Thèse de Doctorat

Elsa QUILLERY

Mémoire présenté en vue de l'obtention du
grade de Docteur d'Oniris - l'École Nationale Vétérinaire Agroalimentaire et de l'Alimentation Nantes-Atlantique
sous le label de L'Université Nantes Angers Le Mans

,
École doctorale : Biologie Santé

Discipline : Agronomie, productions animales et végétales, agroalimentaire
Spécialité : *Biologie de l'environnement, des populations, écologie*
Unité de recherche : ***UMR 1300 INRA Oniris « Biologie, Epidémiologie, Analyse de Risques en santé animale », Oniris, Atlanpole, La Chantrerie, BP 40706, 44307 Nantes***
Soutenue le jeudi 19 décembre 2013
Thèse N° :

Développement de marqueurs génétiques (SNPs) à partir du génome de la tique *Ixodes ricinus* pour l'étude de la structure génétique de ses populations à l'échelle du paysage

JURY

Rapporteurs :	**Muriel VAYSSIER-TAUSSAT**, Directrice de Recherche, USC Bartonella-Tiques, ENVA Maisons-Alfort
	Eric GRENIER, Chargé de recherche, IGEPP, Rennes
Examinateurs :	**Karine HUBER, Chargé de Recherche**, UMR CIRAD-INRA CMAEE, Montpellier
	Xavier BAILLY, Ingénieur de Recherche, UR INRA Epidémiologie Animale, Clermont-Ferrand
Directeur de Thèse :	**Alain CHAUVIN**, Professeur, UMR 1300 Oniris-INRA « BioEpAR », Nantes
Co-encadrant de Thèse :	**Olivier PLANTARD**, Chargé de Recherche, UMR 1300 Oniris-INRA « BioEpAR », Nantes

Remerciements

En premier lieu, je remercie Alain Chauvin, pour avoir accepté d'être mon directeur de thèse et ce malgré les distances de nos thématiques.

Je remercie également Olivier Plantard, pour m'avoir confié ce projet, accompagné dans sa réalisation et pour la confiance accordée tout au long de ces travaux, ce qui m'a permis de gérer ce projet avec une certaine autonomie. Même si ça n'a pas toujours été facile, les longues discussions scientifiques que l'on a pu avoir, m'ont beaucoup apporté.

Je remercie Muriel Vayssier-Taussat et Eric Grenier d'avoir accepté le rôle de rapporteur et de juger ce travail de thèse.
Je remercie également Karine Huber et Xavier Bailly d'avoir accepté celui d'examinateur.

Je remercie les membres de mes comités de thèse, Karen McCoy, Solenn Stoeckel et Jean-François Cosson pour leurs conseils avisés lors des différentes réunions que l'on a pu avoir.

Je remercie François Beaudeau, pour ces qualités scientifiques et humaines, son aide et son soutien durant ces trois années.

Je remercie également mes encadrants de stage de master, Christine, Louis, Alison, Simon et Oliver, pour tout ce qu'ils m'ont apporté, le goût de la recherche, c'est grâce (ou à cause) à vous que j'en suis arrivée là!

Ségo et Manue, je ne peux citer l'une sans l'autre tellement on a partagé de choses ensemble, quelques lignes ne suffiraient pas pour exprimer tous mes remerciements. Il faudrait rédiger une thèse entière pour cela… mais là tout de suite, je n'ai pas vraiment envie de remettre le couvert! Bref, ces longues discussions de tout de rien, ces soirées et toutes ces bières, les pauses de 16h (et…l'attente quand Ségo apparaissait à 16h07), un certain soir à Beaulieu…à la sortie des vestiaires pour y voir apparaitre Niko… l'hystérie qui a suivie, les weekend à St-Brévin, la découverte du wakeboard, les fous rires, votre inestimable soutien dans les moments durs (et il y en a eu !!!). Ces dernière semaines constitueraient un chapitre à part aussi tellement vous m'avez apporté, toute votre aide, vos relectures, la chasse au 's' qui apparaissent malencontreusement un peu n'importe où dans mes écrits. C'est merveilleux d'avoir des amies comme vous !

Erwann, à ton tour d'être remercié. En quelques lignes ça va être compliqué aussi de tout dire, mais merci d'être là, d'être toi (même si des fois… c'est pas facile de LoLiser avec toi ;)). Bref, merci pour l'ami que tu es devenu au labo comme en dehors ! Merci pour toute ton aide informatique, de m'avoir fait une installation de fou, ton aide pour maitriser la bête qu'est Linux, ton aide fabuleuse pour les scripts et tout et tout…

Thierry, c'est l'heure de te remercier !! Mon ronchon préféré, merci d'être là. Tes « IL EST QUELLE HEURE ? » n'auront pas eu raison de mon « c'est quelle heure ?» mais ton écoute, nos discussions, ton aide, m'auront apporté bien autre chose qu'un bon français ;). Les folles soirées à Barcelone ou Saragosse, ces bières à la Cervoiserie, les sorties piscines qui pouvaient aussi bien être totalement 'sportives' comme finir à faire du toboggan, tes petites visites dans mon bureau dans l'aprem… tous ces moments ont rendu encore plus agréable ces trois années !

Je remercie également, de manière un peu plus générale, toutes les filles de l'étage : Sylvie, Claire, Axelle, Agnès, Nathalie, Marie, Nadine, Laurence, Maggy, Qingli (et re-Manue) pour m'avoir fait passer de supers moments à vos côtés, les pauses cafét', les discussions, tout le soutien que vous avez pu me montrer aussi bien dans mon travail que dans ma vie personnelle… Spéciale dédicace à Claire, pour ton aide l'été 2011, ces casse-têtes géants de ACTG qu'on essayait de résoudre à en rester tellement tard le soir qu'on en déclenchait les alarmes ☺, et pour tous les bons moments qu'on a pu partager! Maggy, je te remercie aussi, pour m'avoir supporté 3 ans dans ton bureau… toi qui aime tant l'ordre, je pense avoir mis tes nerfs à rude épreuve avec ma façon si particulière de ranger mes affaires ! Merci d'avoir été là aussi pour tout le reste ☺ Je n'oublie pas non plus le coq de la basse-cour de parasito : j'ai nommé Albert! Merci pour tout, les fous-rires, ton aide pour R, pour ArcGis et tout le reste !

Je remercie aussi, le gang ‘Cervoiserie’, et le gang ‘piscine du midi’… C’était vraiment chouette tous ces moments partagés ☺

Je remercie Chloé, pour le super stage que tu as pu faire, ta bonne humeur constante. Merci pour la super stagiaire que tu as pu être et la super amie que tu es devenue !!

Je remercie également les autres d’jeuns du labo, pour leur présence, leur soutien, et les bons moments que j’ai pu passer en leur compagnie ! Puis également tous ceux qui sont passés par les murs du labo, avec qui j’ai passé de supers moments de boulot, de collecte de tique (KassDédi à Hélène Et la Team Tiques Véto !) de discussions, de sorties et j’en passe… Je pense entre autres à Françoise (mon soutien lyonnais), Ionut (à notre super semaine à Barcelone entre autre), Hélène (des collectes de tiques aux supers soirées en passant par les urgences de Toulouse), Benjamin et Olivier, Floriane, Roxane, Lekan, Imane, Chloé, Soura…… et j’en oublie surement involontairement!

Je remercie également toutes les personnes extérieures au labo, qui m’ont apporté tellement d’aide, de discussions intenses, guidé dans mes choix….
Pierre et Olivier, pour l’aide informatique de folie et m’avoir aidé à dompter Linux et fait apprécier l’écriture de lignes de commandes incompréhensibles, à moi qui n’y connaissait absolument rien il y a encore deux ans !!
Alexandra et Sophie de la plateforme Biogenouest, pour toute votre aide, nos prises de tête ‘pool qui pool quoi pool comment’.
Véronique et Lydia notamment, de Gentyane, pour toute votre aide, vos conseils pour le génotypage !!
Yann pour ton aide, tu es un m*ArcGis*sien !! Toutes les personnes du projet OSCAR pour toutes les discussions scientifiques et également les supers moments passés sur le terrain !!

Tous les copains aussi, je vous dis un énorme merci !!!! Aux copains de toujours : Gabi, Déb’, Val, Marie, Yann, Rom, Erhan !!! A Rémi, à qui je dois tellement ! Aux copains ‘locaux’ qui m’ont fait apprécier Nantes (et ce n’était pas gagné à la base), Tibo, LNeuh, Johann, Cédric, Antoine, Tiphaine, Rémi et tous les autres… j’ai passé de supers moment avec vous !!!

Toute ma famille qui m’a soutenue, Ma sœurette Tiphaine, les tontons et les tatas, les cousins, vous êtes formidables !!! Et une pensée pour ma Satine, à qui j’aurais aimé raconter la fin de l’histoire de mes tiques !!

Et pour finir, le meilleur… mes géniteurs !!!

Pour citer Grand corps malade…

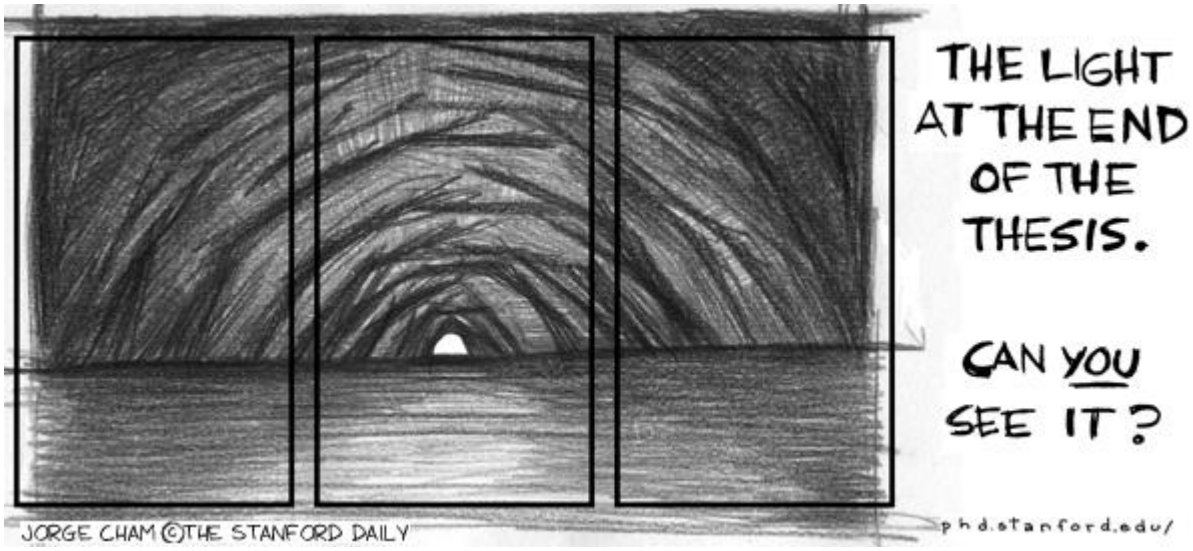
Moi mon père et ma mère sont carrément Hors-pairs

Et au milieu de ce récit

Je prends quelques secondes je tempère

Pour dire à mon père et à ma mère merci

Merci pour tous mes parents chéris !! Votre soutien depuis toujours, tout ce que vous m'avez appris! Et promis, j'arrête les études !! Je vous aime fort !!!!



Un grand MERCI général à vous tous, car c'est grâce à vous tous que ce travail a été concrétisé. Pour tous les moments 'fun' passés avec vous, pour les coups de pied au c** que vous avez pu me donner au moment nécessaire, à votre soutien sans faille dans les moments difficiles (et il y en a eu …), à me montrer que même quand je voyais tout en noir, du positif se cachait dedans… Bref, vous avez illuminé mon tunnel de thésarde…

Je n'oublie pas non plus de remercier la bière et le chocolat d'exister… Vous m'avez bien aidée ces derniers jours !!!

Sommaire

Chapitre 1 : Introduction générale

Chapitre 2 : Développement de marqueurs génétiques (SNPs) dans le génome d'Ixodes ricinus

Chapitre 3 : Analyses de génétique des populations d'*I. ricinus* à l'échelle du paysage

Chapitre 4 : Discussion générale, Perspectives & Conclusions

Annexes 213

Liste des figures

Chapitre 1

Chapitre 2

Figure 2.10 : Distribution du nombre de contigs générés par MIRA3 en fonction de la profondeur de séquençage obtenue lors de l'assemblage.

Figure 2.11: Exemple d'utilisation du graphe de De Bruijn dans DiscoSnp avec des k-mers de taille k=4. Un premier k-mer 'ATCT' correspond au premier nœud, ce k-mer est chevauchant avec deux autres qui présentent une différence sur la dernière base 'TCTT' versus 'TCTG'. De ce fait une bouche s'ouvre. De proche en proche les k-mers successifs se chevauchant sur k-1 bases sont identifiés jusqu'à obtenir un second nœud 'AGCT' qui permet de refermer la 'bouche'. Ainsi deux séquences, correspondant aux deux chemins sont reconstruites avec un SNP identifié.

Figure 2.12 : Fréquence relative de k-mers unique dans le jeu de données en fonction de la taille des k-mers.

Figure 2.13: Exemple de sortie à l'issue du module KisSnp du logiciel DiscoSnp pour deux SNPs.

Figure 2.14 : distribution du nombre de SNPs identifiés (tronquée à une profondeur de 100) par DiscoSnp en fonction de la profondeur de séquençage de chaque SNP.

Figure 2.15 : Représentation graphique de la profondeur de séquençage par SNP dans le jeu de données initial des 1768 SNPs (a) et dans le jeu de données des 384 SNPs finaux (b).

Figure 2.16 : (a) Le système Biomark HD de Fluidigm ; (b) exemple de puce IFC de Fluidigm. Ces puces disponibles en 48*48 ou 96*96 permettent de charger d'un côté de la puce 48 ou 96 échantillons – puits ADN (selon le format de la puce) -, et de l'autre côté 48 ou 96 amorces KASPar (ou Taqman)- puits SNPs-. Un réseau de micro canaux relie l'ensemble des puits 'ADN' et des puits 'SNPs' jusqu'à des micro-chambres réactionnelles dans lesquelles la réaction qPCR a lieu.

Figure 2.17 : Exemple de design d'amorce pour un SNP et des vérifications effectuées relatives à l'hybridation entre amorces.

Figure 2.18 : Photo de gel d'électrophorèse des produits d'amplification WGA pour 7 échantillons d'ADN.

Figure2.19 : Fluorogramme (Scatter plot) obtenu suite à la lecture au LC480 du génotypage de 8 individus pour 12 SNPs. Les points verts et bleus correspondent à des individus homozygotes, les points rouges correspondent à des hétérozygotes, les points roses à des points de génotypage non assignés dû à une lecture ambiguë et les points gris à des points de génotypage de témoins et/ou non amplifié ou présentant un signal trop faible de lecture.

Figure 2.20 : A gauche, Fluorogramme (Scatter Plot) permettant de visualiser les 2304 point de génotypage selon l'intensité de la fluorescence émise. A droite, le plan de la puce, où les individus correspondent aux lignes et les SNPs aux colonnes. Les homozygotes sont présentés en vert (YY) et rouge (XX) et les hétérozygotes (XY) en bleu.

Figure 2.21 : Exemple d'un Scatter Plot pour une puce 96*96 représentant 9216 points de génotypage.

Figure 2.22 : Exemple d'une représentation graphique synthétique pour une puce 96*96 représentant 9216 points de génotypage (points bleus : hétérozygotes ; points verts et rouges : homozygotes ; points gris : données manquantes [signal trop faible]).

Figure 2.23 : Exemple d'un graphique de fluorescence pour un SNP sur une puce (96 points de génotypage)

Figure 2.24 : Le même individu (L008-T10) mis en surbrillance dans le premier cas sur la puce48 où il apparait 'YY' et dans le 2ème cas sur la puce96 où il apparait 'XX'.

Figure 2.25 : Relation entre le nombre de différences observées entre les deux réplicats des 27 individus génotypés en fonction de la quantité d'ADN initiale de chaque individu pré-WGA.

Figure 2.26 : Histogramme présentant le nombre de SNPs concernés par l'observation de loci différents entre les 2 génotypages.

Figure 2.27 : Représentation du pourcentage de données manquantes observées entre les 5 différentes plaques d'extraction (et génotypage) d'ADN (plaque 1, 3, 4, 5, 6 du premier run de génotypage, plaque 24 du deuxième run) pour 31 SNPs d'une même plaque.

Figure 2.28 : A gauche, fluorogramme issu du génotypage de la puce 8 ; à droite, image synthétique de la lecture de la puce 8 par le logiciel d'analyse Fluidigm. On observe dans les deux cas, une distribution anormale des points de génotypage.

Figure 2.29 : Image synthétique de la puce 11 à la sortie du logiciel d'analyse.

Figure 2.30 : Image de la puce 11 suite à mon interprétation manuelle.

Figure 2.31 : Exemple du typage d'un SNP sur la puce 9 après analyse et assignation manuelle des points de génotypage.

Figure 2.32 : Image de la puce 9 à la sortie du logiciel d'analyse, qui correspond également à l'analyse effectuée manuellement.

Figure 2.33 : Graphique représentant le nombre de lignes (=individus) par puce présentant un stretch composé d'au minimum une répétition entre 4 et 40 fois du motif 'XY/No Call' sur les 96 points de génotypage de chaque individus).

Figure 2.34 : histogramme représentant le nombre de données manquantes par individu et par classe de pourcentage de données manquantes. Le cadre vert représente l'ensemble des individus conservés pour la suite (individus portant moins de 40% de données manquantes).

Figure 2.35 : Exemple d'analyse de la ségrégation des allèles à la génération suivante. L'identifiant de chaque locus SNPs est indiqué dans la première colonne. Les colonnes suivantes correspondent aux génotypes des dix descendants analysés. Dans la colonne suivante figure le génotype concaténé des deux parents puis le nombre de données manquantes au sein de la descendance, puis des génotypes des deux parents. Pour finir est répertorié le nombre de descendants présentant une ségrégation des SNPs mendélienne et non-mendélienne (sans faire l'hypothèse d'un allèle nul chez un ou les deux parents). Dans cet exemple les parents sont tous deux homozygotes pour les deux allèles du SNP (XX ou YY), de ce fait la descendance doit être hétérozygote pour l'ensemble des marqueurs présentés dans cet exemple.

Figure 2.36 : Histogramme représentant l'ensemble des SNPs en fonction du nombre d'individus parmi la descendance des 5 croisements réalisés (N=47). Le cadre vert représente l'ensemble des individus conservés pour la suite (paragraphe suivant). En gris (partie supérieur de la barre d'histogramme correspondant à 0 individu non mendélien) sont représentés les loci inclus des données manquantes chez les parents.

Figure 2.37 : Histogramme présentant la répartition des SNPs en fonction de leur pourcentage de données manquantes sur l'ensemble des 408 individus conservés. Le cadre vert représente l'ensemble des SNPs conservés pour la suite (SNPs portant moins de 25% de données manquantes).

Figure 2.38 : Histogramme présentant la répartition des SNPs en fonction des fréquences alléliques minimales sur l'ensemble des 408 individus conservés. Le cadre vert représente l'ensemble des SNPs conservés pour la suite (SNPs présentant plus de 5% de l'allèle minimum).

Figure 2.39 : Histogramme présentant la répartition des 143 SNPs en fonction des pourcentages de descendants mendélien dans l'ensemble des 47 descendants. Le cadre vert représente l'ensemble des SNPs conservés pour la suite (SNPs présentant plus de 50% de descendants mendéliens).

Figure 2.40 : Représentation des individus dans le plan factoriel tridimensionnelle des 7 populations (les noms correspondent aux noms des populations indiqués dans le tableau 2.10).

Figure 2.41 : Résultats du logiciel STRUCTURE pour un nombre de sous-populations inférées à K=2, 3,5 et 7. Les sous populations présentées de 1 à 7 correspondent respectivement aux populations C212, C214, C243, C30, H4I, CG et ZAA.

Chapitre 3

Figure 3.1 : **(a)** Origine des 60 individus d'*I. ricinus* étudiés pour l'étude de Noureddine *et al.* (2011). L'aire de répartition *d'I. ricinus* est représentée en vert. Les couleurs des points correspondent aux différentes échelles considérées : rose (locale), bleu (régionale), orange (européenne), vert (Iran), rouge (Afrique du nord). **(b)** Arbre phylogénétique (NeigbourJoining) entre les 60 individus séquencés lors de l'étude à partir des séquences concaténées des 5 gènes polymorphes (arbre de gauche). L'arbre de droite correspond aux séquences d'un seul gène (*TrospA)* pour lequel 20 individus supplémentaires de Tunisie ont été rajoutés. Les couleurs représentant les individus correspondent à celles utilisées pour l'origine des échantillons dans la figure 3.1.a.

Figure 3.2 : Analyse en composantes principales basée sur le polymorphisme de huit populations d'*I. uriae* échantillonnées dans différents sites européens (chaque point du graphique représentant un site). Les tiques ont été prélevées dans des nids de différentes espèces hôtes (Mouette tridactyle, Guillemot de troil, Guillemot d'hornoya et Macareux moine). Sur ce graphique on voit bien que les différents échantillons se regroupent essentiellement par espèces hôtes (point de même couleur) et non selon leurs localisations géographiques.

Figure 3.3 : Représentation schématique des éléments de base d'un structure paysagère : le patch (habitat), le corridor et la matrice (D'après Burel & Baudry, 2003)

Figure 3.4 : Situation géographique de la Zone Armorique Atelier (ZAA)

Figure 3.5 : cartographie de ZAA. L'ensemble des lignes (bleu et jaune) correspond aux 89 lignes de collecte échantillonnées dans le cadre du projet OSCAR. En bleu sont représentées les 71 lignes de collecte qui correspondent à celles analysées dans cette présente étude.

Figure 3.6 : Représentation du nombre de tiques par ligne de collecte.

Figure 3.22 : Analyse factorielle des correspondances (AFC), réalisée avec GENETIX, sur les fréquences alléliques des populations des 9 biotopes définis. Chaque couleur représente une population, étant donné le pattern observé, il est inutile de définir dans ce manuscrit la correspondance.

Figure 3.23 : Résultat du logiciel STRUCTURE pour un nombre K de 9 sous-populations, les différents clusters de 1 à 9 correspondent aux identifiants défini dans le tableau 3.6 pour chacun des 9 biotopes.

Figure 3.24 : Répartition des 18 différents clusters géographiques dans le bocage dense (BD) et les secteurs forestiers (CF et LF)

Figure 3.25 : Répartition des 7 différents clusters géographiques dans le bocage ouvert (BO)

Figure 3.26 : Histogramme représentant l'ensemble des valeurs de θ calculées à l'aide du logiciel Genepop 4.0.2 pour l'ensemble des clusters géographiques représentés par plus de 6 individus. Les valeurs observées sont regroupées en classe comprenant l'ensemble des valeurs supérieures

Figure 3.27 : Estimation des *Fis* calculés avec Genepop 4.2 pour les 34 lignes, comportant plus de six individus

Figure 3.28 : Histogramme représentant les estimations de valeurs de *Fst*, calculées avec Genepop4.2, pour l'ensemble des 34 lignes représentées par plus de 6 individus prises deux à deux.

Figure 3.29 : Analyse factorielle des correspondances (AFC), réalisée avec GENETIX, sur les fréquences alléliques des deux groupes d'individus porteurs de la bactérie *A. phagocytophilum* (bleu) et non porteurs (jaune)

Figure 3.30 : Analyse factorielle des correspondances (AFC), réalisée avec GENETIX, sur les fréquences alléliques des 15 groupes d'individus (porteurs et non porteurs de la bactérie *A. phagocytophilum*, N=2) regroupés selon leurs lignes de collecte, chaque couleur représentant un groupe différent

Figure 3.31 : Analyse factorielle des correspondances (AFC), réalisée avec GENETIX, sur les fréquences alléliques des 2 groupes d'individus porteurs (bleu) ou non porteurs (jaune) de la bactérie *Borrelia spp*.

Figure 3.32 : Evolution de l'estimation du *Fis* moyen en fonction de l'échelle investiguée lors de l'analyse. Les barres noires présentent les écart-types des moyennes calculées

Liste des tableaux

Chapitre 2

Chapitre 3

population unique prise ensemble. Les estimations ont été calculées à l'aide du logiciel Genepop 4.2 ; N= nombre d'individus génotypés ; H_{att}= hétérozygotie attendue et non biaisée ; H_{obs} = hétérozygotie observées ; *Fis* = indice de fixation intra-population

Tableau 3.4 : Estimation des *Fst* entre les 4 différents secteurs

Tableau 3.5: Matrice des estimations de *Fst* entre les différentes lignes de collecte du secteur CF représentées par plus de 3 individus

Tableau 3.6 : Hétérozygotie des groupes de tiques de la zone atelier séparées selon les biotopes; H_{obs} = hétérozygotie observées ; H_{att}= hétérozygotie attendue et non biaisée ; *Fis* = indice de fixation intra-population

Tableau 3.7 : Matrice des estimations de *Fst* entre les différents biotopes, les numéros de groupe allant de 1 à 9 correspondent aux identifiants des différents biotopes indiqués dans le tableau 3.2

Tableau 3.8 : Hétérozygotie des populations de tiques de la zone atelier séparées selon les clusters géographiques; *Fis* = indice de fixation intra-population

Tableau 3.9 : Matrice des estimations de *Fst* pour l'ensemble des clusters géographiques représentés par plus de 6 individus. Les numéros identifiant chacun des clusters correspondent aux identifiants attribués dans le tableau 3.8

Tableau 3.10 : Analyse Moléculaire de la Variance (AMOVA) à 3 niveaux hiérarchiques (secteurs, lignes, individus).

Tableau 3.11 : Résumé des différents taux d'infection en fonction des 3 agents pathogènes étudiés dans le jeu de données global du projet OSCAR et dans la subdivision réalisée par le génotypage.

Liste des abréviations

ADN : Acide Desoxyribonucléique
AFC : Analyse Factorielle des Correspondances
AFLP : Amplification Fragment Length Polymorphism
AMOVA : Analysis of MOlecular Variance
BD : bocage dense
BO : bocage ouvert
CF : cœur de forêt
DHPLC : *Denaturing High Performance Liquid Chromatography*
DM : Données Manquantes
Fis : indice de fixation des individus dans les sous-populations
Fst : indice de différenciation de Wrigt
GPV : Graphics Processing Unit
HTS : High Throughput Sequencing
H_{att} *: hétérozygotie attendue*
H_{obs} *: hétérozygotie observée*
HW : Hardy Weinberg
IDT : Integrated Dna Technologies
IFC : micro-Circuits Intégrés de Fluidique
ILTER : International Long Term Ecological Research site
Indel: Insertion Deletion
LCPA : Least Cost Path Analysis
LF : lisière de forêt
MID : Multiplex Identifier adaptors
NCBI : National Center for Biotechnology Information
NGS : Next Generation Sequencing
Pb : paires de base
PCA : *Principal Component Analysis*
PCR : Polymerase Chain Reaction
qPCR : quantitative Polymerase Chain Reaction
RAD seq : Restriction site associated DNA sequencing
RAPD : Random of Amplication of Polymorphism Dna
RFLP : Restriction Fragment Length Polymorphism
RNA seq : RNA Sequencing
RRL : Reduced Representation Librairies
SBS : Sequencing By Synthesis
SIG : Système d'Information Géographique
SNP : Single Nucleotide Polymorphism
SOLID : Sequencing by Oligonucleotide LIgation and detection
SRA : Sequencing Reads Archive
SSCP : Simple-Strand Conformation Polymorphism
WGA: Whole Genome Amplification
ZAA : Zone Atelier Armorique

Valorisation du travail de thèse

Publication:

- **Quillery E**., Quenez O., Peterlongo P., and Plantard O. Development of genomic resources for the tick *Ixodes ricinus:* isolation and characterization of Single Nucleotide Polymorphisms (2013) Molecular Ecology Resources **doi: 10.1111/1755-0998.12179**

- Raluca Uricaru R, Rizk G, Lacroix V, **Quillery E**, Plantard O, Chikhi R, Lemaitre C and Peterlongo P . *in prep* Reference-free detection of genotypable SNPs.

Communications orales:

Quillery E & Plantard O (2013) Isolation and characterization of SNPs from the genome of *Ixodes ricinus*. JAS INRA, Cap d'Agde, France

- **Quillery E** & Plantard O (2013) Isolation and characterization of SNPs from the genome of *Ixodes ricinus.* 13ème ICLB, Boston, États-Unis

- **Quillery E** & Plantard O (2013) Isolation of highly resolutive genetic markers (SNPs) from the *Ixodes ricinus* genome for population genetics studies. 2ème réunion annuelle du projet EDENEXT, Barcelona, Espagne

Posters:

- Quillery E & Plantard O (2012) **Development of SNP markers in the genome of *Ixodes ricinus* for population genetics studies at the landscape scale, E-SOVE, Montpellier, France**

- Quillery E & Plantard O (2011) **Characterizing Ixodes ricinus dispersal using landscape genetics,** TTP7, Zaragozza, Espagne

Chapitre 1

Introduction générale

I. Contexte général

A. Les maladies vectorielles

Avec environ un tiers des maladies transmises via un vecteur, les maladies vectorielles figurent parmi les plus préoccupantes en santé humaine et animale, tant par la morbidité que par la mortalité qu'elles entraînent. C'est ainsi que certaines de ces maladies comme le paludisme, la dengue, les leishmanioses et bien d'autres, représentent les fléaux les plus graves pour l'humanité. Bien que l'on assimile souvent les maladies vectorielles à des maladies « tropicales », elles posent également de graves problèmes sanitaires sous nos latitudes tempérées.

C'est le cas des maladies transmises par les tiques, dont l'importance en santé publique augmente depuis les années 1980 avec l'émergence de la maladie de Lyme notamment (Fritz 2009; Coipan *et al.* 2013).

L'émergence et la diffusion d'une maladie vectorielle met en jeu un système complexe résultant des interactions entre l'ensemble des différents acteurs, à savoir l'agent pathogène (microparasite à l'origine de l'infection), le vecteur et l'hôte vertébré (homme ou animal) (Gubler 1998). L'environnement influence fortement chacun de ces acteurs et les interactions entre eux (Keesing *et al.* 2010).

Bien que l'on assimile souvent le vecteur à une seringue vivante, il n'a pas, en réalité, qu'un simple rôle de transporteur : de nombreux agents pathogènes réalisent une partie de leur cycle de reproduction sexuée dans l'organisme vecteur (comme *Plasmodium falciparum* pour le paludisme, ou encore *Babesia spp.* pour les babésioses).

De plus, outre la transmission horizontale de l'agent infectieux entre les différentes stases du vecteur, il arrive qu'il puisse être également transmis *« verticalement »* à la descendance via la contamination des gonades et des œufs. Cela permet la multiplication et la survie de l'agent pathogène sans nécessité de multiplication dans un hôte vertébré. Le vecteur devient alors réservoir : c'est le cas de *Babesia spp.* chez sa tique vectrice (Iori *et al.* 2010).

B. Des vecteurs et des tiques

Les tiques, acariens hématophages, sont considérées comme les deuxièmes vecteurs de maladies humaines et animales au niveau mondial, après les moustiques. Elles transmettent la plus grande diversité d'agents pathogènes (900 espèces réparties en 31 genres) (Parola & Raoult 2001; Pérez-Eid 2007).

Les tiques sont retrouvées quasiment sur l'ensemble de la planète. Comme c'est le cas pour de nombreux vecteurs, les régions tropicales (Afrique, Amérique ou Asie) sont les régions où l'on trouve le plus de tiques, de genres très diversifiés tels que *Rhipicephalus*, *Haemaphysalis*, *Hyalomma* ou *Amblyomma*. Dans les régions plus froides et tempérées, ce sont les tiques du genre *Ixodes* ou *Dermacentor* qui sont principalement observées.

Les tiques posent de graves problèmes sanitaires, d'une part par les effets néfastes directs qu'elles occasionnent (spoliation sanguine,…) mais également par leurs capacités à transmettre un nombre important d'agents pathogènes.

C. Maitrise des maladies vectorielles : un enjeu majeur

Depuis de nombreuses années, du fait de l'augmentation des cas de maladies vectorielles en lien avec les changements globaux, différentes méthodes de lutte contre ces maladies ont été développées.

Contre les maladies à tiques, 2 stratégies principales peuvent être appliquées :

- la lutte contre les agents pathogènes mais les tiques pouvant en transmettre plusieurs, il paraît difficile de pouvoir contrôler l'ensemble

- la maîtrise des populations de vecteurs

La lutte chimique constitue une des options possibles grâce à des molécules acaricides telles que l'arsenic, des produits organochlorés (DDT, Toxaphène…), des organophosphorés, ou des amidines et pyrethrinoïdes appliqués plus récemment. Cependant, malgré de réels progrès dans les modes d'administration et dans le contrôle de la toxicité résiduelle, il apparait régulièrement des phénomènes de résistance. De ce fait d'autres méthodes de lutte sont étudiées telles que :

- le développement de nouveaux produits acaricides

- la lutte biologique, qui consiste à détruire les tiques (phase parasitaire et libre) par des prédateurs ou des parasitoïdes.

-la lutte écologique, qui consiste à modifier le biotope de la tique pour rendre plus difficile (voire impossible) la réalisation de son cycle de développement.

-la vaccination anti-tique en induisant une réaction immunitaire acquise de l'hôte vis-à-vis des tiques.

D. Les outils de la génétique des populations

Afin de mettre en place et d'optimiser des techniques de contrôle ou de lutte les plus efficaces possibles contre un vecteur, il est nécessaire de comprendre et de définir la structuration de ses populations, sa capacité de dispersion, le niveau de sa variabilité génétique et la distribution dans l'espace de cette variabilité génétique. Ceci permet de mieux appréhender comment ce vecteur interagit avec les agents pathogènes qu'il transmet et ses hôtes.

Par exemple, afin de tester l'efficacité d'une molécule acaricide contre une espèce de vecteur, il est nécessaire d'utiliser un panel de populations représentatif de sa variabilité génétique. Ces connaissances sont aussi importantes pour la lutte contre les agents pathogènes vectorisés. En effet, la dispersion de ces agents est probablement largement conditionnée par celle de ses vecteurs, en particulier si une part importante de leur cycle se déroule en leur sein, comme c'est le cas par exemple pour *B. divergens* chez *I. ricinus.* La variabilité génétique intra ou inter-populationnelle des tiques peut aussi avoir des conséquences sur des traits de vies impliqués dans la vection de ce pathogène (compétence vectorielle, capacité à multiplier le parasite…). La génétique des populations doit permettre de décrire la distribution de ces variants et d'identifier les forces évolutives à l'origine des variations dans le temps et dans l'espace des fréquences alléliques. La mesure de la dispersion par les outils de la génétique des populations apparaît donc comme une des approches les plus pertinentes car elle permet d'une part de mesurer la dispersion efficace, c'est à dire celle des individus ayant effectivement contribués (de par leur reproduction) à la génération suivante et d'autre part de mettre en évidence l'existence de flux de gènes.

Par ailleurs, la modélisation de l'épidémiologie des maladies vectorielles constitue un outil important pour tester l'efficacité de méthodes de lutte. Or la dispersion des vecteurs apparaît être un paramètre central dans l'élaboration de ces modèles. Jusqu'à une période récente, l'épidémiologie ne prenait en compte que la présence/absence d'un agent pathogène dans un espace donné à un moment donné. Des outils, notamment de biologie moléculaire, sont maintenant disponibles pour distinguer des variants génétiques dont certains peuvent correspondre à des variants phénotypiques, de virulence différente par exemple. La prise en compte, dans les modèles épidémiologiques, de cette variabilité génétique et phénotypique peut permettre une meilleure compréhension de la progression dans l'espace et dans le temps de la maladie. C'est l'objet d'une approche récente appelée « épidémiologie moléculaire ».

II. Modèle d'étude, la tique *Ixodes ricinus*

A. Phylogénie et classification

Les tiques sont des arthropodes appartenant à la classe des *Arachnida* et à l'ordre des Acariens parasitiformes. Au sein de cet ordre, regroupant environ 80 familles d'Acariens, les tiques se répartissent en trois familles : les *Argasidae* également appelées 'tiques molles', les *Ixodidae* appelées 'tiques dures' et les *Nuttallielidae*.

Ixodes ricinus, est une tique dure appartenant à la famille des *Ixodidae*. Au sein de cette famille, le genre *Ixodes* se trouve dans la sous-famille des *Ixodinae*. Ce genre *Ixodes* compte à ce jour plus de 200 espèces selon Horak *et al*. (2002), dont la tique *Ixodes ricinus*.

B. Biologie et écologie

1. Cycle de vie

Ixodes ricinus a un cycle de vie à 3 stades actifs (larve, nymphe et adulte), durant entre 2 à 6 ans, en fonction de la rencontre des hôtes, du phénomène de diapause et des conditions climatiques environnantes (Sonenshine 1993; Gray 1998; Eisen *et al.* 2002) (Figure 1.1).

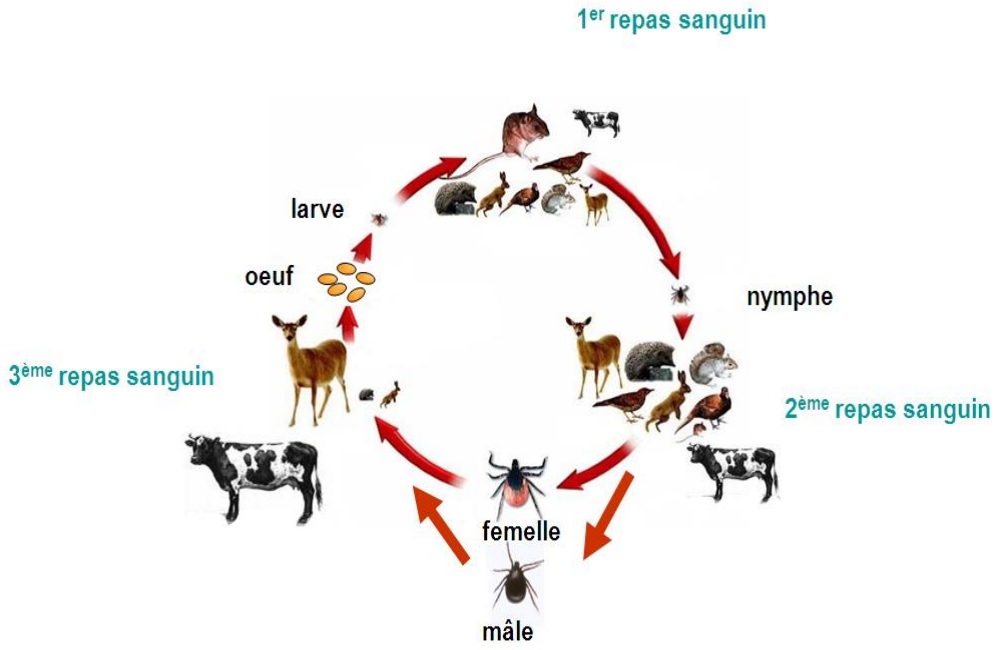


Figure 1.1 : Cycle de développement d'*Ixodes ricinus.* La taille représentée des différents hôtes est proportionnelle à leur importance lors des différents repas sanguins *(d'après J. Gray et B. Kaye)*

Les stades larvaire et nymphal ne présentent pas de dimorphisme sexuel et d'orifice génital à l'inverse des adultes. Les femelles d'*Ixodes ricinus* sont de taille plus importante que les mâles (4 mm versus 2 mm) et morphologiquement, les femelles ont un scutum qui recouvre uniquement la moitié antérieure du corps alors qu'il couvre l'ensemble du corps des mâles (Figure 1.2).

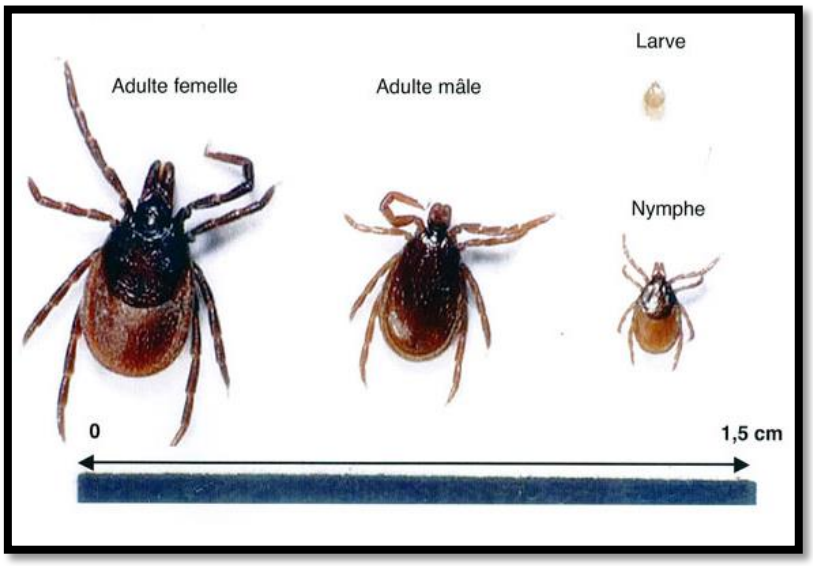


Figure 1.2 : Les différents stades de vie d'*Ixodes ricinus*, larve, nymphe adulte (mâle et femelle), *(source : collection Philippe Parola, extrait de la 16ème conférence de Consensus en thérapeutique anti-infectieuse, 2006)*

Ixodes ricinus est considérée comme un parasite exophile triphasique, car elle doit réaliser un repas de sang sur un hôte vertébré à chaque stade pour se développer (à l'exception des mâles qui se gorgent rarement au stade adulte). De plus, elle est polytrophe, affichant des préférences pour des espèces hôtes différentes en fonction de son stade. *Ixodes ricinus* posséde une large variété d'hôtes vertébrés, avec plus de 300 espèces hôtes répertoriées, allant des mammifères (micromammifères, cervidés, ruminants) aux oiseaux et aux lézards (Anderson 1991). On considère que les larves et les nymphes privilégient, pour leurs repas, les petits mammifères, les oiseaux ou les lézards (Anderson 1991), alors que les adultes préfèrent les grands mammifères, comme des bovins ou des chevreuils. Cependant, les tiques sont considérées comme des ectoparasites opportunistes qui parasitent les hôtes les plus abondants localement (Klompen *et al.* 1996).

Pour réaliser son repas sanguin, la tique est à l'affut sur la végétation et attend qu'un hôte passe à proximité d'elle. Une fois sur un hôte, elle se déplace activement afin de trouver une partie du corps où les vaisseaux sanguins sont le plus affleurant afin de faciliter son repas sanguin. Elle peut alors se

fixer et ingérer le sang. Le gorgement est plus ou moins long en fonction des stades ; les larves se gorgent un ou deux jours, les nymphes trois à quatre jours et les femelles une semaine. Une fois le repas sanguin achevé, les tiques se détachent et se laissent tomber au sol où, dans la litière, elles réaliseront leurs mues ou leur ponte.

2. Distribution et habitat

Comme nous pouvons le voir sur la Figure 1.3, *Ixodes ricinus* est présente du Portugal à la Russie, et de la Scandinavie au nord du Maghreb (Tunisie, Algérie, Maroc) (Gern & Humair 2002).

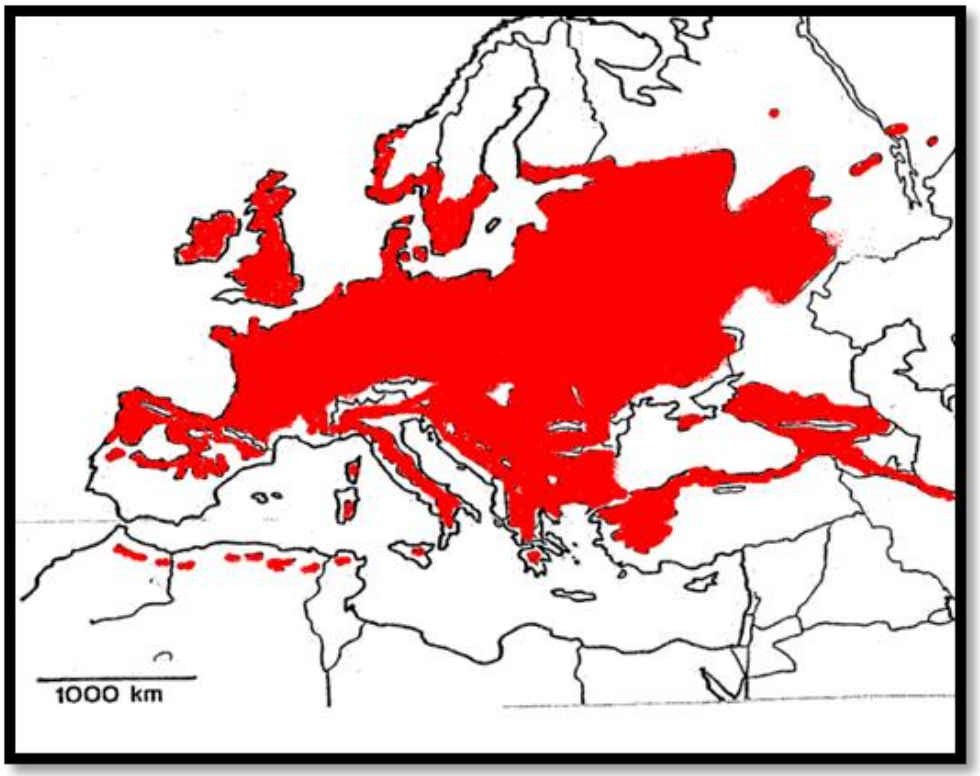


Figure 1.3 : Aire de distribution d'*Ixodes ricinus* en Europe (représentée en rouge)

Bien que cette espèce de tique possède une aire de répartition assez vaste, sa distribution est très hétérogène, une conséquence de son écologie stricte. En effet, *Ixodes ricinus* est inféodée à des microenvironnements lui permettant une bonne hydratation : on la retrouve dans des milieux forestiers, des haies, des landes, dès lors que la végétation maintient un microclimat avec une forte hygrométrie, de l'ordre de 80% (Kahl & Knülle 1988; Boyard *et al.* 2007). *I. ricinus* ne survit pas dans

un environnement où l'hygrométrie est inférieure à 70%, ce qui explique qu'on ne l'a retrouve pas sur le pourtour méditerranéen. Elle est également peu résistance au froid et sa limite de répartition n'excède pas les 1500 m d'altitude.

C. Micro-organismes vectorisé par *Ixodes ricinus*.

Bien qu'impactant directement la santé animale en induisant une spoliation sanguine, une diminution de la croissance (perte de poids) ou de la production laitière (lésions des trayons), ou encore des blessures (abcès, myiases), c'est leur rôle de vecteur d'agents pathogènes qui est le plus préoccupant.

Comme nous l'avons évoqué précédemment, les tiques transmettent divers pathogènes, notamment des virus, des bactéries, des protozoaires et même des nématodes.

Ainsi, parmi les agents infectieux transmis par *I. ricinus*, nous pouvons citer pour leur rôle en santé humaine, le virus de l'encéphalite à tique que l'on retrouve principalement en Europe centrale, de l'Est et du Nord ; la bactérie spirochète *Borrelia burgdorferi*, agent de la maladie de Lyme, qui est présente dans toute l'Europe. Diverses rickettsies sont également vectorisées par *I. ricinus* comme *Rickettsia helvetica* (Parola et al. 2005) qui appartient à la famille des Rickettsiaceae ou *Bartonella henselae* (Cotte et al. 2008) qui appartient à la famille des Anaplasmataceae.

En santé animale, le protozoaire *Babesia divergens*, responsable de la piroplasmose bovine ou la bactérie *Anaplasma phagocytophilum*, responsable de l'anaplasmose granulocytaire peuvent être cités en exemple.

D. Structure génétique et dispersion

Malgré sa place de vecteur majeur en Europe, peu de choses sont connues actuellement sur la dispersion des tiques et leur variabilité génétique. Nous savons que les tiques ont un déplacement actif très limité, estimé par certains auteurs de l'ordre de quelques mètres (voir par exemple l'étude sur une espèce ayant une biologie similaire, *Ixodes pacificus* (Lane *et al.* 2009)).

Du fait de sa faible capacité de dispersion, les populations de tiques pourraient être fortement structurées dans l'espace. Mais étant donné la durée d'un repas de sang (plusieurs jours), les tiques pourraient être fortement dispersées par les hôtes sur lesquels elles se gorgent. Ainsi la structuration (spatiale et temporelle) et les dynamiques (dispersion, reproduction) des populations de tiques pourraient être intimement liées à celles de leurs hôtes. Ces éléments pourraient, par cascade,

affecter aussi la transmission des microparasites transmis par ces vecteurs (Gómez-Díaz & González-Solís 2010) .

E. Facteurs influençant une structure génétique à une échelle locale

C'est à une échelle locale que l'intensité des flux de gènes peut affecter le degré d'adaptation des populations d'agents pathogènes et ainsi permettre l'extension d'une maladie dans un lieu donné (Gandon & Michalakis 2002). Néanmoins aucune étude chez *I. ricinus* n'a été réalisée à une échelle fine (de l'ordre de quelque kilomètres carrés) et encore moins en prenant en compte les hôtes, alors que l'hétérogénéité du paysage pourrait être plus ou moins propice à ces vecteurs ou/et à leurs hôtes et à leurs déplacements.

En effet, si la structuration des populations de tiques et les flux de gènes qui en découlent seraient principalement dus aux mouvements des hôtes, et que les hôtes quant à eux, de par leur distribution et leurs mouvements dépendraient de l'agencement du paysage, le paysage influencerait la structuration des populations de tiques.

III. Objectifs de la thèse

Mon travail de thèse a donc un double objectif qui structure ce manuscrit. Dans un premier temps, nous avons développé de nouveaux marqueurs pour étudier la variabilité génétique d'*I. ricinus*. Il s'agit de SNPs (Single Nucleotide Polymorphisms). Le développement de ces outils a nécessité l'utilisation de techniques de séquençage haut débit et le recours à des outils bioinformatiques puissants. Dans un deuxième temps, nous avons utilisé ces marqueurs afin de décrire la structure génétique des populations de tiques à une échelle fine de quelques dizaines de km^2. Par la suite, nous avons mis en relation les données géographiques, les caractéristiques du paysage, afin de comprendre l'influence du paysage sur la structuration d'*Ixodes ricinus* dans les agro-écosystèmes. Ce travail de thèse s'est inscrit dans le projet ANR OSCAR (voir encadré).

Le projet OSCAR

Le projet OSCAR se déroule de 2012 à 2015 et a pour objectif d'explorer les conséquences des changements d'utilisation des terres à l'échelle du paysage sur le risque acarologique (correspondant à la densité de tiques - de l'espèce *Ixodes ricinus* - infectées par trois agents pathogènes vectorisés, traduisant localement le risque pour un hôte de contracter cet agent pathogène) à travers un outil de simulation cartographique basé sur un modèle de dynamique des populations de tiques spatialisées. Cette étude est réalisée dans deux zones ateliers correspondant à deux niveaux différents de fragmentation forestière : zone atelier armorique autour de Pleine-Fougères en Ille et Vilaine et zone atelier des Vallées et coteaux de Gascogne en Haute Garonne. L'hétérogénéité du paysage de ces sites est bien connue grâce aux Systèmes d'Information Géographique (SIG) qui y ont été développés, ainsi que les populations de chevreuils et les communautés de micromammifères qui y sont étudiées depuis de nombreuses années. La première étape consiste tout d'abord à analyser - à partir d'échantillonnages réalisés dans les deux sites - la distribution observée des tiques, de trois agents pathogènes (*Anaplasma phagocytophilum*, *Babesia divergens* et *Borrelia spp*) et des principaux hôtes domestiques (bovins) et sauvages (chevreuils, micromammifères). Les principaux facteurs biotiques et abiotiques permettant d'expliquer les patrons de distribution observés sont recherchés. Parallèlement, les mouvements des différents acteurs de ce système sont évalués par des approches directes (collier GPS sur les chevreuils) ou indirectes (génétique des populations de tiques et analyse de la variabilité génétique des agents pathogènes). Dans une seconde étape, un modèle spatialisé de dynamique des populations de tiques, intégrant explicitement les mouvements des hôtes, sera développé. Ce modèle permettra alors de simuler l'évolution du risque acarologique à l'échelle du paysage agricole en fonction de différents scenarios de changement d'utilisation des terres et de structure du paysage. Cet outil de simulation cartographique apportera des connaissances importantes pour l''adaptation de l'agriculture aux changements globaux, afin de limiter le développement des maladies à tiques.

Chapitre 2 :

Développement de marqueurs génétiques (SNPs) dans le génome d'*Ixodes ricinus*

I. Introduction

A. Les marqueurs génétiques, outils de la génétique des populations

La génétique des populations est une discipline née au début du 20ème siècle de la synthèse des théories de Mendel, Darwin et de biométriciens comme Fisher, Wright ou encore Haldane. Ils ont ainsi posé les bases conceptuelles de l'évolution de la variation génétique dans les populations par formalisation mathématique. Au cours de l'évolution d'une population, différentes forces, telles que la mutation, la migration, la sélection ou la dérive génétique agissent de façon simultanée. Ces forces évolutives dont certaines augmentent les variations -comme la mutation ou la migration-, d'autres au contraire les diminuent -comme la sélection ou la dérive, façonnent ainsi la constitution génétique d'une population. Comme la description et la compréhension de ces mécanismes augmentant ou diminuant la variabilité génétique au sein et entre populations sont d'un grand intérêt en évolution, les outils et concepts de la génétique des populations ont largement diffusé dans d'autres champs de la biologie.

Dans le cas des maladies vectorielles, il est essentiel de pouvoir définir et caractériser ce qu'est une population –d'agent pathogène, d'hôte ou de vecteur- afin de comprendre et prédire leurs dispersions, leurs dynamiques et les interactions qu'elles peuvent entretenir entre elles, ceci afin de prévoir les impacts épidémiologiques engendrés par d'éventuelles modifications biotiques ou abiotiques de leurs environnements et mettre au point des méthodes de lutte. Dans ce but, il est nécessaire de mesurer la diversité génétique dans les populations naturelles par l'estimation des fréquences alléliques et génotypiques.

Dans les années 1940, avec Ford notamment, les scientifiques, conscients de l'existence d'une grande variabilité entre et à l'intérieur d'espèces, se basaient sur des caractères phénotypiques (couleur, morphologie,…), seule forme de variabilité à laquelle ils avaient alors accès, afin de caractériser le polymorphisme présent entre ou au sein de population (Conn *et al.* 1940).

En 1953, avec la découverte de la structure en double hélice de l'ADN par Watson, Crick et Franklin, la voie de la biologie moléculaire fut ouverte, révolutionnant les champs d'investigation de la génétique des populations. Le développement des techniques de biochimie, de cytogénétique et de biologie moléculaire ont permis d'étudier la variabilité génétique à des échelles plus fines, jusqu'au niveau de la séquence d'ADN, permettant ainsi l'étude du polymorphisme génétique des régions non codantes qui ne pouvaient pas jusqu'alors être étudiées.

Ainsi, au fil du développement de nouvelles techniques de biologie moléculaire et des connaissances des génomes, de nombreux marqueurs génétiques ont été développés (Schlötterer 2004). Les

principaux utilisés en génétique des populations sont répertoriés dans le tableau 2.1. Les marqueurs génétiques peuvent être décrits comme représentatifs d'une position dans le génome: 'un locus' de taille variable (en fonction des types de marqueurs). Ces différents loci présentent différents variants, les allèles, partagés par l'ensemble des individus d'une même espèce ou d'espèces différentes qui sont transmis de génération en génération. La description de la distribution de ces variants permet d'identifier les forces évolutives à l'origine des variations dans le temps et dans l'espace des fréquences alléliques observées.

Plusieurs auteurs s'accordent à définir ce que doit être un 'bon' marqueur (Vienne 1998) : être polymorphe, sélectivement neutre, facilement observable et ce sans ambiguïté, être dispersé le long du génome, ne pas avoir d'effet pléiotropique ou épistasique, être codominant afin de distinguer les homozygotes et les hétérozygotes, ne pas être liés entre eux et pour finir être d'un coût de développement et de typage modeste.

Cependant, le choix du type de marqueur est primordial dans une étude de génétique des populations. Comme nous l'avons vu, il existe plusieurs types de marqueurs ayant des propriétés différentes, plus ou moins compatibles avec les objectifs d'une étude (Tableau 2.1).

Tableau 2.1 : Récapitulatif des principaux marqueurs moléculaires développés et utilisés en génétique des populations.

Marqueur	Date de 1ère utilisation	Description	Particularité des locus	Polymorphisme	Reproductibilité
allozymes	1966	formes variables d'une enzyme, codées par différents allèles à un même locus	codominants, faible neutralité, requiert du matériel biologique frais	faible	bonne
RFLP	1981	Restriction Fragment Length Polymorphism : digestion d'un fragment d'ADN par des enzymes de restriction	dominants, neutres ou sélectionnés selon le choix du locus	limité (dépend de l'enzyme)	bonne
Minisatellites	1989	répétitions en tandem de motif nucléotidique	codominants et neutres	élevé	bonne
Microsattelites	1989	répétitions en tandem de motif nucléotidique	codominants et neutres	elévé	bonne
AFLP	1990	Amplified Frangment Length Polymorphism: amplification de fragments digérés par des enzymes de restriction	sélectionné selon le choix du locus	limité (dépend de l'enzyme)	bonne
RAPD	1991	Random Amplification of Polymorphic DNA: amplification aléatoire de fragments d'ADN non choisis	principalement dominant, difficile à automatiser et analyser	faible	faible
SNP	1998	Single Nucleotide Polymorphism: polymorphisme nucléotidique ponctuel	codominants, neutres ou sélectionnés selon le choix du locus	Faible (pour un seul SNP)	bonne

Ainsi, les différents marqueurs utilisés peuvent être plus ou moins polymorphes par rapport aux processus biologiques que l'on souhaite étudier et à l'échelle à laquelle on se place. Par exemple, des marqueurs trop polymorphes pour des études de phylogéographie peuvent conduire à des problèmes d'homoplasie et à l'inverse un marqueur trop peu polymorphe n'apportera pas suffisamment d'informations pour permettre de conclure sur l'évaluation du rôle respectif des différentes forces évolutives.

B. Les marqueurs génétiques développés chez la tique Ixodes ricinus

Chez *Ixodes ricinus,* les rares études de génétique des populations sont basées sur un faible nombre de marqueurs (allozymes et des microsatellites).
En 1997, Delaye *et al.* ont développé 18 marqueurs allozymes (pour la contraction des termes 'allèle' et 'enzyme'). Ces marqueurs biochimiques sont des protéines et les différents variants d'un même allozyme se distinguent par leur taille et leur charge lors d'une électrophorèse. Du fait de leur faible coût et qu'ils soient relativement universels, les allozymes ont été fréquemment employés dans des études de populations de grande taille. Cependant ces marqueurs restent peu nombreux. Par ailleurs, les variants peuvent être fonctionnellement différents, ce qui les écarte de la neutralité souhaitée pour les études de génétique des populations. Enfin, dans le cas de l'étude de Delaye *et al.* (1997), sur les 18 marqueurs allozymes utilisés, seuls deux se sont montrés polymorphes.

Par la suite, les marqueurs microsatellites ont été mis au point, avec le développement d'un premier set de six marqueurs microsatellites (Delaye *et al.* 1998), puis Roed *et al*. (2006) en ont développé dix-sept. Puis en 2012, Noel *et al.* (2012) ont mis au point un nouveau set de neuf microsatellites. Ces marqueurs mettent en évidence un polymorphisme basé sur la variabilité du nombre de répétition d'un motif de deux à six nucléotides (Litt & Luty 1989; Taylor *et al.* 1989) générant des allèles de taille entre 50 et 500 pb.
De manière générale, leurs répartitions plutôt uniformes et leur caractère hypervariable (on peut parfois observer plusieurs dizaines d'allèles à un même locus), en font des marqueurs de choix en génétique des populations. Chez *I. ricinus* et chez de nombreuses autres espèces, ces marqueurs sont les plus fréquemment utilisés (de Meeûs *et al.* 2002, 2004; Kempf *et al.* 2009, 2010).

Cependant, chez *I. ricinus,* certains problèmes sont apparus lors de l'utilisation de ces marqueurs microsatellites. De Meeus *et al.* (2004), par l'analyse de la transmission des allèles chez huit familles

de tiques, ont permis de mettre en évidence une transmission non-mendélienne de trois loci (ainsi qu'un locus lié au chromosome sexuel X), explicable par la présence d'allèles nuls, la duplication du locus, une empreinte parentale ou un biais d'amplification en faveur de l'allèle le plus court.

Les mêmes problèmes ont été constatés dans les 17 marqueurs microsatellites développés par Roed *et al.* (2006). L'analyse de la descendance de cinq femelles (sans connaître le génotype du/des père(s)) indique la présence de biais à la transmission mendélienne chez tous les loci, pouvant être expliquée par la présence d'allèles nuls (sauf dans cinq familles et pour trois loci). Une autre étude basée sur dix de ces 17 loci a analysé la descendance de trois femelles et la présence d'allèles nuls a été également constatée (Hasle *et al.* 2008).

De plus, du fait de leurs fort taux de mutations (de l'ordre de 10^{-3}), les microsatellites peuvent présenter de l'homoplasie. Etant donné que les microsatellites correspondent à du polymorphisme de longueur, on peut ainsi retrouver indépendamment deux allèles de tailles identiques qui sont issus de mécanismes différents (pas d'identité par descendance).

Des difficultés techniques sont également rencontrées avec les loci microsatellites :

- des difficultés d'interprétation et d'assignation des génotypes liés au grand nombre d'allèles par locus, dont la longueur diffère parfois que de quelques nucléotides
- des problèmes de lecture, avec la présence de bandes avant ou/et après le pic principal.

Pour finir, bien que les microsatellites se montrent abondants dans un grand nombre de génome, chez *Ixodes scapularis*, il a été montré que ces marqueurs étaient présents en faible nombre (Fagerberg *et al.* 2001; Pannebakker *et al.* 2010). Ceci a également été observée chez deux autres acariens *Tetranicus urticae* et *Amblyseius fallacis* (Navajas *et al.* 1998) et chez des insectes, comme *Bombyx mori*, *Aedes aegypti*, et *Drosophila simulans* (Pannebakker *et al.* 2010).

De ce fait, au vu des problèmes rencontrés avec les marqueurs utilisés à l'heure actuelle chez *Ixodes ricinus*, il apparait nécessaire d'en développer de nouveaux. En 2003, Gordon Luikart a écrit: *« the ideal molecular approach for population genomics should uncover hundreds of polymorphic markers that cover the entire genome in a single, simple and reliable experiment. Unfortunately, at present there is no such approach*". Avec les avancées technologiques réalisées au cours des années 2000, et notamment l'apparition du séquençage haut débit, les SNPs (Single Nucleotide Polymorphisms) sont devenus des marqueurs de choix, répondant ainsi aux marqueurs rêvés par Luikart *et al.* (2003). Plusieurs approches qui permettent de découvrir, séquencer et génotyper des centaines et même des milliers de marqueurs à travers n'importe quel génome d'intérêt (espèces modèles ou non) en une seule étape (Stapley *et al.* 2010) sont maintenant disponibles.

C. Les SNPs, marqueurs d'avenir dans l'air du temps

1. Définition et propriétés des SNPs

Les SNPs (Single Nucleotide Polymorphisms) sont des substitutions d'un seul nucléotide qui se produisent à des positions ponctuelles spécifiques dans le génome (Figure 2.1), générant des allèles différents (Brookes 1999). Les SNPs correspondant à une substitution, un polymorphisme d'un seul nucléotide pourrait être en principe biallélique, triallélique ou tétraallélique. Cependant, de manière générale, les SNPs présentant 3 ou 4 allèles sont très rares. C'est la raison pour laquelle ils sont en général considérés comme des marqueurs bialléliques. Cette caractéristique s'explique par un taux de mutation faible (Nielsen 2000) induisant une probabilité quasi-nulle de double mutation sur un même locus. En moyenne, on trouve un SNP toutes les 500 à 1000 pb chez l'Homme et les SNPs forment plus de 90% du polymorphisme génétique humain (Venter *et al.* 2001).

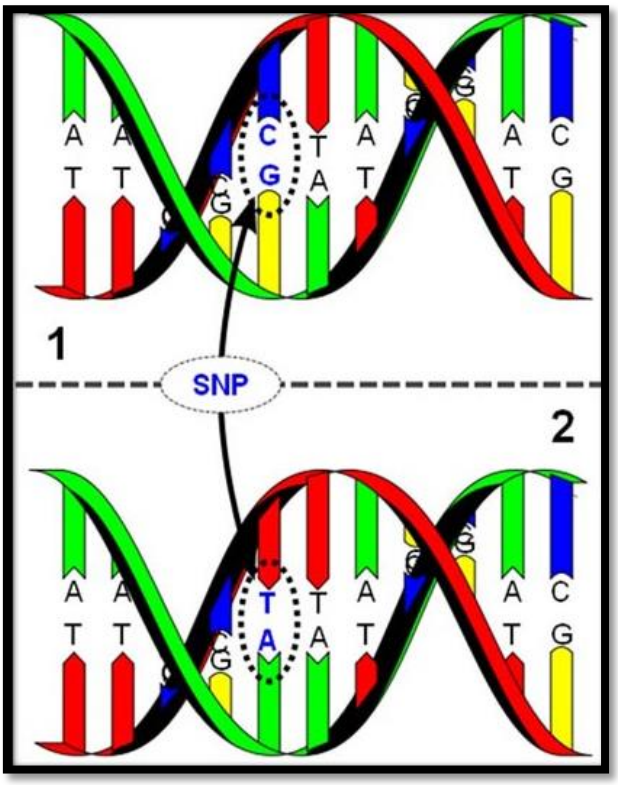


Figure 2.1 : Exemple d'un polymorphisme d'un seul nucléotide (SNP) ; la molécule d'ADN 1 diffère de la 2 par un seul nucléotide C/T. (source : http://blog.neogandalf.com/).

Par définition, la fréquence de l'allèle le moins représenté doit être au moins égale à 1%. Les petites insertions et délétions (indels), souvent comprises dans la dénomination SNP utilisée de manière très large, ne correspondent pas strictement à la définition ci-dessus puisqu'ils traduisent un autre type de mutation qu'une substitution. Ces polymorphismes sont aussi bien présents sur l'ADN nucléaire que sur l'ADN mitochondrial.

A l'origine, les SNPs sont issus de mutations survenues dans les cellules germinales et ayant échappé au système de réparation de l'ADN. Ainsi, la mutation a pu être transmise à la descendance, à condition bien sûr qu'elle ne soit pas létale. Parmi les substitutions possibles (Figure 2.2), on distingue :

- les transitions : remplacement d'une purine par l'autre purine (G<->A) ou d'une pyrimidine par l'autre pyrimidine (C<->T).

- les transversions : remplacement d'une purine par une pyrimidine ou inversement (C<->A, C<->G, T<->A, T<->G).

Bien qu'il y ait deux fois plus de transversions possibles que de transitions, on observe que la fréquence des transitions est supérieure. Ce biais est lié au fait que les transitions n'impliquent pas le changement de noyau des bases azotées. Ainsi la taille et la structure globale du nucléotide ne sont pas modifiées et le système de réparation ne détecte pas le changement. De même, il semble que la prédominance de substitutions C<->T (G<->A sur le brin complémentaire) soit aussi liée aux réactions de déamination de méthylcytosines (Holliday & Grigg 1993).

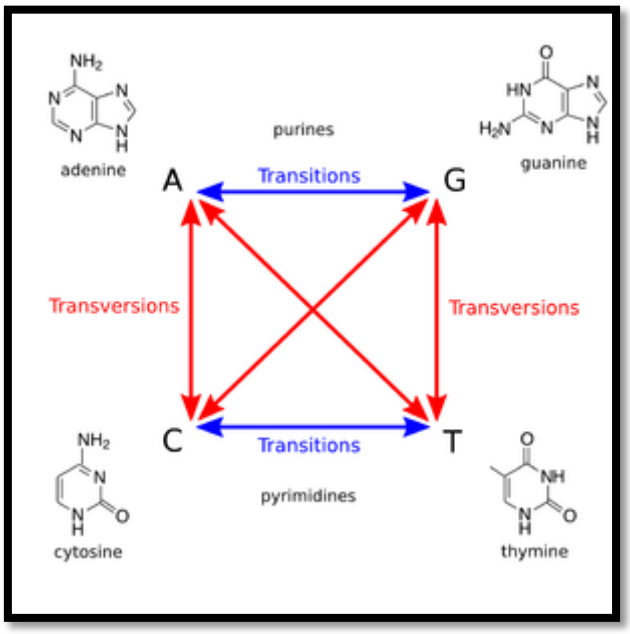


Figure 2.2 : Définition des transversions et transitions (source : http://en.wikipedia.org/).

Selon l'endroit où apparaissent des SNPS, ces polymorphismes peuvent avoir des effets différents sur le phénotype. Dans les exons, la diversité nucléotidique est très inférieure à celle observée dans les régions non-codantes du génome (Nei & Li 1979). De ce fait, les SNPs exoniques n'aboutissent à des changements non synonymes de codons que dans 50% des cas (Nickerson 1998). Les SNPs des régions non-codantes peuvent aussi affecter la régulation des gènes ou présenter un déséquilibre de liaison avec un autre SNP, ceci ayant des conséquences phénotypiques. Cependant, la grande majorité des SNPs est localisée dans les régions non codantes du génome et n'a pas d'impact détectable sur le phénotype d'un individu. Ces SNPs neutres sont très utiles en tant que marqueurs en génétique des populations (Marth *et al.* 2004; Nielsen 2004).

Toutefois comme nous l'avons déjà évoqué, les SNPs, par leurs caractère généralement biallélique, fournissent, pris isolément, une information relativement faible.

2. Le 'boom ' des SNPs

Jusqu'au début des années 2000, les marqueurs SNPs, du fait de ce caractère biallélique et de la faible quantité d'information délivré par un seul marqueur, étaient peu utilisés. Par ailleurs, la phase de découverte/d'isolement des SNPs était laborieuse en raison des faibles débits du séquençage. En effet, les SNPs étaient isolés à partir de séquençage direct ou de polymorphisme détecté par analyse de conformation simple brin (SSCP – single-strand conformational polymorphism) ou par dénaturation par chromatographie liquide de haute performance (DHPLC - denaturing high-performance liquid chromatography).

Cependant, depuis le milieu des années 2000, les techniques de séquençage, toujours en perpétuelle évolution, accroissent l'automatisation et réduisent le coût du typage des SNPs, ce qui permet l'analyse parallèle d'un grand nombre de marqueurs à un coût réduit (Davey *et al.* 2011). Les SNPs correspondant à du polymorphisme entre seulement deux variants, leurs détection, analyse et interprétation à haut débit sont relativement aisées, adaptables à toutes les plateformes existantes d'étude de l'ADN et automatisables (Gut 2001). L'analyse en très haut débit rend plus facile les études de populations nécessaires à l'estimation précise des fréquences alléliques. Leur faible taux de mutation et leur distribution homogène dans l'ensemble du génome (International Human Genome Mapping Consortium, 2001) font des SNPs des marqueurs attractifs l'étude de variabilité génétique pour les espèces modèles comme l'Homme mais aussi pour des espèces non modèles. En effet, les récentes améliorations, aussi bien dans la vitesse, le coût et la précision des NGS (next-generation sequencing), alliées aux progrès en bioinformatique, ont révolutionné les opportunités de développer des ressources génétiques comme les SNPs pour les organismes non-modèles (Ekblom & Galindo 2011). Nous aborderons plus en détail les progrès de séquençage dans la partie suivante.

3. Comparaison des SNPs versus les microsatellites

Comparé aux microsatellites, les SNPs sont intéressants pour leur reproductibilité entre laboratoire (Helyar *et al.* 2011).

Pour un locus microsatellite, la région analysée peut correspondre à plusieurs centaines de nucléotides, région dans laquelle différentes mutations ont pu se produire aussi bien dans le motif microsatellite à proprement parlé que dans les régions flanquantes (risque d'homoplasie). Par ailleurs, les microsatellites étant souvent situés dans des « points chauds mutationnels (Jarem *et al.* 2009), il y a une probabilité élevée pour que des mutations se produisent dans les régions flanquantes où sont définies les amorces, ce qui peut générer des allèles nuls et de l'homoplasie (Estoup *et al.* 2002; Krenke *et al.* 2002). Au contraire, l'analyse d'un SNP peut se limiter à l'investigation d'une région très réduite du génome (typiquement de l'ordre de 50 nucléotides, comprenant au milieu le SNP à proprement parlé entouré de 2 régions flanquantes). Cette caractéristique en fait un outil de choix pour les échantillons de mauvaise qualité comme des ADN dégradés ou des échantillons historiques. De plus, le taux de mutation de ces marqueurs est d'environ 10^{-8} (Kondrashov 2003) contre 10^{-3} pour les marqueurs microsatellites (Estoup *et al.* 2002; Schlötterer 2004). Ainsi, ces marqueurs portent l'information relative à des périodes de temps plus longues en comparaison des microsatellites puisque les allèles seront transmis sans altération sur un plus grand nombre de générations.

Pour la même raison, ces marqueurs sont des outils très intéressants en termes de résolution d'échelle spatiale car avec un nombre important de SNPs, il est possible de travailler à des échelles encore plus fines spatialement que celles qui pouvaient être appréhendée par les microsatellites (Ekblom & Galindo 2011).

Enfin, contrairement à l'analyse des microsatellites, celle des SNPs n'étant pas basée sur l'étude de la taille de fragments composés de séquences répétées, elle n'est pas confrontée au phénomène de "bégayement" de la polymérase lors de l'amplification et de ce fait ne souffre pas des difficultés d'interprétation pouvant être induites par la présence de bande échos (stutters).

Cependant, malgré tous ces avantages, comme nous l'avons évoqué précédemment, les SNPs, étant bialléliques, contiennent une information génétique très inférieure à celle accessible par les marqueurs multialléliques de type microsatellites. Selon certains auteurs, il est nécessaire d'analyser 4 à 5 marqueurs SNPs pour égaler l'information contenue dans un microsatellite (Chakraborty *et al.* 1999)) et selon d'autres, 50 marqueurs SNPs seraient équivalents à 20 marqueurs microsatellites (Smouse 2010). Des évaluations statistiques plus précises montrent qu'en excluant la possibilité d'allèles nuls, l'analyse d'un nombre relativement réduit (50) de marqueurs bialléliques serait

suffisante pour atteindre les valeurs de rapports de vraisemblance obtenues grâce à l'étude de 12 marqueurs microsatellites, si leurs fréquences alléliques sont comprises entre 0,2 et 0,8 (Smouse 2010; Santure *et al.* 2010). Dans tous les cas, cette difficulté d'obtenir un nombre important de marqueurs SNPs est maintenant largement contrecarrée par les outils de séquençage haut-débit qui permettent d'en obtenir et d'en utiliser des centaines voire des milliers (Helyar *et al.* 2011; Seeb *et al.* 2011; Davey *et al.* 2011).

D. L'avènement des nouvelles technologies de séquençage et leurs répercussions en termes d'analyse

1. Le séquençage

a) Le séquençage des origines à 2005

Le séquençage consiste à déterminer l'enchainement des nucléotides le long d'un fragment d'ADN, et de manière plus générale, sur l'ensemble d'un génome. Son invention remonte aux années 1970, où en 1977, une équipe de recherche américaine menée par Maxam et Gilbert développe une méthode basée sur le marquage radioactif de fragments et leur coupure sélective par dégradation chimique (Maxam & Gilbert 1977). La même année, de manière indépendante, une équipe anglaise menée par Sanger met au point une méthode basée sur une synthèse enzymatique des fragments d'ADN après leur amplification par clonage (Sanger *et al.* 1977). Pour ces avancées révolutionnant la biologie, Sanger et Gilbert ont reçu le prix Nobel de chimie en 1980. Depuis, la méthode de séquençage de Gilbert a été écartée du fait de la dangerosité des composants radioactifs utilisés pour la réaction de synthèse. Le séquençage Sanger a, quant à lui, été automatisé. Dans un premier temps les premières machines développées utilisaient des gels de polyacrylamide mais rapidement les séquenceurs capillaires ont remplacé ces machines, accélérant la vitesse d'exécution du séquençage.

Ces appareils, principalement représentés par le 3730 DNA Analyzer de Applied Biosystems (maintenant Life Technologies) commercialisé en 2002, sont considérés comme la première génération de séquenceurs haut-débit. Il permet de générer jusqu'à 48 séquences en parallèle (allant jusqu'à 500 / 700 nucléotides) en environ 2h. Ainsi, le génome humain a été déterminé suite à un consortium international ('International Human Genome Sequencing Consortium') qui débuta au début des années 1990 et qui avait pour objectif de séquencer les 3.5 milliards de bases composant le génome humain. J. Craig Venter, en 1998, avec sa société Celera Genomics, s'est lancé

également dans l'aventure, avec pour ambition de parvenir à séquencer le génome humain avant les laboratoires publics du consortium. En 2000, les deux acteurs ont annoncé avoir atteint leur but et ont publié simultanément leurs résultats (Lander *et al.* 2001; Venter *et al.* 2001). Les génomes complets furent présentés en 2004 par le Consortium (International Human Genome Sequencing Consortium 2004) pour un cout total de 2.7 milliards de $ et en 2007 par le J. Craig Venter Institute (Levy *et al.* 2007) pour un coût de 70 millions de $.

b) La deuxième génération de séquenceurs

Dans les années 2000, les séquenceurs de deuxième génération entrent en scène. C'est l'apparition du séquençage haut-débit, où des dizaines de milliers de séquences sont traitées ensemble, en parallèle. Le "high throughput sequencing" (HTS) ou encore le "next-generation sequencing" (NGS) sont les termes les plus utilisés pour en parler. L'avancée majeure offerte par les NGS est la capacité de produire un énorme volume de données à moindre coût, dû à la parallélisation des différents traitements effectués. Comme on peut le voir sur la figure suivante, les coûts de séquençage pour un génome a considérablement diminué à partir de 2007, date qui coïncide avec l'apparition des NGS sur le marché (Figure 2.3).

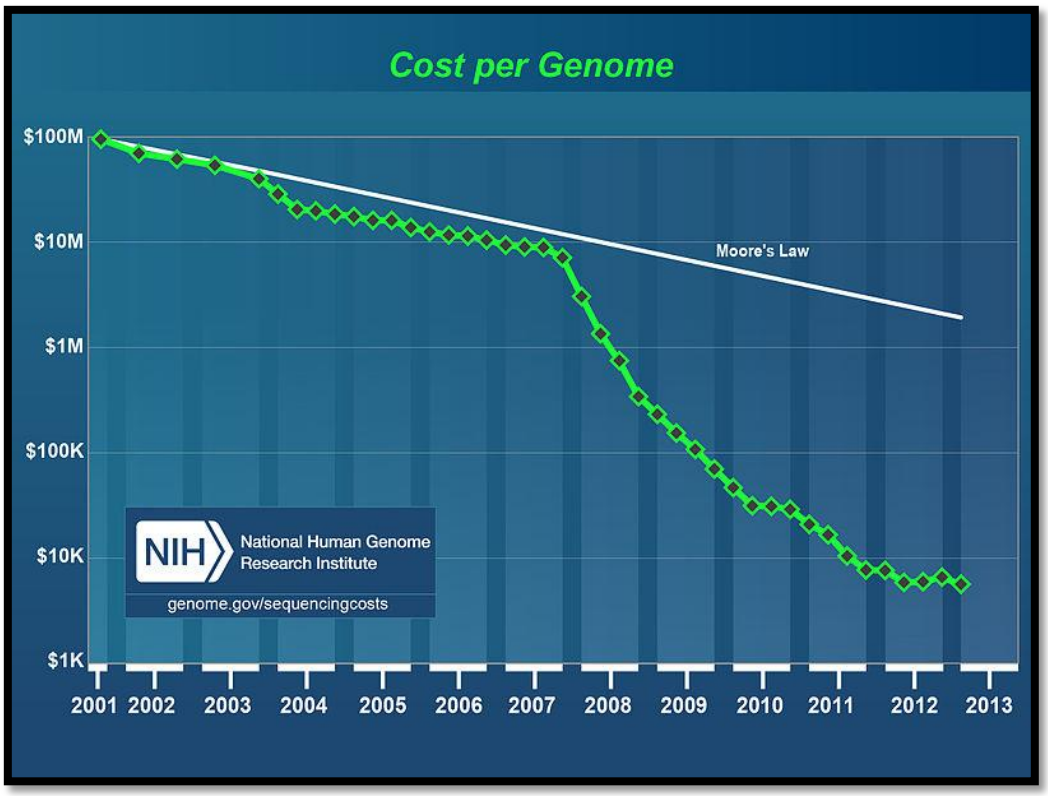


Figure 2.3 : Représentation du coût de séquençage par génome depuis 2000 (source : http://massgenomics.org/).

Actuellement, trois plateformes sont considérées comme de deuxième génération : SOLiD, Illumina et le 454. La première à voir le jour est le 454, développée par Jonathan M. Rothberg en 2005 (Margulies *et al.* 2005), commercialisée alors par Life Sciences puis rachetée par Roche. En 2006, c'est le Genome Analyzer (GA) de Solexa qui arrive sur le marché, commercialisé maintenant par Illumina. L'année suivante, en 2007 Le SOLiD (Sequencing by Oligonucleotide Ligation and Detection) fait son apparition, mis sur le marché alors par Applied Biosystems (aujourd'hui Life Technologies). Ces appareils utilisent des chimies propres à chacun d'eux (Tableau 2.2).

En 2010-2011, période où nous avons choisi la plateforme de séquençage pour notre étude, deux dominaient le marché : Illumina et Roche. Nous avons donc choisi de ne décrire que très brièvement ces deux technologies dans le corps du manuscrit. Cependant les explications des techniques de séquençage de ces deux technologies sont disponibles en annexe (Annexes 1&2).

i. Le 454

La technologie du 454 de Roche est basée sur la technique du pyroséquençage. Cette technique introduite en 1988 par Hyman (Hyman 1988) a été perfectionnée par Ronaghi *et al.* (1996), notamment avec l'intégration de la PCR (Polymerase Chain Reaction). Le principe est un séquençage par synthèse (SBS, Sequencing By Synthesis) qui se base sur l'addition d'un seul nucléotide révélé en temps réel par détection de la luminescence émise. Le pyroséquençage a ensuite été adapté au haut débit en 2005 par Margulies *et al.* (2005).

ii. L'Illumina

Cette technologie est basée sur le même principe que la méthode de séquençage 'traditionnelle' de Sanger (utilisation de nucléotides marqués qui bloquent l'élongation). Elle a la particularité d'utiliser un "terminateur" de chaine irréversible, ce qui permet une détermination plus précise des bases (Bentley *et al.* 2008). A partir de ce nouveau principe, au début des années 2000, le premier prototype haut débit, basé sur la formation de colonies d'ADN sur une puce de silice, voit le jour. En 2006, Illumina commercialise ses premiers séquenceurs de nouvelle génération basés sur cette technologie.

c) Les technologies de séquençage, une perpétuelle évolution

Comme nous l'avons évoqué, ces technologies de séquençage évoluent à une vitesse fulgurante. Ainsi nous entrons actuellement dans l'ère de la troisième génération de séquenceurs (Rusk 2011). La différence majeure qui caractérise cette évolution réside dans les capacités des appareils à séquencer directement les molécules d'ADN de manière individuelle sans aucun recours à une amplification préalable (Pareek *et al.* 2011; Venkatesan & Bashir 2011). Je ne détaillerai pas dans ce manuscrit ces différentes techniques de séquençage. Cependant je les présente dans le tableau suivant (Tableau 2.2). Nous qualifierons la plateforme Ion Torrent de Life Technologies de technologie 'intermédiaire' entre la deuxième et la troisième génération de séquençage, car elle présente un débit bien supérieur aux technologies de deuxième génération, mais comme ces dernières, elle nécessite une phase d'amplification au préalable, ce qui ne permet pas de la considérer comme une technologie de troisième génération (Rusk 2011). J'ai choisi de les présenter à titre indicatif, afin de mettre en avant l'évolution de ces technologies.

Tableau 2.2 : Description des principales plateformes de séquençage de seconde et troisième génération, selon différents critères. Chiffres d'après Davey *et al.* (2011); Glenn (2011) et http://www.biorigami.com/.

	1ère génération	2ème génération								*Intermédiaire*		3ème génération		
Plateforme	**3730**	**454**			**Illumina**			**SOLiD**		**Ion Torrent**		**HeliScope**	**PacBio**	**Starlight**
Entreprise actuelle	Life Technologies	Roche			Illumina			Life Technologies		Life Technologies		Helicos	Pacific Biosciences	Life Technologies
Entreprise d'origine	Applied Biosystems	454			Solexa			Applied Biosystems		Ion Torrent		Helicosbio	Pacific Biosciences	Life Technologies
Date de la première commercialisation	2002	2005			2006			2007		2010		2010	2011	2010
Méthode de séquençage	Sanger	Synthèse (pyroséquençage)			Synthèse			Ligation		Synthèse		Synthèse	Synthèse	Synthèse
Méthode d'amplification	PCR	PCR en émulsion			PCR en ponts			PCR en émulsion		PCR en émulsion		Aucune	Aucune	Aucune
Instrument	**3730**	**GS Junior Titanium**	**FLX Titanium**	**FLX +**	**MiSeq**	**GAIIx**	**HiSeq 2000 v3**	**SOLiD 4**	**SOLiD 5500x1**	**Ion PGM puce 314**	**Ion PGM puce 318**	**Helicos**	**PAcBio RS**	**Starlight**
Temps de run	2h	10h	10h	18-20h	26h	14jours	10jours	12jours	8jours	2h		nd	0.5-2h	nd
Millions de reads par run	0,000096	0.10	1	1	3.4	320	<3000	840	1410	0,10	4-8	12 à 20 000 000	0,01	0,01
Nb de bases par run	650	400	400	700	150 x2	150 x2	100 x2	50+35	75+35	100	>100	35	860-1100	>1000
Rendement (Mb/run)	0.06	50	500	900	1 020	96 000	<600 000	71 400	155 100	>10	>1000	>35 000 000	5 à 10 000 000	nd
Erreur principales	Substitutions	Insertions, délétions			Substitutions			Biais A-T		Insertions, délétions		Séquences répétées, bruit de fond	Délétion C-G	nd
Taux d'erreur (%)	0,1-1	1			>0,1			>0,06	>0,01	1		0,5	16	nd

d) 454 versus Illumina

Les deux technologies présentent des différences sur la quantité et la longueur des séquences produites.

