

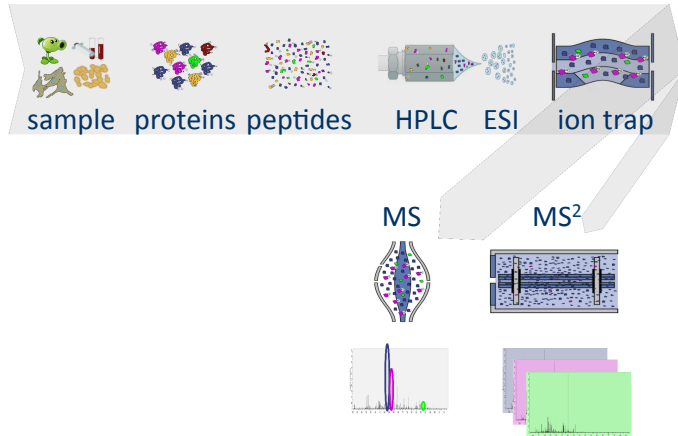
Statistical Methods for Quantitative MS-Based Proteomics:

1. Identification

Lieven Clement

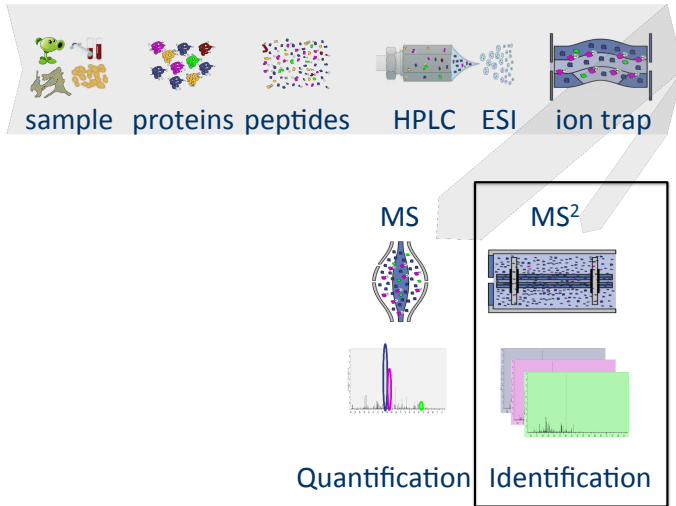
Statistics and Genomics Seminar, UC Berkeley, California

Challenges in Label Free MS-based Quantitative proteomics



Quantification Identification

Challenges in Label Free MS-based Quantitative proteomics



Identification

```

1:PI:1P10029757-51:THEHBL:Q981P7|REFSEQ_HP:HP_089479|ENCLHBL
0:0528|01:000004541|01:00000777|WEGA:01:00000007612 Tax
H0C2020
H0MPP0D0T0P0MPL0G0H0M0E0V0S0L0H0E0H0E0T0A0I0F0E0D0T0P0V0L0S0H0E0C0I0V
0T0E0A0H0G0V0V0A0H0U0H0G0H0G0V0E0H0G0V0P0A0H0D0T0E0L0G0H0L0P0H
0H0E0S0C0L0V0L0Q0E0T0E0S0M0L0L0G0H0L0L0S0F0S0L0T0Q0I0P0G0K0H0M0L0L0H
0L0A0L0H0S0M0L0L0A0L0H0C0V0A0Q0I0H0F0T0H
1:PI:1P10029757-51:THEHBL:Q981P7|REFSEQ_HP:HP_089479|ENCLHBL
0:0528|01:000004541|01:00000777|WEGA:01:00000007612 Tax
H0C2020
H0MPP0D0T0P0MPL0G0H0M0E0V0S0L0H0E0H0E0T0A0I0F0E0D0T0P0V0L0S0H0E0C0I0V
0T0E0A0H0G0V0V0A0H0U0H0G0H0G0V0E0H0G0V0P0A0H0D0T0E0L0G0H0L0P0H
0H0E0S0C0L0V0L0Q0E0T0E0S0M0L0L0G0H0L0L0S0F0S0L0T0Q0I0P0G0K0H0M0L0L0H
0L0A0L0H0S0M0L0L0A0L0H0C0V0A0Q0I0H0F0T0H
1:PI:1P10029757-51:SWISS-PROT:P04278|THEHBL:Q620H0|OSPP0:07
0:0000000000000000|0:1M0:01:00000654 Tax_0:0608 T30 kDa 3
H0H0L0S0H0L0R0A0R0P0L0S0L0L0P0G0R0H0A0S0L0P0A0R0G0P0G0L0S0A0L0V0
H0M0E0D0E0E0S0T0S0M0P0H0A0H0L0S0M0T0P0H0L0Q0H0P0T0S0L0G0
H0H0L0L0R0S0C0L0P0L0K0E0R0T0A0H0I0V0L0Q0L0A0V0V0S0H0L0K0V0L0H0E0V0S
0T0P0L0K0E0A0H0Q0M0I0V0L0I0S0V0C0D0E0S0A0I0G0H0K0L0P0T0A0M0S0L0T
0H0M0G0H0M0E0H0L0T0M0H0E0I0P0P0Y0L0L0M0H0E0D0T0M0H0L0E0K0E0L0H
H0M0L0B0I0F0S0M0G0V0P0S0E0L0E0V0C0E0R0Y0P0H0M0L0L0U0T0E0L0E0A0H0L0L0A
0S0K0S0D0S0F0S0L0M0C0T0P0H0L0T0V0K0H0M0H0S0P0L0H0L0H0L0M0T
H0H0L0M0H0M0E0S0F0P0H0Y0M0L0H0M0E0M0Q0I0C0L0H0M0L0C0M0P0Y0Y0H
0P0C0S0M0G0H0L0Q0H0C0L0S0M0S0M0L0H0E0A0H0A0V0L0S0L0K0M0L0P0L0S0I
H0C0L0L0P0H0M0H0M0E0Y0L0L0Y0D0V0Q0P0P0T0A0M0V0L0H0I0S0D0V0H0
S0H0L0Q0Y0H0L0K0H0M0I0P0H0I0V0T0H0L0S0V0M0P0L0H0H0L0E0S0A0D0Q0K0V0Y
0S0L0E0T0L0K0H0P0I0V0L0Q0L0L0C0E0H0M0L0K0M0S0M0T0G0V0A0L0H0
C0M0M0H0M0L0K0E0T0A0H0S0M0D0G0V0L0H0H0M0L0H0L0K0H0M0L0
0V0T0A0S0F0H0M0H0A0L0E0T0U0L0H0L0L0G0P0S0H0S0P0L0V0H0L0K0H0S0I
0L0H0D0C0V0P0I0H0H0L0C0L0K0E0L0L0Q0M0V0S0Q0H0M0V0L0P0L0
0M0K0A0I0C0T0E0H0M0L0H0C0M0M0H0L0E0V0L0T0L0T0E0C0H0M0H0

