# PathoStat User Manual

Sandro L. Valenzuela, Eduardo Castro-Nallar, Solaiappan Manimaran

2016-10-17

Package version: PathoStat 1.0.0

## Contents

1	Introduction	1
2	Installation and Run	2
3	Relative Abundance         3.1       Taxonomy level         3.2       Heatmap         3.3       Summary         3.4       RA Table (%)         3.5       Count table	<b>2</b> 3 4 5 6
4	4.2       Beta Diversity         4.3       Exploratory Tree         4.4       Biplot	<b>6</b> 7 7 8 9 9
5	5.1       Expression Plots         5.2       Summary, Table and LIMMA	
7		13
1		13 14
8	Time Series	15
9	Core OTUs	16

## **1** Introduction

Welcome! This is the manual for the PathoStat package. PathoStat is a Shiny App interactive package that will let you explore metagenomic datsets, e.g., microbiome abundance tables, for exploratory data analysis, differential abundance hypothesis testing, and more. PathoStat can take metagenomic abundance data produced by any taxonomic profiling pipeline, however, to get the most out of PathoStat we recommend using PathoScope2. PathoStat is not limited to whole-metagenomic shotgun data but also can take metataxonomic data, i.e., 16S rRNA, ITS, etc. Plots can be exported in vector-based file formats (svg, PDF) for sharing or further editing. For detailed installation instructions check the introductory and advanced vignettes.

```
require(PathoStat)
vignette("PathoStatIntro")
vignette("PathoStatUserManual")
vignette("PathoStatAdvanced")
```

## 2 Installation and Run

While there are a few functions that can work from the console, PathoStat is designed to be run interactively. From the R console simply type:

```
source("http://bioconductor.org/biocLite.R")
biocLite("PathoStat")
```

If all went well you should now be able to load PathoStat:

```
require(PathoStat)
runPathoStat()
```

PathoStat functionality is grouped into eight tabs that comprise all analyses and visualizations. Here, we will go over each of those eight tabs and their main features.

## 3 Relative Abundance

The relative abundance tab is further subdivided into five subtabs where you can explore your data in the form of a stacked bar chart, a heatmap, summary, and searchable relative abundance and count tables. You can download these data both in read counts or in relative proportions.

#### 3.1 Taxonomy level

Within the Taxonomy level RA subtab, you can hove over different stacked bars and get the taxonomy membership and relative abundance. In addition you can sort the stacked bars (samples) by different factors in your data and download a svg image for further editing.



The image above shows the user interface of the stacked bar chart. On the left you can select the appropriate taxonomy level to plot, and by clicking on the gear icon on the upper right side of the chart, you can select the file format for your download.

#### 3.2 Heatmap

This subtab shows you read count abundance of selected taxonomic levels by sample. As with the previous subtab, you can select what taxonomic level you wish to plot by simply using the dropdown menu at the left of the user interface.



The legend on the right side of the plot shows the scale and the color gradient.

## 3.3 Summary

The Summary table shows summary statistics that may come in handy for detecting trends in the data and identifying outliers. As with the previous subtabs, Summary offers redundant information that points to the distribution of the taxonomic composition of your samples.

