UCSF cancer center

Integrated Data Systems for Interpreting Genome-Focused Data in Cancer

Ajay N. Jain, PhD

ajain@cc.ucsf.edu

http://jainlab.ucsf.edu

UCSF cancer center

Paradigm shift in biological research

Biology is shifting from being an observational science to being a quantitative molecular science

Old biology: measure one/two things in two/three conditions

- High cost per measurement
- Analysis straightforward
- Enormously difficult to work out pathways

New biology: measure 10,000 things under many conditions

- Low cost per measurement
- Analysis no longer straightforward, but payoff can be bigger
- Biology as a complex system: Can we work out biological pathways this way?

UCSF cancer center

Cancer biology as a complex system:
The marriage of experimental data with annotation information

Phenotype

Proliferation

Measureable phenotypes.

Apoptosis

Protein

P_1 … P_n

Protein status for over multiple conditions.

RNA

G_1 … G_n

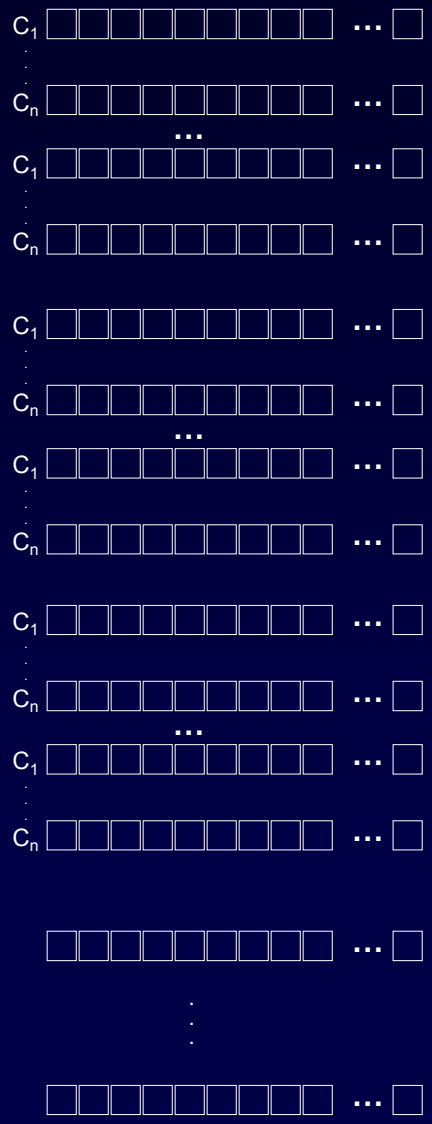
Gene expression levels over multiple conditions.

DNA

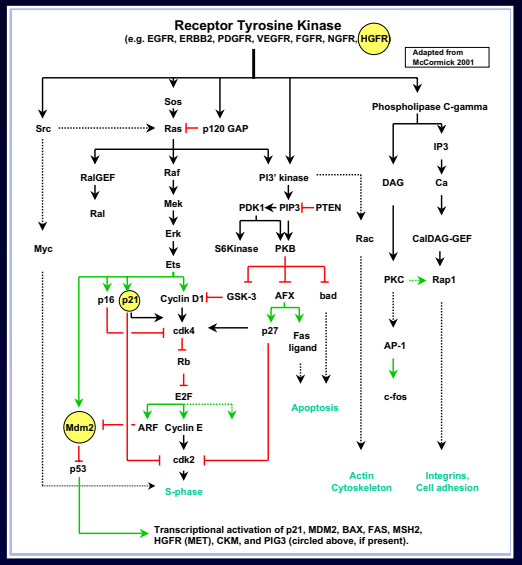
L_1 … L_n

DNA copy number over the entire genome.

Cell Lines



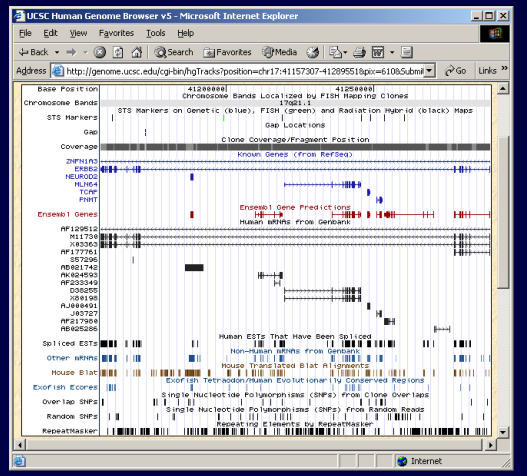
Pathway Structure →



ERBB2:
EC Number: 2.7.1.112
oncogenesis
cell proliferation
Neu/ErbB-2 receptor
protein phosphorylation
protein dephosphorylation
cell growth and maintenance
receptor signaling tyrosine kinase

← Gene Annotations

Genomic Mapping + Context →



UCSF cancer center

We must represent both experimental and annotation data in a generalizable, scalable data system with rich query capabilities and rich analytical and visualization methods

SampleNames
- SampleNumber
- SampleName
- Experiment
- Upload

Annotation
- **Type**
- **Value**
- **DataPosition**
- **DataTypeNum**
- Experiment
- Upload

Data
- **OrdinalPosition**
- **Value**
- **DataTypeNumber**
- **SampleNumber**
- Experiment
- Upload

Experiment
- Experiment
- Upload
- UserName
- LabName
- NumberSamples
- DataType
- Delimiter
- EntryDate

DataTypes
- DataTypeNumber
- DataTypeName
- NumberEntries
- Experiment
- Upload

UploadSpecifications
- TableInformationID
- Upload
- Name
- AnnotationType
- AnnotationStart
- AnnotationStop
- RowColNumber
- Format
- Experiment

Phenotype

Proliferation … Apoptosis — Measureable phenotypes.

