

# **ZIPF'S LAW IN IMPORTANCE OF GENES FOR CANCER CLASSIFICATION USING MICROARRAY DATA**

Second Conference of Chinese Bioinformatics Society (June 28-30, 2002)

*Wentian Li, Ph.D*

Center for Genomics and Human Genetics  
North Shore LIJ Research Institute  
Manhasset, NY 11030, USA

web: <http://linkage.rockefeller.edu/wli/>  
email: [wli@linkage.rockefeller.edu](mailto:wli@linkage.rockefeller.edu), [wli@nshs.edu](mailto:wli@nshs.edu)

**One of the main tasks in microarray analysis is to select genes that are differentially expressed in diseased tissues vs. normal tissues; one subtype of disease vs. another; or any two clinical/phenotypical conditions whose differences are of interests.**

**If we examine genes one at the time (rather than clusters, or conditionally over another gene), they can be ranked according to how differentially expressed they are. “Top” gene is the most differentially expressed gene.**

Do we pick top 10 genes? top 100? If we rank genes by some quantitative measurement, how this measurement decreases with the rank? What is the functional form of this decrease?

Zipf's law

Discriminant  
Analysis

Likelihood

City Population

Keyword Search

Power-law

Gene Selection

# OUTLINE

- + Discriminant analysis/Logistic regression (binomial, multinomial, matched)
- + Maximum likelihood as measure of the importance of genes
- + Zipf's law/Luhn's keyword design principle
- + Zipf's law in cancer classifications (leukemia, colon, lymphoma, breast, bladder)
- + Discussions

# Discriminant Analysis

**Tests:** assuming a gene doesn't discriminate two types, test how incorrect the assumption is. [Based on normal distribution:] t-test; [Do not assume normal distribution:] other non-parametric tests.

**Modeling:** assume a gene contributes to the discrimination via some model. there are parameters in the model. adjusting the parameter values to reach the best fit

## Discriminant Analysis → Modeling → Logistic Regression

sample  $i$  (1,2,..N), gene  $j$  (1,2,..p)

sample label  $y$ , gene expression (usually on log level)  $x$

$$Prob(y_i = 1 | x_{ji}) = \frac{1}{1 + e^{-a_j - b_j x_{ji}}}$$

## Discriminant Analysis → Modeling → Logistic Regression

### → Maximum likelihood

multiply over all samples ( $\prod_i$ ).

if the sample is actually labeled as 1, use  $\max P(y = 1)$ , if the sample is

actually labeled as 0, use  $\max P(y = 0)$

$$\hat{L}_j = \max \prod_i \text{Prob}(y_i = 1 | x_{ji})^{y_i} (1 - \text{Prob}(y_i = 1 | x_{ji}))^{1-y_i}$$

$$= \prod_i \left( \frac{1}{1 + e^{-\hat{a}_j - \hat{b}_j x_{ji}}} \right)^{y_i} \left( 1 - \frac{1}{1 + e^{-\hat{a}_j - \hat{b}_j x_{ji}}} \right)^{1-y_i}$$

**Discriminant Analysis → Modeling → Logistic Regression  
→ Maximum likelihood → Measure of importance of a gene**

$\hat{L}_j$  represents the overall discrimination/classification performance of gene  $j$  on the data at hand.

To get a per sample performance, remember that  $\hat{L}_j$  is derived from a product, we use

$$\bar{p}_j = \hat{L}_j^{1/N}$$

or ( $H$  is the entropy):

$$\bar{p}_j = e^{\frac{1}{N} \sum_i \log(P(y_i))} \approx e^{-H_j}$$

**Discriminant Analysis → Modeling → Logistic Regression  
→ Multinomial Logistic Regression**

sample  $i$  (1,2,...N), gene  $j$  (1,2,...p)

sample label  $y$  (whose value can be 1,2..l), gene expression  
(on log level)  $x$

$$Prob(y_i = I | x_{ji}) = \frac{e^{-a_I - b_I x_{ji}}}{\sum_{K=1}^C e^{-a_K - b_K x_{ji}}}$$

**Discriminant Analysis → Modeling → Logistic Regression  
→ Matched Case-Control Logistic Regression**

sample  $i$  ( $1, 2, \dots, n=N/2$ ), gene  $j$  ( $1, 2, \dots, p$ )

the direction of condition is switched

$$Prob(x_{ji} - x'_{ji} | y_i = 1, y'_i = 0) = \frac{1}{1 + e^{b_j(x_{ji} - x'_{ji})}}$$