• • • • • • • • • • • • • • • • • • •	2	/Library/Frame	works/karamework/v	ersions/3.3/Resources	library/PathoStat/shiny/P	atriuotat - oniny		💁 Publish
								J 10000
PathoStat	Relative Abundance	Diversity [	Differential Expression	n Confidence Regi	on PCA PCoA	Time Series Co	re OTUs	
Taxonomy Level	Tax	onomy level RA	Heatmap St	ummary RA Table	%) Count Table			
phylum		mple_01_1 . :0.00000	Sample_01_2 Min. :0.0000	Sample_01_3 Min. :0.00000	Sample_02_1 Min. :0.000000	Sample_02_2 Min. :0.000008	Sample_02_3 Min. :0.00000	
& Download RA CSV	1st	Qu.:0.00000	1st Qu.:0.0000 Median :0.0000	1st Qu.:0.00000 Median :0.00000	1st Qu.:0.000000 Median :0.001926	1st Qu.:0.001229 Median :0.003149	1st Qu.:0.00000	
A Download Count CSV		n :0.14286 Qu.:0.06453	Mean :0.1429 3rd Qu.:0.1564	Mean :0.14286 3rd Qu.:0.01554	Mean :0.142857 3rd Qu.:0.148268	Mean :0.142857 3rd Qu.:0.135501		
Sort By	Ma> Sa	. :0.87093	Max. :0.6873 Sample_03_2	Max. :0.96892 Sample_03_3	Max. :0.701538 Sample_06_1	Max. :0.723392 Sample_06_2	Max. :0.93721 Sample_06_3	
<ul> <li>None</li> </ul>		. :0.00000	Min. :0.00000			Min. :0.0000	Min. :0.0000	
<ul> <li>Condition</li> </ul>		Qu.:0.05325	1st Qu.:0.06331 Median :0.09046			1st Qu.:0.0000 Median :0.0000	1st Qu.:0.0000 Median :0.0000	
O Batch	Mea		Mean :0.14286			Mean :0.1429	Mean :0.1429	
		Qu.:0.22417	3rd Qu.:0.13609	9 3rd Qu.:0.2454		3rd Qu.:0.1309	3rd Qu.:0.2214	
	Max		Max. :0.51074			Max. :0.7382	Max. :0.5572	
		mple_07_1 . :0.0000	Sample_07_2 Min. :0.0000	Sample_07_3 Min. :0.00000	Sample_08_1 Min. :0.00000	Sample_08_2 Min. :0.00000	Sample_08_3 Min. :0.0000	
		Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.00000			1st Qu.:0.0000	
		ian :0.0000	Median :0.0000	Median :0.01242			Median :0.0000	
		n :0.1429	Mean :0.1429	Mean :0.14286	Mean :0.14286		Mean :0.1429	
		Qu.:0.2250	3rd Qu.:0.1040 Max. :0.7919	3rd Qu.:0.08685 Max. :0.81389			3rd Qu.:0.2544 Max. :0.4911	
		mple_09_1	Sample_09_2	Sample_09_3	Sample_10_1	Sample_10_2	Sample_10_3	
	Mir							
		Qu.:0.008806						
		ian :0.029445 n :0.142857						
		Qu.:0.095886						
	Max	. :0.761171					7 Max. :0.83968	
		mple_12_1	Sample_12_2	Sample_12_3	Sample_13_1	Sample_13_2	Sample_13_3	
		. :0.00000 Ou.:0.00000	Min. :0.0000 1st Qu.:0.0000	Min. :0.0000 1st Qu.:0.0000			Min. :0.0000 1st Qu.:0.0000	
		ian :0.01049	Median :0.0000	Median :0.0000			Median :0.0000	
		n :0.14286	Mean :0.1429	Mean :0.1429			Mean :0.1429	
		Qu.:0.10697	3rd Qu.:0.1630	3rd Qu.:0.2025			3rd Qu.:0.1698	
		. :0.77558	Max. :0.6741	Max. :0.5949	Max. :0.86614	Max. :0.86580	Max. :0.6604	
	Mir	mple_15_1 . :0.00000	Sample_15_2 Min. :0.00000	Sample_15_3 Min. :0.0000				
		Qu.:0.04619	1st Qu.:0.02639					
		ian :0.10964	Median :0.06523	7 Median :0.0810				
	Mea		Mean :0.14286					
		Qu.:0.22644	3rd Qu.:0.15475 Max. :0.57245					
	Haz							

## 3.4 RA Table (%)

This subtab presents a searchable and sortable table of taxa abundance by sample. At user-selected taxonomy levels, you get information about the taxonomy ID (taxid) of the organism as per NCBI's taxonomy database and their relative abundance. All columns are sortable, and at the top right corner of the table you can use the search case to check for specific taxa.