La technologie Illumina est maintenant la plus utilisée des techniques de séquençage haut débit. Cela est dû à certains avantages qu'offre cette plateforme. C'est la méthode de séquençage qui coûte le moins cher et permet d'atteindre un niveau de profondeur de séquençage assez important (*i.e.* la profondeur étant le nombre de fois qu'une base est séquencée).

De façon générale, le 454 produit un plus petit nombre de séquences (de l'ordre d'1 million) mais celles-ci sont de grandes tailles (500 à 600 pb). Ces séquences plus longues constituent un avantage lorsqu'on cherche à reconstituer la disposition initiale des séquences d'un génome (assemblage). Au contraire, l'Illumina produit des séquences de tailles plus réduites, mais en plus grand nombre. Avec des longueurs de reads de 100 à 150 pb au maximum selon la machine, cette technologie connaît sa première limite.

Le nombre d'erreurs de séquençage tend à diminuer au fil des améliorations des protocoles. Le 454 fait assez peu d'erreur de substitution (1/1000b) mais de nombreuses erreurs de type indels (90%), en particulier générées par le problème des homopolymères où plusieurs bases peuvent être incorporées en même temps lorsqu'elles sont du même type. Ceci conduit à des erreurs de quantification du signal et par conséquent, de séquençage (taux d'erreur ~ 1/250b). Malgré une correction des erreurs de séquençage dues aux homopolymères, cette technique connaît aussi d'autres types de biais. En effet, la détection parallèle peut conduire à des erreurs de substitution (95% ; 1/300b ; Harismendy *et al.* 2009), surtout à la fin des reads (Brockman *et al.* 2008).

e) Estimation de la qualité de séquençage

Comme nous venons de l'aborder, l'un des problèmes avec le séquençage automatique est de déterminer la fiabilité des résultats acquis, notamment en évaluant le risque d'erreur de séquençage. Avec les NGS, diverses erreurs de séquençage sont connues, qui sont inhérentes à chaque système.

Il a été développé des modèles probabilistes afin de modéliser et donc déterminer la probabilité qu'une base détectée ne soit pas la base exacte de la séquence, comme le programme Phred décrit par Ewing *et al* (Ewing *et al.* 1998). Phred calcule différents paramètres relatifs à la forme et la résolution (intensité, hauteur,…) de chaque lecture. Par la suite, le programme utilise ces paramètres pour rechercher un score de qualité dans des tables de correspondance établies.

Le score de qualité est calculée suivant les formules (1) et (2) pour les plateformes 454 et Illumina (Cock *et al.* 2010) respectivement :

(1) Qphred = $10 \times \log_{10}{}^{P}$

(2) Qsolexa = $10 \times \log_{10}{}^{(P/1-P)}$

P étant probabilité que la base lue soit fausse.

Par exemple, si Phred assigne un score de qualité de 30 à une base, la probabilité que cette base ait été identifiée incorrectement est de 1 pour 1000, soit 99.9% de précision. De manière générale, un seuil consensus a été établi à un score de 20 (soit 99% de précision) pour considérer une lecture comme bonne.

D'une autre manière, la profondeur permet de palier à ce taux d'erreurs ainsi que le score de qualité de séquençage. Harismendy *et al.* (2009) estiment que pour un fragment d'ADN séquencé dix fois (couverture 10X) avec un score de qualité de séquençage supérieur à 20 (Q20), la probabilité d'erreur est proche de zéro. Cependant ces taux d'erreurs changent selon la complexité de l'ADN séquencé et des différentes versions de protocoles utilisées.

2. Analyse des données NGS, l'apport de la bioinformatique

a) Les défis bioinformatiques posés par les NGS

Les nouvelles méthodes de séquençage s'accompagnent d'un « déluge »* (*expression issue du titre : « The data deluge » en couverture du numéro de « *The economist* » du 25 février 2010) de données générées en une seule expérimentation. De ce fait il parait inimaginable de pouvoir et devoir analyser manuellement les jeux de données générés pouvant aller jusqu'à des Téra bases. Ainsi l'analyse bioinformatique est indispensable pour permettre le traitement de millions de séquences. Les NGS ayant mis quelques temps à s'implanter, jusqu'en 2009, très peu de logiciels permettaient d'analyser des données générées par ces technologies. Actuellement, avec l'évolution et le succès des NGS, le nombre de logiciels d'analyse de telles données a augmenté de manière exponentielle. De ce fait devant la multitude de logiciels existant, il est parfois difficile de trouver l'outil approprié pour répondre à des questions spécifiques (en fonction de la question de recherche et de l'organisme d'étude -organisme modèle ou non par exemple-). Certains sites internet comme Seqanswers (http://seqanswers.com/wiki/Software/list) ou encore Array Directory (http://arraydirectory.com/) proposent une liste à jour des logiciels développés. De plus chaque logiciel a été développé dans le but d'accomplir une tâche et il n'existe pas vraiment de suite logicielle où chaque opération peut se faire en un seul tenant avec le même outil. Une des grosses

difficultés pour les biologistes réside également dans le fait que ces logiciels ont été, pour leur très grande majorité, développés en open-source et nécessite l'utilisation de Linux/Unix et l'acquisition de certaines bases en langages de programmation (généralement C, C++, Perl, Python). Dans cette jungle de logiciels, les informaticiens commencent à mettre en place des solutions et outils utilisables par des biologistes. Ainsi certaines applications web permettent de simplifier le traitement des données grâce au développement d'interface graphique, de pipeline préétabli en fonction des questions de recherche, ce qui est beaucoup plus facile d'accès, évite les phases de paramétrages et ne requiert pas de compétences spécifiques pour utiliser des lignes de commande et donc la maitrise d'un langage spécifique. Ainsi The Center for Comparative Genomics and Bioinformatics a développé Galaxy (http://galaxyproject.org/) qui est une plateforme qui propose une "constellation" d'outils pour analyser, manipuler et visualiser des données génomiques, sans avoir besoin de connaissances en programmation. Galaxy est une solution pratique pour l'analyse d'un jeu de données classiques (comme des analyses génomique d'organismes modèles). Cependant cet outil présente de nombreuses limites lorsque le jeu de données est plus complexe.

A cette difficulté d'analyse vient s'ajouter la gestion de la masse importante de données générées et qui ne cesse de s'accroitre. Ces ressources ont donc besoin d'être stockées et réclament des performances de calcul importantes pour leur traitement.

En effet, les données NGS nécessitent un temps de calcul et de l'espace mémoire importants par rapport à des données classiques. Néanmoins, au vu des besoins suscités par le traitement de ces séquences, des efforts sont faits pour développer des clusters de calculs ayant la capacité de répondre correctement aux demandes d'espace mémoire et d'accélération des calculs. De même, des formats de compression comme le SRA (Sequencing Read Archive) de la NCBI (National Center for Biotechnology Information) ont vu le jour pour optimiser le stockage et les temps de transfert. Par ailleurs, des algorithmiciens se penchent également sur l'optimisation des calculs par l'utilisation du parallélisme à travers les GPU (Graphics Processing Unit) par exemple mais également sur l'évolution des architectures distribuées comme dans « eoulsan » (mode cloud) (Jourdren *et al.* 2012).

b) La recherche de SNPs (« SNP calling »)

Dans le cas le plus simple, l'identification de SNPs se base sur un alignement sur une référence (par exemple un génome connu) des reads issus du jeu de données généré. Les positions nucléotidiques où les reads divergent sont repérées (mismatch) et les SNPs peuvent ainsi être identifiés.

Cependant, les génomes de références (ou justes des séquences) ne sont pas toujours disponibles, comme c'est le cas pour *Ixodes ricinus*, un organisme non-modèle. Dans ce cas, une séquence référence d'un organisme phylogénétiquement proche peut être utilisé. Si aucune séquence n'est disponible, la solution pour contourner cette difficulté consiste à réaliser un assemblage dit « *de novo* » des reads issus du jeu de données afin de déterminer une séquence consensus représentative du jeu de données qui sera de ce fait considérée comme 'séquence de référence' (Miller *et al.* 2010).

Plusieurs logiciels permettent de réaliser un assemblage *de novo* , comme Newbler (associé à la suite de logiciel proposé par Roche pour les analyses de données 454) (Margulies *et al.* 2005), MIRA3 (Chevreux *et al.* 1999) ou encore Velvet (Zerbino & Birney 2008). Chacun de ces logiciels est basé sur un algorithme d'assemblage qui recherche des chevauchements entre séquences. Il en existe deux types principaux : l'algorithme basé sur l'analyse du graphe de Bruijn et celui sur l'analyse de l' 'Overlaps Layout Consensus' également appelé 'string graph' (Miller *et al.* 2010).

Pour l'assemblage par le graphe de Bruijn, les séquences générées lors du séquençage sont fragmentées en séquences plus petite de longueur « k » se chevauchant sur une longueur de k-1. Chaque fragment est appelé « k-mer ».

Dans le graphe (Figure 2.4), chaque k-mer constitue un nœud et est lié à d'autres nœuds par des flèches. Ainsi tous les k-mers chevauchant sur une taille de k-1 sont recherchés et assemblés afin de constituer un contig le plus long possible par enchainements de k-mers chevauchants provenant de reads distincts.

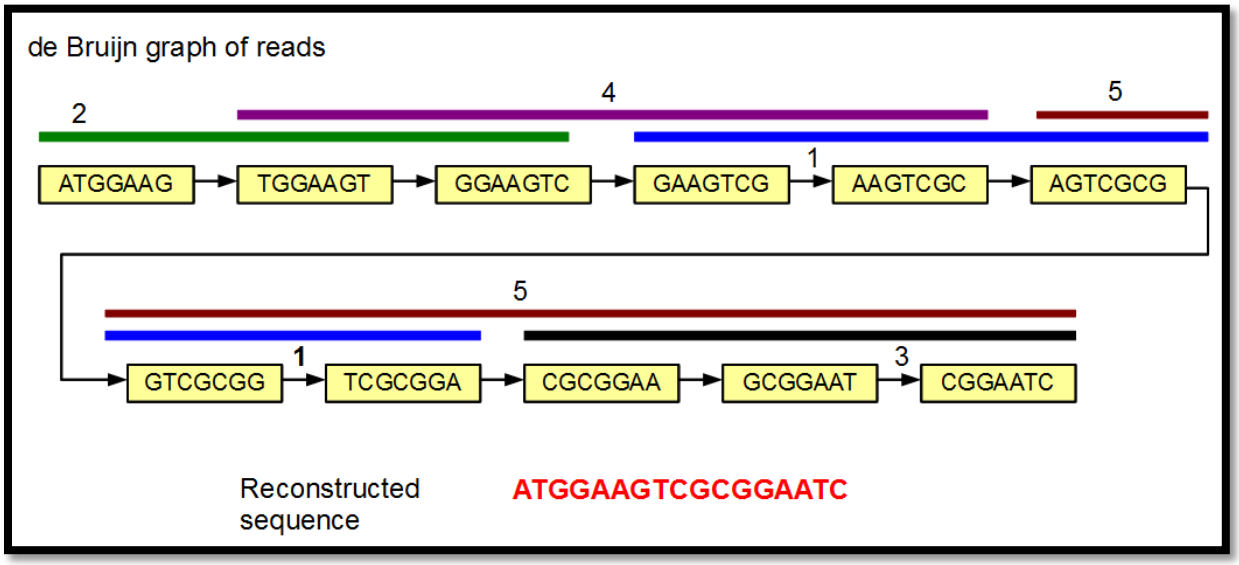


Figure 2.4 : Exemple de l'utilisation du graphe de Bruijn avec 5 reads (représentés par les traits de couleurs). Chaque reads a été découpé en k-mers de taille k=7. Par chevauchement des k-mers sur une distance de k-1, une séquence de 17 bases a pu être assemblée à partir des 5 reads (source : http://www.homolog.us/blogs/).

Pour l'assemblage par le String graph, les séquences ne sont pas redécoupées mais les chevauchements sont recherchés directement entre les reads issus du séquençage (Figure 2.5)

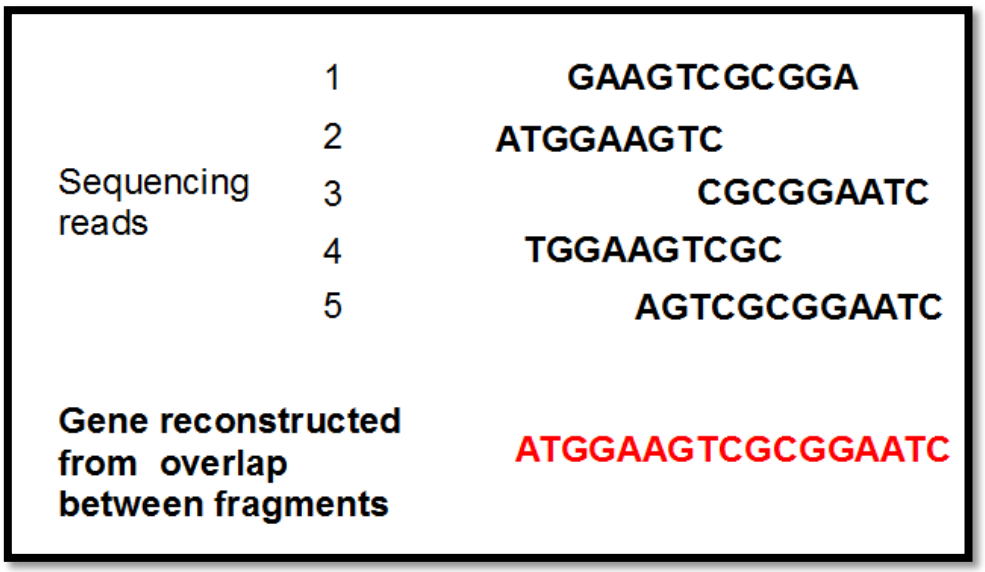


Figure 2.5 : Exemple de l'utilisation du String graph avec 5 reads (les mêmes que ceux utilisé avec le graphe de Bruijn figure 2.4) Par chevauchement des reads la séquence de 17 bases a pu être assemblée (source : http://www.homolog.us/blogs/).

L'objectif de ces outils d'assemblage est d'obtenir une séquence consensus la plus longue possible et la plus 'vraie' possible' afin d'être considérée comme 'référente'. La recherche de SNPs se fait *a posteriori*, en alignant l'ensemble des reads issus du séquençage sur cette séquence 'référente' (étape de 'Mapping').

Néanmoins, dans certains cas (par exemple une couverture trop faible pour obtenir un alignement satisfaisant), la tâche de reconstruction d'une séquence de référence en l'absence de ressources génomiques disponibles suffisantes, s'avère extrêmement complexe. Dans ce sens, des outils ont été développé comme KisSnp (Peterlongo *et al.* 2010), Cortex (Iqbal *et al.* 2012) ou encore Bubbleparse (Leggett *et al.* 2013) qui permettent d'identifier des SNPs directement dans un jeu de données sans passer par les étapes d'assemblage ou/et de mapping, ce qui est extrêmement intéressant pour les organismes non-modèles sans génome de référence.

II. Développement de SNPs dans le génome d'Ixodes ricinus par l'utilisation de technologies à haut-débit

Comme nous l'avons vu en introduction de ce chapitre, les marqueurs utilisés actuellement en génétique des populations chez *Ixodes ricinus* présentent des biais et des difficultés d'interprétation. De ce fait, il apparait nécessaire de développer de nouveaux marqueurs afin d'investiguer au mieux la biologie d'*I. ricinus.*

Nous avons choisi de développer des marqueurs SNPs, jusqu'alors inexistant dans le génome d'*I. ricinus*. Comme nous l'avons présenté, grâce aux développements récents de nouvelles technologies de séquençage, mais également de génotypage, l'accession à ces marqueurs en grand nombre est devenue possible.

Selon les différents formats de génotypage proposés par les plateformes de génotypage (en multiplexe), l'objectif a été de définir un set de 384 marqueurs. Ces marqueurs serviront, dans le troisième chapitre de ce manuscrit, à décrire le fonctionnement et la structure génétique des populations d'*I. ricinus* à l'échelle du paysage.

Le développement des SNPs a impliqué différentes étapes qui sont :

- la création d'une banque d'ADN
- le séquençage de la banque d'ADN par l'utilisation des NGS
- la recherche des SNPs par des outils bioinformatiques
- le design d'amorces pour 384 SNPs
- le génotypage à haut-débit des 384 SNPs d'individus d'*I. ricinus* (qui serviront aux analyses de la partie suivante)
- la sélection et la validation des SNPs développés.

Dans cette partie, je développerai les différentes étapes et les outils utilisés qui nous ont conduits à développer un set de marqueurs SNPs chez *I. ricinus.*

A. Séquençage

Le choix de la technologie de séquençage et de la méthodologie à employer – parmi les nombreuses disponibles - se fait en fonction de la question de recherche mais également des contraintes biologiques de l'organisme d'étude. *Ixodes ricinus* est un organisme de petite taille et de ce fait, la quantité d'ADN disponible pour un individu varie entre quelques nanogrammes à 2-3µg d'ADN pour les femelles, les plus riches en ADN. *I. ricinus* possède un génome de très grande taille – estimé à 2,1 Gb - qui présente de plus un grand nombre de séquences répétées. Par ailleurs, nous ne disposons pas de génome de référence et le nombre de ressources génomiques disponibles est très limité (1969 EST soit environ 1Mb disponible sur Genbank en 2011 - 2012).

Le séquençage de longs fragments est nécessaire pour résoudre les problèmes de séquences répétitives des génomes de grande taille et complexes et, donc permettre un assemblage de haute qualité. De ce fait, afin de faciliter l'étape critique d'analyse des données en l'absence de génome de référence, le 454, générant les reads les plus longs, nous est apparu comme la meilleure stratégie de séquençage à adopter.

Bien qu'un run de 454 génère beaucoup moins de données qu'un run d'illumina, quasiment 200 fois moins, les 500 Mb générées en moyenne par le 454 sont amplement suffisantes, afin d'identifier les 384 SNPs nécessaires à notre travail. De nombreuses études peuvent en témoigner (Bundock *et al.* 2009; Hyten, Song, *et al.* 2010; Hyten, Cannon, *et al.* 2010; Seeb *et al.* 2011; Fu & Peterson 2012).

La méthodologie à employer est également importante lorsque l'on souhaite identifier des SNPs. Dans le but de réaliser une étude de génétique des populations, nous cherchons à isoler des marqueurs monocopies. Il est donc nécessaire de pouvoir discriminer un vrai polymorphisme issu de séquences uniques dans le génome d'erreurs de séquençage, de polymorphisme dû à des séquences répétées ou de polymorphisme artefactuel lié à un mauvais alignement. Pour ceci, une profondeur de séquençage supérieure à 2 ou 3X est nécessaire (Voir Encadré 2.1).

Encadré 2.1 : Couverture *versus* profondeur de séquençage

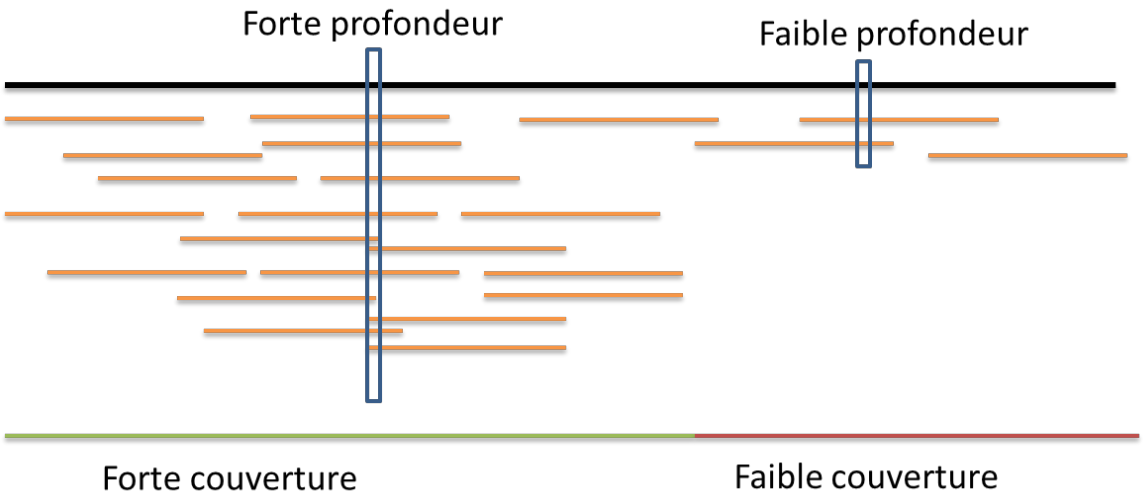
La profondeur de séquençage est le nombre de fois qu'une base est lue au cours du séquençage.

La couverture de séquençage correspond à la moyenne du nombre de reads représentant un nucléotide donné de la séquence reconstruite.

La couverture de séquençage dans le cas d'un génome peut-être calculée par la formule suivante :

$$N \times L / G$$

N=Nombre de reads
L= longueur moyenne des reads
G= taille du génome



Dans notre cas, étant donné que la taille du génome de *I. ricinus* est estimée à 2,1 Gb, et qu'un run de 454 produisant 500 Mb avec une longueur moyenne des fragments de 450 pb génèrerait environ 1 100 000 reads, nous obtiendrions une couverture de X0,24 du génome. De ce fait, il est nécessaire de réduire la taille du génome à séquencer afin d'augmenter la couverture et la profondeur de séquençage pour identifier du polymorphisme de façon fiable.

Altshuler a développé une technique, nommée *Reduced Representation Libraries* (RRL) en 2000 (Altshuler *et al.* 2000). Cette technique a ensuite été adaptée par Van Tassel en 2008 afin de pouvoir être compatible et utilisable avec les technologies de séquençage haut débit (Van Tassell *et al.* 2008). Cette technique, comme de nombreuses autres méthodes de réduction génomique, est basée sur l'utilisation d'enzyme de restriction et a été utilisée avec succès dans différentes études de développement de SNPs chez d'autres espèces (Wiedmann *et al.* 2008; Hyten *et al.* 2010a; Fu &

Peterson 2012). La RRL permet de digérer l'ADN de différents individus par une enzyme de restriction choisie selon l'objectif de l'étude. Les fragments résultant de la digestion sont ensuite sélectionnés selon leur taille et séquencés. Les sites de restriction étant des sites supposés conservés dans le génome entre différents individus, leur utilisation permet une reproductibilité sans faille da la portion réduite.

Ce protocole permet, d'une part de produire une partie représentative de l'ensemble du génome car les fragments issus de la digestion sont répartis de manière aléatoire dans le génome, et d'autre part d'avoir une reproductibilité dans le choix des séquences entre les différents individus séquencés et donc augmenter la couverture des fragments séquencés.

La principale contrainte de cette technique dans notre cas, est la quantité d'ADN par individu qui est très limitée. En ne prenant qu'une fraction des fragments digérés, on diminue encore la quantité d'ADN récupérée par individu alors que le 454 nécessite une quantité importante d'ADN (500ng) pour construire une librairie.

1. Construction d'une librairie réduite représentative du génome d'*Ixodes ricinus*

Afin de construire une banque d'ADN dans le but d'identifier des SNPs, nous avons opté pour une stratégie de réduction génomique basée sur une digestion enzymatique. Cette stratégie nécessite une bonne qualité de l'ADN, une reproductibilité lors de l'expérimentation entre les différents échantillons et le choix crucial de l'enzyme de restriction.

a) L'ADN, un facteur limitant

Le principal prérequis afin de construire une banque d'ADN basée sur la digestion enzymatique est la qualité de l'ADN et la conservation de son intégrité pour la digestion. La méthode utilisée pour réaliser le broyage mécanique des tissus de tiques influence fortement l'intégrité de l'ADN extrait. Les tiques possédant une cuticule en chitine, le broyage se fait habituellement de manière mécanique à l'aide du Tissue Lyser de Qiagen, dans des tubes contenant les tiques à broyer auxquelles ont été ajoutées des billes en silice réduisant ainsi une tique entière en une poudre fine.

Cependant cette technique n'est pas sans conséquences pour l'intégrité de la molécule d'ADN. En effet l'ADN issu d'un tel broyage est dégradé et fragmenté de manière totalement aléatoire, générant un smear avant même la digestion enzymatique (Figure 2.6a). De ce fait, l'utilisation d'enzyme de restriction *a posteriori* ne permet pas d'obtenir une bonne reproductibilité expérimentale entre différents échantillons. Nous avons opté pour le broyage au pilon sur des tiques congelées dans de l'azote liquide dans des cryotubes individuels, permettant de conserver l'intégrité de l'ADN extrait (Figure 2.6b).

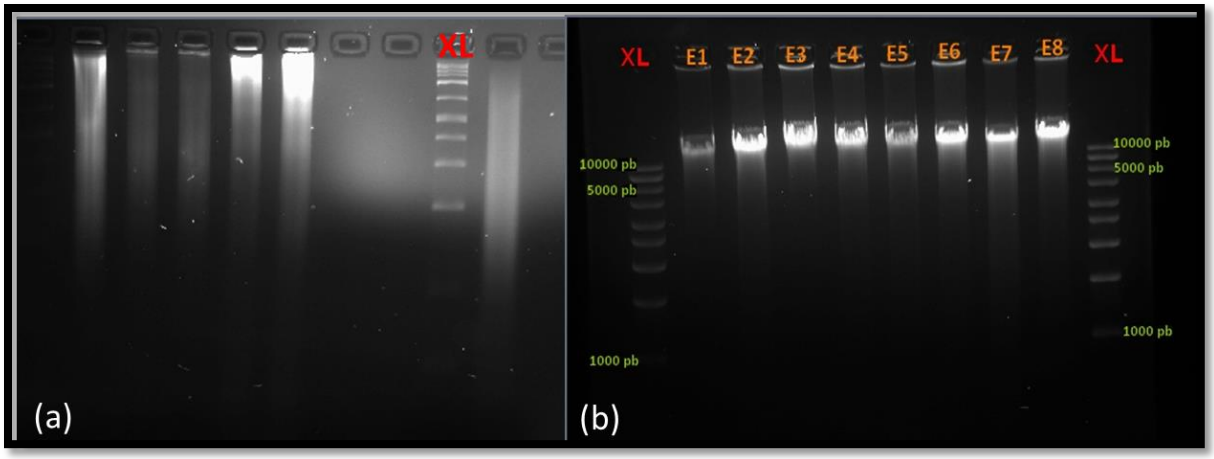


Figure 2.6 : Electrophorèse en gel d'agarose 1% d'ADN extrait à l'aide du kit NucleoSpin de Macherey-Nagel suite à un broyage de tique individuelle (a) au Tissue Lyser (b) à l'azote liquide.

Afin de récupérer le maximum d'ADN, nous avons choisi de ne travailler qu'à partir de femelles qui constituent le matériel biologique le plus riche en ADN (comparativement aux autres stades/sexe : mâles, nymphe ou larve). L'ADN a été extrait à l'aide du kit NucleoSpin tissue XS (Macherey-Nagel). Ce kit utilise des colonnes qui fixent l'ADN sur une membrane. Le kit XS, grâce à une membrane très fine, permet de concentrer l'ADN au maximum lors de la phase d'élution ce qui s'avère utile pour les échantillons de faible quantité d'ADN comme nos tiques.

b) Sélection de l'enzyme de restriction

Afin de maximiser le rendement du séquençage, qui permettait alors de séquencer des fragments d'une taille moyenne de 450 pb, nous souhaitions utiliser une enzyme qui maximisait le nombre de fragments autour de cette longueur optimale. Nous souhaitions également utiliser une enzyme qui ne montre pas de bandes répétées dans le smear, du moins au niveau de la zone ciblée pour être séquencée. En effet, une bande discrète au sein du smear suggère la présence d'éléments répétés dont le polymorphisme peut difficilement être utilisable pour des études de génétique des populations (locus présent en plus de deux copies dans un génome diploïde).

De ce fait nous avons testé différentes enzymes de restriction, disponibles au sein de notre laboratoire ou ayant déjà été utilisées lors d'études précédentes utilisant des techniques similaires aux nôtres (Wiedmann *et al.* 2008; Ramos *et al.* 2009; Sanchez *et al.* 2009; Hyten *et al.* 2010a).

Le choix final s'est arrêté sur l'utilisation d'une enzyme unique, *MseI* (5'...T↓T A A...3' 3'...A A T↑T...5') qui répondait à nos attentes en termes d'intensité du smear dans la région ciblée (500-600 pb) et d'absence de bandes répétées dans cette région (Figure 2.7a).

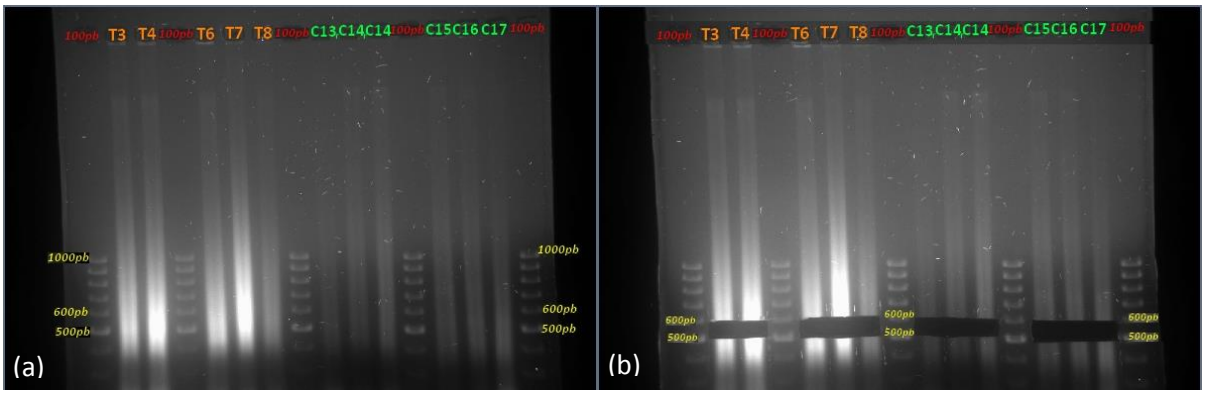


Figure2.7 : Electrophorèse en gel d'agarose 1% d'ADN de différents individus digéré par l'enzyme *MseI* (a)avant excision ; (b) partie excisée.

c) Sélection des individus et création de banque d'ADN.

L'objectif du séquençage dans notre projet est d'une part identifier des SNPs qui serviront à génotyper des tiques issues de populations naturelles pour répondre à des questions de génétique des populations à l'échelle du paysage, et d'autre part de développer des SNPs utilisables à une plus large échelle géographique. De ce fait nous avons opté pour le séquençage de deux populations de tiques d'origine géographique différente : Malville (proche de Nantes) et Gardouch (proche de Toulouse)

Les tiques de Gardouch (que l'on appellera Population T, -pour Toulouse-), proviennent du sud-ouest de la France ('Gardouch' – Haute-Garonne - 43° 23' 27.88"N, 1° 41' 1.67"E). Les tiques ont été récoltées semi-gorgées (1/4 de gorgement maximum) sur chevreuil durant l'hiver 2010 puis ont été conservées à -80°C, jusqu'au moment de l'extraction durant l'hiver 2012.
Les tiques de Malville (que l'on appellera Population M), proviennent quant à elle du nord-ouest de la France (Malville– Loire-Atlantique -47° 21' 30.10" N, 1° 51' 41.59"W). Les tiques ont été récoltées par la technique du drapeau sur la végétation, non gorgées, au printemps 2012 et ont été ensuite conservées vivantes à 4°C jusqu'à l'extraction.

L'ADN de chaque individu a été extrait de manière individuelle à l'aide du kit d'extraction NuleoSpin XS (Macherey –Nagel), afin de conserver l'information individuelle jusqu'au séquençage.
Les échantillons ont ensuite été digérés individuellement pendant 8 heures (2.5U/µg d'ADN) suivant les instructions du fabricant. Les échantillons digérés ont ensuite été séparés sur un gel d'agarose à 1% pendant 4h à 80V (Figure 2.7a). Un fragment de gel de chaque échantillon, contenant les fragments d'ADN entre 500 et 600 pb sur la base de l'échelle du marqueur de taille 100 pb DNA ladder (Eurobio) a été excisé sous lampe UV (Figure 2.7b).

2. Réalisation du pyroséquençage 454, constitution des pools d'individus

Après avoir obtenu la fraction d'ADN cible pour les tiques de notre étude, les échantillons ont été envoyés et traités individuellement à la plateforme BioGenouest de Rennes.

La quantité d'ADN a été dosée grâce au picogreen (Quant-iT™ PicoGreen® dsDNA) et la qualité de l'ADN extrait a été vérifiée pour chaque individu à l'aide du Bioanalyzer 2100 d'Agilent Technologies, basé sur une plaque micro-fluidique, qui permet l'analyse qualitative et quantitative de l'ADN. Ces puces permettent notamment de visualiser la taille des fragments d'ADN des échantillons. (Figure 2.8).

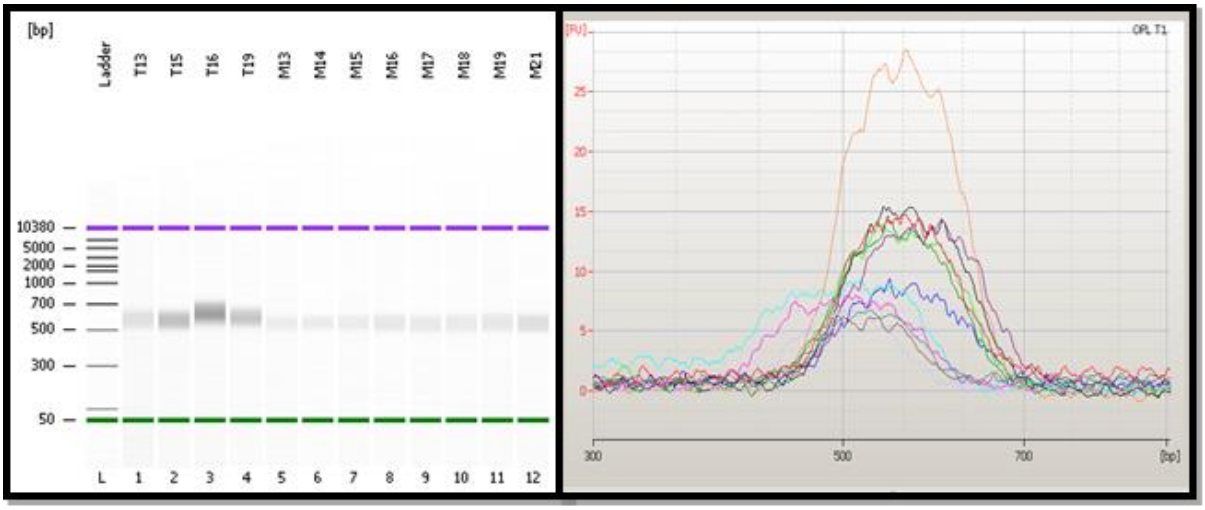


Figure 2.8 : Exemple de résultats obtenus grâce au BioAnalyser 2100 testant la qualité de l'ADN extrait ; à gauche, un électrophoregramme permettant d'estimer la taille globale de l'échantillon et à gauche la quantification de chaque échantillon (de couleur différente) selon la taille des fragments et leur quantification relative.

Les échantillons ont été sélectionnés sur la base de leur quantité d'ADN mais également selon la taille et le profil des fragments obtenus suite à l'excision. Les échantillons ayant des profils similaires ont été favorisés. Par exemple, sur la figure 2.8, les échantillons rose et bleu turquoise n'ont pas été conservés pour le séquençage, la taille des fragments issus du smear n'étant pas la même taille que pour les autres échantillons. Ceci peut être dû aux difficultés du découpage des bandes sur gel d'agarose qui peuvent ne pas être réalisés exactement au même endroit pour tous les individus.

Les individus présentant une quantité et une qualité d'ADN suffisantes (ADN non dégradé et taille de fragments de 500 pb environ) ont été regroupés de manière la plus équimolaire possible afin d'éviter une surreprésentation de certains individus au cours du séquençage.

Les individus provenant de la population T, étant des tiques semi-gorgées, étaient riches en ADN (en moyenne 79,8 ng d'ADN par tique avec des valeurs allant de 43,8 à 161,1ng). De ce fait nous avons sélectionné l'ADN de 10 individus de la population T, que nous avons poolé de manière équimolaire en harmonisant les échantillons à une concentration de 2,6 ng/µl dans 30µl, et ainsi obtenu une librairie de 780 ng d'ADN.

Pour les tiques de la population M, on a obtenu une moyenne de 28,8 ng d'ADN avec des valeurs allant de 53,4 à 16,47 ng. Les quantités d'ADN étant plus faibles pour cette population, nous avons sélectionné 20 individus à pooler. Du fait des quantités faibles et hétérogènes d'ADN, nous avons réalisé trois ajustements de dilution à 0,6 ; 0,8 et 1 ng/µl pour être le plus équimolaire possible. Ainsi, nous avons obtenu une librairie d'ADN de 515ng.

Comme la digestion avec l'enzyme *MseI* génère des bouts cohésifs TAA lié au site de restriction, un étiquetage 'MID' (Multiplex IDentifier adaptator) a été ajouté à l'extrémité de chaque brin d'ADN. En effet, des erreurs d'assignation peuvent être observées lors du séquençage avec l'appareil 454 avec les fragments d'ADN démarrant par un T.

3. Résultats du pyroséquençage

Le séquençage a été réalisé sur la plateforme génomique Biogenouest à Rennes par le Genome Sequencer FLX version Titanium. Les deux pools ont été déposés séparément dans un secteur différent de la puce 454, générant ainsi deux jeux de données différents, soit un premier pour la population M et un second pour la population T.

Pour l'ensemble des 2 jeux de données issus du pyroséquençage, nous avons obtenu un total de 1 389 201 reads (=séquence courte résultant du séquençage), 658 719 reads pour la population T et 730 482 reads pour la région M. Sur un total de 528 Mbp générées, les séquences obtenues avaient une taille moyenne de 425 pb (Figure 2.9), et une qualité de Q33 avec 73,7% de bases ayant un score de qualité supérieur ou égal à 20.

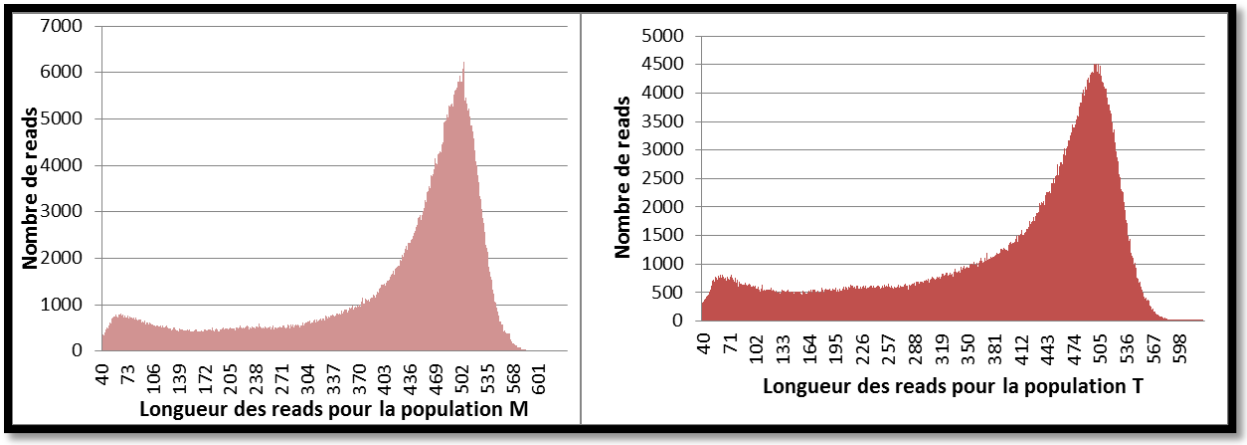


Figure 2.9 : Distribution de la longueur des reads obtenue suite au séquençage pour la population M (à gauche) et pour la population T (à droite).

B. Analyse bioinformatique des données issues du pyroséquençage

1. 'Trimming' et sélection des reads pour l'identification de SNPs

Avant de réaliser l'analyse du jeu de données, un nettoyage ('trimming') des séquences est nécessaire afin de retirer les séquences de mauvaise qualité, les séquences issues d'erreurs de séquençage et les séquences des adaptateurs MIDs.

Pour ceci deux étapes de nettoyage des données ont été effectuées. Un premier filtre a été effectué directement par la plateforme de séquençage. Ce filtre, un script implémenté en langage Python, a permis d'éliminer les séquences avec des problèmes inhérents au séquençage. Ainsi les séquences (i) ne contenant pas les quatre bases (A, C, T, G), (ii) contenant plus de 7% de base indéterminée, (iii) contenant des motifs répétés, (iv) de plus de 950 bp et de taille inférieure à 150 pb ont été supprimées.

Une deuxième étape de nettoyage a été effectuée en utilisant un script implémenté en langage Perl développé par les bioinformaticiens de l'INRA de Jouy-en-Josas. Ce script a permis de retirer l'ensemble des reads ou extrémités de reads présentant une qualité moyenne inférieure à 20. Suite à cette opération, les MIDs ont été retirés et les séquences commençant par le motif 'TAA' issu de la digestion enzymatique ont été recherchées pour ne sélectionner que les séquences issues de la digestion enzymatique. Quatre-vingt-quinze pourcents des séquences obtenues correspondaient à ce critère.

Ainsi 392 693 reads ont été supprimés. La taille moyenne des reads obtenus suite à cette étape de 'trimming' nous a permis d'obtenir 996 508 reads d'une longueur moyenne de 529 pb (Tableau 2.3). Ces 996 508 reads ont été utilisés par la suite afin d'identifier des SNPs.

Tableau 2.3 : Récapitulatif des données issues du séquençage (raw) pour les deux populations (M et T) et des deux étapes de trimming ('Passed 1' et 'Passed 2).

Population	Number of Individuals	Number of reads			Length of reads (Passed 2 reads)			
		Raw	Passed 1	Passed 2	Mean	Maximum	Minimum	Total nt
M	20	730,482	638,228	536,061	528	914	167	283,554,541
T	10	658,719	563,986	460,447	530	825	30	244,272,285
Total	30	1,389,201	1,202,214	996,508	529	914	30	527,826,826

2. Recherche de SNPs : approche par assemblage *do novo*

Afin d'identifier les SNPs dans notre jeu de données, nous avons dans un premier temps opté pour la méthode la plus couramment utilisée à savoir, le mapping sur un génome de référence (même partiel) pour rechercher des polymorphismes. Pour cela, nous avons réalisé un assemblage du jeu de données dans le but d'obtenir une séquence de référence sur laquelle nous pourrions par la suite aligner l'ensemble des reads et identifier les SNPs. Nous avons utilisé deux logiciels d'assemblage (Newbler (Margulies *et al.* 2005) et MIRA3 (Chevreux *et al.* 1999)) afin de les comparer et choisir le plus adapté. Indépendamment de la recherche de SNPs, l'assemblage permettait aussi d'estimer la couverture génomique des deux banques RRL et ainsi la couverture de séquençage et la densité des SNPs dans le génome d'*I. ricinus*.

Cependant, les deux logiciels utilisés n'ont pas permis d'obtenir un assemblage des reads satisfaisant (Tableau 2.4). Le nombre de reads utilisé pour l'assemblage dans les deux cas est assez faible (moins de 50%), et la longueur de certains contigs (plus de 2000 pb) n'est pas compatible avec la stratégie adoptée de réduction génomique par enzyme de restriction. De plus la vérification des contigs générés montre l'ajout d'un grand nombre d'insertions/délétions afin d'obtenir des chevauchements entre les reads. De ce fait, nous ne pouvons qu'accorder une faible confiance à ces assemblages générés.

Tableau 2.4 : Récapitulatif des résultats d'assemblage obtenu par Newbler et MIRA3.

Logiciel	Nb de contigs	Taille des contigs	Taille moyenne	Nb de reads utilisés
Newbler	11 149	3-2629pb	579pb	111 712 (24%)
MIRA3	136 345	47-2163pb	568pb	397 058 (40%)

Bien que les deux assembleurs n'aient pas donné de résultats satisfaisants, MIRA3 s'est avéré donner de meilleurs résultats que Newbler, d'une part en assemblant un plus grand nombre de reads mais également en générant un nombre de contig beaucoup plus élevé. Cependant, la couverture moyenne obtenue de l'ensemble des contigs est de 2,68 (avec une valeur médiane de 2,3) (Figure 2.10). Nous avons pu estimer à 78,3 Mbp la taille obtenue de couverture du génome par l'ensemble des contigs générés par MIRA3, représentant 3,8% du génome complet d'*I. ricinus.*

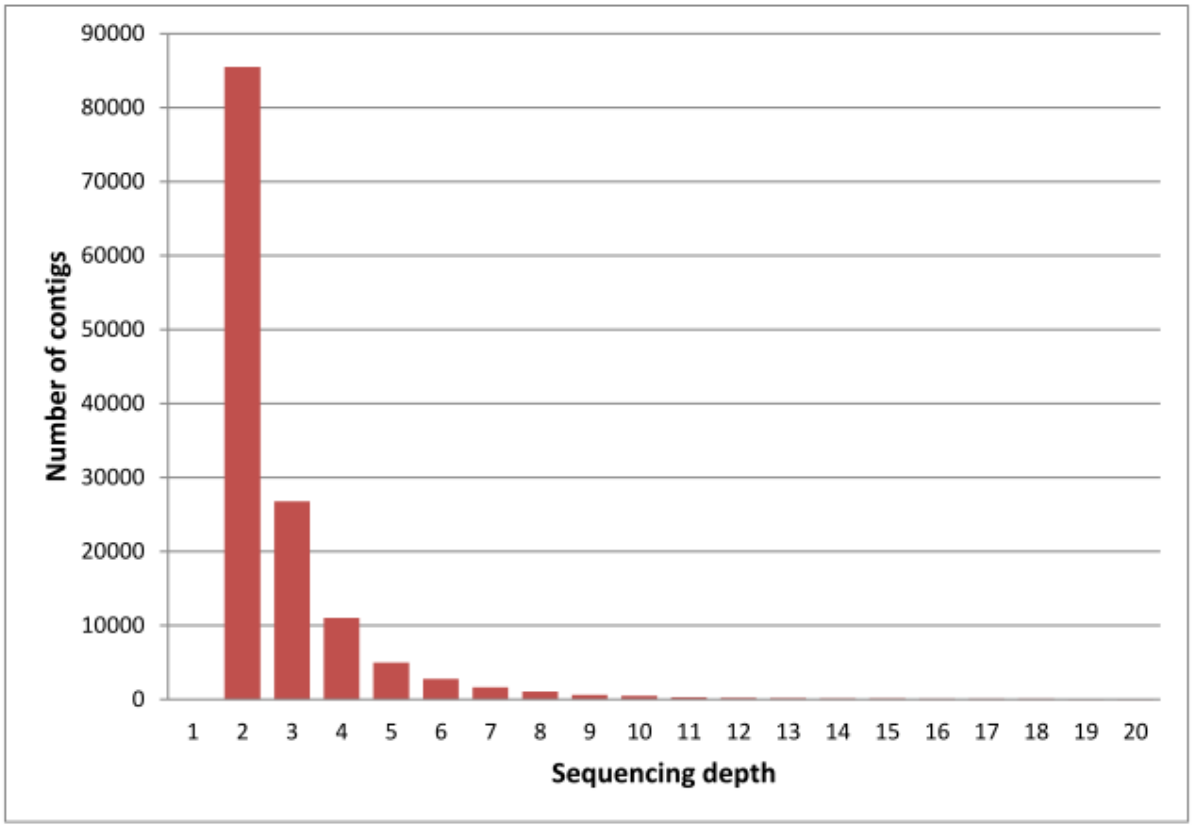


Figure 2.10 : Distribution du nombre de contigs générés par MIRA3 en fonction de la profondeur de séquençage obtenue lors de l'assemblage.

3. Recherche de SNPs : approche *de novo* avec le logiciel DiscoSnp

Étant donné les difficultés rencontrées lors de l'étape d'assemblage, nous avons opté pour l'utilisation du logiciel KisSnp, développé par Pierre Peterlongo et collaborateurs (Peterlongo *et al.* 2010) permettant de détecter les SNPs à partir des séquences brutes, sans utiliser de génome de référence. Les échanges que nous avons eus avec Pierre Peterlongo pour utiliser ce logiciel – dont notre jeu de données a constitué la première application sur données réelles – a contribué à l'évolution de KisSnp et ajuster différents paramètres qui permettaient de valider les SNPs isolés. De ce fait, nous avons collaboré avec Pierre Peterlongo et Olivier Quenez (ingénieur bioinformaticien sur la plateforme Biogenouest et à l'INRIA de Rennes), afin de d'améliorer KisSnp, ce qui a permis de développer un nouveau logiciel –DiscoSnp (encadré 2.2)- et d'identifier avec fiabilité des SNPs dans notre jeu de données. Une publication, dont je suis co-auteur, sur cet outil et son utilisation est en finalisation de rédaction (Annexe 9).

Encadré 2.2 :DiscoSnp

DiscoSnp est l'un des rares logiciels (avec Bubbleparse (Leggett *et al.* 2013), NIKS (Nordström *et al.* 2013) et Cortex (Iqbal *et al.* 2012)) à proposer la recherche de SNPs dans un jeu de données sans génome de référence. L'enjeu de ce logiciel est d'identifier des SNPs sans réaliser d'assemblage, avec un bon niveau de fiabilité et en écartant les SNPs issus d'erreurs de séquençage. DiscoSnp est un outil composé de deux modules KisSnp2 et KissReadss. KisSnp2 permet de détecter les SNPs dans un ou plusieurs jeux de données. KissReadss évalue la qualité des SNPs et leur couverture. En effet, l'isolement de SNPs doit se faire sur un nombre de séquences déterminés (au minimum 2 pour mettre en évidence du polymorphisme). Différents modules optionnels permettent de sélectionner les SNPs d'intérêt en fonction des objectifs de l'utilisateur de DiscoSnp. Ces modules complémentaires ne seront pas détaillés dans la suite de cette partie.

KisSnp2

KisSnp2 est basé sur une analyse des reads, divisés en k-mer (séquence de taille k), par le graphe de De Bruijn. Pour ceci, les reads sont découpés en k-mers chevauchants sur k-1 bases. Ensuite un algorithme parcourt le graphe de De Bruijn et détecte des régions où deux k-mers présentent une base de différence. A partir de là, une 'bouche' (d'où le nom de "Kiss") est ouverte. De part et d'autre de cette 'bouche', l'extension de proche en proche sur des k-mers successifs se chevauchant sur

k-1 bases va permettre d'obtenir les séquences flanquantes à gauche et à droite jusqu'à atteindre une longueur totale de 2k-1 et permettre à la 'bouche' de se refermer (Figure 2.11).

De ce fait chaque 'bouche' est composée de deux chemins commençant et terminant par le même nœud, et tous les deux possèdent k+2 nœuds. KisSnp2 recherche les 'bouches' ainsi formées dans un ou plusieurs jeux de données.

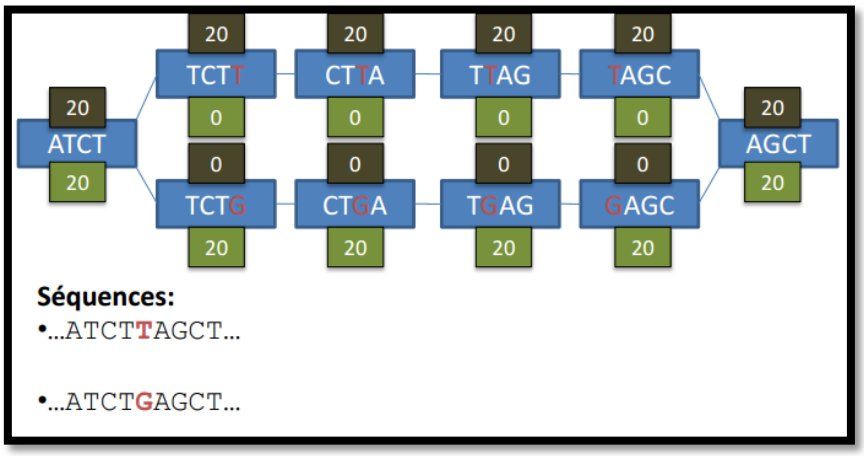


Figure 2.11: Exemple d'utilisation du graphe de De Bruijn dans DiscoSnp avec des k-mers de taille k=4. Un premier k-mer 'ATCT' correspond au premier nœud, ce k-mer est chevauchant avec deux autres qui présentent une différence sur la dernière base 'TCTT' versus 'TCTG'. De ce fait une bouche s'ouvre. De proche en proche les k-mers successifs se chevauchant sur k-1 bases sont identifiés jusqu'à obtenir un second nœud 'AGCT' qui permet de refermer la 'bouche'. Ainsi deux séquences, correspondant aux deux chemins sont reconstruites avec un SNP identifié.

KissReadss

KissReadss est un prolongement du module KisSnp2. Ce dernier met en évidence toutes les 'bouches' du jeu de données analysées et ne fait aucune vérification. La recherche de toutes les 'bouches' du jeu de données peut entraîner la comparaison de k-mers n'ayant absolument rien à voir les uns avec les autres, car les k-mers sont issus de reads fragmentés. Il est donc possible d'obtenir la formation de SNPs n'existant pas dans le jeu de données, par chevauchement de k-mers présentant la même séquence mais issus de reads originaux différents. Le rôle de KissReadss est de filtrer les SNPs qui ne sont pas cohérents avec les reads (=faux positifs). Une séquence 2k-1 identifiée par KisSnp2 est dite « read-coherent » si elle peut être retrouvée dans les reads originaux du jeu de données. Pour chaque 'bouche', KissReadss évalue la « k-read-coherency » c'est à dire la cohérence de chaque k-mer avec un read donné, pour chaque chemin de la 'bouche' et par rapport au jeu de données utilisé en entrée de KisSnp2. En plus de cette évaluation, KissReadss ajoute des informations pour chaque bulle (qui correspond à un SNP) telles que la qualité moyenne du SNP, la complexité de la séquence, etc., qui peuvent être utilisées dans les modules complémentaires.

a) paramétrage de ‘k’

Dans un premier temps, nous avons cherché à déterminer la taille de k-mer la plus favorable à l'obtention de SNPs dans notre jeu de données afin d’éviter au maximum les faux positifs. En effet, les k-mers courts générés par DiscoSnp vont être beaucoup plus facilement superposables que de longs k-mers. La conséquence des k-mers courts est d'obtenir un grand nombre de SNPs qui n'existeraient pas, dû à une superposition non spécifique. A l'inverse, avec l’utilisation de k-mers très grands, une partie des SNPs existant réellement n’est pas trouvée du fait de la difficulté à assembler ces longs k-mers. Cependant les SNPs obtenus de cette manière sont très fiables.

Dans le but de choisir le ‘k’ optimal, nous avons utilisé le premier module de DiscoSnp en faisant varier la taille des k-mer de 10 à 80 (Figure 2.12). Les fichiers obtenus ont ensuite été analysés avec la suite logicielle GenomeTools (Gremme *et al.* 2013). Le programme ‘tallymer’ permet de compter, d'indexer et de chercher les k-mers dans un fichier donné et le programme ‘Occratio’ permet d'obtenir la répartition des k-mers uniques en fonction de leur taille (Kurtz *et al.* 2008).

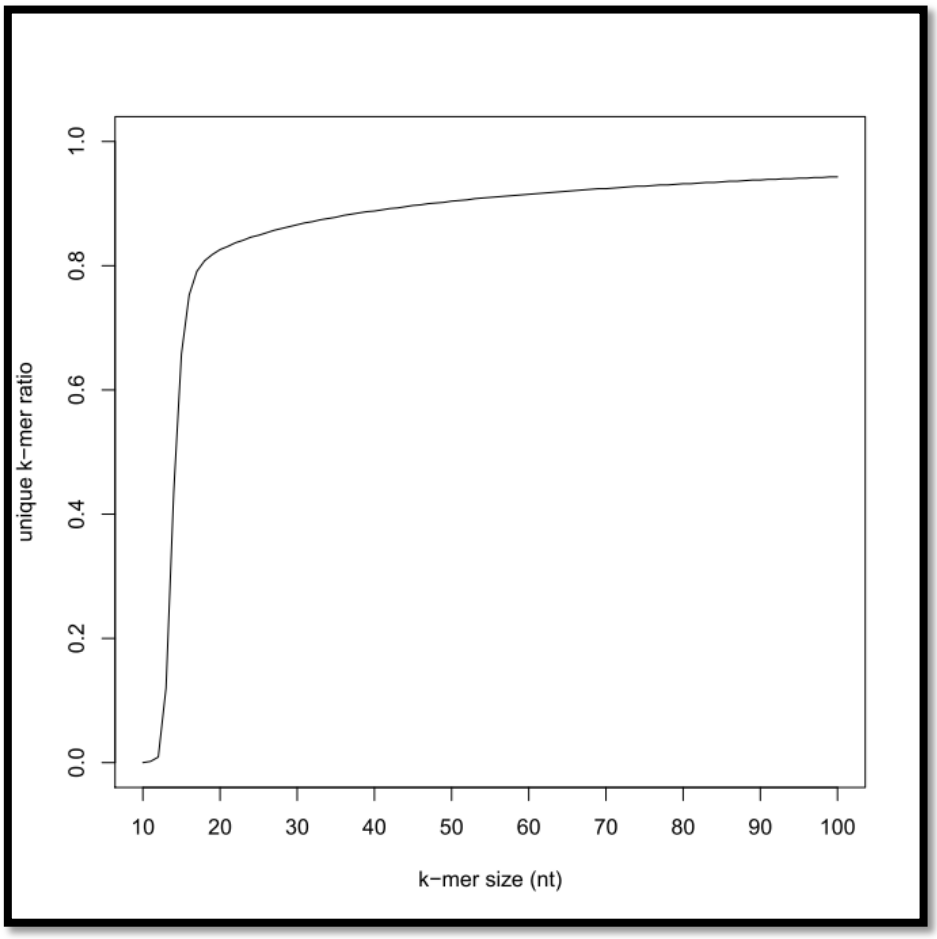


Figure 2.12 : Fréquence relative de k-mers unique dans le jeu de données en fonction de la taille des k-mers.

La proportion de k-mers uniques dans notre jeu de données augmente très rapidement de 0 à 80% entre des tailles de k-mer de 10 à 20, puis un palier est atteint au-delà de k = 40. De ce fait, nous avons choisi de travailler avec une taille de k-mer de 29, taille correspondant au début du plateau de la courbe de distribution et le logiciel ne prenant en compte que des nombres impairs.

b) Identification de SNPs par le module KisSnp2

Nous avons utilisé l'ensemble du jeu de données, soit un total de 996 508 séquences (536 061 pour la population T et 460 447 pour la population M) en sélectionnant pour taille de k-mer 30 (ce qui correspond à k=29 en considérant k-1). L'identification des SNPs se fait donc par chevauchement de 2 séquences de 29 bases. La séquence de 2k-1 bases obtenue par le logiciel est d'une longueur de 57 bases et contient le SNP en position 29. Ainsi, grâce à KisSnp2, 791 803 bulles ont été trouvées, correspondant à autant de SNPs.
Le fichier de sortie en .FASTA contient les résultats avec, pour chaque SNP, les deux séquences qui ont permis d'identifier une bouche (higher et lowerpath), ainsi que des indications sur le score de qualité (score-XX) et la complexité de la séquence (high) (Figure 2.13).

```
>SNP_higher_path_1|score_42|high|left_contig_length_6|right_contig_length_1
tactacAACCACCCGATCAAGCAGTGGATTGCGAGATACCTAGGCCACAACCCTCAGGAAAACTAt
>SNP_lower_path_1|score_42|high|left_contig_length_6|right_contig_length_1
tactacAACCACCCGATCAAGCAGTGGATTGCGAGGTACCTAGGCCACAACCCTCAGGAAAACTAt

>SNP_higher_path_2|score_66|high|left_contig_length_12|right_contig_length_12
atgtagtccatgGCGAGGGCGTGGTTGGCGGCGCTCACATCAAGGCTCGTCCTGATATCGTGCAGGG
>SNP_lower_path_2|score_66|high|left_contig_length_12|right_contig_length_12
atgtagtccatgGCGAGGGCGTGGTTGGCGGCGCTCACATCGAGGCTCGTCCTGATATCGTGCAGGG
```

Figure 2.13: Exemple de sortie obtenue à l'issue du module KisSnp2 du logiciel DiscoSnp pour deux SNPs.

c) Validation des SNPs par le module KissReads

Comme KisSnp2 sépare complètement les SNPs détectés dans des séquences reconstruites des reads d'origine, il est nécessaire de rapporter les SNPs sur les reads d'origine. Après élimination par KissReads des SNPs « faux positifs » (SNPs identifiés par KisSnp2 appartenant à 2 reads d'origine différente), 321 088 SNPs, soit 40% des SNPs identifiés par KisSnp2, ont été retenus.

4. Sélection des SNPs identifiés par DiscoSnp

Notre objectif étant d'obtenir un set de 384 SNPs pour des études de génétique des populations à l'échelle du paysage, nous avons effectué un tri des 321 088 SNPs obtenus afin de sélectionner ceux pouvant être génotypés avec la plus grande confiance, selon différents critères.

a) Critère de profondeur de séquençage

Comme nous l'avons vu précédemment, le génome d'*Ixodes ricinus* est très riche en séquences répétées. De ce fait nous avons analysé la distribution des SNPs (Figure 2.14) en fonction de leur profondeur de séquençage.

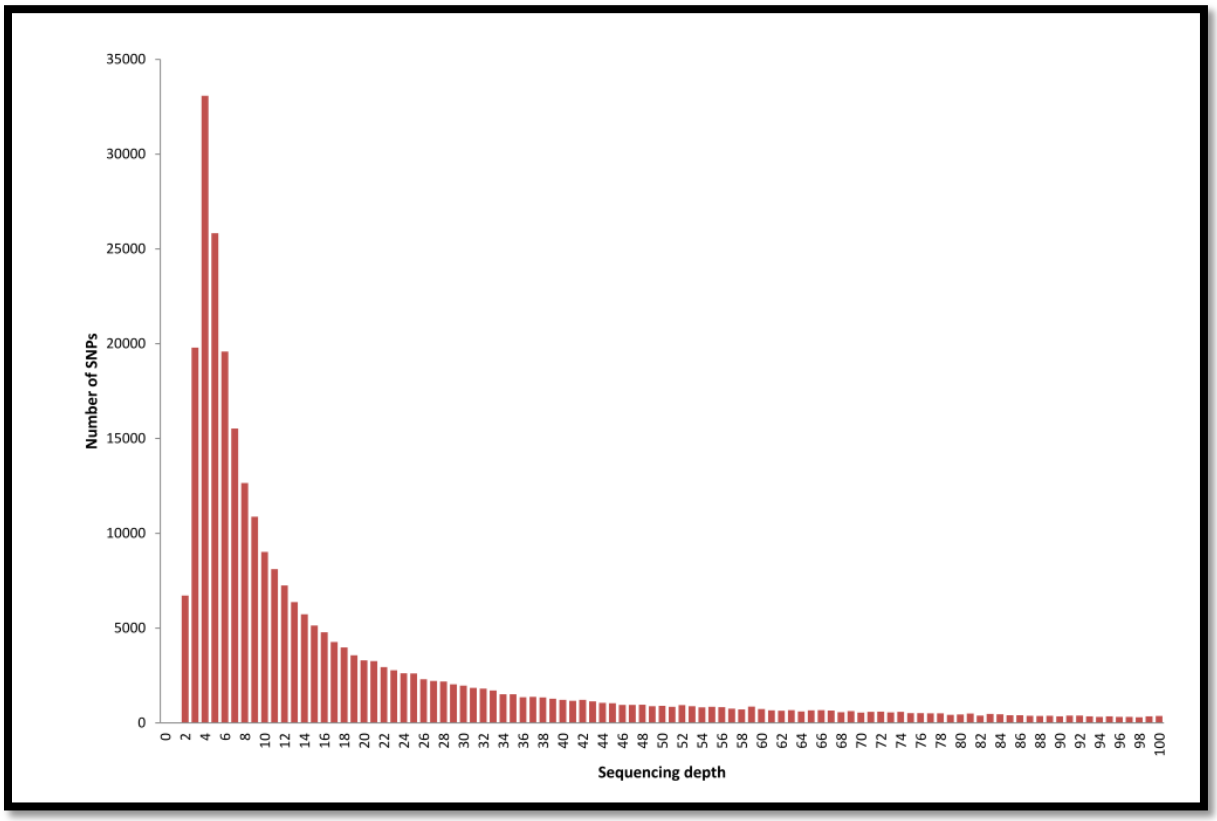


Figure 2.14 : distribution du nombre de SNPs identifiés (tronquée à une profondeur de 100) par DiscoSnp en fonction de la profondeur de séquençage de chaque SNP.

Un SNP identifié à partir d'un nombre élevé de reads aurait de forte chance de faire partie d'une famille de gènes multi-copies ou d'un élément répété.
A l'inverse, une couverture trop faible ne permet pas de distinguer, avec une confiance suffisante, un vrai polymorphisme allélique d'une erreur de séquençage, et est peu informative sur la fréquence de chaque allèle dans la population. La couverture moyenne est de 6 et la médiane se situe à 7. De ce fait nous avons choisi de conserver uniquement les SNP présentant une couverture entre 4 et 10 et avec au minimum 2 représentants de chaque allèle.
Cette restriction nous a permis de conserver 126 567 SNPs, soit 39.4% des SNPs identifiés par DiscoSnp.

b) Critère de qualité de séquençage

Comme nous l'avons déjà évoqué, chaque base séquencée par le 454 possède un indice de qualité évaluant la fiabilité de la lecture. Nous avons choisi de ne conserver que les SNPs d'une qualité de lecture supérieure ou égale à 30 et dont la séquence flanquante de 2k-1 (57 bases) possède une qualité moyenne supérieure ou égale à 30 (PHRED score).

c) Critère de qualité de séquences

Afin définir des amorces spécifiques à chacun des deux allèles dans les régions immédiatement adjacentes au SNP, la séquence proche du SNP doit être de bonne qualité pour permettre une bonne hybridation des amorces. Le principal problème posé par le 454 est la gestion des homopolymères (stretchs de nucléotides identiques). Ainsi nous avons choisi d'éliminer les SNPs dont les séquences flanquantes présentaient des homopolymères dans la séquence de 57 bases. Pour ceci nous avons utilisé une fenêtre glissante de 8 nucléotides et les séquences contenant plus de 5 nucléotides identiques dans cette fenêtre balayant l'ensemble de la séquence de 57 pb ont été éliminés.

Avec l'aide d'Olivier Quenez, qui a développé un script en python, ces tâches de sélection (tri sur la qualité et les homopolymères) ont pu être automatisées sur l'ensemble du jeu de données. Ainsi 9537 SNPs répondaient à nos critères.

d) critère de similarité

Pour éviter les allèles nuls liés à la présence de mutation dans les régions flanquantes au SNP, zones où sont définies les amorces spécifiques à chaque allèle, il est important de ne pas avoir de variations

ni d'indels dans les zones de design des amorces. Pour ceci nous avons utilisé le logiciel GASSST (Rizk & Lavenier 2010), logiciel permettant d'effectuer une recherche de similarité entre deux jeux de séquences. Pour ceci, l'ensemble des séquences reconstruites pour la découverte des SNPs, les séquences 2k-1, ont été confrontées aux séquences brutes issues du séquençage 454, considérées comme séquences référentes. Les séquences 2k-1 ont été donc repositionnées (par mapping) sur les séquences référentes, afin d'identifier des variations et les séquences 2k-1 pouvant provenir de mêmes reads originels. Pour la sélection de SNPs fiables, nous avons paramétré la détection de séquences à 80% de similarité et toléré la présence d'au maximum 2 gaps dans la séquence de 57pb (2k-1).

De plus, nous n'avons conservé qu'un seul SNP par séquences 'référentes' au sein de notre représentation réduite du génome d'*I. ricinus* afin de s'assurer d'une distribution maximale de nos SNPs.

Après cette étape finale de sélection, parmi les 996 508 séquences initiales et les 321 088 SNPs identifiés, 1 768 SNPs répondaient à l'ensemble de nos critères, soit 0.0055% des SNPs initiaux.

e) Restriction des SNPs à une sélection finale de 384 SNPs

384 SNPs ont été sélectionnés de manière à conserver un ratio de profondeur de séquençage similaire à celui des 1768 SNPs retenus jusqu'à présent (Figure 2.15). Ainsi, 58% des SNPs retenus parmi les 384 finaux présentent une profondeur de séquençage de 4 dans le jeu de données, alors que parmi les 1768 SNPs retenus initialement, on en observait 52%.

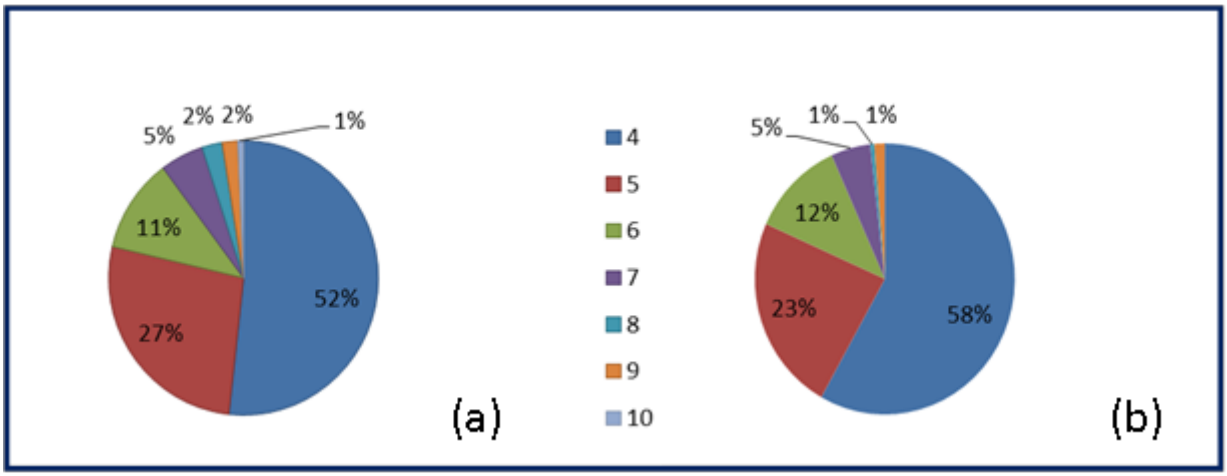


Figure 2.15 : Représentation graphique de la profondeur de séquençage par SNP dans le jeu de données initial des 1768 SNPs (a) et dans le jeu de données des 384 SNPs finaux (b).

f) Description des 384 SNPs

Parmi les 384 SNPs retenus, 254 SNPs, soit 66%, correspondent à des transitions et 130, soit 34%, à des transversions.

Vingt-deux présentent un polymorphisme uniquement au sein de la population M et 62 uniquement au sein de la population T. Les 300 autres SNPs présentent les deux allèles de chaque variant dans les deux populations.

La fréquence allélique minimale (MAF) moyenne est de 0,46 avec des valeurs allant de 0,22 à 0.50. Comme les individus séquencés n'ont pas pu être identifiés individuellement, ces valeurs sont à prendre avec précaution car elles sont basées sur des reads de pools de 10 et 20 individus

Dans le but d'identifier des SNPs provenant d'éventuelles séquences contaminantes (comme par exemple des séquences issues d'organismes hébergées par les tiques : virus, bactéries, protozoaires…) ou d'identifier des SNPs dans des séquences de tiques (notamment *Ixodes scapularis* pour laquelle il existe plus de 450 Mb séquencés) déjà connues, les 384 contigs réalisés par chevauchement des reads contenant les SNPs identifiés ont été analysés à l'aide de BLAST sur la banque de données Genbank. Pour identifier des homologies de séquences entre nos séquences et celles déposées sur Genbank, nous avons recherché toutes les séquences présentant un alignement d'au minimum de 10% entre deux séquences avec une similarité supérieure à 80%.

- 56.51% des séquences présentent, selon nos critères de recherche, de très fortes similarités avec *Ixodes scapularis*. Ce résultat suggère qu'un nombre important des SNPs que nous avons isolé pourrait être conservé chez *I. ricinus*. Par ailleurs, un nombre important de ces séquences correspondent à des régions codantes. Pourtant, on pouvait s'attendre a priori à une faible représentation des séquences codantes dans notre banque RRL constitué à partir d'ADN génomique où la fraction codante doit être relativement faible.

- 0.78% avec la tique *Rhipicephalus microplus* (espèce phylogénétiquement plus distante car appartenant non seulement à un autre genre mais aussi à une autre sous famille, celle des Rhipicephaline)

- aucune homologie n'a été trouvée avec le reste des séquences suggérant qu'aucun des SNPs retenus ne correspond à des génomes de microorganismes contaminants (dont beaucoup ont été complètement séquencés et sont donc présents dans les banques de données comme GenBank).

C. Génotypage haut-débit de 553 individus à l'aide du set de 384 SNPs

Suite à une première étape de mise au point et d'optimisation de la technique, deux runs de génotypage ont été réalisés en collaboration avec la plateforme GENTYANE située à Clermont-Ferrand. Le premier run a visé à génotyper des tiques issues de populations naturelles collectées sur le terrain dans le cadre du projet OSCAR. Le second run a permis de compléter le jeu de données des tiques génotypées lors du premier run, mais également de génotyper des tiques issues de croisements contrôlés afin de valider la ségrégation des SNPs à la génération suivante. L'ensemble de ces runs sera détaillé par la suite.

A l'image des technologies de séquençage, le génotypage a également connu une révolution au cours des années 2000 (Ding & Jin 2009; Davey *et al.* 2011). Différentes technologies de génotypage haut-débit sont maintenant proposées par diverses plateformes. Comme pour le séquençage, la principale évolution est la miniaturisation et la diminution des volumes de réactifs, qui permet de traiter en parallèle un grand nombre d'échantillons, de diminuer considérablement le coût et d'améliorer la sensibilité par rapport aux systèmes traditionnels (Maresso & Broeckel 2008).

Notre choix s'est arrêté sur l'utilisation du système Biomark HD de Fluidigm (Figure 2.16a). Ce système est un instrument de PCR en temps-réel, combiné à des puces 'Dynamic Array'. Il permet d'effectuer jusqu'à 9216 réactions de génotypage dans des volumes réactionels de quelques nanolitres (Wang *et al.* 2009b). Ces puces 'Dynamic Array' sont des puces composées de micro-circuits Intégrés de Fluidique (IFC) (Figure 2.16b) ce qui permet une diminution drastique des nombre de pipetages et des volumes, notamment d'ADN puisque quelques nanolitres suffisent.

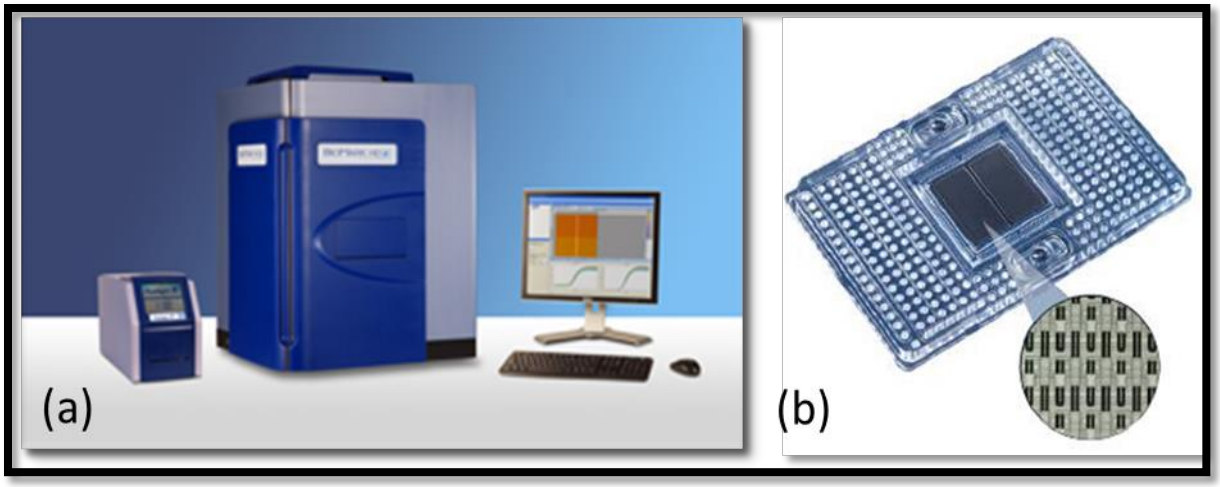


Figure 2.16 : (a) Le système Biomark HD de Fluidigm ; (b) exemple de puce IFC de Fluidigm. Ces puces disponibles en 48*48 ou 96*96 permettent de charger d'un côté de la puce 48 ou 96 échantillons – puits ADN (selon le format de la puce) -, et de l'autre côté 48 ou 96 amorces KASPar (ou Taqman)- puits SNPs-. Un réseau de micro canaux relie l'ensemble des puits 'ADN' et des puits 'SNPs' jusqu'à des micro-chambres réactionnelles dans lesquelles la réaction qPCR a lieu.

Afin de réaliser le génotypage, nous avons choisi d'utiliser la chimie KASPar de KBiosciences (KBiosciences Competitive Allele-Specific PCR SNP genotyping system)) (Cuppen 2007). Cette chimie est basée sur le principe de la PCR allèle-spécifique compétitive qui part de l'idée qu'une ADN polymérase peut difficilement incorporer un nucléotide dans un brin en cours de synthèse si le nucléotide n'est pas complémentaire de la matrice ou si la base extrême en 3' de l'amorce n'est pas correctement hybridée sur la matrice. Ainsi, en mettant en compétition deux amorces PCR marquées différemment (deux fluorochromes différents), correspondant chacune à leur extrémité 3' à l'un des allèles attendus à un SNP, il est possible de génotyper ce SNP.

1. Design des amorces compatibles avec la chimie KASPar

Le principe de la chimie KASPar étant basée sur une compétition d'allèle-spécifique, trois amorces sont nécessaire à la réaction de génotypage d'un SNP : deux amorces 'allèle-spécifique' à chacun des allèles du SNP et une amorce 'allèle non spécifique' qui s'hybride sur le brin complémentaire.

Le design des amorces, a été réalisé selon les critères recommandés par Fluidigm et Kbiosciences afin d'être compatible avec la technologie utilisée. J'ai également intégré mes propres critères afin d'optimiser au mieux l'hybridation des amorces, en supprimant notamment:

- les SNPs non-bialléliques, la chimie utilisée ne mettant en évidence que deux allèles différents.

- les SNPs proche de microsatellites, afin d'éviter de définir des amorces dans des régions répétées qui pourraient s'hybrider à plusieurs endroits dans la séquence.
- les SNPs dans des séquences trop riches ou trop pauvres en GC%, le génotypage se réalisant sur une seule et même puce et donc les températures d'hybridation devant être assez homogènes pour l'ensemble des amorces (~60°C).

A notre connaissance et à l'heure actuelle, aucun logiciel de design d'amorces en libre d'accès ne permet d'automatiser cette tâche fastidieuse, soumise à de nombreuses contraintes. En effet pour être compatible avec la chimie KASPar, les amorces dites spécifiques doivent finir en 3' par le SNP et l'amorce dite non spécifique doit être distante que de quelques nucléotides. Il est donc parfois difficile de pouvoir ajuster les températures d'hybridation, la longueur de l'amorce et d'éviter les problèmes d'hybridation des amorces. Les amorces pour 384 SNPs ont donc été définies manuellement en utilisant le logiciel Perl Primer (Marshall 2004) afin d'être compatible avec la chimie Kaspar (Figure 2.17).

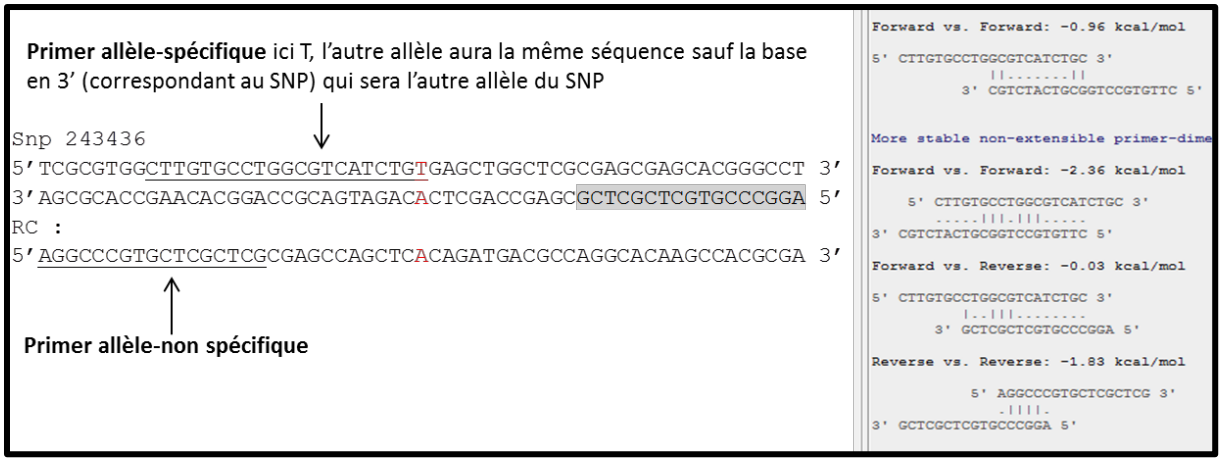


Figure 2.17 : Exemple de design d'amorce pour un SNP et des vérifications effectuées relatives à l'hybridation entre amorces.

Après ces vérifications, les séquences des amorces ont été envoyées chez IDT (Integrated DNA Technologies) qui a réalisé la synthèse des oligonucléotides, en ajoutant une queue en 5' compatible avec la chimie KASPar et incorporant les fluorochromes VIC ou FAM pour les amorces spécifiques des deux allèles de chaque SNP.

2. Echantillons

Dans le cadre du projet OSCAR, des tiques ont été collectées dans la zone atelier Armorique (ZAA) près de Pleine-Fougères, en Bretagne. Parmi l'ensemble des tiques collectées, 550 tiques, toutes au stade nymphal, ont été sélectionnées de façon à être représentatives du maximum de points de prélèvements échantillonnées. Sur ces nymphes, les études comprenaient l'analyse des repas sanguin, le génotypage SNP, la recherche de trois agents pathogènes et le génotypage microsatellite, le tout réalisé par les différents laboratoires partenaires du projet.

Du fait des différentes études de ce projet entre plusieurs équipes de recherche, les extraits d'ADN ont été divisés en six. Une première extraction a été réalisée à l'ammoniac (Humair *et al.* 2007), permettant de récupérer la majorité de l'ADN extrayable d'une tique (environ 80%) sans étape de broyage mécanique des tissus de la tique. Cet extrait a ensuite été partagé en quatre. Dans un second temps, la carcasse de chaque individu restant à l'issue de l'extraction à l'ammoniaque, a été broyée au Tissue Lyser de Qiagen et l'ADN extrait à l'aide du kit NucleoMag de Macherey-Nagel. La solution d'ADN issue de cette extraction a ensuite été divisée en deux et nous avons utilisé un des deux aliquots, l'autre étant destiné à une analyse de génétique des populations basé sur des marqueurs microsatellites.

Au final nous avons obtenu par dosage au picogreen une moyenne pour l'ensemble de nos individus (demi-carcasse de nymphe) de 0,57ng/µl d'ADN pour un volume avoisinant les 25µl. Nous avons donc en moyenne une quantité d'ADN de 14,25ng par individu. Pour le génotypage de 384 SNPs en Fluidigm, une quantité de 2,4µg est préconisée, soit quasiment 170 fois plus. Nous avons donc dû réaliser une 'pré-amplification' de l'ADN.

3. Mise au point de la technique d'amplification du génome

Cette technique de pré-amplification du génome n'ayant jusqu'à présent jamais été testée sur l'ADN de tique ni combinée avec la technologie Fluidigm et les amorces KASPar, une étape de mise au point de la technique a été nécessaire. L'amplification du génome a donc été testée, ainsi que la compatibilité avec les amorces. Les premiers tests ont été réalisés en qPCR puis en puce Fluidigm 48*48 afin de vérifier l'efficacité de la méthode ainsi que sa compatibilité avec le système Fluidigm.

a) Validation de l'amplification d'ADN à l'aide d'un kit WGA (Whole Genome Amplification)

Afin d'amplifier l'ADN génomique des nymphes de tique dont nous disposions, nous avons choisi d'utiliser le kit Primer extension pré-amplification (PEP-PCR) (Kbiosciences) qui permet l'amplification de l'entièreté du génome. Pour tester le kit, l'ADN de huit individus -provenant de tiques maintenues au laboratoire- a été extrait dans les mêmes conditions que pour l'expérimentation finale (cf. extraction à l'ammoniac suivie d'une extraction de l'ADN de la carcasse par le kit NucleoMag de Macherey-Nagel).

