Gene Expression Analysis – Candidate Gene Approach CANB 7640

Aik Choon Tan, Ph.D.
Associate Professor of Bioinformatics
Division of Medical Oncology
Department of Medicine
aikchoon.tan@ucdenver.edu
9/18/2018
http://tanlab.ucdenver.edu/labHomePage/teaching/CANB7640/

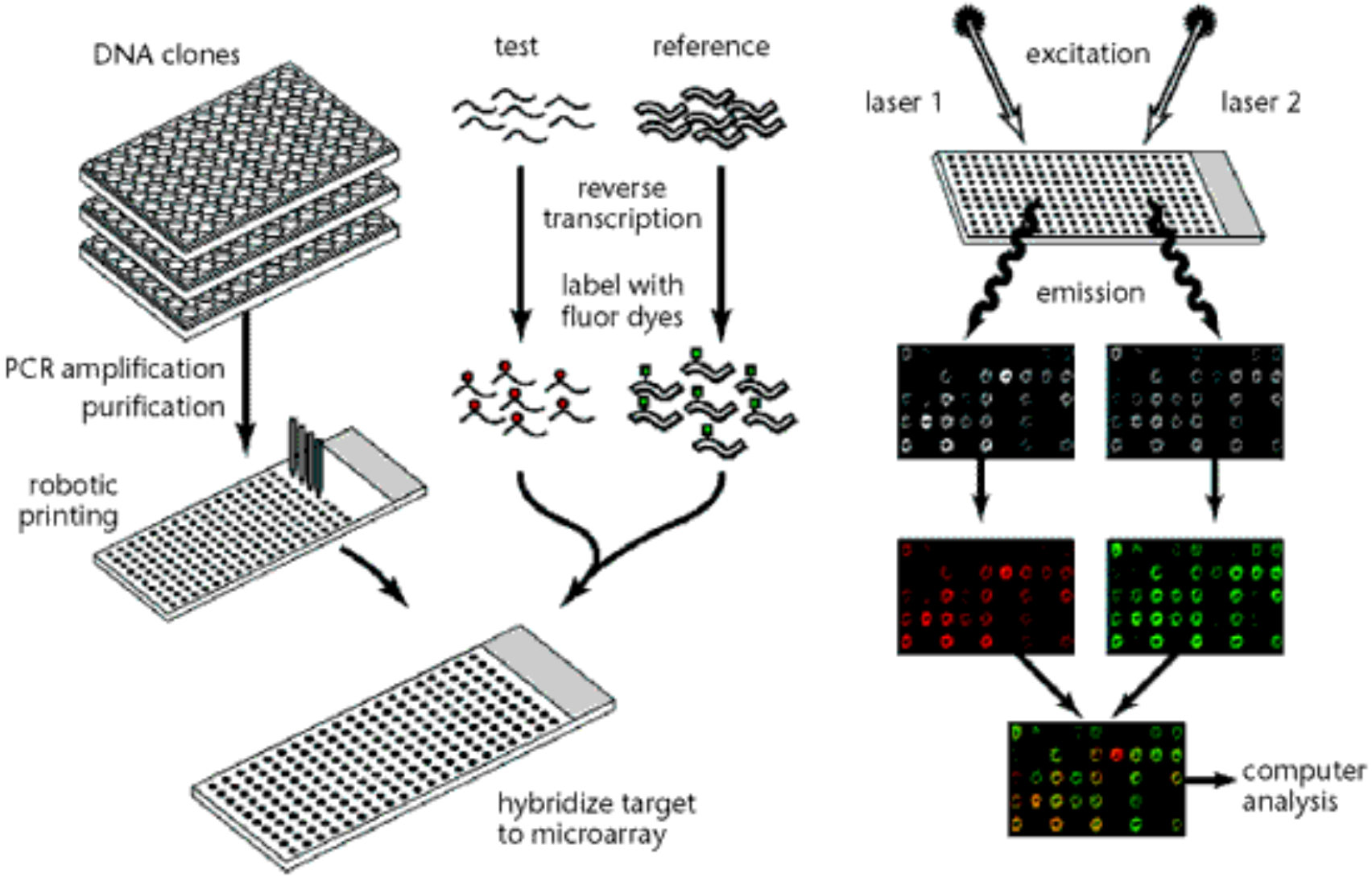
Outline

- Introduction to microarray technology
- Methods for Identify Differentially Expressed Genes
  - Signal-to-noise Ratio (SNR) (Golub et al 1999)
  - T-test
  - False Discovery Rate (FDR)
  - Significance Analysis of Microarrays (SAM) (Tusher et al 2001)
  - Methods Comparisons

Types of Microarray

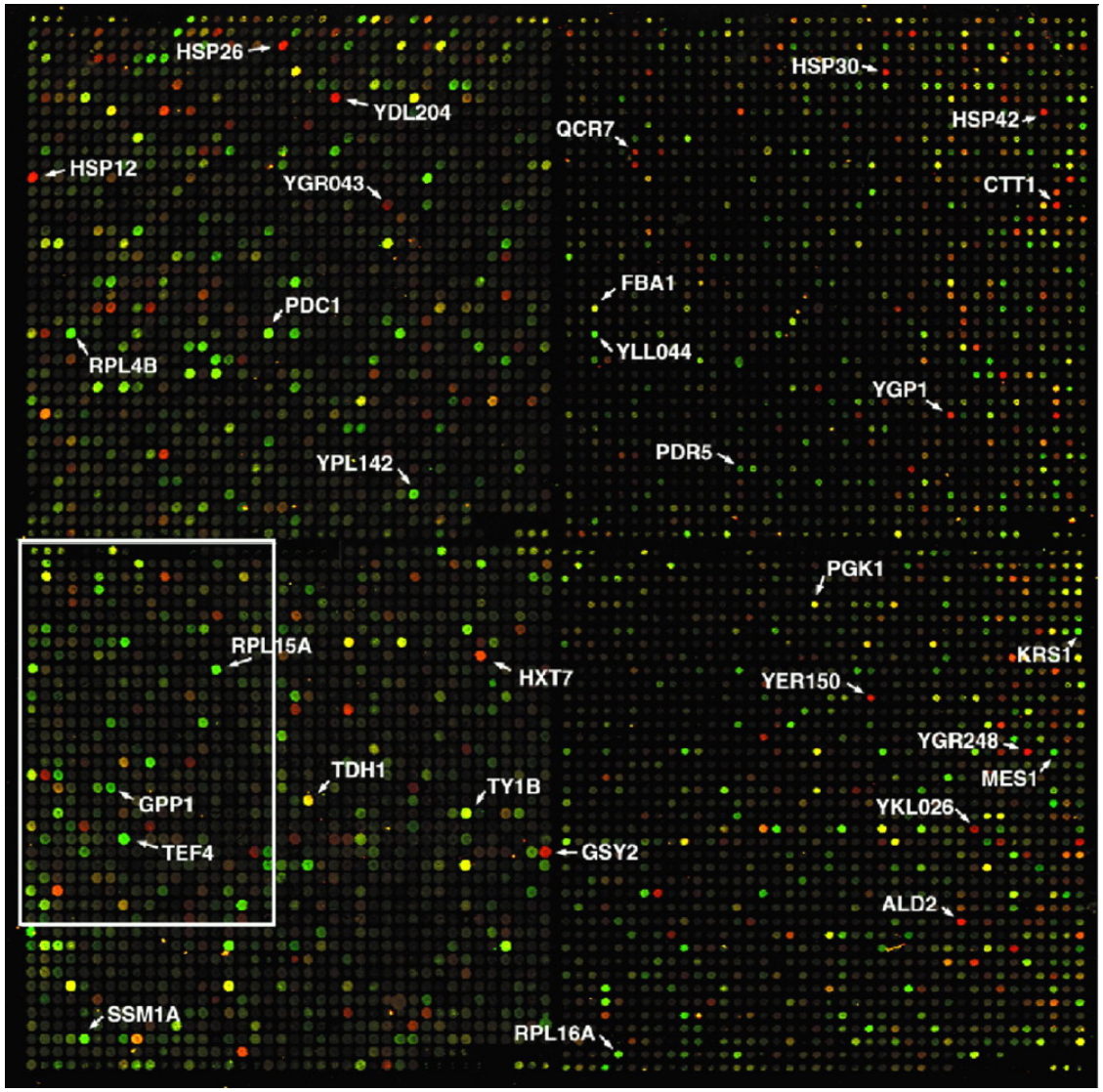
- Two-channel microarrays
  - Pre-synthesised cDNA arrays (Glass/Nylon/Plastic slides)
  - DIY
- Single channel microarrays
  - *In situ* synthesised oligonucleotide arrays (Chips)

cDNA microarray schema



From Duggan *et al. Nature Genetics* **21**, 10 – 14 (1999)

cDNA microarray of the Yeast genome



Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. (DeRisi, Iyer & Brown, Science, 268: 680-687, 1997)

