Discovery, Validation, and Use of Genomic Markers in Cancer Clinical Trials

Pan American Advanced Studies Institute August 19—20, 2011

Cheng Cheng, Ph.D.
Department of Biostatistics
St. Jude Children's Research Hospital
Memphis, TN, 38105, USA

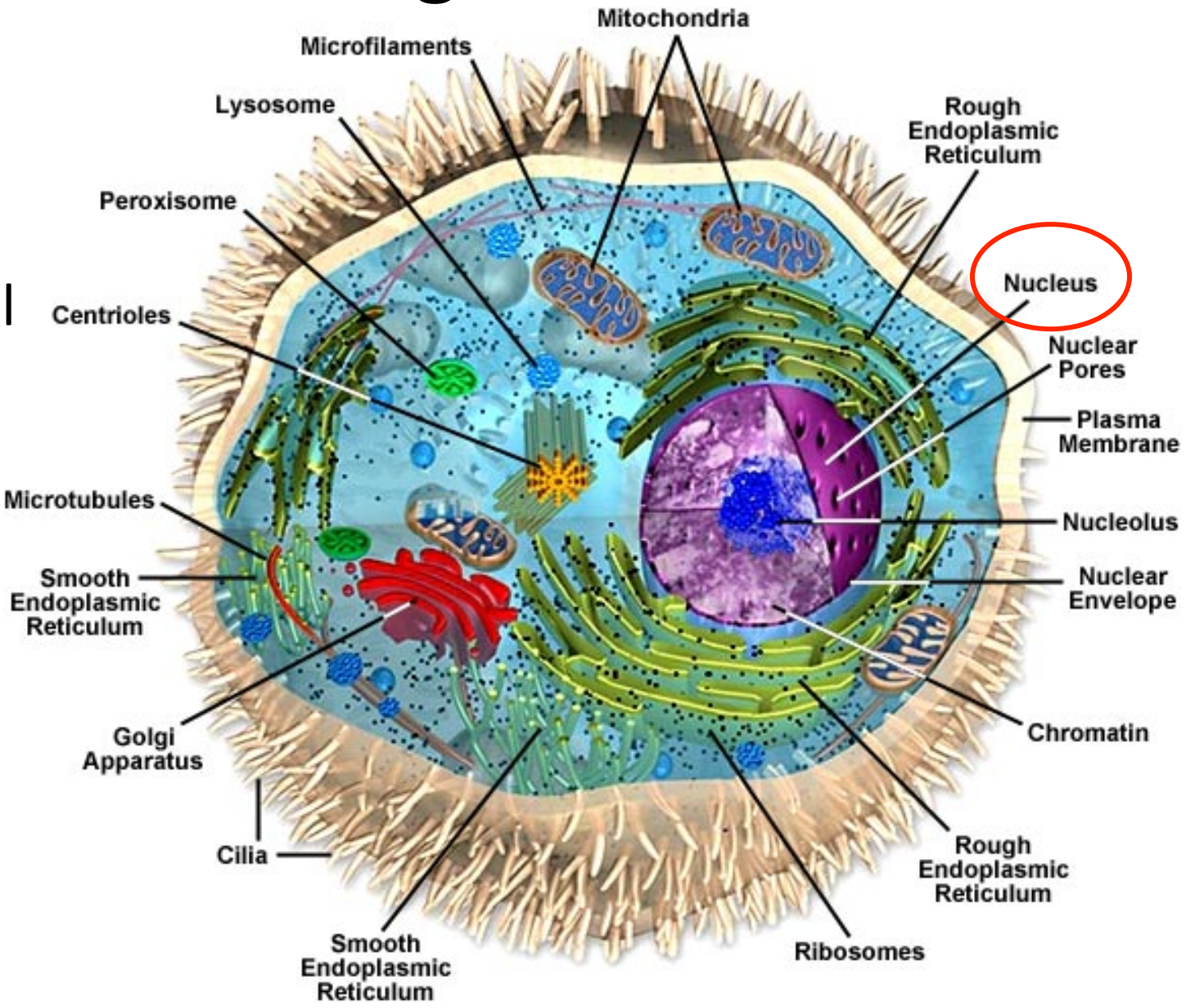
Outline

- Background: Cells, DNA, etc.
- Genomic Markers
- Gene Expression Profiling
- Massive Multiple Hypothesis Tests
- Genome-wide Association Study (GWAS)
- Validation
- [Genomic Classifiers]
- Genomic Markers in Cancer Clinical Trials

Background

- Cell

Animal cell



Background

- *Deoxyribonucleic Acid* (DNA) and *chromosomes*
  - In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called *chromosomes*.
  - Each chromosome is made up of DNA tightly coiled many times around proteins called histones that support its structure.

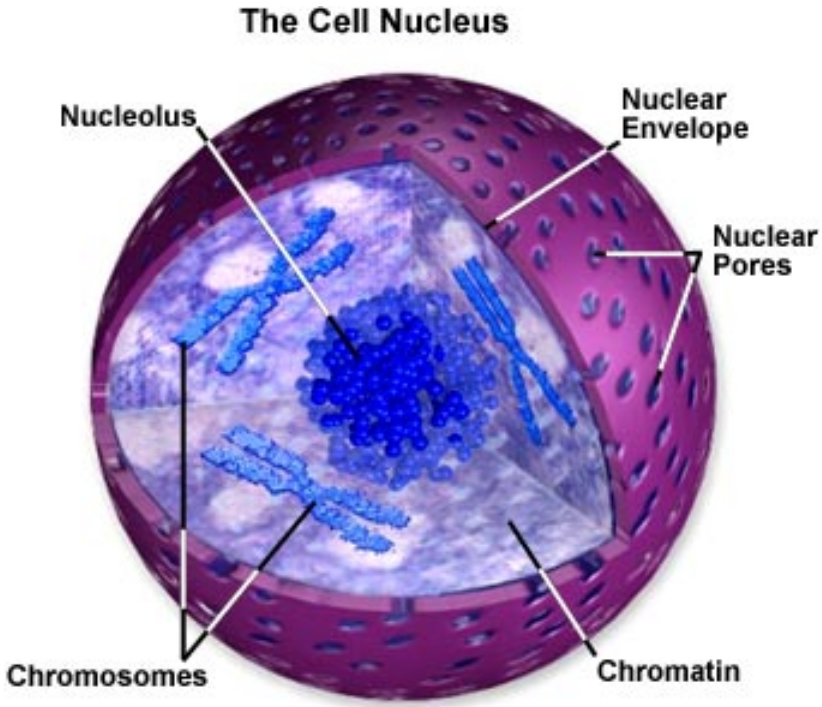
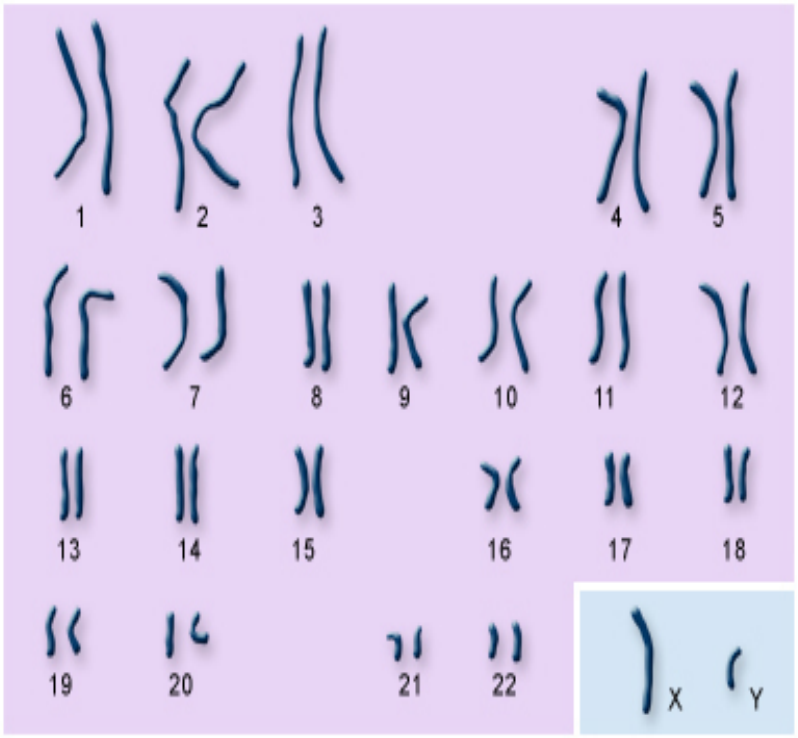
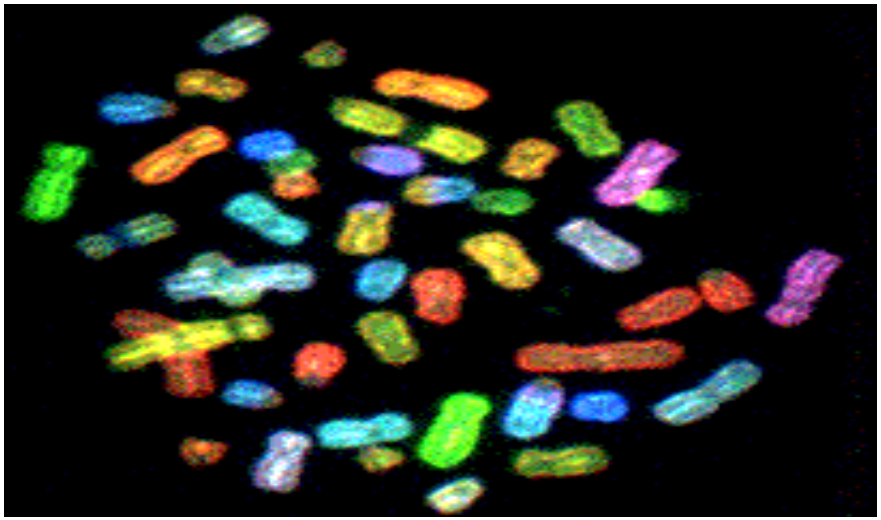


Figure 1

- In humans, each cell normally contains 23 pairs of chromosomes, for a total of 46.
- Twenty-two of these pairs, called autosomes, look the same in both males and females.
- The 23rd pair, the sex chromosomes, differ between males and females.
  - Females have two copies of the X chromosome
  - males have one X and one Y chromosome.



U.S. National Library of Medicine



DNA structure

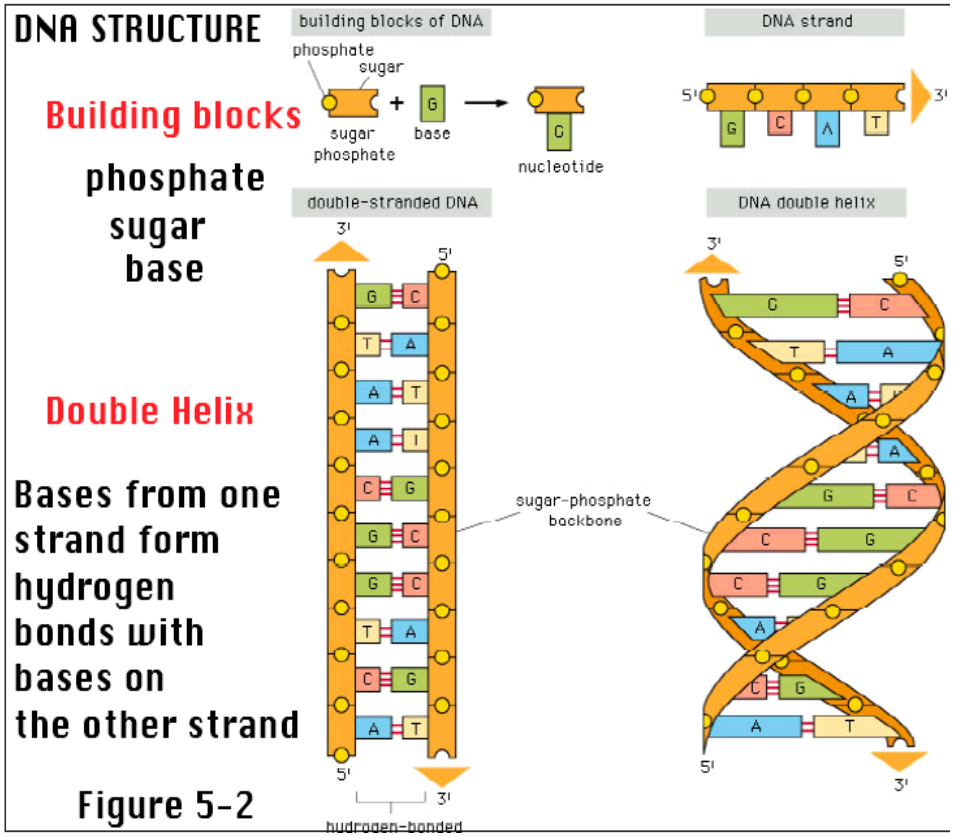


Figure 5-2

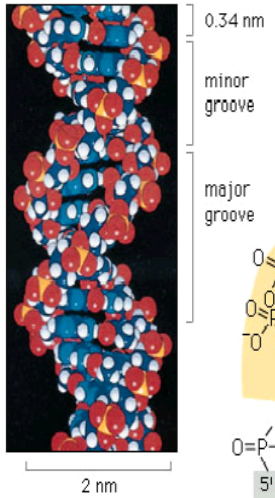


Figure 5-8

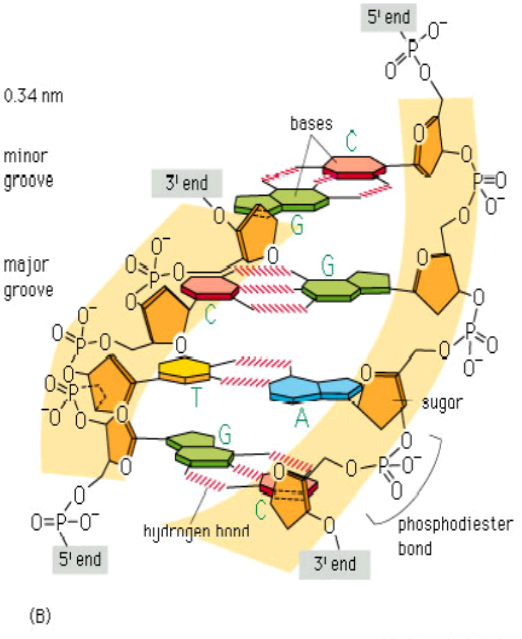
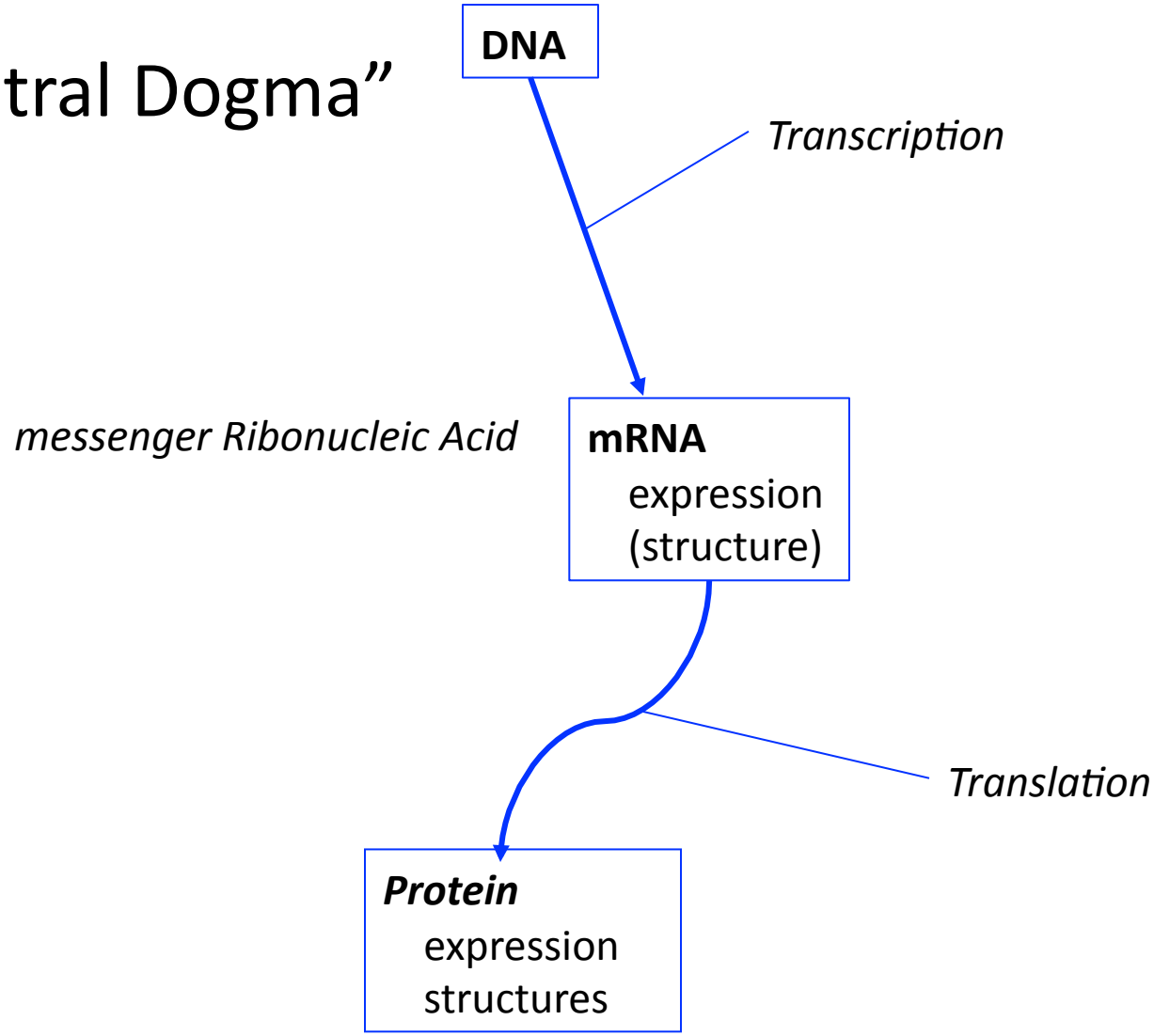


Figure 5-7

DNA is a double-stranded molecule twisted into a helix (think of a spiral staircase).
Each spiraling strand, comprised of a sugar-phosphate backbone and attached bases, is connected to a complementary strand by non-covalent hydrogen bonding between paired bases.
The bases are adenine (**A**), thymine (**T**), cytosine (**C**) and guanine (**G**).

