Are a Set of Microarrays Independent of Each Other?

Bradley Efron
Stanford

A Cardiovascular Microarray Study

(Dr. Tom Quertermous)

- $\boldsymbol{n = 63}$ stent recipients: 44 "low risk", 19 "high risk"

- $\boldsymbol{m = 20426}$ genes on each patient's microarray

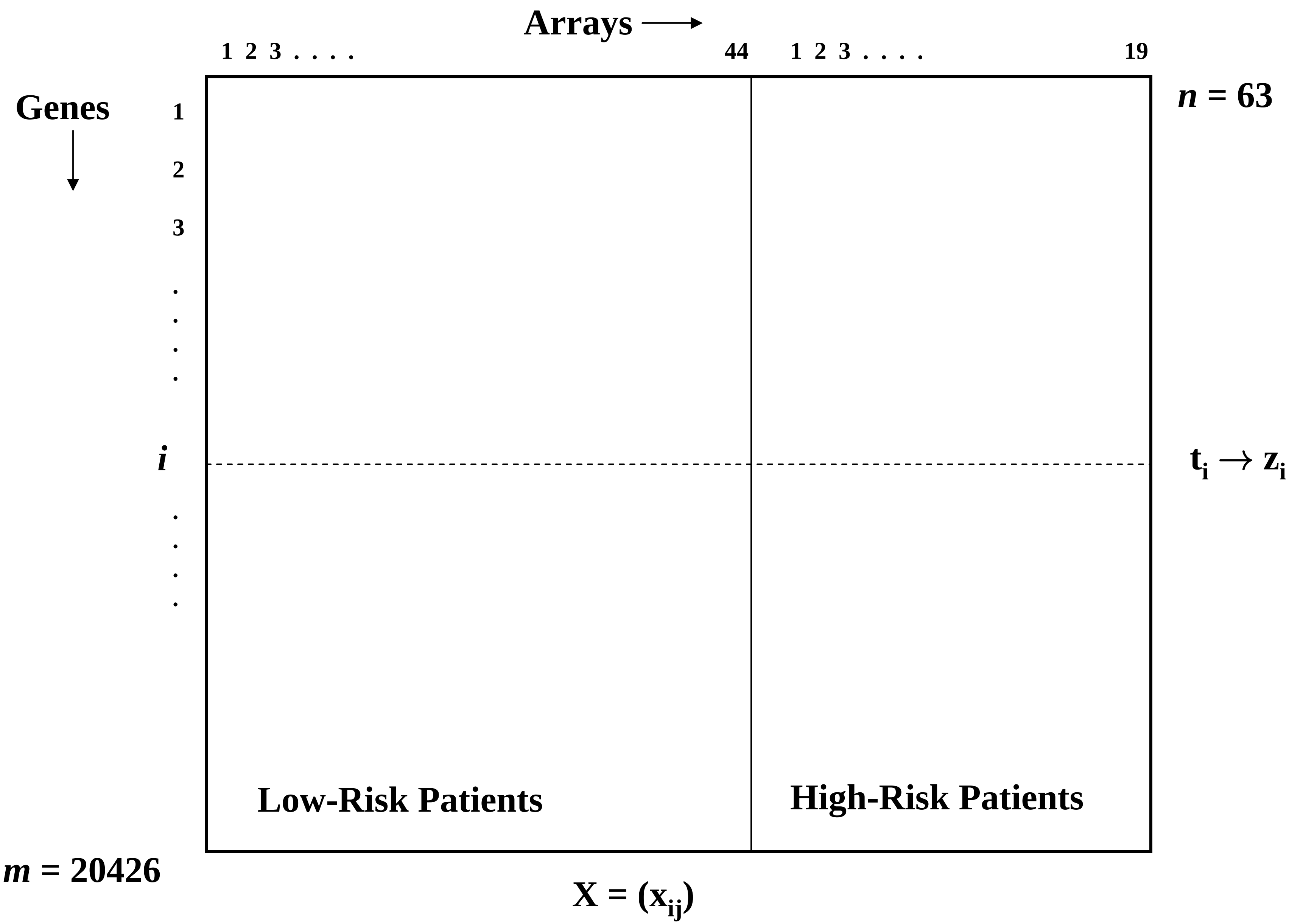
- **Data Matrix** $$\underset{m \times n}{X} = \begin{pmatrix} & \vdots & \\ \cdots & x_{ij} & \cdots \\ & \vdots & \end{pmatrix}$$

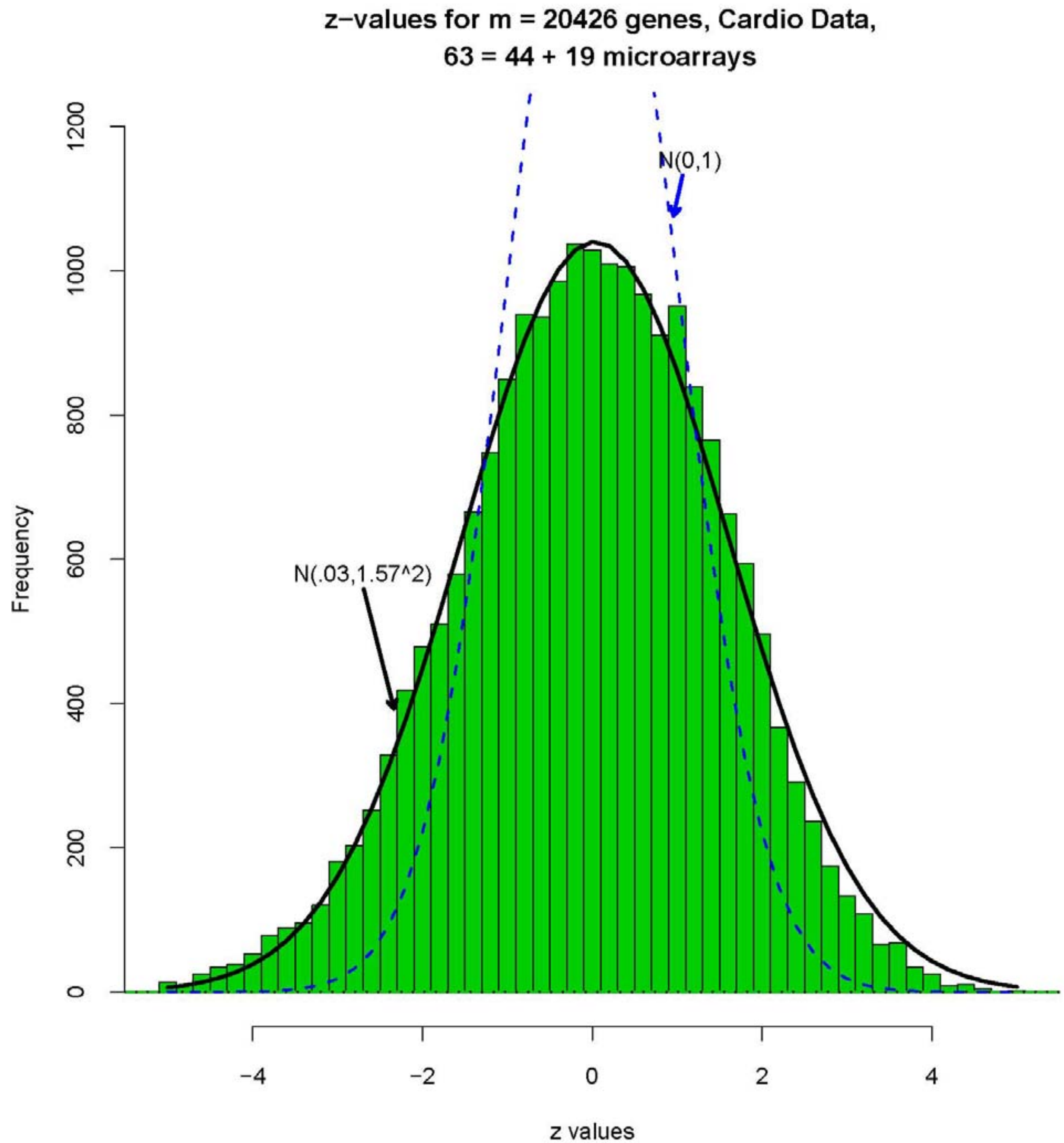
- *t statistics* t_i compares high vs. low risk expression values, patient i