```

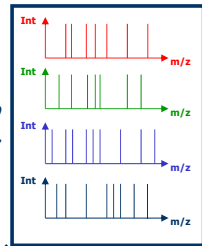
protein sequence database

in silico
digest

YSFVATAER
HETSINGK
MILQEESTVYR
SEFASTPINK
...

peptide sequences

in silico
MS/MS

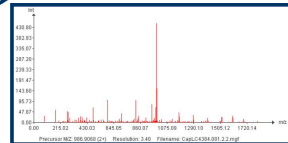


theoretical MS/MS spectra

1) YSFVATAER 34
2) YSFVSAIR 12
3) FFLIGGGGK 2

peptide scores

in silico
matching



experimental MS/MS spectrum

(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 unknown! (FDP: false discovery proportion)

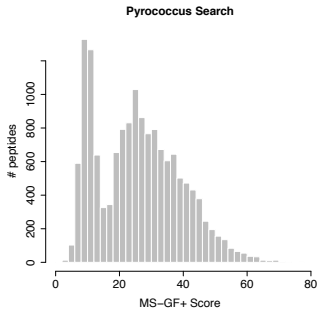
Table of Outcomes

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

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

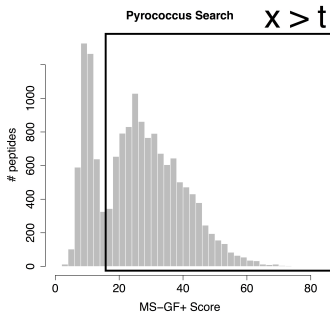
What does it mean?

Search engines return score that discriminates good from bad matches

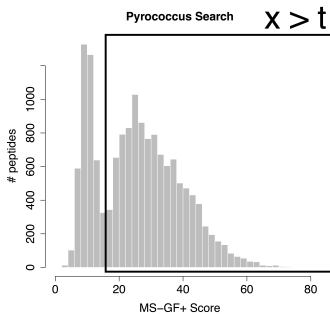


Search engines return score that discriminates good from bad matches

Score threshold t ?



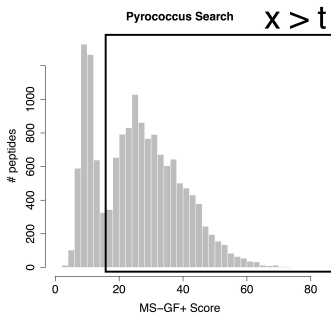
Search engines return score that discriminates good from bad matches



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

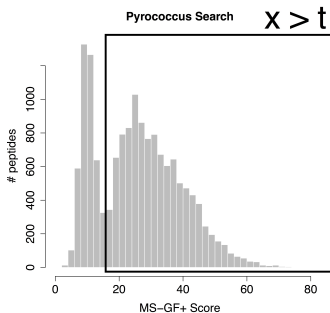


Score threshold t ?

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

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

Search engines return score that discriminates good from bad matches



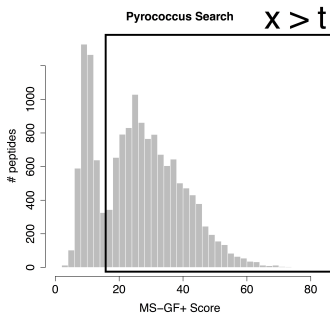
Score threshold t ?

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

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

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

Search engines return score that discriminates good from bad matches



Score threshold t ?

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

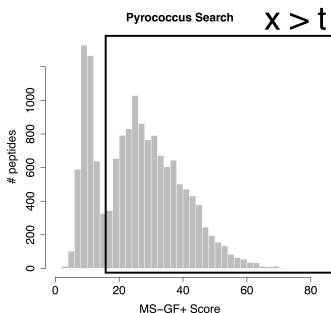
$$\text{FDR}(t) = E \left[\frac{FP}{FP + TP} \right]$$

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

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

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

Search engines return score that discriminates good from bad matches



Score threshold t ?

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

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

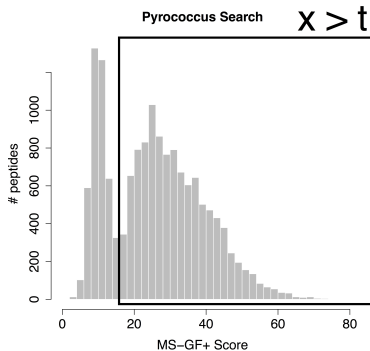
$$\text{FDR}(t) = \frac{mPr[FP]Pr[x>t|FP]}{mPr[x>t]}$$

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

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

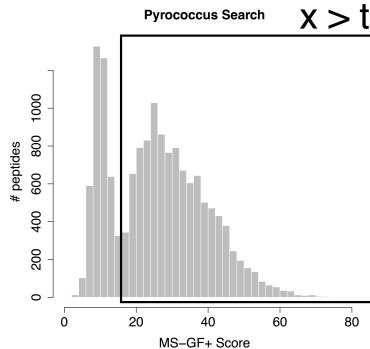
$$\text{FDR}(t) = \frac{\pi_0 [1 - F_0(t)]}{1 - F(t)} \text{ with } F.(t) = \int_{-\infty}^t f.(x) dx$$

How to estimate FDR?



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

How to estimate FDR?

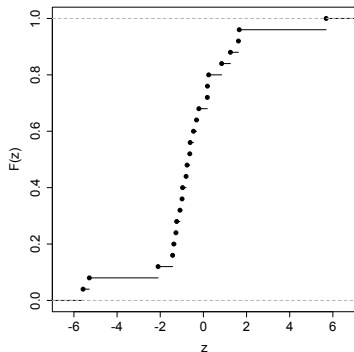


$$FDR(t) = \frac{\pi_0 [1 - F_0(t)]}{1 - F(t)} = \frac{\pi_0 Pr[x > t | FP]}{Pr[x > t]}$$

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

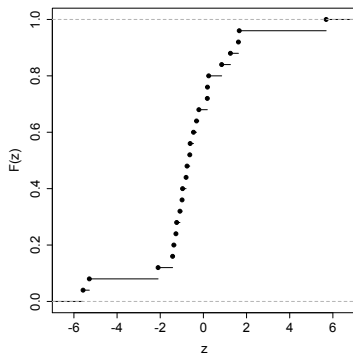
How to estimate FDR?

- $F(t)$ using the ECDF



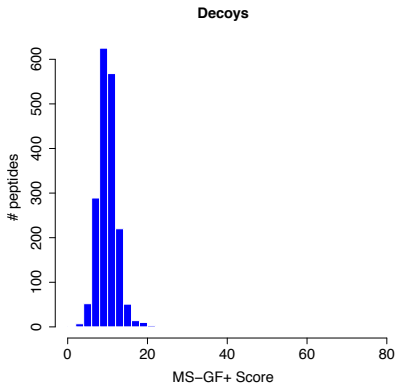
How to estimate FDR?

