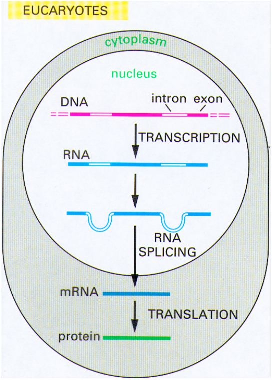
UNIVERSITEIT GENT

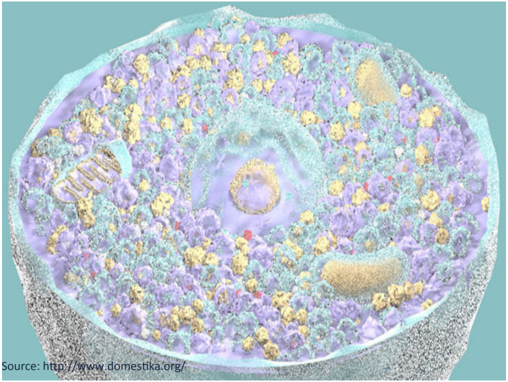
BIG N2N
Bioinformatics Institute Ghent
from nucleotides to networks

# Statistical Methods for Quantitative MS-Based Proteomics: 1. Identification & False discovery rate

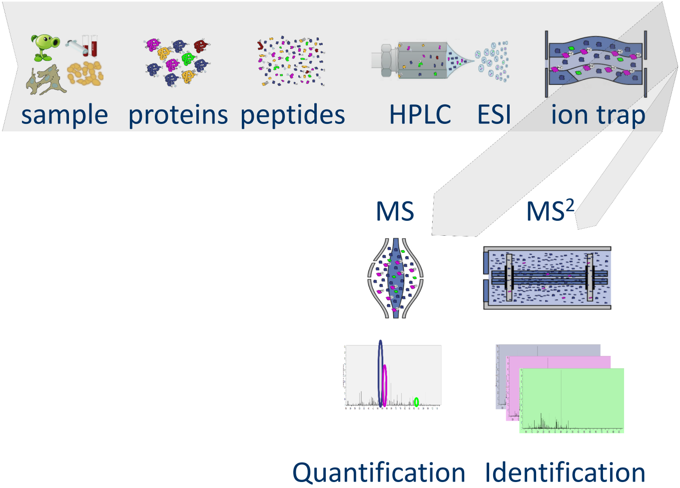
Lieven Clement

Proteomics Data Analysis Shortcourse

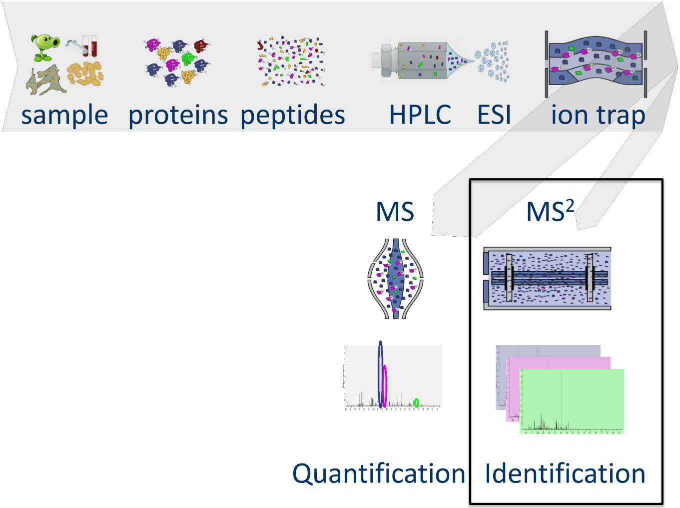

EUCARYOTES
cytoplasm
nucleus
DNA
intron
exon
TRANSCRIPTION
RNA
RNA
SPLICING
mRNA
TRANSLATION
protein


Source: http://www.domestika.org/

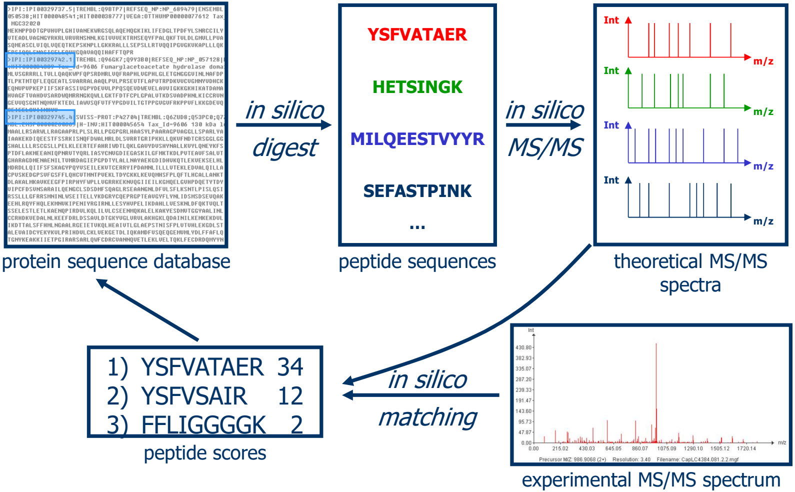
# Challenges in Label Free MS-based Quantitative Proteomics



# Challenges in Label Free MS-based Quantitative Proteomics



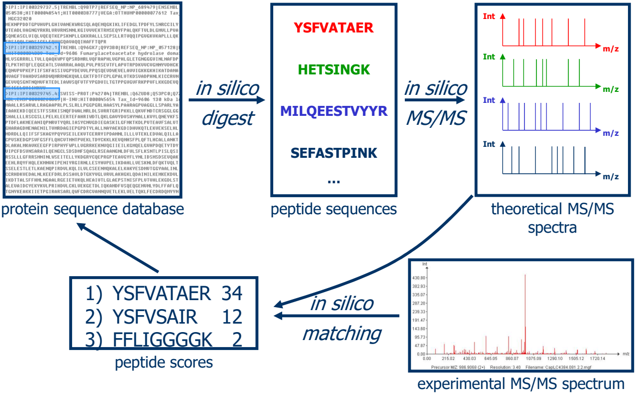
# Identification




(slide courtesy to Lennart Martens)

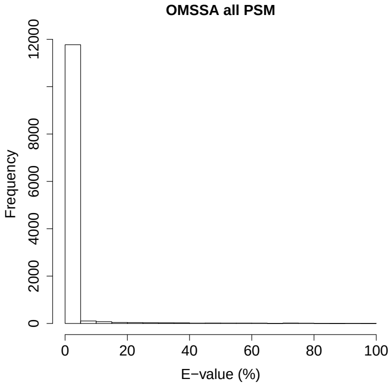
# E-values

Probability that a random candidate peptide produces a higher score that the observed PSM score.



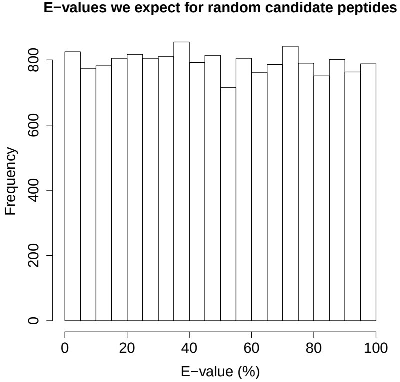
# E-values

Probability that a random candidate peptide produces a higher score that the observed PSM score.



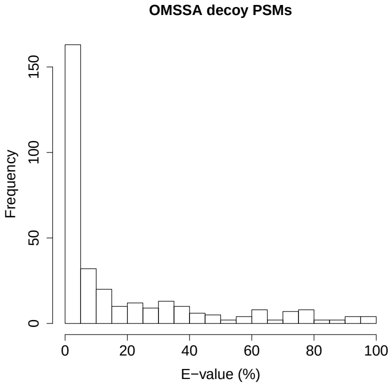
# E-values

Probability that a random candidate peptide produces a higher score that the observed PSM score.