- *z values* $\boxed{z_i = \Phi^{-1}(F_{61}(t_i))}$ [F_{61} cdf for t_{61}]

- *Theoretical Null* $H_0 : z_i \sim N(0,1)$

Cardio Study: *m* = 20426 genes, *n* = 63 microarrays

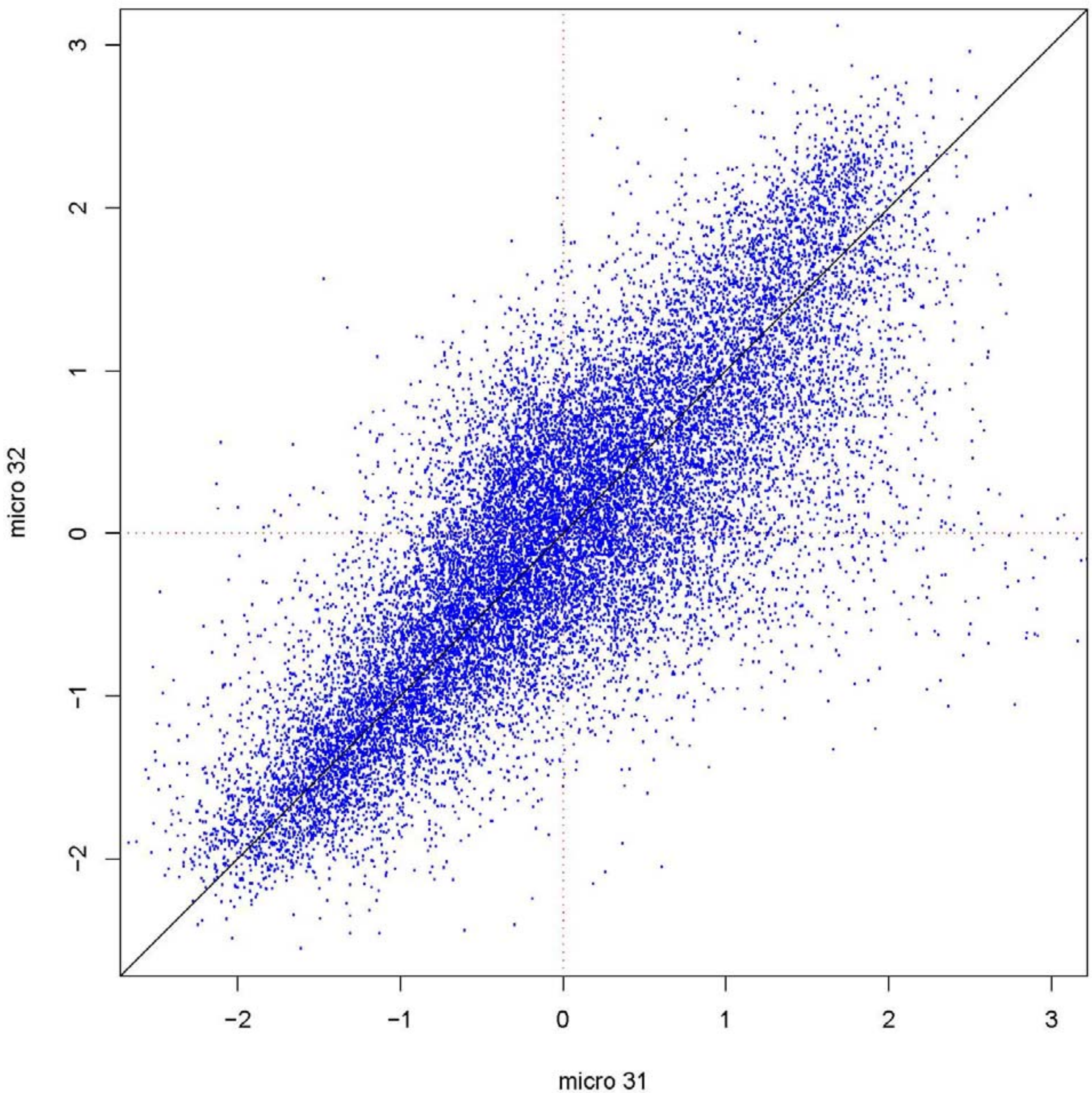



z−values for m = 20426 genes, Cardio Data,
63 = 44 + 19 microarrays
Frequency
0
200
400
600
800
1000
1200
N(0,1)
N(.03,1.57^2)
−4
−2
0
2
4
z values

