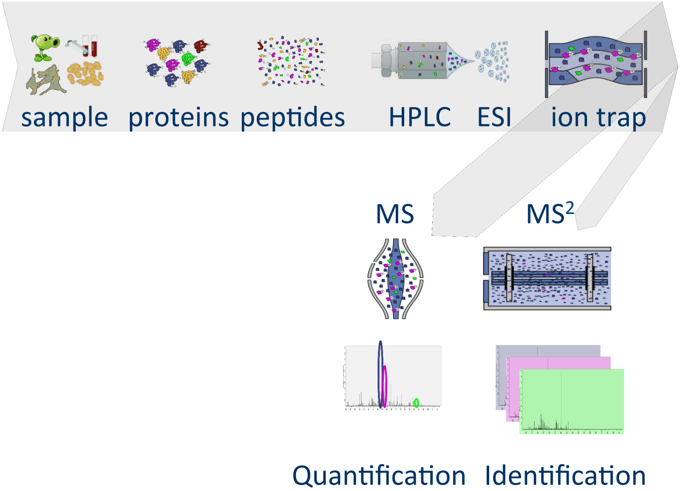
Statistical Methods for Quantitative MS-Based Proteomics:
1. Identification & False discovery rate

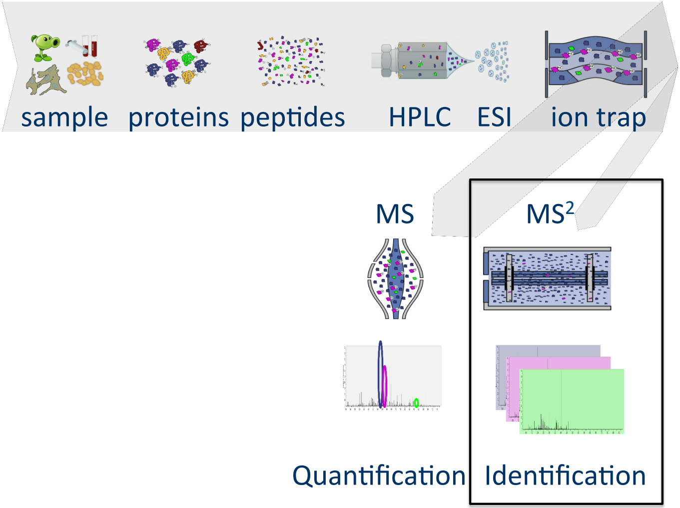
Lieven Clement

Statistics and Genomics Seminar, UCBerkeley, California

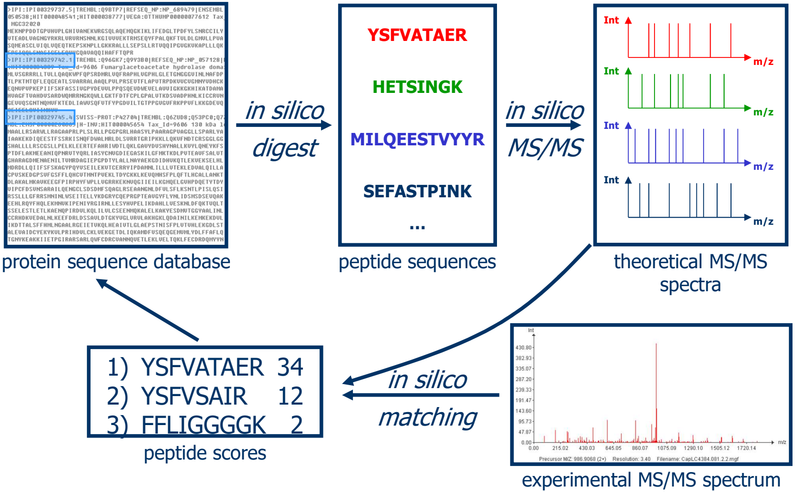
Challenges in Label Free MS-based Quantitative Proteomics



Challenges in Label Free MS-based Quantitative Proteomics



Identification



(slide courtesy to Lennart Martens)

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	
Bad hit	TN	FP	m_0
Correct hit	FN	TP	m_1
Total	NR	R	m

Random Variables

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

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

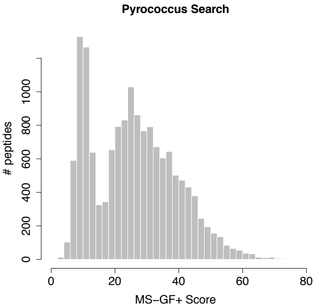
$FDP = \frac{FP}{FP+TP}$. But is unkown! (FDP: false discovery proportion)

Table of Outcomes

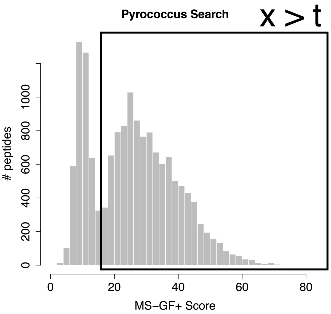
		Called Bad	Called Correct	
Unobservable	Bad hit	TN	FP	m_0
	Correct hit	FN	TP	m_1
Observable	Total	NR	R	m

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

Search engines return score that discriminates good from bad matches

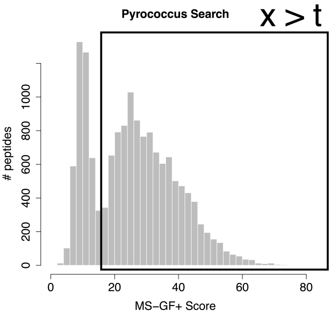


Search engines return score that discriminates good from bad matches



Score threshold t?

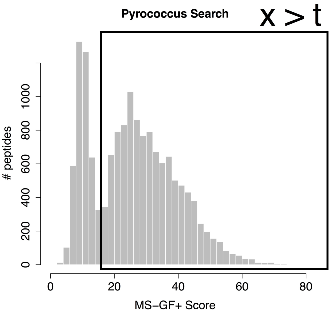
Search engines return score that discriminates good from bad matches



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

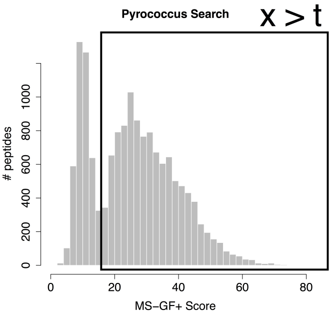


Score threshold t?

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

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

Search engines return score that discriminates good from bad matches



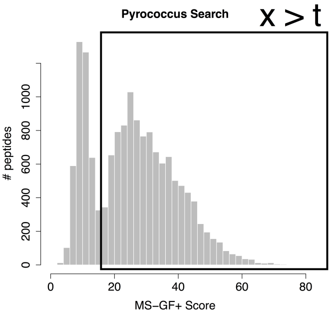
Score threshold t?

