

TUTORIEL SHAMAN

SHiny Application for Metagenomic ANalysis

21/05/2017

The screenshot shows the 'Welcome to SHAMAN' page. On the left, a sidebar lists navigation links: Home, Tutorial, Download/Install, Raw data, and Upload your data. The main content area is titled 'Welcome to SHAMAN' and contains a detailed workflow diagram. The workflow starts with 'Input files' (Counts, Annotation, Metadata (target file) or BIOM) leading to 'Statistical Analysis'. This involves 'Normalization at OTU level (modified PLE approach)', 'Merging normalized counts at the user selected level', 'Filtering the features (optional step)', and 'Run DESeq2'. The output of this analysis is used for 'Differential Analysis' (Define contrast vector (comparisons), Get differential abundance, Results). This leads to 'Diagnostic plots' (PCA, NMDS, Phylogenetic tree, Taxonomic plot, Box plot, Violin plot) and finally 'Visualization plots' (Treemap, Bar chart, Scatter plot, Line plot). A text box explains the software's compatibility with standard formats for metagenomic analysis and its use of DESeq2 R package. To the right, a 'What's new in SHAMAN' section lists updates: March 30th 2017 (Krona, Phylogeny and bug fixes), Dec 9th 2016 (Phylogenetic tree and stress plot), Nov 22th 2016 (New visualization and bug fix), and Oct 12th 2016 (Filtering step and bugs fix). Logos for Institut Pasteur and CIBIO are at the bottom.

shaman.pasteur.fr

Auteur : Stevenn Volant

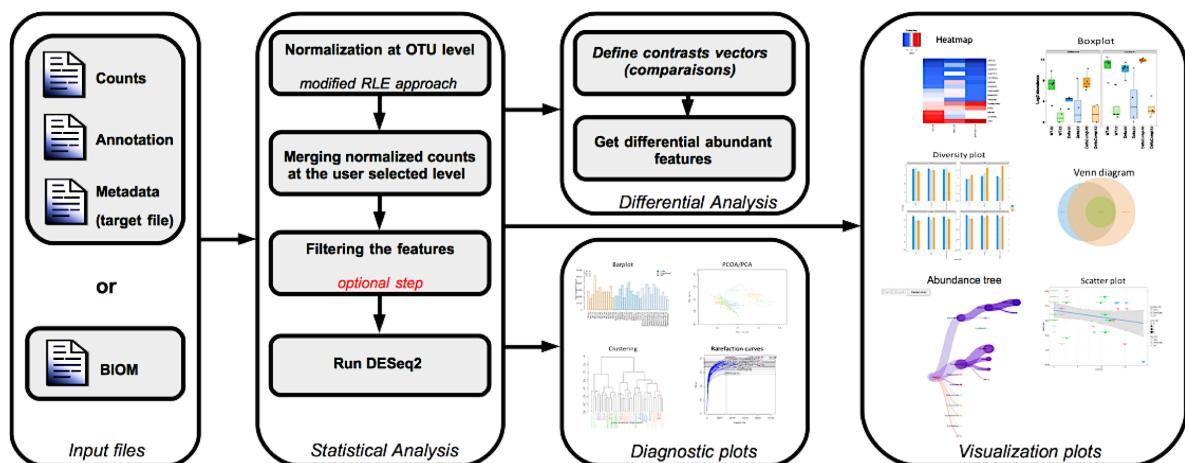
Présentation.....	3
L'interface	4
Page d'accueil	4
Tutoriel et téléchargement.....	5
Chargement des données	6
Analyse statistique.....	10
Définir le modèle statistique.....	10
<i>Chargement du design experimental.....</i>	<i>11</i>
<i>Création du modèle.....</i>	<i>12</i>
Options du modèle et de la normalisation.....	14
Appliquer un filtre sur les données (optionnel).....	17
Définir un vecteur de contraste.....	18
Vérification du modèle statistique	19
Résultats de l'analyse différentielle.....	24
Visualisations.....	26
Visualisations des résultats.....	28
<i>Composition globale.....</i>	<i>28</i>
<i>Fold-change.....</i>	<i>30</i>
<i>Liens entre les variables.....</i>	<i>30</i>
<i>Abondance et taxonomie.....</i>	<i>31</i>
Comparaisons de résultats	32
Bibliographie	34
Annexe A	35

PRESENTATION

SHAMAN est une application web développée à partir des packages shiny et shinydashboard de R. Cette application permet de réaliser une analyse différentielle complète des données de métagénomique (16S, 18S, 23S et MGS) ainsi que de visualiser les résultats par le biais de nombreux types de graphiques (barplot, boxplots, heatmap,...). L'utilisateur peut également utiliser SHAMAN pour réaliser l'analyse bioinformatique, c'est-à-dire obtenir l'abondance et l'annotation des OTU à partir des fichiers fastq.

L'analyse statistique réalisée par SHAMAN est fondée sur le package DESeq2 de R [Anders et Huber 2010 ; Love et al. 2014] offrant de bonnes performances pour l'analyse de données métagénomiques ([McMurdie et Holmes 2014, Jonsson 2016]).

SHAMAN est compatible avec les formats standard utilisés en métagénomique (.csv, .tsv, ou biom). L'ensemble des figures et tableaux générés sont exportables dans plusieurs formats. La figure ci-après représente le workflow de SHAMAN :



L'INTERFACE

Page d'accueil

Bienvenu sur <http://shaman.pasteur.fr/>. Une fois l'application chargée, voici la page d'accueil qui s'affiche.

The screenshot shows the SHAMAN application homepage. On the left, there is a sidebar with a logo and links: Home, Tutorial, Download/Install, Raw data, and Upload your data. A green circle labeled '1' is positioned next to the 'Upload your data' link. The main content area has a dark header 'Welcome to SHAMAN'. Below it, there's a 'What's new in SHAMAN' section with three items:

- March 30th 2017 - Krona, Phylogeny and bug fixes**: Describes the addition of Krona and phylogenetic tree plots, and fixes for import float count matrices.
- Dec 9th 2016 - Phylogenetic tree and stress plot**: Describes the addition of a phylogenetic tree to calculate unifrac distance and a stress plot for NMDS.
- Nov 22th 2016 - New visualization and bug fix**: Describes a new visualization called 'tree'.

In the center, there is a large diagram titled 'Hereafter is the global workflow of the SHAMAN application:' which illustrates the data processing pipeline from input files to differential analysis and finally to visualization plots.

- 1 Barre de menu. 5 onglets sont disponibles (avant chargement des données) : Home, Tutorial, Download/Install, Raw data, Upload your data.
- 2 Description de l'application
- 3 Liste des nouveautés dans SHAMAN.

Tutoriel et téléchargement

Les deux onglets « Tutorial » et « Download/Install » sont les suivants.

Screenshot 1: Tutorial

The screenshot shows the 'Tutorial' section of the SHAMAN web interface. It features a study design diagram illustrating two groups of participants over a 40-day period. The timeline is divided into phases: a 7-day control period, a 5-day diet period ("D1"), a 15-day washout period ("W"), a 5-day diet period ("D2"), and a final 7-day washout period. Participants are numbered 1 through 6. A legend indicates that solid black bars represent 40g of fiber per day, while dashed black bars represent 10g of fiber per day. Sample sequencing details are also provided at the bottom.

Screenshot 2: Download/Install

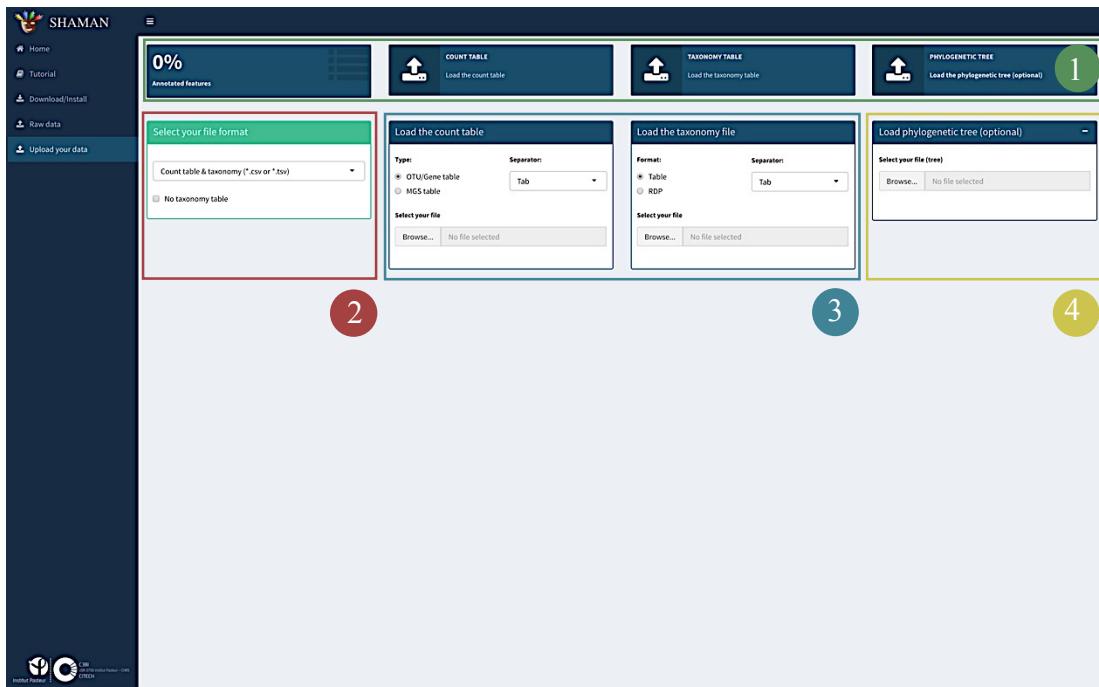
The screenshot shows the 'Download/Install' section of the SHAMAN web interface. It provides instructions for installing SHAMAN using Docker or R. The Docker install section includes commands for pulling the image and running it with port mapping. The R install (RC) section includes a link to the R package documentation. A red box highlights the Docker installation steps.