Comme nous pouvons le voir dans le tableau ci-dessous (Tableau 2.5), le recours au WGA a permis d'augmenter substantiellement la quantité d'ADN disponible pour réaliser les génotypages ultérieurs : en moyenne nous obtenons une quantité 42 fois supérieure qu'au départ. Cependant malgré l'amplification réalisée, la quantité d'ADN reste inférieure aux préconisations de Fluidigm.

Tableau 2.5 : Récapitulatif des concentrations et quantité d'ADN initiales, suite à l'amplification WGA et après purification pour les 8 individus (Ech) ; [ng/µl] correspond à la concentration,

Ech	[ng/µl] (15µl)	Quantité initiale	[ng/µl] (50µl)	Quantité post-WGA (ng)	Taux d'amplification	[ng/µl] après Ampure (30µl)	Quantité après Ampure (ng)
1	0,686	10,29	59,6	2980	**289,60**	24,6	23,91
2	0,314	**4,71**	**45,5**	**2275**	**483,01**	**10,2**	21,66
3	0,154	**2,31**	**50,2**	**2510**	**1086,58**	**16,4**	71
4	0,123	**1,845**	**49**	**2450**	**1327,91**	**10,3**	55,83
5	0,283	**4,245**	**56,2**	**2810**	**661,955**	**13,1**	30,86
6	0,169	**2,535**	**54,5**	**2725**	**1074,95**	**19,8**	78,11
7	0,115	**1,725**	**35**	**1750**	**1014,49**	**4,5**	26,09
8	0,208	**3,12**	**39,5**	**1975**	**633,012**	**9,4**	30,13

Les ADNs amplifiés ont été déposés sur gel d'agarose 1% pour les 7 premiers échantillons afin de vérifier que le WGA avait amplifié l'ensemble des fragments des différentes tailles (Figure 2.18). Comme nous pouvons le voir sur la photo de gel, l'amplification a généré un smear et ne montre pas de bandes spécifiques qui auraient été synonymes d'amplification préférentielle de certains fragments. L'amplification a donc été homogène.

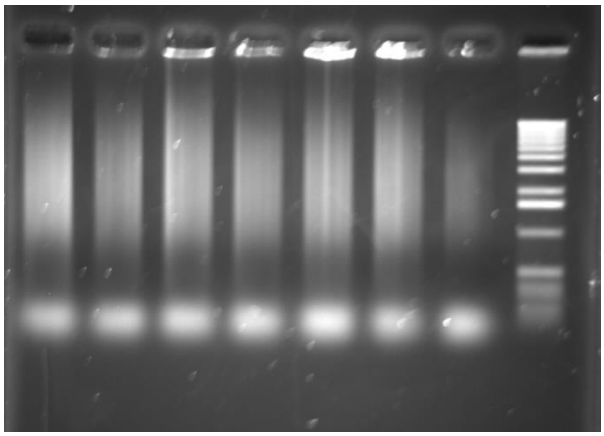


Figure 2.18 : Photo de gel d'électrophorèse des produits d'amplification WGA pour 7 échantillons d'ADN.

Cependant, comme nous pouvons le voir sur la photo du gel (Figure 2.18), il reste beaucoup de dimères d'amorces dans les échantillons qui conduisent à une surestimation de la quantité réelle d'ADN génomique amplifiée de chaque échantillon lors du dosage. De plus, les sels et résidus des réactifs de PCR peuvent également poser des problèmes lors du génotypage avec le système de puces en canaux microfluidiques qui sont très sensibles aux impuretés pouvant rester dans les solutions analysées. De ce fait nous avons purifié les ADNs avec le kit Ampure, un système de purification sur microbilles et qui peut être réalisé en plaque 96 puits. Nous obtenons une quantité finale de 17 fois supérieure à celle initiale, suite à l'étape de purification.

b) Validation de la compatibilité du WGA et des amorces KASPar

Afin de tester la compatibilité des amorces KASPar avec les échantillons issus du WGA, nous avons réalisé un premier test sur un appareil de PCR quantitatif, le LC480. Ce test permet de valider d'une part la compatibilité WGA/amorce KASPar et d'autre part le design et hybridation correcte des amorces.

Les 8 échantillons testés ont été dilués à 10ng/µl et 12 marqueurs ont été utilisés. Les résultats sont présentés sur la figure 2.19, chaque point correspondant à un individu pour un SNP.

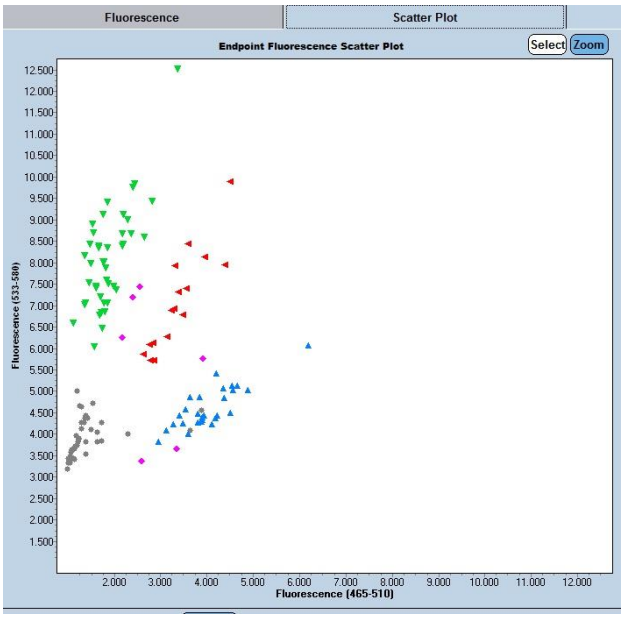


Figure2.19 : Fluorogramme (Scatter plot) obtenu suite à la lecture au LC480 du génotypage de 8 individus pour 12 SNPs. Les points verts et bleus correspondent à des individus homozygotes, les points rouges correspondent à des hétérozygotes, les points roses à des points de génotypage non assignés dû à une lecture ambiguë et les points gris à des points de génotypage de témoins et/ou non amplifié ou présentant un signal trop faible de lecture.

Les 12 marqueurs testés ont tous fonctionnés et montrent l'hybridation des deux amorces spécifiques sur chaque variant (présence d'hétérozygotes ou des deux types d'homozygotes). De ce fait, le protocole WGA semble adapté avec les amorces KASPar.

c) Validation de la combinaison WGA-KASPar avec le système Biomark HD de Fluidigm

Afin de tester la compatibilité de nos amorces et de notre protocole utilisant le WGA sur le système Fluidigm, nous avons réalisé un génotypage sur une puce 'Dynamic Array' 48x48. Pour ceci, 42 échantillons des tiques OSCAR ont été sélectionnés ainsi que 48 nouveaux marqueurs. Parmi ces individus, quatre ont été dupliqués pour tester la répétabilité de l'amplification WGA.

Chaque ADN a été amplifié selon le protocole WGA, puis purifié par Ampure. Les concentrations étant faibles, les échantillons ont été concentrés par évaporation au Speed vac et les ADNs ont été repris dans 20µl d'eau Ultra Pure. Après dosage au Picogreen, la concentration des échantillons a été normalisée (dans la mesure du possible, certains échantillons étaient trop faibles) à 60 ng/µl. La puce 48x48 a été préparée selon le protocole « Genotypage SNP par la Chimie KASPar » sur le Biomark.

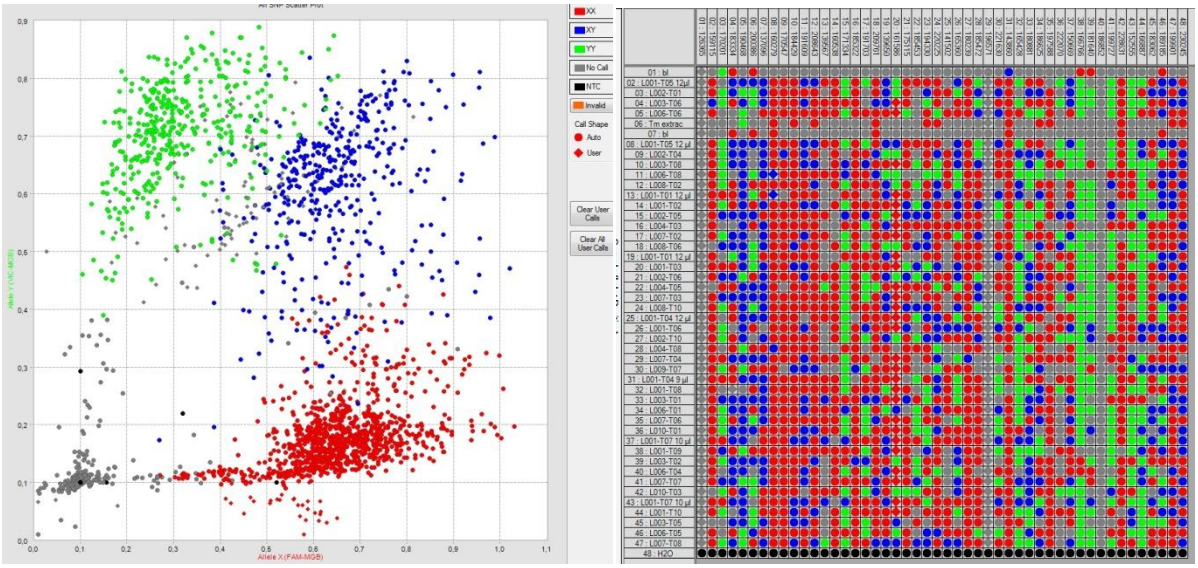


Figure 2.20 : A gauche, Fluorogramme (Scatter Plot) permettant de visualiser les 2304 points de génotypage selon l'intensité de la fluorescence émise. A droite, le plan de la puce, où les individus correspondent aux lignes et les SNPs aux colonnes. Les homozygotes sont présentés en vert (YY) et rouge (XX) et les hétérozygotes (XY) en bleu, les points gris correspondent aux non assigné (NC – pour No Call-).

Comme nous pouvons le voir sur la figure 2.20, ce test a été très concluant puisque les points de fluorescence sur le fluorogramme sont dans l'ensemble dissociables en fonction des hétérozygotes et des homozygotes. Un seul SNP n'a donné aucune amplification, il sera par la suite remplacé par un autre. Cependant certains points de génotypage ont été inférés à un génotype par le logiciel d'analyse de Fluidigm mais ne correspondant pas au signal de fluorescence (présence de quelques points bleus dans le 'nuage' de rouge par exemple). Pour cette raison, une vérification manuelle point par point a été nécessaire.

Afin de valider définitivement la technique et le protocole mis en place, nous avons réalisé un deuxième test, également sur une puce 48*48. En suivant toujours le même protocole, 36 nouveaux marqueurs ainsi que les 12 premiers (testés en qPCR avec le LC480) ont été testés sur 37 nouveaux individus (tiques OSCAR) ainsi que sur les huit premiers échantillons testés en qPCR. Ce deuxième génotypage de ces huit individus a permis de valider la reproductibilité des résultats. Les génotypages obtenus au cours de ce deuxième test utilisant une puce Fluidigm ont été satisfaisant puisque l'ensemble des 48 marqueurs et des 48 individus ont pu être génotypés.

4. Réalisation du premier run de génotypage sur 464 tiques *I. ricinus*

Pour le premier run de génotypage en puce Fluidigm 96*96, 464 individus ont été sélectionnés parmi les 550 tiques dont l'ADN a été extrait. La description biologique de ces 464 individus sera abordée dans le chapitre trois.

Nous avons réalisé 20 puces Fluidigm 96x96 afin de croiser l'ensemble des individus (464-répartis sur cinq puces) avec l'ensemble des marqueurs (384-répartis sur quatre puces).
Les 464 individus ont été choisis en fonction de leur quantité d'ADN avant amplification au WGA parmi les 550 disponibles. En effet, nous avions, pour l'ensemble des individus, une quantité moyenne d'ADN de 14,25ng (concentration de 0,57ng/µl) alors qu'il est préconisé une quantité de 2400ng afin de génotyper cinq puces de 96x96 à une concentration de 60ng/µl.
L'amplification WGA a permis d'obtenir des quantités d'environ 800ng d'ADN par individus (en moyenne), soit 57 fois plus qu'initialement, mais tout de même très en dessous des préconisations des constructeurs.

Les individus ont été répartis sur cinq plaques 96 : pour l'ensemble des 480 puits disponibles nous avions 464 individus et 16 témoins. Au premier témoin, nommé NTC (No Template Control) qui permet le calibrage et l'étalonnage de la puce par le système Biomark, des témoins d'extraction et des témoins de WGA ont été ajoutés pour contrôler l'absence de contaminations. Nous avons effectué le même protocole que précédemment, qui est :

- dosage au Picogreen et sélection des 464 individus les plus concentrés
- amplification WGA
- dosage de l'ADN post-WGA
- purification de l'ADN à l'Ampure
- concentration de l'ADN
- génotypage

5. Réalisation d'un deuxième run de génotypage

Un second run de génotypage en puce 96x96 a été réalisé pour différents objectifs :

- compléter le jeu de données du premier run en incluant 29 tiques OSCAR non analysées lors du premier run.
- analyser la ségrégation des SNPs en incluant 57 individus issus de cinq croisements réalisés en conditions contrôlées au laboratoire (Tableau 2.6).
- tester la reproductibilité des résultats avec ou sans WGA en incluant trois femelles *Ixodes ricinus* (origine de Chizé).

L'objectif du séquençage d'individus issus de cinq croisements réalisés en conditions contrôlées était :

- analyser la ségrégation des allèles à la génération suivante en se basant sur le génotype des deux parents

- reconstituer une carte génétique des marqueurs développés par analyse d'association entre marqueurs dans la ségrégation à la génération suivante.

- valider les marqueurs à une échelle plus large, les tiques provenant d'origines variées (tiques tunisiennes et de différentes origines françaises).

Tableau 2.6 : Description des croisements réalisés en conditions contrôlées.

	Nom du croisement	Gorgement		Père			Mère				descendance	
		Date	Hôte	Origine	Extraction		origine	extraction		Test WGA	nb	stade
					Matériel	Date		matériel	date			
1	C030	2008	Veau	Tunisie	Carcasse sans pattes (-80°C)	?	Toulouse	Carcasse sans patte (-80°C)	05/2013	oui	10	Nymphe vivante
2	C212	09/2008	Veau	Tunisie	Carcasse sans pattes (-80°C)	05/2013	Belle-Ile	Carcasse sans pattes (-80°C)	05/2013	oui	10	Nymphe vivante
3	C214	09/2008	Veau	Tunisie	Carcasse sans pattes (-80°C)	05/2013	Belle-Ile	Carcasse sans pattes (-80°C)	05/2013	non	9	Nymphe vivante
4	C243	09/2008	Veau	Chizé	Pattes (-20°C)	2010	Chizé	Pattes (-20°C)	2010	non	9	Larve congelée à -80°C (7) + nymphe vivante (2)
5	H4I	12/2012	Lapin	Chizé terrain 2012	Carcasse entière (-80°C)	05/2013	Chizé (C243)	Carcasse entière (-80°C)	05/2013	oui	9	Larve vivante

6. Résultats obtenus suite au génotypage de 553 individus *I. ricinus*

Les résultats du génotypage peuvent être représentés graphiquement de façon synthétique à l'aide du logiciel **Fluidigm SNP Genotyping Analysis.**

Les résultats de génotypage sont représentés de différentes façons :

- Un graphique (Scatter Plot)

Le scatter Plot représente l'ensemble des points de génotypage de la puce (9216 points de génotypage) (Figure 2.21). Il est donc possible de voir très rapidement la qualité du génotypage en fonction de la distribution des points de génotypage et du nombre de chacun.

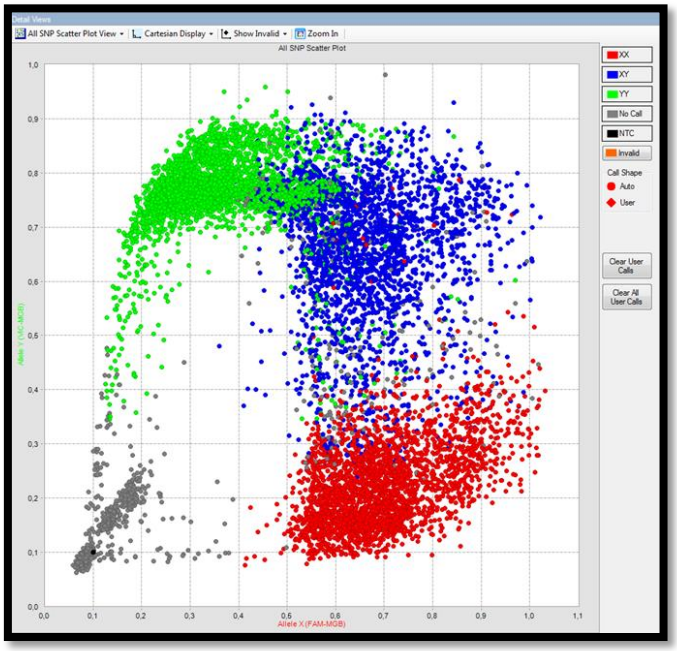


Figure 2.21 : Exemple d'un Scatter Plot pour une puce 96*96 représentant 9216 points de génotypage ; Les homozygotes sont présentés en vert (YY) et rouge (XX) et les hétérozygotes (XY) en bleu, les points gris correspondent aux non assigné (NC –pour No Call-).

- L'image de la puce globale

L'image de la puce combine l'ensemble des points de génotypage avec les SNPs en abscisses et les individus en ordonnées. Cette visualisation permet de distinguer de manière très rapide les SNPS (en colonne) ou les individus (en ligne) qui n'ont pas donné de résultats ou qui accumulent beaucoup de données manquantes (points gris) (Figure 2.22).

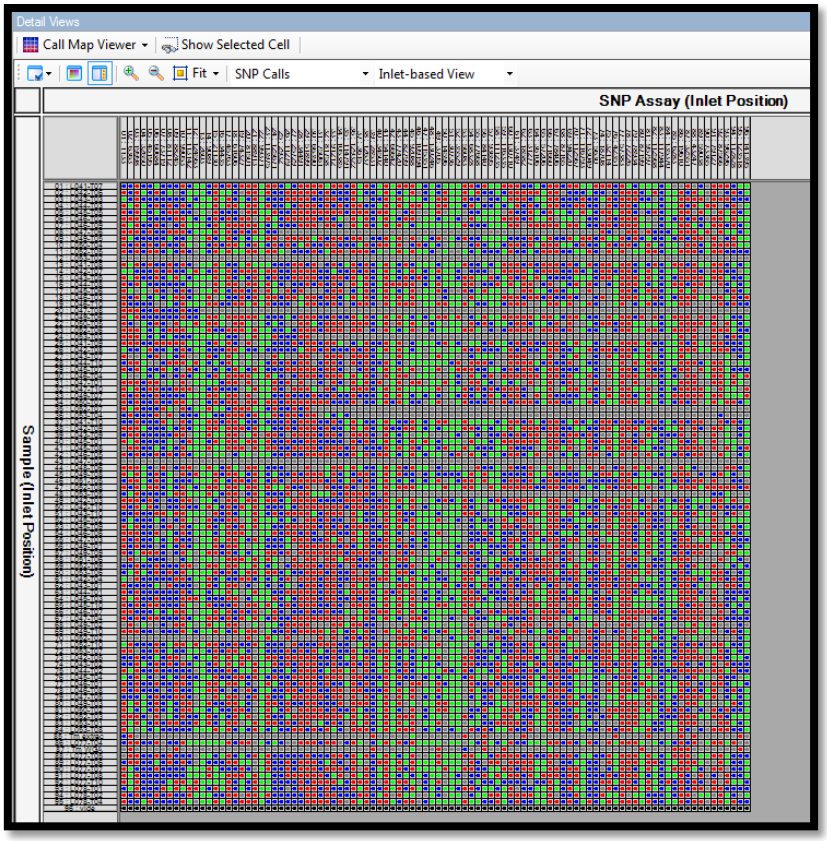


Figure 2.22 : Exemple d'une représentation graphique synthétique pour une puce 96*96 représentant 9216 points de génotypage (points bleus : hétérozygotes ; points verts et rouges : homozygotes ; points gris : données manquantes [signal trop faible]).

- Un fluorogramme ou graphique de fluorescence

La lecture des résultats peut également se faire pour chacun des SNPs génotypés par puce (96 points par graphique) (Figure 2.23). Comme le montre la figure 2.23, des points gris ('No Call') sont visibles au sein d'un nuage de points où il semble ne pas y avoir d'ambiguïté quant à l'assignation du génotype de ces points. Une lecture assidue de chaque graphique est donc nécessaire afin de vérifier les résultats générés par la lecture du logiciel.

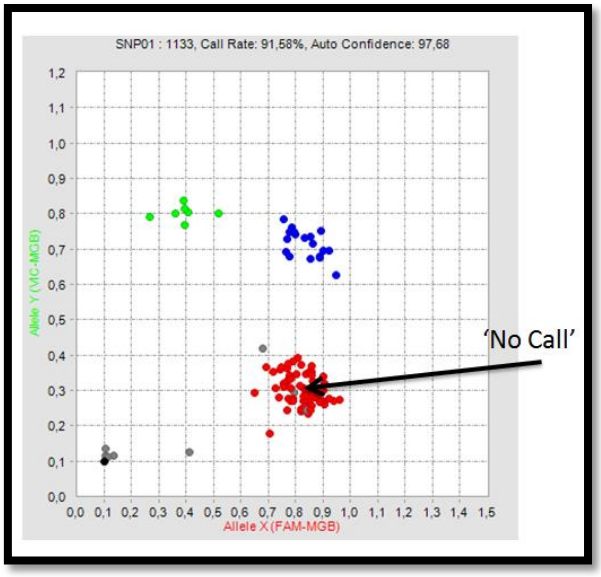


Figure 2.23 : Exemple d'un graphique de fluorescence pour un SNP sur une puce (96 points de génotypage). Les homozygotes sont présentés en vert (YY) et rouge (XX) et les hétérozygotes (XY) en bleu, les points gris correspondent aux non assigné (NC –pour No Call-).

La sortie des données de génotypage se fait dans un fichier csv, où l'information est résumée par ligne pour chaque point de génotypage. Une ligne correspondant donc à un SNP et à un individu, les fichiers en sortie de run de génotypage 96*96 font donc 9216 lignes.

Pour pouvoir traiter ces données plus facilement, un script en language Bash a été rédigé afin de convertir ces fichiers de sortie en tableau croisé (96 SNPs x 96 individus) et de donner le génotype de chaque SNP par individu sur une seule et même ligne (script en annexe XX).

L'analyse des 24 puces, soit des 221 184 points de génotypage, a montré :

- pour le premier run de génotypage : sur l'ensemble des 384 marqueurs génotypés pour les 464 individus, 368 marqueurs montrent une amplification des deux allèles, cinq marqueurs n'ont pas fonctionné et 11 ont amplifié un seul allèle du SNP.

- pour le second run de génotypage, 47 SNPs n'ont pas donné d'amplification ou de signal interprétable. L'ensemble des autres SNPs a montré une amplification des deux allèles.

D. Analyses des résultats du génotypage des 384 SNPs

Les technologies de génotypage haut débit des SNPs sont récentes. Elles ont été essentiellement développées pour des organismes modèles pour lesquels les SNPs ont été précédemment validés par différentes approches. Par ailleurs, elles sont utilisées sur des échantillons ne présentant pas de quantité d'ADN limitante. Notre cas était donc assez différent puisque nous avons développé de nouveaux marqueurs pour *I. ricinus* et nous disposions d'échantillons avec de faible quantité d'ADN.

Pour pallier le problème des quantités faibles d'ADN, un protocole expérimental complexe couplant amplification du génome et les technologies de génotypage haut-débit a été développé. Il est nécessaire d'explorer les résultats obtenus afin de valider ces marqueurs pour qu'ils puissent être utilisés par la suite.

Pour valider ce protocole, différents duplicats ont été réalisés au cours du génotypage afin d'effectuer une analyse critique de nos résultats. Les points critiques du projet sont (i) l'amplification WGA, (ii) le génotypage par le Biomark de Fluidigm.

1. Effet du WGA

Nous n'avions aucun recul sur cette technique d'amplification car elle était utilisée pour la première fois sur le génome de la tique *I. ricinus*. Kbiosciences, le fournisseur du kit utilisé, préconise une quantité d'ADN minimale de 50 ng pour une réaction afin d'obtenir une amplification de 500 à 1000 fois.

Etant donné la quantité initiale d'ADN extrêmement faible, environ 15 ng par échantillon, nous n'avons eu qu'une amplification de 57 fois plus d'ADN en moyenne par échantillon, soit un rendement de 11% du kit. Il était donc nécessaire de vérifier si ce faible rendement, bien qu'il se soit avéré suffisant pour obtenir des points de génotypage de 384 SNPs, n'apportait pas de biais dans l'analyse. Ainsi nous avons réalisé des duplicats afin de vérifier la reproductibilité du WGA :

- duplicats d'individus génotypés à partir de deux PCR WGA différents.
- duplicats d'individus génotypés avec et sans WGA.

Pour ceci, nous avons utilisé comme outil de comparaison, le nombre de loci différents pour un même SNP entre les duplicats mais également le nombre de données manquantes.

a) Duplicats de quatre individus génotypés sur une même puce à partir de deux pré-amplifications WGA différentes

Lors de la réalisation du premier run de génotypage en puce Fluidigm 48x48, quatre individus, génotypés pour 48 SNPs, ont été dupliqués. Chaque duplicat a subi une pré-amplification avant d'être génotypé.

Une grande variabilité (rapport de 1 à 10) dans les concentrations des ADN de tiques a été obtenue. Cette variabilité est principalement due à l'étape d'extraction d'ADN par l'ammoniac. En effet comme elle est hétérogène, il était difficile d'obtenir un rendement similaire entre individu lors de l'extraction d'ADN des carcasses de tiques. De plus on ne remarque pas de proportionnalité entre la quantité d'ADN initiale et la quantité d'ADN obtenue après WGA. Par exemple les individus L001-T01 à 0,23ng/µl et L001-T05 à 2,68ng/µl présente une concentration d'ADN WGA du même ordre – 23,3 et 27,9 ng/µl respectivement– alors qu'il y avait initialement 11,6 fois plus d'ADN chez T05 que chez T01) (Tableau 2.7).

Tableau 2.7 : Récapitulatif, pour les 4 individus dupliqués, des concentrations d'ADN initiale, suite aux 2 pré-amplifications des duplicats et du nombre de loci différents observés entre les génotypages des deux duplicats.

	[ng/µl] initiale (~25µl)	[ng/µl] WGA du premier duplicat	[ng/µl] WGA du second duplicat	nombre de loci différents entre les 2 génotypages
L001-T01	0.23	23.3 (12µl)	38.7 (12µl)	4/48
L001-T04	0.71	27.9 (12µl)	39.3 (9µl)	1/48
L001-T07	0.49	31.5 (10µl)	38.3 (10µl)	0/48
L001-T05	2.68	27.9 (12µl)	39.8 (12µl)	1/48

Malgré cette hétérogénéité d'extraction et l'utilisation de quantités d'ADN inférieures à celles préconisées par les fabricants, nous observons entre zéro et quatre loci différents entre les deux duplicats des différents individus sur les 48 SNPs analysés, avec une moyenne de 1,5 loci différents, soit une variabilité de 3,1% dans les résultats (Tableau 2.8).

Cependant, l'effectif reste tout de même trop faible pour pouvoir généraliser ces résultats à l'ensemble des marqueurs utilisés dans le reste de l'étude.

Tableau 2.8 : Récapitulatif pour les 27 individus dupliqués des concentrations d'ADN initiale, suite aux deux pré-amplifications des duplicats, des données manquantes observées pour les deux duplicats et du nombre de loci différents observés entre les génotypages des deux duplicats.

individu	Concentration initiale [ng/µl]	concentration du premier duplicat [ng/µl]	concentration du second duplicat [ng/µl]	données manquantes observées pour le premier duplicat	données manquantes observées pour le second duplicat	nombre de loci différents entre les 2 génotypages
L001-T09	0,5	37,2	10,72	11	9	0
L002-T10	1,21	31,5	8,59	6	7	0
L001-T02	1,57	29,4	23,74	8	8	1
L002-T06	0,74	40,6	15,45	4	8	1
L004-T05	1,06	29,4	10,31	9	12	1
L007-T03	0,39	35,9	14,57	8	9	1
L001-T03	0,5	41,6	14,52	7	14	2
L001-T10	0,41	42	10,33	9	11	2
L002-T04	0,56	39	10,76	10	10	2
L003-T01	0,85	42,4	10,14	8	10	2
L006-T08	0,19	39	7,36	12	15	2
L007-T02	0,79	35,8	9,06	9	11	2
L007-T08	1,1	34,6	10,69	5	6	2
L001-T06	1,06	33,7	15,94	11	9	3
L003-T02	0,24	36,6	12,06	9	11	3
L003-T08	1,43	36,5	12,12	7	7	3
L007-T04	0,4	38,1	9,14	9	9	3
L007-T07	0,41	34,6	9,64	7	7	3
L002-T01	0,69	29,5	12,86	10	9	4
L003-T05	0,19	43,6	6,87	15	25	5
L001-T08	0,78	42,8	11,41	6	7	6
L009-T07	0,22	37,1	3,75	15	10	7
L010-T03	0,44	38,7	9,15	10	12	7
L008-T06	0,32	35,7	11,47	11	9	8
L008-T02	0,63	37,6	8,27	7	12	13
L008-T10	0,34	33,1	6,17	7	16	15
L010-T01	0,52	30,6	6,31	7	7	17
moyenne		***36,540***	***10,797***	***8,7***	***10,3***	***4,2***

b) Duplicats de 27 individus génotypés à partir de deux pré-amplifications WGA différentes sur deux puces.

Vingt-sept individus ont été génotypés deux fois pour 45 marqueurs. Deux duplicats ont été effectués, une première série sur une puce 48*48 lors des tests préliminaires et une seconde série sur une puce 96*96 (Tableau 2.8). Pour 74 % des individus, le nombre de loci dont le génotypage diffère entre les deux runs est inférieur ou égale à cinq (soit 10% des loci testés). En revanche, trois individus présentent un taux élevé (plus de 27%) de loci présentant une différence entre les deux runs.

Le nombre de loci différents entre les deux duplicats peut être dû à une mauvaise interprétation des graphiques de fluorescence à la sortie du logiciel. Cependant ces derniers ont été vérifiés pour l'ensemble des deux puces concernées. Comme on peut le voir sur l'exemple suivant (Figure 2.24) il ne semble pas y avoir de biais dans la lecture et l'interprétation du signal émis.

Pour la première PCR WGA de l'individu pris en exemple (L008-T10), 15µl ont été utilisés pour l'amplification, soit une quantité de 5,1ng alors qu'une quantité minimale de 50ng est préconisée pour une amplification optimale par WGA.

Pour ce même individu, lors de la deuxième amplification (pour la puce 96), 5,6µl ont été utilisé pour la PCR WGA, soit une quantité inférieure à 2ng (1,904ng) pour réaliser l'amplification. Si cet individu est hétérozygote XY, les différences observées entre les 2 runs pourraient être dues à amplification préférentielle d'un allèle pour ce SNP donné du fait des faibles quantités d'ADN initiale. L'amplification étant exponentielle, le signal pour l'autre allèle serait d'une intensité trop faible pour pouvoir être détecté.

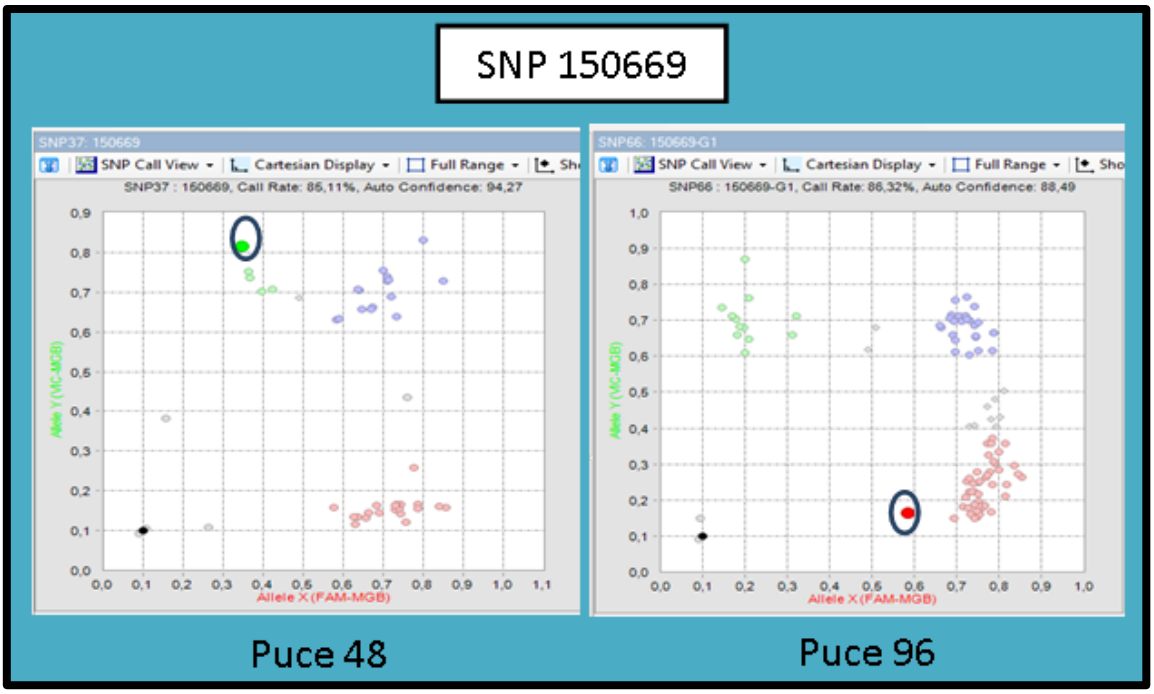


Figure 2.24 : Le même individu (L008-T10) mis en surbrillance dans le premier cas sur la puce48 où il apparait 'YY' et dans le 2ème cas sur la puce96 où il apparait 'XX'.

Comme nous venons de l'évoquer une possible explication de ces différences observées pour un même individu pourrait provenir des faibles quantités d'ADN. Ainsi, une corrélation négative (mais faible) est observée entre la quantité d'ADN initiale et le nombre de loci différents (Figure 2.25). Cette faible corrélation pourrait aussi être liée à un effet 'seuil' car les individus qui avaient le plus d'ADN au départ ne sont jamais les individus avec le plus grands nombre de loci différents.

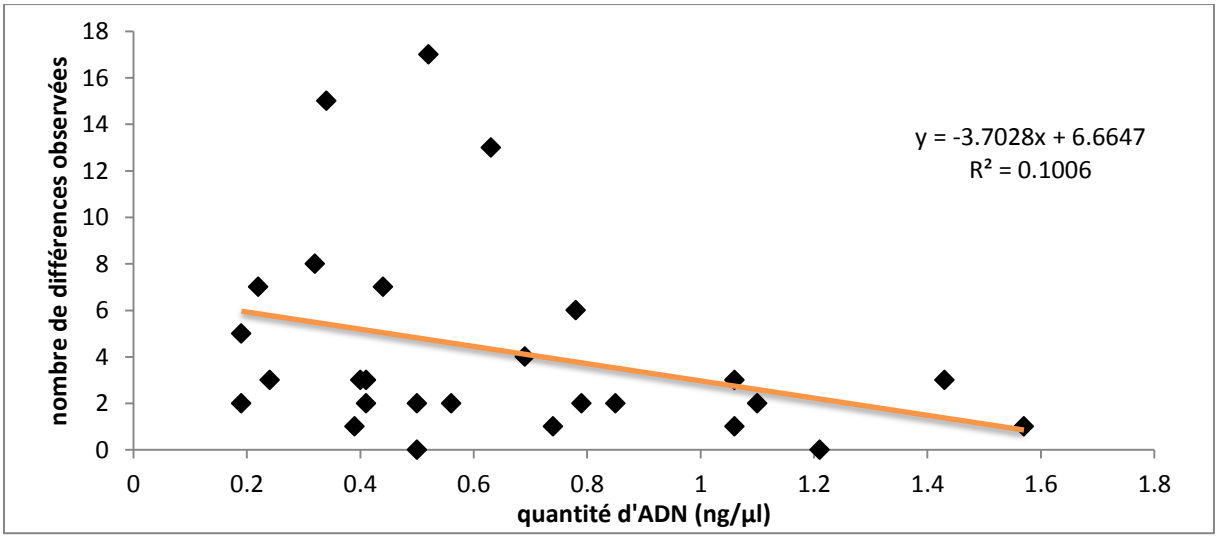


Figure 2.25 : Relation entre le nombre de différences observées entre les deux réplicats des 27 individus génotypés en fonction de la quantité d'ADN initiale de chaque individu pré-WGA.

Cette observation suggère que les différences constatées entre les deux WGA pourraient surtout avoir lieu lorsque la PCR est réalisée sur une quantité faible d'ADN. Le WGA pourrait ainsi être sensible à un effet du type 'amplification préférentielle de l'un ou de l'autre allèle' de la solution de départ (au gré des dilutions et du 'tirage' aléatoire d'une ou l'autre de ces rares copies dans la solution de départ).

On peut se demander également s'il n'y a pas un 'effet SNP' qui pourrait expliquer ces résultats (c'est-à-dire que tous les loci ne seraient pas sensibles de la même façon aux biais potentiellement engendrés par le WGA). Cet 'effet SNP' pourrait être dû à un design d'amorces qui ne serait pas dans une séquence unique mais dans une séquence multi-copie, ce qui pourrait expliquer les différences observées entre deux génotypages. Cependant l'ensemble des différences observées ne semblent pas être porté par un ou plusieurs SNPs (Figure 2.26). En effet, mis à part deux SNPs qui concentrent 11 et 12 individus présentant un génotypage différent entre les duplicats, on observe une diminution progressive du nombre de loci présentant des différences. La distribution ne correspond cependant pas exactement à une distribution binomiale négative comme attendue (avec un déficit de loci à seulement 1 ou 2 différences et un excès de loci à 3 différences).

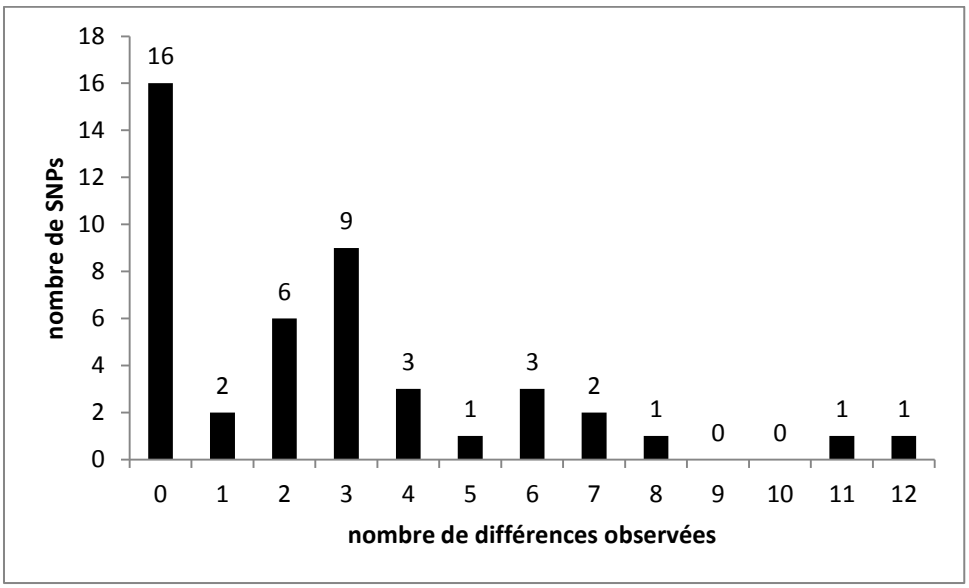


Figure 2.26 : Histogramme présentant le nombre de SNPs concernés par l'observation de loci différents entre les 2 génotypages.

c) Duplicats de 6 femelles génotypées avec et sans pré-amplification WGA

Afin de tester la reproductibilité des résultats avec ou sans WGA, trois des femelles issues des croisements réalisés en conditions contrôlées ainsi que les trois femelles *Ixodes ricinus* (origine de Chizé) génotypées lors du deuxième run de génotypage ont été génotypées avec ou sans une pré-amplification WGA.

Les résultats obtenus sont répertoriés dans le tableau 2.9. De manière générale, nous pouvons voir qu'on obtient une reproductibilité d'un minimum de 95% entre les génotypages avec ou sans WGA.

Pour trois femelles (C212, C214, H4I) on observe une diminution du nombre de données manquantes (DM) avec la pré-amplification WGA ainsi qu'un nombre plus faible de différences observées entre le génotypage avec ou sans WGA (taux de différences entre 1,04 et 1,82%).

Les trois autres femelles (CGB05, CGB07 et CGC04), contrairement, présentent plus de données manquantes à l'issue de la pré-amplification et un plus grand nombre de différences observées entre les génotypage avec ou sans WGA (entre 3,65 et 5,47% de différences observées). Ceci pourrait être expliqué par une concentration ou une qualité plus faible de l'ADN pour ces trois dernières femelles, comme nous avons pu voir précédemment. Cependant ces tests étant réalisés sur un faible nombre d'individus, ils demanderaient à être vérifiés.

Tableau 2.9 : Récapitulatif des résultats obtenus pour le génotypage de six femelles avec ou sans WGA ; DM=données manquantes.

individus	nb de loci identiques entre avec et sans WGA	nb de loci différents entre avec et sans WGA	nb de DM sans WGA	nb de DM avec WGA	nb de DM observés dans les deux cas	%différence	%similitude
C212	279	4	92	83	74	1,04	98,96
C030	275	7	90	85	73	1,82	98,18
H4I	284	5	83	80	68	1,30	98,70
CGB05	247	18	85	98	64	4,69	95,31
CGB07	252	14	89	100	71	3,65	96,35
CGC04	240	21	86	104	67	5,47	94,53

2. Problèmes techniques rencontrés lors du génotypage et de l'analyse des résultats

Lors de l'analyse des puces issues du génotypage, nous avons rencontré diverses difficultés d'analyse, généralement liées à des problèmes techniques. Nous nous sommes ainsi confrontés à des problèmes d'hétérogénéité entre puces, des problèmes techniques lors du déroulement du génotypage et de l'intégration des résultats par le logiciel de lecture Fluidigm. Ces difficultés, ainsi que les solutions proposées pour résoudre ces problèmes, sont abordées par la suite.

a) Hétérogénéité entre puces du génotypage par le Biomark HD de Fluidigm

Pour le génotypage du premier run, 20 puces ont été utilisées permettant de croiser cinq plaques d'individus (464 individus + 16 témoins) et quatre plaques de SNPs (384SNPs). Il est possible de voir si le même résultat est obtenu en fonction des puces d'individus pour une même puce de SNP et donc si un effet puce est visible.

Dans l'exemple illustré dans la figure 2.27, chaque valeur sur l'axe des abscisses représente un locus SNP donné. Ici, seuls 31 SNPs d'une même puce SNP sont représentés pour une meilleure lisibilité. Pour chaque SNP, six points sont représentés correspondant au pourcentage de données manquantes observé sur les six plaques d'individus typés pour ces SNPs.

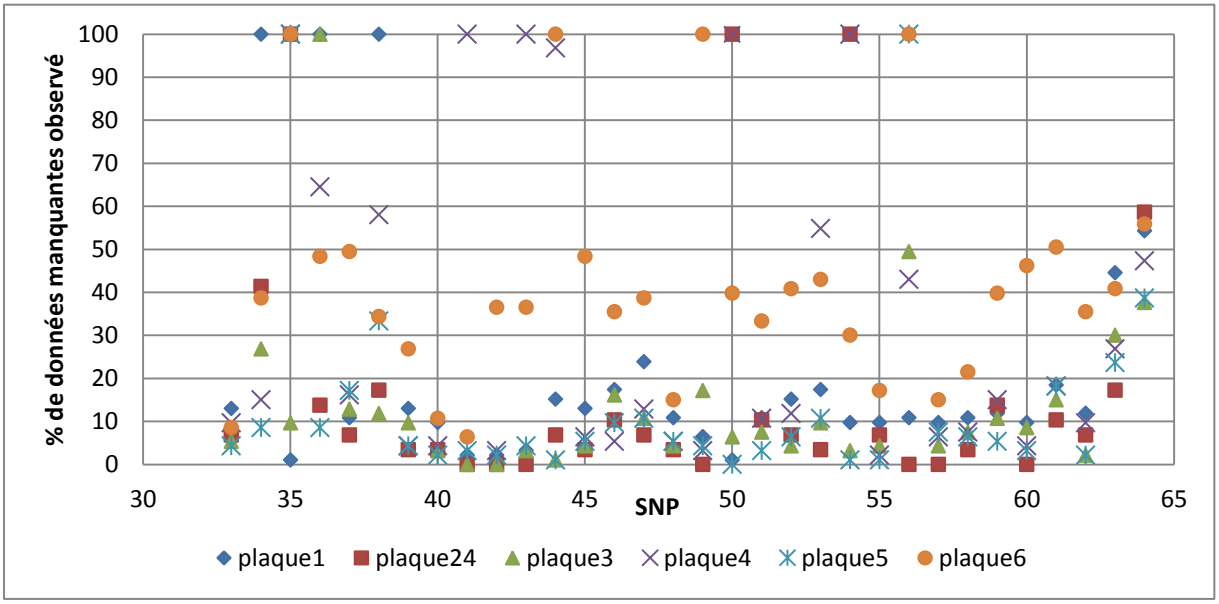


Figure 2.27 : Représentation du pourcentage de données manquantes observées entre les 5 différentes plaques d'extraction (et génotypage) d'ADN (plaque 1, 3, 4, 5, 6 du premier run de génotypage, plaque 24 du deuxième run) pour 31 SNPs d'une même plaque.

La figure 2.27 montre que la plaque 6 comporte beaucoup plus de données manquantes que les autres (39.16% contre 18.64% en moyenne pour les autres plaques). Ceci pourrait être dû à une hétérogénéité de la qualité d'extraction d'ADN, les extractions ayant été réalisées en plaque. Ainsi les extractions de la plaque 6 seraient de moins bonne qualité.

b) Problème technique de lecture et d'assignation des points de génotypage

Lors de l'analyse de la puce 8, chaque graphique de fluorescence des SNPs présentaient un profil similaire avec la grande majorité des points s'alignant sur la diagonale (où l'on retrouve normalement les hétérozygotes), comme représenté sur la figure 2.28a. L'image synthétique de la puce après lecture (Figure 2.28b) montrait une distribution des points de génotypage très atypique, sans points correspondant aux 2 types d'homozygotes (XX en vert et YY en rouge). Il y a donc eu un problème technique lors de la lecture de la puce.

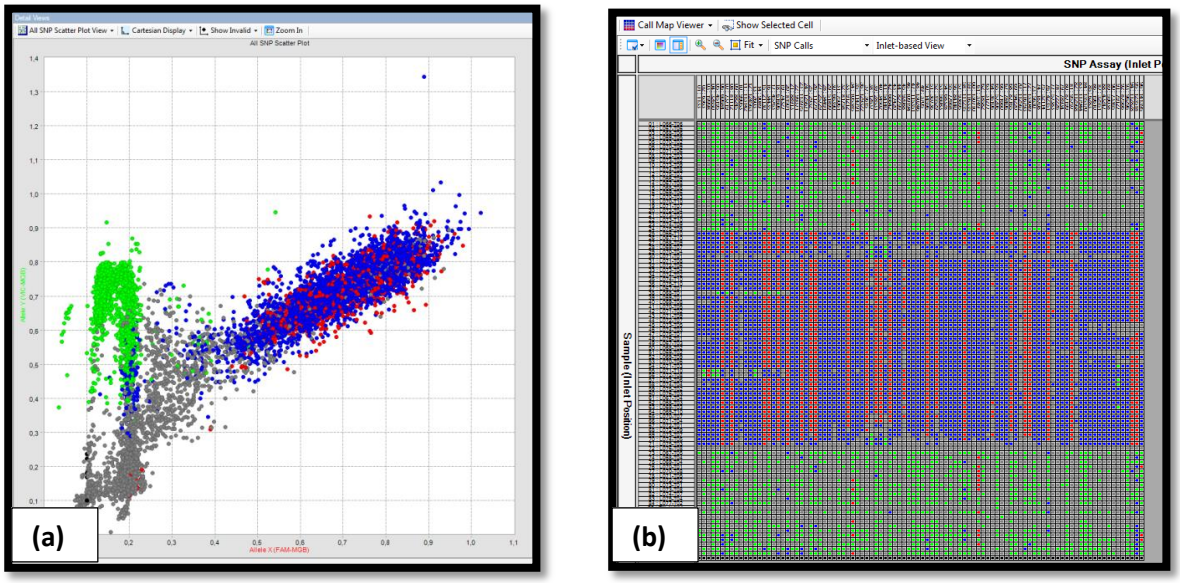


Figure 2.28 : (a) Fluorogramme issu du génotypage de la puce 8 ; (b) Image synthétique de la lecture de la puce 8 par le logiciel d'analyse Fluidigm. On observe dans les deux cas, une distribution anormale des points de génotypage.

c) Série de points de génotypage anormaux sur une puce

i. Premier cas

Un premier cas rencontré est représenté figure 2.29. Comme on peut le voir sur l'image de la puce 11 (Figure 2.29), les données manquantes (No Call) ne sont pas réparties de manière aléatoire au sein de la puce et se concentrent en série (stretch) par ligne, et donc par individus.

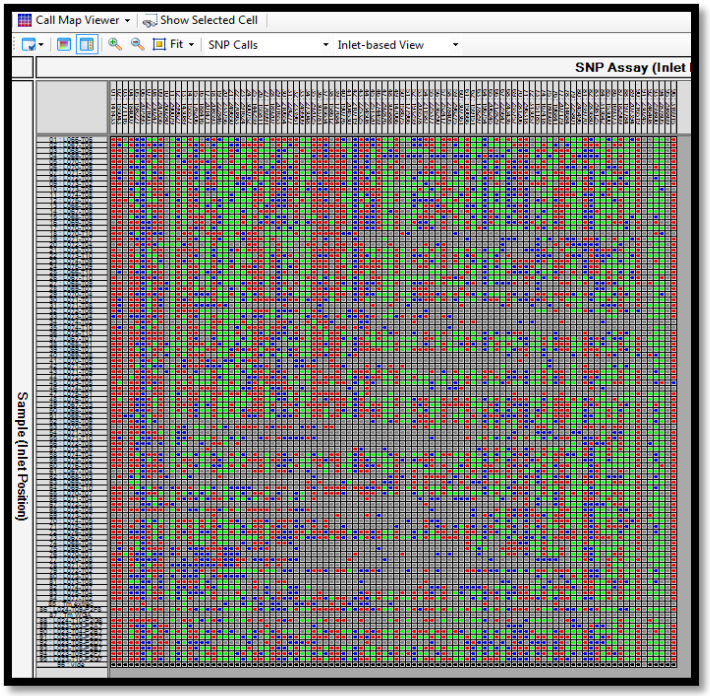


Figure 2.29 : Image synthétique de la puce 11 à la sortie du logiciel d'analyse.

Suite à l'analyse de la puce 11 manuellement, afin de vérifier les inférences données par le logiciel, ces stretchs de No Call (gris) ont été considérés comme des hétérozygotes (bleu) (Figure 2.30). Or il ne parait pas vraisemblable qu'un individu soit réellement hétérozygote pour un si grand nombre de marqueurs consécutifs sur une même puce.

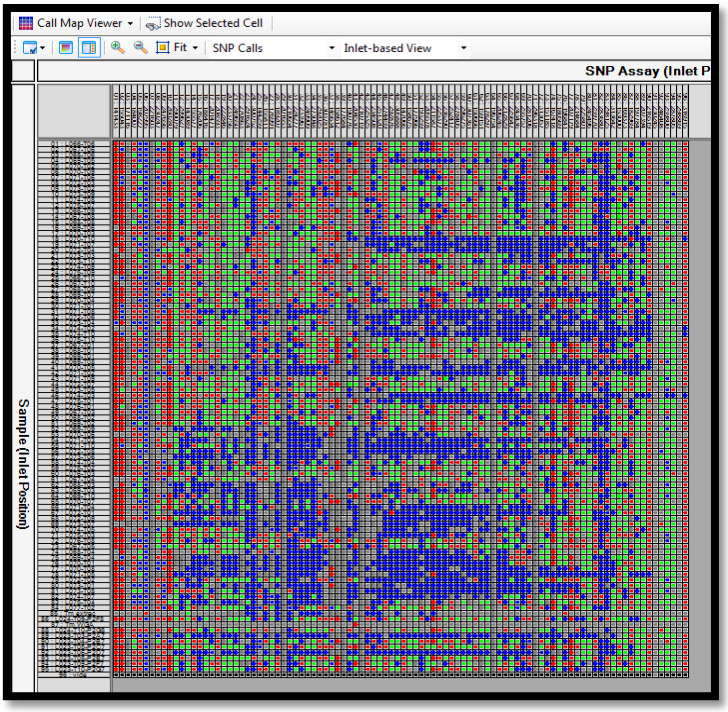


Figure 2.30 : Image de la puce 11 suite à mon interprétation manuelle.

Pourtant, comme nous pouvons le voir sur la figure suivante (Figure 2.31), lors de l'analyse de chaque SNP sur les graphiques de fluorescence, l'assignation réalisée par le logiciel à partir de l'analyse de la fluorescence semble correcte et le génotypage semble s'être déroulé de façon satisfaisante.

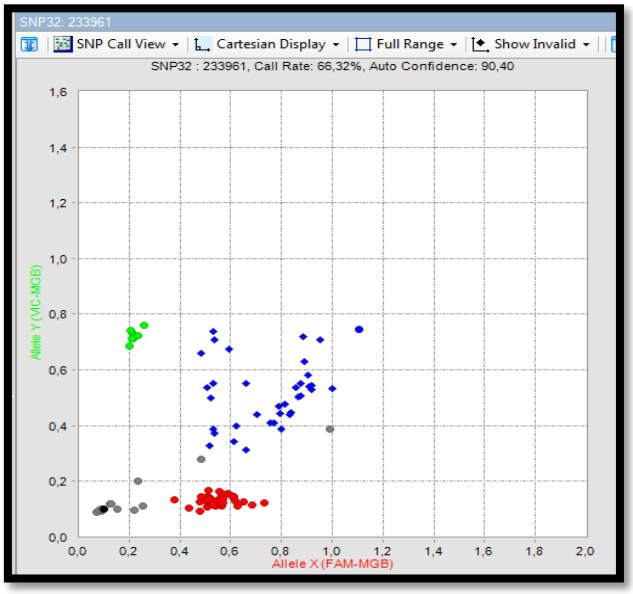


Figure 2.31 : Exemple du typage d'un SNP sur la puce 9 après analyse et assignation manuelle des points de génotypage.

Afin d'expliquer ce problème de série de points de génotypage anormaux, rencontré sur 4 puces (4 sur 24), deux hypothèses sont possibles :

- la première serait due aux conditions d'extraction de l'ADN. L'ensemble des nymphes a été extrait avec le kit d'extraction Macherey-Nagel NucleoMag. Ce kit, qui permet d'extraire en plaque 96 puits de manière rapide, utilise des microbilles magnétiques afin de fixer l'ADN à leur surface (recouverte d'un produit à forte affinité pour l'ADN). Au moment de l'élution, l'ADN se détache des billes magnétiques pour se retrouver élué dans le tampon. Afin de prélever uniquement l'ADN, les microbilles sont retenues sur un support aimanté. Cependant lors de l'aspiration de l'ADN, des microbilles peuvent se retrouver dans la solution avec l'ADN. Etant donné que pour le génotypage, les puces utilisées sont basées sur un système de plaque en système microfluidique, le diamètre des canaux est extrêmement faible. De ce fait, ces billes magnétiques pourraient venir obstruer les canaux et ainsi empêcher la bonne circulation de l'ADN dans les circuits de la plaque jusqu'aux puits où a lieu l'amplification et donc limiter l'amplification et la quantification de la fluorescence émise.

- une deuxième possibilité (très proche de la première explication développée ci-dessus) serait due à l'étape de purification pré-génotypage. Cette purification est basée sur le kit Ampure, qui repose également sur une purification via un système de billes magnétiques. Ainsi, comme pour l'hypothèse 1, ces billes auraient également pu être aspirées et se bloquer par la suite dans les canaux des puces Fluidigm.

ii. Deuxième cas

Le deuxième cas de série de points de génotypage anormaux observés est le suivant :
L'analyse de la puce 9 a posé quelques problèmes au moment de la lecture de la plaque totale. Comme pour le cas de la puce 11 (prise en exemple précédemment) à la lecture de l'ensemble des point de génotypage sur la 'Map viewer', on remarque un « quadrillage » régulier sur la plaque où des individus ressortent à 100% hétérozygotes pour l'ensemble des marqueurs de la puce (cf lignes bleues horizontales) mais également pour quelques marqueurs qui sont tous hétérozygotes pour l'ensemble des individus génotypés (Figure 2.32).

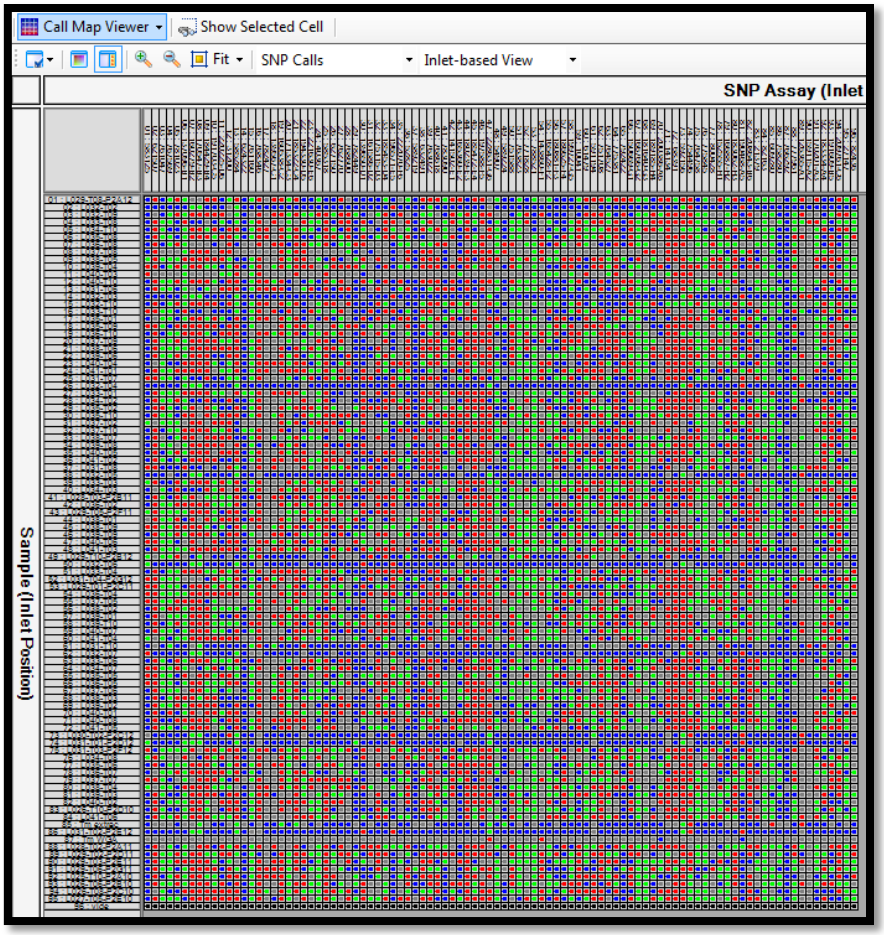


Figure 2.32 : Image de la puce 9 à la sortie du logiciel d'analyse, qui correspond également à l'analyse effectuée manuellement.

Cependant, la lecture des graphiques pris individuellement ne posait pas de difficultés majeures dans l'interprétation des points de génotypage (les différents génotypes étaient facilement dissociables et il ne semblait pas apparaitre d'excès d'hétérozygotes comme on aurait pu s'attendre à la vue de l'ensemble de la puce). Suite à mon assignation manuelle, l'image finale de la puce correspondait strictement à celle issue de l'analyse faite par le logiciel avec le même quadrillage.

Cette régularité observée toutes les 12 lignes correspond aux individus qui étaient disposés sur le bord de la plaque d'ADN. Cet emplacement entrainerait probablement un mauvais pipetage qui est automatisé par robots.

d) Résolution des problèmes techniques rencontrés

Le logiciel d'analyse **Fluidigm SNP GenotypingAnalysis** ne permet pas une grande flexibilité ; il n'autorise notamment pas d'effectuer des corrections « manuelles » des assignations de génotypes directement dans le logiciel. De ce fait, la résolution des problèmes rencontrés s'est faite *a posteriori* directement dans les fichiers de sorties de chaque puce.

Pour le problème de la puce 8 : L'ensemble de la puce étant ininterprétable, toutes les données générées sur cette puce ont été considérés comme des données manquantes. La part de données manquantes portée par cette puce correspond à 4.7% de l'ensemble des données finales du jeu de données OSCAR (9 024 points de génotypage/ 189 312).

Pour le problème de série de points de génotypage anormaux: Nous avons constaté que ces problèmes se caractérisaient pour une répétition de 'XY'/'No Call' de manière anormale. De ce fait nous avons cherché à savoir à partir de combien de répétitions du motif 'XY/No Call' nous pouvions considérer cette suite comme dû à un problème.
Pour ceci nous avons analysé deux puces, la puce 9, très affectée par le problème des séries anormales, et la puce 2, peu affectée, afin de définir la longueur de chaine à considérer comme anormale (Figure 2.33)

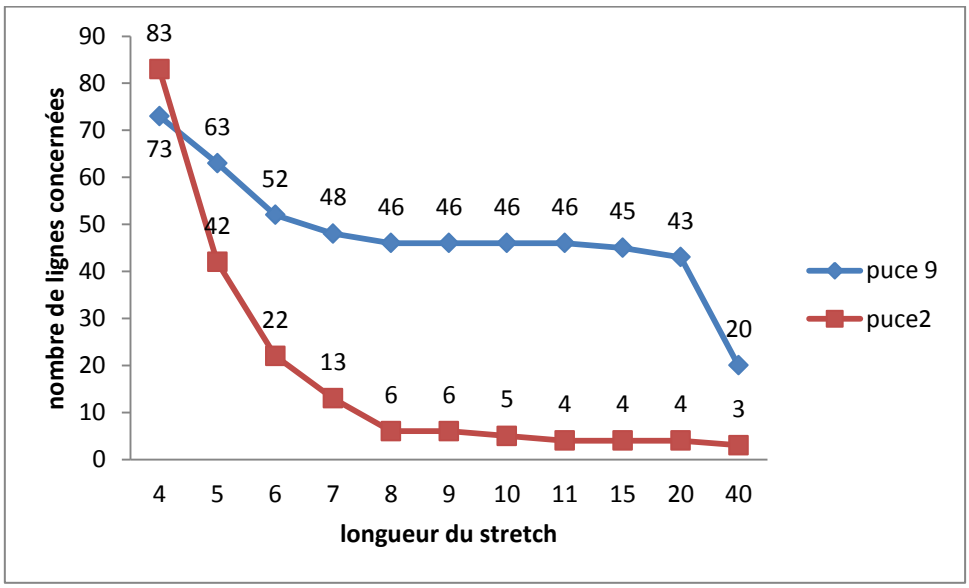


Figure 2.33 : Graphique représentant le nombre de lignes (=individus) par puce présentant un stretch composé d'au minimum une répétition entre 4 et 40 fois du motif 'XY/No Call' sur les 96 points de génotypage de chaque individus).

Suite à cette analyse, nous avons établi comme 'dû à un problème de série anormale' une répétition de minimum sept répétitions consécutives du motif 'XY/No Call' sur une même ligne, cette valeur de sept correspondant à la valeur proche du plateau pour les deux puces analysées (Figure 2.33).
Par la suite, nous avons réalisé un script implémenté en Perl (voir Annexe 3) afin d'automatiser la tâche de correction sur l'ensemble des puces. Ainsi, chaque valeur d'un stretch reconnu comme anormal, a été recodé comme étant une donnée manquante. Par cette correction, 4.6% des données ont été corrigées et donc considérées comme données manquantes.

III. Validation et sélection des marqueurs

Les 384 marqueurs développés dans cette partie permettent d'enrichir les ressources génétiques concernant *Ixodes ricinus*, et permettront par la suite, diverses analyses de sa variabilité génétique. Dans le cadre de ma thèse, ces marqueurs seront utilisés afin de décrire la structuration de la tique *I. ricinus* à l'échelle du paysage, avec l'ensemble des individus issus du projet OSCAR génotypés pour les 384 marqueurs (cf chapitre 3). Nous avons sélectionné une partie des marqueurs, ceux présentant toutes les caractéristiques requises pour apporter des informations fiables et utiles aux études de génétique des populations (codominants pour pouvoir discriminer les homo- des hétérozygotes, ne pas présenter d'allèles nuls et présenter un transmission mendélienne à hérédité simple (Vienne 1998)). Pour sélectionner les meilleurs marqueurs, trois critères de sélection ont été utilisés : (1) la ségrégation des allèles étudiées à partir de l'analyse de croisements afin d'écarter les loci montrant des allèles nuls ou un caractère non-mendélien ; (2) le pourcentage de données manquantes afin de ne pas conserver les loci ou les individus pour lesquels ils manquaient trop d'informations; (3) la fréquence allélique minimale (MAF) afin de ne conserver que les loci dont l'allèle le plus rare est présent à plus de 5% dans le jeu de données .

Cependant avant de réaliser un tri sélectif sur les SNPs, nous avons effectué un filtre au niveau des données manquantes liées aux individus. Le jeu de données final pour les tiques OSCAR de ZAA comporte un grand nombre de données manquantes (27,50% des données), dû notamment aux problèmes techniques rencontrés (puce 8 ou présence de séries anormales, toutes été considéré comme données manquantes) mais également à un nombre important de données manquantes réparties dans l'ensemble du jeu de données du fait des faibles quantités d'ADN disponibles pour certains individus.
De ce fait, le premier filtre appliqué a été de supprimer les individus présentant trop de données manquantes. Avant toute sélection, la fréquence minimum de données manquantes observées était

de 11,20% (soit 43 données manquantes sur les 384 SNPs) pour les individus et la fréquence maximum de 69,01% (soit 265 données manquantes sur les 384 SNPs) (Figure 2.34).

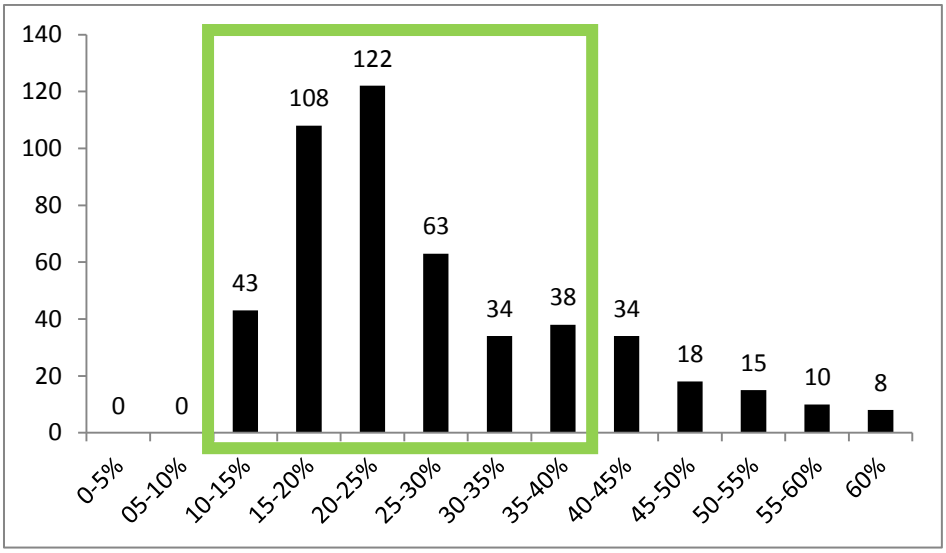


Figure 2.34 : histogramme représentant le nombre de données manquantes par individu et par classe de pourcentage de données manquantes. Le cadre vert représente l'ensemble des individus conservés pour la suite (individus portant moins de 40% de données manquantes).

Comme nous pouvons le voir sur la figure 2.34, un certain nombre d'individus comportent un nombre de données manquantes conséquent (33 individus portent plus de 50% de données manquantes dans leurs jeux de données). L'ensemble des individus portant plus de 40% de données manquantes a été éliminé de l'analyse. Ce seuil a été choisi afin de ne pas éliminer des individus qui présentent un fort taux de données manquantes induit par un nombre important de SNPs ayant également un grand nombre de données manquantes. A ce stade notre jeu de données est représenté par 408 individus et 384 marqueurs. Les individus seront retriés par la suite en fonction des données manquantes de chacun, suite à la sélection des SNPs considérés comme bon candidats.

A. La ségrégation des allèles étudiée à partir de l'analyse de croisements

Pour vérifier l'absence d'allèle nul et le caractère mendélien de la ségrégation des allèles, la transmission des allèles des marqueurs SNPs entre parents et descendants au sein des cinq croisements (croisements C212, C214, C243, C30 et H4I) a été analysée. Ces croisements ont été effectués en 2009 et 2012 au laboratoire, en conditions contrôlées : chacune des femelles (dont nous étions sûr de la virginité car elles étaient issues de nymphes gorgées ayant évolué isolément dans un

tube Eppendorf) a été fécondée par un seul mâle (accouplement réalisé dans un tube Eppendorf). Lors de la réalisation des croisements, après avoir vérifié que la fécondation ait eu lieu (présence de spermatophore), les mâles ont été isolés et conservés (à -80°ou -20°C selon les croisements) jusqu'à l'extraction d'ADN. Les femelles ont été gorgées (sur veau ou lapin selon les différents croisements) puis isolées jusqu'à la ponte des œufs. Suite à la ponte, le corps des femelles a été conservé à -80°C ou -20°C jusqu'à l'extraction d'ADN et les œufs ont évolué jusqu'au stade larvaire. Pour l'ensemble des croisements (mis à part le croisement H4I) les larves ont également réalisé un gorgement (sur gerbille), conduisant à obtenir une descendance analysée au stade nymphe (la descendance du croisement H4I a été analysée au stade larvaire). Une dizaine d'individus de chacun des cinq croisements a été sélectionnée et génotypée. Etant donné la quantité insuffisante d'ADN extrait, une amplification WGA a été faite au préalable afin de pouvoir réaliser l'ensemble des génotypages.

Lors du deuxième run de génotypage, l'ensemble des parents et des descendants sélectionnés a été génotypé pour les 384 SNPs. Ainsi, pour chacun des croisements, la ségrégation des 384 marqueurs a été analysée.

Les génotypes ont été codés de la façon suivante pour chacun des SNP:

- XX pour les homozygotes d'un allèle
- YY pour les homozygotes de l'autre allèle
- XY pour les homozygotes

Comme le montre l'exemple suivant (Figure 2.35) pris pour 18 SNPs analysés pour le croisement C212, les génotypes des parents correspondaient à des homozygotes pour chacun des SNPs présents (XX pour la mère et YY pour le père). De ce fait, suivant les lois de Mendel, l'ensemble de la descendance devait être hétérozygote, en ayant hérité de chacun d'un des deux allèles des deux parents. Parmi les 18 SNPs montré en exemple, seulement huit montrent une hérédité mendélienne à la génération suivante. Pour les autres SNPs, on remarque des génotypes avec un seul allèle observé pour au moins un des descendants. Ces résultats peuvent s'expliquer par l'existence d'allèles nuls chez les parents. Ainsi, pour le locus 56083, le père était probablement Y0, donnant à la descendance 50% de XY (ici 4 sur 10) et 50% de X0 (ici 5 sur 10).

	C212D01	C212D02	C212D03	C212D04	C212D05	C212D06	C212D07	C212D13	C212D15	C212D16	parents concaténés	nb NC/descendance	C212MER (WGA)	C212PER (WGA)	nb de XY	nb de non mendélien
116335	XY	XY	XY	XY	XY	XY	XY	XY	XY	XY	XXYY	0	XX	YY	10	0
145634	XY	XY	XY	XY	XY	XY	XY	XY	XY	XY	XXYY	0	XX	YY	10	0
210654	XY	XY	XY	XY	XY	XY	XY	XY	XY	XY	XXYY	0	XX	YY	10	0
230247	XY	XY	XY	XY	XY	XY	XY	XY	XY	XY	XXYY	0	XX	YY	10	0
243436	XY	XY	XY	XY	XY	XY	XY	XY	XY	XY	XXYY	0	XX	YY	10	0
356395	XY	XY	XY	XY	XY	XY	XY	XY	XY	XY	XXYY	0	XX	YY	10	0
379138	XY	XY	XY	XY	XY	XY	XY	XY	XY	XY	XXYY	0	XX	YY	10	0
571455	XY	XY	XY	XY	XY	XY	XY	XY	XY	XY	XXYY	0	XX	YY	10	0
9089	YY	XY	YY	YY	XY	YY	YY	XX	YY	XY	XXYY	0	XX	YY	3	7
56083	XX	XY	XY	XY	XX	XX	XY	XX	NC	XX	XXYY	1	XX	YY	4	5
113975	XY	XY	XY	NC	XY	XY	XY	XY	XX	XY	XXYY	1	XX	YY	8	1
155043	NC	XY	XX	XY	NC	NC	XY	XY	XY	NC	XXYY	4	XX	YY	5	1
278405	XY	YY	YY	YY	XY	YY	XY	XY	YY	YY	XXYY	0	XX	YY	4	6
300752	XY	XY	XY	XY	XY	XY	XY	XY	XY	YY	XXYY	0	XX	YY	9	1
324697	YY	XY	XX	XX	YY	XX	YY	XY	YY	XY	XXYY	0	XX	YY	3	7
392019	YY	XX	XX	YY	XY	XY	YY	XY	XY	XY	XXYY	0	XX	YY	5	5
139567-C1	XX	YY	YY	YY	XX	YY	XY	XY	YY	YY	XXYY	0	XX	YY	2	8
183334-A4	XX	XX	NC	XY	NC	XY	XX	NC	NC	NC	XXYY	5	XX	YY	2	3

Figure 2.35 : Exemple d'analyse de la ségrégation des allèles à la génération suivante. L'identifiant de chaque locus SNPs est indiqué dans la première colonne. Les colonnes suivantes correspondent aux génotypes des dix descendants analysés. Dans la colonne suivante figure le génotype concaténé des deux parents puis le nombre de données manquantes au sein de la descendance, puis des génotypes des deux parents. Pour finir est répertorié le nombre de descendants présentant une ségrégation des SNPs mendélienne et non-mendélienne (sans faire l'hypothèse d'un allèle nul chez un ou les deux parents). Dans cet exemple les parents sont tous deux homozygotes pour les deux allèles du SNP (XX ou YY), de ce fait la descendance doit être hétérozygote pour l'ensemble des marqueurs présentés dans cet exemple.

L'ensemble des résultats pour l'ensemble des 47 descendants analysés est présenté dans le graphique suivant (Figure 2.36). Ainsi nous pouvons voir que pour 151 des 384 marqueurs, aucun allèle n'apparait être transmis de manière non-mendélienne. Parmi ces 151 marqueurs, 52 correspondent à des croisements avec des données manquantes chez les parents qui ne sont donc pas informatifs pour mettre en évidence des biais de ségrégations. Cette liste de loci inclus également des SNPs pour lesquels les génotypes des deux parents étaient hétérozygotes. De ce fait, ce type de croisement est peu discriminant pour révéler la présence de loci non-mendéliens car les trois types de génotypes (XX, XY et YY) sont attendus parmi les descendants. Par ailleurs, étant donné les petits effectifs de descendants étudiés (maximum 10), nous n'avons pas utilisé l'information

quantitative liée à la fréquence des différents génotypes qui aurait pu être comparée aux fréquences attendus suivant une ségrégation Mendélienne.

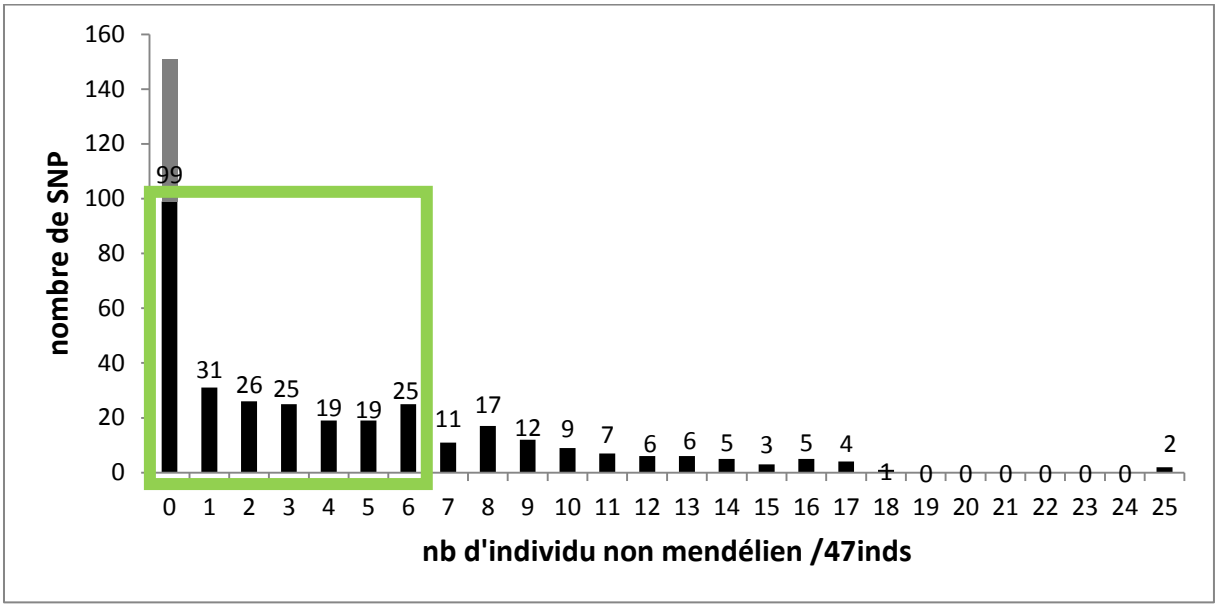


Figure 2.36 : Histogramme représentant l'ensemble des SNPs en fonction du nombre d'individus parmi la descendance des 5 croisements réalisés (N=47). Le cadre vert représente l'ensemble des individus conservés pour la suite (paragraphe suivant). En gris (partie supérieur de la barre d'histogramme correspondant à 0 individu non mendélien) sont représentés les loci inclus des données manquantes chez les parents.

Suite à la vérification du caractère mendélien ou non des SNPs développés, nous avons choisi de conserver les SNPs dont l'analyse présentaient moins de 6 individus douteux dans la descendance, du fait du caractère non mendélien de ces SNPs (soit <15%). Ainsi 244 SNPs ont été conservés (Figure 2.36).

Nous avons toléré un seuil jusqu'à 15% d'individus non mendéliens dans la descendance étant donné que des erreurs d'assignation des génotypes (interprétation des graphiques) ou des erreurs liés à l'amplification WGA des larves et/ou nymphes ont pu se produire. Par ailleurs, l'hypothèse de la présence d'allèle nul chez au moins un des parents permet de respecter une ségrégation mendélienne dans l'immense majorité des cas.

B. Le pourcentage de données manquantes

A ce stade, notre jeu de données est constitué de 244 SNPs et de 408 individus. Cependant les SNPs retenus présentent toujours une grande variation en terme de données manquantes. De ce fait il a été choisi de ne conserver que les SNPs portant moins de 25% de données manquantes (sur les 408 individus). La distribution des données manquantes est très hétérogène, allant de 0,98% à 99,26% de données manquantes par SNP (Figure 2.37).

Cette sélection par le filtre des données manquantes a permis de conserver 182 SNPs présentant entre 0 et 25% de données manquantes.

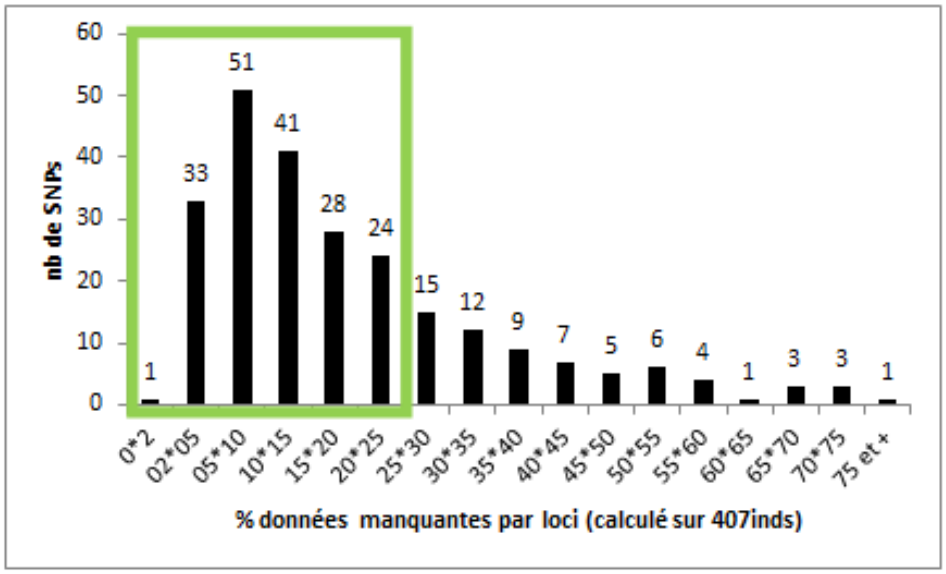


Figure 2.37 : Histogramme présentant la répartition des SNPs en fonction de leur pourcentage de données manquantes sur l'ensemble des 408 individus conservés. Le cadre vert représente l'ensemble des SNPs conservés pour la suite (SNPs portant moins de 25% de données manquantes).

Le jeu de données est maintenant constitué de 182 SNPs. Après les différents filtres utilisés, ces 182 SNPs sélectionnés présentent tous moins de 25% de données manquantes et sont associés au minimum à 6 individus présentant un SNPs ayant potentiellement ségrégué de manière non mendélienne.

C. la fréquence allélique minimale (MAF)

Nous avons également considéré les fréquences alléliques minimales. Ces fréquences servent de filtre pour le contrôle de la qualité car pour un algorithme d'inférence tel que celui utilisé avec le système Fluidigm, il est très difficile de génotyper un SNP dont un des allèles est présent à faible fréquence. En effet, dans l'espace de représentation des intensités lors de l'inférence des génotypes, les fréquences faibles produisent de petits ilots d'un ou deux individus pour les homozygotes portant l'allèle mineur. De ce fait ils risquent d'être assimilés à tort au groupe le plus proche, à savoir les hétérozygotes. Même si chaque point de génotypage a été de vérifié à l'œil, les groupes d'intensités étant parfois difficiles à interpréter, ces erreurs ont également pu être effectuées.

Plusieurs seuils peuvent être utilisés selon les effectifs de la population à l'étude, mais les valeurs sont en général fixées entre 1% et 5%. Nous avons choisi de conservé uniquement les SNPs présentant plus de 5% (0,05) de l'allèle minimum (Figure 2.38).

Cette sélection nous a conduits à restreindre le jeu de données à 143 SNPs, répondant à l'ensemble des critères définis jusqu'à présent.

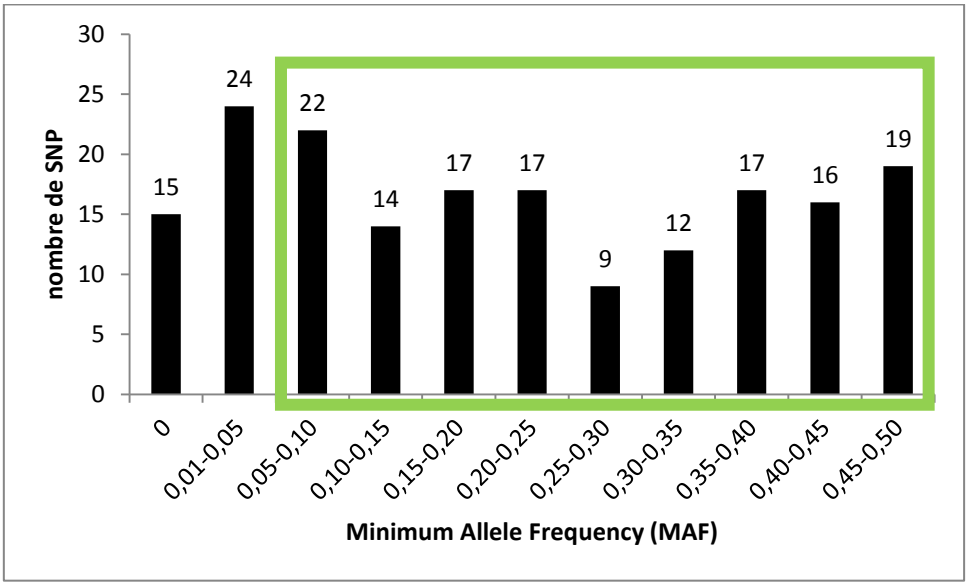


Figure 2.38 : Histogramme présentant la répartition des SNPs en fonction des fréquences alléliques minimales sur l'ensemble des 408 individus conservés. Le cadre vert représente l'ensemble des SNPs conservés pour la suite (SNPs présentant plus de 5% de l'allèle minimum).

