Diplôme Universitaire en Bioinformatique Intégrative (DU-Bii)

Module 3 : Analyse statistique avec R

Séance 5: Détection de gènes différentiellement exprimés

Teachers:

- Jacques van Helden
- Olivier Sand

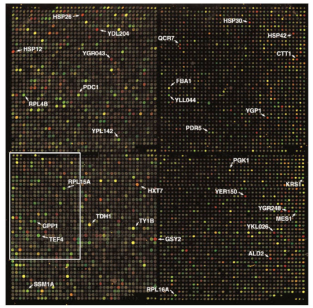
Short link: http://tinyurl.com/dubii19-m6-intro

Article pionnier de la transcriptomique (biopuces)

- **DeRisi, J.L. (1997).** Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. Science 278, 680–686.
- Méthodologie
  - Puces à 2 canaux (rouge / vert), co-hybridation de 2 échantillons (test vs contrôle).
  - Intensité reflète la concentration d'ARN.
  - Couleur reflète **expression différentielle**
    - Rouge: sur-exprimé
    - Vert: sous-exprimé
- But: caractériser la réponse transcriptionnelle des gènes lors d'un changement de milieu de culture.
  - Objet d(intérêt : gène,
  - Variables: mesures d'expression au cours du temps
- Preuve de concept: **shift diauxique**
  - Série temporelle (19h) croissance démarre sur milieu riche, qui s'appauvrit progressivement.
  - **Figure 4.** En partant de listes de gènes fonctionnellement liés, on constate que certains augmentent et d'autres diminuent en fonction de l' évolution du milieu (Figure 4).
  - **Figure 5.** En partant des profils on peut détecter des groupes de co-expression.

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*



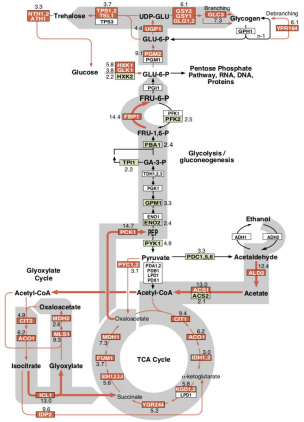


Figure 3: metabolic reprogramming during diauxic shift

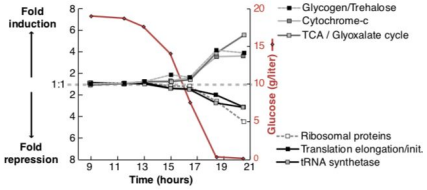


Figure 4: coordinated regulation of functionally related genes.

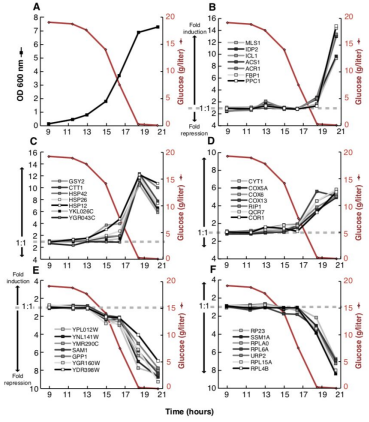
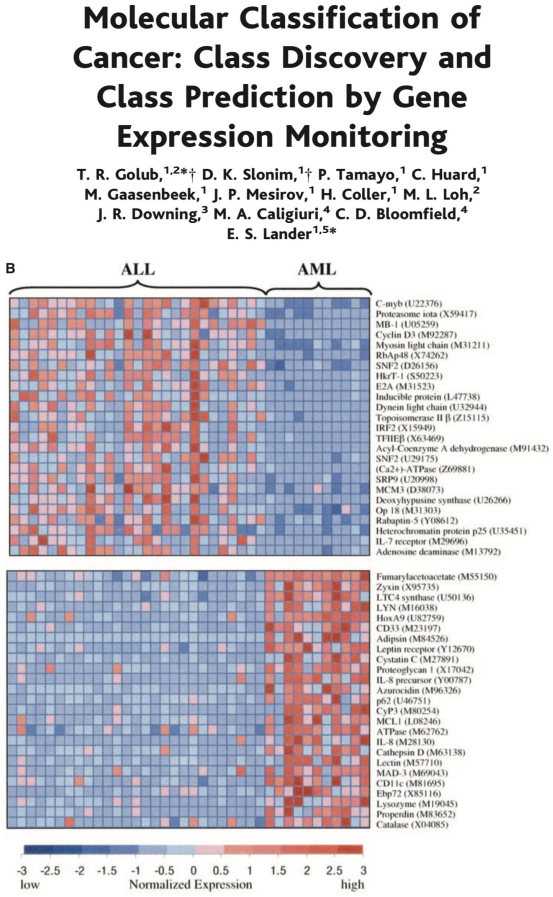


