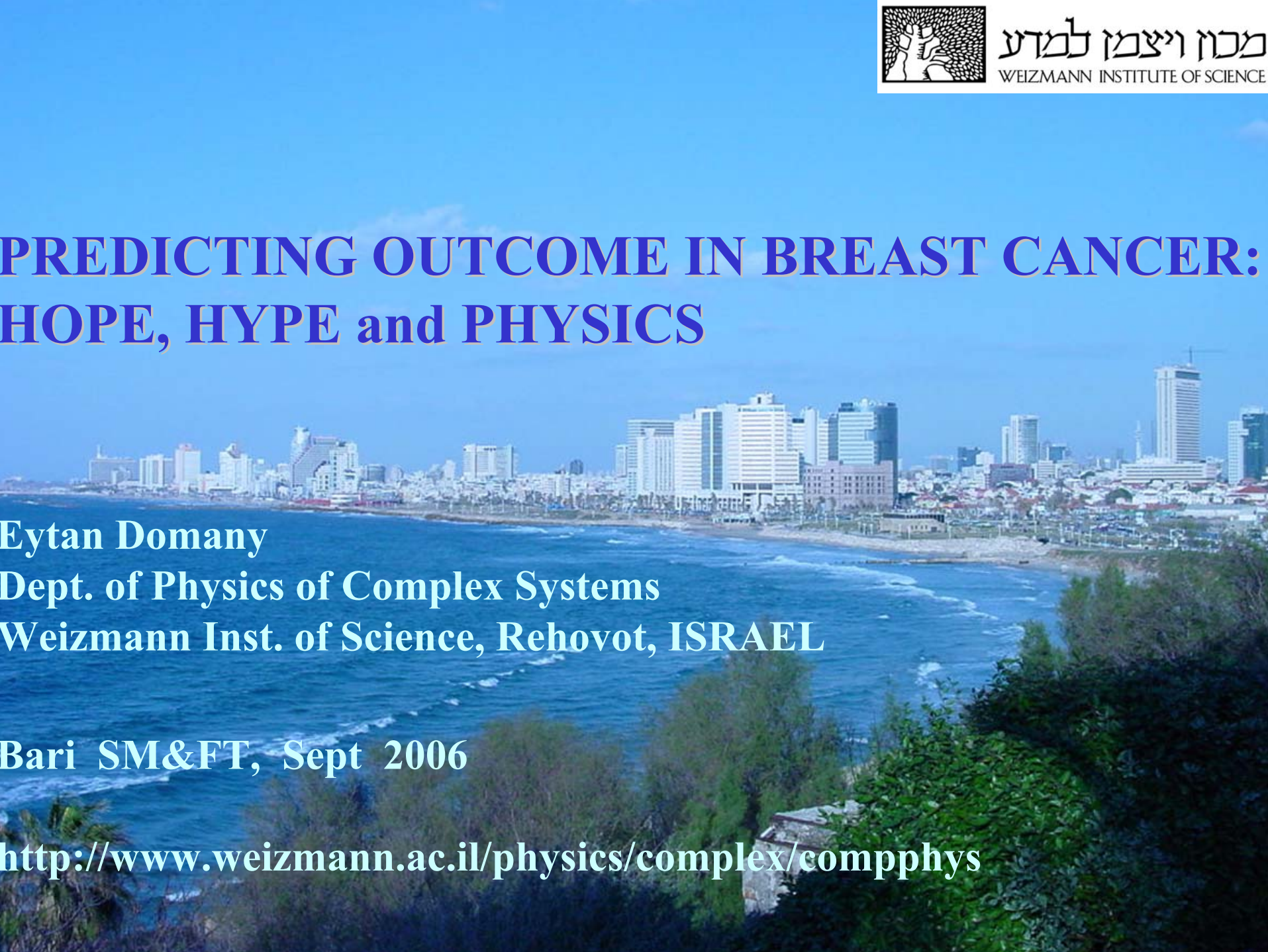


מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

PREDICTING OUTCOME IN BREAST CANCER: HOPE, HYPE and PHYSICS

Eytan Domany
Dept. of Physics of Complex Systems
Weizmann Inst. of Science, Rehovot, ISRAEL

Bari SM&FT, Sept 2006

http://www.weizmann.ac.il/physics/complex/compphys

OUTLINE:

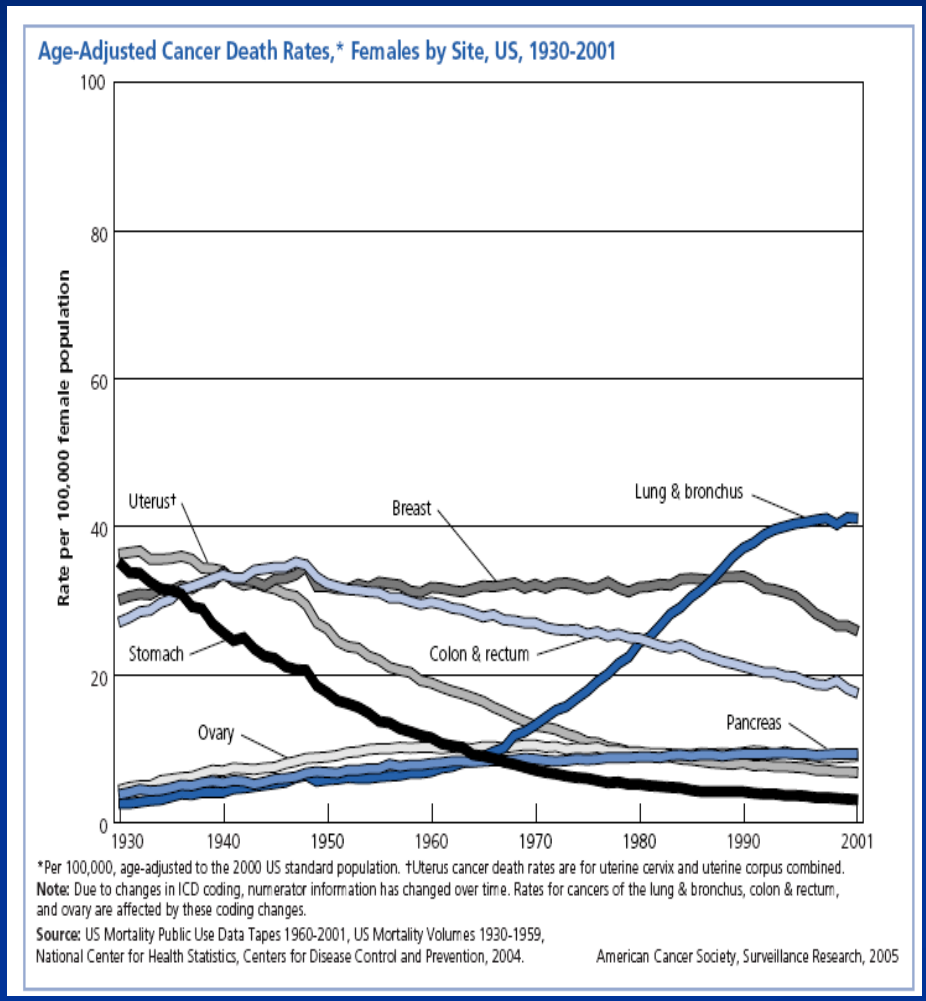
1. THE PROBLEM: EARLY-DISCOVERY BREAST CANCER -- OUTCOME PREDICTION.

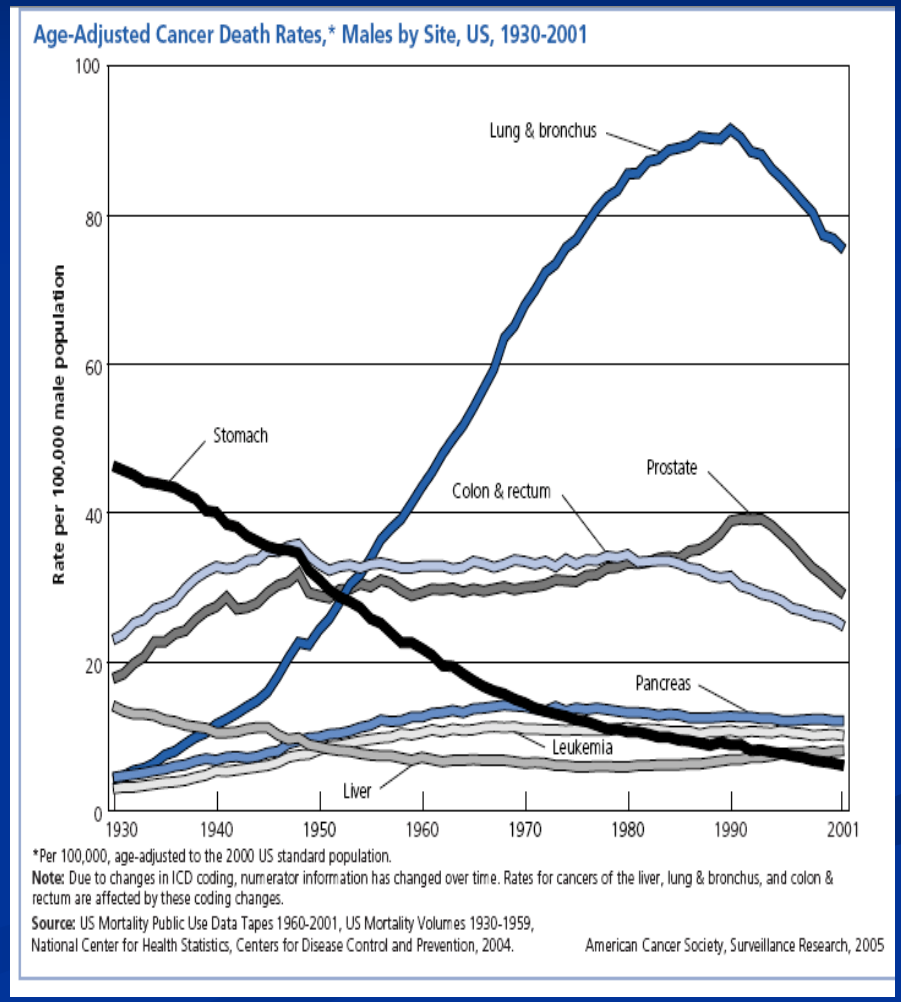
2. THE HOPE: GENE EXPRESSION, DNA MICROARRAYS

3. HYPE: 70 GENES PREDICT OUTCOME! (ALSO 76, 21, 64,…)
   OUTCOME SIGNATURE GENES IN BREAST CANCER:
   IS THERE A UNIQUE SET?

