# Analyse de données compositionnelles par recherche de sous-graphes disjoints

Emmanuel Curis, Bruno Saubaméa, Cynthia Marie-Claire

# Introduction — Données compositionnelles

- Données sur la composition d'un système
  - ⇨ K constituants, 1 à K — $q_i$ : quantité du $i$-ème
- Exprimées en fraction d'un tout
  - ⇨ Elles sont corrélées : somme imposée !
  - ⇨ Changer l'une modifie toutes les autres

## *Applications en chimie, biologie...*

- Composition d'un mélange : fractions molaires, massiques
  - ⇨ Composition des minéraux en oxydes (20 % de $SiO_2$...)
  - ⇨ Composition des huiles en acides gras
  - ⇨ Composition en A. R. N. d'une cellule

# Introduction — Origine de ces données

## *Contrainte intrinsèque du système...*

- Une quantité prédéfinie de matière se répartit entre diverses possibilités
  - ⇨ Spéciation d'un élément en quantité totale imposée
  - ⇨ Traceur entre divers organes, tissus...

## *Contrainte expérimentale « externe »*

- Chaque constituant est libre de varier
- Le dosage utilise une quantité imposée du mélange
  - ⇨ Dosage de fractions, plus de quantités
  - ⇨ A. R. N., protéines...
  - ⇨ Populations lymphocytaires

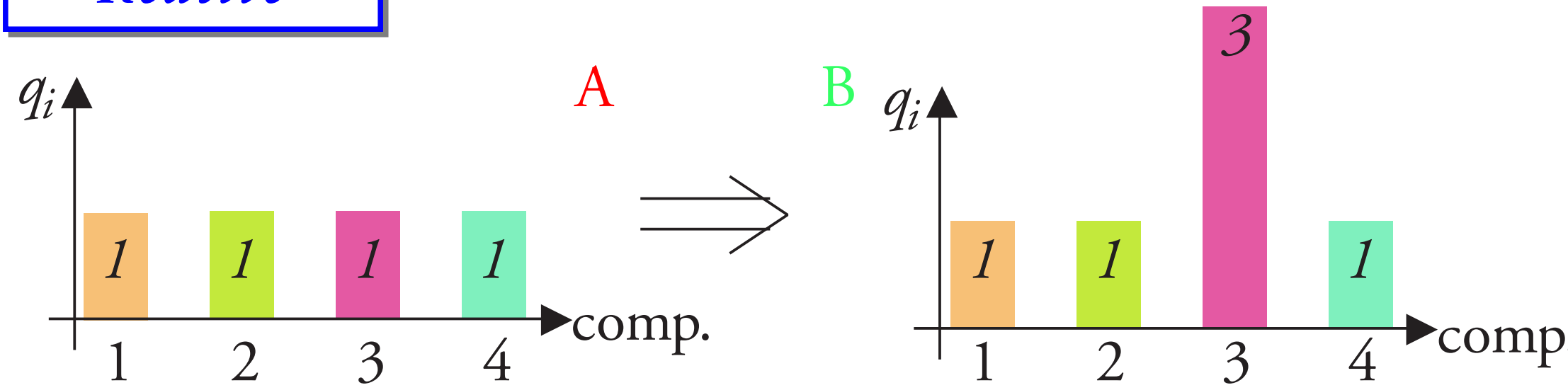
# Introduction — Exemple des données d'expression

- Des cellules sont isolées, mise en culture…
  - ⇨ K A. R. N. différents, [ARN $i$] = $q_i$
- Les A. R. N. sont extraits de la culture
  - ⇨ K A. R. N. différents, [ARN $i$] = $q_i$
- On isole une masse totale M d'A. R. N. pour quantification
  - ⇨ K A. R. N. différents, [RNA $i$] = $x_i = \mathrm{M} \dfrac{q_i}{\sum_{j=1}^{\mathrm{K}} q_j}$
- K* < K A. R. N. différents sont quantifiés
  - ⇨ K* A. R. N. différents, [ARN $i$] = $d_i = \lambda_i x_i = \lambda_i \mathrm{M} \dfrac{q_i}{\sum_{j=1}^{\mathrm{K}} q_j}$

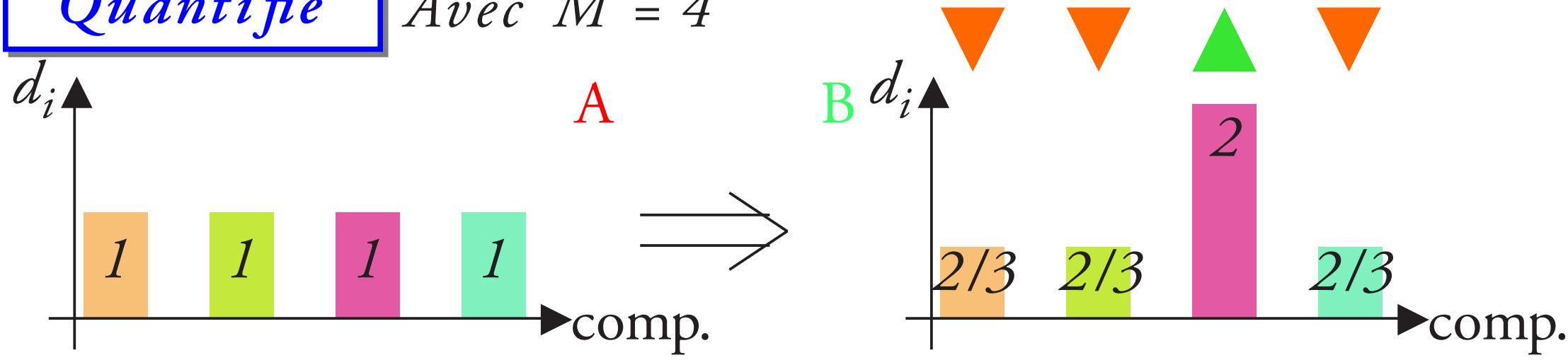
# Illustration des données compositionnelles ①