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties.

Classification des cancers par biopuces transcriptomiques

- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.
- Classification des types de cancers sur base de "***signatures transcriptomiques***".
- Approche supervisée.
  - Individus: échantillons
  - Variables: gènes
  - Jeu d'entraînement
    - 27 échantillons ALL (Acute Lymphoblastic Leukemia)
    - 11 échantillons AML (Acute Myeloblastic Leukemia)
- Sélection de variables
  - 25 gènes sur-exprimés dans un groupe
  - 25 gènes sur-exprimés dans l'autre.
- Construction d'une **fonction discriminante** pour prédire le statut de nouveaux échantillons.



Biopuces: classification des types de biopuces

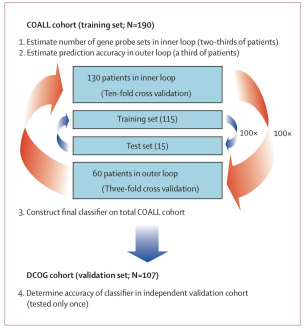
- Une étude illustrative: Den Boer et al. (2009) Lien entre sous-types de leucémies (en fonction des mutations associées, signature transcriptomiques, et résultats d'un traitement.
- Protocole plus élaboré
  - Sélection de variables basée sur test de Wilcoxon pour chaque sous-type versus les autres
  - 50 itérations d'apprentissage / test
  - Cohorte indépendante pour validation
  - Classification non-supervisée pour comparer classes connues (marge) et découvertes.
- La carte de température montre un **bi-clustering** des classes (colonnes) et des échantillons (ligne).
- Source: Den Boer et al. (2009). A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. *Lancet. Oncology* vol. 10,2: 125-34.

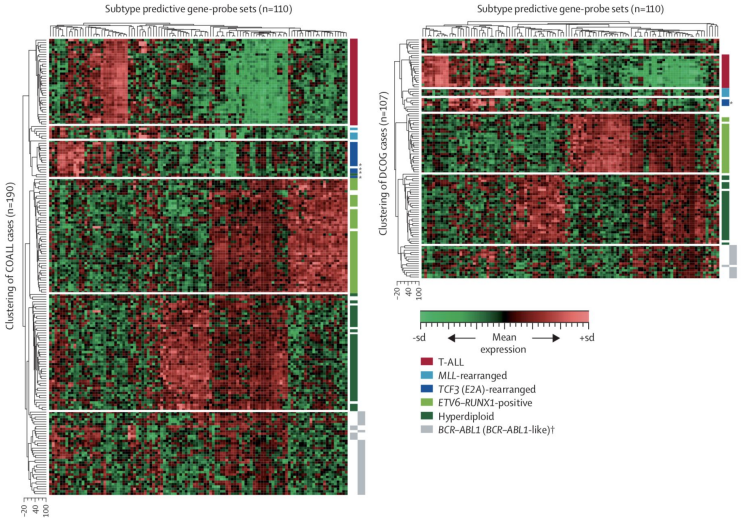
Note: Nous reviendrons sur la classification supervisée lors de la 6ème séance de ce module.

