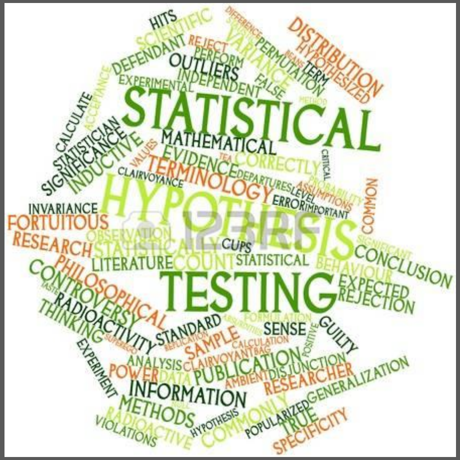
Workflow4metabolomics



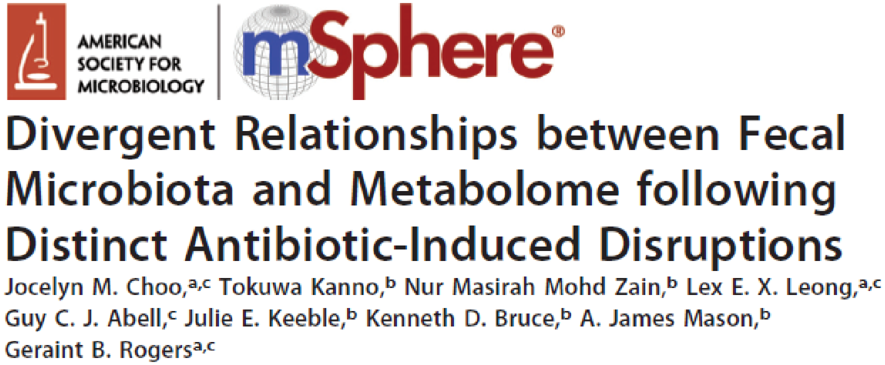
How to perform statistical analysis?

Hypothesis testing

Marie Tremblay-Franco et Gildas Le Corguillé

10/10/2018
v 1.0.0

USED DATA



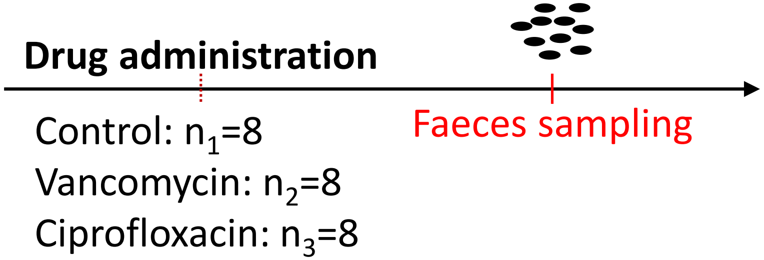
AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

Divergent Relationships between Fecal Microbiota and Metabolome following Distinct Antibiotic-Induced Disruptions

Jocelyn M. Choo,[a,c] Tokuwa Kanno,[b] Nur Masirah Mohd Zain,[b] Lex E. X. Leong,[a,c] Guy C. J. Abell,[c] Julie E. Keeble,[b] Kenneth D. Bruce,[b] A. James Mason,[b] Geraint B. Rogers[a,c]

MTBLS422: Faecal metabolome (Choo et al.)

- **Objective**: characterize how two substantially different antibiotic regimens have an impact on the same mouse gut microbiota and metabolome
- **Experimental design**:
  - Female C57/BL6 mice
  - Treatment: control group / vancomycin-imipenem / ciprofloxacin
  - Sampling of fecal pellet after 14 days on antibiotics



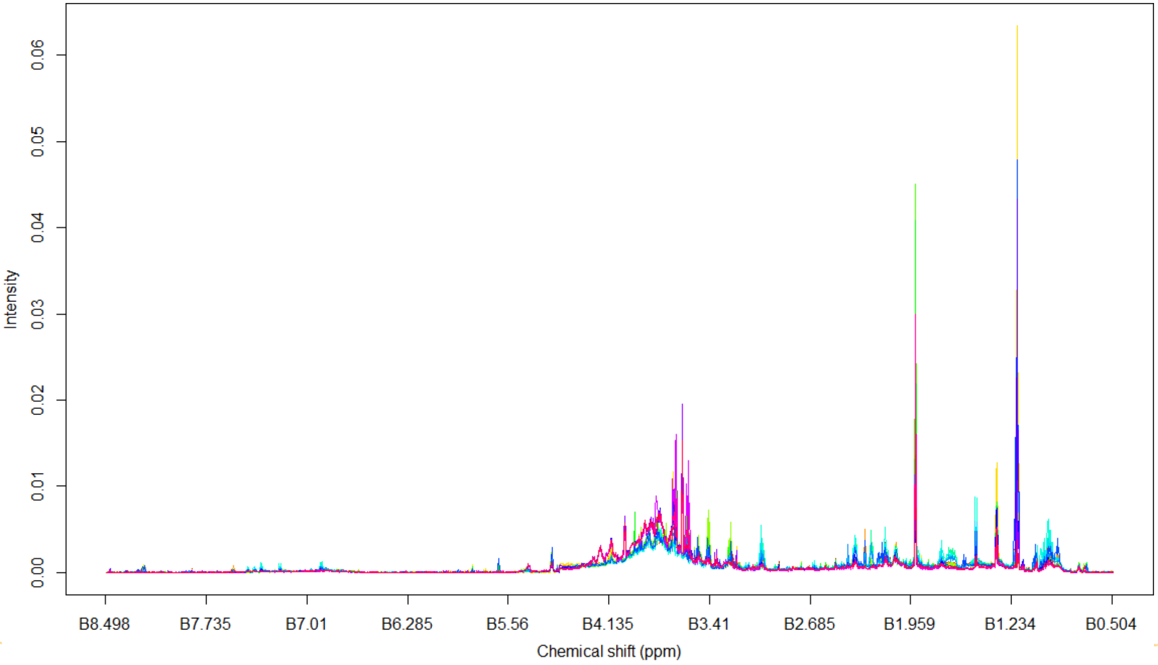
Drug administration

Faeces sampling

Control: n_1=8
Vancomycin: n_2=8
Ciprofloxacin: n_3=8

MTBLS422: Faecal metabolome (Choo et al.)

- **Analytics**: Bruker Avance II 700 NMR spectrometer (Bruker)
  - One-dimensional (1D) Carr-Purcell-Meiboom-Gill presaturation experiment
  - Fourier transformation, calibration to 0.0 ppm, phase correction and automatic baseline correction were applied using the W4M nmr_preprocessing tool
  - Probabilistic quotient normalization (PQN – normalization tool)

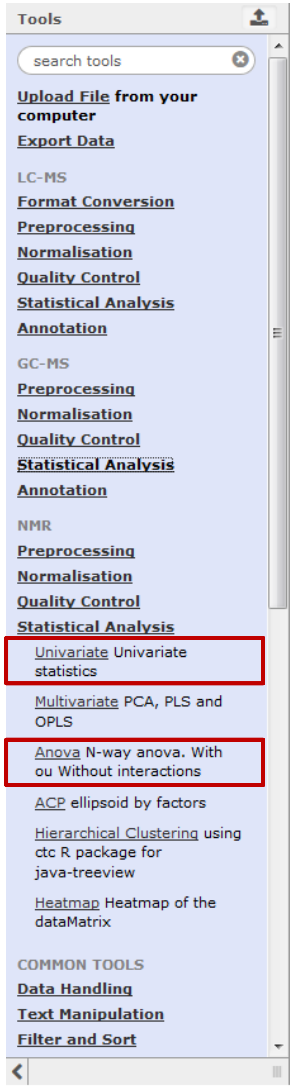


The univariate tools

Univariate

Anova (N-way)

- “**Univariate**” and “**Anova (N-way)**” tools: perform
  - Student / Wilcoxon test
  - N-ways ANOVA / Kruskal-Wallis test
  - Pearson / Spearman correlation test
- Available in the "Statistical Analysis" sections of LC-MS, GC-MS, and NMR
- Can be used either
  - Before multivariate analysis to select significant features


Tools

search tools

Upload File from your computer

Export Data

LC-MS

Format Conversion

Preprocessing

Normalisation

Quality Control

Statistical Analysis

Annotation

GC-MS

Preprocessing

Normalisation

Quality Control

Statistical Analysis

Annotation

NMR

Preprocessing

Normalisation

Quality Control

Statistical Analysis

Univariate Univariate statistics

Multivariate PCA, PLS and OPLS

Anova N-way anova. With ou Without interactions

ACP ellipsoid by factors

Hierarchical Clustering using ctc R package for java-treeview

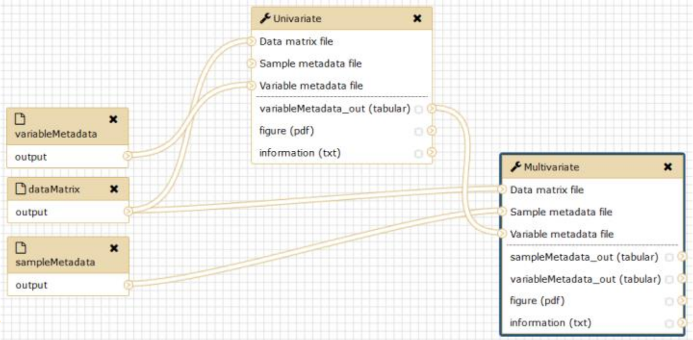
Heatmap Heatmap of the dataMatrix

COMMON TOOLS

Data Handling

Text Manipulation

Filter and Sort

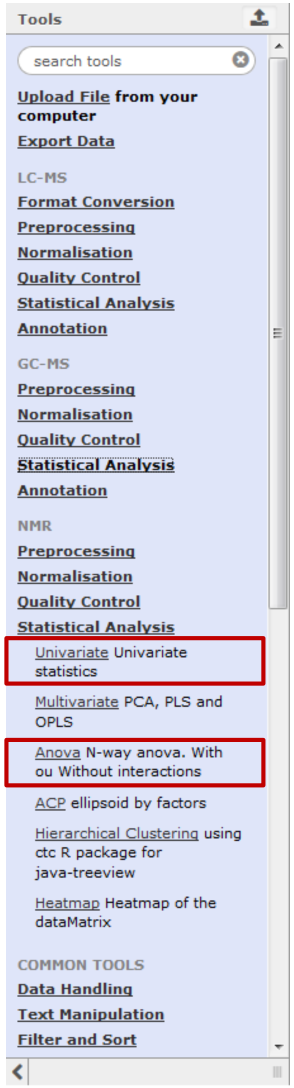


The univariate tools

Univariate

Anova (N-way)

- “**Univariate**” and “**Anova (N-way)**” tools: perform
  - Student / Wilcoxon test
  - N-ways ANOVA / Kruskal-Wallis test
  - Pearson / Spearman correlation test
- Available in the "Statistical Analysis" sections of LC-MS, GC-MS, and NMR
- Can be used either
  - Before multivariate analysis to select significant features
  - After multivariate analysis to test discriminant features


Tools

search tools

Upload File from your computer

Export Data

LC-MS

Format Conversion

Preprocessing

Normalisation

Quality Control

Statistical Analysis

Annotation

GC-MS

Preprocessing

Normalisation

Quality Control

Statistical Analysis

Annotation

NMR

Preprocessing

Normalisation

Quality Control

Statistical Analysis

Univariate Univariate statistics

Multivariate PCA, PLS and OPLS

Anova N-way anova. With ou Without interactions

ACP ellipsoid by factors

Hierarchical Clustering using ctc R package for java-treeview

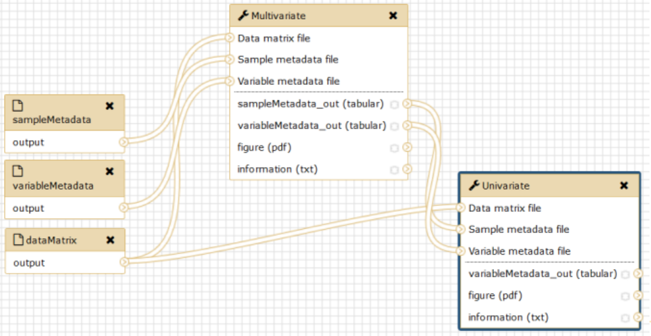
Heatmap Heatmap of the dataMatrix

COMMON TOOLS

Data Handling

Text Manipulation

Filter and Sort



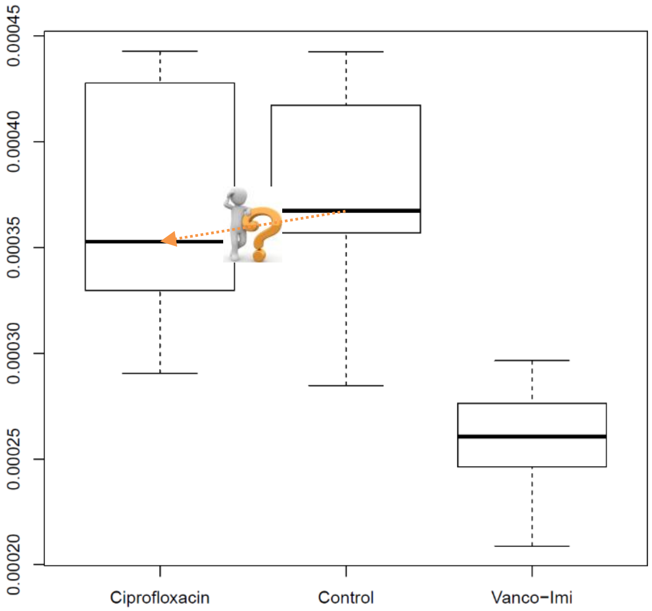
THEORY

INTRODUCTION (1)

