

# État de l'art

L'apport des NGS pour inférer la structure de population

Pierre-Louis STENGER

# Table des figures

1	Les NGS permettent de travailler à différentes échelles(Shendure et Aiden, 2012) . . . . .	4
2	Evolution très rapide des instruments, des débits et des coûts (premier séquençage sur 454 en Janvier 2008) . . . . .	5
3	Aperçu de la technologie 454 (Sengenes, 2012), PTP : PicoTiter Plate, PPi : pyrophosphate inorganique, APS : adénosine phosphosulphate, ATP : adénosine triphosphate. . . . .	7
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 <i>et al.</i> , 2012) . . . . .	8
5	étapes pour obtenir des séquences d'ADN via la technique Rad seq . . . . .	9
6	Comparaison synthétique 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 <i>et al.</i> , 2011). . . . .	9
7	Phylogéographie de <i>N. vectensis</i> : Nova Scotia (NS), Massachusetts (MA), Maryland (MD), and South Carolina (SC) (Reitzel <i>et al.</i> , 2013). . . . .	10
8	Phylogénie des carabes obtenues par ADN nucléaire (A), ADN mitochondrial (B) et via la technique de RAD-Seq (C)(Cruaud <i>et al.</i> , 2014) . . . . .	11
9	Le MinION . . . . .	12
10	Séquençage de l'ADN via Nanopore . . . . .	12

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

Au cours du XX<sup>e</sup> siècle, les avancées technologiques permettant d'accéder directement à l'ADN (**Acide Desoxyribonucléique**) 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 individu 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 utilisés.

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) (Avisé 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 élevée 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 corrélérer 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.

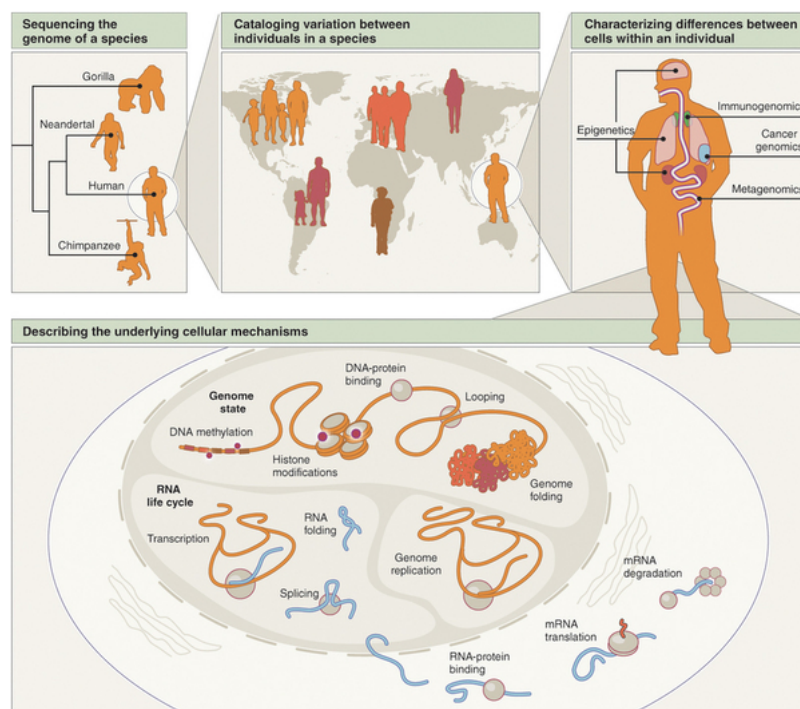


FIGURE 1 – Les NGS permettent de travailler à différentes échelles (Shendure et Aiden, 2012)

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.

