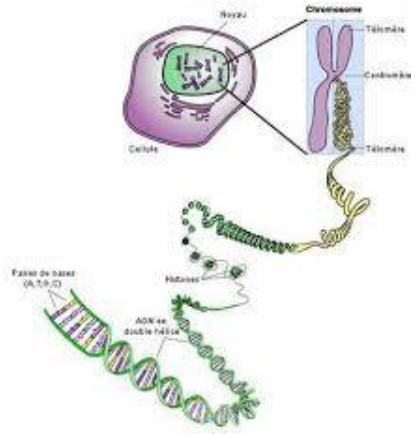
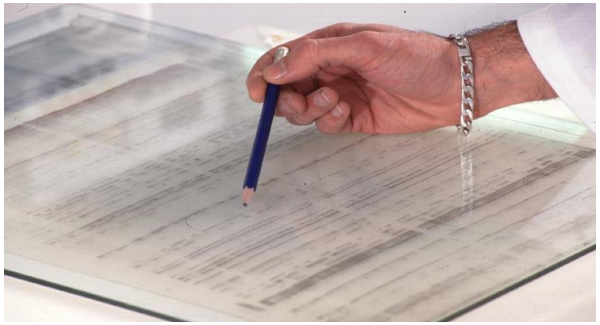
Les marqueurs moléculaires

Philippe Barre

INRA Centre Poitou-Charentes
UR4 : Unité de Recherche Pluridisciplinaire, Prairies et Plantes Fourragères
(URP^{3}F)





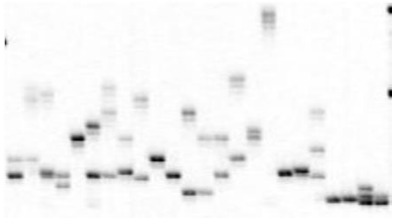
Poitiers M1 Biologie Ecologie 10 Février 2016

INRA
SCIENCE & IMPACT

Plan

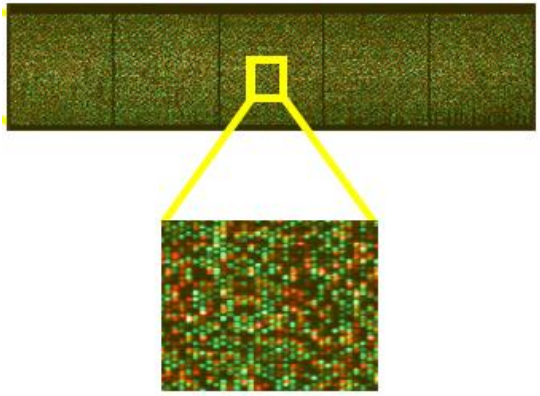
Å Introduction

Å Les marqueurs « bas- moyen débit »



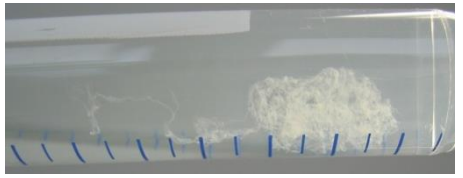
Å Les marqueurs « haut-très haut débit »

Å Conclusion



INTRODUCTION

LôADN ?

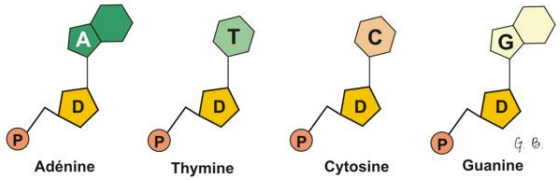


Acide
DésoxyriboNucléique

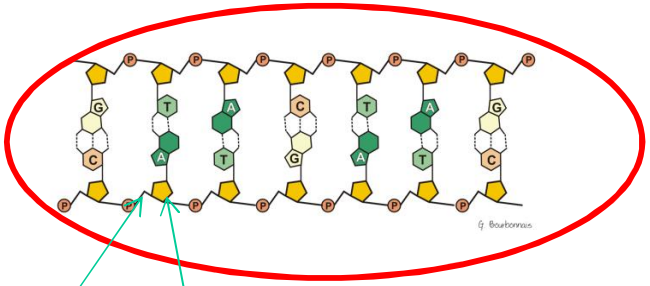
C’est le support de l’information génétique

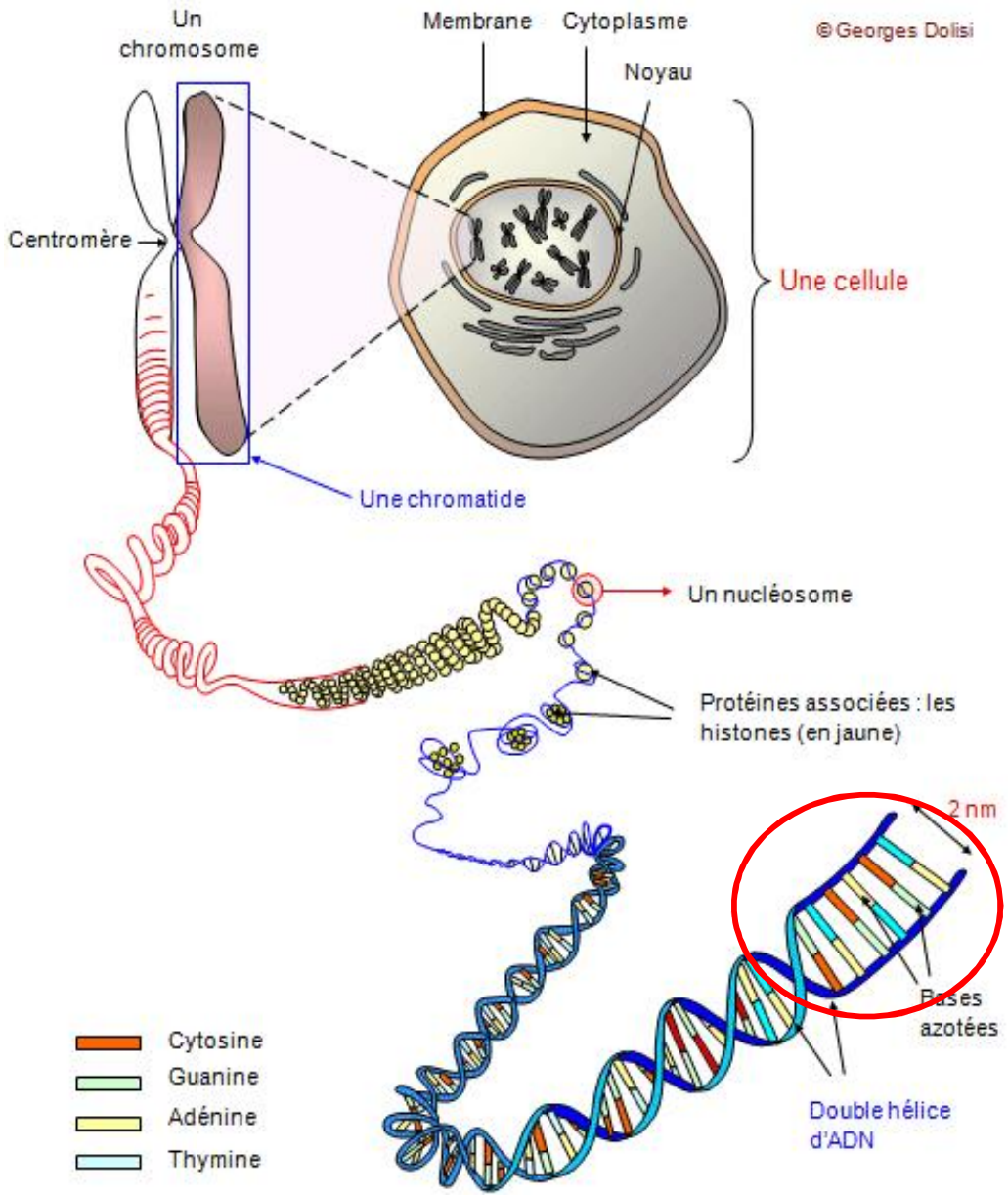
Chromosome (métaphase) : → Forme condensée de l ’ADN (= forme visible)

Éléments constitutifs : **4 nucléotides**



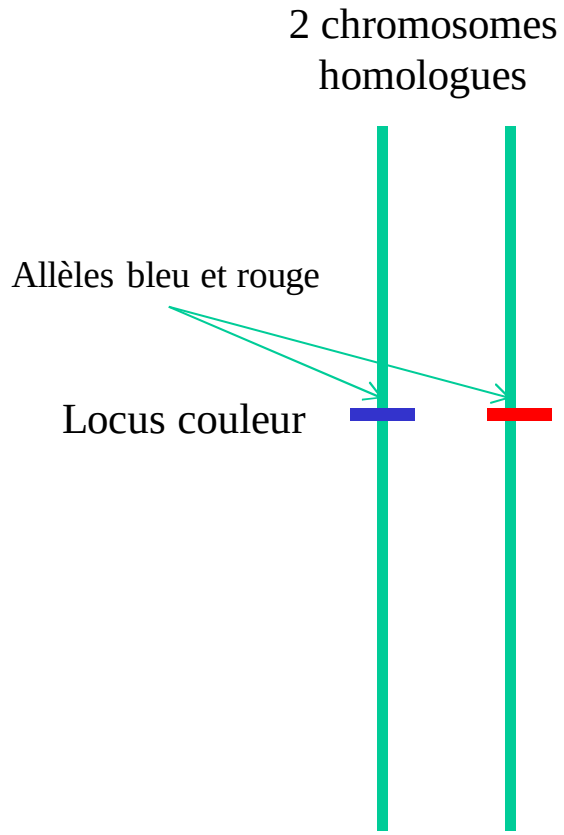
Organisation en double brin :





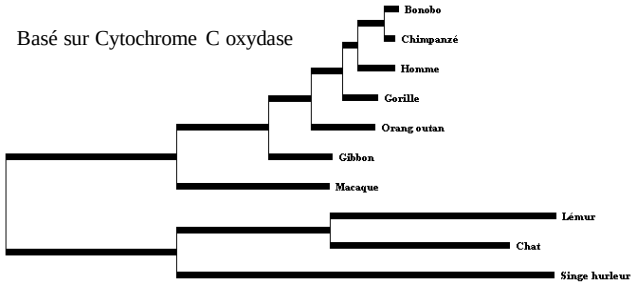
Définition des marqueurs moléculaires

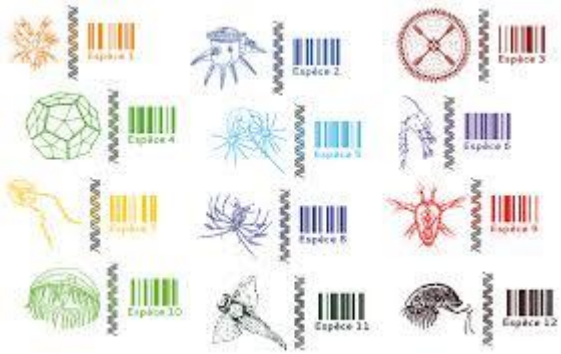
- Un endroit sur le génome : **locus**
- Présentant du polymorphisme entre et au sein des individus : **allèles**
- Les différents allèles sont identifiables en totalité (**co-dominant**) ou en partie (**dominant**) pour chaque individu

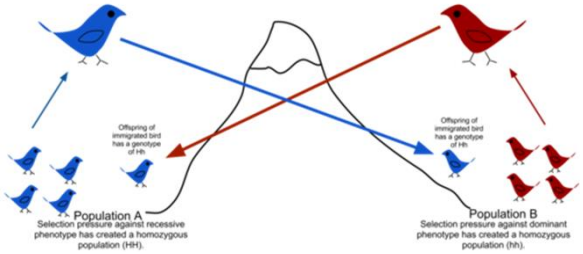


A quoi servent les marqueurs ?

