Ministère de l'enseignement supérieur et de la recherche scientifique
Université Saad Dahleb, Blida
Faculté des sciences de la nature et de la vie
Département de Biologie
Mémoire de fin d'étude
En vue de l'obtention du diplôme de Master
Domaine Science de la Nature et de la vie
Filière : Biologie
Option : Génétique et physiologie

Thème :

Etude comparative génomique entre la race ovine domestique (Ovis aries) et la race ovine sauvage (Ovis musimon) au niveau des cellules musculaires via la nouvelle technologie à haut débit

Soutenu le : *18 /12/2014*

Présenté par :
KERRAOUCH Saida

Devant le jury :

Nom	*Grade*	*Lieu*	*Qualité*
Mme CHAKHMA . A			*Président*
Mr LAFRI M	*Professeur*	*FSAV*	*Promoteur*
Mme Ait Ouaamer. Y			*Examinatrice*
Mr BELHOCINE M.	*MAA*	*TAGC*	*Co-promoteur*

Promotion:

2011/2012

Remerciements :

A mon directeur de projet de fin d'étude de Master:
Monsieur M. Lafri
Grade : Professeur
Pour m'avoir proposé de travailler sur ce projet de thèse, pour m'avoir apporté l'aide nécessaire afin de mener à bien celui-ci. Merci pour votre disponibilité.
Sincères remerciements.

A mon co-directeur de projet de fin d'étude de Master :
Monsieur M. Belhocine
Grade : Maitre-assistant classe B
Pour son encouragement, ses conseils bienveillants, sa rigueur scientifique et sa grande disponibilité.
Qu'il trouve ici l'expression de ma sincère reconnaissance et de ma profonde considération.

A notre jury de projet de fin d'étude de Master

A mes parents
A qui je dois ce que je suis devenue aujourd'hui.
Pour ces nombreuses années de dévouement, de soutien et d'encouragement. Sans vous, je pense que je n'en serai pas là. Ce mémoire est la finalité de mes études mais aussi de celle de vos efforts.
Avec toute ma reconnaissance et ma profonde affection.

Dédicace :

Que Dieu le tout puissant soit loué de m'avoir permis d'atteindre les prémisses du savoir après un long chemin, et couronné mon effort par la joie de l'arrivée dont je l'espère, ce travail en sera témoin.

A ceux qui m'ont tout donné avec amour

A ceux qui m'ont encouragé et soutenu dans les moments les plus difficiles

Et à ceux qui je dois tant

A Mes chers parents, pour leur amour et leur support.

A Tout mes frères et ma chère et unique sœur

A mon cher mari Amine qui m'a toujours encouragé et soutenu

A mes beaux-parents ainsi que mes belles-sœurs sans oublier mes beaux-frères

A tous ceux qui me sont chers

En témoignage de ma profonde affection.

Sommaire :

Listes des figures et tableaux

Liste des abréviations

ADN : Acide **D**ésoxyribon**N**ucléique.

Chip: Chromatin **I**mmun**oP**recipitation

DAVID: **D**atabase for **A**nnotation**, V**isualization and **I**ntegrated **D**iscovery.

SAM: Significance **A**nalysis of **M**icroarrays

Seq: séquençage**.**

Tmev : **M**ulti **e**xperiment **v**iewer.

Résumé

Les principales races ovines algériennes constituent de par leur effectif et leur diversité, une richesse mondiale. Afin de préserver ce patrimoine, il convient préalablement de le caractériser précisément. A plus long terme, la compréhension de la diversité génétique et phénotypique ovine devrait permettre la mise en place de programmes de sélection optimisés et soucieux de préserver l'originalité de chaque race.

Dans cette étude on a essayé de faire une comparaison entre deux races différentes *Ovis aries* (domestique) et *Ovis musimon* (sauvage) avec l'aide de la bioinformatique afin de maitriser cette nouvelle technologie à haut débit en travaillant avec des logiciels tel que R, MeV, David , comprendre leurs fonctionnements et savoir interpréter leurs résultats.

Pourquoi ne pas avoir un résultat positif qui nous mène à faire une sélection génétique entre les deux races par rapport à la qualité de viande dont l'intérêt du consommateur.

Mots clés : *Ovis aries , Ovis musimon*, Bioinformatique , R , MeV, David, sélection génétique.

Abstract

The main Algerian sheep breeds are by their size and diversity, a global wealth. To preserve this heritage, it should first be accurately characterize. In the longer term, the understanding of the genetic and phenotypic diversity sheep should allow the establishment of breeding programs optimized and anxious to preserve the originality of each race.

In this study we tried to make a comparison between two different races Ovis aries (domestic) and Ovis musimon (wild) with the help of bioinformatics in order to control this new broadband technology by working with software such as R, MeV, David, understand how they work and interpreting their results.

Why not have a positive outcome that leads to genetic selection between the two races over the quality of meat that the consumer interest.

Keywords: *Ovis aries, Ovis musimon*, Bioinformatics, R MeV, David, genetic selection.

الملخص

اهم السلالات الجزائرية في فئة الاغنام , من حيث عددها و تنوعها تمثل ثروة عالمية.

للحفاظ على هذا التراث, اولا يجب تمييزها بدقة كبيرة. و على المدى الأطول, ينبغي فهم التنوع الوراثي و الفيزيولوجي للأغنام, و هذا يسمح ب تحديد برامج كل دورها الحماية و الحفاظ على أصل كل سلالة.

في هده الدراسة حاولنا إجراء مقارنة بين اثنين من مختلف الأعراق الغنمية الخروف"المربى" الغنم "البري " بمساعدة المعلوماتية الحيوية, من أجل التحكم في هذه التكنولوجيا الجديدة ذات النطاق العريض ا من خلال العمل على البرامج

الالكترونية مثل *R, MEV,David*

وفهم طريقة عملها و تفسير نتائجها.

لماذا لا , قد تكون النتيجة ايجابية و تؤدي الى الانتقاء الجيني بين سباقين على نوعية اللحوم و كل هذا في مصلحة المستهلك.

الكلمات المفتاحية الخروف المربى, الغنم البري, المعلوماتية الحيوية , الاختيار الوراثية ,

R ,MEV,David

Introduction :

Connue par l'immensité de son territoire et la riche diversité de ses milieux, l'Algérie recèle des ressources dont l'importance tant qualitative que quantitative est à même de lui assurer un développement agricole et rural d'une durabilité indéniable.

Nonobstant cette importance, ces ressources ne sont guère exploitées de façon appropriée. Les espèces animales et végétales, avec toutes les races et les variétés et populations qui les caractérisent, non seulement sont peu connues mais sont en voie d'extinction, voire disparues pour certaines avec toutes conséquences négatives que cela induit tant sur le plan écologique qu'économique.

Les espèces animales, en particulier, représentent non seulement des ressources vitales pour le pays mais également pour le patrimoine génétique universel. Malgré leur importance primordiale pour la sécurité alimentaire et pour le développement économique et social, ces ressources sont sujettes actuellement au processus d'érosion génétique qui va en s'accentuant.

Avec l'avènement des nouvelles technologies à haut débit, on a pu résoudre beaucoup de questions dans le domaine scientifique. C'est dans ce contexte et dans l'optique de viser des mesures appropriées pour la préservation de ces ressources, qu'on a voulu appliquer cette nouvelle technologie sur diverses races ovines dans le but de la tester prochainement sur nos races ovines algériennes dont l'objectif est l'amélioration génétique des races.

Ce projet se base beaucoup plus sur la compréhension de cette technologie et sa maitrise pour l'établissement d'un protocole d'analyse transcriptomique.

Partie bibliographique

Chapitre I : Mouton en Algérie et les deux races d'étude

1. Place de l'élevage ovin dans l'économie mondiale :

En 2006, le monde comptait 1.1 milliard d'ovins soit une proportion d'environ un mouton pour cinq habitants. Ce cheptel est en recul ; il a perdu 5% en 15 ans (Pictoris, 2008). Il est surtout exploité actuellement pour sa viande et pour sa laine. La production laitière demeure très limitée en quantité et localisée autour du bassin méditerranéen.

La Chine rassemble le premier cheptel ovin au monde avec près de 160 millions têtes d'ovins (soit 15 % du cheptel mondial). L'Océanie (Australie et Nouvelle-Zélande) arrive en deuxième place, rassemblant 13 % des reproducteurs, suivie de l'Union européenne (11 % du cheptel)
(Anonyme ,2007).

Le mouton a des capacités d'adaptation remarquables. A l'origine, animal des pays chauds et secs, il est présent aujourd'hui sous toutes les latitudes, depuis le nord de l'Europe jusqu'aux zones tropicales (Bourguignon, 2006).

2. Situation de l'élevage ovin dans l'économie algérienne :

La production annuelle de viande contrôlée est estimée à 16500 tonnes soit 65% de la production nationale (Chemmam, 2007). A cela s'ajoute les quantités provenant de l'abattage non contrôlé (estimées à 40% de cette quantité) et les sacrifices des fêtes et périodes religieuses. En Algérie, la production de viande reste insuffisante pour la demande locale, elle est complétée par l'importation annuelle de 19.7 tonnes de viande bovine et ovine (Chemmam, 2007).

L'élevage ovin représente la spéculation agricole la plus importante. Le secteur de la production animale fournie près de 5 billions de dollars. L'élevage des petits ruminants contribue avec 52% et représente 35% de la production agricole totale (Benaissa, 2001). Il occupe ainsi une place importante sur le plan économique et social. Sa contribution à l'économie nationale est importante dans la mesure où il représente un capital de plus d'un milliard de dinars. C'est une source de revenu pour de nombreuses familles à l'échelle de plus de la moitié du pays. (Mohammedi, 2006 cité par Saidi et al., 2009).

3. Effectif du cheptel en Algérie :

Les effectifs : ovin, caprin, bovin et camelin se sont accrus respectivement de 4.20%, 3.54%, 1.11%, et 2.75% par rapport à l'année 2005.

L'élevage ovin domine avec un effectif de 19 .6 millions de têtes, en deuxième position les caprin avec 3.7 millions de têtes, suivi du bovin avec 1.6 millions de têtes et en dernier le camelin avec 0.3 million de têtes (Ministère de l'Agriculture et du Développement Rural, 2006).

En Algérie, il y a une spécialisation des zones agro-écologiques en matière d'élevage. L'élevage bovin reste cantonné dans le Nord du pays avec quelques incursions dans les autres régions.

Les parcours steppiques sont le domaine de prédilection de l'élevage ovin et caprin avec plus de 90 % des effectifs qui y vivent, entraînant une surexploitation de ces pâturages (Nedjraoui, 2001).

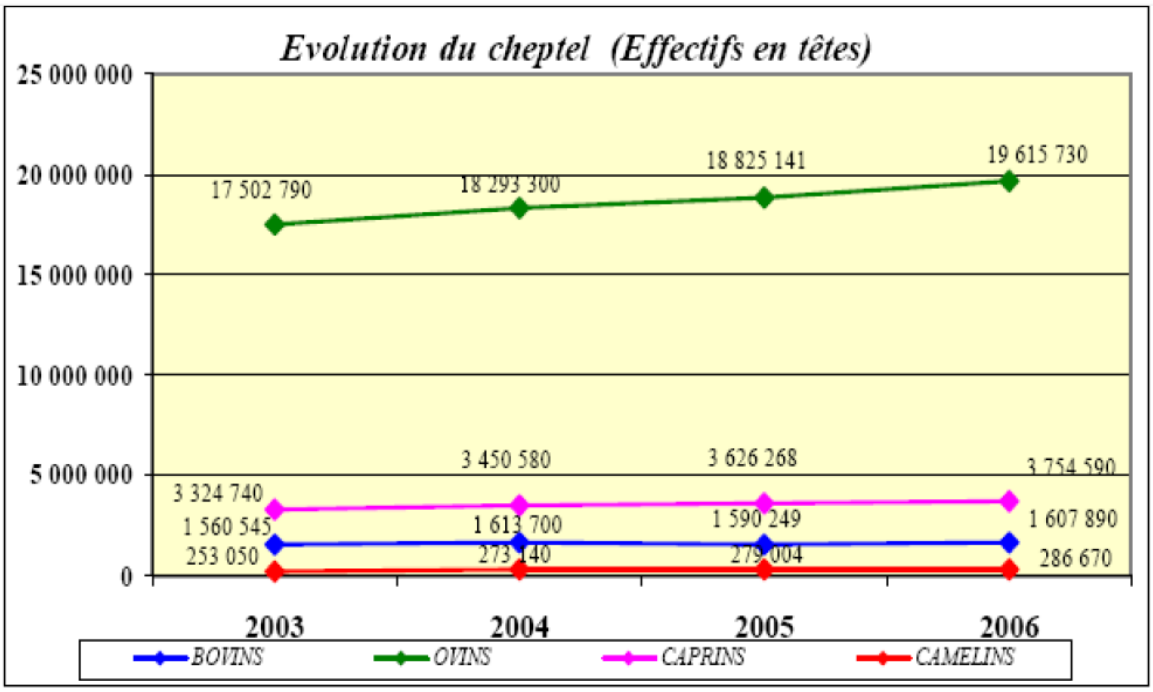


Figure 1. Evolution du cheptel (millions de têtes) MADR, 2006

4. Les deux races ovines d'étude :

4.1. ***Ovis aries :***

Le **mouton** (*Ovis aries*) est un mammifère domestique herbivore de la famille des bovidés, de la sous-famille des Caprinés et du genre *Ovis*. L'animal jeune est l'agneau (féminin : agnelle), la femelle est la brebis et le mâle est le bélier.(Budiansky, p. 97–98)

C'est un mammifère ruminant qui est présent aujourd'hui, surtout, sous sa forme domestiquée, bien que six espèces sauvages existent toujours. Sous le nom de Mouton, les ovins sauvages ont été domestiqués très précocement, sans doute il y a une dizaine de millénaires. Après de nombreuses interrogations sur le groupe originel, les études génétiques récentes pointent vers plusieurs phénomènes de domesticatio (deux ou trois selon les auteurs) au sein des sous-espèces du groupe le plus occidental, le mouflon à 54 chromosomes (*O. orientalis* ou *O.gmelini* selon les noms les plus utilisés), ce qui pointe vers la zone Moyen-Orient - Iran, dont l'espèce est originaire. (Budianksy, p. 100–01)

4.2. ***Ovis Musimon :***

La sous-espèce dénommée traditionnellement « Mouflons de Corse » (*ovis ammon musimon*) prend la dénomination « *ovis gemelini musimon*) .

Le Mouflon de Corse est un des plus petits mouflons d'Eurasie. Comme tous ses congénères, il présente un dimorphisme sexuel et saisonnier très prononcé. Mâle adulte : poids, environ 35-50 kg ; longueur, environ 130-140cm ; hauteur au garrot, environ 75cm. Femelle adulte : poids, environ 25-35kg ; longueur, environ 120-130cm ; hauteur au garrot, environ 65cm. (DUBRAY D,1988)

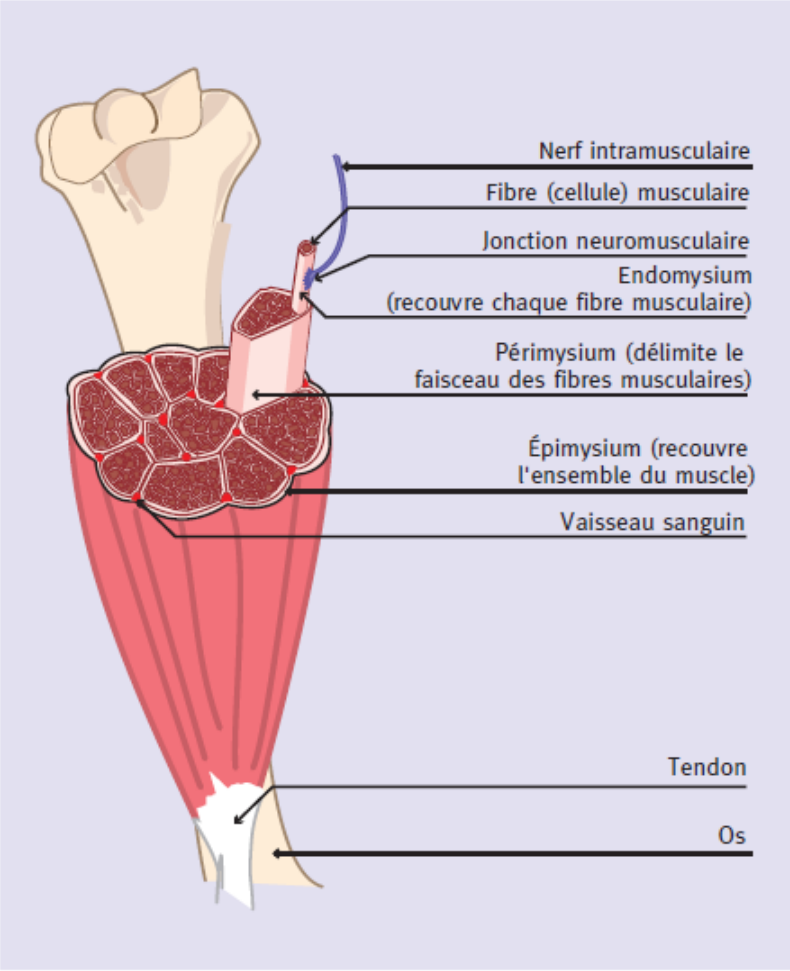
Le corps est trapu mais d'allure élégante, les membres sont terminés par de petits sabots dépourvus de membrane interdigitale ; la queue est toujours très courte. (DUBRAY D , 1988).

