

**EXACT POWER UNDER
INDEPENDENCE FOR THE FALSE
DISCOVERY RATE IN GENE
EXPRESSION ARRAY EXPERIMENTS**

**Deborah H. Glueck¹, Keith E. Muller² &
Lawrence Hunter³**

¹UCHSC, Dept. of Prev. Med. and Biometrics

²UNC-Chapel Hill, Department of Biostatistics

³UCHSC, Department of Pharmacology

Support. Glueck: NCI K07CA88811.

Hunter: NIAAA 1U01, AA13524-02 &
NCI 5 P30 CA46934-15.

Muller NCI P01 CA47 982-04,

R01 CA095749-01A1, NIAID 9P30 AI 50410.

Problem

- Many genes
- Many hypotheses
- Multiple testing problem

Literature Review

- Benjamini & Hochberg (1995, JRSS)
- Efron, Storey and Tibshirani (2001) JASA
- Lee and Whitmore (2002)
Stats in Med
- Storey (2002) JRSS
- Zien, Fluck, Zimmer and Lengauer (2002) RECOMB, Proc.Comp.Bio.

Decisions for an Experiment

		Decision		
		Reject	Accept	
True State				
H_0 True	j	$n - j$	n	$m - n$
H_0 False	$k - j$	$(m - n) -$ $(k - j)$	$m - n$	
	k	$m - k$	m	

Rates

False Discovery Rate

$$g(J, K, \text{FDR}) = \begin{cases} 0 & k = 0 \\ J/K & k > 0, \end{cases}$$

$$\text{FDR} = E[g(J, K, \text{FDR})] .$$

Positive False Discovery Rate

$$E[(J/K) | K > 0]$$

Family-wise Error Rate

$$\Pr\{J \geq 1\}.$$

Per Comparison Error Rate

$$E(J/m)$$

Expected Values

Negative Predictive Value

$$g(J, K, \text{NPV}) = \begin{cases} 0 & k = m \\ (n - J)/(m - K) & 0 \leq k < m . \end{cases}$$

Positive Predictive Value

$$g(J, K, \text{PPV}) = \begin{cases} 0 & k = 0 \\ (K - J)/K & 1 \leq k \leq m . \end{cases}$$

Sensitivity

$$g(J, K, \text{SENS}) = \begin{cases} 0 & m = n \\ (K - J)/m - n & 0 \leq n < m . \end{cases}$$

Specificity

$$g(J, K, \text{SPEC}) = \begin{cases} 0 & n = 0 \\ (n - J)/n & 1 \leq n \leq m . \end{cases}$$

B&H (1995) Procedure

1. assume *independent* tests and $\alpha_* \in (0, 1)$ is control target.
2. m statistics $\{t_i\}$, p-values, $\{p_i\}$.
3. Rank: $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$
4. $p_{(1)} \leq 1 \cdot \alpha_* / m \Rightarrow$ reject
 $p_{(2)} \leq 2 \cdot \alpha_* / m \Rightarrow$ reject
 $\vdots \leq \quad \Rightarrow$ reject
 $p_{(k)} \leq k \cdot \alpha_* / m \Rightarrow$ reject
 $p_{(k+1)} > (k+1) \cdot \alpha_* / m \Rightarrow$ do not reject
5. \Rightarrow False Discovery Rate $\leq \alpha_*$

End literature review.

New work follows.

INTUITIVE IDEA: NEED

- Density of the order statistics of the $k + 1$ smallest p-values.
- Regions of integration corresponding to Benjamini and Hochberg (1995) procedure

THEN WE GET

- Joint probability of total number of rejections and the number of false rejections.
- Exact False Discovery Rate

Assumptions

- m independent hypotheses
- N independent people
- Each test statistic density, $f_{T_i}(t_i)$, exists and is known under the null and the alternative

$$\Pr\{T_i \leq t_i\} = \begin{cases} F_{T_i}[t_i; \boldsymbol{\theta}_i(0)] & H_0 \\ F_{T_i}[t_i; \boldsymbol{\theta}_i(A)] & H_A \end{cases}$$

(big T_i rejects) p-value

$$P_i = 1 - F_{T_i}[T_i; \boldsymbol{\theta}_i(0)]$$

is AC with computable density

m independent, AC random
test statistics $\{T_1, T_2, \dots, T_m\}$
 \Rightarrow p-values $\{P_1, P_2, \dots, P_m\}$