z-value Overdispersion

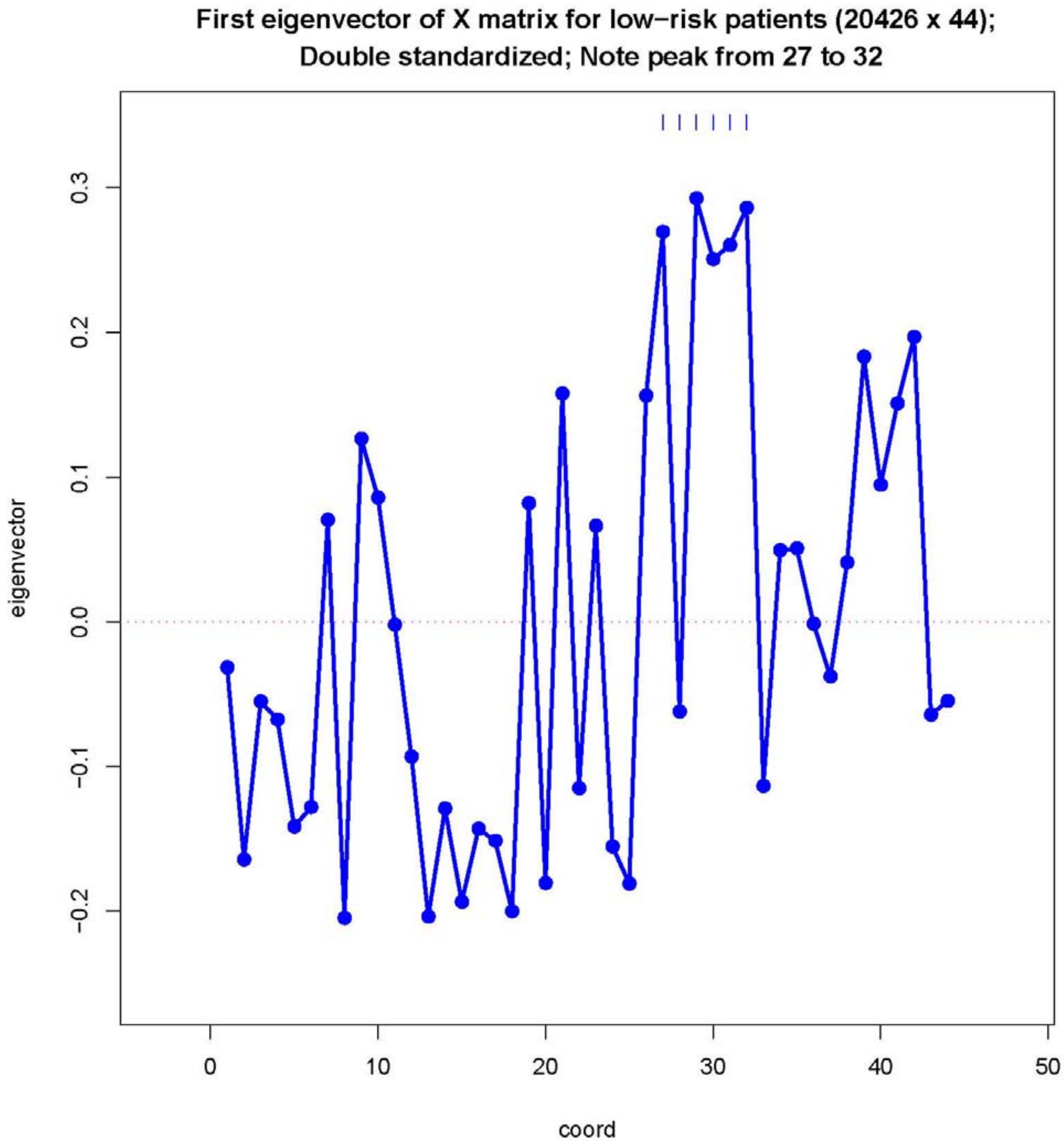
- *z*-value histogram overdispersed near center compared to $N\,(0, 1)$ theoretical null – more like $N\,(.03, 1.57^2)$
- **Possible causes:**
  - Unobserved covariates (Efron 2004)
  - Correlation between genes (Efron 2007)
  - *Correlation across microarrays* (Today: df really < 61)
- Look at low risk group: X: 20426 x 44
- **“Doubly Standardized”:**
  - Row and column means 0
  - Row and column variances 1
- No “signal”

Scatterplot of microarrays 31 and 32 (gene means removed); Correlation .805



Permutation Tests for Independence

- "v_1" *first eigenvector* of X (or of $\widehat{\Delta} = X'X/m$); dim n.
- If columns X i.i.d then v_1 should look "random" plotted versus 1, 2, …, n.
- $S(v_1)$ some statistic measuring apparent structure, for example, slope of linear fit to v_1.
- Compare S to S^* values obtained by permuting entries of v_1.



First eigenvector of X matrix for low−risk patients (20426 x 44);
Double standardized; Note peak from 27 to 32
eigenvector
coord
0.3
0.2
0.1
0.0
−0.1
−0.2
0
10
20
30
40
50

