État de l'art

L'apport des NGS pour inférer la structure de population

Pierre-Louis Stenger

Table des figures

Mots clés : *NGS, Population, Génétique, Pyroséquençage, RAD-Seq, Illumina, MinION*

1. Évoquer les enjeux de conservation et de protection de la biodiversité (notion d'anthropocène, 6ème crise d'extinction d'espèces, e.g. Pimm et al. 2014, Science)
2. Expliquer en quoi l'inférence de la structure de populations est cruciale pour la conservation. (citer des exemples concrets)

Estimer le taux d'échange d'individus au sein d'une population est un point central de l'écologie de l'évolution, ainsi que de ses applications en conservation des espèces et de leur management().

Regarder les flux de gènes en utilisant des marqueurs neutres permettent de voir les niveaux de différenciation entre les populations. Ces méthodes intègrent les effets des forces évolutives, sauf ceux des mutations, car leurs effets sont négligeable au regard de ceux de la migration ().

La migration peut être indirectement inférée en observant la structure génétique de la population. Par exemple, les proportions génomiques des longs segments qui sont identiques entre individus provenant d'une même ou d'une différente population sont directement liés au taux de migration (**?**). La seconde option pour détecter les événements de dispersion des individus et de trouver les distances de dispersion de la population qui peut être donné à travers l'inférence de généalogie (e.g. assignement parental) ou l'analyse de cluster (de "groupe") ().

Les approches génomiques ont été suggérées comme un outil prometteur pour la pratique de la conservation pouvant améliorer la compréhension de l'inférence génétique et de fournir des nouvelles idées pour la gestion des espèces Shafer *et al.* (2015)Garner *et al.* (2015).

Les données génétiques permettent maintenant la détection de sous-structure d'une population, de mesurer la connectivité génétique, et d'identifier les risques potentiels liés à l'évolution démographique et la consanguinité **?**.

Les approches génétiques ont fait des percées qui influent sur les efforts de conservation comme par exemple, l'augmentation de la population de la panthère de Floride *Felis concolor coryi* qui avaient frôlés l'extinction **?** ou encore la détection de braconnage au sein de populations menacées **?**.

L'utilisation traditionnelle des données génétiques en biologie de la conservation a été historiquement délimité en deux domaines interdépendants **?** : Tout d'abord la compréhension des processus évolutifs tels que la dérive génétique, la sélection, et la migration, mais aussi la variation génétique et phénotypique des populations naturelles et de déterminer la structure de la population **?** ; et ensuite, plus spécifiquement, décrire les effets d'une petite taille de la population sur la variation génétique et la viabilité de la population **?**Shafer *et al.* (2015).

La génomique ouvre en outre la possibilité de filtrer les individus et les populations par les "loci adaptatifs", qui est suggéré par certains comme la plus grande contribution potentielle de la génomique à la conservation **?**Shafer *et al.* (2015). Par ailleurs, l'exploration de l'adaptation locale dans le monde sauvage a augmenté de manière considérable (par exemple dans des populations de corbeaux **?** ou encore chez les populations de phasmes américain *Timema cristinae* **?**)

L'analyse ADN d'échantillons anciens (de plusieurs dizaines à plusieurs milliers d'années) peut renseigner sur des données génétiques de base dans les populations ancestrales avant des baisses démographiques actuellement **?**. Le moment de la fragmentation de la population et comment cela est lié à l'évolution passée de l'environnement (par exemple, l'impact anthropique ou le changement climatique) peut fournir des informations précieuses sur les processus actuels qui influent sur la viabilité de la population Shafer *et al.* (2015).

L'inclusion de marqueurs qui reflètent l'adaptation locale augmenterait l'identification des unités de conservation et pourrait améliorer la détection de régions génomiques qui entraînerait une dépression de consanguinité **???**. Identifier des marqueurs d'adaptation seraient également utile pour la conservation des processus évolutifs (**?** parle d'ailleurs d'ESU ("Evolutionarily Significant Units") pour classer les populations) et des gènes associés à l'amélioration de la santé de la population pourraient être propagées via une assistance humaine, y compris réaliser de l'hybridation afin de maximiser la capacité d'adaptation des populations à des environnements changeants **??**Shafer *et al.* (2015). Cette facette de la génomique de la conservation a clairement le potentiel de fournir des informations sur les espèces,les populations, et au niveau individuel qui était inaccessible en utilisant des marqueurs génétiques traditionnels Shafer *et al.* (2015).

Le diagnostic par les loci outliers peut être un outil précieux pour les études de suivi comme la pêche au saumon **?**, indépendamment de la signification adaptative de ces loci outliers Shafer *et al.* (2015).

Près d'une décennie de travail sur la génomique du condor de Californie (*Gymnogyps californianus*) ; une espèce en danger critique d'extinction ; a permis de découvrir la base génétique de leur chondrodystrophie, une forme récessive et mortelle de nanisme **?**.

Les outils génomiques peuvent aussi aider à la gestion de la pêche. Par exemple, le projet FishPopTrace s'est intéressé aux marqueurs SNP de quatre espèces de poissons commerciaux **?** : la morue *Gadus morhua*, le hareng *Clupea harengus*, le merlu *Merluccius merluccius* et la sole *Solea solea*. Ces données ont été utilisées pour décrire les populations au sein des espèces et par la suite d'identifier et de retracer l'origine géographique du poisson dans le commerce grâce aux SNP **?**Shafer *et al.* (2015). FishPopTrace à révélé la structure de leurs populations à une échelle géographique non reconnues précédemment, ce qui conduit directement à la conservation appliquée (par exemple, l'identification du commerce illégal et les erreurs d'étiquetage) Shafer *et al.* (2015).

3. Montrer en quoi c'est un challenge, particulièrement dans le domaine marin où les individus sont souvent difficile à observer, très mobiles et/ou très féconds. (citer des exemples concrets)

L'inférence de structure des populations repose donc sur la différenciation génétique (communication personnelle ; Benoit Simon-Bouhet). Cependant, l'inférence en connectivité génétique à un succés limité à cause du manque en structure génétique spatial pour les espèces à haute fécondité et à grande capacité de dispersion. Mais aussi et surtout dans les régions biogéolimités par des hotspots de différenciation génétique ().

La majorité des espèces marines combinent de nombreux traits d'histoire de vie (e.g.

haute fécondité, grande taille de population, fort taux de dispersion, cycle de vie complexe...) qui produisent de mauvais patterns de différenciation génétique, avec même parfois, pas de différenciation du tout(Ward *et al.*, 1994)(Palumbi, 1994)(Hedgecock *et al.*, 2007) ().

De nouvelles perspectives sont offertes par l'accroissement du nombre de marqueurs dans les études de génomique des populations, et tout particulièrement celles avec un focus sur les loci influencés par la sélection .

4. Montrer que dans ce cadre la génétique est un outil puissant et utile. Mais montrer que dans le milieu marin, l'étude de la structure des populations se heurte à certaines limites (citer des exemples sur l'inférence de la connectivité, voir Gagnaire 2015, parler des échelles de temps/géographiques auxquelles les marqueurs classiques permettent d'accéder ou non).

Les approches en génétique des populations offrent des méthodes d'évaluation de taux et d'échelle de dispersion (ou de migration) quand le mouvement des individus ne peut pas être mesuré par d'autres champs d'expérimentation comme la technique de CMR (Capture-Marquage-Recapture). Ces autres techniques (e.g. CMR) ne sont pas aisément applicable dans l'environnement marin, où les distributions et les voies de migrations des organismes sont cachés au yeux des humains sous la surface des océans(Hellberg, 2009) (Selkoe et Toonen, 2011) ().

Néanmoins, les marqueurs neutres auto-stoppeur avec des loci sous sélection peuvent donner de l'information à propos de la connectivité des patrons (patterns) dans des régions biogéographique bien mixé ().

De plus, les loci outlier peuvent permettre de délimiter de manière pertinente des unités de conservations, ainsi que de mesurer des taux de connectivité entre des régions éco-géogrpahiques ().

Alors, les grandes bases de données de génomique des populations peuvent maintenant être utiliser pour étudier la connectivité des populations marines ().

Il est aussi possible d'inférer la connectivité génétique en usant des méthodes indirectes. Le modèle d'IBD (Isolation By Distance) peut être utilisé pour estimer l'augmentation de différenciation génétique avec un accroissement des distances géographiques entre les populations **?**, ou entre les individus (**?**) quand la dispersion est spatiallement limité. Cependant, les mesures de la structure génétique ne traduisent pas facilement les taux de migration (**?**)(**?**)(). Un FST bas ne doit pas nécessairement signifier que la migration est forte comme la différenciation génétique est influencée par la taille efficace (Ne) et le taux de migration (m). De plus, l'estimation de la dispersion dépends d'autres paramètres comme les autres forces évolutives. Dans ce cas, les effets de la dérive génétique doivent être estimés indépendamment (généralement en estimant la densité), ce qui permet d'inférer la dispersion sous le modèle d'IBD (**?**). L'avantage de l'IBD est qu'il ne requière qu'un petit jeu de données pour tourner. Les méthodes d'inférence de population avec des méthodes indirectes échouent néanmoins sur certains points. Si la dérive génétique est trop faible pour générer une différenciation de la population, la dispersion ne peut pas être inférée en usant un modèle qui relie la balance migration/dérive (). Mais, comme nous l'avons vu plus

haut, ce problème est souvent rencontré avec des espèces avec des tailles de population très large, comme les poissons marins ou les invertébrés (**?**)(**?**). Par exemple, il n'y a pas de différenciation génétique détectable dans les populations californiennes de la moule *Mytilus californianus* sur près de 4000km (**?**). En outre, des espèces avec des tailles de populations importantes peuvent montrer des pattern de structure génétique qui ne présentent pas d'équilibre mutation-migration-dérive. Les estimateurs indirects de leur dispersion sont alors basés sur différentes méthodes statistiques qui évoluent à leur propre vitesse de différenciation ().

