoup, trop d'ailleurs pour les représenter de manière intelligible : 45% de la séquence totale était masquée par *RepeatMasker*, dont 11% était composés de rétroéléments (8% de LINEs) et 26% de transposons. Leur présence pourraient expliquer la position de certains *breakpoints* et leur diversité, et la quasi-totalité des points de cassure que j'ai identifiés sont justement situés sur des zones répétées ou annotées comme des éléments transposables. Toutefois, en regard de la faible précision de prédiction des *breakpoints,* il est difficile d'identifier un élément transposable spécifique ayant pu faciliter leur formation.

Malgré mes réserves exprimées plus haut quand à l'identification et la distribution des copies R et S, je soupçonne les structures δ, γ et ε^{+}* de correspondre à des duplications homogènes (allèles Cp-R^x). En effet, j'ai plusieurs fois retrouvé ces structures dans le génome d'individus pour lesquels je ne retrouvais pour *ace-1* que des *reads* porteurs de la mutation de résistance (haplotypes R). J'identifie donc ces individus avec une certitude raisonnable comme des [RR] (la profondeur de couverture de cette base diagnostique était égale ou supérieure à la *DoC* moyenne du gène, ce qui rend peu plausible l'idée qu'une hypothétique copie S n'ait pas été séquencée). Néanmoins, j'ai aussi retrouvé les structures δ, γ et ε^{+} chez des individus ayant des *reads* porteurs de la mutation de résistance et des *reads* porteurs de l'allèle sensible (haplotypes R et S). La couverture relative des deux mutations varie énormément, et je ne peux donc pas affirmer que ces individus ne sont que des R^x/S ; il n'est pas exclu qu'ils soient en réalité porteurs d'allèles D ayant la même structure que ces allèles

[9] justement parce que cela renforçait encore l'évolution parallèle de *Cx. pipiens s.l.* et *An. gambiae s.l.*, dont je compte faire le coeur d'une publication à venir, voir plus bas.

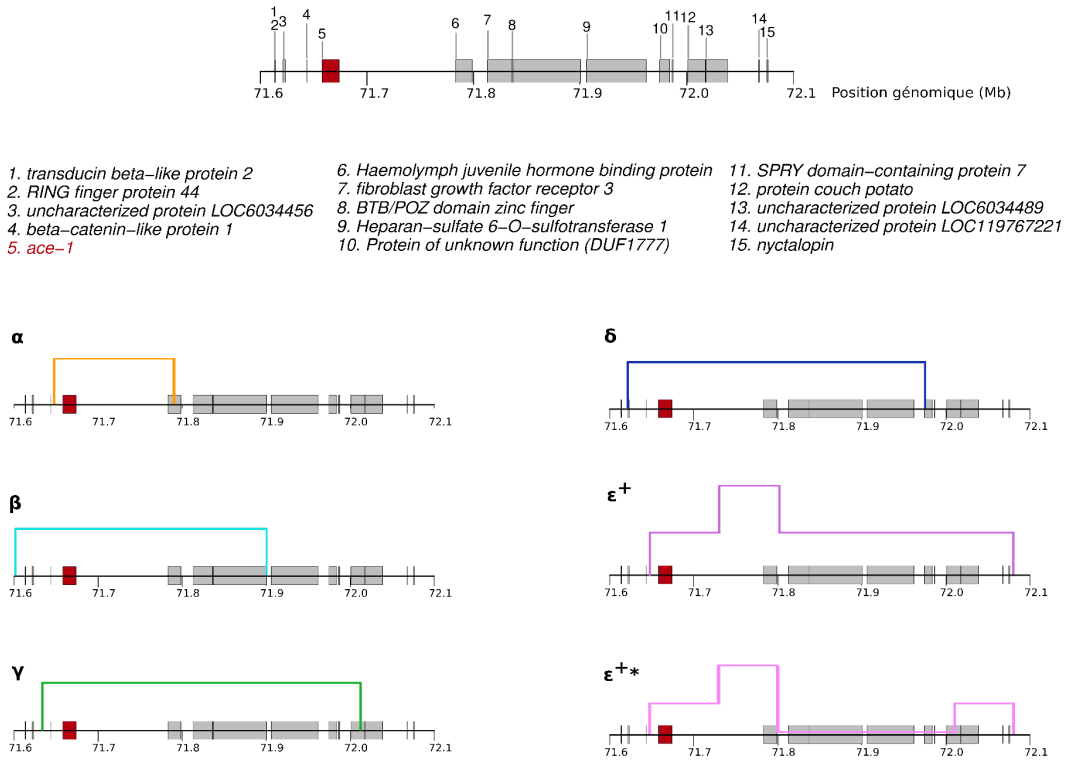


Fig II.1. Structures identifiées dans les données du *PipPop Project*. Les bornes de chaque structure sont indiquées par les lignes verticales en couleur. La position des gènes embarqués (schéma en haut) est représentée par les rectangles gris à l'exception d'*ace-1* qui est en rouge. La hauteur des structures n'est pas indicative : les structures **γ** et **δ** présentaient des nombres de copies variables (de 3 à 4 copies et 3 à 5 copies, resp.) contre 2 copies seulement pour les autres structures. Les structures ε partagent les mêmes bornes et une origine taxonomique commune, mais sont porteuses de réarrangements : une possible duplication interne ($\boldsymbol{\varepsilon}^{+}$), parfois associée à une délétion ($\boldsymbol{\varepsilon}^{+*}$). Ces structures sont toutes différentes des structures décrites pour Cp-D_1, Cp-D_2 et Cp-D_3 (voir Chap I, Fig. I.4).

R^x (comme observé pour Cp-D_1 et Cp-R_2[3] ou encore *An. gambiae s.l.* pour les allèles Ag-D et Ag-R^x, Assogba *et al.* 2016, **Claret *et al.*, soumis**). Je n'ai pas la possibilité de distinguer ces deux hypothèses avec les outils à ma disposition, puisque je ne peux pas savoir quel haplotype est associé à quel allèle (pour cela il faudrait pouvoir faire des croisements ou avoir des haplotypes continus sur l'ensemble de la zone dupliquée, par exemple avec du séquençage *long reads*).

On peut donc observer parmi les 76 individus dupliqués une grande diversité de structures pour ces duplications autour du locus *ace-1*, que ce soit dans la taille de la duplication (de 142 kb à 432 kb), dans la position des *breakpoints*, ou encore dans le nombre de copies (2 à 5 ; Tab. II.1). Cela suggère des origines indépendantes pour chacune de ces structures, résultant probablement des mécanismes différents (*crossing-over* inégaux indépendants ou réversion d'un allèle R^x en allèle D, confirmant ainsi les intuitions des études précédentes, voir Labbé et al., 2007), mais toujours ces duplications sont toutes sélectionnées en réponse à l'application d'insecticides pour contrôler les populations de ces moustiques. On a donc ici un extraordinaire exemple d'évolutions adaptatives indépendantes et répétées au sein du complexe *C. pipiens s.l.*

I.2. Origine géographique.

Grâce aux données du *PipPop Project*, j'ai pu m'intéresser à la répartition taxonomique et géographique des structures. Sans surprise, la plupart d'entre elles semblent être particulièrement associées à un taxon (Tab. II. 1) : je n'ai retrouvé chez *Cx. quinquefasciatus* que α et ε, et dans les populations *Cx. pipiens* que β, γ et δ, à l'exception de deux individus d'Algérie qui présentent la structure α (Fig. II.2.). Étonnamment, malgré leur forte représentation dans le jeu de données (259 sur 830), aucune structure ne semble particulièrement associée aux populations *Cx. p. molestus* : j'en retrouve plusieurs, partagées avec les autres taxons, suggérant soit des acquisitions convergentes indépendantes, ou plus probablement des introgressions.

Comme pour la répartition taxonomique, la distribution géographique des duplications est nettement structurée. Ainsi β et γ sont présentes sur de larges régions du pourtour Méditerranéen et au Moyen-Orient (Fig. II.2.), tandis que d'autres structures sont retrouvées de manière plus ponctuelle en des points éloignés, comme ε qui est présente au Moyen Orient, en Afrique du Sud et en Asie du Sud-Est (Fig. II.2.). Là-encore, ces résultats peuvent s'expliquer par une migration (éventuellement favorisée par des transports humains, Medley *et al.,* 2015 ; Milesi *et al.,* 2018 ; Rose *et al*., 2023), mais il est difficile d'exclure des

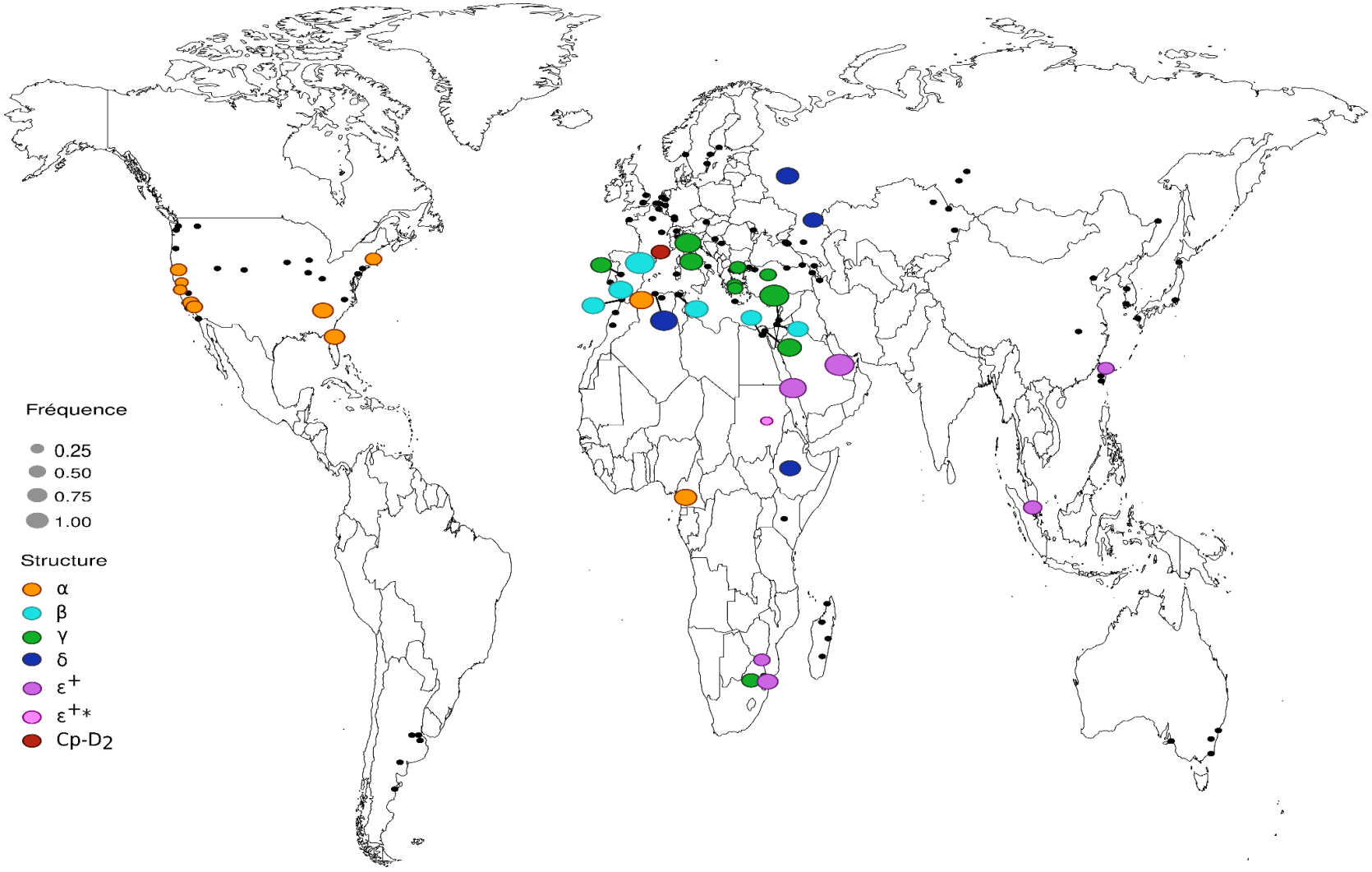


Figure II.2. Carte de l'échantillonnage du *PipPop Project*, et fréquence des structures dupliquées dans les populations prélevées. Les sites échantillonnés sont représentés par les points noirs. Les populations dans lesquelles des structures dupliquées ont été retrouvées sont indiquées par les points de couleur dont la taille correspond à la fréquence de la structure dans la population.

apparitions indépendantes, liées à des *breakpoints* particuliers qui favoriseraient localement la recombinaison et les *crossing-over* inégaux.

Pour démêler ces hypothèses d'introgressions/migrations *vs* origines indépendantes des différentes structures de duplication, il faudrait être en mesure de comparer les différents haplotypes portés par ces structures. Malheureusement, obtenir des haplotypes à partir de données d'alignement *short read* est particulièrement complexe. Il faut établir l'existence d'un déséquilibre de liaison (*linkage disequilibrium* ou *LD*) entre *SNPs* (*i.e.* identifier des *SNPs* fréquemment associés), et pour cela disposer de données phasées. Dans le meilleur des mondes[10], j'aurai exploré cette piste plus en détail (et je m'y intéresserais sans doute plus tard), mais à défaut, j'ai réalisé une ACP (*PLINK* ; Purcell *et al.,* 2007) sur les données de *SNPs* obtenues avec l'outil GATK (voir Box I.1), en me restreignant à une petite zone autour d'*ace-1*, comprenant la duplication et quelques kb des zones flanquantes (de 71,3 à 72,02 Mb ; Fig. II.3.A). On voit dans cette ACP que les individus sont d'abord groupés par taxon : l'axe 1 (PC1 : 36% de la variation totale) discrimine nettement *Cx. quinquefasciatus* de *Cx. p. pipiens*/*Cx. p. molestus*, ces deux derniers taxons étant phylogénétiquement plus proches (Haba *et al.,* 2022). Le second axe semble vaguement regrouper les individus par population (PC2 : 16% variation totale, Fig. II.3.A), mais l'espèce semble donc être le facteur le plus discriminant au vu du reste des composantes principales. Les deux individus *Cx. p. pipiens* d'Algérie porteurs de la structure α sont d'ailleurs bien regroupés avec le reste des individus *Cx. pipiens,* ce qui suggère une apparition indépendante de cette structure dans les deux taxons plutôt qu'une introgression. J'ai par ailleurs identifié plusieurs hybrides entre taxons: les hybrides *molestus*/*pipiens* portaient des structures associées à *pipiens*, les hybrides *molestus*/*quinquefasciatus* des structures associées à *quinquefasciatus*. Ces observations supportent l'idée de potentielles introgressions des allèles dupliqués de *Cx. p. pipiens* ou *Cx. quinquefasciatus* vers *Cx p. molestus.* Les ACP réalisées indépendamment sur les taxons (*Cx. p. pipiens* et *Cx. p. molestus*, Fig. II.3.B ; et *Cx. quinquefasciatus,* Fig. II.3.C) ne permettent pas de regrouper les individus par structure de manière plus efficace que l'ACP globale, ou alors très marginalement seulement. Cette analyse est quoi qu'il en soit impactée par le fait qu'une partie des individus dupliqués sont probablement hétérozygotes, et que nous observons par conséquent ici un mélange d'haplotypes dupliqués et sauvages. De plus, il est possible que les haplotypes contenus dans les allèles dupliqués recombinent avec des haplotypes d'allèles sauvages, ou même avec d'autres haplotypes d'allèles dupliqués

[10] où ma thèse aurait duré plus de 3 ans.

Figure II.3. ACP de la zone entourant *ace-1* pour les individus dupliqués du *PipPop Project*: A. pour tous les échantillons dupliqués; B. chez *Cx. p. pipiens et Cx. p. molestus*; C. chez *Cx. quinquefasciatus*. Réalisée avec *PLINK* sur les données de *SNPs* de 71,3 à 72,02 Mb sur le chromosome 3 (correspondant à la zone dupliquée la plus large et quelques kb autour). Le taxon auquel appartient chaque individu est indiqué par différentes couleurs, et la structure de la duplication qu'il porte par des formes. Le pourcentage de variance expliquée par chaque composante principale (PC) est indiqué entre parenthèses.

ségrégeant dans les mêmes populations (ce dont nous avons discuté pour *An. gambiae* dans **Claret *et al.*, soumis**). Considérant tous ces facteurs, il devient alors compliqué d'étudier les structures sous l'angle des *SNPs* qui leur sont associés.

Au final, grâce aux données mondiales collectées dans le *PipPop Project,* j'ai donc mis en évidence plusieurs nouvelles structures d'allèles dupliqués *ace-1*. Ces structures varient en taille, parfois en nombre de copies des amplicons, et peuvent comporter des modifications secondaires (délétion, duplication interne), ce qui semble indiquer que plusieurs allèles sont apparus indépendamment au sein du complexe *Cx. pipiens s.l.* en réponse aux pressions de sélection exercées par les insecticides.

II. Evolution parallèle : *An. gambiae s.l.* et *Cx. pipiens s.l.*

L'évolution parallèle est définie comme l'acquisition indépendante par deux espèces d'un même trait, suite à la modification d'une même structure, et en réponse à une pression de sélection similaire. Les tribulations évolutives du locus *ace-1* observées dans les complexes d'espèces *An. gambiae s.l.* et *Culex pipiens s.l.*, depuis la mutation ponctuelle à l'origine de la résistance jusqu'aux larges duplications multi-géniques, correspondent donc bien à cette définition. Je vous propose d'étudier l'étendue de leurs similarités, et leurs différences notables. Pour plus de clarté, la table II.2 résume l'ensemble des informations apportées par les paragraphes suivants.

II.1. Mutations de résistance

Dans l'introduction, j'ai détaillé la similarité des adaptations par mutation ponctuelle au locus *ace-1* chez les moustiques des complexes d'espèces *An. gambiae s.l.* et *Cx. pipiens s.l.*. Dans les deux cas, la résistance est due à une même mutation (G119S), le changement d'une adénine en guanine sur le site actif de l'AChE1 (Weill *et al.,* 2003 ; Weill *et al.,* 2004). Bien sûr, l'identité des mutations de résistance et la forte similitude de l'AChE (81% d'homologie au niveau protéique ; Weill *et al.,* 2003) peuvent s'expliquer par la forte contrainte qui s'exerce sur le gène *ace-1*, puisqu'il code pour une molécule essentielle du système nerveux. Néanmoins, au regard de la divergence de ces deux complexes d'espèces (>1G générations, *cf* Aparté taxonomique), cette similarité reste remarquable.

Dans le complexe *An. gambiae s.l.*, on ne connaît qu'un seul allèle R (*i.e.* un seul haplotype), d'ailleurs toujours trouvé dans des duplications homogènes. Il aurait d'abord émergé chez *An. coluzzii* avant de passer chez *An. gambiae s.s.* (Weill *et al.,* 2003; Djogbenou *et al.,* 2008 ; Assogba *et al.,* 2016 ; Grau-Bové *et al.,* 2021). A l'inverse, il existerait plusieurs allèles R chez *Culex pipiens s.l.,* correspondant à plusieurs émergences indépendantes de la mutation de résistance G119S. Dans leur étude de 2018, Milesi *et al.* identifient quatre haplotypes : R_3 uniquement retrouvée dans une population *Cx. quinquefasciatus* des Philippines ; R_4 qui provient de *Cx. pipiens* d'Israël, et enfin R_1 et R_2 plus largement répandus respectivement dans *Cx. pipiens* et *Cx. quinquefasciatus* ; R_1 est trouvé dans tous les pays méditerranéens de l'Ouest, et R_2 de l'Amérique à l'Afrique, en passant par l'océan Indien et l'Australie. Enfin, une autre mutation générant de la résistance aux OP (F290V) est retrouvée dans les populations *Cx. pipiens* de Méditerranée, mais reste plus rare (Alout *et al.,* 2007, 2009 ; Arich *et al.,* 2021).

II.2. Synténie des zones dupliquées

On l'a vu, les duplications au locus *ace-1* sont, chez *An. gambiae s.l.* comme chez *Cx. pipiens s.l.*, de larges structures de plusieurs kilobases, embarquant plusieurs gènes (Fig II.1). Dans la recherche de parallèles entre les complexes *Anopheles* et *Culex*, il m'a semblé intéressant d'en comparer le contenu et de voir si les gènes embarqués étaient similaires entre les deux complexes, ou si le milliard de générations les séparant avait fait son œuvre. Arensburger *et al.* (2010) avaient déjà réussi à mettre en évidence des similarités entre les génomes de ces espèces (en se servant du précédent assemblage *Cx. quinquefasciatus*), et notamment entre les chromosomes portant *ace-1* sur lesquels j'ai décidé de restreindre mon analyse. J'ai d'abord choisi une approche très générale, en utilisant des outils de recherche de synténie (Mummer, Kurtz *et al.,* 2004 ; Syri, Goel *et al.,* 2019) pour comparer ces chromosomes (chromosome 2 pour *An. gambiae* et 3 pour *Cx. quinquefasciatus*). Rapidement, j'ai aligné les régions d'intérêt (*i.e.* les chromosomes) à l'aide de *Minimap2* (Li, 2021), puis j'ai utilisé l'outil *Syri* pour identifier des zones partagées. En parallèle, j'ai recherché l'existence de zones synténiques en utilisant *Mummer*, qui est basé sur un protocole similaire à celui que je viens de décrire, mais permet d'effectuer ces recherches au niveau protéique (algorithme *Promer*). Cette approche s'est rapidement révélée trop imprécise : je n'ai pu identifier qu'une seule zone partagée entre les assemblages, mais elle était située loin des zones dupliquées dans les deux cas. Déçu mais pas vaincu, j'ai revu mes attentes à la baisse et me suis concentré sur une zone de 800 kb entourant les duplications *ace-1*. J'ai isolé les gènes présents dans ces régions pour chacun des deux assemblages, récupéré les séquences protéiques correspondantes sur NCBI, et utilisé *Blastn* pour en identifier la position sur le génome opposé. J'ai réussi à trouver un total de 36 paires de gènes homologues, tous synténiques mais pas colinéaires : si les gènes identifiés sur le chromosome 2R d'*An. gambiae* sont bien retrouvés sur le chromosome 3 de *Cx. quinquefasciatus*, leur ordre n'est généralement pas conservé. Seuls les gènes *ace-1* et de la *ß-catenin like protein 1* sont partagés entre les zones dupliquées d'*An. gambiae s.l* et *Cx. pipiens s.l.*, le premier se situant en amont du second. Bien que préliminaires, je trouve ces résultats intéressants quand on considère la similarité des impacts phénotypiques des duplications *ace-1* dans les deux complexes d'espèces. La duplication de gènes différents mène à des impacts phénotypiques remarquablement similaires. Il est d'ailleurs "amusant" de voir que le gène que j'avais proposé comme piste potentielle pour expliquer le phénotype HS chez *Culex pipiens s.l.* (voir

chapitre I. Tab.I.2), codant pour la *ß-catenin like protein 1*, est aussi présent dans toutes les duplications *An. gambiae s.l.*, pour lesquelles on ne connaît pas d'allèles sublétaux[11].