Protein

$P_1 \ldots P_n$ — Protein status for over multiple conditions.

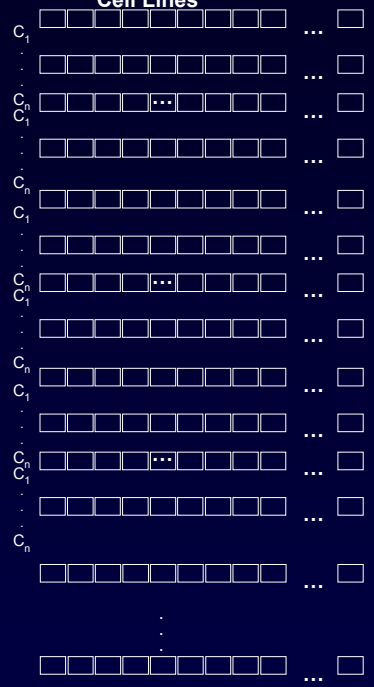
RNA

$G_1 \ldots G_n$ — Gene expression levels over multiple conditions.

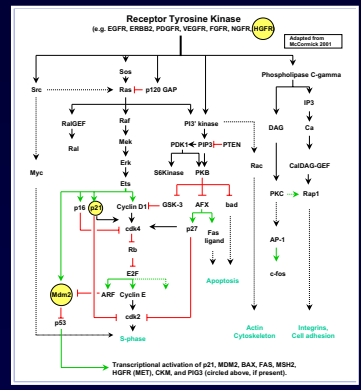
DNA

$L_1 \ldots L_n$ — DNA copy number over the entire genome.

Cell Lines



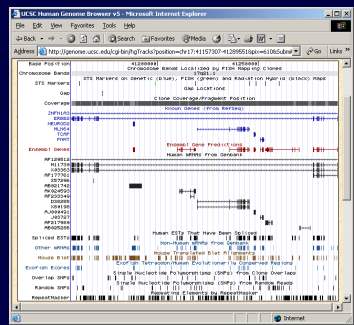
Pathway Structure



ERBB2:
EC Number: 2.7.1.112
oncogenesis
cell proliferation
Neu/ErbB-2 receptor
protein phosphorylation
protein dephosphorylation
cell growth and maintenance
receptor signaling tyrosine kinase

← **Gene Annotations**

Genomic Mapping + Context →



UCSF cancer center

DNA Copy Number Aberrations in Cancer

- Some important cancer genes are sometimes present in altered copy number in some tumors
- Increases over express oncogenes, decreases help inactivate suppressor genes, dosage changes affect expression
- These copy number alterations can be predictive of tumor phenotype and patient outcome

Quantitative analysis of chromosomal CGH in human breast tumors associates copy number abnormalities with p53 status and patient survival

Ajay N. Jain*†, Koei Chin*†, Anne-Lise Børresen-Dale‡, Bjorn K. Erikstein‡, Per Eystein Lonning§, Rolf Kaaresen¶, and Joe W. Gray*‖

*UCSF Cancer Center, University of California, San Francisco, Box 0128, San Francisco, CA 94143-0128; ‡Departments of Genetics and Oncology, Institute for Cancer Research, Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway; §Department of Oncology, Haukeland Hospital, 5021 Haukeland Sykehus, Norway; and ¶Department of Surgery, Ullev Hospital, 0407 Oslo, Norway

Communicated by James E. Cleaver, University of California, San Francisco, CA, May 15, 2001 (received for review February 5, 2001)

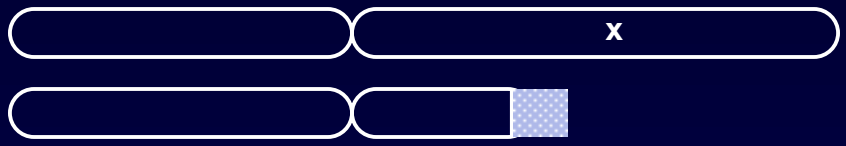
7952–7957 | PNAS | July 3, 2001 | vol. 98 | no. 14

UCSF cancer center

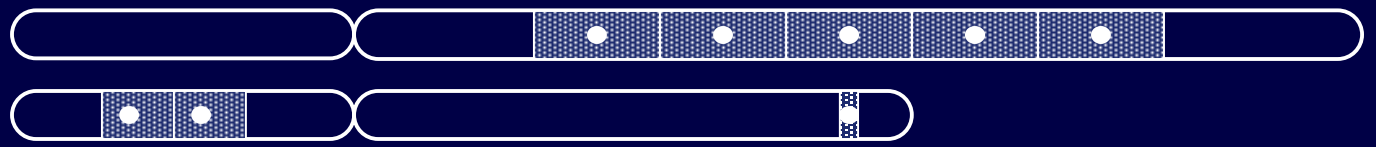
Characteristics of Copy Number Changes in Cancer

