Rôle de la régulation génique dans l'adaptation : approche par analyse comparative du transcriptome de drosophile

François Wurmser

UNIVERSITE PARIS-SUD 11

ÉCOLE DOCTORALE Gènes, Génomes, Cellules

Biologie évolutive

THÈSE DE DOCTORAT

Pour obtenir le grade de Docteur en Sciences de l'Université Paris 11

par

François WURMSER

Rôle de la régulation génique dans l'adaptation : approche par analyse comparative du transcriptome de drosophile

Le 25 novembre 2011

Composition du jury :

Président du jury :	Dominique De Vienne
Rapporteurs :	Frédéric Fleury
	Bruno Lemaître
Examinatrice :	Carole Smadja
Directrice de thèse :	Dominique Joly
Co-directrice de thèse :	Catherine Montchamp-Moreau

Remerciements

Enfin, j'arrive à la meilleure partie. La plus belle à écrire, la plus personnelle.

Je remercie d'abord les membres du jury de ma thèse, notamment les deux rapporteurs, Frédéric Fleury et Bruno Lemaître. Mais aussi Dominique De Vienne et Carole Smadja.

Je remercie Dominique Joly, pour m'avoir supporté, dans les hauts et les bas, avoir corrigé mes grossièretés (saperlipopette!), et m'avoir soutenu de bout en bout, de A à Z dans cette thèse, de mes doutes à mes certitudes, du début à la fin, du haut en bas comme de gauche à droite. Merci!

Je remercie ensuite Catherine Montchamp-Moreau toujours à l'affut d'une idée intéressante, d'un balayage sélectif impromptu. Elle a suivi elle aussi cette thèse sur sa longueur, et l'a co-encadré avec doigté. Merci!

Je remercie Arnaud Le Rouzic, pour sa patience, sa patience et sa patience. Et puis ses compétences multiples dans le domaine informatique, statistique. Et son goût pour le saucisson et les rillettes. Merci!

Je remercie David Ogereau. Parce que son plus beau surnom reste "l'incontournable David", lui qui peut dépanner tout le monde et pour tout, et surtout toujours en souriant. Il rigole même à mes blagues. Et c'est quand même lui qui m'a mis la puce à l'étrier. Merci!

Je remercie les stateux avec qui nous avons collaboré, pour avoir réussi à faire du béotien que j'étais (et que je suis encore), un petit quelque chose de statistiques, saupoudré d'un petit peu de R, arrosé d'estimations de variances. Jean-Jacques à coup sûr, mais surtout Tristan. Merci!

Je remercie Béatrice Denis, la mésestimée. Parce que maman Béatrice m'a approvisionné en petit gâteau pendant toute ma rédaction. En plus, elle se fâche même pas quand je me moque d'elle. Merci!

Je remercie Delphine Sicard, ma tutrice de Thèse. Un tuteur, par définition, ça soutient (demandez à votre plante préférée)! Merci!

Je remercie Gildas Lepennetier, parce que sans lui, je ne me serai pas remis à UR. Pour aimer

aller manger un Kebab en catastrophe à 9h du soir. Et parce qu'il aime râler, ce qui me donne l'occasion de le taquiner un peu. Merci !

Je remercie tous les thésards qui sont passés au labo pendant cette thèse, parce qu'on rame tous ensemble dans la même galère (vous êtes sûrs que c'est par là ?). Silvan et Héloïse, conseillers en thèse +1. Marine, blonde vénitienne. Céline. Lucie (surtout pour ses petites blagounettes dessinées sur les paillasses). Quentin (tour de babel en boîtes de cônes !), et Bastien, son alter ego. Gwenaëlle, Thibaud, Delphine, Mélanie, Fabrice, Antoine, François, Magaly, Floriane, Marion, Julie, Pierre, Antoine, Mohamed, Mariangela, Jessy (qui ne connait la France que enneigée), Nicolas, Clara, Chris (music maestro !). Merci !

Tous les compagnons du midi : Amir, notre égyptien athée (pardon Franco-égyptien maintenant !), Émilie, Aurélie (et ses histoires de chauffage), Isa (la seule à connaître le menu de la cantine avant qu'on y arrive), Céline (délirante et délurée). Sylvie, qui en plus nourrit avec Magalie nos mouches à la petite cuillère. Pierre G, sans qui je ne serais pas appelé "François le Français" plus d'une fois par jour, et son compagnon geek, Julien. Fred et son goût pour le beau temps. Nancy. JB, l'ours qui a fait de moi un homme libre. Jonathan. Sylvie (la com'). Merci !

En vrac : Claude, Jean-Luc, Didier, Marie-Louise, Jean, Bastien, Sylvie (gestionnaire qui ne perd pas le nord), Véronique, Gaëlle (Gaïeïlle selon les jours), Amandine (une tite pétanque ?), Odile, Patrick (notre olivier de la paix à nous), Hélène, Florence, Jean-Christophe, Jean-François, Isabelle. Wilfried Haerty. Michele Schiffer, pour m'avoir initié à l'argot australien (Gidday mate !). Francesco Catania, *paramecium* power ! Et puis le labo. Tout le labo. Du bâtiment 13 au bâtiment 5, la mer rouge nous a séparé. Mais nous sommes toujours le LEGS, labo néo-historico évolutif. Merci !

Côté finances : le CNRS, et plus particulièrement l'INEE qui m'a financé moi. L'ANR, qui a financé mes manips via le programme AdaptAnthrop. L'IDEEV. GATC Biotech pour le séquençage haut-débit (Yadhu Kumar et Benjamin Moingeon), TAGC pour les puces (Béatrice Loriod, Michel Piovant). Merci !

L'École doctorale Gènes, Génomes, Cellules et l'université Paris-sud.

Pour le matériel biologique : l'INRA de Gotheron, Bruno Le Rü, Roland Allemand, Daniel Lachaise. Merci !

Je remercie enfin Pierre Capy, qui a tout fait pour que je puisse faire cette thèse dans son laboratoire. Merci !

Je remercie mes amis sur une roue : Léo, Tonio, Élise, Krokko, Florent, Xav, Yann, Manu,

Violaine, Hélène, Pauline, Arthur, Greg, Fabrice, Pierre, Olivier, Chouch, Alain, Zip, Teddy, Didier, la bande de ptits Suisses, Aline l'Africaine, Baptiste le Chinois, Seb, Amo alias Kaerael. Mes amis bloqués au vélo : Seb (on a commencé en même temps, on va finir en même temps). Anthony. Merci !

Je remercie ma famille. Notamment mes parents néo-Rochelais, toujours là pour moi. Un jour il faudra que j'apprenne à inverser les rôles, et que je sois là pour eux. Mes grands parents et leurs tomates cerises. Mon grand frère, émigré doré. Mes cousines : Noëlle (accompagnée du Nagoya et du Petit Josselin), Cécile la dormeuse et son Jules-ien. Et son frère Sylvain. Merci !

Table des matières

Table des figures

Liste des tableaux

I Introduction

Il fut un temps, où l'on pensait qu'avec l'ensemble des gènes, on comprendrait la globalité du fonctionnement des organismes. Puis cette attente s'est transférée vers les séquences de génomes complets. Chaque étape a montré une complexité des interactions toujours grandissante, et mis en évidence les rôles cruciaux de nombreux aspects des organismes, non seulement dans leur fonctionnement, mais aussi dans leur évolution. L'intérêt pour décrypter le rôle de la séquence en soi dans les fonctions biologiques est toujours présent, mais cette dernière décennie, l'impact de l'expression génique dans l'aspect évolutif des organismes a pris une place grandissante. Quel rôle l'expression des gènes joue-t-elle précisément dans l'adaptation ? Dans la différenciation des populations ? Des espèces ? Cette introduction, à partir de la notion d'espèce, nous amènera peu à peu à montrer l'importance de l'expression génique, tel qu'elle a été révélée en relation avec la variabilité des conditions naturelles.

1.1 L'espèce au centre de l'évolution

La notion d'espèce est à la base des problématiques évolutives actuelles, que ce soit pour la détermination des espèces (approches de phylogénie moléculaire et morphologiques), des processus qui conduisent à leur séparation (différenciation des populations, spéciation), ou la compréhension des relations qu'elles entretiennent entre elles (co-évolution, relations de mutualisme / parasitisme, prédation, compétition), ou plus généralement avec leur environnement (adaptation, écosystèmes). Cette question est également au cœur de cette thèse, car ce que nous avons cherché à étudier ici, ce sont les différenciations de l'expression, parfois en relation avec l'adaptation, conduisant éventuellement à une séparation des populations suffisante pour conduire à la spéciation. Sans oublier cependant la part stochastique des variations de l'expression entre individus, taxons.

La notion d'espèce comporte une part subjective, due à la fois aux nombreuses définitions de

l'espèce au cours des âges, et à l'intérêt de la personne pour les espèces en question. Un agriculteur ou un horticulteur n'aura pas la même compréhension de l'espèce qu'un biologiste moléculaire.

La notion d'espèce n'est pas nouvelle, on peut remonter jusqu'à Platon pour une première approche ($IV^{ème}$ siècle avant J.C.). Celui-ci définit des "types" différents, et constate leur incapacité à s'hybrider, et définit par conséquent l'espèce comme une forme idéale vers laquelle est contrainte chaque animal (il ne commente pas de cas connu d'hybridation possible, comme le tigron, le mulet, etc). La notion d'espèce est ancestrale à celle d'évolution, et elle a été abordée indépendamment de celle-ci pendant la majeure partie de notre histoire. Jusqu'au $XVII^{ème}$ siècle, on considère qu'il y autant d'espèces à l'heure actuelle qu'à la création du monde. Cette conception est essentiellement due à l'omniprésence de la religion dans la science. Lorsque Linné (1707-1778) cherche à classer les espèces, il cherche en fait à reconstituer les groupes créés aux origines. Il se base sur des critères morphologiques, et notamment en ce qui concerne les animaux, sur les variations morphologiques de l'appareil génital. Ces classifications des êtres vivants ont beaucoup de sens dans un contexte fixiste, mais dans un contexte évolutionniste, le problème devient beaucoup plus complexe (Lherminier et Solignac, 2005).

Il faut attendre le $XVIII^{ème}$ et les premiers évolutionnistes (Lamarck, Darwin) pour voir apparaître la définition basée sur l'interfécondité. Des êtres vivants constituent une espèce, si et seulement si ils sont capables de produire des descendants fertiles. C'est la définition biologique de l'espèce, la seule vraiment connue du grand public. Ernst Mayr l'a énoncée ainsi :

> *"Species are groups of actually or potentially interbreeding natural populations, which are isolated from other such groups."* "Systematics and the origins of species", Mayr (1942)

Selon cette définition, tous les êtres vivants capables de produire des descendants interfertiles sont de la même espèce. *A contrario*, ne sont donc pas de la même espèce tous les êtres vivants qui sont isolés d'un point de vue reproductif, et donc incapable de produire des descendants fertiles. Il existe cependant diverses définitions de l'espèce, et si elles sont toutes basées sur la cohésion intra-groupe et l'isolement inter-groupes, elles correspondent à différentes approches : biologique (voir Ernst Mayr Mayr (1942)), écologique (Rice, 1987; Lherminier et Solignac, 2005; Abbott *et al.*, 2008), phylogénétique (Mishler et Brandon, 1987). La définition écologique s'appuie sur les interactions de l'espèce avec son environnement. Selon cette définition, constituent une même espèce tous les êtres vivants présentant une unité écologique, occupant la même niche écologique,

la même place dans l'écosystème. Du point de vue phylogénétique, on définit généralement un degré de ressemblance / dissemblance pour constituer des groupes distincts. Cette mesure peut être basée sur des critères morphologiques, moléculaires, voire même s'appuyer sur les deux. Plusieurs auteurs se sont essayés à des définitions consensuelles de la notion d'espèce. Parmi ceux-ci, je citerai celle énoncée par Rama Singh :

> *"A species is a group of actually or potentially interbreeding natural populations with a sufficient degree of genetical, ecological, or spatiotemporal isolation that would maintain the developmental or reproductive compatibility (or cohesiveness) of the gene pool within groups and incompatibility (or distinctness) between groups."* "Toward a Unified Theory of Speciation", in "Evolutionay Genetics. From molecules to morphology", Singh (2000)

Cette définition prend en compte des données moléculaires, écologiques et phylogénétiques, articulées autour de la notion d'interfertilité. Certes, elle n'a pas la simplicité de la définition biologique, mais elle se veut plus universelle, applicable à l'ensemble du monde vivant.

1.2 Isolement reproducteur, définition et caractéristiques

La définition biologique de l'espèce mène naturellement à la notion d'isolement reproducteur. Cet isolement fait référence à l'indépendance des pools de gènes, c'est-à-dire à l'absence de flux de gènes entre groupes, et donc à l'évolution indépendante des organismes, en terme de fréquences alléliques, mutations (Wu et Ting, 2004),... Cette notion, simple en apparence comporte de nombreuses facettes, qui la rendent plus complexe que de prime abord. En effet, l'isolement reproducteur peut être prézygotique ou postzygotique, c'est-à-dire intervenant avant ou après la fécondation (Abbott *et al.*, 2008). Durant cette thèse, nous avons examiné l'influence de l'expression des gènes sur la stérilité ou l'inviabilité hybride (isolement postzygotique).

1.2.1 Isolement prézygotique

Au niveau prézygotique, les incompatibilités peuvent être écologiques, c'est-à-dire liées à l'habitat (par exemple plantes hôtes différentes pour des champignons), liées aux vecteurs pollinisateurs (attractivité plus ou moins grande de certaines plantes pour les insectes), ou encore à des décalages temporels des périodes de fertilité (floraison décalée de quelques heures, ou de quelques semaines, période de reproduction décalée de quelques mois, voire même une population qui se reproduit de

façon bisannuelle). Pour ce dernier cas, on peut citer l'exemple de l'araignée *Araneus diadematus*, qui dans les régions les plus nordiques se reproduit non plus annuellement mais une fois tous les deux ans, créant un isolement reproducteur partiel (Johannesen et Toft, 2002). L'isolement prézygotique peut aussi être dû à des problèmes de reconnaissance des partenaires sexuels (Andersson, 1994; Arbuthnott, 2009). Cela peut être via l'évolution des signaux de reconnaissance, qu'ils soient visuels (couleur), auditifs (chants de cour), tactiles ou chimiques. Les chants de cour ont une importance qui dépend de l'espèce, y compris à l'intérieur même d'un groupe d'espèces proches. La femelle de l'espèce spécialiste *Drosophila sechellia* accepte moins un mâle qui produit le chant de cour d'une autre espèce qu'un mâle qui ne produit pas de chant de cour, alors que *D. simulans* et *D. melanogaster* acceptent mieux un mâle qui produit un chant de cour d'une autre espèce qu'un mâle qui n'en produit pas, suggérant un rôle fort de ce caractère dans l'isolement de *D. sechellia* (Tomaru *et al.*, 2004). Les incompatibilités mécaniques liées à l'accouplement peuvent aussi jouer un rôle crucial dans l'isolement prézygotique (Eberhard, 1996). Ainsi, Richmond *et al.* (2011) ont montré que des différences de taille chez le lézard provoquaient des difficultés mécaniques lors de l'accouplement. Ils ont ensuite montré par modélisation que ce facteur peut avoir une importance majeure dans les premières étapes de spéciation. Enfin, l'isolement prézygotique peut être ultimement lié à des problèmes de reconnaissance entre l'ovocyte et le spermatozoïde (Dobzhansky, 1951; Palumbi, 2009).

1.2.2 Isolement postzygotique

L'isolement peut aussi se situer au niveau post-zygotique, c'est-à-dire qu'il y a formation du zygote, mais les hybrides nés du croisement seront soit non viables, soit stériles. La drosophile est un animal de choix pour les études sur l'isolement reproducteur : faible temps de génération, croisements souvent possibles et relativement faciles entre espèces proches. Les gènes d'isolement reproducteur actuellement décrits, l'ont généralement été chez la drosophile. Le plus célèbre est le gène *Odysseus*, dont la co-introgression avec une région avoisinante de *D. mauritiana* vers *D. simulans* crée la stérilité (Perez *et al.*, 1993).

Les hybrides suivent généralement la "Règle de Haldane" (Haldane, 1922). Elle postule que si un seul des deux sexes est stérile ou non viable, ce sera le sexe hétérogamètique. Différentes théories ont été avancées pour expliquer la règle de Haldane :

- la théorie de dominance liée à la présence d'interactions délétères récessives sur le chromosome X (dans un système XY)

– l'évolution plus rapide des gènes dont la fonction / l'expression est biaisée vers un sexe, par rapport aux gènes sans biais de sexe (Jagadeeshan et Singh, 2005; Musters *et al.*, 2006). Ce patron est plus marqué pour les gènes intervenant dans les fonctions mâles

Ce second processus ne peut expliquer à lui seul la règle de Haldane, car les animaux dont le mâle est homogamétique présentent également une évolution plus rapide des gènes mâles (Malone *et al.*, 2006).

Une autre caractéristique du dysfonctionnement des hybrides est l'impact majeur du chromosome X. Là aussi, plusieurs théories non exclusives expliquent cette observation :

– la stérilité ou la létalité sont souvent dues à des incompatibilités entre le chromosome X et les autosomes (Orr et Irving, 2001). Il est donc cohérent que le X porte plus de locus liés à une incompatibilité qu'un autosome (voir premier point sur la règle de Haldane)
– une hypothèse liée à des distorteurs de sexe-ratio cryptiques a été proposée (Tao *et al.*, 2001; Orr et Irving, 2005), mais n'a pu être confirmée (Presgraves, 2008)
– des perturbations de la compensation de dosage de l'hémizygote chez l'hybride (Coyne et Orr, 1989)
– des perturbations de l'inactivation du chromosome X en début de spermatogenèse (Hense *et al.*, 2007)

Dans notre étude, nous allons examiner l'expression chez des hybrides en comparaison avec leurs espèces parentales. Cette expérience devrait nous permettre de révéler des gènes impliqués dans les incompatibilités chez les hybrides, en lien avec la stérilité des mâles.

1.3 Expression et évolution

La question centrale de cette thèse est le rôle de l'expression génique dans les processus de différenciation des populations et des espèces, en relation avec l'isolement reproducteur, mais aussi avec l'adaptation au milieu local et l'invasion de nouveaux milieux.

L'expression génique a un très bon potentiel adaptatif. Les mutations proposées peuvent être ajustées par des boucles de régulation, augmentant largement la part de variance explorée par ces mutations par rapport aux mutations de la séquence codante. Prenant pour exemple les patrons de coloration des ailes chez différents insectes, Prud'homme *et al.* (2007) soulignent l'importance de la régulation génique, notamment via la régulation *cis* (régulation par des éléments physiquement liés au gène) dans la création de nouveautés évolutives. Les exemples démontrés d'adaptation via la régulation génique sont pourtant rares, et limités à la drosophile (Sucena et Stern, 2000; Daborn

et al., 2002; Prud'homme *et al.*, 2006) et à l'épinoche (Miller *et al.*, 2007). Cependant, le potentiel de l'expression pour l'adaptation est important, et représente un axe de recherche important pour la compréhension des processus d'adaptation (Prud'homme *et al.*, 2007; Fraser *et al.*, 2010).

La recherche des mutations (au sens large de modifications, changements) permettant l'évolution est un centre d'intérêt majeur en biologie évolutive. A quel niveau interviennent ces changements : séquences codantes, modification des nucléotides, expression, influence des petits ARNs, régulation de la traduction, modifications post-traductionnelles, conformation protéique ? On s'intéresse également à leur impact : une mutation d'impact majeur permettra-t-elle une évolution plus rapide qu'une ou plusieurs mutations d'impact plus faible ? Pour que ces mutations soient pertinentes du point de vue évolutif, il faut qu'elles proposent une variation phénotypique héréditaire. Historiquement, les évolutionnistes ont d'abord accordé plus d'importance aux micromutations. Des mutations d'impact majeur sur des gènes de développement fortement pleiotropes ont cependant été mises en évidence, plaidant en faveur du rôle des macromutations dans les sauts évolutifs (Stern, 2000). Il est pourtant probable que des mutations ayant des effets trop extrêmes soient rapidement éliminées du processus évolutif, qui progresse fondamentalement à pas menus (d'autant que ce genre de mutations n'est que rarement observée). Il est aussi possible qu'une nouvelle mutation, ou l'invasion d'un nouvel environnement par un génotype étranger permette la révélation d'une variabilité génétique cryptique (c'est-à-dire sans conséquence sur le phénotype dans la zone d'origine) (Gibson et Wagner, 2000; Gibson et Dworkin, 2004). Comment la sélection naturelle agit-elle sur ces différentes mutations ? Quel est le rôle de la dérive dans l'évolution de l'expression ? Nous nous intéresserons à ces questions dans cette thèse.

Dans ce contexte de recherche de mutations intéressantes au point de vue évolutif, les changements d'expression génique sont d'excellents candidats pour l'adaptation à un nouvel environnement. Potentiellement plus malléables que la séquence codante, l'impact phénotypique des changements de régulation peut cependant être majeur (Macintyre, 1982). L'importance des mutations dans les séquences régulatrices par rapport aux mutations codantes, longtemps supposée, est maintenant fortement soutenue par la littérature (Wray, 2007; Wittkopp *et al.*, 2008a; Genissel *et al.*, 2008). Les mutations dans les séquences régulatrices sont-elles qualitativement différentes des mutations dans les séquences codantes ? Les arguments en faveur de cette théorie sont principalement de deux natures : d'abord, la probabilité que les mutations *cis* (mutation sur la séquence régulatrice physiquement liée au gène) affectent les phénotypes est intrinsèquement plus grande, enfin, la sélection agit plus efficacement sur les mutations régulatrices (Wray, 2007).

Le premier argument s'appuie sur le caractère dynamique, finement et continuellement régulé de la transcription. L'évolution constante du nombre de transcrits permet une adaptation bien plus rapide que la structure qui est par nature plus figée (bien qu'on retrouve fréquemment des variants de structure, l'amplitude des variations est généralement beaucoup moins large) (Prud'homme *et al.*, 2007). Ces caractéristiques dynamiques permettent également des adaptations temporaires en réponse à des environnements variables. Le second argument (sélection plus facile sur des mutations régulatrices) s'appuie sur plusieurs constats. D'abord, des études ont montré que la régulation est souvent spécifique de chaque allèle (Wittkopp *et al.*, 2004; Ronald *et al.*, 2005; Wittkopp *et al.*, 2008a), ce qui donne donc des mutations codominantes, et donc donnant plus facilement prise à la sélection (car l'hétérozygote présente déjà un phénotype différent). Enfin, la régulation a souvent une spécificité tissulaire, ce qui la rend plus souple et plus adaptable que la mutation codante, qui a un impact dans l'ensemble des tissus où le gène s'exprime.

1.3.1 Variance d'expression au sein des populations

L'ampleur des variances intra-population montre la nécessité d'utiliser, pour des études centrées sur des organismes prélevés dans la nature, plusieurs lignées / individus, afin d'obtenir un résultat représentatif de la population dans son ensemble. Seule une prise en compte de la variance intra-population permet d'évaluer les différences réelles entre populations (Whitehead et Crawford, 2006b). Des études, sur la souris, les primates, la levure et les poissons ont relativement bien pris en compte cette contrainte. Mais cette nécessité a pourtant été peu suivie par d'autres études autour du sous-groupe *melanogaster* en général, et du complexe *simulans* en particulier. Michalak et Noor (2003), tout comme Moehring *et al.* (2007), ont utilisé une seule lignée de *D. simulans*, issue d'une population dérivée (Floride) collectée en 1985, soit une quinzaine d'années d'évolution en laboratoire, ainsi qu'une combinaison de six lignées de *D. mauritiana* collectée à l'île Maurice (dont cette espèce est endémique), en 1981. Haerty et Singh (2006) ont utilisé une lignée de *D. simulans* d'Arizona récemment fondée (2004), mais une lignée de *D. melanogaster* datant de 1955. L'utilisation d'une seule lignée pour représenter une espèce ne permet pas de prendre en compte la variance intra-population, et *a fortiori* intra-espèce. Des lignées collectées 15 à 20 ans avant la réalisation de l'étude ont évolué un minimum de 250 générations en laboratoire, dans des conditions de nutrition et d'élevage optimales, sans stress de prédation, ou lié à l'environnement (stress chimique, physiologique). On peut se demander à quel point ces lignées sont encore représentatives des populations dont elles proviennent.

De la même manière, une comparaison d'espèces devra idéalement être effectuée via plusieurs populations des espèces étudiées, d'autant plus si celles-ci présentent une structuration. Il s'agit comme ci-dessus de la nécessité de mesurer la variance intra et la variance inter : ce qui permet la comparaison des groupes, c'est la mesure de la variation à l'intérieur de ces groupes. Ainsi, Enard *et al.* (2002) ont montré une très forte variation entre individus, et de faibles différences entre espèces de primates (dont l'homme).

Afin de comprendre les variations entre taxons (populations, espèces), plusieurs études ont d'abord cherché à évaluer les variations d'expression à l'intérieur d'une population : soit en observant directement la variation inter-individuelle (Cavalieri *et al.*, 2000; Pritchard *et al.*, 2001; Oleksiak *et al.*, 2002; Cheung *et al.*, 2003; Whitney *et al.*, 2003; Morley *et al.*, 2004; Whitehead et Crawford, 2005; Cobb *et al.*, 2005; Holloway *et al.*, 2007), soit (lorsque par exemple l'organisme est trop petit pour qu'un seul individu fournisse assez d'ARN) entre lignées fortement consanguines (Jin *et al.*, 2001; Townsend *et al.*, 2003; Brem *et al.*, 2002; Hutter *et al.*, 2008; Muller *et al.*, 2011). Whitehead et Crawford (2006a) proposent également une très bonne revue sur le sujet. Ces études évaluent la variation inter-individuelle / inter-lignée (intra-populationnelle) chez différents organismes : levure, drosophile, souris, poisson et humain. Globalement, dans les études précédentes, le polymorphisme inter-individuel est apparu variable, évoluant généralement en dessous de 7% de gènes significativement variables entre individus ou lignées. Des variations beaucoup plus fortes ont cependant été montrées (de l'ordre de 25%) (Brem *et al.*, 2002; Whitehead et Crawford, 2005). Cependant, les méthodes utilisées (notamment statistiques) sont extrêmement diverses, et il est difficile de comparer ces études. Ce qu'il faut en retenir, c'est que les sources de variations inter-individuelle sont nombreuses. Le sexe a notamment un effet extrêmement fort (Meiklejohn *et al.*, 2003; Haerty et Singh, 2006; Ellegren et Parsch, 2007; Zhang *et al.*, 2007; Muller *et al.*, 2011), qui n'est pas limité aux seuls tissus reproducteurs (Yang *et al.*, 2006). L'expression dépend également très largement des tissus examinés (Pritchard *et al.*, 2001; Whitney *et al.*, 2003; Morley *et al.*, 2004; Catron et Noor, 2008). Les conditions de stress jouent également un rôle fort, que ce soit via l'induction de gènes de résistance, du système immunitaire, de gènes liés au stress (Pritchard *et al.*, 2001; Daborn *et al.*, 2002; Oleksiak *et al.*, 2005), etc.

1.3.2 Expression : reflet de la fonction ?

Pourquoi étudier l'expression des gènes ? Certes, on peut considérer l'expression comme le "premier phénotype", dans le sens du phénotype qui est une conséquence quasi directe du génotype.

Une question intéressante est alors : à quel point les données d'expression sont-elles représentatives d'un état fonctionnel, en d'autres termes, quelle corrélation y-a-t il entre la quantité d'ARNm et la quantité de protéine ? Ainsi, dans une étude sur l'expression à l'échelle du génome entier Townsend *et al.* (2003) ont examiné quatre isolats naturels de la levure *Saccharomyces cerevisiae* prélevés en Italie, dans un vignoble (Townsend *et al.*, 2003). Ils ont utilisé des puces portant l'ensemble des gènes de la levure connus à ce moment, pour mesurer l'ampleur des différences d'expression entre ces isolats, et identifier les gènes et réseaux de régulation impliqués. Ils ont montré des divergences au niveau de gènes impliqués dans la biosynthèse des acides aminés, la dégradation des protéines, le transport des ions métalliques et le phénotype de croissance. Les différences mises en évidence ici concernent entre autre la traduction et la dégradation des protéines, et peuvent donc avoir un impact sur la quantité de protéines produites ; cela concerne donc des gènes qui ne sont pas ceux directement révélés par l'analyse transcriptomique. Cette question a été la source de nombreux débats (Townsend *et al.*, 2003; Greenbaum *et al.*, 2003; Hack, 2004; Lu *et al.*, 2007). Toujours chez la levure, Greenbaum *et al.* (2003) ont montré une corrélation relativement faible entre quantité d'ARNm et quantité de protéines (r = 0,66). Ils ont proposé trois hypothèses pour expliquer cette observation :

1. une forte régulation se fait au niveau post-transcriptionnel et traductionnel
2. il y a une forte variance dans la vitesse de dégradation et de biosynthèse des protéines, et ainsi une forte variance dans la demi-vie des protéines (comme le laisse supposer le patron observé par Townsend *et al.* (2003))
3. il y a une erreur, ou un bruit significatif dans les mesures globales des quantités d'ARNm et de protéines, qui rendent difficile une bonne mesure de corrélation

Poussant plus loin la recherche, Greenbaum *et al.* (2003) ont émis l'hypothèse que les gènes les plus variables au niveau de l'expression devraient montrer une meilleure corrélation avec la quantité de protéines. La cellule ayant "investi de l'énergie" dans la régulation génique, il est en effet souhaitable d'un point de vue évolutif que cela se répercute au niveau de l'effecteur, c'est-à-dire la protéine. Et c'est effectivement ce qu'ils ont constaté, montrant une corrélation de 0,89 pour les transcrits les plus variables en expression, et de 0,2 pour les moins variables. Pour ces derniers, la régulation se situe donc au niveau protéique (Greenbaum *et al.*, 2003). Une limite de ces études réside dans le fait que les quantifications des protéines se limitent généralement à 500-1000 protéines parmi les plus abondantes. Il reste à découvrir à quel point les observations faites sur celles-ci sont pertinentes pour l'ensemble du protéome. Lu *et al.* (2007) ont développé une

nouvelle technique de quantification protéique basée sur les techniques de spectrométrie de masse (APEX : Absolute Protein EXpression), et l'ont appliquée à l'examen des contributions relatives des régulations transcriptionnelle et traductionnelle. Ils ont ainsi montré que chez la bactérie *Escherichia coli*, presque la moitié de la régulation des protéines a lieu au niveau transcriptionnel, et que ce chiffre monte à plus de 70% chez *S. cerevisiae*. Cet écart est probablement dû à la régulation transcriptionnelle par opéron chez la bactérie (Lu *et al.*, 2007). Ces observations donnent néanmoins toute leur justification aux analyses différentielles d'expression, puisque par définition, ce sont alors des gènes variables au niveau de l'ARNm que nous repèrerons, c'est-à-dire d'après Greenbaum *et al.* (2003) des gènes pour lesquels la corrélation avec le fonctionnel (les protéines) est bonne. Il est aussi probable que la drosophile suive le chemin de *S. cerevisiae* plus que celui de *E. coli*, son système de régulation étant beaucoup plus proche de celui de la levure (pas de système d'opérons).

1.3.3 Expression et éléments transposables

Les éléments transposables sont d'excellents candidats pour un apport mutagène au niveau de l'expression. Ainsi, Chen et Li (2007) ont montré une forte corrélation entre la variance d'expression et la présence d'éléments transposables dans des parties non codantes proches du gène. Townsend *et al.* (2003) ont également examiné les différences d'expression de dix insertions de l'élément transposable *Ty* entre leurs quatre isolats naturels de levure. Le niveau d'expression des éléments transposables reste très stable à l'intérieur d'une lignée pour l'ensemble des dix copies de l'élément en question. Les différences d'expression des éléments transposables peuvent révéler un niveau mutationnel de l'expression de la lignée, c'est-à-dire, une propension au changement de niveau d'expression. Une activité plus grande de ces éléments laisse supposer une variation phénotypique de l'expression plus fréquente (Paigen, 1986). Certes les variations générées par ces éléments sont le plus souvent délétères (voir par exemple Lynch (2007)), mais les mutations de l'expression proposées pourront parfois être avantageuses pour l'hôte. Le potentiel des éléments transposables pour modifier l'expression est énorme. En effet, ils possèdent leurs propres promoteurs et éléments régulateurs, ils peuvent aussi s'insérer dans des régions régulatrices déjà présentes, proposer des variations d'épissage,...

1.3.4 Signification biologique des variations d'expression

L'analyse statistique permet de révéler des différences d'expression statistiquement significatives. Mais qu'en est-il de leurs significations biologiques ? En effet, certaines différences observées (quelle que soit leur ampleur) n'auront peut-être aucun impact sur la valeur sélective (fitness) et sur le phénotype en général (changement d'expression neutre), et leur différenciation n'est peut-être déterminée que par des processus stochastiques comme la dérive génétique, un effet de goulot d'étranglement lors de la migration vers un nouvel environnement (= effet fondateur) (Hartl *et al.*, 1985; Townsend *et al.*, 2003),... De même, une certaine "course à la puissance" s'était installée dans ce genre d'analyse, dans le but de déceler des différences de plus en plus fines. Pourtant, quelle peut être la signification biologique d'une différence d'expression de 1,2 par exemple ? Ces questions légitimes n'ont probablement pas de réponse unique, l'analyse devant être adaptée à chaque cas. Il est toujours plus facile d'interpréter la signification biologique d'une grande différence, surtout si l'on peut faire le lien avec une fonction bien décrite.

1.3.5 Hérédité et expression

Nous avons discuté à quel point l'expression est capable de proposer de nouveaux phénotypes, d'évoluer plus rapidement que la séquence codante. Cependant, pour que l'expression soit partie intégrante de l'évolution et de l'adaptation, il faut que les changements soient transmissibles d'une génération à l'autre. De nombreux travaux ont étudié chez les hybrides intra comme inter spécifiques les patrons d'expression par rapport aux parents (Gibson et Dworkin, 2004), recherchant l'additivité (= expression correspondant à la moyenne d'expression des parents), la non-additivité (= expression plus proche de celle d'un des deux parents) et l'hétérosis (= expression supérieure ou inférieure à celle des deux parents). Les résultats de ces études sont largement contrastés montrant dans certains cas une additivité fortement limitée (2% ou moins, Gibson et Dworkin 2004; Haerty et Singh 2006, avec tout de même environ 20% pour les hybrides *D. simulans* / *D. mauritiana* dans cette dernière étude), ou beaucoup plus importante (71%, Hughes *et al.* 2006; Rottscheidt et Harr 2007).

1.3.6 Variance d'expression entre taxons (populations, espèces)

De nombreuses études ont porté sur les différences entre populations et/ou espèces. Les objectifs de ces études sont divers, allant de la compréhension de la divergence et de l'adaptation locale des

populations (Oleksiak *et al.*, 2002; Meiklejohn *et al.*, 2003; Hutter *et al.*, 2008; Muller *et al.*, 2011), à une estimation de la diversité et de la structure des populations dans un but de conservation (Vandersteen Tymchuk *et al.*, 2010), en passant par une recherche de contraintes développementales liées à la divergence des espèces (Rifkin *et al.*, 2003). Wolf *et al.* (2010) ont même utilisé l'expression comme un marqueur de spéciation : ils ont comparé l'expression chez deux espèces de corbeaux très faiblement différenciées au niveau moléculaire. Utilisant six individus par espèce, ils ont révélé une assez bonne différenciation de l'expression entre les deux espèces, démontrant ainsi l'utilité de l'expression pour observer des espèces en cours de spéciation.

A l'intérieur même d'une population, les variations d'expression peuvent être importantes (c'est extrêmement variable en fonction des études et des taxons), et les changements d'expression sont donc peu appropriés en tant que marqueur phylogénétique. En effet, des approches de classification (= clustering) ont été utilisées pour regrouper les lignées / individus par patrons d'expression proche. Mais l'expression est trop dépendante à la fois de l'environnement et du sexe, et les lignées regroupées par ce type d'approche le sont plus généralement selon ces facteurs (Oleksiak *et al.*, 2002). Certains auteurs ont cependant suggéré que la variance d'expression sous l'hypothèse d'évolution neutre pourrait être un bon marqueur de phylogénie (Khaitovich *et al.*, 2005).

La première étude qui a réalisé une comparaison d'expression sur des populations naturelles a porté sur deux espèces proches de poissons téléostéens du genre *Fundulus* (Oleksiak *et al.*, 2002). Les auteurs ont examiné par puces à ADN le transcriptome de cinq individus $\times$ deux populations de *F. heteroclitus*, ainsi que cinq individus d'une espèce sœur : *F. grandis*. Ils ont pu révéler 161 gènes différentiellement exprimés entre individus d'une même population, et 15 gènes différentiellement exprimés entre populations. Cela soulève des questions sur la part de variance intra par rapport à la variance inter (comme nous l'avons vu page 15). Nous verrons plus loin que le détail de l'analyse permet d'interpréter ces résultats au regard de l'adaptation locale.

L'étude présentée par Oleksiak *et al.* (2002) a montré une structure des populations au niveau de l'expression. Ils ont révélé des patrons d'expression proches pour *F. grandis* et la population méridionale de *F. heteroclitus*, mais divergents de celui de la population septentrionale de *F. heteroclitus*. Cette divergence semble due à une adaptation locale à l'environnement, notamment ici la température de l'eau. Cette étude est donc la première à avoir montré une adaptation à l'environnement via l'expression, de manière beaucoup plus rapide que l'adaptation des séquences, puisque au niveau de celles-ci, les deux populations de *F. heteroclitus* restent bien plus proches entre elles que de *F. grandis*. Cependant, le patron observé ici peut également être dû à une conservation de

l'induction d'expression du milieu originel par des phénomènes épigénétiques (les individus utilisés ont été prélevés dans la nature avant d'être élevés en laboratoire pendant environ six mois). Cependant, si on a montré que les processus épigénétiques permettent la transmission de la régulation génique après stimulation par l'environnement, y compris à travers les générations, ce phénomène reste ponctuel (Jaenisch et Bird, 2003).

Un bémol de cette étude d'Oleksiak *et al.* (2002) est le fait qu'ils ont choisi de ne pas appliquer de correction pour les tests multiples (qui faisait chuter de façon conservative le nombre de gènes différentiellement exprimés à 37 au lieu de 161). Ce choix (que l'on peut voir dans d'autres études, notamment Moehring *et al.* 2007) est défendable si l'on souhaite commenter l'ampleur des différences, en la comparant avec le taux d'erreur attendu de 5%. Cependant, si on souhaite effectuer une analyse fonctionnelle de gènes, le nombre important de faux positifs aura un impact sur les résultats. Par exemple, Oleksiak *et al.* (2002) ont utilisé des puces comportant 907 gènes au total, soit 907 tests statistiques réalisés avec une p-value de 0,01. On a donc $907 \times 0,01 = 9,07$, soit environ neuf gènes faux positifs parmi la liste de 161. Cela reste assez peu, cependant, si l'on souhaite analyser dans le détail la fonction des gènes, cela implique que neuf gènes ne sont dans cette liste que par hasard. Enfin, cela fait également une moyenne de neuf faux positifs parmi le set de 15 gènes différentiellement exprimés entre populations, ce qui représente un pourcentage important de faux positifs.

Une étude s'est particulièrement concentrée sur la recherche de différences d'expression dans les populations naturelles (Vandersteen Tymchuk *et al.*, 2010). Les auteurs ont relâché des saumons d'élevage dans la nature avant de les recapturer pour l'échantillonnage, les livrant ainsi à la sélection de leur environnement naturel. Ils ont examiné plusieurs populations, à deux points dans le temps. Les auteurs ont ainsi observé des divergences liées à une adaptation locale (due à des différences de clarté et propreté des eaux) ainsi qu'une divergence neutre relativement proportionnelle à la distance séparant les populations, comme attendu sous un modèle d'isolement par la distance.

Globalement, les études ayant examiné les divergences d'expression entre taxons ont pu montrer une part d'évolution adaptative de l'expression, gouvernée principalement par deux forces : la sélection liée à l'adaptation à un milieu local, et la sélection sexuelle (au vu des fortes différences notées entre les sexes Meiklejohn *et al.* 2003; Haerty et Singh 2006; Zhang *et al.* 2007; Muller *et al.* 2011). Cependant, une grande part de la divergence d'expression semble aussi être neutre. Plus les taxons sont éloignés phylogénétiquement, plus la probabilité que les différences d'expression observées soient neutres sera grande.

1.3.7 Notre modèle : drosophiles invasive versus endémique

Le complexe *simulans* est un taxon de choix pour les études évolutives, à la fois par la richesse des outils disponibles pour la drosophile en général, par la proximité entre les trois espèces du complexe (*D. simulans*, *D. sechellia* et *D. mauritiana*) qui peuvent encore s'hybrider, et par les caractéristiques extrêmement variées de l'écologie de ces espèces. *D. simulans* a été décrite pour la première fois en 1919 (Sturtevant, 1919). Son nom provient de la confusion qui a d'abord été faite avec *D. melanogaster*. Cette espèce appartient au sous-groupe *melanogaster* avec huit autres espèces qui montrent une grande diversité écologique, comportementale, et d'aires de répartition. Ces différences sont très bien illustrées par les deux espèces qui ont été l'objet de nos études : *D. simulans* et *D. sechellia*. *D. simulans* est devenue récemment cosmopolite (quelques dizaines à quelques centaines d'années Lachaise *et al.* 2004). La zone d'origine de l'espèce pourrait être située entre Madagascar et l'Afrique de l'est (Kenya / Tanzanie). Madagascar est maintenant l'hypothèse la mieux soutenue (Schöfl et Schlötterer, 2006). *D. sechellia* est une endémique des îles Seychelles (Lachaise *et al.*, 1988, 2004; Dean et Ballard, 2004; Kopp *et al.*, 2006). *D. simulans* est généraliste (se reproduit et se nourrit sur des ressources variées), alors que *D. sechellia* est spécialiste (elle ne se nourrit et ne se reproduit que sur le Nonni, alias *Morinda citrifolia*, plante toxique pour toutes les autres drosophiles, R'kha *et al.* 1991). Ces deux espèces se sont séparées de *D. melanogaster* il y a 2 à 3 millions d'années (Hey et Kliman, 1993; Kliman *et al.*, 2000; Lachaise et Silvain, 2004; Cutter, 2008). Les premières estimations dataient la séparation de *D. simulans*, *D. sechellia* et *D. mauritiana* il y a 400 000 ans (Hey et Kliman, 1993; Kliman *et al.*, 2000; Lachaise et Silvain, 2004), mais des données plus récentes la datent à environ 250 000 ans (McDermott et Kliman, 2008). Quoi qu'il en soit, la séparation de *D. simulans* et de ses deux jumelles est extrêmement récente à l'échelle de l'évolution des *Drosophilidae*, l'invasion du monde par l'espèce l'est donc également. Au niveau de la séquence codante, *D. simulans* et *D. sechellia* divergent d'environ 1,5% (McDermott et Kliman, 2008).

D. sechellia est donc une espèce spécialisée sur la plante *Morinda citrifolia*. Des élevages de différentes espèces sur cette plante ont donc montré que seule *D. sechellia* était capable d'y survivre, les autres espèces présentant plus de 50% de mortalité de l'adulte dès la première heure, et dans tous les cas, 100% de mortalité à sept jours. De même, une exposition de plus de 30 minutes suffit pour avoir plus de 90% de mortalité embryonnaire chez *D. simulans* (R'kha *et al.*, 1991). La toxicité est principalement due à l'acide octanoïque sécrété en grande quantité par la plante. Cette écologie particulière a suscité bien des études, pour comprendre comment *D. sechellia* n'est

pas repoussée par cette plante, comment sa physiologie s'est adaptée aux toxines de la plante, pourquoi *D. sechellia* se limite à cette plante,... Mise en compétition sur une ressource classique avec d'autres drosophiles, *D. sechellia* est peu compétitive (réduction du nombre d'ovarioles et de la production d'œufs comparée à *D. simulans* et *D. mauritiana*) (Lachaise *et al.*, 1988; R' kha *et al.*, 1997). C'est donc cette attraction pour *Morinda citrifolia*, ainsi que sa capacité à tolérer les toxines de cette dernière qui ont donné à *D. sechellia* un avantage sur cette ressource, et lui ont conféré une niche écologique dans laquelle elle a pu se développer.

Structuration des populations

La structuration géographique de *D. simulans* a été longtemps source de débat. Des études basées sur des allozymes (Hyytia *et al.*, 1985; Choudhary et Singh, 1987) ou sur des critères phénotypiques (Capy *et al.*, 1993) ont conclut à une structuration faible. Cependant, des études basées sur le polymorphisme moléculaire (Begun et Aquadro, 1995; Irvin *et al.*, 1998; Andolfatto, 2001) ont montré une nette différenciation entre les populations africaines et non-africaines. En Afrique, Veuille *et al.* (2004) différencient trois groupes : un groupe d'Afrique de l'est correspondant à la zone ancestrale, et comprenant les populations de Mayotte, Madagascar, du Kenya et de Tanzanie ; un groupe d'Afrique du sud (Zimbabwe) ; et enfin un groupe d'Afrique centrale (représenté ici par une population du Cameroun). Ces groupes sont bien séparés moléculairement en Afrique, et les populations du Cameroun et du Zimbabwe semblent avoir subit un effet de fondation, qui a réduit leur diversité moléculaire (Veuille *et al.*, 2004). Les populations européennes et de la Réunion portent également les caractéristiques de populations fondées par migration à partir de la zone ancestrale. Ces observations sont confirmées par des études de microsatellites, qui montrent une variabilité réduite des locus microsatellites liés à l'X, probablement due à des balayages sélectifs associés à la "sortie d'Afrique" de l'espèce (Schöfl et Schlötterer, 2004). Les auteurs ont montré par simulation que l'effet fondateur ne pouvait pas être seul responsable des patrons observés, et qu'il avait probablement été associé à des sélections sur les locus microsatellites observés. Étendant cette étude à un total de huit populations (deux populations en Ouganda, et une au Malawi, Zimbabwe, Tunisie, Italie, Autriche et en Israël), Schöfl et Schlötterer (2006) ont montré une structuration entre populations africaines et non-africaines, mais également à l'intérieur de l'Afrique (population Tunisienne plus proche des populations non-africaines). Ils ont également montré que les populations d'Afrique de l'ouest et du sud sont des populations dérivées. Ces résultats ont été également confirmés par une étude sur des séquences codantes (Baudry *et al.*, 2006).

L'ensemble de ces travaux montre une structuration nette des populations chez *D. simulans*. En Afrique, on peut séparer les populations en trois groupes principaux : la zone d'origine à l'est, l'Afrique de l'ouest et l'Afrique du sud, ces deux derniers étant fondés lors de la dissémination de l'espèce, tout comme les populations dérivées que l'on retrouve sur les autres continents.

L'aire géographiques de *D. sechellia* est limitée à l'archipel des Seychelles. Jusqu'à récemment, on pensait cette espèce constituée d'une seule population (Legrand *et al.*, 2009), mais des analyses plus fines ont montré une légère différenciation, qui divise *D. sechellia* en deux populations (Legrand *et al.*, 2011).

Les études d'expression chez *D. simulans*

Quelques études se sont intéressées à l'expression dans cette espèce. Deux d'entre elles ont examiné la corrélation entre le polymorphisme de séquence et le polymorphisme d'expression entre les six lignées séquencées de *D. simulans* (Holloway *et al.*, 2007; Lawniczak *et al.*, 2008). Les auteurs ont recherché les corrélations entre différentes parties des gènes (UTR = "untranslated region", introns, exons) et l'expression. Cette approche est limitée par la faible couverture du séquençage de *D. simulans*, qui oblige à exclure les gènes pour lesquels la séquence n'est pas disponible pour assez de lignées. Ils ont cependant pu montrer que les gènes plus variables au niveau de l'expression tendaient vers un plus grand polymorphisme de séquence, notamment codante et 3'UTR. Ils ont également observé une moindre variation d'expression des gènes liés au chromosome X par rapport aux autosomes. Ils ont enfin observé une sur-représentation des gènes biaisés vers les mâles (expression significativement plus importante chez les mâles que chez les femelles) parmi les plus variables en expression. Toutes ces observations sont cohérentes avec les patrons généraux d'expression lié à l'X et au sexe (Haerty et Singh, 2006; Zhang *et al.*, 2007; Haerty *et al.*, 2007). La nouveauté de ces études (Holloway *et al.*, 2007; Lawniczak *et al.*, 2008) réside donc plus dans la corrélation entre polymorphisme et variabilité d'expression. Cette corrélation pourrait signifier que la sélection s'exerce parallèlement sur l'expression du gène, sur sa séquence régulatrice et sur sa séquence codante. Les gènes les plus contraints en expression seront donc peu polymorphes au niveau de la séquence, et ceux qui évoluent pour l'expression de façon neutre seront globalement plus libres d'évoluer sans contraintes sélectives en ce qui concerne la séquence. Ces résultats soulignent également l'importance de l'extrémité 3' dans la régulation génique (Holloway *et al.*, 2007; Lawniczak *et al.*, 2008).

D'autres études ont réalisé des comparaisons entre *D. simulans* et d'autres drosophiles. Nuzhdin

et al. (2004) ont montré une évolution adaptative de l'expression entre *D. simulans* et *D. melanogaster*, avec 534 gènes différentiellement exprimés, et notamment une divergence rapide de gènes liés à la reproduction. La divergence rapide des gènes liés à la reproduction est attendue sous l'hypothèse qu'ils évoluent sous de fortes pressions liées à la sélection sexuelle. La compétition pour la reproduction provoque une sorte de course à l'armement, à la fois entre individus de même sexe, mais aussi souvent antagonistes entre les deux sexes (Andersson, 1994). En promouvant une divergence rapide des phénotypes liés à la reproduction, la sélection sexuelle provoque alors celles des gènes impliqués dans ces processus. Ces observations sont intimement liées à l'évolution de gènes à l'expression biaisée vers un sexe déjà évoquée plus haut.

Dworkin et Jones (2009) ont montré des divergences dans la régulation de gènes liés à la spécialisation entre *D. simulans* et *D. sechellia*, et nous analyserons plus avant cette étude dans la partie résultat. Enfin le laboratoire de Mohammed Noor (Duke University) s'est beaucoup intéressé à l'expression chez les drosophiles du sous-groupe *melanogaster*, et s'ils ont souvent eu des approches statistiques discutables (pas de correction pour les tests multiples, voir page 21), les résultats sur les hybrides et les espèces soutiennent les hypothèses concernant les gènes biaisés vers un sexe (notamment les mâles) (Michalak et Noor, 2003, 2004; Moehring *et al.*, 2007).

Globalement *D. simulans* semble suivre les patrons d'expression classique : évolution plus rapide de gènes liés au sexe, gènes sur le chromosome X contraints par une sélection dirigée (moins de polymorphisme, mais une divergence rapide), bonne corrélation entre l'évolution des 3'UTR et l'expression. Cependant, l'expression de *D. simulans* dans son environnement naturel, entre populations issues de la nature n'a pas été étudiée, et c'est un des aspects abordé dans cette thèse.

Notre étude s'est ainsi placée dans le contexte des populations naturelles chez la drosophile, afin d'apporter une connaissance en profondeur de la diversité de l'expression à l'intérieur des espèces. Pour cela, l'utilisation des méthodes haut débit s'est imposée. Nous avons donc d'abord comparé par puces à ADN quatre populations de *D. simulans* (trois africaines et une française), une de *D. sechellia*, et leurs hybrides inter-spécifiques. Chaque population était représentée par quatre lignées. Puis nous avons approfondi la comparaison de population (une de la zone ancestrale versus une dérivée) via les techniques de séquençage haut débit qui permettent un excellente couverture du transcriptome tout en s'affranchissant des méthodes basées sur l'hybridation. Dans cette seconde partie, chaque population était représentée par 100 individus indépendants. Contrairement aux travaux précédents, généralement basés sur une seule lignée collectée de longue date, notre étude apporte des connaissances originales qui privilégient la relation avec l'environnement local.

Nous allons d'abord présenter chacune des deux études de transcriptome entreprises au cours cette thèse, avant de discuter nos résultats de façon globale, puis de s'intéresser aux perspectives ouvertes par ce travail. Les objectifs étaient multiples : en ce qui concerne la première étude, observer la différenciation entre espèces (potentiellement lié à la spécialisation) et populations (de la zone d'origine versus dérivées), ainsi que les incompatibilités d'expression chez les hybrides mâles stériles. Pour ce qui est de la deuxième étude, nous souhaitions observer l'adaptation de l'expression des populations naturelles, tout en distinguant l'expression dans leur milieu naturel par rapport à l'expression dans un milieu nouveau. Pour ces études, nous avons développé une étroite collaboration avec une équipe de statisticiens (équipe Statistiques et Génomes, AgroParisTech / INRA), ce qui nous a permis d'approfondir les approches statistiques.

II Matériel et Méthodes

2.1 Comparaisons de *D. sechellia*, *D. simulans* et leurs hybrides par puces à ADN

2.1.1 Lignées de drosophiles

Nous avons utilisé des lignées fondées chacune à partir d'une femelle fécondée, prélevée dans la nature (lignée isofemelle) :

Drosophila simulans

- quatre lignées isofemelles de Mazoe au Zimbabwe, collectées en 1997
- quatre lignées isofemelles de Nairobi au Kenya, collectées en 2001
- quatre lignées isofemelles de Mahé et Praslin, deux îles de l'archipel des Seychelles, collectées en 2003
- quatre lignées isofemelles de la Vallée du Rhône en France, collectées en 2003

Drosophila sechellia

- quatre lignées isofemelles de Mahé et Praslin, deux îles de l'archipel des Seychelles, collectées en 2003

2.1.2 Obtention des échantillons

Les drosophiles ont été élevées en tube à densité contrôlée, sur milieu axénique (David, 1962) et à 25°C, avec un cycle de lumière naturelle. Pour chaque lignée isofemelle, un minimum de six tubes réplicats ont été réalisés, contenant chacun huit mâles et huit femelles (dix pour *D. sechellia*). Pour une population donnée de *D. simulans*, quatre croisements ont été réalisés, chaque fois entre une femelle d'une lignée différente, et un mâle d'une lignée différente de *D. sechellia* (avec un minimum

de trois réplicats par croisement). Nous avons ainsi obtenu quatre "populations" hybrides pour chaque population de *D. simulans*. Le plan d'expérience comprenait donc neuf "populations", chacune avec quatre lignées correspondant à des réplicats biologiques (voir table 1) :

- Quatre populations de *D. simulans*
- Une population de *D. sechellia*
- Quatre "populations" hybrides

Nous avons donc utilisé 36 puces monocanales, une par lignée (9 populations $\times$ 4 réplicats).

***Tableau 1** – Plan d'expérience des puces : comparaison entre quatre populations de* D. simulans, *une de* D. sechellia *et quatre "populations" hybrides*

		Mâle *Drosophila sechellia*			
Femelle *Drosophila simulans*	Lignée isofemelle	Sech 1	Sech 2	Sech 3	Sech 4
France	F1	■			
	F2		■		
	F3			■	
	F4				■
Zimbabwe	Z1				■
	Z2			■	
	Z3	■			
	Z4		■		
Kenya	K1	■			
	K2		■		
	K3			■	
	K4				■
Seychelles	S1	■			
	S2		■		
	S3			■	
	S4				■

16 hybrides ont été obtenus en croisant des mâles de D. sechellia *avec des femelles de* D. simulans. *Chaque lignée isofemelle de* D. sechellia *était impliquée dans un croisement avec une lignée de* D. simulans *de chaque population.*

Des mâles issus de chaque lignée ont été prélevés à l'émergence dans au moins trois tubes réplicats différents, afin de constituer des groupes de 25 individus, mis à vieillir pendant 7 jours dans un tube contenant du milieu neuf.

2.1.3 Extraction d'ARN

A sept jours, les groupes de 25 mâles ont été congelés à -80°C. Nous avons ensuite utilisé le Kit Nucleospin RNA II de Macherey-Nagel pour extraire l'ARN du corps entier de nos groupes de 25 mâles, obtenant ainsi environ 3μg d'ARN par condition pour l'hybridation sur les puces. L'ARN a ensuite été rétrotranscrit en présence de nucléotides marqués au P33.