Sorting smallest to largest creates
dependent order statistics

$$\left\{ P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(m)} \right\}$$

AC order statistic density known
(Vaughan and Venables, 1972)

Joint density of $k + 1 \leq m$ smallest is

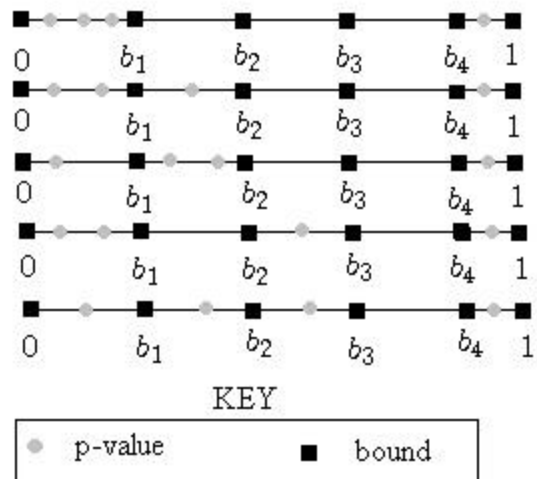
$$f_{P_{(1)}, P_{(2)}, \dots, P_{(k+1)}}(p_{(1)}, p_{(2)}, \dots, p_{(k+1)}) = \frac{1}{c} \begin{vmatrix} \overset{+}{f_{P_1}(p_{(1)})} & \overset{+}{f_{P_2}(p_{(1)})} & \cdots & \overset{+}{f_{P_m}(p_{(1)})} \\ \vdots & \vdots & & \vdots \\ f_{P_1}(p_{(k+1)}) & f_{P_2}(p_{(k+1)}) & \cdots & f_{P_m}(p_{(k+1)}) \\ \underset{+}{S_{P_1}(p_{(k+1)})} & \underset{+}{S_{P_2}(p_{(k+1)})} & \cdots & \underset{+}{S_{P_m}(p_{(k+1)})} \\ \vdots & \vdots & & \vdots \\ S_{P_1}(p_{(k+1)}) & S_{P_2}(p_{(k+1)}) & \cdots & S_{P_m}(p_{(k+1)}) \end{vmatrix}$$

$c = m - (k + 1)!$ and $S_{P_i}(p_{(i)}) = 1 - F_{P_i}(p_{(i)})$

Permanent is like determinant with all signs positive (Minc, 1978)

$m \times m$ with $k + 1$ rows in top } and $m - (k + 1)$ rows in bottom }

Bounds and P-values



Inequality Representation

$$p_{(1)} \leq p_{(2)} \leq b_1 \leq b_2 \leq p_{(3)} \leq b_3; p_{(4)} \geq b_4$$

$$p_{(1)} \leq p_{(2)} \leq b_1 \leq p_{(3)} \leq b_2 \leq b_3; p_{(4)} \geq b_4$$

$$p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq b_1 \leq b_2 \leq b_3; p_{(4)} \geq b_4$$

$$p_{(1)} \leq b_1 \leq p_{(2)} \leq b_2 \leq p_{(3)} \leq b_3; p_{(4)} \geq b_4$$

$$p_{(1)} \leq b_1 \leq p_{(2)} \leq p_{(3)} \leq b_2 \leq b_3; p_{(4)} \geq b_4$$

Ordered List Representation

$$\mathcal{L}_{3,1} = \{0, p_{(1)}, p_{(2)}, b_1, b_2, p_{(3)}, b_3, b_4, p_{(4)}, 1\}$$