- $F(t)$ using the ECDF



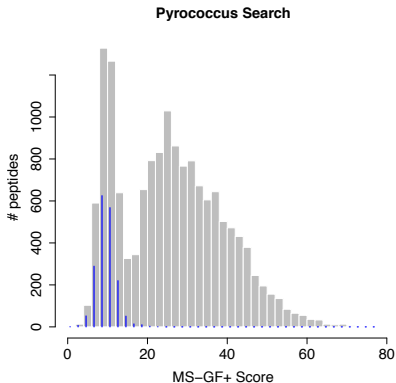
- How to characterize $F_0(t)$ and π_0 in proteomics?

Target-Decoy approach to establish null distribution



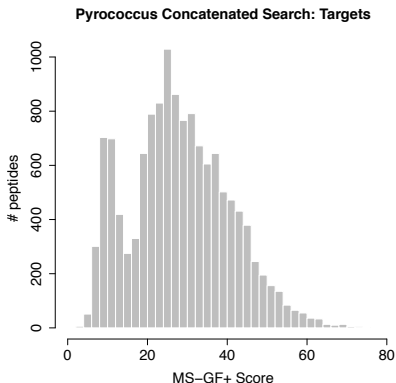
- Searching against decoy databases to generate representative bad hits
- Reversed databases are a popular choice

Target-Decoy approach to establish null distribution



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

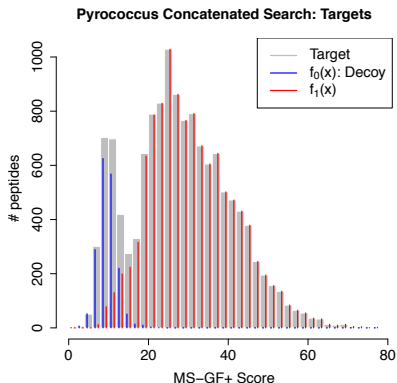
Target-Decoy approach to establish null distribution



- 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)

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

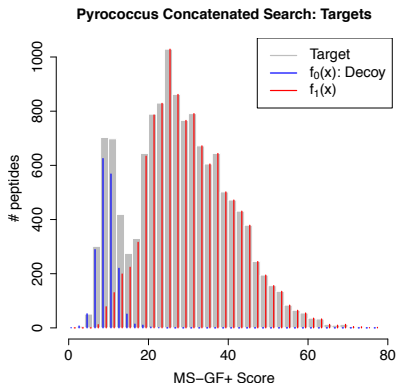
Target-Decoy approach to establish null distribution



- 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



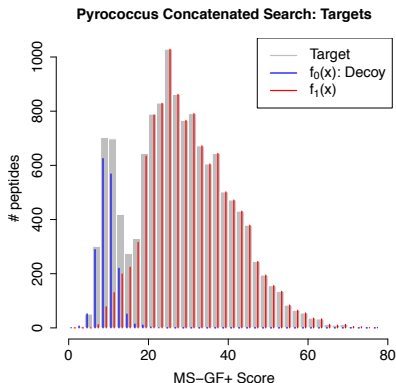
- Score cutoff?

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

- Competitive Target - decoy:

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

Target-Decoy approach to establish null distribution



- Score cutoff?

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

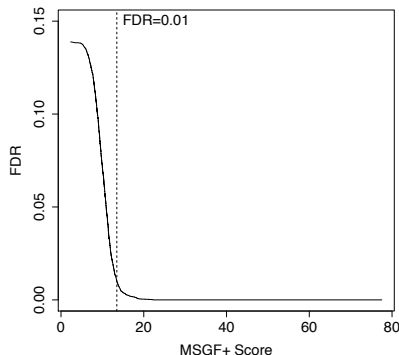
- Competitive Target - decoy:

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

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

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

Target-Decoy approach to establish null distribution



- Score cutoff?

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

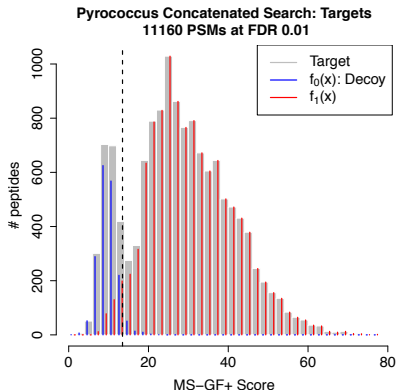
- Competitive Target - decoy:

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

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

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

Target-Decoy approach to establish null distribution



- Score cutoff?

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

- Competitive Target - decoy:

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

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

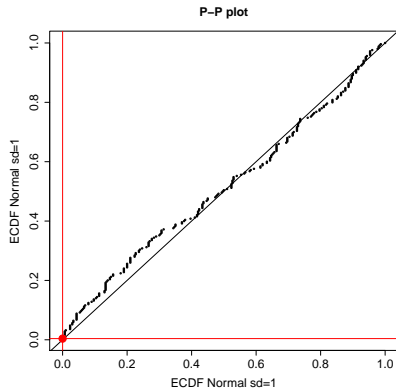
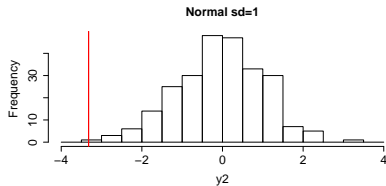
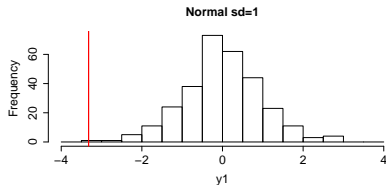
$$\widehat{\text{FDR}}(x) = \hat{\pi}_0 \frac{1 - \bar{F}_0(x)}{1 - \bar{F}(x)}$$

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.

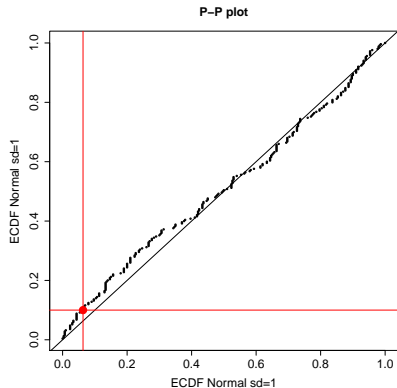
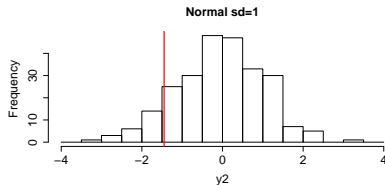
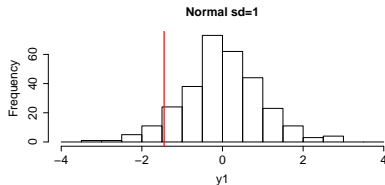
PP-plot