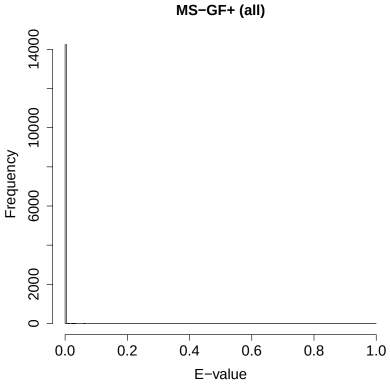
# E-values

Probability that a random candidate peptide produces a higher score that the observed PSM score.



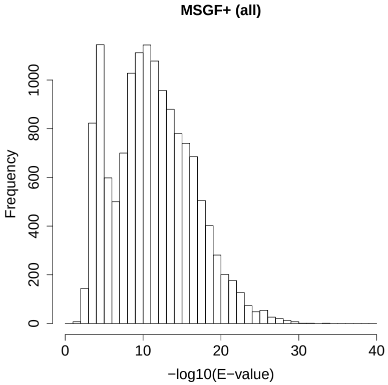
# E-values

Probability that a random candidate peptide produces a higher score that the observed PSM score.



# E-values

Probability that a random candidate peptide produces a higher score that the observed PSM score.



# E-values

Probability that a random candidate peptide produces a higher score that the observed PSM score.

- A bad hit is the random hit with the best score so it is also bound to have a low E-value.

# E-values

Probability that a random candidate peptide produces a higher score that the observed PSM score.

- A bad hit is the random hit with the best score so it is also bound to have a low E-value.
- If we look at E-values for all PSMs they are only useful as a score.

# E-values

Probability that a random candidate peptide produces a higher score that the observed PSM score.

- A bad hit is the random hit with the best score so it is also bound to have a low E-value.
- If we look at E-values for all PSMs they are only useful as a score.
- We should know the distribution of the maximum score of random candidate peptides when we want to do the statistics.

# Table of Outcomes

| | Called Bad | Called Correct | |
|---|---|---|---|
| Bad hit | TN | FP | $m_0$ |
| Correct hit | FN | TP | $m_1$ |
| Total | NR | R | $m$ |

- TN: number of true negatives
- FP: number of false positives
- FN: number of false negatives
- TP: number of true positives
- NR: number of non-rejections, R: number of rejections

# Table of Outcomes

| | | Called Bad | Called Correct | |
|---|---|---|---|---|
| Unobservable | Bad hit | TN | FP | $m_0$ |
| | Correct hit | FN | TP | $m_1$ |
| Observable | Total | NR | R | $m$ |

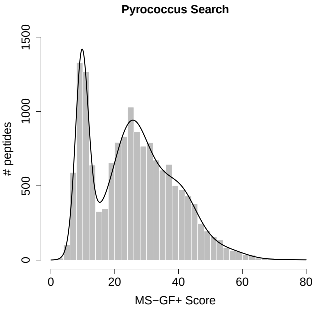
$FDP = \frac{FP}{FP+TP}$. But is unkown! (FDP: false discovery proportion)

# Table of Outcomes

| | | Called Bad | Called Correct | |
|---|---|---|---|---|
| Unobservable | Bad hit | TN | FP | $m_0$ |
| | Correct hit | FN | TP | $m_1$ |
| Observable | Total | NR | R | $m$ |

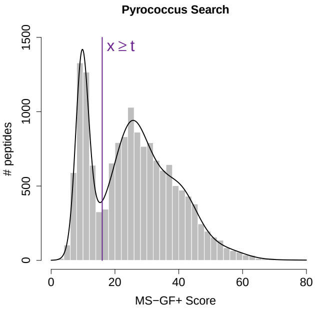
$FDR = E\left[\frac{FP}{FP+TP}\right]$. (FDR: false discovery rate)

# Search engines return score that discriminates good from bad matches

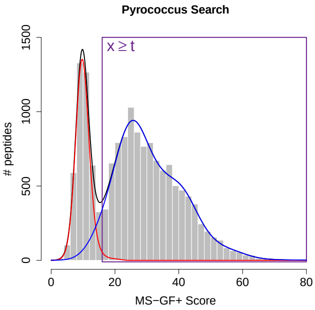


# Search engines return score that discriminates good from bad matches

Score threshold $t$?



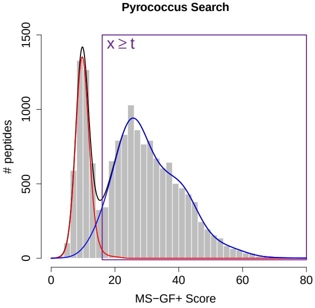
# Search engines return score that discriminates good from bad matches




Score threshold $t$?

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_1(x)$$

# Search engines return score that discriminates good from bad matches

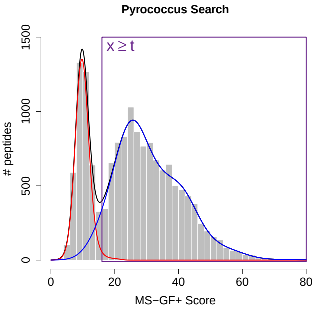



Score threshold $t$?

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_1(x)$$

$$\text{FDR}(t) = E\left[\frac{FP}{FP + TP}\right]$$

# Search engines return score that discriminates good from bad matches




Score threshold $t$?

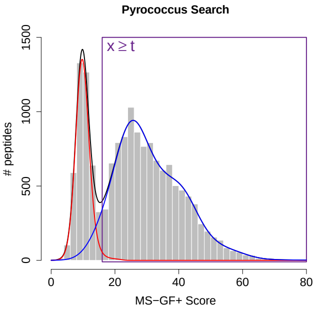
$$f(x) = \pi_0 f_0(x) + (1-\pi_0) f_1(x)$$

$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

$$\text{FDR}(t) = \frac{m_0 P[x \geq t | FP]}{m P[x \geq t]}$$

$$= \frac{m P[FP] P[x \geq t | FP]}{m P[x \geq t]}$$

# Search engines return score that discriminates good from bad matches




Score threshold $t$?

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_1(x)$$

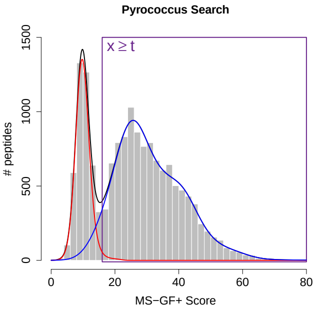
$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

$$\text{FDR}(t) = \frac{m_0 P[x \geq t | FP]}{m P[x \geq t]}$$

$$= \frac{m P[FP] P[x \geq t | FP]}{m P[x \geq t]}$$

$$\text{FDR}(t) = \frac{\pi_0 P_0[x \geq t]}{P[x \geq t]}$$

# Search engines return score that discriminates good from bad matches




Score threshold $t$?

$$f(x) = \pi_0 f_0(x) + (1-\pi_0) f_1(x)$$

$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

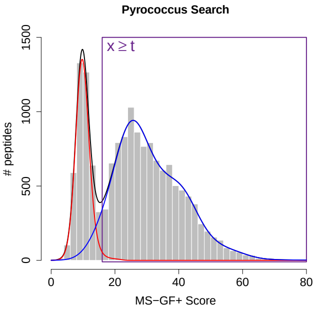
$$\text{FDR}(t) = \frac{m_0 P[x \geq t|FP]}{mP[x \geq t]}$$

$$= \frac{mP[FP]P[x \geq t|FP]}{mP[x \geq t]}$$

$$\text{FDR}(t) = \frac{\pi_0 P_0[x \geq t]}{P[x \geq t]}$$

$$P_{.}[x \geq t] = \int_{x=t}^{+\infty} f_{.}(x)dx$$

# Search engines return score that discriminates good from bad matches




Score threshold $t$?