Copy number changes usually involve a DNA segment that is substantially larger than the critical gene(s)

Mutation plus terminal deletion of tumor a suppressor gene

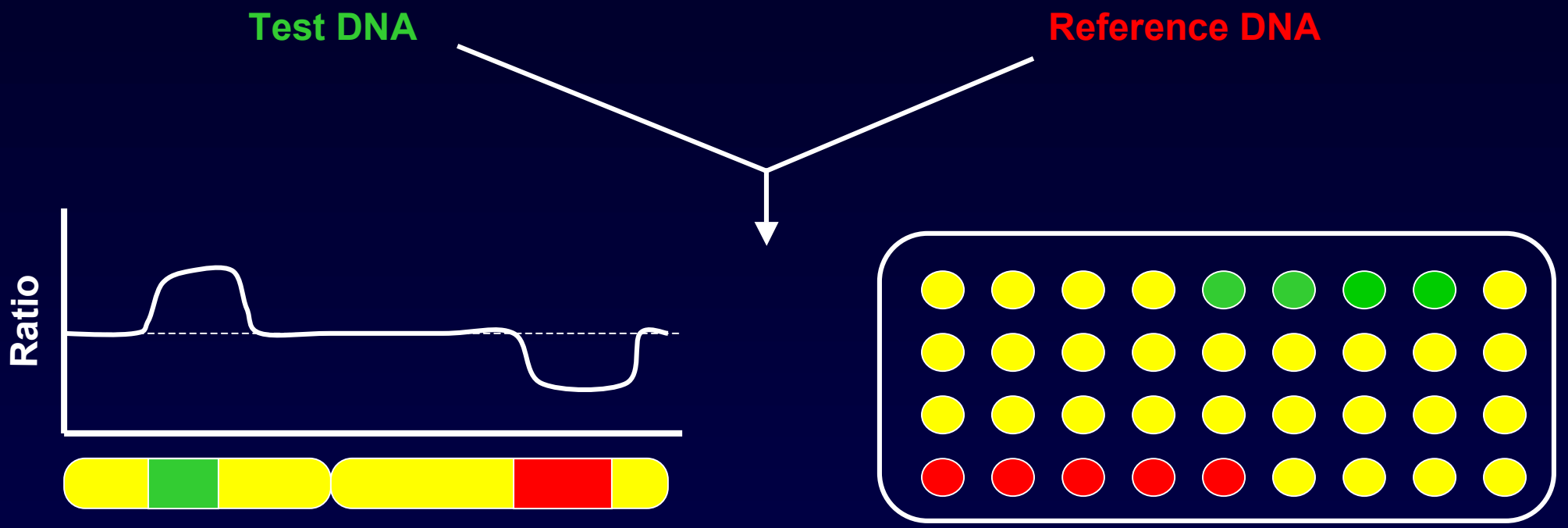


Amplification of an oncogene and surrounding DNA; extra copies can be located anywhere in the genome



UCSF cancer center

Comparative Genomic Hybridization (CGH): Chromosomal targets versus array targets



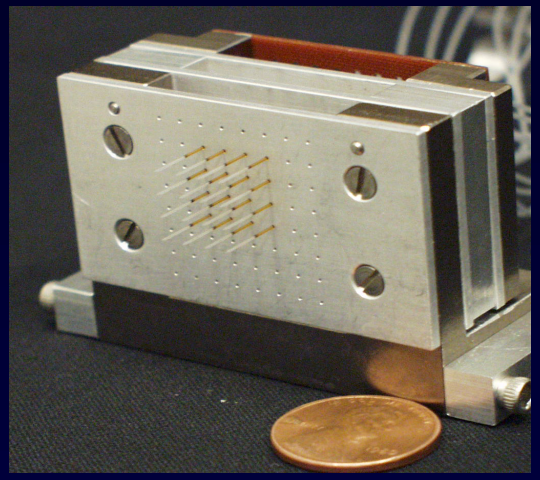
Chromosome CGH provides "cytogenetic" resolution ~ 10 Mb

