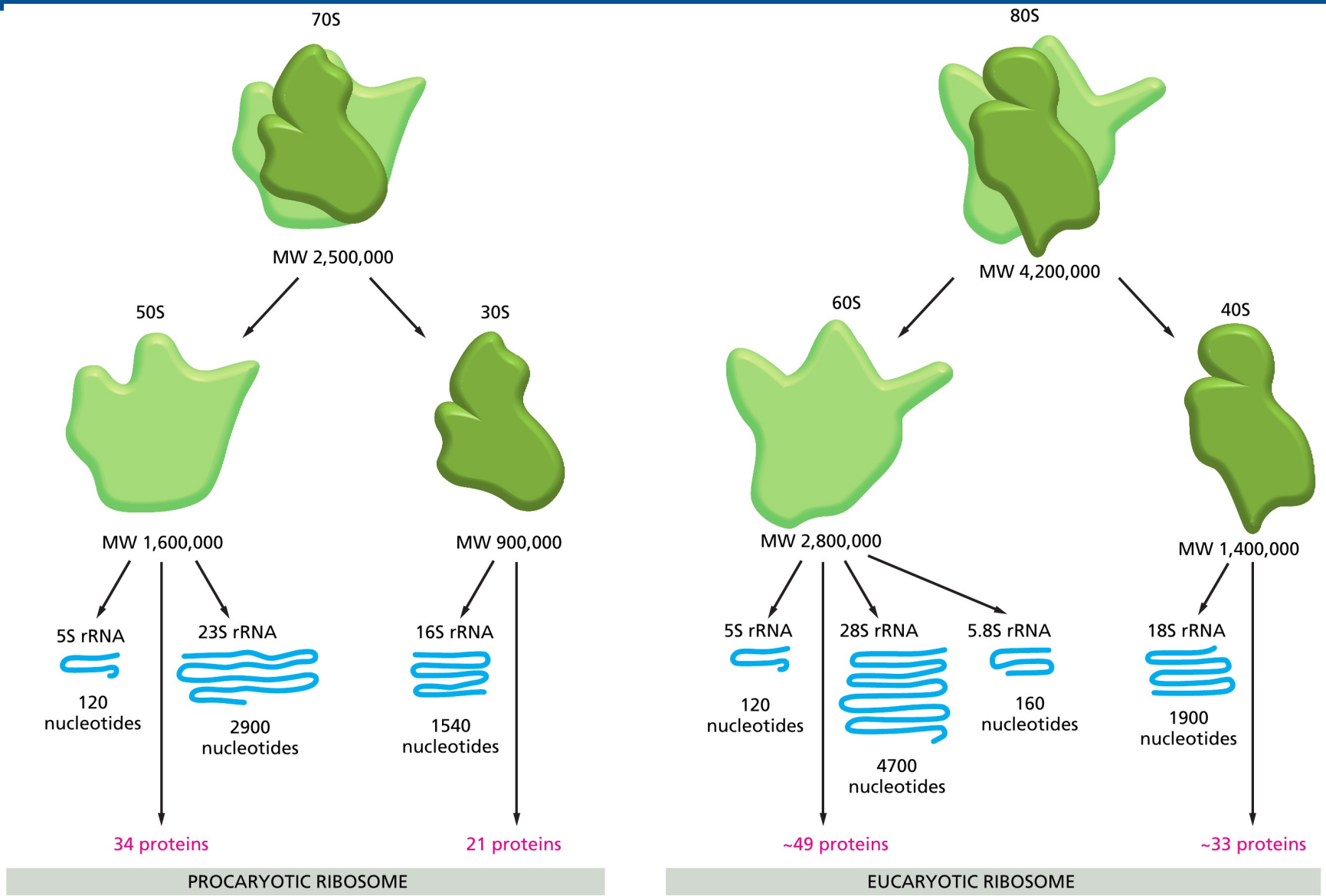
SHAMAN : SHiny Application for Metagenomic ANalysis

Stevenn Volant, Amine Ghozlane
Hub Bioinformatique et Biostatistique – C3BI, USR 3756 IP CNRS
Biomics – CITECH