Les deux cornes du male, présentes systématiquement, sont triangulaires à la base et généralement symétriques ; elles peuvent atteindre 85cm de longueur ; leurs

courbures sont très prononcées et différentes entre les populations du nord et du sud de l'ile. Chez les femelles cornues (environs 10% seulement des femelles sur le Cinto, mais environ 75% sue Bavella), les cornes sont courtes et souvent dissymétriques.

La dentition est adaptée au régime herbivore (formule dentaire : 0.0.3.3/3.1.3.3 = 32 dents).

5. La cellule musculaire :tissu musculaire strié



Le mot muscle vient du mot latin *musculus* qui signifie « petite souris ». les muscles peuvent etre considérés comme les « moteurs » de l'organisme. Les propriétés des muscles : excitabilité, contractilité, élasticité,…. Leur permettent de générer force et mouvement. Le système nerveux est indispensable à leur fonctionnement.

Les muscles striés squelettique sont constitués de cellules allongées : les fibres musculaires.

Figure 2 : Anatomie du muscle strié squelettique

Associées en faisceaux, ces fibres sont rendues solidaires par des enveloppes élastiques. Chaque fibre musculaire présente de nombreux noyaux répartis à la périphérie de la cellule. Elle est délimitée par une membrane (sarcolemme) et contient dans son cytoplasme (sarcoplasme) des myofibrilles qui constituent le support de la contraction musculaire.

Les myofibrilles présentent une structure filamentaire régulière (myofilaments) qui donne au muscle son aspect strié au microscope.

Une fibre musculaire résulte de la fusion de plusieurs cellules non différenciées à noyaux uniques appelées myoblaste. Le myotube, formé par la fusion des myoblastes , est caractérisé par des noyaux en position centrale.

Puis, lors de la différenciation du myotube en fibres musculaire, les noyaux vont se placer en périphérie de la cellule musculaire.

Chapitre II : La génomique

1. La génomique :

La biologie moléculaire est donc entrée depuis 1995 dans l'ère de la génomique: on dispose maintenant de l'information génétique exhaustive sur un nombre croissant d'organismes vivants ; et il est aujourd'hui possible d'aborder de manière globale un certain nombre de problèmes complexes dont on n'avait jusqu'à présent qu'une connaissance fragmentaire : voies métaboliques, interaction de la cellule avec l'extérieur, mécanismes globaux de régulation et de contrôle. Une nouvelle discipline est également née de la connaissance de ces séquences complètes de chromosomes : la génomique comparée. Il est maintenant possible de comparer deux organismes vivants à l'échelle de leur génome, de déterminer les gènes qu'ils ont en commun ou qui leur sont propres. Ce type d'analyse est très prometteur dans le contexte de l'identification sélective de gènes correspondant à des cibles thérapeutiques : en comparant par exemple une bactérie pathogène et une proche cousine non-pathogène, on peut essayer de repérer les gènes impliqués dans la virulence de la souche infectieuse. (Broder, S. & Venter, J. C. (2000)

L'accélération du séquençage, permise en particulier par la robotisation et la parallélisation des méthodes d'analyse, nécessite un soutien de plus en plus important de l'outil informatique. Dans un premier stade, celui-ci est indispensable pour permettre l'assemblage du gigantesque « puzzle » que constituent les milliers ou millions de fragments de génome issus des automates de séquençage. Ensuite l'informatique est un outil incontournable pour extraire et analyser l'information contenue dans ces gigabases (1 Gbase = 109 nucléotides) de séquence. Il est clairement impossible de caractériser expérimentalement tous les gènes contenus dans ces séquences et c'est pourquoi l'analyse *in silico* doit venir au secours des biologistes pour compléter et guider les approches *in vitro* et *in vivo*.(Dear, S. & Staden, R. (1991)

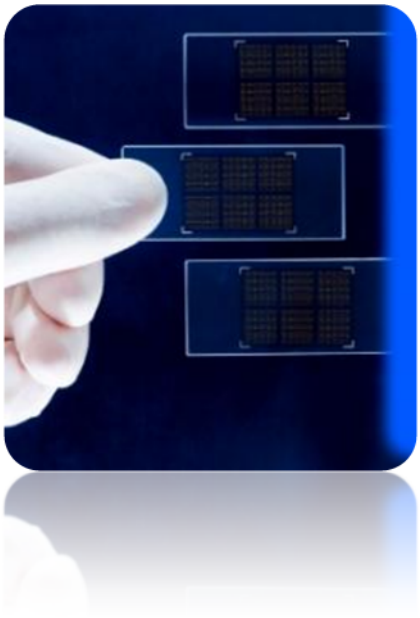
La bioinformatique, discipline récente, traite des différents aspects de ce nouveau champ de la connaissance et s'appuie bien sûr à la fois sur les concepts de la biologie et de l'informatique, et sur des outils issus de la chimie et de la physique.

Chapitre III : Les nouvelles technologies en Biologie.

1. Les puces à ADN :

1. 1 Principe et types de puces à ADN

D'abord conçues sur des membranes poreuses de nylon (appelées parfois « macroarrays » par opposition aux « microarrays »), les puces à ADN ont été progressivement mises au point sur lames de verre à la fin des années 90. La miniaturisation, rendue possible par l'utilisation d'un support solide, de marqueurs fluorescents et par les progrès de la robotique, permet aujourd'hui de fabriquer des puces comportant une très haute densité de spots, susceptibles de recouvrir l'intégralité du génome d'un organisme sur une simple lame de microscope. (*J.-M. BIDART et L. LACROIX)*



Les puces à ADN sont une technique d'identification moléculaire qui regroupe des techniques de biologie, de microélectronique, d'analyse d'image et de bioinformatique. Leurs domaines d'application sont très étendus et intéressent de nombreux secteurs de la recherche biologique notamment la génomique fonctionnelle, la recherche pharmaceutique, le génotypage, le diagnostic et le contrôle alimentaire (O. CROCE ;2005).

Une puce à ADN, aujourd'hui communément appelée « DNA microarray » en anglais (de « array » = rang ordonné), est constituée de fragments d'ADN immobilisés sur un support solide selon une disposition ordonnée. Son fonctionnement repose sur le même principe que des technologies telles que le Southern blot ou le northern blot, qui sont couramment utilisées pour détecter et quantifier la présence d'une séquence nucléique spécifique au sein d'un échantillon biologique complexe, par hybridation à une sonde de séquence complémentaire portant un marquage radioactif. La confection des puces à ADN a permis d'étendre ce principe à la détection simultanée de milliers de séquences en parallèle.

Une puce comporte quelques centaines à plusieurs dizaines de milliers d'unités d'hybridation appelées « spots » (de l'anglais spot=tache), chacune étant constituée d'un dépôt de fragments d'ADN ou d'oligonucléotides correspondant à des sondes de séquences données. L'hybridation de la puce avec un échantillon biologique, marquée par un radioélément ou par une molécule fluorescente, permet de détecter et de quantifier l'ensemble des cibles qu'il contient en une seule expérience.

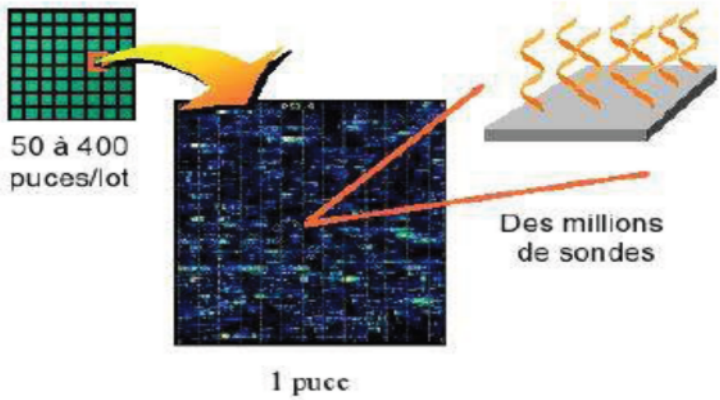


Figure 3 : Une puce peut contenir de quelques dizaines de rapporteurs (sondes) à plusieurs millions *(schéma de Zintilini C., 2003).*

On distingue plusieurs types de puces selon la densité des spots, le mode de fabrication, la nature des fragments fixés à la surface et les méthodes d'hybridation. Les caractéristiques des puces les plus courantes sont résumées dans le Tableau **(**Puces à ADN. Méthodes d'étude du génome et du transcriptome (*J.-M. BIDART et L. LACROIX)***.**

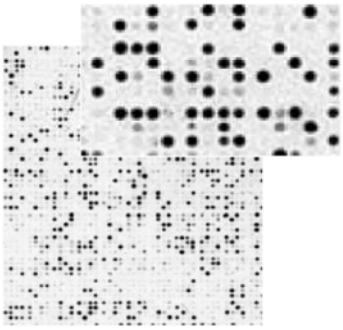
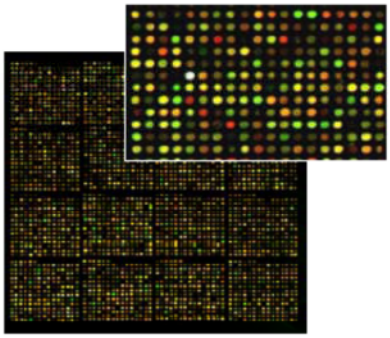
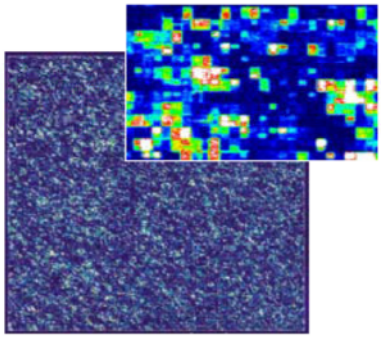
« Macroarray »	« Microarray spottée »	« GeneChips » de Affymetrix
		
- support : membrane de nylon - taille des spots : 0,5-1mm - densité : quelques centaines de spots/cm^2 - sondes : produits de PCR - cibles : ADNc avec marquage radioactif au ^{32}P - principales applications : analyse de l'expression des gènes	- support : lame de verre à revêtement chimique - taille des spots : ~100µm - densité : 1000-10000 spots/cm^2 - sondes : produits de PCR ou oligonucléotides longs (30-70mers) - cibles : ADNc ou produits de PCR avec marquage fluorescent au Cy3 et Cy5 - principales applications : analyse de l'expression, ChIP-on-Chip, CGH-array	- support : lame de verre à revêtement chimique - taille des spots : ~20µm - densité : jusque 250000 spots/cm^2 - sondes : oligonucléotides courts (20-25 mers) synthétisés *in situ* - cibles : ARNc ou produits de PCR avec marquage fluorescent à la biotine-streptavidine - principales applications : analyse de l'expression, détection de marqueurs moléculaires

Table 1 : Principaux types de puces à ADN

Les deux principales utilisations des puces à ADN sont:

– les puces d'expression où chaque rapporteur reconnaît des gènes ou transcrits distincts au sein d'un même génome;

– les puces d'identification où chaque rapporteur reconnaît les variants d'un même gène issu de génomes différents (*Olivier CROCE ;2005)*.

1. 2. La fabrication des puces à ADN :

1. 2. 1. La fabrication de la puce :

La méthode de fabrication des puces « spottées » a été développée par l'équipe de P. Brown à l'Université de Stanford, aux Etats-Unis (DeRisi et al ; 1997). Elle est aujourd'hui bien établie et de nombreuses plate-formes de production sont implantées dans les laboratoires académiques. Des solutions d'ADN sont préparées soit par amplification PCR à partir du génome ou de banques d'ADN complémentaires, soit par synthèse d'oligonucléotides longs (30-70 mers). Des micro-gouttelettes de ces solutions sont ensuite déposées par un robot, selon une matrice d'emplacements définis, sur une lame de verre traitée par un revêtement chimique qui permet de fixer l'ADN.

En général, chaque spot de la matrice correspond à un gène donné. Les robots nécessaires à la fabrication de ces puces étaient construits à l'origine de manière artisanale dans chaque laboratoire selon le modèle conçu par J. DeRisi et dont les plans de montage et le logiciel de pilotage sont disponibles sur Internet (**Puces à ADN. Méthodes d'étude du génome et du transcriptome** *J.-M. BIDART et L. LACROIX*).

1. 2. 2. La fixation des sondes :

Il existe deux méthodes : le transfert de brins d'ADN (« off chip synthetis ») ou sa synthèse *in situ* (« Light directed in situ synthesis »).

- *Transfert de brins d'ADN :*

La synthèse préalable à la fixation des sondes permet de fixer des sondes relativement longues, atteignant 40 à 60 bases. Le transfert de ces sondes sur la puce peut se faire au moyen de micropipettes, de micropointes ou par des dispositifs de type jet d'encre.

- *La synthèse in situ :*

Dans ce cas, la construction des sondes se fait par dépôt de couches successives des quatre bases de l'ADN sur un support en verre. C'est un masque, dont la configuration varie à chaque dépôt d'une couche, qui permet aux bases de s'empiler correctement. Avec ce procédé, utilisé par Affymetrix, les sondes comportent au maximum 30 bases (Viard Bruno et Errachid Céline ; 2006).

1. 3. Préparation des échantillons, amplification, marquage et hybridation des cibles :

1. 3. 1. La préparation des acides nucléiques :

Pour l'étude du transcriptome, l'ARN doit tout d'abord être purifié afin de limiter la contamination par l'ADN génomique, les protéines et les débris cellulaires qui peuvent perturber les différentes étapes enzymatiques d'amplification ou de marquage ainsi que l'hybridation. A l'heure actuelle, il existe de nombreux protocoles d'extractions bien standardisés qui permettent l'obtention d'acides nucléiques de bonne qualité (mécanisme)

1. 3. 2. Amplification et marquage :

En règle générale, 1 à 5 µg d'ARN total sont suffisants. L'utilisation de quantités plus faibles (100 ng à 1 µg) nécessite une double amplification qui induira une augmentation du bruit de fond, donc à une diminution de la sensibilité. Le marquage des cibles peut être réalisé directement lors de la réaction de transcription inverse (*reverse transcription* ou RT) par incorporation de cytosine marquée (fluorescence ou radioactivité) dans les acides nucléiques néosynthétisés. Cependant, ce type de marquage nécessite une quantité de matériel de départ importante, ou indirect par incorporation d'un nucléotide modifié amino allyl dUTP puis couplage du flurochrome dans un second temps. Ceci se fait suite à la transcription inverse des ARNms permettant l'obtention d'un brin d'ADNc fluorescent.