2.1.4 Puces à ADN et hybridation

Les puces utilisées dans cette étude n'ont rien de classique à plusieurs égards. Ce sont des filtres en nylon, monocanaux (hybridation d'une seule condition par puce, ce qui facilite les comparaisons directes, mais nécessite une normalisation stricte), utilisant la radioactivité (et non ce qui est maintenant le plus utilisé, la fluorescence), et les dépôts sont des ADN complémentaires (ADNc) et non des oligonucléotides. L'utilisation d'ADNc permet de limiter l'influence des mutations ponctuelles sur l'hybridation, ce qui dans un contexte inter-spécifique (on utilise des puces prévues pour *D. melanogaster* pour examiner *D. simulans* et *D. sechellia*) est précieux. Chaque puce comporte 7041 spots :

- 5931 ADNc complet de *D. melanogaster*
- 789 spots blancs (témoins négatifs)
- 319 spots portant un ADNc étranger à la drosophile, en l'occurrence une chlorophylle synthetase de *Arabidopsis thaliana* (témoins positifs)

Les ADNc ont été clonés dans un vecteur, amplifiés et déposés sur la membrane de nylon de la puce. Chaque fragment déposé contient à la fois l'ADNc, et un fragment du vecteur qui va servir à normaliser par la quantité de dépôt à chaque spot (première hybridation). Ces puces sont hybridées deux fois :

1. La première hybridation est réalisée avec une sonde marquée au P33 complémentaire de la séquence du vecteur qui accompagne les ADNc. Comme chaque molécule spottée contient ce fragment de vecteur, le signal lu est proportionnel à la quantité d'ADNc à chaque spot (signal d'hybridation vecteur).
2. Après déshybridation et lavage, la seconde hybridation est réalisée avec les ADNc d'intérêt. On peut alors lire le signal d'hybridation "complexe".

2.1.5 Normalisation

La procédure de normalisation a été définie par le constructeur des puces (plate-forme TAGC, Marseille). Tous les spots pour lesquels le signal d'hybridation vecteur était inférieur à cinq fois la médiane des témoins négatifs ont été éliminés de l'analyse. Pour chacun des deux signaux, le bruit de fond, mesuré par la médiane des spots blancs a été soustrait au signal. Cette étape est souvent controversée, mais elle ne changeait pas radicalement les résultats ici, et nous avons donc préféré nous conformer aux instructions du fabricant. La normalisation entre les spots (à l'intérieur de

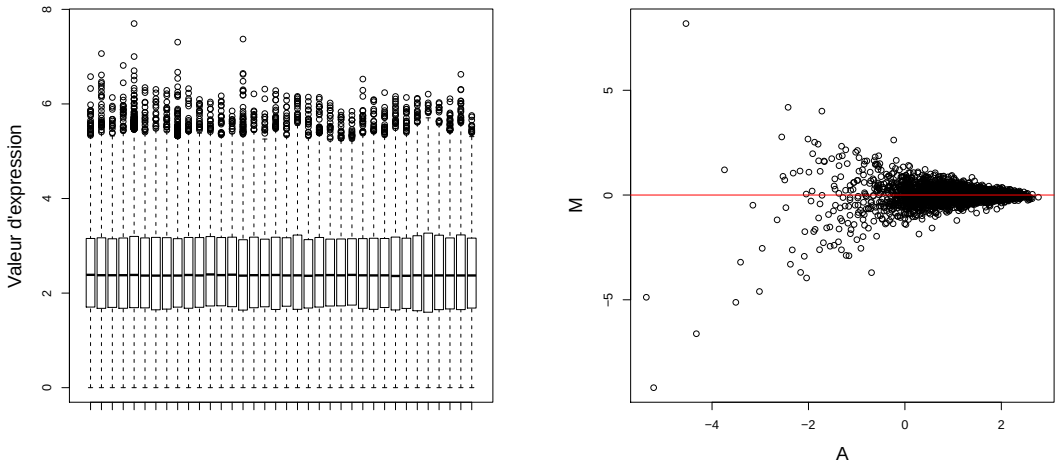


Figure 1 – *Vérification visuelle de la normalisation.* ***A gauche****, un diagramme en boite de la répartition des données pour les 36 puces (unité d'expression arbitraire).* ***A droite****, un exemple de MA plot. M est la différence de signal, A est le signal moyen. Ceci après transformation logarithmique (base 2). On observe une bonne répartition des données autour de la ligne horizontale du 0, témoin d'un bon équilibre entre les signaux.*

chaque puce) a été réalisée en divisant pour chaque spot le signal d'hybridation complexe par le signal d'hybridation vecteur. La dernière étape a été la normalisation entre puces. Deux solutions s'offraient à nous, qui conduisaient au même résultat : diviser le signal à chaque spot par la médiane des signaux de l'ensemble de la puce (approche finalement adoptée) ou par la médiane des contrôles positifs. La qualité de la normalisation a été visuellement vérifiée par MA plot et diagramme en boîte (figure 1). Tous les gènes pour lesquels manquaient quatre données ou plus sur les 36 puces ont été éliminés de l'analyse.

2.1.6 Analyse statistique

L'analyse statistique est basée sur la considération globale de l'ensemble des 36 puces (notamment pour estimer la variance de chaque gène). Pour révéler les gènes différentiellement exprimés, nous avons utilisé une Analyse de Variance (ANOVA) pour chaque gène (Kerr *et al.*, 2000, 2002). Le modèle utilisé est tel que :

$$Y_{ij} = \mu_i + E_{ij},$$

où i est l'indice de population (i = 1,...,9 : 4 *D. simulans*, 1 *D. sechellia* et 4 "populations" hybrides), j est l'indice du réplicat biologique (j= 1,...,4), Y_{ij} est le signal après transformation

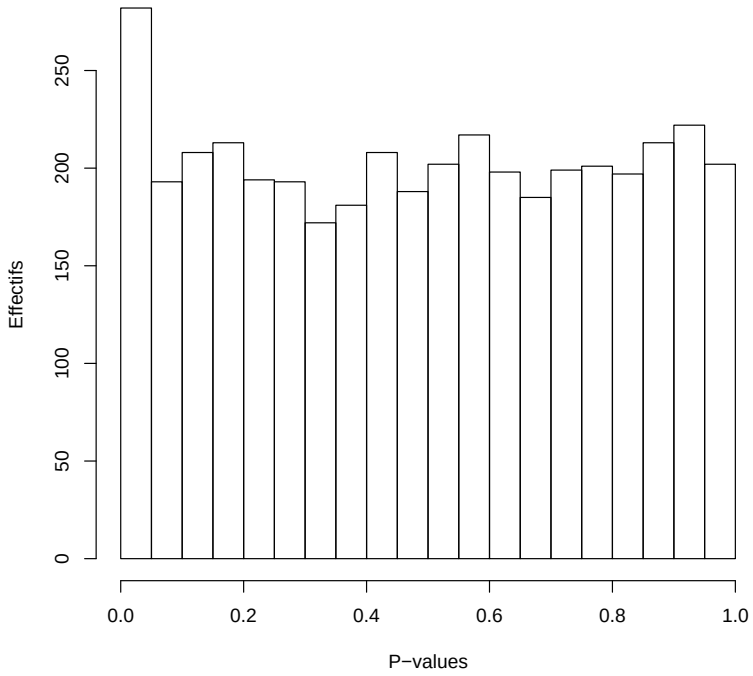


Figure 2 – *Distribution des p-values du test de Levene (Levene, 1960) d'homogénéité de variance.*

logarithmique et normalisation, μ_i est l'expression moyenne du gène pour la population i, et E_{ij} la variabilité résiduelle. Ce modèle suppose une variance commune à toutes les populations, estimée avec $36 - 9 = 27$ degrés de liberté pour la plupart des gènes (quelques gènes ont des données manquantes dues à la normalisation). L'homogénéité de variance entre groupe a été vérifiée via le test de Levene (Levene, 1960). L'homoscédasticité a été confirmée pour tous les gènes exceptés 20 (Taux de faux positifs = False Discovery Rate (FDR) = 0.1, voir figure 2 pour la distribution des p-values). Nous avons réalisé un test d'égalité de moyenne entre les populations de *D. simulans* et la population de *D. sechellia*, et entre les quatre populations de *D. simulans*. Toutes ces comparaisons ont été réalisées via un simple test t dans le contexte du modèle ANOVA. En ce qui concerne les comparaisons hybrides vs. parents, nous avons réalisé un test de comparaison entre les moyennes de chaque population hybride et de ses deux populations parentales indépendamment, et également la moyenne des populations hybrides et la moyenne des populations parentales. Pour chaque comparaison, les p-values brutes ont été ajustées par la méthode de contrôle du taux de faux positifs de Benjamini-Hochberg (Benjamini et Hochberg, 1995), en utilisant un FDR de O,1.

2.1.7 Hétérosis

Pour examiner le mode d'hérédité, nous avons examiné la distribution des effets hybrides en utilisant le rapport d/a, où d est la différence d'expression entre les hybrides et la moyenne des deux parents, et a est la moitié de la différence d'expression entre les deux populations parentales (*D. sechellia* et chaque population de *D. simulans* respectivement) (Falconer et Mackay, 1996).

- Si $d/a = 0$, alors, il n'y a pas d'effet hybride ($d = 0$)
- Si $0 < |d/a| < 1$, il y a effet hybride non transgressif
- Si $|d/a| > 1$, il y a effet hybride transgressif (en dehors des valeurs parentales)

Nous avons réalisé cette analyse seulement avec les gènes différentiellement exprimés entre les parents, afin d'éviter tout biais causé par l'égalité d'expression entre les parents.

2.1.8 Comparaisons de variance

L'objectif était ici de comparer les variances globales des populations. Afin de comparer ces variances entre une population A et une population B, nous avons testé l'existence d'un excès de gènes ayant une variance plus (ou moins) grande dans une population par rapport à l'autre. Pour un gène g, on note $\sigma^2_{g,A}$ et $\sigma^2_{g,B}$ les variances d'expression génique respectivement dans les populations A et B. R_g sera supérieur à 1 pour environ la moitié des gènes dans le cas où les variances de populations seront proches, alors que si la population A montre une plus grande variance d'expression que la population B, alors, le rapport $R_g = \frac{\sigma^2_{g,A}}{\sigma^2_{g,B}}$ sera supérieur à 1 pour la majeure partie des gènes. On peut donc baser le test sur le nombre de gènes N_1^{AB} pour lesquels le ratio empirique $\widehat{R}_g$ pour le gène g sera supérieur à 1. On note p_{AB} la vraie proportion (la proportion théorique dans la population statistique, que l'on va estimer via les gènes de nos puces) de gènes pour lesquels $\sigma^2_{g,A} > \sigma^2_{g,B}$. On teste $H_0 = \{p_{AB} = 1/2\}$ (variances de A et B comparables) contre $H_1 = \{p_{AB} > 1/2\}$ (variance d'expression supérieure chez A). N_1^{AB} suit une distribution binomiale $\mathcal{B}(G, p_{AB})$, avec G le nombre de gènes. La p-value du test est alors :

$$P(N_1^{AB} > n_{1,obs}^{AB} | p_{AB} = 1/2).$$

Cette p-value, c'est la probabilité d'observer $n_{1,obs}^{AB}$ gènes pour lesquels la variance est supérieure dans la population A, sous l'hypothèse H_0 sous laquelle on a en moyenne 50% des gènes qui ont une variance plus grande dans la population A. $n_{1,obs}^{AB}$ étant le nombre de gènes pour lesquels on a effectivement dans notre échantillon $\widehat{\sigma^2_{g,A}} > \widehat{\sigma^2_{g,B}}$.

Effet du nombre de lignées sur la puissance et l'erreur

Sous l'hypothèse nulle Cette analyse prend en compte le nombre de lignées disponibles pour estimer chaque variance. Nous avons effectué des simulations avec différents nombres de lignées pour chaque population (n = 2, 5, 10 et 50), avec une variance égale entre populations pour mesurer l'impact d'un nombre de lignées faible sur les comparaisons de variance. Dans ces quatre cas, 5% des tests étaient significatifs. Le nombre de lignées n'affecte donc pas la distribution des p-values, et par conséquent, l'erreur de première espèce (voir figure 3).

Sous une hypothèse alternative Nous avons répété ces simulations, mais avec 1% des gènes avec un rapport d'écart-type $R_g = 1{,}5$ entre les deux populations. La proportion de p-values inférieures à 0,001 (seuil arbitraire pour observer la puissance du test en fonction du nombre de lignées) était respectivement de 80,2%, 99,1%, 99,2%, 99,5% pour n = 2, 5, 10 and 50. En d'autres termes, on a répété le test, et à ce seuil arbitraire, on détecte la différence de variance dans les données entre 80,2% et 99,5% des fois selon le nombre de lignées utilisées. Ceci montre que la puissance du test augmente très rapidement avec le nombre de lignées, cinq lignées procurant déjà une très bonne puissance.

2.1.9 Ontologies de gènes

Nous avons comparé les gènes présents sur la puce avec le génome en utilisant l'outil en ligne FuncAssociate (Berriz *et al.*, 2003), qui analyse les termes de "Gene Ontology". Cette analyse a révélé de nombreux biais dans les termes d'ontologie de la puce par rapport au génome, ce qui a rendu essentielle l'utilisation de notre puce comme référence pour les analyses des listes de gènes différentiellement exprimés. Nous avons donc ensuite comparé nos listes de gènes différentiellement exprimés avec le groupe de gènes de références constitué par l'ensemble des gènes présents sur la puce. Pour explorer plus avant les termes, leur hiérarchie et les gènes correspondants, nous avons utilisé la base de donnée de "Gene Ontology"du "Gene Ontology consortium" (Ashburner *et al.*, 2000), analyses effectuées pour cette étude avec la base de données en l'état d'avril 2008. Ces termes décrivent de façon standardisée, la fonction moléculaire, le processus biologique, et la localisation cellulaire du gène. Nous recherchons donc une sur-représentation de termes d'ontologie dans la liste de gènes différentiellement exprimés. Ces termes étant organisés de façon hiérarchisée, généralement plusieurs termes sur-représentés correspondent à un même groupe de gènes. Nous avons utilisé un des nombreux outils disponibles en ligne pour examiner ces représentations de termes : FuncAsso-

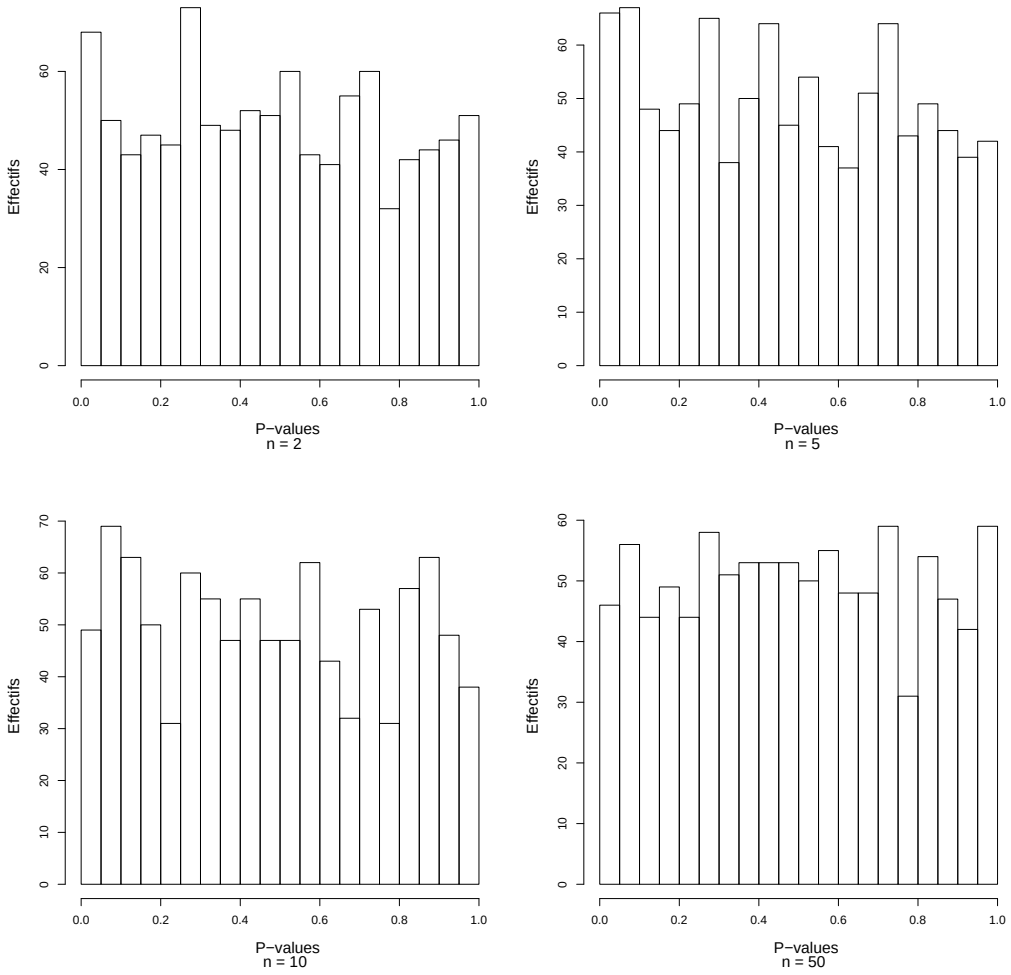


Figure 3 *– Distribution des p-values du test binomial sur les variances de populations, données simulées avec n = 2, 5, 10 et 50 lignées pour estimer la variance, identique entre les populations. Comme il n'y a pas de différence de variance dans les données d'origine, on attend une distribution des p-values uniforme. La distribution n'est pas affectée par le nombre de lignées utilisées, et l'erreur ne l'est donc pas non plus.*

ciate (Berriz *et al.*, 2003). Pourquoi celui-ci ? Au moment où nous avons commencé cette analyse, peu d'outils permettaient de fournir une liste de gènes références pour la sur-représentation, et FuncAssociate le permet. De plus, il est simple d'utilisation, les résultats sont facilement interprétables, et cohérents avec d'autres logiciels testés (notamment DAVID, Huang *et al.* (2009a,b)). Enfin, il est régulièrement mis à jour, ainsi que sa base de donnée. Attention, ces bases de données comme les logiciels d'analyse évoluant avec le temps, chaque analyse étant effectuée à une date précise, elle ne donnera pas nécessairement des résultats parfaitement identiques avec une analyse réalisée avec une autre version des bases de données / logiciels. C'est pourquoi nous disséquons les résultats au niveau des gènes eux-mêmes et de leur(s) fonction(s).

2.2 Analyse intra-spécifique chez *D. simulans* par séquençage de nouvelle génération

2.2.1 Collecte des échantillons de *D. simulans*

Les mouches ont été collectées dans leur habitat naturel. Nous avons collecté un premier échantillon de drosophiles en France métropolitaine, dans un verger de pommiers de l'Institut National de Recherche Agronomique, sur le site de Gotheron près de Valence, dans la Vallée du Rhône (latitude 44°58'20"N et longitude 4°55'39"E). Cette parcelle est dite écologique, c'est-à-dire non traitée. Cependant, elle est environnée de parcelles agricoles classiques, qui subissent des traitements réguliers, il est donc très probable que la population prélevée à cet endroit soit malgré tout soumise à une exposition aux pesticides. Nous avons collecté notre second échantillon dans l'île de Mayotte, située entre Madagascar et la côte africaine, dans une clairière à mi-hauteur, entre les localités de Vahibé et Combani (latitude 12°48'25"S et longitude 45°9'12"E). Cette population a été prélevée dans un environnement relativement répandu à Mayotte : bois de faible densité, correspondant à une forêt secondaire parsemée de parcelles défrichées, avec une grande diversité de végétation. Les prélèvements ont été réalisés à l'aide de pièges, ou au filet directement sur la ressource naturelle (pomme pour Gotheron, banane pour Mayotte). Nous avons fondé pour chaque population 200 lignées isofemelles. 100 sur la ressource naturelle, 100 sur le milieu standard axénique (David, 1962). A l'émergence, les descendants mâles de première génération ont "vieilli" par groupes de 25 au maximum sur le même milieu que le milieu maternel, c'est-à-dire leur milieu de développement. Après cinq jours, les mâles ont été congelés à -80°C. Nous obtenons donc 100 mâles

issus chacun d'une lignée isofemelle différente, et ce pour chaque condition : Gotheron élevée sur Axénique (AG), Gotheron élevée sur pomme (PG), Mayotte élevée sur Axénique (AM), Mayotte élevée sur Banane (BM).

2.2.2 Extraction d'ARN

Pour chaque condition, nous avons réalisé quatre extractions répliques d'ARN, chacune à partir d'un groupe de 25 mâles. Nous avons utilisé le kit Macherey-Nagel Nucleospin RNA II pour l'extraction d'ARN. La concentration et la qualité de l'ARN ont été vérifiées en utilisant à la fois le Nanodrop (Thermo scientific) et une électrophorèse en micropuce (Experion, Biorad). Les quatre extractions d'une même condition ont alors été regroupées pour le séquençage. Pour le transport, les échantillons ont été précipités dans de l'éthanol 100%.

2.2.3 Préparation de la librairie et séquençage

La préparation de la librairie et le séquençage ont été réalisés par l'entreprise GATC Biotech (GATC inc.). A partir des échantillons d'ARN total, les ARN poly(A) ont été sélectionnés pour la synthèse d'ADN complémentaire. L'ADNc a été synthétisé en utilisant une amorce liée à un oligo-(dT) et une transcriptase inverse M-MLV H pour la synthèse du premier brin. Les conditions de réaction ont été choisies de telle manière que les brins d'ADNc était de taille située entre 100 et 500 nucléotides. Pour le séquençage Illumina, une deuxième sélection de la taille a été effectuée via un gel d'agarose préparatoire, afin de se limiter aux fragments mesurant entre 250 et 450 nucléotides. La qualité de la librairie a ensuite été vérifiée par électrophorèse en micropuce (Shimadzu MultiNA). Le séquençage des fragments d'ADNc 3' a ensuite été réalisé sur une machine de séquençage de nouvelle génération "Illumina Genome Analyzer IIx", selon les instructions du fabricant. Nous avons choisi d'utiliser la stratégie 3' Digital Gene Expression (3'DGE). Cette technique sélectionne les extrémités 3' des transcrits pour le séquençage, ce qui augmente la profondeur de quantification par rapport à un séquençage de fragments aléatoires. Elle ne fournit pas contre que peu de couverture sur les séquences, et permet donc mal l'analyse de cette dernière, d'autant que les fragments séquencés étaient ici limités à 32pb.

2.2.4 Cartographie des séquences

Les échantillons ont produits respectivement 31 226 795, 30 960 524, 31 418 537, 27 126 870 de lectures brutes pour les conditions PG, BM, AM et AG. Les séquences ont d'abord été nettoyées

Tableau 2 *– Statistiques de séquençage et cartographie pour les quatre échantillons*

Échantillon	AG		PG		BM		AM	
	#	%	#	%	#	%	#	%
Lectures brutes	27 126 870		31 226 795		31 418 537		30 960 524	
Lectures propres	23 637 316	87,14	26 422 748	84,62	27 497 071	87,52	27 271 618	88,09
Hits sur Dmel	16 397 701	69,37	16 918 600	64,03	17 325 827	63,01	15 915 428	58,36
Hits uniques	14 816 520	62,68	15 509 862	58,7	15 867 700	57,71	14 357 319	52,65
Hits multiples	1 581 181	6,69	1 408 738	5,33	1 458 127	5,3	1 558 109	5,71
Pas de hit	7 239 615	30,63	9 504 148	35,97	10 171 244	36,99	11 356 190	41,64
Hits sur Dsim[1]	4 473 681	18,93	5 430 971	20,55	6 057 828	22,03	7 048 138	25,84
Hits uniques	1 686 322	7,13	2 016 699	7,63	2 284 697	8,31	2 652 715	9,73
Hits multiples	2 787 359	11,79	3 414 272	12,92	3 773 131	13,72	4 395 423	16,12
Pas de hit	2 765 934	11,7	4 073 177	15,42	4 113 416	14,96	4 308 052	15,8

AG : Axénique Gotheron, PG : Pomme Gotheron, BM : Banane Mayotte, AM : Axénique Mayotte. [1]*La cartographie a été réalisée sur* D. simulans *uniquement pour les lectures n'ayant pas pu être cartographiées sur* D. melanogaster *(catégorie "Pas de hit" de* D. melanogaster*). # : nombre.*

avec le logiciel Seqclean, en retirant les adaptateurs de séquençage, et en éliminant les fragments de trop mauvaise qualité. La cartographie a été réalisée par GATC Biotech. Nous avons choisi de cartographier nos séquences d'abord sur le génome de *D. melanogaster*, puis pour les transcrits pour lesquels cela n'avait pas donné de résultat, sur le génome de *D. simulans*. Ce choix est discuté dans la partie de résultat spécifique de cette étude. La cartographie a été réalisée en utilisant le logiciel ELAND fourni par Illumina, en utilisant des séquences de 32 bases de longueur, et en autorisant jusqu'à deux mutations (taux d'erreur de 6,25%). La table 2 montre les statistiques du séquençage, du nettoyage et de la cartographie sur les deux génomes. Les échantillons ont ensuite été normalisés pour le nombre de lectures totales.

2.2.5 Recherche d'insertion d'éléments transposables par PCR

Pour le gène d'intérêt *Cyp6g1* (gène candidat de notre étude, déjà connu dans la littérature, Daborn *et al.* 2002), nous avons examiné des fréquences d'insertion d'éléments transposables par Polymerase Chain Reaction (PCR). Pour cela, nous avons utilisé une PCR en triplex, avec un polymorphisme d'amplification dépendant de la présence de l'élément transposable (voir figure 4). Nous avons ensuite vérifié tous les hétérozygotes, ainsi que les homozygotes par séquençage Sanger.

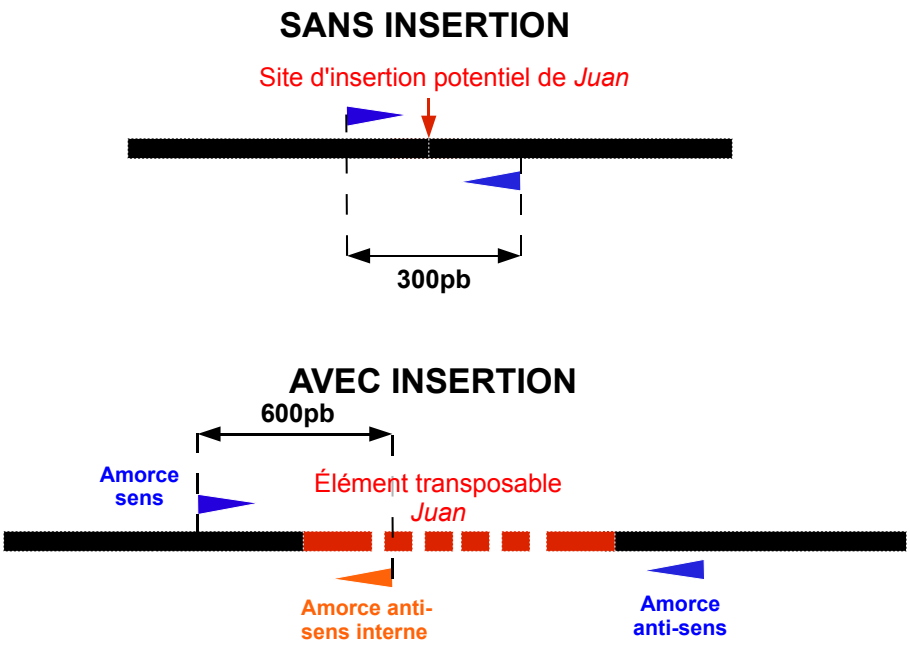


Figure 4 – *Détection de l'insertion de* Juan *par PCR en triplex (trois amorces). En cas d'insertion, nous obtenons un amplifiat de 600 paires de bases (pb), alors que, en absence de l'insertion, nous obtenons un amplifiat de 300pb. L'élément Juan fait environ 4300pb.*

2.2.6 Analyse statistique

L'analyse statistique a été élaborée par Tristan Mary-Huard et Jean-Jacques Daudin de l'équipe INRA / AgroParisTech "Statistiques et Génomes". Les analyses statistiques appliquées aux données de séquençage de nouvelle génération sont généralement basées sur l'utilisation d'une Loi de Poisson surdispersée (Bullard *et al.*, 2010; Robinson et Oshlack, 2010; Salzman *et al.*, 2011). Ces procédures prennent en compte à la fois le caractère discret (comptages) et le caractère surdispersé de ces données. Une loi de poisson surdispersée comprend deux paramètres : λ, la moyenne des comptages, et Φ, le paramètre de surdispersion. On obtient la variance de cette loi en multipliant ces paramètres : $\text{Var}(X) = \lambda \times \Phi$. Elles nécessitent généralement des réplicats biologiques afin d'estimer la variance et le facteur de surdispersion pour chaque gène. Nous avons ici choisi pour des raisons de coût de ne pas répliquer le séquençage (c'est un choix que nous avons regretté *a posteriori*) ; nous avons donc dû adapter les analyses statistiques en fonction de notre plan d'expérience. Nous avons effectué une analyse en deux étapes, sous les hypothèses suivantes :

(1) la large majorité des gènes n'est pas différentiellement exprimée

(2) les gènes avec des niveaux similaires d'expression entre les quatre conditions montrent aussi le même niveau de surdispersion

Dans un premier temps, une analyse gène par gène a été réalisée, en utilisant le modèle suivant une Loi de Poisson surdispersée :

$$X_{gi} \sim \mathcal{P}(\lambda_g, \phi_g),$$

avec X_{gi} l'expression du gène g pour la condition i, λ_g et ϕ_g, respectivement la moyenne et le paramètre de surdispersion associés au gène g. Dans ce modèle, λ_g ne dépend pas de la condition, ce qui est pertinent sous l'hypothèse (1). Les différentes conditions peuvent alors être utilisées comme des réplicats biologiques pour obtenir une estimation $\widehat{\phi}_g$ du paramètre de dispersion ϕ_g. Pour les gènes non différentiellement exprimés, la variance est estimée sans biais, alors qu'elle est sur-estimée pour les gènes différentiellement exprimés. Sous l'hypothèse (2), une estimation plus robuste $\tilde{\phi}_g$ de ϕ_g peut être obtenue en utilisant une estimation locale par Loess (méthode d'estimation de moyenne locale, Cleveland (1979)) de ϕ_g sur les gènes d'expression moyenne similaire. La figure 5 (gauche) montre le paramètre de dispersion en fonction de la moyenne d'expression des gènes, ainsi que la courbe Loess d'estimation de $\tilde{\phi}_g$ (violet). Le Loess est extrêmement proche d'une fonction polynomiale de degré deux (en bleue), qui correspond à la relation trinômiale entre moyenne

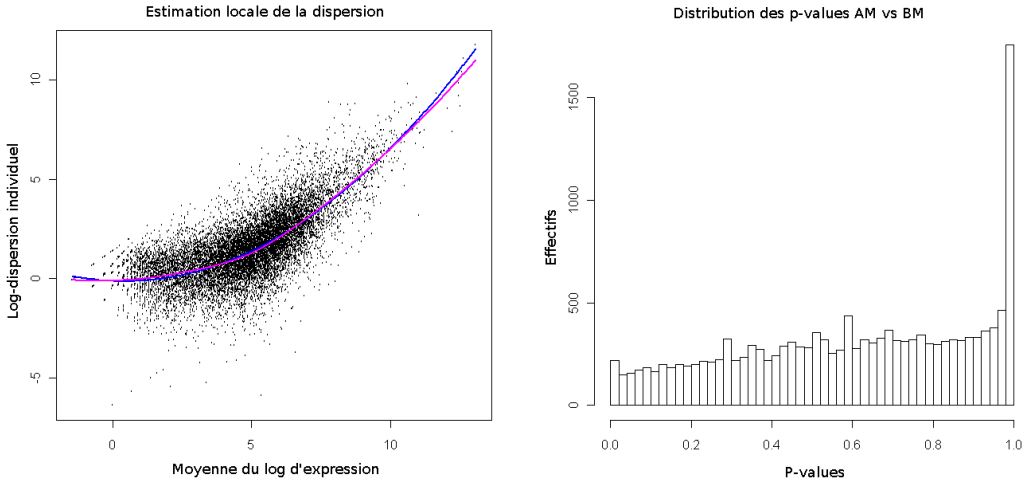


Figure 5 – A gauche : *estimation du paramètre de surdispersion $\widehat{\phi}_g$ en fonction de la moyenne d'expression $\widehat{\lambda}_g$ (échelle logarithmique). Chaque point correspond à un gène. Les courbes violette et bleue représentent respectivement la régression Loess, et l'estimation via une régression polynomiale de degré deux.* ***A droite :*** *histogramme de répartition des p-values pour la comparaison AM vs BM, à titre d'exemple de répartition des p-values dans les comparaisons deux à deux des conditions. On remarque une déviation vers la droite, due à une surestimation du paramètre de surdispersion, ayant pour conséquence une perte de puissance, sans affecter l'erreur.*

et variance de la loi de Poisson surdispersée qui est posée comme hypothèse lors de procédures alternatives (Anders et Huber, 2010; Robinson et Oshlack, 2010).

Dans un deuxième temps, nous avons utilisé une analyse gène par gène utilisant une seconde loi de Poisson surdispersée prenant en compte le paramètre de surdipersion estimé à la première étape :

$$X_{gi} \sim \mathcal{P}(\lambda_{gi}, \tilde{\phi}_g),$$

où λ_{gi} est l'expression moyenne du gène g à la condition i, et $\tilde{\phi}_g$ est le paramètre de surdispersion estimé à l'étape précédente. Nous avons ensuite testé les combinaisons linéaires du paramètre λ_{gi} en utilisant des tests de rapport de vraisemblance et ainsi avons obtenu une p-value. La figure 5 (droite) montre l'histogramme de répartition p-values du test : $H_0 : \{\lambda_{gAM} = \lambda_{gBM}\}$ vs. $H1 : \{\lambda_{gAM} \neq \lambda_{gBM}\}$. On peut observer un décalage des p-values vers 1. Cela montre que la première étape de notre analyse surestime le paramètre de surdispersion (en raison de la présence de gènes différentiellement exprimés). Cette surestimation diminue la puissance de notre test pour détecter des gènes différentiellement exprimés, mais n'affecte pas l'erreur de Type I (Anders et Huber, 2010). Une fois les p-values obtenues, nous avons effectué une correction classique pour les tests multiples

qui contrôle le taux de faux positifs (False Discovery Rate, FDR) (Benjamini et Hochberg, 1995) (FDR = 0,05).

2.2.7 Ontologies de gènes

Nous avons soumis les listes de gènes différentiellement exprimés à une recherche de surreprésentation de terme d'ontologie (version de septembre 2010). Nous avons là aussi utilisé l'application en ligne FuncAssociate (Berriz *et al.*, 2003). Nous avons également examiné manuellement les listes de gènes à la recherche de patrons non détectés par les termes d'ontologies. En effet, ceux-ci peuvent passer à côté de fonctions corrélées, etc. Cependant, l'analyse manuelle reste fortement subjective.

III Résultats et discussion

3.1 Comparaisons intra-, inter-spécifiques et hybrides via des puces à ADN

Le plan d'expérience utilisé ici a permis de mener de front comparaisons d'espèces, comparaisons de populations, et comparaisons entre hybrides et parents. Pour chaque condition, nous avons utilisé quatre lignées représentant quatre réplicats biologiques, et nous nous sommes efforcés d'utiliser des lignées fondées récemment afin de limiter les effets de l'évolution en laboratoire. Nos résultats seront donc représentatifs des populations étudiées, et non simplement de ces lignées. Après toutes les éliminations de gènes pendant la normalisation et l'analyse statistique, nous avons pu évaluer l'expression pour 4398 gènes, soit environ un quart du génome de la drosophile.

Biais inhérent à l'utilisation de puces hétérospécifiques

Une limite potentielle de cette étude résidait dans l'utilisation de puces portant des séquences d'ADN complémentaires de *D. melanogaster* pour comparer des espèces proches : *D. sechellia* et *D. simulans*. En effet, la divergence de séquence entre les sondes et l'ADNc d'intérêt peut créer des biais dans les mesures d'expression (Gilad *et al.*, 2005; Oshlack *et al.*, 2007). Gilad *et al.* (2005) ont évalué l'importance du problème en construisant des puces multi-espèces pour quatre espèces de primates. Ils ont montré de sévères biais d'hybridation dus à la divergence entre sonde et ADNc d'intérêt. Cependant, bien que le problème soit important si on souhaite comparer l'espèce sur la quelle la puce a été dessinée avec une autre, Oshlack *et al.* (2007) ont aussi montré que la mesure d'expression d'une espèce sur les puces d'une espèce proche ne causait "pas de perte d'information discernable". Ici, nous comparons *D. simulans* et *D. sechellia* en utilisant des puces de *D. melanogaster*. Cette dernière n'étant pas impliquée dans la comparaison, le biais est probablement faible, à condition que la divergence entre *D. melanogaster* et *D. simulans* d'un côté

et *D. melanogaster* et *D. sechellia* de l'autre soit équivalente. La divergence entre *D. simulans* et *D. sechellia* est datée d'environ 250 000 ans, alors que la séparation de ces deux espèces avec *D. melanogaster* est estimée à deux à trois millions d'années (Lachaise *et al.*, 1988; Hey et Kliman, 1993; McDermott et Kliman, 2008). La séparation de nos deux espèces est huit à douze fois plus récente que leur séparation avec *D. melanogaster*, ce qui nous permet de supposer que le signal est affecté de façon similaire pour les deux espèces. De plus, afin de limiter ce biais, nous avons choisi des puces portant des cDNA complets plutôt que des oligonucléotides, ce qui limite l'influence des mismatchs (Hsieh *et al.*, 2003). Malgré cela, il est clair que si en moyenne, le biais est peut-être évité, ou à tout le moins limité, il est aussi probable que certains gènes soient fortement affectés par l'hybridation inter-spécifique.

Nous avons effectué plusieurs vérifications grâce à des scripts *ad hoc*, en utilisant le langage de programmation Perl, souvent couplé au logiciel R pour les analyses statistiques. Pour déterminer si la divergence avec *D. melanogaster* est ressentie au niveau de la mesure d'expression, nous avons d'abord cherché une corrélation entre 1- la différence de divergence de séquences entre chacune des deux espèces et *D. melanogaster* et 2- la différence d'expression moyenne entre les deux espèces. Certes, cette analyse n'est pas parfaite, puisqu'elle ne différencie pas le polymorphisme partagé par *D. simulans* et *D. sechellia* et une divergence spécifique de chaque espèce. Cependant, c'est une première indication. Nous avons réalisé cette analyse sur 2303 gènes (ceux des puces pour lesquels l'orthologie était bien annotée, ce qui peut également créer un biais, au moins dans le taux de divergence), en utilisant les séquences issues des bases de données (http ://flybase.org, Tweedie *et al.* (2009)). Nous n'avons pas observé de corrélation (Coefficient de corrélation de Spearman : $\rho = -0,0054$; p-value = 0,8056). *D. simulans* et *D. sechellia* divergent de moins de 2%, ce qui est comparable au niveau de divergence entre des populations disparates de *D. melanogaster*. 80% des bases qui divergent avec *D. melanogaster* sont identiques chez *D. sechellia* et *D. simulans* (Dworkin et Jones, 2009). Dans cette étude, les auteurs n'ont pas observé de biais entre *D. sechellia* et *D. simulans* sur la mesure d'expression, en utilisant des puces dessinées à partir de *D. melanogaster*. Mezey *et al.* (2008) ont montré que la divergence de séquences a pour effet d'augmenter la variance, ce qui conduit à une diminution de la puissance du test, sans affecter l'erreur.

Nous avons enfin effectué une dernière vérification de l'influence de ce biais. Nous avons calculé la différence de divergence de séquence codante (*D. melanogaster* vs. *D. simulans* moins *D. melanogaster* vs. *D. sechellia*). Nous avons effectué ce calcul pour 232 gènes différentiellement exprimés entre *D. sechellia* et *D. simulans* (parmi les 304 gènes communs à toutes les comparaisons

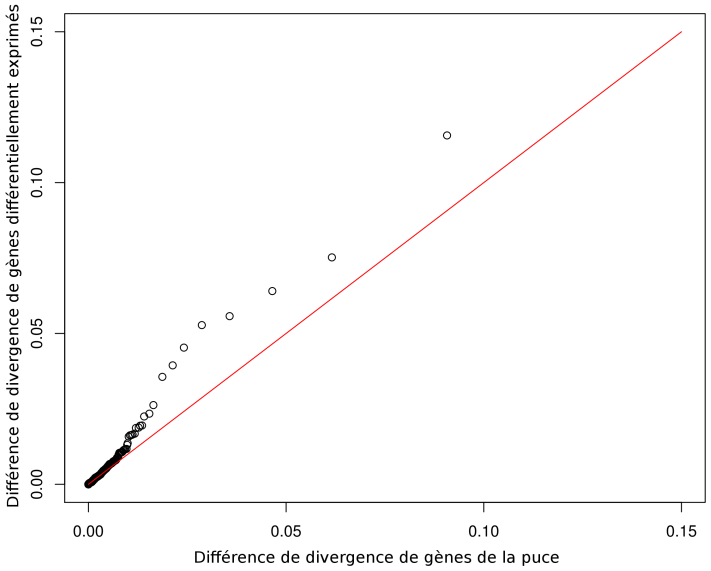


Figure 6 *– Ce diagramme représente les quantiles d'une distribution par rapport à ceux de l'autre. En abscisse, la différence de divergence (*D. melanogaster *vs.* D. simulans *moins* D. melanogaster *vs.* D. sechellia*) pour 2303 gènes parmi ceux de la puce, et en ordonnée, la différence de divergence pour 232 gènes parmi les gènes différentiellement exprimés entre* D. sechellia *et* D. simulans*. Au delà de environ 1,3% de différence de divergence, on observe une déviation vers les gènes différentiellement exprimés.*

D. sechellia/ *D. simulans*, quelle que soit la population), et pour le set de 2303 gènes de la puce précédemment décrit. Bien que la p-value soit faible, la différence de divergence moyenne n'est pas significativement différente entre les gènes différentiellement exprimés et les gènes pris aléatoirement (Mann-Whitney, W = 286,527, p-value = 0,06643). Cependant, la faible p-value nous a incité à pousser cette analyse un peu plus loin. Nous avons donc utilisé un diagramme de quantiles (Q-Q plot) pour comparer les deux distributions. Ce diagramme trace les quantiles d'une distribution par rapport aux quantiles de l'autre (figure 6). On observe une très bonne cohérence jusqu'à environ 1,3%. A partir de 1,3%, il y a une déviation en direction des gènes différentiellement exprimés, et la déviation est de plus en plus forte lorsque la divergence augmente. La différence de divergence est plus forte pour les gènes différentiellement exprimés que pour les gènes aléatoires. Attention, il ne s'agit pas de 1,3% de divergence entre *D. simulans* et *D. sechellia*, mais bien de différence de divergence entre *D. melanogaster* et *D. simulans* d'un côté, *D. melanogaster* et *D. sechellia* de l'autre. Ce biais concerne une trentaine de gènes parmi les 232. Ce résultat a deux explications possibles. D'abord, il peut être dû à la divergence de séquences qui affecterait l'hybridation, générant quelques faux positifs. Mais il peut également s'expliquer par une plus grande divergence

d'expression des gènes qui ont aussi plus divergé en séquences. Il s'avère que tous les gènes étudiés ensuite en détail sont situés dans la partie non biaisée de ce graphique, c'est-à-dire avec une différence de divergence inférieure à 1,3%. Ce résultat est néanmoins une des raisons qui nous a poussées à utiliser les techniques de séquençage de nouvelle génération dans la seconde partie de cette thèse.

3.1.1 Comparaisons hybrides contre populations parentales

Amplitude de la perturbation d'expression chez les hybrides

Nous avons comparé l'expression chez les "populations" hybrides. Un gène a été considéré comme différentiellement exprimé, si et seulement si il était différent de chacun des parents, et de la moyenne des parents. Aucun gène révélé (différentiellement exprimés) n'avait d'expression intermédiaire entre les deux parents. Nous nous attendions à montrer des effets assez forts, les hybrides étant stériles (Cabot *et al.*, 1994; Joly *et al.*, 1997). De plus, Haerty et Singh (2006) ont révélé pas moins de 241 gènes différentiellement exprimés entre hybrides et parents chez ces mêmes espèces. Pourtant, nous n'avons observé que huit gènes différentiellement sur-exprimés chez les hybrides par rapport aux parents (soit 1,81%), et aucun sous-exprimé, soit significativement plus de sur-expression que de sous-expression (test de χ^2 à 1 degré de liberté, p-value $= 3,08 \times 10^{-16}$) (table 3). Ces gènes ont été révélés dans deux comparaisons uniquement : celle impliquant la population d'origine française, et celle impliquant la population d'origine zimbabwéenne (soit les deux populations les plus éloignées de la zone d'origine de *D. simulans* et de la zone de répartition de *D. sechellia*). Un seul gène sortait dans ces deux comparaisons : *Cp110*.

Pourquoi une aussi faible différenciation dans notre étude par rapport à celle de Haerty et Singh (2006), qui ont révélé pas moins de 241 gènes ? Plusieurs hypothèses peuvent expliquer cette différence :

- Haerty et Singh (2006) n'ont pas différencié les effets additifs. Peut être certains de leurs gènes ont donc une expression proche de la moyenne de leurs parents, une possibilité que nous avons exclue ici
- la lignée de *D. simulans* utilisée par Haerty et Singh (2006) est une lignée collectée en Arizona, population fortement éloignée de la zone d'origine de l'espèce, qui pourrait donc avoir subi une adaptation locale, ainsi qu'un effet de la dérive plus important lié à une séparation plus ancienne avec les populations de la zone ancestrale, voire à l'histoire évolutive de la population (migration par fondations successives par exemple)

– l'étude de Haerty et Singh (2006) a examiné uniquement l'expression testiculaire

Ce dernier point présente l'explication la plus plausible. En effet, les hybrides étant stériles (Cabot *et al.*, 1994; Joly *et al.*, 1997), il est probable que l'expression dans cet organe soit fortement perturbée, d'autant que les testicules sont atrophiés et les spermatozoïdes mal formés (Joly *et al.*, 1997). En règle générale, les testicules montrent un patron d'expression extrêmement spécifique (Wolgemuth et Watrin, 1991; Grimes, 2004; Catron et Noor, 2008). L'étude de Haerty et Singh (2006) aura donc montré des différences qui ne seraient pas détectables dans notre étude sur corps entier. Notre but était d'examiner une divergence ubiquitaire et globale, sans *a priori*, c'est pourquoi nous nous sommes intéressés au corps entier.

De même, l'étude de Michalak et Noor (2003) a montré 51 gènes différentiellement exprimés (sur $\approx$ 14000 examinés, soit 0,36%) entre hybrides et parents (l'étude portait sur *D. mauritiana* et *D. simulans*). Cette étude n'a pas utilisé de correction pour les tests multiples. Sans test multiple, nous obtenons en moyenne 67,5 gènes différentiellement exprimés entre hybrides et parents sur nos 4398 examinés, soit une proportion de gènes différentiellement exprimés significativement supérieure à celle observée par Michalak et Noor (2003) (test de χ^2 à 1 degré de liberté, p-value $= 2,57 \times 10^{-17}$). Aucun gène différentiellement exprimé n'est commun entre cette étude et la nôtre. Les test ont été réalisés à 5%, il devrait donc y avoir par hasard 5% des tests qui seront significatifs (en moyenne), soit 700 gènes. Or, sur leurs deux comparaisons de l'hybride avec un parent, ils observent un total de 435 gènes différentiellement exprimés sur les deux comparaisons, soit un déficit de gènes différentiellement exprimés par rapport à ce qui est attendu par hasard. Comment expliquer cela ? Une perte de puissance due à la technique que ce soit au niveau des puces, de l'analyse statistique utilisée,... ? Lors de leur analyse statistique, ils corrigent la variance biologique de leurs lignées afin d'éviter qu'un gène sorte artificiellement simplement à cause d'une mauvaise estimation de la variance biologique. Il est probable qu'ils aient surestimé la variance biologique, causant alors une forte perte de puissance. Moehring *et al.* (2007) ont également comparé les hybrides avec les parents sur corps entier, excluant toute additivité comme dans notre analyse. Cependant, là aussi, ils n'ont pas utilisé de correction pour les tests multiples, laissant donc un plus grand nombre de faux positifs. Leur analyse a révélé 220 gènes différentiellement exprimés sur une puce qui en comportait 11 896. Ce qui n'est pas significativement différent des 67,5 gènes observés dans notre étude en absence de correction pour test multiple (test de χ^2 à 1 degré de liberté, p-value $= 0,16$).

Si on analyse tous ces résultats globalement, il semble que les études utilisant le corps entier

Tableau 3 *– Gènes significativement surexprimés chez les hybrides par rapport aux parents, et localisation chromosomique*

Populations parentales	Bras chromosomique
F *D. simulans* x *D. sechellia*	
Cp110	X
CG14785	X
CG4558	X
Es2	X
Z *D. simulans* x *D. sechellia*	
Cp110	X
sm	2R
r-cup	X
CG3795	X
CG31108	3R

Parmi les gènes différentiellement exprimés, on note une sur-représentation de gènes liés au chromosome X. F : France, Z : Zimbabwe. 2R, 3R : bras droit des chromosomes 2 et 3. Il n'y a aucun gène différentiellement exprimé entre hybrides issus des populations du Kenya et des Seychelles et leurs parents.

révèlent peu de différences au niveau de l'expression (Michalak et Noor, 2003; Moehring *et al.*, 2007; Wurmser *et al.*, 2011) alors que l'étude portant sur les testicules (Haerty et Singh, 2006) révèle de fortes perturbations d'expression. Il semble donc que les perturbations chez l'hybride soient concentrées sur les testicules, et notamment sur les gènes liés à la spermatogenèse tardive (Moehring *et al.*, 2007; Catron et Noor, 2008), ce qui expliquerait les différences de patron d'expression observé entre les études.

Gènes différentiellement exprimés entre hybrides et parents

La table 3 montre les gènes différentiellement exprimés entre les populations hybrides et leurs populations parentales. Il y avait respectivement quatre et cinq gènes différentiellement exprimés entre hybrides et parents pour les populations de France et de Zimbabwe et aucun pour les populations du Kenya et des Seychelles. Un seul gène est commun à ces deux comparaisons : *Cp110*.

***Cp110* : un gène impliqué dans la réplication du centriole** D'après Dobbelaere *et al.* (2008), *Cp110* est impliqué dans la duplication du centriole. Sa perturbation d'expression implique peut-être une perturbation de la formation du fuseau au moment de la méiose, dans laquelle le centriole est impliqué (Gogendeau et Basto, 2010). Le centriole est également impliqué dans la construction et l'utilisation d'un flagelle fonctionnel chez les spermatozoïdes. Des perturbations dans la duplication du centriole entraînent la stérilité (Gogendeau et Basto, 2010), et le rôle de *Cp110* dans la duplication du centriole semble central (Dobbelaere *et al.*, 2008). Dans cette étude,

les auteurs ont montré que la sur-expression de ce gène entraine la formation de structures "fibre-like" dans le cytoplasme. Tout cela reste toutefois très spéculatif, d'autant que la perturbation d'expression observée dans notre étude est présente dans seulement deux populations sur quatre.

Autres gènes différentiellement exprimés Parmi les autres gènes différentiellement exprimés, un seul a une fonction complètement inconnue (*CG4558*). Deux d'entre eux semblent avoir une fonction plus ou moins bien définie, liée au cytosquelette (*CG14785* et *CG31108*), et ce rôle peut être lié au rôle de *Cp110*, même si cela reste hautement spéculatif. Deux gènes semblent avoir un rôle dans le développement du système nerveux (*Es2* et *smooth*). Le gène *r-cup* est impliqué dans la méiose mâle et est donc un excellent candidat pour un rôle dans la stérilité hybride (Benson *et al.*, 2006; Barreau *et al.*, 2008). Enfin le gène *CG3795* est une endopeptidase de type chymotrypsine. Aucun de ces gènes n'était différentiellement exprimé entre hybrides et espèces parentales dans l'étude de Haerty et Singh (2006).

Localisation des gènes perturbés : sur-représentation du chromosome X

Sur les huit gènes différentiellement exprimés entre hybrides et parents, six sont situés sur le chromosome X, soit significativement plus qu'attendu sous l'hypothèse d'une répartition homogène sur les bras chromosomiques par rapport aux gènes de la puce (test de Fisher exact, p-value $= 4,36 \times 10^{-4}$). Les changements d'expression observés sont des changements dominants chez l'hybride, et on peut penser qu'ils sont délétères chez le mâle (stérilité), mais pas forcément chez la femelle (celle-ci étant généralement fertile). Sous cette hypothèse, on attend conformément à ce que nous avons observé, un excès de ces mutations sur le chromosome X (Ellegren et Parsch, 2007). Ces observations supportent l'hypothèse d'effets antagonistes de ces gènes entre les deux sexes. Ce patron d'expression est en adéquation avec la théorie dite du "faster-X", propriété du processus de spéciation toujours fortement débattue (Betancourt *et al.*, 2002; Thornton et Long, 2002; Musters *et al.*, 2006; Begun *et al.*, 2007; Masly et Presgraves, 2007; Presgraves, 2008; Vicoso et Charlesworth, 2009). Selon cette théorie, les gènes liés au chromosome X évoluent en moyenne plus rapidement que les gènes autosomaux, probablement à cause d'une sélection plus efficace sur le X hémizygote chez les mâles. Cependant, si on regarde dans le détail, il semble que les gènes liés à l'X soient biaisés vers les extrêmes : ils sont généralement sous sélection, que celle-ci soit positive ou négative. Parmi ceux-ci, les gènes à l'expression notamment testiculaire sont sous forte sélection positive (Vicoso et Charlesworth, 2009). Les gènes liés au chromosome X ont une évolution particulière, due à leur présence deux tiers du temps chez les femelles, et un tiers chez les mâles, ainsi qu'à leur taille efficace

de population plus faible. Cependant, par une approche d'introgression, Hollocher et Wu (1996) n'ont pas montré de densité plus grande de facteurs de stérilité sur le X que sur les autosomes. En résumé, il y a donc 1- une sur-représentation de l'X parmi les gènes dont l'expression est perturbée chez les hybrides (ici, sur-expression). 2- une évolution plus rapide en terme de séquences des gènes liés à l'X, et notamment ceux liés à la reproduction mâle et à l'expression testiculaire. 3- une densité de facteur de stérilité homogène entre les différents chromosomes. Tout cela semble soutenir l'hypothèse de perturbations multi-factorielles, perturbations liées à des interactions entre le chromosome X et les autosomes, hypothèse soutenue par de nombreux travaux sur la stérilité (Coyne *et al.*, 1991; Hollocher et Wu, 1996; Masly et Presgraves, 2007).

Rôle de la régulation *cis*

Notre étude, comme d'autres études a révélé très peu de perturbations chez les hybrides inter-spécifiques de la drosophile (Michalak et Noor, 2003; Moehring *et al.*, 2007; Wurmser *et al.*, 2011), ce qui signifie peu de dérégulation en *trans*, c'est-à-dire causée par un autre gène non lié. Par ailleurs, nous avons montré de fortes divergences d'expression entre espèces parentales (voir plus loin). Tout cela est cohérent avec un modèle de fortes contraintes sur la régulation *trans*, souvent pléiotrope, et de contraintes moins fortes sur la régulation *cis*, qui par essence est généralement limitée à un seul locus. Cette observation est confirmée par d'autres études (Wittkopp *et al.*, 2004, 2008a,b).

Hétérosis

Nous avons examiné les patrons d'hétérosis. Nous n'avons inclus dans cette étude que les gènes pour lesquels l'expression était différente entre les populations parentales. Pour cela, nous avons utilisé le rapport d/a (voir méthode page 32). Environ 45% de nos gènes montrent un patron qui va de l'absence totale d'hétérosis à la à un effet hybride léger ($0 < d/a < 0,5$, gènes ultérieurement désignés comme ne montrant pas d'effet hybride), environ 26% montrent un effet hybride non transgressif fort ($0,5 < d/a < 1$), et 29% montrent un effet hybride transgressif ($d/a > 1$) (figure 7).