A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study

Monique L Den Boer, Marjon van Slegtenhorst*, Renée X De Menezes, Meyling H Cheok, Jessica G C A M Buijs-Gladdines, Susan T C J M Peters, Laura J C M Van Zutven, H Berna Beverloo, Peter J Van der Spek, Gaby Escherich†, Martin A Horstmann†, Gritta E Janka-Schaub†, Willem A Kamps‡, William E Evans, Rob Pieters‡*

Den Boer et al. (2009). Lancet. Oncology vol. 10,2: 125-34.





Des biopuces au RNA-seq

- Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research 18, 1509–1517.
- Comparaison de biopuces et RNA-seq
- Test: foie versus rein de souris
- Résultats
  - Les fold-changes diffèrent, surtout pour les faibles comptages (Figure 4)
  - Les listes de gènes différentiellement exprimés diffèrent (Figure 5)
  - Plus y'a de répliques mieux ça vaut (Table 1)
- **Leurs conclusions**
  - Forte reproductibilité
  - Les données se prêtent à un modèle de Poisson
- ***Mais ...***
  - Il ne s'agissait que de répliques techniques (le même échantillon passé plusieurs fois au séquençage ou sur biopuces)
  - Les études ultérieures montrent qu'entre "répliques biologiques" (échantillons différents résultant d'expériences séparées) la variabilité est bien plus élevée.

RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

John C. Marioni,[1,6] Christopher E. Mason,[2,3,6] Shrikant M. Mane,[4] Matthew Stephens,[1,5,7] and Yoav Gilad[1,7]

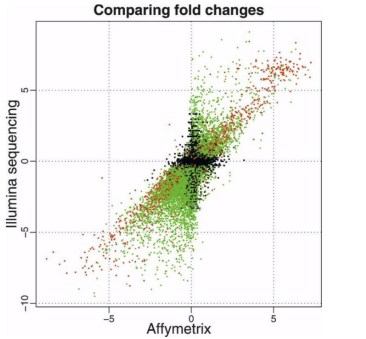


Figure 4. Comparison of estimated $\log_2$ fold changes (liver/kidney)

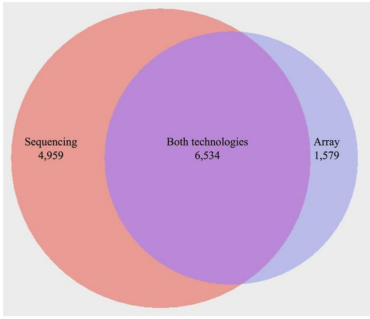


Figure 5. A Venn diagram summarizing the overlap between genes

Table 1. Power to detect differentially expressed genes depends on the number of lanes used for each sample

No. of lanes compared	Differentially expressed genes	Overlap with genes called from the array	Correlation of fold changes between the sequence data and the array
One vs. one	5670	4208	0.67
Two vs. two	7994	5340	0.70
Three vs. three	9482	5909	0.71
Four vs. four	10,580	6278	0.72
Five vs. five	11,493	6534	0.73

The first column shows, for different numbers of lanes, the average number of genes called differentially expressed between the liver and kidney samples (at an FDR of 0.1%) from the Illumina sequencing data. The average overlap with the 8113 genes identified as differentially expressed from the Affymetrix array data is also shown, as is the average correlation between the estimated $\log_2$ fold changes.

We find that the sequencing data are highly reproducible, with few systematic differences among technical replicates. Statistically, we find that the variation across technical replicates can be captured using a Poisson model, with only a small proportion (~0.5%) of genes showing clear deviations from this model. This Poisson model can be used to identify differentially expressed genes, and using this approach, the sequence data identified 30% more differentially expressed genes than were obtained from a standard analysis of the array data at the same false discovery rate. We also illustrate the potential for sequence-based approaches to identify alternative-spliced forms.