★ K = 4 composants différents, 2 conditions A & B

**Réalité**



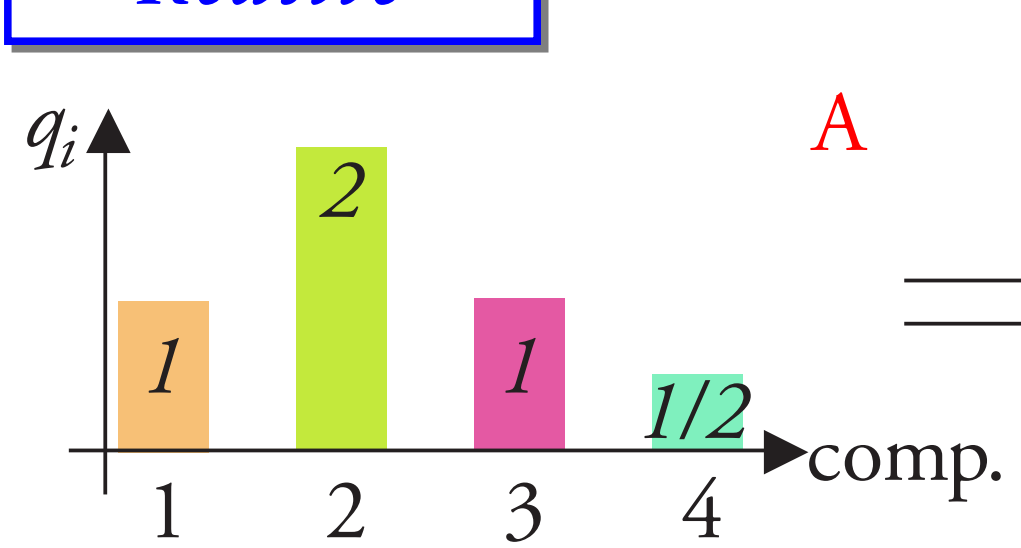

**Quantifié** Avec $M = 4$



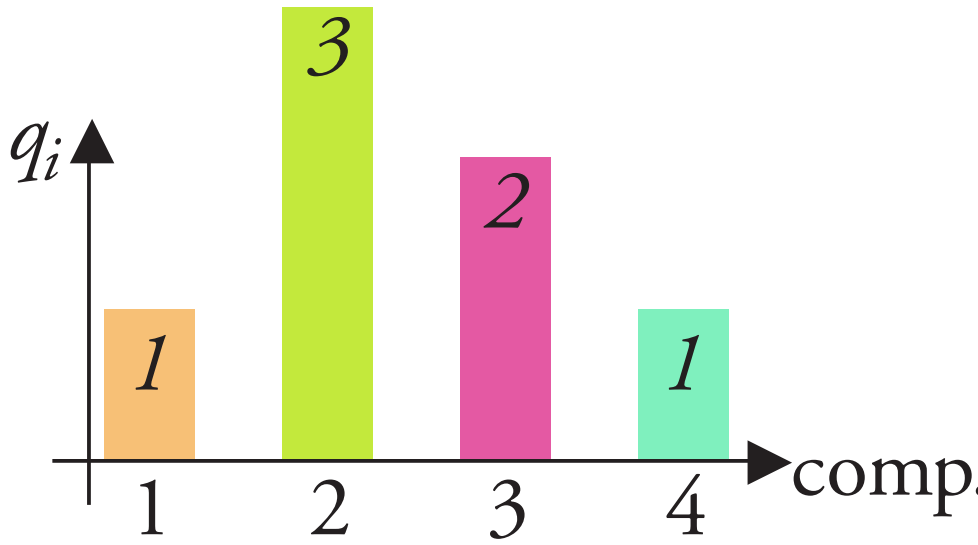
# Illustration des données compositionelles ②

★ K = 4 composants différents, 2 conditions A & B

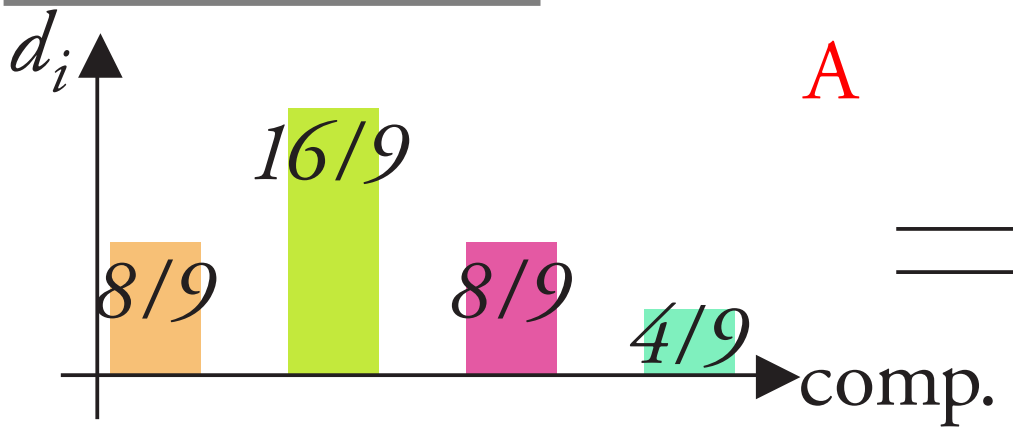
**Réalité**



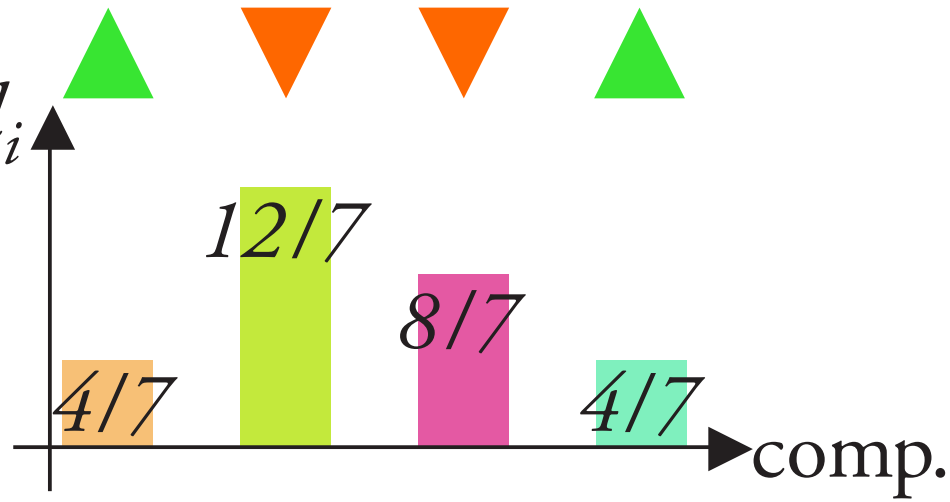

A ⟹ B




**Quantifié** Avec $M = 4$




A ⟹ B



# Méthodes classiques — « alr »

- Données brutes non-analysables
  - ⇨ Corrélées, sens de variation fortuit…

Aitchison, 1982 ; 1984…
Egozcue, 2003
Filzmoser, 2009
…

- On choisit un constituant « de référence » (p. ex. le 1)
- On calcule $r_{i,1} = q_i/q_1$ pour les K – 1 autres constituants
- On travaille sur $\ln r_{i,1}$
  - ⇨ Méthode du log des rapport : « alr »
- Les résultats dépendent de la référence choisie
  - ⇨ Comment choisir la référence ?
- Le nombre de tests augmente avec K : perte de puissance !