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



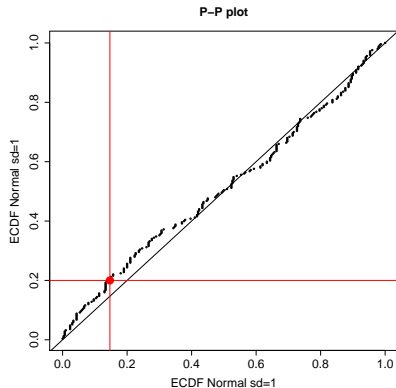
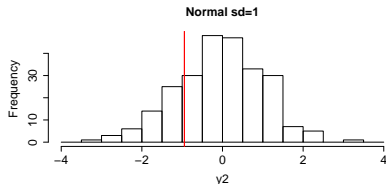
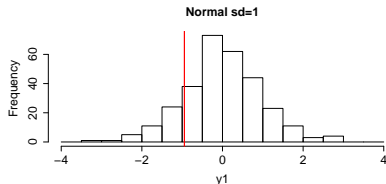
PP-plot

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



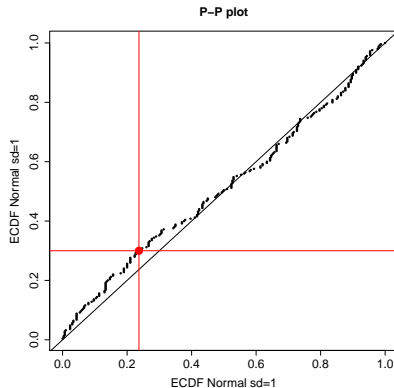
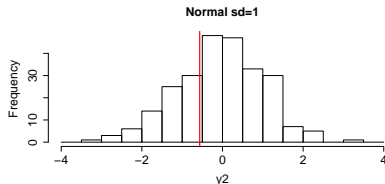
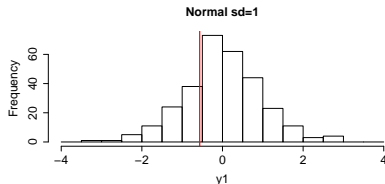
PP-plot

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



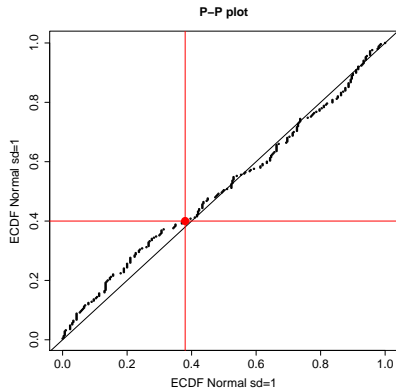
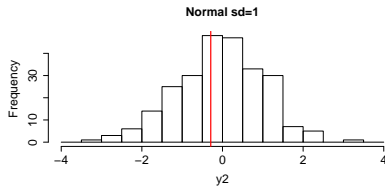
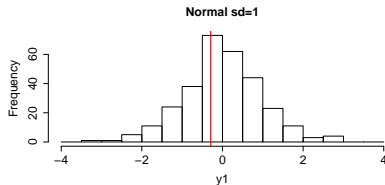
PP-plot

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



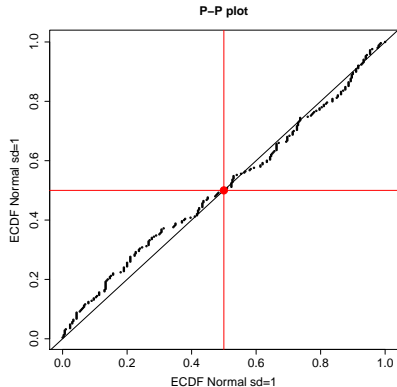
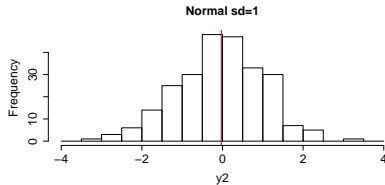
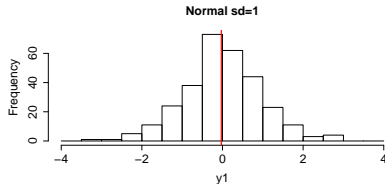
PP-plot

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



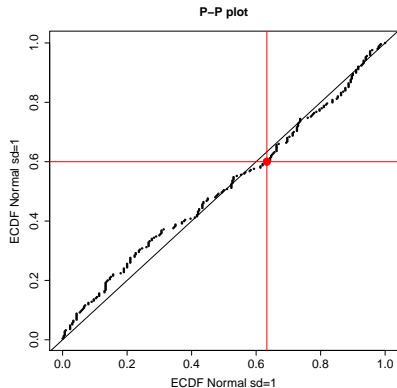
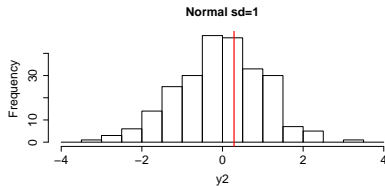
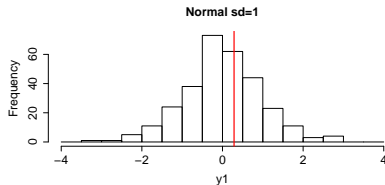
PP-plot

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



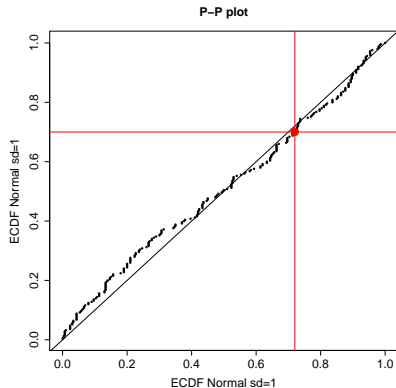
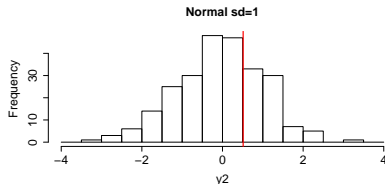
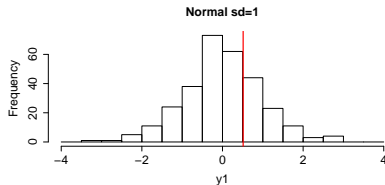
PP-plot

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



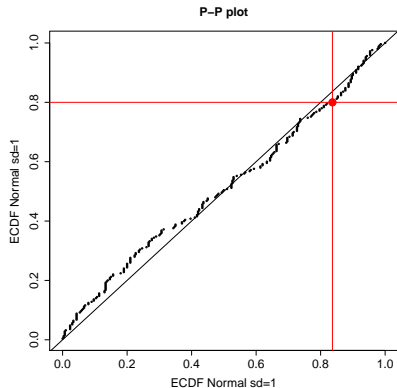
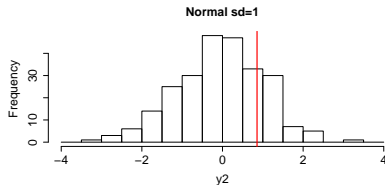
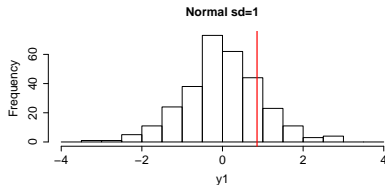
PP-plot

