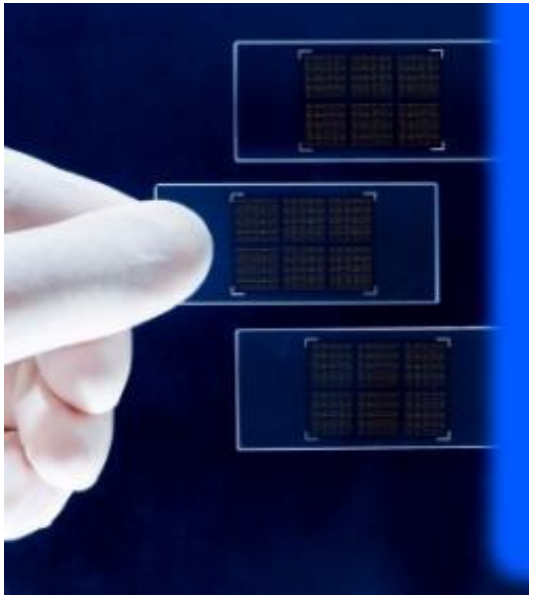
Introduction à l'analyse différentielle d'expression de gènes pour les puces à ADN

Les puces à ADN

- Une **puce à ADN** est un ensemble de molécules d'ADN fixées en rangées ordonnées sur une petite surface qui peut être du verre, du silicium ou du plastique.
- Ce dispositif permet, par exemple, d’analyser le niveau d’expression des gènes dans une cellule à un moment donné par rapport à un échantillon de référence.

source : wikipedia

Les puces à ADN



Exemples de Puce à ADN

Les puces à ADN

- sondes : fragment d’ADN synthétique représentatif des gènes dont on cherche à étudier l'expression.
- cibles : ARNm que l’on cherche à identifier et/ou à quantifier.
- Le nombre de sondes peut varier de quelques milliers à plus de 1 million suivant les puces et les technologies (1 ou 2 conditions par puces).

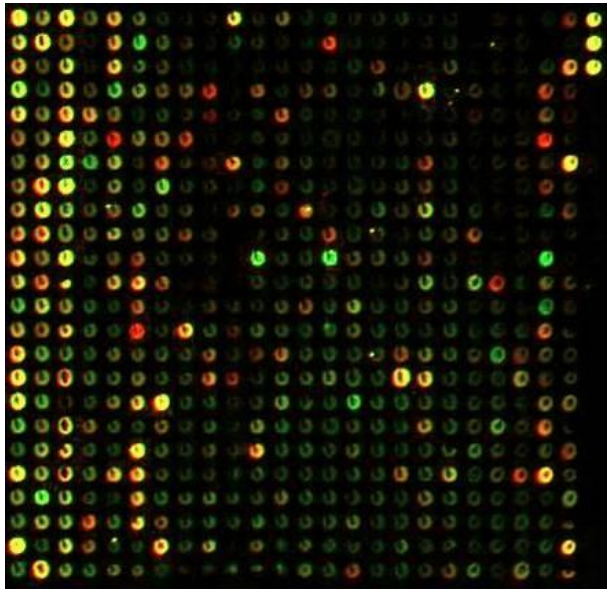
Les puces à ADN

Les étapes d'une expérience de puce à ADN sont les suivantes :

- Extraction de l'ARNm des cellules et amplification
- Transformation en ADNc par rétrotranscription
- Marquage par une molécule fluorescente
- Hybridation des brins d'ADNc avec les sondes
- Analyse de l'hybridation par scanner

Les puces à ADN

Résultat



- L'image scannée est alors analysée informatiquement afin d'associer une valeur d'intensité à chaque sonde
- Ce sont ces intensités que l’on va analyser par la suite.