1. 3. 3. L'hybridation et lavage:

Du fait de la complémentarité des nucléotides, le dépôt des cibles marquées sur la puce déclenche l'appariement des séquences sondes/cibles complémentaires. Cette hybridation, qui dure quelques heures en milieu liquide est suivie d'un lavage du substrat qui permet d'éliminer les cibles non fixées, ou fixées non spécifiquement et qui consiste en une immersion dans une série de tampons. (*J.-M. BIDART et L. LACROIX)*

Les paramètres essentiels pour limiter toute hybridation non spécifique sont la durée d'hybridation, la température, la stringence du tampon, la concentration de la sonde, l'encombrement stérique des sondes, la densité des dépôts et la séquence des sondes. L'hybridation est généralement réalisée sous agitation à 60 °C, pendant 10 à 17 heures, dans un tampon de haute Stringence. Après séchage la puce est passée au scanner pour repérer les hybridations (J.-M. BIDART et al ; 2007)

1. 4. La lecture ou acquisition des images :

La puce est alors révélée par lecteur (scanner) muni de lasers qui permettent d'exciter les molécules de fluorochrome et de détecter par microscope confocal le signal émis dans chaque spot. Dans le cas du marquage avec deux fluorochromes (vert Cy3 et rouge Cy5), une image numérique est acquise pour l'échantillon marqué avec le Cy3 et une en Cy5. Un spot de couleur verte indique un gène dont le niveau d'expression est plus élevé dans l'échantillon marqué avec le Cy3 que celui marqué avec le Cy5, et inversement pour un spot de couleur rouge. Le spot apparait jaune lorsque le gène est exprimé de manière identique dans les deux échantillons comparés. L'analyse des données numériques issues de l'acquisition est effectuée par un logiciel qui prétraite, segmente et quantifie les différents niveaux d'activité dans les spots.

1. 5. L'analyse des données :

L'analyse d'image est un aspect central des expériences menées à l'aide des puces à ADN. Dans le cas d'un double marquage, le but de cette phase est de quantifier, de manière relative, le niveau d'expression des gènes. Cette mesure,

basée sur un rapport d'intensité entre les deux niveaux de fluorescence détectés, est fonction de nombreux paramètres dépendant des méthodes utilisées, des mesures expérimentales et des conditions biologiques. Globalement, cette étape d'analyse a un impact considérable sur l'interprétation biologique des données et repose sur quatre phases.

❖ La localisation des spots :

Il s'agit, ici, de déterminer les coordonnées de chaque spot de la puce. Cette étape est normalement effectuée à l'aide d'une grille théorique définie lors du plan de dépôt des sondes. Pour localiser un spot sur une image, c'est-à-dire correspondre un modèle idéal de puce avec une image acquise, un nombre important de paramètres doit être estimé (espaces entre les spots, espaces entre blocs d'une puce…). De même que la précision du spotter.

Le repérage des spots doit être simple et rapide à réaliser. L'automatisation de cette tâche permet une accélération considérable de l'analyse.

Enfin, la définition des grilles doit être aussi juste que possible. En effet, la précision de cette étape dépend de l'efficacité des mesures ultérieures.

❖ L'extraction des indices :

Maintenant que les pixels « signal » sont identifiés, il s'agit d'extraire un ou plusieurs indices pertinents permettant de quantifier le niveau d'expression de chaque gène. Ce niveau correspond à une mesure relative des intensités de fluorescences en rouge et vert. Un filtrage des données permet de sélectionner des variations significatives (tri des spots sur un barème de critères), à partir desquelles un rapport de fluorescence est calculé. La valeur de ce ratio indique l'induction ou la répression du gène.

Dans le cas de la comparaison de données issues de plusieurs expériences (données temporelles, sain/pathologique…), il faut de plus nécessairement

normaliser les données avant quantification pour éliminer les artefacts dus par exemple au protocole expérimental.

❖ La classification des données :

Enfin, les données préparées par les étapes précédentes permettent de regrouper les gènes par familles ayant des comportements semblables, en analysant les niveaux d'expression puis en classant les gènes de proche en proche (technique pairwise) ou en faisant appel à des techniques plus sophistiquées tel que l'analyse de la composante principale ou des réseaux neuronaux. La représentation des résultats se fait par claustring sous forme d'une barre verticale où chaque colonne correspond à une expérience et chaque ligne correspond à un gène. Les ratios vert/rouge sont représentés par une échelle à couleur.

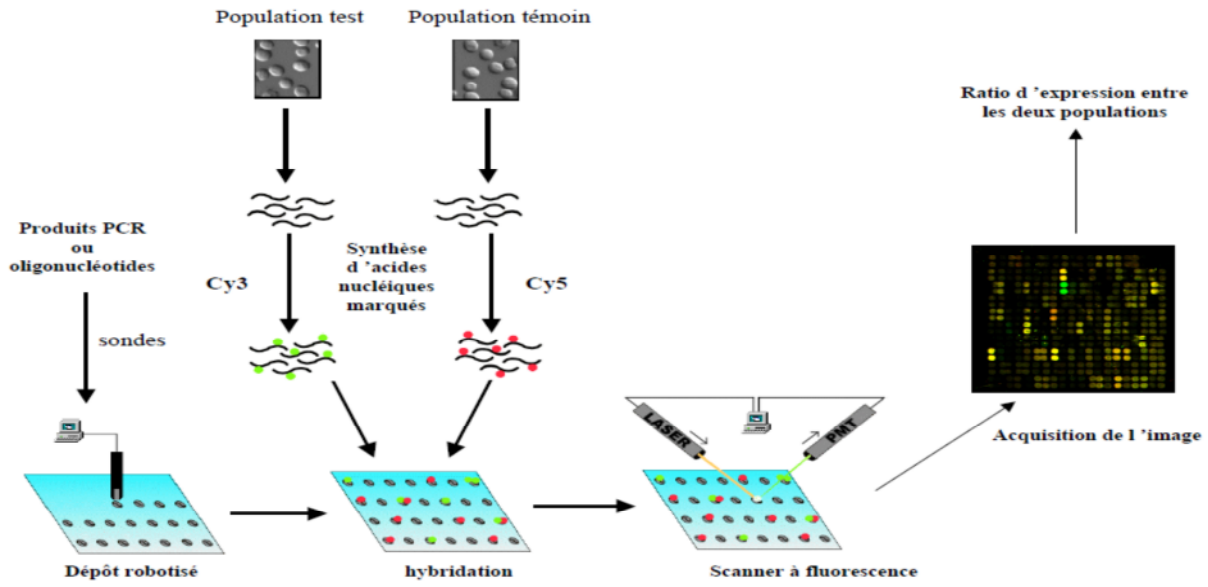


Figure 4 : Principe générale des puces à ADN. (Eisen et Brown, Methods in Enzymology 1999)

1. 6. Exemples d’application des puces à ADN :

Le cancer du sein est le second cancer le plus fréquent, avec une incidence de 1 million de nouveaux cas par an dans le monde (OMS 2003). De nombreuses équipes ont utilisé les méthodes d’analyse du transcriptome pour essayer de mieux comprendre la physiopathologie de cette maladie et en améliorer les stratégies thérapeutiques.

Les premières études ont permis de confirmer l’importance du statut des récepteurs d’oestrogène (ER) et l’hétérogénéité moléculaire de cette pathologie, en regroupant les tumeurs en plusieurs groupes : les *basal-like* (principalement ER négatives) et les *luminal-like* (principalement ER positives), dont le profil d’expression inclut respectivement de nombreux marqueurs des cellules myo-épithéliales (dites basales), ou des cellules luminales, et les erbB2 (HER2) positives surexprimant cet oncogène.

Chapitre IV : Bioinformatique et ses outils :

Ces dernières années, la recherche en biologie et tout particulièrement en génétique a connu un formidable essor et continue sur sa lancée. Les avancées réalisées sont considérables. Cependant, si à l'époque de Gregor Mendel1, il suffisait de faire certains croisements entre diverses espèces de pois comestibles pour faire de grandes avancées dans le domaine de la génétique, de nos jours les chercheurs utilisent d'autres techniques : techniques qui demandent à traiter une très importante somme d'informations que nous ne pouvons traiter sans l'aide de l'informatique. *(BOUKADIDA Jawer ,DENIS Julien 2004)*

A tel point, qu'il y a un peu plus d'une dizaine d'années, une nouvelle discipline a été créée : la bio-informatique. Située au carrefour de la biologie, des mathématiques, des statistiques et de l'informatique, elle consiste à utiliser les possibilités offertes par l'informatique afin d'acquérir, traiter, organiser et interpréter l'information concernant la vie comme l'indique son préfixe « bio » qui a pour racine « *Bio »* signifiant « la vie » en grec ancien.

Cette notion englobe l'ensemble des applications de l'informatique aux sciences de la vie, domaine très vaste qui recouvre tous les axes de recherche, allant des applications en robotique aux techniques les plus avancées en intelligence artificielle. Pour la plupart des membres de la communauté scientifique, cette notion semble dans la pratique s'adapter, plus particulièrement, aux outils informatiques qui permettent de stocker, d'analyser et de visualiser les informations contenues dans les séquences des gènes et des protéines des êtres vivants.

L'histoire de la bioinformatique est donc étroitement liée à celle de la biologie moléculaire, l'étude des molécules du vivant.

Il est intéressant de constater que l'essor des connaissances en biologie moléculaire progresse parallèlement à celle de l'informatique. En ce qui concerne la biologie moléculaire, un tournant important a été impulsé par la mise au point de techniques de séquençage de l'ADN en 1977 (conjointement par Frederick Sanger d'une part, et par Allan Maxam et Walter Gilbert d'autre part). Il faudra attendre le

milieu des années 1980 pour voir apparaître le développement des premiers robots séquenceurs.

Dans les mêmes années l'informatique connaîtra de grandes avancées : avènement des micro-ordinateurs personnels et création de langages de programmation évolués (comme le langage C). Les biologistes s'apercevront rapidement du bénéfice qu'ils pourront tirer de tels outils. Il faut, en effet, se rappeler que les premières recherches en biologie moléculaire ont été menées avec des moyens très limités : en 1962, la résolution de la structure de la myoglobine a demandé à Max Perutz le traçage manuel de plus de deux mille cercles. La visualisation tridimensionnelle des molécules nécessitait par ailleurs la construction fastidieuse d'imposantes structures à base de tiges métalliques qui encombraient les bureaux des chercheurs.

Un des pionniers de la bio-informatique, certainement Rodger Staden, a très vite ressenti l'intérêt de développer des programmes pour analyser les séquences. Dès 1977, il propose ainsi des outils informatiques qui servent encore aujourd'hui (un package très utilisé porte son nom).

La définition exacte du terme bio-informatique constitue une source récurrente de dissensions au sein de la communauté scientifique. Deux approches peuvent être relevées : la première consiste à concevoir la bio-informatique comme un procédé nouveau d'investigation biologique ; la deuxième approche associe cette discipline à un ensemble d'outils mis à la disposition des biologistes pour valider des expériences biologiques.

Les tenants de la première approche considèrent que l'informatique bouleverse fondamentalement la recherche en biologie moléculaire. Elle apporte un nouveau paradigme de recherche défini par la conception de modèles mathématiques sur lesquels peuvent se mener des expériences *in silico*. Ce néologisme a été créé par analogie avec le terme latin *in vivo* pour désigner des simulations numériques dont le but est la découverte de nouvelles lois ou fonctions biologiques, par opposition aux manipulations expérimentales classiques. Un exemple basique de découverte est l'inférence de la fonction d'une protéine à partir

de sa séquence primaire, par identification de séquences similaires dont la fonction est connue. Cette conception de la bio-informatique est à rapprocher du terme anglo-saxon *computational biology*, qui accorde une grande part à la théorie et à la modélisation. La deuxième approche consiste à concevoir l'informatique comme un outil d'analyse de données adapté aux besoins des biologistes.

Cette discipline conçoit et développe des méthodes et des logiciels pour le stockage et le traitement de données biologiques. Ce dernier point de vue, partagé par de nombreux biologistes, s'explique aisément par un souci pragmatique de gérer la masse d'informations nouvelles extraite d'expériences à haut débit. La création de bases de données a très certainement été un des premiers objectifs des chercheurs. Les données biologiques progressent en outre à un rythme accru et cette accumulation ne semble pas connaître de limites. Depuis plus d'une décennie, on observe une augmentation gigantesque du volume de données disponibles. Des bases de données, telles que GenBank [Benson 2001] pour les acides nucléiques et SwissProt [Bairoch 2000] pour les protéines, ont vu leurs données doubler de taille tous les quinze mois. Début 2010, GenBank renfermait pas moins de 106 533 156 756 bases correspondant à 108 431 692 de séquences ; SwissProt contenait 180 900 945 acides aminés provenant de 186 149 références annotées (la version basée sur de l'annotation automatique (base TrEMBL) contenait 10 158 056 de séquences). L'avènement de nouvelles technologies comme le *Next Generation Sequencing* [Margulies 2005] a considérablement compliqué la donne, puisque certains séquenceurs peuvent engendrer jusqu'à 3 Gigabases par jour [Richter 2009].

Le coût du séquençage devrait baisser et permettre le développement de la génomique individuelle, et donc une véritable révolution dans les pratiques des chercheurs en biologie et des cliniciens. La Table 2 présente quelques-unes des techniques utilisées en biologie moléculaire.

Technologie	But
Séquençage d'ADN	Détermination de l'ordre exact des nucléotides dans un fragment d'ADN, assemblage automatique des fragments, reconstitution d'un génome.
Spectrométrie de masse	Caractérisation des molécules d'après leurs masses, identification de protéine.
Chromatine Immuno-Précipitation on chip (chip-chip)	Identification des sites de fixation des facteurs de transcription.
Chip-seq	Technique alternative au chip-chip utilisant le séquençage des fragments d'ADN en liaison avec un facteur de transcription.

Table 2. Eventail de technologie de Biologie Moléculaire

La bio-informatique est née il y a presque une vingtaine d'années pour subvenir aux biologistes qui avaient besoin d'un support permettant de stocker un nombre de données ne cessant d'augmenter, et d'un outil y facilitant l'accès et en simplifiant le traitement.

Certains pays ont un rôle prépondérant dans le développement de cette discipline, alors que d'autres tentent de rattraper le retard qu'ils ont accumulé.

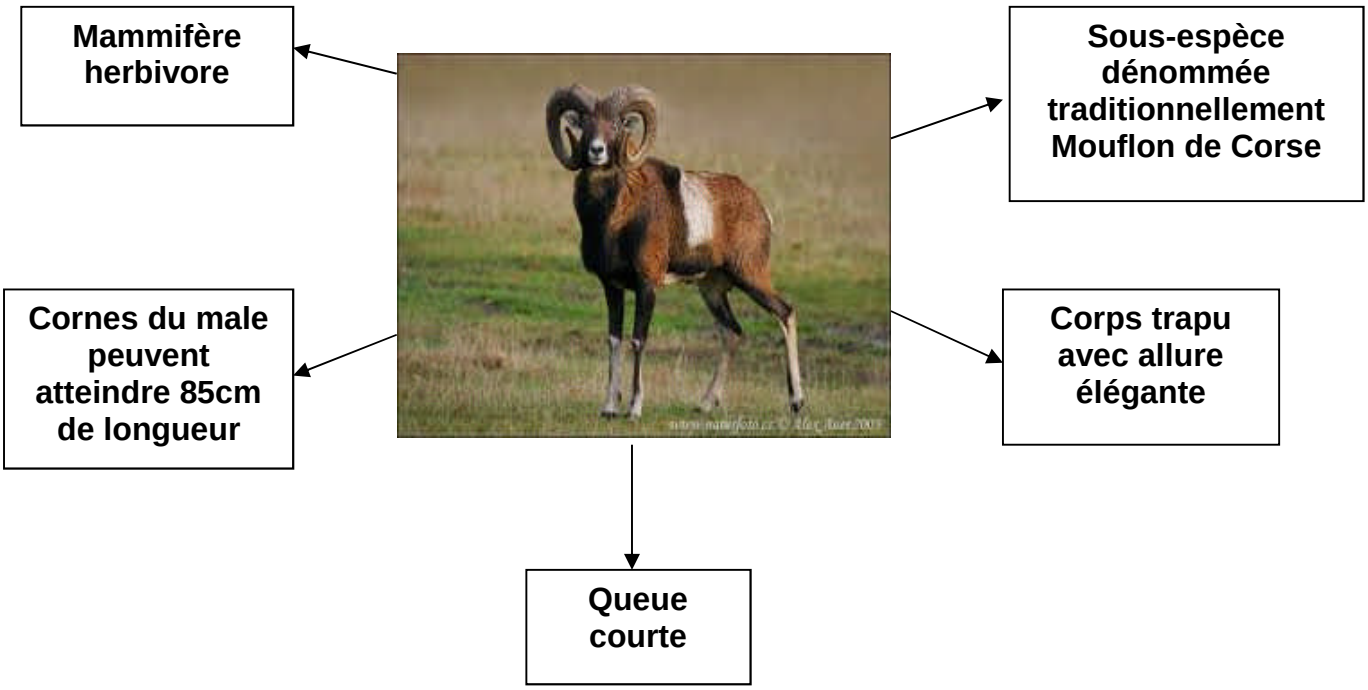
La raison pour laquelle des bio-informaticiens ont été formés, est qu'il était nécessaire de disposer d'individus ayant une double compétence : d'une part, biologiste afin d'avoir les connaissances nécessaires pour comprendre les problèmes soulevés par la génétique moderne et autres branches de la biologie; et d'autre part, informaticien afin de pouvoir créer des bases de données, mettre au point des logiciels et développer des algorithmes permettant de résoudre les problèmes précédents.