Les patrons d'hétérosis mis en évidence dans cette étude sont intermédiaires entre l'effet hybride très faible (quelques pour cent) qui a pu être observé dans certaines études (Gibson *et al.*, 2004; Haerty et Singh, 2006), et l'hétérosis forte observée dans deux autres études (Hughes *et al.*, 2006;

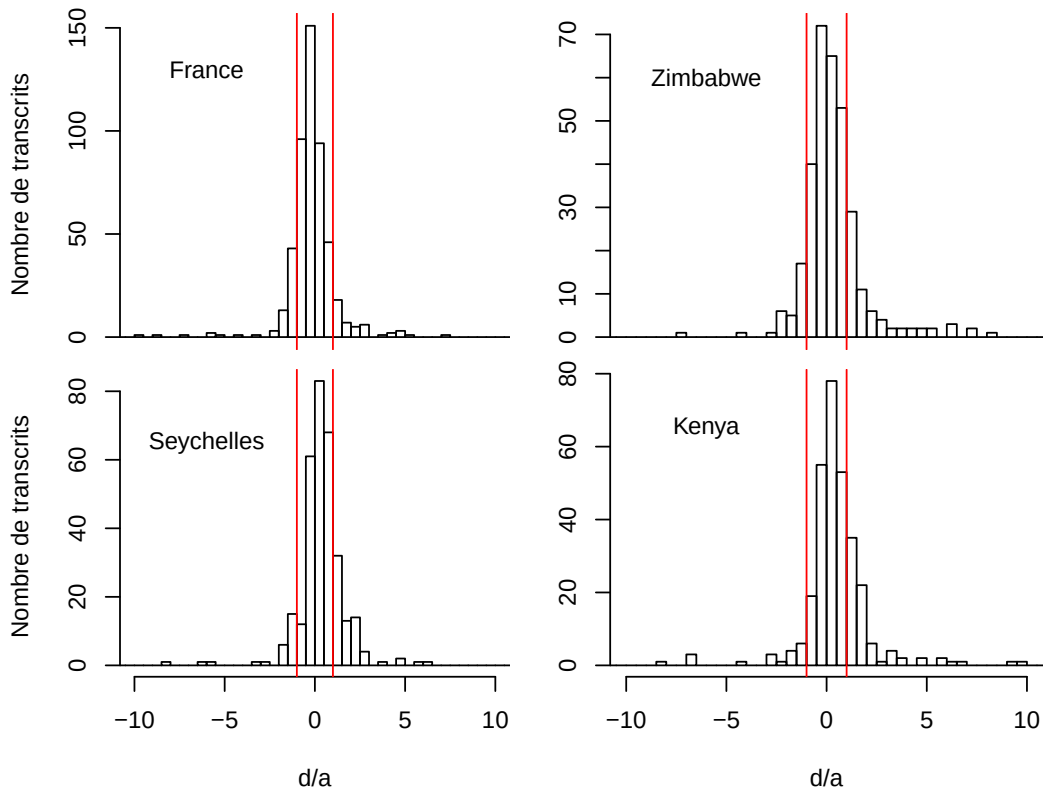


Figure 7 *– Distributions du rapport d/a pour les quatre croisements hybrides issus de nos quatre populations de* D. simulans. *Quand $d = 0$, il n'y a pas d'effet hybride. De 0 à 1, nous passons progressivement de l'absence d'hétérosis à un effet hybride non transgressif. Au dessus de 1 (ou en dessous de -1), il s'agit d'effet hybride transgressif. Les barres rouges correspondent aux seuils -1 et 1.*

Rottscheidt et Harr, 2007). Les raisons pouvant expliquer ces différences sont nombreuses, et ont été très bien détaillées par Rottscheidt et Harr (2007) :

- le taux de divergence des lignées parentales : plus les lignées se seraient séparées récemment, plus l'absence d'hétérosis serait la règle (Hughes *et al.*, 2006)
- le taux de consanguinité des lignées parentales : en effet, si celles-ci sont fortement consanguines, il y aura beaucoup plus d'effets hybrides transgressifs
- la méthodologie utilisée pour déterminer l'hétérosis. Bien que nous ayons essayé de nous conformer au maximum aux méthodes suivies dans les études précédentes, il est possible que des différences méthodologiques aient une influence sur les résultats

L'utilisation de corps entiers ou d'organes particuliers crée probablement aussi des biais, comme nous l'avons vu précédemment. Il peut également y avoir des biais dus aux puces elles mêmes : simple ou double canal, à fluorescence ou à radioactivité, mais surtout oligonucléotides ou cDNAs complets.

3.1.2 Comparaison d'espèces

Amplitude des différences

Nous avons comparé l'expression entre *D. sechellia* et *D. simulans* pour chacune des quatre populations de *D. simulans*. Nous avons révélé 518, 347, 353 et 337 gènes différentiellement exprimés pour les populations respectivement de France, du Zimbabwe, des Seychelles et du Kenya (table 4). Parmi ces gènes, 304 sont communs à toutes les comparaisons, et seront par la suite désignés comme "gènes différentiellement exprimés entre *D. sechellia* et *D. simulans*". Les trois populations africaines montrent autant de gènes différentiellement exprimés, alors que la population française en présente significativement plus (test de χ^2 à un degré de liberté, p-value $= 1,89 \times 10^{-9}$). Cette observation est en accord avec la structuration de *D. simulans* observée sur les séquences codantes (Veuille *et al.*, 2004; Baudry *et al.*, 2006) ou microsatellites (Schöfl et Schlötterer, 2004, 2006). Cependant, ces études ont montré une différenciation des populations européennes par rapport aux populations de la zone ancestrale (du Kenya à Madagascar), mais elles ont également montré une différenciation des populations d'Afrique du sud, notamment du Zimbabwe. Nous ne retrouvons pas cette différenciation au niveau des comparaisons d'expression avec *D. sechellia* (pas plus d'ailleurs qu'en comparaison directe de populations, voir page 57). Pourquoi la population du Zimbabwe, considérée comme dérivée par les études de séquences, se comporte-t-elle au niveau de l'expression comme les populations de la zone ancestrale ? Est-ce une question d'environnement, proche entre

Tableau 4 – *Nombre de gènes différentiellement exprimés entre* D. sechellia *et* D. simulans, *pour les différentes populations de* D. simulans

Population	Total	Sur-exprimés[a]	Sous-exprimés[b]
Zimbabwe	347	148	199
Kenya	337	144	193
Seychelles	353	158	195
France	518	214	304
Communs	304	131	173

[a] *Gènes sur-exprimés chez* D. simulans *par rapport à* D. sechellia ; [b] *Gènes sous-exprimés chez* D. simulans *par rapport à* D. sechellia.

cette population et celles du Kenya et des Seychelles ? Le seul indice qui peut laisser penser que la population du Zimbabwe n'est pas du point de vue de l'expression parfaitement équivalente aux autres populations africaines, c'est son comportement lors des comparaisons hybrides (voir plus haut) qui montre des différenciations chez les hybrides, comme pour la population française, et contrastant avec les populations d'Afrique de l'est.

Nous avons comparé notre liste de 304 gènes différentiellement exprimés avec les gènes révélés dans une autre étude comparant *D. simulans* et *D. sechellia* (Dworkin et Jones, 2009). 28% des gènes différentiellement exprimés dans leur étude étaient présents sur notre puce (soit 36 gènes). Parmi ceux-ci, 30.5% (11 gènes) sont également présents dans notre liste de gènes différentiellement exprimés. Ces différences peuvent avoir plusieurs explications :

- le choix des lignées : nous avons utilisé des lignées récemment collectées, alors qu'ils ont utilisé des lignées entretenues longtemps en laboratoire
- la puissance des puces : il est possible qu'il y ait une différence de puissance entre leurs puces et les nôtres (ils ont révélé 136 gènes différentiellement exprimés sur 11880 sur la puce, nous 304 sur 4398, soit une différence significative, test de χ^2 à un degré de liberté, p-value = $2,78 \times 10^{-90}$). Cependant, si nos puces sont plus puissantes, nous aurions dû retrouver leurs gènes différentiellement exprimés
- les sexes étudiés : ils ont travaillé sur des mélanges de femelles et de mâles, alors que nous avons travaillé sur des mâles uniquement, et diverses études ont montré que le sexe a une influence très forte sur les patrons d'expression (Meiklejohn *et al.*, 2003; Haerty et Singh, 2006; Zhang *et al.*, 2007; Muller *et al.*, 2011)

Tableau 5 *– Termes de Gene Ontology sur-représentés dans les gènes sur-exprimés chez* D. simulans *par rapport à* D. sechellia *quelque soit la population de* D. simulans

Type de Terme	Terme GO	Identifiant GO
MF	electron carrier activity	GO :0009055
BP	lipid metabolism	GO :0006629
BP	hormone catabolism	GO :0042447
CC	vesicular fraction	GO :0042598
CC	microsome	GO :0005792

MF : Fonction moléculaire, BP : processus biologique, CC : localisation cellulaire. Ces termes sont dans l'ordre décroissant de la significativité de leur sur-représentation. Nous ne pouvons fournir de p-value, car celles-ci ont été calculées indépendamment pour chaque population.

Analyse d'ontologie

Nous avons donc étudié nos 304 gènes différentiellement exprimés en utilisant une approche basée sur les termes de Gene Ontology. Nous avons analysé séparément les gènes sur-exprimés chez *D. simulans* et ceux sous-exprimés.

La table 5 présente les termes de Gene Ontology sur-représentés d'après FuncAssociate pour les gènes significativement plus exprimés chez *D. simulans* par rapport à *D. sechellia*, quelle que soit la population de *D. simulans* concernée. Nous avons examiné les gènes correspondant à ces termes dans notre liste de gènes différentiellement exprimés. La fonction moléculaire "electron carrier activity", et les localisations cellulaires "vesicular fraction" et "microsome" se réfèrent à des gènes de la superfamille des cytochromes P450. Cette superfamille est impliquée notamment dans les processus de détoxification. Les deux autres termes d'ontologie se réfèraient au moment de notre analyse (mai 2008) à cinq gènes, dont trois sont présents dans notre liste de gènes différentiellement exprimés : *Juvenile hormone epoxide hydrolase 1* (*Jheh1*), *Juvenile hormone epoxide hydrolase 3* (*Jheh3*) et *Dopamine N-acétyltransférase* (*Dat*). Ce dernier gène ne figure plus actuellement dans les gènes désignés par ce terme. Nous n'avons pas détecté de différences d'expression sur les Odorant Binding Protein (Obp), mais trois gènes mentionnés dans des études précédentes (*Odorant binding protein 57d*, *Obp57e*, Matsuo *et al.* (2007) et *Obp56e*, Dworkin et Jones (2009)) ne sont pas présents sur notre puce, et leur expression n'a donc pas pu être examinée. Ces trois gènes semblent jouer un rôle dans l'écologie particulière de *D. sechellia*.

Cytochrome P450 : relâchement de la sélection pour la détoxification chez *D. sechellia* ?

Différentes études se sont intéressées à la génétique de cette tolérance et de cette attraction, ou à tout le moins de cette absence de répulsion. Des études de génétique ont montré que la résistance était dominante, et qu'elle était due à différents facteurs répartis sur tous les chromosomes exceptés les deux chromosomes très pauvres en gènes (4 et Y) (Jones, 1998, 2005). Matsuo *et al.* (2007) ont montré que les *Obp57d* et *Obp57e* étaient impliqués dans la répulsion des autres espèces au *Morinda*. Nous nous attendions donc à trouver des gènes impliqués dans ces phénomènes parmi les gènes différentiellement exprimés entre *D. sechellia* et *D. simulans*, mais également des gènes s'étant différenciés après la spécialisation de *D. sechellia*, et pas forcément en rapport direct avec l'écologie particulière de *D. sechellia*. Par une simple étude de transcriptome, il n'est pas possible de distinguer parmi les gènes différentiellement exprimés ceux qui seraient la cause de la différenciation, et ceux qui en seraient la conséquence. Autrement dit, nous ne pouvons déterminer quels gènes ont été directement impliqués dans la séparation des espèces, et quels gènes ont divergé après coup par des processus plus ou moins stochastiques.

Parmi les gènes significativement moins exprimés chez *D. sechellia* par rapport à *D. simulans*, on trouve donc sept cytochromes P450. Les cytochromes P450 composent une large famille de gènes composée d'environs 83 (Tijet *et al.*, 2001) à 87 (Wu *et al.*, 2011) gènes fonctionnels chez *D. melanogaster*. Il y en aurait 85 fonctionnels pour 7 pseudogènes chez *D. simulans*, et 75 fonctionnels pour 17 pseudogènes chez *D. sechellia* (Clark *et al.*, 2007; Wu *et al.*, 2011). Il y a donc probablement eu une dizaine de pseudogénisations chez *D. sechellia* depuis la séparation avec *D. simulans*. Nous avons donc une chute de l'expression de gènes de détoxification, accompagnée d'une large pseudogénisation dans cette famille, ce qui permet de supposer un rôle moins important de cette dernière, ou au moins d'une partie des gènes qui la composent, chez *D. sechellia*. La chute d'expression de gènes de la famille des cytochromes P450 a aussi été observée par Dworkin et Jones (2009). Ils interprètent cette chute d'expression comme une conséquence de la spécialisation de *D. sechellia*. En s'associant de façon très stricte à la plante *Morinda citrifolia*, *D. sechellia* aurait réduit la variété de toxines auxquelles elle est exposée, relâchant ainsi la pression de sélection sur certains gènes de détoxification comme ces cytochromes P450, mais aussi une Glutathion transférase (GST) et une Glucuronosyl-transférase (deux gènes de détoxification, nous reviendrons sur ces familles dans la seconde étude). S'il est possible que ces observations soient liées à la spécialisation, il est aussi possible qu'elles soient dues à une autre caractéristique de *D. sechellia*, que ce soit au niveau de

son environnement, ou de son génome. Cependant, ces différences étant constantes quelle que soit la population de *D. simulans* avec laquelle *D. sechellia* est comparée, il est plus probable qu'elles soient constitutives et causées par l'évolution de *D. sechellia* lors de la spécialisation.

Régulation hormonale

Parmi nos gènes sur-exprimés chez *D. simulans* par rapport à *D. sechellia*, les termes d'ontologies désignent deux à trois (selon la version de la base de données d'ontologies) gènes impliqués dans les régulations hormonales. *Jheh1* et *Jheh3* catabolysent l'hormone juvénile (JH). Selon Gruntenko et Rauschenbach (2008), la quantité de JH peut être estimée via son taux de dégradation. La plus forte expression de deux *Jheh* chez *D. simulans* implique donc une plus faible quantité de JH. Parmi les gènes sur-exprimés chez *D. simulans* par rapport à *D. sechellia*, se trouve également un gène induit par la JH (Juvenile hormone-inducible protein 26, *JhI-26*), qui n'a pas de rôle précisément décrit. On trouve également un gène induit par l'ecdysone, une hormone aux interactions nombreuses avec la JH (Ecdysone-inducible gene L2, I *mpL2*). Ce gène a un rôle dans la régulation de l'insuline. La JH est une hormone dont le mécanisme moléculaire n'est pas encore connu. Elle a un rôle prépondérant chez la larve, notamment pour la régulation des mues, et chez la femelle, notamment pour la vitellogenèse et l'ovogenèse (Gruntenko et Rauschenbach, 2008; Liu *et al.*, 2008). Cependant, son rôle chez les mâles est très mal connu, il est donc difficile d'interpréter les différences observées ici. Une approche physiologique a mis en évidence le rôle de la JH dans l'accumulation des protéines du fluide séminal dans les glandes accessoires (Wolfner *et al.*, 1997), et ce rôle a pu être confirmé par une étude basée sur des mutants (Wilson *et al.*, 2003). Les mutants dont la réceptivité à la JH est diminuée, montrent une moindre accumulation de protéines, ce qui affecte donc la fertilité mâle. Ces mutants montrent également une altération du comportement de cour, ce qui suggère un rôle de la JH dans ce comportement, que ce soit directement ou indirectement via la perturbation de la synthèse protéique dans les glandes accessoires.

Cependant, René Feyereisen (communication personnelle), spécialiste des détoxifications (Feyereisen, 1999; Tijet *et al.*, 2001; Feyereisen, 2006; Le Goff *et al.*, 2006), met en doute la spécificité des *Jheh*. Selon lui, ce sont effectivement des epoxydes hydrolases, mais pas limitées à l'hormone juvénile. Il est donc possible qu'elles soient impliquées dans une autre branche du métabolisme, et non directement avec la JH. Il est possible que ces gènes soient également impliqués dans les détoxifications.

Deux gènes impliqués dans le métabolisme de la dopamine sont sur-exprimés chez *D. simulans*

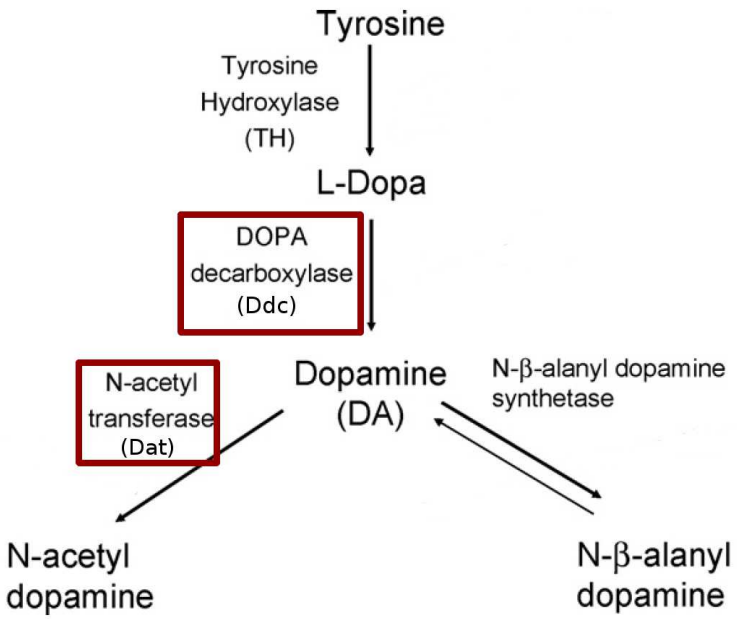


Figure 8 *– Métabolisme de la dopamine. Sont encadrées en rouge les deux enzymes dont les gènes sont moins exprimés chez* D. sechellia *que chez* D. simulans. *La Ddc a également un rôle dans la dernière étape de synthèse de la sérotonine. Figure modifiée d'après Wicker-Thomas et Hamann (2008).*

par rapport à *D. sechellia*. Le gène de la Dopamine N-acétyl-tranfèrase (*Dat*) est impliqué dans la transformation de la dopamine en N-acétyldopamine (figure 8, Wicker-Thomas et Hamann (2008)). Le gène de la DOPA-décarboxylase (*Ddc*) transforme la L-DOPA en dopamine. Cependant, cette enzyme catalyse également la dernière étape de la réaction menant à la sérotonine. L'observation de deux gènes liés au métabolisme de la dopamine suggère un rôle de celle-ci dans la différenciation entre les espèces, que ce soit en amont ou en aval de la spéciation. Cependant, la différence d'expression de ces deux gènes peut être indépendante et liée d'un côté à la production de sérotonine, et de l'autre à celle de la N-acétyldopamine. Chez la drosophile, la dopamine affecte la durée de vie et la fertilité, la pigmentation, la mobilité, la réponse à un certain nombre de drogues (notamment cocaïne, nicotine et alcool), l'apprentissage et la mémoire (Wicker-Thomas et Hamann, 2008). Elle joue également un rôle important dans la reproduction, rôle bien décrit chez la femelle : régulation de la synthèse des phéromones, et donc de la stimulation de la cour du mâle, etc. Chez les mâles, la dopamine a un rôle dans la cour (cour mâles - mâles immatures lors de la diminution de la dopamine) (Neckameyer, 1998; Liu *et al.*, 2009). En résumé, si la dopamine est impliquée dans de nombreux mécanismes, ils sont encore très mal décrits chez le mâle, et il est donc difficile d'interpréter ces différences observées au niveau de l'expression de gènes impliqués dans le métabolisme

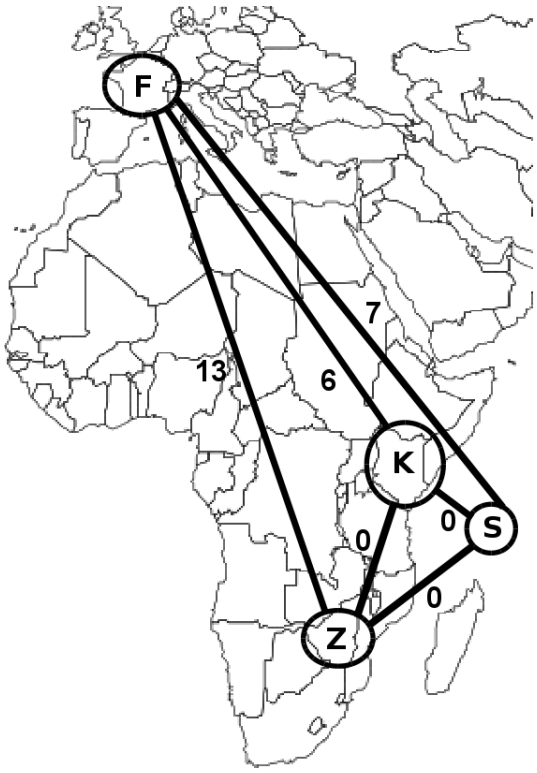


Figure 9 *– Gènes différentiellement exprimés entre les populations de* D. simulans. *Nous n'avons pas révélé de gènes différentiellement exprimés entre les populations africaines, et nous avons révélé respectivement 13, 6 et 7 gènes différentiellement exprimés entre la population de France (F) et les populations du Zimbabwe (Z), du Kenya (K) et des Seychelles (S).*

de la dopamine en terme de conséquences comportementales, d'implication dans la reproduction, dans la détoxification de toxines, dans l'apprentissage ou dans l'espérance de vie.

3.1.3 Comparaison de populations chez *D. simulans*

Faible différenciation d'expression entre populations de *D. simulans*

Nous avons examiné les différences d'expression entre nos quatre populations de *D. simulans*. Deux populations proviennent de la zone ancestrale de l'espèce, il s'agit de la population des Seychelles et de la population du Kenya. Les deux autres (Zimbabwe et France) sont des populations dérivées, d'après différentes études de génétique des populations utilisant des microsatellites (Schöfl et Schlötterer, 2004, 2006) ou des séquences codantes (Hamblin et Veuille, 1999; Veuille *et al.*, 2004; Baudry *et al.*, 2006). La figure 9 récapitule le nombre de gènes différentiellement exprimés entre populations, et la table 6 donne leur nom.

Nous n'avons pas observé de différences d'expression entre les trois populations africaines.

Tableau 6 *– Gènes différentiellement exprimés entre populations de* D. simulans

France >			**France <**		
au Kenya	aux Seychelles	au Zimbabwe	au Kenya	aux Seychelles	au Zimbabwe
Jheh1	***Jheh1***	*sm*	***CG9636***	***CG9636***	*trbd*
Cyp6w1	*Cyp6d5*	*wal*			*CG1942*
Cyp305a1	***Cyp6w1***	***slmo***			*Gprk2*
bw	***slmo***	*Edem1*			*Akap200*
Cyp12d1-p	*CG4729*	***Cyp305a1***			*CG11911*
	Cyp12d1-p	*trr*			*Aats-ala*
					CG12428

Il n'y a des gènes différentiellement exprimés qu'entre la population française et les populations africaines. Aucun gène n'est retrouvé dans les trois comparaisons, mais plusieurs sont présents dans deux (en ***gras****).*

Seules les comparaisons avec la population française ont pu révéler des différences d'expression, avec respectivement 13, 6 et 7 gènes pour les populations du Zimbabwe, du Kenya et des Seychelles. C'est peu, très peu (respectivement 0,30%, 0,16% et 0,14%). Pourquoi ? Problème de puissance de l'analyse ? C'est improbable, car en utilisant une analyse identique, nous avons révélé entre 337 et 518 gènes différentiellement exprimés entre *D. sechellia* et *D. simulans*. Il faut garder à l'esprit que ces puces ne couvraient que environ un quart du génome. On observerait plus de différences si on avait une couverture plus grande (mais pas quatre fois plus, la correction pour les tests multiples serait plus stringente). Mais nous trouverions aussi plus de différences entre *D. sechellia* et *D. simulans*.

En comparant des populations naturelles de *Fundulus* (à partir de tissu cardiaque), Oleksiak *et al.* (2002) ont révélé 37 gènes différentiellement exprimés (avec correction pour les tests multiples), sur un total de 907 gènes examinés, soit significativement plus que dans notre étude (test de χ^2 à un degré de liberté, p-value $< 10^{-16}$). Chez les humains, Storey *et al.* (2007) ont comparé l'expression de deux lignées cellulaires provenant d'individus africains (Nigéria) et d'individus européens. Ils ont montré une structuration des populations au niveau de l'expression (17% de gènes différentiellement exprimés). Cependant, leurs résultats sont toujours controversés : Davis et Kohane (2009) ont pointé des différences dans le processus d'immortalisation des lignées cellulaires utilisées dans cette étude. On peut également se demander à quel point l'expression de lignées cellulaires est le reflet d'une expression dans l'organisme, qui est elle-même fortement variable selon les tissus, les conditions environnementales, les individus,... La faible différenciation observée ici entre les populations de *D. simulans* est une des raisons pour laquelle nous avons souhaité pousser plus loin l'exploration dans une seconde étude avec une meilleure puissance, une meilleure couverture du génome et une plus grande profondeur de quantification.

Parmi les 12 gènes significativement plus exprimés en France que dans au moins une population africaine, se trouvent quatre cytochromes P450, dont trois se retrouvent dans deux comparaisons. Ces gènes ont un rôle dans la détoxification. Il est possible que l'environnement de la population française l'ait soumise à une plus forte pression pour l'efficacité des processus de détoxification. L'Europe a une agriculture intensive qui utilise plus de pesticides, et depuis plus longtemps, que l'Afrique. Cependant, nous ne disposons pas de donnée sur les polluants dans les zones de collectes, et nous ne pouvons donc que formuler cette hypothèse.

Le gène *Jheh1* est également sur-exprimé dans la population française par rapport au Kenya et aux Seychelles (les deux populations africaines proches de la zone ancestrale). La JH serait donc différentiellement régulée entre ces populations et la France. Il est possible que cette différenciation soit la base d'un changement dans le comportement de cour, et pourquoi pas d'un futur isolement reproducteur entre ces populations. Les gènes liés à la reproduction évoluent rapidement et conjointement dans les deux sexes sous l'effet de la sélection sexuelle (Singh et Kulathinal, 2000), il n'est donc pas surprenant de les retrouver dans les premiers gènes différenciés entre populations, comme entre espèces. C'est le cas également pour le gène *slowmo* (slmo), également surexprimé en France par rapport à deux populations (Seychelles et Zimbabwe). Ce gène est impliqué dans la spermatogenèse, et est nécessaire à l'établissement d'une lignée germinale (Reeve *et al.*, 2007).

Le dernier gène présent dans deux comparaisons est sous-exprimé dans la population française par rapport aux populations du Kenya et des Seychelles. Pour chacune de ces deux populations, c'est le seul gène sous-exprimé. Il s'agit de *CG9636*, dont la fonction moléculaire comme les processus biologiques dans lesquels il serait impliqué sont inconnus. Il est donc difficile d'aller plus loin dans l'interprétation de sa différenciation d'expression.

Variance intra-population : différences liées à l'histoire des populations ?

Nous avons examiné grâce à un test binomial la variance d'expression globale des quatre populations de *D. simulans* ainsi que celle de la population de *D. sechellia* (table 7). La variance de la population française est significativement plus élevée que celle de toutes les autres populations à l'exception de la population du Zimbabwe ($p < 0,005$, seuil avec la correction de Bonferroni), et ces deux populations ont une variance plus élevée que les trois autres. Toutes les autres comparaisons deux à deux ont révélé des différences significatives, avec des p-values très variables. Attention, cette observation n'est pas contradictoire avec l'homogénéité de variance gène par gène testée grâce

Tableau 7 *– P-values et direction des variations pour les comparaisons de variance globale entre populations*

		D. simulans				***D. sechellia***
		France	Zimbabwe	Seychelles	Kenya	Seychelles
D. simulans	F		$4,78.10^{-2}$	$2,94.10^{-69}$	$3,75.10^{-200}$	$9,16.10^{-128}$
	Z			$1,20.10^{-53}$	$1,81.10^{-148}$	$1,03.10^{-81}$
	S				$7,51.10^{-41}$ *	$1,83.10^{-4}$ *
	K					$1,20.10^{-25}$ *
D. sechellia	Sech					
Ordre des variances		France ≈ Zimbabwe > Seychelles > ***D. sechellia*** > Kenya				

*Au dessus de la diagonale : p-values pour la comparaison de variance globale entre les deux populations correspondantes. * significatif ($p < 0,005$, seuil avec la correction de Bonferroni). La dernière ligne montre l'ordre de classement des variances observées entre les populations.*

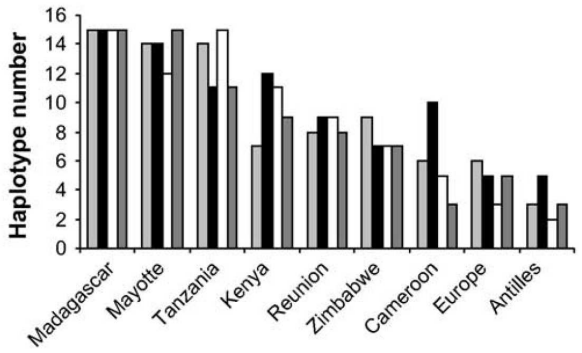


Figure 10 *– D'après Baudry* et al. *(2006). Nombre d'haplotypes en fonction de la population. Pour chaque population, les quatre couleurs correspondent dans l'ordre aux locus suivants :* Sex-lethal, vermilion, sevenless, runt. *On remarque une décroissance progressive du nombre d'haplotypes avec l'éloignement par rapport à la zone présumée ancestrale.*

au test de Levene. Il est possible d'avoir une variance homogène gène par gène, mais montrant des variations à un niveau global.

L'observation d'une variance globale plus importante chez les populations dérivées est intéressante à plus d'un point de vue. D'abord, c'est à l'opposé du patron classiquement observé au niveau des séquences, où les effets de fondation ainsi que du régime démographique instable réduisent le polymorphisme dans les populations dérivées par rapport aux populations ancestrales. Celles-ci, présentes depuis longtemps dans la même zone, ont eu le temps d'y développer un polymorphisme plus important, lié à leur large taille de population et à leur stabilité démographique. C'est effectivement ce qui est observé dans le cas de *D. simulans* (Hamblin et Veuille, 1999; Schöfl et Schlötterer, 2004; Veuille *et al.*, 2004; Schöfl et Schlötterer, 2006; Baudry *et al.*, 2006), comme illustré pour quatre gènes particuliers par la figure 10 (tirée de Baudry *et al.* (2006)).

Le polymorphisme de séquence diminue lorsque l'on s'éloigne de la zone ancestrale, alors que nous observons ici que la variance d'expression augmente dans le même temps. Comment expliquer ce patron surprenant ? Une hypothèse est fournie par le processus de décanalisation, c'est-à-dire la révélation dans un nouvel environnement de variations cryptiques pré-existantes (Gibson et Wagner, 2000; Gibson et Dworkin, 2004). Cette hypothèse prédit que dans la zone ancestrale, il existe des contraintes phénotypiques fortes dues à la sélection stabilisante, liées à l'équilibre démographique de l'espèce et à l'équilibre mutation / sélection / dérive qui en découle à grande échelle dans le génome. Au niveau génétique, des variations cryptiques s'accumulent, mais ne sont pas exprimées dans le phénotype, à cause d'un effet "tampon" qui peut avoir des causes diverses (interactions épigénétiques, contraintes environnementales ou développementales, etc.). Lors de l'invasion d'un environnement nouveau, les pressions de sélection changent. Ces perturbations peuvent alors conduire à l'expression de tout ou partie de la variation génétique cryptique, créant une augmentation de la variance phénotypique (dans notre cas, la variance d'expression). Cette explication ne vaut que si l'invasion est récente, ce qui est le cas ici. En effet, avec le temps, le patron s'atténuera, et tendra vers ce qui est observé pour les zones ancestrales. Ce genre de phénomène a été précédemment décrit pour des phénotypes particuliers, par exemple pour le nombre de soies mesosternales chez *Zaprionus indianus*, où des phénotypes atypiques apparaissent lorsque les individus sont placés à des températures inhabituelles (Yassin *et al.*, 2007).

Si le processus de décanalisation a été observé au niveau de phénotypes particuliers, il n'a jamais été montré à un niveau global. Cette hypothèse est basée sur peu de choses ici, il faut donc être très prudent quant à sa pertinence. Nous aurions souhaité vérifier par modélisation si cette hypothèse était plausible pour expliquer ces patrons de polymorphisme de séquence et de variance d'expression à l'échelle du génome. Une telle modélisation étant un travail en soi au vu de sa complexité, elle mériterait de faire l'objet d'un travail théorique futur. Nous avons cependant mené d'autres investigations sur des données publiées afin de tester cette hypothèse. Ainsi, une étude sur différents traits phénotypiques de *D. simulans* a montré des clines morphométriques (Capy *et al.*, 1993), mais nous n'avons pas détecté de différence de variance des traits entre populations dérivées et populations de la zone ancestrale dans ces données. De même, nous avons recherché un patron similaire dans les données de transcriptomique de Hutter *et al.* (2008) (qui compare une population européenne avec une africaine chez *D. melanogaster*), sans succès. Cependant, l'expansion géographique de *D. melanogaster* est beaucoup plus ancienne que celle de *D. simulans* (Lachaise *et al.*, 1988; Hey et Kliman, 1993; Kliman *et al.*, 2000; Lachaise *et al.*, 2004), ce qui a

pu permettre une diminution secondaire de la variance d'expression dans les populations dérivées. Nous n'avons donc pas pu confirmer ou infirmer d'une manière ou d'une autre cette hypothèse.

L'observation sur les variances semble être biologiquement significative, puisqu'elle est cohérente avec l'histoire invasive de l'espèce, ce qui semble éliminer l'hypothèse d'un biais technique. Ce résultat mériterait de plus amples investigations, par la recherche de données d'expression publiées sur une autre espèce à l'invasion récente, et la modélisation théorique pour déterminer la pertinence de cette hypothèse.

3.2 Utilisation des nouvelles techniques de séquençage pour étudier l'expression chez des populations de *D. simulans*

Dans cette seconde partie, nous avons utilisé les nouvelles technologies de séquençage haut débit. Après les faibles différences d'expression observées en puces, nous cherchions à explorer les différenciations d'expression des populations par une technique plus sensible que nos puces. En outre, les nouvelles technologies de séquençage nous ont permis de nous libérer de tout biais lié à l'hybridation sur des puces portant des sondes d'une autre espèce / population, de procéder sans *a priori* sur les transcrits que nous pouvions obtenir, et enfin d'augmenter nettement la profondeur de quantification. Nous avons choisi de nous limiter cette fois-ci à une population de la zone ancestrale (Mayotte), et une population dérivée, que nous avons prélevée dans une zone agricole de la vallée du Rhône en France métropolitaine (Gotheron). Chaque population est représentée par 100 individus indépendants, c'est-à-dire issus de 100 femelles différentes de la nature. Cette seconde partie de la thèse s'est inscrite dans le cadre d'une étude plus large sur l'adaptation des insectes aux changements du climat et de ressources (programme Adaptanthrop de l'Agence National pour la Recherche, concernant des insectes d'intérêt agronomique et/ou médical, ainsi que la drosophile comme insecte modèle). Dans ce cadre, nous avons donc décidé d'étudier l'expression des insectes à la fois sur leur ressource naturelle et sur une ressource de laboratoire standardisée, en l'occurence le milieu axénique.

Stratégie de double cartographie : pourquoi, comment ?

Nous avons choisi de cartographier les transcrits séquencés dans un premier temps sur le génome de *D. melanogaster*, puis pour ceux pour lesquels cela n'avait pas abouti à un résultat, sur le génome de *D. simulans*. Pourquoi cette stratégie ? Le génome de *D. melanogaster* est beaucoup mieux

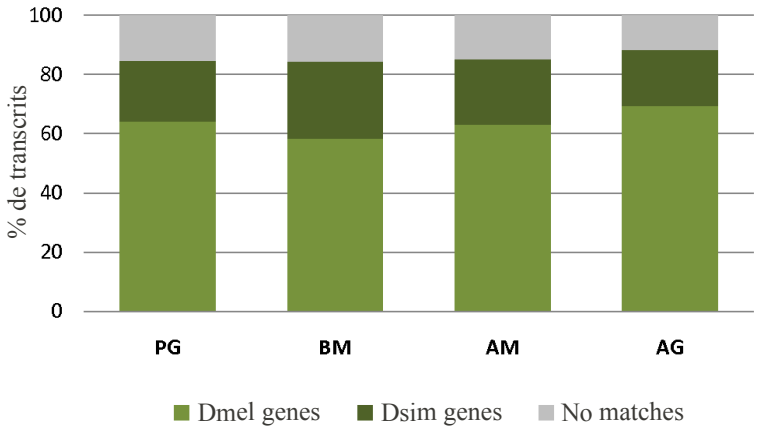


Figure 11 *– Pourcentage de transcrits cartographiés sur* D. melanogaster, *sur* D. simulans, *et sur aucun des deux génomes pour nos quatre conditions. On remarque que cela est relativement variable : entre 58 et 70% de transcrits cartographiés sur* D. melanogaster, *entre 19 et 26% sur* D. simulans, *entre 11 et 16% de transcrits non cartographiés.*

séquencé, annoté et assemblé que le génome de *D. simulans*. La cartographie sur ce génome en priorité permet donc d'obtenir des informations de fonction beaucoup plus nombreuses et fiables qu'avec le génome de *D. simulans*. Cependant, certains gènes trop divergents sont difficiles à cartographier sur le génome de *D. melanogaster*, et c'est pourquoi nous avons choisi d'effectuer la deuxième cartographie sur le génome de *D. simulans*. Le nombre exact de gènes pour lesquels nous avons obtenu des données de transcription est impossible à déterminer, à cause de cette stratégie de double cartographie des transcrits, d'abord sur le génome de *D. melanogaster*, puis sur celui de *D. simulans*. Cependant, les transcrits ont été assignés à 15359 identifiants de gènes différents (après correction, voir plus bas). Parmi ceux-ci, il y a une certaine redondance due à la double cartographie, ainsi qu'une redondance due à des problèmes d'annotations sur Flybase. La figure 11 montre pour chaque échantillon la proportion de séquences qui ont mappé sur *D. melanogaster*, sur *D. simulans*, et sur aucune des deux espèces. Cette stratégie de double cartographie a posé plusieurs problèmes.

Le premier problème concerne les gènes dont certains variants alléliques ont pu être cartographiés sur le génome de *D. melanogaster*, et d'autres, plus divergents par rapport au génome de référence n'ont été cartographiés que sur le gènome de *D. simulans*. Nous avons donc dans

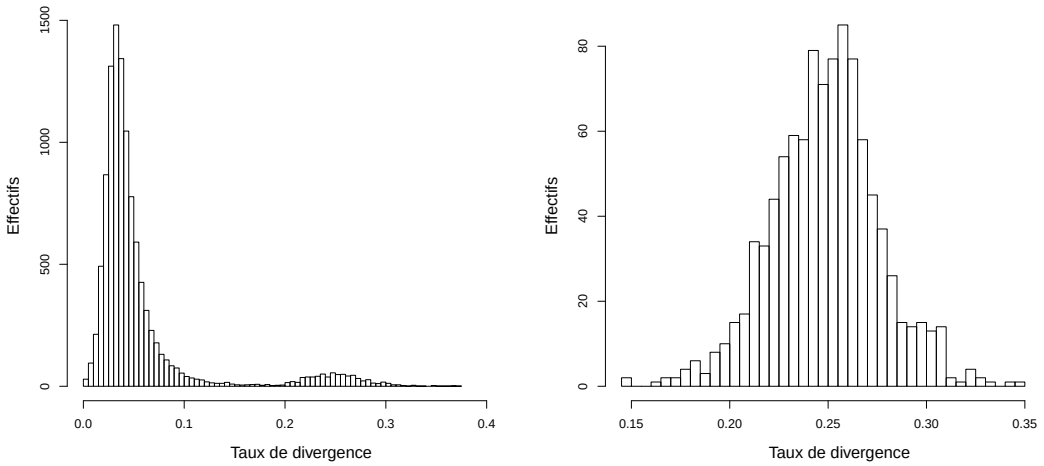


Figure 12 – A gauche : *Distribution du taux de divergence des gènes annotés comme orthologues entre* D. melanogaster *et* D. simulans *sur Flybase (Tweedie* et al.*, 2009).* ***A droite :*** *distribution du taux de divergence pour des gènes aléatoirement associés de* D. melanogaster *et* D. simulans. *On remarque que la distribution du taux de divergence des gènes orthologues est bimodale, le deuxième mode correspondant à la distribution de divergence de gènes associés aléatoirement.*

nos données deux entrées correspondant au même gène. Pour corriger cela, nous avons utilisé les annotations d'orthologie fournie par Flybase (Tweedie *et al.*, 2009). Cependant, en observant ces données, nous nous sommes rapidement aperçus que les annotations d'orthologie étaient relativement approximatives. Nous avons donc procédé à une vérification des données d'orthologie, en alignant systématiquement les gènes annotés orthologues entre les deux espèces, et en calculant leur taux de divergence (scripts *ad hoc*, langages perl et R). Nous avons procédé de même avec un set de 990 gènes de *D. melanogaster* et *D. simulans*, associés aléatoirement. La figure 12 montre la distribution du taux de divergence entre gènes orthologues avant correction (à gauche) et entre gènes associés aléatoirement (à droite). On remarque que la distribution du taux de divergence pour les gènes annotés comme orthologue est bimodale. La partie gauche de la distribution correspond à de vrais orthologues. Le second mode est à environ 25% de divergence, ce qui est aussi le mode approximatif de la distribution pour les gènes aléatoires. Il s'agit donc probablement de gènes à l'annotation d'orthologie erronée, bien qu'il y ait probablement aussi parmi ceux-ci quelques gènes orthologues très divergents entre les deux espèces. Nous avons donc souhaité vérifier l'orthologie des gènes les plus divergents. Nous avons pris un seuil arbitraire pour désigner les gènes à vérifier, en l'occurence, les gènes avec une divergence supérieure à 21%, ce qui correspond à 93% de la distribution de la divergence pour les gènes associés aléatoirement. Pour ces gènes, nous avons vérifié l'orthologie par la procédure de meilleur BLAST (Altschul *et al.*, 1990) réciproque, et obtenu

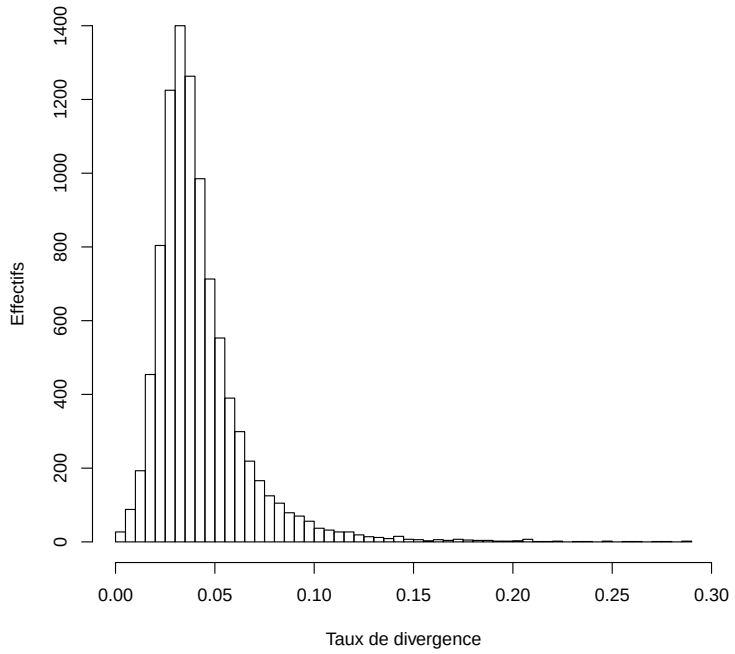


Figure 13 – *Distribution du taux de divergence des gènes orthologues après correction. On observe une disparition de la bimodalité.*

ainsi un fichier d'orthologie corrigé. La figure 13 montre la distribution du taux de divergence des gènes orthologues après correction. Il y a disparition de la seconde gaussienne. Grâce à ce fichier, nous avons ensuite rassemblé les gènes ayant été cartographiés sur les deux génomes en une seule entrée, à laquelle nous avons attribué le numéro d'identifiant du gène chez l'espèce *D. melanogaster*. Nous sommes ainsi passés de 20170 entrées à 15359 gènes. Il est probable cependant que ce problème persiste pour certains gènes, que ce soit des gènes peu connus ou trop divergents. Mais nous pensons que ce problème est, après correction, devenu marginal.

Le second problème causé par cette approche, c'est que les listes de gènes différentiellement exprimés étaient constitués d'un mélange d'identifiants de *D. melanogaster* et de *D. simulans*. Nous avons donc là aussi écrit un script *ad hoc* pour retrouver le gène de *D. melanogaster* correspondant à chaque identifiant de *D. simulans*. Toutefois, nous n'avons pas toujours pu retrouver un équivalent *D. melanogaster*, et il reste donc 14,2% de gènes avec un identifiant *D. simulans* dans nos listes de gènes différentiellement exprimés. Parmi ces gènes, il y a à la fois des gènes à l'évolution rapide pour lesquels l'assignation d'orthologie n'a pas été possible après correction, et des gènes n'ayant pas d'orthologues chez *D. melanogaster*, c'est-à-dire des gènes potentiellement spécifiques

Tableau 8 *– Nombre de gènes différentiellement exprimés entre chaque comparaison deux à deux des conditions*

Conditions impliquées	>	<
AG / AM	203	110
AM / BM	8	46
AG / PG	88	99
AG / BM	233	178
AM / PG	239	264
BM / PG	269	242
Naturel / Artificiel	3	0
G / M	66	38

AG : Axénique Gotheron. AM : Axénique Mayotte. BM : Banane Mayotte. PG : Pomme Gotheron. G : Gotheron. M :Mayotte. Comparaisons avec un seul facteur variant (trois premières lignes), avec deux facteurs variants (trois suivantes). Naturel vs. artificiel présente les gènes communs entre les comparaisons des deux populations élevées sur milieu axénique ou sur milieu naturel. G vs. M présente les gènes communs entre les quatre comparaisons impliquant d'un coté la population de Gotheron, de l'autre celle de Mayotte, et ce indépendamment du milieu.

de *D. simulans*. A la publication des 12 génomes, 80% des gènes décrits chez *D. simulans* ont été retrouvés chez *D. melanogaster*, et 67% ont été retrouvés chez l'ensemble des 12 espèces. Cela laisse donc 20% de gènes sans orthologue connu chez *D. melanogaster*, notre 14,2% de gènes sans orthologues chez *D. melanogaster* semble donc raisonnable. Ces gènes n'ont pas pu être inclus dans les analyses d'ontologies, ce qui n'est pas une perte d'information importante, puisqu'ils sont généralement sans aucune annotation liée à la fonction ou au rôle. Dans l'ensemble des analyses, nous identifions par simplicité les gènes par leur nom chez *D. melanogaster*, alors qu'il s'agit en fait d'orthologues putatifs chez *D. simulans*.

Cette cartographie en deux temps a été bénéfique à notre analyse : en effet, si nous avions cartographié uniquement sur *D. simulans*, nous n'aurions pas eu énormément d'informations de fonction, ce qui nous aurait empêchés de faire des analyses pertinentes sur les termes d'ontologie. L'assemblage du génome de *D. simulans* étant pauvre, nous aurions également eu plus de problèmes pour l'analyse des localisations géniques, bien que le fait de réaliser celles-ci sur *D. melanogaster* soit également discutable puisque l'organisation des gènes que nous observons est bien celle de *D. melanogaster* et non celle de *D. simulans*. Cette analyse se base donc sur la conservation de la synténie entre les deux espèces (Ranz *et al.*, 2007). Si nous avions simplement cartographié nos transcrits sur *D. melanogaster*, nous aurions perdu l'information supplémentaire apportée par la deuxième étape.

Nous avons donc assigné les transcrits à 15359 gènes différents (bien que ce nombre soit discutable, voire page 63). Le nombre de gènes différentiellement exprimés est extrêmement variable,

puisqu'il va de 54 pour la comparaison entre AM et BM à 313 entre AG et AM (en se limitant aux comparaisons avec un seul facteur variant, c'est-à-dire les comparaisons biologiquement significatives, et non celles où ressource et population sont différentes). De même, les gènes sur-/sous-exprimés ne sont pas en nombre égal dans chaque comparaison. Pour la comparaison de populations sur Axénique (AG vs. AM), on voit un excès de gènes sur-exprimés à Gotheron par rapport à Mayotte (table 8). En ce qui concerne les comparaisons de ressources, la tendance est à l'excès de gènes sur-exprimés sur le milieu naturel. Ces patrons s'expliquent par les grandes tendances d'induction et/ou adaptation, que nous examinerons par la suite au cas par cas.

3.2.1 Comparaison de populations : adaptation locale à l'environnement

Nous avons cherché les gènes communs aux quatre listes de gènes différentiellement exprimés pour les quatre comparaisons possibles entre d'un côté la population de Gotheron, de l'autre la population de Mayotte, et ce indépendamment du milieu d'élevage. Nous avons trouvé 66 gènes sur-exprimés à Gotheron par rapport à Mayotte, et 38 sous-exprimés à Gotheron par rapport à Mayotte. Nous avons analysé les localisations chromosomiques de ces gènes (figure 14). Ces gènes sont généralement regroupés physiquement sur les chromosomes, par groupe allant jusqu'à cinq gènes. Comment peut-on expliquer cette observation ? Il y a deux hypothèses, non exclusives :

- des gènes physiquement proches sont co-régulés
- des gènes impliqués dans des processus similaires et généralement adaptés aux mêmes phénomènes sont proches physiquement en raison de l'histoire de la famille multi-génique. Ces gènes sont issus de multiples duplications en tandem

On observe ainsi onze cytochromes P450 sur les treize sur-exprimés à Gotheron sur le bras droit du chromosome deux. Il est à noter que d'autres gènes de la même famille font parfois partie du groupe, sans être différentiellement exprimés. Ainsi, les gènes différentiellement exprimés *Cyp6a8*, *Cyp6a17*, *Cyp6a20*, *Cyp6a21* et *Cyp6a23* sont regroupés avec quatre gènes non différentiellement exprimés (*Cyp6a9*, *Cyp6a19*, *Cyp6a22* et *Cyp317a1*). On favorisera donc la seconde hypothèse, qui est cohérente avec l'histoire évolutive de la famille (Feyereisen, 2006).

Nous avons effectué des analyses de représentation de termes d'ontologie sur ces listes de gènes en utilisant FuncAssociate (Berriz *et al.*, 2003). Aucun terme n'est sur-représenté dans la liste de gènes sous-exprimés à Gotheron par rapport à Mayotte, alors que de nombreux termes sont sur-représentés dans la liste de gènes sur-exprimés à Gotheron par rapport à Mayotte (voir table 9). Une analyse détaillée des gènes correspondant à ces termes révèle essentiellement deux familles de

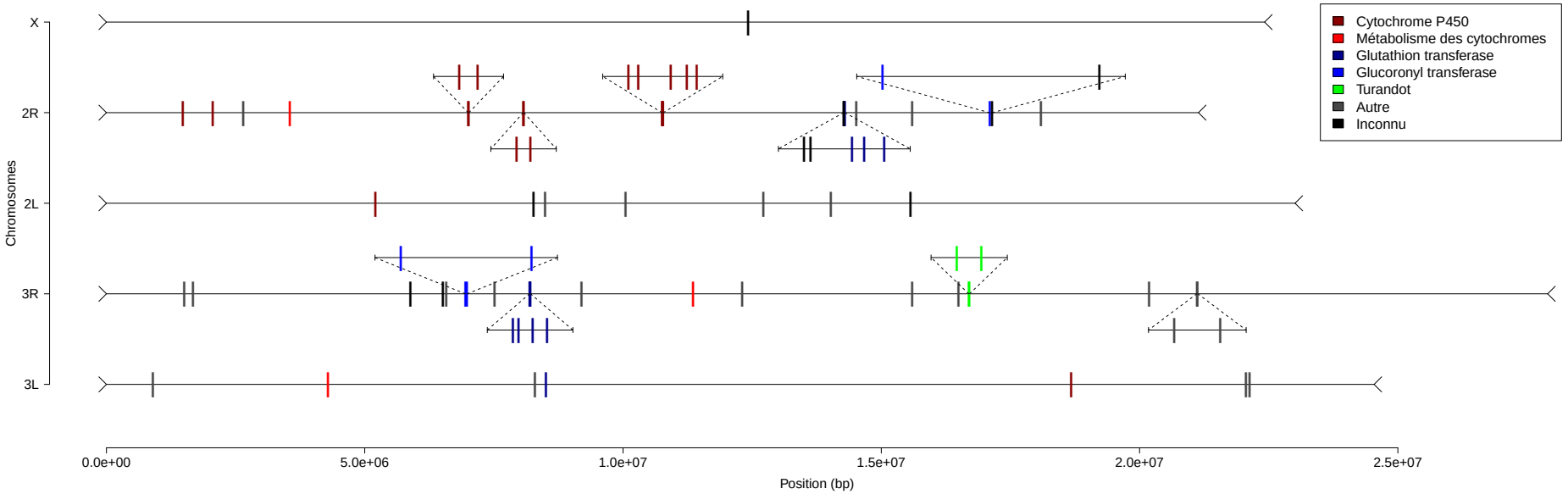


Figure 14 *– Localisation chromosomique des gènes sur-exprimés à Gotheron par rapport à Mayotte. On note des regroupements physiques des gènes différentiellement exprimés*

Tableau 9 *– Termes d'ontologie sur-représentés pour les gènes sur-exprimés à Gotheron par rapport à Mayotte*

N	X	P-value	ID GO	Terme GO
14	148	9.81×10^{-16}	GO :0009055	electron carrier activity
13	117	1.39×10^{-15}	GO :0004497	monooxygenase activity
12	138	3.84×10^{-13}	GO :0020037	heme binding
12	139	4.19×10^{-13}	GO :0046906	tetrapyrrole binding
13	199	1.45×10^{-12}	GO :0005506	iron ion binding
8	38	2.74×10^{-12}	GO :0004364	glutathione transferase activity
18	633	5.03×10^{-11}	GO :0016491	oxidoreductase activity
9	85	7.21×10^{-11}	GO :0005792	microsome
9	85	7.21×10^{-11}	GO :0042598	vesicular fraction
8	63	2.00×10^{-10}	GO :0016765	transferase activity (alkyl or aryl groups)
9	98	2.64×10^{-10}	GO :0005624	membrane fraction
9	101	3.47×10^{-10}	GO :0005626	insoluble fraction
9	102	3.80×10^{-10}	GO :0000267	cell fraction
14	474	6.85×10^{-09}	GO :0055114	oxidation reduction
35	4053	6.54×10^{-07}	GO :0003824	catalytic activity

Avec N le nombre de gènes annotés avec ce terme dans la requête ; X le nombre de gènes annotés avec ce terme dans le génome ; P-value de la siginificativité de la sur-représentation du terme dans la requête par rapport au génome, calculée avec FuncAssociate (Berriz et al.*, 2003) ; ID GO et Terme GO, respectivement l'identifiant, et le terme d'ontologie de gène correspondant.*

gènes dont plusieurs représentants sont sur-exprimés dans la population de métropole par rapport à celle de Mayotte. La première famille est celle des Cytochromes P450, et la seconde est celle des glutathion transférases. La figure 15 présente les gènes sur-exprimés à Gotheron par rapport à Mayotte, avec les ratios d'expression moyens ($\frac{AG+PG}{2}/\frac{AM+BM}{2}$) entre les deux populations. On remarque la forte présence des deux familles de gènes mentionnés précédemment. Ces deux familles ne sont pas révélées chez *D. melanogaster* lors d'études de puces comparant une population zimbabwéenne et une population des Pays-Bas, que ce soit pour les mâles (Hutter *et al.*, 2008), ou les femelles (Muller *et al.*, 2011). Cependant, le cytochrome P450 *Cyp6g1* est le seul gène sur-exprimé significativement à la fois dans ces deux études et dans la nôtre (il ne présentait pas de données exploitables dans l'étude de puces). Il peut y avoir plusieurs raisons expliquant ces différences observées entre ces deux espèces. D'abord, il est possible que les divergences de méthode soient une source de différence (fluorescence versus radioactivité, puces bicanales versus monocanales). Ensuite, il est probable que *D. melanogaster* et *D. simulans* se soient adaptées différemment à leur nouvel environnement durant l'invasion, c'est-à-dire que cette adaptation n'ait pas concerné les mêmes gènes. Il est également possible que les patrons d'expression aient été similaires après l'invasion, mais il s'est écoulé beaucoup plus de temps depuis l'invasion de *D. melanogaster* que depuis celle de *D. simulans*, les études comparent donc des états différents des espèces dus à leurs

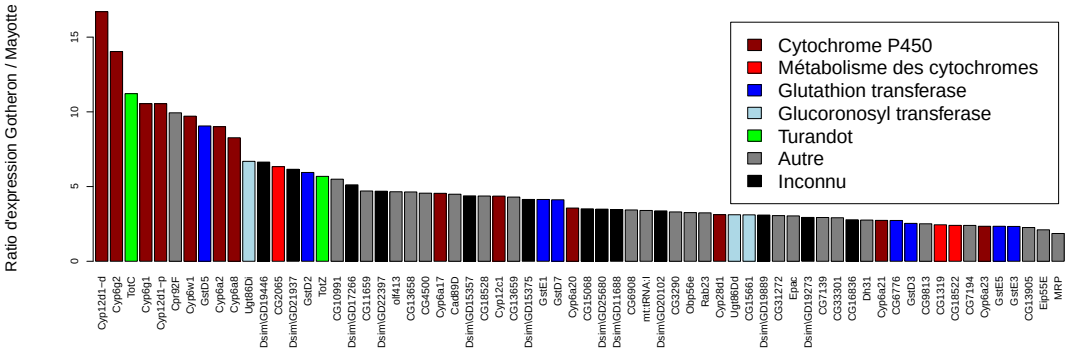


Figure 15 *– Rapports d'expression Gotheron sur Mayotte pour les gènes plus exprimés à Gotheron. Forte représentation des Cytochrome P450 et des glutathion transférases.*

histoires évolutives. Il est également possible que ce soit un effet de l'environnement dans lequel nous avons prélevé nos populations (nous avons prélevé en zone agricole, peut-être les lignées de *D. melanogaster* utilisées proviennent-elles d'une zone soumise à des pressions de sélection différentes, comme une zone urbaine par exemple). L'expression étant fortement dépendante de l'environnement, cela conduirait alors à des différences de patron d'expression. Enfin, il se peut que ces espèces soient soumises à une évolution neutre qui a fait évoluer leurs populations dans des directions différentes, simplement par les processus stochastiques de dérive.