GeneChip® Affymatrix

GeneChip® Single Feature

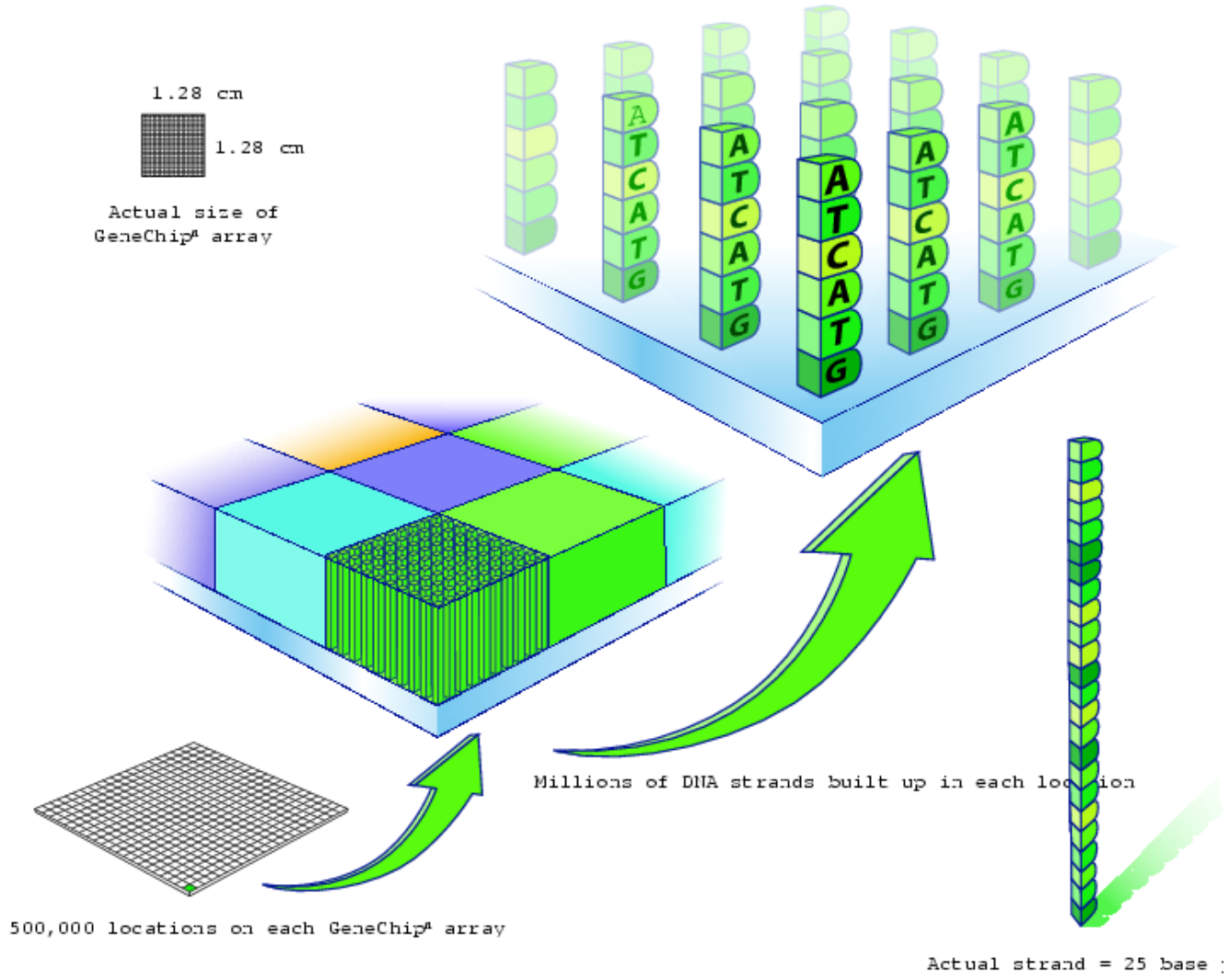


Image courtesy of Affymetrix.

GeneChip® Hybridization

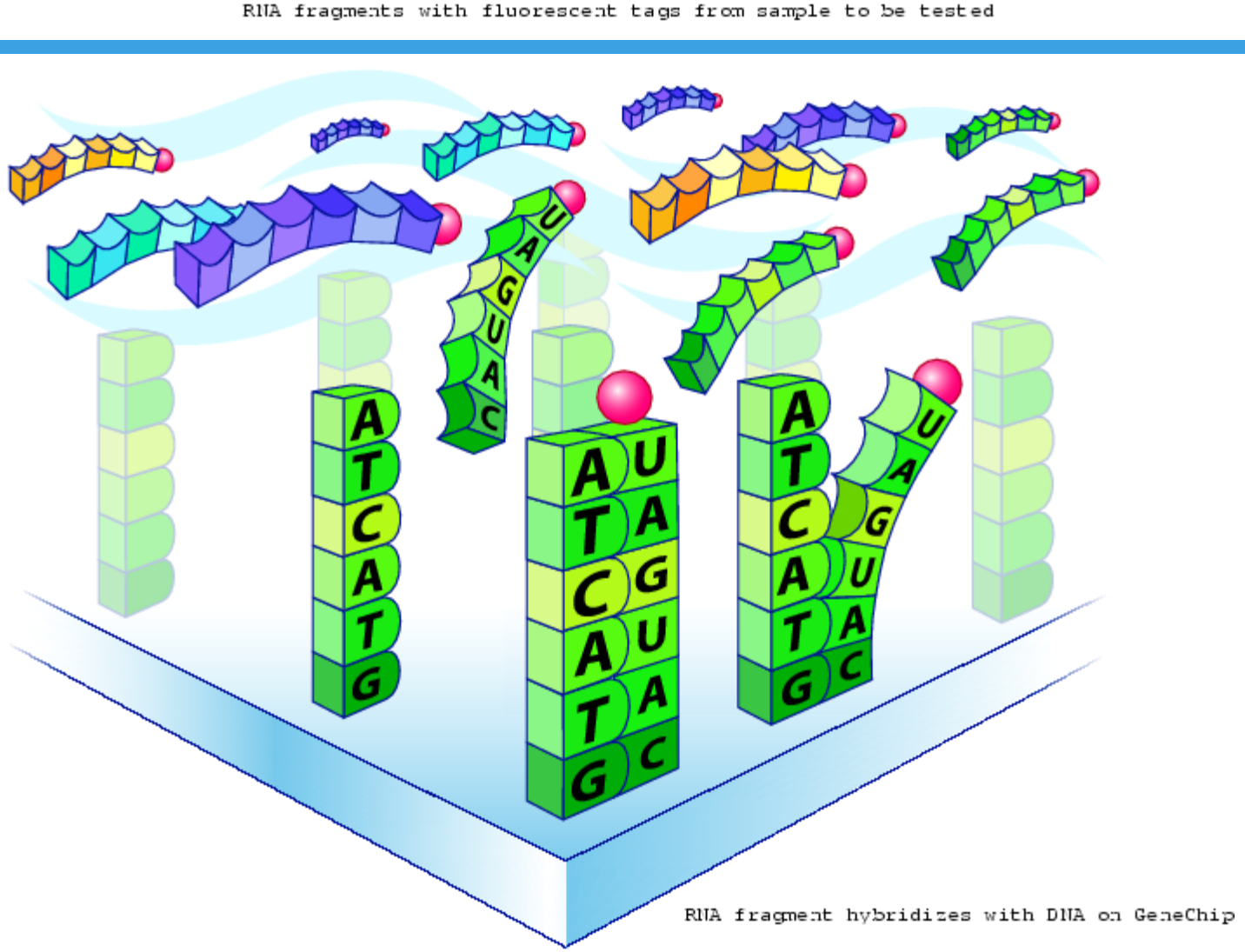


Image courtesy of Affymetrix.