# Zipf's Law

George Kingsley Zipf (1902-1950), linguistic professor at Harvard, observed that frequency of usage of English words  $f_{(r)}$  ( $r$  is the rank), follows a power-law function

$$f_{(r)} = \frac{c}{r^\alpha}, \quad \alpha \approx 1$$

## Zipf's law → city population

$p(r)$  is the population of rank- $r$  city:

$$p(r) = \frac{c}{r^\alpha} \quad \alpha \text{ may not be } 1$$

Similarly, **company sizes in terms of revenue, in terms of profit,...**

## **Zipf's law → webpage statistics**

The number of access of the rank- $r$  webpage...

Similarly, **keyword search in database, number of times library books are borrowed, citation of scientific publications...**

## **Zipf's law → information retrieval → Luhn's principle for keywords selection**

Hans Peter Luhn (1896-1964), IBM, pioneer of information retrieval systems, recognized that both high-ranking words (common) and low-ranking words (rare) are not good candidates for keywords used for searching. Middle-ranked words are the best candidates.

# Zipf's Law in Microarray Data

*public domain data:*

cancer	types	source	model	N	total genes
leukemia	ALL,AML	<i>Golub'99</i>	LR	72	7129
colon	colon, norm	<i>Alon'00</i>	LR	62	2000
lymphoma	DLBCL,FL,CLL,norm	<i>Alizadeh'00</i>	mult.LR	96	4026
lymphoma	GC- ,A-DLBCL,norm	<i>Alizadeh'00</i>	mult.LR	72	4026
breast	before,after (chemo.)	<i>Perou'00</i>	matched LR	20p	8102

