The tail-rank test for finding biomarkers in microarray data

Kevin Coombes
Joint work with Jing Wang and Keith Baggerly
Section of Bioinformatics
Department of Biostatistics and Applied Mathematics
UT M. D. Anderson Cancer Center
kcoombes@mdanderson.org

11 October 2004

Overview

- Background and motivation
- The Tail-Rank Test
- Application to a published data set
- Looking at the results
- Sample size and power computations

Background and motivation

Answer, easy; question, very very hard.

– “Old Chinese proverb”, quoted by Spencer Bloch

in the preface to **Lectures on Algebraic Cycles**.

Consider a simple microarray experiment:

- Two groups of samples, “healthy” and “cancer”.
- Thousands of genes.
- Typical question: which genes are differentially expressed?

What is differential expression?

Standard answer: Difference in mean expression.

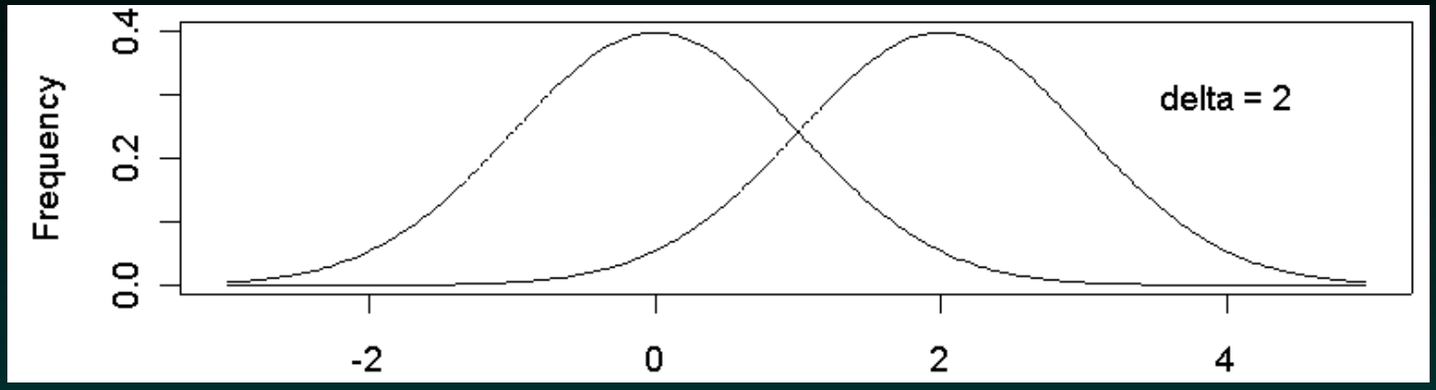
Most statistical tests for differential expression are based on the two-sample t-test or the Wilcoxon rank-sum test. Both tests focus on the center of the distribution.

At last year's CAMDA competition, the second-best entry (from Kerby Shedden and Jeremy Taylor) introduced the idea of "differential correlation", which looks at different properties of the distributions of gene expression values.

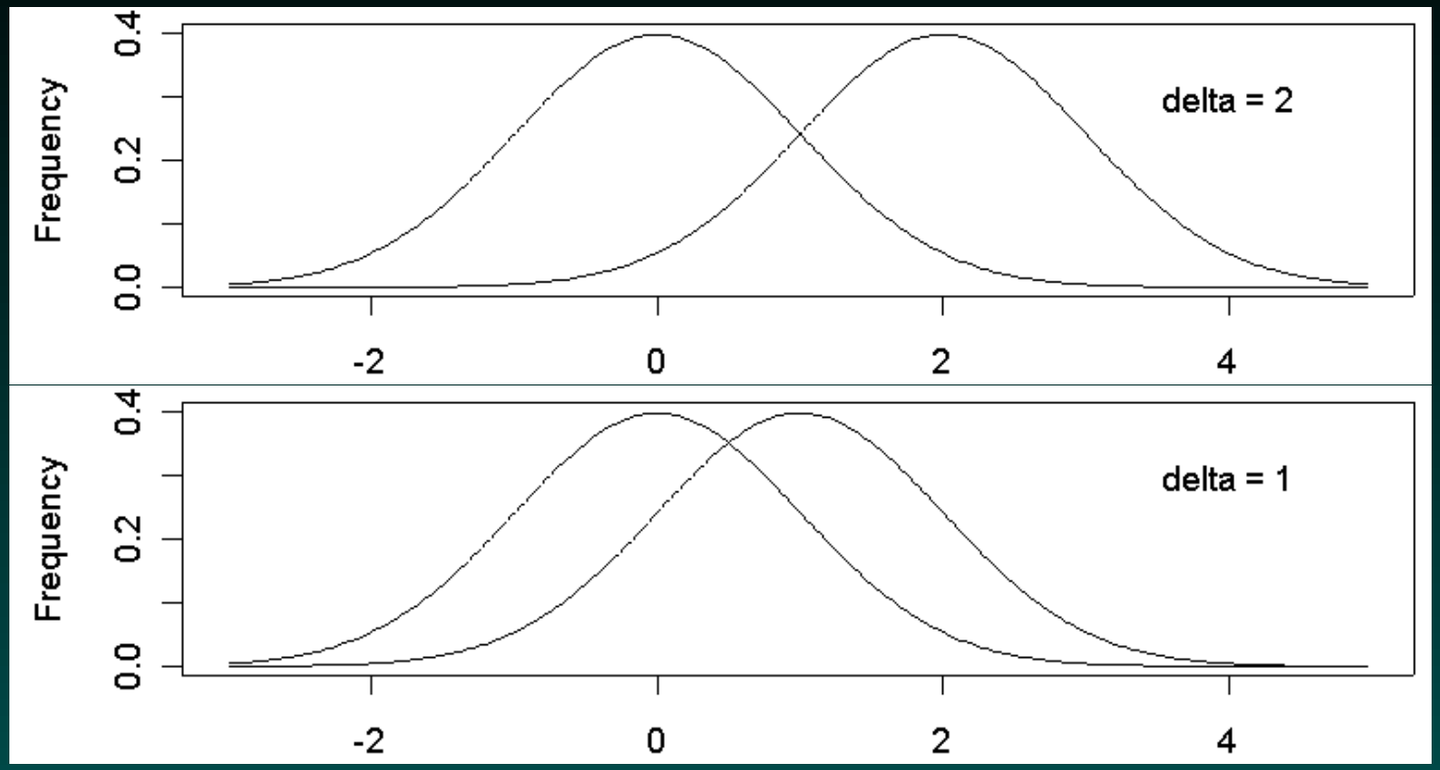
Maybe we should broaden our notion of "differential expression" by thinking about some other questions. . ..

For instance, one of the most interesting problems at present is: How do we detect potential biomarkers for cancer?