Background

- The “Central Dogma”

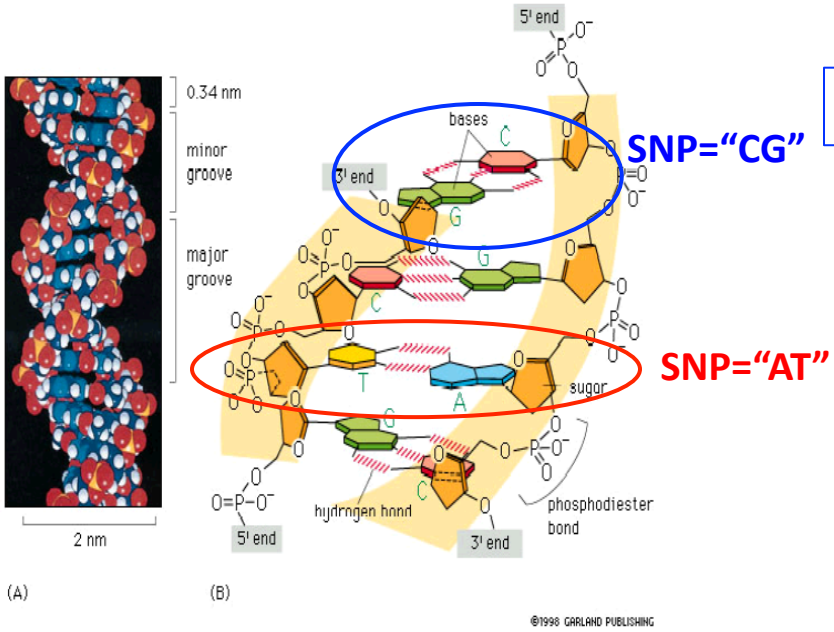


Background

- Gene: a nucleotide (base pair) sequence in DNA coding for a functional product (usually protein and certain types of RNAs)
- -- a segment of DNA that is transcribed into mRNA and then translated into protein which carries out biological function.

Questions?

Genomic Markers



SNP="CG"

SNP="AT"

DNA

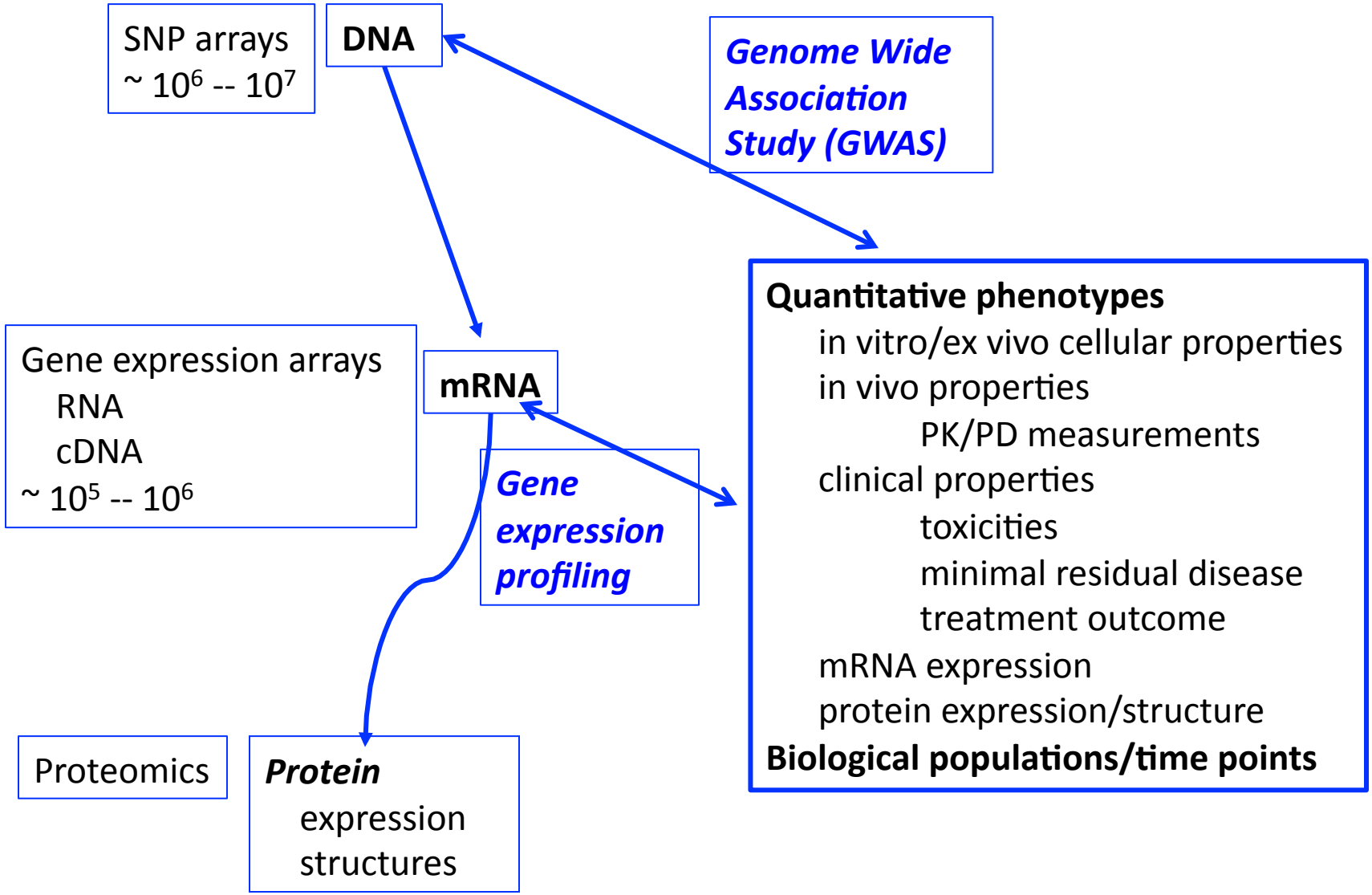
Single nucleotide polymorphism (SNP)

mRNA
expression
(structure)

"gene expression"

Protein
expression
structures

Genomic Markers



Gene Expression Profiling

- Gene Differential Expression and Association: Structures of the data and statistical problems

Each gene expression is a (random) variable – m ~ 10^5 variables
$X_1, X_2, \ldots, X_m$

n subjects, a microarray is run for each subject -- each subject has m observed gene expression data points

Each gene expression has n observations: $X_{i1}, X_{i2}, \ldots, X_{in}$; i=1,2,…,m

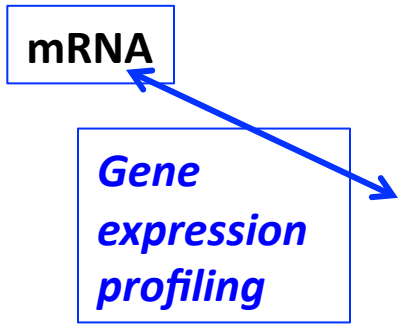
A "phenotype" (biological feature) Y, is a variable observed on each subject
$Y_1, Y_2, \ldots, Y_n$

Statistical problem: Detect the stochastic dependence between $(X_1, Y), \ldots, (X_m, Y)$
Test m (~ 10^5) hypothesis. The i^{th} test uses the data $(X_{i1}, Y_1), \ldots, (X_{in}, Y_n)$; i=1,…,m

-- **A massive multiple hypothesis tests problem**

Gene Expression Profiling

- Test statistics



Quantitative phenotypes
- in vitro/ex vivo cellular properties
- in vivo properties
  - PK/PD measurements
- clinical properties
  - toxicities
  - minimal residual disease
  - treatment outcome
- mRNA expression
- protein expression/structure

Biological populations/time points

Detect the stochastic dependence between $(X_1, Y), \ldots, (X_m, Y)$
Test m (~ 10^5) hypothesis. The ith test uses the data $(X_{i1}, Y_1), \ldots, (X_{in}, Y_n)$; i=1,…,m

ANOVA, t, Wilcoxon, regression coefficients, ranks statistics, Bayesian, and variants…

Gene Expression Profiling

Gene	1	2	…	g	…	m
Hypo.	(H_{01}, H_{A1})	(H_{02}, H_{A2})		(H_{0g}, H_{Ag})		(H_{0m}, H_{Am})
Test Stat.	S_1	S_2	…	S_g	…	S_m
P values	P_1	P_2	…	P_g	…	P_m

	Declared Diff. Expr.	Declared not Diff. Expr.	Total
True H_0	V	U	m_0
True H_A	$R - V$	$m - R - U$	m_1
Total	R	$m - R$	m

Massive Multiple Hypothesis Tests

Detect the stochastic dependence between $(X_1, Y), \ldots, (X_m, Y)$
Test m ($\sim 10^5$) hypothesis.
-- **A massive multiple hypothesis tests problem**

	Reject H_0	Fail to reject H_0	Tot
True H_0	V	U	m_0
True H_A	R-V	m-R-U	m_1
	R	m-R	m

Benjamini and Hochberg (1995) *JRSS-B*, **57**, 289-300

Significance criteria: FWER, gFWER, FDR/q-value, local FDR, adaptive significance threshold

Massive Multiple Hypothesis Tests

	Reject H_0	Fail to reject H_0	Tot
True H_0	V	U	m_0
True H_A	R-V	m-R-U	m_1
	R	m-R	m

Family-Wise Error Rate (FWER) and generalized FWER (gFWER)

FWER = Pr(V>0)

gFWER = Pr(V≥k|R>0) for a specified k

Pr(V/R≥γ|R>0) for a specified γ

Van der Laan, Dudoit, Pollard (2004) *SAGMB* **3**, Article 15

//www.bepress.com/sagmb/vol13/iss1/art15

Ramano and Shaikah (2006) *Annals of Statistics*

Massive Multiple Hypothesis Tests

	Reject H_0	Fail to reject H_0	Tot
True H_0	V	U	m_0
True H_A	R-V	m-R-U	m_1
	R	m-R	m

False discovery rate (FDR)

$$FDR = E(V/R|R>0)Pr(R>0)$$

Benjamini and Hochberg (1995) *JRSS-B*, **57**, 289-300

Massive Multiple Hypothesis Tests

False discovery rate (FDR)

FDR = $E(V/R|R>0)Pr(R>0)$

Benjamini and Hochberg (1995) *JRSS-B*, **57**, 289-300

FDR Control: Sig. criteria assuring FDR≤η

$P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ ordered P values

$H_{[01]}, H_{[02]}, \ldots, H_{[0m]}$ corresponding sequence of H_0's

$P^*_{(i)} = P_{(i)}/(i/m)$ ("adjusted" P values)

$i^* = \max\{ j: P^*_{(j)} \leq \eta \}$

Reject the null hypotheses $H_{[01]}, \ldots, H_{[0i^*]}$

- Can improve power if m is replaced by m_0 (# of H_0's) in $P^*_{(i)}$
- The number of null hypotheses m_0 is unknown!

Massive Multiple Hypothesis Tests

q-*values* pFDR=E(V/R|R>0)

$$P_1,\ldots,P_m$$

$$\widetilde{F}(t)=\frac{1}{m}\sum_{i=1}^{m} I(P_i \le t), \quad t\in[0,1]$$

$$F(t):=\mathrm{E}[\widetilde{F}(t)]=\frac{m_0}{m}t+\frac{1}{m}\sum_j F_j(t)=\pi_0 t+\frac{1}{m}\sum_{j\in A} F_j(t); \quad \pi_0=\frac{m_0}{m}$$

Rejection rule - - Reject H_{0i} if $P_i \le \alpha$. Then for small α

$$FDR(\alpha)\approx\frac{\pi_0\alpha}{F(\alpha)}$$

$$r_i=\frac{\hat{\pi}_0 P_{(i)}}{\widetilde{F}(P_{(i)})}, i=1,2,\ldots,m$$

$$q_i=\min\{r_j, j\ge i\}, i=1,2,\ldots,m$$

Rejecting all the corresponding H0's with q ≤η controls *the positive FDR* (pFDR) at level η.

Storey (2002) *JRSS-B*, **64**, 479-498
Storey, Taylor, Siegmund (2003) *JRSS-B*, **66**, 187-205
Storey (2003) *Annals of Statistics*, **31**, 2103-2035

Massive Multiple Hypothesis Tests

FDR estimation

$P_1,\ldots,P_m$

$$\widetilde{F}(t) = \frac{1}{m}\sum_{i=1}^{m} I(P_i \le t), \quad t \in [0,1]$$

$$F(t) := \mathrm{E}[\widetilde{F}(t)] = \frac{m_0}{m}t + \frac{1}{m}\sum_j F_j(t) = \pi_0 t + \frac{1}{m}\sum_{j\in A} F_j(t); \quad \pi_0 = \frac{m_0}{m}$$

Rejection rule - - Reject H_{0i} if $P_i \le \alpha$. Then for small α

$$FDR(\alpha) \approx \frac{\pi_0\alpha}{F(\alpha)} \qquad FDR\hat{}(\alpha) \approx \frac{\hat{\pi}_0\alpha}{\hat{F}(\alpha)}$$

Storey (2002) *JRSS-B*, **64**, 479-498

Cheng et al. (2004) *SAGMB*, **3**, Article 36. www.bepress.com/sagmb/vol13/iss1/art36

Pounds and Cheng (2003) *Bioinformatics*, **20**, 1737-1745

Pounds and Cheng (2006) *Bioinformatics*, **22**, 1979-1987

Cheng (2006) *Optimality, The Second Lehmann Symposium* 77-99

Hunt, Cheng, Pounds (2009) *Comput. Statist. Data Anal.* **53**, 1688-1700

Allison et al (2002) *Comput. Statist. Data Anal.* **39**, 1-20

Massive Multiple Hypothesis Tests

FDR estimation: The estimator $\hat{\pi}_0$ $\pi_0 = \frac{m_0}{m}$ the proportion of true null hypotheses

$P_1, \ldots, P_m$

$$\widetilde{F}(t) = \frac{1}{m}\sum_{i=1}^{m} I(P_i \le t), \quad t \in [0,1]$$

$$F(t) := \mathrm{E}[\widetilde{F}(t)] = \frac{m_0}{m}t + \frac{1}{m}\sum_{j} F_j(t) = \pi_0 t + \frac{1}{m}\sum_{j \in A} F_j(t); \quad \pi_0 = \frac{m_0}{m}$$

$$F(t) = \pi_0 t + \frac{m - m_0}{m}\frac{1}{m - m_0}\sum_{j \in A} F_j(t) = \pi_0 t + (1 - \pi_0)F_A(t);$$

$$F_A(t) = \frac{1}{m - m_0}\sum_{j \in A} F_j(t)$$

$$\pi_0 = \frac{F_A(t) - F(t)}{F_A(t) - t} \le \frac{1 - F(t)}{F_A(t) - t} \approx \frac{1 - F(t)}{1 - t} \quad \text{for } t \text{ close to } 1.$$

$$\hat{\pi}_0 = \frac{1 - \hat{F}(\lambda)}{1 - \lambda} \quad \text{for some chosen } \lambda \text{ -- the "slope estimator"}$$

-- slope between the points $(\lambda, \hat{F}(\lambda))$ and $(1,1)$

Massive Multiple Hypothesis Tests

FDR estimation: The estimator $\hat{\pi}_0$ $\pi_0 = \frac{m_0}{m}$ the proportion of true null hypotheses

$P_1, \ldots, P_m$

$$\widetilde{F}(t) = \frac{1}{m}\sum_{i=1}^{m} I(P_i \le t), \quad t \in [0,1]$$

$$F(t) := \mathrm{E}[\widetilde{F}(t)] = \frac{m_0}{m}t + \frac{1}{m}\sum_{j} F_j(t) = \pi_0 t + \frac{1}{m}\sum_{j \in A} F_j(t); \quad \pi_0 = \frac{m_0}{m}$$

$$F(t) = \pi_0 t + \frac{m - m_0}{m}\frac{1}{m - m_0}\sum_{j \in A} F_j(t) = \pi_0 t + (1 - \pi_0)F_A(t);$$

$$F_A(t) = \frac{1}{m - m_0}\sum_{j \in A} F_j(t)$$

$$f(t) = \pi_0 + (1 - \pi_0) f_A(t)$$

π_0 is identifiable and $\pi_0 = f(1)$ iff $f_A(1) = 0$ for monotone f.

$\hat{\pi}_0 = \hat{f}(1)$

References in previous slide, plus

Langaas et al. (2005) *JRSS-B*, **67**, 555-572

SUMMARY & additional references

$$\pi_0 = \frac{m_0}{m}; \quad \hat{FDR}(\alpha) = \frac{\hat{\pi}_0 \alpha}{\hat{F}(\alpha)}$$

Benjamini & Hochberg (2000) *JEBS*: Adaptive FDR control

Estimate π_0 by a graphical (simple slope) method;
use the EDF for *F*.

Storey, Taylor, Siegmund (2003) *JRSS-B*

FDR control

Storey (2002) *JRSS-B*: Estimate π_0 by an empirical pdf (histogram/spacing);
use the EDF for *F*.

Pounds & Morris (2003) *Bioinformatics*: Beta-Uniform mixture model (BUM)

Estimate π_0 and beta parameters by constrained maximum likelihood

Cheng, Pounds et al. (2004): *Stat. Appl. Genetics Mol. Bio.*

Estimate *F* and *f*=*F'* by adaptive smoothing (B-spline series) under stochastic order constraint, and estimate π_0 by the minimum of *f*.

Pounds & Cheng (2004) *Bioinformatics*

Estimate *F* and *f* by adaptive smoothing (LOESS), and estimate π_0 by the minimum of *f*.