- **Graphical charts**: useful to study metabolomic features

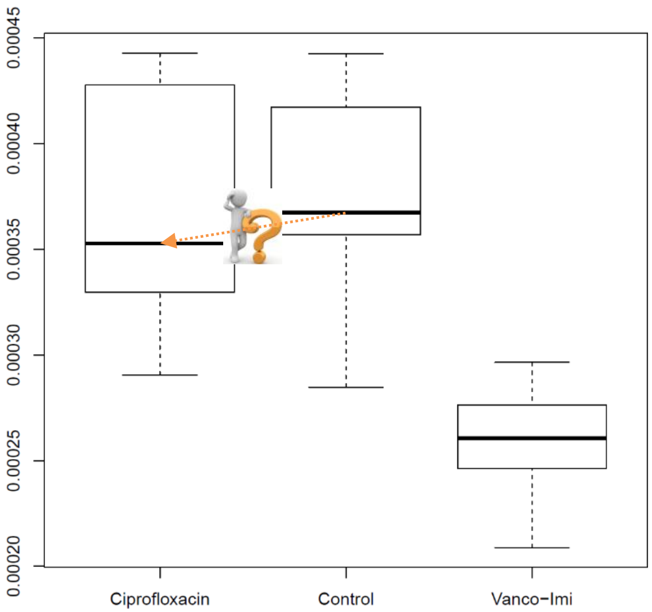
INTRODUCTION (1)

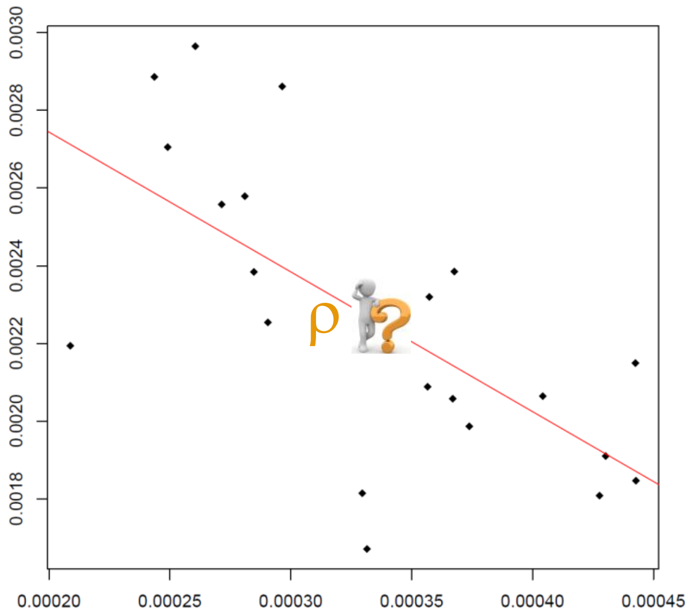
- **Graphical charts**: useful to study metabolomic features but unusable to quantify difference between groups



INTRODUCTION (1)

- **Graphical charts**: useful to study metabolomic features but unusable to quantify difference between groups or relationship between quantitative variables





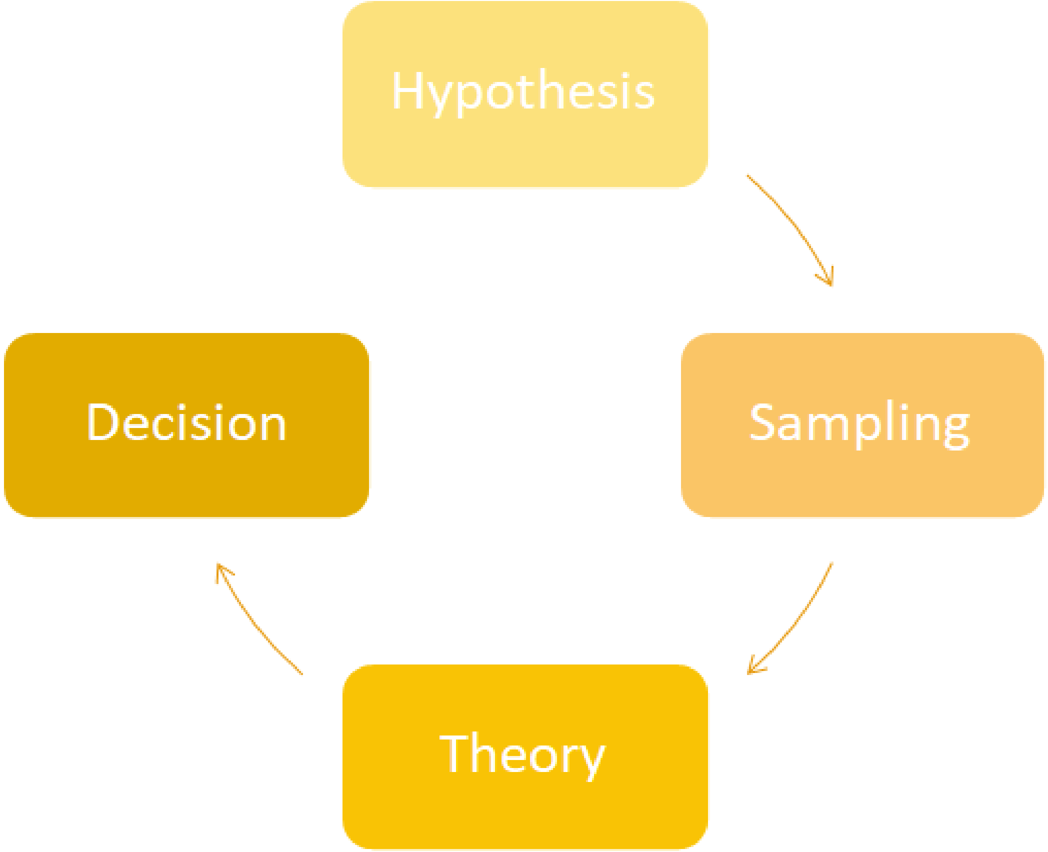
⇒ Inferential methods: draw reliable conclusions

INTRODUCTION (2)

- **Hypothesis testing**: method used to decide whether the observed difference between two means (or the relationship between two features) is *real (significant)* or is simply *due to chance* (sampling fluctuations: physiology, technical variability, …)
    - Ex: comparison of asparagine concentration measured in fecal samples of individuals in a control population and in a vancomycin-treated population

INTRODUCTION (3)

- **Hypothesis testing**: 4-steps procedure
  - Working hypothesis statement
  - Sample collection
  - Theory
  - Decision



Hypothesis

Sampling

Theory

Decision

HYPOTHESIS STATEMENT (1)

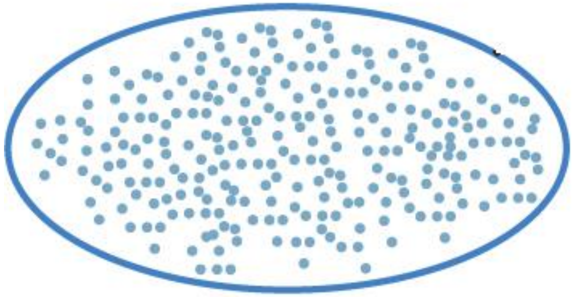
- **Null hypothesis H_0**: statement of the biological question
  - Prior hypothesis: usually hypothesis of no difference or relationship
  - Observed difference results purely from chance (sampling fluctuations)
- **Alternative hypothesis H_1** (or research hypothesis): stated hypothesis if H_0 is rejected
  - Observed difference does not purely result from chance but from factor of interest
- Example: Asparagine intensity in the control group and in the vancomycin-treated group
  - H_0: $\mu_C = \mu_V$
  - H_1: $\mu_C \neq \mu_V$

HYPOTHESIS STATEMENT (2)

- **Null hypothesis H_0**: statement of the biological question
  - Prior hypothesis: usually hypothesis of no difference or relationship
  - Observed difference results purely from chance (sampling fluctuations)
- **Alternative hypothesis H_1** (or research hypothesis): stated hypothesis if H_0 is rejected
  - Observed difference does not purely result from chance but from factor of interest
- Example: Asparagine concentration in the control group and in the vancomycin-treated group
- H_0 is assumed to be true until proven otherwise: data (evidence) is collected to see if H_0 may be rejected and the H_1 may be supported

SAMPLE COLLECTION

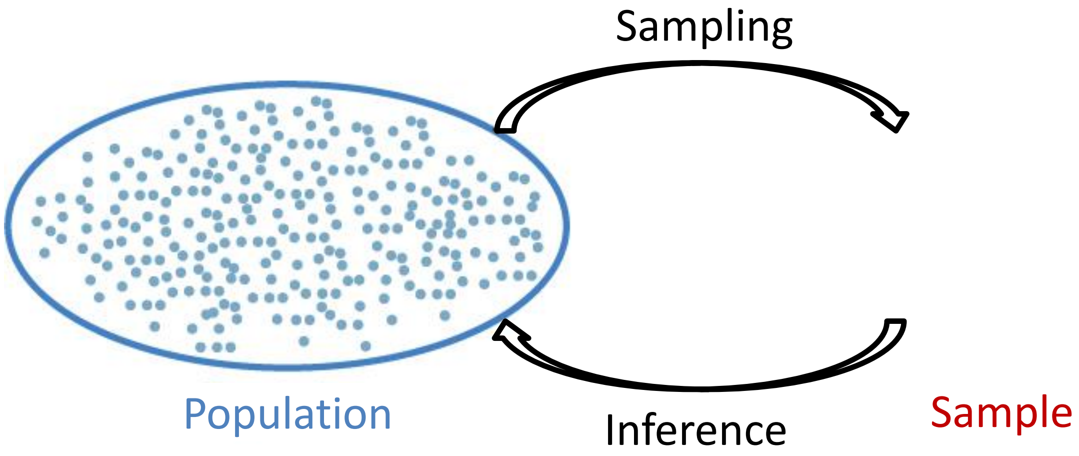
- Hypothesis testing = **inferential method**: extrapolation of conclusion drawn on the studied sample to the general population



Population

SAMPLE COLLECTION

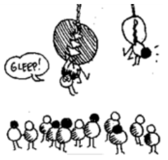
- Hypothesis testing = **inferential method**: extrapolation of conclusion drawn on the studied sample to the general population



⇒ Selection of the sample is essential

  - Sufficient size
  - Representative of population variability

- **Randomization**: simplest way to draw a random sample
  - Each individual of the population has an equal probability of selection



THEORY

- **Test statistic (Z)**: variable used to take a decision = reject H_0 or not
  - Example: comparison of asparagine concentration in the control group and in the vancomycin-treated group = **Student statistic**

$$z = \frac{m_1 - m_2}{\sqrt{s^2\left(1/n_1 + 1/n_2\right)}}$$

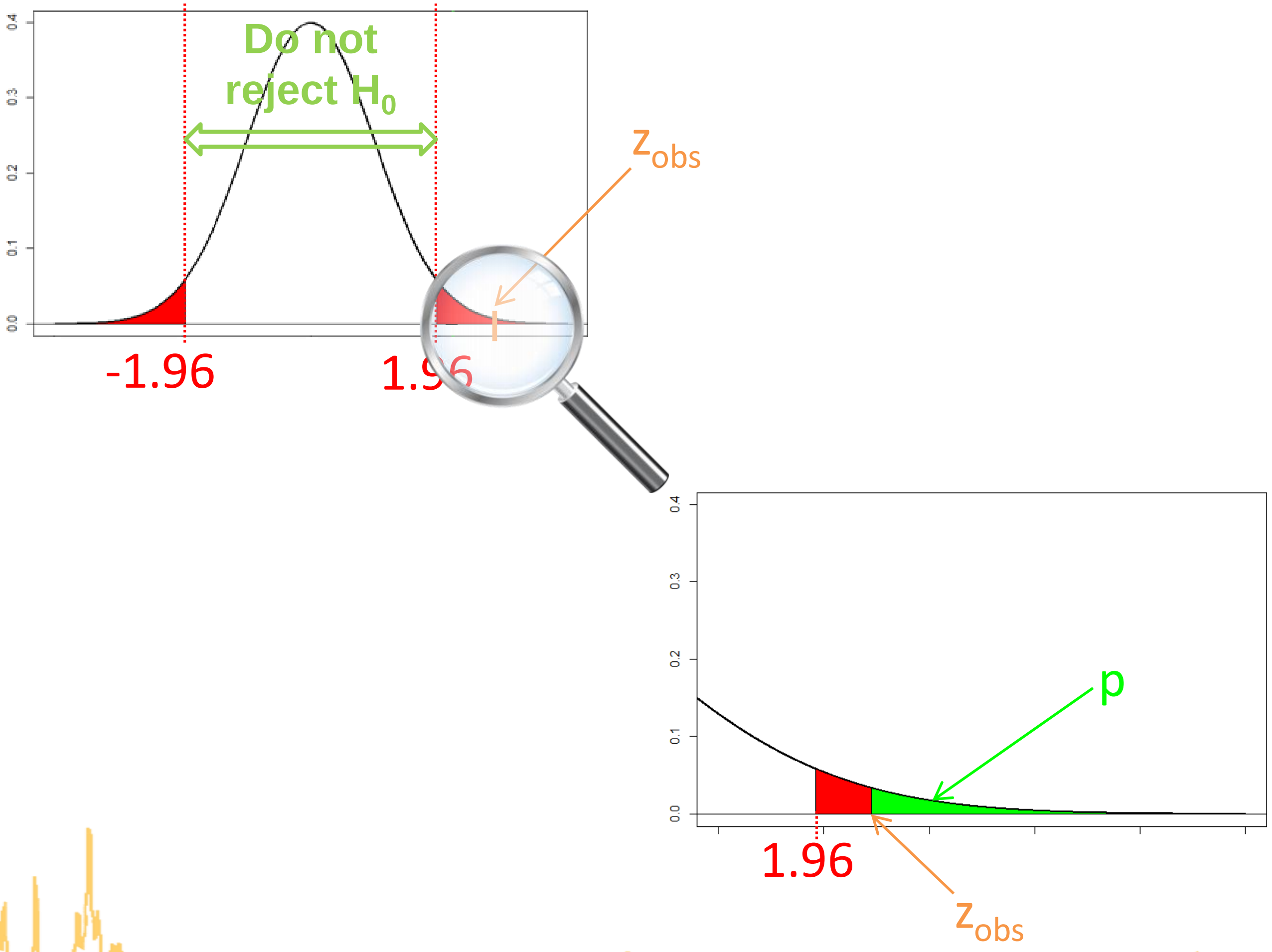
- If H_0 is true, probability distribution of Z is known

$\Rightarrow$ Comparison of the computed value to the expected value

  - Example: z_{obs}=5.70; z_{exp}=2.16

DECISION

- **p-value**: probability of observing, if H_0 is true, a value of test statistic as far from 0 as the value actually observed on the collected sample