Compte-tenu de la présence élevée parmi les gènes différentiellement exprimés de ces deux familles fortement impliquées dans les détoxifications, il n'aurait pas été surprenant de révéler également des gènes de la famille des estérases. Cette famille est fortement impliquée dans la détoxification des pesticides organiphosphorés chez le moustique *Culex pipiens*, via des séries de duplications en tandem, ainsi que des augmentations constitutives - par constitutive, nous entendons directement liée à la structure génétique, en opposition à plastique (induite) ou même à des phénomènes de régulation par l'épigénétique - de l'expression, et cela indépendamment dans plusieurs populations (pour une revue, voir Raymond *et al.* 1998). Pourtant, aucune estérase n'est différentiellement exprimée entre populations dans notre étude haut débit, pas plus d'ailleurs que dans notre comparaison de puces.

Rôle des glutathion transférases dans l'adaptation locale rapide

Parmi nos gènes sur-exprimés à Gotheron par rapport à Mayotte, se trouvent huit gènes de la famille des glutathion transférases, ainsi que trois gènes d'une famille très proche : les glucuronosyl transférases. Ces gènes ont un rôle dans la détoxification des xénobiotiques. Chez *D. melanogaster*,

la famille des glutathion transférases (GST) est composée de 36 (dont trois pseudogènes) (Wu *et al.*, 2011) à 38 gènes (Low *et al.*, 2007), et chez *D. simulans* de 40 gènes fonctionnels. Ces gènes sont divisés en classes en fonction de l'homologie de séquence et de leurs caractéristiques immunologiques (Sheehan *et al.*, 2001; Enayati *et al.*, 2005). Parmi ces classes, les δ et ε GST sont spécifiques des insectes, et elles ont subi une expansion majeure dans leur génome par duplication en tandem. *D. melanogaster* possède dans son génome neuf δ et quatorze ε GSTs fonctionnelles. Ces familles se sont fortement développées indépendamment chez *D. melanogaster* et chez le moustique *Anopheles gambiae*, ce qui suggère un rôle de ces enzymes dans l'adaptation rapide de ces espèces à un nouvel environnement. La multiplication des copies des gènes de ces familles aurait étendu la gamme des cibles qu'elles sont capables de métaboliser (Low *et al.*, 2007). Cette hypothèse est cohérente avec nos observations : notre population métropolitaine, à l'environnement plus anthropisée (zone agricole, exposée à des pesticides) montre de fortes sur-expressions de ces gènes, et notamment quatre δ GSTs, et trois ε GSTs (la dernière étant une ω GST). Cette différence d'expression semble donc liée à une adaptation locale à l'environnement, notamment ici la capacité à détoxifier des pesticides variés. Certes, ces variations pourraient avoir été causées par des processus stochastiques, mais de telles variations, à la fois dans le nombre de gènes de la famille et dans l'expression de ces gènes soutiennent fortement l'hypothèse adaptative.

Cytochromes P450

Parmi les gènes sur-exprimés à Gotheron par rapport à Mayotte, on trouve également treize cytochromes P450, ainsi que trois gènes contenant des domaines protéiques proches. Ces gènes présentent les rapports d'expression parmi les plus marqués (figure 15) : parmi les dix gènes avec les ratios d'expression les plus élevés, on trouve pas moins de sept cytochromes P450, avec une sur-expression dans la population française qui va de $\times 8$ à $\times 16$ pour ces sept gènes (pour rappel, il y a environ 85 cytochromes P450 fonctionnels chez *D. simulans*). On assiste donc dans la population de Gotheron à une très forte sur-expression constitutive de cette famille de gènes, qui a de nombreux rôles. Parmi ces rôles, il y a la détoxification des xénobiotiques, et nos cytochromes P450 sur-exprimés sont pour la plupart connus pour être spécifiquement impliqués dans ce rôle (Daborn *et al.*, 2002; Clark *et al.*, 2007; Feyereisen, 1999), tout comme les glutathion transférases. Il semble donc que l'adaptation des populations de *D. simulans* à l'environnement de la France métropolitaine soit passée essentiellement par une adaptation à l'exposition à des produits utilisés pour l'agriculture, c'est-à-dire les pesticides. Cette adaptation est constitutive, et n'a pas besoin d'être stimulée par

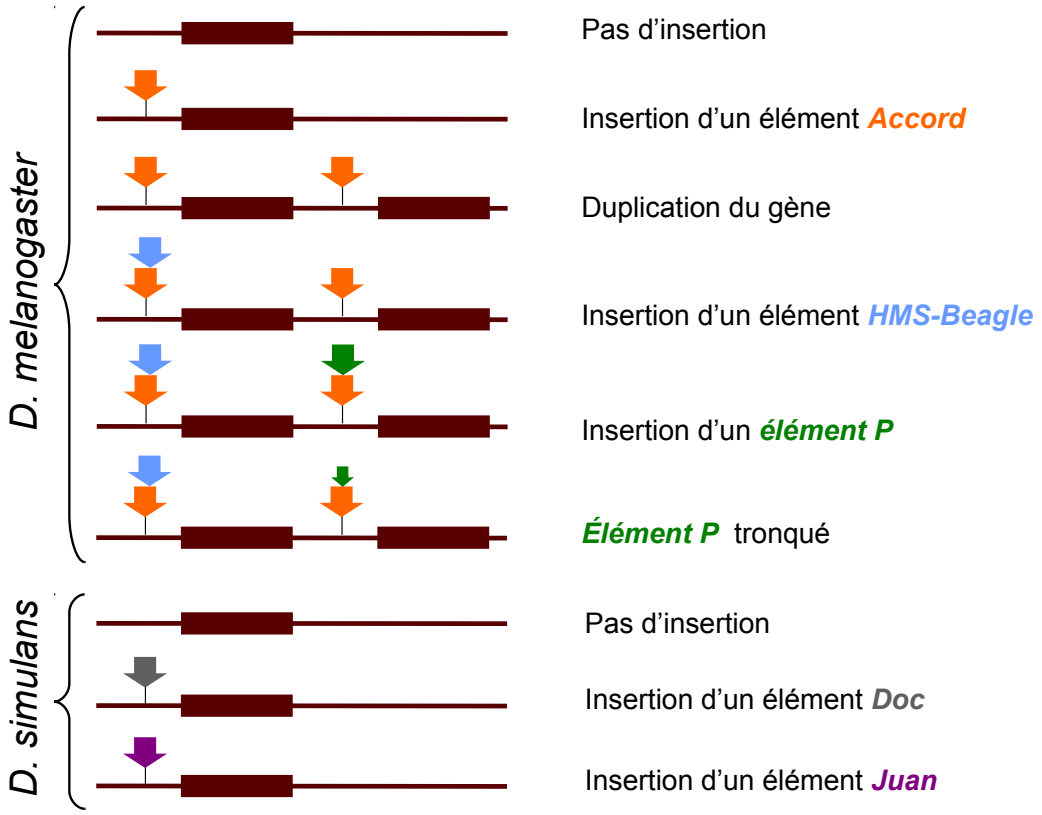


Figure 16 *– Récapitulatif des allèles de Cyp6g1 connus chez* D. melanogaster *(Schmidt* et al.*, 2010) et* D. simulans. *L'insertion de l'élément* Doc *dans une population californienne a été publiée par Schlenke et Begun (2004). Notre étude a permis de montrer que d'autres populations dérivées possèdent des insertions différentes, puisque notre population de Gotheron montre une insertion de l'élément transposable* Juan *à huit paires de bases du lieu d'insertion de* Doc. *C'est un remarquable exemple de convergence évolutive.*

l'environnement chez chaque individu, puisque même les mâles élevés sur le milieu axénique, et n'ayant donc jamais été exposés aux pesticides, montrent ces sur-expressions de gènes (sauf si ces gènes sont induits, et que cette induction est transmise à la F1 via des processus épigénétiques).

Cyp6g1, ou l'histoire d'une co-évolution

Parmi les gènes de la famille des cytochromes P450 significativement sur-exprimés à Gotheron par rapport à Mayotte, se trouve le gène *Cyp6g1*. Ce gène localisé sur le bras droit du chromosome 2 (2R) est exprimé environ dix fois plus à Gotheron qu'à Mayotte (voir figure 15). Ce gène montre également des différences d'expression entre les populations européenne et africaine de *D. melanogaster*. Il est exprimé environ 4,5 fois plus chez les mâles européens que chez les mâles africains (Hutter *et al.*, 2008), et trois fois plus en ce qui concerne les femelles (Muller *et al.*, 2011). C'est

le seul gène qui est sur-exprimé en Europe par rapport à l'Afrique dans ces deux études ainsi que dans la nôtre. Ce gène a une histoire intéressante : Daborn *et al.* (2002) ont montré que sa sur-expression était nécessaire et suffisante pour permettre à *D. melanogaster* de résister non seulement au DDT, insecticide interdit en France depuis 1972, mais aussi à d'autres insecticides (imidaclopride, nitenpyram et lufenuron). Or, dans certaines populations naturelles de *D. melanogaster*, la sur-expression est due à l'insertion d'un élément transposable *Accord* en amont de *Cyp6g1* (Daborn *et al.*, 2002; McCart et Ffrench-Constant, 2008). Des études plus poussées ont ensuite montré l'existence d'au moins six allèles différents, allant d'un gène *Cyp6g1* sans insertion (populations africaines) à un gène *Cyp6g1* dupliqué, avec de multiples insertions dans chaque copie (la figure 16 récapitule les différents allèles connus à ce jour chez *D. melanogaster* et *D. simulans*). Chez *D. melanogaster*, les six allèles sont corrélés à une expression de plus en plus forte au fur et à mesure de leur apparition, dont l'ordre a été établi par des études de lignées anciennes (Schmidt *et al.*, 2010), et à une résistance aux insecticides (testées généralement sur le DDT) de plus en plus marquées (Catania *et al.*, 2004; Chung *et al.*, 2007; McCart et Ffrench-Constant, 2008; Schmidt *et al.*, 2010). L'expression est restreinte à quelques tissus chez *D. melanogaster* : les tubes de Malpighi, l'intestin moyen, le cæcum gastrique et le corps gras (McCart et Ffrench-Constant, 2008; Chung *et al.*, 2007). Chez *D. simulans*, Schlenke et Begun (2004) ont montré une remarquable convergence évolutive en découvrant également l'insertion d'un élément transposable en amont de *Cyp6g1* dans une population naturelle californienne, élément différent de celui observé chez *D. melanogaster* (*Doc* au lieu de *Accord*). *A contrario*, leur population africaine ne montrait pas d'insertion, alors que l'allèle avec une insertion de l'élément transposable *Doc* est présent en Californie à une fréquence de 98%. Vu la sur-expression de ce gène à Gotheron, nous avons recherché l'existence d'une insertion dans nos deux populations (métropolitaine et africaine). Nous n'avons pas détecté l'insertion *Doc*, mais nous avons mis en évidence l'insertion d'un autre élément transposable, *Juan*, à huit paires de bases d'écart de l'emplacement du site d'insertion de *Doc* en Californie.

Pour estimer les fréquences allèliques de l'insertion dans nos deux populations, nous avons génotypé 47 individus issus de 47 lignées isofemelles différentes pour chaque population. Dans la population de Gotheron, nous avons trouvé 43 individus homozygotes pour l'insertion et quatre hétérozygotes, soit une fréquence de l'insertion estimée entre 90% et 99% (intervalle de confiance à 95%, estimé à partir de différentes méthodes de calculs d'intervalle de confiance sur une loi binomiale, via le logiciel R). Dans la population africaine, nous avons observé 45 individus homozygotes sans insertion, et deux individus hétérozygotes (sans insertion / *Juan*), soit une fréquence de l'in-

sertion estimée entre 0,3% et 7% (intervalle de confiance à 95%). Schlenke et Begun (2004) n'ont pas révélé d'insertion dans leur population africaine ($n = 10$). Il peut y avoir plusieurs raisons à cela :

- la faible taille de leur échantillon n'a pas permis de détecter d'insertion (pour $n = 10$, il y a 48% de risque de ne pas détecter une insertion présente à hauteur de 7%)
- la présence de cet allèle en Afrique est extrêmement récente, ou restreinte à certaines sous populations
- leur méthode recherchait uniquement l'insertion de l'élément *Doc* et ils sont donc passés à côté d'une autre insertion.

Étant donné qu'ils ont séquencé ces lignées, il est peu probable qu'ils soient passés à côté d'une insertion alternative, ce qui nous fait pencher vers une des deux premières hypothèses.

Vu la remarquable convergence évolutive entre *D. melanogaster* et *D. simulans* (mais aussi à l'intérieur des deux espèces), la corrélation avec la résistance aux insecticides (bien établie chez *D. melanogaster*, controversée chez *D. simulans*), il est clair que nous ne sommes pas en présence d'un simple point chaud d'insertion d'élément transposable, mais bien de génotypes fortement sélectionnés. Cela est confirmé par des études du polymorphisme environnant les insertions. Schlenke et Begun (2004) ont montré chez *D. simulans* une très forte baisse du polymorphisme s'étendant jusqu'à 100 kilobases autour de l'élément *Doc*, et Catania *et al.* (2004) ont montré chez *D. melanogaster* le même patron, s'étendant au minimum jusqu'à 20 kilobases autour de l'insertion. Ces baisses importantes de polymorphisme sont les marques d'un balayage sélectif autour de la mutation.

Nous avons également utilisé ces fréquences pour tester si nos populations étaient à l'équilibre de Hardy-Weinberg à ce locus. La forte sélection pourrait contraindre les populations et provoquer un écart à cet équilibre (un locus sous sélection devrait théoriquement s'écarter de l'équilibre de Hardy-Weinberg). Devant la faiblesse de certains effectifs, nous avons dû estimer les p-values du test de χ^2 par simulation. D'après ces tests, nos populations ne s'écartent pas significativement de l'équilibre de Hardy-Weinberg (p-value $= 0,33$ pour la population de Mayotte et $0,39$ pour la population de Gotheron). Pourquoi cela, vu les traces de sélection décrites dans les populations dérivées (Catania *et al.*, 2004; Schlenke et Begun, 2004) ? Si la sélection agit essentiellement sur les homozygotes sans insertion, ceux-ci étant maintenant de toute façon très rares dans la population métropolitaine, la sélection n'a plus beaucoup d'impact sur les fréquences alléliques. Cependant, la sélection ne s'est pas exercée sur les femelles une fois élevées en laboratoire, ainsi que sur leur

descendance. Il est donc logique que les fréquences alléliques estimées sur les mâles ne montrent pas d'écart à l'équilibre de Hardy-Weinberg.

On peut également s'interroger sur la faible fréquence de l'insertion dans la population africaine. En effet, autant l'avantage conféré par l'insertion dans les populations dérivées est clair, autant la rareté de cette insertion en Afrique peut être surprenante, d'autant que McCart *et al.* (2005) ont montré un certain avantage au niveau de la reproduction pour les femelles de *D. melanogaster* porteuses d'allèles de résistance. Plusieurs hypothèses peuvent expliquer ces observations. D'abord, il est possible que l'insertion soit apparue récemment dans la population de métropole (ou dans une autre population dérivée), et que sa présence à Mayotte s'explique par un apport récent par migration (les échanges aériens à Mayotte se font essentiellement avec la France métropolitaine, ou les îles proches, il ne serait donc pas étonnant dans ce cas de retrouver à Mayotte l'allèle présent en métropole).

La seconde hypothèse est celle d'une contre sélection de l'insertion dans les populations africaines, notamment chez les mâles (antagonisme sexuel). En effet, la présence d'une insertion provoque une augmentation de la valeur sélective chez les femelles, y compris en l'absence de pesticide (McCart et Ffrench-Constant, 2008). S'il y a un coût (par exemple reproductif), celui-ci est donc peut-être lié aux mâles. Smith *et al.* (2011) ont étudié des aspects variés de la fitness chez les mâles, et ont montré des effets très variables, et dépendant du fond génétique dans lequel l'allèle de résistance est introduit. Ils n'ont pas pu montrer de coût stable associé, mais seulement une forte épistasie, une forte interaction avec d'autres locus du génome. Le coût de la mutation est donc très variable, et peut parfois provoquer le maintien en faible fréquence de l'insertion, dans un environnement génomique adéquat.

Turandot, une famille de gènes qui répond aux stress

Fortement sur-exprimés à Gotheron par rapport à Mayotte, se trouvent deux gènes appartenant à la famille des *Turandot* : *TotC* et *TotZ*. Cette famille est composée de huit gènes chez *D. melanogaster*. Le premier membre de la famille décrit a été *TotA*. Ce gène code pour un facteur humoral, de réponse notamment aux chocs provoqués par de hautes températures. Contrairement aux Heat Shock Proteins (HSP), ces gènes répondent au stress après un temps de latence, via le relargage de petits peptides dans l'hémolymphe (Ekengren et Hultmark, 2001). Des données plus récentes montrent chez *D. melanogaster* que ces gènes (en l'occurence, *TotA* et *TotC*) peuvent également être activés par d'autres stimuli, comme par exemple les pesticides organochlorés (Sharma *et al.*,

2011). Ces études ont montré une réponse inductible persistante dans le temps. A contrario, nous montrons ici une sur-expression probablement constitutive de *TotC* et *TotZ* dans la population de Gotheron comparée à la population africaine. Nous avons ici le premier indice d'adaptation de l'expression dans cette famille de gènes. Cela dit, ces différences peuvent être simplement stochastiques, c'est-à-dire une conséquence liée au hasard de la divergence des populations. Cependant, la présence de deux gènes de la famille favorise l'argument adaptatif. Il serait intéressant d'observer l'induction de ce gène chez *D. simulans* afin de comparer avec ce qui est connu chez *D. melanogaster*.

3.2.2 Changements de ressources, conséquences variables

Nous avons élevé nos deux populations à la fois sur leur milieu d'origine (pomme pour Gotheron, banane pour Mayotte) et sur un milieu normalisant, standardisé : le milieu axénique. Ce milieu, bien que stérile au départ, peut être contaminé, par exemple par la flore microbienne apportée par les drosophiles elle-mêmes. L'objectif était d'observer des changements d'expression liés au changement de ressource (nouveau milieu, nouveau stress ? nouveau métabolisme ?), afin de comprendre comment les organismes s'adaptent à une modification de leur ressource naturelle. Nous avons notamment observé un décalage d'environ 24h à l'éclosion pour les drosophiles sur axénique par rapport aux drosophiles sur milieu naturel. Nous n'avons pas mesuré ce décalage précisément puisque nous nous en sommes rendu compte *a posteriori*. Il est donc difficile de savoir s'il s'agit d'une ponte retardée (rétention d'oeufs) par les femelles sur un milieu qui n'est pas familier, ou d'un retard de développement sur le milieu axénique, milieu plus pauvre, notamment en sucres.

De la banane à l'axénique : impact limité

Nous avons comparé la population de Mayotte élevée d'un côté sur axénique, et de l'autre sur banane, une des ressources naturelles de l'île. Nous avons observé 54 gènes différentiellement exprimés entre ces deux conditions (table 8). Parmi ces gènes, seuls neuf sont plus exprimés sur l'axénique, les 46 autres étant sur-exprimés sur la banane. Aucun terme d'ontologie n'est surreprésenté parmi les neuf gènes sur-exprimés sur axénique. Dans cette liste, il y a deux gènes de la famille des *Jonah* (Carlson et Hogness, 1985) : *Jon99Ci* et *Jon99Fi*. Ces gènes ont un rôle dans le développement larvaire, mais ils sont aussi exprimés dans l'intestin moyen des adultes. Ce sont des peptidases de type chymotrypsine, impliquées dans la digestion. Leur sur-expression est donc peut-être la trace d'une plasticité lors du changement de nourriture. Ces deux gènes ne sont pas

concernés par d'autres fonctions classiques des protéases, comme la défense immunitaire : ils ne sont pas induits par les infections (De Gregorio *et al.*, 2001). Les autres gènes ont des fonctions très diverses et/ou mal décrites.

De l'axénique à la banane : et si les microorganismes s'en mêlaient ?

46 gènes sont plus exprimés sur la banane que sur l'axénique. C'est significativement plus que le nombre de gènes sur-exprimés sur l'axénique (test de Fisher exact, p-value = $2,4 \times 10^{-7}$). Pourquoi ce patron ? L'axénique ne semble pas stressant pour les drosophiles. En revanche, la banane provoque une forte induction du système immunitaire. C'est pour cela qu'un plus grand nombre de gènes sont différentiellement exprimés. La table 10 montre les 26 termes d'ontologies sur-représentés dans la liste de gènes sur-exprimés sur banane par rapport à l'axénique. Tous ces termes sont liés à la défense immunitaire. Nous avons examiné dans le détail les gènes désignés par ces termes. La catégorie la plus représentée est la famille de gènes codant les Peptides Anti-Microbiens (PAM), véritables effecteurs de l'immunité. Le milieu axénique est un milieu stérile, ce qui n'est pas le cas de notre milieu "banane", rapidement envahi par des microorganismes, tant bactériens que fongiques. Il n'est donc pas surprenant que le milieu banane ait fortement induit le système immunitaire des drosophiles. Nous nous sommes donc intéressés aux caractéristiques de cette induction.

Les drosophiles ont une réponse immunitaire innée très robuste. Elles produisent de grandes quantités de PAM via l'activation de deux voies principales de régulation : la voie de signalisation Toll, dirigée contre les champignons et les bactéries Gram positive, et la voie de signalisation Immune Deficiency (IMD) - nommée de par son identification via une mutation hypomorphe (= qui affecte l'expression ou l'activité sans la supprimer totalement) - dirigée contre les bactéries Gram négative (De Gregorio *et al.*, 2001; Hultmark, 2003; Cherry et Silverman, 2006; Lemaitre et Hoffmann, 2007; Aggarwal et Silverman, 2008). Deux autres voies sont concernées par la réponse immunitaire : une voie de réponse via des ARN interférences (anti-virale), et la voie de signalisation JAK/STAT (Janus Kinase-Signal Transducers and Activators of Transcription) qui régule notamment les gènes de la famille des Turandots. Cette dernière voie est encore assez mal connue, mais il semble qu'elle ait surtout un rôle antiviral (Lemaitre et Hoffmann, 2007; Wang *et al.*, 2010).

Une vingtaine de PAM ont été identifiés à ce jour. Leur action est dirigée soit vers les bactéries Gram négative (*Diptericin*, *Attacin*, *Drosocin*, *Cecropin*, *Listericin* (Goto *et al.*, 2010)), soit vers les Gram positive (*Defensin*), soit vers les champignons (*Drosomycin* et *Metchnikowin*) (Lemaitre

Tableau 10 *– Termes d'ontologie de gènes sur-représentés parmi les gènes sur-exprimés sur banane par rapport à l'axénique*

N	X	P-value	ID GO	Terme GO
14	77	$7{,}95\times10^{-22}$	GO :0009617	response to bacterium
11	29	$3{,}00\times10^{-21}$	GO :0019731	antibacterial humoral response
16	175	$6{,}33\times10^{-20}$	GO :0006952	defense response
12	68	$1{,}38\times10^{-18}$	GO :0042742	defense response to bacterium
14	135	$3{,}21\times10^{-18}$	GO :0051707	response to other organism
14	141	$6{,}02\times10^{-18}$	GO :0009607	response to biotic stimulus
14	162	$4{,}41\times10^{-17}$	GO :0006955	immune response
14	184	$2{,}69\times10^{-16}$	GO :0002376	immune system process
11	74	$3{,}29\times10^{-16}$	GO :0019730	antimicrobial humoral response
11	89	$2{,}78\times10^{-15}$	GO :0006959	humoral immune response
14	236	$8{,}86\times10^{-15}$	GO :0051704	multi-organism process
8	32	$5{,}63\times10^{-14}$	GO :0050829	defense response to Gram-negative bacterium
16	551	$4{,}71\times10^{-12}$	GO :0006950	response to stress
14	376	$5{,}25\times10^{-12}$	GO :0005576	extracellular region
8	57	$8{,}34\times10^{-12}$	GO :0005615	extracellular space
6	27	$2{,}16\times10^{-10}$	GO :0050830	defense response to Gram-positive bacterium
8	136	$9{,}89\times10^{-09}$	GO :0044421	extracellular region part
18	1337	$4{,}19\times10^{-08}$	GO :0050896	response to stimulus
4	27	$1{,}55\times10^{-06}$	GO :0050832	defense response to fungus
4	29	$2{,}09\times10^{-06}$	GO :0009620	response to fungus
3	12	$6{,}71\times10^{-06}$	GO :0008745	N-acetylmuramoyl-L-alanine amidase activity
3	14	$1{,}11\times10^{-05}$	GO :0000270	peptidoglycan metabolic process
3	14	$1{,}11\times10^{-05}$	GO :0006027	glycosaminoglycan catabolic process
3	14	$1{,}11\times10^{-05}$	GO :0009253	peptidoglycan catabolic process
3	14	$1{,}11\times10^{-05}$	GO :0042834	peptidoglycan binding
4	54	$2{,}62\times10^{-05}$	GO :0045087	innate immune response

Avec N le nombre de gènes annotés avec ce terme dans la requête ; X le nombre de gènes annotés avec ce terme dans le génome ; P-value de la sur-représentation du terme dans la requête par rapport au génome, calculé avec FuncAssociate (Berriz et al.*, 2003) ; ID GO et Terme GO, respectivement l'identifiant, et le terme d'ontologie de gène correspondant. Tous ces termes désignent des gènes impliqués dans les processus immunitaires.*

et Hoffmann, 2007). Sur cette vingtaine de gènes, nous avons montré que 12 sont sur-exprimés chez les drosophiles élevées sur la banane par rapport à celles élevées sur axénique. Les niveaux de sur-expressions sont extrêmement variables, allant d'un rapport de 2,3 pour la *Diptericin B* à une rapport de 47 pour la *Metchnikowin*. Nos tubes d'élevage avec du milieu banane étant envahis de bactéries et champignons, cela pourrait, comme nous l'avons vu plus haut, expliquer cette stimulation. Il est à noter que la présence de microorganismes n'a pas affecté la mortalité des mâles adultes pendant les 5 jours de vieillissement utilisés dans notre protocole, cette mortalité ayant été extrêmement faible dans toutes les conditions (pas plus de un à deux mâles par groupe de 100). L'activation du système immunitaire semble donc avoir répondu efficacement aux sollicitations environnementales, au moins au stade adulte. On ne peut cependant pas exclure que la présence de microorganismes ait provoqué une mortalité larvaire : peut-être n'ont émergé que les drosophiles les plus résistantes aux microorganismes.

Les études sur le système immunitaire chez la drosophile ont reposé sur des infections bactériennes via des injections ponctuelles, ou une infection fongique via une exposition ponctuelle des insectes (les insectes sont secoués dans une boite de petri contenant les champignons). A notre connaissance, notre étude est la première à décrire une induction du système immunitaire dans le cadre d'une exposition prolongée "semi-naturelle" aux pathogènes. Si cela peut-être intéressant pour observer globalement les processus immunitaires et leur évolution lors d'une infection continue, il y a cependant un problème majeur pour pouvoir réellement comparer ces données à celle de la littérature : nous ne savons pas quels microorganismes, et à quelle concentration, ont stimulé nos drosophiles. Elles ont probablement été exposées à des cocktails divers, mélanges de champignons et de bactéries, Gram positives et négatives, soit une situation probablement semblable à ce qu'elles peuvent rencontrer dans la nature. L'induction serait donc plus naturelle dans notre cas, à la fois par son caractère continu et par sa diversité. Pour une analyse détaillée de la réponse des insectes, il aurait cependant fallu caractériser les microorganismes présents, par exemple par une étude métagénomique, ce qui n'était pas l'objectif de cette étude. Néanmoins, nous avons comparé l'expression des gènes du système immunitaire observés dans notre étude avec les patrons d'expression décrits dans la littérature.

Nous avons comparé les niveaux d'induction des PAMs ainsi que de deux gènes impliqués dans la reconnaissance des parois bactériennes (PeptidoGlycan Recognition Protein, PGRP) avec l'induction observée dans l'étude de De Gregorio *et al.* (2001), qui a fait un bilan de l'induction des gènes en fonction du type d'infection et du temps. Les PGRPs ont un rôle bien en amont

par rapport aux PAM, puisqu'ils sont impliqués dans la reconnaissance des parois bactériennes, et stimulent ensuite l'ensemble de la cascade de régulation menant à une réponse immunitaire. La figure 17 montre l'activation des 12 PAMs et des deux PGRPs, après une infection bactérienne ou fongique (d'après l'étude de De Gregorio *et al.* 2001), et dans notre étude. Attention, nous n'avons pas de réplicat dans notre étude, et nous ne pouvons donc pas estimer d'erreur sur les ratios d'induction ; nos valeurs sont donc forcément approximatives, puisque basées sur une seule mesure (mesure effectuée sur des groupes d'individus). Cependant, elles permettent d'avoir un ordre de grandeur de l'induction réelle. Les inductions mesurées par De Gregorio *et al.* (2001) comportent quant à elles deux à cinq réplicats. Certains de ces gènes montrent un niveau d'induction similaire à celui provoqué par une piqûre septique, notamment les *Attacins*, *PGRP-SC2* et peut-être la *Diptericin*. D'autres ont réagi comme lors d'une stimulation fongique, notamment la *Diptericin B*, la *Listericin*, la *Defensin* et *PGRP-SB1*, qui montrent une induction relativement limitée (un maximum d'environ $\times 4$). Enfin, deux autres gènes (*Drosocin*, *Metchnikowin*) sont beaucoup plus induits dans notre étude que dans celle de De Gregorio *et al.* (2001). Ces résultats contrastés sont peut-être dus aux modalités d'exposition aux microorganismes dans notre étude : à la fois diversifiées, continues, et avec des concentrations variables.

De l'axénique à la pomme : un coût pour la reproduction ?

88 gènes étaient sur-exprimés sur l'axénique par rapport à la pomme, alors que 99 gènes étaient sur-exprimés sur pomme par rapport à l'axénique. Nous avons également examiné les sur-représentations d'ontologies de gènes dans ces listes (comparaison de la population de Gotheron sur son milieu naturel et sur milieu axénique). La table 11 présente les termes d'ontologies sur-représentés pour les gènes plus exprimés chez les mouches élevées sur axénique par rapport à celles élevées sur pomme. Une analyse détaillée de ces termes montre qu'ils désignent deux groupes de gènes : des gènes liés à la reproduction, et une fois de plus, des gènes de la famille des cytochromes P450.

Le premier groupe est composé de huit gènes ayant une fonction dans la reproduction. Ce sont des inhibiteurs de Serine Protease, des protéines des glandes accessoires (ACP), des protéines du fluide séminal (SFP), une "Odorant Binding Protein" (Obp) et un gène de fonction moléculaire inconnu. A cela, s'ajoute deux autres ACPs, qui ne sont paradoxalement pas annotées comme impliquées dans la reproduction, mais qui montrent clairement le même patron d'expression que les huit autres. Ce patron est caractéristique pour ces gènes : l'expression génique est analogue

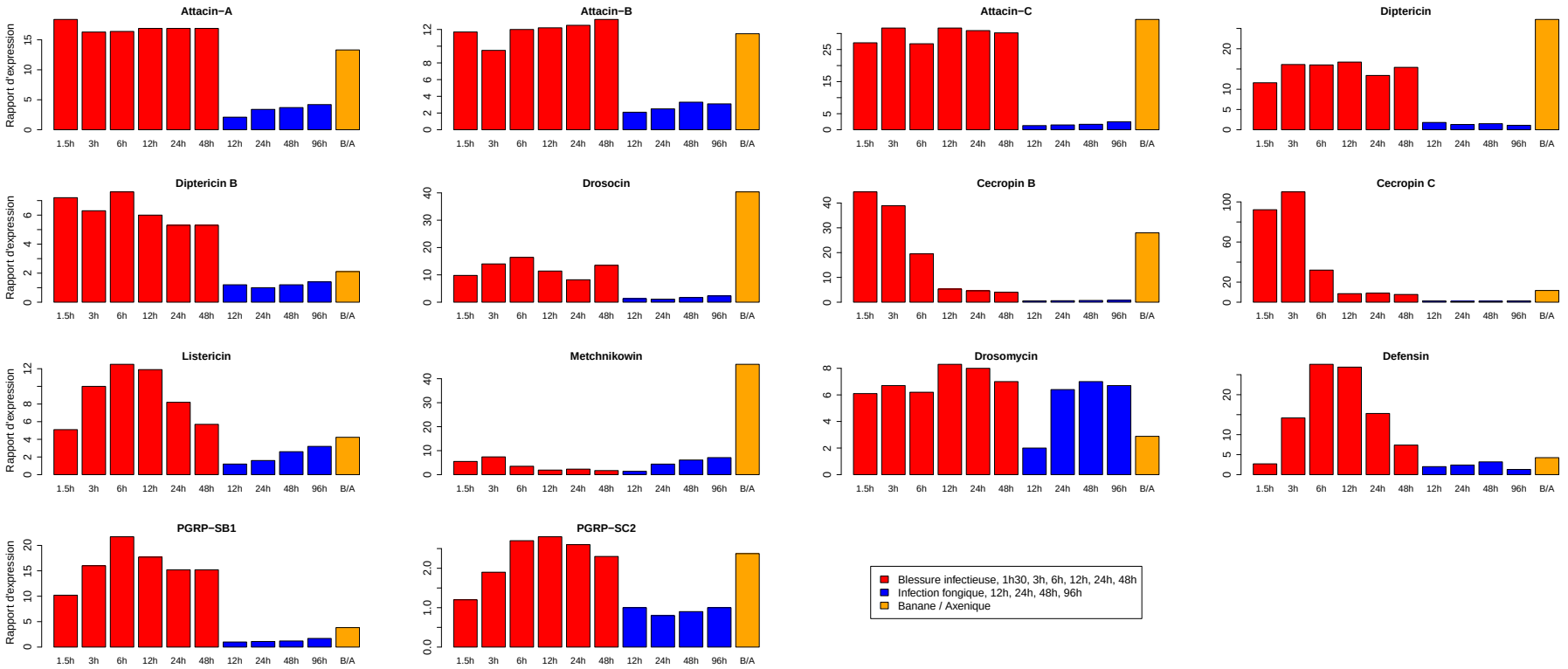


Figure 17 *– Induction des 12 peptides anti-microbiens sur banane par rapport à l'axénique, mais aussi après infection bactérienne ou fongique par rapport au contrôle (données issues de De Gregorio* et al. *2001).*

Tableau 11 *– Termes d'ontologies de gènes sur-représentés dans la liste de gènes sur-exprimés sur l'axénique par rapport à la pomme*

N	X	P-value	ID GO	Terme GO
8	75	2.65×10^{-08}	GO :0032504	multicellular organism reproduction
8	77	3.27×10^{-08}	GO :0000003	reproduction
10	144	2.90×10^{-08}	GO :0005615	extracellular space
6	85	1.81×10^{-05}	GO :0005792	microsome
6	85	1.81×10^{-05}	GO :0042598	vesicular fraction
11	223	1.91×10^{-07}	GO :0044421	extracellular region part
7	143	3.86×10^{-05}	GO :0020037	heme binding
7	144	4.03×10^{-05}	GO :0046906	tetrapyrrole binding

Avec N le nombre de gènes annotés avec ce terme dans la requête ; X le nombre de gènes annotés avec ce terme dans le génome ; P-value de la siginificativité de la sur-représentation du terme dans la requête par rapport au génome, calculé avec FuncAssociate (Berriz et al.*, 2003) ; ID GO et Terme GO, respectivement l'identifiant, et le terme d'ontologie de gène correspondant.*

pour trois conditions (Banane Mayotte, Axénique Mayotte et Axénique Gotheron), mais elle chute pour les mouches élevées sur pomme, ce qui suggère un coût reproducteur de l'élevage sur pomme. Ce coût est directement dû au milieu d'élevage, puisque la même population élevée sur axénique ne présente pas ces caractéristiques. Pourquoi ? Les raisons peuvent être diverses.

Il est possible que les pommes récoltées sur le terrain pour servir de milieu, bien que provenant d'une parcelle écologique (c'est-à-dire non traitée), aient été contaminées par des pesticides des parcelles alentours. La résistance aux pesticides est coûteuse pour les drosophiles, ce qui mènerait alors à une baisse de la reproduction. Chez les humains, l'exposition aux pesticides peut affecter la qualité du sperme (morphologie, mobilité, et concentration en spermatozoïdes) (pour des revues du sujet, voir Joly *et al.* 2008 et Tiwari *et al.* 2011). Chez *D. melanogaster*, au moins une partie des gènes d'ACPs montrent une chute de l'expression suite à une exposition aux pesticides (Gupta *et al.*, 2007). Les glandes accessoires mâles sont alors nécrosées (Gupta *et al.*, 2007; Tiwari *et al.*, 2011).

Le milieu pomme présentait par ailleurs les indices d'une fermentation (dégagement gazeux). La présence d'alcool dans le milieu peut être à l'origine du coût reproductif. Selon cette hypothèse, nous devrions trouver parmi les gènes sur-exprimés sur pomme des gènes du métabolisme de l'éthanol. Or, parmi les gènes plus exprimés sur pomme que sur axénique on ne retrouve pas les gènes les plus impliqués dans le métabolisme de l'éthanol : *Adh* notamment (mais son activité chez *D. simulans* est plus faible que chez *D. melanogaster* créant une résistance moindre à l'éthanol (David et Bocquet, 1976)), mais aussi *Aldh* et *AcCoAS* (Chakir *et al.*, 1993), ou encore *slowpoke* (Cowmeadow *et al.*, 2006). Cependant, deux autres gènes sont sur-exprimés qui soutiennent cette hyptohèse : *astray*, un

gène de réponse à l'éthanol, et *CG1600*, qui est annoté comme un gène de réponse à l'hypoxie (et qui possède un domaine homologue à l'*Adh*), celle-ci pouvant être causée par l'abondance d'alcool dans l'air contenu par le tube (même si les bouchons ne sont pas hermétiques, la circulation d'air est probablement relativement limitée). D'autres gènes différentiellement exprimés dans notre étude ont été signalés par de précédents travaux comme impliqués dans la résistance ou la sensibilité à l'éthanol : *Cyp6a20*, *CG6830*, *CG9459*, *CG16926* (sous-exprimés sur pomme) et *Thor*, et *Ect4* (sur-exprimés sur pomme) (Morozova *et al.*, 2011). Cependant, aucune étude ne s'est intéressée à l'effet de la fermentation sur la reproduction chez la drosophile.

Ce coût pour la reproduction peut aussi être causé par d'autres phénomènes physiologiques, liés à des différences du milieu : pH, consistance, concentration en sucre et plus généralement en nutriments divers, etc.

De l'axénique à la pomme encore : gènes liés au vol ?

Parmi les gènes plus exprimés sur la pomme que sur l'axénique, on trouve également un florilège de gènes plus ou moins impliqués dans le vol et la formation des muscles des ailes (*Actin 88F*, *Strn-Mlck*, *fln*, *ade2*, *Actin 42A*, *kel*, *Alg-2* et le pléiotrope *Thor*). Ce patron n'est pas révélé par l'analyse des termes d'ontologie, et seule une analyse détaillée des gènes présents nous a permis de le mettre en évidence. Il faut prendre ces considérations avec précaution, car autant le rôle de certains de ces gènes dans la fonction musculaire est clair, autant pour d'autres, il s'agit de gènes fortement pléiotropes, il est donc délicat d'attribuer leur sur-expression à une fonction particulière. D'autres enfin ont simplement un allèle annoté avec le mot "wing" dans Flybase (Tweedie *et al.*, 2009), sans plus de détails. Cependant, le nombre de gènes potentiellement reliés au vol nous a poussés à le mentionner ici. Hutter *et al.* (2008) ont montré chez *D. melanogaster* une sur-expression de gènes liés à la musculature du vol chez une population africaine par rapport à une population française. Ils ont émis l'hypothèse que cela est dû à la différence de taille des ailes par rapport à la taille du corps observée entre les populations. Chez *D. teissieri* plus les populations vivent près de l'équateur, plus leur ailes sont petites par rapport à la taille de leur corps (Paillette *et al.*, 1997). Une sur-expression de ces gènes pourrait donc être associée à une fréquence plus rapide des battements (Hutter *et al.*, 2008). Ici, nous n'avons pas de données sur la taille des mouches ou des ailes en fonction de la ressource (des données montrent cependant l'absence de cline latitudinal pour ce caractère chez *D. simulans*, Capy *et al.* 1993). Il est donc difficile de conclure sur cette hypothèse.

IV Discussion générale

Cette étude s'est penchée avant tout sur l'évolution des populations naturelles. Comment ces populations s'adaptent-elles à leur environnement ? Comment vivent-elles les interactions avec d'autres organismes, dans des éco-systèmes complexes ? Quels processus sont impliqués dans l'évolution des populations et des espèces ? Quel est le rôle de l'expression dans cette évolution ? Le sous-groupe *melanogaster* est étudié depuis longtemps dans notre laboratoire d'un point de vue biogéographique, écologique et historique, ainsi que du point de vue de la génétique des populations. Mais ces approches globales sont nouvelles au laboratoire, et ouvrent la voie à de nouvelles études écologiques et génétiques.

Notre étude de puces nous a permis de caractériser la divergence d'expression entre espèces et entre hybrides inter-spécifiques, afin d'examiner à la fois le rôle de l'expression dans la différenciation des espèces, et des perturbations de l'expression chez les hybrides qui pourraient être la manifestation d'incompatibilités dans le génome hybride. Elle nous a permis également de comparer l'expression entre populations de *D. simulans*, et l'observation assez surprenante d'une différenciation très faible entre populations nous a poussé à obtenir sur le sujet des résultats plus approfondis. Nous avons utilisé pour cela le séquençage haut débit qui permet une forte couverture des transcrits. Nous nous sommes limités à une population de la zone ancestrale et une population dérivée, afin de mettre en évidence les processus mettant en œuvre l'adaptation locale des populations lors de la colonisation d'un nouvel environnement. Nous discuterons dans un premier temps la différenciation d'expression entre populations de *D. simulans*. Nous reviendrons ensuite sur la famille des cytochromes P450, et son rôle majeur dans l'adaptation à l'environnement. Puis nous discuterons les modes d'évolution de l'expression : sélection versus neutralité. Enfin, nous évoquerons les différentes techniques d'étude de l'expression génique et nous tirerons les principaux enseignements de cette thèse.

4.1 Différenciation d'expression entre populations de *D. simulans*

4.1.1 Cohérence et divergences entre nos deux études

Nous avons révélé 104 gènes différentiellement exprimés entre populations française et africaine via l'étude de haut débit, et seulement une vingtaine via l'étude de puces (et encore, en prenant tous les gènes différentiellement exprimés entre la population française et au moins une population africaine). Pourquoi ces différences entre nos deux études, qui ont toutes deux utilisées de l'ARNm issu de mâles, extrait du corps entier ? L'âge des mâles était légèrement différent : ils avaient sept jours dans l'étude par puces, cinq dans l'étude haut débit. Mais il n'y a pas de raison que cela affecte les différences entre populations. Cependant, notre étude via les puces n'a pu examiner qu'une partie du génome (environ un quart) et le nombre de gènes différentiellement exprimés semble beaucoup plus proche une fois cet élément pris en compte. Parmi les gènes différentiellement exprimés entre les populations dans l'étude de séquençage, seulement 19,3% étaient représentés sur la puce. L'étude de puces révèle tout de même significativement moins de gènes différentiellement exprimés dans les comparaisons de population (Test de Fisher exact, p-value $= 2,2 \times 10^{-16}$). De plus, le taux de faux positifs était fixé à 0,1 dans l'étude de puces, contre 0,05 dans l'étude de séquençage. Qu'est-ce qui peut expliquer ces différences ? Les hypothèses sont multiples :

- la quantification par puces a une puissance inférieure à celle par séquençage (Marioni *et al.*, 2008; Fu *et al.*, 2009), d'autant plus avec l'augmentation rapide du nombre de lecture sur les séquenceurs haut débit
- la méthode statistique utilisée pour l'analyse des données de séquençage a sous-estimé la variance biologique, ce qui créerait donc des faux positifs. Cependant, nous avons utilisé une approche conservative en estimant la variance biologique sur l'ensemble des conditions, cette hypothèse est donc peu vraisemblable
- la conservation de lignées dans le laboratoire pendant plusieurs générations dans le cas des puces a réduit les différences entre les populations de *D. simulans* étudiés par puces. Cependant, une étude sur *D. melanogaster* utilisant des lignées conservées plusieurs années en laboratoire a révélé, via un plan d'expérience puissant, un nombre bien plus grand de gènes différentiellement exprimés entre une population africaine et une dérivée (Meiklejohn *et al.*, 2003; Hutter *et al.*, 2008; Muller *et al.*, 2011). Peut-être l'invasion plus récente de *D. simulans*

n'a pas permis de différenciation stable de l'expression, ainsi, ces populations placées dans le même environnement montreraient moins de différences que chez *D. melanogaster*, pour laquelle la séparation est plus ancienne

Aucune de ces trois hypothèses n'est réellement convaincante, cependant, nous n'avons pas d'autre explication à ces observations. Il se peut également que plusieurs hypothèses soient impliquées. La deuxième hypothèse étant peu crédible, il s'agirait alors d'un impact à la fois de la puissance, et de la normalisation par le milieu sur des populations récemment différenciées.

Seuls deux gènes étaient différentiellement exprimés entre populations à la fois dans notre étude de puces et dans notre étude par séquençage. Il s'agit de *Cyp6w1* et de *Cyp12d1-p*. Ces deux cytochromes P450 sont impliqués dans l'adaptation locale, et nous analyserons plus avant les connaissances sur ces gènes dans la partie sur les cytochromes (page 88).

4.1.2 Expression et adaptation locale

Nos études montrent essentiellement une adaptation locale via des processus de détoxification, et/ou peut-être de régulation hormonale. Cela est-il spécifique de l'organisme étudié ? Peut-on retrouver ces patrons chez d'autres organismes ?

Nous avons déjà vu que la sur-expression de gènes de détoxification est un patron également révélé chez *D. melanogaster* (Hutter *et al.*, 2008; Muller *et al.*, 2011), cependant, les auteurs notent d'autres phénomènes physiologiques impliqués dans la différenciation entre populations. Ainsi, ils ont observé une sur-expression en Europe de gènes impliqués dans le métabolisme des acides gras ; un gène sur-exprimé à Gotheron est potentiellement impliqué dans ce métabolisme (*CG4500*). Le corps gras est un organe qui a un rôle majeur dans la détoxification en général, ce qui explique ce patron observé par ailleurs dans des comparaisons de l'expression entre lignées sensibles et résistantes au DDT (Pedra *et al.*, 2004).

Les auteurs ont également observé la sur-expression en Afrique de gènes impliqués dans la formation des muscles. Ce patron est potentiellement dû à la plus petite taille des ailes par rapport à la taille du corps observée en Afrique, qui serait alors compensée par une fréquence de battement des ailes plus élevée, permise par la sur-expression des gènes liés à la formation des muscles (Hutter *et al.*, 2008). Chez *D. simulans*, ces différences latitudinales sur les ailes ne sont pas observées (Capy *et al.*, 1993).

Des différences au niveau des processus de détoxification ont également été révélées par des études d'expression dans d'autres espèces, comme la souris (Rottscheidt et Harr, 2007), ou encore

l'huitre (Chapman *et al.*, 2011). Cette dernière étude a révélé l'induction de glutathion transférases suite à l'exposition à des ions métalliques. En comparant trois sous-espèces de *Mus musculus*, des différences pour des gènes impliqués dans la détoxification ont également été montrées, au niveau du foie, mais aussi, de façon plus surprenante, et pour l'instant toujours inexpliquée, au niveau testiculaire (Rottscheidt et Harr, 2007). Ces patrons sont assez peu cohérents avec ceux observés dans une autre étude sur la souris (Voolstra *et al.*, 2007). Dans cette étude, les auteurs ont observé des patrons fonctionnels très divers, parmi lesquels on retrouve des termes d'ontologie suggérant une variation dans les processus de régulation de la quantité de protéines.

Les variations dans le métabolisme protéique sont observées également dans deux études chez la levure (Cavalieri *et al.*, 2000; Townsend *et al.*, 2003). Les auteurs ont étudié la divergence d'expression entre quatre isolats de levures prélevées dans la nature (dans un espace vinicole), comme nous l'avons déjà évoqué dans l'introduction. Leurs résultats suggèrent que ces différences au niveau du métabolisme, et dans la régulation protéique, puissent être dues à de la sélection sur ces isolats utilisés en œnologie. Chez l'huitre Chapman *et al.* (2011) ont également observé des modifications du métabolisme des protéines, et ont relié ce phénomène à l'adaptation locale à des facteurs environnementaux. Peut-être est-ce également la raison de ce phénomène, chez la levure comme chez la souris (Townsend *et al.*, 2003; Voolstra *et al.*, 2007).

Si l'adaptation à l'environnement passe souvent par l'adaptation aux polluants rencontrés, le climat, et notamment la température, est un autre paramètre environnemental majeur auquel les espèces sont exposées. Une étude sur le saumon a montré l'adaptation de gènes impliqués dans le repliement protéique, en réaction à la température (Evans *et al.*, 2011). Ils ont également révélé l'induction de gènes connus pour réagir au stress thermique : les heat shock protein (HSP). Comme nous l'avons évoqué dans l'introduction, Oleksiak *et al.* (2002) ont montré que des populations sympatriques des espèces proches *F. grandis* et *F. heteroclitus* divergeaient moins en expression que des populations de *F. heteroclitus* vivant dans deux environnements différents. Ils ont révélé des différences pour des gènes impliqués dans le transport du cholestérol, ainsi que dans d'autres aspects du métabolisme. Ils ont proposé que ces différences étaient dues à une adaptation, notamment à la température de l'eau. Ces observations ne sont pas en accord avec les modèles d'évolution neutre de l'expression génique (voir page 95), puisqu'alors, les changements d'expression suivraient globalement la phylogénie des taxons. D'autres études, toujours dans le genre *Fundulus*, ont montré des différences au niveau des gènes de détoxification, et notamment des cytochromes (Whitehead *et al.*, 2010).

Tableau 12 *– Résumé des fonctions associées aux quatre clades de cytochromes P450 chez* D. melanogaster

Clade	# gènes	Fonctions associées
CYP2	6	Régulation hormonale (notamment ecdysone)
CYP6[1]	36	Détoxification des xénobiotiques ; résistance aux pesticides
CYP4	32	Inductible par les xénobiotiques ; Olfaction, phéromones
CYPM[2]	11	Résistance aux insecticides ; Fonctions physiologiques majeures

[1] *le clade CYP6 est issu des clades CYP3 et CYP5 chez les vertébrés.* [2] *CYPM est le clade de cytochromes mitochondriaux. Les autres clades sont microsomaux. Les données de cette table proviennent de la revue de Feyereisen (2006).*

Globalement, ces études montrent des patrons très divers, et dépendant de l'écologie des espèces. Les variations que l'on retrouve dans plusieurs études concernent notamment le métabolisme des acides gras, de la régulation protéique, de l'adaptation à la température et de la détoxification. L'adaptation locale passe donc généralement par l'adaptation à des milieux pollués, ou à des climats différents. Il serait intéressant d'analyser en parallèle chez différents organismes des cas d'invasion d'environnements plus ou moins anthropisés, afin de déterminer à quel point la réponse à l'environnement est spécifique, ou au contraire, commune à différents taxons.

4.2 Les cytochromes P450 : une famille fortement liée à l'adaptation locale ?

4.2.1 Classification et rôles

La famille des cytochromes P450 est une des familles multigéniques les plus importantes en nombre de gènes comme en fonction, et ce dans l'ensemble des domaines du vivant. Ces gènes ont tous une origine commune et se sont développés dans les génomes par duplications successives (Feyereisen, 1999, 2006). Leurs rôles sont extrêmement divers, allant de la biosynthèse des ecdystéroïdes et de l'hormone juvénile, à la détoxification des xénobiotiques. C'est surtout ce dernier rôle qui les a désignés comme sujet d'étude privilégié, la métabolisation des drogues par l'être humain touchant à de nombreux sujets, qui vont de l'utilisation de drogues communes, à la métabolisation des médicaments.

Chez les insectes, cette famille comprend une centaine de gènes. On les divise en quatre principaux clades (Feyereisen, 2006). La table 12 résume le nombre de gènes de chaque clade connus chez la drosophile, ainsi que les fonctions associées.