Resolution of array CGH depends on spacing and length of clones

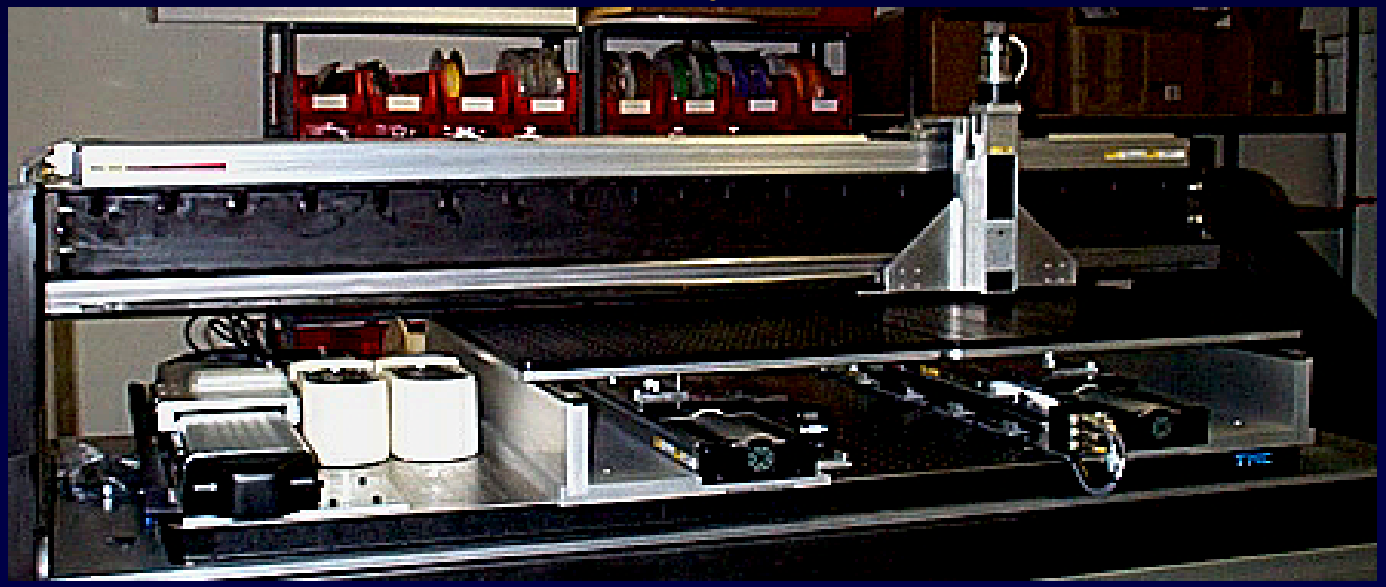
UCSF cancer center

UCSF Cancer Center Investigators Albertson and Pinkel have developed in-house technology for printing genomic clone arrays

Capillary Print head

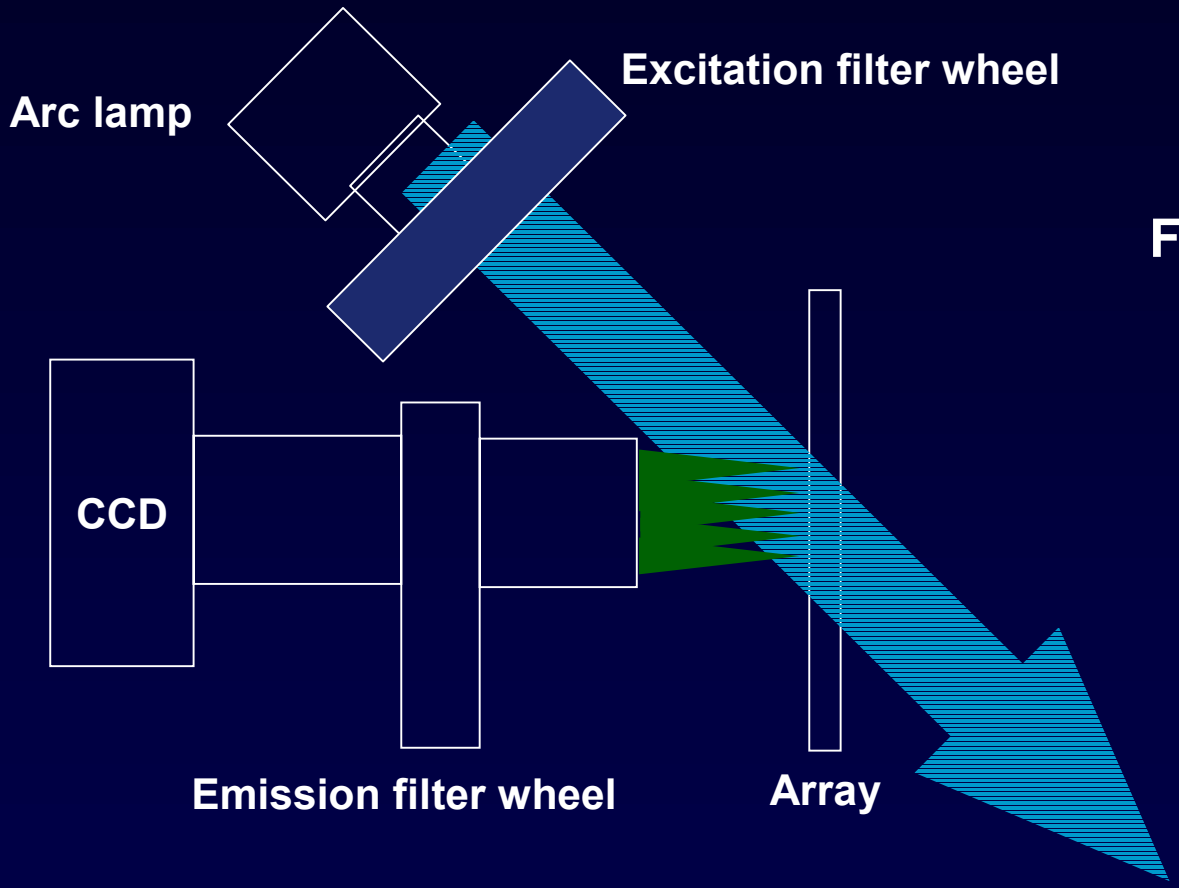


Overview of Layout



UCSF cancer center

Custom CCD Imaging System: Allows for imaging multiple distinct nucleic acid species



