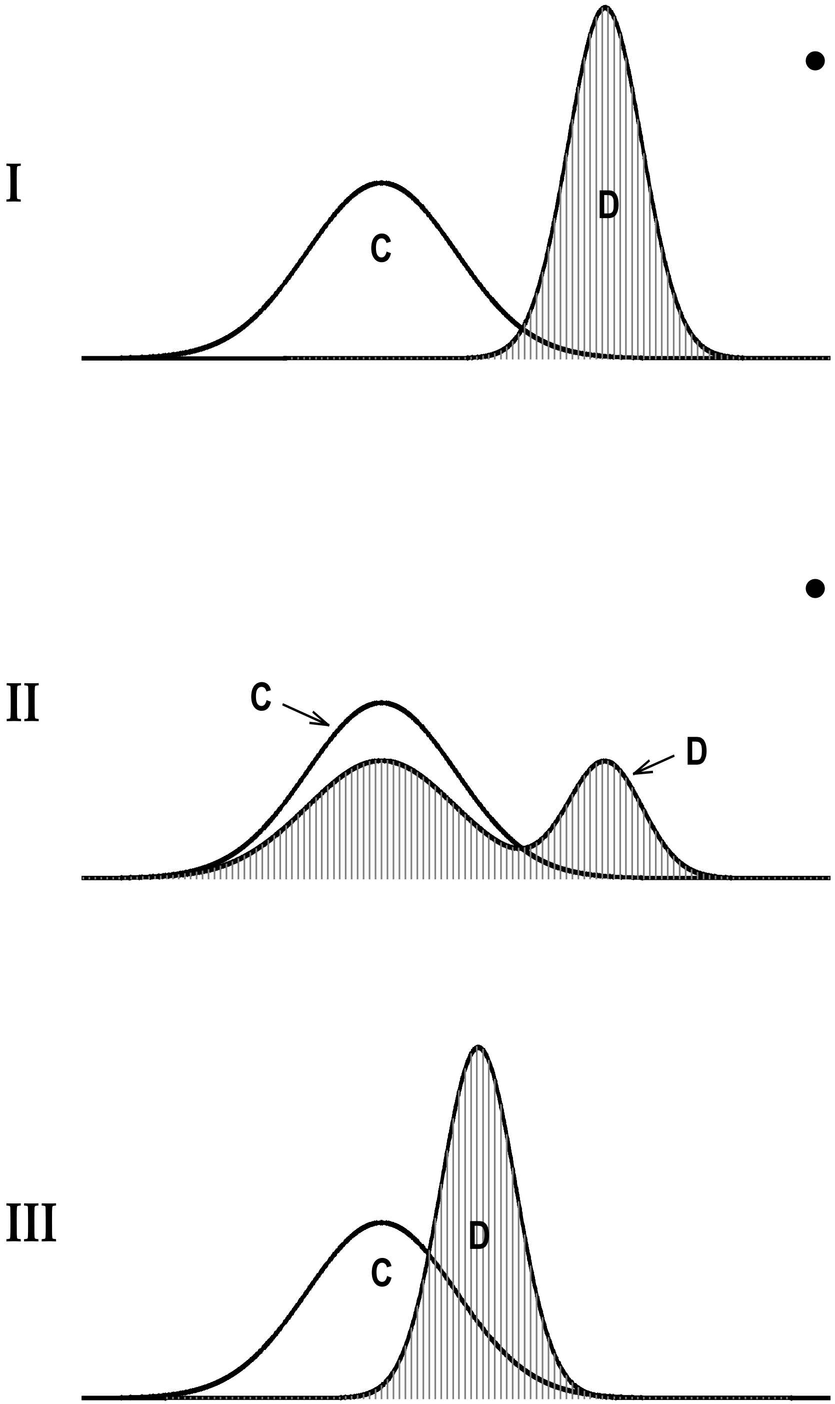
Univariable Screening by ROC curve analysis

Binary response:

- rank genes according to their differential expression between control sample C and target sample D

- use summary measures based on Receiver Operating Characteristic (ROC) curves as described by Pepe et al., Biometrics 2003.



- **Panel I:**

  Almost complete separation between the distributions of controls (C) and disease (D).

  Classify with almost 100% accuracy.

