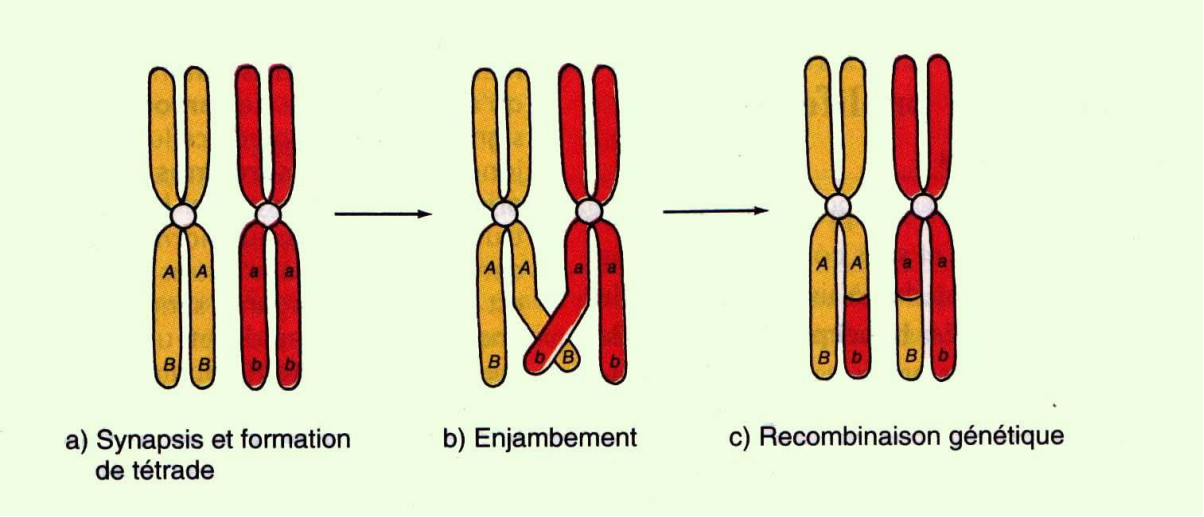
Rapport de stage

Construction de cartes de recombinaison chez la Cione

Travaux menés par Aurélie Fischer

Encadrée par Laurent Duret et Julien Joseph



4ème année Biosciences

Spécialité BiM Année 2020/2021

Tuteur INSA : M. Nicolas Parisot

Table des matières

I. Préambule

Le Laboratoire de Biométrie et Biologie Evolutive (LBBE) dans lequel s'est déroulé mon stage est rattaché à l'université Claude Bernard Lyon I, mais aussi au CNRS et à VetAgroSup. Il s'organise en plusieurs unités dans la région lyonnaise, notamment une localisée sur le campus de la Doua de Villeurbanne, accueillant jusqu'à 122 permanents et autant de doctorants et de post-doctorants. Le laboratoire y traite les thématiques principales de biométrie et de biologie évolutive sous différents aspects grâce au travail de plusieurs départements de recherche :

- GECO : génomique computationnelle et évolutive ;
- COEVOL : étude de la co-évolution à multi-échelle ;
- Ecologie évolutive ;
- BioStatistiques et Modélisation pour la santé et l'environnement.

Le stage que j'ai réalisé a eu lieu dans l'équipe Bioinformatique, Phylogénie et Génomique évolutive du département GECO.

II. Introduction

La recombinaison est un mécanisme qui se produit chez les individus lors de leur méiose. Elle joue un rôle essentiel dans le bon fonctionnement méiotique, permettant notamment d'assurer une bonne ségrégation des chromosomes dans les gamètes. Cela évite ainsi l'apparition de possibles aneuploïdies chez les individus, comme c'est le cas avec la trisomie 21 par exemple, où la mauvaise ségrégation des chromosomes induit la présence d'un chromosome en 3 exemplaires. Le phénomène de recombinaison a lieu pendant la prophase I de la méiose, où deux chromosomes peuvent échanger une partie de leur séquence d'ADN, ce qui peut conduire à un enjambement (ou crossing-over en anglais, cf. Figure 1). En cela, la recombinaison participe à la diversité génétique des individus et au sein des espèces. C'est

pourquoi il est intéressant de caractériser ce phénomène, notamment où les crossing-over se produisent sur le génome et à quelle intensité.

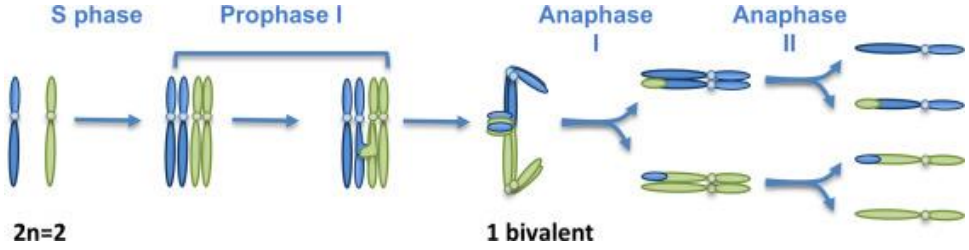


Figure 1 - Les différentes étapes de la méiose conduisant à des gamètes haploïdes [1]. Lors de la prophase I, deux chromosomes de deux paires différentes échangent de l'ADN : un crossing-over se produit entre ces deux chromosomes.

En effet, le taux de crossing-over peut varier tout au long du génome, et ces variations peuvent être différentes à plusieurs échelles [2]. D'une part elles peuvent être caractérisées différemment à grande échelle (de l'ordre du Mb) et à fine échelle (de l'ordre de kb) sur les chromosomes. Et d'autre part, elles peuvent être différentes à l'échelle de l'individu, des populations, des espèces, et même des groupes d'espèces.

Chez certaines espèces, comme la drosophile, le taux de recombinaison ne varie que très peu à fine échelle le long du génome, ce sont plutôt des variations à grande échelle qui sont observées. *A contrario*, chez d'autres espèces la distribution des taux de recombinaison est moins homogène à fine échelle, et on retrouve des points chauds de recombinaison (appelés hotspots). C'est le cas chez le chien, l'homme et la souris [3]. Parmi ces espèces qui ont des paysages de recombinaison avec hotspots, la position des hotspots est déterminée chez certaines par la liaison à l'ADN de la protéine Prdm9 de manière séquence-spécifique [4], c'est le cas par exemple de l'homme et de la souris, mais pas des canidés [5] pour lesquels le déterminisme des positions de hotspots n'est pas encore connu.

Prdm9 est un gène ancien, qui était déjà présent chez l'ancêtre commun des animaux. On retrouve ce gène dans de nombreux groupes taxonomiques (vertébrés, mollusques, insectes, cnidaires, éponges, etc.). Mais curieusement, ce gène a été perdu à de multiples reprises au cours de l'évolution des animaux (e.g. chez les insectes diptères et hyménoptères, chez les oiseaux, chez les canidés). Pour l'instant la fonction de Prdm9 n'a été caractérisée que chez l'homme et la souris. Donc nous ne savons pas si les homologues de Prdm9 trouvés en dehors des vertébrés ont une fonction identique à celle décrite chez les mammifères. Dans notre étude, nous allons nous intéresser à des animaux marins du

genre Ciona qui font partie du sous-embranchement des Urochordés, appelés plus récemment Tunicata, pour nous intéresser à des animaux très proches des vertébrés dans la phylogénie. On sait de ces animaux marins qu'ils possèdent la protéine Prdm9, mais sa fonction n'a pas encore été déterminée chez ces animaux. Ils ont un petit génome d'environ 100Mb, avec 14 chromosomes de tailles très différentes qui varient entre 1,5 et 10Mb.

Nous allons donc chercher à étudier et caractériser le paysage de recombinaison de la Cione, afin d'en caractériser la distribution et la position des zones de recombinaison. Pour cela, nous allons reconstruire une carte représentant la distribution des taux de recombinaison tout au long du génome des individus de Cione. Plusieurs méthodes existent pour construire ces cartes de recombinaison. Des méthodes de reconstruction par pédigré ont été souvent utilisées, car elles sont fiables et se basent sur des expériences de croisement entre individus ou bien sur des données de séquençage d'individus d'une même famille. Les méthodes par pédigré ne permettent d'obtenir que des cartes de recombinaison à faible résolution, car on étudie qu'un nombre limité d'individus, donc de méioses et moins de crossing-overs. Elles ne permettent donc de reconstruire que des cartes à grande échelle. Des méthodes plus récentes donnent la possibilité de réaliser des cartes à fine échelle, ce qui nous intéresse pour constater la présence ou non de hotspots dans le paysage de recombinaison de la Cione. Ces méthodes récentes s'appuient sur les données de polymorphisme des individus et cherchent à estimer le déséquilibre de liaison entre les marqueurs de polymorphisme pour inférer le taux de recombinaison. Ainsi, on arrive à obtenir des taux de recombinaison sur des petites fenêtres du génome (de l'ordre du kb) et avoir une information sur les phénomènes de recombinaison à fine échelle. Toutefois, les méthodes se basant sur le déséquilibre de liaison pour réaliser des cartes de recombinaison ont aussi leurs limites. Effectivement, elles sont moins performantes lorsqu'il est difficile d'estimer le déséquilibre de liaison, comme c'est le cas pour de petits chromosomes ou encore lorsque la taille efficace de la population étudiée est grande.