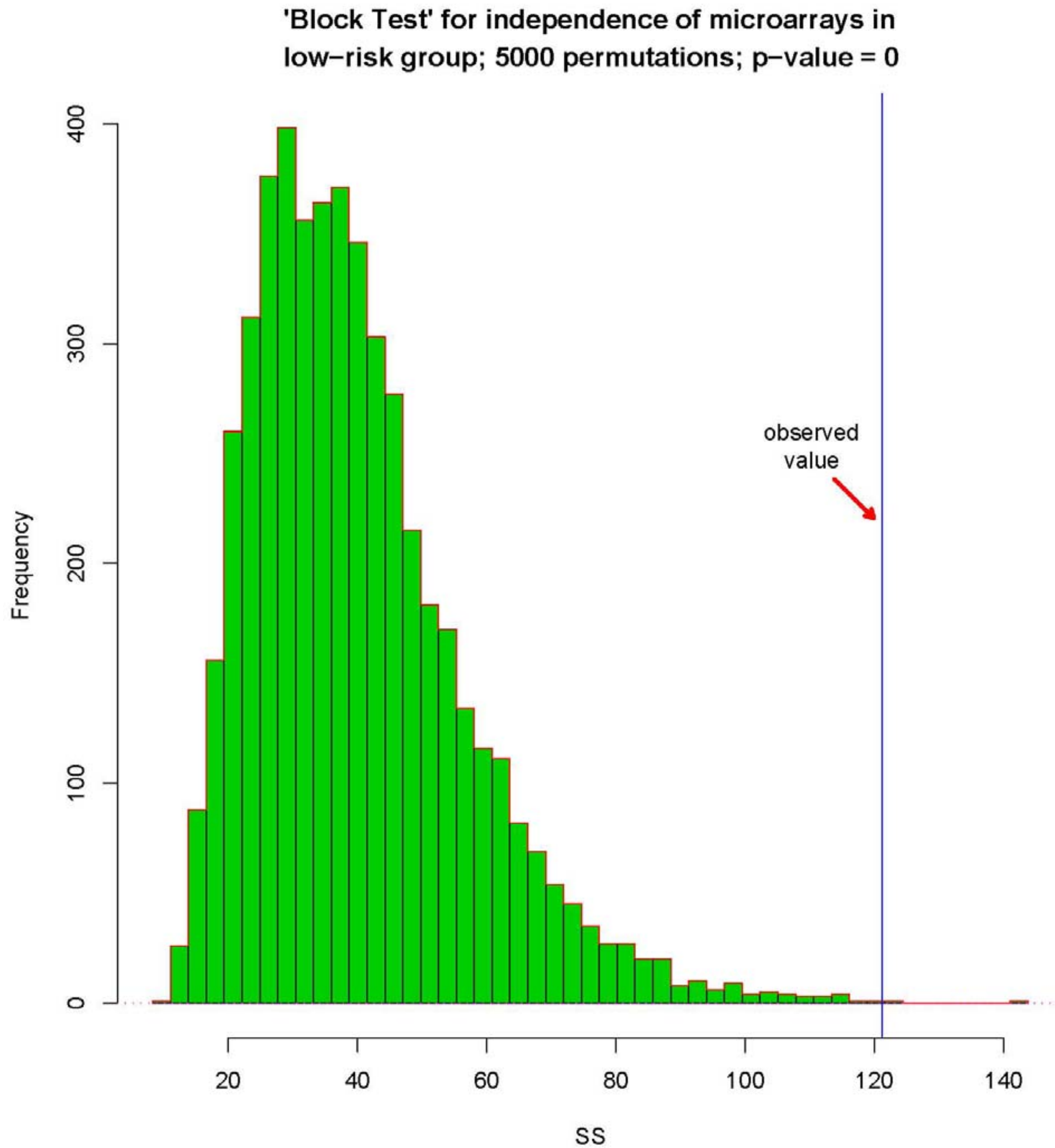
Block Tests

- $S = v_1' \, B \, v_1$ where $B = \sum_k \beta_k \, \beta_k'$,

$$\beta_k' = (0, 0, \cdots, 0, 1, 1, \cdots, 1, 0, 0 \cdots 0)$$

- All blocks, lengths 2 through 10

- Test strongly rejects independence, $p \doteq 2/5000$


'Block Test' for independence of microarrays in
low−risk group; 5000 permutations; p−value = 0
Frequency
0
100
200
300
400
observed
value
20
40
60
80
100
120
140
SS

Row and Column Correlations

- $\underset{m\times n}{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n) = \begin{pmatrix} x_1' \\ \vdots \\ x_m' \end{pmatrix}$
- *Column Correlations* $\widehat{\mathbf{cor}}_{jj'} = \mathbf{x}_j \cdot \mathbf{x}_{j'}$
- *Row Correlations* $\widehat{\mathrm{cor}}_{ii'} = x_i \cdot x_{i'}$
- *Variances* $\mathbf{A}^2 = \sum_{jj'} \widehat{\mathbf{cor}}^2_{jj'}/n^2$ and $A^2 = \sum_{ii'} \widehat{\mathrm{cor}}^2_{ii'}/m^2$

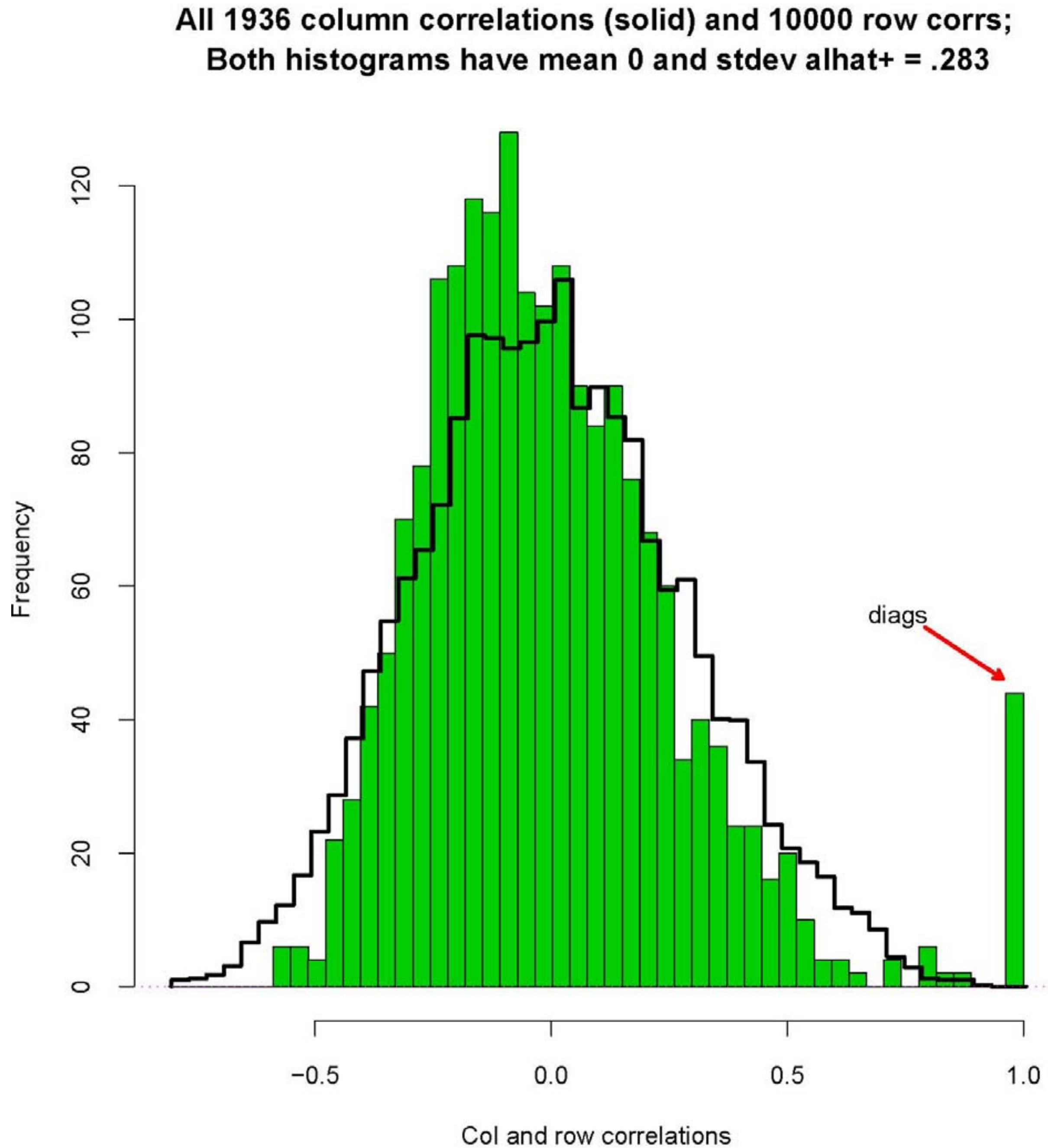
- Doubly standardized
- $\sum_{jj'} \widehat{\mathbf{cor}}_{jj'} = 0$
- $\sum_{ii'} \widehat{\mathrm{cor}}_{ii'} = 0$

A Theorem

Theorem Let $e_1, e_2, \cdots, e_n$ be ordered eigenvalues of $X'X$. Then

$$\boxed{\mathbf{A}^2 = A^2 = \sum_1^n e_k^2/(mn)^2} \qquad (= .283^2)$$

- Column $\widehat{\text{cors}}$ as dispersed as row $\widehat{\text{cors}}$, even if columns (microarrays) independent!