Hunt, Cheng, Pounds (2009) *Comp. Stat. Data. Anal.*

FDR estimation for substantially dependent tests

Langaas et al. (2005) *JRSS-B*

Massive Multiple Hypothesis Tests

- Local FDR

$$\Pr(H_0) = \pi_0; \Pr(H_A) = 1 - \pi_0$$

$$g(\cdot): \text{ pdf of test statistic}$$

$$g_0(\cdot): \text{ conditional pdf of test statistic given } H_0$$

$$g_1(\cdot): \text{ conditional pdf of test statistic given } H_A$$

$$g(x) = \pi_0 g_0(x) + (1 - \pi_0) g_1(x),\ x \in R$$

The empirical Bayes posterior probability of H_0 given test statistic S:

$$EBP(H_0 \mid S) = locFDR(S) = \frac{\hat{\pi}_0 \hat{g}_0(S)}{\hat{g}(S)}$$

Efron et al. (2001) *JASA* **96**, 1151-1160
Efron (2004) *JASA* **99**, 96-104

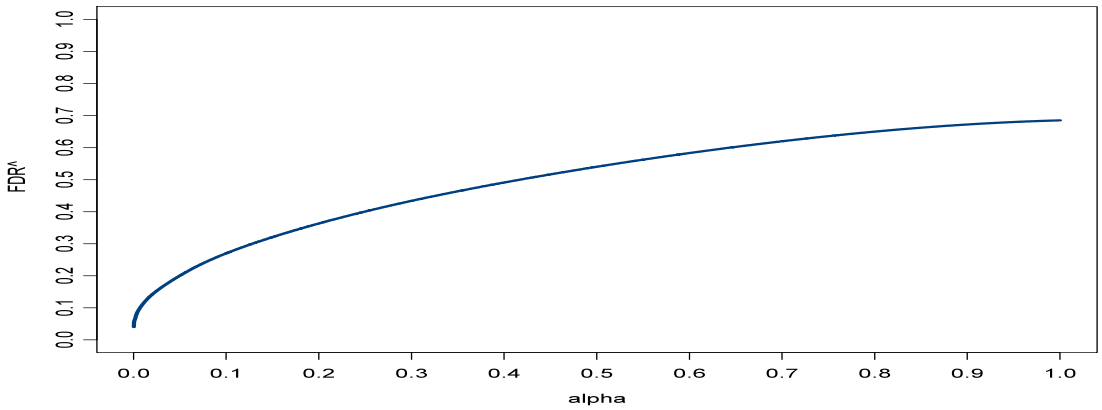
Questions?

Massive Multiple Hypothesis Tests

Significance criteria: adaptive significance threshold

- Previous criteria are concerned only with false positive errors
  - FDR control – control at what level?
  - FDR estimation – estimate at which α?
- Flip side: false negative errors are equally important in large microarray experiments
- Consider determination of significance threshold (P value cutoff) adaptively

$$F\hat{D}R(\alpha) = \frac{\hat{\pi}_0 \alpha}{\hat{F}(\alpha)}$$



FDR control:

For a preset FDR level η, find the largest possible alpha, α^*, such that

$$F\hat{D}R(\alpha^*) \leq \eta$$

Reject all nulls with P value ≤ α^*.

(1) Controlling FDR addresses only one type of error because it ignores the proportion of false negatives.

(2) In practice, there may not be much clue for how to balance between the FDR level and the statistical significance level – which alpha level?

Outline

- Background: Cells, DNA, etc.
- Genomic Markers
- Gene Expression Profiling
- Massive Multiple Hypothesis Tests
- Genome-wide Association Study (GWAS)
- Validation
- [Genomic Classifiers]
- Genomic Markers in Cancer Clinical Trials

Massive Multiple Hypothesis Tests

Detect the stochastic dependence between $(X_1, Y), \ldots, (X_m, Y)$
Test m ($\sim 10^5$) hypothesis.
-- **A massive multiple hypothesis tests problem**

	Reject H_0	Fail to reject H_0	Tot
True H_0	V	U	m_0
True H_A	R-V	m-R-U	m_1
	R	m-R	m

Benjamini and Hochberg (1995) *JRSS-B*, **57**, 289-300

Significance criteria: FWER, gFWER, FDR/q-value, local FDR, adaptive significance threshold

Massive Multiple Hypothesis Tests

Significance criteria: adaptive significance threshold

- *Profile Information Criteria* (I_p): Cheng *et al*. (2004), Cheng (2006)

Provide a guide for determination of significance level complementary to FDR control/estimation

Borrow idea from model selection criteria.

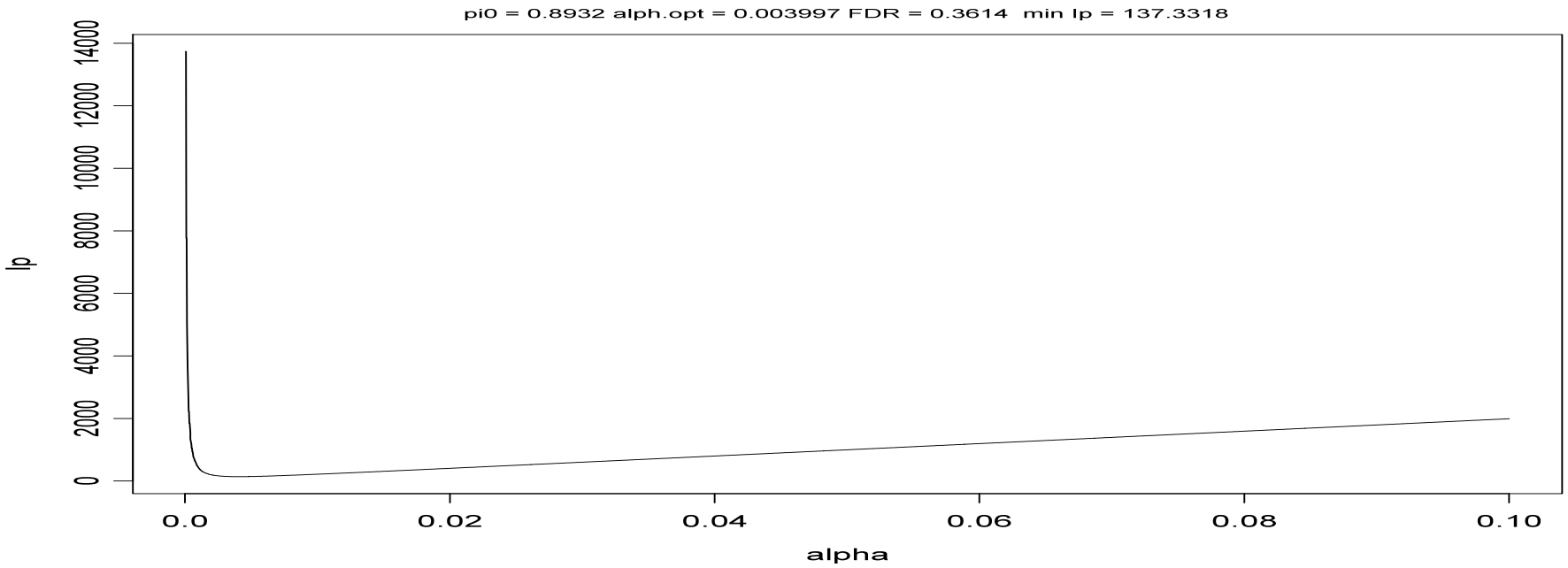
Form a criterion function with two terms: a term represents the "significance" (i.e., small magnitude of the P values) in terms of a deviation from uniformity, and a penalty term monotonically increasing in the expected number of false positives.

$$I_P(\alpha) = \frac{1}{\widetilde{D}(\alpha)} + \lambda\hat{\pi}_0 m\alpha$$

The first term measures the **amount of aggregate evidence against nulls (supporting alternative hypotheses) as reflected by the small P values; it is non-increasing in α: $\alpha\uparrow$, $D(\alpha)\uparrow$, $1/D(\alpha)\downarrow$**

The second term ↑ in α and penalizes with the estimated number of false discoveries from rejecting too many nulls (calling too many genes differentially expressed).

The optimal significance threshold is α^* that minimizes $I_p(\alpha)$. Reject all nulls with P values ≤ α^*.


pi0 = 0.8932 alph.opt = 0.003997 FDR = 0.3614 min lp = 137.3318
lp
0
2000
4000
6000
8000
10000
12000
14000
0.0
0.02
0.04
0.06
0.08
0.10
alpha

$$I_P(\alpha) = \frac{1}{\widetilde{D}(\alpha)} + \lambda \hat{\pi}_0 m \alpha$$

$$\widetilde{D}(\alpha) = \{ m \int_0^{\alpha} [t - \widetilde{Q}(t)]_+^2 dt \}^{1/2}$$

λ

A recent development: Use an AUC of ROC type statistic for the first term
-- work in progress

For $0 \leq t \leq 1$,

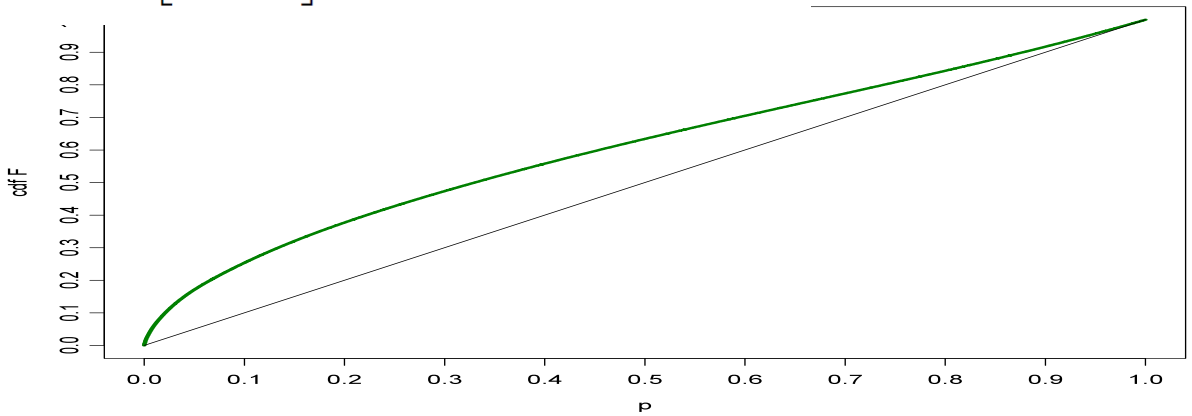
$$\widetilde{F}_m(t) = \frac{1}{m} \sum_{i=1}^{m} I\left(P_i \leq t\right)$$

$$F(t) = \mathbf{E}\left[\widetilde{F}_m(t)\right]$$

$$F(t) = \pi_0 t + (1-\pi_0)\frac{1}{m_1} \sum_{i \in \mathcal{H}_1} F_{1i}(t) = \pi_0 t + (1-\pi_0) F_1(t)$$

$$F_1(t) = \frac{1}{m_1} \sum_{i \in \mathcal{H}_1} F_{1i}(t)$$

$$F(t) = \mathbf{E}\left[\widetilde{F}_m(t)\right] = \pi_0 t + (1-\pi_0)F_1(t) \;\; 0 \le t \le 1$$

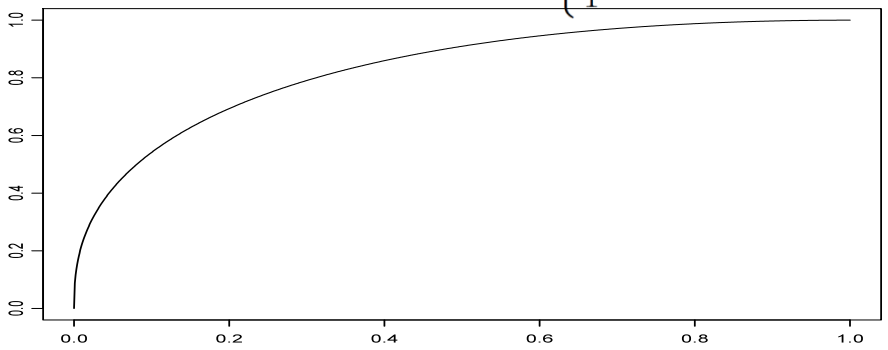


τ_0

t

$$\tau_0 = \begin{cases} \operatorname{argmax}\{F(t) - t,\ 0 \le t \le 1\} & \text{if } \pi_0 < 1 \\ 1 & \text{if } \pi_0 = 1 \end{cases}$$

$$F(t) = \mathbf{E}\left[\widetilde{F}_m(t)\right] = \pi_0 t + (1-\pi_0)F_1(t) \qquad \tau_0 = \begin{cases} \operatorname{argmax}\{F(t)-t,\ 0 \le t \le 1\} & \text{if } \pi_0 < 1 \\ 1 & \text{if } \pi_0 = 1 \end{cases}$$



F_1

α

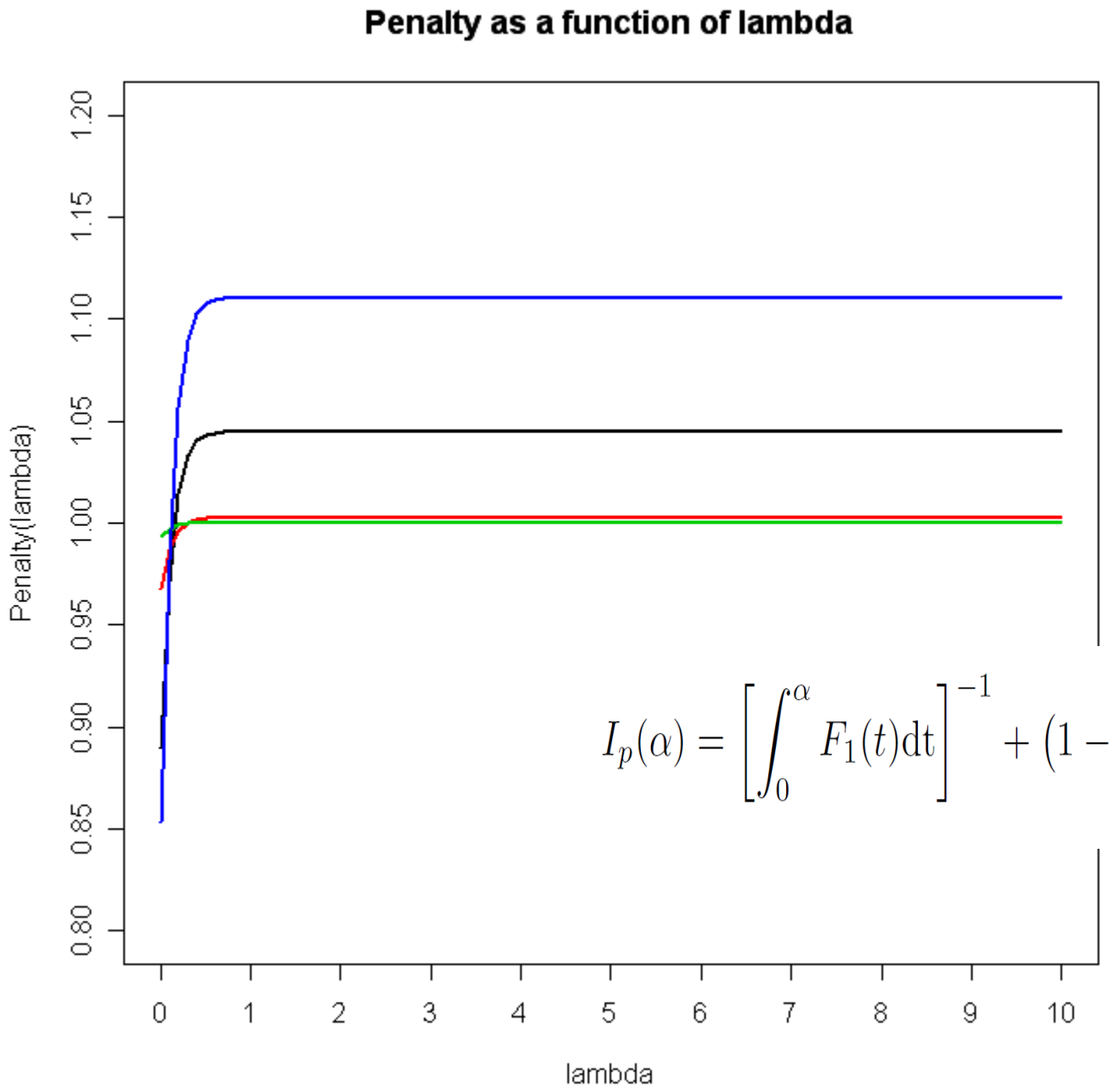
t

A Receiver Operating Characteristic (ROC) Curve

AUC measures performance

$$I_p(\alpha) = \left[\int_0^{\alpha} F_1(t)\mathrm{dt}\right]^{-1} + \left(1 - \pi_0\tau_0 + m^{-\lambda}\right)^{-\pi_0\tau_0} m\pi_0\alpha$$

$$\alpha^* = \operatorname{argmin}\{I_p(\alpha),\ 0 < \alpha \le \bar{\alpha}_0\}$$



m=25,000
π_0=0.99

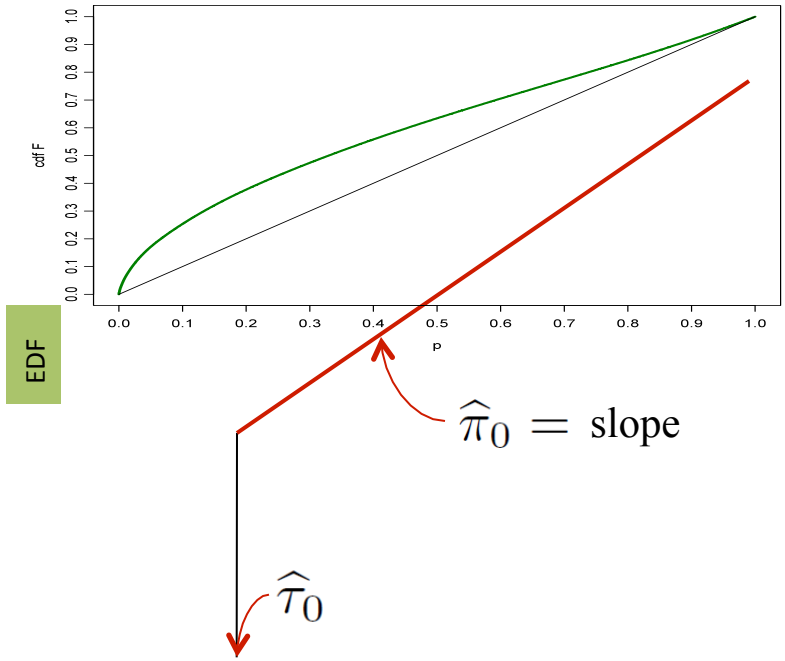
tau0=0.3: blue
tau0=0.2: black
tau0=0.05: red
tau0=0.01: green

$$I_p(\alpha) = \left[\int_0^{\alpha} F_1(t)\mathrm{dt}\right]^{-1} + \left(1 - \pi_0\tau_0 + m^{-\lambda}\right)^{-\pi_0\tau_0} m\pi_0\alpha$$

$$\widehat{I}_p(\alpha) = \left[\int_0^{\alpha} \widehat{F}_1(t)\mathrm{dt}\right]^{-1} + \left(1 - \widehat{\pi}_0\widehat{\tau}_0 + m^{-\lambda}\right)^{-\widehat{\pi}_0\widehat{\tau}_0} m\widehat{\pi}_0\alpha$$

$$\widehat{\tau}_0 = \operatorname{argmax}\left\{\widetilde{F}_m(t) - t,\ 0 \le t \le 1\right\}$$

$$\widehat{\pi}_0 = \frac{1 - \widetilde{F}_m(\widehat{\tau}_0)}{1 - \widehat{\tau}_0}$$



$$\widehat{F}_1(t) = \min\left\{\frac{\widetilde{F}_m(t) - \widehat{\pi}_0 t}{1 - \widehat{\pi}_0},\ 1\right\}$$

$$F(t) = \pi_0 t + (1 - \pi_0)F_1(t)$$

λ ?!