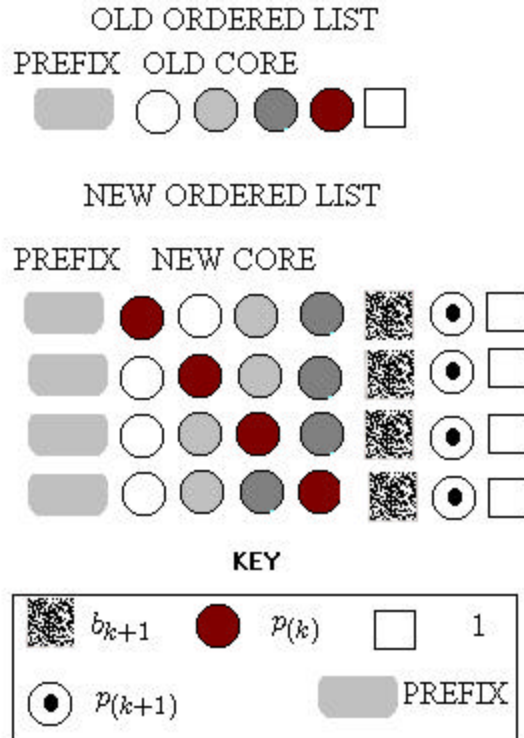
$$\mathcal{L}_{3,2} = \{0, p_{(1)}, p_{(2)}, b_1, p_{(3)}, b_2, b_3, b_4, p_{(4)}, 1\}$$

$$\mathcal{L}_{3,3} = \{0, p_{(1)}, p_{(2)}, p_{(3)}, b_1, b_2, b_3, b_4, p_{(4)}, 1\}$$

$$\mathcal{L}_{3,4} = \{0, p_{(1)}, b_1, p_{(2)}, b_2, p_{(3)}, b_3, b_4, p_{(4)}, 1\}$$

$$\mathcal{L}_{3,5} = \{0, p_{(1)}, b_1, p_{(2)}, p_{(3)}, b_2, b_3, b_4, p_{(4)}, 1\}$$

Construction Algorithm for Ordered Lists



First Three Sets of Ordered Lists

$$\mathbf{k} = 1$$

$$\mathcal{L}_{1,1} = \{0, p_{(1)}, b_1, b_2, p_{(2)}, 1\}$$

$$\mathbf{k} = 2$$

$$\mathcal{L}_{2,1} = \{0, p_{(1)}, b_1, p_{(2)}, b_2, b_3, p_{(3)}, 1\}$$

$$\mathcal{L}_{2,2} = \{0, p_{(1)}, p_{(2)}, b_1, b_2, b_3, p_{(3)}, 1\}$$

$$\mathbf{k} = 3$$

$$\mathcal{L}_{3,1} = \{0, p_{(1)}, p_{(2)}, b_1, b_2, p_{(3)}, b_3, b_4, p_{(4)}, 1\}$$

$$\mathcal{L}_{3,2} = \{0, p_{(1)}, p_{(2)}, b_1, p_{(3)}, b_2, b_3, b_4, p_{(4)}, 1\}$$

$$\mathcal{L}_{3,3} = \{0, p_{(1)}, p_{(2)}, p_{(3)}, b_1, b_2, b_3, b_4, p_{(4)}, 1\}$$

$$\mathcal{L}_{3,4} = \{0, p_{(1)}, b_1, p_{(2)}, b_2, p_{(3)}, b_3, b_4, p_{(4)}, 1\}$$

$$\mathcal{L}_{3,5} = \{0, p_{(1)}, b_1, p_{(2)}, p_{(3)}, b_2, b_3, b_4, p_{(4)}, 1\}$$

Ordered List to Ordered Set of Ordered Pairs

$$\mathcal{L}_{3,3} = \{0, p_{(1)}, p_{(2)}, p_{(3)}, b_1, b_2, b_3, b_4, p_{(4)}, 1\}$$

.

$$\{0, p_{(1)}, p_{(2)}\} \Rightarrow (0, p_{(2)})$$

$$\{0, p_{(2)}, p_{(3)}\} \Rightarrow (0, p_{(3)})$$

$$\{0, p_{(3)}, b_1\} \Rightarrow (0, b_1)$$

$$\{b_1, b_2\} \Rightarrow \{b_1, b_2, b_3\} \Rightarrow \{b_2, b_3, b_4\} \Rightarrow \{b_3, b_4, p_{(4)}\}$$

$$\Rightarrow \{b_4, p_{(4)}, 1\} \Rightarrow (b_4, 1)$$

$$\mathcal{B}_{3,3} = \left\{ (0, p_{(2)}), (0, p_{(3)}), (0, b_1), (b_4, 1) \right\}$$

