High-throughput gene profiling in neurobiology

Pietro Paolo Sanna

High-throughput gene profiling in neurobiology

- Gene array technologies
- Strategy and analysis
- Applications to neurobiology

High-throughput gene profiling in neurobiology

- **Gene array technologies**
- **Strategy and analysis**
- **Applications to neurobiology**

High-throughput approaches in basic science

- Transcriptomics
- Proteomics
- Metabolomics

- **Microarrays**
- **Forerunners (a partial list):**

- slot/dot blots
- differential display
- macro arrays
-Subtractive hybridization and cloning

…and alternatives:

- TOGA (Total Gene exp. Analysis)
- SAGE (Serial Analysis of Gene Exp.)
- NEXT GEN SEQ

High-throughput approaches in basic science

- Transcriptomics
- Proteomics
- Metabolomics

Complementary and/or alternative to Transcriptomics:
- 2D gel electrophoresis
- multidimensional chromatography coupled with MassSpec

High-throughput approaches in basic science

- Transcriptomics
- Proteomics
- Metabolomics

Small-molecule metabolite profiling (the metabolome)

The “modern” microarrays

QuickTime and a
decompressor
are needed to see this picture.

QuickTime and a
decompressor
are needed to see this picture.

45 cDNAs from the small flowering plant Arabidopsis thaliana were spotted in duplicate on glass slides, hybridized to fluorescently labeled cRNAs and laser-scanned. Quantitation by two color competitive hybridization: root (flurescein) leaf (lissamine).

Proof of principle, scalable, quantitative

A late arrival

QuickTime and a
decompressor
are needed to see this picture.

High-throughput strategies were implemented relatively recently and address a long-standing unmet need in the biology tool box.

Main types of microarrays for transcription analysis

- High-density microarrays
- Randomly ordered microarrays
- Slide arrays
- PCR arrays

Main types of micorarrays for transcription analysis - I

Affymetrix GeneChip

•High-density microarrays
•*In situ* DNA synthesis by photolithography and masking
• 11 µm features
• GeneChip format
• Requires dedicated scanner
• 25mer oligonucleotides probes
• 11-40 different probes per gene

Illumina

•Randomly ordered microarrays
•3-µm silica beads on one of two substrates: fiber optic bundles or planar silica slides. Spacing of ~5.7 µm
•BidArray format
•Requires dedicated scanner
•50mer oligonucleotides
•High redundancy (> 20-fold) but mostly single probe per gene.

Main types of micorarrays for transcription analysis - II

Slide format microarrays

Low-density arrays

Spotted (cDNA)

Agilent
•60-mer oligonucleotides
•inkjet process
•up to 44k probes about 1 probe/gene

Nimblegen
•60mer oligonucleotides
•Digital Micromirror Device, UV de-protecting
•40k genes x 8 unique probes per gene; or 24k genes x 3 unique probes x 4plex

PCR arrays

TaqMan (Abi) and **SuperArray (Bioscience)**
•Custom and focused arrays e.g: pathways, Ion Channels, Apoptosis, Proteases, miRNA et cetera
•35mer probes (ABI)
•96-384 well format
•Large number of assays per species available
•Pre-designed focused panels

Pros & cons of main microarray types

- Affymetrix :

pros: genome coverage in single array
high signal/noise
redundancy of design (multiple probes per gene)
scalable experimental design
large selection of species

cons: more expensive than other platforms (-> low sample size); short oligonucleotides

---

- Illumina :

pros: genome coverage in single array
very high density and low input size (≥50 ng)
high redundancy (mostly single probes >30 folds)
lowest priced of main commercial arrays

cons: limited selection of species (human, mouse, rat)
mostly single probe per gene

Slide microarrays:
(Agilent, Nimblegen)

pros: genome coverage in single array
highly customizable (design, oligo length)

cons: more expensive than Illumina

- Spotted array

pros: customizable, suited for specialized applications

cons: labor intensive (cDNA library/manufacturing)
limited to competitive hybridization

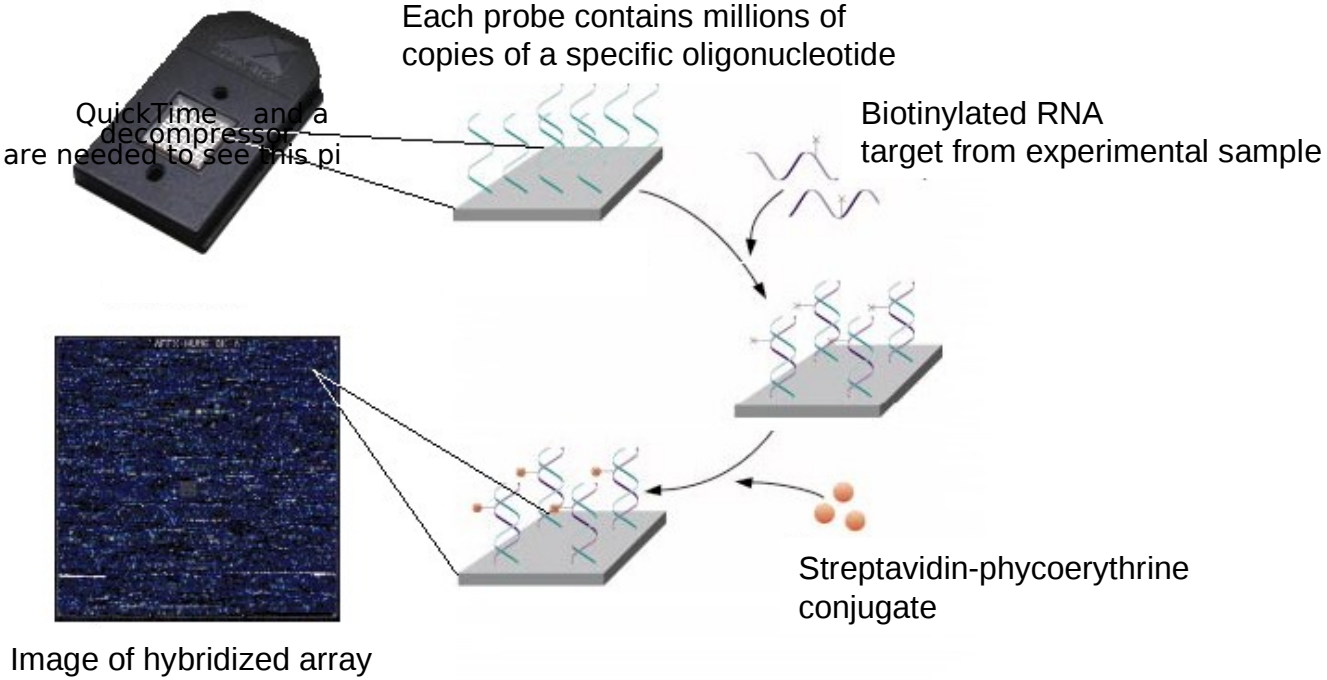
- PCR arrays:

pros: well-suited for specialized applications

cons: limited number of probes

The **Affymetrix GeneChip** expression analysis process

GeneChip arrays



Affymetrix expression analysis process

One probe set is composed of between 11 and 20 **PM/MM** pairs, as a function of the sequence complexity of the gene - *New design uses **GC bin** design instead of MM*

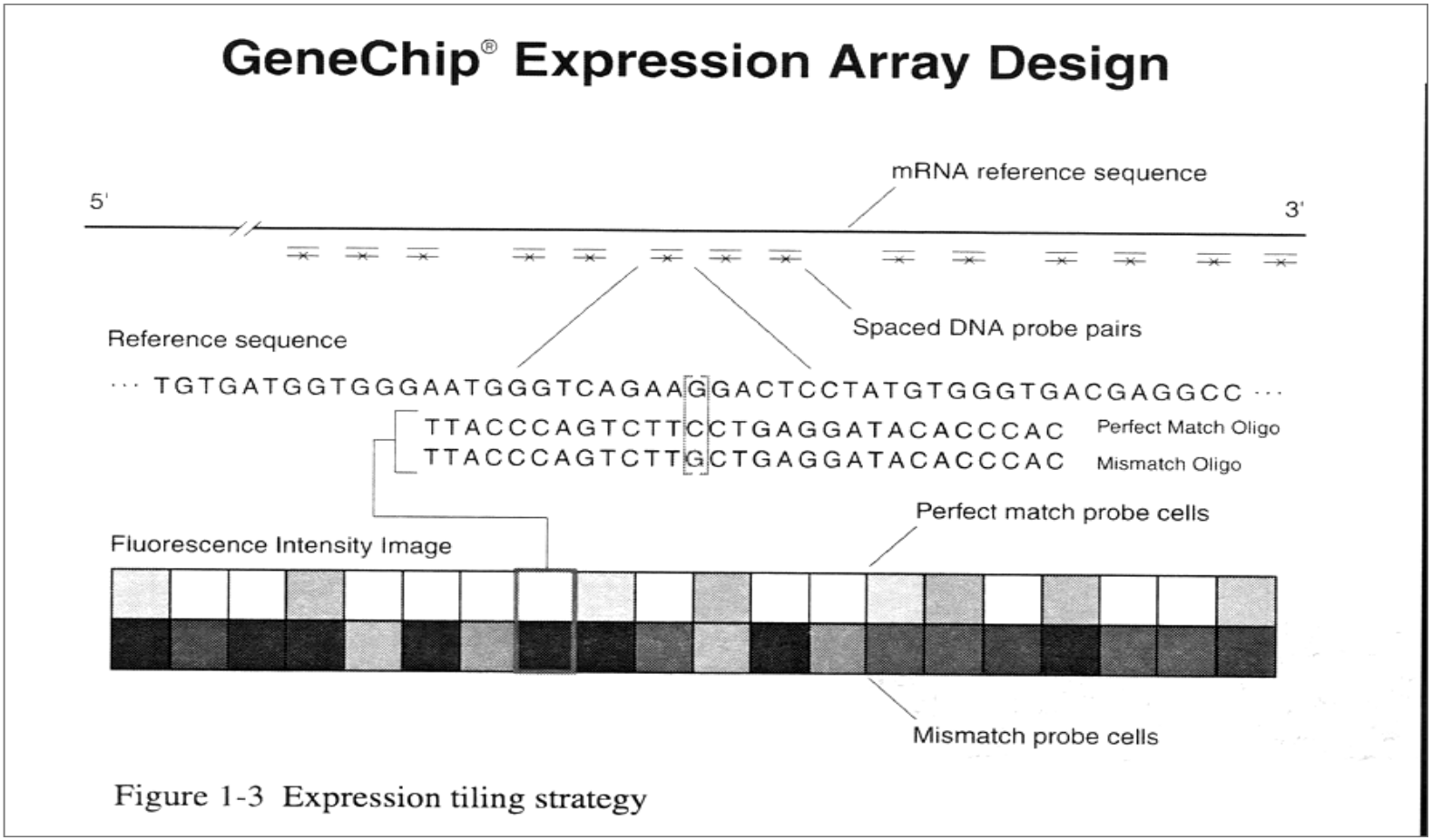


Figure 1-3 Expression tiling strategy

Affymetrix expression analysis process

QuickTime and a
decompressor
are needed to see this picture.

Exon arrays contain up to four probes for each putative exonic region. In addition to probes targeting each exon Supported by RefSeq mRNA evidence (core probes), Exon arrays also have probes that target exons supported solely by expressed sequence tag evidence (extended probes) or by purely computational predictions (full probes).

QuickTime and a
decompressor
are needed to see this picture.

QuickTime and a
decompressor
are needed to see this picture.

QuickTime and a
decompressor
are needed to see this picture.

QuickTime and a
decompressor
are needed to see this picture.

2007

Target RNA labelling and hybridization

QuickTime and a
decompressor
are needed to see this p

QuickTime and a
decompressor
are needed to see this picture.

QuickTime and a
decompressor
are needed to see this pic

Target RNA synthesis

QuickTime and a
decompressor
are needed to see this picture.