Filters permit use of:

- **Dapi etc.**
- **Fluorescein etc.**
- **Cy3 etc.**
- **Texas red etc.**
- **Cy5**

UCSF cancer center

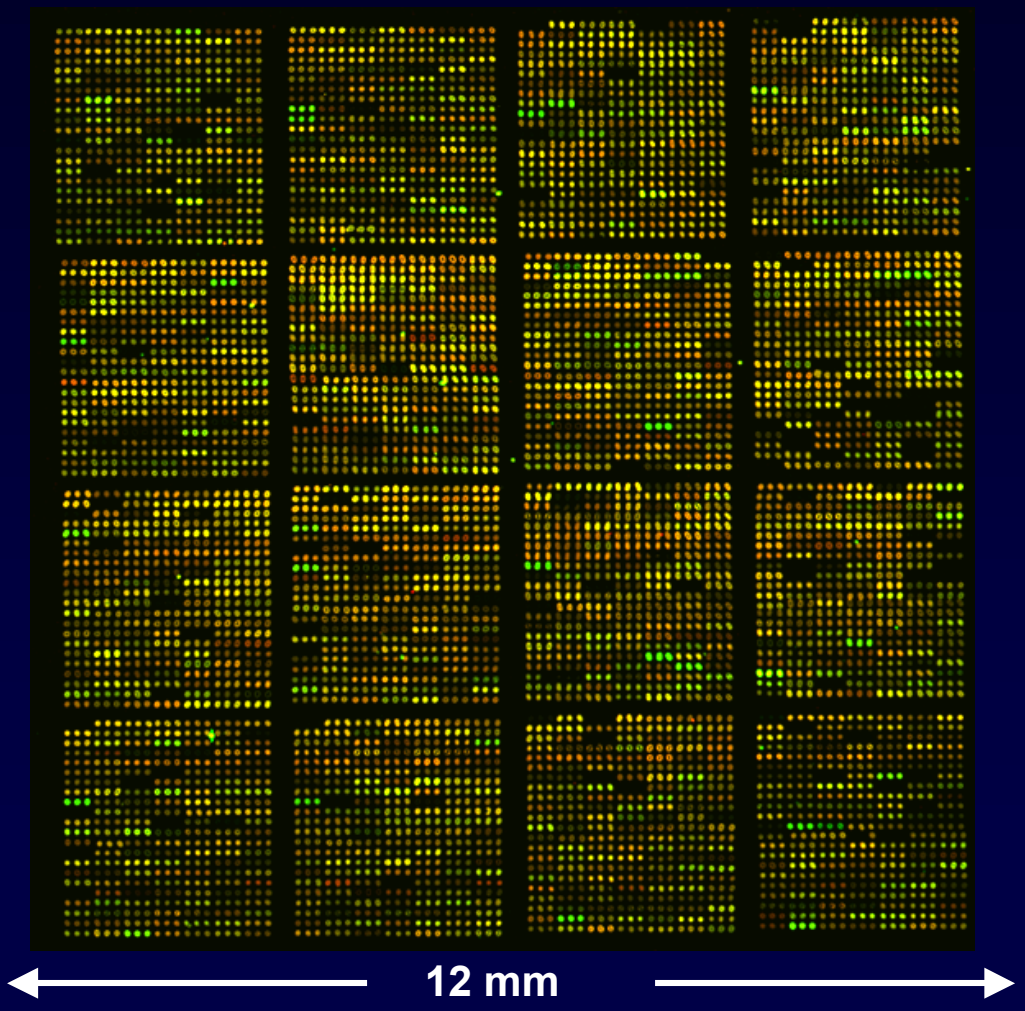
Genome Scanning Array with ~ 1.4 Mb Resolution

DNA obtained from peripheral blood, fresh, frozen or fixed tissue

3 ng to 0.5 μg input DNA

Random Prime Nick Translation FITC, Cy3, Cy5

16 - 40 hr Hyb.