Les cytochromes différentiellement exprimés dans notre première étude (puces) entre *D. simulans* et *D. sechellia* appartiennent essentiellement aux clades CYP4 et CYP3 (parmi ceux-ci, particulièrement le groupe des CYP6), et semblent être essentiellement impliqués dans les détoxifications de xénobiotiques / résistances aux pesticides. Cela dit, nos puces comportent un biais d'identification, puisque seulement 26 des 85 P450 y étaient représentés, dont dix CYP4, quatorze CYP3, et un seul représentant pour les clades CYP2 et CYPM. Cependant, l'analyse par haut débit à également montré une différence d'expression essentiellement limitée à ces clades, alors que pas moins de 68 P450 ont pu être analysés dans l'étude par séquençage (parmi ces 68, on trouve 22 CYP4, 34 CYP3, 3 CYP2 et 9 CYPM). Les seuls gènes différentiellement exprimés à la fois dans notre comparaison de population par séquençage et dans au moins une comparaison de population française contre une population africaine par puce, sont deux cytochromes P450 : *Cyp6w1* (qui appartient au clade CYP3)et *Cyp12d1-p* (qui appartient au clade CYPM).

4.2.2 Cytochromes, gènes induits et/ou à l'expression modifiée par l'environnement

Une étude de la réponse aux xénobiotiques via des puces chez *D. melanogaster* a montré la spécificité de réponse des cytochromes P450 (Le Goff *et al.*, 2006). Cette étude a utilisé des puces développées localement, et centrées sur les trois grandes familles de gènes de détoxification : les P450, les glutathion transférases, et les estérases. Les auteurs ont testé l'effet de deux xénobiotiques différents : le psychotrope phénobarbital, et l'herbicide atrazine (utilisés largement respectivement en médecine et agronomie). Ils ont montré que les patrons d'induction de ces gènes diffèrent à la fois selon le xénobiotique utilisé pour l'induction, mais aussi selon le sexe. Cependant, quelques gènes montrent un patron relativement constant, et sont généralement également différentiellement exprimés dans nos deux études. Seuls trois gènes sont significativement induits chez mâles et femelles par les deux xénobiotiques de Le Goff *et al.* (2006) : *Cyp6a2*, *Cyp6w1*, *Cyp12d1-d*. Nous avons montré pour ces trois gènes une expression constitutive plus forte dans la population de Gotheron qu'en Afrique, et *Cyp6w1* a aussi une sur-expression constitutive chez *D. simulans* par rapport à *D. sechellia* (étude de puces). La figure 18 montre les différences d'expression observés par Le Goff *et al.* (2006) et ici, pour ces gènes ainsi que quelques autres qui sont communs à nos deux études.

Globalement, le phénobarbital induit une réponse plus forte que l'atrazine. Si on compare les inductions observées par Le Goff *et al.* (2006) et l'expression issue de nos données, on observe

des patrons variés, qui tendent à montrer que l'adaptation de l'expression ne suit pas les patrons observés lors d'une exposition ponctuelle à des xénobiotiques. Sur les sept gènes constitutivement sur-exprimés dans notre étude, trois sont induits par l'atrazine, et tous sont induits à différents niveaux par le phénobarbital. L'adaptation locale montrée par nos populations a été contrainte par une diversité de produits inconnus. Il est possible et probable que ces produits soient différents de ceux testés dans l'étude de Le Goff *et al.* (2006). Nous avons également cherché à faire le parallèle avec les données obtenues pour ces gènes sur nos puces, malheureusement certains de ces gènes n'y sont pas représentés, d'autres ont trop de données manquantes. Un seul (*Cyp6w1*) montre des données exploitables. Il est induit avec un rapport moyen de 1,40 dans la population française par rapport aux trois populations africaines. Ceci cache des différences entre les trois populations, puisque la population française a une induction identique à celle du Zimbabwe, alors que le rapport d'expression est de respectivement 1,65 et 1,77 avec les populations des Seychelles et du Kenya. Comme nous l'avons vu, la figure 18 montre peu de cohérence entre les données d'induction de cytochromes tirées de Le Goff *et al.* (2006) et l'adaptation via une augmentation de l'expression constitutive que nous avons observée. 27 cytochromes, soit environ un tiers des P450 de la drosophile sont inductibles par les xénobiotiques, alors que 12 ont montré une expression constitutive plus haute dans les lignées qui montrent une résistance à ces substances (Giraudo *et al.*, 2010). Huit gènes sont à la fois inductibles et montrent des allèles de résistance (*Cyp6a2*, *Cyp6g1*, *Cyp12d1*, *Cyp4e2*, *Cyp4p1*, *Cyp6a17*, *Cyp6a8* et *Cyp6w1*). Parmi ces huit gènes, six montrent effectivement une sur-expression à Gotheron dans notre seconde analyse. Mais notre étude a également révélé six autres cytochromes constitutivement sur-exprimés dans la population française plus exposée au pesticides, ce qui suggère une adaptation. Il s'agit des gènes *Cyp6a20*, *Cyp12c1*, *Cyp6a21*, *Cyp28d1* (clade CYP3) et *Cyp6a23*. Ce dernier montre cependant une sur-expression chez certaines lignées de *D. melanogaster* résistantes au DDT (Pedra *et al.*, 2004). Cette dernière étude a également révélé le gène *Cyp314A1*, qui est également induit dans une lignée résistante au DDT chez le moustique (Vontas *et al.*, 2005). Les six nouveaux gènes de notre étude, peuvent ne pas avoir été révélés jusqu'à maintenant pour différentes raisons :

- pour des raisons méthodologique
- leur sur-expression est spécifique de *D. simulans* ou même de la population de Gotheron
- les différences observées ont été induites dans la nature à la génération parentale, et transmises aux descendants par des processus épigénétiques

En effet, les autres études ont essentiellement porté sur l'espèce modèle *D. melanogaster*. Notre

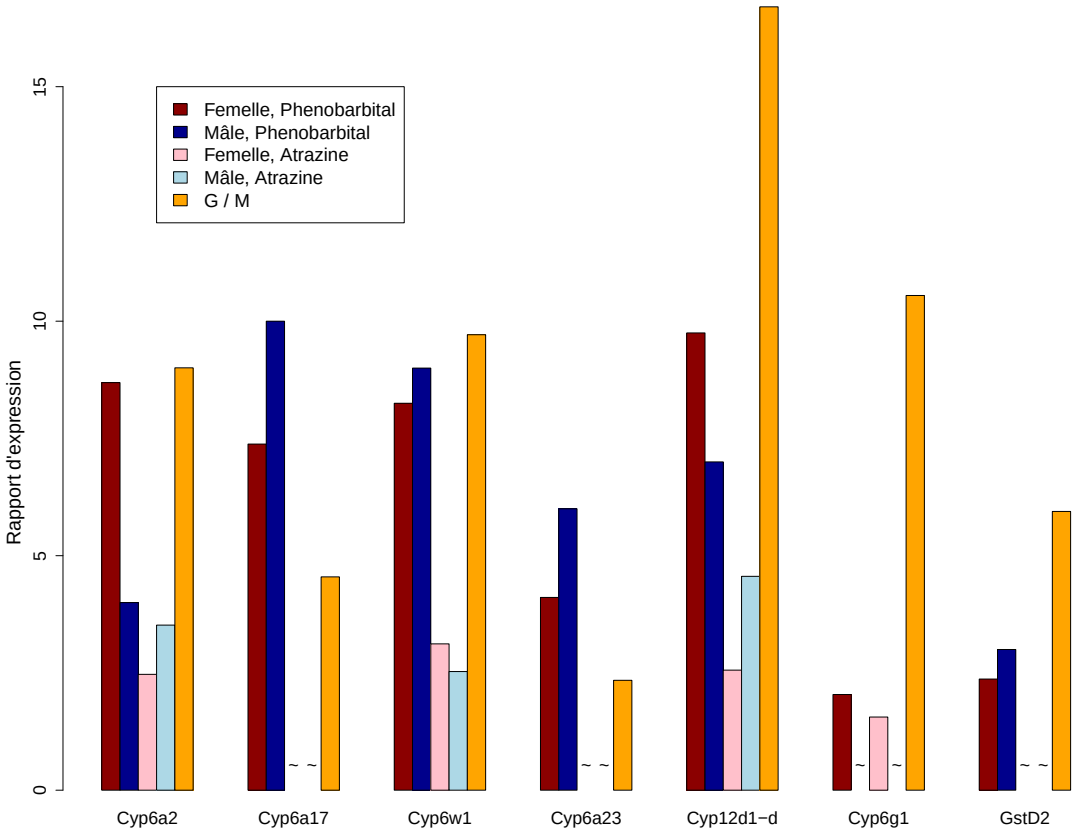


Figure 18 *– Ratio d'expression pour 7 cytochromes impliqués dans les détoxifications, pour des mâles et des femelles stimulés soit au phénobarbital, soit à l'atrazine. En orange, le ratio d'expression constitutive dans la population de Gotheron par rapport à la population de Mayotte. Les colonnes manquantes, repérées par ˜ correspondent à des inductions d'expression qui ne sont pas significatives (le ratio n'est pas significativement différent de 1). Pour chaque gène, les quatre premières données sont tirées de Le Goff* et al. *(2006).*

étude montre donc à la fois que les gènes majeurs impliqués dans l'adaptation locale à la résistance sont identiques entre *D. melanogaster* et *D. simulans*, mais qu'il y a également d'autres gènes candidats chez *D. simulans*, qui mériteraient peut-être d'être examinés chez *D. melanogaster*. Cependant, l'étude de puce de Le Goff *et al.* (2003) qui a pris en compte l'espèce *D. simulans*, n'a pas montré de grandes différences entre lignées résistantes et lignées contrôles, notamment pour les nouveaux gènes observés dans notre analyse. Cette étude a surtout confirmé le rôle prépondérant de *Cyp6g1*, et notamment la grande diversité des substrats que ce gène est capable de métaboliser. Si les nouveaux gènes candidats de notre étude peuvent correspondre à une spécificité de notre population métropolitaine, ils peuvent également être dus aux conditions plus naturelles de notre analyse, puisque les drosophiles n'ont pas subi de génération complète en laboratoire.

L'expression des cytochromes P450 a été examinée chez de nombreux organismes. Chez l'humain, plus de 200 études ont mesuré leur expression, avec des applications très diverses allant de la pharmacologie à la régulation hormonale, en passant par les impacts physiologiques de la grossesse (Rezen *et al.*, 2007). Chez d'autres organismes, des études ont montré des réactions diverses à des facteurs environnementaux et physiologiques. Ainsi, une réponse à des milieux d'élevage plus ou moins riches en azote a été montrée sur 27 cytochromes P450 chez le champignon *Phanerochaete chrysosporium*, un champignon enrichi en gènes de cytochromes P450, puisqu'il en compte environ 150 (Doddapaneni et Yadav, 2005). Comme dans notre étude, les auteurs ont ici montré que les cytochromes différentiellement exprimés étaient généralement regroupés physiquement sur les chromosomes. Les cytochromes réagissent aux polluants, qu'ils soient d'origine biotique ou abiotique. Narusaka *et al.* (2004) ont ainsi testé un vaste répertoire d'agents stressants, qu'ils soient chimiques, environnementaux (température, ultraviolets), physiologiques ou infectieux chez la plante *Arabidopsis thaliana*. Ces différents stress ont permis l'induction de 29 cytochromes P450 différents (sur 49 étudiés), certains d'entre eux réagissant à de multiples stress. Les cytochromes P450 montrent également une diversité d'expression tissulaire. Toujours chez *A. thaliana*, Xu *et al.* (2001) ont montré que des gènes dont la séquence est proche s'expriment généralement dans les mêmes tissus. Ce patron peut être dû à une co-régulation des groupes de gènes multipliés par duplications en tandem successives. L'induction des cytochromes P450 pourrait même être utilisée comme témoin de la pollution d'un environnement, comme cela a été suggéré en étudiant le nématode *Caenorhabditis elegans* (Custodia *et al.*, 2001).

4.2.3 Exemples de cytochromes montrant une adaptation locale

Cette famille est fortement impliquée dans l'adaptation locale aux xénobiotiques rencontrés par les populations. En dehors du très bel exemple de *Cyp6g1* dont nous avons déjà parlé, d'autres gènes montrent des exemples d'adaptation par sélection locale d'allèles influant sur l'expression. Waters *et al.* (1992) ont étudié l'expression et la structure du gène maintenant appelé *Cyp6a2* sur des lignées de *D. melanogaster* sélectionnées en laboratoire pour leur résistance au DDT et des lignées contrôles. Les lignées résistantes montrent une expression augmentée par rapport aux contrôles (Waters *et al.*, 1992; Brun *et al.*, 1996). Ils ont mis en évidence quelques mutations entre les deux lignées (sensible et résistante), et notamment la présence d'éléments répétés issus d'un élément transposable à l'extrémité 3' du gène *Cyp6a2* dans la souche sensible, provoquant une instabilité potentielle du transcrit. La lignée résistante montre en aval du gène l'insertion d'un fragment provenant d'un petit ARN situé ailleurs dans le génome (Waters *et al.*, 1992). La structure du locus a été confirmée sur d'autres lignées collectées dans la nature.

Le cytochrome *Cyp12a4* a été relié à la résistance au lufenuron (un pesticide), et bien que la mutation responsable n'ait pas été identifiée, c'est là aussi une sur-expression qui confère la résistance (Bogwitz *et al.*, 2005). Chen et Li (2007) ont cependant montré une insertion d'élément transposable dans la région 3' du gène dans une lignée résistante. Les cytochromes ont participé à l'adaptation à plusieurs étapes de l'histoire de l'espèce humaine, comme par exemple une réduction de la vitesse du métabolisme lors de la transition des populations de chasseurs-cueilleurs aux populations agricoles. L'adaptation s'est faites entre autre via des variants alléliques du gène *Cyp2d6*. Une autre étude a montré l'implication du cytochrome *Cyp11b2* dans l'adaptation à la haute altitude, toujours chez *Homo sapiens* (Rajput *et al.*, 2006). En comparant des populations locales avec des visiteurs, cette étude a montré la plus grande prévalence d'un allèle particulier dans les populations locales, en association avec une baisse du taux d'aldostérone (molécule importante dans le mal de l'altitude, Fuselli *et al.* 2010). Ces adaptations chez l'humain permettent de penser que le rôle des cytochromes P450 dans l'adaptation est loin de s'arrêter aux détoxifications.

4.2.4 Les éléments transposables en tant qu'éléments mutagènes liés à l'adaptation locale

L'observation de Chen et Li (2007), ainsi que celle de la présence d'un élément transposable en aval de *Cyp6g1* (dans notre étude et dans d'autres), suggère un rôle important des éléments

transposables dans la production de variants adaptatifs des gènes de cytochromes. Chen et Li (2007) ont vérifié cette hypothèse en recherchant *in silico* la présence d'éléments transposables à proximité de treize gènes de cytochromes chez *D. melanogaster* : huit avec un rôle dans la détoxification des xénobiotiques, et cinq impliqués dans la biosynthèse de l'ecdysone, ou dans le contrôle du développement. Les auteurs ont observé de multiples évènements d'insertion d'éléments transposables autour de sept cytochromes sur les huit impliqués dans les détoxifications, mais aucun autour des cinq cytochromes impliqués dans d'autres processus. Ces éléments sont généralement insérés en amont du gène, mais on peut parfois en détecter également en aval. Ils ont fait une observation similaire sur le papillon *Helicoverpa zea* (Chen et Li, 2007).

Ces observations mettent en avant le rôle majeur des éléments transposables en tant qu'éléments mutagènes de l'expression liés à une adaptation locale. L'activité des éléments transposables est généralement stimulée en cas de stress (Capy *et al.*, 2000). L'exposition à des xénobiotiques nouveaux activerait donc ces éléments, et permettrait une adaptation rapide de la régulation de gènes à impacts majeurs comme les cytochromes P450, ou encore les glutathion transférases. La sous-représentation des éléments transposables à proximité de gènes de ménage fortement contraints montre que leur rôle est plutôt au niveau d'une adaptation rapide, leur faculté à transposer ne garantissant pas la stabilité de l'expression requise pour les gènes de ménage. La réaction à un stress se ferait donc dans un premier temps via l'induction de gènes de résistance, puis par l'adaptation via des mutations régulatrices, provoquées préférentiellement par des éléments transposables, et enfin via des modifications de la séquence codante. Cette gradation de réponse s'appuie sur l'ensemble des observations décrites au-dessus, mais reste hautement spéculative. On peut le voir en terme de probabilité de réponse.

González *et al.* (2008) ont examiné les insertions d'éléments transposables référencés sur Flybase pour rechercher des cas d'insertion adaptatives. Sur les treize insertions qu'ils ont observées, au moins quatre d'entre elles provoquent un changement dans l'expression. Ce nombre d'insertions peut paraître relativement faible, mais ils ont posé des critères stringents de sélection sur les insertions. Warnefors *et al.* (2010) ont comparé l'expression entre humains et chimpanzés. Ils ont montré une divergence d'expression plus forte pour les gènes qui montraient une insertion d'élément transposable dans l'une des deux espèces. Cependant, ils ont également révélé que les gènes porteurs d'une même insertion d'élément transposable dans les deux lignées (insertion ancestrale) ont également une plus grande divergence d'expression. Ils concluent donc que l'insertion n'est pas la cause de la divergence d'expression, mais que les éléments transposables ont tendance à

s'insérer ou à être retenus dans des gènes qui par ailleurs montrent une plus grande divergence d'expression (gènes moins contraints ?). Cependant, d'autres études ont mis en évidence l'importance des éléments transposables dans l'évolution de la régulation. Avec une approche similaire à celle adoptée par Warnefors *et al.* (2010), Pereira *et al.* (2009) ont montré chez les rongeurs que les éléments transposables étaient responsables d'environ 20% de la divergence d'expression entre la souris et le rat. Enfin, Jordan *et al.* (2003) ont montré la contribution importante des éléments transposables aux séquences régulatrices chez l'humain : ≈25% des séquences régulatrices seraient issues des éléments transposables.

Tous ces exemples de domestication d'éléments transposables ne doivent pas cacher le fait que ces éléments sont avant tout des parasites génomiques, dont l'insertion est généralement délétère.

4.2.5 Bilan sur l'adaptation

L'ensemble de ces observations liées à l'adaptation à l'environnement montrent que les patrons d'adaptation de l'expression sont divers, avec quelques familles qui réagissent particulièrement, notamment celles impliquées dans la détoxification. Cela montre aussi à quel point les environnements ont été transformés par l'être humain, puisque le type d'adaptation que nous pouvons détecter maintenant, est une adapatation à notre pollution des écosystèmes. On détecte aussi des changements métaboliques, notamment au niveau des acides gras, etc. Enfin, l'enseignement que l'on peut tirer de la comparaison d'études qui ont examiné la différenciation de l'expression entre populations naturelles, c'est la diversité des patrons observés (en dehors des quelques familles susmentionnées). Cela permet de penser qu'une part non négligeable de la divergence d'expression n'est pas adaptative, mais bien gouvernée par l'évolution neutre, la dérive génétique.

4.3 Expression et évolution neutre

Selon la théorie neutraliste de l'évolution, les polymorphismes qui correspondent à des mutations soumises à sélection sont négligeables en nombre au regard des mutations qui ont une évolution neutre ou proche de la neutralité (Kimura, 1983). En dehors des adaptations claires chez certains gènes ou familles de gènes, on peut observer entre taxons de nombreux gènes différentiellement exprimés (Cavalieri *et al.*, 2000; Oleksiak *et al.*, 2002; Townsend *et al.*, 2003; Meiklejohn *et al.*, 2003; Oleksiak *et al.*, 2005; Haerty et Singh, 2006; Oshlack *et al.*, 2007; Michalak *et al.*, 2007; Rottscheidt et Harr, 2007; Voolstra *et al.*, 2007; Hutter *et al.*, 2008; Dworkin et Jones, 2009;

Muller *et al.*, 2011). Il est donc possible que ces gènes, ou certains de ces gènes aient évolué de façon neutre. On peut alors s'interroger sur la part d'évolution neutre dans les changements de l'expression génique.

Pour analyser la part d'évolution neutre dans l'expression, il faut d'abord décrire l'hypothèse nulle, c'est-à-dire comprendre comment évoluerait l'expression sous cette hypothèse de neutralité. De nombreuses études se sont penché sur ce problème (Khaitovich *et al.*, 2004, 2005; Ogasawara et Okubo, 2009). Ces modèles plus ou moins complexes sont ensuite généralement testés au regard de données réelles d'expression. D'autres études enfin, se sont bornées à tester des conséquences très simple de l'hypothèse neutre, comme la relation linéaire qui existerait alors entre phylogénie et divergence d'expression.

Khaitovich *et al.* (2004) ont analysé la corrélation entre divergence d'expression et temps de divergence entre taxons chez les primates. Ils ont montré une relation linéaire positive entre ces deux variables, ce qui est compatible avec un modèle neutre, mais n'exclut pas un modèle adaptatif. Ils ont également montré que les gènes évoluant le plus vite à l'intérieur d'une espèce sont ceux qui montrent le plus de différences entre espèces (que ce soit chez les primates ou chez la souris). Pour vérifier si ces gènes évoluent de façon neutre, ils ont comparé l'évolution de leur expression à celle de gènes supposés neutres : un ensemble de pseudogènes exprimés. La distribution de la divergence d'expression est similaire chez ces pseudogènes et les autres gènes, ce qui conforte l'hypothèse neutre (Khaitovich *et al.*, 2004). Cependant, il est probable que ces pseudogènes exprimés n'évoluent pas eux-mêmes de façon neutre. Les résultats observés dans cette étude sont contrariés par une étude similaire, qui n'a pas montré de relation linéaire entre divergence d'expression et temps de divergence chez sept espèces de drosophiles (Bedford et Hartl, 2009). Les raisons de ces incohérences peuvent être multiples (statistiques utilisées, types de puces et biais d'utilisation, etc). Devant la difficulté à conclure, Khaitovich *et al.* ont élaboré un second modèle plus complexe. Ils ont utilisé des données d'expression, également chez les primates (Enard *et al.*, 2002), et ont observé, conformément aux prédictions du modèle, une relation strictement linéaire de la divergence des taxons avec la variance d'expression. Ces résultats soutiennent l'idée d'un rôle fort de l'évolution neutre de l'expression des gènes. Il aurait été intéressant de tester une corrélation similaire avec nos données. Cependant, nous n'aurions alors que trois points (divergence entre *D. simulans* et *D. sechellia*, divergence à l'intérieur de *D. simulans*, et divergence à l'intérieur de *D. sechellia*) ce qui est trop peu pour rechercher une corrélation, d'autant que les populations de *D. simulans* n'ont divergé que très récemment, tout comme les lignées de *D. sechellia* (pour rappel, cette espèce ne

comporte qu'une seule population, Legrand *et al.* 2009, bien qu'elle soit finement structurée entre les différentes îles de l'Océan Indien, Legrand *et al.* 2011). Khaitovich *et al.* (2005) voient ce modèle comme l'hypothèse nulle permettant de tester la neutralité d'évolution du transcriptome, tout écart à leur prédiction témoignant de sélection, quelle qu'elle soit. Mais si cette méthode permet de tester la neutralité d'évolution de l'expression, elle ne permet pas de quantifier la part de gènes évoluant de façon neutre, et la part de gènes évoluant de façon adaptative. De plus, elle dépend fortement des hypothèses de départ, qui sont plus ou moins lourdes (Khaitovich *et al.* 2005, ont par exemple ignoré l'existence de la régulation *trans*).

A contrario, d'autres études ont montré une influence forte de la sélection stabilisante sur de l'expression. Ainsi, Lemos *et al.* (2005) ont trouvé que 61 à 100% des gènes évoluaient plus lentement qu'attendu sous une hypothèse neutre, suggérant une action à grande échelle de la sélection stabilisante. Ils ont pour cela ré-analysé plusieurs jeux de données d'expression précédemment publiés, et les ont mis à l'épreuve de modèles neutres, également précédemment publiés. Une autre étude sur le transcriptome de plusieurs espèces de drosophiles a conclu que 67% des gènes sont sous sélection stabilisante, 25% sous sélection positive, 1% montrent une sélection dépendante du taxon, et enfin 7% auraient une divergence cohérente avec un modèle neutre. Ils ont également montré que les facteurs de transcription étaient plus contraints (sélection stabilisante) que leurs cibles (Rifkin *et al.*, 2003).

Une autre approche consiste à rechercher les corrélations de l'expression de certains gènes avec des facteurs environnementaux. Oleksiak *et al.* (2002) ont montré que l'expression de 22% des gènes examinés entre populations étaient corrélés avec la température. Ceci ne prend en compte que l'adaptation à un seul paramètre environnemental, bien que celui-ci soit majeur. Un problème important de ces approches par corrélation, c'est qu'elles ignorent la corrélation entre gènes. En effet, il est probable qu'une partie seulement de ces 22% ait changé sa régulation de façon adaptative, l'autre partie étant entrainée, par des facteurs de transcriptions communs et/ou des contraintes développementales, bref par une co-régulation quelle qu'en soit la cause. De plus, un changement de l'expression de ces gènes peut être la simple conséquence d'une pression environnementale, sans que le changement soit réellement adaptatif (Fay et Wittkopp, 2008).

Les modèles développés par Khaitovich *et al.*(2004,2005), et précédemment décrits, sont basés sur des divergences entre taxons très récentes. Abordant la question par un autre point de vue, Ogasawara et Okubo (2009) ont développé un modèle prenant en compte l'expression individuelle de chaque locus, ainsi que quelques aspects cytologiques (nombre total de transcrits dans une cellule

limité, nombre de gènes exprimés constant). Ces hypothèses semblent relativement raisonnables (Bishop *et al.*, 1974). Le modèle de Ogasawara et Okubo (2009) a, d'après les auteurs, une bonne valeur prédictive pour l'évolution de l'expression à long terme. Tout comme pour Khaitovich *et al.*(2004,2005), ce modèle suppose que la probabilité pour les transcrits d'un gène de passer de une à deux copies est la même que celle de passer de 10 à 20 copies, et non pas de 10 à 11 copies. Ce modèle est essentiellement stochastique, mais il suppose également qu'un gène qui perd son expression de façon stochastique provoquera l'élimination de l'individu porteur par sélection. Les contraintes cytologiques posent également un cadre à la stochasticité. Leur modèle permet d'expliquer plusieurs phénomènes précédemment observés dans la littérature :

- la distribution des transcrits selon une loi de Zipf (Ogasawara *et al.*, 2003), loi qui dit que l'expression d'un gène dépend de son rang d'expression, tel que $Y_i = \frac{K}{n_i}$, avec K une constante, et n_i le rang ($n^{ième}$ gène le plus exprimé), Y_i étant l'expression du gène i ; si le gène le plus exprimé l'est K fois, le dixième l'est $\frac{K}{10}$ fois
- la relation linéaire entre divergence d'expression et temps de divergence (Khaitovich *et al.*, 2004)
- la saturation de la divergence d'expression après un certain temps de divergence (Bedford et Hartl, 2009)

En conclusion, ces résultats sont variés, il n'y a pour l'instant pas de consensus sur la part de sélection (stabilisante et directionnelle) et la part d'évolution neutre en ce qui concerne l'expression des gènes. Si de rares cas de sélection positive comme celui de *Cyp6g1* sont très bien documentés, l'expression de la grande majorité des gènes évolue sous des pressions dont la nature est toujours controversée. Il est probable que l'expression des gènes soit le résultat d'un équilibre fin entre sélection stabilisante et dérive bornée (par les contraintes cytologiques, comme décrit par Ogasawara et Okubo 2009). Peut-être l'arrivée des données de haut débit permettra-t-elle d'avancer dans ce débat en affranchissant les données de transcriptome des techniques de quantification par hybridation, permettant ainsi une évaluation directe du nombre de transcrits.

4.4 Techniques d'étude d'expression

Il existe actuellement plusieurs outils pour étudier l'expression des gènes. De la PCR quantitative en temps réel, aux puces à ADN (monocanales ou bicanales), en passant par les techniques de séquençage haut-débit, les outils pour étudier l'expression sont de plus en plus performants. La PCR quantitative est un peu à part parmi ces outils, puisqu'elle ne permet l'étude simultanée que

d'un nombre restreint de gènes. C'est l'apparition des puces à ADN miniaturisées dans les années 90 qui a permis les premières études sur un grand nombre de gènes simultanément. Les puces à ADN ont eu leur heure de gloire dans les années 2000, mais avec l'apparition des techniques de séquençage haut débit, quel avenir ont-elles maintenant ? Nous allons faire une analyse comparative rapide de ces deux techniques, et tenter de faire ressortir les avantages et inconvénients de chacune.

4.4.1 Principes généraux des puces / du séquençage haut débit

Les puces à ADN utilisent l'hybridation spécifique d'ADN complémentaire marqué pour quantifier l'ARNm d'origine. Elles permettent la quantification des gènes de l'ensemble d'un génome, mais sont conditionnées par les sondes déposées sur le support solide. Elles seront donc limitées à des transcrits attendus, et donc à des génomes (ou transcriptomes) séquencés.

Le séquençage haut débit a pour principe de base de compter directement les fragments d'ADNc obtenus à partir de l'ARNm. Si cette technique peut permettre une quantification sans *a priori* sur les transcrits rencontrés, elle est cependant dépendante d'une référence, que celle-ci soit un génome séquencé ou une banque d'EST (Expressed Sequence Tag) précédemment séquencée.

Les avantages des puces sont de plusieurs ordres. Le premier s'atténuera rapidement dans les prochaines années : actuellement, l'utilisation des puces est très bien maitrisée, et très bien cadrée, que ce soit d'un point de vue expérimental ou d'un point de vue statistique. Elles permettent de très bonnes quantifications des transcrits, via une réplication généralement peu coûteuse, et faisable à grande échelle. Cependant, par la nature de la quantification via une hybridation, elles sont sensibles à plusieurs biais :

- l'expression ne peut être quantifiée que jusqu'à un certain palier au dessus duquel le spot sature
- des gènes avec une séquence proche (par exemple des paralogues) peuvent s'hybrider sur le mauvais spot, faussant ainsi la quantification (hybridation croisée)
- l'étude est limitée aux transcrits déjà bien connus et référencés
- elles ne permettent pas l'accès à la séquence du transcrit, ce qui ne permet pas de différencier les variants alléliques, ni même parfois les paralogues
- l'hybridation peut être affectée par la divergence de séquence, ce qui limite les applications en évolution

A contrario, les techniques de séquençage haut débit ne saturent pas, même si elles sont limitées par la quantité totale de transcrits lus (ceci dit, avec les quantités en perpétuelle augmentation

produite, cette limite n'en n'est déjà quasiment plus une). Si les séquences permettent de distinguer les différents variants, il est possible que des variations de séquences affectent la cartographie vers les gènes références sur le génome, un petit peu à la manière des hybridations croisées sur les puces. Les principaux avantages résident dans l'absence d'*a priori* sur les transcrits attendus, ce qui permet également de comparer des espèces proches sans soucis de biais d'hybridation, rendant cette technique particulièrement intéressante en biologie évolutive. Enfin, les techniques de séquençage haut débit permettent l'accès à la séquence, ou au moins à une partie de la séquence du transcrit.

Dans notre deuxième étude, nous avons utilisé une technique appelé 3'DGE (Digital Gene Expression). Cette technique est peu utilisée en transcriptomique, à cause de son inconvénient majeur : son coût, qui est lié à l'isolement des fragments 3' non codants des transcrits. Le second inconvénient, c'est la moindre couverture en terme de longueur de séquence proposée par cette technique, qui permet donc peu d'analyse au niveau des séquences. L'avantage de cette technique réside dans une grande profondeur de quantification, ainsi que dans l'affranchissement de la nécessité de normaliser par la longueur des transcrits.

La technique de haut débit la plus utilisée en transcriptomique s'appelle le RNA-seq. Elle consiste à séquencer les fragments d'ADNc issus de l'ARNm après les avoir cassés aléatoirement, sans aucune sélection préalable, si ce n'est la taille. Cette technique nécessite donc une normalisation des comptages par la taille du transcrit. Sa profondeur de quantification est moindre par rapport au 3'DGE. Cependant, vu les capacités des séquenceurs actuels, c'est un moindre mal. D'autant que le RNA-seq donne accès à la séquence des transcrits, et donc à des analyses poussées, par exemple allèle par allèle (Wittkopp *et al.*, 2008a), approche qui est plus limitée pour le 3'DGE, puisque la taille de la séquence du transcrit sera bien plus courte.

A posteriori, notre choix motivé par l'obtention d'une meilleure profondeur de quantification de nous orienter vers les 3'DGE nous semble discutable. Pour un coût équivalent, il aurait été préférable de s'orienter vers une meilleure réplication des échantillons, afin de réaliser une meilleure analyse statistique. Cette erreur a cependant pu être compensée par l'élaboration d'un protocole statistique par étape (voir méthode page 39). Enfin, lorsque nous avons élaboré notre expérience, les technologies haut débit n'en étaient qu'à leurs balbutiements, et le choix a donc été orienté par les caractéristiques des technologies à ce moment, caractéristiques qui ont ensuite fortement évolué, notamment en terme de nombre de lectures, et donc de profondeur de quantification.

4.5 Principaux enseignements et apports de cette thèse

D'un point de vue technique et humain, cette thèse a été très enrichissante, et a suscité des liens étroits avec les statisticiens, concrétisés par une collaboration. De plus, de nombreux pipeline *ad hoc* ont été développés via des scripts dédiés en langage R ou Perl, langages qui nous étaient inconnus avant cette thèse, ou via les nombreux outils disponibles en ligne pour la drosophile. Enfin, l'utilisation des nouvelles techniques de séquençage a procuré un défi extrêmement intéressant, à la fois d'un point de vue pratique et conceptuel. Avez-vous déjà essayé d'ouvrir un fichier texte pesant pas moins de sept gigaoctets ?

Le principal résultat de cette thèse, c'est l'implication majeure des gènes de détoxification dans l'évolution des populations et espèces, sous l'effet généralement de la sélection positive forte créée par des environnements pollués, et en ce qui concerne *D. sechellia*, d'un relâchement de la sélection lié à l'installation dans une niche écologique particulière. Il est difficile d'évaluer à partir de nos données la part adaptative de l'évolution de l'expression, et la part qui évolue de manière neutre. Nos données d'expression et de variance d'expression sont cohérentes avec ce qui est connu de l'histoire biogéographique de *D. simulans*. Les changements d'expression concerne des gènes regroupés sur les chromosomes par petits groupes, que ce soit dû à l'histoire des familles de gènes ou à des phénomènes de co-régulation.

V Perspectives

5.1 Critiques : ce que nous aurions pu faire mieux

Comme dans toute expérimentation, il y a *a posteriori* des regrets, des choses que nous aurions changées, ou réalisées en plus, si nous y avions pensé *a priori*. Les problèmes de l'étude de puces étaient principalement liés au fait que les données dataient de 2004, ce qui dans un domaine qui évolue aussi rapidement que les techniques d'analyse du transcriptome, était parfois difficile à gérer. Cependant, ces puces étaient très faciles à normaliser, et proposaient des données particulièrement propres.

Les améliorations se situent plus au niveau de la seconde analyse. Nous ne reviendrons pas sur le choix du 3'DGE plutôt que le RNA-seq, et ses aspects positifs comme négatifs, notamment les problèmes de réplication liés au coût (voir page 100). Par ailleurs, nous nous sommes limités quasi strictement (mis à part l'analyse de *Cyp6g1*) à l'analyse transcriptomique, pourtant, il y a de nombreux aspects, notamment physiologiques, qui auraient pu constituer un complément intéressant. Le décalage d'émergence des adultes sur milieu naturel et milieu axénique (voir page 76) aurait été très intéressant à quantifier, mais aussi à examiner, via le comptage du nombre d'œufs (différence dans le temps de développement et/ou délai de ponte). Nous aurions pu également tester la résistance de nos lignées africaines et européennes aux insecticides, via des tests de mortalité. Malheureusement, nous n'avons pas conservé les lignées françaises, et elles ne sont vraiment possibles à collecter qu'à l'automne, ce qui a restreint les expérimentations possibles (ou aurait retardé considérablement la soutenance de cette thèse). La dernière interrogation que nous avons porte sur l'opportunité de l'utilisation d'un milieu dit "naturel". En effet, en laboratoire, l'utilisation de ce type de milieu se heurte à certains obstacles. Soit on choisit de stériliser le milieu en amont par autoclavage, mais à ce moment là, en quoi reste-t-il naturel ? Soit on choisit de l'utiliser tel quel, permettant alors le développement incontrôlé de microorganismes. Notre solution a été intermédiaire : nous avons passé le milieu 24h au congélateur avant de l'utiliser pour nos

drosophiles. Cependant, ceci n'a pas empêché les développements bactériens et fongiques, qui ont eu des conséquences majeures sur cette partie de l'expérience. La stimulation du système immunitaire était très intéressante à observer, mais ce n'était pas l'objectif premier de la manipulation. La notion de milieu naturel s'oppose fondamentalement à la notion de milieu standardisé, et il est donc difficile de savoir à quoi les insectes réagissent.

Cependant, le principal atout de cette étude, qui est la forte représentativité de nos échantillons vis-à-vis des populations naturelles, la rend unique en son genre actuellement.

5.2 Développements futurs

Plusieurs orientations sont possibles et intéressantes, pour poursuivre l'étude présentée dans cette thèse.

5.2.1 Du transcriptome, encore du transcriptome, toujours du transcriptome

Continuer sur des analyses de transcriptome, avec plusieurs objectifs pas forcément exclusifs :

- étudier la part adaptative et la part de gènes évoluant sous l'hypothèse de neutralité, ce qui nécessite de multiplier les taxons avec différents temps de divergence (ou de repérer des pseudogènes exprimés, Khaitovich *et al.* 2005), voire même de réaliser des points dans le temps de l'expression des mêmes populations
- faire l'analyse sur les femelles. Plusieurs études ont montré de fortes divergences de l'expression entre les sexes (Meiklejohn *et al.*, 2003; Zhang *et al.*, 2007; Muller *et al.*, 2011), ce qui serait intéressant à observer dans nos populations
- étudier de façon spécifique certains organes, pour se concentrer sur un phénomène particulier (par exemple, la détoxification via l'étude du corps gras, de l'intestin moyen, ou la sélection sexuelle via l'étude des appareils reproducteurs). En effet, diverses études ont montré à quel point l'expression était tissu spécifique (Oleksiak *et al.*, 2005; Haerty et Singh, 2006; Rottscheidt et Harr, 2007; Catron et Noor, 2008). Il est donc légitime de s'interroger sur la pertinence pour le futur d'étudier seulement le corps entier.

5.2.2 Détoxification et populations naturelles

Notre laboratoire possède actuellement une collection unique de lignées naturelles de *D. simulans*, issues du monde entier. Il serait intéressant d'examiner les différents allèles du gène *Cyp6g1* dans ces populations afin de retracer l'histoire évolutive du locus dans l'espèce, et refaire le parallèle avec *D. melanogaster*. De même, on pourrait étudier les coûts potentiels de l'insertion *Juan*, susceptibles d'expliquer sa faible fréquence dans les populations africaines. On pourrait également suivre l'évolution des fréquences alléliques de l'insertion dans les mêmes populations en refaisant l'analyse à quelques années d'intervalle, afin d'observer si l'allèle augmente en fréquence en Afrique, ce qui sous-entendrait soit qu'il y est apporté par migration, soit que le développement de l'usage des pesticides en Afrique favorise l'envahissement de l'allèle dans la population. À l'inverse, son absence ou son maintien à faible fréquence favoriserait l'hypothèse d'un coût de l'allèle dans un environnement moins contraint par les pesticides.

Nous pourrions également examiner d'autres gènes candidats révélés par la transcriptomique (d'autres cytochromes P450, quelques glutathion transférases), afin de comprendre comment s'est développée la résistance, comment a été modifié le niveau d'expression. Des patrons surprenants tel celui de *Cyp6g1* nous attendent probablement dans de telles études.

5.2.3 Test de l'hypothèse de décanalisation

Il serait intéressant dans le futur d'examiner de plus près l'hypothèse de canalisation / décanalisation (voir page 59) que nous avons proposée pour expliquer les patrons de variance d'expression. Deux approches pourraient être utilisées. La première consisterait à modéliser l'expression dans la population ancestrale avant l'invasion, sous des hypothèses de population grande et stable, et de phénotypes assez contraints. On simulerait ensuite l'invasion d'un nouvel environnement par un sous ensemble de cette population, afin de déterminer si on observe une augmentation subséquente de la variance d'expression. Le second moyen serait de tester cette hypothèse sur une autre espèce, en utilisant un jeu de données d'expression provenant d'un échantillon de lignées d'une population de la zone ancestrale de l'espèce en question, et d'une population dérivée, et ce à condition que l'invasion de l'espace "dérivé" soit récente. Celle de *D. simulans* est extrêmement récente (quelques dizaines à centaines d'années). Quelques organismes pourraient se prêter à cela : *D. suzukii*, dont l'invasion de l'Europe est en cours, mais aussi le poisson ange *Centropyge acanthops*, qui a récemment envahi l'atlantique depuis l'océan indien (cette invasion est néanmoins beaucoup plus

ancienne que celle de *D. simulans*, Bowen *et al.* 2006). Ces deux espèces constituent des exemples, mais il y en a bien d'autres.

Bibliographie

Abbott R. J., Ritchie M. G., et Hollingsworth P. M. 2008. Introduction. speciation in plants and animals : pattern and process. *Philos Trans R Soc Lond B Biol Sci*, 363(1506) : 2965–2969.

Aggarwal K., et Silverman N. 2008. Positive and negative regulation of the *Drosophila* immune response. *BMB Rep*, 41(4) : 267–277.

Altschul S. F., Gish W., Miller W., Myers E. W., et Lipman D. J. 1990. Basic local alignment search tool. *J Mol Biol*, 215(3) : 403–410.

Anders S., et Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol*, 11(10) : R106.

Andersson M. 1994. *Sexual selection*. Princeton University Press, Princeton, New Jersey.

Andolfatto P. 2001. Contrasting patterns of x-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Mol Biol Evol*, 18(3) : 279–290.

Arbuthnott D. 2009. The genetic architecture of insect courtship behavior and premating isolation. *Heredity*, 103(1) : 15–22.

Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., *et al.* 2000. Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1) : 25–29.

Barreau C., Benson E., Gudmannsdottir E., Newton F., et White-Cooper H. 2008. Post-meiotic transcription in *Drosophila* testes. *Development*, 135(11) : 1897–1902.

Baudry E., Derome N., Huet M., et Veuille M. 2006. Contrasted polymorphism patterns in a large sample of populations from the evolutionary genetics model *Drosophila simulans*. *Genetics*, 173 (2) : 759–767.

Bedford T., et Hartl D. L. 2009. Optimization of gene expression by natural selection. *Proc Natl Acad Sci USA*, 106 : 1133–1138.

Begun D. J., et Aquadro C. F. 1995. Molecular variation at the *vermilion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. *Genetics*, 140(3) : 1019–1032.

Begun D. J., Holloway A. K., Stevens K., Hillier L. W., Poh Y. P., Hahn M. W., Nista P. M., Jones C. D., Kern A. D., Dewey C. N., *et al.* 2007. Population genomics : whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*, 5(11) : e310.

Benjamini Y., et Hochberg Y. 1995. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J Roy Stat Soc B*, 57 : 289–300.

Benson E., Klyne G., Gudmannsdottir E., Shotton D., et White-Cooper H. 2006. The *Drosophila* testis gene expression database. In *A drosophila research conference*, number 494A.

Berriz G. F., King O. D., Bryant B., Sander C., et Roth F. P. 2003. Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18) : 2502–2504.

Betancourt A. J., Presgraves D. C., et Swanson W. J. 2002. A test for faster x evolution in *Drosophila*. *Mol Biol Evol*, 19(10) : 1816–1819.

Bishop J. O., Morton J. G., Rosbash M., et Richardson M. 1974. Three abundance classes in HeLa cell messenger RNA. *Nature*, 250(463) : 199–204.

Bogwitz M. R., Chung H., Magoc L., Rigby S., Wong W., O'Keefe M., McKenzie J. A., Batterham P., et Daborn P. J. 2005. *Cyp12a4* confers lufenuron resistance in a natural population of *Drosophila melanogaster*. *Proc Natl Acad Sci USA*, 102(36) : 12807–12812.

Bowen B. W., Muss A., Rocha L. A., et Grant W. S. 2006. Shallow mtDNA coalescence in Atlantic pygmy angelfishes (genus *Centropyge*) indicates a recent invasion from the Indian Ocean. *J Hered*, 97(1) : 1–12.

Brem R. B., Yvert G., Clinton R., et Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568) : 752–755.

Brun A., Cuany A., Mouel T. L., Berge J., et Amichot M. 1996. Inducibility of the *Drosophila melanogaster* cytochrome P450 gene, *Cyp6a2*, by phenobarbital in insecticide susceptible or resistant strains. *Insect Biochem Mol Biol*, 26(7) : 697–703.

Bullard J. H., Purdom E., Hansen K. D., et Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11 : 94.

Cabot E. L., Davis A. W., Johnson N. A., et Wu C. I. 1994. Genetics of reproductive isolation in the *Drosophila simulans* clade : complex epistasis underlying hybrid male sterility. *Genetics*, 137(1) : 175–189.

Capy P., Gasperi G., Biémont C., et Bazin C. 2000. Stress and transposable elements : co-evolution or useful parasites ? *Heredity*, 85 (Pt 2) : 101–106.

Capy P., Pla E., et David J. 1993. Phenotypic and genetic variability of morphometrical traits in natural populations of *Drosophila melanogaster* and *D. simulans*. I. geographic variations. *Gen Sel Evol*, 25 : 517–536.

Carlson J. R., et Hogness D. S. 1985. The Jonah genes : a new multigene family in *Drosophila melanogaster*. *Dev Biol*, 108(2) : 341–354.

Catania F., Kauer M. O., Daborn P. J., Yen J. L., Ffrench-Constant R. H., et Schlotterer C. 2004. World-wide survey of an accord insertion and its association with DDT resistance in *Drosophila melanogaster*. *Mol Ecol*, 13(8) : 2491–2504.

Catron D. J., et Noor M. A. F. 2008. Gene expression disruptions of organism versus organ in *Drosophila* species hybrids. *PLoS ONE*, 3(8) : e3009.

Cavalieri D., Townsend J. P., et Hartl D. L. 2000. Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. *Proc Natl Acad Sci USA*, 97(22) : 12369–12374.

Chakir M., Peridy O., Capy P., Pla E., et David J. R. 1993. Adaptation to alcoholic fermentation in *Drosophila* : a parallel selection imposed by environmental ethanol and acetic acid. *Proc Natl Acad Sci USA*, 90(8) : 3621–3625.

Chapman R. W., Mancia A., Beal M., Veloso A., Rathburn C., Blair A., Holland A. F., Warr G. W., Didinato G., Sokolova I. M., *et al.* 2011. The transcriptomic responses of the eastern oyster, *Crassostrea virginica*, to environmental conditions. *Mol Ecol*, 20(7) : 1431–1449.

Chen S., et Li X. 2007. Transposable elements are enriched within or in close proximity to xenobiotic-metabolizing cytochrome p450 genes. *BMC Evol Biol*, 7 : 46.

Cherry S., et Silverman N. 2006. Host-pathogen interactions in *Drosophila* : new tricks from an old friend. *Nat Immunol*, 7(9) : 911–917.

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

Choudhary M., et Singh R. S. 1987. A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. III. Variations in genetic structure and their causes between *Drosophila melanogaster* and its sibling species *Drosophila simulans*. *Genetics*, 117(4) : 697–710.

Chung H., Bogwitz M. R., McCart C., Andrianopoulos A., Ffrench-Constant R. H., Batterham P., et Daborn P. J. 2007. *Cis*-regulatory elements in the Accord retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *Cyp6g1*. *Genetics*, 175(3) : 1071–1077.

Clark A. G., Eisen M. B., Smith D. R., Bergman C. M., Oliver B., Markow T. A., Kaufman T. C., Kellis M., Gelbart W., Iyer V. N., *et al.* 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167) : 203–218.

Cleveland W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*, 74 : 829–836.

Cobb J. P., Mindrinos M. N., Miller-Graziano C., Calvano S. E., Baker H. V., Xiao W., Laudanski K., Brownstein B. H., Elson C. M., Hayden D. L., *et al.* 2005. Application of genome-wide expression analysis to human health and disease. *Proc Natl Acad Sci USA*, 102(13) : 4801–4806.

Cowmeadow R. B., Krishnan H. R., Ghezzi A., Al'Hasan Y. M., Wang Y. Z., et Atkinson N. S. 2006. Ethanol tolerance caused by slowpoke induction in *Drosophila*. *Alcohol Clin Exp Res*, 30 (5) : 745–753.

Coyne J. A., et Orr H. A. 1989. Two rules of speciation, *Speciation and its consequences*. Sinauer associates, 180–207.

Coyne J. A., Rux J., et David J. R. 1991. Genetics of morphological differences and hybrid sterility between *Drosophila sechellia* and its relatives. *Genet Res*, 57(2) : 113–122.

Custodia N., Won S. J., Novillo A., Wieland M., Li C., et Callard I. P. 2001. *Caenorhabditis elegans* as an environmental monitor using DNA microarray analysis. *Ann N Y Acad Sci*, 948 : 32–42.

Cutter A. D. 2008. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. *Mol Biol Evol*, 25(4) : 778–786.

Daborn P. J., Yen J. L., Bogwitz M. R., Le Goff G., Feil E., Jeffers S., Tijet N., Perry T., Heckel D., Batterham P., *et al.* 2002. A single P450 allele associated with insecticide resistance in *Drosophila*. *Science*, 297(5590) : 2253–2256.

David J. 1962. A new medium for rearing *Drosophila* in axenic conditions. *Dros Inf Serv*, 93 : 28.

David J., et Bocquet C. 1976. Compared toxicities of different alcohols for two *Drosophila* sibling species : *D. melanogaster* and *D. simulans*. *Comp Biochem Physiol C*, 54(2) : 71–74.

Davis A. R., et Kohane I. S. 2009. Expression differences by continent of origin point to the immortalization process. *Hum Mol Genet*, 18(20) : 3864–3875.

De Gregorio E., Spellman P. T., Rubin G. M., et Lemaitre B. 2001. Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. *Proc Natl Acad Sci USA*, 98 (22) : 12590–12595.

Dean M. D., et Ballard J. W. O. 2004. Linking phylogenetics with population genetics to reconstruct the geographic origin of a species. *Mol Phylogenet Evol*, 32(3) : 998–1009.

Dobbelaere J., Josué F., Suijkerbuijk S., Baum B., Tapon N., et Raff J. 2008. A genome-wide RNAi screen to dissect centriole duplication and centrosome maturation in *Drosophila*. *PLoS Biol*, 6(9) : e224.

Dobzhansky T. 1951. *Genetics and the Origin of Species*. Colombia University Press, New York.

Doddapaneni H., et Yadav J. S. 2005. Microarray-based global differential expression profiling of P450 monooxygenases and regulatory proteins for signal transduction pathways in the white rot fungus *Phanerochaete chrysosporium*. *Mol Genet Genomics*, 274(5) : 454–466.

Dworkin I., et Jones C. D. 2009. Genetic changes accompanying the evolution of host specialization in *Drosophila sechellia*. *Genetics*, 181(2) : 721–736.

Eberhard W. G. 1996. *Sexual selection by cryptic female choice*. Princeton University Press, Princeton, New Jersey.

Ekengren S., et Hultmark D. 2001. A family of turandot-related genes in the humoral stress response of *Drosophila*. *Biochem Biophys Res Commun*, 284(4) : 998–1003.

Ellegren H., et Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet*, 8(9) : 689–698.

Enard W., Khaitovich P., Klose J., Zöllner S., Heissig F., Giavalisco P., Nieselt-Struwe K., Muchmore E., Varki A., Ravid R., *et al.* 2002. Intra- and interspecific variation in primate gene expression patterns. *Science*, 296(5566) : 340–343.

Enayati A. A., Ranson H., et Hemingway J. 2005. Insect glutathione transferases and insecticide resistance. *Insect Mol Biol*, 14(1) : 3–8.

Evans T. G., Hammill E., Kaukinen K., Schulze A. D., Patterson D. A., English K. K., Curtis J. M. R., et Miller K. M. 2011. Transcriptomics of environmental acclimatization and survival in wild adult Pacific sockeye salmon (*Oncorhynchus nerka*) during spawning migration. *Mol Ecol*.

Falconer D. S., et Mackay T. F. C. 1996. *Introduction to quantitative genetics*. Fourth Edition.

Fay J. C., et Wittkopp P. J. 2008. Evaluating the role of natural selection in the evolution of gene regulation. *Heredity*, 100(2) : 191–199.

Feyereisen R. 1999. Insect p450 enzymes. *Annu Rev Entomol*, 44 : 507–533.

Feyereisen R. 2006. Evolution of insect p450. *Biochem Soc Trans*, 34(Pt 6) : 1252–1255.

Fraser H. B., Moses A. M., et Schadt E. E. 2010. Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc Natl Acad Sci USA*, 107(7) : 2977–2982.

Fu X., Fu N., Guo S., Yan Z., Xu Y., Hu H., Menzel C., Chen W., Li Y., Zeng R., *et al.* 2009. Estimating accuracy of RNA-seq and microarrays with proteomics. *BMC Genomics*, 10 : 161.

Fuselli S., Filippo d. C., Mona S., Sistonen J., Fariselli P., Destro-Bisol G., Barbujani G., Bertorelle G., et Sajantila A. 2010. Evolution of detoxifying systems : the role of environment and population history in shaping genetic diversity at human cyp2d6 locus. *Pharmacogenet Genomics*, 20 (8) : 485–499.

Genissel A., McIntyre L. M., Wayne M. L., et Nuzhdin S. V. 2008. *Cis* and *trans* regulatory effects contribute to natural variation in transcriptome of *Drosophila melanogaster*. *Mol Biol Evol*, 25 (1) : 101–110.

Gibson G., et Dworkin I. 2004. Uncovering cryptic genetic variation. *Nat Rev Genet*, 5(9) : 681–690.

Gibson G., Riley-Berger R., Harshman L., Kopp A., Vacha S., Nuzhdin S., et Wayne M. 2004. Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. *Genetics*, 167(4) : 1791–1799.

Gibson G., et Wagner G. 2000. Canalization in evolutionary genetics : a stabilizing theory ? *Bioessays*, 22(4) : 372–380.

Gilad Y., Rifkin S. A., Bertone P., Gerstein M., et White K. P. 2005. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res*, 15(5) : 674–680.

Giraudo M., Unnithan G. C., Le Goff G., et Feyereisen R. 2010. Regulation of cytochrome P450 expression in *Drosophila* : genomic insights. *Pestic Biochem Physiol*, 97(2) : 115–122.

Gogendeau D., et Basto R. 2010. Centrioles in flies : the exception to the rule ? *Semin Cell Dev Biol*, 21(2) : 163–173.

González J., Lenkov K., Lipatov M., Macpherson J. M., et Petrov D. A. 2008. High rate of recent transposable element-induced adaptation in *Drosophila melanogaster*. *PLoS Biol*, 6(10) : e251.

Goto A., Yano T., Terashima J., Iwashita S., Oshima Y., et Kurata S. 2010. Cooperative regulation of the induction of the novel antibacterial Listericin by peptidoglycan recognition protein LE and the JAK-STAT pathway. *J Biol Chem*, 285(21) : 15731–15738.

Greenbaum D., Colangelo C., Williams K., et Gerstein M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol*, 4(9) : 117.

Grimes S. R. 2004. Testis-specific transcriptional control. *Gene*, 343(1) : 11–22.

Gruntenko N. E., et Rauschenbach I. Y. 2008. Interplay of JH, 20E and biogenic amines under normal and stress conditions and its effect on reproduction. *J Insect Physiol*, 54(6) : 902–908.

Gupta S. C., Siddique H. R., Mathur N., Mishra R. K., Mitra K., Saxena D. K., et Chowdhuri D. K. 2007. Adverse effect of organophosphate compounds, dichlorvos and chlorpyrifos in the reproductive tissues of transgenic *Drosophila melanogaster* : 70kDa heat shock protein as a marker of cellular damage. *Toxicology*, 238(1) : 1–14.

Hack C. J. 2004. Integrated transcriptome and proteome data : the challenges ahead. *Brief Funct Genomic Proteomic*, 3(3) : 212–219.

Haerty W., Jagadeeshan S., Kulathinal R. J., Wong A., Ravi Ram K., Sirot L. K., Levesque L., Artieri C. G., Wolfner M. F., Civetta A., *et al.* 2007. Evolution in the fast lane : rapidly evolving sex-related genes in *Drosophila*. *Genetics*, 177(3) : 1321–1335.