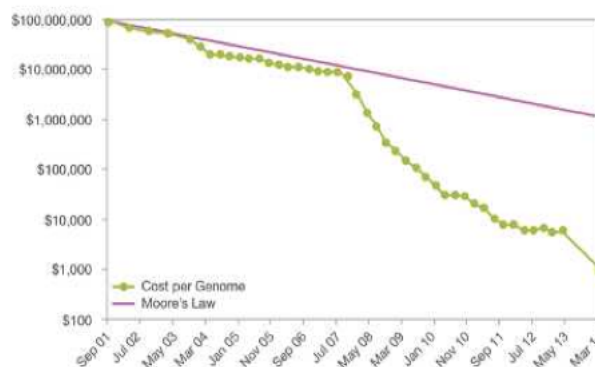


FIGURE 2 – Evolution très rapide des instruments, des débits et des coûts (premier séquençage sur 454 en Janvier 2008)

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.

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 colonies, comme dans la technique Illumina), puis ces clones sont **séquencés** par une des techniques vu précédemment. 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 genitalium* (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  $\mu\text{m}$  (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  $\mu\text{m}$ ) (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  $\mu\text{m}$ ) (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 créée 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 malgré 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 noyau 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 adjacent (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 ambiguës* peuvent apparaître 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 répliques artificiels :** ils correspondent 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 même 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 où s'effectue la réaction de pyroséquencage (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 a 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).

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),

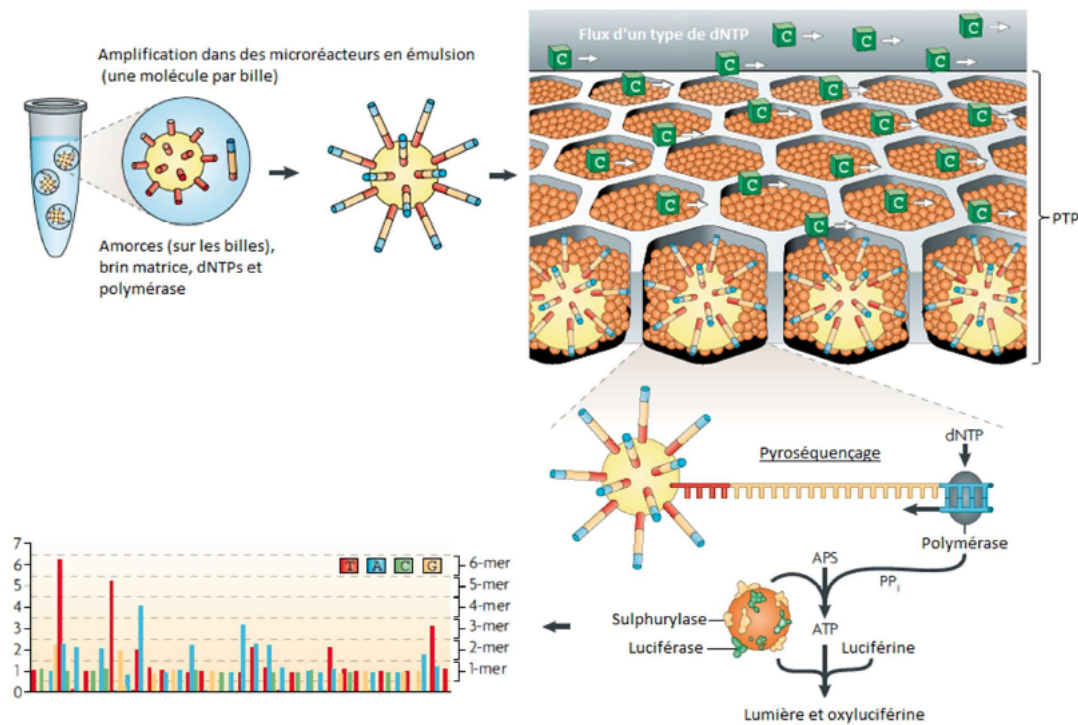


FIGURE 3 – Aperçu de la technologie 454 (Sengenes, 2012), PTP : PicoTiter Plate, PPi : pyrophosphate inorganique, APS : adénosine phosphosulphate, ATP : adénosine triphosphate.

mais aussi dans la génétique des populations, et donc dans la compréhension de l'évolution (Davey et Blaxter, 2010).