II.3. Duplications homogènes

Chez *An. gambiae s.l.*, les duplications homogènes, comme toutes les duplications du gène *ace-1*, partagent une structure commune (Assogba *et al.,* 2016 ; Assogba *et al.,* 2018 ; **Claret *et al.* soumis**). Elles varient donc uniquement par le nombre de copies R associées, de 2 à 7 au moins en populations naturelles (Assogba et al. 2016), puisque tous les haplotypes R, *i.e.* la séquence *ace-1* porteuse de la mutation de résistance, sont identiques. Des analyses ont été réalisées sur des souches fixées pour un nombre de copies différent (R^3 *vs* R^5), à l'aide de bioassays (niveau de résistance) et en les mettant en compétition dans des cages à populations pendant une dizaine de générations sans exposition aux insecticides (mesure les désavantages sélectifs relatifs). Ces analyses ont montré que la résistance comme les désavantages sélectifs semblent augmenter avec le nombre de copies (Assogba *et al.,* 2015 & 2016).

Article 2 : "Evolutionary trade-offs associated with copy number variations in resistance alleles in *Culex pipiens* mosquitoes." Pascal Milesi, Jean-Loup Claret, Sandra Unal, Mylène Weill et Pierrick Labbé. 2022. Parasite and Vectors, 15:484.

Au début de ma thèse, on ne connaissait pas de duplication homogène du gène *ace-1*. J'ai ainsi participé à la description de l'allèle Cp-R_2^3 (**Milesi, Claret *et al.* 2022**), originellement échantillonné en Martinique et également retrouvé à Mayotte dans des populations *Cx. quinquefasciatus*, comme Cp-D_1 (Milesi *et al.,* 2018). Il a été fixé dans une souche à l'ISEM en suivant le même protocole de *backcrossing* que celui qui avait été utilisé pour les allèles Cp-D. Cela a permis de comparer l'impact de cet allèle en termes de *fitness* à la souche sensible Slab, mais aussi à l'allèle monocopie R_1 (lui aussi introgressé sur fond Slab, souche SR) qui est largement répandu sur tout le pourtour méditerranéen dans les populations *Cx. pipiens s.s.* Ces deux allèles diffèrent par la séquence de l'haplotype R et par leur nombre de copies: des qPCR ont en effet révélé une copie unique pour Cp-R_1 et trois copies pour Cp-R_2^3 (**Milesi, Claret *et al.* 2022**). Les capacités de résistance de ces deux allèles ont été testées par des bioassays dans lesquels votre dévoué serviteur a exposé des larves des différentes souches (Slab, Cp-R_1 et Cp-R_2^3) à des doses croissantes d'insecticides OP et CX : l'allèle

[11] L'amplification d'un même gène peut certainement avoir des impacts différents dans des espèces ayant divergé depuis si longtemps, mais je voulais noter l'ironie de la chose.

Cp-R_2^3 permet à son porteur de résister à de plus fortes concentration d'insecticide que Cp-R_1 (Cp-R_1/Cp-R_2^3 resistance ratio RR_{50} = 3.2 [2.54.2], χ^2 = 61, df = 1, p < 0.001 ; **Milesi, Claret *et al.* 2022**). Bien que la relation entre le nombre de copies R et l'activité enzymatique des allèles ne soit pas strictement additive, nous avons trouvé qu'elle était deux fois supérieure chez Cp-R_2^3 par rapport à Cp-R_1 (**Milesi, Claret *et al.* 2022**). Nous avons ensuite évalué l'impact relatif sur la *fitness* des deux allèles de résistance en les mettant directement en compétition dans des cages : il en ressort qu'en absence d'insecticide, l'allèle Cp-R_2^3 est plus délétère que l'allèle de résistance mono-copie (augmentation de la fréquence Cp-R_1 de 0.5 à 0.63 en 6 générations), et que les hétérozygotes présentent une fitness supérieure à celle des homozygotes (**Milesi, Claret *et al.* 2022**). On observe donc de fortes convergences entre Cp-R_2^3 (dans cet article, mais aussi dans les analyses de structure du Chap. I), et les duplications homogènes Ag-R^x. i) Les duplications homogènes sont dans les deux cas de larges structures de plusieurs kilobases (203 kb pour Ag-R^x et 411 kb pour Cp-R_2^3), avec des variations du nombre d'amplicons en populations naturelles (dans les mêmes populations pour *An. gambiae s.s.*, à notre connaissance dans des populations différentes pour *Cx. pipiens s.l.*) ; ii) L'augmentation du nombre copies R s'accompagne d'une capacité de résistance accrue, mais également de désavantages sélectifs plus lourds. Remarquablement, dans les deux cas, il existe un faisceau d'indices soutenant que l'augmentation du coût sélectif ne serait pas lié à la duplication des copies R de *ace-1*, mais plutôt aux gènes embarqués. En effet, l'activité enzymatique de l'AChE des porteurs de duplications homogènes augmente avec le nombre de copies, ce qui atténuerait l'effet délétère de la mutation G119S (pour rappel, elle induit une diminution ~60% de l'activité de l'AChE1, Alout et al. 2007).

En analysant les données du *PipPop Project*, j'ai probablement décrit au moins deux allèles Cp-R^x, représentés par les structures δ et γ. Si cela reste à confirmer formellement (une simple PCR suffirait, mais je n'ai pas encore les ADN…), il semblerait donc que plusieurs allèles R^x soient apparus indépendamment au sein du complexe *Cx. pipiens s.l.*, alors qu'une unique structure est retrouvée chez *An. gambiae s.l.* Là-encore il s'agit de larges structures (363 à 432 kb ; Tab.II.1) avec des variations du nombre d'amplicons en populations naturelles.

En ce qui concerne les duplications homogènes R^x, on observe donc une différence entre les deux complexes d'espèces : d'une part un unique haplotype R variant en nombres de copies mais une structure unique pour *An. gambiae s.l. ;* de l'autre plusieurs haplotypes, des nombres de copies variables et plusieurs structures (avec des aires de distribution contrastées) chez *Cx. pipiens s.l.* Cela signifie notamment que les impacts phénotypiques des duplications

sont sans doute plus différents entre allèles chez *Cx. pipiens s.l.*, puisque ce ne sont pas toujours les mêmes locus qui sont concernés par la duplication au sein de ce complexe d'espèces. Toutefois, il est probable que ces différences s'expliquent simplement par l'étendue et la structuration des aires de distributions des deux complexes d'espèces, continue et restreinte à l'Afrique pour le premier, et discontinue à l'échelle planétaire pour le second. Les réponses évolutives restent malgré tout similaires dans les deux complexes : les allèles R^x sont principalement sélectionnés dans des populations exposées à de fortes pressions insecticides, où la résistance est le critère déterminant.

II.4. Duplications hétérogènes

Dans des populations exposées à des doses plus modérées ou de façon intermittente/hétérogène spatialement, ce sont les hétérozygotes RS qui semblent représenter le meilleur compromis évolutif : une situation de superdominance, irréductible à cause du fardeau de ségrégation associé à l'hétérozygotie. Là encore, les deux complexes d'espèces semblent parvenir à des solutions convergentes pour contourner cette impasse, *via* la sélection d'allèles dupliqués hétérogènes, ou allèles D.

On recense plusieurs duplications hétérogènes dans chaque complexe. Elles partagent toutes une unique structure chez *An. gambiae s.l.*, identique à celle des allèles Ag-R^x, et comportent une seule copie S et une seule copie R (**Claret *et al.* soumis**). Le scénario le plus parcimonieux quant à l'origine des allèles Ag-D semble donc celui d'une unique recombinaison inégale générant un allèle Ag-R^2, suivi de recombinaisons secondaires entre un des amplicons et des allèles S simple-copie présents dans les populations, qui modifient l'haplotype S associé à la copie R (**Claret *et al.* soumis**). A l'inverse, les allèles Cp-D présentent à la fois des haplotypes et des structures différentes, en général différentes de celles des allèles Cp-R^x. Ces allèles varient dans le nombre de copies R et S qu'ils associent, bien que la plupart n'en comportent qu'une seule de chaque (Chap. I). Il est donc raisonnable de penser que plusieurs événements de recombinaisons inégales ont généré cette diversité. Par exemple, bien que retrouvés dans les mêmes populations autour de Montpellier, les allèles Cp-D_2 et Cp-D_3 présentent des structures différant de plusieurs kb en taille, avec des *breakpoints* différents (voir Chap I). De même, tous deux comportent le même haplotype R (allèle R_1), mais celui-ci n'existe qu'en simple-copie dans ces mêmes populations, et les haplotypes S portés par ces allèles sont très différents de R_1 (et différents entre eux; Labbé *et al.* 2007). Le scénario le plus probable dans ce cas est donc celui de deux événements de recombinaison inégale indépendants, chez deux hétérozygotes avec des allèles S différents.

L'allèle Cp-D_1 semble suggérer que plusieurs mécanismes sont à l'origine de la diversité de duplications hétérogènes observée au sein de *Cx. pipiens s.l.* Il est en effet le seul connu à ce jour avec trois copies, deux R et une S. Par ailleurs, comme indiqué plus haut, l'haplotype S ne diffère de l'haplotype R en deux copies que par la seule mutation de résistance: le fait que l'allèle Cp-R_2^3 retrouvé dans les même populations que Cp-D_1 comporte lui aussi les mêmes copies R, en trois exemplaires, suggère que l'allèle Cp-D_1 dérive probablement de l'allèle Cp-R_2^3 par réversion d'une des copies R en copie S (mutation S119G), plutôt que d'un événement de duplication indépendant (Labbé *et al.* 2007 et Milesi *et al.* 2018, **Milesi, Claret et al. 2022**). Pour les autres allèles Cp-D, les haplotypes S et R sont suffisamment différents pour exclure un tel processus (Milesi et al. 2018). On ne peut toutefois exclure qu'une partie de la diversité des allèles Cp-D puisse, comme pour *An. gambiae s.l.,* venir de recombinaisons secondaires si des allèles partagent la même structure (*e.g.* les structures δ et γ) et le même haplotype R (ce qui est difficile à établir avec les données dont je dispose, là-encore il faut disposer des haplotypes). Comme pour les duplications homogènes, on trouve donc une plus grande diversité de duplications hétérogènes au sein du complexe *Cx. pipiens s.l.*, à la fois en termes de structures, de nombres de copies et d'haplotypes *ace-1* embarqués. Il est fort probable aussi qu'il existe des recombinaisons secondaires chez *Cx. pipiens s.l.* similaires à celles observées en Côte-d'Ivoire pour *A. coluzzii*, mais hélas les données du *PipPop Project* ne permettent pas de tester cette hypothèse, il faudrait un échantillonnage plus important à des échelles plus localisées pour les détecter, et utiliser d'autres approches plus ciblées que le *whole genome sequencing* (voir **Claret *et al.,* soumis**).

Le plus probable est aussi qu'il existe une plus grande diversité de mécanismes moléculaires à l'origine de cette diversité chez *Cx. pipiens s.l.*, de même qu'on attend là-encore une plus grande variété d'impacts phénotypiques pour ces allèles Cp-D, puisque le nombre et la nature des locus embarqués dans la duplication est variable également. A l'inverse, les allèles Ag-D sont beaucoup plus homogènes, et donc leurs impacts phénotypiques sont sans doute plus proches (même si des indices populationnels nous amènent à penser qu'ils ne sont pas tous aussi fonctionnels, voir **Claret *et al.,* soumis**). Là-encore on peut penser que la moindre variation observée chez *A. gambiae s.l.* par rapport à *Cx. pipiens s.l.* pourrait n'être qu'un effet d'échelle/homogénéité au niveau des aires de répartition.

II.5. Réarrangements secondaires dans les duplications

Chez *A. gambiae s.s.*, une délétion supprimant tous les gènes embarqués à l'exception d'*ace-1* dans au moins un amplicon avait été documentée dans un allèle R^x (Assogba *et al.,* 2016). Elle serait sans doute adaptative parce qu'elle réduirait les effets des perturbations du *gene dosage*. Cette délétion augmentait en fréquence dans les populations naturelles exposées aux insecticides au Togo (de 2012 à 2017 ; Assogba *et al.,* 2018). J'en ai découvert une similaire dans la structure ε^+*, qui affecte tous les gènes embarqués à l'exception d'*ace-1*, du gène de l'HJHBP et d'un autre gène non identifié (Fig. II.2 et Tab. II.1), mais cette fois-ci pour l'ensemble des amplicons et non plus un seul. Cette délétion a été identifiée dans un unique individu ne portant que des copies R, ce qui ferait de ε^+* un allèle Cp-R^x. Il est donc probable que des contraintes similaires existent pour les allèles R^x dans les deux complexes d'espèces, avec encore une fois la sélection de réponses similaires.

II.6. Take home message

Les complexes *Cx. pipiens s.l.* et *An. gambiae s.l.* constituent un excellent exemple d'évolution parallèle: on y retrouve les mêmes mutations de résistance, les mêmes types de duplications, homogènes et hétérogènes, qui ségrègent dans les mêmes populations, jusqu'aux mêmes réarrangements secondaires (dont le caractère adaptatif reste encore à prouver chez *Culex*). Il existe toutefois des différences notables : que ce soit en nombre de copies, de structures et d'haplotypes *ace-1* embarqués, ou encore dans la diversité des mécanismes ayant permis leur émergences, les variations semblent plus importantes chez *Cx. pipiens s.l.*. Ce constat reste toutefois à relativiser au regard des différences de taille et de d'homogénéité des aires de répartition des deux complexes : on retrouve en effet beaucoup moins de variation dans les populations de *Cx. pipiens s.l.* à une échelle plus réduite (échelle continentale par exemple), comparable alors avec celle de *An. gambiae s.l.* Il semble donc que pour la résistance aux insecticides OP/CX, si l'apparition des mutations restent aléatoires, après quelques années, les allèles sélectionnés sont très similaires pour les deux complexes : si l'histoire ne se répète pas, la sélection semble bien la faire converger vers les rares solutions adaptées.

Table II.2. Résumé de l'évolution parallèle au locus *ace-1* d'*An. gambiae s.l.* et *Cx. pipiens s.l.*. En noir ce qui était connu avant ma thèse, en vert ce que j'ai apporté, y compris en consolidant des hypothèses antérieures ou en étendant des propriétés connues à d'autres allèles (*).

Caractéristiques	*Cx. pipiens s.l.*	*An. gambiae s.l.*
Mutations ponctuelles	- *ace-1* G119S - apparue indépendamment dans plusieurs taxons du complexe (*Cx. pipiens* et *Cx. quinquefasciatus*) - *ace-1* F290V (Mediterrannée, rare)	- *ace-1* G119S - passée entre taxons du complexe par introgression (*An. coluzzi* vers *An. gambiae s.s.*)
Duplications homogènes	**Cp-R^x** - 4 allèles décrits - structures différentes - nombres de copies variables - résistance et désavantages sélectifs croissants avec le nombre de copies	**Ag-R^x** - 1 allèle décrit - structure unique - nombres de copies variables - résistance et désavantages sélectifs croissants avec le nombre de copies
Duplications hétérogènes	**Cp-D** - >27 allèles décrits - au moins 6 structures différentes (certaines identiques, potentiels R^x?) - locus embarqués (hors *ace-1*) différents - 2 (R+S) à 3 (2R+S) copies -* probablement plusieurs événements de duplication indépendants, mais aussi des réversions d'allèles R^x vers D - phénotype similaire à un hétérozygote standard - certains allèles délétères à l'état homozygote, d'autres non	**Ag-D** - 9 allèles décrits -structure unique (identique R^x)* - locus embarqués (hors *ace-1*) identiques - 2 (R+S) copies* - probablement un unique événement de duplication suivi de recombinaisons secondaires - phénotype similaire à un hétérozygote standard - 1 allèle non délétère à l'état homozygote, statuts inconnus pour les autres
Réarrangement secondaires	- délétion au sein de structure probablement R^x (reste *ace-1*, HJHBP et un gène non-identifié), effets sur la fitness inconnus	- délétion au sein d'allèles R^x (reste *ace-1*), probablement adaptative

Parasites & Vectors

RESEARCH Open Access

Evolutionary trade-offs associated with copy number variations in resistance alleles in *Culex pipiens* mosquitoes

Pascal Milesi[1,2*], Jean-Loup Claret[3], Sandra Unal[3], Mylène Weill[3] and Pierrick Labbé[3,4]

Abstract

Organophosphate and carbamate insecticides have largely been used worldwide to control mosquito populations. As a response, the same amino acid substitution in the *ace-1* gene (G119S), conferring resistance to both insecticides, has been selected independently in many mosquito species. In *Anopheles gambiae*, it has recently been shown that the G119S mutation is actually part of homogeneous duplications that associate multiple resistance copies of the *ace-1* gene. In this study, we showed that duplications of resistance copies of the *ace-1* gene also exist in the *Culex pipiens* species complex. The number of copies is variable, and different numbers of copies are associated with different phenotypic trade-offs: we used a combination of bioassays and competition in population cages to show that having more resistance copies conferred higher resistance levels, but was also associated with higher selective disadvantage (or cost) in the absence of insecticide. These results further show the versatility of the genetic architecture of resistance to organophosphate and carbamate insecticides around the *ace-1* locus and its role in fine-tuned adaptation to insecticide treatment variations.

Background

Whether for sanitary or economic (agriculture, tourism) reasons, pest and vector species have been the target of intense xenobiotic exposure to control their populations. As a response, resistances have been selected for and have spread worldwide in many diverse organisms [1–3]. Resistance has been the focus of many studies with a management perspective, in particular in vector species where long-term monitoring and resistance evolution surveys took place, for instance in mosquitoes ([3] and references therein). However, resistance to insecticides also provides a wealth of information for evolutionary biologists as it is an iconic example of rapid adaptation to a new environment.

Mutations conferring resistance to insecticides are adaptive in the sense that they provide greater fitness in presence of insecticide. However, in mosquitoes, most of the mutations conferring resistance described so far are also disadvantageous in the insecticide-free environment [4–9]. For instance, the mutated allele could encode for a protein that is less efficient than the susceptible one, or metabolic equilibria could be dysregulated. Both can result in strong deleterious effects [3], affecting many different life history traits (hereafter termed 'selective disadvantage;' see useful criticism of use of the term 'cost' for these selective disadvantages by Lenormand et al. [10]). Because both advantages and disadvantages can vary between resistance mutations, they convey different evolutionary trade-offs in different environments (e.g. [11, 12], in particular regarding insecticide treatment intensities [5, 13]). For instance, intermediate treatment intensities would favor heterogeneous duplications over single-copy resistance alleles, despite the fact that

*Correspondence: pascal.milesi@scilifelab.uu.se

[1] Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen, 18D, SE-752 36, Uppsala, Sweden
Full list of author information is available at the end of the article

they confer a lower resistance level, because they are also associated with a much lower fitness cost. However, heterogeneous duplications are outcompeted by single-copy resistance alleles when selective pressure is high because of a too low resistance level [5, 7]. In a constant environment, and as time passes, new alleles, with a more favorable evolutionary trade-off, and/or compensatory mutations are expected to be selected for (e.g. [14] in *Lucilia cuprina*, [11, 15–17] in *Culex pipiens* species complex and [18] in *Anopheles gambiae*). These contemporary evolutions of insecticide resistance are thus iconic examples of adaptive trajectories ("adaptive walk" in Orr [19]). The resistance alleles selected along these adaptive walks can result from simple nucleotide substitutions (e.g. affecting the target of insecticides), but various genetic architectures can also be selected for, for instance homogeneous duplications (aka gene amplification) or heterogeneous duplications [3, 20]. Different genetic architectures of resistance can be selected for precisely because they are associated with different evolutionary trade-offs [5, 21, 22]. Finally, the variation of treatment intensities in time and/or space can also generate balancing selection patterns that could (i) allow for the maintenance of susceptible and resistance alleles polymorphism in natural populations, (ii) select for alleles with more generalist trade-offs (e.g. [9, 10, 16]), but also (iii) maintain resistance allele polymorphism [23].

Among the most commonly used families of insecticides worldwide, organophosphate (OPs) and carbamate (CXs) insecticides target acetylcholinesterase (AChE1), encoded by the *ace-1* locus, which terminates cholinergic neurotransmission by hydrolysis of acetylcholine (ACh). These insecticides bind to AChE1, thereby impeding ACh degradation and inducing death by tetany. Point mutations modifying the conformation of the AChE1 active site and preventing the binding of the insecticide have been selected for in many vector and pest species. In particular, the same amino acid substitution (G119S) has been repetitively and independently selected in many mosquito species [24, 25]. While conferring resistance to OPs and CXs, the G119S mutation has also been shown to decrease the activity of the resistant acetylcholinesterase (AChE1R) by about 60% compared with the wild-type protein (AChE1S) in both the West Nile virus vector *Culex pipiens* sensu lato and the malaria vector *An. gambiae* [26, 27]. In both species, this drastic reduction in affinity with its natural subtract probably explains a large part of the selective disadvantage endured by resistant mosquitoes in absence of insecticides compared to susceptible ones [4, 7, 12, 28, 29].

In recent studies of *An. gambiae*, all alleles carrying the G119S mutation (R alleles) have been found to be part of homogeneous duplications (several resistance copies in tandem, R^x alleles) or of heterogeneous duplications (pairing a susceptible and a resistance copy, D alleles), i.e. they were never found in the natural population in a single-copy state [18, 22, 30].