- **Panels II and III:**

  Overlapping distributions.

  Cancer screening:

  Panel II is of more practical interest than panel III.

  Panel II: clearly distinguishes a subset of D from C

  Panel III: values for D are entirely within the range of those for C.

(Pepe et al., Biometrics 2003)

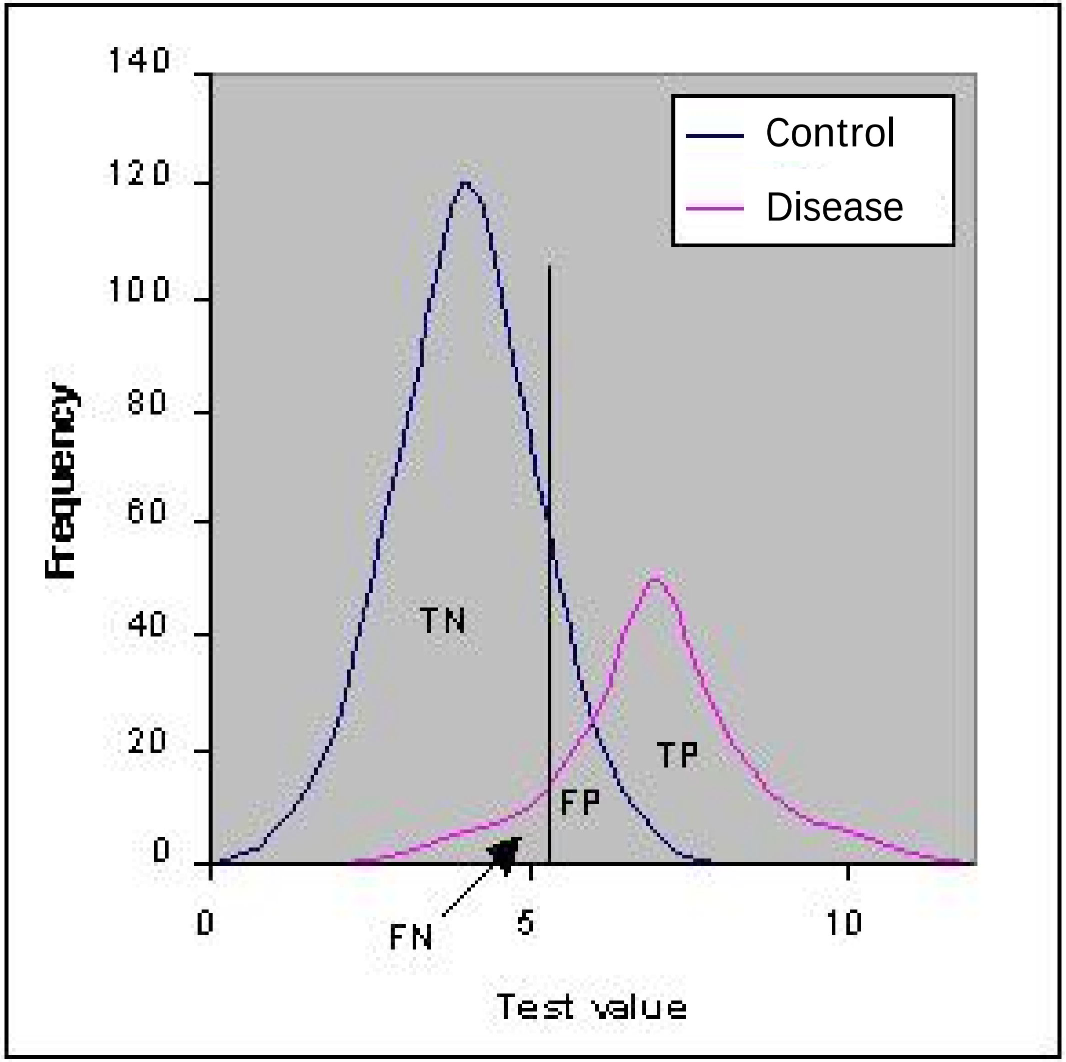
- **Notes**:

  - Panels I and III correspond to assumptions of standard two-sample tests of a location alternative like the t test or the Mann-Whitney test.

  - Panel II correspond to situations where the observed target distribution is a mixture of the distributions of controls (C) and disease (D).