QuickTime and a
decompressor
are needed to see this pic

QuickTime and a
decompressor
are needed to see this p

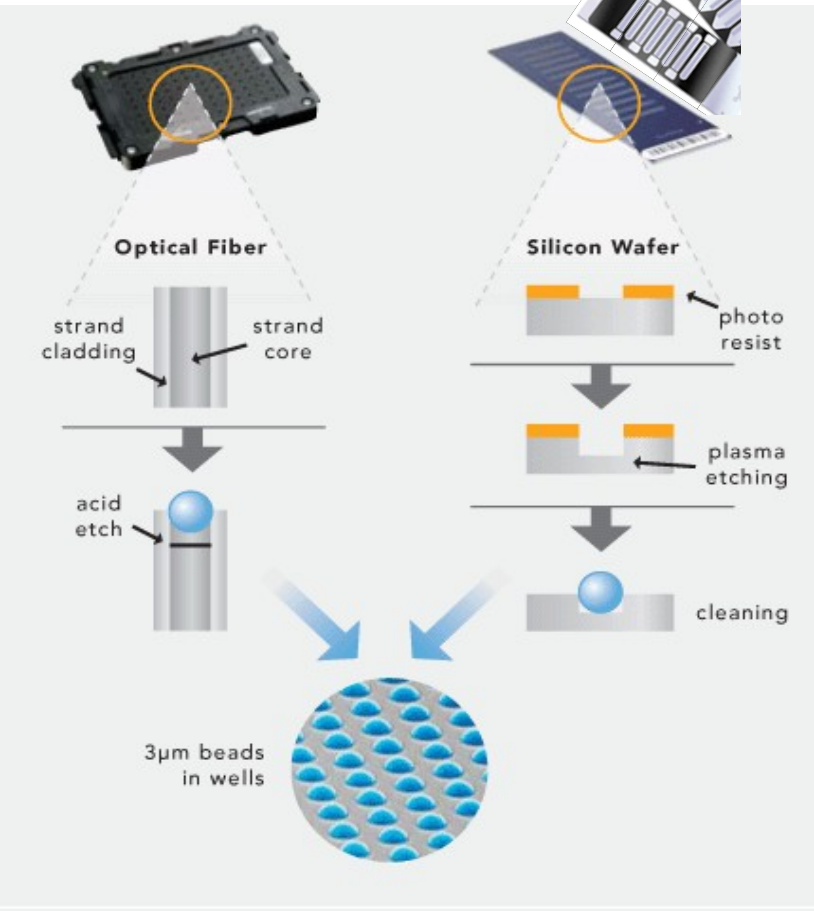
QuickTime and a
decompressor
are needed to see this picture.

QuickTime and a
decompressor
are needed to see this picture.

To fragment the single-stranded DNA dUTP is incorporated in the DNA during the second-cycle, first-strand reverse transcription reaction. This single-stranded DNA sample is then treated with a combination of uracil DNA glycosylase (UDG) and apurinic/apyrimidinic endonuclease 1 (APE 1) that specifically recognizes the unnatural dUTP residues and breaks the DNA strand.

QuickTime and a
decompressor
are needed to see this p

The **ILLUMINA** expression analysis process

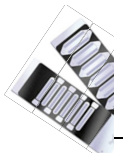


Illumina multi-sample array formats

The **96-sample Array Matrix** format is used in Illumina's DASL Gene Expression, Focused Arrays, and GoldenGate Genotyping, applications.

The **BeadChip** format is used in Illumina's gene expression arrays, DASL Gene Expression, Infinium Genotyping, and Focused Arrays applications.

The **ILLUMINA** expression analysis process



QuickTime and a
decompressor
are needed to see this picture.

QuickTime and a
decompressor
are needed to see this picture.

QuickTime and a
decompressor
are needed to see this picture.

Microarray Experimental objective

A Mouse microarray experiment examines 45,000 probe sets, representing 39,0000 transcripts including 34,000 genes and other transcripts.

Question: Which genes are the truly differentially expressed genes?

Using multiple analysis strategies helps to identify the most robust changes in gene expression and to compensate for the biases that affect each strategy.

Analysis Worflow

I. Image processing: determination of signal of a hybridization (intensity measurements) for a defined physical location on the chip.

II. Scaling/Transformation: to permit comparison of chip values, we can multiplying the signals for all genes by a common scale factor and/or fit numerical values to the assumptions of a mathematical model. The main methods we use are MAS5, dChip and RMA.

II. Testing: statistical testing for significance is done using parametric tests which assume a distribution for the values, e.g. t tests or ANOVA, or nonparametric tests, e.g. sign rank test or Kruskal Wallis test, which do not assume a distribution of values. Often an additional statistical correction for multiple testing is also performed, e.g. Bonferroni correction or Tukey test.

Overall Analysis

I. Image processing: determination of signal of a hybridization (intensity measurements) for a defined physical location on the chip.

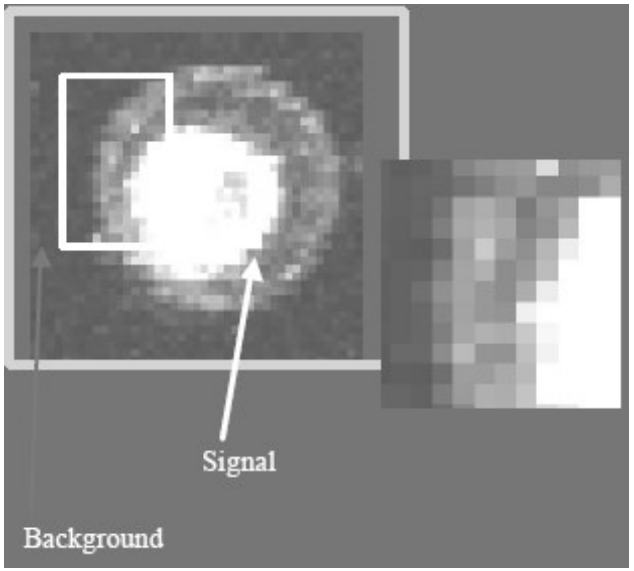
II. Scaling/Transformation: to permit comparison of chip values, we can multiplying the signals for all genes by a common scale factor and/or fit numerical values to the assumptions of a mathematical model. The main methods we use are MAS5, dChip and RMA.

II. Testing: statistical testing for significance is done using parametric tests which assume a distribution for the values, e.g. t tests or ANOVA, or nonparametric tests, e.g. sign rank test or Kruskal Wallis test, which do not assume a distribution of values. Often an additional statistical correction for multiple testing is also performed, e.g. Bonferroni correction or Tukey test.

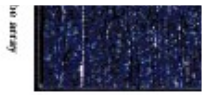
I. Image Processing

Assumption: signal intensity ~ mRNA level

•Gridding: locate spots
•Segmentation: classify pixels as signal or background.
•Measurement: for each spot of the array, calculate signal intensity (mean, median, mode) background and quality measures.



Problems with microarray image acquisition



QuickT
deco
are needed

Optimal

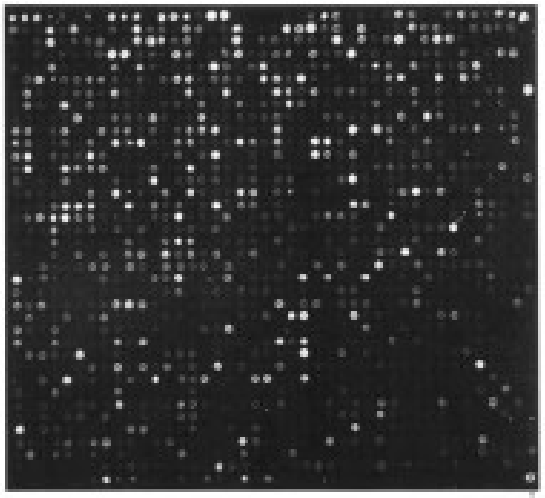


FIG. 4. Scanned image. Image from laser scanning confocal microscope of slide pictured in Fig. 3 (only one quadrant shown). A gray value is assigned to each pixel on the basis of the observed fluorescence intensity (black represents no fluorescence).

problematic

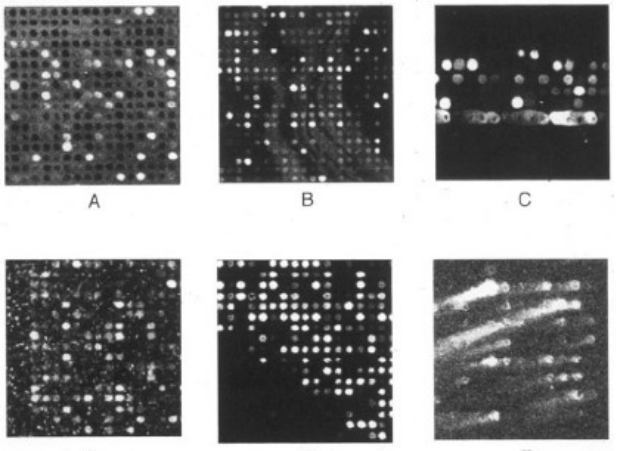


FIG. 5. Array problems: Representative problems: (A) antiprobe; (B) locally high background; (C) spot overlap; (D) precipitate; (E) locally low signal; and (F) comet tails.

From: Methods in Enzymology 1999, vol 303 [11] and [12]

Overall Analysis

I. Image processing: determination of signal of a hybridization (intensity measurements) for a defined physical location on the chip.

II. Scaling/Transformation: to permit comparison of chip values, we can multiplying the signals for all genes by a common scale factor and/or fit numerical values to the assumptions of a mathematical model. The main methods we use are MAS5, dChip and RMA.

II. Testing: statistical testing for significance is done using parametric tests which assume a distribution for the values, e.g. t tests or ANOVA, or nonparametric tests, e.g. sign rank test or Kruskal Wallis test, which do not assume a distribution of values. Often an additional statistical correction for multiple testing is also performed, e.g. Bonferroni correction or Tukey test.

II. Transformation

2 steps: normalization & fitting

- Normalization removes extraneous signal (noise) that may obscure information content but also removes some biological information.

- Therefore a balance of the negative effect versus the positive effect of normalization is required for effective subsequent high level data analysis.

- The data is then fit to an expected distribution, usually a symmetrical distribution (normal) often by taking the logarithm of the values.

- *The purpose of data transformation is to increase confidence in the high level analysis – the results of statistical testing.*

Commonly used methods

MAS5 - MicroArray Suite

provided by Affymetrix as part of the

GeneChip Operating System (GCOS)

dChip - DNA Chip Analyzer

Li & Wong 2001 PNAS 98:31-6

RMA - Robust Multiarray Averaging

Irizarry et al. 2003 NAR 31(4):e15

Comparison of Transformation Methods

MAS5 assumes the intensity value is a reflection of the efficiency of hybridization. Therefore each value is corrected by the difference between the perfect match and mismatch values. The normalization is done by chip through scaling so the correction is independent of the experiment (Scaling a chip means multiplying the signals intensity measures for all genes by a common scale factor in order to obtain the same mean intensity across the experiment).

RMA assumes quantile normalization across the entire experiment will provide more reliable values for comparison, therefore each value is adjusted by assigning corrected values by rank. Statistical testing is done by gene so the values compared now reflect the distribution across the entire experiment and by gene.

dChip Uses an invariant set normalization method, which chooses a subset of PM probes with small within-subset rank difference in the experiment, to serve as the basis for fitting a normalization curve.

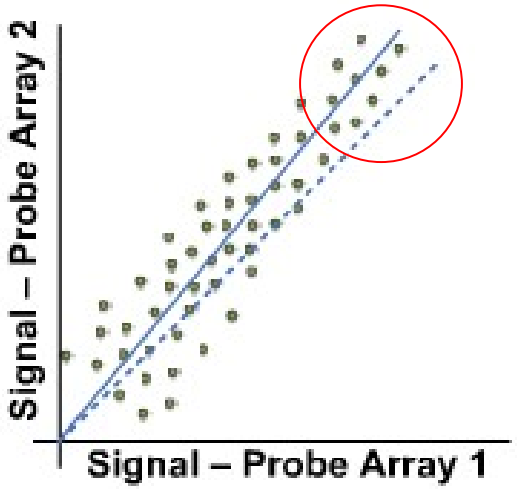
The above methods involve addition, subtraction, replacement and taking the log. Other statistical methods are considerably more complex and include: Loess, Splines, Kernel smoothing, and Support Vector Regression.