Hybridized GeneChip® Microarray

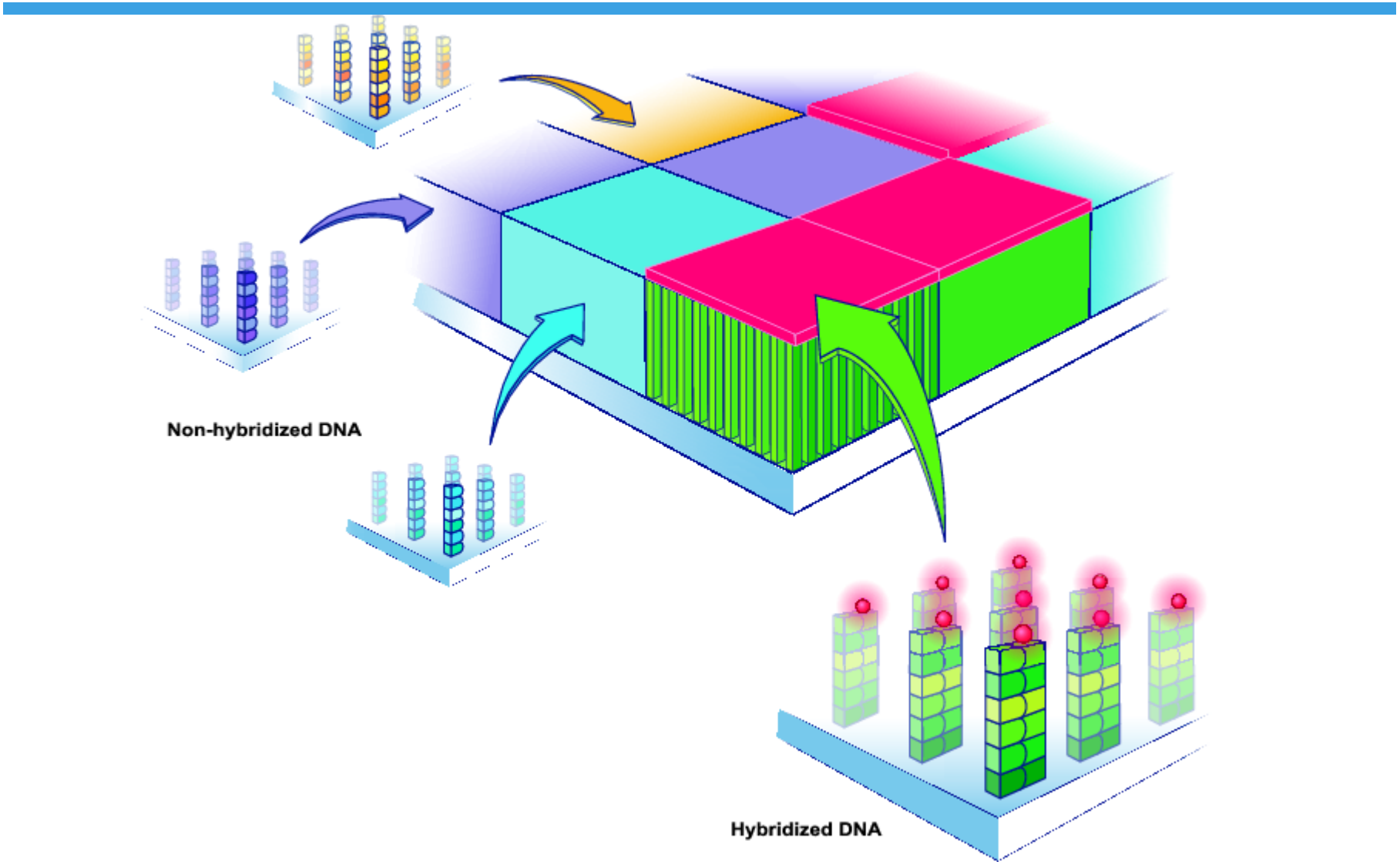
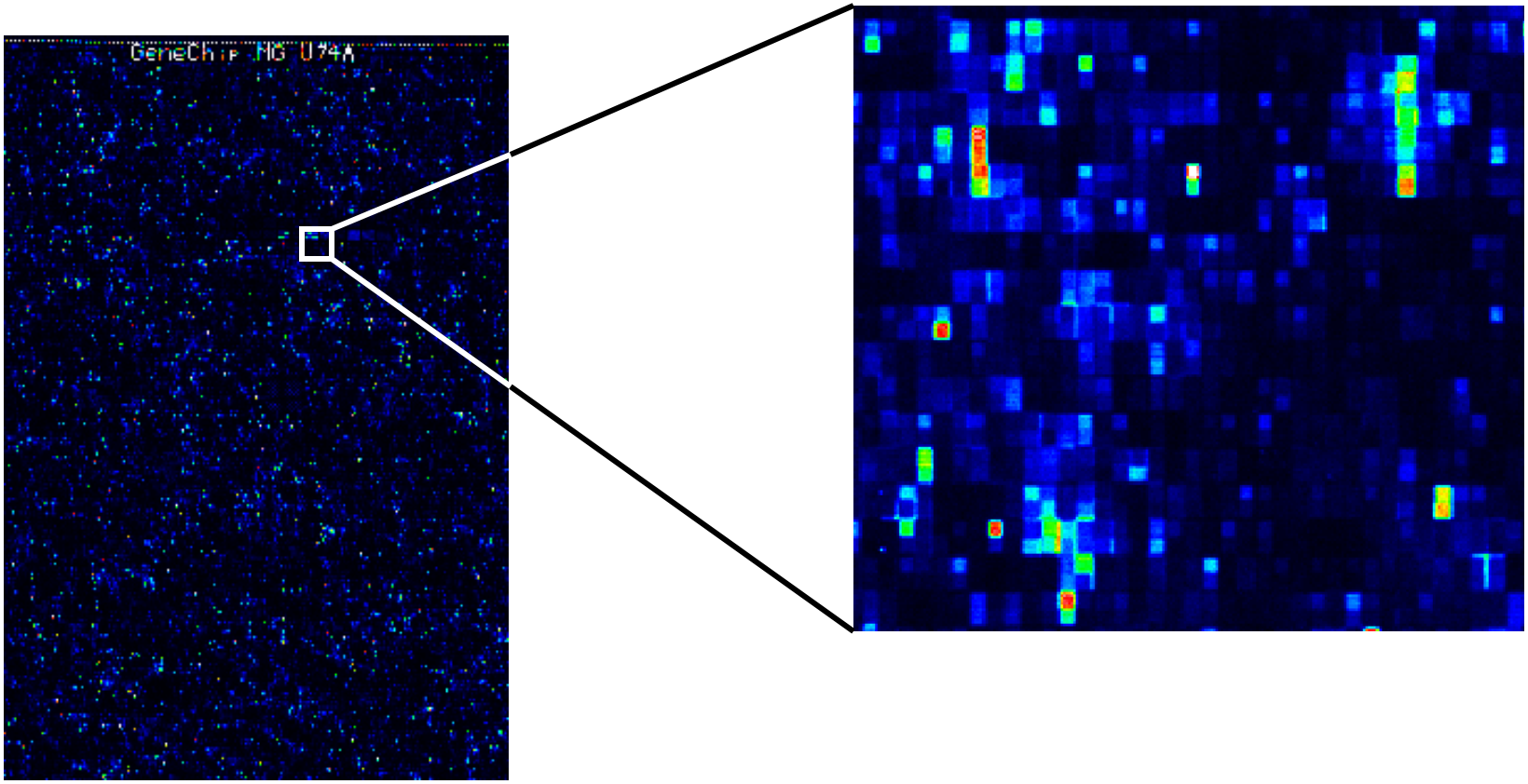


Image courtesy of Affymetrix

Spotting the arrays

RED = Present (P) = highly expressed, detected by the detector
YELLOW = Marginal (M) = expressed, “not sure” for the detector
GREEN = Absent (A) = maybe expressed, not detected by the detector