DECISION

- **p-value**: probability of observing, if H_0 is true, a value of test statistic as far from 0 as the value actually observed on the collected sample
  - Quantifies the confidence you can have in your decision: the smaller the *p*-value, the more confident we can be in the conclusions drawn from it
  - ❗ $p \neq P[H_0\ true]$

- **Decision**: reject H_0 when p-value < α

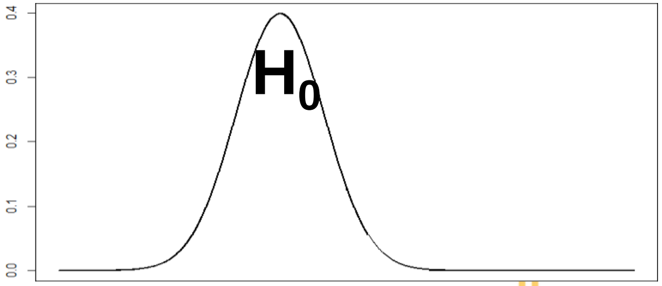
ERROR TYPES

		REALITY (unknown)	
		H_0 TRUE	H_0 FALSE
DECISION	NO REJECT	No error	
	REJECT		No error

ERROR TYPES

		REALITY (unknown)	
		H_0 TRUE	H_0 FALSE
DECISION	NO REJECT	No error	
	REJECT	α	No error

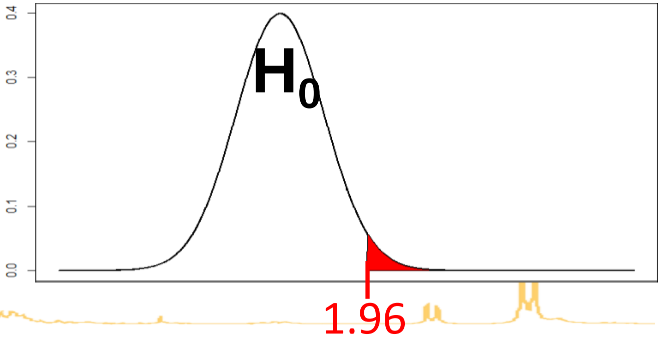
- Two error types:
  - *False positive error* (type I error): asserting a difference which doesn’t exist!
    - $\alpha = P[\text{Reject } H_0 \mid H_0 \text{ true}]$



ERROR TYPES

		REALITY (unknown)	
		H_0 TRUE	H_0 FALSE
DECISION	NO REJECT	No error	
	REJECT	α	No error

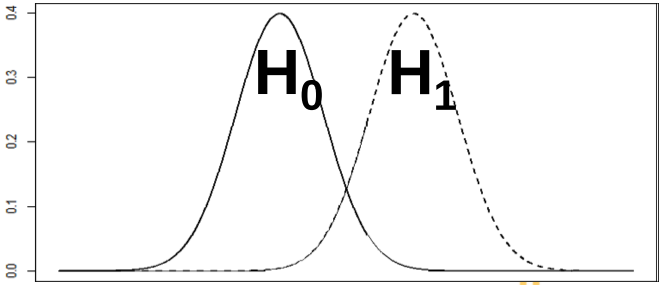
- Two error types:
  - *False positive error* (type I error): asserting a difference which doesn't exist!
    - α = P[Reject H_0 | H_0 true]
    - α**= significance threshold,** generally 0.05



ERROR TYPES

		REALITY (unknown)	
		H_0 TRUE	H_0 FALSE
DECISION	NO REJECT	No error	β
	REJECT	α	No error

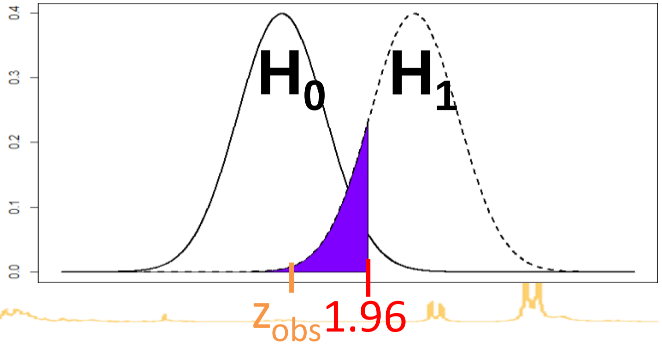
- Two error types:
  - False positive error (type I error): asserting a difference which doesn't exist!
    - α = P[Reject H_0 | H_0 true]
    - α= significance threshold, generally 5%
  - **False negative error** (type II error): failing to assert a difference which exists!
    - β = P[No reject H_0 | H_0 false]



ERROR TYPES

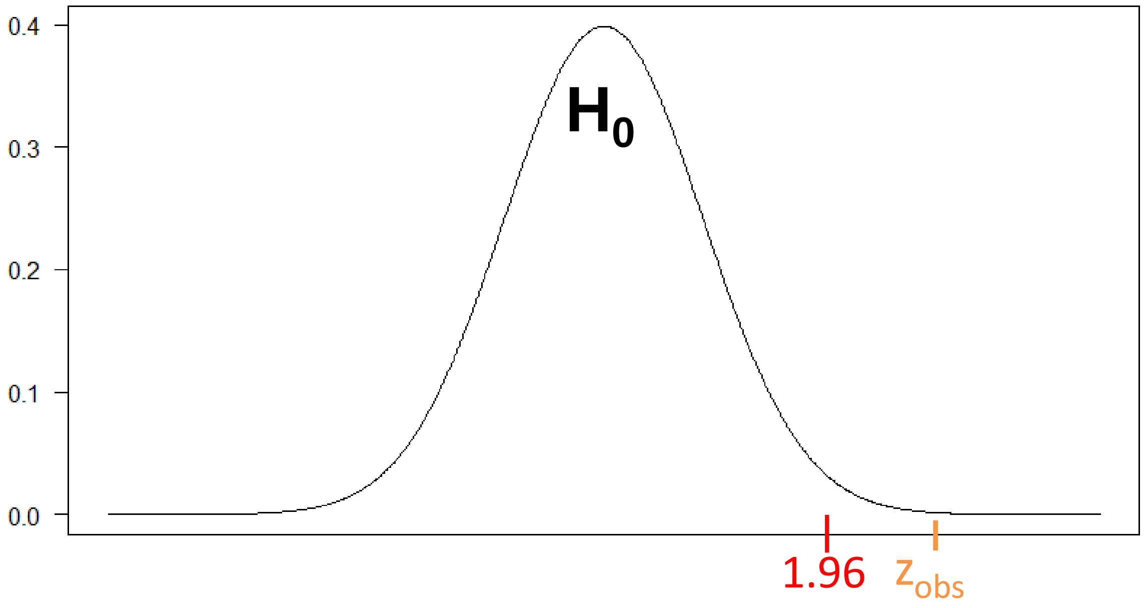
		REALITY (unknown)	
		H_0 TRUE	H_0 FALSE
DECISION	NO REJECT	No error	β
	REJECT	α	No error

- Two error types:
  - False positive error (type I error): asserting a difference which doesn't exist!
    - α = P[Reject H_0 | H_0 true]
    - α= significance threshold, generally 5%
  - **False negative error** (type II error): failing to assert a difference which exists!
    - β = P[No reject H_0 | H_0 false]
    - **Power** lack



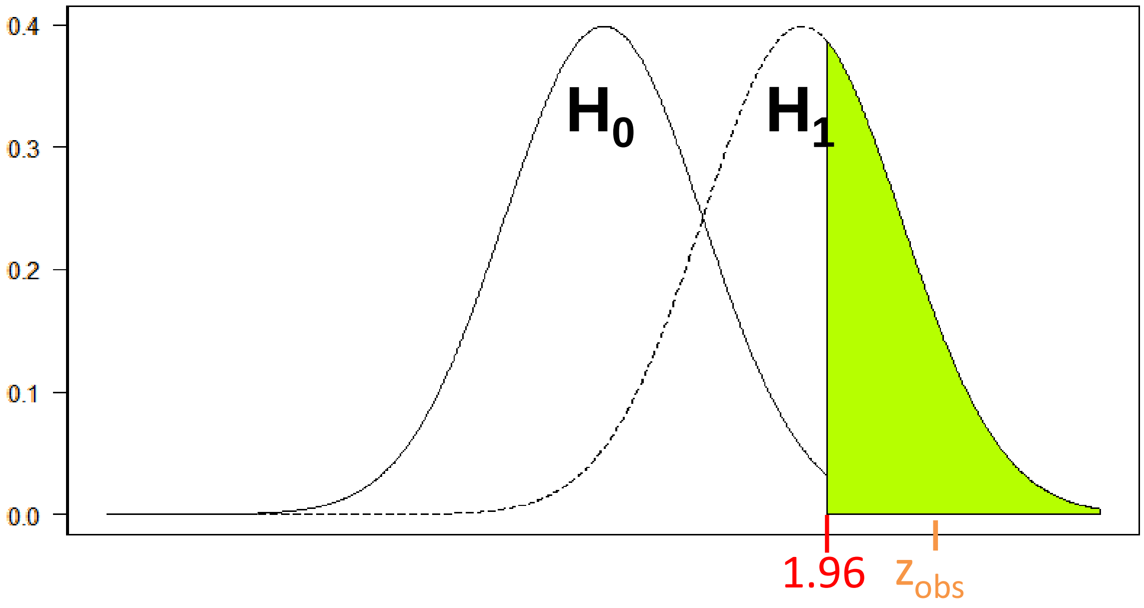
POWER

- Probability to reject H_0



POWER

- Probability to reject H_0 when H_0 is false (true positive)
- $\pi = P[\text{Reject } H_0 \mid H_0 \text{ false}]$
- Depends on the sample size: $n \uparrow \Rightarrow \pi \uparrow \Rightarrow \beta \downarrow$



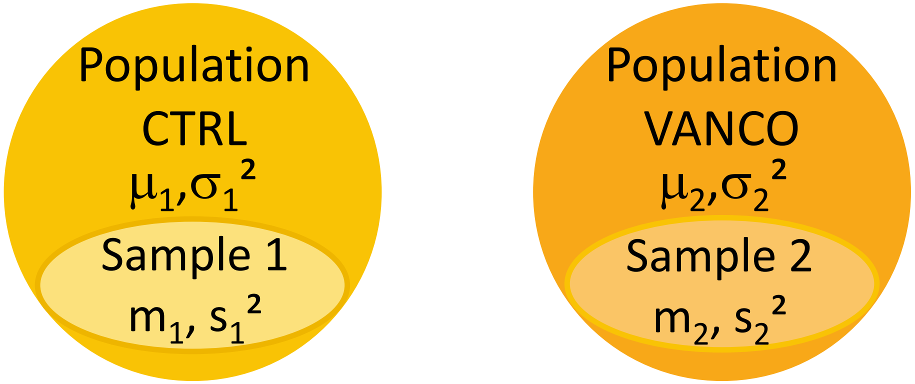
TESTS

- **Comparison test**: comparison of a parameter of interest (mean, proportion, …) computed on one, two or more samples
  - Ex: Is the ciproflaxin-treated group different from the vancomycin treated group?

- **Independence (or association) test** between variables: are observations on two variables independent of each other?
  - Ex: does the asparagine concentration depend on the lactic acid concentration?

STUDENT TEST (1)

- Comparison of the mean value of a variable of interest measured in **two populations**
  - Ex: concentration of Asparagine



Population CTRL μ_1,σ_1^2

Sample 1 m_1, s_1^2

Population VANCO μ_2,σ_2^2

Sample 2 m_2, s_2^2

- Null Hypothesis H_0 : $\mu_{CTRL} = \mu_{VANCO}$
- Test statistic
  - $z = \frac{m_1 - m_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, with $s^2 = \frac{(n_1 - 1)s_1^2 - (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
- **Decision**: Reject H_0 if p-value < 0.05

STUDENT TEST (2)

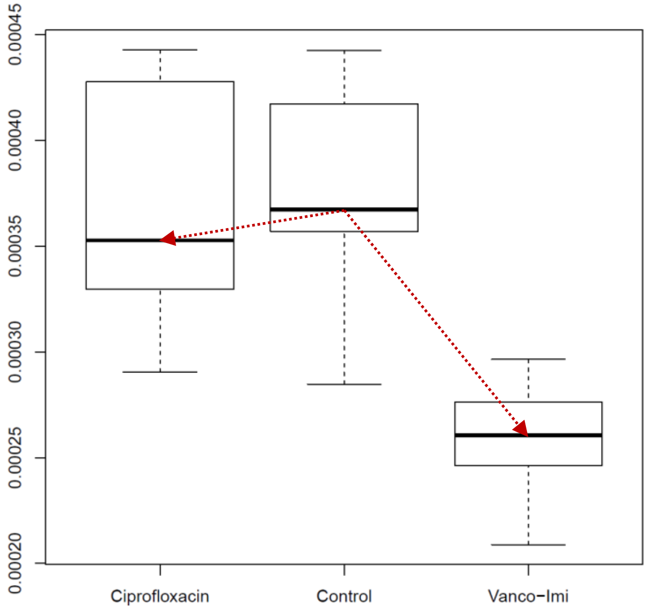
- Ex: concentration of Asparagine in Control and vancomycin-treated populations

variable	ppm	Treatment_ttest_Vanco-Imi.Control_dif	Treatment_ttest_Vanco-Imi.Control_fdr
B2.95	2.95	-0.000117571862586912	0.000526995159001272

⇒ p-value < 0.05: H_0 can be rejected = concentration of Asparagine in vancomycin-treated population is not equal to concentration in control population

ONE WAY ANALYSIS OF VARIANCE (1)