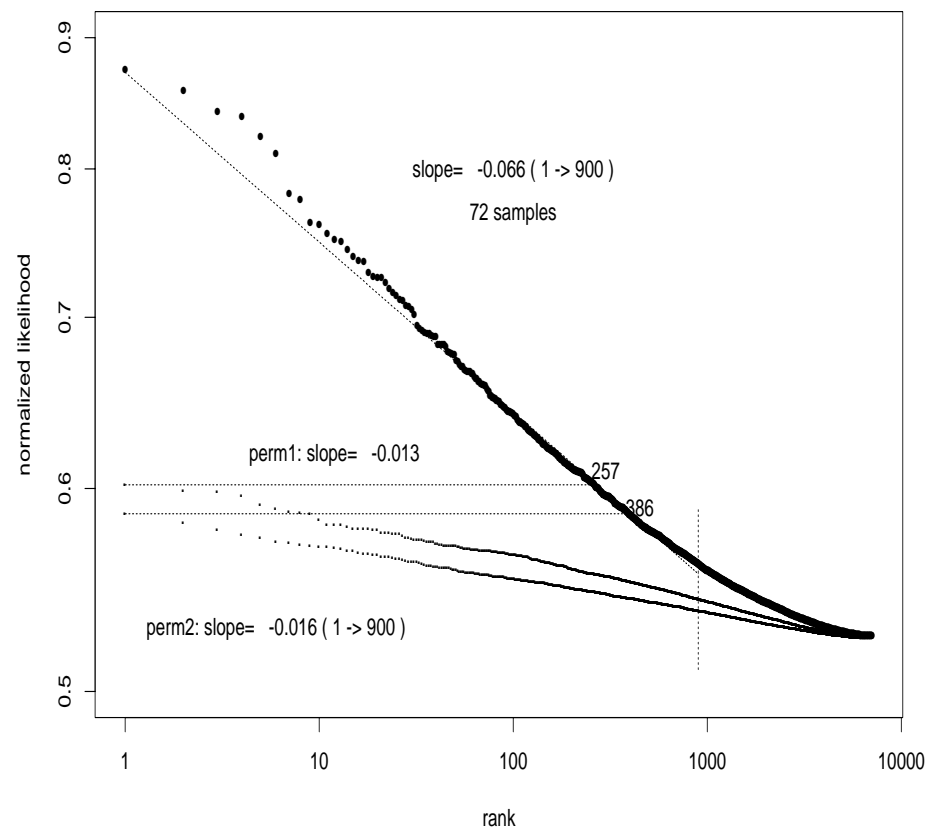
*unpublished data:*

bladder cancer, less than 20 samples, more than 15000 genes/ESTs

rheumatoid arthritis, 42 sampels, 8344 genes/ESTs are used

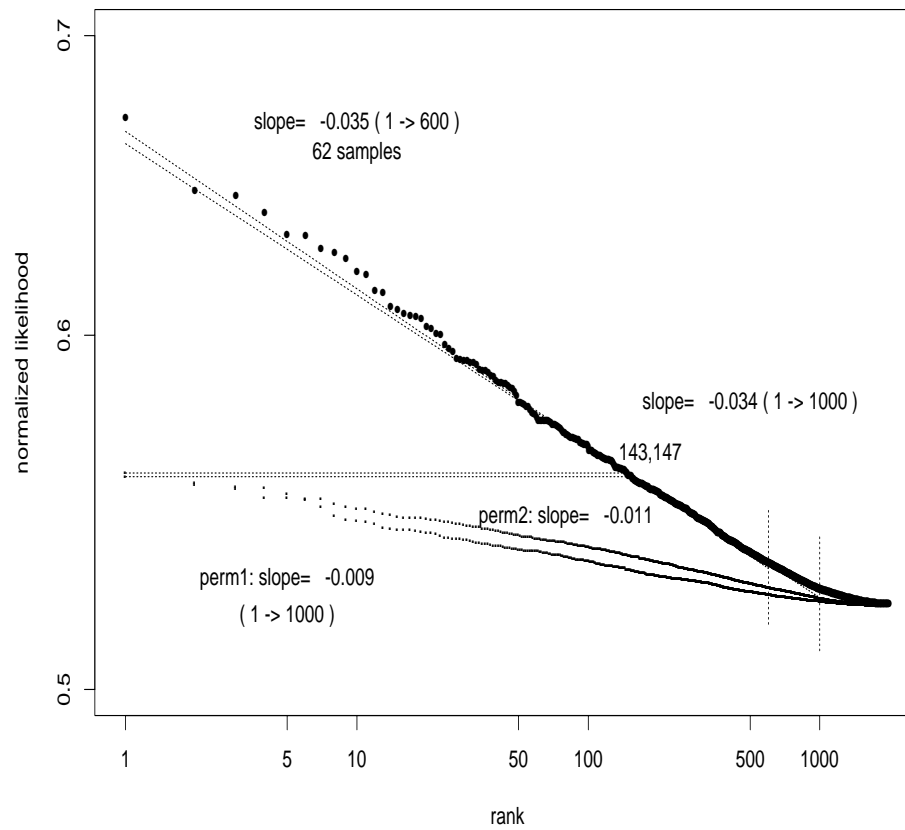
# cancer classification → two classes → two leukemia subtypes

classifying two leukemia subtypes



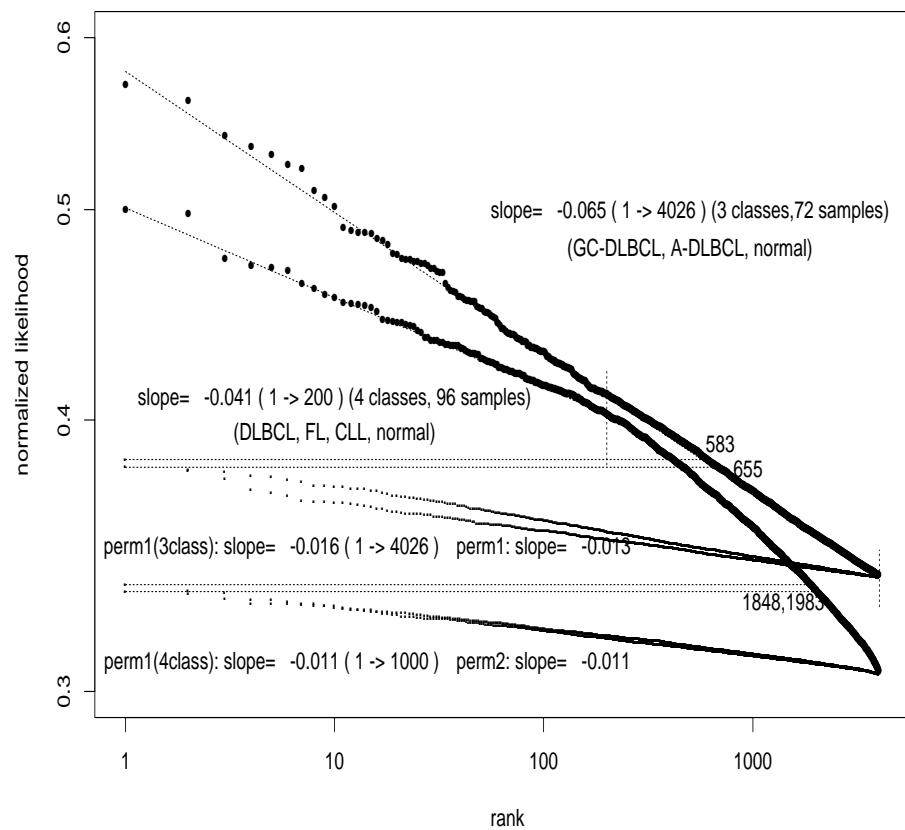
# cancer classification → two classes → colon cancer vs. normal