Il est aussi possible d'inférer la connectivité génétique en utilisant des méthodes directes. Ces méthodes dont généralement plus intuitives que les approches par méthodes indirectes, il faut faire attention à bien regarder les erreurs de type I (mauvaise identification d'un individu comme étant un immigrant par exemple). Mais elles sont appliquées avec succés pour les espèces marines (). Par exemple, l'assignation génétique ("genetic assignment"), basé sur des équations de fréquences alléliques est utilisé pour connaitre la dispersion des phoques gris (**?**). Ces approches requièrent une forte densité d'échantillonnage sur une large échelle géographique, et leur application en environnement marin est limité pour les populations avec des grandes tailles ou qui ont une distribution géographique mal documenté pour permettre un échantillonnage pertinent (). Bien que des études récentes ont démontré que la dispersion larval avait parfois des dispersions d'échelles spatiales beaucoup plus petites de prévues (**?**) (**?**). Etant donné que beaucoup d'espèces marines ont typiquement de forts taux de fécondité, des distributions larges et de grandes tailles de population (Palumbi, 1994), ces méthodes sont donc peu applicables pour la majorité des espèces marines ().

L'inférence de la connectivité génétique peut aussi se réaliser avec la méthode des clusters. Quand les espèces sont subdivisées en population discrètes, il faut d'abord évaluer le nombre de population avant d'évaluer le flux de gènes (**?**). La méthode des clusters peut détecter les discontinuités génétiques et les limites entre les flux de gènes pour identifier les populations (ou les "stocks") et les migrants (**?**)(Broquet et Petit, 2009). Les différentes approches du clustering ((**?**),(**?**)) ont aussi leur limites, qui partent de modèles sous-jacents (**?**). Par exemple, les modèles d'IBD peuvent conduire à du clustering (**?**). La puissance du clustering augmente avec la différenciation génétique au sein des population (**?**). Pour cette raison, cette méthode est adapté à l'inférence de la connectivité génétique pour des espèces qui ont des capacités de dispersion faible et une taille de population locale relativement petite, par exemple **?** sur le corail rouge de Méditerranée et **?** sur les serpents marins. Cependant ces espèces ne sont pas représentatives de la majorité des espèces marines.

5. Introduire une solution possible pour dépasser ces difficultés : les NGS (c'est évidemment le gros morceau de l'état de l'art). Dans cette partie, il faudra notamment :

a. Expliquer de quoi il s'agit (faire un rapide point sur les principales méthodes, pourquoi pas dans un encadr

b. Évoquer la chute des coûts qui explique que ces techniques deviennent abordables pour les études à l'échell

Nous approchons du génome à 1000 dollars **?**, ce qui signifie que la génération d'informations génomique est devenue de plus en plus accessible, même pour les organismes non-modèles avec de grandes tailles grande de génome **?**Shafer *et al.* (2015).

¡¡ moore$_{s}criptBSB_{m}odifié.R[, options]$ >>= $gg4$@

¡¡echo=FALSE¿¿= setwd("/Users/pierre-louisstenger/Documents/Cours fac/Master/S3/Stage M2/R/Moore") Importation des données cost ¡- read.table('costseq.txt',header=T, dec=",") cost

library(ggplot2)

Get costs according to Moore's law and the real cost in 2000 The cost should be divided by 2 every 2 years y ¡- cost$Couts[1]/(2^(0:6))x <- -seq(2000, 2012, by = 2)$

Compute the missing values (i.e. values for years 2001, 2003, 2005, etc) out ¡- lm(log(y) x) y ¡- exp(predict(out, data.frame(x=2000 :2015))) x ¡- 2000 :2015

moore ¡- data.frame(y,x) ggplot(data = moore, aes(x=x, y=y)) + geom$_{l}ine()+geom_{p}oint()$

all ¡- data.frame(Cost = c(cost$Couts, moore$y), Year = rep(2000 :2015, 2), Source = rep(c("Observé","Loi de Moore"), each=16)) all

gg4 ¡- ggplot(data = all, aes(x=Year, y=Cost, colour=Source)) + geom$_{l}ine()+scale_{yl}og10()gg4 <- -gg4+xlab("Années")+ylab("Coût")+ggtitle("Évolutiondescoûtsdesséquenceursentre2000et2015comp$

@

¡¡echo=FALSE,cache=FALSE¿¿= gg4 @

c. Expliquer l'intérêt de disposer de 10aines de milliers de marqueurs répartis dans tout le génome plutô

Plus il y a de marqueurs, plus la précision des mesures et la puissance des statistiques augmentent (**?**) ().

Les approches par NGS facilitent la découverte de nouveaux marqueurs qui sont influencés par la sélection (**?**)(**?**) (). Les loci outlier peuvent révéler des patterns de différenciation génétique, à la place des marqueurs neutres qui sont parfois peu informatifs. Ainsi, nous avons vu que les signaux des loci outlier pourraient être utilisé pour délimiter des stocks adaptés localement et redefinir des unités de conservation (Nielsen *et al.*, 2011) (**?**) (). Cette approche est tentante car la sélection peut être plus efficace que la dérive en opposition à l'effet d'homogénéisation des populations par la migration, en particulier quand les populations ont de grande taille efficace. Cependant, les loci outlier peuvent apparaitre suite à une large variété de mécanismes évolutifs en dehors de l'adaptation locale. Ces mécanismes évolutifs doivent donc être identifiés avant d'utiliser les loci outlier pour évaluer la connectivité ().

Il faut garder à l'esprit que les hotspots de différenciation génétique peuvent se trouver auprès des barrières naturelles qui limitent la dispersion (**?**), ou au travers de barrières exogènes ou endogènes de reproduction (**?**).

De plus, les effets indirects de la sélection peuvent révéler des structures génétiques cryptiques dû à de l'introgression originaire d'une zone de contact géographiquement distante (**?**), ou alors dû aux gènes autostoppeurs qui sont générés pendant la propagation d'un balayage sélectif (**?**) ().

Utiliser des loci sous influence de la selection, et des loci auto-stoppeurs peut être une

approche alternative à l'inférence de la connectivité marine. Les données avec nombreux marqueurs ont considérablement amélioré la puissance du scan génomique pour identifier les loci avec des niveaux de différenciation extrêmes (**?**). Les "FST Outliers" supposent être directement ou, plus probablement, indirectement affecté par la sélection (**?**)(**?**). De recentes études en génétique de la conservation ont proposé de délimiter localement des unités de gestion basées sur les signaux de ces loci outliers (**?**)(**?**). Un problème commun rencontré dans les études de génomique des populations est que les différentes méthodes pour identifier les FST outliers détectent seulement partiellement les lots de chevauchement des loci. La méthode de detection des FST la plus commune a un fort taux de détection faux-positif sous des scenari non équilibrés (**?**), dans des populations hiérarchisés (**?**), et dans les patterns d'IBD **?**. Pour contourner ces problèmes, combiner les méthodes basées sur la différenciation avec l'association du génotype et de l'environnement est suggéré comme plus fiable que l'approche des outliers (**?**). En outre, la sélection peut aussi avoir un effet dominant sur la diversité génétique des espèces marines, tel est le cas de populations d'épinoches *Gasterosteus aculeatus* (**?**) ou de encore de bar *Dicentrarchus labrax* **?**.

Les clines génétiques peuvent aussi estimer les distances de dispersion. Les études sur les scans génomiques dans le milieu marin ont reporté des exemples de patterns en forme de clines sur les loci outliers, qui pourraient coincider dans l'espace avec un changement de gradient environnemental, d'écotones, de barrières biogéographiques... ((**?**)(**?**)(**?**)(**?**)) Les formes de clines peuvent être déterminés basiquement par la balance entre la migration et la sélection. Les études de génomique des populations ont maintenant le pouvoir de détecter des loci montrant des variations clinales dans des espèces que l'on croyaient être génétiquement homogènes. La probabilité de découvrir de nouveaux cas d'adaptation locale avec les clines, et des zones hybrides cryptiques est donc grande (**?**).

La fréquence des allèles varient selon une courbe sigmoïde de la distance géographique, sans nécessairement trouver une fixation d'un allèle si la sélection ne peut pas purger l'afflux de génotypes maladaptés. Les clines d'adaptation local peuvent être utilisés pour estimer la distance de dispersion, si le coefficient de sélection peut être mesuré, ce qui représente actuellement un challenge. La mesure de la sélection peut parfois être obtenue en utilisant des populations expérimentales ou en comparant la fréquence du génotype entre les larves et les adultes collectés d'une même cohorte. Ces clines d'adaptation locale peuvent offrir des alternatives pour estimer la migration des espèces marines avec un grand flux de gènes, en gardant à l'esprit que les modèles sous-jacents supposent que chaque cline évolue indépendamment. Par exemple, un scan génomique de grande densité réalisé chez *Drosophila melanogaster* a révélé la présence de clines latitudinaux (**?**) qui se chevauchent géographiquement, avec des clines qui ont été attribués à de l'adaptation locale (**?**). Comme pour la drosophile, des clines dits "classiques" ont été retrouvés dans les organismes marins, comme le cline du gène Ldh chez le poisson-killy *Fundulus heteroclitus* (**?**) qui s'est finalement avéré être dû à une zone de contact secondaire (**?**).

Les clines dû au contact de zones entre des taxa hybridés sont des zones hybrides (**?**). Dans ces clines, chaque locus cumule les effets indirects de la sélection d'un autres loci et les effets de son propre coefficient de sélection (**?**)(**?**). L'association parmis les allèles sous

sélection dans les zones hybrides peuvent alors être utilisés pour inférer la dispersion (**?**).