# Méthodes classiques — « clr »

- On pose $m_g$ la moyenne géométrique des K $q_i$
- On calcule $r^c_{i,c} = q_i / m_g$ pour les K constituants
- On travaille sur $\ln r^c_{i,c}$

  ⇨ Méthode du log des rapport centrés : « clr »
- Les $r^c_{i,c}$ sont toujours compositionnels !
- $r^c_{i,c}$ dépend de tous les constituants : s'il change, pourquoi ? est-ce dû au constituant $i$ ?
- Donne une information globale : « la composition a changé »

# Méthodes classiques — « ilr »

- On pose $m_{g,k}$ la moyenne géométrique des $k$ derniers $q_i$
- On calcule $r_i^i = q_i / m_{g,i+1}$ pour les K – 1 premiers constituants
- On travaille sur $[(K - i)/(K - i + 1)]^{0,5} \ln r_i^i$

  ⇨ Méthode du log des rapport isométrique : « ilr »
- Transformation complexe qui assure l'indépendance
- $r_i^i$ dépend des K – $i$ constituants suivants : s'il change, pourquoi ? est-ce dû au constituant $i$ ?

  ⇨ Information globale : « la composition a changé »

# Rationnel : les rapports ne sont pas faussés

$$x_i = \frac{q_i}{\sum_{k=1}^{K} q_k}$$

- Soient deux composants, $i$ ($q_i$) et $j$ ($q_j$)
- Rapport dans l'échantillon père : $r_{i,j} = \frac{q_i}{q_j}$
- Rapport dans l'échantillon quantifié : $r^*_{i,j} = \frac{x_i}{x_j} = \frac{q_i}{q_j} = r_{i,j}$
- Rapport quantifié : $r^{Q}_{i,j} = \frac{\lambda_i\, x_i}{\lambda_j\, x_j} = \frac{\lambda_i}{\lambda_j} r_{i,j}$

## Conclusion

- L'information sur les quantités relatives persiste
  - ⇨ Peut être celle cherchée (équilibre chimique…)
- Comment interpréter cette information sinon ?
  - ⇨ Que signifie un changement du rapport moyen ?

# Interpréter les modifications des rapports ①

- Exemple : 2 composants, 3 changements possibles chacun

| composant $j$ \ composant $i$ | *Inchangé* | *Doublé* | *Divisé par 2* |
|---|---|---|---|
| *Inchangé* | $r_{i,j}$ inchangé | $r_{i,j}$ doublé | $r_{i,j}$ divisé par 2 |
| *Doublé* | $r_{i,j}$ divisé par 2 | $r_{i,j}$ inchangé | $r_{i,j} \times 1/4$ |
| $\times$ *1/2* | $r_{i,j}$ doublé | $r_{i,j} \times 4$ | $r_{i,j}$ inchangé |

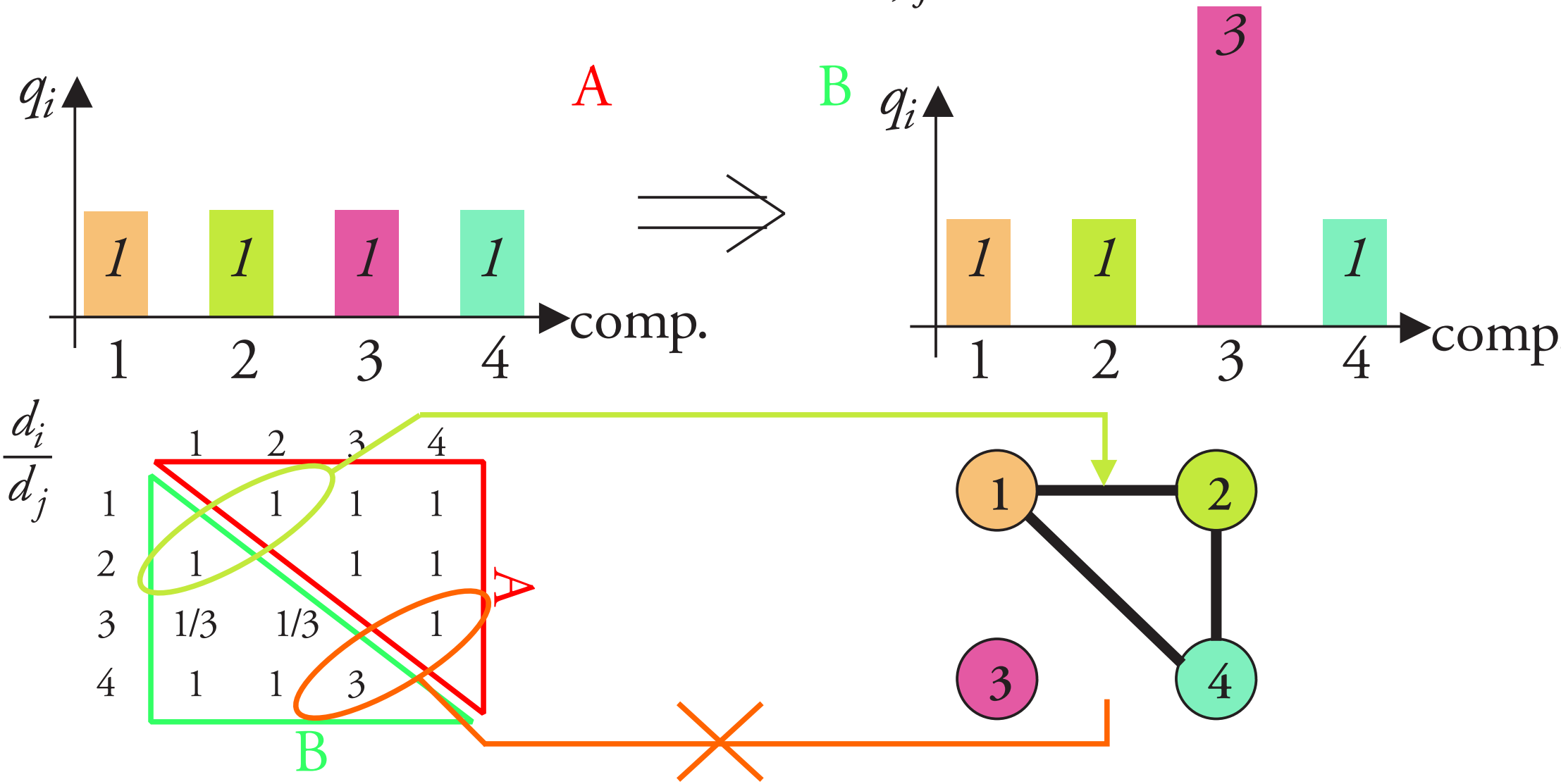
**Le rapport est inchangé**

- Les quantités des deux composants sont inchangées…
- … ou les deux ont été modifiées par le même facteur.

Si le passage de A à B ne modifie pas $r_{i,j}$, alors il a le même effet sur les quantités des composants $i$ et $j$.

