

Cross-study differential gene expression

`giovanni_parmigiani@dfci.harvard.edu`

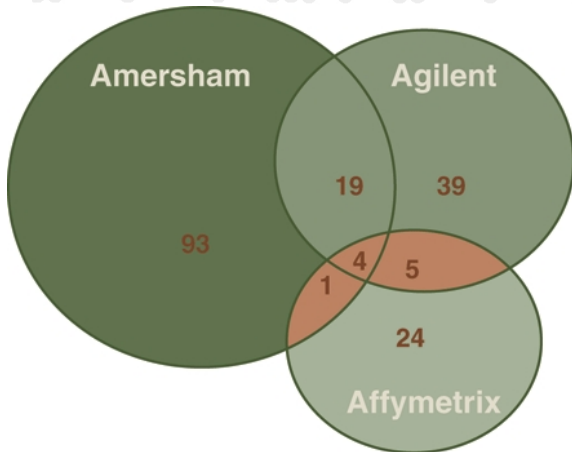
PQG, November 2009

GENES IN ACTION

NEWS

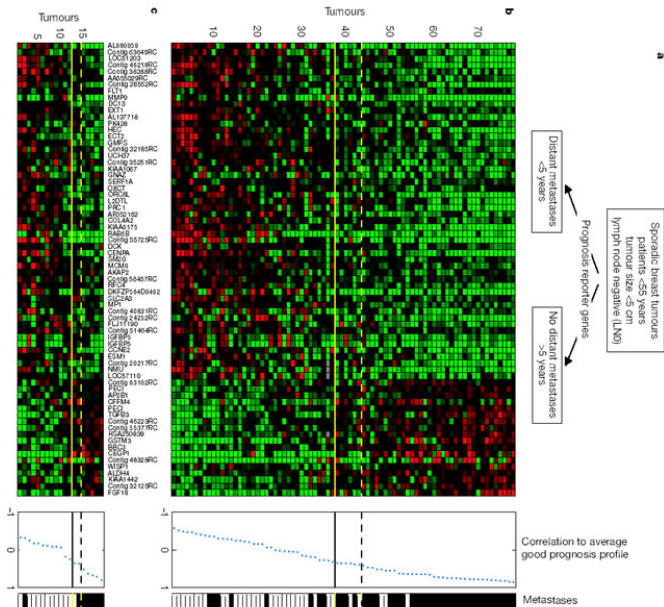
Getting the Noise Out of Gene Arrays

Thousands of papers have reported results obtained using gene arrays, which track the activity of multiple genes simultaneously. But are these results reproducible?



Marshall: “Little overlap. Three array systems rated the activity of 185 genes differently in one test.”

FDA approves Mammaprint in Feb 2007



outline: three related questions

- Which aspects of gene expression can be consistently measured across studies and platforms?
[Integrative Correlation \(Parmigiani et al JCO 2004\)](#)
- To what extent are the biological conclusions confirmed across studies?
[Integrative Association \(Zhong et al almost done\)](#)
- How do we perform a joint analysis?
[Hierarchical Modeling \(Scharpf et al JASA 2009\)](#)

INTEGRATIVE CORRELATION
A PROFILE FOR BRCA1-LINKED TUMORS?

a profile for BRCA1-linked tumors?

- Studies:
 - van't Veer, Nature 2002 (Rosetta, Agilent long oligos)
 - Hedenfalk, NEJM 2001 (NHGRI, cDNA)
- The overlap among the lists of BRCA1-related genes is meager, and reproducibility has been criticized.
- Does breast cancer in BRCA1 germline mutation carriers have a specific molecular profile?

integrative correlation (IC)

Expression matrices,
studies a and b:

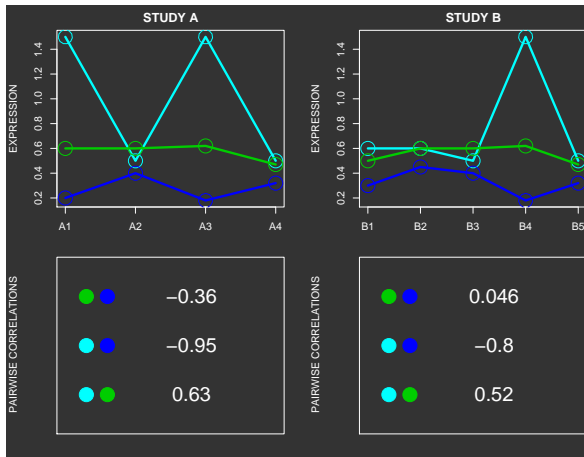
$$A^*, B^*$$

Gene by Gene
correlation matrices:

$$C_a, C_b$$

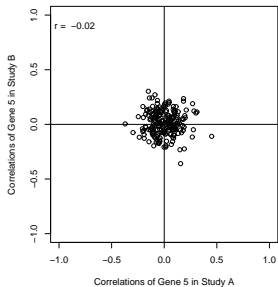
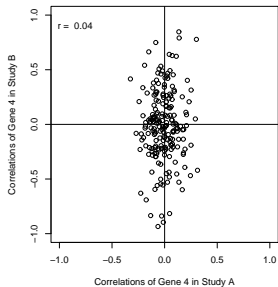
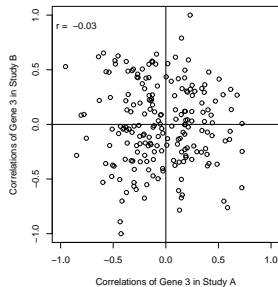
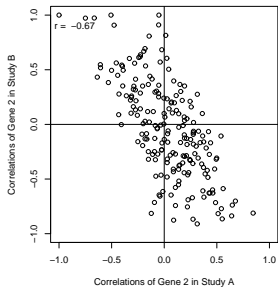
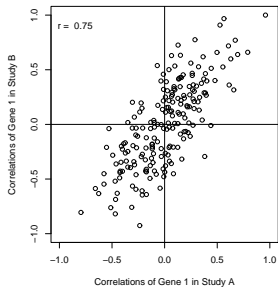
Integrative Correlation
for Gene g

$$Cor(C_{ag}, C_{bg})$$



GP *etal* CCR 2004, P Pavlidis, JK Lee

integrative correlations: examples



integrative correlations: alternative representation

x_a, x_b standardized gene expression of gene g in the two studies. integrative covariance of x is