classifying colon cancer vs normal



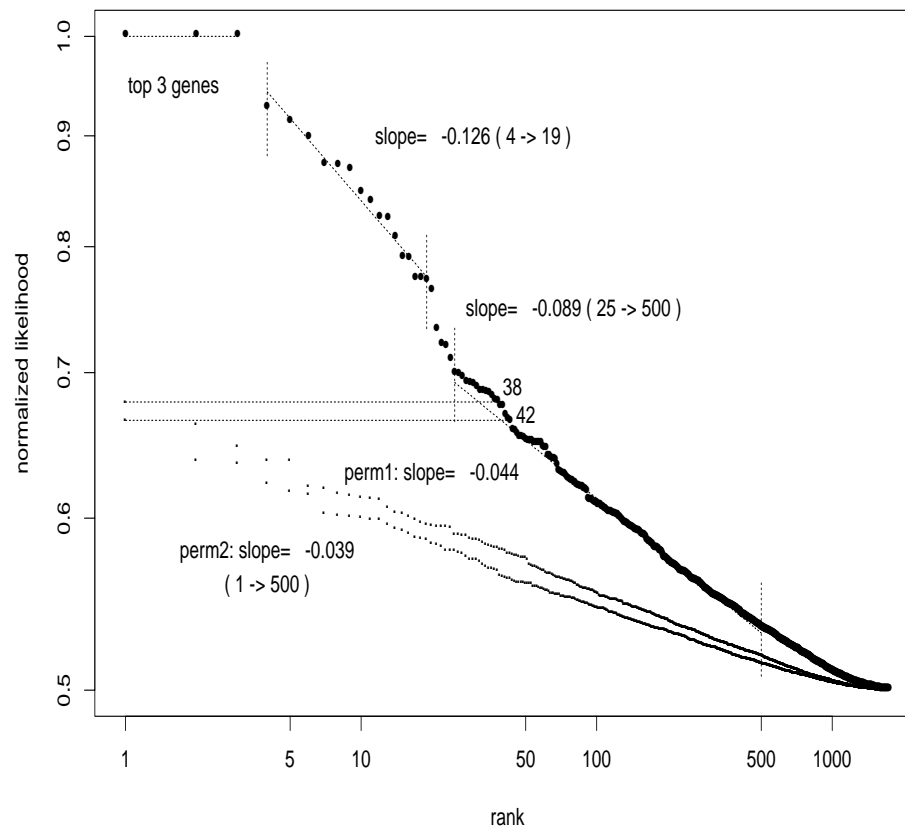
# cancer classification → multiple classes → lymphoma subtypes