Les étapes de l’analyse

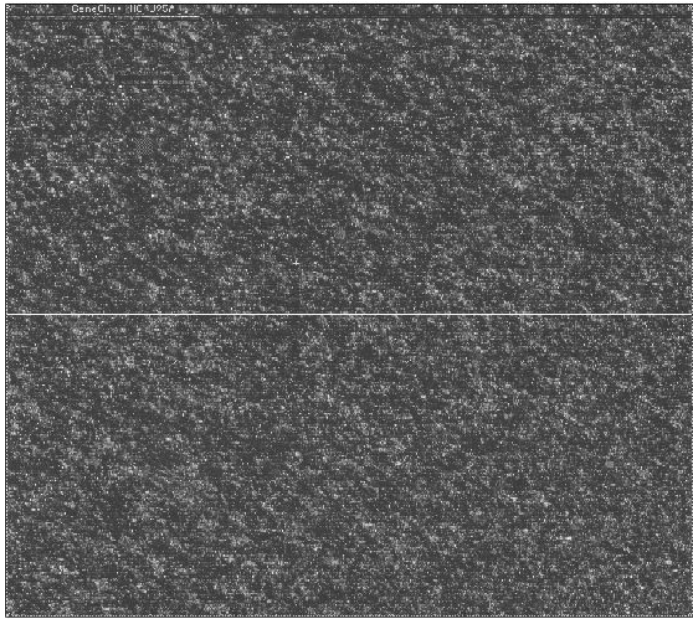
- La phase de normalisation
  - permet de “nettoyer” les données et de les rendre comparables.
- La phase d’analyse différentielle
  - permet à l’aide de méthodes statistiques, d’établir quels sont les gènes différentiellement exprimés entre plusieurs conditions.

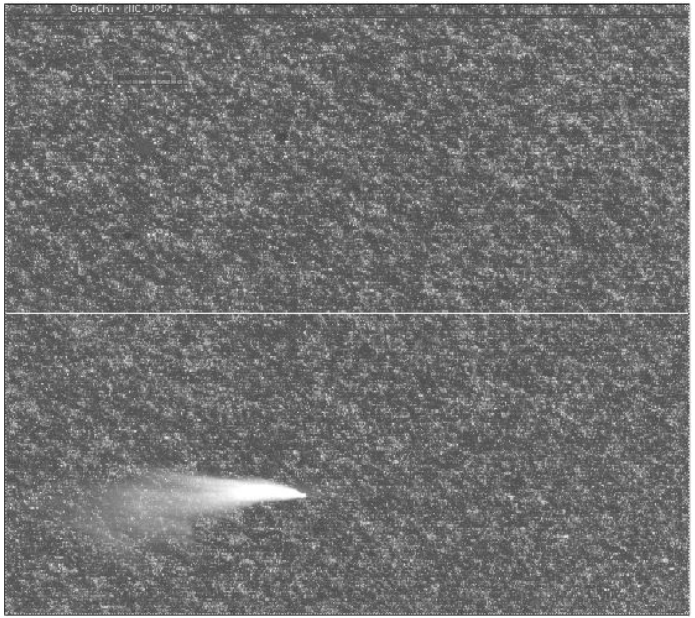
Normalisation et qualité des données

- La phase de normalisation permet
  - de s’assurer que les données sont exploitables
  - de réduire les biais techniques expérimentaux
  - de pouvoir comparer les données des différentes puces entre elles
  - de s’approcher des hypothèses favorables pour l’analyse différentielle (distribution gaussienne des données)

Normalisation et qualité des données

Visualisation de l’image des puces :





Normalisation et qualité des données

Il existe de nombreuses méthodes de normalisation.

Elles se fondent généralement sur 2 hypothèses :

- Seule une minorité de gènes est différentiellement exprimés
- Les nombres de gènes sous-exprimés et sur-exprimés sont équivalents

Normalisation et qualité des données

Une des méthodes la plus répandue est la normalisation rma (Robust Multi-Array Average) :

- background correction : supprime le bruit et les artefacts locaux, les mesures ne sont plus affectées par les mesures voisines
- normalization : supprime les effets liés aux puces, permet de comparer les mesures de puces différentes
- summarization : combine les mesures entre plusieurs sondes pour donner une mesure d’expression au niveau du gène

Normalisation et qualité des données

De plus, les valeurs sont transformées via la fonction log2

- Afin de se rapprocher d'une distribution gaussienne
- De faciliter l'interprétation des résultats.

En effet, le Fold change ou (FC), ratio qui mesure la variation de l'expression d'un gène entre 2 conditions, est symétrisé par passage au log. Le logFC sera donc négatif pour un gène sous-exprimé et positif pour un gène sur-exprimé.