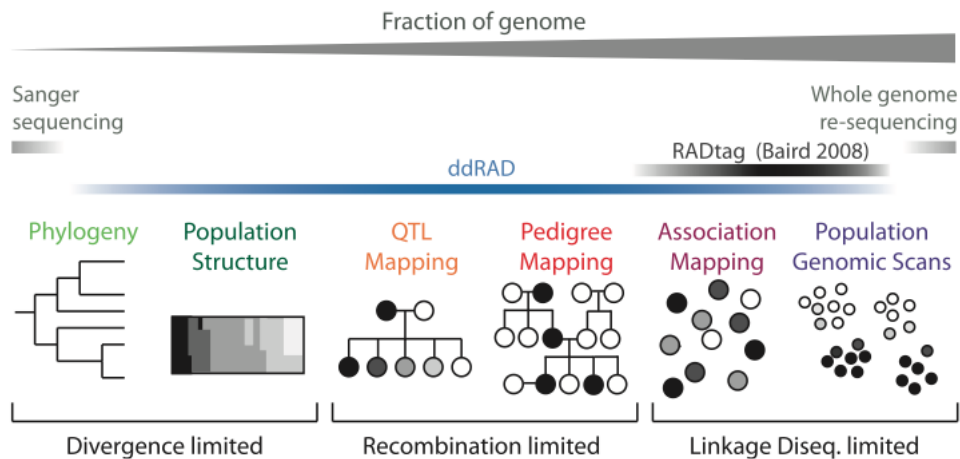


FIGURE 4 – Résolution des marqueurs de type RAD : Un procédé de génotypage souple peut être utilisé pour optimiser le nombre de marqueurs génétiques pour une approche expérimentale spécifique dans un système biologique donné (Peterson *et al.*, 2012)

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).

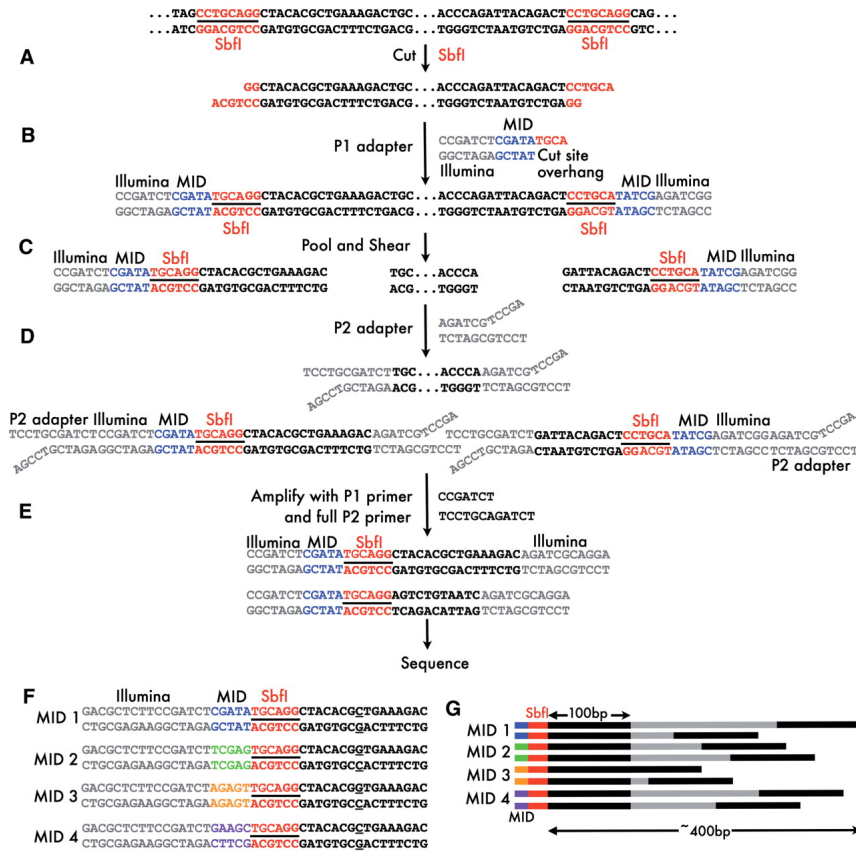


FIGURE 5 – étapes pour obtenir des séquences d'ADN via la technique Rad seq

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

- (Hohenlohe *et al.*, 2010) ont utilisé la **technique Illumina-sequenced RAD tags** 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 permis 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 été réalisé sur des *Lepisosteus* (*Lepisosteus oculatus*) (Amores *et al.*, 2011), ce qui a permis de découvrir que c'est une lignée de poissons qui a divergé 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).