Utiliser l'assignation génétique individuel ("individual genetic assignment") est une approche conceptuellement différente pour estimer la connectivité dans les zones de contact. Cette approche est très similaire à l'estimation directe de la connectivité génétique et prend l'avantage considérable de prendre en compte les différences génétiques entre des populations ou des espèces des deux côtés de la zone hybride (). Cette approche est utilisé pour estimer l'origine de la dispersion de larves, comme par exemple celles de la moule bleue *Mytilus edulis* (**?**).

Estimer la connectivité d'une population est souvent nécessaire en dehors de ces régions singulières, par exemple quand il est nécessaire de déterminer s'il y a une dispersion limitée entre les populations dans les zones délimitées par des limites écologiques ou biogéographiques, ce qui est une préoccupation relativement commune pour les questions de gestion de conservation (**?**). Une solution potentielle, quand l'équilibre migration-dérive n'est pas informatif, est de chercher des preuves de la structure spatiale dû par une introgression (**?**). Utiliser le ratio introgression/homogénéisation permet de montrer que les taux d'introgression sont variables entre les loci et peut donner les moyens de détecter une faible barrière au flux de gènes, même lorsque l'introgression a commencé des milliers de générations dans le passé ().

Les queues d'introgressions peuvent être aussi influencés par une sélection agissant hors de la zone de tension. Dans ce cas, le gradient de la fréquence de l'allèle dans les espèces introgressés peut être accentuée par un gradient de sélection (par exemple un gradient de l'environnement). En effet, les zones de contacts secondaires coïncident généralement avec des gradients environnementaux **?**. De plus, les queues d'introgression peuvent être couramment rencontrées dans les régions biogéographiques séparées par des limites environnementales, comme pour la mer Baltique (). Ces mécanismes montrent combien il est important d'échantillonner non seulement l'ensemble de l'aire de répartition d'une espèce, mais aussi les populations divergentes, ou alors les espèces qui sont étroitement liés et qui vivent en parapatrie ou en sympatrie, avant d'interpréter les modèles de variation génétique spatiale (**?**)(**?**)(**?**)(). Maintenant que les outils des NGS commencent à révéler les îlots génomiques de différenciation entre les espèces cryptiques qui étaient auparavant considérés comme des populations de la même espèce (**?**)(**?**)(**?**). Les polymorphismes situés dans la périphérie de ces îles pourraient devenir un nouveau type de marqueurs relativement puissant qui permettraient de déduire la connectivité au sein des espèces ().

Cependant la création de loci outlier peut aussi être du aux allèles autostoppeur dans une population spatiallement subdivisée. Ce processus laisse une empreinte au niveau des marqueurs neutres dans le voisinage chromosomique de l'allèle qui a été balayé. Lorsque la différenciation génomique globale est faible (ce qui est généralement le cas dans les espèces marines), ce processus génère un niveau élevé de différenciation des deux côtés du locus sélectionné (**?**). En effet, la recombinaison rompt progressivement l'association entre le lieu choisi et le lieu neutre auto-stoppeur, tandis que l'onde de balayage se propage (). Par conséquent, l'effet de l'auto-stop est fort dans le berceau de la mutation qui est favorable, alors qu'il se ramollit progressivement à mesure que l' onde se déplace (). Il

n'existe que quelques exemples d'études s'ayant penché sur ce sujet, dont celui de la moule bleue *Mytilus edulis* de **?** et celui de l'épinoche *Gasterosteus aculeatus* de **?** explicités plus haut. En ajustant le modèle de l'auto-stop global aux données de la moule bleue *Mytilus edulis*, il a été possible d'estimer le taux de migration minimale nécessaire pour obtenir la valeur de FST observée entre les deux populations géographiquement éloignées selon **?**, qui, finalement se révéla être étonnamment faible ($m < 10^8$) (**?**). Ceci demontre donc que deux populations de moules qui sont démographiquement indépendantes depuis des milliers d'années ne présentent pas de panmixie génétique (**?**) ().

La question qui reste en suspend est de déterminer si toutes les différences génétiques révélées par loci outliers sont pertinentes pour la gestion de la conservation et de l'espèce ( ).

6. Évoquer enfin les difficultés liées aux NGS. Dans le désordre :

c. outils statistiques classiques pas forcément adaptées à la quantité de données (e.g. overfitting lié très gra

Plus le nombre de marqueurs augmente, plus la non-indépendance des loci d'un jeu de donnée génétique d'une grande population deviens un problème qui requière encore plus d'investigation (**?**). Malgré leurs limitations bien connues, il y a bon espoir que les jeux de données en génomique des populations vont améliorer l'utilité des méthodes indirectes en augmentant la puissance et la précision des petites différenciations génétiques (). Les ACP (Analyses en Composantes Principales) ont le bénéfice de donner des informations sur de variations qui sont rares, difficiles à détecter sur les populations à échelles fines (**?**), spécialement dans le cas de larges populations qui n'échangent que quelques migrants par an ().

a. méthodes encore coûteuses pour une utilisation en routine, mais probablement plus pour longtemps

Sur le long terme, il sera essentiel de développer une gamme de protocoles de laboratoire comme celui proposé par **?**.

b. problèmes liés au traitement des données (assemblage des contigs, détection des erreurs de séquençage, pipe

d. difficulté liées aux quantités/qualité d'ADN nécessaire pour ces méthodes : pas toujours simple pour des

e. lien entre génétique et conservation

Il n'est cependant pas évident que les approches via la génomique soient vues par les sociétés de conservation comme étant la clé du succés à la gestion des espèces Shafer *et al.* (2015). Une raison évidente de cette déconnexion est que bon nombre des problèmes liés à la pression de conservation (par exemple, pour les grands carnivores **?** ou pour la cohabitation entre la création de routes et la gestion des espèces au parc du Serengeti (Tanzanie) **?**) n'ont tout simplement pas besoin de la génomique, mais plutôt besoin d'une volonté politique.

Utiliser les données génomiques pour déduire les paramètres démographiques est encore un domaine quasi exclufif de la recherche universitaire actif **??**Garner *et al.* (2015). Un logiciel convivial qui accueille les données génomiques - qui est une clé pour une application pratique - reste limitée **???** et des ordinateurs à haute performance sont généralement nécessaires pour le stockage et l'analyse de ces données Shafer *et al.* (2015)Garner *et al.* (2015).

Les communautés scientifiques et politiques (de la gestion et de la conservation) opèrent dans des domaines largement séparés Shafer *et al.* (2015)Garner *et al.* (2015) et il est possible que l'introduction de la génomique dans l'équation va augmenter cet écart Shafer *et al.* (2015).

Un problème récurant dans ce dans ce domaine est que la recherche en génomique de la conservation n'est généralement pas renforcée dans les programmes de financement actuels Shafer *et al.* (2015). Développer un outil génomique qui exige rigueur et répétition n'est pas favorisé dans un climat de "publish or perish" Shafer *et al.* (2015). Il faudrait donc repenser comment les founds académiques et des communautés de conservation fonctionnent, et les fusionner, tout en impliquant les ONGs protectrices de l'environnement Garner *et al.* (2015), et ainsi avoir une politique de recherche commune, pour avoir plus d'impact **?**Shafer *et al.* (2015).

DON'T READ BELLOW THIS LINE (ancient script)

Valeurs	Connectivité adaptative	Connectivité de consanguinité	Connectivité de dérive
Critère	Flux génétique suffisant pour propager les allèles avantageux	Flux génétique suffisant pour éviter les effets néfastes de la consanguinité locale	Flux génétique suffisant pour maintenir une même fréquence allèlique
m	?	?	?
Nm	> 0.1	> 1.0	> 10
F_{st}	< 0.35	< 0.20	< 0.02

TABLE 1 – Définition des trois types de connectivité génétique

Au cours du XX^{e} siècle, les avancées technologiques permettant d'accéder directement à l'**ADN (Acide .)** et les nombreux progrès améliorant notre compréhension des mécanismes de l'**hérédité** ont posé les bases d'une nouvelle discipline : la **génomique** (Eggen, 2003).

La **séquence nucléotidique** (succession d'acides nucléiques qui composent l'ADN et qui est porteur de l'information génétique) permet dorénavant d'appréhender les variations phénotypiques (les différences de l'ensemble des caractères observables d'un individu) entre des êtres vivants (Champe et Benzer, 1962).

L'objectif majeur de la génomique consiste à obtenir une connaissance et une compréhension de la structure et de la fonction des génomes (ensemble du matériel génétique d'un organisme) (Eggen, 2003).

La diversité génétique est un concept central qui lie l'évolution biologique d'une espèce avec la complexité de l'organisme via son génome (Lynch et Conery, 2003), le rétablissement de l'écosystème (Reusch *et al.*, 2005) et l'habilité des espèces à répondre aux changements environnementaux(O'Brien, 1994).

En effet, selon Lynch et Conery (2003), les séquences génomiques de diverses lignées phylogénétiques révèlent des augmentations importantes dans la complexité du génome de procaryotes par rapport aux eucaryotes. Ceci serait due à l'augmentation graduelle du nombre de gènes, résultant de la rétention de gènes en double, et des augmentations plus brusques de l'abondance des introns (parties de l'ADN non exprimées) et des éléments génétiques mobiles. (Lynch et Conery, 2003) soutiennent que le nombre de ces modifications est expliqué par des réductions à long terme la taille de la population.

En cas de résilience d'un écosystème suite à une pression, la **biodiversité** est expliquée par la complémentarité des génomes ("Genotypic complementarity") plutôt que par la sélection de génotypes particulièrement robustes (Reusch *et al.*, 2005). Il est donc important de maintenir la diversité des espèces pour améliorer la résilience des écosystèmes dans un monde d'incertitudes croissantes environnementalement parlant (Reusch *et al.*, 2005).