All 1936 column correlations (solid) and 10000 row corrs;
Both histograms have mean 0 and stdev alhat+ = .283
Frequency
0
20
40
60
80
100
120
diags
−0.5
0.0
0.5
1.0
Col and row correlations

The Off-Diagonal Correlations

- The $n(n-1)/2$ off-diagonal column correlations have (mean, variance)

$$\{\widehat{\mathbf{cor}}_{jj'} : j < j'\} \ \sim \ (-\frac{1}{n-1},\ \hat{\alpha}^2)$$

where $\boxed{\hat{\alpha}^2 = \frac{n}{n-1}(A^2 - \frac{1}{n-1}).}$ $[= .241^2]$

- **Total True Row-wise Correlation** defined to be

$$\alpha = \left[\sum_{i<i'} \mathrm{cor}_{ii'}^2 \Big/ \binom{m}{2} \right]^{1/2}$$

- If columns (microarrays) independent then $\hat{\alpha}$ almost unbiased for α.

- Compute all row $\widehat{\mathrm{cor}}_{ii'}$ values
- Shrink to account for sampling variability
   - Shrunken distribution has sample sd $\doteq \hat{\alpha} = .241$.

Is ".805" a Significantly Large Correlation?

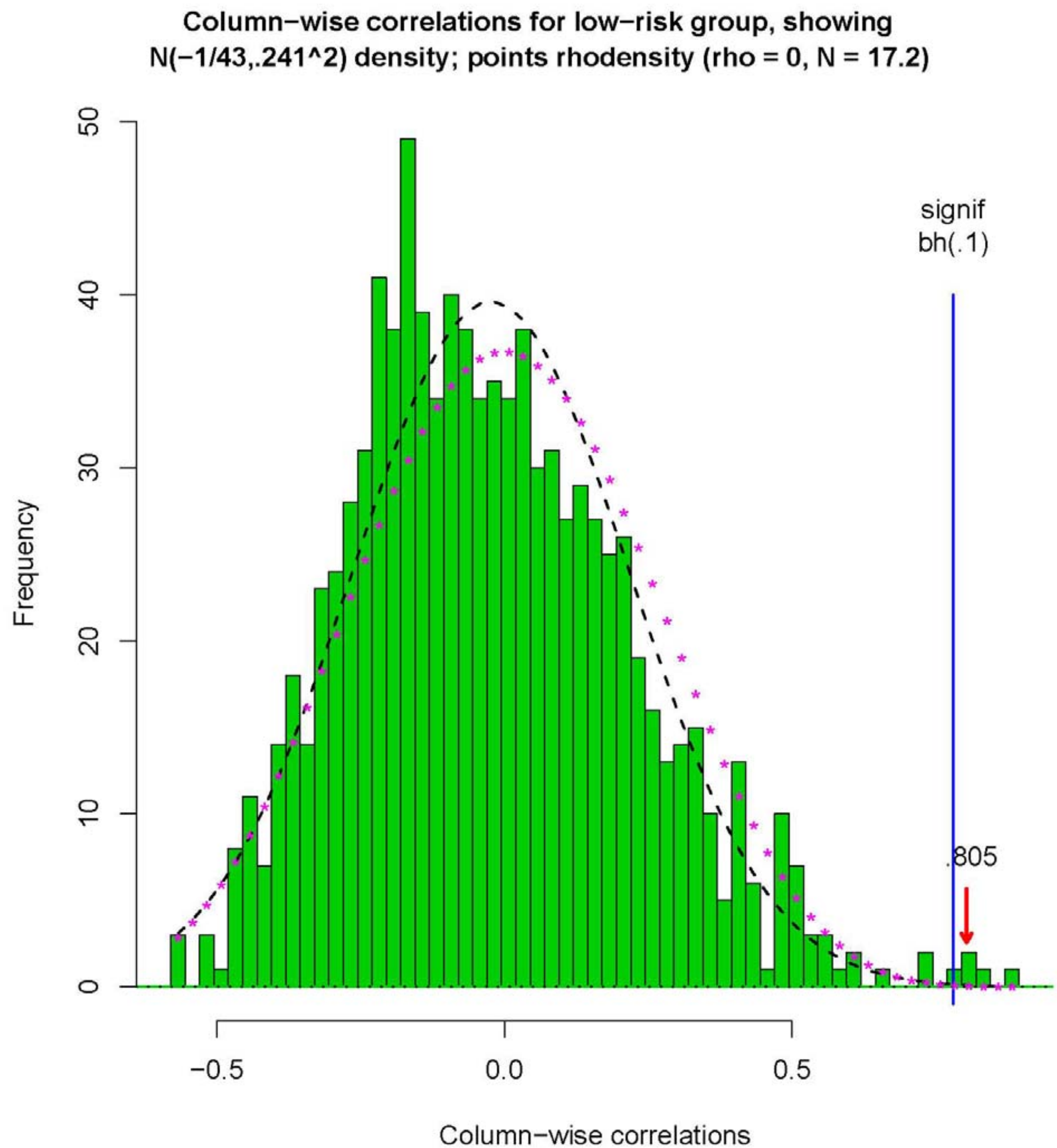
- There are 936 off-diagonal column correlations.

- If columns actually independent, $\alpha \doteq .241$, we expect

$$\{\widehat{\mathbf{cor}}_{jj'} : j < j'\} \;\sim\; (-.023, .241^2)$$

- Benjamini-Hochberg Fdr (.1) test, assuming normality, declares "significant" the 5 pairs with $\widehat{\mathbf{cor}}_{jj'} > .78$ (all 5 from $27:32$).

- *No "Slam Dunk"!* Correlation makes effective sample size much smaller than $m = 20426 : m_{eff} = 17.2$.


Column−wise correlations for low−risk group, showing
N(−1/43,.241^2) density; points rhodensity (rho = 0, N = 17.2)
signif
bh(.1)
.805
Frequency
0
10
20
30
40
50
−0.5
0.0
0.5
Column−wise correlations

Normal Theory

- $\underset{m\times n}{X} \sim N_{m,n}(0, \underset{m\times m}{\not\Sigma} \otimes \underset{n\times n}{\Delta})$ [“Demeaning” makes mean $=0$.]

  - $\begin{cases} \text{rows} & x_i \sim N_n(0, \sigma_{ii}\Delta) \\ \text{columns} & \mathbf{x}_j \sim N_m(0, \Delta_{ii}\not\Sigma) \end{cases}$ (not independent) (”)

- Assume $\sigma_{ii} \equiv 1$, so $E\{x_i x_{i'}'\} = \Delta$

- Sample covariance $\widehat{\Delta} = X'X/m$ has expectation Δ.

- *Independence Hypothesis* $\Delta = I$

Effective Sample Size

- *If rows independent*: $x_i \overset{\text{iid}}{\sim} N(0, \Delta)$ and $\widehat{\Delta} \sim \text{Wishart}\ (\Delta, m)/m$, with mean and covariance

$$\widehat{\Delta} \sim (\Delta, \Gamma/m) \qquad [\Gamma_{jk,lh} = \Delta_{jl}\Delta_{kh} + \Delta_{jh}\Delta_{kl}].$$

Theorem $\widehat{\Delta} \sim (\Delta, \Gamma/m_{\text{eff}})$ where effective sample size is

$$\boxed{m_{\text{eff}} = m/[1 + (m-1)\alpha^2)}$$

- α is the total row-wise correlation
- If $\alpha = .241$ then $m_{\text{eff}} = 17.2$!
- *Wishart Approximation* $\widehat{\Delta} \sim \text{Wishart}\ (\Delta, m_{\text{eff}})/m_{\text{eff}}$.