In the *C. pipiens* species complex, two different R alleles have been found widely spread across natural populations [23]: R_1 is found in *C. pipiens* sensu stricto all over Europe and the Mediterranean area, and R_2 is found worldwide in *Culex quinquefasciatus*. They are also found in many D alleles associated with local susceptible variants [23]. However, the possibility that some R alleles could, as in *An. gambiae*, actually be part of homogeneous duplications, as well as what phenotypic effects these different genomic architectures could induce, has not yet been investigated in *C. pipiens* species complex.

In the present study, we isolated the R_1 and R_2 alleles in laboratory strains sharing the genetic background of the susceptible reference strain, SLAB. We first showed that, while R_1 is found in a single-copy state, R_2 is part of a homogeneous duplication carrying three R copies. We then investigated the phenotypes conferred by these different alleles (protein activity, resistance level and dynamics in absence of insecticides) and showed that different evolutionary trade-offs are associated with the different genomic architectures. We finally discuss the implication of the present study from both evolutionary biology and more applied perspectives.

Materials and methods

Mosquito strains

Three mosquito laboratory strains were used in this study: SLAB [31], SR [32] and SRQ (this study). SLAB is fixed for a single-copy susceptible allele (S_{SLAB}, isolated in California, *C. quinquefasciatus*). SR is fixed for R_1 [24], a resistance allele isolated from Southern France and found in *C. pipiens* s.s. all over Europe and around the Mediterranean Sea [23, 33]. SRQ is fixed for R_2 [24], a resistance allele found worldwide in *C. quinquefasciatus* [23] and isolated from a population from Martinique Island. The two resistance alleles were introgressed into the genetic background of the SLAB strain through at least 15 rounds of back-crossing. All strains thus share the same genetic background (>99%) and differ from each other almost only in their *ace-1* locus (although recombination around the *ace-1* gene is not complete, most of the background effects would be eliminated).

All strains were regularly checked for contamination: DNA was extracted from pools of first-instar larvae (~200 individuals per pool) and molecular tests, specific for each *ace-1* allele (detailed below), were used to check the homogeneity of each strain.

Genotyping

The various strains can be easily distinguished using a single PCR and different restriction fragment length polymorphisms (RFLPs). After DNA extraction, following the protocol in [34], a ~ 600-bp fragment of the *ace-1* gene, including intron 2 and most of exon 3 (with the resistance G119S mutation), was amplified using two generalist primers, Intron2dir1 and CpEx3rev, according to [35].

Susceptible vs. resistant

The G119S mutation creates an *AluI* restriction site [33] so that three genotypes can be distinguished (*AluI* RFLP test): susceptible homozygote (SS; one fragment, 597 bp), resistant homozygote (RR; two fragments, 496 and 101 bp) and heterozygote (RS; three fragments, 597, 496 and 101 bp); 5 µl of the PCR product was incubated for 2 h at 37 °C.

R_1 *vs.* R_2

The two different resistance alleles can be further distinguished by taking advantage of another single-nucleotide polymorphism (SNP) between R_1 and R_2 creating a *Bfa1* restriction site in R_2 (Additional File 1). This second RFLP test (*Bfa1* RFLP test) distinguishes three genotypes, the homozygotes R_1R_1 (one fragment, 597 bp) and R_2R_2 (three fragments, 73, 132 and 392 bp) and the heterozygotes R_1R_2 (four fragments, 597, 73, 132 and 392 bp); 5 µl of the PCR product was incubated for 2 h at 37 °C.

Gene copy number quantification

ace-1 gene copy number was estimated for ten individuals of each resistant strain using quantitative real-time PCR (qRT-PCR). Two individuals from the SLAB-susceptible strain were also used as controls. After DNA extraction, we dispensed 250 ng of genomic DNA and 1.5 µl of reaction mixture containing specific primers, each at a concentration of 0.8 µM and 0.75 µl of Master Mix (LightCycler 480 SYBR Green I Master, Roche), into the wells of a 384-well plate with a Labcyte Echo525 dispenser. We performed qPCR as follows: activation at 95 °C for 8 min followed by 45 cycles of 95 °C for 4 s, 67 °C for 13 s and 72 °C for 19 s. Melting curves were generated by a post-amplification melting step between 70 °C and 95 °C for Tm analysis. All quantifications were replicated three times for each DNA template. Two loci were amplified for each individual: *ace-1* (primers: 'Culexace1univdir3' AGA AGG TGG ACG CAT GGA TG; 'Culexace1univrev3' ATC TGG ACG CAG GAG TTG G) and *ace-2*, a locus known to be in a single copy in these species (primers: 'acequantidir' GCA GCA CCA GTC CAA GG; 'acequantirev' CTT CAC GGC CGT TCA AGT AG) [36]. *ace-1* over *ace-2* copy-number ratios were determined by the advanced quantification method (LightCycler 480 software v.1.5.0). Standard reference curves were constructed with tenfold dilutions of a PCR product previously amplified with specific primers for each locus from SLAB DNA.

Phenotyping

Protein activity

We measured acetylcholinesterase (AChE1) activity for 48 individuals of each resistance strain, using spectrophotometry [37]. Adult mosquitoes were decapitated, and each head was individually homogenized in 400 µl of a phosphate buffer (0.25 M, pH7) supplemented with 1% Triton X-100. Homogenates were centrifuged (9.3 g for 3 min), and 100 µl of the supernatant was dispensed into each of two wells of a 96-well microtitration plate. We added 10 µl of propoxur, a carbamate insecticide, at 10^{-3} M and 10^{-1} M (diluted in ethanol) into the first and second well, respectively. The plate was incubated for 15 min at room temperature. We then added 100 µl of substrate solution (25 mM sodium phosphate, pH 7.0, 0.2 mM DTNB, 0.35 mM sodium bicarbonate, 2.5 mM acetylthiocholine) to each well. AChE1 activity was estimated by measuring the change in optical density following the cleavage of acetylthiocholine, as described by [38]. Optical density at 412 nm was recorded every minute for 15 min with an EL 800 microplate reader (Bio-Tek Instruments, Inc.). The mean slope of each reaction was calculated with KCjunior v1.41.4 analysis software (Bio-Tek Instruments, Inc.) and was used as a measurement of AChE1 activity in each well. Individual AChE1 activity was computed as the average activity between the two wells. To avoid any block or sex confounding effects, individuals from both sexes and the two strains were evenly distributed in the plates.

Resistance level and bioassays

We used bioassays to assess the three strains' resistance to an OP insecticide, temephos (PESTANAL®,96% purity). We incubated 20 late third-instar larvae for 24 h at 27 °C ± 2 °C in plastic cups containing 99 ml of distilled water to which we added 1 ml of insecticide solution at the required concentration (1 ml of ethanol in controls). Four replicates were performed for each concentration (from 0 to 0.07 $g\mu.ml^{-1}$ see Additional File 2 for the complete dataset). Larval mortality was recorded after 24 h of exposure. We used the *BioRssay* R package (v.1.0.0 [39], https://CRAN.R-project.org/package=BioRssay) to analyze the dose-mortality responses of the different *ace-1* alleles and calculate the LD_{50} of the different strains, i.e. the lethal dose for 50% of the sample.

Experimental evolution in population cages

Population cages were used to set up a competition experiment between the two resistance alleles in absence of insecticides. R_1R_1 and R_2R_2 individuals were crossed, and the resulting F1 (100% R_1R_2 individuals) was reared until adulthood under standard conditions (25 °C, > 60% humidity, 12:12 h light:dark). Adults were released into a new cage to mate freely and reproduce. Their offspring were raised and released in new cages to ensure discrete generations. The process was repeated 11 times (i.e. 11 generations) with three independent cages (i.e. replicates). Almost each generation, and for each cage, about a hundred second-instar larvae were genotyped using the *Bfa1* RFLP test (see above) to measure the frequency of each genotype (R_1R_1, R_1R_2, R_2R_2). Allelic frequencies were then computed from genotypic frequencies.

We estimated the relative fitness of the various genotypes (R_1R_1, R_1R_2 and R_2R_2) using a deterministic genetic model (reproduction and selection, 11 cycles, no drift). The model was adjusted to the data and optimized using a maximum-likelihood approach as in Milesi et al. [5, 23]. For the reproduction step, the frequency of each genotype in the larvae of generation i was computed from the allelic frequencies (p) in the gametes of the previous generation, assuming panmixia (Eq. 1):

$$\begin{aligned} f(R_1R_1) &= p_1^2 \\ f(R_1R_2) &= 2 \times p_1 \times p_2 \\ f(R_2R_2) &= p_2^2 \end{aligned} \quad (1)$$

For each genotype g selection was then computed between larval and adult stages of generation i using the following genotype fitness: $w_{R1R1} = 1$, $w_{R1R2} = 1 + h.s$ and $w_{R2R2} = 1 + s$, with h the dominance coefficient and s the selection coefficient, both varying between -1 and 1 (Eq. 2):

$$f'_{gi} = \frac{f_{gi} \times W_g}{\sum (f_{gi} \times W_g)} \quad (2)$$

The genotypic frequencies after selection were used to calculate the allelic frequencies in the gametes produced by the surviving adults (Eq. 3).

$$\begin{aligned} p_1^{'} &= f(R_1R_1) + \frac{f(R_1R_2)}{2} \\ p_2^{'} &= (1 - p_1^{'}) \end{aligned} \quad (3)$$

The first run of 100,000 simulations was used to explore the parameter space and provide the likelihood profile associated with different random pairs of h and s values (Eq. 4):

$$L = \sum_g \sum_i \left(n_{gi} \times \ln \left(f_{gi} \right) \right) \quad (4)$$

where n is the number of individuals of genotype g observed in the cages at generation i, and f is the frequency of the genotype g at generation i calculated for a given pair of h and s values using our deterministic genetic model. One million additional simulations were run with parameter ranges more limited around the maximal likelihood h and s pair to precisely estimate the coefficients and their support limits (rough equivalents to 95% confidence intervals), defined as h and s maximal and minimal values, resulting in a likelihood equal to the maximum likelihood minus 1.96, as in [5].

Statistical analyses

All the statistical analyses were conducted using the R software (R Core Team, https://www.r-project.org/):

We used the following linear model to compare *ace-1* copy number between the various strains:

$$\gamma_{ij} = \mu + \alpha_j + \varepsilon_{ij} \quad \text{(mod.1)}$$

with γ_{ij} the number of copies of the *ace-1* gene in replicate i from strain j, μ the population mean, α is the fixed effect of strain j (SLAB, SR or SRQ) and ε_{ij} the error term following a normal distribution $\mathcal{N}(0, 1)$.

We used the following linear model to test the significance of the difference in AChE1 activity between the SR and SRQ strains:

$$\gamma_{ijkl} = \mu + \alpha_j + \beta_k + \delta_l + \varepsilon_{ijkl} \quad \text{(mod.2)}$$

with γ_{ijkl} the AChE1 activity for individual i of strain j and sex k measured in plate l, μ the population mean and α the fixed effect of strain j (SR or SRQ). β and δ are control for the fixed effects of sex k and plate l, respectively, and ε_{ijkl} is the error term following a normal distribution $\mathcal{N}(0, 1)$.

For both models, the significance of the various terms was tested using likelihood ratio tests (LRTs) comparing the full model with a model without the tested effect (*'anova'* function, R [40]). For both models, we also confirmed the absence of significant heteroskedasticity (*'bptest'* function, *'lmtest'* R package [41]) and that the models' residuals followed a normal distribution (*'shapiro.test'* function, *'stats'* R package).

Finally, we used binomial proportion tests (*'prop.test'* function, *'stats'* R package) to assess whether the allele frequencies at the end of the experimental evolution in cages (i.e. after 11 generations) differed from initial frequencies of 0.5 (100% R_1R_2 individuals).

Results and discussion

The goal of the present study was to investigate the potential existence of homogeneous duplications of the *ace-1* locus in the *C. pipiens* species complex, similar to those found in *An. gambiae*, to assess the phenotypic effects of different genetic architectures and their role in adaptation to insecticides.

Higher ace-1 copy number partly restores protein activity levels

We first quantified the number of *ace-1* copies in the three different strains, with the susceptible reference strain SLAB as a control. SLAB was found carrying a single-copy allele (mean = 1 ± 0.007 SD), and this was also the case for the resistant strain SR, carrying the R_1 allele (mean = 1.03 ± 0.03 SD; mod. 1, $t = 0.34$, $p = 0.74$). However, we detected three copies of the *ace-1* locus in the SRQ strain (mean = 3 ± 0.14 SD), indicating that the R_2 allele is a homogeneous duplication (R^x), i.e. R_2^3 allele (mod. 1, $t = 25.8$, $p < 0.001$, Fig. 1A). While several R^x alleles (aka homogeneous duplications) have recently been described in *An. gambiae* populations, with two to nine *ace-1* copies [22, 30], this study is the first to report homogeneous duplications of the *ace-1* resistance allele in mosquitoes from the *C. pipiens* species complex.

We then investigated whether having more *ace-1* resistance copies would lead to higher activity of the resistant acetylcholinesterase (AChE1R, encoded by the R alleles). The activity A_R of the resistant acetylcholinesterase was significantly higher in the SRQ strain ($A_{R2} = 10.8 \pm 1.4$, three copies) than in the SR strain ($A_{R1} = 6.3 \pm 1.2$; mod. 2, LRT: $F = 269$, $df = 1$, $p < 0.001$, Fig. 1B, Additional file 3: Table S1).

As in *An. gambiae* [22], it thus clearly appears that having a higher resistance copy number does increase the AChE1 protein activity. However, the protein activity did not increase in a strictly additive way with the number of copies of *ace-1*; R_2^3 activity is 1.67 times higher than that of R_1, not three times as was expected. In previous studies in *An. gambiae* and *C. pipiens* species complex conducted on heterogeneous duplications, D, A_R was indeed found roughly proportional to the number of R copies [7, 21]. Here, the R_1 and R_2^3 alleles differ not only by their copy number but also in their *ace-1* sequences (Additional file 1). As all R_2^3 copies are identical (no variation over 3 kb of the *ace-1* sequence in [23]), the departure from additivity observed could thus be explained by a lower per-copy activity for the proteins encoded by the *C. quinquefasciatus* R_2^3 allele compared to the protein encoded by the *C. pipiens* s.s. R_1 allele. Alternatively, the expression of the *ace-1* gene in the SRQ strain could be somehow regulated. Finding more variation in copy number for the R_2 allele, if any, would help settle this issue. For instance, in *An. gambiae*, for the homogeneous duplication R^x, the relation between the number of resistance copies and ACHE1R activity is not strictly additive, even though all copies are identical [22], strongly suggesting

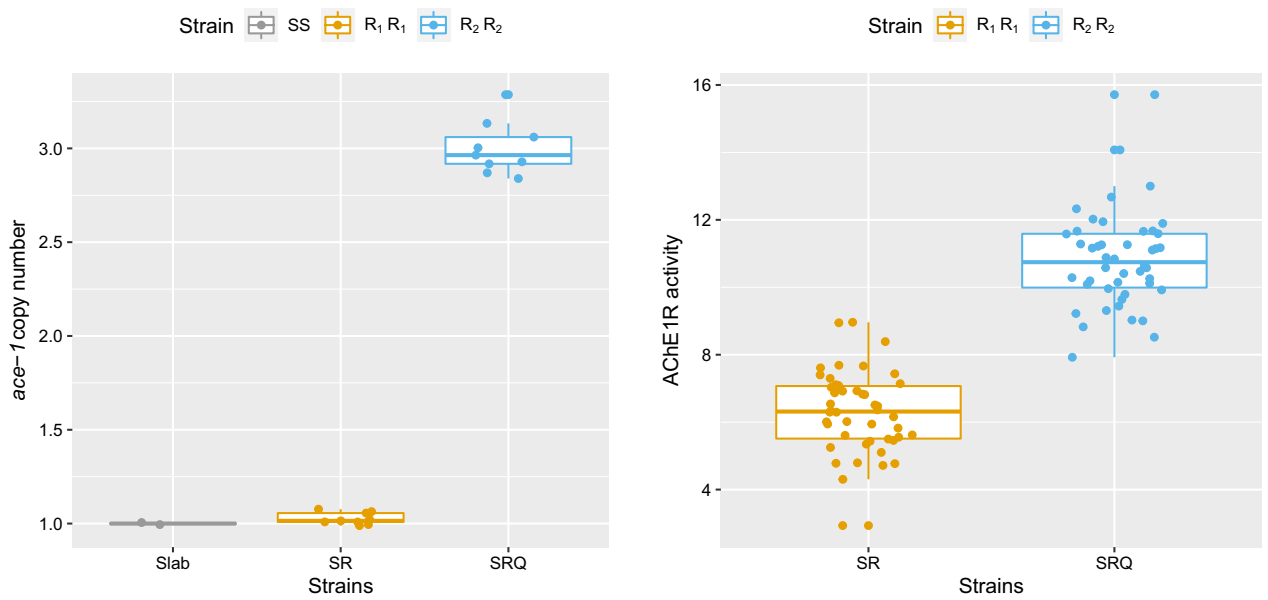


Fig. 1 *ace-1* copy number and resistant acetylcholinesterase activity. Left panel is the number of copies of *ace-1* gene estimated using quantitative PCR for the susceptible strain, SLAB and the resistance strains SR and SRQ. Right panel shows the activity of the resistant acetylcholinesterase (AChE1R) estimated through spectrophotometry for the resistant strains SR (R_1 allele, orange) and SRQ (R_2^3 allele, blue), respectively

that *ace-1* resistance copy expression is further regulated. Both hypotheses are not exclusive, and further studies are required to identify if and when regulation happens.

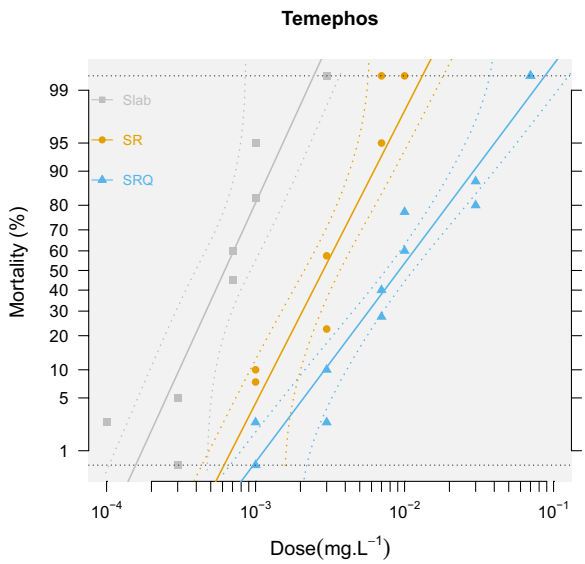


Fig. 2 Bioassay analysis. Probit-transformed mortality rates as a function of the log-dose of insecticide [temephos (R), an OP insecticide]. Colors and point shapes indicate the different strains. The fit from the linear model is provided (solid line) along with the 95% confidence intervals (dotted lines). The linearity of the log-responses has been checked using *test.validity* = *T* in the "*mort.plot*" function from the "*BioRssay*" R package

Alternative genetic architectures confer different evolutionary trade-offs

We then compared the phenotypic consequences of copy number variation at the *ace-1* locus and first quantified the resistance levels associated with the different genotypes. As expected, both R_1 and R_2^3 alleles confer insecticide resistance compared with the susceptible strain (RR_{50} = 4.6 [3.7–5.7], 95% CI and 15 [12–18], respectively), but more importantly, the duplicated allele conferred a significantly higher resistance level than the single copy allele (R_1/R_2^3 resistance ratio RR_{50} = 3.2 [2.5–4.2], X_2 = 61, df = 1, $p < 0.001$; Fig. 2 and Additional file 3: Table S2). We then addressed the selective disadvantages associated with the resistance alleles. Despite its obvious advantage in presence of insecticides, the SR strain was indeed repeatedly shown in previous studies to incur a strong selective disadvantage compared to SLAB in their absence (e.g. [28]). Rather than comparing both resistant strains to the susceptible one, we thus chose to directly assess whether R_2^3 incurred a stronger or lesser disadvantage than R_1 in absence of insecticide through a competition experiment in population cages: this set-up allows an integrative assessment of their relative fitness over the full life cycle and ensures that genetic background effects associated to each strain (e.g. resulting from their fixation process) are strongly reduced, as the alleles are mixed in the individuals of each generation [5].

After 11 generations of direct competition, the frequency of the R_1 allele rose significantly, from 0.5 to ~ 0.63 in all three replicates, out-competing the R_2^3 duplicated allele (binomial test, all $p < 0.004$, Fig. 3A). To quantitatively estimate the fitness of the different genotypes, we then adjusted a model of reproduction-selection to the

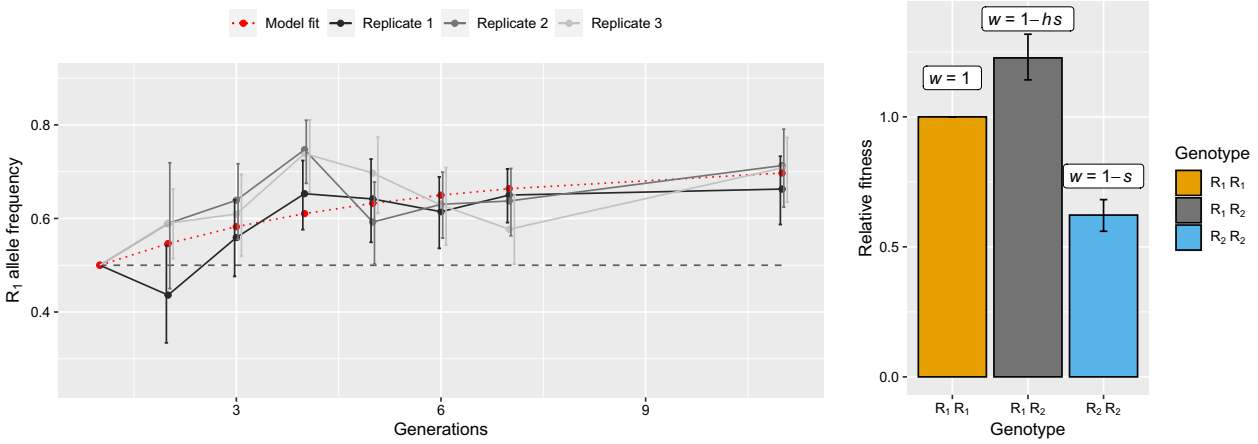


Fig. 3 Dynamics of the resistance alleles and selection coefficients estimations. Left panel represents the dynamics of the R_1 allele (SR strain) frequency (± SD) over generations, in competition with the R_2 allele (SRQ strain) in absence of insecticide (the dots of the 3 replicates are slightly offset for easier reading). The R_1 allele dynamics predicted by the maximum likelihood fit of the reproduction-selection model is in red. Right panel represents the relative fitnesses along with their corresponding support limits (≈ 95% confidence intervals) of the different genotypes (R_1R_1 was used as the reference, hence fixed to 1), estimated through a maximum likelihood approach for the parameters' values in the reproduction-selection model (see methods)

temporal genotypic data: we found that the heterozygous genotype (R_1/R_2^3) conferred the highest fitness ($w_{R1R2}=1.23$ [1.14–1.21] support limits) and that the R_1/R_1 genotype ($w_{R1R1}=1$) had much higher fitness than R_2^3/R_2^3 (0.62 [0.56–0.68], Fig. 3 and Additional file 3: Fig. S1).