classifying lymphoma subtypes and normal



# cancer classification → matched case control → breast cancer before vs after chemotherapy treatment

classifying conditions before and after treatment of breast cancer



## A summary of these ranked plots

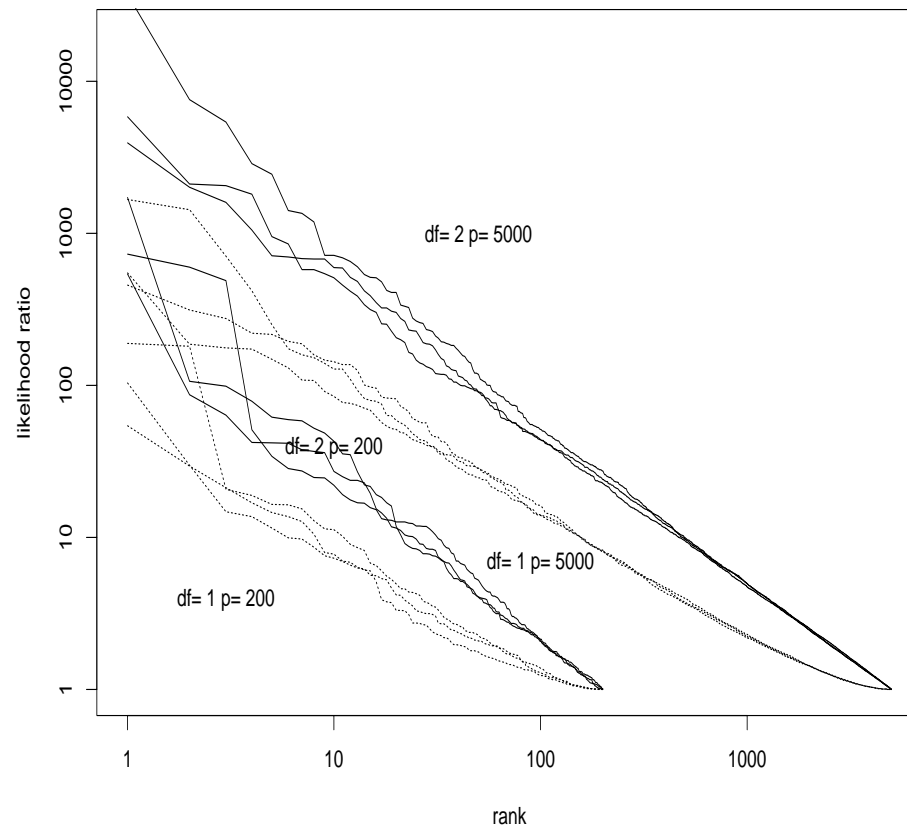
cancer (size)	best likelihood (ratio over baseline)	baseline likelihoods (G0, G1, P, P)	$\alpha$ ( $\alpha/\alpha_{perm}$ )	num. gene (total)
leukemia (72)	0.87 (1.7, 1.7, 1.5, 1.5)	0.5, 0.52, 0.59, 0.60	0.066 ( $\sim$ 4-5)	-, -, 386,257 7129
colon (62)	0.67 (1.3, 1.3, 1.2, 1.2)	0.5, 0.52, 0.56, 0.56	0.034 ( $\sim$ 3-4)	-, -, 143,147 2000
lym(4) (96)	0.50 (2.0, 1.6, 1.5, 1.5)	0.25, 0.31, 0.34, 0.33	0.04? ( $\sim$ 3-4 ?)	-, -, 1848,1983 4026
lym(3) (72)	0.57 (1.7, 1.7, 1.5, 1.5)	0.33, 0.34, 0.38, 0.38	0.065 ( $\sim$ 4-5)	-, -, 583,655 4026
breast (20)	1.00 (2.0, 2.0, 1.5, 1.5)	0.5, 0.5, 0.68, 0.67	0.13? 0.09? ( $\sim$ 2-3?)	-, -, 38,42 8102

**What about random data? Do they also follow Zipf's law?**

Under null hypothesis, (2, log, max) likelihood ratio  $\hat{L}_1/\hat{L}_0$  follows a  $\chi^2$  distribution with  $df$  degrees of freedom ( $df = p_1 - p_0$ , where  $p_1$  and  $p_0$  are the number of parameters in model  $L_1$  and null  $L_0$ ).

YES, they do!

likelihood ratios generated by  $\chi^2(df=1, 2)$



# Unpublished data (1): bladder cancer

+ 19 samples: 4 “normal” (non-bladder-cancer), 7 superficial, 3 invasive (though one of the samples is suspected to be superficial), 5 metastatic