4. PHYSICS: HOW MANY BREAST CANCER SAMPLES ARE NEEDED TO PRODUCE A ROBUST PREDICTIVE GENE LIST?
   Probably Approximately Correct (PAC) – ranking

Cancer Death Rates in USA





ABOUT 600,000 DEATHS PER YEAR IN THE USA

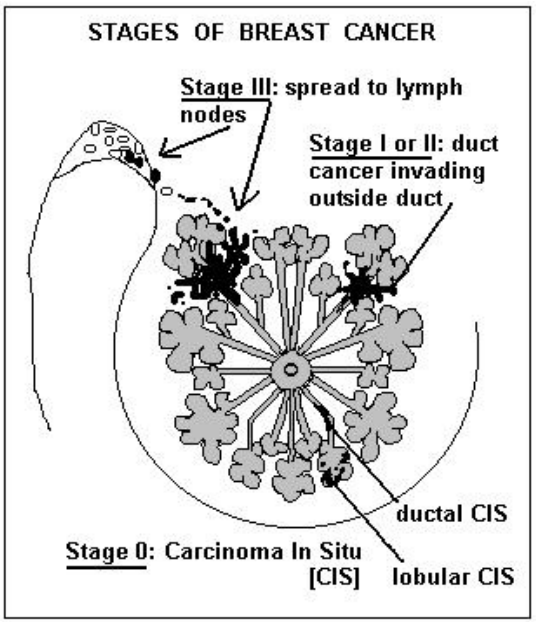
BREAST CANCER:

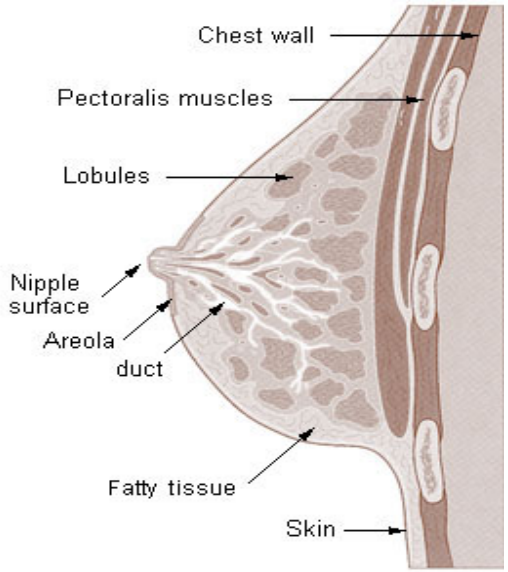
DEATH RATE 30/100,000 per year

INCIDENCE: ABOUT 1 OUT OF 9 WOMEN AFFECTED.

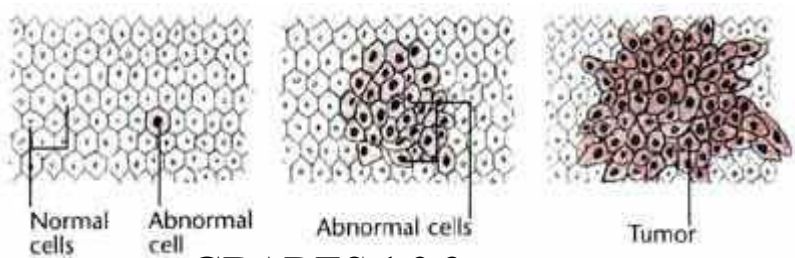
EARLY DISCOVERY: SMALL TUMOR (< 2cm),
NO SPREADING TO LYMPH NODES
LOWEST GRADE, STAGE

TREATMENT: SURGICAL REMOVAL OF TUMOR + RADIOTHERAPY
+ HORMONAL THERAPY (IF ER+)





CHEMOTHERAPY ???



GRADES 1,2,3

NO CHEMOTHERAPY IF PATIENT IS ***LOW RISK:***

ST. GALLEN

Low risk[a]	*Node negative* AND *all* of the following features: pT ≤2 cm, AND Grade 1,[b] AND Absence of peritumoral vascular invasion,[c] AND HER2/*neu* gene neither overexpressed nor amplified,[d] AND Age ≥35 years

NIH

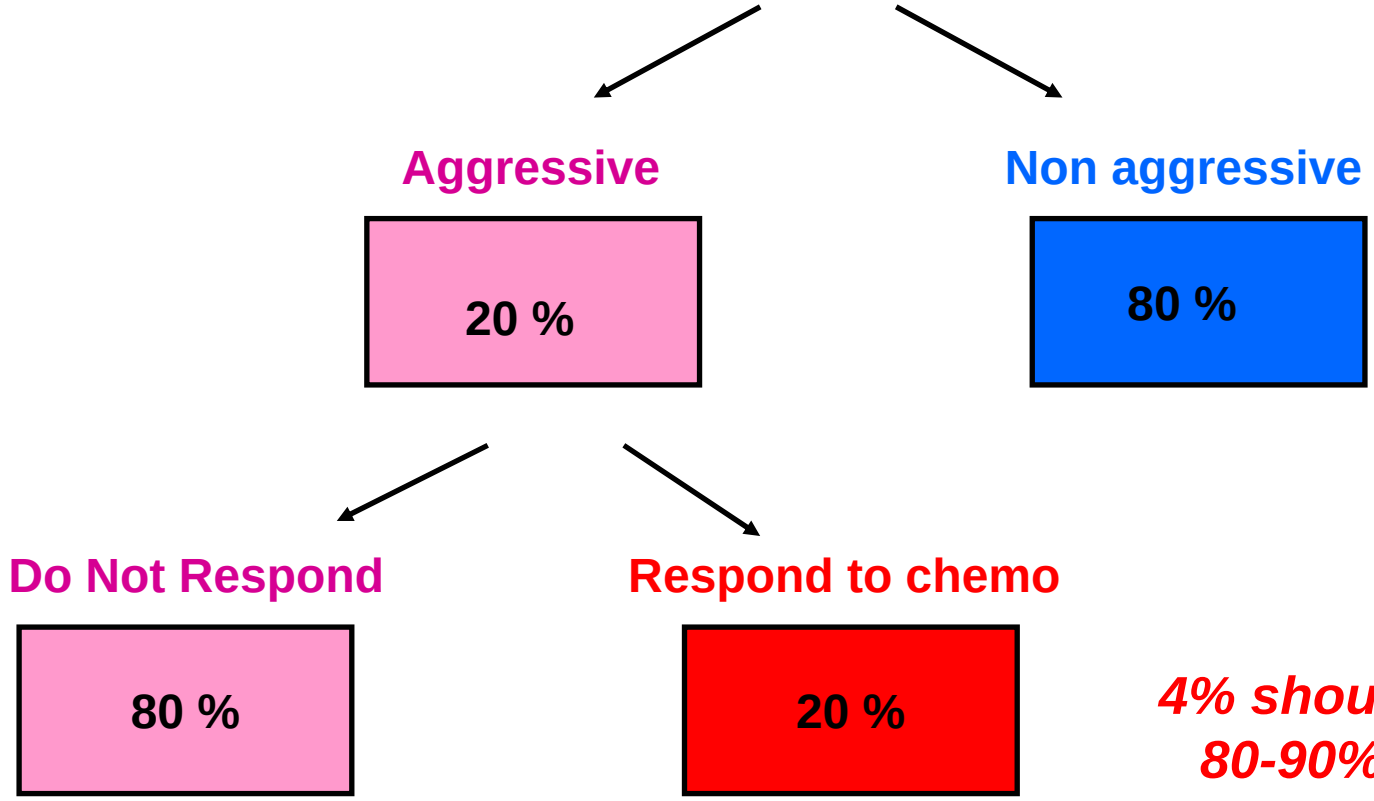
Minimal/low risk†	ER- and/or PgR-positive, and all of the following features: pT‡ ≤ 2 cm, and Grade 1§, and Age‖ ≥ 35 years

NOTTINGHAM

NPI = (0.2 x tumor diameter in cms) + lymph node stage + tumour grade < 3.4

HOW WELL DO THESE CRITERIA WORK?

Early discovery:
Small tumors (few cm), lymph-node negative



4% should get chemo
80-90% get chemo

Can we do better in identifying patients at high risk – and avoid chemotherapy for low-risk?

OUTLINE:

1. THE PROBLEM: EARLY-DISCOVERY BREAST CANCER -- OUTCOME PREDICTION.

2. THE HOPE: GENE EXPRESSION, DNA MICROARRAYS

3. 70 GENES PREDICT OUTCOME! (ALSO 76, 21, 64,…)
   OUTCOME SIGNATURE GENES IN BREAST CANCER:
   IS THERE A UNIQUE SET?

4. HOW MANY BREAST CANCER SAMPLES ARE NEEDED TO PRODUCE A ROBUST PREDICTIVE GENE LIST?
   Probably Approximately Correct (PAC) – ranking

AIM: Identifying patients at high risk

METHOD: Measure gene expression profile of primary tumor and find signature of bad outcome tumors

THE BASIC PARADIGM:
GENE EXPRESSION REFLECTS STATE

THE STATE OF A CELL AND THE ONGOING

BIOLOGICAL PROCESSES ARE REFLECTED

IN ITS EXPRESSION PROFILE:

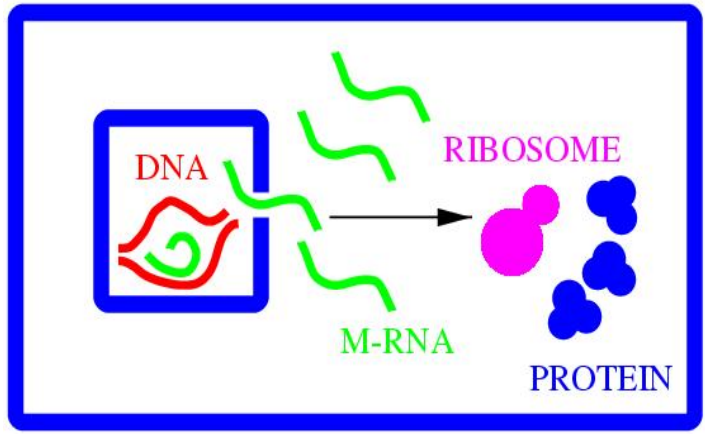
THE EXPRESSION LEVEL OF EACH GENE.

(HUMAN GENOME – 40,000? 24,000? NUMBERS)

HOW DO WE MEASURE THEM?

MEASURING GENE EXPRESSION PROFILE

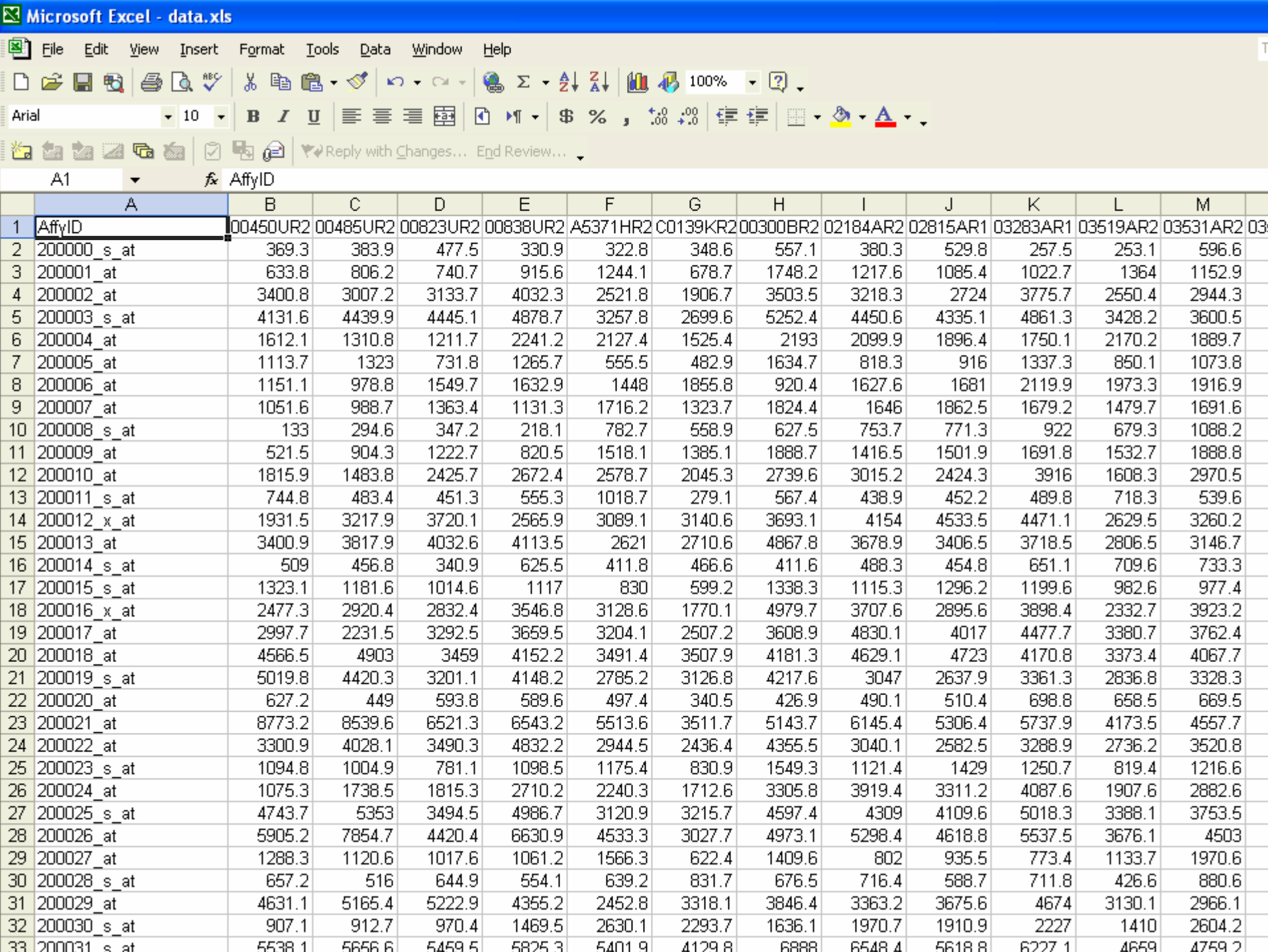
WHEN A PARTICULAR GENE IS EXPRESSED THE CONCENTRATIONS OF ITS CORRESPONDING MESSENGER RNA AND PROTEIN ARE HIGH.