    This result if e.g. no false positive diagnosis is possible (a diagnosed “D” is always a true disease), but that a negative diagnosis (i.e. “C”) is really a diseased case which simply could not be identified.

    Example: Diagnostic system with 100% sensitivity but only 50% specificity



TN: true negative (specificity)
FP: false positive (1-spec.)
FN: false negative (1-sens.)
TP: true positive (sensitivity)

	Null hypothesis H_0 true	Null hypothesis H_0 false
H_0 rejected	FP (α)	TP ($1-\beta$)
H_0 accepted	TN	FN

Gene screening by ROC analysis

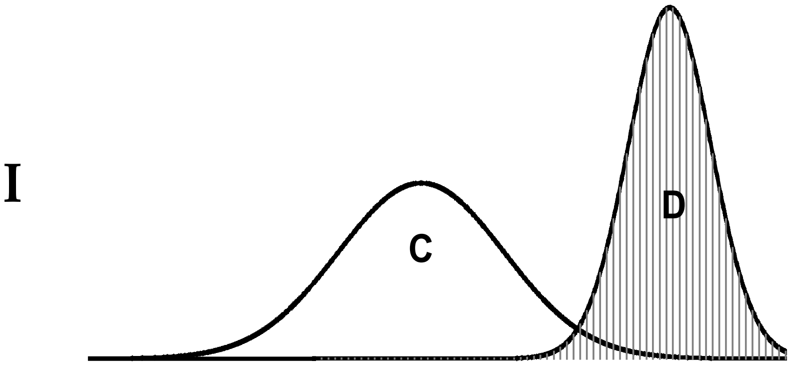
Let Y_g^i denote the relative expression level of gene g in sample i=C,D after normalization.

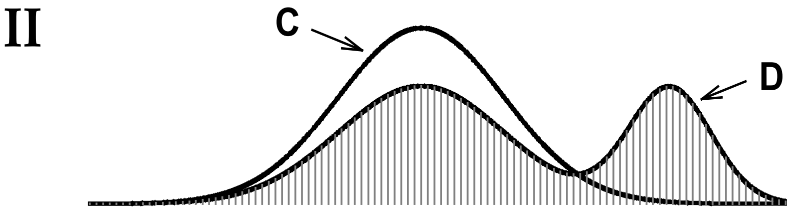
Each point on the ROC- curve, $\{t, ROC(t)\}$, corresponds to a different gene expression level u with

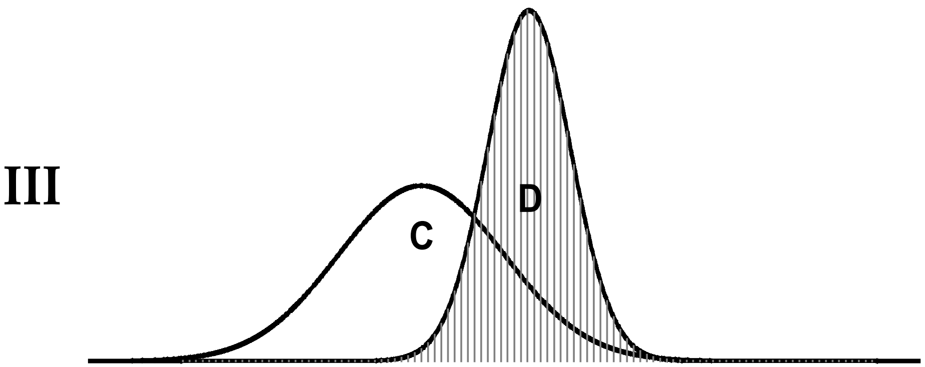
$$t = 1 - P[Y_g^C < u]$$ (1-specificity/false positive)

