Case Studies on High Throughput Gene Expression Data

Kun Huang, PhD
Raghu Machiraju, PhD

Department of Biomedical Informatics
Department of Computer Science and Engineering
The Ohio State University

Wexner Medical Center

Review

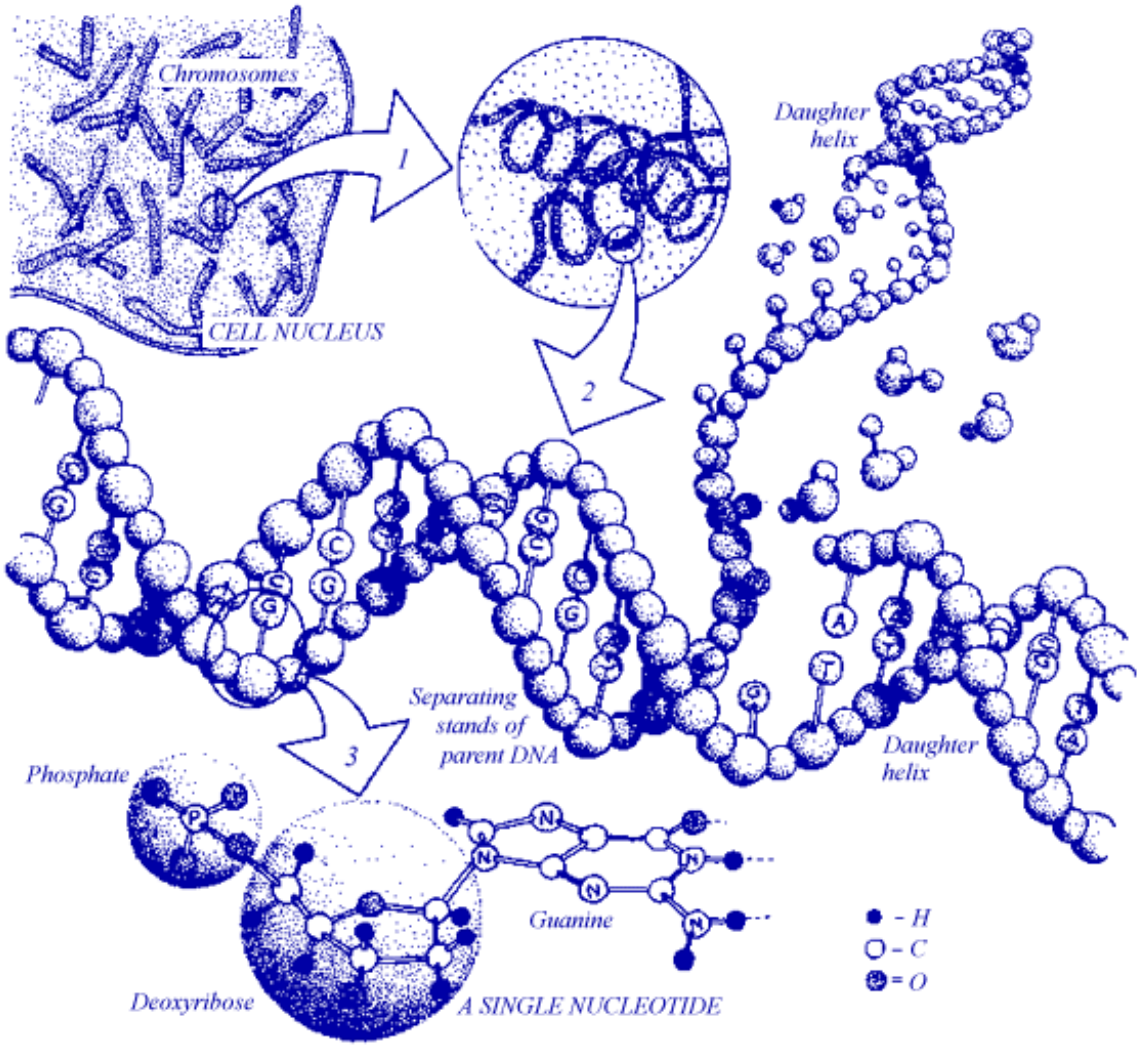


Figure by Lawrence Berkeley Lab Human Genome Center, Berkeley, California, USA

Wexner Medical Center

Synopsis -
http://www.jsonline.com/news/health/1122482

Causes -
http://media.journalinteractive.com/documents/dna2gr.pdf

JSONLINE.COM
MILWAUKEE · WISCONSIN
JOURNAL SENTINEL
PULITZER PRIZE WINNER 2008 · 2010 · 2011

Wexner Medical Center

http://cbm.msoe.edu/scienceOlympiad/module2012/xiapHome.html

Glossary - http://www.jsonline.com/news/health/105196904.html

Reading Assignment

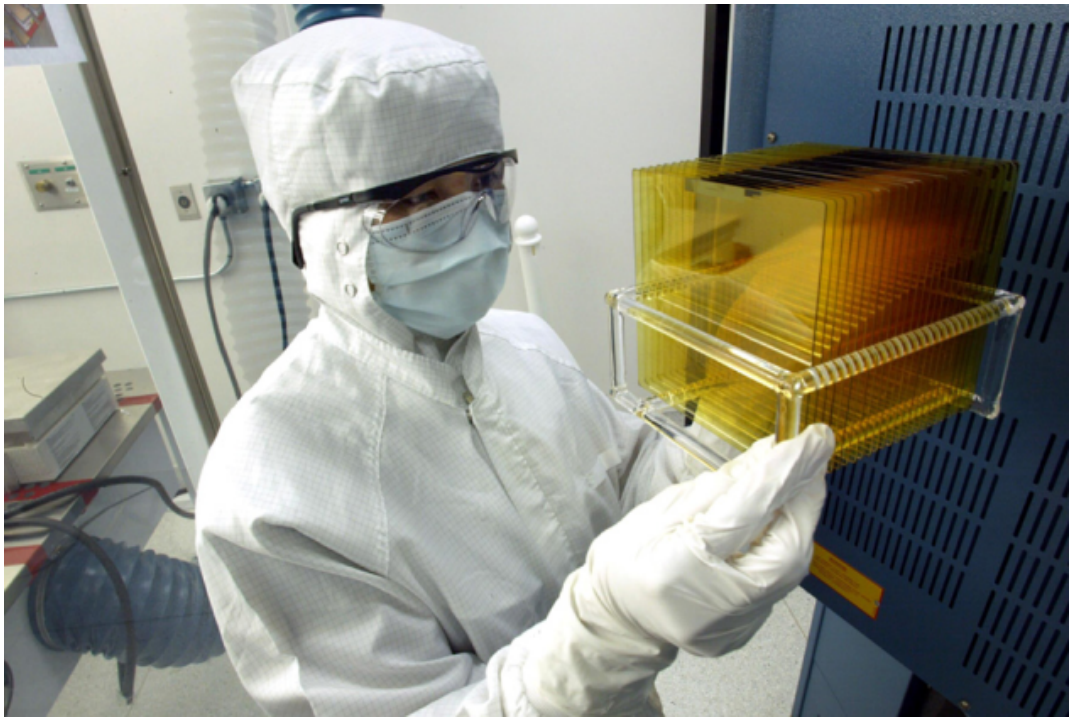
Chapter 2, Text

Outline

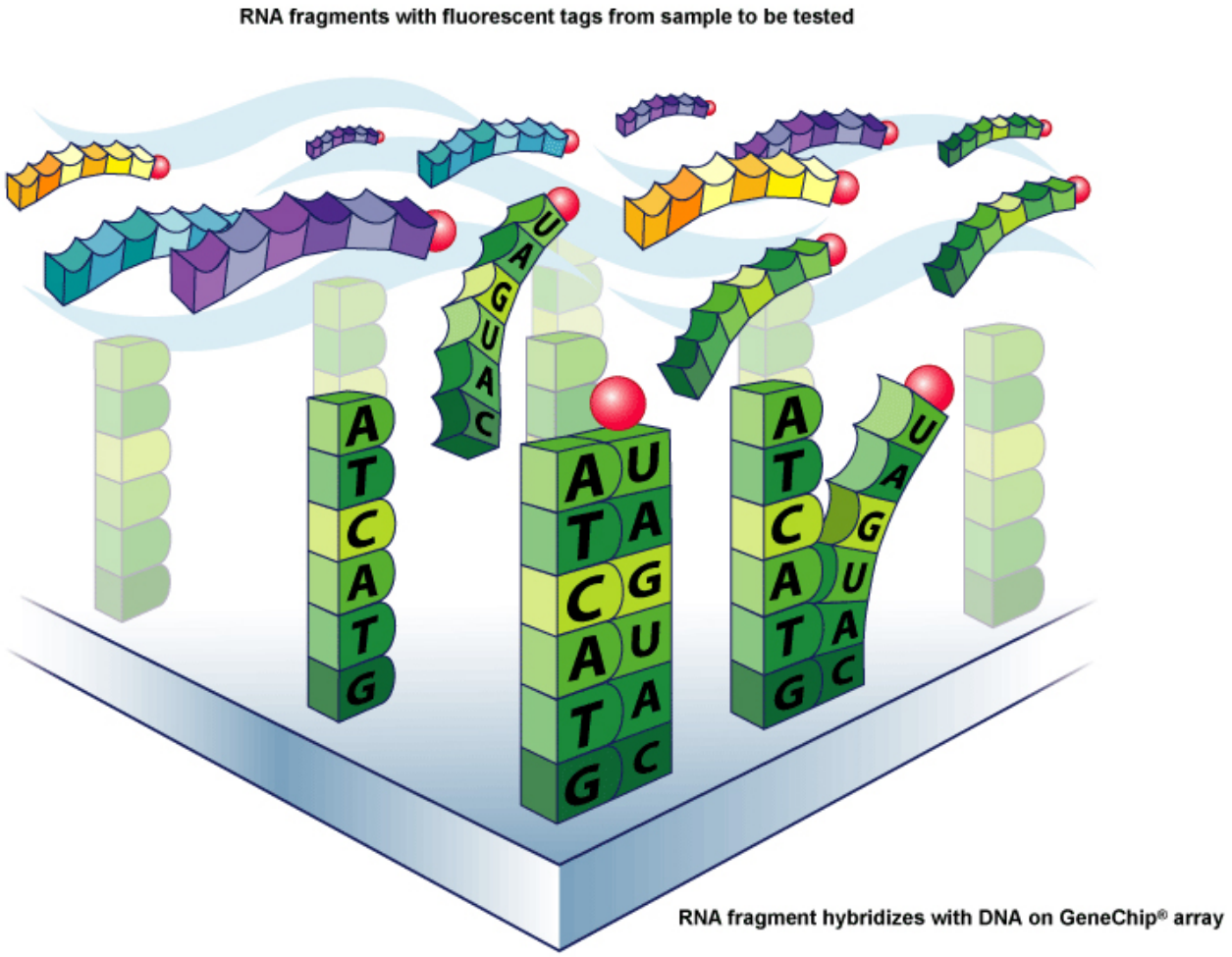
- What can we do with high throughput gene expression data?
- Profiling and comparative studies – gene signatures
- Machine learning approaches – biomarker discovery
- Correlation analysis – gene function prediction using co-expression
- Network analysis – network biomarkers
- Genomic variants – the drivers (?)

