StatOmique

Partage d'expérience sur l'analyse statistique de données Omiques

Julie Aubert (AgroParisTech-INRA, UMR MIA Paris)
Marie-Agnès Dillies (Institut Pasteur, C3BI-USR 3756 IP & CNRS, HUB Bioinformatique et Biostatistique)
Christelle Hennequet-Antier (INRA, Centre Val de Loire Tours, UMR BOA)

StatOmique

Création en 2008

- ~ 70 bioinformaticiens et statisticiens
- Groupe de travail du GdR BIM
- Site Web : http://www.sfbi.fr/statomique
- Liste de diffusion : statomique@agroparistech.fr

EPST, EPIC, fondations
INRA (17), CNRS, INRIA, INSERM,
CEA, IFP Energies nouvelles,
Institut Pasteur (9), Institut Curie (8).

Universités, écoles d'ingénieur
Lyon, Strasbourg, Lille, UPMC, Paris-Diderot, Rouen, Toulouse, Paris-Saclay,
AgroParisTech, AgroCampus Ouest, ENS, INSA, Supelec.

Centres hospitaliers, instituts de recherche médicaux
CHU Lyon, CHRU Lille, Centre Léon Bérard de lutte contre le cancer,
Institut de Génétique et de biologie Moléculaire et Cellulaire, Institut de
Pharmacologie Moléculaire et Cellulaire, Institut Imagine.

Autres
Entreprises privées (Pharnext), étudiants ou personnes en recherche d'emploi.

RNA-seq : Normalisation

Consortium StatOmique

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 ... Show more
Author Notes

Briefings in Bioinformatics, Volume 14, Issue 6, 1 November 2013, Pages 671–683,
https://doi.org/10.1093/bib/bbs046

Méthodes de normalization : quelles performances ?

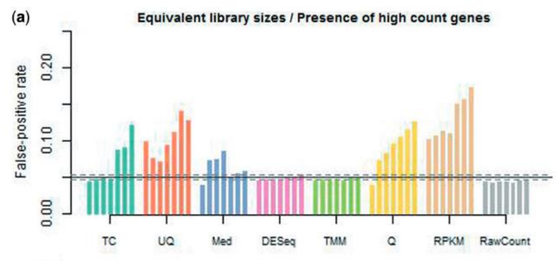
Données réelles

Organism	Type	Number of genes	Replicates per condition	Minimum library size	Maximum library size	Correlation between replicates	Correlation between conditions	% most expressed gene	Library type	Sequencing machine
H. sapiens	RNA	26,437	{3,3}	2.0×10^7	2.8×10^7	(0.98,0.99)	(0.93,0.96)	≈ 1%	SR 54, ND	GaIIx
A. fumigatus	RNA	9,248	{2,2}	8.6×10^6	2.9×10^7	(0.92,0.94)	(0.88,0.94)	≈ 1%	SR 50, D	HiSeq2000
E. histolytica	RNA	5,277	{3,3}	2.1×10^7	3.3×10^7	(0.85,0.92)	(0.81,0.98)	6.4-16.2%	PE 100, ND	HiSeq2000
M. musculus	miRNA	669	{3,2,2}	2.0×10^6	5.9×10^6	(0.95,0.99)	(0.09,0.75)	17.4-51.1%	SR 36, D	GaIIx

Table 1: Summary of datasets used for comparison of normalization methods, including the organism, type of sequencing data, number of genes, number of replicates per condition, minimum and maximum library sizes, Pearson correlation between replicates and between samples of different conditions (minimum, maximum), percentage of reads associated with the most expressed RNA (minimum, maximum), library type (SR = single-read or PE = paired-end read, D = directional or ND = non-directional), and sequencing machine.