Comparison of Transformation Methods

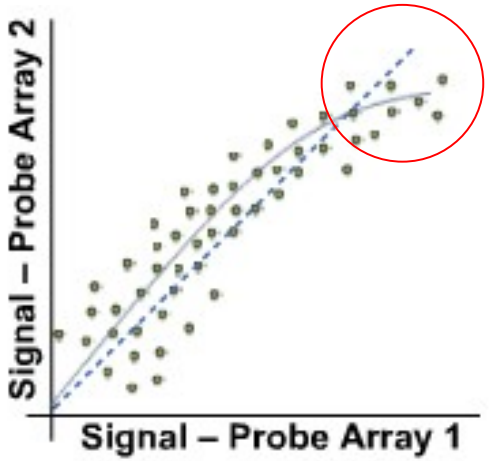
STEP\METHOD	MAS5	RMA	dChip
Normalization	Global scaling	Quantile	invariant set of genes
Probes	**PM** & **MM**	**PM** only	**PM** only
Expression	Antilog Tukey Biweight {log2 (PMij - MM*ij)}	Log2	Log2
Source	Statistical Algorithm Description Document www.affymetrix.com	Irizarry et al. 2003 NAR 31(4):e15	Li & Wong 2001 PNAS 98:31-6

Bias - I

Bias: “tendency or preference towards a particular ideology or result that interferes with the ability to be impartial, unprejudiced or objective”.

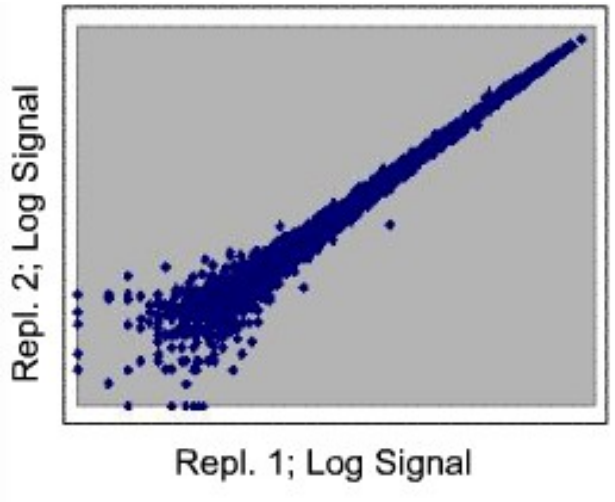


Linear bias



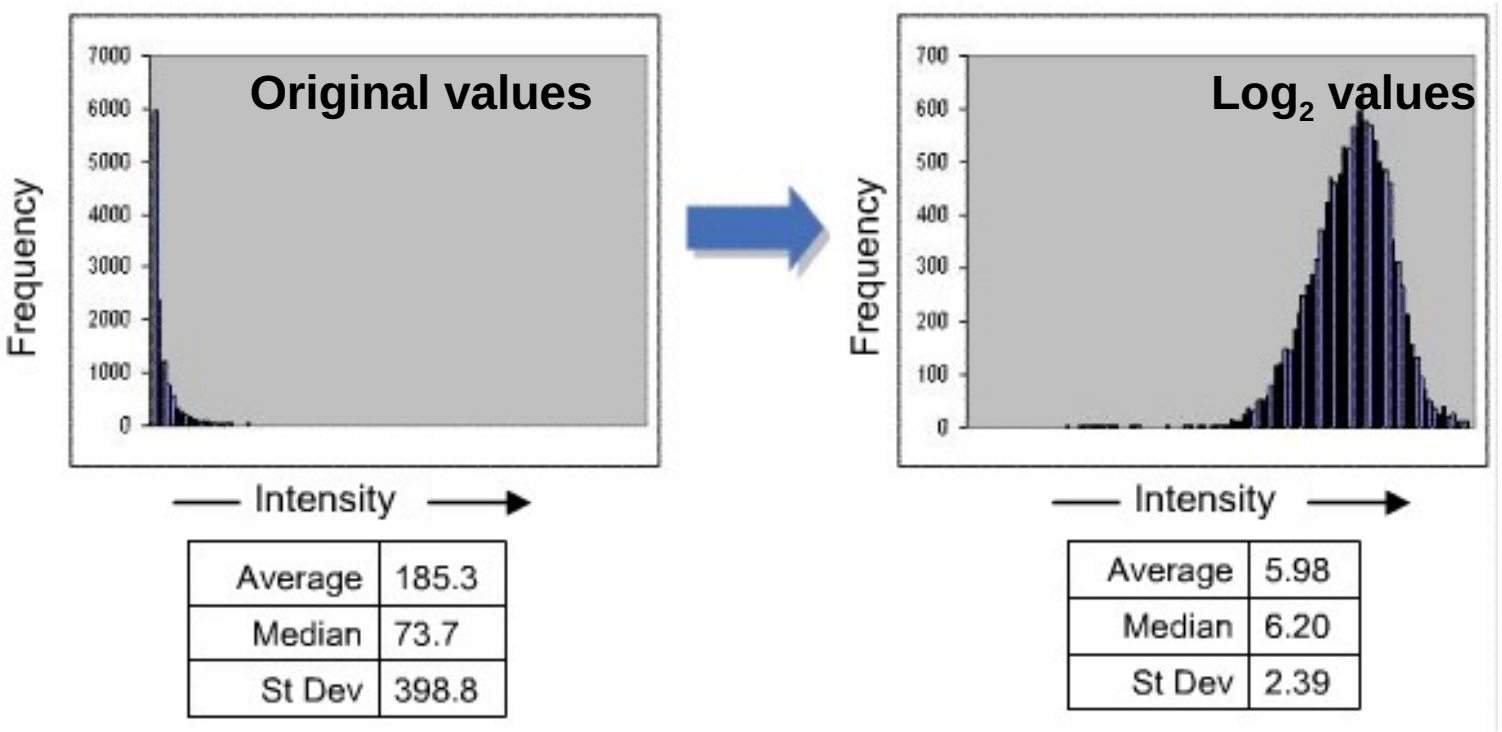
Non-Linear bias

Typical result;
Minimal bias
Intensity level-dependent variation



Effect of taking the logarithm of a value

Parametric tests (for example t test, ANOVA) assume that the data is distributed symmetrically around the mean. Taking the log maps the distribution of values closer to a gaussian shape and reduces the bias toward the lower intensity values.

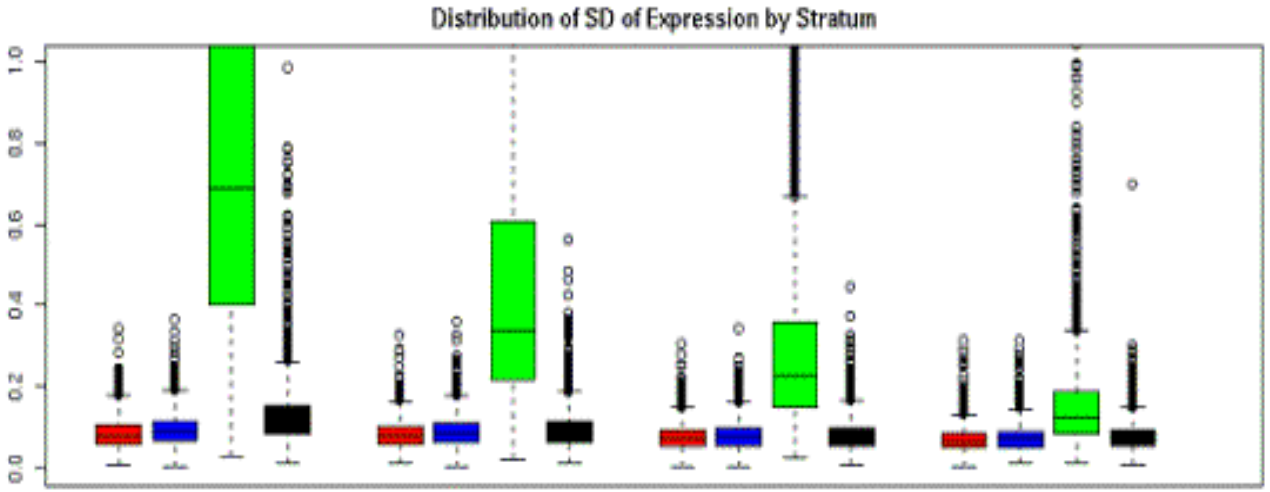


Average	185.3
Median	73.7
St Dev	398.8

Average	5.98
Median	6.20
St Dev	2.39

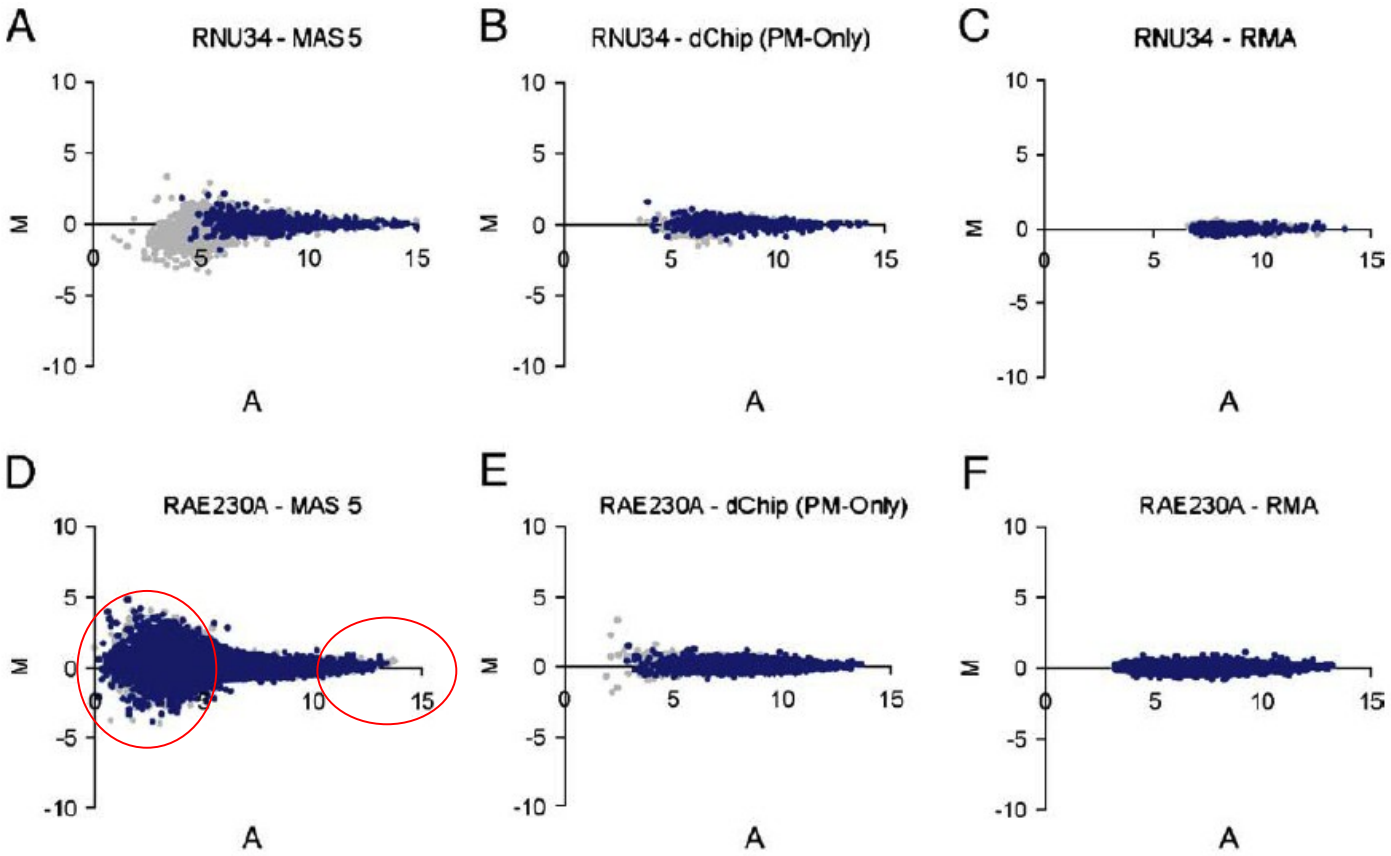
Comparison of MAS5, dChip, and RMA transformation methods

- The genes have been divided into strata (quarters) based on average expression. **Each box plot represents the standard deviation of genes in one stratum**. Note that the multi-chip models (dChip and RMA) have less variance than MAS5 on the low-abundance genes (which includes most transcription factors and signaling proteins) which is important for statistical analysis.