- Etude de la diversité entre individus de la même espèce ou d'espèces différentes
  - Evolution des espèces
  - Identification des espèces (barcoding)
  - Identification des individus (criminalité)
  - Structuration de la diversité au sein des espèces (flux de gènes, sélection, dérive, mutation)

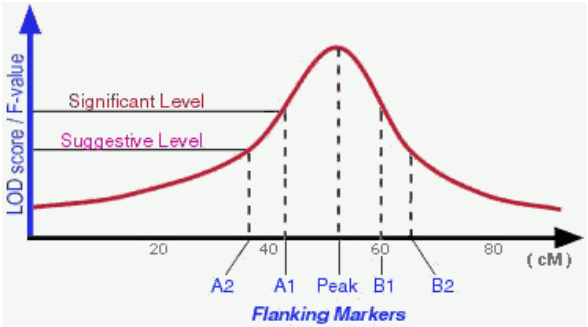


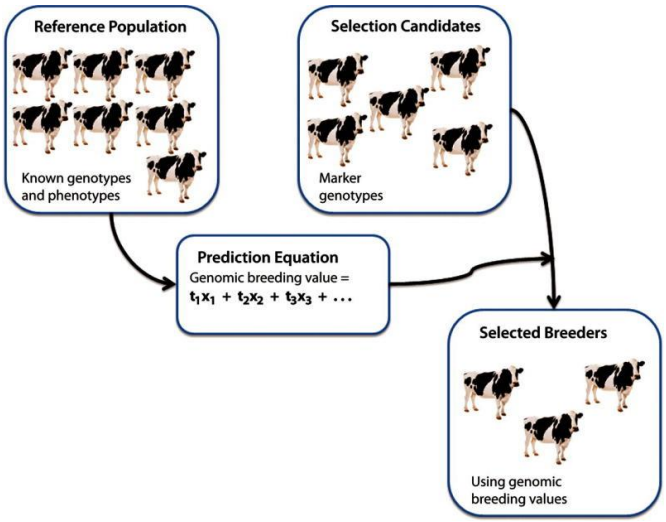




A quoi servent les marqueurs ?

- Relation entre polymorphisme moléculaire et polymorphisme phénotypique
  - Identification de régions du génome ou de gènes impliqués dans la variation d'un caractère quantitatif (QTL)
    - Construction d'individus sur la base des allèles aux QTL
  - Prédiction de la valeur génétique d'un individu à l'aide de son génome
    - Sélection génomique

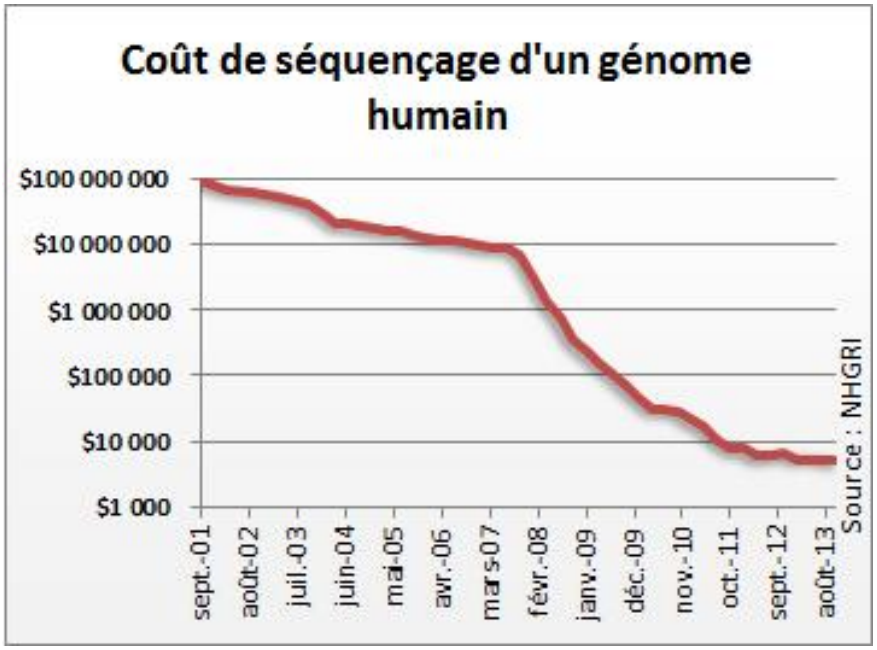


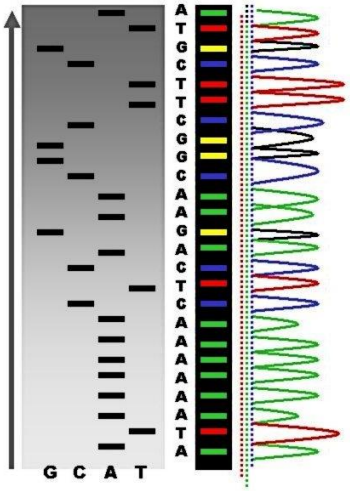


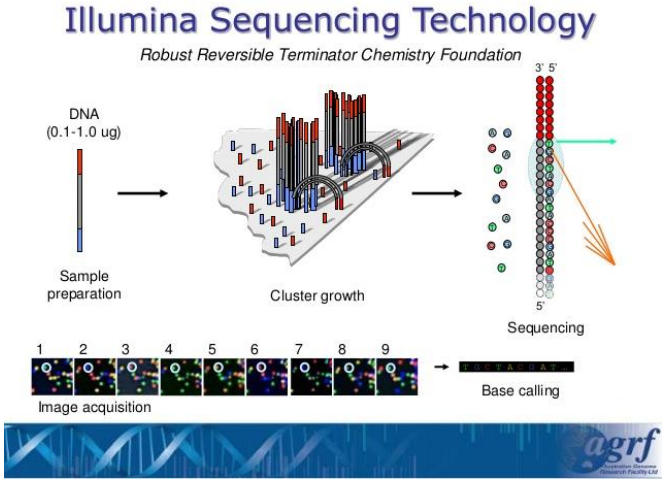
Une évolution fulgurante!!!

- Découverte de la double hélice de lôADN 1953 (James Watson and Francis Cricks)
- Techniques dôhybridation (années 1970)
- Séquençage de lôADN 1955 (Fred Sanger)
- Invention de la PCR 1993 (Kary Bank Mullis)

Une évolution fulgurante!!!







http://www.ncbi.nlm.nih.gov/genome/probe/doc/Technologies.shtml
http://www2.warwick.ac.uk/fac/sci/lifesci/research/vegin/geneticimprovement/geneticmarker/
http://get.genotoul.fr/
http://gentyane.clermont.inra.fr/

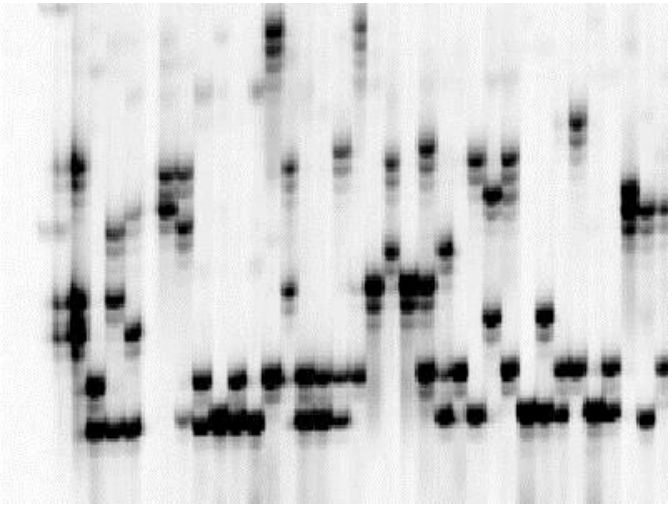
Une évolution fulgurante!!!

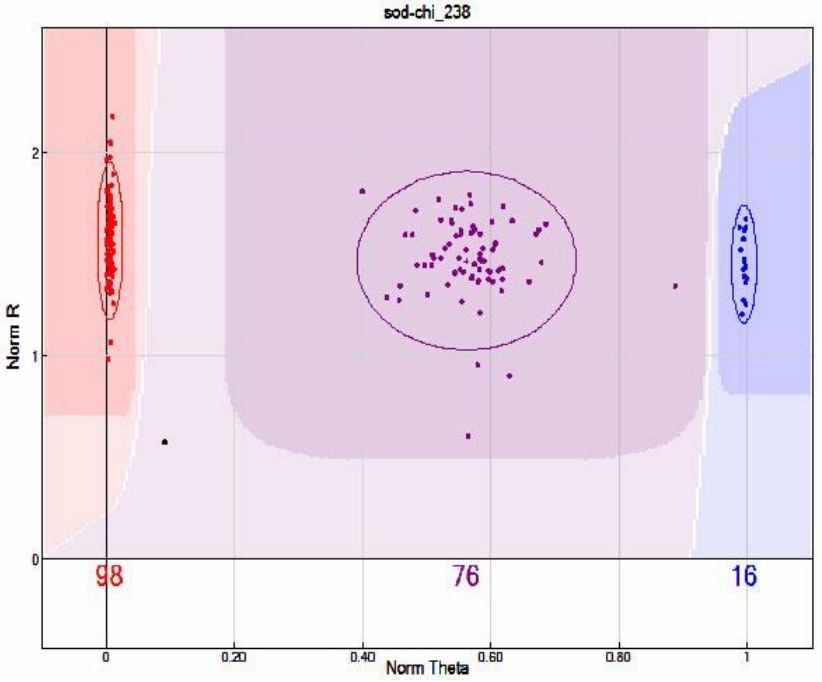
Coût du génotypage dépend beaucoup du nombre de marqueurs et du nombre dôindividus

Exemples:

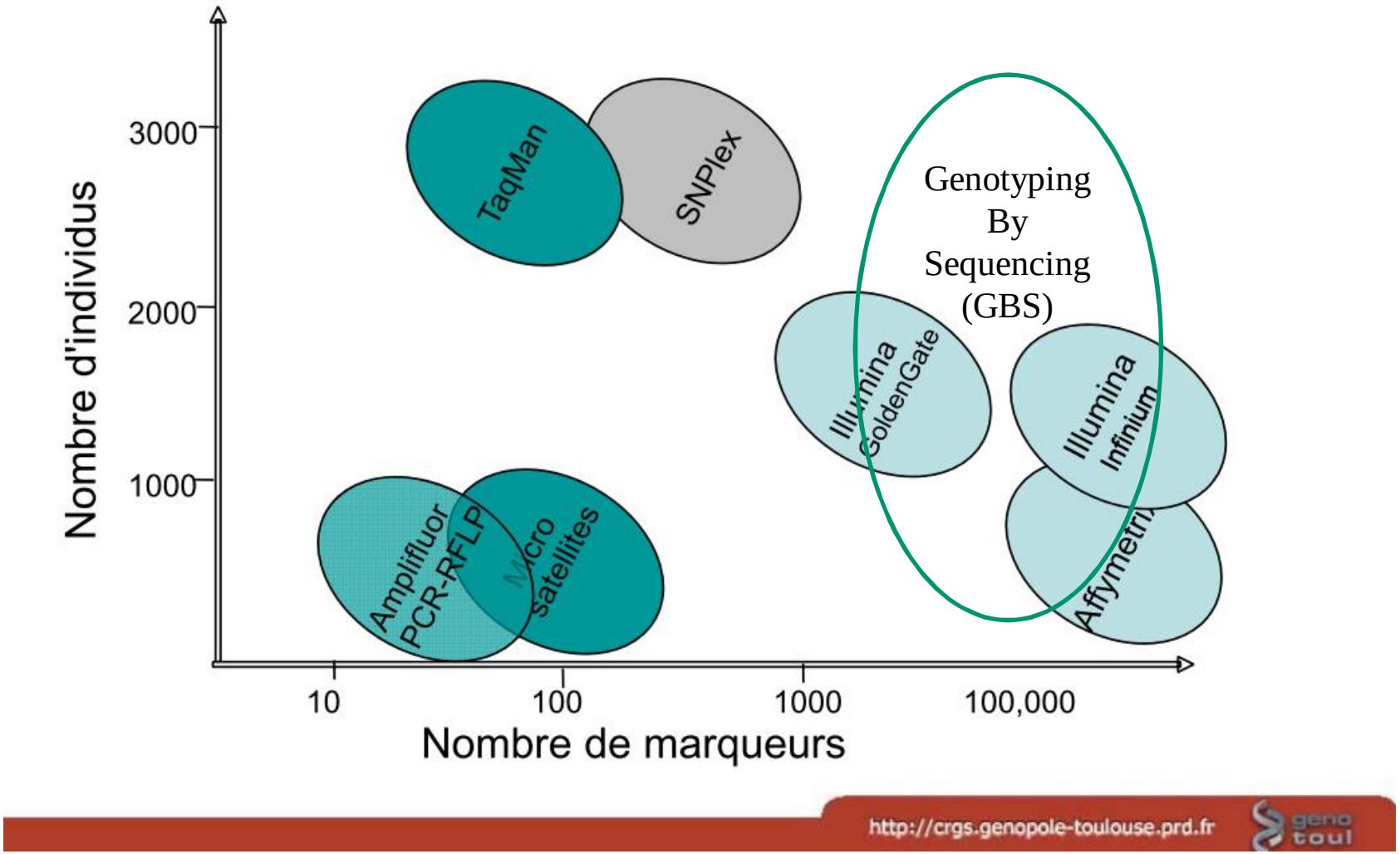
- 0,3 ú/point pour un marqueur sur un génotype
- 0,05ú/point pour 384 marqueurs sur 1536 individus
- 0,003ú/point pour 20000 marqueurs sur 500 individus

Moins le coût du point est cher plus il faut génotyper de marqueurs et dôindividus





De multiples méthodes de génotypage



A définir selon les besoins !!!