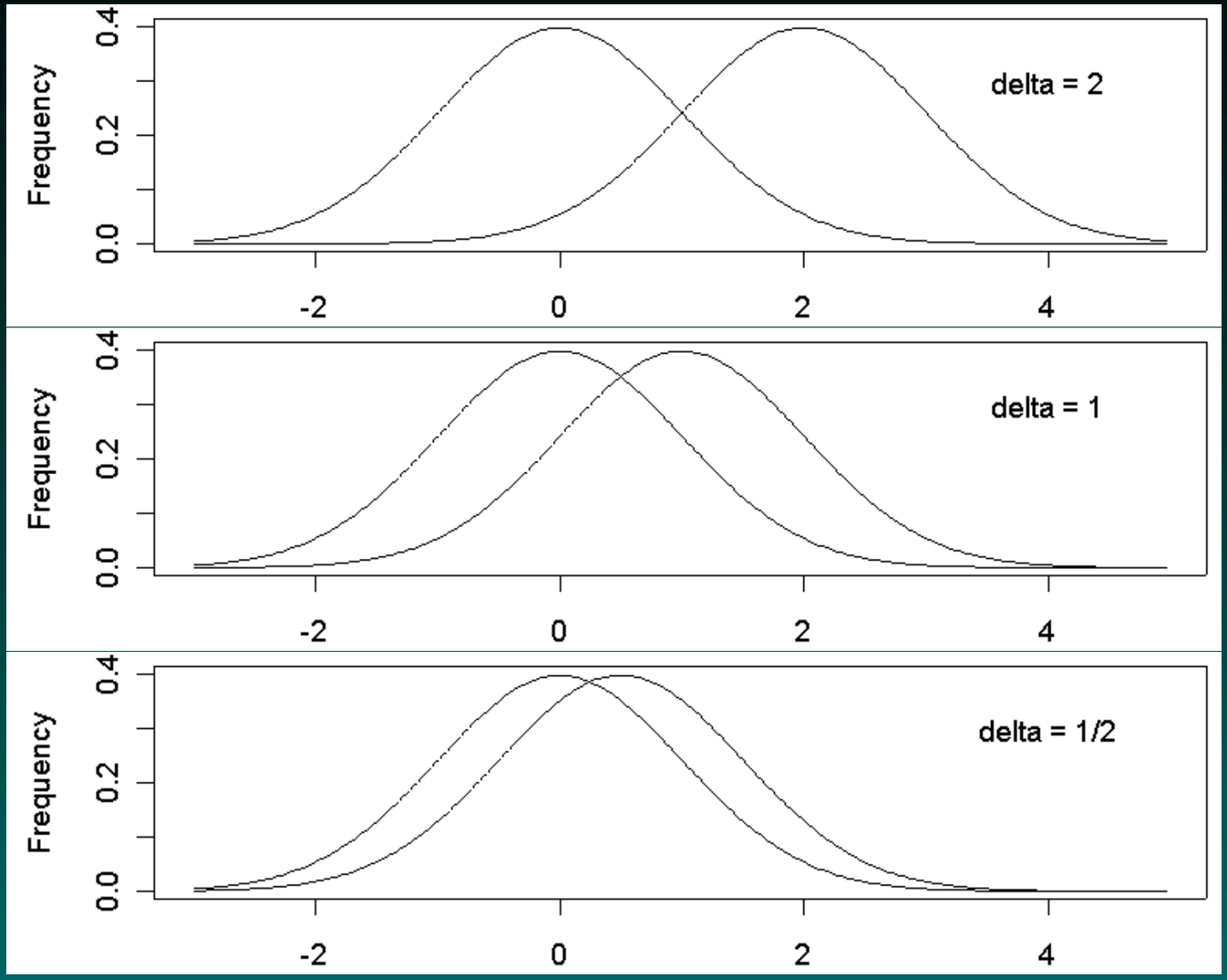
Central differences may be real but not useful



Central differences may be real but not useful



Central differences may be real but not useful



Cancer is heterogeneous

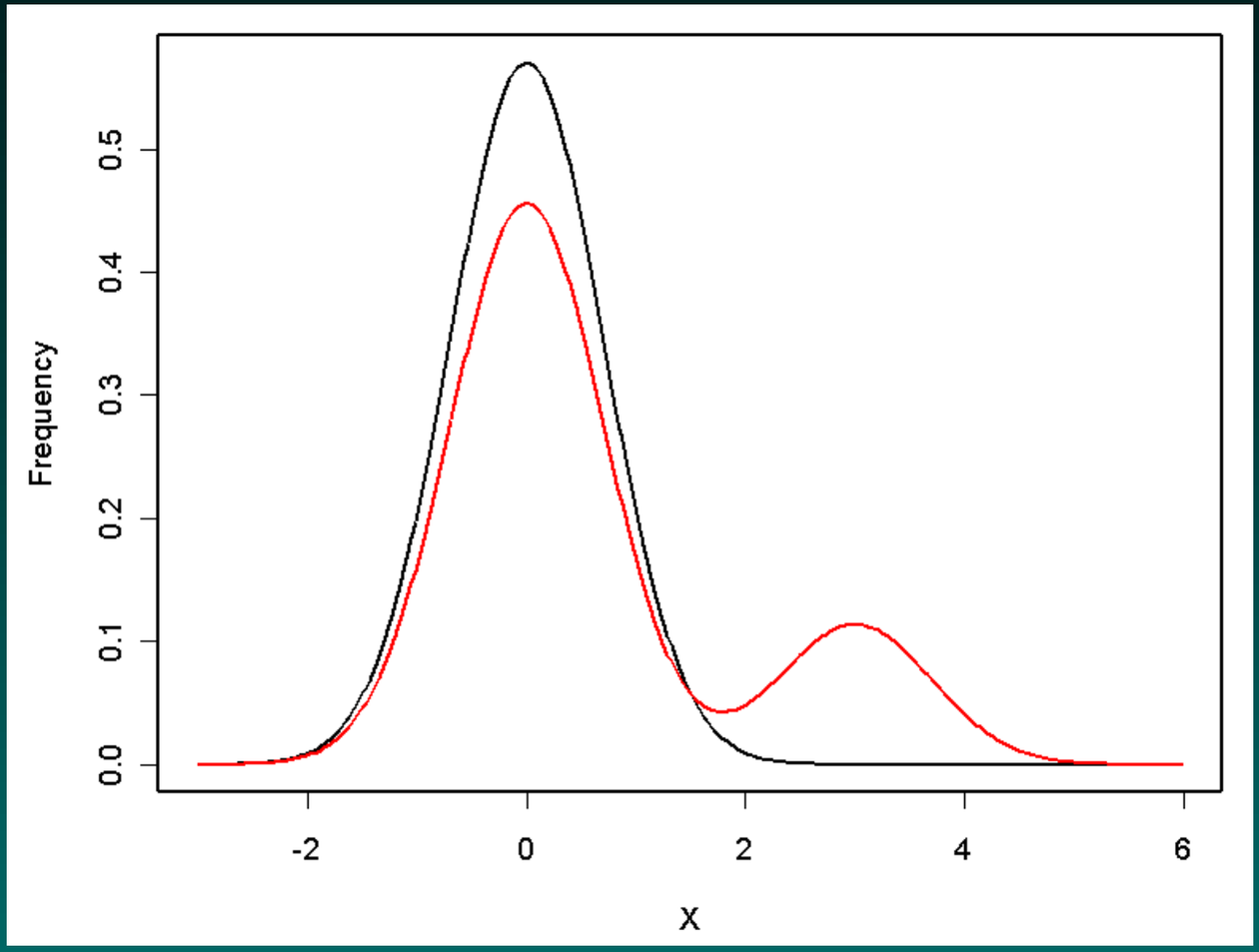
Cancers that are histologically the same are not identical. Some examples:

- Deletion of part of chromosome 3 (3p14-p23) is found in 50% of non-small-cell lung cancers;
- MYC amplification is found in 14% of stomach cancers;
- BRCA1 mutations are found in a subset of breast cancers;
- A translocation between chromosomes 11 and 14 occurs in 35% of mantle cell lymphomas.