$$x_a A^t \mathcal{J} B x_b^t$$

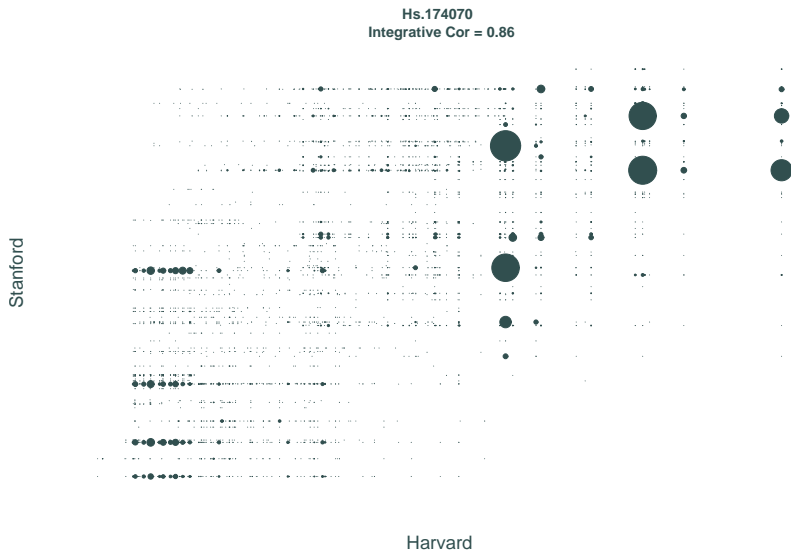
and the integrative variance of x_a is

$$x_a A^t \mathcal{J} A x_a^t.$$

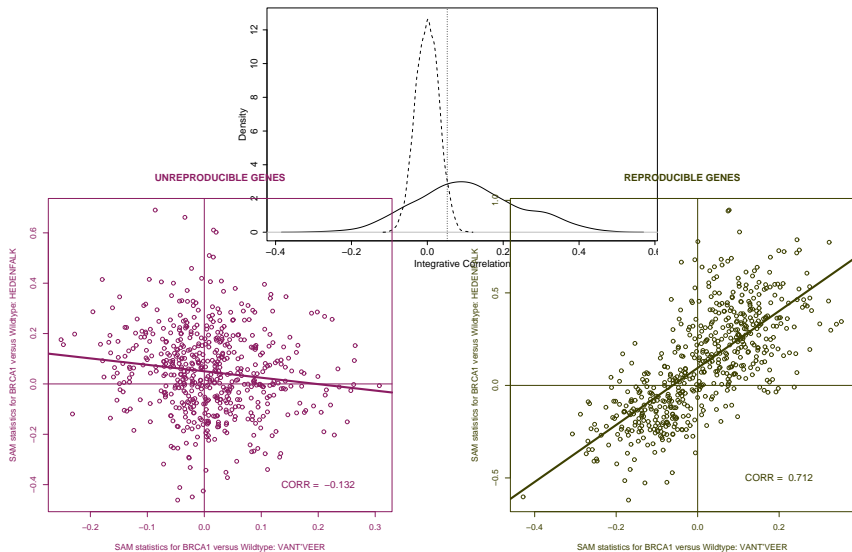
The integrative correlation of gene g is

$$\frac{x_a A^t \mathcal{J} B x_b^t}{\sqrt{x_a A^t \mathcal{J} A x_a^t} \sqrt{x_b B^t \mathcal{J} B x_b^t}}.$$

IC and sample correlations

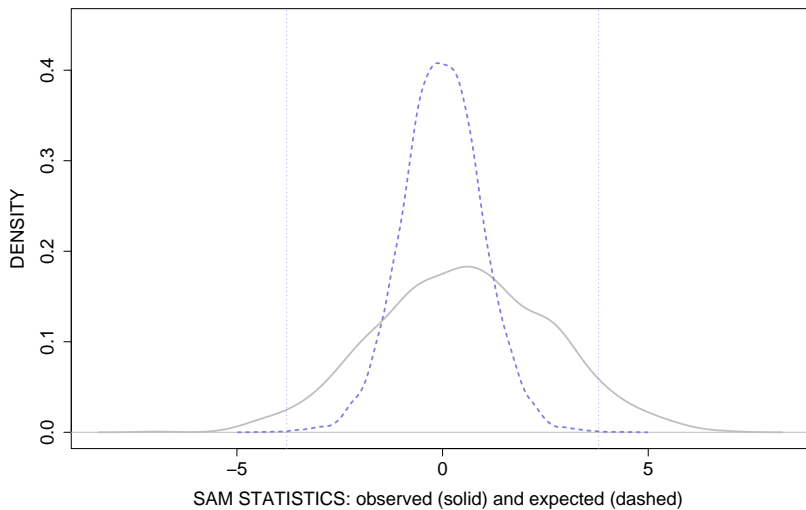


integrative correlation and integrative association



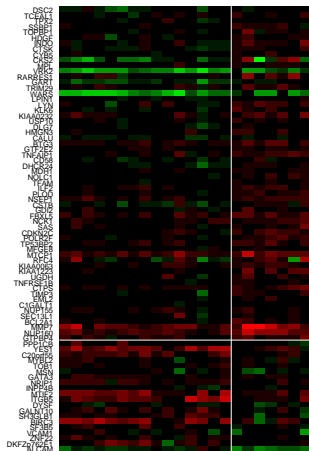
Here “Reproducible” means high IC

observed vs expected IC and false discovery rates

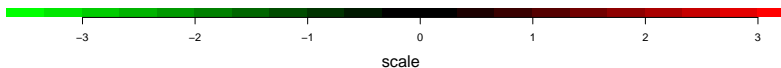
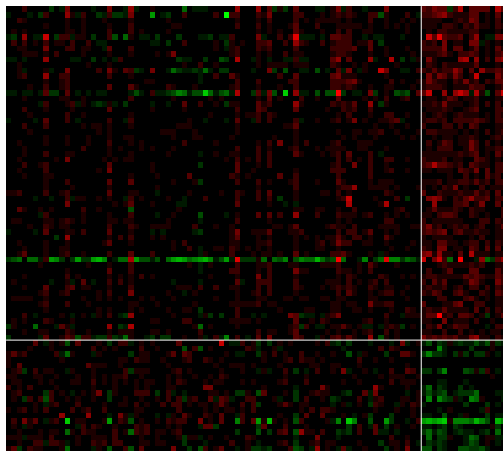