Quelles régions du génome?

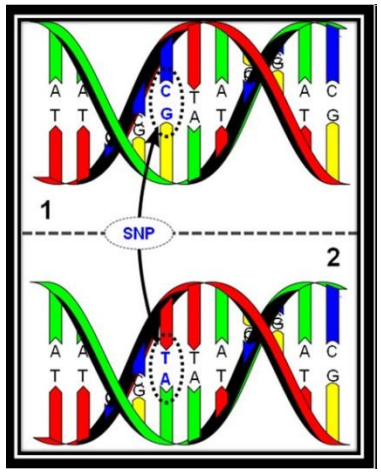
- Répartis sur lœnsemble du génome
- Ciblés sur des régions particulières:
  - Gènes candidats
  - Ensemble des gènes

- Connaissance préalable du génome
  - Séquençage complet / partiel du génome
  - Séquençage du transcriptome

LES MARQUEURS « BAS-MOYEN DÉBIT »

De multiples méthodes

- Polymorphisme de longueur
  - par hybridation: Restriction Fragments Length Polymorphism (pour mémoire)
  - par PCR: Sequence Tagged Site (STS) including Simple Sequence Repeat (SSR), Cleaved Amplified Polymorphic Sequences (CAP), Amplified Fragment Length Polymorphism (AFLP).é
- Polymorphisme de séquence : Single Nucleotide Polymorphism (SNP)
  - Taqman
  - Kompetitive Allelic Specific PCR (KASP)é

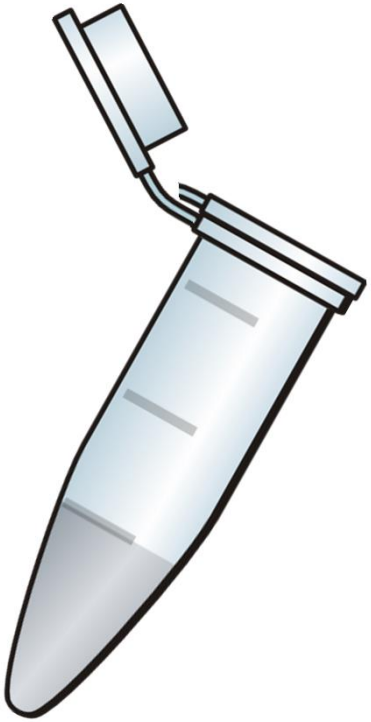


La PCR

Multiplier un fragment d ’ADN un grand nombre de fois afin de le repérer.

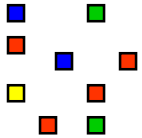
Technique de la PCR : **P**olymerase **C**hain **R**eaction

Réactifs nécessaires à la PCR



tampon PCR 10X

$MgCl_2$



dNTPs

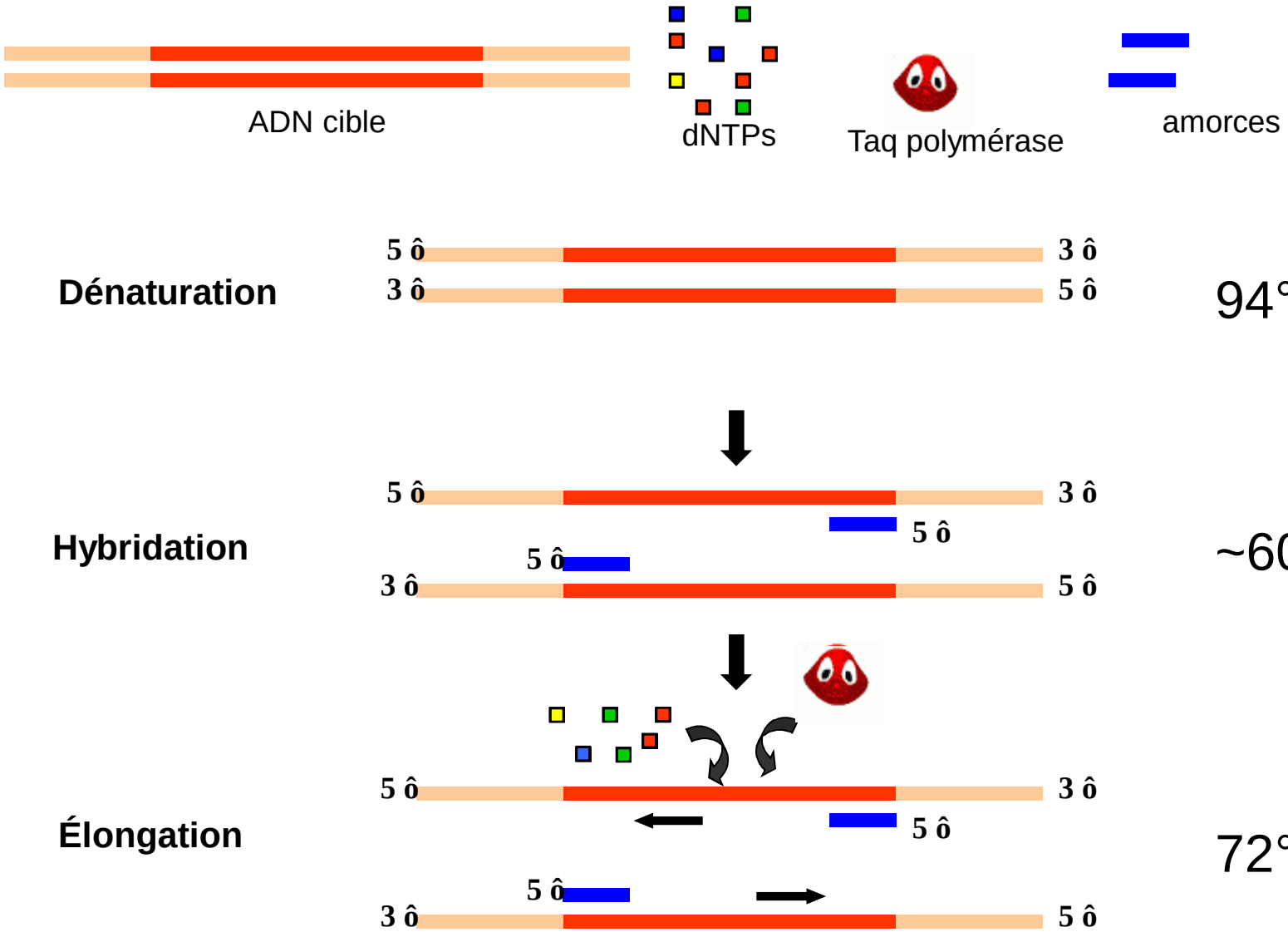
amorces

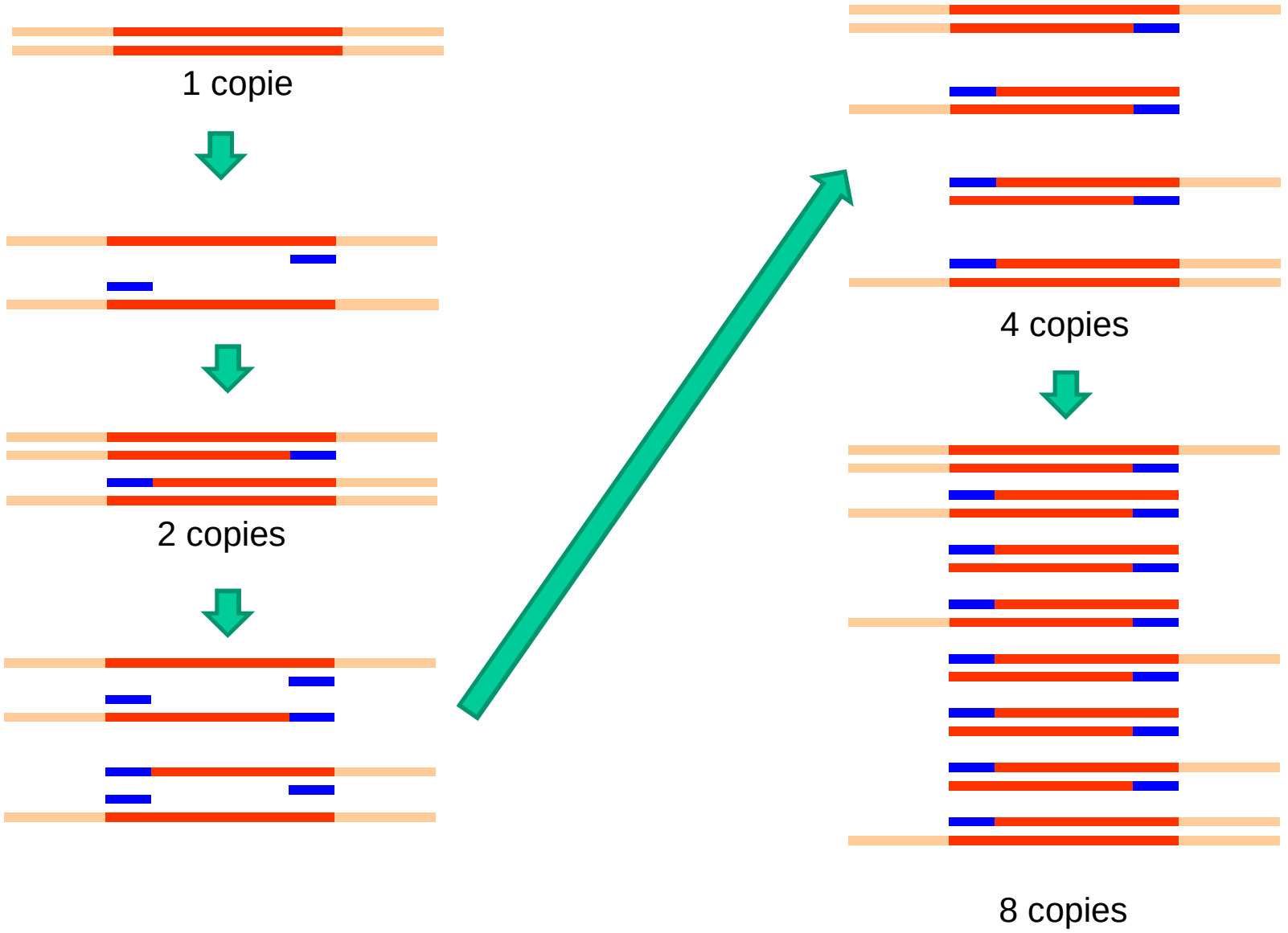
Taq polymérase

ADN

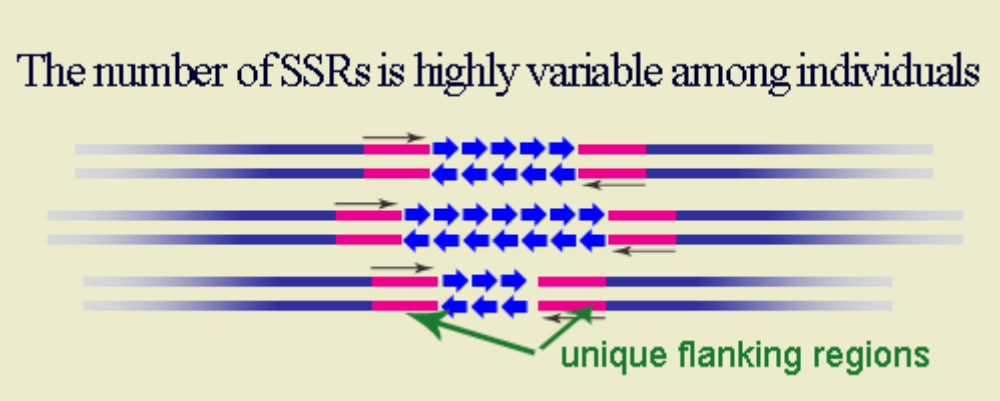
eau

Thermocycleur


ADN cible
dNTPs
Taq polymérase
amorces
Dénaturation
5 ô
3 ô
3 ô
5 ô
94°C
Hybridation
5 ô
3 ô
5 ô
5 ô
3 ô
5 ô
~60°C
Élongation
5 ô
3 ô
5 ô
5 ô
3 ô
5 ô
72°C

