LSTAT2340: Traitement statistique de données -OMICS

Professeurs Céline Bugli
& Bernadette Govaerts
celine.bugli@uclouvain.be
bernadette.govaerts@uclouvain.be
Janvier 2018

Analyse de données transcriptomiques: données qPCR

Plan du cours

- Introduction: outils de base et introduction au traitement de données omiques
- Concepts biologiques utiles en omiques (par JL Gala)
- Analyse de données transcriptomiques : technologie q-PCR et prétraitement.
- Multiple testing
- Méthodes de modélisation multivariée (Multiple Regression stepwise/PCR/PLS/O-PLS/Lasso et PLS-DA et O-PLS DA)
- Traitement de données micro-damier et de high sequencing (participation de J. Ambroise)
- Données métabolomiques: Prétraitement de données RMN
- Outils statistique pour la validation de modèles multivariés. Evaluation de la répétabilité de spectres.
- Intégration de données mutisources + méthodes de type ASCA et APCA
- Traitement de données métagénomique (J. Ambroise)

Données transcriptomiques

- Rappel: L'ARN messager (ARNm) est produit à partir de la molécule d'ADN d'un gène et sert à la construction d'une protéine encodée par ce gène.

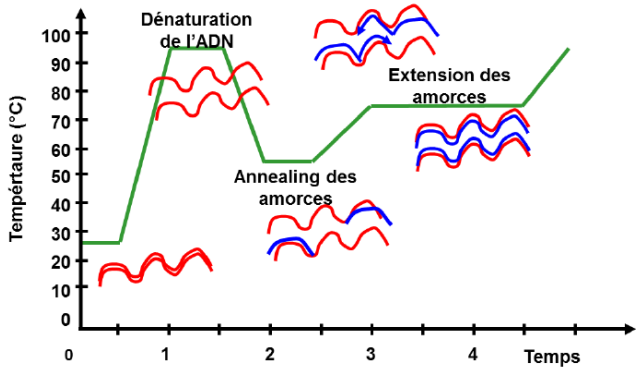
- Un gène est exprimé dans une cellule ou un groupe de cellules lorsqu'il est transcrit en ARNm ou que la protéine produite résultante est détectée. La synthèse des protéines comporte deux étapes : la transcription et la traduction. Lors de la transcription, l'ADN est transcrit en ARNm, celui-ci est ensuite traduit en protéine.

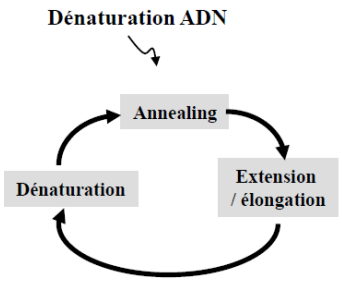
Données transcriptomique: quantification des ARNm

- Différentes méthodes existent pour détecter et mesurer la quantité d'ARNm, et donc l'expression d'un gène, dans un échantillon biologique. Les éléments principaux pour la détection et la quantification d'une quantité spécifique d'ARNm sont une quantité d'ARN total ou messager suffisante, une sonde spécifique à la séquence cible, une méthode de détection sensible et des contrôles pour l'interprétation des résultats (Roth, 2002).
- Différences techniques existent pour la mesure de l'expression de gènes :
  - les techniques électrophorétiques (comme le nothern blot),
  - les puces à ADN (voir cours 4)
  - et les méthodes basées sur la PCR.

Réaction en Chaîne par Polymérase (PCR)

- Développée dans les années 80 par Kary Mullis et associés, la PCR (Polymerase Chain Reaction) est une réaction enzymatique permettant de produire de multiples copies de séquences spécifiques d'ADN.





Source: JL Gala

PCR quantitative: qPCR

LA PCR quantitative est une méthodologie utilisée pour évaluer la nombre de molécules cibles d'ADN, ARN ou encore ARNm présent dans un échantillon (van Pelt-Verkuil et al., 2008). En réalité, la PCR quantitative mesure le nombre d'amplicon (portion d'ADN définie par un couple d'amorces).

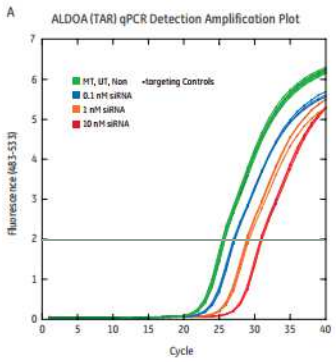
Deux stratégies de quantification existent :

- relative : décrit les changements dans la quantité du gène cible par rapport à un contrôle interne. Les résultats sont exprimés en ratio.
- et absolue: détermine le nombre exact de copie du gène d'intérêt et cela en comparant habituellement le signal PCR à une courbe de calibration.