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

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

$$\text{FDR}(t) = \frac{mPr[FP]Pr[x>t|FP]}{mPr[x>t]}$$

Search engines return score that discriminates good from bad matches



Score threshold t?

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

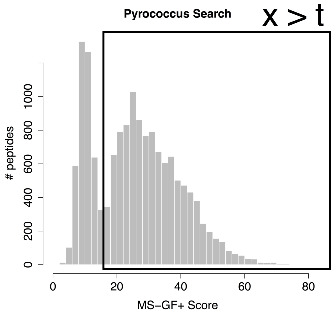
$$\text{FDR}(t) = E\left[\frac{FP}{FP+TP}\right]$$

$$\text{FDR}(t) = \frac{mPr[FP]Pr[x>t|FP]}{mPr[x>t]}$$

$$\text{FDR}(t) = \frac{\pi_0 Pr[x>t|FP]}{Pr[x>t]}$$

$$\text{FDR}(t) = Pr\left[FP|x > t\right]$$

Search engines return score that discriminates good from bad matches



Score threshold t?

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

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

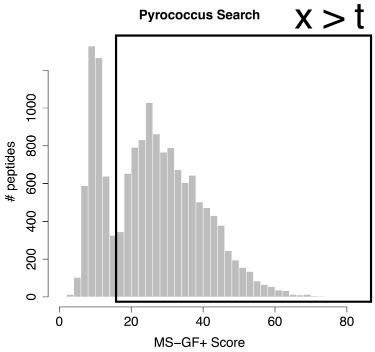
$$\text{FDR}(t) = \frac{mPr[FP]Pr[x>t|FP]}{mPr[x>t]}$$

$$\text{FDR}(t) = \frac{\pi_0 Pr[x>t|FP]}{Pr[x>t]}$$

$$\text{FDR}(t) = Pr\left[FP|x>t\right]$$

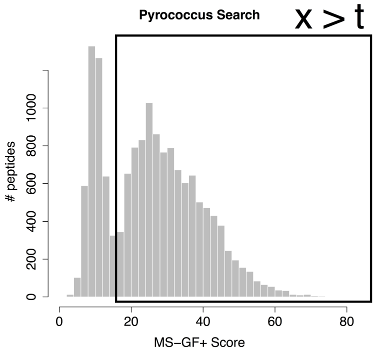
$$\text{FDR}(t) = \frac{\pi_0[1-F_0(t)]}{1-F(t)} \text{ with } F.(t) = \int\limits_{-\infty}^{t} f.(x)dx$$

How to estimate FDR?



$$\mathrm{FDR}(t) = \frac{\pi_0\,[1 - F_0(t)]}{1 - F(t)} = \frac{\pi_0 Pr[x > t|FP]}{Pr[x > t]}$$

How to estimate FDR?

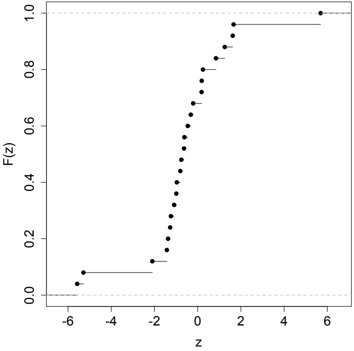


$$\text{FDR}(t) = \frac{\pi_0 \left[1 - F_0(t)\right]}{1 - F(t)} = \frac{\pi_0 Pr[x > t|FP]}{Pr[x > t]}$$

$$FDR(t) = \frac{\pi_0 \left[1 - F_0(t)\right]}{1 - \frac{\#x \leq t}{m}} = \frac{\pi_0 Pr[x > t|FP]}{\frac{\#x > t}{m}}$$

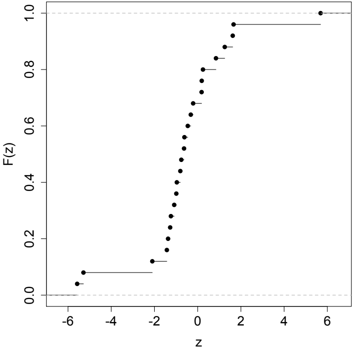
How to estimate FDR?

- $F(t)$ using the Empirical cumulative distribution function (ECDF)



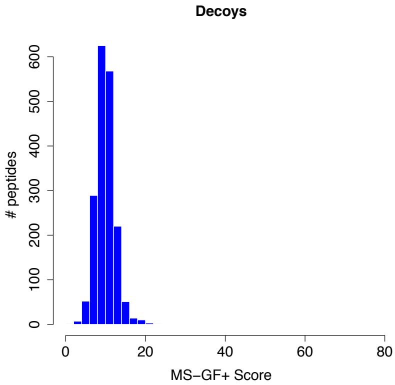
How to estimate FDR?

- $F(t)$ using the Empirical cumulative distribution function (ECDF)



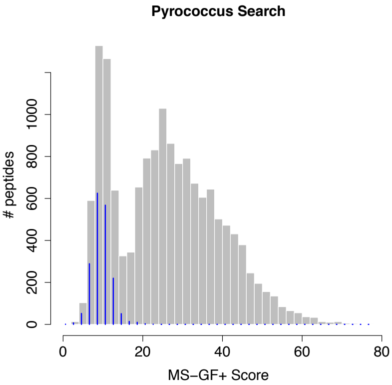
- How to characterize $F_0(t)$ and π_0 in proteomics?

Target-Decoy approach to establish null distribution



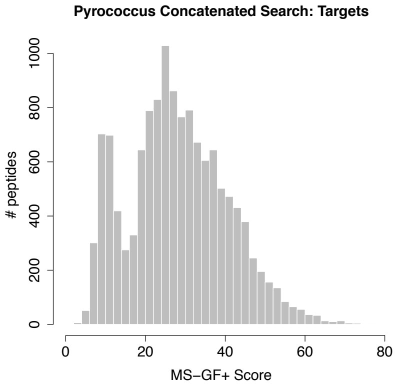
- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice

Target-Decoy approach to establish null distribution



- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice
- Concatenated search

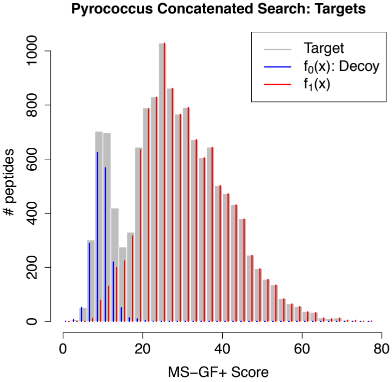
Target-Decoy approach to establish null distribution



- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice
- Concatenated search
- Assumption that bad hits have an equal probability to map on forward (target) and reverse database (decoy)

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

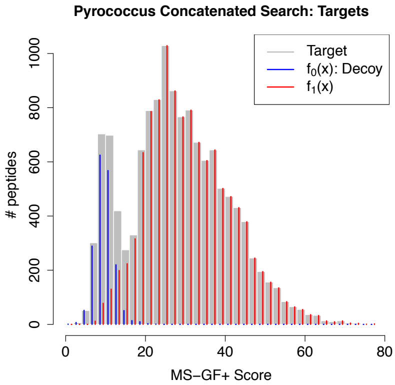
Target-Decoy approach to establish null distribution



- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice
- Concatenated search
- Assumption that bad hits have an equal probability to map on forward (target) and reverse database (decoy)

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

Target-Decoy approach to establish null distribution



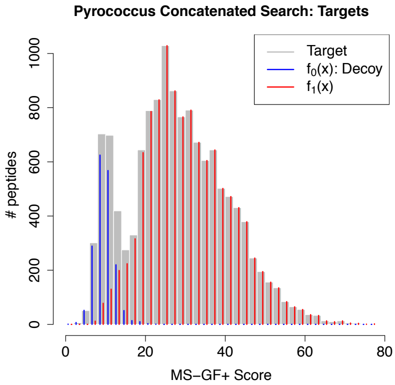
- Score cuttoff?

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

- Competitive Target - decoy:

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

Target-Decoy approach to establish null distribution



- Score cuttoff?

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

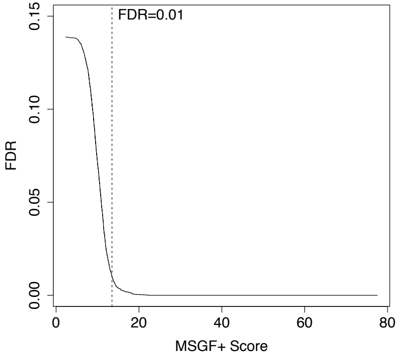
- Competitive Target - decoy:

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

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

$$\widehat{\text{FDR}}(x) = \hat{\pi}_0 \frac{1 - \bar{F}_0(x)}{1 - \bar{F}(x)}$$

Target-Decoy approach to establish null distribution



- Score cuttoff?

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

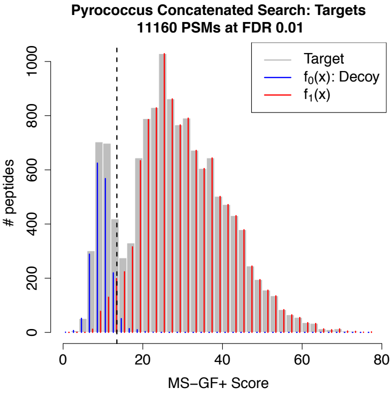
- Competitive Target - decoy:

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

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

$$\widehat{\text{FDR}}(x) = \hat{\pi}_0\frac{1-\bar{F}_0(x)}{1-\bar{F}(x)}$$

Target-Decoy approach to establish null distribution



- Score cuttoff?

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

- Competitive Target - decoy:

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

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

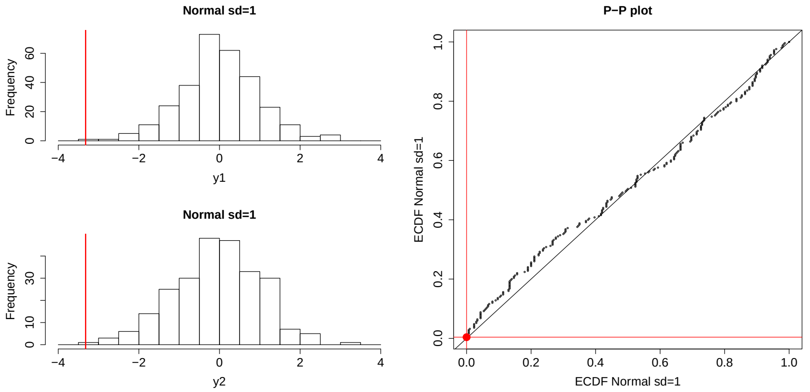
$$\widehat{\text{FDR}}(x) = \hat{\pi}_0 \frac{1 - \bar{F}_0(x)}{1 - \bar{F}(x)}$$

We have to evaluate that

- The decoys are good simulations of the targets: compare $\bar{F}_0(x)$ with $\bar{F}(x)$
- $\hat{\pi}_0 = \frac{\#decoys}{\#targets}$ is a good estimator for π_0.
- We will use Probability-Probability-plots for this purpose.
- They plot the ECDFs from two samples in function of each other.