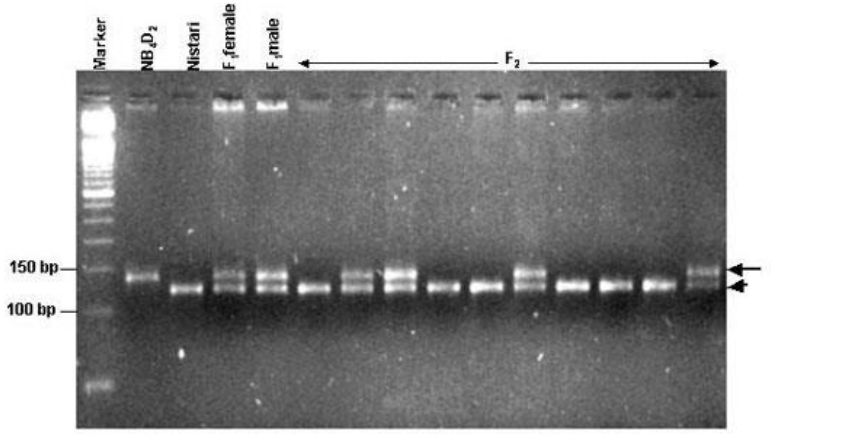

1 copie
2 copies
4 copies
8 copies

Exemple 1: SSR, microsatellite



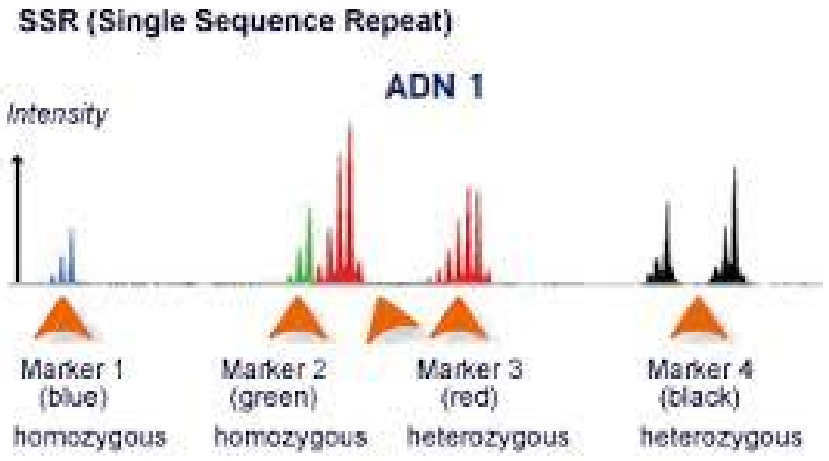
Nécessité de connaitre la séquence pour définir des amorces :

- Banques enrichies en SSR
- Résidus du séquençage de transcriptomes



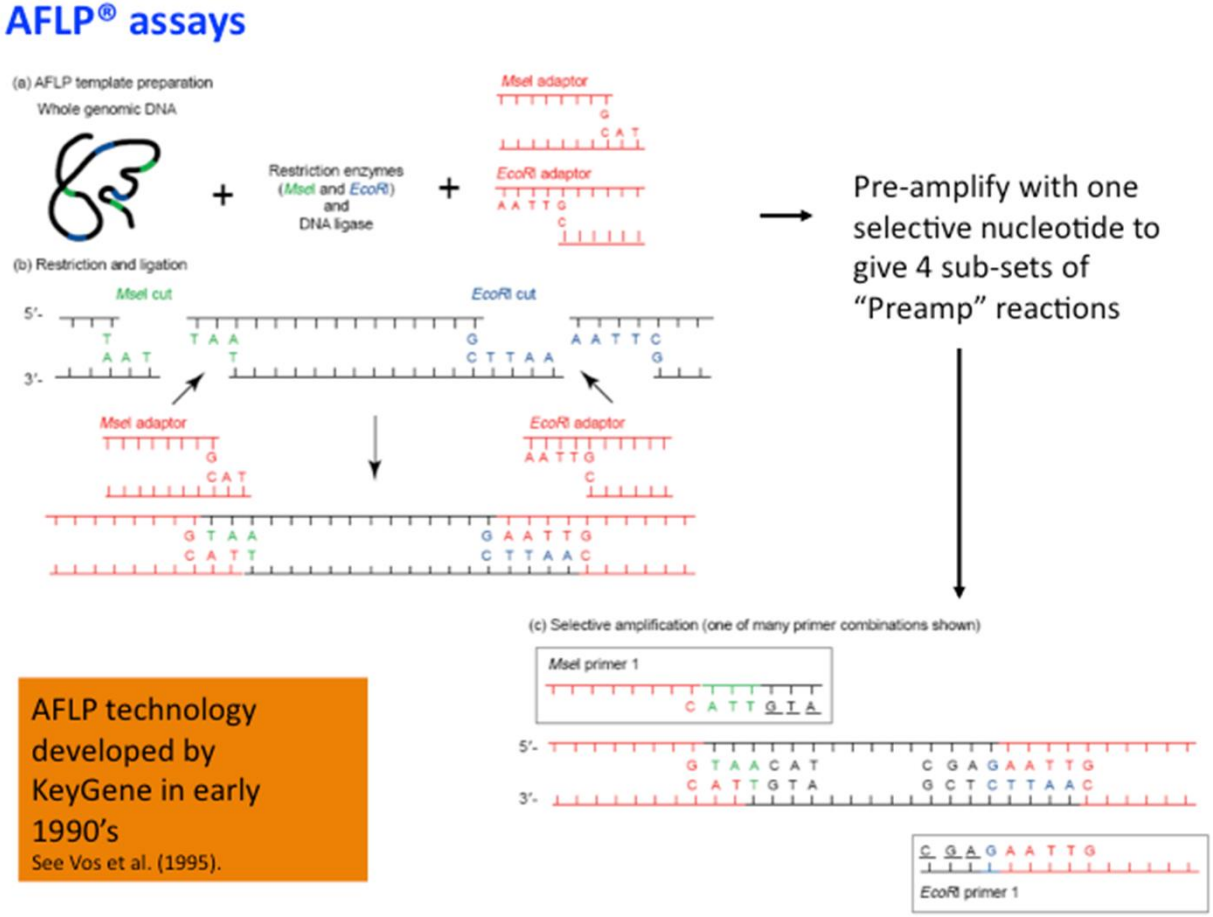
Inheritance and segregation of SSR loci in the silkworm .
SSR profiles obtained using Bmsat070 on two parental strains,Nistari and NB_4D_2 and their $F_1$and F_2 offspring (10Nos). The segregation of NB_4D_2 (Female) and Nistari (Male) specific polymorphic loci is shown by arrow and arrow heads respectively.

Electrophorèse sur gel dôagarose

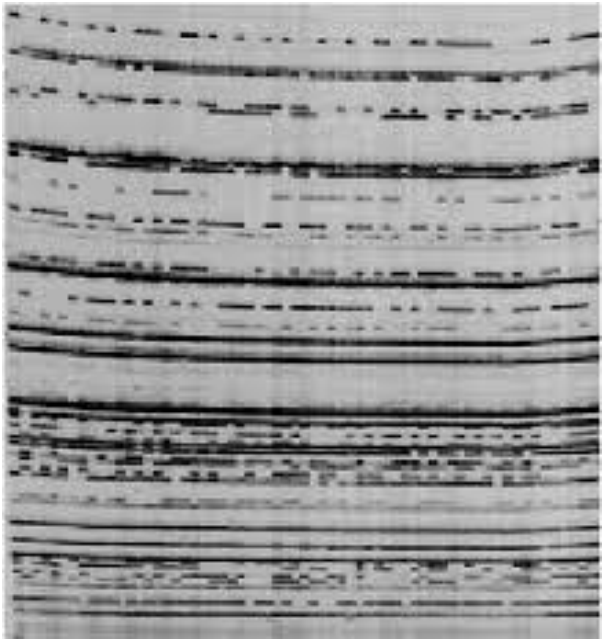


Electrophorèse sur séquenceur capillaire

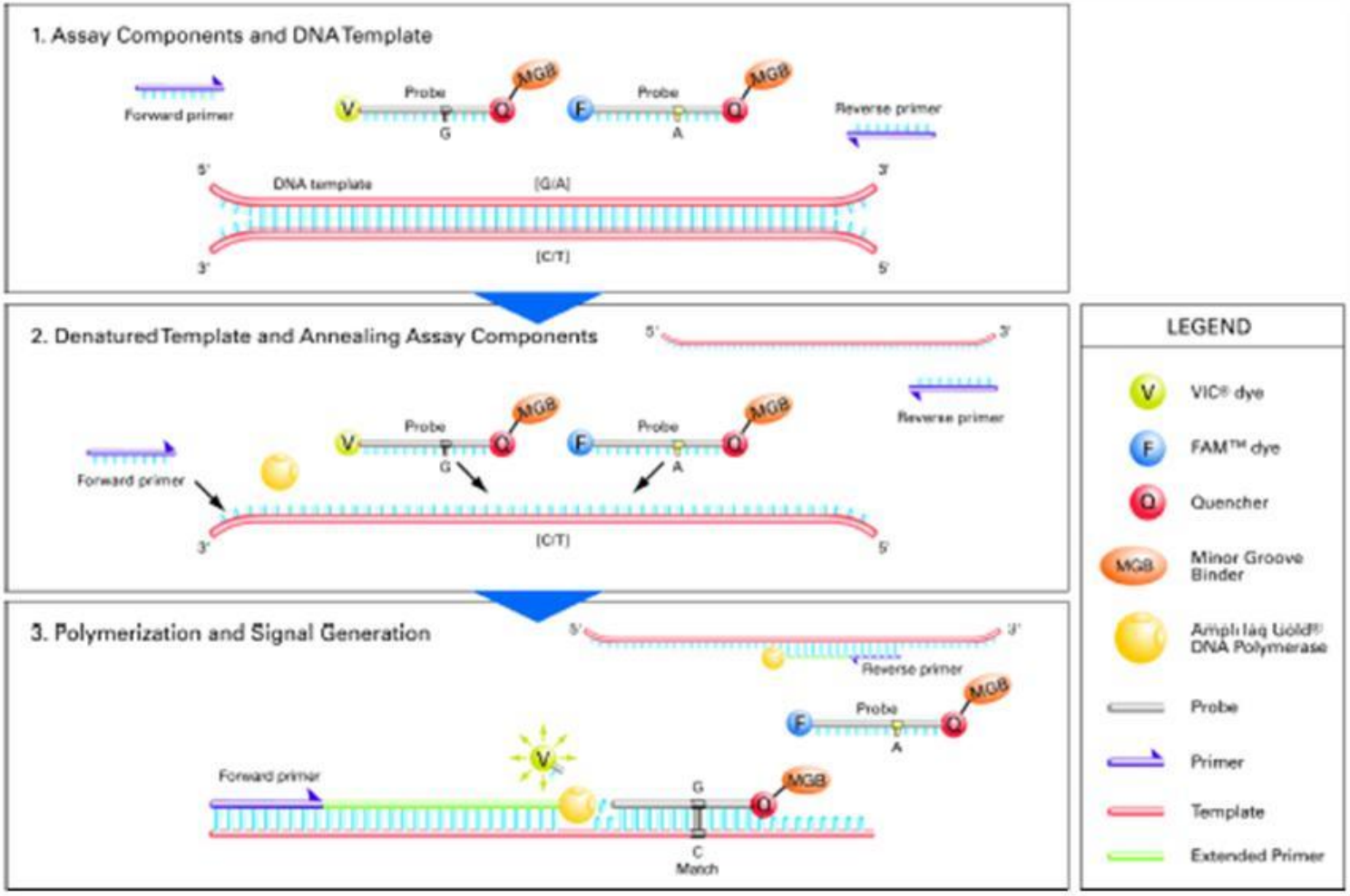
Exemple 2: Ampified Fragment Length Polymorphism (AFLP)



AFLP technology developed by KeyGene in early 1990's
See Vos et al. (1995).

