Harvard-MIT Division of Health Sciences and Technology
HST.512: Genomic Medicine
Prof. Zoltan Szallasi

Children's Hospital Informatics Program Harvard Medical School

Limitations of massively parallel technologies

Zoltan Szallasi, MD

Children’s Hospital Informatics Program
www.chip.org

New technology

↓

All problems will be solved within a couple of years

↓

Realistic Expectations (limitations)

Limitations: (you want to make predictions)

Accuracy – noise

Sensitivity - completeness

Inherent limitations –
(think about unpredictability > chaos)

NOISE:

- what is noise ? (and what is signal ?)
- noise as an inherent feature of complex systems
- noise in continuous and discrete measurements
- noise as the limitation of the technology
- what can be done about noise ?
    - Statistics
    - Normalization as a way to deal with systematic errors

c : an unwanted signal or a disturbance (as static or a variation of voltage) in an electronic device or instrument (as radio or television); *broadly* **:**
a disturbance interfering with the operation of a usually mechanical device or system
d : electromagnetic radiation (as light or radio waves) that is composed of several frequencies and that involves random changes in frequency or amplitude
e : irrelevant or meaningless data or output occurring along with desired information

Noise may turn out to be an important signal !!!!

-Penzias and Wilson >>> cosmic background radiation

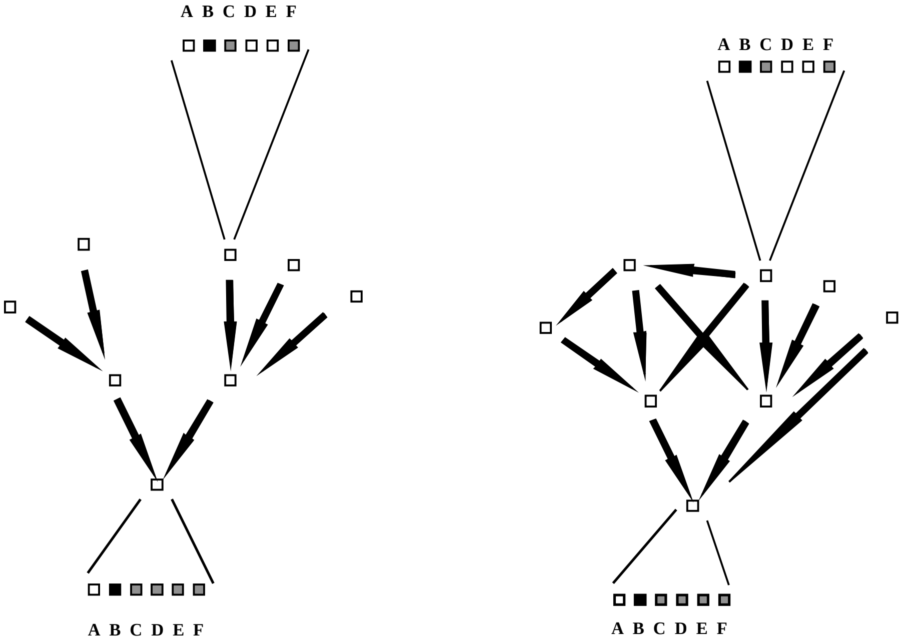
- discovery of the chemotherapeutic agent cis-platinum

What we perceive as noise/error might be a key component of biological processes:

1) Mutations in evolution
2) “Junk” DNA
3) Asymmetric cell division may contribute to differentiation
4) Stochastic fluctuations may be important for the stability of complex physicochemical systems

Genetic networks are stochastic systems:

1) A couple of hundred copies of a given transcription factor/nucleus
2) Intracellular environment is the not a free solution
3) Reaction kinetics is often slow etc.



Please see Science. 2002 Aug 16; 297(5584):1183-6.

Comment in:

Science. 2002 Aug 16; 297(5584):1129-31.

Stochastic gene expression in a single cell.

Elowitz MB, Levine AJ, Siggia ED, Swain PS.

-measuring population averaged data.

That is true even if single cells are quantified due to stochasticity > two cells can get from a given state to another one via different paths

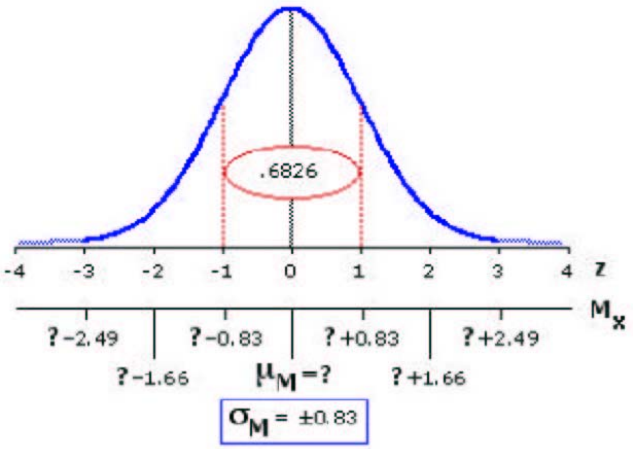
Noise in measurements

There is no measurement without noise - (it is the accuracy/sensitivity of your measurement that is low)

For continuous variables it is expected to obtain data with a certain “spread”

Consequently: Statistics was invented

- 0.5, -0.3, 0.2, 1.4, -1.5…..etc what is the true value of the observed variable ?
- Did the variable change due to a given treatment? Etc.



Lots of measurements and/or fairly good idea about the nature of the noise (e.g. normal distribution)

Statistical analysis in biology:

1) What is the true value of a given parameter ?

2) the most common analysis – Bayesian

3) You don’t believe the measurements >> normalization

4) There are too many numbers >> permutation etc.

Biological measurements are often expensive !!!!!!!!!

A large number of papers relating to cancer were published in Nature/Science ….. based on single microarray measurements