Haerty W., et Singh R. S. 2006. Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of *Drosophila*. *Mol Biol Evol*, 23(9) : 1707–1714.

Haldane J. B. S. 1922. Sex ratio and unisexual sterility in animal hybrids. *J Genet*, 12 : 101–109.

Hamblin M. T., et Veuille M. 1999. Population structure among african and derived populations of *Drosophila simulans* : evidence for ancient subdivision and recent admixture. *Genetics*, 153 (1) : 305–317.

Hartl D. L., Dykhuizen D. E., et Dean A. M. 1985. Limits of adaptation : the evolution of selective neutrality. *Genetics*, 111(3) : 655–674.

Hense W., Baines J. F., et Parsch J. 2007. X chromosome inactivation during *Drosophila* spermatogenesis. *PLoS Biol*, 5(10) : e273.

Hey J., et Kliman R. M. 1993. Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. *Mol Biol Evol*, 10(4) : 804–822.

Hollocher H., et Wu C. I. 1996. The genetics of reproductive isolation in the *Drosophila simulans* clade : X vs. autosomal effects and male vs. female effects. *Genetics*, 143(3) : 1243–1255.

Holloway A. K., Lawniczak M. K., Mezey J. G., Begun D. J., et Jones C. D. 2007. Adaptive gene expression divergence inferred from population genomics. *PLoS Genet*, 3(10) : 2007–2013.

Hsieh W.-P., Chu T.-M., Wolfinger R. D., et Gibson G. 2003. Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. *Genetics*, 165(2) : 747–757.

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

Huang D. W., Sherman B. T., et Lempicki R. A. 2009b. Bioinformatics enrichment tools : paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1) : 1–13.

Hughes K. A., Ayroles J. F., Reedy M. M., Drnevich J. M., Rowe K. C., Ruedi E. A., Cáceres C. E., et Paige K. N. 2006. Segregating variation in the transcriptome : *cis* regulation and additivity of effects. *Genetics*, 173(3) : 1347–1355.

Hultmark D. 2003. *Drosophila* immunity : paths and patterns. *Curr Opin Immunol*, 15(1) : 12–19.

Hutter S., Saminadin-Peter S. S., Stephan W., et Parsch J. 2008. Gene expression variation in african and european populations of *Drosophila melanogaster*. *Genome Biol*, 9(1) : R12.

Hyytia P., Capy P., David J. R., et Singh R. S. 1985. Enzymatic and quantitative variation in european and african populations of *Drosophila simulans*. *Heredity*, 54 (Pt 2) : 209–217.

Irvin S. D., Wetterstrand K. A., Hutter C. M., et Aquadro C. F. 1998. Genetic variation and differentiation at microsatellite loci in *Drosophila simulans*. evidence for founder effects in new world populations. *Genetics*, 150(2) : 777–790.

Jaenisch R., et Bird A. 2003. Epigenetic regulation of gene expression : how the genome integrates intrinsic and environmental signals. *Nat Genet*, 33 Suppl : 245–254.

Jagadeeshan S., et Singh R. S. 2005. Rapidly evolving genes of *Drosophila* : differing levels of selective pressure in testis, ovary, and head tissues between sibling species. *Mol Biol Evol*, 22 (9) : 1793–1801.

Jin W., Riley R. M., Wolfinger R. D., White K. P., Passador-Gurgel G., et Gibson G. 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat Genet*, 29(4) : 389–395.

Johannesen J., et Toft S. 2002. A test for reproductive separation of alternate generations in a biennial spiders *Araneus diadematus* (Araneae, Araneidae). *J Arachnol*, 30(1) : 65–69.

Joly D., Bazin C., Zeng L. W., et Singh R. S. 1997. Genetic basis of sperm and testis length differences and epistatic effect on hybrid inviability and sperm motility between *Drosophila simulans* and *D. sechellia*. *Heredity*, 78 (Pt 4) : 354–362.

Joly D., Luck N., et Dejonghe B. 2008. Diversité morphologique et fonctionnelle des spermatozoïdes chez les drosophiles. *J Soc Biol*, 202(2) : 103–112.

Jones C. D. 1998. The genetic basis of *Drosophila sechellia*'s resistance to a host plant toxin. *Genetics*, 149(4) : 1899–1908.

Jones C. D. 2005. The genetics of adaptation in *Drosophila sechellia*. *Genetica*, 123(1-2) : 137–145.

Jordan I. K., Rogozin I. B., Glazko G. V., et Koonin E. V. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet*, 19(2) : 68–72.

Kerr M., Afshari C., Bennett L., Bushel P., Martinez J., Walker N., et Churchill G. 2002. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, 12 : 203–217.

Kerr M. K., Martin M., et Churchill G. A. 2000. Analysis of variance for gene expression microarray data. *J Comput Biol*, 7(6) : 819–837.

Khaitovich P., Pääbo S., et Weiss G. 2005. Toward a neutral evolutionary model of gene expression. *Genetics*, 170(2) : 929–939.

Khaitovich P., Weiss G., Lachmann M., Hellmann I., Enard W., Muetzel B., Wirkner U., Ansorge W., et Pääbo S. 2004. A neutral model of transcriptome evolution. *PLoS Biol*, 2(5) : E132.

Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.

Kliman R. M., Andolfatto P., Coyne J. A., Depaulis F., Kreitman M., Berry A. J., McCarter J., Wakeley J., et Hey J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics*, 156(4) : 1913–1931.

Kopp A., Frank A., et Fu J. 2006. Historical biogeography of *Drosophila simulans* based on Y-chromosomal sequences. *Mol Phylogenet Evol*, 38(2) : 355–362.

Lachaise D., Capy P., Cariou M. L., Joly D., Lemeunier F., et David J. R. 2004. Nine relatives from one African ancestor : population biology and evolution of the *Drosophila melanogaster* subgroup species, Singh R. S., et Uyenoyama M. K. (ed), *The evolution of population biology*. Cambridge University Press, 315–343.

Lachaise D., Cariou M., David J., Lemeunier F., Tsacas L., et Ashburner M. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup, Hecht N. K., Wallace B., et Prance G. T. (ed), *Evolutionary biology*. Plenum Pub. Co, 22 : 159–225.

Lachaise D., et Silvain J. F. 2004. How two afrotropical endemics made two cosmopolitan human commensals : the *Drosophila melanogaster-D. simulans* palaeogeographic riddle. *Genetica*, 120 (1-3) : 17–39.

Lawniczak M. K. N., Holloway A. K., Begun D. J., et Jones C. D. 2008. Genomic analysis of the relationship between gene expression variation and DNA polymorphism in *Drosophila simulans*. *Genome Biol*, 9(8) : R125.

Le Goff G., Boundy S., Daborn P. J., Yen J. L., Sofer L., Lind R., Sabourault C., Madi-Ravazzi L., et Constant f. R. H. 2003. Microarray analysis of cytochrome P450 mediated insecticide resistance in *Drosophila*. *Insect Biochem Mol Biol*, 33(7) : 701–708.

Le Goff G., Hilliou F., Siegfried B. D., Boundy S., Wajnberg E., Sofer L., Audant P., Ffrench-Constant R. H., et Feyereisen R. 2006. Xenobiotic response in *Drosophila melanogaster* : sex dependence of P450 and GST gene induction. *Insect Biochem Mol Biol*, 36(8) : 674–682.

Legrand D., Tenaillon M. I., Matyot P., Gerlach J., Lachaise D., et Cariou M.-L. 2009. Species-wide genetic variation and demographic history of *Drosophila* sechellia, a species lacking population structure. *Genetics*, 182(4) : 1197–1206.

Legrand D., Vautrin D., Lachaise D., et Cariou M.-L. 2011. Microsatellite variation suggests a recent fine-scale population structure of *Drosophila sechellia*, a species endemic of the Seychelles archipelago. *Genetica*, 139(7) : 909–919.

Lemaitre B., et Hoffmann J. 2007. The host defense of *Drosophila melanogaster*. *Annu Rev Immunol*, 25 : 697–743.

Lemos B., Meiklejohn C. D., Cáceres M., et Hartl D. L. 2005. Rates of divergence in gene expression profiles of primates, mice, and flies : stabilizing selection and variability among functional categories. *Evolution*, 59(1) : 126–137.

Levene H. 1960. *Contributions to Probability and Statistics : Essays in Honor of Harold Hotelling*. Stanford University Press.

Lherminier P., et Solignac M. 2005. *De l'espèce*. Éditions Sylepse, Paris.

Liu T., Dartevelle L., Yuan C., Wei H., Wang Y., Ferveur J.-F., et Guo A. 2009. Reduction of dopamine level enhances the attractiveness of male *Drosophila* to other males. *PLoS One*, 4(2) : e4574.

Liu Z., Li X., Prasifka J. R., Jurenka R., et Bonning B. C. 2008. Overexpression of *Drosophila* juvenile hormone esterase binding protein results in anti-jh effects and reduced pheromone abundance. *Gen Comp Endocrinol*, 156(1) : 164–172.

Low W. Y., Ng H. L., Morton C. J., Parker M. W., Batterham P., et Robin C. 2007. Molecular evolution of glutathione S-transferases in the genus *Drosophila*. *Genetics*, 177(3) : 1363–1375.

Lu P., Vogel C., Wang R., Yao X., et Marcotte E. M. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25(1) : 117–124.

Lynch M. 2007. *The origins of genome architecture*. Sinauer associates, Sunderland, Massachusetts.

Macintyre R. J. 1982. Regulatory genes and adaptation—past, present, and future, *Evolutionary biology*. Plenum Publishing Corporation, New York, 247–285.

Malone J. H., Hawkins D. L., et Michalak P. 2006. Sex-biased gene expression in a ZW sex determination system. *J Mol Evol*, 63(4) : 427–436.

Marioni J. C., Mason C. E., Mane S. M., Stephens M., et Gilad Y. 2008. RNA-seq : an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9) : 1509–1517.

Masly J. P., et Presgraves D. C. 2007. High-resolution genome-wide dissection of the two rules of speciation in *Drosophila*. *PLoS Biol*, 5(9) : e243.

Matsuo T., Sugaya S., Yasukawa J., Aigaki T., et Fuyama Y. 2007. Odorant-binding proteins *OBP57d* and *OBP57e* affect taste perception and host-plant preference in *Drosophila sechellia*. *PLoS Biol*, 5(5) : e118.

Mayr E. 1942. *Systematics and the origin of species*. Columbia University Press, New York.

McCart C., Buckling A., et Ffrench-Constant R. H. 2005. DDT resistance in flies carries no cost. *Curr Biol*, 15(15) : R587–R589.

McCart C., et Ffrench-Constant R. H. 2008. Dissecting the insecticide-resistance associated cytochrome P450 gene *Cyp6g1*. *Pest Manag Sci*, 64(6) : 639–645.

McDermott S. R., et Kliman R. M. 2008. Estimation of isolation times of the island species in the *Drosophila simulans* complex from multilocus DNA sequence data. *PLoS ONE*, 3(6) : e2442.

Meiklejohn C. D., Parsch J., Ranz J. M., et Hartl D. L. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. *Proc Natl Acad Sci USA*, 100(17) : 9894–9899.

Mezey J. G., Nuzhdin S. V., Ye F., et Jones C. D. 2008. Coordinated evolution of co-expressed gene clusters in the *Drosophila* transcriptome. *BMC Evol Biol*, 8 : 2.

Michalak P., Malone J. H., Lee I. T., Hoshino D., et Ma D. 2007. Gene expression polymorphism in *Drosophila* populations. *Mol Ecol*, 16(6) : 1179–1189.

Michalak P., et Noor M. A. 2003. Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. *Mol Biol Evol*, 20(7) : 1070–1076.

Michalak P., et Noor M. A. 2004. Association of misexpression with sterility in hybrids of *Drosophila simulans* and *D. mauritiana*. *J Mol Evol*, 59(2) : 277–282.

Miller C. T., Beleza S., Pollen A. A., Schluter D., Kittles R. A., Shriver M. D., et Kingsley D. M. 2007. *cis*-regulatory changes in kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell*, 131(6) : 1179–1189.

Mishler B. D., et Brandon R. N. 1987. Individuality, pluralism, and the phylogenetic species concept. *Biology and Philosophy*, 2 : 397–414.

Moehring A. J., Teeter K. C., et Noor M. A. 2007. Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. *Mol Biol Evol*, 24(1) : 137–145.

Morley M., Molony C. M., Weber T. M., Devlin J. L., Ewens K. G., Spielman R. S., et Cheung V. G. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430 (7001) : 743–747.

Morozova T. V., Mackay T. F. C., et Anholt R. R. H. 2011. Transcriptional networks for alcohol sensitivity in *Drosophila melanogaster*. *Genetics*, 187(4) : 1193–1205.

Muller L., Hutter S., Stamboliyska R., Saminadin-Peter S. S., Stephan W., et Parsch J. 2011. Population transcriptomics of *Drosophila melanogaster* females. *BMC Genomics*, 12(1) : 81.

Musters H., Huntley M. A., et Singh R. S. 2006. A genomic comparison of faster-sex, faster-X, and faster-male evolution between *Drosophila melanogaster* and *Drosophila pseudoobscura*. *J Mol Evol*, 62(6) : 693–700.

Narusaka Y., Narusaka M., Seki M., Umezawa T., Ishida J., Nakajima M., Enju A., et Shinozaki K. 2004. Crosstalk in the responses to abiotic and biotic stresses in *Arabidopsis* : analysis of gene expression in cytochrome P450 gene superfamily by cDNA microarray. *Plant Mol Biol*, 55 (3) : 327–342.

Neckameyer W. S. 1998. Dopamine and mushroom bodies in *Drosophila* : experience-dependent and -independent aspects of sexual behavior. *Learn Mem*, 5(1-2) : 157–165.

Nuzhdin S. V., Wayne M. L., Harmon K. L., et McIntyre L. M. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol Biol Evol*, 21(7) : 1308–1317.

Ogasawara O., Kawamoto S., et Okubo K. 2003. Zipf's law and human transcriptomes : an explanation with an evolutionary model. *C R Biol*, 326(10-11) : 1097–1101.

Ogasawara O., et Okubo K. 2009. On theoretical models of gene expression evolution with random genetic drift and natural selection. *PLoS One*, 4(11) : e7943.

Oleksiak M. F., Churchill G. A., et Crawford D. L. 2002. Variation in gene expression within and among natural populations. *Nat Genet*, 32(2) : 261–266.

Oleksiak M. F., Roach J. L., et Crawford D. L. 2005. Natural variation in cardiac metabolism and gene expression in *Fundulus heteroclitus*. *Nat Genet*, 37(1) : 67–72.

Orr H. A., et Irving S. 2001. Complex epistasis and the genetic basis of hybrid sterility in the *Drosophila pseudoobscura* Bogota-USA hybridization. *Genetics*, 158(3) : 1089–1100.

Orr H. A., et Irving S. 2005. Segregation distortion in hybrids between the Bogota and USA subspecies of *Drosophila pseudoobscura*. *Genetics*, 169(2) : 671–682.

Oshlack A., Chabot A. E., Smyth G. K., et Gilad Y. 2007. Using DNA microarrays to study gene expression in closely related species. *Bioinformatics*, 23(10) : 1235–1242.

Paigen K. 1986. Gene regulation and its role in evolutionary processes, Karlin S., et Nevo E. (ed), *Evolutionary processes and theory*. Harcourt Brace Jovanovich, Orlando, 3–23.

Paillette M., Joly D., et Bizat N. 1997. Differentiation of dialects and courtship strategies in allopatric populations of *Drosophila teissieri*. *J Insect Physiol*, 43(9) : 809–814.

Palumbi S. R. 2009. Speciation and the evolution of gamete recognition genes : pattern and process. *Heredity*, 102(1) : 66–76.

Pedra J. H. F., McIntyre L. M., Scharf M. E., et Pittendrigh B. R. 2004. Genome-wide transcription profile of field- and laboratory-selected dichlorodiphenyltrichloroethane (DDT)-resistant *Drosophila*. *Proc Natl Acad Sci USA*, 101(18) : 7034–7039.

Pereira V., Enard D., et Eyre-Walker A. 2009. The effect of transposable element insertions on gene expression evolution in rodents. *PLoS One*, 4(2) : e4321.

Perez D. E., Wu C. I., Johnson N. A., et Wu M. L. 1993. Genetics of reproductive isolation in the *Drosophila simulans* clade : DNA marker-assisted mapping and characterization of a hybrid-male sterility gene, *Odysseus* (Ods). *Genetics*, 134(1) : 261–275.

Presgraves D. C. 2008. Sex chromosomes and speciation in *Drosophila*. *Trends Genet*, 24(7).

Pritchard C. C., Hsu L., Delrow J., et Nelson P. S. 2001. Project normal : defining normal variance in mouse gene expression. *Proc Natl Acad Sci USA*, 98(23) : 13266–13271.

Prud'homme B., Gompel N., et Carroll S. B. 2007. Emerging principles of regulatory evolution. *Proc Natl Acad Sci USA*, 104 Suppl 1 : 8605–8612.

Prud'homme B., Gompel N., Rokas A., Kassner V. A., Williams T. M., Yeh S.-D., True J. R., et Carroll S. B. 2006. Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. *Nature*, 440(7087) : 1050–1053.

R' kha S., Moreteau B., Coyne J. A., et David J. R. 1997. Evolution of a lesser fitness trait : egg production in the specialist *Drosophila sechellia*. *Genet Res*, 69(1) : 17–23.

Rajput C., Arif E., Vibhuti A., Stobdan T., Khan A. P., Norboo T., Afrin F., et Pasha M. A. Q. 2006. Predominance of interaction among wild-type alleles of cyp11b2 in himalayan natives associates with high-altitude adaptation. *Biochem Biophys Res Commun*, 348(2) : 735–740.

Ranz J. M., Maurin D., Chan Y. S., Grotthuss v. M., Hillier L. W., Roote J., Ashburner M., et Bergman C. M. 2007. Principles of genome evolution in the *Drosophila melanogaster* species group. *PLoS Biol*, 5(6) : e152.

Raymond M., Chevillon C., Guillemaud T., Lenormand T., et Pasteur N. 1998. An overview of the evolution of overproduced esterases in the mosquito *Culex pipiens*. *Philos Trans R Soc Lond B Biol Sci*, 353(1376) : 1707–1711.

Reeve S., Carhan A., Dee C. T., et Moffat K. G. 2007. *Slowmo* is required for *Drosophila* germline proliferation. *Genesis*, 45(2) : 66–75.

Rezen T., Contreras J. A., et Rozman D. 2007. Functional genomics approaches to studies of the cytochrome P450 superfamily. *Drug Metab Rev*, 39(2-3) : 389–399.

Rice W. R. 1987. Speciation via habitat specialization : the evolution of reproductive isolation as a correlated character. *Evolutionary Ecology*, 1 : 301–314.

Richmond J. Q., Jockusch E. L., et Latimer A. M. 2011. Mechanical reproductive isolation facilitates parallel speciation in western north american scincid lizards. *Am Nat*, 178(3) : 320–332.

Rifkin S. A., Kim J., et White K. P. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat Genet*, 33(2) : 138–144.

R'kha S., Capy P., et David J. R. 1991. Host-plant specialization in the *Drosophila melanogaster* species complex : a physiological, behavioral, and genetical analysis. *Proc Natl Acad Sci USA*, 88(5) : 1835–1839.

Robinson M. D., et Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, 11(3) : R25.

Ronald J., Brem R. B., Whittle J., et Kruglyak L. 2005. Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet*, 1(2) : e25.

Rottscheidt R., et Harr B. 2007. Extensive additivity of gene expression differentiates subspecies of the house mouse. *Genetics*, 177(3) : 1553–1567.

Salzman J., Jiang H., et Wong W. 2011. Statistical modeling of RNA-seq data. *Statistical Science*, 26(1) : 62–83.

Schlenke T. A., et Begun D. J. 2004. Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc Natl Acad Sci USA*, 101(6) : 1626–1631.

Schmidt J. M., Good R. T., Appleton B., Sherrard J., Raymant G. C., Bogwitz M. R., Martin J., Daborn P. J., Goddard M. E., Batterham P., *et al.* 2010. Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genet*, 6(6) : e1000998.

Schöfl G., et Schlötterer C. 2004. Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-african *D. simulans*. *Mol Biol Evol*, 21(7) : 1384–1390.

Schöfl G., et Schlötterer C. 2006. Microsatellite variation and differentiation in african and non-african populations of *Drosophila simulans*. *Mol Ecol*, 15(13) : 3895–3905.

Sharma A., Mishra M., Ram K. R., Kumar R., Abdin M. Z., et Chowdhuri D. K. 2011. Transcriptome analysis provides insights for understanding the adverse effects of endosulfan in *Drosophila melanogaster*. *Chemosphere*, 82(3) : 370–376.

Sheehan D., Meade G., Foley V. M., et Dowd C. A. 2001. Structure, function and evolution of glutathione transferases : implications for classification of non-mammalian members of an ancient enzyme superfamily. *Biochem J*, 360(Pt 1) : 1–16.

Singh R. S. 2000. Toward a Unified Theory of Speciation, Singh R. S., et Krimbas C. B. (ed), *Evolutionary genetics. From molecules to morphology*. Cambridge University Press, Cambridge, 570–604.

Singh R. S., et Kulathinal R. J. 2000. Sex gene pool evolution and speciation : a new paradigm. *Genes Genet Syst*, 75(3) : 119–130.

Smith D. T., Hosken D. J., Rostant W. G., Yeo M., Griffin R. M., Bretman A., Price T. A. R., Ffrench-Constant R. H., et Wedell N. 2011. DDT resistance, epistasis and male fitness in flies. *J Evol Biol*, 24(6) : 1351–1362.

Stern D. L. 2000. Evolutionary developmental biology and the problem of variation. *Evolution*, 54(4) : 1079–1091.

Storey J. D., Madeoy J., Strout J. L., Wurfel M., Ronald J., et Akey J. M. 2007. Gene-expression variation within and among human populations. *Am J Hum Genet*, 80(3) : 502–509.

Sturtevant A. H. 1919. A new species closely resembling *Drosophila melanogaster*. *Psyche*, 26 : 153–156.

Sucena E., et Stern D. L. 2000. Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by *cis*-regulatory evolution of ovo/shaven-baby. *Proc Natl Acad Sci USA*, 97(9) : 4530–4534.

Tao Y., Hartl D. L., et Laurie C. C. 2001. Sex-ratio segregation distortion associated with reproductive isolation in *Drosophila*. *Proc Natl Acad Sci USA*, 98(23) : 13183–13188.

Thornton K., et Long M. 2002. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. *Mol Biol Evol*, 19(6) : 918–925.

Tijet N., Helvig C., et Feyereisen R. 2001. The cytochrome P450 gene superfamily in *Drosophila melanogaster* : annotation, intron-exon organization and phylogeny. *Gene*, 262(1-2) : 189–198.

Tiwari A. K., Pragya P., Ram K. R., et Chowdhuri D. K. 2011. Environmental chemical mediated male reproductive toxicity : *Drosophila melanogaster* as an alternate animal model. *Theriogenology*, 76(2) : 197–216.

Tomaru M., Yamada H., et Oguma Y. 2004. Female mate recognition and sexual isolation depending on courtship song in *Drosophila sechellia* and its siblings. *Genes Genet Syst*, 79(3) : 145–150.

Townsend J. P., Cavalieri D., et Hartl D. L. 2003. Population genetic variation in genome-wide gene expression. *Mol Biol Evol*, 20(6) : 955–963.

Tweedie S., Ashburner M., Falls K., Leyland P., McQuilton P., Marygold S., Millburn G., Osumi-Sutherland D., Schroeder A., Seal R., *et al.* 2009. Flybase : enhancing *Drosophila* gene ontology annotations. *Nucleic Acids Res*, 37(Database issue) : D555–D559.

Vandersteen Tymchuk W., O'Reilly P., Bittman J., Macdonald D., et Schulte P. 2010. Conservation genomics of atlantic salmon : variation in gene expression between and within regions of the bay of fundy. *Mol Ecol*, 19(9) : 1842–1859.

Veuille M., Baudry E., Cobb M., Derome N., et Gravot E. 2004. Historicity and the population genetics of *Drosophila melanogaster* and *D. simulans*. *Genetica*, 120(1-3) : 61–70.

Vicoso B., et Charlesworth B. 2009. Effective population size and the faster-X effect : an extended model. *Evolution*, 63(9) : 2413–2426.

Vontas J., Blass C., Koutsos A. C., David J.-P., Kafatos F. C., Louis C., Hemingway J., Christophides G. K., et Ranson H. 2005. Gene expression in insecticide resistant and susceptible *Anopheles gambiae* strains constitutively or after insecticide exposure. *Insect Mol Biol*, 14(5) : 509–521.

Voolstra C., Tautz D., Farbrother P., Eichinger L., et Harr B. 2007. Contrasting evolution of expression differences in the testis between species and subspecies of the house mouse. *Genome Res*, 17(1) : 42–49.

Wang J.-H., Valanne S., et Rämet M. 2010. *Drosophila* as a model for antiviral immunity. *World J Biol Chem*, 1(5) : 151–159.

Warnefors M., Pereira V., et Eyre-Walker A. 2010. Transposable elements : insertion pattern and impact on gene expression evolution in hominids. *Mol Biol Evol*, 27(8) : 1955–1962.

Waters L. C., Zelhof A. C., Shaw B. J., et Ch'ang L. Y. 1992. Possible involvement of the long terminal repeat of transposable element 17.6 in regulating expression of an insecticide resistance-associated P450 gene in *Drosophila*. *Proc Natl Acad Sci USA*, 89(24) : 12209.

Whitehead A., et Crawford D. L. 2005. Variation in tissue-specific gene expression among natural populations. *Genome Biol*, 6(2) : R13.

Whitehead A., et Crawford D. L. 2006a. Neutral and adaptive variation in gene expression. *Proc Natl Acad Sci USA*, 103(14) : 5425–5430.

Whitehead A., et Crawford D. L. 2006b. Variation within and among species in gene expression : raw material for evolution. *Mol Ecol*, 15(5) : 1197–1211.

Whitehead A., Triant D. A., Champlin D., et Nacci D. 2010. Comparative transcriptomics implicates mechanisms of evolved pollution tolerance in a killifish population. *Mol Ecol*, 19(23) : 5186–5203.

Whitney A. R., Diehn M., Popper S. J., Alizadeh A. A., Boldrick J. C., Relman D. A., et Brown P. O. 2003. Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci USA*, 100(4) : 1896–1901.

Wicker-Thomas C., et Hamann M. 2008. Interaction of dopamine, female pheromones, locomotion and sex behavior in *Drosophila melanogaster*. *J Insect Physiol*, 54(10-11) : 1423–1431.

Wilson T. G., DeMoor S., et Lei J. 2003. Juvenile hormone involvement in *Drosophila melanogaster* male reproduction as suggested by the methoprene-tolerant(27) mutant phenotype. *Insect Biochem Mol Biol*, 33(12) : 1167–1175.

Wittkopp P. J., Haerum B. K., et Clark A. G. 2004. Evolutionary changes in *cis* and *trans* gene regulation. *Nature*, 430(6995) : 85–88.

Wittkopp P. J., Haerum B. K., et Clark A. G. 2008a. Independent effects of *cis*- and *trans*-regulatory variation on gene expression in *Drosophila melanogaster*. *Genetics*, 178(3) : 1831–1835.

Wittkopp P. J., Haerum B. K., et Clark A. G. 2008b. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat Genet*, 40(3) : 346–350.

Wolf J. B. W., Bayer T., Haubold B., Schilhabel M., Rosenstiel P., et Tautz D. 2010. Nucleotide divergence vs. gene expression differentiation : comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Mol Ecol*, 19 Suppl 1 : 162–175.

Wolfner M. F., Partridge L., Lewin S., Kalb J. M., Chapman T., et Herndon L. A. 1997. Mating and hormonal triggers regulate accessory gland gene expression in male *Drosophila*. *J Insect Physiol*, 43(12) : 1117–1123.

Wolgemuth D. J., et Watrin F. 1991. List of cloned mouse genes with unique expression patterns during spermatogenesis. *Mamm Genome*, 1(4) : 283–288.

Wray G. A. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet*, 8 (3) : 206–216.

Wu C.-I., et Ting C.-T. 2004. Genes and speciation. *Nat Rev Genet*, 5(2) : 114–122.

Wu D.-D., Irwin D. M., et Zhang Y.-P. 2011. Correlated evolution among six gene families in *Drosophila* revealed by parallel change of gene numbers. *Genome Biol Evol*, 3 : 396–400.

Wurmser F., Ogereau D., Mary-Huard T., Loriod B., Joly D., et Montchamp-Moreau C. 2011. Population transcriptomics : insights from *Drosophila simulans*, *Drosophila sechellia* and their hybrids. *Genetica*, 139(4) : 465–477.

Xu W., Bak S., Decker A., Paquette S. M., Feyereisen R., et Galbraith D. W. 2001. Microarray-based analysis of gene expression in very large gene families : the cytochrome P450 gene superfamily of *Arabidopsis thaliana*. *Gene*, 272(1-2) : 61–74.

Yang X., Schadt E. E., Wang S., Wang H., Arnold A. P., Ingram-Drake L., Drake T. A., et Lusis A. J. 2006. Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res*, 16(8) : 995–1004.

Yassin A., Abou-Youssef A. Y., Bitner-Mathe B., Capy P., et David J. R. 2007. Mesosternal bristle number in a cosmopolitan drosophilid : an X-linked variable trait independent of sternopleural bristles. *J Genet*, 86(2) : 149–158.

Zhang Y., Sturgill D., Parisi M., Kumar S., et Oliver B. 2007. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature*, 450(7167) : 233–237.

Annexes

Dans les annexes, vous trouverez la première publication (dans Genetica), et le second papier qui est en cours de préparation et est loin d'être abouti. Vous trouverez également une liste des contributions orales et écrites lors de congrès, ainsi qu'un fac-similé des posters présentés pendant la thèse. Vous trouverez enfin les listes de gènes différentiellement exprimés.

Contributions

Présentations orales :

- Evosud, février 2009 (Gif-sur-Yvette)
- Evosud, octobre 2011 (Gif-sur-Yvette)
- Integrative Ecological Genomics, octobre 2011 (Roscoff)

Posters :

- International Conference of Zoology, août 2008 (Paris)
- Challenges of speciation research, septembre 2008 (Sheffield)
- European Society of Evolutionary Biology conference, août 2009 (Turin)
- Society for Molecular Biology and Evolution conference, juillet 2010 (Lyon)
- European Society of Evolutionary Biology conference, août 2011 (Tübingen)

Genetica
DOI 10.1007/s10709-011-9566-0

Population transcriptomics: insights from *Drosophila simulans*, *Drosophila sechellia* and their hybrids

François Wurmser · David Ogereau · Tristan Mary-Huard · Béatrice Loriod · Dominique Joly · Catherine Montchamp-Moreau

Received: 22 November 2010 / Accepted: 7 March 2011

Abstract Sequence differentiation has been widely studied between populations and species, whereas interest in expression divergence is relatively recent. Using microarrays, we compared four geographically distinct populations of *Drosophila simulans* and a population of *Drosophila sechellia*, and interspecific hybrids. We observed few differences between populations, suggesting a slight population structure in *D. simulans*. This structure was observed in direct population comparisons, as well as in interspecific comparisons (hybrids vs. parents, *D. sechellia* vs. *D. simulans*). Expression variance is higher in the French and Zimbabwean populations than in the populations from the ancestral range of *D. simulans* (Kenya and Seychelles). This suggests a large scale phenomenon of decanalization following the invasion of a new environment. Comparing *D. simulans* and *D. sechellia*, we revealed 304 consistently differentially expressed genes, with striking overrepresentation of genes of the cytochrome P450 family, which could be related to their role in detoxification as well as in hormone regulation. We also revealed differences in genes involved in Juvenile hormone and Dopamine differentiation. We finally observed very few differentially expressed genes between hybrids and parental populations, with an overrepresentation of X-linked genes.

Keywords Expression differentiation · Population structure · Ecology · CYP450 · Hormone regulation · *Drosophila*

Introduction

Genetic variation at the sequence level has been widely studied within and between species, to identify the forces that drive evolution (e.g., natural selection, genetic drift). The interest has been now turning to changes in gene regulation leading to genetic isolation, and ultimately speciation. Indeed, gene expression can have a strong influence on downstream phenotypes, and therefore its variation is likely to be a target of natural selection (Pavey et al. 2010). Large-scale technologies such as microarrays provide genome wide information that can be used to assess expression evolution (Gilad and Borevitz 2006). Several comparisons are possible: among populations of the same species, among species, as well as between interspecific hybrids and their parents. Natural populations have been previously studied for regulatory variations in several organisms from yeast (Townsend et al. 2003) to hominids (Storey et al. 2007). Two studies have reported differences in gene expression between African and European populations of *Drosophila melanogaster* with respect to sex biased genes (Meiklejohn et al. 2003) and genes involved in toxicity resistance or flight musculature which are potentially involved in adaptation (Hutter et al.

Electronic supplementary material The online version of this article (doi:10.1007/s10709-011-9566-0) contains supplementary material, which is available to authorized users.

F. Wurmser (✉) · D. Ogereau · D. Joly · C. Montchamp-Moreau
Laboratoire Evolution, Génomes et Spéciation, CNRS UPR9034 Avenue de la Terrasse, Gif-sur-Yvette F-91198 Cedex, and Univ Paris-Sud, 91405 Orsay, France
e-mail: francois.wurmser@legs.cnrs-gif.fr

T. Mary-Huard
Statistics and Genomes Team, INRA UMR518/AgroParisTech, 16 rue Claude Bernard, 75231 Paris, France

B. Loriod
U928 INSERM, Parc scientifique de Luminy, 163 Avenue de Luminy, case 928, 13288 Marseille Cedex 09, France

Published online: 19 March 2011

2008). No evolutionary studies so far have compared expression between closely related species and their hybrids using a set of distinct parental populations (contrasting with studies based on a single strain for each population or species). This is the approach we have adopted in the present work, studying closely related *Drosophila* species belonging to the *melanogaster* subgroup. It will allow us to assess at the expression level, not only the extent of population structure but also the importance of misregulation and additivity in hybrids.

We used here different populations of *Drosophila simulans* and a population of the sibling species *Drosophila sechellia*. *Drosophila simulans* and *D. sechellia* have long been thought to have separated about 400.000 years ago (Hey and Kliman 1993; Kliman et al. 2000; Lachaise and Silvain 2004), but more recent data pointed to a split occurring around 250.000 years ago (McDermott and Kliman 2008). On average, the DNA sequence divergence between these two sister species is around 1.5% (McDermott and Kliman 2008).

While these two species are phylogenetically very close, they are ecologically strongly different. *D. simulans,* which originates from eastern Africa or Madagascar (Lachaise et al. 1988; Lachaise and Silvain 2004; Dean and Ballard 2004; Kopp et al. 2006), is now a cosmopolitan generalist. *D. sechellia* is an endemic specialist in the Seychelles islands, and breeds exclusively on *Morinda citrifolia*, a plant highly toxic for other drosophilids (including *D. simulans*) (R'Kha et al. 1991). According to Oleksiak et al. (2002), gene expression differences between populations arise mainly from genetic drift (they did not show more differences within population than between). Our design already takes into account the variation within populations, thus differences shown there may equally represent drift or adaptation. Notably we expect a higher expression variation in *D. simulans* because of the higher diversity of environments occupied by the species compared to *D. sechellia*. The geographic structure of *D. simulans* has been studied using nuclear gene sequences or microsatellites (Hamblin and Veuille 1999; Schöfl and Schlötterer 2006; Baudry et al. 2006). These studies have revealed a differentiation between east African populations, other African populations (notably Zimbabwean), and European populations. In contrast, *D. sechellia* harbours little nucleotide sequence variation, which ranks this species as the least genetically diverse drosophilid (Legrand et al. 2009).

Previous studies which have examined divergence in gene expression between *D. simulans* and closely related species, have used samples isolated from a single isofemale line (i.e. consisting of the offspring of a single wild-caught female) (Michalak and Noor 2003; Haerty and Singh 2006; Moehring et al. 2007). Thus, they have not taken into account the intraspecific variation. An exception from this is the recent parallel studies of sequence polymorphism and expression of six *D. simulans* lines originating from five different locations within the species range (Holloway et al. 2007; Lawniczak et al. 2008). The authors revealed interesting insights into the evolution of regulatory sequences between populations. However, each of their population was represented by only one (or two in the case of Madagascar) line, thus they could not take into account intra-population variation.

The present study includes four populations of *D. simulans* each represented by four isofemale lines, four isofemale lines of *D. sechellia* and four hybrid "populations" (crosses between *D. sechellia* males and *D. simulans* females). This design allowed us to take into account two different types of variances: intraspecific variance (between the different populations of *D. simulans)*, and intra-population variance (between the different lines which represent biological replicates within each population). We considered the four isofemale lines of *D. sechellia as a single* population as this species does not show geographic structure (Legrand et al. 2009). We used males because they are the most affected sex in *Drosophila* hybrids (since they are heterogametic they show more hybrid breakdown), therefore suitable to highlight differences linked with species divergence. Regulation breakdown in hybrids can occur due to incompatibilities between alleles at a given locus It can also result from negative epistasis between loci according to Dobzhansky-Muller's model of hybrid incompatibility (Dobzhansky 1936; Muller 1942). Differences in hybrid expression can also simply result from regulation divergence between genomes. Hybrids have been previously shown to harbour strong regulation breakdown compared to parents in different organisms (Gibson et al. 2004; Haerty and Singh 2006; Moehring et al. 2007). However, other studies have shown large scale additivity in the patterns of expression of the hybrids (Hughes et al. 2006; Rottscheidt and Harr 2007). These contrasting results may be caused by methodological differences (organ-specific vs. whole body/ differences in microarray platform) or by different degree of divergences and inbreeding between species (Rottscheidt and Harr 2007). These two non-mutually exclusive hypotheses will be considered here. We extracted RNA from whole body males to consider only general differences, and not tissue specific differences.

Our populations of *D. simulans* were from Zimbabwe, Kenya, the Seychelles (the last two likely represent the ancestral range) (Lachaise et al. 1988; Lachaise and Silvain 2004; Dean and Ballard 2004; Kopp et al. 2006), and from France (a derived population). This approach allowed us to determine to what extent gene expression shows geographic structure in *D. simulans*, as well as differentiate between population effects (possibly linked with recent invasion) and species divergence. We observed a population structure in *D. simulans*, and consistent expression divergence

between *D. simulans* and *D. sechellia*. We notably observed a strong involvement of the Cytochrome P450 gene family, as well as genes regulating the juvenile hormone (JH).

Materials and methods

Drosophila stocks

We studied four populations of *D. simulans*. Each population was represented by four isofemale lines (biological replicates). The populations came from France (the Rhone Valley, collected in 2003), Kenya (Nairobi, collected in 2001), Zimbabwe (Mazoe, collected in 1997) and the Seychelles Islands (Mahé and Praslin, collected in 2003). The four *D. sechellia* lines also originated from Mahé and Praslin in the Seychelles archipelago (collected in 2003).

RNA samples

The lines were mass reared in uncrowded culture, on axenic standard medium at 25°C, with a natural light cycle. For each isofemale line, $\geq$6 replicate cultures were raised in vials containing each 8 males and 8 virgin females (10 of each for *D. sechellia*). For a given *D. simulans* population, four different crosses were performed, each between females of one of the line and males from a different line of *D. sechellia* (at least 3 replicates per cross). We thus obtained four "populations" of hybrids corresponding to the four *D. simulans* populations (Table 1). The experimental design thus included a total of nine "populations", each consisting of four biological replicates. Virgin male offspring from at least 3 replicate vials were collected within a few hours of emergence to create pools of 25 individuals that were transferred into fresh vials. Seven days later, each pool was frozen at −80°C. We used the Nucleospin RNA II kit from Macherey–Nagel to extract the RNA from the pools of 25 whole-body adult males, yielding to a total of about 3 μg RNA to hybridize on the arrays. RNA was then reverse transcribed in presence of alpha d'CTP p33.

Arrays

The arrays were nylon filters spotted with long amplicons from the species *D. melanogaster*. They were hybridised in the TAGC platform in Marseille. There were 7,041 spots: 5,931 different whole cDNA of *D. melanogaster* and 1,100 control spots, either negative, or positive controls (a cDNA of *Arabidopsis thaliana*: chlorophylle synthetase). cDNAs were cloned into a vector, amplified and spotted on the array. Each spotted fragment contained both the cDNA and a specific part of the vector for spotting normalization (first hybridisation). These arrays were hybridised twice. Firstly, hybridisation was performed with a P33 labeled oligonucleotide probe specific to the vector sequence spotted. As every molecule spotted contained this sequence, the radioactive signal (vector hybridisation signal) read was proportional to the quantity of spotted cDNA. Secondly, after deshybridisation of the vector probe, we proceeded with the hybridisation of the cDNA samples. A second radioactive signal was read. It will be further designated as complex hybridisation signal.

Data normalization

The normalization procedure was defined by the manufacturer of the arrays. Every spot for which the vector signal was smaller than 5 times the negative spots' median was eliminated from the analysis. For both signals, background (measured by the median of negative spots) was subtracted. Inter-spots/intra-arrays normalization was then performed by dividing the complex hybridisation signal by the vector hybridisation signal. The last normalization step was to divide the obtained expression value in each array by the corresponding median of all spots (or by the median of positive controls), effectively normalizing the signal between arrays. The two approaches (median of controls/median of all spots) led to the same results. Normalization quality was assessed visually by MA plot and box-plot of normalized expression values for each array (Supplementary Fig. 1). All genes with four or more missing data throughout the 36 arrays were discarded.

Statistical analysis

The statistical analysis was based on the whole set of 36 arrays. To determine differentially expressed genes, an Analysis of Variance (ANOVA) model was fitted for each gene (Kerr et al. 2000, 2002). The fitted model was the following:

$$Y_{ij} = \mu_i + E_{ij},$$

where i is the population index (i = 1,…,9: 4 *D. simulans*, 1 *D. sechellia* and 4 hybrid "populations"), j is the biological replicate index (j = 1,…,4), Y_{ij} is the normalized signal (log-transformed), μ_i is the mean expression for the gene in population i and E_{ij} is the residual variability. This model assumes a common variance for all populations, that is consistently estimated with $36 - 9 = 27$ degrees of freedom for most genes (a few genes have missing values due to normalization). The homogeneity of variance between groups was verified using Levene's test (Levene 1960). We showed a variance homogeneity for all genes but 20 (FDR = 0.1, see Supplementary Fig. 2 for the distribution of *P*-value). We tested the equality of mean expression between *D. sechellia* and *D. simulans*

Table 1 Experimental design to obtain hybrid males

♀ *Drosophila simulans*	Isofemale lines	♂ *Drosophila sechellia* Sech 1	Sech 2	Sech 3	Sech 4
France	F1	■			
	F2		■		
	F3			■	
	F4				■
Zimbabwe	Z1				■
	Z2			■	
	Z3	■			
	Z4		■		
Kenya	K1	■			
	K2		■		
	K3			■	
	K4				■
Seychelles	S1	■			
	S2		■		
	S3			■	
	S4				■

16 hybrids were obtained from crosses of *D. sechellia* males with *D. simulans* females. Each isofemale line of *D. sechellia* was involved in a cross with a different isofemale line of each population of *D. simulans*

populations as well as between *D. simulans* populations. As for the comparisons between a population of hybrids and its two parental populations, we performed tests to compare the mean expression of the hybrid to the mean expression of each parental population, and we also compared the mean expression of the hybrid to the average of the mean expressions of the two parental populations. All these comparisons were performed using usual contrast *t* tests within the ANOVA model. For each comparison, raw *P*-values were adjusted by the Benjamini-Hochberg method, which controls the false discovery rate (FDR) (Benjamini and Hochberg 1995). We used a FDR of 0.1.

A mixed-effects model does not suit our analysis, since estimates of some parental effects would only have been from two values. However, we assessed the effect of our analysis assuming correlated data by simulating data with parental effect. We did not find any increased number of false positive even with a high biological/technical variability ratio, the only consequence was a decreased power (Supplementary Fig. 3).

Patterns of inheritance: additivity, dominance, overdominance

To assess patterns of inheritance, we analyzed the distribution of dominance effects using the ratio d/a, where a is half the difference in expression between the parental populations (*D. sechellia* and respectively each population of *D. simulans*), and d is the expression difference between F1 hybrid and the parental average. If d/a = 0, it means perfect additivity (d = 0), if |d/a|=1, complete dominance and if |d/a| >1, overdominance (Falconer and Mackay 1996). We performed this analysis for our four parental populations, only to those genes which were differentially expressed between the parents, to avoid any bias due to equality of expression between the parents.

Variance comparison

To compare the genomic variability in two populations A and B, we propose the following test. This test looks for an excess of genes with higher (or lower) variance in one population relative to another. For a given gene g, we note $\sigma^2_{g,A}$ and $\sigma^2_{g,B}$ the gene expression variances in populations A and B. If population A harbours more genetic variability than population B, then, for most of the genes, the ratio $R_g = \frac{\sigma^2_{g,A}}{\sigma^2_{g,B}}$ will be higher than 1, whereas R_g should be higher than 1 for roughly 50% of the genes if the two populations are comparable. Therefore a test can be based on the number N_1^{AB} of genes for which the empirical ratio R_g of gene *g* is higher than 1. We note p_{AB} the true proportion of genes for which $\sigma^2_{g,A} > \sigma^2_{g,B}$. We test $H_0 = \{p_{AB} = 1/2\}$ (A and B are comparable) versus $H_1 = \{p_{AB} > 1/2\}$ (variability is higher in A). N_1^{AB} has a binomial distribution $B(G, p_{AB})$ with G the number of genes. The P-value of the test is thus:

$$P(N_1^{AB} > n_{1,obs}^{AB} | p_{AB} = 1/2).$$

This analysis takes into account the number of lines available to estimate each variance. We simulated data with different number of lines (n = 2, 5, 10 and 50) with equal population variance, to measure the impact of a small number of lines on variance estimates. This did not affect the *P*-value distribution and thus the error. Significant tests were around 5% (Supplementary Fig. 4). Further simulation showed this only affects the power of the test.

Gene ontology

Lists of differentially expressed genes were examined for statistical over/under representation of Gene Ontology

(Ashburner et al. 2000) terms using FuncAssociate (Berriz et al. 2003) with a reference background consisting of all genes in the arrays. Our array itself was compared to the whole genome, revealing several ontology biases in the construction of the array. This made essential the use of our array as background when examining differentially expressed genes for gene ontology bias. To further explore the terms and the corresponding genes, we used the Gene Ontology database provided by the Gene Ontology consortium (in May 2008).

Results

The experimental design allowed us to perform multiple comparisons, within *D. simulans*, and between *D. simulans* and *D. sechellia*, as well as their hybrids. By maximizing the biological source of variation (using biological and not merely technical replicates, Altman 2005) in populations (and species when applied), we revealed strongly significant variations. After all gene filtering during the normalization process, we assessed expression for 4,398 genes, which is about a fourth of the *Drosophila* genome. The differences observed are thus non exhaustive, but their consistency between all populations, through our large sampling, is supported by the power of the cross-design.

Comparison between populations of *D. simulans*

We did not detect any significant difference in gene expression between the three African populations. Contrasting with this result, all the comparisons between the French population and each of the three African populations showed differential expression. Respectively 6, 7 and 13 genes were found to be differentially expressed between the French population and the Kenyan, the Seychelles and the Zimbabwean populations (Supplementary Table 1). Six genes were differentially expressed in two pairwise comparisons. No gene was differentially expressed in all three pairwise comparisons. Out of the twelve genes which were over-expressed in the French population compared to at least one of the African population, four are cytochrome P450 genes (significantly over-represented, Fisher's exact test, $P < 0.05$).

We compared variability using a binomial test based on the fact that variance ratios of genes are expected to be half of the time above 1 under the assumption of similar variances, using only pairwise comparisons. *P*-values of the binomial test are provided in Table 2. The variance from the French population is significantly higher than the variance from any other population but the Zimbabwean ($P < 0.005$, Bonferroni corrected threshold). In terms of variance, we observe a differentiation of the French and the Zimbabwean populations compared to other African populations (from the zone of origin of *D. simulans*). All other pairwise comparisons revealed significant differences, even though there is a wide range of *P*-values. It is important to note that this analysis is independent from the test of variance homogeneity gene by gene performed with Levene's test. It is possible to have homogeneity gene by gene, whereas on a global scale, heterogeneity can be observed.

Drosophila simulans vs. *Drosophila sechellia*

The comparisons between *D. sechellia* and *D. simulans* yielded to 347, 337, 353 and 518 genes differentially expressed with the populations of Zimbabwe, Kenya, Seychelles and France, respectively. Details of over/under-expressed genes are shown in Table 3. The striking result is that 304 genes are consistently differentially expressed between all four populations of *D. simulans* and *D. sechellia* (Supplementary Table 2). We can therefore assume that these genes present constitutive expression differences between the two species.

Table 2 Above the diagonal: *P*-values of binomial tests under the assumption of equality of variance between the populations

		D. simulans				*D. sechellia*
		France	Zimbabwe	Seychelles	Kenya	Seychelles
D. simulans	F		4.78e-2	2.94e-69 *	3.75e-200 *	9.16e-128 *
	Z	F ≈ Z		1.20e-53 *	1.81e-148 *	1.03e-81 *
	SimS	SimS < F	SimS < Z		7.51e-41 *	1.83e-4 *
	K	K < F	K < Z	K < SimS		1.20e-25 *
D. sechellia	Sech	Sech < F	Sech < Z	Sech < SimS	Sech > K	

Below the diagonal: direction of the variance change

F France, *Z* Zimbabwe, *SimS* Seychelles, *K* Kenya, *Sech*: *D. sechellia*

* significant ($P < 0.005$, Bonferroni corrected threshold)

Table 3 Number of genes over-/under-expressed in *D. simulans* compared with *D. sechellia*

Population	Total	Over-expressed[a]	Under-expressed[b]
Zimbabwe	347	148	199
Kenya	337	144	193
Seychelles	353	158	195
France	518	214	304

[a] Genes over-expressed in *D. simulans* compared with *D. sechellia*

[b] Genes under-expressed in *D. simulans* compared with *D. sechellia*

Five terms were consistently over-represented in the subset of genes under-expressed in *D. sechellia* compared to every population of *D. simulans*. The molecular function "electron carrier activity" and the cellular components "vesicular fraction" and "microsome" refer to cytochrome P450 genes, as was assessed by examining the intersection of the genes with this annotation in *D. melanogaster*, and our set of differentially expressed genes.

The two other terms (namely "lipid metabolism" and "hormone catabolism") refer to three genes: *Juvenile hormone epoxide hydrolase 1* (*Jheh1*), *Juvenile hormone epoxide hydrolase 3* (*Jheh3*) and *Dopamine N-acetyltransferase* (*Dat*). These three genes are highly pleiotropic as they are directly involved in the regulation of key hormones: Juvenile hormone and Dopamine (DA). *Jheh1* and *Jheh3* are involved in JH regulation by degrading it. According to Gruntenko and Rauschenbach (2008), the JH titre can be assessed by the JH degradation level. Thus, we can assume that an over-expression of *Jheh1* and *Jheh3* in *D. simulans* compared with *D. sechellia* implies a lower JH titre in *D. simulans*. The gene *Dat* is involved in the degradation of DA.

Hybrids vs. parental populations

Expression in male hybrids was compared with males of both parental species, and the mean expression of the parents. By this process, the differentially expressed genes were those different from the parents and therefore not showing the dominance of a parental allele on the other. This also excluded genetic additive effects since their expression had to be significantly different from the parents' mean. A significant difference could be due to underdominance, i.e. failed interaction between the two alleles, or misregulation through negative epistasis. We found few genes perturbed in hybrids. No gene disruption was detected in hybrids obtained with the populations of *D. simulans* from Kenya and from Seychelles. Significant over-expression was detected for four and five genes in hybrids offspring of the French and Zimbabwean populations, respectively (Table 4). X-linked genes were significantly over-represented in this set of genes (Fisher exact test, $P < 0.001$). One gene was common in both comparisons (*Cp110*). This gene is involved in centriole replication, although its precise function is unknown (Dobbelaere et al. 2008). Another gene (*r-cup*) is involved in male meiosis, and its disturbance could have a role in hybrid sterility (Barreau et al. 2008).

Table 4 Genes over-expressed in hybrids compared to parents

Offspring of	Gene localisation
F *D. simulans* × *D. Sechellia*	
Cp110	X
CG14785	X
CG4558	X
Es2	X
Z *D. simulans* × *D. Sechellia*	
Cp110	X
sm	2R
r-cup	X
CG3795	X
CG31108	3R

No genes were found differentially expressed with offspring of the Kenyan and Seychelles population

F French, *Z* Zimbabwean

We also examined patterns of dominance and additivity. For this analysis, we included only genes for which expression is different in the two parental populations. We used the ratio dominance over additivity (d/a). About 45% of genes showed additivity to partial dominance ($0 < d/a < 0.5$), about 26% showed partial to complete dominance ($0.5 < d/a < 1$), and about 29% showed overdominance (Fig. 1).

Discussion

Our multiple comparisons of transcriptomes revealed three main features: 1-geographic differentiation in *D. simulans;* 2- expression divergence between *D. simulans* and *D. sechellia* (about 7% of the genes), notably cytochrome P450 genes, and genes involved in hormone metabolism (JH and DA); 3- only eight genes misregulated in hybrids among which X-linked genes were over-represented.