La quantification est effectuée sur base de la fluorescence émise durant la phase exponentielle de l'amplification.

Quantification qPCR

- Cq : « cycle seuil » ou « quantification cycle »
- aussi connu sous les appellations "threshold cycle"(C_t), "crossing point"(C_p), "take-off point" (TOP)
- Le seuil de quantification représente la fraction du nombre de cycle PCR pour lesquels l'émission de fluorescence est supérieur au seuil de détection.



Calcul de la concentration relative entre 2 échantillons

La concentration relative de la cible dans 2 échantillons se calcule par comparaison des Cq (ou C_t). Si on considère des efficacités de réaction E_{PCR} égales dans les échantillons, le rapport des concentrations des acides nucléiques dans ces échantillons vaut :

$$\textbf{Ratio=(Ech\#2)/(Ech\#1)} = (1+E_{PCR})^{Cq1-Cq2} = (1+E_{PCR})^{\Delta Ct}$$

Dans le cas où l'efficacité de réaction est maximale ($\mathbf{E_{PCR}}=1$) la formule devient :

$$\textbf{Ratio=(Ech\#2)/(Ech\#1)} = 2^{Cq1-Cq2} = 2^{\Delta Ct}$$

Normalisation par gène de référence (ou une combinaisson de gènes de référence)

On va toujours comparer la cible dans un échantillon pour un patient (cible2 dans l'échantillon 2) par rapport à échantillon considéré comme un calibrateur (échantillon 1) et que l'on va garder tout au long de l'expérience (càd pour tous les patients).

De plus, souvent, on normalise les données par rapport à un ARNm de référence: càd un ARNm que l'on pense stable d'un échantillon à l'autre (Ref1 dans l'échantillon 1 càd dans le calibrateur et Ref2 dans l'échantillon 2 càd pour le patient).

La concentration de la cible dans deux échantillons après correction par le signal de la référence (Ref: un ARNm de référence que l'on pense stable d'un échantillon à l'autre) vaut :

$$\textbf{ratio= (cible2/Ref2) / (cible1/Ref1)}$$

Calcul par la méthode des ΔΔCq

Si E_{PCR}=1,

ΔCq1 = Cq (cible1) – Cq (Ref1),

ΔCq2 = Cq (cible2) – Cq (Ref2)

et ΔΔCq =ΔCq1 – ΔCq2

⇒ **ratio= $2^{\Delta\Delta Cq}$**

Cette méthode s'appelle la Méthode de Livak.

Si E_{PCR}≠1, la formule est plus compliquée…

Attention, on ne fait pas de statistiques sur les Cq directement!

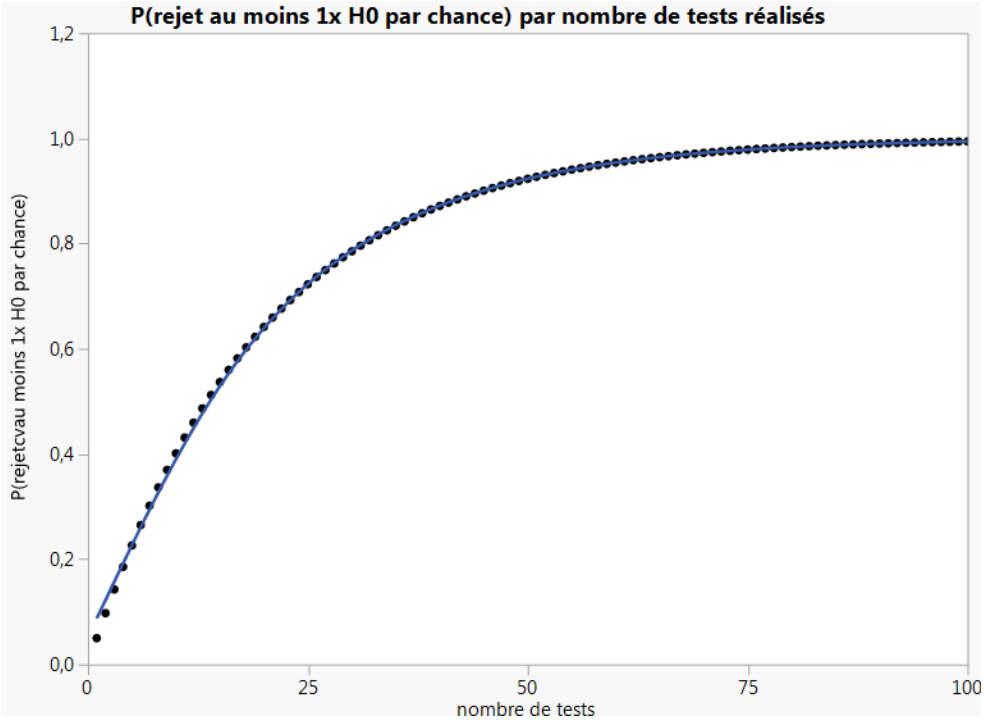
- Le Cq est fréquemment exprimé sous la forme Cq moyen ± SD.
- Le Cq étant une fonction logarithmique, il est mathématiquement incorrecte d'exprimer les Cq sous forme de Cq moyen ± SD car cela ne rend compte ni de la moyenne du nombre de copies dans les échantillons ni de la déviation standard sur cette mesure.
- Le Cq ne peut pas non plus être utilisé dans des tests statistiques classiques tels que *t-tests* et ANOVAs.