These genetic abnormalities directly causes specific differences in gene expression that only occur in a subset of cancers.

Motivation: Subset Biomarkers

If a biomarker is only present in 20% of the cancer samples, then the distributions might look something like this.



The Tail-Rank Test

- Collect data on G genes from n_H healthy individuals. Write $X_{g,i}$ for measurement of gene g on individual i. Assume for fixed g that $X_{g,i} \sim X_g$ are IID.
- Specify a target value ψ for specificity.
- Estimate, for each g, a threshold τ_g such that $Prob(X_g < \tau_g) = \psi$.
- Collect data from n_C cancer patients. Count the number Y_g of cancer patients for which the measured expression level of gene g exceeds τ_g; we call Y_g the tail-rank statistic.
- Call g a biomarker if Y_g exceeds a certain threshold.

The null distribution

Null hypothesis: gene g is not a useful biomarker.

More precisely: the measurements on cancer patients have the same distribution as the measurements from healthy individuals.

Then: all Y_g have identical binomial distributions,

$$Y_g \sim Y = Binom(n_C, 1 - \psi).$$

The point here is that the probability of being in the tail is the same for healthy and cancer, and is given by $1 - \psi$, where ψ was the desired specificity.

Even when we perform the same test for G genes, the expected maximum value of G independent instances of Y_g remains small.

Let $M_G = max_{g=1...G}(Y_g)$ be the maximum over G IID binomial random variables. Also, let

$$\begin{aligned} \alpha &= \alpha(m) = Prob(Y > m) \\ \gamma &= Prob(M_G > m) \end{aligned}$$

Then

$$\begin{aligned} 1 - \gamma &= Prob(M_G \leq m) \\ &= Prob(Y_1 \leq m, \ldots, Y_G \leq m) \\ &= Prob(Y \leq m)^G = (1 - \alpha)^G. \end{aligned}$$

The maximum value expected by chance

Solving,

$$\alpha = 1 - (1 - \gamma)^{1/G}.$$

and m is the $(1 - \alpha)^{\text{th}}$ quantile of a single binomial distribution:

	$\gamma = 0.01$, $\psi = 0.99$			
n_C	$G = 100$	1000	10000	100000
10	3	3	4	4
20	3	4	5	5
50	5	6	6	7
100	6	7	8	9
250	10	12	13	14
500	15	17	19	20

Cutoff depends on ψ and γ

	$\gamma = 0.05, \psi = 0.95$			
n_C	$G = 100$	1000	10000	100000
10	4	5	5	6
20	5	6	7	8
50	9	10	11	13
100	13	15	17	19
250	25	28	30	32
500	42	46	49	52

Interpretation

One needs to specify two parameters in order to apply the tail-rank test.

1. ψ, the desired specificity of the biomarker

2. γ, the desired bound on the FWER

Then, given the number of genes and the number of cancer samples, the values m in a table like the previous one represent the largest value of Y_g that we would expect to see by chance over the entire microarray. Any gene where we observe $Y_g > m$ is a potential biomarker.

Application to a published data set

Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR: **Gene expression profiling identifies clinically relevant subtypes of prostate cancer**. *Proc Natl Acad Sci USA* 2004, **101**:811–816.

- 41 normal prostate, 62 prostate cancer, 9 lymph node metastases.
- Two-color glass arrays with $42,129$ spots
- Common reference channel
- Loess normalization, take log ratios

Tail-rank and real data

With:

- Number of healthy samples, $n_H = 41$
- Number of cancer samples, $n_C = 71$
- Specificity, $\psi = 0.95$
- Significance, $\gamma = 0.05$

a gene was called a biomarker if at least 16 of the 71 cancer samples were above the threshold representing the 95^{th} percentile in the 41 healthy samples.

Tail-rank and real data

We assumed that the log ratios of the normal prostate samples were normally distributed. We computed 90% tolerance bounds for the 5^{th} and 95^{th} percentiles, and counted the number of combined prostate cancer samples whose log ratios fell outside these boundaries.

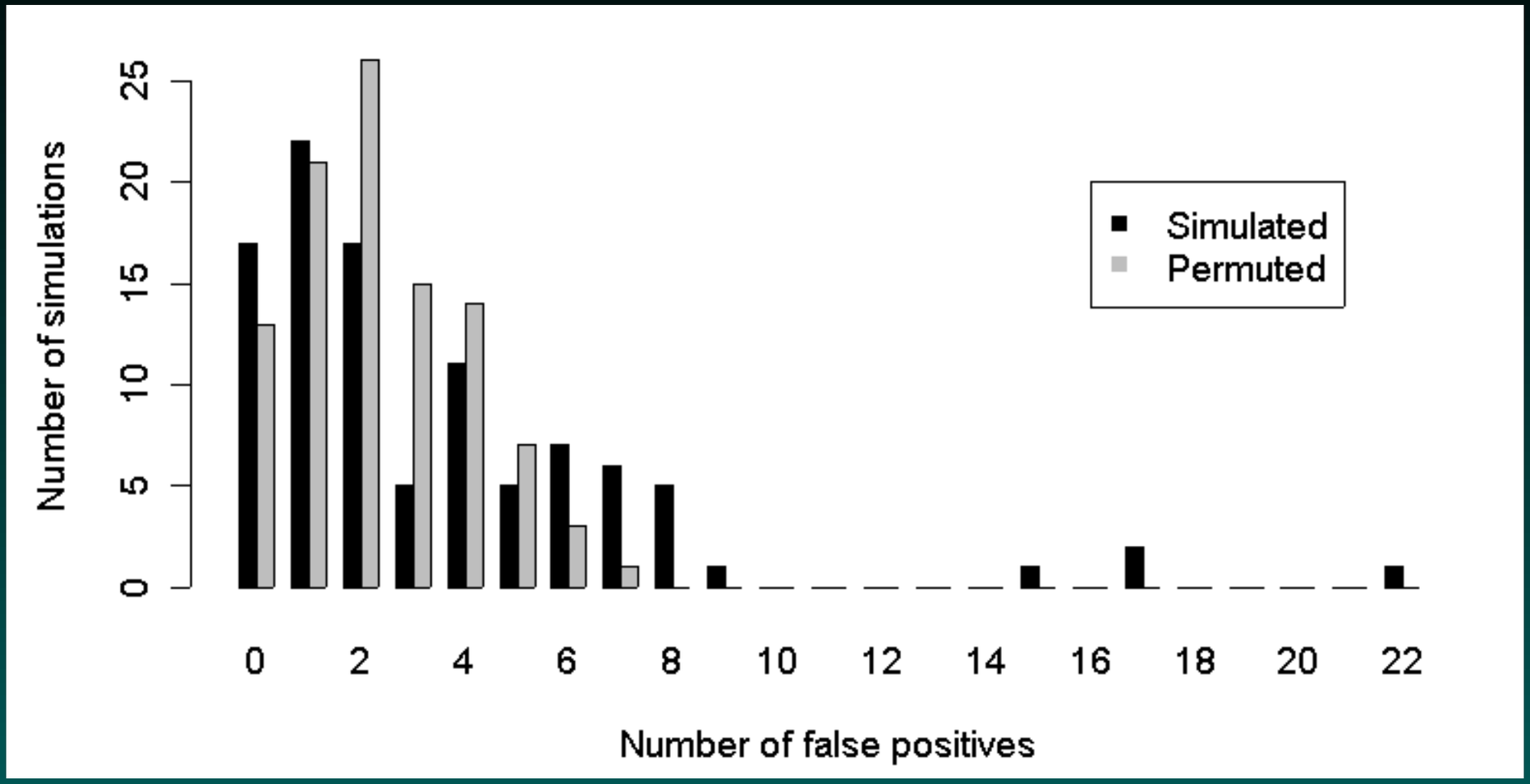
Alternatively, one could estimate the percentiles in the distribution of healthy individuals empirically, possibly using a bootstrap.