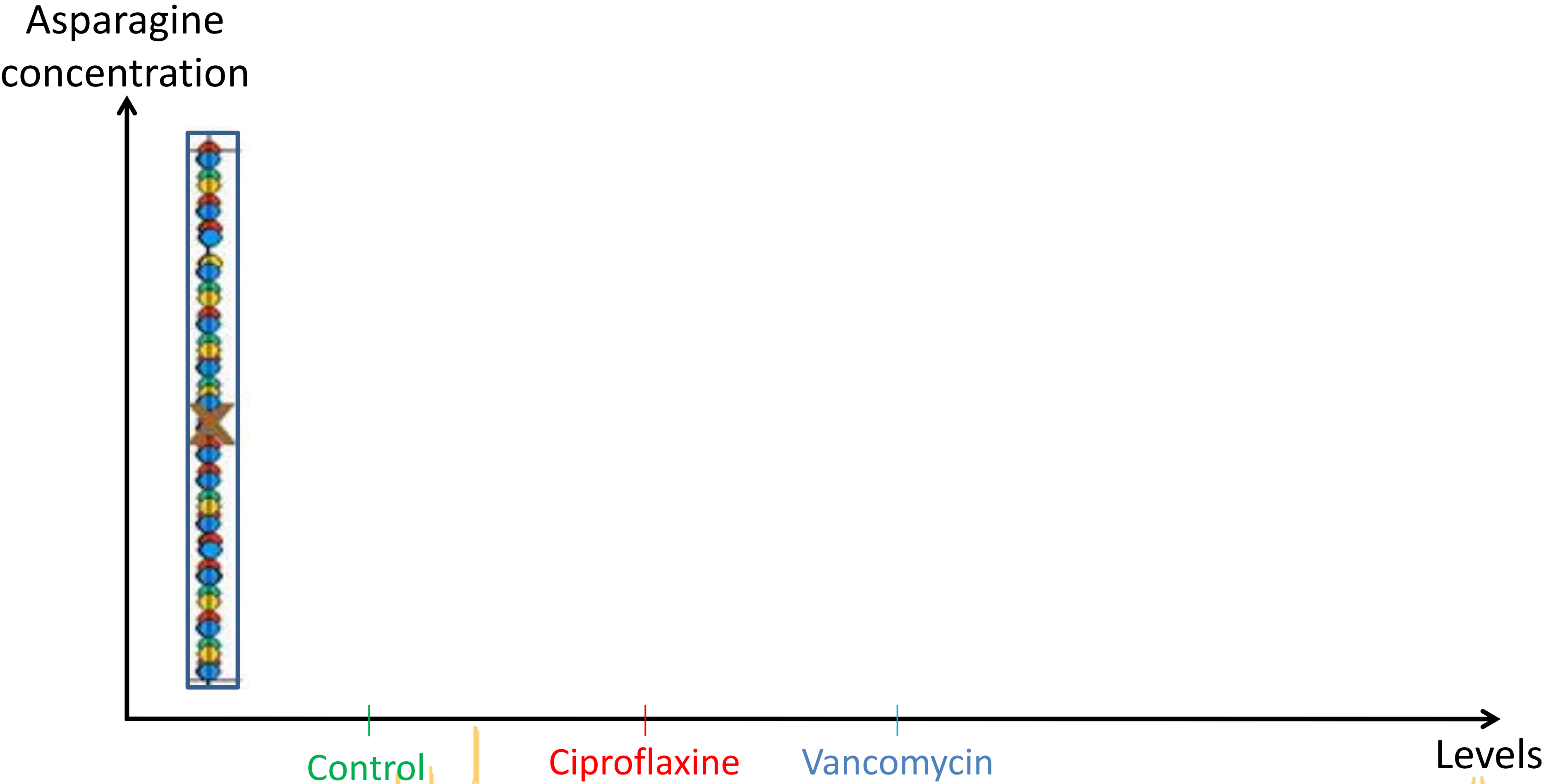
- Comparison of **means of 3 or more populations**
  - Ex: Is Asparagine concentration identical in control population, ciprofloxacin-treated population and vancomycin-treated population?



- Student test generalization
- Hypothesis
  - H_0 : $\mu_{CTRL} = \mu_{CIPRO} = \mu_{VANCO}$
  - H_1 : at least 2 group means are different

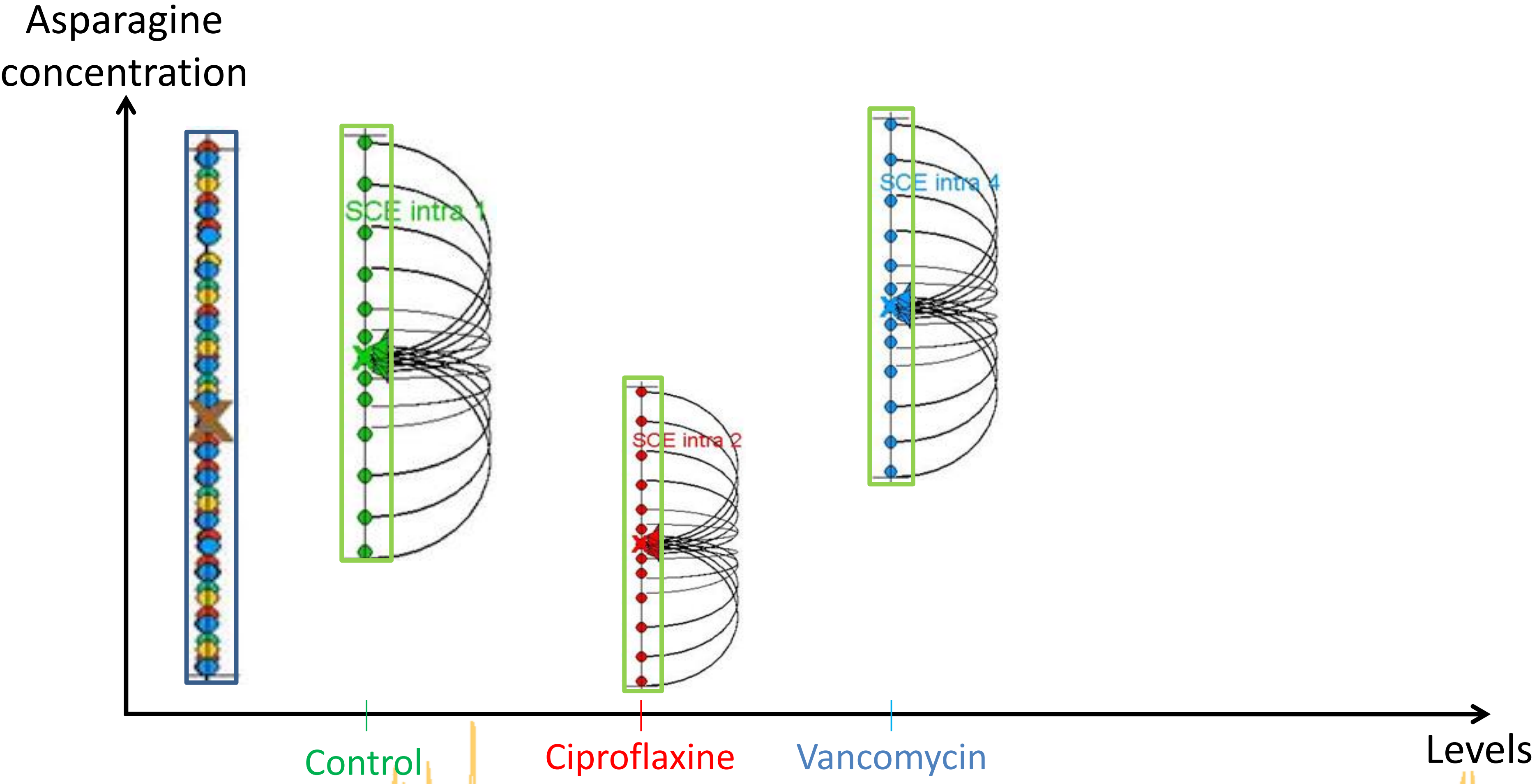
ONE-WAY ANALYSIS OF VARIANCE (2)

- Decomposition of total variability (SST)



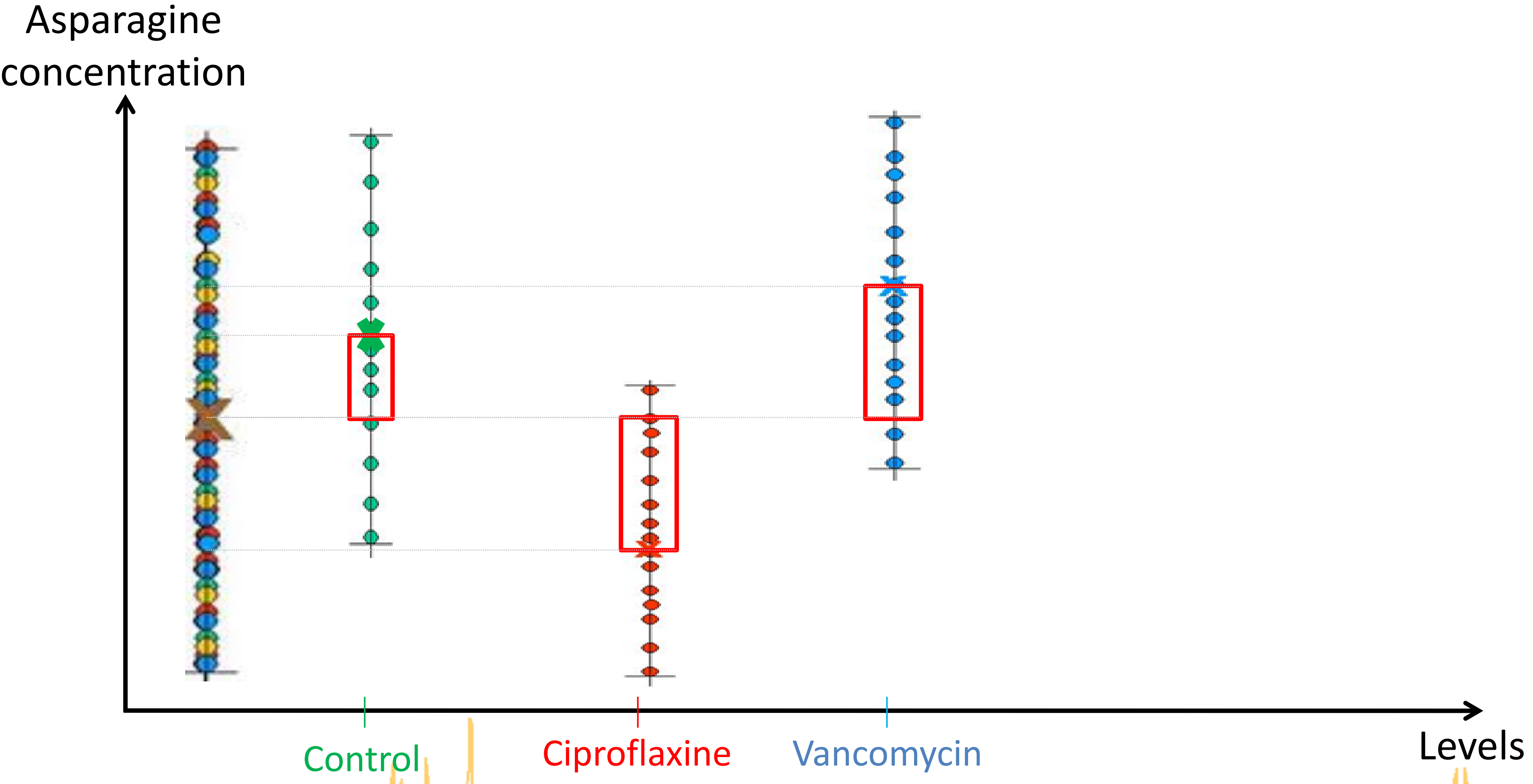
ONE-WAY ANALYSIS OF VARIANCE (3)

- Decomposition of total variability (SST)
  - SST = **Within group variability** (SSW)



ONE-WAY ANALYSIS OF VARIANCE (4)

- Decomposition of total variability (SST)
  - SST = **Within group variability** (SSW) + **Between group variability** (SSB)

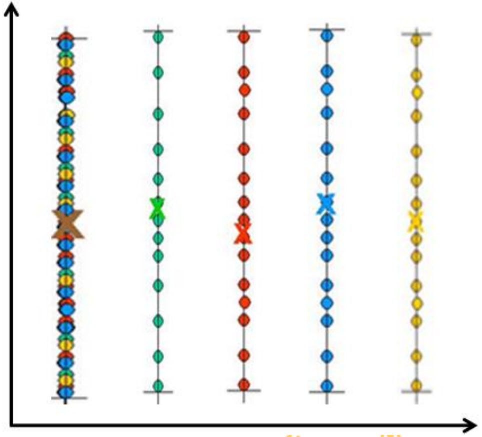


ONE-WAY ANALYSIS OF VARIANCE (5)

- ANOVA Table

Variability	Sum of Squares	Degrees of freedom	Mean Squares	Fisher statistic test	P-Value
Between	SSB	G-1	SSB / (G-1)	$F = \frac{SSB}{SSW} * \frac{n-G}{G-1}$	**p**
Within	SSW	n-G	SSW / (N-G)		
Total	SST	n-1			

- **Decision**: H_0 is rejected if $p < 0.05$
  - H_0 rejected = total variability >> within variability
  - H_0 no rejected = total variability ≈ within variability



ONE-WAY ANALYSIS OF VARIANCE (6)

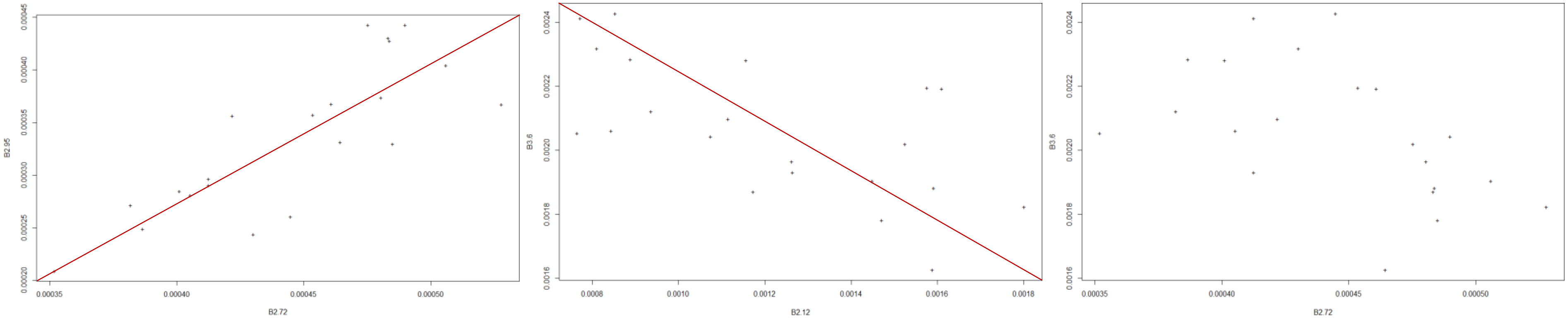
- Ex: comparison of Asparagine concentration in control population, ciprofloxacin-treated population and vancomycin-treated population