Affymetrix GeneChip

- silicon chip
- oligonucleiotide probes lithographically synthesized on the array

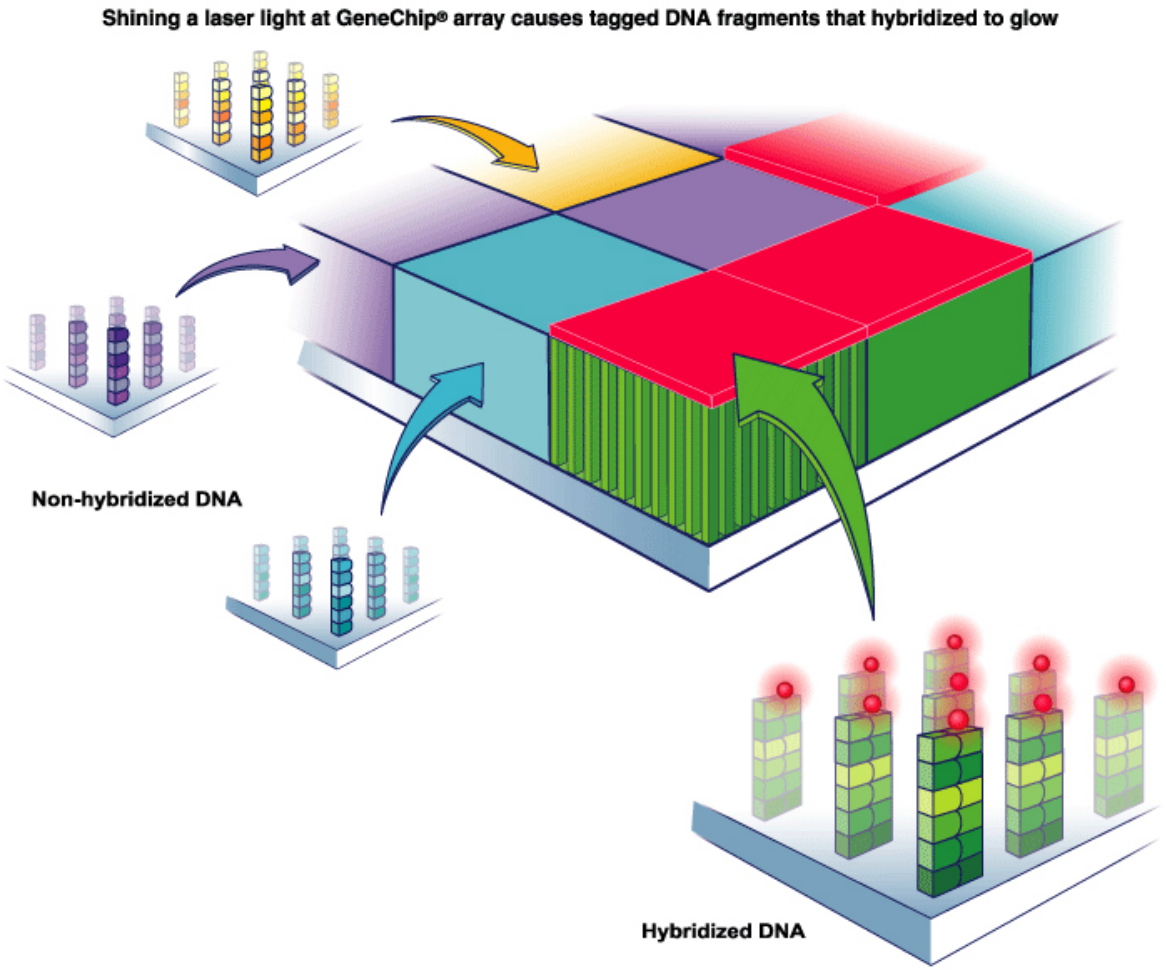


How does microarray work?



Wexner Medical Center

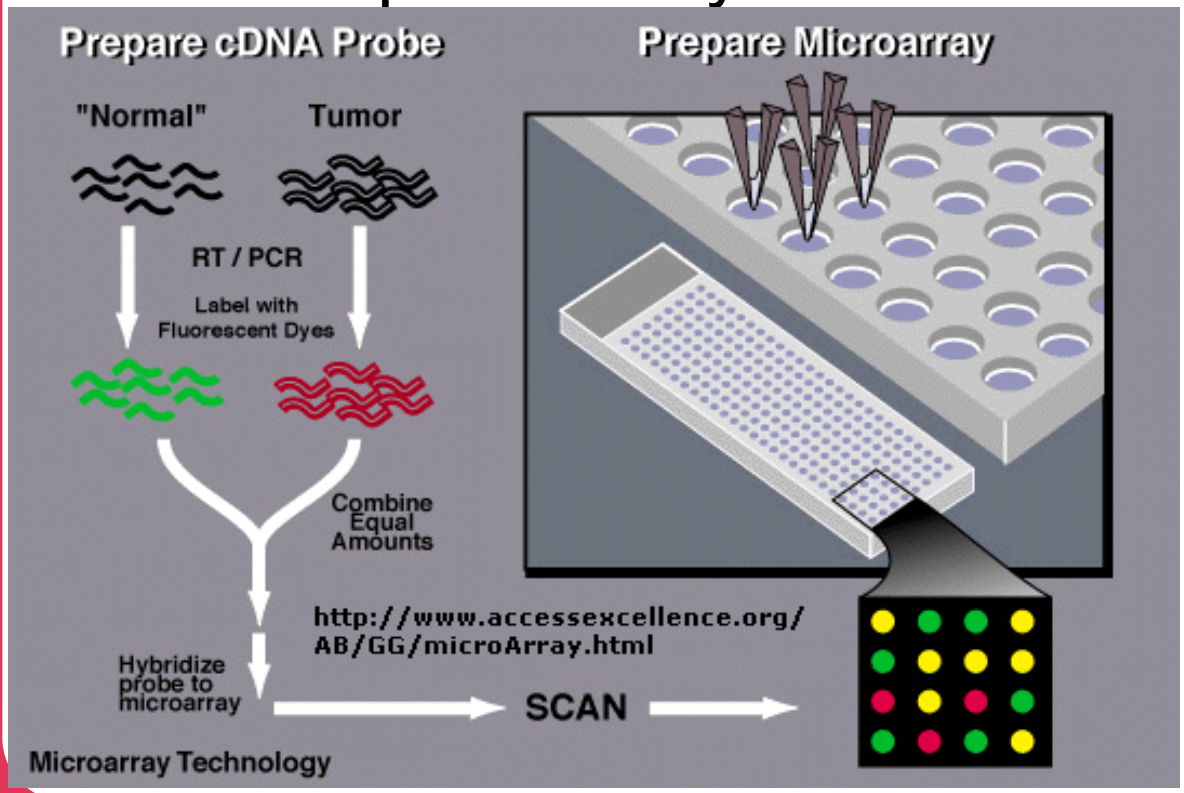
How does microarray work?

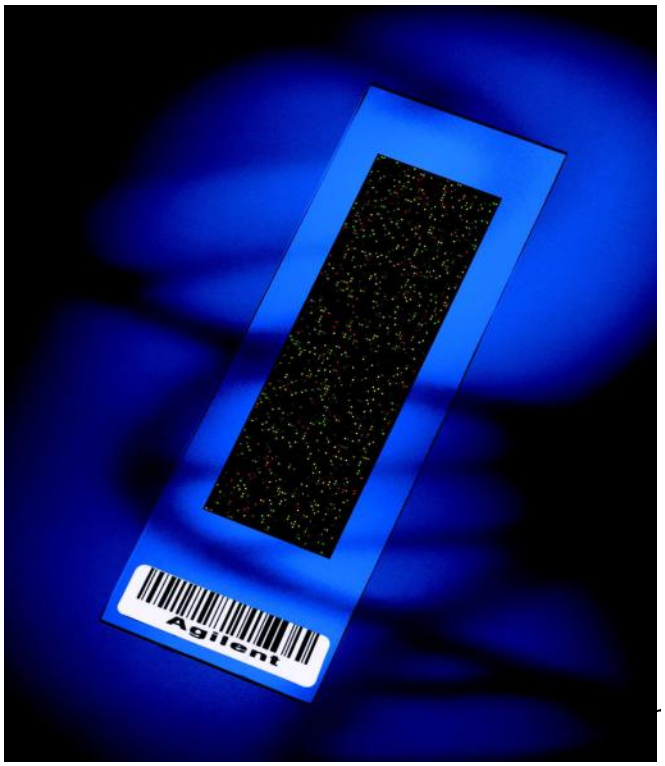


Wexner Medical Center

Two-channel microarray

- Printed microarrays
- Long probe oligonucleotides (80-100) long are “printed” on the glass chip
- Comparative hybridization experiment



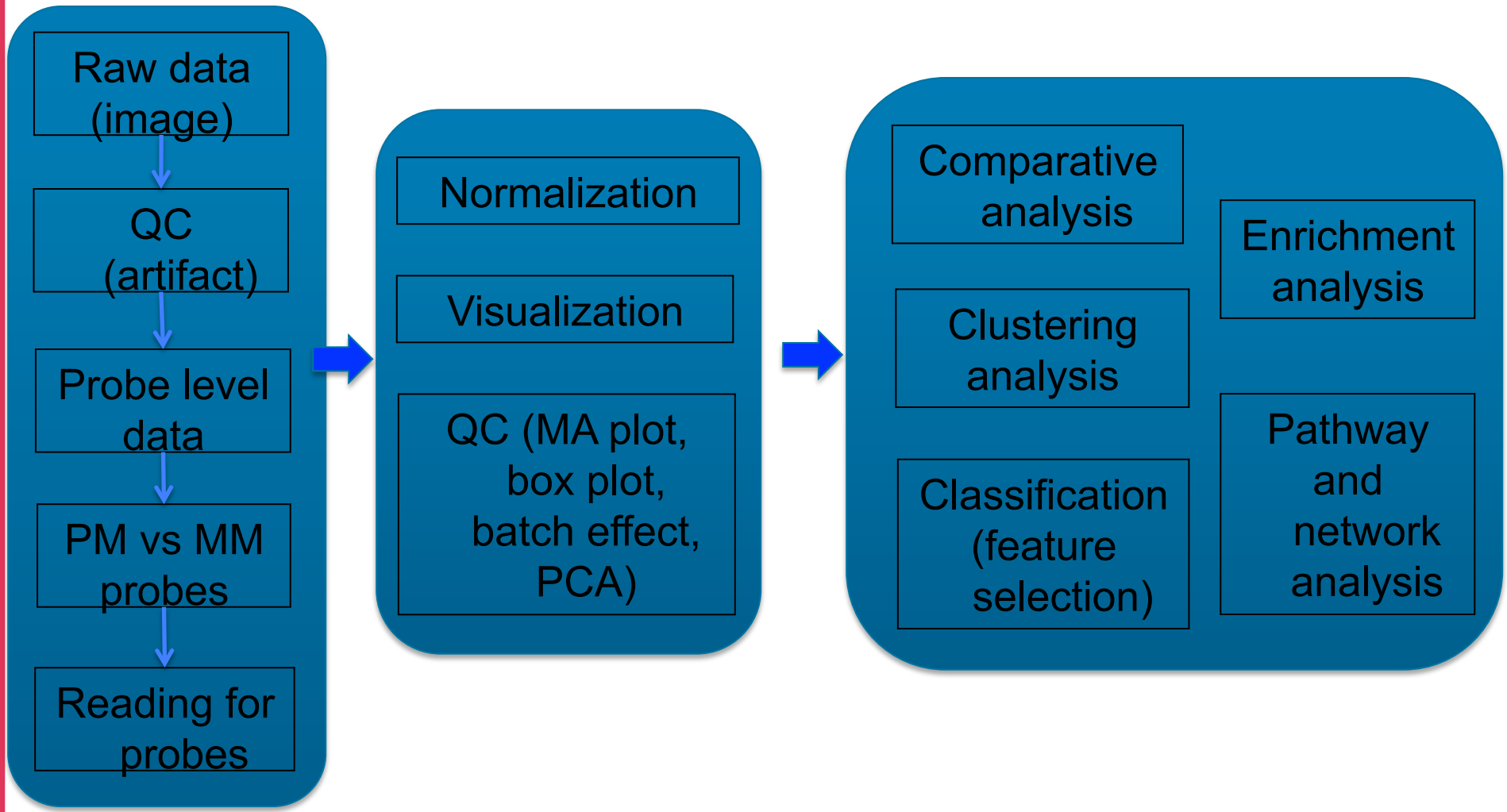


Reading Material – Chapter 3, Text

Outline

- What can we do with high throughput gene expression data?
- Profiling and comparative studies
- Machine learning approaches – biomarker discovery
- Correlation analysis – gene function prediction using co-expression
- Network analysis – network biomarkers