BRCA1 profile

HEDENFALK



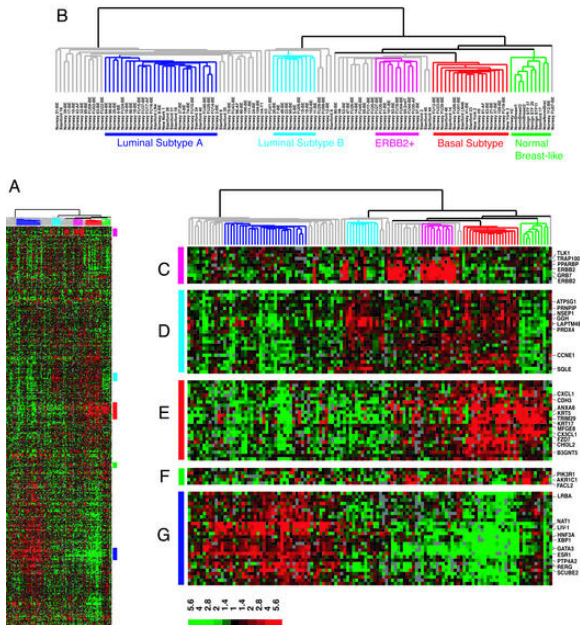
VAN'T VEER



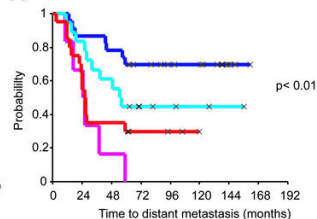
- Concordance between .78 and .94 by “square one cross-validation”.
- More informative than family history
- Useful complement to genetic testing?
- Useful surrogate for mutation analysis?

**INTEGRATIVE ASSOCIATION:
BASAL SUBTYPE AND SURVIVAL
IN BREAST CANCER**

Sortie PNAS 03: subtypes, profiles, prognosis

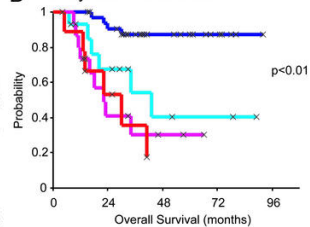


A van't Veer data set

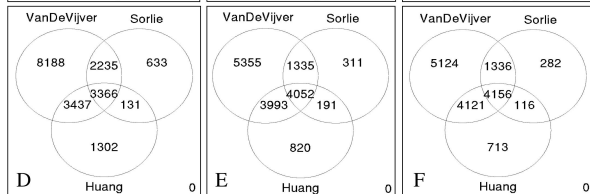
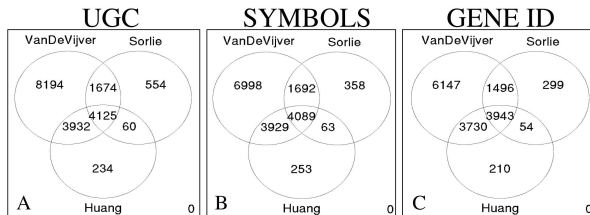


× Censored, — Luminal A, — Luminal B, — Basal, — ERBB2+

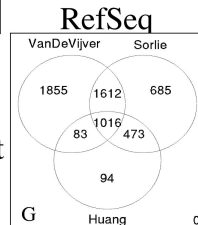
B Norway/Stanford data set



overlap of cross-referencing approaches



**BLAST
alignment**



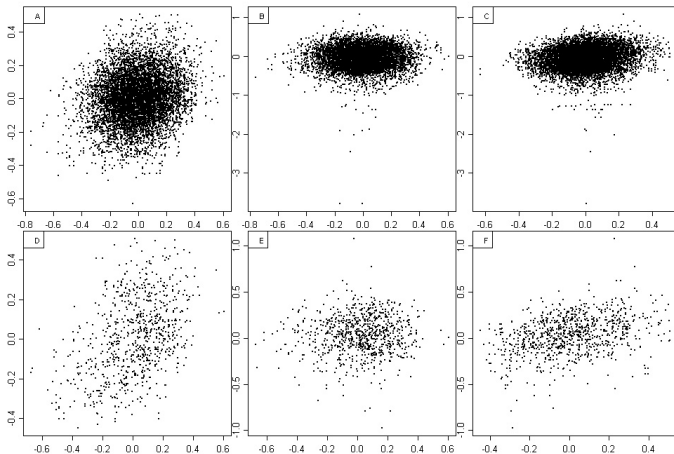
integrative association and filtering

V vs S

V vs H

S vs H

All
genes



High
IC

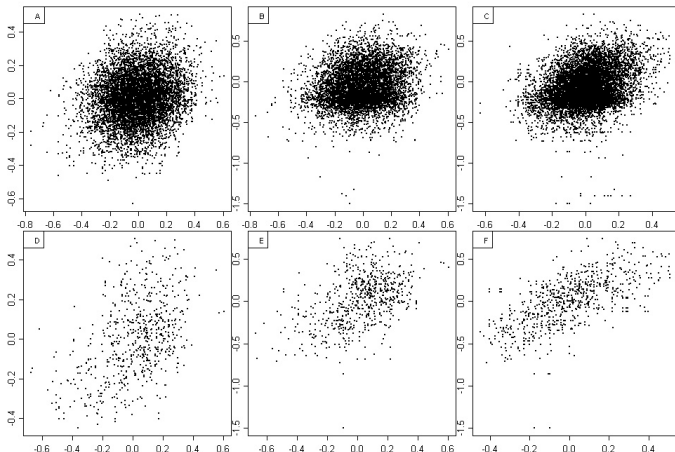
integrative association and Taiwan "batch 2"