Johnson and Graybill's Model (1972)

- $y_{ij} = \mu + A_i + B_j + a_i\beta_j + \epsilon_{ij} \begin{cases} a_i \sim N(0, \sigma_a^2) \\ \epsilon_{ij} \overset{\text{ind}}{\sim} N(0, \sigma_\epsilon^2) \end{cases}$

  (generalizes "one df for non-additivity")

- *Remove* row and column means: $x_{ij} = a_i\beta_j + \epsilon_{ij}$

- *rows* $x_i \sim N_n(0, \sigma_\epsilon^2 I + \sigma_a^2 \beta\beta')$

- $\Delta = \sigma_\epsilon^2 I + \sigma_a^2 \beta\beta'$

- H_0 : Independence $\Leftrightarrow \sigma_a^2 = 0$

- *J & G* Likelihood Ratio Test rejects for large values e_1/e_+

- Simulating from $\widehat{\Delta} \sim W(I, m_{\text{eff}})/m_{\text{eff}}$
  strongly rejects independence

Best Linear Test: β Known

- Assume $\widehat{\Delta} \sim (\Delta, \Gamma/m_{\text{eff}})$, $\qquad \Delta = I + \lambda\beta\beta'$

- *Independence Hypothesis* $\quad H_0 : \lambda = 0 \;\; [\Delta = I, \Gamma = 2I\,]$

- *Linear Test Stat* $\quad S_v = v'(\widehat{\Delta} - I)v$

  Best Choice $\quad \boxed{S_\beta = \beta'(\widehat{\Delta} - I)\beta}$

  [maximizes $(E_\lambda S - E_0 S)^2/\text{Var}_0(S)$]

Omnibus Test

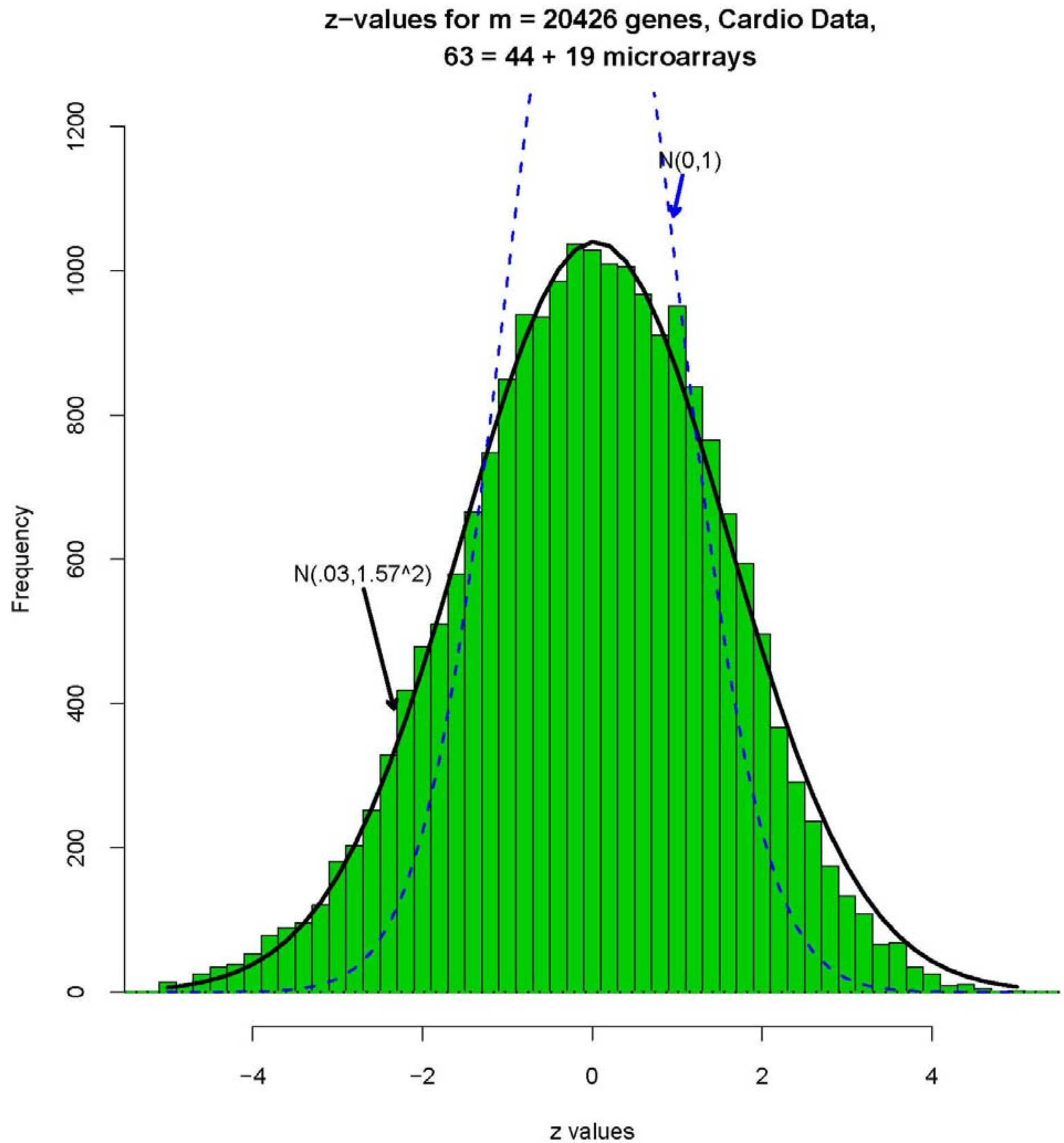
- Don't know "β"
- **Catalog** $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_H)$
- *Omnibus Test*

$$\begin{aligned} \mathbf{S} &= \sum_h \beta_h' (\widehat{\Delta} - I) \beta_h \\ &= \operatorname{tr} \widehat{\Delta} B + \text{constant} \end{aligned}$$

where $B = \sum_h \beta_h \beta_h'$

- If $\widehat{\Delta} = V \text{ diag(e) } V'$ then $\boxed{\mathbf{S} = \sum_{i=1}^{n} e_i v_i' B v_i}$
- *J & G* First eigenvector v_1 of Δ is crucial $\Rightarrow$

$$S = v_1' B v_1, \text{ the "block test"}$$


z−values for m = 20426 genes, Cardio Data,
63 = 44 + 19 microarrays
N(0,1)
N(.03,1.57^2)
Frequency
0
200
400
600
800
1000
1200
−4
−2
0
2
4
z values

Does Correlation Account for Cardio Overdispersion?