Data Analysis



Hypothesis Testing

Pick a dataset from GEO

Molecular Systems Biology 8; Article number 596; doi:10.1038/msb.2012.25
Citation: *Molecular Systems Biology* 8:596

www.molecularsystemsbiology.com

molecular systems biology

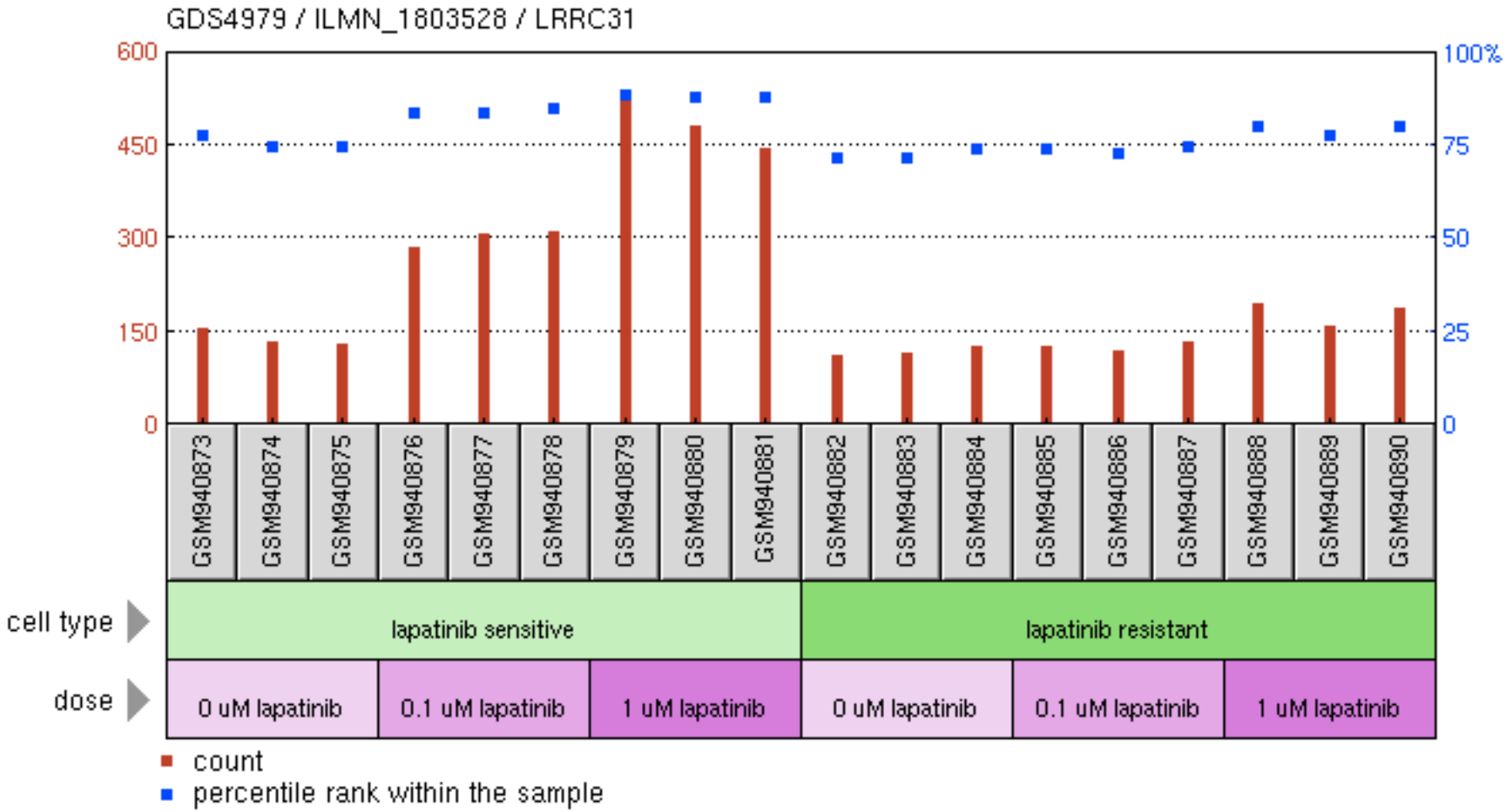
The glucose-deprivation network counteracts lapatinib-induced toxicity in resistant ErbB2-positive breast cancer cells

Kakajan Komurov[1], Jen-Te Tseng[1], Melissa Muller[1], Elena G Seviour[1], Tyler J Moss[1], Lifeng Yang[2], Deepak Nagrath[2] and Prahlad T Ram[1,*]

[1] Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA and [2] Chemical and Biomolecular Engineering Department, Rice University, Houston, TX, USA
* Corresponding author. Department of Systems Biology, The University of Texas MD Anderson Cancer Center, 7435 Fannin Street, Unit 0950, Houston, TX 77054, USA. Tel.: +1 713 563 4227; Fax: +1 713 563 4235; E-mail: pram@mdanderson.org

Hypothesis Testing

Two set of samples sampled from two distributions



New Resource

https://insilicodb.com/browse

Outline

- What can we do with high throughput gene expression data?
- Profiling and comparative studies
- Machine learning approaches – biomarker discovery
- Correlation analysis – gene function prediction using co-expression
- Network analysis – network biomarkers

A Case Study

Breast Cancer Prognosis Marker

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‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡

* *Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands*
‡ *Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034, USA*
† *These authors contributed equally to this work*

Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome. The strongest predictors for metastases (for example, lymph node status and histological grade) fail to classify accurately breast tumours according to their clinical behaviour[1–3]. Chemotherapy or hormonal therapy reduces the risk of distant metastases by approximately one-third; however, 70–80% of patients receiving this treatment would have survived without it[4,5]. None of the signatures of breast cancer gene expression reported to date[6–12] allow for patient-tailored therapy strategies. Here we used DNA microarray analysis on primary breast tumours of 117 young patients, and applied supervised classification to identify a gene expression signature strongly predictive of a short interval to distant metastases ('poor prognosis' signature) in patients without tumour cells in local lymph nodes at diagnosis (lymph node negative). In addition, we established a signature that identifies tumours of *BRCA1* carriers. The poor prognosis signature consists of genes regulating cell cycle, invasion, metastasis and

tumours are the dominant feature in this two-dimensional display (top and bottom of plot, representing 62 and 36 tumours, respectively), suggesting that the tumours can be divided into two types on the basis of this set of ~5,000 significant genes. Notably, in the upper group only 34% of the sporadic patients were from the group who developed distant metastases within 5 years, whereas in the lower group 70% of the sporadic patients had progressive disease (Fig. 1b). Thus, using unsupervised clustering we can already, to some extent, distinguish between 'good prognosis' and 'poor prognosis' tumours.

To gain insight into the genes of the dominant expression signatures, we associated them with histopathological data; for example, oestrogen receptor (ER)-α expression as determined by immunohistochemical (IHC) staining (Fig. 1b). Out of 39 IHC-stained tumours negative for ER-α expression (ER negative), 34 clustered together in the bottom branch of the tumour dendrogram. In the enlargement shown in Fig. 1c, a group of downregulated genes is represented containing both the ER-α gene *(ESR1)* and genes that are apparently co-regulated with ER, some of which are known ER target genes. A second dominant gene cluster is associated with lymphocytic infiltrate and includes several genes expressed primarily by B and T cells (Fig. 1d).

Sixteen out of eighteen tumours of *BRCA1* carriers are found in the bottom branch intermingled with sporadic tumours. This is consistent with the idea that most *BRCA1* mutant tumours are ER negative and manifest a higher amount of lymphocytic infiltrate[15]. The two tumours of *BRCA2* carriers are part of the upper cluster of tumours and do not show similarity with *BRCA1* tumours. Neither high histological grade nor angioinvasion is a specific feature of either of the clusters (Fig. 1b). We conclude that unsupervised clustering detects two subgroups of breast cancers, which differ in ER status and lymphocytic infiltration. A similar conclusion has also been reported previously[7,16].

The 78 sporadic lymph-node-negative patients were selected specifically to search for a prognostic signature in their gene expression profiles. Forty-four patients remained free of disease

http://www.nature.com/nature/journal/v415/n6871/full/415530a.html

Wexner Medical Center

Derive Prognostic Markers

- Gene expression microarray data for 295 breast cancer patients including 69 ER-negative ones
- Survival time is a major evaluation criteria for cancer treatment, thus developing prognosis markers is particularly important for cancers
- Full clinical information – ER, PR, HERR2 status, survival status and time, metastasis status, lymph node status, etc

Unsupervised Analysis the Data

Clustering

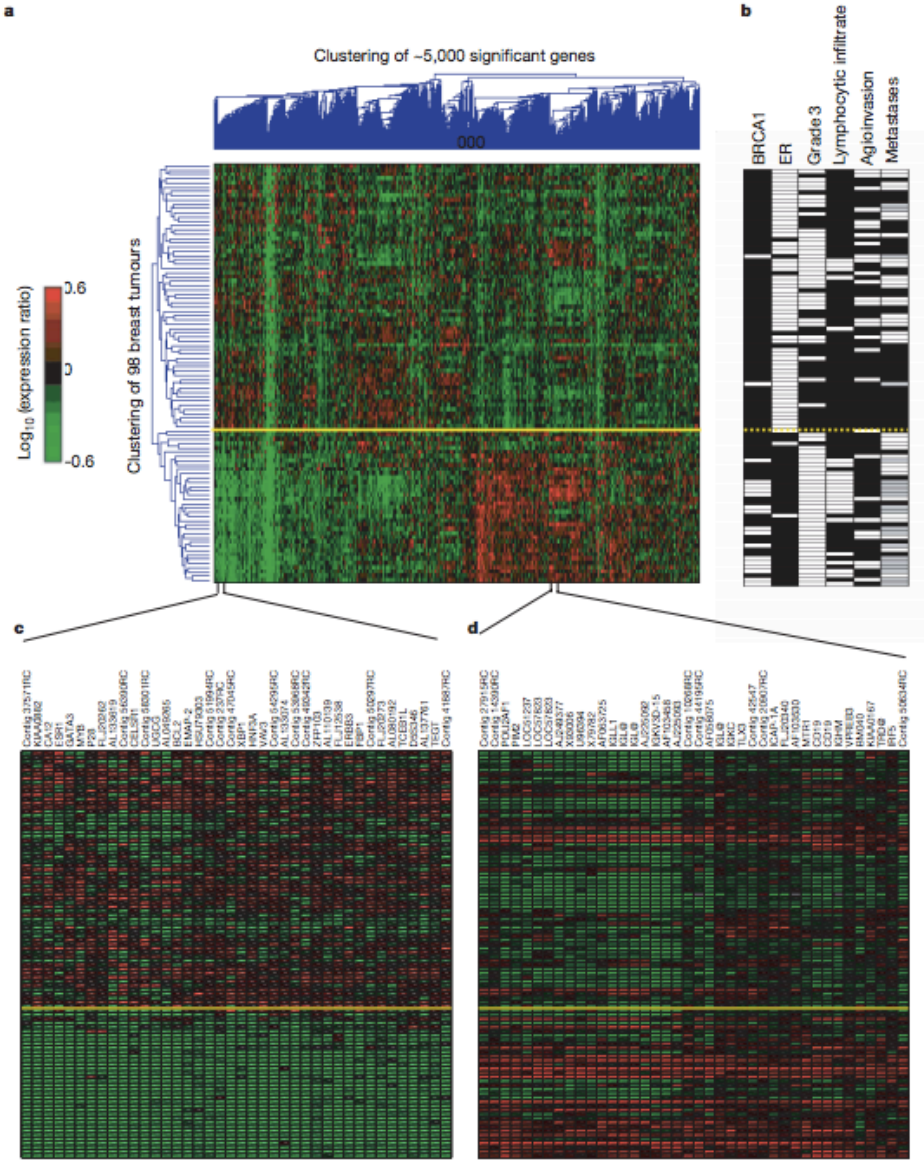


Figure 1 Unsupervised two-dimensional cluster analysis of 98 breast tumours. **a**, Two-dimensional presentation of transcript ratios for 98 breast tumours. There were 4,968 significant genes across the group. Each row represents a tumour and each column a single gene. As shown in the colour bar, red indicates upregulation, green downregulation, black no change, and grey no data available. The yellow line marks the subdivision into two dominant tumour clusters. **b**, Selected clinical data for the 98 patients in **a**: *BRCA1* germline mutation carrier (or sporadic patient), ER expression, tumour grade 3 (versus grade 1 and 2), lymphocytic infiltrate, angioinvasion, and metastasis status. White indicates positive, black negative and grey denotes tumours derived from *BRCA1* germline carriers who were excluded from the metastasis evaluation. The cluster below the yellow line consists of 36 tumours, of which 34 are ER negative (total 39 ER-negative) and 16 are carriers of the *BRCA1* mutation (total 18). **c**, Enlarged portion from **a** containing a group of genes that co-regulate with the ER-α gene (*ESR1*). Each gene is labelled by its gene name or accession number from GenBank. Contig ESTs ending with RC are reverse-complementary of the named contig EST. **d**, Enlarged portion from **a** containing a group of co-regulated genes that are the molecular reflection of extensive lymphocytic infiltrate, and comprise a set of genes expressed in T and B cells. (Gene annotation as in **c**.)