Ribosome

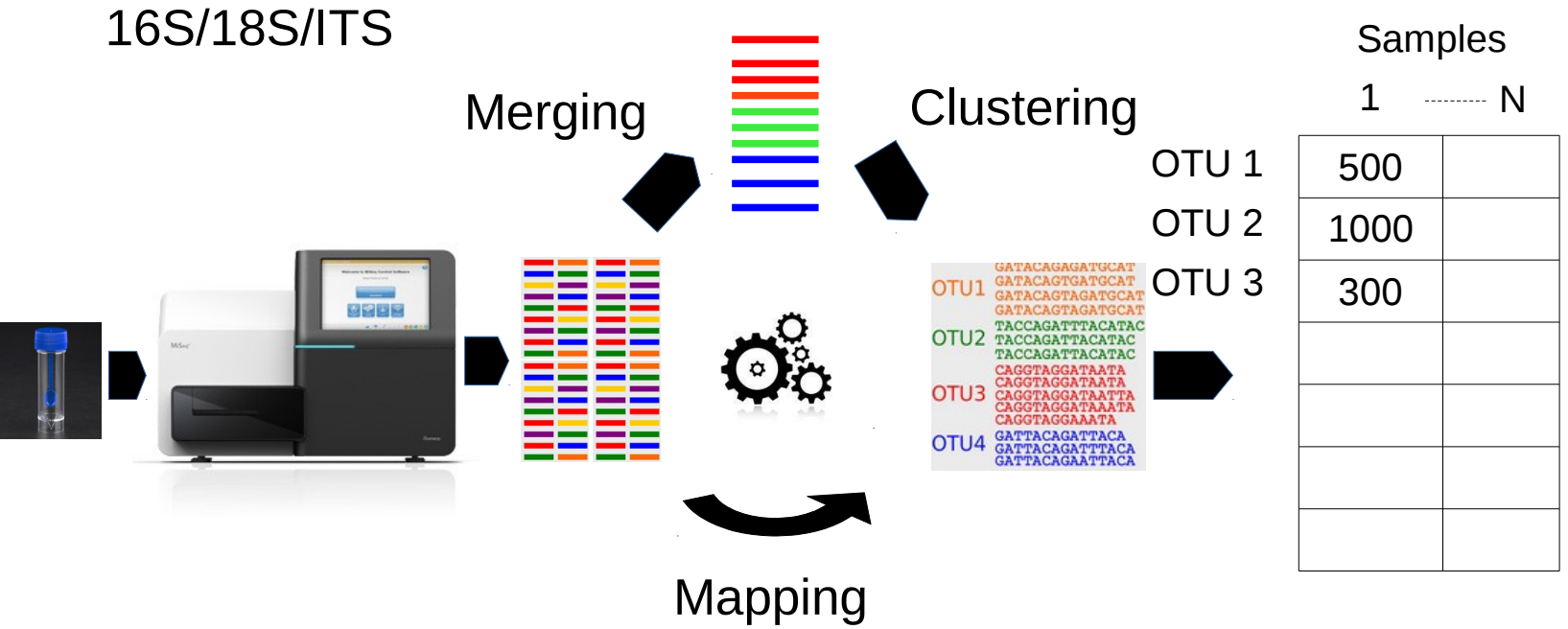


ITS (1) : located between 18S and 5.8S rRNA genes

Image from Alberts Molecular Biology of the Cell 5th

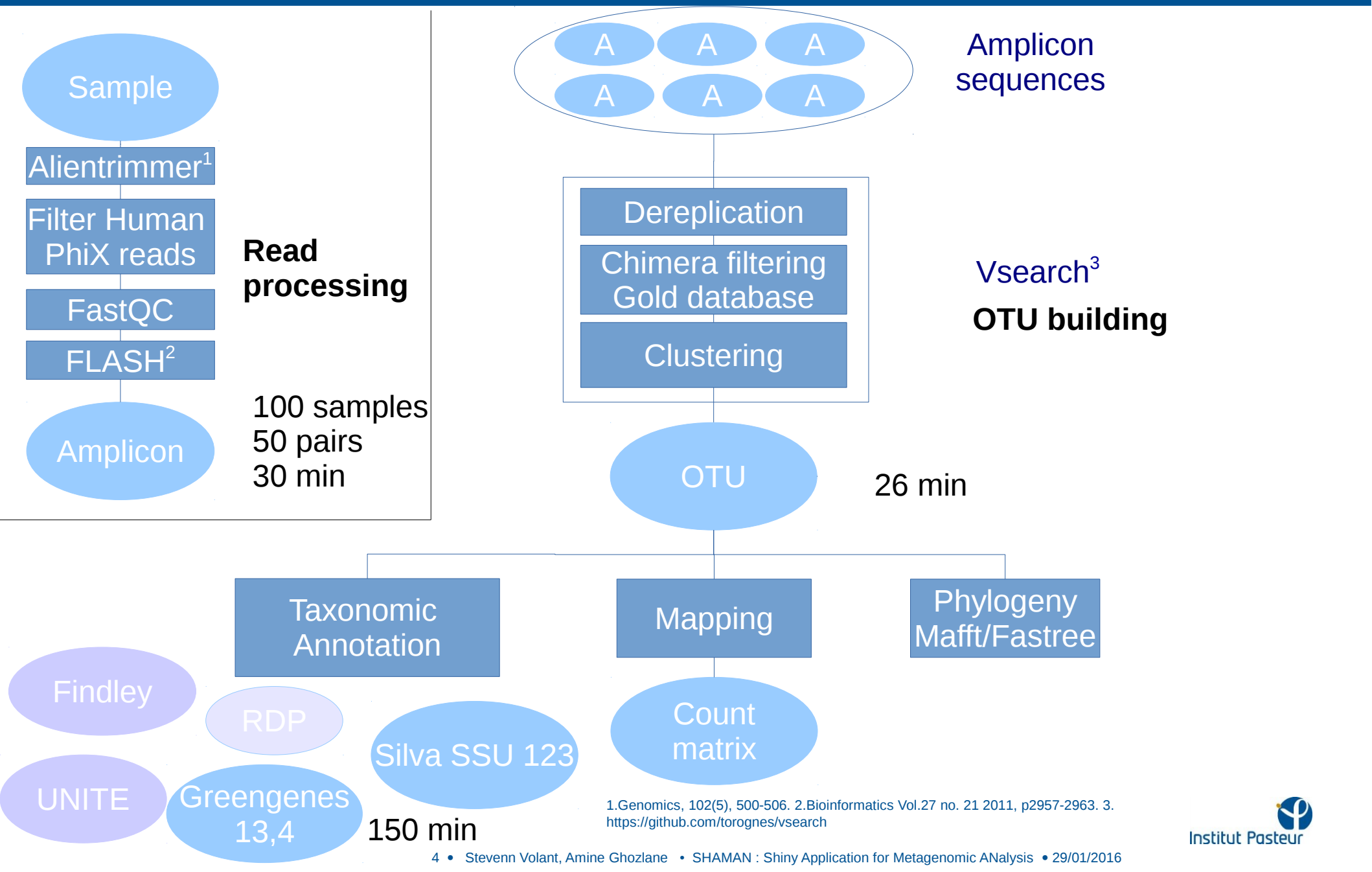
Institut Pasteur

Quantitative metagenomics pipeline



Institut Pasteur

HUB – 16S/18S/ITS pipeline