A DNA-CHIP MEASURES CONCENTRATIONS OF THOUSANDS OF DIFFERENT

MESSENGER RNA

LATEST AFFYCHIP: U133P2 – 54,675 (probesets)



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	AffyID	00450UR2	00485UR2	00823UR2	00838UR2	A5371HR2	C0139KR2	00300BR2	02184AR2	02815AR1	03283AR1	03519AR2	03531AR2
2	200000_s_at	369.3	383.9	477.5	330.9	322.8	348.6	557.1	380.3	529.8	257.5	253.1	596.6
3	200001_at	633.8	806.2	740.7	915.6	1244.1	678.7	1748.2	1217.6	1085.4	1022.7	1364	1152.9
4	200002_at	3400.8	3007.2	3133.7	4032.3	2521.8	1906.7	3503.5	3218.3	2724	3775.7	2550.4	2944.3
5	200003_s_at	4131.6	4439.9	4445.1	4878.7	3257.8	2699.6	5252.4	4450.6	4335.1	4861.3	3428.2	3600.5
6	200004_at	1612.1	1310.8	1211.7	2241.2	2127.4	1525.4	2193	2099.9	1896.4	1750.1	2170.2	1889.7
7	200005_at	1113.7	1323	731.8	1265.7	555.5	482.9	1634.7	818.3	916	1337.3	850.1	1073.8
8	200006_at	1151.1	978.8	1549.7	1632.9	1448	1855.8	920.4	1627.6	1681	2119.9	1973.3	1916.9
9	200007_at	1051.6	988.7	1363.4	1131.3	1716.2	1323.7	1824.4	1646	1862.5	1679.2	1479.7	1691.6
10	200008_s_at	133	294.6	347.2	218.1	782.7	558.9	627.5	753.7	771.3	922	679.3	1088.2
11	200009_at	521.5	904.3	1222.7	820.5	1518.1	1385.1	1888.7	1416.5	1501.9	1691.8	1532.7	1888.8
12	200010_at	1815.9	1483.8	2425.7	2672.4	2578.7	2045.3	2739.6	3015.2	2424.3	3916	1608.3	2970.5
13	200011_s_at	744.8	483.4	451.3	555.3	1018.7	279.1	567.4	438.9	452.2	489.8	718.3	539.6
14	200012_x_at	1931.5	3217.9	3720.1	2565.9	3089.1	3140.6	3693.1	4154	4533.5	4471.1	2629.5	3260.2
15	200013_at	3400.9	3817.9	4032.6	4113.5	2621	2710.6	4867.8	3678.9	3406.5	3718.5	2806.5	3146.7
16	200014_s_at	509	456.8	340.9	625.5	411.8	466.6	411.6	488.3	454.8	651.1	709.6	733.3
17	200015_s_at	1323.1	1181.6	1014.6	1117	830	599.2	1338.3	1115.3	1296.2	1199.6	982.6	977.4
18	200016_x_at	2477.3	2920.4	2832.4	3546.8	3128.6	1770.1	4979.7	3707.6	2895.6	3898.4	2332.7	3923.2
19	200017_at	2997.7	2231.5	3292.5	3659.5	3204.1	2507.2	3608.9	4830.1	4017	4477.7	3380.7	3762.4
20	200018_at	4566.5	4903	3459	4152.2	3491.4	3507.9	4181.3	4629.1	4723	4170.8	3373.4	4067.7
21	200019_s_at	5019.8	4420.3	3201.1	4148.2	2785.2	3126.8	4217.6	3047	2637.9	3361.3	2836.8	3328.3
22	200020_at	627.2	449	593.8	589.6	497.4	340.5	426.9	490.1	510.4	698.8	658.5	669.5
23	200021_at	8773.2	8539.6	6521.3	6543.2	5513.6	3511.7	5143.7	6145.4	5306.4	5737.9	4173.5	4557.7
24	200022_at	3300.9	4028.1	3490.3	4832.2	2944.5	2436.4	4355.5	3040.1	2582.5	3288.9	2736.2	3520.8
25	200023_s_at	1094.8	1004.9	781.1	1098.5	1175.4	830.9	1549.3	1121.4	1429	1250.7	819.4	1216.6
26	200024_at	1075.3	1738.5	1815.3	2710.2	2240.3	1712.6	3305.8	3919.4	3311.2	4087.6	1907.6	2882.6
27	200025_s_at	4743.7	5353	3494.5	4986.7	3120.9	3215.7	4597.4	4309	4109.6	5018.3	3388.1	3753.5
28	200026_at	5905.2	7854.7	4420.4	6630.9	4533.3	3027.7	4973.1	5298.4	4618.8	5537.5	3676.1	4503
29	200027_at	1288.3	1120.6	1017.6	1061.2	1566.3	622.4	1409.6	802	935.5	773.4	1133.7	1970.6
30	200028_s_at	657.2	516	644.9	554.1	639.2	831.7	676.5	716.4	588.7	711.8	426.6	880.6
31	200029_at	4631.1	5165.4	5222.9	4355.2	2452.8	3318.1	3846.4	3363.2	3675.6	4674	3130.1	2966.1
32	200030_s_at	907.1	912.7	970.4	1469.5	2630.1	2293.7	1636.1	1970.7	1910.9	2227	1410	2604.2
33	200031_s_at	5538.1	5656.6	5459.5	5825.3	5401.9	4129.8	6888	6548.4	5618.8	6227.1	4659	4759.2

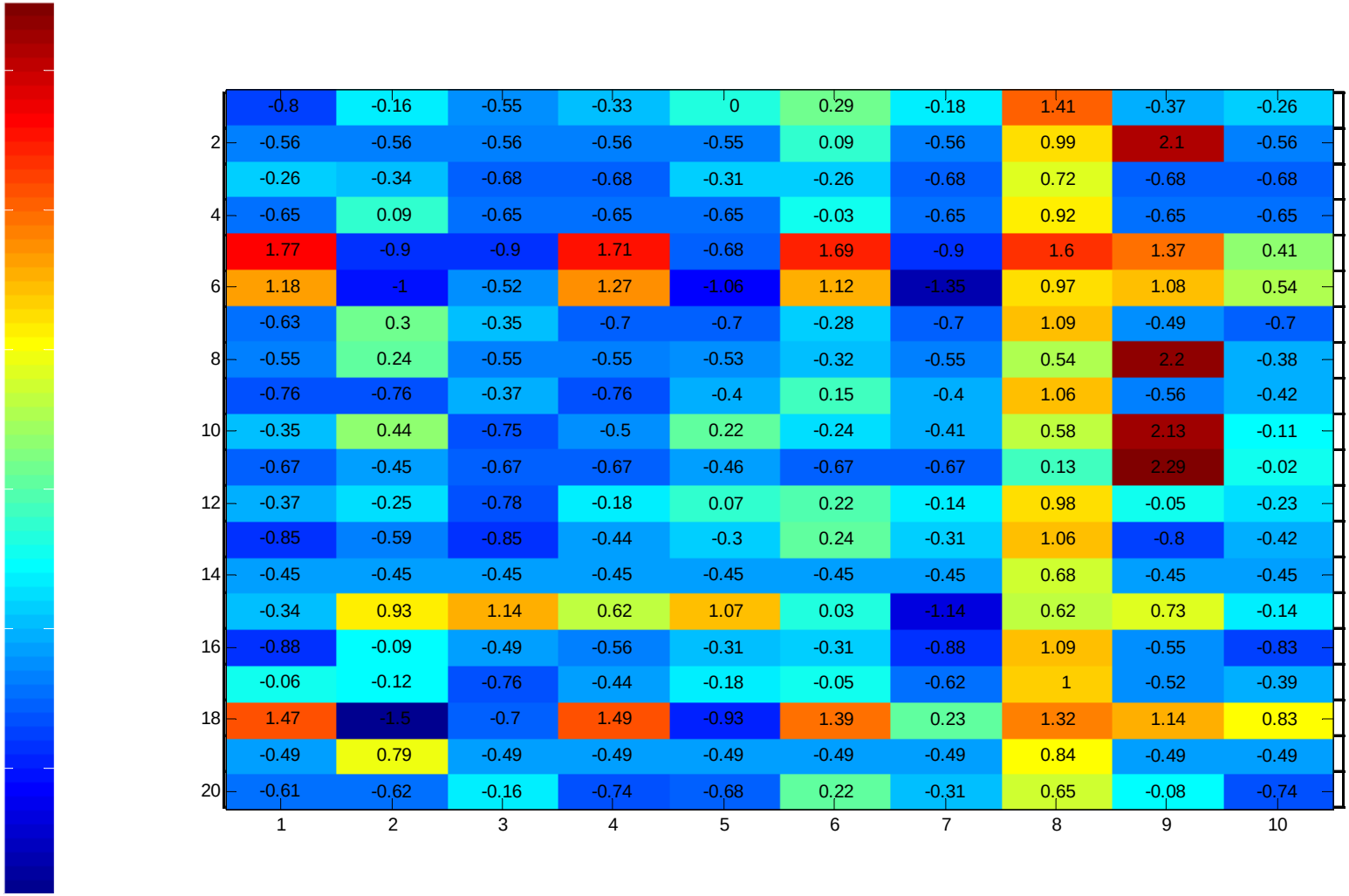
VISUALIZATION: HOW DOES ONE SHOW SO MANY NUMBERS?