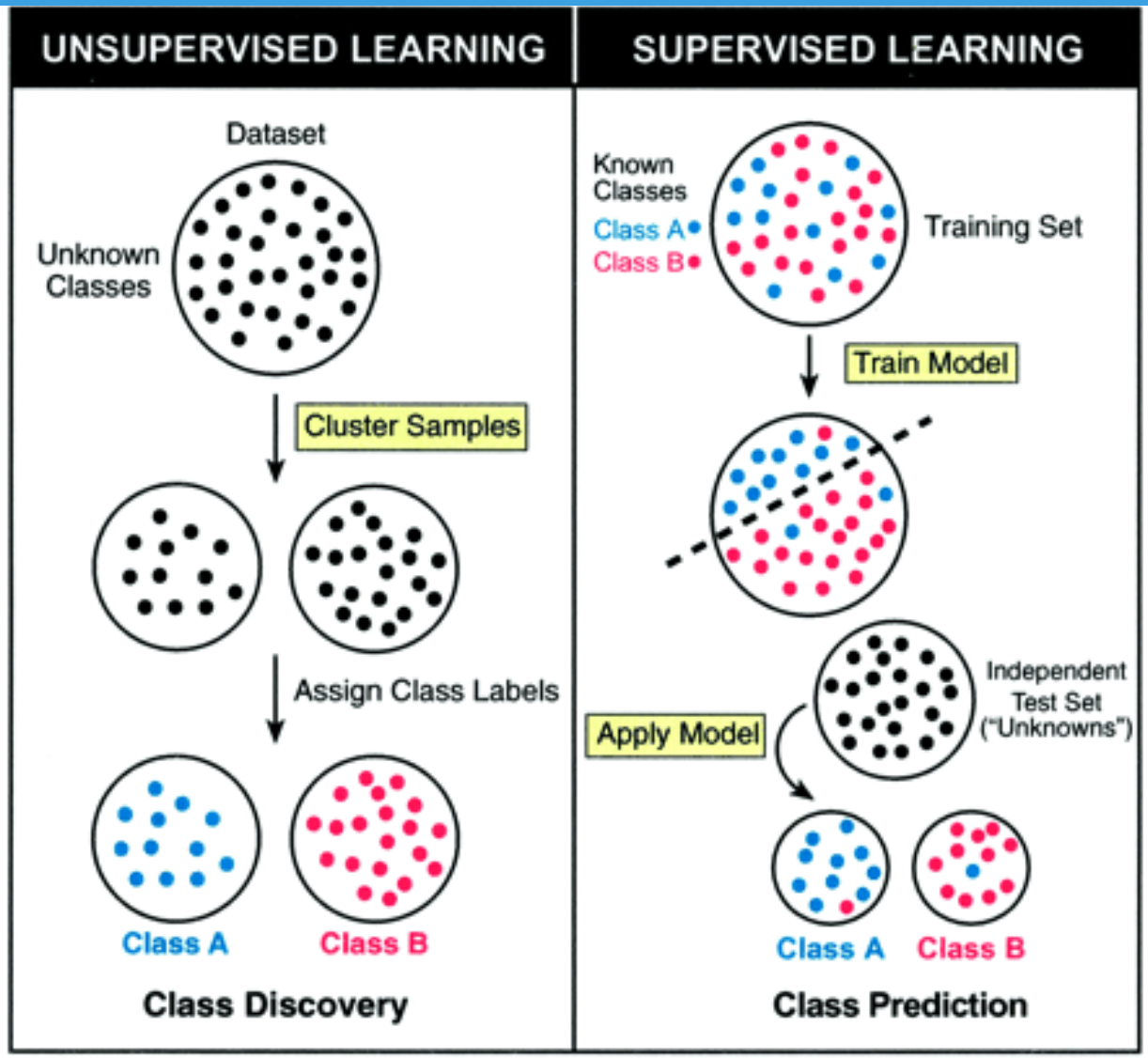
Gene Expression Profile

m samples

n genes

Geneid	Condition 1	Condition 2	…	Condition *m*
Gene1	103.02	58.79	…	101.54
Gene2	40.55	1246.87	…	1432.12
…	…	…	…	…
Gene *n*	78.13	66.25	…	823.09

Gene expression data analysis



Clustering (last week)

This week

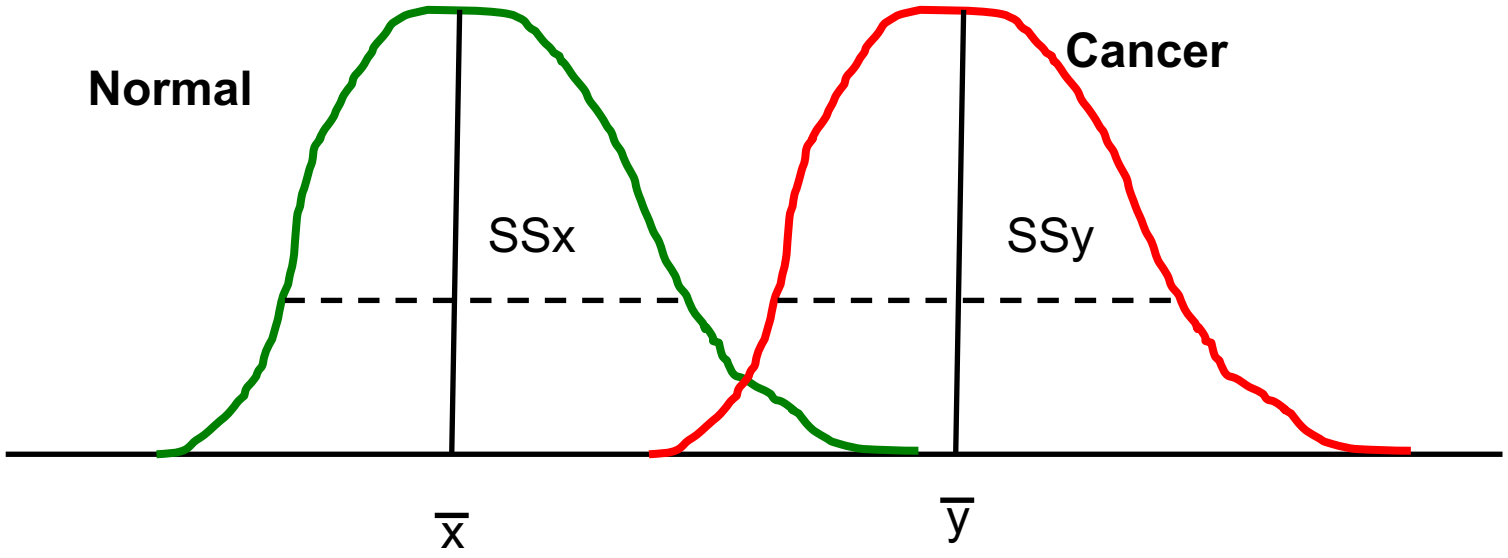
(Ramaswamy and Golub 2002)

Classification (Supervised Learning)

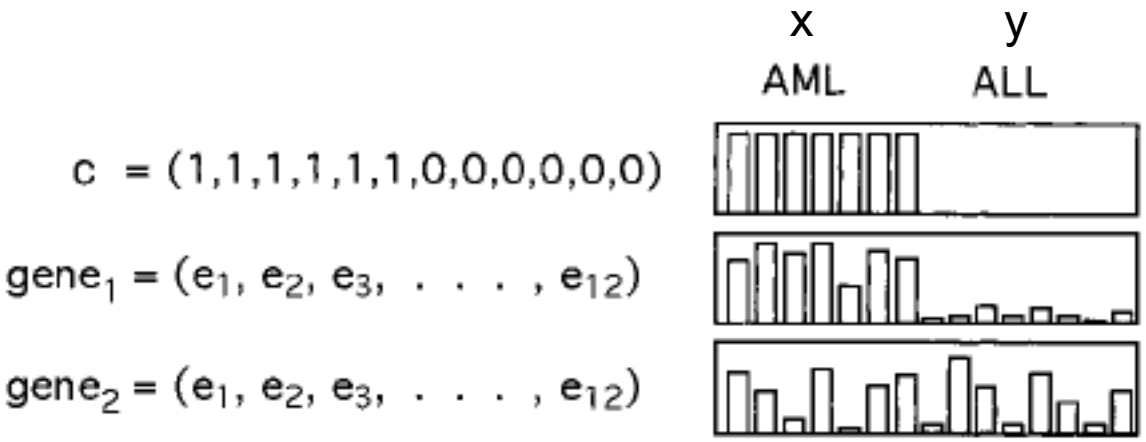
- Input: List of gene expressions and samples with known phenotypes (e.g. cancer vs. normal)
- Goal: Find gene markers (features) that distinguish one class from another class (e.g. cancer vs. normal).
- Selected gene markers will be used in the model for predicting new and unseen samples.