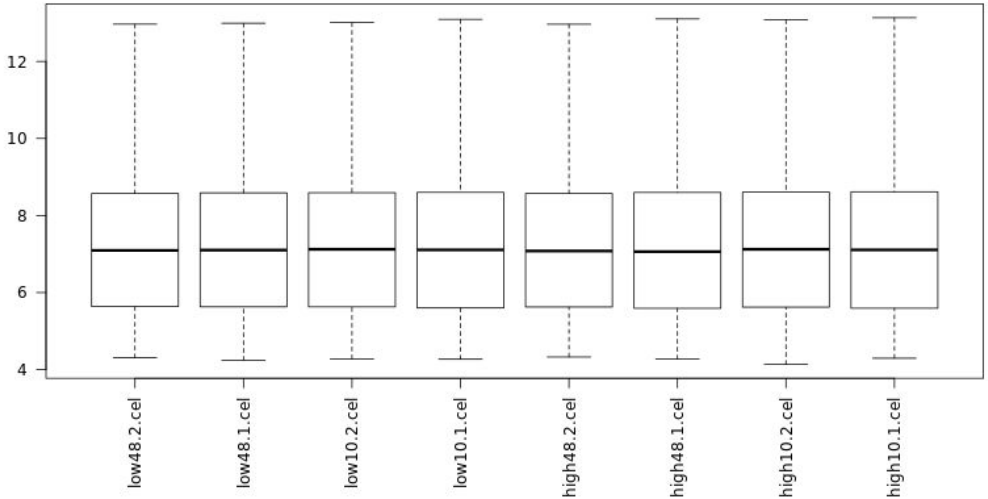
Normalisation et qualité des données

Afin de vérifier l’efficacité de la phase de normalisation, on utilise différents types de visualisation des données.

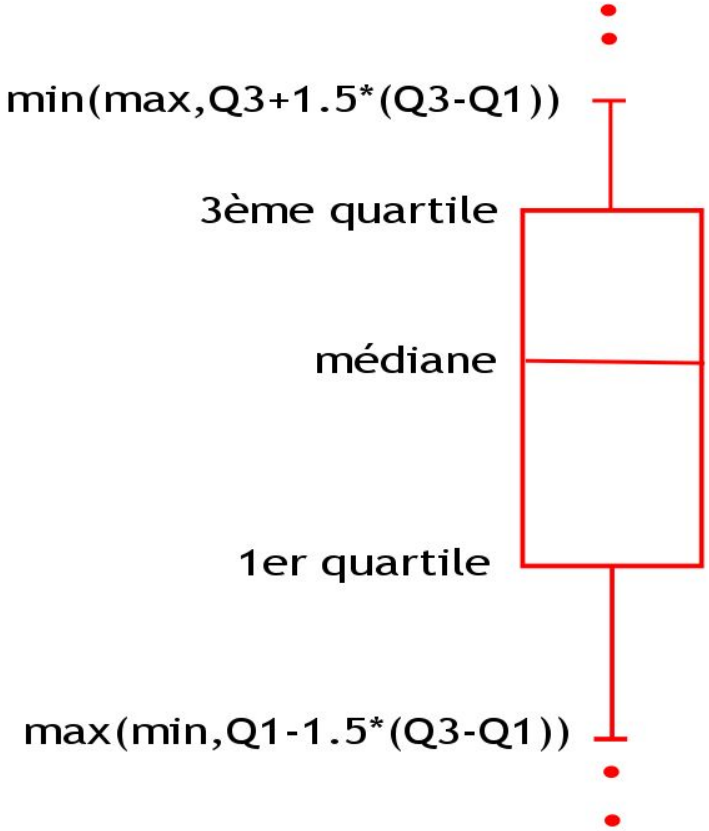
Voici 2 types de graphiques utilisés :