λ

$$Q_1(u) = F_1^{-1}(u), \text{ so for small } u$$

$$Q_1(u) = Q_1(0) + q_1(0)u + \frac{1}{2}q_1'(0)u^2 + \frac{1}{6}q_1''(\xi)u^3$$

$$F_1(t) = \beta t^{\delta} \qquad \text{for small } t$$

$$\beta \geq 1,\ 0 < \delta \leq 1$$

$$I_p(\alpha) = \left[\int_0^{\alpha} F_1(t)\mathrm{dt}\right]^{-1} + \left(1 - \pi_0\tau_0 + m^{-\lambda}\right)^{-\pi_0\tau_0} m\pi_0\alpha$$

$$\alpha^* = [\beta(1+\delta)]^{\frac{1}{\delta+2}} \left(1 - \pi_0\tau_0 + m^{-\lambda}\right)^{\frac{\pi_0\tau_0}{\delta+2}} \pi_0^{-\frac{1}{\delta+2}} m^{-\frac{1}{\delta+2}}$$

Complete uniformity: $\pi_0 = \tau_0 = \beta = \delta = 1$

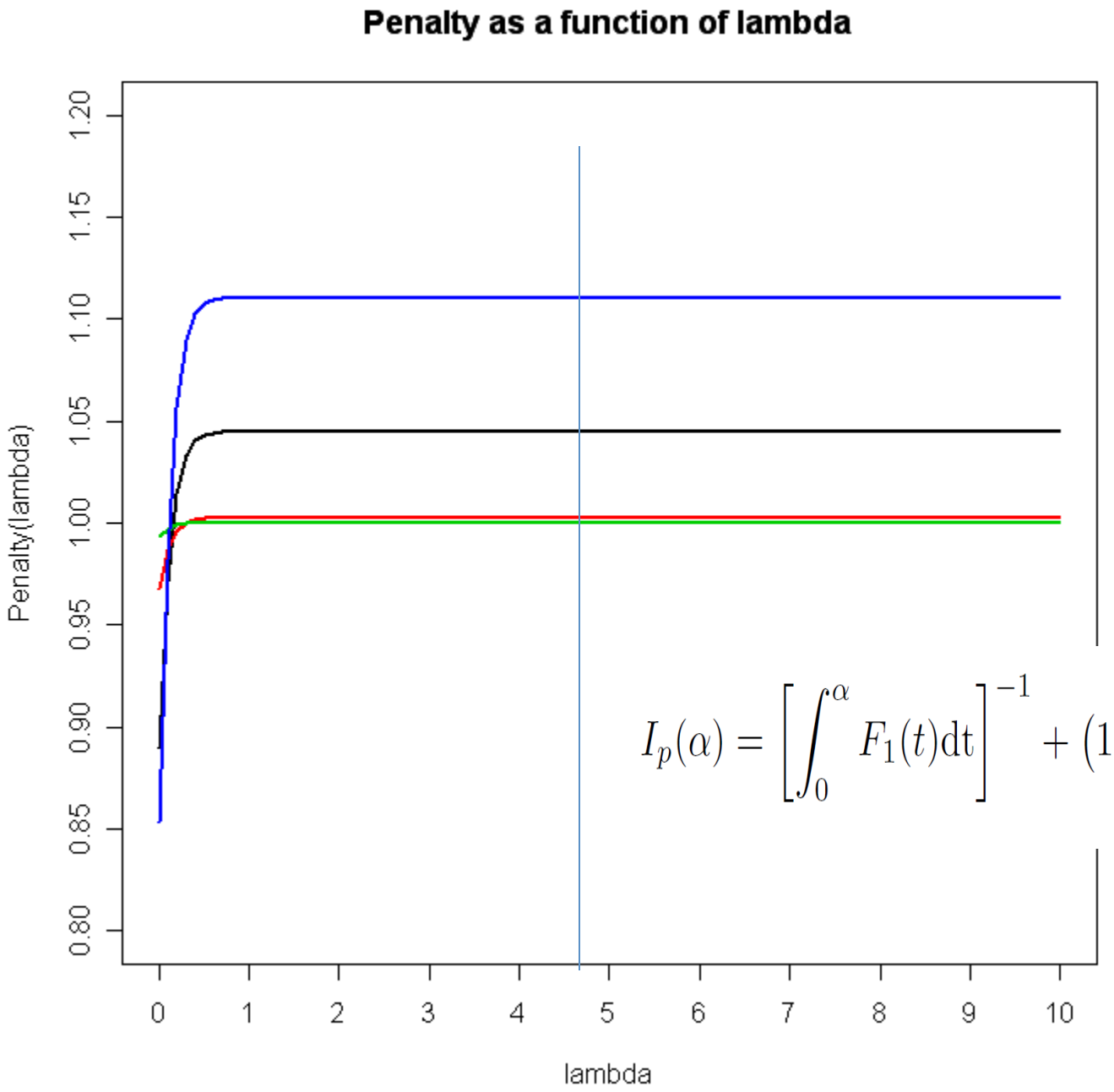
$$\alpha^* = 4^{1/3} m^{-(\lambda+1)/3}$$

λ≈4.7 for α_0=0.05

Set $\lambda = \frac{5}{3} - \log\alpha_0 + \frac{\log 4^{1/3}}{\log m}$ gives

$\alpha^* = \alpha_0 m^{-1}$ under complete uniformity

Bonf. adjustment



m=25,000
π_0=0.99

tau0=0.3: blue
tau0=0.2: black
tau0=0.05: red
tau0=0.01: green

$$I_p(\alpha) = \left[\int_0^{\alpha} F_1(t)\mathrm{dt}\right]^{-1} + \left(1 - \pi_0\tau_0 + m^{-\lambda}\right)^{-\pi_0\tau_0} m\pi_0\alpha$$

Massive Multiple Hypothesis Tests

A Simulation Study

FDR Control (Storey et al. 2003) at 1, 5, 10, 15, 20, 25, 30, 40, 60, 75 percent level

I_p in Cheng et al. (2004)

New I_p

Bias of FDR estimator in Cheng et al. (2004)

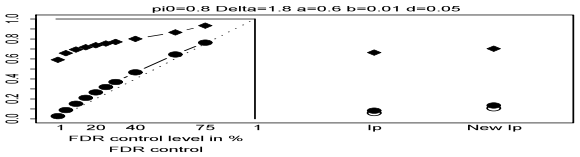
(Estimated) Actual FDR = Average proportion of false positives across simulations
(Estimated) Average Power = Average proportion of captured true H_1's

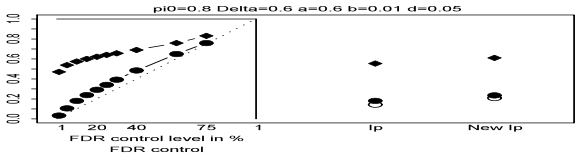
Two Group Differential Expression, *m* Genes

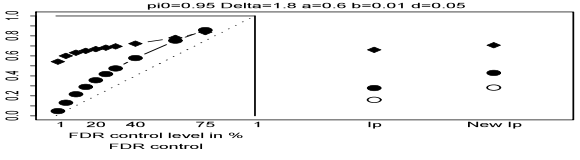
$Z_1,\ Z_2,\ Z_3$ i.i.d. $N(0,\sigma^2)$

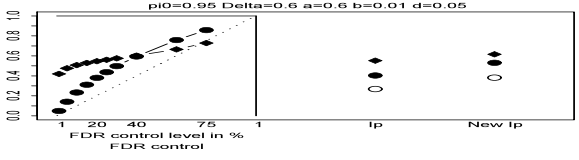
i=1,...,m; j=1,...,n

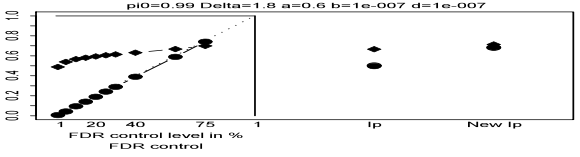
	1	n_1 n_1+1	*n*	
1	$\Delta(i/m_1)+Z_1+\varepsilon_{ij}$ ε_{ij} i.i.d. $N(0,a^{-1}\sigma^2)$	$Z_2+\varepsilon_{ij}$ ε_{ij} i.i.d. $N(0,a^{-1}\sigma^2)$		$\frac{a}{a+1}$
m_1	$Z_1+\varepsilon_{ij}$ ε_{ij} i.i.d. $N(0,b^{-1}\sigma^2)$	$Z_2+\varepsilon_{ij}$ ε_{ij} i.i.d. $N(0,b^{-1}\sigma^2)$		$\left[\left(1+\frac{1}{a}\right)\left(1+\frac{1}{b}\right)\right]^{-\frac{1}{2}}$ $\frac{b}{b+1}$
m_2	$Z_3+\varepsilon_{ij}$ ε_{ij} i.i.d. $N(0,d^{-1}\sigma^2)$	i.i.d. $N(0,\sigma^2(d+1)/d)$		$\frac{d}{d+1}$
m_3	i.i.d. $N(0,\sigma^2)$	i.i.d. $N(0,\sigma^2)$		
m				

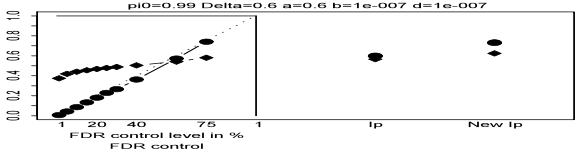


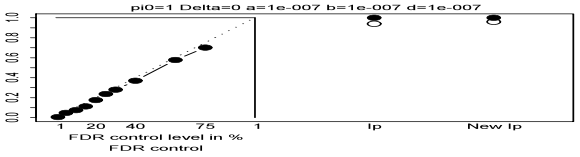


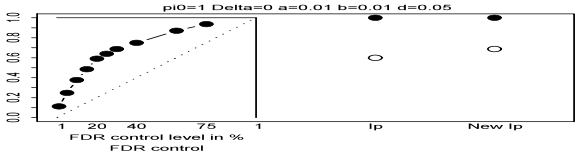












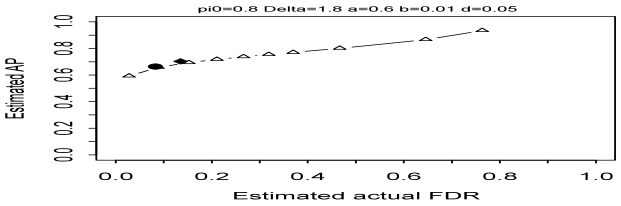
Dot: actual FDR
Diamond: avg. power
Circle: mean FDR estimate

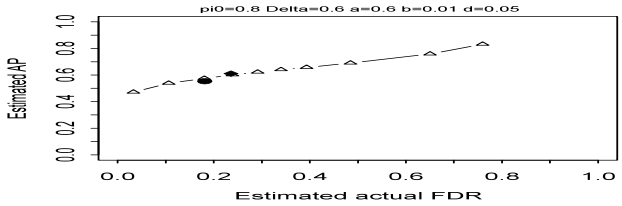
Two Group Differential Expression, *m* Genes

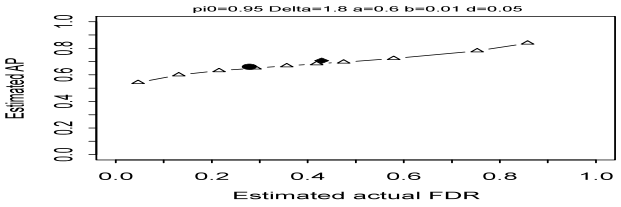
$Z_1,\ Z_2,\ Z_3$ i.i.d. $N(0,\sigma^2)$

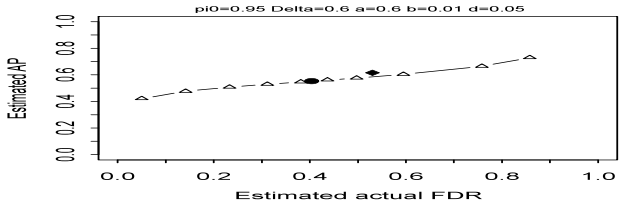
i=1,...,m; j=1,...,n

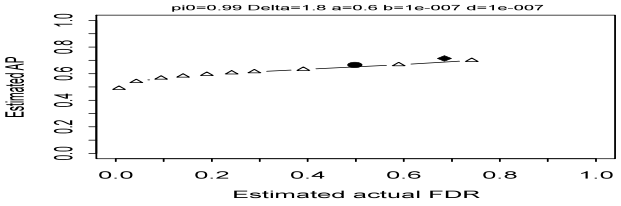
	1	n_1 n_1+1	*n*	
1	$\Delta(i/m_1)+Z_1+\varepsilon_{ij}$ ε_{ij} i.i.d. $N(0,a^{-1}\sigma^2)$	$Z_2+\varepsilon_{ij}$ ε_{ij} i.i.d. $N(0,a^{-1}\sigma^2)$		$\frac{a}{a+1}$
m_1	$Z_1+\varepsilon_{ij}$ ε_{ij} i.i.d. $N(0,b^{-1}\sigma^2)$	$Z_2+\varepsilon_{ij}$ ε_{ij} i.i.d. $N(0,b^{-1}\sigma^2)$		$\left[\left(1+\frac{1}{a}\right)\left(1+\frac{1}{b}\right)\right]^{-\frac{1}{2}}$ $\frac{b}{b+1}$
m_2	$Z_3+\varepsilon_{ij}$ ε_{ij} i.i.d. $N(0,d^{-1}\sigma^2)$	i.i.d. $N(0,\sigma^2(d+1)/d)$		$\frac{d}{d+1}$
m_3	i.i.d. $N(0,\sigma^2)$	i.i.d. $N(0,\sigma^2)$		
m				

