

In the beginning there was nothing.  
God said : Let there be light!  
There was still nothing, but now you could see it.  
(Dave Thomas)

# Etat de l'art: "Apport des NGS pour inférer la structure de population"

Pierre-Louis STENGER<sup>1</sup>

<sup>1</sup>Université de La Rochelle, Master 2 Gestion de l'Environnement et Ecologie Littorale (GEEL)

## Résumé

Résumé

**Key words :** *NGS, Population, Génétique*

## Introduction

### 1 Mémo notes

Apport des NGS pour inférer la structure de population

-> Qu'est ce que ça nous apprend de plus ?

Méthodes traditionnelles en génétiques des populations -> Dérive, mutation et migration. -> "On était bien" BSB(C)

NGS ont permis de prendre en compte la sélection et ça à mis 30 ans de recherches en question.

Basin (2006) -> Taille de la population décorrélé au nombre de diversité mitochondriale -> Il faudrait faire un balayage sélectif

RAD SEQ -> Moins de marqueur que 454 Amélia -> Dauphin commun -> Structure population en Atlantique -> Augmentation du nombre de marqueur -> Augmentation de la puissance statistique Car génétique ne montrait qu'une seule population, alors que écotoxicologie et isotopes en montraient plusieurs. On va travailler sur marqueur de sélection -> Adaptation locale -> Ce sont les marqueurs Outlayer -> FST -> Avec les FST on revient dans un système circulaire (on a deux populations, on veut les différencier, on a un FST, on regarde si c'est supérieur. Aurait-on le même résultats avec 5 population ? Si on a 1000 marqueurs, il y en a forcement -> Statistiques fréquentistes (c'est à l'opposé des statistiques bayésiennes)) -> On coupe le jeu de données en deux.

Do it yourself ABC (Marie Louis, usé à la fin de thèse) -> Histoire évolutive

Eric :

— Séquenceur nanoport (Oxford)

— Micness -> Séquence en NGS les microsatellites, Marie Suez et al. 2015

Séquenceur à plaques -> N'est pas un NGS

### 2 A creuser

PDF -> SKIM

<http://www.sciencemag.org/content/312/5773/570.short> -> E. Bazin, S. Glemin, N. Galtier, Science 312, 570 (2006).

Au sein des espèces, la diversité génétique est pensée pour refléter la taille de la population , l'histoire , l'écologie et capacité d'adaptation. Avec l'utilisation d'une collection complète d'ensembles de données de polymorphisme couvrant 3000 espèces animales , 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 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, en revanche, est adapté à ces attentes intuitives. La distribution inattendue de la diversité mitochondrial est expliquée par

l'évolution adaptative récurrente, contestant la théorie neutraliste de l'évolution et de remettre en question la pertinence de l'ADNmt dans les études de la biodiversité et de conservation.

Thèse HDR de Didier AURELLE :

La majorité des études de génétique des populations réalisées jusqu'à présent sont faites sous l'hypothèse d'une évolution neutre ou quasi neutre des marqueurs employés. Or d'une part la neutralité de divers marqueurs couramment utilisés peut être remise en cause qu'il s'agisse de l'ADN mitochondrial (Bazin et al., 2006) ou des microsatellites (Kashi et King, 2006). D'autre part l'étude de l'impact des changements environnementaux et notamment du changement global sur les populations naturelles nécessite de prendre en compte les processus sélectifs, ou au moins de se poser la question des capacités d'évolution des organismes par rapport à la vitesse de ces changements (Berteaux et al., 2004) et donc de s'intéresser à la diversité génétique adaptative et à ses relations avec la diversité neutre.

Via Thèse HDR de Didier AURELLE : Calderon I., Garrabou J., Aurelle D. (2006) Evaluation of the utility of COI and ITS markers as tools for population genetic studies of temperate gorgonians. Journal of Experimental Marine Biology and Ecology, 336, 184-197. doi :10.1016/j.jembe.2006.05.006. :

L'ADN mitochondrial présente des propriétés qui en font a priori un marqueur de choix pour une étude de phylogéographie : généralement pas de recombinaison, un taux de dérive plus important que l'ADN nucléaire et un niveau élevé de polymorphisme (Avise, 2000). Cependant comme l'ont noté Ballard et Whitlock (2004) ses propriétés biologiques peuvent aussi représenter des inconvénients : il ne reflète que l'histoire des lignées maternelles et l'absence de recombinaison associée à la forte densité en gènes le rend aussi plus sensible à la perte de variabilité suite aux fixations de mutations avantageuses ("genetic draft" ; Bazin et al., 2006). Malgré tout, ce marqueur permet de fournir une première image de la structuration génétique d'une espèce sous réserve de garder à l'esprit les contraintes précédentes.

Review de l'article Bzin et al. 2006 : Kitchen, A., Miyamoto, M. M., Mulligan, C. J. (2008). A three-stage colonization model for the peopling of the Americas. PLoS One, 3(2), e1596.

Bazin et al . (Rapports, le 28 avril 2006 , p . 570 ) n'a trouvé aucune relation entre ADNmt diversité et la taille de la population lorsque l'on compare entre grands groupes d'animaux . Nous montrons empiriquement que les espèces avec de petites populations , telle que représentée par les mammifères euthériens , présentent une corrélation positive entre l'ADNmt et variation des allozymes , ce qui suggère que la diversité peut ADNmt corrélation avec la taille de la population de ces animaux.