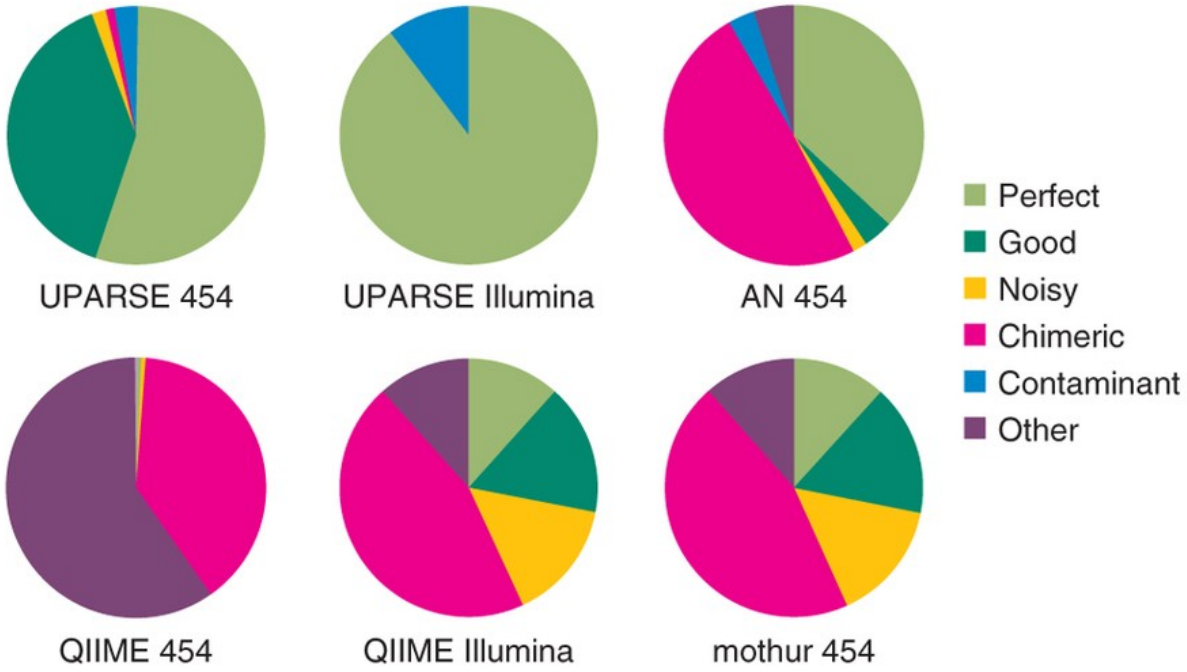


1.Genomics, 102(5), 500-506. 2.Bioinformatics Vol.27 no. 21 2011, p2957-2963. 3. https://github.com/torognes/vsearch

Institut Pasteur

Uparse/Vsearch

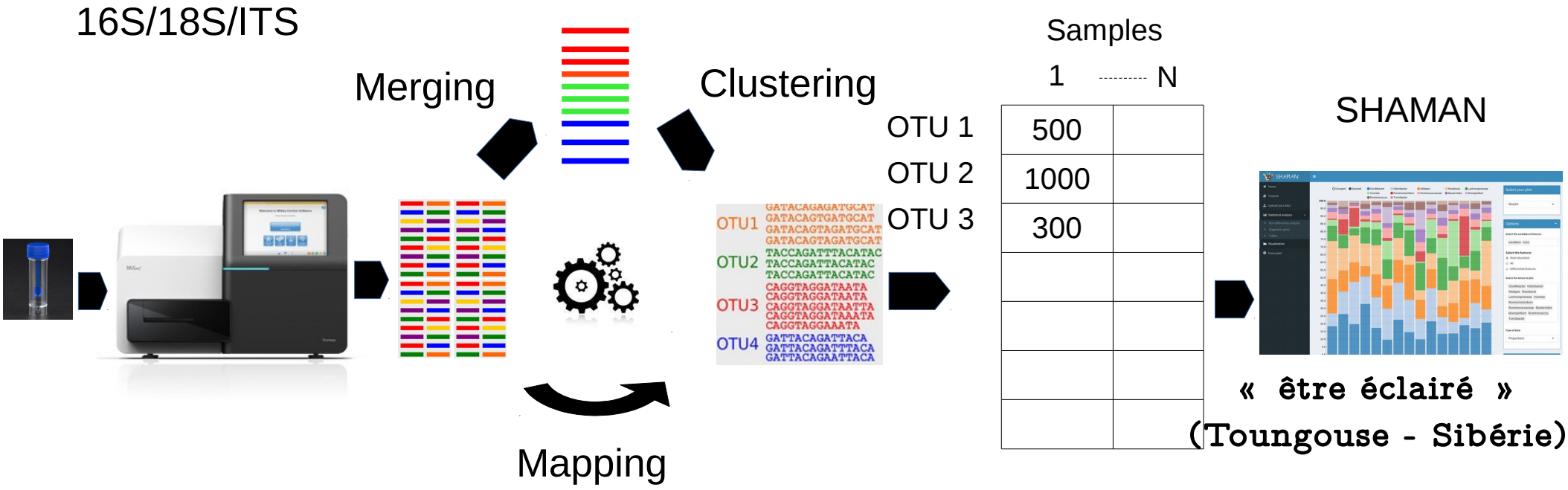
- **Integrated in QIIME, MOTHUR and LotuS**



Nature methods, 10(10), 996-998.