Nous allons nous appuyer sur ces méthodes récentes pour tenter de construire une carte de recombinaison de la Cione et estimer si les logiciels utilisant le déséquilibre de liaison permettent de réaliser une carte de recombinaison fiable chez la Cione.

III. Matériels et méthodes

Les données de départ

Pour reconstruire le paysage de recombinaison de la Cione, on dispose comme données de départ des données de polymorphisme pour deux espèces de Cione : un échantillon de 11 individus *Ciona robusta* et un échantillon de 13 individus *Ciona intestinalis*. Ces individus ont été séquencés, puis les séquences obtenues ont été alignées sur un génome de référence (*C. robusta* de 2011). Ensuite a été effectuée une étape de détection de variants (« variant calling » en anglais) qui répertorie les positions des marqueurs de polymorphisme sur le génome. Les marqueurs étudiés ici sont des SNPs, des Single Nucleotide Polymorphism, c'est-à-dire du polymorphisme concernant une seule base d'ADN pour chaque position de marqueur. Le travail réalisé pendant ce stage part de ces données de polymorphisme.

Filtrer les marqueurs

La première étape consiste à traiter les données d'entrée. On filtre les marqueurs selon plusieurs critères pour ne travailler que sur les plus fiables. Ici, quatre critères ont été choisis, ainsi on ne garde que les marqueurs :

- Bialléliques : uniquement deux allèles représentés dans l'échantillon, un allèle dit de référence et un allèle dit alternatif ;
- Qui ne sont pas des indels : les possibles sites d'insertions et de délétions peuvent apporter de la divergence entre les individus, mais les indels sont retirés car ils sont sources de possibles erreurs de séquençage ;
- Avec une MAF (Minor Allele Frequency) > 10 % : la MAF est ajustée pour avoir des sites polymorphes qui sont représentatifs de l'échantillon étudié ;
- Avec au maximum 50 % de génotypes manquants : cela signifie qu'on ne garde que les marqueurs dont on a le génotype au moins chez la moitié des individus de la population. Ici, on aurait pu garder uniquement les marqueurs qui n'avaient pas de génotype manquant, mais cela réduisait trop le nombre de marqueurs (Tableau 1), nous avons choisi de faire un compromis.

Lors de cette étape, on utilise un programme nommé VCFtools [6]. On précise que l'analyse des données est effectuée séparément pour *C. intestinalis* et *C. robusta*, les deux espèces de

Cione représentées dans nos données. On traite également les données chromosome par chromosome, puisque les logiciels que l'on utilise par la suite fonctionnent ainsi.

Phaser les marqueurs

Durant la seconde étape, on va phaser les marqueurs qui ont été filtrés précédemment. Le but du phasage est de déterminer les haplotypes des individus à partir des marqueurs que l'on a (Figure 2). Il existe plusieurs méthodes de phasage.

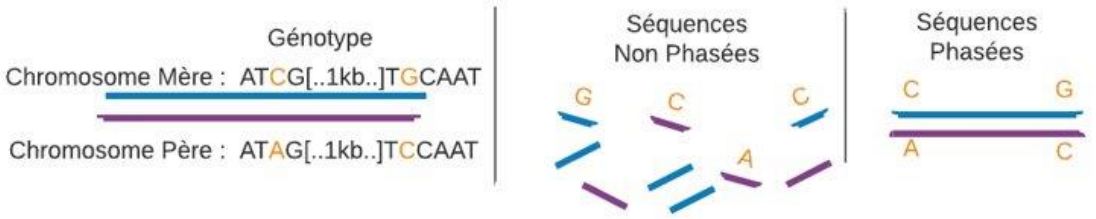


Figure 2 – Le phasage des SNPs (Single Nucleotide Polymorphism). Tout individu reçoit à la naissance un chromosome de sa mère et un chromosome de son père. Il a donc deux copies d'un même chromosome, sur lesquelles on peut trouver des allèles différents. Si tel est le cas, c'est à ces positions que l'on trouvera un marqueur de polymorphisme chez l'individu. En d'autres termes, lorsque l'on séquence l'individu et que l'on obtient les données de polymorphisme, on connaît les positions des marqueurs. Le phasage consiste alors à trouver comment les allèles sont agencés les uns par rapport aux autres dans la réalité du génome de l'individu.

La première méthode de phasage que l'on utilise est le phasage par trio. Ce phasage requiert de connaître le génotype de la descendance de l'individu ou celui de ses parents pour retrouver la configuration des allèles dans son génome. Le phasage par trio est un phasage fiable puisqu'il repose sur des données physiques. Malheureusement, c'est une méthode qui ne permet pas nécessairement de connaître le phasage de tous les marqueurs de polymorphisme répertoriés chez un individu.

Dans notre cas, les résultats de ce phasage par trio sur les données d'entrée ont déjà été obtenus grâce à des expériences de croisement. Ils sont déjà connus et le phasage par trio est explicité dans les données. On a donc simplement filtré avec VCFtools les marqueurs pour lesquels l'information du phasage était mentionnée.

La seconde méthode de phasage utilisée dans notre étude est le phasage statistique. Cette méthode implique d'inférer le phasage de chaque marqueur à partir de données de polymorphisme apportées en entrée. Il s'appuie également sur une carte génétique de référence, mentionnant les taux de recombinaison associés à chaque paire de marqueurs,

pour ajuster l'estimation du phasage. Nous avons utilisé le logiciel SHAPEIT (version 2) [7] pour effectuer le phasage statistique de nos données. En guise de carte génétique de référence, nous avons commencé par construire la carte de recombinaison à l'aide des marqueurs phasés par trio et du logiciel LDhat (que l'on détaillera ci-dessous).

Le phasage statistique est un phasage moins fiable que le phasage par trio, mais comme on peut le voir sur le Tableau 1, il permet de construire une carte de recombinaison avec une meilleure résolution.

Tableau 1 – Le nombre de SNPs baisse à chaque étape de traitement et influe sur la résolution de la carte de recombinaison que l'on obtient à la fin du processus. Le phasage statistique permet une meilleure résolution que le phasage par trio, puisqu'il estime le phasage de tous les SNPs, tandis que l'on a besoin de l'information sur la descendance pour phaser les SNPs par trio.

		C. intestinalis	C. robusta
Nombre de SNPs au départ		9 315 517	9 315 517
Nombre de SNPs après les 4 filtres	0 % max de génotypes manquant	178 802	211 356
	50 % max de génotypes manquants	885 526	868 903
Nombre de SNPs après phasage	Phasage par trio	67 531	58 098
	Phasage statistique	885 526	868 903
Résolution de la carte	Phasage par trio	1 SNP tous les 1.7 kb	1 SNP tous les 2.0 kb
	Phasage statistique	1 SNP tous les 0.1 kb	1 SNP tous les 0.1 kb

Inférence des taux de recombinaison

Après avoir filtré les marqueurs et les avoir phasés, on considère que les données d'entrée sont prêtes à être utilisées. Elles vont servir à estimer les taux de recombinaison grâce à deux logiciels : LDhat et LDhelmet. Les deux programmes ont en commun d'estimer les taux sur une fenêtre de deux marqueurs consécutifs à partir du déséquilibre de liaison. On obtient en sortie un taux de recombinaison associé à chacune de ces paires de marqueurs et de cette manière on construit une carte de recombinaison à fine échelle.

LDhat

Une première option employée pour estimer les taux de recombinaison est le logiciel LDhat [8]. Pour commencer, on crée une table de vraisemblance (*likelihood look-up table*) ajustée à nos données à partir du programme *complete* du logiciel, avec les paramètres suivants :

- Nombre de chromosomes (n) : 22 pour *C. robusta* et 26 pour *C. intestinalis* ;
- Nombre de points sur la grille (n_pts) : 101 ;
- Theta : 0.0153244 pour *C. robusta* et 0.0533579 pour *C. intestinalis* [9]. On note que theta est un estimateur de la diversité génétique d'une population donnée.

Une fois la table de vraisemblance obtenue pour chaque espèce de Cione, on estime les taux de recombinaison avec la commande *rhomap*. Plusieurs programmes dans LDhat permettent d'inférer les taux de recombinaison, mais *rhomap* est un choix plus adapté lorsque l'on s'attend à avoir un paysage de recombinaison à hotspots. Ce programme détaille mieux les variations à fine échelle sur la carte de recombinaison, or c'est ce que l'on cherche à étudier chez la Cione, voilà pourquoi nous l'avons choisi.

Ce programme utilise une méthode rjMCMC (*reversible-jump Monte Carlo Markov Chain*) pour inférer les taux de recombinaison. Cette méthode bayésienne réalise une exploration stochastique de l'espace des paramètres afin de les adapter au mieux au jeu de données mis en entrée. On choisit de faire tourner cette méthode sur un nombre de 1 100 000 itérations au total avec 100 000 itérations dites *burn-in*, qui vont permettre à la méthode rjMCMC d'ajuster au mieux les paramètres permettant l'inférence des taux, et 100 itérations dites *samp*, itérations laissées entre deux échantillons de la chaîne rjMCMC.

LDhelmet

Une seconde option pour estimer les taux de recombinaison grâce à la méthode du déséquilibre de liaison est LDhelmet [10]. En entrée du programme, nous allons utiliser des fichiers fasta pour chacun de nos chromosomes. Nous avons donc procédé à une conversion de notre fichier d'entrée au format vcf en utilisant la commande *vcf2fasta* de la librairie vcflib [11]. Pour cette conversion, le génome de référence utilisé est celui de *C. robusta* réalisé en 2011 [12]. Les fichiers fasta correspondent à un chromosome à la fois puisque LDhelmet traite les données de cette manière.