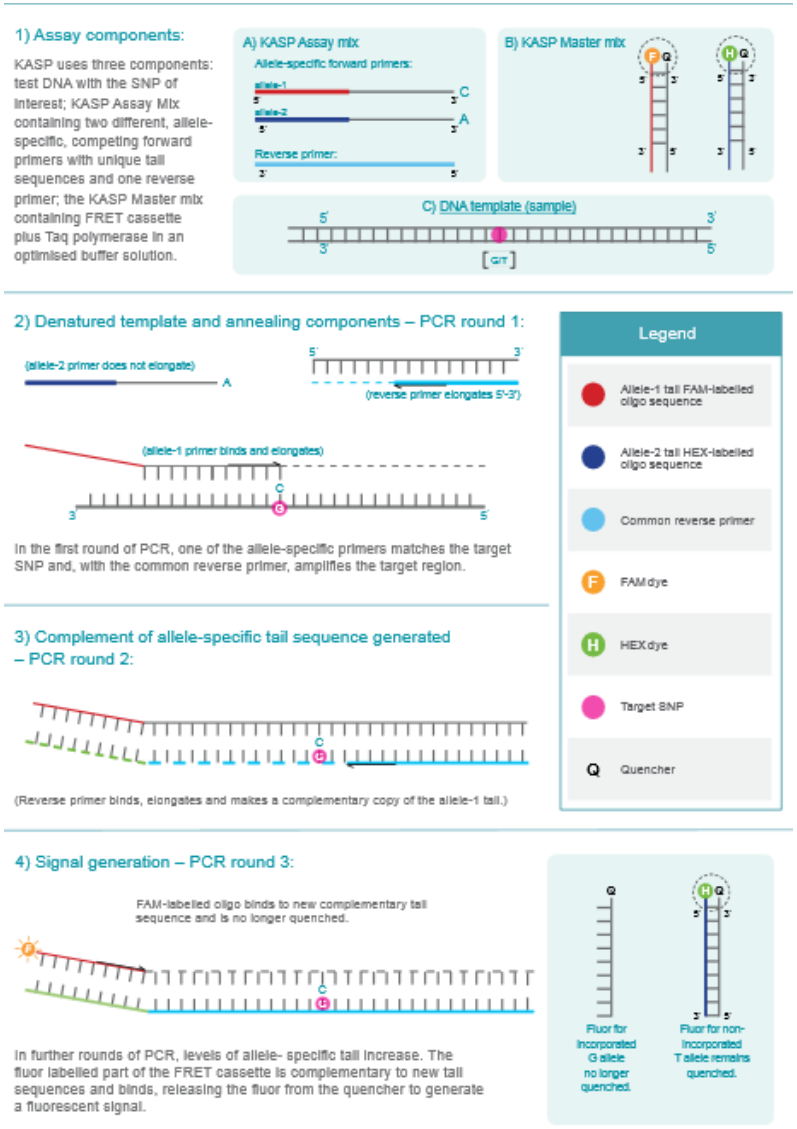


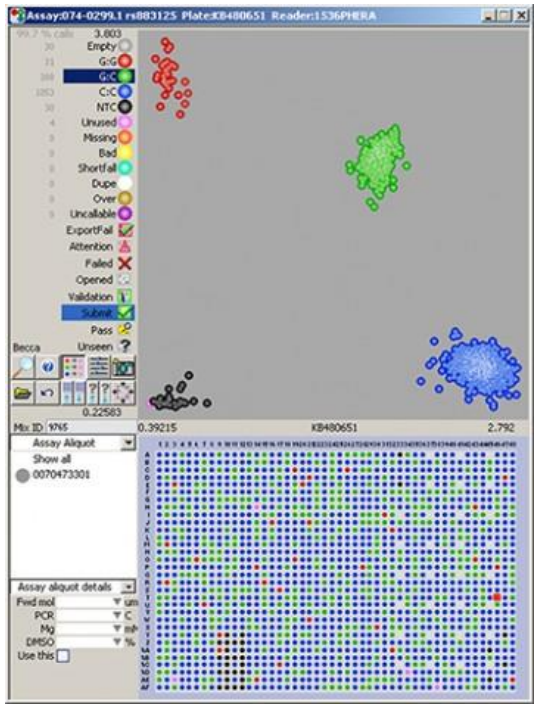
Exemple 3: Système Taqman



Cher à développer mais une fois développer très peu cher mais pas « multiplexable » !

Exemple 4: Système KASP KBioscience

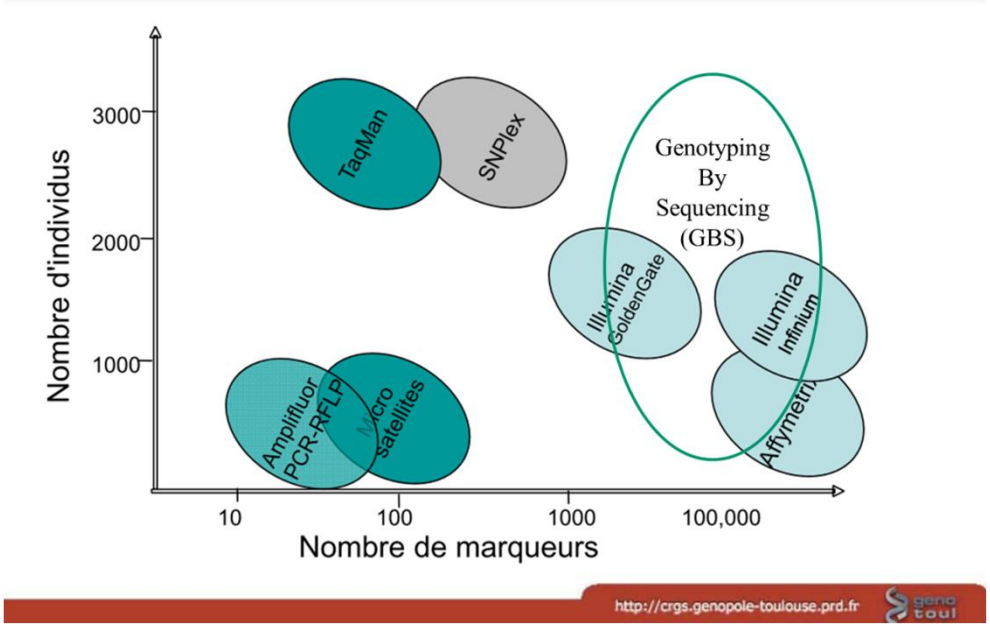




- Utilisation de la PCR + lecteur fluorescence
- Kit pour 96, 384, 1536 individus

LES MARQUEURS « HAUT-TRES HAUT DÉBIT »

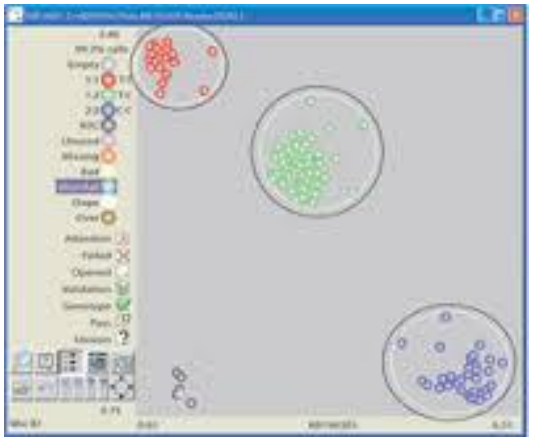
Haut ï Très haut débit ?



- A la fois beaucoup de marqueurs (>1000) et beaucoup dôindividus (>1000)
- Nécessité de robot de pipetage
- Nécessité de mémoire informatique et puissance de calcul

Deux grands types

- Génotypage par codage
  - Illumina golden gate
  - Illumina Infinium
  - Affymetrixé
- Génotypage par séquençage : GBS
  - Différentes régions du génome séquencées
  - De multiples techniques de séquençage



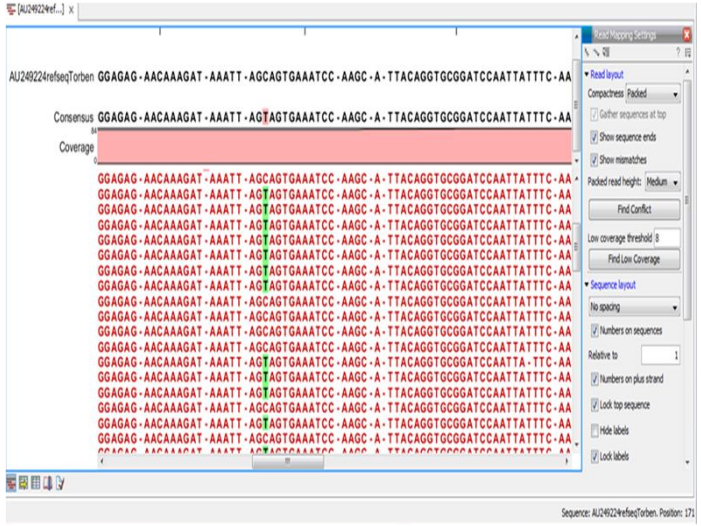
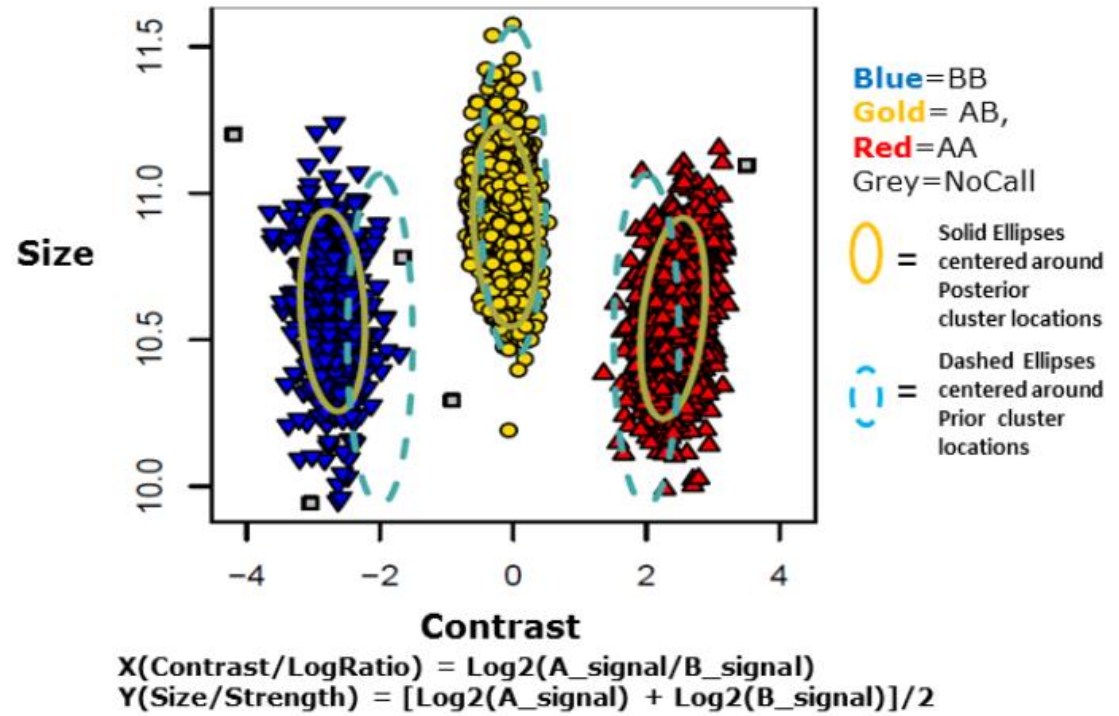


Figure 2.2 SNP Cluster Plot produced by the SNPolisher package, via the *Ps_Visualization* function.



GÉNOTYPAGE PAR CODAGE

Exemple 1: KASP

	LC480®	Biomark®	Tecan Infinite
Support	Plaque 384	Puces : - 48*48 - 96*96	Plaque 96, 394 ou 1536
Type de projet	Quelques SNPs	Au moins 48 SNPs	De 1 à 1000 SNPs
Volume réactionnel	µL	nL	µL
Débit	20 plaques/jour : 7680 données	2 à 3 puces/jour : 18240 données (puces 96*96)	60 plaques/jour : 23040 données
Coût / data	0,16 € HT	0,1 € HT	0,08 € HT

Moyen débit

Haut ï très haut débit

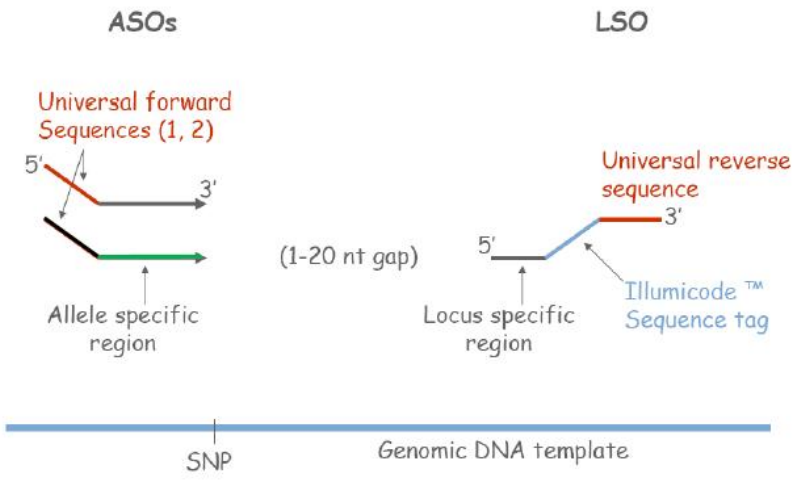
Exemple 2: Illumina Golden Gate Veracode (BeadXpress) http://uthscsa.edu/csb/CSBPDFfiles/genomics-GoldenGateGenotypingOverview.pdf

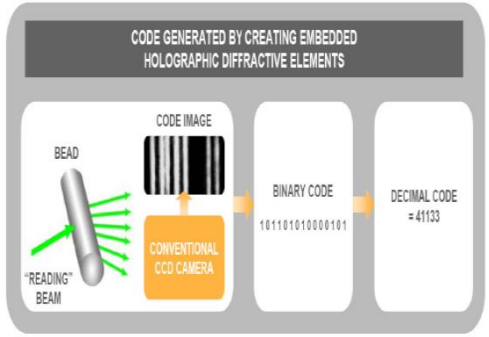
- Plusieurs degrés de multiplexage disponibles :
  - 48plex
  - 96plex
  - 144plex
  - 192plex
  - 384plex
- Nombre minimum dôindividus génotypés : 480 (5*96) puis par multiple de 96
- Coût à la data : de 0,045ú HT à 0,22ú HT selon configuration (nbre ech/nbre de SNP)