1 Tutoriel en ligne : guide sur l'utilisation de SHAMAN pour des données 16S. Le jeu de données utilisé pour ce tutoriel est disponible en téléchargement.

2 Guide d'installation de SHAMAN. Outre la version disponible en ligne, SHAMAN vous propose une installation via docker et vous guide également pour une installation locale dans R.

Chargement des données

Avant de lancer une analyse différentielle, l'utilisateur doit charger ses données. Si l'utilisateur dispose d'un fichier contenant **une table de comptage et un autre contenant l'annotation taxonomique**, voici l'affichage de SHAMAN :



- 1 Information sur l'annotation, la table de comptage, le fichier d'annotation taxonomique et l'arbre phylogénétique (optionnel)
- 2 Sélection du type de fichiers d'entrée. L'utilisateur a la possibilité de charger soit une table de comptage et un fichier d'annotation taxonomique soit un fichier au format biom.

L'utilisateur peut également réaliser une étude sans fichier d'annotation, dans ce cas l'analyse est réalisée au même niveau que le fichier de comptage.
- 3 Zones de chargement des fichiers d'entrée. L'utilisateur peut choisir le type de donnée (OTU/Gene ou MGS), le séparateur du fichier et, pour l'annotation RDP, le seuil de probabilité.
- 4 Chargement de l'arbre phylogénétique .tree (optionnel)

Une fois que le chargement des données a été réalisé, SHAMAN affiche les données de l'utilisateur ainsi quelques représentations graphiques permettant de mesurer la qualité de l'annotation.

The screenshot shows the SHAMAN software interface. At the top, there are four status indicators: 'Annotated features' (61.49%, green), 'COUNT TABLE' (green), 'TAXONOMY TABLE' (green), and 'PHYLOGENETIC TREE' (orange with a warning icon). On the left, a sidebar lists navigation options: Home, Tutorial, Download/Install, Raw data, Upload your data, Statistical analysis, and Visualization. The main area has tabs for Count table, Taxonomy, Summary, and Phylogeny. A red box highlights the 'Count table' tab, which displays a table of sample counts across various OTUs. The table includes columns for ATB03 through ATB18 and A1. The first few rows of data are:

	ATB03	ATB04	ATB03.4	ATB05	ATB06	ATB05.6	ATB07	ATB08	ATB07.8	ATB09	ATB10	ATB09.10	ATB12	ATB13	ATB14	ATB15	ATB17	ATB18	A1
OTU_1	3535	322	1829	489396	5860	247628	82518	337147	209833	2947	199	1573	306	148	74	214	25	55	
OTU_10	0	0	0	0	0	0	118	108	113	678	492	585	0	0	16	3732	0	2	
OTU_100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	110	0	0	
OTU_104	0	9	5	0	0	0	0	0	9	5	0	12	6	0	0	0	3	0	80
OTU_105	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTU_106	5	0	3	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
OTU_107	126	0	63	0	0	0	0	0	0	0	0	0	0	0	0	0	80	0	0

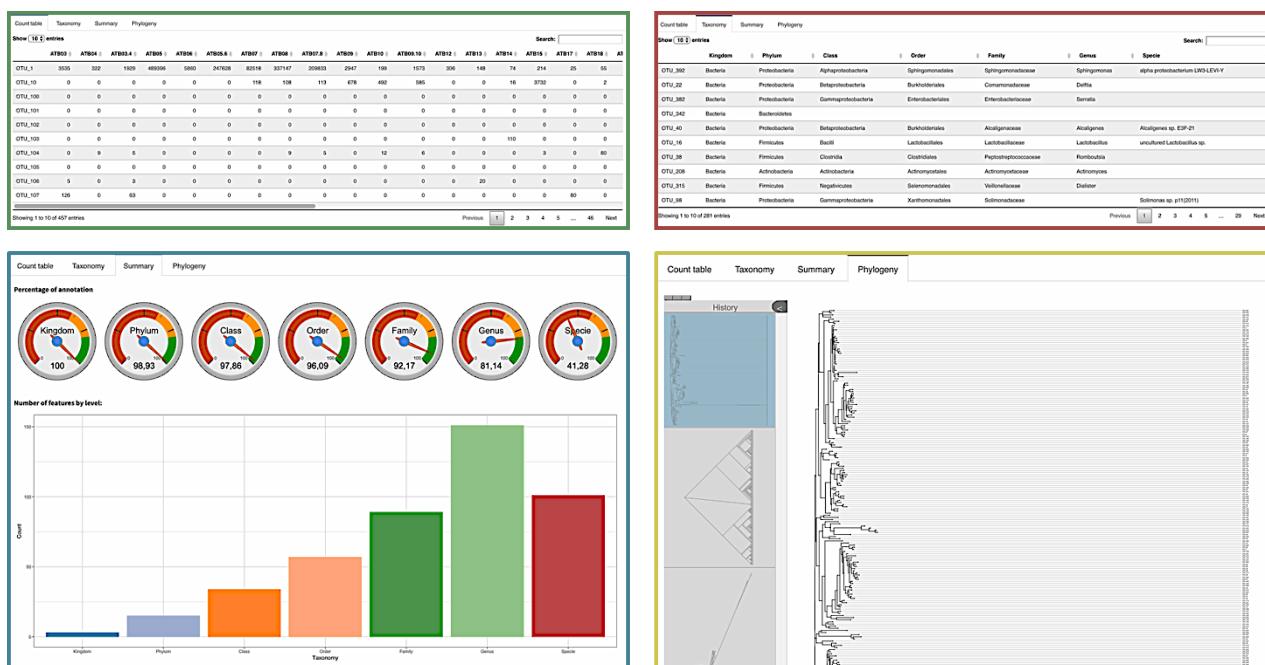
Below the table, it says 'Showing 1 to 10 of 457 entries'. To the right, there are buttons for 'Previous' and 'Next' with page numbers 1, 2, 3, 4, 5, ..., 46. A red circle labeled '1' is at the top right of the status bar, and a red circle labeled '2' is at the bottom right of the status bar.

1 Si le chargement s'est bien passé, les boîtes deviennent vertes et indiquent le pourcentage d'annotation.

2 Zone présentant un aperçu des fichiers chargés et des statistiques de base. Le détail des 4 onglets est donné en page 8.

La table de comptages doit se présenter sous le format suivant : échantillons en colonne et les individus (OTU, ...) en ligne.

Les 4 onglets de la boîte précédente sont les suivants



- 1 Table des comptages avec les échantillons en colonnes et les OTU/gènes en ligne
- 2 Table d'annotation taxonomique avec les différents niveaux en colonnes les OTU/gènes en ligne
- 3 Représentation le pourcentage d'OTU/gènes annotés à chaque niveau ainsi que le nombre d'individus à chaque niveau
- 4 Représentation de l'arbre phylogénétique (uniquement représenté si le fichier *.tree* a été chargé)

Une fois que les données ont été chargées, deux nouveaux onglets apparaissent « Statistical analysis » et « Visualization ».

Statistical analysis - Cet onglet se décompose en 3 sous-onglets :

- Run differential analysis : permet de définir le modèle statistique ainsi que les différentes options, le niveau taxonomique auquel l'utilisateur souhaite réaliser l'étude et de créer les vecteurs de contrastes nécessaires pour définir les comparaisons souhaitées
- Diagnostic plots : propose de nombreuses visualisations (barplots, boxplots, clustering, PCA, PCoA, NMDS, ...) permettant de contrôler le design de l'expérience, de mesurer l'impact de la normalisation et d'identifier d'éventuelles incohérences.
- Tables : présente les résultats de l'analyse différentielle pour chaque vecteur de contraste défini.

Visualization - Cet onglet se décompose en 2 sous-onglets :

- Global views: fournit de nombreuses représentations interactives (barplots, heatmap, boxplots, krona, arbre d'abondance, courbes de rarefaction, diversités) afin de visualiser, en fonction du design expérimental, le contenu des échantillons ainsi que le résultats de l'analyse différentielle.
- Comparaison plots : si au moins 2 vecteurs de contrastes ont été définis, cet onglet propose un venn diagram et un heatmap comparer les résultats de l'analyse pour chaque contraste.

ANALYSE STATISTIQUE

L’analyse statistique de SHAMAN repose sur l’utilisation du package DESeq2. L’utilisation de ce package requiert la définition d’un modèle statistique ainsi que de fixer de nombreuses options (présentées dans la suite de ce tutoriel). Les différentes étapes se résument ainsi :

1. Normalisation des données (calcul des size factors)
2. Estimation de la dispersion à l’aide d’une modélisation entre la moyenne des comptages normalisés et la dispersion « empirique »
3. Ajustement d’un modèle linéaire généralisé avec une loi négative binomiale et un lien log.
4. Tests statistiques sur les paramètres du modèle (test de Wald)
5. Filtrage des « outliers »
6. Corrections des tests multiples

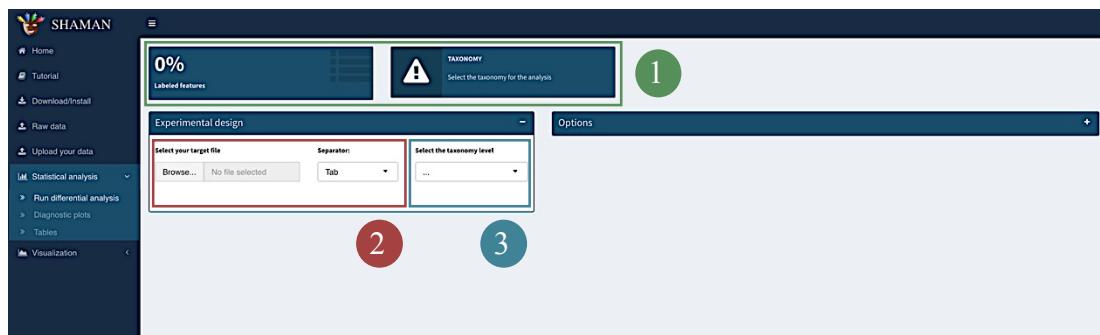
Il a été montré que cette méthode est la plus adaptée pour identifier les éléments différenciellement abondants.