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_1(x)$$

$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

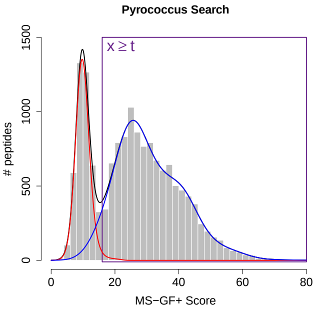
$$\text{FDR}(t) = \frac{m_0 P[x \geq t | FP]}{m P[x \geq t]}$$

$$= \frac{m P[FP] P[x \geq t | FP]}{m P[x \geq t]}$$

$$\text{FDR}(t) = \frac{\pi_0 P_0[x \geq t]}{P[x \geq t]}$$

FDR is a set property: $FDR(t) = \dfrac{\pi_0 \int_{x=t}^{+\infty} f_0(x)dx}{\int_{x=t}^{+\infty} f(x)dx}$

# Search engines return score that discriminates good from bad matches




Score threshold $t$?

$$f(x) = \pi_0 f_0(x) + (1-\pi_0) f_1(x)$$

$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

$$\text{FDR}(t) = \frac{m_0 P[x \geq t|FP]}{mP[x \geq t]}$$

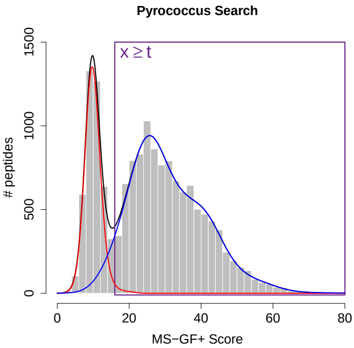
$$= \frac{mP[FP]P[x \geq t|FP]}{mP[x \geq t]}$$

$$\text{FDR}(t) = \frac{\pi_0 P_0[x \geq t]}{P[x \geq t]}$$

local fdr (posterior error probability, PEP): $fdr(x) = \frac{\pi_0 f_0(x)}{f(x)}$

Probability that an individual PSM is a bad hit.

# How to estimate FDR?

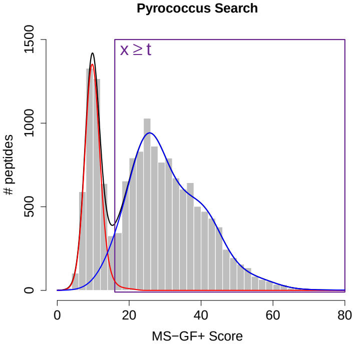



$$\mathrm{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

$$= \frac{\pi_0 P_0[x \geq t]}{P[x \geq t]}$$

$$P_{\cdot}[x \geq t] = \int_t^{\infty} f_{\cdot}(x)dx$$

# How to estimate FDR?



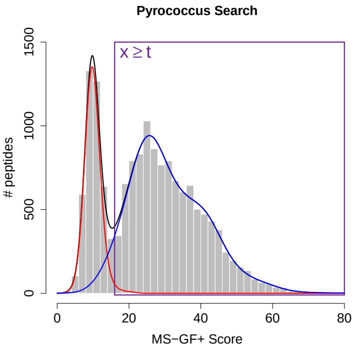

$$\mathrm{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

$$= \frac{\pi_0 P_0[x \geq t]}{P[x \geq t]}$$

$$P_{\cdot}[x \geq t] = \int_t^{\infty} f_{\cdot}(x)dx$$

$$\hat{P}[x \geq t] = \frac{\#x \geq t}{m} \quad \Rightarrow \quad \widehat{\mathrm{FDR}}(t) = \frac{\pi_0 P_0[x \geq t]}{\frac{\#x \geq t}{m}}$$

# How to estimate FDR?




$$\mathrm{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

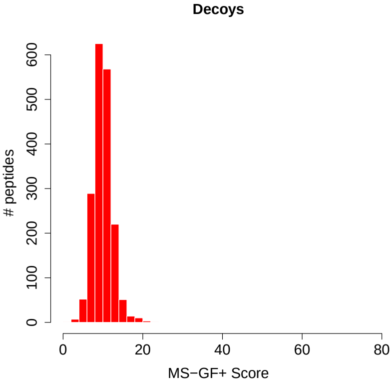
$$= \frac{\pi_0 P_0[x \geq t]}{P[x \geq t]}$$

$$P_{\cdot}[x \geq t] = \int_t^{\infty} f_{\cdot}(x)dx$$

$$\hat{P}[x \geq t] = \frac{\#x \geq t}{m} \quad \Rightarrow \quad \widehat{\mathrm{FDR}}(t) = \frac{\pi_0 P_0[x \geq t]}{\frac{\#x \geq t}{m}}$$

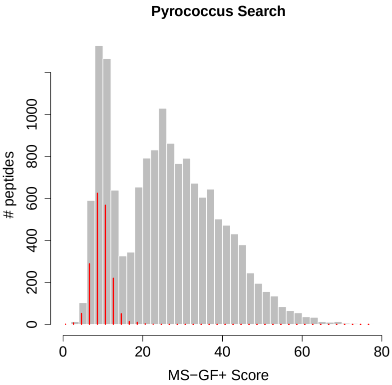
How to characterize $f_0(t)$ and $\pi_0$ in proteomics?

# Target-Decoy approach to establish null distribution



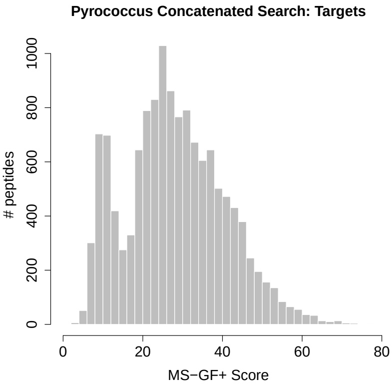

- Search against decoy database to generate representative bad hits
- Reversed databases are popular

# Target-Decoy approach to establish null distribution



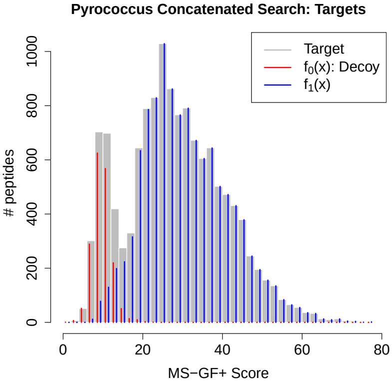

- Search against decoy database to generate representative bad hits
- Reversed databases are popular
- Concatenated search

# Target-Decoy approach to establish null distribution




- Search against decoy database to generate representative bad hits
- Reversed databases are popular
- Concatenated search

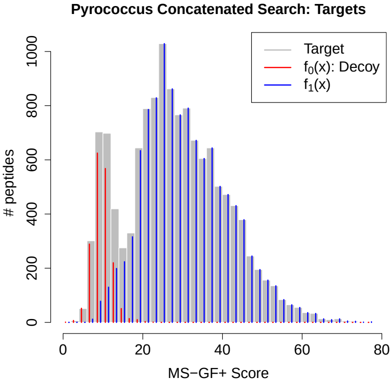
# Target-Decoy approach to establish null distribution




- Search against decoy database to generate representative bad hits
- Reversed databases are popular
- Concatenated search
- Assumption: bad hits has equal probability to map on target and decoy

$$\hat{\pi}_0 = \frac{\#decoys}{\#targets}$$

# Target-Decoy approach to establish null distribution

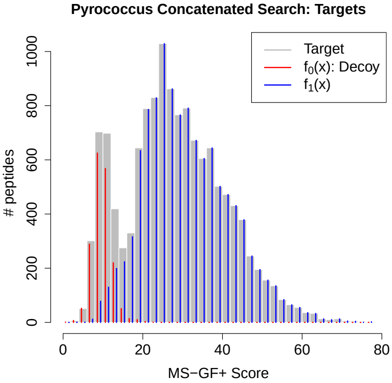



- Search against decoy database to generate representative bad hits
- Reversed databases are popular
- Concatenated search
- Assumption: bad hits has equal probability to map on target and decoy

$$\hat{\pi}_0 = \frac{\#decoys}{\#targets}$$

- Score cuttoff: $\text{FDR}(x) = E\left[\frac{FP}{FP+TP}\right]$

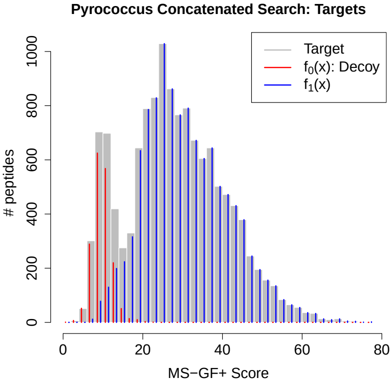
# Target-Decoy approach to establish null distribution