How to Identify Differentially Expressed Genes? (2-class problem)

Expression distribution of Gene *i*



Signal-to-Noise Ratio (SNR) (Golub et al 1999)



For every gene *i*

$$SNR = \frac{\bar{x} - \bar{y}}{(SS_x + SS_y)}$$

Positive SNR = correlates with Class x (e.g. AML)
Negative SNR = correlates with Class y (e.g. ALL)

T-test

For gene *i,* compute t-score

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}{\left(n_x + n_y - 2\right)}\left(SS_x + SS_y\right)}}$$

n_x = number of samples in x (e.g. Cancer)
n_y = number of samples in y (e.g. Normal)
$\bar{x}$ = mean gene expression of x samples
$\bar{y}$ = mean gene expression of y samples
SS_x = standard deviation of gene expression in x samples
SS_y = standard deviation of gene expression y samples

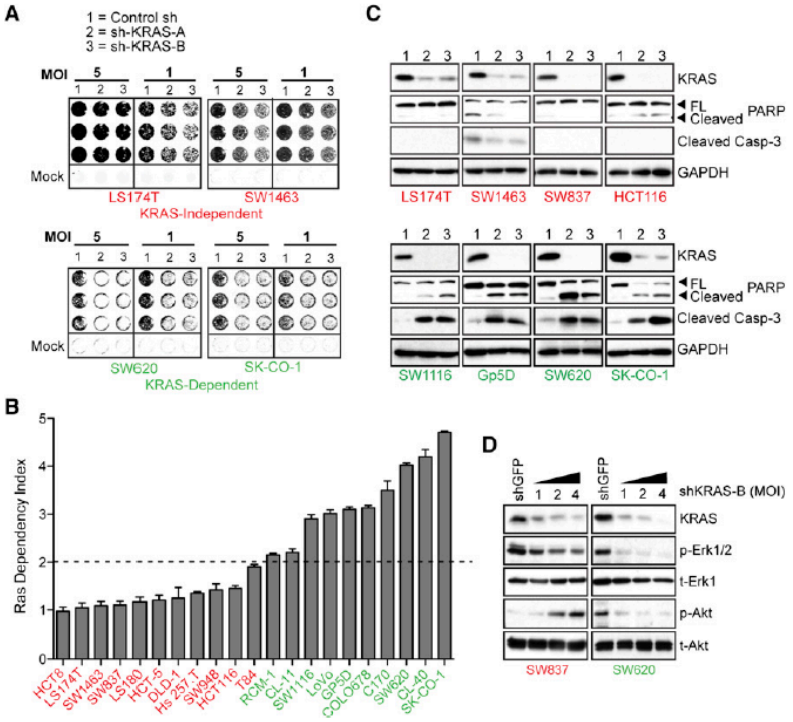
Example

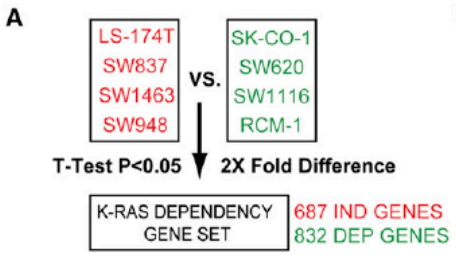
TAK1 Inhibition Promotes Apoptosis in KRAS-Dependent Colon Cancers

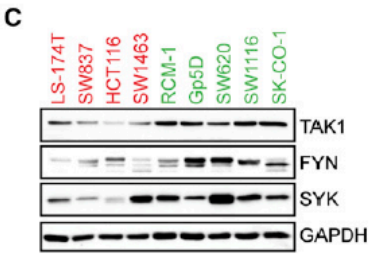
Anurag Singh,[1,3] Michael F. Sweeney,[1] Min Yu,[1] Alexa Burger,[1] Patricia Greninger,[1] Cyril Benes,[1] Daniel A. Haber,[1,2,*] and Jeff Settleman[1,4,*]

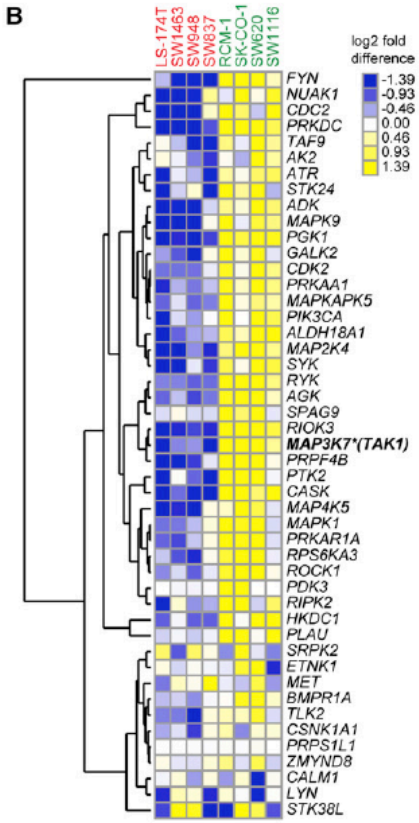
[1]Massachusetts General Hospital Cancer Center and Harvard Medical School, Charlestown, MA 02129, USA
[2]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[3]Present address: Department of Pharmacology and Experimental Therapeutics, Division of Medical Oncology and Hematology, Cancer Research Center, Boston University School of Medicine, Boston, MA 02118, USA
[4]Present address: Discovery Oncology, Genentech, Inc., South San Francisco, CA 94080, USA
*Correspondence: haber@helix.mgh.harvard.edu (D.A.H.), settleman.jeffrey@gene.com (J.S.)
DOI 10.1016/j.cell.2011.12.033

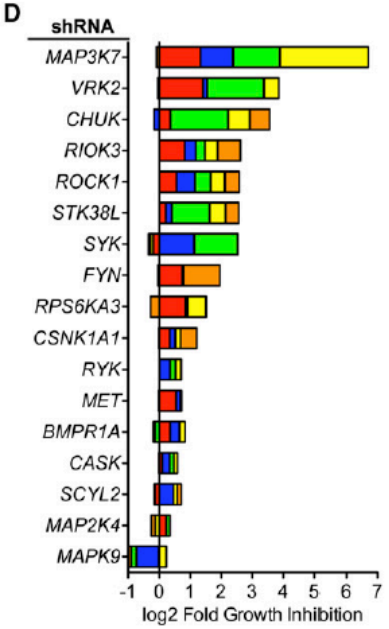
Cell (2012). 148, 639–650

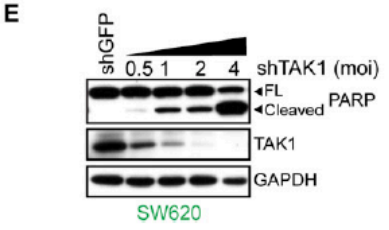






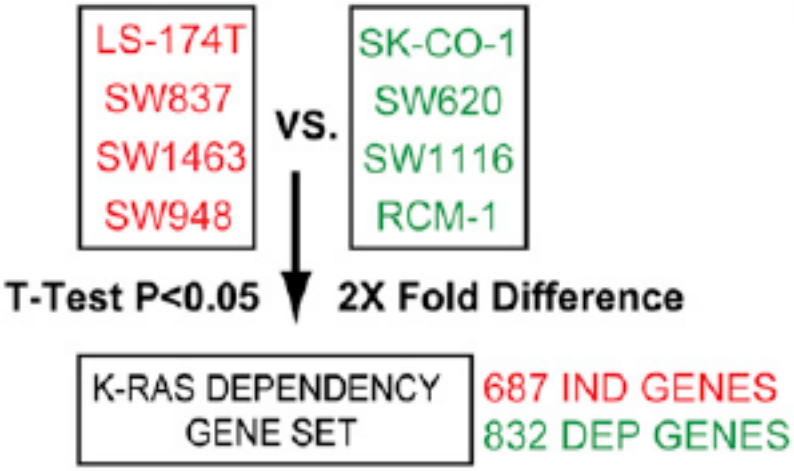






Druggable list

Is t-test an appropriate approach?



Cell (2012). 148, 639–650

687 + 832 = 1519 DEG
Assume 23,000 genes in human genome

1519/23000 = 0.066 (~6.6 % of the genes are differentially expressed, more than 5%, p= 0.05)

Is it by chance? False positive? False Discovery? How to control it?

False Discovery Rate (FDR)

- FDR control is a statistical method used in *multiple hypothesis testing* to correct for *multiple comparisons*.
- FDR procedures are designed to control the expected proportion of incorrectly rejected null hypotheses (“false discoveries”).
- Less stringent than family wise error rate (FWER) procedures (such as the Bonferroni correction).

Remember the 2x2 Contingency Table?