# Construire un graphe des composants quantifiés

- Nœuds du graphe : les K* composants quantifiés
- Les nœuds $i$ et $j$ sont reliés ssi $r_{i,j}$ est inchangé



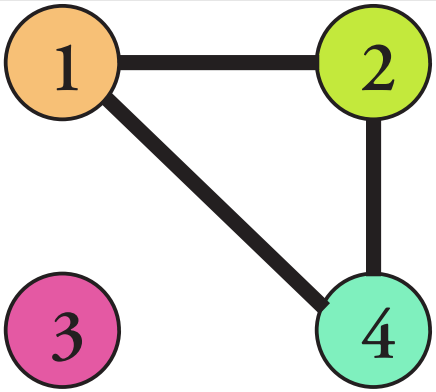
# Interpréter les changements des rapports ②

- Plusieurs sous-graphes disjoints
  - ⇨ Chacun est complètement connexe
- Chaque sous-graphe correspond à une variation différente
  - ⇨ Au plus un pour « pas de changement »
  - ⇨ Mais impossible de savoir lequel…

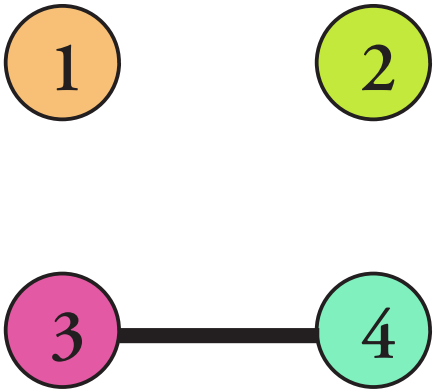
*Limitation des données compositionelles*

- En pratique, les résultats ne sont pas aussi tranchés
  - ⇨ Connexions indues entre les nœuds…
  - ⇨ Connexions absentes entre les nœuds…

**Exemple 1**




**Exemple 2**



# Déterminer des groupes de composants

*Comment gérer les fausses connexions ?*

- Ne considérer que les sous-graphes disjoints
  - ⇨ même s'il manque des connexions dedans
  - ⇨ très sensible aux variations non-détectées
- Ne considérer que les ensembles connexes de nœuds
  - ⇨ cliques et cliques maximales
  - ⇨ très sensible aux fausses variations
  - ⇨ calculs longs pour les grands graphes (RNAseq...)
- Recherche de communautés
  - ⇨ Plusieurs définitions & algorithmes...

# Déterminer si un rapport a été modifié

- ***Étape clef de la méthode***
  - ⇨ Détermine la structure du graphe !
- Estimer la variation du rapport
  - ⇨ Modèle statistique adapté au plan expérimental
  - ⇨ Variation cherchée : l'un des paramètres du modèle, $\theta$
  - ⇨ Typiquement : modèle log-linéaire
- Les nœuds $i$ et $j$ sont déliés si $r^{\mathrm{B}}_{i,j} / r^{\mathrm{A}}_{i,j}$ est significativement différent de 1
  - ⇨ Dans le modèle, si le test de $\theta$ est significatif
- Quel niveau (« seuil de $p$ ») utiliser pour ce test ?

# Choix du niveau du test — sous-graphes disjoints ①

- Observation : « Au moins deux graphes disjoints »
  - ⇨ Si aucun changement, doit arriver avec prob. $< \alpha$ ($H_0$)
  - ⇨ Quel niveau $\alpha_0$ utiliser dans le test du rapport ?

- Se produit si un nœud (le 1) n'a aucune connexion
  - ⇨ Si *tous* les rapports $r_{1,\,j}$ sont significativements modifiés
  - ⇨ Aucune correction de multiplicité nécessaire

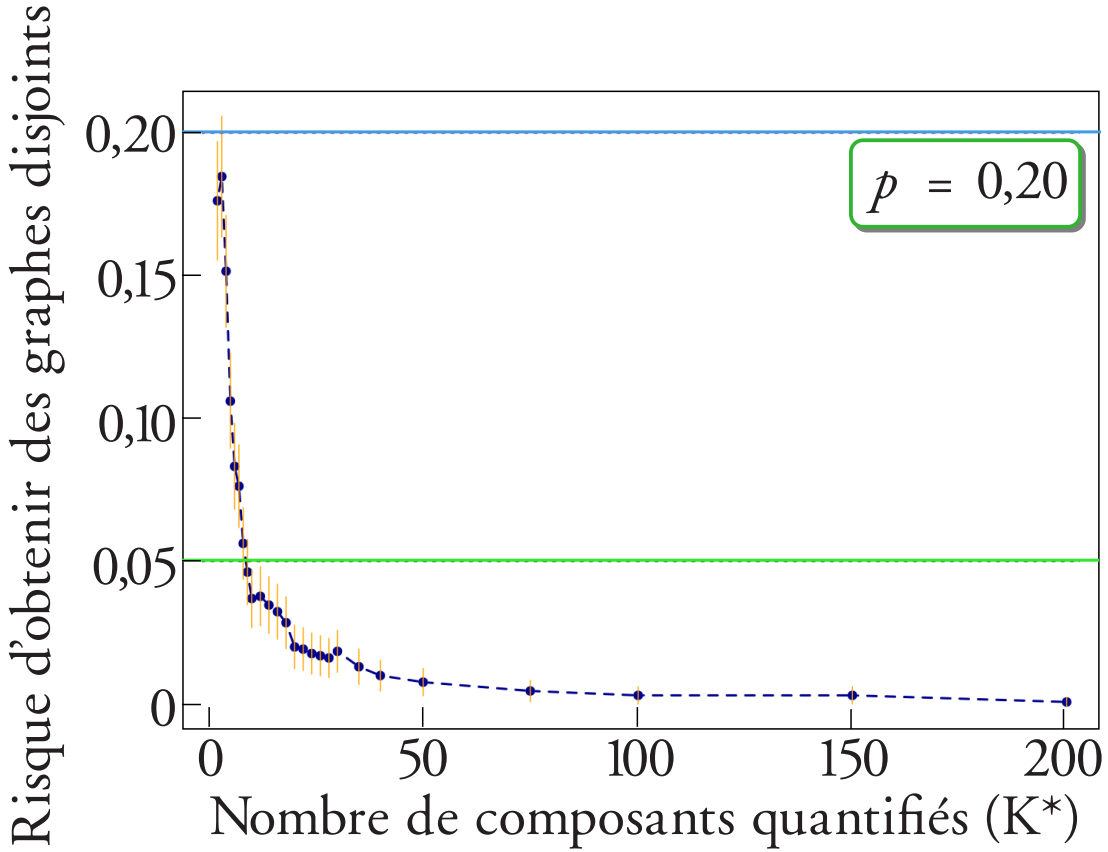
- Si les tests étaient indépendants, $\Pr(\text{TSM}|H_0) = \alpha = \alpha_0^{K^*-1}$
  - ⇨ $\alpha_0$ doit être (bien) plus grand que $\alpha$