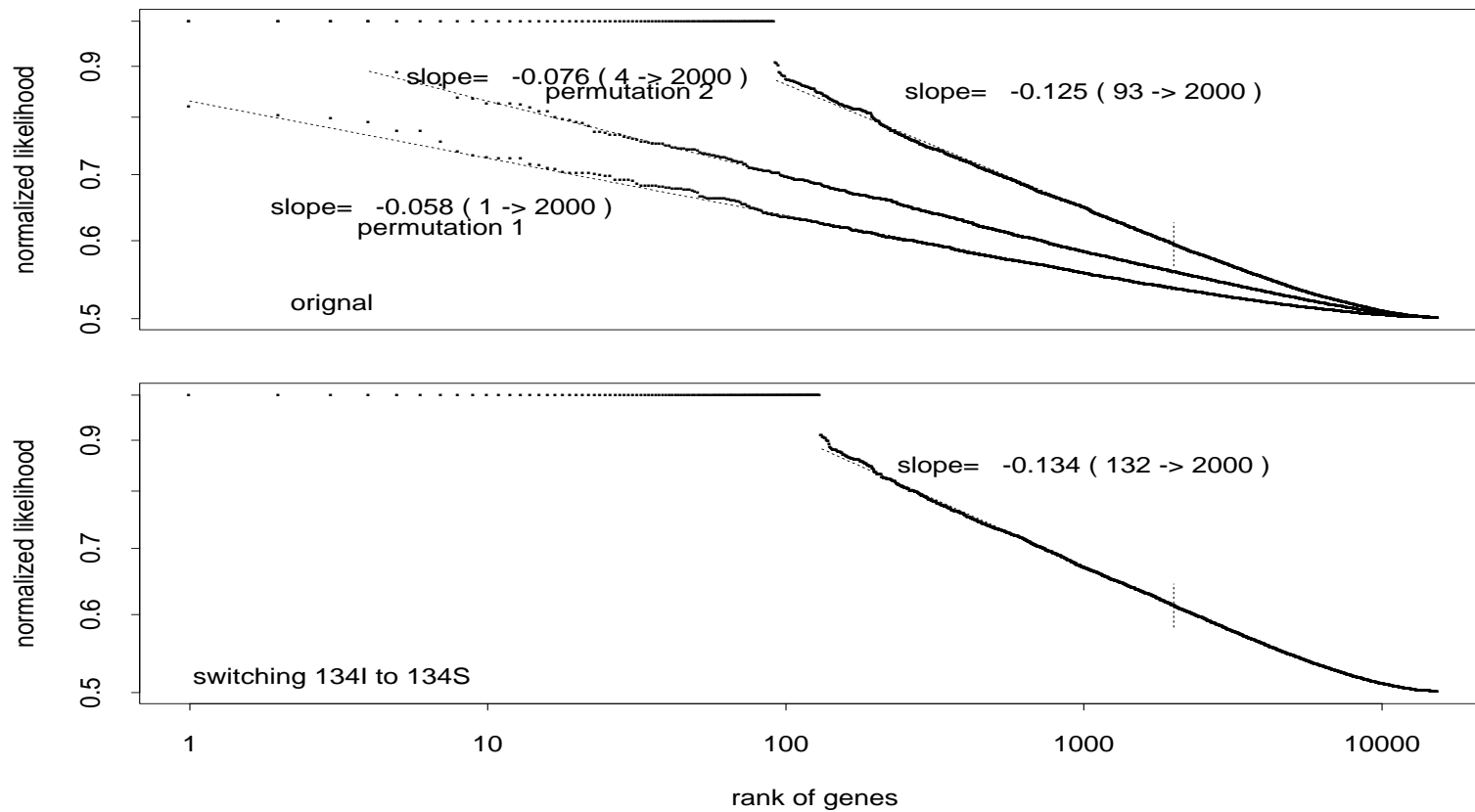
+ Number of genes/ESTs (spots) is 15,650

+ cDNA chip, plot  $(\log I_{red} - \log I_{green})$  vs.  $(\log I_{red} + \log I_{green})/2$ . regression line provides the mean to be subtracted.  $(\log I_{red} - \log I_{green} - mean)$  is the log-expression used.

+ “Normal” samples are not really normal (patients with other cancers), more interested in superficial vs. (invasive+metastatic), and invasive vs. metastatic