- Competitive Target - decoy:

$$\widehat{\text{FDR}}(x) = \frac{\#decoys|X \geq x}{\#targets|X \geq x}$$

# Target-Decoy approach to establish null distribution




- Competitive Target - decoy:

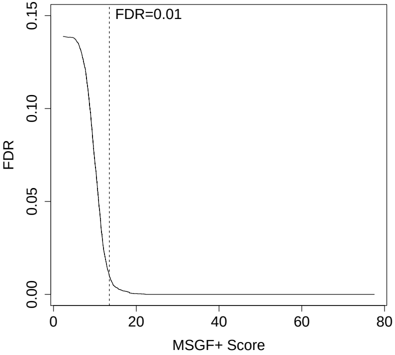
$$\widehat{\text{FDR}}(x) = \frac{\#decoys|X \geq x}{\#targets|X \geq x}$$

$$\widehat{\text{FDR}}(x) = \frac{\#decoys}{\#targets} \frac{\frac{\#decoys|X\geq x}{\#decoys}}{\frac{\#targets|X\geq x}{\#targets}}$$

$$\widehat{\text{FDR}}(x) = \hat{\pi}_0 \frac{\widehat{\int_t^{+\infty} f_0(x)dx}}{\widehat{\int_t^{+\infty} f(x)dx}}$$

$$\widehat{\text{FDR}}(x) = \hat{\pi}_0 \frac{\hat{P}_0[X \geq x]}{\hat{P}[X \geq x]}$$

# Target-Decoy approach to establish null distribution




- Competitive Target - decoy:

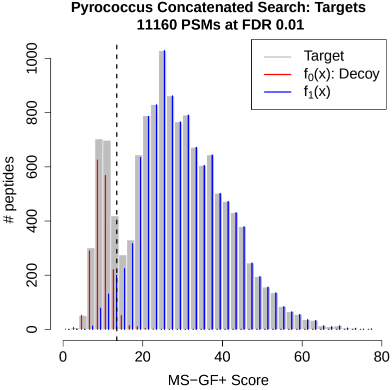
$$\widehat{\text{FDR}}(x) = \frac{\#decoys|X \geq x}{\#targets|X \geq x}$$

$$\widehat{\text{FDR}}(x) = \frac{\#decoys}{\#targets} \frac{\frac{\#decoys|X\geq x}{\#decoys}}{\frac{\#targets|X\geq x}{\#targets}}$$

$$\widehat{\text{FDR}}(x) = \hat{\pi}_0 \frac{\widehat{\int_t^{+\infty} f_0(x)dx}}{\widehat{\int_t^{+\infty} f(x)dx}}$$

$$\widehat{\text{FDR}}(x) = \hat{\pi}_0 \frac{\hat{P}_0[X \geq x]}{\hat{P}[X \geq x]}$$

# Target-Decoy approach to establish null distribution




- Competitive Target - decoy:

$$\widehat{\mathrm{FDR}}(x) = \frac{\#decoys|X \geq x}{\#targets|X \geq x}$$

$$\widehat{\mathrm{FDR}}(x) = \frac{\#decoys}{\#targets} \frac{\frac{\#decoys|X \geq x}{\#decoys}}{\frac{\#targets|X \geq x}{\#targets}}$$

$$\widehat{\mathrm{FDR}}(x) = \hat{\pi}_0 \frac{\widehat{\int_t^{+\infty} f_0(x)dx}}{\widehat{\int_t^{+\infty} f(x)dx}}$$

$$\widehat{\mathrm{FDR}}(x) = \hat{\pi}_0 \frac{\hat{P}_0[X \geq x]}{\hat{P}[X \geq x]}$$

# Assess TDA assumptions

We have to evaluate that

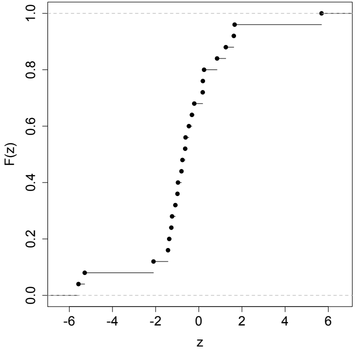
- The decoys are good simulations of the bad target hits: compare distributions $F_D(x)$ with $F(x)$

$$F_D(x) = \int_{-\infty}^{t} f_D(x)dx \quad \leftrightarrow \quad F(x) = \int_{-\infty}^{t} f(x)dx$$

- $\hat{\pi}_0 = \frac{\#decoys}{\#targets}$ is a good estimator for $\pi_0$.

- We will use Probability-Probability-plots (PP-plot) for this purpose.

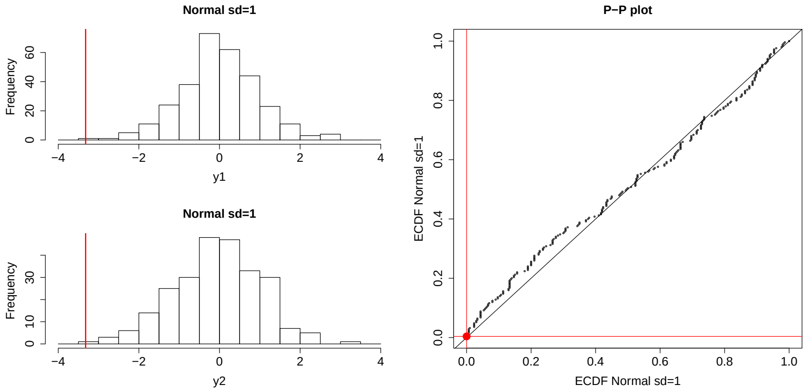
- To make PP-plots we need estimates for $F_D(x)$ and $F(x)$.
- The empirical cumulative distribution (ECDF) is used for that purpose




$$\hat{F}_D(x) = \frac{\#decoys|X \leq x}{\#decoys}, \quad \hat{F}(x) = \frac{\#targets|X \leq x}{\#targets}$$

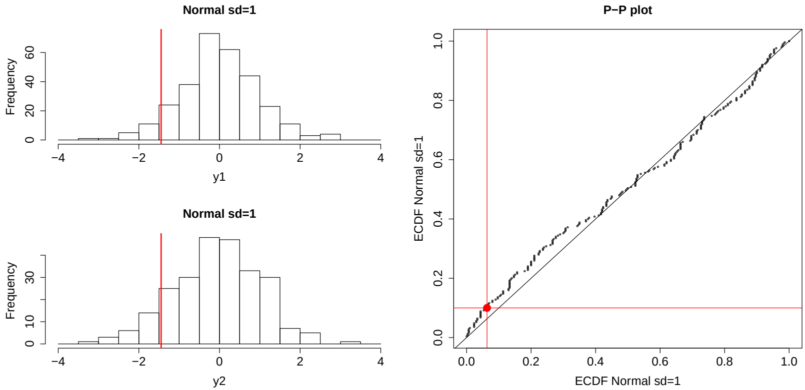
# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



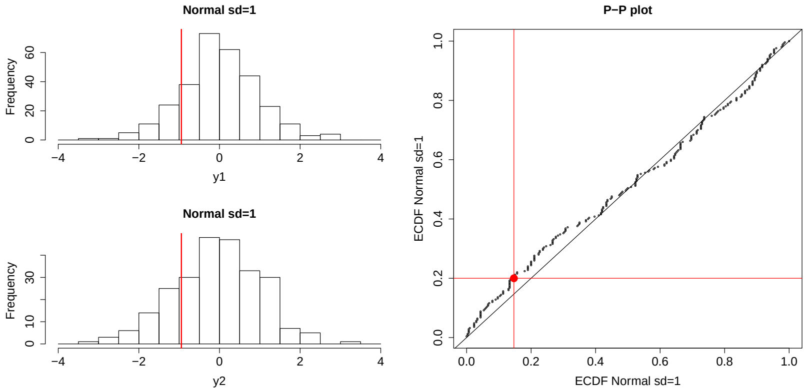
# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