		Test Result	
		Positive	Negative
Actual	Positive	True Positive	False Positive (Type I Error)
	Negative	False Negative (Type II Error)	True Negative

FDR

- Instead of controlling type I error (false positive), FDR controls the expected proportion of false positives.
- FDR definition:
  - R is observable random variable.
  - V is non-observable random variable.
  - FDR is the expectation of random variable V/R

$$FDR = E(V/R)$$

		Test Result		**Total**
		Positive	Negative	
Actual	Positive	**U** (True Positive)	**V** (False Positive) (Type I Error)	$\mathbf{m_0}$
	Negative	**T** (False Negative) (Type II Error)	**R** (True Negative)	$\mathbf{m - m_0}$
Total		**m – R**	**R**	**m**

Significance Analysis of Microarrays (SAM)

Significance analysis of microarrays applied to the ionizing radiation response

Virginia Goss Tusher*, Robert Tibshirani†, and Gilbert Chu*‡

5116–5121 | PNAS | April 24, 2001 | vol. 98 | no. 9

- "assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements."
- SAM uses permutations of repeated measurements to estimate the False Discovery Rate

- Paper available online: http://www-stat.stanford.edu/~tibs/SAM/pnassam.pdf

- Today's practical and assignment

Overview of SAM

- Calculate “relative difference” – a value that incorporates the change in expression between conditions and the variation of measurements in each condition
- Calculate “expected relative difference” – derived from controls generated by permutations of data
- Plot against each other, set cutoff to identify deviating genes
- Calculate FDR for chosen cutoff from the control permutations

(Adapted from OHRI Bioinformatics Slides 2006)

Relative Difference

For gene *i*

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

(Adapted from OHRI Bioinformatics Slides 2006)

Relative Difference

For gene *i*

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

$\bar{x}_I(i), \bar{x}_U(i)$	Mean expression of gene *i* in condition I or U (e.g. Cancer vs Normal)

(Adapted from OHRI Bioinformatics Slides 2006)

Relative Difference

For gene *i*

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

$s(i)$	Gene-specific scatter

(Adapted from OHRI Bioinformatics Slides 2006)

Relative Difference

For gene *i*

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

s_0	Small positive constant calculated to minimize coefficient of variation.

(Adapted from OHRI Bioinformatics Slides 2006)

T-test vs. SAM

(1) $$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_p^{\,2}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

(2) $$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{df_x(s_x^{\,2}) + df_y(s_y^{\,2})}{df_x + df_y}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

(3) $$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{df_x\left(\frac{\sum_i (x_i - \bar{x})^2}{df_x}\right) + df_y\left(\frac{\sum_i (y_i - \bar{y})^2}{df_y}\right)}{df_x + df_y}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

(4) $$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(SS_x + SS_y)}{df_x + df_y}\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$

(5) $$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}{(n_x + n_y - 2)}(SS_x + SS_y)}}$$

(6) $$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

(7) $$s(i) = \sqrt{a\left\{\sum_m [x_m(i) - \bar{x}_I(i)]^2 + \sum_n [x_n(i) - \bar{x}_U(i)]^2\right\}}$$

(8) $$s(i) = \sqrt{a(SS_I + SS_U)}$$

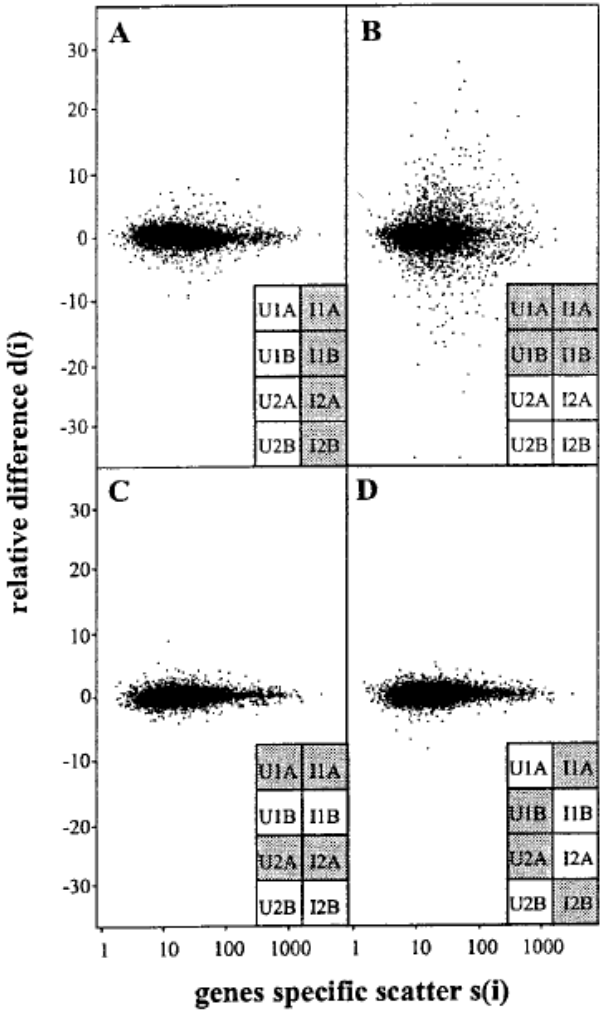
$$a = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{(n_1 + n_2 - 2)}$$

(9) $$s(i) = \sqrt{\frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{(n_1 + n_2 - 2)}(SS_I + SS_U)}$$

(10) $$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{\sqrt{\frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{(n_1 + n_2 - 2)}(SS_I + SS_U)} + s_0}$$

(Adapted from OHRI Bioinformatics Slides 2006)

Relative difference vs. Gene scatter



- Plotting *d(i)* vs *s(i)*

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

- Comparing 4 shaded vs 4 non-shaded samples
- A: Relative differences between irradiated and unirradiated states
- B: Relative differences between cell lines
- C: Relative differences between hybridizations (technical replicates)
- D: Relative differences between 'balanced' permutation (Extra control)

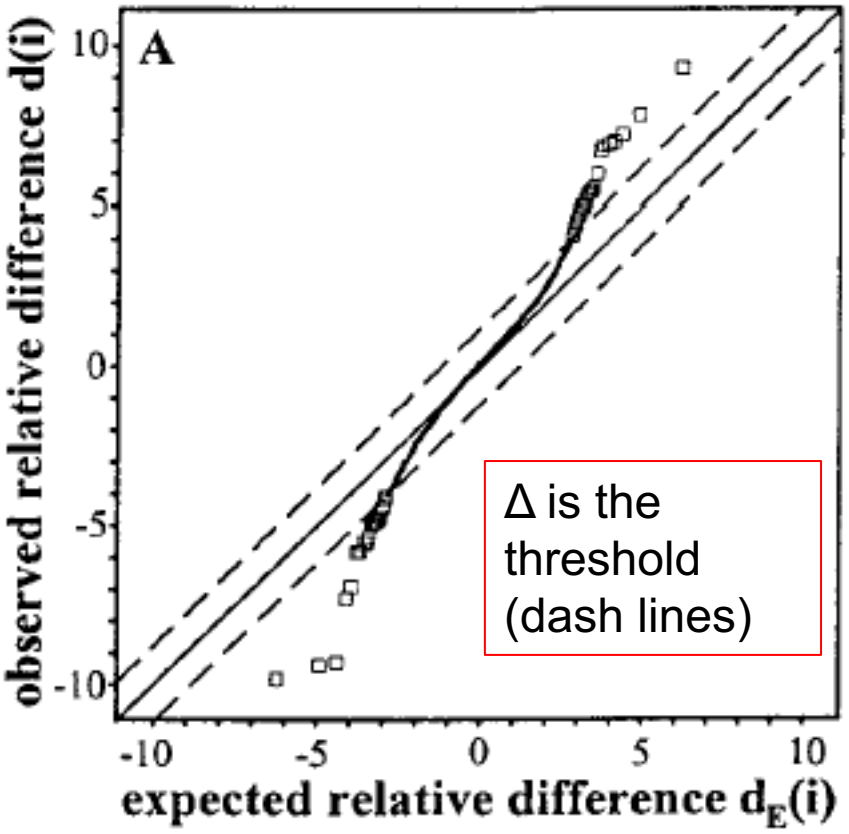
(Adapted from OHRI Bioinformatics Slides 2006)

SAM creates controls via permutation

- Consider permutations of the samples used.
- Calculate $d_p(i)$ for each permutation p
- Average all $d_p(i)$ to get ‘expected relative difference’ : $d_E(i)$

(Adapted from OHRI Bioinformatics Slides 2006)