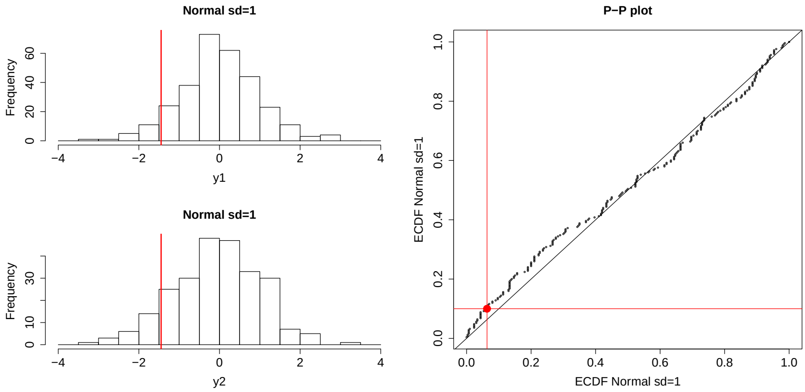
PP-plot

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



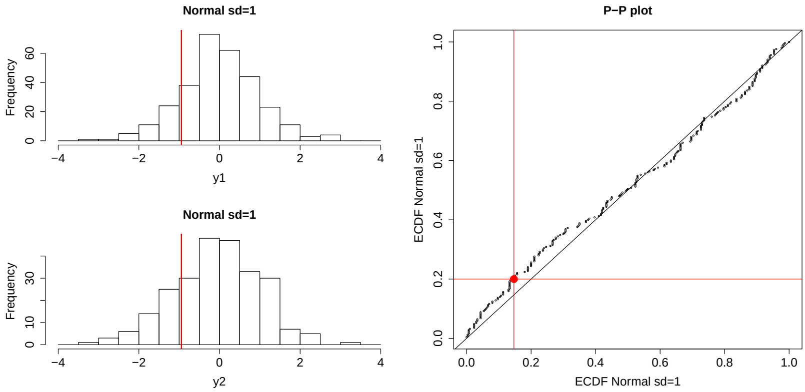
PP-plot

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



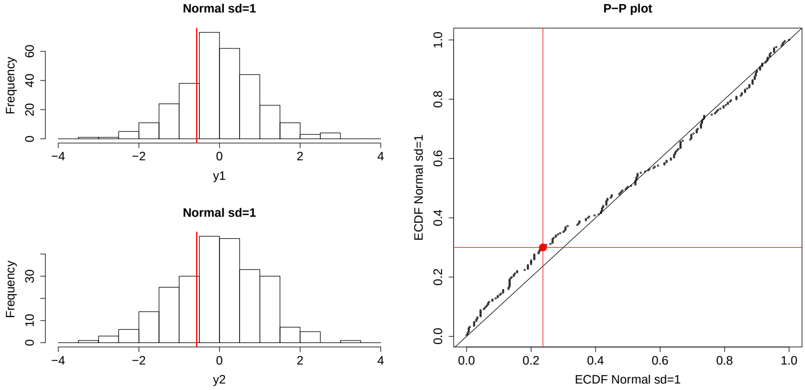
PP-plot

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



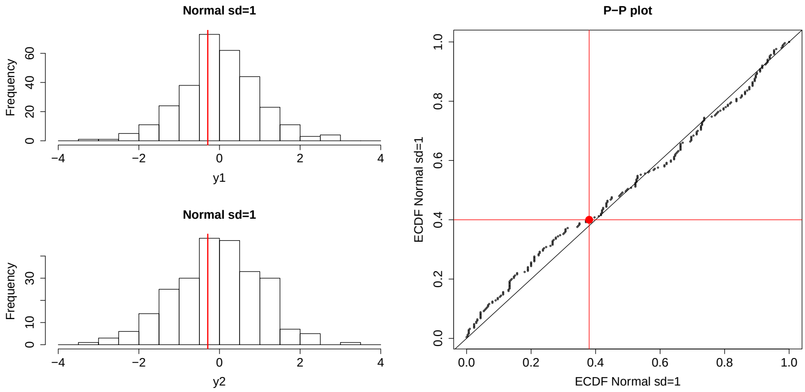
PP-plot

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



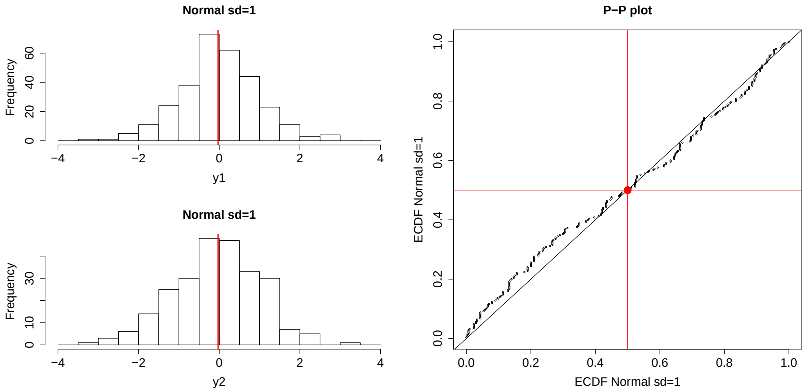
PP-plot

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



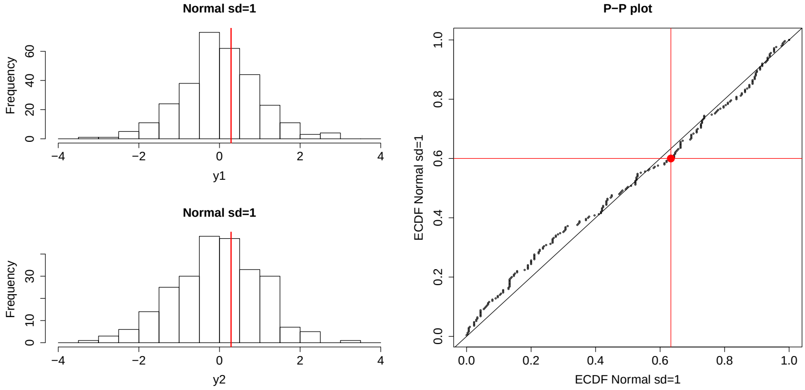
PP-plot

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