D. retour sur la ségrégation des allèles étudiée à partir de l'analyse de croisements

Même si le caractère mendélien ou non des loci a déjà été utilisé à ce stade de la sélection des loci, celui-ci a été utilisé à nouveau de façon encore plus stringente afin d'éliminer des loci potentiellement porteurs d'allèles nuls. Alors que les loci présentant des données manquantes chez les parents avaient été exclus, ils restaient des loci présentant un nombre important de données manquantes parmi les descendants. Ces loci se révèlent eux aussi peu discriminants pour exclure ceux qui pourraient présenter des allèles nuls ou une ségrégation non-mendélienne. Dans cette ultime phase de sélection, je n'ai retenu, parmi les 143 SNPs sélectionnés jusqu'ici, que les SNPs ayant donné un génotype pour plus de 50% des 47 individus (Figure 2.39). Ainsi, à ce stade, 10 SNPs supplémentaires ont été retirés.

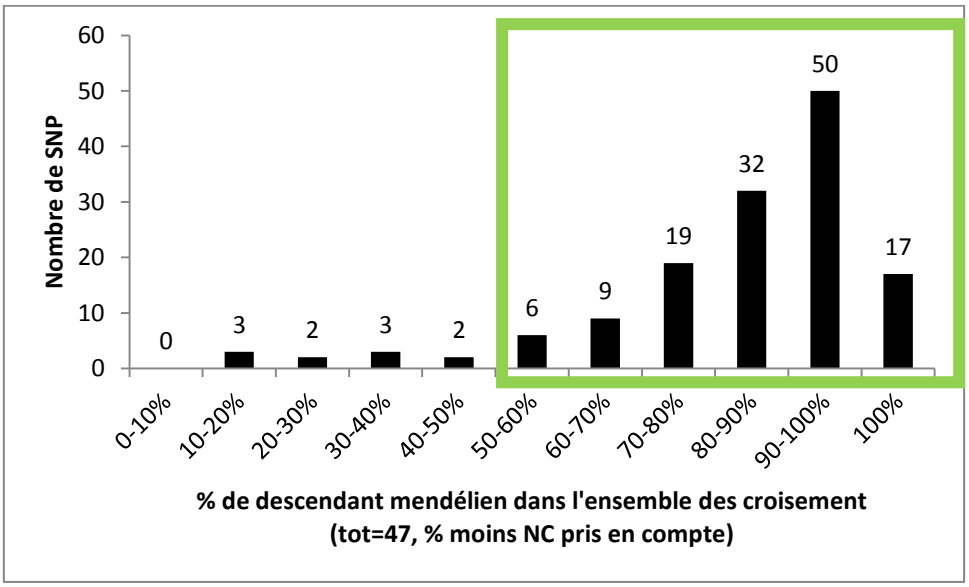


Figure 2.39 : Histogramme présentant la répartition des 143 SNPs en fonction des pourcentages de descendants mendélien dans l'ensemble des 47 descendants. Le cadre vert représente l'ensemble des SNPs conservés pour la suite (SNPs présentant plus de 50% de descendants mendéliens).

Le nombre de SNPs répondant à l'ensemble des critères de sélection est de 133. Finalement, parmi ces 133 SNPs restant, cinq ont également été retirés car lors des tests sur les duplicats d'individus génotypés (voir chapitre 2 partie D.1.b), ils présentaient plus de 15 % de différences dans le génotype inféré. Ces différences de génotypage entre duplicat pouvant être dû à un problème d'allèle nul ou encore de séquences répétées, ces SNPs ont été exclus des analyses ultérieures.

Le set définitif de marqueurs qui sera de ce fait utilisables pour la suite des analyses de génétique des populations s'élève donc à 128 SNPs.

E. Validation des marqueurs à une échelle intercontinentale par une analyse de génétique des populations

Afin d'estimer le niveau d'informations génétiques apporté par les 128 SNPs précédemment validés pour l'analyse de la ségrégation d'allèles, les génotypes de différentes familles (père, mère et descendants – larves/nymphes -) issues des croisements réalisés en conditions contrôlées ont été analysés (Tableau 2.10).

1. Matériels et méthodes

Sept populations ont été utilisées :

- 5 populations issues de croisements, entre parents de différentes origines géographiques.
- une population de trois femelles non gorgées issues de nymphes gorgées sur chevreuil récoltées à Chizé et ramenées au laboratoire où elles ont mué en adulte.
- une population de 10 nymphes récoltées dans la Zone Armorique Atelier (Pleine Fougère) et sélectionnées de manière aléatoire parmi l'ensemble des 493 tiques génotypées.

Tableau 2.10 : Récapitulatif des origines géographiques des différentes tiques analysées.

Population	Origine géographique	Stade des individus étudiés	Nombre d'individus
C212	**Tunisie x Belle-Ile**	**Mâle, femelle, nymphes**	**12**
C214	**Tunisie x Belle-Ile**	**Mâle, femelle, nymphes**	**11**
C30	**Tunisie x Toulouse**	**Mâle, femelle, nymphes**	**12**
C243	**Chizé x Chizé**	**Mâle, femelle, nymphes et larves**	**11**
H4I	**Chizé x Chizé**	**Mâle, femelle, larves**	**11**
CG	**Chizé (terrain)**	**Femelle adulte**	**3**
ZAA	**Pleine-Fougère (terrain)**	**Nymphe**	**10**

Parmi les 128 SNPs sélectionnés précédemment, six n'ont pas donné de résultats pour l'ensemble des croisements, réduisant le nombre de marqueurs à 122 pour cette analyse.

La différenciation génétique entre paires de populations a été estimée par l'indice *Fst* de Wright (Wright 1965), avec son estimateur non-biaisé θ de Weir and Cockerham (1984). Cet indice varie entre zéro (absence totale de différenciation) et un (traduisant une absence totale de flux de gènes entre les deux populations). La significativité du test a été établie au seuil de 5% après 1000 permutations. L'analyse a été effectuée avec le logiciel Genepop (Rousset 2008).

La différenciation génétique des populations a également été explorée par deux autres méthodes. D'une part, une méthode basée sur une analyse factorielle des correspondances (AFC) a été effectuée sur les fréquences alléliques des populations à l'aide du logiciel Genetix (Belkhir *et al.* 2004). Les objets analysés, ici des individus, sont projetés dans un hyper-espace qui a autant de dimensions qu'il y a de marqueurs génétiques, et leur position dans cet espace sera fonction des allèles qu'ils portent aux différents loci. Cette analyse permet de représenter graphiquement le niveau de proximité génétique entre individus. Un code couleur permet de visualiser l'appartenance de chaque individu à sa population d'origine.

Par ailleurs, la structure génétique a également été inférée par une méthode d'analyse bayésienne de classification, réalisée à l'aide du logiciel STRUCTURE (Pritchard *et al.* 2000). Ce logiciel fait l'hypothèse que les fréquences génotypiques sont à l'équilibre de Hardy-Weinberg et les loci à l'équilibre de liaison. La méthode d'inférence ré-échantillonne de façon probabiliste chaque individu à partir de son génotype multi locus, afin de l'assigner à une sous population considérée (dont le nombre K peut varier et est fixé par l'expérimentateur). Aucune information sur l'origine géographique des individus n'a été fournie au logiciel. La probabilité a posteriori des données a été estimée avec 100 itérations pour chaque valeur de K, sous un modèle avec introgression (« admixture model ») et avec corrélation des fréquences alléliques entre les populations. Une première phase « d'apprentissage » (Burn-in) de la simulation a été réalisée pour 1000 répétitions et la phase d'acquisition a été réalisée pour 10000 répétitions de la chaine de Markov. Les résultats sont résumés dans un diagramme où chaque individu est analysé et représenté par une barre d'histogramme verticale découpée en K segments colorés. La hauteur de ces segments est proportionnelle à la fraction des marqueurs du génome de l'individu assigné à chacun des K groupes génétiques inférés par l'analyse. Pour l'évaluation de la différenciation génétique, nous avons fait varier K entre 2 et 10 afin de trouver la valeur de K optimale.

2. Résultats

Les valeurs de *Fst* observées sont inférieures à 0,1 lorsque l'on compare les populations de tiques provenant d'une population naturelle (de ZAA –nord de la Bretagne - ou de CG –Chizé près de Niort dans les Deux-Sèvres -) à des tiques issues de croisements contrôlés impliquant des parents issus de la population de Chizé (Tableau 2.11). Une valeur extrême (θ = 0,00) indique aucune différenciation génétique entre les tiques de terrain collectées à Pleine-Fougère ou à Chizé, deux sites distants pourtant de 300 km (mais la population de Chizé ne comprenait que trois individus). De fortes différenciations génétiques sont observées entre les différents croisements (θ entre 0,170 et 0,305), ce qui pourrait s'expliquer par l'étroite base génétique de ces croisements (seulement deux parents – la femelle n'ayant pas pu être fécondée par plus d'un mâle- et neuf ou dix descendants) et la grande distance génétique des individus utilisés pour réaliser le croisement (population tunisienne versus française).

Tableau 2.11 : Matrice de distance représentant la différenciation génétique entre les différentes populations, selon le θ de Weir et Cockerham (1984).

	C212	**C214**	**C243**	**C30**	**CG**	**H4I**
C214	0,2139					
C243	0,2808	0,2788				
C30	0,3055	0,2683	0,2603			
CG	0,2378	0,2173	0,1358	0,2170		
H4I	0,2603	0,2496	0,1811	0,2720	0,0872	
ZAA	0,1879	0,1708	0,1159	0,1735	0,000	0,0695

La diversité génétique des populations a également été étudiée par une analyse factorielle des correspondances (AFC). Les axes 1,2 et 3 ont été choisis pour la représentation graphique car ils expliquaient la plus grande part de la variabilité génétique observée (presque 25% au total), avec respectivement 10,28 / 8,03 et 6,58 % expliquée par chacun des axes (Figure 2.40).

Trois groupes de populations sont observés. Les individus des croisements C212 (bleu) et C214 (jaune) sont regroupés entre eux et fortement différenciés des deux autres groupes. Un groupe intermédiaire est constitué uniquement de la population du croisement C30 (gris). Dans un groupe moins différencié, on retrouve les individus des populations de Chizé (population naturelle issu du terrain (CG) et des deux croisements impliquant aussi des tiques de Chizé : C243 et H4I) ainsi que les tiques issues du terrain de Pleine-Fougère (ZAA).

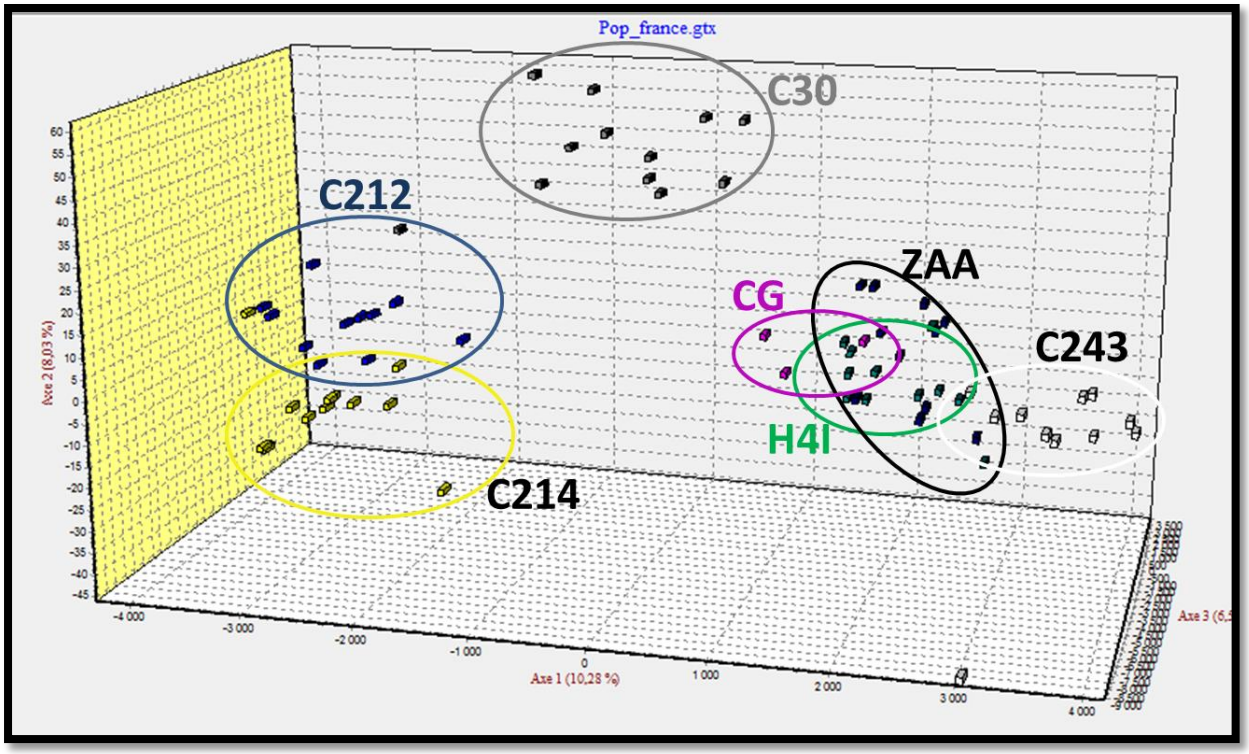


Figure 2.40 : Représentation des individus dans le plan factoriel tridimensionnelle des sept populations (les noms correspondent aux noms des populations indiqués dans le tableau 2.10).

Pour finir, la structure des populations a également été inférée par une méthode bayésienne à l'aide du logiciel STRUCTURE, à partir des données des génotypes en faisant varier le nombre de sous-population (K) entre 2 et 10 (Figure 2.41). A K=2, on remarque deux groupes distincts, un premier cluster regroupant les populations C212 et C214 (correspondant aux deux croisements Tunisie x Belle Ile) et un autre regroupant les cinq autres populations. Il est à noter que les tiques de Belle Ile correspondent à une lignée de tiques maintenue au laboratoire depuis plusieurs années, à l'origine probablement d'une moins grande diversité génétique au sein de cette lignée que pour les tiques des autres croisements. Lorsque K prend une valeur de 3, la population C30 (impliquant un croisement Tunisie x Toulouse) se détache des autres, confirmant les résultats de l'AFC où trois groupes très distincts sont représentés. A K=5, cinq sous-groupes sont observés : le groupement des populations C212 et C214 reste inchangé, l'isolement de C30 également, les deux populations issus de croisement de tiques de Chizé se différencient également chacune en une sous-population (C243 et H4I). Les deux populations 'sauvages' prélevées sur le terrain (il ne s'agit plus de lignées de tiques à la base génétique étroite) se retrouvent dans le même cluster (CG et ZAA). Ensuite, même si on augmente le nombre de sous-populations constituant notre analyse, l'ensemble des regroupements reste inchangé (exemple à K=7). Seules les deux populations 'sauvages' (ZAA et CG) subissent une subdivision en sous-populations intra-population.

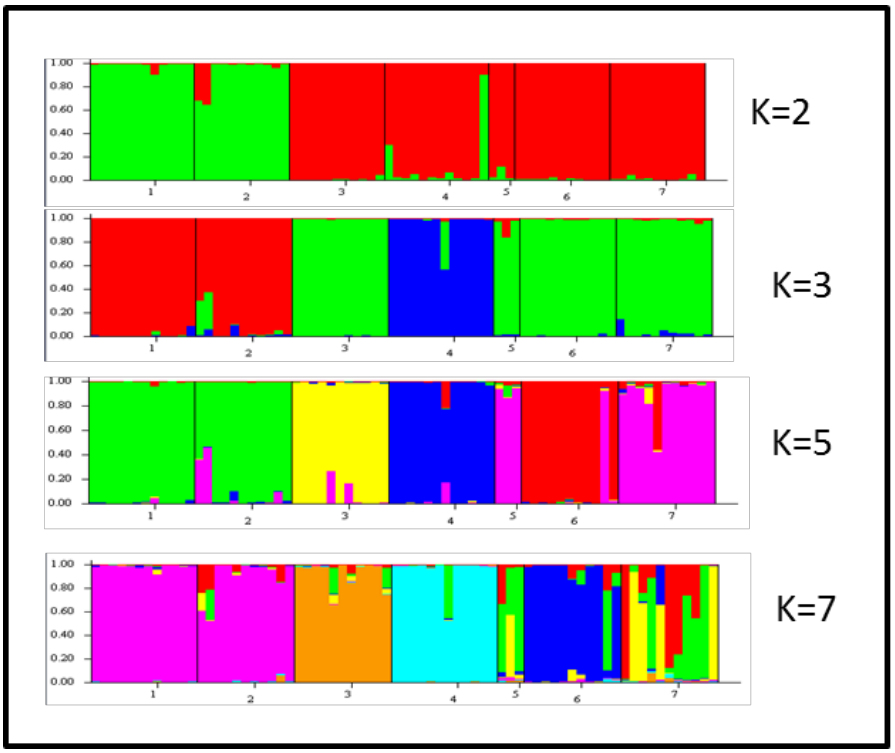


Figure 2.41 : Résultats du logiciel STRUCTURE pour un nombre de sous-populations inférées à K=2, 3,5 et 7. Les sous populations présentées de 1 à 7 correspondent respectivement aux populations C212, C214, C243, C30, H4I, CG et ZAA.

Même si ces analyses préliminaires sont réalisées sur un nombre limités d'individus de nature diverse (croisements, populations naturelles…), elles montrent que les marqueurs SNPs développés permettent de distinguer différentes populations d'*I. ricinus* en fonction de leur origine géographique et que l'information sur le polymorphisme génétique ainsi caractérisée est fiable.

IV. Discussion

Les résultats présentés dans ce chapitre constituent la première étude ayant permis l'identification et la validation de SNPs au sein d'un génome de tiques utilisant des technologies haut débit. Ces résultats, qui font l'objet d'une première publication dans le cadre de ma thèse (Quillery *et al.* 2013 – Annexe 8) permettent de fournir un set de marqueurs utilisables dans le futur pour des études analysant la variabilité génétique d'*I.ricinus*, mais également d'enrichir les données génomiques du genre *Ixodes* avec les 500Mb de données générées.

Dans le cas de la tique *Ixodes ricinus*, en raison des nombreuses difficultés rencontrées avec les marqueurs utilisés jusqu'à présent (cf les trois sets de marqueurs microsatellites : Delaye *et al.* 1998; Røed *et al.* 2006; Noel *et al.* 2012), il était urgent de produire un nouveau type de marqueur génétique hautement résolutif afin de pouvoir mieux comprendre la biologie mais également d'analyser la variabilité génétique de cette tique.

Le travail mené durant cette première partie de thèse a permis de mettre au point un set de 384 marqueurs SNPs, dont 368 ont été validés par génotypage, du polymorphisme ayant été observé pour chacun de ces loci. Finalement, une étape de sélection nous a permis d'isoler un set de 128 marqueurs pour lesquels nous pouvions accorder une légitime confiance dans les résultats de génotypage obtenus. Ces 128 marqueurs seront utilisés dans partielle chapitre 3 de ma thèse afin d'analyser la variabilité génétique et la structure génétique des populations d'*I. ricinus* à l'échelle du paysage.

Le développement de ces marqueurs a impliqué les cinq étapes suivantes :

- le pyroséquençage de deux banques d'ADN issues d'une réduction génomique de pools d'individus
- l'isolement de SNPs à partir du jeu de données NGS par le développement d'un pipeline original
- le design d'amorces
- le génotypage des SNPs
- la sélection et validation des SNPs.

Chacune de ces étapes constituera les différentes parties de cette discussion

A. Une stratégie de séquençage réussie

A notre connaissance, il n'existe pas de lignée consanguine homozygote d'*I. ricinus*. Ainsi, un polymorphisme intra-individuel est attendu, même si des individus uniques sont séquencés. Ce polymorphisme peut-être particulièrement élevé dans le cas d'*I. ricinus.* Différentes études basées sur un séquençage Sanger de gènes connus ont rapportées de fortes densités de SNPs dans le génome d'*Ixodes ricinus* (Noureddine *et al.* 2011) mais également dans le génome d'*I. scapularis* (Van Zee *et al.* 2013). Ainsi, théoriquement, le séquençage d'un seul ou deux individus permettraient d'identifier du polymorphisme chez *I. ricinus*. Cependant, en raison de ce polymorphisme non réductible, les erreurs de séquençage ainsi que les difficultés durant l'assemblage ne sont qu'accrues (*a fortiori* en l'absence de génome de référence, comme c'est notre cas). Par ailleurs, les quantités d'ADN disponibles pour un seul individu tique sont limitées. Avec en plus un génome de grande taille, les risques d'erreurs d'assemblage dus principalement à une faible couverture du génome sont élevés. De ce fait, plusieurs individus doivent être poolés pour permettre la réalisation d'un pyroséquençage. Dans le même temps en poolant plusieurs individus hétérozygotes, la probabilité d'observer du polymorphisme est maximisée. Cependant, pour estimer avec précision les fréquences alléliques, l'identification des individus au sein d'un pool est nécessaire (Cutler & Jensen 2010). Pour passer outre cette difficulté, il est possible de tagguer chaque individu, avec des étiquettes 'MID' (MID-tag) (séquence unique d'une vingtaine de paires de bases ajoutées aux fragments d'ADN avant l'étape de séquençage) qui, après séquençage, sont retrouvées aux extrémités de chaque read. Mais cette méthode n'a pas pu être utilisée car pour réaliser cet étiquetage, une librairie de 500ng d'ADN est nécessaire. Une approche similaire a été tentée au laboratoire, en dessinant des adaptateurs permettant d'identifier chaque individu. Ceux-ci étaient ajoutés sur les extrémités cohésives issues de la digestion de l'ADN génomique. Cependant, cette stratégie a dû être abandonnée suite à des difficultés de reproductibilité de l'expérimentation.

De ce fait, sans pouvoir obtenir l'information individuelle, suite au run de génotypage, nous ne pouvions pas directement estimer les fréquences alléliques des différents SNPs détectés. Travailler sur un pool d'individu permet de restreindre plus facilement les portions de génome à séquencer. La profondeur de séquençage doit en effet être réglée selon le compromis entre le nombre minimum de séquences pour un locus donné afin de détecter un polymorphisme (avec au minimum deux séquences nécessaire) et le 'gaspillage' de l'effort de séquençage de régions non-variables, ou de régions déjà représentées un grand nombre de fois pour lesquels un polymorphisme a déjà pu être identifié. Dans notre cas, la stratégie du regroupement de 10 à 20 individus dans chacune des deux librairies a montré son efficacité à fournir un niveau de polymorphisme avec une profondeur adéquat pour l'isolement de SNPs.

La stratégie *Reduced Representation Librairies* (RRL) qui a été employée a permis de maximiser la profondeur de séquençage de la portion génomique séquencée. Bien que cette profondeur de séquençage soit très faible (estimée à 2,5 grâce au logiciel MIRA3), un nombre suffisant de séquences présentaient une couverture permettant d'identifier 384 SNPS. De plus, nous avons pu utiliser ces paramètres de profondeur, afin d'exclure les erreurs de séquençage mais également, afin d'éviter les séquences d'ADN répétitives ou de loci dupliqués. Les SNPs ont par la suite été sélectionnés uniquement pour des séquences ayant une couverture entre 4 et 10.

La réussite de cette stratégie est directement liée à la qualité des séquences obtenues (qualité moyenne obtenue de 33). L'identification du site de restriction aux extrémités des séquences sur 93% des séquences nous a également permis de confirmer la réussite de la stratégie employée.

B. Une stratégie d'identification de SNP validée

En raison de l'absence d'un génome de référence, de la très grande taille du génome d'*I. ricinus*, de la forte densité de SNPs et d'éléments répétés au sein du génome, et de par la fréquence des erreurs de séquençage obtenues avec la technologie 454, la découverte de SNPs typables s'est avérée être un véritable challenge. Comme l'a montré l'utilisation des logiciels d'assemblage *de novo*, la tâche était fastidieuse. La stratégie avec l'utilisation du logiciel DiscoSnp nous a permis d'identifier un grand nombre de SNPs (321 088) ce qui confirme la large densité de SNPs observées précédemment dans le genre *Ixodes* (Noureddine *et al.* 2011; Van Zee *et al.* 2013). Cependant, du fait de la réduction génomique réalisée par RRL et de la sélection d'une fraction du jeu de données avec des profondeurs de séquençage différentes en fonction des différents loci, nous ne pouvons pas estimer la densité de SNPs dans le génome *d'I. ricinus*. La stratégie de sélection effectuée sur les critères de qualité, de profondeur, d'absence d'homopolymères et de similarité, nous a permis de ne conserver qu'une part restreinte de 1768 SNPs. Les filtres utilisés pour la sélection de ces 1768 SNPs ainsi que le choix de l'enzyme de restriction et le choix de la taille des fragments d'ADN utilisés pour la librairie qui évite les séquences répétées (comme les éléments transposables) nombreux dans le génome d'*I. scapularis* (Hill & Wikel 2005; Ullmann *et al.* 2005; Meyer *et al.* 2010) et qui sont inadaptées pour l'isolement de marqueurs mendéliens et co-dominants nous permettent de s'assurer de la qualité des SNPs mais également de leurs séquences flanquantes pour le design des amorces. De ce fait, nous en avons sélectionné 384, mais les 1384 autres SNPs pourraient également être appropriés pour la conception d'amorce et être génotypables.

Il faut tout de même noter que pour le pipeline développé et utilisé pour l'identification de SNPs, nous avons appliqué une certaine stringence dans nos critères. Dans notre cas, nous souhaitions identifier 384 loci SNPs et de ce fait nous avons pu établir des paramètres très sélectifs. Cependant, de nombreux autres SNPs isolés par DiscoSnp auraient pu être utilisés pour concevoir des amorces dans l'ensemble restant des sites polymorphes identifiés.

C. La validation des SNPs par génotypage

Parmi les 384 SNPs sélectionnés et génotypés dans un premier temps sur 464 individus issus d'une population naturelle de la Zone Armorique Atelier en Bretagne, nous avons obtenu un taux de validation de 96%, 368 ayant affiché à la fois une amplification et présenté du polymorphisme au cours du génotypage. Par contre, aucune amplification n'a été observée pour cinq SNPs tandis que 11 autres loci SNPs présentaient un seul type d'individu homozygote. Le taux de validation obtenu est plus élevé que dans d'autres études où des taux de SNPs n'ayant pas amplifié (ou donné de polymorphisme) variaient entre 6 et 52% (Sanchez *et al.* 2009; Hyten *et al.* 2010; Fu & Peterson 2012). Ceci reflète l'efficacité et la rigueur du pipeline mis en place pour la sélection des SNPs et le design des amorces.

Différentes hypothèses peuvent expliquer l'insuccès des 16 SNPs.

Tout d'abord, le SNP peut être présent dans le génome des individus testés, mais les amorces conçues pour ces loci n'ont pas permis d'amplification, dû à un mauvais design ou une mutation dans la séquence flanquante empêchant l'amorce de s'hybrider. Deuxièmement, ces SNPs pourraient être due à des erreurs de séquençage et donc aucune amplification n'est obtenue par ce SNP. Cependant, cette hypothèse semble peu probable en raison des différents filtres de sélection que nous avons appliqués. Troisièmement, ces 16 SNPs pourraient correspondre à allèles rares qui étaient seulement présents dans l'une des deux populations dont ils ont été isolés (proche de Toulouse et proche de Nantes), mais pas dans la population utilisée pour le génotypage (proche de Pleine-fougères en Bretagne- situé respectivement à 800 et 200 km des populations séquencées). Enfin, l'importante quantité de données manquantes pourrait expliquer l'absence de signal ou de polymorphisme obtenu pour ces différents loci.

D. la sélection et validation des SNPs.

La stratégie employée a permis d'isoler 368 SNPs montrant du polymorphisme. Ces marqueurs ont été choisis après différentes sélections (filtres) afin de pouvoir les utiliser avec confiance dans la partie suivante de ma thèse (Chapitre 3). L'important taux de données manquantes (27,5%) observé dans le jeu de données initial nous a incité à sélectionner les SNPs en éliminant ceux pouvant poser des problèmes dans l'utilisation de logiciels de génétique des populations qui tolèrent un seuil restreint de données manquantes. Pour ceci nous avons sélectionné les SNPs ainsi que les individus du jeu de données qui sera exploité par la suite de manière séquentielle. L'objectif était d'obtenir un set de 100-150 marqueurs afin de conserver la puissance du nombre de marqueurs, certaines études ayant montré que l'ajout de marqueur supplémentaire n'apportait pas d'informations complémentaires (Smouse 2010). Santure *et al.* (2010) ont montré que 20 microsatellites donnaient environ la même résolution que l'utilisation de 50 SNPs. En conservant les SNPs présentant moins de 25% de données manquantes, les analyses effectuées ne seront que plus fiables. En parallèle, l'analyse portant sur le comportement des différents loci lors de la ségrégation des allèles par l'analyse des croisements nous permet de nous affranchir de divers problèmes qui ont pu être constatés avec l'utilisation des marqueurs microsatellites. Les marqueurs microsatellites développés jusqu'à présent chez *Ixodes ricinus* ont montré, pour certains, une transmission non mendélienne mais également une large proportion d'allèles nuls. De ce fait, par l'analyse de la ségrégation des allèles pour chaque loci SNP et en ne conservant que ceux correspondant à une ségrégation mendélienne, les biais précédemment observés pour les marqueurs microsatellites ont pu être réduits au maximum. Les 128 SNPs sélectionnés au final, permettent d'obtenir un set de marqueur considérable afin de réaliser les analyses de génétique des populations à l'échelle du paysage, qui constituent le chapitre suivant de ce manuscrit.

L'analyse réalisée à l'échelle continentale, permet de plus de valider ces marqueurs. Ils ont en effet permis de montrer une différenciation génétique en fonction des origines géographiques des différentes populations analysées, notamment entre les populations ayant une origine tunisienne et les populations d'origine française. Ces résultats sont, de plus, concordants avec ceux apportés par les études précédentes (de Meeûs *et al.* 2002; Noureddine *et al.* 2011). L'étude de De Meeus *et al.* (2002) montrait une différenciation élevée entre les populations suisses et tunisiennes (θ entre 0,1 et 0,13). Cette forte différenciation entre ces populations situées sur deux continents différents a également pu être montrée par l'analyse des différents croisements que nous avons réalisés. Les trois croisements analysés présentant un des parents d'origine tunisienne présentent de fortes différenciations génétiques par rapport aux autres populations (aussi bien issues de croisements

contrôlés qu'issues du terrain) (θ variant entre 0,21 et 0,30). Bien que cette forte différenciation soit en partie due à une base génétique étroite (les populations étant représentés par le couple parent et une dizaine de représentants de leurs descendances), elle reste tout de même plus élevées que pour la différenciation génétique observée entre croisement réalisés entre deux parents d'origine française (θ= 0,18). Ces résultats confirment donc les divergences génétiques ayant pu être observés dans les études de De Meeus *et al.* (2002) et de Noureddine *et al.* (2011).

Chapitre 3 :

Analyses de génétique des populations d'*I. ricinus* à l'échelle du paysage

I. Introduction

Comme nous l'avons vu en introduction (voir Chapitre I.C), afin de mettre en place des moyens de lutte efficace contre *Ixodes ricinus*, il est nécessaire de connaitre ses capacités de dispersion mais également son niveau de variabilité génétique et comment celle-ci est distribuée dans l'espace.
La génétique des populations, dont un des objectifs principal est de décrire la distribution de la variabilité au sein des populations, constitue une approche incontournable par les outils qu'elle propose, d'une part pour détailler la répartition de la diversité génétique dans l'espace et dans le temps et d'autre part pour estimer le poids relatif des différentes forces évolutives ayant abouti à cette distribution. Malgré l'importance de cette approche, il existe encore de nombreuses lacunes dans nos connaissances sur le fonctionnement génétique des populations d'*Ixodes ricinus*.
Après une synthèse des connaissances disponibles sur ce sujet, je présenterai les différents facteurs biotiques et abiotiques agissant potentiellement sur la dispersion des tiques puis je mettrais en avant l'apport potentiel d'une approche de génétique des populations à l'échelle du paysage (« landscape genetics ») pour la compréhension du fonctionnement génétique des populations d'*I. ricinus*

A. Etat de l'art de nos connaissances actuelles sur la structure génétique de tiques

Plutôt qu'une liste exhaustive de toutes les publications sur la génétique des populations de tiques, j'ai choisi de ne présenter qu'une sélection d'articles représentatifs des différentes échelles spatiales étudiées, en prenant des exemples chez différentes espèces de tiques.

1. Etudes multi-échelle chez *Ixodes ricinus*

a) A l'échelle Eurasienne

Les premières investigations sur la structure génétique d'*I. ricinus* ont débuté en 1997 avec une étude réalisée en Suisse à une échelle régionale sur cinq lieux d'échantillonnage distants au maximum de 60km (Delaye *et al.* 1997). Un set de 18 marqueurs allozymes a été utilisé, permettant de caractériser une absence de structure génétique au sein de l'échantillonnage avec une faible variabilité allozymique (θ=-0,004) entre les cinq lieux. Cependant, parmi les 18 marqueurs utilisés,

seulement deux se sont montrés polymorphes, avec 90% de la variation due à deux allèles de chaque locus. Ce résultat, bien que basé sur seulement deux marqueurs, suggère que les populations à l'échelle régionale sont panmictiques. L'homogénéité génétique observée entre les différents échantillons pourrait être associée à la large gamme d'hôtes d'*I. ricinus* en Suisse, dont les oiseaux qui pourraient contribuer au brassage génétique et de ce fait à la diminution du niveau de différenciation entre populations (Delaye *et al.* 1997).

Les travaux suivants ont été réalisés à l'aide de marqueurs microsatellites, toujours à une échelle régionale (populations échantillonnées en Suisse, de part et d'autre du massif alpin (De Meeûs *et al.* 2002)). L'analyse, basée sur un set de six marqueurs microsatellites (développés par Delaye *et al.* (1998)), a montré une faible différenciation, même pour des populations éloignées de 200 km (θ compris entre 0,001 et 0,004) mais également pour des tiques échantillonnées de part et d'autre d'un versant dans les Alpes (θ=0,01) (De Meeûs *et al.* 2002). Ces résultats confirment l'absence de structure à l'échelle régionale identifiée lors de l'analyse précédente de Delaye *et al.* (1998).

A une échelle plus large, Casati *et al.* (2008) ont analysé le polymorphisme de cinq gènes mitochondriaux dans 26 populations européennes (provenant d'Italie, de Suisse, d'Autriche, du Danemark, de Suède et de Finlande). Le nombre de substitutions observées pour ces cinq marqueurs entre les différents haplotypes reconstitués (plus de 3 400 nucléotides) varie de 1,6 à 5 %. Cependant aucun patron n'a été trouvé entre la distribution des différents haplotypes et leur localisation géographique, les différents haplotypes se répartissant au nord comme au sud du transect européen échantillonné.

Noureddine *et al.* (2011) a également caractérisé la variabilité génétique d'*Ixodes ricinus* à l'échelle de l'ensemble de l'aire de répartition de l'espèce (ouest paléarctique, c'est-à-dire Europe, Afrique du Nord ainsi que l'Iran) au travers de 20 populations originaires de 14 pays différents et représentant 60 individus (Figure 3.1). Cette analyse basée sur cinq gènes présentant des niveaux de polymorphisme variables (gènes mitochondiaux -*Co1 et 16S*-, gène de ménage –*EF1α*-, gènes soumis à de fortes pressions de sélection potentiellement diversifiantes -*Trospa* et *Defensin*-) a également étudié les échelles locale (représenté par 20 individus provenant d'une même forêt) et régionale (représenté par 20 individus issus chacun de populations couvrant l'ensemble de la France). Les résultats de cette étude ont montré que le niveau de variabilité génétique (183 des 6963 nucléotidiques, soit 2,63 %, présentent du polymorphisme) est similaire entre les différentes échelles spatiales analysées (locale, régionale et européenne). Cependant, tout comme avait pu le montré Casati *et al.* (2008), cette variabilité génétique n'est absolument pas structurée spatialement.

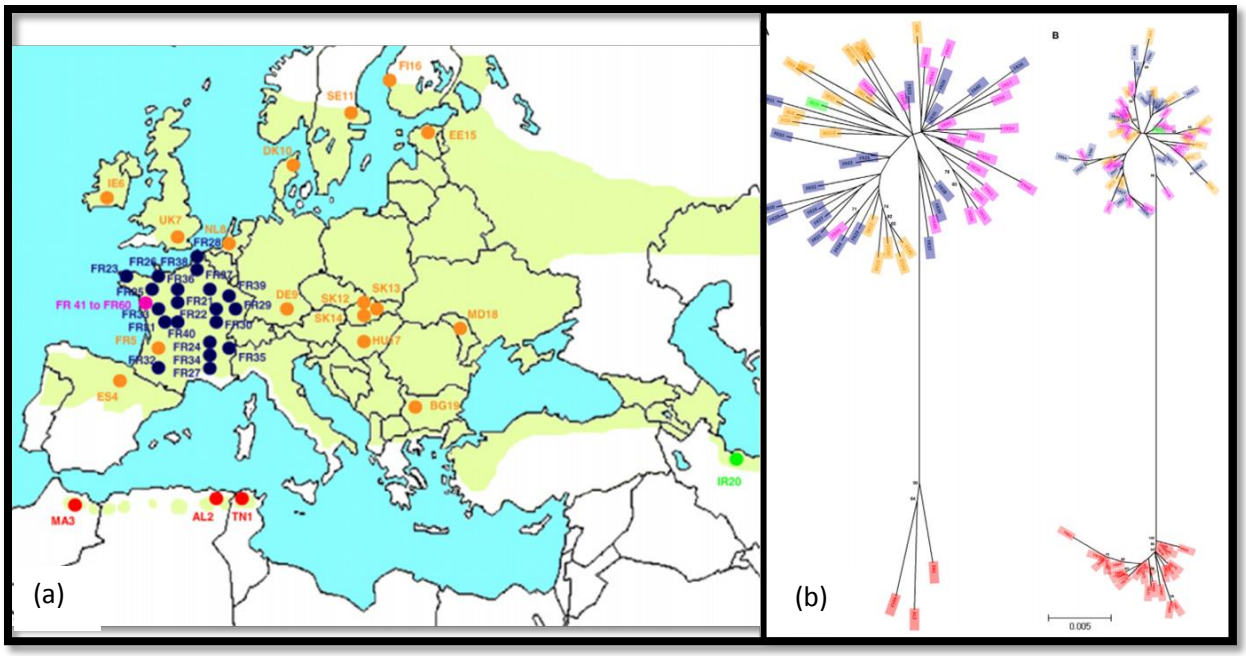


Figure 3.1 : **(a)** Origine des 60 individus d'*I. ricinus* étudiés pour l'étude de Noureddine *et al.* (2011). L'aire de répartition *d'I. ricinus* est représentée en vert. Les couleurs des points correspondent aux différentes échelles considérées : rose (locale), bleu (régionale), orange (européenne), vert (Iran), rouge (Afrique du nord). **(b)** Arbre phylogénétique (NeigbourJoining) entre les 60 individus séquencés lors de l'étude à partir des séquences concaténées des 5 gènes polymorphes (arbre de gauche). L'arbre de droite correspond aux séquences d'un seul gène (*TrospA)* pour lequel 20 individus supplémentaires de Tunisie ont été rajoutés. Les couleurs représentant les individus correspondent à celles utilisées pour l'origine des échantillons dans la figure 3.1.a.

Un travail similaire a été conduit par Porretta *et al.* (2013), basé sur le séquençage de quatre gènes (dont trois utilisés par Noureddine *et al.* (2011)) d'individus de 22 populations récoltées dans toute l'Europe, dont 9 en Italie. Cette étude ne montre pas non plus de structuration en Europe ni de pattern génétique en Italie, où il y aurait pu en avoir, révélant l'existence de refuge glaciaire dans la péninsule italienne durant la dernière période glacière. Cependant, elle suggère que, même pendant les périodes glaciaires en Europe, il y a eu des flux de gènes importants entre les refuges du sud de l'Europe.

L'ensemble de ces travaux réalisés à différentes échelles spatiales en Europe, montre donc une absence de différenciation génétique entre les différentes populations analysées, mais également une absence de structure phylogéographique à l'échelle européenne.

Différentes hypothèses peuvent être émises pour expliquer cette absence de structure à l'échelle européenne.

Elle pourrait être due à une récente évolution de l'espèce (Noureddine *et al.* 2011; Porretta *et al.* 2013). En effet un faible nombre de substitutions est observé entre les différentes populations. Cette récente évolution de l'espèce, couplée avec l'expansion des populations d'*I. ricinus* du sud vers le Nord de l'Europe à partir de refuges glaciaires datant du Pléistocène, pourrait expliquer la faible différenciation observée (McLain *et al.* 2001; Casati *et al.* 2008; Noureddine *et al.* 2011). Dans ce sens les travaux réalisés par Porretta *et al.* (2013) sont compatibles avec le scenario de l'existence d'un refuge glaciaire en Italie, même si les flux de gènes s'étant produits depuis auraient gommés tout pattern phylogéographique observable actuellement.

Enfin, les importants flux de gènes, caractéristiques d'une faible différence de niveau de variabilité nucléotidique, observés aux échelles régionale ou européenne peuvent s'expliquer par les mouvements des hôtes d'*I. ricinus* :

- d'une part, par les déplacements des hôtes sauvages tels les ongulés sauvages ou les oiseaux qui peuvent établir des flux de gènes efficaces, les oiseaux pouvant disperser les tiques sur de très longues distances (Ogden *et al.* 2008)

- d'autre part, par les échanges commerciaux d'animaux en Europe. Les politiques actuelles d'élevage en Europe favorisent les échanges commerciaux de bovins, avec des pays 'naisseurs', des pays 'engraisseurs' et des échanges de 'broutard'. Ainsi, environ deux millions de bovins circulent à travers différents pays d'Europe chaque année. Ces échanges, effectués également pour d'autres espèces hôtes d'*Ixodes ricinus* (moutons, faisans …), peuvent également favoriser les flux de gènes à travers l'Europe (Kurtenbach *et al.* 1998).

b) A l'échelle intercontinentale

L'analyse de la structure génétique a également été étudiée sur l'ensemble de l'aire de répartition d'*Ixodes ricinus*, à une échelle intercontinentale, en ajoutant aux populations européennes les populations d'Afrique du Nord. De Meeûs *et al.* (2002) ont comparé huit populations suisses et une population tunisienne avec cinq marqueurs microsatellites. Ces comparaisons ont montré une forte différenciation entre les populations suisse et tunisienne, qui est aussi plus marquée chez les femelles (θ = 0,13 pour les femelles et 0,10 pour les mâles). Les travaux menés par Noureddine *et al.* (2011) ont également été réalisés à l'échelle intercontinentale en intégrant des populations marocaine, tunisienne, algérienne et iranienne (Figure 3.1). Les résultats de cette étude montrent également une forte différenciation intercontinentale. Une forte divergence entre les populations Européennes et d'Afrique du Nord est donc observée, suggérant que la mer Méditerranée séparant ces deux aires géographiques, agirait comme une barrière aux flux de gènes entre les deux

continents, ce qui aurait conduit à la divergence entre les tiques de part et d'autre de la méditerranée.

Cette divergence entre continent peut s'expliquer également par des différences biologiques d'*I. ricinus* dues à son adaptation à chacun des milieux : en Afrique du nord, le climat est globalement plus chaud qu'en Europe, ce qui laisse à penser que les tiques africaines se sont adaptées différemment à ces régions. Elles présentent aussi des différences de phase d'activité, en Europe I*xodes ricinus* présente deux pics d'activité, au printemps et à l'automne, alors qu'en Afrique du Nord, sa période d'activité est à l'automne et en hiver. De plus les hôtes ne sont pas exactement les mêmes entre les deux continents. En Afrique du nord, les lézards semblent plus utilisés comme hôtes par les nymphes qu'en Europe.

Enfin cette divergence entre les populations peut être la conséquence de la dérive génétique qui agit largement sur les populations de petites tailles, telles que les populations nord africaines qui sont cantonnées aux massifs montagneux de cette zone. Une des hypothèses possibles serait que les tiques nord africaines auraient été isolées des tiques européennes lors de la dernière glaciation du Pléistocène (Noureddine *et al.* 2011), ce qui aurait engendré une évolution allopatrique des deux populations d'*I. ricinus* de part et d'autre de la Méditerranée.

Cependant, l'ensemble de ces résultats doit être interprété en prenant en compte les possibles biais liés aux marqueurs. Bien que les séquences nucléotidiques puissent mettre en évidence une différenciation intracontinentale (Noureddine *et al.* 2011), l'absence de structuration à l'échelle européenne observée par Casati *et al.* (2008) et Noureddine *et al.* (2011) peut-être due au faible polymorphisme de ces séquences. L'étude de De Meeûs *et al*. (2002) est, quant à elle, basée sur des marqueurs microsatellites qui, notamment par les problèmes d'allèles nuls, rendent difficiles leur interprétation et l'estimation d'une possible divergence (voir Chapitre 2.B) l'explication des biais portés par les marqueurs microsatellites).

2. Etudes chez d'autres tiques

Les investigations sur la génétique des populations des tiques ont porté sur un nombre limité d'espèces : *I. uriae, I. scapularis* et *Rhipicephalus (Boophilus) microplus*. Un rapide état des lieux est présenté dans ce manuscrit.

a) *Ixodes uriae*

Cette tique associée aux oiseaux de mer est surement l'espèce de tique pour laquelle le fonctionnement génétique des populations est le mieux connu. *Ixodes uriae* est vectrice de différents

agents pathogènes dont diverses *Borrelia* et notamment *B. burgdorferi*. Comme cette tique est inféodée aux colonies d'oiseaux de mer que l'on trouve dans des îles, les interprétations et analyses de la structure génétique des populations sont facilitées car cette situation correspond au modèle théorique de migration appelé « modèle en île» introduit par Wright (1948). McCoy *et al.* (2001; 2003) ont observé une forte diversité génétique intra-population et une relativement faible différenciation entre populations à l'aide de marqueurs microsatellites. Par ailleurs, l'existence de races d'hôtes à l'intérieur de cette espèce de tiques a été montrée. Ainsi, les individus récoltés dans des nids d'une même espèce d'oiseau (macareux moine ou mouette tridactyle) apparaissent plus proches génétiquement, même lorsqu'ils sont issus de colonies distantes géographiquement (Norvège, Royaume-Uni, France, Canada…), que les tiques prélevées dans la même région mais sur une autre espèce d'oiseau (Figure 3.2). Par ailleurs, le pattern d'isolement génétique par la distance est différent suivant les races d'hôtes, avec une différenciation plus forte chez la race « mouette tridactyle » que chez la race « macareux », même entre sites distants. La différenciation génétique des races d'hôtes chez les tiques se traduit aussi par des différences phénotypiques montrant l'adaptation des tiques à leur hôte (Dietrich *et al.* 2013).

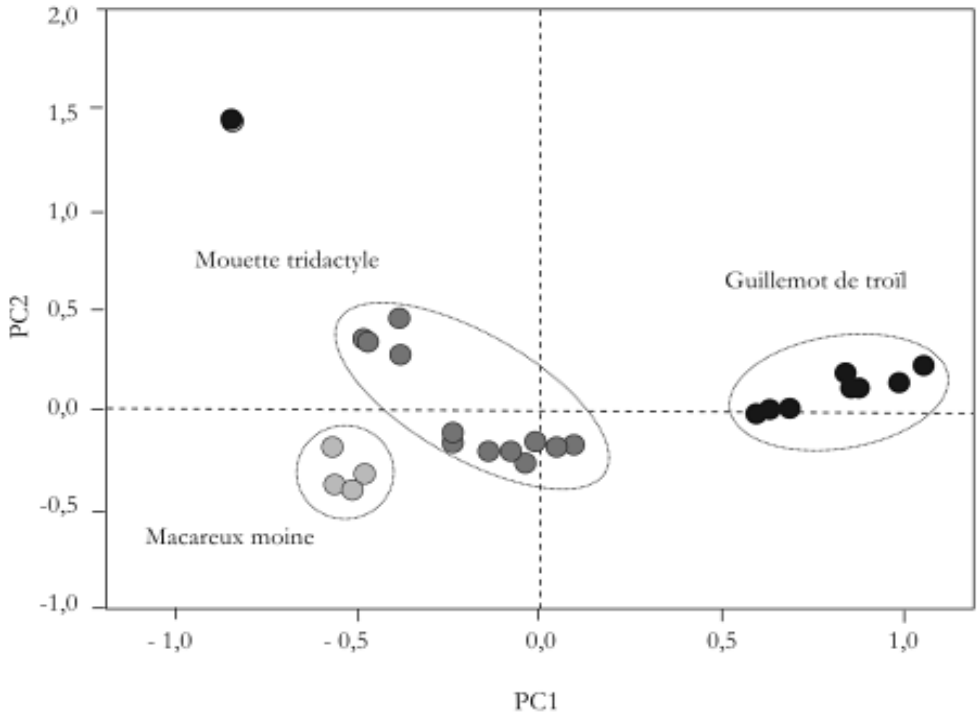


Figure 3.2 : **Analyse en composantes principales basée sur le polymorphisme de huit populations d'*I. uriae* échantillonnées dans différents sites européens** (chaque point du graphique représentant un site). Les tiques ont été prélevées dans des nids de différentes espèces hôtes (Mouette tridactyle, Guillemot de troil, Guillemot d'hornoya et Macareux moine). Sur ce graphique on voit bien que les différents échantillons se regroupent essentiellement par espèces hôtes (point de même couleur) et non selon leurs localisation géographiques *(d'après McCoy- communication personnelle).*

Ces études montrent donc bien l'importance des mouvements des hôtes, notamment à longue distance, pour les flux de gènes entre populations de tiques. La démonstration de l'existence de races d'hôtes montre aussi l'importance de la pression de sélection que constitue l'hôte pour structurer la diversité génétique des tiques. Cependant, étant donné la biologie très particulière de cette tique (qui contrairement à *Ixodes ricinus*, par exemple, ne s'alimente pas sur d'autres hôtes que des oiseaux), on peut se demander si une telle structuration liée à l'hôte est transposable chez d'autres espèces beaucoup plus polyphages.

b) *Ixodes scapularis*

En Amérique du Nord, malgré l'importance d'*I. scapularis* dans la transmission de la maladie de Lyme, très peu d'investigations relevant de la génétique des populations ont été effectuées. Une publication sur cette espèce avait utilisé un unique locus microsatellite avec des résultats forcément limités (Rosenthal and Spielman, 2004). Une étude menée par Qiu et al. (2002) analysant la séquence du gène ribosomique mitochondrial *16S* a mis en évidence 25 haplotypes différents, suggérant l'existence de deux clades assez divergents, un du sud (Caroline du Sud et certaines populations de Caroline du Nord) et un du Nord (Maryland, New-Jersey, New-York, Pennsylvanie, Connecticut, Massachusetts…). Qiu et al. (2002) ont également observé une forte divergence entre les deux populations de Caroline du Nord échantillonnées, l'une sur le littoral constituée uniquement d'individus de l'haplotype du nord (sauf un), l'autre à 80 km dans les terres avec un mélange des deux haplotypes du nord et du sud. Cette différence pourrait s'expliquer par la présence d'oiseaux migrateurs à l'origine d'un flux de tiques à l'automne du nord vers le sud.
Ils semblent qu'au sud, le spectre d'hôtes utilisé par *I. scapularis* soit différent, avec une forte proportion de lézards utilisés par les nymphes pour leur repas sanguins. Cette différence de comportement expliquerait aussi pourquoi les collectes au drap sont inefficaces dans cette région des Etats-Unis pour collecter *I. scapularis.*

c) *Rhipicephalus (Boophilus) microplus*

Cette espèce de tique tropicale ou sub-tempérée présente un cycle de vie qui se déroule sur un seul hôte, avec l'ensemble des stades (larves, nymphes, adultes) muant et se gorgeant sur le même individu (absence de détachement sauf pour les femelles gorgées). Cela a pour conséquence une multiplication très rapide, pouvant aller jusqu'à six générations par an. De ce fait, on s'attend à observer une forte consanguinité, favorisée par la présence d'individus apparentés sur le même animal. Cependant des études basées sur des loci microsatellites ont montré un assez faible déficit

en hétérozygote (*Fis* compris entre 0,03 et 0,071) (Koffi *et al*. 2006). Par ailleurs, une différenciation génétique entre troupeaux ainsi qu'un pattern d'isolement par la distance ont aussi été observés. Celui-ci pourrait être lié à un événement d'introduction unique à partir duquel les tiques se sont dispersées. Enfin, une forte différenciation a été détectée entre les tiques collectées sur bovins et celles sur cerf rufa (De Meeûs *et al.* 2010). Les spécificités de la biologie de cette espèce et de son histoire dans cette zone, où elle a été introduite, rendent difficilement extrapolables à d'autres espèces les enseignements tirés de ces études sur le fonctionnement génétique des populations de tiques.

B. Définir la structure génétique des populations d'*Ixodes ricinus* : une question d'échelle

Un des prérequis à l'estimation de la structure génétique et à l'estimation de la dispersion en génétique des populations est de se placer aux échelles spatiales et temporelles les plus adaptées à la réalité biologique de l'espèce considérée. De ce fait, définir spatialement une population est capital afin de s'intéresser à la dispersion et à la variabilité génétique. Ce concept de population est d'ailleurs un concept central en biologie et de nombreuses définitions peuvent en être trouvées dans la littérature (Waples & Gaggiotti, 2006). Une population peut être définie comme un groupe d'individus de la même espèce vivant dans une zone géographique de taille finie de manière à ce que chaque individu constituant cette population puisse se reproduire avec un membre du sexe opposé de cette même population (Hartl & Clark 1997). Cependant, la définition d'une population reste difficile notamment pour les espèces largement réparties (comme *Arabidopsis thaliana* ou *Drosophila melanogaster*). De plus, l'existence de structures géographiques ou paysagères conduit à des patrons de distribution spatiale des individus non aléatoires qui complexifie la reconnaissance de populations.

Dans ce contexte, la structure du paysage est un facteur clé, notamment pour *I. ricinus* qui a une écologie très stricte qui influe fortement sur sa distribution spatiale. Les notions de paysage et de structure du paysage se sont considérablement développées ces dernières années et font désormais l'objet d'une discipline scientifique à part entière, l'écologie du paysage. Néanmoins cette notion de paysage aborde différentes définitions selon qu'elle soit définie par des écologues ou des géographes par exemple. En écologie du paysage, la définition prédominante est de considérer le paysage comme la réalité physique d'un espace. Sans faire l'historique des différentes visions du paysage et de sa définition, l'ensemble des définitions est convergente et complémentaire. Le paysage est le

niveau d'organisation où les interactions entre organismes vivants, espaces et sociétés prennent toute leur signification. Sa structure nous renseigne sur l'histoire des relations entre les sociétés et leur environnement et elle est déterminante pour les processus écologiques. L'hétérogénéité spatio-temporelle du paysage résulte des interactions entre facteurs naturels mais également entre ces facteurs et les modes d'utilisation de l'espace par les sociétés. L'hétérogénéité du paysage peut être considérée comme « l'interprétation » de la structure spatiale et dépend donc de la nature des éléments paysagers. De manière générale, une structure paysagère est caractérisée par trois éléments essentiels (Figure 3.3) basés sur la configuration spatiale des unités paysagères : les taches (patch), les corridors écologiques et la matrice (Forman 1995).

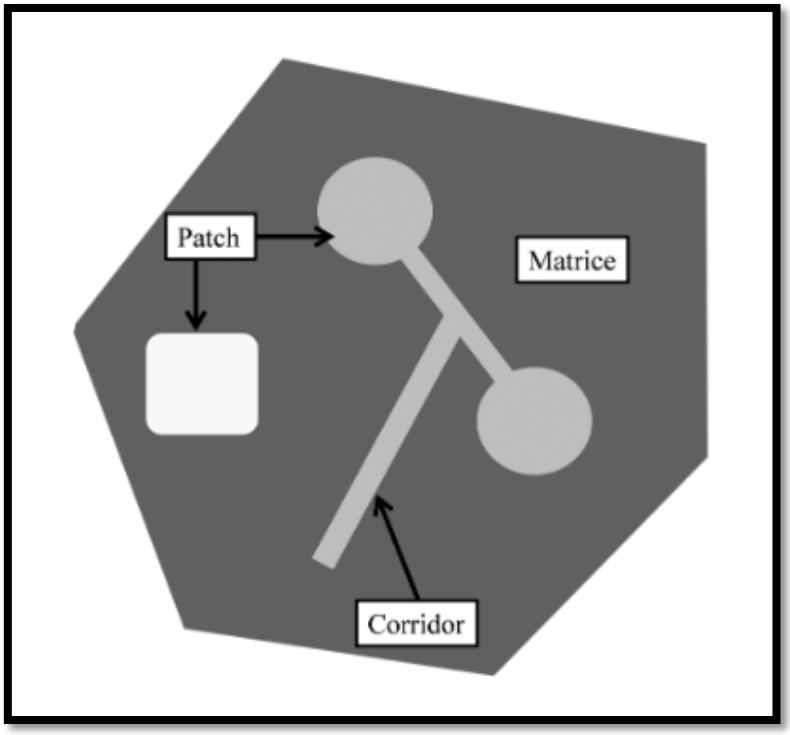


Figure 3.3 : Représentation schématique des éléments de base d'une structure paysagère : le patch (habitat), le corridor et la matrice (D'après Burel & Baudry 2003)

La structure du paysage peut donc se caractériser par des patches d'habitats favorables et d'autres moins favorables au maintien des populations d'*I. ricinus* mais également à celle de ses hôtes. Un paysage peut être homogène sur un grand espace ou à l'inverse être très fragmenté. Cette structuration paysagère avec des patches plus ou moins favorables conditionne de ce fait la taille des populations mais également les interactions entre ces dernières en fonction de leurs connectivités. Cela joue sur la dérive génétique ou sur les flux géniques qui agissent sur la différenciation génétique des populations. De plus, la structure du paysage influe également sur la répartition spatiale des hôtes d'*Ixodes ricinus* et sur leurs mouvements au sein du paysage.

De ce fait, il apparait nécessaire de prendre en compte les différents facteurs pouvant influencer la structure génétique d'*I. ricinus* à l'échelle du paysage.

C. Facteurs pouvant influencer la structuration génétique d'*Ixodes ricinus* à l'échelle du paysage

1. Les facteurs abiotiques : température et hygrométrie

Du fait de leurs exigences pour leur survie en terme de température et d'humidité, les tiques sont très sensibles aux conditions microclimatiques (Gray 1991; Estrada-Peña *et al.* 2004). Celles-ci vont donc affecter la dynamique des populations de tiques et leur distribution.
De plus les conditions climatiques agissent de manière indirecte sur la distribution des tiques en agissant sur la végétation. Le type de végétation conditionne le microclimat qui conviendra aux besoins des tiques. La végétation peut par ailleurs changer dans le temps, notamment en lien avec le réchauffement climatique, et agir de ce fait sur l'aire de distribution des tiques (Porretta *et al*. 2013) et sur la structure et la composition génétique des populations de tiques.

2. Les facteurs biotiques

a) La fragmentation du paysage

La fragmentation du paysage décrit un ensemble de processus qui transforme une surface continue d'habitat naturel en un nombre plus ou moins important de fragments de taille variable. L'ensemble de milieux qui en résulte – souvent hétérogène – séparant les fragments est communément désigné par le terme de « matrice » (Wilcove *et al*. 1986). La perte d'habitat peut se faire de plusieurs manières avec des conséquences très variables pour la configuration spatiale. La fragmentation peut se manifester par : (i) la réduction en surface d'un habitat, (ii) l'isolement de parcelles/fragments de l'habitat dans le paysage, (iii) l'augmentation du nombre de parcelles, (iv) la réduction de la taille de ces parcelles, (v) de plus grandes distances entre celles-ci, et (vi) une modification des propriétés de la matrice qui affecte le déplacement des individus. A l'heure actuelle, rares sont les habitats non fragmentés, notamment dû à leur transformation par des activités humaines telles que l'urbanisation, la construction de routes, la déforestation ou l'intensification de l'agriculture. Ces

activités réduisent les effectifs de certaines espèces, modifient leur distribution dans l'espace et les possibilités d'échanges entre populations, ou mettent au contraire en contact des espèces d'habitats différents jusque-là isolées. De manière générale, les espèces trouvées dans des paysages fragmentés forment souvent des métapopulations, c'est à dire un ensemble de sous-populations habitant des patches d'habitat isolés spatialement mais connecté via leurs migration (Kindlmann *et al.* 2005). L'isolement des habitats résultant de la fragmentation peut réduire l'effectif total des populations et les niveaux de diversité génétique, augmenter la fréquence des croisements consanguins, mais également diminuer le potentiel d'adaptation et avoir une influence négative sur la persistance à long terme de ces populations (Horskins *et al.* 2006). Ces effets sont une conséquence d'une réduction ou de l'arrêt de la dispersion, ce qui diminue le flux génétique (Burgman & Lindemayer 1998). La fragmentation du paysage pourrait ainsi agir sur la structuration génétique des populations de tiques, mais également induire une forte différenciation génétique à une échelle locale en isolant les populations les unes des autres (comme deux forêts isolées). Ceci pourrait également expliquer l'absence de tiques localement dans des biotopes 'idéaux' à un moment donné en raison des phénomènes fréquents d'extinction et de colonisation locale du aux possibles petites tailles de population.

b) La connectivité du paysage

Le terme connectivité, introduit par Merriam (1984), désigne le degré avec lequel un paysage facilite ou empêche le mouvement entre différents patches de ressources, ou l'intégration de sous-populations dans une unité fonctionnelle (Horskins *et al.* 2006 ; Taylor *et al.* 2006). Ainsi, la configuration d'un paysage, en termes d'usage de terre, de types et de quantité d'éléments paysagers, a une influence sur le mouvement des organismes qui s'y trouvent et par conséquent, sur la dynamique des populations et structures de communautés (Taylor *et al.* 2006). De ce fait, la connectivité tend à être considérée dans de nombreux travaux comme un paramètre directement lié à la présence de certains éléments paysagers qui facilitent la dispersion, comme les corridors biologiques (Figure 3.3). D'après Taylor *et al.* (2006), les mesures de connectivité les plus souvent utilisées ne prennent en compte que la taille des patches (habitats) et les distances inter-patches, ignorant par conséquent la complexité des réponses des organismes à l'hétérogénéité environnementale, qui peut avoir une influence sur leurs capacités de colonisation et de dispersion. Pour une estimation correcte de la connectivité au sein d'un paysage, il est nécessaire de prendre en compte son aspect fonctionnel. Cet aspect comprend, en particulier, les éléments qui favorisent le déplacement de chaque espèce. Par exemple, les haies représentent un élément paysager qui facilite

la dispersion des hôtes d'*I. ricinus* comme les micrommamifères, agissant comme des corridors biologiques pour la dispersion d'*I. ricinus* dans des paysages. L'estimation de la connectivité dans la dispersion d'*I. ricinus* doit prendre en compte notamment le mouvement des hôtes entre des différents éléments paysagers, leur capacité de dispersion mais aussi la mortalité de tiques liée à la dispersion dans un patch non favorable, comme un cœur de prairie (Taylor *et al.* 2006). Des méthodes génétiques sont également utilisées dans l'évaluation de la connectivité, en quantifiant les flux de gènes (Frankham 2006 ; Angelone & Holderegger 2009).

c) Les hôtes d'*I. ricinus*

i. Dispersion via les hôtes

Bien que pour certains auteurs, les communautés d'hôtes ne jouent qu'un rôle très marginal par rapport aux conditions climatiques (Klompen *et al.* 1996; Cumming 1999), leur présence est primordial au cycle de développement des tiques. De ce fait, l'abondance des tiques serait davantage déterminée par les variations d'effectifs des hôtes que par le climat (Randolph 2008).

Ixodes ricinus est une tique considérée comme généraliste (Chapitre 1.B et D), son spectre d'hôte étant très vaste allant des micromammifères aux cervidés, en passant par les oiseaux ou les lézards. Cependant chacun de ces hôtes n'est pas impliqué de la même façon dans la dynamique des populations de tiques, en fonction des stases des tiques qui se gorgent mais également de leur dispersion locale.

Les micromammifères connus pour être un hôte important pour les stases pré-adultes (Matuschka *et al.* 1991) se déplacent peu, mais peuvent malgré tout véhiculer des tiques d'un habitat à un autre et ainsi contribuer à leur dispersion à l'échelle locale. Dans ce sens, une étude réalisée dans le centre de la France suggère que le mulot sylvestre serait en partie responsable de la dissémination d'*I. ricinus* des milieux boisées aux prairies (Boyard *et al.* 2008).

Le chevreuil, qui est le grand mammifère sauvage le plus abondant en Europe de l'ouest (Morellet *et al.* 2013), est considéré comme un acteur primordial dans la dynamique des populations de tiques (Medlock *et al.* 2013). En effet, il est un des hôtes principal des adultes femelles qui, en se gorgeant, peuvent clore leur cycle de développement. En France et en Europe, la population de chevreuils a augmenté d'environ 50% en 50 ans (Apollonio *et al.* 2010), principalement du fait de l'instauration de quotas de chasse, de l'augmentation de zone boisées ou encore de la fragmentation du milieu forestier. En Europe, la distribution des tiques et leur distribution semblent se calquer sur celles des

chevreuils (Pichon *et al.* 1999; Vor *et al.* 2010; Jaenson *et al.* 2012a; b). Les chevreuils, de par leurs comportements, peuvent jouer un rôle très important dans la dispersion des tiques (Ruiz-Fons & Gilbert 2010). En effet, le domaine vital des chevreuils est composé d'éléments boisés (forêts, bois, haies, bosquets). Lorsque l'habitat est fragmenté, les chevreuils peuvent être amenés à réaliser des déplacements fréquents entre zones boisées. De plus lors des déplacements en période de reproduction, lors de la dispersion natale ou encore lors des migrations entre leurs domaines estivaux et hivernaux, les chevreuils sont amenés à réaliser des déplacements sur de longues distances (> 10km) (Debeffe *et al.* 2012). Tous ces comportements favorisent la dissémination des tiques entre zones boisées et sur de grandes distances.

Bien que les chevreuils semblent jouer un rôle clé dans la dispersion des tiques, les oiseaux, ressource importante des nymphes, pourraient aussi contribuer aux mouvements des tiques. Dans une étude, Plantard *et al.* 2010 [communication à ICOPA 2011 – Australie] ont réalisé un échantillonnage au niveau de quatre fleuves français (Loire, Garonne, Rhône et Rhin) en réalisant un transect de quatre populations (deux de part et d'autres de chaque fleuve). Cette étude, investiguant la variabilité génétique des différentes populations échantillonnées le long de ces transects, n'a pas montré de différenciation plus significative entre les populations échantillonnées sur la même rive qu'entre les populations séparées par un fleuve. Ceci suggère un rôle important des oiseaux dans la dispersion des tiques, les mouvements des oiseaux n'étant pas contraints par la présence d'un fleuve, à l'inverse des chevreuils par exemple.

Différentes études se sont intéressées au portage de tiques par des oiseaux, des merles noirs, des bruants (Comstedt *et al.* 2006; Ogden *et al.* 2008) ou encore des faisans de Colchide (Hoodless *et al.* 2002; Whitfield 2002), rapportant de forts portages de tiques par ces oiseaux (par le nombre d'individus parasités et par le nombre de tiques portées par individu). Ceci pourrait expliquer le brassage génétique des tiques et donc l'absence de structure génétique observé.

À l'échelle de l'Europe, les oiseaux migrateurs joueraient peut être ce même rôle. Des études ont rapporté des taux de portage des oiseaux migrateurs de 1 à 3% (Olsén *et al.* 1995; Comstedt *et al.* 2006; Ogden *et al.* 2008). Malgré ces taux de portage relativement faibles, la dispersion engendrée par les oiseaux migrateurs pourrait suffire à limiter la différenciation génétique entre les populations de tiques distantes.

A la vue de l'ensemble de ces études, les hôtes semblent jouer un rôle crucial dans la dissémination des tiques à différentes échelles et par conséquent pourraient influencer leur structure génétique.

ii. Race d'hôte

Bien qu' *I. ricinus* soit considérée comme un parasite généraliste, elle pourrait se spécialiser pour un hôte particulier à une échelle locale, en fonction des hôtes disponibles. Cette spécialisation pourrait structurer génétiquement les tiques par espèce hôte exploitée pour le gorgement. Dans ce sens, Kempf *et al.* (2011) ont comparé des tiques prélevées sur différents hôtes (chevreuils, sangliers, micromammifères, oiseaux et lézards) en Europe de l'ouest et centrale et ont montré des variations significatives de structure génétique entre les tiques collectées sur les différents hôtes. Ce résultat suggère que les tiques semblent évoluer en race d'hôtes lorsque les conditions locales sont favorables (Kempf *et al.* 2011).

d) Comportement des mâles et des femelles *I. ricinus*

Kempf *et al.* (2009) ont montré une importante tendance à l'homogamie chez certaines populations qui pourrait être une signature d'une sous-structuration au sein d'une population. Basées sur 7 loci microsatellites, les études réalisées ont démontré que les couples appariés collectés sur le terrain étaient apparentés, suggérant l'existence d'accouplements préférentiels ('assortative mating') au sein des populations.
De ce fait, l'assortative mating induirait une sous-structuration au sein des populations à une échelle fine, générant un effet Walhund lors d'analyse prenant en compte l'ensemble de la population.

Lors de leur étude de différentes populations suisse et tunisienne, De Meeûs *et al.* (2002), ont identifié une variabilité génétique différente entre les mâles et les femelles dans leur échantillonnage. Les femelles se sont révélées plus apparentées entre localités d'une même région que les mâles. Ceci s'expliquerait par une différence de comportement entre les sexes, avec une plus grande dispersion des mâles, due à une préférence trophique différente entre les stases larvaire ou nymphal des mâles ou des femelles. Les futurs mâles pourraient préférer se gorger sur des oiseaux alors que les futures femelles opteraient pour des micromammifères et ainsi la dispersion de ces dernières serait beaucoup plus faible. Ces mêmes données ont été analysées également à l'intérieur de chaque lieu d'échantillonnage (Kempf *et al.* 2010). Les auteurs ont identifié des sous-groupes au sein de chaque localité avec un patron d'appariement de couples en fonction de leur génotype. Cependant, à l'inverse de l'étude réalisée par De Meeus *et al.* (2002), les mâles sont plus apparentés entre eux que les femelles à l'échelle du site d'échantillonnage. Ce résultat renforce l'hypothèse de préférence trophique différente en fonction du sexe qui agit sur la dispersion des tiques.

e) Effet des agents pathogènes

De Meeûs et al. (2004) ont cherché à voir si le portage de parasites par les tiques pouvait jouer sur la dispersion des tiques. Ils ont montré une différentiation génétique plus forte entre individus porteurs de la bactérie *B. burgdorferi* qu'entre individus non-porteurs (θ = 0,004 versus 0,046 pour les femelles et -0,005 versus 0,14 pour les mâles). Il en ressort que les tiques infectées par *B. burgdorferi* sont plus dispersées que les tiques non infectées et ceci est plus marqué pour les mâles que pour les femelles. Il est possible qu'en agissant sur la mortalité et/ou la survie des tiques, les agents pathogènes jouent indirectement sur leur structure génétique.

D. Structuration génétique des populations d'*Ixodes ricinus* à l'échelle du paysage

Comme nous avons pu le voir, un grand nombre de facteurs peuvent influencer la structure génétique des populations d'*Ixodes ricinus* à l'échelle du paysage.

I. ricinus est une espèce inféodée au milieu forestier mais est retrouvée dans divers autres milieux (haies, landes, …), sa distribution étant corrélée au microclimat créé par la végétation des différents habitats. Ses hôtes (micromammifères, chevreuils, oiseaux, bovins) évoluent également dans des habitats très diversifiés, connectant des forêts, haies, cultures, pâtures, au sein de ce que l'on appelle un agro-écosystème. Les changements actuels des modes d'usage des terres entraînent la destruction, la transformation et la fragmentation de ces habitats naturels. Or, la structure paysagère, qui conditionne la connectivité des habitats, influence fortement la structuration spatiale de la variabilité génétique des espèces présentes.

L'étude de la variabilité génétique spatiale fournit donc de précieux renseignements sur les flux migratoires et la dynamique des populations. L'enjeu majeur de l'étude de l'organisation spatiale des populations réside dans la détection de barrières à la dispersion et dans l'identification des contraintes environnementales qui régulent les flux géniques dans ou entre les populations. L'organisation spatiale des populations est directement déterminée par la dispersion des individus qui est elle-même influencée par les propriétés et les éléments qui composent un paysage. La connectivité des différents éléments constituant le paysage, ayant des conséquences sur les mouvements des hôtes, apparaît donc importante pour comprendre le fonctionnement des populations de tiques.

Les changements de pratiques agricoles, qui accentuent la fragmentation du paysage, peuvent également avoir des répercussions sur l'abondance des tiques dans certains milieux et sur la circulation de pathogènes entre les espèces sauvages et domestiques.

Dans ce sens, plusieurs études menées aux Etats-Unis, ont pu montrer que l'épidémiologie de la maladie de Lyme semble être fortement influencée par les modifications du paysage (Bouchard *et al.* 2013). En effet, les pratiques agricoles ayant accentuées la fragmentation des forêts, seraient à l'origine d'une augmentation des cas de la maladie de Lyme (Rogic *et al.* 2013). Ce phénomène s'explique par la disparition progressive des différents hôtes d'*I. scapularis*, comme le cerf de Virginie ou les écureuils roux et gris favorisant d'autres hôtes, comme *Peromyscus leucopus* (également appelée 'souris à patte blanche') qui se trouve être un meilleur réservoir de *B. burgdorferi* par rapport au cerf de Virginie chez qui la bactérie se multiplie mal.

De manière générale, on assiste ces dernières années à une augmentation du nombre de cas d'infections humaines par des agents pathogènes transmis par les tiques, qui peut être expliquée en partie par une augmentation de la densité et une expansion de l'aire de distribution des tiques. L'exemple de cas d'infections humaines par le virus de l'encéphalite à tiques en Europe dans des zones nouvellement colonisées par les tiques va dans ce sens (Léger *et al.* 2013).

La circulation des agents pathogènes transmis par des tiques dépend de la densité et des espèces d'hôtes compétentes ainsi que de la présence de la tique dans un environnement donné. Il est donc nécessaire d'investiguer la structuration génétique d'*Ixodes ricinus* à une échelle fine, entre et au sein de localités, paramètre déterminant dans la circulation des agents pathogènes et dans l'épidémiologie des maladies associées.

Pour ceci, travailler à l'échelle du paysage est pertinent. La fragmentation du paysage peut influencer directement l'abondance et la structure génétique des communautés d'hôtes, réduisant la biodiversité en un lieu donné. Ceci peut conduire à une diminution de la transmission locale (via l'effet de dilution) mais également augmenter la transmission (abondance d'hôtes compétents). La fragmentation du paysage peut également conduire à un isolement des populations de tiques et d'hôtes entre différentes localités qui pourraient conduire, par des effets d'isolement mais également de dérive génétique, à une diminution de la transmission d'agents pathogènes à une échelle régionale.

La génétique du paysage ou 'landscape genetics' permet de décrire et de comprendre l'influence des structures paysagères et des facteurs environnementaux sur la structuration spatiale de la variabilité génétique (Manel *et al.* 2003, 2010 ; Holderegger & Wagner 2008). Jusqu'à présent, en génétique des populations, afin d'estimer la variabilité génétique, les populations étaient définies *a priori* (prédéfinie par des limites visibles).

Avec l'émergence de la génétique du paysage, des outils ont été développés, permettant de réaliser des inférences précises sur le nombre de populations et les éléments paysagers ou environnementaux pouvant les délimiter. Ces méthodes dites de 'clustering' (ou partitionnement) des individus sont basées sur des modèles de statistiques Bayésiennes (Pritchard *et al.* 2000; Manel *et al.* 2003, 2005) reposant sur l'hypothèse de l'existence d'un nombre K de populations dans un périmètre défini. Chaque individu, basé sur son génotype multilocus, est assigné à une population. Ainsi un nombre K de populations est alors identifié afin de minimiser les écarts à l'équilibre d'Hardy-Weinberg (Pritchard *et al*. 2000).

Certains outils développés récemment permettent également de prendre en compte la cadre spatial, via les coordonnées spatiales des individus, dans les analyses de génétique du paysage (Guillot *et al.* 2005; François *et al.* 2006; Chen *et al.* 2007). Ces outils permettent de mettre en relation la localisation de discontinuités génétiques et les contraintes spatiales qui influencent la dispersion, comme des rivières , des routes ou des montagnes en empêchant les flux de gènes, ou à l'inverse comme la présence de corridor reliant des forêt pouvant accentuer les flux de gènes.

De ce fait, la connaissance de la structure des populations en relation avec le paysage est d'un intérêt capital pour comprendre le fonctionnement des populations, et plus particulièrement identifier les éléments dans l'espace qui façonnent la dispersion.

Se placer à l'échelle du paysage permet de faire la synthèse des différents facteurs potentiels évoqués précédemment pour expliquer la distribution de la variabilité génétique des tiques. Les marqueurs SNPs, développés dans le Chapitre 2 du présent manuscrit, s'avèrent particulièrement intéressants pour une analyse robuste de la différenciation et de la structure génétique des populations naturelles de tiques à l'échelle du paysage. Dans ce contexte, cette partie de mes travaux de thèse consiste à estimer la diversité génétique, mais également la structuration des populations à une échelle fine, dans une zone atelier où le paysage est déjà connu et décrit avec une grande précision.

II. Structure des populations d'*I. ricinus* à l'échelle du paysage

Les marqueurs SNPs étant validés pour une étude de variabilité génétique des populations d'*I. ricinus* (cf Chapitre 2.III.E), nous avons étudié l'influence du paysage sur la structuration des populations d'*I. ricinus.* Cette étude a été réalisée dans une zone atelier en Bretagne.

A. Matériels et méthodes

1. Populations naturelles échantillonnées

a) Description de la zone atelier

La Zone Atelier Armorique (ZAA) dans laquelle s'est déroulée l'étude se trouve aux alentours de Pleine-Fougères, en Bretagne (Figure 3.4). Cette zone atelier est située plus précisément au Nord-Est du département d'Ille-et-Vilaine (35) et au Sud de la baie du Mont-Saint-Michel. Cette zone statuée ILTER (International Long Term Ecological Research site), est une zone étudiée depuis longtemps puisqu'elle a été constituée zone atelier en 1990. Son paysage ainsi que toutes ses informations relatives sont consignées dans une base de données SIG.

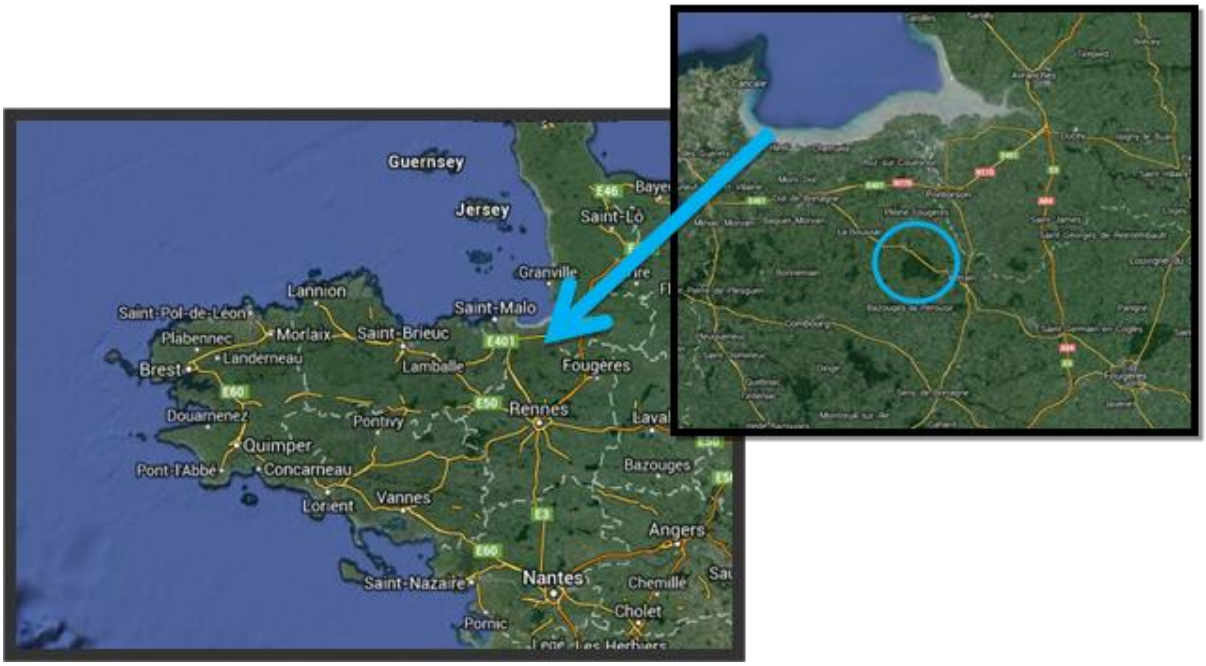


Figure 3.4 : Situation géographique de la Zone Armorique Atelier (ZAA)

Cette zone atelier couvre une superficie de 130 km². Le climat y est océanique, avec une pluviométrie moyenne annuelle de 750 mm, et une température moyenne de 6°C en janvier et de 18°C en août (site OSUR).

Dans le cadre du projet ANR OSCAR, dans lequel cette thèse s'inscrit, quatre secteurs ont été définis au sein de ZAA selon des caractéristiques paysagères : le cœur de forêt (CF), la lisière de forêt (LF), le bocage dense (BD) et le bocage ouvert (BO).

Les deux premiers secteurs, CF et LF, sont localisés au Sud-Est de la Zone Atelier. CF correspond au cœur de la forêt domaniale de Villecartier (980 hectares) et LF à la lisière de cette forêt. Le bocage dense (BD) se situe à l'Ouest de la forêt de Villecartier, et le bocage ouvert BO au Nord. Les deux bocages sont caractérisés par un paysage de pâtures, de haies et de bois. Le bocage ouvert se distingue du bocage dense par des pâtures de plus grande taille et un maillage de haies moins dense lié au remembrement ayant eu lieu dans cette zone.

Le socle géologique est granitique sur l'ensemble de la zone d'étude, sauf au Sud du bocage ouvert où il est schistique et l'altitude moins importante.

La Zone Atelier est favorable à la présence des tiques, avec une forêt pouvant être un habitat source pour ces acariens, et les réseaux de haies constituant des corridors jusqu'aux pâtures (Hoch *et al.* 2010). Ce type de paysage est également favorable à la présence de micromammifères et de chevreuils qui sont des hôtes de prédilection pour *I. ricinus*, leur permettant de se nourrir et de se déplacer dans les différents écotones (Boyard *et al.* 2008).

Par ailleurs, ces hôtes sauvages sont considérés comme des réservoirs d'agents pathogènes vectorisés par les tiques. Les bovins présents sur certaines pâtures de la Zone Atelier sont aussi des hôtes privilégiés pour ces ectoparasites, tout en étant, avec l'homme, des victimes potentielles de la transmission vectorielle.

b) Échantillonnage des tiques

La campagne de capture a été réalisée au printemps 2012, saison correspondant à une période d'activité des tiques. Nous avons échantillonné 90 lignes sur l'ensemble de la Zone Atelier, une "ligne" d'échantillonnage de tiques étant composée de 10 "tirages" de 10 m linéaires séparés les uns des autres par une distance de 20 m. L'ensemble de ces 90 lignes est réparti comme suit : 10 lignes dans le secteur CF, 20 lignes dans le secteur LF, 30 lignes dans le secteur BD et 30 lignes dans le secteur BO (figure 3.5). Les tiques ont été collectées à l'affut sur la végétation par la « méthode du drap » (un drap d'1 m² (1 m x 1 m)). Les lignes de collecte de tiques ont été géo-référencées grâce

aux coordonnées GPS obtenues sur le terrain. Chaque tique a été également géo-référencée de manière individuelle en prenant comme référentiel le centre du tirage de provenance.
Chaque ligne est identifiée suivant le code L0*XX*, selon la localisation de la ligne (Tableau 3.1).
Parmi les 90 lignes établies préalablement pour l'analyse, une ligne de cœur de forêt n'ayant pas été échantillonnée, le jeu de données est constitué en réalité de 89 lignes (Figure 3.5)

Au total, la campagne d'échantillonnage a permis de prélever 5108 larves, 2663 nymphes et 84 adultes sur l'ensemble des 89 lignes parcourues. La répartition entre les différents secteurs s'effectue comme suit (Tableau 3.1).

Tableau 3.1: Répartition des prélèvements effectués en fonction des différents secteurs d'études, CF, LF, BD et BO.

secteur	Nb de ligne		Larves	Nymphes	Adultes
Cœur de forêt (CF)	9	L001 -> L010	174	162	18
Lisière de forêt (LF)	20	L011 -> L030	344	405	19
Bocage dense (BD)	30	L031 -> L060	2838	1236	26
Bocage ouvert (BO)	30	L061 -> L090	1752	860	21

Pour les analyses génétiques, nous nous sommes intéressés uniquement aux nymphes. En effet, les larves présentent trop peu d'ADN pour être analysées et portent un moins grand nombre de pathogènes étant donné qu'elles n'ont pas encore réalisé de repas sanguins et les adultes sont trop peu nombreux pour constituer des échantillons de taille suffisante. Parmi les 89 lignes, aucune tique (nymphes du moins) n'a été prélevée sur six lignes. Pour les 83 lignes de collecte restantes, nous avons sélectionné une nymphe par tirage (10m) de chaque ligne (10 tirages/ligne) où au moins une nymphe a été prélevée. Ainsi chaque ligne est donc représentée dans le jeu de données avec une variation entre un et dix nymphes représentant chacune de ces lignes. 550 nymphes ont été sélectionnées de manière à obtenir un nombre le plus représentatif de la surface étudiée.