Integral Representation

$$\int_{b_2}^{b_3} \int_0^{b_1} \int_0^{p(2)} f_{P(1),P(2),P(3),P(4)} \left(p(1), p(2), p(3), p(4) \right) \prod_{i=1}^4 d\chi$$

$$\int_{b_1}^{b_2} \int_0^{b_1} \int_0^{p(2)} f_{P(1),P(2),P(3),P(4)} \left(p(1), p(2), p(3), p(4) \right) \prod_{i=1}^4 d\chi$$

$$\int_0^{b_1} \int_0^{p(3)} \int_0^{p(2)} f_{P(1),P(2),P(3),P(4)} \left(p(1), p(2), p(3), p(4) \right) \prod_{i=1}^4 d\chi$$

$$\int_{b_2}^{b_3} \int_{b_1}^{b_2} \int_0^{b_1} f_{P(1),P(2),P(3),P(4)} \left(p(1), p(2), p(3), p(4) \right) \prod_{i=1}^4 dp_i$$

$$\int_{b_1}^{b_2} \int_{b_1}^{p(3)} \int_0^{b_1} f_{P(1),P(2),P(3),P(4)} \left(p(1), p(2), p(3), p(4) \right) \prod_{i=1}^4 d\chi$$

General Form of Region of Integration

$$\mathcal{B}_{k,p} = \begin{cases} \{(b_1, 1)\} & k = \\ \{(0, u_{k,p,1}), \dots, (b_{k+1}, 1)\} & k \neq \\ \{(0, u_{m,p,1}), \dots, (l_{m,p,k}, u_{m,p,m})\} & k = \end{cases}$$

$$\mathcal{B}_k = \{\mathcal{B}_{k,1}, \mathcal{B}_{k,2}, \dots, \mathcal{B}_{k,c_k}\} .$$

**LISP program generates integration regions.
Catalan number of them:**

$$c_k = \frac{(2k)!}{k!k!(k+1)}$$

Probability of k rejections is

$$\Pr\{K = k\} =$$

$$\left\{ \begin{array}{l} \prod_{i=1}^m [1 - F_{P_i}(b_1)] \\ \sum_{p=1}^{c_k} \int_{\mathcal{B}_{k,p}} f_{P_{(1)}, \dots, P_{(k+1)}}(p_{(1)}, \dots, p_{(k+1)}) \prod_{i=1}^{k+1} dp_{(i)} \\ \sum_{p=1}^{c_m} \int_{\mathcal{B}_{m,p}} m! \prod_{i=1}^m f_{P_i}(p_{(i)}) \prod_{i=1}^m dp_{(i)} \end{array} \right.$$

$$k = 0$$

$$k \neq 0, m$$

$$k = m$$

Joint Probability of Total Rejections and False Rejections

- Assume without loss of generality that lower # p-values correspond to null hypotheses that are true.
- Joint density function is sum of products of $k + 1$ terms
- In each product Count number of terms of form $f_{P_d}(p_{(e)})$, with $d \in \{1, \dots, n\}$ and $e \in \{1, \dots, k\}$
- Let \mathcal{C}_j be the set of terms that have j factors that fulfill the condition.

Joint Probability of Total Rejections and False Rejections

$$\Pr\{J = j, K = k\} = \begin{cases} \prod_{i=1}^m [1 - F_{P_i}(b_1)] \\ \sum_{p=1}^{c_k} \int_{\mathcal{B}_{k,p}} \sum_{v \in \mathcal{C}_j} a_v dp_{(1)} \dots dp_{(k+1)} \\ m! \sum_{p=1}^{c_m} \int_{\mathcal{B}_{k,m}} \sum_{v \in \mathcal{C}_j} a_v dp_{(1)} \dots dp_{(m)} \end{cases}$$

$$k = 0, j = 0$$

$$1 \leq k \leq m - 1$$

$$k = m .$$

Exact Calculations

Law of total probability

- positive False Discovery Rate
- False Discovery Rate
- per comparison Error rate
- Expected Sensitivity
- Expected Specificity
- Expected Positive Predictive Value
- Expected Negative Predictive Value