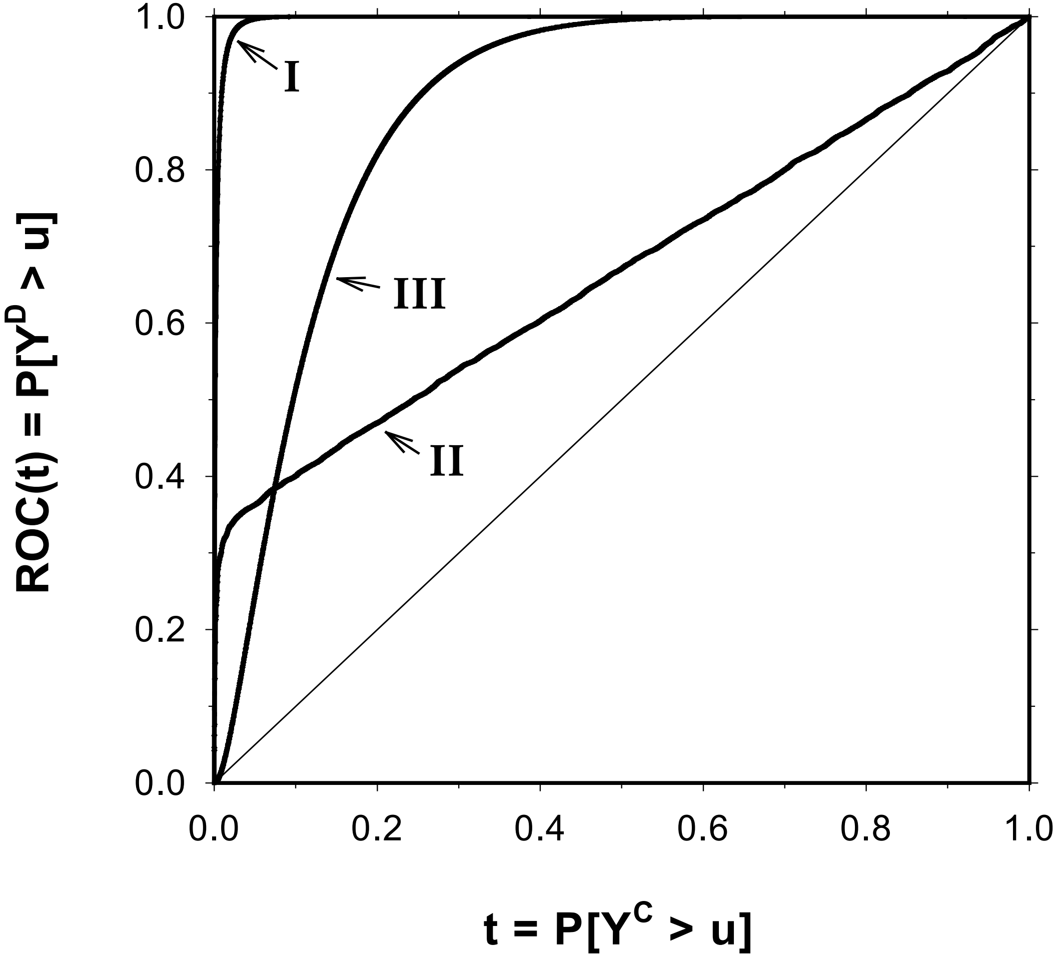
and

$$ROC(t) = P[Y_g^D \geq u]$$ (sensitivity/true positive).