://127.0.0.1:3386 🛛 🖉 Open in Browser 🛛 🤕	9									🤧 Publis
PathoStat	Relative Abundance	Diversity	Differenti	al Expression C	onfidence Region	PCA PCoA	Time Series	Core OTUs		
Taxonomy Level	Tax	onomy level R	IA Hea	tmap Summary	RA Table(%)	Count Table				
phylum	- Show	10 🔁 entr	ies						Search:	
	phy	lum 🕴	taxid 🕴	Sample_01_1 🕴	Sample_01_2	Sample_01_3 🕴	Sample_02_1 0	Sample_02_2	Sample_02_3 🕴	Sample_03_1
Download RA CSV     Download Count CSV	Firmi	cutes	515619	12.69	30.84	3.09	70.15	72.34	93.72	28.76
Sort By	Verru	comicrobia	349741	0.00	0.00	0.00	0.00	0.63	0.00	33.70
None	Actin	obacteria	367928	0.00	0.00	0.00	0.00	0.00	0.00	10.82
Condition Batch	Eury	archaeota	420247	0.00	0.00	0.00	0.00	0.00	0.00	16.07
O Batch	other	s	NA	0.22	0.43	0.02	0.19	0.25	1.60	4.02
	Bact	eroidetes	435590	87.09	68.73	96.89	28.96	26.47	2.85	6.63
	Prote	obacteria	536056	0.00	0.00	0.00	0.69	0.31	1.83	0.00
	_									
	Showi	ng 1 to 7 of 7	entries						Previous	1 Next

#### 3.5 Count table

This table is similar to the previous one but it shows abundance data as raw read counts instead of proportions. This is useful when you want to test for differential abundance of taxa between conditions. Statistical models such as those implemented in EdgeR and DESeq2 explicitly require count data to ensure specificity.

PathoStat	Relative Abundance	Diversity	Differenti	al Expression C	onfidence Region	PCA PCoA	Time Series 0	Core OTUs		
Taxonomy Level	1	axonomy level F	RA Hea	tmap Summary	RA Table(%)	Count Table				
phylum	- Sho	ow 10 [ ont	ries					5	Search:	
Lownload RA CSV	5	ohylum 🔶	taxid 🔅	Sample_01_1 0	Sample_01_2 0	Sample_01_3 0	Sample_02_1 0	Sample_02_2 \$	Sample_02_3 0	Sample_03_1
& Download Count CSV	Fi	rmicutes	515619	19047	58672	6474	141663	133928	165268	4441
Sort By	Ve	errucomicrobia	349741	0	0	0	0	1166	0	5204
<ul> <li>None</li> </ul>	A	ctinobacteria	367928	0	0	0	0	0	0	1670
Condition Batch	E	ıryarchaeota	420247	0	0	0	0	0	0	2481
batch	ot	hers	NA	332	822	46	389	455	2820	620
	B	acteroidetes	435590	130765	130748	203260	58479	49007	5031	1025
	Pr	oteobacteria	536056	0	0	0	1401	583	3221	
	Sho	wing 1 to 7 of 7	entries						Previous	1 Next

## 4 **Diversity**

The second tab in PathoStat is the Diversity tab. Here, users can obtain estimates of alpha and beta diversity, as well as explore a clustering dendrogram decorated with the abundance of certain taxonomic groups by sample. Also, the Biplot subtab illustrates the relationships between user-selected variables using multidimensional scaling. Finally, the Co-Occurrence subtab shows the relationships among taxa at user-defined distances allowing the exploration of potential biological associations.

#### 4.1 Alpha Diversity

Alpha diversity considers presence and absence of taxonomit units as well as their homogeneity or evenness. In this subtab, you obtain estimates for alpha diversity using the Shannon, Simpson, and Inverse Simpson metrics as a function of a factor.



#### 4.2 Beta Diversity

The beta diversity subtab provides the user with a heatmap of sample to sample variation using the Bray-Curtis distance metric. The color key indicates a normalized score (the row z-score) and the dendrogram is constructed using hierarchical clustering.



### 4.3 Exploratory Tree

This subtab shows a dendrogram (tree) among all the samples in your dataset. In the figure below, the tree is decorated with the most abundant taxon and their relative abundance by condition. This tree can reveal not only similarity among samples but also what taxon dominates each sample.



#### 4.4 Biplot

Another useful feature of PathoStat is the biplot. Here, the user can explore potential associations among taxa, and among taxa and factors in the dataset, e.g., condition, treatment, etc. The menu on the left side of the panel lets you select the variables to explore as well as the method to estimate the distance among variables.