Do it yourself : Thèse Marie Louis Chapitre 6, page 162 :

We investigated the demographic history best describing the genetic dataset of the combined microsatellite and mtDNA markers using a coalescent-based Approximate Bayesian Computation (ABC) approach (Beaumont et al. 2002 ; Bertorelle et al. 2010 ; Csillary et al. 2010, the general principle of this analysis is presented in Chapter 2.2c).

RAD Seq Amélia :

Amélia Viricel, Eric Pante, Willy Dabin, Benoit Simon-Bouhet. Applicability of RAD-tag geno- typing for inter-familial comparisons : empirical data from two cetaceans. Molecular Ecology Resources, Blackwell, 2014, 14 (3), pp.597-605. <10.1111/1755-0998.12206>. <hal-00908459>

Viricel, A., Rosel, P. E. (2014). Hierarchical population structure and habitat differences in a highly mobile marine species : the Atlantic spotted dolphin. Molecular ecology, 23(20), 5018-5035.

blabla [?] blabla  
blabla [?]

### 3 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 sequençeur automatique : Applied Biosystems Prism 373
- 1990 : Séquençage par mesure de la fluorescence
- 1995 : Puces à ADN (microarray)
- 1996 : Sequençeur 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. Nature 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"

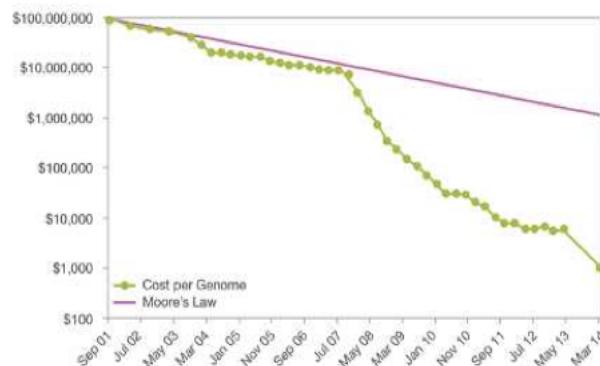


FIGURE 1 – Evolution très rapide des instruments, des débits et des coûts (premier séquençage sur 454 en Janvier 2008)

10 ans → Prix divisé par 100 000

## 4 Définition NGS

NGS → Séquençage à haut débit, Next Generation Sequencing, NextGen Sequencing, NGS, Massively Parallel Sequencing

## 5 Grands principes NGS

Sanger : 96 ADNs différents analysés en une fois NGS : Millions d'ADNs différents analysés en une fois → Saut technologique

- Intégration (système combinant les avantages de la PCR et des puces)
- Parallélisation (PCR multiplex)
- Miniaturisation

NGS à permis le séquençage entier du génome humain → Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., ... Gomes, X. (2008). The complete genome of an individual by massively parallel DNA sequencing. nature, 452(7189), 872-876.

Voelkerding 2010 Matériel biologique : ADN ou ARN Etapes communes aux différentes technologies :

- Fragmentation enzymatiques de l'ADN
- Préparation d'une banque (library) par ligation d'adaptateurs
- Amplification clonale
- Séquençage générant des signaux (luminescent ou fluorescent)
- Détection des signaux émis et conversion en séquence

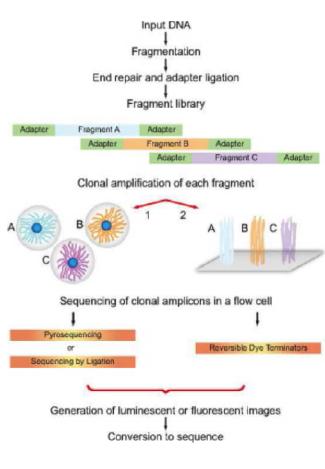


FIGURE 2 – Grands principes NGS

	454 FLX/Roche	Solexa/Illumina	Solid/Life Tech
Préparation de banques de NGS	<b>ADN génomique</b> → <b>Fragmentation</b> → <b>Ligation d'adaptateurs</b> But ajout d'adaptateurs : Utiliser toujours la même amorce pour PCR et Séquençage quelque soit l'ADN à séquencer		
Amplification clonale	<b>Système miniature d'amplification</b> PCR en émulsion sur une bille  Amplification clonale dans des microréacteurs	<b>« Bridge » PCR sur un support plan (=Flow Cell)</b>  Colonies d'ADN = colonies	PCR en émulsion sur une bille  Amplification clonale dans des microréacteurs
Support pour le séquençage	Billes insérées dans des micropuits	Flow Cell ayant servi à l'amplification	Billes fixées à une lame de verre
Séquençage en temps réel : 1 bille ou 1 cluster-> 1 Read	*Séquençage par synthèse *Pyroséquençage *Longs fragments	*Séquençage par synthèse *dye-terminator reversible *petits fragments	*Séquençage réversible par ligation *encodage de 2 bases *petits fragments