Tail-rank results

We identified $1,766$ spots that were “positive” biomarkers, since they were present at higher than normal levels in at least 16 cancer samples. We also identified $1,930$ spots that were “negative” biomarkers, since they were expressed at lower than normal levels in at least 16 samples. In total, we identified 3,692 spots as candidate biomarkers.

Although the theory told us the number of false positives should be close to zero, we decided to test this using both simulations and a permutaion test. We simulated completely random (IID normal) data 100 times, and we permuted the samples labels on the real data 100 times.

Reviewing significance



Pretty good, when you consider that the test with the real data detected a few thousand potential markers!

Differential expression results

We performed two-sample t-tests on the prostate cancer data set (adjusing for multiple testing using BUM). With $FDR < 0.05$, we used a cutoff at $p < 0.000045$ or $|t| > 4.25$. We detected 3,522 differentially expressed spots. Of these, $1,415$ spots were overexpressed in prostate cancer and $2,107$ spots were underexpressed.

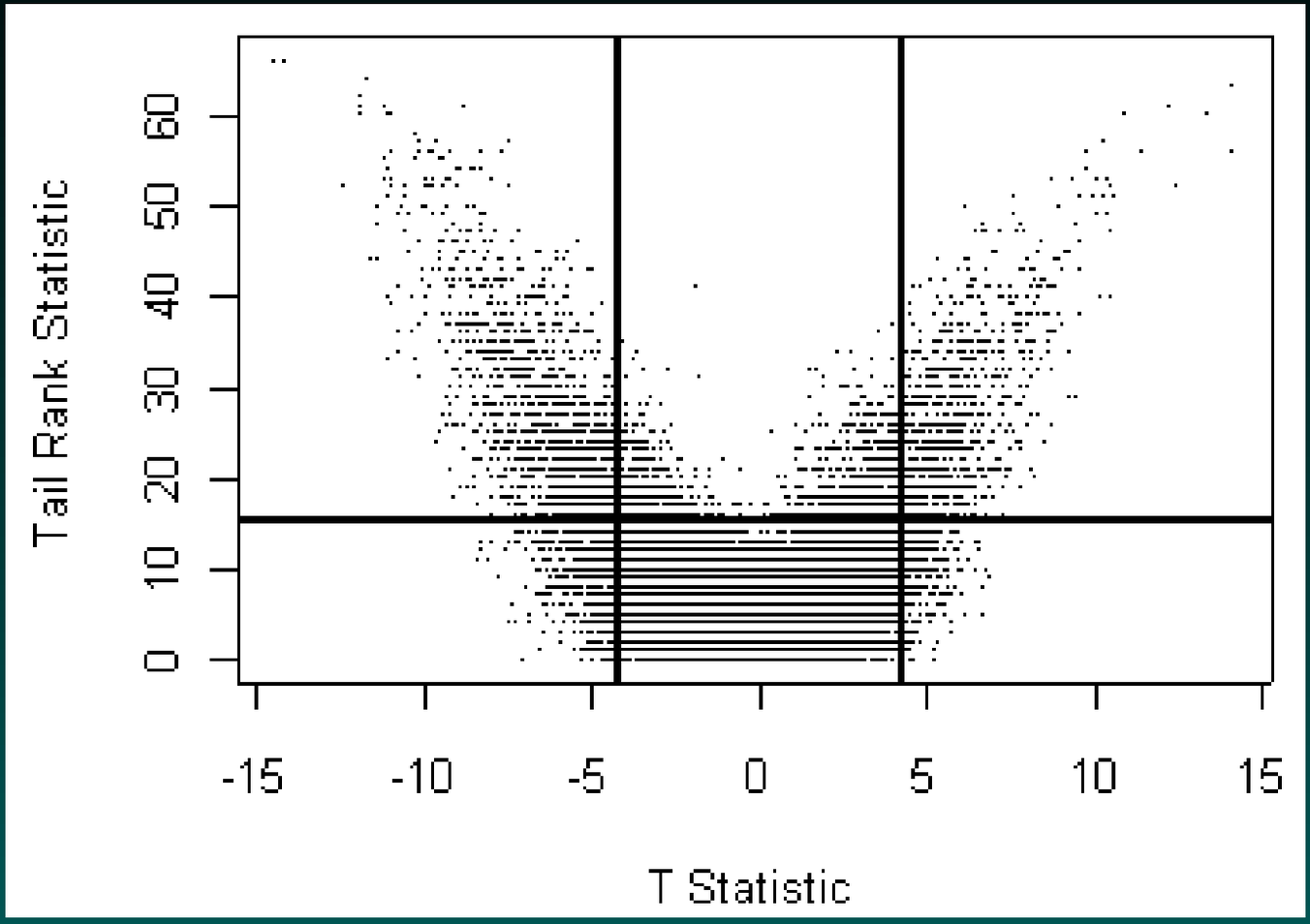
We also performed a Wilcoxon rank-sum test with the empirical Bayes approach. In order to get comparable results, we selected a cutoff corresponding to a posterior probability of 99.9%. We detected 3,627 differentially expressed spots. Of these, $1,498$ spots were overexpressed and $2,129$ spots were underexpressed in prostate cancer.

Comparing tests

The number of genes found by the three tests was very similar. Are they finding the same things?

There was good agreement between the t-test and the Wilcoxon test. More than 90% $(1,905)$ of underexpressed and 88% $(1,244)$ of overexpressed spots that were found by the t-test were also detected by the Wilcoxon test. So, we only need to compare one of these to the tail-rank test.

Comparing tests



Lower left and right = different by T, not by tail-rank

Upper center = different by tail-rank, not by T.