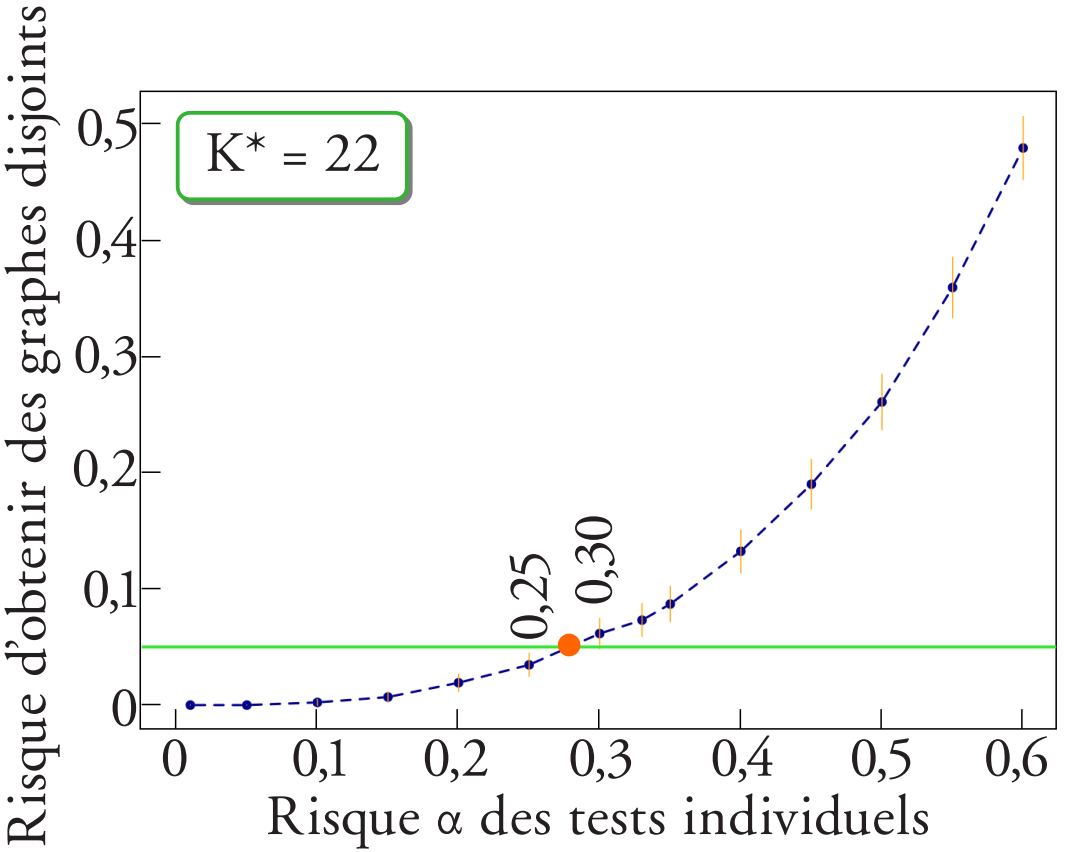
- Les tests ne sont **pas** indépendants. Et doit être vrai pour tous les nœuds…

# Choix du niveau du test — graphes disjoints ②

## *Résultats de simulation*

- Valeurs log-normales, sous $H_0$, somme valant 1
- 10 000 simulations, avec K = 200 composants



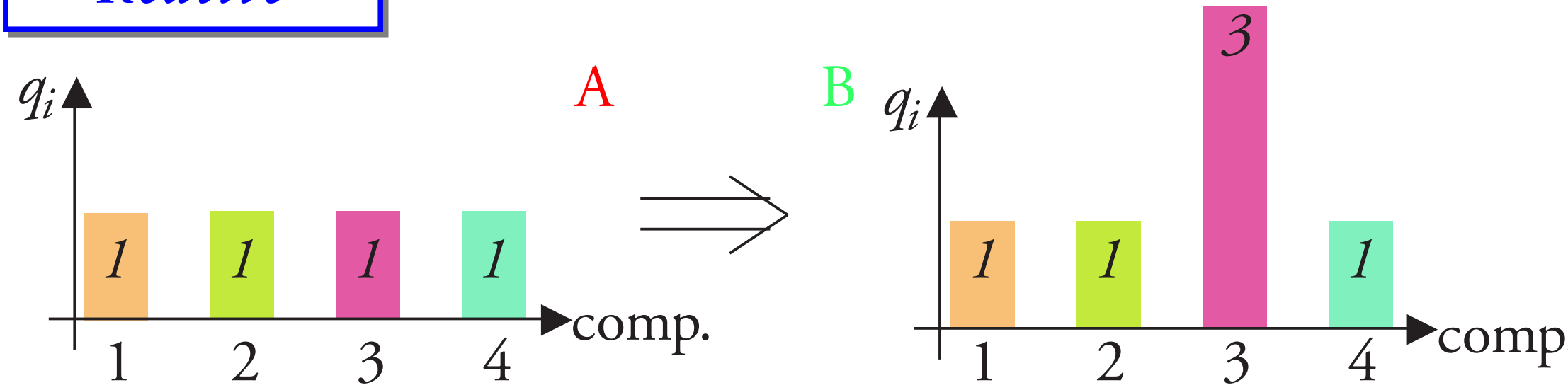
# Et pour la puissance ?

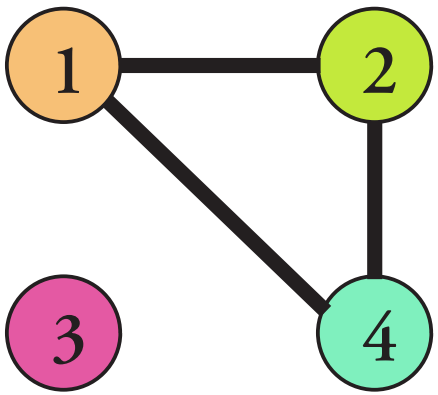
- Simulations précédentes : risque de détecter des composants se comportant différemment, quand tous se comportent de façon identique (« rejet de $H_0$ quand elle est vraie »)
- Comment se comporte la méthode quand certains composants se comportent réellement différemment ?
  - ⇨ Détecte-t-elle des graphes disjoints ? (*puissance*)
  - ⇨ Détecte-t-elle les bons groupes de composants ?
- Dépend de la puissance de chaque test individuel
- Difficile de comparer à la méthode classique alr
  - ⇨ Posent des questions légèrement différentes...