COLOR CODE: REPRESENT NUMBERS BY COLORS

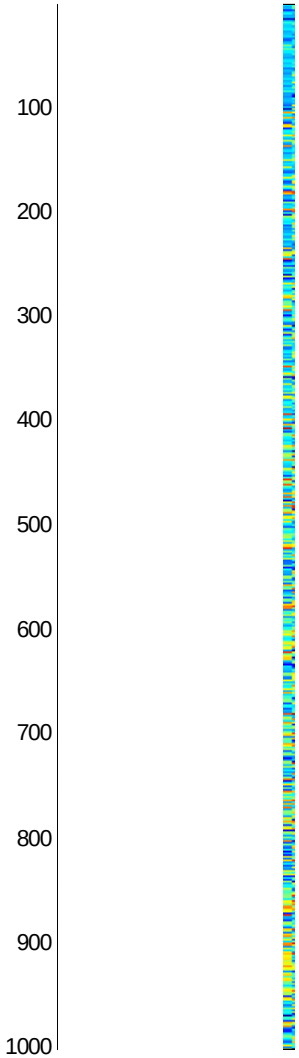
-0.80	-0.16	-0.55	-0.33	0.00	0.29	-0.18	1.41	-0.37	-0.26
-0.56	-0.56	-0.56	-0.56	-0.55	0.09	-0.56	0.99	2.10	-0.56
-0.26	-0.34	-0.68	-0.68	-0.31	-0.26	-0.68	0.72	-0.68	-0.68
-0.65	0.09	-0.65	-0.65	-0.65	-0.03	-0.65	0.92	-0.65	-0.65
1.77	-0.90	-0.90	1.71	-0.68	1.69	-0.90	1.60	1.37	0.41
1.18	-1.00	-0.52	1.27	-1.06	1.12	-1.35	0.97	1.08	0.54
-0.63	0.30	-0.35	-0.70	-0.70	-0.28	-0.70	1.09	-0.49	-0.70
-0.55	0.24	-0.55	-0.55	-0.53	-0.32	-0.55	0.54	2.20	-0.38
-0.76	-0.76	-0.37	-0.76	-0.40	0.15	-0.40	1.06	-0.56	-0.42
-0.35	0.44	-0.75	-0.50	0.22	-0.24	-0.41	0.58	2.13	-0.11
-0.67	-0.45	-0.67	-0.67	-0.46	-0.67	-0.67	0.13	2.29	-0.02
-0.37	-0.25	-0.78	-0.18	0.07	0.22	-0.14	0.98	-0.05	-0.23
-0.85	-0.59	-0.85	-0.44	-0.30	0.24	-0.31	1.06	-0.80	-0.42
-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	-0.45	0.68	-0.45	-0.45
-0.34	0.93	1.14	0.62	1.07	0.03	-1.14	0.62	0.73	-0.14
-0.88	-0.09	-0.49	-0.56	-0.31	-0.31	-0.88	1.09	-0.55	-0.83
-0.06	-0.12	-0.76	-0.44	-0.18	-0.05	-0.62	1.00	-0.52	-0.39
1.47	-1.50	-0.70	1.49	-0.93	1.39	0.23	1.32	1.14	0.83
-0.49	0.79	-0.49	-0.49	-0.49	-0.49	-0.49	0.84	-0.49	-0.49
-0.61	-0.62	-0.16	-0.74	-0.68	0.22	-0.31	0.65	-0.08	-0.74

VISUALIZATION: HOW DOES ONE SHOW SO MANY NUMBERS?

COLOR CODE: REPRESENT NUMBERS BY COLORS



COLON CANCER DATA:



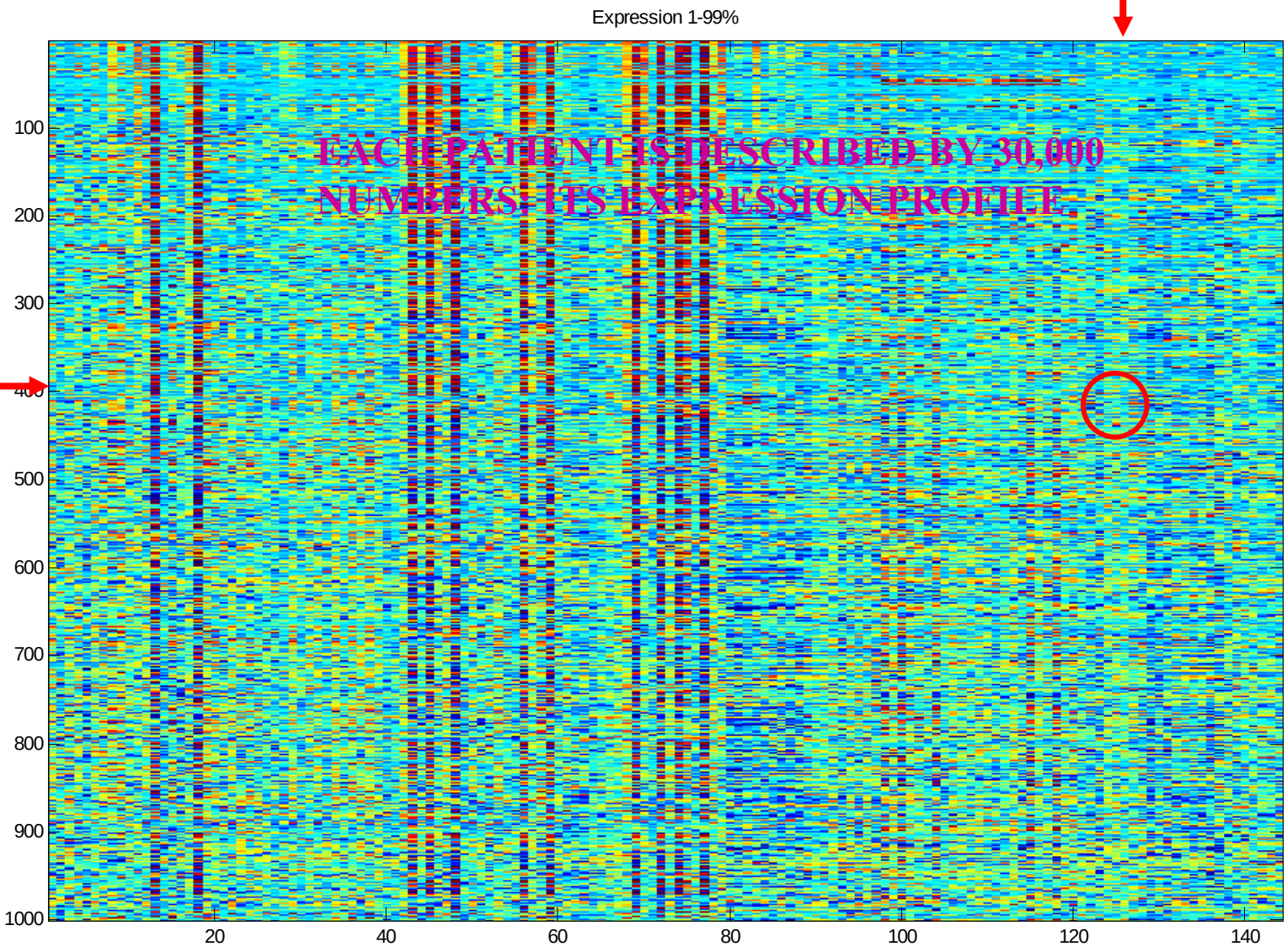
EACH PATIENT IS DESCRIBED BY 30,000 NUMBERS: ITS EXPRESSION PROFILE

COLON CANCER DATA:

E_{ij} = EXPRESSION LEVEL OF GENE i IN SAMPLE j

Sample # 127

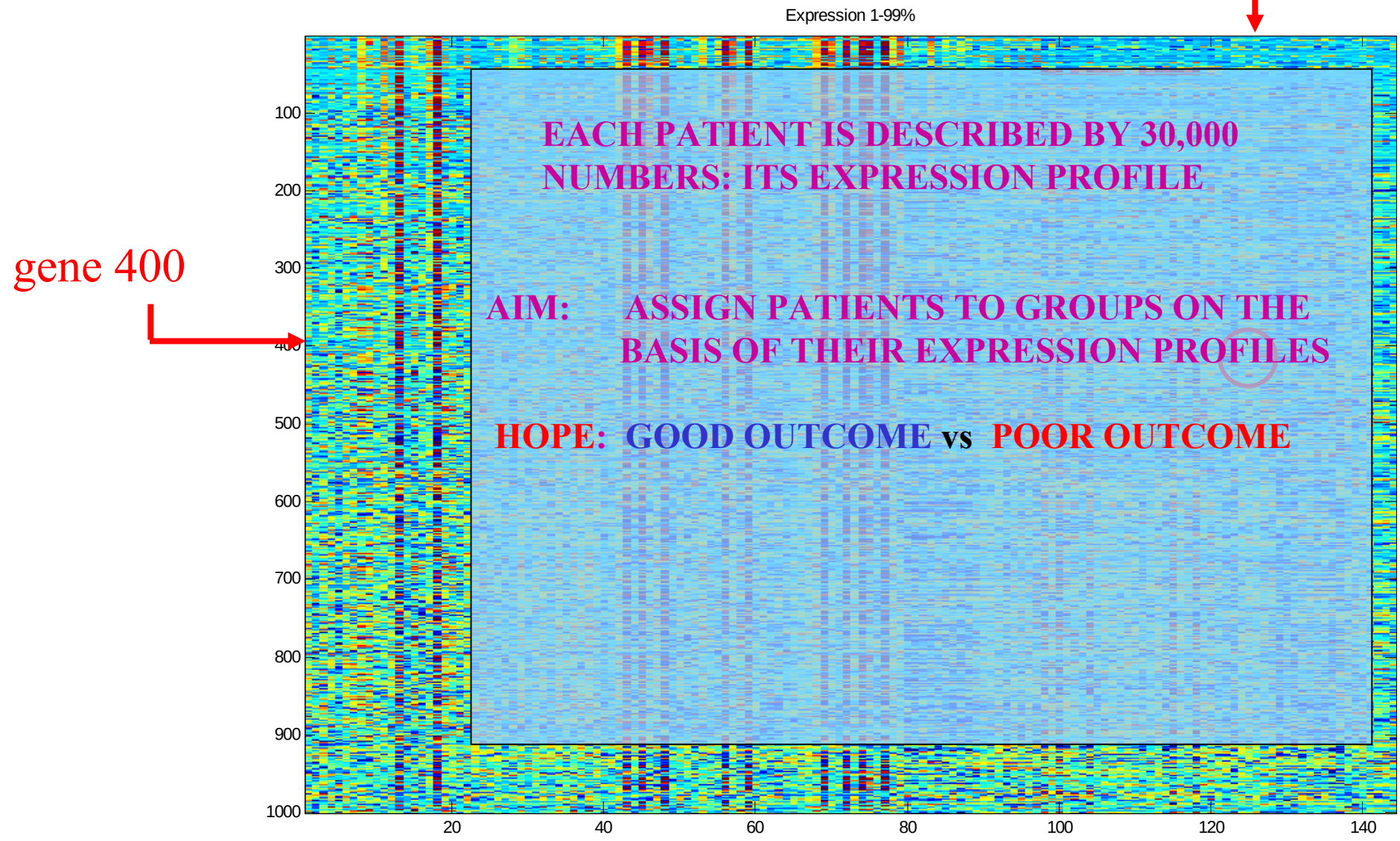
gene 400



COLON CANCER DATA:

E_{ij} = EXPRESSION LEVEL OF GENE i
IN SAMPLE j

Sample # 127



OUTLINE:

1. THE PROBLEM: EARLY-DISCOVERY BREAST CANCER -- OUTCOME PREDICTION.

2. THE HOPE: GENE EXPRESSION, DNA MICROARRAYS

3. 70 GENES PREDICT OUTCOME! (ALSO 76, 21, 64,…)
   OUTCOME SIGNATURE GENES IN BREAST CANCER:
   IS THERE A UNIQUE SET?