A partir de ces fichiers fasta, on utilise la commande *find_confs* de LDhelmet pour trouver les configurations des haplotypes présents chez l'individu. Ensuite, de la même manière que pour LDhat, on crée une table de vraisemblance pour chacune des deux espèces de Cione. Elle est créée avec la commande *table_gen* de LDhelmet en rentrant les mêmes valeurs de theta qu'avec LDhat et avec la grille de valeurs de ρ recommandée par le manuel (0.0 0.1

10.0 1.0 100.0). Cette grille représente les taux de recombinaison à l'échelle de la population en 1/bp.

Les deux étapes qui suivent sont optionnelles mais aident à préciser l'estimation des taux. On effectue tout d'abord une table de coefficients Padé avec la commande *pade* du logiciel. On donne les mêmes valeurs de theta que précédemment et on indique que l'on veut 11 coefficients Padé (valeur recommandée par le manuel). Deuxièmement, on construit la matrice de mutation de chacune de nos deux espèces de Cione. Cette matrice (Figure 3) répertorie la probabilité de passer d'un nucléotide à un autre lors d'une mutation, et cela pour les quatre bases possibles A, T, C, G. À partir des données de polymorphisme de départ, c'est-à-dire les données non filtrées, nous avons cherché les singletons. Un singleton est considéré comme un SNP n'apparaissant qu'une seule fois dans l'échantillon d'individus que nous étudions. Puis, on compte pour chacun de ces singletons le nombre de fois où l'on passe d'un allèle de référence, par exemple l'allèle A, à un allèle alternatif, par exemple l'allèle C. On convertit ensuite cela en probabilité en divisant le comptage par le nombre de singletons concernés par le même allèle de référence. Et c'est ainsi que l'on remplit la matrice de mutation, en répétant l'opération pour tous les allèles de référence et alternatifs possibles. La somme de chaque ligne de la matrice doit être égale à la fin à 1.

		Allèle alternatif				
		A	C	G	T	
Allèle de référence	A	$p(A \rightarrow A)$	$p(A \rightarrow C)$	...		$\Sigma\, p(A \rightarrow ...) = 1$
	C	$p(C \rightarrow A)$	...			$\Sigma\, p(C \rightarrow ...) = 1$
	G	...				$\Sigma\, p(G \rightarrow ...) = 1$
	T				$p(T \rightarrow T)$	$\Sigma\, p(T \rightarrow ...) = 1$

Figure 3 – Représentation d'une matrice de mutation.

Une fois que ces premières étapes ont été réalisées, le pré-processus est fini. On peut alors lancer la commande de LDhelmet pour inférer les taux de recombinaison. Celle-ci se nomme *rjmcmc* et tel que son nom l'indique, elle se base sur le même type de méthode que LDhat,

c'est-à-dire sur un algorithme bayésien rjMCMC. On donne alors en entrée les paramètres suivants :

- La fenêtre de SNPs w = 50 ;
- Le block-penalty b = 0, celui-ci indique que l'on veut bien calculer les taux de recombinaison sur des fenêtres de deux marqueurs consécutifs ;
- Le nombre d'itérations total de l'algorithme rjMCMC = 1 100 00 ;
- Le nombre d'itérations dites *burn-in* = 100 000.

Le fichier de sortie de cette commande étant un fichier binaire, il faut le convertir en fichier texte lisible, ce que l'on peut faire avec la commande *post_to_text* de LDhelmet. On utilise le paramètre -m en faisant tourner cette commande afin d'avoir la moyenne du taux de recombinaison sur chaque fenêtre de SNPs.

IV. Résultats

En vue de savoir si les méthodes basées sur le déséquilibre de liaison permettent de construire une carte de recombinaison fiable chez la Cione, nous avons investigué quatre protocoles de construction de carte (voir Tableau 2).

Tableau 2 – Les quatre protocoles utilisés pour construire la carte de recombinaison de la Cione

	PHASAGE		ESTIMATION DES TAUX DE RECOMBINAISON	
	Par trio	Statistique	LDhat	LDhelmet
Carte n°1	X		X	
Carte n°2		X	X	
Carte n°3	X			X
Carte n°4		X		X

Pour chaque carte, nous avons d'abord visualisé les variations des taux de recombinaison de manière globale à l'échelle du chromosome. A partir des taux obtenus par LDhat et LDhelmet entre chaque paire de marqueurs donnés en entrée, on visualise la carte en affichant le taux de recombinaison en fonction de la position des marqueurs. Sur les quatre cartes du chromosome 1 (Figure 4), on peut déjà constater un paysage de recombinaison avec beaucoup de variations. On peut notamment identifier un plateau au niveau du centromère et des pics de recombinaison au niveau des télomères, ce qui correspond à ce

que l'on s'attend à observer le long du chromosome. Les quatre cartes se différencient néanmoins les unes des autres. Par exemple, les cartes construites à partir des jeux de marqueurs phasés statistiquement sont plus denses en pics parce qu'elles contiennent plus d'information. Les cartes de LDhelmet et celles de LDhat indiquent également des variations différentes. Il sera intéressant donc de les comparer par la suite, car cette visualisation graphique ne suffit pas à observer ce qu'il se passe à fine échelle.

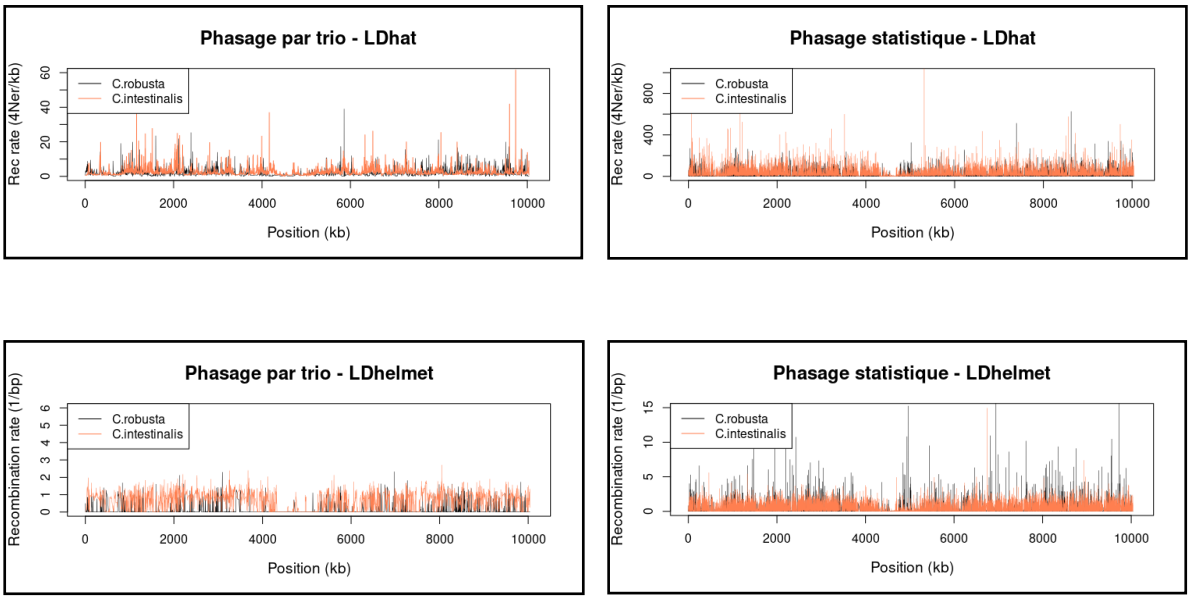


Figure 4 – La carte de recombinaison du chromosome 1 de Ciona intestinalis et Ciona robusta sous quatre protocoles différents.

Pendant la réalisation de ces protocoles, nous avons étudié trois facteurs qui peuvent exercer une influence sur la carte produite finalement. Le premier facteur est la reproductibilité des cartes. En effet, toutes les cartes ont été construites à partir d'un algorithme rjMCMC. De ce fait, nous n'obtenons pas les mêmes résultats de carte à partir d'un même jeu de données au départ si l'on fait tourner plusieurs fois l'algorithme. Les cartes ne sont donc pas parfaitement reproductibles avec les méthodes que nous avons employées, il convient donc d'observer l'influence que cela a sur les résultats. Le second facteur pouvant influencer le résultat des cartes est le phasage choisi et le dernier facteur est le logiciel utilisé pour estimer les taux de recombinaison.

Afin d'étudier ces divers facteurs, nous avons lissé les cartes de recombinaison à trois différentes échelles : 10kb, 100kb et 1Mb. Cela consiste à moyenner le taux de recombinaison sur des fenêtres de la taille de l'échelle fixée. Ainsi, on peut constater les

variations du taux de recombinaison de manière globale comme de manière plus fine sur les chromosomes.