FIGURE 3 – Comparaison des principes des trois premières technologies de NGS

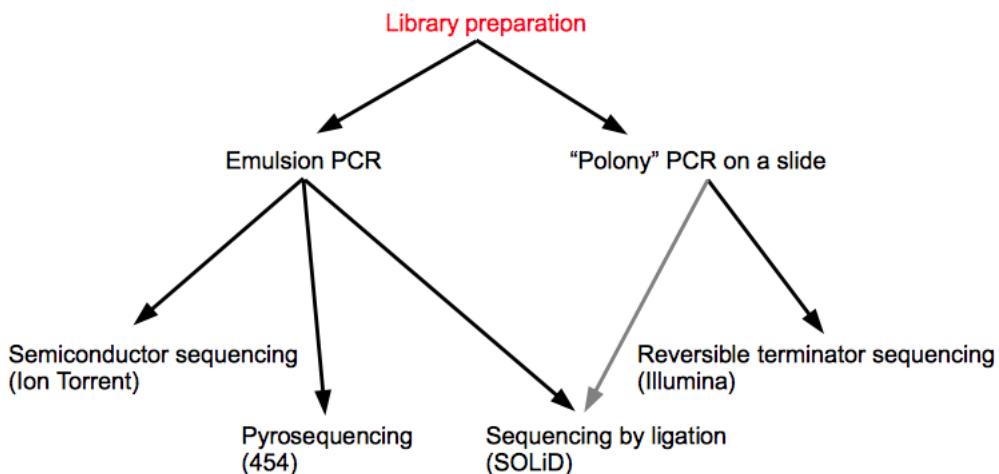


FIGURE 4 – Next Generation Sequencing, Amplified Single Molecule Sequencing

## 6 Library preparation

## 7 Emulsion PCR

## 8 "Polony" PCR on a slide

## 9 Les différentes plateformes (Sommaire)

454 Sequencing / Roche :(1)

- GS Junior System
- GS FLX+ System
- Illumina (Solexa)(1) :
- HiSeq System
- Genome analyzer IIx
- MySeq

Applied Biosystems - Life Technologies :(1)

- SOLiD 5500 System
- SOLiD 5500xl System

Ion Torrent - Life Technologies :(1)

- Personal Genome Machine (PGM)

— Proton

Helicos :(2)

- Helicos Genetic Analysis System

Pacific Biosciences :(2)

- PacBio RS

Oxford Nanopore Technologies :(2)

- GridION System

— MinION

(1) = Next Generation Sequencing, Amplified Single Molecule Sequencing

(2) = Third Generation Sequencing, Next Next Generation Sequencing, Single Molecule Sequencing

## Which Next Generation Sequencer to choose for your project ?

	Capacity	Speed	Read Length	Homopolymers	Cost/run	Amplification
<b>454 Roche</b>	35-700 Mb	10-23 hours	400-700 bp	-	5.000 €	Yes
<b>SOLiD</b>	90-180 Gb	7-12 days	75 bp	+	5.000 €	Yes
<b>Illumina</b>	6-600 Gb	2-14 days	100-250 bp	+	10.000-20.000 €	Yes
<b>Ion Torrent</b>	20 Mb- 1Gb	4,5 hours	200 bp	-	1.000-2.000 €	Yes
<b>Helicos</b>	35 Gb	8 days	35 bp	+	20.000 €	No
<b>PacBio</b>	1Gb	30 minutes	3000 bp	+	600-800 €	No

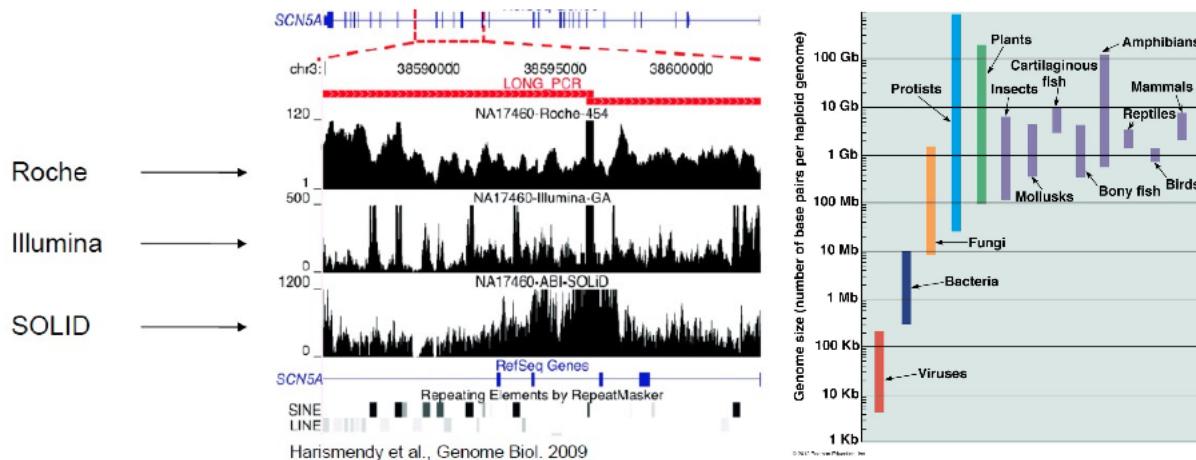


FIGURE 5 – Que choisir ?