- $X \quad 20246 \times (44+19)$
- Contrast $\mathbf{c}' = (-\frac{1}{44}, -\frac{1}{44}, \cdots, \frac{1}{19}, \frac{1}{19}, \cdots)/(\frac{1}{44}+\frac{1}{19})^{1/2}$
- Suppose rows x_i have cov matrix Δ
- $\boxed{z_i = c'x_i = \frac{\bar{x}_{2i}-\bar{x}_{1i}}{(\frac{1}{44}+\frac{1}{19})^{1/2}}}$
- $var(z_i) = c'\Delta c =$ "τ^2"
- *Estimated variance* $\hat{\tau}^2 = c'\widehat{\Delta}c = \Sigma z_i^2/m$
- *Cardio* $\hat{\tau} = \mathbf{1.48}$
- $\hat{\tau}^2 \dot{\sim} \tau^2 \frac{\chi^2_{m_{\text{eff}}}}{m_{\text{eff}}} \quad \Rightarrow \quad \widehat{CV}(\hat{\tau}) = .17$
- $\tau^2 = 1$ if $\Delta = I$

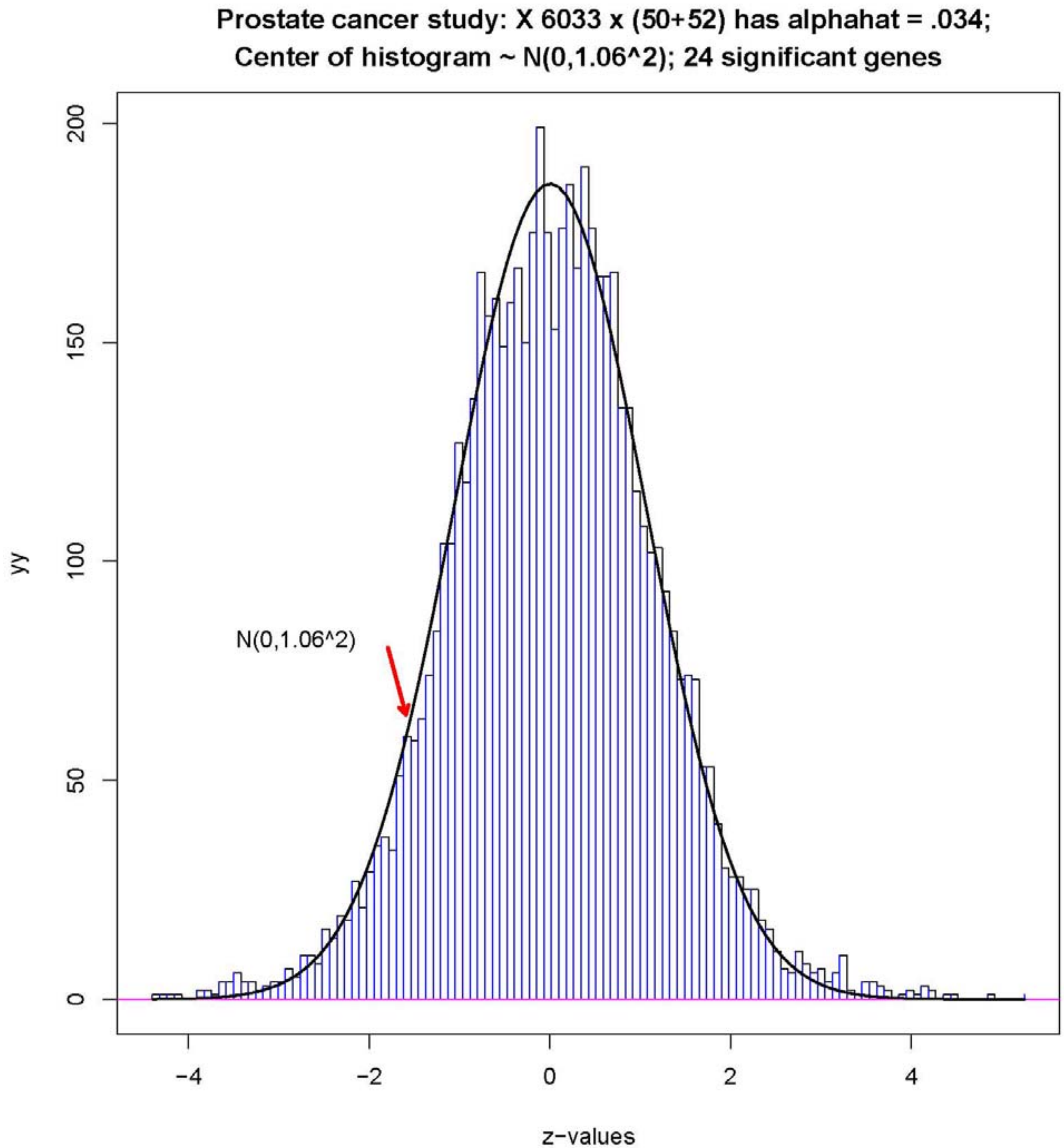
A “Nicer” Microarray Study

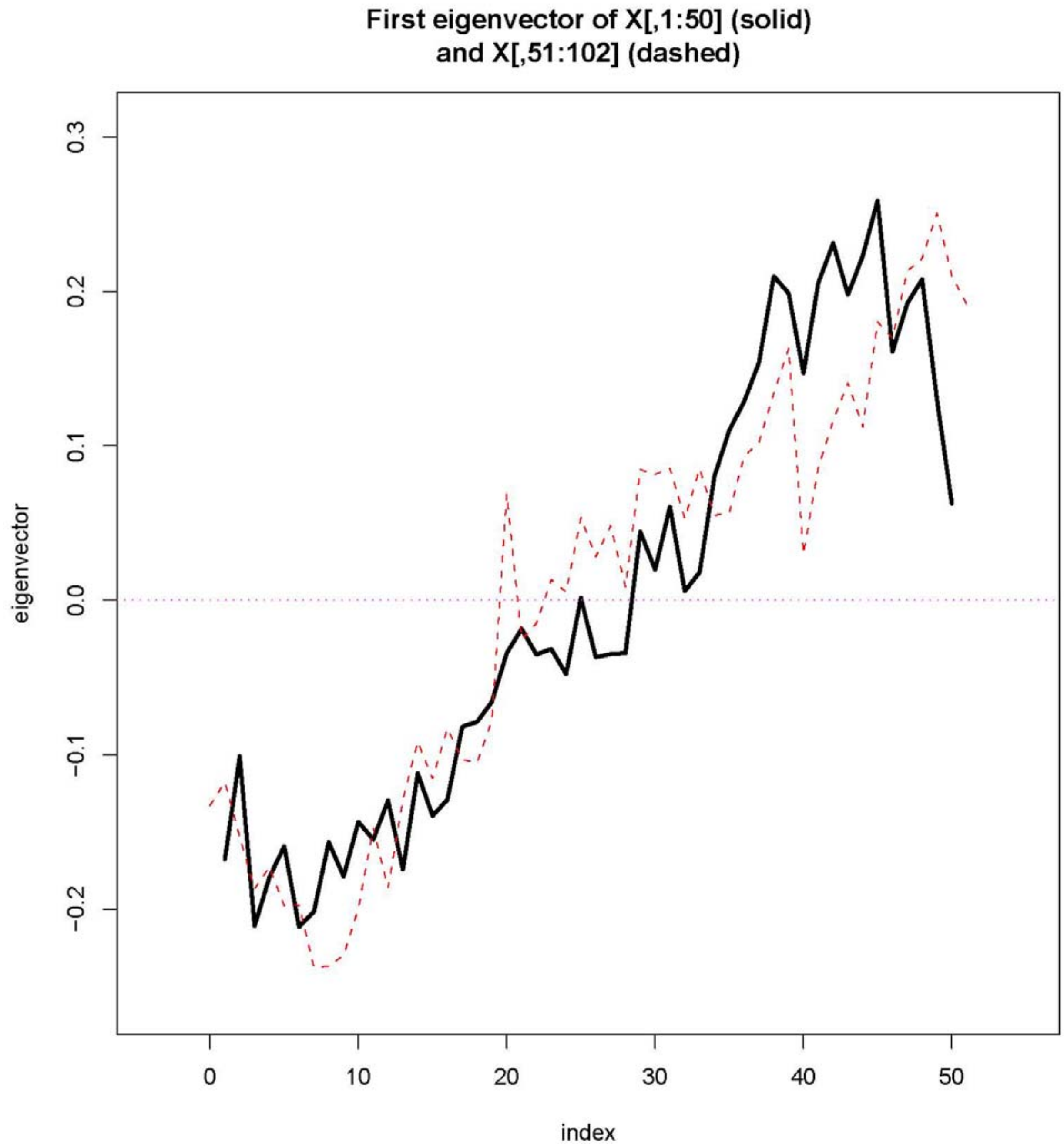
- 102 men, 50 normals, 52 prostate cancer