Plateforme GENTYANE INRA Clermont-Ferrand (C. Poncet)

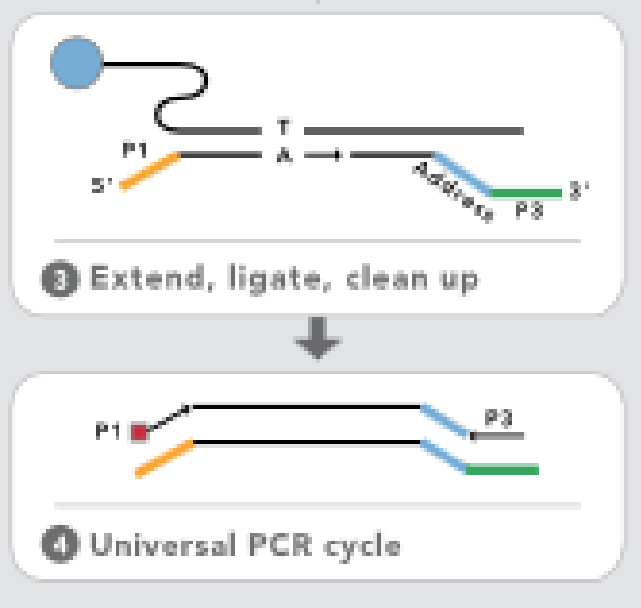
Choix des SNP

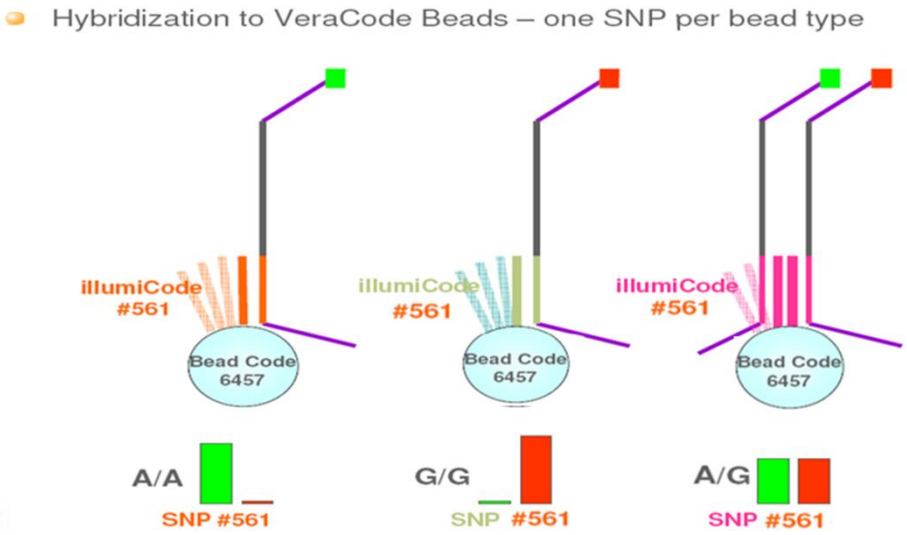
- Soumission à Illumina dôun fichier en ligne
- Besoin de 50bp autour du SNP
- Obtention dôun score de qualité par SNP (Illumina, résultat instantané)
- Choix définitif des SNP intégrant score et position
- Commande de lôOPA et design des amorces par ILLUMINA (2 mois)
- Réception OPA et génotypage (1 semaine pour 480 individus)

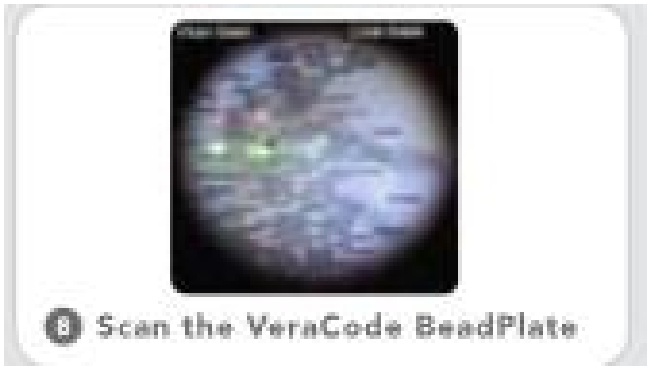


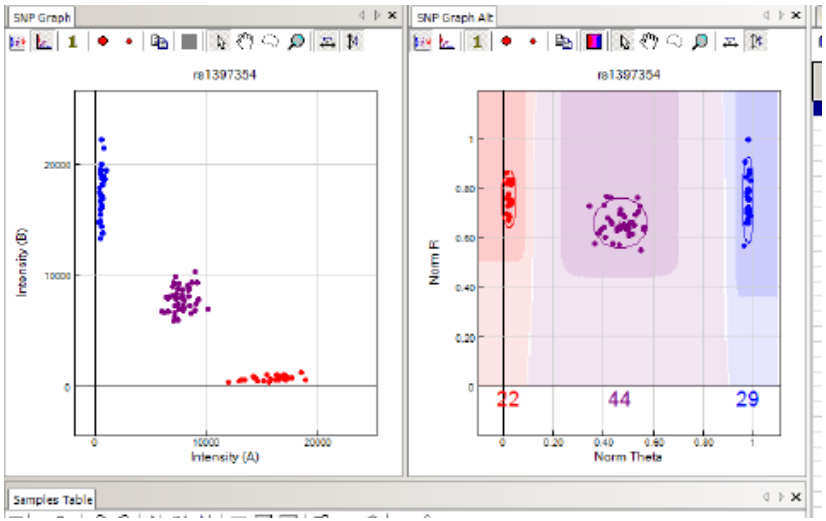


Présentation de la Chimie GoldenGate









Exemple 3:Axiom Affymetrix Genetitan

Solution automatisée de génotypage SNP par hybridation

Genetitan Affymetrix

Hybridation / lavage / lecture des plaques (PEG) 96 puces

Biomeck FxP Beckman

Préparation des échantillons avant passage sur GeneTitan

Permet un débit de 8 plaques (PEG)/semaine

http://media.affymetrix.com/support/downloads/manuals/axiom_genotyping_solution_analysis_guide.pdf

Plateforme GENTYANE INRA Clermont-Ferrand (C. Poncet)

Axiom™ Genotyping Solution

DNA Amplification | Fragmentation | Precipitation | Resusp. & Hyb. Prep | Hyb. | Wash, Ligation, Stain & Imaging

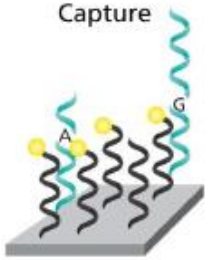
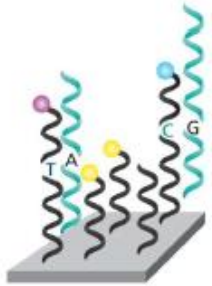
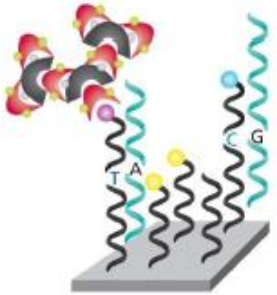
Mon | Tue | Wed | Thu

Throughput	• Highest throughput/operator &Fewest interventions • **768 samples per week**
Hands-on-time	• Lowest in industry • **Less than 2.5 hr/plate**
Ease of use	• Color coded reagent tubes, templates • **Reagents, buffers, GeneTitan consumables**
Support	• Partnership with Beckman • **Methods supported by Affymetrix FAS**

Affymetrix

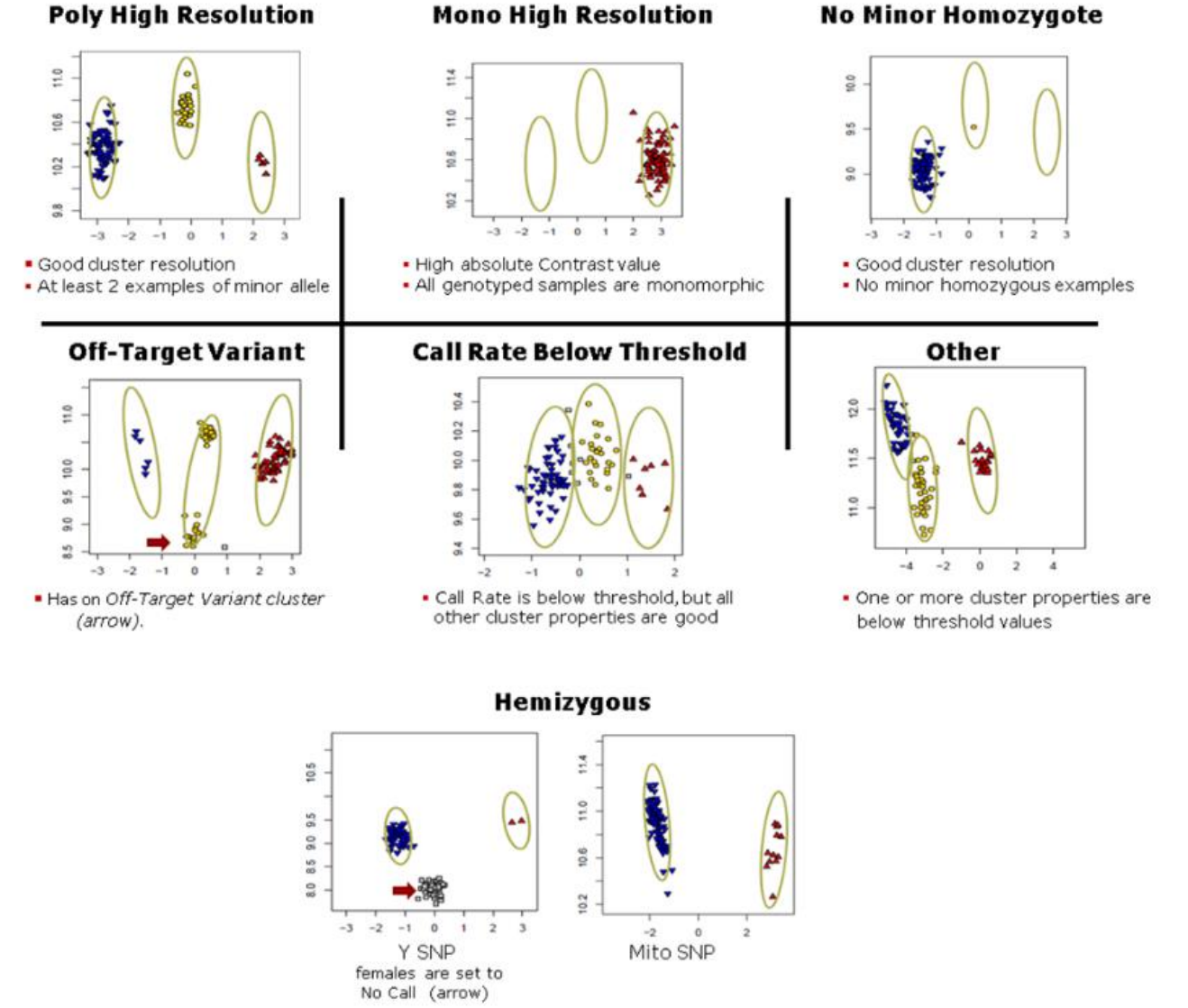
Schematic of the Axiom Assay




Target prep	Hybridization	Ligation	Signal amplification
Amplify	Capture	Differentiate	Stain and image
			
Fragment	+		
	Label		
			
	Labeled solution probe		

Le codage en génotype

Figure 3.3 Cluster Plot examples and descriptions of the seven SNP classification categories. OTV SNPs are discussed further in *Adjust Genotype Calls for OTV SNPs* on page 30.