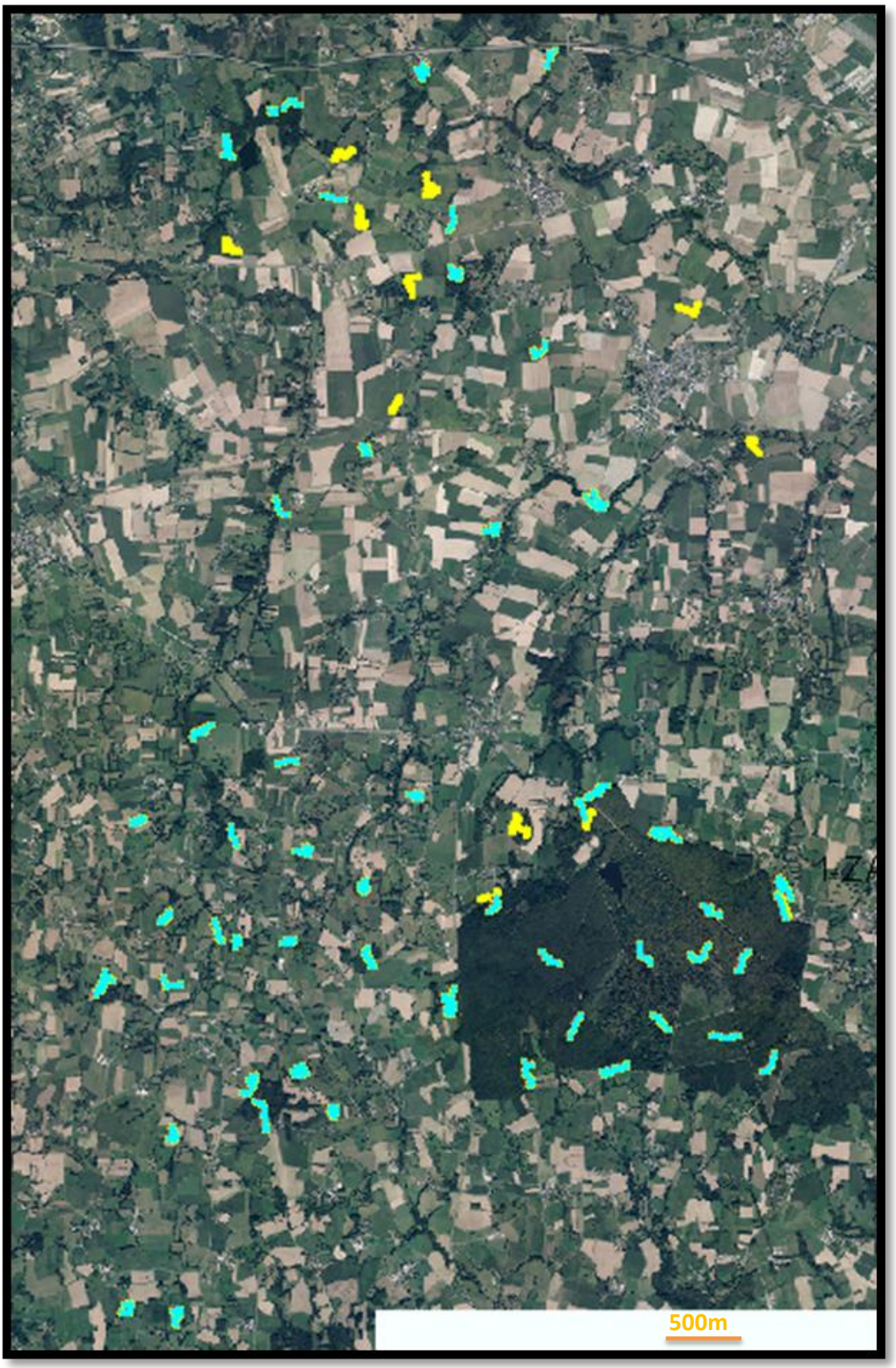


Figure 3.5 : Cartographie de ZAA. L'ensemble des lignes (bleu et jaune) correspondent aux 89 lignes de collecte échantillonnées dans le cadre du projet OSCAR. En bleu sont représentées les 71 lignes de collecte qui correspondent à celles analysées dans cette présente étude.

c) Extraction d'ADN et génotypage

L'extraction d'ADN de chaque individu a été réalisée en deux temps, une première extraction à l'ammoniac (dont l'extrait était réservé à la recherche de pathogène et l'identification des repas sanguins par les différents laboratoires partenaires du projet OSCAR) puis une deuxième extraction à partir de la carcasse de tique, suivant le protocole NucleoMag de Macherey-Nagel (extrait dédié au génotypage des tiques).

Le génotypage de 384 SNPs a été effectué à la plateforme génomique GENTYANE (Centre INRA de Clermont-Ferrand) via la chimie KASPar et la technologie Fluidigm (Biomark) pour un ensemble de 493 individus [cf. Chapitre 2.II.C]. Ces 493 individus, parmi les 550 préalablement sélectionnés, ont été choisis en fonction des quantités d'ADN disponibles pour chaque individu (les 57 individus présentant de faibles quantités d'ADN ont été écartés, les quantités ne permettant pas d'assurer un génotypage optimal).

d) Jeu de données final

Sur l'ensemble des 89 lignes de l'échantillonnage, 6 lignes ne sont pas représentées dans le jeu de données final (4 LF; 2 BO) car aucune nymphe n'y a été récoltée.

Parmi les 493 individus génotypés, les quantités d'ADN disponibles pour chaque individu ayant entrainé des variations interindividuelles dans le succès du génotypage et donc sur le nombre de données manquantes (variation entre 11,2% et 69% de données manquantes entre les individus), nous avons choisi de ne conserver pour l'analyse, que les individus présentant moins de 40% de données manquantes sur l'ensemble des SNPs sélectionnés (128 SNPs – *cf.* Chapitre 2.III). Ce critère avait réduit à 408 le nombre d'individus analysables. Cependant suite à l'étape de sélection des SNPs, le jeu de données de 408 individus/108 SNPs présentait encore un grand nombre de données manquantes. De ce fait nous avons réduit à nouveau le nombre d'individus, en retirant les individus présentant plus de 25% de données manquantes, réduisant.

Le jeu de donnée final est donc constitué de 371 individus et de 128 SNPs dont la distribution selon les secteurs est détaillée dans le tableau suivant (Tableau 3.2). Cette restriction a permis d'obtenir un jeu de données final présentant 10% de données manquantes, ce qui permet une analyse plus robuste.

Tableau 3.2 : Répartition des effectifs génotypés en fonction du secteur de collecte

Secteur	Nombre de lignes sélectionnées	Lignes de collecte	Nombre d'individus génotypés	Nombre d'individus analysés
CF	9	L001-L010	33	31
LF	16	L011-L030	80	66
BO	28	L031-L060	200	185
BD	30	L061-L090	180	89

La répartition des individus en fonction des lignes de collecte s'établit comme suit (Figure 3.6)

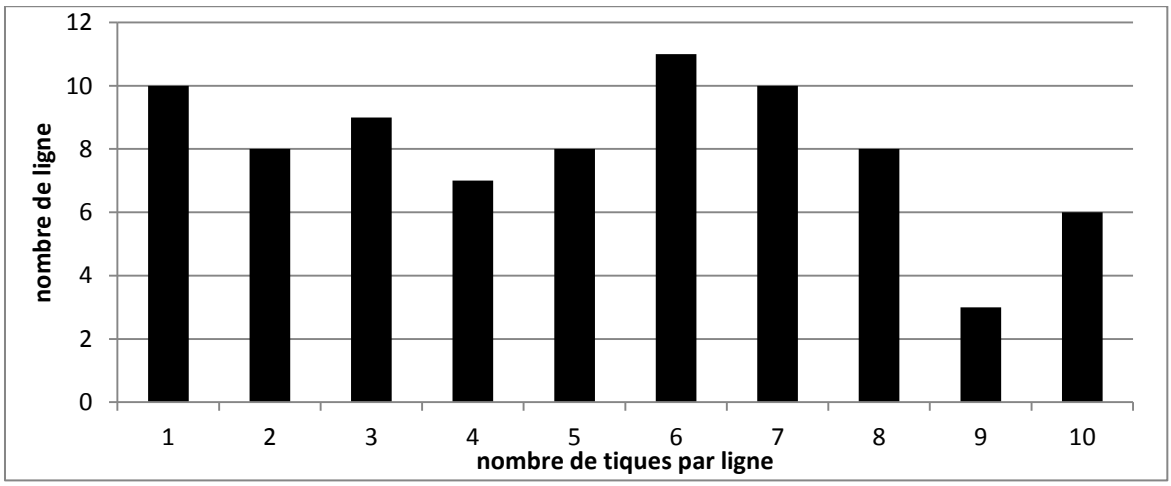


Figure 3.6 : Représentation du nombre de tiques par ligne de collecte.

2. Analyses génétiques

a) Consanguinité et fonctionnement des populations

La diversité génétique intra-population a été estimée à partir du calcul de l'hétérozygotie observée (H_{obs}) et de son estimation non biaisée attendue sous l'hypothèse d'équilibre d'Hardy-Weinberg (HW) (H_{att}), par le « test exact de Hardy-Weinberg » (Guo & Thompson 1992). L'hétérozygotie observée (H_{obs}) est le pourcentage réel d'individus hétérozygotes dans un échantillon. La population est « à l'équilibre » si H_{obs} est identique à H_{att}. Une valeur négative de cet indice exprime un déficit en hétérozygotes alors qu'une valeur positive exprime un excès par rapport à l'attendue à l'équilibre (Encadré 3.2). Une valeur de H_{obs} significativement plus faible que celle de H_{att} peut avoir plusieurs significations :

- il existe des pressions évolutives sur les locus étudiés

- la consanguinité est significative

- l´échantillon testé est formé de deux ou plusieurs populations distinctes au niveau de leurs reproduction («effet Wahlund »).

De plus, au sein de chaque population l'écart à l'équilibre d'HW a été estimé par calcul de l'indice *Fis* ('indice de fixation des individus dans les sous-populations') (Wright 1965). Le *Fis* représente le ratio d'hétérozygotie, en plus ou en moins, observé par rapport à l'hétérozygotie attendue (H_{att}), sous les hypothèses d'HW. Ce paramètre varie entre -1 et 1, les valeurs négatives correspondant à un excès d'hétérozygotes, les valeurs positives à un déficit en hétérozygotes et une valeur nulle correspondant à l'attendu sous l'équilibre d'HW. La significativité des tests a été établie au seuil de 5% après 1000 permutations. L'ensemble de ces analyses a été réalisée à l'aide du logiciel Genepop (Rousset 2008).

Encadré 3.2 : Population théorique idéale, population panmictique (d'après F. Fleury)

Une population théorique idéale est définie par les caractéristiques suivantes:

- une population d'organismes diploïdes à reproduction sexuée où les croisements sont entièrement aléatoires et à générations non chevauchantes

- une population d'effectif infini ne subissant donc pas les effets de la dérive génétique

- une population close génétiquement, donc présentant une absence de flux migratoire

- une population où tous les individus, quel que soit leurs génotype, ont la même capacité à se reproduire et à engendrer une descendance viable : absence de sélection

-absence de mutation et de distorsion des allèles. (Un individu ***Aa*** produira toujours 50% de gamètes ***A*** et 50% de gamètes ***a***).

Dans ce cas, la population est à l'équilibre de Hardy-Weinberg c'est-à-dire que les fréquences alléliques et génotypiques sont constantes au fil des générations.

Le croisement au hasard des individus, appelé système de reproduction panmictique, est une hypothèse essentielle, car elle permet d'estimer les fréquences génotypiques à partir des fréquences alléliques (Figure 3.7). Cette hypothèse suppose que les individus ne choisissent pas leur partenaire sexuel en fonction de leur génotype (panmixie) et que la rencontre des gamètes se fait au hasard (pangamie). Par exemple, dans le cas d'un locus à deux allèles **A** de fréquence p et **a** de fréquence q, si la reproduction est panmictique, on a les fréquences génotypiques $P(AA)=p^2$; $P(Aa)=2pq$, $P(aa)=q^2$

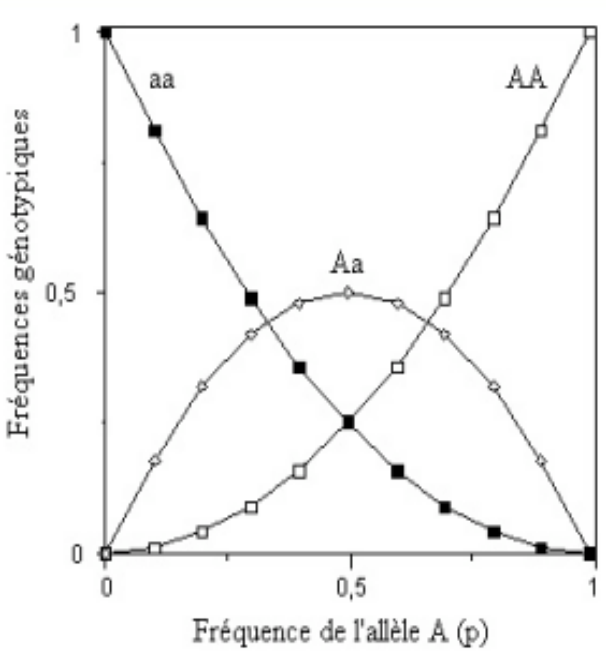


Figure 3.7 : Représentation des fréquences génotypiques en fonction des fréquences alléliques attendues sous l'équilibre d'Hardy-Weinberg.

b) Différenciation génétique et structure génétique des populations

La diversité génétique entre paires de populations a été estimée par l'indice de différenciation *Fst* de Wright (Wright, 1969), selon l'estimateur θ de Weir and Cockerham (1984). Cet indice mesure la variance des fréquences alléliques entre populations. Plus la valeur de cet indice s'éloigne de zéro, plus la différenciation génétique entre les populations est élevée. L'analyse a été effectuée avec le logiciel Genepop et le seuil de significativité du test a été établi au seuil de 5% après 1000 permutations.

La structure génétique des populations a également été explorée par deux méthodes différentes. Dans un premier temps, par une analyse factorielle des correspondances (AFC) sur les fréquences alléliques des populations a été effectuée à l'aide du logiciel Genetix (Belkhir *et al.* 2004). Puis dans un second temps par une méthode d'analyse bayésienne de classification, réalisée à l'aide du logiciel STRUCTURE (Pritchard *et al.* 2000).

c) Isolement par la distance

Le test de Mantel d'isolement par la distance a été effectué (Mantel 1967) avec le logiciel Genepop (Rousset 2008). Pour cela, les distances génétiques entre paires de populations sont habituellement estimées par l'indice de différenciation *Fst* (Rousset 1997) qui sont ramenées à *Fst*/1-*Fst*.

Cependant, au vu de l'hétérogénéité de notre échantillonnage avec des lignes de collecte, qui sont la plus petite représentation d'une population au sein de notre échantillonnage, représentées par un nombre d'individus allant de un à dix, j'ai choisi d'utiliser la procédure d'isolement par la distance entre individus (Rousset 2000 ; Watts *et al.* 2007) implémentée par le logiciel Genepop. Pour ceci, chaque individu est considéré comme une sous-population, et le test peut être effectué avec la statistique *â* (un équivalent du *Fst*/(1-*Fst*) pour la différenciation entre individus). Pour constituer la matrice de distance géographique, les coordonnées spatiales uniques de chaque individu ont été utilisées. La significativité du test a été choisie au seuil de 5% après 1000 permutations.

d) Analyse Moléculaire de la Variance (AMOVA)

Chacune des méthodes d'analyses évoquées ci-dessus a été effectuée à différentes échelles au niveau de la zone atelier, de l'échelle la plus large (la zone atelier) à l'échelle la plus fine (la ligne de collecte). Ces différentes échelles d'études, correspondant à autant de niveaux hiérarchiques, structureront la partie concernant les résultats.

Afin de déterminer la partition de la variabilité génétique aux différentes échelles considérées (secteurs, lignes, individus au sein des lignes), une AMOVA (Analysis of Molecular Variance) a été réalisée avec le logiciel ARLEQUIN (Excoffier *et al.* 1992). Pour ce faire, une matrice de distance entre les différents génotypes est calculée et ARLEQUIN détermine la partition des sommes des carrés des écarts de cette matrice de distance en fonction des différentes composantes hiérarchisées de la variance.

B. Résultats

1. A l'échelle de la zone atelier

Sur la totalité de l'échantillon étudié (N=371) représentatif des tiques collectées à l'échelle de la zone atelier, on observe un déficit significatif en hétérozygotes par rapport aux proportions attendues selon l'équilibre d'Hardy-Weinberg (HW) (*Fis* = 0,18) lorsque l'ensemble des loci est considéré. Lorsque l'on considère les marqueurs pris de manière individuelle, une variation très forte est observée (valeur de *Fis* comprise entre -1 (SNP207995) et 0,86 (SNP166766)). Sur les 128 marqueurs, 106 présentent un déficit significatif en hétérozygotes (p>0,05). Un écart aussi conséquent entre les valeurs de *Fis* observées ainsi que le nombre important de loci présentant des déficits en hétérozygotes pourraient être dus à la présence d'allèles nuls. Cependant, cette explication reste tout de même peu vraisemblable étant donné le caractère biallélique des SNPs. Il parait plus plausible de privilégier l'hypothèse d'un effet Walhund dû à un échantillonnage englobant plusieurs populations fonctionnelles (*i.e*. groupe d'individus échangeant librement des gènes).

La différenciation génétique globale, sur l'ensemble des populations, confirme également cette hypothèse. A l'échelle de la zone atelier, en prenant en compte l'ensemble des 371 individus comme constituant une seule population, l'analyse factorielle des correspondances (AFC), réalisée avec GENETIX 4.0.5.2, ne montre aucune structuration (aucun pattern de points ne semble se regrouper permettant de différencier certains groupes dans l'échantillonnage) (Figure 3.8). Les axes de l'AFC expliquent 4,88% de la différenciation (Axe1 : 2,58% ; Axe2 : 2,30%), ce qui confirme l'absence de pattern observée à cette échelle.

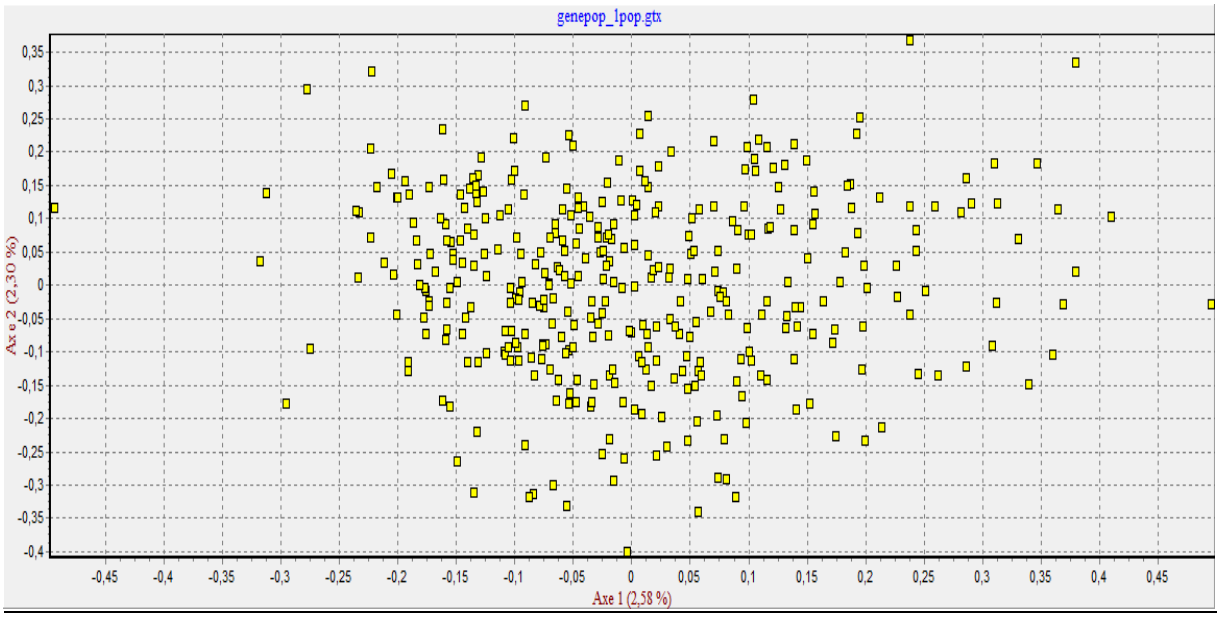


Figure 3.8 : Analyse factorielle des correspondances (AFC), réalisée avec GENETIX, en prenant l'ensemble des individus à l'échelle de la zone atelier.

Les différents sites échantillonnés (lignes de collecte) sont répartis dans la zone atelier sur un large espace géographique, les lignes les plus éloignées étant distantes d'environ 15 kilomètres. De plus, les sites choisis pour l'échantillonnage au cœur de la zone atelier sont issus de différents secteurs environnementaux (pour rappel : le cœur de forêt 'CF', la lisière de forêt 'LF', le bocage dense 'BD' et le bocage ouvert 'BO'). Par exemple, les deux secteurs 'bocage' (BO et BD) étudiés sont différents par leurs maillage dans le réseau de haies, la présence de bois, du nombre de culture et/ou prairies les composant.

De ce fait on peut faire l'hypothèse d'une sous-structuration de la diversité génétique due à la composition paysagère 'fine'. Les tiques d'un milieu forestier, plus adaptées au milieu humide, pourraient être différentes génétiquement de celles que l'on retrouve au niveau des pâtures, du fait des conditions environnementales différentes mais également des différences d'hôtes entre ces différents environnements. Ces différents secteurs identifiés peuvent donc structurer la diversité génétique et être à l'origine d'un effet Walhund.

Pour cette raison, nous allons considérer par la suite séparément les différents secteurs afin de savoir si on observe toujours un tel déficit d'hétérozygotes.

2. A l'échelle des quatre secteurs (CF, LF, BD, BO)

Sur l'ensemble des quatre populations représentant les différents secteurs de la zone atelier, l'écart à l'équilibre d'HW montre toujours un déficit significatif en hétérozygotes (*Fis* = 0,118 à 0,170). De plus l'hétérozygotie observée est toujours plus faible que celle attendue sous l'équilibre d'HW (Tableau 3.3).

Tableau 3.3 : Hétérozygotie des populations *d'Ixodes ricinus* échantillonnées au sein de la zone atelier séparées selon les 4 secteurs (CF, LF, BD, BO) ou étudiées en ne considérant qu'une population unique prise ensemble. Les estimations ont été calculées à l'aide du logiciel Genepop ; N= nombre d'individus génotypés ; H_{att}= hétérozygotie attendue et non biaisée ; H_{obs} = hétérozygotie observées ; *Fis* = indice de fixation intra-population

Population	**N**	**Nombre de lignes**	**H_{att}**	**H_{obs}**	***Fis***
CF	31	9	0,349	0,295	0,151
LF	66	15	0,352	0,307	0,118
BD	185	30	0,347	0,298	0,148
BO	89	17	0,353	0,291	0,170
Ensemble des populations (=1 seule population)	371	71	0,350	0,298	0,180

La différenciation génétique globale sur l'ensemble des populations est significativement différente de zéro. L'estimation de l'indice θ de Weir et Cockerham (1984) varie entre 0,001 et 0,006 (Tableau 3.4). La valeur la plus élevée est trouvée lors de la comparaison entre les populations du bocage ouvert (BO) et ceux de la lisière de forêt (LF).

De manière générale, le secteur CF présente les valeurs de θ les plus élevées. La valeur observée est la plus faible lors de la comparaison entre les populations des secteurs LF et BD (θ= 0,001).

Tableau 3.4 : Estimation des *Fst* entre les 4 différents secteurs

	LF	BD	BO
CF	0,0036	0.0056	0.0054
LF		0,0010	0.0062
BD			0,0020

La structure des populations a également été étudiée par une analyse factorielle des correspondances (AFC) (Figure 3.9). Pour l'analyse, les axes 1 et 2 ont été choisis pour la représentation car ils étaient les plus explicatifs de la différenciation. Les deux axes de l'AFC

expliquent 4,88% (Axe1 : 2,58% ; Axe2 : 2,30%) de la variance de fréquences alléliques entre les différents secteurs au sein de la zone atelier.

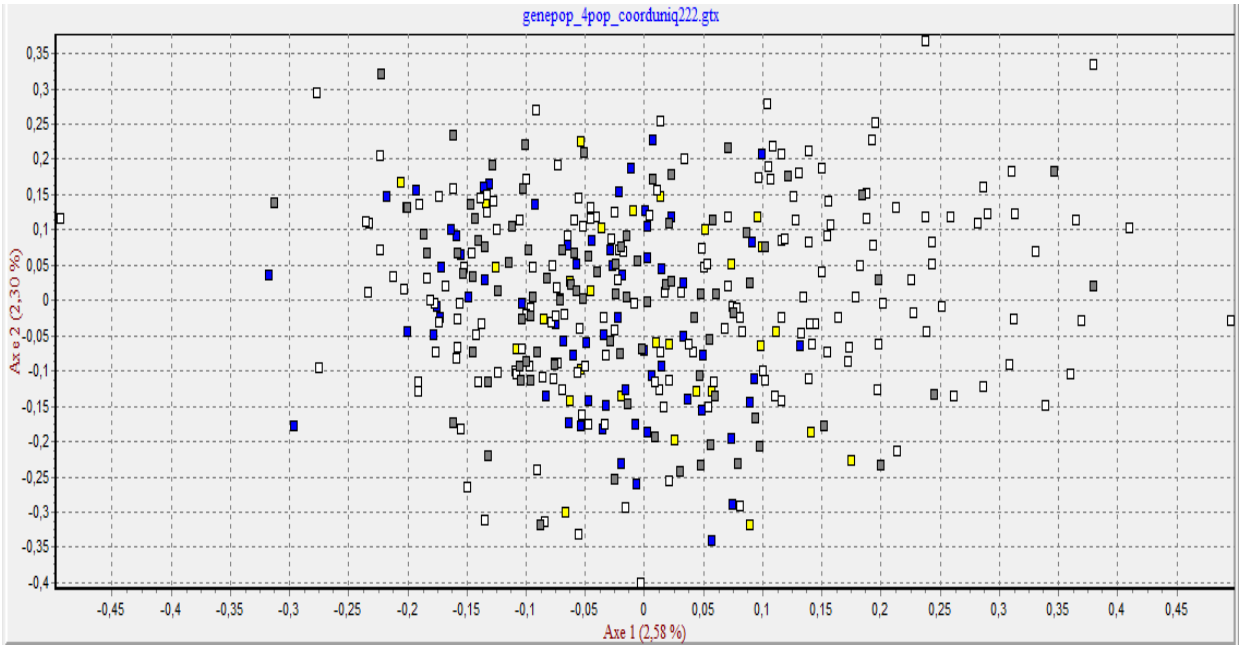


Figure 3.9 : Analyse factorielle des correspondances (AFC), réalisée avec GENETIX, sur les fréquences alléliques des populations des 4 secteurs échantillonnés : CF (jaune), LF (bleu),BD (blanc), BO (gris).

Aucun groupe correspondant à une population (secteur) n'est observé. L'ensemble des points des différentes populations constituent le même nuage de points, ne montrant aucune différenciation à l'échelle des quatre secteurs étudiés. Les points de la population BO (en blanc) sont les plus dispersés dans cette représentation graphique, ce qui peut s'expliquer par la plus grande couverture géographique de ce secteur par rapport aux autres mais également par le nombre d'individus plus important constituant cette population (185 contre 31 individus du secteur CF par exemple).

Etant donné que l'AFC n'a pas mis en évidence une possible différenciation entre les populations, nous avons testé une méthode différente basée sur une méthode bayésienne implémentée dans le logiciel STRUCTURE, à partir des données des génotypes en définissant quatre sous-populations au sein de notre échantillonnage (K=4) (Figure 3.10).

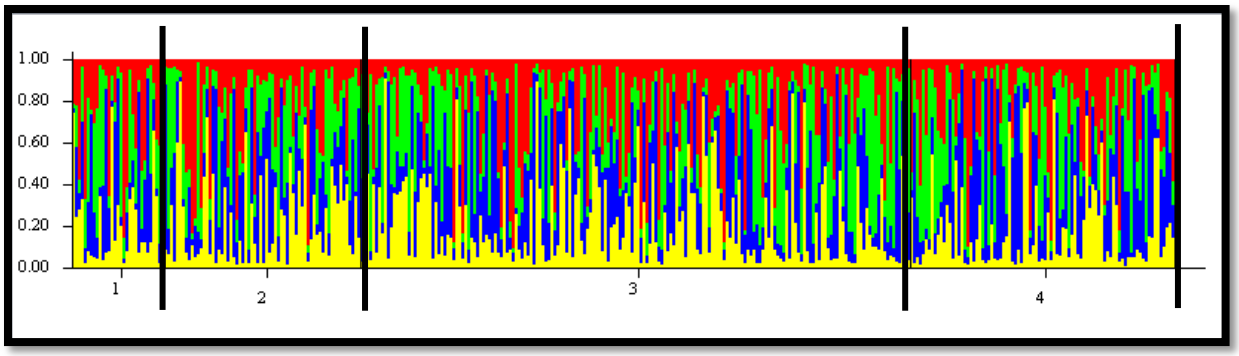


Figure 3.10 : Résultat du logiciel STRUCTURE pour un nombre K de 4 sous-populations. Les groupes 1, 2, 3 et 4 correspondent respectivement aux secteurs CF, LF, BD et BO.

Chaque groupe génétique testé étant représenté par une couleur différente, si les différents individus (représenté par une barre verticale) de la même population sont plus apparentés entre eux qu'avec les individus des autres populations, on doit s'attendre à avoir une couleur majoritaire par groupe génétique observé, ce qui n'est pas le cas ici. Le nombre le plus probable de groupes génétiques a été estimé de deux manières :

- d'une part, en utilisant la méthode initiale de STRUCTURE avec un critère ad hoc basé sur la distribution du logarithme des probabilités L(K), la valeur la plus élevée correspondant au K optimal (Pritchard et al., 2000) (Figure 3.11.a).

- d'autre part, en estimant K à partir du taux de changements de second ordre du logarithme des probabilités a posteriori entre les valeurs successives de K (Evanno *et al.* 2005) (Figure 3.11.b).

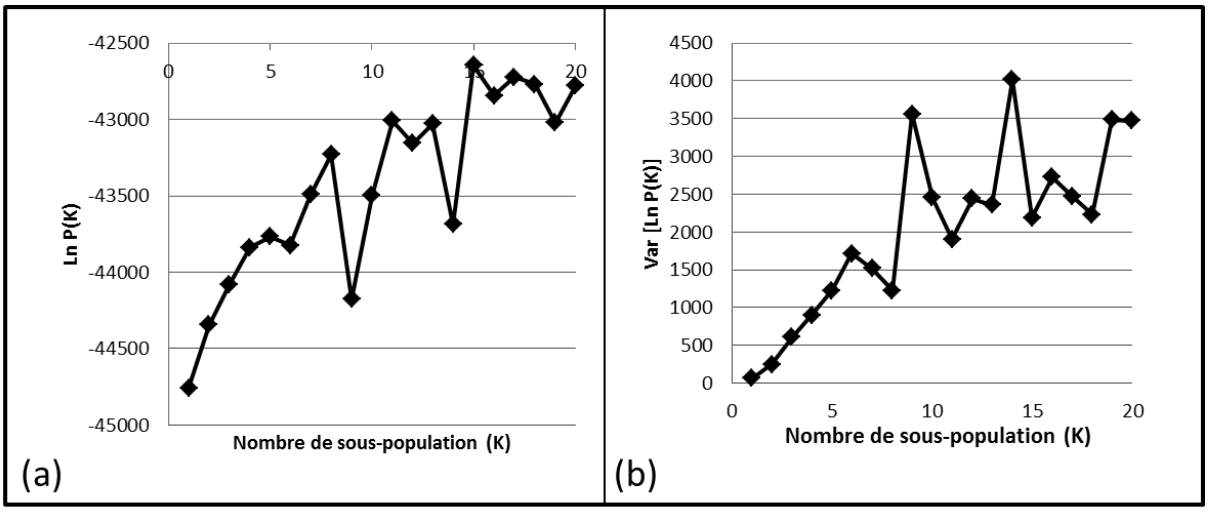


Figure 3.11 : Détermination du nombre de sous-populations (K) optimal dans notre échantillonnage par deux méthodes, (a) la moyenne L(K) du logarithme des probabilités ; (b) Variation de second ordre du logarithme des probabilités ΔK calculé selon la formule d'Evanno et al. (2005).

Ces courbes présentent un premier pic pour une valeur de K=9 sous populations (Figure 3.11). Cependant l'inférence avec neuf sous-populations ne permet pas de distinguer de sous-populations réelles au sein du jeu de données, comme on peut le voir sur le diagramme ci-dessous (Figure 3.12).

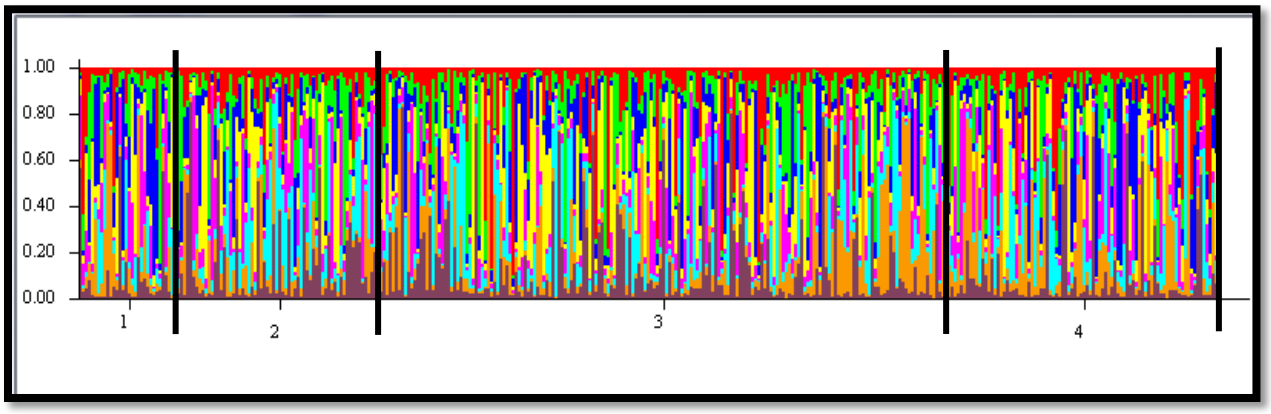


Figure 3.12 : Résultat du logiciel STRUCTURE pour un nombre K de 9 sous-populations. Les groupes 1, 2, 3 et 4 correspondent aux secteurs CF, LF, BD et BO respectivement.

Tout comme à l'échelle de l'ensemble de la zone atelier, il n'est pas possible de mettre en évidence de différenciation génétique entre les différents secteurs. Les fortes valeurs de *Fis* observées, accompagnées de l'absence de différenciation génétique entre les secteurs, laissent suggérer l'existence d'une sous-structuration au sein de ces quatre secteurs environnementaux que nous avons définis.

A la vue de ces résultats, confirmant l'absence de différenciation génétique entre les quatre populations, nous avons choisi de tester l'existence de sous-populations au sein de ces secteurs.

3. Relation entre les différents secteurs : rôle de la connectivité du paysage

a) Cœur de forêt

Etant donné l'écologie d'*Ixodes ricinus* qui est avant tout une espèce forestière cherchant l'humidité atmosphérique du sous-bois, on peut faire l'hypothèse que le secteur cœur de forêt correspond à une population « source » étant à l'origine des individus trouvés dans les autres secteurs (avec des conséquences sur la dynamique comme la génétique des populations de ces autres secteurs). Nous pouvons aussi le prendre comme secteur de référence pour la suite des analyses étant donné que

c'est le secteur d'un point de vue paysager le plus homogène de par sa végétation (au moins en terme de strates – herbacée, arbustive, arborescente…-). Ainsi, nous pouvons supposer que la fréquentation des hôtes est similaire entre les différents sites de collecte à l'intérieur de ce secteur (il n'y a pas de barrières majeures à leurs dispersions, ni de corridors préférentiels). Les neuf sites de collecte sont répartis dans le cœur de forêt sur une surface d'environ 7,2 km^2 et les lignes les plus distantes (L009 et L002) sont éloignées de trois kilomètres (Figure 3.13). Le calcul de l'indice de différenciation (*Fst*) entre les différentes lignes montre une grande homogénéité au sein des différents sites échantillonnés. Seules trois valeurs apparaissent significativement différentes de zéro, correspondant aux croisements de lignes de collecte L001, L002, L003 et L007 (Tableau 3.5).

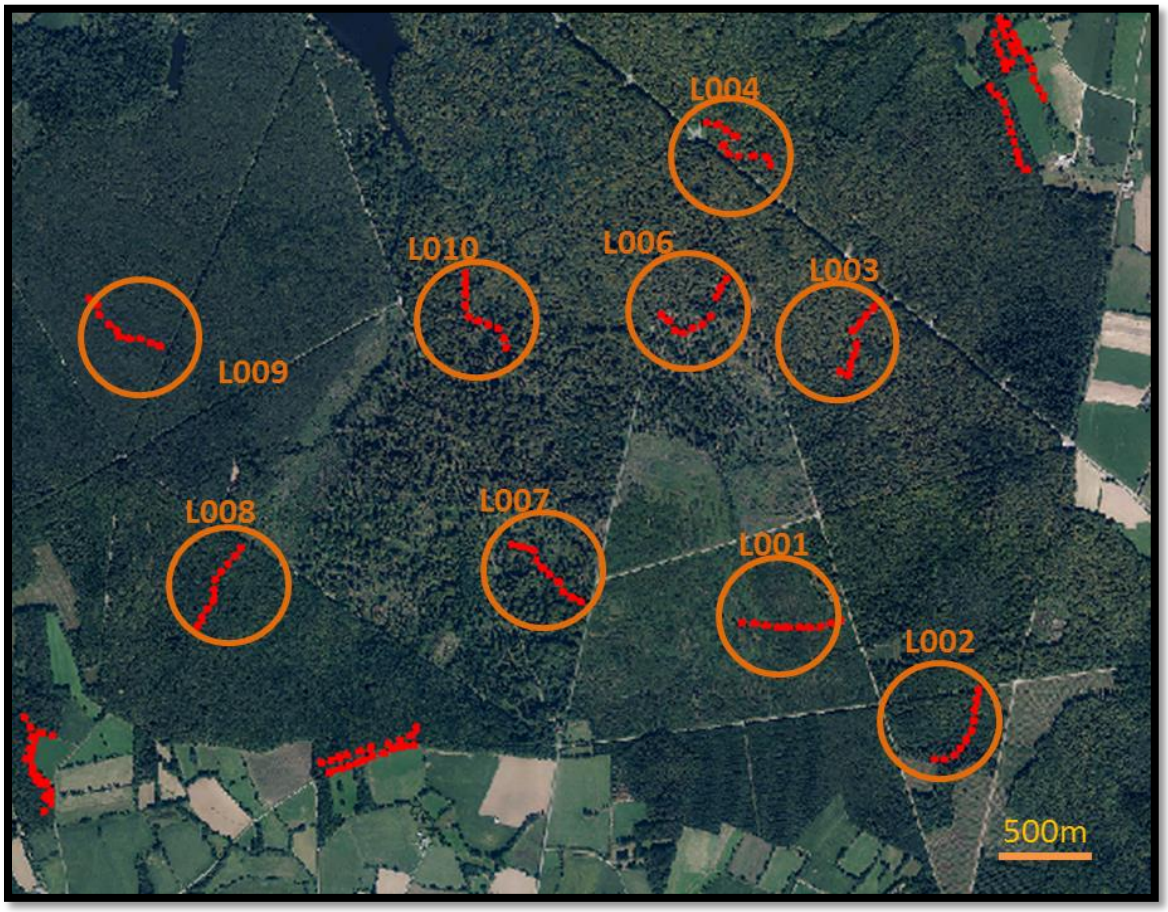


Figure 3.13 : Localisation des neuf lignes de collecte de tiques du secteur cœur de forêt (CF). Chaque point rouge correspond à un tirage de 10m.

Tableau 3.5: Matrice des estimations de *Fst* entre les différentes lignes de collecte du secteur CF représentées par plus de trois individus

	L001	L002	L003	L007	L008
L002	0,017				
L003	0,032	0			
L007	0	0	0,018		
L008	0	0	0	0	
L010	0	0	0	0	0

Le test de Mantel d'isolement par la distance a été effectué en prenant l'ensemble des individus du cœur de forêt (N= 33) et en utilisant la procédure d'isolement par la distance entre individus (Rousset 2000; Watts *et al.* 2007). L'analyse montre une absence d'isolement par la distance au sein du secteur cœur de forêt, le coefficient de la droite de régression étant de 0,195 (R^2= 0,0008, p=0,124) (Figure 3.14).

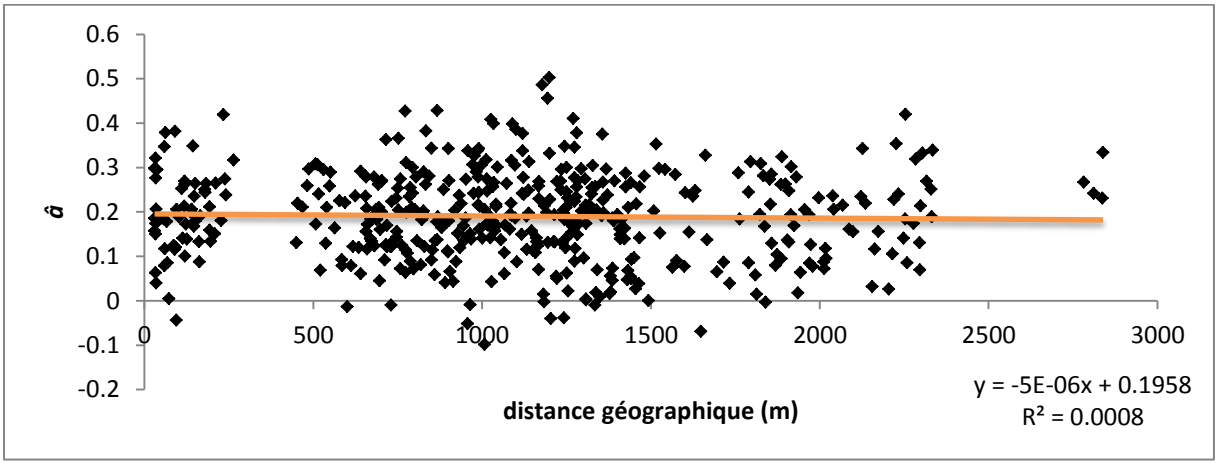


Figure 3.14 : Isolement par la distance entre les différents individus (N=33) du secteur cœur de forêt

b) Cœur de forêt (CF) et Lisière de forêt (LF)

En considérant le cœur de forêt comme secteur source, on peut tester l'hypothèse que plus on s'éloigne du cœur de forêt vers la lisière (aussi bien du côté forêt que prairie), plus la différenciation génétique est importante. Cependant l'estimation de la différenciation entre les trois populations considérées (CF, LF côté forêt et LF côté prairie) ne montre qu'une très faible différenciation entre ces trois populations. Les valeurs de θ les plus fortes sont observées entre les deux populations de LF et la population CF (*Fst* de 0,0029 entre la population CF et la population LF côté forêt et *Fst* de

0,0016 entre la population CF et la population LF côté prairie). Les deux populations de part et d'autre de la lisière de forêt ne montrent aucune différenciation (valeur proche de 0).

Le test de Mantel d'isolement par la distance réalisé avec Genepop sur l'ensemble des individus (N=97) des secteurs CF et LF est non significatif, le coefficient de la droite de régression étant de 0,156, ce qui confirme l'absence d'isolement par la distance entre les différentes lignes (R^2= 0,00005, p= 0,008) (Figure 3.15).

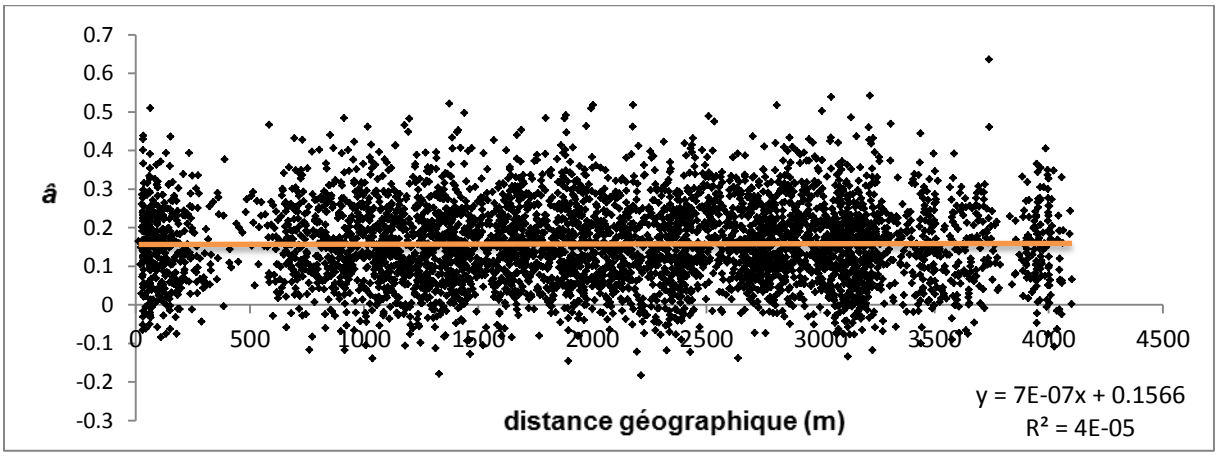


Figure 3.15 : Isolement par la distance entre les différents individus (N=97) du secteur CF et LF

Une analyse en composante principale a également été réalisée avec GenAlEx 6.5 (Peakall & Smouse 2012)et n'a pas permis d'identifier de groupes distincts (Figure 3.16). La contribution relative des trois axes à l'explication de la variation génétique est respectivement de 4,96%, 4,08% et 3,33%, soit un total expliqué par les trois axes de 12,37% de la variation génétique observée. Bien que cette valeur reste assez faible et que les groupes ne soient pas très différents, on peut tout de même observer une moindre dispersion des points correspondant aux individus du secteur CF comparé aux deux autres secteurs dont les points sont plus excentrés dans la représentation graphique (Figure 3.16). Ceci suggère une plus grande hétérogénéité génétique de ces individus qui pourraient s'expliquer notamment par leur plus grande « dispersion géographique ».

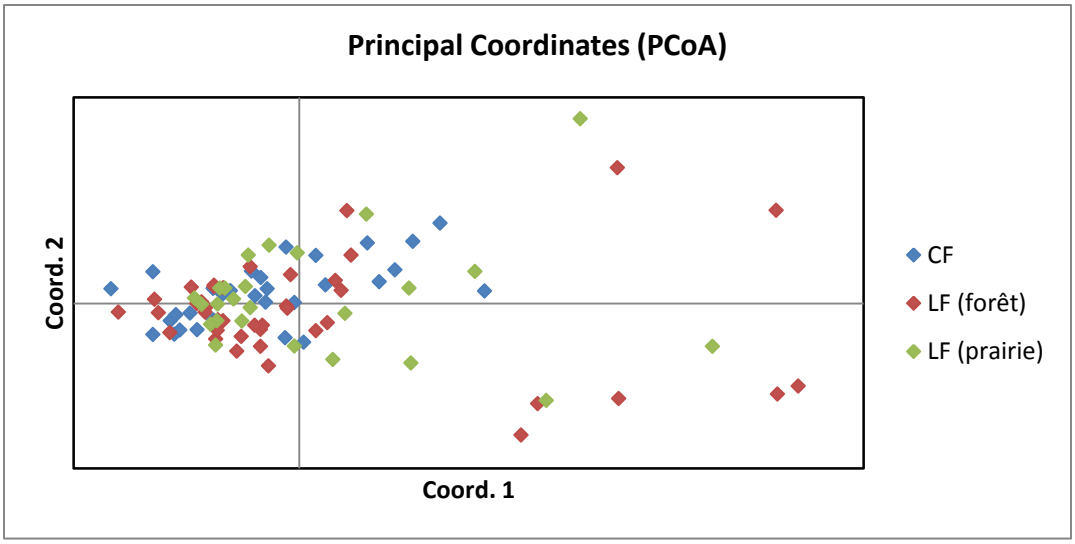


Figure 3.16 : Analyse en composante principale réalisée avec le logiciel GenAlEx 6.5. Chaque point représente un individu selon son lieu d'échantillonnage (bleu en cœur de forêt, rouge et vert en lisière de forêt selon le côté prairie ou forêt)

c) Du cœur de forêt vers les secteurs bocagés, effet de la connectivité paysagère

Afin de tester l'effet de la connectivité du paysage sur la structuration des populations de tiques et leurs interactions, notamment les flux de gènes, on peut s'intéresser au réseau de haies et de boisements de la zone atelier en considérant qu'ils peuvent constituer des corridors facilitant la dispersion des hôtes et des tiques. Sur la représentation schématique de la zone atelier (Figure 3.17), l'ensemble des haies et des boisements influant sur la connectivité des différents habitats est représenté à l'aide d'ArcGis. Le secteur BD, plus proche de la forêt, présente un maillage beaucoup plus dense en haies que le secteur BO, qui est géographiquement plus éloigné de la forêt. De ce fait on peut s'attendre à observer une plus grande différenciation génétique entre les lignes du secteur CF et BO qu'entre les lignes du secteur CF et BD puisqu'elles sont plus connectées et peuvent donc faciliter les mouvements des hôtes dans le paysage. Comme nous l'avons montré précédemment, les estimations de *Fst* entre les secteurs BD et BO par rapport au secteur CF n'ont pas révélé de valeurs significativement différentes de zéro (Tableau 3.4). Cependant une différence plus marquée a été observée entre les secteurs LF et BO (θ=0,006) qu'entre les secteurs LF et BD (θ=0,001).

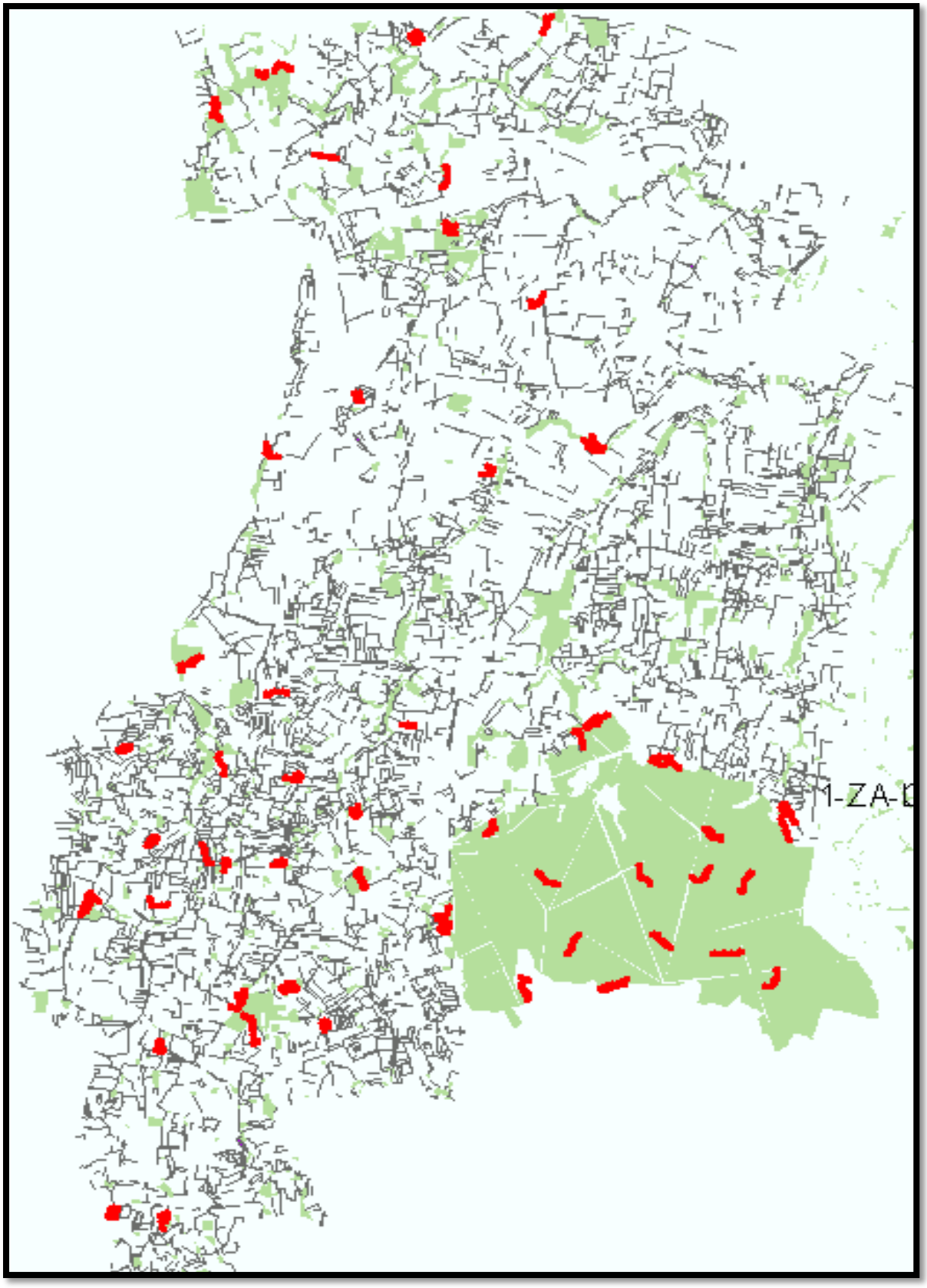


Figure 3.17 : Cartographie schématique de la zone atelier réalisée sous ArcGis représentant la connectivité du paysage de la zone atelier, les lignes de collectes (N=71) dont représentées en rouge, les zones boisées sont représentées en vert, les haies sont représentées en noir.

Un test de Mantel a été réalisé afin de vérifier si la connectivité du paysage plus faible du secteur BO accompagné d'un éloignement géographique plus marqué pouvait induire un isolement par la distance plus marqué entre le secteur CF et BO qu'entre le secteur CF et BD. Pour ceci nous avons considéré d'une part les individus des secteurs CF, LF et BD soit 282 individus et d'autre part les 186 individus des secteurs CF, LF et BO. Aucune différence n'a pu être observée : pour l'analyse concernant le secteur BD (Figure 3.18.a) on obtient une pente de la droite de régression de 0,17 (R^2= 0,005, p= 0,004) alors que pour le secteur BO on obtient une pente de la droite de régression de 0,18 (R^2= 0,004, p= 0,009) (Figure 3.18.b).

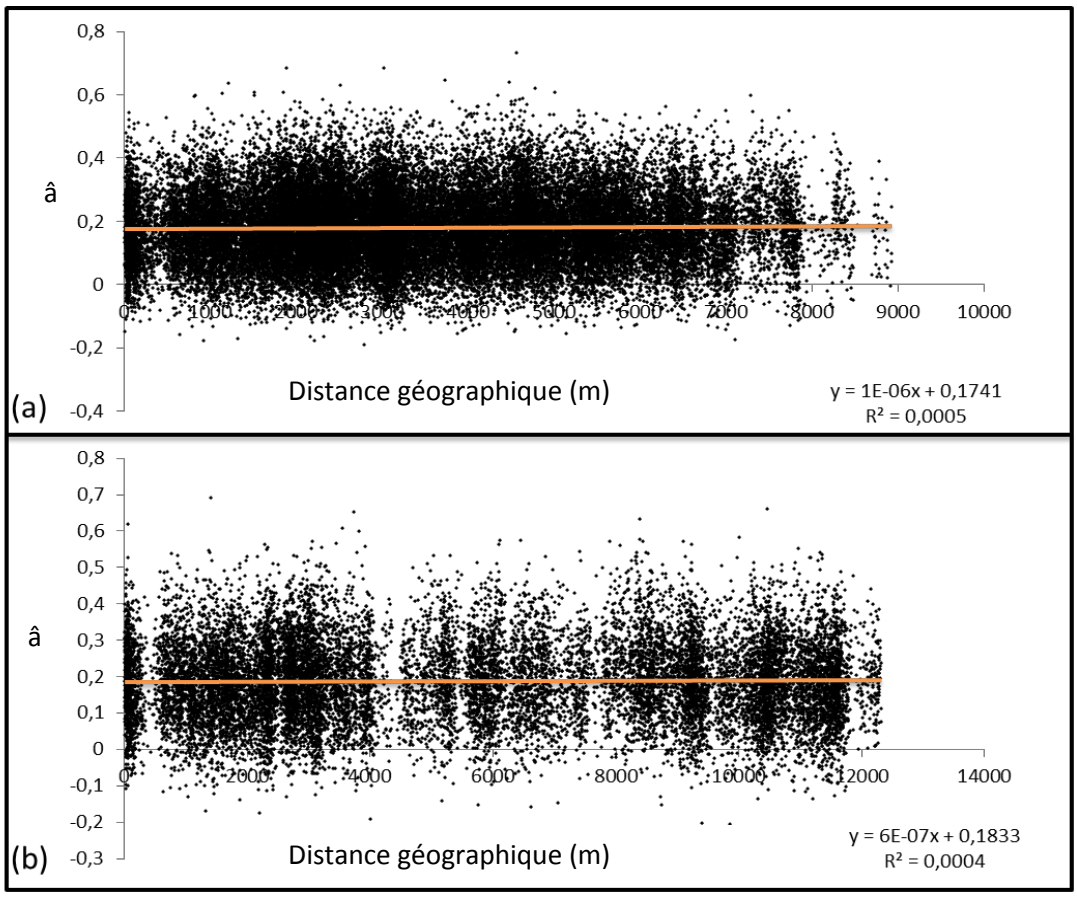


Figure 3.18 : Isolement par la distance réalisé avec Genepop, pour les individus des secteurs cœur de forêt et lisière de forêt et bocage dense (N=282) (a) ou bocage ouvert (N=186)(b).

En partant toujours de l'hypothèse que le secteur de forêt soit le secteur source, on peut donc penser que c'est ce secteur qui est à l'origine des tiques trouvées dans les autres secteurs, influençant le brassage génétique qui conduit à l'absence de structure observée entre les différents secteurs. De ce fait les secteurs BO et BD dans cette hypothèse seraient moins connectés et on

pourrait observer un isolement par la distance entre ces deux secteurs. De plus ce sont les deux secteurs les plus éloignés géographiquement (distance maximale d'environ 15km).

Cependant, l'analyse effectuée en prenant en compte les 274 individus des secteurs BD et BO n'a pas pu montrer d'isolement par la distance, la pente de la droite de régression étant de 0,189 (R^2= 0,002 ; p<0,001), donc très semblable aux valeurs observées précédemment (Figure 3.19).

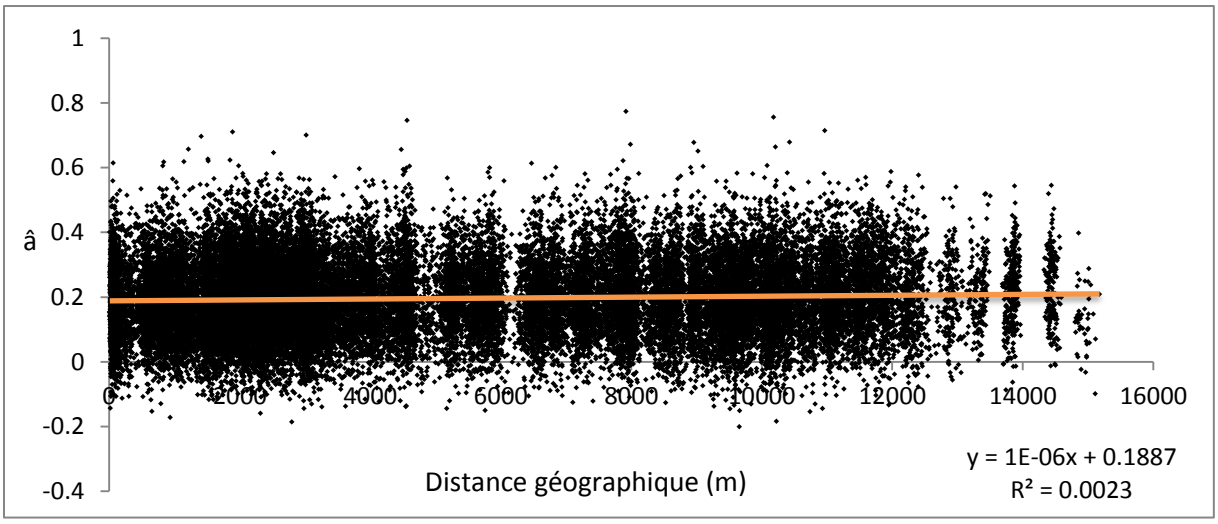


Figure 3.19 : Isolement par la distance réalisée avec Genepop, pour les individus des secteurs BO et BD (N= 274).

4. A l'échelle des différents biotopes identifiés

Au sein des différents secteurs, les collectes de tiques ont été définies dans différents biotopes présentant une végétation plus ou moins hétérogène. Notre échantillonnage n'est pas homogène mais plutôt composite du fait de la structure paysagère de chacun de ces secteurs. De ce fait les lignes de tiques peuvent être regroupées en fonction du milieu dans lequel les échantillonnages ont eu lieu si on fait l'hypothèse que ces milieux peuvent influencer la structure des populations, en traduisant par exemple une adaptation des tiques aux conditions microclimatiques et aux hôtes vivants dans ces milieux. Dans les deux secteurs bocagés (BO et BD), les lignes de collecte se répartissent dans trois milieux distincts : dans un bois, à l'interface prairie/ bois, et à l'interface prairie/haie (Figure 3.20, pour les effectifs des lignes échantillonnées voir Tableau 3.2). Pour le secteur LF, les échantillonnages ont été réalisés à l'interface forêt/prairie soit du côté prairie soit du côté forêt (Figure 3.21, pour les effectifs des lignes échantillonnées voir Tableau 3.2).

Figure 3.20 : Exemple d'échantillonnage dans les différents biotopes constituant le bocage (BD et BO) *NB* : la ligne L040 est bien située à l'extérieur du bois, dans la prairie située à l'ouest du bois mais la zone sombre à gauche des tirets jaunes indiquant la localisation des tirages correspond à l'ombre portée sur le sol liée à la présence des arbres.

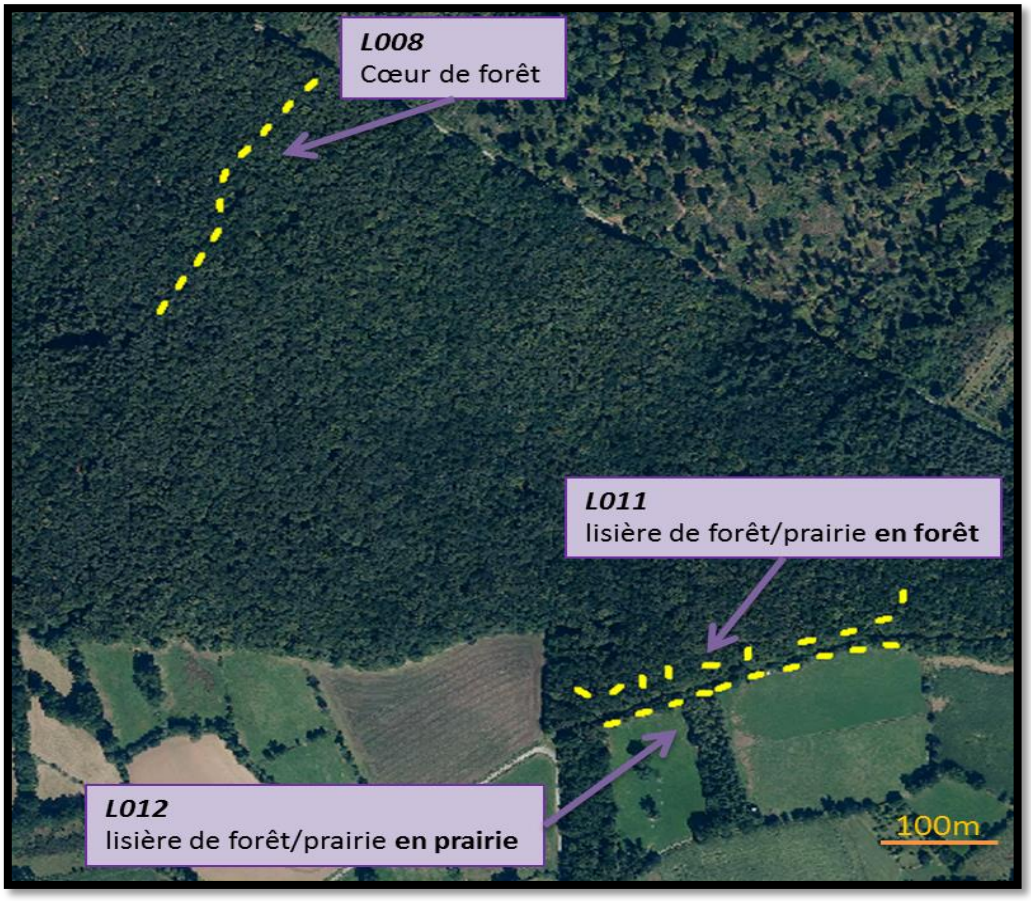


Figure 3.21 : Exemple d'échantillonnage dans les différents biotopes constituant le cœur et la lisière de forêt.

Sur l'ensemble des neufs groupes correspondant aux différents biotopes de chaque secteur (un groupe dans le cœur de forêt, deux groupes en lisière de forêt, trois groupes dans le bocage dense et trois groupes dans le bocage ouvert), les valeurs de *Fis* (Tableau 3.6) apparaissent toujours aussi élevées (entre 0,107 et 0,188) similaire à celles observées à une échelle plus large. De ce fait les différents milieux composants le paysage ne permettent pas de mettre en évidence une sous-structuration.

Tableau 3.6 : Hétérozygotie des groupes de tiques de la zone atelier séparées selon les biotopes; H_{obs} = hétérozygotie observées ; H_{att}= hétérozygotie attendue et non biaisée ; *Fis* = indice de fixation intra-population

Secteur	Biotope	identifiant	Nb lignes	Nb individus	H_{obs}	H_{att}	*Fis*
CF	**CF**	**1**	9	31	0,295	0,349	0,151
LF	**Lisière côté forêt**	**2**	8	41	0,312	0,352	0,107
	Lisière côté prairie	**3**	7	25	0,299	0,351	0,141
BD	**Prairie/Haies**	**4**	10	52	0,294	0,344	0,146
	Prairie/Bois	**5**	9	61	0,294	0,347	0,163
	Bois	**6**	10	72	0,304	0,347	0,132
BO	**Prairie/Haies**	**7**	7	23	0,282	0,344	0,164
	Prairie/Bois	**8**	6	31	0,288	0,357	0,188
	Bois	**9**	5	35	0,299	0,353	0,154

La différenciation génétique globale sur l'ensemble des différents groupes constitués selon le biotope, pris deux à deux, a été estimée par le θ de Weir et Cockerham (1984). De manière générale, une faible différenciation génétique a été observée, les valeurs de θ variant entre -0,004 et 0,009 (Tableau 3.7). On observe tout de même une plus forte différenciation entre le biotope constitué par les lignes du secteur CF (biotope 1) et l'ensemble des autres biotopes, qu'entre ces derniers (valeurs de θ entre 0,002 et 0,009 lorsque le biotope CF est impliqué dans la comparaison). De plus les valeurs de θ les plus fortes (0,009) correspondent à une différenciation génétique moyenne impliquant les biotopes constitués des lignes du cœur de forêt (biotope 1) et de la lisière de forêt côté forêt (biotope 2) d'une part et les lignes du bocage dense de l'interface Prairie-haies (biotope 7) d'autre part.

Tableau 3.7 : Matrice des estimations de *Fst* entre les différents biotopes, les numéros de groupe allant de 1 à 9 correspondent aux identifiants des différents biotopes indiqués dans le tableau 3.2

Groupe	1	2	3	4	5	6	7	8
2	0,003							
3	0,002	-0,004						
4	0,005	0,003	0,006					
5	0,007	0,001	-0,001	0,005				
6	0,007	-0,001	-0,001	0,005	-0,002			
7	0,009	0,009	0,003	0,007	0,004	0,005		
8	0,005	0,004	0,004	0,000	0,002	0,003	0,003	
9	0,006	0,006	0,008	0,002	0,004	0,004	0,001	-0,002

Par ailleurs, une analyse factorielle des correspondances (AFC) a été réalisée pour mettre en évidence d'éventuels clusters différenciés suivant les biotopes (Figure 3.22).

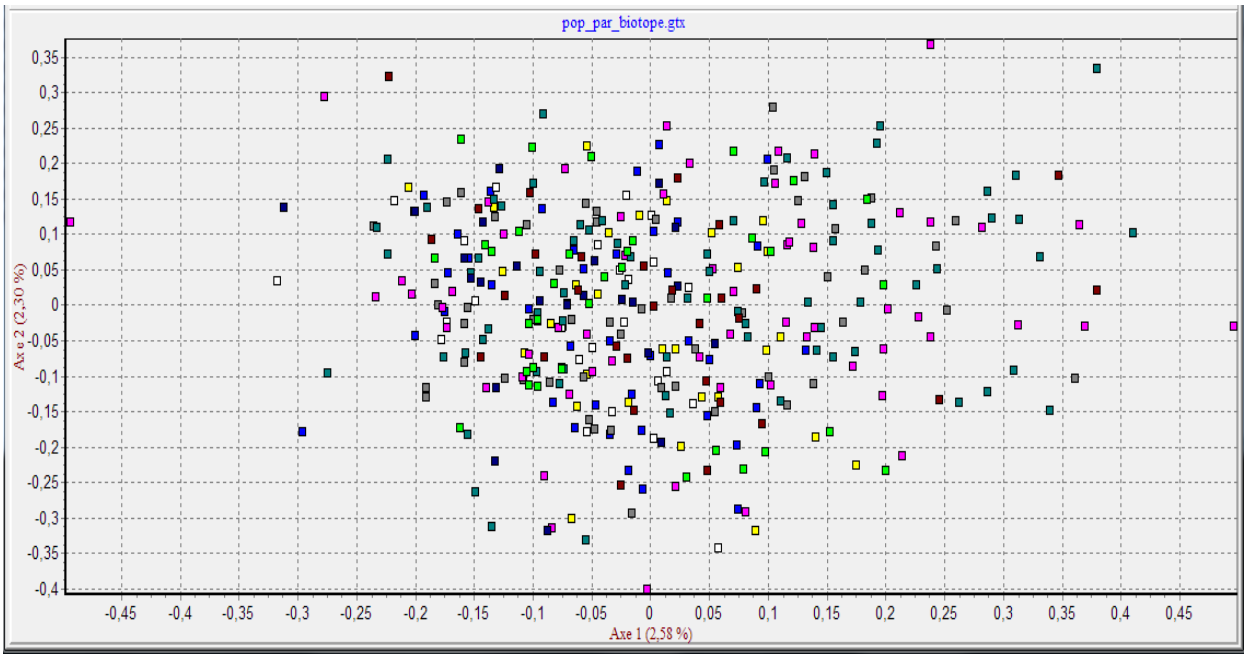


Figure 3.22 : Analyse factorielle des correspondances (AFC), réalisée avec GENETIX, sur les fréquences alléliques des populations des 9 biotopes définis. Chaque couleur représente une population, étant donné le pattern observé, il est inutile de définir dans ce manuscrit la correspondance.

La sous structure en fonction des différents biotopes a également été analysée avec le logiciel STRUCTURE, pour un K définit à 9, correspondant toujours aux neufs biotopes du jeu de données (Figure 3.23). De plus, la recherche du nombre de population K, testé avec les quatre secteurs comme populations, avait identifié comme le plus probable un K de valeur de 9 (K=9, Figure 3.11), ce qui pourrait coïncider avec les différents groupes constituants les différents biotopes.

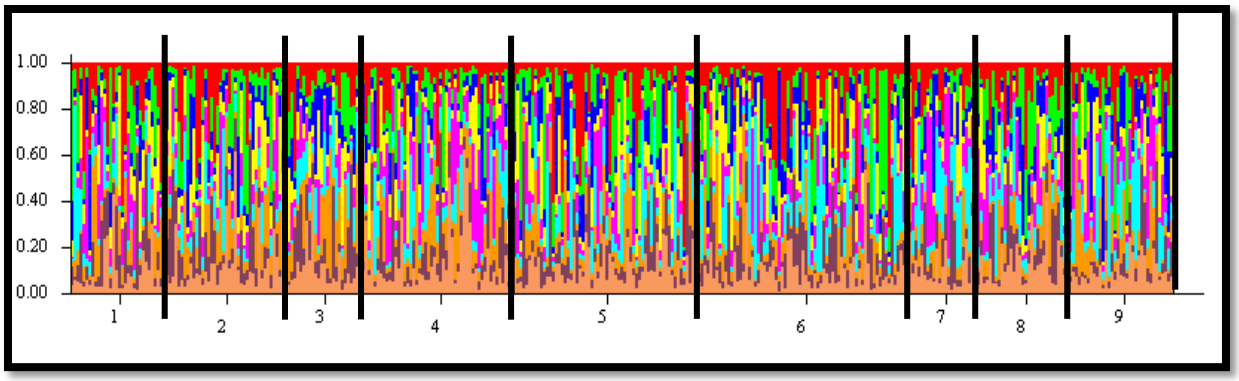


Figure 3.23 : Résultat du logiciel STRUCTURE pour un nombre K de 9 sous-populations, les différents clusters de 1 à 9 correspondent aux identifiants défini dans le tableau 3.6 pour chacun des 9 biotopes.

L'assignation des individus à des groupes définis selon le biotope au sein des différents secteurs par le logiciel STRUCTURE confirme les résultats de l'analyse factorielle des correspondances, à savoir l'absence de structuration génétique et de forts déficits en hétérozygotes.

Cependant, les secteurs couvrent une large zone géographique, la plus grande étant représentée par le 'bocage ouvert' où les sites d'échantillonnages se répartissent sur une zone de 32 km^2. En conséquence, des lignes affiliées à une même population dans cette analyse peuvent être distantes de plusieurs kilomètres (distance maximale de 7,2 kilomètres). De ce fait, pour définir au plus près la diversité génétique des populations d'*I. ricinus* à l'échelle du paysage, il parait pertinent d'analyser des populations collectées sur le terrain dans des environnements semblables (à l'échelle des secteurs) et géographiquement proches. La fréquentation des hôtes et leurs mouvements dans ces différents biotopes pourraient en effet être à l'origine d'une sous-structuration de la diversité génétique sur des courtes distances géographiques.

5. A l'échelle de différents clusters géographiques

Afin de constituer des clusters réunissant les lignes les plus proches géographiquement ('clusters géographiques'), nous avons choisi, au sein d'un même secteur (CF, LF, BD et BO), de regrouper les lignes d'un environnement similaire, distantes au maximum de 1600 mètres et connectées entre elles. Ainsi des lignes en lien avec le même massif boisé seront considérées dans le même cluster géographique, même si ces lignes sont dans le bois ou en lisière du bois, comme c'est le cas pour le cluster 15 par exemple (Figure 3.24). Certains sites, étant trop distants des autres ou n'étant pas connectés par un réseau de haies ou par un bois (dans le bocage) constituent des clusters uniques

(cluster 6, 9 et 21). Les 25 différents clusters géographiques ainsi définis sont présentés dans les figures suivantes : Figure 3.24 pour les 18 clusters de CF, LF et BD et Figure 3.25 pour les sept clusters de BO.

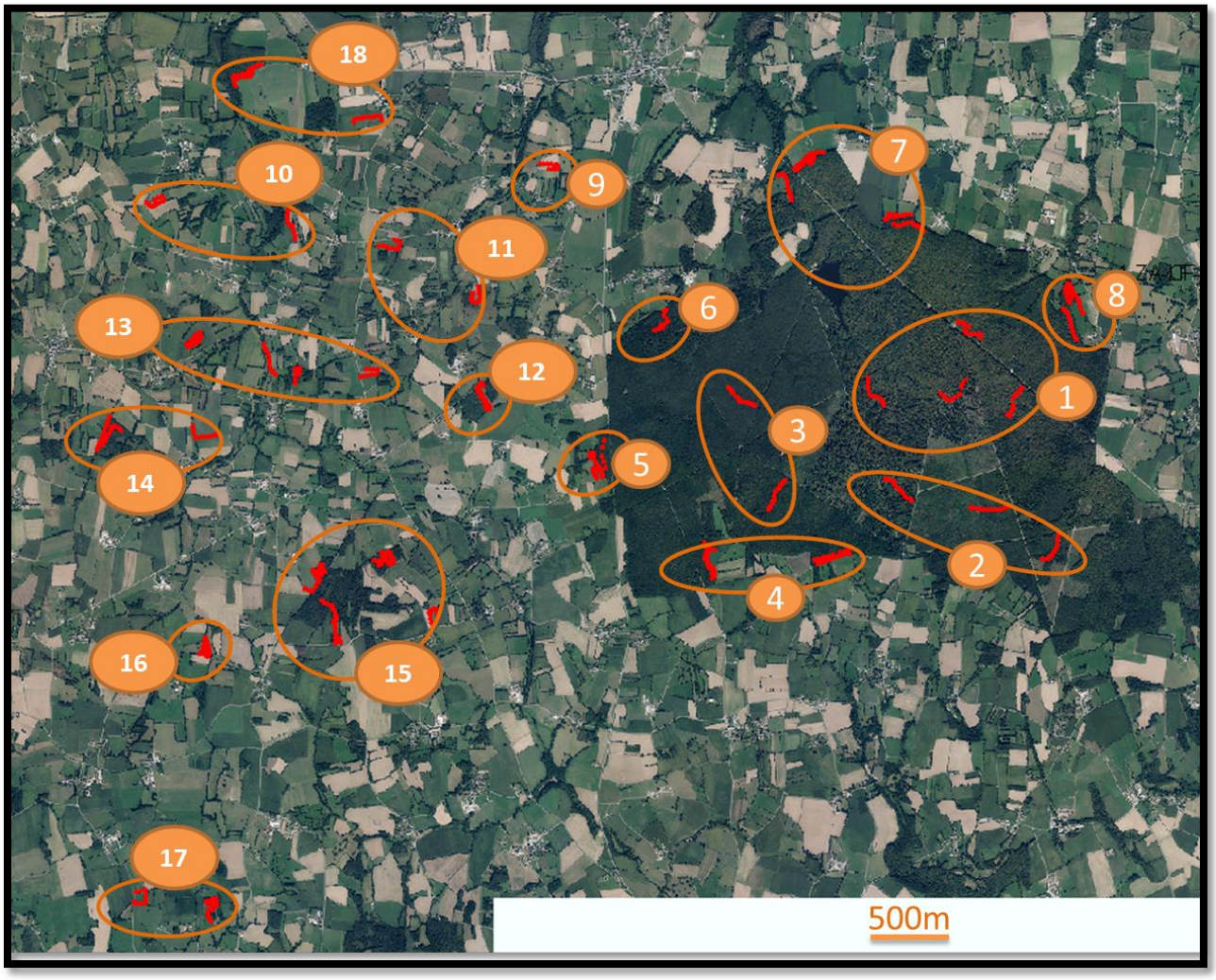


Figure 3.24 : Répartition des 18 différents clusters géographiques dans le bocage dense (BD) et les secteurs forestiers (CF et LF),

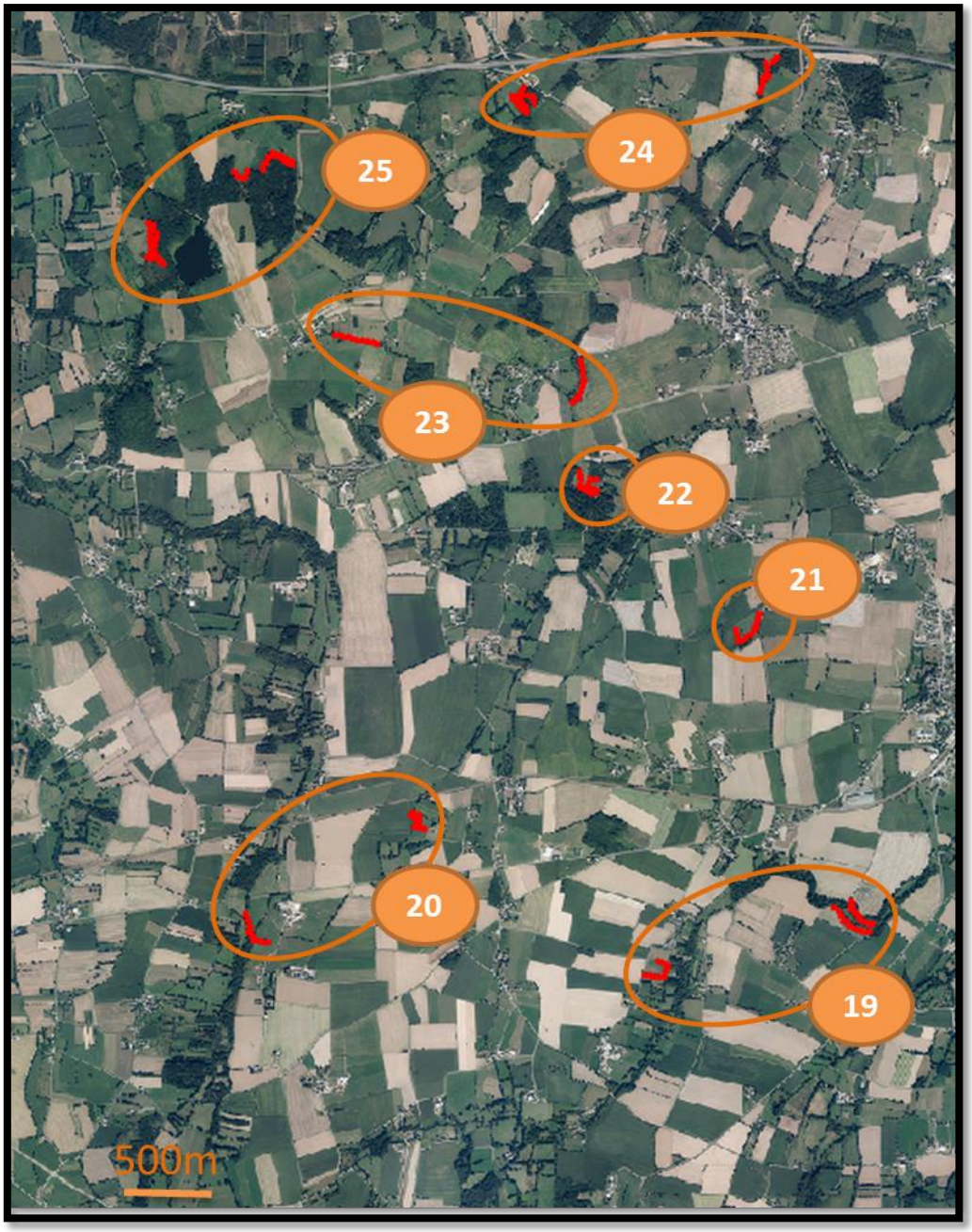


Figure 3.25 : Répartition des 7 différents clusters géographiques dans le bocage ouvert (BO)

Sur l'ensemble des 25 populations définies au sein de ces clusters géographiques, les valeurs de *Fis* apparaissent assez hétérogènes avec des valeurs variant entre 0,072 et 0,250 (Tableau 3.8). Deux des clusters n'étant représentés que par un seul individu (cluster 6 et 21), aucune estimation du *Fis* n'a pu être établi. Parmi les 23 autres clusters, trois ont des valeurs de *Fis* inférieures à 0,1, notamment les clusters 4 et 8 du secteur LF (*Fis*= 0,081 et *Fis*= 0,090 respectivement). Les valeurs de *Fis* observées sont majoritairement inférieures à celle observées précédemment.

Cela suggère que l'effet Walhund a été réduit suite à ce regroupement, et donc ce dernier aurait plutôt une base géographique/spatiale plutôt qu'une origine liée aux milieux. Les tiques collectées sur une ligne proche (distant de 450 mètres au maximum pour le cluster 8) de la lisière de forêt, aussi bien du côté prairie que du côté forêt, présente un moins grand écart à l'équilibre d'HW que l'ensemble des tiques collectées dans le secteur LF prises séparément (Tableau 3.8).

Tableau 3.8 : Hétérozygotie des populations de tiques de la zone atelier séparées selon les clusters géographiques; *Fis* = indice de fixation intra-population

n° cluster	secteur	nb de ligne	nb d'individu	ID ligne	distance maximale	*Fis*
1	CF	4	11	L033-L006-L004-L010	1200	0,187
2	CF	3	16	L002-L001-L007	1400	0,131
3	CF	2	4	L008-L009	850	0,145
4	LF	4	22	L014-L013-L011-L012	1200	0,081
5	LF	2	12	L015-L016	330	0,160
6	LF	1	1	L017	0	-
7	LF	5	20	L022-L023-L024-L026-L025	1300	0,152
8	LF	3	12	L028-L027-L029	450	0,090
9	BD	1	2	L055	0	0,133
10	BD	2	15	L052-L053	1000	0,164
11	BD	2	10	L056-L054	950	0,170
12	BD	2	14	L037-L038	250	0,250
13	BD	5	24	L033-L034-L036-L035-L057	1500	0,129
14	BD	3	25	L058-L039-L040	1000	0,123
15	BD	7	58	L045-L046-L041-L042-L043-L044-L059	1100	0,135
16	BD	2	7	L047-L048	200	0,192
17	BD	3	15	L060-L049-L050	650	0,173
18	BD	3	15	L051-L031-L032	1200	0,072
19	BO	3	14	L079-L080-L089	1400	0,182
20	BO	2	6	L087-L088	1250	0,150
21	BO	1	1	L084	0	-
22	BO	2	13	L077-L078	200	0,121
23	BO	2	11	L082-L083	1500	0,204
24	BO	3	14	L061-L062-L081	1600	0,161
25	BO	4	30	L065-L066-L064-L063	1000	0,149

A l'inverse, certains clusters, comme le cluster 12 qui présente une valeur de *Fis* de 0,250, montrent que la distance géographique entre les lignes ne suffit pas pour expliquer le déficit en hétérozygotes au sein des populations d'*I. ricinus* à l'échelle du paysage. Ce cluster géographique (12) du secteur BD situé à l'interface bois – prairie et constitué de deux lignes de collecte (L037 et L038) éloignées de 250 mètres maximum, montre, par son *Fis* élevé, que d'autres éléments sont à prendre en compte pour expliquer la structuration génétique des tiques. Il apparait nécessaire de réduire l'échelle spatiale d'investigation afin de pouvoir prendre en compte d'éventuelles divergences locales.