Références

- Notes de cours de « LA QUANTIFICATION D'ACIDES NUCLEIQUES PAR PCR ET RT-PCR » , Charles Lambert (Ulg).

Multiple testing

Le problème du multiple testing



Types of errors

RECALL: *To show an effect, we want to reject the null-hypothesis: hypothesis of no-effect or no-change.*

H0	True (No effect)	False (effect exists)
Rejected (significant test)	Type I error ⇒ False positive ⇒ probability=α	Ok ⇒ True positive ⇒ Probability=1-β
Not rejected (not significant test)	Ok ⇒ True negative ⇒ Probability=1-α	Type II error ⇒ False negative ⇒ probability=β (lack of power ?)

Suppose H0 is true….

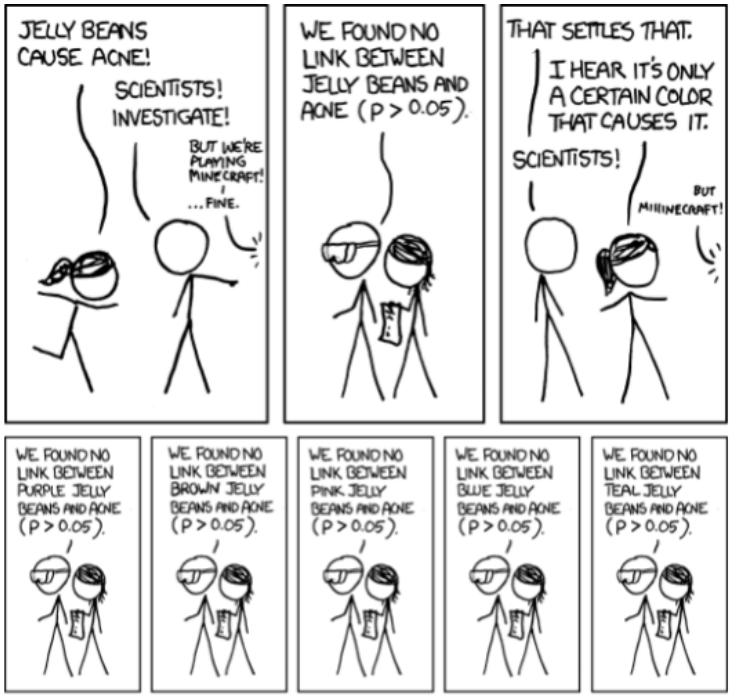
P(Making an error) = α

P(Not making an error) = 1 - α

P(Not making an error in N tests) = $(1 - \alpha)^N$

P(Making at least 1 error in N tests) = $1 - (1 - \alpha)^N$

Jelly beans and acne?!



To be continued…(with a lot of possible colors…)

Source: https://en.wikipedia.org/wiki/Xkcd

Back to the jelly beans problem…

1 test for all colors mixed up

20 tests for the 20 colors…

- **21 statistical tests**
- **Probability of Type I error=**1-(1-0.05)^21=0.66!!!

Also known as
« Data dredging »,
« Cherry picking », …

		Declared significant: True	Declared significant: False
H0	True	FP	TN
H0	False	TP	FN

TP	True positive	TP
FP	False positive	FP
FN	False negative	FN
TN	True negative	TN
KP	Known Positive	TP+FN
KN	Known Negative	TN+FP
PP	Predicted Positive	TP+FP
PN	Predicted Negative	FN+TN
N	Total	TP + FP + FN + TN
Prev	Prevalence	(TP + FN)/N
ODP	Overall Diagnostic Power	(FP + TN)/N
CCR	Correct Classification Rate	(TP + TN)/N
Sn	**Sensitivity**	**TP/(TP + FN)**
Sp	Specificity	TN/(FP + TN)
FPR	False Positive Rate	FP/(FP + TN)
FNR	False Negative Rate	FN/(TP + FN) = 1-Sn
PPV	**Positive Predictive Value**	**TP/(TP + FP)**
FDR	False Discovery Rate	FP/(FP+TP)
NPV	Negative Predictive Value	TN/(FN + TN)
Mis	Misclassification Rate	(FP + FN)/N
Odds	Odds-ratio	(TP + TN)/(FN + FP)