V vs S

V vs H2

S vs H2

All
genes

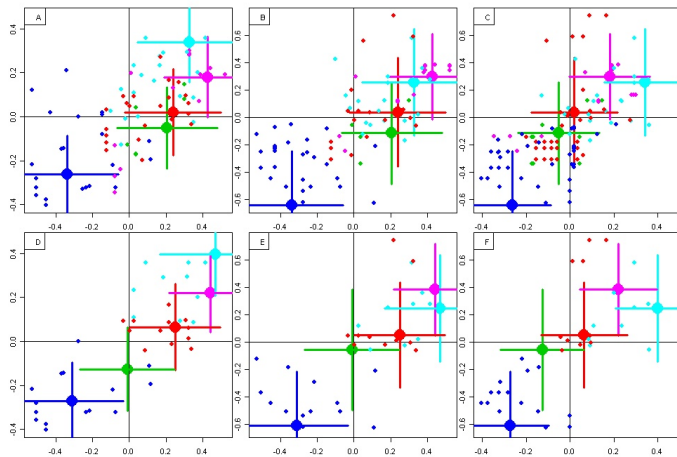


High
IC

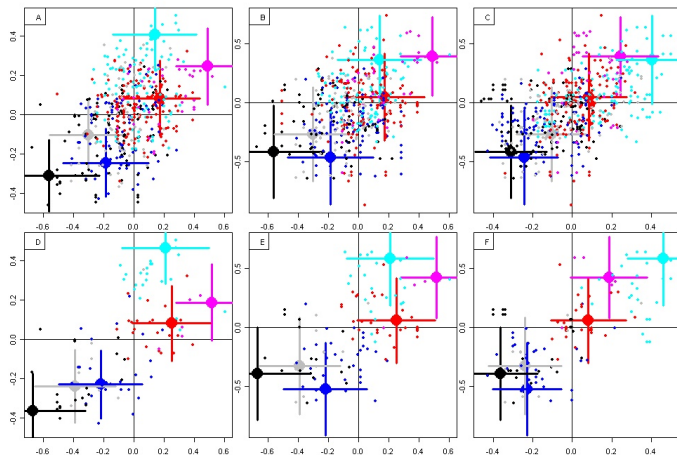
integrative association and filtering lots of ways

	Procedures	Num of Gene Mappings	Num of UGLs	SV	SH	SH2	VH	VH2
Common Refseqs	all the common genes	1226	1016	0.132	-0.036	NA	0.083	NA
	all the common genes, Std	1226	1016	0.178	0.001	NA	0.168	NA
Common UGCs	all the common genes	11531	4125	0.115	-0.011	0.061	0.038	0.1
	all the common genes, Std	11531	4125	0.165	0.005	0.152	0.127	0.3
	reproducible genes, FDR .1, Std	7419	3359	0.202	-0.002	NA	0.197	NA
	reproducible genes, variance filtering, FDR .1, Std	3912	1935	0.228	0.013	NA	0.2	NA
	reproducible genes, FDR .1, Std	7215	3305	0.202	NA	0.204	NA	0.425
	reproducible genes, FDR .01, Std	5065	2612	0.234	NA	0.239	NA	0.498
	reproducible genes, ICOR > .25, Std	865	513	0.355	NA	0.437	NA	0.616
Common intrinsic genes, UGCs	intrinsic genes, Std	1087	354	0.198	0	0.217	0.246	0.423
	reproducible genes, FDR .1, Std	889	334	0.237	NA	0.244	NA	0.453
	reproducible genes, FDR .01, Std	760	295	0.359	NA	0.311	NA	0.551
	reproducible genes, ICOR > .25, Std	224	100	0.382	NA	0.425	NA	0.545
Intrinsic gene clusters, UGCs	intrinsic gene clusters, Std	119	56	0.639	0.357	0.686	0.305	0.623
	reproducible gene clusters, FDR .1, Std	104	52	0.731	NA	0.706	NA	0.652
	reproducible gene clusters, FDR .01, Std	93	50	0.744	NA	0.712	NA	0.657
	reproducible gene clusters, ICOR > .25, Std	40	30	0.763	NA	0.716	NA	0.698
	gene clusters centroids, Std	5 centroids	5 centroids	0.851	0.459	0.977	0.361	0.93
New intrinsic gene clusters, UGCs	new intrinsic gene clusters, Std	796	296	0.337	0.064	0.241	0.113	0.41
	reproducible gene clusters, FDR .1, Std	547	265	0.39	NA	0.323	NA	0.526
	reproducible gene clusters, FDR .01, Std	462	243	0.418	NA	0.34	NA	0.563
	reproducible gene clusters, ICOR > .25, Std	129	89	0.592	NA	0.472	NA	0.661
	gene clusters centroids, Std	5 centroids	5 centroids	0.831	0.706	0.88	0.496	0.97

genes and profiles, “old” intrinsic gene list



genes and profiles, “new” intrinsic gene list



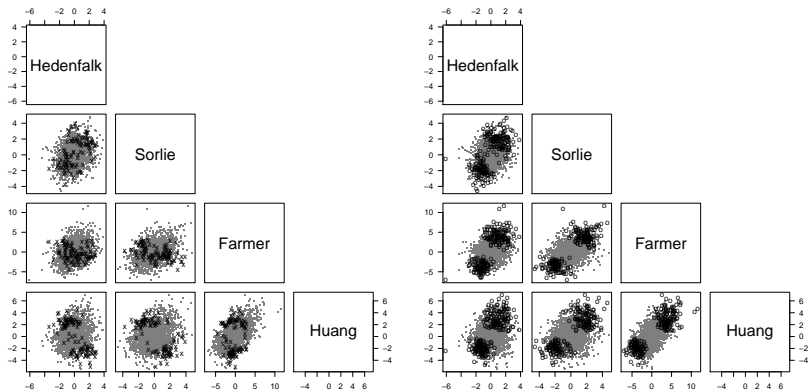
**XDE: A BAYESIAN MODEL
FOR COMBINED ANALYSIS**

Important insight about technology, study design or the genetics of alternative splicing can potentially be gained by identifying and following discordant genes

Discordant patterns of expression could emerge from

- genetic heterogeneity of samples across studies
- alternative splicing
 - two technologies that measure a gene's expression by targeting portions of a gene that are associated with different transcripts that are negatively correlated with each other.