Looking at the results

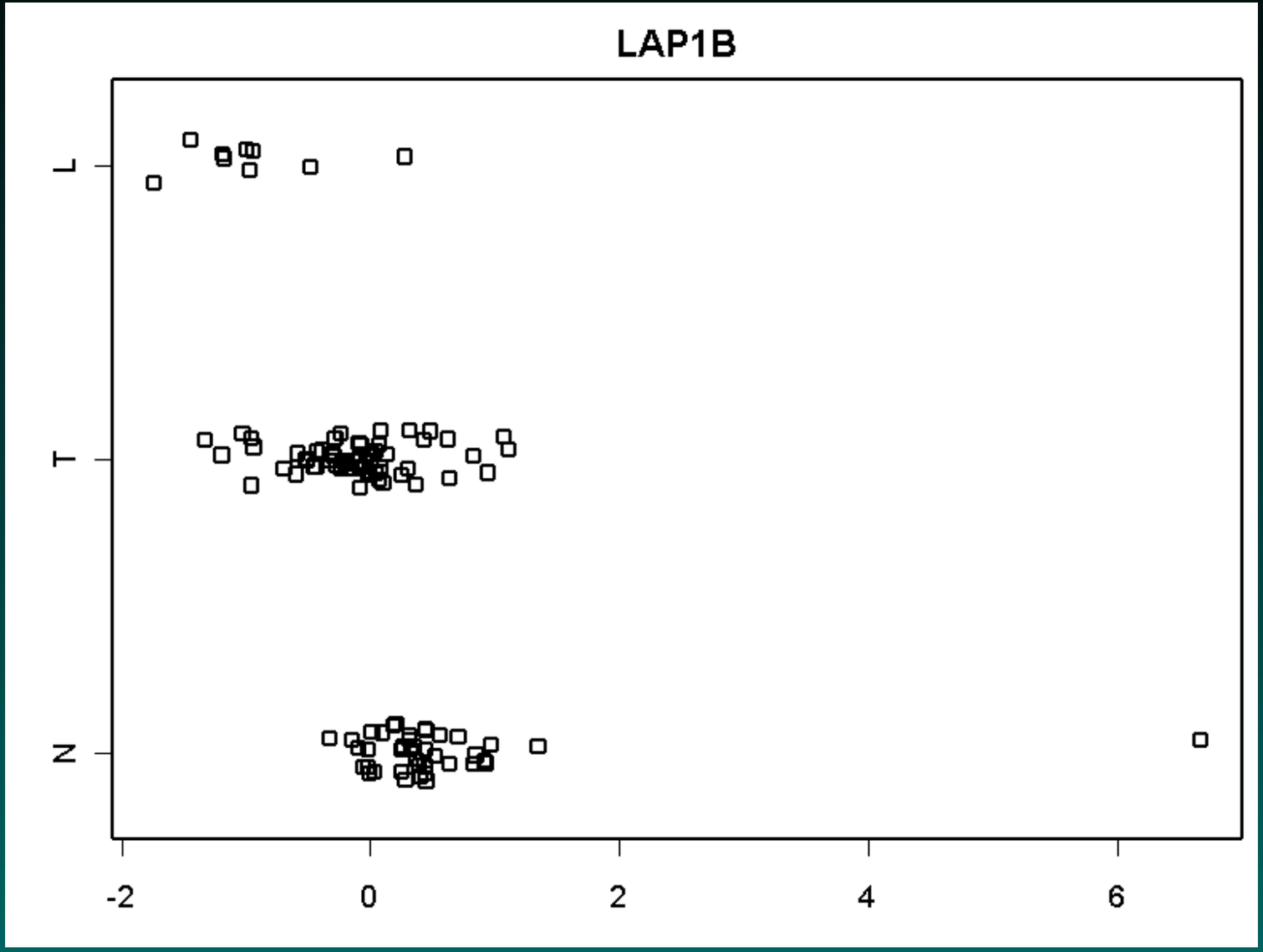
Since the tests give different answers, which one should we believe?

Both, since they are giving the answers to different questions.

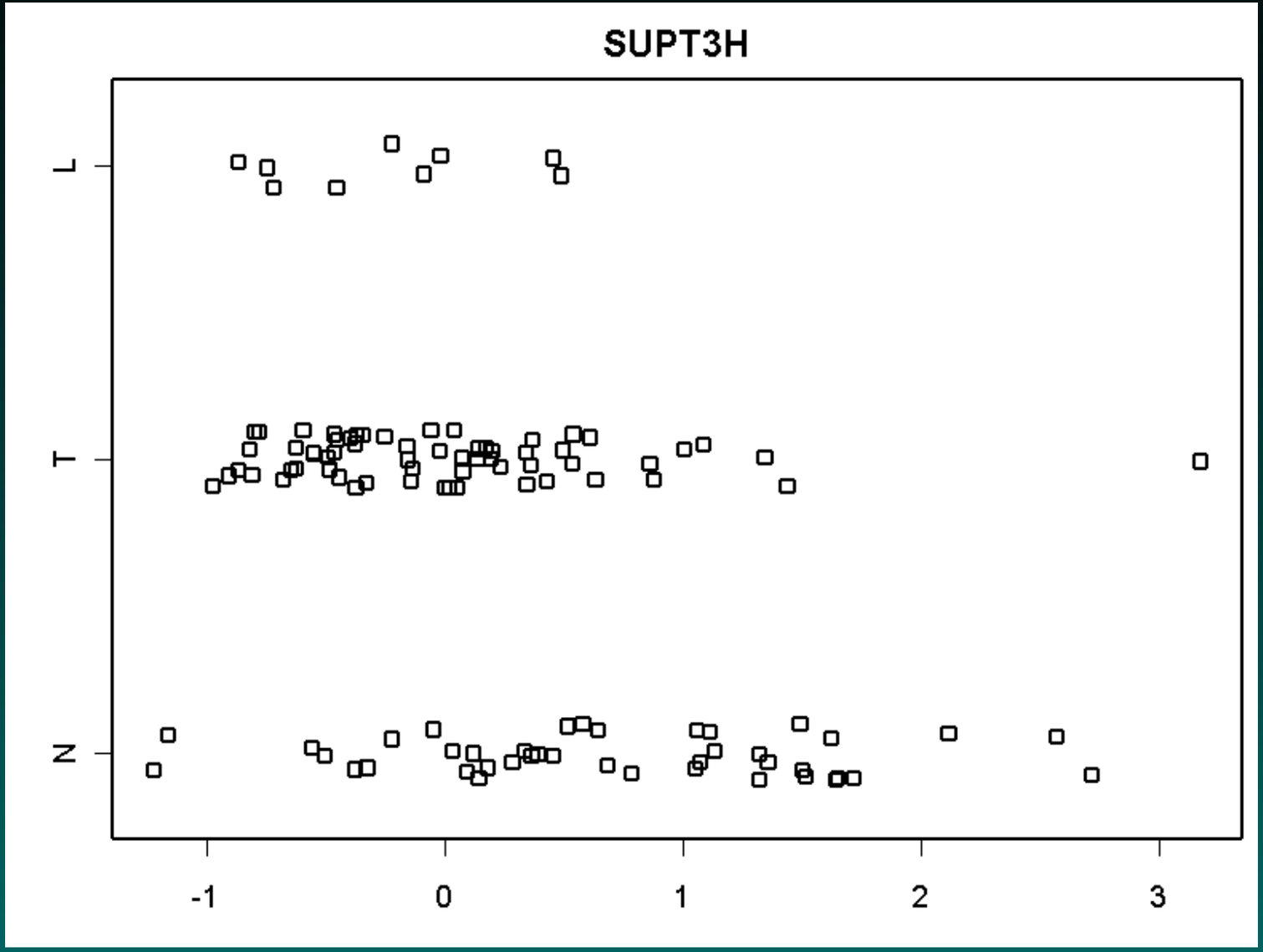
Whether you perform one test or many, however, it is useful to look at the expression values for some of the genes that you find, if only to make sure you believe the results.

First, we will look at genes with small tail-rank statistics (< 2) and significant t-statistics.