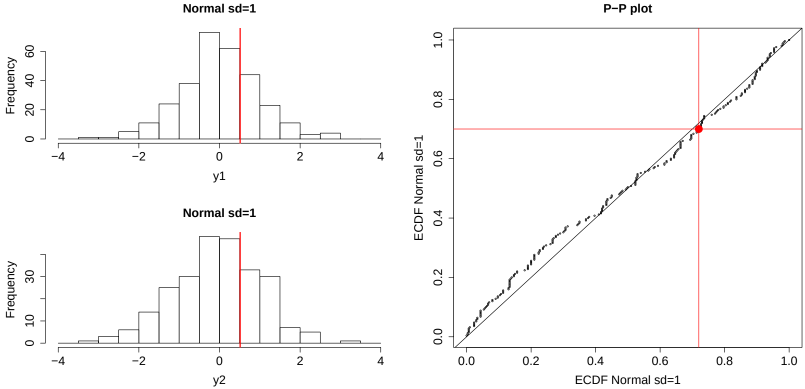
PP-plot

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



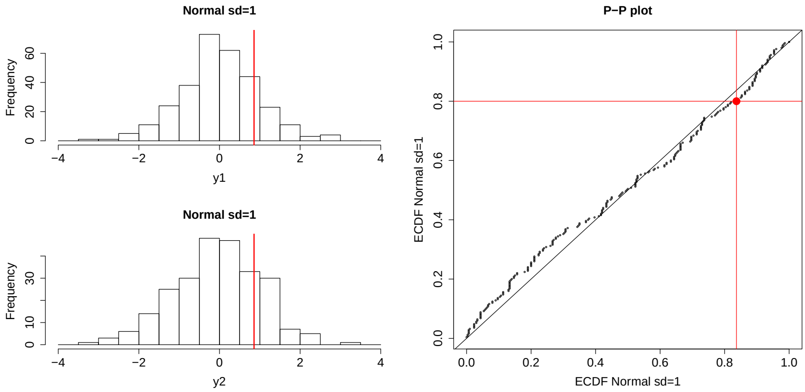
PP-plot

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



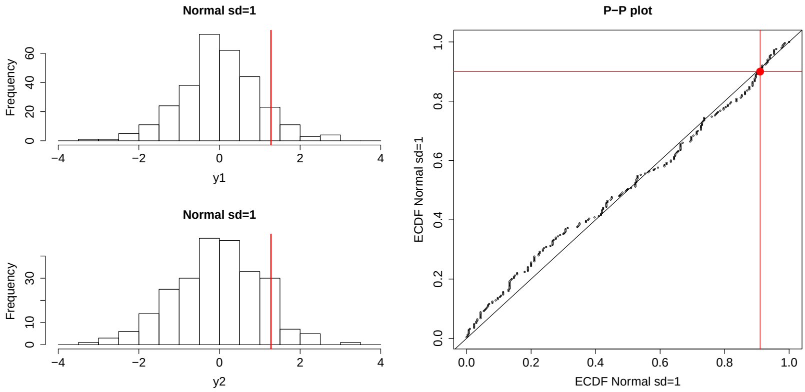
PP-plot

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



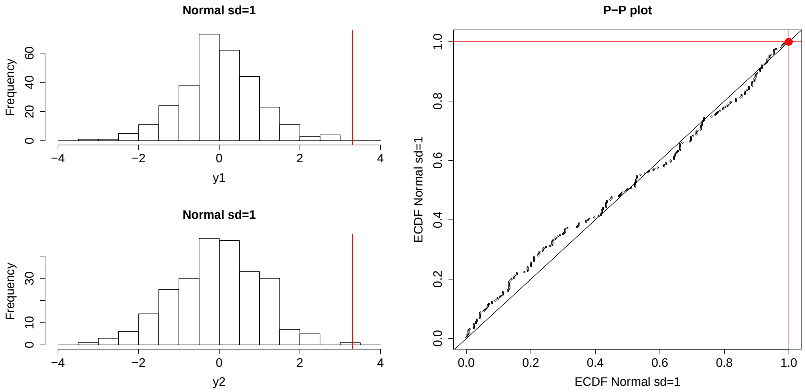
PP-plot

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

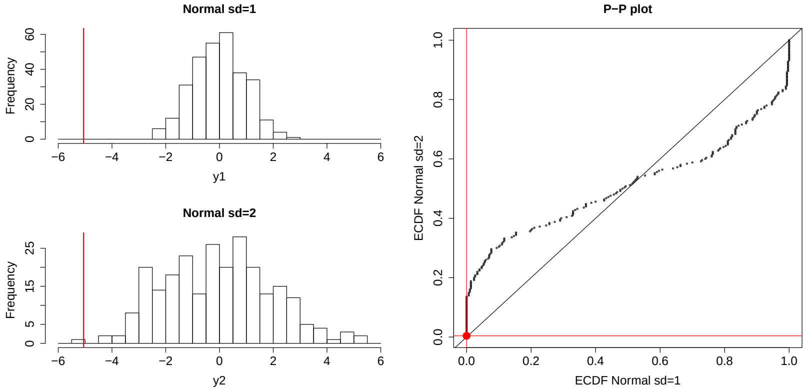


PP-plot

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

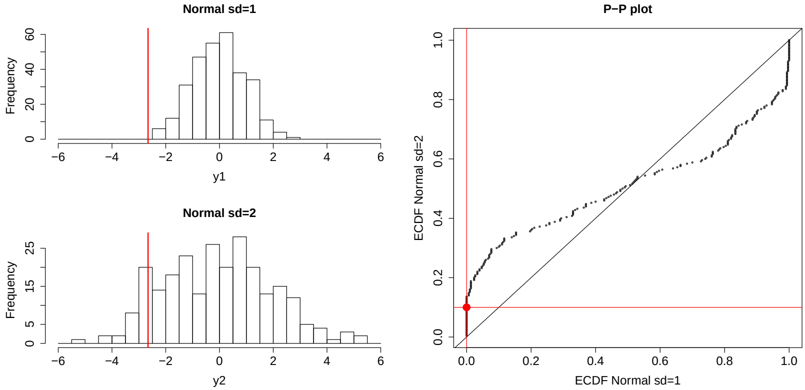


PP-plot

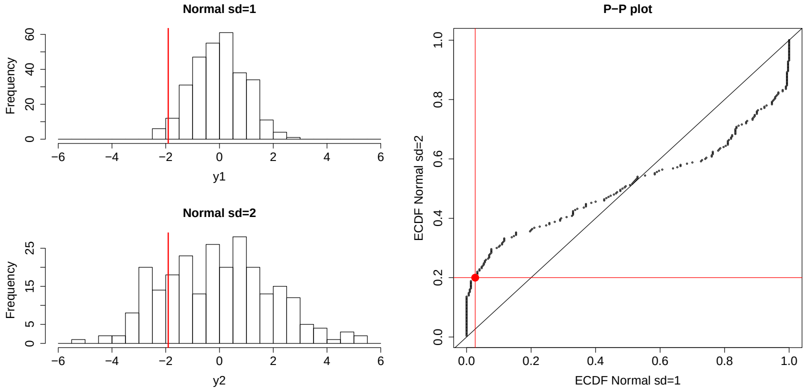