Définir le modèle statistique

Afin de définir le modèle statistique, l’utilisateur doit tout d’abord fournir le design expérimental (*target file*) et sélectionner le niveau taxonomique auquel il souhaite réaliser l’étude. Il est ensuite nécessaire de sélectionner les variables d’intérêt décrivant le phénomène biologique étudié ainsi que les éventuelles interactions entre ces variables. Avant de lancer l’analyse, l’utilisateur peut régler les différentes options relatives à DESeq2 telles que l’*independent filtering*, la forme de la modélisation de la dispersion, la méthode de correction des tests multiples … (pour plus d’information voir [Love et al., 2014]).

Ces différentes opérations doivent être réalisées dans le sous-onglet « Run differential analysis ».

Chargement du design expérimental

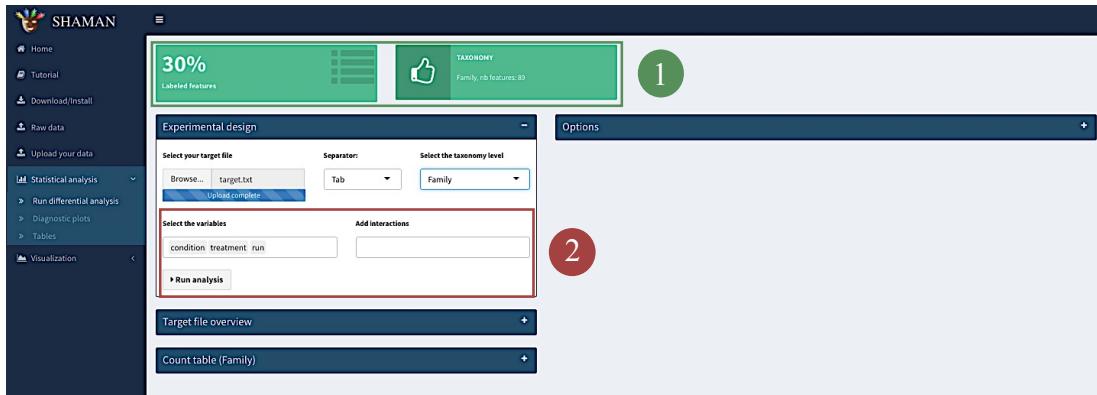


- 1 Information sur le design expérimental et le choix du niveau taxonomique.
- 2 Zone permettant de charger le design expérimental. L'utilisateur peut également définir le séparateur utilisé dans son fichier d'entrée. Ce fichier doit respecter un certain format (cf Annexe A pour plus d'information).
- 3 Choix du niveau d'annotation taxonomique. Ces niveaux correspondent à ceux fournis dans le fichier d'annotation.

Si aucun fichier d'annotation n'a été fourni, le fichier a uniquement la possibilité de choisir « OTU/gene », ce qui correspond au niveau auquel la table de comptage a été calculée.

Création du modèle

Un fois le fichier « target » chargé et le niveau taxonomique sélectionné, SHAMAN vérifie que le nom des échantillons correspond aux noms de la table de comptage et fusionne les **comptages** des OTU/gènes ayant la même annotation.



- 1 Si le chargement s'est bien passé, les boites deviennent vertes et indiquent le pourcentage d'échantillon présent dans la table de comptage pour lequel SHAMAN dispose d'une information provenant du design chargé par l'utilisateur.
- 2 Sélection des variables d'intérêt de l'étude (provenant du design). Les variables peuvent être qualitatives et quantitatives. Il est également possible d'ajouter des interactions entre les variables.

Dans l'exemple présenté, le modèle comporte 3 variables :

- condition : 2 modalités « WT » et « KO »
- treatment : 2 traitements « A » et « B »
- run : effet « batch » qui permet de prendre en compte le fait que le séquençage des échantillons n'a pas été fait en même temps

Dans ce cas, il serait judicieux d'ajouter une interaction entre les variables condition et treatment. En effet, on peut supposer que l'effet du traitement dépend de la condition et donc vouloir tester les effets par sous-groupes. Ici, pour simplifier la notation cet effet a été négligé.

Attention : il est recommandé d'éviter d'utiliser des chiffres pour les variables qualitatives. Par exemple, pour numérotter le « run », il faut préférer la notation « r1 » et « r2 » à « 1 » et « 2 ».

The image shows two side-by-side tables from a bioinformatics application.

Target file overview:

sampleID	condition	treatment	run
SN03.4	WT	A	r1
SN07.8	WT	A	r1
SN14	WT	A	r2
ATB03.4	ATB03.4	WT	B
ATB07.8	ATB07.8	WT	B
ATB12	ATB12	WT	B
ATB13	ATB13	KO	B
ATB15	ATB15	KO	B
ATB05.6	ATB05.6	KO	B
SN13	SN13	KO	A

Count table (Family):

	SN03.4	SN07.8	SN14	ATB03.4	ATB07.8	ATB12
Acetobacteraceae	9580	41940	494	12196	21174	1
Aeromonadaceae	2868	108150	74	46	63	1
Alcaligenaceae	0	0	1	115	0	1
Burkholderiaceae	0	6	1	87	3179	1
Carnobacteriaceae	0	17	194	0	0	1
Comamonadaceae	4	1	74	462	1	1
Corynebacteriaceae	0	2	65	22	2	1
Enterobacteriaceae	175589	678417	128624	3164	219289	521
Family XI	4	18	5	256	24	1
Flavobacteriaceae	14333	16782	20885	18519	9546	1983

1

Aperçu du fichier target chargé par l'utilisateur. Ce fichier contient les informations permettant de décrire chaque échantillon par une ou plusieurs variables (qualitatives et/ou quantitatives).

L'utilisateur a la possibilité de supprimer certains échantillons de mauvaise qualité et d'exporter ce nouveau fichier target.

2

Aperçu de la table des comptages normalisés « agrégée » au niveau taxonomique sélectionné par l'utilisateur.

L'utilisateur a la possibilité d'exporter la table des comptages normalisés et/ou les abondances relatives

Options du modèle et de la normalisation

Options			
Statistical model	Normalization	Filtering	
Type of transformation	Independent filtering	p-value adjustement	Level of significance
<input checked="" type="radio"/> VST <input type="radio"/> rlog	<input checked="" type="radio"/> True <input type="radio"/> False	<input checked="" type="radio"/> BH <input type="radio"/> BY	0.05
Cooks cut-off	Local function	Relationship	
<input checked="" type="radio"/> Auto <input type="radio"/> No cut-off <input type="radio"/> Value	<input checked="" type="radio"/> Median <input type="radio"/> Shorth	<input checked="" type="radio"/> Parametric <input type="radio"/> Local	

- **Type of transformation**

Deux types de transformation sont disponibles dans DESeq2, VST (Variance Stabilizing Transformation) ou rlog (regularized log transformation). Les données étant hétéroscédastiques, cette transformation permet de supprimer la dépendance de la variance à la moyenne. Elle est uniquement appliquée pour représenter et/ou classifier les données (elle n'intervient pas dans la modélisation).

Lorsque le nombre d'échantillons est grand, il est recommandé d'utiliser la transformation VST qui est plus rapide que le rlog.

- **Independent filtering**

Ce filtre basé sur la moyenne des comptages sur tous les échantillons permet de filtrer les individus (espèces, genres, ...) qui ont très peu de chance d'être différentiellement abondants. Le seuil utilisé pour le filtre est déterminé tel que le nombre d'individus significatifs soit maximum pour un FDR donné.

- **p-value adjustment**

Deux méthodes classiques de correction des tests multiples par control du FDR sont proposées : Benjamini-Hochberg et Benjamini-Yekuteli.

- **Level of significance**

Seuil de significativité (risque de 1ère espèce). Par défaut, cette valeur est fixée à 5%.

- **Cooks cut-off**

Le calcul de log-fold change est fortement influencé par la présence de valeurs aberrantes. Le calcul de la distance de Cooks, qui correspond à l'impact de supprimer un échantillon sur l'estimation des paramètres, permet de détecter ces valeurs aberrantes. Le seuil choisi pour cette distance correspond au 99^{ème} percentile de la distribution de Fisher $F(p, m-p)$, avec p le nombre de paramètres et m le nombre d'échantillons)

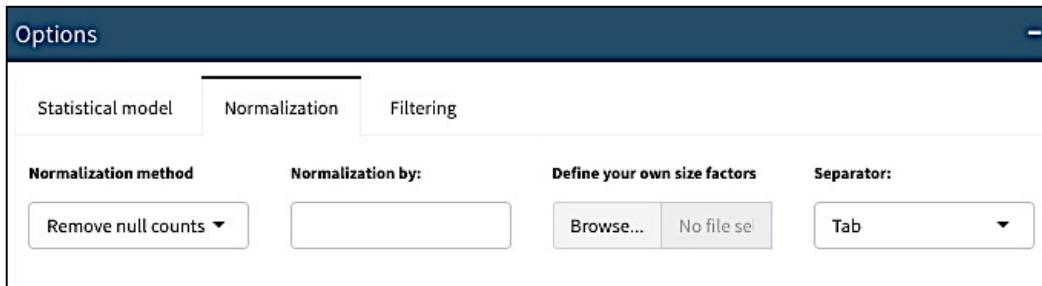
- **Local function**

Pour calculer les size factors (utilisés pour normaliser les données), deux options sont possibles « median » ou « shorth ». La première, « median », correspond à la méthode par défaut qui consiste à calculer la médiane du rapport entre les comptages et la moyenne géométrique (voir la section normalisation pour plus de détails). La seconde, « shorth », calcule la moyenne du plus petit intervalle qui couvre la moitié des valeurs. Cette option est surtout recommandée pour les faibles comptages.