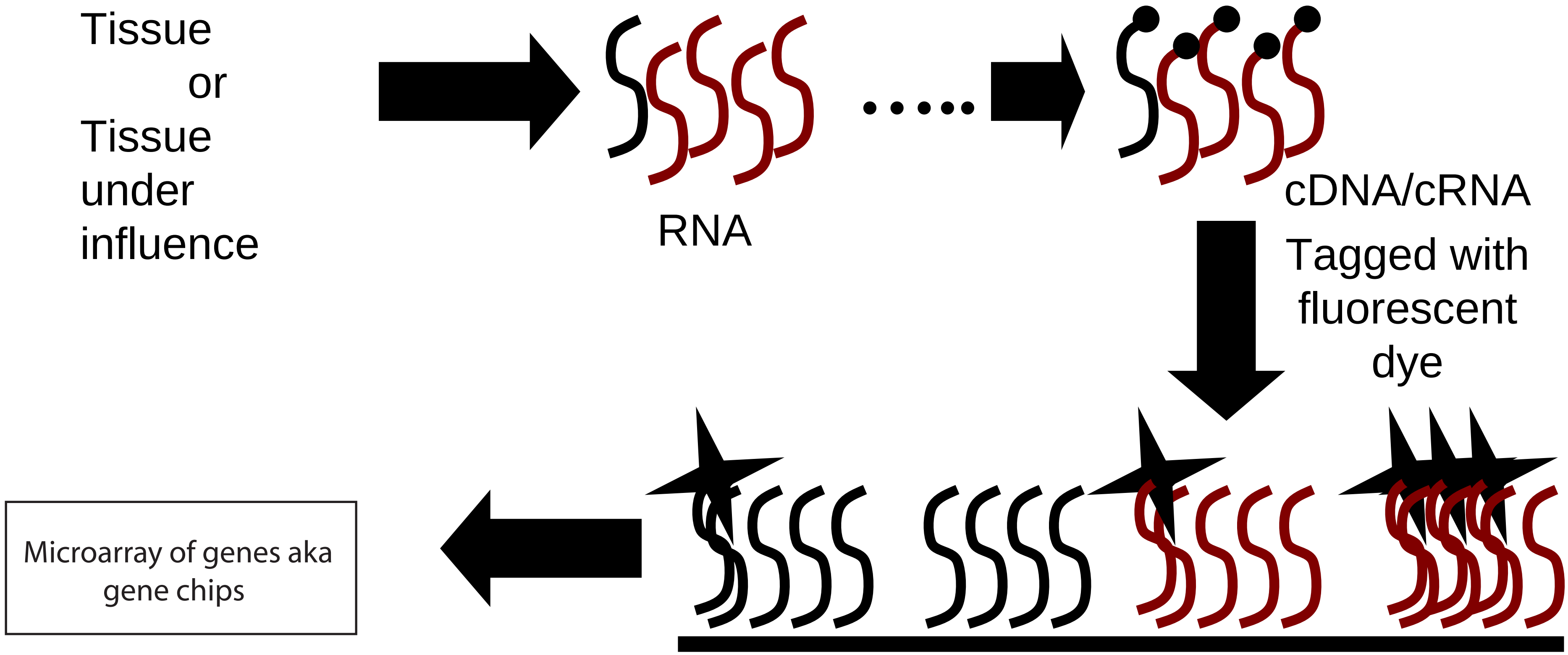
~~STATISTICS~~

Reliable numbers cannot be produced without replicates

The central problem :

In massively parallel biological measurements quantitative or qualitative calls are supposed to be made on a large number of heterogeneous variables using only a few replicates.

Noise of continuous variables, e.g. microarray measurements



Tissue
or
Tissue
under
influence

RNA

cDNA/cRNA

Tagged with
fluorescent
dye

Microarray of genes aka
gene chips

Ideally: 1 copy of a given RNA will produce 1 unit of a specific signal !!!!!!!!!!!!!!

1) cDNA produced from RNA (initiation of RT step, RT might drop off etc.)

2) cRNA produced in the presence of fluorescent dyes (cRNA production in not linear, Dye incorporation)

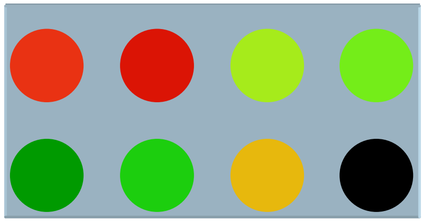
3) Breaking down cRNA into small pieces

4) hybridization/cross hybridization

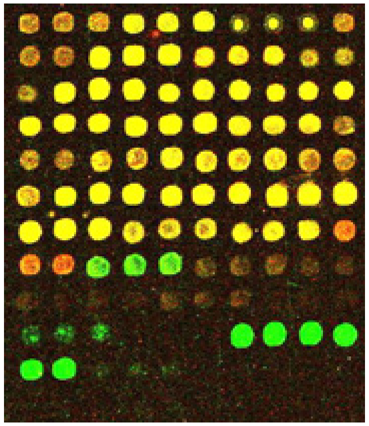
$$\text{final signal} = \sum (\text{all of the above})$$

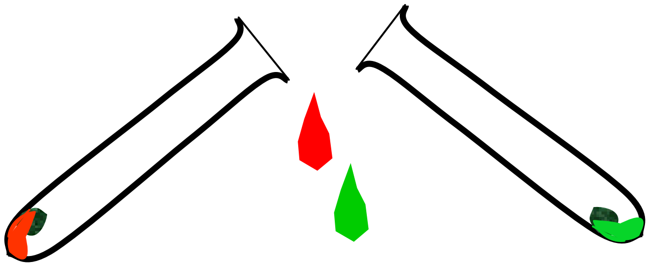
The situation is further complicated by other experimental issues >>> two-color cDNA microarray

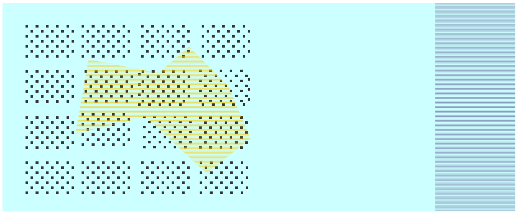
Ratio is influenced on background calculations



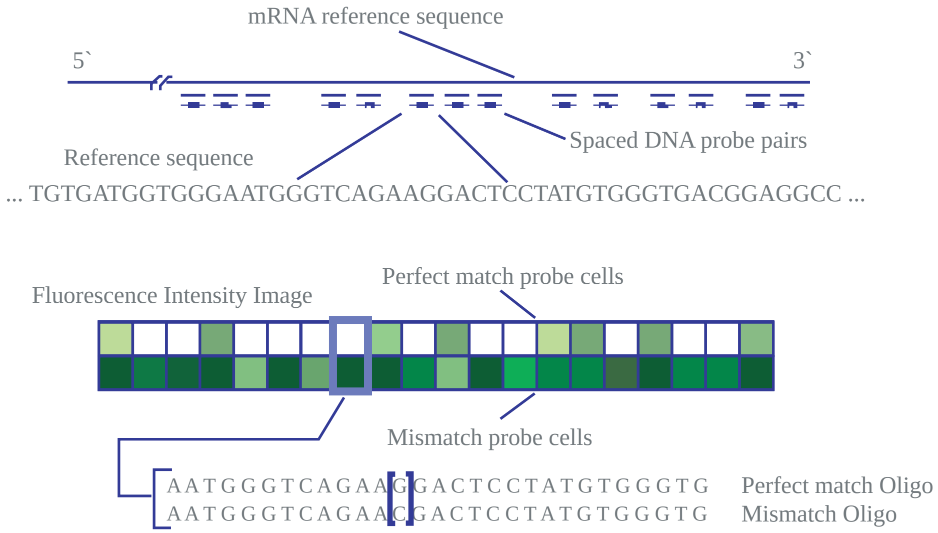
equal amounts of labelled cDNA samples







There is no truly blank spot !!!! Background



mRNA reference sequence

5`

3`

Spaced DNA probe pairs

Reference sequence

... TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGGAGGCC ...