Les petites populations sont soumises aux risques de la **consanguinité**, de la **dérive génétique**, de la **perte de la variation génomique** globale due à la perte allélique, ou encore à la réduction de l'**hétérozygotie** (quand un individus possède deux allèles différents d'un même gène) (O'Brien, 1994). Les conséquences de ces **déplétions génétiques** (O'Brien, 1994) peuvent être catastrophiques pour ces populations, voir pour l'espèce. Il est donc important d'appliquer les outils moléculaires de la génétique des populations pour la **conservation de ces espèces**.

Au sein des espèces, la **diversité génétique** est donc pensée pour refléter **la taille de la population, l'histoire, l'écologie et capacité d'adaptation.**

De nombreuses techniques de biologie moléculaire existent pour obtenir ces informations. Pour ce faire, l'**ADN mitochondrial** (ADN provenant des mitochondries, et légué uniquement par la mère), les microsatellites (Séquences d'ADN formée par une répétition continue des mêmes bases) et l'**ADN nucléaire** (ADN provenant du noyau de la cellule) sont très utilisé.

L'ADN mitochondrial semble être un marqueur de choix ; il n' a généralement **pas de recombinaison**, a un **taux de dérive plus important** que l'ADN nucléaire et un

niveau élevé de **polymorphisme** (propriété des espèces à se présenter sous plusieurs formes différentes) (Avise et Walker, 2000; Aurelle, 2009).

Cependant, (Bazin *et al.*, 2006) montrent que l'ADN mitochondrial (ADNmt) est un marqueur couramment utilisé qui ne reflète pas l'abondance des espèces ou de l'écologie : la diversité de l'ADNmt n'est pas plus élevé chez les invertébrés que chez les vertébrés, en milieu marin que dans chez les espèces terrestres, ou que dans les petits organismes par rapport aux grands. Le loci nucléaire (ADN nucléaire), en revanche, est adapté à ces attentes intuitives. La distribution inattendue de la diversité mitochondrial est expliquée par l'**évolution adaptative récurrente** (évolution répétée d'un caractère particulier), contestant la **théorie neutraliste de l'évolution** (selon laquelle la plupart des mutations ont une influence négligeable sur la fitness (succès reproducteur) de l'individu, et ne donne donc qu'une influence ponctuelle à la sélection naturelle) et de remettre en question la pertinence de l'ADNmt dans les études de la biodiversité et de conservation ; ce qui correspond à environ trente ans de recherche scientifique (communication personnelle, Benoit Simon-Bouhet). Malgré tout, ce marqueur permet de fournir une première image de la structuration génétique d'une espèce (Aurelle, 2009). De plus, (Kitchen *et al.*, 2008) ont démontré une corrélation positive entre l'ADNmt et les variations des allozymes, ce qui suggère que la diversité de l'ADNmt peut corréler avec la taille de la population.

Néanmoins, (Kashi et King, 2006) remettent aussi en cause l'utilisation des **microsatellites**, ainsi que l'hypothèse d'une évolution neutre des marqueurs employés (Aurelle, 2009). Les microsatellites, en vertu de leurs qualités de mutations et de leurs qualités fonctionnelles, jouent un rôle majeur dans la génération de la variation génétique impliquant une évolution adaptative (Kashi et King, 2006).

Depuis le début des **années 2000**, de nouvelles technologies apparaissent, permettant de répondre aux questions de la génétique des populations, en limitant les coûts, le temps et les biais. Le **séquençage** (qui permet la détermination de la séquence des gènes) haut débit, appelé **NGS** pour **Next-Generation Sequencing**, a radicalement changé le paysage de la génomique, et pour inférer la structure de population.

En effet, les méthodes traditionnelles en génétiques des populations ne mettaient en lumière que les effets de la **dérive génétique** (fixation d'un allèle (une des versions différentes d'un même gène) dans la population), les **mutations** et les **migrations**. Mais c'est l'avènement des NGS qui a permis de prendre en compte les effets de la sélection, mais aussi à (Bazin *et al.*, 2006) et à (Kashi et King, 2006) de remettre en question ces anciennes pratiques.

De nouvelles formes de NGS apparaissent rapidement, baissant toujours plus les coûts et les temps de séquençage. Cette évolution est encore plus rapide que la conjecture (loi) de Moore (voir figure 2), en effet, en dix ans, le prix d'un séquençage a été divisé par 100 000.

¡¡ $\text{moore}_{s}criptBSB_{m}odifié.R[, options] >>= gg4@$

La qualité et la quantité des résultats a donc explosé. Avec la **technique Sanger** (première technique de séquençage ADN, en 1977 par Frédérick Sanger) 96 ADNs différents pouvaient être analysés en une fois, alors qu'avec les NGS ce sont des millions d'ADNS différents qui peuvent être analysés en une fois. C'est un véritable saut technologique.

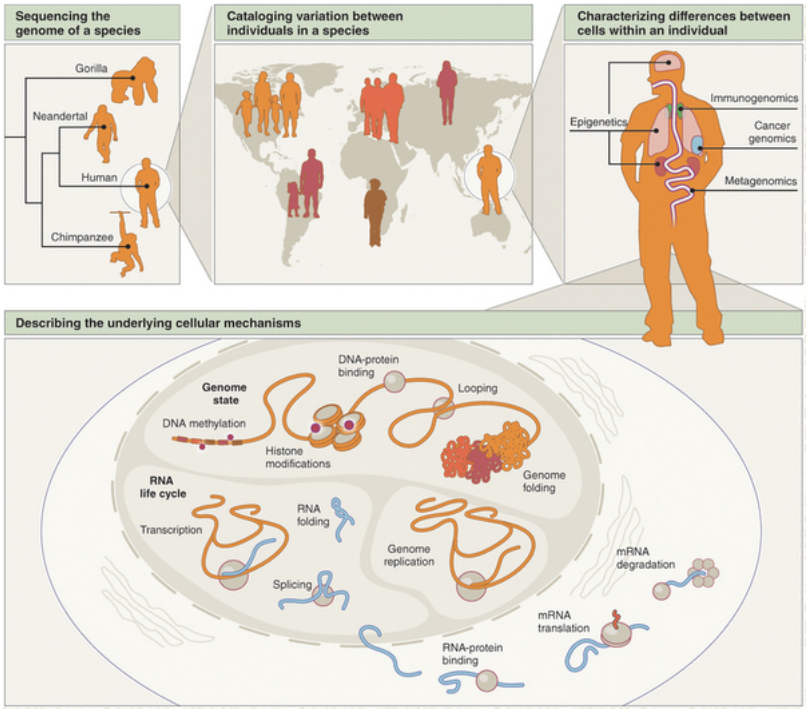

FIGURE 1 – Les NGS permettent de travailler à différentes échelles(Shendure et Aiden, 2012)

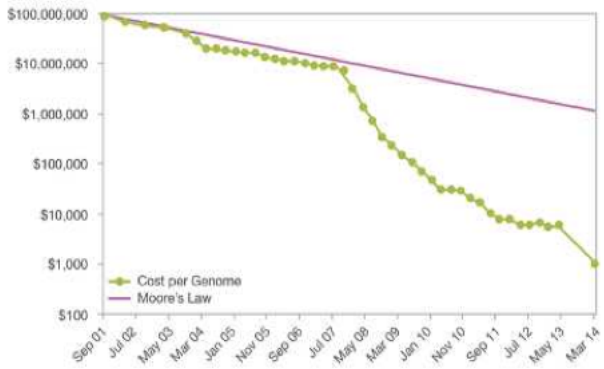

FIGURE 2 – Evolution très rapide des instruments, des débits et des coûts (premier séquençage sur 454 en Janvier 2008)

Il existe de nombreuses technologies et plateformes dont les principales sont la technique de **pyroséquençage 454** (Société Roche, avec les appareils GS Junior System, et GS FLX+ System), la technique d'**Illumina** (Société Solexa avec les appareils HiSeq System, Genome analyser Ilx ou encore MySeq), la technique **Applied Biosystems** de Life Technologies avec les appareils SOLID 5500 System ou encore la technique **Ion Torrent** de la même société (Life Technologies, avec les appareils Personal Genome Machine et Proton). À partir de 2012, une nouvelle génération de séquenceur encore plus performant arrive sur le marché, appelé "troisième génération de séquenceur" ou encore "**Next-Next Generation Sequencing**". Les principales techniques sont celles d'**Helicos** (avec l'appareil Helicos Genetic Analysis System), la technique **Pacific Biosciences** avec PacBio RS ou encore la technique d'**Oxford Nanopore Technologies** avec GridION System et MinION.

Il ne sera présenté ici que la technique de pyroséquençage 454, la technique d'Illumina via la technique RAD-Seq et la technique d'Oxford Nanopore via le MinION.

Toutes ces NGS se basent sur le même principe. Il faut **fragmenter** l'ADN, créer des **banques d'ADN** par ligation d'adaptateurs, les **cloner** (soit par PCR en émulsion sur une bille dans des microréacteurs (décrit plus bas, dans la technique 454) comme pour les techniques 454 ou SOLID, ou alors par "Bridge" PCR (Polymerase Chain Reaction) sur un support plan (Flow cell) ce qui permet de créer des colonies d'ADn se nommant polonies, comme dans la technique Illumina), puis ces clones sont **séquencés** par une des techniques vu précedemment. Par réaction chimique, de la luminescence est émise (décrit plus bas, dans la technique 454) ce qui permet de convertir les séquences en fichier informatique.

En 2005, Jonathan M. Rothberg a élaboré la technologie de **pyroséquençage 454** Life Sciences (depuis racheté par Roche) et a prouvé la robustesse de ce séquençage en séquençant le génome de *Mycoplasma génitalium* (Margulies *et al.*, 2005) et qui ne nécessite pas de clonage (donc gain de temps et d'argent), et permet une lecture directe de la séquence obtenue après le séquençage.(Sengenes, 2012).