# Étude de l'exemple 1 — conditions de simulation

- K = 4 composants, 2 conditions A & B, 1 composant triple

***Réalité***



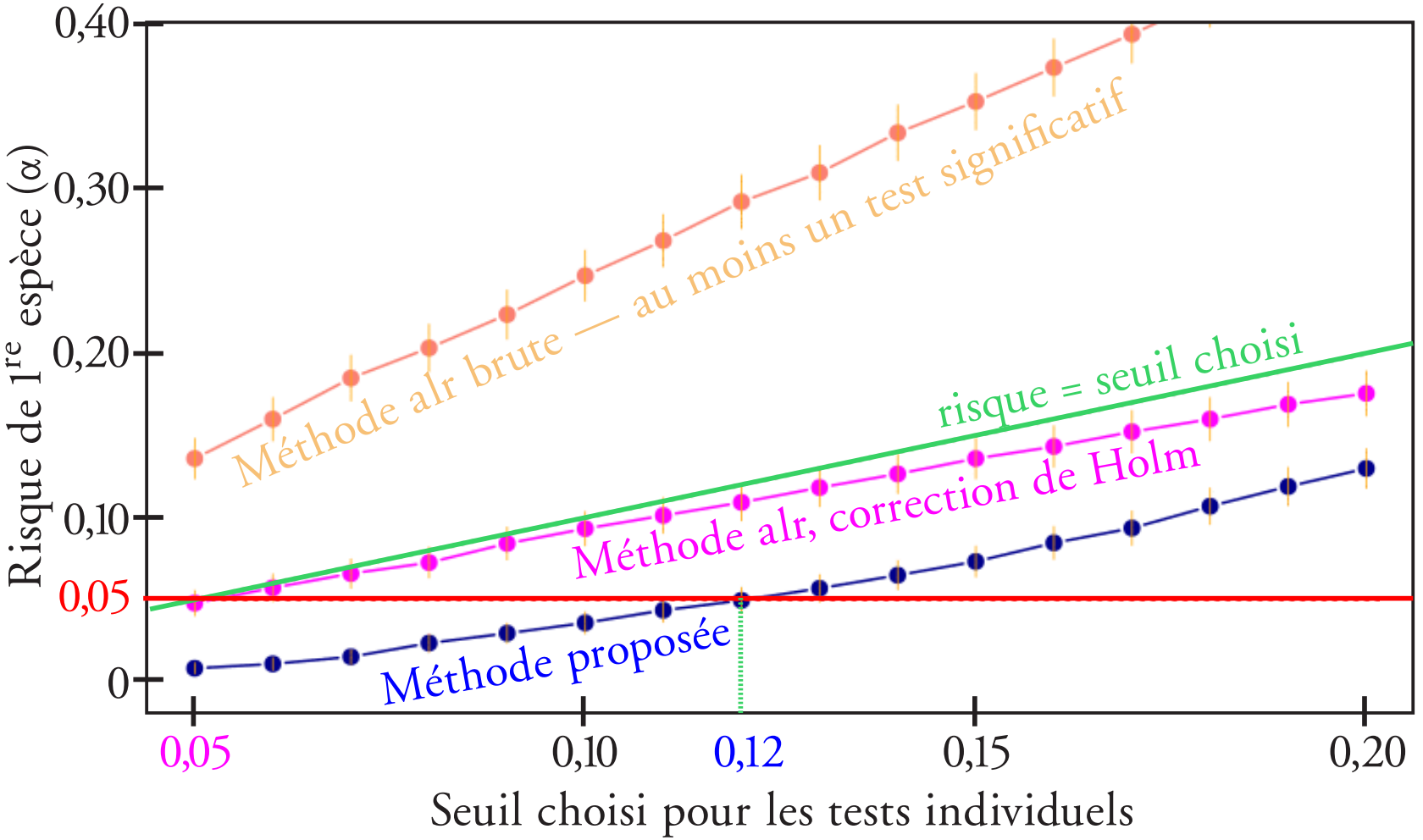




Graphe théorique

- Distribution log-normale ; CV = 20 %
- $n$ = 3 par condition (puissance individuelle > 80 %)
- Méthode alr : référence = composant 1

# Étude de l'exemple 1 — choix du seuil (risque α)

- La méthode alr « brute » ne contrôle pas α
  - ⇨ Correction de multiplicité (ici, Holm) indispensable

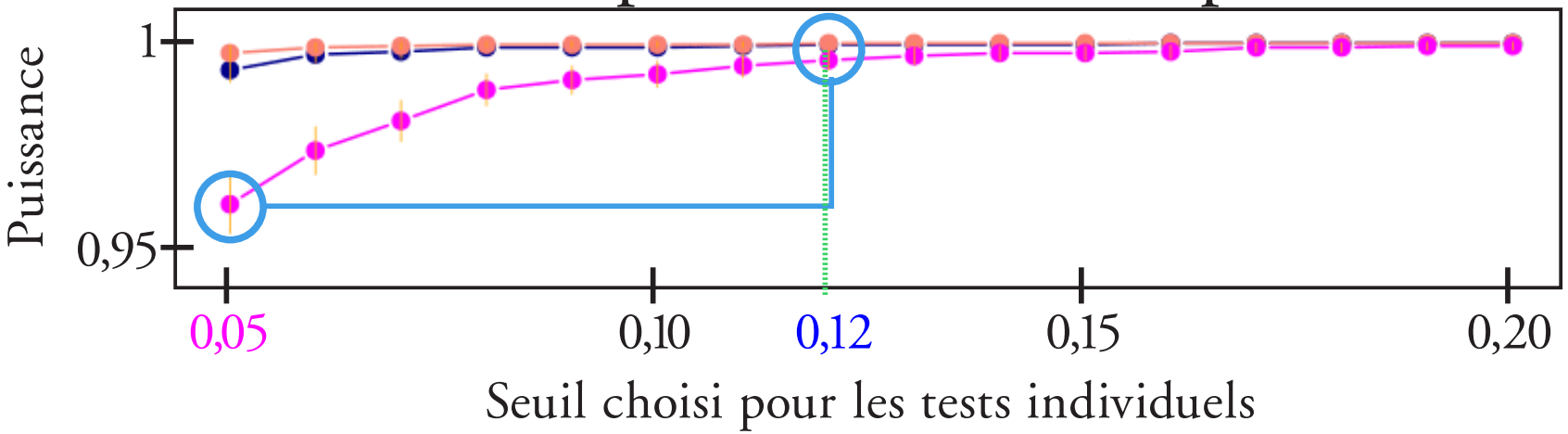


# Étude de l'exemple 1 — étude de la puissance totale

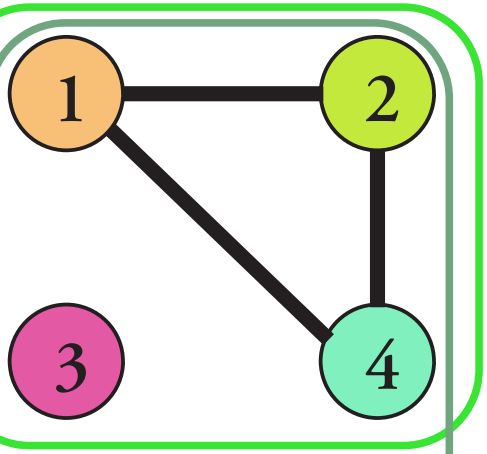
- On rejette $H_0$ pour n'importe quelle raison
  - ⇨ Il se passe quelque chose (pas forcément ce qui est observé !)