Institut Pasteur

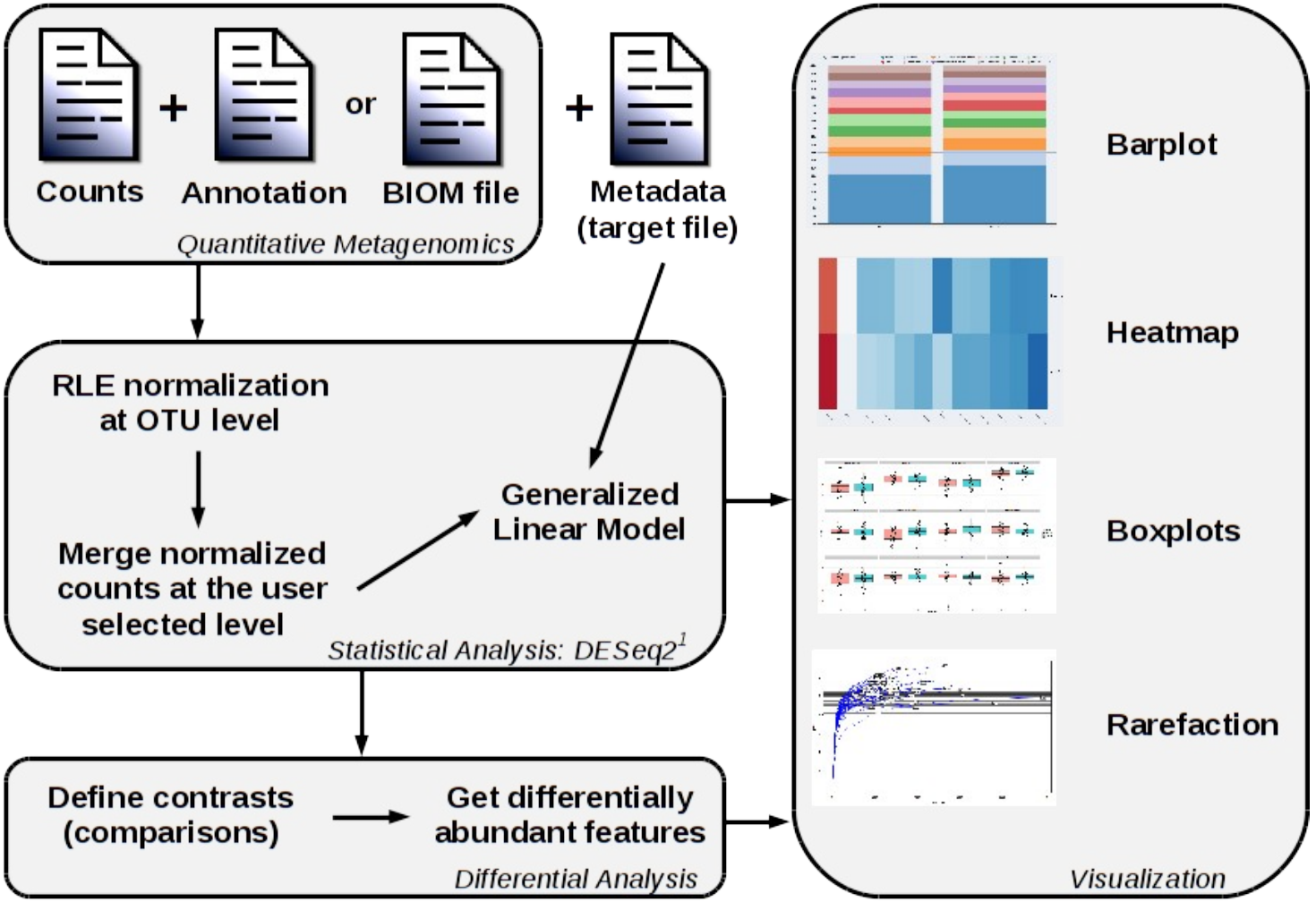
Quantitative metagenomics pipeline



Institut Pasteur

SHAMAN : shaman.c3bi.pasteur.fr

« There is no disputing the importance of statistical analysis in biological research, but too often it is considered only after an experiment is completed, when it may be too late. »



[1] Love MI, Huber W and Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, pp. 550

Institut Pasteur

SHAMAN : shaman.c3bi.pasteur.fr

Counts

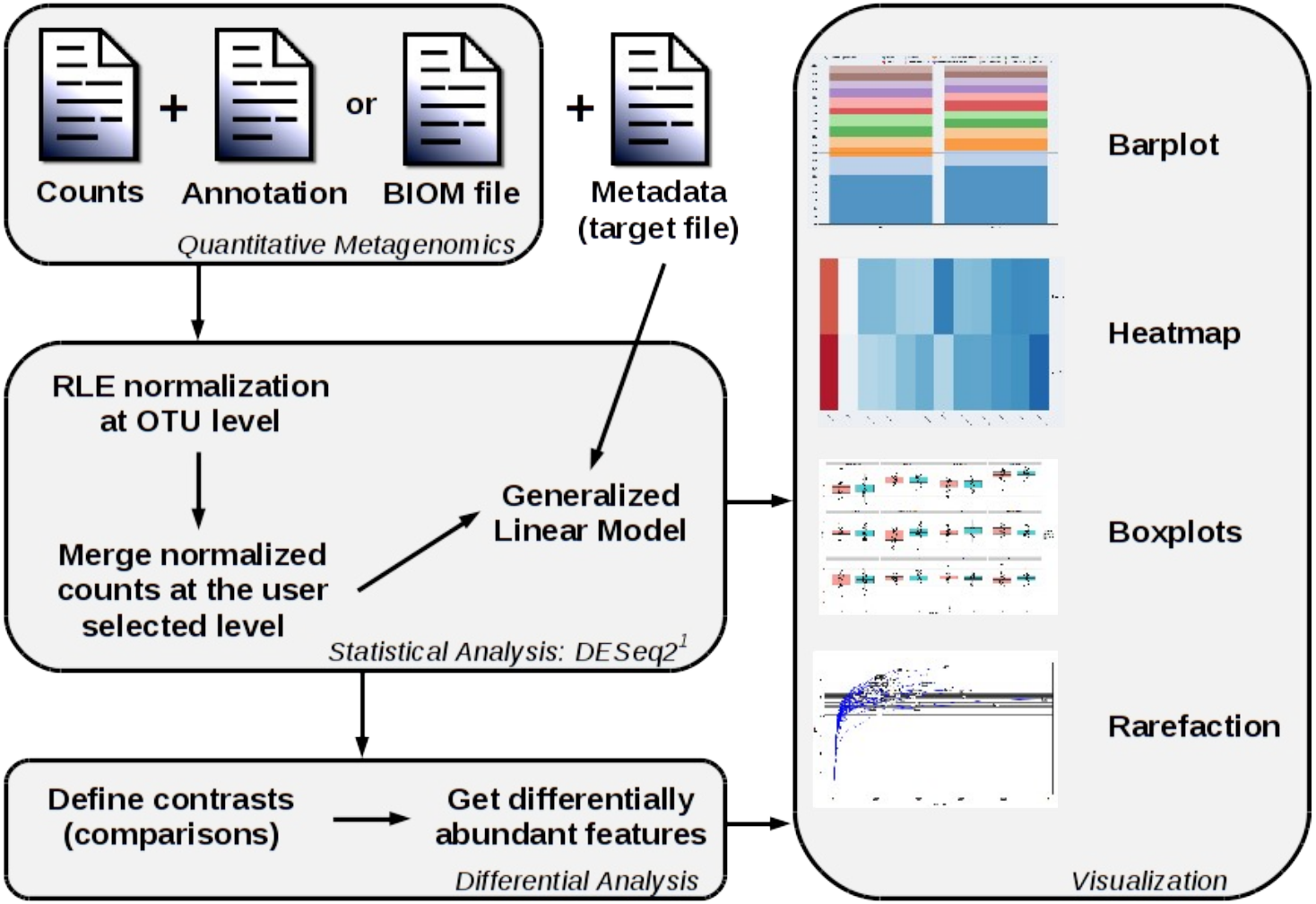
OTUId	Delta.Compl1.13_S13	Delta.Compl1.31_S31	Delta.Compl1.49_S49	Delta.Compl1.67_S67
OTU_41131	50	19	47	11
OTU_21509	641	356	1526	447
OTU_26144	204	88	32	68
OTU_34025	130	47	18	6
OTU_4597	1820	1628	16	4
OTU_40251	11	22	63	74
OTU_35066	156	85	570	168
OTU_39472	17	1	32	8
OTU_35326	297	51	61	47
OTU_2526	946	282	70	32
OTU_23642	303	106	65	40
OTU_44238	0	1	2	5
OTU_53265	6	9	7	3
OTU_31446	799	237	28	47
OTU_39136	28	235	179	152
OTU_8534	807	225	1973	267
OTU_38289	183	82	106	42
OTU_37452	95	41	132	70
OTU_53906	85	25	45	55
OTU_30585	828	319	49	46
OTU_51805	1	0	1	2
OTU_1	1316	532	573	1182
OTU_27211	422	131	61	59
OTU_41302	126	39	3	0
OTU_16427	8351	893	75	865
OTU_49006	0	0	0	0
OTU_51874	0	1	0	0
OTU_48435	0	1	0	0
OTU_20150	234	189	834	4055
OTU_24853	225	81	50	4
OTU_36396	448	81	20	111
OTU_27700	358	84	35	71
OTU_29553	186	149	273	1019
OTU_46484	3	0	0	0