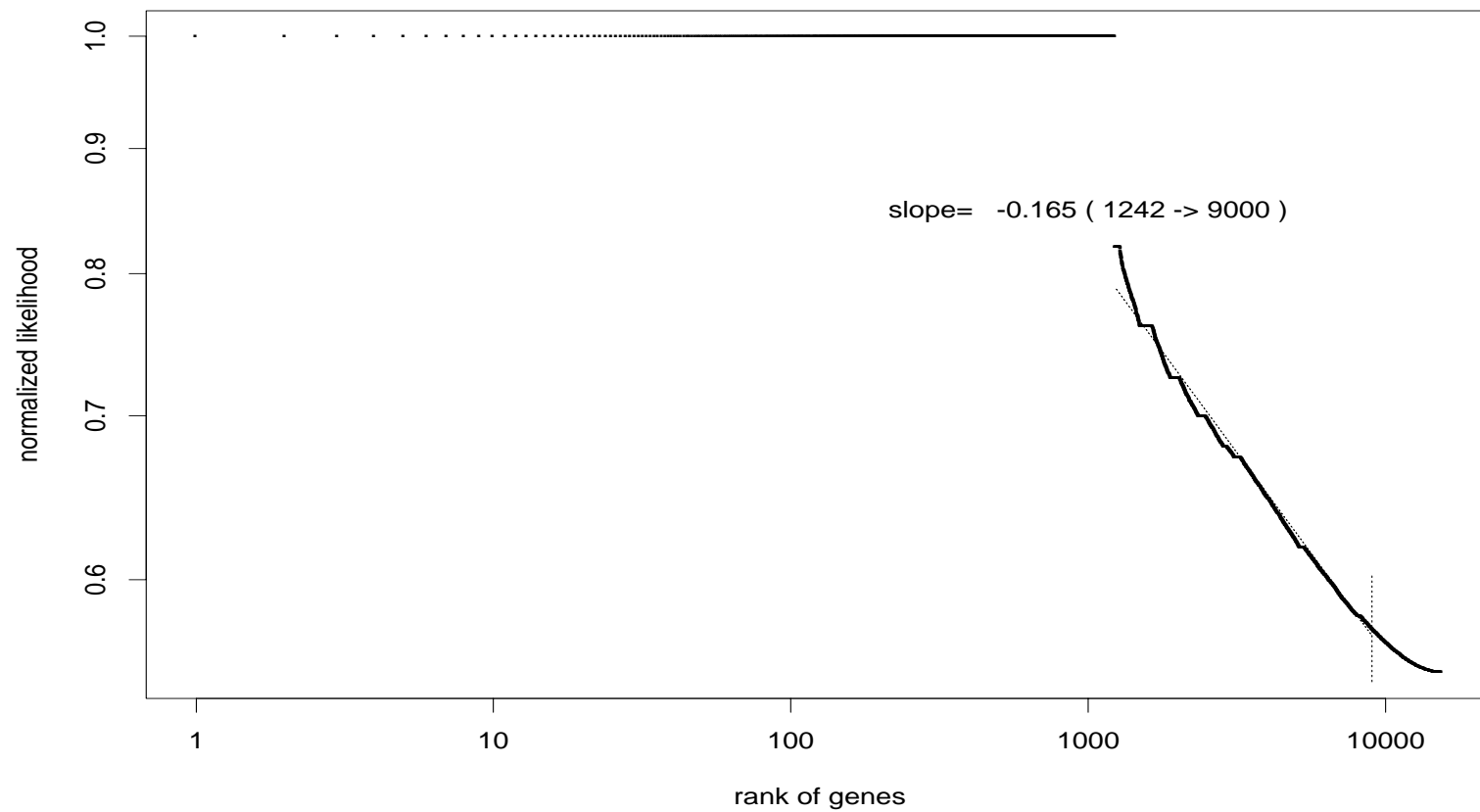
# cancer classification → two classes → bladder cancer → S vs. I+M