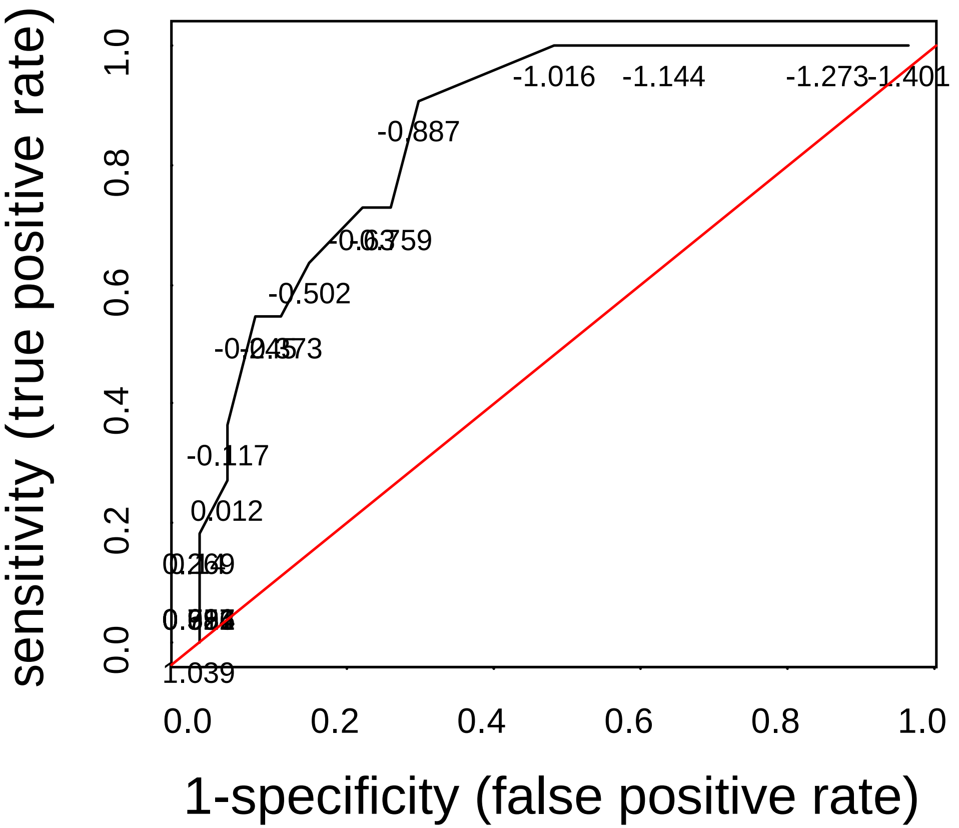


Resulting ROC curves for panels I- III

Pepe et al., Biometrics 2003

- ROC curve:

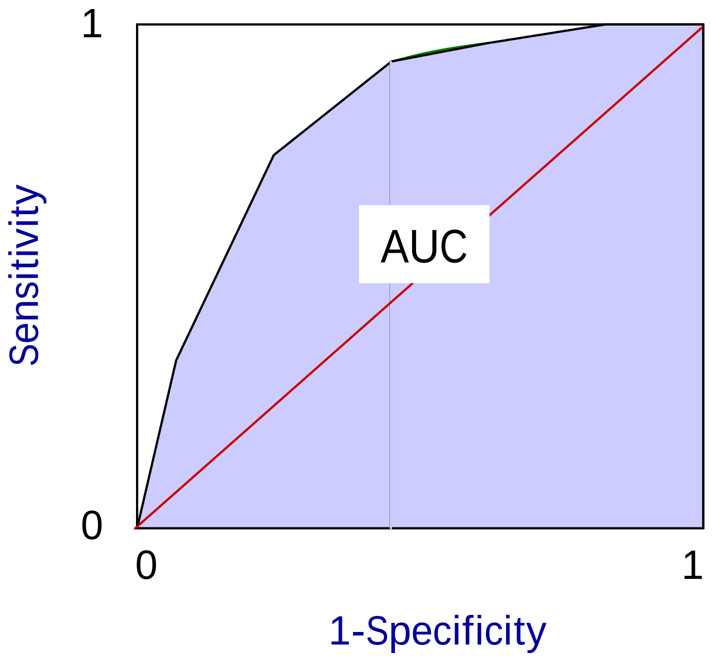
plot of the true versus false positive rates associated with all possible expression level cutpoints for classifying a sample as belonging to the target sample D based on the values of Y_g.