Fluorescence Intensity Image

Perfect match probe cells

Mismatch probe cells

AATGGGTCAGAAGGACTCCTATGTGGGTG Perfect match Oligo

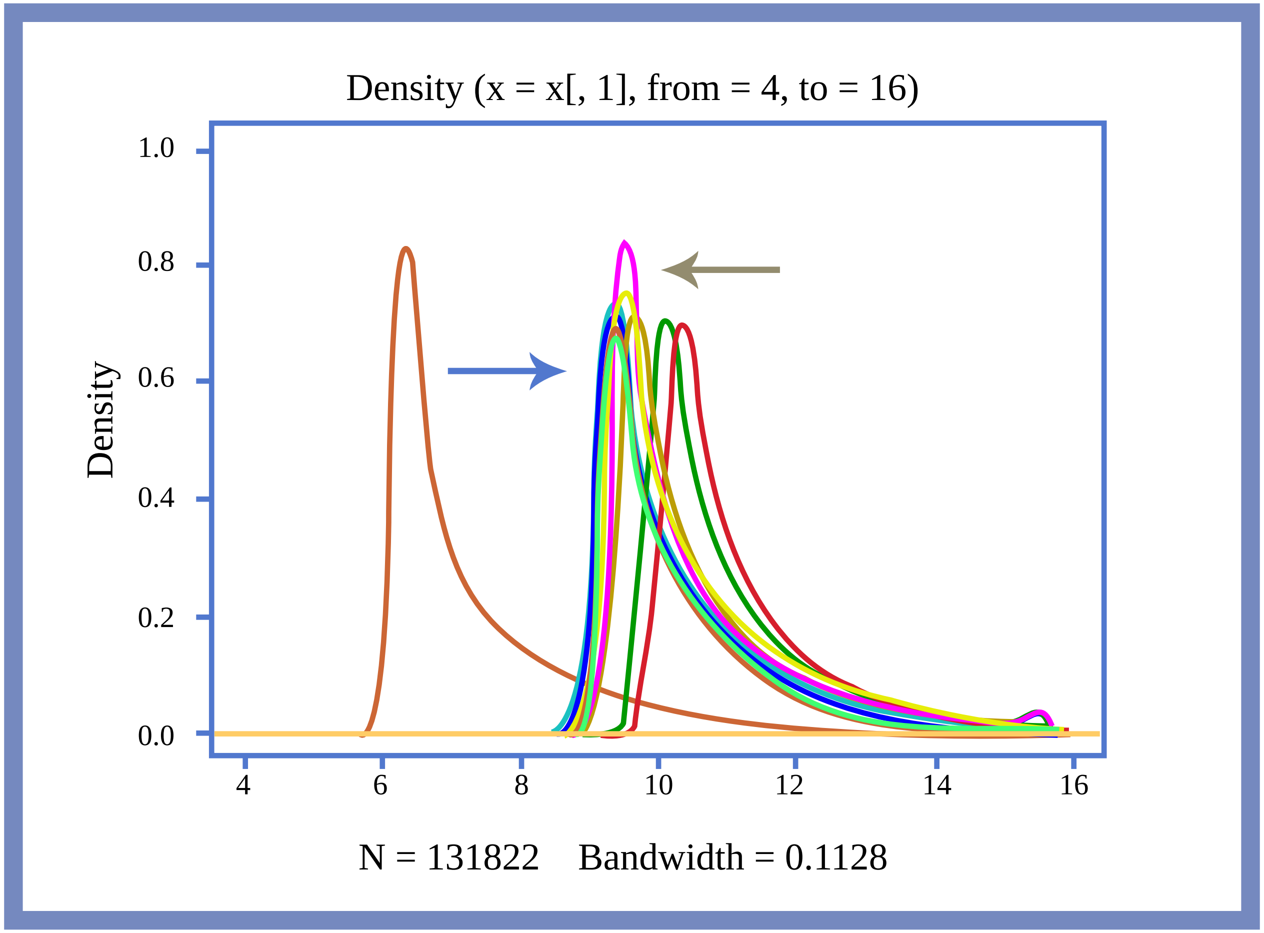
AATGGGTCAGAACGACTCCTATGTGGGTG Mismatch Oligo

Data representation

If we express gene expression measurements as “per unit RNA” then decrease in the level of a given message unavoidably leads to a relative increase in the level of other messages.

Distribution of probe intensities of several Affymetrix data sets belonging to the same set of experiment.