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



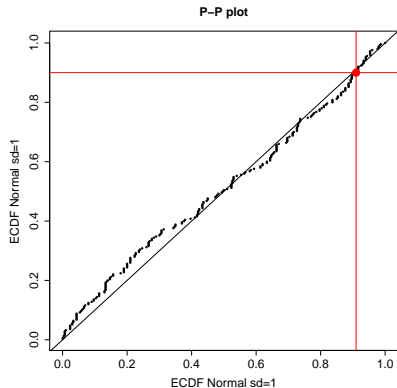
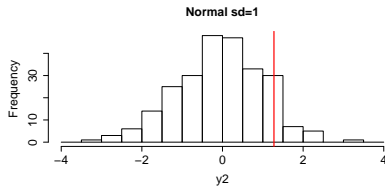
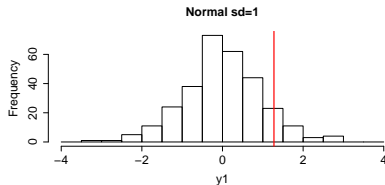
PP-plot

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



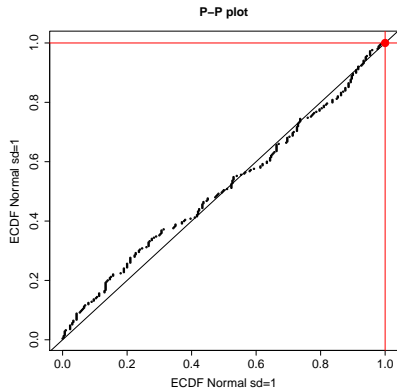
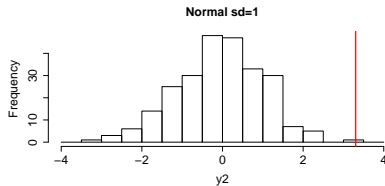
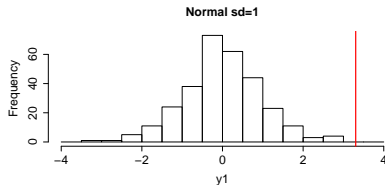
PP-plot

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

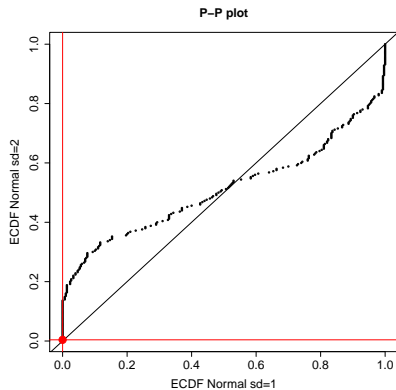
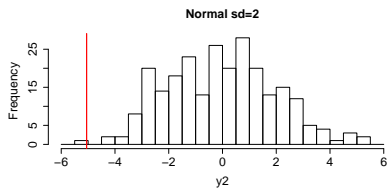
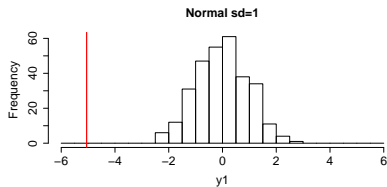


PP-plot

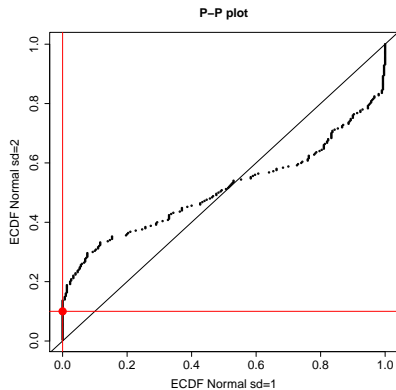
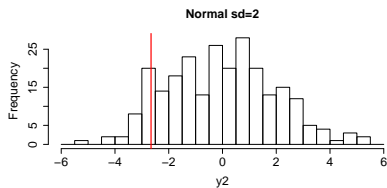
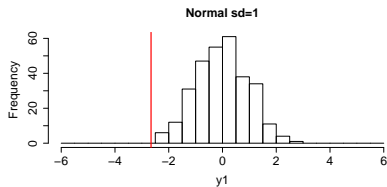
PP-plots have the property that they show a straight 45 degree line through the origin if and only if both distributions are equivalent.



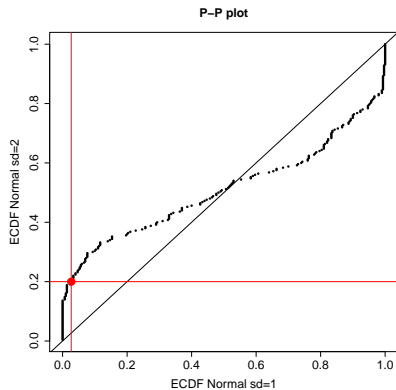
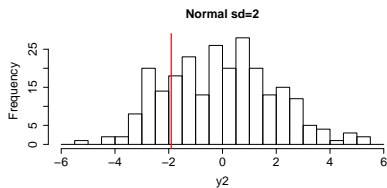
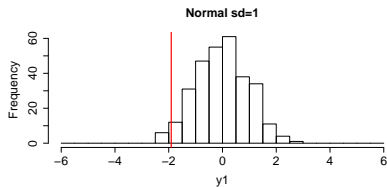
PP-plot



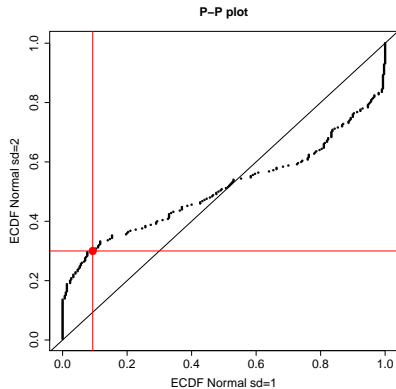
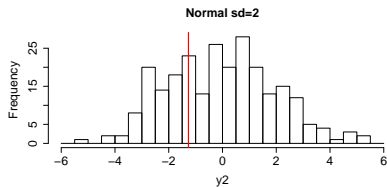
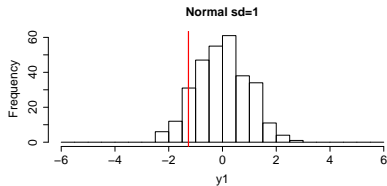
PP-plot