2500 BACs

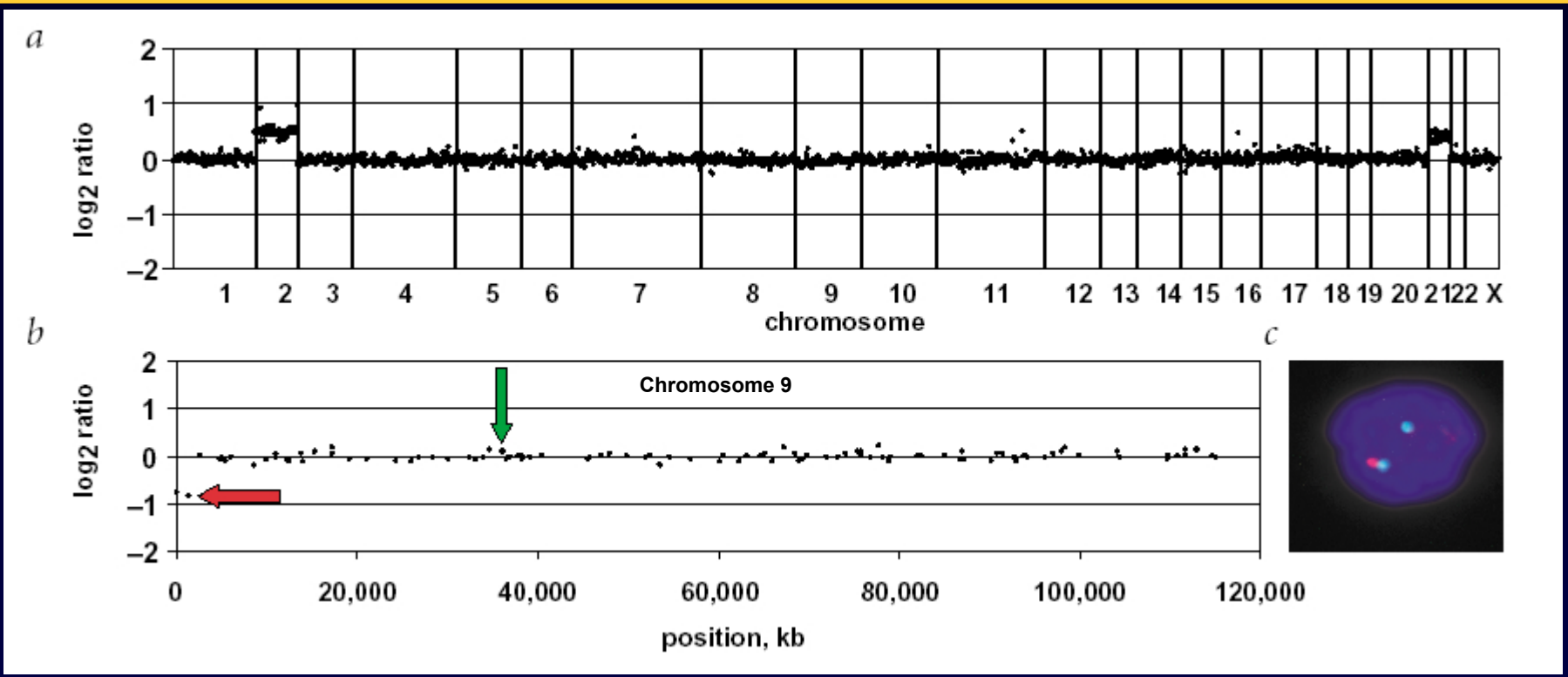
Triplicate spots

130 μm centers

864 well plates

UCSF cancer center

Accurate detection of single copy changes



nature genetics • volume 29 • november 2001

Assembly of microarrays for genome-wide measurement of DNA copy number

Published online: 30 October 2001, DOI: 10.1038/ng754

We have assembled arrays of approximately 2,400 BAC clones for measuremen DNA copy number across the human genome. The arrays provide precise meas ment (s.d. of $\log_2$ ratios=0.05–0.10) in cell lines and clinical material, so that we reliably detect and quantify high-level amplifications and single-copy alteration diploid, polyploid and heterogeneous backgrounds.

Antoine M. Snijders[1,2], Norma Nowak[4], Richard Segraves[1], Stephanie Blackwood[1,2], Nils Brown[1], Jeffrey Conroy[4], Greg Hamilton[1], Anna Katherine Hindle[1,2], Bing Huey[1], Karen Kimura[1], Sindy Law[1,2], Ken Myambo[1], Joel Palmer[1,2], Bauke Ylstra[1,2], Jingzhu Pearl Yue[1], Joe W. Gray[1,3], Ajay N. Jain[1–3], Daniel Pinkel[1,3] & Donna G. Albertson[1–3]