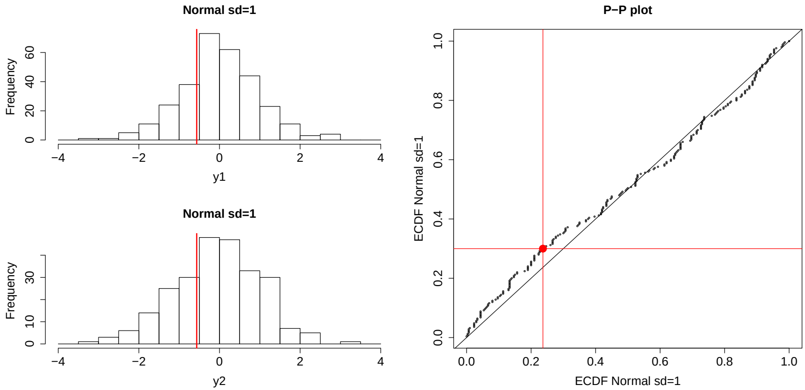
# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



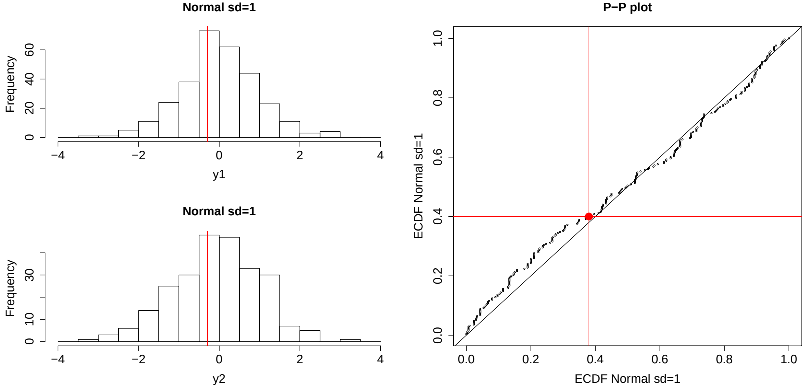
# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



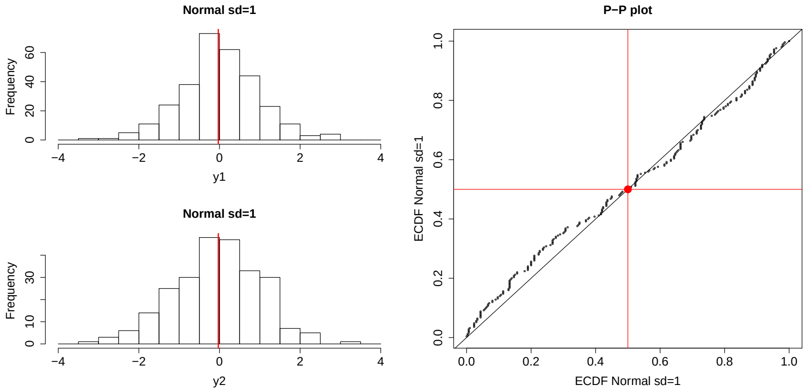
# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



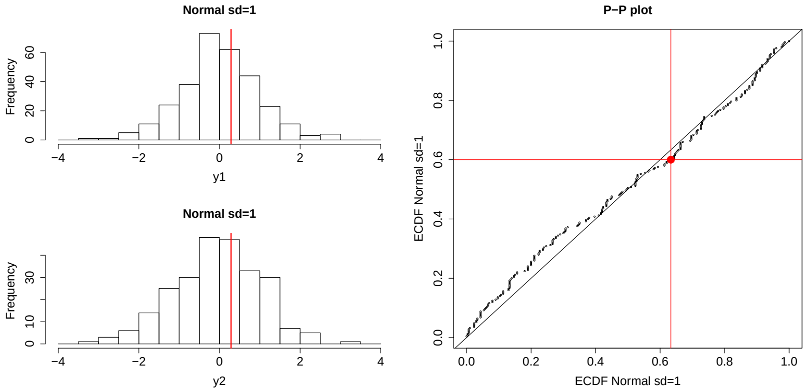
# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



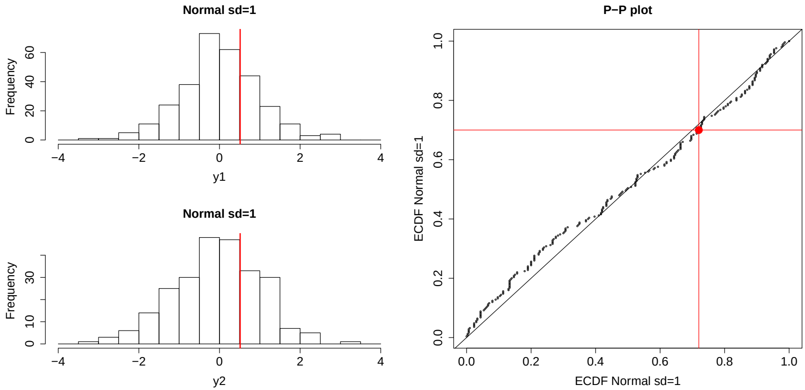
# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



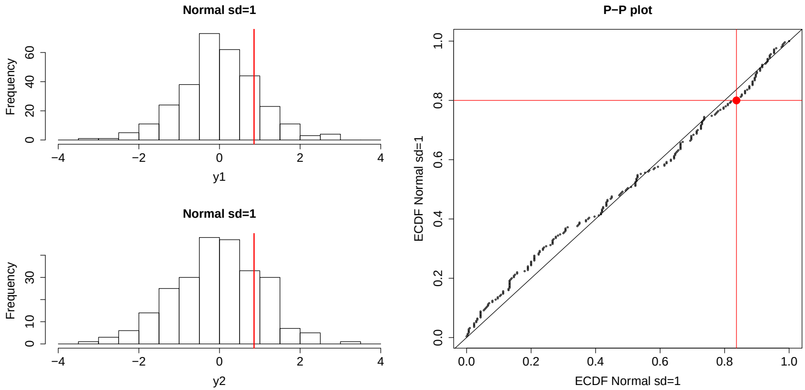
# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



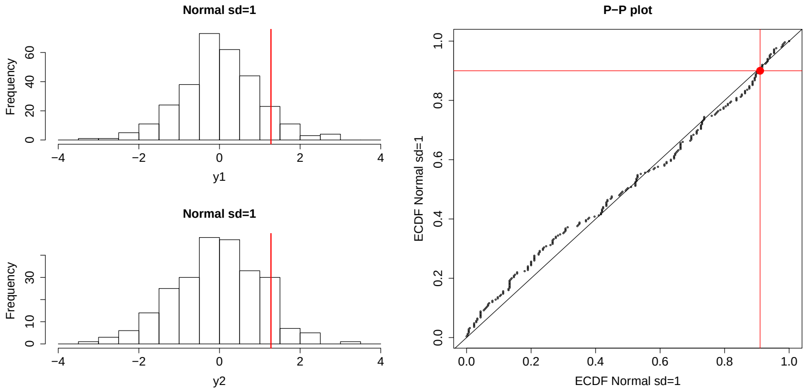
# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



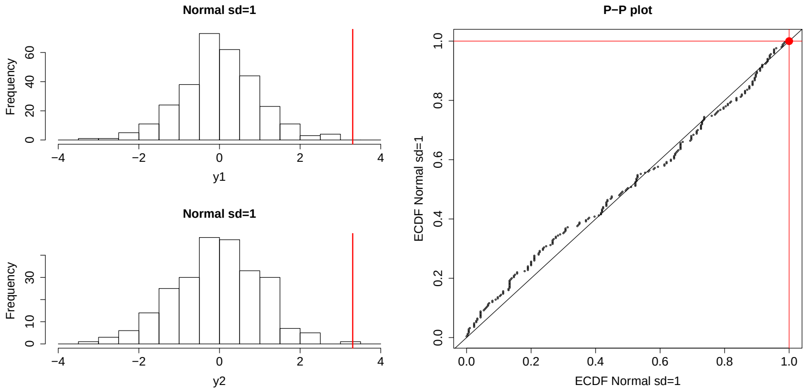
# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.