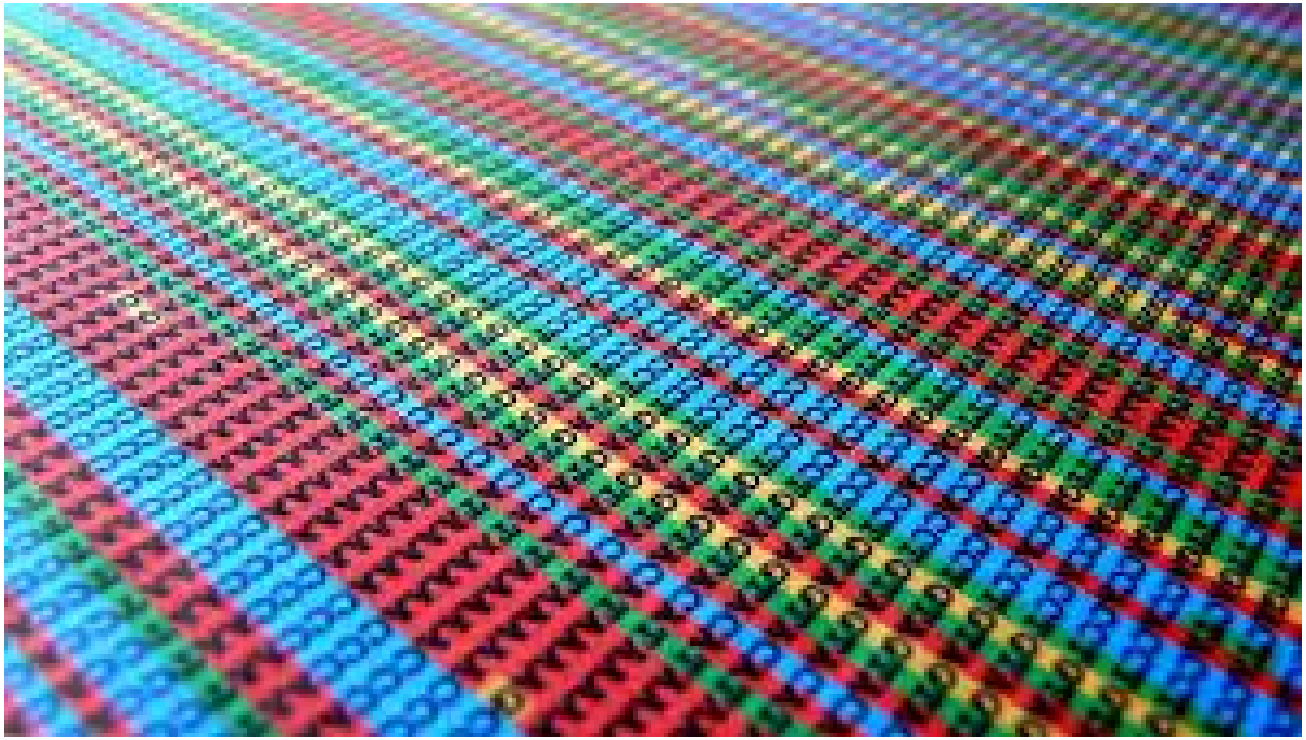


Avantages et Inconvénients de cette technologie

+ Débit très important
+ Besoin humain limité / quantité de données produites
+ Analyse rapide sous R
+ Coût à la data très faible de 0,08€ HT à 0,00006€ HT selon config
+ Souplesse nombre de SNP de 1500 à 675000 par puce

- **Format 96 individus imposé**
- **Minimum de commande pour 480 individus**
- **Ticket d'entrée minimum de 60k€ HT**

Technologie proposée sur la plateforme Gentyane depuis Septembre 2013



GÉNOTYPAGE PAR SÉQUENÇAGE

Génotypage par séquençage

- Séquencer tout le génome : oui mais cher !!
- Séquencer une partie du génome :
  - Les gènes : transcriptome
  - Des régions réparties sur le génome mais non « choisies » : GBS avec enzyme de restriction
  - Des régions choisies : capture

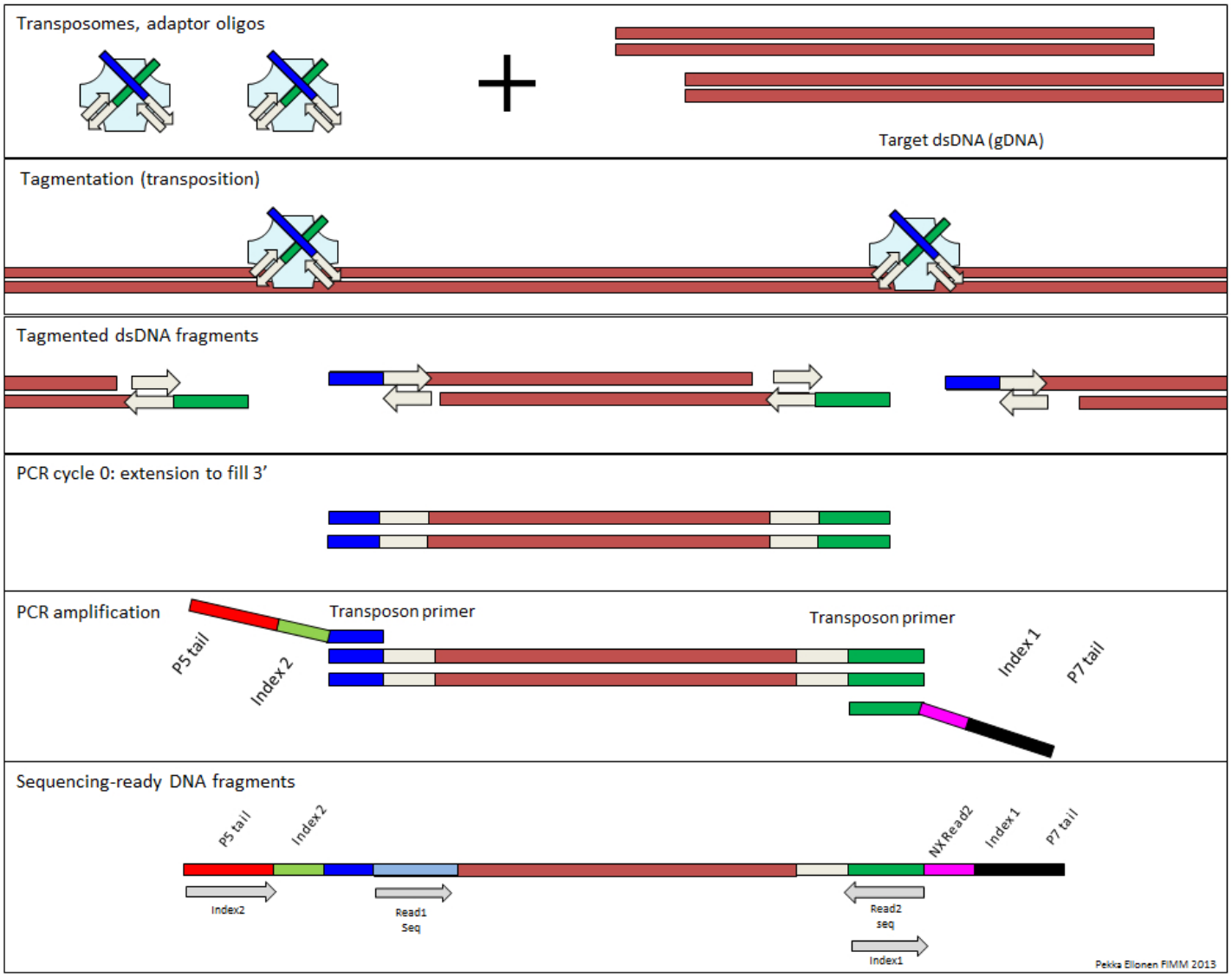
Génotypage par séquençage

- Trois étapes:
  - Fabrication de la « banque » : mélange de fragments dôADN : peut être réalisée dans un labo de biologie moléculaire classique + compétence BM
  - Séquençage : nécessité des machines très chères et qui évoluent très vite
  - Analyse bio-informatique : nécessité des ordinateurs très puissants + compétence informatique

Les banques

- Banques dôADNc : attention aux tissus choisis selon la question et si besoin standardisation
- Banque dôADN génomique :
  - ADN total fragmenté de manière aléatoire
  - ADN issu de lôamplification dôune partie du génome par PCR: séquences cibles (ex. barcoding) ou enzymes de restriction
  - ADN issu de lôhybridation sur des séquences connues : capture

Les banques

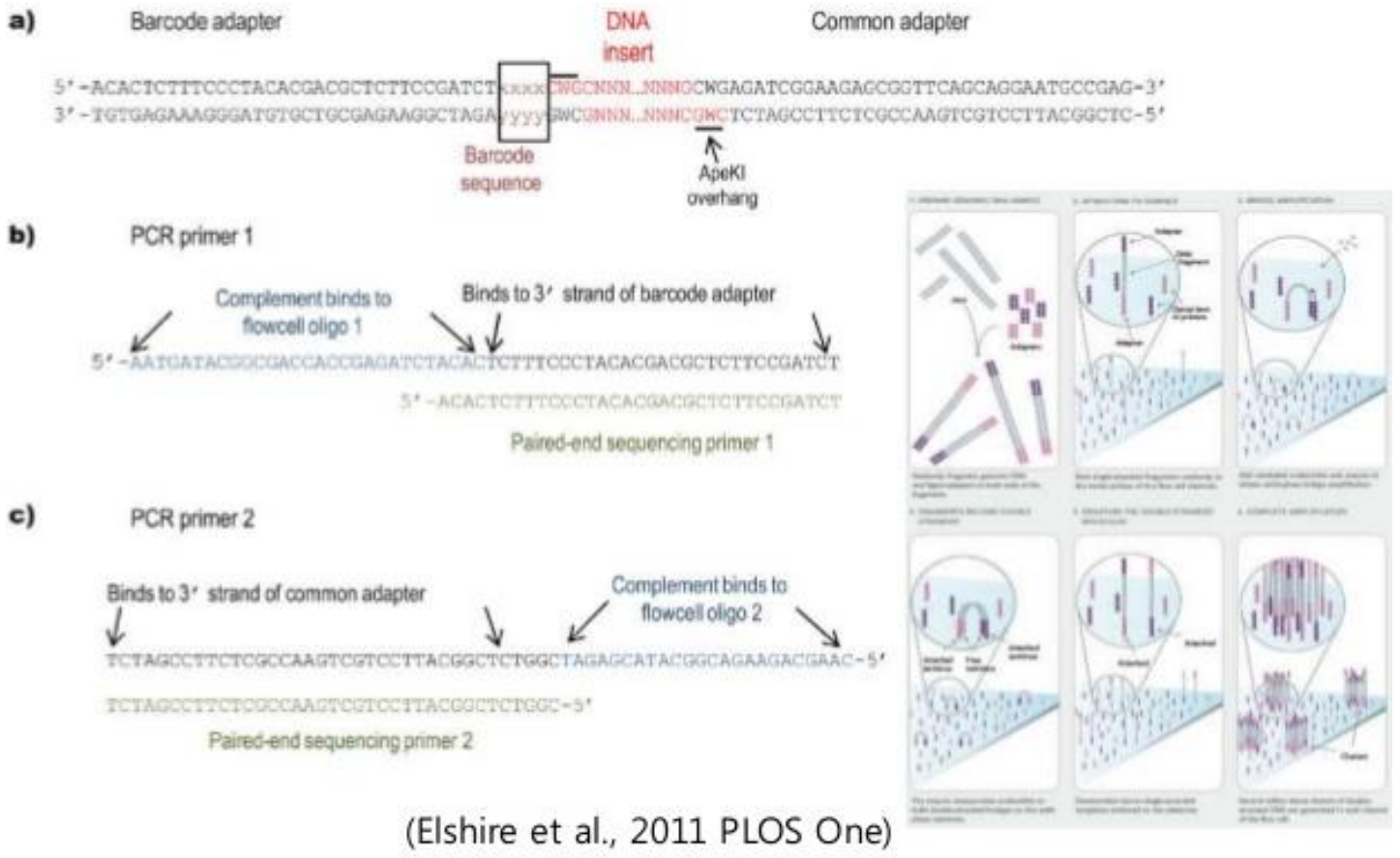


Système NEXTERA de Illumina

https://www.fimm.fi/en/services/technology-centre/sequencing/next-generation-sequencing/dna-library-preparation