Wexner Medical Center

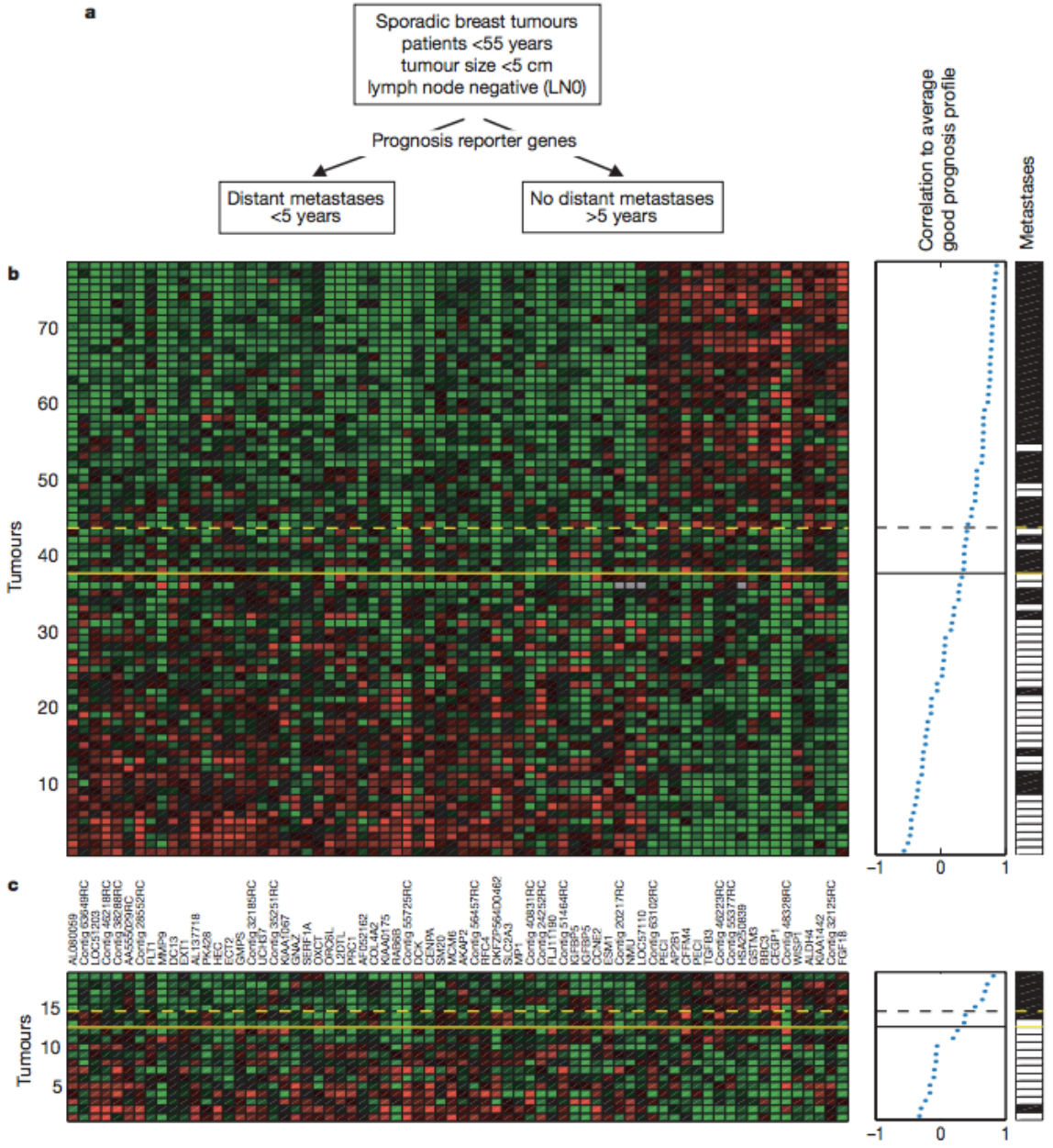
Gene Feature Selection

mean time to metastases 2.5 years) (Fig. 2a). To identify reliably good and poor prognostic tumours, we used a powerful three-step supervised classification method, similar to those used previously[8,17,18]. In brief, approximately 5,000 genes (significantly regulated in more than 3 tumours out of 78) were selected from the 25,000 genes on the microarray. The correlation coefficient of the expression for each gene with disease outcome was calculated and 231 genes were found to be significantly associated with disease outcome (correlation coefficient <-0.3 or >0.3). In the second step, these 231 genes were rank-ordered on the basis of the magnitude of the correlation coefficient. Third, the number of genes in the 'prognosis classifier' was optimized by sequentially adding subsets of 5 genes from the top of this rank-ordered list and evaluating its power for correct classification using the 'leave-one-out' method for cross-validation (see Supplementary Information). Classification was made on the basis of the correlations of the expression profile of the 'leave-one-out' sample with the mean expression levels of the remaining samples from the good and the poor prognosis patients, respectively. The accuracy improved until the optimal number of marker genes was reached (70 genes).

Supervised Analysis

Test

Validation



ner Medical Center

Gene Markers

- The 70 gene signature was later incorporated into a test called Mammaprint and is used clinically (FDA approved) for breast cancer prognosis

- Other clinically used gene signatures also exists such as OncoDX, PAM50

- Prognosis signature is not enough for personalized treatment. Ongoing work is more and more shifted toward Predictive maker discovery (e..g, to predict drug response).

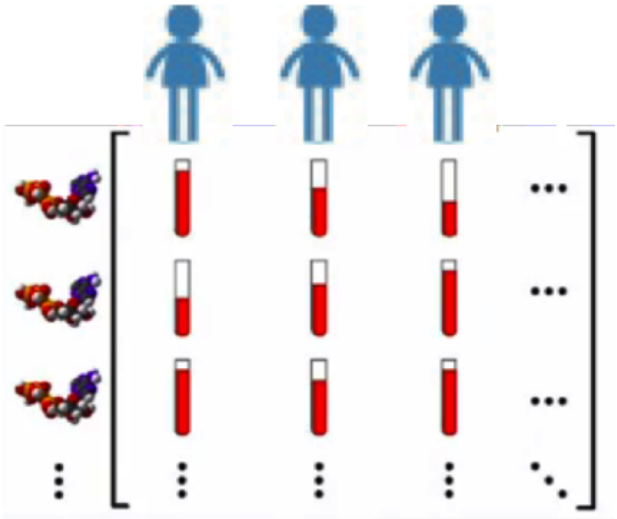
Another Case Study

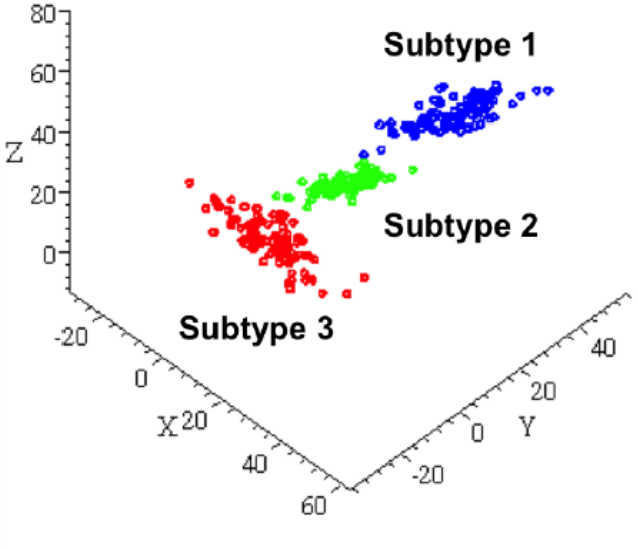
Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub *et al.*
Science **286**, 531 (1999);
DOI: 10.1126/science.286.5439.531

http://www.ncbi.nlm.nih.gov/pubmed/10521349

Goal - Stratification





Wexner Medical Center

www.google.com ☺

PubMed Id: 10521349

Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

Also Available @

http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43

http://www.broadinstitute.org/cancer/software/genepattern/datasets/

The Premise

- Class discovery to automatically distinguish between
  - acute myeloid leukemia: AML
  - acute lymphoblastic leukemia: ALL
- No knowledge of these classes
- Genetic causes known - specific chromosomal translocations
- Use microarrays

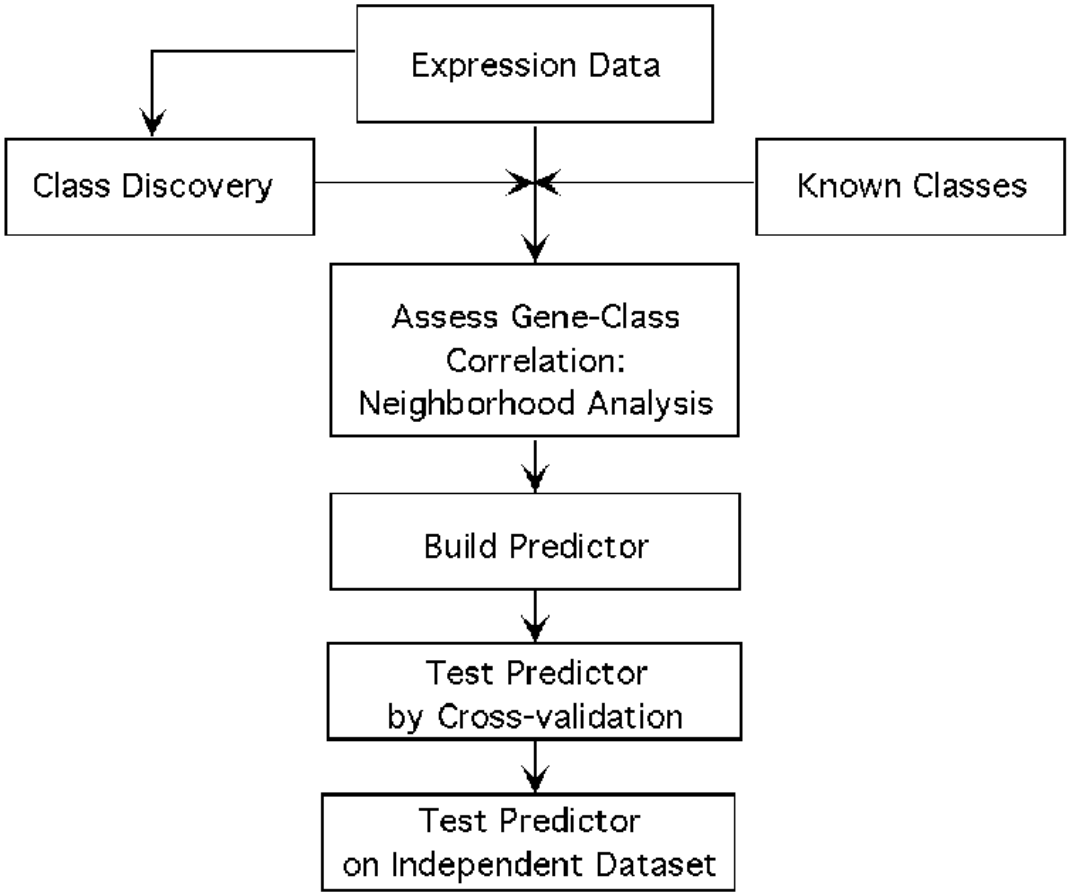
The Data

- 38 bone marrow samples (27 ALL, 11 AML)
- Obtained from patients at time of diagnosis
- RNA from bone marrow mononuclear cells