Company/Platform	Sequencing	Amplification	Read length	Max. Output	Run time	Pros/Cons
Roche 454 GS FLX+	SBS Pyro	emPCR	700 bp* (SE, PE)	700 Mb	10-23 h	Pro: Long reads, short run time Con: High Mb cost, homopolymer errors
Roche 454 GS Junior	SBS Pyro	emPCR	400 bp* (SE, PE)	35 Mb	10 h	Same as GS FLX+ Additional Con: Lowest output of small scale instruments
Illumina HiSeq 1000	SBS RDT	Bridge PCR	36-101 bp (SE, PE)	≤150 Gb	1.5-8.5 days	Pro: Ultra high output, ease of use Con: No run scalability like SOLiD 5500
Illumina HiSeq 2000	SBS RDT	Bridge PCR	36-101 bp (SE, PE)	≤300 Gb	2.5-11 days	Same as HiSeq 1000
Illumina GAIIx	SBS RDT	Bridge PCR	36-151 bp (SE, PE)	≤95 Gb	2-14 days	Pro: Mature platform Con: Superseded by HiSeq, higher Mb cost
Illumina MiSeq	SBS RDT	Bridge PCR	36-151 bp (SE, PE)	>1 Gb	4-27 h	Pro: Proven chemistry, fully automated workflow Con: Unproven instrument
Illumina HiScanSQ	SBS RDT	Bridge PCR	100 bp (SE, PE)	≤150 Gb	1.5-8.5 days	Pro: Dual use instrument (microarray) Con: Higher Mb cost than HiSeq
Life Technologies 5500	SBL	emPCR	35-75 bp (SE, PE)	77 Gb	2-7 days	Pro: Ultra high output, scalable runs allow sequencing on part flow cell Con: Shorter reads than other platforms, longer time to clonal template prep than Illumina
Life Technologies 5500XL (4bp)	SBL	emPCR	35-75 bp (SE, PE)	155 Gb	2-7 days	Same as 5500
Life Technologies Ion Torrent	SBS H+	emPCR	316+318 chip >100 bp (SE)	316->100 Mb 318->1 Gb	2 h+	Pro: Label-free chemistry—cheap and fast, highly scalable, long read length potential Con: Homopolymer errors, no PE yet, laborious template preparation but semi-automatable

Note: Specifications for all platforms were derived from company websites. \*Mode read length: the individual fragment read length is variable. SBS = Sequencing-by-synthesis; Pyro = Pyrosequencing; RDT = reverse dye terminator chemistry; H+ = Hydrogen ion detection; SBL = Sequencing-by-ligation; emPCR = emulsion PCR; SE = Single-end read; PE = Paired-end read; Mb = Megabases; Gb = Gigabases; bp = base pairs.

Totthill 2011

Séquenceur de pailleasse |

FIGURE 6 – Comparaison NGS

## 10 454 Sequencing/ Roche

### 10.1 GS Junior System

### 10.2 GS FLX + System

## 11 Illumina

### 11.1 HiSeq System

### 11.2 Genome analyser IIx

### 11.3 MySeq

## 12 Applied Biosystems - Life Technologies

### 12.1 SOLiD 5500 System

### 12.2 SOLiD 5500-v1 System

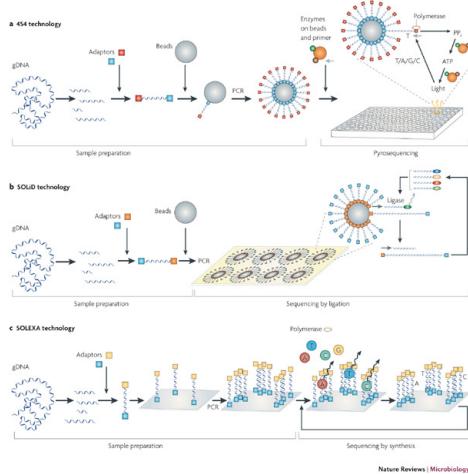


FIGURE 7 – Plusieurs méthodes

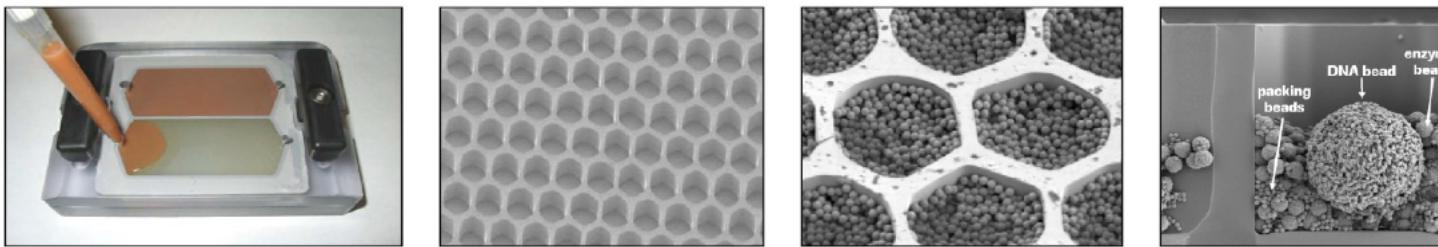


FIGURE 8 – sequencage454

- Pas de système optique
- Support de séquençage = surface semi-conducteur
- Mesure du pH (ion H<sup>+</sup> produit par l'ADN polymerase lors de l'incorporation de chaque base)
- Erreurs de séquençage recurrentes spécifiques comme des erreurs dans les homopolymers -> indels artéfactuels