4. HOW MANY BREAST CANCER SAMPLES ARE NEEDED TO PRODUCE A ROBUST PREDICTIVE GENE LIST?
   Probably Approximately Correct (PAC) – ranking

Flood of signatures

Available online http://breast-ca

Research article

Gene expression profiling spares early breast ca from adjuvant therapy: derived and validated in t population-based cohorts

Yudi Pawitan[1], Judith Bjöhle[2], Lukas Amler[3], Anna-Lena Borg[2], Suzanne
Xia Han[4], Lars Holmberg[5], Fei Huang[4], Sigrid Klaar[2], Edison T Liu[6], Lance
Ale
Sa

PLOS BIOL

Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer*†, Hongyue Dai†‡, Marc J. van de Vijver*†, Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse*, Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen*, George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡,

Gene Expression Signature of Fibroblast Seru Response Predicts Human Cancer Progressior Similarities between Tumors and Wounds

Howard Y. Chang[1,2], Julie B. Sneddon[2], Ash A. Alizadeh[2¤1], Ruchira Sood[2], Rob B. West[3], Kelli Montgomery[3], Jen-Tsan Chi[2], Matt van de Rijn[3], David Botstein[4¤2], Patrick O. Brown[2,5*]

PNAS

Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival

Howard Y. Chang[a,b,c], Dimitry S. A. Nuyten[c,d,e], Julie B. Sneddon[b], Trevor Hastie[f], Robert Tibshirani[f], Therese Sørlie[b,g], Hongyue Dai[h,i], Yudong D. He[h,i], Laura J. van't Veer[d,j], Harry Bartelink[e], Matt van de Rijn[i], Patrick O. Brown[b,k,j], and Marc J. van de Vijver[d,i]

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

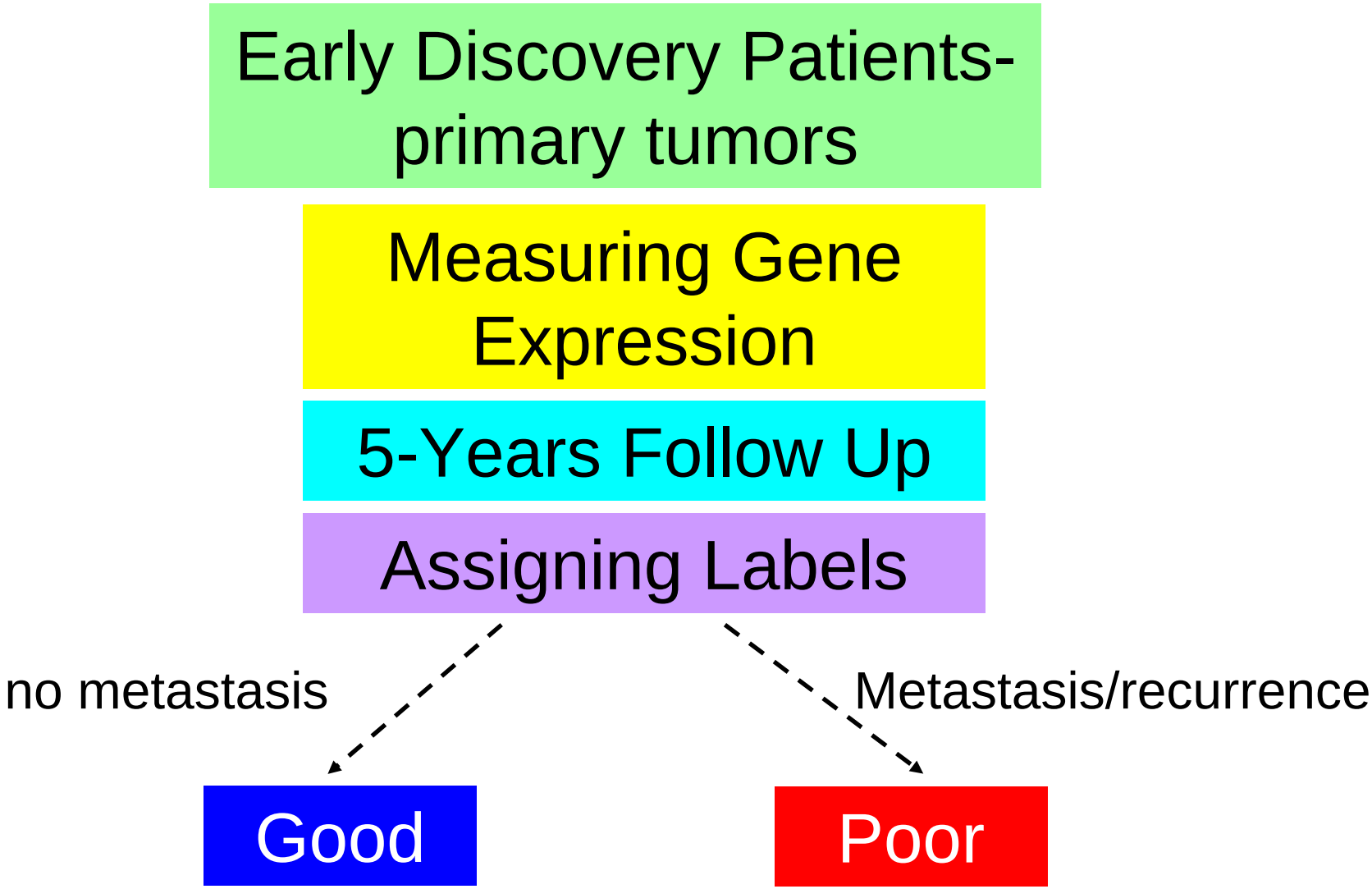
Gene-expression profiles to predict dis lymph-node-negative primary breast

Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri
Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foeker

A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer

Soonmyung Paik, M.D., Steven Shak, M.D., Gong Tang, Ph.D., Chungyeul Kim, M.D., Joffre Baker, Ph.D., Maureen Cronin, Ph.D., Frederick L. Baehner, M.D., Michael G. Walker, Ph.D., Drew Watson, Ph.D., Taesung Park, Ph.D., William Hiller, H.T., Edwin R. Fisher, M.D., D. Lawrence Wickerham, M.D., John Bryant, Ph.D., and Norman Wolmark, M.D.

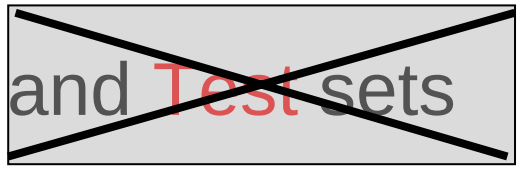
Prepare Training Set for outcome prediction:



Predicting Outcome with Expression Profiling

Step I - Grouping

Divide the samples into Training and Test sets



Step II - Feature selection

Find ***the*** subset of N_{TOP} predictive genes

Feature Selection

Selecting a short list of predictive genes
Typically ~ 100 samples for training, hence use
~ 100 genes (out of ~ 10,000 on chip)

Why short list?

1. Avoid overtraining/reduce generalization error
2. Gain insight into the biological mechanism underlying outcome.
3. Less genes – simpler chip, easier prediction.

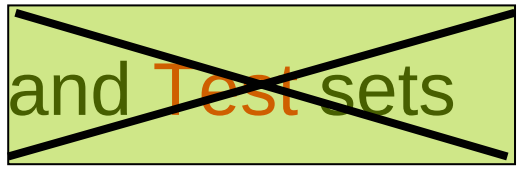
How? Two steps:

- **Rank** genes according to their individual predictive power (correlation with outcome).
- Select the N_{TOP} **highest ranked** genes.

Predicting Outcome with Expression Profiling

Step I - Grouping

Divide the samples into Training and Test sets



Step II - Feature selection

Find ***the*** subset of N_{TOP} predictive genes

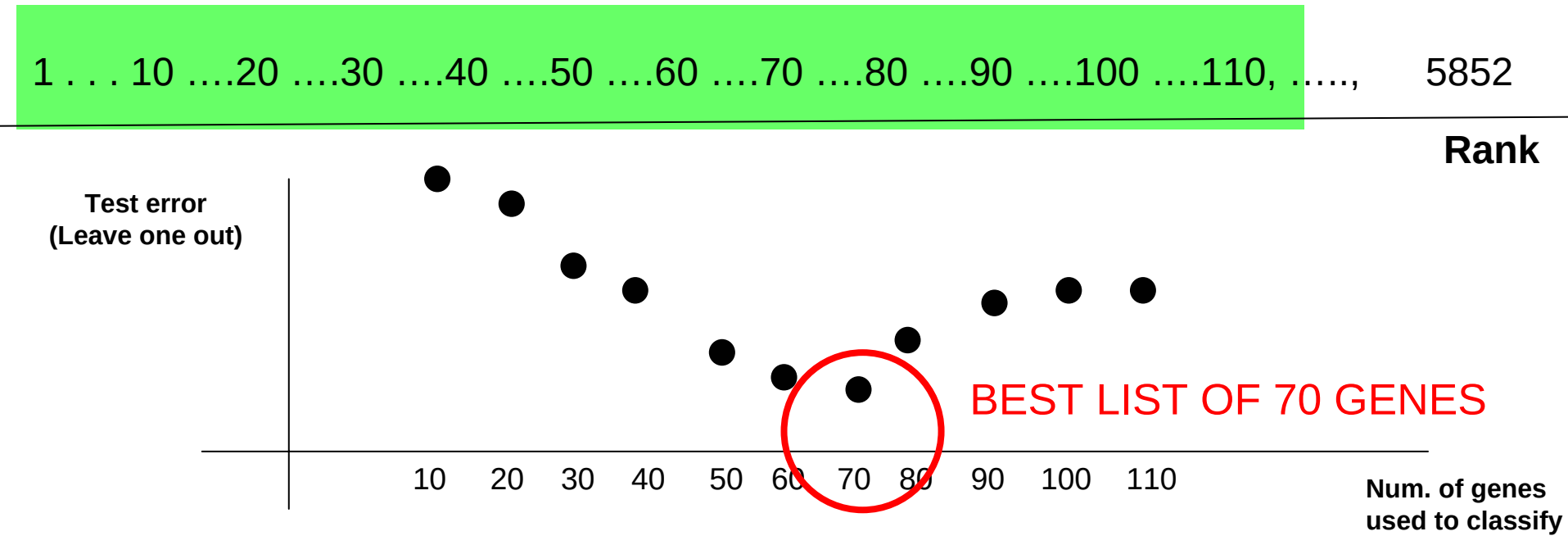
Step III - classification rule

Develop prediction rule using the selected genes

Determine N_{TOP}

Van't Veer's Analysis

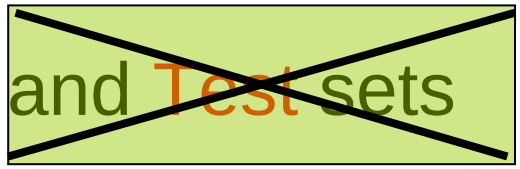
1. Rank the genes according to their correlation to disease outcome.
2. Search from the top for the set of genes that has the best performance to predict outcome



Predicting Outcome with Expression Profiling

Step I - Grouping

Divide the samples into Training and Test sets



Step II - Feature selection

Find ***the*** subset of N_{TOP} predictive genes

Step III - classification rule

Develop prediction rule using the selected genes

Determine N_{TOP}

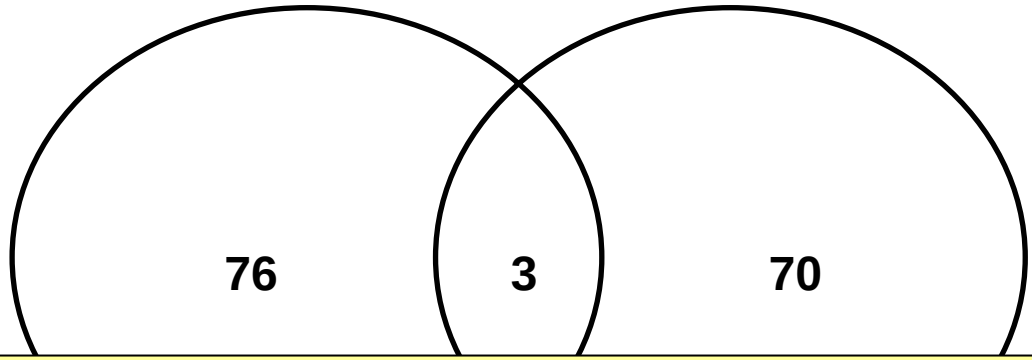
Step IV – prediction error

Check classifier performance

Two Successful Analyses

Wang et al. Lancet 2005, List = 76 top-ranked genes

Van't Veer. et al. Nature 2002, List = 70 top-ranked genes



VERY SMALL OVERLAP!!! POOR TRANSFERABILITY!!!
WHY ???