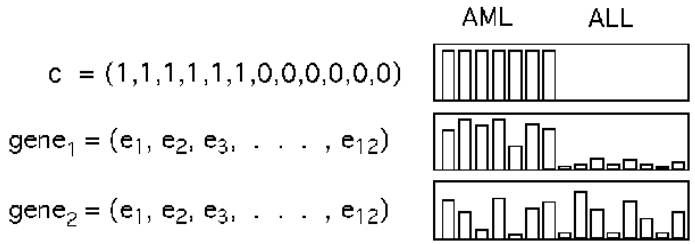
- Affymetrix microarrays
- Probes for 6817 human genes

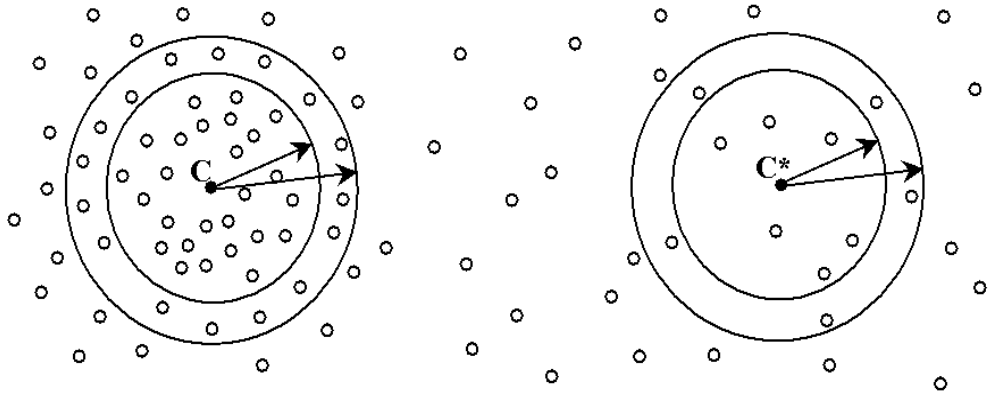
The Methods

Schematic illustration of methodology. (a) Strategy for cancer classification. Tumor classes may be known a priori or discovered on the basis of the expression data by using Self-Organizing Maps (SOMs) as described in the text. Class Prediction involves assignment of an unknown tumor sample to the appropriate class on the basis of gene expression pattern. This consists of several steps: neighborhood analysis to assess whether there is a significant excess of genes correlated with the class distinction, selection of the informative genes and construction of a class predictor, initial evaluation of class prediction by cross-validation, and final evaluation by testing in an independent data set.



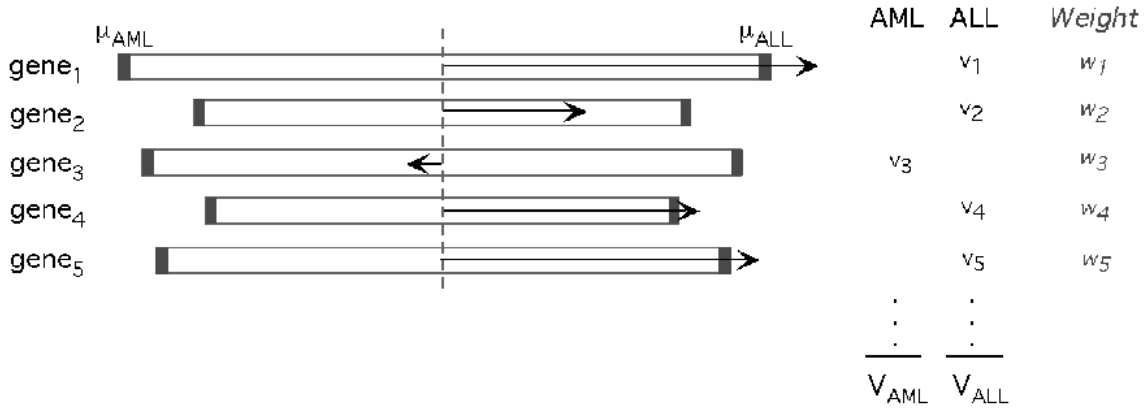
Supervised Analysis





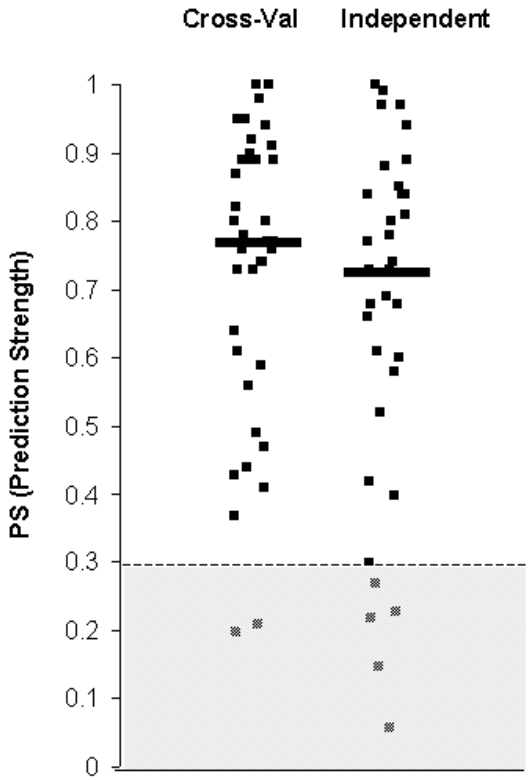
Neighborhood Analysis. The class distinction is represented by an 'idealized expression pattern c, in which the expression level is uniformly high in class 1 and uniformly low in class 2. Each gene is represented by an expression vector, consisting of its expression level in each of the tumor samples. In the figure, the dataset consists of 12 samples comprised of 6 AMLs and 6 ALLs. Gene g1 is well correlated with the class distinction, while g2 is poorly correlated. Neighborhood analysis involves counting the number of genes having various levels of correlation with c. The results are compared to the corresponding distribution obtained for random idealized expression patterns c*, obtained by randomly permuting the coordinates of c. An unusually high density of genes indicates that there are many more genes correlated with the pattern than expected by chance. The precise measure of distance and other methodological details are described in notes (16,17) and on our web site.

Prediction



The prediction of a new sample is based on 'weighted votes' of a set of informative genes. Each such gene gi votes for either AML or ALL, depending on whether its expression level x_i in the sample is closer to mu_AML or mu_ALL (which denote, respectively, the mean expression levels of AML and ALL in a set of reference samples). The magnitude of the vote is w_i v_i, where w_i is a weighting factor that reflects how well the gene is correlated with the class distinction and v_i = |x_i - (mu_AML + mu_ALL)/2| reflects the deviation of the expression level in the sample from the average of mu_AML and mu_ALL. The votes for each class are summed to obtain total votes V_AML and V_ALL. The sample is assigned to the class with the higher vote total, provided that the prediction strength exceeds a predetermined threshold. The prediction strength reflects the margin of victory and is defined as (V_win-V_lose)/(V_win+V_lose), where as V_win and V_lose are the respective vote totals for the winning and losing classes. Methodological details are described in the paper (notes 19,20).

Testing

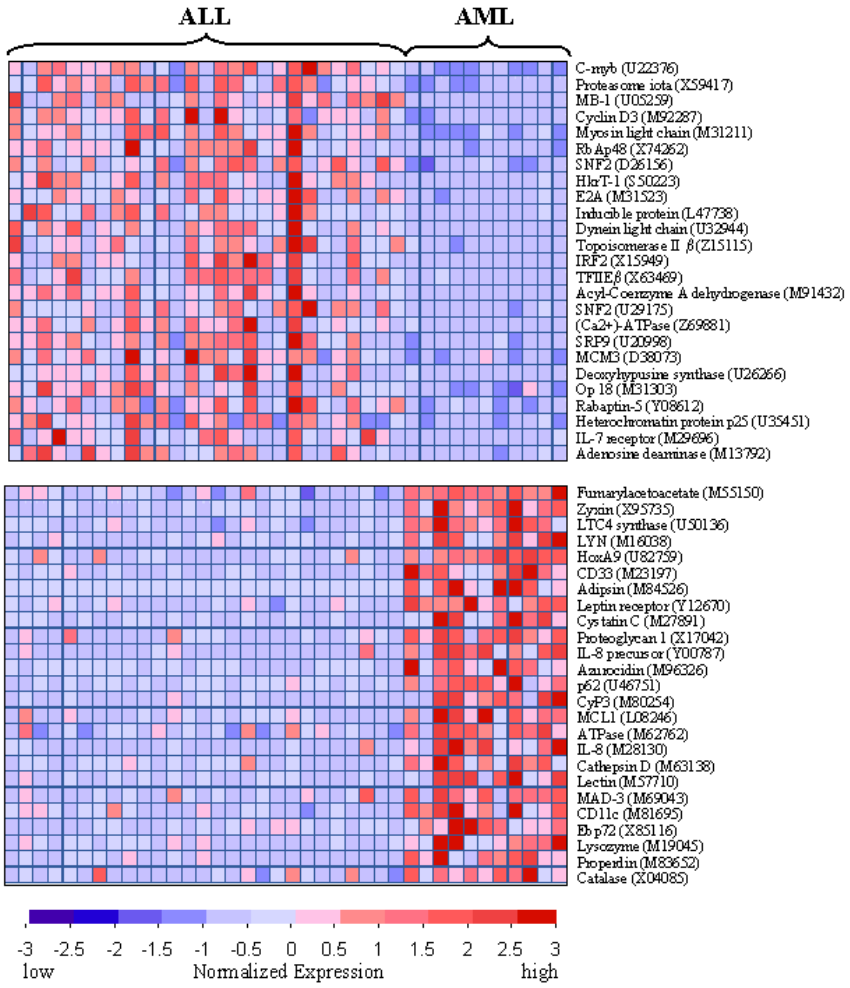


Prediction strengths. The scatterplots show the prediction strengths (PS) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS below 0.3 are considered as uncertain.

Wexner Medical Center

Results

Genes distinguishing ALL from AML. The 50 genes most highly correlated with the ALL/AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and the standard deviation is 1. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue. The scale indicates standard deviations above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML. Note that while these genes as a group appear correlated with class, no single gene is uniformly expressed across the class, illustrating the value of a multi-gene prediction method.



Wexner Medical Center

Outline

- What can we do with high throughput gene expression data?
- Profiling and comparative studies
- Machine learning approaches – biomarker discovery
- Correlation analysis – gene function prediction using co-expression
- Network analysis – network biomarkers

Gene Co-Expression

ARTICLES

nature genetics

Network modeling links breast cancer susceptibility and centrosome dysfunction

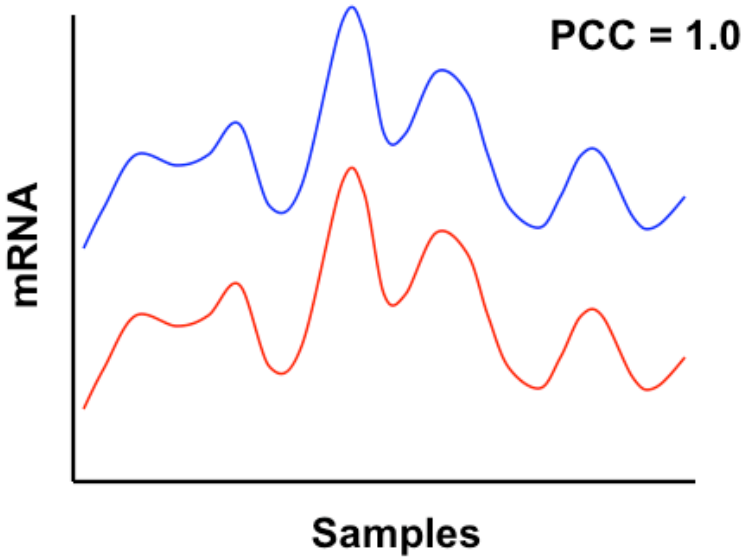
Miguel Angel Pujana[1,2,16,17], Jing-Dong J Han[1,2,16,17], Lea M Starita[3,16,17], Kristen N Stevens[4,17], Muneesh Tewari[1,2,16], Jin Sook Ahn[1,2], Gad Rennert[5], Víctor Moreno[6,7], Tomas Kirchhoff[8], Bert Gold[9], Volker Assmann[10], Wael M ElShamy[2], Jean-François Rual[1,2], Douglas Levine[8], Laura S Rozek[6], Rebecca S Gelman[11], Kristin C Gunsalus[12], Roger A Greenberg[2], Bijan Sobhian[2], Nicolas Bertin[1,2], Kavitha Venkatesan[1,2], Nono Ayivi-Guedehoussou[1,2,16], Xavier Solé[7], Pilar Hernández[13], Conxi Lázaro[13], Katherine L Nathanson[14], Barbara L Weber[14], Michael E Cusick[1,2], David E Hill[1,2], Kenneth Offit[8], David M Livingston[2], Stephen B Gruber[4,6,15], Jeffrey D Parvin[3,16] & Marc Vidal[1,2]