- H_0 rejected ⇔ at least, 2 means are different

  ⇒ pairwise comparisons

PEARSON TEST (1)

- Used to test **dependence** between two variables *X* and *Y*



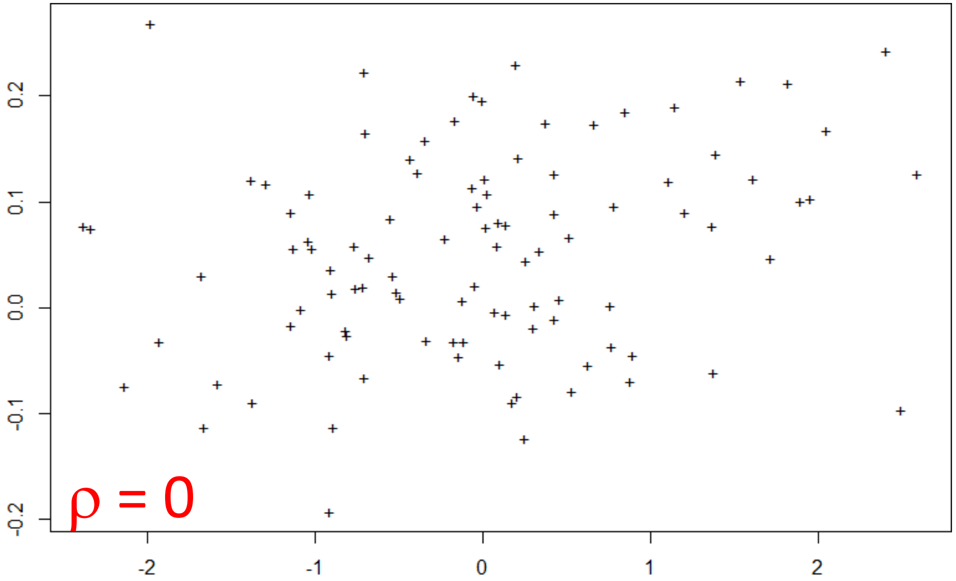
- Null Hypothesis H_0 : $\rho = 0$ (independence)
  - Ex: are concentrations of buckets 2.95 and 2.12 ppm correlated?
- Test statistic

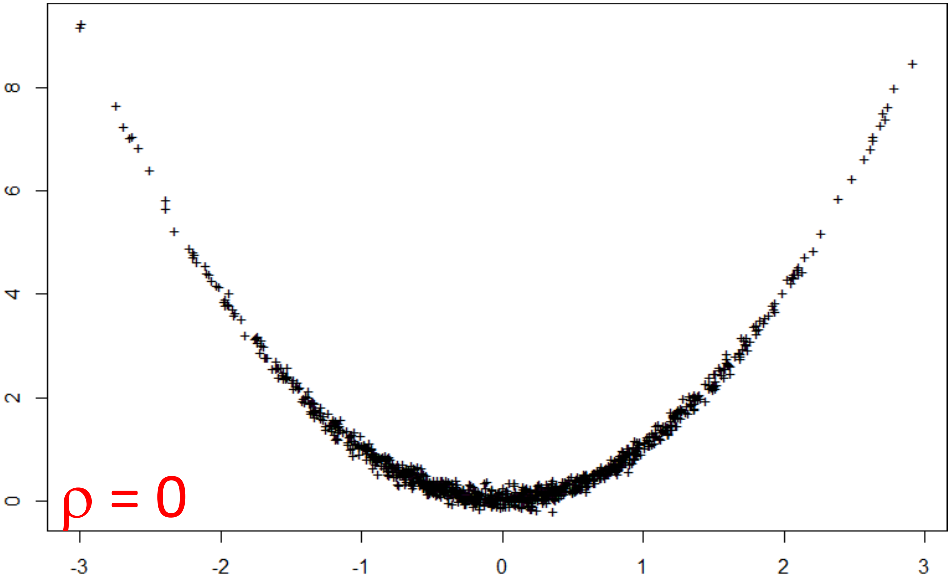
$$Z = \frac{r}{\sqrt{(1-r^2)/(n-2)}}, \text{ with } r = \frac{\sum_{i=1}^{n} (x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y}$$

- **Decision**: Reject H_0 if p-value < 0.05

PEARSON TEST (2)

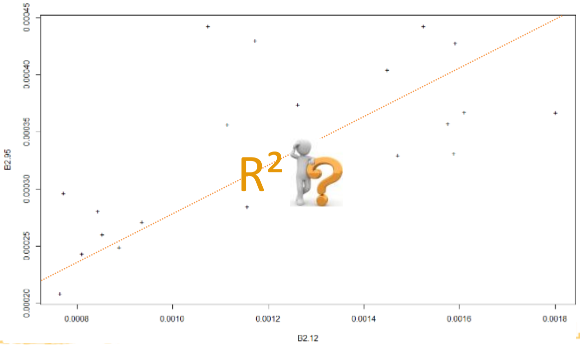
- Pearson correlation = **linear** relationship





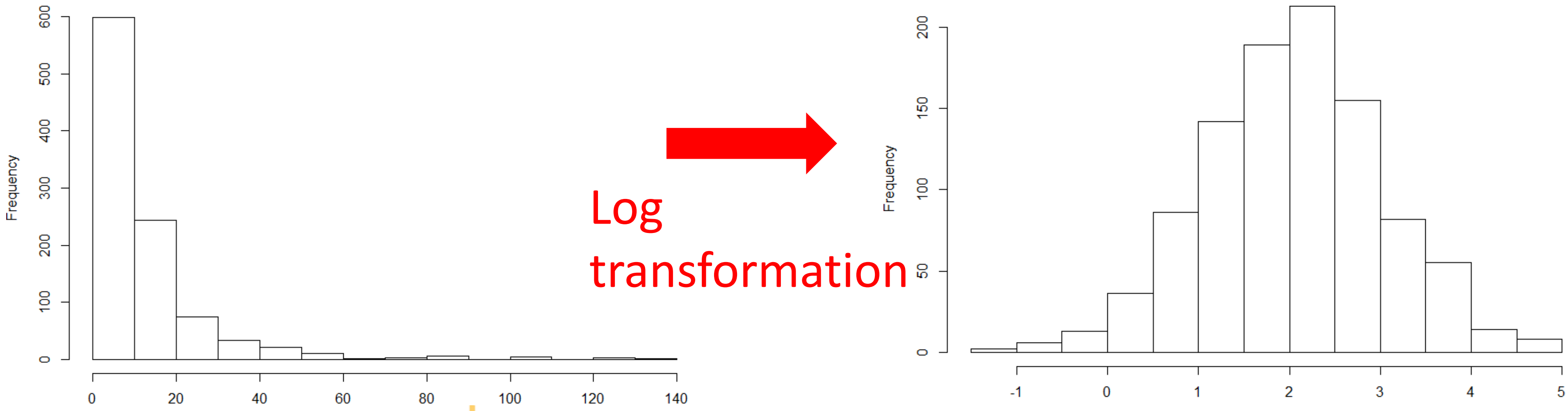
⇒ Does not measure non linear relationship

- Ex: are intensities of buckets 2.95 and 2.12 ppm correlated?



WHEN ASSUMPTIONS ARE NOT SATISFIED…

- Previous tests = parametric tests
  - Normal distribution
  - Homogeneous variance
- When assumptions are not satisfied?
  - Data transformation for normality



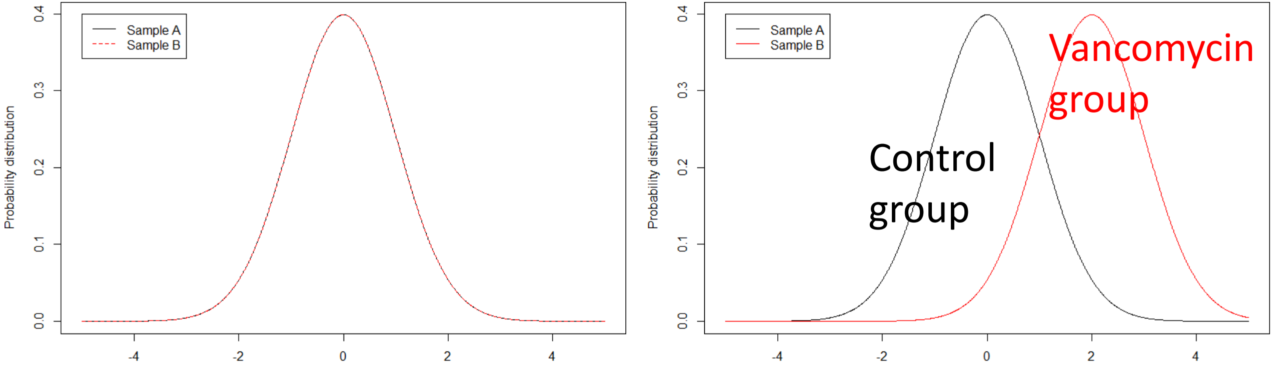
Log transformation

WHEN ASSUMPTIONS ARE NOT SATISFIED…

- Previous tests = parametric tests
  - Normal distribution
  - Homogeneous variance
- When assumptions are not satisfied?
  - If any transformation works, parametric tests are unusable
- Non parametric tests
  - No assumption about probability distribution, variance, …
  - Based on observation ranks

WILCOXON TEST (1)

- Used to compare **probability distribution** of a quantitative variable observed on two samples
- Hypothesis
  - H_0 : the 2 samples come from the same population
  - H_1 : the 2 samples are not from the same population (one sample tends to have larger values) = at least, sample medians are not equal



- Test statistic

$$U_{n_1} = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_{n_1}$$

- **Decision**: Reject H_0 if p-value < 0.05

WILCOXON TEST (2)

- Ex: concentration of Asparagine in Control and vancomycin-treated populations

- Reminder:

Treatment_**ttest**_Vanco-Imi.Control_fdr = 0.00052

⇒ Non parametric tests: less powerful, but more robust

KRUSKAL-WALLIS TEST

- Used to compare **probability distribution** of a quantitative variable observed on three and more samples
- Null hypothesis
  - H_0: the samples come from the same population
  - H_1: sample medians are not equal for 2 samples at least
- Test statistic

$$S = \frac{12}{N(N+1)} \sum_{g=1}^{G} \left[\frac{R_g^2}{n_g} - 3*(N+1) \right]$$

- **Decision**: Reject H_0 if p-value < 0.05
- Ex: comparison of Asparagine concentration in control population, ciprofloxacin-treated population and vancomycin-treated population

SPEARMAN TEST

- Used to test **dependence** between two variables (monotonic relationship)
- Null Hypothesis H_0 : $\rho = 0$ (independence)
- Test statistic

$$Z = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

- Ex: are concentrations of buckets 2.95 and 2.12 ppm correlated?

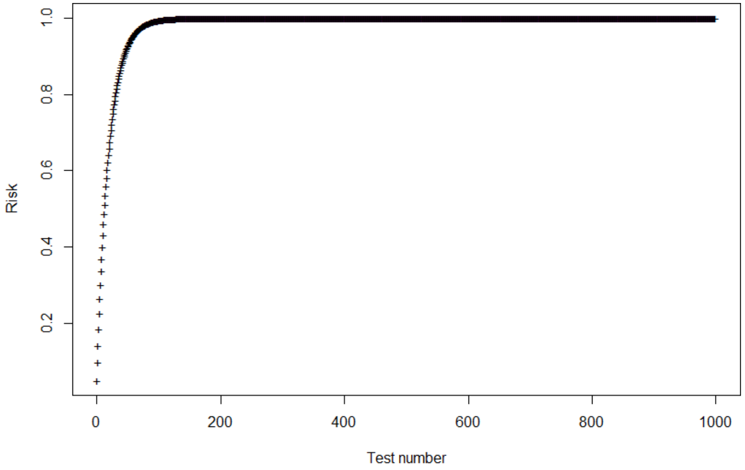
MULTIPLE TESTING (1)

- Univariate test ⇒ each feature is individually tested
- Metabolomic dataset: hundreds or thousands of features

⇒ hundreds tests $H^j_0: \mu_1 = \mu_2$ vs $H^j_1: \mu_1 \neq \mu_2$ simultaneously made

⇒ Multiple testing problem

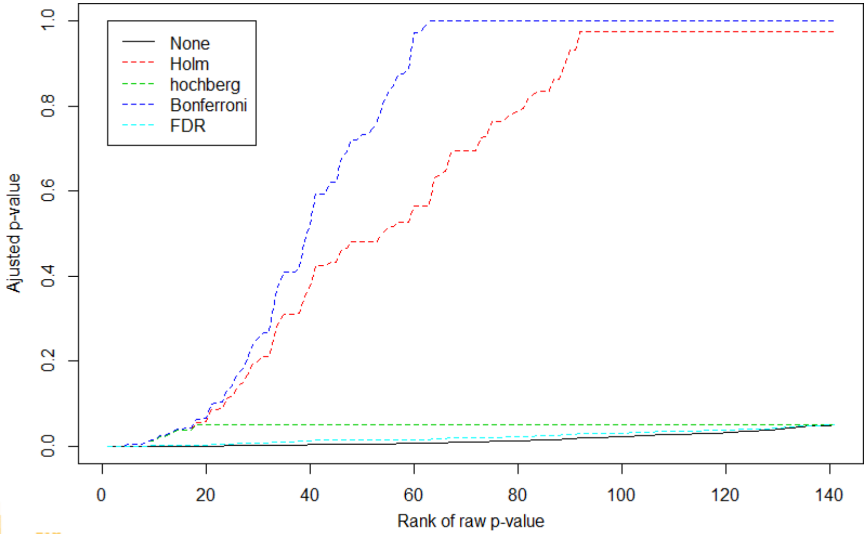
  - ↑ α risk: the probability of getting a significant result simply due to chance keeps going up (false positive)



⇒ $\alpha = P[\text{Reject } H_0 \mid H_0 \text{ true}]$ ↑ with the number of simultaneous univariate tests