Outliers can throw off the estimates



Some genes are normally variable

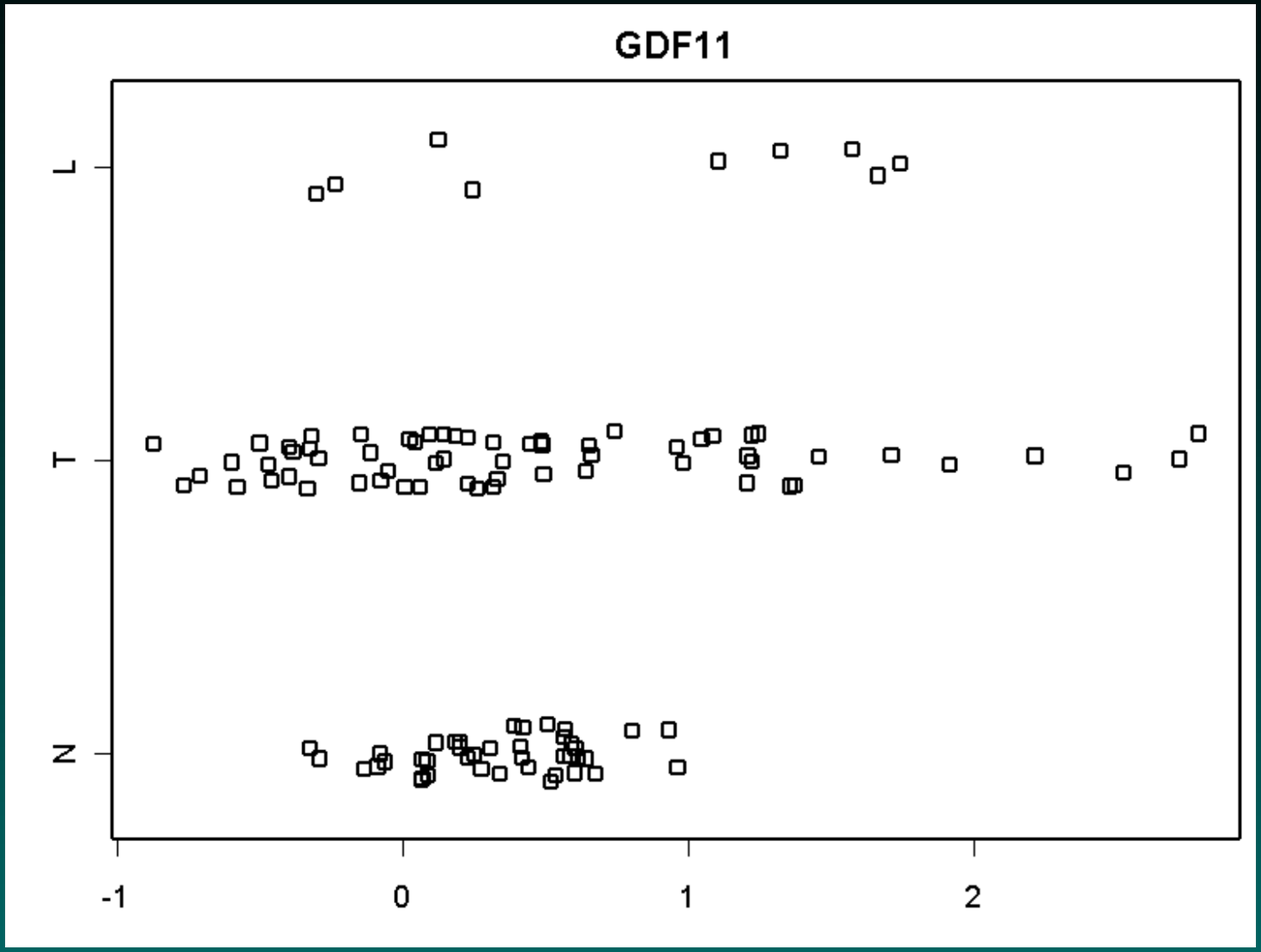


Selecting interesting biomarkers

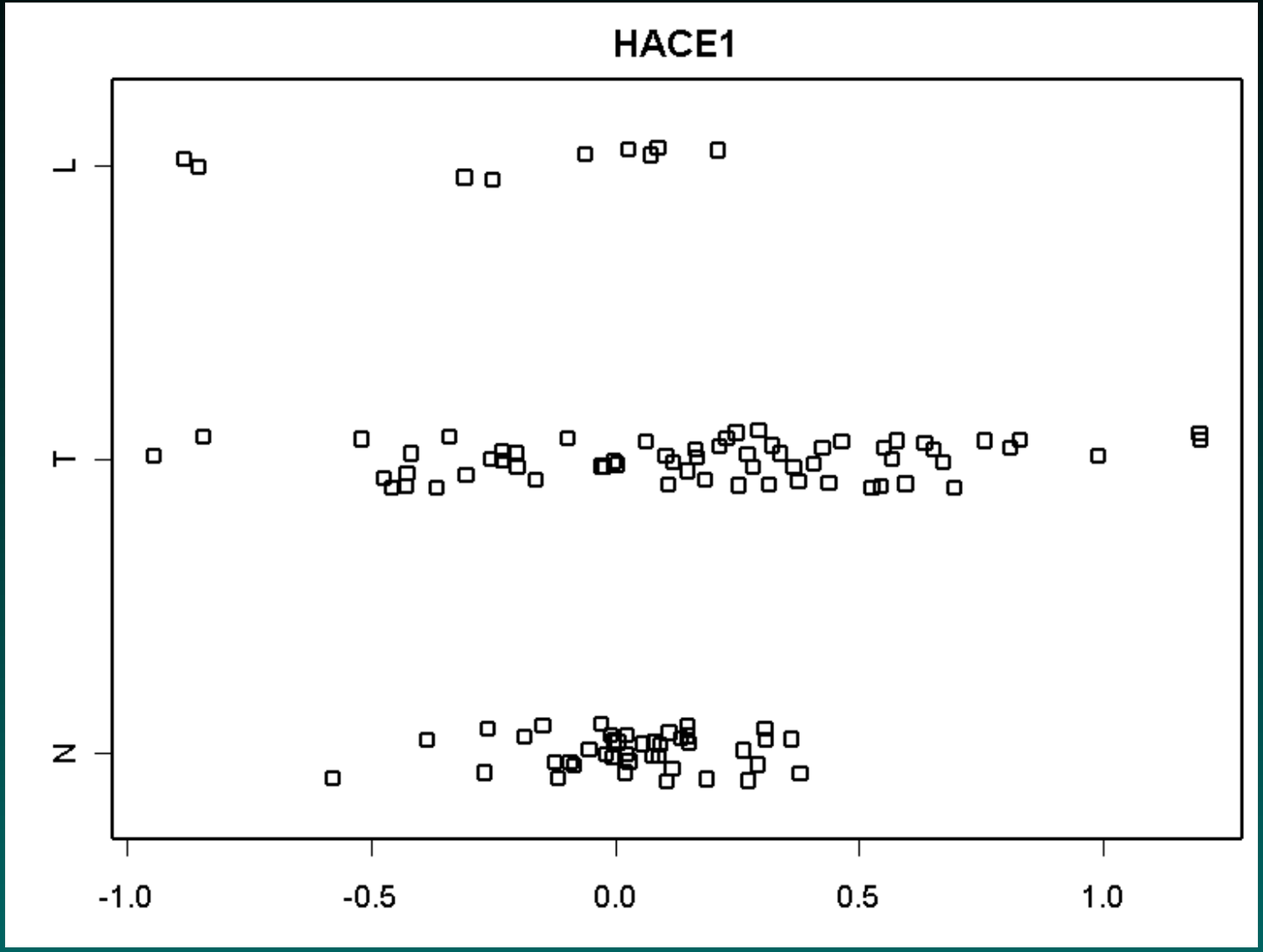
Next, we look at genes with significant tail-rank statistics and small t-statistics ($|t| < 1.25$).

If we're right about the method, these genes should make potentially interesting biomarkers even though a t-test or Wilcoxon test wold never find them.