Particularités des comptages RNA-seq

- Non-normalité des donnés: comptages = variables discrètes
  - → *Tests paramétriques (Student, Welch) ne s'appliquent pas*
- Le nombre de "répliques" (en stat: taille des échantillons) est souvent très faible (3 réplicats par conditions, voire moins).
  - → *Faible puissance des tests non-paramétriques*
- Présence de valeurs très très fortement aberrantes ("outliers")
  - → *normalisation des librairies demande traitement particulier*
- Présence d'un très grand nombre de valeurs nulles
  - → *"Zero-inflated distributions"*
- Sur-dispersion des données par rapport à un modèle de Poisson
  - → *La distribution de Poisson ne s'applique pas*
- Les lectures (short reads) résultent d'un échantillonnage aléatoire des transcrits
  - → *A concentration égale, les petits gènes sont associés à des comptages plus faibles que les grands*

Cas d'étude: sporulation de la levure du boulanger

- Sporulation chez la levure
- 2 génotyypes: sauvage versus mutant
- Série temporelle 3 points (0h, 4h, 8h)

Series GSE89530	Query DataSets for GSE89530
Status	Public on Jan 01, 2017
Title	RNA-SEQ analysis of bdf1-Y187F-Y354F mutant strain during yeast sporulation
Organism	Saccharomyces cerevisiae
Experiment type	Expression profiling by high throughput sequencing
Summary	The aim of this study is to compare NGS-derived yeast transcriptome profiling (RNA-seq) of wild-type and bdf1-Y187F-Y354F mutant strains after sporulation induction (time points: 0h 4h and 8h).
Overall design	S. cerevisiae wild-type and bdf1-Y187F-Y354F mutant strains were collected 0h, 4h and 8h after sporulation induction in triplicates. mRNA were purifired, prepared and sequencedusing Illumina HiSeq 2000 sequencer.
Contributor(s)	Govin J, Garcia-Oliver E, Battail C, Boland A, Deleuze J
Citation(s)	García-Oliver E, Ramus C, Perot J, Arlotto M et al. Bdf1 Bromodomains Are Essential for Meiosis and the Expression of Meiotic-Specific Genes. *PLoS Genet* 2017 Jan;13(1):e1006541. PMID: 28068333

Design d'une analyse différentielle

- 2 groupes
  - $H_0: \mu_1 = \mu_2$
- Multi-groupes
  - $H_0: \mu_1 = \mu_2 = \ldots = \mu_g$
- Séries temporelles
- Combinaison de facteurs: effet de chaque facteur + interactions
- Échantillons appariés
- ...

GSE89530 f1-Y187F-Y354F mutant strain during yeast sporulation

Sample descriptions

ID	Condition	Replicate	Genotype	Time.point	Label
GSM2375281	bdf1_Y187F_Y354F_mutant_0	1	bdf1-Y187F-Y354F mutant	0	bd1bd2mut_0h_1
GSM2375287	bdf1_Y187F_Y354F_mutant_0	2	bdf1-Y187F-Y354F mutant	0	bd1bd2mut_0h_2
GSM2375293	bdf1_Y187F_Y354F_mutant_0	3	bdf1-Y187F-Y354F mutant	0	bd1bd2mut_0h_3
GSM2375278	Wild_type_0	1	Wild-type	0	Sc_WT_0h_1
GSM2375284	Wild_type_0	2	Wild-type	0	Sc_WT_0h_2
GSM2375290	Wild_type_0	3	Wild-type	0	Sc_WT_0h_3
GSM2375282	bdf1_Y187F_Y354F_mutant_4	1	bdf1-Y187F-Y354F mutant	4	bd1bd2mut_4h_1
GSM2375288	bdf1_Y187F_Y354F_mutant_4	2	bdf1-Y187F-Y354F mutant	4	bd1bd2mut_4h_2
GSM2375294	bdf1_Y187F_Y354F_mutant_4	3	bdf1-Y187F-Y354F mutant	4	bd1bd2mut_4h_3
GSM2375279	Wild_type_4	1	Wild-type	4	Sc_WT_4h_1
GSM2375285	Wild_type_4	2	Wild-type	4	Sc_WT_4h_2
GSM2375291	Wild_type_4	3	Wild-type	4	Sc_WT_4h_3
GSM2375283	bdf1_Y187F_Y354F_mutant_8	1	bdf1-Y187F-Y354F mutant	8	bd1bd2mut_8h_1
GSM2375289	bdf1_Y187F_Y354F_mutant_8	2	bdf1-Y187F-Y354F mutant	8	bd1bd2mut_8h_2
GSM2375295	bdf1_Y187F_Y354F_mutant_8	3	bdf1-Y187F-Y354F mutant	8	bd1bd2mut_8h_3
GSM2375280	Wild_type_8	1	Wild-type	8	Sc_WT_8h_1
GSM2375286	Wild_type_8	2	Wild-type	8	Sc_WT_8h_2
GSM2375292	Wild_type_8	3	Wild-type	8	Sc_WT_8h_3