Sn = TP/(TP+FN)
Declared significant

H0	True	False
True	FP	TN
False	TP	FN

PPV=TP/(TP+FP)
Declared significant

H0	True	False
True	FP	TN
False	TP	FN

Sp=TN/(FP+TN)
Declared significant

H0	True	False
True	FP	TN
False	TP	FN

NPV=TN/(FN+TN)
Declared significant

H0	True	False
True	FP	TN
False	TP	FN

FPR=FP/(FP+TN)
Declared significant

H0	True	False
True	FP	TN
False	TP	FN

FDR=FP/(FP+TP)
Declared significant

H0	True	False
True	FP	TN
False	TP	FN

FN/(FN+TN)
Declared significant

H0	True	False
True	FP	TN
False	TP	FN

FNR=FN/(TP+FN)
Declared significant

H0	True	False
True	FP	TN
False	TP	FN

Corrections classiques – correction de Bonferroni

Stratégie 1: diviser le alpha par le nombre de comparaisons (N)

$$\alpha_b = \frac{\alpha_n}{N} << \frac{1}{N}$$

where
α_n =Nominal p-value
N = Number of tests
α_b = Bonferonni-corrected threshold

Stratégie 2: garder le même alpha mais multiplier les p-valeurs par le nombre de comparaisons

$$P_{Bonferroni} = \begin{cases} N.Pval & \text{if } Pval < 1/N \\ 1 & \text{otherwise} \end{cases}$$

correction de Bonferroni (suite)

Note: on parle aussi de E-value qui représente le nombre de faux positifs attendus

$$Eval = N \cdot Pval$$

Avantages de la méthode de Bonferroni: facile à mettre en œuvre, contrôle du taux de faux positifs (FPR)

Inconvénients: Correction très forte si beaucoup de comparaison menant à une perte de sensitivité (Sn) => risque de ne plus avoir grand-chose de significatif…

Il existe 2 stratégies de correction pour données de grandes tailles en contrôlant l'erreur de type I:

- Family-Wise Error Rate Control (FWER): probabilité d'observer au moins un faux positif :

$$P(FP \geq 1) = 1 - P(FP = 0)$$

- False Discovery Rate Control (FDR): proportion de faux positifs parmi tous les tests significatifs

$$FDR = FP / (FP + TP)$$

FDR and FWER control of type I error

	number declared non-significant	number declared significant	total
true null hypotheses	U	V	m_0
false null hypotheses	T	S	$m - m_0$
	m - R	R	m

V = # Type I errors [false positives]

Family-Wise Error Rate Control (FWER): probability of observing at least one false positive :

$$FWER = P(V \geq 1)$$

=> FWER is appropriate when you want to guard against ANY false positives

False Discovery Rate Control (FDR): proportion of false positive tests amongst all positive tests

$$FDR = E[V/R]$$ (Benjamini and Hochberg, 1995)

=> designed to control the proportion of false positives among the set of rejected hypotheses (R)

Contrôle du FWER

- Contrôle du seuil:
  - Approche de Sidak: considérer un seuil $\alpha_{FWER} = 1 - (1-\alpha)^{1/N}$
  - Approche de Bonferroni: On calcule toutes les p-valeurs et on les trie de la plus petite à la plus grande. Ensuite, on évalue quel doit être le seuil de significativité pour obtenir un taux de FWER = α. On va considérer comme significatifs les tests 1 à k tels que la p-valeur du kième test : $P_{(k)} \leq \frac{\alpha}{N}$
  - Approche de Holm: Classer les p-valeur de la plus petite à la plus grande, considérer comme significatif les tests 1 à k tels que : $P_{(k)} \leq \frac{\alpha}{N+1-k}$
- Corriger les p-valeurs:
  - Approche de Bonferroni: p-valeur corrigée= min(1,p-valeur*N)
  - Approche de Sidak : $FWER = 1 - (1 - Pval)^N$

FWER in practice (with *Sidak* approach)