The strong fitness reduction incurred by resistant mosquitoes is thought to be associated with multiple pleiotropic deleterious effects, affecting many different life history traits, because of the reduced activity of the AChE1R [4, 7, 12, 28, 29, 42]. Accordingly, D alleles are thought to be selected because they reduce these deleterious effects by pairing a resistance copy (R, low AChE1 activity) and a susceptible copy (S, high AChE1 activity) in a heterogeneous duplication, thereby partly restoring the AChE1 activity to levels closer to those of susceptible alleles [21, 22]. However, the case of the homogeneous duplication R^x is less clear: we show in the present study that despite a higher global activity for R_2^3 compared to R_1, the former allele induces higher selective disadvantages (Fig. 3), which is also what was observed for R^5 *vs* R^3 alleles in *An. gambiae* [22].

The less-than-strictly-additive AChE1R activity for the R_2^3 allele suggests that some deleterious effects could be associated with specific resistance alleles, potentially resulting from background deleterious mutations in the gene or its vicinity in the haplotype where the resistance mutation occurred. The fact that R_2^3/R_1 heterozygotes appear to incur a higher fitness than both homozygotes supports this hypothesis (Fig. 3): if both alleles are weighted by linked deleterious mutations, they can complement each other, i.e. if they are different between the two alleles, the heterozygote would incur a higher fitness (a similar explanation has been proposed for the complementation of strongly deleterious D alleles in *C. pipiens* s.l. [17, 23].). However, the overall activity remains higher for this allele compared to R_1 (Fig. 1), and in *An. gambiae* it is the same *ace-1* sequence that is present in five or three copies [22]. It thus suggests that the architecture itself, i.e. the mere fact of carrying more copies, induces selective disadvantages. This structural "cost" could result from deleterious mutations trapped into the amplicons, from the breakpoints of the duplication being located in functional regions or from dosage imbalance for other genes embedded in the duplicated alleles that might disrupt biochemical equilibrium, as previously proposed [22, 23, 35]. Though none of these hypotheses are exclusive, the latter has been favored in the case of *An. gambiae* duplications: in this species, a ~ 200-kb amplicon encompassing 11 genes in addition to *ace-1* has been described, and a variant with a deletion of these other genes appears to be favored by selection in natural populations, probably because the deletion restores the gene balances [18]. Note that deleterious mutations in these closely linked genes could also explain the higher fitness of R_2^3/R_1 heterozygotes (Fig. 3). There is thus a strong incentive to characterize the genomic structure of the *ace-1* duplications (either heterogeneous or homogeneous) in *C. pipiens* species complex too.

To summarize, the two *ace-1* R alleles present different evolutionary trade-offs: while having a higher copy number of resistance allele confers a higher resistance level, and thus higher selective advantage in presence of OP and CX insecticides, it is also associated with higher selective disadvantages, revealed in absence of insecticide. Although the mutations occurred independently in the different species, the same relationships among R copy number, resistance level and selective disadvantages have been described in *An. gambiae* [22]. Similarly, the R_2^3 duplicated allele would likely tend to be selected for in areas of intense selective pressure, its higher resistance surpassing its higher disadvantages, while the R_1 single-copy allele would be favored in areas with more moderate intensity of treatment. This can reflect the ecology of the mosquito populations where these alleles were found: in the tropical areas where R_2^3 was found, *C. quinquefasciatus* is the year-long vector of several viruses and thus probably subjected to more intense and regular treatments than in the Mediterranean area where R_1 is found and where *C. pipiens* s.s. is less a vector than a summer nuisance (the female diapauses in winter). The genotyping of natural populations to look specifically for the presence of R^x homogeneous duplications of the *ace-1* locus could confirm this hypothesis. It would also allow us to understand whether the number of copies is as variable as in *An. gambiae* (at least up to 6 copies [22]), whether the R^x alleles are only found in *C. quinquefasciatus* or are also found in *C. pipiens* s.s. and, if so, if they are found in populations experiencing higher treatment intensities. Finally, the recurrent selection of homogeneous duplications of the *ace-1* resistance copies in phylogenetically distant species complexes (e.g. in *Anopheles* [18, 22, 30] and in *Culex*, this study), along with the high diversity of heterogeneous duplications already described in both species [7, 17, 23, 30, 35, 43–45], provides further support for a very high duplication rate of the *ace-1* loci. It also highlights the versatility of adaptive responses that can result from such structural variants (i.e. as opposed to simple SNP): from a more quantitative resistance advantage resulting, at least in part, from the increased amount of protein produced for the homogeneous duplications (as also seen for metabolic resistances like esterases or P450 monooxygenases [46]) to a more qualitative advantage for the heterogeneous duplications that allow the fixation of a heterozygote advantage selected in more variable environments [5, 23]. Note however that

the recurrent selection of architectures such as homogeneous duplications in distant lineages calls for more functional research to understand how producing more AChE1R proteins leads to higher resistance levels.

Conclusion

In *C. pipiens* species complex many different genetic architectures encompassing the *ace-1* locus exist for resistance to OPs and CXs insecticides, which are each associated with a different evolutionary trade-off: the single-copy resistance allele provides resistance but is associated with a high selective disadvantage in absence of insecticides, while homogeneous duplications provide even higher resistance levels but are associated with higher selective disadvantages. Not only different genetic architectures could represent various steps along an adaptive walk, but there also are many ways to answer to the various intensities of selective pressure. While inspiring from an evolutionary perspective, the vector management view is clearly worrying, as this 'toolbox' allows mosquito populations to finely and quickly adjust to local treatment strategies in natural populations (particularly if one considers the various heterozygous combinations between the different alleles), which definitely represents a hindrance to vector control policies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13071-022-05599-8.

Additional file 1: Sequence alignment for R1 (SR) and R2 (SRQ). The primers used for the PCR amplification are indicated (light gray) as well as the restriction sites for the *BfaI* and *AluI* enzymes (darker gray, the triangles indicate restriction cuts). Mutations between the two sequences are in bold.

Additional file 2. This file basically contains all the raw data supporting each analysis.

Additional file 3: Table S1. Analysis of variance of model 1. **Table S2.** Bioassay analyses. **Figure S1.** Reproduction-selection model likelihood profiles.

Acknowledgements
We thank Patrick Makoundou for his technical help. The computations were enabled by resources in project SNIC 2022/22-23 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

Author contributions
PL, PM and MW designed the study. JLC and SU performed the experiments. JC and PM analyzed the data. PM wrote the first draft of the manuscript and received input from all authors. All authors read and approved the final manuscript.

Funding
Open access funding provided by Uppsala University. This work was funded by the Agence Nationale de la Recherche (ANR) ArchR project (ANR-20-CE34-0007) and PL's grant from the Institut Universitaire de France (IUF).

Availability of data and materials
All data generated or analyzed during this study are included in this published article and its supplementary information files: "Additional file 2."

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen, 18D, SE-752 36, Uppsala, Sweden. [2]Science for Life Laboratory (SciLifeLab), Uppsala, Sweden. [3]Institut Des Sciences de L'Évolution de Montpellier (UMR 5554, CNRS-UM-IRD- EPHE), Université de Montpellier, Cedex 05, Montpellier, France. [4]Institut Universitaire de France, 1 Rue Descartes Cedex 05, 75231 Paris, France.

Received: 11 September 2022 Accepted: 30 November 2022
Published online: 22 December 2022

References

1. Georghiou GP. The evolution of resistance to pesticides. Annu Rev Ecol Syst. 1972. https://doi.org/10.1146/annurev.es.03.110172.001025.
2. Coleman M, Hemingway J, Gleave KA, Wiebe A, Gething PW, Moyes CL. Developing global maps of insecticide resistance risk to improve vector control. Malar J. 2017. https://doi.org/10.1186/s12936-017-1733-z.
3. Labbé P, David J-P, Alout H, Milesi P, Djogbénou L, Pasteur N, et al. 14 - Evolution of resistance to insecticide in disease vectors. In: tibayrenc MBT-G and E of ID Second E editor. London: Elsevier 2017 313–39. https://doi.org/10.1016/B978-0-12-799942-5.00014-7.
4. Duron O, Labbé P, Berticat C, Rousset F, Guillot S, Raymond M, et al. High *Wolbachia* density correlates with cost of infection for insecticide resistant *Culex pipiens* mosquitoes. Evolution. 2006;60:303–14.
5. Milesi P, Weill M, Lenormand T, Labbé P. Heterogeneous gene duplications can be adaptive because they permanently associate overdominant alleles. Evol Lett. 2017;1:169–80.
6. Berticat C, Bonnet J, Duchon S, Agnew P, Weill M, Corbel V. Costs and benefits of multiple resistance to insecticides for *Culex quinquefasciatus* mosquitoes. BMC Evol Biol. 2008;8:104.
7. Assogba BS, Djogbénou LS, Milesi P, Berthomieu A, Perez J, Ayala D, et al. An *ace-1* gene duplication resorbs the fitness cost associated with resistance in *Anopheles gambiae*, the main malaria mosquito. Sci Rep. 2015;5:1–12.
8. Tantely ML, Tortosa P, Alout H, Berticat C, Berthomieu A, Rutee A, et al. Insecticide resistance in *Culex pipiens* quinquefasciatus and aedes albopictus mosquitoes from La Réunion Island. Insect Biochem Mol Biol. 2010;40:317–24.
9. Ffrench-Constant RH, Steichen JC, Rocheleau TA, Aronstein K, Roush RT. A single-amino acid substitution in a gamma-aminobutyric acid subtype a receptor locus is associated with cyclodiene insecticide resistance in drosophila populations. Proc Natl Acad Sci USA. 1993;90:1957–61.
10. Lenormand T, Harmand N, Gallet R. Cost of resistance: an unreasonably expensive concept. Rethink Ecol. 2018;3:51–70.
11. Labbé P, Sidos N, Raymond M, Lenormand T. Resistance gene replacement in the mosquito *Culex pipiens*: fitness estimation from long-term cline series. Genetics. 2009;182:303–12.
12. Lenormand T, Bourguet D, Guillemaud T, Raymond M. Tracking the evolution of insecticide resistance in the mosquito *Culex pipiens*. Nature. 1999;400:861–4.
13. Milesi P, Lenormand T, Lagneau C, Myl E. Relating fitness to long-term environmental variations in natura. Mol Ecol. 2016;21:5483–99.

14. Clarke GM. The genetic and molecular basis of developmental stability: The *Lucilia* story. Trends Ecol Evol. 1997;12:89–91.
15. Guillemaud T, Lenormand T, Bourguet D, Chevillon C, Pasteur N, Raymond M. Evolution of resistance in *Culex pipiens* : allele replacement and changing environment. Evolution. 1998;52:443–53.
16. Lenormand T, Guillemaud T, Bourguet D, Raymond M. Appearance and sweep of a gene duplication: adaptive response and potential for new functions in the mosquito *Culex pipiens*. Evolution. 1998;52:1705–12.
17. Labbé P, Berticat C, Berthomieu A, Unal S, Bernard C, Weill M, et al. Forty years of erratic insecticide resistance evolution in the mosquito *Culex pipiens*. PLoS Genet. 2007;3:e205.
18. Assogba BS, Alout H, Koffi A, Penetier C, Djogbénou LS, Makoundou P, et al. Adaptive deletion in resistance gene duplications in the malaria vector anopheles gambiae. Evol Appl. 2018;11:1245–56. https://doi.org/10.1111/eva.12619.
19. Orr HAA. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution. 1998;52:935–49.
20. Heckel DG. Perspectives on gene copy number variation and pesticide resistance. Pest Manag Sci. 2022;78:12–8. https://doi.org/10.1002/ps.6631.
21. Labbé P, Milesi P, Yébakima A, Pasteur N, Weill M, Lenormand T. Gene-dosage effects on fitness in recent adaptive duplications: *ace-1* in the mosquito *Culex pipiens*. Evolution. 2014;68:2092–101.
22. Assogba BS, Milesi P, Djogbénou LS, Berthomieu A, Makoundou P, Baba-Moussa LS, et al. The *ace-1* locus is amplified in all resistant *Anopheles gambiae* mosquitoes: fitness consequences of homogeneous and heterogeneous duplications. PLOS Biol. 2016;14:e2000618.
23. Milesi P, Assogba BS, Atyame CM, Pocquet N, Berthomieu A, Unal S, et al. The evolutionary fate of heterogeneous gene duplications: a precarious overdominant equilibrium between environment, sublethality and complementation. Mol Ecol. 2018;27:493–507.
24. Weill M, Lutfalla G, Mogensen K, Chandre F, Berthomieu A, Berticat C, et al. Insecticide resistance in mosquito vectors. Nature. 2003;423:423–6.
25. Weill M, Berthomieu A, Berticat C, Lutfalla G, Nègre V, Pasteur N, et al. Insecticide resistance: a silent base prediction. Curr Biol. 2004;14:R552-3.
26. Bourguet D, Roig A, Toutant JP, Arpagaus M. Analysis of molecular forms and pharmacological properties of acetylcholinesterase in several mosquito species. Neurochem Int. 1997;31:65–72.
27. Alout H, Djogbénou L, Berticat C, Chandre F, Weill M. Comparison of *Anopheles gambiae* and *Culex pipiens* acetycholinesterase 1 biochemical properties. Comp Biochem Physiol B-Biochemistry Mol Biol. 2008;150:271–7.
28. Bourguet D, Guillemaud T, Chevillon C, Raymond M. Fitness costs of insecticide resistance in natural breeding sites of the mosquito *Culex pipiens*. Evolution. 2004;58:128–35.
29. Djogbénou L, Noel V, Agnew P. Costs of insensitive acetylcholinesterase insecticide resistance for the malaria vector *Anopheles gambiae* homozygous for the G119S mutation. Malar J. 2010;9:12.
30. Grau-Bové X, Lucas E, Pipini D, Rippon E, van Hof't AE, Constant E, et al. Resistance to pirimiphos-methyl in West African Anopheles is spreading via duplication and introgression of the Ace1 locus. PLOS Genet. 2021. https://doi.org/10.1371/journal.pgen.1009253.
31. Georghiou GP, Metcalf RL, Gidden FE. Carbamate-resistance in mosquitos Selection of Culex pipiens fatigans Wiedemann (*C quinquefasciatus* say) for resistance to Baygon. Bull World Health Organ. 1966;35:691–708.
32. Berticat C, Boquien G, Raymond M, Chevillon C. Insecticide resistance genes induce a mating competition cost in *Culex pipiens* mosquitoes. Genet Res. 2002;79:41–7.
33. Alout H, Weill M. Amino-acid substitutions in acetylcholinesterase 1 involved in insecticide resistance in mosquitoes. Chem Biol Interact. 2008;175:138–41.
34. Roger SO, Bendich AJ. Extraction of DNA from plant tissues. In: Gelvin SB, Schilperoort RA, editors. Plant Molecular Biology Manual. Cambridge: Academic Publishers; 1988.
35. Labbé P, Berthomieu A, Berticat C, Alout H, Raymond M, Lenormand T, et al. Independent duplications of the acetylcholinesterase gene conferring insecticide resistance in the mosquito *Culex pipiens*. Mol Biol Evol. 2007;24:1056–67.
36. Weill M, Berticat C, Raymond M, Chevillon C. Quantitative polymerase chain reaction to estimate the number of amplified esterase genes in insecticide-resistant mosquitoes. Anal Biochem. 2000;285:267–70.
37. Bourguet D, Pasteur N, Bisset J, Raymond M. Determination of Ace 1 genotypes in single mosquitoes : toward an ecumenical biochemical test. Pestic Biochem Physiol. 1996;55:122–8.
38. Ellman GL, Courtney KD, Andres V, Featherstone RM. A new and rapid colorimetric determination of acetylcholinesterase activity. Biochem Pharmacol. 1961;7:88–95.
39. Karunarathne P, Pocquet N, Labbé P, Milesi P. BioRssay: an R package for analyses of bioassays and probit graphs. Parasit Vectors. 2022. https://doi.org/10.1186/s13071-021-05146-x.
40. Crawley MJ. The R book. Chichester: John Wiley & Sons Ltd.; 2007.
41. Zeileis A, Hothorn T. Diagnostic checking in regression relationships. R News. 2002;2:7–10.
42. Berticat C, Duron O, Heyse D, Raymond M. Insecticide resistance genes confer a predation cost on mosquitoes. Culex pipiens Genet Res. 2004;83:189–96.
43. Liebman KA, Pinto J, Valle J, Palomino M, Vizcaino L, Brogdon W, et al. Novel mutations on the *ace-1* gene of the malaria vector *Anopheles albimanus* provide evidence for balancing selection in an area of high insecticide resistance in Peru. Malar J. 2015;14:1–10.
44. Alout H, Labbé P, Pasteur N, Weill M. High incidence of *ace-1* duplicated haplotypes in resistant *Culex pipiens* mosquitoes from Algeria. Insect Biochem Mol Biol. 2011;41:29–35.
45. Osta MAMMA, Rizk ZZJ, Labbé P, Weill M, Knio K. Insecticide resistance to organophosphates in *Culex pipiens* complex from Lebanon. Parasit Vectors. 2012. https://doi.org/10.1186/1756-3305-5-132.
46. Dang K, Doggett SL, Veera Singham G, Lee C-Y. Insecticide resistance and resistance mechanisms in bed bugs, *Cimex* spp. (Hemiptera: Cimicidae). Parasit Vectors. 2017. https://doi.org/10.1186/s13071-017-2232-3.

Publisher's Note

Chapitre III. Duplications et résistance aux insecticides : du local au global

Après avoir caractérisé les structures des allèles dupliqués du locus *ace-1* dans des souches de laboratoire et dans des populations provenant des quatre coins du monde, j'ai voulu un peu changer d'air. J'ai donc décidé de m'atteler à l'identification de nouvelles duplications touchant d'autres gènes de résistance aux insecticides. Tout en me permettant de brosser dans le sens du poil ma fibre complétionniste, je voyais plusieurs intérêts à cette analyse. Tout d'abord, elle me donnait l'occasion de tester la robustesse du *pipeline* d'identification que j'ai développé en le confrontant à de nouveaux gènes. Ensuite et surtout, elle me permettait de prendre du recul sur mon sujet de thèse, et de commencer à considérer la question de l'importance des duplications dans la réponse adaptative rapide à une pression de sélection forte, non plus par le prisme d'un gène unique mais en tant que processus adaptatif. Dans ce chapitre, je présenterai essentiellement des résultats préliminaires. J'ai passé beaucoup (peut-être trop ?) de temps focalisé sur l'étude des structures des duplications *ace-1* et je suis loin d'avoir terminé les analyses que je veux réaliser sur le reste du jeu de données que j'ai généré. Ce chapitre sera donc constitué de certaines de mes analyses en chantier, et si elles ne sont pas prêtes pour une publication, je suis confiant de l'intérêt qu'elles présentent.

Je me suis donc intéressé à trois gènes (mais quatre locus) largement étudiés pour leur rôle dans la résistance aux insecticides[1] : le gène du récepteur de l'acide γ-aminobutyrique (GABA), *Rdl* ; le gène codant pour les protéines membranaires des canaux sodium voltage-dépendants, *vgsc* (*Voltage Gated Sodium Channel*) ; et enfin le supergène de détoxication *Ester* (qui comprend deux locus, les α et les β-*esterases*).

I. Les nouveaux gènes cibles de mon analyse

Je vous épargnerai une longue et douloureuse présentation de ces différents gènes impliqués dans la résistance[2], mais j'en esquisserai tout de même un court résumé et j'exposerai surtout les raisons qui m'ont poussé à les choisir pour mes analyses.

[1] Un autre mécanisme de résistance est bien connu chez les moustiques et implique plusieurs gènes de la famille CYP450 (e.g. Gong *et al.,* 2021). Dans le cas présent, considérant la qualité de l'assemblage du génome à ma disposition, j'ai préféré me consacrer à des cas plus simples, n'impliquant *a priori* pas de famille multigénique (donc riches en duplications passées et indépendantes de l'adaptation que j'étudie). Un jour peut-être...

[2] Non, vraiment, ne me remerciez pas, ça me fait plaisir.

I.1. *Rdl*

GABA est un neurotransmetteur essentiel chez les insectes, actif dans tout le système nerveux. Les récepteurs GABA, codés par le gène *Rdl*, sont associés à des canaux activés par l'ion chlore Cl^- : leur hyperpolarisation bloque les signaux neuronaux (Ffrench-Constant *et al.,* 2000). La cyclodine dieldrine (CD) et certains pyréthrinoïdes (PYR) se lient à ces récepteurs et les bloquent en position fermée, ce qui entraîne la mort des organismes par convulsion. La résistance aux insecticides CD et PYR résulte d'une mutation ponctuelle dans le gène *Rdl*. D'abord identifiée chez *Drosophila melanogaster*, une mutation engendrant le changement d'une alanine en sérine sur le 302^e acide aminé (A302S) de la protéine affecte la liaison des insecticides aux récepteurs et induit la résistance (Ffrench-Constant *et al.,* 2000). L'allèle de résistance est co-dominant et les hétérozygotes présentent un phénotype intermédiaire entre résistance et sensibilité. La résistance a été trouvée dans diverses espèces d'insectes, la plupart partageant la même mutation (la position est partagée, mais l'acide aminé muté peut être une sérine ou une glycine), et elle est notamment retrouvée chez *Cx. pipiens s.l.* (Tantely *et al.,* 2010 ; Taskin *et al.,* 2016 ; Pocquet *et al.,* 2013) et *An. gambiae s.l.*. (Asih *et al.,* 2012). Dans ces deux complexes d'espèces, la mutation semble coûteuse et est principalement retrouvée dans les populations traitées, mais son impact exact n'est pas complètement caractérisé (Labbé et al. 2017). Comme pour *ace-1* une duplication hétérogène associant un haplotype sensible et résistant a été identifiée, d'abord chez *Myzus persicae* (Anthony *et al.,* 1998), puis une seconde fois chez *D. melanogaster* (Remnant *et al.,* 2013). Jusqu'à présent, les duplications du gène *rdl* sont inconnues chez les moustiques, et ce malgré des études populationnelles de très grande ampleur dans des populations résistantes chez *An. gambiae s.l.* (Lucas *et al.,* 2019 ; Clarkson *et al*., 2020).