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

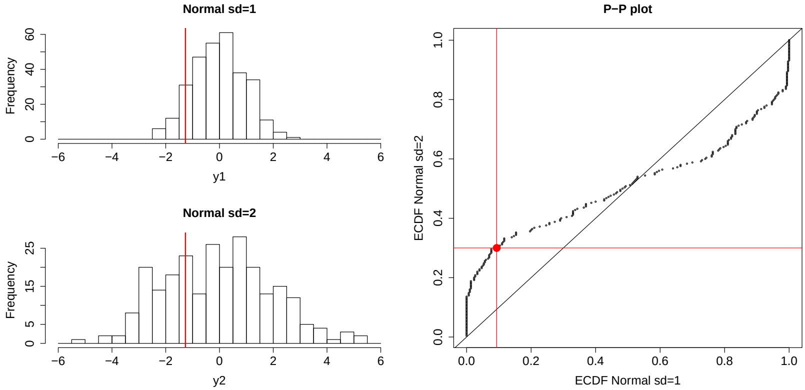
PP-plot


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

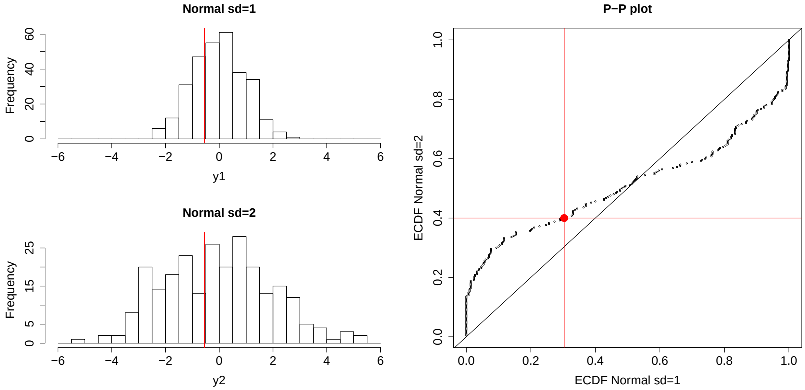
PP-plot


Normal sd=1
Frequency
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

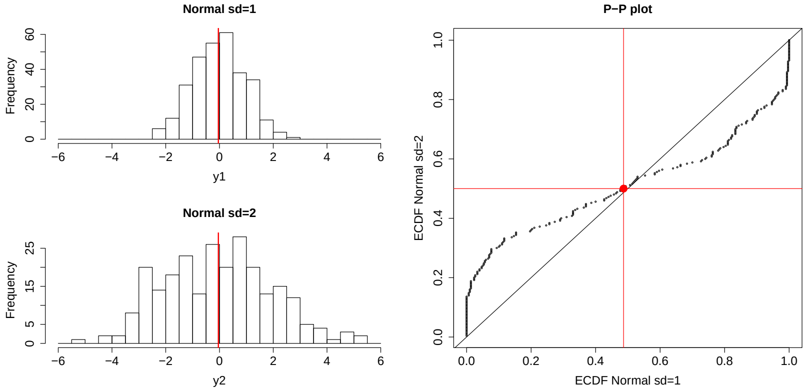

Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

PP-plot

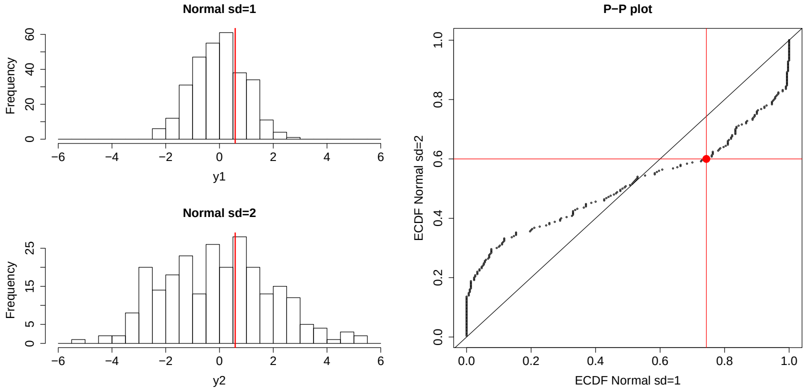


Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

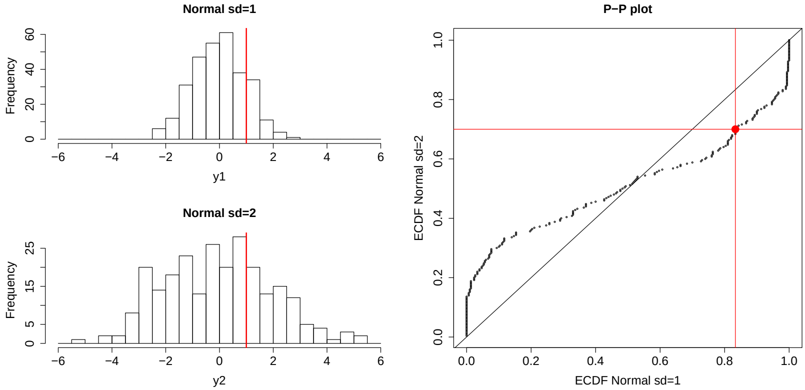
PP-plot


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
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