#### 4.5 Co-Occurrence

One recurrent question in metagenomics is whether two or more taxa tend to co-occur among a set of samples. In general, one might suspect that if two or more taxa co-occur then there might be a functional relationship among them. In the Co-Occurrence subtab, the user selects relevant taxonomic levels to be compared and the maximum distance that will connect two taxa together. The result is a network graph where edges are proportional to the strength of co-occurrence.



# 5 Differential Expression

The third tab in PathoStat allows you to test for differential abundance of taxa between conditions. This test is analogous to differential expression of genes in transcriptomic experiments and uses similar statistical models.

#### 5.1 Expression Plots

In the Expression Plot subtab, you first select the taxonomy level of choice for the comparison and then you select the primary and secondary covariates. The results are shown as BoxPlots colored by covariate.



## 5.2 Summary, Table and LIMMA

The Summary subtab shows text summary of the data used for the BoxPlots in the Expression Plots subtab. The next two subtabs summarize the results as a table of raw values and as a Fold Change table, respectively.

//127.0.0.1:3386 🛛 🔊 Open in Browser 🛛 🚱																Publisi
Taxonomy Level	Express	sion Plots	Sun	nmary	Table	LIMMA										
no rank 🔹	3	6	7	11	14	18	20	22	27	30	31	1	5	9	12	15
Primary Covariate	1.25	1.50	1.70	15.98	15.94	13.77	1.47	14.12	14.74	13.54	1.66	1.74	1.43	1.62	17.19	15.84
Diet	1.25	12.26	1.70	1.64	1.35	1.25	1.47	1.43	1.41	1.39	1.66	1.74	1.43	1.62	1.41	12.60
Diet	1.25	1.50	1.70	1.64	1.35	1.25	1.47	1.43	1.41	1.39	1.66	1.74	1.43	1.62	1.41	1.22
Secondary Covariate	8.57	1.50	1.70	1.64	1.35	1.25	1.47	1.43	1.41	1.39	1.66	1.74	1.43	1.62	1.41	1.22
SubjectID ·	1.25	18.87	1.70	18.32	16.92	17.93	18.15	1.43	17.09	16.42	1.66	1.74	18.51	1.62	16.29	16.68
No. of Sample(s) Per Primary	1.25	1.50	18.36	1.64	1.35	1.25	15.88	1.43	1.41	1.39	17.07	1.74	12.62	18.01	1.41	1.22
Covariate	1.25	1.50	16.72	1.64	1.35	1.25	1.47	1.43	1.41	1.39	16.50	1.74	1.43	15.31	1.41	1.22
11 0	6.68	1.50	1.70	1.64	1.35	1.25	1.47	1.43	1.41	1.39	1.66	11.20	1.43	1.62	1.41	1.22
No. of Sample(s) Per Secondary Covariate	1.25	1.50	1.70	1.64	1.35	1.25	1.47	1.43	1.41	1.39	16.88	1.74	1.43	1.62	1.41	1.22
	1.25	10.95	13.60	12.73	1.35	1.25	14.13	1.43	1.41	14.52	15.25	1.74	13.80	14.78	11.96	1.22
3	1.25	1.50	17.29	1.64	1.35	1.25	15.69	1.43	1.41	1.39	1.66	1.74	1.43	17.79	1.41	1.22
Sort By Secondary Covariate First	1.25	12.93	1.70	1.64	1.35	1.25	1.47	1.43	1.41	1.39	1.66	1.74	1.43	1.62	1.41	1.22
(Default: Sort By Primary Covariate	1.25	14.60	1.70	1.64	1.35	1.25	1.47	1.43	1.41	1.39	1.66	1.74	1.43	1.62	1.41	1.22
First)	1.25	1.50	1.70	1.64	1.35	1.25	1.47	1.43	1.41	1.39	1.66	1.74	1.43	1.62	1.41	1.22
<ul> <li>Color By Secondary Covariate (Default: Color By Primary Covariate)</li> </ul>	1.25	1.50	1.70	1.64	1.35	1.25	1.47	1.43	13.82	1.39	1.66	10.40	1.43	1.62	1.41	1.22
	1.25	1.50	1.70	1.64	1.35	1.25	1.47	1.43	1.41	1.39	1.66	1.74	1.43	1.62	1.41	1.22
No. of top Differentially Expressed Taxons to display	1.25	1.50	1.70	1.64	1.35	1.25	1.47	13.73	1.41	1.39	1.66	1.74	1.43	1.62	1.41	1.22
10 0	1.25	1.50	14.53	16.90	13.42	13.97	13.51	1.43	1.41	16.94	15.32	1.74	11.04	1.62	16.47	13.84
	1.25	1.50	1.70	1.64	1.35	1.25	1.47	1.43	1.41	1.39	1.66	1.74	1.43	1.62	1.41	1.22
	1.25	1.50	1.70	1.64	14.27	1.25	1.47	1.43	13.99	1.39	1.66	14.24	1.43	1.62	1.41	13.75