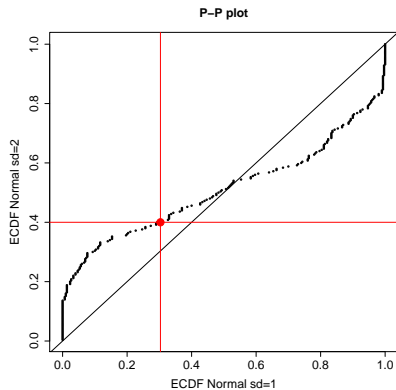
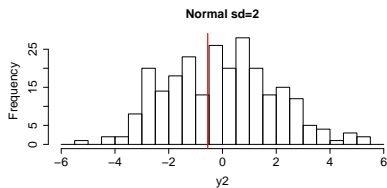
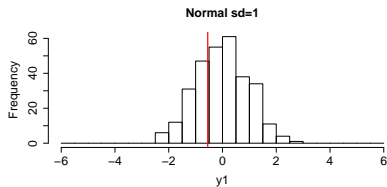
PP-plot



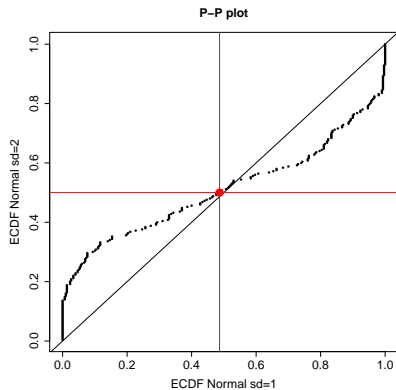
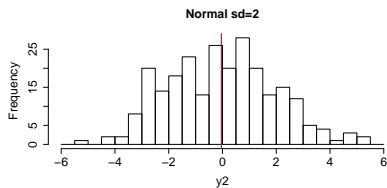
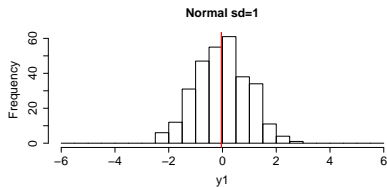
PP-plot



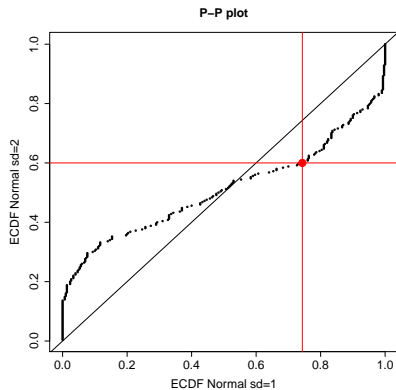
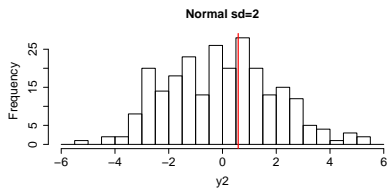
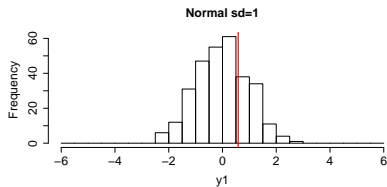
PP-plot



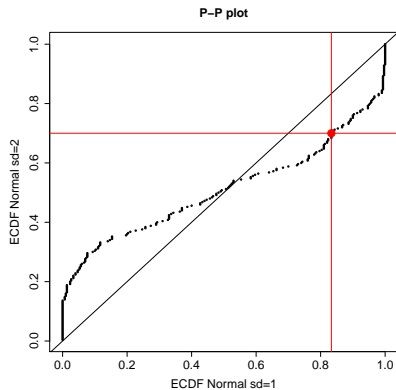
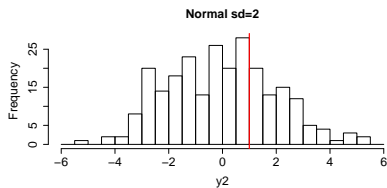
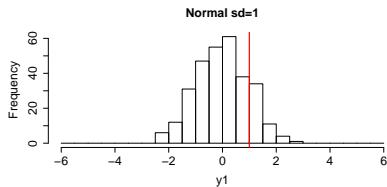
PP-plot



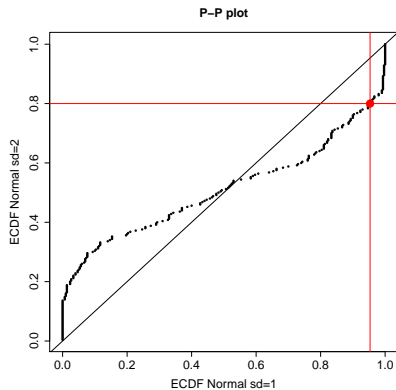
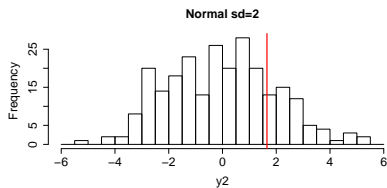
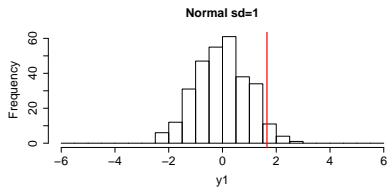
PP-plot



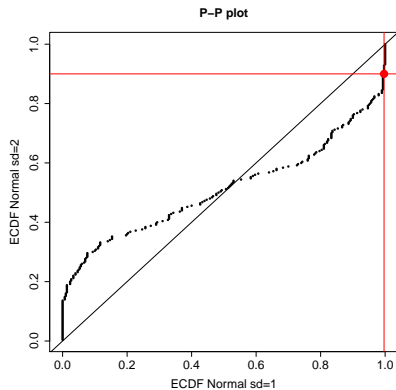
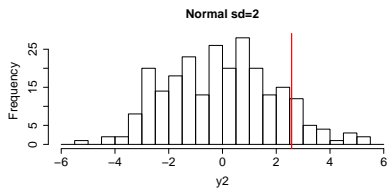
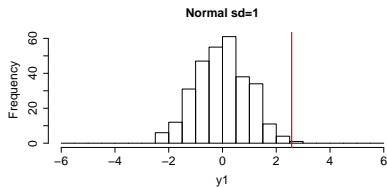
PP-plot



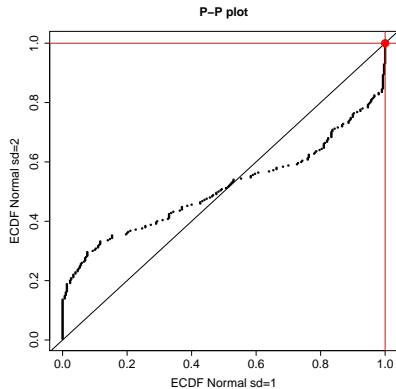
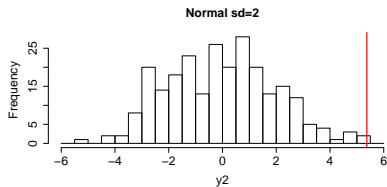
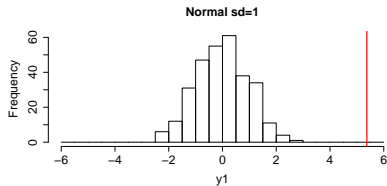
PP-plot