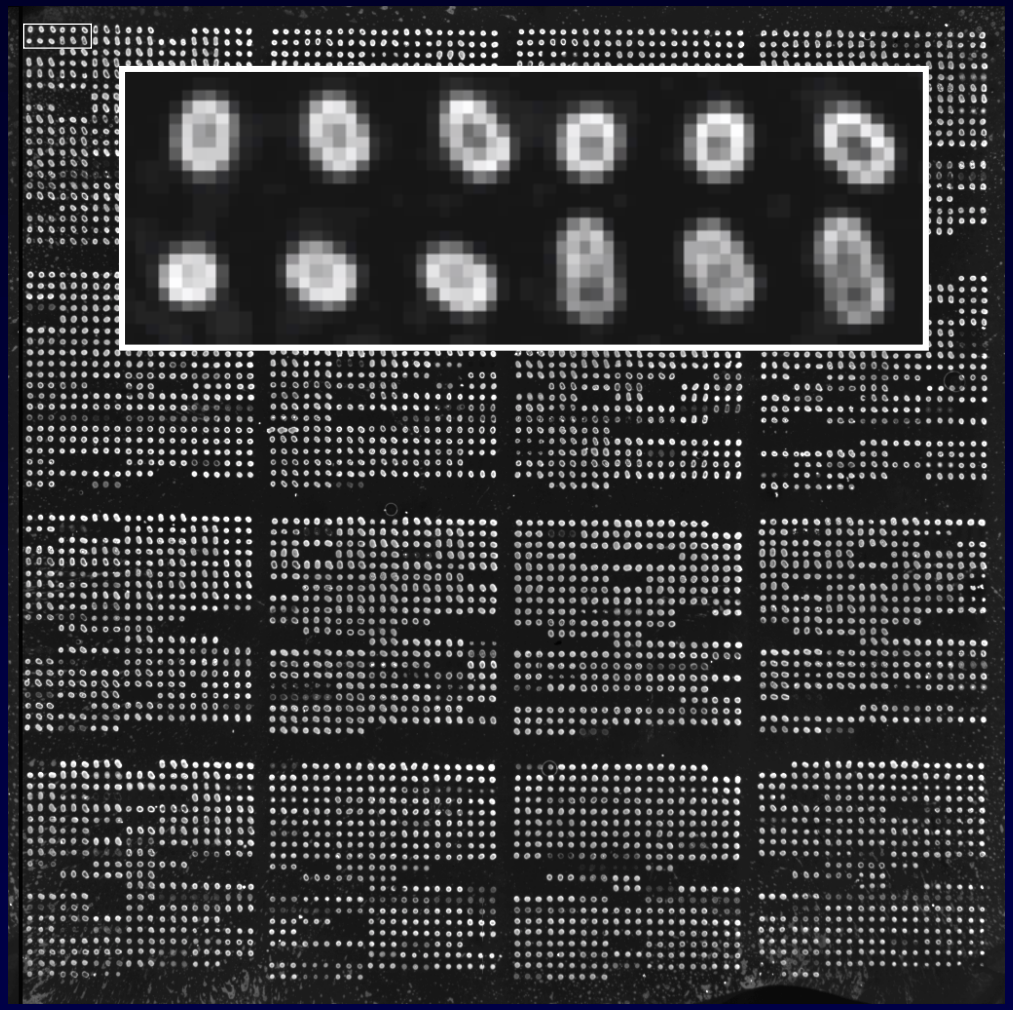
[1]Comprehensive Cancer Center, [2]Cancer Research Institute and [3]Department of Laboratory Medicine, University of California San Francisco, San Francisco, California 94143. [4]Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, New York 14263. Correspondence should be addressed to D.G.A. (e-mail: albertson@cc.ucsf.edu).

ArrayCGH can detect single copy gains and losses. Top shows genome-wide copy number for cell strain GM03576.

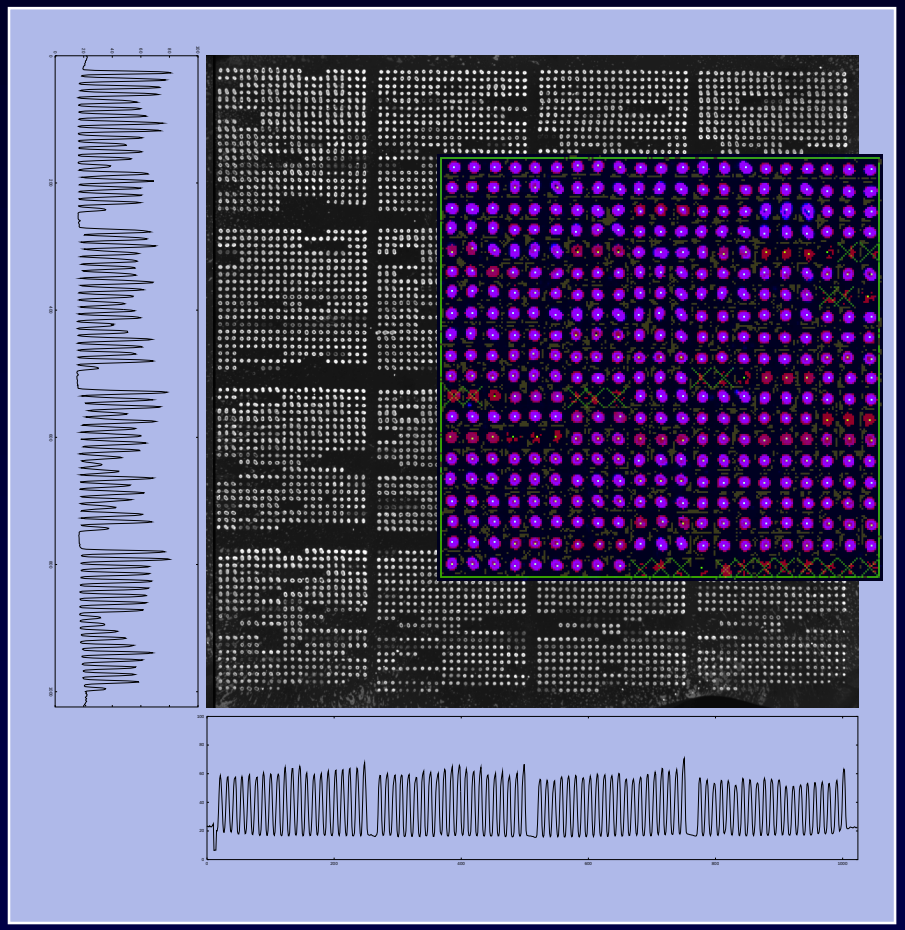
Bottom left shows copy number for cell strain GM03563 on chromosome 9. Bottom right: single copy deletion verified by FISH.