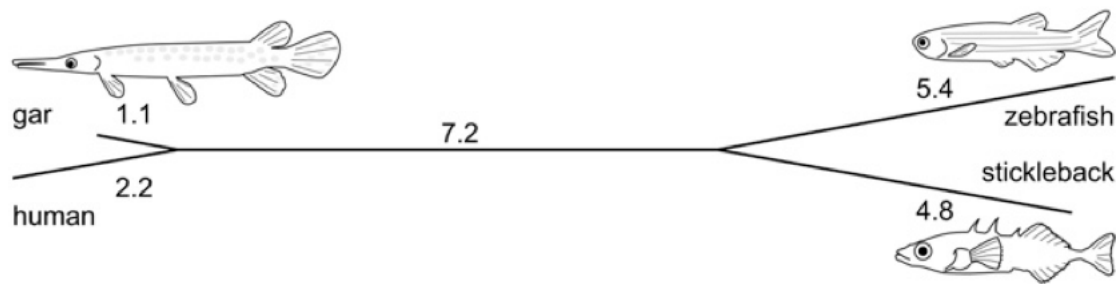


FIGURE 6 – Comparaison synthétique de deux téléostéens (zebrafish and stickleback) par rapport au lepisosteus et à l’homme. La longueur des branches est proportionnelle à l’estimation du nombre de divergence en terme de chromosome entre les espèces (Amores *et al.*, 2011).

- 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’à du polymorphisme ancestral partagé. Ces résultats permettent donc d’expliquer l’origine de ces ”génomés 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 parmi 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).
- 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 permis 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 espèce 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

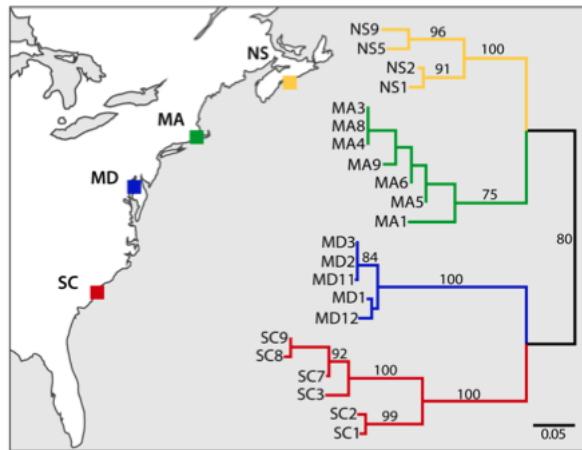


FIGURE 7 – Phylogéographie de *N. vectensis* : Nova Scotia (NS), Massachusetts (MA), Maryland (MD), and South Carolina (SC) (Reitzel *et al.*, 2013).

nucléaire et mitochondrial ne donnaient pas les mêmes résultats (voir figure 8).