Systematic error



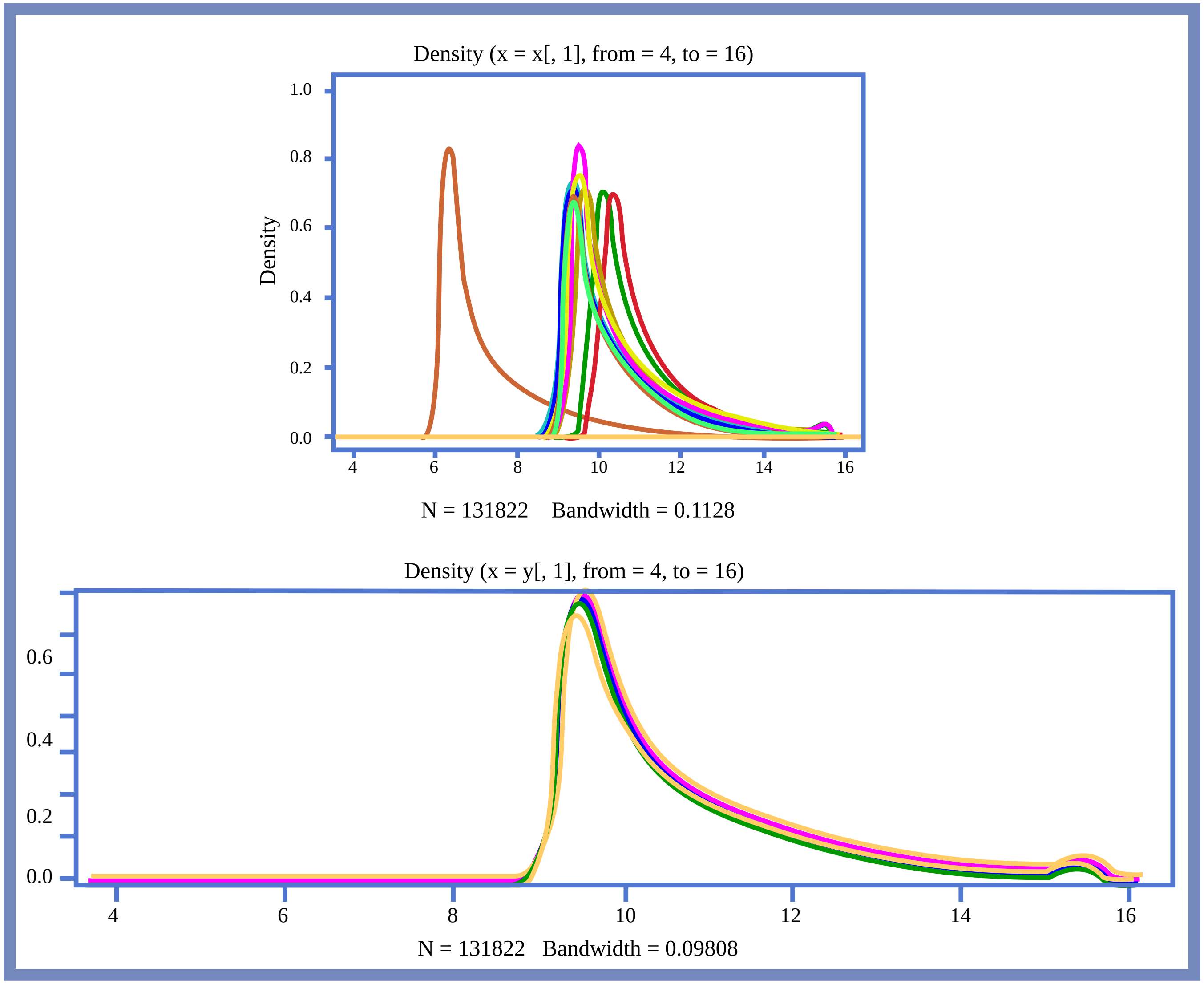
Normalization

Normalization – You don’t believe the numbers

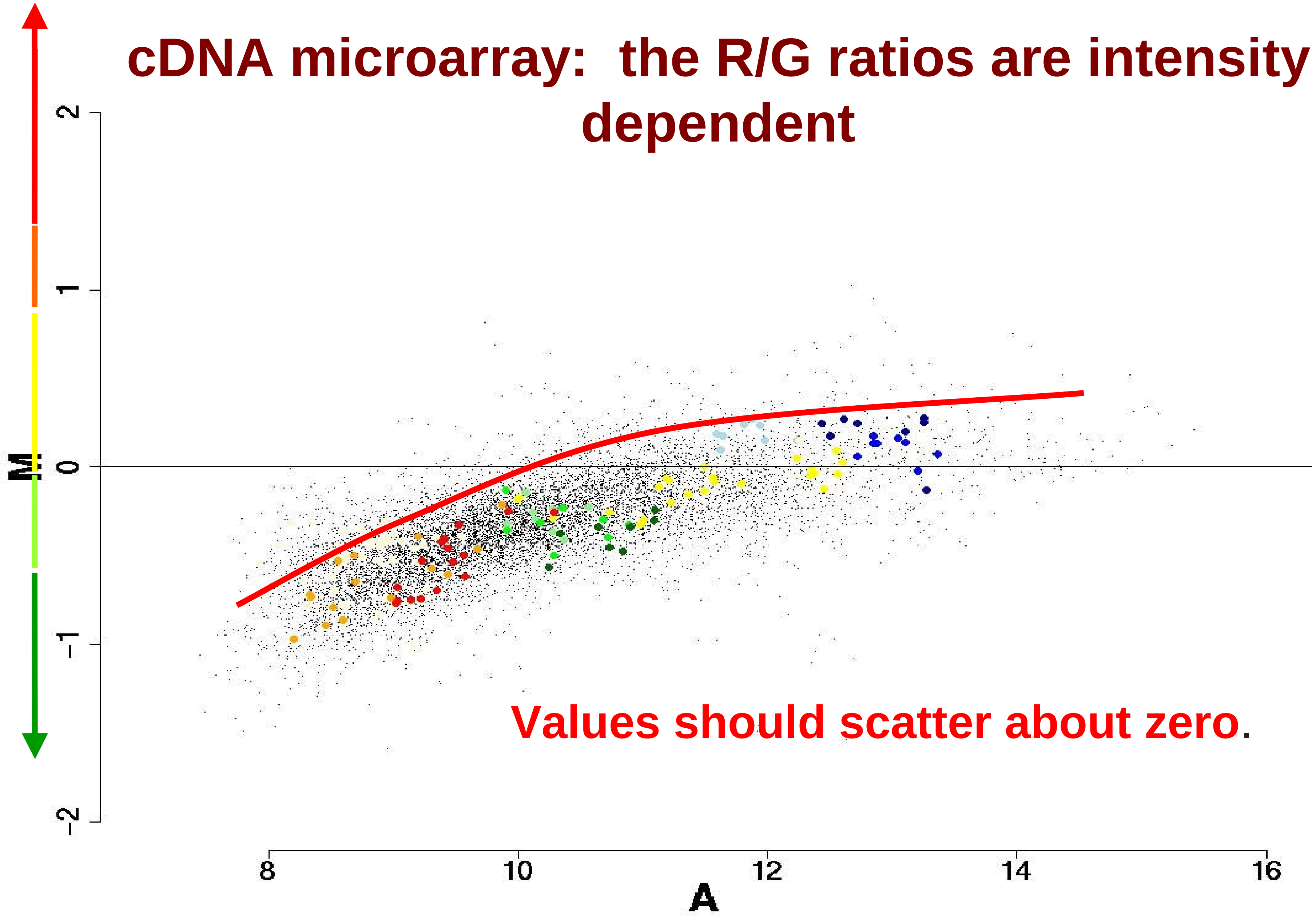
1) “most or certain things do not change”

2) Error model

Shifting the means or medians and adjusting the distributions by Cubic spline fit/ Lowess etc. (Overfitting !!!)



cDNA microarray: the R/G ratios are intensity dependent



Values should scatter about zero.

Courtesy of Natalie Thorne. Used with permission.

Overview of normalization:

- to correct for systematic errors

1) Choose a set of elements that will be used

- **housekeeping genes**
- **special control genes etc.**

2) Determine the normalization function

- **global mean/median normalization**
- **intensity dependent normalization**

Microarray Gene Expression Data Society

www.mged.org

Intensity dependent normalization by error models

Error model: Rocke, Vingron

Low concentrations ⟶ $\mathbf{x} = \mu + \varepsilon$

High concentrations ⟶ $x = \mu e^{\eta}$

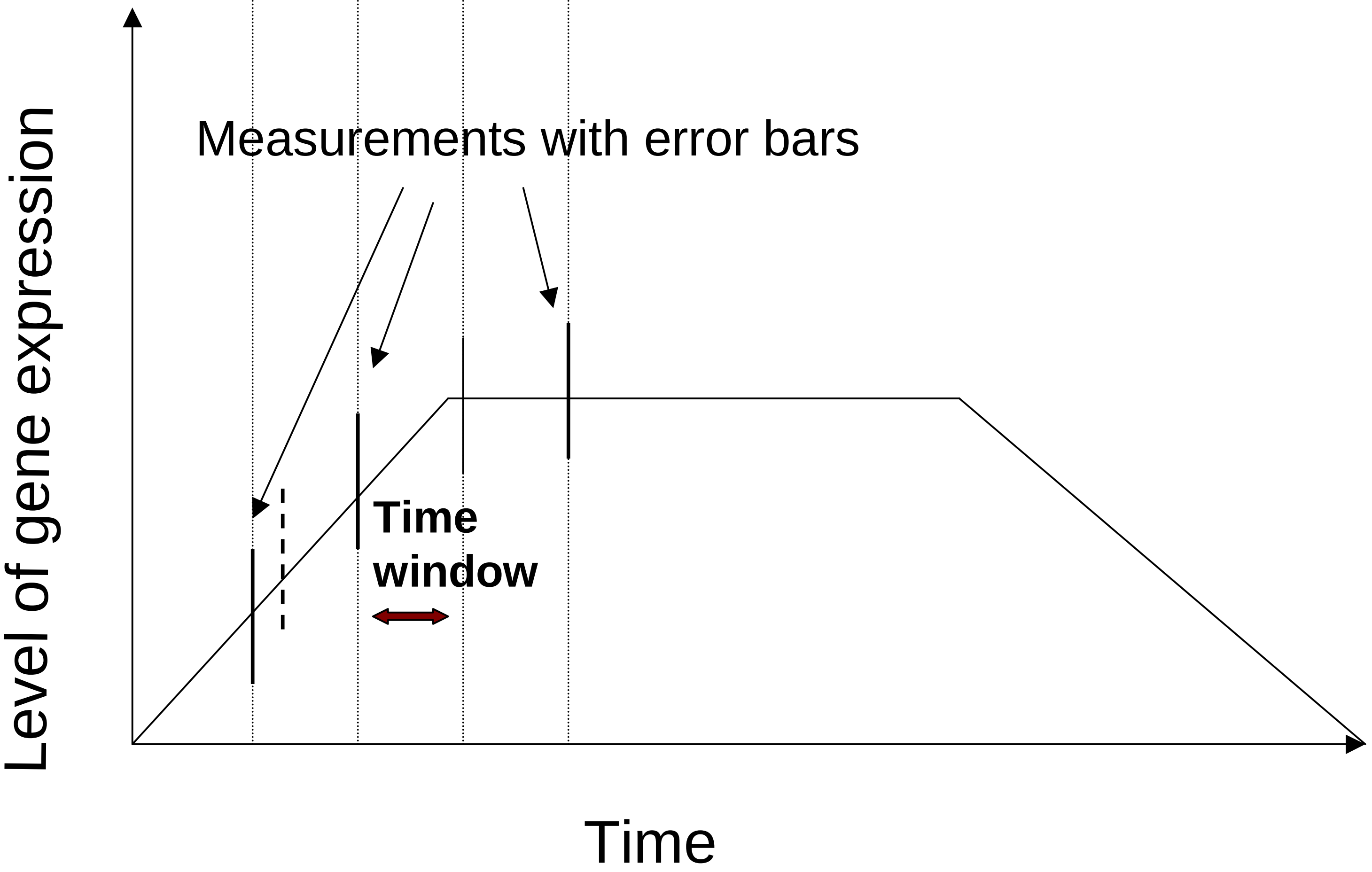
$$x = \mu e^{\eta} + \varepsilon$$

$\eta \sim N(0, \sigma_{\eta}^{2})$ $\varepsilon \sim N(0, \sigma_{\varepsilon}^{2})$

Noise will limit the useful information content of measurements:

A reliable detection of 2-fold differences seems to be the practical limit of massively parallel quantitation.

(estimate: optimistic and not cross-platform)



A rational experiment will sample gene-expression according to a time-series in which each consecutive time point is expected to produce at least as large expression level difference as the error of measurement: approximately 5 min intervals in yeast, 15-30 min intervals in mammalian cells.

Limitations: (you want to make predictions)

Accuracy – noise

Sensitivity - completeness

Inherent limitations –
(think about unpredictability > chaos)

Sensitivity – completeness

How many parameters are we measuring ?

↕

How many parameters should we measure ?

How many bionodes ?

Cautious estimate: on the order of 1-2x10^5

10,000-20,000 active genes per cell

< 3 posttranslational modifications/protein in yeast

3-6 (?) posttranslational modifications/protein in humans

The number of bionodes is probably less than 10 times the number of genes

Splice variants < > modules

The coverage of microarray chips and proteomics keeps increasing >>>> complete genome

Holland MJ. Transcript abundance in yeast varies over six orders of magnitude.

J Biol Chem. 2002

Sensitivity : 2 copies/cell

MOST transcripts are not seen by microarray

Please see J Biol Chem. 2002 Apr 26; 277(17): 14363-6. Epub 2002 Mar 06.

Transcript abundance in yeast varies over six orders of magnitude.

Holland MJ.

The utmost goal of technology :

Single copy/ single cell

BUT even if you measure everything accurately there might be problems with predictions

Even a relatively simple set of ODEs can produce a rather strange behavior.

Edward Lorenz – 3 linked ODEs produced a behavior very sensitive to the initial conditions.
(Chaos theory, Bifurcations etc.)

Small changes in the initial conditions can cause huge changes at later time points

The problem of way too many correlated numbers:

Can this be
due to chance ?

-Analytical solution

- Computational solution:
  Permutate and look for similar patterns

In some cases analytical solution may exist

Six breast cancer cell lines yielded 13 consistently mis-regulated genes (H-cadherin, S1002A, keratin 5 etc.)

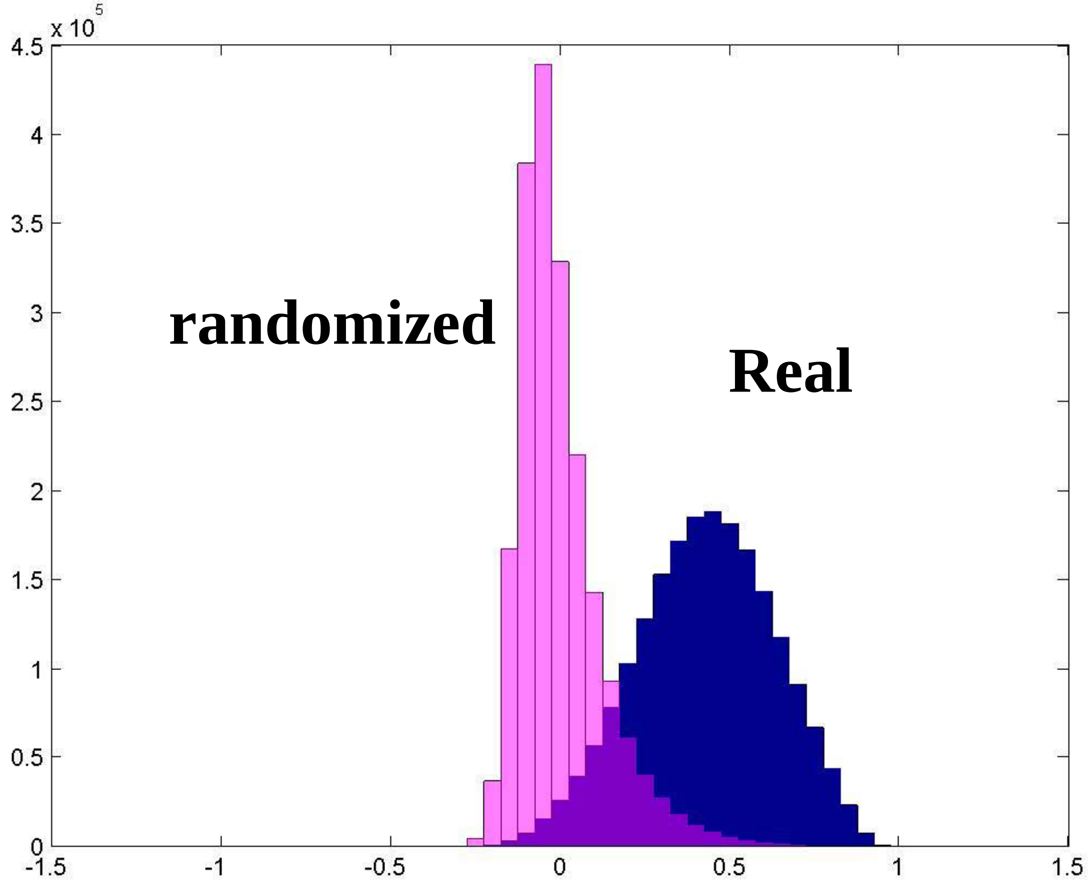
Can this be due to chance ?