Corrections pour les tests multiples

- ***P-valeur nominale*** (“non-ajustée”) et ***FPR (False Positive Rate)***
  - Probabilité d’obtenir, *sous hypothèse nulle*, un résultat aussi extrême que celui observé.
  - Pour un test bilatéral $Pvalue = P(X >= abs(x) \mid H_0)$
  - Interprétation: indication du risque de faux-positif (FPR) attaché à un **test individuel** (d’où la qualification de “nominale”)
- ***E-valeur*** (Expectation, Expected number of FP)
  - Nombre de faux-positifs attendus *sous hypothèse nulle*, dans une batterie de tests.

    $Evalue = \langle FP \rangle = FPR \bullet n_{tests}$
- ***FWER, Family-Wise Error Rate***
  - Probabilité d’obtenir au moins un FP dans une batterie de tests *sous* H_0.
  - $FWER = P(FP >= 1 \mid H0)$
- ***FDR, False Discovery Rate***
  - Taux attendu de FP *parmi les tests déclarés positifs* (parmi les éléments découverts)

Résultats d'une analyse différentielle

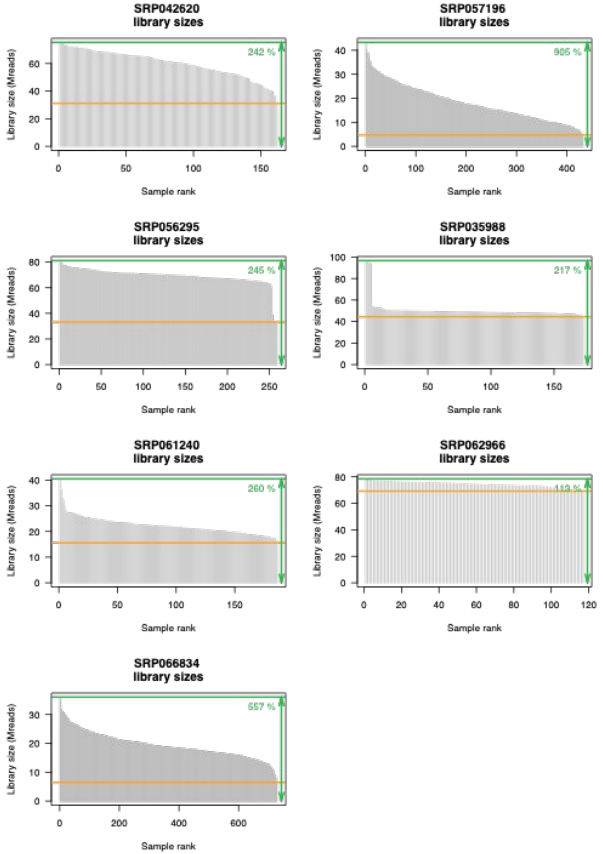
- Montrer un bout de tableau
- Expliquer le sens des différentes colonnes
- ***A faire en TP***
  (je ne veux pas vous spolier)

Plots diagnostiques

- Types de plots
  - MA plot
  - Volcano plot
  - Histogramme des P-valeurs
- Comparer résultats réels (sporulation chez la levure) et contrôles
  - Négatif: données artificielles sous H0
  - Positif: données artificielles sous H1
  - Mélange des deux (avec proportions contrôlées)
- ***A découvrir en TP***
  (je ne veux pas vous spolier)

Pourquoi faut-il normaliser les tailles de librairies ?