- **Relationship**

La mesure de dispersion pour chaque individu est estimée en modélisant la relation entre la dispersion empirique et la moyenne comptages. Pour modéliser cette relation, trois méthodes sont proposées :

- ✓ une régression paramétrique de la forme $\alpha_{tr}(\bar{\mu}) = \alpha_0 + \frac{a_1}{\bar{\mu}}$
- ✓ une régression locale qui permet d'obtenir une meilleure modélisation lorsque la forme en « 1/x » n'est pas bien adaptée.
- ✓ la moyenne qui associe la dispersion moyenne à tous les individus. Cette approche peut être utilisée lorsque le nombre d'individus est faible.



- **Normalization method**

SHAMAN propose 3 méthodes de normalisation :

- ✓ Usual : méthode par défaut de DESeq2 qui consiste à calculer la médiane du rapport entre les comptages et la moyenne géométrique

$$s_j = \text{median}_i \frac{K_{ij}}{\left(\prod_{k=1}^n K_{ik}\right)^{1/n}}$$

- ✓ Remove null counts : Le calcul est uniquement réalisé sur les comptages non nuls

$$s_j = \text{median}_i \frac{K_{ij}}{\left(\prod_{k \in S_i} K_{ik}\right)^{1/n_i}}$$

- ✓ Weighted : version pondérée de la méthode précédente (les poids correspondent au pourcentage d'échantillons avec une valeur non nulle)
- ✓ Total count : cette méthode consiste à calculer un size factor en divisant le total des comptages de chaque échantillon par la moyenne des totaux sur tous les échantillons. Cette approche est à utiliser lorsque la composition d'une condition à l'autre est très différente, c'est-à-dire lorsque plusieurs espèces sont présentes dans une condition et absentes dans l'autre (et inversement).

Dans le cadre d'une analyse métagénomique, les matrices de comptage sont très creuses (beaucoup de 0). Il est donc recommandé d'éviter l'utilisation de la méthode « usual ».

- **Normalization by**

Cette option permet de réaliser une normalisation par groupe en fonction d'une variable du design expérimental.

Cette option peut s'avérer utile lorsque l'utilisateur souhaite visualiser les résultats de plusieurs études sans influer sur la normalisation.

Attention : il faut éviter de sélectionner une variable d'intérêt pour normaliser les données car les différences entre les groupes peuvent être accentuées.

- **Define your own size factors**

L'utilisateur peut charger son propre vecteur de size factors (obtenus par une autre méthode par exemple).

- **Separator**

Séparateur du fichier contenant les size factors.

Appliquer un filtre sur les données (optionnel)

Dans certains cas, lorsque beaucoup d'éléments ont un faible nombre comptages et/ou ne sont détectés que pour un faible nombre d'échantillons, il peut alors être judicieux de les filtrer avant de lancer l'analyse. Ceci n'aura pas une forte influence sur les résultats de l'analyse différentielle. En effet, l'independant filtering réalisé par DESeq2 permet déjà de filtrer ces éléments. Pour les études avec beaucoup de données cela permet de diminuer les temps de calcul.



- 1 Seuil sur le nombre d'échantillons. Par défaut, le seuil fixé par SHAMAN correspond à 20% des échantillons.
- 2 Seuil sur la moyenne d'abondance. Pour calculer un seuil automatiquement, SHAMAN réalise une régression linéaire entre le nombre d'individus avec une abondance moyenne d'au moins x et l'abondance moyenne.
- 3 Graphique représentant le nombre d'échantillons en fonction de la moyenne de l'abondance. Les points rouges correspondent aux éléments qui seront filtrés une fois les 2 filtres appliqués.

Définir un vecteur de contraste

Une fois l'analyse statistique réalisée, SHAMAN fournit une liste de paramètres correspondant aux variables incluses dans le modèle. A partir de l'estimation de ces paramètres, l'utilisateur a la possibilité de tester différents effets liés aux différentes modalités des variables. On note β_i le vecteur de paramètres. Soit c un vecteur de contrastes, on peut définir une combinaison linéaire des paramètres qui permet de tester les effets souhaités. On utilise ensuite un test de Wald dont la statistique de test $\beta_i^c = c^t \beta_i$ est distribuée selon une loi normale centrée réduite :

$$\frac{\beta_i^c}{\sqrt{c^t \Sigma_i c}} \sim N(0, 1)$$

La matrice Σ_i correspond à la matrice de variance-covariance des paramètres du modèle.

Exemple 1 :

Si l'utilisateur souhaite tester s'il existe un effet du traitement (A vs B), cela revient à tester si la différence A-B est nulle. Il suffit alors de créer un vecteur composé d'un 1 et d'un -1 pour les paramètres associés aux traitements A et B. Les hypothèses du test seront les suivantes :

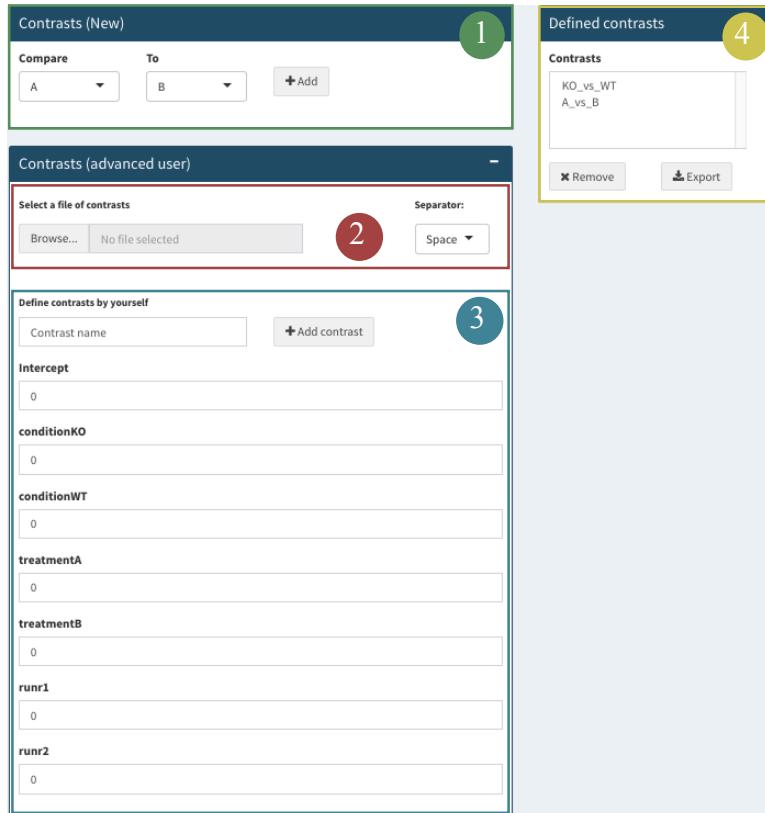
$$\begin{cases} H_0: \beta_A - \beta_B = 0 \\ H_1: \beta_A - \beta_B \neq 0 \end{cases}$$

Dans un plan d'expérience complexe, l'écriture des vecteurs de contrastes permet de tester de nombreuses hypothèses entre les différentes variables.

Exemple 2 :

On suppose qu'on dispose d'un plan d'expérience avec des mesures pour 3 traitements A, B et C. L'utilisateur souhaite savoir si l'effet du traitement C correspond à la moyenne des traitements A et B. Dans ce cas, 3 paramètres seront estimés par le modèle et il faudra créer le vecteur de contraste suivant pour réaliser la comparaison souhaitée :

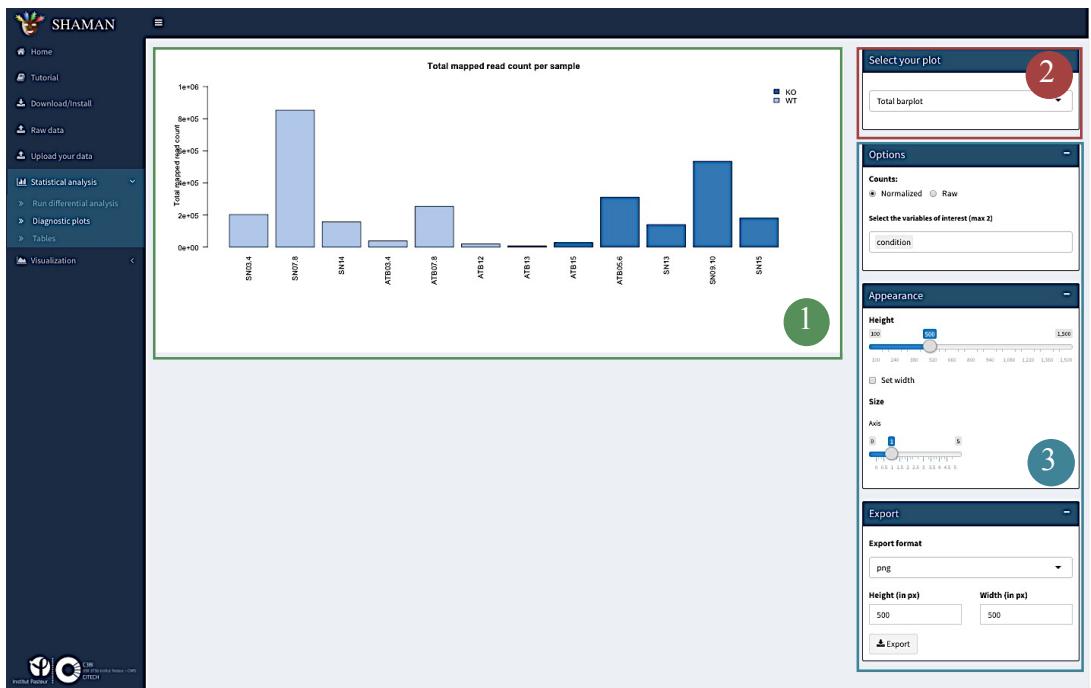
$$c^t \beta = \left[\begin{array}{ccc} \frac{1}{2} & \frac{1}{2} & -1 \end{array} \right] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \frac{1}{2}\beta_1 + \frac{1}{2}\beta_2 - \beta_3 = 0$$