"E" different cell lines
"N"-gene microarray
M_i genes mis-regulated in the "i"-th cell line,
K consistently mis-regulated across all E cell lines.

What is the probability that the K genes were mis-regulated by chance ?
This translates into a simple combinatorics problem

BUT !!! - what if more genes are involved

Distribution of pair-wise correlation coefficients in cancer associated gene expression data



The problem of way too many correlated numbers is a particularly nasty one.

Significance can be off by orders of magnitude when comparing completely random permutations with “structural permutations”

Noise in discrete measurements: DNA sequences

Measurement error: Sequencing errors (0.1%-1%)

Solution: sequence a lot

AAATAACTCGGTGACCAAAAAAGAGTGTGAGGATAGATGTCA
GAATGGTTGCTAAGGCACCTATTATTAGGTCGCTTATTAGTTTT
CATGCCGTACATTGCACCTGGCAGACCTTGCCTTATTTCTCTGT
ACATTTTTATTTTCCCGCGTGCTGCGCGGTGTTACACTGCGTTG
TGTATTGCGCTGTGCACGGGGTCTGCGTAAGCGATGTTTTAGG
GCACGGTTTGCTTCTAGAGTGGCCTCTCGCTCTTTTATTACCTCG
CGCTTGTCAATTAGCTTTTTACCTCGCGCAAGGGATATAAGAA
GCTTCGCGCGGCCGTTCCTGAAATAAAACTTGATGGGCACCAG
GGTTATACCAGG……………………

3 billion

-Find genes, introns, exons, transcription factor binding sites etc.

Help can be found --- cDNA libraries etc.
BUT

1) Yelin et al. Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol. 2003:379-86.

~1600 ACTUALLY transcribed antisense transcriptional Units

2) Kapranov et al. Large-scale transcriptional activity in chromosomes 21 and 22. Science, 2002

As much as one order of magnitude more of the genomic sequence is transcribed than accounted for by the predicted and characterized exons.

TF binding site: TGGACT

It can also be: TGCACT

TGG/CACT

TCG/CNCT

Try to add constraints –

1) Within –500 bp from the ATG
2) Tends to cluster in the same region

Even if you do all this you will find that many "obviously" TF binding site-looking sequences do not function as such.
(due to higher level DNA organization etc.)

AND

You often do not know what sequence to start with.

1. Statistical overrepresentation

2. Cross-species conservation

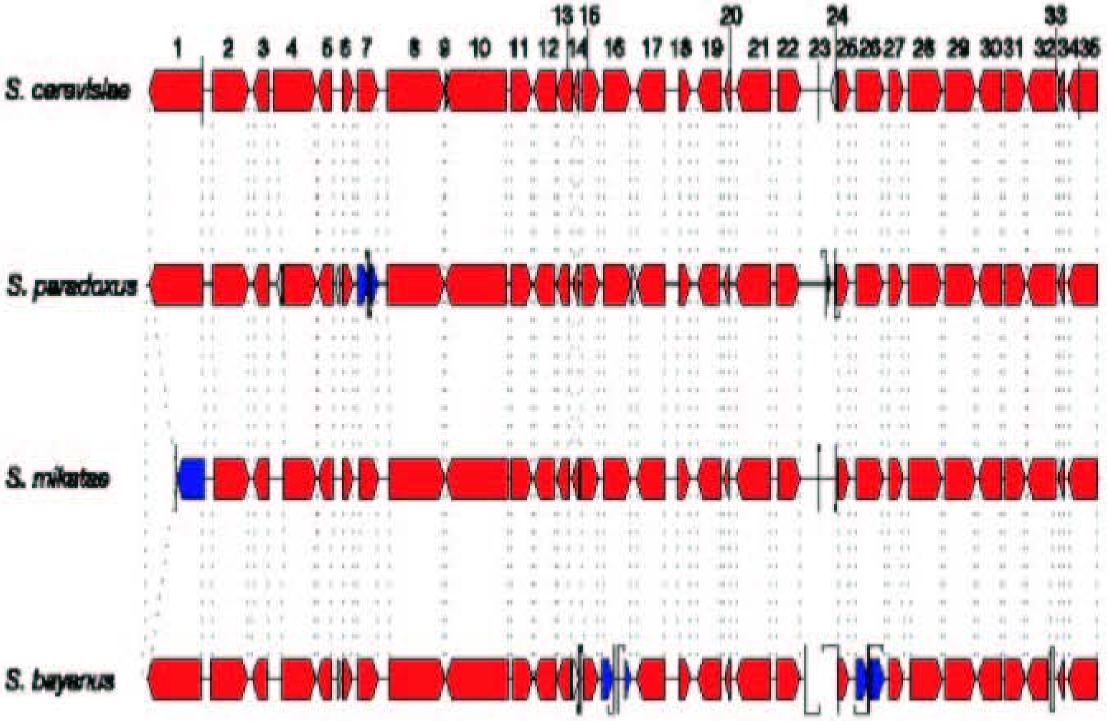
You define the rules

3. Using artificial intelligence/Machine learning
   Hidden Markov models for exon/intron/gene identification (GENIE)

Please see Nature. 2003 May 15; 423(6937): 241-54.
Sequencing and comparison of yeast species to identify genes and regulatory elements.
Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES.

S. cerevisiae S. bayanus S. mikatae S. paradoxus

Number of genes ~ 5,500



1	YGL140C	19	YGL122C (*NAB2*)
2	YGL139W	20	YGL121C
3	YGL138C	21	YGL120C (*PRP43*)
4	YGL137W (*SEC27*)	22	YGL119W (*ABC1*)
5	YGL136C	23	tW(CCA)G1
6	YGL135W (*RPL1B*)	24	YGL118C
7	YGL134W (*PCL10*)	25	YGL117W
8	YGL133W	26	YGL116W (*CDC20*)
9	YGL132W	27	YGL115W (*SNF4*)
10	YGL131C	28	YGL114W
11	YGL130W (*CEG1*)	29	YGL113W
12	YGL129C	30	YGL112C (*TAF60*)
13	YGL128C	31	YGL111W
14	YGL127C (*SOH1*)	32	YGL110C
15	YGL126W (*SCS3*)	33	YGL109W
16	YGL125W (*MET13*)	34	YGL108C
17	YGL124C	35	YGL107C
18	YGL123W (*RPS2*)		

Figure 1 Aligned ORFs across four species. A 50-kb segment of *S. cerevisiae* chromosome VII aligned with orthologous contigs from each of the other three species. Predicted ORFs are shown as arrows pointing in the direction of transcription. Orthologous ORFs are connected by dotted lines and are coloured by the type of correspondence: red for 1-to-1 matches, blue for 1-to-2 matches and white for unmatched ORFs. Sequence gaps are indicated by vertical lines at the ends of contigs, with the estimated size of each gap shown by the length of the hook. See Supplementary Information for 250 such figures tiling the complete *S. cerevisiae* genome.