### 13.1 Personal Genome Machine (PGM)

### 13.2 Proton

## 14 Next nexte gen :

- Séquençage sur molécule unique
- Pas d'amplification clonale
- Inconvénient actuel : Taux élevé d'erreurs de séquences (taux d'erreur de séquençage 10 fois plus élevé qu'avec le séquençage Sanger selon <http://www.math-info.univ-paris5.fr/rozen/Analyse-Genome-Tumoral/Exposes>)

## 15 Helicos

### 15.1 Helicos Genetic Analysis System

## 16 Pacific Biosciences

### 16.1 PacBio RS

## 17 Oxford Nanopore Technologies

MinION PromethION



FIGURE 9 – Nanoport



FIGURE 10 – Nanoport

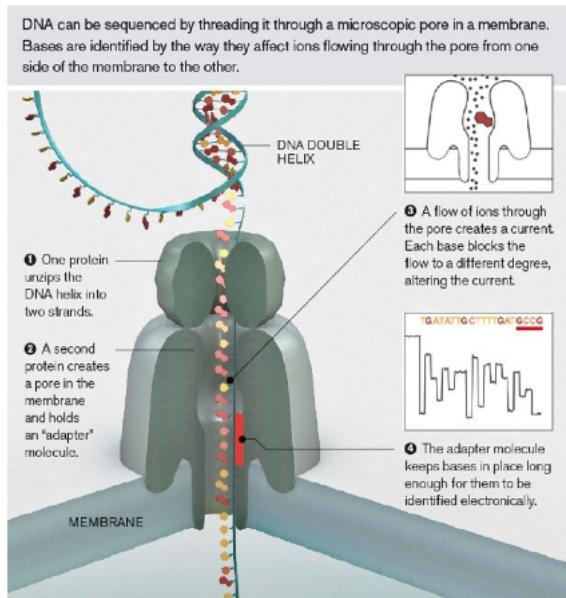


FIGURE 11 – Nanoport

Hargreaves and Mulley (2015), Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. PeerJ 3 :e1441 ; DOI 10.7717/peerj.1441 -> "Le MinION deviendra l'approche par défaut du séquençage d'ADN circulaire dans la variété des espèces"

Une plaque 900 dollars 48 plaques 24 00 dollars

## 17.1 GridION System

## 17.2 MinION

## 18 MicNeSs

Suez, M., Behdenna, A., Brouillet, S., Graça, P., Higuet, D. and Achaz, G. (2015), MicNeSs : genotyping microsatellite loci from a collection of (NGS) reads. Molecular Ecology Resources. doi : 10.1111/1755-0998.12467  
Les microsatellites sont largement utilisés dans la génétique des populations pour découvrir événements évolutifs récents. Ils sont généralement génotypés en utilisant un séquenceur à capillaire, dont la capacité est

généralement limitée à 9, au plus 12 loci pour chaque terme, et dont l'analyse est une tâche fastidieuse qui est effectué à la main. Avec la montée de séquençage de nouvelle génération (NGS), un plus grand nombre de lieux et de personnes sont disponibles à partir de séquençage : par exemple, sur un seul passage d'un GS junior, 28 loci de 96 personnes sont séquencées avec une couverture 30X. Suez et al 2015 ont développé un algorithme pour génotyper automatiquement et efficacement les microsatellites à partir d'un recueil de lectures triés par individu (par exemple amplifications par PCR spécifiques d'un locus ou d'une collection de lit qui englobent un locus d'intérêt). Comme le séquençage et l'amplification par PCR introduisent des insertions ou délétions artefactuelles, l'ensemble de lit à partir d'un seul allèle microsatellite montre plusieurs variantes de longueur. Les déduits de l'algorithme, sans alignement, la vraie inconnue allèle (s) de chaque individu à partir des distributions observées de microsatellites longueur de tous les individus. MicNeSs, une implémentation de Python de l'algorithme, peut être utilisé pour le génotype toute locus microsatellite de tout organisme et a été testé sur 454 données de pyroséquençage de plusieurs loci de mouches des fruits (espèce de modèle) et cerfs rouges (espèce nonmodel). Sans aucune parallélisation, il génotype automatiquement 22 loci de 441 personnes en 11 heures sur un ordinateur standard. La comparaison des inférences MicNeSs la méthode standard montre un excellent accord, avec quelques différences illustrant les avantages et les inconvénients des deux méthodes.