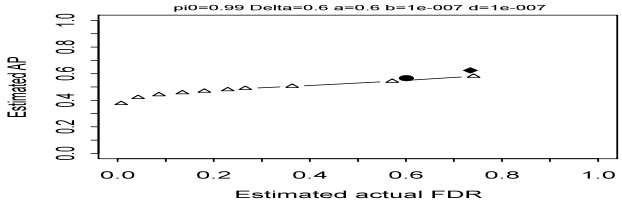








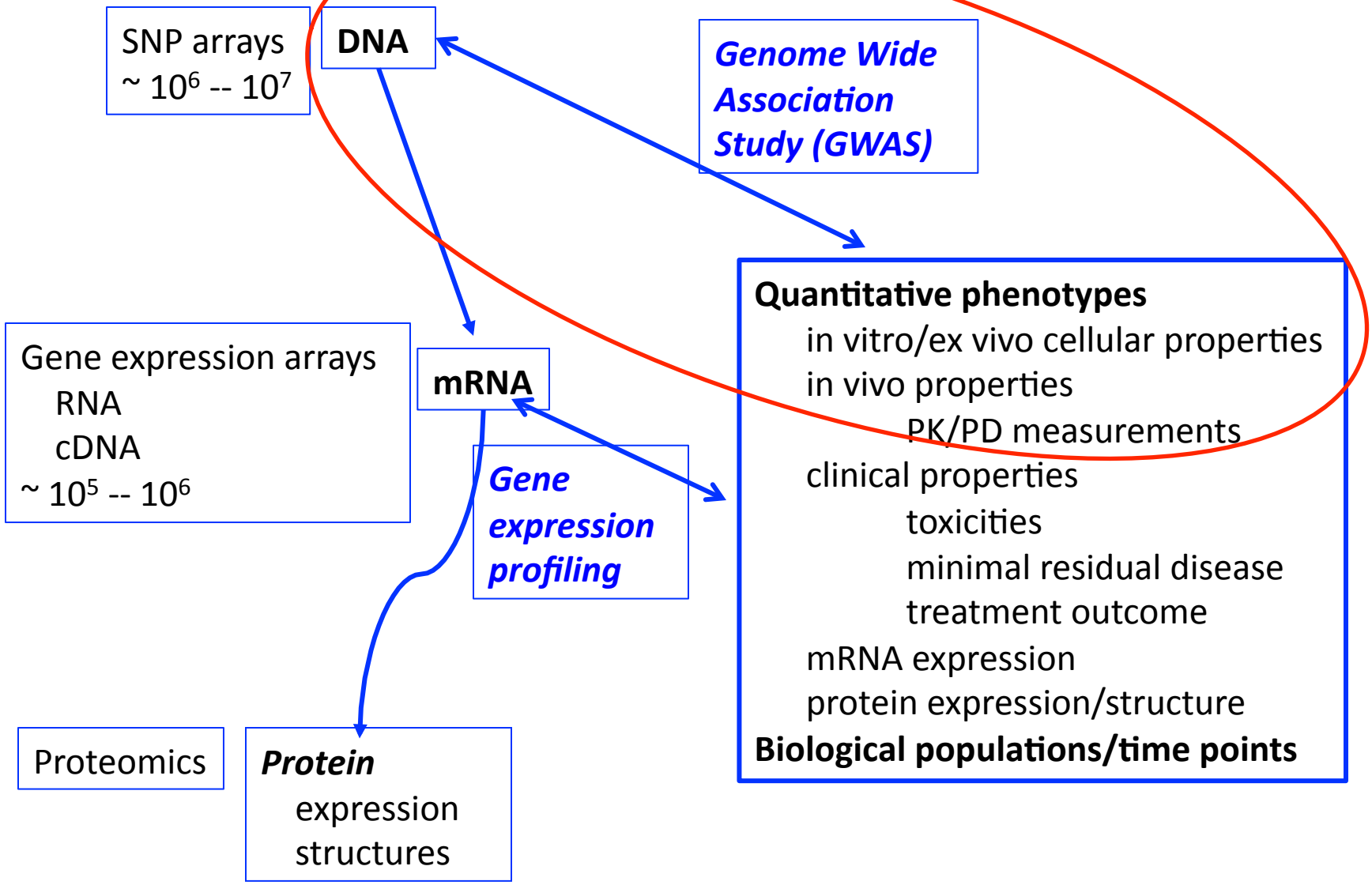


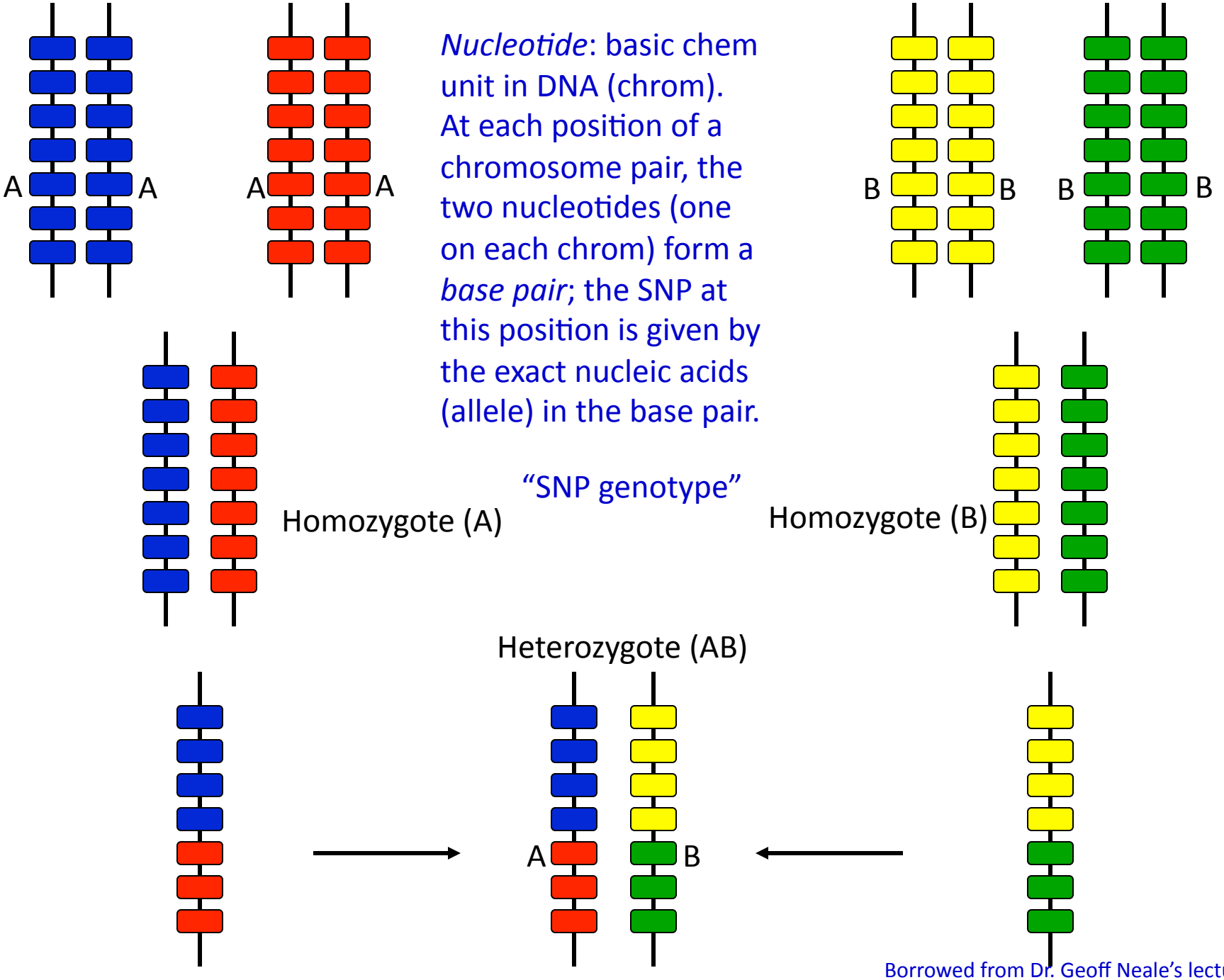


Dot: Ip
Diamond: New Ip
Triangle: FDR control

Questions?

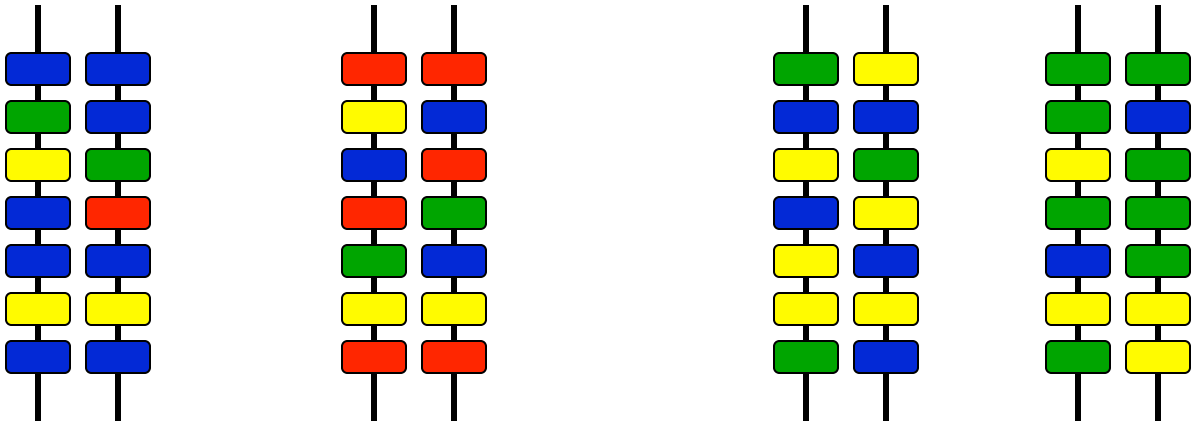
Genome-wide Association Study





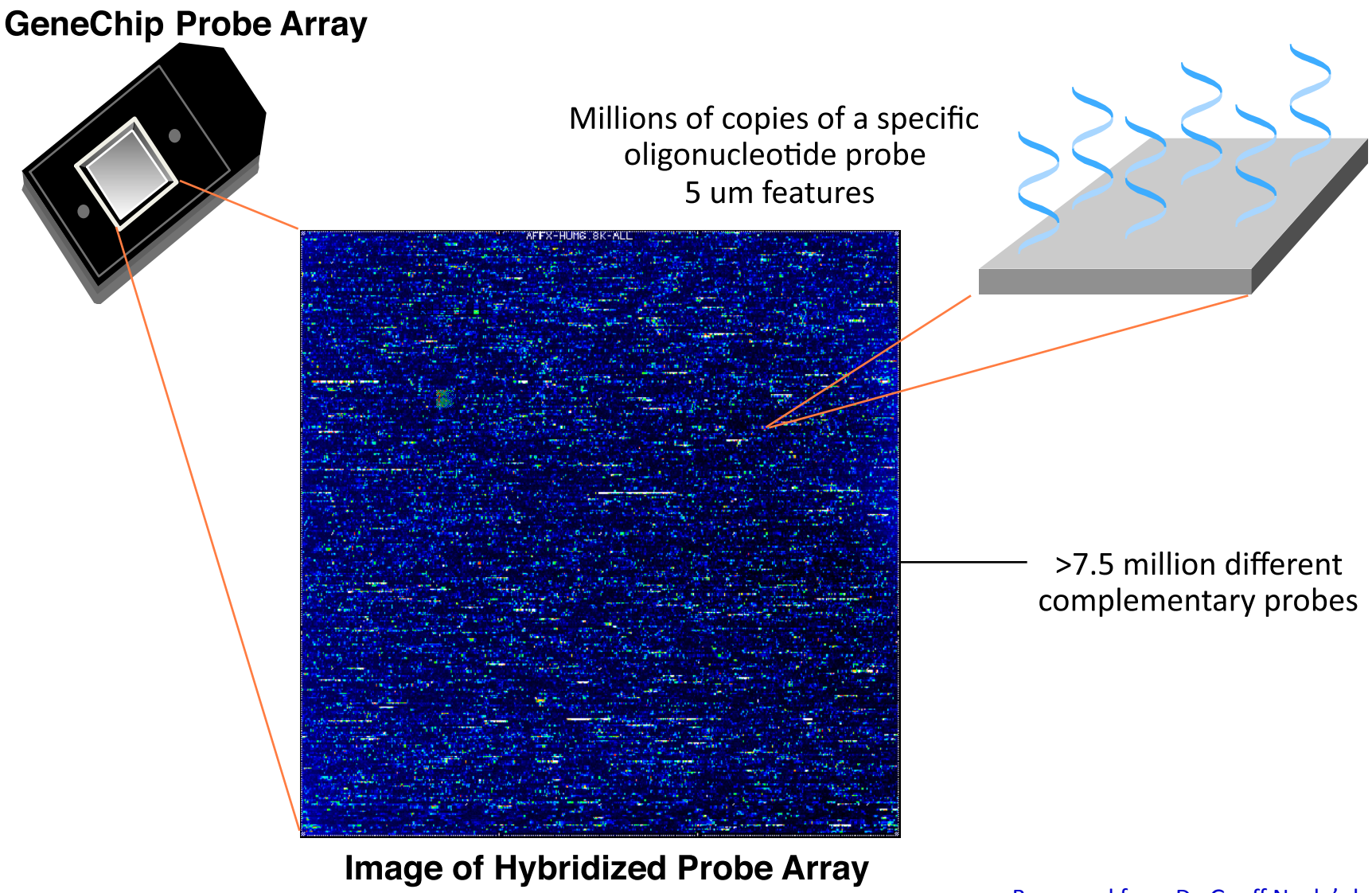
Borrowed from Dr. Geoff Neale's lecture

Traces of genetic ancestry



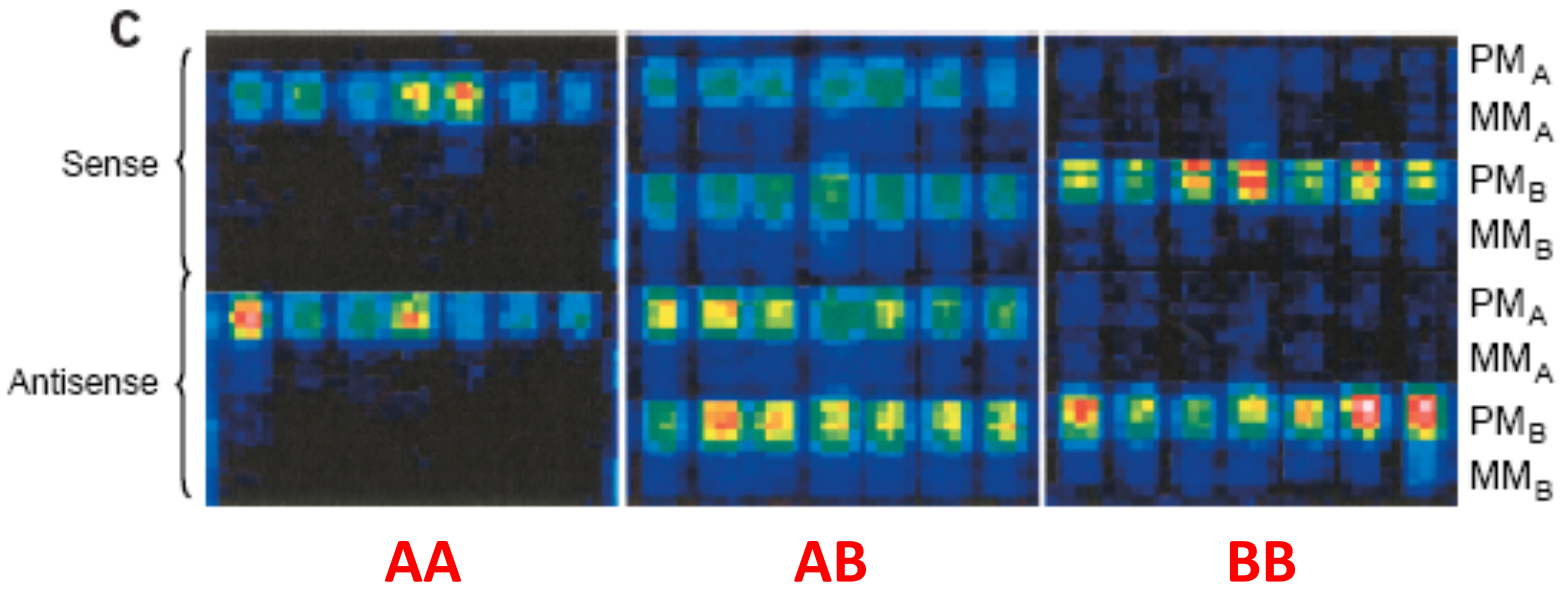
Borrowed from Dr. Geoff Neale's lecture

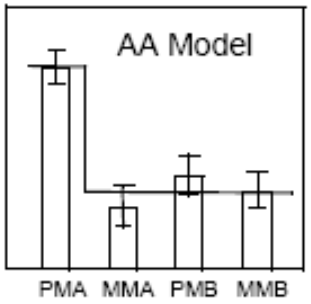
SNP Array
Affymetrix GeneChip Technology

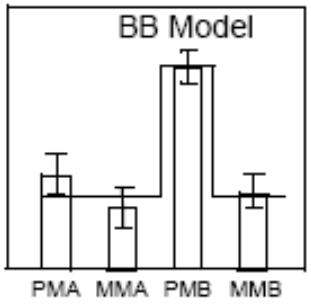


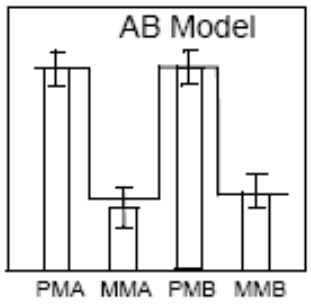
Borrowed from Dr. Geoff Neale's lecture

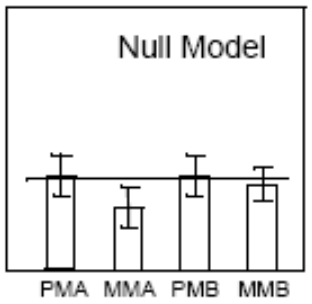
Genotype Calling











Borrowed from Dr. Geoff Neale's lecture

SNP Assay Performance

SNP Assay	No. SNPs	Inter-SNP Median	Inter-SNP Mean	Hetero-zygosity Mean	Genome within 10kb of SNP	Call Rate Concordance
10K	10,204	113 kb	258 kb	0.38		>99.6%
100K	116,204	8 kb	22.5 kb	0.30	40%	>99.7%
500K	500,000	2.5 kb	5.8 kb	0.29	85%	>99.6%

Borrowed from Dr. Geoff Neale's lecture

Genome-Wide Association Study (GWAS)

- Affy SNPChip6.0: Assay 1.2 million SNPs, 900,000 for genotypes and 300,000 for copy number variation
- Whole-genome sequencing: many, many more
- Goal: Discover genotype-phenotype association
- Procedure: Test each SNP's association with the trait of interest and determine by some criteria which SNPs are statistically (and hopefully, biologically as well) significant in explaining the phenotypic variation in the population under study.
- ~10^6 to 10^8 SNPs – about the same number of statistical tests in a single study

GWAS Designs and Statistical Models

- Case-Control Study
  - Binary phenotype (present/1, absent/0)
  - Cases: phenotype=1; Controls: phenotype=0
  - Obtain (large) numbers of cases and controls from the study population
  - Often retrospective: identify cases and controls from what's available in the tissue bank(s)
  - Run SNP array on each case and each control

GWAS Designs and Statistical Models

- Case-Control Study
  - Run SNP array on each case and each control
  - Each case/control has say 600,000 SNPs typed
  - Test SNP association for each typed SNP (600K tests)

	AA	AB	BB
Case	1010	3218	5126
Control	100287	40125	1021

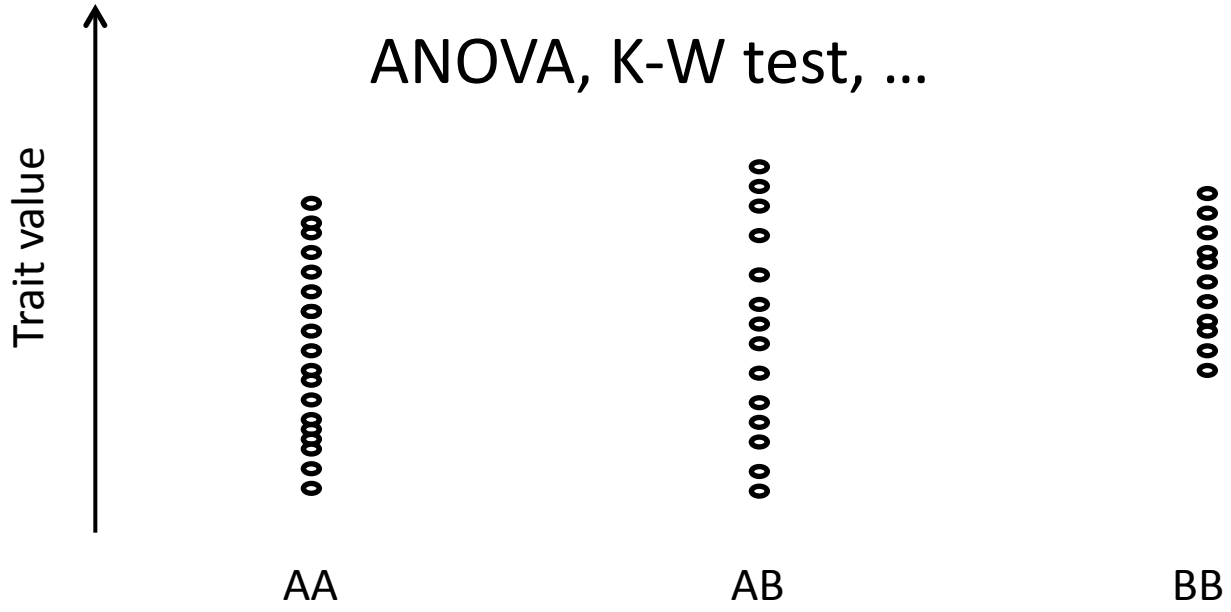
Chi-square test, Cochran-Armitage trend test

GWAS Designs and Statistical Models

- Cohort Study (in Oncology)
  - "Arbitrary" phenotype: binary, ordinal, continuous
  - Obtain (a large) number of subjects from the study population (e.g., "uniformly" treated patients): the study cohort
  - Can be retrospective or prospective
  - Run SNP array on each subject in the cohort

GWAS Designs and Statistical Models

- Cohort Study (in Oncology)
  - Run SNP array on each subject in the cohort
  - Each subject has say 600,000 SNPs typed
  - Test SNP association for each SNP (600K tests)



Genome-wide Association Study

- Multiple Hypothesis Tests
  - Statistical errors
  - Genome-wide significance (FWER)
  - False discovery rate
    - Estimation
    - Control
  - Adaptive significance thresholds

- *Genome-wide significance*: Race/population specific!