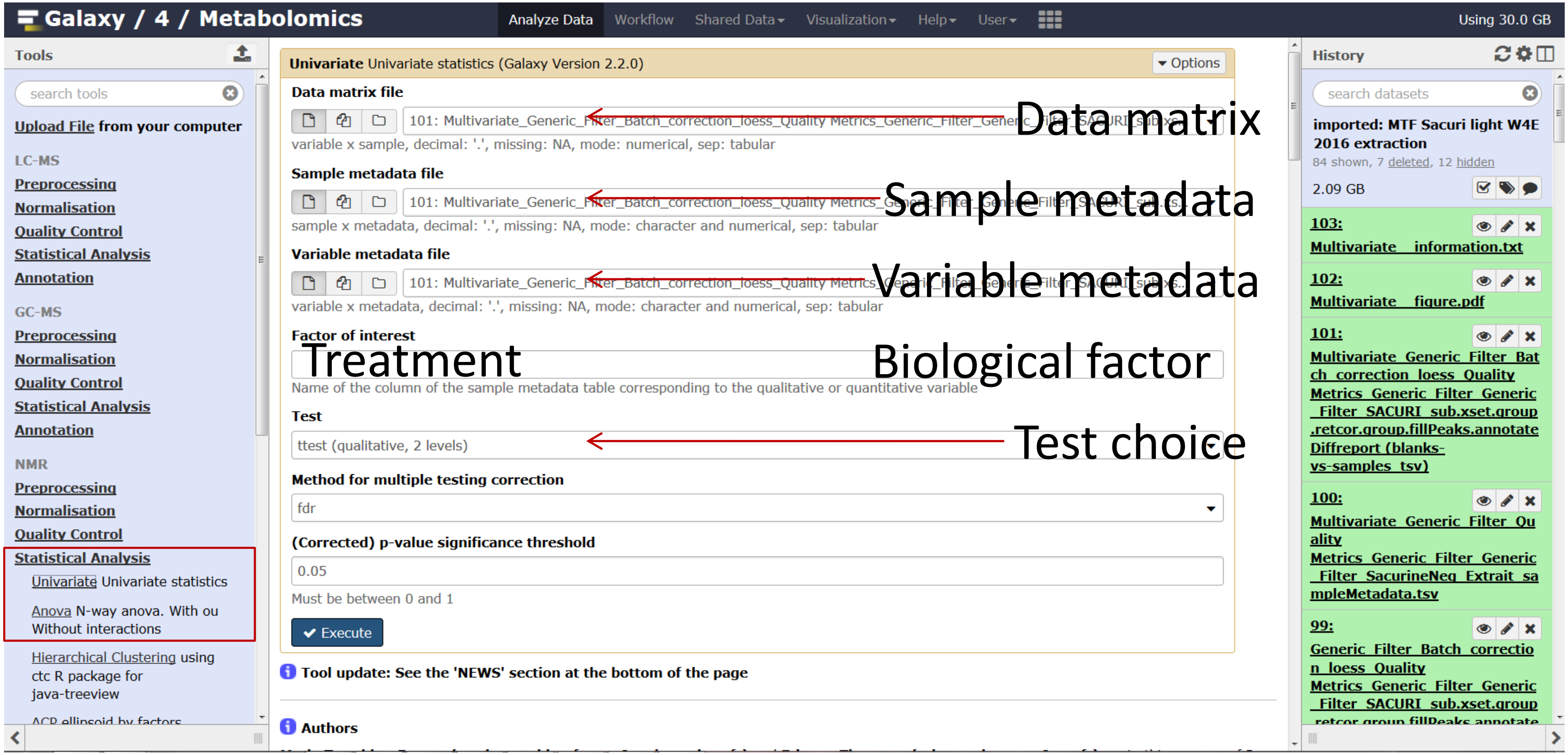
MULTIPLE TESTING (2)

- Several correction methods proposed: p-value correction depends on the number of comparisons (the probability of observing at least one significant result due to chance remains below your desired significance level)
  - Bonferroni: divide α by the number of performed tests
    - Ex: p=100 variables, α'=0.0005
  - **False Discovery Rate**



HOW TO DO WITH GALAXY?

GALAXY FORM (1)



Univariate Univariate statistics (Galaxy Version 2.2.0)

Data matrix file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.x

variable x sample, decimal: '.', missing: NA, mode: numerical, sep: tabular

Sample metadata file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.s...

sample x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

Variable metadata file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.x...

variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

Factor of interest

Treatment

Name of the column of the sample metadata table corresponding to the qualitative or quantitative variable

Test

ttest (qualitative, 2 levels)

Method for multiple testing correction

fdr

(Corrected) p-value significance threshold

0.05

Must be between 0 and 1

Tool update: See the 'NEWS' section at the bottom of the page

imported: MTF Sacuri light W4E 2016 extraction

84 shown, 7 deleted, 12 hidden

2.09 GB

103: Multivariate_information.txt

102: Multivariate_figure.pdf

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xset.group.retcor.group.fillPeaks.annotate Diffreport (blanks-vs-samples_tsv)

100: Multivariate_Generic_Filter_Quality Metrics_Generic_Filter_Generic_Filter_SacurineNeg_Extrait_sampleMetadata.tsv

99: Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xset.group.retcor.group.fillPeaks.annotate

Data matrix

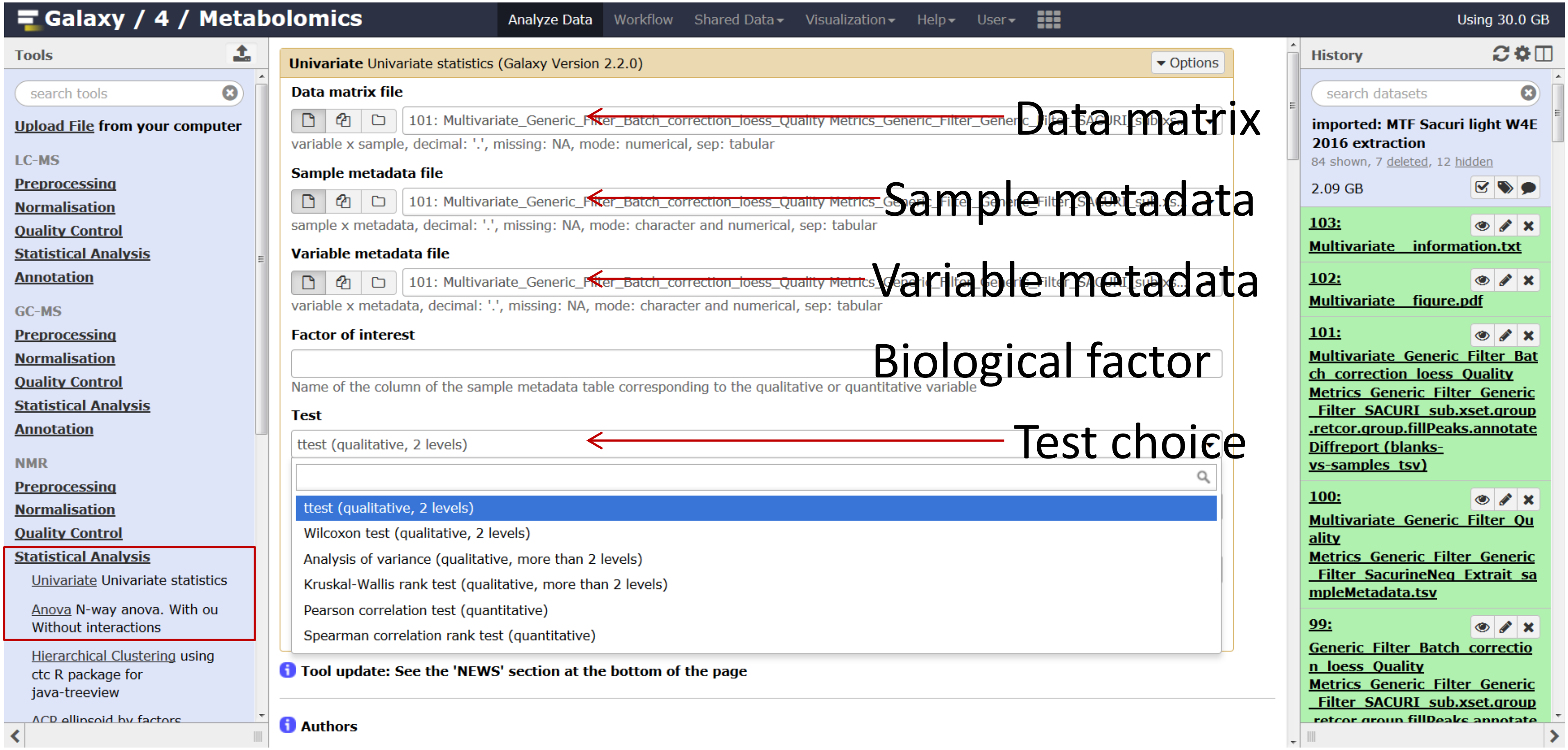
Sample metadata

Variable metadata

Biological factor

Test choice

GALAXY FORM (2)



Univariate Univariate statistics (Galaxy Version 2.2.0)

Data matrix file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xs...

variable x sample, decimal: '.', missing: NA, mode: numerical, sep: tabular

Sample metadata file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xs...

sample x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

Variable metadata file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xs...

variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

Factor of interest

Name of the column of the sample metadata table corresponding to the qualitative or quantitative variable

Test

ttest (qualitative, 2 levels)

- ttest (qualitative, 2 levels)
- Wilcoxon test (qualitative, 2 levels)
- Analysis of variance (qualitative, more than 2 levels)
- Kruskal-Wallis rank test (qualitative, more than 2 levels)
- Pearson correlation test (quantitative)
- Spearman correlation rank test (quantitative)

Tool update: See the 'NEWS' section at the bottom of the page

Authors

imported: MTF Sacuri light W4E 2016 extraction

84 shown, 7 deleted, 12 hidden

2.09 GB

103: Multivariate_information.txt

102: Multivariate_figure.pdf

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xset.group.retcor.group.fillPeaks.annotate Diffreport (blanks-vs-samples_tsv)

100: Multivariate_Generic_Filter_Quality Metrics_Generic_Filter_Generic_Filter_SacurineNeg_Extrait_sampleMetadata.tsv

99: Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xset.group.retcor.group.fillPeaks.annotate

Data matrix

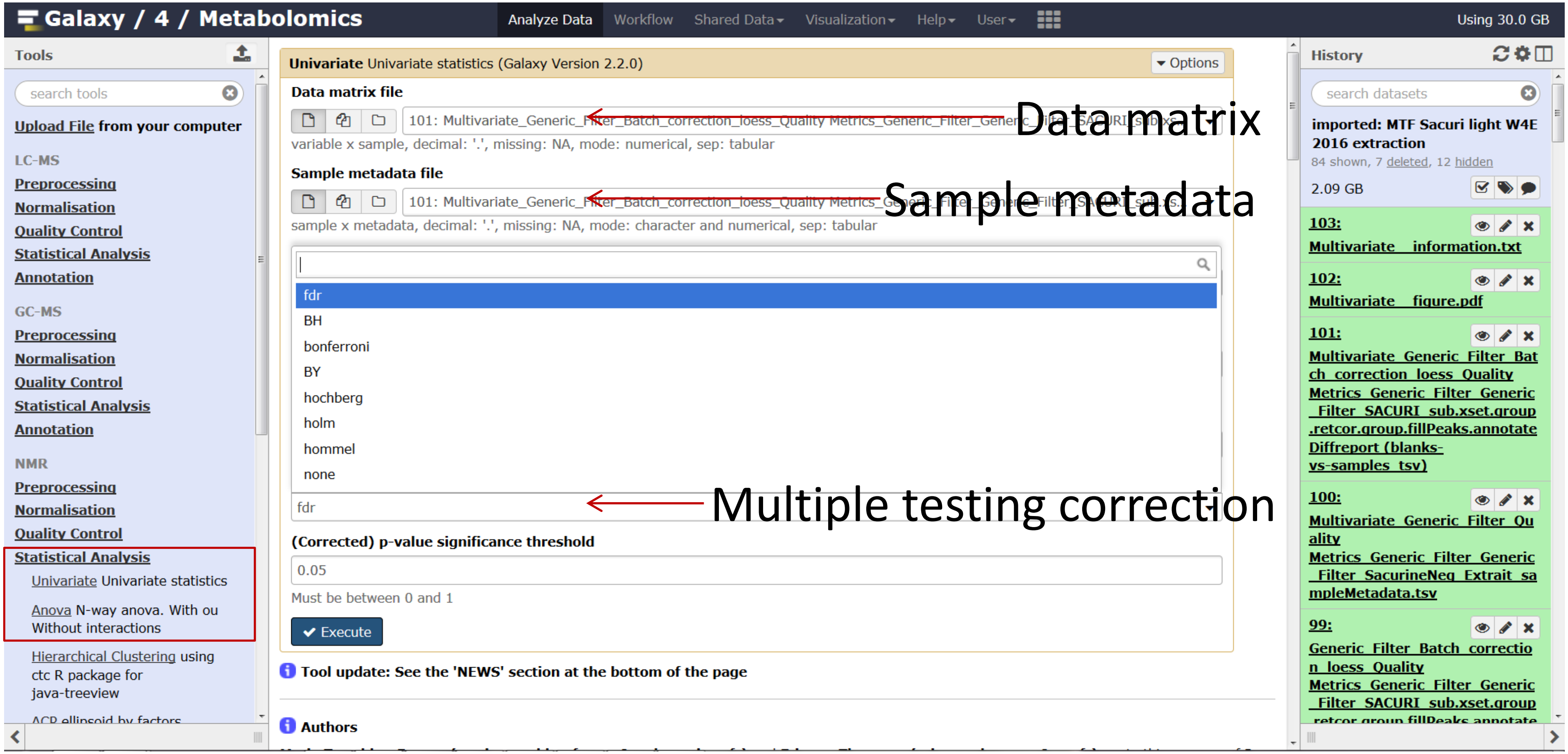
Sample metadata

Variable metadata

Biological factor

Test choice

GALAXY FORM (3)



Univariate Univariate statistics (Galaxy Version 2.2.0)

Data matrix file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xs...

variable x sample, decimal: '.', missing: NA, mode: numerical, sep: tabular

Sample metadata file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xs...