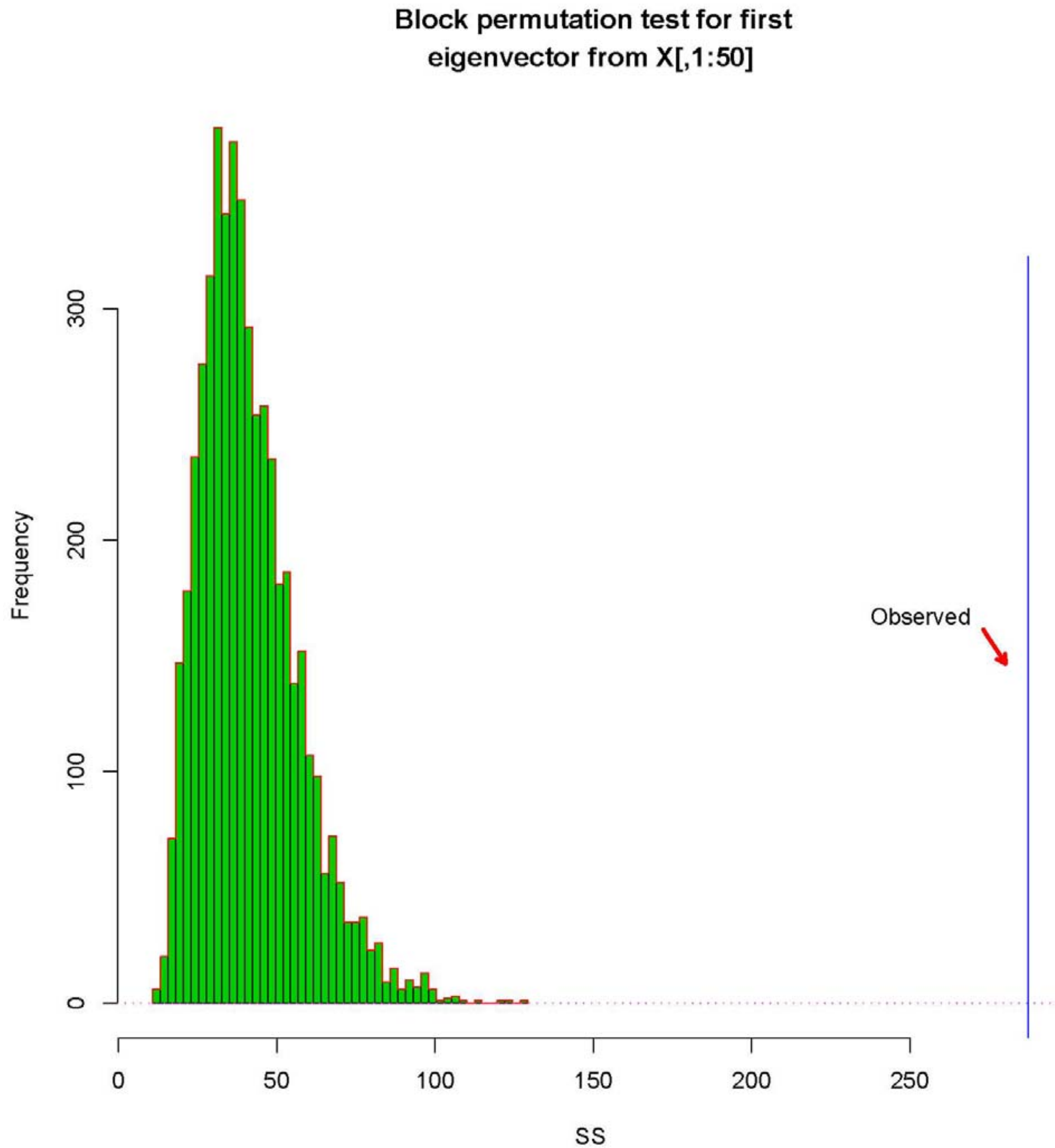
- 6033 genes
- $\underset{6033\times 102}{X}$ has $\hat{\alpha} = .034$

- $z_i = \Phi^{-1}F_{100}(t_i)$ has center of histogram $\dot{\sim}$ $N(0, 1.06^2)$

- 24 low-fdr genes


Prostate cancer study: X 6033 x (50+52) has alphahat = .034;
Center of histogram ~ N(0,1.06^2); 24 significant genes
yy
200
150
100
50
0
N(0,1.06^2)
-4
-2
0
2
4
z-values


First eigenvector of X[,1:50] (solid)
and X[,51:102] (dashed)
eigenvector
index
0.3
0.2
0.1
0.0
−0.1
−0.2
0
10
20
30
40
50


Block permutation test for first
eigenvector from X[,1:50]
Frequency
0
100
200
300
Observed
0
50
100
150
200
250
SS

References

Dallas E. Johnson & Franklin A. Graybill (1972). An Analysis of a Two-Way Model with Interaction and No Replication. *JASA* **67**, 862–868.

Efron (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *JASA* **99**, 96–104.

Efron (2007). Correlation and Large-Scale Simultaneous Significance Testing. *JASA* **102**, 93–103.