- Les tailles de librairies de séquences (profondeurs de séquençage) peuvent varier entre échantillons pour une même expérience.
-



Source: AbuElQumsan, Ghattas & van Helden, *in prep.*

Impact de la normalisation sur la détection de gènes différentiellement exprimés

- **A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis.**
  Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloe, Caroline Le Gall, Brigitte Schaeffer, Stéphane Le Crom, Mickael Guedj, Florence Jaffrezic and on behalf of The French StatOmique Consortium (2013). Brief. Bioinformatics 14, 671–683.
- Marie-Agnès Dillies (Institut Pasteur) et le consortium statomique (un tas de gens).
- Evaluation de 7 méthodes de normalisation sur la détection de gènes différentiellement exprimés.
- Jeux de données réels (haut)
  - Les boîtes à moustaches montrent les distributions de comptage après normalisation, pour 7 méthode.
  - Le dendrogramme (clustering hiérarchique) montre la similarité entre résultats obtenus par ces méthodes.
- Données simulées (bas)
  - Mélange de données sous hypothèse nulle (moyennes égales) et une proportion croissante de données sous alternative alternative (moyennes différentes).
- Résultats
  - Toutes les méthodes ont des sensibilités similaires, mais 2 méthodes ont un taux de faux-positif fiable: **DESeq** et **TMM** (edgeR).
  - Note: éviter les Reads Per Kilobase Million Pairs (RPKM)

Real data

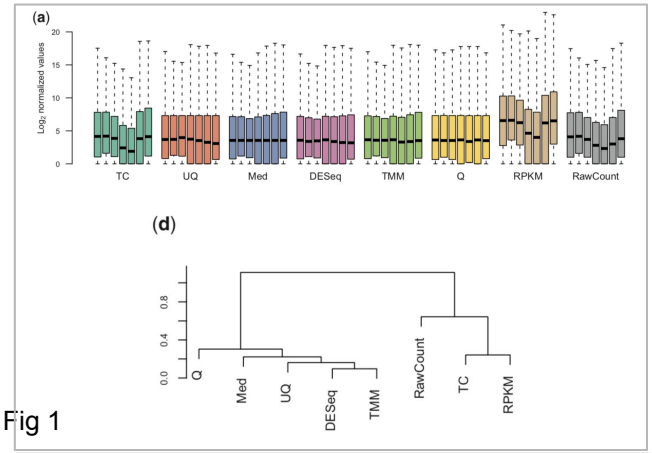
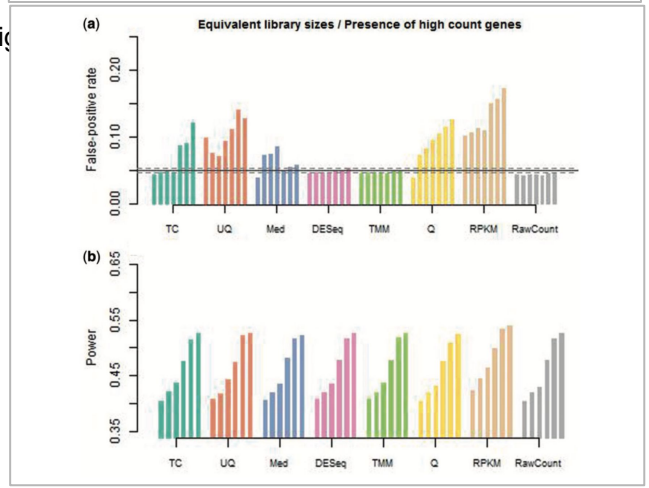