I.2. *vgsc*

Les *vgsc* permettent le passage d'ions sodium Na^+ au travers de la membrane axonale. Ces mouvements ioniques engendrent une dépolarisation permettant le transfert de l'influx nerveux. Le dichlorodiphényltrichloroéthane[3] (DDT) et les insecticides PYR ciblent spécifiquement les *vgsc* et retardent leur fermeture, ce qui stoppe le potentiel d'action. Les organismes touchés par ces insecticides sont alors temporairement incapables de se mouvoir (effet *knockdown*). Là-encore, plusieurs mutations ponctuelles réduisant l'affinité des

[3] à mes souhaits.

insecticides pour leur cible ont été découvertes. On les regroupe souvent sous l'appellation de mutations *kdr* (pour *knockdown resistance*). Comme pour *ace-1* et *rdl*, ces mutations sont partagées par plusieurs espèces, notamment deux mutations affectant le même acide aminé, L1014F et L1014S. Elles sont liées à des désavantages sélectifs et des niveaux de résistance différents, qui suggèrent autant de *trade-off* dans les populations soumises aux insecticides, notamment chez les moustiques (Labbé *et al.,* 2009). Ces mutations L1014F/S sont connues chez plusieurs espèces de moustiques, dont *An. gambiae* (Ranson *et al.,* 2000), *Cx. quinquefasciatus* (Norris & Norris, 2011) et *Cx. p. pipiens* (Martinez *et al*., 1999). Chez *Cx quinquefasciatus,* une duplication hétérogène associant un haplotype S et R L1014F a été identifiée, mais la copie S ne serait plus fonctionnelle des suites d'une large délétion d'un fragment de sa séquence et peut être perçue comme un pseudo-gène (Xu *et al.,* 2011). Enfin, chez *Ae. aegypt*i, une potentielle duplication associerait plusieurs haplotypes sensibles et résistants de deux autres mutations de résistance, I1011M et V1016I (Martins *et al.,* 2013).

I.3. Supergène *Ester*

Ester est composé de deux locus, codant pour les *α* et les *β*-esterases (que je noterai par la suite *est-α* et *est-β*). Les locus codant *est-α* et *est-β* sont très proches (<1cM de recombinaison) et se comportent en général comme un supergène transmis d'un bloc (Berticat *et al.,* 2001). Ils appartiennent à la famille génique des carboxylestérases. Contrairement aux gènes que j'ai présentés jusqu'à présent, ce sont des gènes de détoxication, hydrolysant les groupements *ester* des xénobiotiques en composés moins toxiques pour les organismes. La résistance est donc métabolique, par augmentation de la quantité de protéines produites, et non plus liée à la modification de la cible des xénobiotiques. Les *α* et *β-esterases* jouent donc un rôle dans la résistance à plusieurs classes d'insecticides (principalement OP et CX, mais aussi PYR bien que le niveau de résistance soit réduit, Labbé *et al.,* 2017). Les allèles de résistance sont de deux types : i) pour certains l'augmentation de la quantité de protéines produites est seulement au niveau de la régulation de l'expression (ex. $Ester^{1}$) ; ii) pour d'autres, elle résulte de multiples duplications, ou amplifications, d'un (ex. allèle $Ester^{B1}$) ou des deux locus (ex. allèles $Ester^{2}$ et $Ester^{4}$), avec des nombres de copies très variables, de 2-3 copies à ≈ 80 (ex. Weill et al. 2000). Ces gènes ont été largement étudiés, depuis longtemps, notamment chez *Cx. pipiens* pour lequel la résistance aux OP est surveillée depuis la fin des années 1960 dans la région de Montpellier. Cette surveillance à long terme a montré que plusieurs allèles de résistance *Ester* se sont succédés au fil du temps, et qu'ils correspondent à différents compromis en termes de valeurs

sélectives, et donc associés à différentes intensités de traitements (Labbé et al. 2009 ; Milesi *et al.* 2016) : *Ester*^{A} confère une résistance modérée mais n'est associé qu'à des désavantages sélectifs mineurs en absence d'insecticides (donc sélectionné si les traitements sont modérés ou plus intermittents), alors qu'*Ester*^{2} confère un niveau de résistance et des désavantages sélectifs plus élevés (donc sélectionné si les traitements sont plus intenses).

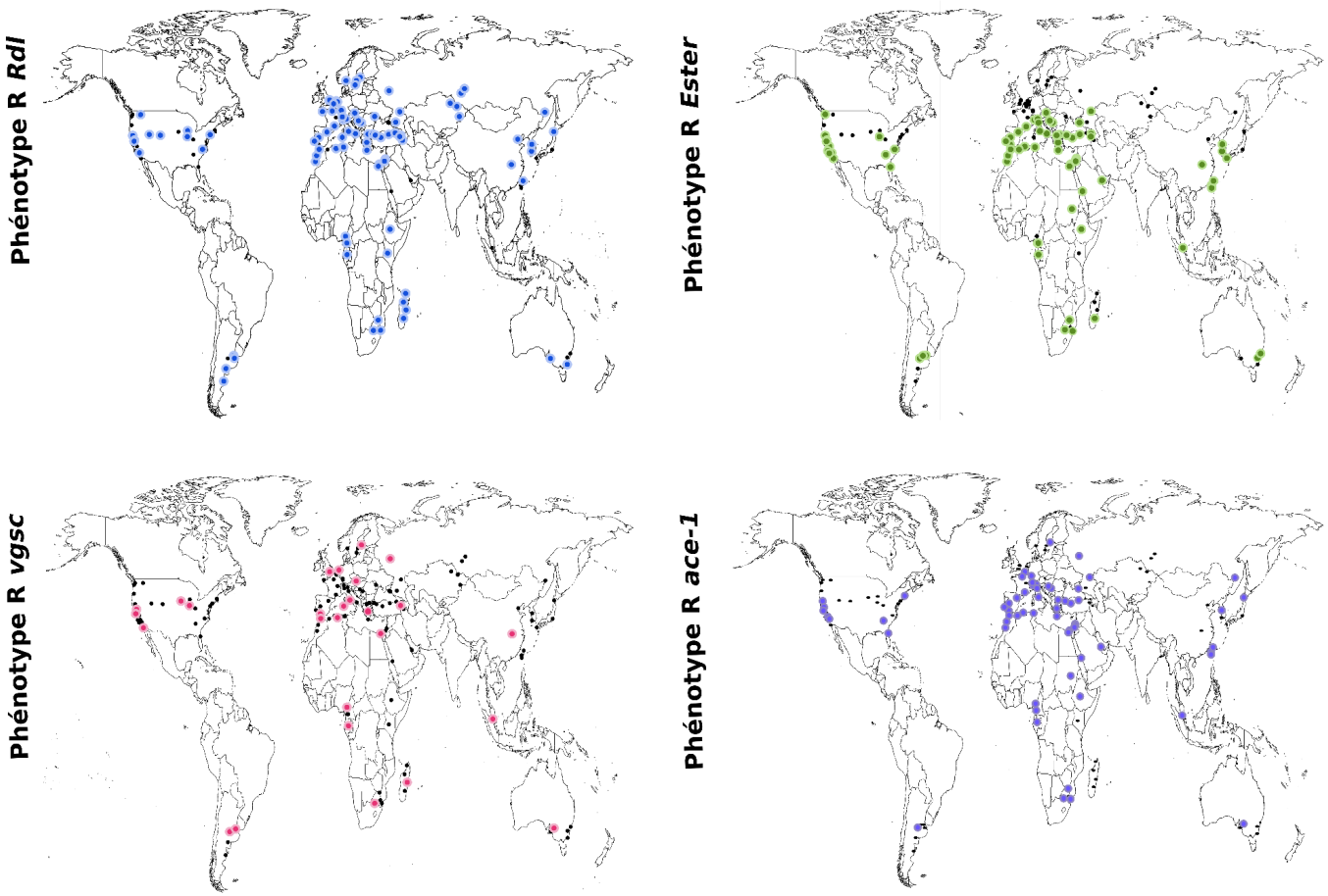


Figure III.1. Localisation des phénotypes résistants (R) du *PipPop Project* pour trois gènes : *ace-1*, *Rdl, vgsc* et le supergène *Ester*. Les populations échantillonnées sont indiquées en noir. Les points de couleurs représentent celles où des allèles de résistance (R) ont été identifiés. Les allèles R de différents gènes sont fréquemment retrouvés ensemble dans les mêmes populations, et même au sein de nombreux individus (394 en portent au moins un, 159 en portent deux, 38 en portent trois et un individu possède les quatre à la fois).

II. Identification des duplications.

II.1. *Rdl*

Référence et normalisation de DoC du gène cible.

J'ai d'abord utilisé la fonction d'*ArDu* qui permet de *screener* pour une mutation ponctuelle, afin d'identifier les porteurs de l'allèle de résistance A302S. J'en ai découvert un nombre effarant: 444 parmi les 830 individus du *PipPop Project*, soit plus d'un moustique échantillonné sur deux, répartis dans la quasi-totalité des populations du jeu de données (Fig. III.1, en bleu)[4]. Étonnamment, A302S semble majoritairement retrouvé à l'état homozygote (301/444), confirmant des fréquences élevées dans les populations échantillonnées. De manière intéressante, l'haplotype de résistance A302S est assez fréquemment associé à d'autres mutations de résistance (70 avec *ace-1*, 22 avec *vgsc L*1014S/F et 104 avec des duplications du supergène *Ester*). Bien que ces chiffres me paraissent particulièrement élevés, ils sont cohérents avec une étude récente signalant des fréquences élevées de cette mutation dans des populations sauvages (~80% à La Réunion chez *Cx. quinquefasciatus ;* Lebon *et al.,* 2022). Idéalement il m'aurait fallu rechercher d'autres mutations compensatoires, connues pour alléger le coût de l'allèle A302S, mais là-encore, ce sera pour plus tard…

Le gène *Rdl* aura bouleversé les certitudes que j'avais acquises sur le fonctionnement de mon *pipeline*. Comme je l'ai précisé dans le chapitre II, j'ai dû essentiellement me reposer sur la *DoC* pour caractériser l'étendue des zones dupliquées sur les données de populations naturelles, les autres indices de présence d'une duplication (tailles d'*insert* et *soft-clipped reads*) étant souvent absents (surement du fait de la relativement faible profondeur de séquençage). Pour *ace-1*, j'avais néanmoins une tendance claire : les deux références que j'utilise pour calculer le nombre de copies du gène-cible ne sont pas équivalentes, la *DoC* médiane du chromosome (DoC_{CHROM}) produisant des nombres de copies systématiquement plus élevés que la *DoC* moyenne du gène de référence (DoC_{REF}). J'avais remarqué que tous les individus désignés comme dupliqués d'après la DoC_{CHROM} mais pas d'après la DoC_{REF} étaient à chaque fois des faux positifs, ce qui m'avait amené à croire que cette dernière était généralement à privilégier. Toutefois, il semble que le gène *Rdl* soit en moyenne moins couvert qu'*ace-1* et que la DoC_{REF} soit trop stringente pour être utilisée ici. A ce stade et en utilisant la DoC_{CHROM}, j'ai identifié neufs individus potentiellement dupliqués (parmi les 830

[4] Soupçonnant la présence d'une gonade dans le potage, j'ai repris plusieurs fois ces analyses. Elles sont rigoureusement exactes, et j'ai aussi pu constater que la souche ayant servi à établir le génome de référence, JHB, est porteuse de la mutation de résistance A302S.

du *PipPop Project*), avec *a priori* une structure unique pour les neufs. Je dis "potentiellement" parce que je n'ai pas eu le temps de m'intéresser assez en détail à cette duplication pour en confirmer l'existence. J'observe bien une faible augmentation de *DoC* sur plusieurs locus en amont de *Rdl,* et j'ai trouvé une position marquée par des *soft-clipped reads* et des tailles d'*insert* très élevées (>600 kb), ce qui semble confirmer son existence; mais pour en être assuré, il me faudra reprendre les analyses sur ces individus plus en détail et tenter d'identifier d'autres signes de duplication (position avec triples SNPs, signaux de *breakpoints,* ou idéalement demander à Yuki Haba, qui a réalisé les extractions pour le séquençage, s'il resterait de quoi faire une toute petite qPCR).

II.2. *vgsc*

Parmi les 830 individus du *PipPop project***,** je n'ai identifié que 42 individus porteurs de mutations de résistance sur le codon de l'acide aminé L1014 (Fig. III.1). Parmi eux, très peu semblent être homozygotes : j'ai identifié deux R_{L1014F}/R_{L1014F} et un seul R_{L1014S}/R_{L1014S}. Le reste des allèles de résistance semblent être portés par des hétérozygotes. L1014S/F est systématiquement trouvée associée à d'autres allèles de résistance, dont huit fois avec *ace-1* G119S, 14 fois avec des duplications du supergène *Ester* et 22 fois avec *Rdl* A302S.

En revanche, je n'ai découvert aucune duplication du gène *vgsc*. Quatre individus présentaient un nombre de copies légèrement supérieur au seuil de détection quelque soit la référence utilisée pour son calcul, DoC_{REF} ou DoC_{CHROM} (moyenne de 1.61±0.18 copies, établies depuis la DoC_{REF}), mais en approfondissant leur analyse ils se sont avérés être de faux positifs : je n'ai trouvé aucun *soft-clipped read* ou taille d'*insert* inattendue permettant de soutenir la présence d'une duplication, et je me suis rendu compte que l'augmentation de leur *DoC* n'était pas retrouvée sur tous leurs exons (moins de la moitié dépassaient le seuil de 1.4).

II.3. Supergène *Ester*

Contrairement au deux autres locus, les duplications du supergène *Ester* sont bien connues chez *Cx. pipiens s.l.,* puisqu'elles constituent le principal mécanisme de résistance à ce locus. En explorant le *PipPop Project*, je savais que je devrais retrouver des duplications affectant ces gènes. Le moins que l'on puisse dire c'est que mes attentes ont été comblées: 246 des 830 individus échantillonnés portaient une duplication, et sont retrouvés dans 34 des 46 populations sur les cinq continents échantillonnés. Elles sont elles aussi fréquemment retrouvées associées à d'autres allèles de résistance, 61 fois avec *ace-1* G119S, 104 fois avec

Rdl A302S et 28 fois avec *vgsc* L1014F/S. A ce stade, et contrairement aux autres gènes de résistance, je ne me suis pas du tout intéressé à la diversité nucléotidique, et je ne peux donc pas identifier les allèles dupliqués retrouvés.

Des structures différentes chez *α* et *β-esterases* ?

Puisque j'ai bien trouvé des duplications du supergène *Ester*, se posait ensuite la question de leur caractérisation (taille, nombre de copies, réarrangements secondaires, etc.). Je savais déjà que les niveaux d'amplifications pouvaient être très différents, de deux à >75 copies ; Weill *et al.* (2000) avait même trouvé des individus portant 80 copies dans les îles Jasmin au Vietnam. Mais qu'en était-il des structures englobant ces amplifications ?

Le premier indice de la pluralité des structures m'est apparu avant même d'observer les graphes de *DoC:* en effet, si pour 199 individus sur 246 j'ai bien le même nombre de copies pour les deux locus, j'ai également retrouvé que les deux locus du supergène ne sont pas toujours co-amplifiés. Pour vingt individus provenant de populations aux USA, en Russie, Chine et au Japon, seul le locus *est-β* était dupliqué, et là-aussi, le nombre de copies était variable, allant de deux à vingt. Ces résultats sont cohérents avec la présence de l'allèle $Ester^{B1}$, largement répandu en Amérique du Nord et présent en Chine, et effectivement connu pour être un amplification du seul locus *est-β* (Qiao & Raymond, 1995). A noter, pour les variants où un seul locus est amplifié, c'est à ce stade toujours *est-β* dans les échantillons traités, conformément à la diversité déjà connue des allèles *Ester* chez *Cx. pipiens s.l.* (Weill et al, 2003 ; Cui *et al.,* 2007).

Plus étonnant, chez 47 individus, j'ai observé des différences entre les nombres de copies des locus *est-α* et *est-β,* allant de une à quatre copies en plus pour un des locus. Elles pourraient être attribuées à une incertitude technique quand le nombre total de copies dépasse une dizaine. Quand c'est le locus *est-α* qui est en plus grand nombre de copies (trois individus sur 47), c'est d'ailleurs le cas systématiquement : on ne retrouve guère plus d'une copie supplémentaire par rapport au locus *est-β*, et des nombres moyens de copies élevés pour les deux. Néanmoins, cela semble plus significatif quand il n'y a que quelques copies de chaque locus avec une bonne couverture, comme c'est le cas pour certains individus où le nombre de copies de *est-β* dépasse celui des copies de *est-α* (44 individus sur 47). Puisqu'on trouve des allèles où seul *est-β* est amplifié, une explication possible serait qu'il s'agisse d'hétérozygotes, porteurs d'un allèle dupliqué avec même nombre de copies pour *est-α* et *est-β*, associé à un autre allèle dupliqué type $Ester^{B1}$. Avec les outils dont je dispose, *i.e.* en

l'absence d'haplotypes sur l'ensemble de la zone dupliquée/amplifiée, je ne peux cependant pas rejeter l'hypothèse alternative que les deux copies soient réellement différentiellement amplifiées, même si cela n'a jamais été documenté à ma connaissance. Je me retrouve donc avec un problème que j'avais déjà rencontré sur *ace-1*, et lié à la nature de mes données : comme j'observe des individus de populations naturelles, dont beaucoup sont hétérozygotes, je ne peux pas être certain que les variations que j'observe sont le fait d'une structure unique ou d'un mélange. Ce problème est d'autant plus impactant ici que les duplications du supergène *Ester* semblent très nombreuses et sont probablement polymorphes au sein d'une même population (ex. dans la région de Montpellier ; Labbé *et al.,* 2009 et Milesi *et al.,* 2016).

Cependant, tout s'est évidemment compliqué lorsque j'ai voulu caractériser les *breakpoints* de ces structures. Loin des structures nettes que j'espérais observer, les graphes de *DoC* tiennent plus de l'électrocardiogramme d'une personne souffrant d'arythmie chronique (Fig.III.2). Ils présentent une série de zones très largement couvertes, mais qui ne sont pas toutes partagées entre individus dupliqués. J'espérais que la partie du *pipeline ArDu* permettant d'identifier les positions candidates des *breakpoints* m'aiderait à y voir un peu plus clair, mais j'ai rapidement déchanté : les *soft-clipped reads* abondent dans toute la zone et ne convergent pas sur une position particulière. Les tailles d'*insert* anormales ne m'ont également été d'aucun secours, puisque comme pour les structures *ace-1* identifiées dans le *PipPop Project*, je n'ai pas trouvé de congruence entre leurs positions et celles des *soft-clipped reads*...

Structures fragmentées : mauvais assemblage, soucis bio-informatiques ou réarrangements secondaires ?