Annotation

OTU	Kingdom	Phylum	Class	Order	Family	Genus	Specie
OTU_47937	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae		
OTU_50499	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae		
OTU_50493	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae		
OTU_52457	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	
OTU_54350	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	
OTU_48079	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae	
OTU_51367	Bacteria	Firmicutes	Clostridia	Clostridiales			
OTU_53666	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae	
OTU_53912	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae	
OTU_45606	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia	
OTU_47565	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae		
OTU_53991	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	
OTU_51235	Bacteria	Bacteroidetes	Bacteroidia				
OTU_46289	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae	
OTU_53310	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae	
OTU_47779	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae		
OTU_38495	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	
OTU_52264	Bacteria	Bacteroidetes	Bacteroidia				
OTU_54136	Bacteria	Bacteroidetes	Bacteroidia				
OTU_54531	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae		
OTU_41172	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillibacter	
OTU_54407	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae		
OTU_44950	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Odoribacter	
OTU_54051	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Odoribacter	
OTU_54274	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae		
OTU_51992	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Coprococcus	
OTU_26872	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	
OTU_47012	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae		
OTU_48135	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	
OTU_48860	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae		
OTU_52609	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	
OTU_53138	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae		
OTU_53305	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	
OTU_53604	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales			
OTU_53951	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae		
OTU_53964	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	
OTU_53990	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	
OTU_54067	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae		
OTU_54079	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	
OTU_54080	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	
OTU_54268	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae		
OTU_52265	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae	

SHAMAN : shaman.c3bi.pasteur.fr

« There is no disputing the importance of statistical analysis in biological research, but too often it is considered only after an experiment is completed, when it may be too late. »



[1] Love MI, Huber W and Anders S (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biology, 15, pp. 550

Institut Pasteur

Metagenomic vs RNA-seq

DESeq2 approach is usually used for RNA-seq dataset

	Metagenomic	RNA-seq
Distribution	Overdispersed counts → Negative binomial	Overdispersed counts → Negative binomial
Constraints	Highly abundant species	Highly expressed genes
Goal	Find differentially abundant features (species, familly, …): OTU distributions and abundances vary between conditions	Find differentially expressed genes: Distributions and expression vary between conditions

→ Metagenomic data are similar to RNA-seq data

Data normalization

Why ?

- To correct technical biaises and make samples comparables.

How ?

- Fitting the distributions (Total Read Count, UpperQuartile, Median, Full Quantile)
- Account for the feature length (RPKM)
- **Concept of « effective reads number »** (TMM, **DESeq2**)

Remarks?

- Some methods normalize the counts, others the library sizes
- Some are designed for differential analysis

DESeq2 normalization (OTU level)

Assumption

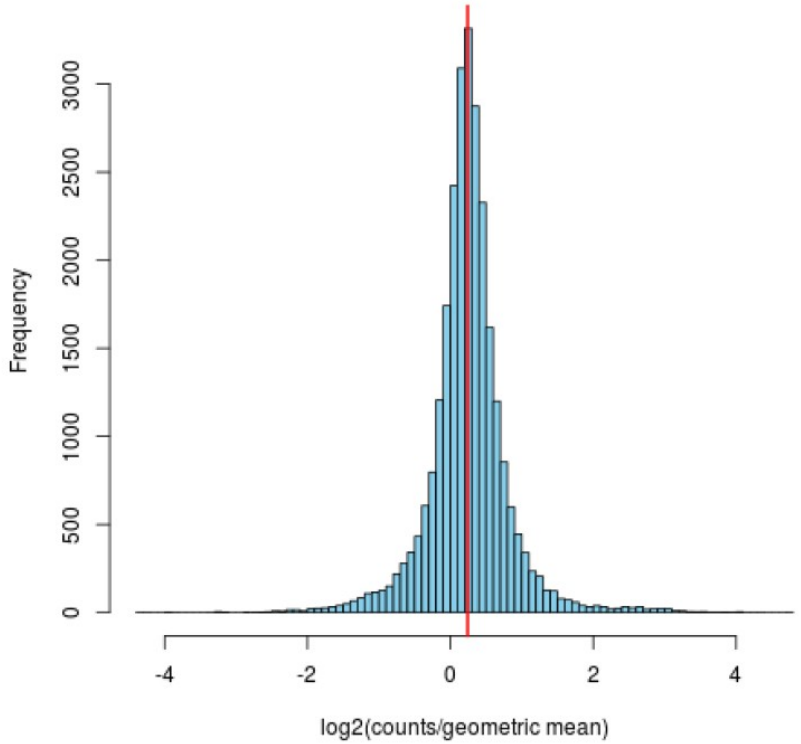
- Most of the OTU have the « same » abundance between 2 conditions

Normalization factor :