Normalisation et qualité des données

Les boxplots



Exemple de boxplots après normalisation



Normalisation et qualité des données

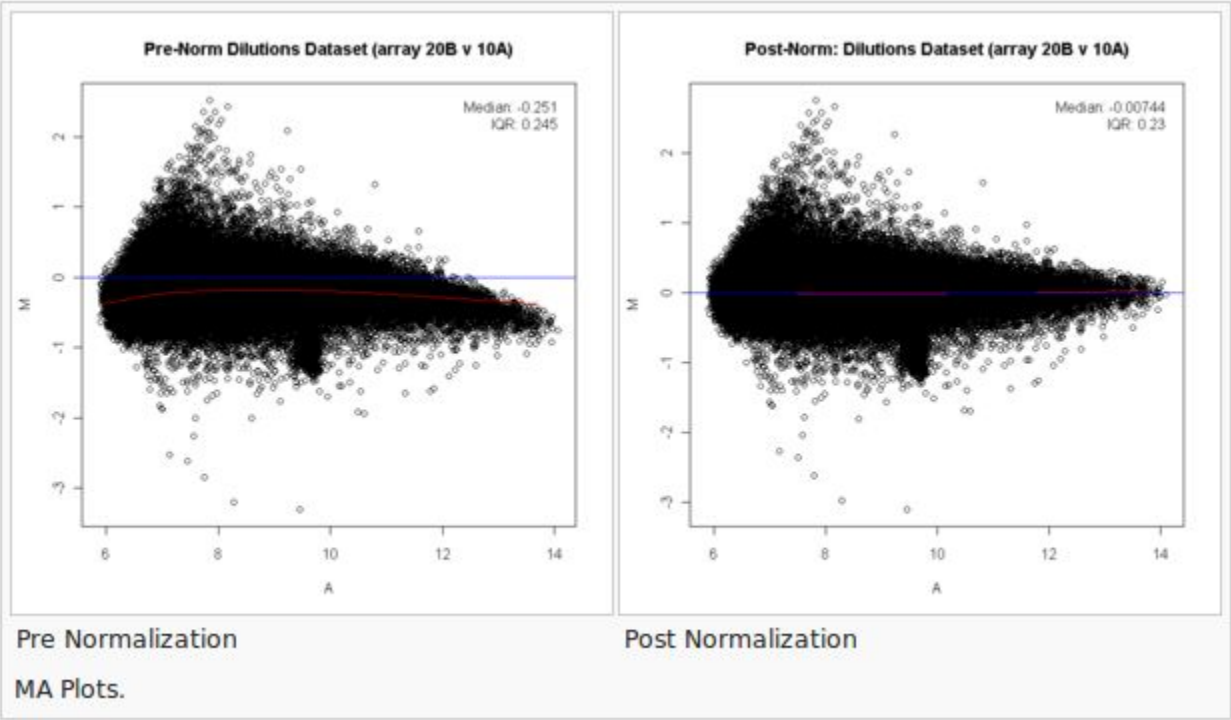
Les MA-plots (pour 1 condition par puce) :

- en ordonnée M = log2_array – log2_medianarray
- en abscisse A = (log2_array + log2_medianarray)/2

Les MA-plots montrent dans quelle mesure la variabilité de l'expression dépend du niveau d'expression.

Le nuage de points doit être centré en 0 et la ligne rouge proche de la ligne bleue.

Normalisation et qualité des données



source : wikipedia

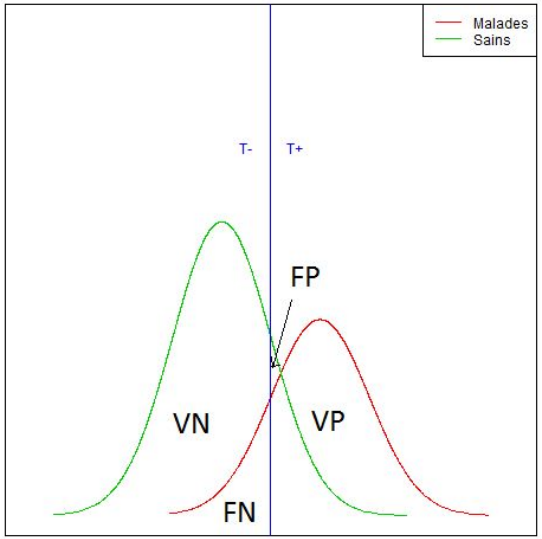
Analyse différentielle

Rappel sur les Tests statistiques :

- Hypothèse nulle H0 : hypothèse testée.
- Exemple : les niveaux d’expression d’un gène entre 2 conditions sont égaux.
- p-value : probabilité d'obtenir la même valeur (ou une valeur encore plus extrême) du test si l'hypothèse nulle était vraie.

Analyse différentielle

	H0 rejetée	H0 non rejetée
H0 vraie	Faux Positif	Vrai Négatif
H0 fausse	Vrai Positif	Faux Négatif



Analyse différentielle

Problème des tests multiples :

- Dans le cas de l'analyse différentielle de plusieurs milliers de gènes le nombre de faux positifs peut devenir très grand.

Une correction pour tests multiples est nécessaire.

Analyse différentielle

Correction pour les tests multiples :

- Family Wise Error Rate : rejeter à tort au moins une hypothèse nulle (e.g Bonferroni). Très stringeant.
- False Discovery Rate : contrôler la proportion attendue de faux positifs parmi les positifs (e.g. Benjamini Hochberg).