nce Diversity								😏 Publish
Diversity	Differential Expression	Confidence Region	PCA PCoA	Time Serie	is Co	ore OTUs		
Express	on Plots Summary	Table LIMMA						
Primary	Covariate: Simple (logFC)	Primary Covariate	: Unrefined (logFC)	AveExpr	F	P.Value	adj.P.Val	
	-0.04		3.73	2.73	5.53	0.01	0.44	
	-5.98		-2.96	6.57	4.20	0.03	0.57	
	-5.48		-5.22	4.46	3.19	0.06	0.65	
	-0.09		-3.79	15.53	3.12	0.06	0.65	
	-5.89		-5.94	6.04	2.45	0.11	0.65	
	-1.95		13.97	2.43	0.11	0.65		
	4.65		1.15	4.48	2.39	0.11	0.65	
	-0.83		-3.35	3.41	2.22	0.13	0.65	
	-0.04		1.97	2.14	2.00	0.16	0.65	
	0.75		2.68	2.64	1.99	0.16	0.65	
	Primary	Primary Covariate: Simple (logFC) -0.04 -5.98 -5.48 -0.09 -5.89 -1.95 -1.95 -0.83 -0.04 -0.04 -0.75	Primary Covariate: Simple (logFC)         Primary Covariate           -0.04         -           -5.98         -           -5.48         -           -0.09         -           -5.89         -           -1.95         -           -0.63         -           -0.04         -	Primary Covariate: Simple (logFC)         Primary Covariate: Unrefined (logFC)           -0.04         3.73           -0.05         -2.96           -5.98         -2.96           -5.48         -5.22           -0.09         -3.79           -5.89         -5.94           -1.95         -0.94           4.65         1.15           -0.83         -3.35           -0.04         1.97           0.75         2.68	Primary Covariate: Simple (logFC)         Primary Covariate: Unrefined (logFC)         AveExpr           -0.04         3.73         2.73           -5.98         -2.96         6.57           -5.48         -5.22         4.46           -0.09         -3.79         15.53           -5.58         -5.94         6.04           -1.95         -0.94         13.97           4.65         1.15         4.48           -0.03         -3.35         3.41           -0.04         1.97         2.14           0.75         2.68         2.64	Primary Covariate: Simple (logFC)         Primary Covariate: Unrefined (logFC)         AveExpr         F           -0.04         3.73         2.73         5.53           -5.98         -2.96         6.57         4.20           -5.48         -5.22         4.46         3.19           -0.09         -3.79         15.53         3.12           -5.89         -5.94         6.04         2.45           -1.95         -0.94         13.97         2.43           4.65         1.15         4.48         2.39           -0.83         -3.35         3.41         2.22           -0.04         1.97         2.14         2.00           0.75         2.68         2.64         1.99	Primary Covariate: Simple (logFC)         Primary Covariate: Unrefined (logFC)         AveExpr         F         P.Value           -0.04         3.73         2.73         5.53         0.01           -5.98         -2.96         6.57         4.20         0.03           -5.48         -5.22         4.46         3.19         0.06           -0.09         -3.79         15.53         3.12         0.06           -5.89         -5.94         6.04         2.45         0.11           -5.89         -5.94         6.04         2.45         0.11           -1.95         -0.94         13.97         2.43         0.11           -1.95         -0.94         13.97         2.43         0.11           -0.83         -3.35         3.41         2.22         0.13           -0.04         1.97         2.14         2.00         0.16           -0.05         2.68         2.64         1.99         0.16	Primary Covariate: Simple (logFC)Primary Covariate: Unrefined (logFC)AveExprFPValueadj,PVal-0.043.732.735.530.010.44-5.98-2.966.574.200.030.57-5.48-5.224.463.190.060.65-0.09-3.7915.533.120.060.65-5.89-5.946.042.450.110.65-1.95-0.9413.972.430.110.65-0.641.154.482.390.110.65-0.83-3.353.412.220.130.65-0.041.972.142.000.160.650.752.682.641.990.160.65