NATURE | VOL 423 | 15 MAY 2003 | www.nature.com/nature © 2003 Nature Publishing Group 243

Courtesy of Eric Lander. Used with permission.

Slow and rapid evolution:

YBR184W – 32% nucleotide and 13% aa identity

MATa2 - 100 % nucleotide and 100 % aa identity !!!!!!!!

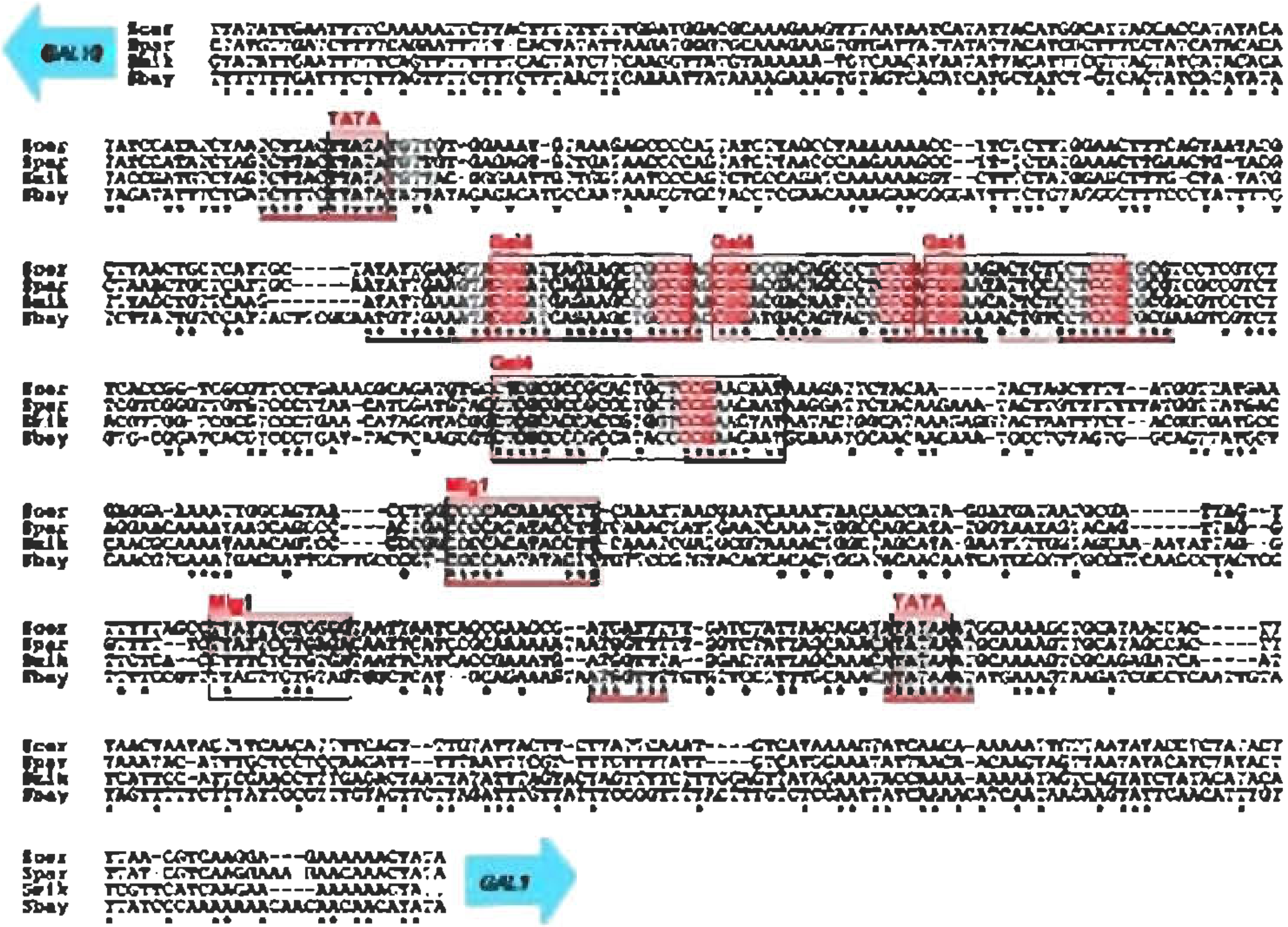


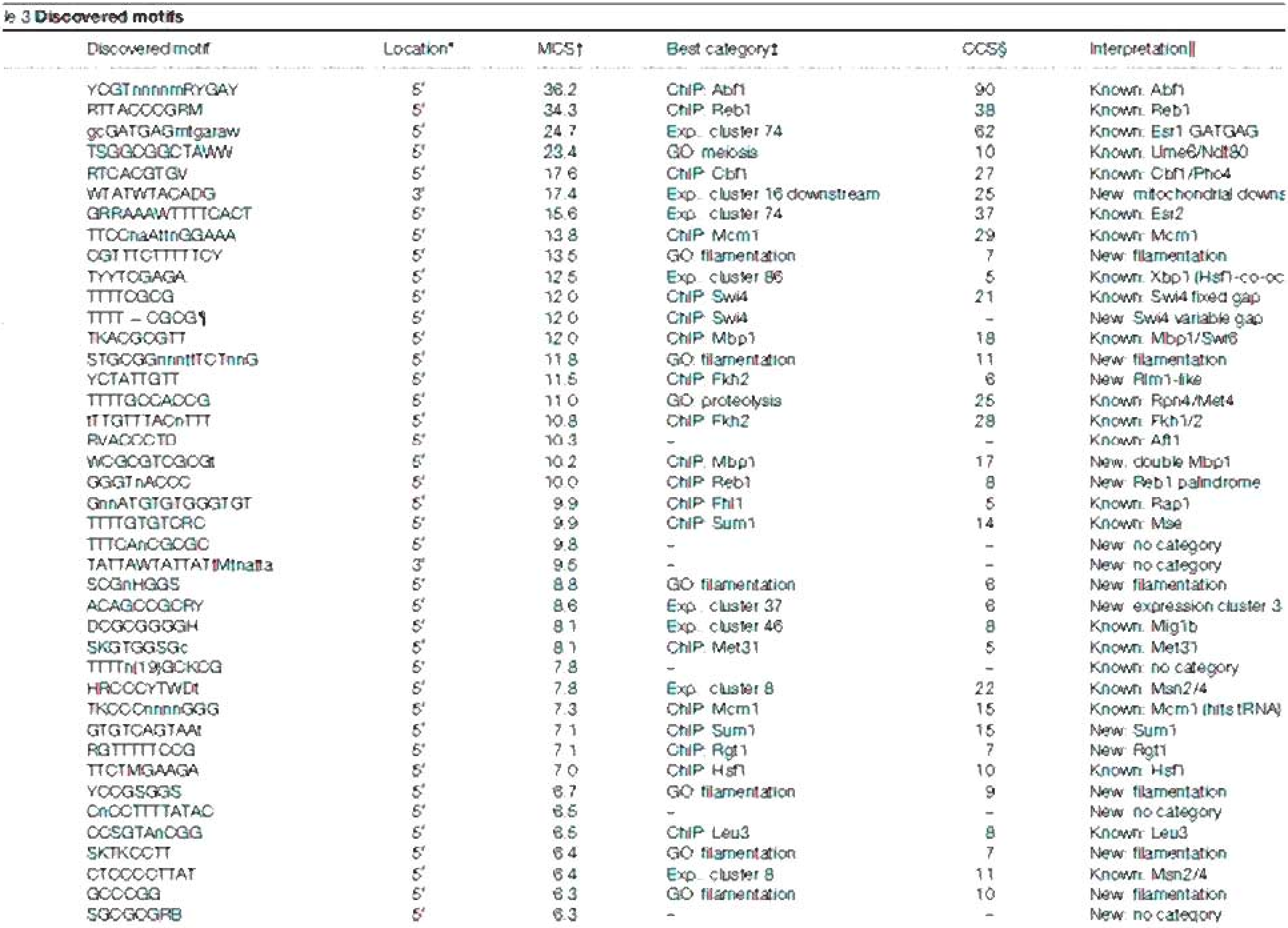
Figure 6 Conservation in the *GAL1–GAL10* intergenic region. Multiple alignment of the four species shows a strong overlap between functional nucleotides and stretches of conservation. Asterisks denote conserved positions in the multiple alignment. Blue arrows denote the start and transcriptional orientation of the flanking ORFs. Experimentally validated factor-binding footprints are boxed and labelled according to the bound factor. Stretches of conserved nucleotides are underlined. Nucleotides matching the published Gal4 motif are shown in red. The fourth experimentally validated site Gal4 L 8 shows a longer footprint and a non-standard consensus motif (boxed). This variant motif is also conserved across all four species. Scer, *S. cerevisiae*, Spar, *S. paradoxus*, Smik, *S. mikatae*, Sbay, *S. bayanus*.

Courtesy of Eric Lander. Used with permission.

$XYZn_{(0-21)}ABC$

Intergenic conservation
Intergenic vs. genic conservation
Upstream vs. downstream conservation

A given motif is also enriched in front of genes with similar function



le 3 **Discovered motifs**

Discovered motif	Location*	MCS†	Best category‡	CCS§	Interpretation‖
YCGTnnnnnRYGAY	5′	36.2	ChIP: Abf1	90	Known: Abf1
RTTACCCGRM	5′	34.3	ChIP: Reb1	38	Known: Reb1
gcGATGAGmtgaraw	5′	24.7	Exp.: cluster 74	62	Known: Esr1 GATGAG
TSGGCGGCTAWW	5′	23.4	GO: meiosis	10	Known: Ume6/Ndt80
RTCACGTGV	5′	17.6	ChIP: Cbf1	27	Known: Cbf1/Pho4
WTATWTACADG	3′	17.4	Exp.: cluster 16 downstream	25	New: mitochondrial downs