La particularité de cette technologie est qu'elle repose sur trois phases, dont une phase centrale de PCR en émulsion pour l'amplification des fragments à séquencer. (Margulies *et al.*, 2005) (Sengenes, 2012)

La première phase correspond à la **préparation de la banque :** L'ADN que l'on souhaite séquencer est dans un premier temps fragmenté par nébulisation (Loman *et al.*, 2012) (pour les petits ADN de $0.5 - 5$ µm(Prodromou *et al.*, 2007), et utilisé avec de l'azote comprimé)(Syed *et al.*, 2009) , par audition hydrodynamique (Poptsova *et al.*, 2014) (pour les ADN de taille moyenne $5 - 10$ µm)(Prodromou *et al.*, 2007) qui est obtenu à l'aide de pressions aquatiques exercées sur les molécules d'ADN (Poptsova *et al.*, 2014) ou par "sonication" (Knierim *et al.*, 2011) (pour les gros ADN de $10 - 100$ µm)(Prodromou *et al.*, 2007) où les échantillons sont soumis à des ondes ultrasonores, dont les vibrations produisent des cavitations gazeuses dans le liquide qui cisaillement les molécules d'ADN par vibration de résonance (Knierim *et al.*, 2011). afin d'obtenir des fragments d'environ 300pb. Les extrémités cohésives créent lors de la coupure vont être réparées afin d'obtenir des extrémités franches permettant l'ajout des adaptateurs.(Prodromou *et al.*, 2007) (Sengenes, 2012)

Par le biais d'une ADN ligase (Sengenes, 2012), on ajoute ensuite des adaptateurs qui contiennent notamment une séquence MID (Multipled IDentifier) qui permettra d'identifier chaque échantillon lors du séquençage. (Margulies *et al.*, 2005)

La seconde phase est l'**amplification clonale par PCR (Polymerase Chain Reaction) en émulsion (emPCR) :**

La PCR à lieu dans une microgoutte avec une microbille d'agarose en phase aqueuse, baignant dans de l'huile avec plusieurs autres millions de billes. Chaque bille comprend donc plusieurs copies d'un même brin d'ADN. Chacune des billes est disposée dans un des 1,6 millions de puits d'un support nommé PTP (PicoTiterPlate) (Sengenes, 2012)

La dernière phase correspond à la **réaction de séquençage en elle même qui est le pyroséquençage :** ces réactions se produisent dans chaque puits avec les nucléotides qui traversent l'un après l'autre la PTP (voir figure 3). Quand la polymérase incorpore un nucléotide, un pyrophosphate (PPi) est alors libéré, et par le biais de cascades enzymatiques, une adénosine triphosphate (ATP) est crée qui va ensuite convertir la luciférine en oxyluciférine via une luciférase. C'est cette dernière réaction qui produit de la lumière, et plus de nucléotides seront incorporés, plus la réponse lumineuse sera importante. Ces signaux lumineux sont captés par la machine qui les convertis en information numérique. (Sengenes, 2012)

Cette technique présente malgrè tout des limitations :

— **Pour l'extraction des données :** un fichier SFF (Standard Flowgram File) est créé en sortie standard du 454 (Fichier binaire qui est humainement illisible(Pey, 2010)), pour le lire il faut un exécutable fourni par la société Roche (sffinfo, uniquement pour ordinateurs avec noyeau Unix (Linux, iOs...))

— **Les erreurs de séquençage peuvent être nombreuses :**

1. Il existe des *insertions ou des délétions* qui rendent difficile à déterminer le nombre de nucléotides entrant dans la composition d'un homopolymère (suite d'un même nucléotide), possible perte de la relation de linéarité entre l'intensité lumineuse émise et le nombre de nucléotides incorporé (Pey, 2010). Il peut aussi y avoir détection d'un signal provenant d'un puits adajacent (Pey, 2010). Selon (Balzer *et al.*, 2011), le **phénomène de CAFIE** (CArry Forward/incomplete Extension) est une erreur de séquençage (qui rend les séquences incomplètes) relativement commune.
2. Des *bases ambigües* peuvent apparaitre dans la séquence sous forme de code particulier (par exemple un "N" correspond à une base inconnue, ou encore un Y correspond à une hésitation entre les bases C et T...(Communication personnelle ; Amélia Viricel-Pante))
3. Des *erreurs de prédiction* sont aussi possible comme un signal surestimé suivi d'un signal sous-estimé ou vice versa. (Gilles *et al.*, 2011)

— **Par des réplicas artificiels :** ils correspond 4 à 44 pourcents des erreurs selon (Niu *et al.*, 2010) et 11 à 35 pourcents selon (Gomez-Alvarez *et al.*, 2009)). Il y a plusieurs billes dans une meˆme goutte d'émulsion dont une seule porte un fragment d'ADN. La caméra détecte une émission de lumière dans un ou plusieurs puits vides provenant d'un puits adjacent ou` s'effectue la réaction de pyroséquençage (Pey, 2010).

— Et enfin il existe des **séquences chimériques (amplicons)**, qui sont des séquences qui se sont hybridées(Haas *et al.*, 2011).

En 2005, Rothberg annonça que cette technologie permettrait de séquencer le génome de James D. Watson pour seulement 1 millions de dollars (Margulies *et al.*, 2005) (Sengenes, 2012). Ils réussirent en 2007 en respectant le budget annoncé, et en le réalisant en deux mois (Wheeler *et al.*, 2008). En 2006, c'est avec cette technique que le génome de l'homme de Néanderthal à pu être séquencé (Green *et al.*, 2006), (Noonan *et al.*, 2006), même si

(Wall et Kim, 2007) ont démontré qu'il y avait majoritairement de l'ADN humain moderne contaminant (Sengenes, 2012).

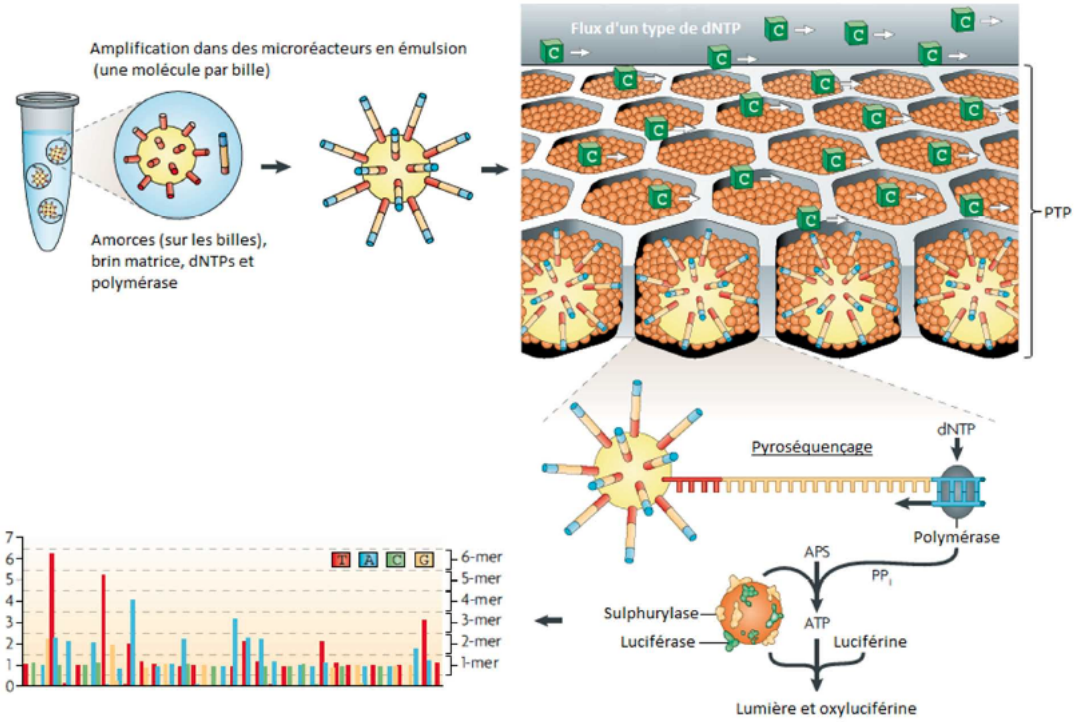

FIGURE 3 – Aperçu de la technologie 454 (Sengenes, 2012), PTP : PicoTiter Plate, PPi : pyrophosphate inorganique, APS : adénosine phosphosulphate, ATP : adénosine triphosphate.

Les **marqueurs des sites de restriction** (séquence particulière de nucléotides qui est reconnue par une enzyme de restriction comme un site de coupure dans la molécule d'ADN) associés à l'ADN (**RAD-Seq** : Restriction site Associated DNA) sont utilisés pour la cartographie génétique, dont la cartographie des QTL (Quantitative Trait Loci), mais aussi dans la génétique des populations, et donc dans la compréhension de l'évolution (Davey et Blaxter, 2010).

La **technique RAD-Seq** est un séquençage de type NGS qui lie les séquences aux sites de restriction (Baird *et al.*, 2008), puis fragmente le génome par digestion enzymatique, réalise des ligation d'amorces et un code barre (pour distinguer les différents échantillons (Davey *et al.*, 2013)). Il faut isoler les balises RAD (**RAD-tags**), puis les séquences ADN flanquent immédiatement dans chaque site de restriction dans tout le génome. Il y a donc deux fois plus de RAD-tags que de sites de restriction(Davey *et al.*, 2013). Une fois les balises RAD isolées, on identifie et recherche les SNP (Single Nucleotide Polymorphism) pour voir le polymorphisme (Hohenlohe *et al.*, 2010). Il est utilisé sur des séquenceurs de type Illumina, SOLID ou encore Ion Torrent PGM (Hohenlohe *et al.*, 2010).