Matériels et méthodes

I. Matériels :

Matériels biologique :

***Ovis musimon* (sauvage):**



Ovis aries* (domestique) *:

Mammifère

Herbivore

Femelle: Brebis
Male: Bélier

54 chromosomes

Famille : bovidés
Sous-famille: Caprinés
Genre: *Ovis*

Les études génétiques récentes pointent vers plusieurs phénomènes de domestications

Materiels informatiques :

1) Les puces à ADN :

Les puces à ADN des cellules musculaires des deux races ovines.
(Base de données sur internet open source).

2) Le logiciel R :

R est un système créé par Ross Ihaka et Robert Gentleman.il comporte de nombreuses fonctions pour les analyses statistiques et les graphiques ; celle-ci sont visualisées immédiatement dans une fenêtre propre et peuvent être exportées sous divers formats (jpg, png, bmp, ps, pdf, emf, pictex, xfig ; les formats disponibles peuvent dépendre du système d'exploitation).

Les résultats des analyses statistiques sont affichés à l'écran, certains résultats partiels (valeurs de P, coefficients de régression, résidus, . . .) peuvent être sauvées à part, exportées dans un fichier ou utilisées dans des analyses ultérieures (Emmanuel Paradis ; 2005).

3) MeV :

MeV (MultiExperiment Viewer ; http://www.tm4.org/mev.html) permet de faire l'analyse des données filtrées et normlisées. Il permet, entre autre, de faire la visualisation des hybridations et de leurs patrons d'expressions correspondantes. De nombreux algorithmes de 'clustering' (Bootstrapping, Jacknifing et K-means par exemple) sont disponibles pour identifier et travailler facilement avec des gènes d'intérêts.

Il est aussi possible d'ajouter des annotations personnelles ou publiques aux données d'expression à l'aide des fichiers EASE. MeV permet la mise en place et l'échange de protocole d'analyse sous forme de fichier facilement échangeable et utilisable.

Il intéressant de préciser que MeV accepte beaucoup de format de fichiers comme fichier d'entrée donc il n'est pas nécessaire, bien que souvent recommandé, de faire

la normalisation et standardisation des données avec MIDAS avant de faire l'analyse avec MEV.

Il accepte les fichiers en format TIGR MeV (*.mev), les fichiers délimitées par des tabulations (*.TDMS), les fichiers TIGR Array viewer (*.tav), les formats Affymetrix, Genepix et Agilent.

4) DAVID:

Datatabase for ***A***nnotation, ***v***isualization and ***I***ntegrated ***D***iscovery (Denis et al.; 2003. Hang and al.; 2009) regroupe un ensemble d'outils web destinés à l'annotation fonctionnelle d'ensemble de gènes à l'aide de sa propre banque de données, DAVID knowledgebase. Cette dernière intègre les identifiants de gènes ou de protéines de plusieurs espèces, ainsi que leur annotation, à partir d'une grande variété de banque de données publiques (NCBI, PIR, SWISS-PROT, GO, OMIM, Pub-Med, KEGG, BIOCARTA, AffyMetrix, TIGR, Pfam, BIND, MINT, DIP…).

Les outils fournis par DAVID analysent des listes de gènes fournies par l'utilisateur et sont disponibles à l'adresse (david.abcc.ncifcrf.gov/). Ils comprennent l'outil d'annotation fonctionnelle (analyse de l'enrichissement en catégories fonctionnelles, cartographie sur les voies métaboliques, résumé d'annotation sous forme de graphiques…), l'outil de classification fonctionnelle de gènes (regroupement gènes ayant une annotation fonctionnelle similaire et l'outil de conversion d'identifiants. Dans notre étude nous avons utilisé l'outil de classification fonctionnelle dans nos listes de gènes.

II. Méthodes :

a) Objectif de l'étude :

C'est l'étude des différences entre les deux races ovines domestique et sauvage dans les cellules musculaires, pour expliquer leurs effets sur la qualité de la viande, et pouvoir arriver à améliorer la meilleure race ovine dans l'intérêt du consommateur.

b) Présentation des données :

Notre travail a pris pour support des échantillons des cellules musculaires de deux races ovines *Ovis aries* et *Ovis musimon* .

Les cellules récoltées ont été conservées à une température de -80°C puis décongelées pour que la totalité des ARN soit extraite à l'aide d'un RNeasy Mini Kit (Qiagen) (annexe n°1) et mises immédiatement dans un conservateur d'ARN ethanol (Qiagen). Elle a été par la suite rétrotranscrite en ADNc en utilisant SuperScript III Reverse Transcriptase (Invitrogen). La purification de l'ADNc synthétisé s'est faite avec Qiagen PCR Purification Kit et fragmenté par une DNase I (Promega). L'ADNc fragmenté a été marqué à son extrémité 3' par un réactif de marquage (Affymetrix) et la terminal deoxynucleotidyl transferase (Promega).

L'ADNc marqué a été hybridé avec des sondes sur une puce Affymetrix et coloré sur GeneChip Fluidics Station 450 (figure n°9). Les sondes ont finalement été scannées avec GeneChip Scanner 3000(figure n9). Les données brutes (fichiers de format .CEL) ont été obtenues avec Affymetrix GeneChip Operating System 1.4 puis prétraitées par Bioconductor en utilisant la méthode ***RMA***(annexe n°3pour les protocoles des kits utilisés).

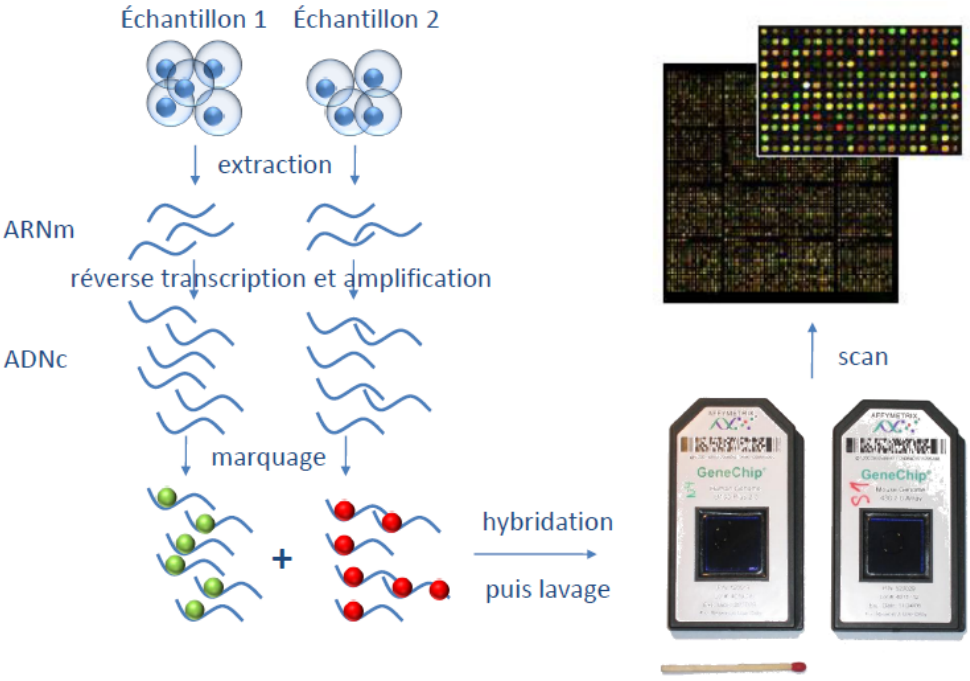


Figure 5: acquisition des données

c) Prétraitement et normalisation des données avec le logiciel R :

La première étape de notre étude s'est faite avec le logiciel R ; à commencer par la visualisation des données brutes avec la fonction ***arrayQualityMetrics()***, puis un prétraitement des données par la réalisation d'une normalisation avec la fonction ***RMA ()*** de la librairie ***(Affy)***.

On a ensuite réalisé avec ces données un tableau -avec les références d'échantillons comme noms de colonnes et les identifiants de gènes comme noms de lignes- sur lequel on a pu calculer la moyenne d'expression (M) de chaque gène à part dans tous les échantillons avec la fonction ***mean()***. Le tableau contenait 24128 gènes (lignes) et 9 colonnes (3 prises × 3 réplicas) (figure n°6).

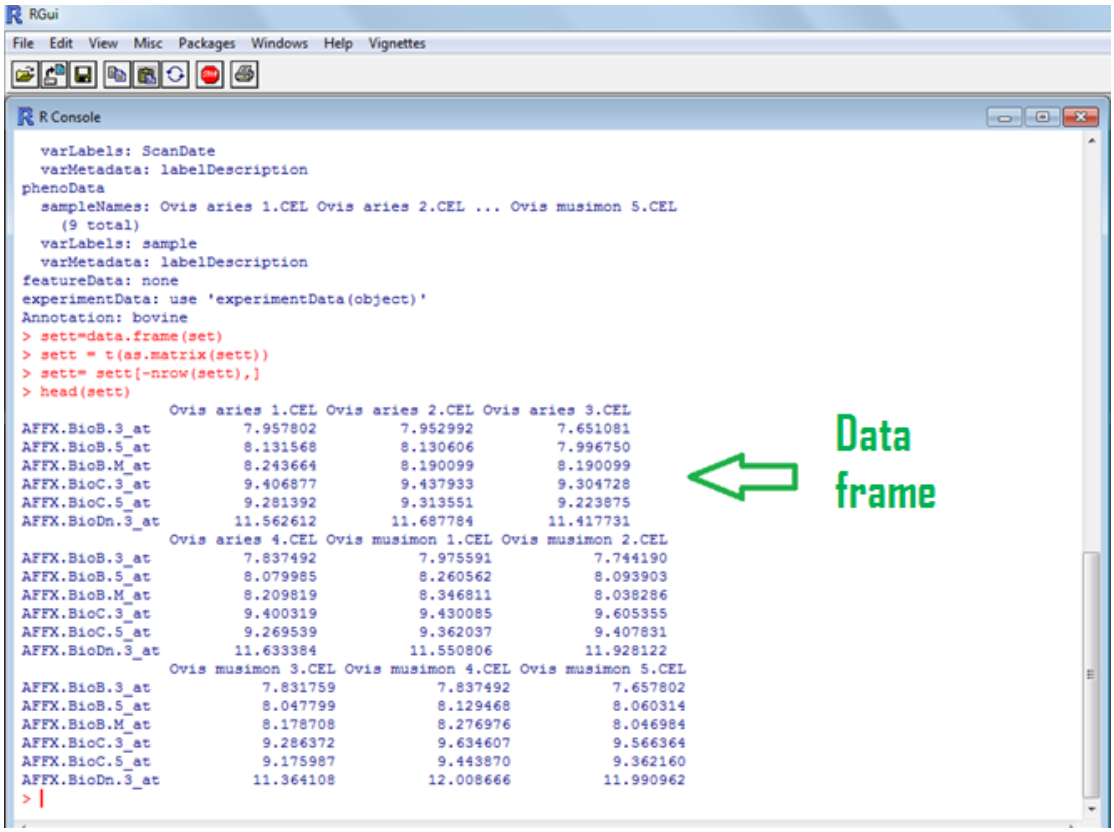


Figure 6: Tableau des expressions génique obtenues par R.

L'étape suivante était le calcul des taux d'expression relative pour chaque gène. Les taux d'expression relative se calcule en divisant -pour chaque gène- une à une ses valeurs d'expression par leur moyenne générale (M). Cette fonction nous permet d'ajuster les valeurs sur une courbe Gaussienne et de ce fait repérer les valeurs hors de l'intervalle de confiance, ce qui représente une expression génique réprimée ou activée. Les résultats on été traduits en boxplot grâce à la fonction ***boxplot()***.

Le deuxième volet visait une comparaison entre les différentes expressions dans tous les échantillons et pour tous les gènes. Pour réaliser celui-ci, il a fallu faire appel à la librairie ***(limma)*** qui contient des fonctions pouvant détecter les gènes différentiellement exprimés.

Quatre étapes précédaient la comparaison : (*i*) établissement d'une matrice contenant le nombre d'échantillons et le nombre de réplicas pour chacun (ex : 1,1,1,2,2,2,3,3,3 ; ce qui veut dire 3 échantillons ou prises avec 3 réplicas à chaque prise), (*ii*) nommer chaque échantillon (ex : A, B,C...), (*iii*) préciser toutes les combinaisons de comparaison possibles entre les échantillons (ex :A-B, A-C, B-C), (*iiii*) réaliser un alignement des données avec la fonction ***ImFit()*** pour enfin établir les comparaisons une à une par le biais de la fonction ***eBayes()***.

Une fois la comparaison établie, on a organisé les gènes selon un ordre croissant de P-value, et donc les gènes en haut du classement sont ceux qui ont les meilleures probabilités d'être différentiellement exprimés (annexe n°). Se basant sur ce raisonnement, on a sélectionné les 2000 gènes les mieux classés dont on a récupéré les valeurs d'expression avec lesquelles on va continuer notre étude avec le logiciel TMeV.

d) Clustering et identification des signatures spécifiques de l'activation et la répression :

L'utilisation du logiciel a nécessité le chargement du tableau issu du prétraitement avec R. en lui fournissant ces données sous format de fichier.txt MeV a pu réaliser, avec sa fonction Hierarchical Clustering (figure n°), un clustering des 1000 gènes en utilisant le paramètre Pearson correlation pour la corrélation (un coefficient de corrélation dont les données sont transformées en rang, ce qui convient parfaitement à une étude de valeurs correspondant à une évolution).

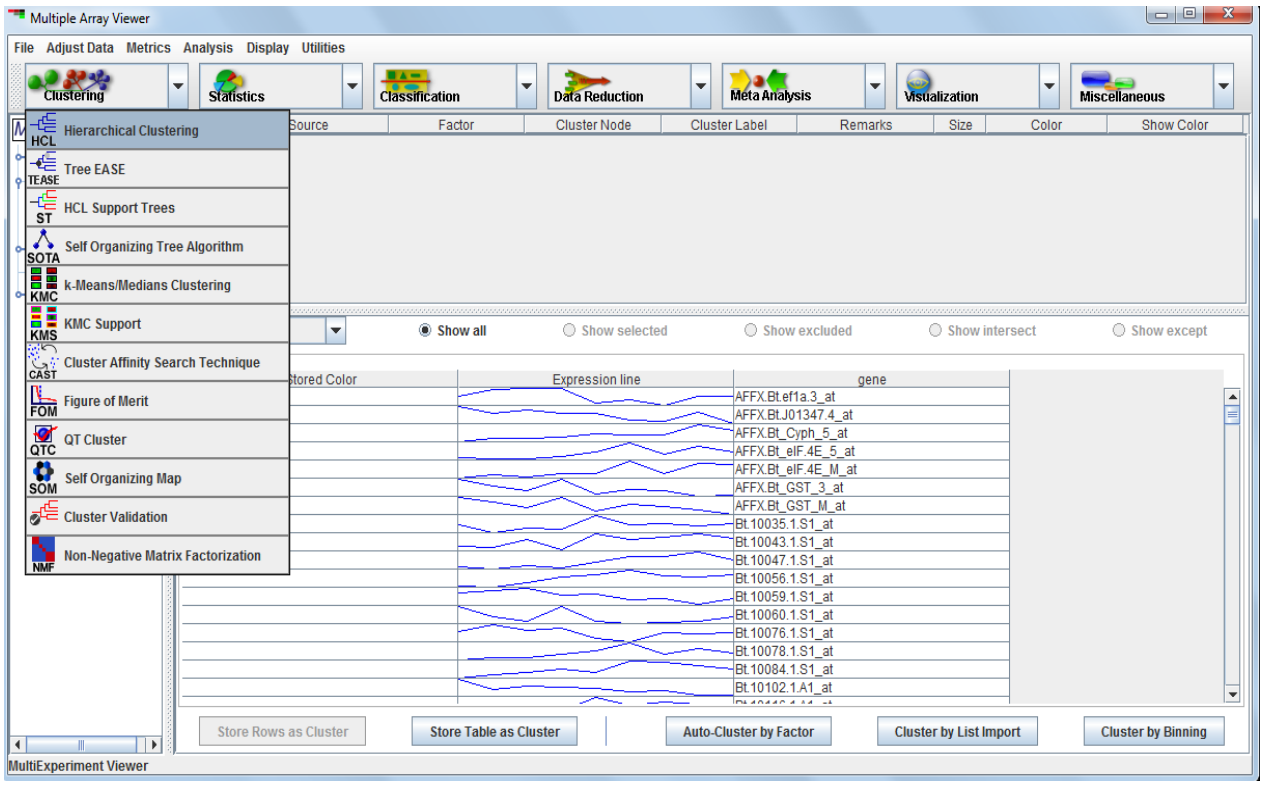


Figure 7 : Interface du logiciel MultiExperiment Viewer.