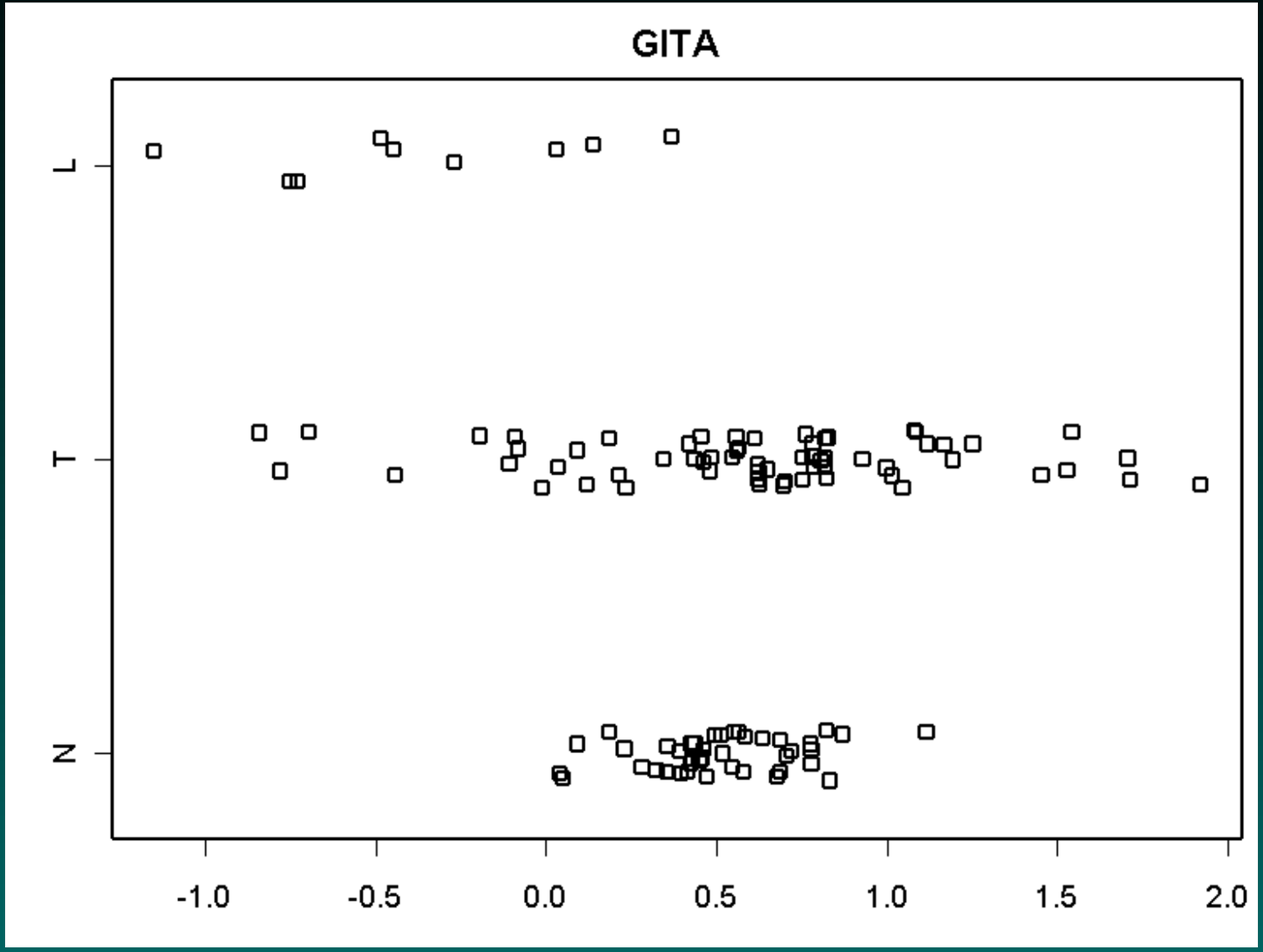
GDF11



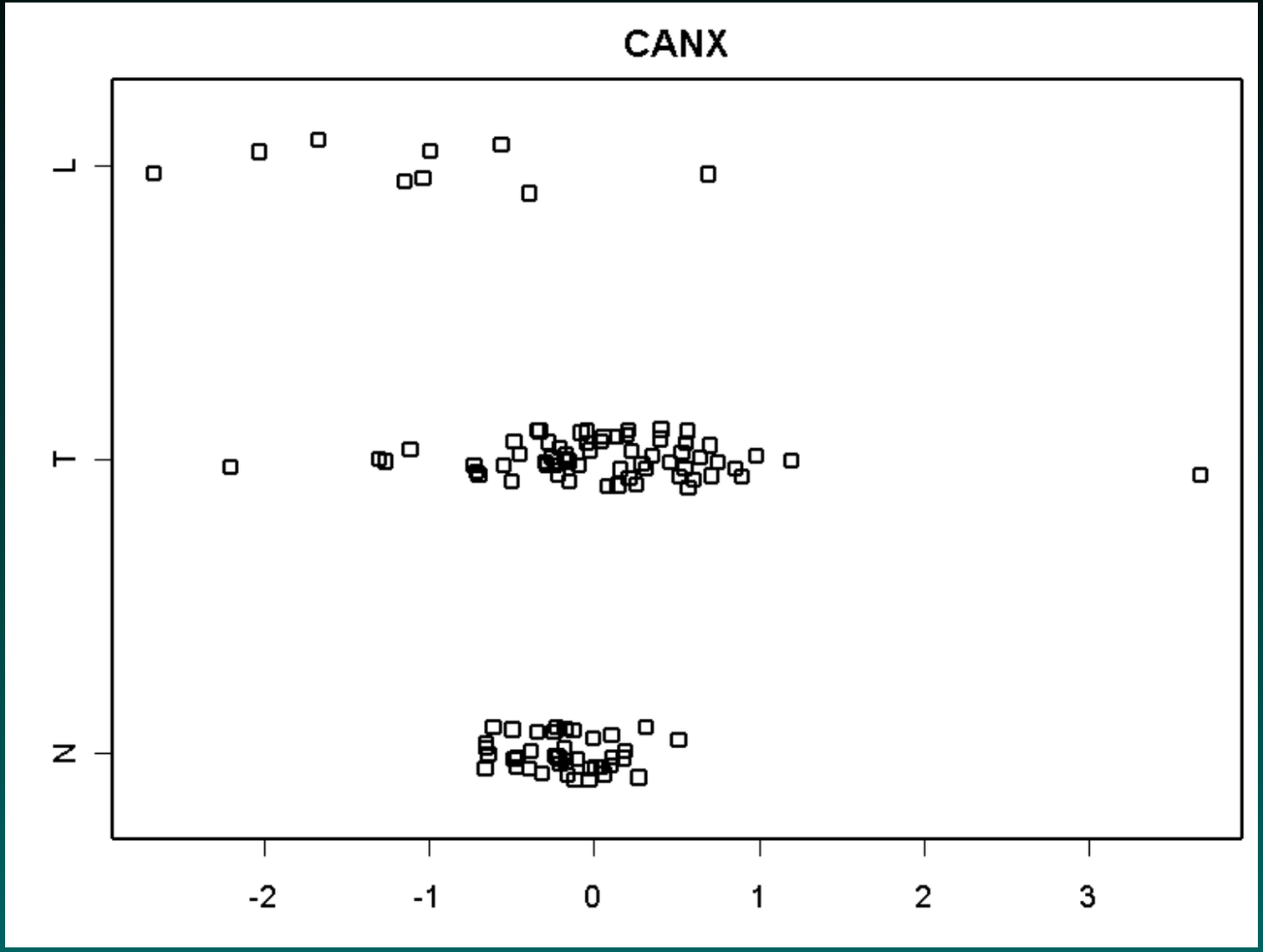
HACE1



GITA



CANX



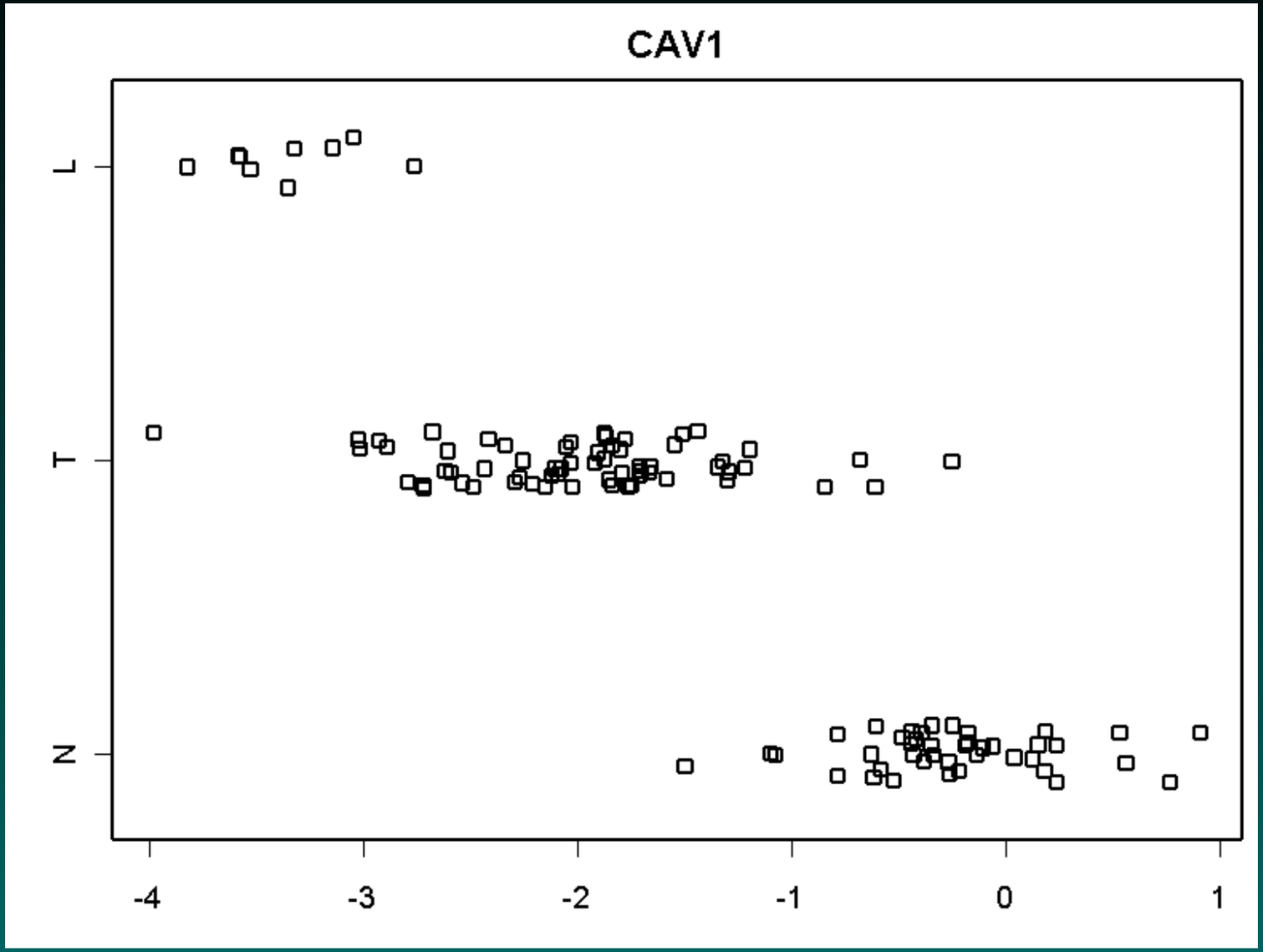
Consistency

CANX is particularly interesting. Five different clones represented calnexin on these microarrays. All five spots containing these clones were selected by the tail-rank test, even though the t-statistics were insignificant (0.80, 0.82, 0.89, 0.92, and 2.40).

Depending on the clone, between 16 and 20 of the prostate cancer samples had expression levels that were higher than the 95^{th} percentile of the expression in normal prostate.

Interestingly, between 6 and 8 of the 9 lymph node metastases had levels that were *below* the 5^{th} percentile of normal, and 8 lymph node metastases had levels that were well below the mean for the primary prostate cancers. This finding is particularly intriguing since it has recently been reported that downregulation of calnexin increases the metastatic potential of melanoma cells.

When T and Tail-Rank agree



Sample size and power computations

To compute the power, fix

- the significance level γ,
- the specificity ψ,
- the number G of genes.

Given these values, we estimate the expected maximum value of G independent instances of Y_g under the null hypothesis as a function of the sample size n_C. As before, we compute this estimate as the $(1-\alpha) = (1-\gamma)^{1/G}$ quantile of the null binomial distribution $Binom(n_C, 1-\psi)$.

Sample size and power computations

For a binomial random variable $X \sim Binom(N, p)$, we write $F(x \mid N, p) = Prob(X \leq x)$ for its cumulative distribution function. Now, the expected maximum value $m = E[M_G]$ of Y_g over G genes satisfies

$$F(m \mid n_C, 1 - \psi) = (1 - \gamma)^{1/G},$$