Reproductibilité des cartes

Pour tester la reproductibilité des cartes de recombinaison, nous avons réalisé dix réplicats de carte de recombinaison avec LDhat. Ces réplicats ont été faits, pour des raisons de temps, uniquement sur le chromosome 1 de *C. intestinalis* et *C. robusta*, chromosome le plus long de leur génome d'une taille d'environ 10Mb. Les réplicats de carte ont réalisés avec le jeu de SNPs dont on connaît le phasage par trio (6996 SNPs chez *C.intestinalis* et 6776 SNPs chez *C. robusta*, donnant une résolution de carte d'1 SNP tous les 1.5 kb*)*. Nous avons phasé ces SNPs d'une part par trio et d'autre part statistiquement. Nous cherchons à regarder quelle est la corrélation entre les différents réplicats de carte de recombinaison, lorsqu'ils ont été faits en suivant un même processus, de manière à constater quel est le défaut de reproductibilité des cartes associé à l'algorithme rjMCMC.

Dans ce but, nous avons comparé les réplicats deux par deux et quantifié la corrélation entre eux en calculant le coefficient de Pearson associé à la corrélation, comme on peut le voir sur l'exemple de la Figure 5. Sur cette figure sont seulement affichés les deux premiers réplicats du chromosome 1 de *C. intestinalis*, réalisés à partir des SNPs phasés par trio. On peut constater que les taux de recombinaison des deux réplicats sont globalement très bien corrélés, en particulier à grande échelle (1Mb) avec un coefficient de Pearson à 0.997. Cependant cette corrélation n'est pas parfaite, comme on le voit à l'échelle plus fine de 10kb. Le coefficient de Pearson vaut alors 0.955, il n'est donc pas égal à 1, mais nous indique que le défaut de reproductibilité des cartes semble minime sur cet exemple.

Après avoir comparé tous les réplicats d'une même méthode et récupéré les coefficients de Pearson associé à chaque corrélation, on peut observer la distribution de ces coefficients. Cela a donc été faits pour deux méthodes comme dit précédemment : avec d'un côté les SNPs phasés par trio et d'un autre avec les SNPs phasés statistiquement (respectivement les histogrammes bleu et rose de la Figure 6 et de la Figure 10 en annexe). On voit alors qu'aux échelles 1Mb et 100kb, les taux de recombinaison entre réplicats corrèlent fortement, puisque la majorité des coefficients sont estimés autour de 0.99. Comme on a pu le voir avec l'exemple précédent, c'est à plus fine échelle (10kb) que l'on constate des coefficients

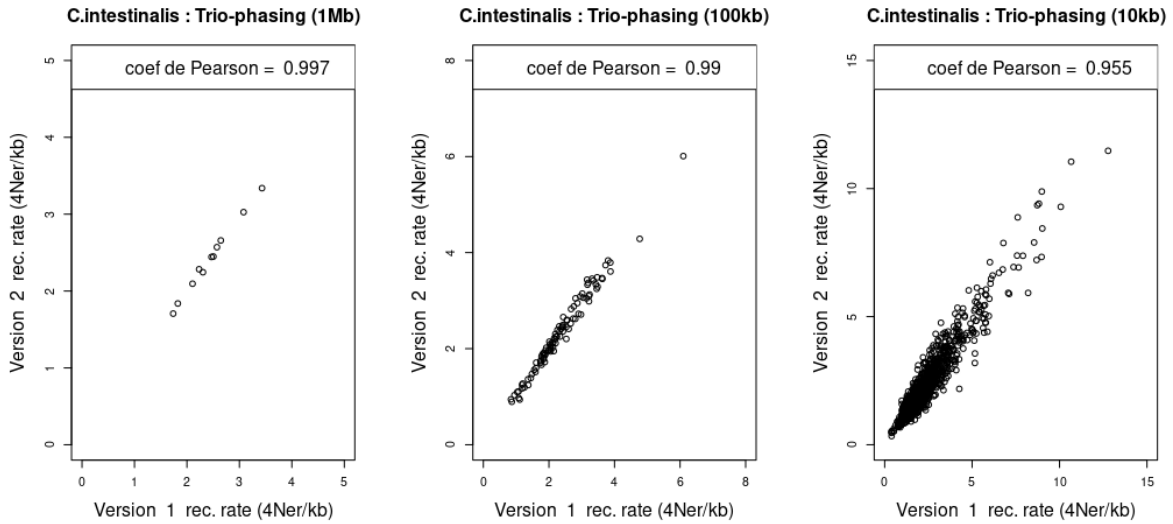


Figure 5 – Graphique de corrélation entre les taux de recombinaison des deux premiers réplicats de carte du chromosome 1 de C. intestinalis sous LDhat avec les SNPs phasés par trio. Les p-values des coefficients de Pearson sont toutes significatives au risque α de 5 %

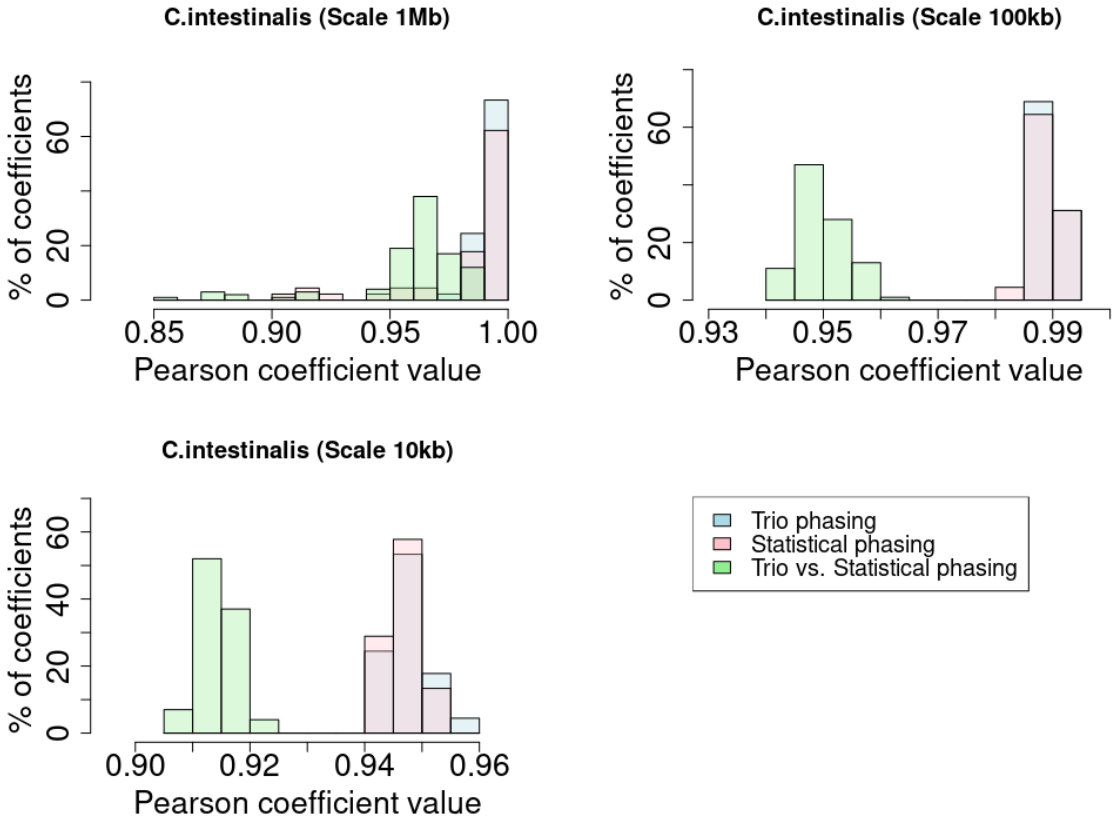


Figure 6 – Distribution des coefficients de Pearson associés à la corrélation entre réplicats du chromosome 1 de C. intestinalis faits sous LDhat. Les distributions en bleu et en rose correspondent à la comparaison entre réplicats faits à partir de la même méthode, respectivement l'une avec les SNPs phasés par trio et l'autre avec les SNPs phasés statistiquement. La distribution en vert correspond à la comparaison entre réplicats faits à partir de méthodes de phasage différentes. Les p-value des coefficients de Pearson sont toutes significatives au risque α de 5 %.

un peu plus faibles de l'ordre de 0.95. L'algorithme rjMCMC de LDhat ne permet donc pas de reproduire exactement les mêmes cartes de recombinaison à partir des mêmes données d'entrée, mais ces cartes tendent tout de même à indiquer globalement la même information et ne sont pas très différentes les unes des autres à fine échelle.

Comparaison entre phasage statistique et phasage par trio

Nous avons dans nos protocoles utilisé deux méthodes de phasage différentes pour phaser les marqueurs : le phasage par trio et le phasage statistique. Si l'on sait que le premier est plus fiable que le second par définition, on peut se demander à quel point les estimations du phasage statistique sont erronées par rapport au phasage par trio. En comparant les deux phasages, on trouve un pourcentage d'erreur de 11.30 % pour *C.intestinalis* et 13.66 % pour *C. robusta* (Tableau 5 en annexe). Il est intéressant de voir à quel point cela impacte les corrélations entre les cartes de recombinaison produites selon les deux méthodes.