Analyse différentielle

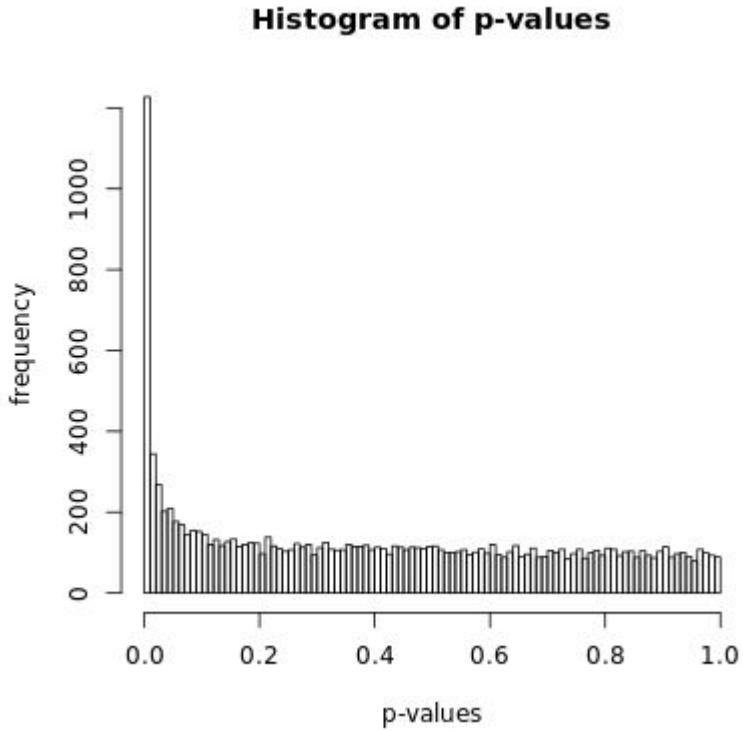
Plusieurs approches pour les tests :

- Un test par gène : manque de puissance
- Hypothèse que la variance est commune à tous les gènes
  - beaucoup de faux positifs
- Tests modérés : compromis entre approches gène à gène et variance commune. Méthode utilisée par le package R limma.

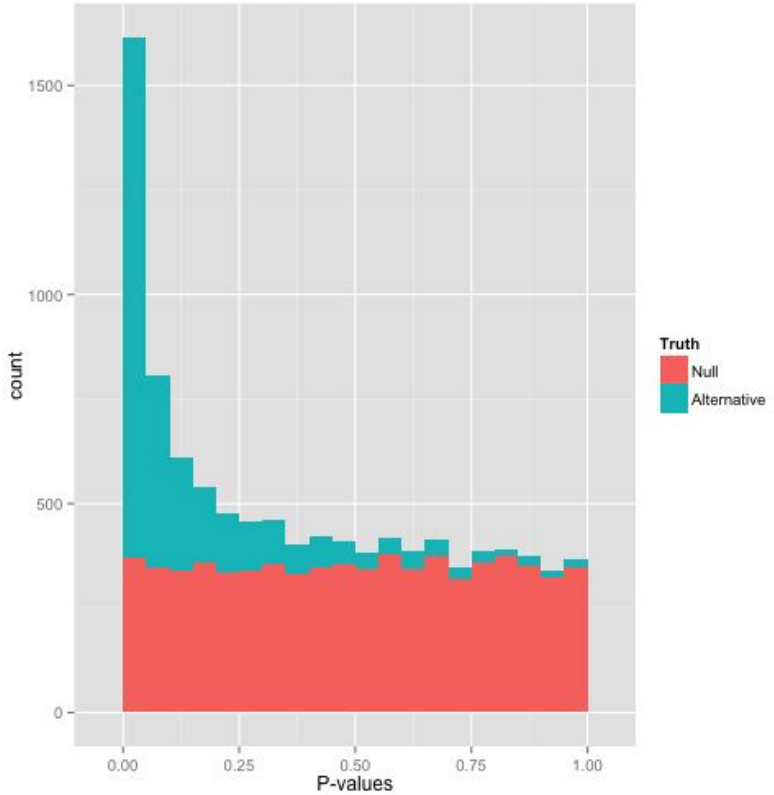
Analyse différentielle

L’histogramme de p-values :

il représente la distribution des p-values brutes (avant la correction pour tests multiples) et permet de s’assurer que les tests se comportent de façon attendue.



Analyse différentielle



Sous H0, la distribution des p-value est uniforme.

La forme attendue de l'histogramme est donc globalement plate avec un pic proche de 0, représentant les gènes différentiellement exprimés (qui rejettent H0).

Analyses complémentaires

D’autres analyses peuvent êtres effectuées :

- Gene Set Enrichement Analysis (GSEA) : Détecter des groupes (prédéfinis) de gènes qui sont sur ou sous-représentés dans l’ensemble des gènes différentiellement exprimés.
- Clustering