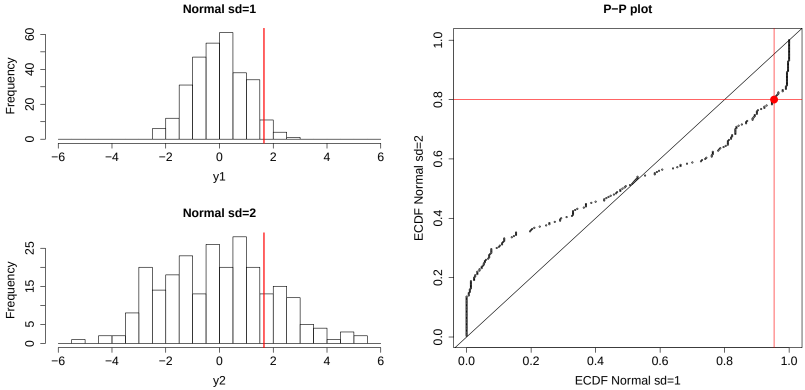
PP-plot


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

PP-plot

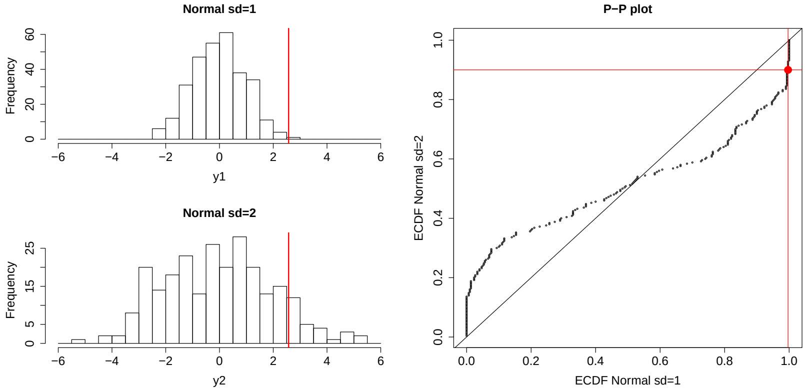

Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

PP-plot

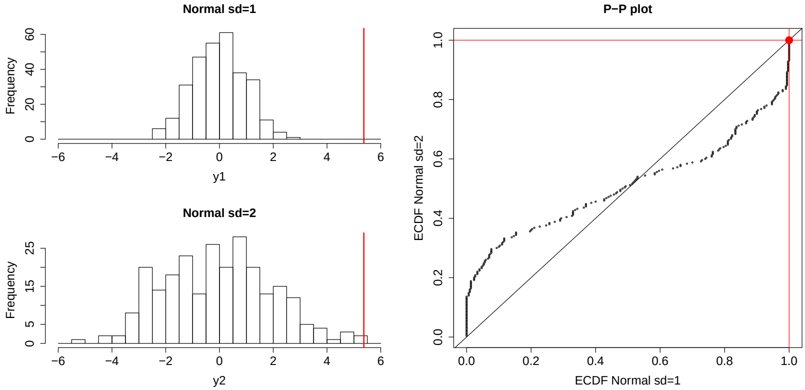


Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

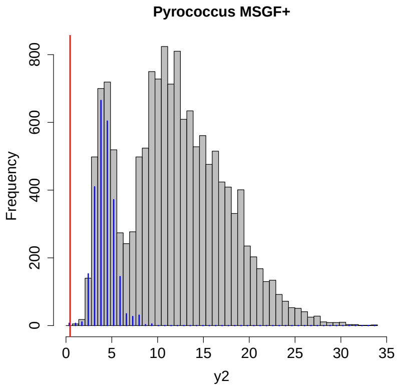
PP-plot

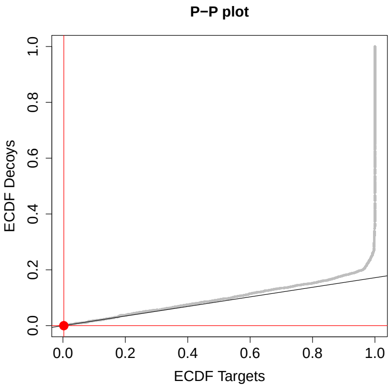

Normal sd=1
Frequency
0
20
40
60
−6
−4
−2
0
2
4
6
y1
Normal sd=2
Frequency
0
5
15
25
−6
−4
−2
0
2
4
6
y2
P−P plot
ECDF Normal sd=2
0.0
0.2
0.4
0.6
0.8
1.0
ECDF Normal sd=1

PP-plot

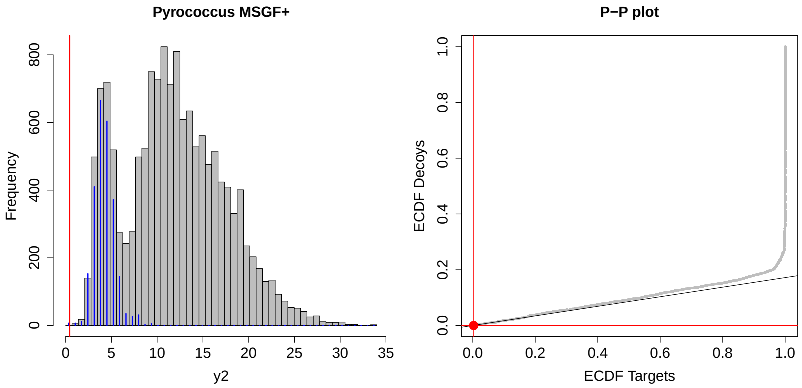

Normal sd=1
Frequency
y1
Normal sd=2
Frequency
y2
P−P plot
ECDF Normal sd=2
ECDF Normal sd=1

PP-plot: pyrococcus



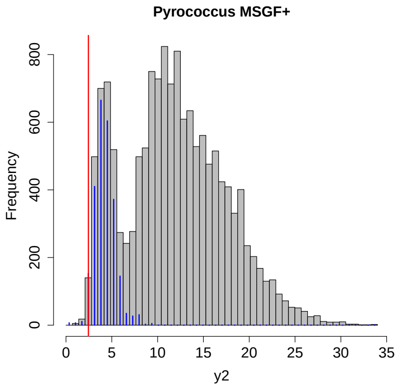


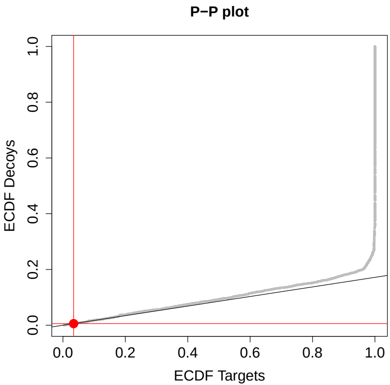
PP-plot: pyrococcus



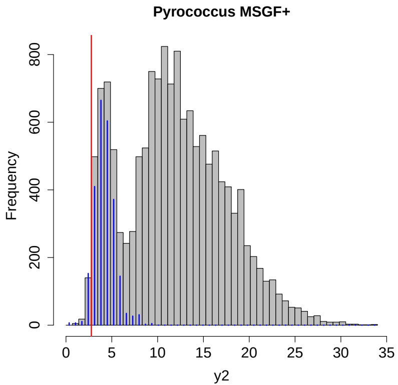
What about $\hat{\pi}_0$?