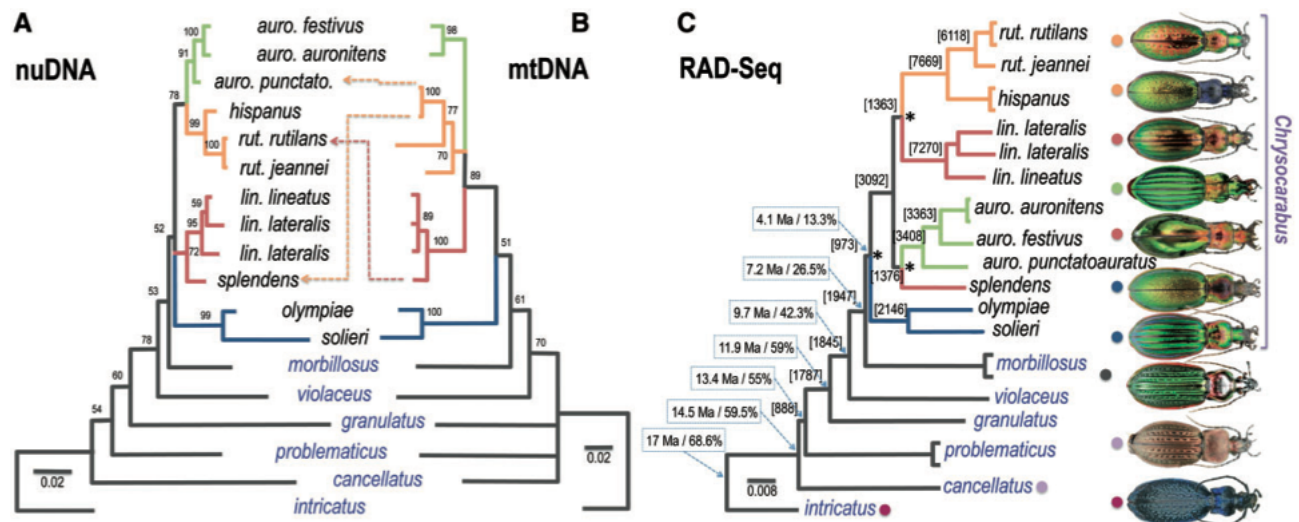


FIGURE 8 – Phylogénie des carabes obtenues par ADN nucléaire (A), ADN mitochondrial (B) et via la technique de RAD-Seq (C) (Cruaud *et al.*, 2014).

— C'est aussi une technique très utilisée 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 a 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 caractérisés par un même haplotype mitochondrial peuvent appartenir à des espèces distinctes. 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 est 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

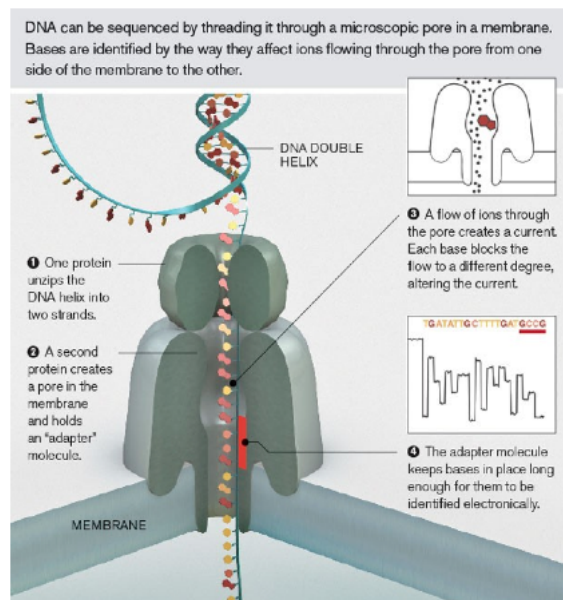


FIGURE 10 – Séquençage de l'ADN via Nanopore

# 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 MinION™ portable nanopore sequencer. *bioRxiv*, page 021758.
- BONIN, A., BELLEMAIN, E., BRONKEN EIDSEN, 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.
- CHAMPE, S. P. et BENZER, S. (1962). Reversal of mutant phenotypes by 5-fluorouracil : 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**.
- 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.
- 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) : e1000862.
- 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) : e28240.
- LOMAN, N. J., MISRA, R. V., DALLMAN, T. J., CONSTANTINIDOU, C., GHARBIA, S. E., WAIN, J. et PALLAN, 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., ATTIIYA, 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.
- 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) : e37135.
- 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.
- 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.

- 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.
- 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 : Développement des PCR (Polymerase Chain Reaction)
- 1987 : Premier séquenceur automatique : Applied Biosystems Prism 373
- 1990 : Séquençage par mesure de la fluorescence
- 1995 : Puces à ADN (microarray)
- 1996 : Séquenceur 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"