Fig 1

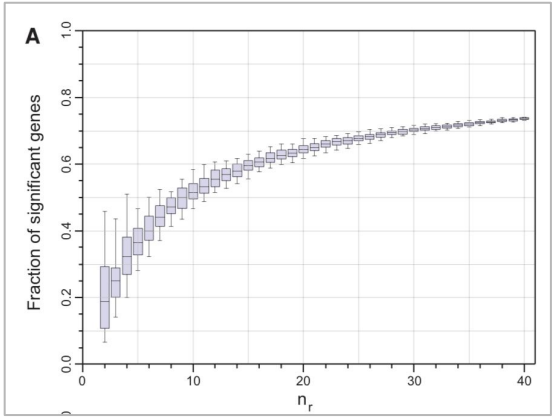
Simulated data

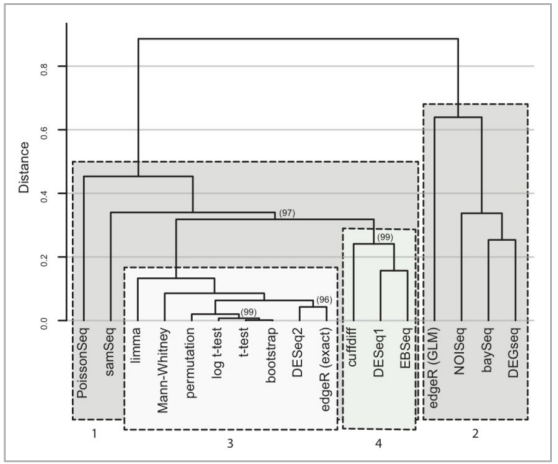


Dillies et al (2013). Briefings in Bioinformatics 14:671-83.

Combien de répliques ?

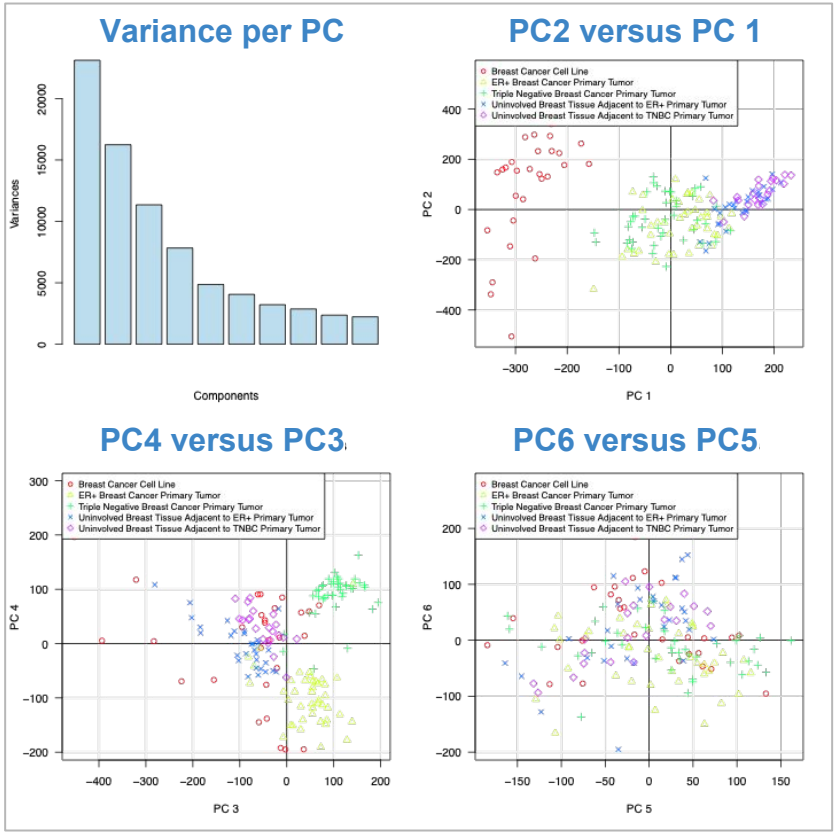
- ***How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?***
  Schurch, N.J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G.G., Owen-Hughes, T., et al. (2016). RNA 22, 839–851.
- Objet de l'étude: tout est dans le titre ;-)
- Expérience chez la levure du boulanger
  - 2 génotypes (WT eet Snf2)
  - 48 répliques
- Résultats
  - Plus y en a plus on a de puissance
  - Avec 3 répliques on ne va vraiment pas très loin.
  - Les gènes déclarés positifs dépendent fortement de la méthode !
- Remarques
  - Le mutant Snf2 intervient dans le remodelage de la chromatine → quasiment tous les gènes sont affectés par la mutation. Ceci pose problème pour les méthodes de normalisation.
  - La variabilité inter-individuelle d'une levure en condition est bien moindre que celle d'humains en liberté