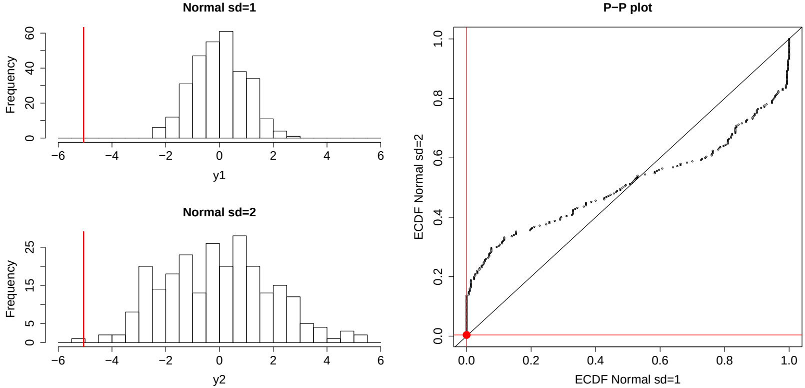


# PP-plot

PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.

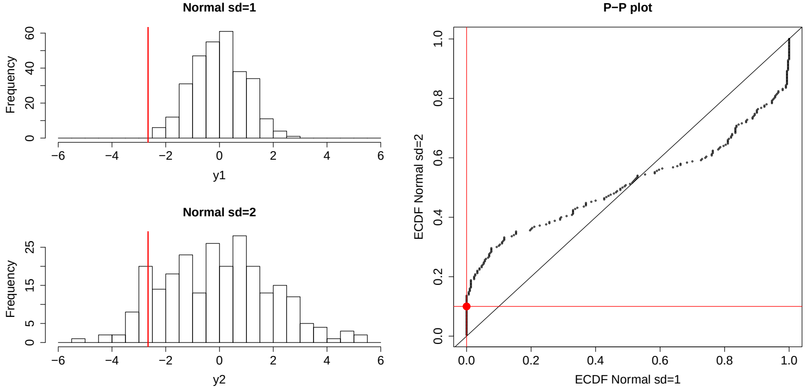


# PP-plot



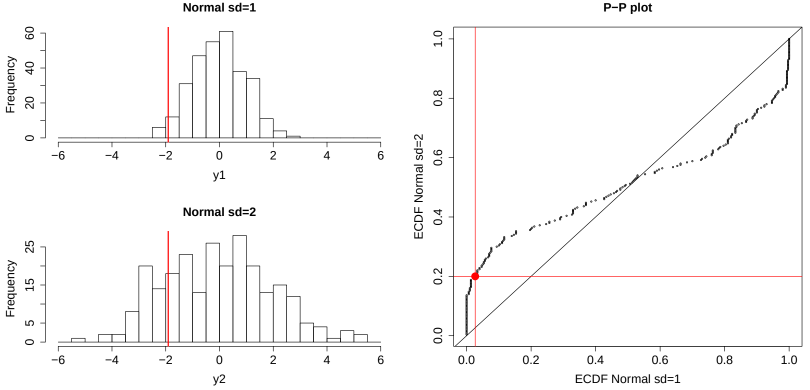
Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

# PP-plot

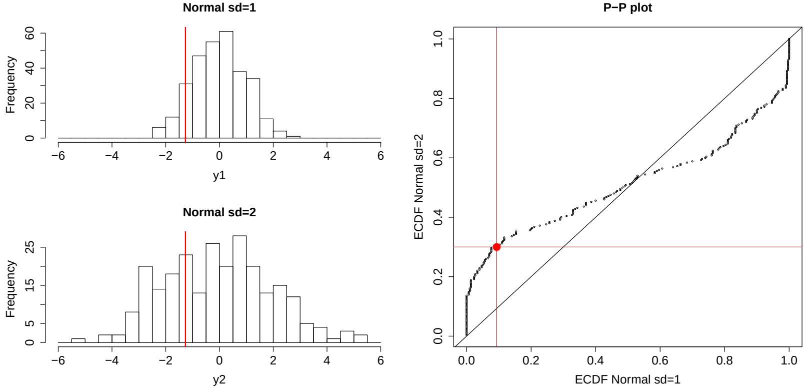


Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

# PP-plot

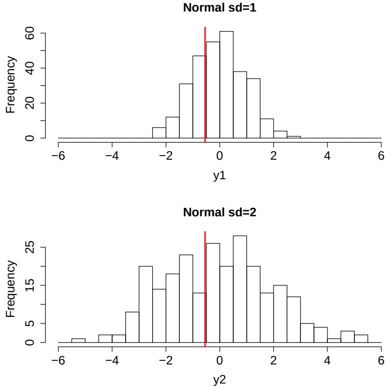

Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

# PP-plot

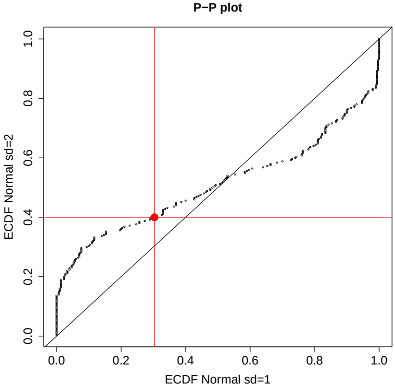


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

# PP-plot

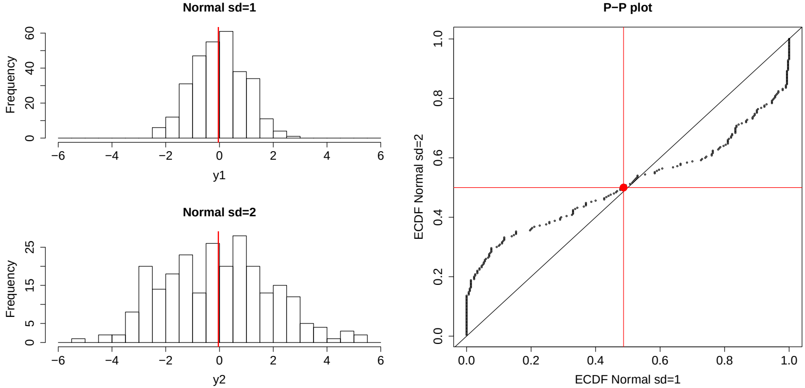


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2

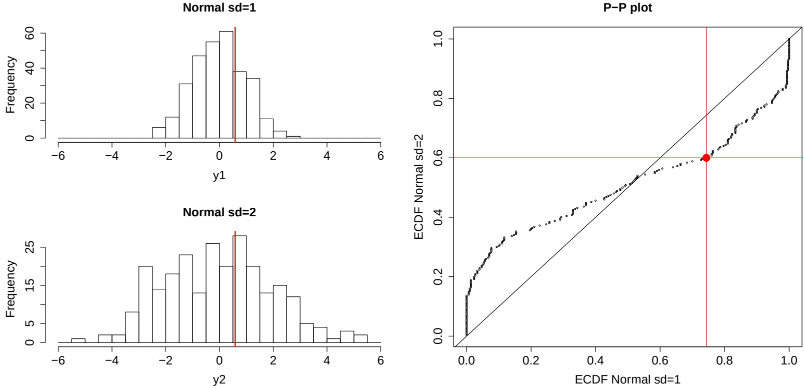
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

# PP-plot



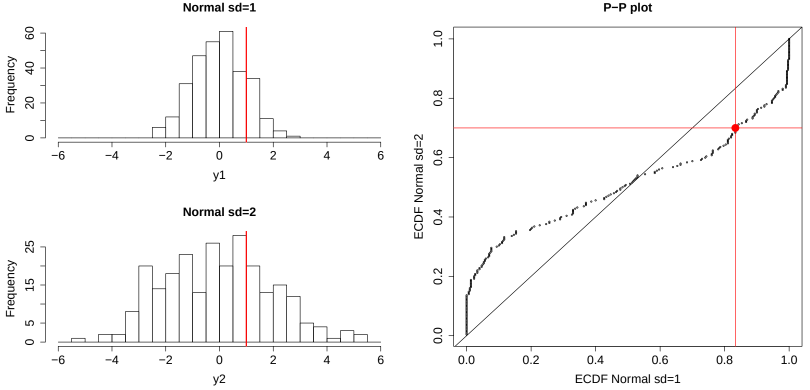
Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