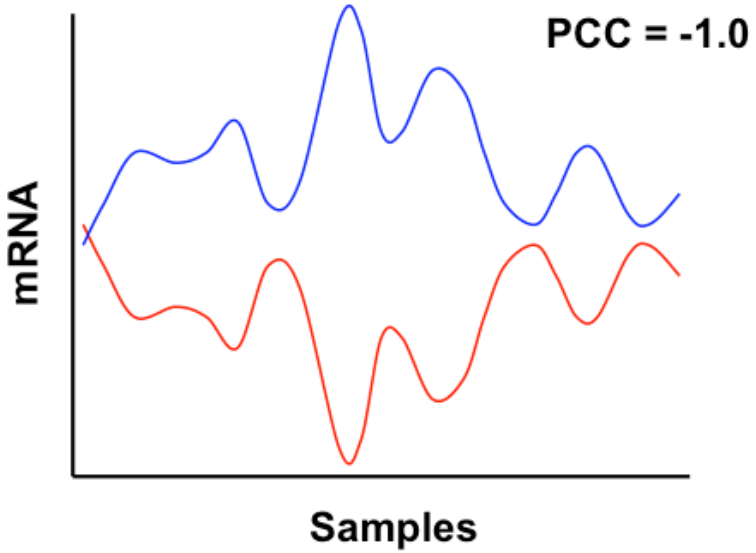
Many cancer-associated genes remain to be identified to clarify the underlying molecular mechanisms of cancer susceptibility and progression. Better understanding is also required of how mutations in cancer genes affect their products in the context of complex cellular networks. Here we have used a network modeling strategy to identify genes potentially associated with higher risk of breast cancer. Starting with four known genes encoding tumor suppressors of breast cancer, we combined gene expression profiling with functional genomic and proteomic (or 'omic') data from various species to generate a network containing 118 genes linked by 866 potential functional associations. This network shows higher connectivity than expected by chance, suggesting that its components function in biologically related pathways. One of the components of the network is ***HMMR***, encoding a centrosome subunit, for which we demonstrate previously unknown functional associations with the breast cancer–associated gene ***BRCA1***. Two case-control studies of incident breast cancer indicate that the ***HMMR*** locus is associated with higher risk of breast cancer in humans. Our network modeling strategy should be useful for the discovery of additional cancer-associated genes.

http://www.ncbi.nlm.nih.gov/pubmed/17922014

Gene Co-Expression

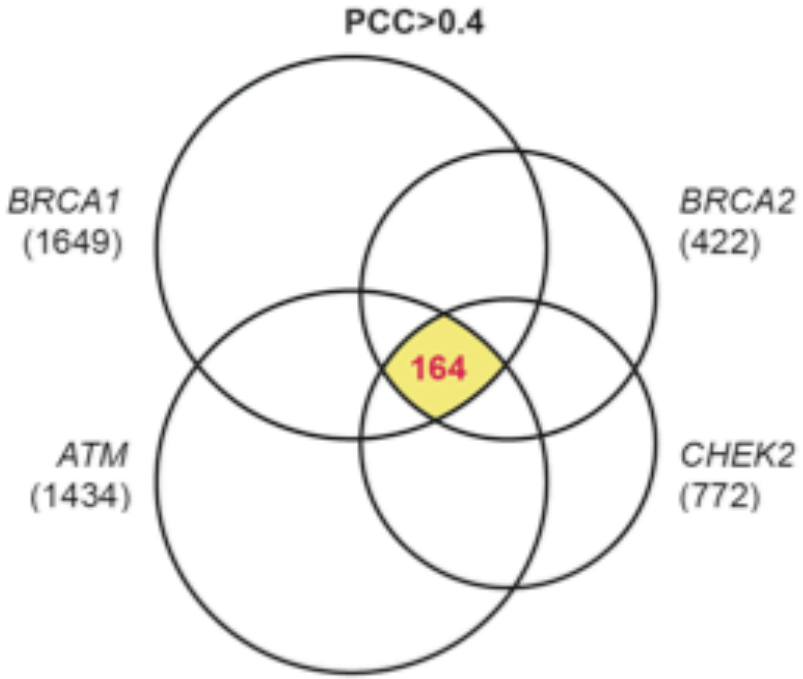
- Correlated gene expression profiles
  - Both positive and negative correlation
- Why do genes co-express?
  - Functional reasons?
  - Mechanistic reasons?
  - Genetic reasons?





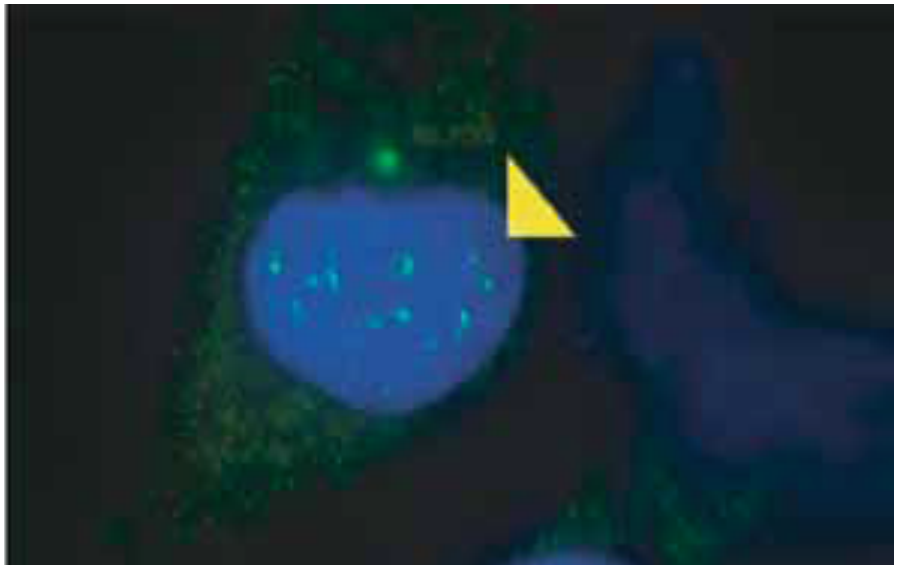
Gene Co-Expression

- Genes co-express with multiple “anchor” genes
- Anchor genes were selected from BRCA1 pathways
- Goal – using co-expression relationships to identify new breast cancer genes



Gene Co-Expression

- HMMR was ranked highest
- Experimental validation – siRNA silencing of HMMR led to similar phenotypes in HeLa and breast cancer cells as knockout of BRCA1 – multiple centrosomes
- GWAS study (using CGEMS data) showed that certain mutation on HMMR is associated with increased breast cancer risk



HMMR siRNA

Outline

- What can we do with high throughput gene expression data?
- Profiling and comparative studies
- Machine learning approaches – biomarker discovery
- Correlation analysis – gene function prediction using co-expression
- Network analysis – network biomarkers

Integrative Genomics – Multiple Phenotypes

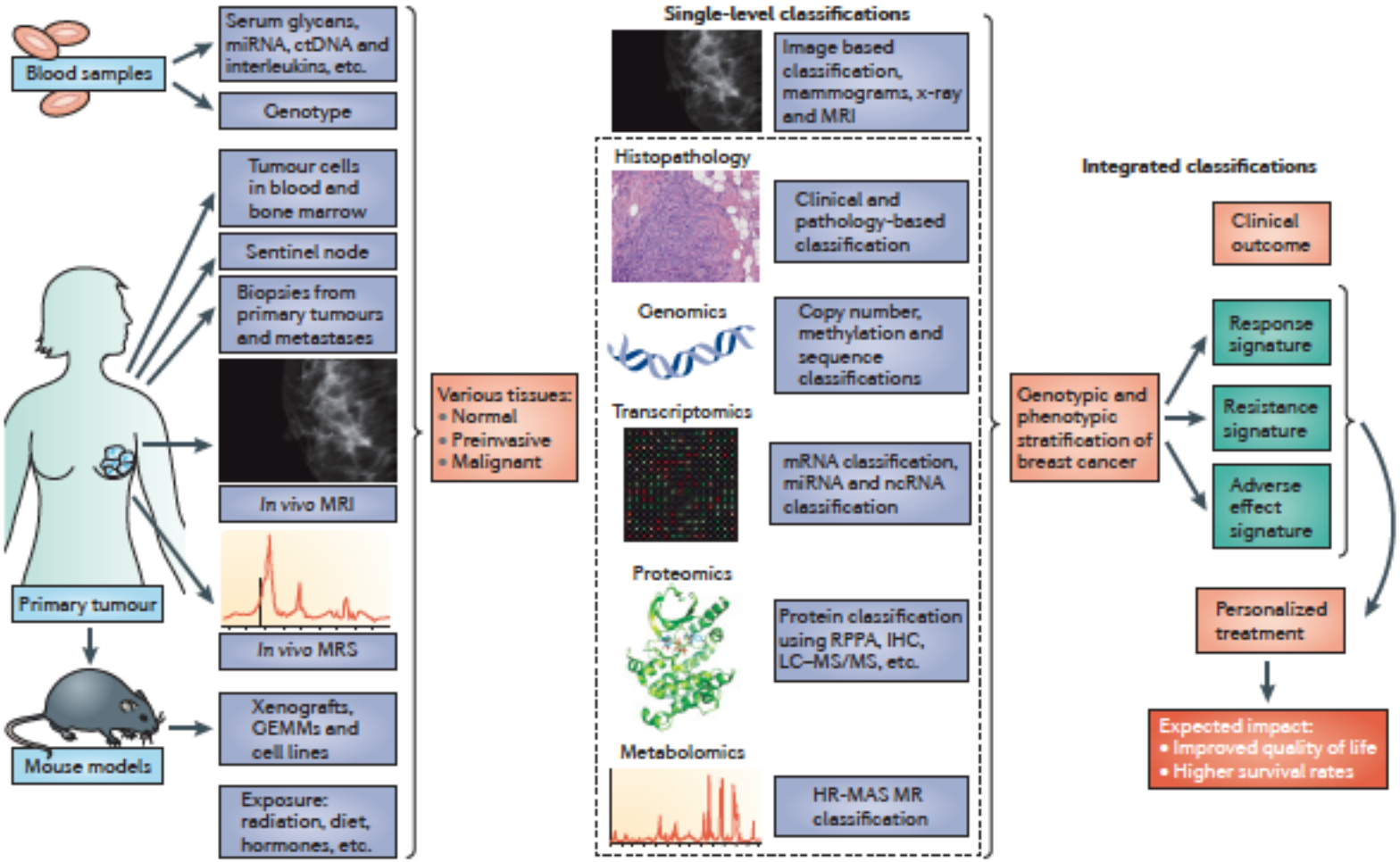
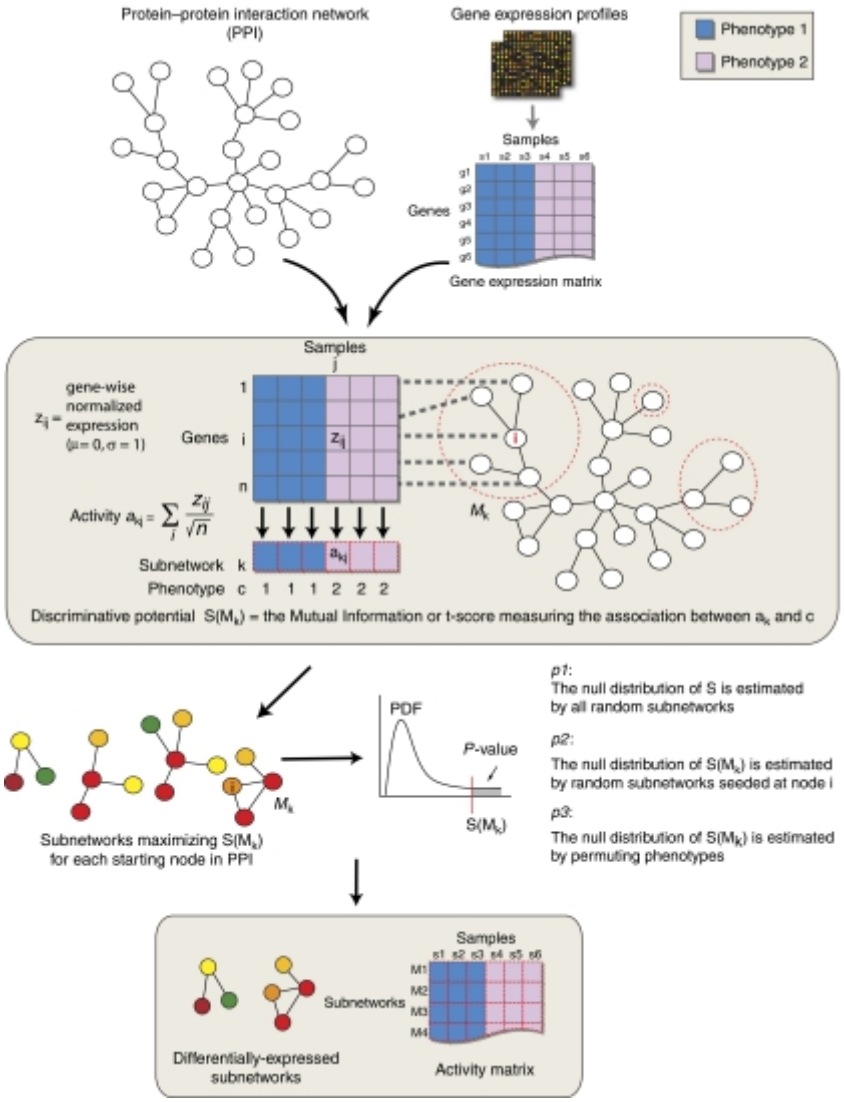