## 19 Les limitations des NGS

PDF NGS

### Table des figures

### Liste des tableaux

### ANNEXE : Emulsion

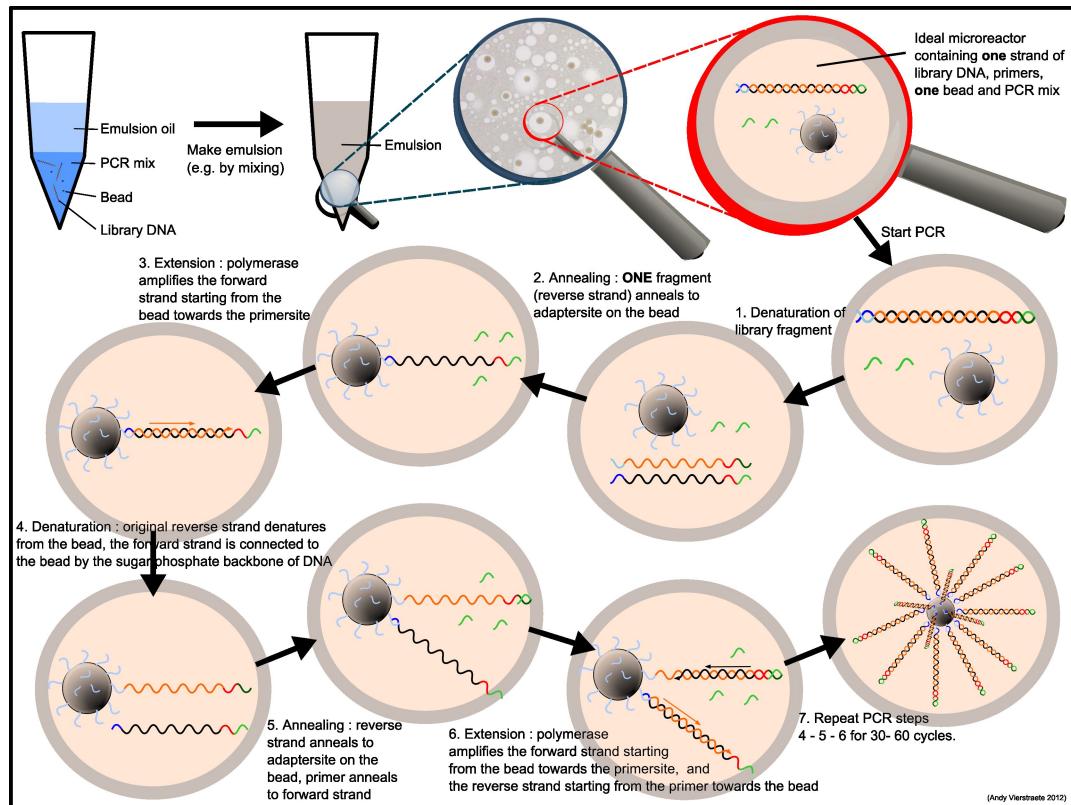


FIGURE 12 – Emulsion PCR expliquée en détail2

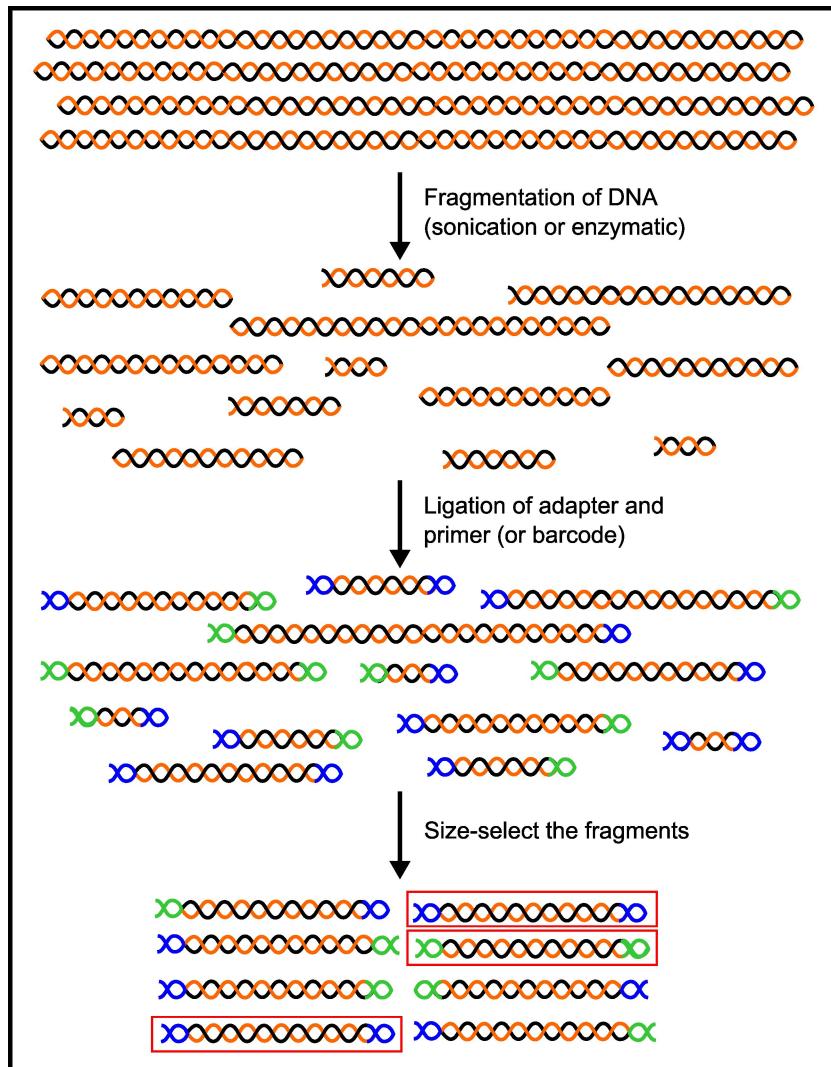