La différenciation génétique globale sur l'ensemble des clusters géographiques, pris deux à deux, a été estimée par le θ de Weir et Cockerham (1984). Comme certains clusters sont constitués d'une seule ligne de collecte (clusters 3, 6, 9 et 21), le faible nombre de tiques pourrait biaiser les estimations de θ. De ce fait, ces clusters n'ont pas été pris en compte dans l'analyse. De manière générale, une faible différenciation génétique a été observée, les valeurs de θ variant entre -0,011 et 0,043 (Tableau 3.9).

Tableau 3.9 : Matrice des estimations de *Fst* pour l'ensemble des clusters géographiques représentés par plus de 6 individus. Les numéros identifiant chacun des clusters correspondent aux identifiants attribués dans le tableau 3.8

clusters	**1**	**2**	**4**	**5**	**7**	**8**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**22**	**23**	**24**
2	0																			
4	-0,008	-0,003																		
5	-0,003	0,002	-0,009																	
7	0,004	0,008	-0,005	0,003																
8	0,001	0,004	-0,004	0,003	0,002															
10	-0,005	0,002	-0,007	0,007	0,008	-0,003														
11	0,021	0,006	0,006	0,001	-0,001	0,015	0,011													
12	0,015	0,022	0,005	0,001	0,01	0,008	0,004	0,004												
13	0,014	0,01	0,003	0,012	0,002	0,002	0,005	0,003	0,008											
14	0,011	0,018	0,006	0,009	0	0,004	0,01	0	0,003	0,004										
15	0,003	0,007	-0,003	0,004	0,002	0,003	-0,005	0,005	0,005	0,006	0,006									
16	0,007	0,006	0,003	0,014	0	0,008	0,003	0,022	0,011	0,01	0,029	0,005								
17	-0,001	0,006	0	0	0,005	0,007	-0,004	0,015	0,013	0,01	0,015	0,003	0,019							
18	-0,007	0,005	-0,004	0,004	0	-0,005	-0,003	0,018	0,01	0,003	0,004	0,006	0,017	0,003						
19	0,006	0,01	0,001	-0,001	0,008	0,019	0,006	0,012	0,005	0,019	0,016	0,006	-0,001	0,01	0,016					
20	0,03	0,027	0,02	0,025	0,028	0,028	0,036	0,018	0,014	0,036	0,032	0,021	0,018	0,025	0,04	0,012				
22	0,014	0,015	0,014	0,02	0,01	0,024	0,005	0,014	0,022	0,013	0,009	0,009	0,03	0,007	0,012	0,013	0,043			
23	0,004	0,011	0,006	0,002	0,005	0,021	0,017	-0,011	0,01	0,015	0	0,012	0,039	0,009	0	0,026	0,021	0,022		
24	0,004	0,007	0,01	0,017	0,013	0,013	0,004	-0,006	0,012	0,012	0,009	0,013	0,015	0,004	0,002	0,012	0,019	0,008	0,001	
25	0,01	0,009	0,004	0,004	0,005	0,013	0,005	-0,001	0,004	0,011	0,002	0,003	0,016	0,006	0,01	0,004	0,01	0,021	-0,004	0,008

Aucun pattern de différenciation génétique induit par la clusterisation géographique n'est détecté. En effet les deux clusters géographiques les plus distants (clusters 17 et 24 distant d'environ 15km), présentent un *Fst* de 0,004 alors que les clusters 22 et 23 distant de moins d'un kilomètre présentent un *Fst* de 0,022. De plus la valeur la plus élevée (θ =0,043) correspond à la différenciation génétique observée entre deux clusters relativement proche du secteur BO (cluster 20 et 22). Ainsi une faible différenciation génétique entre les différents clusters est observée, la majorité des estimations de *Fst* ayant des valeurs inférieures ou égales à 0,02 (Figure 3.26).

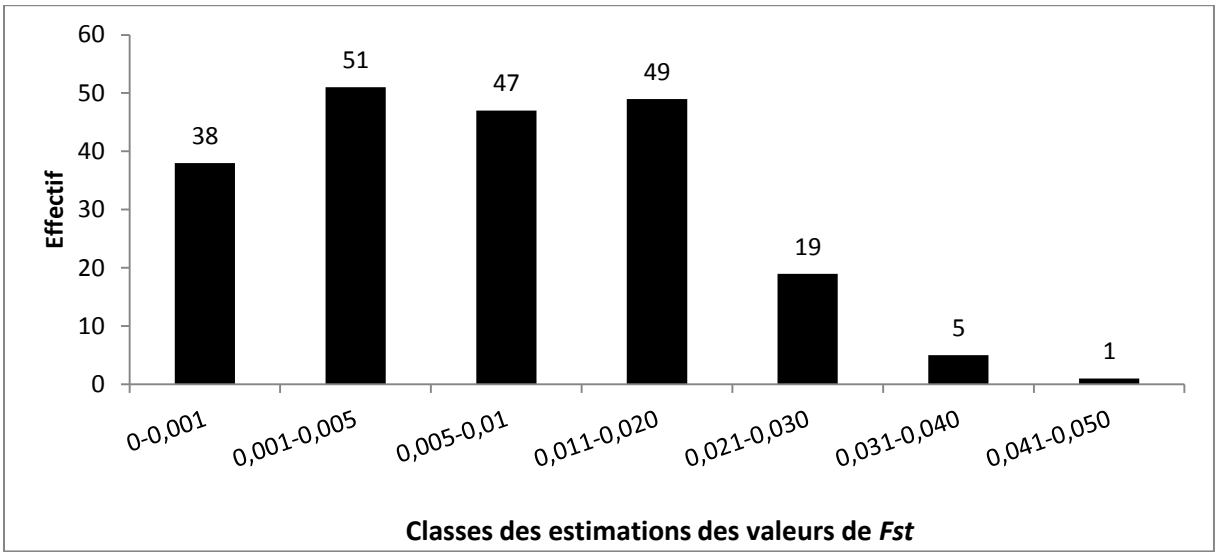


Figure 3.26 : Histogramme représentant l'ensemble des valeurs de θ calculées à l'aide du logiciel Genepop 4.0.2 pour l'ensemble des clusters géographiques représentés par plus de 6 individus. Les valeurs observées sont regroupées en classe comprenant l'ensemble des valeurs supérieures

6. A l'échelle des différentes lignes de collecte

Pour notre échantillonnage, la plus petite échelle que l'on peut considérer est celle de la ligne de collecte. Pour rappel, les tiques ont été collectées sur des transects de 300 m de distance, avec un tirage de 10 m effectué tous les 20 m. Ainsi, une distance de 300 m peut séparer deux nymphes collectées sur la même ligne. Ces lignes ont été effectuées dans des milieux homogènes, où la fréquentation des hôtes doit également être identique entre les différentes zones de tirage. Dans l'échantillonnage final, les tiques se répartissent en 71 lignes de collecte. Cependant, comme nous l'avons vu précédemment, de un à dix individus représentent la même ligne. Nous avons choisi, pour cette analyse, de garder uniquement les lignes comportant plus de six individus. Ce choix réduit le

nombre de lignes analysées à 34. Cette sélection conduit à une représentation biaisée du nombre de lignes considérées pour chaque secteur ou cluster de l'échantillonnage global, mais permet d'obtenir une moins grande variance dans le calcul des différents estimateurs du fait des échantillons de petites tailles.

Comme lors des précédentes analyses à des échelles plus larges, des déficits en hétérozygotes pour la grande majorité des lignes sont observés (Annexe 6). Ainsi, une seule ligne semble être à l'équilibre d'HW (ligne L031 ; *Fis*= -0,03). Pour l'ensemble des autres lignes analysées, le taux d'hétérozygotes observé est systématiquement inférieur à celui attendu sous l'équilibre d'HW. Ceci s'accompagne de fort déficit en hétérozygotes pour l'ensemble des lignes.

L'estimation des *Fis* varie entre -0,03 et 0,27 avec une moyenne de 0,14. Pour les lignes CF, le *Fis* estimé est en moyenne de 0,15 +/- 0,03; pour les lignes LF, le *Fis* moyen est de 0,11 +/-0,02 ; pour les lignes BD, le *Fis* moyen est de 0,13 +/- 0,07 et pour les lignes BO, le *Fis* moyen est de 0,17 +/-0,05 (Figure 3.27 -l'ensemble des résultats est présenté dans l'annexe 6). Ainsi, les valeurs de *Fis* semblent relativement homogènes pour les lignes des secteurs CF et pour celles de LF et plus hétérogènes dans le secteur BD (écart-type de 0,07 pour BD contre 0,02 ; 0,03 et 0,05 pour les autres secteurs).

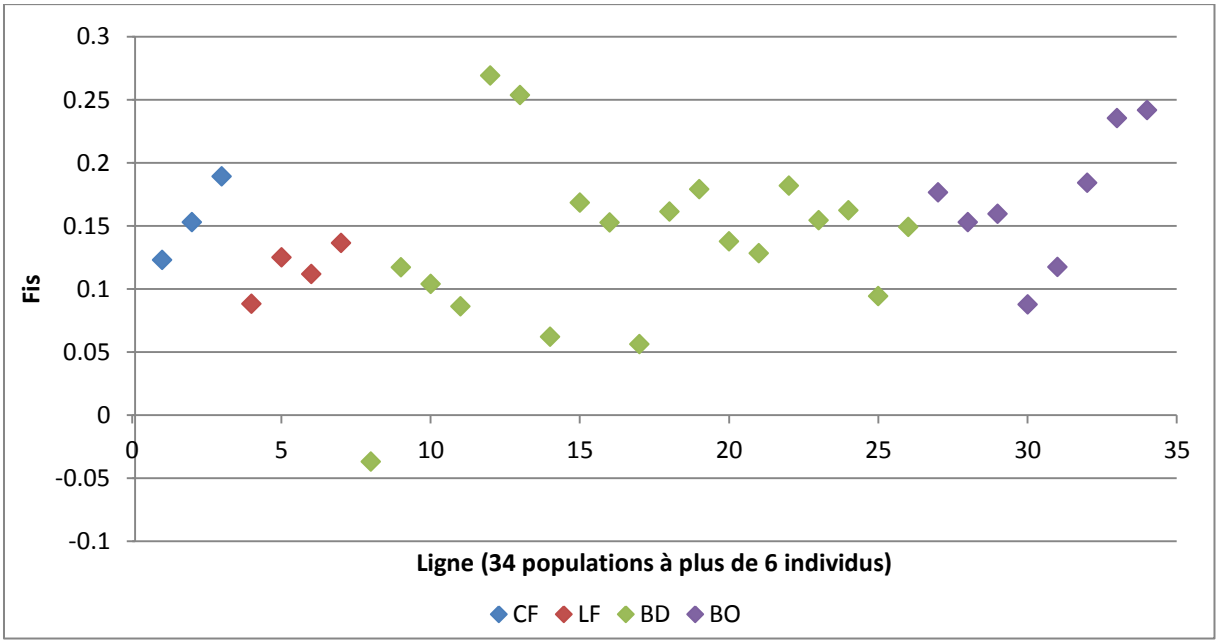


Figure 3.27 : Estimation des *Fis* calculés avec Genepop pour les 34 lignes, comportant plus de six individus

La différenciation génétique globale sur l'ensemble des lignes prises deux à deux a été estimée par le θ de Weir et Cockerham (1984) qui varie entre -0,038 et 0,057 (Annexe 7). Tout comme pour les estimations des *Fis*, seules les lignes représentées par plus de six individus ont été considérées, soit 34 lignes. Comme on peut le voir sur l'histogramme présentant l'ensemble des valeurs en fonction des *Fst,* les valeurs de θ sont globalement extrêmement faibles (médiane à 0,005) (Figure 3.28). Comme nous avons pu le constater dans les précédents regroupements effectués (biotope, clusters géographiques), nous n'observons aucun pattern quant aux valeurs de *Fst* observées en fonction du paysage ou de l'éloignement géographique, des lignes très proches géographiquement pouvant avoir des valeurs de *Fst* très élevées. Par exemple, les deux lignes les plus éloignées, L049 et L061 présentent une différenciation génétique moyenne (θ de 0,033) qui pourrait s'expliquer par l'éloignement géographique et les différents environnements de chacune de ces deux lignes. Cependant des valeurs similaires sont observées entre des lignes extrêmement proches comme les lignes L035 et L036 (θ de 0,017).

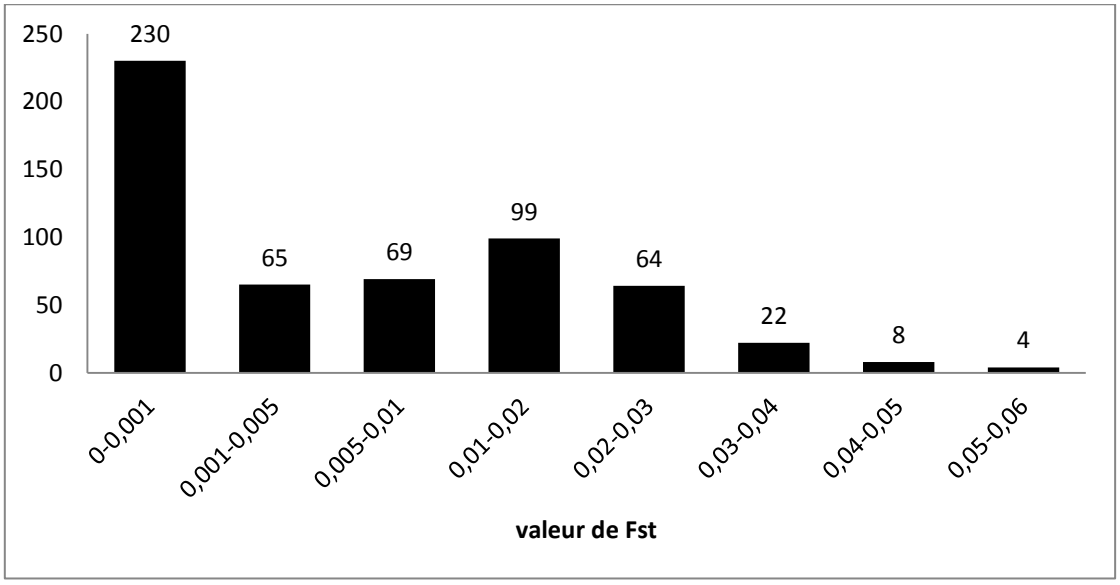


Figure 3.28 : Histogramme représentant les estimations de valeurs de *Fst*, calculées avec Genepop4.2, pour l'ensemble des 34 lignes représentées par plus de 6 individus prises deux à deux.

L'assignation des individus réalisée par une AFC et par la méthode bayésienne implémentée dans le logiciel ne met en évidence aucune différenciation génétique entre les différentes lignes. Les patterns obtenus suite aux analyses étant semblables à ceux obtenus aux échelles des biotopes ou des secteurs, et étant donné le nombre d'entités testées pour cette analyse (N=71), les représentations graphiques ne sont pas présentées dans ce présent manuscrit.

7. Analyse de la partition de la variabilité génétique aux différentes échelles par une AMOVA

Finalement, afin de déterminer la partition de la variabilité génétique aux différentes échelles considérées (secteurs, lignes, individus au sein des lignes), une AMOVA (Analysis of Molecular Variance) a été réalisée avec le logiciel ARLEQUIN (Excoffier *et al.* 1992).

Cette analyse révèle qu'une très faible partie de la variabilité génétique est observée entre les secteurs (0,13 % de la variabilité génétique totale), ou entre les lignes (0,90% de la variabilité génétique totale) et que l'immense majorité de la variabilité s'observe entre les individus appartenant à une même ligne (98,97%) (Tableau 3.10). Ce résultat converge donc bien avec l'ensemble de ceux obtenus avec les autres outils d'analyse de la génétique des populations qui indiquent que nous n'observons pas de structuration de la diversité génétique aux différentes échelles ou regroupements géographiques considérés.

Tableau 3.10 : Analyse Moléculaire de la Variance (AMOVA) à trois niveaux hiérarchiques (secteurs, lignes, individus).

Source de variation	Nombre de degrés de liberté	Somme des carrés	Pourcentage de variation
Entre groupes (entre secteurs)	3	68.116	0.13
Entre populations à l'intérieur des groupes (entre lignes à l'intérieur des secteurs)	67	1265.294	0.90
A l'intérieur des populations (à l'intérieur des lignes)	671	11579.301	98.97
Total	741	12912.712	

8. Structuration génétique liée aux agents pathogènes

Dans le cadre du projet OSCAR, parallèlement au génotypage SNP que nous avons réalisé durant cette thèse, 1222 nymphes collectées sur le terrain ont également été analysées pour le portage de différents agents pathogènes : *Babesia spp.*, *Anaplasma phagocytophilum* et *Borrelia spp.* (Tableau 3.11).

Tableau 3.11 : Résumé des différents taux d'infection en fonction des trois agents pathogènes étudiés dans le jeu de données global du projet OSCAR et dans la subdivision réalisée par le génotypage. Du au très faible effectif, les tiques présentant une coinfection Ba/Bo (N=2) ne seront pas analysées, mais sont présentées ici à titre indicatif.

		Anaplasma phagocytophilum	***Babesia spp.* (Ba)**	***Borrelia spp.* (Bo)**	**Coinfection Ba/Bo**
Jeu de données global OSCAR	Nb de tiques testées	1222	1222	1222	561
	Nb Tiques positives	17	67	56	3
	Nb Tiques négatives	458	1748	1157	507
	Nb de données manquantes	747	7	9	2
	% infection	3,58%	5,51%	4,61%	0,005%
Tiques génotypées	nb de tiques totales	371	371	371	371
	nb de tiques infectées	15	21	14	2
	% infection	4,04%	5,66%	3,77%	0,005%

a) *Anaplasma phagocytophilum*

Afin de tester l'influence de l'infection des tiques par *A. phagocytophilum* sur la structure génétique *d'Ixodes ricinus*, une analyse factorielle des correspondances (AFC) a été réalisée. Pour ceci deux groupes d'individus ont été constitués, un premier composé de tiques présentant un portage de la bactérie *A. phagocytophilum* (N=15) et un second considéré comme groupe 'témoin' ne présentant aucun portage de pathogène. Ce second groupe est également composé de 15 individus, sélectionnés chacun sur les mêmes lignes de collecte ou sur des lignes les plus proches de celles des tiques infectées. Les deux axes de l'AFC expliquent 13,55% (Axe1 : 7,12% ; Axe2 : 6,43%) de la variance de fréquences alléliques entre les deux groupes constitués (Figure 3.29). Ceci explique l'absence de pattern distincts entre les tiques infectées (représentées en bleu) et non-infectées

(représentées en jaune). L'estimation du *Fst* par le calcul de θ est très proche de zéro (-0,0063), ce qui confirme l'absence de différentiation observée.

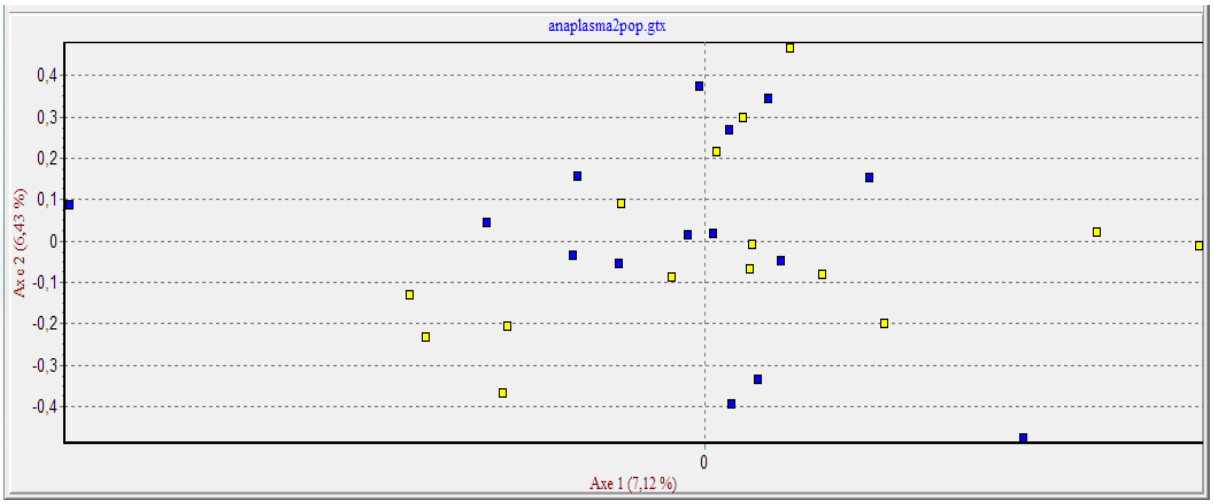


Figure 3.29 : Analyse factorielle des correspondances (AFC), réalisée avec GENETIX, sur les fréquences alléliques des deux groupes d'individus porteurs de la bactérie *A. phagocytophilum* (bleu) et non porteurs (jaune)

L'analyse par AFC a été également réalisée en considérant les individus infectés et témoins provenant des mêmes lignes comme un seul groupe. Aucune différenciation génétique entre l'ensemble des tiques analysées dans cet échantillonnage n'est observée (Figure 3.30).

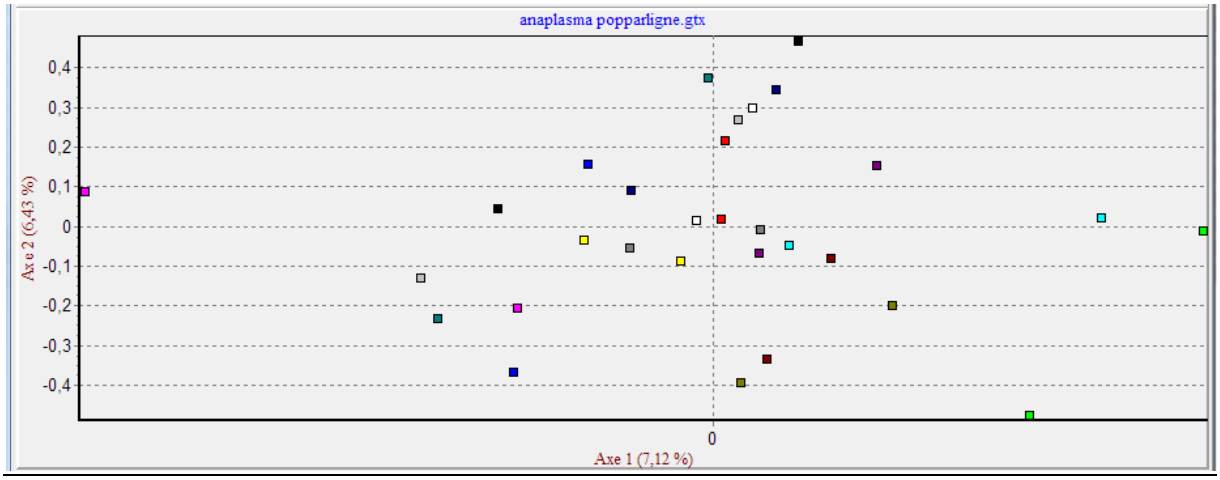


Figure 3.30 : Analyse factorielle des correspondances (AFC), réalisée avec GENETIX, sur les fréquences alléliques des 15 groupes d'individus (porteurs et non porteurs de la bactérie *A. phagocytophilum*, N=2) regroupés selon leurs lignes de collecte, chaque couleur représentant un groupe différent

b) Babesia spp.

Pour les analyses de l'influence de l'infection des tiques par *Babesia spp.* (mais également pour *Borrelia spp.*-présenté par la suite) sur la structure génétique *d'I. ricinus*, le même protocole et analyses ont été effectués que pour *A. phagocytophilum.*

Le groupe 'infecté' est constitué de 21 individus tout comme le groupe 'témoin'. Cependant sur la ligne L006, une seule nymphe a été collectée et s'est révélée positive à *B. divergens*. L'individu génotypé le plus proche géographiquement a alors été pris comme son homologue témoin, soit l'individu L003-T08.
L'analyse factorielle des correspondances réalisée, tout comme pour *A. phagocytophilum*, ne permet pas de distinguer de sous-structuration génétique due au portage par *B. divergens*. Les deux axes de l'AFC expliquent 10,52% (Axe1 : 5,40% ; Axe2 : 5,12%) de la variance de fréquences alléliques entre les deux groupes constitués. L'estimation du *Fst* par le calcul de θ est de 0,0014, ce qui confirme l'absence de différentiation observée.

c) Borrelia spp.

Les analyses de l'influence de l'infection des tiques par *Borrelia spp.* sur la structure génétique *d'I. ricinus* ont suivis le même protocole que pour les deux autres pathogènes précédemment investigués.
Le groupe d'individus 'infectés' est constitué de 14 individus tout comme le groupe 'témoin'. L'analyse factorielle des correspondances ne permet pas de distinguer de sous-structuration génétique due au portage de *Borrelia*. L'ensemble des trois axes de l'AFC expliquent 21,85% (Axe1 : 8,23% ; Axe2 : 7,25 % ; Axe3 : 6,37%) de la variance de fréquences alléliques (Figure 3.31) et l'estimation du *Fst* par le calcul de θ est proche de zéro, (-0,00927). Cependant des clusters composés d'individus porteurs ou non de la bactérie *Borrelia spp.* peuvent être observés (Figure 3.22). Mais mis à part le cluster portant une étoile rouge sur la représentation graphique (Figure 3.22), qui est composé des deux individus prélevés sur la même ligne, les autres clusters identifiables sont composés d'individus qui n'ont aucune relation géographique.

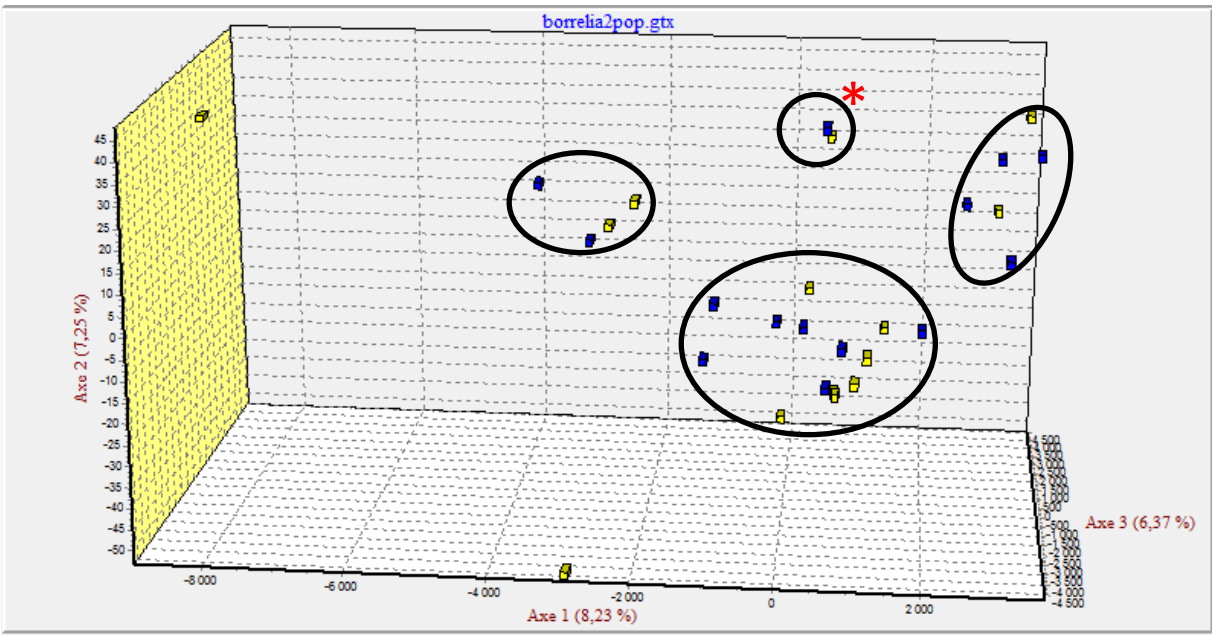


Figure 3.31 : Analyse factorielle des correspondances (AFC), réalisée avec GENETIX, sur les fréquences alléliques des 2 groupes d'individus porteurs (bleu) ou non porteurs (jaune) de la bactérie *Borrelia spp*.

Suite à l'ensemble des analyses réalisées, de manière générale, nous avons observé de faibles différenciations entre les différentes populations étudiées (selon l'échelle à laquelle nous nous placions). Cependant aucune structuration génétique n'a pu être identifiée au sein des populations de tiques dans la zone étudiée quelle que soit l'échelle géographique considérée (de la plus large : ZAA, à la plus fine : la ligne). Aucun lien n'a également pu être mis en évidence entre le statut d'infection et une potentielle influence des pathogènes sur la structuration génétique des populations de tiques.

III. Discussion

La présente étude consistait en une analyse spatiale de la différenciation génétique des tiques *Ixodes ricinus* à l'échelle du paysage. Il s'agissait d'une part de définir le fonctionnement des populations d'*I. ricinus* à cette échelle, et d'autre part d'identifier une éventuelle structuration de ces populations et les éléments du paysage pouvant l'influencer. Pour cela 128 marqueurs SNPs développés dans la seconde partie de ce manuscrit ont été utilisés. De manière générale, une forte consanguinité a été observée entre les différents individus constituant l'échantillonnage. De plus, une absence de structure génétique des populations accompagnées d'une absence d'isolement par la distance ont été identifiées, suggérant une forte dispersion de la population d'*I. ricinus* à l'échelle du paysage. Nous discuterons par la suite de chacun de ces résultats à la lumière de nos connaissances sur la biologie d'*I. ricinus* et des données de la littérature.

A. La consanguinité

Aux différentes échelles étudiées dans cette partie, les valeurs de *Fis* observées présentent des valeurs significativement supérieures à zéro, traduisant un déficit en hétérozygotes. En effet les taux d'hétérozygotes observés sont très majoritairement inférieurs à ceux attendus sous l'équilibre d'HW. Ces déficits observés à l'échelle de la zone atelier d'une surface d'environ 100 km^2 pourraient suggérer un effet Walhund. Cependant même à la plus petite échelle spatiale considérée, la ligne de collecte de 300m, de forts déficits en hétérozygotes ont également été observés. Dans la Figure 3.32, les moyennes des valeurs de *Fis* aux différentes échelles investiguées sont récapitulées. De manière générale, nous pouvons observer qu'en affinant l'échelle d'étude, l'estimation du *Fis* diminue. A l'échelle de la zone atelier, nous obtenons l'estimation la plus forte (*Fis*=0.18), alors qu'en considérant les individus des différents secteurs ou des différentes subdivisions analysées (biotope, clusters géographiques ou lignes de collecte) les estimations des *Fis* varient entre 0,151 et 0,170.

De manière générale, les valeurs estimées de *Fis* sont plus faibles pour la/les population(s) du secteur LF. A l'inverse au sein du secteur BO, la/les population(s) présente(nt) les plus fortes valeurs et de fortes variations de Fis entre les différentes lignes échantillonnées dans ce secteur sont observées. Ceci pourrait s'expliquer par la grande couverture géographique de ce secteur (32km^2) regroupant des environnements très différents.

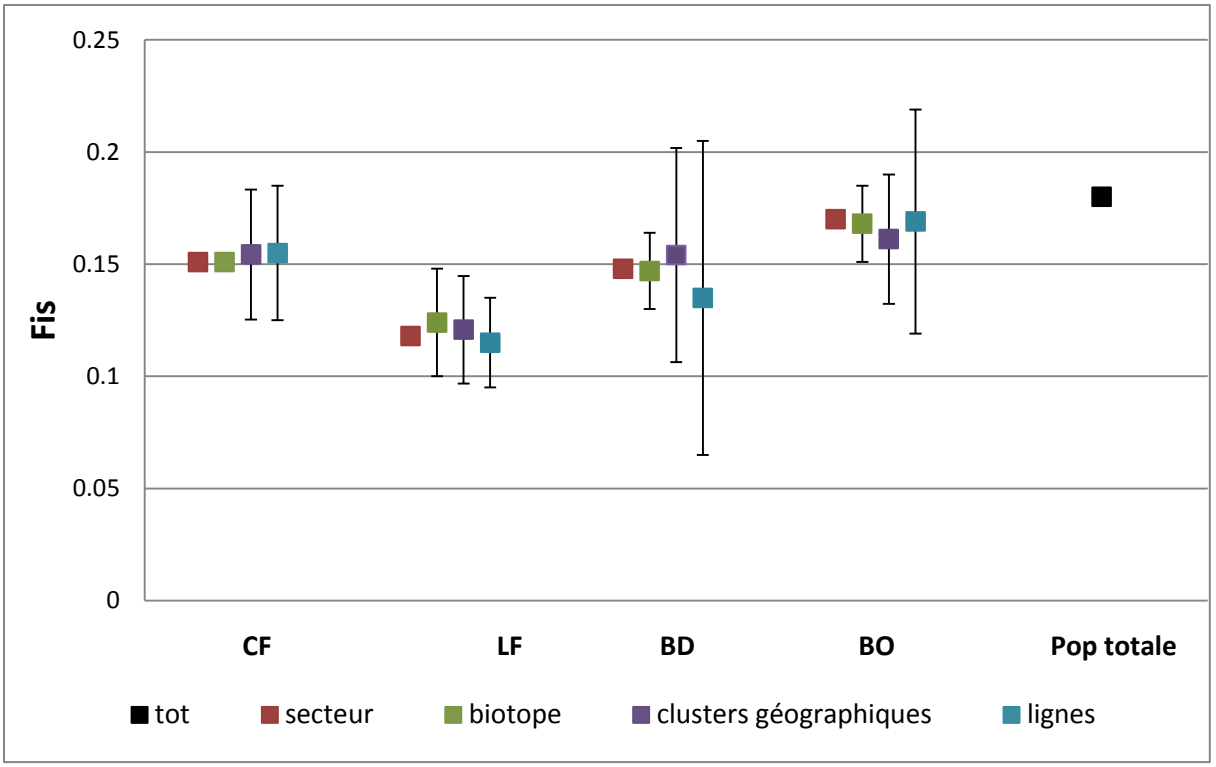


Figure 3.32 : Evolution de l'estimation du *Fis* moyen en fonction de l'échelle investiguée lors de l'analyse. Les barres noires présentent les écart-types des moyennes calculées

Nous observons des valeurs plus faibles de *Fis* que celles qui avaient pu être observées dans d'autres études (*Fis* de l'ordre de 0,15 alors que dans les études précédentes, les *Fis* observés étaient de l'ordre de 0,40). Ceci pourrait être dû au fait que nous avons travaillé ici à une échelle spatiale relativement fine, limitant ainsi l'effet Wahlund observé par d'autres auteurs.

En effet, ces déficits en hétérozygotes avaient déjà été observés dans des études antérieures réalisées à des échelles géographiques beaucoup plus vastes et avec d'autres marqueurs génétiques (De Meeus *et al.* 2002, 2004 ; Kempf *et al.* 2010). De Meeus *et al.* (2002) avaient observé un Fis moyen de 0,4 sur l'ensemble des 24 échantillons suisses analysés dans leurs études. Ces valeurs observées semblaient être dues à un échantillonnage de sous-populations génétiquement différenciées, créant un effet Wahlund. De plus ces études étaient basées sur des marqueurs microsatellites présentant de forts niveaux de déficit en hétérozygotes pour les différents locus analysés (valeurs de *Fis* variant entre 0,13 et 0,62) (De Meeus *et al*. 2002). Une analyse de la ségrégation d'allèles de cinq loci microsatellites dans la descendance de huit femelles avait conclu à différents biais en fonction des loci pour expliquer ces déficits (amplification préférentielle de l'allèle

le plus court, locus lié aux chromosomes sexuels, empreinte maternelle, allèles nuls…) (De Meeus *et al.* 2004).

Bien que les marqueurs SNPs utilisés dans nos analyses aient été sélectionnés selon différents critères, dont la ségrégation mendélienne des allèles, nous observons de fortes valeurs de *Fis* pour certains loci (Annexe 8). Ces importants déficits peuvent être expliqués, comme pour les microsatellites, par la présence d'allèles nuls, ce qu'il faudra tester, mais également par les problèmes d'assignations des allèles dû aux faibles quantités d'ADN disponibles pour le génotypage.

A la lumière des résultats obtenus dans ce manuscrit, nous pouvons conforter l'existence réelle de déficits en hétérozygotes au sein des populations de tiques, dus à la biologie d'*I. ricinus* et non à des artefacts liés aux marqueurs. En effet, nos marqueurs SNPs sont d'une nature différente de ceux utilisés pour mettre en évidence ces déficits jusqu'à présent : ils sont bialléliques, leur polymorphisme est lié à des substitutions ponctuelles et il ne s'agit pas de polymorphisme de longueur.

Malgré ces différences importantes, les résultats de notre étude et des études précédentes, acquis indépendamment, convergent vers la même interprétation.

Plusieurs hypothèses biologiques peuvent également être évoquées afin d'expliquer ces déficits en hétérozygotes : l'existence de races d'hôtes ou d'accouplements préférentiels.

L'existence de races d'hôtes a été abordée dans des études précédentes, notamment menées par Kempf *et al.* (2010, 2011). Bien qu'*I. ricinus* soit considérée comme un parasite généraliste par excellence, avec une vaste gamme d'hôtes potentiels (Sonenshine 1993), les importants déficits en hétérozygotes peuvent être un indicateur d'une sous-structuration locale (de Meeûs *et al.* 2002; Kempf *et al.* 2010) due à un choix non aléatoire des hôtes. Les mécanismes de l'interaction physiologique entre tiques et hôtes sont complexes, ce qui favorise l'émergence d'un polymorphisme dans l'exploitation des hôtes, précurseur de la spécialisation d'hôte. Ceci peut sous-structurer les populations de tiques à une échelle locale. Kempf *et al*. (2011) a montré que des tiques collectées directement sur différents types d'hôtes (oiseaux, micromammifères, lézards, sangliers et chevreuils) dans différents sites européens présentaient une structure génétique significative en fonction du type d'hôtes.

Ces observations ne sont pas uniquement inhérentes à *I. ricinus*. Dans de nombreuses autres investigations chez d'autres tiques, les mêmes observations sont relatées. Chez *I. uriae*, les populations d'*I. uriae* se structurent génétiquement en fonction des différents hôtes sur l'ensemble

de son aire de distribution (McCoy *et al.* 2001, 2005, 2012; Dietrich *et al.* 2012). Chez une autre tique d'oiseaux de mer, *Ornithodoros capensis*, une tique molle, les mêmes patterns de structuration génétique en fonction des espèces-hôtes ont été observés (Gómez-Díaz *et al.* 2012). Chevillon *et al.* (2007) ont également identifié une sous-structuration génétique liée aux races d'hôtes chez la tique *Rhipicephalus microplus* en Nouvelle-Calédonie. Cette tique, présente sur l'ile depuis environ 60 ans, soit l'équivalent d'environ 240 générations, montre une évolution tendant vers la divergence en deux races d'hôtes, une race se spécialisant dans l'exploitation des bovins (hôte originel de *R. microplus*) et une race se spécialisant dans l'exploitation du cerf rusa (*Rusa timorensis*) (De Meeus *et al*. 2010).
A la lumière de ces constats, les forts déficits en hétérozygotes que nous observons pourraient être induits par l'existence de races d'hôte. Différentes méthodes moléculaires récemment développées (Reverse Line Blot par exemple ; Humair *et al.* 2007) permettent d'identifier l'hôte ayant été exploité par les tiques lors du dernier repas sanguin. Il est ainsi prévu que l'ensemble des tiques, ici génotypées, soient analysées également par la méthode du Reverse Line Blot (Humair *et al.* 2007) afin de tester l'hypothèse d'une sous-structuration génétique d'*I. ricinus* dans la Zone Atelier Armorique due aux races d'hôtes.

Une autre hypothèse pour expliquer des déficits en hétérozygotes observés est l'existence d'accouplements préférentiels, notamment entre individus appartenant à la même race d'hôtes. Kempf *et al.* (2009) a en effet montré un patron d'appariement des couples mâles-femelles en fonction de leurs identités génétiques. Ces accouplements préférentiels expliqueraient la consanguinité observée au sein de nos populations et permettraient de maintenir la cohabitation de races d'hôtes (dans le cas où plusieurs types d'hôtes seraient présents à une échelle locale) en évitant les flux de gènes entre elles. Les comportements (et notamment la reconnaissance d'un partenaire sexuel issu de la même race d'hôtes) qui permettraient des accouplements préférentiels ne sont pas connus. L'accouplement pourrait par exemple avoir lieu sur l'hôte, comme le suggère plusieurs études (Graf 1975) favorisant la mise en place de ces races. Dans notre étude, les tiques ont été prélevées sur la végétation. Or les tiques, et notamment le stade larvaire, sont particulièrement agrégées dans le paysage. Ceci est facilement observable lorsqu'on récolte des tiques au drapeau. En effet, les larves sont issues de la ponte d'une femelle (2000 à 3000 œufs) et en raison de leur faible dispersion active, elles se retrouvent distribuées sur une toute petite surface. De ce fait les larves apparentées, issues d'une même ponte, pourraient se gorger sur un même hôte et se décrocher simultanément. Ainsi, elles resteraient agréger au fil de leurs développements dans le paysage, pouvant expliquer l'accouplement préférentiel au sein d'une race.

B. Structure génétique des populations d'*Ixodes ricinus*

L'analyse génétique de 471 tiques collectées dans la Zone Armorique Atelier (ZAA) n'a pas permis d'identifier de réelle différenciation génétique liée à la structure du paysage au sein de l'échantillonnage réalisé. Nous avons pu voir une absence de structuration génétique entre les différents secteurs étudiés et ce, malgré les distances géographiques de plusieurs kilomètres entre les populations comparées et la variabilité des structures paysagères. De manière générale, nous avons pu observer une bonne convergence dans les résultats obtenus des différents outils d'analyse de la génétique des populations utilisés.

Ce travail est le premier à étudier la structure génétique d'*Ixodes ricinus* à une échelle aussi fine. Les études menées jusqu'alors n'avaient été réalisés qu'à des échelles plus larges, ne montrant pas de divergences génétiques significatives (pour des populations suisses, θ= -0,004 : Delaye *et al.* 1997 ; θ entre 0,001 et 0,01 : De Meeus *et al.* 2002). A l'échelle européenne, l'absence de structure génétique observée suggérait un brassage génétique important malgré les distances géographiques importantes.

Cette absence de structure génétique a également été observée à l'échelle du paysage dans notre étude. Bien que nous ayons pu observer une faible différenciation génétique entre les différentes populations étudiées, aucune structure génétique ne ressort. Nous avons pu observer une faible différenciation génétique à l'échelle de la zone atelier entre les différents secteurs étudiés (CF, LF, BD et BO) en observant des valeurs *Fst* variant entre 0,001 et 0,006. Cependant en subdivisant ces différents secteurs en différentes populations (biotopes, lignes de collecte…), les valeurs Fst sont extrêmement variables. Par exemple, en considérant les lignes de collecte représentées par au moins 6 individus dans le secteur BO, les valeurs varient entre -0,008 et 0,048.

Dans le même sens, les analyses d'isolement par la distance réalisée entre les tiques provenant de divers secteurs n'ont pas permis de mettre en évidence une corrélation entre les distances génétiques des tiques et leurs distances géographiques.

De plus, du fait de la forte fécondité des femelles, même un faible taux de dispersion efficace dû aux mouvements des hôtes dans le paysage pourrait induire un brassage génétique suffisant pour empêcher l'action de la dérive génétique et ainsi empêcher d'occasionner une structuration génétique observable. En effet, une femelle peut pondre plusieurs milliers d'œufs. En raison de la grande taille des populations de tiques (lors de collecte au drapeau, on observe jusqu'à 150 nymphes

sur 10m2 et cette collecte est très partielle) et de leur large distribution dans le paysage (forêt, boisements, haies…), la dérive génétique serait beaucoup moins efficace.

L'ensemble de ces résultats et ceux obtenus dans de précédentes études indiquent la présence de flux de gènes important au niveau de la zone atelier. Cette interprétation est aussi confirmée par le test AMOVA qui indique que seule une très petite partie de la variabilité génétique est observée entre secteurs ou entre lignes.

Etant donnée la très faible dispersion active d'*I. ricinus*, les flux de gènes agissant à l'échelle de la zone atelier peuvent être attribués aux déplacements des hôtes. La zone atelier est fréquentée par un grand nombre d'espèces hôtes d'*I. ricinus*, comme les chevreuils, les oiseaux, les micromammifères. Ces hôtes, lors des périodes de repas sanguins des tiques, peuvent agir sur la dispersion de ces dernières et permettent des flux de gènes efficaces entre populations de tiques. De ce fait, même si le paysage apparait fragmenté, les déplacements d'hôtes, comme ceux des chevreuils entre différents habitats boisés de la zone atelier, pourraient contribuer au brassage génétique de leurs ectoparasites.

Bien que la zone atelier présente une importante fragmentation paysagère due à l'anthropisation du milieu, aucune barrière physique majeure à la dispersion, pouvant contraindre les mouvements des hôtes, comme une rivière, la présence d'une autoroute, ou encore des différences d'altitude, n'est observée. Les mouvements des hôtes sont donc peut être trop homogènes pour mettre en évidence une différenciation génétique des populations de tiques. A ce titre, il serait intéressant d'étudier un autre site où de telles barrières à la dispersion des hôtes existeraient.

La fragmentation du paysage de la zone atelier est assez récente par rapport à l'échelle des temps au cours desquels une différenciation génétique aurait pu se mettre en place. L'anthropisation du milieu date de 200 ans environ, ce qui correspondrait pour *I. ricinus* à seulement 60 générations en considérant le temps d'une génération à environ trois ans. De ce fait, bien que le paysage soit fragmenté, la fragmentation pourrait être trop récente pour qu'il y ait eu une dérive génétique des tiques des différents habitats qui leur seraient favorables dans le paysage.

Enfin, une autre hypothèse, testée dans cette étude, est l'influence du portage de différents agents pathogènes (*Borrelia spp., Anaplasma phagocytophilum, Babesia spp.*) sur la structuration génétique des tiques. Bien que des influences de l'infection par *Borrelia bugdorferi* sur la différenciation génétique des tiques aient pu être montrées précédemment par De Meeus *et al.* (2004) (θ = 0.004 versus 0.046 pour les femelles et -0.005 versus 0.14 pour les mâles), nous n'avons pas mis en évidence de différenciation génétique entre les tiques infectées ou non.

Chapitre 4 :

Discussion générale,

Perspectives

&

Conclusions

I. Discussion générale - Perspectives

Malgré l'importance en santé humaine et animale du vecteur *Ixodes ricinus*, de nombreuses lacunes demeurent quant à nos connaissances de la biologie de cette tique, lacunes préjudiciables au développement de nouvelles méthodes de lutte. Ainsi, nous disposons d'une connaissance très limitée de la variabilité génétique à l'intérieure de cette espèce, pourtant distribuée sur une large surface de l'ouest paléoarctique (5000 km d'est en ouest et 3000 km de nord au sud) et composée de milliards d'individus. Nous pouvons supposer qu'une telle taille de population, soumise à des climats et des hôtes relativement variés et sur une telle surface, doit laisser place à la coexistence de nombreux variants. Les travaux présentés dans les chapitres précédents tirent profit des évolutions technologiques spectaculaires ayant eu lieu depuis quelques années en matière de séquençage de l'ADN mais également de génotypage pour faire progresser nos connaissances sur la variabilité génétique de cette espèce et la compréhension des facteurs structurant cette diversité.

A. Développement de SNPs à partir de données génomiques chez *I. ricinus* : intérêts *versus* limites, par rapport à d'autres méthodes et perspectives d'utilisation

1. Etat des lieux sur les données génomiques chez *Ixodes ricinus* / *Ixodes scapularis*

Pour des organismes comme *Ixodes ricinus* ou *I. scapularis,* les investigations génomiques se sont avérées jusqu'à présent compliquées par la taille conséquente de leurs génomes, le nombre important de séquences répétées (entre 50 et 70% du génome chez *I. scapularis* selon les auteurs (Ullmann *et al.* 2005; Van Zee *et al.* 2013)) , et leur fort polymorphisme (Van Zee *et al.* 2013 estiment une fréquence de 1 SNP toutes les 14pb dans le génome d'*I. scapularis).* Ainsi, bien qu'un projet de séquençage du génome complet d'*I. scapularis* ait été initié en 2004 (Hill & Wikel 2005), aucune donnée n'est publiée sur le sujet. Un plan de relance du projet a vu le jour en décembre 2010 (« Tick and mites genome white paper », consultable sur le site Vectorbase.org) sans non plus aboutir à ce jour à la publication du génome d'*I. scapularis.* L'assemblage actuel, d'une couverture estimée

de 4X, présente 369 492 supercontigs (d'une taille N50 de 72Kb) pour une taille totale de 1,76 Gb, soit couvrant les deux tiers du génome. Un projet de développement de SNPs avait été également annoncé depuis quelques années chez *Ixodes scapularis*. A l'heure actuelle, aucune donnée n'a été rendue publique sur ce projet. La seule évolution visible a été un nouveau projet déposé par Catherine Hill, en 2012 sur Vectorbase concernant un projet de séquençage RAD-Seq dans le but d'identifier des SNPs dans le génome d'*I. scapularis* (https://www.vectorbase.org/). Pour *I. ricinus*, il n'existe pas encore d'initiative portant sur le séquençage complet du génome de cette espèce (en dehors du travail d'une doctorante d'un laboratoire d'immunologie Luxembourgeois qui a présenté ses résultats dans un poster présenté au dernier congrès sur la Borreliose de Lyme à Boston en août 2013). Cependant, depuis quelques années, les données génomiques s'accumulent notamment grâce à des projets de séquençage du transcriptome des glandes salivaires d'*I. ricinus* (cf l'article de Chmelar *et al.* (2008) qui était encore basé sur le séquençage par la technique Sanger d'ESTs, c'est-à-dire de cDNA clonés et très récemment, le séquençage par NGS de cDNA paru dans Schwarz *et al.* (2013).

Dans ce contexte en évolution rapide, notre avions choisi de générer notre propre jeu de données génomiques (et non transcriptomique) afin d'obtenir un nombre important de séquences dans des régions non codantes du génome, donc a priori moins variables et aussi appartenant moins à des régions dupliquées du génome (comme c'est le cas de nombreux gènes qui appartiennent à des familles multigéniques). Etant donné les résultats produits, cette stratégie s'est avérée finalement un bon choix.

2. Une perspective accessible à court terme : le développement de SNPs dans le transcriptome d'*I. ricinus*

Différentes stratégies sont possibles afin de réduire la complexité d'un génome. Il est en effet possible de réduire l'ADN génomique, comme nous l'avons fait par RRL, ce qui permet d'obtenir la plus grande partie des SNPs dans de l'ADN non codant, SNPs moins soumis à sélection que des SNPs identifiés dans des régions codantes (la fraction codante étant estimé à 5% du génome d'*I. ricinus*). De ce fait les SNPs que nous avons développés peuvent être considérés comme des marqueurs neutres et permettent de mieux décrire le fonctionnement des populations ou encore étudier la phylogéographie de l'espèce. Il reste encore à identifier dans notre set de SNPs lesquels peuvent être

inclus dans des gènes. Une autre stratégie possible consiste à réduire le jeu de données en ne séquençant que la partie codante, par une approche du type RNA-Seq (Nagalakshmi *et al.* 2008; Wang *et al.* 2009a). Ceci permet de réduire la complexité génomique, en s'affranchissant aussi des séquences répétées présentes dans le non-codant.

Durant ma thèse, j'ai pu avoir accès à un jeu de données 454 (non publié à l'heure actuelle) du transcriptome de glandes salivaires d'*I. ricinus* développé par Xiang Ye Liu lors de sa thèse réalisée dans l'USC INRA Bartonella-Tiques, à l'école vétérinaire de Maisons-Alfort, sous la direction de Sarah Bonnet. Avec Chloé Riou, une stagiaire du M1 Bioinformatique de l'Université de Nantes, que j'ai encadrée pendant trois mois, et en collaboration avec Pierre Peterlongo et Olivier Quenez, nous avons adapté à ce jeu de données transcriptomiques la démarche bioinformatique que j'avais développé pour mon propre jeu de données génomiques afin d'identifier des SNPs. Ce travail initié en avril 2013 a permis d'identifier un grand nombre de nouveau SNPs, tous situés dans des régions codantes du génome d'*I. ricinus*. Ce travail doit être finalisé notamment par l'identification des gènes dont sont issus les reads pour lesquels des SNPs ont été identifiés. L'analyse du polymorphisme de ces marqueurs couplée par exemple à des études sur la variabilité de vection de lignées de tiques permettrait de cibler des gènes d'intérêts dans la lutte anti-vectorielle.

Une autre application intéressante de l'utilisation des marqueurs SNPs serait le développement de carte génétique, qui n'existe pas chez *I. ricinus* jusqu'à présent. La seule carte génétique disponible à ce jour est celle qui concerne *I. scapularis* (Ullmann *et al.* 2003) mais elle est basée sur un nombre limité de marqueurs, principalement des marqueurs de type RAPDs et microsatellites. Ce travail de réalisation de carte génétique chez *I. ricinus* est en cours de développement au laboratoire à partir de l'analyse des familles évoquées dans ce manuscrit.

3. Les RAD-tags : une méthode alternative issue des NGS pour l'isolement de SNPs

Durant ma thèse (démarrée en novembre 2010), les technologies de séquençage ont fortement évoluées. Ainsi en prenant comme exemple la technologie Illumina, en 2010 le GAIIx générait 200000Mb en un run, alors qu'actuellement le HiSeq2000 permet d'en produire quatre fois plus. De plus, le séquenceur HiSeq2500 sera bientôt capable de séquencer un génome entier en seulement un

jour. Ainsi des cibles spécifiques, comme des SNPs, peuvent être recherchés directement par séquençage, on parle alors de « génotypage par séquençage ». Le RAD-Seq (Baird *et al.* 2008) surfe actuellement sur cette vague, avec les quantités de données de plus en plus importantes, les profondeurs de séquençage rendent possible la recherche de SNPs en s'assurant d'éviter la prise en compte d'erreurs de séquençage. Ainsi des études récentes, et de plus en plus nombreuses, recherchent des SNPs directement dans les jeux de données issus du séquençage et les analysent directement (sans étape de génotypage). Le RAD-Seq a d'ailleurs montré son efficacité dans de nombreuses investigations, aussi bien dans des études de génétiques des populations (Hohenlohe *et al.* 2012; Andersen *et al.* 2012), de phylogéographie et phylogénétique (Emerson *et al.* 2010; Rubin *et al.* 2012) que de construction de carte génétique (Baxter *et al.* 2011; Amores *et al.* 2011). Mais cette technique a également permis de générer un grand nombre de SNPs chez diverses espèces, comme le saumon (Houston *et al.* 2012), la truite (Amish *et al.* 2012) ou encore l'artichaut (Scaglione *et al.* 2012). Cependant cette technique, contrairement à celle que nous avons appliquée, ne permet pas de générer de ressources génomiques réutilisables par la suite ; elle est qualifiée de 'one shot'. De plus, cette technique est basée sur une digestion enzymatique de plusieurs échantillons auxquels sont ajoutés des adaptateurs qui contiennent un barcode unique par échantillons. Pour les organismes comme *I. ricinus,* les mêmes problèmes que ceux auxquels nous nous sommes confrontés se posent, et notamment les faibles quantités d'ADN obtenues après extraction. En effet, comme pour les autres techniques utilisant des outils haut-débit, le RAD-Seq demande de grande quantité d'ADN (3µg ; http://floragenex.com/). Cette contrainte pour des organismes de petite taille oblige à amplifier le génome car dans le cas *d'I. ricinus,* 3µg correspond au maximum d'ADN que l'on peut obtenir d'une femelle adulte. De ce fait cette technologie n'est pas très utilisable dans notre cas et le caractère 'one shot' ne permet pas de développer des ressources réutilisables.

Des études récentes ont montré une bonne corrélation entre l'estimation des fréquences alléliques de SNPs estimés par RAD-seq à partir de pool d'individus comparés à des individus marqués individuellement, surtout lorsque la couverture de séquençage est importante (Gautier *et al.* 2013). En revanche, ces approches ne permettent pas d'estimer l'hétérozygotie à l'échelle individuelle et donc de réaliser l'analyse de certains paramètres classiques en génétique de population (cf *Fis*…). Le nombre de loci pour lesquels on obtient les informations sur tous les individus risque aussi d'être réduit en raison de la quantité importante de donnée manquantes générées par cette méthode qui séquence un très grand nombre de fragments répartis sur tout le génome.

Finalement, notre stratégie de développer des amorces spécifiques à chacun des loci étudiés, couplées au génotypage FLUIDIGM permettant de réaliser les génotypages sur un grand nombre de loci par individu, nous a permis de connaitre le génotype (homozygotes ou hétérozygote) de chaque individu et pour chaque locus. Etant donné les objectifs de cette étude de génétique des populations pour lesquels un nombre relativement limité de marqueurs (100 à 150) suffisait, notre stratégie nous a permis d'utiliser au maximum l'ensemble des outils classiques d'analyse de la génétique des populations. En revanche, pour des études cherchant à identifier des marqueurs liés à des gènes particuliers, l'approche RAD-seq apparait plus puissante (et maintenant plus accessibles étant donné le recul qu'on commence à avoir sur cette méthode très récente).

4. Le Whole Genome Amplification (WGA) : un outil original et utile mais nécessitant des études complémentaires

En raison des contraintes du projet OSCAR, qui nous ont conduit à travailler à partir de faibles quantités d'ADN, nous avons eu recours au WGA. Même si cette méthode est assez récente, limitant le recul que l'on peut avoir sur elle, plusieurs études ont déjà permis de valider la recherche et l'identification de SNPs suite à un WGA (Xing *et al.* 2008; Indap *et al.* 2013). Ces études ont montré une bonne reproductibilité de la technique (He *et al.* 2012). Toutefois, elle nécessite d'être mieux validée, étant très dépendante de la complexité génomique des différents organismes. Bien que le recours à ce type d'outils soit utile et parfois nécessaire, de nombreux biais peuvent être introduits. En premier lieu des erreurs faites par la Taq lors de l'amplification, mais également, dans le cas de faibles concentrations d'ADN, l'amplification de fragments préférentiels peut conduire à des biais d'interprétation lors des analyses. Pour ces raisons, il semble préférable de s'affranchir de cette méthode. Une étude récente qui s'est attachée à comparer les techniques d'amplification du WGA et de la PCR-nichée pour l'amplification d'un gène en particulier, ont pu montrer que le WGA était moins sensible et spécifique que la PCR-nichée et que la répétabilité du WGA restait discutable et ne permettait pas la garantie de résultats fiables (Michalska *et al.* 2013). Cependant cette étude n'est pas comparable à la nôtre car une cible en particulier était recherchée, contrairement à nous. Pour notre travail, les tests de validation que nous avons effectués ont également montré une reproductibilité discutable des génotypages SNPs issus de deux pré-amplifications WGA réalisées de manière concomitante.

Etant donné que nos quantités d'ADN de départ étaient faibles (et en dessous du seuil préconisé) nous ne pouvons pas réellement conclure sur les effets réels du WGA et de sa reproductibilité, à savoir si les différences de reproductibilité que nous avons pu observer sont dues à l'utilisation du WGA ou aux faibles quantités d'ADN qui induisent des biais. Pour ceci il serait nécessaire de réaliser l'expérience avec des témoins et en conditions appropriées (quantité d'ADN préconisée) afin de valider cette technique.

De manière générale, les quantités d'ADN très faibles d'*I. ricinus*, soulèvent de réels problèmes, encore plus mis en lumière avec les technologies haut-débit qui nécessite des quantités d'ADN de plus en plus importantes pour les analyses. Dans notre cas, une solution consisterait à ne travailler qu'à partir de tiques adultes dont la quantité d'ADN est supérieure.

B. Les enseignements tirés de l'analyse des génotypages SNPs sur la structure génétique des populations d'*Ixodes ricinus*

Parmi les 384 SNPs développés dans ce projet de thèse, 128 ont été sélectionnés pour étudier la génétique des populations à l'échelle du paysage. Ce travail est le premier à s'intéresser à une échelle aussi fine pour *Ixodes ricinus*

1. La consanguinité chez *I. ricinus*

Les résultats obtenus à l'échelle la plus fine (correspondant à sur 300m^2), montrent des écarts significatifs à l'équilibre d'Hardy-Weinberg. Cependant les déficits en hétérozygotes que nous avons pu observer sont certes présents mais beaucoup moins élevés que ceux qui ont pu être observés dans des études précédentes (de Meeûs *et al.* 2002; Kempf *et al.* 2010). Deux hypothèses sont possibles pour expliquer ceci. Une première hypothèse serait liée à l'échelle géographique investiguée, les études précédentes ayant été réalisées à de plus large échelles que celles étudiées dans le présent travail. La deuxième hypothèse serait en lien avec notre sélection de marqueurs. En effet, en ne retenant que 128 loci, nous avons écartés les SNPs présentant des biais de ségrégation (caractère non-mendélien ou suspicion d'allèle nul), ce qui n'est pas le cas des marqueurs microsatellites utilisés jusqu'à présent. Il faut tout de même noter que du fait des problèmes

techniques rencontrés durant l'analyse des puces du génotypage (stretchs anormaux de points de génotypage correspondant à des hétérozygotes) et du filtre appliqué afin de le résoudre, il est probable que nous ayons écarté des vrais points de génotypage hétérozygotes en les considérant comme des erreurs et de ce fait sous-estimer l'hétérozygotie.

Afin de vérifier ces hypothèses, à savoir, si c'est la résolution spatiale ou la résolution des marqueurs qui explique le mieux nos résultats obtenus, une analyse comparative de notre jeu de données sera réalisée par la suite, les tiques analysées dans cette étude ayant été génotypées par des marqueurs microsatellites en parallèle du génotypage SNP dans le cadre du projet OSCAR.

Les déficits en hétérozygotes peuvent s'expliquer par un effet Walhund, due à la présence de sous-populations au sein de notre échantillonnage. Sur une échelle aussi fine que celle que nous avons étudié ($300m^2$), une structuration génétique due à l'environnement/le paysage ne peut pas être expliquée : le microclimat, la végétation ne paraissent pas être une explication plausible pouvant réduire à une échelle encore plus faible (quelques mètres carrés) les différentes populations. Pour vérifier l'existence d'un effet Wahlund à cette échelle, il serait intéressant de génotyper un plus grand nombre d'individus provenant d'un même tirage ($10m^2$), et de répéter cette opération sur plusieurs tirages d'une même ligne ($300m^2$) afin de voir si une structuration génétique opère à une échelle plus fine. En effet, les nymphes observées sur ces surfaces réduites sont issues de larves qui se sont gorgées à proximité (par exemple sur un micromammifère comme le mulot sylvestre, qui est considéré comme un des hôtes majeurs des larves pour leur gorgement) et ont évolué en nymphe à proximité du lieu où la larve gorgée s'est décrochée. Comme les mouvements de cette espèce de rongeurs sont limités, il est donc tout à fait possible que les nymphes présentes localement soient apparentées (en étant issues de larves elles-mêmes issues d'une même ponte). Pour approfondir cette hypothèse, il serait aussi intéressant de génotyper - toujours à partir de tiques prélevées sur une surface réduite de quelques mètres ou dizaines de mètres carré - des adultes afin de savoir si le niveau de consanguinité est identique ou inférieure aux nymphes (par exemple en raison de la dispersion par d'autres hôtes comme des chevreuils sur lesquels les nymphes auraient pu se gorger).

Une hypothèse alternative à ce déficit en hétérozygotes observé consisterait à considérer l'existence de races d'hôtes, comme l'a montré Kempf *et al.* (2010) ou encore d'accouplements préférentiels (De Meeûs *et al.* 2004; Kempf *et al.* 2010, 2011), ces derniers pouvant d'ailleurs se réaliser sur la base des races d'hôtes. L'analyse des repas sanguins des tiques génotypées apporterait des éléments de réponse particulièrement utiles pour tester cette hypothèse. Malheureusement, ces informations –

prévues pour être acquises dans le cadre du projet OSCAR - n'ont pas pu être obtenues à temps pour être exploitées dans le cadre de ce travail de thèse.

2. Mesurer les flux de gènes à différentes échelles spatiales

Par l'ensemble des résultats que nous avons pu obtenir, nous pouvons en conclure qu'à l'échelle de quelques kilomètres carrés, les flux de gènes sont très importants. L'ensemble des analyses réalisées (*Fst*, isolement par la distance, AMOVA, AFC…), convergent vers cette hypothèse. L'absence de structure observée ne nous a donc pas permis de conclure quant au rôle potentiel du paysage dans la structuration génétique des populations de tiques.

La caractérisation de la dispersion et la délimitation de populations distinctes chez des espèces à large distribution et présentant des populations de grandes tailles constituent un défi pour la génétique des populations, comme c'est le cas pour des organismes modèles pourtant très étudiés comme *Drosophila melanogaster* (Nunes *et al.* 2008) ou *Arabidopsis thaliana* (Platt *et al.* 2010). Dans le cas d'*A. thaliana*, au moins dans sa zone d'origine, l'Eurasie, ces auteurs ont analysés le polymorphisme de 149 SNPs et ils n'ont observé aucun isolement par la distance et aucune rupture correspondant à la notion classique de populations chez cette espèce.

Il serait intéressant maintenant de tester les SNPs développés chez *I.* ricinus sur une large gamme de populations issues de l'ensemble de l'aire de répartition de l'espèce (Tunisie, France, …), afin de rechercher l'existence éventuelle de rupture génétique entre populations. Nous avons abordé brièvement le sujet des populations distantes géographiquement en évoquant les résultats de génotypage d'individus issus de croisement avec des individus Tunisiens qui montrent effectivement une plus grande distance génétique par rapport aux individus français.

Les difficultés pour mettre en évidence les limites de populations sont aussi accrues pour les espèces considérées comme « hyperdiverses » (définies comme présentant des valeurs intra-populationnelles de π [diversité nucléotidique] supérieures à 0.05) comme c'est le cas de certains nématodes ou ascidies par exemple (Cutter *et al.* 2012). *Ixodes ricinus* pourrait appartenir à cette catégorie, étant donné la forte densité de SNPs observée par Van Zee chez *I. scapularis* sur un panel de 10 gènes séquencés par Sanger (Van Zee *et al.* 2013). Cette hypothèse pourrait aussi être testée par l'analyse d'un plus grand nombre de gènes à travers des données NGS comme celle acquises pour le développement des marqueurs SNPs ou de données RNAseq.

Pour estimer les capacités de dispersion et les flux de gènes « contemporains », de nouvelles approches prometteuses viennent d'être développées avec l'analyse de clines génétiques (mis en évidence avec des marqueurs neutres) au niveau de zone de contact (Bermond et al. 2013). Il est ainsi possible de mesurer la pente de ces variations génotypiques régulières et continues pour estimer la dispersion. Cette méthode serait d'ailleurs encore plus précise avec des marqueurs bialléliques comme les SNPs comparé à l'utilisation de marqueurs à plus grand nombre d'allèles comme les microsatellites. Pour être transposé à *I. ricinus*, il reste cependant à identifier des espaces où l'on observerait un tel cline. L'identification de telles zones permettrait aussi d'utiliser des outils de la « landscape genetics » comme je l'ai déjà introduit précédemment. Ainsi certains outils basés sur la théorie des circuits électriques (McRae 2006; McRae & Beier 2007) (McRae 2006, McRae and Beier 2007) ou des LCPA (Least Cost Path Analyses) (Storfer *et al.* 2007; Wang *et al.* 2009c) permettraient de prendre en compte la structure génétique des hôtes afin d'évaluer leurs dispersion et la part que chacun peut jouer.

II. Conclusions

Ce travail constitue le premier développement d'une nouvelle catégorie de marqueurs génétiques, les SNPs (Single Nucleotide Polymorphisms), qui n'avaient jamais été isolés jusqu'à présent chez *Ixodes ricinus*. Pour atteindre cet objectif, étant donné les particularités biologiques de cette espèce (très grand génome, absence de génome de référence, très importante hétérozygotie, petite quantité d'ADN disponible par individu…), de nombreux obstacles ont dû être franchis. L'utilisation d'une méthode de réduction de la complexité génomique (Reduced Representation Libraries), d'un pipeline bioinformatique original développé avec des bioinformaticiens pour s'affranchir de l'absence de génome de référence (DiscoSnp) et le recours à une pré-amplification de l'ensemble du génome (Whole Génome Amplification) pour typer plusieurs centaines de marqueurs par individu tique ont notamment permis de contourner ces difficultés.

Même si de nombreux témoins et contrôles ont été introduits tout au long du processus ayant abouti au génotypage de plus de 500 individus avec 384 SNPs, des expérimentations complémentaires devront être conduites pour valider le bon comportement de l'ensemble des marqueurs (absence

d'allèles nuls…) et leur reproductibilité (notamment pour l'étape de WGA). Cependant, grâce à la sélection qui nous a conduits à ne retenir qu'un tiers des loci initialement isolés, nous sommes confiants sur la qualité des génotypages réalisés.

Dans un deuxième temps, nous avons analysé ces données afin de décrire la structure génétique des populations d'*I. ricinus* à l'échelle du paysage à partir d'un échantillonnage réalisé au printemps 2012 dans 89 sites, répartis dans une zone atelier de 100 km^2. Un déficit en hétérozygotes a été observé chez les nymphes étudiées, à toutes les échelles étudiées, même celle de lignes constitués de 10 individus distribués sur 300 m de long.

Ce résultat confirme les observations réalisées avec d'autres marqueurs (loci microsatellites) même si aucune étude n'était descendue à une échelle aussi fine. Cette consanguinité des nymphes pourrait être dû à une faible dispersion des larves jusqu'au stade nymphe. Cette hypothèse pourra être testée avec des échantillonnages menés à une échelle encore plus fine. Finalement, aucun pattern d'isolement par la distance, ni structuration des populations n'ont été observées à l'intérieur du site atelier, quel que soit les secteurs étudiés (depuis un massif forestier de 1000 hectares d'un seul tenant jusqu'à une zone de bocage ouvert présentant un faible maillage de haies bordant les parcelles agricoles). L'ensemble des analyses menées suggère l'existence d'importants flux de gènes à l'échelle de l'ensemble de la zone géographique étudiée.

En parallèle aux connaissances apportées, ce travail ouvre aussi de nombreuses perspectives pour une meilleure caractérisation de la diversité génétique de cette espèce de vecteur d'importance majeure. Ainsi, il pourra être étendu à l'isolement de marqueurs SNPs liés à des gènes d'intérêts (impliqués dans la vection, la résistance aux acaricides, codant pour des protéines cibles pour développer des vaccins anti-tiques…). Il va aussi permettre l'établissement d'une carte génétique qui constituera un outil de première importance pour l'étude du génome de cette tique et de son évolution.

Références

bibliographiques

Altshuler D, Pollara VJ, Cowles CR *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Amish SJ, Hohenlohe PA, Painter S *et al.* (2012) RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. *Molecular ecology resources*, **12**, 653–660.

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, **188**, 799–808.

Andersen EC, Gerke JP, Shapiro JA *et al.* (2012) Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nature Genetics*, **44**, 285–290.

Anderson JF (1991) Epizootiology of Lyme borreliosis. *Scandinavian journal of infectious diseases. Supplementum*, **77**, 23–34.

Angelone S, Holderegger R (2009) Population genetics suggests effectiveness of habitat connectivity measures for the European tree frog in Switzerland. *Journal of Applied Ecology*, **46**, 879–887.

Apollonio M, Andersen R, Putman R (2010) *European Ungulates and Their Management in the 21st Century*. Cambridge University Press.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, **3**, e3376.

Baxter SW, Davey JW, Johnston JS *et al.* (2011) Linkage Mapping and Comparative Genomics Using Next-Generation RAD Sequencing of a Non-Model Organism. *PLoS ONE*, **6**.

Belkhir K, P Borsa, L Chikhi, N Raufaste & F Bonhomme (2004) GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France).

Bentley DR, Balasubramanian S, Swerdlow HP *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Bermond G, Blin A, Vercken E *et al.* (2013) Estimation of the dispersal of a major pest of maize by cline analysis of a temporary contact zone between two invasive outbreaks. *Molecular ecology*, **22**, 5368–5381.

Bouchard C, Beauchamp G, Leighton PA *et al.* (2013) Does high biodiversity reduce the risk of Lyme disease invasion? *Parasites & vectors*, **6**, 195.

Boyard C, Barnouin J, Gasqui P, Vourc'h G (2007) Local environmental factors characterizing Ixodes ricinus nymph abundance in grazed permanent pastures for cattle. *Parasitology*, **134**, 987–994.

Boyard C, Vourc'h G, Barnouin J (2008) The relationships between Ixodes ricinus and small mammal species at the woodland-pasture interface. *Experimental & applied acarology*, **44**, 61–76.

Brockman W, Alvarez P, Young S *et al.* (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome research*, **18**, 763–770.

Brookes AJ (1999) The essence of SNPs. *Gene*, **234**, 177–186.

Bundock PC, Eliott FG, Ablett G *et al.* (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant biotechnology journal*, **7**, 347–354.

Burel F, Baudry J (2003) *Landscape Ecology: Concepts, Methods, and Applications*. Science Publishers.

Burgman, M.A et Lindenmayer, D.B. *(1998) Conservation Biology for the Australian Environment.* Surrey Beatty and Sons, Chipping Norton, Australie.

Casati S, Bernasconi MV, Gern L, Piffaretti J-C (2008) Assessment of intraspecific mtDNA variability of European Ixodes ricinus sensu stricto (Acari: Ixodidae). *Infection, Genetics and Evolution*, **8**, 152–158.

Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B (1999) The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis*, **20**, 1682–1696.

Chen C, Durand E, Forbes F, FrançOis O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, **7**, 747–756.

Chevillon C, Koffi BB, Barré N *et al.* (2007) Direct and indirect inferences on parasite mating and gene transmission patterns. Pangamy in the cattle tick Rhipicephalus (Boophilus) microplus. *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases*, **7**, 298–304.

Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. In:, pp. 45–56.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, **38**, 1767–1771.

Coipan EC, Jahfari S, Fonville M *et al.* (2013) Spatiotemporal dynamics of emerging pathogens in questing Ixodes ricinus. *Frontiers in cellular and infection microbiology*, **3**, 36.

Comstedt P, Bergström S, Olsen B *et al.* (2006) Migratory passerine birds as reservoirs of Lyme borreliosis in Europe. *Emerging infectious diseases*, **12**, 1087–1095.

Conn HJ, Wolfe GE, Ford M (1940) Taxonomic Relationships of Alcaligenes spp. to Certain Soil Saprophytes and Plant Parasites 1. *Journal of Bacteriology*, **39**, 207–226.

Cumming G s. (1999) Host distributions do not limit the species ranges of most African ticks (Acari: Ixodida). *Bulletin of Entomological Research*, **89**, 303–327.

Cuppen E (2007) Genotyping by Allele-Specific Amplification (KASPar). *CSH protocols*, **2007**, pdb.prot4841.

Cutler DJ, Jensen JD (2010) To Pool, or Not to Pool? *Genetics*, **186**, 41–43.

Cutter AD, Wang G-X, Ai H, Peng Y (2012) Influence of finite-sites mutation, population subdivision and sampling schemes on patterns of nucleotide polymorphism for species with molecular hyperdiversity. *Molecular Ecology*, **21**, 1345–1359.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Debeffe L, Morellet N, Cargnelutti B *et al.* (2012) Condition-dependent natal dispersal in a large herbivore: heavier animals show a greater propensity to disperse and travel further. *The Journal of animal ecology*, **81**, 1327.

Delaye C, Aeschlimann A, Renaud F, Rosenthal B, De Meeûs T (1998) Isolation and characterization of microsatellite markers in the Ixodes ricinus complex (Acari: Ixodidae). *Molecular ecology*, **7**, 360–361.

Delaye C, Béati L, Aeschlimann A, Renaud F, De Meeûs T (1997) Population genetic structure of Ixodes ricinus in Switzerland from allozymic data: No evidence of divergence between nearby sites. *International Journal for Parasitology*, **27**, 769–773.

Dietrich M, Beati L, Elguero E, Boulinier T, McCoy KD (2013) Body size and shape evolution in host races of the tick Ixodes uriae. *Biological Journal of the Linnean Society*, **108**, 323–334.

Dietrich M, Kempf F, Gómez-Díaz E *et al.* (2012) Inter-oceanic variation in patterns of host-associated divergence in a seabird ectoparasite. *Journal of Biogeography*, **39**, 545–555.

Ding C, Jin S (2009) High-throughput methods for SNP genotyping. *Methods in molecular biology (Clifton, N.J.)*, **578**, 245–254.

Eisen L, Eisen RJ, Lane RS (2002) Seasonal activity patterns of Ixodes pacificus nymphs in relation to climatic conditions. *Medical and veterinary entomology*, **16**, 235–244.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196–16200.

Estoup A, Jarne P, Cornuet J-M (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*, **11**, 1591–1604.

Estrada-Peña A, Martinez JM, Sanchez Acedo C, Quilez J, Del Cacho E (2004) Phenology of the tick, Ixodes ricinus, in its southern distribution range (central Spain). *Medical and Veterinary Entomology*, **18**, 387–397.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, **14**, 2611–2620.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research*, **8**, 175–185.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Fagerberg AJ, Fulton RE, Black Iv WC (2001) Microsatellite loci are not abundant in all arthropod genomes: analyses in the hard tick, Ixodes scapularis and the yellow fever mosquito, Aedes aegypti. *Insect Molecular Biology*, **10**, 225–236.

Forman RTT (1995) *Land Mosaics: The Ecology of Landscapes and Regions*. Cambridge University Press.

François O, Ancelet S, Guillot G (2006) Bayesian Clustering Using Hidden Markov Random Fields in Spatial Population. *Genetics*, **174**, 805–816.

Frankham, R *(2006)* Genetics and landscape connectivity*. Dans : Crooks, K.R et Sanjayan, M. Connectivity Conservation. Cambridge University Press, Cambridge, Royaume-Uni.* 72-93.

Fritz CL (2009) Emerging tick-borne diseases. *The Veterinary clinics of North America. Small animal practice*, **39**, 265–278.

Fu Y-B, Peterson GW (2012) Developing genomic resources in two Linum species via 454 pyrosequencing and genomic reduction. *Molecular Ecology Resources*, **12**, 492–500.

Gandon S, Michalakis Y (2002) Local adaptation, evolutionary potential and host–parasite coevolution: interactions between migration, mutation, population size and generation time. *Journal of Evolutionary Biology*, **15**, 451–462.

Gautier M, Foucaud J, Gharbi K *et al.* (2013) Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology*, **22**, 3766–3779.

Gern L & Humair O.F (2002) Ecology of Borrelia burgdorferi sensu lato in Europe, in Lyme borreliosis: biology, epidemiology and control. *Gray J.S., Kahl O., Lane R.S., Stanek G. (eds.), CAB International, Wallingford, Oxon, United Kingdom*, 149-174.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular ecology resources*, **11**, 759–769.

Gómez-Díaz E, González-Solís J (2010) Trophic structure in a seabird host-parasite food web: insights from stable isotope analyses. *PloS one*, **5**, e10454.

Gómez-Díaz E, Morris-Pocock JA, González-Solís J, McCoy KD (2012) Trans-oceanic host dispersal explains high seabird tick diversity on Cape Verde islands. *Biology letters*, **8**, 616–619.

Graf JF (1975) [Ecology and ethology of Ixodes ricinus L. in Switzerland (Ixodoidea: Ixodidae). III: copulation, nutrition and oviposition]. *Acarologia*, **16**, 636–642.

Gray JS (1991) The development and seasonal activity of the tick Ixodes ricinus: a vector of Lyme borreliosis. *Review of Medical and Veterinary Entomology*, **79**, 323–333.

Gray JS (1998) Review The ecology of ticks transmitting Lyme borreliosis. *Experimental & Applied Acarology*, **22**, 249–258.

Gremme G, Steinbiss S, Kurtz S (5555) GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **99**, 1.

Gubler DJ (1998) Resurgent vector-borne diseases as a global health problem. *Emerging Infectious Diseases*, **4**, 442–450.

Guillot G, Mortier F, Estoup A (2005) Geneland: a computer package for landscape genetics. *Molecular Ecology Notes*, **5**, 712–715.

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361–372.

Gut IG (2001) Automation in genotyping of single nucleotide polymorphisms. *Human mutation*, **17**, 475–492.

Harismendy O, Ng PC, Strausberg RL *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome biology*, **10**, R32.

Hartl DL, Clark AG (1997) *Principles of Population Genetics*. Sinauer Associates, Incorporated.

He YJ, Misher AD, Irvin W Jr *et al.* (2012) Assessing the utility of whole genome amplified DNA as a template for DMET Plus array. *Clinical chemistry and laboratory medicine: CCLM / FESCC*, **50**, 1329–1334.

Helyar SJ, Hemmer-Hansen J, Bekkevold D *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, **11**, 123–136.

Hill CA, Wikel SK (2005) The Ixodes scapularis Genome Project: an opportunity for advancing tick research. *Trends in Parasitology*, **21**, 151–153.

Hoch T, Monnet Y, Agoulon A (2010) Influence of host migration between woodland and pasture on the population dynamics of the tick Ixodes ricinus: A modelling approach. *Ecological Modelling*, **221**, 1798–1806.

Hohenlohe PA, Catchen J, Cresko WA (2012) Population genomic analysis of model and nonmodel organisms using sequenced RAD tags. *Methods in molecular biology (Clifton, N.J.)*, **888**, 235–260.

Holderegger R, Wagner HH (2008) Landscape Genetics. *BioScience*, **58**, 199–207.

Holliday R, Grigg GW (1993) DNA methylation and mutation. *Mutation research*, **285**, 61–67.

Hoodless AN, Kurtenbach K, Nuttall PA, Randolph SE (2002) The impact of ticks on pheasant territoriality. *Oikos*, **96**, 245–250.

Horak IG, Camicas J-L, Keirans JE (2002) The Argasidae, Ixodidae and Nuttalliellidae (Acari: Ixodida): a world list of valid tick names. *Experimental & applied acarology*, **28**, 27–54.

Horskins K, Mather PB, Wilson JC (2006) Corridors and connectivity: when use and function do not equate. *Landscape Ecology*, **21**, 641–655.

Houston DD, Elzinga DB, Maughan PJ *et al.* (2012) Single nucleotide polymorphism discovery in cutthroat trout subspecies using genome reduction, barcoding, and 454 pyro-sequencing. *BMC Genomics*, **13**, 724.

Humair P-F, Douet V, Morán Cadenas F *et al.* (2007) Molecular identification of bloodmeal source in Ixodes ricinus ticks using 12S rDNA as a genetic marker. *Journal of medical entomology*, **44**, 869–880.

Hyman ED (1988) A new method of sequencing DNA. *Analytical Biochemistry*, **174**, 423–436.

Hyten DL, Cannon SB, Song Q *et al.* (2010a) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics*, **11**, 38.

Hyten DL, Song Q, Fickus EW *et al.* (2010b) High-throughput SNP discovery and assay development in common bean. *BMC genomics*, **11**, 475.

Hyten DL, Song Q, Fickus EW *et al.* (2010c) High-throughput SNP discovery and assay development in common bean. *BMC Genomics*, **11**, 475.

Indap AR, Cole R, Runge CL, Marth GT, Olivier M (2013) Variant discovery in targeted resequencing using whole genome amplified DNA. *BMC genomics*, **14**, 468.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Iori A, Gabrielli S, Calderini P *et al.* (2010) Tick reservoirs for piroplasms in central and northern Italy. *Veterinary parasitology*, **170**, 291–296.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*, **44**, 226–232.

Jaenson TGT, Hjertqvist M, Bergström T, Lundkvist A (2012a) Why is tick-borne encephalitis increasing? A review of the key factors causing the increasing incidence of human TBE in Sweden. *Parasites & vectors*, **5**, 184.

Jaenson TGT, Hjertqvist M, Lundkvist A (2012b) [2011 peaks the TBE incidence. The deer tribe variation in size and the weather are key factors]. *Läkartidningen*, **109**, 343–346.

Jarem DA, Wilson NR, Delaney S (2009) Structure-dependent DNA damage and repair in a trinucleotide repeat sequence. *Biochemistry*, **48**, 6655–6663.

Jourdren L, Bernard M, Dillies M-A, Le Crom S (2012) Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. *Bioinformatics (Oxford, England)*, **28**, 1542–1543.

Kahl O, Knülle W (1988) Water vapour uptake from subsaturated atmospheres by engorged immature ixodid ticks. *Experimental & applied acarology*, **4**, 73–83.

Keesing F, Belden LK, Daszak P *et al.* (2010) Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature*, **468**, 647–652.

Kempf F, McCoy KD, De Meeûs T (2010) Wahlund effects and sex-biased dispersal in Ixodes ricinus, the European vector of Lyme borreliosis: New tools for old data. *Infection, Genetics and Evolution*, **10**, 989–997.

Kempf F, de Meeûs T, Arnathau C, Degeilh B, McCoy KD (2009) Assortative Pairing in <I>Ixodes ricinus</I> (Acari: Ixodidae), the European Vector of Lyme Borreliosis. *Journal of Medical Entomology*, **46**, 471–474.