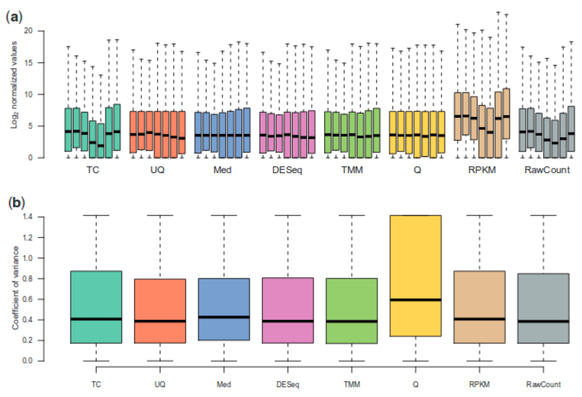
Données simulées (RNA-seq M. musculus)

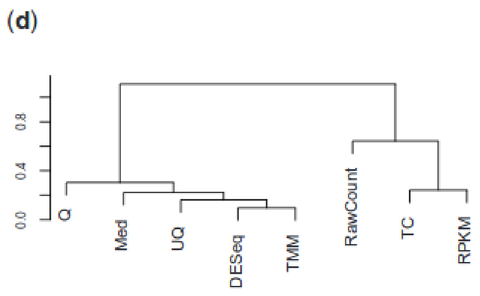
Proportion de gènes différentiellement exprimés : 0 à 30%
Taille des banques : équivalentes ou non
présence / absence de gènes à fort comptage



Sept méthodes de normalisation comparées

Taille de la banque : RPKM - Reads Per KiloBase Per Million Mapped (Mortazavi et al. 2008), Total Count (Marioni et al. 2008)
Distribution : Median, Upper Quartile (Bullard et al. 2010), Full Quantile (Robinson and Smyth 2008), Trimmed Mean of M-values (Robinson and Oshlack 2010, edgeR), Relative Log-Expression (Anders and Huber 2010, DESeq2)



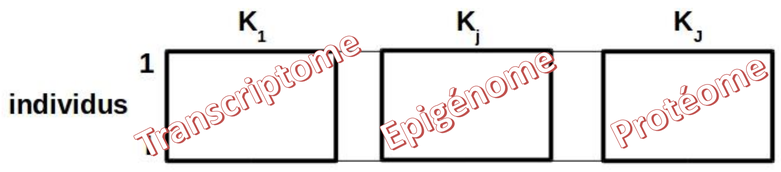


Conclusions

- Différences entre les méthodes en présence de gènes à fort comptage et de tailles différentes des banques.
- Méthodes performantes et robustes dans un contexte d'analyse différentielle :
  - ✓ **Trimmed Mean of M-values**, (Robinson and Oshlack 2010, edgeR)
  - ✓ **Relative Log-Expression**, (Anders and Huber 2010, DESeq2)

  Hypothèse : les gènes sont majoritairement invariants entre les conditions
- La normalisation est **nécessaire et non triviale**
- Ne pas normaliser par la longueur du gène dans un contexte d'analyse différentielle
- **Autres méthodes et hypothèses : Evans et al. 20017**.

Intégration statistique de données hétérogènes



Données mesurées sur les mêmes individus et réparties en J tableaux (individus x variables)

Défis :
- Hétérogénéité
- Fléau de la dimension p variables >> n individus
- Données manquantes

Challenge Integraal : "Intégration de données hétérogènes et de haute dimension pour la découverte de biomarqueurs prédictifs" https://inra-ph.wixsite.com/workshop-ppc/challenge

Groupe	Nb de variables
Méthylation	15163
miARN	852
CNA	17130
RNASeq	20016

Groupe	Nb de variables
Transcriptome	24415
miARN	362
Métabolites	54

Fonctionnement et perspectives

2 à 3 journées d'échanges par an

Partages bibliographiques

Expression	Epigénétique	Et aussi
• Single cell • RNA-seq • Analyse différentielle • Réseaux de co-expression	• ChIP-seq • Détection de pics • méthylome	• Métagénomique • CRISPR • Planification d'expérience • GWAS • Long reads

Formations en interne : RGCCA, Tidyverse, …

Single Cell OMICS Data Analysis Workshop

December 1, 2017
> Institut de Pharmacologie Moléculaire et Cellulaire, Sophia Antipolis, France

Rejoignez-nous :
christelle.hennequet-antierATinra.fr,
marie-agnes.dilliesATpasteur.fr,
Julie.AubertATagroparistech.fr

Merci à tous les StatOmiciens, pour leurs contributions passées, présentes et futures.