Different Platforms ?

Different Populations of Patients ?

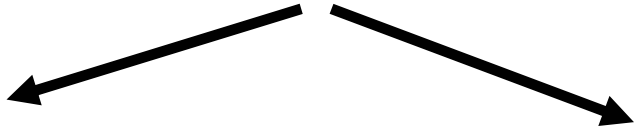
Different Types of Analysis?

NO!!

Selecting 70 genes: Van't Veer's dataset

Nature, 2002

96 breast sporadic tumors



46

Poor prognosis patients
(developed distant metastases Within 5 years)

50

Good prognosis patients
(Did not develop distant metastases Within 5 years)

5852 genes

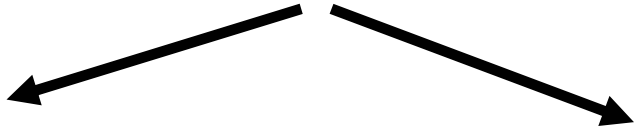
Significantly regulated

1. Select 77 patients for training set
2. Measure, over the training set, the correlation of each genes' expression levels with outcome
3. Rank 5852 genes by correlation, **take top 70**

Selecting 70 genes: Van't Veer's dataset

Nature, 2002

96 breast sporadic tumors



46

Poor prognosis patients
(developed distant metastases Within 5 years)

50

Good prognosis patients
(Did not develop distant metastases Within 5 years)

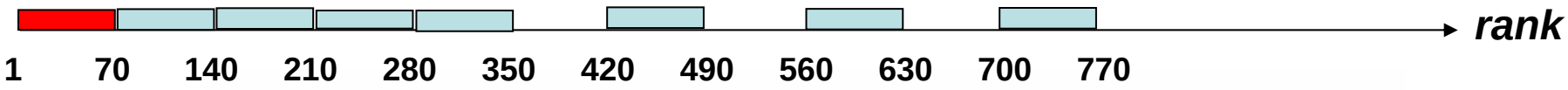
5852 genes

Significantly regulated

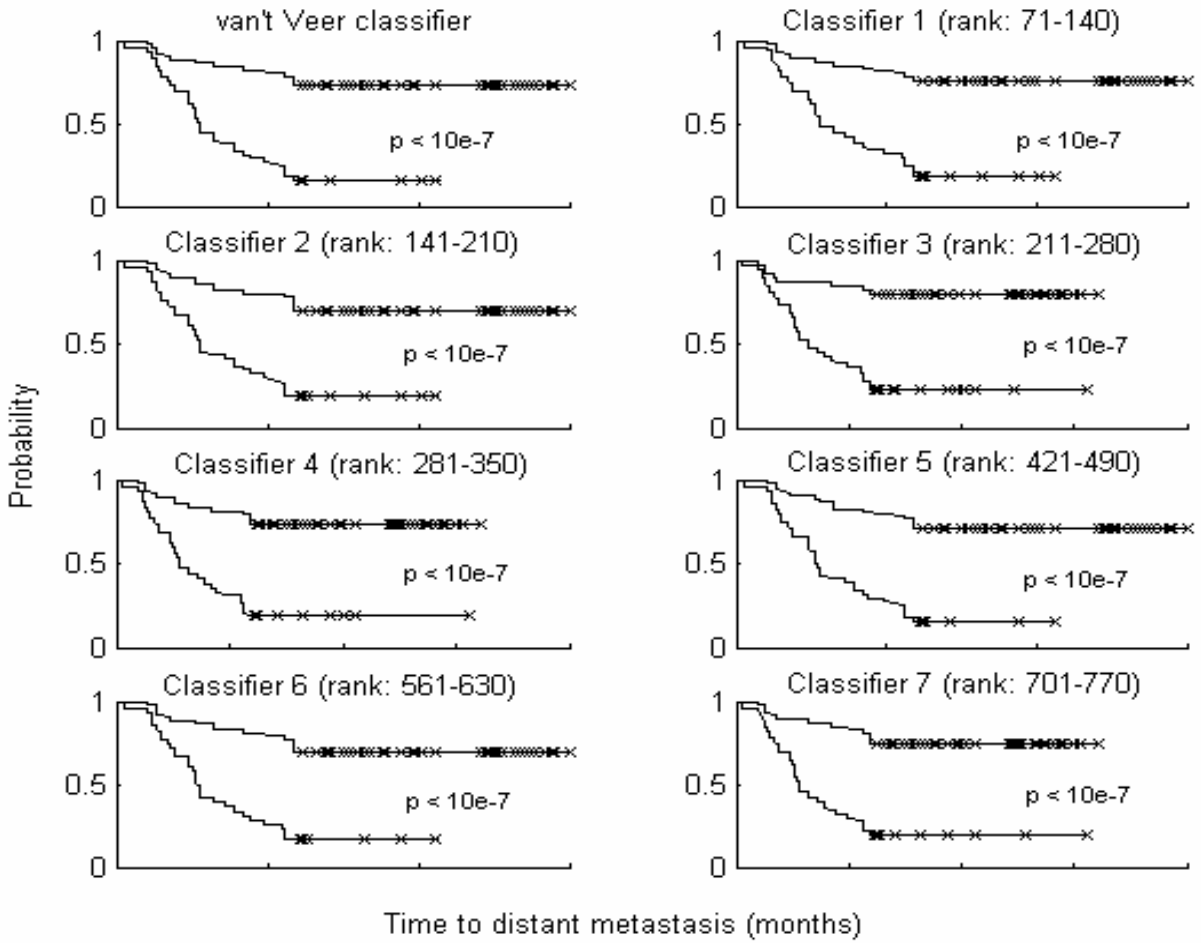
1. Select 77 patients for training set
2. Measure, over the traini[illegible]relation of each genes' [illegible]vels with outcome
3. Rank 5852 [illegible]by correlation, **take top 70**

Is this list unique?