# PP-plot



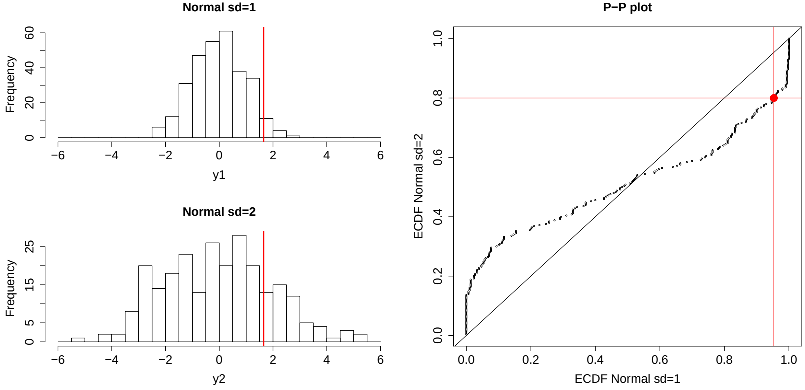
Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

# PP-plot



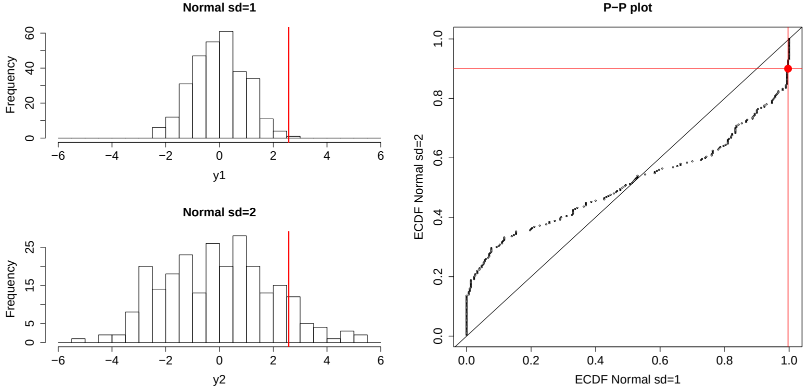
Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

# PP-plot

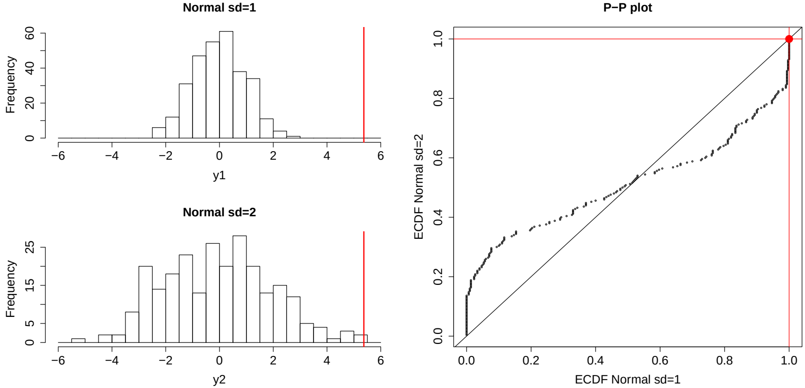


Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

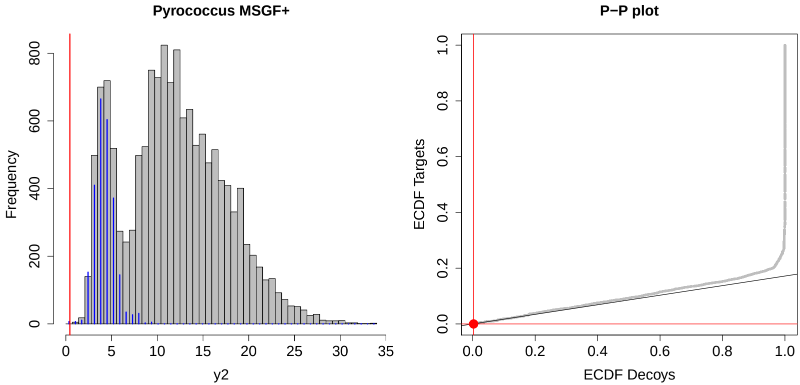
# PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

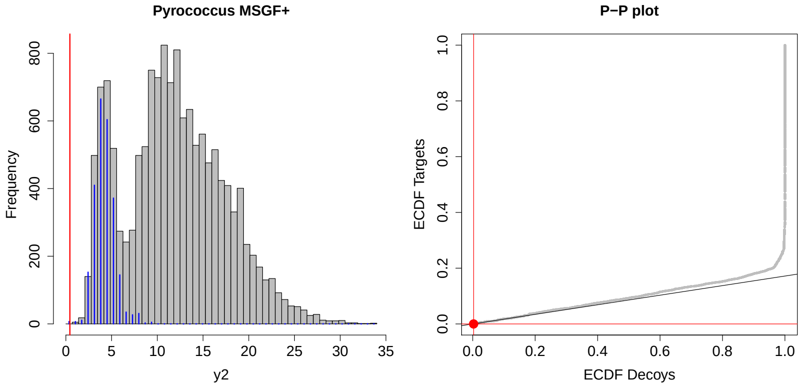
# PP-plot


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

# PP-plot: pyrococcus

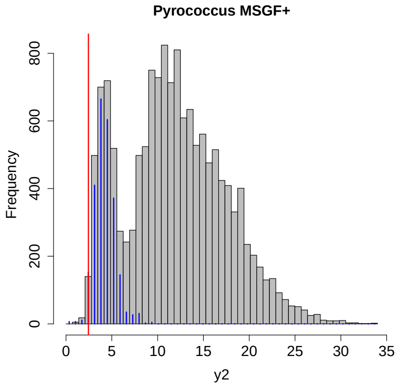


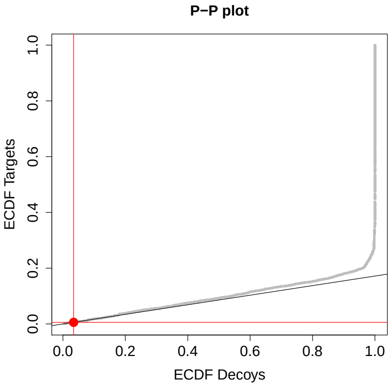
# PP-plot: pyrococcus



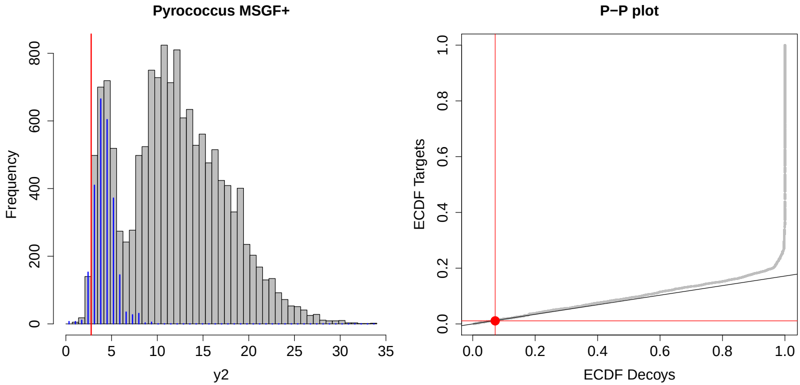

What about $\hat{\pi}_0$?