Assessment of the use of interspecific array

We used arrays carrying cDNA from *D. melanogaster*. The use of interspecific arrays has been shown to introduce a bias in comparisons, due to differential hybridisation caused by sequence divergence (Gilad et al. 2005; Oshlack et al. 2007). However, *D. melanogaster*, one of the closest sister species of *D. simulans* and *D. sechellia*, is approximately as

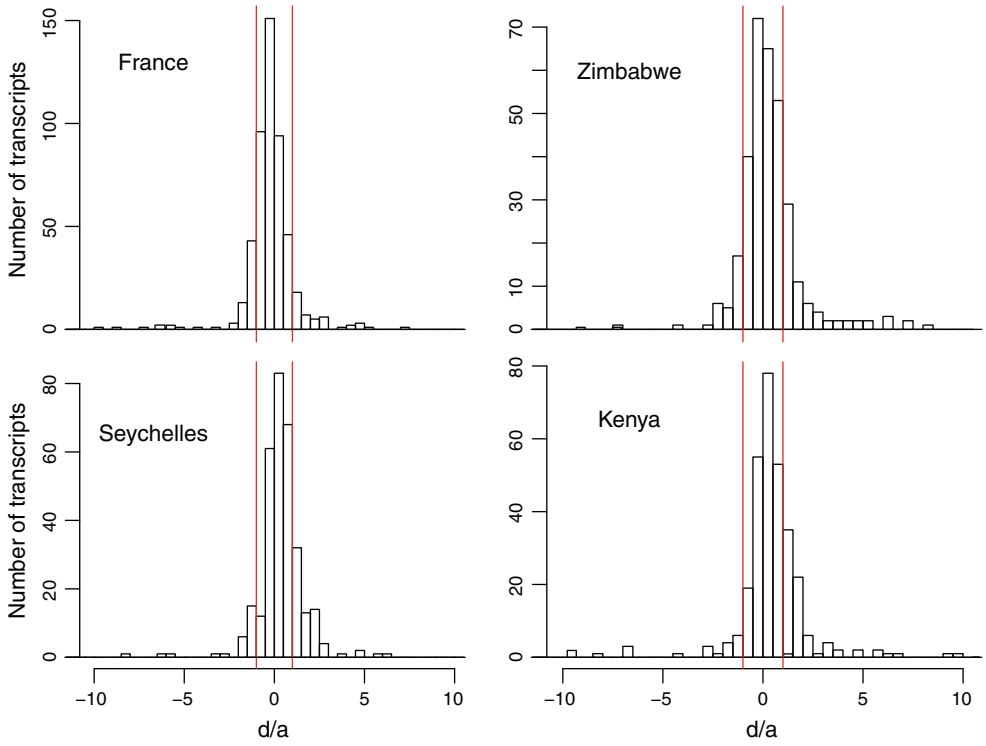


Fig. 1 Distribution of the ratio d/a for hybrids crossed from our four different populations of *D. simulans.* When d/a = 0, d = 0, we are thus showing perfect additivity. From 0 to 1 (or −1), we shift from complete additivity to complete dominance. Over 1 (under −1), we show overdominance

divergent from *D. simulans* and *D. sechellia*. The divergence between *D. sechellia* and *D. simulans* is thought to have occurred around 250.000 years ago, while the divergence between the two species and *D. melanogaster* dates back 2 to 3 million years (Lachaise et al. 1988; Hey and Kliman 1993; McDermott and Kliman 2008). We can thus assume the hybridisation signal to be similarly affected by the use of a *D. melanogaster* array. Moreover, in order to limit the possible bias, we chose a long cDNA array rather than oligonucleotide, which limited the effect of possible mismatches in the sequence. To assess whether there was a hybridisation bias, we looked for a correlation between: 1- the sequence divergence difference between each species and *D. melanogaster* and 2- the mean differences of expression between the two species. We assessed this in a large subset of genes included in the array (2,303 randomly chosen genes), using the sequences from the database (http://www.flybase.org). We observed no correlation (Spearman correlation coefficient: Rho = −0.0054; $P = 0.8056$). *D. simulans* and *D. sechellia* diverge by less than 2%, which is comparable with the level of divergence of disparate populations of *D. melanogaster.* 80% of divergent bases with *D. melanogaster* are shared between *D. simulans* and *D. sechellia* (Dworkin and Jones 2009). In this study, they found little evidence of any bias between *D. sechellia* and *D. simulans*. Although on a global scale, the influence of mismatch is likely negligible, we agree that individual genes can be affected. However, Mezey et al. (2008) showed that the effect of sequence divergence is mainly an increased variance leading to a decreased power of the test. Our tests are thus likely conservative. To assess a possible bias, we used sequences from Flybase to calculate the difference in coding sequence divergence (*D. melanogaster* vs. *D. simulans* minus *D. melanogaster* vs. *D. sechellia*). This was calculated first for differentially expressed genes between *D. sechellia* and *D. simulans*, and second for the random set of genes previously mentioned. Differentially expressed genes are represented by a subset of 232 genes taken from our list of 304 genes consistently differentially expressed between *D. sechellia* and *D. simulans*. We could not consider all genes because of incomplete/erroneous ortholog annotation in Flybase. However, in both sets (random and differentially expressed), genes were sorted the same way. Although the P-value is relatively low, the mean difference in divergence between random genes and differentially expressed genes is not significant (Mann–Whitney, W = 286,527, *P*-value = 0.06643). We assessed further the extent of the possible bias using a Q-Q plot of the distributions (Supplementary Fig. 5). This revealed that about 30 genes out of the 232 are biased toward a higher divergence compared to the random genes distribution. This bias can have two explanations. The first one is that

sequence divergence affected hybridisation, leading to false positive. The second explanation is biological: genes with a higher coding sequence divergence can also be genes evolving faster in terms of expression changes. Although there might be some bias, cytochrome genes as well as *Jheh1*, *Jheh3*, and *Dat* show a divergence difference below 1.3%, situated in the unbiased part of the Q-Q plot. This bias may have two consequences: a decreased power for the test, and a slight increase in the false positive rate.

Differentiation between *D. simulans* populations

Expression differentiation

D. simulans originates from Eastern Africa or Madagascar (Lachaise et al. 1988; Lachaise and Silvain 2004; Dean and Ballard 2004; Kopp et al. 2006). This ancestral area is consistent with several observations of our study. We would like to point out the difference between the African populations of *D. simulans*, and the French population which are revealed with all comparisons. Each of the African population revealed around 350 expression differences with *D. sechellia*, while the French one revealed 518, thus showing a significantly stronger differentiation compared with *D. sechellia*. The direct comparison revealed no differential expression between African populations of *D. simulans*, while we detected several differentially expressed genes between the French population and each African population. This is consistent with the weak but existing population structure observed on microsatellites by Schöfl and Schlötterer (2006) and on nuclear genes by Baudry et al. (2006). Gene flow should be higher between the three African populations than with the French population, as expected from an isolation by distance model and the evolutionary history of the species. A low level of existing geographic differentiation was also described for morphological traits (Gibert et al. 2004). Among African populations, the study of Schöfl and Schlötterer (2006) shows a differentiation between sub-Saharan populations and non sub-Saharan populations. Within the sub-Saharan population a differentiation was observed between the lines from Zimbabwe and Malawi on the one hand, and the lines from Uganda on the other hand, the latter being geographically closer to the likely region of origin of *D. simulans*. Our study suggests a similar differentiation between the Zimbabwean populations and the population from the ancestral zone. Indeed, when we compare F1 interspecific hybrids with their parents, the Zimbabwean population leads to 5 differentially expressed genes (Table 4). This result is comparable with what is observed for the French population (4 genes differentially expressed), but contrasts with the lack of differences observed for the two populations close to the likely origin of *D. simulans*: the Kenyan and the Seychelles populations. The Zimbabwean population shows a very specific pattern. It has been shown to be clearly apart from the eastern populations at the genetic level, but it still has quite high polymorphism (Baudry et al. 2006). Although results in terms of expression differences are contrasted, this population will be further considered as derived.

The study of Baudry et al. (2006) revealed significant differentiation between Eastern African populations and both the French and the Zimbabwean population using four X-linked loci. The authors did not detect any differentiation between populations from Madagascar, Mayotte, Kenya and Tanzania, which is consistent with our observation of similar expression profiles for *D. simulans* of the Seychelles Islands and the Kenya. However we did not detect gene expression differences between the population from Zimbabwe and the two other African populations with direct comparisons. It is likely the differentiation observed on DNA sequences (consistent with hybrid and variance analysis of the present study) is not strong enough to appear on a direct expression comparison, or does not affect expression. This result is consistent with what is known of *D. simulans* biogeographic history: a strong intra-population variation and a relatively weak differentiation between populations (Lachaise et al. 1988; Lachaise and Silvain 2004).

Population structure in expression has been shown in other organisms. In humans, Storey et al. (2007) showed population structure in expression between European and Nigerian cell cultures, although these results are still controversial due to possible differences in the cell immortalisation process (Davis and Kohane 2009). Adaptation to local environment at the level of gene expression has also been shown in other organisms such as *Saccharomyces cerevisae* (Townsend et al. 2003) or on a teleost fish which showed adaptation to temperature (Oleksiak et al. 2002; Whitehead and Crawford 2006).

Population variance, an evidence for decanalization?

We observed a higher interspecific expression variance in derived populations compared with populations from the ancestral area (Eastern Africa/Madagascar). It is interesting to note that DNA sequence variation shows the opposite trend (Schöfl and Schlötterer 2004; Baudry et al. 2006). This resembles the pattern produced by the process of decanalization, i.e. the revelation in a new environment of existing cryptic variation (Gibson and Wagner 2000; Gibson and Dworkin 2004). This hypothesis predicts phenotypic constraints in the ancestral area due to stabilizing selection. At the genetic level, variation can accumulate and not be expressed in the phenotype, because of various buffering mechanisms such as epigenetic interactions,

environmental constraints, etc. When the genotype is transferred in a new environment (for example when there is invasion of a new area), the environment changes, and so does the selective pressure. This can in turn result in the expression of the cryptic genetic variation, which increases the phenotypic variance (in our case, the variance in expression level). This was notably observed for mesosternal bristle number in a drosophilid (Yassin et al. 2007). This process happens despite a reduction in sequence polymorphism due to bottleneck and founding effects (Lachaise et al. 1988; Lachaise and Silvain 2004). However, decanalization has been only described with specific traits, and never at a general scale. Whether this process can be seen by assessing expression variance in all genes is still unknown. Moreover, studying different traits in several populations of *D. simulans*, Capy et al. (1993) revealed morphometric clines, but no difference in variance of traits between derived populations and those from the ancestral range. Adopting a similar approach, we checked expression data from *D. melanogaster* for a similar pattern (data from Hutter et al. 2008). We could not detect any significant variance difference between European and African populations. As *D. melanogaster*'s invasion of the world is older than *D. simulans*'s, it is possible stabilizing selection has acted on regulation, effectively dropping the expression variance in derived populations. The variance difference in *D. simulans* seems biologically significant since it matches the invasion pattern of the species. This tends to eliminate the hypothesis of a technical bias. However, we could not confirm our observation, either on other expression data, or on phenotypic data. This result therefore deserves further investigation.

Gene expression divergence between species

Our study shows 304 genes consistently differentially expressed between the two species. We compared the genes differentially expressed in our study with those revealed by Dworkin and Jones (2009). Thirty-four percent of their differentially expressed genes were present in our array. Out of these, 16% (60 genes) were also in our list of 304 genes differentially expressed between *D. simulans* and *D. sechellia*. The discrepancies can be explained by differences in the design (we compared males while they compared mixed sexes), lines choice (ancient laboratory lines versus recent wild lines), and power of the arrays.

Cytochrome P450 gene family: detoxification or hormone regulation

Cytochrome P450 is a large family of 83 functional genes in *D. melanogaster* (Tijet et al. 2001). The down-regulation of cytochrome P450 genes in *D. sechellia* compared with *D. simulans* (also observed by Dworkin and Jones 2009) can probably be explained by the role of this gene family in detoxification. It could be related to the specialization of the species on the toxic plant *Morinda citrifolia* (Dworkin and Jones 2009). The strict association between *D. sechellia* and its host could have reduced the variety of toxins *D. sechellia* is exposed to, releasing selective constraints on detoxification genes such as the cytochrome P450 gene family. It is also possible that some differences in the need of detoxification genes arose from different environmental conditions (not related to the specialization on *M. citrifolia*). However, we observe cytochrome P450 regulation divergences in comparisons of *D. sechellia* with all four populations of *D. simulans*, despite the fact that they come from four different geographic areas. If the latter hypothesis was supported, we would likely observe differences between all *D. simulans* populations.

Interestingly, the cytochrome P450 genes are also significantly over-represented in the twelve genes that are over-expressed in the French population compared with at least one African population (Supplementary Table 1). Four genes are cytochrome P450s, and one (*Walrus*) has a similar molecular function. It is possible the French population encounters a wider set of toxins due to anthropization processes (Dworkin and Jones 2009) leading to stronger constraints on the cytochrome P450 gene family. However, without data concerning pollutants at the sites of sampling, it is difficult to verify this hypothesis. Alternatively, the divergence of expression of these genes could be related with the involvement of these genes in hormone metabolism (Feyereisen 1999; Tijet et al. 2001), and especially JH regulation.

Divergence of hormonal regulation

We have observed divergence of expression for three genes involved in hormone (notably juvenile hormone and Dopamine) regulation: *jheh1*, *jheh3*, and *dat*. Although the role of JH has been widely described in females of *D. melanogaster* (Gruntenko and Rauschenbach 2008; Liu et al. 2008), it is still poorly known in adult males. However, a physiologic approach has shown a role of JH in seminal fluid protein accumulation in the male reproductive accessory glands (Wolfner et al. 1997), a role also supported by the mutant-based study of Wilson et al. (2003). Mutants with weak receptivity to JH show lower protein accumulation in these glands, and this can in turn affect male fertility. Mutant males also show very little interest in courtship, suggesting a role (either direct or indirect via the perturbation of accessory glands' protein synthesis) of JH in courtship behavior. Dopamine (DA) also plays a role in courtship behavior, a role that could be consistent with the misregulation of *Dat*. The changes in the regulatory

pathway of JH between the two species suggests a change of reproductive behavior in males, which could possibly correspond to a change in females, i.e. coevolution of both sexes via sexual selection. While highly speculative, this hypothesis could be a source (as much as a consequence) of the reproductive/behavioural isolation between the two species. DA has been shown to be involved in JH regulation in females, but has apparently little effect on JH in males as has been shown in *D. virilis* (Gruntenko and Rauschenbach 2008). It is however likely that a change in DA level will affect reproduction in males, but this won't be via JH.

Sterile hybrids, yet weak gene expression differences

We observed 8 genes over-expressed in interspecific hybrids compared to their parental populations, and none under-expressed. Six of these genes are located on the X chromosome. Genes were differentially expressed only for offspring of the two most differentiated hybrid populations: Zimbabwe and France. One gene was common in the two comparisons (*Cp110*).

Impact of the X-chromosome

An interesting result of our analysis is that out of the 8 genes mis-regulated in hybrids, six are located on the X chromosome, a number significantly higher than expected under the assumption of random localization of the differentially expressed genes. This observation is consistent with the so-called "faster-X" effect, which is a commonly mentioned but still controversial characteristic of speciation (Betancourt et al. 2002; Thornton and Long 2002; Musters et al. 2006; Begun et al. 2007; Masly and Presgraves 2007; Presgraves 2008; Vicoso and Charlesworth 2009). According to this theory, X-linked genes evolve more rapidly than genes on autosomes, perhaps due to higher efficiency of selection on the hemizygous X in males. X-linked genes have a specific evolution, due to their presence two-thirds of the time in females and one-third in males and due to their smaller population size than autosomes. However, using introgression, Hollocher and Wu (1996) found no higher density of sterility factor in the X chromosome than on autosomes. This suggests that the X-linked disturbance causing sterility is linked to divergence in regulation and not directly to sequence divergence of X-linked genes.

A low hybrid/parent regulation differentiation

The weak hybrid/parents differentiation observed in our study can be somewhat surprising, compared to results obtained in other studies. For example, using gene expression data on testes, Haerty and Singh (2006) found 241 genes differentially expressed between hybrid and parents. Three differences in the experimental design can explain this discrepancy. First, Haerty and Singh (2006) did not differentiate additive effects. Thus, it is possible that some of their differentially expressed genes represent in fact an averaged expression between the two parental genomes, a possibility we have excluded here. Second (and main) point, their study was on testes. They focused on an organ that is strongly affected as hybrid males between these two species are sterile (Cabot et al. 1994; Joly et al. 1997). Therefore they must have revealed genes involved in this sterility. Our goal was more to examine a global divergence. Therefore, we adopted a whole-body approach, which limited the detection of organ specific divergences, but highlighted more ubiquitous and global changes. Testes have a very specific expression pattern, likely perturbed in hybrids, but these particular differences would be hidden by our approach (Wolgemuth and Watrin 1991; Grimes 2004). Finally, the *D. simulans* line used by Haerty and Singh (2006) was a laboratory strain originating from Arizona, a population geographically far from the African native area of the species where we collected our samples.

The study of Michalak and Noor (2003) revealed 51 genes differentially expressed between hybrids and parents using *D. mauritiana* and *D. simulans*. None of these genes were found differentially expressed in our study. However, it has been previously shown that factors involved in hybrid sterility are different between *simulans/sechellia* vs. *simulans/mauritiana* hybrids (Coyne et al. 1991; Cabot et al. 1994). Another study detected 220 differentially expressed genes (Moehring et al. 2007). No gene is commonly differentially expressed between their study and ours. In this study, there is no correction for multiple testing, potentially allowing for a large number of false positives. If we adopted the same approach, our number of differentially expressed genes would jump to an average of about 67 (Supplementary Table 3), which is similar ($Chi^2 = 1.97$, $df = 1$, $P = 0.16$) to what is observed by Moehring et al. (2007). Interestingly, all these studies, as well as others, in *Drosophila* (Michalak and Noor 2003, 2004; Ranz et al. 2004; Landry et al. 2005; Haerty and Singh 2006) or various other organisms (Wang et al. 2006; Malone et al. 2007; Renaut et al. 2009; Mavarez et al. 2009) detected a large number of genes under-expressed in hybrids compared to parents, and only a few over-expressed. Contrasting with this observation, the present study only showed over-expressed genes in hybrids with FDR correction applied. Without the FDR correction, the representation of over/under-expressed genes is not consistent between the different hybrid populations. In fact, we have strong contrasts between the four different comparisons of hybrids with parents (Supplementary Table 3). This could be related with differences in allele profile between the

parental *D. simulans* populations, therefore allowing for different misregulation patterns in hybrids. Furthermore, the populations we used may lead to hybrids with different properties, owing to the fact that we chose African populations, while other studies have chosen more recently derived American isofemale lines. This aspect remains to be more thoroughly explored.

Intermediate additivity of expression

We observed about 45% of genes showing additivity in expression in the hybrid. This suggests an intermediate pattern compared to what was observed in other studies, from a few percent (Gibson et al. 2004; Haerty and Singh 2006; Moehring et al. 2007) up to 71% (Hughes et al. 2006; Rottscheidt and Harr 2007). The possible reasons for these discrepancies are numerous, as detailed by Rottscheidt and Harr (2007): amount of inbreeding of the parental lines, methodological differences and phylogenic distance between the parents. The last argument probably explains why, using the same method, we observed more dominance than Hughes et al. (2006). Their parental lines are isofemale lines of the same species, while we used populations of two closely related species; it is therefore expected that we would observe more expression disturbances in the form of overdominance (Hughes et al. 2006). It is worth noting that this overdominance is probably still quite small for each individual gene, since we observed very few genes showing expression outside the range of their parental populations.

Cis-regulation, a major player?

Our observations of very few cases of misexpression in hybrids (=very little perturbation of *trans*-regulation) and a large set of expression changes between the two species suggest a major role of *cis*-regulation on the divergence between the two species. This is consistent with an evolutionary model of stronger constraints on *trans*-regulation: the pleiotropic role of transcription factors makes them likely to be more constrained than *cis*-factors, which usually affect only one locus, or even one allele of a given locus. This observation is consistent with other studies on *Drosophila* species (Wittkopp et al. 2004, 2008a, b).

Acknowledgments FW was supported by a PhD fellowship from the "Institut Ecologie et Environnement" of the "Centre National de la Recherche Scientifique". The authors wish to thank Dr Wilfried Haerty for advice both during the analysis and about the manuscript. We also would like to thank Dr Amir Yassin and Dr Francesco Catania for fruitful discussions improving this manuscript. We thank Dr Jean-Jacques Daudin for advice in the statistical analysis. We also thank Dr Michel Piovant for giving access to the array platform. We thank people who provided us with the biological material: Dr Bruno Le Rü, Dr Roland Allemand and Dr Daniel Lachaise (deceased in 2006).

Conflict of interest The authors declare no conflict of interest.

References

Altman N (2005) Replication, variation and normalisation in microarray experiments. Appl Bioinformatics 4:33–44

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25:25–29

Barreau C, Benson E, Gudmannsdottir E, Newton F, White-Cooper H (2008) Post-meiotic transcription in *Drosophila* testes. Development 135:1897–1902

Baudry E, Derome N, Huet M, Veuille M (2006) Contrasted polymorphism patterns in a large sample of populations from the evolutionary genetics model *Drosophila simulans*. Genetics 173:759–767

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP et al (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5:e310

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 57:289–300

Berriz GF, King OD, Bryant B, Sander C, Roth FP (2003) Characterizing gene sets with funcassociate. Bioinformatics 19:2502–2504

Betancourt AJ, Presgraves DC, Swanson WJ (2002) A test for faster X evolution in drosophila. Mol Biol Evol 19:1816–1819

Cabot EL, Davis AW, Johnson NA, Wu CI (1994) Genetics of reproductive isolation in the *Drosophila simulans* clade: complex epistasis underlying hybrid male sterility. Genetics 137:175–189

Capy P, Pla E, David J (1993) Phenotypic and genetic variability of morphometrical traits in natural populations of *Drosophila melanogaster* and *D. simulans*. I. geographic variations. Genet Sel Evol 25:517–536

Coyne JA, Rux J, David JR (1991) Genetics of morphological differences and hybrid sterility between *Drosophila sechellia* and its relatives. Genet Res 57:113–122

Davis AR, Kohane IS (2009) Expression differences by continent of origin point to the immortalization process. Hum Mol Genet 18:3864–3875

Dean MD, Ballard JWO (2004) Linking phylogenetics with population genetics to reconstruct the geographic origin of a species. Mol Phylogenet Evol 32:998–1009

Dobbelaere J, Josué F, Suijkerbuijk S, Baum B, Tapon N et al (2008) A genome-wide RNAi screen to dissect centriole duplication and centrosome maturation in *Drosophila*. PLoS Biol 6:e224

Dobzhansky T (1936) I. Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. Genetics 21:113–135

Dworkin I, Jones CD (2009) Genetic changes accompanying the evolution of host specialization in *Drosophila sechellia*. Genetics 181:721–736

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Longman, London

Feyereisen R (1999) Insect p450 enzymes. Annu Rev Entomol 44:507–533

Gibert P, Capy P, Imasheva A, Moreteau B, Morin JP et al (2004) Comparative analysis of morphological traits among *D. melanogaster* and *D.* simulans: genetic variability, clines and phenotypic plasticity. Genetica 120:165–179

Gibson G, Dworkin I (2004) Uncovering cryptic genetic variation. Nat Rev Genet 5:681–690

Gibson G, Wagner G (2000) Canalization in evolutionary genetics: a stabilizing theory? Bioessays 22:372–380

Gibson G, Riley-Berger R, Harshman L, Kopp A, Vacha S et al (2004) Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. Genetics 167:1791–1799

Gilad Y, Borevitz J (2006) Using DNA microarrays to study natural variation. Curr Opin Genet Dev 16:553–558

Gilad Y, Rifkin SA, Bertone P, Gerstein M, White KP (2005) Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. Genome Res 15:674–680

Grimes SR (2004) Testis-specific transcriptional control. Gene 343: 11–22

Gruntenko NE, Rauschenbach IY (2008) Interplay of JH, 20e and biogenic amines under normal and stress conditions and its effect on reproduction. J Insect Physiol 54:902–908

Haerty W, Singh RS (2006) Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of *Drosophila*. Mol Biol Evol 23: 1707–1714

Hamblin MT, Veuille M (1999) Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. Genetics 153:305–317

Hey J, Kliman RM (1993) Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol Biol Evol 10:804–822

Hollocher H, Wu CI (1996) The genetics of reproductive isolation in the *Drosophila simulans* clade: X vs. autosomal effects and male vs. female effects. Genetics 143:1243–1255

Holloway AK, Lawniczak MK, Mezey JG, Begun DJ, Jones CD (2007) Adaptive gene expression divergence inferred from population genomics. PLoS Genet 3:2007–2013

Hughes KA, Ayroles JF, Reedy MM, Drnevich JM, Rowe KC et al (2006) Segregating variation in the transcriptome: *cis* regulation and additivity of effects. Genetics 173:1347–1355

Hutter S, Saminadin-Peter SS, Stephan W, Parsch J (2008) Gene expression variation in African and European populations of *Drosophila melanogaster*. Genome Biol 9:R12

Joly D, Bazin C, Zeng LW, Singh RS (1997) Genetic basis of sperm and testis length differences and epistatic effect on hybrid inviability and sperm motility between *D. simulans* and *D. sechellia*. Heredity 78:354–362

Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comput Biol 7:819–837

Kerr M, Afshari C, Bennett L, Bushel P, Martinez J et al (2002) Statistical analysis of a gene expression microarray experiment with replication. Stat Sinica 12:203–217

Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M et al (2000) The population genetics of the origin and divergence of the *Drosophila simulans* complex species. Genetics 156:1913–1931

Kopp A, Frank A, Fu J (2006) Historical biogeography of *Drosophila simulans* based on Y-chromosomal sequences. Mol Phylogenet Evol 38:355–362

Lachaise D, Silvain JF (2004) How two afrotropical endemics made two cosmopolitan human commensals: the *D. melanogaster-D. simulans* palaeogeographic riddle. Genetica 120:17–39

Lachaise D, Cariou M, David J, Lemeunier F, Tsacas L et al (1988) Historical biogeography of the *Drosophila melanogaster* species subgroup. Evolutionary biology, pp 159–225

Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG et al (2005) Compensatory *cis-trans* evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. Genetics 171:1813–1822

Lawniczak MKN, Holloway AK, Begun DJ, Jones CD (2008) Genomic analysis of the relationship between gene expression variation and DNA polymorphism in *Drosophila simulans*. Genome Biol 9:R125

Legrand D, Tenaillon MI, Matyot P, Gerlach J, Lachaise D et al (2009) Species-wide genetic variation and demographic history of *Drosophila sechellia*, a species lacking population structure. Genetics 182:1197–1206

Levene H (1960) Contributions to probability and statistics: essays in honor of Harold Hotelling. In: Olkin I (ed). Stanford University Press, Stanford, pp 278–292

Liu Z, Li X, Prasifka JR, Jurenka R, Bonning BC (2008) Overexpression of *Drosophila* juvenile hormone esterase binding protein results in anti-JH effects and reduced pheromone abundance. Gen Comp Endocr 156:164–172

Malone JH, Chrzanowski TH, Michalak P (2007) Sterility and gene expression in hybrid males of *X. laevis* and *X. muelleri*. PLoS One 2:e781

Masly JP, Presgraves DC (2007) High-resolution genome-wide dissection of the two rules of speciation in *Drosophila*. PLoS Biol 5:e243

Mavarez J, Audet C, Bernatchez L (2009) Major disruption of gene expression in hybrids between young sympatric anadromous and resident populations of brook charr (*Salvelinus fontinalis mitchill*). J Evol Biol 8:1708–1720

McDermott SR, Kliman RM (2008) Estimation of isolation times of the island species in the *Drosophila simulans* complex from multilocus DNA sequence data. PLoS One 3:e2442

Meiklejohn CD, Parsch J, Ranz JM, Hartl DL (2003) Rapid evolution of male-biased gene expression in *Drosophila*. Proc Natl Acad Sci USA 100:9894–9899

Mezey JG, Nuzhdin SV, Ye F, Jones CD (2008) Coordinated evolution of co-expressed gene clusters in the *Drosophila* transcriptome. BMC Evol Biol 8:2

Michalak P, Noor MA (2003) Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. Mol Biol Evol 20: 1070–1076

Michalak P, Noor MA (2004) Association of misexpression with sterility in hybrids of *D. simulans* and *D. mauritiana*. J Mol Evol 59:277–282

Moehring AJ, Teeter KC, Noor MA (2007) Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. II. Examination of multiple-species hybridisations, platforms, and life cycle stages. Mol Biol Evol 24:137–145

Muller HJ (1942) Isolating mechanisms, evolution, and temperature. Biol Symp 6:71–125

Musters H, Huntley MA, Singh RS (2006) A genomic comparison of faster-sex, faster-X, and faster-male evolution between *Drosophila melanogaster* and *Drosophila pseudoobscura*. J Mol Evol 62:693–700

Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. Nat Genet 32:261–266

Oshlack A, Chabot AE, Smyth GK, Gilad Y (2007) Using DNA microarrays to study gene expression in closely related species. Bioinformatics 23:1235–1242

Pavey SA, Collin H, Nosil P, Rogers SM (2010) The role of gene expression in ecological speciation. Ann N Y Acad Sci 1206:110–129

Presgraves DC (2008) Sex chromosomes and speciation in *Drosophila*. Trends Genet 7:336–343

R'Kha S, Capy P, David JR (1991) Host-plant specialization in the *Drosophila melanogaster* species complex: a physiological, behavioral, and genetical analysis. Proc Natl Acad Sci USA 88:1835–1839

Ranz JM, Namgyal K, Gibson G, Hartl DL (2004) Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*. Genome Res 14:373–379

Renaut S, Nolte AW, Bernatchez L (2009) Gene expression divergence and hybrid misexpression between lake whitefish

species pairs (*Coregonus spp.* salmonidae). Mol Biol Evol 26:925–936

Rottscheidt R, Harr B (2007) Extensive additivity of gene expression differentiates subspecies of the house mouse. Genetics 177:1553–1567

Schöfl G, Schlötterer C (2004) Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-African *D. simulans*. Mol Biol Evol 21:1384–1390

Schöfl G, Schlötterer C (2006) Microsatellite variation and differentiation in African and non-African populations of *Drosophila simulans*. Mol Ecol 15:3895–3905

Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J et al (2007) Gene-expression variation within and among human populations. Am J Hum Genet 80:502–509

Thornton K, Long M (2002) Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. Mol Biol Evol 19:918–925

Tijet N, Helvig C, Feyereisen R (2001) The cytochrome p450 gene superfamily in *Drosophila melanogaster*: annotation, intron-exon organization and phylogeny. Gene 262:189–198

Townsend JP, Cavalieri D, Hartl DL (2003) Population genetic variation in genome-wide gene expression. Mol Biol Evol 20:955–963

Vicoso B, Charlesworth B (2009) Effective population size and the faster-X effect: an extended model. Evolution 63:2413–2426

Wang J, Tian L, Lee H, Wei NE, Jiang H et al (2006) Genomewide non-additive gene regulation in *Arabidopsis* allotetraploids. Genetics 172:507–517

Whitehead A, Crawford DL (2006) Neutral and adaptive variation in gene expression. Proc Natl Acad Sci USA 103:5425–5430

Wilson TG, DeMoor S, Lei J (2003) Juvenile hormone involvement in *Drosophila melanogaster* male reproduction as suggested by the methoprene-tolerant (27) mutant phenotype. Insect Biochem Mol 33:1167–1175

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in *cis* and *trans* gene regulation. Nature 430:85–88

Wittkopp PJ, Haerum BK, Clark AG (2008a) Regulatory changes underlying expression differences within and between *Drosophila* species. Nat Genet 40:346–350

Wittkopp PJ, Haerum BK, Clark AG (2008b) Independent effects of *cis*- and *trans*-regulatory variation on gene expression in *Drosophila melanogaster*. Genetics 178:1831–1835

Wolfner MF, Partridge L, Lewin S, Kalb JM, Chapman T et al (1997) Mating and hormonal triggers regulate accessory gland gene expression in male *Drosophila*. J Insect Physiol 43:1117–1123

Wolgemuth DJ, Watrin F (1991) List of cloned mouse genes with unique expression patterns during spermatogenesis. Mamm Genome 1:283–288

Yassin A, Abou-Youssef AY, Bitner-Mathe B, Capy P, David JR (2007) Mesosternal bristle number in a cosmopolitan drosophilid: an x-linked variable trait independent of sternopleural bristles. J Genet 86:149–158

Population and resources: consequences on gene expression in *Drosophila simulans*

François Wurmser Tristan Mary-Huard Jean-Jacques Daudin
Dominique Joly Catherine Montchamp-Moreau

Background

Expression is a major, but still poorly known, contributor to adaptation of populations to changing environments and plasticity. Its direct and strong influence on phenotypes makes expression a major target of natural selection (Pavey *et al.*, 2010). Large scale technologies such as microarrays and more recently RNA-seq have allowed whole transcriptome studies of population expression (Gilad and Borevitz, 2006; Marioni *et al.*, 2008). This study aims at exploring large scale changes in expression linked with invasion of a new environment, or with adaptation to environmental shifts due to climate changes. Differentiation of populations in expression has been more and more explored throughout the last few years, ranging from yeast (Townsend *et al.*, 2003) to hominids (Storey *et al.*, 2007). Here we used *D. simulans*.

Three studies have previously studied gene expression in natural populations of *Drosophila*. Studying Zimbabwean and cosmopolitan *D. melanogaster*, Meiklejohn *et al.* (2003) showed a strong bias towards an acceleration of evolution of male biased genes. Hutter *et al.* (2008) and Muller *et al.* (2011) focused on a strong microarray design comparing European (Nederlands) and African (Zimbabwe) *D. melanogaster*, respectively using males and females. They showed a strong discrepancy between male patterns and female patterns, suggesting a sex-specific regulatory adaptation in populations. They also showed low polymorphism within populations and strong differentiation between populations, identifying 153 and 569 differentially expressed genes between populations, respectively for males and females.

D. simulans belongs to the *melanogaster* subgroup. It originates from eastern Africa, between Kenya and Madagascar (Lachaise *et al.*, 1988, 2004; Dean and Ballard, 2004; Kopp *et al.*, 2006). It separated with *D. melanogaster* about two to three million years ago (Lachaise *et al.*, 1988; Hey and Kliman, 1993; Kliman *et al.*, 2000), and with its two sister species *D. sechellia* and *D. mauritiana* about 250 000 years ago (Kliman *et al.*, 2000; McDermott and Kliman, 2008). *Drosophila simulans* and *melanogaster* are both cosmopolitan and generalist, although the invasion of the world by *D. simulans* is thought to be more recent than those of *D. melanogaster* (Lachaise *et al.*, 2004). The recentness of the invasion first led to the idea that *D. simulans* was only slightly structured, an idea supported by allozymes based studies (Choudhary and Singh, 1987) as well as phenotypic data (Capy *et al.*, 1993). This pattern strongly contrasts with what has been shown by several studies on sequence evolution (Irvin *et al.*, 1998; Hamblin and Veuille, 1999; Schöfl and Schlötterer, 2006; Baudry *et al.*, 2006). Working on microsatellite, Schöfl and Schlötterer (2006) showed distinct structure between several populations of *D. simulans*, including between southern Africa and the zone of origin of the species. This pattern was confirmed on nuclear loci by study of Baudry *et al.* (2006). Overall, it seems *D. simulans* shows no structure between populations of Kenya, Tanzania, Madagascar and Mayotte (the presumed ancestral area), but harbors structure between other populations, either from southern or western Africa, Europe, Middle East, North or South America (Irvin *et al.*, 1998; Hamblin and Veuille, 1999; Schöfl and Schlötterer, 2006; Baudry *et al.*, 2006).

In a preliminary study, we showed strong differentiation at the expression level between *D. sechellia* and a French population (518 differentially expressed genes), and a weaker but still strong

expression differentiation of *D. sechellia* with three African popultions (337, 347 and 353 genes differentially expressed) (Wurmser *et al.*, 2011). However, direct comparisons of populations revealed no differences among African populations and few differences between the three African populations and the French one. Here we used next generation sequencing as a powerful tool to check on this result and examine in a powerful manner the differentiation of expression between a French population from the Rhne Valley and an African population from Mayotte, the ancestral range of the species.

Methods

Fly collection Flies were collected directly from their natural habitat. French flies were from an untreated apple orchard located at 44°58'20"N latitude and 4°55'39"E longitude (Rhone Valley). Flies from Mayotte were from a clearing in mid-height of Mayotte main Island, located at 12°48'25"S latitude and 45°9'12"E longitude. They were collected on bananas, one of their local natural resource. After collection single females were placed either on axenic medium, or on the natural resource (apple for french flies, banana for Mayotte flies), at 25°C. Offspring males were placed in vials (medium was the same as for females, either axenic or original resource) for aging. At 5 days old, males were instantly frozen at -80C. We have therefore four different conditions: french flies raised on axenic medium, french flies raised on apple, Mayotte flies raised on axenic medium, Mayotte flies raised on banana.

RNA extraction For each condition, RNA was extracted from 4 pools of 25 males, offspring of 25 different females. RNA was extracted using Nucleospin RNA II kit from Macherey-Nagel. RNA extractions were checked for concentration and quality using both Nanodrop (Thermo Scientific) and microchip electrophoresis (Experion, Bio-Rad). Extractions of the same conditions were then pooled for sequencing. RNA was precipitated in 100% ethanol for transport.

Library preparation and sequencing Library preparation and sequencing were performed by the biotechnological company GATC Biotech (GATC inc.). From the total RNA samples, poly(A) RNA was prepared which was used for cDNA synthesis. cDNA was synthesized using an oligo(dT)-linker primer and M-MLV H reverse transcriptase for first strand synthesis. The reaction conditions were chosen such that the length of the first-strand cDNAs was limited to about 100-500 nt. For Illumina sequencing, the cDNAs in the size range of 250-450 bp were eluted from preparative agarose gels. Library quality was verified on the Shimadzu MultiNA microchip electrophoresis system. 3' cDNA sequencing was performed on Illumina Genome Analyzer according to manufacturer's instructions.

Mapping Mapping was performed by GATC Biotech. We chose to map the sequences first to the *D. melanogaster* genome, and only secondly, for those of the sequences that did not map at the first step, on the *D. simulans* genome. There are two reasons for this two-step mapping. First, the *D. melanogaster* genome is very well annotated, strongly contrasting with *D. simulans* genome. Successful mapping on *D. melanogaster* genes was then much more informative in terms of function. Second, the *D. simulans* genome is poorly sequenced and poorly assembled. This two step mapping should then have increased the number of overall sequences successfully mapped. Mapping was done using ELAND software supplied by Illumina using 32 kmer and allowing up to two mismatches (6.25% error rate). Reads mapped to *D. simulans* were reassociated to there *D. melanogaster* ortholog to simplifiy the analysis (notably the Gene Ontology analysis). Flybase orthology was verified using a divergence analysis, and when estimated necessary (divergence over 21%, corresponding to 93% of alignment of random genes) was checked / corrected with best reciprocal Blast (Altschul *et al.*, 1990). After all these corrections, we had expression information for 15359 genes, 12942 with a *D. melanogaster* ID, and 2417 with a *D. simulans* ID, with no ortholog annotated and no strong result using reciprocal blast. According to Flybase (http://flybase.org) release notes, the genome of *D. simulans* is composed of around 15000 to 17000 genes. We therefore

have a good coverage of the genome, although it is likely some genes are still described by two IDs, despite our efforts to get rid of this possible bias. However, we feel this problem remains marginal, and concerns mainly poorly annotated genes for which we will not be able to analyse the function anyway.

PCR assessment of transposable element insertion We first realised a long PCR with primers flanking the insertion. Sanger sequencing of the beginning of the insert in samples from the french populations showed the insertion was not *Doc* as previously published (Schlenke and Begun, 2004). We used BLAST (Altschul *et al.*, 1990) on this sequences, and results showed another insertion by the name of *Juan*. We then designed a triplex PCR, with two primers flanking the insertion site, and one primer inside the element. The primers were designed so that without insertion, the fragment would be 300bp long, whereas in the presence of the element, the amplified fragment would be 600bp long. All heterozygotes along with two homozygotes of each category were verified by Sanger sequencing.

Statistical analysis In the recent literature, several articles advocated the use of overdispersed or extended Poisson distribution procedures for the analysis of Next Generation Sequencing data (Bullard *et al.*, 2010; Robinson *et al.*, 2010; Salzman *et al.*, 2011). These procedures takes into account both the discrete (counting) and overdispersed nature of the data to handle. Most of these procedures require biological replicates to estimate the variance/overdispersion parameter associated with each gene.

In the present experiment, no biological replicates are available, and the statistical analysis has to be adapted accordingly. A 2 step analysis is performed, under the following hypotheses:

(i) most of the genes are non differentially expressed,

(ii) genes with similar mean expression levels have similar dispersion levels.

In the first step, a gene-per-gene analysis is performed using the following overdispersed Poisson model:

$$X_{gi} \sim \mathcal{P}(\lambda_g, \phi_g) \quad ,$$

where X_{gi} is the observed expression of gene g in condition i, and λ_g and ϕ_g are the mean and dispersion parameters associated with gene g, respectively. Note that in this model the mean expression level λ_g does not depend on the condition, which is relevant for most genes under hypothesis (i). Condition replicates may then be used as biological replicates to obtain an estimate $\widehat{\phi}_g$ of the dispersion parameter ϕ_g. For non differentially expressed genes the variance is unbiasedly estimated, while it is over-estimated for differentially expressed genes. Under hypothesis (ii), a more robust estimation $\tilde{\phi}_g$ of ϕ_g can be obtained using a Loess local estimation of ϕ_g on genes with similar average expression levels. Figure 1 (left) displays the dispersion parameter as a function of the mean expression of the genes, along with the Loess curve of estimates $\tilde{\phi}_g$ (in purple). The Loess is very close to the quadratic curve (in blue) that corresponds to the quadratic relationship between mean and variance of the overdispersed Poisson that is usually assumed in many alternative procedures (Anders and Huber, 2010; Robinson *et al.*, 2010).
In the second step, a gene-per-gene analysis is performed using another overdispersed Poisson model:

$$X_{gi} \sim \mathcal{P}(\lambda_{gi}, \tilde{\phi}_g) \quad ,$$

where λ_{gi} is the mean expression of gene g in condition i, and $\tilde{\phi}_g$ is the dispersion parameter estimated in the previous step. Likelihood Ratio Tests (LRT) can then be performed to test any linear combination of parameters λ_{gi} and obtain a p-value. Figure 1 (right) displays the p-value histogram for the test $H_0 : \{\lambda_{gAM} = \lambda_{gBM}\}$ vs $H1 : \{\lambda_{gAM} \neq \lambda_{gBM}\}$. One can observe a shift of the p-values toward 1 (see values on the right of the histogram). This shows that the first step of

the analysis leads to an overestimation of the dispersion parameter (since most but not all genes are non differentially expressed). Importantly, as mentioned in Anders and Huber (2010), this overestimation decreases the power of the procedure, but does not affect the control of Type I error. Once the p-values are obtained, a classical Benjamini-Hochberg correction (Benjamini and Hochberg, 1995) is performed to control the False Discovery Rate (FDR = 0.05).

Results and Discussion

We used next generation sequencing to assess the expression of two populations (ancestral area versus invaded area) times two resources (natural resource versus artificial resource). The selection of 3'UTRs prior to sequencing allowed us to significantly enhance the depth of quantification. It also got rid of the need to normalize according to the transcript size. After elimination of genes without any read, we assessed expression for 15359 genes.

Population differentiation: candidate genes for expression adaptation

Our analysis revealed 106 consistently differentially expressed genes between our French population and our Mayotte population, independently of the medium. 67 genes are overexpressed in France, while 39 are overexpressed in Mayotte. This difference may be due to the stronger environmental selection the French population is exposed to. We analysed the genes differentially expressed using Gene Ontology tools (FuncAssociate, Berriz *et al.* 2003) to reveal overrepresented attributes. In the 67 genes overexpressed in France, several terms were significantly overrepresented (Table **??**). A detailed analysis of the terms showed that two sets of genes, representing two gene families are described by the ontology terms. The first family is the Cytochrome P450 gene family. The second is the Gluthatione transferase gene family. These two gene families are not revealed in *D. melanoagster* (Hutter *et al.*, 2008; Muller *et al.*, 2011) although the gene*Cyp6g1* is the only one overexpressed in all three studies. There might be two reasons for this discrepancy. First, it is possible that *D. melanogaster* and *D. simulans* have adapted differently, and that their process of invasion did not involve the same genes (at least in terms of expression differentiation). The second hypothesis is that our European population from the Rône Valley was more exposed to pesticides than their populations from Leiden (Netherlands). This hypothesis does not seem likely, as Leiden itself is surrounded by agricultural areas.

In terms of chromosomal location, genes of the same gene family tend to cluster together (Figure 3), especially on the 2R chromosomal arm for Cytochromes, and 2R / 3R concerning Gluthatione Transferase genes. These co-location support a co-regulation hypothesis for these genes. However, other genes of the same family are at the same location. Three hypothesis can explain that these other genes are not found differentially expressed. First, there is indeed co-regulation, but these genes have their own regulation that counteracts the global regulation of the cluster. Second, these genes are co-localised merely for historical basis of the constitution of the family (i.e. tandem gene duplication), but do not retain common regulation. Third, the power of our test could not detect differentiation of other genes of the cluster. The latter hypothesis can be ruled out: we checked expression for other genes in the cluster, and it is clear that expression is stable for these genes.

Gluthatione transferase enzymes, an adaptation to local environment In *D. melanogaster*, Gluthatione transferases (GSTs) are a large gene family composed of 38 members (Low *et al.*, 2007). These genes are divided into classes according to sequence homology and immunological reactions (Sheehan *et al.*, 2001; Enayati *et al.*, 2005). Two of these classes, namely Delta and Epsilon GSTs, are insect specific and have undergone a major expansion in insects via local gene duplication. *D. melanogaster* has 9 Delta and 14 Epsilon functional GSTs (Low *et al.*, 2007). The expansion of these specific GSTs is thought to be associated with local environmental adaptation. Indeed, this gene family has expanded its number independently in *D. melanogaster* and *Anopheles gambiae*, which suggests that these enzymes play a major role in the species local adaptation to their environment. The multiplication of the gene copies should have expanded the range of target GSTs

are able to detoxify. It is believed that adaptation between insect species also occur via adaptation of transcription. It is indeed what we observe here. Our French population, more exposed to pesticides (it was collected from an agricultural area) show stronger expression than our Mayotte population for 4 deltas and 3 epsilons GSTs, independently of the food resource. Thus, this change of transcription is constitutive of a local adaptation of the population to its environment, here selective pressure due to regular use of pesticides.

The Cytochrome P450 gene family This family is composed of approximately 85 functional genes in *D. simulans* (Feyereisen, 1999; Wu *et al.*, 2011). Although it is a very pleiotropic family, its main role is in detoxification of xenobiotics. For example, many genes of this family are underexpressed in the specialist species *Drosophila sechellia* compared with *D. simulans* (Dworkin and Jones, 2009; Wurmser *et al.*, 2011). This species has a strict association with *Morinda citrifolia*, a plant toxic for other *Drosophila*. The number of toxins *D. sechellia* is exposed to is reduced by this strict specialisation, thus relaxing constraints on this gene family and allowing a breakdown of expression (Dworkin and Jones, 2009; Wurmser *et al.*, 2011), as well as a large number of pseudogenisation (Clark *et al.*, 2007; Wu *et al.*, 2011). Cytochrome genes are strongly overexpressed in the population from the Rhône Valley compared with the population from Mayotte (Figure 2). 13 genes are Cytochrome P450, and they are among the genes with the biggest fold ratio (7 out of 10 of the biggest fold ratios). Three additional genes have Cytochrome related functions. Although our *Drosophila* were collected on a field where pesticides are not used, this area is surrounded by regular fields where pesticides are spread on a regular basis. Our population has thus adapted its expression to resist this wide pesticide exposure. The adaptation is genetic since even flies raised on axenic medium (i.e. pesticide free) show the same expression difference.

***Cyp6g1*: altered transcription due selection of pesticide resistance** One gene consistently overexpressed in the French population compared to the Mayotte population was *Cyp6g1*, a cytochrome gene located on the chromosome arm 2R. This gene was expressed approximately ten times more in France than in Mayotte. The history of this gene has been thoroughly studied in *D. melanogaster*. It is strongly linked to broad pesticide resistance, including resistance to DDT (Daborn *et al.*, 2002; Catania *et al.*, 2004) which is considered a good marker of a general role in insecticide resistance (Schmidt *et al.*, 2010). A recent detailed analysis of the locus in *D. melanogaster* revealed five major conformation of the locus, and described the progressive apparition of the different alleles. The ancestral allele (the only one detected in strains collected in the 1930s) is formed by the original *Cyp6g1*, without insertion. In derived populations, they observed first, the insertion of an *Accord* transposable element 300bp upstream of the transcription start site, followed by a tandem duplication of the locus, then the insertion of a *HMS-Beagle* in one of the copy, and finally a *P-element* insertion in the second copy. All these alleles lead to a better fitness in the presence of pesticides than the previous one (Schmidt *et al.*, 2010). It is also correlated with an increase in expression, notably localised in the gastric cecum, the midgut, the Malpighian tubule and the fat body (McCart and Ffrench-Constant, 2008; Chung *et al.*, 2007). Using transgenic constructs, Chung *et al.* (2007) showed that the*Accord* element is likely carrying enhancers that lead to tissue specific increased expression. *Accord* can by itself confer the specificity, (the endogenous *Cyp6g1* shows the same pattern of expression). In the present study, we used whole flies, but the difference in expression was obvious anyway (10-fold). This gene was also shown to have a 4.5 fold ratio in males between European and African population in the study of Hutter *et al.* (2008), and a 3 fold ratio in females (Muller *et al.*, 2011). *Cyp6g1* is the only gene overexpressed in Europe in our study as well as in both of these papers working on *D. melanogaster*.

The derived alleles are present in north Africa, but rare or even absent from eastern / southern Africa. In American, European and Asian populations, they are close to fixation (Catania *et al.*, 2004; Schmidt *et al.*, 2010). A strong reduction of DNA sequence variation in a 20kb zone around the gene suggests a recent selective sweep at the locus (Catania *et al.*, 2004). In a great example of parallel evolution, *D. simulans* has developed a similar response to pesticide selection. Populations from California are nearly fixed (98% frequency) for a *Doc* transposable element insertion, an

insertion correlated with an increase of *Cyp6g1* expression as well as a relative resistance to DDT (direct evidence are still controversial) (Schlenke and Begun, 2004). We designed a PCR assay to investigate the frequency of the *Doc* insertion in our population. We could not find the insertion in any of the population, however, we detected the insertion of a *Juan* transposable element 8 base pairs away from the insertion site of the *Doc* element. We then assessed by PCR the frequency of the insertion in our two populations. Out of 47 males (each from a different female), 43 showed the insertion in France, while four were heterozygotes. In Mayotte, 45 showed no insertion, and two males were heterozygotes. We have 96% of chromosomes harboring the insertion in France and 2% in Mayotte. This stringent pattern suggests that the importance of the insertion in pesticide exposed populations is major, as has been shown in *D. melanogaster* (Catania *et al.*, 2004; Schmidt *et al.*, 2010). The low prevalence of the insertion in Mayotte suggests a cost of the insertion for pesticides free populations, as indeed has been discussed in *D. melanogaster*, but this hypothesis is quite controversial (McCart and Ffrench-Constant, 2008).

Such a strong example of paralell evolution raises questions about the variety of ways to achieve a new phenotype. Indeed, this locus has been consistently selected in different populations and species. Le Goff *et al.* (2003) hypothesized that the singularity of *Cyp6g1* is its broad range of substrates. Indeed selecting populations for DDT resistance, they were able to increase the expression of other cytochrome P450 genes. However, field resistant isolates consistently show overexpression of *Cyp6g1*. Such an example of paralell evolution due to a strong selection for resistance is not unique: the *Resistance to dieldrin* locus, which harbors a mutant linked with insecticide resistance was shown to have arised independtly in different insect species, and even multiple times in *Tribolium casteneum* (Andreev *et al.*, 1999; Ffrench-Constant, 1994).

As for *D. simulans*, the role of *Cyp6g1* in DDT resistance still remains controversial (Schlenke and Begun, 2004). *Increased expression in simulans* (Le Goff *et al.*, 2003)

***Turandot*: stress induced gene family** Strongly differentially expressed in France compared with Mayotte are two genes of the *Turandot* family: *TotC* and *TotZ*. This family is composed of 8 genes in *D. melanogaster*. *TotA* was the first described as encoding a humoral factor in response to high temperature. Contrasting with Heat Shock Proteins, these factors respond slowly to stress via a humoral response of small peptides released in the hemolymph (Ekengren *et al.*, 2001). These genes react to several different form of stress, such as oxidative stress, severe temperature, mechanical shock (Ekengren and Hultmark, 2001). Recent data point to an activation due to endosulfan, an organochloride pesticide (Sharma *et al.*, 2011). All these studies however, have described a plastic although persistent response to external aggression. Here we show a constitutive overexpression of *Tot C* and *Tot Z* in the French population compared with the African. This is the first evidence of adaptive expression in this gene family.

Plasticity: reaction to medium is strongly linked with immunity

We raised both of our populations on their natural medium as well as on axenic medium. The main goal was to investigate changes linked with the invasion of a new medium. However, expression changes showed that the medium shift mainly leads to a relaxation of constraints, at least concerning Banana. Indeed the axenic medium is a uninfected medium, while natural media are easily infected by bacteria and fungi. Within a population, expression changes due to medium shift is therefore strongly linked with stimulation of the immune system. Studies of the immune system in drosophila have relied on injection of bacteria, fungi infection via short time exposure to a culture. To our knowledge, our study is the first showing stimulation of the immune system in a natural and long term exposure. Drosophila have a robust innate immune response. They produce a large number of anti-microbial peptides via the activation of two main regulating pathways: the Toll pathway, directed against fungi and Gram-positive bacteria, and the immune deficiency (IMD) pathway, directed against Gram-negative bacteria (De Gregorio *et al.*, 2001; Lemaitre and Hoffmann, 2007; Aggarwal and Silverman, 2008). Two others pathway regulate immunity: a RNAi pathway, and the JAK/STAT (Janus kinase-signal transducers and activators of transcription) pathway.

Anti-Microbial Peptides About 20 Anti-Microbial Peptides (AMPs) have been identified thus far. Their action is either directed at Gram-negative bacteria (*Diptericin*, *Attacin*, *Drosocin*, *Cecropin* and *Listericin* (Goto *et al.*, 2010)), Gram-positive bacteria (*Defensin*) or fungi (*Drosomycin* and *Metchnikowin*) (Lemaitre and Hoffmann, 2007). Out of these 20 or so genes, 12 are significantly overexpressed in banana fed flies contrasting with axenic fed flies. Overexpression ranges from an induction of *Metchnikowin* 47 times more in banana to the induction of *Diptericin B* with 2.3 fold ratio. These peptides are considered as the effectors of the innate immunity response in *Drosophila*. Therefore their strong and generalized induction shows our individuals are under strong infectious pressure. This was predictable as our banana medium was strongly infected by both bacteria and Fungi. We compared their induction with expression measured on another study (De Gregorio *et al.*, 2001) (see supplementary Figure 1). Some of the genes have an induction similar to a bacterial induced immune stimulation (notably *Attacins, PGRP-SC2, and maybe Duptericin*. Others react as for a fungi infection (*Diptericin B, Listericin, Defensin, PGRP-SB1*), showing only a limited induction. Finally others harbor a very contrasted pattern (notably *Drosocin and Metchnikowin*). These last two genes show a very strong stimulation. These contrasted inductions are likely due to this unprecedented mode of infection: both continuous, and likely with several microbes at a time. The defence response of flies seemed very efficient, since we could not reveal a higher mortality rate on banana.

Apple orchard and pesticides: a cost to fertility ? We used FuncAssociate to examine overrepresented terms within differentially expressed genes of the French population raised either on apple (natural medium) or on axenic. Among genes more expressed on axenic than on apple, overrepresented terms are shown in table 2. Two sets of genes are designated by these terms. The first set is composed of eight genes which have a function in the process of reproduction (Serine protease inhibitors, Accessory Gland Proteins (ACPs), Seminal fluid proteins, an Odorant binding protein and a gene with an unknown molecular function). These genes are not overexpressed in axenic, they are underexpressed on apple. Indeed, their expression level is the same in all conditions, except on apple raised flies, where it suffers a strong breakdown. Two others genes can be added to this set: two ACPs not annotated as involved in reproduction, and showing the same expression pattern. This could be due to the reproductive cost of actually being exposed to pesticides. Indeed our apple medium was composed of apples taken from the field, where they could have suffered from exposure to pesticides. The resistance to these pesticides is costly for the drosophila, leading to a breakdown of some reproductive genes. In humans, it has been shown that exposure to pesticides can affect sperm quality (morphology, motility and sperm counts)(for a review, see Tiwari *et al.* (2011)). Gupta *et al.* (2007) showed that at least some ACPs are affected (expression breakdown) by exposure to pesticides in *D. melanogaster*. Necrosis in male accessory gland was also observed (Gupta *et al.*, 2007; Tiwari *et al.*, 2011). However, other hypothesis are possible, such as differences in medium pH, sugar contents, fermentation of the apples,...

The second set of genes overexpressed in axenic-raised compared with apple-raised flies is the Cytochrome P450 gene family. Five genes of this family were also differentially expressed between the populations. Their expression profile is similar: they show a very strong difference (8 to 16 fold ratio) between populations and a fold ratio of about two between axenic and apple raised flies. Two other Cytochrome P450 genes show an upregulation on axenic compared with apple. The axenic medium contains nipagine, an anti-microbial chemical. It is possible the slight upregulation of these genes on axenic is a response to this chemical.

Three Turandot genes were also overexpressed on axenic *TotC, TotM and TotA*. These genes may be responding to the stress generated by medium change.

No term of Gene Ontology was overrepresented in genes overexpressed in apple-raised flies compared with axenic raised flies.

Conclusion

We compared a population from the ancestral range with a population from an invaded area of *D. simulans*, in natural medium vs. artificial medium. Detoxification genes (Cytochrome P450 and Gluthationes Transferases) were strongly overexpressed in the French population. *Cyp6g1*, a very well known gene was strongly overexpressed, due to a so far unpublished insertion of a *Juan* transposable element.

Flies raised on banana harboured a strong and complex (continuous and from multiple microbes) stimulation of the immune system. Specifically, anti-microbial peptides were strongly inducted.

Finally, apple raised flies showed a reproductive cost, that could be due to plastic exposure to pesticide.