De la même manière que pour contrôler la reproductibilité des cartes, nous avons utilisé les dix réplicats de carte du chromosome 1 construits à partir des deux méthodes de phasage pour étudier ces corrélations. La distribution des coefficients de Pearson des corrélations entre les deux phasages (courbe verte de la Figure 6 et de la Figure 10 en annexe) nous indique le coût que peut apporter de changer de phasage sur les résultats d'inférence des taux de recombinaison. A grande échelle (1Mb et 100kb), les taux de recombinaison obtenus avec les deux phasages chez *C. intestinalis* sont très fortement corrélés, avec un coefficient de Pearson valant 0.95 environ. On remarque que la valeur de la majorité des coefficients a baissé par rapport aux coefficients entre réplicats d'une même méthode. Cette baisse est plus marquée chez *C. robusta*, avec des valeurs de coefficients à 0.85. Cependant, cela est cohérent avec le pourcentage d'erreur du phasage statistique que nous avons calculé en amont et qui est plus élevé chez cette espèce. À plus fine échelle (10kb), on observe également que les coefficients de corrélation ont globalement une moins bonne valeur. Chez *C. intestinalis*, ils valent environ 0.91 alors que chez C. robusta, les valeurs diminuent à 0.80. On ne peut donc pas ignorer que le choix de la méthode de phasage a un impact sur la carte de recombinaison obtenue finalement.

Ces réplicats de carte de recombinaison ne montrant que ce qui a lieu sur le chromosome 1, on peut également regarder ce qu'il se passe sur le génome entier des deux espèces. La

corrélation entre les cartes de recombinaison faites sous LDhat avec chacune une méthode de phasage différente semble être bonne (Figure 7 et Figure 11 en annexe). En effet, à l'échelle de 1Mb, les points du graphique sont à peu près alignés en une droite, tout comme ils le sont sur le graphe à l'échelle 100kb. Les coefficients de Pearson confirment cette corrélation positive significative. A l'échelle 10kb, cette corrélation est moins nette et le coefficient de Pearson baisse un peu à 0.67 pour *C. intestinalis* et à 0.44 pour *C. robusta*. Modifier le phasage a donc un impact sur l'estimation des taux de recombinaison et conduit à des cartes de recombinaison différentes. Cependant, les cartes sont au vu de l'analyse des corrélations quand même cohérentes et corrèlent bien entre elles.

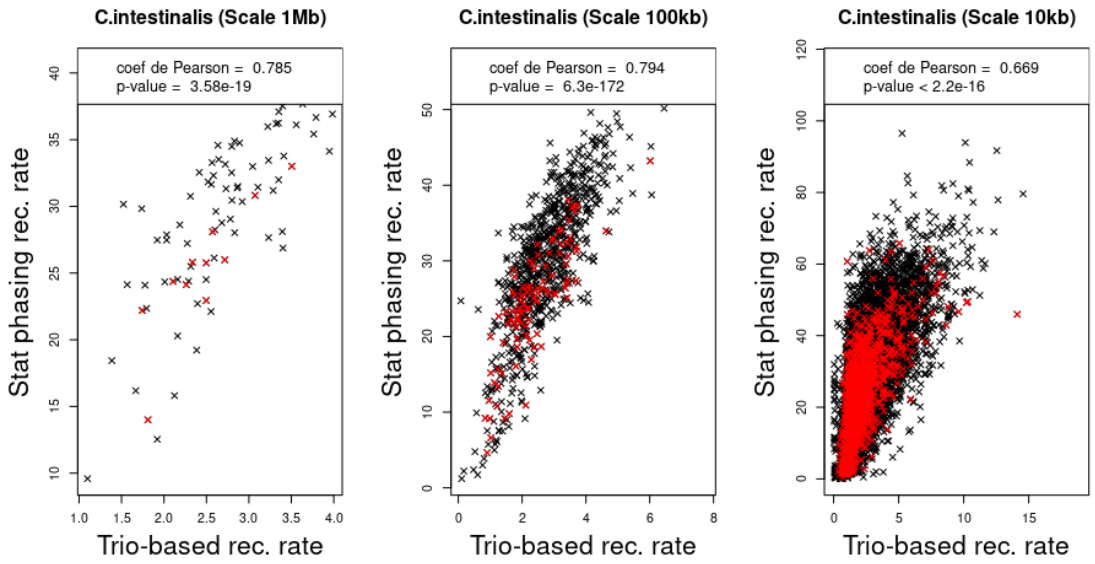


Figure 7 – Graphiques de corrélation entre la carte de recombinaison de C. intestinalis réalisée avec les SNPs phasés par trio et celle avec les SNPs phasés statistiquement sous LDhat. Les points en rouge sont associés au chromosome 1.

Comparaison des 2 programmes : LDhat et LDhelmet

Enfin, le programme choisi pour estimer les taux de recombinaison influe sur la construction de la carte de recombinaison. Cela se voit par exemple lorsque l'on compare les cartes réalisées avec la même méthode de phasage (les marqueurs phasés statistiquement) mais l'une faite avec LDhat et l'autre avec LDhelmet (Figure 8 et Figure 12 en annexe). Les corrélations sont positives et significatives, avec une meilleure corrélation à grande échelle (1Mb) qu'à petite échelle (10kb) chez *C. intestinalis*. Les deux cartes de recombinaison faites avec les 2 outils corrèlent donc globalement, mais comme on peut le voir à l'échelle 100kb, les deux programmes ne donnent pas d'estimations similaires pour les forts taux de

recombinaison. Chez *C. robusta*, les corrélations sont positives et significatives, mais les valeurs des coefficients de Pearson sont plutôt faibles autour de 0.1-0.2. Afin d'avoir une vue plus nette sur ce qu'il se passe pour des taux plus faibles, on peut regarder les corrélations à échelle logarithmique (Figure 9 et Figure 13 en annexe). Les coefficients de Pearson ont de bien meilleures valeurs sur ces graphes, notamment sur la carte de *C. robusta* lissée à 10kb avec un coefficient de Pearson à 0.76. Les points correspondant aux taux de recombinaison les plus faibles suivent d'ailleurs un bon alignement. Cela incite à penser que LDhat et LDhelmet ont des estimations de taux de recombinaison similaires sur les faibles taux.

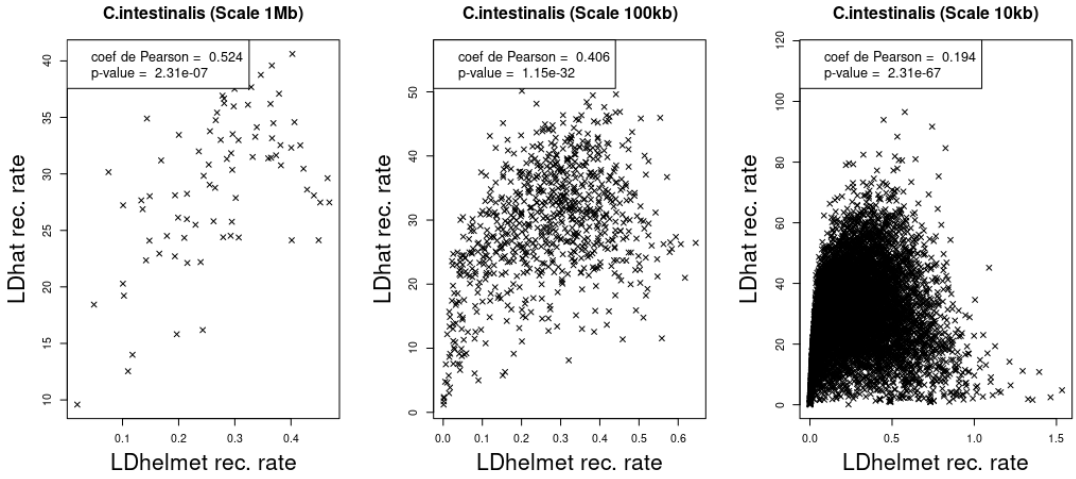


Figure 8 – Graphiques de corrélation entre la carte de recombinaison de C. intestinalis obtenue sous LDhat et celle obtenue sous LDhelmet. La méthode de phasage des marqueurs est identique : le phasage statistique.

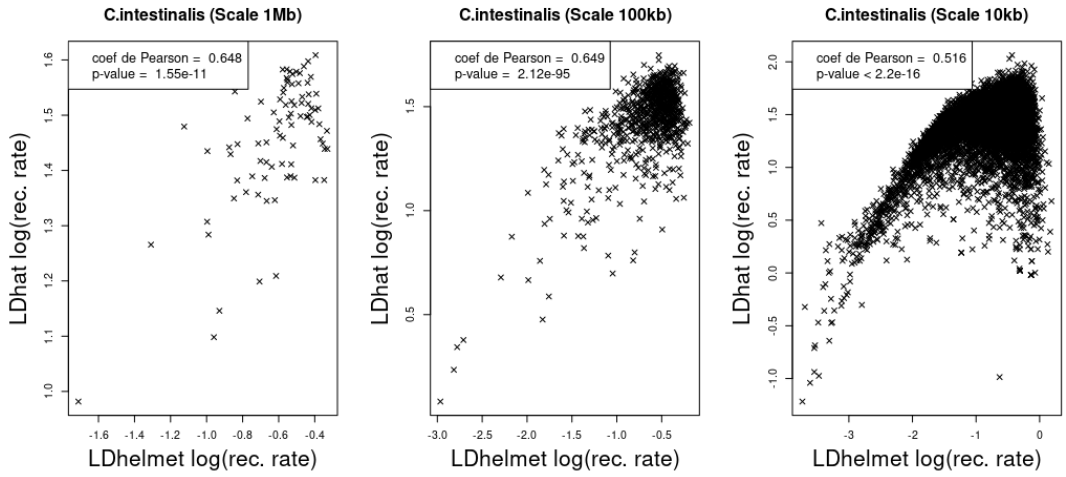


Figure 9 - Graphiques de corrélation entre la carte de recombinaison de C. intestinalis obtenue sous LDhat et celle obtenue sous LDhelmet. La méthode de phasage des SNPs est identique : le phasage statistique. Les taux de recombinaison des cartes sont affichés à échelle logarithmique.