Genome-wide Association Study

- Multiple Hypothesis Tests
  - *Family-Wise Error Rate* (FWER): The probability of making one or more type-I errors in *m* tests (say *m*=600K) – the "alpha level" of multiple tests
  - Based on the Binomial model:
    - If the probability of type-I error on a single SNP is $5x10^{-2}$ (5%), the FWER for 600K SNPs is essentially 1.00 (100%)
    - If the probability of type-I error on a single SNP is $8x10^{-8}$, the FWER for 600K SNPs is 0.047.

    Has to be extremely conservative on single test in order to control the FWER

Genome-wide Association Study

- Multiple Hypothesis Tests
  - *Genome-Wide Significance* (an α level for GWAS)

Genetic Epidemiology 32: 227–234 (2008)

Estimation of Significance Thresholds for Genomewide Association Scans

Frank Dudbridge* **and Arief Gusnanto**

MRC Biostatistics Unit, Institute for Public Health, Cambridge, United Kingdom

Abstract

The question of what significance threshold is appropriate for genomewide association studies is somewhat unresolved. Previous theoretical suggestions have yet to be validated in practice, whereas permutation testing does not resolve a discrepancy between the genomewide multiplicity of the experiment and the subset of markers actually tested. We used genotypes from the Wellcome Trust Case-Control Consortium to estimate a genomewide significance threshold for the UK Caucasian population. We subsampled the genotypes at increasing densities, using permutation to estimate the nominal *P*-value for 5% family-wise error. By extrapolating to infinite density, we estimated the genomewide significance threshold to be about 7.2×10^{-8}. To reduce the computation time, we considered Patterson's eigenvalue estimator of the effective number of tests, but found it to be an order of magnitude too low for multiplicity correction. However, by fitting a Beta distribution to the minimum *P*-value from permutation replicates, we showed that the effective number is a useful heuristic and suggest that its estimation in this context is an open problem. We conclude that permutation is still needed to obtain genomewide significance thresholds, but with subsampling, extrapolation and estimation of an effective number of tests, the threshold can be standardized for all studies of the same population. *Genet. Epidemiol.* 32: 227–234, 2008.

Genome-wide significance control FWER and is *population-specific*!

For 600K tests, the Bonferroni adjustment to achieve 5% FWER is $0.05/600{,}000=8.3\times10^{-8}$. The estimated genome-wide significance for 5% FWER is essentially the same as Bonferroni adjustment.

In very large GWAS (say 5 million tests), 7.2×10^{-8} is appropriate for controlling the FWER at 5% for studies "**of the same population**."

Discussion

We have shown that previous proposals for genomewide significance have been in the right order of magnitude. It seems clear that, in a Western population, any P-value less than say 5×10^{-8} can be regarded as convincingly significant. We rely on permutation testing to estimate significance thresholds, but these should be adjusted to reflect the genomewide multiplicity. Estimation of an effective number of tests remains an open problem but one which has potential to considerably reduce the computational burden. The next generation of genotyping chips should allow more accurate estimation of significance thresholds with application to a wider range of genomic variation.

Often YOUR study population (patients, mice, cell lines …) is NOT the same as "the Western Population."

Do not blindly follow the "genome-wide significance" HYPE!

Genome-wide Association Study

- Multiple Hypothesis Tests
  - Family-Wise Error Rate (FWER): The probability of **making one** or more type-I errors in *m* tests (say *m*=600K)
  - Based on the Binomial model:
    - If the probability of type-I error on a single SNP is 5×10^{-2} (5%), the FWER for 600K SNPs is essentially 1.00 (100%)
    - If the probability of type-I error on a single SNP is 8×10^{-8}, the FWER for 600K SNPs is 0.047 (Bonferonni adjustment).
  - Genome-wide significance 7.2×10^{-8}

Is FWER a reasonable measurement of error level for GWAS?
NO!

Genome-wide Association Study

- Multiple Hypothesis Tests
  - Family-Wise Error Rate (FWER): The probability of making **one** or more type-I errors in *m* tests (say *m*=600K)
  - FWER is unnecessarily conservative, unfit for exploratory studies such as GWAS.
  - *False Discovery Rate* (FDR)
    - Look directly at the number of false positive errors in all positive findings, instead of the probability of making a false positive error in all tests.
  - Adaptive significance thresholds

Questions?

Validation of Positive Findings

- Biological Validation
  - Demonstrate the biological relationship discovered from microarray experiments by additional wet-lab experiments, in e.g., cell lines, animal models, (prospective clinical trials?), etc.

- Statistical Validation
  - **Validate discovered genomic associations**
    - Internal "validation": based only on the study dataset
    - External validation: based on independent dataset(s)
  - **Validate classifiers**
    - Internal validation: cross validation using the study dataset only
    - External validation: assess the classifier using independent data

- Internal validation -- Split dataset once: bad idea!

Internal Validation of Discovered Genomic Associations: What does this mean?

Gene Expression Profiling of the Development of
Secondary Myeloid Leukemia (t-ML)

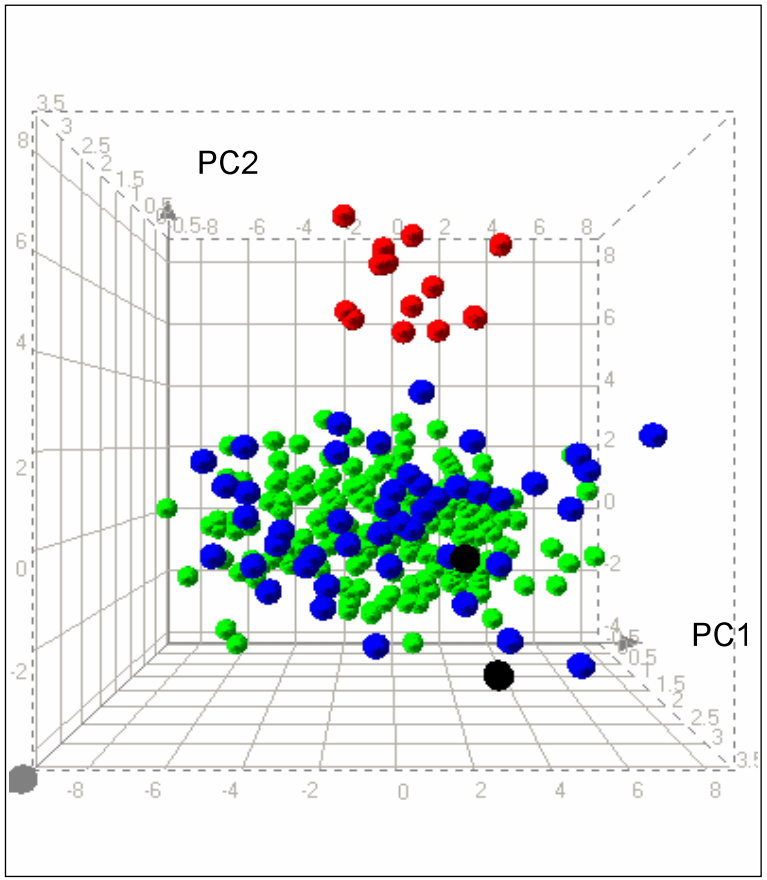
Bogni, Cheng, Liu, … Relling *Leukemia,* 20:239-246, 2006

TOTAL-XIIIA & TOTAL-XIIIB Patients
267 with gene expression data (ALL blasts at diagnosis)
Affymetrix U95Av.2 GeneChip
14 have developed t-ML at the time of analysis
Median length of follow-up 6.1 years.

First Stage Analysis:

1. Probesets (genes) were selected by hazard rate regression modeling
   241 probes were identified at α=0.01.
2. The 267 patients were clustered into 3 groups using the 241 probesets as features by Ward's hierarchical clustering and cutting the tree at the 3-cluster level --> **<u>all 14 t-MLs were clustered together</u>**

3D plot of PCA for selected 241 genes from CRR model for t-ML in T13 group (N=267)



Event-free survival ● t-ML ● Other 2nd malignancy ● Other adverse events ●

Must demonstrate that the observed cluster is NOT completely attributable to CHANCE; otherwise we would have made a …

Circular Reasoning!

Dupuy A, Simon R (2007) JNCI, 99:147-157

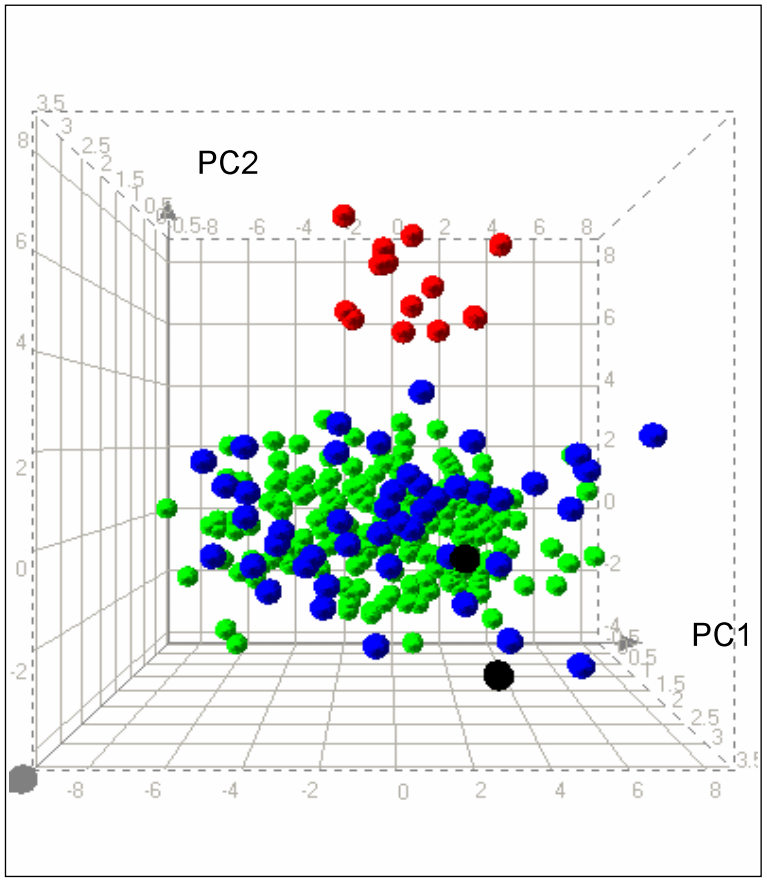
Validation of discovered genomic associations: Definition

- Internal validation
  - "validation" using only the study dataset
- ***Validation***: A convincing statistical argument (evidence) that the discovery of the (statistically) significant findings is not entirely due to chance – evidence in addition to error control for multiple hypothesis tests
  - Significance: gFWER, FDR, adaptive threshold …
  - Model for chance (null hypothesis)

Internal validation of discovered genomic associations: Definition

- ***Validation***: A convincing statistical argument (evidence) that the discovery of the (statistically) significant findings is not entirely due to chance – evidence in addition to error control for multiple hypothesis tests
- A higher level hypothesis test demonstrating a global association between the discovered set of genomic variables (genes) and the "response variable" (phenotype); e.g. association between clusters generated by the discovered gene expression profile and the phenotype.

3D plot of PCA for selected 241 genes from CRR model for t-ML in T13 group (N=267)



Event-free survival ● t-ML ● Other 2nd malignancy ● Other adverse events ●

Split data validation (one-round cross validation)
Bullinger … Pollack *NEJM* 350(16), April 2004

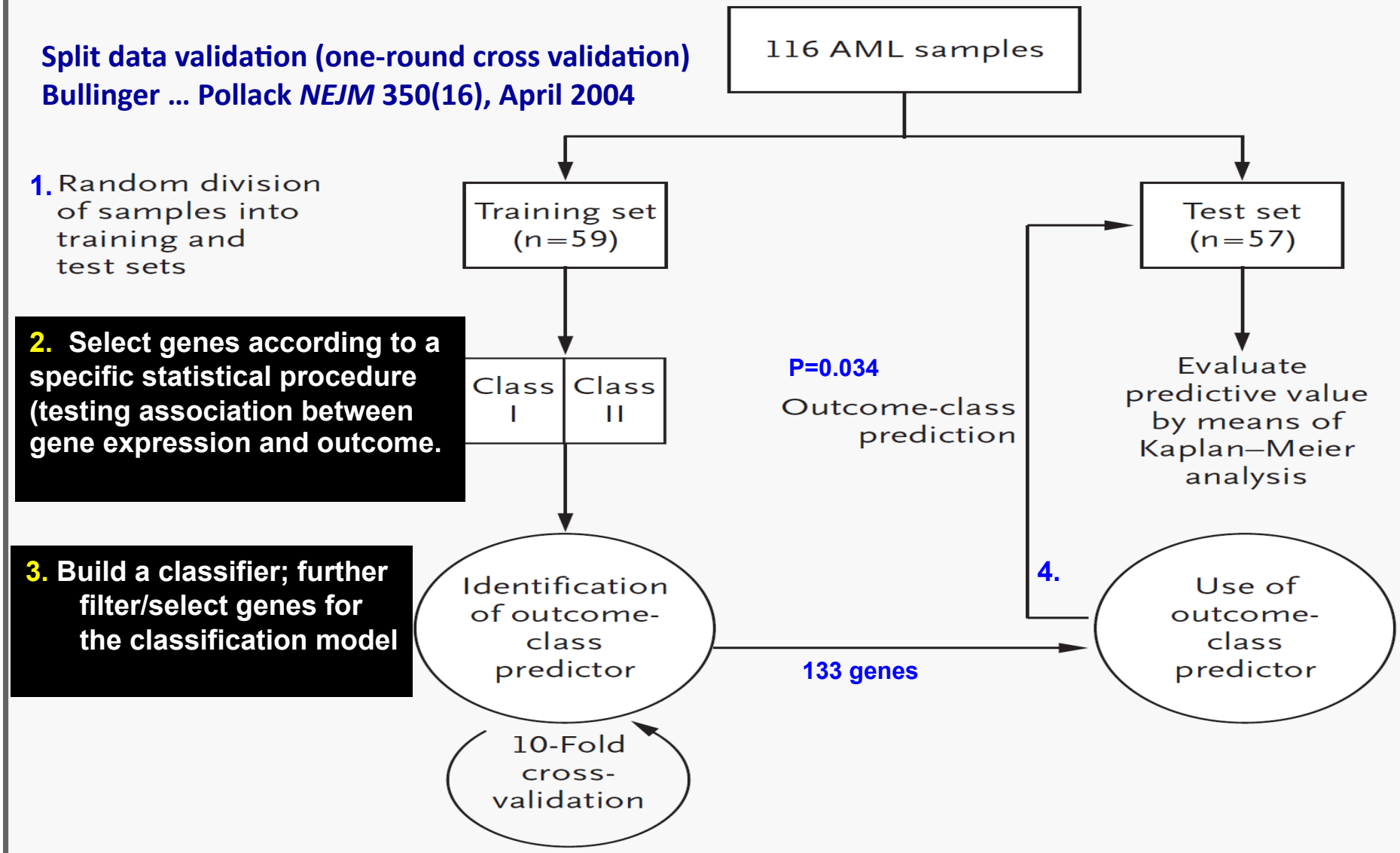


Figure 3. Overview of the Strategy Used for the Development and Validation of an Outcome Predictor Based on Gene-Expression Signatures.

SAM denotes significance analysis of microarrays, and PAM prediction analysis of microarrays.

Training vs. Test sets split from entire cohort

Reduction in Power Castaldi *et al.* (2010) Briefings in Bioinformatics, 12(3):189-202

- Splitting cohort reduces statistical power to detect important genes on the training set.
- The relatively small "test set" reduces power to test the association between the expression profile and the outcome (phenotype), and it reduces the accuracy of estimating the classification error.

Lack of Stability and interpretability Michiels *et al.* (2005) The Lancet, 365:488-492
Cheng (2009) CSDA, 53:788-800

- Will the same 133 genes "PAM" out if the 116-pts cohort is split again?
  for third time? the fourth time? …
- PAM prediction of individuals into good vs. bad outcome on test set is weakly significant (P=0.034)
  Will one get the same if the cohort is randomly split again,
  third time, …?

WHAT HAS BEEN VALIDATED?

Back to the t-ML example…

Internal Validation: A convincing statistical argument (evidence) that the discovery of the (statistically) significant findings (cluster-tML association) is not entirely due to chance.

If there is no gene (represented by the probes) associated with t-ML, how often one would observe the glorious clusters by exactly the same analyses in Steps 1 and 2?

First Stage Analysis:

1. Probesets/Genes were selected by hazard rate regression
   241 probes were identified at α=0.01 (FDR=*high!*)
2. The 267 patients were clustered into 3 groups using the 241 probesets as features by Ward's hierarchical clustering and cutting the tree at the 3-cluster level --> **all 14 t-MLs were clustered together**

One can conclude no more nor any stronger by one-round split than what can be achieved by a straightforward association analysis without splitting the whole dataset into two.

How to meaningfully validate the significant findings in an association analysis internally (using the entire study set alone)? Cheng (2009) *Comput. Statist. Data Anal.* **53**, 788-800

Computational Statistics and Data Analysis 53 (2009) 788–800

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Internal validation inferences of significant genomic features in genome-wide screening

Cheng Cheng*

Department of Biostatistics, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105-2794, United States

If there is no gene (represented by the probes) associated with t-ML, how often one would observe the glorious clusters by exactly the same analyses in Steps 1 and 2?

Permutation Internal Validation (Basic Idea):

Answer the above "how often" question by estimating the probability via permutation/re-sampling methods

Test the strength of the signal, i.e. global stochastic dependence between the set of significant findings and the phenotype against the completely null hypothesis (white noise background):

a novel family of test statistics coupled with permutation tests.