La deuxième utilisation de MeV était pour diviser les 1000 gènes en deux clusters (groupes) : Up et Down, ce qui veut dire respectivement activés et réprimés. Pour ce, MeV met à la disposition de l'utilisateur une fonction nommée significance Analyses microarrays SAM .

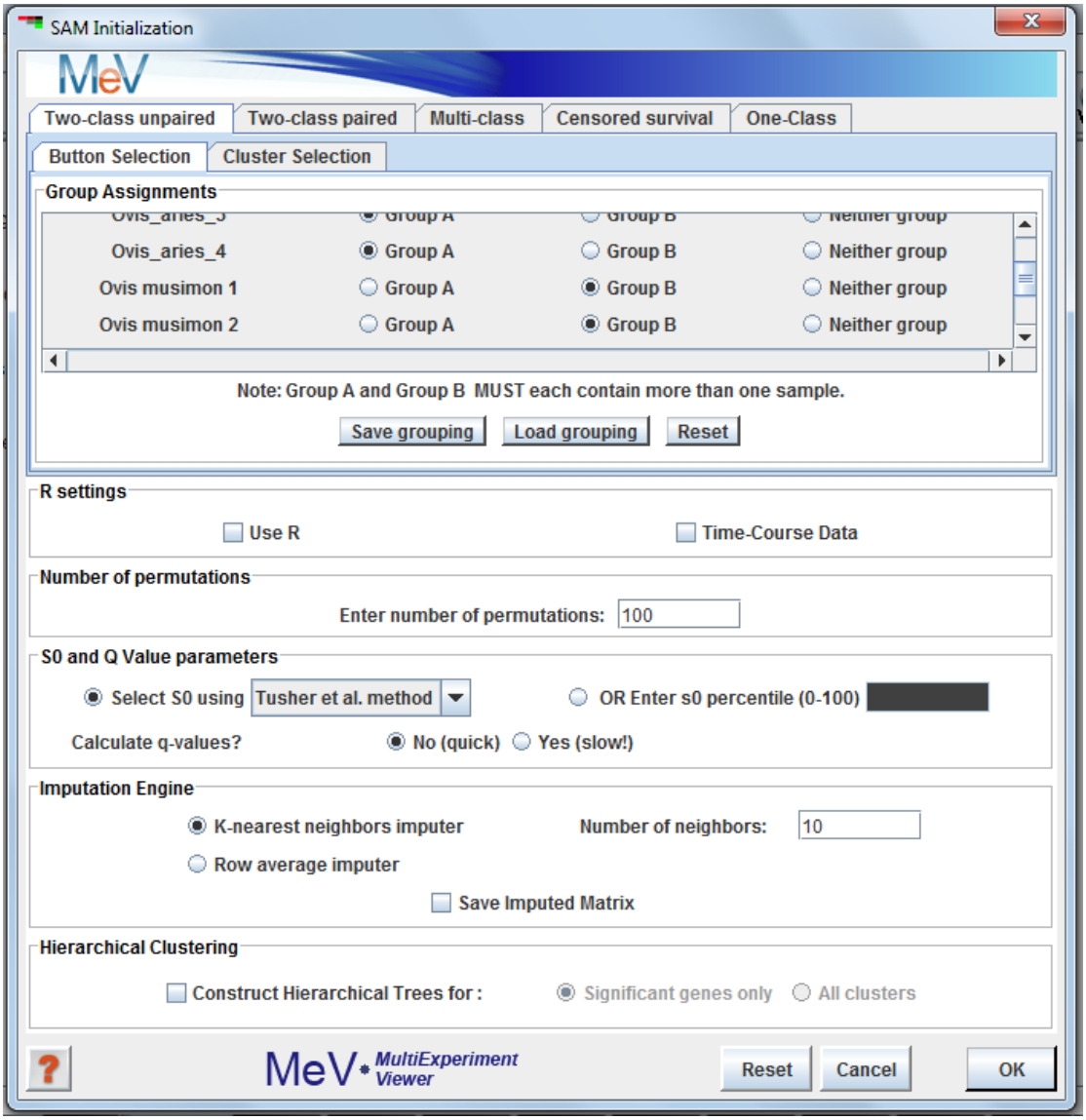


Figure 8 : Initiation de SAM.

Après isolement des deux clusters, MeV fourni les listes de gènes pour les deux groupes, qu'on a gardé pour une utilisation ultérieure avec DAVID. La liste des gènes réprimés contenait 552 gènes, alors la deuxième contenait 298 gènes sur les 1000 gènes présélectionnés.

e) Annotation fonctionnelle et définitions des voies de signalisations avec DAVID :

L'utilisation des outils bioinformatiques de la base de données DAVID est possible grâce à une interface web ouverte au public. Pour bénéficier des services d'annotation fonctionnelle proposés par le site, nous avons dû soumettre, dans l'espace alloué à ce fait, l'une après l'autre, les deux listes de gènes qu'on avait obtenus au préalable après clustering à l'aide de MeV.

Il a fallu aussi paramétrer l'analyse en précisant le type d'identifiant de nos gènes qui était « AFFYMETRIX_3PRIME_IVT_ID », puis identifier nos listes comme étant « gene list »(figure n°9). La deuxième étape consistait en une sélection d'un outil d'analyse parmi la panoplie proposée dans le site (figure n°10).

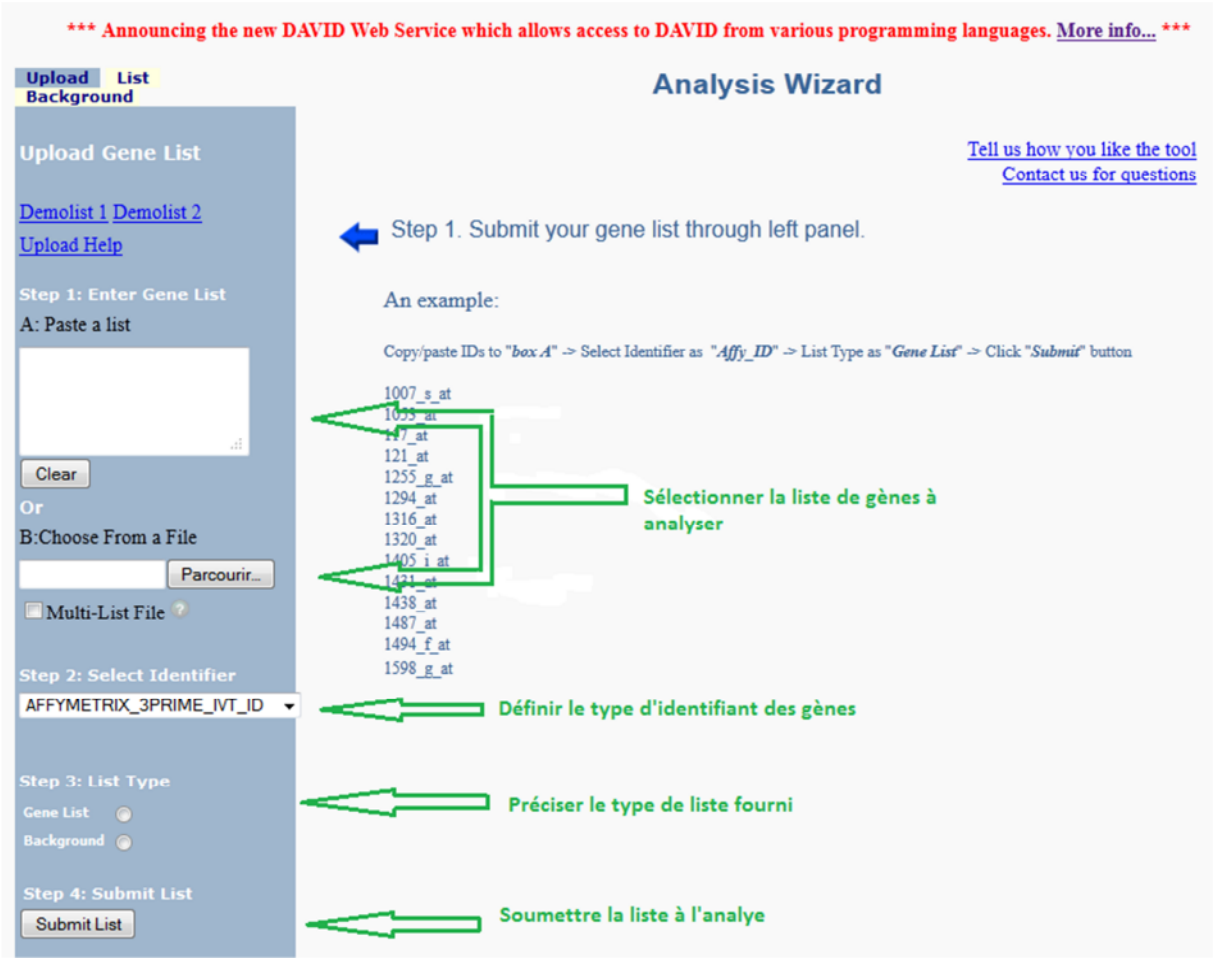


Figure 9 : Première étape avec DAVID : chargement de la liste à analyser.

Figure 10 : Deuxième étape de DAVID : le choix de l'outil d'analyse.

L'analyse nous a fourni pour les deux listes, un ensemble de voies de signalisation avec les P-values qui leurs sont assignées ainsi que le nombre de gènes inclus dans chacune des voies sous forme de tableaux (figure 11 ; 12).

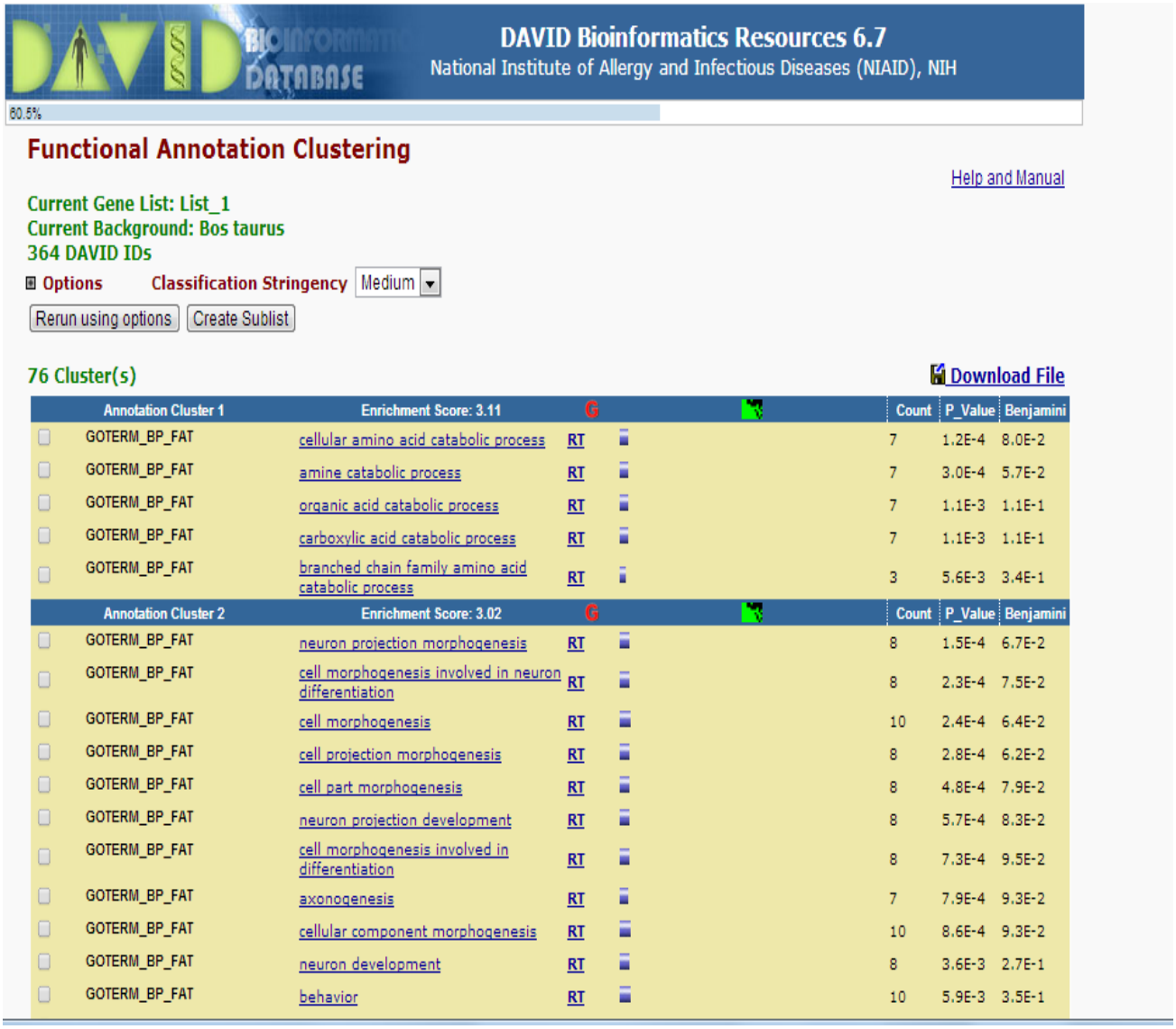
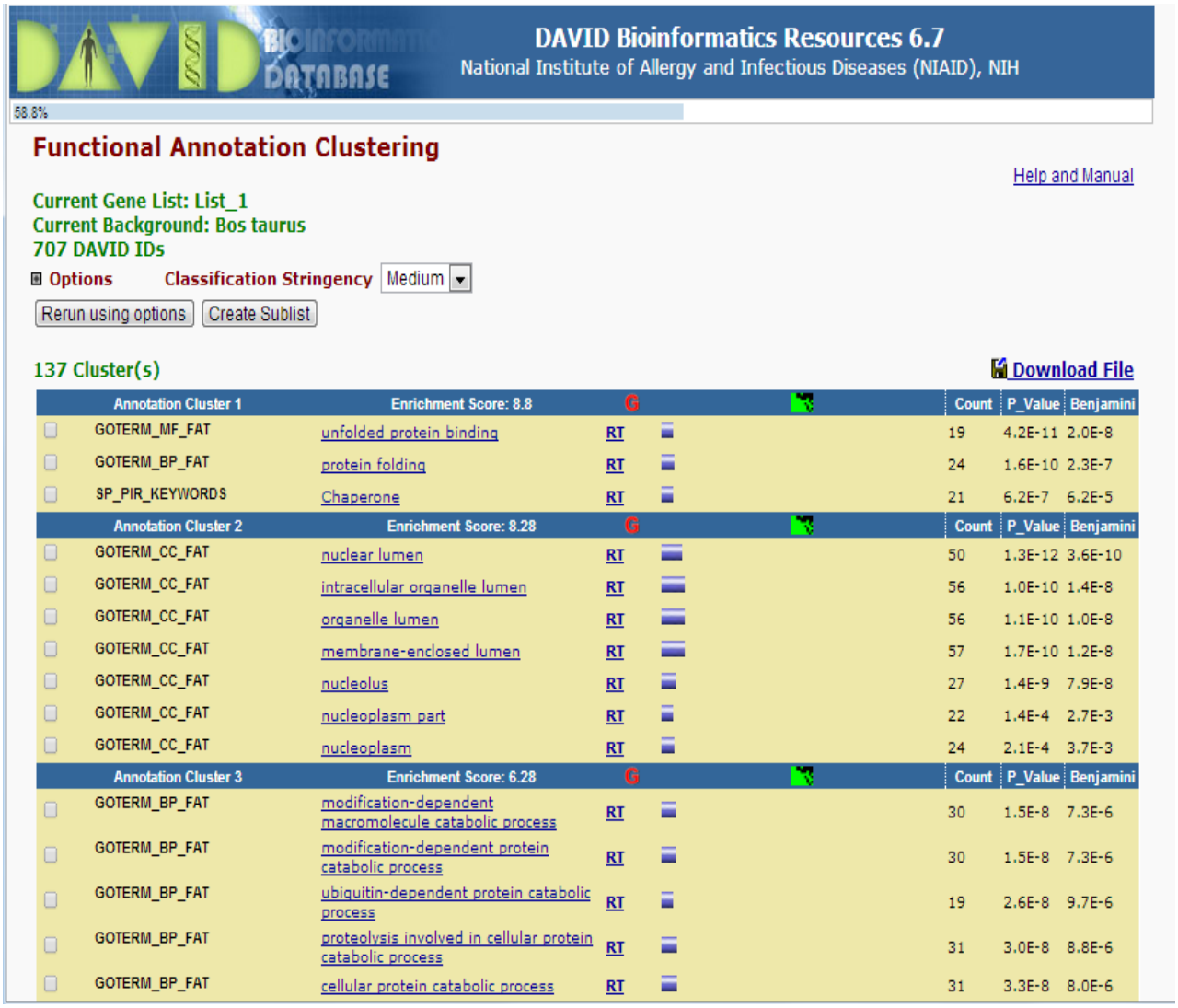


Figure 11 : Voies de signalisation régulées à la baisse par DAVID.



DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

58.8%

Functional Annotation Clustering

Help and Manual

Current Gene List: List_1
Current Background: Bos taurus
707 DAVID IDs

Options Classification Stringency: Medium

Rerun using options | Create Sublist

137 Cluster(s) Download File

Annotation Cluster 1	Enrichment Score: 8.8				Count	P_Value	Benjamini
GOTERM_MF_FAT	unfolded protein binding	RT			19	4.2E-11	2.0E-8
GOTERM_BP_FAT	protein folding	RT			24	1.6E-10	2.3E-7
SP_PIR_KEYWORDS	Chaperone	RT			21	6.2E-7	6.2E-5
Annotation Cluster 2	Enrichment Score: 8.28				Count	P_Value	Benjamini
GOTERM_CC_FAT	nuclear lumen	RT			50	1.3E-12	3.6E-10
GOTERM_CC_FAT	intracellular organelle lumen	RT			56	1.0E-10	1.4E-8
GOTERM_CC_FAT	organelle lumen	RT			56	1.1E-10	1.0E-8
GOTERM_CC_FAT	membrane-enclosed lumen	RT			57	1.7E-10	1.2E-8
GOTERM_CC_FAT	nucleolus	RT			27	1.4E-9	7.9E-8
GOTERM_CC_FAT	nucleoplasm part	RT			22	1.4E-4	2.7E-3
GOTERM_CC_FAT	nucleoplasm	RT			24	2.1E-4	3.7E-3
Annotation Cluster 3	Enrichment Score: 6.28				Count	P_Value	Benjamini
GOTERM_BP_FAT	modification-dependent macromolecule catabolic process	RT			30	1.5E-8	7.3E-6
GOTERM_BP_FAT	modification-dependent protein catabolic process	RT			30	1.5E-8	7.3E-6
GOTERM_BP_FAT	ubiquitin-dependent protein catabolic process	RT			19	2.6E-8	9.7E-6
GOTERM_BP_FAT	proteolysis involved in cellular protein catabolic process	RT			31	3.0E-8	8.8E-6
GOTERM_BP_FAT	cellular protein catabolic process	RT			31	3.3E-8	8.0E-6

Figure 12 : Voies de signalisation régulées à la hausse par DAVID.

Par la suite on a eu besoin d'un outil qui est le Gene ID conversion, qui nous a permis de convertir les deux listes de gènes positive et négative en choisissant le type de ciblage génique Identifier

Figure 13 : Gene ID conversion

Une autre information que DAVID peut nous conférer c'est les gènes de l'ontologie pour les deux listes, on a trois groupes de ces derniers :

Gènes BP ------ processus biologique

Gènes CC------ composition cellulaire

Gènes MF------ fonction moléculaire

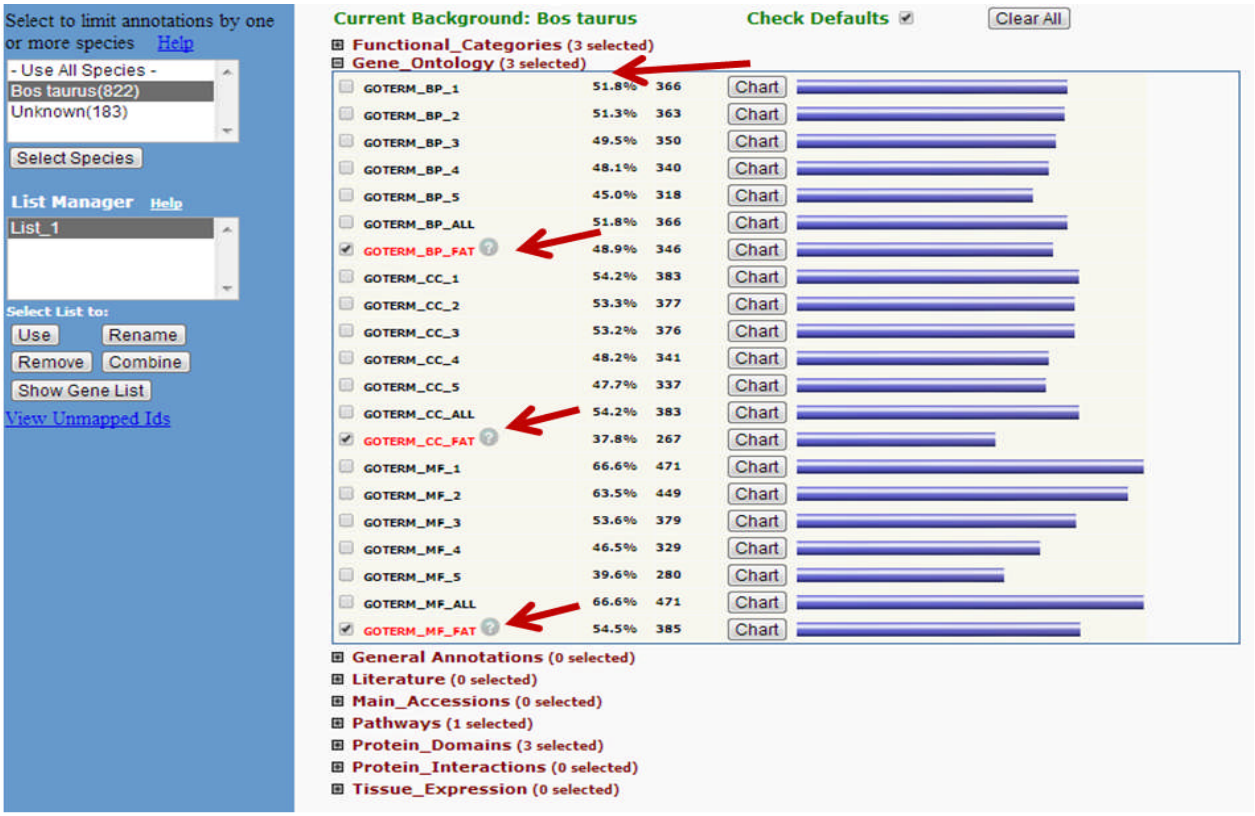


Figure 14 : Gènes d'ontologies

Nb : Dans notre étude on a essayé de se baser sur les gènes de processus biologique.

Résultats

(A) Les résultats de R :

R est un langage de développement bioinformatique et statistique contenant un nombre très important de fonctions mathématiques et statistiques. Ces fonctions sont contenues dans des packages qui, eux même, font partie des librairies. Elles sont mises à disposition sous forme d'un logiciel et utilisables par le biais d'une console. Cet outil nous a permis de visualiser –avant tout traitement- les données de nos puces à ADN qu'il a traduit en boxplots .

La deuxième fonction que nous avons utilisée était la fonction ***RMA()***, C'est une fonction de normalisation qui nous a permis de faire un ajustement des données afin de minimiser les écarts non significatifs (erreurs de manipulation, bruits de fonds…). Une deuxième visualisation des données en boxplots matérialise l'efficacité de cette fonction (figure n15).

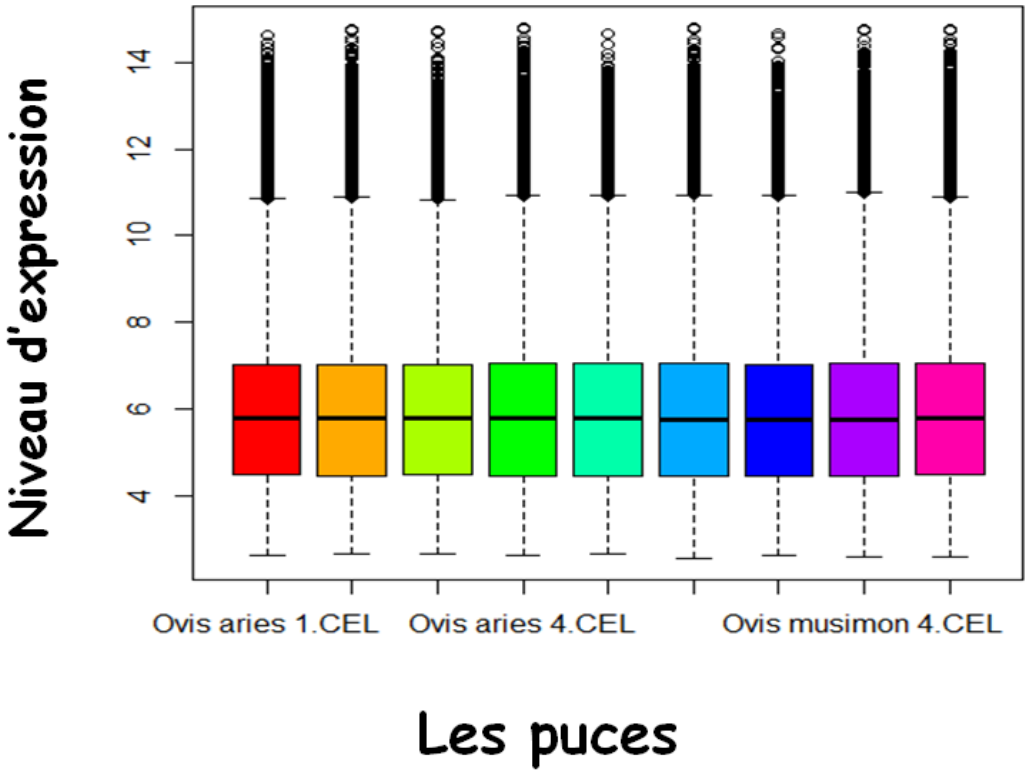


Figure 15: Boxplots des données après la normalisation RMA.

Les statistiques sont le fondement de chaque étude bioinformatique. La figure suivante est issue de la transformation des données en une courbe en cloche grâce au calcul des expressions relatives.

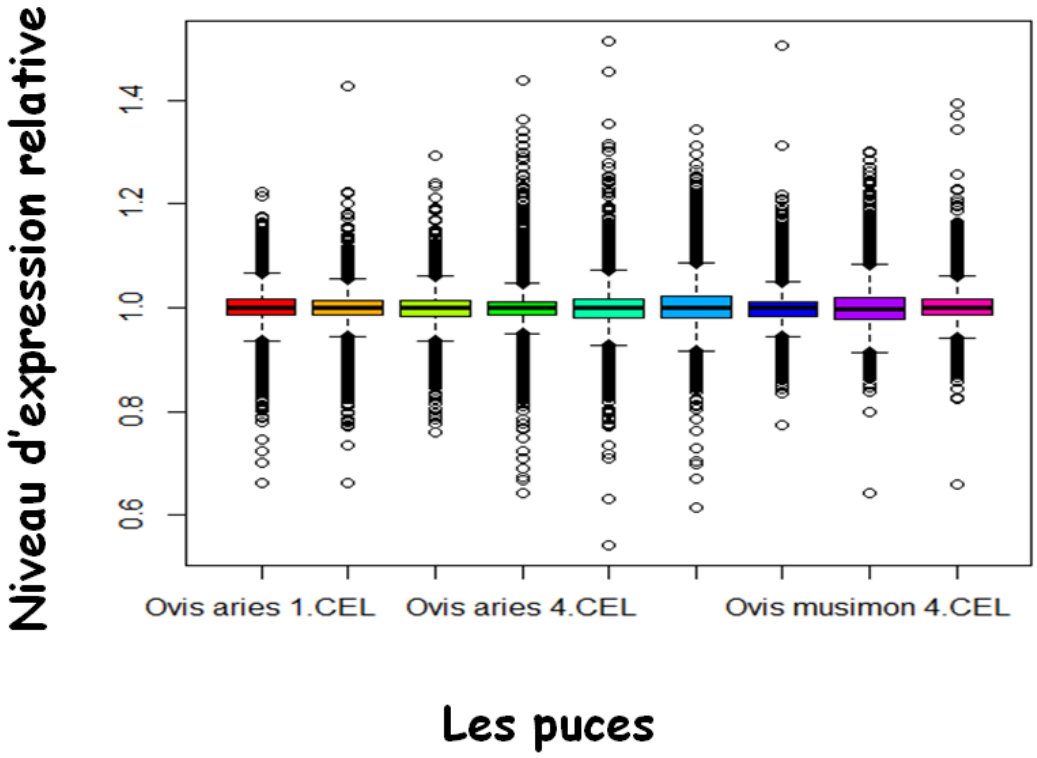


Figure 16 :Boxplots des expressions relatives.

(B) Les résultats de Mev :

✓ Clustering :

Les résultats de TMev présentés par la figure 17 regroupent des cellules en cluster selon leur profil de transcription. Les dendrogrammes, sur le dessus, indiquent les relations entre les deux races ovines. Le taux d'expression est indiqué par l'échelle de couleurs, avec rouge pour les plus élevé, vert pour les faibles, et noir pour le signal de transcription intermédiaire.