Fiabilité des cartes de recombinaison

Une fois les cartes de recombinaison construites grâce à LDhat et LDhelmet, il est intéressant dans notre étude de voir si ces cartes créées à partir de méthodes basées sur le déséquilibre de liaison sont fiables. Comme nous n'avions pas de carte génétique de référence de la Cione déjà réalisée et à laquelle on aurait pu comparer nos cartes, nous avons choisi deux critères pour évaluer la fiabilité des cartes que nous avons produites. Le premier critère est le taux de GC. Il a été prouvé chez un bon nombre d'espèces eucaryotes que le taux de recombinaison est corrélé positivement avec le taux de GC [13]. Nous avons donc calculé les coefficients de Pearson associés à la corrélation avec le taux de GC pour chacune des quatre cartes réalisées (Tableau 3).

Tableau 3 – Coefficients de Pearson des corrélations entre les quatre cartes de recombinaison et le taux de GC. (: la p-value de ces coefficients est significative au risque α de 5 %)*

	Echelle / Espèce	1Mb	100kb	10kb
Carte n°1 Phasage par trio LDhat	*C. intestinalis*	0.19	0.31*	0.25*
	C. robusta	0.02	-0.04	-0.09
Carte n°2 Phasage statistique LDhat	*C. intestinalis*	0.37*	0.30*	0.22*
	C. robusta	0.10	-0.13*	-0.19*
Carte n°3 Phasage par trio LDhelmet	*C. intestinalis*	0.13	0.005	-0.03*
	C. robusta	0.03	0.04	0.02
Carte n°4 Phasage statistique LDhelmet	*C. intestinalis*	0.07	0.01	-0.003
	C. robusta	-0.04	-0.08*	-0.04*

Les coefficients de corrélation avec le taux de GC des cartes LDhelmet sont très faibles, voire tendent à être nuls et indiquent rarement une corrélation significative. Au contraire, pour les cartes LDhat, la corrélation avec le taux de GC est détectée, mais les résultats diffèrent entre les deux espèces de Cione : chez *C. intestinalis*, une corrélation positive et significative est trouvée aux échelles 100kb et 10kb, tandis que chez *C. robusta* les coefficients sont négatifs et significatifs aux mêmes échelles pour la carte avec phasage statistique. Il est étonnant d'ailleurs d'avoir des coefficients de corrélation négatifs pour ce critère est étonnant et ne coïncide pas à ce que l'on attendait à obtenir. Cela remet en question donc

la fiabilité des cartes construites avec *C. robusta* sous LDhat, tout comme la fiabilité des cartes réalisées avec LDhelmet puisqu'aucune corrélation avec le taux de GC n'est repérée.

Nous avons également exploré un second critère de fiabilité : la longueur des chromosomes. En effet, on sait qu'un chromosome plus long a des taux de recombinaison moins forts qu'un chromosome plus petit [13]. On regarde donc si les quatre cartes de recombinaison corrèlent négativement avec la longueur des chromosomes et si elles confirment donc ce critère (Tableau 4).

Tableau 4 - Coefficients de Pearson des corrélations entre les quatre cartes de recombinaison et la longueur des chromosomes. (: la p-value de ces coefficients est significative au risque α de 5 %)*

	Espèce	Coefficient de Pearson
Carte n°1 Phasage par trio LDhat	*C. intestinalis*	0.45
	C. robusta	-0.51
Carte n°2 Phasage statistique LDhat	*C. intestinalis*	-0.21
	C. robusta	-0.71*
Carte n°3 Phasage par trio LDhelmet	*C. intestinalis*	-0.64*
	C. robusta	-0.75*
Carte n°4 Phasage statistique LDhelmet	*C. intestinalis*	-0.72*
	C. robusta	-0.63*

Les cartes de LDhat ne corrèlent pas significativement avec la longueur des chromosomes, à part la carte faite avec *C. robusta* et les SNPs phasés statistiquement. Le coefficient de Pearson étant de -0.71, une belle corrélation négative est repérée, ce qui va dans le sens de ce à quoi l'on s'attendait. Les cartes de LDhelmet quant à elles corrèlent toutes significativement et négativement avec la longueur des chromosomes. Ce critère de fiabilité est validé pour les cartes de LDhelmet mais pas pour les cartes de LDhat.

Les corrélations avec les deux critères de fiabilité choisis donnent des résultats qui ne vont pas dans le même sens, les cartes sous LDhat corrélant bien avec le taux de GC et les cartes sous LDhelmet corrélant mieux avec la longueur des chromosomes. Seulement si on réunit les informations apportées par l'évaluation de ces critères, aucune des quatre cartes construites ne valide tous les critères de fiabilité.

V. Discussion

Dans le but de construire une carte de recombinaison fiable de la Cione, nous avons exploré quatre protocoles où variaient la méthode de phasage et le logiciel utilisé pour estimer les taux de recombinaison. En plus de cela, nous avons remarqué que peu importe le programme choisi, les cartes de recombinaison ne pouvaient pas être parfaitement reproduites à partir des mêmes données de départ. Les résultats de nos corrélations nous indiquent que deux cartes suivant un même protocole ne seront certes pas parfaitement identiques, mais corrèleront très fortement. Le défaut de reproductibilité des cartes introduit par les algorithmes rjMCMC des logiciels d'estimation des taux n'impacte donc pas grandement la construction des cartes de recombinaison.

En ce qui concerne les méthodes de phasage, il était certain au départ que le phasage le plus fiable était le phasage par trio, car il s'appuie sur des données physiques de pédigré. Seulement, ce phasage a comme inconvénient de diminuer considérablement le nombre de marqueurs de polymorphisme que l'on peut apporter aux logiciels pour estimer les taux de recombinaison. Ainsi, la résolution de la carte de recombinaison est moins bonne. Le phasage statistique quant à lui permet d'avoir une meilleure résolution en phasant plus de marqueurs, mais ce phasage est a priori moins fiable puisqu'il est estimé. Après avoir analysé les corrélations entre une carte avec marqueurs phasés par trio et une carte avec marqueurs phasés statistiquement, on constate que les corrélations sont très bonnes et indiquent que les cartes sont cohérentes entre elles. De plus, le pourcentage d'erreur moyen du phasage statistique chez les deux espèces de Cione n'est pas très élevé (aux alentours de 12%). En somme le phasage statistique paraît être un compromis fiable pour gagner en résolution de carte et pouvoir obtenir des cartes de recombinaison à plus fine échelle qu'avec le phasage par trio.

Enfin, le point le plus important de mes résultats concernent les différences que nous avons détectées entre LDhat et LDhelmet. A partir du même jeu de données, les deux logiciels réalisent des cartes de recombinaison des deux espèces de Cione qui ne se ressemblent pas. Au moins une des deux méthodes nous donnent donc de faux résultats. L'analyse des corrélations ne permet pas de trancher entre les deux logiciels sur celui qui permet de

construire une carte plus fiable que l'autre. La première raison est que les cartes de LDhat et LDhelmet ne corrèlent pas très bien surtout à fine échelle. Elles sont cohérentes sur les faibles taux de recombinaison, mais ne donnent pas la même information pour les taux de recombinaison plus forts, c'est-à-dire les zones à hotspots. Deuxièmement, les critères de fiabilité du GC et de la longueur de chromosomes ne permettent pas de mettre en avant un des deux logiciels. Les cartes LDhat semblent mieux corréler avec le taux de GC alors que les cartes LDhelmet donnent de meilleures corrélations avec la longueur de chromosomes.

D'après ces résultats, on peut dire qu'au moins une des deux méthodes utilisées n'est pas fiable pour réaliser des cartes de recombinaison de *C. intestinalis* et *C. robusta*. On ne peut pas conclure qu'un de nos quatre protocoles a conduit à construire une carte de recombinaison fiable de la Cione. Cela peut s'expliquer par le fait que ces méthodes basées sur le déséquilibre de liaison ne sont applicables que dans la mesure où le rapport entre le taux de recombinaison r (nombre de crossing-over par génération et par paire de base) et le taux de mutation μ (nombre de mutation par génération et par paire de base) n'est pas trop élevé. La valeur de ce rapport influe sur le fonctionnement des méthodes basées sur le déséquilibre de liaison, car le taux de mutation μ impacte la densité en marqueurs de polymorphisme que l'on va trouver chez les individus de l'espèce, et le taux de recombinaison r est lié directement au déséquilibre de liaison. Si un faible taux de crossing-overs a lieu sur un locus, alors on va repérer du déséquilibre de liaison au niveau de ce locus. Si ce taux devient trop fort, on ne repérera plus de déséquilibre de liaison, les crossing-overs n'étant plus rares sur ce locus. Pour ces raisons, les méthodes d'estimation des taux de recombinaison basées sur le déséquilibre de liaison saturent lorsque le taux r est grand et qu'il y a une grande différence entre le taux de mutation μ et le taux de recombinaison r qui conduit à un rapport r/μ important. Il se trouve que chez l'homme, espèce pour laquelle on a déjà prouvé que les méthodes basées sur le déséquilibre de liaison sont fiables pour construire des cartes, le rapport r/μ est de l'ordre de 1, alors que chez la Cione on estime que ce rapport est de 30 à 100 fois plus fort (chez l'homme r est environ égal à 1cM/Mb alors que chez la Cione r vaut plutôt 38cM/Mb [14]). On est en dehors du champ d'applicabilité des méthodes basées sur le déséquilibre de liaison pour estimer les taux de recombinaison.