Finding significant genes

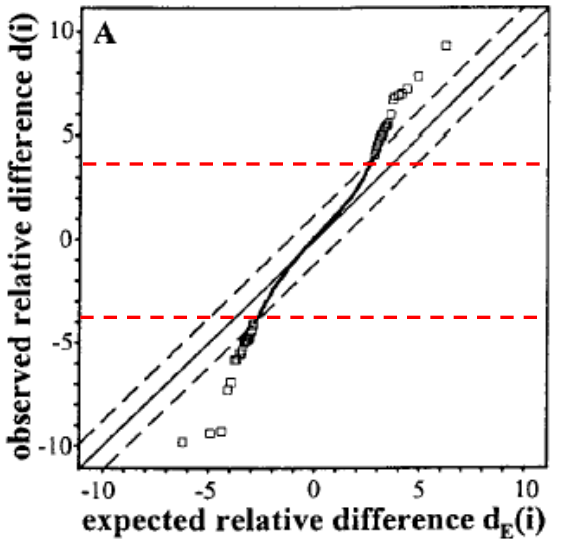


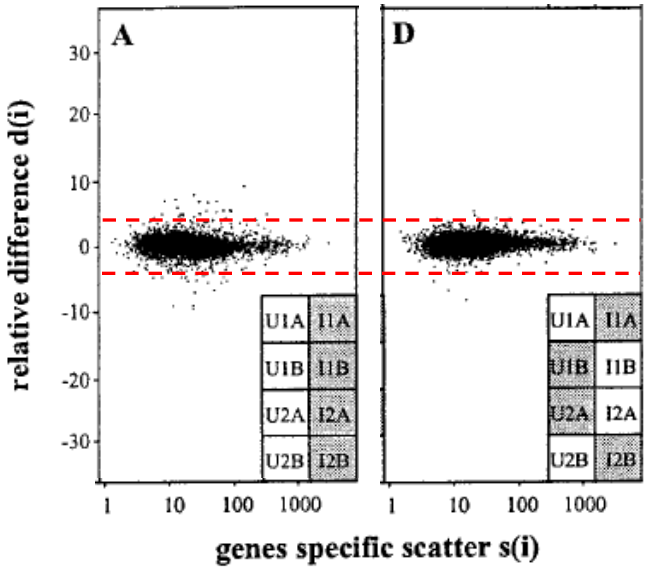
- Plot $d(i)$ vs $d_E(i)$
- Identify genes which deviate from $d(i)=d_E(i)$ by more than a threshold, Δ
- These do not necessarily have the largest change in expression.
- Can optimize Δ with estimate of false positive rate

(Adapted from OHRI Bioinformatics Slides 2006)

False Discovery Rate

- Take observed *d(i)* values for upper and lower cutoffs
- Find the mean number of genes exceeding these cutoffs in the permuted data - this gives an estimate for FDR

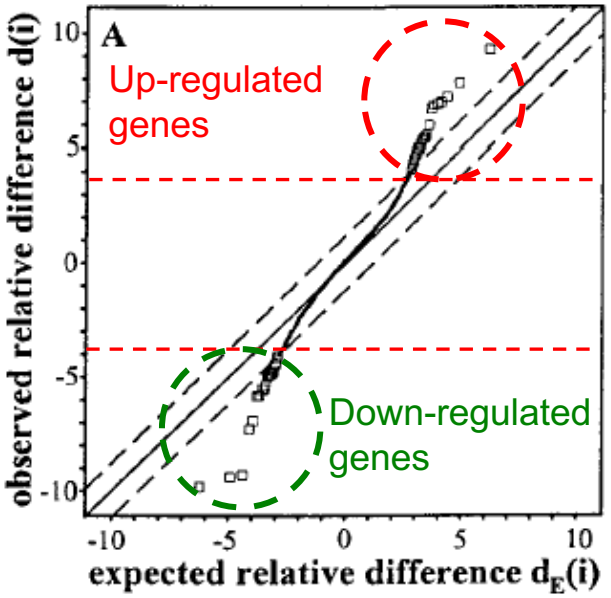




(Adapted from OHRI Bioinformatics Slides 2006)

SAM Output

- List of significantly changing genes
  - Fold changes may be asymmetric
- Estimated false positive rate for the list



(Adapted from OHRI Bioinformatics Slides 2006)

Which method is better?

OPEN ACCESS Freely available online

PLoS one

Should We Abandon the *t*-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies

Marine Jeanmougin[1,2,3,4]*, Aurelien de Reynies[1], Laetitia Marisa[1], Caroline Paccard[2], Gregory Nuel[3], Mickael Guedj[1,2]

1 Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre le Cancer, Paris, France, **2** Department of Biostatistics, Pharnext, Paris, France, **3** Department of Applied Mathematics (MAP5) UMR CNRS 8145, Paris Descartes University, Paris, France, **4** Statistics and Genome Laboratory UMR CNRS 8071, University of Evry, Evry, France

Abstract

High-throughput post-genomic studies are now routinely and promisingly investigated in biological and biomedical research. The main statistical approach to select genes differentially expressed between two groups is to apply a *t*-test, which is subject of criticism in the literature. Numerous alternatives have been developed based on different and innovative variance modeling strategies. However, a critical issue is that selecting a different test usually leads to a different gene list. In this context and given the current tendency to apply the *t*-test, identifying the most efficient approach in practice remains crucial. To provide elements to answer, we conduct a comparison of eight tests representative of variance modeling strategies in gene expression data: Welch's *t*-test, ANOVA [1], Wilcoxon's test, SAM [2], RVM [3], limma [4], VarMixt [5] and SMVar [6]. Our comparison process relies on four steps (gene list analysis, simulations, spike-in data and re-sampling) to formulate comprehensive and robust conclusions about test performance, in terms of statistical power, false-positive rate, execution time and ease of use. Our results raise concerns about the ability of some methods to control the expected number of false positives at a desirable level. Besides, two tests (limma and VarMixt) show significant improvement compared to the *t*-test, in particular to deal with small sample sizes. In addition limma presents several practical advantages, so we advocate its application to analyze gene expression data.

[33 citations]

Methods

- **Welch t-test**
- **ANOVA**
- **Wilcoxon's test**
- **SAM** (Tusher et al 2001) – Significance Analysis of Microarrays (www-stat.stanford.edu/~tibs/SAM/) [8954 citations]
- **RVM** (Wright & Simon, 2003) – Random Variance Model [BRB-ArrayToolshttp://linus.nci.nih.gov/BRB-ArrayTools.html).] [393 citations]
- **Limma** (Smyth 2005) – Linear Models for Microarray Data (available as limma package in R/Bioconductor) [1883 citations]
- **VarMixt** (Delmar, Robin, Daudin 2004) – Variance Mixture model (formerly available as varmixt package in R/Bioconductor) [89 citations]
- **SMVar** (Jaffrézic et al 2007) – Structural Mixed Model for Variance(available as SMVar package in R/Bioconductor) [28 citations]

Data Sets and Testing Methods

Table 1. Data sets used for the gene list analysis.

Data-set	Groups	Sample size	Publication
Lymphoid tumors	Disease staging	37	Lamant et al. 2007 [26]
Liver tumors	TP53 mutation	65	Boyault et al. 2007 [27]
Head and neck tumors	Gender	81	Rickman et al. 2008 [28]
Leukemia	Gender	104	Soulier et al. 2006 [29]
Breast tumors	ESR1 expression	500	Bertheau et al. 2007 [30]

The five data sets come from the *Cartes d'Identité des Tumeurs* (CIT, http://cit.ligue-cancer.net) program and are publicly available. All the microarrays are Affymetrix U133A microarrays with 22,283 genes.
doi:10.1371/journal.pone.0012336.t001

Testing Methods:

1. Gene selection – compare the common genes ($p < 0.05$)
2. Simulation – test for False Discovery Rate (FDR)
3. Spike-in data – test for true fold-change and FDR
4. Re-sampling – test for small samples