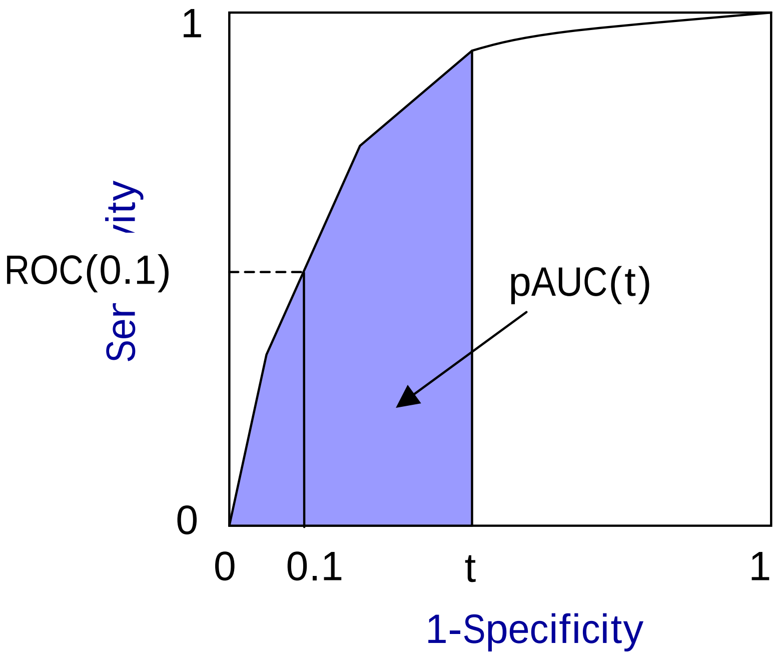


Example:
gene expression levels range from -1.401 to 1.039 (possible cutpoints)

- AUC (~Mann-Whitney statistic) scores for discrimination ability (and equals 0.5 for a random classifier)

- Besides AUC, the area under the full ROC curve, more interest is on the ROC curve at low values of *t*, corresponding to a maximum tolerable false positive rate t_0.





- Summary measures are defined by $AUC = \int_0^1 ROC(t)dt$,

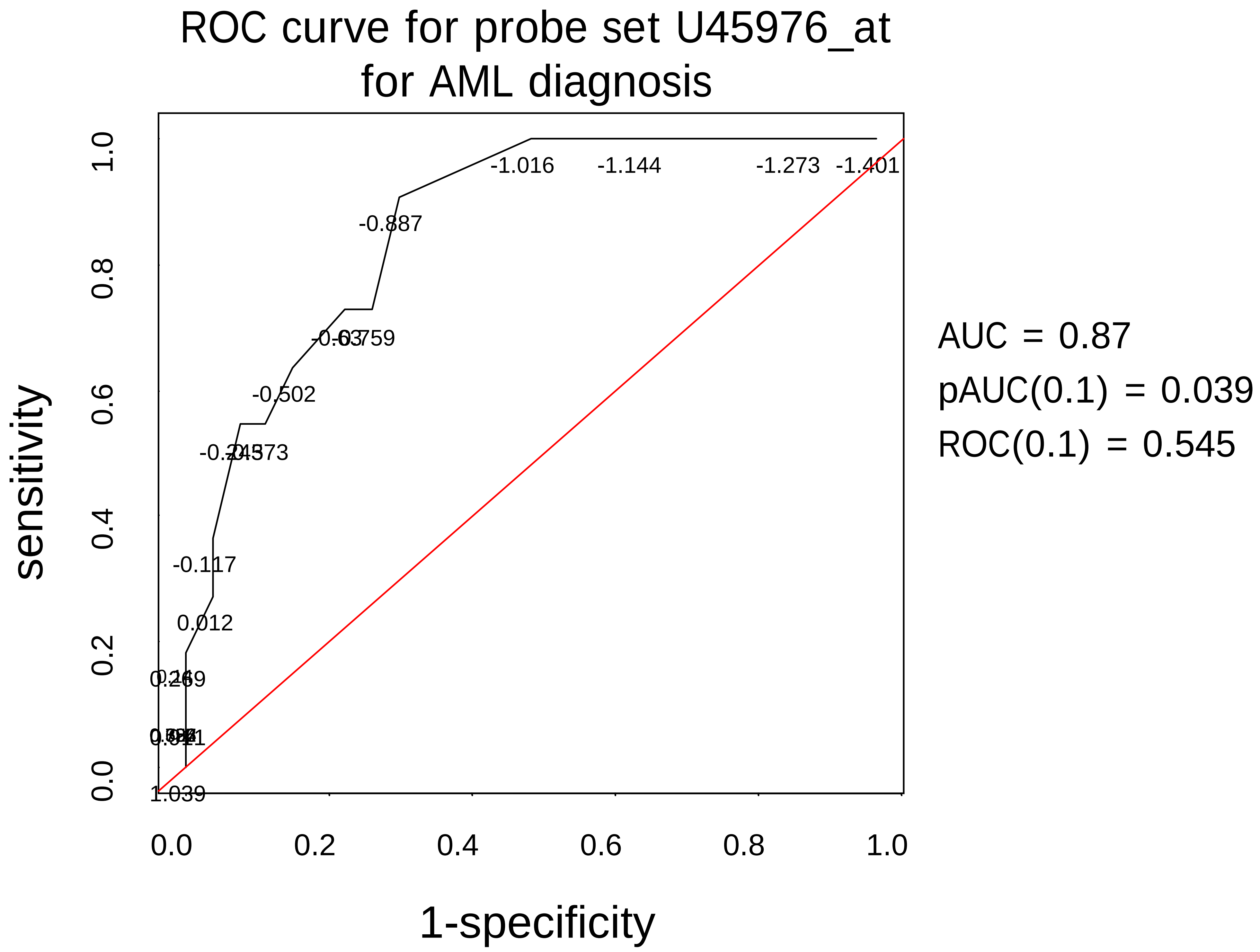
  $ROC(t_0) = P[Y_g^D \geq y_{(1-t_0)}^C]$ and $pAUC(t_0) = \int_0^{t_0} ROC(t)dt$