C'est une technique très utilisée pour la **biologie évolutive** :

— (Hohenlohe *et al.*, 2010) ont utilisé la **technique Illumina-sequenced RAD tags**

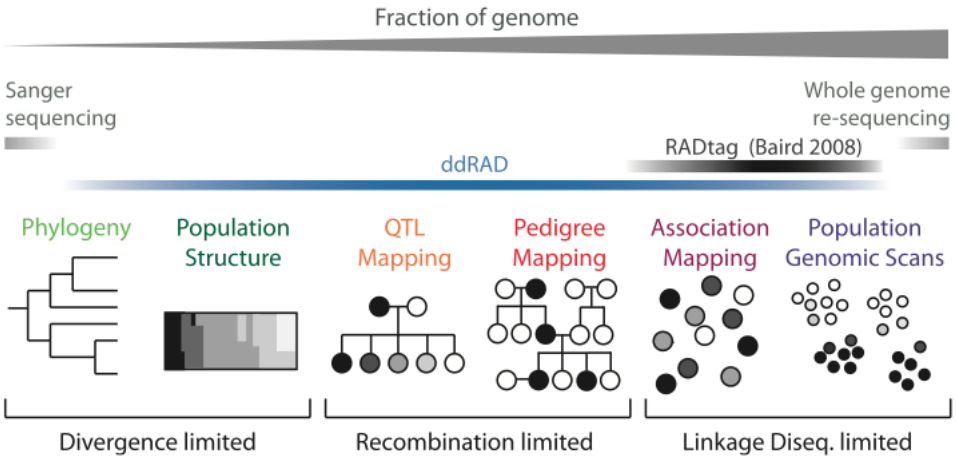

FIGURE 4 – Résolution des marqueurs de type RAD : Un procédé de génotypage souple peut être utilisé pour optimiser le nombre de marqueurs génétiques pour une approche expérimentale spécifique dans un système biologique donné(Peterson *et al.*, 2012)

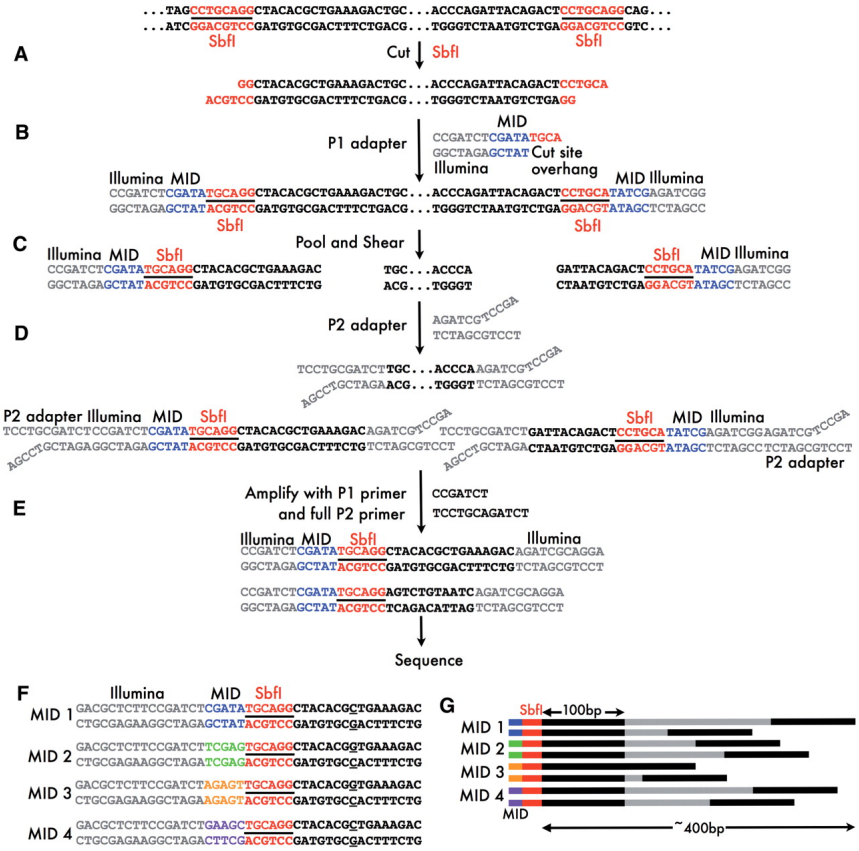

FIGURE 5 – étapes pour obtenir des séquences d'ADN via la technique Rad seq

pour identifier plus de 45 000 SNP chez 100 épinoches (*Gasterosteus aculeatus*) provenant de la mer ou de rivières. Cette étude est une première en terme **scannage génomique** de haute densité basé sur des SNP permettant de calculer la diversité génétique et la différenciation de ces populations d'épinoches dans la nature. Ceci a permit de d'identifier les régions génomiques, d'élucider la part évolutive

et démographique de ces populations naturelles et donc de trouver **des gènes candidats de signification évolutive**.

— De la **cartographie à l'aide de marqueurs** a était réalisé sur des Lepisosteus (*Lepisosteus oculatus*)(Amores *et al.*, 2011), ce qui a permit de découvrir que c'est une lignée de poissons qui a divergée avant la duplication du génome des téléostéens (ce qui correspond à un "Outgroup"). De plus, cette technique a mis en lumière que leur génome est plus proche de celui des hommes que celui des autres téléostéens (voir figure 6).

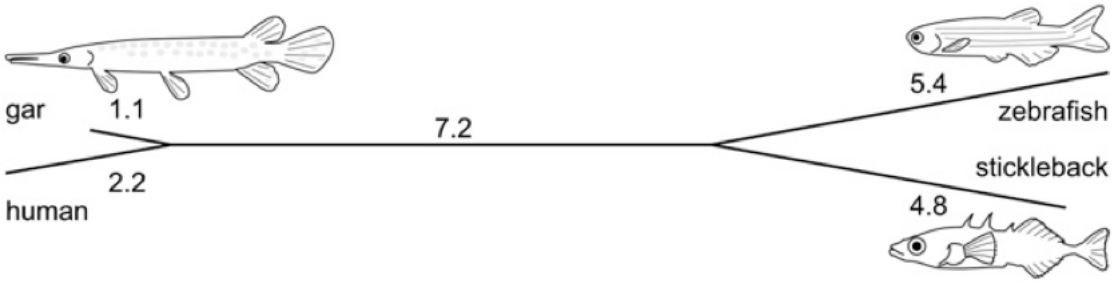

FIGURE 6 – Comparaison synthetique de deux téléostéens (zebrafish and stickleback) par rapport au lepisosteus et à l'homme. La longueur des branches est proportionnelle à l'estimation du nombre de divergence en terme de chromosome entre les espèces (Amores *et al.*, 2011).

— Les **scans génomiques** sur des milliers de SNP peuvent permettre de découvrir un patron de divergence et/ou un flux de gènes entre des espèces écologiquement divergente, comme pour *Populus tremula* avec *Populus trichocarpa* (Stölting *et al.*, 2013). Stölting et ses collègues ont scanné le génome de ces deux arbres hybrides différents d'un point de vue écologique. Ils ont utilisé plus de 38 000 SNP en utilisant la méthode de RAD-seq et ils ont découvert une grande divergence génétique (e.g. la proportion de SNP fixé) entre les espèces sur 11 des 19 chromosomes. Ceci correspondrait plus à un flux de gènes régulier qu'a du polymorphisme ancestral partagé. Ces résultats permettent donc d'expliquer l'origine de ces "génomes mosaïques" (Stölting *et al.*, 2013) vu dans ces taxa avec des génomes dits "poreux" (Stölting *et al.*, 2013) et suggèrent une introgression ou une conservation extensive naissante parmis les espèces des chromosomes sexuels chez ces végétaux.

— Le RAD-seq permet aussi de réaliser de la **phylogéographie**. En effet, un nombre très large de SNP à travers le génome a le pouvoir d'affiner nos connaissances sur l'histoire démographique d'une population et d'identifier les régions du génome où la sélection naturelle a agit. Reitzel *et al.* (2013) ont utilisé cette technique sur une anémone américaine (*Nematostella vectensis*) en guise de modèle. Des centaines de SNP contenant des "tags" ont été identifiés dans des milliers de RAD loci provenant de 30 individus barcodés de quatre lieux différents sur la côte Est des Etats-Unis d'Amérique. Malgré le manque d'information sur cet espèce (e.g. de génome de référence), un arbre phylogénétique a pu être créé (voir figure 7) (Reitzel *et al.*,

2013).

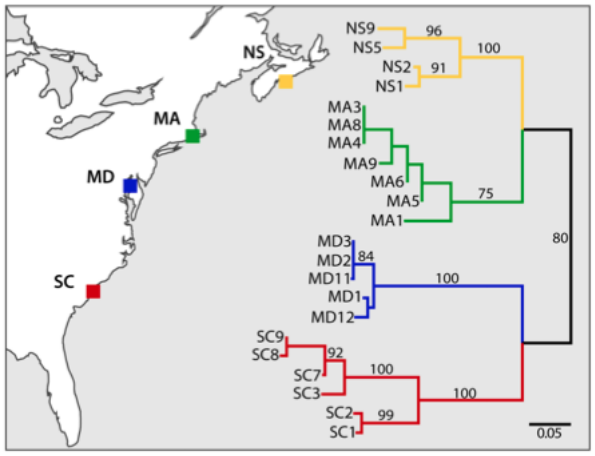

FIGURE 7 – Phylogéographie de *N. vectensis* : Nova Scotia (NS), Massachusetts (MA), Maryland (MD), and South Carolina (SC) (Reitzel *et al.*, 2013).

— La **phylogénomique** a vu ses capacités augmenter grâce à la technique RAD-seq. L'analyse des bibliothèques RAD en utilisant des outils bioinformatiques et phylogénétique a permit d'avoir 400 fois plus de sites que l'approche de Sanger et d'avoir par exemple une phylogénie basée sur un alignement de 2 262 825 nucléotides par espèces chez des coléoptères (Cruaud *et al.*, 2014). Ainsi les relations entre 18 espèces de carabes qui ont divergé il y a 17 millions d'années ont pu être déterminés avec précision (Cruaud *et al.*, 2014), alors que les techniques de Sanger via l'ADN nucléaire et mitochondrial ne donnaient pas les mêmes résultats (voir figure 8).