Kempf F, De Meeûs T, Vaumourin E *et al.* (2011) Host races in Ixodes ricinus, the European vector of Lyme borreliosis. *Infection, Genetics and Evolution*, **11**, 2043–2048.

Kindlmann P, Aviron S, Burel F (2005) When is landscape matrix important for determining animal fluxes between resource patches? *Ecological Complexity*, **2**, 150–158.

Kiszewski AE, Matuschka F-R, Spielman A (2001) Mating Strategies and Spermiogenesis in Ixodid Ticks. *Annual Review of Entomology*, **46**, 167–182.

Klompen JS, Black WC 4th, Keirans JE, Oliver JH Jr (1996) Evolution of ticks. *Annual review of entomology*, **41**, 141–161.

Kondrashov AS (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Human mutation*, **21**, 12–27.

Krenke BE, Tereba A, Anderson SJ *et al.* (2002) Validation of a 16-locus fluorescent multiplex system. *Journal of forensic sciences*, **47**, 773–785.

Kurtenbach K, Peacey M, Rijpkema SG *et al.* (1998) Differential transmission of the genospecies of Borrelia burgdorferi sensu lato by game birds and small rodents in England. *Applied and environmental microbiology*, **64**, 1169–1174.

Kurtz S, Narechania A, Stein JC, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics*, **9**, 517.

Lander ES, Linton LM, Birren B *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lane RS, Mun J, Stubbs HA (2009) Horizontal and vertical movements of host-seeking Ixodes pacificus (Acari: Ixodidae) nymphs in a hardwood forest. *Journal of vector ecology: journal of the Society for Vector Ecology*, **34**, 252–266.

Léger E, Vourc'h G, Vial L, Chevillon C, McCoy KD (2013) Changing distributions of ticks: causes and consequences. *Experimental & applied acarology*, **59**, 219–244.

Leggett RM, Ramirez-Gonzalez RH, Verweij W *et al.* (2013) Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de bruijn graphs. *PloS one*, **8**, e60058.

Levy S, Sutton G, Ng PC *et al.* (2007) The diploid genome sequence of an individual human. *PLoS biology*, **5**, e254.

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American journal of human genetics*, **44**, 397–401.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature reviews. Genetics*, **4**, 981–994.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in ecology & evolution*, **20**, 136–142.

Manel S, Joost S, Epperson BK *et al.* (2010) Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Molecular Ecology*, **19**, 3760–3772.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer research*, **27**, 209–220.

Maresso K, Broeckel U (2008) Genotyping platforms for mass-throughput genotyping with SNPs, including human genome-wide scans. *Advances in genetics*, **60**, 107–139.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Marshall OJ (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics*, **20**, 2471–2472.

Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, **166**, 351–372.

Matuschka FR, Fischer P, Musgrave K, Richter D, Spielman A (1991) Hosts on which nymphal Ixodes ricinus most abundantly feed. *The American journal of tropical medicine and hygiene*, **44**, 100–107.

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 560–564.

Merriam G (1984) Connectivity: a fundamental ecological characteristic of landscape pattern. *Brandt J, Agger P (eds)roceedings of first international seminar on methodology in landscape ecology research and planning, vol I. Roskilde Universitessforlag GeoRue, Roskilde, Denmark,* 5–15

McCoy KD (2003) Sympatric speciation in parasites – what is sympatry? *Trends in Parasitology*, **19**, 400–404.

McCoy KD, Beis P, Barbosa A *et al.* (2012) Population genetic structure and colonisation of the western Antarctic Peninsula by the seabird tick Ixodes uriae. *Marine Ecology Progress Series*, **459**, 109–120.

Mccoy KD, Boulinier T, Tirard C, Michalakis Y (2001) Host specificity of a generalist parasite: genetic evidence of sympatric host races in the seabird tick Ixodes uriae. *Journal of Evolutionary Biology*, **14**, 395–405.

McCoy KD, Chapuis E, Tirard C *et al.* (2005) Recurrent evolution of host-specialized races in a globally distributed parasite. *Proceedings of the Royal Society B: Biological Sciences*, **272**, 2389–2395.

McLain DK, Li J, Oliver JH (2001) Interspecific and geographical variation in the sequence of rDNA expansion segment D3 of Ixodes ticks (Acari: Ixodidae). *Heredity*, **86**, 234–242.

McRae BH (2006) Isolation by resistance. *Evolution; international journal of organic evolution*, **60**, 1551–1561.

McRae BH, Beier P (2007) Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19885–19890.

Medlock JM, Hansford KM, Bormane A *et al.* (2013) Driving forces for changes in geographical distribution of Ixodes ricinus ticks in Europe. *Parasites & vectors*, **6**, 1.

De Meeûs T, Béati L, Delaye C *et al.* (2002) Sex-biased genetic structure in the vector of Lyme disease, ixodes ricinus. *Evolution*, **56**, 1802–1807.

De Meeûs T, Koffi BB, Barré N, de Garine-Wichatitsky M, Chevillon C (2010) Swift sympatric adaptation of a species of cattle tick to a new deer host in New Caledonia. *Infection, genetics and evolution: journal of molecular epidemiology and evolutionary genetics in infectious diseases*, **10**, 976–983.

De Meeûs T, Lorimier Y, Renaud F (2004) Lyme borreliosis agents and the genetics and sex of their vector, Ixodes ricinus. *Microbes and Infection*, **6**, 299–304.

De Meeûs T, Humair P-F, Grunau C, Delaye C, Renaud F (2004) Non-Mendelian transmission of alleles at microsatellite loci: an example in Ixodes ricinus, the vector of Lyme disease. *International Journal for Parasitology,* 34, 943–950.

Meyer JM, Kurtti TJ, Zee JPV, Hill CA (2010) Genome organization of major tandem repeats in the hard tick, Ixodes scapularis. *Chromosome Research*, **18**, 357–370.

Michalska D, Jaguszewska K, Liss J *et al.* (2013) Comparison of whole genome amplification and nested-PCR methods for preimplantation genetic diagnosis for BRCA1 gene mutation on unfertilized oocytes--a pilot study. *Hereditary cancer in clinical practice*, **11**, 10.

Navajas MJ, Thistlewood HM, Lagnel J, Hugues C (1998) Microsatellite sequences are under-represented in two mite genomes . *Insect Molecular Biology*, **7**, 249–256.

Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315–327.

Morellet N, Bonenfant C, Börger L *et al.* (2013) Seasonality, weather and climate affect home range size in roe deer across a wide latitudinal gradient within Europe. *Journal of Animal Ecology*, **82**, 1326–1339.

Nagalakshmi U, Wang Z, Waern K *et al.* (2008) The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science (New York, N.Y.)*, **320**, 1344–1349.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 5269–5273.

Nickerson JA (1998) Nuclear dreams: the malignant alteration of nuclear architecture. *Journal of cellular biochemistry*, **70**, 172–180.

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.

Nielsen R (2004) Population genetic analysis of ascertained SNP data. *Human genomics*, **1**, 218–224.

Noel V, Léger E, Gómez-Díaz E, Risterucci A-M, McCoy KD (2012) Isolation and characterization of new polymorphic microsatellite markers for the tick Ixodes ricinus (Acari, Ixodidae). *Acarologia*, **52**, 123–128.

Nordström KJV, Albani MC, James GV *et al.* (2013) Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nature Biotechnology*, **31**, 325–330.

Noureddine R, Chauvin A, Plantard O (2011) Lack of genetic structure among Eurasian populations of the tick Ixodes ricinus contrasts with marked divergence from north-African populations. *International Journal for Parasitology*, **41**, 183–192.

Nunes MDS, Neumeier H, Schlötterer C (2008) Contrasting patterns of natural variation in global Drosophila melanogaster populations. *Molecular ecology*, **17**, 4470–4479.

Ogden NH, Lindsay LR, Hanincova K *et al.* (2008) Role of Migratory Birds in Introduction and Range Expansion of Ixodes scapularis Ticks and of Borrelia burgdorferi and Anaplasma phagocytophilum in Canada. *Applied and Environmental Microbiology*, **74**, 1780–1790.

Olsén B, Jaenson TG, Bergström S (1995) Prevalence of Borrelia burgdorferi sensu lato-infected ticks on migrating birds. *Applied and environmental microbiology*, **61**, 3082–3087.

Pannebakker BA, Niehuis O, Hedley A, Gadau J, Shuker DM (2010) The distribution of microsatellites in the Nasonia parasitoid wasp genome. *Insect Molecular Biology*, **19**, 91–98.

Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, **52**, 413–435.

Parola P, Raoult D (2001) Tick-borne bacterial diseases emerging in Europe. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, **7**, 80–83.

Pc W, F. R, Ij S *et al.* (2007) Compatibility of genetic and demographic estimates of “neighbourhood size” in insect populations: analysis of Coenagrion mercuriale (Odonata: Zygoptera) using an improved estimator of genetic divergence. *Mol Ecol*, 737–751.

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research--an update. *Bioinformatics (Oxford, England)*, **28**, 2537–2539.

Pérez-Eid C (2007) *Les tiques: Identification, biologie, importance médicale et vétérinaire*. Tec & Doc Lavoisier.

Peterlongo P, Schnel N, Pisanti N, Sagot M-F, Lacroix V (2010) Identifying SNPs without a reference genome by comparing raw reads. In: *Proceedings of the 17th international conference on String processing and information retrieval* SPIRE’10., pp. 147–158. Springer-Verlag, Berlin, Heidelberg.

Pichon B, Mousson L, Figureau C, Rodhain F, Perez-Eid C (1999) Density of deer in relation to the prevalence of Borrelia burgdorferi s.l. in Ixodes ricinus nymphs in Rambouillet forest, France. *Experimental & applied acarology*, **23**, 267–275.

Platt A, Horton M, Huang YS *et al.* (2010) The scale of population structure in Arabidopsis thaliana. *PLoS genetics*, **6**, e1000843.

Porretta D, Mastrantonio V, Mona S *et al.* (2013) The integration of multiple independent data reveals an unusual response to Pleistocene climatic changes in the hard tick Ixodes ricinus. *Molecular Ecology*, **22**, 1666–1682.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Quillery E, Quenez O, Peterlongo P, Plantard O (2013) Development of genomic resources for the tick Ixodes ricinus: isolation and characterization of Single Nucleotide Polymorphisms. *Molecular ecology resources*.

Ramos AM, Crooijmans RPMA, Affara NA *et al.* (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PloS one*, **4**, e6524.

Randolph SE (2008) Tick-borne encephalitis virus, ticks and humans: short-term and long-term dynamics. *Current opinion in infectious diseases*, **21**, 462–467.

Rizk G, Lavenier D (2010) GASSST: global alignment short sequence search tool. *Bioinformatics*, **26**, 2534–2540.

Røed KH, Hasle G, Midthjell V, Skretting G, Leinaas HP (2006) Identification and characterization of 17 microsatellite primers for the tick, Ixodes ricinus, using enriched genomic libraries. *Molecular Ecology Notes*, **6**, 1165–1167.

Rogic A, Tessier N, Legendre P, Lapointe F-J, Millien V (2013) Genetic structure of the white-footed mouse in the context of the emergence of Lyme disease in southern Québec. *Ecology and evolution*, **3**, 2075–2088.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Analytical Biochemistry*, **242**, 84–89.

Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.

Rousset (2000) Genetic differentiation between individuals. *Journal of Evolutionary Biology*, **13**, 58–62.

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Rubin BER, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. *PloS one*, **7**, e33394.

Ruiz-Fons F, Gilbert L (2010) The role of deer as vehicles to move ticks, Ixodes ricinus, between contrasting habitats. *International journal for parasitology*, **40**, 1013–1020.

Rusk N (2011) Torrents of sequence. *Nature Methods*, **8**, 44–44.

Sanchez CC, Smith TP, Wiedmann RT *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.

Santure AW, Stapley J, Ball AD *et al.* (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular ecology*, **19**, 1439–1451.

Scaglione D, Acquadro A, Portis E *et al.* (2012) RAD tag sequencing as a source of SNP markers in Cynara cardunculus L. *BMC genomics*, **13**, 3.

Schlötterer C (2004) The evolution of molecular markers — just a matter of fashion? *Nature Reviews Genetics*, **5**, 63–69.

Seeb JE, Carvalho G, Hauser L *et al.* (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, **11**, 1–8.

Smouse PE (2010) How many SNPs are enough? *Molecular Ecology*, **19**, 1265–1266.

Sonenshine DE (1993) Biology of ticks. Volume 2. , xvii + 465 pp.

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Storfer A, Murphy MA, Evans JS *et al.* (2007) Putting the "landscape" in landscape genetics. *Heredity*, **98**, 128–142.

Van Tassell CP, Smith TPL, Matukumalli LK *et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247–252.

Taylor GR, Noble JS, Hall JL, Stewart AD, Mueller RF (1989) Hypervariable microsatellite for genetic diagnosis. *Lancet*, **2**, 454.

Taylor, P.D, Fahrig, L, et With, K. (2006) Landscape connectivity: A return to basics. *Crooks, K.R et Sanjayan, M. Connectivity Conservation. Cambridge University Press, Cambridge, Royaume-Uni.* 29-43.

Ullmann AJ, Lima CMR, Guerrero FD, Piesman J, Black WC 4th (2005) Genome size and organization in the blacklegged tick, Ixodes scapularis and the Southern cattle tick, Boophilus microplus. *Insect molecular biology*, **14**, 217–222.

Ullmann AJ, Piesman J, Dolan MC, Iv WCB (2003) A preliminary linkage map of the hard tick, Ixodes scapularis. *Insect Molecular Biology*, **12**, 201–210.

Venkatesan BM, Bashir R (2011) Nanopore sensors for nucleic acid analysis. *Nature nanotechnology*, **6**, 615–624.

Venter JC, Adams MD, Myers EW *et al.* (2001) The sequence of the human genome. *Science (New York, N.Y.)*, **291**, 1304–1351.

Vienne D de (1998) *Les marqueurs moléculaires en génétique et biotechnologies végétales*. Editions Quae.

Vor T, Kiffner C, Hagedorn P, Niedrig M, Rühe F (2010) Tick burden on European roe deer (Capreolus capreolus). *Experimental & applied acarology*, **51**, 405–417.

Wang Z, Gerstein M, Snyder M (2009a) RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, **10**, 57–63.

Wang J, Lin M, Crenshaw A *et al.* (2009b) High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. *BMC Genomics*, **10**, 561.

Wang IJ, Savage WK, Shaffer HB (2009c) Landscape genetics and least-cost path analysis reveal unexpected dispersal routes in the California tiger salamander (Ambystoma californiense). *Molecular ecology*, **18**, 1365–1374.

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular ecology*, **15**, 1419–1439.

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, **38**, 1358.

Whitfield J (2002) Lovelorn pheasants spread ticks. *Nature News*.

Wiedmann RT, Smith TP, Nonneman DJ (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genetics*, **9**, 81.

Wilcove, D.S*.,* McLellan, C.H. *&* Dobson, A.P. (1986) Habitat fragmentation in the temperate zone*. Soule, M.E. (ed.) Conservation Biology. The science of scarcity and diversity:* 237–256*.*

Wright S (1965) The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution*, **19**, 395.

WRIGHT S (1948) On the roles of directed and random changes in gene frequency in the genetics of populations. *Evolution; international journal of organic evolution*, **2**, 279–294.

Xing J, Watkins WS, Zhang Y, Witherspoon DJ, Jorde LB (2008) High Fidelity of Whole-Genome Amplified DNA on High-Density Single Nucleotide Polymorphism Arrays. *Genomics*, **92**, 452–456.

Van Zee J, Black IV WC, Levin M *et al.* (2013) High SNP density in the blacklegged tick, Ixodes scapularis, the principal vector of Lyme disease spirochetes. *Ticks and Tick-borne Diseases*, **4**, 63–71.

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

Annexes

Annexe 1 : description des différentes étapes du séquençage 454

La réalisation du séquençage 454 se déroule en trois étapes majeures, (i) la préparation de la librairie d'ADN à séquencer, (ii) l'étape d'amplification de l'ADN par PCR en émulsion (emPCR), (iii) le séquençage. Ces différentes étapes sont détaillées ci-dessous. Le séquençage est réalisé par le Genome Sequencer FLX system (figure 1).

Figure 1 : Le Genome Sequencer FLX system

i. Préparation de la librairie

L'ADN est fragmenté mécaniquement (fragmentation aléatoire) en utilisant par exemple la nébulisation (cassure physique de l'ADN par haute pression) ou par ultrasons. Afin d'ajouter le couple d'adaptateurs (adaptateur A et adaptateur B) nécessaires à l'étape d'amplification, les fragments d'ADN double brin doivent être réparés pour posséder des extrémités franches compatibles avec les extrémités des adaptateurs. Pour ceci, l'ADN est traité avec le fragment de

Klenow, qui, grâce à son activité exonucléase et polymérase, permet d'obtenir les extrémités franches assurant la ligature avec les adaptateurs.

Les adaptateurs, préalablement dé-phosphorylées sur leurs extrémités 5' (afin d'éviter l'association de plusieurs adaptateurs) sont ensuite ajoutés à l'ADN en présence de ligase (figure 2). La ligation n'étant pas orientée, les fragments d'ADN peuvent être bornés par deux adaptateurs A, deux adaptateurs B ou un de chaque. Seuls les fragments comportant l'adaptateur A et B pourront être amplifiés et par la suite séquencés. Les fragments sont ensuite purifiés. La première étape de purification se fait par des billes sur lesquelles se trouve de la streptavidine ; l'adaptateur B étant biotinylé sur l'un de ses brins, les fragments d'ADN portant un adaptateur B s'accrocheront, les adaptateurs comportant deux adaptateurs A seront, quant à eux, élués. La dernière étape de purification est une dénaturation ; seuls les fragments simple brin ayant un adaptateur A et un B pourront se décrocher des billes. En effet, les deux brins d'ADN des molécules possédant deux adaptateurs B restent piégés par les billes.

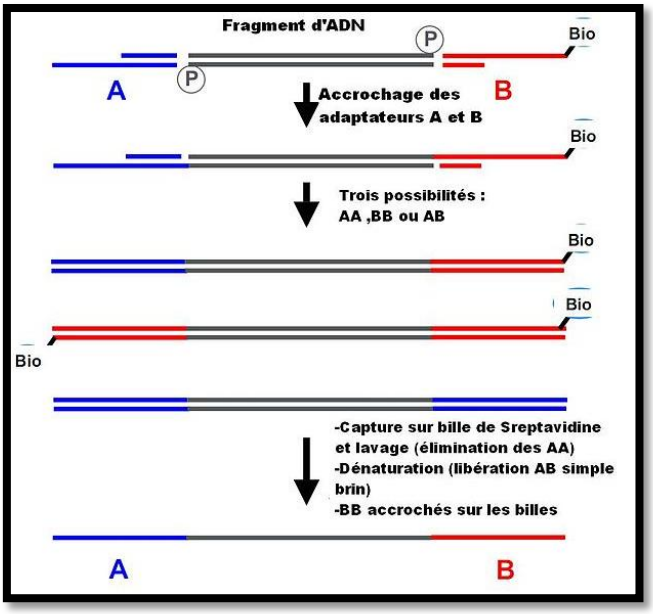


Figure 2: Préparation de fragments pour le séquençage

ii. PCR en émulsion

Les fragments d'ADN purifiés sont mis dans une solution avec des microbilles sur lesquelles sont fixées des amorces A ou B complémentaires des séquences des adaptateurs A et B. Les microbilles

étant en en large excès par rapport aux molécules d'ADN, un seul fragment d'ADN se fixe en théorie sur chaque microbille.

Les microbilles sont ensuite mises dans une solution lipidique dont les gouttelettes vont jouer le rôle de microréacteur au sein duquel la réaction de PCR va se réaliser (figure 3). Chaque microgoutte d'huile renferme alors une microbille d'agarose contenant un fragment d'ADN, séparée des autres billes en phase aqueuse. Chacune de ces gouttelettes contenant les composés nécessaires à la réaction de PCR, chacune d'elles devient un microréacteur dans lequel une amplification a lieu (plusieurs millions de réactions simultanées). Cette étape d'enrichissement des microbilles en séquences d'ADN permettra d'obtenir un signal de fluorescence suffisant pour être détecté par la caméra lors du séquençage.

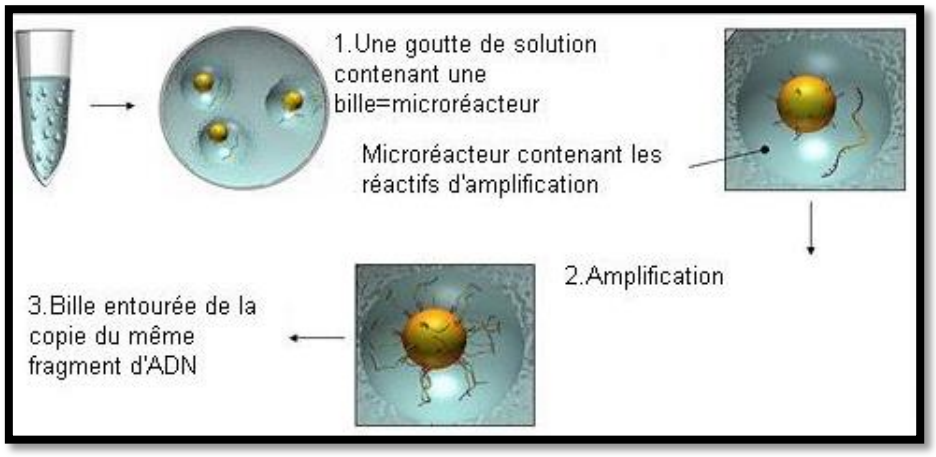


Figure 3 : Réaction de PCR en émulsion (emPCR)

iii. Pyroséquençage

Après amplification, les billes enrichies sont déposées sur la plaque servant au séquençage. Cette plaque, appelée PTP (PicoTiter Plate) contient plusieurs millions de puits dont le diamètre ne permet de récupérer qu'une seule microbille par puits (diamètre de 44µm). Dans chacun de ces puits, de minuscules microbilles contenant les réactifs nécessaires à la réaction de pyroséquençage sont également présentes (figure 4).

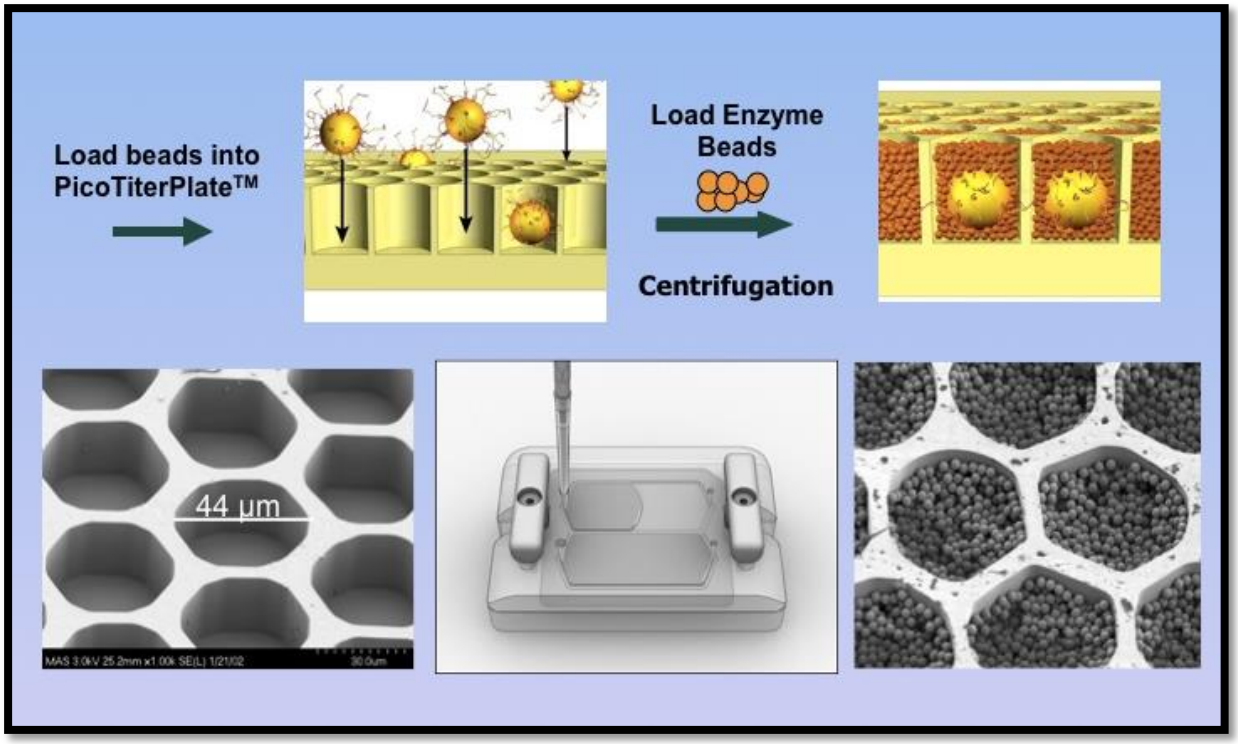


Figure 4 : Représentation du chargement des microbilles dans la PicoTiterPlate.

Un cycle de séquençage, également appelé 'flow cycle' se déroule de la façon suivante (figure 5):

- ➔ Injection d'un nucléotide A marqué
- ➔ Complémentation du/de plusieurs nucléotides A
- ➔ Elimination des nucléotides non complémentés par un traitement par une apyrase
- ➔ Excitation des nucléotides par un laser et réception de la fluorescence

Il en est ainsi de suite pour les trois autres nucléotides. Le passage des quatre nucléotides constitue un cycle de séquençage. Dans la version actuelle, 200 cycles se succèdent.

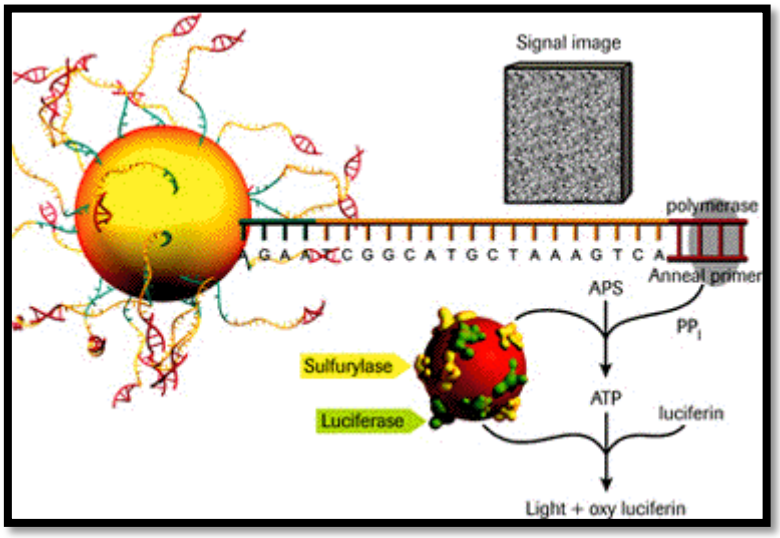


Figure 5 : Réaction de pyroséquençage.

A l'issue du pyroséquençage, l'ensemble des images de l'émission de la fluorescence observée pour la PTP est synthétisé et converti en un flowgram (figure 6) qui permet d'obtenir la séquence d'ADN de chaque read.

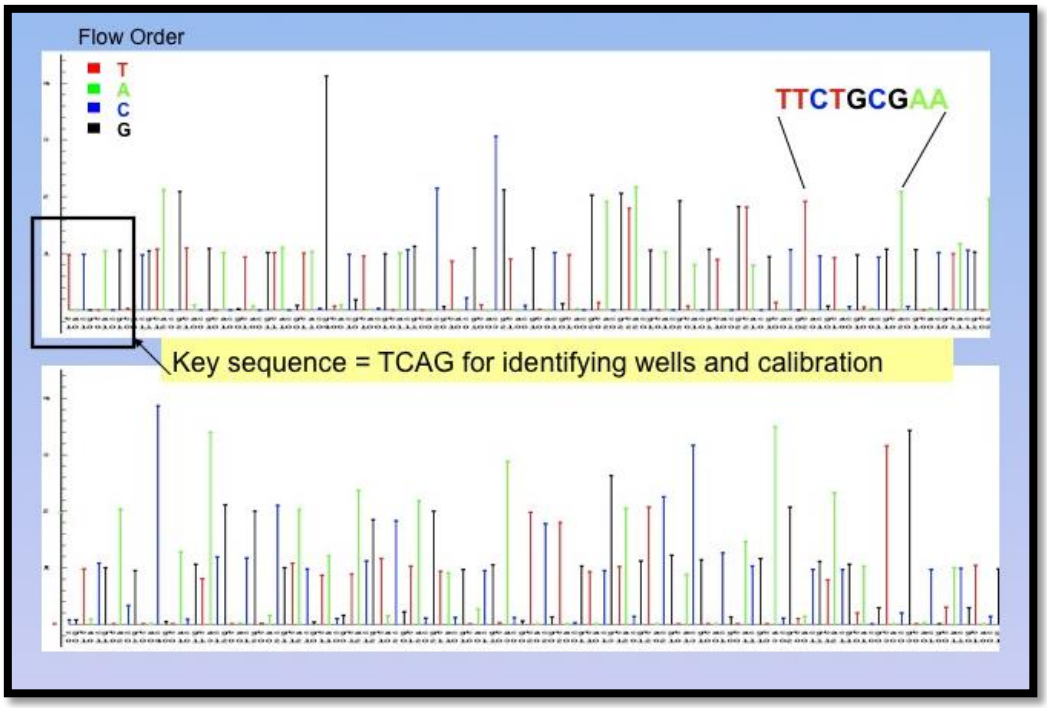


Figure 6 : Exemple de Flowgram, la hauteur de chaque pic dépend du nombre de nucléotides incorporé à chaque injection de nucléotides.

Annexe 2 : description des différentes étapes du séquençage Illumina

La réalisation du séquençage se déroule en trois étapes majeures, (i) la préparation de la librairie d'ADN à séquencer, (ii) l'étape d'amplification de l'ADN par PCR sur phase solide (amplification par pont- bridge PCR), (iii) le séquençage. Ces différentes étapes sont détaillées ci-dessous. Le séquençage est réalisé par le Genome Analyser (figure 1).

Figure 1 : Le Genome Analyzer

i. Préparation de la librairie

L'ADN extrait est fragmenté en séquences de plus petites tailles (typiquement 400 bases (HiSeq) de façon aléatoire par nébulisation ultra-sons. Une sélection fine des fragments d'une taille de 400 bases (plus ou moins 100) est effectuée sur gel d'agarose. Des adaptateurs A et B sont ensuite fixés à chaque séquence (figure 2a).

ii. Préparation du séquençage

Les molécules d'ADN simple brin sont fixées à l'aide des adaptateurs sur une plaque de silice comportant les adaptateurs complémentaires (figure 2b). Les brins d'ADN sont ensuite amplifiés par PCR afin de former des colonies ou des clusters sur la plaque de silice autour de la zone de fixation du simple brin d'ADN. Cette formation de cluster est favorisée par des cycles de fixation-dénaturation. Chaque molécule simple brin qui a un bout libre (bout lié à un adaptateur) se fixe sur son complémentaire et forme un pont. L'amplification se fait puis les deux brins se séparent, chacun ayant un bout libre et un bout fixé sur le support. Ce cycle est ensuite répété un grand nombre de fois. Lors du séquençage, chaque séquence composant la colonie va émettre de la fluorescence (4 couleurs pour chacun des 4 nucléotides) permettant au signal d'être suffisamment fort pour qu'il soit mesurable par les détecteurs.

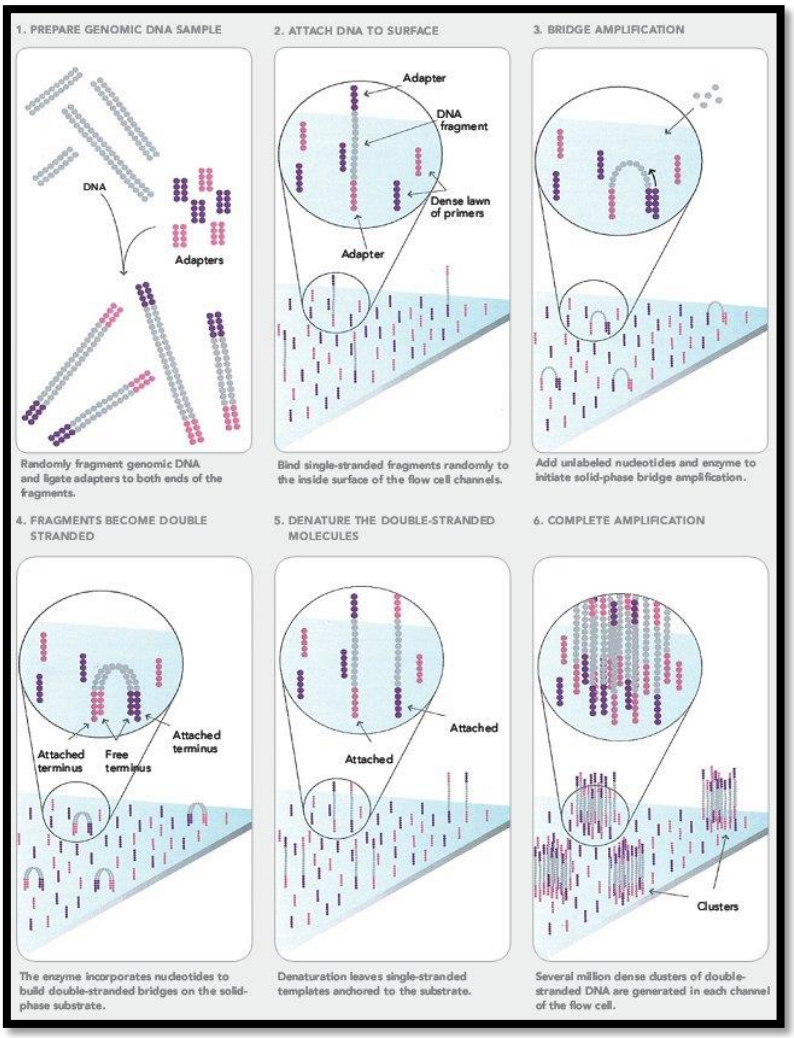


Figure 2 : synthèse des différentes étapes de la préparation de la librairie à l'amplification par pont des fragments d'ADN.

iii. Séquençage

Les colonies ainsi formées sont tout d'abord dénaturées pour n'avoir qu'un seul brin d'ADN. Les cycles de séquençage se déroulent de la façon suivante (figure 3) :

1-Injection d'un nucléotide fluorescent

2-Complémentation d'un seul nucléotide

3-Elimination des nucléotides excédentaires

4-Excitation des nucléotides complémentés par un laser et mesure de la fluorescence

Des ddNTP liés à des fluorophores sont ajoutés. La polymérase incorpore une base dans le brin en cours de synthèse et celle-ci émet une lumière qui est détectée. Ensuite, un clivage permet de retrouver une base dNTP qui pourra être liée à la suivante. La différence entre un dNTP et un ddNTP réside dans leur capacité à se lier ou pas en 3'. Un ddNTP ne peut être suivi par une autre base contrairement à dNTP.

5-Ce cycle est répété de même pour les 3 autres nucléotides

Par rapport à la technique de pyroséquençage, ce blocage de l'incorporation de nouvelles bases permet d'éviter le problème de détection des homopolymères car un seul nucléotide est incorporé à la fois. La taille du fragment séquencé est ici égale au nombre de cycles de séquençage. Ainsi pour 100 cycles, le fragment séquencé fera exactement 100 bases.

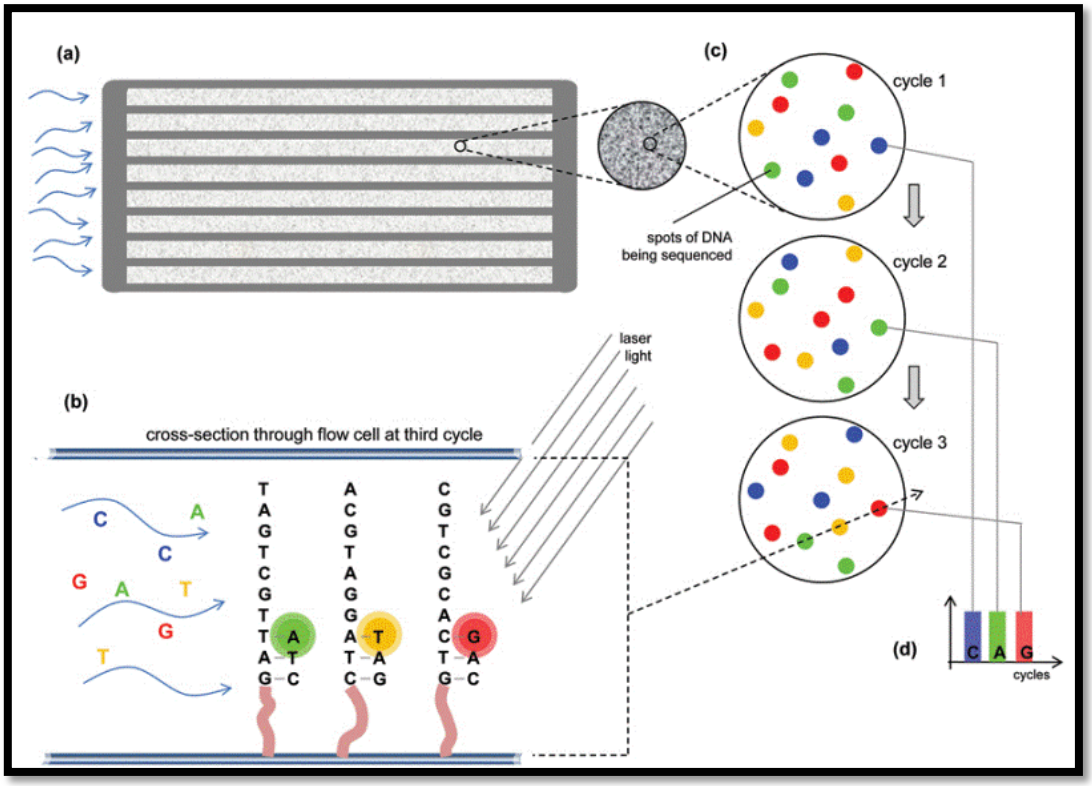


Figure 3 : Etape de séquençage

Annexe 3 : Script implémenté en Bash pour la conversion des fichiers .csv issus du logiciel Fluidigm pour combiner toutes les données SNP d'une puce sur une ligne par individu.

```
#!/bin/bash
ENTREE=$1
NB_COL=$2
NB_LIG=$3
SORTIE="/mnt/montagewindows/genotypage/puce22.csv"
touch $SORTIE
LIGNE=";"
for (( i=1 ; i <= $NB_COL ; i++ ))
        do
                SNP=`cat $ENTREE | head -n$i | tail -n1 | cut -
d";" -f2`
                LIGNE=$LIGNE$i"_"$SNP";"
        done
echo $LIGNE > $SORTIE
for (( i=1 ; i<=$NB_LIG ; i++ ))
        do
                NR_LIGNE=$((i*NB_COL))
                IND=`cat $ENTREE | head -n$NR_LIGNE | tail -n1 |
cut -d"-" -f1`
                echo "individu : $IND"
                IND2=`cat $ENTREE | head -n$NR_LIGNE | tail -n1 |
cut -d";" -f3`
                IND=$IND"_"$IND2
                LIGNE="$IND"
                for (( j=1 ; j<=$NB_COL ; j++ ))
                        do
                                NR_LIGNE=$(((i-1)*NB_COL+j))
                                SNP=`cat $ENTREE | head -n
$NR_LIGNE | tail -n1 | cut -d";" -f4`
                                LIGNE=$LIGNE";"$SNP
                        done
                echo $LIGNE >> $SORTIE
        done
```

Annexe 4: Script implémenté en Perl pour la gestion des stretchs anormaux de nucléotides identiques successifs

```
#!/usr/bin/perl -w

my $nb_elem = 7;
my $nb_pass = 13; #idealement c'est (nombre_snp / $nb_elem)
my $motif = "TRUC";

my $fichier = $ARGV[0];

open FICHIER, "<$fichier";

@lignes=<FICHIER>;

my $sortie = $ARGV[1];
open SORTIE, ">$sortie";

my $j = 0;

for (my $x = 0 ; $x < $nb_pass ; $x++) {
     my @lignes2;
     foreach $ligne (@lignes) {
          $j++;
          if (($un)= $ligne =~ /((No Call\;|XY\;|XY|No\
Call\;){$nb_elem,})/) {
               #print "($j) $un";
               my @snps = split(/\;/,$un);
               my $taille = @snps;
               my $chaine = "";
               for (my $i=0 ; $i< $taille ; $i++) {
                    $chaine .= "$motif;";
               }
               $nouvelleligne = str_replace($un,$chaine,$ligne);
               push(@lignes2, $nouvelleligne);
           } else {
               push(@lignes2, $ligne);
          }
     }
@lignes = @lignes2;
}

foreach $ligne (@lignes) {
     print SORTIE $ligne;
}

close FICHIER;
close SORTIE;
```

```
substr_replace {
     my $replace_this = shift;
     my $with_this= shift;
     my $string = shift;

     my $length = length($string);
     my $target = length($replace_this);

     for(my $i=0; $i<$length - $target + 1; $i++) {
          if(substr($string,$i,$target) eq $replace_this) {
               $string = substr($string,0,$i) . $with_this
.substr($string,$i+$target);
               return $string; #Comment this if you what a global
replace
          }
     }
     return $string;
}
```

Annexe 5: Protocole d'amplification de l'ensemble du génome (WGA)

Le mix d'amplification WGA a été réalisé selon le protocole fourni par Kbiosciences :

ReagentName	
gDNA	X
ddH20	37.41- X
10x PCR Buffer B	4.3
50mMMgCl2	1.2
2.5mMdNTPMix	2.63
Primer Mix (Poly N)	2.19
KleartaqPolymerase@ 5units	2.19

Pour chaque échantillon, sauf le 1 (5 µl d'ADN + 10 µl de ddH2O), 15 µl d'ADN ont été utilisé.

L'amplification a été réalisée sur un thermocycleur Veriti, selon le programme suivant :

Step1 - 94°C for 15 minutes,

Step2 - 94°C for 1 minute,
Step3 - 37°C for 2 minute,
Step4 - 55°C for 4 minutes,
(Repeat 10 times steps 2-4),

Step 5 - 94°C for 1 minute,
Step 6 - 37°C for 2 minutes,
Step 7 - 55°C for 4 minutes,
(Repeat 36 times steps 5-7, however in step 6 increase the temperature by0.5°C and duration by 15 sec for every repeat so that on the 36th repeat step6 the setting are 55°C for 11minutes).

Step 8 - 94°C for 30minutes.

Annexe 6: Tableau présentant l'hétérozygotie des populations de tiques de chaque lignes de collecte représentées par plus de six individus (N=34) ; N= nombre d'individus par ligne ; Hatt= hétérozygotie attendue et non biaisée ; Hobs = hétérozygotie observée ; *Fis* = indice de fixation intra-population

Population	identifiant (figure XX)	N	H_{obs}	H_{att}	*Fis*
L001	1	6	0,296	0,335	0,123
L007	2	6	0,299	0,350	0,1528
L010	3	6	0,288	0,348	0,1891
L013	4	7	0,320	0,355	0,0881
L015	5	9	0,299	0,339	0,125
L025	6	8	0,300	0,342	0,1117
L029	7	7	0,315	0,365	0,1363
L031	8	6	0,348	0,336	-0,0371
L032	9	7	0,322	0,357	0,117
L035	10	7	0,294	0,333	0,1039
L036	11	6	0,298	0,328	0,086
L037	12	7	0,248	0,336	0,2692
L038	13	7	0,262	0,347	0,2536
L039	14	8	0,319	0,345	0,0619
L040	15	9	0,273	0,327	0,1684
L041	16	10	0,303	0,349	0,1525
L042	17	8	0,309	0,327	0,0561
L045	18	10	0,261	0,312	0,1612
L046	19	10	0,269	0,330	0,179
L047	20	6	0,282	0,328	0,1377
L049	21	8	0,265	0,304	0,1282
L052	22	6	0,249	0,301	0,1817
L053	23	9	0,292	0,338	0,1545
L054	24	6	0,299	0,354	0,1623
L058	25	8	0,302	0,338	0,0942
L059	26	10	0,302	0,357	0,1491
L061	27	7	0,282	0,345	0,1766
L063	28	10	0,295	0,349	0,153
L064	29	10	0,288	0,345	0,1596
L065	30	7	0,316	0,346	0,0877
L078	31	8	0,321	0,359	0,1173
L079	32	6	0,293	0,350	0,1842
L080	33	6	0,256	0,334	0,2354
L083	34	8	0,253	0,337	0,2416

Annexe 7: Matrice des estimations de *Fst* pour l'ensemble des lignes de collecte représentées par plus de 6 individus (N=34).

pop	L001	L007	L010	L013	L015	L025	L029	L031	L032	L035	L036	L037	L038	L039	L040	L041	L042	L045	L046	L047	L049	L052	L053	L054	L058	L059	L061	L063	L064	L065	L078	L079	L080
L007	-0,010																																
L010	-0,028	-0,015																															
L013	-0,016	-0,016	-0,038																														
L015	-0,010	-0,003	-0,029	-0,010																													
L025	-0,008	0,010	-0,018	-0,008	-0,005																												
L029	-0,002	-0,009	-0,035	-0,017	-0,007	-0,002																											
L031	0,021	0,004	-0,012	0,001	0,006	0,025	-0,001																										
L032	0,021	-0,009	-0,006	-0,005	0,007	-0,001	-0,015	0,015																									
L035	0,024	-0,001	0,005	0,014	0,009	0,015	-0,009	0,027	0,006																								
L036	0,003	0,019	-0,019	-0,007	0,005	0,001	-0,002	0,009	0,006	0,017																							
L037	-0,001	0,004	0,004	-0,011	0,005	0,005	-0,010	-0,003	0,012	0,012	0,003																						
L038	0,012	0,014	-0,016	-0,002	-0,017	-0,007	-0,018	0,013	-0,006	0,003	0,004	-0,028																					
L039	0,023	0,028	-0,001	0,000	0,007	0,003	0,018	0,023	0,027	0,043	0,004	0,008	-0,008																				
L040	0,020	0,007	0,000	0,016	0,011	-0,014	-0,002	0,012	0,005	-0,005	-0,005	-0,003	-0,016	0,017																			
L041	0,001	0,005	-0,001	-0,009	0,000	-0,002	-0,005	0,008	0,017	0,004	0,002	-0,017	-0,020	0,003	-0,009																		
L042	-0,003	-0,001	-0,011	0,004	0,001	-0,007	-0,005	-0,009	0,014	0,023	-0,015	0,006	0,004	0,019	-0,002	-0,015																	
L045	0,004	0,000	-0,007	-0,001	0,006	0,000	-0,003	0,007	0,009	0,024	0,014	0,003	-0,002	0,014	0,012	0,007	-0,009																
L046	-0,018	0,002	-0,021	-0,011	0,001	-0,015	-0,016	0,008	0,002	0,014	0,009	-0,007	-0,012	0,012	0,012	-0,011	-0,004	-0,016															
L047	-0,001	-0,012	-0,011	-0,002	0,016	0,001	0,012	0,030	0,002	0,032	0,004	0,024	0,004	0,033	0,024	0,000	0,002	0,019	-0,003														
L049	-0,001	0,031	0,000	0,010	0,001	-0,003	0,018	0,037	0,020	0,056	0,026	0,016	0,008	0,043	0,030	0,007	0,021	0,015	-0,008	0,020													
L052	-0,011	0,011	-0,017	-0,010	-0,006	0,001	-0,009	-0,008	0,002	0,019	0,002	-0,005	-0,017	0,010	0,010	-0,013	-0,013	-0,008	-0,013	0,000	0,004												
L053	-0,009	-0,014	-0,016	0,001	0,001	0,007	-0,006	-0,002	0,015	0,020	0,007	0,000	-0,003	0,024	0,012	-0,002	-0,003	0,001	-0,013	0,007	0,002	-0,002											
L054	-0,002	-0,001	-0,001	0,006	0,001	0,008	0,002	0,017	0,013	0,012	0,010	-0,010	-0,006	0,012	0,009	0,000	-0,004	0,010	0,005	0,038	0,037	0,005	-0,004										
L058	0,018	0,032	0,014	0,022	0,025	0,009	-0,001	0,022	0,015	0,032	0,023	0,022	0,023	0,042	0,008	0,031	0,019	0,019	0,027	0,056	0,047	0,023	0,022	0,000									
L059	0,000	-0,005	-0,017	0,004	-0,003	0,004	0,007	0,015	0,008	0,012	0,015	0,013	0,002	0,010	-0,004	0,009	0,002	0,003	0,006	0,028	0,028	0,003	-0,005	-0,019	0,000								
L061	0,005	0,003	-0,010	0,000	0,015	0,017	0,008	0,010	0,020	0,046	0,004	0,012	0,021	0,022	0,029	0,027	0,006	0,012	0,034	0,025	0,033	0,023	0,018	-0,006	0,026	0,011							
L063	-0,006	0,014	-0,019	0,010	-0,001	-0,002	0,005	0,019	0,022	0,023	0,004	0,010	-0,004	-0,003	-0,001	-0,003	-0,008	0,001	0,002	0,014	0,025	0,010	0,007	-0,008	0,019	0,004	-0,007						
L064	0,003	0,009	-0,008	0,005	-0,005	-0,001	0,003	0,006	0,017	0,023	0,001	-0,017	-0,012	0,006	-0,009	-0,003	-0,001	-0,001	-0,003	0,031	0,015	-0,004	0,002	-0,023	0,016	-0,013	0,010	-0,010					
L065	0,022	0,012	0,021	0,017	0,024	0,024	0,030	0,026	0,030	0,035	0,021	-0,011	0,025	0,041	0,012	0,008	0,002	0,014	0,024	0,034	0,039	0,007	0,022	0,006	0,035	0,014	0,030	0,014	0,005				
L078	0,006	0,003	-0,011	0,005	0,006	0,004	0,011	0,020	0,012	0,020	0,011	0,032	-0,003	0,023	0,006	0,001	-0,011	0,008	0,008	0,022	0,022	0,000	0,010	0,013	0,017	-0,001	0,000	0,003	0,013	0,036			
L079	-0,010	-0,001	-0,030	-0,024	-0,008	-0,023	-0,007	0,023	0,007	0,012	-0,008	-0,013	-0,016	0,003	-0,002	-0,019	-0,005	-0,002	-0,006	0,016	0,008	0,004	0,001	-0,009	0,036	-0,005	-0,016	-0,005	-0,019	0,003	-0,008		
L080	-0,031	0,012	-0,003	-0,008	0,005	0,017	0,007	0,040	0,024	0,057	0,023	0,007	0,011	0,010	0,046	0,014	0,012	0,018	-0,011	-0,006	0,009	-0,008	0,011	0,024	0,036	0,027	0,027	0,011	0,010	0,039	0,011	-0,007	
L083	0,016	0,002	-0,012	0,009	0,004	0,012	0,014	0,008	-0,002	0,027	0,019	-0,003	-0,001	0,013	-0,004	0,017	0,009	0,005	0,022	0,052	0,026	0,013	0,008	-0,026	0,014	-0,017	-0,006	-0,003	-0,010	-0,012	0,019	-0,008	0,048

Annexe 8: Résultats des tests exacts de divergence de l'équilibre de Hardy-Weinberg pour les 128 marqueurs étudiés pour six populations de 10 individus (L041, l045, L046, L059, L063, L064). Les valeurs des erreurs types des calculs sont toutes inférieures à 0,003 (valeurs non présentées). Les marqueurs montrant des valeurs de p inférieures à la valeur de p critique (0,05) sont en rouge.

	L041	L045	L046	L059	L063	L064
locus	P	P	P	P	P	P
1133	0,132	0,160	0,053	1,000	1,000	0,169
3705	0,091	No	No	0,044	0,056	0,010
6283	0,477	0,216	1,000	1,000	0,084	1,000
6363	1,000	0,138	1,000	0,501	0,496	0,003
10041	1,000	0,503	0,478	0,459	0,540	0,544
19998	0,048	0,030	0,012	0,086	0,077	0,142
21130	1,000	1,000	0,245	1,000	0,242	0,031
30736	No	No	No	No	No	No
31200	0,244	0,475	1,000	0,499	1,000	0,011
32114	0,306	No	0,572	1,000	0,179	No
32551	No	No	No	No	No	No
34502	1,000	1,000	1,000	0,047	1,000	No
42351	1,000	1,000	0,174	0,478	0,534	0,022
57206	No	No	0,092	No	No	0,335
60684	0,478	1,000	0,129	0,386	0,246	1,000
61606	No	No	No	1,000	1,000	No
66390	1,000	No	No	0,159	No	0,058
68328	0,564	1,000	0,117	1,000	0,564	1,000
68391	0,341	0,047	No	0,067	1,000	1,000
72226	1,000	0,080	No	0,481	0,081	0,486
77668	0,527	0,058	0,003	0,200	0,091	0,007
78934	0,172	1,000	0,179	0,482	1,000	1,000
81501	1,000	0,059	No	0,175	No	0,011
81758	No	No	No	No	No	No
87199	1,000	No	1,000	1,000	1,000	0,083
93695	0,397	No	No	No	No	No
96296	1,000	1,000	1,000	0,132	0,174	0,523
105385	0,068	No	0,066	No	No	No
113142	1,000	0,246	0,177	1,000	1,000	0,003
114791	No	No	No	No	0,093	No
116335	No	No	No	No	No	No

125671	0,089	No	No	0,010	0,006	0,007
129322	0,055	0,343	1,000	0,172	0,048	0,341
133049	No	No	No	No	No	No
137096	0,431	No	No	0,007	0,055	0,044
143089	1,000	No	No	0,222	1,000	1,000
	L041	L045	L046	L059	L063	L064
locus	**P**	**P**	**P**	**P**	**P**	**P**
144259	0,046	No	No	No	No	No
145634	No	No	No	0,046	0,157	0,053
150669	0,054	0,305	0,159	0,045	No	No
155043	0,093	No	No	No	No	No
159151	No	No	No	No	0,054	No
166766	No	No	No	0,489	0,037	0,016
167418	No	No	No	No	No	No
175115	0,052	0,381	0,176	0,338	0,132	0,135
176991	0,082	0,571	0,573	1,000	1,000	0,574
180239	No	1,000	1,000	1,000	1,000	1,000
189207	1,000	0,222	1,000	1,000	0,246	1,000
197784	0,395	No	No	0,173	0,172	0,046
198227	1,000	0,343	1,000	0,481	1,000	0,478
205578	1,000	1,000	1,000	1,000	No	1,000
207995	No	No	No	No	No	No
208593	No	No	No	No	No	No
209761	0,076	No	No	No	0,077	No
210654	No	No	No	No	No	No
212829	1,000	0,173	0,046	0,563	1,000	0,245
214684	1,000	1,000	1,000	0,573	1,000	0,480
221603	1,000	1,000	0,437	No	No	No
224277	No	No	1,000	0,272	0,112	0,335
225377	No	No	No	No	No	No
230247	1,000	1,000	No	1,000	1,000	1,000
233961	1,000	1,000	0,220	0,480	1,000	1,000
234508	0,246	1,000	0,520	0,243	1,000	0,075
236290	No	No	No	No	No	No
243436	No	1,000	1,000	No	1,000	No
251320	0,135	0,491	0,437	0,015	0,014	0,503
255757	No	No	0,308	1,000	No	1,000
259770	0,247	0,047	1,000	0,169	0,016	0,006
281206	0,001	0,015	1,000	1,000	0,046	0,532
283680	No	No	No	0,077	0,065	No
287805	No	No	No	No	No	No
292025	No	No	No	No	No	No

296275	No	No	No	No	No	No
298125	No	No	No	0,176	No	No
299627	No	0,014	0,077	0,087	0,229	No
300752	No	No	No	No	1,000	No
303781	No	No	No	No	No	No
305888	1,000	1,000	1,000	No	0,306	1,000
307361	0,492	0,273	1,000	0,059	0,033	0,015
	L041	L045	L046	L059	L063	L064
locus	**P**	**P**	**P**	**P**	**P**	**P**
313057	No	No	No	No	No	No
320000	1,000	1,000	1,000	1,000	0,305	No
329834	No	No	No	No	1,000	0,157
333882	No	No	No	No	No	No
336267	No	No	No	No	No	No
339272	0,307	No	0,478	No	0,341	0,046
340581	0,482	1,000	1,000	1,000	1,000	1,000
356074	No	No	No	No	No	No
356395	1,000	No	No	No	No	No
371093	0,172	1,000	0,480	1,000	0,078	0,574
374382	No	No	1,000	No	No	No
376474	No	No	No	No	No	No
380487	No	No	0,077	No	No	0,052
393248	1,000	No	No	No	No	No
399212	0,036	0,014	0,216	0,119	0,214	0,047
411541	No	No	No	0,067	No	No
419658	0,482	0,493	1,000	0,479	0,177	1,000
428503	No	No	No	No	No	No
438644	0,010	0,015	0,002	1,000	1,000	0,304
441042	0,398	0,225	0,055	No	0,032	No
446758	No	No	No	No	No	No
450975	No	No	No	No	No	No
465604	0,006	1,000	0,133	0,016	0,014	0,012
465892	No	0,178	No	1,000	No	0,307
468480	0,029	0,007	0,172	0,503	0,386	0,158
480915	No	No	No	No	No	No
487540	No	No	No	No	No	No
493429	0,491	1,000	No	No	No	No
552113	1,000	1,000	1,000	1,000	1,000	0,242
558063	No	No	No	No	No	No
561492	No	No	No	No	No	No
580716	1,000	1,000	0,305	0,166	No	0,052
583125	1,000	0,082	1,000	0,574	No	1,000

585284	0,564	1,000	0,520	0,477	1,000	0,015
585318	0,160	0,155	0,176	0,309	1,000	1,000
589219	1,000	0,046	1,000	1,000	0,576	0,504
627150	No	No	No	0,480	1,000	No
751708	No	No	0,065	No	0,232	0,111
754496	No	0,060	No	No	0,066	0,075
761047	No	0,105	No	0,021	0,118	0,093
763022	0,172	0,173	0,479	0,174	0,220	0,478
764527	0,159	No	No	No	1,000	No
	L041	L045	L046	L059	L063	L064
locus	**P**	**P**	**P**	**P**	**P**	**P**
767569	No	No	No	No	No	No
768618	No	No	No	No	No	No
771828	No	No	1,000	No	No	No
775381	0,060	0,011	0,217	0,091	0,061	0,305
777961	0,478	1,000	1,000	No	No	No
781023	No	No	No	No	No	No
783090	1,000	0,077	0,494	1,000	0,157	No
792422	1,000	No	No	0,049	1,000	0,016

Annexe 9 : Publication 1 parue dans Molecular Ecology Resources

Quillery E, Quenez O, Peterlongo P, Plantard O (2013) Development of genomic resources for the tick *Ixodes ricinus*: isolation and characterization of Single Nucleotide Polymorphisms. *Molecular ecology resources*.

Molecular Ecology Resources (2013) doi: 10.1111/1755-0998.12179

Development of genomic resources for the tick *Ixodes ricinus*: isolation and characterization of single nucleotide polymorphisms

E. QUILLERY,*† O. QUENEZ,‡ P. PETERLONGO§ and O. PLANTARD*†

*Epidemiology and Risk Analysis in Animal Health, UMR1300 Biology, INRA, BP 40706, F-44307 Nantes, France, †UMR BioEpAR, LUNAM Université, Oniris, Nantes-Atlantic College of Veterinary Medicine and Food Sciences and Engineering, F-44307 Nantes, France, ‡GenOuest Core Facility, UMR6074 IRISA CNRS/INRIA/Université de Rennes1, Campus de Beaulieu, 35042 Rennes Cedex, France, §GenScale, INRIA Rennes Bretagne-Atlantique, IRISA, Rennes, France

Abstract

Assessing the genetic variability of the tick *Ixodes ricinus*—an important vector of pathogens in Europe—is an essential step for setting up antitick control methods. Here, we report the first identification of a set of SNPs isolated from the genome of *I. ricinus*, by applying a reduction in genomic complexity, pyrosequencing and new bioinformatics tools. Almost 1.4 million of reads (average length: 528 nt) were generated with a full Roche 454 GS FLX run on two reduced representation libraries of *I. ricinus*. A newly developed bioinformatics tool (READ2SNPS), which isolates SNPs without requiring any reference genome, was used to obtain 321 088 putative SNPs. Stringent selection criteria were applied in a bioinformatics pipeline to select 1768 SNPs for the development of specific primers. Among 384 randomly SNPs tested by Fluidigm genotyping technology on 464 individuals ticks, 368 SNPs loci (96%) exhibited the presence of the two expected alleles. Hardy–Weinberg equilibrium tests conducted on six natural populations of ticks have shown that from 26 to 46 of the 384 loci exhibited significant heterozygote deficiency.

Keywords: 454, *de novo* SNP calling, *Ixodes ricinus*, nonmodel organism, reduced representation library, SNP

Received 31 May 2013; revision received 17 September 2013; accepted 20 September 2013

Introduction

Ticks (Acari, Ixodidae) are haematophagous vectors of numerous pathogens. In Europe, *Ixodes ricinus* is the most widespread species and is responsible for both human and animal diseases (Gubler 1998; Parola & Raoult 2001). This tick transmits pathogenic agents responsible for zoonotic diseases such as Lyme disease, tick-borne-encephalitis, babesioses or rickettsioses (Gray 2002).

An accurate knowledge of tick dispersal and genetic diversity is required to set-up efficient vector control methods, such as acaricides or antitick vaccines, but is still lacking (Philipp *et al.* 1997; Gillet *et al.* 2009). However, all three sets of microsatellite markers developed independently to date (6 by Delaye *et al.* 1998; 17 by Røed *et al.* 2006 and 9 by Noel *et al.* 2012) exhibit high heterozygote deficiency (De Meeûs *et al.* 2004; Røed *et al.* 2006; Kempf *et al.* 2009, 2011; Noel *et al.* 2012) and, for several loci, nonmendelian transmission patterns that could not be fully explained by the high frequency of null alleles (De Meeûs *et al.* 2004; Røed *et al.* 2006). Moreover, the large number of alleles—sometimes differing by a single nucleotide, stutter bands or short allele dominance makes genotype assignation difficult (De Meeûs *et al.* 2004). To circumvent the difficulties associated with microsatellites, single nucleotide polymorphisms may constitute markers of choice for the investigation of genetic variability in *I. ricinus* (Noureddine *et al.* 2011; Porretta *et al.* 2013; Van Zee *et al.* 2013). SNPs have been successfully developed in other nonmodel organisms (Ekblom & Galindo 2011; Helyar *et al.* 2011). SNPs are the most abundant markers within genomes and exhibit a uniform distribution across chromosomes (Schlötterer 2004). Because of their lower mutation rates (compared with microsatellites; Estoup *et al.* 2002), SNPs exhibit less homoplasy (Morin *et al.* 2004). As they are biallelic, a lower error rate in genotyping and allele assignation can be expected. Moreover, due to their large numbers, they are very useful and informative for the investigation of genetic polymorphism (Lao *et al.* 2006; Paschou *et al.*

Correspondence: Elsa Quillery and Olivier Plantard, fax: 0240 687 751; E-mails: elsa.quillery@oniris-nantes.fr and olivier.plantard@oniris-nantes.fr

Journal Name MEN 12179 Manuscript No. WILEY Dispatch: 14.10.13 Journal: MEN No. of pages: 9 CE: Swathi Lakshmi PE: Iniya Selvi

2007). Fifty biallelic SNPs are considered to be equivalent to 20 highly polymorphic microsatellites (Smouse 2010).

Several methodological constraints have to be overcome during the development of SNPs in *I. ricinus.* First, because of the large estimated genome size (about 2.1 Gb based on the closely allied north American species *I. scapularis*), it is more difficult to get a higher sequencing depth. However, when looking for SNPs, a minimum depth is necessary to discriminate between true polymorphism and sequencing errors. Second, no reference genome is available for *I. ricinus*. Indeed, most of the available softwares developed to date and designed to call SNPs from NGS data sets make use of a reference genome (Li *et al.* 2009; McKenna *et al.* 2010). They map the reads on this reference genome to look for differences between this sequence and the reads (Nielsen *et al.* 2011).

In the case of *I. ricinus*, the *de novo* assembly from NGS data sets would thus be especially difficult to build because of the large size of the genome and the high proportion of repeated elements within this genome (Meyer *et al.* 2010).

We overcame the first difficulty using a method employing reduced representation libraries (RRL) to reduce genome complexity (Altshuler *et al.* 2000; Van Tassell *et al.* 2008). This method is based on the use of restriction enzymes and the selection of digested DNA fragments of a given range size. Finally, we developed an original method based on the Read2Snps tool (Uricaru et al. unpublished) that can identify SNPs by comparing their reads in raw NGS data sets using a de-Bruijn graph and without any genome assembly.

Materials and methods

Tick collections and DNA extraction

The DNA libraries were constructed by sampling two *I. ricinus* populations. Only adult females (the lifecycle stage with the largest amount of DNA) were used. The T population corresponds to 10 partially engorged females that were collected on roe deer in southwest France ('Gardouch'; 43° 23′ 27.88″N, 1° 41′ 1.67″E) during the winter of 2010 and kept at −80°C until DNA extraction. The M population corresponds to 20 nonengorged females that were collected in spring 2011 (and kept alive at 4°C in the laboratory until DNA extraction) as they were questing for a host on vegetation in northwest France ('Malville'; 47° 21′ 30.10″ N, 1° 51′ 41.59″W). Each individual was extracted, and the DNA concentration measured separately.

The ticks were frozen in liquid nitrogen and crushed with a pestle in individual tubes. DNA extraction was conducted according to manufacturer instructions using the NucleoSpin Tissue XS kit (Macherey-Nagel).

For the genotyping and SNP validation, *I. ricinus* nymphs have been collected by the drag method (Schulze *et al.* 1997) in 83 different sampling sites covering an area of 130 km^2 of the « Zone Atelier Armorique» (Bretagne, France; 48° 28′ 28.25″N, 1° 33′ 49.29″W). Between 1 to 10 *I. ricinus* individuals, nymphs were extracted in each sampling sites, resulting in a total of 464 individuals. Sixteen controls were added in the genotyping assay including 'no template controls' (NTC) used for Fluidigm technology, Whole-Genome Amplification controls and DNA extraction controls.

As the amount of DNA obtained from a single individual was insufficient for genotyping of all the SNPs isolated, a Whole-Genome Amplification (WGA) using the primer extension preamplification method (PEP-PCR) (LGC-Kbio) was performed on each individual DNA extract prior to genotyping.

Genome reduction and sequencing

After testing several restriction enzymes, *MseI* (T ^ AAT; New England Biolabs) was selected because a high concentration of DNA fragments of the target size (*id est* 500–600b, to maximize the efficiency of 454 pyrosequencing relative to reads length) could be obtained, and no discrete band (that would reveal the presence of repeat elements) was observed in the target range of DNA fragments selected. Each sample was digested for 8 hours (2.5 U/μg of DNA) according to manufacturer's instructions. The digested DNA was then separated on a 1% agarose gel (4 h, 80v). A gel piece of each sample, containing DNA fragments from 500 to 600 bp (according to the 100 pb marker size ladder; Eurobio), was sheared under a UV lamp. The DNA was then extracted and cleaned with the gel clean-up kit (Macherey-Nagel), eluted in 40 μL of EB Buffer (3.33 mM Tris, pH 8.5), then quantified using Qubit™ with the dsDNA HS Assay™ kit (Invitrogen). Each sample was then sent to the Biogenouest Genomic platform (Rennes, France) where the 454 sequencing was conducted. The samples from each of the 2 populations (T and M) were pooled separately. Pools were prepared from an equimolar quantity of DNA from each of the 10 to 20 samples, respectively, corresponding to 500 ng of DNA in 10 μl, tagged with a unique barcode (multiplex identifier MID) and sequenced using the Roche 454 GS FLX and Titanium chemistry.

The reads obtained with the 454 sequencer were trimmed by default filters. Another filter was performed directly by the sequencing platform (the Biogenouest Genomic platform, scripts are available upon request on website galaxy: galaxy.genouest.org) to delete reads that (i) did not contain the four nucleotides A, C, T, G; (ii) contained a high frequency of undetermined base (>7.0%); (iii) was less than 150 bp or greater than 950 bp

in size; (iv) contained repeat motifs (cf. 'passed 1' reads in Table 1). Finally, reads (or ends of reads) with a quality <20 (PHRED Quality score) or that did not contain the expected restriction site were deleted (script available upon request from INRA MIGALE bioinformatics platform: http://migale.jouy.inra.fr), as well as the MID tags used for the pyrosequencing (cf. 'passed 2' reads in Table 1).

Identification and checking of SNPs

A pipeline for SNP calling was developed based on the Read2Snps tool (Uricaru et al. unpublished). The Read2Snps source code is available under CeCILL licence, and it can be downloaded from colibread.inria.fr/read2snps/. This tool calls SNPs from one or several reads sets without using any reference genome.The Read2Snps tool is composed of two main modules. The first module, KisSnp2, constructs a de-Bruijn graph by extracting all words of length k (k-mers) from the reads. The de-Bruijn graph organizes the k-mers as follows: a node stores a k-mer and an oriented edge connects two nodes if the suffix of length k-1 of the source node is equal to the prefix of length k-1 of the target node. KisSnp2 then detects patterns in the graph that reveal the presence of a SNP in the read set(s). This first module output is a FASTA file containing sets of pairs of heterozygous sequences of length 2k-1. Because sequences are constructed by assembling k-mers from reads, it is necessary to map the initial reads on them. This step is performed by the second Read2Snps module called KissReads. This step (i) removes spurious sequences not existing in reads but only in k-mers and (ii) quantifies the average PHRED quality and the read sequencing depth of each position of each sequence and for each read set.

A draft assembly was conducted with MIRA3 (Chevreux *et al.* 1999) to estimate the coverage of the *I. ricinus* genome by the sequencing of our two reduced representation libraries.

SNP assay design and genotyping

For each SNP, read alignments containing the isolated SNPs were checked visually with the Tablet software (Milne *et al.* 2013). Only biallelic SNPs were considered. SNP loci located in the vicinity of microsatellite loci were excluded because such loci are known to exhibit high mutation rates that would prevent hybridization of the primers designed for genotyping. Among the SNPs satisfying all the above criteria, 384 loci (corresponding to 4 Fluidigm chips 96 × 96) were selected at random, for which primers were designed with the Perl Primer software (Marshall 2004) with length, annealing temperature and GC content as recommended for their use with Fluidigm technology and Kaspar chemistry (Table S1, Supporting information).

The WGA and the genotyping were conducted by the GENTYANE platform (INRA, Clermont-Ferrand, France) using the Biomark HD system (Fluidigm technology) and Kaspar chemistry (Wang *et al.* 2009).

Minor allele frequencies (MAF) were calculated on the 464 individuals genotyped (Table S1, Supporting information). Expected and observed heterozygote frequencies were calculated, and Hardy–Weinberg exact tests were conducted using GENEPOP'007 (Rousset 2008). The tests were conducted on the six largest populations of the sample, with 10 individuals for each population.

Results

A workflow illustrating the steps involved in data processing is given in Fig. S1.

Pyrosequencing of the reduced representation libraries

454 GS FLX sequencing generated 1 389 201 reads for the 2 reduced representation libraries (730482 and 658 719 for populations M and T, respectively) corresponding to 556 Mbp (Fig. S2, Supporting information). The reads had a mean length of 401 nt, with an average quality score of 33.2. Most of the reads (95%) began with 'TAA ', as expected for DNA fragments digested with the *MseI* enzyme. After application of the initial trimming steps, 392 693 reads were excluded due to insufficient quality score, presence of repeat sequences or absence of the restriction site.

Table 1 Summary statistics for 454 pyrosequencing of the 2 reduced representation libraries. The number of reads and their lengths are indicated in the raw data and after the first (Passed 1), and second criteria of trimming (Passed 2)

Population	Number of Individuals	Number of reads			Length of reads (Passed 2 reads)			
		Raw	Passed 1	Passed 2	Mean	Maximum	Minimum	Total nt
M	20	730 482	638 228	536 061	528	914	167	283 554 541
T	10	658 719	563 986	460 447	530	825	30	244 272 285
Total	30	1 389 201	1 202 214	996 508	529	914	30	527 826 826

SNPs discovery with READ2SNPS

A total of 996 508 reads (536 061 and 460 447 for populations T and M, respectively) with a mean length of 529 nt were used for the isolation of SNPs (Table 1). READ2SNPS was run on two read sets corresponding to the M and T populations. The main parameter in READ2SNPS is the k value. The best k value was determined by plotting the k-mer counting histograms (Fig. S3) for distinct k values. Finally, k = 29 was used, corresponding to the beginning of the plateau of the curve with more than 80% of unique K-mer and giving the optimal sequence size as output for READ2SNPS.

After applying the KisSnp2 module of the READ2SNPS software using k = 29, 791 803 SNPs were obtained. The KissReads module was then used to check whether the reconstructed sequences of length (2k-1) really corresponded to reads existing in the original data set. At this stage, 321 088 SNPs (corresponding to 40% of the initial number of SNPs found) were identified as true positive SNPs.

The SNPs generated by read2snps were then sorted according to their sequencing depth (Fig. 1).

This curve of distribution of sequencing depth for the SNPs was similar to that observed for the sequencing depth of contigs obtained during the assembly with MIRA3 (Fig. S4, Supporting information). Only those SNPs for which the sequencing depth was between 4 and 10 were retained for subsequent analysis, corresponding to 126 567 SNPs (i.e. 39.4% of the 321 088 SNPs previously isolated). This selection was conducted to eliminate SNPs that could be due to sequencing errors (considered to be present in SNPs with a sequencing depth <4) or located in duplicated loci or repeat elements (considered to be present in SNPs with a sequencing depth > 10). The SNPs sequence vicinity was used to filter out SNPs close to homopolymers as these concentrate sequencing errors in 454 data and can therefore affect primer hybridization. For this, we used a sliding window of 8 nucleotides and eliminated the sequence if 6 identical nucleotides were detected within this window. We also filtered out SNPs with PHRED sequence quality < 30. Only 9537 SNPs with a quality score >30 (PHRED) and without any homopolymer in the sequence of 2k-1 were kept. We additionally filtered out SNPs in sequences that could have been generated by erroneous

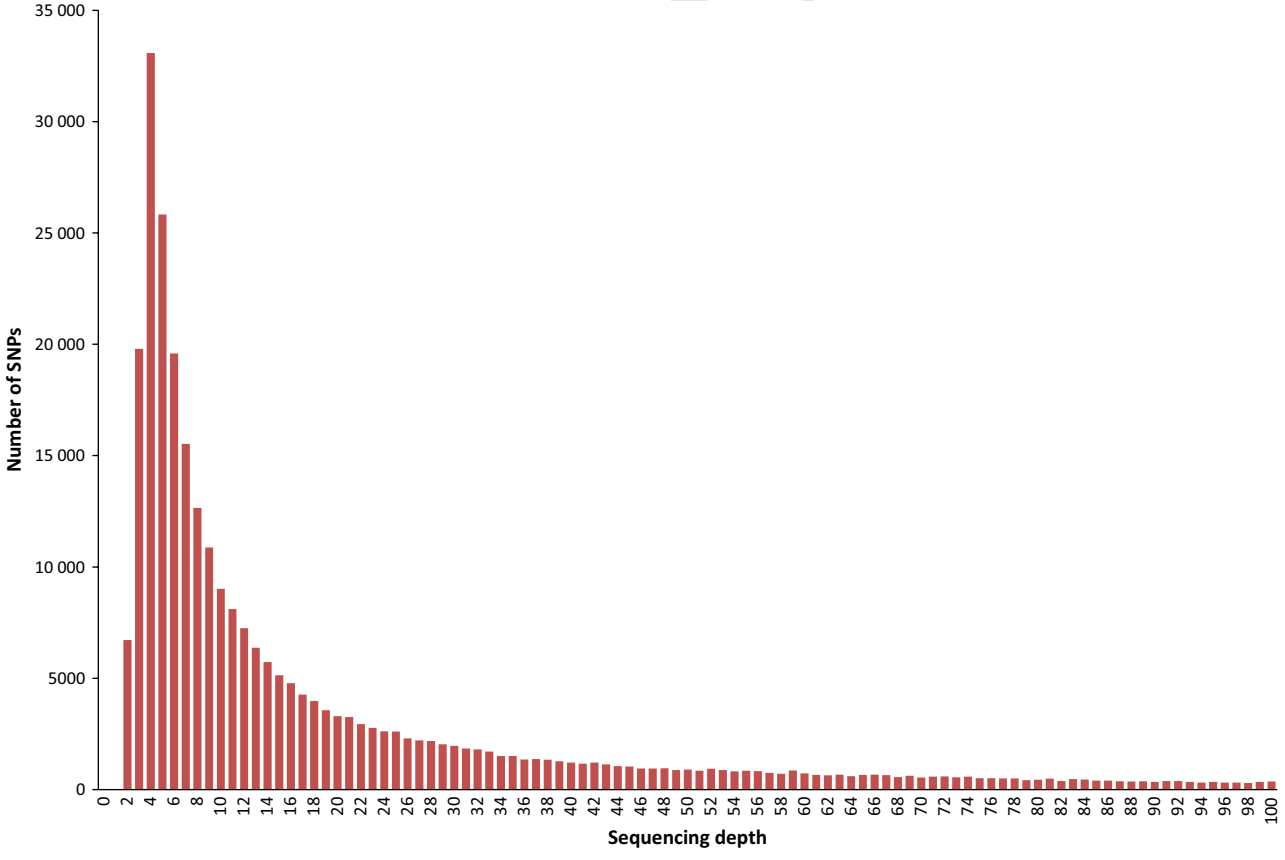


Fig. 1 Number of SNPs identified by READ2SNPS according to sequencing depth (the distribution was truncated; sequencing depths above 100 are not illustrated).

reads. To do this, we mapped back reads on all SNPs sequences (of length 2k-1) using GASSST (Global Alignment Short Sequence Search Tool; Rizk & Lavenier 2010) with a similarity threshold set at 80%. In this final step, SNPs for which the alignment between sequence and reads included gaps or substitutions could be detected and filtered out. As a result, 1768 SNPs were retained.

Description of the 384 SNPs isolated

From those 1768 SNPs, 384 SNPs were selected at random. The number of SNPs selected in each of the seven sequencing depth classes (4 to 10) was proportional to the initial sequencing depth distribution observed for the 1768 SNPs. Among those 384 SNPs, 254 SNPs (66%) corresponded to transitions while 130 (34%) corresponded to transversions. Twenty-two SNPs showed some polymorphism only in the M population and 62 only in the T population, while 300 SNPs exhibited the two alleles in each of the two RRL analysed. Homology of the 384 contigs (corresponding to the consensus of reads for a given SNP locus) was investigated by BLAST using sequences already deposited in GenBank. 56.51% of the retained sequences exhibited homology with sequences from *I. scapularis* (% identity > 80% and coverage > 10%) and 0.78% with another tick species (*Rhipicephalus* (*Boophilus*) *microplus*). No other significant match with any other organism was observed.

Genotyping results

Because of the small amount of DNA available for each individual tick nymph (~17 ng), a Whole-Genome Amplification method (WGA) was applied to obtain a x30 increase in available DNA.

Twenty Fluidigm chips (« dynamic array » 96.96 BioMark™) were used for the genotyping. One of them was not used in subsequent analyses for technical reasons. Finally, 168 378 genotyping points were analysed, of which 35 363 (21%) were not interpretable (lack of amplification or ambiguous genotype assignation) and were therefore considered as missing data.

Among the 384 SNPs, 5 loci showed no amplification (corresponding to 5.4% of the 35363 missing data). Eleven SNPs among the 484 individuals investigated were only found in the homozygous state. The remaining 368 SNPs provided suitable amplification results and exhibited the two alleles. The minor allele frequency (MAF) varied between 0.04 and 0.5, with a mean value of 0.23 (Table S2, Supporting information). In the 6 populations investigated for Hardy–Weinberg equilibrium (those with a population size of ten individuals), from 6.77% to 11.97% of the 384 loci were found with a significant heterozygote deficiency (Table S3, Supporting information).

Discussion

Due to the numerous difficulties encountered with the three sets of microsatellite loci developed in *I. ricinus* to date (Delaye *et al.* 1998; Røed *et al.* 2006; Noel *et al.* 2012), it was urgent to produce a new set of highly resolutive genetic markers for the analysis of genetic variability in this important tick species.

The SNP discovery workflow described here permitted the identification of 321 088 putative SNPs in the *I. ricinus* genome, the successful design of primers for 384 SNP loci and polymorphism was observed in 96% of the loci.

The development of these SNPs involved the following four steps: pyrosequencing of two reduced representation libraries from a pool of *I. ricinus* individuals, isolation of SNPs from this NGS data set with the Read2Snps pipeline, design of SNPs primers and SNPs genotyping.

To our knowledge, no inbred homozygous *I. ricinus* strain is currently available. Thus, within-individual polymorphism is expected even if a single individual is sequenced. Due to this nonreducible polymorphism, the difficulties encountered in the genome assemblage and the avoidance of sequencing errors are therefore increased. Moreover, as the amount of DNA available from a single individual is limited, several individuals must be pooled together for pyrosequencing.

The sequencing depth must be tuned according to the trade-off between the minimum number of reads for a given sequenced locus required to detect polymorphism (with a minimum of two reads) and the wastage of sequencing efforts due to sequencing of nonvariable genomic regions or regions that are already represented by a sufficient number of reads in the data set. We made use of sequencing depth to exclude sequencing errors (expected to be found in sequences represented by a single or a few reads). We also used the sequencing depth to avoid loci in repetitive DNA or duplicated loci (that are not suitable for SNP design).

An RRL strategy was employed to maximize the sequencing depth of the genomic region sequenced. After a draft assembly of our data set using MIRA3, we estimated that our libraries represented 78.3 Mpb, corresponding to a coverage of about 3.8% of the *I. ricinus* genome (considering a genome size of 2.1 Gb), with a mean sequencing depth of 2.58 and a median of 2.3. Although this sequencing depth could be considered as weak, SNPs were subsequently only selected in reads with a sequencing depth between 4 and 10 to avoid sequencing errors and repeat loci. Moreover, the choice of restriction enzyme and the size of the DNA fragments used for the library avoided the sequencing of repeat loci (such as transposable elements), which are known to be

common in the *I. scapularis* genome (Ullmann *et al.* 2005; Hill *et al.* 2009; Meyer *et al.* 2010) and unsuitable for the isolation of codominant and mendelian molecular markers.

Because of (i) the absence of reference genome, (ii) the large size of the *I. ricinus* genome, (iii) its high density of SNPs, (iv) its large proportion of repeat elements and (v) the frequency of sequencing errors obtained with the 454 technology, the genome assembly required by most SNP calling software would have been especially difficult to conduct. We thus argue that our bioinformatics tool, because it avoids genome assembly, is especially useful and relevant for the isolation of SNPs from the *I. ricinus* genome. This assembly step is in fact precluded using Read2Snps. The strategy adopted here revealed a large number of SNPs (321 088) in our data set. Using GASSST, we conserved only 18.5% (1768) of the 9537 SNPs meeting all the requested quality criteria (quality of read, sequencing depth, absence of homopolymers). Because of our RRL strategy (and the selection of only a fraction of the whole-genome in our data set, with varying sequencing depths for each locus), the SNP density cannot be directly inferred at the whole-genome scale. Moreover, the final number of SNPs isolated in this pipeline is highly dependent on the stringency of the filters used to avoid SNPs that might correspond to artefacts or be located in unsuitable loci (duplicated loci, occurrence of other SNPs or microsatellite loci in the vicinity of the SNP targeted…). In our case, as we wanted to select only a limited number of SNP loci (384) for population genetics studies, we used parameter values that were highly selective to keep only those SNP loci that were the most suitable for primer design and genotyping. However, numerous additional SNPs isolated could have been used to design new primers in the remaining set of polymorphic sites identified.

Among 1768 SNPs selected as candidates for primer design, 384 SNPs were tested by genotyping 464 individuals of *Ixodes ricinus* from an area (130 km^2) located in Brittany (North of France). These 384 SNPs were successfully amplified at 96%, as 368 displayed both amplification and polymorphism during genotyping. No amplification was observed for five SNPs, while 11 other SNP loci exhibited only one kind of homozygous individuals. The obtained validation rate is higher than in other studies where missing SNPs were reported to represent between 6 and 52% (Sanchez *et al.* 2009; Hyten *et al.* 2010; Fu & Peterson 2012). This reflects the effectiveness and stringency of the pipeline developed to select only those SNPs with the most suitable loci for primer design.

Among the 6 populations investigated, a deviation from Hardy–Weinberg equilibrium was observed in 6.77 to 11,97% of the 384 loci genotyped. This is in agreement with previous investigations based on microsatellite loci that have all found high heterozygote deficiency (De Meeûs *et al.* 2004; Røed *et al.* 2006; Noel *et al.* 2012). The limited active dispersal of those arthropods, the existence of host-specialized races as well as associative mating may be responsible for this pattern, through a Wahlund effect (Kempf *et al.* 2009, 2011). For a better understanding of *I. ricinus* population structure, complementary investigations conducted at a limited spatial scale and using ticks for which the host used is known will be needed (Quillery *et al.* in prep).

The strategy reported here, which combines high throughput sequencing (HTS), genome reduction with the RRL technique and a powerful bioinformatics pipeline to isolate SNPs without requiring any reference genome, illustrates the feasibility of SNP discovery for nonmodel organisms like *Ixodes ricinus*. The SNP loci described here will be useful for a variety of applications such as the assessment of the genetic structure of *I. ricinus* populations or the building of a genetic map.