Figure 1 | The systems biology of breast cancer. Exploring the systems biology of breast cancer and strategies to investigate multi-dimensional interactions by integration of data from various sources at the indicated levels. ctDNA, circulating tumour DNA; GEMMs, genetically engineered mouse models; HR-MAS MR, high-resolution magic angle spinning magnetic resonance; IHC, immunohistochemistry; LC–MS, liquid chromatography–mass spectrometry; miRNA, microRNA; MRI, magnetic resonance imaging; MRS, magnetic resonance spectroscopy; ncRNA, non-coding RNA; RPPA, reverse phase protein array. Mammography image courtesy of M. M. Holmen of Oslo University Hospital, Oslo, Norway.

Microarray Data + Protein Interaction Networks

- Systems approach
- Instead of “which genes changed” to “which part of the network is perturbed?”
    - “Network-based classification of breast cancer metastasis.” (Chung et al, Mol. Sys. Bio., 2007)

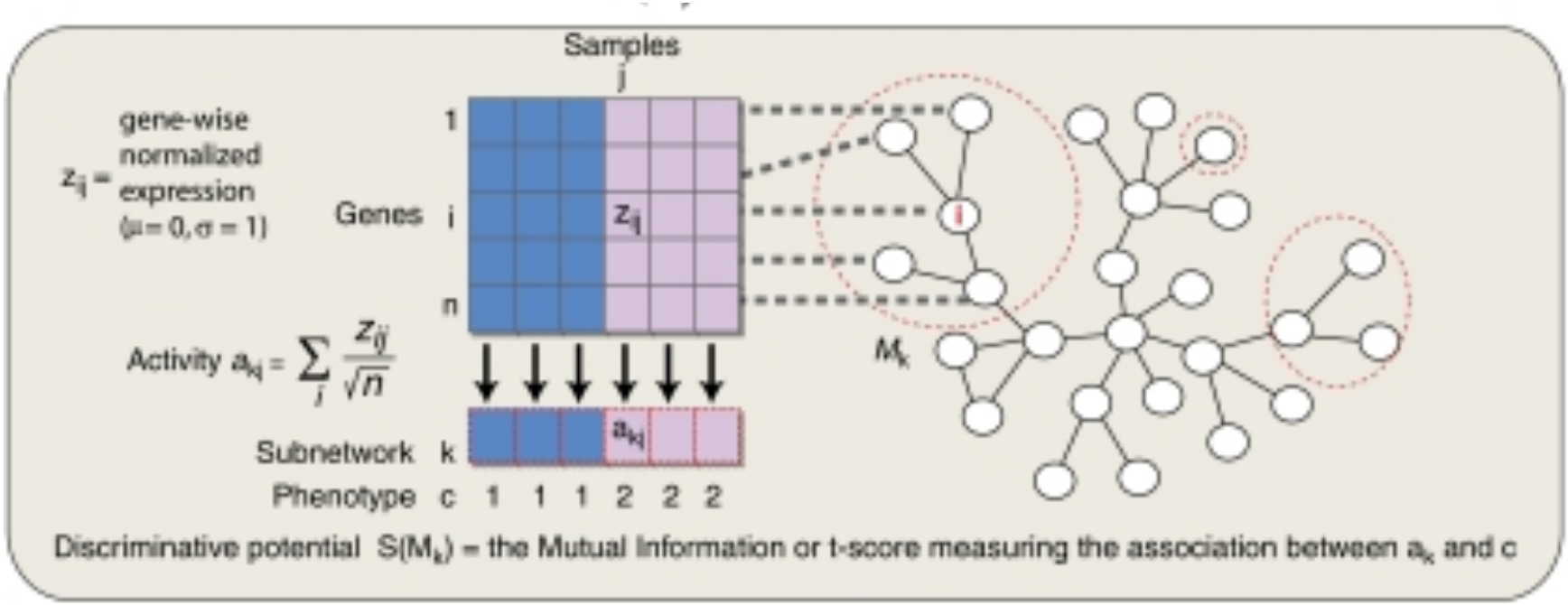
Microarray Data + Protein Interaction Networks



Wexner Medical Center

Microarray Data + Protein Interaction Networks

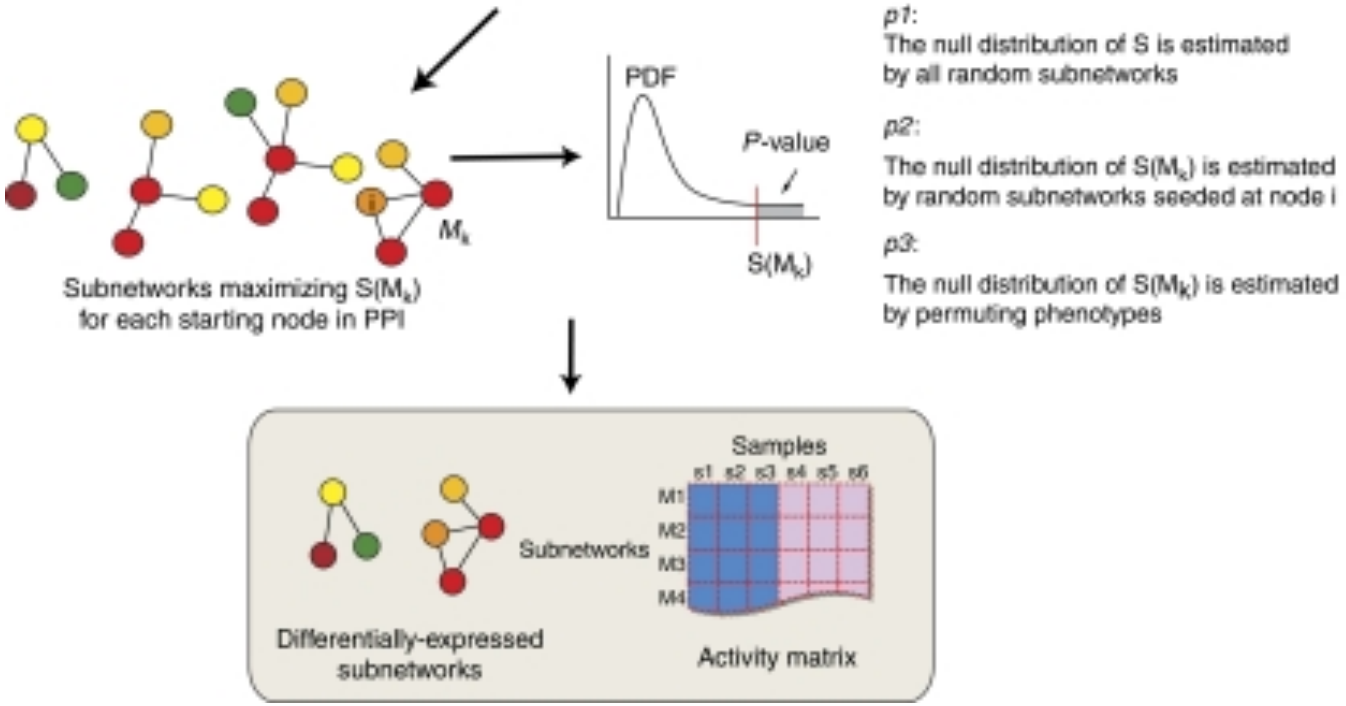
- Defining gene network “activity” score
- Correlate the activity score with different disease conditions (e.g., metastasis vs non-metastasis)
- Network mining – a greedy approach



Wexner Medical Center

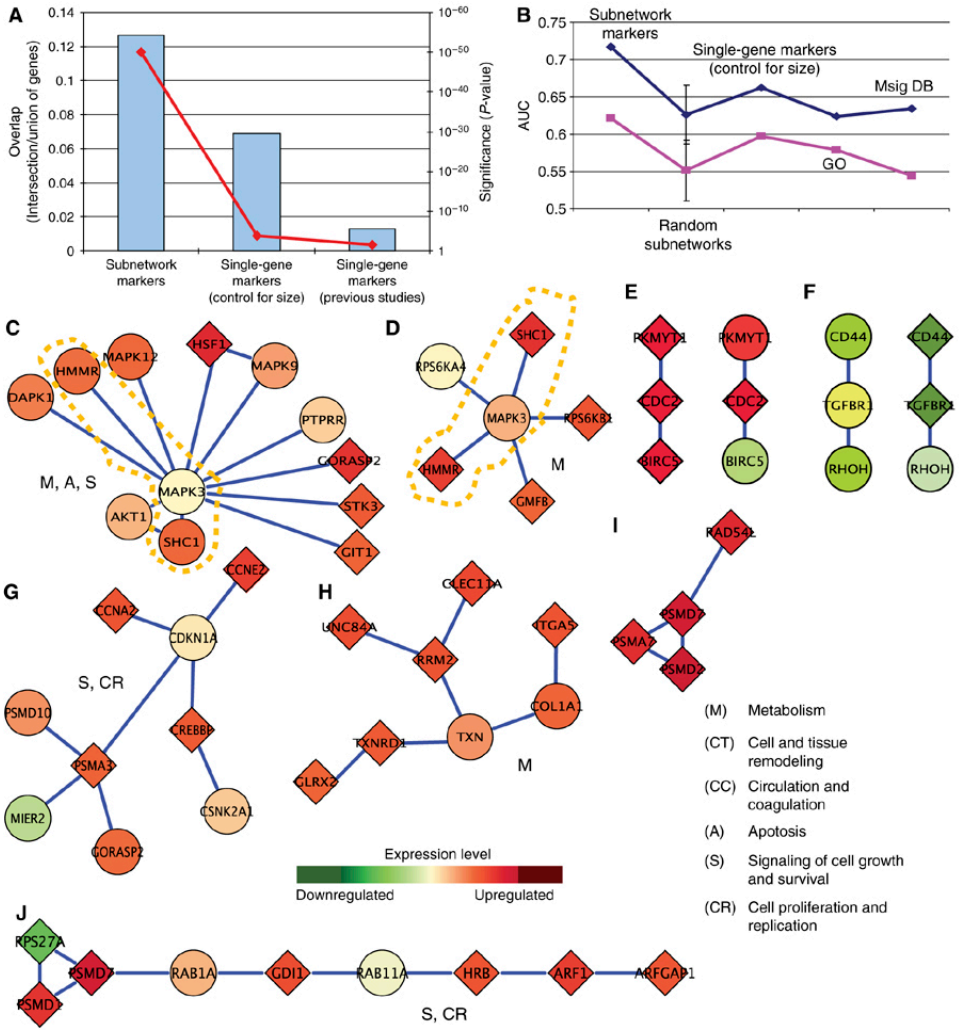
Microarray Data + Protein Interaction Networks

Statistical evaluation of the identified gene modules based on random permutations



Microarray Data + Protein Interaction Networks

Marker reproducibility and metastasis prediction performance.