FIGURE 13 – Emulsion PCR expliquée en détail2

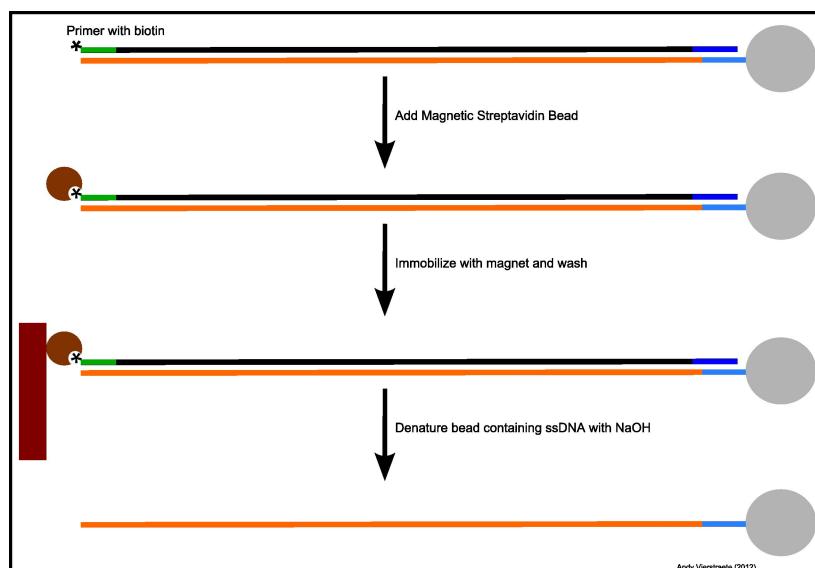


FIGURE 14 – Emulsion PCR expliquée en détail2

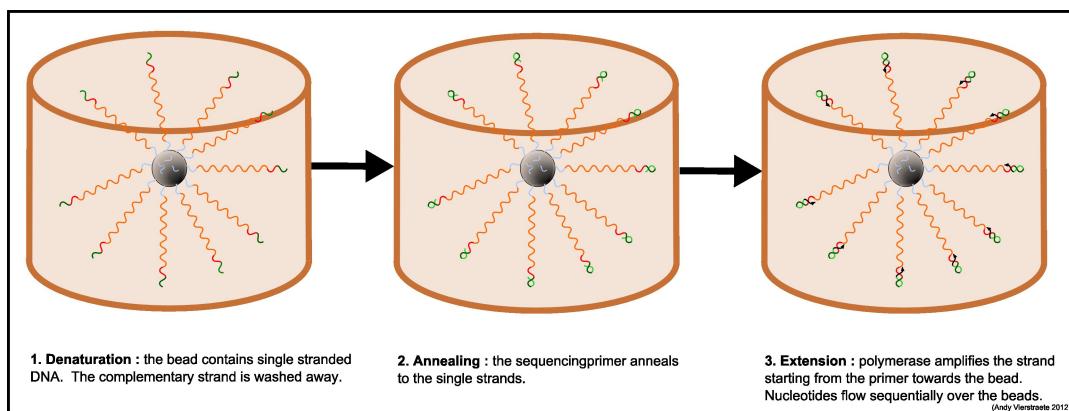


FIGURE 15 – Emulsion PCR expliquée en détail2

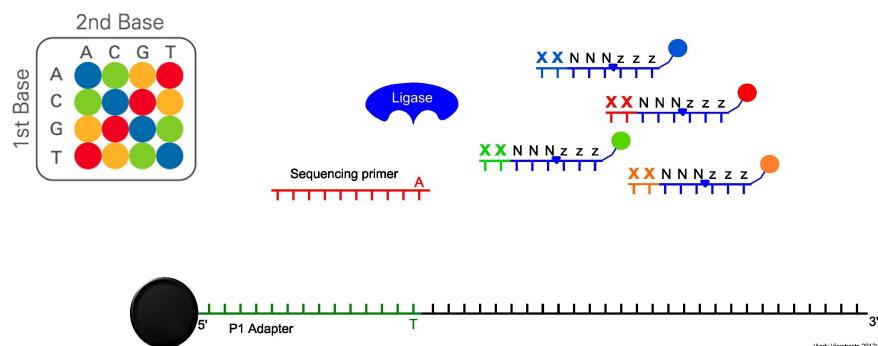