# 6 Confidence Region

Many times researchers are interested in the accuracy of taxon abundance estimates. In this subtab, we provide a way to compare within-sample taxa in terms of their abundance estimate and 95% confidence interval.



On the left-side menu, you can select both the sample and the taxa to be used in the comparison. The results are plotted as a jitter plot that indicates the 95% confidence interval between the selected taxa.

## 7 PCA and PCoA

The next two tabs in PathoStat calculate multidimensional scaling using Principal Components and Principal Coordinates Analysis. These tools can help you odentify overall trends in the data. For instance, you can explore whether your samples are related by some biological condition of interest or by technical batch, in which case you would need to denoise the data first using methods such as surrogate variable analysis.



#### 7.1 Explained Variation

An important aspect of PCA analysis is to understand to what extend the resulting vectors can explain the observed multidimensional variation. In this subtab, you can get an idea of that by exploring the different columns in terms of Proportion of Variance, Cumulative Variance and percent variation, among others.

//127.0.0.1:3386 🛛 🔊 Open i	in Browser 🛛 🚱							📀 Pub			
PathoStat Relative A	Abundance	Diversity Differe	ntial Expression	Confidence Region PCA	PCoA Tim	PCoA Time Series Core OTUs					
Principal Component (x-	axis)	PCA Sur	nmary Table	Explained Variation							
1 Component (y-axis)		Proportion of Variance (%)	Cumulative Proportion of Variance (%)	Percent Variation Explained by Either Condition or Batch	Percent Variation Explained by Condition	Condition Significance (p-value)	Percent Variation Explained by Batch	Batch Significance (p-value)			
2	٢	14.51	14.51	90.30	0.10	0.93	90.20	0.00			
Color By Batch (Default	t: Color By	10.66	25.17	91.80	0.10	0.89	91.70	0.00			
Condition)		9.15	34.32	19.40	11.20	0.27	8.20	0.99			
		7.67	41.99	75.90	1.00	0.66	74.90	0.00			
		6.43	48.41	70.00	4.00	0.29	66.00	0.00			
		5.56	53.97	72.40	4.20	0.24	68.20	0.00			
		5.33	59.30	70.50	1.40	0.63	69.10	0.00			
		4.82	64.13	66.80	3.60	0.36	63.20	0.01			
		4.33	68.46	43.70	17.30	0.07	26.40	0.52			
		4.05	72.50	61.40	11.10	0.08	50.20	0.03			
		3.82	76.32	31.10	6.80	0.39	24.30	0.71			
		3.20	79.51	42.40	13.50	0.12	29.00	0.47			
		2.82	82.33	37.30	3.00	0.62	34.20	0.41			
		2.80	85.14	22.80	7.50	0.40	15.30	0.93			

# 8 Time Series

The Time Series tab allows you to understand variation in relative abundance as a function of time. Strictly speaking, you could use any discrete numerical variable from your dataset for visualization. In the example below, we use SubjectID (not a numerical variable though) and agglomerate the data by Phylum.



# 9 Core OTUs

Many times we would like to understand what is particular and what is general about a set of metagenomic samples. Analogous to pangenome analysis in comparative genomics, the Core OTU tab allows you to identify shared taxa or core taxa among samples.



You first set up a detection threshold, meaning a minimum abundance proportion a OTU must have in order to be considered in the analysis, and the number of samples selected from your dataset. You can also select the taxonomic level of interest at which you would like to perform the analysis.