$$\hat{s}_j = median_i \frac{x_{ij}}{(\prod_{\nu=1}^{n} x_{i\nu})^{1/n}}$$

where

X_{ij} : Number of mapped reads of the OTU i in sample j
n: Number of samples



Institut Pasteur

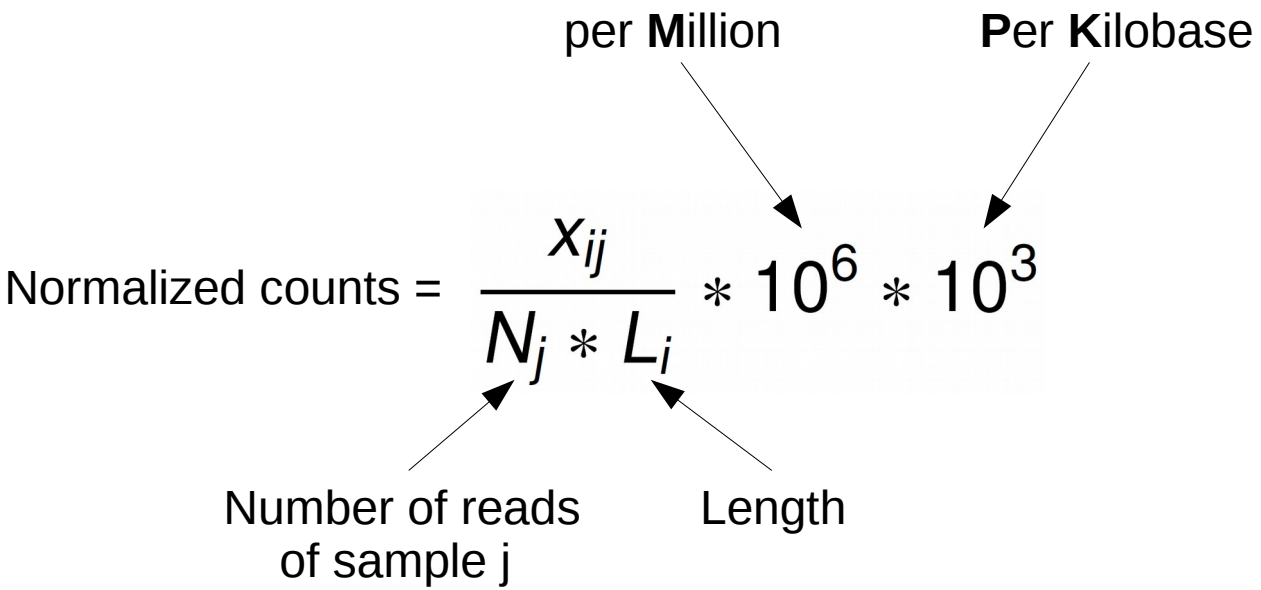
Comparison with RPKM (1/3)

- **RPKM** : **R**eads **P**er **K**ilobase per **M**illion mapped reads

Assumption

- ✗ Counts are proportional to abundance, the length and the sequencing deepth.

- **Method**



Institut Pasteur

Comparison with RPKM (2/3)

Briefings in Bioinformatics Advance Access published September 17, 2012
BRIEFINGS IN BIOINFORMATICS. page 1 of 13 doi:10.1093/bib/bbs046

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 Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

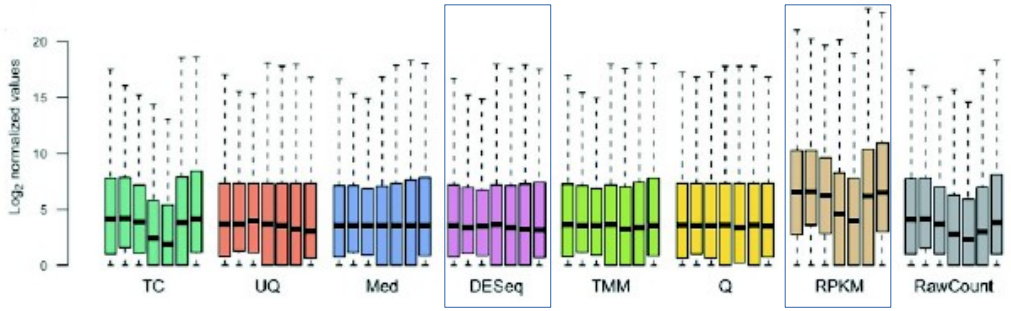
Submitted: 12th April 2012; Received (in revised form): 29th June 2012

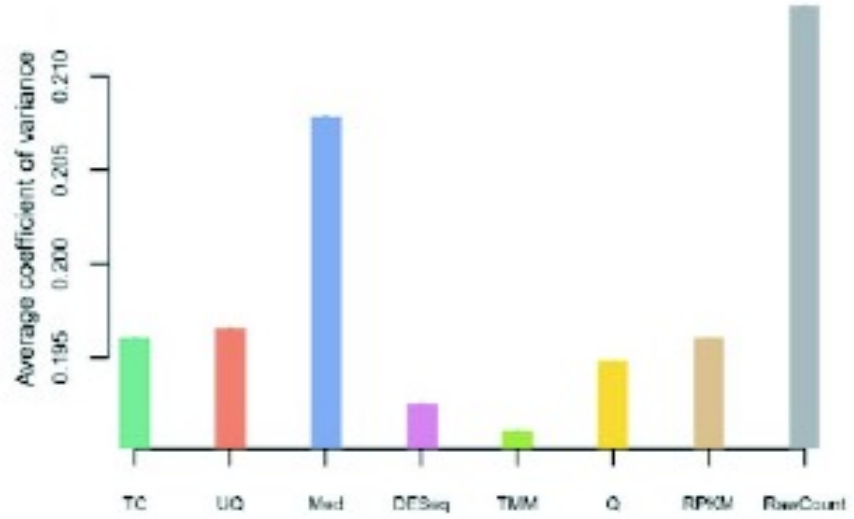
→ Comparison of 7 normalization methods

Institut Pasteur

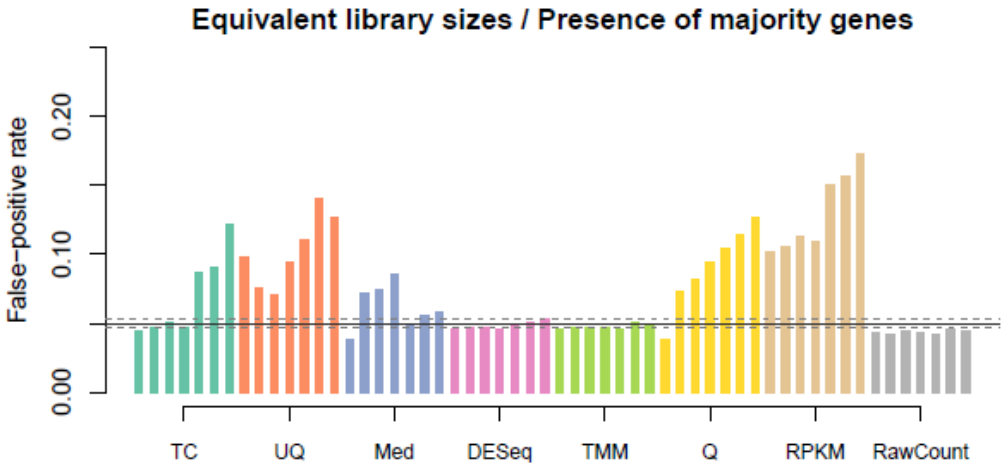
Comparison with RPKM (3/3)

Results on real data (7 samples)