4 breast cancer studies



t-statistics for estrogen receptor status
in bold are genes significant in 3 out of 4 studies

Hierarchical model for gene expression

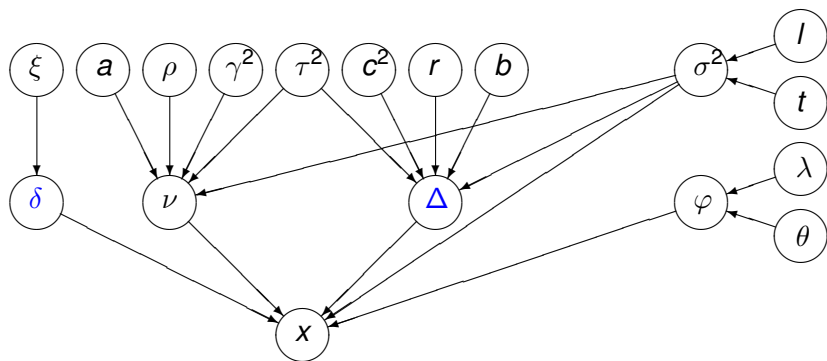


Figure: A graphical representation of the hierarchical Bayesian model

Δ_{gp} : effect size or 'offset' (indexed by gene and platform)

δ_g : binary indicator for differential expression (indexed by gene)

Hierarchical model for gene expression

Level 1:

$$x_{gsp} | \nu_{gp}, \delta_g, \Delta_{gp}, \sigma_{g0p}^2, \sigma_{g1p}^2 \sim N(\nu_{gp} + \delta_g(2\psi_{sp} - 1)\Delta_{gp}, \sigma_{g\psi_{sp}p}^2)$$

Level 2:

$$P(\delta_g = 1 | \xi) = \xi, \text{ where } \xi \sim \text{Beta}(\alpha_\xi, \beta_\xi)$$

$$\nu_g \sim N(0, \Sigma_g), \quad (\Sigma)_{pq} = \gamma^2 \rho_{pq} \sqrt{\tau_p^2 \tau_q^2 \sigma_{gp}^{2a_p} \sigma_{gq}^{2a_q}} \text{ and } \prod_p \tau_p^2 = 1$$

$$\Delta_g \sim N(0, R_g), \quad (R_g)_{pq} = c^2 r_{pq} \sqrt{\tau_p^2 \tau_q^2 \sigma_{gp}^{2b_p} \sigma_{gq}^{2b_q}}$$

$$\sigma_{gp}^2 = \sqrt{\sigma_{g0p}^2 \sigma_{g1p}^2}, \quad \sigma_{g0p}^2 = \sigma_{gp}^2 \varphi_{gp}, \quad \sigma_{g1p}^2 = \frac{\sigma_{gp}^2}{\varphi_{gp}}$$

$$\sigma_{gp}^2 | t_p, l_p \sim \text{Gamma}(t_p, l_p), \quad \varphi_{gp} | \theta_p, \lambda_p \sim \text{Gamma}(\theta_p, \lambda_p)$$

Level 3:

$$P(a_p = 0) = p_a^0, \quad P(a_p = 1) = p_a^1, \quad a_p | a_p \in (0, 1) \sim \text{Beta}(\alpha_a, \beta_a)$$

$$P(b_p = 0) = p_b^0, \quad P(b_p = 1) = p_b^1, \quad b_p | b_p \in (0, 1) \sim \text{Beta}(\alpha_b, \beta_b)$$

$$\text{Barnard } et al. \text{ priors for } r_{pq} \text{ and } \rho_{pq}: \text{ joint uniform for } \tau_1^2, \dots, \tau_p^2 \text{ and } \prod_p \tau_p^2 = 1$$

$$t_p \sim \text{Unif}(0, \infty), \quad l_p \sim \text{Unif}(0, \infty), \quad \gamma^2 \sim \text{Unif}(0, \infty), \quad c^2 \sim \text{Unif}(0, \infty)$$

Hierarchical model for gene expression

Level 1:

$$x_{gsp} | \nu_{gp}, \delta_g, \Delta_{gp}, \sigma_{g0p}^2, \sigma_{g1p}^2 \sim N\left(\nu_{gp} + \delta_g(2\psi_{sp} - 1)\Delta_{gp}, \sigma_{g\psi_{sp}p}^2\right)$$

Level 2:

$P(\delta_g = 1 | \xi) = \xi$, where $\xi \sim \text{Beta}(\alpha_\xi, \beta_\xi)$

$$\nu_{\mathbf{g}} \sim N(0, \Sigma_{\mathbf{g}}), \quad (\Sigma)_{pq} = \gamma^2 \rho_{pq} \sqrt{\tau_p^2 \tau_q^2 \sigma_{gp}^{2ap} \sigma_{gq}^{2aq}} \quad \text{and} \quad \prod_p \tau_p^2 = 1$$

$$\Delta_{\mathbf{g}} \sim N(0, R_{\mathbf{g}}), \quad (R_{\mathbf{g}})_{pq} = c^2 r_{pq} \sqrt{\tau_p^2 \tau_q^2 \sigma_{gp}^{2bp} \sigma_{gq}^{2bq}}$$

$$\sigma_{gp}^2 = \sqrt{\sigma_{g0p}^2 \sigma_{g1p}^2}, \quad \sigma_{g0p}^2 = \sigma_{gp}^2 \varphi_{gp}, \quad \sigma_{g1p}^2 = \frac{\sigma_{gp}^2}{\varphi_{gp}}$$

$$\sigma_{gp}^2 | t_p, l_p \sim \text{Gamma}(t_p, l_p), \quad \varphi_{gp} | \theta_p, \lambda_p \sim \text{Gamma}(\theta_p, \lambda_p)$$

Level 3:

$$P(a_p = 0) = p_a^0, \quad P(a_p = 1) = p_a^1, \quad a_p | a_p \in (0, 1) \sim \text{Beta}(\alpha_a, \beta_a)$$

$$P(b_p = 0) = p_b^0, \quad P(b_p = 1) = p_b^1, \quad b_p | b_p \in (0, 1) \sim \text{Beta}(\alpha_b, \beta_b)$$

Barnard *et al.* priors for r_{pq} and ρ_{pq} : joint uniform for $\tau_1^2, \dots, \tau_p^2$ and $\prod_p \tau_p^2 = 1$

$$t_p \sim \text{Unif}(0, \infty), \quad l_p \sim \text{Unif}(0, \infty), \quad \gamma^2 \sim \text{Unif}(0, \infty), \quad c^2 \sim \text{Unif}(0, \infty)$$

Hierarchical model for gene expression

Level 1:

$$x_{gsp} | \nu_{gp}, \delta_g, \Delta_{gp}, \sigma_{g0p}^2, \sigma_{g1p}^2 \sim N\left(\nu_{gp} + \delta_g(2\psi_{sp} - 1)\Delta_{gp}, \sigma_{g\psi_{sp}p}^2\right)$$