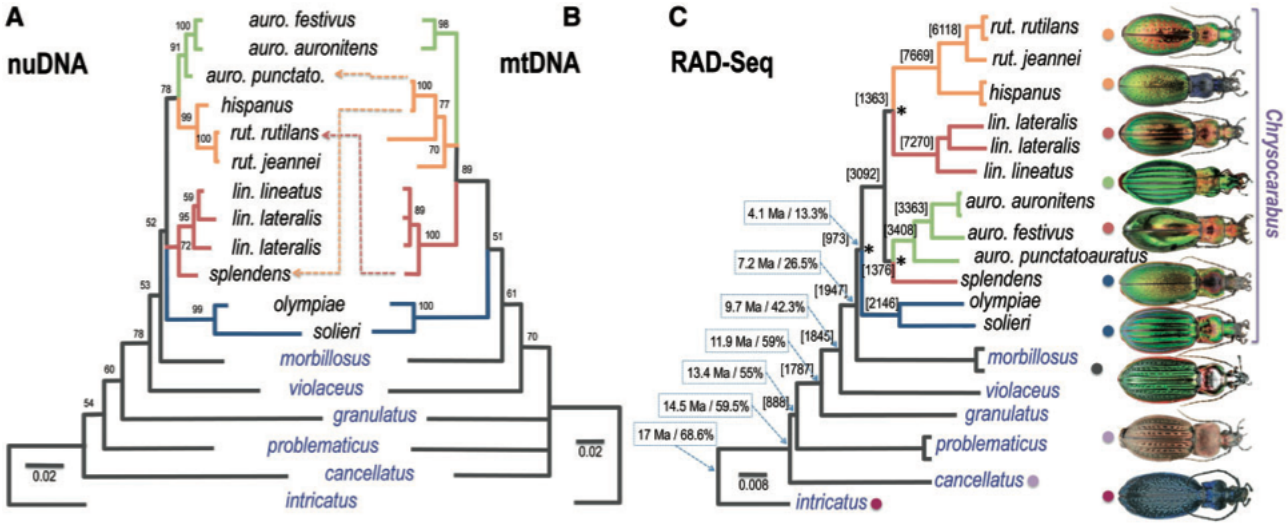

FIGURE 8 – Phylogénie des carabes obtenues par ADN nucléaire (A), ADN mitochondrial (B) et via la technique de RAD-Seq (C)(Cruaud *et al.*, 2014) .

— C'est aussi une technique très utilisé pour la **délimitation d'espèces** (Herrera et Shank, 2015)(Pante *et al.*, 2015) et la **structure des populations** (Pante *et al.*, 2015). Selon les techniques classiques de génétiques, les *Chrysogorgia* sont des coraux profonds dont les espèces sont délimitées par un seul haplotype mitochondrial

(Herrera et Shank, 2015)(Pante *et al.*, 2015). Avec la technique de RAD-Seq le nombre de loci homologues RAD à décru dramatiquement avec une baisse de la divergence. Plus de 70 pourcents des loci étaient perdus lors de la comparaison de spécimens séparés par deux mutations sur un brin mitochondrial de 700 nucléotides. Ainsi, six espèces sur neuf ont été confirmées, et il se peut que des individus caractérisé par un même haplotype mitochondrial peuvent appartenir à des espèces disctinctes. À l'inverse, trois haplotypes mitochondriaux forment un clade bien supporté dans lequel aucune structure de population n'a été détecté, ce qui suggère une possible variation intraspécifique de l'ADN mitochondrial chez *Chrysogorgia*. Ainsi, les données RAD ont permis à peaufiner les interprétations des marqueurs mitochondriaux classiques utilisés dans les octocoraux pour délimiter les espèces et de découvrir la diversité détectée auparavant (Pante *et al.*, 2015).

Il existe cependant quelques difficultés liées au génotypage de SNPs RAD-tags (Pan, 2014) (Mastretta-Yanes *et al.*, 2015). En effet, **en laboratoire**, la qualité des réactifs peut-être hétérogène, les risques de contamination sont possible, des erreurs de pipetage peuvent subvenir, la sensibilité de l'enzyme à la qualité de l'ADN n'est pas toujours la même, ou encore des biais liés aux PCR. (Bonin *et al.*, 2004), (Baird *et al.*, 2008), (Peterson *et al.*, 2012), (Hohenlohe *et al.*, 2012), (Pan, 2014) De plus, il peut y avoir des erreurs **de séquençage** ou encore un séquençage aléatoire d'allèles et de loci (Meacham *et al.*, 2011), (Nielsen *et al.*, 2011), (Hohenlohe *et al.*, 2012), (Loman *et al.*, 2012), (Pan, 2014). Il peut aussi y avoir des erreurs **intrinsèques au génome** comme le polymorphisme sur les sites de restriction ou la méthylation du site de restriction (Davey *et al.*, 2013), (Gautier *et al.*, 2013), (Pan, 2014).

Les **techniques de Next-Next generation sequencing** qui ont vu le jour en 2012 permettent un séquençage sur une molécule unique et de ce passer de l'amplification clonale (Boy, 2014). Cependant, le taux d'erreur de séquençage est 10 fois plus élevé qu'avec le séquençage de type Sanger (Boy, 2014). L'appareil **MinION** (voir figure 9) d'**Oxford Nanopore Technologies** "deviendra l'approche par défaut du séquençage d'ADN circulaire pour étudier la variété des espèces" selon (Hargreaves et Mulley, 2015).

Le MinION est un dispositif portable pour les analyses moléculaires grâce à la technologie nanopore. Il est adaptable pour l'analyse de l'ADN, de l'ARN, des protéines ou de petites molécules avec un flux de production simple (nanoporetech).

Le principe de la technologie nanopore consiste à faire passer l'ADN le long d'un pore qui est formé par une première protéine qui permet de séparer les deux brins d'ADN (voir figure 10) (Boy, 2014). Puis le passage de l'ADN simple brin au sein de la seconde protéine provoque un **courant électrique** caractéristique de chaque base de l'ADN (Boy, 2014). Ce courant électrique et ensuite traduit en information numérique (Hargreaves et Mulley, 2015).

Cet appareil à déjà fait ses preuves, notamment pour séquencer rapidement les ARN et ADN de pathogènes comme celui du virus Ebola, donnant des informations scientifiques et

sur la santé publique en un temps record (Hoenen *et al.*, 2015) mais aussi pour réaliser de la taxonomie microbienne de haute résolution et en même temps dans diverses analyses de diversité microbienne via l'étude de l'ADN 16S (Benitez-Paez *et al.*, 2015). Ou encore pour la détection en temps réel des gènes de résistance aux antibiotiques, par exemple en cas de pic de Salmonelles dans un hôpital (Quick *et al.*, 2015).

FIGURE 9 – Le MinION

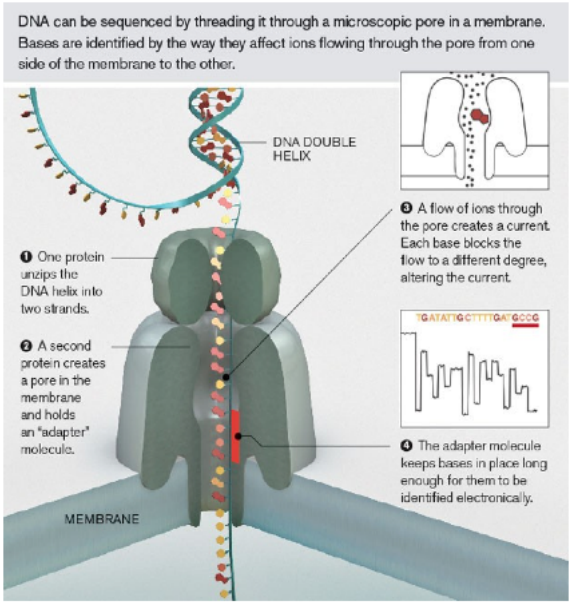

FIGURE 10 – Séquençage de l'ADN via Nanopore

Bibliographie

(2010). *Qualité des séquences produites par 454 : exemple de traitement*. Equipe d'accueil CIDAM, Faculté de Pharmacie, Université d'Auvergne.

(2014). *Le séquençage de Nouvelle Génération*.

(2014). *Séquençage de RAD tags : mise en oeuvre et applications*. Eric Pante laboratoire LIENSs UMR 7266 CNRS-Université de La Rochelle.

Amores, A., Catchen, J., Ferrara, A., Fontenot, Q. et Postlethwait, J. H. (2011). Genome evolution and meiotic maps by massively parallel DNA sequencing : spotted gar, an outgroup for the teleost genome duplication. *Genetics*, **188**(4) : 799–808.

Aurelle, D. (2009). *De l'évolution moléculaire à l'adaptation : approches de génétique des populations en milieu aquatique*. These, UNIVERSITE DE LA MEDITERRANEE CENTRE D'OCEANOLOGIE DE MARSEILLE.

Avise, J. C. et Walker, D. (2000). Abandon all species concepts ? A response. *Conservation Genetics*, **1**(1) : 77–80.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A. et Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**(10) : e3376.

Balzer, S., Malde, K. et Jonassen, I. (2011). Systematic exploration of error sources in pyrosequencing flowgram data. *Bioinformatics*, **27**(13) : i304–i309.

Bazin, E., Glémin, S. et Galtier, N. (2006). Population size does not influence mitochondrial genetic diversity in animals. *Science*, **312**(5773) : 570–572.