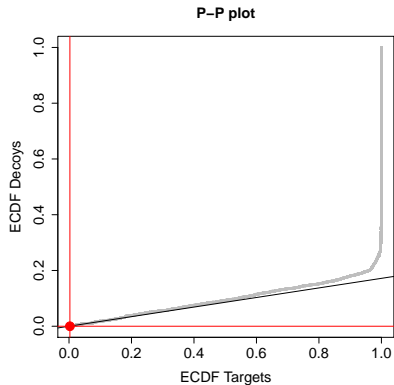
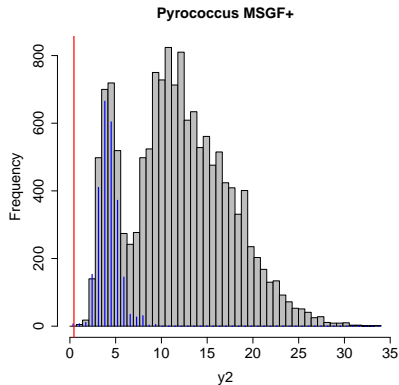
PP-plot



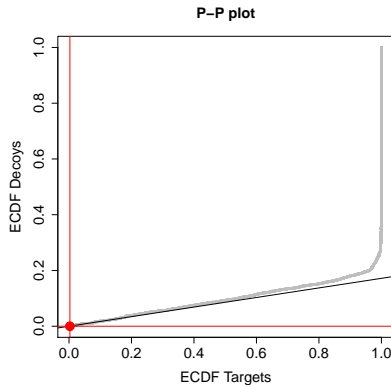
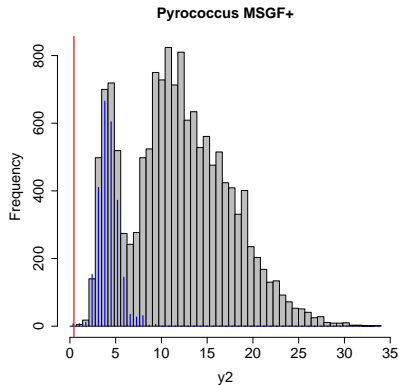
PP-plot



PP-plot: pyrococcus

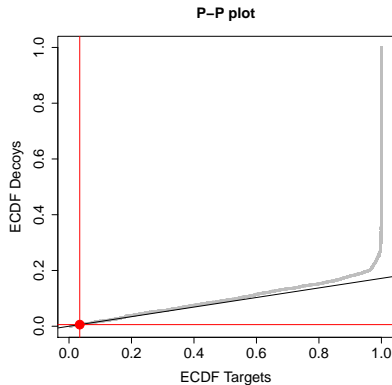
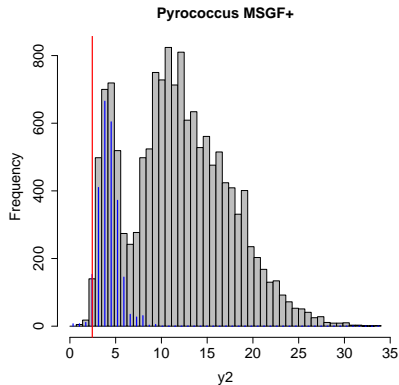


PP-plot: pyrococcus

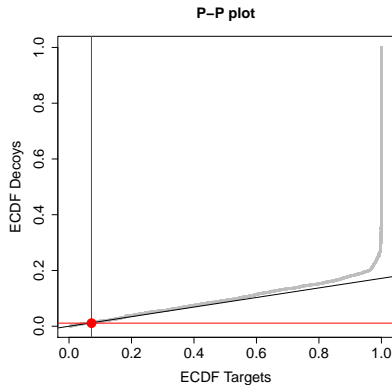
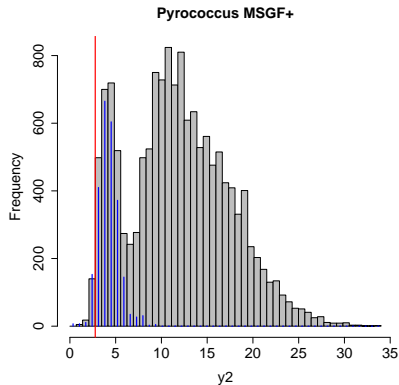


What about $\hat{\pi}_0$?

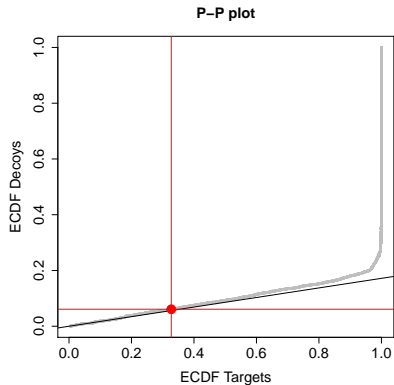
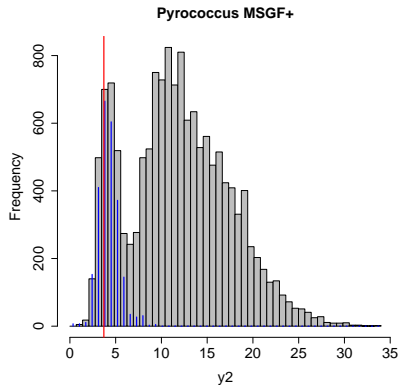
PP-plot: pyrococcus



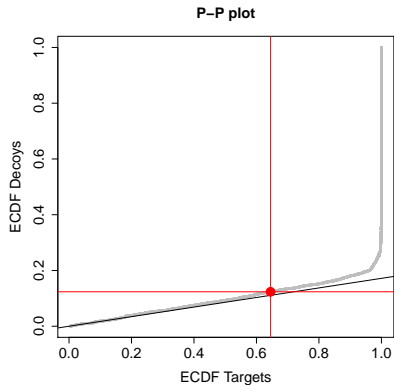
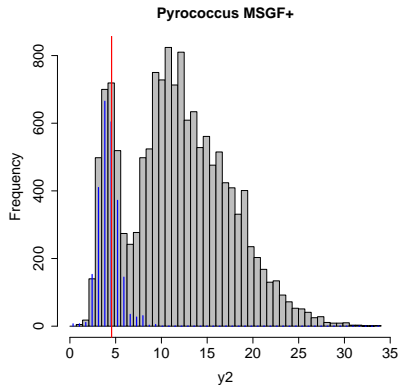
PP-plot: pyrococcus



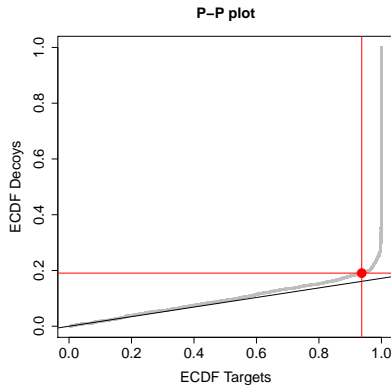
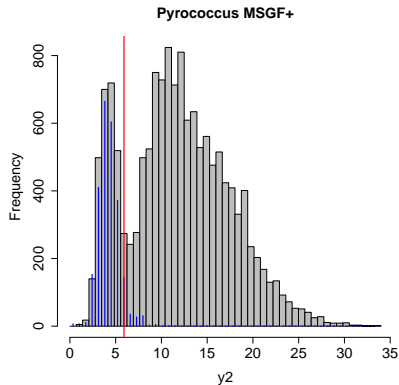
PP-plot: pyrococcus