Back to the t-ML Example – Second Stage Analysis:

Permutation internal validation (test) indicated that if the completely null (white noise) hypothesis is true, the chance (probability) of obtaining a gene expression profile as strong as shown by the cluster analysis was 4/1000=0.4%!

Permutation Internal Validation

Randomly assign the 267 (time, censor) pairs to the 267 patients
Repeat EXACTLY Steps 1 and 2 on permuted data
1. Probes selection by CIN regression (Fine & Gray 1999)
   death, relapse, and SMN as competing risks
2. Cluster the 267 patients into 3 groups using the
   selected probes as features and Ward's algorithm;
   cut the tree at the 3-cluster level

Compute Gray's test statistic over the clusters
Repeat the above steps 1000 times

A permutation round is considered significant if and only if

- the number of selected probes at 0.01 level is at least 1, and
- the cluster with the highest percentage of t-ML events must contain at least ***k*** t-ML events, and
- Gray's statistic is greater than or equal to 267.8, the value observed on the original dataset.

***Profile Significance* = #(*significant permutation rounds*)÷1000**

k	Prof. Sig.
7	0.004
8	0.004
9	0.004
10	0.004
11	0.004
12	0.004

k = cluster coverage of t-ML

Questions?

External Validation of Discovered Genomic Associations

- External validation: study design?
- Replicate (some of) the discovered gene-phenotype relationships on an independent dataset according to specified criteria
- Similar to a single-stage phase-II clinical trial
  - Population (and patient cohort/treatment)
  - Effect size (estimated using the discovery data)
  - Power/probability of replication
  - Sample size required
  - Data collection: Genome-wide scan or focused (on "discovered genes" only)? Scale of multiplicity?

External Validation of Discovered Genomic Associations

- "Replicate" (some of) the discovered gene-phenotype relationships on an independent dataset according to some specified criteria
- Illustrative example: criteria/inference for "replication" -- Validation of gene co-expression clusters using related phenotype/outcome

An Example:

Leukemia blast gene expression and minimal residual disease (MRD)

NEOPLASIA

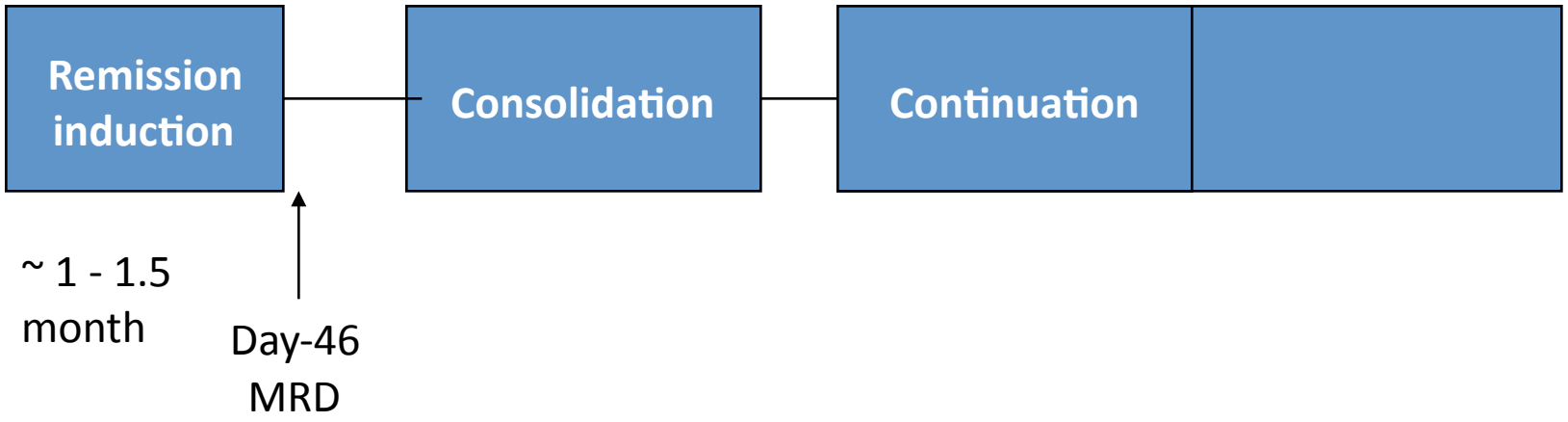
Genes contributing to minimal residual disease in childhood acute lymphoblastic leukemia: prognostic significance of *CASP8AP2*

Christian Flotho, Elaine Coustan-Smith, Deqing Pei, Shotaro Iwamoto, Guangchun Song, Cheng Cheng, Ching-Hon Pui, James R. Downing, and Dario Campana

In childhood acute lymphoblastic leukemia (ALL), early response to treatment is a powerful prognostic indicator. To identify genes associated with this response, we analyzed gene expression of diagnostic lymphoblasts from 189 children with ALL and compared the findings with minimal residual disease (MRD) levels on days 19 and 46 of remission induction treatment. After excluding genes associated with genetic subgroups, we identified 17 genes that were significantly associated with MRD. The caspase 8–associated protein 2 *(CASP8AP2)* gene was studied further because of its reported role in apoptosis and glucocorticoid signaling. In a separate cohort of 99 patients not included in the comparison of gene expression profiles and MRD, low levels of *CASP8AP2* expression predicted a lower event-free survival ($P = .02$) and a higher rate of leukemia relapse ($P = .01$) and were an independent predictor of outcome. High levels of *CASP8AP2* expression were associated with a greater propensity of leukemic lymphoblasts to undergo apoptosis. We conclude that measurement of *CASP8AP2* expression at diagnosis offers a means to identify patients whose leukemic cells are highly susceptible to chemotherapy. Therefore, this gene is a strong candidate for inclusion in gene expression arrays specifically designed for leukemia diagnosis.
(Blood. 2006;108:1050-1057)

An Example (Cont.)

Contemporary chemotherapy of childhood acute lymphoblastic leukemia (ALL)



Minimal Residual Disease/leukemia (MRD) in bone marrow is detected by flow cytometry or PCR. Sensitivity of flow is 1/10,000 (0.01%). The flow is extremely specific.

Clinical MRD Positive (Negative): ≥ 0.01% (< 0.01%)
This is the phenotype of interest: represents the sensitivity/resistance of the leukemia to the given chemotherapy *in vivo*.

Dichotomized MRD level (MRD+ vs. MRD-)

An Example (Cont.)

- Want to find genes associated with *in vivo* sensitivity/resistance to the given chemotherapy as represented by the Day-46 MRD status: positive vs. negative (< 0.01%)

- Genome-wide gene differential expression analysis

- Gene (mRNA) expression in leukemia blasts at diagnosis
  Affymetrix U133A GeneChip®: 22,200+ features (probe sets)

Initial Analysis:

- Probeset-by-probeset test for differential expression between the two MRD status by the rank-sum (Wilcoxon) test

- Declare statistically significant differential expression by the Ip criterion for massive multiple tests (Cheng et al. 2004)

An Example (Cont.)

- At α=0.003997 significance level, 223 probe sets were declared as significantly differentially expressed between Day 46 MRD+ and MRD- samples

Questions:

- Are there "co-differentially-expressed" genes that did not make to the 0.003997 cut?
  - **Co-expressed genes may be biologically related**
    Supervised Sequential Clustering Algorithm

- How to do a validation?
  - An independent set of 99 samples from earlier trials, incomplete MRD data but high-quality follow-up (outcome) data
  - Well known that MRD is very prognostic
    (Coustan-Smith et al. 2000, Blood; 2002 Blood; 2004 Leukemia)
  - Use these 99 samples and their failure (hematological relapse) time to perform "validation" (assurance assessment) – to be detailed later

The Clustering Algorithm

Step 1: Gene-by-gene screening; obtain a P value for each gene

Step 2: Sort the genes by their P values in increasing order;

Step 3: Select a number (N) of top genes by a criterion

Step 4: $X_1, \ldots, X_m$: the expression vectors

$X_1^*, \ldots, X_N^*$: the expression vectors of the N selected top genes ordered by P values

ρ_0: a specified correlation (similarity) threshold

nc current number of clusters

SET $nc = 0$

REPEAT for each X_j^* $(j = 1, \ldots N)$

  IF X_j^* is already in a cluster THEN skip ELSE

  REPEAT for each X_i $(i = 1, \ldots, m)$

    IF X_i is already in a cluster THEN skip ELSE

    COMPUTE $\rho(X_j^*, X_i)$

    IF $\rho(X_j^*, X_i) \geq \rho_0$ THEN put X_i into cluster nc (the current cluster)

  END

  $nc = nc + 1$

END

Back to the Example

- Used the 223 probesets as leads to generate clusters of probesets/genes whose expressions are highly correlated
  - Supervised sequential clustering algorithm
  - Similarity measure ρ: rank correlation ($\rho_0 = 0.8$)

- 188 clusters

Size	1	2	3	4	5	6	10	11	13
Num	162	11	4	4	3	1	1	1	1

All together containing 267 probesets

Back to the Example (Cont.)

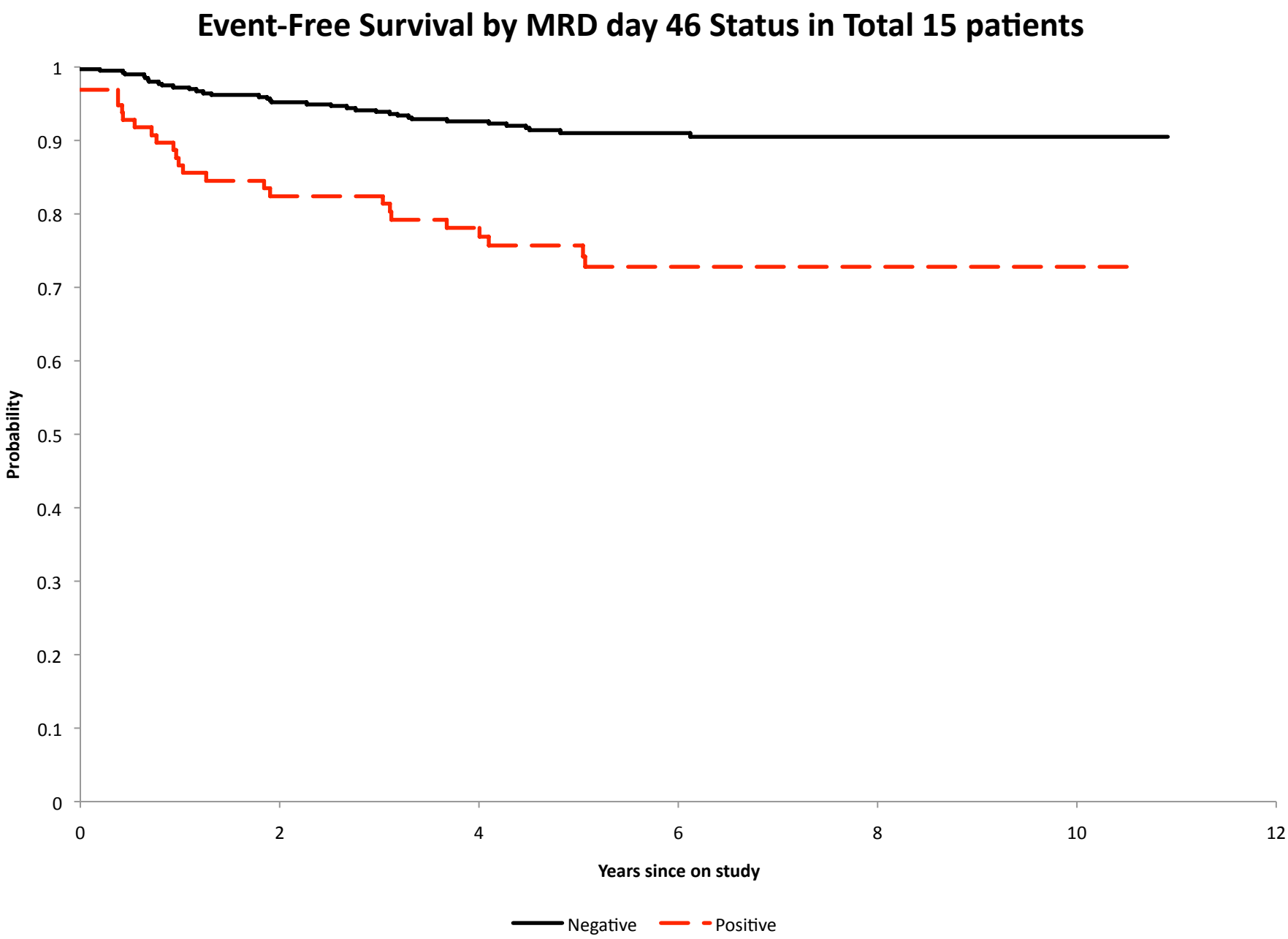
- 223 probesets were declared as "significant" at α=0.003997

Questions:

- Are there "co-differentially-expressed" probesets/genes that did not make to the 0.003997 cut?
  - Co-expressed genes may be biologically related

- How to do a "validation"?
  - An independent set of 99 samples from earlier trials, incomplete MRD data but high-quality follow-up (outcome) data
  - Well known that MRD is very prognostic
    (Coustan-Smith et al. 2000, Blood; 2002 Blood; 2004 Leukemia)
  - Use these 99 samples and their failure (hematological relapse) time to perform validation

    Validation inference for the identified co-expression clusters


Event-Free Survival by MRD day 46 Status in Total 15 patients
Probability
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
Years since on study
Negative
Positive

Validation Inference Using the Validation Dataset

Step 1: Gene-by-gene screening on the validation set; obtain a P value for each gene

Step 2: Sort the genes by their P values in increasing order;

Step 3:

$\mathcal{C}_k = \{X_{j_1}, \ldots, X_{j_{N_k}}\}$: the kth cluster; N_k genes

$\mathcal{D}_k^T = \{D_{j_1}^T, \ldots, D_{j_{N_k}}^T\}$: direction of associaiton (diff. expr.) on "training" set

$D = -1,\ 1$: e.g., sign of correlation (regression) coefficient or difference of mean (median)

$\mathcal{R}_k = \{R_{j_1}^V, \ldots, R_{j_{N_k}}^V\}$: Ranks on the validation set

$\mathcal{D}_k^V = \{D_{j_1}^V, \ldots, D_{j_{N_k}}^V\}$: direction of associaiton (diff. expr.) on "validation" set

The *cluster rank score* $S_k = -\sum_{j=1}^{N_k} I(D_j^T D_j^V > 0) \log\left(\frac{R_j^V}{m+1}\right)$

Step 4: Test the null hypothesis that $I(D_j^T D_j^V > 0)$ are $i.i.d.$ $Bernoulli(0.5)$ and that the unitized ranks $(R_j^V/(m+1)$'s) with $I(D_j^T D_j^V > 0) = 1$ come from a homogeneous point process on $[0, 1]$; or equivalently, given $M = \sum_{j=1}^{N_k} I(D_j^T D_j^V > 0)$, such unitized ranks are $i.i.d.$ $U(0, 1)$. Under this null hypothesis, the distribution of S_k has the cdf

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.5^{N_k} & x = 0 \\ 0.5^{N_k} + \sum_{j=1}^{N_k} b(j; N_k, 0.5) G(x; j, 1) & x > 0 \end{cases}$$

where $b(\cdot; n, \pi)$ is the $Binomial(n, \pi)$ pmf and $G(\cdot; \alpha, \beta)$ is the $Gamma(\alpha, \beta)$ cdf.

One-sided P value $P = 1 - F(S_k)$.

Back to the Example (Cont.)

- An independent set of 99 samples from earlier trials, incomplete MRD data but high-quality outcome data
- Well known that MRD is very prognostic

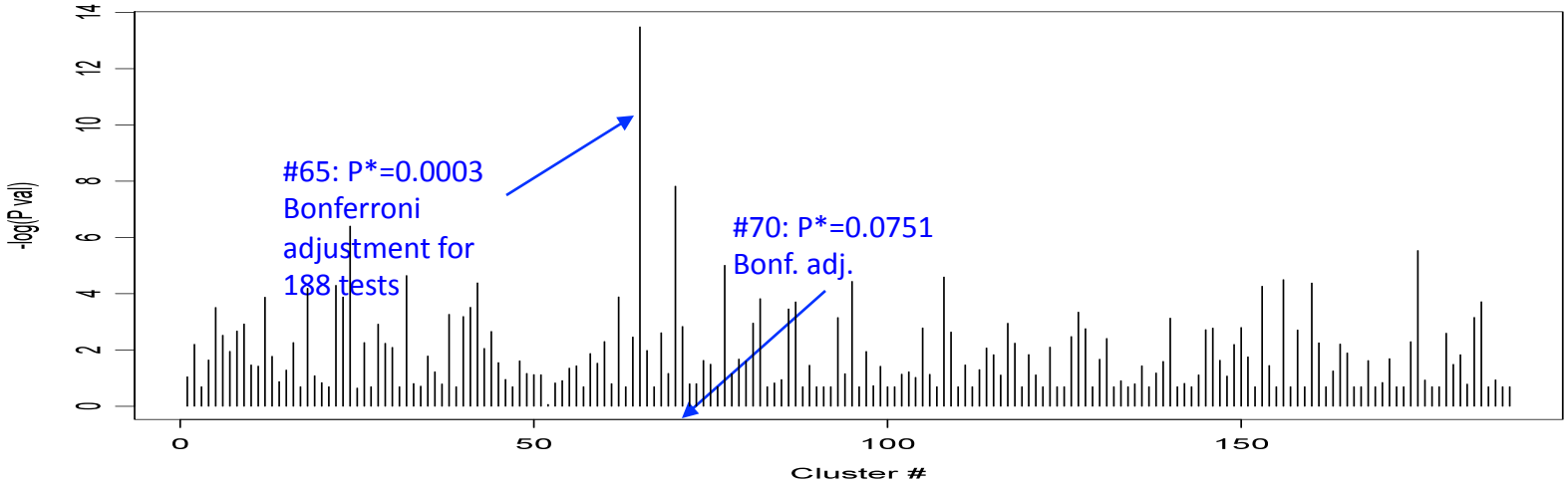
  (Coustan-Smith et al. 2000, Blood; 2002 Blood; 2004 Leukemia)
- Use these 99 samples and their failure (hematological relapse) time to perform validation

1. Perform probeset-by-probeset test for association with failure time using a Cox-type regression model; obtain model coefficient P values