# PP-plot: pyrococcus

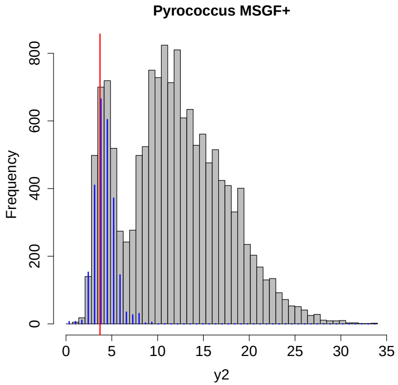


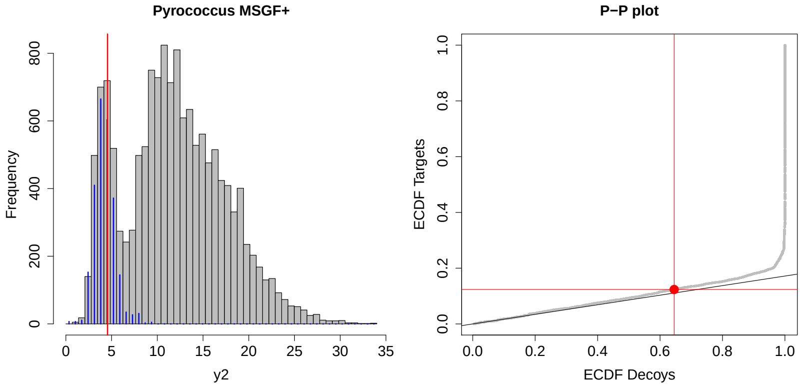
# PP-plot: pyrococcus



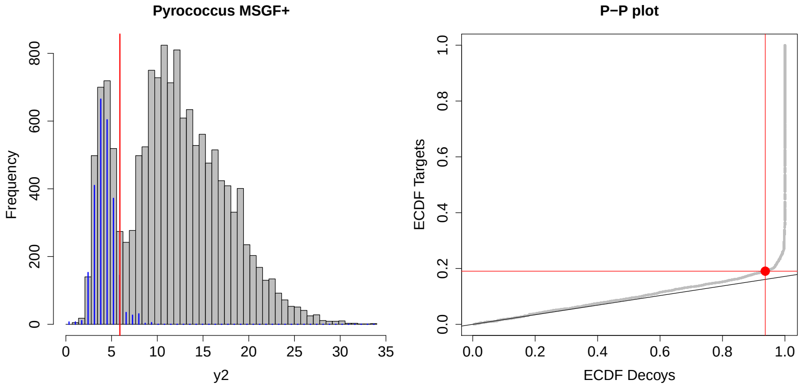
# PP-plot: pyrococcus



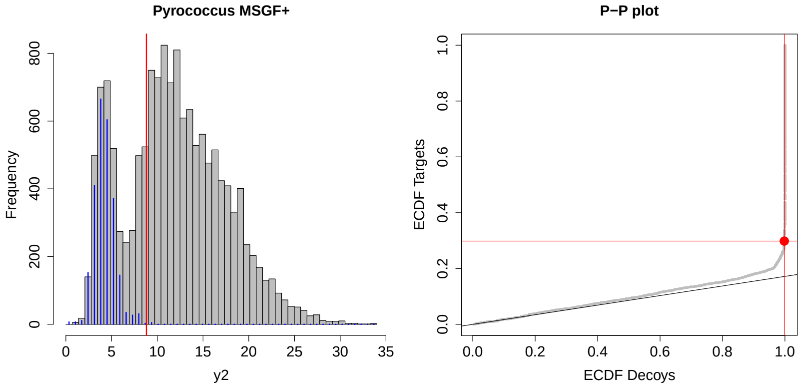
# PP-plot: pyrococcus



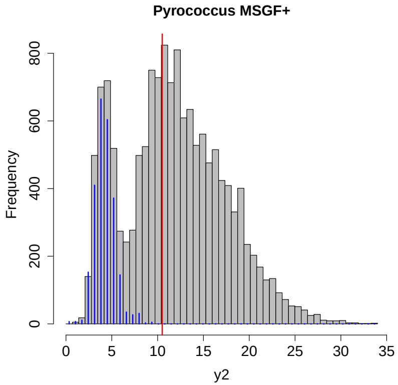
# PP-plot: pyrococcus

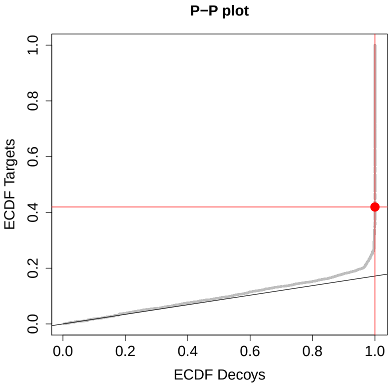


# PP-plot: pyrococcus

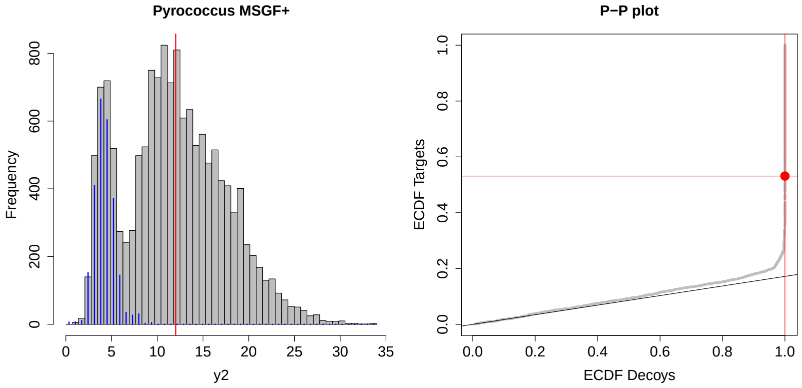


# PP-plot: pyrococcus

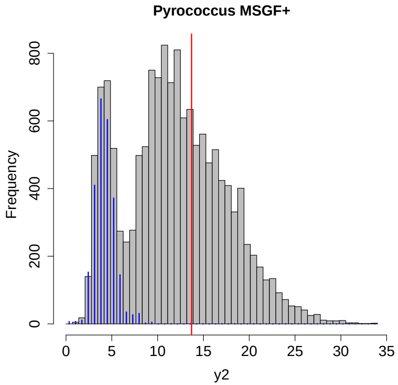


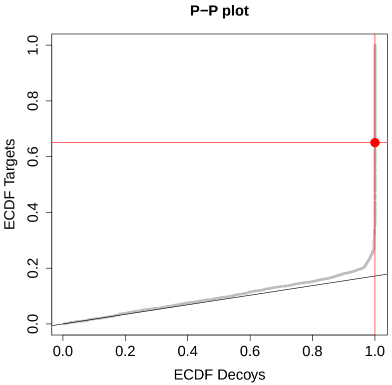



# PP-plot: pyrococcus

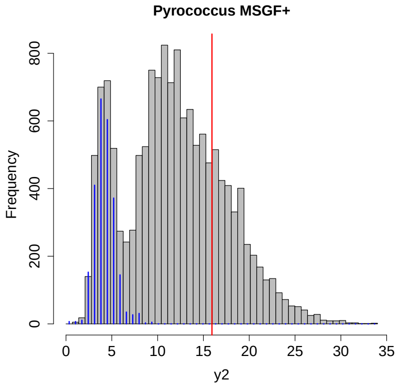


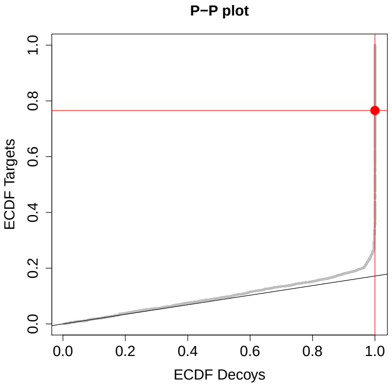
# PP-plot: pyrococcus



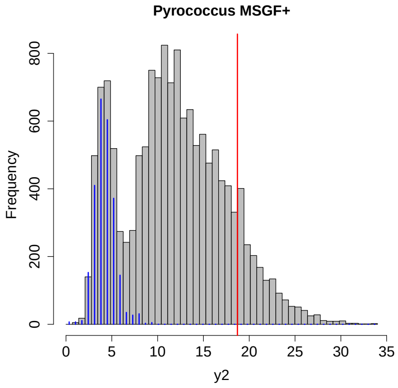



# PP-plot: pyrococcus



# PP-plot: pyrococcus



# PP-plot: pyrococcus