Benitez-Paez, A., Portune, K. et Sanz, Y. (2015). Species level resolution of 16S rRNA gene amplicons sequenced through MinIONTM portable nanopore sequencer. *bioRxiv*, page 021758.

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C. et Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Mol Ecol*, **13**(11) : 3261–73.

Broquet, T. et Petit, E. J. (2009). Molecular estimation of dispersal for ecology and population genetics. *Annual Review of Ecology, Evolution, and Systematics*, **40** : 193–216.

Champe, S. P. et Benzer, S. (1962). Reversal of mutant phenotypes by 5-fluorouracid : an approach to nucleotide sequences in messenger-RNA. *Proceedings of the National Academy of Sciences of the United States of America*, **48**(4) : 532.

Cruaud, A., Gautier, M., Galan, M., Foucaud, J., Sauné, L., Genson, G., Dubois, E., Nidelet, S., Deuve, T. et Rasplus, J.-Y. (2014). Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular biology and evolution*, **31**(5) : 1272–1274.

Davey, J. W. et Blaxter, M. L. (2010). RADSeq : next-generation population genetics. *Briefings in Functional Genomics*, **9**(5-6) : 416–423.

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K. et Blaxter, M. L. (2013). Special features of RAD Sequencing data : implications for genotyping. *Molecular Ecology*, **22**(11) : 3151–3164.

Eggen, A. (2003). Les approches génomiques pour l'identification de gènes d'intérêt économique : outils, applications et perspectives. *Renc. Rech. Ruminants*, **10**.

Gagnaire, P.-A., Broquet, T., Aurelle, D., Viard, F., Souissi, A., Bonhomme, F., Arnaud-Haond, S. et Bierne, N. (2015). Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evol Appl*, **8**(8) : 769–86.

Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., Morin, P. A., Narum, S. R., O'Brien, S. J., Roffler, G. *et al.* (2015). Genomics in conservation : case studies and bridging the gap between data and application. *Trends in ecology &amp ; evolution.*

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J.-M. et Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**(11) : 3165–3178.

Gilles, A., Meglécz, E., Pech, N., Ferreira, S., Malausa, T. et Martin, J.-F. (2011). Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC genomics*, **12**(1) : 245.

Gomez-Alvarez, V., Teal, T. K. et Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *The ISME journal*, **3**(11) : 1314–1317.

Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., Du, L., Egholm, M., Rothberg, J. M., Paunovic, M. *et al.* (2006). Analysis of one million base pairs of Neanderthal DNA. *Nature*, **444**(7117) : 330–336.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E. *et al.* (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research*, **21**(3) : 494–504.

Hargreaves, A. D. et Mulley, J. F. (2015). Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. *PeerJ*, **3** : e1441.

Hedgecock, D., Barber, P. H. et Edmands, S. (2007). Genetic approaches to measuring connectivity. *OCEANOGRAPHY-WASHINGTON DC-OCEANOGRAPHY SOCIETY-*, **20**(3) : 70.

Hellberg, M. E. (2009). Gene flow and isolation among populations of marine animals.

Herrera, S. et Shank, T. M. (2015). RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. *bioRxiv*, page 019745.

Hoenen, T., Groseth, A., Rosenke, K., Fischer, R., Hoenen, A. et Judson, S. (2015). Nanopore sequencing as a rapidly deployable Ebola outbreak response tool. *Emerg Infect Dis.*

Hohenlohe, P. A., Bassham, S., Currey, M. et Cresko, W. A. (2012). Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society of London B : Biological Sciences*, **367**(1587) : 395–408.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., Cresko, W. A. *et al.* (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet*, **6**(2) : e1000 862.

Kashi, Y. et King, D. G. (2006). Simple sequence repeats as advantageous mutators in evolution. *TRENDS in Genetics*, **22**(5) : 253–259.

Kitchen, A., Miyamoto, M. M. et Mulligan, C. J. (2008). A three-stage colonization model for the peopling of the Americas. *PLoS One*, **3**(2) : e1596.

Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M. et Seelow, D. (2011). Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One*, **6**(11) : e28 240.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J. et Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, **30**(5) : 434–439.

Lynch, M. et Conery, J. S. (2003). The origins of genome complexity. *science*, **302**(5649) : 1401–1404.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei,

M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. et Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057) : 376–80.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D. et Emerson, B. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular ecology resources*, **15**(1) : 28–41.

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D. I., Singer, M. et Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. *BMC bioinformatics*, **12**(1) : 451.

nanoporetech ().

Nielsen, R., Paul, J. S., Albrechtsen, A. et Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**(6) : 443–451.

Niu, B., Fu, L., Sun, S. et Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC bioinformatics*, **11**(1) : 187.

Noonan, J. P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Pääbo, S., Pritchard, J. K. *et al.* (2006). Sequencing and analysis of Neanderthal genomic DNA. *science*, **314**(5802) : 1113–1118.

O'Brien, S. J. (1994). A role for molecular genetics in biological conservation. *Proceedings of the National Academy of Sciences*, **91**(13) : 5748–5755.

Palumbi, S. R. (1994). Genetic divergence, reproductive isolation, and marine speciation. *Annual review of ecology and systematics*, pages 547–572.

Pante, E., Abdelkrim, J., Viricel, A., Gey, D., France, S. C., Boisselier, M. C. et Samadi, S. (2015). Use of RAD sequencing for delimiting species. *Heredity (Edinb)*, **114**(5) : 450–9.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. et Hoekstra, H. E. (2012). Double digest RADseq : an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, **7**(5) : e37 135.

Poptsova, M. S., Il'Icheva, I. A., Nechipurenko, D. Y., Panchenko, L. A., Khodikov, M. V., Oparina, N. Y., Polozov, R. V., Nechipurenko, Y. D. et Grokhovsky, S. L. (2014). Non-random DNA fragmentation in next-generation sequencing. *Scientific reports*, **4**.

Prodromou, C., Savva, R. et Driscoll, P. C. (2007). DNA fragmentation-based combinatorial approaches to soluble protein expression Part I. Generating DNA fragment libraries. *Drug Discov Today*, **12**(21-22) : 931–8.

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., Neal, K., Nye, K., Peters, T., De Pinna, E., Robinson, E., Struthers, K., Webber, M., Catto, A., Dallman, T. J., Hawkey, P. et Loman, N. J. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol*, **16** : 114.

Reitzel, A. M., Herrera, S., Layden, M. J., Martindale, M. Q. et Shank, T. M. (2013). Going where traditional markers have not gone before : utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Mol Ecol*, **22**(11) : 2953–70.

Reusch, T. B., Ehlers, A., Hämmerli, A. et Worm, B. (2005). Ecosystem recovery after climatic extremes enhanced by genotypic diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(8) : 2826–2831.

Selkoe, K. et Toonen, R. J. (2011). Marine connectivity : a new look at pelagic larval duration and genetic metrics of dispersal. *Marine Ecology Progress Series*, **436** : 291–305.

Sengenes, J. (2012). *Développement de méthodes de séquençage de seconde génération pour l'analyse des profils de méthylation de l'ADN*. These, Université Pierre et Marie Curie-Paris VI.

Shafer, A. B., Wolf, J. B., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., Colling, G., Dalén, L., De Meester, L., Ekblom, R. *et al.* (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology &amp ; Evolution*, **30**(2) : 78–87.

Shendure, J. et Aiden, E. L. (2012). The expanding scope of DNA sequencing. *Nature biotechnology*, **30**(11) : 1084–1094.

Stölting, K. N., Nipper, R., Lindtke, D., Caseys, C., Waeber, S., Castiglione, S. et Lexer, C. (2013). Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Mol Ecol*, **22**(3) : 842–55.

Syed, F., Grunenwald, H. et Caruccio, N. (2009). Next-generation sequencing library preparation : simultaneous fragmentation and tagging using in vitro transposition. *Nature Methods*, **6**(11).

Wall, J. D. et Kim, S. K. (2007). Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genet*, **3**(10) : 1862–6.

Ward, R., Woodwark, M. et Skibinski, D. (1994). A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. *Journal of fish biology*, **44**(2) : 213–232.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G. T. *et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. *nature*, **452**(7189) : 872–876.

Annexe : Historique des technologies

— 1953 : Découverte de la molécule d'ADN par Watson et Crick
— 1973 : Première séquence de 24 paires de bases publiée (Walter Gilbert and Allan Maxam 1973. The nucleotide sequence of the lac operator)
— 1975 : Southern Blot
— 1977 : Séquençage Sanger Gilbert
— 1982 : Genbank started
— 1983 : Developpement des PCR (Polymerase Chain Reaction)
— 1987 : Premier sequenceur automatique : Applied Biosystems Prism 373
— 1990 : Séquençage par mesure de la fluorescence
— 1995 : Puces à ADN (microarray)
— 1996 : Sequenceur à capillaires : ABI 310
— 1998 : Genome de *Caenorhabditis elegans* séquencé
— 2000 : Evolution des puces à ADN
— 2003 : Séquençage du génome humain. 3 milliards de dollars, 13 ans
— 2005 : 1st 454 Life Sciences Next Generation Sequencing system : GS 20 System
— 2006 : 1st Solexa Next Generation Sequencer : Genome Analyzer
— 2007 : 1st Applied Biosystems Next Generation Sequencer : SOLiD
— 2007 : Séquençage d'un individu (JC Venter) Méthode Sanger (Levy et al. Plos Bio 2007)
— 2008 : Séquençage d'un individu (J.D Watson) Méthode haut débit (454 Roche) (Wheeler *et al.*, 2008), 1 million de dollar, 2 mois
— 2009 : 1st Helicos single molecule sequencer : Helicos Genetic Analyser System 2011 : 1st Ion Torrent Next Generation Sequencer : PGM
— 2011 : 1st Pacific Biosciences single molecule sequencer : PacBio RS Systems
— 2012 : Oxford Nanopore Technologies demonstrates ultra long single molecule reads
— 2012 : "Next-next generation Sequencing"