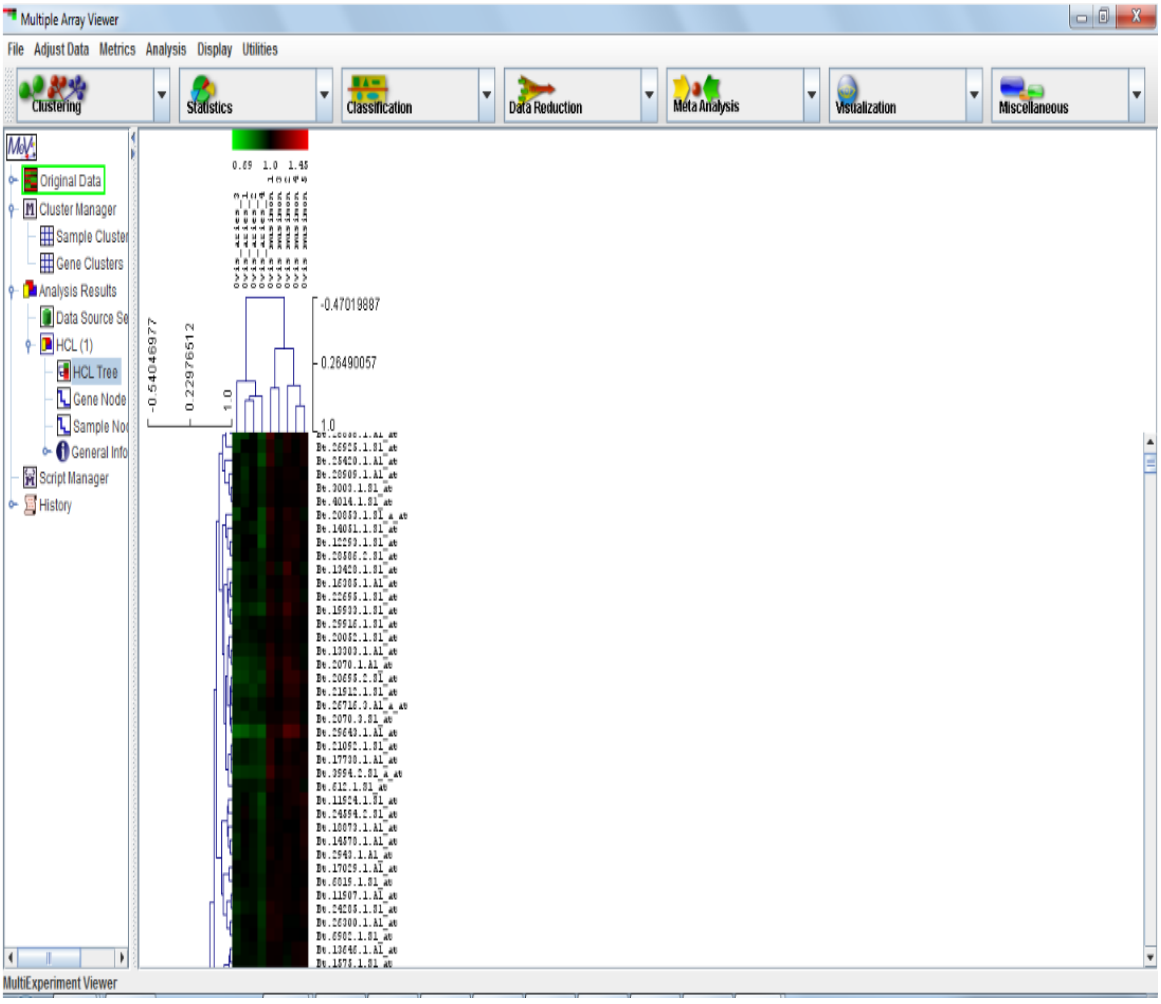


Figure 17 : clustering

✓ SAM :

Le graphe coloré en rouge indique les gènes actifs les plus exprimés d'Ovis aries par rapport aux Ovis musimon par contre le graphe coloré en vert indique les gènes inactifs d'Ovis aries par rapport aux Ovis musimon.

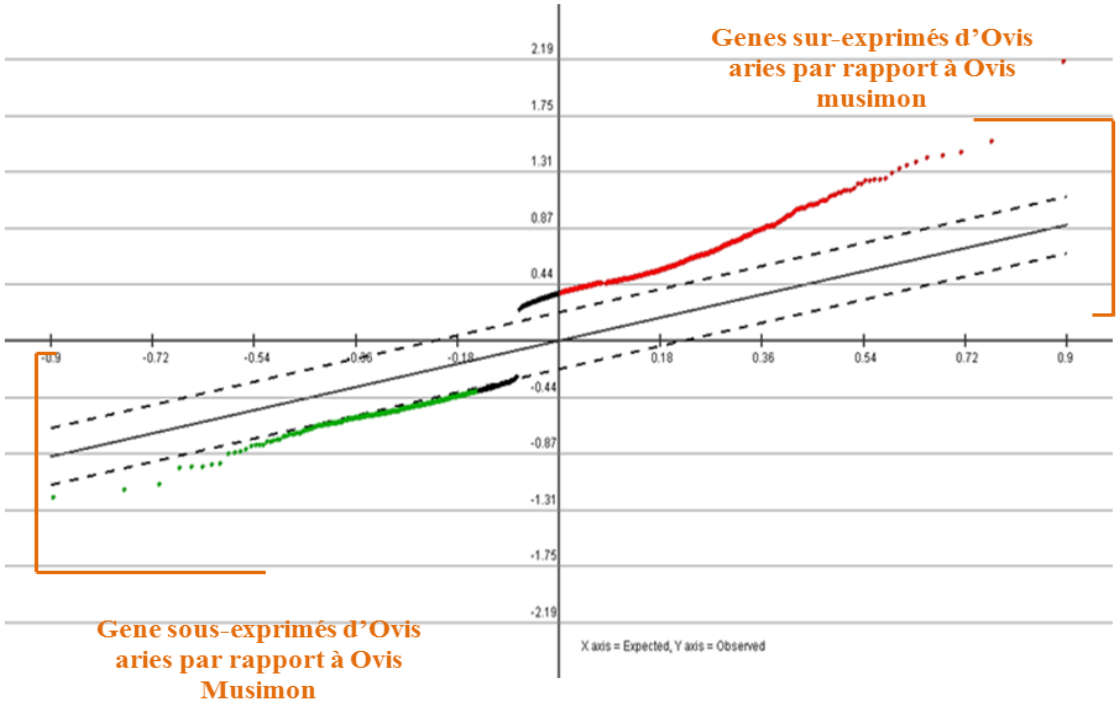


Figure 18 : graphe de Sam

(C) Les résultats de DAVID :

Après chargement de la liste des gènes ainsi que la définition des paramètres, DAVID nous a proposé un certain nombre de voies de signalisation caractérisées à chaque fois par un P-value et le nombre de gènes qui y sont impliqués.

Parmi les voies proposées, nous avons sélectionné les plus pertinentes et ayant les meilleurs P-values que nous avons mis dans le tableau suivant :

Type de Régulation	Processus	Nb de gènes	Exemple de gènes	P-value
Up	**Métabolisme oxydative**	30	**Mdh2** : Malate Dehydrogenase 2	1.5E -8
Up	**Métabolisme de la mitochondrie**	19	**ABCB6** : ATP-Binding cassette sub-familyB	2.6E -8
Up	**Anti-apoptotique**	15	**Nol3** : Nuclear protein3	3.0E -8
Down	**Régulation de croissance**	8	**PA2GA** : phospholipase A2 membrane associated	2.3E -4
Down	**Cycle cellulaire**	10	**PCNP** : Proteolytic Signal Containing Nuclear Protein	2.4E -4

Table 3 : Les voies de signalisation différentiellement exprimées.

Up : régulation à la hausse ; Down : régulation à la baisse

Discussion

Les recherches actuelles visent principalement à mettre en évidence les relations génétiques entre caractères de production et qualités sensorielles.

Vouloir améliorer la qualité de la viande, via la génétique, dépend de notre capacité à discerner dans la variabilité de la qualité la part qui est effectivement d'origine génétique et utilisable par la sélection pour être cumulée au cours des générations.

Cette discrimination sera d'autant moins hasardeuse qu'on aura réduit la part de la variabilité non génétique en standardisant au mieux les conditions de milieu *ante* et *post mortem.*

Toutes choses étant égales par ailleurs (conduite, abattage, traitements, etc), les différences entre animaux expriment, alors, ce qui est communément appelé la variabilité individuelle, et que les généticiens appellent la variabilité phénotypique. Heureusement, les progrès rapides de projets de séquençage de l'ADN ont fait que le génome de la plupart des races ovines soit aujourd'hui connu. Ceci a fait de la technique des puces à ADN un outil d'investigation privilégié par les chercheurs. Cependant, la façon précise d'analyser les données des biopuces et d'en extraire les informations utiles reste un obstacle à surmonter.

L'objet de notre travail était de faire une comparaison génomique entre deux races ovines : l'une domestique ***Ovis aries*** et l'autre sauvage ***Ovis musimon*** dont l'intérêt de trouver les gènes responsables à la sélection entre les deux races au niveau des cellules musculaires, ce qui nous aident, par la suite, à pouvoir améliorer la qualité de viande pour satisfaire le consommateur.

Pour ce qui concerne la qualité de la viande, nos connaissances sur la variabilité individuelle et son déterminisme génétique sont encore limitées. Mis à part le cas des productions sous label, la principale raison réside dans le manque de lisibilité et donc de quantification de la demande des consommateurs.

En l'absence de réelle plus-value économique pour les viandes de qualité supérieure, aucune mesure en abattoir de la qualité n'a donc été développée. De plus, la difficulté d'obtenir des enregistrements, même indirects, de cette qualité limite les possibilités de mettre en place des programmes expérimentaux pour entreprendre des études génétiques de dimension suffisante.

Parmi les critères les plus importants dans la qualité de viande on cite : la tendreté, la jutosité et la flaveur.

Les chercheurs ont identifié un gène de dureté de la viande, DnaJa1, dont l'expression, à elle seule, explique 60% des écarts de qualité sensorielle des viandes observées entre des lots différents.
«L'expression de ce gène constitue donc un bon candidat pour être un marqueur négatif de tendreté en race Charolaise. Ces résultats ont fait l'objet d'un dépôt de brevet», JF Hochette et al,1998.

En effet, les caractéristiques biologiques étudiées jusqu'à présent telles que la teneur en collagène ou en lipides intramusculaires, n'expliquent pas plus d'un tiers de la variabilité de la tendreté de la viande bovine.
Identifier les gènes responsables soulève plusieurs enjeux : les intégrer dans l'avenir au sein des schémas de sélection et mettre au point pour la filière viande un test identifiant les bovins à fort potentiel pour produire une viande tendre.

En utilisant des nouvelles techniques de génomique, les chercheurs de l'Inra ont identifié 112 gènes associés aux différents critères de qualité sensorielle. Cependant, tous ne sont pas impliqués à des mêmes degrés dans la tendreté, la jutosité, la flaveur.

Parmi les gènes révélés, le gène Prkag1, dont l'expression favorise flaveur et jutosité. Un code pour une protéine est impliquée dans le métabolisme des acides gras et du glucose.

Cette étude a permis également d'identifier 58 gènes impliqués dans des différences de tendreté, jutosité, flaveur. Certains ont beaucoup d'influence. Ainsi, 18

gènes expliquent la moitié des différences observées à la dégustation de la viande pour les critères de jutosité et flaveur.

Les analyses ont permis de mettre en évidence dans les lots de viandes tendres la présence de protéines de type lent oxydatif, confirmant les données précédentes de protéomique, contredisant en apparence l'observation selon laquelle le type rapide et glycolytique favorise la tendreté.

En fait, il existe plusieurs catégories de fibres ; et les chercheurs pensent que *« le pourcentage de fibres rapides oxydo-glycolytiques (type IIA) soit un facteur défavorable à la tendreté de la viande, et non pas les proportions de fibres lentes oxydatives (type I) ou rapides glycolytiques (type IIB). »*

Dans cette étude, on a essayé de faire une comparaison entre le génome d'Ovis aries et d'Ovis musimon au niveau des s musculaires en se concentrant surtout au processus biologiques et les voies de signalisation.

Pour l'accomplissement de ce travail, notre fusil d'épaule a été, d'abord, les puces à ADN. Les puces à ADN sont des multicapteurs permettant de caractériser et quantifier un acide nucléique dans un échantillon.

En second lieu, les outils bioinformatiques tel que le logiciel R, le logiciel Tmev (*TranscriptomeMultiExperimentViewer*) et DAVID (Database for Annotation, Visualization and IntegratedDiscovery) nous ont permis tour à tour de :

- *(i)* prétraiter et de normaliser les données brutes issues des puces à ADN (sous format de fichiers .CEL),
- *(ii)* puis isoler les gènes différentiellement exprimés et enfin
- *(iii)* l'annotation fonctionnelle nous a permis de sélectionner les voies de signalisation les plus pertinentes.

Nos résultats montrent que les gènes impliqués dans les processus du catabolisme, métabolisme oxydative, anti-appoptotique et métabolisme de la mitochondrie sont surexprimés chez ***Ovis aries*** par rapport à ***Ovis musimon***.(Tableau n :)

En effet, selon Renand *et al* 2001, les animaux, dont le métabolisme du muscle long dorsal est plus oxydatif, ont tendance à produire une viande qui mature moins vite et qui est donc plus dure.

Tandis que les gènes impliqués dans la régulation de croissance et la régulation du cycle cellulaire sont moins exprimés chez ***Ovis aries*** par rapport à ***Ovis musimon***.

(Les relations génétiques sont peu marquées avec la croissance en vif, mais nettement plus avec la composition du croît. Au vu des corrélations génétiques, une sélection, pour accroître la masse musculaire aux dépens des dépôts adipeux, devrait se traduire par une réduction des teneurs en lipides intramusculaires et en pigments, du diamètre des fibres musculaires et par une augmentation du pH et de la solubilité du collagène.
Les relations avec le type de fibre sont nettement moins marquées.
Quelles conséquences peuvent avoir ces modifications attendues des caractéristiques musculaires sur les qualités de la viande, sachant que seules les qualités organoleptiques sont concernées ?

Nos résultats montrent aussi l'expression de deux gènes déjà cités auparavant et qui ont été découverts dans des recherches récentes ; ils ont un effet important sur la dureté de la viande et la flaveur en jutosité : Dnaja1, Prkag1 chez la race d'Ovis musimon .

Les résultats de cette étude montrent que la Race sauvage est mieux que la race domestique vis-à-vis la qualité de la viande ; mais ceci reste qu'une hypothèse, puisque malheureusement, bien qu'il existe une littérature bien documentée en Amérique du Nord et plus récemment en Australie sur le déterminisme génétique des qualités de la viande conjointement aux caractères de production, il est difficile d'extrapoler leurs résultats car les qualités de la viande, tout comme leur variabilité et leurs corrélations, dépendent très fortement des conditions de production (sexe, âge, vitesse d'engraissement), de transformation (abattage, refroidissement, maturation) et de consommation (cuisson).

Conclusion

Cette nouvelle technologie à haut débit nous a permis d'exploiter le génome de deux espèces et d'essayer de découvrir de nouveaux gènes qui contribuent à l'amélioration génétique de la qualité de viande ; même si cette comparaison n'a pas donné une conclusion précise afin de pouvoir améliorer la qualité de la viande d'une race par rapport à une autre.

Rien n'empêche que l'espoir réside dans la recherche de gènes marqueurs utilisables pour une sélection directe. Il est donc nécessaire de trouver un ou des gènes responsables ou des marqueurs très proches pour exploiter le déséquilibre de liaison au sein des populations élevées.

Toutefois, cette démarche se heurte à la pauvreté des résultats publiés dans le domaine public et à la difficulté d'obtenir des données phénotypiques pertinentes pour étudier finement des régions du génome ou tester d'éventuels gènes candidats mis en évidence dans d'autres études.

Références

- ANONYME, 2007. Le marche des produits laitiers, carnes et avicoles en 2007. OVINS/MONDE. http://www.office elevage.fr/publications/marche2007/Ovins/Ovresume. PDF

- Benaissa, 2001. Mémoires à la dérive, publié aux Éditions Lansman en 2001

- Budiansky, Stephen. New York : H. Holt, c2000. Includes bibliographical references (p. 97-98).

- Broder, S. & Venter, J. C. (2000) Whole genome: the foundation of new biology and medicine. Curr Opin Biotechnol 11, 581-585

- *BOUKADIDA Jawer ,DENIS Julien 2004*

- CHEMMAM, M., 2007. Variation de l'ingestion et des performances chez la brebis « Ouled Djellal » sur pâturage : effet de la saison et de la complémentation. Thèse doctorat (ANNABA) 167p.

- DUBRAY D,1988 Abondance structure et dynamique de la population de mouflons de corse (ovis ammon musimon) de structure EST du massif de Cialo (Haute /corse) et analyse du role de protection de la réserve de l'office national de la chasse d'Asco.Bull.école. 19 : 493-450.

- Dear, S. & Staden, R. *Nucleic Acids Res.* 19, 3907−3911 (1991)

- DeRisi et al ; 1997Exploring the metabolic and genetic control of gene expression on a genomic scale.

- Denis et al.; 2003. Hang and al.; 2009

- Emmanuel Paradis ; 2005.R for Beginners

- Eisen et Brown, Methods in Enzymology 1999) DNA arrays for analysis of gene expression. 303:179-205.