sample x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

- fdr
- BH
- bonferroni
- BY
- hochberg
- holm
- hommel
- none

fdr

(Corrected) p-value significance threshold

0.05

Must be between 0 and 1

Tool update: See the 'NEWS' section at the bottom of the page

Authors

imported: MTF Sacuri light W4E 2016 extraction

84 shown, 7 deleted, 12 hidden

2.09 GB

103: Multivariate_information.txt

102: Multivariate_figure.pdf

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xset.group.retcor.group.fillPeaks.annotate Diffreport (blanks-vs-samples_tsv)

100: Multivariate_Generic_Filter_Quality Metrics_Generic_Filter_Generic_Filter_SacurineNeg_Extrait_sampleMetadata.tsv

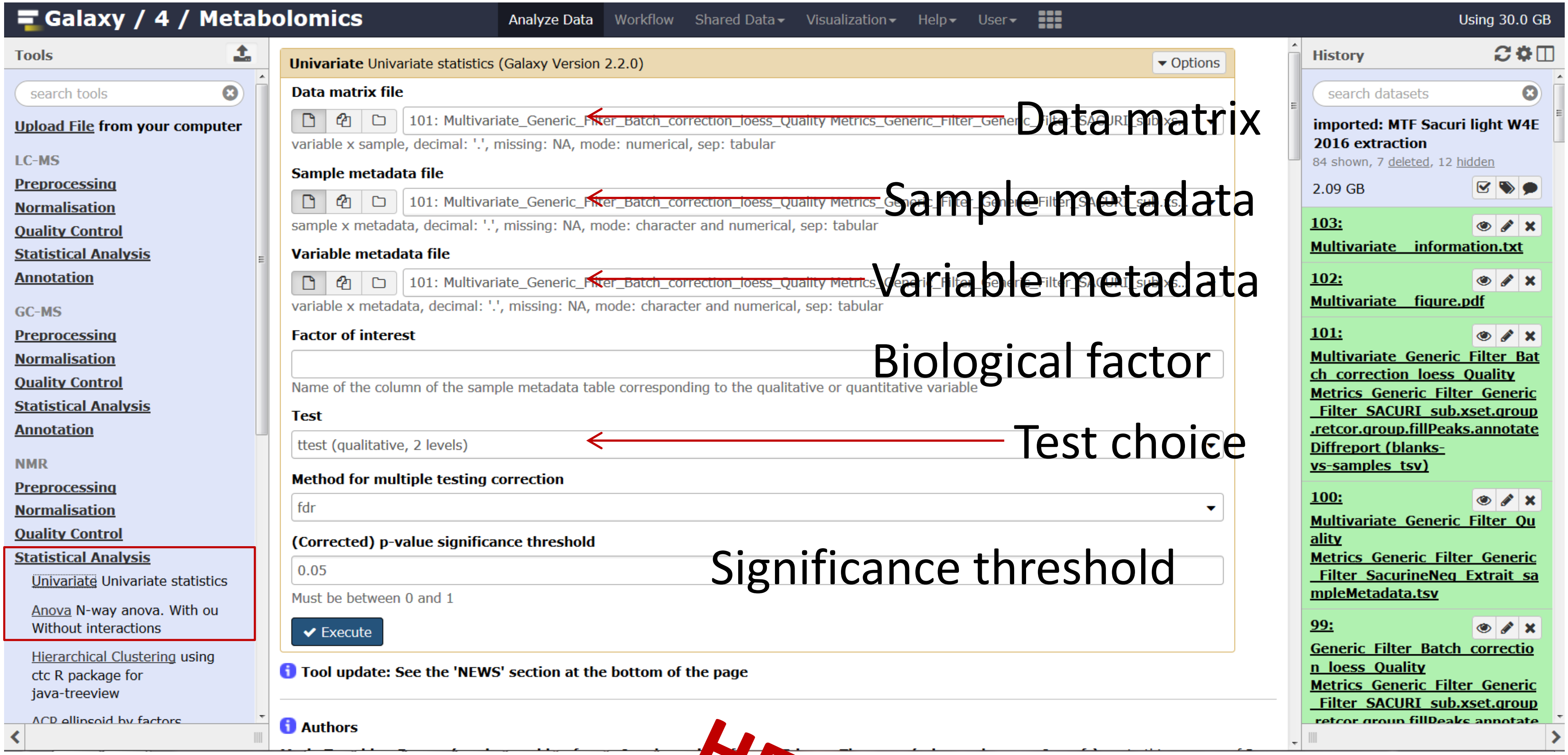
99: Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xset.group.retcor.group.fillPeaks.annotate

Data matrix

Sample metadata

Multiple testing correction

GALAXY FORM (4)



Univariate Univariate statistics (Galaxy Version 2.2.0)

Data matrix file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.x

variable x sample, decimal: '.', missing: NA, mode: numerical, sep: tabular

Sample metadata file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xs...

sample x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

Variable metadata file

101: Multivariate_Generic_Filter_Batch_correction_loess_Quality Metrics_Generic_Filter_Generic_Filter_SACURI_sub.xs...

variable x metadata, decimal: '.', missing: NA, mode: character and numerical, sep: tabular

Factor of interest

Name of the column of the sample metadata table corresponding to the qualitative or quantitative variable

Test

ttest (qualitative, 2 levels)

Method for multiple testing correction

fdr

(Corrected) p-value significance threshold

0.05

Must be between 0 and 1

Tool update: See the 'NEWS' section at the bottom of the page

Authors

Data matrix

Sample metadata

Variable metadata

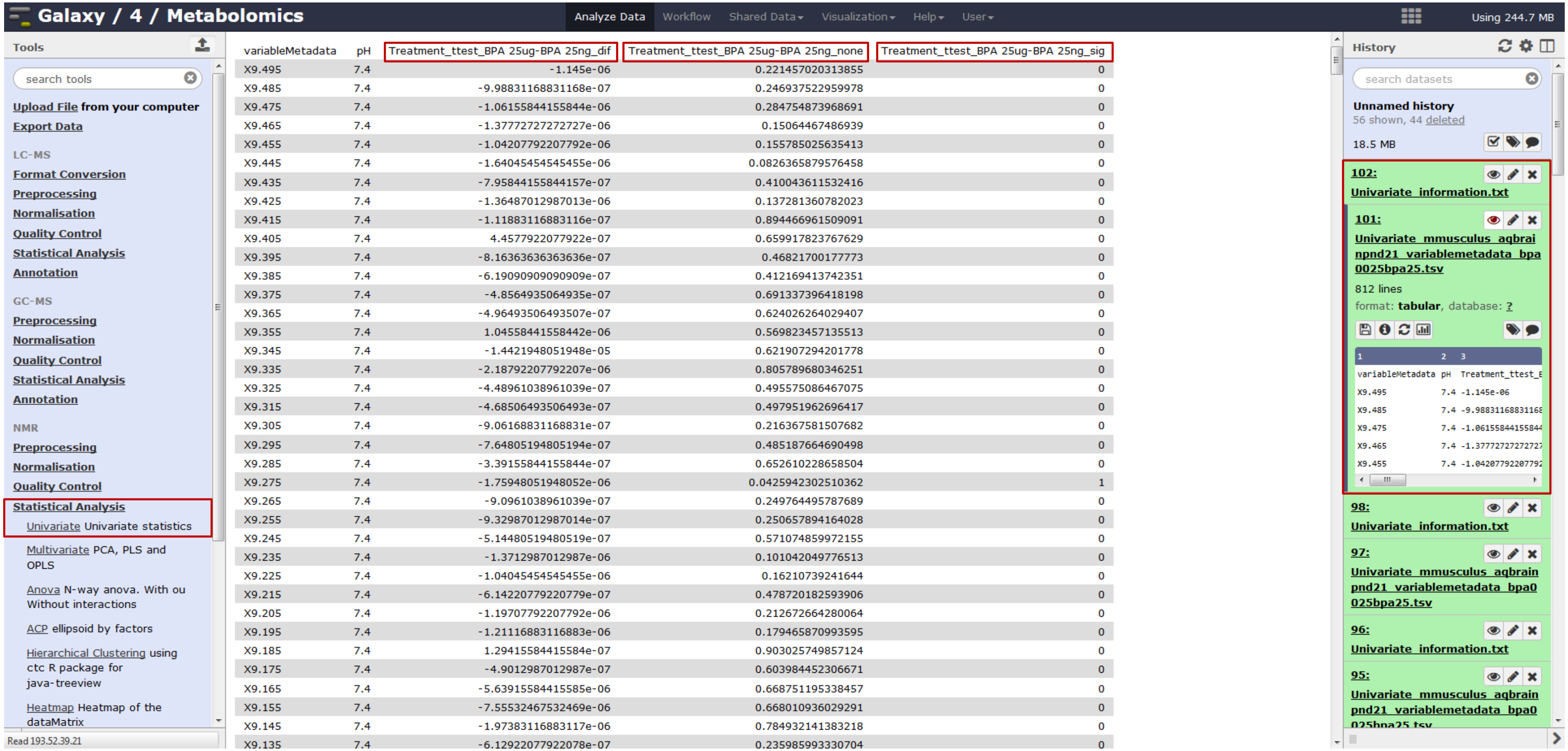
Biological factor

Test choice

Significance threshold

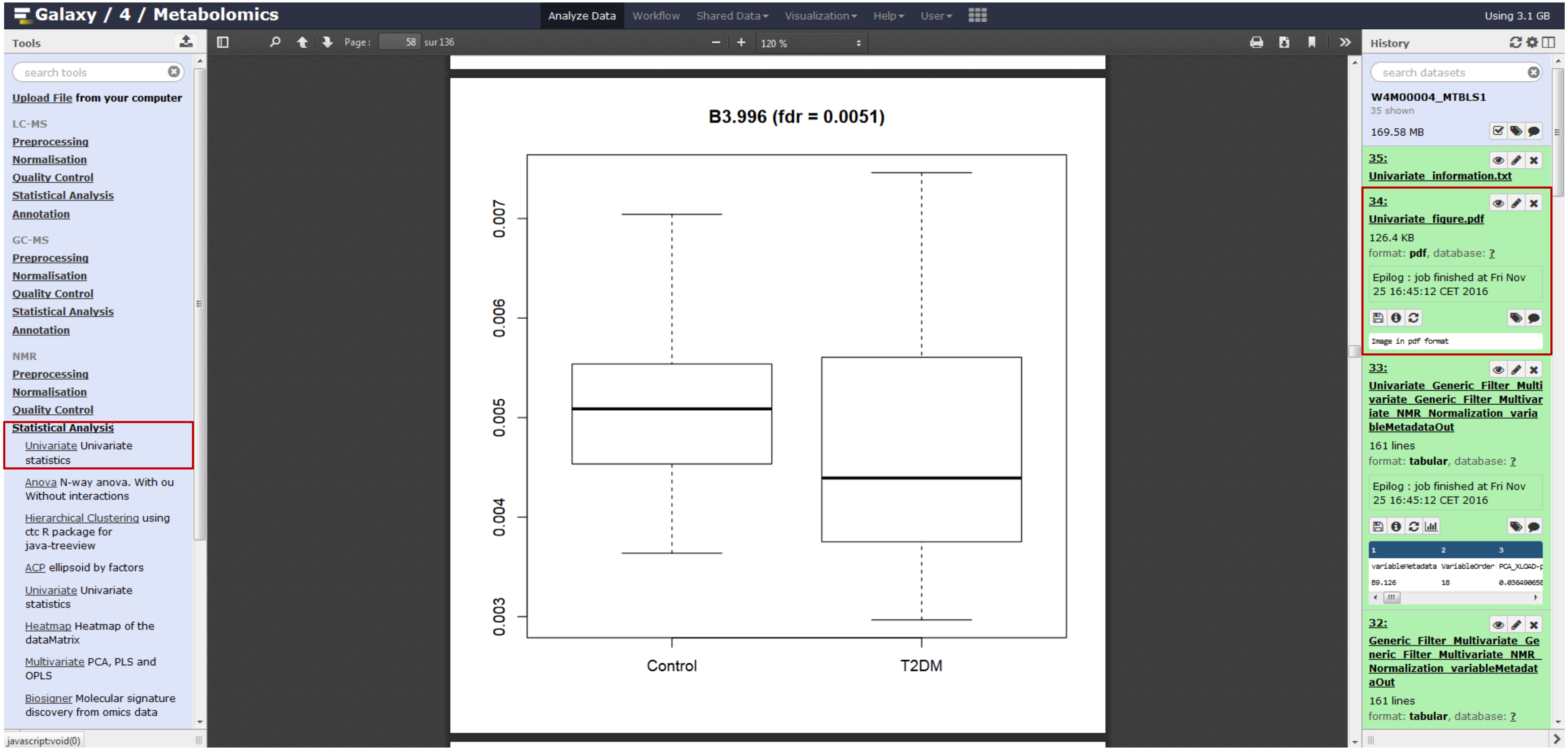
HELP

GALAXY RESULT: VARIABLE METADATA



variableMetadata	pH	Treatment_ttest_BPA 25ug-BPA 25ng_dif	Treatment_ttest_BPA 25ug-BPA 25ng_none	Treatment_ttest_BPA 25ug-BPA 25ng_sig
X9.495	7.4	-1.145e-06	0.221457020313855	0
X9.485	7.4	-9.98831168831168e-07	0.246937522959978	0
X9.475	7.4	-1.06155844155844e-06	0.284754873968691	0
X9.465	7.4	-1.37772727272727e-06	0.15064467486939	0
X9.455	7.4	-1.04207792207792e-06	0.155785025635413	0
X9.445	7.4	-1.64045454545455e-06	0.0826365879576458	0
X9.435	7.4	-7.95844155844157e-07	0.410043611532416	0
X9.425	7.4	-1.36487012987013e-06	0.137281360782023	0
X9.415	7.4	-1.11883116883116e-07	0.894466961509091	0
X9.405	7.4	4.4577922077922e-07	0.659917823767629	0
X9.395	7.4	-8.16363636363636e-07	0.46821700177773	0
X9.385	7.4	-6.19090909090909e-07	0.412169413742351	0
X9.375	7.4	-4.8564935064935e-07	0.691337396418198	0
X9.365	7.4	-4.96493506493507e-07	0.624026264029407	0
X9.355	7.4	1.04558441558442e-06	0.569823457135513	0
X9.345	7.4	-1.4421948051948e-05	0.621907294201778	0
X9.335	7.4	-2.18792207792207e-06	0.805789680346251	0
X9.325	7.4	-4.48961038961039e-07	0.495575086467075	0
X9.315	7.4	-4.68506493506493e-07	0.497951962696417	0
X9.305	7.4	-9.06168831168831e-07	0.216367581507682	0
X9.295	7.4	-7.64805194805194e-07	0.485187664690498	0
X9.285	7.4	-3.39155844155844e-07	0.652610228658504	0
X9.275	7.4	-1.75948051948052e-06	0.0425942302510362	1
X9.265	7.4	-9.0961038961039e-07	0.249764495787689	0
X9.255	7.4	-9.32987012987014e-07	0.250657894164028	0
X9.245	7.4	-5.14480519480519e-07	0.571074859972155	0
X9.235	7.4	-1.3712987012987e-06	0.101042049776513	0
X9.225	7.4	-1.04045454545455e-06	0.16210739241644	0
X9.215	7.4	-6.14220779220779e-07	0.478720182593906	0
X9.205	7.4	-1.19707792207792e-06	0.212672664280064	0
X9.195	7.4	-1.21116883116883e-06	0.179465870993595	0
X9.185	7.4	1.29415584415584e-07	0.903025749857124	0
X9.175	7.4	-4.9012987012987e-07	0.603984452306671	0
X9.165	7.4	-5.63915584415585e-06	0.668751195338457	0
X9.155	7.4	-7.55532467532469e-06	0.668010936029291	0
X9.145	7.4	-1.97383116883117e-06	0.784932141383218	0
X9.135	7.4	-6.12922077922078e-07	0.235985993330704	0