VI. Conclusion

Le but de ces travaux était de réaliser des cartes de recombinaison chez la Cione, grâce en particulier aux méthodes de construction de carte s'appuyant sur le déséquilibre de liaison, et de pouvoir en discuter la fiabilité. Quatre cartes de recombinaison ont été réalisées au total, en changeant soit la méthode de phasage des marqueurs en entrée, soit la méthode elle-même qui infère les taux de recombinaison (LDhat ou LDhelmet). Trois points ont été analysés à partir de ces cartes : la reproductibilité des cartes, l'impact de la méthode de phasage et celui de la méthode d'inférence des taux choisie sur les résultats. Finalement, au vu des résultats de corrélation. Les deux premiers points donnent des résultats satisfaisants : ni le défaut de reproductibilité, ni le choix du phasage statistique ne semblent avoir un impact conséquent sur l'estimation des taux de recombinaison. Le dernier point analysé donne des résultats plus inquiétants, car LDhat et LDhelmet se contredisent sur les estimations des taux de recombinaison. Au moins une des ces deux méthodes indiquent des résultats faux avec les données de polymorphisme de la Cione. Il est possible que cela soit dû au fait que les méthodes basées sur le déséquilibre de liaison ne sont pas adaptées pour reconstruire les cartes de recombinaison chez cette espèce. Les perspectives de travail concernant les cartes de recombinaison de la Cione consisteraient à affiner les protocoles pour tenter de faire des cartes fiables. Une piste à explorer est notamment de réaliser une carte de recombinaison sans phaser les marqueurs en entrée, ce qui est possible avec LDhat, et voir si cela induit des différences dans les résultats des cartes par rapport aux autres phasages. Une autre possibilité serait de travailler sur l'utilisation des logiciels LDhat et LDhelmet et d'effectuer des simulations pour trouver le jeu de paramètres le plus adapté aux données de la Cione et estimer les taux de recombinaison de manière plus fiable.

VII. Bibliographie

[1] M. Grelon, « Meiotic recombination mechanisms », *C. R. Biol.*, vol. 339, nº 7, p. 247-251, juill. 2016, doi: 10.1016/j.crvi.2016.04.003.

[2] M. Zelkowski, M. A. Olson, M. Wang, et W. Pawlowski, « Diversity and Determinants of Meiotic Recombination Landscapes », *Trends Genet.*, vol. 35, nº 5, p. 359-370, mai 2019, doi: 10.1016/j.tig.2019.02.002.

[3] J. Stapley, P. G. D. Feulner, S. E. Johnston, A. W. Santure, et C. M. Smadja, « Variation in recombination frequency and distribution across eukaryotes: patterns and processes », *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 372, nº 1736, p. 20160455, déc. 2017, doi: 10.1098/rstb.2016.0455.

[4] S. Myers, C. Freeman, A. Auton, P. Donnelly, et G. McVean, « A common sequence motif associated with recombination hot spots and genome instability in humans », *Nat. Genet.*, vol. 40, nº 9, p. 1124-1129, sept. 2008, doi: 10.1038/ng.213.

[5] A. Auton *et al.*, « Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs », *PLOS Genet.*, vol. 9, nº 12, p. e1003984, déc. 2013, doi: 10.1371/journal.pgen.1003984.

[6] P. Danecek *et al.*, « The variant call format and VCFtools », *Bioinformatics*, vol. 27, nº 15, p. 2156-2158, août 2011, doi: 10.1093/bioinformatics/btr330.

[7] O. Delaneau, J. Marchini, et J.-F. Zagury, « A linear complexity phasing method for thousands of genomes », *Nat. Methods*, vol. 9, nº 2, p. 179-181, févr. 2012, doi: 10.1038/nmeth.1785.

[8] A. Auton et G. McVean, « Recombination rate estimation in the presence of hotspots », *Genome Res.*, vol. 17, nº 8, p. 1219-1227, août 2007, doi: 10.1101/gr.6386707.

[9] J. Romiguier *et al.*, « Comparative population genomics in animals uncovers the determinants of genetic diversity », *Nature*, vol. 515, nº 7526, p. 261-263, nov. 2014, doi: 10.1038/nature13685.

[10] A. H. Chan, P. A. Jenkins, et Y. S. Song, « Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster », *PLOS Genet.*, vol. 8, nº 12, p. e1003090, déc. 2012, doi: 10.1371/journal.pgen.1003090.

[11] E. Garrison, Z. N. Kronenberg, E. T. Dawson, B. S. Pedersen, et P. Prins, « Vcflib and tools for processing the VCF variant call format », mai 2021. doi: 10.1101/2021.05.21.445151.

[12] R. Pennati *et al.*, « Morphological Differences between Larvae of the Ciona intestinalis Species Complex: Hints for a Valid Taxonomic Definition of Distinct Species », *PloS One*, vol. 10, nº 5, p. e0122879, 2015, doi: 10.1371/journal.pone.0122879.

[13] L. Duret et N. Galtier, « Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes », *Annu. Rev. Genomics Hum. Genet.*, vol. 10, nº 1, p. 285-311, 2009, doi: 10.1146/annurev-genom-082908-150001.

[14] S. Kano, N. Satoh, et P. Sordino, « Primary Genetic Linkage Maps of the Ascidian, Ciona intestinalis », *Zoolog. Sci.*, vol. 23, nº 1, p. 31-39, janv. 2006, doi: 10.2108/zsj.23.31.

[15] S. Singhal *et al.*, « Stable recombination hotspots in birds », *Science*, vol. 350, nº 6263, p. 928-932, nov. 2015, doi: 10.1126/science.aad0843.

[16] A. Auton *et al.*, « A Fine-Scale Chimpanzee Genetic Map from Population Sequencing », *Science*, vol. 336, nº 6078, p. 193-198, avr. 2012, doi: 10.1126/science.1216872.

Annexes

Annexe 1 :

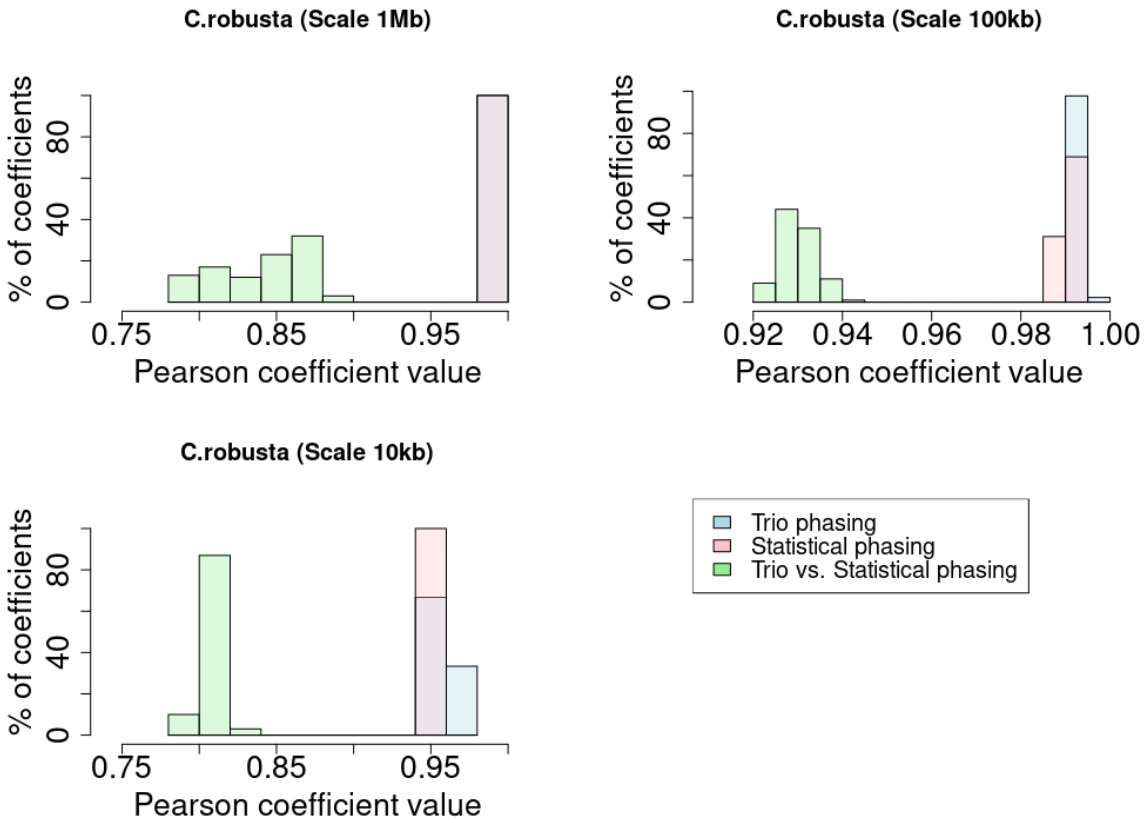


Figure 10 - Distribution des coefficients de Pearson associés à la corrélation entre réplicats du chromosome 1 de C. robusta faits sous LDhat. Les distributions en bleu et en rose correspondent à la comparaison entre réplicats faits à partir de la même méthode, respectivement l'une avec les SNPs phasés par trio et l'autre avec les SNPs phasés statistiquement. La distribution en vert correspond à la comparaison entre réplicats faits à partir de méthodes de phasage différentes. Les p-values des coefficients de Pearson sont toutes significatives au risque α de 5 %.