- François Bourguignon et Martin Browning et Pierre André Chiapori 2006 (Efficient Intra-household Allocations and Distribution Factors: Implications and Identification) CAM Papers . University of Copenhagen. Department of Economics. Centre for Applied Microeconometrics.

- Ministère de l'Agriculture et du Développement Rural, 2006

- O. CROCE ;2005. *Early On - The American Recordings 1993-1998*

- OMS 2003

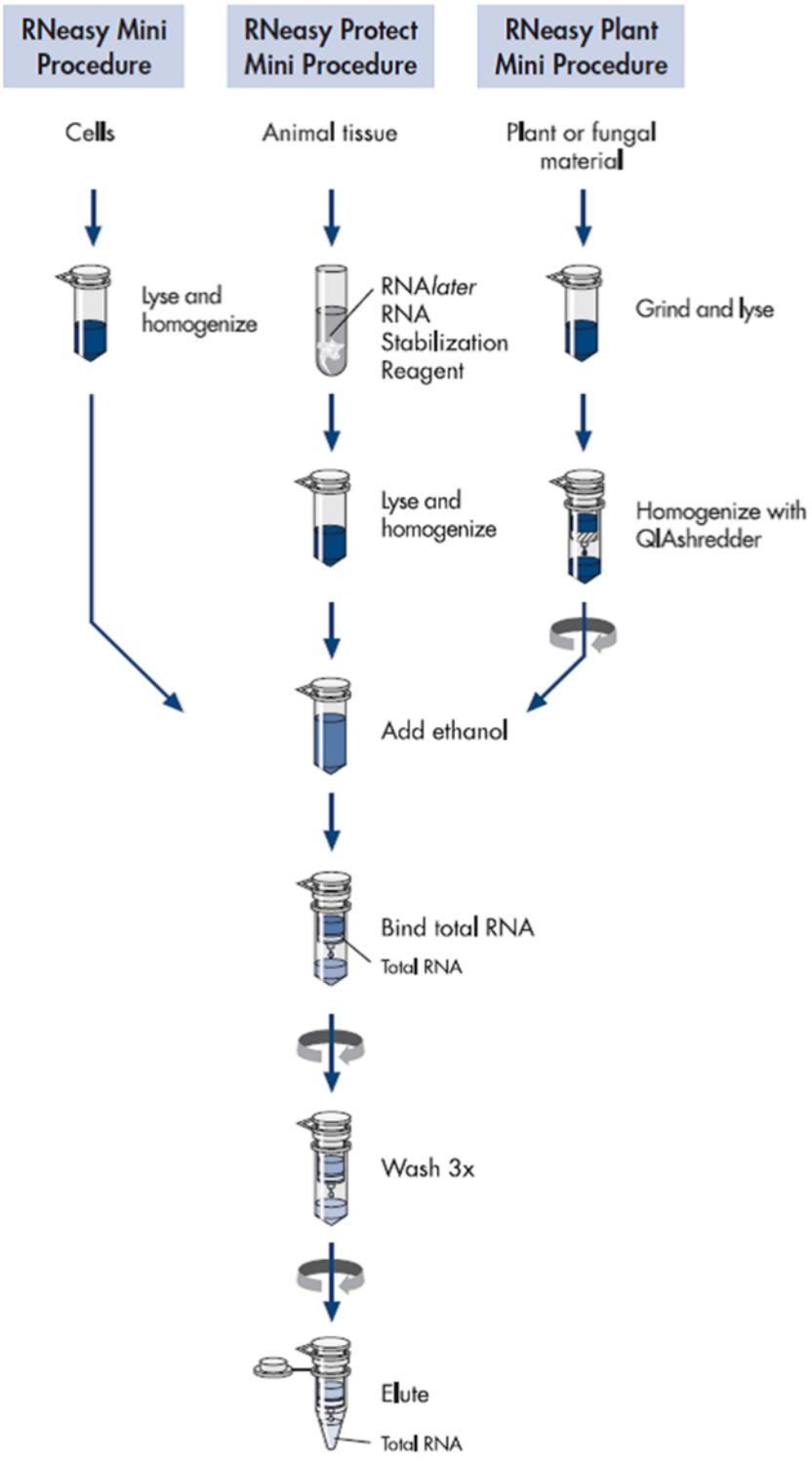
- Puces à ADN. Méthodes d'étude du génome et du transcriptome *J.-M. BIDART et L. LACROIX*.

- Zintilini C (2003). Identification des mycobactéries: Application de la biologie moléculaire. Conservatoire National des Arts et Métiers. Centre d'enseignement de Lyon.

Annexes

Annexe n°1 : Les protocoles des kits utilisés

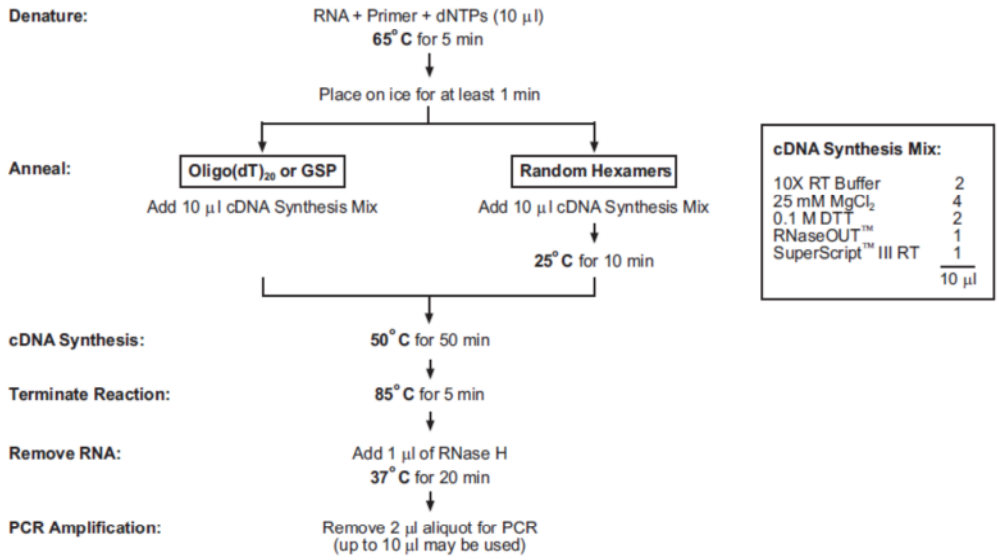
1- **RNeasy Mini Kit (Qiagen) :**



RNeasy Mini Kit . 2001–2010 QIAGEN.

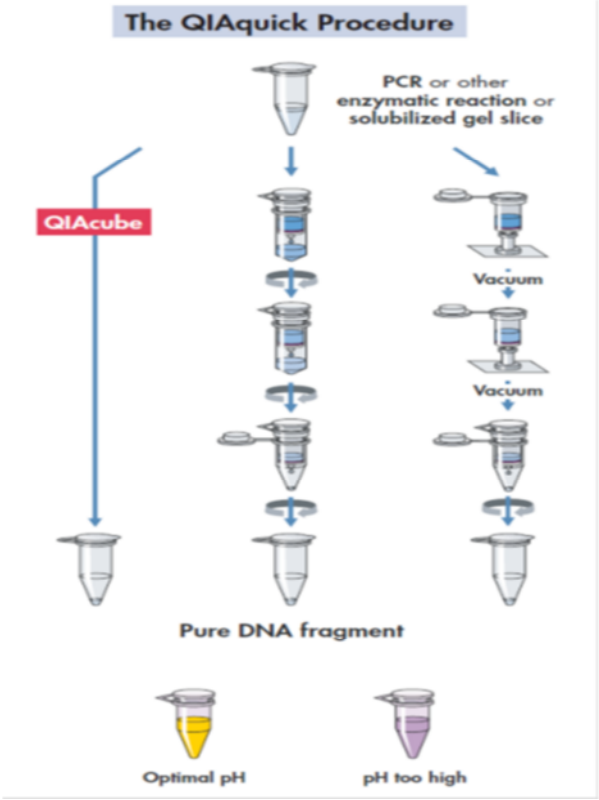
2- SuperScript III Reverse Transcriptase (Invitrogen)

Summary of Procedure



SuperScript™ III First-Strand Synthesis System for RT-PCR. 2003.

3- **Qiagen PCR Purification Kit :**



1. Add 5 volumes Buffer PB to 1 volume of the PCR reaction and mix. If the color ofthe mixture is orange or violet, add 10 μl 3 M sodium acetate, pH 5.0, and mix.The color of the mixture will turn yellow.

2. Place a QIAquick column in □ a provided 2 ml collection tube or into a vacuum manifold. For details on how to set up a vacuum manifold, refer to the *QIAquick Spin Handbook*.

3. To bind DNA, apply the sample to the QIAquick column and □ centrifuge for 30-60 s or □ apply vacuum to the manifold until all the samples have passed through the column. Discard flow-through and place the QIAquick column back in the same tube.

4. To wash, add 0.75 ml Buffer PE to the QIAquick column □ centrifuge For 30-60 s or □ apply vacuum.Discard flow-through and place the QIAquick column back in the same tube.

5. Centrifuge the QIAquick column once more in the provided 2 ml collection tube for 1 min to remove residual wash buffer.

6. Place each QIAquick column in a clean 1.5 ml microcentrifuge tube.

7. To elute DNA, add 50 μl Buffer EB (10 mM Tris·Cl, pH 8.5) QIAquick®
PCR Purification Kit. 2010.
or water (pH 7.0-8.5) to the center of the QIAquick membrane and centrifuge the column for 1 min. For increased DNA concentration, add 30 μl elution buffer to the center of the QIAquick membrane, let the column stand for 1min, and then centrifuge.

8. If the purified DNA is to be analyzed on a gel, add 1 volume of Loading Dye to
5 volumes of purified DNA. Mix the solution by pipetting up and down before loading the gel

4- **DNase I (Promega) :**

1. Set up the DNase digestion reaction as follows:

RNA in water or TE buffer	1-8μl
RQ1 RNase-Free DNase 10X Reaction Buffer	1μl
RQ1 RNase-Free DNase	1u/μg RNA
Nuclease-free water to a final volume of	10μl

Note: Use 1 unit of RQ1 RNase-Free DNase per microgram of RNA. For smaller amounts of RNA, use 1 unit of RQ1 RNase-Free DNase per reaction.

2. Incubate at 37° C for 30 minutes.

Note: If analyzing RNA samples by gel electrophoresis, perform a phenol:chloroform extraction and ethanol precipitation before loading the samples on the gel because salts in the RQ1 DNase Reaction Buffer and Stop Solution may cause aberrant migration or smearing of RNA on gels. Steps 3 and 4 may be omitted if a phenol:chloroform extraction is performed.

3. Add 1μl of RQ1 DNase Stop Solution to terminate the reaction.

4. Incubate at 65° C for 10 minutes to inactivate the DNase.

5. Add all, or a portion of, the treated RNA to the RT-PCR. See the Access RT-PCR System Technical Bulletin #TB220 (5).

5- **La terminal deoxynucleotidyl transferase (Promega) :**

A. Protocol

Materials to Be Supplied by the User :

- 0.2M EDTA
- 0.5M Na_2HPO_4 (pH 6.8)
- Whatman® DE-81 2.3cm circular filters

1. Dilute 1µl of the reaction mixture into 100µl of 0.2M EDTA. Spot 3µl of this solution onto each of four Whatman® DE-81 2.3cm circular filters.

2. Dry the filters briefly under a heat lamp. Set two filters aside for use in determining total cpm.

3.Wash the other two filters in 50ml of 0.5M Na_2HPO_4 (pH 6.8) twice for 5 minutes each to remove unincorporated nucleotides.

4. Dry the washed filters under a heat lamp.

5. Add the appropriate scintillation fluid to each filter and count in a scintillation counter.

B. Example of a Standard Calculation

$$\% \text{ incorporation} = \frac{\text{incorporated}}{\text{cpm total cpm}} \times 100$$

$$\text{Total cpm incorporated} = \text{incorporated cpm} \times \text{dilution factor*} \times \frac{\textbf{total reaction volume}}{\textbf{volume counted}}$$

$$\text{Average number of bases added to each primer} = \frac{\textbf{\% incorporation}}{\textbf{100}} \times \text{molar ratio of nucleotide to primer present in the reaction}$$

Amount of DNA synthesized = average number of bases added per primer × 330pg/pmol base × pmol primer present in reaction

$$\text{Specific activity of probe} = \frac{\textbf{total cpm incorporated}}{\textbf{µg of DNA template + µg DNA synthesized}}$$

Annexe n 2 : le script R

```
######################### start of script
setwd("C:/Users/MOHAMED/Desktop/Stage PFE Blida/saida/données/les puces")
          ### mettre le chemin de vos fichiers .CEL (à modifier selon ton ordinateur)
library(affy)                  ### charger la library affymetrix
df = ReadAffy()                ### lire tout les fichier .CEL
df                           ### voir il contient quoi ce affy batch

set =rma(df)                   ### réaliser une normalisation rapide avec la fonction
RMA
head(set)                    ### voir il correspond à quoi cet objet set

sett=data.frame(set)           ### transforme le affy batch en tableau (data frame)
sett = t(as.matrix(sett))      ### inversé le tableau (colonnes x lignes)
sett= sett[-nrow(sett),]       ### supprime la dernier ligne qui contient des
informations unitile
head(sett)                   ### voir il correspond à quoi cette data frame
dim(sett)                    ### voir la taille de cette data (nb gène x nb expérience)

boxplot(sett)                  ### voir si la normalisation à bien ramené les puces au
même niveau
boxplot(sett , col=rainbow(n=9))  ### même chose mais avec des couleur (option
"col")



sdd= apply(sett, MAR=1 , mean)     ### calculer la moyenne de l'expression de
chaque gène à part dans toutes les expériences
head(sdd,3)                    ### voir il correspond à quoi ces valeurs

settr= sett/sdd                ### diviser les taux d'expression de chaque gènes par la
moyenne pour avoir des taux relative d'epression

boxplot(settr , col=rainbow(n=9)) ### voir à nouveau la distribution mais cette fois
de l'expression relative

write.table(settr, file = "matrix.txt", sep = "\t", row.names = FALSE, col.names =
TRUE)
                                                ### enregister ce tableau dans votre
ordianteur sous le nom "matrix.txt" que on peut ouvrir avec XSL

library(limma)                 ### charger la library limma qui contient les fonction de
recherche des gènes différentiellement exprimés entre les expériences
design <- model.matrix(~ -1 + factor(c(1,1,1,1,2,2,2,2,2)))
                         ### crée la matrice de travaille on lui disant combient
d'expérience (6) et combient de réplicat pour chaque expérience (1,1,1,1 ...etc)

colnames(design) <- c("O","L")
                         ### nommer chaque expérience
```

```
contrast.matrix <- makeContrasts(O-L,levels=design)
                        ### demander a faire toutes les comparaison entre les différents
expériences (exemple entre O et T ...etc)

fit <- lmFit(settr, design)        ### réaliser un ajustement des comparaisons et un
alignement des données

bay= eBayes(contrasts.fit(fit,contrast.matrix))
                        ### réalisation des comparaisons une à une
head(bay)                    ### voir le résulats des comparaison

baytri <- bay[order(bay$F.p.value, decreasing=FALSE),]
                        ### ordonner les gènes selon un ordre décroissant de la P-value
                        ### donc les gènes qui sont en 1er dans le tableau sont ceux qui
sont différentiellement exprimés entre les expériences

head(baytri)                  ### voir ce tableau ordonné

mGenes <- baytri$genes$ID[1:1000]  ### selectionner les 1000 premiers gènes les
plus intéressants à regardés
head(mGenes)                  ### voir cette liste de 1000 gènes

expGenes = subset(settr, subset=rownames(settr)%in% mGenes)
                        ### récupere les valeurs de l'expression de ces 1000 gènes
head(expGenes)                ### voir le tableau isolés de ces 1000 gènes

write.table(x=cbind(rownames(expGenes),expGenes),file="top1000.txt",
sep='\t',row.names=FALSE)
                        ### enregister ce tableau sur le PC pour le visualiser avec
TMeV plutard

######################### The end of script
```