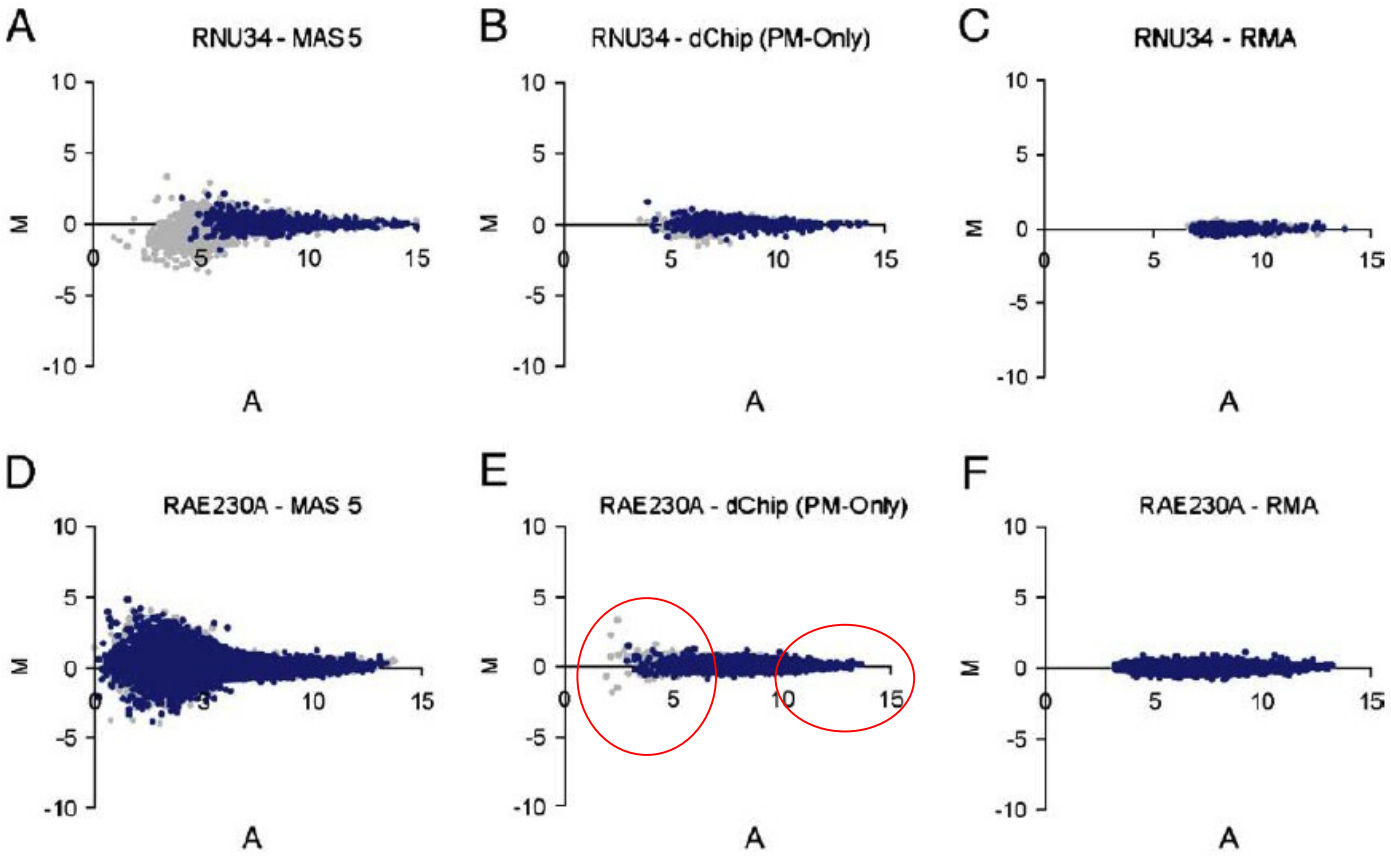
green = MAS5, black = dChip, blue= RMA, red = RMA

Bias - II



Sanna et al. 2005

Bias - II



Sanna et al. 2005

Comparison of Transformation Methods - II

We compared Affymetrix MAS4 and MAS5, which use perfect-match-minus-mismatch, and DCHIP and RMA, which use perfect-match-only models.

To validated the results, approximately 70 genes were tested by qRT-PCR in individual animals from an independent replication of the experiment.

QuickTime and a decompressor are needed to see this picture.

QuickTime and a decompressor are needed to see this picture.

a (LgA)
b (ShA)
c (control)

ESC: a $\neq$ b,c; CSA: a,b $\neq$ c

All of the genes identified as ESC genes with all four analysis strategies were confirmed by RT-PCR as such. Of the genes identified as ESC with at least two analysis strategies, 70% were confirmed by RT-PCR. Of these genes, all of the ones that were not confirmed as ESC genes proved nevertheless to be CSA genes. Of the genes identified as ESC genes with only one of the four analysis strategies, 62% were confirmed as ESC genes, with 50% of those remaining belonging to the CSA class.

Using multiple analysis strategies helps to identify the most robust changes in gene expression and to compensate for the biases that affect each strategy.

Overall Analysis

I. Image processing: determination of signal of a hybridization (intensity measurements) for a defined physical location on the chip.

II. Scaling/Transformation: to permit comparison of chip values, we can multiplying the signals for all genes by a common scale factor and/or fit numerical values to the assumptions of a mathematical model. The main methods we use are MAS5, dChip and RMA.

II. Testing: statistical testing for significance is done using parametric tests which assume a distribution for the values, e.g. t tests or ANOVA, or nonparametric tests, e.g. sign rank test or Kruskal Wallis test, which do not assume a distribution of values. **Often an additional statistical correction for multiple testing is also performed, e.g. Bonferroni correction or Tukey test.**

III - Statistical testing

t tests: The difference between the means is evaluated in terms of the sample variance as a function of the sample size – if the difference between the means is greater than the variance, the groups are statistically different.

ANOVA: Use if 3 or more group means are to be compared. The difference between the means is evaluated in terms of the sample variance as a function of the number of groups and the sample sizes – if the difference between the means is greater than the variance among all the observed values, then at least one mean is statistically different. The magnitude of the difference is expressed by the F statistic – a ratio of the sum of the differences between means divided by the sum of the differences within observations. Following the ANOVA pairwise comparisons are needed to identify the means which differ – these are multiple comparison procedures.

Multiple comparison procedures: essentially are modified t tests which correct the probability distribution for the number of comparisons. The idea is that for the first t test, the probability of error is as expected, but for the second test the probability is changed. Typically we use Fisher's LSD, Tukey HSD or Bonferroni each of which correct the t test differently.

False Discovery Rate (FDR): expected fraction of falsely identified genes in a list selected solely by statistical means.

False Positive Rate (FPR): rate at which unchanged genes appear as false positives (as changed genes).

III - Statistical testing

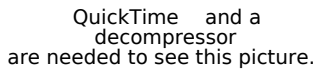
SAM (Significance Analysis of Microarrays) is a statistical technique for finding significant genes in a set of microarray experiments. SAM uses repeated permutations of the data to determine if the expression of any genes are significantly related to the group. The cut-off for significance is determined by a tuning parameter delta, chosen by the user based on the false positive rate. One can also choose a fold change parameter, to ensure that called genes change at least a pre-specified amount.

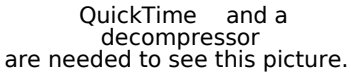
Sample size determination

Power analysis is the ability of a statistical test to reject the null hypothesis when it is false. This is dependent upon the number of measurements, the variability of the measurements, and the minimum detectible difference between the groups. **If the variance is small and the difference between means is large, the minimum sample size can be very small.**

small difference between groups

large difference between groups



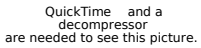


Δ/σ = difference/SD

1.4 fold change

$\alpha = p < 0.05$

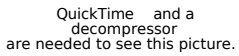
π = *proprotion of changing genes: 5, 10, 50%*



Sample size determination - II

Pooling of samples. Pooling generally reduces the amount of information. However, in microarray studies it may be used to increase power. It is only advisable for relatively homogeneous populations.

e.g., *5 pools of 5 samples afford the same power of a sample size of 20*.



Illumina

Average normalization is used to rescale intensities across multiple arrays and chips. This algorithm is an appropriate choice for experiments that employ a large number of arrays with differences in overall intensity. For average normalization, a scaling factor is calculated by dividing the average intensity of the virtual array by the average intensity for all arrays in a group. *A virtual array comprises the average values from all the arrays in the reference group and is used to determine normalization parameters* (linear - reference-based)

Rank invariant normalization, a subset of probes whose rank does not change across the experiment are identified and serve to define the normalization parameters (linear - quantile).

Cubic spline normalization is implemented to remove curvatures observed in scatter plots that arise from nonlinear relationships between samples or groups of samples when plotted in log space. This method initially divides the intensity distribution into a group of quantiles consisting of a similar number of gene intensities (nonlinear - quantile).

Illumina

linear

QuickTime and a
decompressor
are needed to see this picture.

linear

nonlinear

QuickTime and a
decompressor
are needed to see this picture.

High-throughput gene profiling in neurobiology

- Gene array technologies
- Strategy and analysis
- Applications to neurobiology

Microarray-based approaches in gene profiling

(things I learnt writing grants on microarrays)

- [Traditional hypotheses: one gene at the time]

Microarray based transcription analysis:

- Hypothesis formulation 1: Differential gene expression
- Hypothesis formulation 2: Pathway analyses
- Hypothesis formulation 3: microarray iterations to refine hypothesis
- Hypothesis formulation 4: gene regulatory relationships (“systems biology”)

Microarray-based approaches in gene profiling

- [**Traditional hypothesis: one gene at the time**]

  Pro: hypothesis-driven

  Con: hypothesis-driven

Microarray-based:

- Hypothesis formulation 1: Differential gene expression
- Hypothesis formulation 2: Pathway analyses
- Hypothesis formulation 3: microarray iterations to refine hypothesis
- Hypothesis formulation 4: gene regulatory relationships (“systems biology”)

Microarray-based approaches in gene profiling

- **Traditional hypothesis II:**

  Focused arrays (several genes at the time)

Microarray based:

- Hypothesis formulation 1: Differential gene expression
- Hypothesis formulation 2: Pathway analyses
- Hypothesis formulation 3: microarray iterations to refine hypothesis
- Hypothesis formulation 4: gene regulatory relationships (“systems biology”)

Microarray-based approaches in gene profiling

- [Traditional hypotheses: one gene at the time]

Microarray-based:

- **Hypothesis formulation 1: Differential gene expression**

  Pro: a man's reach exceeds his grasp

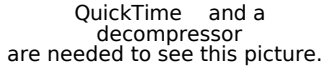
  Con: differential expression is not a correlate of functional significance

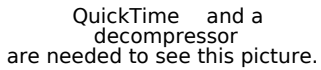
- Hypothesis formulation 2: Transcriptional signature - Pathway analyses

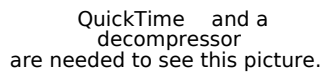
- Hypothesis formulation 3: microarray iterations to refine hypothesis

- Hypothesis formulation 4: gene regulatory relationships ("systems biology")

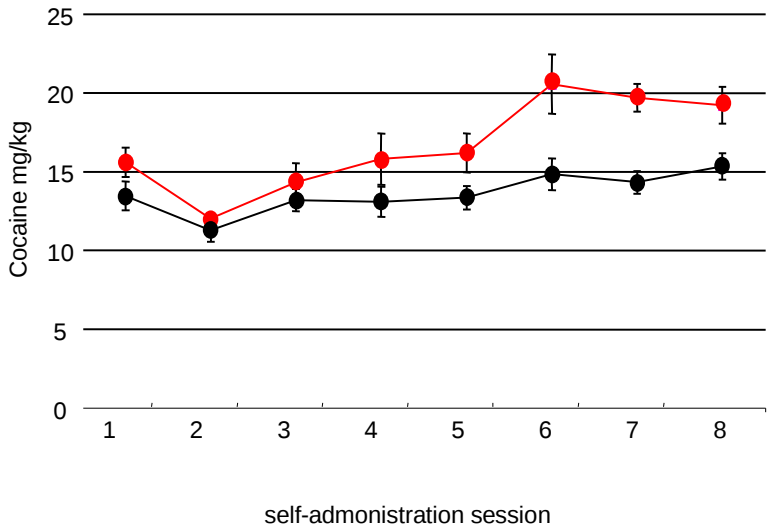
Hypothesis formulation 1:
Differential Gene Expression







Hypothesis formulation 1:
Differential Gene Expression



QuickTime and a
decompressor
are needed to see this picture.