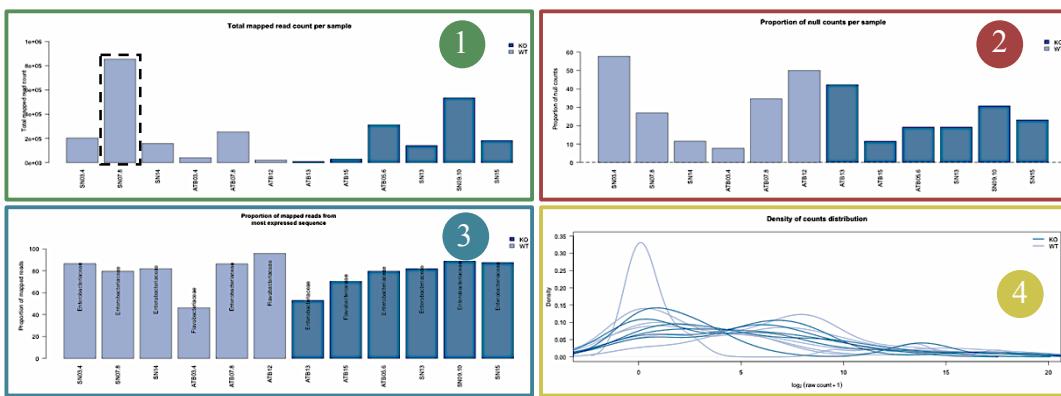
- 1 Création automatique des vecteurs de contrastes en fonction des variables du modèle. Permet de définir la plupart des vecteurs de contrastes usuels.
- 2 Chargement d'un fichier de contraste créer à partir de SHAMAN. Les variables doivent correspondre exactement à celles qui ont été utilisées pour créer le fichier de contrastes.
- 3 Crédit manuelle des vecteurs de contrastes. Cette zone est réservée aux utilisateurs avancés sachant définir un vecteur de contrastes à partir des paramètres du modèle statistique.
- 4 Liste des contrastes créés (par l'une des 3 possibilités). L'utilisateur peut exporter les vecteurs de contrastes en fichier .txt (pour une utilisation ultérieure) et/ou supprimer certains contrastes.

Vérification du modèle statistique

Une fois le modèle défini et l'analyse réalisée, l'utilisateur doit vérifier que la normalisation, l'estimation de la dispersion et des size factors ont correctement été effectuées (barplots, boxplots,...). Au travers de différentes méthodes d'ordination (PCA, PCoA, NMDS) l'utilisateur pourra vérifier qu'il n'y a pas eu d'inversion d'échantillons et que l'effet biologique étudié est bien observé dans les données. L'ensemble de ces représentations se trouvent dans le sous-onglet « Diagnostic plots ».



- 1 Représentation du graphique sélectionné.
- 2 Sélection du graphique à représenter.
- 3 Ensemble des options pour les graphiques (dépendant du type de graphique sélectionné). Possibilité de choisir la ou les variables à représenter, de modifier l'apparence et d'exporter la figure dans différents formats (png, pdf, eps, ...).



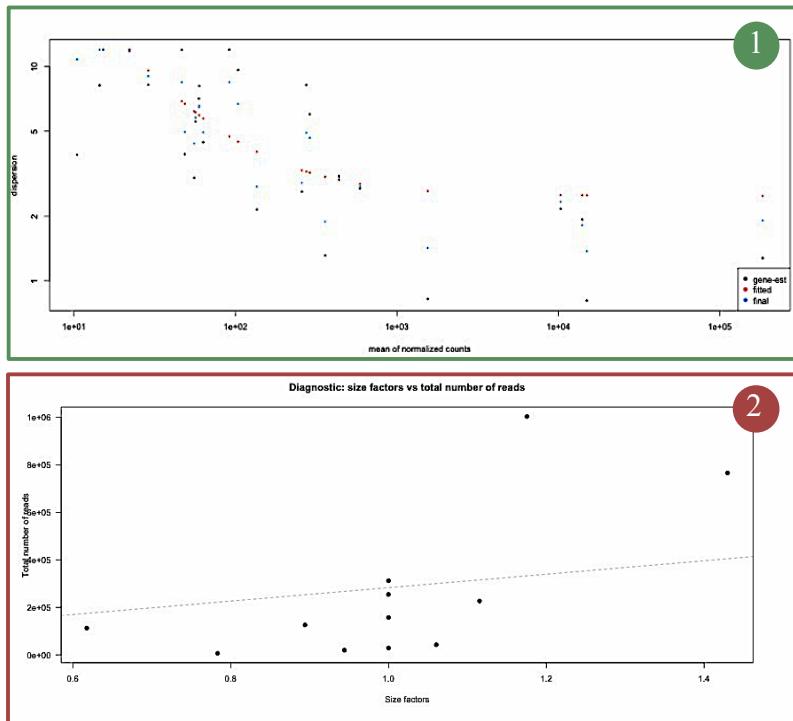
- 1 Diagramme en barres des comptages par échantillon
- 2 Diagramme en barres des comptages nuls par échantillon
- 3 Diagramme en barres des individus les plus abondants
- 4 Représentations de la densité des comptages en log2.

Interprétation :

Ces diagrammes permettent d'identifier d'éventuels problèmes dans les échantillons (problèmes de séquençage par exemple). Dans l'exemple présenté, un des échantillons « SN07-8 » (en pointillés) semble avoir un nombre de reads alignés important par rapport aux autres. Cet échantillon ressort également lorsqu'on observe la représentation des densités.

Dans ce contexte, l'utilisateur doit porter attention à cet échantillon et vérifier si l'ensemble du processus bio-informatique s'est bien déroulé. Il faudra alors décider de supprimer ou non cet échantillon de l'analyse.

Remarque : Les courbes de raréfaction présentées dans la partie « Visualisation » peuvent également être utilisées pour décider de supprimer ou non un échantillon.

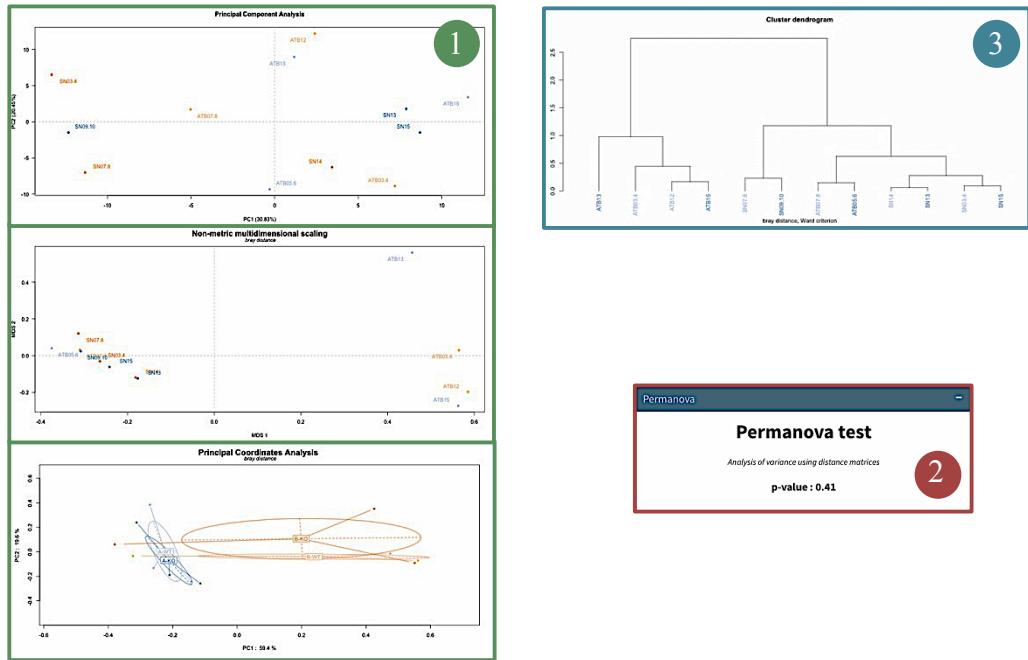


- 1 Représentation de la moyenne des comptages normalisés en fonction de la dispersion.
- 2 Représentation du nombre total de lecture alignés en fonction des size factors.

Interprétation :

Ces deux représentations permettent de vérifier que la normalisation ainsi que l'estimation de la dispersion ont bien été effectuées. Concernant l'estimation de la dispersion, l'utilisateur doit vérifier que la modélisation choisie correspond bien à la forme du nuage de points. Dans l'exemple présenté ici, la fonction par défaut (« parametric ») ne fournissait pas une modélisation satisfaisante, il a été alors décidé d'utiliser l'option « local » pour obtenir une modélisation plus cohérente avec les données.

La représentation du nombre total de lecture alignés en fonction des size factors permet de vérifier la normalisation des données ainsi que de repérer d'éventuels points aberrants. Sur ce graphique, il faut vérifier que les size factors se trouvent dans un intervalle raisonnable (entre 0.5 et 2.5 environ) et que les points ne sont pas trop éloignés de la ligne représentée en pointillés. En effet, les size factors sont sensés corriger les biais dus à la variation des tailles de librairie, il est donc important de vérifier qu'il y a bien un lien entre size factors et taille de librairie.



- 1 Représentation des 2 premiers axes des 3 méthodes d'ordination (PCA, NMDS et PCoA) en fonction des variables d'intérêts sélectionnées par l'utilisateur.
- 2 Résultat du test de permanova. Ce test repose sur un calcul de distance entre les points. Non disponible pour la PCA.
- 3 Classification hiérarchique des échantillons.