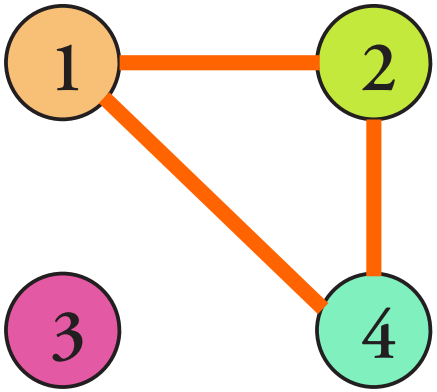
- Méthode alr : composants 2 ou 3 ou 4 détectés
  Méthode proposée : graphe disjoint, quel qu'il soit

- En pratique les deux méthodes sont aussi puissantes
  - ⇨ Mais la correction de multiplicité en alr a un prix !



# Exemple 1 — trouve-t-on le bon composant ?

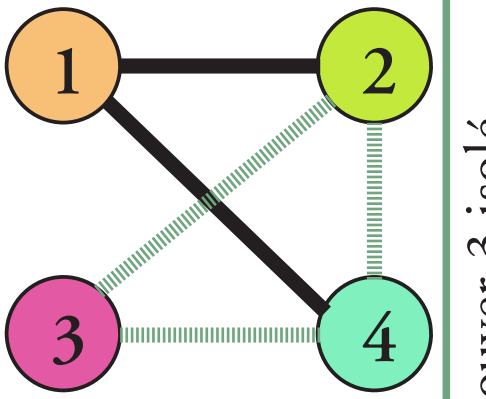



## Méthode proposée

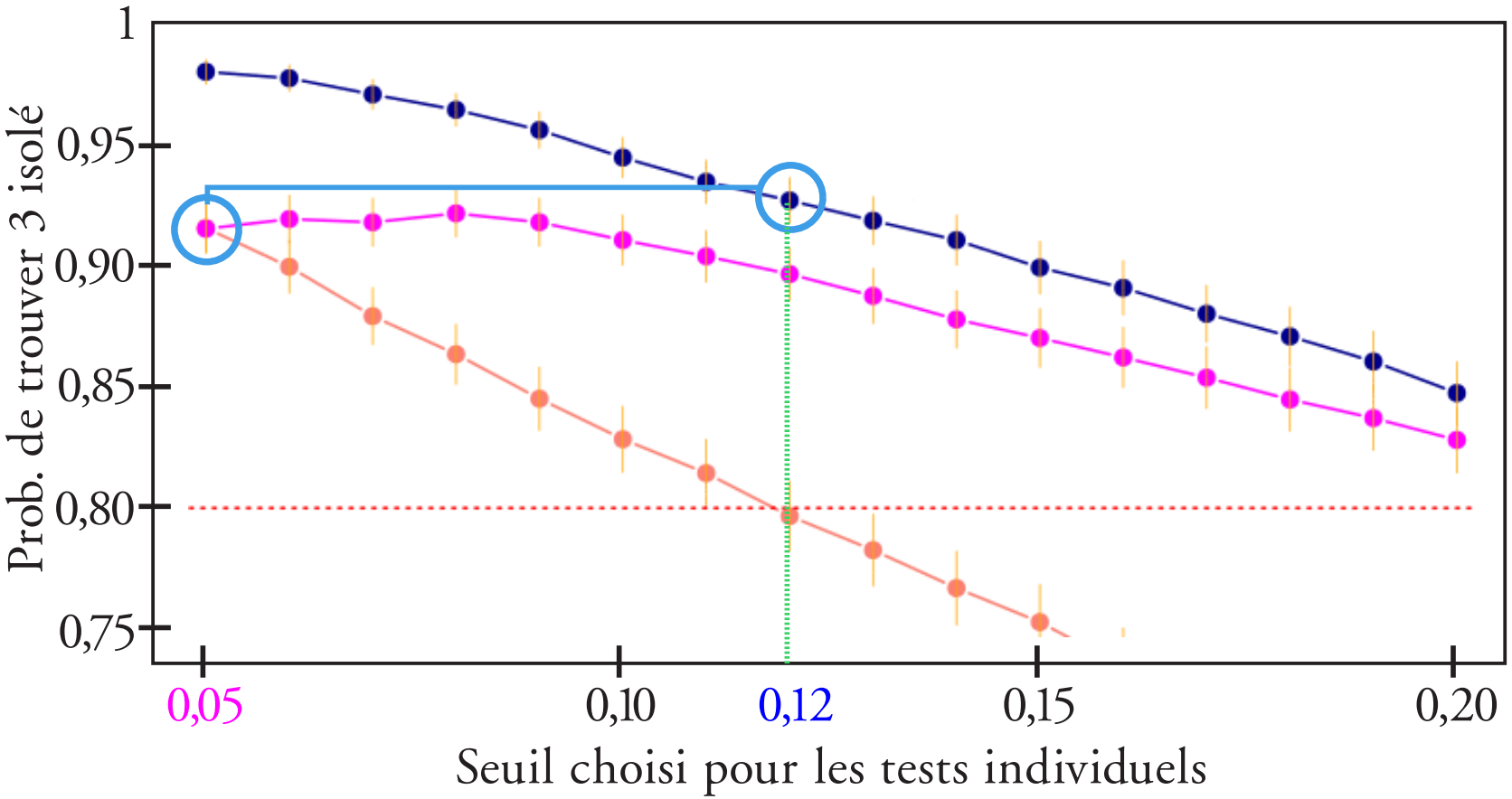
Une seule de ces arêtes peut manquer, au plus, pour conclure correctement

Soit 4 graphes acceptables (sur $2^6$ possibles)