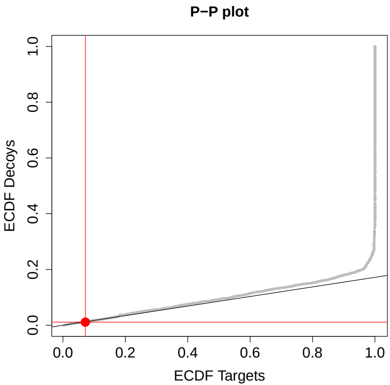
PP-plot: pyrococcus



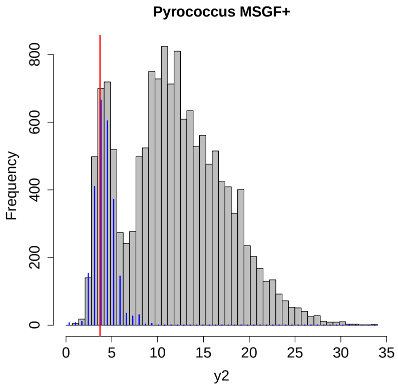


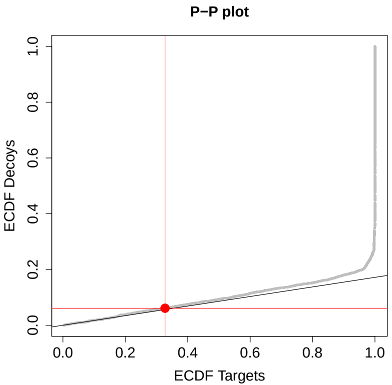
PP-plot: pyrococcus



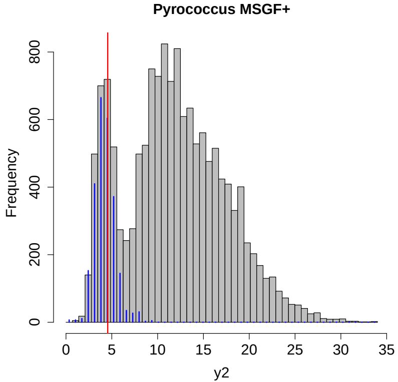


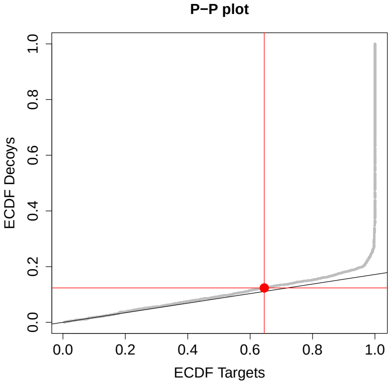
PP-plot: pyrococcus



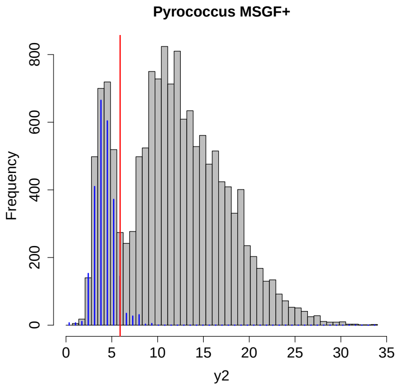


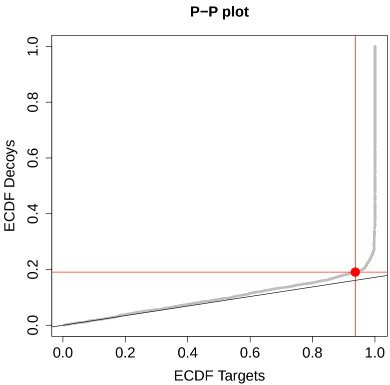
PP-plot: pyrococcus



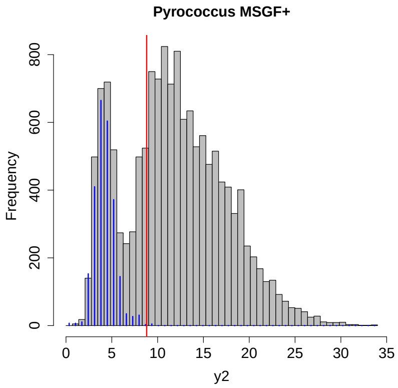


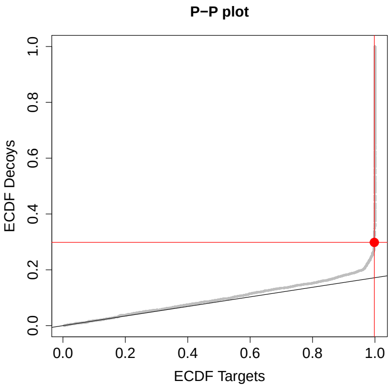
PP-plot: pyrococcus



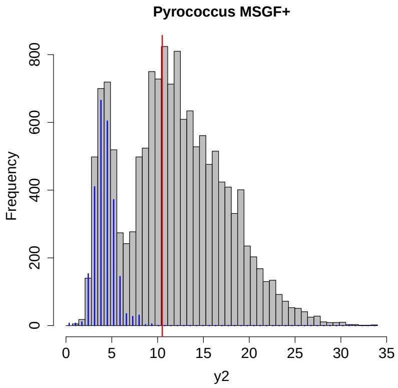


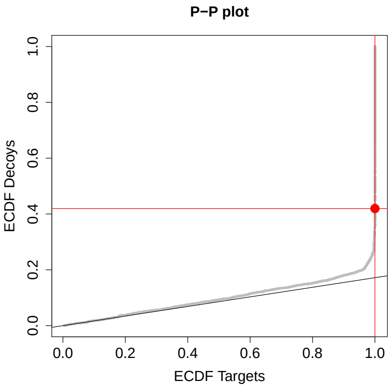
PP-plot: pyrococcus



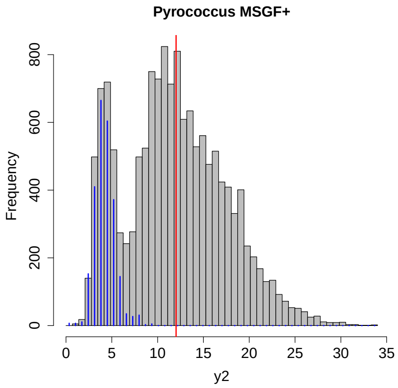


PP-plot: pyrococcus

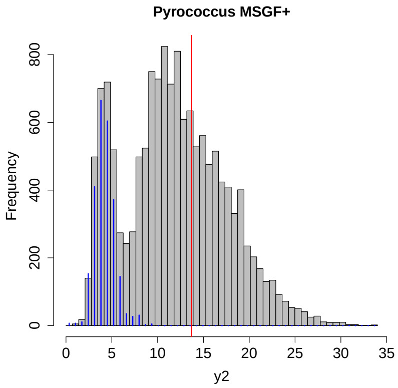




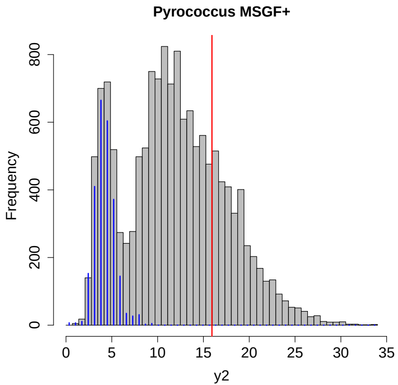
PP-plot: pyrococcus

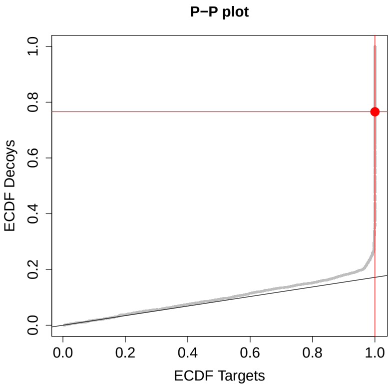


PP-plot: pyrococcus

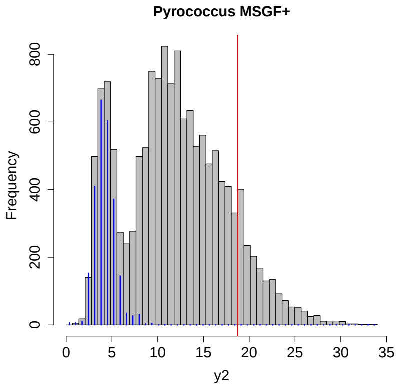


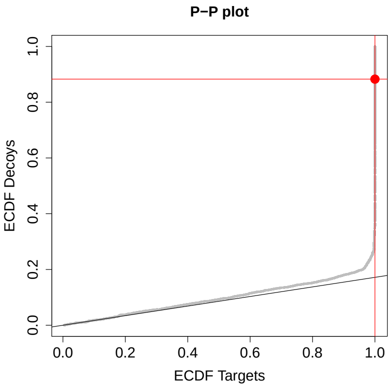
PP-plot: pyrococcus





PP-plot: pyrococcus





PP-plot: pyrococcus