Level 2:

$$P(\delta_g = 1 | \xi) = \xi, \text{ where } \xi \sim \text{Beta}(\alpha_\xi, \beta_\xi)$$

$$\nu_{\mathbf{g}} \sim N(0, \Sigma_{\mathbf{g}}), \quad (\Sigma)_{pq} = \gamma^2 \rho_{pq} \sqrt{\tau_p^2 \tau_q^2 \sigma_{gp}^{2ap} \sigma_{gq}^{2aq}} \text{ and } \prod_p \tau_p^2 = 1$$

$$\Delta_{\mathbf{g}} \sim N(0, R_{\mathbf{g}}), \quad (R_{\mathbf{g}})_{pq} = c^2 r_{pq} \sqrt{\tau_p^2 \tau_q^2 \sigma_{gp}^{2bp} \sigma_{gq}^{2bq}}$$

$$\sigma_{gp}^2 = \sqrt{\sigma_{g0p}^2 \sigma_{g1p}^2}, \quad \sigma_{g0p}^2 = \sigma_{gp}^2 \varphi_{gp}, \quad \sigma_{g1p}^2 = \frac{\sigma_{gp}^2}{\varphi_{gp}}$$

$$\sigma_{gp}^2 | t_p, l_p \sim \text{Gamma}(t_p, l_p), \quad \varphi_{gp} | \theta_p, \lambda_p \sim \text{Gamma}(\theta_p, \lambda_p)$$

Level 3:

$$P(a_p = 0) = p_a^0, \quad P(a_p = 1) = p_a^1, \quad a_p | a_p \in (0, 1) \sim \text{Beta}(\alpha_a, \beta_a)$$

$$P(b_p = 0) = p_b^0, \quad P(b_p = 1) = p_b^1, \quad b_p | b_p \in (0, 1) \sim \text{Beta}(\alpha_b, \beta_b)$$

$$\text{Barnard } et al. \text{ priors for } r_{pq} \text{ and } \rho_{pq}; \quad \text{joint uniform for } \tau_1^1, \dots, \tau_p^2 \text{ and } \prod_p \tau_p^2 = 1$$

$$t_p \sim \text{Unif}(0, \infty), \quad l_p \sim \text{Unif}(0, \infty), \quad \gamma^2 \sim \text{Unif}(0, \infty), \quad c^2 \sim \text{Unif}(0, \infty)$$

Estimates of differential expression

Parameter	<i>XDE</i> estimates
differentially expressed	$\text{PM}_\varepsilon(g)$
concordantly expressed	$\text{PM}_\mathcal{C}(g)$
discordantly expressed	$\text{PM}_\mathcal{D}(g)$

$\text{PM}_\cdot(g)$ denotes the posterior mean of the following indicators:

$$\mathcal{E}_g \equiv \delta_g$$

$$\mathcal{C}_g \equiv \begin{cases} 1 & \delta_g = 1 \text{ and } \mathbf{\Delta}_g \text{ have the same sign,} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{D}_g \equiv \begin{cases} 1 & \delta_g = 1 \text{ and } \mathbf{\Delta}_g \text{ do not have the same sign,} \\ 0 & \text{otherwise} \end{cases}$$

Estimates of differential expression

Parameter	<i>XDE</i> estimates
differentially expressed	$\text{PM}_\varepsilon(g)$
concordantly expressed	$\text{PM}_c(g)$
discordantly expressed	$\text{PM}_D(g)$

$\text{PM}_\cdot(g)$ denotes the posterior mean of the following indicators:

$$\mathcal{E}_g \equiv \delta_g$$

$$\mathcal{C}_g \equiv \begin{cases} 1 & \delta_g = 1 \text{ and } \mathbf{\Delta}_g \text{ have the same sign,} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{D}_g \equiv \begin{cases} 1 & \delta_g = 1 \text{ and } \mathbf{\Delta}_g \text{ do not have the same sign,} \\ 0 & \text{otherwise} \end{cases}$$

Estimates of differential expression

Parameter	<i>XDE</i> estimates
differentially expressed	$\text{PM}_{\mathcal{E}}(g)$
concordantly expressed	$\text{PM}_{\mathcal{C}}(g)$
discordantly expressed	$\text{PM}_{\mathcal{D}}(g)$

$\text{PM}_{\cdot}(g)$ denotes the posterior mean of the following indicators:

$$\mathcal{E}_g \equiv \delta_g$$

$$\mathcal{C}_g \equiv \begin{cases} 1 & \delta_g = 1 \text{ and } \mathbf{\Delta}_g \text{ have the same sign,} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{D}_g \equiv \begin{cases} 1 & \delta_g = 1 \text{ and } \mathbf{\Delta}_g \text{ do not have the same sign,} \\ 0 & \text{otherwise} \end{cases}$$

Estimates of differential expression

Parameter	Estimates	
	<i>XDE</i>	Alternative
differentially expressed	$PM_{\mathcal{E}}(g)$	$u_{\mathcal{E}}(g)$ ¹
concordantly expressed	$PM_{\mathcal{C}}(g)$	z-score ²
discordantly expressed	$PM_{\mathcal{D}}(g)$	$u_{\mathcal{D}}(g)$ ³

¹ $u_{\mathcal{E}}(g) \equiv \alpha_1|U_{g1}| + \dots + \alpha_P|U_{gP}|$, where

$$\alpha_p \equiv \frac{L_p \sqrt{S_p}}{\sum_{i=1}^P L_i \sqrt{S_i}} \text{ for } p \in \{1, \dots, P\}$$

L is the covariance loading from the first principal components and S is the number of samples.