FIGURE 16 – Emulsion PCR expliquée en détail2

## ANNEXE : Polony

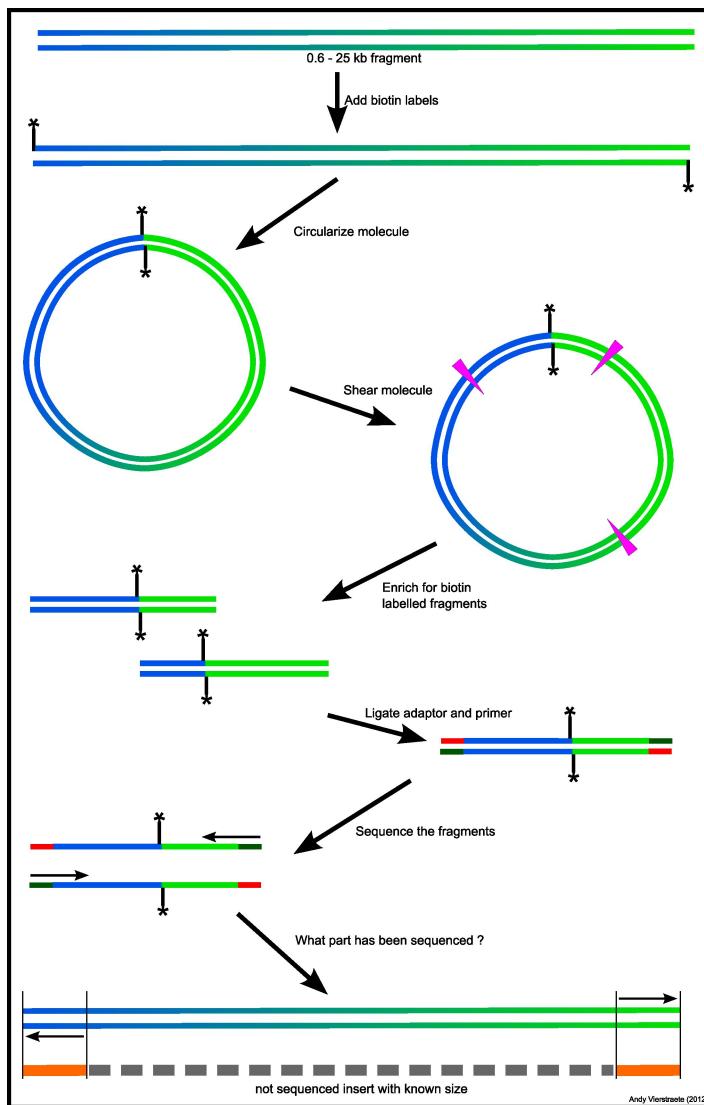


FIGURE 17 – Emulsion PCR expliquée en détail2

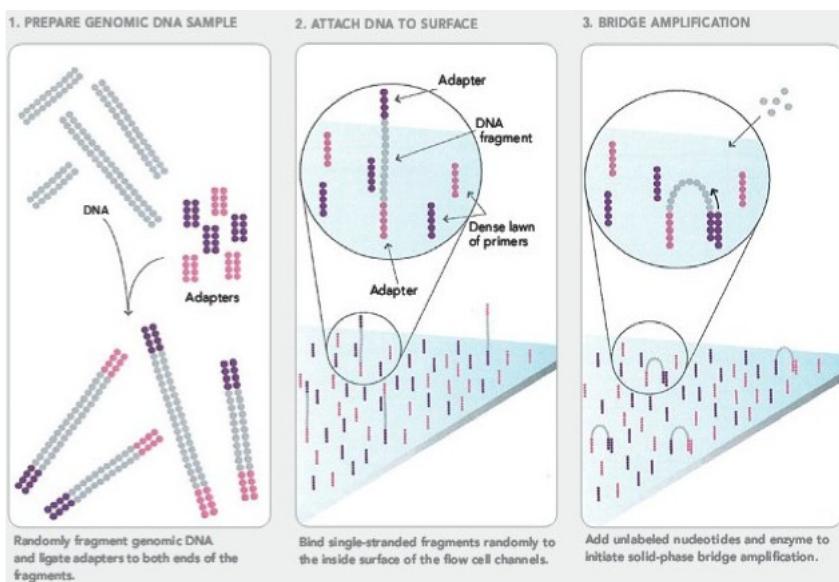


FIGURE 18 – Polony

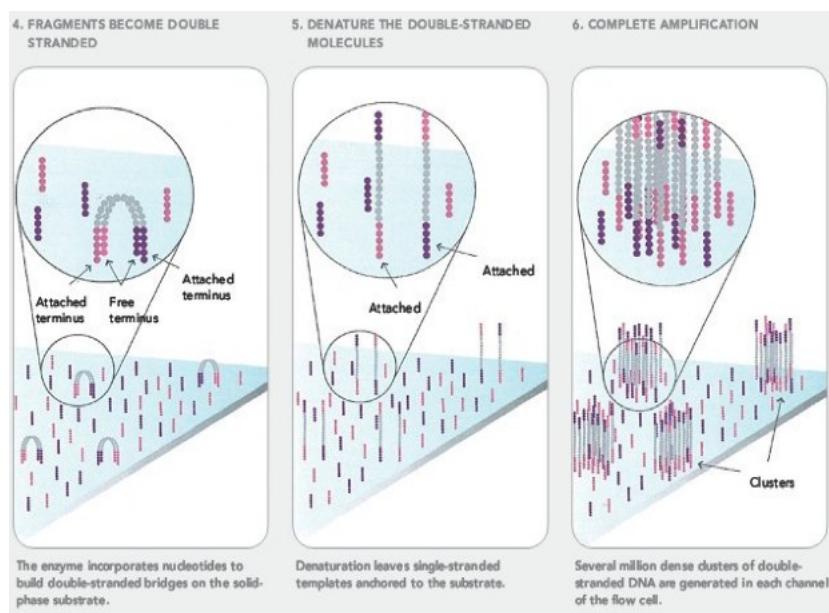


FIGURE 19 – Polony