Observed and Predicted FDR

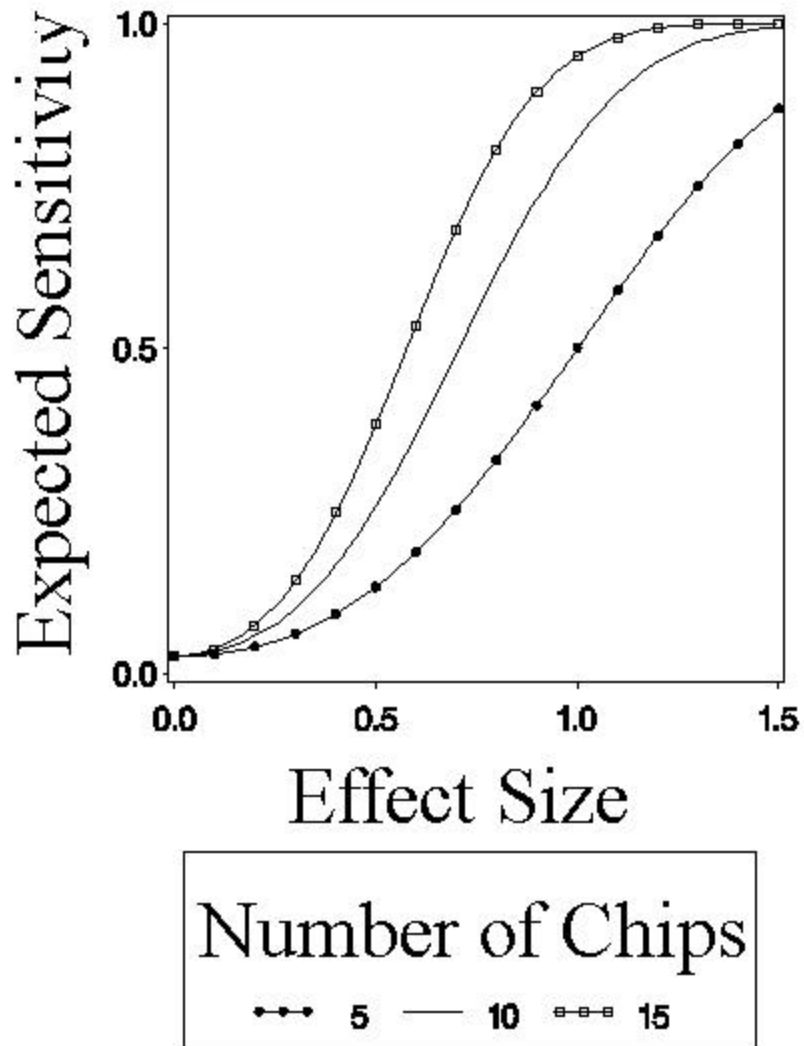
$$T_1 \sim \mathcal{N}(0, 1) \text{ and } T_2 \sim \mathcal{N}(\delta, 1)$$

δ	Rejections		Probability	
	False j	Total k	Theory	Observed (SE $\leq .0016$)
0	0	0	.950	.951
	1	1	.047	.047
	2	2	.002	.002
1.5	0	0	.751	.752
	0	1	.218	.218
	1	1	.017	.017
	1	2	.014	.014

$$\delta = 0 \Rightarrow \text{FDR} = \mathcal{E}\left(\frac{J}{K}\right) \approx \frac{1}{1} \cdot 0.47 + \frac{2}{2} \cdot .002 = .05$$

$$\delta = 1.5 \Rightarrow \text{FDR} = \mathcal{E}\left(\frac{J}{K}\right) \approx \frac{1}{1} \cdot 0.17 + \frac{1}{2} \cdot .014 = .0$$