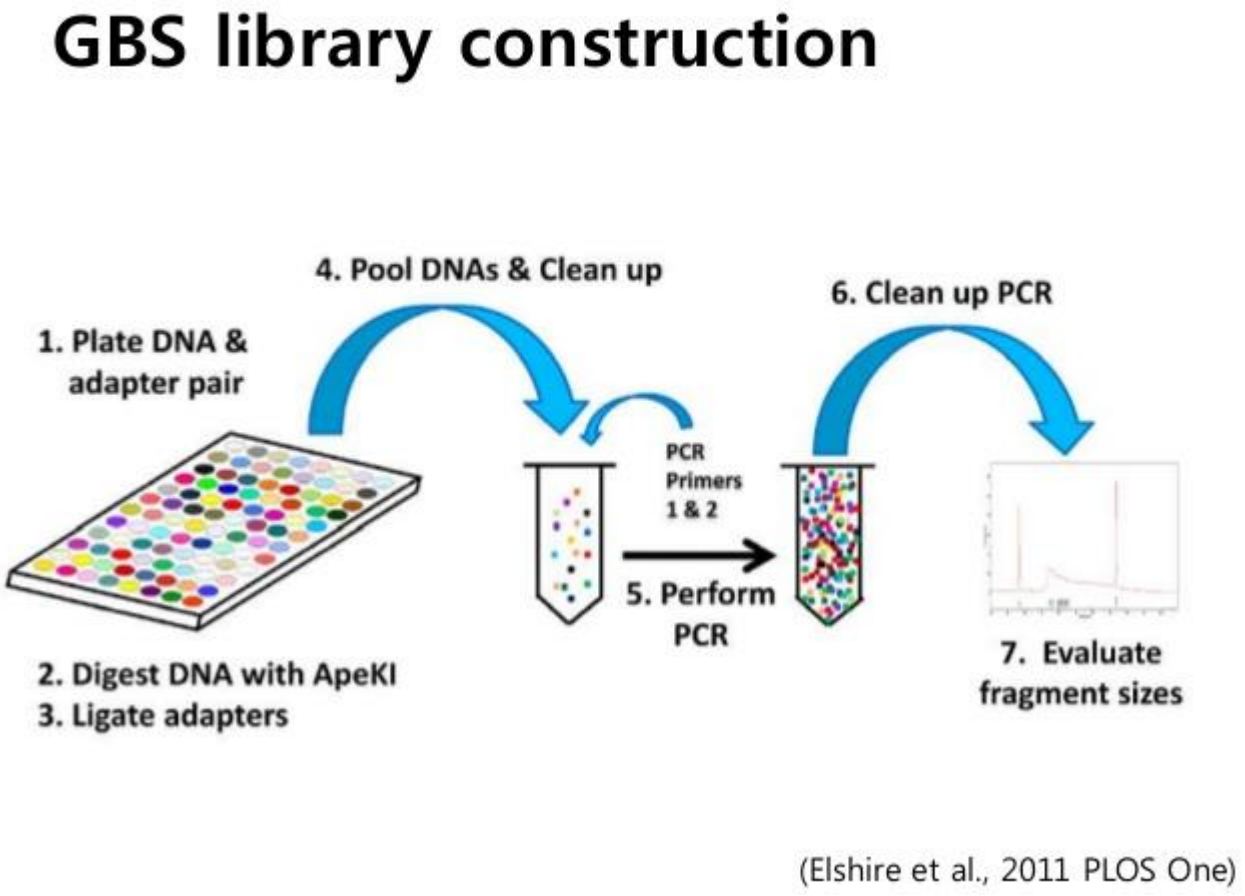
Exemple : GBS Elshire et al.2011 PLOS One

Digestion de lôADN génomique par une enzyme de restriction (ici ApeKI)



(Elshire et al., 2011 PLOS One)

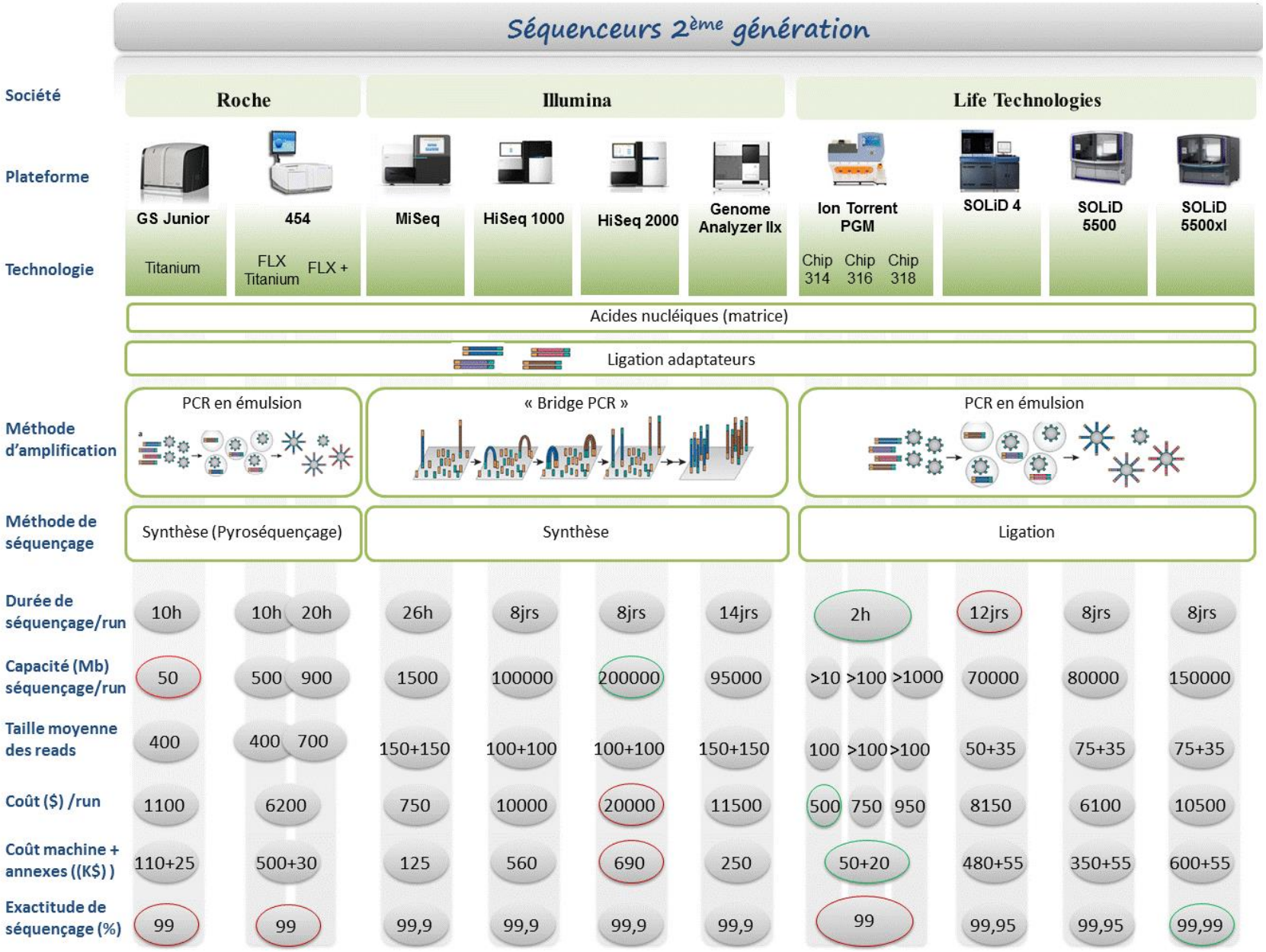
Exemple : GBS Elshire et al.2011 PLOS One



(Elshire et al., 2011 PLOS One)

Il existe beaucoup de variantes RADseqé .

Séquençage: plusieurs plateformes selon les besoins



Séquenceur Illumina : évolution rapide

	 MiniSeq System	 MiSeq Series	 NextSeq Series	 HiSeq Series	 HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million†	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days
Benchtop Sequencer	Yes	Yes	Yes	No	No
System Versions	• MiniSeq System for low-throughput targeted DNA and RNA sequencing	• MiSeq System for targeted and small genome sequencing • MiSeq FGx System for forensic genomics • MiSeqDx System for molecular diagnostics	• NextSeq 500 System for everyday genomics • NextSeq 550 System for both sequencing and cytogenomic arrays	• HiSeq 3000/HiSeq 4000 Systems for production-scale genomics • HiSeq 2500 Systems for large-scale genomics	• HiSeq X Five System for production-scale whole-genome sequencing • HiSeq X Ten System for population-scale whole-genome sequencing

Exemple séquençage Hiseq

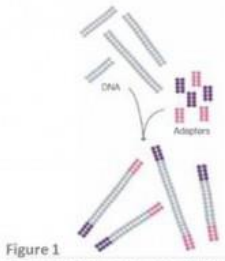


Figure 1
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

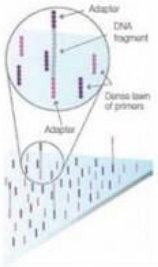


Figure 2
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

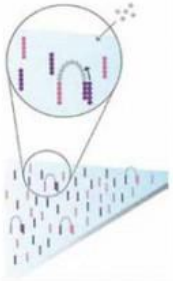


Figure 3
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

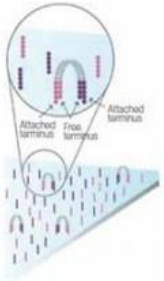


Figure 4
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

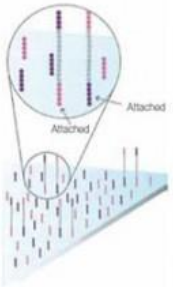


Figure 5
Denaturation leaves single-stranded templates anchored to the substrate.

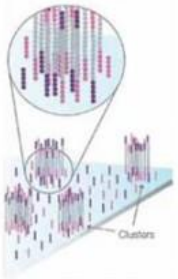


Figure 6
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

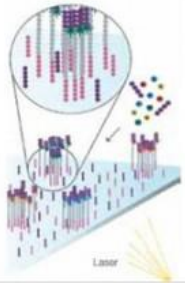


Figure 7
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

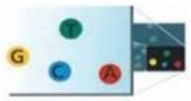


Figure 8
After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

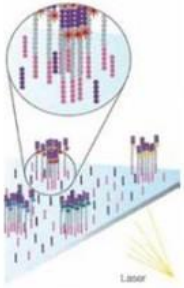


Figure 9
The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

Figure 10
After laser excitation, the image is captured as before, and the identity of the second base is recorded.

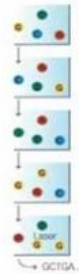


Figure 11
The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

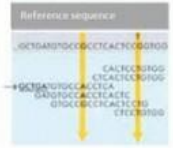


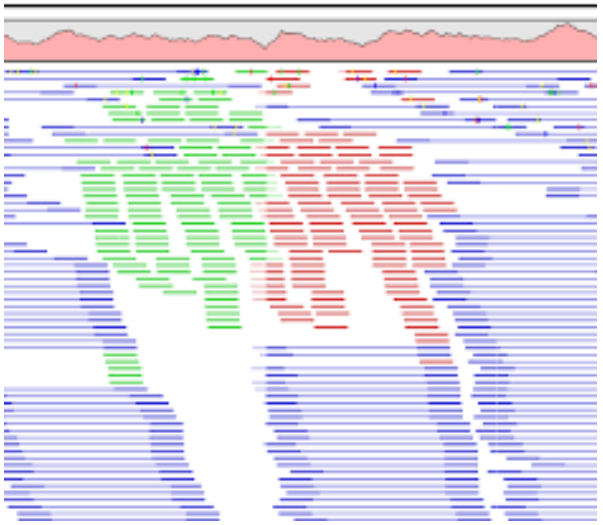
Figure 12
The data are aligned and compared to reference, and sequencing differences are identified.

Bio-informatique

- Enormément de données à traiter : penser à la mémoire !!! : machines dédiées
- Logiciels :
  - Quelques uns sous Windows dont CLC genomics: convivial, payant, peu modulable
  - Beaucoup d'application sous Linux : gratuit, très modulable, moyennement convivial
  - Un système intermédiaire : Galaxy https://usegalaxy.org/

Bio-informatique : plusieurs étapes

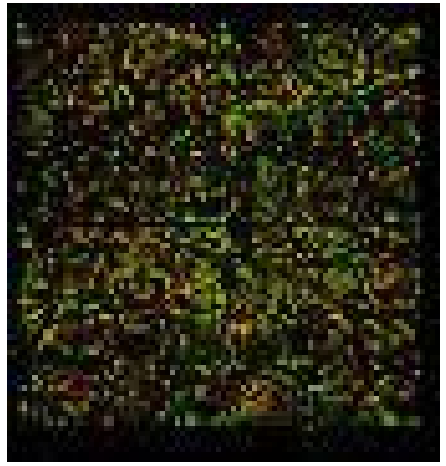
- Démultiplexage : séparation des séquences par génotypes
- Nettoyage : enlever les adaptateurs, bonne qualité, longueur minimale
- Alignement des séquences:
  - Sur une séquence de référence
  - Sur un assemblage de novo
- Détection des SNP
  - Profondeur de séquence
  - Fréquence minimale dôun allèle



CONCLUSIONS

Conclusions

- Une multitude de méthodes avec des avantages et inconvénients
- Adapter la méthode à la question posée et aux connaissances de lôespèce dôintérêt
- Génotypage de plus en plus spécialisé à sous traiter tout ou partie (Macrogen, Keygene, LGC genomicsé)



Conclusions

- Dans lôavenir séquençage complet systématique??
- Outils de bio-informatique cruciaux, problème de stockage
- Changement des outils dôanalyse des données de génotypage et de phénotypage maintenant en très grand nombre
- Des outils impressionnants pour mieux comprendre la diversité du monde vivants !!!!