Many sets of 70 genes can be used to predict time to distance metastasis

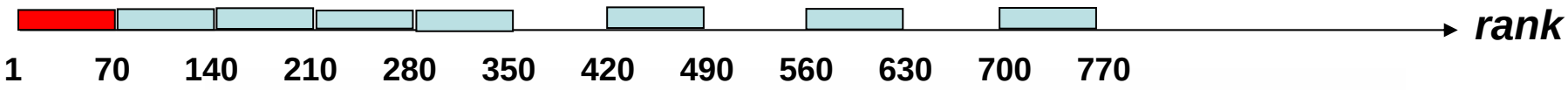


Van't Veer

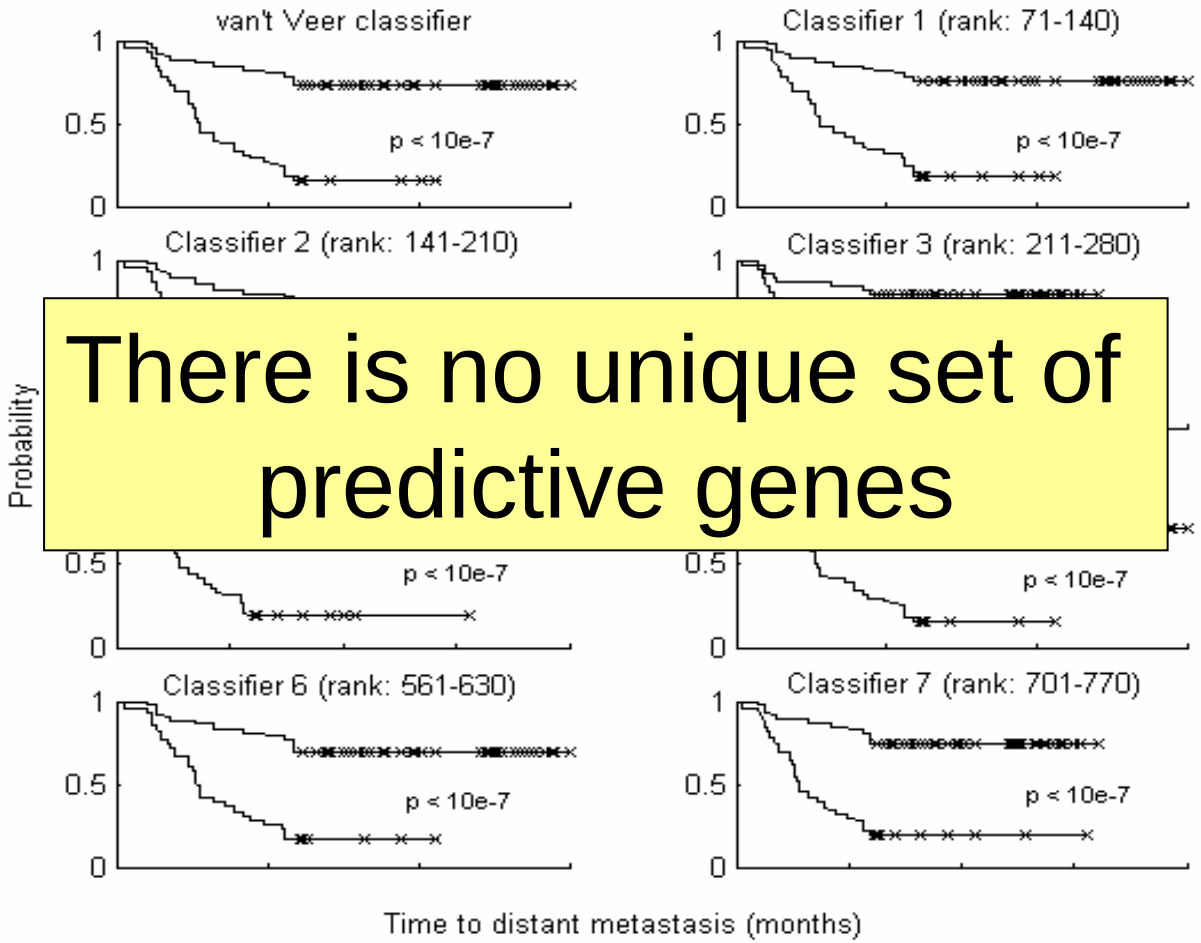


Ein-Dor et al, Bioinformatics 2005

Many sets of 70 genes can be used to predict time to distance metastasis



Van't Veer



There is no unique set of predictive genes

Ein-Dor et al, Bioinformatics 2005

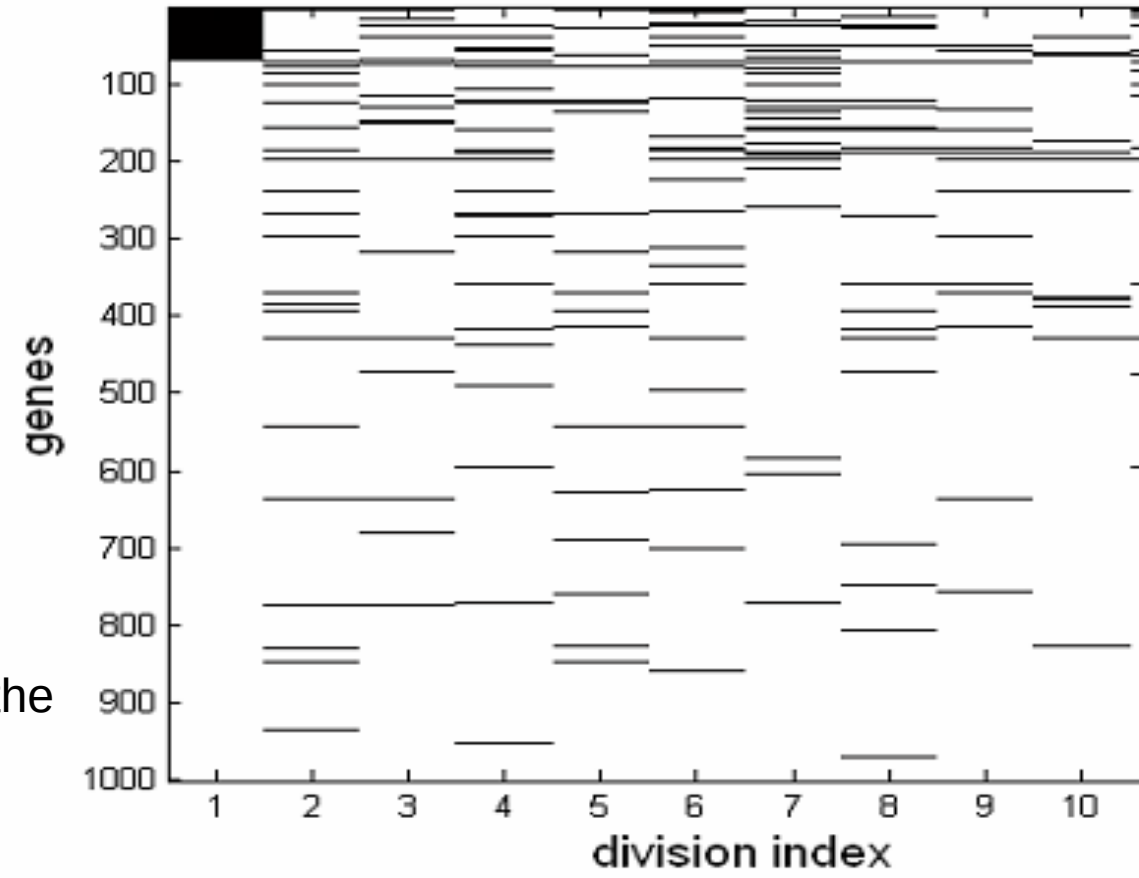
A gene's rank may fluctuate

Step I

1. Choose a group of 77 (out of 96) samples (training set).
2. Order the genes according their correlation to survival.
3. Mark by black lines the top 70 genes.

Step II

1. Choose a different training set (new 77 samples).
2. Order the genes according their correlation to survival (based on the **new training set**).
3. Mark by black lines the top 70 genes of the **first training set**.
4. Do 10 times…



Ein-Dor et al, Bioinformatics 2005

A gene's rank may fluctuate

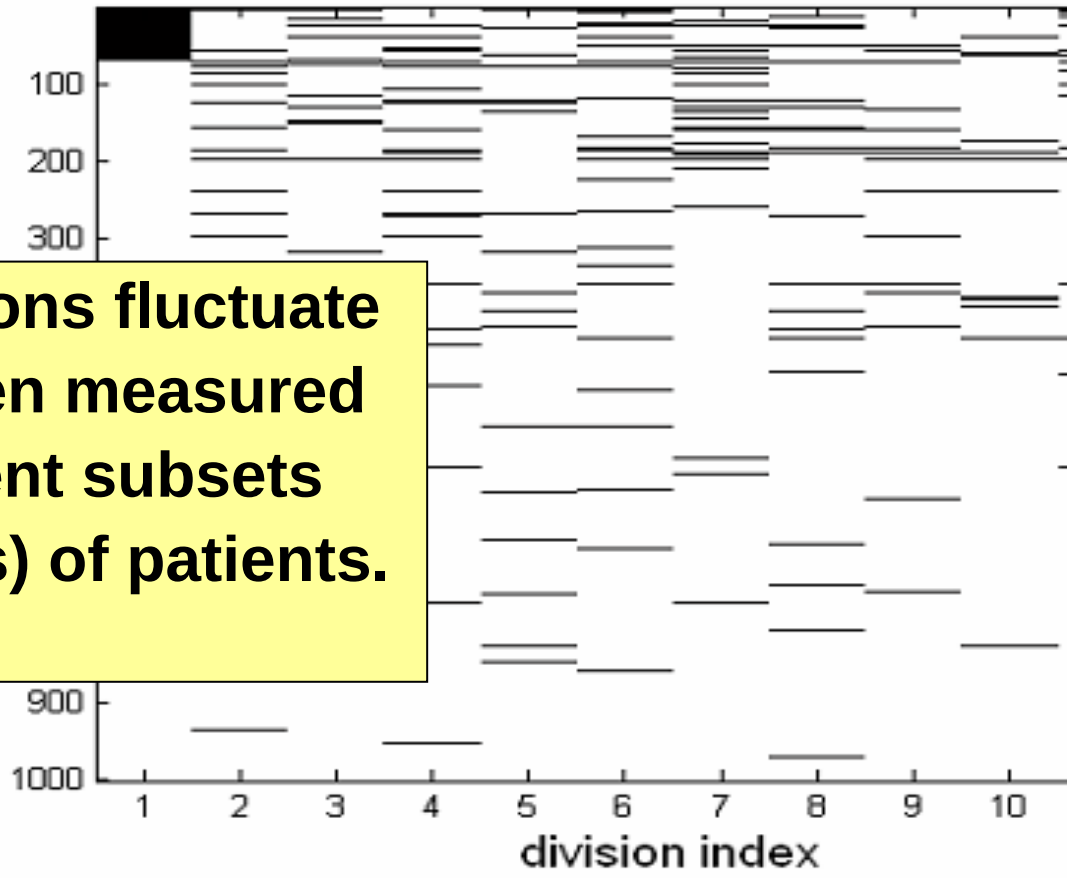
Step I

1. Choose a group of 77 (out of 96) samples (training set).
2. Order the genes according their correlation to survival.
3. Mark by black lines the top 70 genes.

Step II

1. Choose a differ
   (new 77 sample
2. Order the genes
   correlation to survival (based on the **new training set**).
3. Mark by black lines the top 70 genes of the **first training set**.
4. Do 10 times…

The correlations fluctuate strongly when measured over different subsets (Training sets) of patients.



Ein-Dor et al, Bioinformatics 2005

The problem of ranking

An Example – Race (semi-Marathon)

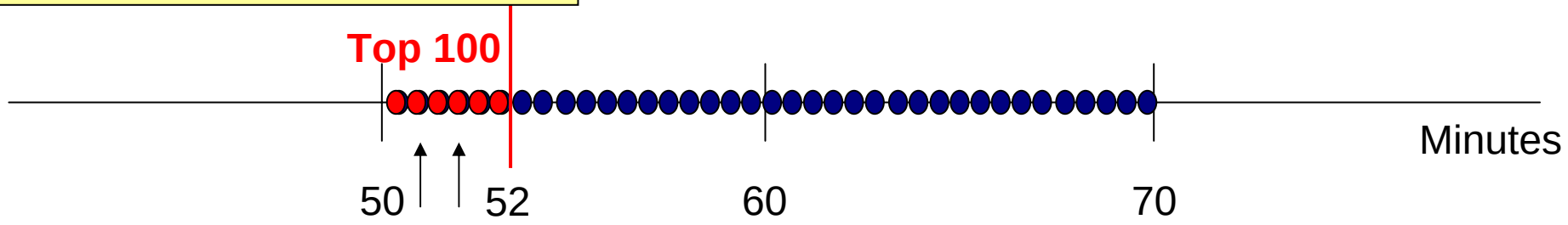
1000 runners (**NON** professional and at about the **SAME LEVEL**)

Each runner can finish the race within 50 - 70 min

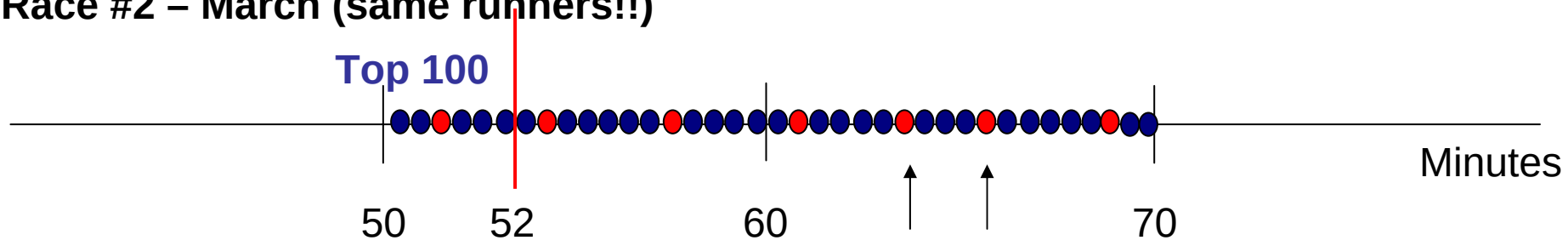
Race #1 - February

The top 100 runners are in the time range:
100*1.2 sec = **2 min**

The average time difference between two consecutive runners is 20*60/1000 = **1.2 sec**



Race #2 – March (same runners!!)



REMEMBER: THE TIME OF EACH RUNNER FLUCTUATES WITHIN A 20 MINUTE INTERVAL

Tumor Biology:

Possible explanation for moderate and highly fluctuating, noisy correlation values: heterogeneity of the tumors.

To get a robust predictive gene list (one that two experimenters will agree on 50% of the genes) one needs a large number of training samples.

how many?

OUTLINE:

1. THE PROBLEM: EARLY-DISCOVERY BREAST CANCER -- OUTCOME PREDICTION.

2. THE HOPE: GENE EXPRESSION, DNA MICROARRAYS

3. 70 GENES PREDICT OUTCOME! (ALSO 76, 21, 64,…)
   OUTCOME SIGNATURE GENES IN BREAST CANCER:
   IS THERE A UNIQUE SET?

4. HOW MANY BREAST CANCER SAMPLES ARE NEEDED TO PRODUCE A ROBUST PREDICTIVE GENE LIST?
   Probably Approximately Correct (PAC) – ranking

INSTABILITY OF GENE LIST IS CAUSED BY FLUCTUATIONS OF THE RANKING OF INDIVIDUAL GENES.

THE RANKING OF GENE g IS DETERMINED BY C_g , ITS CORRELATION WITH OUTCOME.