2. Sort the probesets by the above P values

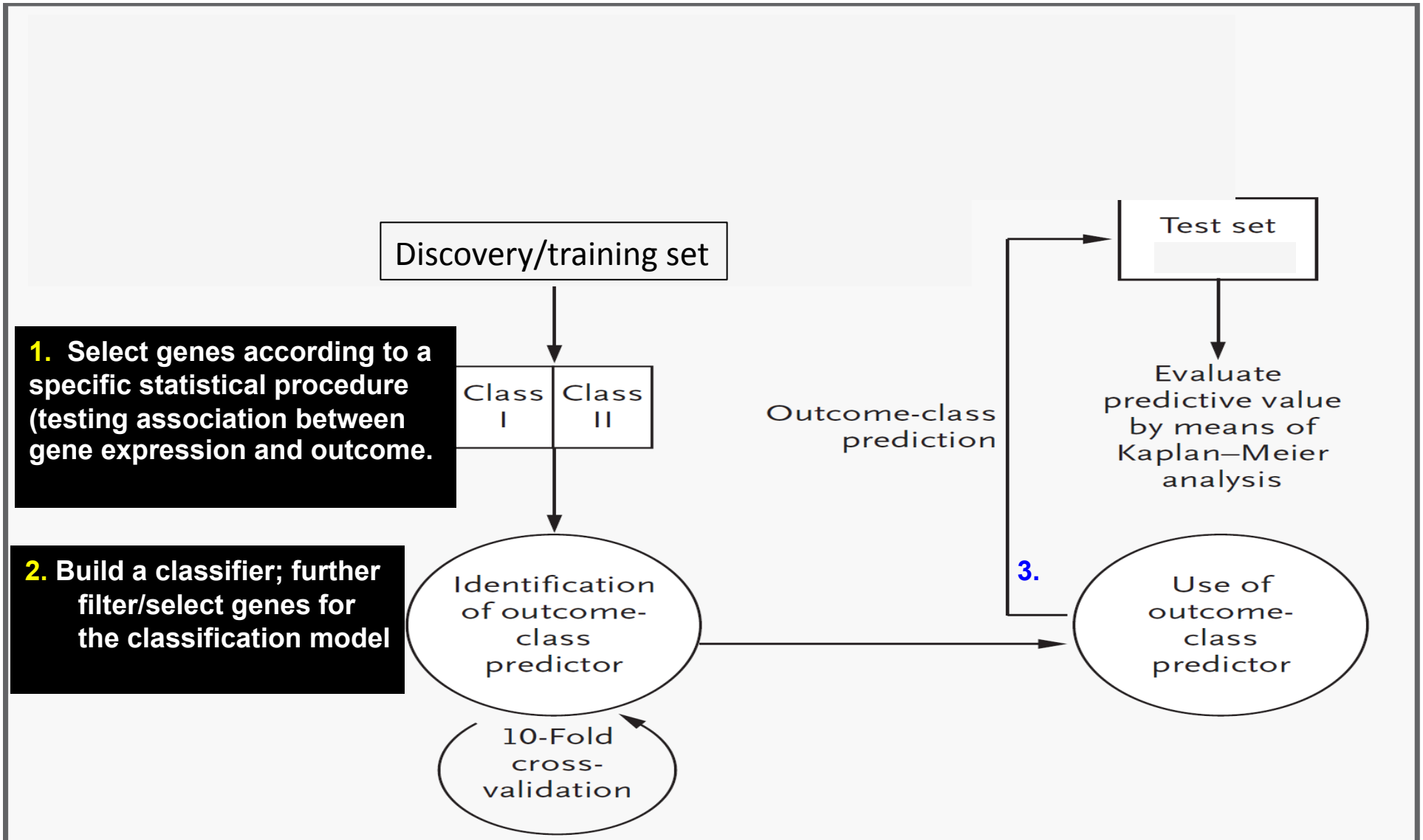
3. Carry out the validation inference on the gene co-expression clusters as described

Back to the Example (Cont.)



Questions?

External Validation of Genomic Prognostic Predictors



External Validation of Genomic Prognostic Predictors

- External validation: study design?
- Assess the predictor's accuracy in prospective clinical trials
  - A predictor working by itself
  - A predictor that potentially improves existing clinical classification
- Similar to a phase-III clinical trial ?
  - Population (treatment and patient cohort)
  - Expected accuracy (clinically useful; estimation from discovery data)
  - Power/probability of achieving the expected accuracy
  - Sample size required
  - Stratified randomization: make sure the distribution of treatments is the same across the predicted good vs. bad prognosis.
  - Evaluation of clinical value: correlation/association of prediction with prospectively observed treatment response/outcome; prediction accuracy

Genomic Markers in Cancer Trials

- **The Duke University fallout:** To make a ~5-year long story short
  - Scientifically flawed and problematic genomic classifiers validated with corrupted validation data lead to three unethical cancer trials that may have put patients at undue risk of over- or under-treatment
- Duke University investigators:

  Anil Potti and Joseph Nevins
- Biostatisticians who discovered flaws and problems in Duke studies:

  Keith Baggerly, Kevin Coombes

  M. D. Anderson Cancer Center, Houston, TX
- US Federal agencies eventually involved:

  National Cancer Institute (NCI)

  Food and Drug Administration (FDA)
- Institute of Medicine (IOM) Committee on constructing guidelines for clinical trials using risk/prognostic/response classifiers based on genomic "signatures" (e.g. gene/mRNA expression profiles)

Genomic Markers in Cancer Trials

- Forensic Bioinformatics:
- Baggerly and Coombes (2010) Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics* **3**, 1309-1334

Genomic Markers in Cancer Trials

- Some materials from Cancer Letter and IOM website:

Cancer Letter:
http://www.cancerletter.com/search?searchtext=Potti

IOM: Baggerly's presentation
http://www.iom.edu/~/media/Files/Activity%20Files/Research/OmicsBasedTests/baggerly_iom11.pdf

IOM: Lisa McShane's (NCI) presentation
http://www.iom.edu/~/media/Files/Activity%20Files/Research/OmicsBasedTests/PAF%20Document%2021.pdf

IOM: Nevins' presentation
http://www.iom.edu/~/media/Files/Activity%20Files/Research/OmicsBasedTests/Meeting-2-March-2011/Nevins.pdf

Genomic Markers in Cancer Trials

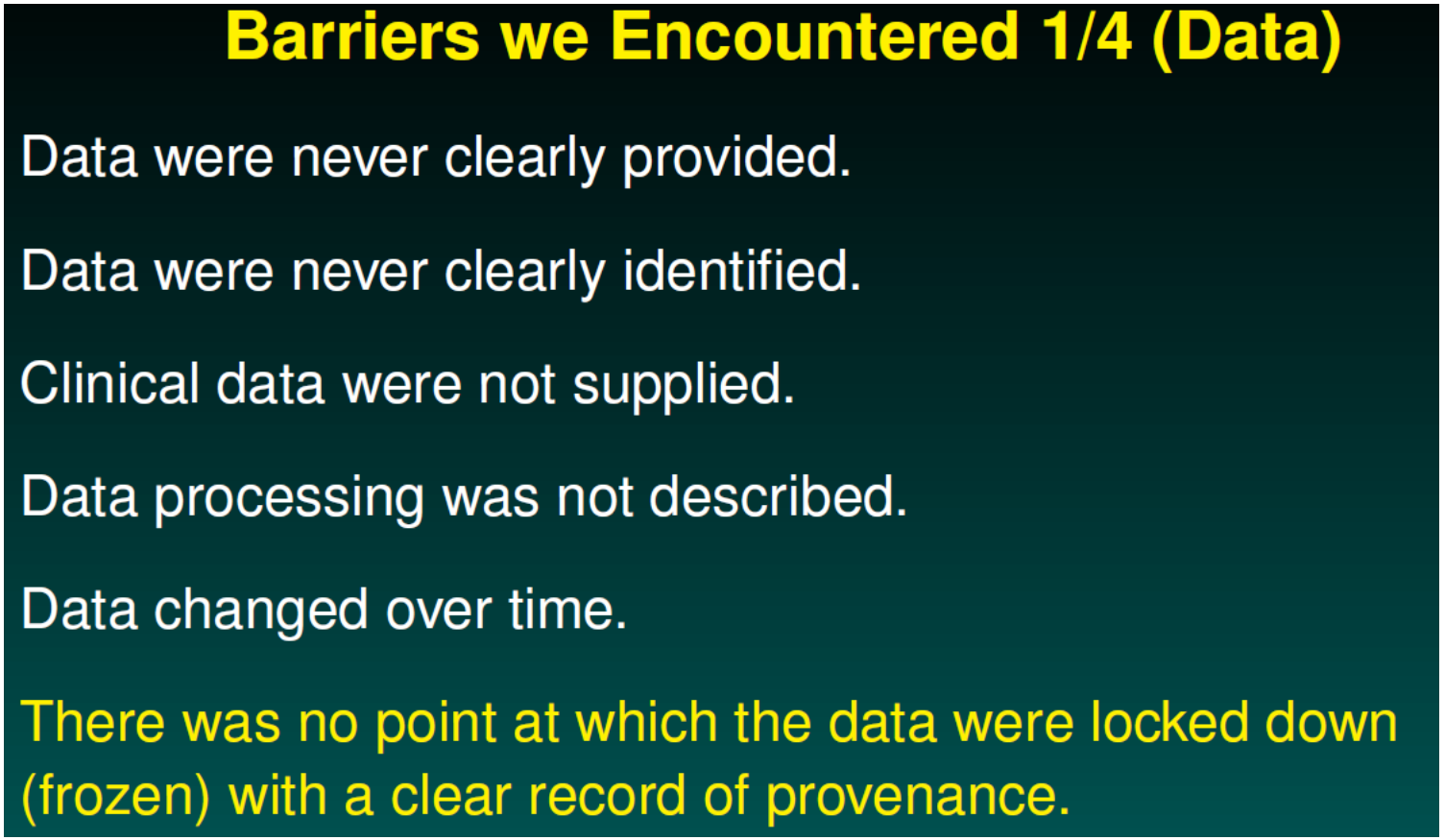
- **Lessons:** Dr. Nevins

Lessons Learned and Questions

- Importance of an infrastructure for collecting samples and generating data even in early observational studies
  - Data and samples for validation are the limiting resources in further developing and validating biomarkers
  - Critical to develop high quality data to facilitate the goal of further development and validation
  - An opportunity for this committee to establish guidelines and standards to best achieve this goal?

Genomic Markers in Cancer Trials

- **Issues:** Dr. Baggerly



Genomic Markers in Cancer Trials

- **Issues:** Dr. Baggerly

Barriers we Encountered 2/4 (Questions)

Specific questions were unanswered.

Specific algorithms were not supplied.

Worked examples were not supplied.

Specific and documented objections were countered with assertions without evidence.

Authors were never required to substantiate their claims,
(a) to us,
(b) to the journals, or
(c) to Duke's internal review.

Genomic Markers in Cancer Trials

- **Issues:** Dr. Baggerly

Barriers we Encountered 3/4 (Duke Review)

The Duke reviewers didn't verify provenance.

The Duke report wasn't published.

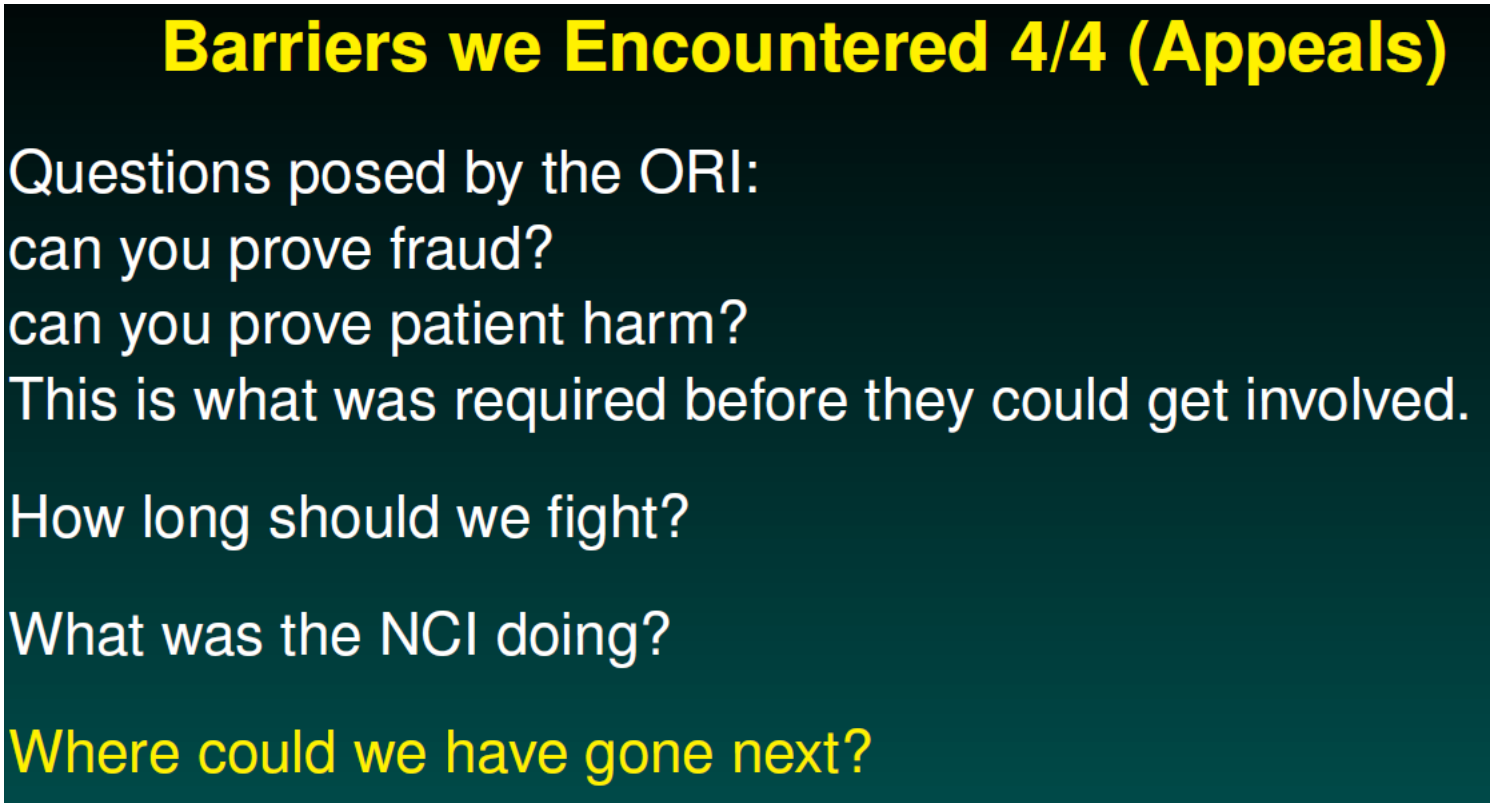
The Duke data weren't released.

Members of the Duke administration and IRB withheld information from the reviewers.

The review was neither complete nor transparent.

Genomic Markers in Cancer Trials

- **Issues:** Dr. Baggerly



Genomic Markers in Cancer Trials

- **What's Needed:** Dr. Baggerly

What are Our Recommendations?

We've outlined some in recent notes: Nature letter, Clin Chem editorial, ENAR notes

We need data.

We need metadata (clinical information, run order, design information).

We need evidence of provenance.

We need the code (MAQC II, NCI experience).

We need auditability before trials begin (Duke TMQF docs).

We need reproducibility.

Genomic Markers in Cancer Trials

- **What's Needed:** Dr. Baggerly

How Do We Get There?

Investigators need to think of reproducibility as a goal from the outset.

Journals need to ask (and check) for code and data deposition (and be prepared to host code and clinical data).

Agencies need to provide data repositories. They need to check for data and code availability at renewal time. They need to budget for reproducibility audits.

Institutions need to help with training and infrastructure.

Genomic Markers in Cancer Trials

- **Issues and What's Needed**: Dr. McShane

Reflection on NCI's experiences with the Duke genomic predictors identifies several issues to be considered as the committee begins its deliberations. Investigators conducting studies using omics technologies face many challenges. Frequently investigators must acquire additional expertise themselves or find collaborators with the relevant expertise to address these challenges that include handling high volumes of data and use of new and complex bioinformatic, computational and statistical tools. Massive amounts of data are publicly available, and data analysis software is often freely shared. Errors can be introduced in data handling, and poorly documented data can be misinterpreted. Computer software might be "research-grade" and highly complex and can be misunderstood or used inappropriately. There has to be a level of trust in the competence, carefulness, and integrity with which individual members of the research team carry out their responsibilities because it is a rare individual who possesses all of the required types of expertise to carefully monitor and fully understand all aspects of a project.

Genomic Markers in Cancer Trials

- **Issues and What's Needed:** Dr. McShane

Sometimes the glamour of the technology or the sheer volume of omics data seem to make investigators forget basic scientific principles. In addition, mistakes or other unfortunate events can occur, and their occurrence is harder to detect when the data sets are bigger and the data and analysis methods are more complex. If we are going to move clinical tests based on omics technologies into clinical trials where they will have an impact on patient treatment and outcome, we need to instill more rigor into the development and validation process. When we conduct clinical trials of new therapeutics, we would not accept situations in which data sources could not be verified or trusted, drug formulations were not clearly specified or documented, and drug delivery mechanisms only delivered the right dose 75% of the time. We don't assume that a drug that has been studied in ovarian cancer only would automatically work in lung cancer. We have long understood the valuable role that blinding can play and the importance of pre-specified analysis plans in therapy trials. There are many common sense principles that could be applied but have not been consistently applied in the development of omics-based tests.

ACKNOWLEDGEMENTS

National Cancer Institute and National Institute of General Medical Sciences of the National Institutes of Health (U01GM 92666, U19HL065962)

Cancer Center Support Grant P30 CA-21765, NIH

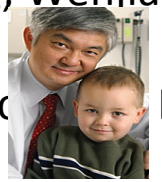
The American Lebanese and Syrian Associated Charities (ALSAC).

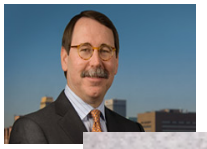





Stan Pounds, Deqing Pei, Xueyuan Cao

Mary Relling, William Evans, Jun Yang, Wenjian Yang – Pharma



Ching-Hon Pui, Dario Campana – Onc

Charles Mullighan, James Downing -- Patholog


Geoff Neale, Yiping Fan -- Bioinformatics

Javier Rojo – My first and most favorite Math Statistics teacher


THANK YOU!