Analyse de données génomiques

Session 1

David Causeur

Agrocampus Ouest, IRMAR CNRS UMR 6625

Objectifs pédagogiques

A la fin du module, vous serez capables de :

- choisir des méthodes d'analyse adaptées aux problématiques de l'analyse de données -omiques ;
- mettre en œuvre ces méthodes avec `R` ;
- évaluer la qualité des règles de décision mises en œuvre ;
- analyser des données de grande dimension.

Evaluation

- ~~Contrôle de connaissance (1h) - 50%~~
- Etude de cas - ~~50 %~~ 100%

Plan du cours

Mesure d'expression de gènes

Technologie	Principe
RT-qPCR (1992)	Nombre de cycles d'amplification pour lesquels la quantité de transcrits dépasse un seuil définissant le bruit de fond
Microarray (1997)	Intensité relative de fluorescence indiquant la quantité de transcrits
NGS (2008)	Nombre de copies d'une séquence d'ADN dans une région du génome

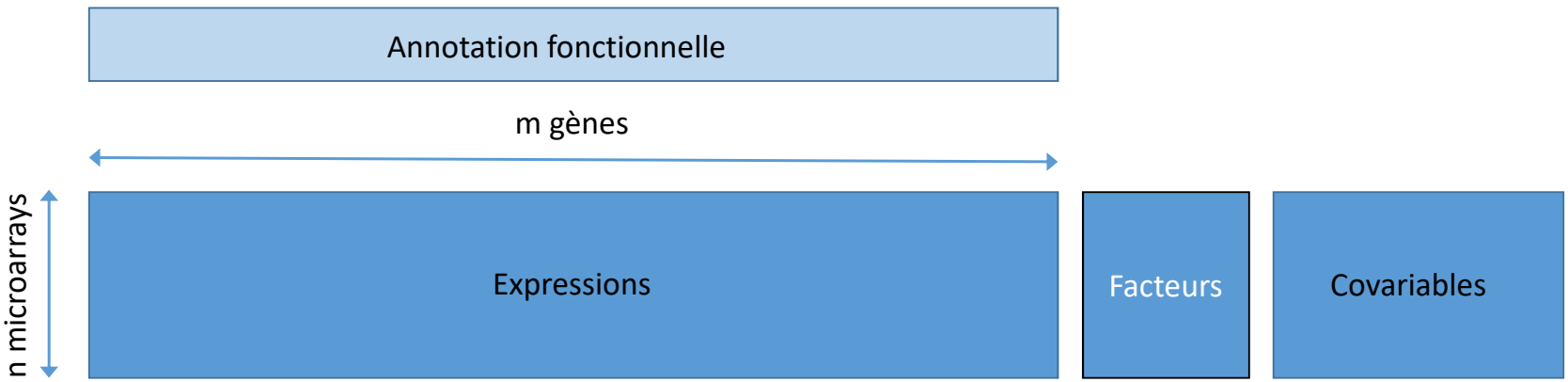
Mesure d'expression de gènes

Technologie	Principe
RT-qPCR (1992)	Nombre de cycles d'amplification pour lesquels la quantité de transcrits dépasse un seuil définissant le bruit de fond
Microarray (1997)	Intensité relative de fluorescence indiquant la quantité de transcrits
NGS (2008)	Nombre de copies d'une séquence d'ADN dans une région du génome

Données à haut débit

- moins précises à l'échelle du gène,
- nécessitant des pré-traitements.

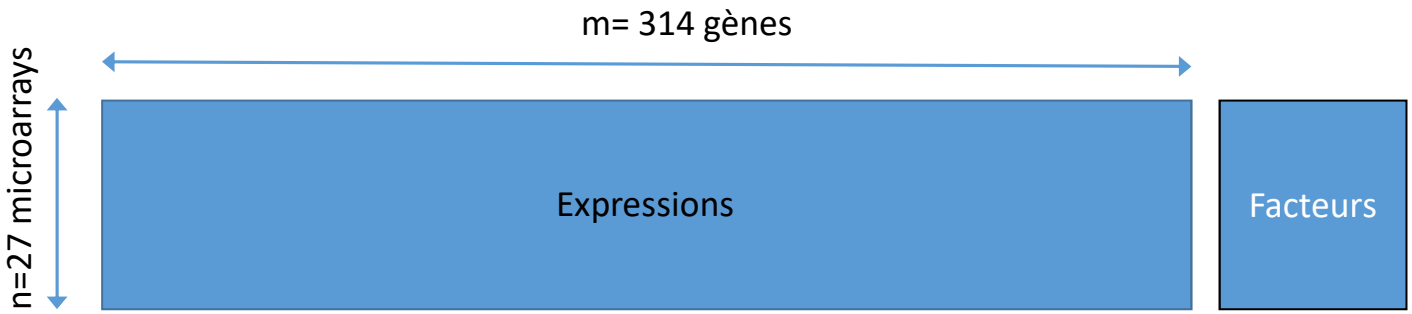
Dispositifs expérimentaux



Problématiques de la statistique génomique

- Identification de gènes d'intérêt
  - Sélection de gènes dont l'expression dépend des conditions expérimentales
  - Identification de signature moléculaire pour la prédiction
- Modélisation des processus biologiques
  - Interprétation des groupes de gènes par leur annotation fonctionnelle
  - Intégration de données -omiques
  - Inférence de réseau de régulation

Illustration par une analyse différentielle



Régime : A jeun, Nourri
Lignée : Maigre, Gras

Illustration par une analyse différentielle

Quels sont les gènes dont l'expression dépend du régime ?