Interprétation :

Les méthodes d'ordination (PCA, NMDS et PCoA) ont un double rôle, elles permettent à la fois de détecter des points aberrants ou des inversions d'échantillons ainsi que de vérifier si l'effet biologique étudié apporte une forte variabilité entre les échantillons. Dans l'exemple présenté, on constate que le premier axe de la PCoA (celui avec le plus fort pourcentage de variance) permet de séparer les 2 traitements A et B.

Lorsque le séquençage des échantillons n'a pas été réalisé au même moment, il est possible de voir apparaître de voir un effet « run ». Dans ce cas, il faut prendre cet effet en variable d'ajustement dans le modèle (il suffit d'ajouter une variable « run » dans le fichier target et de l'incorporer dans le modèle). Le plan d'expérience doit être réfléchi en amont de l'analyse afin d'éviter des problèmes de confusion entre les variables (par exemple : séquencer tous les échantillons du traitement A ensemble et, dans un second temps, ceux du traitement B).

Concernant la p-value associée au test de permanova, elle ne permet pas d'affirmer qu'au moins un des groupes est significativement différent des autres (au risque 5%). Lorsque le test est significatif cela signifie qu'au moins des groupes est différent des autres. Lorsque l'effet biologique étudié est fort, la classification hiérarchique permet de repérer d'éventuelle(s) inversion(s) d'échantillon(s).

Résultats de l'analyse différentielle

Une fois l'analyse réalisée, les résultats sont disponibles sous forme de tableau dans l'onglet « Table ».

Id	baseMean	FoldChange	log2FoldChange	pvalue_adjusted
Aeromonadaceae	10317.61	8.449966e+01	6.401	0.00001358166402451139
Propionibacteriaceae	359.06	6.061851e-02	-4.044	0.00257263629427126
Enterobacteriaceae	184018.57	1.094042e+01	3.452	0.00985817723662584
Moraxellaceae	1547.76	8.210246e+00	3.037	0.00985817723662584
Burkholderiaceae	274.18	2.982792e-02	-5.057	0.0118630789397626
Mitochondria	438.25	7.214072e-02	-3.793	0.020192

1 4 tables de résultats sont disponibles :

- ✓ Significant : liste des éléments détectés comme différentiellement abondants.
- ✓ Complete : liste de tous les éléments.
- ✓ Up, down : listes des éléments détectés comme différentiellement abondants en fonction du signe du fold change.

2 Tableau des résultats :

- ✓ baseMean : moyenne des comptages normalisés sur tous les échantillons.
- ✓ FoldChange : fold change estimé pour le contraste sélectionné
- ✓ Log2FoldChange : valeur précédente transformée en log2
- ✓ Pvalue_adjusted : valeur de la p-value ajustée (en fonction de la méthode sélectionnée par l'utilisateur, BH par défaut)

3 Sélection de la comparaison souhaitée en fonction des vecteurs de contrastes définis par l'utilisateur.

4 Export des résultats. L'utilisateur peut choisir la table à exporter ainsi que le séparateur de valeur.

Interprétation :

Les tables présentées ci-dessus permettent d'identifier rapidement les éléments qui ont une abondance significativement différente entre les conditions étudiées. Les p-values présentées correspondent aux p-values du test de Wald (après correction pour les comparaisons multiples). Ces p-values sont à comparer au seuil de significativité choisi (e.g 5%).

En métagénomique, les effets importants sont généralement recherchés. Les plus faibles sont moins étudiés. La lecture de la table de résultat doit donc être faite en tenant compte de la valeur du baseMean qui renseigne sur l'importance de l'élément étudié (espèce, genre, ...).

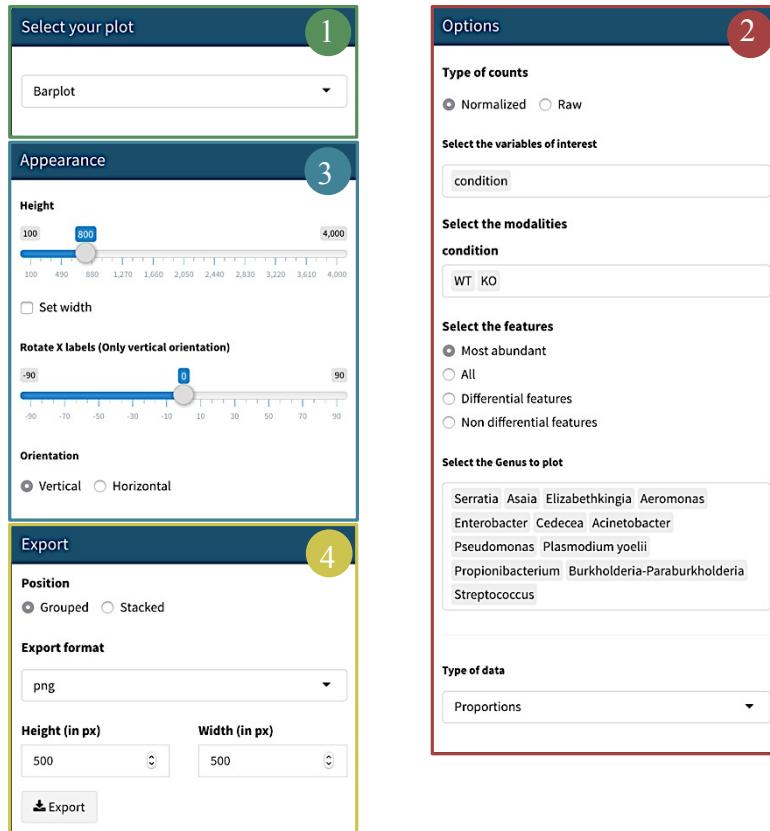
Remarque : le log2FoldChange est obtenu suite à l'estimation des paramètres du modèle et à un « shrinkage » effectué pour minimiser la valeur des log2FoldChange pour les éléments les plus variants (impacte fortement les faibles comptages).

VISUALISATIONS

SHAMAN propose à la fois une analyse statistique robuste des données et un éventail de représentations (barplots, heatmap, boxplots, arbre d'abondance, scatterplot, diversités, courbes de rarefaction, krona et diagrammes de Venn) permettant une visualisation complète des résultats de l'analyse. Chacune des visualisations disponibles dans SHAMAN sont présentées dans la suite de ce tutoriel. Elles sont divisées en 2 catégories :

- ✓ Global views : représentations permettant de visualiser l'abondance des individus en fonction des variables définies dans le target.
- ✓ Comparison plots : représentations spécifiques permettant de comparer les résultats obtenus pour 2 contrastes ou plus.

Quel que soit le type de visualisation utilisé, SHAMAN propose de nombreuses options à l'utilisateur (qui resteront actives d'une représentation à l'autre). Ces options se trouvent dans une colonne située à la droite des visualisations et constituée de 4 éléments.



- 1 Menu déroulant permettant de sélectionner la visualisation souhaitée. 8 visualisations différentes sont proposées : barplot, heatmap, boxplot, Tree, scatterplot, diversity, rarefaction, Krona.
- 2 Liste des options disponibles pour la visualisation sélectionnée. Pour la majorité des visualisations disponibles, l'utilisateur a la possibilité de sélectionner les données qu'il souhaite représenter (normalisées or not), de définir les variables ainsi que de sélectionner la totalité ou seulement une partie des modalités et de sélectionner les éléments à représenter (les plus abondants, tous, seulement les différentiellement abondants ou non pour une des comparaisons définies).
- 3 Liste des options disponibles pour modifier l'apparence du graphique (dépend de la visualisation sélectionnée)
- 4 Export des résultats. L'utilisateur peut choisir le format (png, svg, eps, pdf) ainsi que la taille de la figure à exporter.

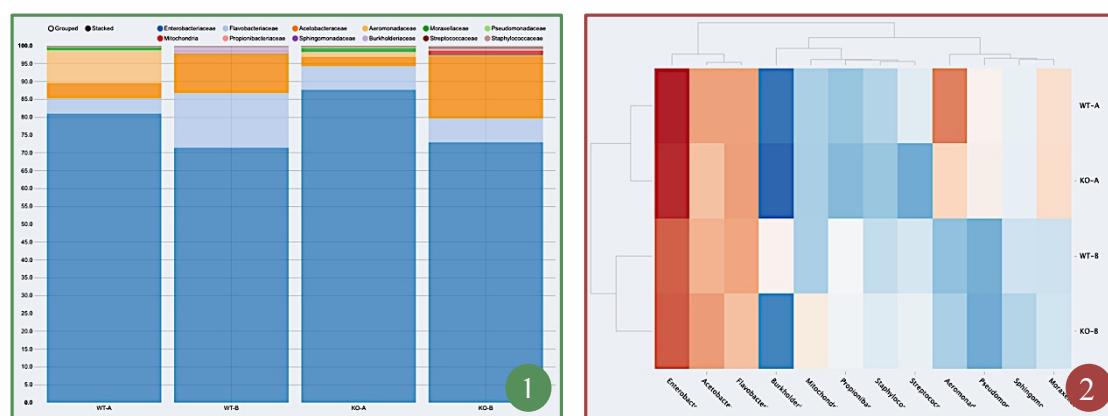
Visualisations des résultats

Une fois l'analyse différentielle réalisée, SHAMAN propose 8 visualisations interactives permettant d'illustrer les effets étudiés sous différentes formes (disponibles dans l'onglet « Global views »).

Remarque : l'interprétation des données 16S à partir des visualisations doit être effectuée avec précaution. En effet, il est possible d'analyser l'abondance d'un élément d'un échantillon (ou condition) à l'autre mais comparer l'abondance de 2 éléments au sein d'un même échantillon (ou condition) n'a pas de sens. Les espèces n'ayant pas le même nombre de copies de leur 16S, il est possible que les différences ne soient dues qu'à une différence dans leur nombre de copies.