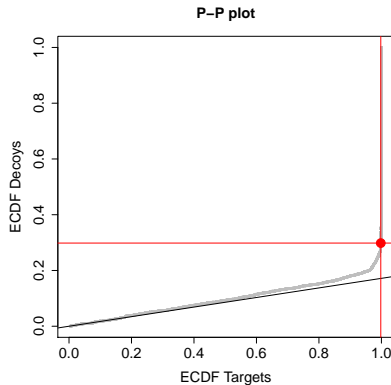
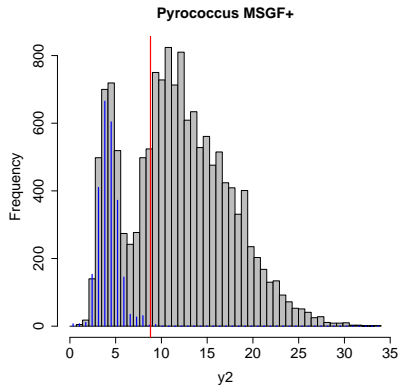
PP-plot: pyrococcus



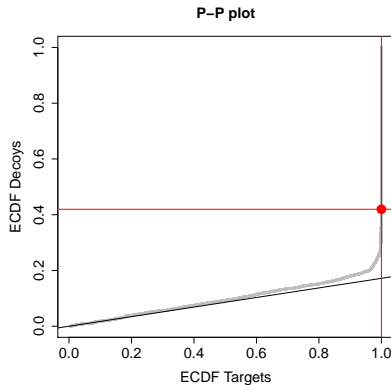
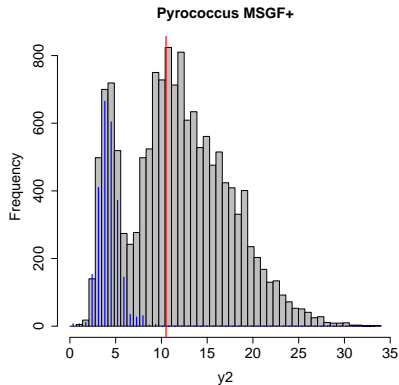
PP-plot: pyrococcus



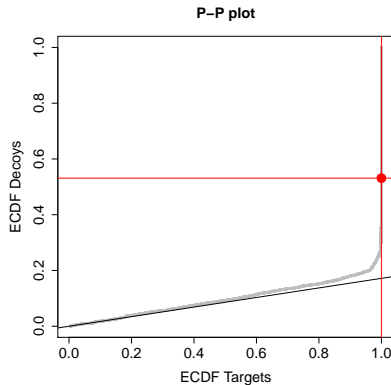
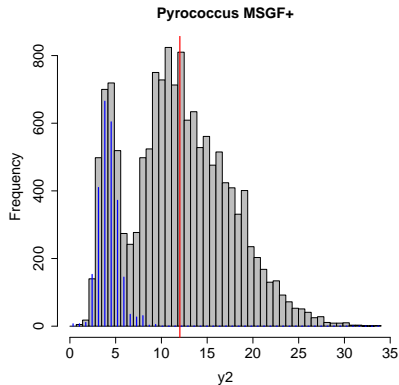
PP-plot: pyrococcus



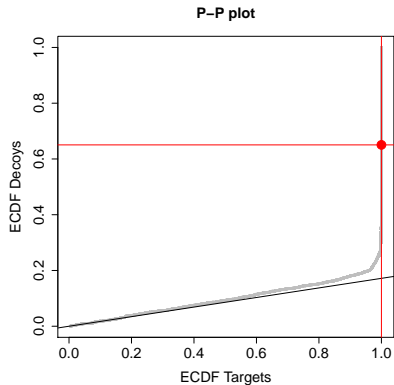
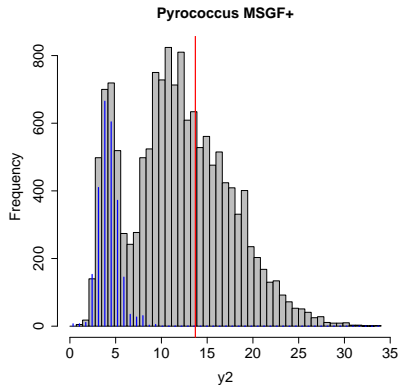
PP-plot: pyrococcus



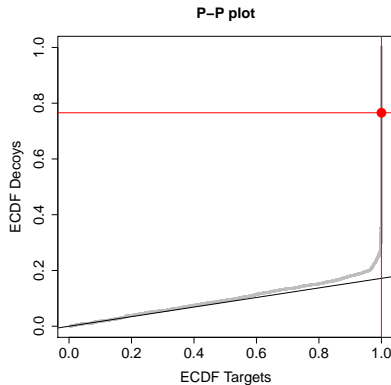
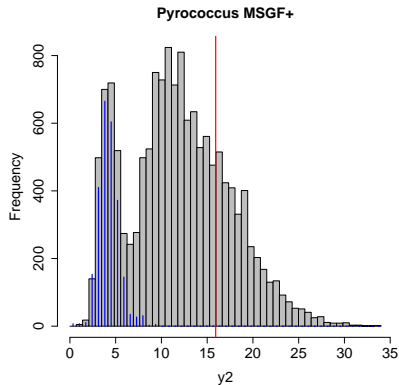
PP-plot: pyrococcus



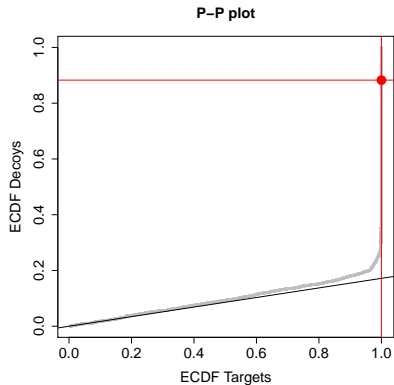
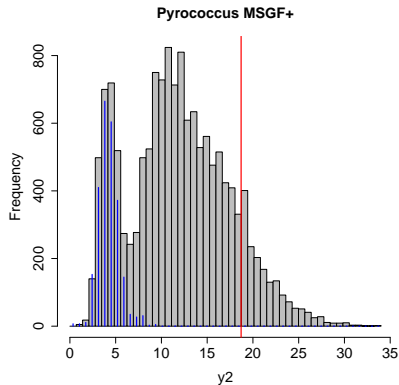
PP-plot: pyrococcus



PP-plot: pyrococcus



PP-plot: pyrococcus



PP-plot: pyrococcus