Gene List Analysis

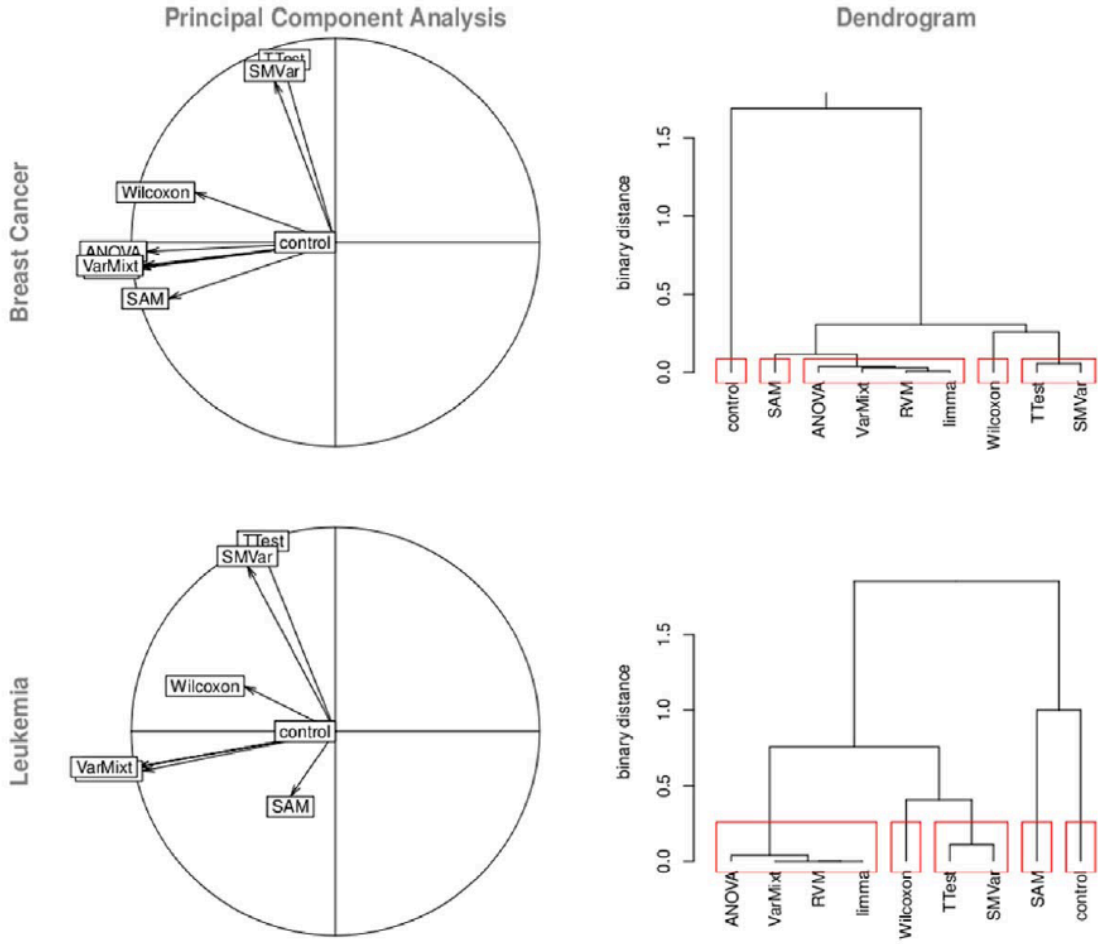


Figure 2. Gene list analysis. PCAs and dendrograms are generated based on the gene lists resulting from the application of the eight tests of interest and the control-test. Here we show results for two data sets comparing ESR1 expression in breast cancer and gender in leukemia. Both outline five clusters of tests.

doi:10.1371/journal.pone.0012336.g002

False Positive Rate

Table 2. False-positive rate study from simulations.

	M1		M2		M3		M4	
Sample size	**n=5**	**n=100**	**n=5**	**n=100**	**n=5**	**n=100**	**n=5**	**n=100**
t-test▼	3.8–4.6	4.5–5.4	4.0–4.8	4.6–5.5	3.8–4.6	4.7–5.6	3.9–4.7	4.4–5.3
ANOVA	4.5–5.2	4.5–5.4	4.7–5.6	4.6–5.5	4.5–5.4	4.7–5.6	4.5–5.3	4.4–5.3
Wilcoxon▼	2.8–3.5	4.6–5.5	2.6–3.3	4.5–5.4	2.8–3.5	4.7–5.6	2.7–3.4	4.5–5.4
SAM	4.6–5.5	4.5–5.3	4.2–5.1	4.5–5.4	4.7–5.6	4.7–5.6	4.3–5.2	4.4–5.3
RVM▲	5.7–6.7	4.5–5.4	5.6–6.5	4.5–5.4	5.4–6.3	4.7–5.6	5.3–6.2	4.7–5.5
limma	4.6–5.5	4.6–5.5	4.2–5.1	4.5–5.4	4.7–5.6	4.7–5.6	4.4–5.3	4.3–5.1
SMVar▲	7.0–8.1	4.7–5.6	–	–	5.9–6.8	4.8–5.7	4.6–5.5	4.5–5.3
VarMixt	4.7–5.5	4.6–5.5	4.3–5.2	4.6–5.5	4.8–5.6	4.6–5.5	4.5–5.4	4.5–5.3

For small and large samples, this table presents the 95% confidence-interval of false-positive rate obtained by applying a threshold of 0.05 to the *p*-values. Up triangles ▲ (resp. down triangles ▼) indicate an increase (resp. a decrease) of the false-positive rate compared to the expected level of 5%. Two triangles inform of a deviation in both small and large sample sizes.

doi:10.1371/journal.pone.0012336.t002

Summary

Table 3. Summary table.

	False-positive rate		Power		In practice	
	Small samples	Large samples	Small samples	Large samples	Ease of use	Execution time
t-test	+	+++	+	+++	+++	+++
ANOVA	+++	+++	+	+++	+++	+++
Wilcoxon	+	+	+	++	+++	++
SAM	+++	+++	+	++	++	++
RVM	+	++	+++	+++	++	+
limma	+++	+++	+++	+++	++	+++
VarMixt	+++	+++	+++	+++	+	+
SMVar	+	+	++	+++	++	+++

This table summarizes the results of our study in terms of false-positive rate, power and practical criteria. The number of "+" indicates the performance, from weak (+), to very good one (+++).

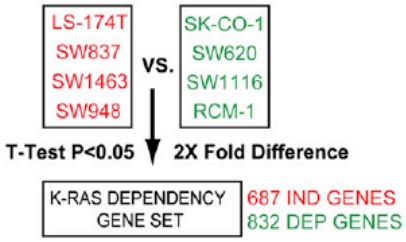
doi:10.1371/journal.pone.0012336.t003

Biology trumps statistics - Example

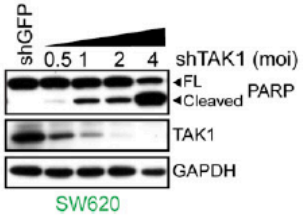
TAK1 Inhibition Promotes Apoptosis in KRAS-Dependent Colon Cancers

Anurag Singh,[1,3] Michael F. Sweeney,[1] Min Yu,[1] Alexa Burger,[1] Patricia Greninger,[1] Cyril Benes,[1] Daniel A. Haber,[1,2,*] and Jeff Settleman[1,4,*]
[1]Massachusetts General Hospital Cancer Center and Harvard Medical School, Charlestown, MA 02129, USA
[2]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[3]Present address: Department of Pharmacology and Experimental Therapeutics, Division of Medical Oncology and Hematology, Cancer Research Center, Boston University School of Medicine, Boston, MA 02118, USA
[4]Present address: Discovery Oncology, Genentech, Inc., South San Francisco, CA 94080, USA
*Correspondence: haber@helix.mgh.harvard.edu (D.A.H.), settleman.jeffrey@gene.com (J.S.)
DOI 10.1016/j.cell.2011.12.033

Cell (2012). 148, 639–650



A lot of experiments (validation)



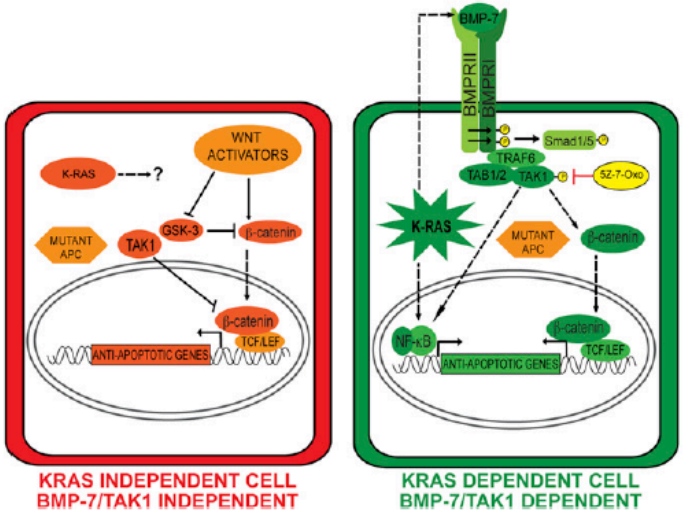


Figure 7. A Model for Context-Specific KRAS Dependency in Colon Cancers

In KRAS-independent colon cancers, APC loss of function results in hyper-activation of canonical Wnt signaling through stabilization of β-catenin in cooperation with upstream Wnt activators. TAK1 can be a negative regulator of canonical Wnt signaling in these cells. In KRAS-dependent cells, oncogenic KRAS upregulates BMP-7 expression/secretion, activating the BMP receptor and resulting in TAK1 activation. KRAS and TAK1 in these cells are activators of Wnt signaling by promoting β-catenin nuclear localization, which is stabilized by virtue of *APC* loss-of-function mutations. KRAS-mediated anti-apoptotic signaling could also be facilitated by NF-κB activation. Dashed lines represent unknown molecular interactions.
See also Figure S6.

Take home message

- Consider FDR in selecting differentially expressed genes
- Compare with multiple methods
- Overlapping genes identified from different methods enhance the real signals
- *Biology trumps statistics* – if you can validate the genes