Acknowledgements

The authors would like to thank the CEFS laboratory (Centre INRA de Toulouse), for providing engorged ticks collected on roe deer from Gardouch, and all the colleagues who took part in tick sampling at Malville and in the Zone Atelier Armorique. The heads of the Zone Atelier Armorique are thanked for providing access to the site. Philippe Vanderkoornhuyse, Alexandra Dheilly and Sophie Coudouel from the 'Biogenouest platform' (Rennes) are acknowledged for the 454 pyrosequencing run. Charles Poncet and the 'GENTYANE platform' is acknowledged for Fluidigm genotyping. Members of the 'Ticks and Tick-Borne Diseases' group (Réseau Ecologie des interactions durables) are thanked for useful discussions as well as François Beaudeau (UMR BioEpAR) for his comments on an earlier version of the manuscript. This work was funded by INRA (Animal Health Division; Call of Proposal Epidemiology and Biology/CADIX; AIP Bioressources: IXOMIC), the Pays de La Loire Region (PhD grant for E.Q.), the French National Research Agency (ANR; Call for Proposal 'Agrobiosphere', OSCAR project) and the European Commission (FP7 Health: EDENext projet).

References

Altshuler D, Pollara VJ, Cowles CR *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. pp. 45–56.

De Meeûs T, Humair P-F, Grunau C, Delaye C, Renaud F (2004) Non-Mendelian transmission of alleles at microsatellite loci: an example in *Ixodes ricinus*, the vector of Lyme disease. *International Journal for Parasitology*, **34**, 943–950.

Delaye C, Aeschlimann A, Renaud F, Rosenthal B, De Meeûs T (1998) Isolation and characterization of microsatellite markers in the *Ixodes ricinus* complex (Acari: Ixodidae). *Molecular Ecology*, **7**, 360–361.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Estoup A, Jarne P, Cornuet J-M (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*, **11**, 1591–1604.

Fu Y-B, Peterson GW (2012) Developing genomic resources in two Linum species via 454 pyrosequencing and genomic reduction. *Molecular Ecology Resources*, **12**, 492–500.

Gillet L, Schroeder H, Mast J *et al.* (2009) Anchoring tick salivary anticomplement proteins IRAC I and IRAC II to membrane increases their immunogenicity. *Veterinary Research*, **40**, ???–???.

Gray JS (2002) Biology of Ixodes species ticks in relation to tick-borne zoonoses. *Wiener Klinische Wochenschrift*, **114**, 473–478.

Gubler DJ (1998) Resurgent vector-borne diseases as a global health problem. *Emerging Infectious Diseases*, **4**, 442–450.

Helyar SJ, Hemmer-Hansen J, Bekkevold D *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, **11**, 123–136.

Hill CA, Guerrero FD, Zee JPV *et al.* (2009) The position of repetitive DNA sequence in the southern cattle tick genome permits chromosome identification. *Chromosome Research*, **17**, 77–89.

Hyten DL, Song Q, Fickus EW *et al.* (2010) High-throughput SNP discovery and assay development in common bean. *BMC Genomics*, **11**, 475.

Kempf F, De Meeus T, Arnathau C, Degeilh B, McCoy KD (2009) Assortative Pairing in *Ixodes ricinus* (Acari: Ixodidae), the European Vector of Lyme Borreliosis. *Journal of Medical Entomology*, **46**, 471–474.

Kempf F, De Meeus T, Vaumourin E *et al.* (2011) Host races in *Ixodes ricinus*, the European vector of Lyme borreliosis. *Infection, Genetics and Evolution*, **11**, 2043–2048.

Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M (2006) Proportioning Whole-Genome Single-Nucleotide–Polymorphism Diversity for the Identification of Geographic Population Structure and Genetic Ancestry. *The American Journal of Human Genetics*, **78**, 680–690.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078–2079.

Marshall OJ (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics*, **20**, 2471–2472.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Meyer JM, Kurtti TJ, Zee JPV, Hill CA (2010) Genome organization of major tandem repeats in the hard tick, *Ixodes scapularis*. *Chromosome Research*, **18**, 357–370.

Milne I, Stephen G, Bayer M *et al.* (2013) Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics*, **14**, 193–202.

Morin PA, Luikart G, Wayne RK, the SNP workshop group, (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Noel V, Léger E, Gómez-Díaz E, Risterucci A-M, McCoy KD (2012) Isolation and characterization of new polymorphic microsatellite markers for the tick *Ixodes ricinus* (Acari, Ixodidae). *Acarologia*, **52**, 123–128.

Noureddine R, Chauvin A, Plantard O (2011) Lack of genetic structure among Eurasian populations of the tick *Ixodes ricinus* contrasts with marked divergence from north-African populations. *International Journal for Parasitology*, **41**, 183–192.

Parola P, Raoult D (2001) Tick-borne bacterial diseases emerging in Europe. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, **7**, 80–83.

Paschou P, Ziv E, Burchard EG *et al.* (2007) PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations. *PLoS Genetics*, **3**, e160.

Philipp MT, Lobet Y, Bohm RP Jr *et al.* (1997) The outer surface protein A (OspA) vaccine against Lyme disease: efficacy in the rhesus monkey. *Vaccine*, **15**, 1872–1887.

Porretta D, Mastrantonio V, Mona S *et al.* (2013) The integration of multiple independent data reveals an unusual response to Pleistocene climatic changes in the hard tick *Ixodes ricinus*. *Molecular Ecology*, **22**, 1666–1682.

Rizk G, Lavenier D (2010) GASSST: global alignment short sequence search tool. *Bioinformatics*, **26**, 2534–2540.

Røed KH, Hasle G, Midthjell V, Skretting G, Leinaas HP (2006) Identification and characterization of 17 microsatellite primers for the tick, *Ixodes ricinus*, using enriched genomic libraries. *Molecular Ecology Notes*, **6**, 1165–1167.

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Sanchez CC, Smith TP, Wiedmann RT *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

Schlötterer C (2004) The evolution of molecular markers — just a matter of fashion? *Nature Reviews Genetics*, **5**, 63–69.

Schulze TL, Jordan RA, Hung RW (1997) Biases associated with several sampling methods used to estimate abundance of *Ixodes scapularis* and Amblyomma americanum (Acari: Ixodidae). *Journal of Medical Entomology*, **34**, 615–623.

Smouse PE (2010) How many SNPs are enough? *Molecular Ecology*, **19**, 1265–1266.

Ullmann AJ, Lima CMR, Guerrero FD, Piesman J, Black WC IV (2005) Genome size and organization in the blacklegged tick, *Ixodes scapularis* and the Southern cattle tick *Boophilus microplus*. *Insect Molecular Biology*, **14**, 217–222.

Van Tassell CP, Smith TPL, Matukumalli LK *et al.* (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247–252.

Van Zee J, Black WC IV, Levin M *et al.* (2013) High SNP density in the blacklegged tick, *Ixodes scapularis*, the principal vector of Lyme disease spirochetes. *Ticks and Tick-borne Diseases*, **4**, 63–71.

Wang J, Lin M, Crenshaw A *et al.* (2009) High-throughput single nucleotide polymorphism genotyping using nanofluidic Dynamic Arrays. *BMC Genomics*, **10**, 561.

E.Q. conceived and designed the project and performed experiments, collected samples, analysed data and wrote the study. O.Q. conceived and performed the bioinformatics analysis and helped correct the text. P.P. conceived and developed the bioinformatics tool and helped correct the text. O.P. conceived and designed the project, collected samples and wrote the study.

Data Accessibility

For SNPs sequences, please see the online supplementary materials. For DNA sequences, see NCBI SRA: SRX327489 For draft DNA sequence assembly (.ace), output files from Read2Snps (.fasta) and Bioinformatics scripts see DRYAD entry doi:10.5061/dryad.1h1f2.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Workflow illustrating the application of 454 pyrosequencing in the *de novo* identification, selection and validation of SNPs.

Figure S2 Distribution of the reads (length in bp) for the two PicoTiterPlate regions in the Roche 454 GS FLX run.

Figure S3 Ratio of unique k-mers obtained for different K-mer size (nt).

Figure S4 Number of contigs identified by MIRA3 with a sequencing depth only between 2 and 20.

Table S1 Excel file containing primers sequences for the 384 SNP loci.

Table S2 Excel file containing the 384 sequences containing SNP and with MAF for each locus.

Table S3 Excel file containing Observed (H_o) and expected (H_e) heterozygosity, p-value and standard error of Hardy–Weinberg Equilibrium test in the 6 populations with the largest population size (L041 to L064).

Author Query Form

Journal: MEN

Article: 12179

Dear Author,
During the copy-editing of your paper, the following queries arose. Please respond to these by marking up your proofs with the necessary changes/additions. Please write your answers on the query sheet if there is insufficient space on the page proofs. Please write clearly and follow the conventions shown on the attached corrections sheet. If returning the proof by fax do not write too close to the paper's edge. Please remember that illegible mark-ups may delay publication.

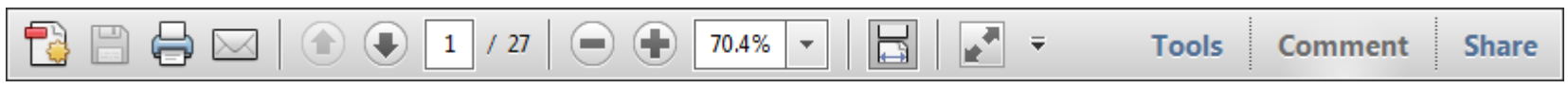
Many thanks for your assistance.

Query reference	Query	Remarks
1	**AUTHOR: Please provide full postal address for 4th affiliation.**	
2	**AUTHOR: Please check the corresponding author names.**	
3	**AUTHOR: Please provide all author names along with their initials for all unpublished data.**	
4	**AUTHOR: Please provide the page range for reference Gillet et al. (2009).**	

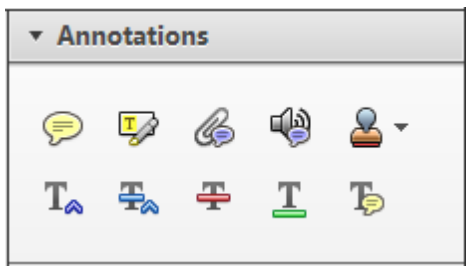
USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

Required software to e-Annotate PDFs: Adobe Acrobat Professional or Adobe Reader (version 8.0 or above). (Note that this document uses screenshots from Adobe Reader X)
The latest version of Acrobat Reader can be downloaded for free at: http://get.adobe.com/reader/

Once you have Acrobat Reader open on your computer, click on the Comment tab at the right of the toolbar:



This will open up a panel down the right side of the document. The majority of tools you will use for annotating your proof will be in the Annotations section, pictured opposite. We've picked out some of these tools below:

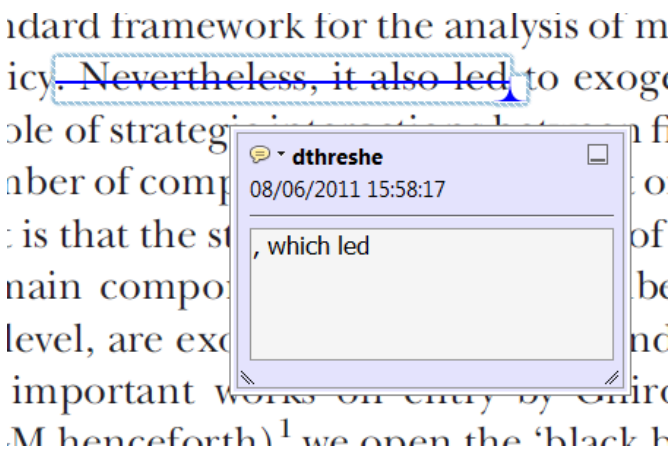


1. Replace (Ins) Tool – for replacing text.

Strikes a line through text and opens up a text box where replacement text can be entered.

How to use it

- Highlight a word or sentence.
- Click on the Replace (Ins) icon in the Annotations section.
- Type the replacement text into the blue box that appears.

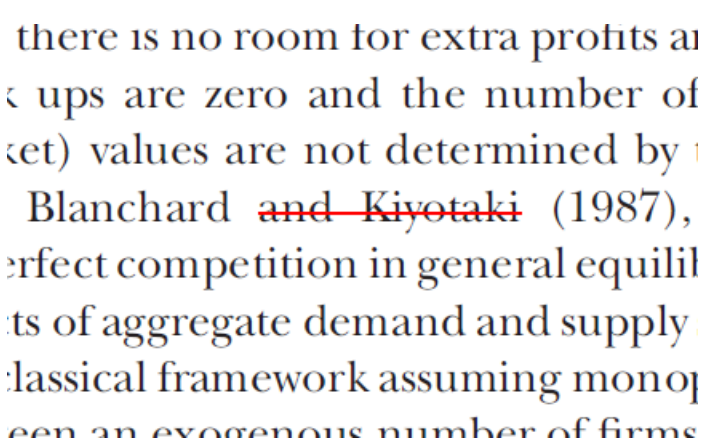


2. Strikethrough (Del) Tool – for deleting text.

Strikes a red line through text that is to be deleted.

How to use it

- Highlight a word or sentence.
- Click on the Strikethrough (Del) icon in the Annotations section.

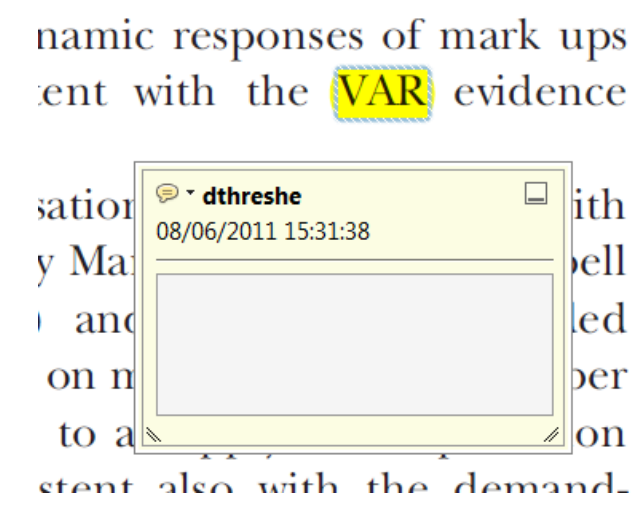


3. Add note to text Tool – for highlighting a section to be changed to bold or italic.

Highlights text in yellow and opens up a text box where comments can be entered.

How to use it

- Highlight the relevant section of text.
- Click on the Add note to text icon in the Annotations section.
- Type instruction on what should be changed regarding the text into the yellow box that appears.

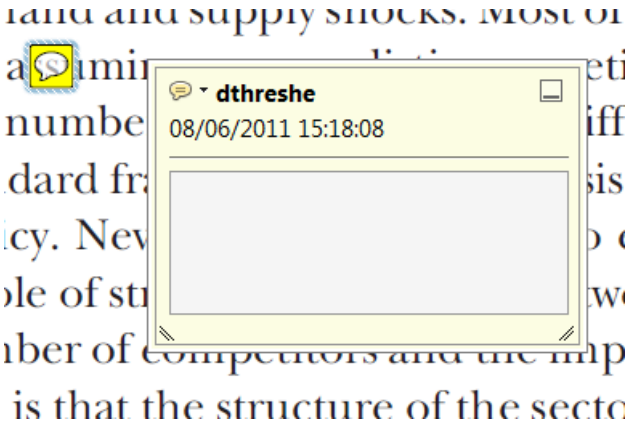


4. Add sticky note Tool – for making notes at specific points in the text.

Marks a point in the proof where a comment needs to be highlighted.

How to use it

- Click on the Add sticky note icon in the Annotations section.
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the yellow box that appears.



WILEY-BLACKWELL

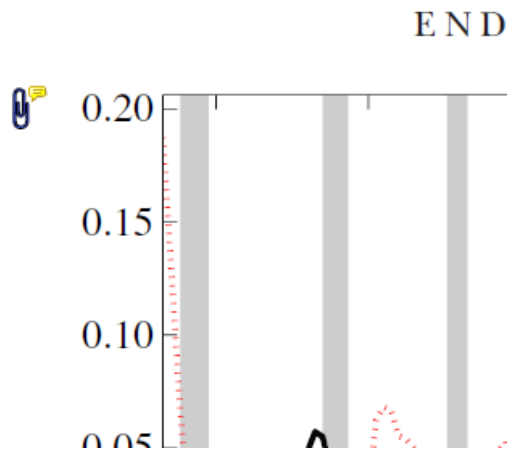
USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

5. Attach File Tool – for inserting large amounts of text or replacement figures.

Inserts an icon linking to the attached file in the appropriate pace in the text.

How to use it

- Click on the Attach File icon in the Annotations section.
- Click on the proof to where you'd like the attached file to be linked.
- Select the file to be attached from your computer or network.
- Select the colour and type of icon that will appear in the proof. Click OK.

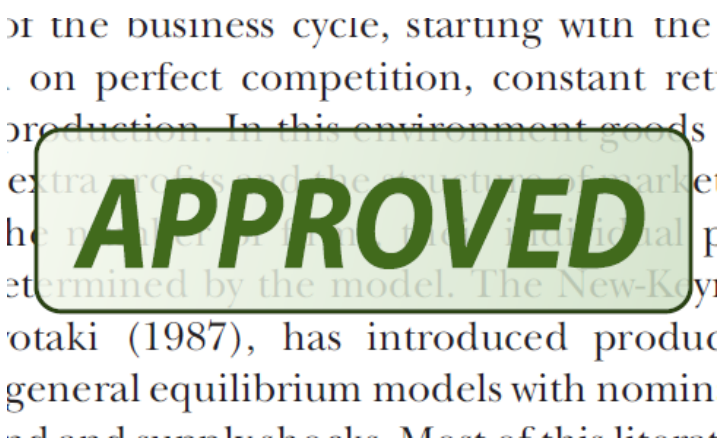


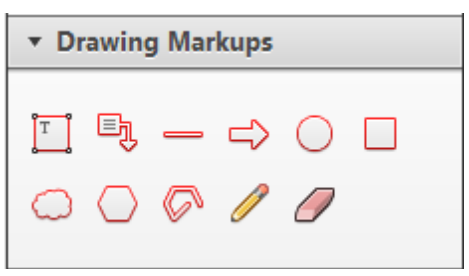
6. Add stamp Tool – for approving a proof if no corrections are required.

Inserts a selected stamp onto an appropriate place in the proof.

How to use it

- Click on the Add stamp icon in the Annotations section.
- Select the stamp you want to use. (The Approved stamp is usually available directly in the menu that appears).
- Click on the proof where you'd like the stamp to appear. (Where a proof is to be approved as it is, this would normally be on the first page).



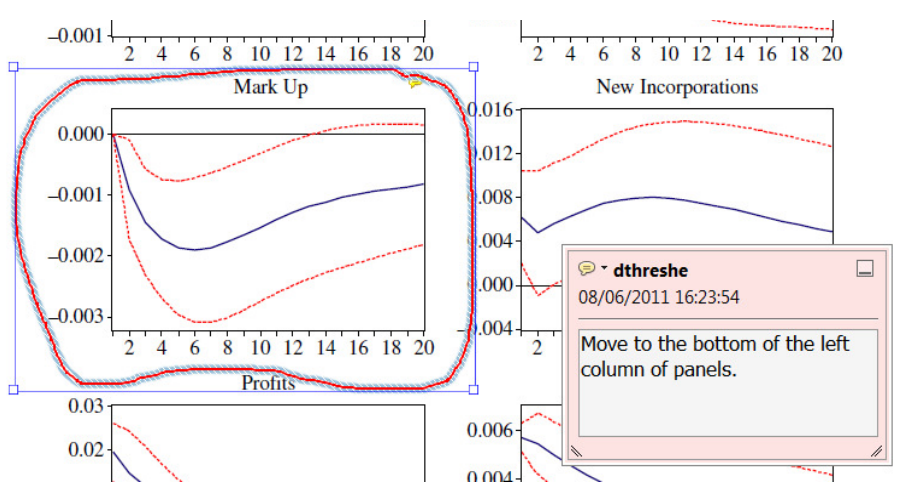


7. Drawing Markups Tools – for drawing shapes, lines and freeform annotations on proofs and commenting on these marks.

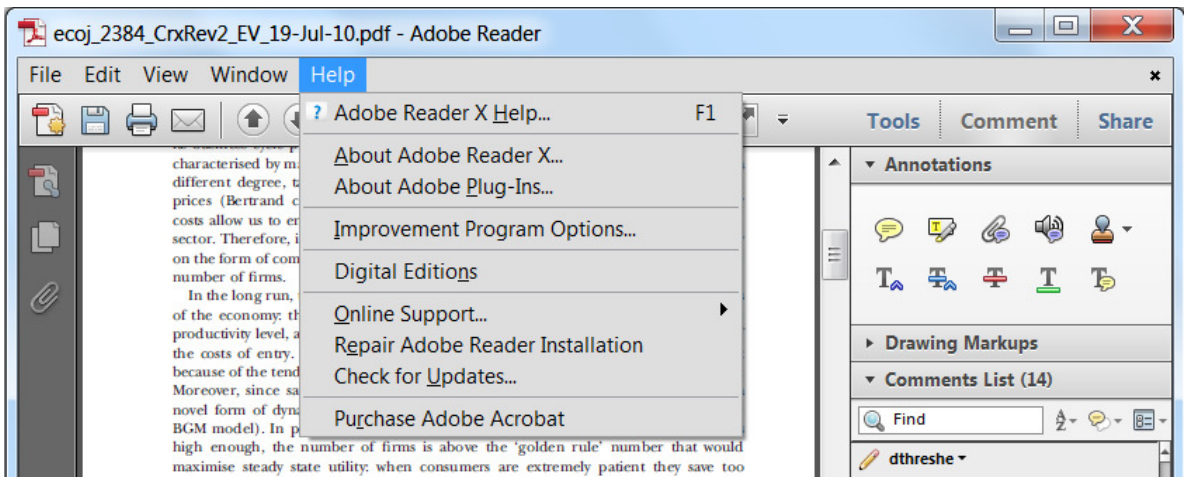
Allows shapes, lines and freeform annotations to be drawn on proofs and for comment to be made on these marks..

How to use it

- Click on one of the shapes in the Drawing Markups section.
- Click on the proof at the relevant point and draw the selected shape with the cursor.
- To add a comment to the drawn shape, move the cursor over the shape until an arrowhead appears.
- Double click on the shape and type any text in the red box that appears.



For further information on how to annotate proofs, click on the Help menu to reveal a list of further options:



Annexe 10 : Publication 2 en cours de préparation

Raluca Uricaru R, Rizk G, Lacroix V, **Quillery E**, Plantard O, Chikhi R, Lemaitre C and Peterlongo P *in prep* Reference-free detection of genotypable SNPs.

N.B: **Cette publication étant encore en préparation, merci de ne pas diffuser la publication ainsi que les résultats présentés.**

RESEARCH

Reference-free detection of genotypable SNPs

Raluca Uricaru[1,2*]
, Guillaume Rizk[3]
, Vincent Lacroix[4]
, Elsa Quillery[5,6]
, Olivier Plantard[5,6]
, Rayan Chikhi[7]
, Claire Lemaitre[3]
and Pierre Peterlongo[3*]

Full list of author information is available at the end of the article
*Corresponding authors:
ruricaru@labri.fr,
pierre.peterlongo@inria.fr

Abstract

We propose a new method calling both heterozygous and homozygous *genotypable* SNPs from any number of read datasets, without a reference genome, and with very low memory and time footprints (billions of reads can be analyzed with a standard desktop computer). We establish for the first time that calling SNPs directly from sequenced reads gives better results than a state-of-the-art assembly and mapping approach, mostly due to the fact that our approach better discriminates SNPs from inexact repeats, while still detecting SNPs in repeated regions. On an arthropod species, 96% of the predicted SNPs that were tested *in vitro* were confirmed.

Keywords: reference-free; SNPs; next-generation sequencing; low memory

Background

Assessing the genetic differences between individuals within a species, or between chromosomes of an individual, is a fundamental task in many aspects of biology. Of specific interest, single nucleotide polymorphisms (SNP) are variations of a single base, either between two homologous chromosomes within a single individual, or between two individuals. Finding biallelic or mendelian SNPs is often done in many biological applications involving SNP genotyping, e.g. population genomics, health, ecology or agronomy research[*]. As they must be easily amplified by PCR, those SNPs must not be surrounded by other polymorphism sources such as other SNPs, repeats and structural variants. We say that such SNPs are *genotypable*. With next-generation sequencing technologies, individuals from virtually any species can be sequenced at a modest cost. State of the art methods to detect SNPs between individual or strains generally require a high-quality reference genome [1, 2]. However, biologists are increasingly working on non-model organisms, for which building high-quality references is extremely challenging [3]. For these reasons, there is a strong need for methods able to detect SNPs, in particular those which are genotypable, without relying on a reference genome.

[*] RC — Needs a citation here, anyone has one?

Methods that detect SNPs can be broadly divided into three categories, based on their use of a reference genome. The first category, to which we refer as *reference-based*, consists of SNP detection methods that map reads of an individual to a high-quality reference genome [1, 2]. Detecting SNPs between two sequenced individuals with a reference-based method is done via mapping both read datasets to a reference sequence. Such approaches are limited by the quality and the mappability of available reference genomes [4]. Furthemore, in many biological experiments, a reference genome for the organism of interest may not yet exist.

In this article, we work under the hypothesis that a high-quality reference sequence of the sequenced organism is not available. We therefore focus on the remaining two categories of SNP detection methods. Methods in the second category perform *de novo* assembly to reconstruct the genome of a sequenced individual A. Then, these methods include, as a sub-module, a reference-based method to map the reads of an individual B to the assembly of A. We refer to such methods as *hybrid*, as they use both *de novo* assembly and mapping techniques to call SNPs. [†] The third category does not make use of a reference genome at any stage. We refer to this methods in category as *de novo*. At a high level, a *de novo* SNP calling method transforms a read dataset into a graph, and detects SNPs between two or more datasets by comparing their graphs.

Several methods that fall into the *de novo* category have been developed recently [5, 6, 7, 8, 9]. However, they generally suffer from a high number of false positive calls [5, 6], due to genomic repetitions and sequencing errors. We also note that all *de novo* methods use similar computational techniques to *de novo* assembly, entailing unpractical computational costs on large genomes. Even with substantial computational resources, most available *de novo* tools still cannot detect variants in mammalian-sized datasets. Therefore, there is still a need for robust tools that can detect genotypable SNPs without a reference genome.

We introduce a new *de novo* method, DISCOSNP. It is designed to call genotypable SNPs directly from sequenced reads, without a reference genome. DISCOSNP focuses on detecting high-quality genotypable SNPs, unlike other methods which aim to detect all SNPs. DISCOSNP has been used to detect heterozygous genotypable SNPs, to build databases of high-quality markers within and across populations. DISCOSNP can also detect homozygous SNPs between individuals/strains, to create discriminatory markers. It introduces several mechanisms to avoid detecting false positive SNPs within this category[‡].

We compared DISCOSNP with other *hybrid* and *de novo* methods.[§] Compared to KISSNP and NIKS, DISCOSNP supports heterozygous and homozygous SNP detection for one, two or more individuals. Compared to BUBBLEPARSE, DISCOSNP shows better precision/recall performance. It also has a faster runtime and is at least two orders of magnitude more efficient in term of memory. We establish for the first time that our *de novo* method gives significantly better results than a typical *hybrid* method composed of SOAPdenovo2 and GATK. We performed experiments on two simulated datasets (bacterial and human chromosome 1) and

[†] RC — Do we have papers to cite for this category?
[‡] RC — Je ne suis pas sur de la veracite de cette phrase, a verifier
[§] RC — Il faudra mettre des chiffres dans ce paragraphe

two real datasets from large genomes (mouse and tick). On the tick data, a wet-lab experiment over a sample of SNPs predicted by DISCOSNP validated 96% of them. Finally, DISCOSNP is designed to reach a wide audience, as it aims to be easy to use, regardless of the size and the complexity of the input data.

Results

We evaluated DISCOSNP to detect homozygous and heterozygous genotypable SNPs from one or more sets of raw reads, without using a reference genome sequence. We also compared DISCOSNP to other *de novo* SNP detection methods: KISSNP, BUBBLEPARSE and NIKS. These tools construct a *de Bruijn graph*, which is a directed graph where nodes are all words of length k from reads (*k-mers*), and where an edge connects two nodes if their k-mers have a k-1 overlap. They then search for specific patterns called *bubbles* in the graph, which correspond to SNPs.

KISSNP [9] is a software that takes as input two sets of reads, and detects SNPs using k-mers that are differentially abundant between the two datasets. CORTEX_VAR [5] constructs a colored de Bruijn graph from n datasets. BUBBLEPARSE [6] is an improvement of CORTEX_VAR in which bubbles that are not clean (i.e. nodes in the bubble may be connected to other nodes in the graph) can also be detected. Furthermore BUBBLEPARSE classifies SNPs into homozygous and heterozygous groups. NIKS [7] finds homozygous SNPs between two datasets by performing local *de novo* assembly around sample-specific k-mers.

In a few words[¶], DISCOSNP constructs the *de Bruijn graphs* from all input read sets pooled together and detects bubble patterns in the graph, which may correspond to SNPs (*first module*). An example of such bubble is proposed in Figure 1, while a full explanation is given in Section "Materials and methods". By mapping reads back on the sequences of the bubbles (*second module*) the average coverage and read quality for each allele and each read set are computed. In practice, from one or more read set(s), DISCOSNP produces a *multi-fasta* file composed of pairs of sequences that are identical except for one substitution, *i.e.* one pair per SNP. These SNPs are ranked according to their read coverage in each condition (or read set), while favoring SNPs that are discriminant between conditions.

Compared to the other tools, DISCOSNP uses recent advances enabling implementation of de Bruijn graphs with as little as 5 GB of memory for full human sequencing data [10, 11]. DISCOSNP has only two main parameters: the size of the k-mer used to construct the de Bruijn graphs and the minimal coverage a k-mer should reach to be inserted in the graph.

Results were obtained with DISCOSNP, version 1.0.0 available online. The tests were performed on the *GenOuest* (genouest.org) cluster, composed by Intel Xeon© core processors with speed varying between 2 and 2.8GHz.

Results on simulated datasets

Assessing exactly the recall and precision of a SNP calling approach requires datasets for which a perfect and exhaustive list of SNPs exists. Thus, in order

[¶] RC — Ces deux paragraphes suivants sont bien mais n'ont pas leur place dans les results, je vous proposer de les deplacer dans les methods ou dans une section "Overview of Methods", qu'en dites vous?

to give a first insight on the quality and performance of DISCOSNP, we carried out extensive tests on simulated datasets enabling exact control of the results quality. The simulation procedure consists in introducing substitutions in a real genome sequence at a fixed rate, based on a uniform distribution. The sequencing process is then simulated by sampling fixed-length sequences (100 bp) at a given coverage (50x) with a fixed substitution error rate (1%).

In order to check the reliability of introducing SNPs based on a uniform distribution, we compare the results we obtain with this approach on human chromosome 1, to the ones we obtain on data simulated using true SNP positions, i.e. referenced in dbSNP. As discussed in *Additional File 1*, the results are roughly the same between the two datasets.

As explained in *Materials and methods*, two distinct simulated datasets were used as case studies in this paper. The first dataset, of reduced size and low complexity, was generated from the genome sequence of a 5 MB bacterium *Syntrophobacter*. The second dataset, of increased size and moderate complexity, was generated from the full human chromosome 1 sequence. These results enabled a discussion on the erroneous and missing calls and on what triggers them.

Additionally, we checked the impact of an increasing number of datasets on the DISCOSNP global results. This was followed by a comparative study with the other reference-free existing tools, as well as with the assembly+mapping-based approaches. Results described in this section were obtained with $k = 31$, a minimal coverage of 10 (k-mers seen less than 10 times are discarded).

Recall and precision

On the bacterial dataset, DISCOSNP results reach high recall and precision: 99.7% of simulated SNPs were recovered, and only 1.7% of predicted SNPs are false positives (98.3% precision). As expected, when the complexity and the repeat content of the input sequenced genomes increase, DISCOSNP makes more erroneous calls and misses more real SNPs. However, on the human chromosome 1 dataset, DISCOSNP achieves a good compromise between recall (86.6%) and precision (90.5%).

Moreover, as DISCOSNP can be applied on any number of input datasets, we checked whether the global quality is affected when increasing the number of datasets. For this purpose we simulated 10 read sets, where each set corresponds to a simulated muted individual of a bacterial species. We observed that recall slightly decreases with the number of datasets but stays above 98% even for the 10 datasets compared simultaneously. On the other hand, DISCOSNP is more specific as the number of datasets increases, reaching a precision of 99.7% for 10 datasets (Figure 2). As detailed in a section below, this experiment also showed that the memory fingerprint was very similar for 1 to 10 datasets, in contrast with other methods.

Finally, DISCOSNP proves to be robust with respect to its two main parameters: the size of the k-mers and the minimal coverage. Indeed (see Additional File 1), the precision and the recall are only slightly affected by the choice of these values, as long as they are consistent with the data (read length and approximate coverage).

False positives and false negatives are mainly due to repeats.

Based on the results of DISCOSNP on the human chromosome 1 dataset, this section proposes a closer examination of the false positives and false negatives and an overview of the possibilities offered by DISCOSNP to filter them out.

Among the 19,405 false positives, 17,165 (88.5%) were inexact repeats of size at least $2k-1$, with both paths of the bubble mapping exactly the non-muted chromosome. The remaining ones seemed to also be repeat-associated, since their coverage was greater than 10 for all paths in all conditions.

We designed mechanisms to optionally filter out likely false positives. DISCOSNP ranks the predictions according to their read coverage in each condition, favoring SNPs that are discriminant between conditions (Phi coefficient). Indeed, bubbles generated by genomic inexact repeats have similar read coverage for both paths in all conditions and are poorly ranked. In fact, we observed that false positives have significantly lower Phi coefficients than true positives (median of 0.040 versus 0.93), as shown on Figure 3(a). Consequently, when ranking the predictions, the false positive rate increases only after having reached almost all true positives (see also the precision-recall curve Figure 4). An analysis of the Phi coefficients showed that filtering out predictions with a Phi coefficient lower than 0.2 enables to remove 99% of the false positives at a small cost, (loosing less than 0.5% of the true positives), thus achieving a precision of 99.9% for a recall of 86.3%.

Note that even though the Phi coefficient is maximal for discriminant homozygous SNPs (see Methods), we confirmed that it still enables to discriminate false from true positives even for SNPs that are heterozygous inside one dataset. To this end, we simulated a heterozygous individual by sampling reads equally from both the initial human chromosome and the mutated one. We show here that even if the average Phi coefficient is smaller for heterozygous SNPs (0.51 vs 0.90 for homozygous SNPs) it still enables to discriminate false from true positives (Figure 3(b)).

Each sequence in the multi-fasta file output by DISCOSNP is composed of the $2k-1$ nucleotides, which correspond to the bubble, together with its left and right contigs that are extending the $2k-1$ sequences. For each such contig, the length of the longest unambiguous context (longest non-branching path in the graph starting from the bubble) is indicated. We call such path an *unambiguous extension*. Long unambiguous extensions reveal SNPs that are isolated from other polymorphisms, while short unambiguous extensions reveal repeated regions or regions with high densities of polymorphisms. Thus, alternatively to the Phi based filtering, one can also filter out repeat induced bubbles by using the left and right unambiguous extension sizes, since false positives have smaller unambiguous extension sizes (median size of 27 bp versus 796 bp for true positives (Additionnal file Fig XX5).

Finally, high copy number repeats typically yield complex bubbles, which may combinatorially increase the number of false positives. To limit this effect, by default, DISCOSNP filters out branching bubbles (see Methods). This filter generates a decrease in the recall from XXX99.8% to 86.6%; however, in the absence of it, the false positives amount is going up dramatically, e.g. from ≈XXX174 million generated bubbles, only XXX211,492 among them are true positives SNPs, i.e. the use of this filter raises the precision from XXX0.12% to 90.5%. Thus, the large gain in precision is counter-balanced by a small loss in recall. This indicates that, while the vast majority of branching bubbles correspond to false positives, some do correspond to real SNPs. When examining the SNPs that were missed due to this filter (false negatives), we observed that a large majority of them had been simulated inside transposable elements or other types of highly repeated regions.

In practice, 67% fall in regions of chromosome 1 with a mappability greater than 10, i.e. a read coming from this locus would map to at least 9 other loci [12]. Note that the branching bubbles filter applied by DISCOSNP is different of the strategies employed by other de-novo SNP callers (see Methods). As this filter proves to be crucial for obtaining a good precision/recall compromise, such differences in the filtering approach explain why DISCOSNP obtains better precision-recall results than the other tools.

DISCOSNP *is faster, uses less memory and provides better quality results than other* de novo *SNP calling tools*

In this section, we discuss the results obtained by DISCOSNP compared to the other reference-free SNP calling tools, i.e. KISSNP, BUBBLEPARSE and NIKS, on the bacterial and human datasets.

In order to make comparisons as fair as possible, we used identical values for the parameters shared between the four tools: the k-mer size was fixed to 31 and the minimal coverage threshold to 10. Note that all four tools are based on *de Bruijn graphs*, thus using the same k-mer size and the same minimal coverage is a fair way to compare results.

The results in terms of precision and recall are presented in Table 1. To summarize, DISCOSNP always outperforms the other tools, except on the bacterial dataset where KISSNP has a better precision. Note that on this dataset, due to its stringency for finding only homozygous SNPs, KISSNP has a slightly better precision than other tools. This comes at the price of lower recall and much higher running time (15 times slower than DISCOSNP) and memory footprint (200 times more than DISCOSNP), which makes this method unsuitable on bigger data.

On the human dataset, DISCOSNP gets better quality results than BUBBLEPARSE, while using much less memory and time (see also the precision-recall curve in Figure 4). Most certainly, DISCOSNP improved results in terms of precision/recall are due to its novel approach to filter branching bubbles. Compared to BUBBLEPARSE, which filters bubbles that accumulate more than a fixed amount of branchings on their paths, DISCOSNP discards a bubble if, at some point, the two paths of the bubble can be extended simultaneously with more than one nucleotide, i.e. generate an additional bubble. Intuitively, BUBBLEPARSE discards branching bubbles regardless of the type of the branchings, while DISCOSNP labels a bubble as being branching and discards it, if it overlaps with other bubbles. This allows us to filter false positives in repeated regions, while keeping our recall values high.

In all the tests we performed, DISCOSNP was faster and used at least 100 times less memory than existing tools (Figures 5(a) and 5(b)). For instance, even on datasets of 250 million reads from human chromosome 1, only BUBBLEPARSE and DISCOSNP managed to complete successfully, consuming respectively 105 and 1 GB of memory, whereas the other methods were terminated when they exceeded 512 GB of memory.

We also conducted experiments with an increasing number of input read sets, from 2 to 10. As only DISCOSNP and BUBBLEPARSE were designed to deal with more than two datasets, experiments were carried out in parallel with these tools. However, BUBBLEPARSE was not completed successfully either because it did not manage to

build the graph, as it exceeded the maximum available memory (512GB) for more than 8 sets, or because computations were stopped due to errors engendered by the SNP discovery module. Indeed, it appears (personal communication with R. Leggett) that BUBBLEPARSE is not stable while dealing with more than 2 datasets. With respect to DISCOSNP performances, this study enabled reaching two main conclusions. First, it empirically showed that the time and the memory needed by DISCOSNP increases linearly with the quantity of input reads (see Figure XX4 Additional File 1). Second, redundant sequences between read sets do not raise the memory consumption of DISCOSNP: the memory requirement was only raised from $\approx$ 8MB for analyzing reads from 2 individuals to $\approx$ 9MB for 10 individuals.

Detecting SNPs de novo with DISCOSNP *works better than the classical assembly+mapping approach*

When no reference genome is available, a frequently used approach consists in first creating a draft reference through de novo assembly, and then running a reference-based SNP detection pipeline (e.g. read mapping and SNP calling with GATK or samtools). We reproduced this approach by pipelining SOAPDENOVO2 [13] (assembly) with BOWTIE2 [14] (mapping of the reads), finally followed by GATK [1] (variant calling). On the human dataset, DISCOSNP achieves a recall of 86.6% while the mapping+assembly pipeline achieves a recall of 85.6% only. The recall difference is therefore somehow negligible. However, the precision difference is remarkable: assembly+mapping achieves a 63.0% precision compared to 90.5% for DISCOSNP.

As Figure 4 shows, when using the DISCOSNP ranking, the false positive rate increases only after having reached almost all true positives. This is not the case for the assembly+mapping approach where false positives are not clearly separated from true positives. Indeed, many of the best ranked SNPs of GATK are false positives, i.e. among the $4,000$ best ranked SNPs, $3,850$ are false positives.

When looking more closely at these best ranked false positives, we note that they show significantly higher read coverage than the true positives (all but one have a total read coverage greater than twice the expected coverage). This suggests that they are due to repeated sequences. Normally, GATK is not supposed to predict SNPs inside repeats, since reads having multiple mapping positions are discarded. In the case of draft de novo assembly however, numerous repeated sequences are merged into single contigs and reads are then mapped uniquely, thus preventing GATK to detect them as repeats. Conversely, when repeated sequences are well assembled, we expect GATK not to find all the SNPs located inside these sequences. This observation is comforted by the results, as we observe that DISCOSNP is able to detect 38% of SNPs simulated inside repeated regions, whereas GATK finds only 28% of these.

Last but not least, conducting experiments with DISCOSNP is much easier than applying an assembly+mapping approach. First, our solution avoids handling multiple large files as *sam/bam* and *vcf*. Second, DISCOSNP consumes much less time and memory. For example, on the human data, DISCOSNP needed less than 7 hours and 1.2 GB of memory , while the full assembly+mapping approach needed 31 hours and 69 GB of memory.

Results on real data: detection of SNPs from two mouse strains

To analyze the behavior of DISCOSNP when applied on real data, we performed several tests on real sequencing data coming from two mouse strains. More precisely, we focused on the detection of SNPs between two publicly available sets of reads produced with the Illumina Genome Analyzer II. The first one was generated from the *FVB/NJ* mouse inbred strain, while the second one was generated from the *C57BL/6NJ* reference line. We applied DISCOSNP on these two read sets, and compared results to those detected by Wong *et al.* study [15] in which reads from the *FVB/NJ* strain were mapped on the *C57BL/6J* reference sequence. Such mapping approach enabled the detection of 4.1 million homozygous SNPs, amongst which we identified $1,967,755$ *genotypable* SNPs. From this point, we refer to this set of genotypable SNPs as *GS*.

Using $k = 31$ and a minimal coverage of 5, DISCOSNP detects $2,065,833$ SNPs. From these SNPs, 79.1% were also present in *GS*, while 74.8% of the SNPs in *GS* were also found by DISCOSNP. Among SNPs that were found by DISCOSNP only, 46.9% are annotated as a SNP in dbSNP [16].

Moreover, as the Wong *et al.* study was performed on inbred strains, only homozygous SNPs are expected. Therefore, as in the Wong *et al.* study, we filtered out heterozygous SNPs from DISCOSNP results by suppressing SNPs whose read coverage was bigger or equal to minimal coverage (5) in both alleles, for at least one of the two strains. In this context, 61.7% of the SNPs in *GS* were recovered by DISCOSNP, while inversely, 95.1% of *non heterozygous* SNPs found by DISCOSNP were also present in *GS*.

It is worth stressing that calling SNPs with mapping-based approaches depends greatly on the reference quality. Moreover, we underline the fact that the results of Wong *et al.* were obtained by running a complex pipeline, involving 6 distinct tools, followed by a filtering step composed of 14 non-automated filters (coverage, quality, heterozygosity, etc.). On the other hand, the DISCOSNP reference-free approach calls SNPs without needing of any third-party tool.

Finally, this large-scale study showed that DISCOSNP behaves remarkably well on big amounts of data. Indeed, applied on the 2.88 billion mouse reads, the KISSNP2 module needed less than 34 hours and used 4.5 GB of memory for calling SNPs. The second module, which assesses the average quality and coverage of the results, took 78.5 hours and 5.7 of GB memory. These results highlight that, even for large-scale studies, DISCOSNP does not require significant computational resources.

Use case example with experimental validation on SNPs in the *Ixodes ricinus* genome

In order to further validate DISCOSNP results, we conducted a study on real data, including an experimental validation on SNPs selected from predictions. This was part of a population genetic study on the tick species *Ixodes ricinus*, which is, in Europe, the main vector species of human or animal vector-borne diseases. Given the stake of tick-borne diseases in public health [17, 18], it is necessary to have an accurate description of the genetic variability within and among populations of ticks, with the aim of developing efficient control methods against this vector. To this end, highly resolutive genetic markers, like SNPs, provide particularly valuable information to estimate genetic variability and also to estimate the dispersal and genetic structure of tick populations.

No genomic resources, nor a reference genome, were available until now for this species. This study fits a case in which DISCOSNP is useful as i/ sequenced material exists but no reference genome is available and 2/ one is interested in detecting a small set of highly confident heterozygous SNPs. Therefore DISCOSNP was applied on a 454 read set obtained from pooling and sequencing of several tick individuals isolated from natural populations [19].

DISCOSNP detected 321,088 SNPs of which 384 were selected, according to their minimal and maximal coverage and context sequences for experimental validation (see online Methods). Note that as in this context there is no need to discriminate SNPs between conditions, the Phi coefficient was not used. Primers were designed for each selected SNP and 464 individuals were then genotyped for these 384 SNPs using the *Fluidigm* technology. Among them, 368 SNPs (95.8%) were retrieved with a MAF (Minor Allele Frequency) varying between 0.04 and 0.5, with a mean value of 0.23. Of the remaining 16 SNPs, 5 SNPs were not amplified and 11 presented only one of the two alleles.

Discussion

An important part of the presented results were obtained on simulated data as it is actually the only way to control exactly the results quality. We paid very close attention to generate realistic simulated dataset. The studies presented in this paper show that DISCOSNP obtains high-quality results for finding genotypable SNPs on both simulated data (as complex as human) and real data (mouse and tick genomes).

Regarding the other *de novo* SNP detection tools, DISCOSNP presents significant advantageous novelties. It is easy to use: the input data is simply an ordered list of any number of read set(s) in *fasta* or *fastq* format, compressed or not, while the output is a *fasta* file containing a list of SNPs together with their coverage and average *phred* quality. Moreover, the two main parameters (k-mer size and minimal coverage threshold) have only few consequences on the results quality (see Additional File 1-XXX), as long as they do not have clearly extreme values and they remain consistent with the data type (read length and approximate coverage). This robustness is the second main advantage of this new method. Moreover, the rank associated to each SNP enables to clearly distinguish between true positives and the few false positives. Such a feature is precious in any biological study where one is interested by finding not all, but only high confidence SNPs. For instance, our results on human data show that, by filtering out lowly ranked SNPs, one achieves XXX86.3% recall for a XXX99.9% precision.

Another essential advantage of our new method lies in its extremely low memory usage. For instance, in the previously presented mouse study framework, nearly 3 billion reads of size 100 were analyzed by DISCOSNP, using at most 5.7 GB of memory. Such memory performances are not achieved at the expense of prohibitive running times, as DISCOSNP remains significantly faster than all other known de novo SNP detection tools. Such performances make DISCOSNP usable on standard desktop computers usually present in any biological lab. It is worth noticing that DISCOSNP uses at least 100 times less memory than other published de novo SNP detection tools. Only DISCOSNP could perform on very large read sets composed of dozens of billions of reads. Thus, even without having access to computational resources with hundreds of GB of memory, studies that were not feasible with known

de novo SNP detection tools become possible if using DISCOSNP. Moreover and contrary to other de novo SNP calling tools, as DISCOSNP performances are not impacted by redundant sequences between read sets, its memory and time requirements are only slightly affected by the number of input read sets. For instance, DISCOSNP was the only tool able to compute SNPs for 10 bacterial individuals.

One of the main limitations of this approach lies in the fact that it produces SNPs, but no genomic location for each SNP. However, in numerous biological applications the localization of polymorphism is not required. For instance, DISCOSNP can be applied to identify SNPs associated to some phenotypic traits or diseases, or alternatively to estimate or compare the genetic diversity among or between natural populations. Sequences obtained around those SNPs of interest can then be genotyped at larger population scales with standard genotyping technologies or used for diagnostic assays. This was actually the case for the tick study, where natural populations were genotyped to characterize their reproductive mode (level of inbreeding) and estimate the gene flow within and among populations at various spatial scales. Moreover, the SNPs isolated thanks to DISCOSNP are currently used to build a genetic map based on the analysis of the segregation of parental alleles in the offspring of several controlled crosses.

A natural future development will consists in integrating de novo assembly with de novo SNPs and other polymorphism detection tools. This solution would unite the power of both approaches, facilitating the assembly by tackling the polymorphism problem and by conserving the recall and precision performances of methods such as DISCOSNP. This idea would lead to assembled genomes represented no more as "simple" linear sequences but as graphs such as suggested by the fastg format `fastg.sourceforge.net/`, in which polymorphisms are conserved. Such a change would open the way to new possibilities of storage and use of polymorphisms.

Conclusions

The proposed method, implemented as the DISCOSNP tool, is to our knowledge the only reference-free approach which i/ proposes the highest-quality (precision and recall) "*genotypable*" SNPs with robust ranking, ii/ can be applied on any number of read sets and iii/ scales-up big data studies thanks to an extremely memory-efficient strategy while remaining faster than other reference free SNP detection tools.

The studies presented in this paper show that DISCOSNP obtains high-quality results for finding genotypable SNPs on both simulated data (as complex as human) and real data (mouse and tick genomes).

Additionally, this study had shown that DISCOSNP gives better results than a state-of-the-art assembly and mapping approach, both in terms of time and memory usage and in terms of quality, by achieving better recall and precision. Such quality results are explained by the fact that assembly+mapping approaches cumulate assembly and mapping imprecisions, thus being highly sensitive to repeats, while DISCOSNP can better discriminate SNPs from inexact repeats and is able to find SNPs in repeated regions.

The DISCOSNP source code, available under CeCILL license, can be downloaded from `colibread.inria.fr/discoSnp/`. Moreover, this web page shows how to integrate DISCOSNP in any galaxy instance using the *GenOuest* tool shed.

Materials and methods

Algorithms

DISCOSNP is composed of two main modules, KISSNP2 and KISSREADS. KISSNP2 detects putative SNPs from one or more set(s) of reads. KISSREADS evaluates the coverage and base quality of the SNPs per read set and ranks them accordingly.

KISSNP2 *module*

Similarly to CORTEX_VAR and KISSNP, the KISSNP2 module is based on a *de Bruijn graph*. A *de Bruijn graph* is a directed graph that contains all the k-mers present in a read dataset as nodes, and all the possible $(k-1)$-overlaps as edges. Such graphs have been widely used in *de novo* assembly [20]. The idea is the following: if a data set contains two sequences that are identical, except for one character, then these *polymorphic* sequences generate a *bubble* in the graph (see an example in Figure 1). Formally, a *bubble* in a *de Bruijn graph* is composed of two distinct paths of $k+2$ nodes, having the start and the end nodes in common. Precisely, KISSNP2 detects couples of paths, says p_1 and p_2, each of length $2k-1$ that are the polymorphic sequences, i.e. sequences of bubbles excepting the two extreme nodes. Formally this two paths can be written as $p_1 = p\alpha q$ and $p_2 = p\beta q$, with p, q being $(k-1)$-mers and with $\alpha \neq \beta$.

In KISSNP2, this idea is exploited by generating the *de Bruijn graph* of all input dataset(s) pooled together, and by searching the previously described bubbles in the graph. For efficiency reasons, the *de Bruijn graph* is not explicitly created. Instead, all k-mers are stored thanks to an exact data-structure [10] based on a Bloom filter. This data-structure enables to implicitly walk the *de Bruijn graph* and thus, to construct sequences longer than k characters.

In the following, we say that a k-mer ω can be *right extended* with a nucleotide α if the k-mer obtained by concatenating the suffix of length $k-1$ of ω with α, exists in the reads. Symmetrically, we say that a k-mer ω can be *left extended* with a nucleotide α, if α concatenated with the prefix of length $k-1$ of ω forms a k-mer existing in the reads.

The KISSNP2 algorithm detects all k-mers that can be right extended with at least two distinct nucleotides. Such k-mers are the starting nodes of the bubbles they engender, as in the example of Figure 1 where the starting k-mers are $CTGA$ and $CTGT$. Then, for each such couple of k-mers starting with the same $k-1$ length prefix, KISSNP2 constructs a bubble by performing successively $k-1$ right extensions on both paths with the same nucleotide (using successively nucleotides C,C and T for the example in Figure 1). If, at one step, both paths cannot be right extended by the same nucleotide, then the bubble is discarded.

Branching bubbles If, during a right extension step, two (or more) distinct nucleotides may be used to extend right both paths, then the bubble is considered as *branching*. Symmetrically if two (or more) distinct nucleotides may be used to extend left both paths, the bubble is also considered as branching. Depending on the user requests, branching bubbles may be filtered out. Note that this branching filter enables to keep bubbles containing branching nodes.

Retrieve sequence contexts For each bubble that we find, we retrieve its sequence contexts from both sides. This is done by iteratively right (respectively left) extending the last (respectively first) k-mer of the bubble, as long as no branching is met in the graph. Thus, we generate the two *unitigs* surrounding the polymorphic sequences.

Minimal k-mer coverage Sequencing errors generate k-mers with a lower coverage than the expected sequencing depth. Thus, k-mers whose support is below a user-defined threshold are not used to construct the *de Bruijn graph* during the KisSnp2 phase. Note that even if rare k-mers are discarded during the bubble detection phase, this does not impact the coverage and quality of the computations made by KissReads. Indeed, as substitutions are authorized during read mapping, reads containing sequencing errors are hopefully correctly mapped.

KisSnp2 ***output*** The KisSnp2 output is a *multi-fasta* file, in which every consecutive couple of sequences corresponds to the two paths of a bubble, surrounded by the left and the right extensions.

KissReads *module*

Based only on a raw *de Bruijn graph*, KisSnp2 cannot output read coverage, nor read quality information. Moreover, as all *de Bruijn graph*-based methods, chimeric sequences may be constructed due to overlapping k-mers never present in the same reads. These sequences are said to be *non-read-coherent*. The KissReads module aims at filtering out bubbles composed of *non-read-coherent* sequences, as well as adding coverage and quality information on the remaining ones, and finally at ranking SNPs (see Section).

Given a sequence s and a set of reads $\mathcal{R}$ mapped on s, we say that s is *read-coherent* if, for each position of s, at least c reads are mapped, with c a user-defined threshold. Note that reads are mapped in a semi-global manner: mapped reads may have a prefix starting before the first position of s and/or may have a suffix ending after the last position of s. By default, one substitution is authorized, fitting actual sequencing features. Moreover, knowing that in the discoSNP framework the sequence s represents one of the two alleles of a SNP, no substitution is authorized on the polymorphic position during the mapping.

Moreover, it may appear that a position of s is mapped only by the *end* of reads (last k positions), and/or only by the *beginning* of reads (first k positions). This reveals a situation where part of the sequence s was created only by k-mers not belonging to the mapped reads, because of a repeat of length bigger or equal to $2k$ and smaller than the read size. Such a sequence is thus chimeric. An example of this situation is illustrated in Figure 6. To overcome this problem, we define the *k-read-coherency*. Given a sequence s and a set of reads $\mathcal{R}$ that can be mapped on s, we consider s as *k-read-coherent* if, for each position i of s except the last $k-1$

ones, there exists at least c mapped reads fully covering the k-mer starting position i. Note that this condition is fully symmetrical, whether s is read in forward of reverse complement strand.

Given a set of sequences $\mathcal{S}$ generated by KissSnp2 and the initial sets of reads, the KissReads algorithm maps the reads (using a classical seed-and-extend approach) on the sequences of $\mathcal{S}$. Once all reads are mapped on all sequences of $\mathcal{S}$, the *k-read-coherency* is computed for each read set and for each sequence of $\mathcal{S}$. Sequences of $\mathcal{S}$ for which at least one read set makes them *k-read-coherent* are conserved. KissReads outputs them together with their read depth per read set and with the average *phred* quality of the polymorphic nucleotide per read set.

Finally the read depth per read is used to rank the SNPs as described Section .

Ranking SNPs

Depending on the application framework, many distinct ranking possibilities can be implemented using the information provided with each bubble found. By default, discoSNP scores each SNP according to the coverage repartition of its alternative paths between the conditions. For a given SNP, the score is the *Phi coefficient* of the table of read counts for each path and each dataset, computed as follows: $\sqrt{\frac{\chi^2}{n}}$. It can be seen as a normalized Chi-squared statistics that varies between 0 and 1. A high score, close to 1, is obtained if the frequencies of the paths are very different between datasets, the best case being for homozygous SNPs between two datasets, where each path is strictly specific to one dataset. Notably, this score ranks poorly bubbles due to sequencing errors and inexact repeats as they are likely to have similar repartitions in all datasets (small frequency for an error, and equal frequency for repeats). Moreover, the normalization prevents from over-scoring highly covered bubbles which are often due to repeats.

Simulation and evaluation

Recall and precision computations

For the tests on simulated data, we provide recall and precision measures. These measures are computed as follows. The SNPs simulation provides a reference SNP list that is an exhaustive and exact list of *genotypable* SNPs to be found (at a distance of at least k nucleotides from the other SNPs).

We call *true positives* (TP), SNPs in the reference list, also found by discoSNP; we call *false positives* (FP), SNPs that are found only by discoSNP, and we call *false negatives* (FN), SNPs in the reference list only. Finally, the *precision* is computed as the number of TP divided by the total number of SNPs found by discoSNP, while the *recall* is given as the number of TP divided by the total number of SNPs in the reference list.

Protocol for experiments on SNP detection between two simulated datasets:

- Selection of an initial genomic sequence, S_i. In this paper, we present results based on the use of two distinct data sources:
  - Bacterial *Syntrophobacter fumaroxidans* MPOB chromosome, ≈ 5 million base pairs;
  - Human genome, hg19 chomosome 1, ≈ 249 million base pairs, which is highly repeated.
- Creation of a mutated copy, S_m, of the initial sequence. The initial sequence and the mutated copy differ by 0.1% of uniformly distributed substitutions, corresponding to homozygous SNPs. A *multi-fasta* file $ref\_snps.fa$ formatted as the DISCOSNP output and containing all the generated SNPs is produced for subsequent use as a reference list to compute precision and recall.
- Sequencing simulation by sampling equal-length reads (100 base pairs) on both the initial and the mutated sequences, with a uniform probability distribution, on a $50x$ coverage basis. Substitution errors are uniformly distributed along each read with a fixed probability (error rate). More sophisticated read simulators were also tested (with more complex error profiles) giving highly similar results (see Additional File 1).
- Running DISCOSNP (and/or eventually another compared tool) together on the reads produced from the initial and from the mutated sequence.
- Comparison using the GASSST [21] mapping tool, with 100% identity of DISCOSNP results against the $ref\_snps.fa$ file.
- Based on the GASSST mapping results, compute the DISCOSNP ability (precision and recall) to detect homozygous SNPs.

Protocol for experiments on SNP detection between more than two simulated datasets

- Selection of the *Syntrophobacter fumaroxidans* MPOB chromosome as the initial sequence.
- Creation of 10 mutated copies $\{Sm_1, Sm_2, \ldots, Sm_{10}\}$ of the initial sequence. The initial sequence and each muted copy Sm_j differ by 0.05% of uniformly distributed substitutions, corresponding to homozygous SNPs. For each Sm_j a *multi-fasta* file $ref\_snps_j.fa$ formatted as the DISCOSNP output and containing all the SNPs generated in sequence Sm_j is produced.
- Sequencing simulation on each Sm_j sequence, as previously described.
- Running DISCOSNP on $\{Sm_1, Sm_2\}$, then on $\{Sm_1, Sm_2, Sm_3\}$, and so forth up to $\{Sm_1, \ldots, Sm_{10}\}$.
- Comparison using GASSST [21] with 100% identity of each DISCOSNP result against the corresponding subset of $ref\_snps_j.fa$ concatenated files.
- Using the GASSST mapping results, compute the DISCOSNP ability (precision and recall) to detecting SNPs.

Comparing DISCOSNP *with other de novo SNP calling methods*

In Section , we compared DISCOSNP to three other de novo SNP calling methods: KISSNP, NIKS and BUBBLEPARSE. In order to make comparisons as fair as possible,

we used identical values for parameters that are shared between the four tools: the k-mer size was fixed to 31 and the minimal coverage threshold to 10. The rest of the parameters of KISSNP were tuned based on DISCOSNP, as they are common between both tools. In the case of BUBBLEPARSE and NIKS we paid particular attention to using the best fitted values respectively. Concretely, except the memory-related parameters that needed to be adjusted in order to satisfy the memory requirements, BUBBLEPARSE was run, based on the recommendations in [6], with additional graph-cleaning parameters and a search depth of 1. Note that we also performed same tests using BUBBLEPARSE with depth 2, leading to a recall/precision respectively of 85.6%/84.3% instead of 82.7%/89.9% with depth 1. We chose to present results using BUBBLEPARSE with depth 1, that seems to be a reasonable trade-off between recall and precision. NIKS, which is based on JELLYFISH [22] and VELVET [20], was run with default parameters that were adjusted for $k = 31$, as described in [7]. Moreover, both BUBBLEPARSE and NIKS required manual intervention to reformat the input data.

Finding SNPs by assembly and mapping

For comparing DISCOSNP to the assembly+mapping approach, we chose to use state-of-the-art *de novo* assembly, mapping and variant caller tools, respectively: SOAPDENOVO2 [13], BOWTIE2 [14] and GATK [1]. This study was performed on the human chromosome 1 dataset previously mentioned. SOAPDENOVO2 was used with $k = 31$ and default values for the other parameters.

For assembling a reference sequence with SOAPDENOVO2, reads produced from the initial sequence were used. However, in order to detect all kinds of SNPs, including potential heterozygous SNPs in the initial sequence, both sets of reads (generated from the mutated sequence, as well as from the initial sequence) were mapped on the reference sequence with BOWTIE2. In the end, GATK was used to call the SNPs from the *bam* files that were produced. The precision vs recall computation was carried-out following the same protocol as the one previously described.

Mouse dataset

In a study by Wong *et al.* [15], reads from *FVB/NJ* mouse strain sequence data, among others, were mapped on the *C57BL/6J* mouse reference genome (NCBIM37). This approach enabled to discover approximately 4.3 million SNPs between *FVB/NJ* mouse strain and *C57BL/6J* reference genome. From this catalogue of 4.3 million SNPs, we kept only *genotypable* ones, i.e. SNPs that are distant of at least k base pairs (31 in our case) from another variant.

To this end, we used the full *vcf* file (mgp.v3.snps.rsIDdbSNPv137) provided by the *Mouse Genomes Project* (`www.sanger.ac.uk/resources/mouse/genomes/`). As this file concerns 18 distinct strains, we conserved only those SNPs with distinct alleles for the strains *C57BL6/NJ* and *FVB/NJ*, and we removed the *non-genotypable* ones. Thus, we conserved $2,181,297$ SNPs that were formatted in a DISCOSNP output like format, thus obtaining a reference SNP list.

Next, we downloaded reads from the European Nucleotide Archive (`www.ebi.ac.uk/ena/`), for the *C57BL/6NJ* strain (ERP000041, 987 million reads) and for the *FVB/NJ* strain (ERP000687, $1,888$ million reads). DISCOSNP was applied on

these two read sets with $k = 31$ and a minimal coverage value of 5. Applied on these ≈ 2.88 billion reads, DISCOSNP found $2,065,833$ SNPs.

We compared the DISCOSNP results to the Wong *et al.* results, by mapping DISCOSNP results on the reference SNP list. The mapping results enabled to compute the number of SNPs that were discovered by both approaches.

Additionally, we analyzed the SNPs that were only found by DISCOSNP. To this end we downloaded the full mouse *dbSNP* from (hgdownload.soe.ucsc.edu/goldenPath/mm10/database/snp137.txt.gz), from which we generated a *multi-fasta* file containing all polymorphic annotated sequences. Among SNPs predicted only by DISCOSNP, those whose both paths map exactly this file was considered as annotated in *dbSNP*.

Tick dataset

DISCOSNP was applied on real sets of reads as part of a population genetic study of ticks *Ixodes ricinus* [19], a species for which no reference genome is available (expected genome size: 2.1 GB). DNA was extracted from two tick pools, one composed of 10 individuals from Gardouch (close to Toulouse), France, and another composed of 20 individuals from Malville (close to Nantes), France. We applied a genomic reduction on these two pools, conserving 3.8% of the initial genome. The DNA was sequenced by 454 Roche pyrosequencing, leading to the generation of $1,389,201$ reads in two libraries ($730,482$ for one and $658,719$ for the second). After quality trimming, a total of $996,508$ reads ($536,061$ for the first pool and $460,447$ for the second) with an average length of 529 bp were used for analysis with DISCOSNP, detecting $321,088$ SNPs.

Detected SNPs were then selected for experimental validation using the following criteria : (1) SNPs with a coverage between 4 and 10 ($126,567$ SNPs) were selected to avoid sequencing errors and repeated sequences, highly frequent in ticks (66% of the genome is repeated [23]), (2) in order to be able to design efficient primers, SNPs had to be distant from homopolymers and other variants (3) SNPs with PHRED sequence quality < 30 were filtered out too. As in this study framework one is not interested by SNPs discriminating the two datasets, we did not use the Phi coefficient for selecting SNPs for experimental validation. Among the 1768 SNPs meeting these criteria, 384 were randomly picked for genotyping validation. This was done by designing primers that are specific to each of the two alleles due to the presence of the SNP.

Primers were designed for each selected SNP. To validate them, we performed a genotyping run using *Fluidigm* technology, where primers are combined with fluorochrome (VIC or FAM for each allele) [24]. Reading the fluorescence allows to determine the genotype of the individual typed at each locus (homozygous XX or YY, heterozygous XY).

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

Author's contributions

PP initiated the work. RU, GR, RC and PP carried out the implementations. CL led the computational experiments. EQ and OP performed the experimental validations on ticks. All authors participated in the simulated data and mouse data experiments. RU, RC, CL and PP wrote the manuscript. All authors read the manuscript and approved its final form.

Acknowledgements

The authors warmly thank Zamin Iqbal and Richard Leggett for interesting discussions and valuable help on using CORTEX_VAR and BUBBLEPARSE tools. We also thank the *GenOuest* (genouest.org) cluster team, who allowed us to perform all the tests. This work was supported by the French ANR-12-BS02-0008 *Colib'read* project and by SOFIPROTEOL under the *PEAPOL* project.

Author details

[1]LaBRI, University Bordeaux 1, Talence, France. [2]INRA, UMR1349 IGEPP, Le Rheu, France. [3]GenScale, INRIA Rennes Bretagne-Atlantique, IRISA, Rennes, France. [4]BAMBOO, INRIA Grenoble Rhone-Alpes, Lyon, France. [5]INRA, UMR1300 Biology, Epidemiology and Risk Analysis in Animal Health, Nantes, France. [6]LUNAM University, Oniris, Nantes Atlantic College of Veterinary Medicine and Food Sciences and Engineering, UMR BioEpAR, Nantes, France. [7]Department of Computer Science and Engineering, The Pennsylvania State University, USA.

References

1. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., *et al.*: A framework for variation discovery and genotyping using next-generation dna sequencing data. Nature genetics **43**(5), 491–498 (2011)
2. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) **25**(16), 2078–9 (2009). doi:10.1093/bioinformatics/btp352
3. Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, İ., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T.R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N.A., Ganapathy, G., Gibbs, R.A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J.B., Ho, I.Y., Howard, J., Hunt, M., Jackman, S.D., Jaffe, D.B., Jarvis, E., Jiang, H., Kazakov, S., Kersey, P.J., Kitzman, J.O., Knight, J.R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., MacCallum, I., MacManes, M.D., Maillet, N., Melnikov, S., Vieira, B.M., Naquin, D., Ning, Z., Otto, T.D., Paten, B., Paulo, O.S., Phillippy, A.M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F.J., Richards, S., Rokhsar, D.S., Ruby, J.G., Scalabrin, S., Schatz, M.C., Schwartz, D.C., Sergushichev, A., Sharpe, T., Shaw, T.I., Shendure, J., Shi, Y., Simpson, J.T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Wang, J., Worley, K.C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., Korf, I.F.: Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. ArXiv e-prints (2013). 1301.5406
4. Lee, H., Schatz, M.C.: Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics (Oxford, England) **28**(16), 2097–105 (2012). doi:10.1093/bioinformatics/bts330
5. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., McVean, G.: De novo assembly and genotyping of variants using colored de bruijn graphs. Nature genetics **44**(2), 226–232 (2012)
6. Leggett, R.M., Ramirez-Gonzalez, R.H., Verweij, W., Kawashima, C.G., Iqbal, Z., Jones, J.D.G., Caccamo, M., MacLean, D.: Identifying and Classifying Trait Linked Polymorphisms in Non-Reference Species by Walking Coloured de Bruijn Graphs. PLoS ONE **8**(3), 60058 (2013). doi:10.1371/journal.pone.0060058
7. Nordström, K.J.V., Albani, M.C., James, G.V., Gutjahr, C., Hartwig, B., Turck, F., Paszkowski, U., Coupland, G., Schneeberger, K.: Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. Nature Biotechnology (October 2012) (2013). doi:10.1038/nbt.2515
8. Sacomoto, G.A., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M.-F., Peterlongo, P., Lacroix, V.: Kissplice: de-novo calling alternative splicing events from rna-seq data. BMC bioinformatics **13**(Suppl 6), 5 (2012)
9. Peterlongo, P., Schnel, N., Pisanti, N., Sagot, M.-F., Lacroix, V.: Identifying snps without a reference genome by comparing raw reads. In: Chavez, E., Lonardi, S. (eds.) String Processing and Information Retrieval. Lecture Notes in Computer Science, vol. 6393, pp. 147–158. Springer, ??? (2010). doi:10.1007/978-3-642-16321-0$_1$4. http://dx.doi.org/10.1007/978-3-642-16321-0$_1$4
10. Chikhi, R., Rizk, G.: Space-efficient and exact de Bruijn graph representation based on a Bloom filter. In: Lecture Notes in Computer Science (ed.) Wabi, vol. 7534, pp. 236–248 (2012). http://www.springerlink.com/index/G031354QQ57464N2.pdf
11. Salikhov, K., Sacomoto, G., Kucherov, G.: Using cascading Bloom filters to improve the memory usage for de Brujin graphs. In: Proc. of the 13th Workshop on Algorithms in Bioinformatics (WABI), 2013. Lecture Notes in Computer Science, vol. 8126. Springer, Sophia Antipolis, France (2013). http://hal.archives-ouvertes.fr/hal-00824697
12. Derrien, T., Estellé, J., Sola, S.M., Knowles, D.G., Raineri, E., Guigó, R., Ribeca, P.: Fast computation and applications of genome mappability. PLoS One **7**(1), 30377 (2012)
13. Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., *et al.*: Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience **1**(1), 18 (2012)
14. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with bowtie 2. Nature methods **9**(4), 357–359 (2012)

15. Wong, K., Bumpstead, S., Van Der Weyden, L., Reinholdt, L.G., Wilming, L.G., Adams, D.J., Keane, T.M.: Sequencing and characterization of the FVB/NJ mouse genome. Genome biology **13**(8), 72 (2012). doi:10.1186/gb-2012-13-8-r72
16. Consortium, M.E., Stamatoyannopoulos, J., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D., Groudine, M., Bender, M., Kaul, R., Canfield, T., Giste, E., Johnson, A., Zhang, M., Balasundaram, G., Byron, R., Roach, V., Sabo, P., Sandstrom, R., Stehling, A.S., Thurman, R., Weissman, S., Cayting, P., Hariharan, M., Lian, J., Cheng, Y., Landt, S., Ma, Z., Wold, B., Dekker, J.: An encyclopedia of mouse dna elements (mouse encode). Genome Biology **13**(8), 418 (2012). doi:10.1186/gb-2012-13-8-418
17. Gubler, D.J.: Resurgent vector-borne diseases as a global health problem. Emerging infectious diseases **4**(3), 442 (1998)
18. Parola, P., Raoult, D.: Tick-borne bacterial diseases emerging in europe. Clinical microbiology and infection **7**(2), 80–83 (2001)
19. Quillery, E., Quenez, O., Peterlongo, P., Plantard, O.: Development of genomic resources for the tick ixodes ricinus: isolation and characterization of single nucleotide polymorphisms. Molecular Ecology Resources, (2013). doi:10.1111/1755-0998.12179
20. Zerbino, D.R., Birney, E.: Velvet: algorithms for de novo short read assembly using de bruijn graphs. Genome research **18**(5), 821–829 (2008)
21. Rizk, G., Lavenier, D.: GASSST: Global Alignment Short Sequence Search Tool. Bioinformatics (Oxford, England) **26**(20), 2534–2540 (2010). doi:10.1093/bioinformatics/btq485
22. Marçais, G., Kingsford, C.: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**(6), 764–770 (2011)
23. Ullmann, A., Lima, C., Guerrero, F., Piesman, J., Black, W.: Genome size and organization in the blacklegged tick, ixodes scapularis and the southern cattle tick, boophilus microplus. Insect molecular biology **14**(2), 217–222 (2005)
24. Wang, J., Lin, M., Crenshaw, A., Hutchinson, A., Hicks, B., Yeager, M., Berndt, S., Huang, W.-Y., Hayes, R., Chanock, S., *et al.*: High-throughput single nucleotide polymorphism genotyping using nanofluidic dynamic arrays. BMC genomics **10**(1), 561 (2009)

Figures

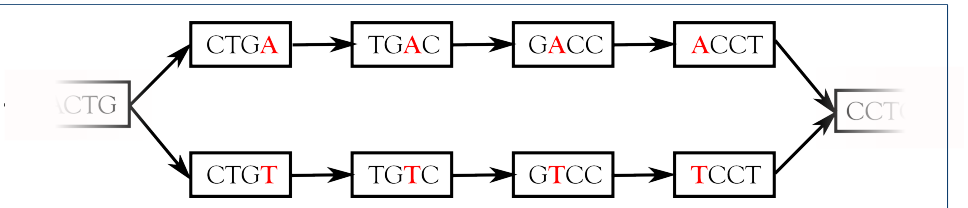


Figure 1 Toy example of a Bubble in the de Bruijn Graph ($k = 4$). Bubble caused by a single nucleotide polymorphism. The two polymorphic sequences are $\ldots CTGACCT \ldots$ and $\ldots CTGTCCT \ldots$

Tables

	Bacterial dataset		Human chr1 dataset	
	Recall (%)	Precision (%)	Recall (%)	Precision (%)
NIKS	90.3	98.5	N/A	N/A
KISSNP	93.2	**99.9**	N/A	N/A
BUBBLEPARSE	98.7	97.5	82.7	89.9
DISCOSNP	**99.7**	98.3	**86.6**	**90.5**

Table 1 Precision and recall results on simulated datasets. The results on the human data are not available (N/A) for NIKS and KISSNP, as these tools did not scale up on this dataset when using a 512GB machine.

Additional Files

Additional data file 1

This file Provides additional results while varying simulation criteria (SNP density and repartition and sequencing depth) or DISCOSNP parameters (size of k-mers and minimal coverage). A section shows memory and time performance when increasing dataset numbers, and finally a section presents results of DISCOSNP and other reference SNP callers while looking for all simulated SNPs, not limited to genotypable ones.

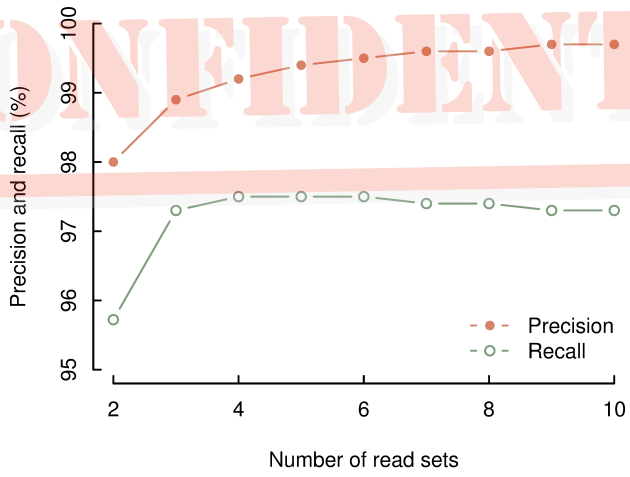


Figure 2 DISCOSNP **behavior when increasing the number of input datasets. XXX refaire avec dernieres valeurs: cf evernote Pierre.** Each read set is composed of 2 million reads of length 100 simulated from a bacterial genome

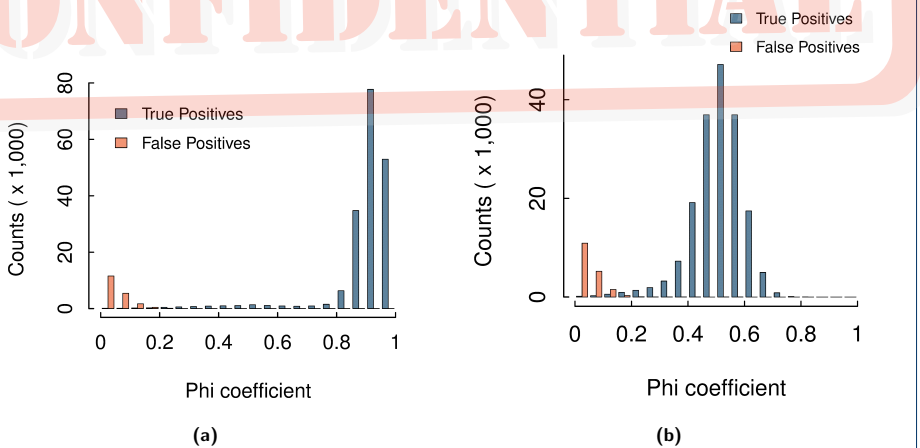


Figure 3 Comparison of Phi coefficient between True and False Positive predictions. Results obtained from human chromosome 1 dataset simulating (a) homozygous and (b) heterozygous SNPs.

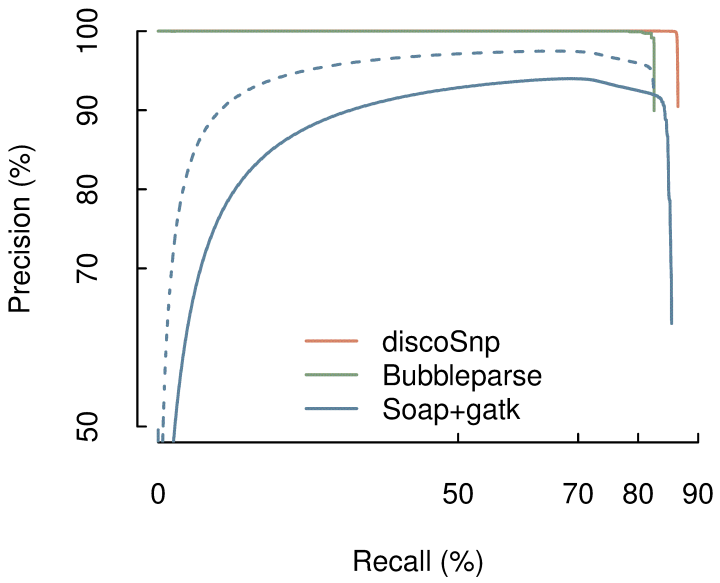


Figure 4 Comparative results of BUBBLEPARSE, DISCOSNP and assembly/mapping approaches. Precision vs recall curves obtained on the human chromosome 1 dataset, on SNPs sorted with respect to their ranks. DISCOSNP is compared to BUBBLEPARSE and to the assembly/mapping approach (SOAPDENOVO+GATK).

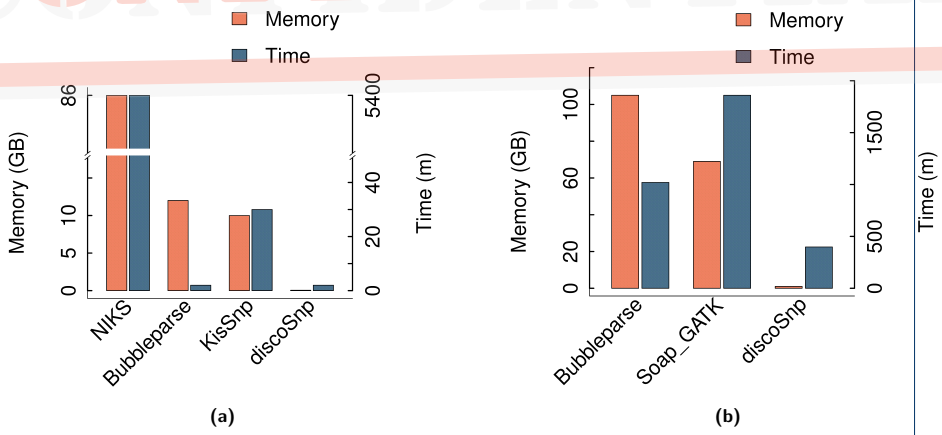


Figure 5 Performances. Comparative memory and time performances both on the bacterial dataset (a) and on the human dataset (b).

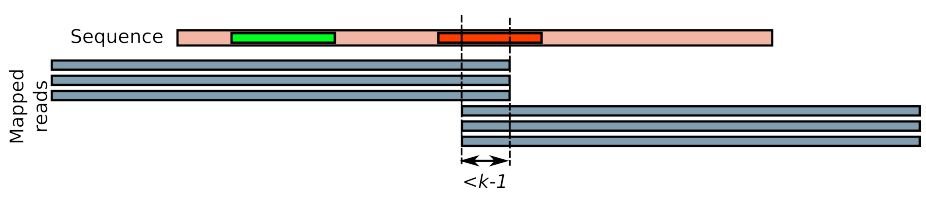


Figure 6 Read-coherency and k-read-coherency example. With coverage threshold=2, schematic example where a sequence is *read-coherent* but not *k-read-coherent*. The leftmost represented k-mer (green) on the sequence is an example where the k-mer starting at this position is covered with three mapped reads. On the other hand, the rightmost represented k-mer (red) is covered with zero mapped read illustrating why the sequence is not *k-read-coherent*.

l'université nantes angers le mans
PÔLE DE RECHERCHE ET D'ENSEIGNEMENT SUPÉRIEUR

Oniris
École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation — Nantes Atlantique
Santé et alimentation au cœur de la vie

Thèse de Doctorat

Elsa QUILLERY

Développement de marqueurs génétiques (SNPs) à partir du génome de la tique *Ixodes ricinus* pour l'étude de la structure génétique de ses populations à l'échelle du paysage

Development of genetic markers (SNPs) from the genome of the tick *Ixodes ricinus* for the landscape genetics study of its populations

Résumé

Ixodes ricinus est la principale espèce vectrice d'agents pathogènes transmises par les tiques en Europe. Une meilleure connaissance de sa variabilité génétique constitue un apport majeur pour le développement de méthodes de lutte (vaccins anti-tiques…). Via les outils de la génétique des populations et la mesure des flux de gènes, elle permet aussi d'estimer la dispersion des tiques et donc une meilleure compréhension de l'épidémiologie de ces maladies vectorielles. Nous avons développé un nouveau type de marqueurs, les Single Nucleotide Polymorphisms en générant par pyroséquençage une quantité importante de séquences (643 millions de nucléotides) du génome d'I. ricinus. En l'absence de génome de référence, nous avons utilisé une suite originale de logiciels bioinformatiques (DiscoSneap) pour isoler 1765 SNPs. Chacun des 384 loci retenus ont été validés par le génotypage de 480 nymphes individuelles collectées dans une zone du nord de l'Ille et Vilaine comprenant un massif forestier de 1000 hectares entouré par du bocage plus ou moins dense. Un déficit en hétérozygotes a été observé, même à l'échelle la plus fine étudiée, indiquant une forte consanguinité qui pourrait être due à la faible dispersion des larves et/ou l'existence de races d'hôtes. L'ensemble des analyses de génétique des populations suggère l'existence d'importants flux de gènes à l'échelle de l'ensemble de la zone d'étude.

Ces marqueurs constituent aussi un apport important pour l'étude de la variabilité génétique de cette tique, depuis l'établissement d'une carte génétique jusqu'à l'identification de gènes du vecteur impliqués dans l'épidémiologie des maladies transmises par les tiques.

Mots clés :

Tique, génétique des populations, single nucleotide polymorphism, NGS, bioinformatique, dispersion, consanguinité,

Abstract

Ixodes ricinus is the main vector species for pathogenic agents transmitted by ticks in Europe. A better knowledge of its genetic variability is especially useful for the design of control methods (against tick vaccine…). Through population genetics and the assessment of gene flow, it also allows to estimate tick dispersal and hence a better understanding of tick-borne diseases. We have developed a new type of molecular markers (Single Nucleotide Polymorphism) by generating by pyrosequencing a large amount of sequencing reads (643 billions of nucleotides) from the I. ricinus genome. Without any reference genome available, we have used a unique bioinformatics pipeline (DiscoSneap) to isolate 1765 SNPs. Each of 384 selected loci have been validated by the genotyping of 480 individual nymphs sampled in an area located in the north of the Bretagne region containing a 1000 hectare forest surrounded by hedges. An heterozygous deficiency was observed, even at the finest spatial scale investigated, indicating a large inbreeding that could be due to the weak dispersal abilities of larvae and/or the existence of host races. All the population genetic analysis conducted suggest a large amount of gene flow within the whole studied area.

Those markers provide an important tool for the investigation of genetic variability in this tick, from the building of a genetic map to the identification of genes of involved in the epidemiology of tick-borne diseases.

Keywords :

ticks, population genetics, single nucleotide polymorphism, NGS, bioinformatics, dispersal, inbreeding
Key Words