²cross-study estimator for differential expression described in Choi *et al.*, 2003

$${}^3u_{\mathcal{D}}(g) \equiv \begin{cases} u_{\mathcal{E}}(g) & \text{sign}(U_{g1}) = \dots = \text{sign}(U_{gP}) \\ -1 \times u_{\mathcal{E}}(g) & \text{otherwise.} \end{cases}$$

Simulation using three lung cancer studies

- 1 Randomly assign a binary covariate (ψ^*) to early stage lung adenocarcinomas
- 2 For each gene, simulate δ^* from a Bernoulli(ξ^*)
- 3 Simulate offsets Δ^*

$$\begin{bmatrix} \Delta_{g1}^* \\ \Delta_{g2}^* \\ \Delta_{g3}^* \end{bmatrix} \sim N \left(k^* \begin{bmatrix} s_{g1} \\ s_{g2} \\ s_{g3} \end{bmatrix}, \frac{1}{c^*} \begin{bmatrix} s_{g1}^2 & r_1^* s_{g1} s_{g2} & r_2^* s_{g1} s_{g3} \\ r_1^* s_{g2} s_{g1} & s_{g2}^2 & r_3^* s_{g2} s_{g3} \\ r_2^* s_{g3} s_{g1} & r_3^* s_{g3} s_{g2} & s_{g3}^2 \end{bmatrix} \right)$$

- 4 Compute the simulated expression values:

$$x_{gsp}^* = \begin{cases} x_{gsp} + (2\psi_{sp}^* - 1)\Delta_{gp}^* & \text{if } \delta_g^* = 1 \\ x_{gsp} & \text{otherwise.} \end{cases}$$

Simulation

We simulated a number of artificial datasets of different sample sizes (S) by varying parameters that affect the location (k^*), precision (c^*), and inter-study correlation (r^*) of the simulated offsets, as well as the proportion of genes that are differentially expressed (ξ^*)

Simulation	k^*	S	c^*	r^*	ξ^*
A [†]	0.5	4	0.5	(0.1, 0.2, 0.4)	0.10
B	0.50
C	.	.	.	(0.8, 0.9, 0.92)	0.10
D	0.50
E [†]	.	8	0.5	(0.1, 0.2, 0.4)	0.10
F	.	.	1	.	0.10
G	0.50
H	.	.	.	(0.8, 0.9, 0.92)	0.10
I	0.50
J [†]	0	16	10	(0.1, 0.2, 0.4)	0.10
K	0.50
L	.	.	.	(0.8, 0.9, 0.92)	0.10
M	0.50
O	.	32	20	(0.1, 0.2, 0.4)	0.10
P	0.50
Q	.	.	.	(0.8, 0.9, 0.92)	0.10
R	0.50

[†] used 10 different seeds to assess sensitivity to randomly generated values

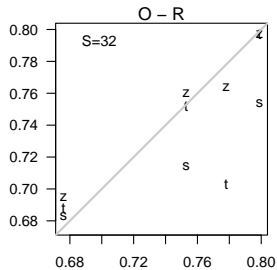
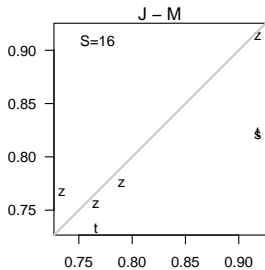
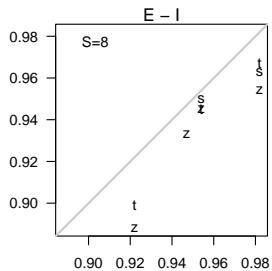
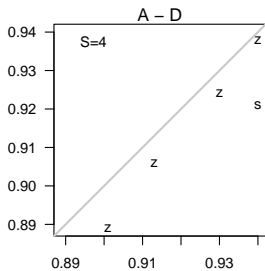
4 genes, 2 studies

gene	δ^*	$\text{sign}(\Delta^*)$	\mathcal{E}^*	\mathcal{C}^*	\mathcal{D}^*
1	0	.	0	0	0
2	1	$\{-, -\}$	1	1	0
3	1	$\{-, +\}$	1	0	1
4	1	$\{+, +\}$	1	1	0

Columns \mathcal{E}^* , \mathcal{C}^* , and \mathcal{D}^* are indicators for true differential expression, concordant differential expression, and discordant differential expression, respectively.

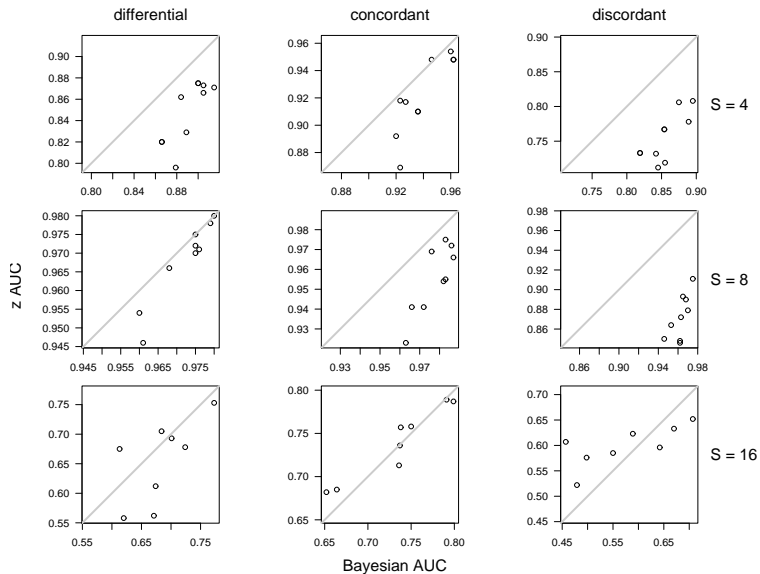
Simulations A - R

AUC - alternatives



AUC - XDE

Random seeds

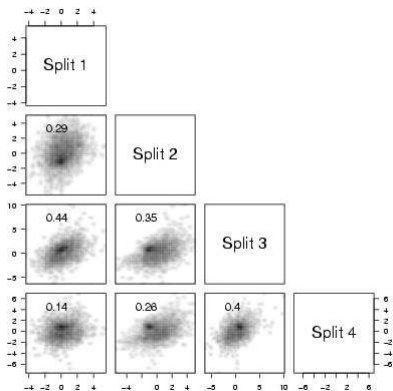


Split Study Validation

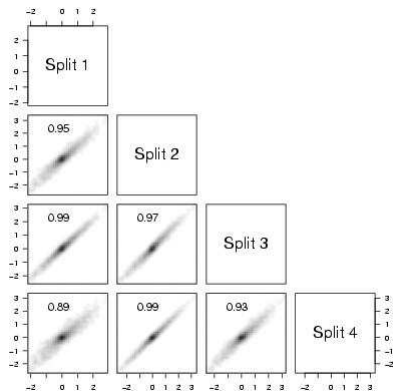
- To assess the baseline behavior of *XDE*, we split the Huang ¹ study into four disjoint parts, treating each part as an independent study
- We randomly assigned 5 estrogen receptor (ER) negative and 16 ER positive samples to each split.
- We denote the Bayesian effect size (BES) for gene g and platform p by $\frac{\delta_g \Delta_{gp}}{c \tau_p \sigma_{gp}}$ and use this as a study-specific Bayesian estimate of differential expression