1. Compute all the N p-values and sort them
   (from smaller to higher)

2. The first k p-values are significiant if
   - $P_{(k)} \leq \alpha_{FWER} = 1 - (1-\alpha)^{1/N}$
   - Or compute a p-value $FWER = 1 - (1 - Pval)^N$

k	1	2	3	4	5	…	1000
p-value	0,00001	0,000013	0,000023	0,000144	0,000294		0,04
P-value bonferroni	0,01	0,013	0,023	0,1444	0,2944		
P-value < $1 - (1-\alpha)^{1/N}$= 0,00005129	yes	yes	yes	No => STOP			
P-value FWER (Sidak)	0,00995	0,012916	0,022738	0,134675			

Contrôle du FDR et Q-valeur

Une q-valeur est une p-valeur corrigée par la méthode FDR. Par exemple, par la procédure de *Benjamini–Hochberg*:

On calcule toutes les p-valeurs et on les trie de la plus petite à la plus grande. Ensuite, on évalue quel doit être le seuil de significativité pour obtenir un taux de FDR = α.

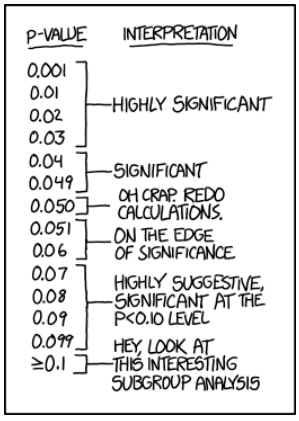
On va considérer comme significatifs les tests 1 à k tels que la p-valeur du kième test : $P_{(k)} \leq \frac{k}{N}\alpha$

De manière équivalente: q-valeur[k]=(pvaleur[k]*N)/k

Remarque: la procédure de *Benjamini–Hochberg–Yekutieli* permet de tenir compte de la corrélation entre les tests

FDR in practice (with *Benjamini–Hochberg* procedure)

1. Compute all the N p-values and sort them (from smaller to higher)
2. The first k p-values are significiant if
   - $P_{(k)} \leq \frac{k}{N}\alpha$
   - Or compute a q-value:q-valeur[k]=(pvaleur[k]*N)/k



k	1	2	3	4	5	…	1000
p-value	0,00001	0,000013	0,000023	0,000144	0,000294		0,04
k/N alpha	0,00005	0,0001	0,00015	0,0002	0,00025		0,05
significant	yes	yes	yes	yes	no => STOP		
qvalue	0,01	0,0065	0,0077	0,0361	0,05888		
P-value bonferroni	0,01	0,013	0,023	0,1444	0,2944		

Inteprétation

- Une p-valeur de 0.05 implique que 5% de tous les tests donneront un faux positif
- Une p-valeur ajustée par FDR (ou q-value) de 0.05 implique que 5% des tests significatifs donneront un faux positif.
- Les p-valeurs corrigées par FDR sont moins conservatives que par l'approche de Bonferroni et permettent d'obtenir une plus grande puissance.

FWER approach ⇒ pvalue < 0,05 => 5% of all tests are false positive

FDR approach ⇒ qvalue < 0,05 => 5% of significant tests are false positive

Méthodes plus avancées

- Utilisation de tests de permutation, bootstrap ou de stat bayesiennes
- Exemple: procédure de Westfall & Young
  - Si on a m tests à réaliser, il y a $O(2^m)$ sous-ensembles I de tests d'hypothèses possibles.

Min P: $\tilde{p}_i = \Pr\left(\min_{1 \le l \le m} P_l \le p_i \mid H_{\mathcal{M}}\right)$.

Max T: $\tilde{p}_i = \Pr\left(\max_{1 \le l \le m} |T_l| \ge |t_i| \mid H_{\mathcal{M}}\right)$.

where H_M denotes the complete null hypothesis and P_l the random variable for the raw p-value of the lth hypothesis

Recommendations

- La correction la plus recommandée dan la littérature est la méthode FDR de Benjamini&Hochberg car elle fourni un bon compromis entre la découvertes de différences potentiellement statistiquement significatives et la limitation de l'occurrence de faux positifs.
- La correction de Bonferroni est la correction la plus forte de toutes, mais fourni l'approche la plus conservative pour le contrôle des faux positifs.
- Le choix de la méthode dépend de l'objectif de l'étude: exploratoire ou confirmatoire

Références

- Multiple Testing with Minimal Assumptions, P. Westfall and J. Troendle, Biom J. 2008 Oct; 50(5): 745–755.

- Multiple Testing Procedures: R multtest Package and Applications to Genomics, K, Pollard, S. Dudoit, M. van der Laan (2004)

Case study 2: Données qPCR

Variables:

Ratios de la concentration relative de 19 ARNm cibles par rapport à un ARNm référence

Voir fichier qPCRmultipletesting.Rmd

Package multtest, limma, p.adjust