  where t_0 is a given false positive rate and $y_{(1-t_0)}^C$ is the corresponding $(1\text{-}t_0)$ quantile of the distribution of Y_g^C.

  The value $ROC(t_0)$ gives the proportion of target samples with expression levels above the $(1\text{-}t_0)$ quantile of control samples.

  The partial area under the curve, $pAUC(t_0)$, averages this proportion across values of $t \leq t_0$.

- **Comments:**
  - Since 1-specificity is comparable to the type I error α and sensitivity is comparable to the power (1-type II error) 1-β of a test statistic as used in clinical trials, the computation of *ROC*(t_0) for a fixed false positive rate t_0 is comparable to the computation of a retrospective power 1-β given type I error α and a fixed sample size *n* in clinical trials.
  - For clinical trials desired values of α and 1-β are 0.05 and 0.8 (0.9), respectively.
  - Due to the relatively small sample size of gene expression studies we recommend to choose t_0 as 0.1 instead of 0.05 and search for genes with high sensitivity (high power) of at least 0.6 to distinguish target samples from control samples.


ROC curve for probe set U45976_at
for AML diagnosis
AUC = 0.87
pAUC(0.1) = 0.039
ROC(0.1) = 0.545
sensitivity
1-specificity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
-1.016
-1.144
-1.273
-1.401
-0.887
-0.63
-0.759
-0.502
-0.245
-0.373
-0.117
0.012
0.269
0.811
1.039

- **Validation:**

- Sampling variability in the gene rankings is quantified using the ‘selection probability function’

$$\mathbf{P_g(G_s) = P[\text{gene } \mathit{g} \text{ ranked in the top } G_s \text{ genes}]}$$

$$\mathbf{= P[Rank(g) \leq G_s]}$$

which is estimated using bootstrap resampling, with the resampling unit being at the tissue/sample level.

- When a tissue is included in the bootstrap sample, the entire vector of data for all genes for that tissue is entered into the bootstrap data set, and genes are ranked according to the statistical measure chosen.

Example

Given a dichotomous element of the phenodata slot/data-frame of an exprSet, a ROC curve may be defined using the expression levels of any gene as the vector of marker values.

```
library(Biobase)
data(eset); pData(eset)
myauc <- function(x) {
  dx <- cov1 - 1
  AUC(rocdemo.sca(truth = dx, data = x, rule = dxrule.sca))
}
nResamp <- 5
nTiss <- ncol(exprs(eset))
nGenes <- nrow(exprs(eset[1:50, ]))
out <- matrix(, nr=nGenes, nc=nResamp)
set.seed(123)
for (i in 1:nResamp) {
  TissInds <- sample(1:nTiss, size=nTiss, replace=TRUE)
  out[, i] <- esApply(eset[1:50, TissInds], 1, myauc)
}
rout <- apply(out, 2, rank)
```

- **Notes**:

  - If no genes are differentially expressed, then the expected value of $P_g(G_s)$ is G_s /G
    where G is the total number of genes analyzed.

  - If sample size increase, the $P_g(G_s)$ will tend to 0 or 1 for differentially expressed genes, according to whether the true asymptotic discriminating measure for the g-th gene ranks below G_s or not.

  - Here we focus on detection of over-expressed genes although adaptation of the methods to detection of under-expressed genes is obvious.