GALAXY RESULT: BOXPLOT



35: Univariate_information.txt

34: Univariate_figure.pdf
126.4 KB
format: pdf, database: ?
Epilog : job finished at Fri Nov 25 16:45:12 CET 2016
Image in pdf format

33: Univariate_Generic_Filter_Multivariate_Generic_Filter_Multivariate_NMR_Normalization_variableMetadataOut
161 lines
format: tabular, database: ?
Epilog : job finished at Fri Nov 25 16:45:12 CET 2016

1	2	3
variableMetadata	VariableOrder	PCA_XLOAD-p
B9.126	18	0.03649065E

32: Generic_Filter_Multivariate_Generic_Filter_Multivariate_NMR_Normalization_variableMetadataOut
161 lines
format: tabular, database: ?

T.P.

Galaxy parameters: Choo dataset

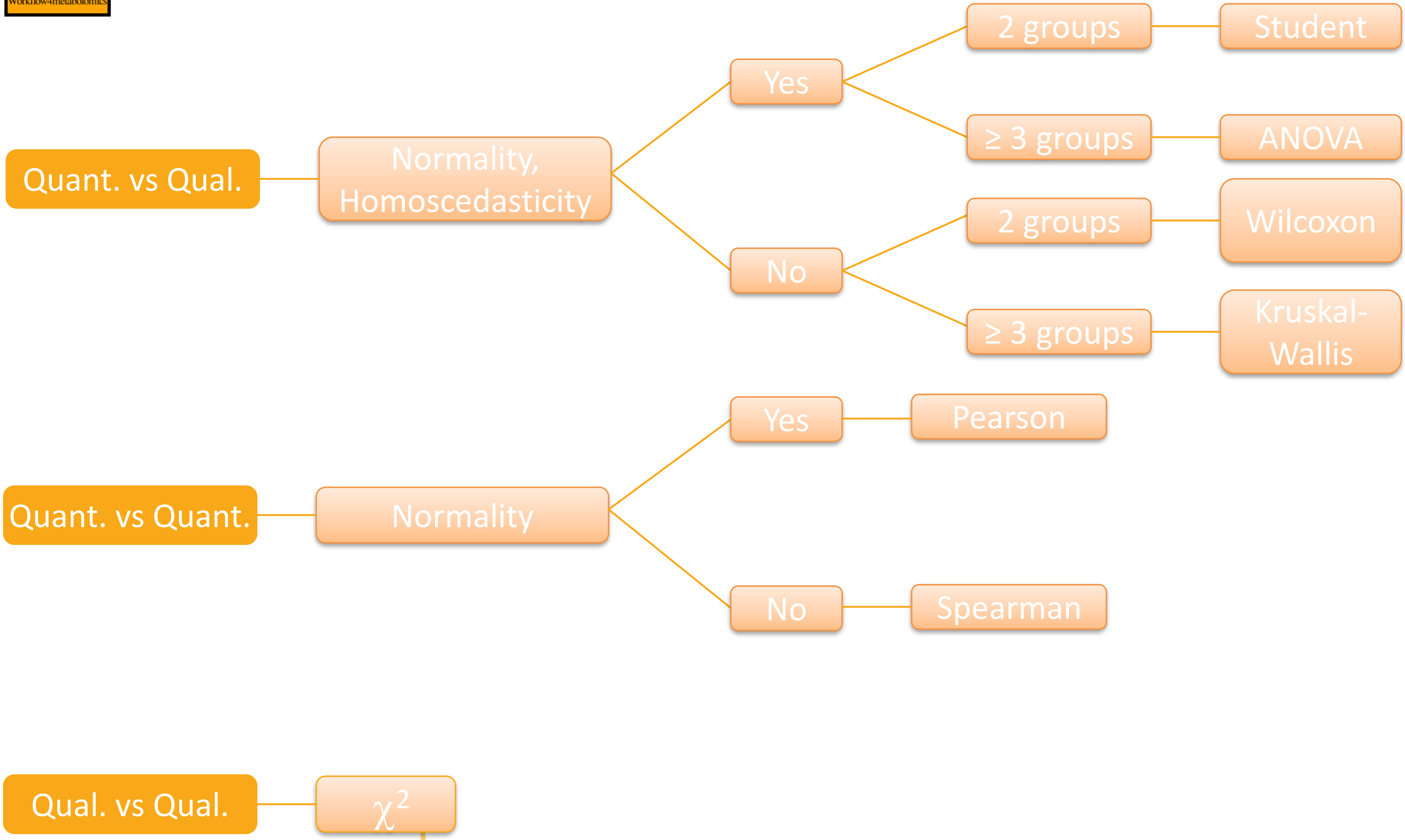
- Bucket width: 0.005
- Left Border: 8.50
- Right Border: 0.50
- Exclusion zone(s):
  - Zone 1: 5.20-4.50

Exercise

- Does Cirpoflaxin influence NMR variable’s relative intensity?
  - Which test would you use (cite parametric or non test)?
  - Run the chosen test with different multiple test correction methods
  - Compare results
- To help in NMR spectrum annotation, we can identify NMR buckets corresponding to the same metabolite(s)
  - Which test do you use?
  - For example, one of Glutamic acid chemical shift is 3.75 ppm
    - Run the parametric and the non parametric versions of the chosen test, using 3.75 ppm as comparison variable
    - Compare results
- Shared history: https://galaxy.workflow4metabolomics.org/u/mtremblayfranco/h/w4e2018---choo-exercice

SUPPLEMENTARY

WHICH TEST FOR WICH DATA?



- Quant. vs Qual.
  - Normality, Homoscedasticity
    - Yes
      - 2 groups
        - Student
      - ≥ 3 groups
        - ANOVA
    - No
      - 2 groups
        - Wilcoxon
      - ≥ 3 groups
        - Kruskal-Wallis
- Quant. vs Quant.
  - Normality
    - Yes
      - Pearson
    - No
      - Spearman
- Qual. vs Qual.
  - χ^2

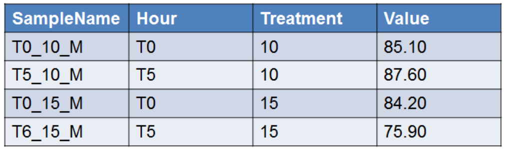
N-WAY ANOVA (1)

SampleName	Hour	Treatment	Glucose value
T0_10_M	T0	10	85.10
T5_10_M	T5	10	87.60
T0_15_M	T0	15	84.20
T6_15_M	T5	15	75.90

- Analysis of variance:

$$SST = SSB + (SSW_1 + SSW_2 + SSW_{ht})$$

N-WAY ANOVA (2)

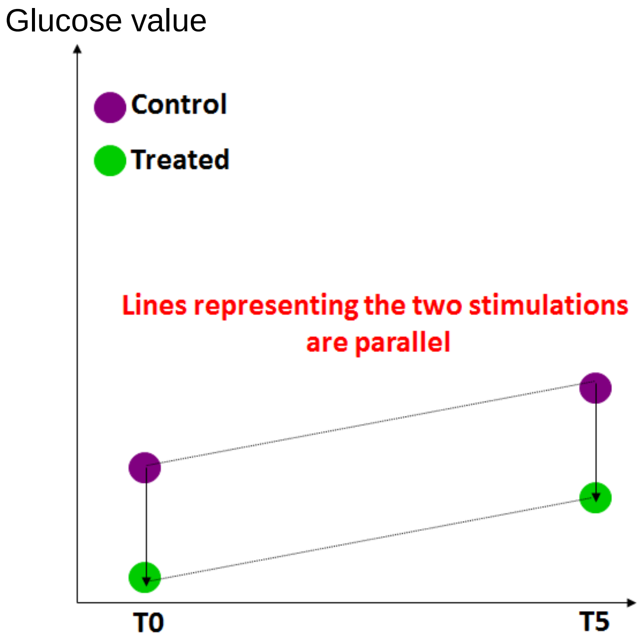


SampleName	Hour	Treatment	Value
T0_10_M	T0	10	85.10
T5_10_M	T5	10	87.60
T0_15_M	T0	15	84.20
T6_15_M	T5	15	75.90

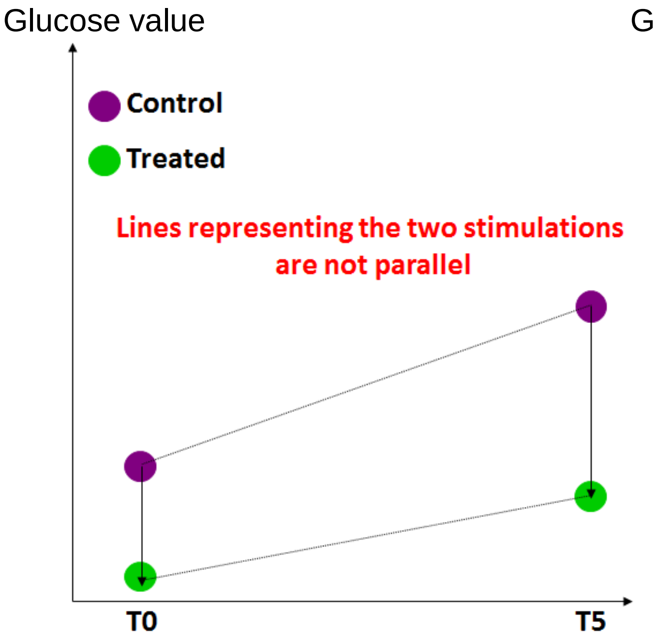
- Several null hypothesis are tested in parallel
  - Relative to Factor 1 (here Hour):
    - H_0: "µ1i = µ1j = ... "
    - H_1: "there is at least 1 average of the F1 different from other"
  - Relative to Factor 2 (here Treatment):
    - H_0: "µ2i = µ2j = ... "
    - H_1: "there is at least 1 average of the F2 different from other"
  - Relative to the interaction between the 2 Factors :
    - H_0: "there is not any interaction between factors 1 and 2"
    - H_1: "there is interactions between factors 1 and 2"

N-WAY ANOVA (3)

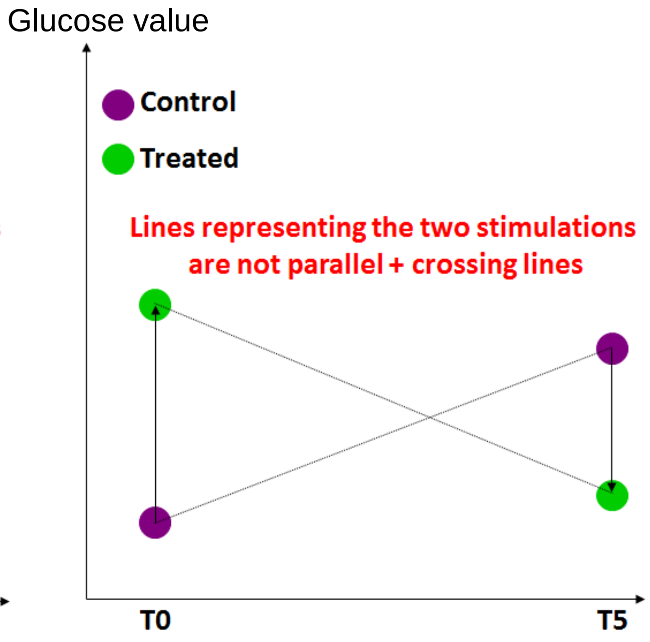
- Interaction: combined effects of factors (qualitative variables) on the quantitative variable



Lines representing the two stimulations are parallel



Lines representing the two stimulations are not parallel



Lines representing the two stimulations are not parallel + crossing lines

N-WAY ANOVA (4)

SampleName	Hour	Treatment	Value
T0_10_M	T0	10	85.10
T5_10_M	T5	10	87.60
T0_15_M	T0	15	84.20
T6_15_M	T5	15	75.90

$$\forall i,j \; n_{ij} = \frac{n_{i+} \times n_{+j}}{n}$$

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \text{ with } \epsilon_{ijk} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$$

$$i \in \{1,\dots,I\}$$

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \sum_j \gamma_{ij} = 0$$

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

$$Y_{ijk} \mid \mu_{ij}, \sigma^2 \overset{i.i.d.}{\sim} \mathcal{N}(\mu_{ij}, \sigma^2)$$

$$\mathrm{P}(X \leq x, Y \leq y \mid Z = z) = \mathrm{P}(X \leq x \mid Z = z) \cdot \mathrm{P}(Y \leq y \mid Z = z)$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Hour	2	31.1	15.55	5.923	0.005
Treatment	1	2.017	2.017	0.3772	0.54
Hour:Treatment	1	63.27	84.20	11.83	< 0.001
Residuals	10	5.346	0.5346