Composition globale

Afin de visualiser globalement les résultats de l'analyse, il convient d'utiliser des représentations du type **barplot** ou **heatmap**. Ces représentations permettent de visualiser rapidement les différences entre les conditions étudiées.

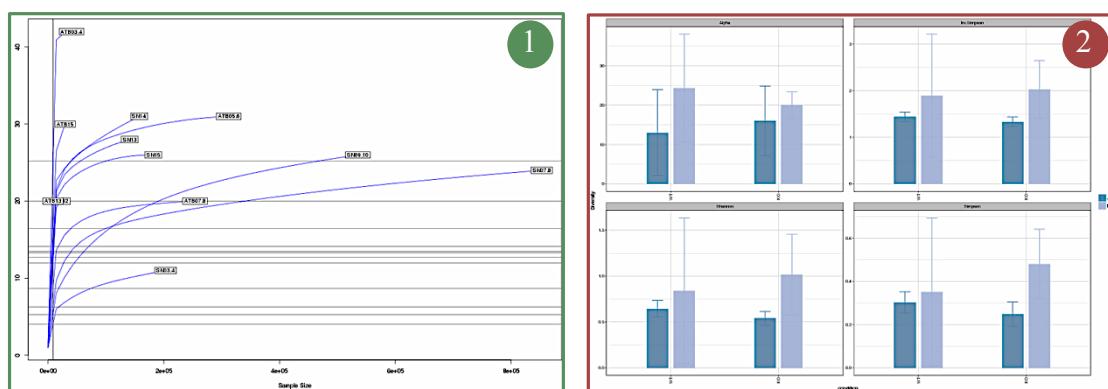


- 1 Barplot des 12 éléments les plus abondants. Chaque couleur représente la proportion de chaque élément.
- 2 Heatmap pour les 12 éléments les plus abondants par variable. La couleur dépend du niveau moyen d'abondance. Rouge foncé = très abondant ; Bleu foncé = Très peu abondant.

Interprétation :

Les 2 figures précédentes montrent que, globalement la composition varie entre les traitements (A et B) mais pas pour les conditions (KO et WT) sur les 12 plus abondants. Ce constat est renforcé par le clustering hiérarchique sur les lignes du heatmap qui regroupe les traitements entre-eux.

Les **courbes de raréfaction** et les mesures de **diversités** permettent également d'avoir un aperçu global de la composition en fonction des métadonnées.



- 1 Courbes de rarefaction pour chaque échantillon. Correspond au nombre d'éléments détectés en fonction de la profondeur de séquençage.
- 2 Mesure de la diversité en fonction des variables d'intérêt. SHAMAN propose 6 mesures différentes : alpha, beta, gamma, shannon, simpson et inverse simpson.

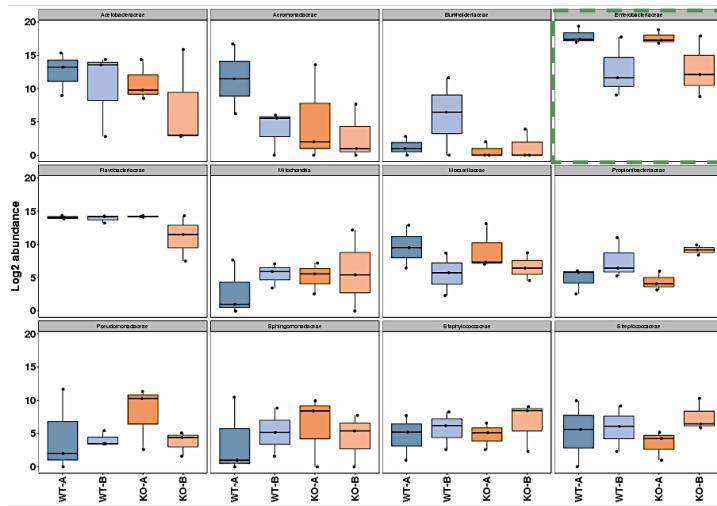
Interprétation :

Les courbes de raréfaction permettent de mettre en avant d'éventuels problèmes de séquençage. Le meilleur cas consiste à obtenir des courbes de raréfaction qui convergent toutes vers un même plateau. Dans l'exemple présenté, les 3 échantillons les plus proches de l'axe des ordonnées montrent que le séquençage n'est pas suffisant. Dans ce cas, il faut vérifier que la normalisation n'est pas la responsable de ce résultat (en affichant les courbes de raréfaction pour les données non normalisées) et, le cas échéant, changer de méthode de normalisation. Si la normalisation n'est pas responsable, il faut vérifier qu'il n'y a pas eu de problème en amont de l'analyse statistique.

Les diversités sont utiles pour illustrer l'impact de la condition biologique étudiée sur la composition des échantillons. Les intervalles de confiance permettent de conclure si la différence est significative ou non (chevauchement ou non des intervalles). L'utilisateur peut également exporter les valeurs utilisées pour les représentations.

Fold-change

Le **boxplot** est la représentation usuelle permettant de visualiser les résultats de l'analyse différentielle et les log2 fold-change estimés.

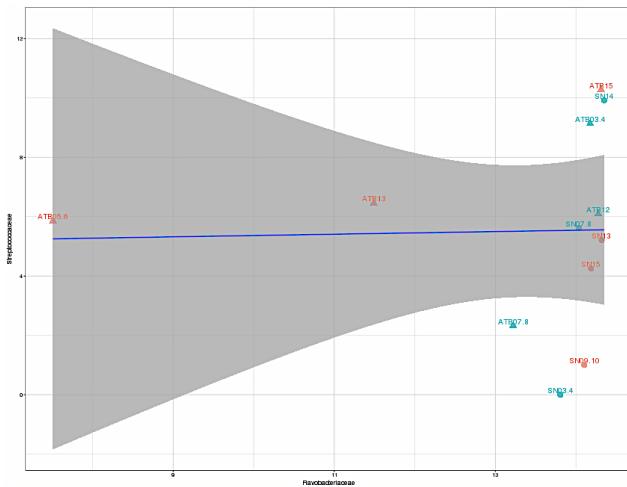


Pour chaque élément sélectionné par l'utilisateur, SHAMAN représente un boxplot pour toutes les modalités des variables d'intérêt (une couleur différente). Cette représentation permet de mettre en avant les différences identifiées lors de l'analyse différentielle. Par exemple, pour la famille *Enterobacteriaceae* (en haut à droite), on voit clairement que le traitement B fait diminuer l'abondance de cette famille (quelle que soit la condition). Cet effet avait été détecté comme significatif dans les tables de résultats (3^{ème} ligne du tableau).

Liens entre les variables

Dans certains cas, il est intéressant de croiser les niveaux d'abondance entre deux éléments, cela permet de mesurer leur corrélation. D'autres possibilités comme modéliser l'abondance d'une espèce par rapport à la diversité ou par rapport à une variable auxiliaire comme l'age, l'imc, ... peut s'avérer informatif notamment pour l'analyse de données cliniques.

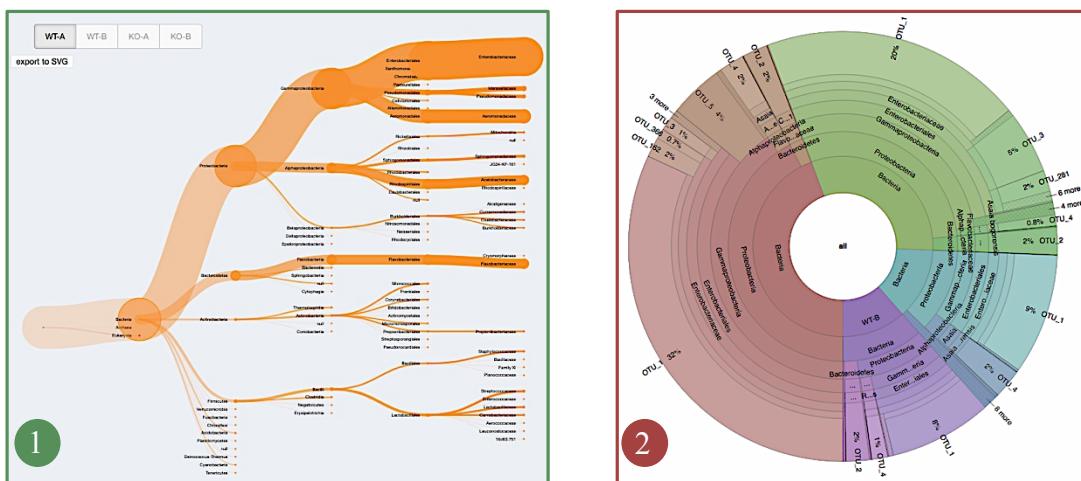
Pour réaliser ces différents croisements, SHAMAN propose un **scatterplot** permettant de représenter tous les couples d'éléments, de changer les points (forme et couleur) en fonction des variables d'intérêt et d'ajouter une troisième variable dont dépend la taille des points.



L'utilisateur a également la possibilité d'ajouter une droite de régression linéaire (ligne bleue) avec la zone de confiance (zone grise). L'équation de la droite, le coefficient de détermination R^2 et les tests pour la pente et l'ordonnée à l'origine sont alors disponibles. L'utilisateur peut également accéder à la table de corrélation (Pearson ou Spearman).

Abondance et taxonomie

Toutes les représentations précédentes permettent de visualiser les résultats de l'analyse au niveau défini par l'utilisateur lors de la définition du modèle statistique.