Seules deux hypothèses me semblent permettre d'expliquer l'apparence fragmentée des duplications *Ester* : i) une fragmentation artéfactuelle liée à un problème d'origine bioinformatique, ou ii) une fragmentation réelle liée à plusieurs réarrangements secondaires touchant des zones différentes des amplicons. Toute la question est donc de savoir quelle hypothèse prévaut[5].

i) Il est possible que la zone entourant les locus *est-α* et *est-β* ne soit pas parfaitement assemblée sur le génome de référence ; on a déjà vu de tels problèmes pour *ace-1* par exemple, parce que le génome de *Cx. quinquefasciatus* n'est pas encore aussi poli que celui

[5] ou si les deux entrent en synergie pour me pourrir la vie…

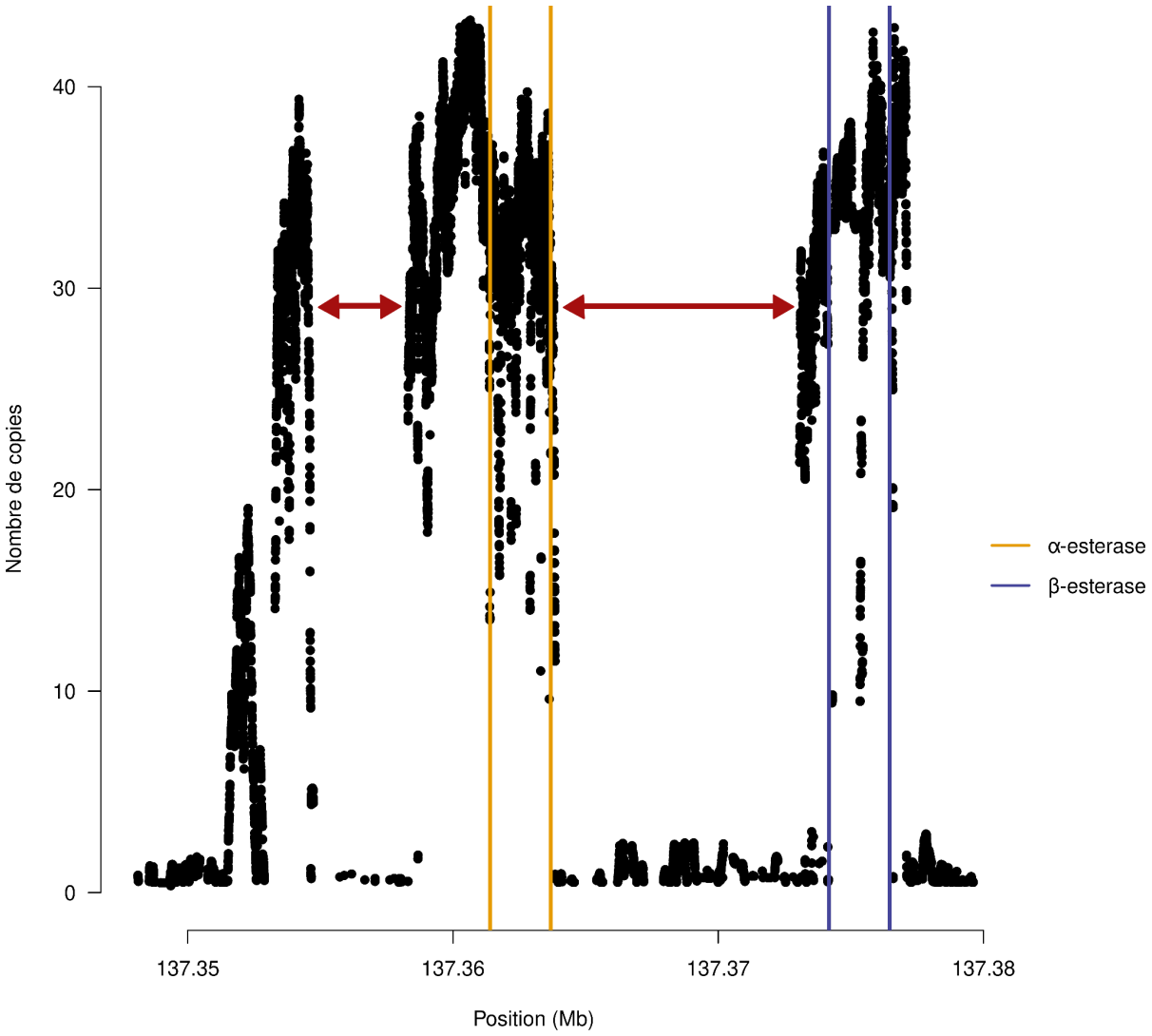


Figure III. 2. Exemple de graphe de *DoC* du supergène *Ester*. La couverture normalisée (obtenue avec la DoC_{REF}; voir texte) est représentée en fonction de la position génomique (Mb). Contrairement aux graphes de *DoC* présentés dans le chapitre I, ces données sont représentées sans transformation (sliding window et bining) pour mieux rendre compte des variations de *DoC* sur la zone. Les positions des locus des *α*- et des *β-esterase* sont indiquées par les lignes de couleur. Ils sont largement amplifiés (~ 32 copies chacun). La structure dupliquée présente de larges lacunes qui pourraient correspondre à différentes délétions secondaires (double flèches rouges).

d'*An. gambiae*. Dans ce cas, une région normalement amplifiée d'un seul bloc pourrait apparaître fragmentée lors de l'alignement. Les *reads* s'aligneraient correctement sur la référence, mais leur position relative serait fausse à cause d'erreurs d'assemblage.

ii) Il est également tout à fait concevable que des délétions et des duplications secondaires (de différentes tailles) aient touché cette zone, comme ça a déjà été observé pour les allèles R^{x}* chez *An. gambiae* ou dans la structure ε^{+}* de *Cx. pipiens*. Un processus analogue à celui observé chez les duplications homogènes d'*ace-1* a peut-être eu lieu, avec une ou des délétions permettant de réduire le déséquilibre du dosage génique. Ceci serait d'autant plus probable que le nombre de copies est nettement plus large (> 10 copies le plus souvent), et donc que la pression de sélection pour restaurer le dosage génique initial serait potentiellement très élevée. De même, le nombre de copies identiques élevé faciliterait l'occurrence de réarrangement(s) secondaire(s).

Il est toutefois impossible de discriminer ces scénarios sur la base des données à ma disposition, puisqu'ils sont à l'origine des mêmes signaux (*i.e.* une accumulation de *soft-clipped reads* et des tailles d'*insert* anormales). Pour tester l'hypothèse d'un mauvais assemblage du génome de référence (i), j'ai tenté d'effectuer un ré-assemblage local de la zone comprenant le supergène *Ester*. Pour ce faire, j'ai aligné des *long reads* issus du séquençage de la souche sensible de référence Slab (8 kb, *Minion long reads*[6]) sur le génome de référence JHB. J'ai isolé tous les *reads* s'alignant sur la zone comprenant le supergène *Ester* (1 Mb centré sur les locus *est-α* et *est-β*) et je les ai ensuite réassemblés pour établir une nouvelle référence sur laquelle aligner le génome d'un individu porteur d'amplification *Ester* (assemblage local *de novo* avec *flye ;* Kolmogorov *et al.,* 2019*)*. Ce faisant, j'ai de nouveau identifié des pics de couverture séparés par des zones non couvertes, similaires à ceux observés sur l'assemblage de référence. Cette analyse, un peu brute j'en conviens, et sur laquelle il me faudra revenir, tend à suggérer que les pics de *DoC* correspondraient bien à une réelle fragmentation de la zone amplifiée, donc certainement due à des réarrangements secondaires.

Nombre et origine des structures.

Il m'a donc fallu me rendre à l'évidence: je ne peux pas identifier les structures du supergène *Ester* comme j'ai pu le faire pour *ace-1*. Les variations de nombre de copies que

[6] Malgré les efforts de Haoues Alout, ces *reads* n'ont pas permis de générer un génome de référence de meilleure qualité (ou même similaire) à celui établi à partir de JHB. Ça m'aurait pourtant bien aidé d'avoir une référence réellement *Cx. quinquefasciatus* (plutôt que ce taxon bizarre et *outlier* dont est issue JHB, voir Aparté taxonomique).

j'ai identifiées sur ces gènes résultent très probablement de duplications successives, et il semblerait qu'elles ne touchent pas toujours les mêmes positions. Puisque tenter d'en identifier les *breakpoints* nécessiterait un temps que je n'ai pas (pour un gain pas si évident), je me suis concentré sur les bornes les plus larges que j'ai pu identifier, pour les délétions et les éventuels gènes embarqués (Fig. III. 3). J'ai ainsi pu identifier deux structures-types : i) une première ne contenant que l'amplification du locus *est-β,* structure notée ci-après *Estβ*[7], certainement liée à l'estérase B1 (Georghiou *et al.,* 1980 ; Raymond *et al.,* 1987), et ii) une seconde, que je noterai *Estαβ*, et qui comprend les deux locus du supergène et un gène embarqué, l'aldéhyde oxydase 1, dont la co-amplification avait déjà été signalée (Hemingway *et al.,* 2000). Ce gène semble toutefois inactivé puisque j'ai découvert une délétion couvrant une majeure partie de sa séquence (exon 2 à 5 ; Fig. III.3).

J'ai à nouveau utilisé *RepeatMasker* pour identifier la position d'éventuels ET dans la zone qui pourraient expliquer la position des *breakpoints* et leur diversité : j'en ai identifié plusieurs dont deux aux bornes de la structure *Estβ* et en limite d'une délétion interne de la structure *Estαβ* (en rouge ; Fig. III.3). Néanmoins, le génome de *Culex* contient tellement d'ET (~45%) qu'il est difficile d'en suggérer certains en particulier comme étant à l'origine des duplications observées sans faire œuvre de *cherry-picking* déplacé…

II.4. Importance des duplications dans l'adaptation aux insecticides.

Sur les trois gènes que j'ai rajouté à mon analyse, en plus de *ace-1* donc, je n'ai pour l'instant *formellement* identifié de duplications que dans le supergène *Ester*, bien qu'un faisceau d'indices soutient l'existence d'une duplication chez *Rdl* (cependant relativement peu fréquente). Pour *Ester* comme pour *ace-1*, mes travaux montrent qu'il existe une grande diversité d'haplotypes, de structures, et de nombre de copies. Contrairement à ce qui a été retrouvé chez *An. gambiae s.l.*, on trouve comparativement peu d'allèles de résistance et aucun allèle dupliqué du gène *vgsc* chez *Cx. pipiens s.l.* Il semble donc que les taxons de ce complexe soient globalement peu exposés aux insecticides de la famille des PYR (ce qui converge avec une étude précédente à Mayotte, Pocquet *et al.,* 2013). Pour *Rdl*, l'allèle dupliqué putatif reste rare, alors que la fréquence des allèles de résistance simple-copie est très élevée dans nos échantillons (environ la moitié des individus échantillonnés portent la mutation de résistance A302S). Ceci suggère que les populations de *Cx. pipiens s.l.* sont

[7] Les joies de la nomenclature décidément, qui me font par là-même reconsidérer la pertinence de l'utilisation de lettres grecques pour nommer les structures du chapitre précédent…

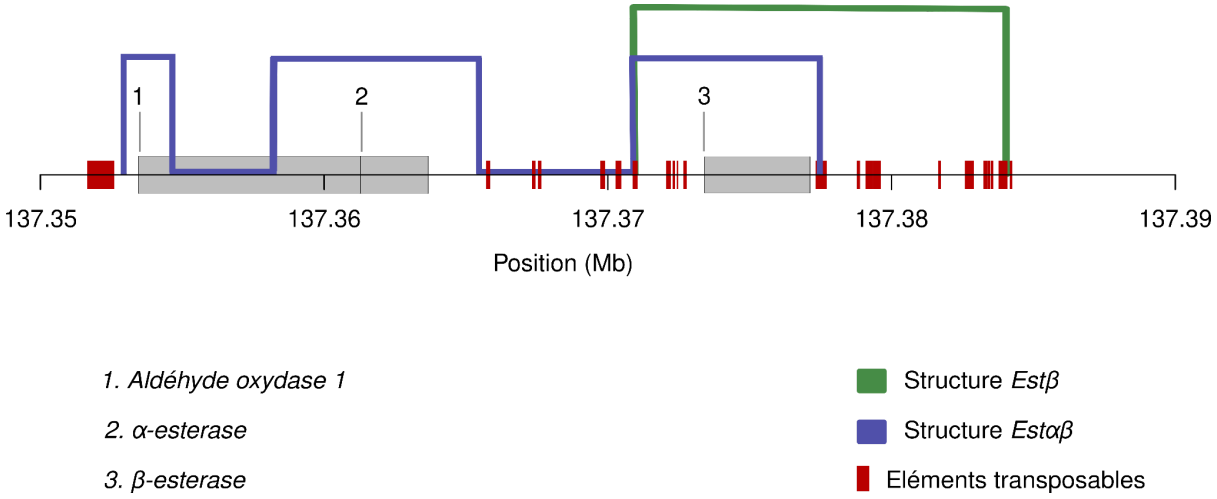


Figure III.3. Structures et *breakpoints* des duplications du supergène *Ester*. La position des gènes est indiquée par les rectangles gris, celles des éléments transposables en rouge. Les bornes représentent les *breakpoints* les plus larges identifiés pour chaque structure, en bleu pour la structure où les deux gènes *esterase* sont co-amplifiés (*Estαβ*), en vert quand seul le locus de la *β-esterase* est amplifié (*Estβ*).

encore régulièrement exposées à des insecticides sélectionnant ces mutations (voir aussi Tantely et *al.* 2010).

Grâce aux données que j'ai acquises, il est donc évident que les duplications ont un rôle important dans le processus d'adaptation aux insecticides. Il reste toutefois à comprendre leur origine, et en particulier à savoir si des caractéristiques locales du génome favorisent leur émergence. C'est une question complexe, qui demande de considérer beaucoup de facteurs différents. Dans un premier temps, je trouverais intéressant d'établir pour *Cx. pipiens* et *Cx. quinquefasciatus* des cartes de recombinaisons précises autour de ces locus, qui pourraient par exemple permettre d'identifier des points chauds de recombinaison: es cartes existent à plus large échelle, mais datent un peu (chez *Cx. pipiens s.l.* : Hickner *et al.,* 2013 ; Unger *et al.,* 2015; et chez *An. gambiae s.l. :* Ranson *et al.,* 2004 ; Sharakhova & Sharakhov, 2010), mais c'est à une échelle plus locale qu'il faudrait les établir.

Les deux autres questions majeures à explorer à mon avis sont i) de mieux comprendre la dynamique de ces allèles en populations naturelles, et ii) de mesurer leurs impacts sur la diversité génomique. Ça tombe bien, j'ai aussi posé des jalons à ce propos, j'en parlerai dans les perspectives.

Discussion générale

Nous voilà donc arrivés à la dernière partie de cette thèse[1], à la fois un bilan et une ouverture. Pendant ces trois dernières années, je me suis intéressé aux cas de plusieurs duplications du locus *ace-1* qui ont été largement sélectionnées dans plusieurs espèces de moustiques depuis une soixantaine d'années. J'ai concentré mes efforts sur l'étude de leur diversité, à la fois en termes de variations de séquences et de structures. Mon objectif était de mettre en lien cette diversité avec l'histoire évolutive des allèles dupliqués, ainsi qu'avec l'impact de leur sélection à l'échelle génomique. J'ai mené mes recherches sur deux modèles distincts, des complexes d'espèces très divergents : i) *An. gambiae s.l.,* le principal vecteur du paludisme, endémique d'Afrique sub-saharienne ; et ii) *Cx. pipiens s.l.*, un vecteur généralement plus discret, à la répartition cosmopolite. Ce faisant, j'ai étendu nos connaissances de la grande diversité des duplications dans chacun des complexes, à la fois en nombre d'allèles mais surtout de structures génomiques, et j'ai mis en lumière l'étendue de la convergence entre ces deux complexes d'espèces pourtant très divergents (pour rappel, ≈1 milliard de générations quand même !) dans l'adaptation aux insecticides OP et CX au locus *ace-1*.

Chez *An. gambiae s.l.,* j'ai démontré que plusieurs duplications hétérogènes ségrègent dans les mêmes populations naturelles (**Claret *et al.,* soumis**), révélant ainsi un polymorphisme insoupçonné pour cette espèce, parallèle aux situations déjà décrites chez *Cx. pipiens s.l.*. Nous avons également décrit un allèle R^x chez *Cx. pipiens s.l.*, une duplication homogène liant plusieurs copies R du gène *ace-1*, alors que ce type de duplication n'était jusqu'à présent connu que chez *An. gambiae s.l.* (**Milesi, Claret *et al.,* 2022**). Au cours de ces travaux, j'ai élaboré un outil bioinformatique automatisé (*ArDu*) permettant d'identifier les duplications d'un gène candidat depuis des données *WGS* et d'en caractériser la structure (nombre de copies, taille, réarrangements secondaires, gènes embarqués, points de cassure). Je l'ai d'abord utilisé pour confirmer l'homogénéité des structures découvertes chez *An. gambiae s.l.* (**Claret *et al.,* soumis**), puis pour caractériser la structure de trois duplications hétérogènes et d'une duplication homogène fixées dans des souches de laboratoire, dont certaines sont sublétales à l'état homozygote (phénotype HS). J'ai montré que ces trois allèles diffèrent par leur taille, et aussi par le nombre et l'identité des gènes embarqués avec *ace-1*. Si ces travaux n'ont malheureusement pas permis d'identifier l'origine du phénotype HS, ils m'ont néanmoins permis de déterminer plusieurs pistes à explorer pour le comprendre. La

[1] Je ne vous cache pas que j'en suis assez soulagé, et mes directeurs de thèse aussi ! Ces derniers mois ont été pour le moins épiques !

mise au point de *ArDu* m'a surtout permis d'étudier les duplications d'*ace-1*, puis d'autres gènes de résistance, dans les génomes de 830 *Cx. pipiens s.l.* en provenance de plusieurs populations naturelles du monde entier grâce au *PipPop Project*, un jeu de données inédit et d'une valeur extraordinaire auquel j'ai eu la chance d'accéder en primeur grâce à la non-moins extraordinaire générosité de Yuki Haba et Lindy McBride. Cela m'a permis de révéler la diversité d'organisation structurelle des duplications chez *Cx. pipiens s.l.,* là où une seule est retrouvée chez *An. gambiae s.l.*

Si le temps m'a manqué au final pour mener à bien tout ce que je souhaitais faire, ces travaux constituent les fondations indispensables à l'étude de certaines questions auxquelles je voudrais m'intéresser par la suite. Je vous propose de les détailler ici, en commençant par celles qui devraient trouver leur réponse le plus rapidement (en tous cas je l'espère !), pour finir par les questions générales qui me demanderont plus de temps (et de travail).

I. A boucler rapidement.

I.1. Diversité des structures *Culex*: quel lien avec les allèles identifiés ?

Mon premier objectif va être de publier les résultats que j'ai présentés dans le chapitre II. Nous avions pensé rédiger un unique article organisé en deux parties : une partie présentant les structures *Cx pipiens s.l.,* et une seconde plus axée revue de littérature dans laquelle nous établirons le parallèle entre *Cx. pipiens s.l.* et *An. gambiae s.l.*[2]. Sans que celà soit nécessaire pour la publication, j'aimerais me pencher sur une question qui me taraude depuis l'identification des structures et qui apporterait selon moi un intérêt supplémentaire à leur description. Je n'ai pour le moment pas essayé de voir si je pouvais associer ces structures aux allèles identifiés dans les travaux de Milesi *et al.* (2018). Dans cette étude, les auteurs ont identifié 27 duplications hétérogènes et quatre allèles de résistance R dans diverses populations *Cx. pipiens s.l.* avec moins de populations échantillonnées (une trentaine), mais à une échelle géographique comparable à celle du *PipPop Project*. Il me faudrait réussir à recréer des haplotypes à partir de données *short reads*, ce qui reviendrait en définitive à phaser les différents variants portés par les individus dupliqués. Cette tâche serait déjà suffisamment complexe si je travaillais sur des données de gènes en simple copie, mais il existe des outils pour le faire (*e.g.* WhatsHap, Martin *et al.,* 2016 ; SDhaP, Das & Mikalo, 2015 ; voir aussi Garg, 2021 pour une revue). Cependant la tâche devient cauchemardesque quand on s'intéresse à une zone effectivement polyploïde, comme celle créée par

[2] Toutefois la rédaction de ma thèse nous aura montré que l'étude des structures *Cx pipiens s.l.* était assez complexe pour mériter un article à part. Qu'en dites-vous ?

l'alignement sur le génome de référence des multiples amplicons d'un individu dupliqué. Heureusement, mon ambition est plus modeste, puisque la diversité des allèles D connus chez *Cx. pipiens s.l.* s'appuie sur un peu moins d'un kb (Milesi *et al.,* 2018). Je pourrais donc tout de même essayer de reconstruire des haplotypes sur ce fragment en cherchant du *linkage disequilibrium* entre *SNPs* de cette zone réduite. Pour cela, j'ai déjà établi le "profil" de chaque allèle identifié par Milesi *et al.* (2018), *i.e.* les *SNPs* propres à chaque allèle sur ce fragment. Si j'arrivais à retrouver ces *SNPs* dans les individus dupliqués et à les phaser je pourrais peut-être identifier ainsi l'allèle qu'ils portent.

Le problème de cette analyse (si tant-est qu'elle puisse être concrétisée) est que je me retrouverais alors face au même problème que j'ai rencontré quant à la distribution des copies dans les individus dupliqués potentiellement hétérozygotes (discuté en longueur dans le chap. **II. Des structures c'est bien, mais des allèles c'est mieux !**). De plus, il est possible que comme chez *An. gambiae s.l.,* des allèles mono-copie partagent l'haplotype de copies embarquées dans les duplications (**Claret *et al.,* soumis**). Un cas simple me permettra au moins de tester la faisabilité de ce projet, puisque je pourrai commencer par l'identification des haplotypes R dans les individus R^x (duplications homogènes). Je reste de toute façon intéressé par ce que je pourrais découvrir en explorant cette piste[3].

I.2. Des points chauds de recombinaison ?

Les diversités des structures découvertes chez *Cx. pipiens s.l.* et des haplotypes associés dans les duplications hétérogènes chez *An. gambiae s.l.* semblent indiquer que la zone génomique entourant *ace-1* présente de forts taux de recombinaison dans les deux complexes. J'aimerai tester l'hypothèse de l'existence de points chauds de recombinaison autour de cette zone, ce qui devrait être possible à partir des données populationnelles du *An. gambiae 1000 genome project* et du *PipPop Project.* A ma connaissance, il n'existe pas encore de carte de recombinaison fine pour les deux assemblages de référence. Il me semble intéressant de déterminer si les zones entourant les différents gènes de résistance présentent des différences en termes de taux de recombinaison, ce qui pourrait participer à expliquer la tendance de certains d'entre eux à être plus souvent dupliqués que d'autres, puisque je ne suis pas totalement convaincu que ces différences puissent être uniquement liées aux conditions de traitements insecticides[4]. Cette analyse pourrait être réalisée relativement rapidement en

[3] Une autre façon de dire : "Je sais que ça risque fort de rater, mais je veux pas lâcher le morceau avant d'avoir essayé."
[4] Oui Pierrick, je persiste !

suivant la méthode préconisée par Raynaud *et al.* (2023), *i.e.* en utilisant l'outil *LdHelmet* (Chan *et al.,* 2012) qui se base sur des approches de *linkage disequilibrium* depuis des données populationnelles pour déterminer l'existence de points chauds de recombinaison[5].

I.3. Duplications *Rdl* et *Ester.*

Grâce à *Ardu*, j'ai mis en évidence des signes soutenant l'existence en faible fréquence (9/830) d'une duplication du gène *Rdl* : des *reads* dont l'alignement est anormal et une faible augmentation de la *DoC* sur l'ensemble des exons du gène *Rdl* (Chap. III partie II.1). Cependant, la faible couverture relative de cette zone du génome (peut-être liée à une plus forte divergence locale avec le génome de référence[6]), et le fait qu'elle soit visiblement uniquement retrouvée à l'état hétérozygote (*i.e.* la *DoC* de *Rdl* est toujours proche de 1.5, valeur caractéristique d'un individu hétérozygote portant un allèle dupliqué et un allèle monocopie), m'empêchent de caractériser convenablement sa structure. Pour confirmer son existence, je pourrais rechercher le long du locus *Rdl*, et entre les positions des *breakpoints* putatifs de cette duplication, la présence de positions multi-alléliques (*i.e.* des positions avec plus de deux variants) : cela reviendrait à reproduire, par des analyses bio-informatiques et à plus large échelle, la méthode appliquée sur séquençage Sanger pour identifier des triple-pics qui m'a permis de trouver les nouveaux allèles Ag-D (voir **Claret *et al.,* soumis, Materials and methods**). Si je découvre plusieurs positions avec de multiples *SNPs* dont la fréquence relative est cohérente avec la présence de trois séquences différentes, je pourrais confirmer l'existence d'une duplication. Cette analyse pourrait d'ailleurs facilement être implantée dans *ArDu*, et elle pourrait donner des informations intéressantes à croiser avec les nombres de copies des gènes cibles et les tailles de zone dupliquée déjà prédits par le *pipeline* que j'ai développé.