Annexe 2 :

Tableau 5 – Pourcentages d'erreur moyens du phasage statistique sur le phasage des SNPs pour les deux espèces de Cione : C. intestinalis et C. robusta. Les pourcentages sont indiqués pour chaque chromosome et pour chaque individu de l'échantillon étudié.

CHROMOSOME	% moyen d'erreur sur le chromosome		*C. intestinalis*		*C. robusta*	
	C. intestinalis	*C. robusta*	*INDIVIDUS*	*% moyen d'erreur*	INDVIDUS	% moyen d'erreur
Chr 1	11.57 %	13.49 %	Ad13	11.71 %	Ad10	15.84 %
Chr 2	11.40 %	13.22 %	Ad15	11.41 %	Ad11	15.31 %
Chr 3	11.59 %	13.26 %	Ad19	6.87 %	Ad12	13.70 %
Chr 4	11.71 %	14.25 %	Ad24	12.67 %	Ad16	3.27 %
Chr 5	11.40 %	14.12 %	Ad25	12.64 %	Ad22	16.57 %
Chr 6	10.47 %	13.01 %	Ad26	12.96 %	Ad23	15.89 %
Chr 7	10.99 %	13.13 %	Ad27	12.19 %	Ad32	16.63 %
Chr 8	11.35 %	13.22 %	Ad29	13.34 %	Ad33	15.67 %
Chr 9	11.22 %	13.81 %	Ad30	0.39 %	Ad34	12.35 %
Chr 10	11.28 %	13.95 %	Ad3	12.78 %	Ad35	11.85 %
Chr 11	11.27 %	13.63 %	Ad4	12.87 %	Ad6	13.15 %
Chr 12	11.49 %	14.54 %	Ad7	13.98 %		
Chr 13	11.51 %	12.96 %	Ad9	13.12 %		
Chr 14	11.01 %	14.60 %				
TOTAL	**11.30 %**	**13.66 %**	TOTAL	**11.30 %**		**13.66 %**

Annexe 3 :

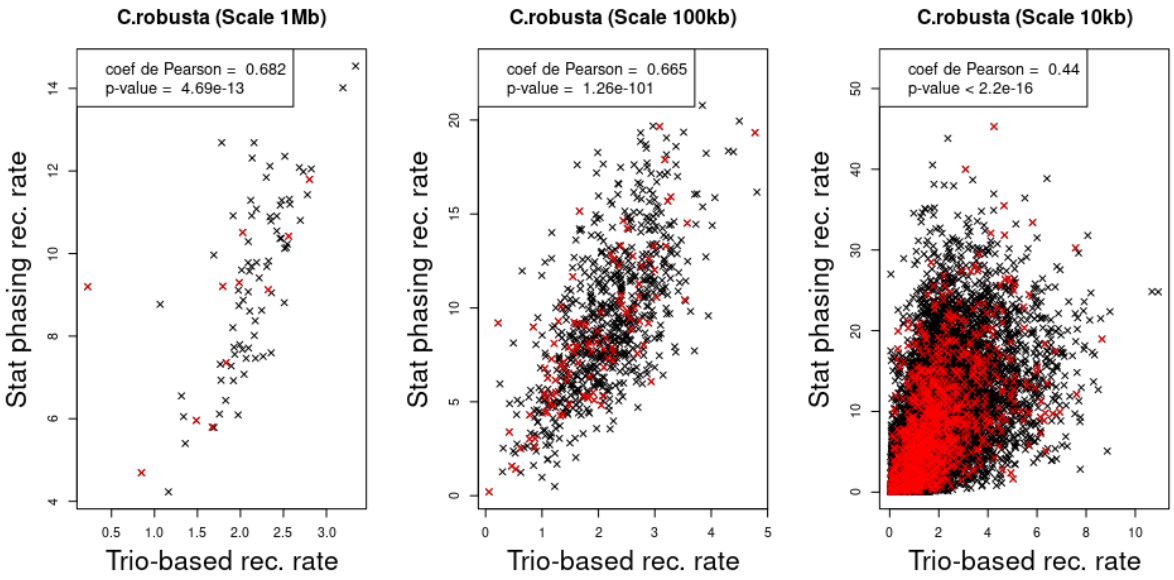


Figure 11 - Graphiques de corrélation entre la carte de recombinaison de C. robusta réalisée avec les SNPs phasés par trio et celle avec les SNPs phasés statistiquement sous LDhat. Les points en rouge sont associés au chromosome 1.

Annexe 4 :

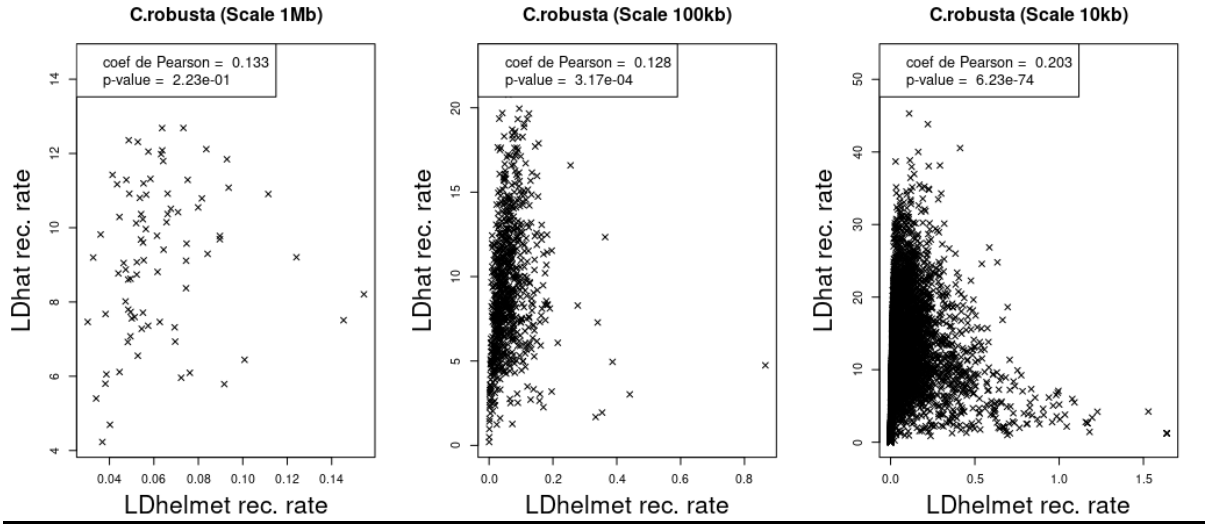


Figure 12 – Graphiques de corrélation entre la carte de recombinaison de C. robusta obtenue sous LDhat et celle obtenue sous LDhelmet. La méthode de phasage des SNPs est identique : le phasage statistique.

Annexe 5 :

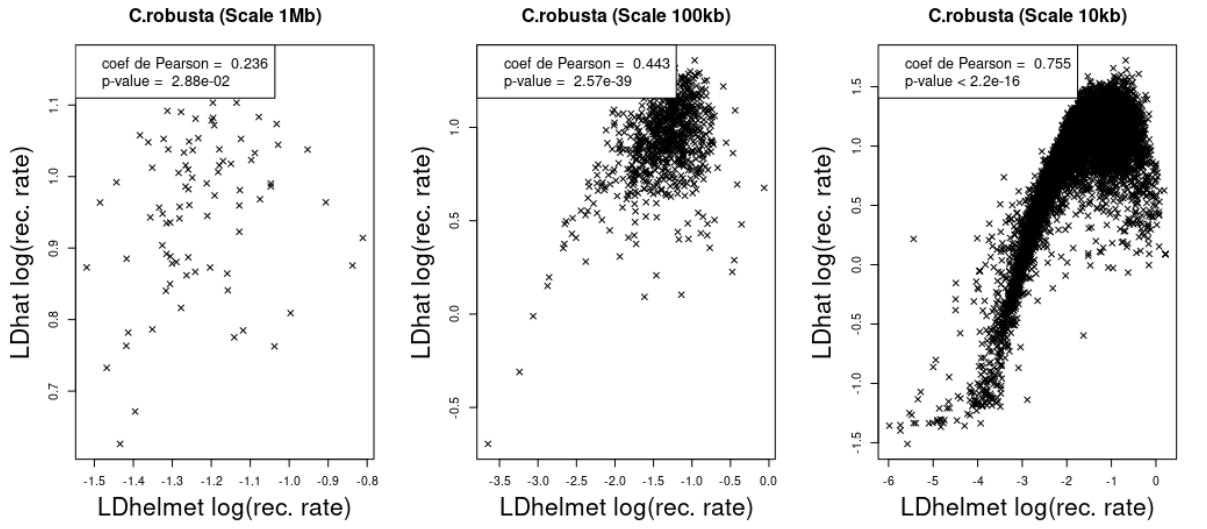


Figure 13 - Graphiques de corrélation entre la carte de recombinaison de C. robusta obtenue sous LDhat et celle obtenue sous LDhelmet. La méthode de phasage des SNPs est identique : le phasage statistique. Les taux de recombinaison des cartes sont affichés à échelle logarithmique.