Power Analog



Research in Progress

Computational Issues

Correlated Data

Approximations

REFERENCES

- Aitken, A. C. (1939), *Determinants and Matrices*. Edinburgh: Oliver and Boyd.
- Benjamini, Y, Drai, D., Elmer, G., Kafkafi, N. and Golani, I. (2001), "Controlling the False Discovery Rate in Behavior Genetics Research", *Behavioural Brain Research*, **125**, 279-284.
- Benjamini, Y., and Hochberg, Y. (1995) "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B*, **57**, 289-300.
- Benjamini, Y. and Liu, W. (1999) "A Step-Down Multiple Hypotheses Testing Procedure That Controls The False Discovery Rate Under Independence," *Journal of Statistical Planning and Inference*, **82**, 163-170.
- Brown, P. O. and Botstein, D. (1999), "Exploring The New World Of The Genome With DNA Microarrays". *Nature Genetics*, **21**, 33-37.
- Clement, K., Viguerie, N., Diehn, M., Alizadeh, A., Barbe, P., Thalamas, C., Storey, J.D., Brown, P.O., Barsh, G.S., Langin, D. (2002), "In Vivo Regulation Of Human Skeletal Muscle Gene Expression By Thyroid Hormone," *Genome Research*, 12(2), 281-91.
- Curran-Everett, D. (2000), "Multiple Comparisons: Philosophies And Illustrations," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, **279**, R1-R8.
- Efron, B., Storey, J. and Tibshirani, R. (2001), "Microarrays, Empirical Bayes Methods, and False Discovery Rates," *Journal of the American Statistical Association*, **96**, 1151-1160.
- Genovese, C. R., and Wasserman, L. (2001), "Operating Characteristics and Extensions of the False Discovery Rate Procedure," *Journal of the Royal Statistical Society: Series B*, **64**, 499-518.
- Hochberg, Y., and Tamhane, A. (1987), *Multiple Comparison Procedures*. New York:Wiley.
- Ji, X., Cheung, R., Cooper, S., Li, Q., Greenberg H.B. and He X.S. (2003), "Interferon Alpha Regulated Gene Expression In Patients Initiating Interferon Treatment For Chronic Hepatitis C". *Hepatology*, **37**(3), 610-621.
- Keselman, H.J., Cribbie, R., and Holland, B. (2002), "Controlling The Rate Of Type 1 Error Over A Large Set Of Statistical Tests," *British Journal of Mathematical and Statistical Psychology*, **55**, 27-39.
- Kohane, I. S., Kho, A., and Butte, A. J. (2002), *Microarrays for an Integrative Genomics*. Cambridge: MIT Press.
- Leithold, L. (1968), *The Calculus with Analytic Geometry*. New York: Harper and Row, Publishers.
- Lindgren, B. W. (1976), *Statistical Theory: Third Edition*. New York:Macmillan Publishing.
- Lee, M. and Whitmore, G. (2002), "Power and Sample Size for DNA Microarray Studies," *Statistics in Medicine*, **21**, 3543-3570.
- Miller, R. (1977), "Developments in multiple comparison procedures 1966-76," *Journal of the American Statistical Association*, **72**, 779-788.
- Minc, H. (1978), *Permanents*. Reading, MA: Addison-Wesley.
- Shenkar R., Elliott J. P., Diener K., Gault J., Hu L. J. , Cohrs R. J., Phang T., Hunter L., Breeze R. E., and Awad I. A. (2003), "Differential Gene Expression In Human Cerebrovascular Malformations," *Neurosurgery*, **52**(2), 465-477.
- Storey, J. (2002), "A Direct Approach to the False Discovery Rate," *Journal of the Royal Statistical Society: Series B*, **64**, 479-598.

- Tusher, V.G., Tibshirani R., and Chu, G. (2001), "Significance Analysis Of Microarrays Applied To The Ionizing Radiation Response". *Proceedings of the National Academy of Sciences of the United States of America*, **98**(9), 5116-5121.
- Vaughan, R. J. and Venables, W. N. (1972), "Comments and Queries: Permanent Expressions for Order Statistic Densities," *Journal of the Royal Statistical Society: Series B*, **34**, 308-310.
- Weisstein, E. W. (1999), *CRC Concise Encyclopedia of Mathematics*, CRC Press: Boca Raton, Fla.
- Williams, V., Jones, L., and Tukey, J. (1999), "Controlling Error in Multiple Comparisons, with Examples from State to State Differences in Educational Achievement," *Journal of Educational and Behavioral Statistics*, **24**, 42-69.
- Yekutieli, D. and Benjamini, Y. (1999), "Resampling Based False Discovery Rate Controlling Procedure for Dependent Test Statistics," *Journal of Statistical Planning and Inference*, **82** 171-196.
- Xiao Y., Segal M. R., Rabert D. , Ahn A. H., Anand P., Sangameswaran L., Hu D., and Hunt C. A. (2002), "Assessment Of Differential Gene Expression In Human Peripheral Nerve Injury," *BMC Genomics* **3**(1):28.
- Zien, A., Fluck, J., Zimmer, R., and Lengauer, T. (2002), "Microarrays: How Many Do You Need?," *RECOMB 2002: Proceedings of the Sixth Annual International Conference on Computational Biology* (ACM Press), 321-330.