FDR and Power



Dillies M. et al., Bioinformatics 2013

Institut Pasteur

To sum up

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	–	+	+	–	–
UQ	++	++	+	++	–
Med	++	++	–	++	–
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
FQ	++	–	+	++	–
RPKM	–	+	+	–	–

DESeq2 normalization provides better results

OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

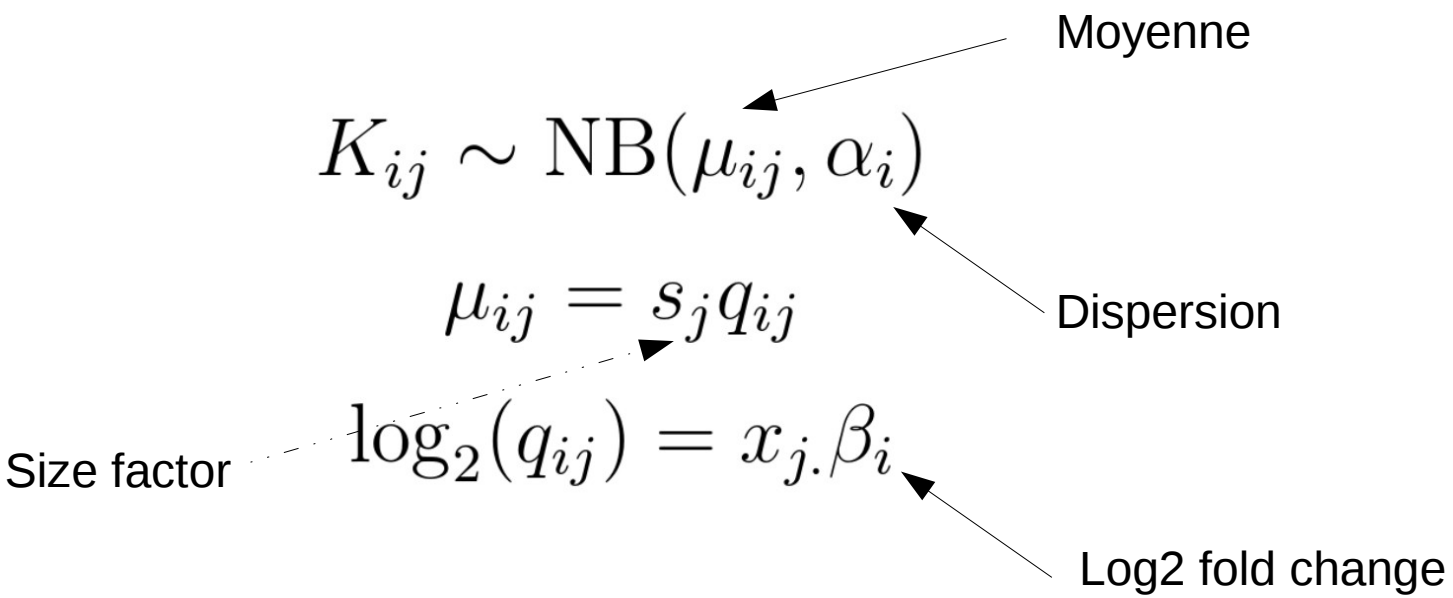
Paul J. McMurdie, Susan Holmes*

Statistics Department, Stanford University, Stanford, California, United States of America

Recommend using DESeq2 to perform analysis of differential abundance

Statistical model of DESeq2

Generalized Linear Model



Advantages

- Allows complex experimental designs.

Dispersion estimation

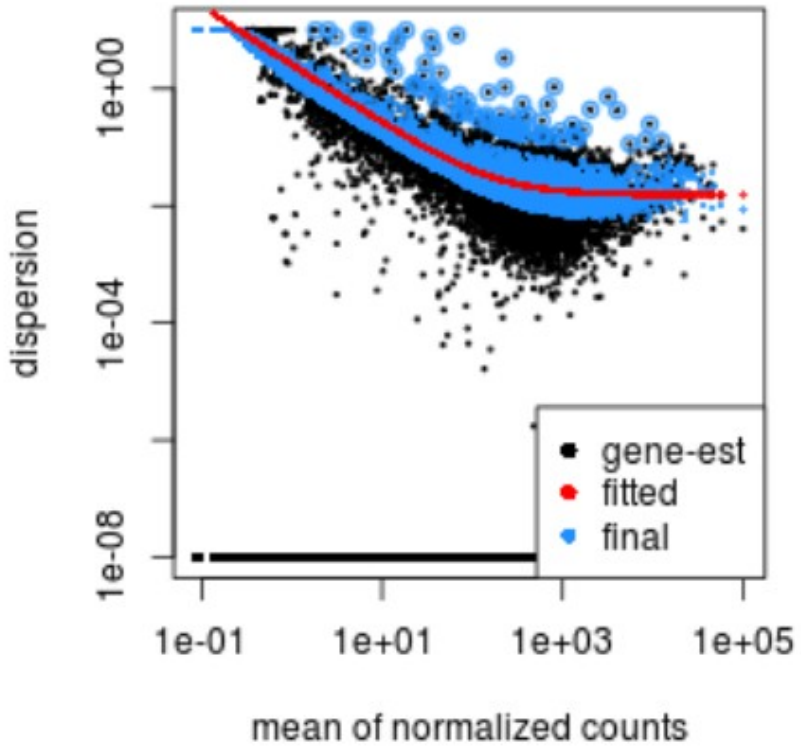
Problem

- Get a good estimate of the dispersion with a small number of samples.

Modelisation of the dispersion :

$$\log \alpha_i \sim N(\log \alpha_{\mathrm{tr}}(\bar{\mu}_i), \sigma^2_{\mathrm{d}})$$