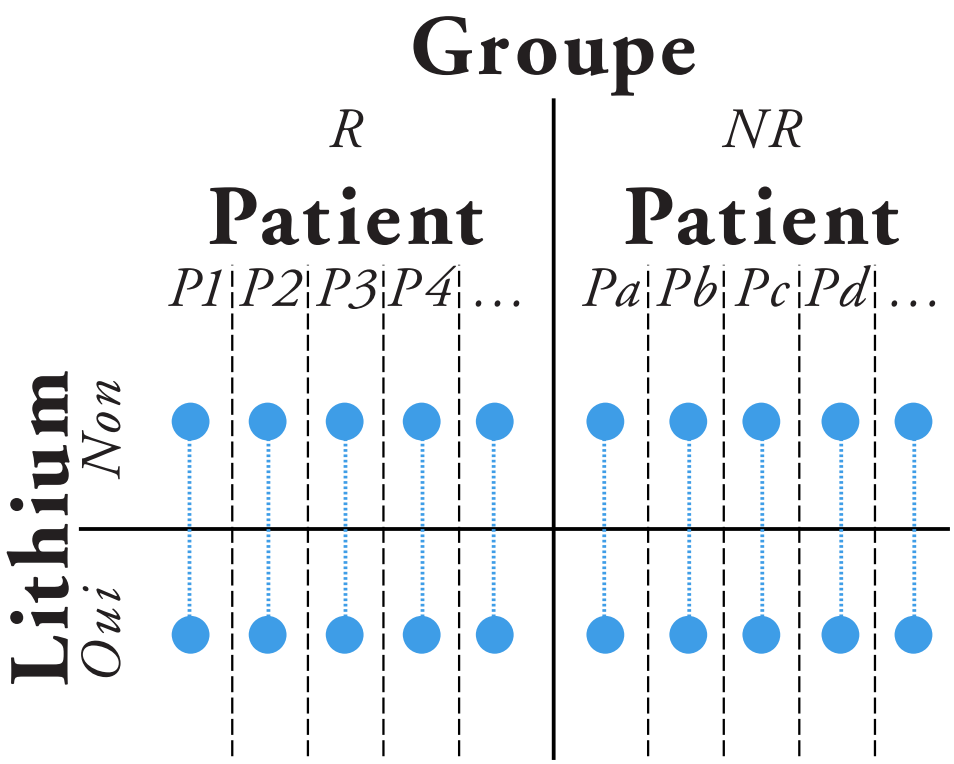

## Méthode alr

Ces arêtes peuvent, ou non, exister — la conclusion est identique



# Application en qRT-PCR: un exemple ①

- 2 groupes de patients bipolaires (répond [R, $n = 19$] ou non [NR, $n = 19$] au lithium)
- Leurs lymphocytes sont cultivés avec et sans lithium
- 17 candidats + 2 références




Geoffroy et coll., 2017

## *Modèle*

$$\ln r_{i,j} = \mu_0 + U_P + \delta_R \mathbf{1}_R + \delta_{Li} \mathbf{1}_{Li} + \delta_I \mathbf{1}_R \mathbf{1}_{Li} + \varepsilon$$

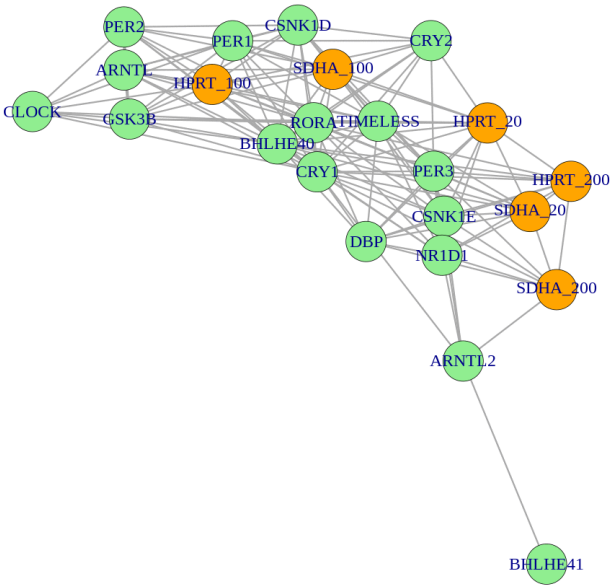
- $U_P$ : *Effet patient*
- $\delta_R \mathbf{1}_R$ : *Différence basale*
- $\delta_{Li} \mathbf{1}_{Li}$ : *Effet du lithium chez les patients NR*
- $\delta_I \mathbf{1}_R \mathbf{1}_{Li}$ : *Effet additionnel du lithium chez les patients R*

# Application en qRT-PCR : un exemple ②

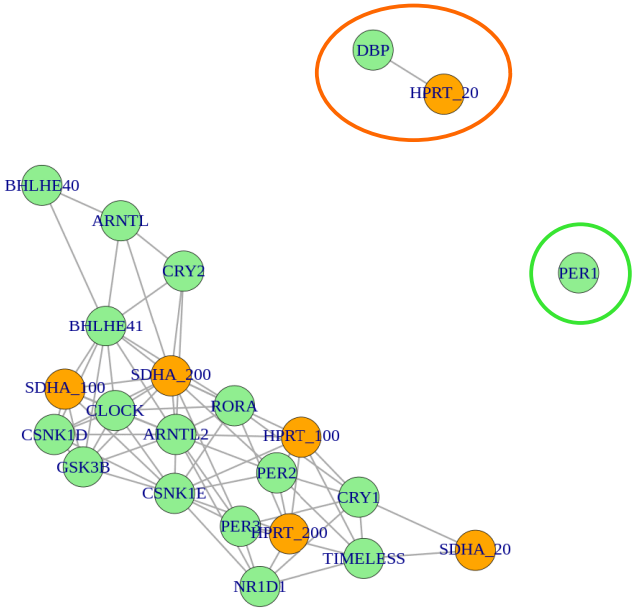
- Modèle linéaire à effets mixtes, avec lme4 (R)
  - ⇨ Coefficients testés par test de Wald asymptotique

- Critère : sous-graphes disjoints
- K* = 23 nœuds ⇨ $\alpha_0$ = 0,25 pour avoir $\alpha$ < 0.05

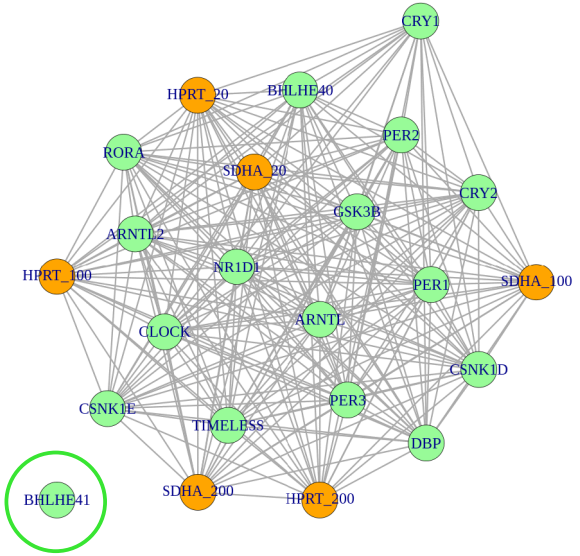
## Comparaison R vs NR, basal




## Effet de Li⁺ chez les NR




## Effet différentiel de Li⁺



# Quelques limitations de la méthode...

- Nombre de tests augmente avec K* comme $K^{*2}$
  - ⇨ Temps et mémoire nécessaires augmentent « vite »
  - ⇨ Peut être problématique en RNAseq ($K^* \approx 2 \times 10^4$)
- Taille du graphe augmente avec K*
  - ⇨ Temps d'analyse du graphe peuvent devenir longs (cliques, communautés...)
- Impossible de savoir comment varie un gène donné
  - ⇨ Limite du plan expérimental, pas de la méthode...
- Gestion des valeurs non-quantifiables délicate
  - ⇨ Division par 0...

# … et quelques avantages

- Pas besoin d'hypothèse sur des gènes de référence invariants
- Pas de correction de multiplicité
  - ⇨ Moins de perte de puissance quand K* augmente…
- Rapports au sein d'une préparation
  - ⇨ Insensible aux erreurs sur la masse M prélevée…
- Insensible aux différences d'efficacité de quantification entre A. R. N.
  - ⇨ tant qu'elles ne dépendent pas des conditions comparées !
- RNASeq : insensible aux profondeurs de séquençage…

*Merci de votre attention !*