¹Huang *et al.* 2003, *Lancet*, 361(9369):1590–6

Split Study Validation



T-test



XDE

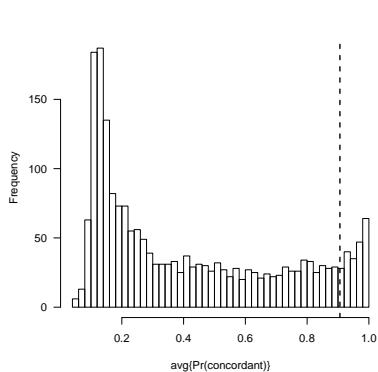
4 breast cancer studies

	platform	ER-	ER+
Hedenfalk	cDNA	6	10
Sorlie	cDNA	30	81
Farmer	Affymetrix hu133a	22	27
Huang	Affymetrix hu95av2	23	65

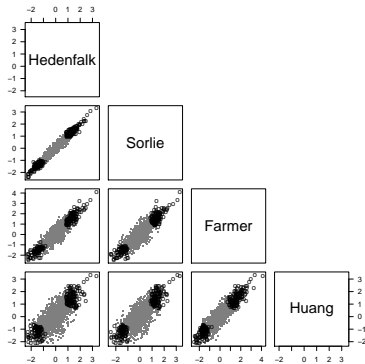
Table: Distribution of the estrogen receptor status in the three studies

Platform-specific annotations for the features were cross-referenced by Entrez gene identifiers to yield a set of 2064 features measured in each platform.

4 breast cancer studies: concordant genes

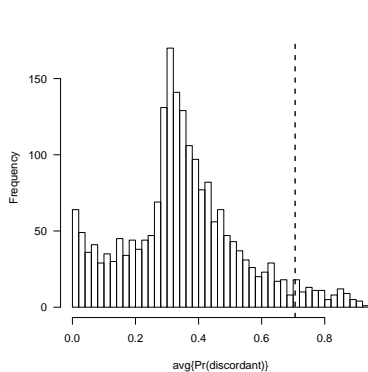


$PM_e(g)$

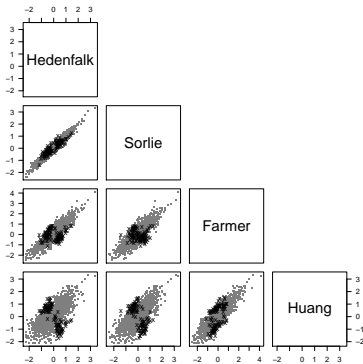


posterior average of BES

4 breast cancer studies: discordant genes

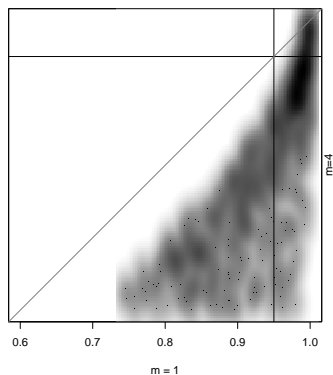
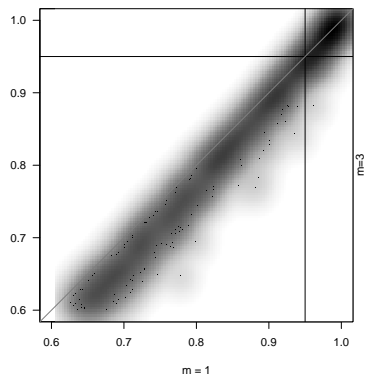


$\text{PM}_{\mathcal{D}}(g)$



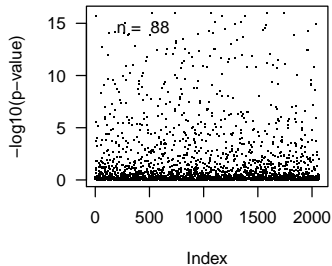
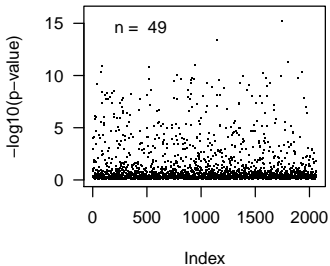
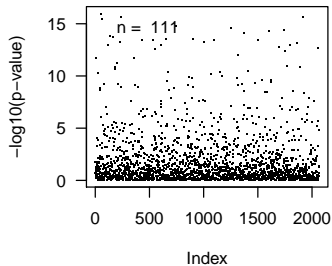
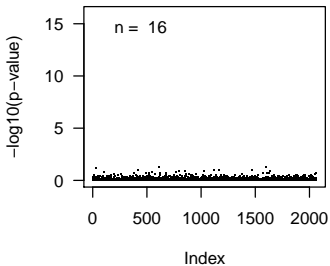
posterior average of BES

4 breast cancer studies: outlying studies



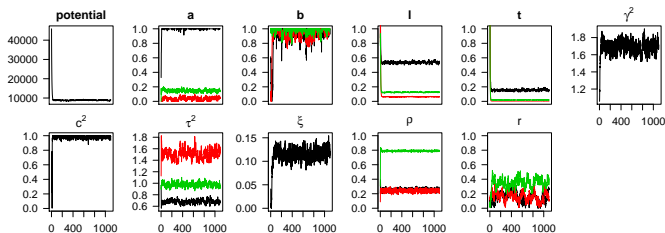
Probability of concordant differential expression in at least m studies

4 breast cancer studies: goodness of fit



R package: XDE

```
> data(expressionSetList)
> params <- new("XdeParameter",
+   esetList = expressionSetList,
+   phenotypeLabel = "adenoVsSquamous")
> fit <- xde(params, expressionSetList)
> plot(fit)
```



Look for it here: www.bioconductor.org

INTEGRATIVE CORRELATION:

Les Cope, Ed Gabrielson, Liz Garrett-Mayer

INTEGRATIVE ASSOCIATION:

Simens Zhong, Luigi Marchionni, Les Cope, Ed Gabrielson, Liz Garrett-Mayer

XDE:

Rob Scharpf, Häkon Tjemeland, Andrew Nobel