Function of the mean
of normalized count



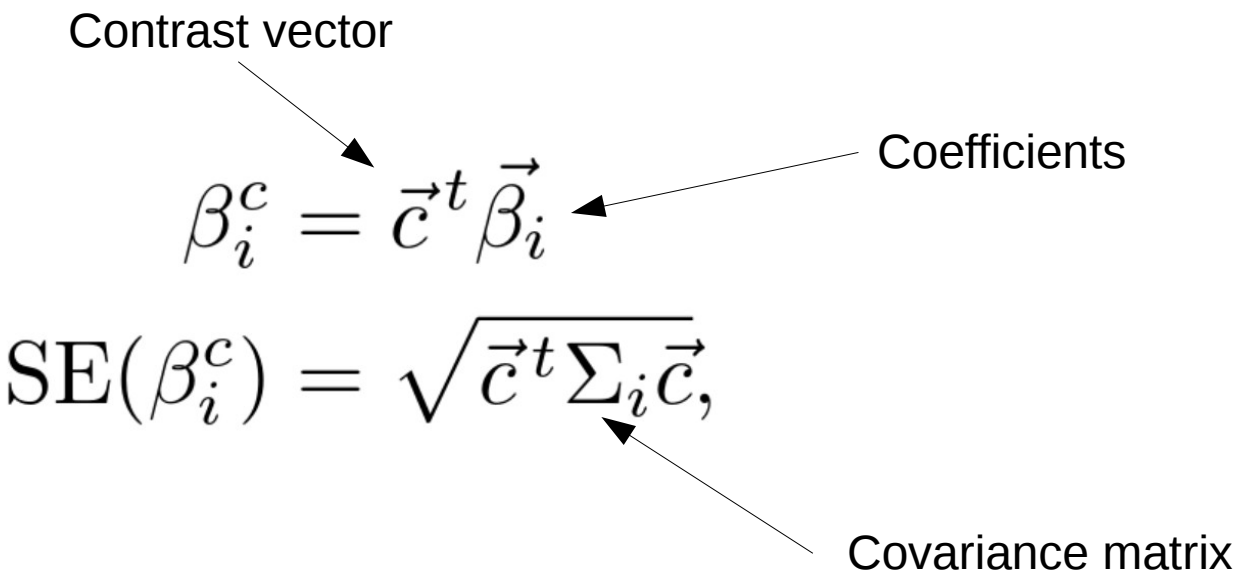
Local parametric regression

Institut Pasteur

Contrasts (comparisons)

Aim

- Testing a specific effect without having to re-fit the model.



Advantages

- Parameters are estimated with all samples.

Institut Pasteur

Conclusions

SHAMAN

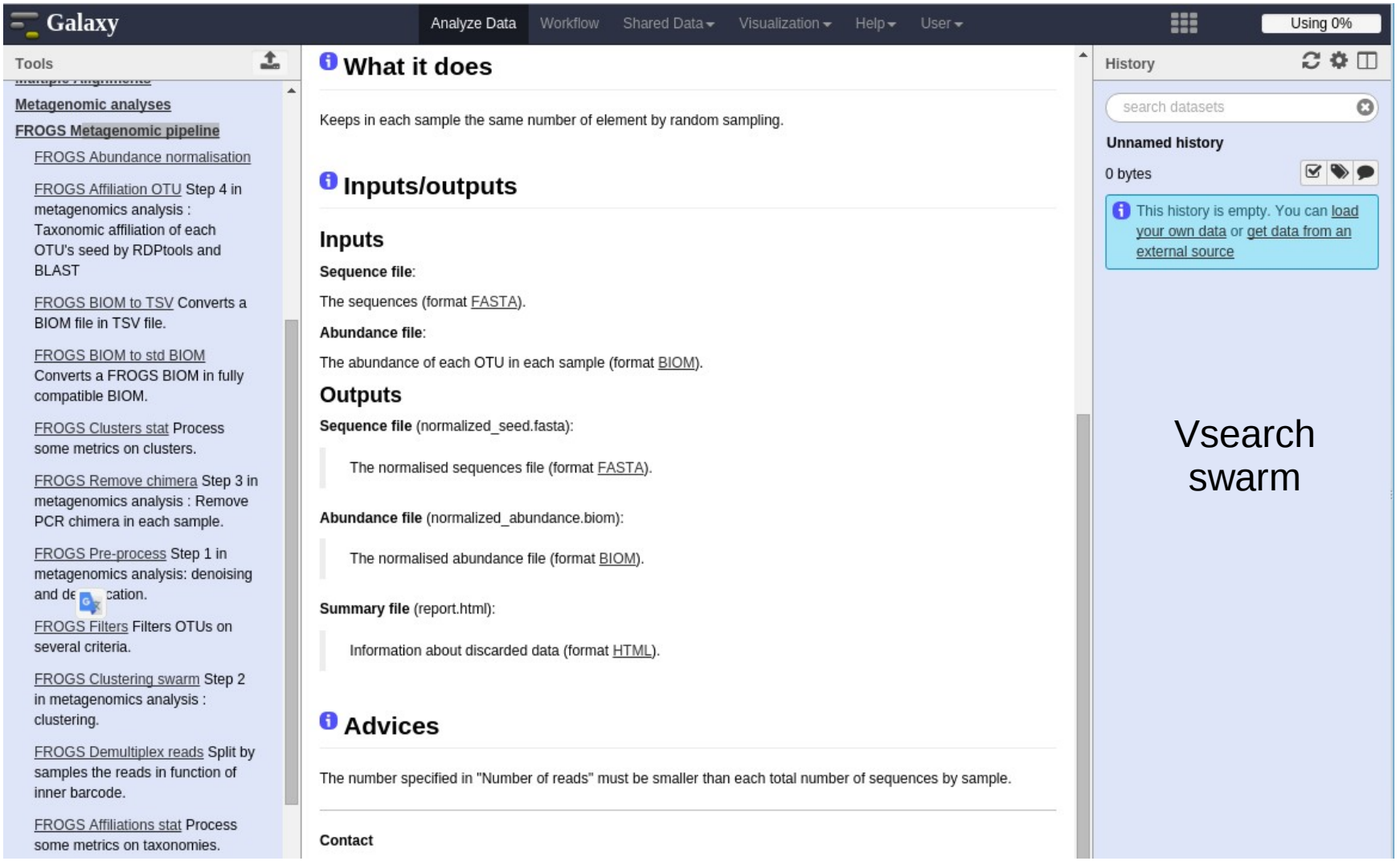
- **16s/18s/its analysis**
- **Strong statistical approach**
- **Several visualizations available**
- **Access : http://shaman.c3bi.pasteur.fr**

Incoming features

- **WGS analysis**
- **New visualizations (Taxonomy plot, Krona, continuous data)**
- **Compatibility with FROGS**

Institut Pasteur

CIB – FROGS 16S/18S – GALAXY Pasteur



Galaxy team : Mathieu Valade, Fabien Mareuil
Emmanuel Quevillon, Eric Deveaud

Institut Pasteur

Acknowledgements

C3BI Bioinformatics Biostatistics Hub

Olivier GASCUEL
Marie-Agnès DILLIES
Christophe MALABAT
Hugo VARET
Nicolas MAILLET
Pierre LECHAT
Anna ZHUKOVA
Rachel TORCHET

CiTech Biomics Pole

Sean KENNEDY
Béatrice REGNAULT

Institut Pasteur