## S vs. I+M classification (N=15)



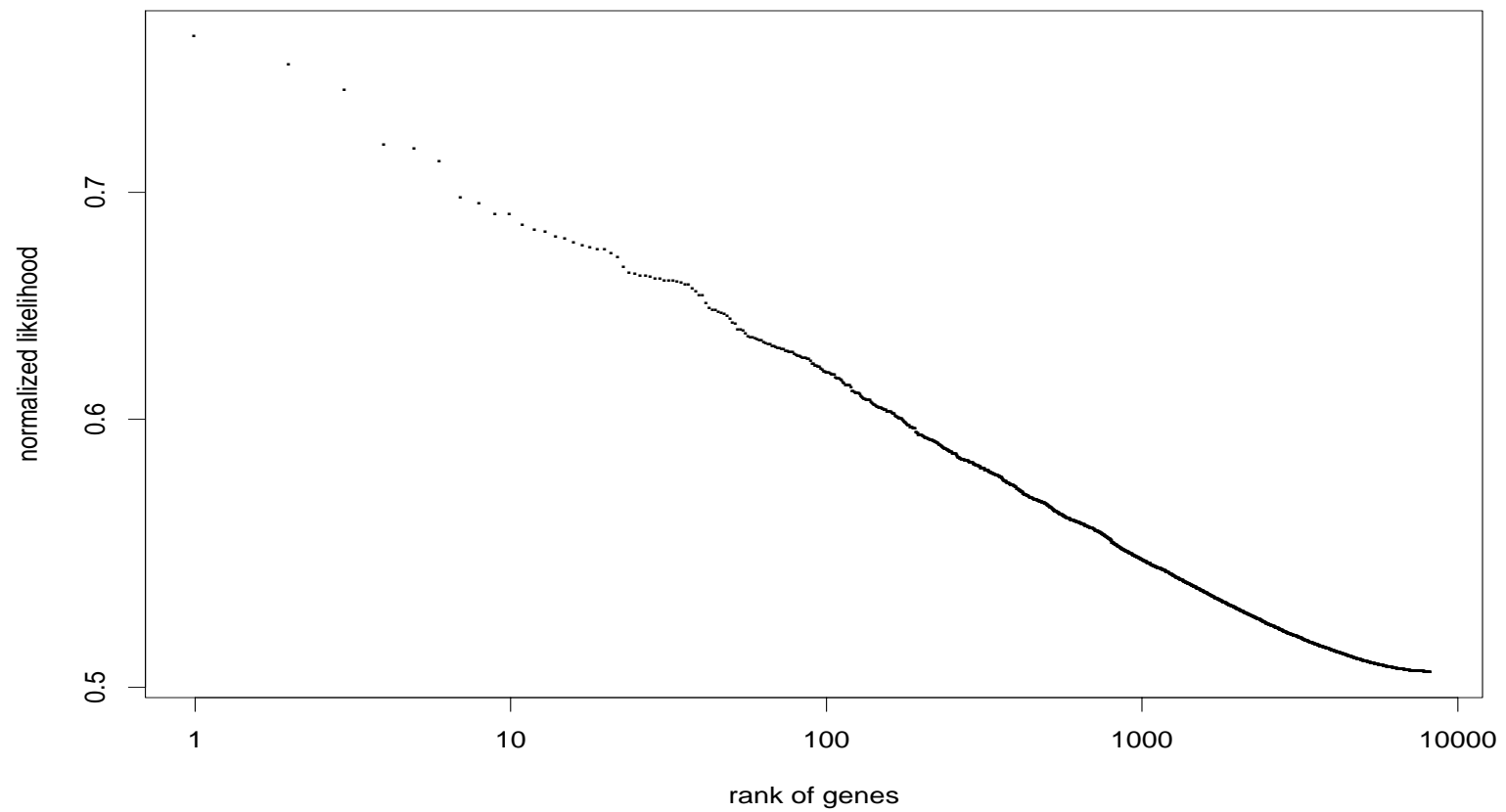
**cancer classification → two classes → bladder cancer → I vs. M**

I vs. M classification (N=7, 2 I's, 5 M's)



## Unpublished data (2): rheumatoid arthritis

classification performance of RA vs controls (42 samples)



## **Discussions → No intrinsic cutoff**

Just as there is no intrinsic cutoff threshold between big cities and median sized cities and small cities, there is no “gap” in the ranked likelihood profile that separates “important” genes and “unimportant” genes. Grey areas always exist.

## Discussions → Shape of the ranked likelihood

Power-law function ( $\log(\hat{L}_{(r)})$  vs.  $\log r$ ) seems to be universally true (if sample size is large enough). Even true for permuted data. The fall-off of the ranked likelihood is perhaps more informative. In contrast to examples of English words and city populations, the slope is not close to 1.

Ranked entropy  $H_{(r)}$  increases with  $r$  as  $\alpha \log(r)$

Ranked p-values and connection with multiple testing?

## **Discussions → Avoid top and bottom genes?**

Following Luhn's principle, we may want to throw away high and low ranking genes. Bottom genes, yes, but top genes? Only in the sense that if these top genes are well known, genes involved in a joint action are more likely in the middle-ranked range. [Logistic regression with two or more genes, variable selection, model selection (AIC,BIC),...]

# *Collaborators*

- *Yaning Yang* (Rockefeller)
- *Marta Sanchez-Carbayo* (Sloan-Kettering)  
bladder cancer data
- *Franak Batliwalla* (North Shore)  
rheumatoid arthritis data

# *References*

W Li, Y Yang (2002), “**How many genes are needed for a discriminant microarray data analysis?**”, in *Methods of Microarray Data Analysis*, pp.137-150 (Kluwer Academic)

W Li, Y Yang, “**Zipf’s plot of importance of genes for cancer classification**”, submitted to *Journal of Theoretical Biology*.

M Sanchez-Carbayo, N Socci, JJ Lozano, W Li, E Charytonowicz, T Belbin, M Preystowski, A Ortiz, G Childs, C Cordon-Cardo, “**Gene expression profiling of bladder cancer tissues using cDNA microarrays segregates superficial and invasive disease**”, submitted.