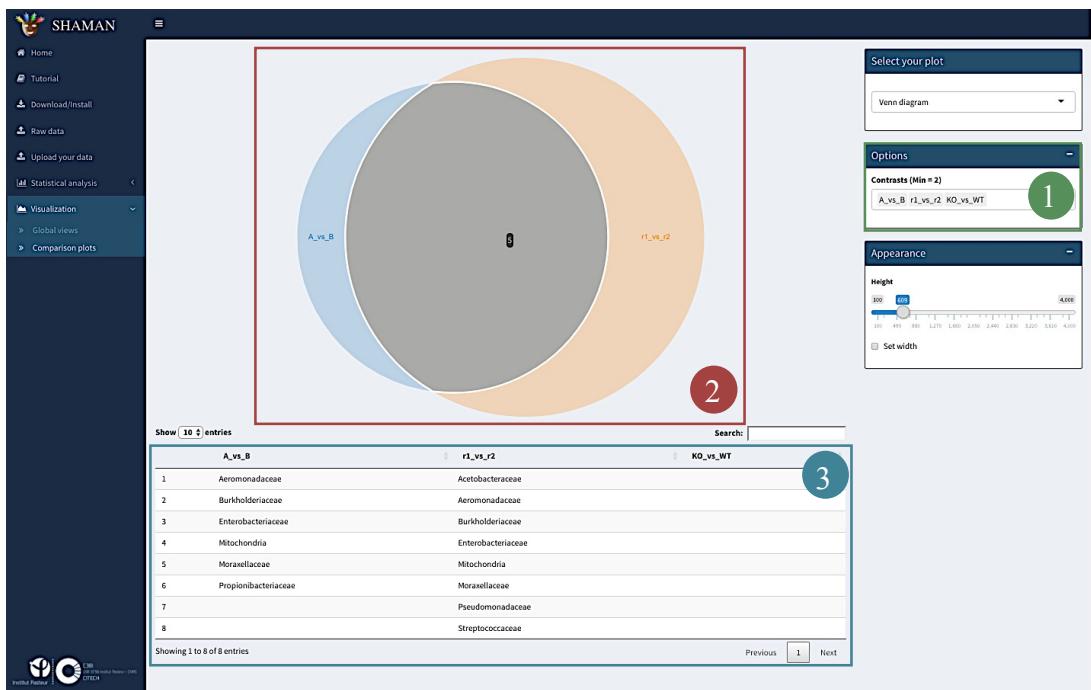


- 1 Arbre d'abondance en fonction du niveau taxonomique. L'utilisateur peut choisir les échantillons qu'ils souhaitent représenter.
- 2 Représentation sous forme de krona¹. Cette visualisation interactive permet de naviguer au travers des différents niveaux taxonomiques.

¹ <https://github.com/marbl/Krona/wiki>

Comparaisons de résultats

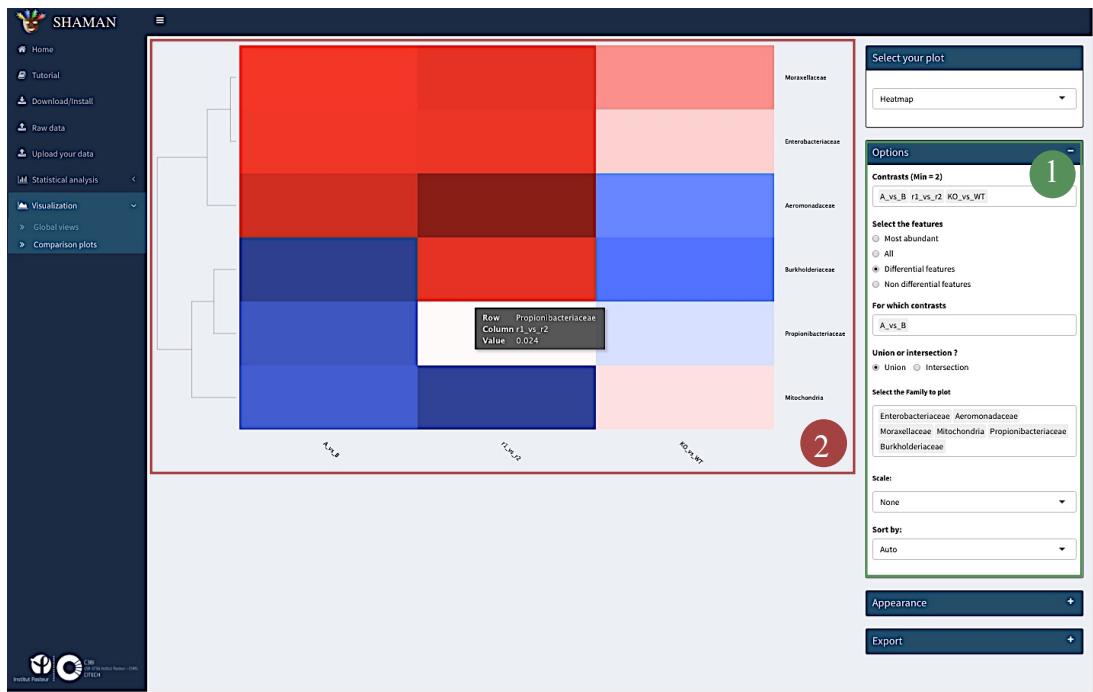
Le principal objectif de l'application SHAMAN est de réaliser une étude quantitative des différences entre des conditions expérimentales. Une approche plus qualitative est proposée dans SHAMAN afin de comparer les résultats de différentes études. Celle-ci s'effectue par le biais de 2 représentations permettant de comparer les éléments détectés comme différentiels pour 2 ou plus comparaisons (diagramme de Venn) ainsi que d'illustrer la force de la différence (heatmap des log₂ fold change). Ce type de représentation permet de comparer les résultats de plusieurs études d'un point de vue qualitatif. Dans de nombreux cas, une analyse quantitative peut également être réalisée (avec 2 comparaisons) en définissant un contraste adapté (mode « advanced user »)



- 1 Sélection des contrastes que l'on souhaite comparer. Au minimum 2 contrastes doivent être sélectionnés.
- 2 Diagramme de Venn entre les éléments différentiellement abondants pour les contrastes sélectionnés.
- 3 Listes des éléments pour chaque contraste.

Interprétation :

Le diagramme de Venn s'interprète simplement en identifiant les éléments communs entre les contrastes. La taille des cercles permet de mesurer le nombre total d'éléments détectés pour chaque contraste.



- 1 Liste des différentes options proposées pour représenter le heatmap.
- 2 Représentation des log2 fold change sous forme de heatmap. La couleur dépend de la valeur du log2 fold change. Rouge foncé : valeur positive élevée ; Bleu foncé : valeur négative élevée ; Blanc : valeur nulle.

Interprétation :

Ce type de représentation permet d'identifier rapidement les éléments communs entre 2 contrastes ou plus. La nuance de couleur donne rapidement le sens et la force de la différence.

BIBLIOGRAPHIE

Anders and Huber, Differential expression analysis for sequence count data, *Genome Biology*, 2010

Jonsson V, Österlund T, Nerman O et al. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics* 2016; 17: 78

Love, Huber and Anders, Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2, *Genome Biology*, 2014

McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLOS Comput. Biol.* 10, e1003531 (2014)

ANNEXE A

How to create a target file for Shaman

format, contents and constraints

The target file is required for each analysis done with Shaman. It must contain all the available information on the samples (corresponding to the metadata) that will be used to build the statistical model and/or to visualize the data. To be loaded in Shaman the target file must respect some properties

- 1. The first column is dedicated to the sample name which must correspond exactly to the column names of the count matrix. At least 2 samples are required.** *Once the count matrix is loaded, check carefully the sample names in the “count table” tab. Sometimes some characters are modified with the loading.*
 - 2. At least one variable must be provided.** *In Example 1, two variables are provided (condition and treatment).*
 - 3. NA or missing values are not allowed.**
 - 4. A variable with the same value for each sample is not allowed.** *This kind of variable should be removed from the target file before loading.*
 - 5. The selected variables for the statistical model must not be collinear.** *It means that if one variable can be determined by another variable or a combination of variables the analysis cannot be done with all the variables. However, the user can use this variable for visualization. (See example 3).*
 - 6. Be careful, numeric variables will be considered as quantitative variable.** *For instance, do not use 1 and 2 to describe two different conditions but C1 and C2 or A and B. (see example 3)*
 - 7. Avoid using special characters such as / \ ? * : < > / + , [] - + { } % @ " &**
-

Example 1: Target file with 2 variables (condition and treatment)

sampleID	condition	treatment
S1	WT	A
S2	WT	A
S3	KO	A
S4	KO	A
S5	WT	B
S6	WT	B
S7	KO	B
S8	KO	B

Error

The model matrix is not full rank. One or more variables or interaction terms are linear combinations of the others and must be removed.

Reminder: Your target file must contain at least 2 columns and 2 rows. NA's values are not allowed and the variables must not be collinear.

This is a usual example in which we have 2 variables to describe the samples (condition and treatment). For instance, the user will be able to define the following model: condition + treatment + condition:treatment and then get differentially abundant features between treatments A and B for each condition.

Example 2: Target file with collinearity problem

sampleID	condition	treatme	group
D		nt	
S1	WT	A	g1
S2	WT	A	g1
S3	KO	A	g2
S4	KO	A	g2
S5	WT	B	g3
S6	WT	B	g3
S7	KO	B	g4
S8	KO	B	g4

In this example, group = condition + treatment, so the variables are collinear.

Note that this file can be loaded in Shaman without error but the error will appear if the user tries to define a model with the three variables condition, treatment and group.

Example 3: Quantitative versus qualitative variable.

Target file

sampleID	condition
S1	1
S2	1
S3	2
S4	2
S5	3
S6	3
S7	4
S8	4

sampleID	condition
S1	C1
S2	C1
S3	C2
S4	C2
S5	C3
S6	C3
S7	C4
S8	C4

Model parameters

condition

(↑)

conditionC1

(↑)

conditionC2

(↑)

conditionC3

(↑)

conditionC4

(↑)

In case 1, condition is considered as a numeric variable which leads to only one parameter in the statistical model. It assumes that the difference between 1 and 3 is two times the difference between 1 and 2 and so on. In case 2, there is no order between the conditions.