Author contribution

FW, DJ, and CMM designed the experiment. FW and CMM captured the flies. FW proceeded with RNA extraction for sequencing. TMH and JJD designed the statistical analysis. FW analysed the data and wrote the paper, in collaboration with DJ, CMM abd TMH.

Acknowledgments and funding

We would like to thank David Ogereau for his precious and unique technical skills. We also thank GATC, especially Benjamin Moingeon and Dr Yadhu Kumar for the sequencing and helpful discussions. This project was funded by "Agence Nationale de la Recherche", Adaptanthrop project. FW was funded by a CNRS fellowship.

References

Aggarwal K., and Silverman N. 2008. Positive and negative regulation of the *Drosophila* immune response. *BMB Rep*, 41(4): 267–277.

Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. 1990. Basic local alignment search tool. *J Mol Biol*, 215(3): 403–10.

Anders S., and Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol*, 11(10): R106.

Andreev D., Kreitman M., Phillips T. W., Beeman R. W., and Constant f. R. H. 1999. Multiple origins of cyclodiene insecticide resistance in *Tribolium castaneum* (coleoptera: Tenebrionidae). *J Mol Evol*, 48(5): 615–624.

Baudry E., Derome N., Huet M., and Veuille M. 2006. Contrasted polymorphism patterns in a large sample of populations from the evolutionary genetics model *Drosophila simulans*. *Genetics*, 173(2): 759–67.

Benjamini Y., and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57: 289300. Controlling the false discovery rate: a practical and powerful approach to multiple testing.

Berriz G. F., King O. D., Bryant B., Sander C., and Roth F. P. 2003. Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18): 2502–4.

Bullard J. H., Purdom E., Hansen K. D., and Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, 11: 94.

Capy P., Pla E., and David J. 1993. Phenotypic and genetic variability of morphometrical traits in natural populations of *Drosophila melanogaster* and d simulans. i. geographic variations. *Genetics, selection, evolution*, 25: 517–536.

Catania F., Kauer M. O., Daborn P. J., Yen J. L., Ffrench-Constant R. H., and Schlotterer C. 2004. World-wide survey of an accord insertion and its association with DDT resistance in *Drosophila melanogaster*. *Mol Ecol*, 13(8): 2491–504.

Choudhary M., and Singh R. S. 1987. A comprehensive study of genic variation in natural populations of drosophila melanogaster. iii. variations in genetic structure and their causes between drosophila melanogaster and its sibling species drosophila simulans. *Genetics*, 117(4): 697–710.

Chung H., Bogwitz M. R., McCart C., Andrianopoulos A., Ffrench-Constant R. H., Batterham P., and Daborn P. J. 2007. *Cis*-regulatory elements in the Accord retrotransposon result in tissue-specific expression of the *Drosophila melanogaster* insecticide resistance gene *cyp6g1*. *Genetics*, 175(3): 1071–1077.

Clark A. G., Eisen M. B., Smith D. R., Bergman C. M., Oliver B., Markow T. A., Kaufman T. C., Kellis M., Gelbart W., Iyer V. N., *et al.* 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167): 203–18.

Daborn P. J., Yen J. L., Bogwitz M. R., Goff G. L., Feil E., Jeffers S., Tijet N., Perry T., Heckel D., Batterham P., *et al.* 2002. A single p450 allele associated with insecticide resistance in *Drosophila*. *Science*, 297(5590): 2253–2256.

De Gregorio E., Spellman P. T., Rubin G. M., and Lemaitre B. 2001. Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. *Proc Natl Acad Sci U S A*, 98(22): 12590–12595.

Dean M. D., and Ballard J. W. O. 2004. Linking phylogenetics with population genetics to reconstruct the geographic origin of a species. *Mol Phylogenet Evol*, 32(3): 998–1009.

Dworkin I., and Jones C. D. 2009. Genetic changes accompanying the evolution of host specialization in *Drosophila* sechellia. *Genetics*, 181(2): 721–36.

Ekengren S., and Hultmark D. 2001. A family of turandot-related genes in the humoral stress response of *Drosophila*. *Biochem Biophys Res Commun*, 284(4): 998–1003.

Ekengren S., Tryselius Y., Dushay M. S., Liu G., Steiner H., and Hultmark D. 2001. A humoral stress response in *Drosophila*. *Curr Biol*, 11(18): 1479.

Enayati A. A., Ranson H., and Hemingway J. 2005. Insect glutathione transferases and insecticide resistance. *Insect Mol Biol*, 14(1): 3–8.

Feyereisen R. 1999. Insect p450 enzymes. *Annu Rev Entomol*, 44: 507–33. Insect P450 enzymes;0066-4170 (Print) Comparative Study Journal Article Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. Review.

Ffrench-Constant R. H. 1994. The molecular and population genetics of cyclodiene insecticide resistance. *Insect Biochem Mol Biol*, 24(4): 335–345.

Gilad Y., and Borevitz J. 2006. Using DNA microarrays to study natural variation. *Curr Opin Genet Dev*, 16(6): 553–8.

Goto A., Yano T., Terashima J., Iwashita S., Oshima Y., and Kurata S. 2010. Cooperative regulation of the induction of the novel antibacterial Listericin by peptidoglycan recognition protein LE and the JAK-STAT pathway. *J Biol Chem*, 285(21): 15731–15738.

Gupta S. C., Siddique H. R., Mathur N., Mishra R. K., Mitra K., Saxena D. K., and Chowdhuri D. K. 2007. Adverse effect of organophosphate compounds, dichlorvos and chlorpyrifos in the reproductive tissues of transgenic *Drosophila melanogaster*: 70kda heat shock protein as a marker of cellular damage. *Toxicology*, 238(1): 1–14.

Hamblin M. T., and Veuille M. 1999. Population structure among african and derived populations of *Drosophila* simulans: evidence for ancient subdivision and recent admixture. *Genetics*, 153 (1): 305–17. Sep;Population structure among African and derived populations of Drosophila simulans: evidence for ancient subdivision and recent admixture;0016-6731 (Print) Comparative Study Journal Article Research Support, Non-U.S. Gov't.

Hey J., and Kliman R. M. 1993. Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. *Mol Biol Evol*, 10(4): 804–22.

Hutter S., Saminadin-Peter S. S., Stephan W., and Parsch J. 2008. Gene expression variation in african and european populations of *Drosophila melanogaster*. *Genome Biol*, 9(1): R12.

Irvin S. D., Wetterstrand K. A., Hutter C. M., and Aquadro C. F. 1998. Genetic variation and differentiation at microsatellite loci in drosophila simulans. evidence for founder effects in new world populations. *Genetics*, 150(2): 777–790.

Kliman R. M., Andolfatto P., Coyne J. A., Depaulis F., Kreitman M., Berry A. J., McCarter J., Wakeley J., and Hey J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics*, 156(4): 1913–31.

Kopp A., Frank A., and Fu J. 2006. Historical biogeography of *Drosophila simulans* based on y-chromosomal sequences. *Mol Phylogenet Evol*, 38(2): 355–62.

Lachaise D., Capy P., Cariou M. L., Joly D., Lemeunier F., and David J. R. 2004. Nine relatives from one African ancestor: population biology and evolution of the *Drosophila melanogaster* subgroup species, Singh R. S., and Uyenoyama M. K. (ed), *The evolution of population biology*. Cambridge University Press, 315–343.

Lachaise D., Cariou M., David J., Lemeunier F., Tsacas L., and Ashburner M. 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup, Hecht N. K., Wallace B., and Prance G. T. (ed), *Evolutionary biology*. Plenum Pub. Co, 22: 159–225.

Le Goff G., Boundy S., Daborn P. J., Yen J. L., Sofer L., Lind R., Sabourault C., Madi-Ravazzi L., and Ffrench-Constant R. H. 2003. Microarray analysis of cytochrome P450 mediated insecticide resistance in *Drosophila*. *Insect Biochem Mol Biol*, 33(7): 701–8.

Lemaitre B., and Hoffmann J. 2007. The host defense of *Drosophila melanogaster*. *Annu Rev Immunol*, 25: 697–743.

Low W. Y., Ng H. L., Morton C. J., Parker M. W., Batterham P., and Robin C. 2007. Molecular evolution of glutathione S-transferases in the genus *Drosophila*. *Genetics*, 177(3): 1363–1375.

Marioni J. C., Mason C. E., Mane S. M., Stephens M., and Gilad Y. 2008. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9): 1509–17.

McCart C., and Ffrench-Constant R. H. 2008. Dissecting the insecticide-resistance associated cytochrome P450 gene *Cyp6g1*. *Pest Manag Sci*, 64(6): 639–645.

McDermott S. R., and Kliman R. M. 2008. Estimation of isolation times of the island species in the *Drosophila simulans* complex from multilocus DNA sequence data. *PLoS ONE*, 3(6): e2442.

Meiklejohn C. D., Parsch J., Ranz J. M., and Hartl D. L. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. *Proc Natl Acad Sci U S A*, 100(17): 9894–9.

Muller L., Hutter S., Stamboliyska R., Saminadin-Peter S. S., Stephan W., and Parsch J. 2011. Population transcriptomics of *Drosophila melanogaster* females. *BMC Genomics*, 12(1): 81.

Pavey S. A., Collin H., Nosil P., and Rogers S. M. 2010. The role of gene expression in ecological speciation. *Ann N Y Acad Sci*, 1206(1): 110–29.

Robinson M. D., McCarthy D. J., and Smyth G. K. 2010. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140.

Salzman J., Jiang H., and Wong W. 2011. Statistical modeling of RNA-Seq data. *Statistical Science*, 26(1): 62–83.

Schlenke T. A., and Begun D. J. 2004. Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. *Proc Natl Acad Sci U S A*, 101(6): 1626–31.

Schmidt J. M., Good R. T., Appleton B., Sherrard J., Raymant G. C., Bogwitz M. R., Martin J., Daborn P. J., Goddard M. E., Batterham P., *et al.* 2010. Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genet*, 6(6): e1000998.

Schöfl G., and Schlötterer C. 2006. Microsatellite variation and differentiation in african and non-african populations of *Drosophila simulans*. *Mol Ecol*, 15(13): 3895–905.

Sharma A., Mishra M., Ram K. R., Kumar R., Abdin M. Z., and Chowdhuri D. K. 2011. Transcriptome analysis provides insights for understanding the adverse effects of endosulfan in *Drosophila melanogaster*. *Chemosphere*, 82(3): 370–376.

Sheehan D., Meade G., Foley V. M., and Dowd C. A. 2001. Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. *Biochem J*, 360(Pt 1): 1–16.

Storey J. D., Madeoy J., Strout J. L., Wurfel M., Ronald J., and Akey J. M. 2007. Gene-expression variation within and among human populations. *Am J Hum Genet*, 80(3): 502–9.

Tiwari A. K., Pragya P., Ram K. R., and Chowdhuri D. K. 2011. Environmental chemical mediated male reproductive toxicity: *Drosophila melanogaster* as an alternate animal model. *Theriogenology*, 76(2): 197–216.

Townsend J. P., Cavalieri D., and Hartl D. L. 2003. Population genetic variation in genome-wide gene expression. *Mol Biol Evol*, 20(6): 955–63.

Wu D.-D., Irwin D. M., and Zhang Y.-P. 2011. Correlated evolution among six gene families in *Drosophila* revealed by parallel change of gene numbers. *Genome Biol Evol*.

Wurmser F., Ogereau D., Mary-Huard T., Loriod B., Joly D., and Montchamp-Moreau C. 2011. Population transcriptomics: insights from *Drosophila simulans*, *Drosophila sechellia* and their hybrids. *Genetica*, 139(4): 465–477.

Table 1: Gene ontology terms for genes overexpressed in France compared to Mayotte

N	X	P-value	GO ID	GO term
14	148	9.81E-16	GO:0009055	electron carrier activity
13	117	1.39E-15	GO:0004497	monooxygenase activity
12	138	3.84E-13	GO:0020037	heme binding
12	139	4.19E-13	GO:0046906	tetrapyrrole binding
13	199	1.45E-12	GO:0005506	iron ion binding
8	38	2.74E-12	GO:0004364	glutathione transferase activity
18	633	5.03E-11	GO:0016491	oxidoreductase activity
9	85	7.21E-11	GO:0005792	microsome
9	85	7.21E-11	GO:0042598	vesicular fraction
8	63	2.00E-10	GO:0016765	transferase activity (alkyl or aryl groups)
9	98	2.64E-10	GO:0005624	membrane fraction
9	101	3.47E-10	GO:0005626	insoluble fraction
9	102	3.80E-10	GO:0000267	cell fraction
14	474	6.85E-09	GO:0055114	oxidation reduction
35	4053	6.54E-07	GO:0003824	catalytic activity

With N the number of genes with the term in the query; X the number of genes with the term in the genome; P-value of the significance of the overrepresentation of the term in query compared to genome, processed with FuncAssociate (Berriz *et al.*, 2003); GO ID and GO term, respectively the identifier and the corresponding term of Gene Ontology

Table 2: Gene ontology terms for genes overexpressed in axenic France compared with apple France

N	X	P	GO ID	GO term
8	75	2.65E-08	GO:0032504	multicellular organism reproduction
8	77	3.27E-08	GO:0000003	reproduction
10	144	2.90E-08	GO:0005615	extracellular space
6	85	1.81E-05	GO:0005792	microsome
6	85	1.81E-05	GO:0042598	vesicular fraction
11	223	1.91E-07	GO:0044421	extracellular region part
7	143	3.86E-05	GO:0020037	heme binding
7	144	4.03E-05	GO:0046906	tetrapyrrole binding

With N the number of genes with the term in the query; X the number of genes with the term in the genome; P-value of the significance of the overrepresentation of the term in query compared to genome, processed with FuncAssociate (Berriz *et al.*, 2003); GO ID and GO term, respectively the identifier and the corresponding term of Gene Ontology

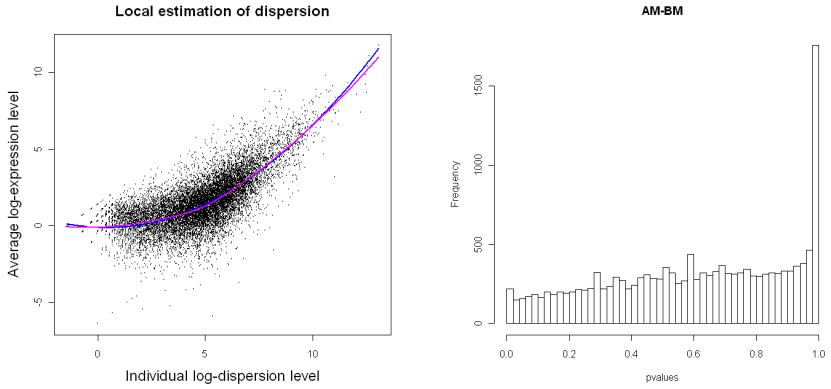


Figure 1: **Left:** Dispersion parameter estimates $\widehat{\phi}_g$ as a function of the mean expression $\widehat{\lambda}_g$ (log-scale). Each point corresponds to a gene. The purple and blue curves represent the Loess and Quadratic Regression estimates, respectively. **Right:** Histogram of the p-values for the AM vs BM comparison.

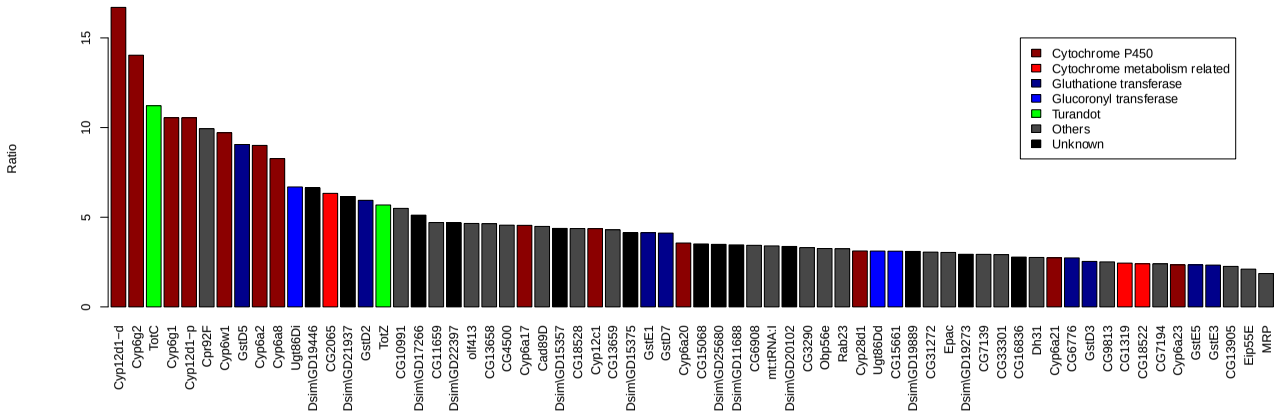


Figure 2: Barplot of the ratio of expression of France over Mayotte for genes overexpressed in France.

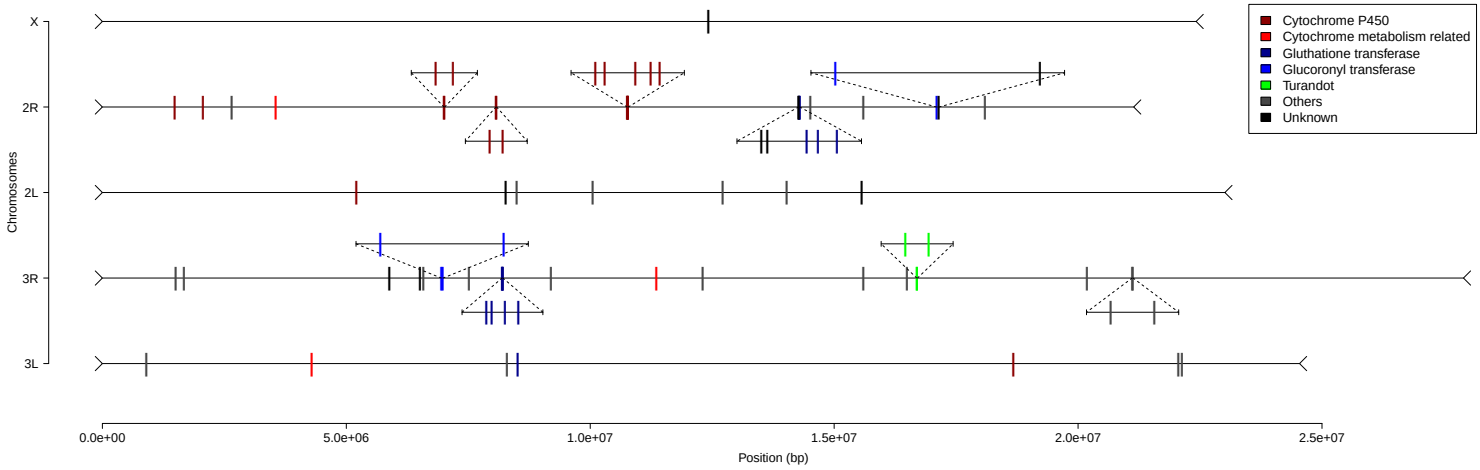


Figure 3: Chromosome location of genes overexpressed in France compared to Mayotte

Role of expression differentiation in adaptation and speciation: comparative analysis of *D. simulans* and *D. sechellia*.

François Wurmser[1], David Ogereau[1], Tristan Mary-Huard[2], Dominique Joly[1], Catherine Montchamp-Moreau[1]

[1]*Laboratory Evolution, Genomes and Speciation, CNRS UPR 9034 Avenue de la Terrasse 91198 Gif-sur-Yvette, France, Université Paris-Sud 11, 91405 Orsay*
[2]*Statistics and Genomes Team, UMR 518 INRA / AgroParisTech, 16 rue Claude Bernard 75231 Paris*

Introduction

Genotype → Gene expression → Proteome → Phenotype

What is the role of the divergence of gene regulation on phenotype ?

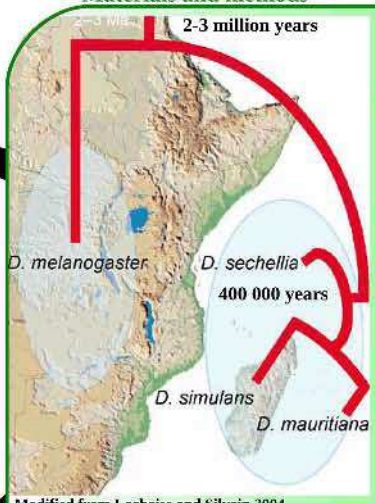
What genes and pathways can be involved in populations and species differentiation at the transcriptome level ?

Comparative transcriptome analysis

Background

Sequence divergence has been widely studied in evolution, especially with the knowledge brought by genome sequences. However gene expression differences are not as well-known. We examined these differences between natural populations and species, as they may explain an important part of the phenotypic variation. Our goal is to better understand how adaptation and/or drift can lead to changes in gene regulation and ultimately to incompatibility between hybrids, and therefore speciation.

Materials and methods



Modified from Lachaise and Silvain 2004

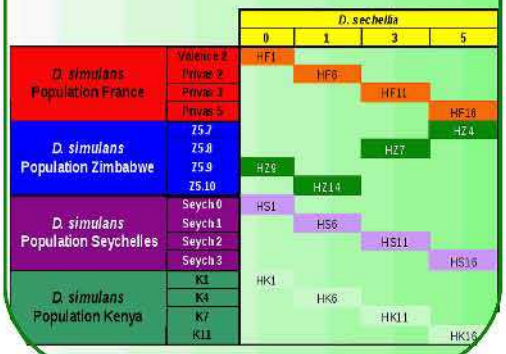
Microarrays 4398 genes

mRNA whole-body adult males

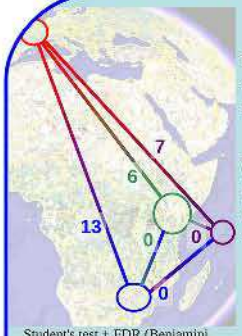
4 isofemale lines / pop = 4 biological replicates / population

16 hybrid F1 lines

D. simulans	*D. sechellia*
Cosmopolitan (invasive)	Endemic
Generalist	Specialist (*Morinda citrifolia*)
4 populations	1 population



D. simulans: population differentiation



Student's test + FDR (Benjamini and Hochberg 1995)

Number of genes differentially expressed between the four *D. simulans* populations.

These data suggest a differentiation of the French population comparing with the African ones.

Adaptation, drift ?

Overrepresentation of Ontology terms

Genes overexpressed in the French population compared with African populations of *D. simulans*.

N	X	P-adj	
5	82	<0.001	MF : electron carrier activity
4	46	<0.001	MF : monooxygenase activity
5	132	0.002	BP : electron transport/electron transfer

5 cytochrome P450 out of 12 genes

MF: **Molecular Function**
BP: **Biological Process**
N: number of genes with this attribute in the list of DE genes
X: number of genes with this attribute on the array
P-adj: multiple test corrected p-value

FuncAssociate : Berriz *et al.* 2003

D. simulans / *D. sechellia* comparison

Student's test + FDR (Benjamini and Hochberg 1995)

In the comparison with *D. sechellia*, the French population of *D. simulans* shows much larger differentiation than the African populations.

304 genes are common to all comparisons, and are thus potentially involved in the divergence.

Overrepresentation of Ontology terms

Genes overexpressed in *D. simulans* compared with *D. sechellia*

MF : electron carrier activity CC : vesicular fraction CC : microsome	Cytochromes P450 (~10 genes)
BP : lipid metabolism BP : hormone catabolism	Juvenile Hormone regulation (3 genes)

Cytochromes are also involved in JH regulation, suggesting a major role of this hormone in the differentiation between species and populations.

BP : Biological Process
CC : Cellular Component
MF : Molecular Function

FuncAssociate : Berriz *et al.* 2003

Hybrids

● : hybrids *D. simulans* / *D. sechellia*

304 genes differentially expressed between the two species…

But…

…only very few genes differentially expressed between interspecific hybrids and parents

This suggest stronger constraints on *trans*-regulation than *cis*-regulation

Conclusion

Originality of the study: examination of both intra- and inter-specific divergence, along with hybrids.

Population differentiation within *D. simulans*

Consistent with published microsatellite structuration (Schöfl and Schlötterer 2006)

We observed differentiation between species and populations, especially in genes involved in JH regulation (*Jheh1* and *Jheh3*, *Dat*, *Cytochrome P450*). This differentiation occurs between France and Africa in *D. simulans*, but also between *D. simulans* and *D. sechellia*, including the *D. simulans* population collected in the Seychelles Islands.
Our results also suggest that the pleiotropic role of transcription factors leads to an easier differentiation of cis-regulation in adaptation.

Role of expression differentiation in adaptation and speciation: comparative analysis of *D. simulans* and *D. sechellia*

François Wurmser[1], David Ogereau[1], Tristan Mary-Huard[2], Dominique Joly[1], Catherine Montchamp-Moreau[1]

[1]*Laboratory Evolution, Genomes and Speciation, CNRS UPR 9034 Avenue de la Terrasse 91198 Gif-sur-Yvette, France, Université Paris-Sud 11, 91405 Orsay*
[2]*Statistics and Genomes Team, UMR 518 INRA / AgroParisTech, 16 rue Claude Bernard 75231 Paris*
francois.wurmser@legs.cnrs-gif.fr

Background

Genome → Transcriptome → Proteome → Phenotype

What genes and pathways can be involved in populations and species differentiation at the transcriptome level?

Comparative transcriptome analysis

Patterns of DNA sequence variation and divergence have been widely studied in evolution, especially with the data brought by whole genomes. However gene expression differences are not as well-known though they may explain an important part of the phenotypic variation. We examined these differences between natural populations and species. Our goal was to better understand how selection and/or drift can lead to changes in gene regulation and ultimately to incompatibility between hybrids, and therefore speciation.

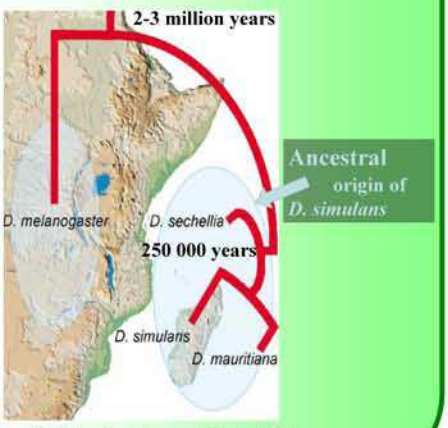
Materials and methods

D. simulans	*D. sechellia*
Cosmopolitan (invasive)	Endemic
Generalist	Specialist (*Morinda citrifolia*)
4 populations: France, Zimbabwe, Kenya, Seychelles	1 population

mRNA
whole-body adult males

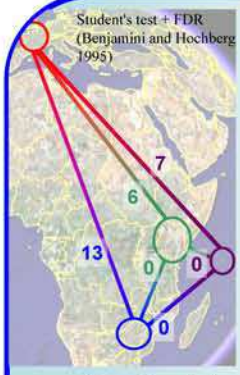
Microarrays
4398 genes

4 isofemale lines / pop
=
4 biological replicates / population



Modified from Lachaise and Silvain 2004

Geographic differentiation of *D. simulans*



Number of genes differentially expressed between the four *D. simulans* populations.

These data suggest a differentiation of the French population comparing with the African ones.

→ **Consistent with known ancestral biogeography of the species**

Overrepresentation of Ontology terms

Genes overexpressed in the French population compared with African populations of *D. simulans*.

→ **4 cytochrome P450 out of 12 genes**

Detoxification?
Regulation of hormones?

→ **Adaptation, drift?**

FuncAssociate : Berriz *et al.* 2003

D. simulans / *D. sechellia* divergence

In the comparison with *D. sechellia*, the French population of *D. simulans* shows much larger differentiation than the African populations.

304 genes are common to all comparisons, and are thus potentially involved in the species divergence.

Overrepresentation of Ontology terms

Genes overexpressed in *D. simulans* compared with *D. sechellia*

+ **Cytochromes P450 (~10 genes)**

→ relaxation of selection on a variety of detoxification functions due to the specialization of *D. sechellia* on *M. citrifolia*.

Dworkin and Jones 2009

+ **Juvenile Hormone regulation (3 genes)**

→ regulation of accessory gland proteins, influence on reproductive behavior,…

Decanalization in *D. simulans*

		D. simulans France	*D. simulans* Zimbabwe	*D. simulans* Seychelles	*D. simulans* Kenya	*D. sechellia* Seychelles
Intrapopulation variance		0.1109	0.1022	0.0760	0.0572	0.0714
D. simulans	France		NS (P = 0.08)	*	*	*
	Zimbabwe			*	*	*
	Seychelles				*	*
	Kenya					*

NS : not significant; * : significant, P < 0.0001

The mean variance of expression is higher for derived populations than for populations close to the original species range. This could be due to a phenomenon of decanalization:
1- old populations accumulate cryptic genetic variation in the ancestral area.
2- during the invasion of a new area, the cryptic variation is suddenly expressed leading to increased expression variance.

Conclusion

Population differentiation within *D. simulans*

→ Coherent with what is known of the biogeography of the species

We observed differentiation between species and populations, especially in genes in cytochrome P450. It is likely divergence on the selection on these genes leaded to the differences observed in expression. Other interesting differences were related with divergences of regulation of Juvenile hormone and Dopamine in males, likely part of the reproductive isolation. We also revealed a decanalization phenomenon in the French population of *D. simulans*.

An analysis of expression variance between populations of *D. simulans* and *D. Sechellia*

François Wurmser[1], David Ogereau[1], Tristan Mary-Huard[2], Dominique Joly[1], Catherine Montchamp-Moreau[1]

[1]*Laboratory Evolution, Genomes and Speciation, CNRS UPR 9034 Avenue de la Terrasse 91198 Gif-sur-Yvette, France, Université Paris-Sud 11, 91405 Orsay*
[2]*Statistics and Genomes Team, UMR 518 INRA / AgroParisTech, 16 rue Claude Bernard 75231 Paris*
francois.wurmser@legs.cnrs-gif.fr

Follow me

Introduction

Genotype → Gene expression → Proteome → Phenotype

It is well known that sequences show more polymorphism in ancestral compared to derived populations, due to founder effects.

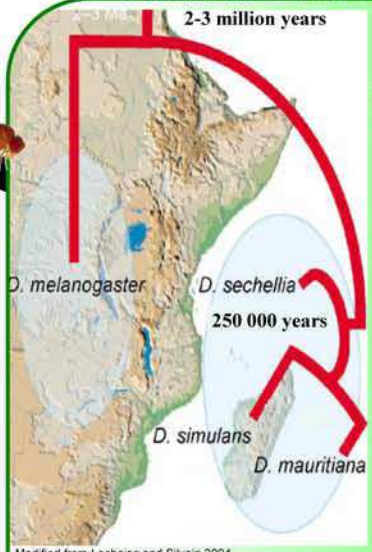
What about expression variance ?

Comparative transcriptome analysis

Background

Sequence polymorphism and divergence has been widely studied in evolution, within and between species. However gene expression differences are not as well-known. Moreover no study examined both differential expression, and global expression variance within populations. We seek a better understanding of the overall mechanisms ruling expression evolution during the extension of a species range. We also examined specific gene differences between closely related species of *Drosophila*.

Materials and methods



Modified from Lachaise and Silvain 2004

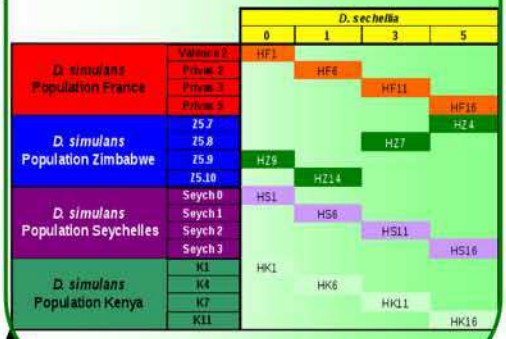
Microarrays 4398 genes

mRNA whole-body adult males

4 isofemale lines / pop = 4 biological replicates / population

16 hybrid F1 lines

D. simulans	*D. sechellia*
Cosmopolitan (invasive)	Endemic
Generalist	Specialist (*Morinda citrifolia*)
4 populations	1 population



Population variance differences

Global scale variances of populations:

Derived populations

(p-val = 4.8^{e}-2)
Var(France) ≈ Var(Zimbabwe)

≫

(p-val = 2.94^{e}-69) (p-val = 3.75^{e}-200)

Var(Seychelles) > Var(Kenya)

Populations from the ancestral range

Contrasting with sequence variation[1], expression variance is significantly higher in derived populations of* D. simulans *than in populations from the ancestral area.

[1]Schöfl and Schlötterer 2006
[1]Baudry et al.2006

Population variance test

Binomial test: principle

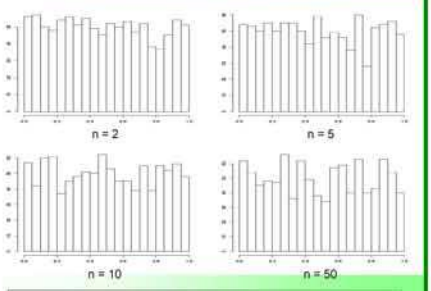
$R_g = \frac{\sigma^2_{g,A}}{\sigma^2_{g,B}}$ with R_g the variance ratio between population A and B

We test $H_0 = (p_{AB} = 1/2)$ with p_{AB} the proportion of genes for which $\sigma^2_{g,A} > \sigma^2_{g,B}$

$N \sim B(G, p_{AB})$ with N the number of genes for which $R_g > 1$ and G the total number of genes

Simulation:

Influence of the number of lines on the test:



Error rate is constant around 5% regardless of the number of lines used.

D. simulans: canalization ?

The variance of expression increases with the distance from the ancestral area of the species. This could be due to the phenomenon of decanalization:

Old populations accumulate cryptic genetic variation in the ancestral area

During the invasion of a new area, the cryptic variation is suddenly expressed leading to increased expression variance

Summary of differential analysis

***304 genes are consistently differentially expressed between* D. simulans *and* D. sechellia.**

Expression differentiation between populations of* D. simulans *are consistent with known structuration[1].

Hybrid expression is only slightly disturbed with an over-representation of X-linked genes.

D. simulans / *D. sechellia* comparison

Notable genes differentially expressed between *D. simulans* and *D. sechellia*

+ **Genes of the cytochrome P450 family underexpressed in *D. sechellia***

Relaxation of selection on a variety of detoxification functions due to the specificity of the interaction with the host[2].

[2]Dworkin and Jones 2009

+ **Genes involved in hormone regulation (Juvenile Hormone and Dopamine)**

Pleiotropic role in reproduction, reproductive behavior, regulation of accessory gland proteins, etc.

[3]Wolfner et al. 1997
[3]Wilson et al. 2003

The END

Adaptation of expression in *Drosophila simulans*: invasion of an anthropised world

Looking for a PostDoc

François Wurmser[1], Tristan Mary-Huard[2], Jean-Jacques Daudin[2], Dominique Joly[1], Catherine Montchamp-Moreau[1]
francois.wurmser@legs.cnrs-gif.fr

Background

Invasion

New environmental constraints
Population bottleneck

Sequence adaptation

Expression adaptation

What genes, which pathways ?

Drosophila simulans

- Originates from eastern Africa
- **Recent** invasion of the world
- Cosmopolitan generalist

Methods

in

X

or

and

and

RNA was extracted
from a 100 of 5-days old males
from a 100 different wild females
for each of the 4 conditions

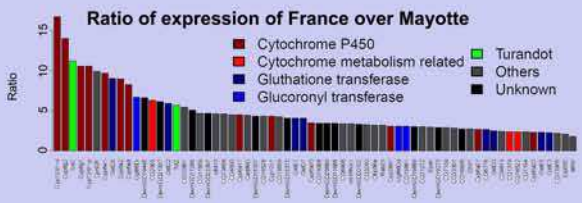
3'-DGE on an Illumina Genome Analyser

2 populations x 2 resources = 4 conditions

Results

POPULATION COMPARISON

Gene families associated with the invasion :



- **Cytochrome P450**
  - **85 functional genes** in *D. simulans*
  - **Detoxification of xenobiotics**
- **Gluthatione transferase**
  - **Detoxification of xenobiotics**
  - **Insect specific GSTs (Delta and Epsilon familly)**
    - **Linked with local adaptation**

Low et al 2007, Genetics

Cyp6g1, or the true story of a striking co-evolution

- **Overexpressed in derived populations of *Drosophila***
- **Correlated with pesticide resistance**

D. melanogaster

Numerous allelic variants
Schmidt *et al.* 2010

Transposable elements, duplication

D. simulans

No insertion
Population from Africa
(Schlenke and Begun 2004, present study)

***Doc* insertion**
Population from California
(Schlenke and Begun 2004)

NEW INSERTION
Transposable element Juan
Population from France
(present study)

Cyp6g1

RESOURCE SHIFT

Relaxation of the immune system induction

12 / 20 anti-microbial peptides overexpressed on banana

Induction: **2.3 times to 47 times**

Apple medium: breakdown of expression of 10 reproduction proteins

Accessory gland proteins, Seminal Fluid Proteins, protease inhibitors,...

Reproductive cost of apple medium

WHY ?

Pesticides ?
pH ?
Sugar contents ?

To be continued

[1]*Laboratory Evolution, Genomes and Speciation, CNRS UPR 9034 Avenue de la Terrasse 91198 Gif-sur-Yvette, France, Université Paris-Sud 11, 91405 Orsay*
[2]*Statistics and Genomes Team, UMR 518 INRA / AgroParisTech, 16 rue Claude Bernard 75231 Paris*

Gènes différentiellement exprimés entre *D. simulans* et *D. sechellia* ($1^{ère}$ étude, puces à ADN)

sech > sim	sech < sim
FBgn0023489	FBgn0039342
FBgn0033132	FBgn0027375
FBgn0034088	FBgn0003380
FBgn0023517	FBgn0039537
FBgn0027601	FBgn0010786
FBgn0027598	FBgn0036022
FBgn0027597	FBgn0053302
FBgn0027592	FBgn0027599
FBgn0037818	FBgn0038052
FBgn0010246	FBgn0050160
FBgn0031327	FBgn0003435
FBgn0035515	FBgn0010053
FBgn0038791	FBgn0038194
FBgn0001316	FBgn0011576
FBgn0051150	FBgn0010516
FBgn0037688	FBgn0000079
FBgn0015799	FBgn0037643
FBgn0031824	FBgn0039114
FBgn0028274	FBgn0026314
FBgn0028491	FBgn0022787
FBgn0034390	FBgn0019643
FBgn0027579	FBgn0033065
FBgn0039560	FBgn0031897
FBgn0036764	FBgn0028536
FBgn0027565	FBgn0032122
FBgn0039635	FBgn0040350
FBgn0003308	FBgn0035871
FBgn0032012	FBgn0029161
FBgn0035035	FBgn0052529
FBgn0027054	FBgn0031461
FBgn0033919	FBgn0034335
FBgn0037215	FBgn0023511
FBgn0027844	FBgn0000422
FBgn0033988	FBgn0032495
FBgn0035107	FBgn0023525
FBgn0031294	FBgn0034883
FBgn0031995	FBgn0037975
FBgn0015573	FBgn0035975
FBgn0000163	FBgn0031401
FBgn0019948	FBgn0035091
FBgn0026144	FBgn0037721
FBgn0035985	FBgn0029990
FBgn0052026	FBgn0035980
FBgn0036671	FBgn0031228

sech > sim	sech < sim
FBgn0034957	FBgn0037090
FBgn0003074	FBgn0031925
FBgn0037939	FBgn0036623
FBgn0019929	FBgn0023479
FBgn0034997	FBgn0051236
FBgn0020385	FBgn0032923
FBgn0030087	FBgn0030082
FBgn0030581	FBgn0032586
FBgn0031423	FBgn0029687
FBgn0031800	FBgn0044817
FBgn0010851	FBgn0026567
FBgn0036325	FBgn0000241
FBgn0082585	FBgn0063492
FBgn0085433	FBgn0030468
FBgn0032618	FBgn0032871
FBgn0039223	FBgn0020255
FBgn0035315	FBgn0036316
FBgn0023520	FBgn0033872
FBgn0038046	FBgn0042174
FBgn0039667	FBgn0031940
FBgn0036834	FBgn0029688
FBgn0028916	FBgn0026160
FBgn0035099	FBgn0004623
FBgn0036968	FBgn0036298
FBgn0037252	FBgn0000253
FBgn0032484	FBgn0034654
FBgn0031240	FBgn0011642
FBgn0036710	FBgn0014906
FBgn0019938	FBgn0039304
FBgn0030744	FBgn0021825
FBgn0028956	FBgn0004169
FBgn0037133	FBgn0024947
FBgn0031347	FBgn0001205
FBgn0036775	FBgn0015582
FBgn0026721	FBgn0029092
FBgn0023077	FBgn0035878
FBgn0051163	FBgn0011606
FBgn0034321	FBgn0030805
FBgn0030245	FBgn0027509
FBgn0020270	FBgn0039141
FBgn0003423	FBgn0037537
FBgn0004049	FBgn0034432
FBgn0005683	FBgn0033337
FBgn0035838	FBgn0002719
FBgn0039638	FBgn0027498
FBgn0034878	FBgn0036332
FBgn0034617	FBgn0037472
FBgn0026319	FBgn0031820

sech > sim	sech < sim
FBgn0030628	FBgn0036302
FBgn0014366	FBgn0020299
FBgn0031696	FBgn0028373
FBgn0033935	FBgn0039698
FBgn0025825	FBgn0050489
FBgn0034961	FBgn0026403
FBgn0000173	FBgn0034406
FBgn0032467	FBgn0036842
FBgn0032450	FBgn0015930
FBgn0030594	FBgn0032597
FBgn0028380	FBgn0030768
FBgn0033427	FBgn0034971
FBgn0039733	FBgn0037447
FBgn0046776	FBgn0023174
FBgn0031314	FBgn0032693
FBgn0050007	FBgn0053174
FBgn0034590	FBgn0028424
FBgn0039488	FBgn0033740
FBgn0036800	FBgn0033853
FBgn0028519	FBgn0015010
FBgn0037556	FBgn0028394
FBgn0026722	FBgn0020506
FBgn0030630	FBgn0040723
FBgn0000629	FBgn0039476
FBgn0019637	FBgn0037684
FBgn0034098	FBgn0036024
FBgn0037116	FBgn0036996
FBgn0034009	FBgn0250815
FBgn0038306	FBgn0003638
FBgn0034618	FBgn0034599
FBgn0030026	FBgn0031362
FBgn0039252	FBgn0033980
FBgn0030912	FBgn0035023
FBgn0031664	FBgn0030478
FBgn0034689	FBgn0030073
FBgn0035850	FBgn0001257
FBgn0038126	FBgn0032147
FBgn0050020	FBgn0026370
FBgn0050467	
FBgn0053554	
FBgn0030013	
FBgn0032290	
FBgn0030456	
FBgn0039182	
FBgn0004901	
FBgn0023522	
FBgn0033538	
FBgn0023216	

sech > sim	sech < sim
FBgn0010416	
FBgn0034270	
FBgn0036053	
FBgn0030963	
FBgn0029118	
FBgn0015907	
FBgn0037182	
FBgn0042106	
FBgn0028688	
FBgn0036314	
FBgn0035464	
FBgn0035238	
FBgn0035199	
FBgn0032798	
FBgn0031249	
FBgn0038244	
FBgn0027094	
FBgn0028343	
FBgn0030674	
FBgn0037391	
FBgn0250850	
FBgn0037968	
FBgn0033156	
FBgn0015763	
FBgn0037619	
FBgn0023000	
FBgn0030648	
FBgn0086365	
FBgn0036640	
FBgn0031201	
FBgn0024329	
FBgn0000487	
FBgn0034909	

Gènes différentiellement exprimés entre pomme et axénique (population de Gotheron)

AG > PG				
FBgn0044812	FBgn0031910	FBgn0033696	FBgn0032726	FBgn0032868
FBgn0032660	FBgn0028986	FBgn0029898	FBgn0085762	FBgn0034846
FBgn0031701	FBgn0002565	FBgn0037764	FBgn0038115	FBgn0051446
FBgn0038394	FBgn0259958	FBgn0040732	FBgn0022355	FBgn0053530
FBgn0261336	FBgn0033980	FBgn0015570	FBgn0039151	FBgn0053346
FBgn0028396	FBgn0250847	FBgn0041337	FBgn0032494	FBgn0001089
FBgn0035916	FBgn0013301	FBgn0013772	FBgn0051205	FBgn0025454
FBgn0003138	FBgn0037389	FBgn0051779	FBgn0039342	FBgn0035922
FBgn0032836	FBgn0037936	FBgn0259961	FBgn0039192	FBgn0261059
FBgn0003863	FBgn0037934	FBgn0034595	FBgn0043825	FBgn0014469
FBgn0035693	FBgn0029765	FBgn0028526	FBgn0051872	FBgn0035770
FBgn0083121	FBgn0039325	FBgn0260932	FBgn0039010	FBgn0039761
FBgn0053503	FBgn0037611	FBgn0043533	FBgn0034153	FBgn0050488
FBgn0030105	FBgn0033216	FBgn0038147	FBgn0020399	**ID non mel**
FBgn0054002	FBgn0034329	FBgn0025683	FBgn0026174	FBgn0054035
FBgn0259146	FBgn0051418	FBgn0039800	FBgn0032782	FBgn0194847
FBgn0038956	FBgn0053202	FBgn0085358	FBgn0031907	FBgn0197125
FBgn0028987	FBgn0085256	FBgn0035290	FBgn0037684	FBgn0044812

AG < PG				
FBgn0000047	FBgn0039209	FBgn0010040	FBgn0031643	FBgn0187025
FBgn0031942	FBgn0031940	FBgn0036997	FBgn0262656	FBgn0187126
FBgn0037167	FBgn0000043	FBgn0040251	FBgn0020269	FBgn0196090
FBgn0010041	FBgn0058002	FBgn0031805	FBgn0262684	FBgn0185240
FBgn0037974	FBgn0259736	FBgn0005563	FBgn0085447	FBgn0186355
FBgn0014454	FBgn0040733	FBgn0034538	FBgn0040237	FBgn0185128
FBgn0032699	FBgn0027660	FBgn0085285	FBgn0031129	FBgn0185720
FBgn0010222	FBgn0040837	FBgn0034756	**ID non mel**	FBgn0194162
FBgn0261560	FBgn0036044	FBgn0034706	FBgn0256344	FBgn0193735
FBgn0038175	FBgn0041588	FBgn0002534	FBgn0256558	FBgn0197173
FBgn0037975	FBgn0024361	FBgn0261625	FBgn0191455	FBgn0187616
FBgn0033188	FBgn0030482	FBgn0036449	FBgn0186055	FBgn0194398
FBgn0038467	FBgn0053192	FBgn0037275	FBgn0197102	FBgn0186634
FBgn0002573	FBgn0050029	FBgn0030482	FBgn0186748	FBgn0196157
FBgn0013988	FBgn0035817	FBgn0262579	FBgn0186167	FBgn0193036
FBgn0023129	FBgn0033830	FBgn0001092	FBgn0195445	FBgn0195177
FBgn0005633	FBgn0038819	FBgn0085503	FBgn0193803	FBgn0195723
FBgn0013984	FBgn0040992	FBgn0001301	FBgn0188249	FBgn0182766
FBgn0015040	FBgn0004244	FBgn0259979	FBgn0190857	FBgn0188459
FBgn0000052	FBgn0015919	FBgn0086378	FBgn0187743	FBgn0186745

Gènes différentiellement exprimés entre banane et axénique (population de Mayotte)

AM > BM			
FBgn0003358	FBgn0013687	FBgn0039778	FBgn0003358
FBgn0037166	FBgn0051789	**ID non mel**	FBgn0037166
FBgn0033592	FBgn0043791	FBgn0193858	FBgn0033592

AM < BM				
FBgn0004240	FBgn0038914	FBgn0034296	FBgn0033327	FBgn0262579
FBgn0014865	FBgn0010388	FBgn0035743	FBgn0032281	**ID non mel**
FBgn0038532	FBgn0039685	FBgn0032507	FBgn0013705	FBgn0036068
FBgn0012042	FBgn0041581	FBgn0043575	FBgn0036947	FBgn0194117
FBgn0041579	FBgn0027584	FBgn0034407	FBgn0067905	FBgn0194618
FBgn0036767	FBgn0000279	FBgn0032087	FBgn0053192	FBgn0190661
FBgn0033593	FBgn0029006	FBgn0034052	FBgn0034160	FBgn0186809
FBgn0034647	FBgn0010381	FBgn0000278	FBgn0036766	FBgn0004240
FBgn0043578	FBgn0010385	FBgn0038790	FBgn0013704	FBgn0014865
FBgn0002869	FBgn0030105	FBgn0030929	FBgn0030482	FBgn0038532

Gènes différentiellement exprimés entre population de Gotheron et de Mayotte

G > M				
FBgn0013772	FBgn0040735	FBgn0039189	FBgn0038819	FBgn0196965
FBgn0044812	FBgn0033696	FBgn0013696	FBgn0035868	FBgn0190777
FBgn0053503	FBgn0033204	FBgn0033978	FBgn0010039	FBgn0187043
FBgn0025454	FBgn0036806	FBgn0032048	FBgn0028519	FBgn0190945
FBgn0033065	FBgn0033981	FBgn0035176	FBgn0040251	FBgn0193803
FBgn0010038	FBgn0040256	FBgn0044809	FBgn0034605	FBgn0191376
FBgn0050489	FBgn0031689	FBgn0034711	FBgn0051272	FBgn0188828
FBgn0033980	FBgn0010043	FBgn0085421	FBgn0038347	FBgn0183428
FBgn0037389	FBgn0027532	FBgn0035904	FBgn0038439	FBgn0187025
FBgn0034335	FBgn0053301	FBgn0038143	FBgn0032456	FBgn0193352
FBgn0015714	FBgn0037153	FBgn0038731	FBgn0037364	FBgn0191582
FBgn0037936	FBgn0039319	FBgn0063495	FBgn0000566	FBgn0013772
FBgn0034471	FBgn0063497	FBgn0040733	**ID non mel**	FBgn0044812
FBgn0039315	FBgn0010041	FBgn0035529	FBgn0068665	FBgn0053503

G < M				
FBgn0020906	FBgn0010425	FBgn0039685	FBgn0034010	FBgn0196598
FBgn0039471	FBgn0035199	FBgn0039472	FBgn0052268	FBgn0188057
FBgn0038009	FBgn0029856	FBgn0030827	FBgn0262005	FBgn0183167
FBgn0051789	FBgn0054026	FBgn0047000	FBgn0031628	FBgn0189943
FBgn0036996	FBgn0039313	FBgn0051515	FBgn0052778	FBgn0186428
FBgn0013704	FBgn0033890	FBgn0035604	**ID non mel**	FBgn0186692
FBgn0036169	FBgn0036343	FBgn0016919	FBgn0190790	FBgn0187829
FBgn0013689	FBgn0035667	FBgn0036831	FBgn0193034	FBgn0020906

Résumé

Cette thèse a été consacrée à l'évolution du transcriptome de *Drosophila simulans*, et à son rôle dans l'adaptation et la spéciation. L'étude a comporté deux parties. La première utilisant des puces à ADN pour comparer les transcriptomes de populations de *D. simulans*, de son espèce jumelle *D. sechellia*, et de leurs hybrides. La seconde basée sur la quantification des transcrits par séquençage haut débit pour comparer une population de la zone d'origine (Afrique), et une population d'une zone récemment envahie (France métropolitaine). Ces analyses ont mis en évidence plusieurs groupes ou familles de gènes montrant des variations d'expression.

Un résultat majeur est l'implication prépondérante de la famille des cytochromes P450 dans l'adaptation. Cette superfamille composée de 85 gènes chez *D. simulans* est notamment importante pour la détoxification des xénobiotiques. L'expression de plusieurs gènes de cette famille est fortement réduite chez *D. sechellia*, probablement à cause de la spécialisation de cette espèce sur la plante *Morinda citrifolia* (plante toxique pour les autres drosophiles). On peut s'attendre alors à un relâchement des contraintes de sélection sur cette famille de gènes, dû à une forte réduction de la diversité des toxines auxquelles cette espèce est exposée.

Ces gènes sont également impliqués dans la différenciation entre les populations de la zone ancestrale de *D. simulans* et les populations dérivées. La zone ancestrale, en Afrique de l'est et dans les îles de l'Océan Indien occidental, est encore peu anthropisée. *A contrario*, la plupart des populations dérivées, comme ici notre population de la vallée du Rhône, sont exposées à des contraintes chimiques sous la forme de pesticides utilisés massivement sur les grandes cultures. Ces pesticides contraignent les populations dérivées à s'adapter, ce qui peut se réaliser par une augmentation de l'expression. Nous avons détecter une augmentation de l'expression de treize P450, dont un gène très bien connu pour ses fonctions de détoxification : *Cyp6g1*. Ce gène montre une augmentation d'expression corrélée à une résistance aux pesticides et à l'insertion d'éléments transposables en 5' ; ceci a été montré en détail chez *D. melanogaster*, et dans une moindre mesure chez *D. simulans*. Nous avons mis en évidence chez cette dernière espèce un nouvel évènement d'insertion.

Nos résultats montrent également que d'autres familles de gènes impliqués dans les détoxifications sont concernées par ces augmentations d'expression liées à l'anthropisation des milieux, notamment les Glutathion transfèrases (GST).

Nous avons également examiné la plasticité d'expression liée au changement de ressource, en élevant une partie de nos populations sur la ressource d'origine (fruits), et une autre partie sur le milieu axénique, milieu d'élevage standard de laboratoire (stérile). Les drosophiles élevées sur un milieu "naturel" montrent une forte activation du système immunitaire, et notamment une forte induction des gènes effecteurs de l'immunité innée, codant les peptides anti-microbiens. Cela est probablement dû à la présence de microorganismes sur ce milieu (ici, la banane). En conclusion, cette thèse a révélé des familles de gènes fortement impliquées dans les différenciations d'expression entre populations, espèces, et ressources, posant les jalons d'une meilleure compréhension des mécanismes d'adaptation du transcriptome.

Abstract

The topic of this thesis was the evolution of the transcriptome of *Drosophila simulans*, and its role in adaptation and speciation. First, we used microarrays to compare transcriptomes of populations of *D. simulans*, its sister species *D. sechellia* and their hybrids. Second, we used a RNA-seq like approach to quantify gene expression of a population from the ancestral range (Africa) on the one side, and a population from a recently invaded zone (France) on the other side. These analyses revealed several gene groups or families showing gene expression variations.

One main result is the major involvement of the cytochrome P450 gene family in adaptation. This superfamily is composed of 85 genes in *D. simulans*, and is notably important in detoxification of xenobiotics. The expression of several genes of this family is strongly reduced in *D. sechellia*, likely due to the specialization of this species on *Morinda citrifolia* (a plant toxic for any other drosophila). This specialization may strongly reduce the diversity of toxines this species is exposed to, thus relaxing selective constraints on this gene family.

These genes are also involved in the differentiation between populations of the ancestral range of *D. simulans* and derived populations. The ancestral range, in eastern Africa and in the occidental islands of the Indian Ocean, is not highly anthropized yet. Contrasting with this pattern, many derived populations (here our population from the Rhône valley) are exposed to chemical pressures due to the massive use of pesticides on agricultural zones. These pesticides force derived populations to adapt, which can happen via increased expression in genes such as the very famous example of *Cyp6g1*. This gene shows a strong increase in expression correlated with pesticide resistance as well as the insertion of transposable elements upstream of the gene ; this was described in details in *D. melanogaster*, and to some extent in *D. simulans*. We have shown a novel insertion event in the latter species.

Our results also reveal the involvement of other gene families in detoxification linked with anthropized environment via increase of gene expression, notably the Glutathione transferases (GST).

We also examined expression plasticity linked with resource change, raising half our flies on their natural food resource (fruits), and the other half on axenic medium (standard sterile laboratory medium). Drosophila raised on their natural medium show a strong activation of their immune system, and notably an induction of the effectors of teir innate immunity, the anti-microbial peptides. This observation can be explained by the presence of microorganisms on this medium (here, banana). To conclude, this thesis revealed gene families strongly involved in expression differentiation among populations, species and food resources, paving the way of a better understanding of mechanisms of adaptation of the transcriptome.