FLUCTUATION OF C_g IS CAUSED BY "SAMPLING ERROR" – DUE TO THE FINITE SIZE n OF THE SAMPLE (OF PATIENTS) THAT WAS USED TO CALCULATE C_g

SELECT n PATIENTS, CALCULATE C_g; SELECT $N_{TOP} = \alpha N_g$ GENES WITH HIGHEST $|C_g|$. REPEAT WITH ANOTHER n TO GET ANOTHER GENE LIST.

f = OVERLAP OF THE TWO GENE LISTS

AIM: CALCULATE THE PROB. DISTRIBUTION $P_{n,\alpha}(f)$, TO ANSWER:

HOW MANY PATIENTS n ARE NEEDED TO HAVE

(PAC, Valiant 1984)

Prob $[f > 1 - \varepsilon] > 1 - \delta$

FLUCTUATION OF C_g IS CAUSED BY "SAMPLING ERROR" – DUE TO THE FINITE SIZE n OF THE SAMPLE (OF PATIENTS) THAT WAS USED TO CALCULATE C_g

DISTRIBUTION OF C_g IS HARD TO CALCULATE

FISHER 1915, 1921: THE VARIABLE $Z = \tanh^{-1}(C)$ IS NORMAL DISTRIBUTED:

True Z_t

$P(Z) = N(Z_t, \sigma_n)$ with variance $\sigma_n^2 = 1/(n-3)$ (under certain assumptions)

Prob [$|Z_m| < x$] = $P(x, Z, \sigma) = \int_{-x}^{x} dZ_m \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(Z_m - Z)^2}{2\sigma^2})$.

$$P_{n,\alpha}(f) = \frac{1}{Nr} \int_0^\infty dx_1 dx_2 \sum_{h,l \in \{0,1\}^{N_g}} \left\{ \delta\left(\sum_{j=1}^{N_g} h_j \, , N_{TOP}\right) \delta\left(\sum_{j=1}^{N_g} l_j \, , N_{TOP}\right) \delta\left(\sum_{j=1}^{N_g} h_j l_j \, , f \; N_{TOP}\right) \right.$$

$$\left. \prod_{j=1}^{N_g} \left[(1-h_j)P(x_1, Z_{tj}, \sigma_n) + h_j(1 - P(x_1, Z_{tj}, \sigma_n))\right] \prod_{k=1}^{N_g} \left[(1-l_k)P(x_2, Z_{tk}, \sigma_n) + l_k(1 - P(x_2, Z_{tk}, \sigma_n))\right] \right\}$$

True Z_{tj}

$h_j = 1$ IF $|Z_j| > x_1$ $\quad l_k = 1$ IF $|Z_k| > x_2$ $\quad \boldsymbol{\delta(n,m) = 1}$ if $\boldsymbol{n=m}$, $\boldsymbol{0}$ if $\neq$

STEPS: USE INTEGRAL REPRESENTATION OF δ

$P(f)$ REWRITTEN AS

$$P_{n,\alpha}(f) = \frac{1}{Nr}\int_0^\infty dx_1 dx_2 \int_{-\pi}^{\pi} \frac{dydzdw}{(2\pi)^3} \exp\left(-N_g F\right),$$

where

$$F(x_1, x_2, y, x, w; f) = -i(1-\alpha)y - i(1-\alpha)z - i(1-\alpha f)w - 2\int_0^\infty q(Z)dZ \ln(A(x_1, x_2, y, z, w, Z)).$$

depends on σ_n^2

Distribution of the TRUE Z_t (normal, with variance V_t)

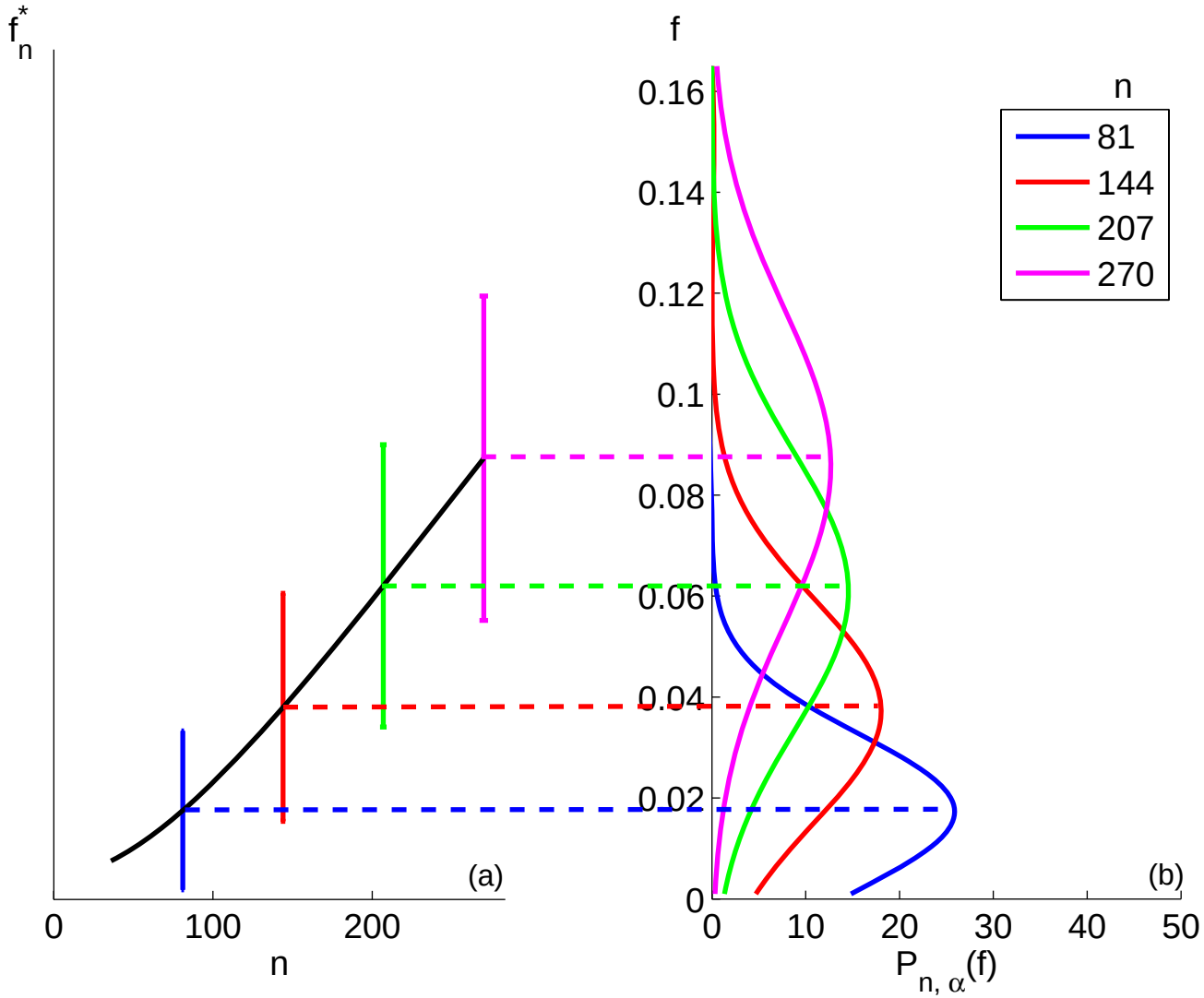
(saddle-point integration, expansion in $1/N_g$)

$$P_{n,\alpha}(f) = \frac{1}{\sqrt{2\pi}\Sigma_n} e^{-\frac{(f-f_n^*)^2}{2\Sigma_n^2}}$$

derived from data

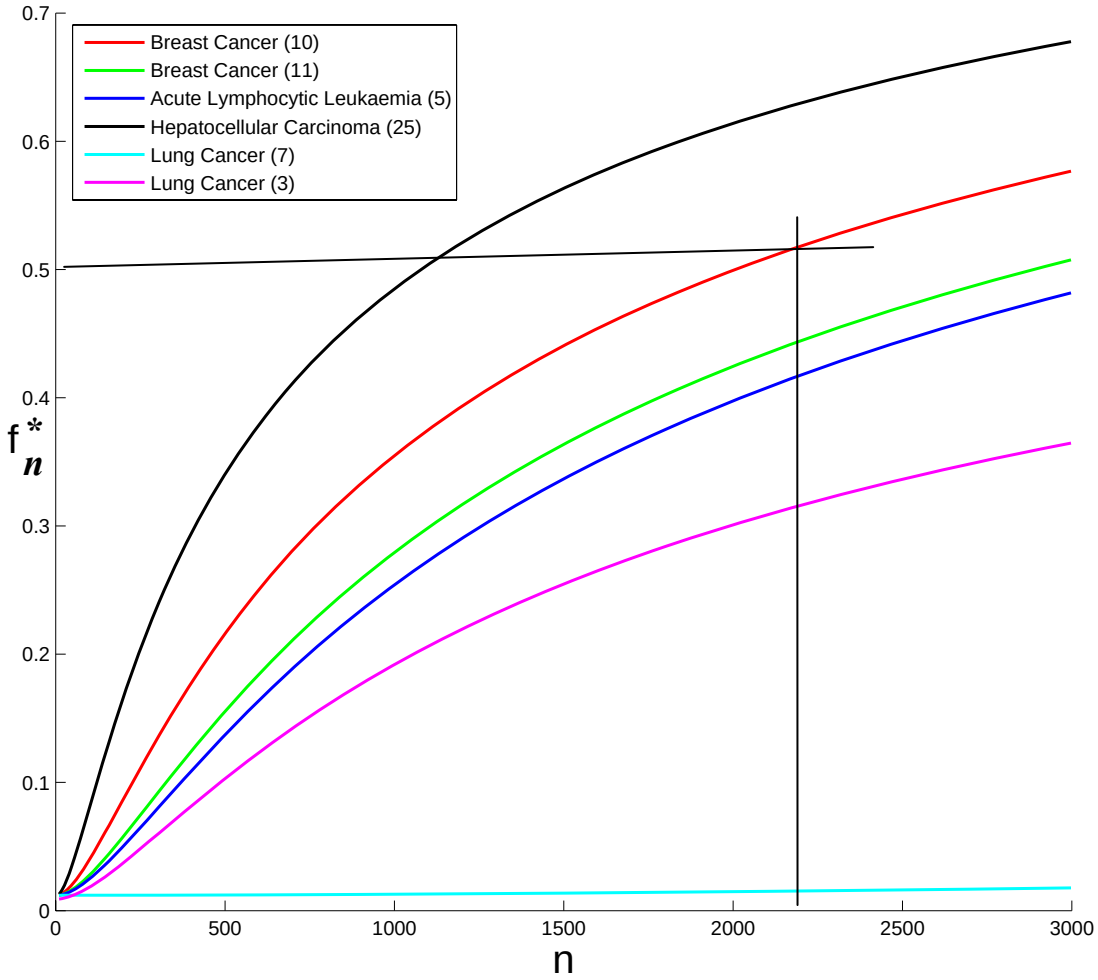
AIM: CALCULATE THE PROB. DISTRIBUTION $P_{n,\alpha}(f)$

$$P_{n,\alpha}(f) = \frac{1}{\sqrt{2\pi}\Sigma_n} e^{-\frac{(f-f_n^*)^2}{2\Sigma_n^2}}$$



HOW MANY PATIENTS ARE NEEDED TO HAVE $f_n^* = 0.5$ *(typical f)* ??

Prob[$f > 0.5$] = 0.5



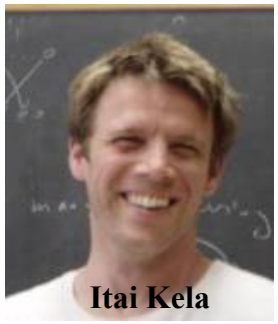
Van't Veer needs 2200 training samples to get 50% typical overlap

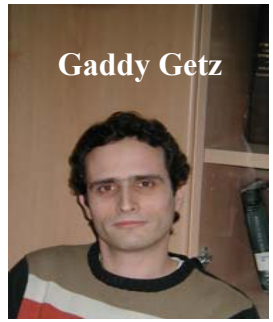
Results for Breast Cancer Data

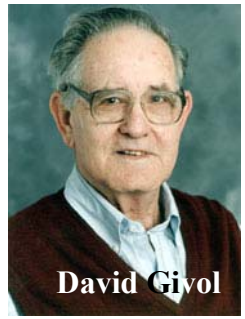
- For a typical overlap of 50% between two lists of 70 genes, more than 2300 patients are needed.

- The expected overlap between van't Veer's list and another list produced from similar experiment is less than 2%

Ein-Dor et al ***PNAS 2006***







Bioinformatics 2005

PNAS 2006

Funding and support:

NIH, EC/RTN, EC/6FW, ISF, GIF, Bikura, Ridgefield, Minerva,Levine, Wolfson Foundations, IMOS,