Microarray-based approaches in basic science

- [Traditional hypotheses: one gene at the time]

Microarray-based:

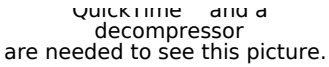
- Hypothesis formulation 1: Differential gene expression
- **Hypothesis formulation 2: Transcriptional signature - Pathway analyses:**

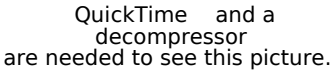
  Pro: bigger picture

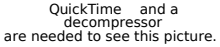
  Con: partial picture
- Hypothesis formulation 3: microarray iterations to refine hypothesis
- Hypothesis formulation 4: gene regulatory relationships (“systems biology”)

Hypothesis formulation 2:

Transcriptional signature







Hypothesis formulation 2:
Transcriptional signature

Molecular Psychiatry (2007) 12, 167–189. doi:10.1038/sj.mp.4001897

Region-specific transcriptional changes following the three antidepressant treatments electro convulsive therapy, sleep deprivation and fluoxetine

B Conti[1], R Maier[2], A M Barr[3], M C Morale[1], X Lu[1], P P Sanna[1], G Bilbe[2], D Hoyer[2] and T Bartfai[1].
1. Molecular and Integrative Neuroscience Department, The Harold L Dorris Neurological Research Institute, The Scripps Research Institute, La Jolla, CA, USA 2.Neuroscience Research, Novartis Institutes for Biomedical Research, Basel, Switzerland

In an attempt to identify common underlying mechanisms for fast- and slow-acting antidepressant modalities, we have examined the transcriptional changes in seven different brain regions of the rat brain induced by three clinically effective antidepressant treatments: electro convulsive therapy (ECT), sleep deprivation (SD), and fluoxetine (FLX), the most commonly used slow-onset antidepressant using the Affymetrix rat genome microarray 230 2.0. The gene chip data were validated using in situ hybridization or autoradiography for selected genes. The major findings of the study are:

1. The transcriptional changes induced by SD, ECT and SSRI display a regionally specific distribution distinct to each treatment.
2. **The fast-onset, short-lived antidepressant treatments ECT and SD evoked transcriptional changes primarily in the catecholaminergic system, whereas the slow-onset antidepressant FLX treatment evoked transcriptional changes in the serotonergic system.**
3. ECT and SD affect in a similar manner the same brain regions, primarily the locus coeruleus, whereas the effects of FLX were primarily in the dorsal raphe and hypothalamus, suggesting that both different regions and pathways account for fast onset but short lasting effects as compared to slow-onset but long-lasting effects. However, the similarity between effects of ECT and SD is somewhat confounded by the fact that the two treatments appear to regulate a number of transcripts in an opposite manner.
4. Multiple transcripts (e.g. brain-derived neurotrophic factor (BDNF), serum/glucocorticoid-regulated kinase (Sgk1)), whose level was reported to be affected by antidepressants or behavioral manipulations, were also found to be regulated by the treatments used in the present study. Several novel findings of transcriptional regulation upon one, two or all three treatments were made, for the latter we highlight homer, erg2, HSP27, the proto oncogene ret, sulfotransferase family 1A (Sult1a1), glycerol 3-phosphate dehydrogenase (GPD3), the orphan receptor G protein-coupled receptor 88 (GPR88) and a large number of expressed sequence tags (ESTs).
5. Transcripts encoding proteins involved in synaptic plasticity in the hippocampus were strongly affected by ECT and SD, but not by FLX.

Hypothesis formulation 2:
Pathway analyses

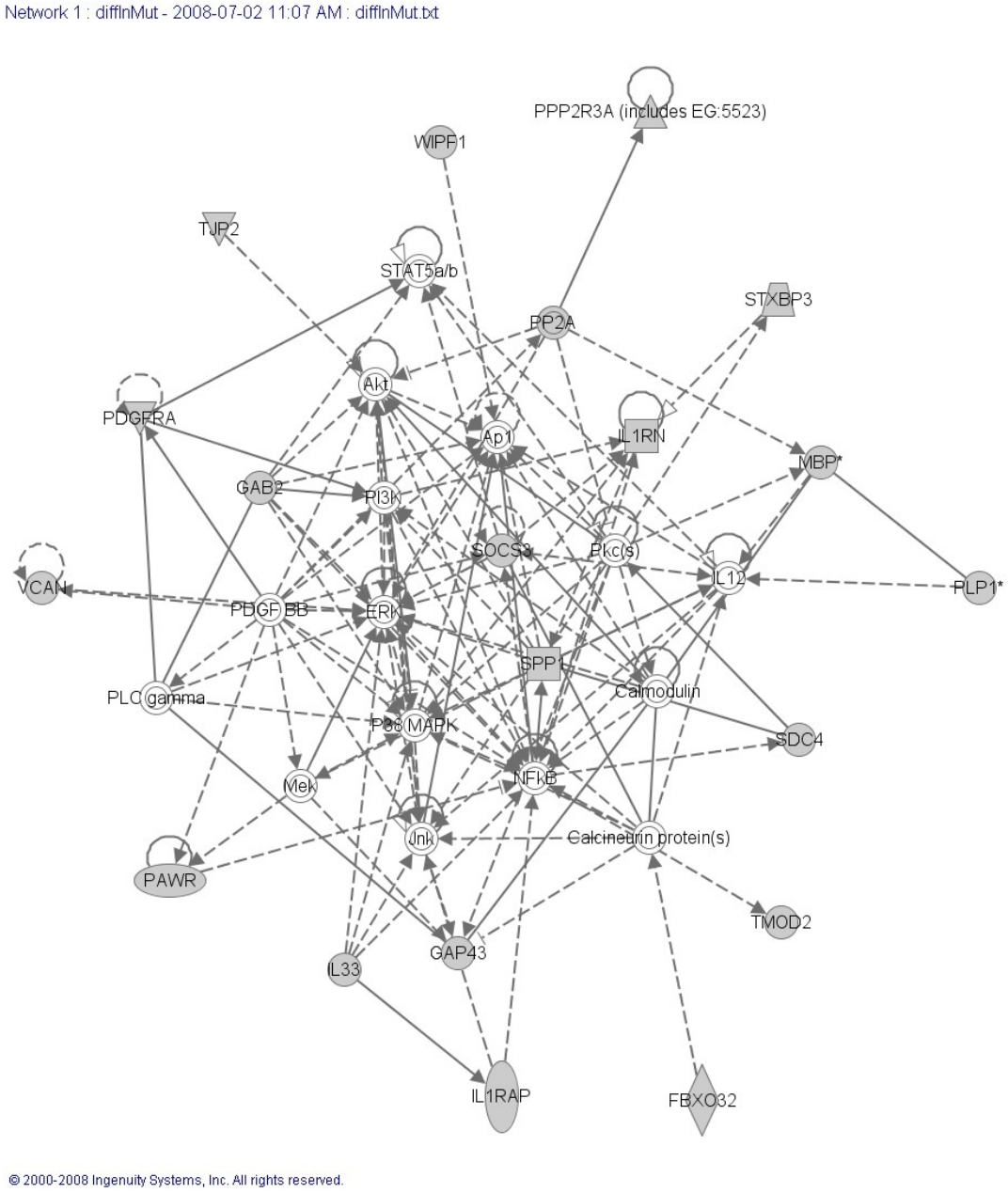
Ingenuity pathway analysis
GeneGo
Pathway studio
GSEA

Create visual representations of differentially expressed genes based on current scientific literature

Hypothesis formulation 2: Pathway analyses

Ingenuity Pathway Analysis (IPA), an integrated systems biology database broadly covering gene regulation, signal transduction, protein-protein interactions, cellular component, tissue, organ, small molecule and relationship with human disease. The IPA is structured and context-based and centers around the Ingenuity Pathways Knowledge Base (IPKB), which is based on manually curated scientific literature (http://www.ingenuity.com). Uploading a gene list or a dataset (genomic and/or proteomic experiments) into IPA will return an analysis of the molecular interactions, functions, and pathways relevant to that list of genes. Genes may be entered to explore their known interactions as well as the knowledge around a biological model. For pathway analyses the a significance threshold can be set.

The main limitation of the IPA is that it is relies exclusively on existing published sources.

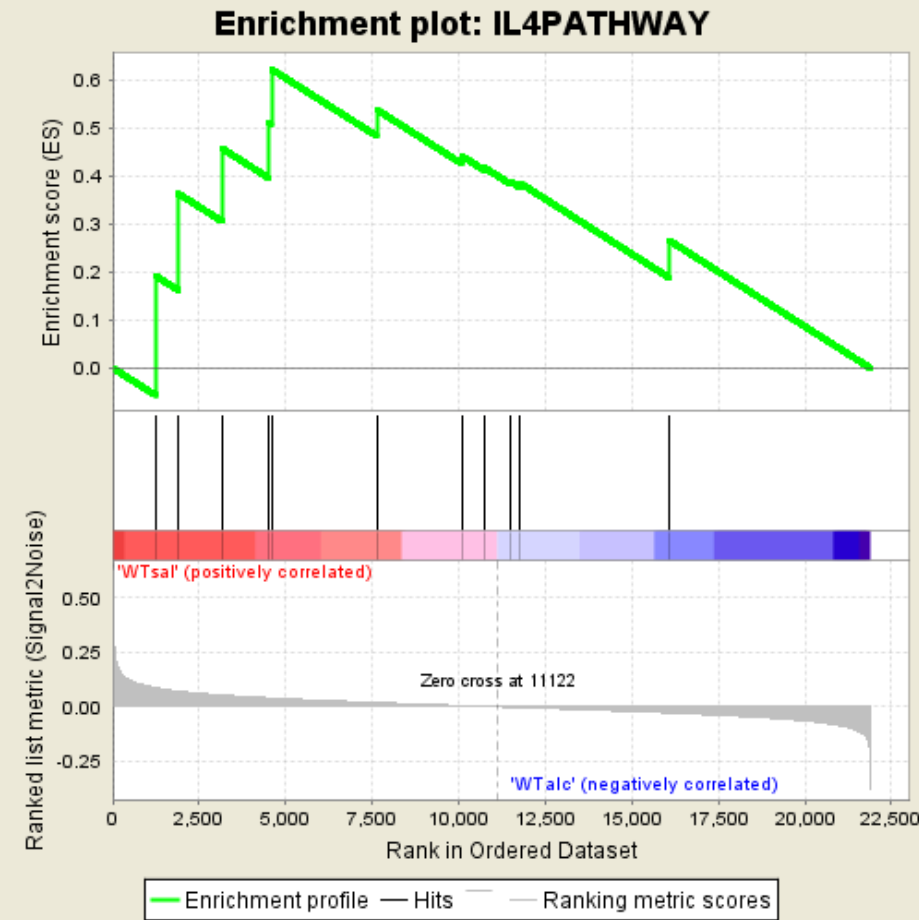


Hypothesis formulation 2: Pathway analyses

Gene Set Enrichment Analysis (GSEA)

GSEA is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

GSEA software uses gene sets from BioCarta, KEGG, GO (Gene Ontology), and the curated gene sets from the Molecular Signature Database among others. (http://www.broad.mit.edu/gsea/msigdb/index.jsp)



Microarray-based approaches in basic science

- [Traditional hypotheses: one gene at the time]

Microarray-based:

- Hypothesis formulation 1: Differential gene expression
- Hypothesis formulation 2: Transcriptional signature -Pathway analyses
- **Hypothesis formulation 3: microarray iterations to refine hypothesis**

  Pro: target identifications, pathway analysis

  Con: basic hierarchical relationships
- Hypothesis formulation 4: gene regulatory relationships (“systems biology”)

Microarray-based approaches in basic science

Hypothesis formulation 3: microarray iterations to refine hypothesis

<:3)~ →  → QuickTime and a decompressor are needed to see this picture. → <:3)~ → 

WT mouse or rat **Mutant mouse**

Microarray-based approaches in basic science

Hypothesis formulation 3: microarray iterations to refine hypothesis

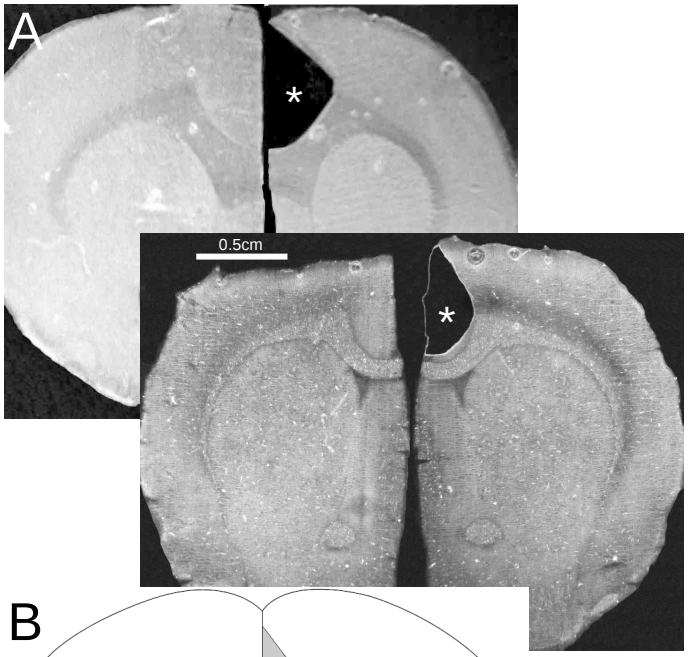
<:3)~ →  → QuickTime and a decompressor are needed to see this picture. → <:3)~ → 

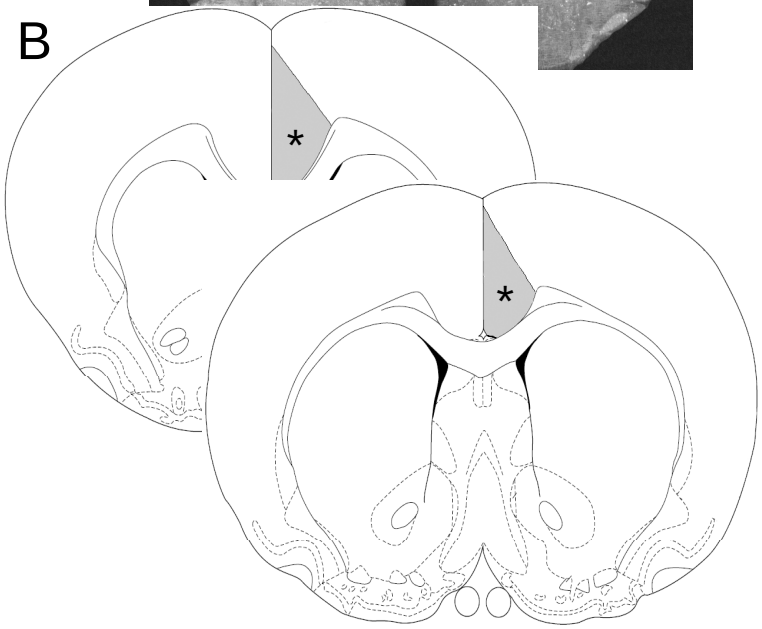
WT mouse or rat **Mutant mouse**

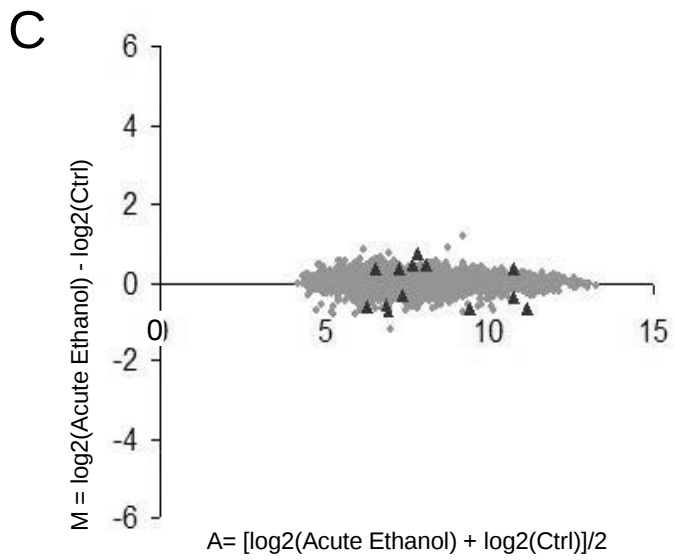
Best combined with:

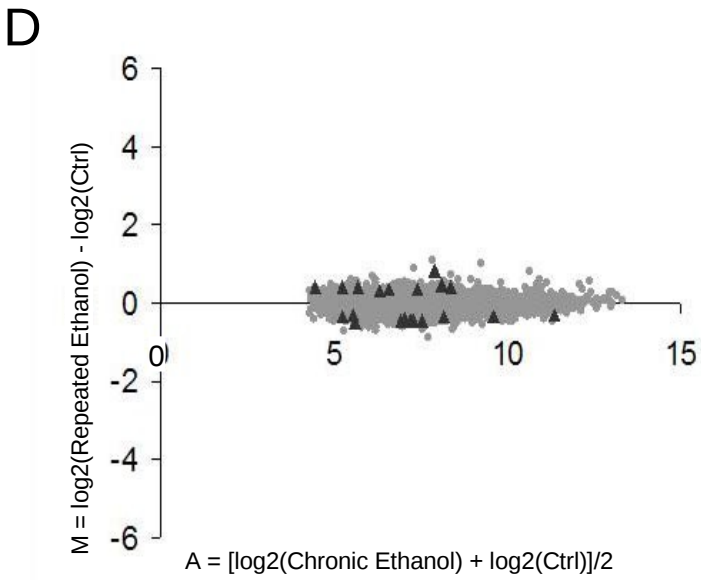
- **pathway analysis**
- **gene network analyses**
- **experimental validation**

Hypothesis formulation 3: microarray iterations to refine hypothesis

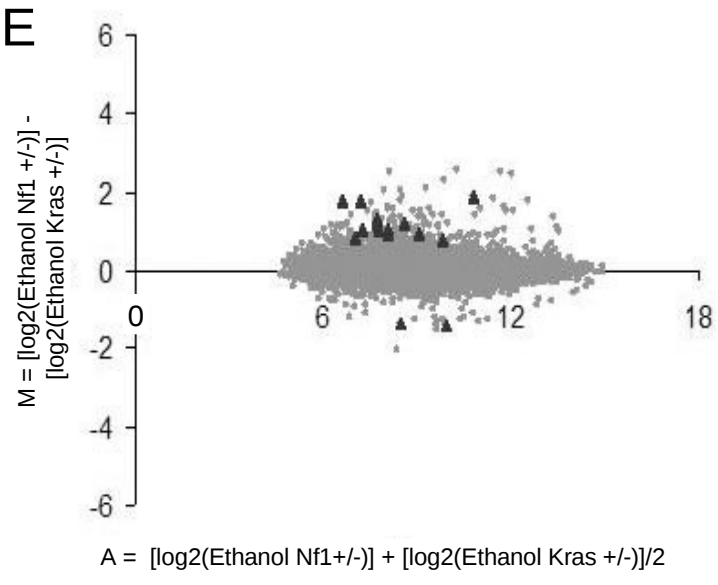


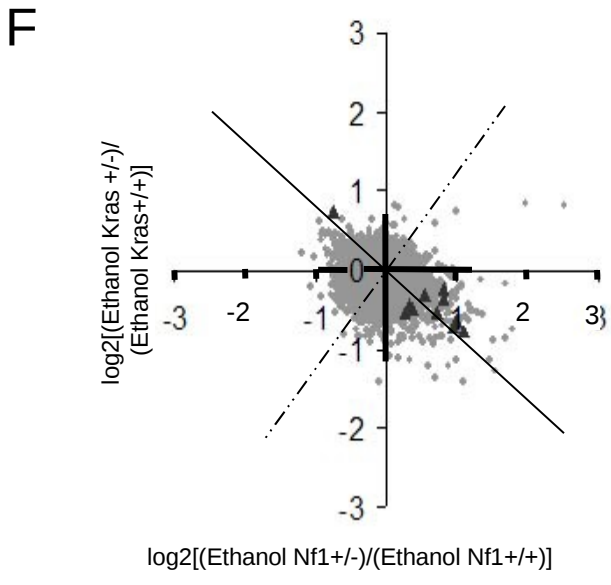


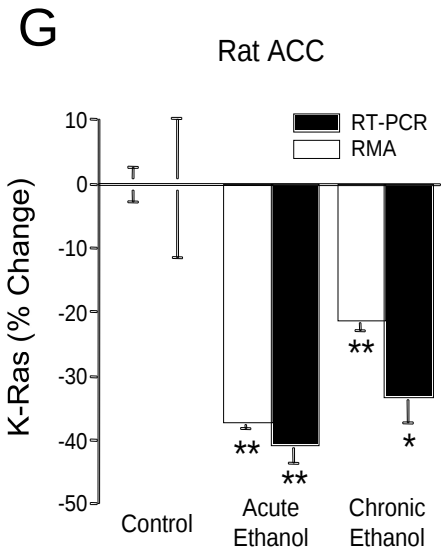


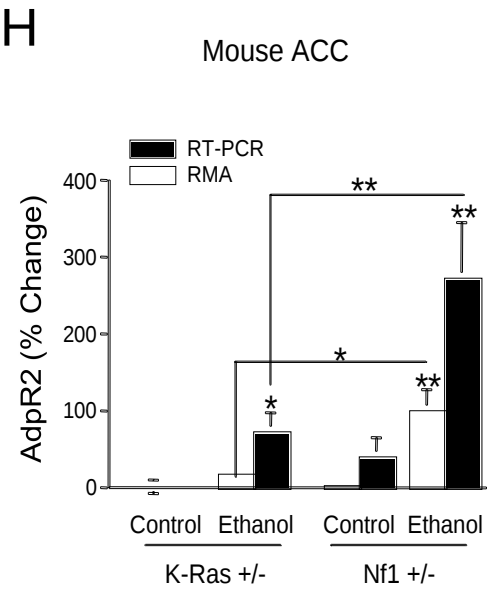


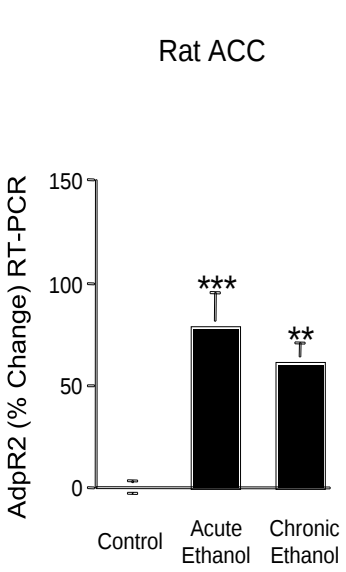
Hypothesis formulation 3: microarray iterations to refine hypothesis