Illustration par une analyse différentielle

Pour chaque gène, une analyse de la variance à un facteur

Y_{ij} = expression du gène pour le poulet j soumis au régime i,

$[i = 1]$: [A jeun]

$[i = 2]$: [Nourri]

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$,

avec $\alpha_1 = 0$ ['A jeun' est le groupe de référence]

Test de l'effet 'régime'. $H_0 : \alpha_2 = 0$

Statistique de test : t-test (Student), F-test (Fisher).

Illustration par une analyse différentielle

Peu de gènes positifs ... 

Pas de différences d'expression entre les régimes ?

Si le signal est inexistant, combien de gènes positifs (en moyenne) ?

Plan du cours

Puissance de l'analyse différentielle

Statistique de Fisher

$$F = \frac{\text{variation entre groupes}}{\text{variation intra-groupe}},$$

Variation intra-groupe : biologique + technologique

objectif : augmenter la puissance du test

... réduire la variation intra-groupe d'origine technologique.

Normalisation

Transformation des données d'expression visant à extraire leur part de variabilité non-biologique (technologique)

Puissance de l'analyse différentielle

Etape 1 de la normalisation: transformation logarithmique

... un peu plus de gènes positifs 

Variabilité technologique

Hypothèse d'invariance
Les répartitions des mesures d'expression par microarray sont identiques.

Hypothèse d'invariance partielle
Les répartitions des mesures d'expression de gènes de contrôle par microarray sont identiques.

Si des variations de répartitions sont observées entre microarrays, elles sont d'ordre technologique.

Correction des biais technologiques

Normalisation par la médiane

Si median(y_i) désigne l'expression médiane sur la *i*ème *microarray*,

$$\tilde{y}_{ij} \quad = \quad y_{ij} - \text{median}(y_i).$$

Correction des biais technologiques

Normalisation par la médiane

Si median(y_i) désigne l'expression médiane sur la *i*ème *microarray*,

$$\tilde{y}_{ij} = y_{ij} - \text{median}(y_i).$$

Illustration

	G_1	G_2	G_3	G_4
Array 1	5	2	3	4
Array 2	4	1	4	2
Array 3	3	4	6	8

Correction des biais technologiques

Normalisation par la médiane

Si median(y_i) désigne l'expression médiane sur la *i*ème *microarray*,

$$\tilde{y}_{ij} = y_{ij} - \text{median}(y_i).$$

Illustration

	G_1	G_2	G_3	G_4
Array 1	5	2	3	4
Array 2	4	1	4	2
Array 3	3	4	6	8

q_{25}	q_{50}	q_{75}	$\bar{y}$
2.75	3.50	4.25	3.50
1.75	3.00	4.00	2.75
3.75	5.00	6.50	5.25

Correction des biais technologiques

Normalisation par la médiane

Si median(y_i) désigne l'expression médiane sur la *i*ème *microarray*,

$$\tilde{y}_{ij} \quad = \quad y_{ij} - \text{median}(y_i).$$

Illustration

	G_1	G_2	G_3	G_4
Array 1	1.50	-1.50	-0.50	0.50
Array 2	1.00	-2.00	1.00	-1.00
Array 3	-2.00	-1.00	1.00	3.00

q_{25}	q_{50}	q_{75}	$\bar{y}$
-0.75	0.00	0.75	0.00
-1.25	0.00	1.00	-0.25
-1.25	0.00	1.50	0.25

Correction des biais technologiques

Normalisation par les quantiles

Illustration

	G_1	G_2	G_3	G_4
Array 1	5	2	3	4
Array 2	4	1	4	2
Array 3	3	4	6	8

Moyennes au			
Rang 1	Rang 2	Rang 3	Rang 4
2.00	3.00	4.67	5.67

Correction des biais technologiques

Normalisation par les quantiles

Illustration

	G_1	G_2	G_3	G_4
Array 1	5.67	2.00	3.00	4.67
Array 2	4.67	2.00	4.67	3.00
Array 3	2.00	3.00	4.67	5.67

Moyennes au			
Rang 1	Rang 2	Rang 3	Rang 4
2.00	3.00	4.67	5.67

Correction des biais technologiques

Normalisation par les quantiles :

Illustration

	G_1	G_2	G_3	G_4
Array 1	5.67	2.00	3.00	4.67
Array 2	5.67	2.00	4.67	3.00
Array 3	2.00	3.00	4.67	5.67

q_{25}	q_{50}	q_{75}	$\bar{y}$
2.75	3.83	4.92	3.83
2.75	3.83	4.67	3.58
2.75	3.83	4.92	3.83

Correction des biais technologiques

Package `limma`: Linear Model for Microrray Analysis

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.

`limma`, un des packages du portail `bioconductor.org`

- Importation de données issues de la plupart des technologies
- Contrôle qualité des données et normalisation
- Analyse différentielle