Un autre point concernant cette fois les duplications du supergène *Ester* : là-aussi, il serait intéressant d'identifier les haplotypes des locus *est-α* et *est-β*. Les structures liant des dizaines de copies de chaque locus sont-elles des amplifications d'un unique haplotype ? Pourrait-on identifier la trace de recombinaison avec d'autres esterases qui créerait des allèles "mixtes" à la manière des duplications hétérogènes *ace-1* ? De telles chimères n'ont pour le moment été détectées que dans des croisements répétés entre souches fixées au laboratoire (Berticat *et al.,* 2001). Avec de telles données, nous pensons pouvoir valoriser rapidement

[5] Une collaboration avec la première auteure de cet article est d'ailleurs prévue à cet effet - collaboration qui devrait se trouver facilitée par le fait que je n'ai que trois portes à franchir pour discuter des détails des analyses avec l'intéressée.

[6] On y revient toujours… pourquoi avoir choisi JHB ?

sous la forme d'une publication le travail que j'ai fait sur les autres gènes de résistance (hors *ace-1*).

II. Ça risque d'être plus long.

A l'origine, ma thèse s'organisait autour de deux axes. Le premier était la caractérisation des structures des allèles dupliqués, et le second consistait en l'étude de l'impact de la sélection de ces allèles sur la diversité génomique. Grâce au *PipPop Project*, j'ai pu rechercher et caractériser des structures sur une échelle bien plus large que ce qui était initialement prévu (uniquement les duplication fixées dans des souches présentées dans le Chap. I), et je me suis ensuite laissé embarquer[7] par l'étude d'autres gènes de résistance. La contrepartie est que je n'ai pas eu le temps de m'intéresser à la deuxième question qui motivait ma thèse, mais je voudrais maintenant y remédier. Ça devrait d'ailleurs être d'autant plus intéressant avec le *PipPop Project*, puisque je vais pouvoir confronter à ces données les attendus théoriques de l'effet de la sélection d'une mutation sur la diversité génomique environnante. Je m'explique : généralement la présence même de *hard-sweep* est utilisée comme marqueur d'événements de sélection récents et intenses (par exemple dans le cadre de scans génomiques et d'approches par gènes candidats). Dans le cas présent, les mutations sous sélection positive sont connues *a priori*, et on peut donc étudier et comparer le comportement de la diversité neutre autour (et à l'intérieur, voir plus bas) des régions sous sélection.

Une des grandes leçons de ma thèse, c'est que l'exploration et la mise en place de nouvelles approches est toujours plus simple sur un jeu de données propre. Et ça tombe bien, avant de me jeter à corps perdu dans les données populationnelles *Culex,* je dispose toujours de celles du *An. gambiae* 1000 *Genome Project*, sur lesquelles des analyses similaires ont déjà été réalisées (Lucas *et al.,* 2019 ; Clarkson *et al.,* 2020 ; Grau-Bové *et al.,* 2021), pour me faire la main et servir de point de comparaison.

II.1. Génomique populationnelle de l'adaptation et sélection des mutations: du local au global.

La pression de sélection liée aux insecticides est à la fois soudaine et extrêmement élevée. On peut donc s'attendre à ce qu'elle impacte la diversité génétique à deux échelles : i) localement, sous forme d'un balayage sélectif (*selective sweep*) lorsque que la sélection d'une

[7] La rédaction de ma thèse finirait-elle par impacter mes choix de mots ?

mutation entraîne une chute de la diversité neutre environnante, et ii) globalement, lorsque la réduction soudaine de la taille réelle de la population (en termes de nombre d'individus, on parle alors de goulot d'étranglement ou *bottleneck*) entraîne une perte de diversité génétique (*i.e.* une baisse de la taille efficace N_e).

La pression de sélection créée par les insecticides a-t-elle été suffisamment forte pour que j'observe son impact à l'échelle des génomes ? Bien que je n'ai pas encore eu le temps de me pencher pleinement sur cette question, une première approche serait de tester la corrélation entre la diversité nucléotidique globale (π) de chacune des populations et la proportion d'individus résistants qu'elles comportent ; dans ce cas la fréquence des résistants est considérée comme un proxy de l'intensité de la pression de sélection. Une des limites de cette approche est qu'avec le temps, et sous l'effet de la recombinaison et du flux de gène entre populations, ce signal devrait s'éroder[8]. Je suis donc beaucoup plus confiant dans l'étude des signatures locales. Pour rappel, la plupart des allèles de résistances décrits à ce jour chez les moustiques sont délétères en l'absence d'insecticide. Leur présence dans les populations naturelles témoigne donc de pressions de sélection contemporaines.

Détecte-t-on, comme attendu, des signatures de sélection autour des gènes de résistance ? Ces signatures sont-elles les mêmes pour les différents allèles de résistances (relation entre l'intensité de la pression de sélection et celle du signal détecté) ? Sont-elles liées au type d'architecture et donc au type de sélection (duplications hétérogènes, sélection balancée, *vs* homogènes, sélection positive d'un effet quantitatif, *vs* mutations ponctuelles, sélection positive d'un effet qualitatif) ? Enfin, dans le cas de résistances impliquant des duplications, la taille de l'architecture génétique (en termes de nombre de copies et de taille des amplicons) affecte-elle l'étendue de la région génomique porteuse d'une signature de sélection (notamment à cause de l'arrêt de la recombinaison "localement[9]") ?

Là-encore une première approche consiste en l'étude des changements de diversité nucléotidique (π), mais cette fois le long du génome, par exemple en comparant les individus résistants aux sensibles. Des approches populationnelles sont tout à fait envisageables et sans doute préférables eu égard aux distances phylogénétiques (voir ci-après), mais par manque de temps j'ai préféré une approche par contrastes pour une analyse préliminaire. J'ai donc commencé par comparer les changements de diversité nucléotidique dans la région du locus *ace-1* entre individus sensibles et résistants, d'abord chez *An. gambiae s.l.* puis chez *Cx.*

[8] Mais qui ne tente rien, n'a rien, non ?
[9] Pour *Ester*, les 40 copies de 100 kb correspondent en fait à 4 Mb ; pour *ace-1,* 5 copies de 400kb correspondent à 2 Mb…

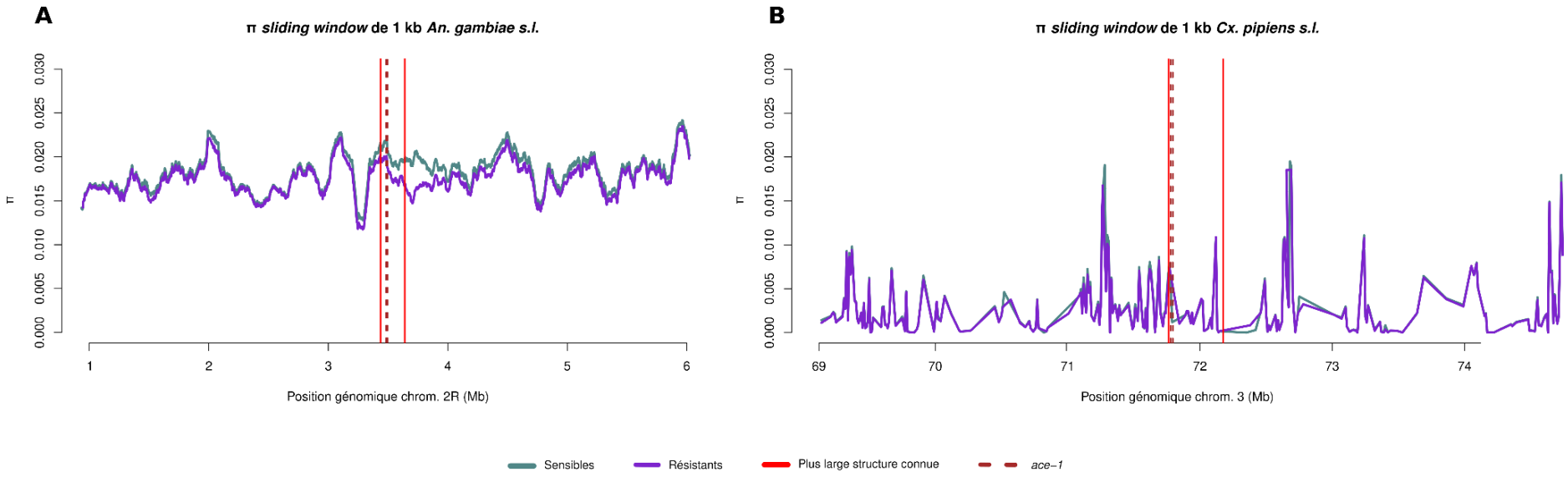


Figure D.1 : Valeurs π calculées par fenêtres glissantes d'1 kb sur le chromosome portant *ace-1* chez *An. gambiae s.l.* (A) et *Cx. pipiens. s.l.* (B). Réalisé depuis les données populationnelles du *An. gambiae 1000 genome project* et du *PipPop Project*. π a été calculé séparément pour les individus résistants (portant au minimum un allèle *ace-1* R) en violet et les individus sensibles en cyan. La position du gène cible (pointillés) et les bornes des plus larges duplications (lignes continues) identifiées pour chaque espèce sont indiquées par les lignes verticales.

pipiens s.l. (Fig. D.1 A et B). Comme vous pouvez le voir, j'ai été déçu: on constate une très légère baisse de diversité chez les résistants par rapport aux sensibles chez *An. gambiae s.l.*, mais pas chez *Cx. pipiens s.l.*. Rétrospectivement, ce résultat n'est peut être pas si étonnant, et mes encadrants et moi-même avons peut-être été un peu naïfs de penser que nous pourrions observer un signal, puisque nous nous intéressons à la sélection de larges variants structuraux. Notre naïveté est toutefois liée à une quasi-absence de littérature abordant ce sujet. Les attendus de π sur ce genre de variants ne sont donc pas si évidents à formuler. Comment a pu évoluer la diversité nucléotidique dans le cas de la sélection de variants liants des copies divergentes et sous sélection balancée, comme dans le cas des allèles *ace-1* D ? Ou bien dans le cas de variants associant de nombreuses copies, comme les allèles *ace-1* R^x ou les estérases ? La façon même dont est calculé π pourrait être à remettre en cause ici : en étudiant des régions dupliquées alignées sur un génome de référence non-dupliqué on fausse totalement les estimations, puisque les différents amplicons s'alignent en un seul endroit et donc sur un segment beaucoup plus court que leur taille réelle. Peut-être serait-il donc plus correct d'adapter le calcul de π au sein des régions dupliquées pour prendre en compte la taille cumulée des amplicons ? De même, comment séparer les "vrai" SNPs des pseudo-SNPs générés par l'empilement de séquences issues de différents amplicons ? J'explorerai ces questions sous peu.

J'ai donc opté pour le moment pour une autre approche en calculant le F_{st} entre individus sensibles et résistants et en espérant détecter une structuration de la diversité génétique, d'abord chez *An. gambiae s.l.*, puis chez *Cx. pipiens s.l.* (Fig.D.2)[10]. Dans le cas d'*An. gambiae s.l.* (Fig. D.2.A), on voit une claire différenciation sur une zone de près d'1 Mb centrée sur *ace-1*. Chez *Cx. pipiens s.l.,* le signal est plus bruité mais tout de même présent, et semble s'étendre sur une zone encore plus large (Fig. D.2.B). Ce résultat est remarquable, puisqu'il montre que dans les deux complexes, on arrive à différencier les résistants des sensibles dans une large zone autour du gène sélectionné malgré i) le polymorphisme des allèles de résistance au locus *ace-1*, ii) la non prise en compte de la structure des populations, et iii) les larges échelles phylogénétiques.

II.2. Dynamique de la résistance et signaux de sélection: un cas plus local.

Avec les données du *PipPop Project* et du *An. gambiae* 1000 *Genome Project*, je dispose d'un outil idéal pour répondre à beaucoup de ces questions, mais il persiste un problème : la

[10] Pour citer une personne qui restera anonyme (un petit barbu quasi-suédois mais avec un fort accent melgorien) : "Tentes le coup, fais ça comme un gros bourrin".

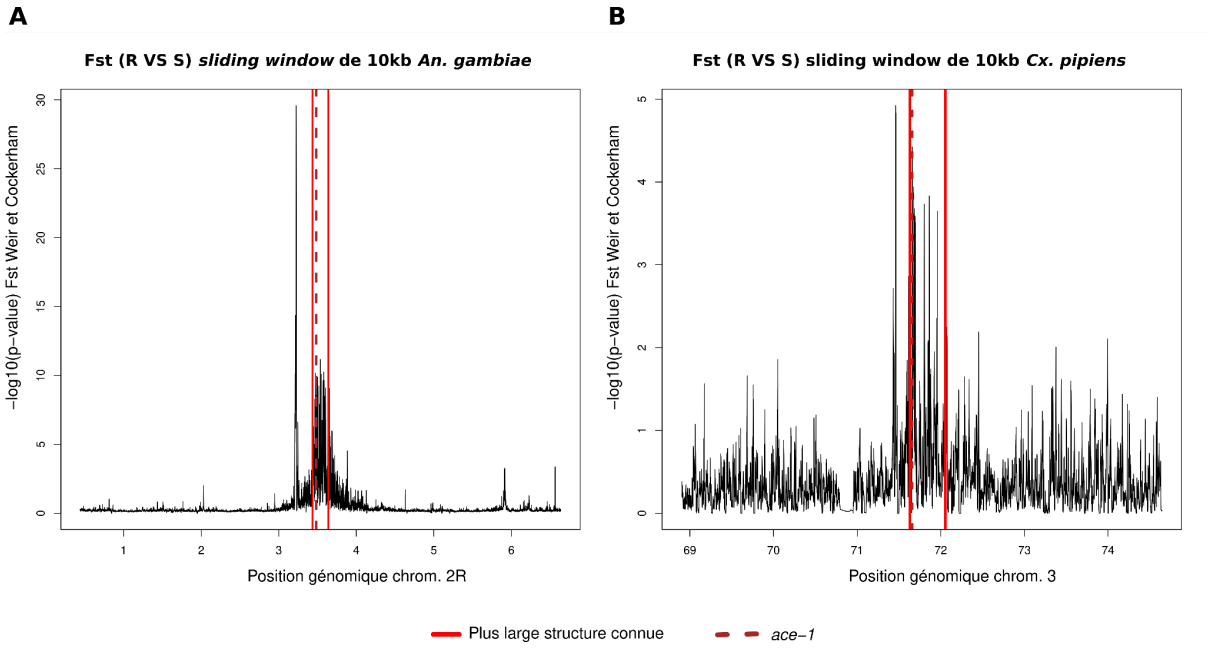


Figure D.2: Valeurs de F_{st}* calculées par fenêtres glissantes sur le chromosome portant *ace-1* chez *An. gambiae s.l.* (A) et *Cx. pipiens. s.l.* (B). Réalisé depuis les données populationnelles du *An. gambiae 1000 genome project* et du *PipPop Project*. La position du gène cible (pointillés) et les bornes des plus larges duplications (lignes continues) identifiées pour chaque espèce sont indiquées par les lignes verticales.
*Pour mieux séparer le bruit du signal, les valeurs reportées correspondent en réalité au $-\log_{10}$ des *p*-values calculées sur des *Z*-scores dérivés des valeurs de F_{st} ayant pour distribution théorique $\mathcal{N}_{(0,1)}$.

contrepartie du grand nombre de populations échantillonnées est que la taille de chaque échantillon est faible et peut être limitante pour certaines analyses. De plus, ces données ne permettent pas d'étudier l'aspect temporel de l'évolution conjointe de la dynamique des variants de résistance et de la diversité neutre. Fort heureusement, nous avons aussi développé notre propre jeu de données populationnel pendant ma thèse[11]. Il comprend 320 génomes de moustiques (*WGS Illumina paired end*, 150 pb, couverture 15X) collectés pendant près d'une quarantaine d'années le long d'un cline dans la région de Montpellier (Lenormand *et al.,* 1999 ; Labbé *et al.,* 2009 ; Milesi *et al.,* 2016). Il consiste en 15 individus par population, pour trois populations situées au Sud (zone traitée, Maurin), au centre (limite de la zone de traitement, St Gely et la distillerie de Prades-le-Lez, "Distill"[12]) et au Nord (hors de la zone traitée, St Bauzille-de-Putois) du cline, en prenant huit points temporels plus ou moins espacés, notamment à proximité de la date d'arrêt dans la région des traitements aux OP et CX, en 2007 (1986, 1995, 2002, 2005, 2008, 2010, 2016 et 2021 ; Fig.D.3). En suivant la dynamique neutre tant autour des gènes de résistance ainsi qu'à l'échelle du génome, je pourrai tester la robustesse des méthodes classiques d'inférence conjointes de sélection et de démographie. En disposant de données concrètes pour mesurer les effets locaux de la propagation des différents allèles de résistance en réponse aux variations environnementales, je devrais pouvoir évaluer si la taille des variants adaptatifs et/ou l'étendue de la sélection ont des effets différents sur la diversité génomique voisine. De plus, je pourrais aussi étudier dans quelle mesure et à quelle vitesse la diversité génétique est récupérée autour des gènes de résistance après l'arrêt de la pression de sélection insecticide.

Comme vous pouvez le voir, toutes ces questions sont encore très préliminaires. Elles nécessiteront du temps pour y répondre correctement, et c'est pourquoi elles sont l'objet de plusieurs demandes de financement, principalement appuyées sur le nouveau jeu de données montpelliérain, qui pourront, je l'espère, me permettre de réfléchir à tout ça tranquillement le temps d'un premier post-doctorat en Suède[13].

[11] Et malgré mon impatience, mes chers directeurs ont réussi à m'empêcher d'y toucher jusqu'à présent (les menaces physiques et psychologiques ont été cette fois efficaces).
[12] Devenue aujourd'hui une station d'épuration et un centre de tri de déchets, au grand désespoir de Pierrick.
[13] J'ai eu la chance d'avoir un avant goût de ce pays au cours de ma thèse, et j'aimerais l'explorer plus avant maintenant qu'elle est terminée !

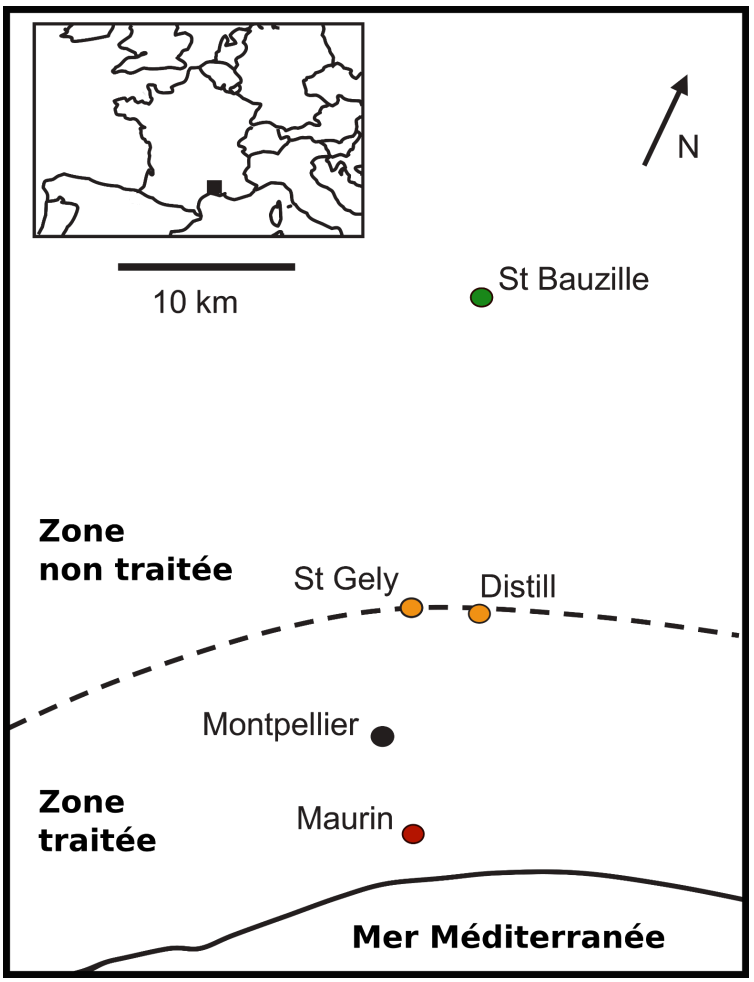


Figure D.3: Plan d'échantillonnage du jeu de données établi dans la région de Montpellier. Pour chaque population échantillonnée, les couleurs correspondent à l'intensité du traitement (du rouge, fort, au orange, moyen, et au vert, aucun traitement) et la période d'échantillonnage est indiquée. Modifié depuis Labbé *et al.* (2007).

III. Conclusion Générale

- " Ouah, mais c'est galère d'écrire une conclusion de thèse !"
- "Ah oui, et en plus t'es tout seul sur ce coup mon gars, ça doit être ce que tu veux raconter, toi.[14]"

Ma thèse coïncide avec le développement et la généralisation de nouvelles méthodes d'étude en biologie évolutive : les approches de génomique et de bioinformatique. Elle marque donc un changement majeur dans la recherche menée dans notre laboratoire, similaire à celui qui a pu s'observer avec le développement du génotypage par PCR et fragment de séquences, succédant à celui basé sur les allozymes[15]. En permettant la caractérisation de la taille, des points de cassures, du contenu en gènes embarqués et du nombre de copies des allèles dupliqués, elle a permis de répondre à des questions restées en suspens pendant longtemps. Ce faisant, elle a également ouvert la porte à une kyrielle d'interrogations que j'ai détaillées dans cette discussion : comment expliquer la diversité de ces duplications ? Quel a été l'impact de leur sélection sur la diversité génomique, et surtout comment parvenir à le mesurer ? Si j'ai bien l'intention de répondre à certaines d'entre elles, d'autres, j'en suis persuadé, trouveront leur réponse par de nouvelles idées, des avancées technologiques, ou grâce à un regard neuf. À qui le tour ?

[14] Ce qui s'est avéré être une demi-vérité, puisqu'"on" m'a donné un petit coup de pouce pour me sortir de ma leucosélophobie.

[15] Merci au passage à Nicole Pasteur, Michel Ramond, Christine Chevillon, Denis Bourguet, Thomas Guillemaud, Thomas Lenormand, Mylène Weill et tout-e-s les autres pour avoir défriché la plaine (à ce stade là, on ne peut plus parler de chemin) !