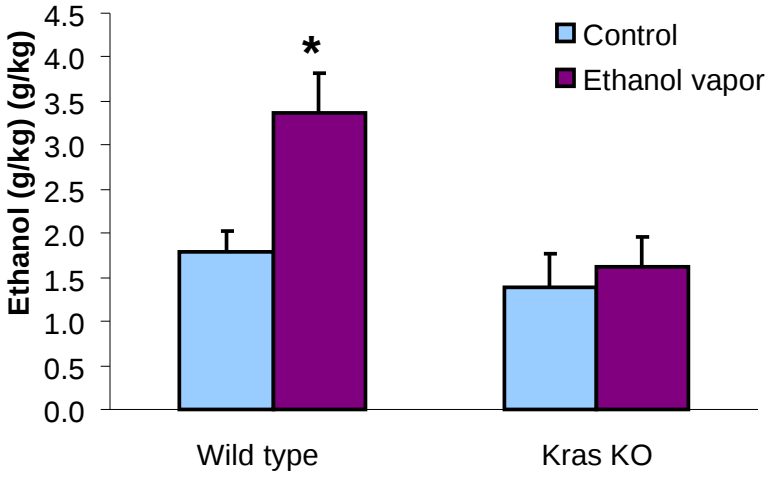


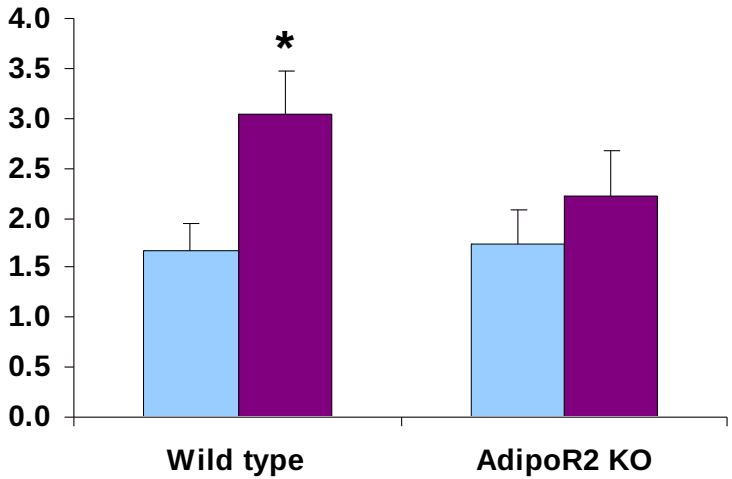




Hypothesis formulation 3: microarray iterations to refine hypothesis

The increased ethanol consumption observed in dependent mice is not present in Kras KO or AdipoR2 mice





Microarray-based approaches in gene profiling

- [Traditional hypotheses: one gene at the time]

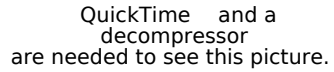
Microarray-based:

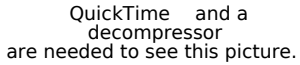
- Hypothesis formulation 1: Differential gene expression
- Hypothesis formulation 2: Transcriptional signature - Pathway analyses
- Hypothesis formulation 3: microarray iterations to refine hypothesis
- **Hypothesis formulation 4: gene regulatory relationships ("systems biology")**

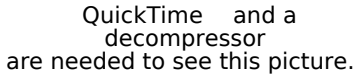
  Pro: Hierarchical relationships inferred/predicted

  Con: Predictions often not robust

Hypothesis formulation 4: gene regulatory relationships (“systems biology”)







Hypothesis formulation 4: gene regulatory relationships ("systems biology")

The systems biology process works through iterations of high-throughput analyses -> network modeling -> development of more specific hypotheses -> hypothesis testing

QuickTime and a
decompressor
are needed to see this picture.

***Caveat**: Most methods have been successful only in the study of organisms with relatively simple genomes, such as Saccharomyces cerevisiae.*

Reverse engineering strategies

four broad categories

- **optimization methods**
- **regression techniques**
- **integrative bioinformatics approaches**
- **statistical/information-theoretic methods**

Reverse engineering strategies

1. optimization methods: produce many possible networks using a relationship model (e.g. binary Boolean where the relationship is scored as present or absent, increase or decrease, etc.) and a score is computed using a function which essentially sums the number of 'correct' relationships to evaluate the network. The network with the maximum score is deemed optimal (Gat-Viks and Shamir, 2003; Liang et al., 1998)

2. regression techniques: fit the data to *a-priori* models where the relationship among the network components is estimated using a well understood mathematical model (linear, exponential, logarithmic) and used to construct the network. The constructed network is the best one using the particular model (de la Fuente et al., 2002; Gardner et al., 2003; Tegner et al., 2003).

Reverse engineering strategies
integrative bioinformatics approaches

3. integrative bioinformatics approaches: combine data from a number of independent experimental clues and maximize the amount of information extracted. This approach requires a series of experiments where **a)** a set of genes of interest is defined (pathway components) for a model system (bacteria, yeast) and an initial model is constructed of the relationships of the genes, **b)** then the set of genes is perturbed (by multiple treatments or mutations) with the effects being quantified then measured (using protein or RNA expression technologies), **c)** and all proposed models are integrated using the measured changes, **d)** followed by a formulation of hypotheses for the changes not predicted by the models, with design of new experiments to test he hypotheses and reiteration of these steps until the modeling is 'complete' (Ideker et al., 2001).

4. statistical/information-theoretic methods: rely on a variety of measures of pairwise gene corr d Kohane, 2000)

MUTUAL INFORMATION RELEVANCE NETWORKS:
FUNCTIONAL GENOMIC CLUSTERING USING
PAIRWISE ENTROPY MEASUREMENTS

A. J. BUTTE, I. S. KOHANE
Children's Hospital Informatics Program and
Division of Endocrinology,
300 Longwood Avenue,
Boston, MA 02115, USA

Reverse engineering strategies
integrative bioinformatics approaches

3. integrative bioinformatics approaches: combine data from a number of independent experimental clues and maximize the amount of information extracted. This approach requires a series of experiments where **a)** a set of genes of interest is defined (pathway components) for a model system (bacteria, yeast) and an initial model is constructed of the relationships of the genes, **b)** then the set of genes is perturbed (by multiple treatments or mutations) with the effects being quantified then measured (using protein or RNA expression technologies), **c)** and all proposed models are integrated using the measured changes, **d)** followed by a formulation of hypotheses for the changes not predicted by the models, with design of new experiments to test he hypotheses and reiteration of these steps until the modeling is 'complete' (Ideker et al., 2001).

4. statistical/information-theoretic methods: rely on a variety of measures of pairwise gene corr d Kohane, 2000)

MUTUAL INFORMATION RELEVANCE NETWORKS:
FUNCTIONAL GENOMIC CLUSTERING USING
PAIRWISE ENTROPY MEASUREMENTS

A. J. BUTTE, I. S. KOHANE
Children's Hospital Informatics Program and
Division of Endocrinology,
300 Longwood Avenue,
Boston, MA 02115, USA

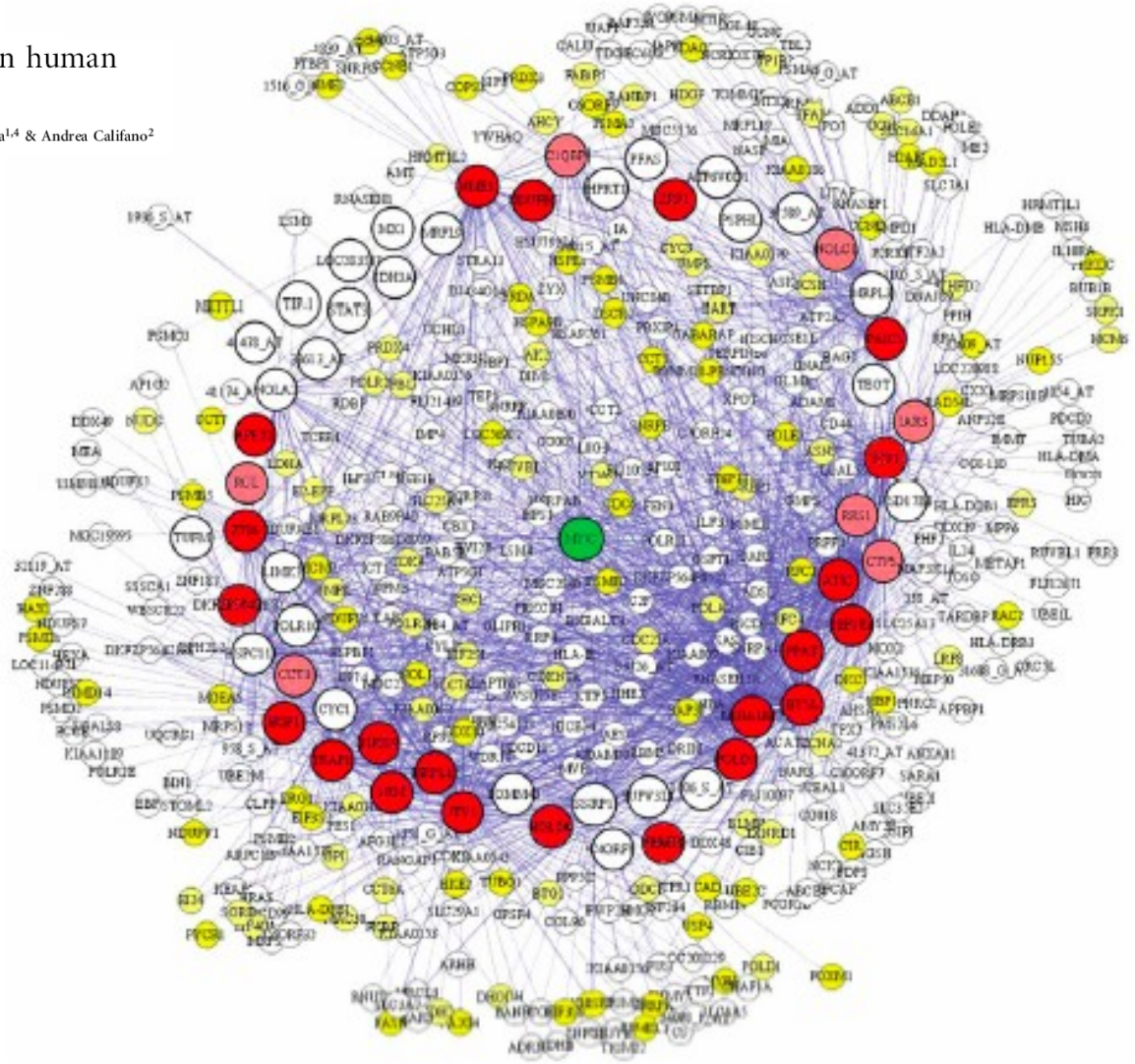
Hypothesis formulation 4: gene regulatory relationships ("systems biology")

nature genetics

Reverse engineering of regulatory networks in human B cells

Katia Basso[1], Adam A Margolin[2], Gustavo Stolovitzky[3], Ulf Klein[1], Riccardo Dalla-Favera[1,4] & Andrea Califano[2]

ARACNe (algorithm for the reconstruction of accurate cellular networks), a new approach for the reverse engineering of cellular networks from microarray expression profiles. ARACNe first identifies statistically significant gene-gene co-regulation by mutual information, an information-theoretic measure of relatedness. It then eliminates indirect relationships, in which two genes are co-regulated through one or more intermediaries, by applying the 'data processing inequality' (DPI). Hence, relationships included in the final reconstructed network have a high probability of representing either direct transcriptional regulatory interactions or interactions mediated by post-transcriptional modifiers that are undetectable from gene-expression profiles.