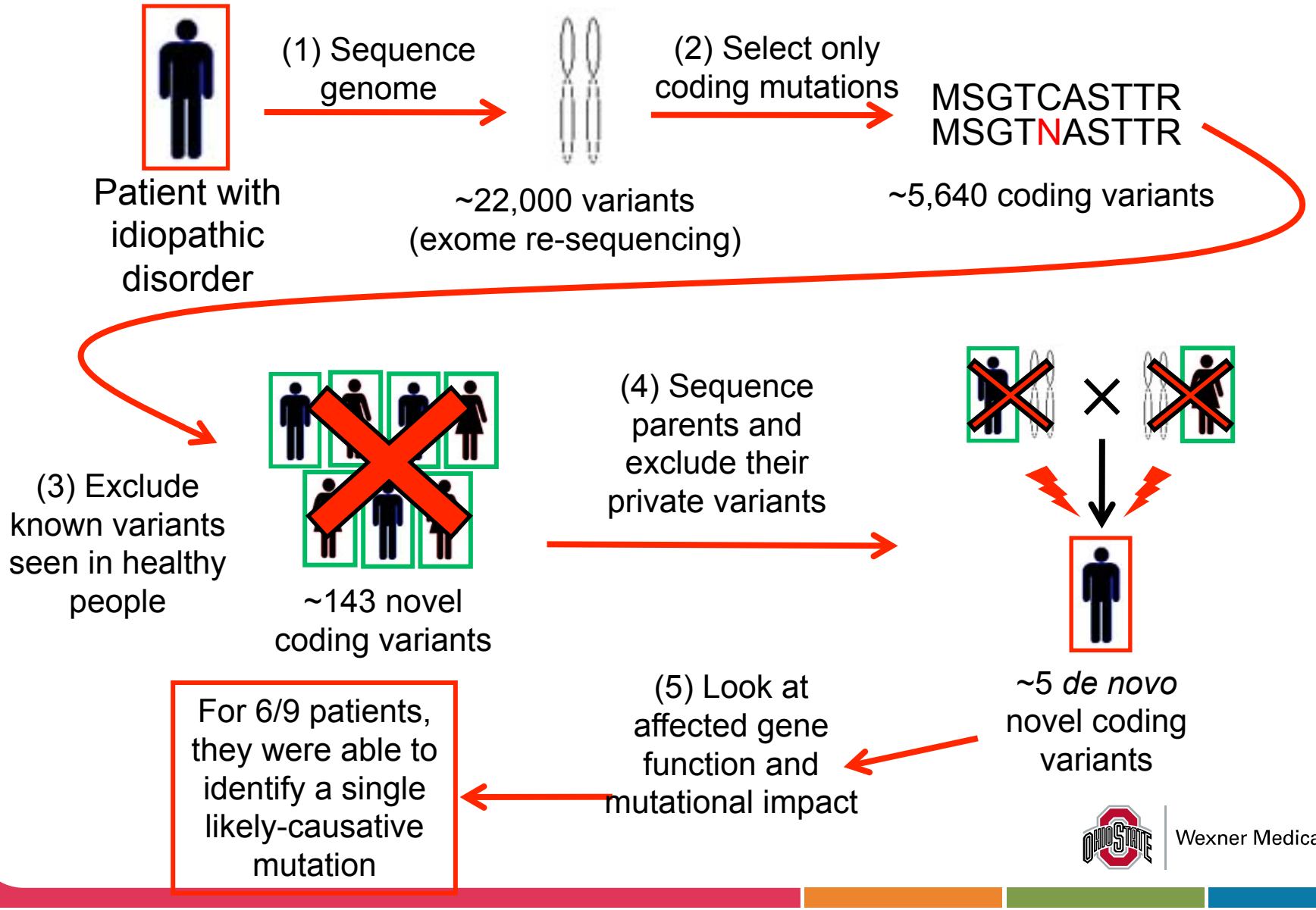
Han-Yu Chuang et al. Mol Syst Biol 2007;3:140

Outline

- What can we do with high throughput gene expression data?
- Profiling and comparative studies
- Machine learning approaches – biomarker discovery
- Correlation analysis – gene function prediction using co-expression
- Network analysis – network biomarkers
- Genomic variants – the drivers

Identifying a causative *de novo* mutation

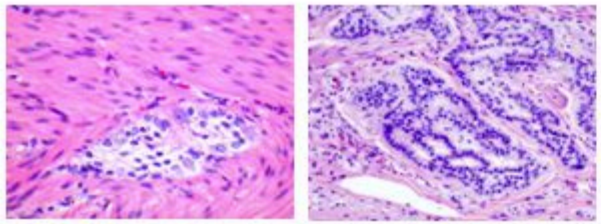
Veltman and colleagues - Nat Genet. 2010 Dec;42(12):1109-12



Wexner Medical Center

RNA sequencing

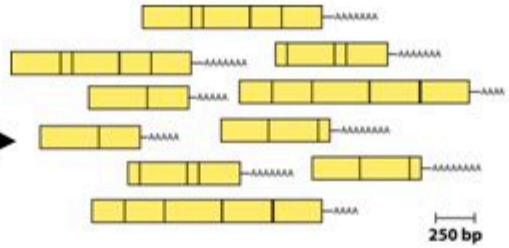
Samples of interest



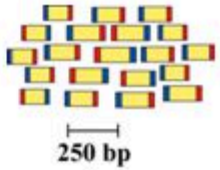
Condition 1 (normal colon)

Condition 2 (colon tumor)

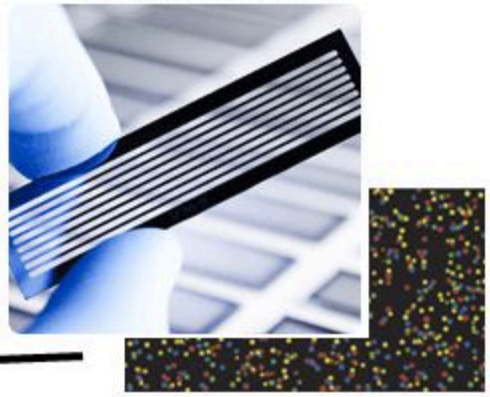
Isolate RNAs



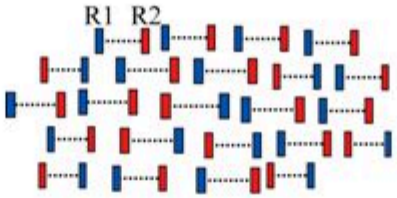
Generate cDNA, fragment, size select, add linkers



Sequence ends



100s of millions of paired reads
10s of billions bases of sequence



Map to genome, transcriptome, and predicted exon junctions

Downstream analysis

DREAM 9.5 contest

Sub-Challenge 1: Fusion Prediction

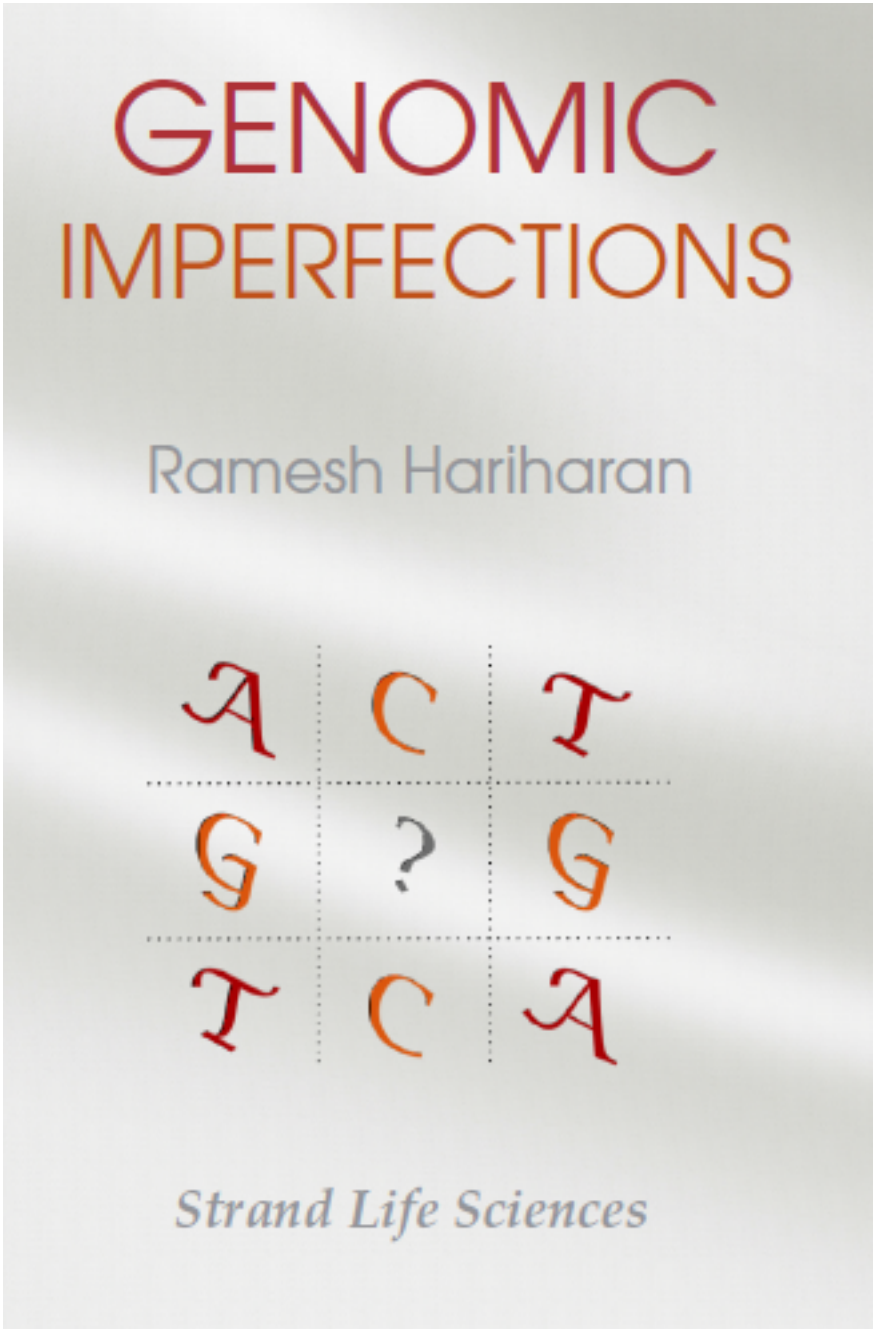
Scientific Rationale: Gene fusions have an important role in the initial steps of tumorigenesis. An increasing number of gene fusions are being recognized as important diagnostic and prognostic parameters in malignant haematological disorders and childhood sarcomas. Gene fusions occur in all malignancies and account for 20% of human cancer morbidity.[4]

Sub-Challenge 2: Isoform Prediction

Scientific Rationale: Cancer cell lines, regardless of their tissue of origin, can be effectively discriminated from non-cancer cell lines at isoform level, but not at gene level. Existence of an isoform signature, rather than a gene signature, could be used to distinguish cancer cells from normal cells.[5]

Sub-Challenge 3: Somatic SNV Calling

Scientific Rationale: Accurate identification of somatic SNVs can contribute to the discovery of critical mutations and expression changes that drive tumorigenesis, drug resistance, or tumor evolution in cancer.



Wexner Medical Center

Synopsis -
http://www.jsonline.com/news/health/1122482

Causes -
http://media.journalinteractive.com/documents/dna2gr.pdf

JSONLINE.COM
MILWAUKEE · WISCONSIN
JOURNAL SENTINEL
PULITZER PRIZE WINNER 2008 · 2010 · 2011

Wexner Medical Center

The End

Next – Dr. Parvin's Lecture on Sequencing T

Wexner Medical Center

57