Visualisation des classes en composantes principales

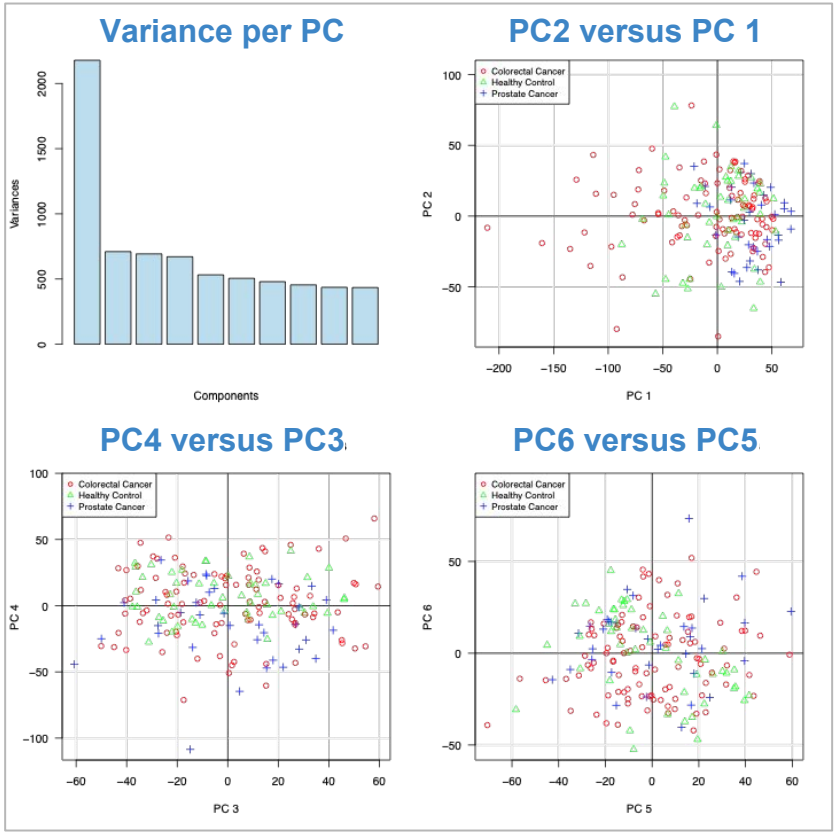
- Quand on a de la chance, l'**analyse en composantes principales** (**ACP**) permet de séparer partiellement les classes d'échantillons.
- Pour ce jeu de données, les composantes 1 à 4 discriminent relativement bien les classes d' échantillons (couleurs).
- Notes
  - L'ACP est non-supervisée, et se base uniquement sur la variance de chaque dimension (variable, gène) entre individu (variance inter-individuelle), sans tenir compte de leur classe.
  - La capacité à discriminer repose sur la variance inter-classes.
  - A priori les variables discriminantes ne sont pas forcément celles associées à la plus grande variance inter-individuelle.
  - Cependant, occasionnellement les variables (gènes) à variance élevée proviennent en partie de la variance intra-classes.



Source: AbuElQumsan, Ghattas & van Helden, *in prep.*

Visualisation des classes en composantes principales

- Dans d'autres cas, l'analyse en composante principales ne permet absolument pas de distinguer les classes.
- Ne perdons pas espoir: ceci n'empêchera pas forcément d'obtenir de bons résultats en classification supervisée, car l'apprentissage repose sur des milliers de variables, et pas sur les quelques premières composantes qu'on peut afficher graphiquement.



Source: AbuElQumsan, Ghattas & van Helden, *in prep.*