Annexe: demande de financement Birgitta Sintring Foundation

Title: How environmental variation shapes genome diversity: from local architectures to global signature. A validation of models from real-word data

Ethical Consideration: All the work will comply with international rules and agreements. No other ethical or gender aspects have to be considered for the project.

Project Aim

Understanding how natural selection shapes adaptive responses *in natura* has been and remains one main challenge of evolutionary biology and ecology (Lewontin 1974; Endler 1986; Barrett and Hoekstra 2011). It is indeed natural selection that explains the match between the phenotypes (*i.e.* the observable characteristics in an individual resulting from the expression of genes) and their environment. The implications are broad (e.g., plant and animal breeding, resistance to antibiotics, conservation biology) and it has never been as relevant and urgent as today in the face of the 6th mass extinction.

Studying natural selection in natural populations is not an easy task. Estimating fitness (i.e. the reproductive success of an individual or a phenotype) is particularly difficult and hindered by many issues and confounding factors. For instance, genetic drift (*i.e.* stochastic changes in allele frequencies between generations) and trait plasticity (*i.e.* capacity of a given genotype to produce different phenotype in different environment) can alter the detection of signature of natural selection (Rausher 1992). Natural selection can vary at different spatial and temporal scales, so repeated sampling over adequate geographical areas and time periods is often required. Even when it is possible to detect and quantify the response to selection, it is generally more difficult to link selection to its actual causes *in natura* (the agents of selection), selective pressures being often multiple and confounded (Siepielski et al. 2009).

To circumvent these issues, indirect methods have been developed to detect the signature that strong selection events on focal loci let into the genomes (aka “selective sweep”) while controlling for confounding factors, as for instance demography (effective population sizes variation, migration). Also, such signatures are transients and are not expected to be maintained over long period of time. A given adaptation can raise to fixation into the populations and the advantage associated with the mutation would hence disappear. Similarly, the environment is constantly changing, both in space and time and so is the selective pressure. A mutation that would be favorable in a given environment is thus not expected to be advantageous in another. The original genetic diversity is thus expected to be recovered and the pace at which former signal for selection would get eroded depends on the mutation and recombination rate, the intensity of genetic drift and gene flow.

To discriminate between these different processes, it is necessary to infer genetic and demographic (demo-genetic) parameters together, with appropriate methods and models. However, these approaches are usually applied without *a priori* knowledge of the population dynamics and selection history, the point being precisely to infer their evolutionary and demographic history. And it is often impossible to connect the inferred events of selection with their proximal causes, as well as to precisely estimate their timing of occurrence. Adaptations to human-caused environmental variations (e.g. insecticide and antibiotic resistance (Whalon et al. 2008; Norris et al. 2015), heavy metal tolerance (Janssens et al. 2009) constitute precious systems for investigating the fitness response to environment changes: the agent of selection is easier to identify and quantify, and the potentially simple genetic determinants of the adaptive responses can be traceable in natural populations.

In this project, I propose to use resistance monitoring data collected over 40 years in the mosquito *Culex pipiens*, in Southern France as an ideal case to study how environmental variation shapes genome diversity. I will leverage extensive genomic resources to assess how the insecticide treatments and the resulting selection of resistance genes impacted the rest of the genome and its diversity. Such data would also provide us with a unique opportunity to test the accuracy of classical selection and demography inference methods, the location, timing and quantity of treatment insecticide used in the populations being known.

Background:

Resistance to organophosphate (OP) and carbamate (CX) insecticides in the mosquito *Culex pipiens* of the Montpellier area (Southern France) is a textbook example of adaptation to extreme selective pressure (Guillemaud *et al.* 1998; Labbé *et al.* 2007a and 2007b; Milesi *et al.* 2016). A yearly sampling effort has been maintained since the mid 80's along a 50 km deep, north-south transect, covering a gradient of heavily treated (southernmost) to non-treated (northernmost) locations. It allowed investigating the long-term dynamics of multiple insecticide resistance alleles in natural populations in response to variations of insecticide quantities used. These data revealed that different alleles of different genes conferring resistance were selected over time (Guillemaud et al. 1998; Labbé et al. 2009). This is explained by the direct relationship between the fitness conferred by the adaptive alleles and the environmental conditions: the advantage-cost trade-off associated with the different resistance alleles is not fixed, but depends directly on the intensity of the selection pressure (i.e. the quantities of treatment used to control the populations of mosquitoes, Milesi et al. 2016). By many aspects this system is thus ideal to answer our aim:

1. The small geographic scale of the study ensures no confounding effect from population structure or demography history. In addition, mosquitoes' effective population size being very large and the selective pressure on the resistance loci extremely strong, confounding effect from the genetic drift should be limited and *we can expect a strong signature of selection into the genomes*.
2. Mosquitoes have a short generation time (i.e. 10 to 15 generations per year) and the 40 years of the survey represents a significant evolutionary time of more than 300 generations in natural populations (~7,500 years in humans given a generation time of about 25 years). Such a number of generations is usually restricted to micro-organisms. *It will allow us to follow the evolution of the signature of selection into the genome "in live".*
3. The agent of selection (i.e. insecticide treatment campaigns) is well known and quantified and we have records about which type of molecules were used, how much, where and when. *It will allow us to match the dynamics of signature with variation in environment.*
4. In 2007, the use of OP and CX insecticides was banned, providing us with a unique *opportunity to study the evolution of former adaptations once the selective pressure is removed.*

This unique dataset would thus allow us to assess the resistance status of the mosquitoes population through space and time, to quantify the impact of the selection regime on the genome diversity globally and surrounding the resistance genes.

Project plan

Data acquisition. Recently the whole genome of 320 mosquitoes from Montpellier's collection were sequenced (Illumina paired end sequencing, 150 pb, 15X coverage): 15 individuals per population, three populations per year (located on the northernmost, center and southernmost parts of the cline) on eight sampling years (1986, 1995, 2002, 2005, 2008 ,2010, 2016 and 2021; Fig. 1).

Classical pipeline for DNA mapping and SNP calling will be used to characterize the genome-wide diversity and genetic architecture of resistance. Briefly, once cleaned, the raw reads will be mapped against *C. pipiens* reference genome using BWA-MEM software. Generated .bam files

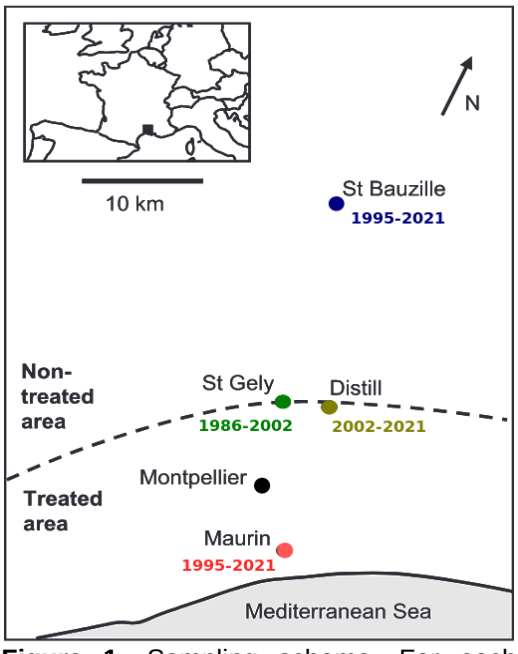


Figure 1: Sampling scheme. For each sampled population colors correspond to treatment intensity (from red, strong to blue no treatment) and period of sampling is indicated. Modified from Labbé et al. 2007

will be filtered and PCR-duplicates marked and removed using dedicated tools in the PICARD suits. SNPs calling and further filtering will be done using GATK software. Finally, I will use a method based on genome depth of coverage analysis and read mapping information (soft-clipped reads and abnormal read mate position) that I have developed during my PhD studies to identify and characterize the genomic architecture (copy number, breakpoints location, overall size and possible gene content) of the four targeted resistance genes (Claret *et al*. 2023 Biorχve).

Task 1: Genomic characterization of resistance dynamics.

I will assess the resistance status of Montpellier's populations over the years and across the treatment transect. *Ester* superlocus encodes for a generalist detoxifying enzyme whose overproduction allAows degrading the insecticide before they reach their target. Being a generalist detoxifying enzyme, we also expect other xenobiotics than those used for mosquito population control (for instance used in agriculture) to select for it. On the other hand, *ace-1* gene encodes for the target of the insecticide, the acetylcholinesterase (AChE1). Hence it is specialist and should only be selected by the treatment campaign. The resistance is achieved through a single point mutation that changes the conformation of the active site of the protein and impedes the binding of the insecticides. The resistance at these two major loci also involve extensive structural variation at the genomic level. For *Ester*, the overproduction of the detoxifying enzyme is achieved through a change in gene copy number (up to 40 copies) (Fig. 2A). For *ace-1*, heterogeneous duplications pairing a susceptible and a resistance copy of the *ace-1* gene segregates in the populations. I recently described them at the genomic level and showed that they span more than 400Kb and encompass seven other genes (Fig. 2B). Finally, I will also focus on two other genes, *vgsc* and *rdl*, known for being involved in resistance to other compounds mostly used in Mosquito's repellent and thus expected to be under lower selective pressure than *ace-1* and *Ester*.

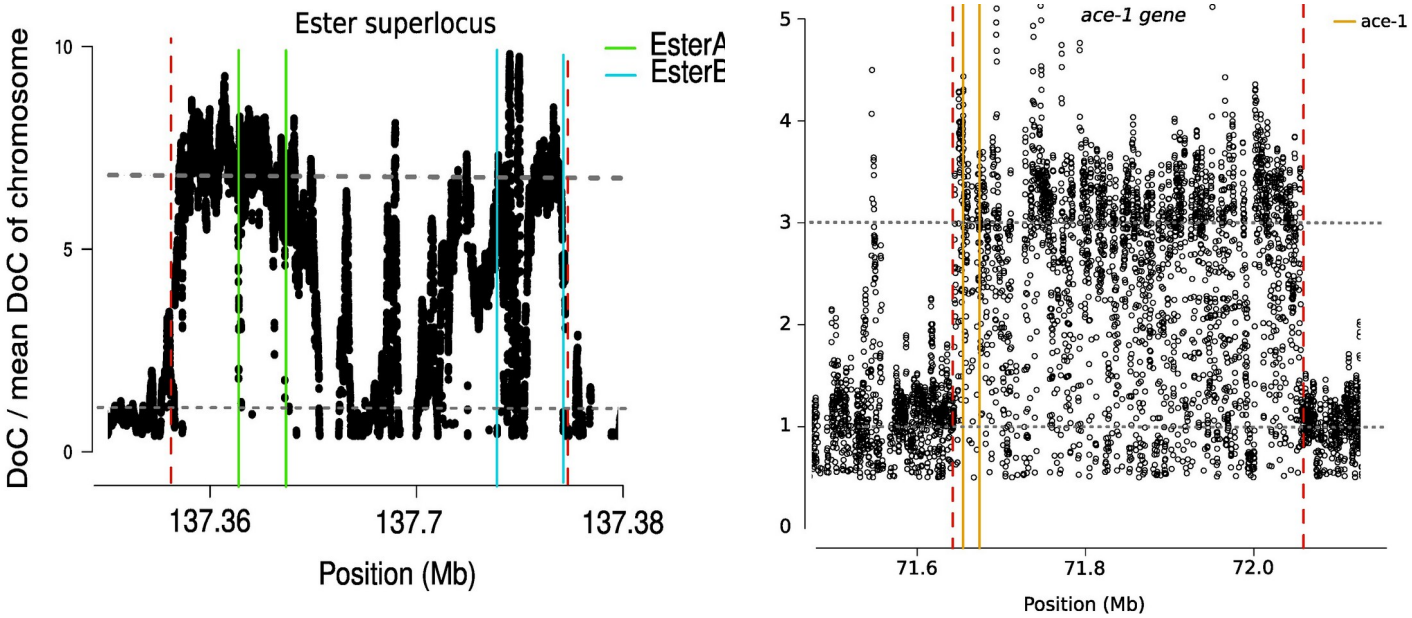


Figure 2: Structural Variants identified at two resistance loci. The change in copy number around the resistance loci is evidenced by the ratio of local coverage over the average coverage of the chromosome (1 for single copy). The breakpoints of the amplicons are shown in red dashed lines. The position of the resistance gene is shown in colored full lines. Left plot is for *Ester* superlocus and the right plot for *ace-1*.

Task 2: Impact of adaptation dynamics on surrounding diversity and model testing.

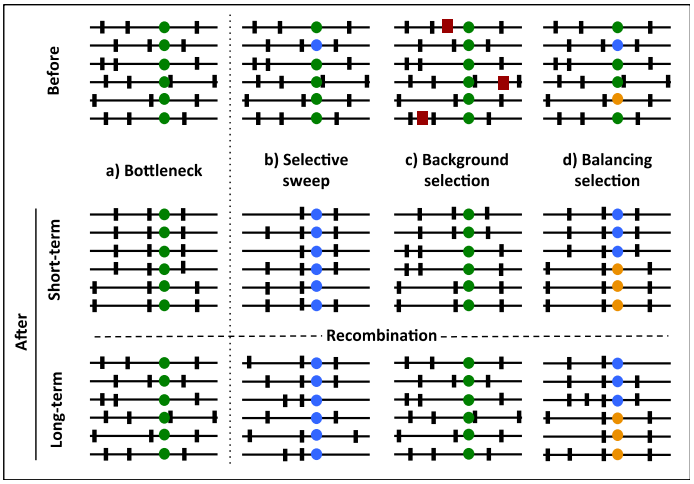


Figure 3: Demo-genetic effects on neutral variation. Six chromosomes are sampled before and after the selection or the demographic event, with the immediate effect (short-term) and the long-term effect of recombination. The different mutations are represented by black vertical lines (neutral), green circle (ancestral allele of the focal gene), blue and orange circles (advantageous alleles of the focal gene), and red squares (deleterious mutations).

Mutation is the ultimate source of polymorphism. Variations in allele frequencies in natural populations are under the influence of different evolutionary forces, either stochastic (history, demography or migration) or determinist (natural selection). Stochastic forces tend to affect the effective size of the population, and thus the diversity of the genome as a whole: a major reduction in effective size (i.e. bottleneck) significantly reduces polymorphism (Fig. 3a). By contrast, the effect of selection is generally more local: i) directional selection reduces polymorphism at the target loci (Fig. 3b), while ii) balancing selection (frequency dependence, overdominance, and variations in selective pressures) prevents the fixation of any allele. Selection can also affect allele frequencies at neighboring loci, in inverse proportion to the recombination rates (i.e. genetic hitch-hiking), through i) the selective sweep of an advantageous mutation (Fig. 3b); ii) purifying selection, eliminating a deleterious mutation (Fig. 3c); iii) or balancing selection (Fig. 3d). In all cases, recombination will tend to restore polymorphism in the long term (Fig. 3, bottom line).

Task 1 will provide unprecedented data to understand how a changing environment affects the adaptability of a species, both through demography and adaptive responses. Following the neutral dynamics both around the resistance genes and at the genome scale, I will thus be able to test for the accuracy of classical selection and demography inference methods. As we have actual data to measure the local effects of the spread of the various resistance alleles in response to environmental variations, I can finally assess whether i) the architecture of these adaptive mutations (point mutations or large-scale duplications) and/or ii) the extent of selection, have different effects on the neighboring genomic diversity. I will also be able to measure the effects of the population size reductions induced by the insecticides, reconstructing the 40+-year demographic history of the mosquito populations in relation with the varying intensities in insecticide selection pressure. Finally, I will use the ban of the OP and CX insecticides in 2007 to measure how much and at what speed original genetic diversity is recovered around the resistance genes.

Anticipated results:

Our project should shed a new light on how environmental variations (here the insecticide treatments) affect the genomic diversity both, locally around the mutations responsible for the adaptations (here the resistance genes) but also globally through a change in demography. It will also be, to my knowledge, the first study to confront models to actual data over so many generations while incorporating both spatial and temporal variation of the environment in natural populations. With a more applied perspective, vector control is designed to negatively affect mosquito population demography: the magnitude of these effects on the population genetic diversity, and thus its adaptability, will thus also be assessed, which will provide invaluable information for the agencies in charge of pest control, to help them design more effective and sustainable strategies.

Time Plan and Project Organization

	2024				2025			
Activity	1	2	3	4	1	2	3	4
Project meeting and orientation								
Genomic data curation								
SNPs calling and Variant calling								
Dynamics of resistance								
Genomics impact of resistance								
Paper writing								

I will focus on this project and regular meetings will be held with all the members of the research group to discuss the project and its developments. The results will be published in international open access scientific journals with peer review committees and will also be presented at conferences. Data will be made accessible in public repositories.

Equipment and infrastructures: Bioinformatic data analysis will be done using Uppsala multidisciplinary center for advanced computational science (UPPMAX). No new equipment will be required for this project. *Please note that I am familiar with that computing environment, which will ease and speed up the start of the project.*

Economy and budget: All the data sequencing necessary for this project has already been performed, therefore the only additional costs concern a computer, mandatory IT costs at IEG and workshop and conference attendance-related costs (respectively ~25000 sek, 11200 sek and ~40,000 sek). These costs will be covered by faculty funds granted to Dr. Milesi.

Working Environment: The study of local adaptation and the role of structural variant in evolution is well represented at Uppsala University and in particular at the department of Ecology and Genetics (e.g. Pr. V. Katju or Dr. A. Husby). It will allow me to broaden my understanding of my research topic by interacting with scientists working with different model species and on related questions. The research developed in Dr. Milesi's group aims at understanding the role of structural variants in evolution with an emphasis on short-term evolution and adaptation. Dr. Milesi is familiar with the model of resistance to insecticide in mosquitoes and was an external member of my PhD committee. Please note that I have already spent two months in Dr. Milesi's group at IEG as a visiting PhD student in 2023.

References

- Barrett, R. D., & Hoekstra, H. E. (2011). *Nat. Rev. Gen*.
- Claret J.L., Di-Liegro M., et al. (2023).bioRxiv.
- Endler, John A. (1986). Princeton Univ. Press.
- Guillemaud, T., Lenormand, *et al.* (1998). *Evolution.*
- Janssens, T. K., Roelofs, D., & Van Straalen, N. M. (2009). *Insect Science*.
- a. Labbé, P., Berticat, et al. (2007) *PLoS Gen*.
- b Labbé, P., Berthomieu, et al. (2007). *Mol. Biol. and Evol.*
- Labbé, P., Sidos, N., Raymond, M., & Lenormand, T. (2009). *Genetics*.
- Lewontin, R. C. (1974). New York: Columbia Univ. Press.
- Milesi, P., Lenormand, T., Lagneau, C., Weill, M., & Labbé, P. (2016). *Mol. Ecol*.
- Rausher, M. D. (1992). *Evolution.*
- Siepielski, A. M., DiBattista, J. D., & Carlson, S. M. (2009). *Ecol. Letters*.
- Whalon, M. E., Mota-Sanchez, D., & Hollingworth, R. M. (Eds.). (2008). Cabi.

A plusieurs, on est meilleur: du rôle des duplications dans l'adaptation aux insecticides chez les moustiques.

Les duplications sont des variants structuraux à l'origine d'une importante source de diversité génétique sur laquelle la sélection naturelle peut agir. Un exemple contemporain de duplications adaptatives est largement étudié chez les moustiques. Suite à l'utilisation à large échelle d'insecticides, une mutation ponctuelle du gène *ace-1* (allèle R) s'est rapidement propagée dans plusieurs espèces de moustiques, où elle est apparue indépendamment. L'allèle (R) permet la résistance à ces insecticides, mais est très délétère en leur absence par rapport à l'allèle sensible (S). Des allèles dupliqués associés à un large éventail de phénotypes représentant différents compromis de valeur sélective ont été découverts: des duplications homogènes associant plusieurs copies R, et des duplications hétérogènes liant des copies R et S, ou allèles D. Nous avons étudié ces allèles en parallèle dans deux complexes d'espèces très divergents (~145 Ma, 1G générations), *Culex pipiens s.l.* et *Anopheles gambiae s.l.*. Grâce à un jeu de données inédit, nous avons caractérisé de nombreuses structures génomiques à l'échelle mondiale chez *Cx. pipiens s.l.*, tandis que nous avons mis en évidence dans des populations sauvages d'*An. gambiae s.l.* un fort polymorphisme d'allèles D partageant tous une même structure génomique. Nos résultats soulignent l'évolution parallèle de ces complexes d'espèces face à la pression de sélection des insecticides, les mêmes types de duplications engendrant des phénotypes similaires. Ces travaux, dont certains restent en cours, nous ont permis de mieux comprendre l'origine de la diversité et le rôle adaptatif de ces duplications, et ont posé les premiers jalons pour comprendre l'impact de leur sélection sur la diversité génomique à différentes échelles.

Mots clefs : *Résistance aux insecticides, duplications de gènes, évolution parallèle, génétique de l'adaptation.*

Better together: the role of duplications in adaptation to insecticides in mosquitoes.

Duplications are structural variants at the origin of an important genetic diversity upon which natural selection can act. A contemporary example of adaptive duplications is studied in mosquitoes. Following the massive use of insecticides, a point mutation in the *ace-1* gene (R allele) quickly spread in several mosquito species, where it appeared independently. This allele (R) allows resistance to these insecticides, but is highly deleterious in their absence compared to the susceptible one (S). Duplicated alleles associated with a wide range of phenotypes representing different fitness trade-offs were discovered: homogeneous duplications associating several R copies, or heterogeneous duplications linking R and S copies, or D alleles. We studied the parallel evolution of these alleles in two species complexes that diverged ~145 Ma, *Culex pipiens s.l.* and *Anopheles gambiae s.l.* Using an unprecedented dataset, we characterised numerous genomic structures on a global scale in *Cx. pipiens s.l.*, and evidenced in wild populations of *An. gambiae s.l.* a high polymorphism of D alleles, all sharing the same genomic structure. These results highlight the convergent response in both species complexes to face the same insecticide selection pressure, where the same types of duplications generate similar phenotypes. This work, some of which remains in progress, has allowed us to better understand the origin of diversity and the adaptive role of these duplications, and has paved the road for understanding the impact of their selection on genomic diversity at different scales.

Keywords: *Insecticide resistance, gene duplications, parallel evolution, adaptation genetics.*