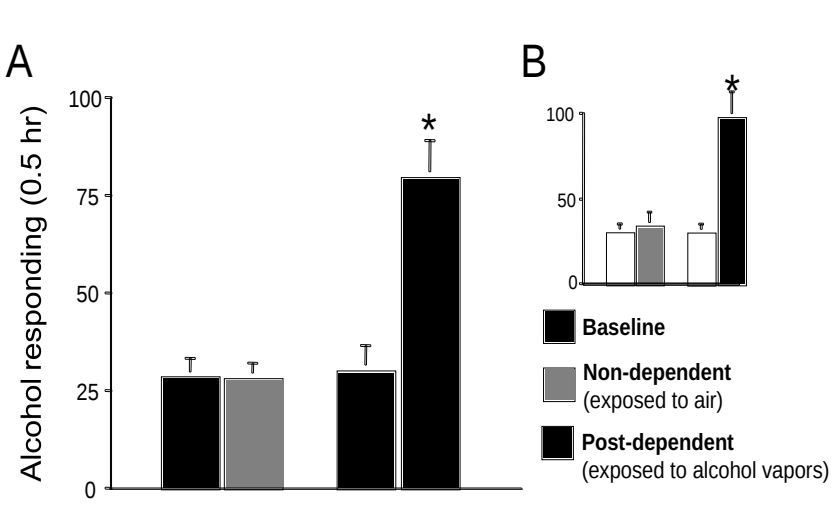


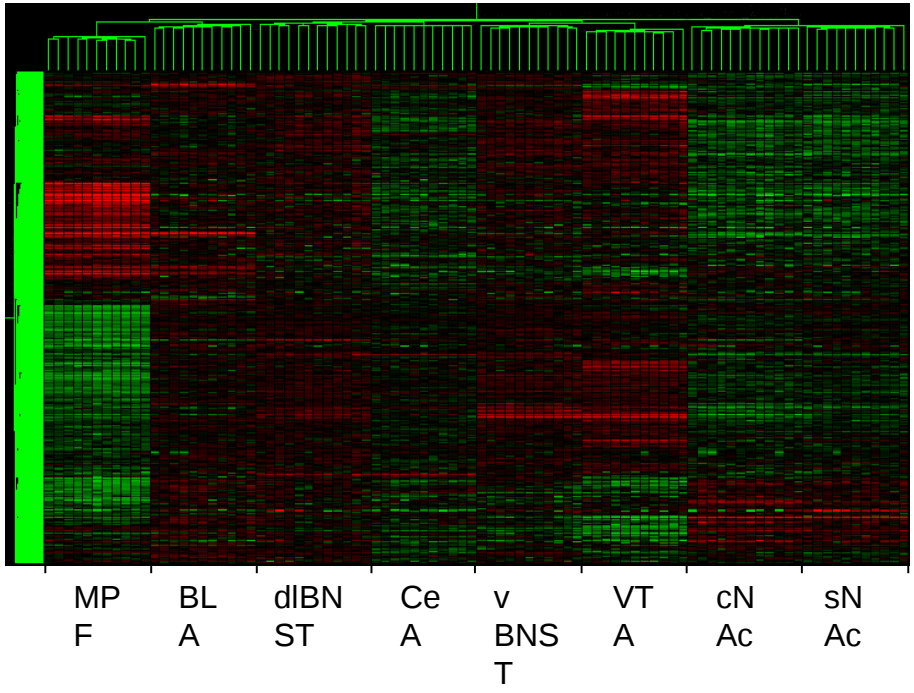
QuickTime and a
decompressor
are needed to see this pi

QuickTime and a
decompressor
are needed to see this picture.

Andrea Califano
Professor
Columbia University Medical Center
Department of Biomedical Informatics (DBMI)Institute of Cancer Genetics (ICG)
Center for Computational Biology and Bioinformatics
C2B2 (C2B2)
Director, Center for the Multiscale Analysis of Genetic Networks (MAGNet)
Associate Director for Bioinformatics, Irving Cancer Research Center (ICRC)
Co-Director, Center for Computational Biology and Bioinformatics (C2B2)

Hypothesis formulation 4: gene regulatory relationships ("systems biology")





Unsupervised hierarchical clustering of a microarray dataset of CNS samples.
MPF: medial prefrontal cortex; BLA: basolateral nucleus of the amygdala;
dlBNST and vBNST: dorsolateral and ventrolateral
bed nucleus of the stria terminalis;
CeA: central nucleus of the amygdala;
cNAc and sNAc: core and shell nucleus accumbens,
VTA: ventral tegmental area. Clustering was performed using GeneSpring GX.

Other microarrays and microarray-based strategies

- Drug effects profiling (pharmacogenomics, toxicogenomics)
  > personalized medicine, drug classification….
- Disease profiling and progression
- Diagnosis of infectious diseases
- Pathogen discovery

Microarray-based approaches in gene profiling

(things I learnt writing grants on microarrays)

- [Traditional hypotheses: one gene at the time]

Microarray based transcription analysis:

- **Hypothesis formulation 1: Differential gene expression**
- **Hypothesis formulation 2: Pathway analyses**
- **Hypothesis formulation 3: microarray iterations to refine hypothesis**
- **Hypothesis formulation 4: gene regulatory relationships (“systems biology”)**

Beyond transcription analysis: Other microarrays and microarray-based strategies

- **Tiling Arrays and Promoter arrays**

  unbiased transcription analysis

  methylation analysis

  ChIP on Chip

- **SNP genotyping arrays**

  High-throughput variation detection and genotyping using microarrays

Other microarrays and microarray-based strategies

- **Tiling Arrays and Promoter arrays**

  unbiased transcription analysis

  methylation analysis

  ChIP on Chip

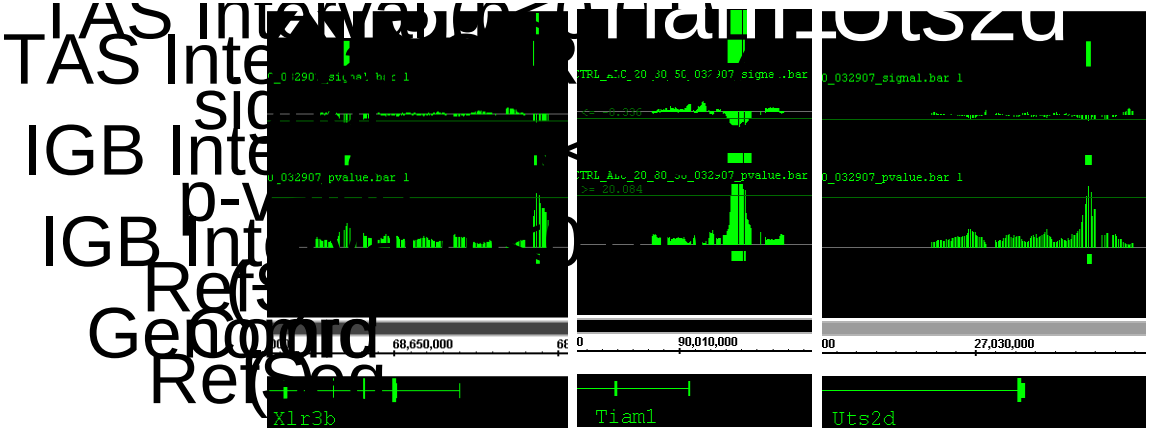
Other microarrays and microarray-based strategies

- **Tiling Arrays and Promoter arrays**

  unbiased transcription analysis

  methylation analysis

  ChIP on Chip



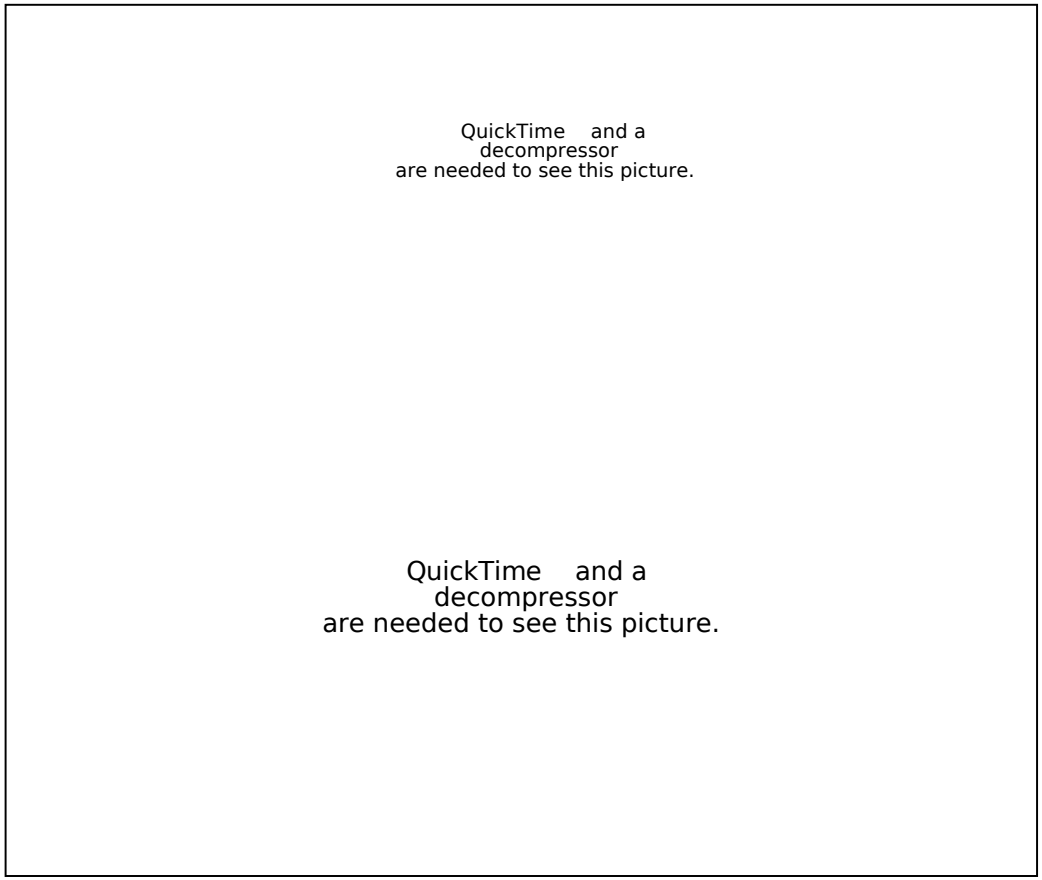
Other microarrays and microarray-based strategies

- **Tiling Arrays and Promoter arrays**

  unbiased transcription analysis

  methylation analysis

  ChIP on Chip



Are there any alternatives to microarray-based strategies for high through-put gene profiling?

QuickTime and a decompressor are needed to see this picture.

“Ultra-high-throughput sequencing is emerging as an attractive alternative to microarrays. Illumina sequencing data are highly replicable, with relatively little technical variation, and thus, for many purposes, it may suffice to sequence each mRNA sample only once (i.e., using one lane). The information in a single lane of Illumina sequencing data appears comparable to that in a single array in enabling identification of differentially expressed genes, while allowing for additional analyses such as detection of low-expressed genes, alternative splice variants, and novel transcripts. “

Other applications include genotyping, analysis of methylation patterns, and identification of transcription factor binding sites

QuickTime and a decompressor are needed to see this picture.

Conclusions

Microarrays

…are a *late arrival* in the biology toolbox that is *rapidly evolving*.

…are a versatile tools to address diverse biological questions.

…do better when combined with appropriate computational strategies.

…require specific experimental design strategies.

…are indispensable in the post-genome world (for the time being).

Conclusions

Microarrays

…are a late arrival in the biology toolbox that is rapidly evolving.

…are a *versatile tools to address diverse biological questions*.

…do better when combined with appropriate computational strategies.

…require specific experimental design strategies.

…are indispensable in the post-genome world (for the time being).

Conclusions

Microarrays

…are a late arrival in the biology toolbox that is rapidly evolving.

…are a versatile tools to address diverse biological questions.

…do better when combined with *appropriate computational strategies*.

…require specific experimental design strategies.

…are indispensable in the post-genome world (for the time being).

Conclusions

Microarrays

…are a late arrival in the biology toolbox that is rapidly evolving.

…are a versatile tools to address diverse biological questions.

…do better when combined with appropriate computational strategies.

…*require specific experimental design strategies*.

…are indispensable in the post-genome world (for the time being).

Conclusions

Microarrays

…are a late arrival in the biology toolbox that is rapidly evolving.

…are a versatile tools to address diverse biological questions.

…do better when combined with appropriate computational strategies.

…require specific experimental design strategies.

…are *indispensable* in the post-genome world (for the time being).

Conclusions

Microarrays

…are a late arrival in the biology toolbox that is rapidly evolving.

…are a versatile tools to address diverse biological questions.

…do better when combined with appropriate computational strategies.

…require specific experimental design strategies.

…are *indispensable* in the post-genome world (for the time being).

Conclusions

Microarrays

…are a late arrival in the biology toolbox that is rapidly evolving.

…are a versatile tools to address diverse biological questions.

…do better when combined with appropriate computational strategies.

…require specific experimental design strategies.

…are *indispensable* in the post-genome world *(for the time being).*

High-throughput gene profiling in neurobiology

Pietro Paolo Sanna