GRRAAAWTTTTCACT	5′	15.6	Exp.: cluster 74	37	Known: Esr2
TTCCnaAttnGGAAA	5′	13.8	ChIP: Mcm1	29	Known: Mcm1
CGTTTCTTTTTCY	5′	13.6	GO: filamentation	7	New: filamentation
TYYTCGAGA	5′	12.5	Exp.: cluster 86	5	Known: Xbp1 (Hsf1-co-oc
TTTTCGCG	5′	12.0	ChIP: Swi4	21	Known: Swi4 fixed gap
TTTT – CGCG¶	5′	12.0	ChIP: Swi4	–	New: Swi4 variable gap
TKACGCGTT	5′	12.0	ChIP: Mbp1	18	Known: Mbp1/Swi6
STGCGGnnnttTCTnnG	5′	11.8	GO: filamentation	11	New: filamentation
YCTATTGTT	5′	11.5	ChIP: Fkh2	6	New: Rlm1-like
TTTTGCCACCG	5′	11.0	GO: proteolysis	25	Known: Rpn4/Met4
tTTGTTTACnTTT	5′	10.8	ChIP: Fkh2	28	Known: Fkh1/2
RVACCCTD	5′	10.3	–	–	Known: Aft1
WCGCGTCGCGt	5′	10.2	ChIP: Mbp1	17	New: double Mbp1
GGGTnACCC	5′	10.0	ChIP: Reb1	8	New: Reb1 palindrome
GnnATGTGTGGGTGT	5′	9.9	ChIP: Fhl1	5	Known: Rap1
TTTTGTGTCRC	5′	9.9	ChIP: Sum1	14	Known: Mse
TTTCAnCGCGC	5′	9.8	–	–	New: no category
TATTAWTATTATtMtnatta	3′	9.5	–	–	New: no category
SCGnHGGS	5′	8.8	GO: filamentation	6	New: filamentation
ACAGCCGCRY	5′	8.6	Exp.: cluster 37	6	New: expression cluster 3
DCGCGGGGH	5′	8.1	Exp.: cluster 46	8	Known: Mig1b
SKGTGGSGc	5′	8.1	ChIP: Met31	5	Known: Met31
TTTTn(19)GCKCG	5′	7.8	–	–	Known: no category
HRCCCYTWDt	5′	7.8	Exp.: cluster 8	22	Known: Msn2/4
TKCCCnnnnGGG	5′	7.3	ChIP: Mcm1	15	Known: Mcm1 (hits tRNA)
GTGTCAGTAAt	5′	7.1	ChIP: Sum1	15	New: Sum1
RGTTTTTCCG	5′	7.1	ChIP: Rgt1	7	New: Rgt1
TTCTMGAAGA	5′	7.0	ChIP: Hsf1	10	Known: Hsf1
YCCGSGGS	5′	6.7	GO: filamentation	9	New: filamentation
CnCCTTTTATAC	5′	6.5	–	–	New: no category
CCSGTAnCGG	5′	6.5	ChIP: Leu3	8	Known: Leu3
SKTKCCTT	5′	6.4	GO: filamentation	7	New: filamentation
CTCCCCTTAT	5′	6.4	Exp.: cluster 8	11	Known: Msn2/4
GCCCGG	5′	6.3	GO: filamentation	10	New: filamentation
SGCGCGRB	5′	6.3	–	–	New: no category

Courtesy of Eric Lander. Used with permission.