or, equivalently,

$$m = m(G, n_C, \psi) = F^{-1}((1 - \gamma)^{1/G} \mid n_C, 1 - \psi).$$

Power depends on marker sensitivity

Since we identify a gene as a biomarker if the observed value of Y_g is larger than m, the power π to detect a biomarker whose true sensitivity equals ϕ is given by

$$\pi = Prob(Binom(n_C, \phi) > m) = 1 - F(m \mid n_C, \phi).$$

Thus, it is straightforward to compute the power provided we are given the sample size, the sensitivity, and the number of genes.

Example of power computations

	$\gamma = 0.05, \psi = 0.95$					
n_C	$\phi = 0.10$	0.20	0.30	0.40	0.50	0.60
10	0.0001	0.0064	0.05	0.16	0.38	0.63
20	0.0004	0.0321	0.23	0.58	0.87	0.98
50	0.0032	0.2893	0.86	0.99	0.99	1.00
100	0.0100	0.7288	0.99	1.00	1.00	1.00
250	0.1247	0.9994	1.00	1.00	1.00	1.00
500	0.5218	1.0000	1.00	1.00	1.00	1.00

From the table, we see that even 500 samples are not enough to detect a biomarker that is present in only 10% of the cancer patients. By contrast, 100 samples have enough power ($> 70\%$) to detect biomarkers with a sensitivity of 20%, fewer than 50 are needed to detect a biomarker with a sensitivity of 30%, and as few as 10 will suffice to detect biomarkers with a sensitivity of 70%.

Using the tail-rank test in R

A preprint describing the tail-rank test, along with an R package, is available on the web at

http://bioinformatics.mdanderson.org/TailRank

Basic usage:

```
> tr.stats <- tail.rank.test(data, status)
```

Future directions

- Find a way to avoid the normal assumption in healthy samples
  - Needed for applications to response or prognosis
- Multivariate analogue
  - Best way to combine markers that pick out specific subsets
- (Relative) rank-based version of the test.