UCSF cancer center

UCSF Spot processes typical arrayCGH hybridizations in less than 20 seconds, fully automatically



Dapi image, ~6000 spots, 135 micron centers
Spots are not perfect filled circles.



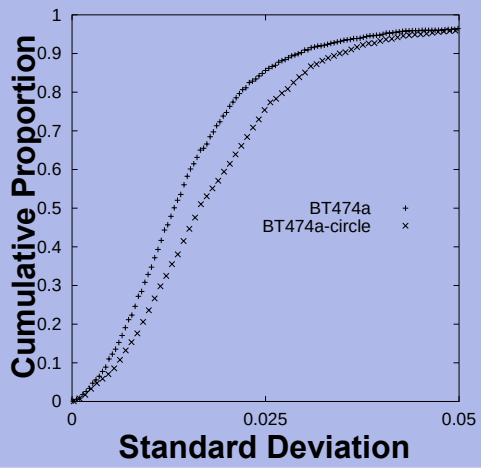
A. N. Jain, T. A. Tokuyasu, A. M. Snijders, R. Segraves, D. G. Albertson, and D. Pinkel. Fully automatic quantification of microarray image data. Genome Research 12: 325-332, 2002.

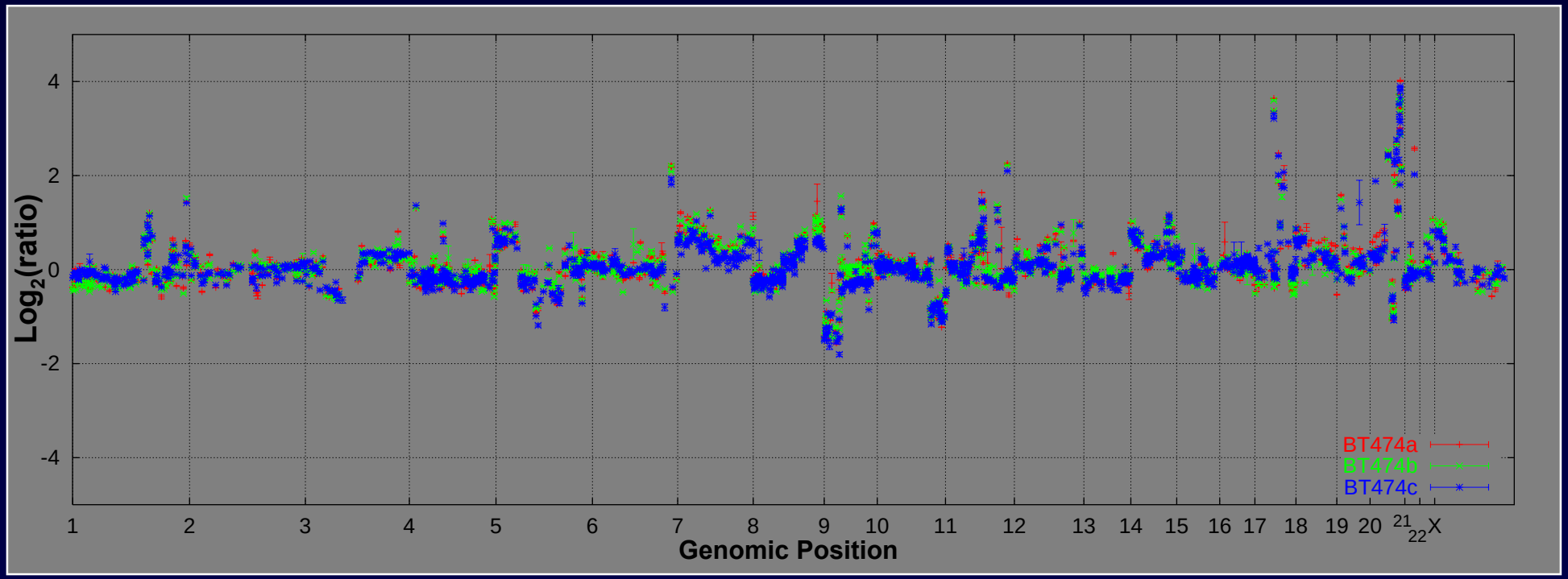
UCSF cancer center

Explicit spot segmentation yields accurate results
Complex patterns of amplification/deletion are common

Circularity assumption leads to poorer spot replicate stddev distributions. Higher signal reduces the difference.

Experimental replicates have larger deviation than that due to image quantification noise (BT474x3, Albertson and Pinkel Labs)





UCSF cancer center

Cancer biology as a complex system:
The marriage of experimental data with annotation information

Phenotype

Proliferation

Measureable phenotypes.

Apoptosis

Protein

P_1 … P_n

Protein status for over multiple conditions.

RNA

G_1 … G_n

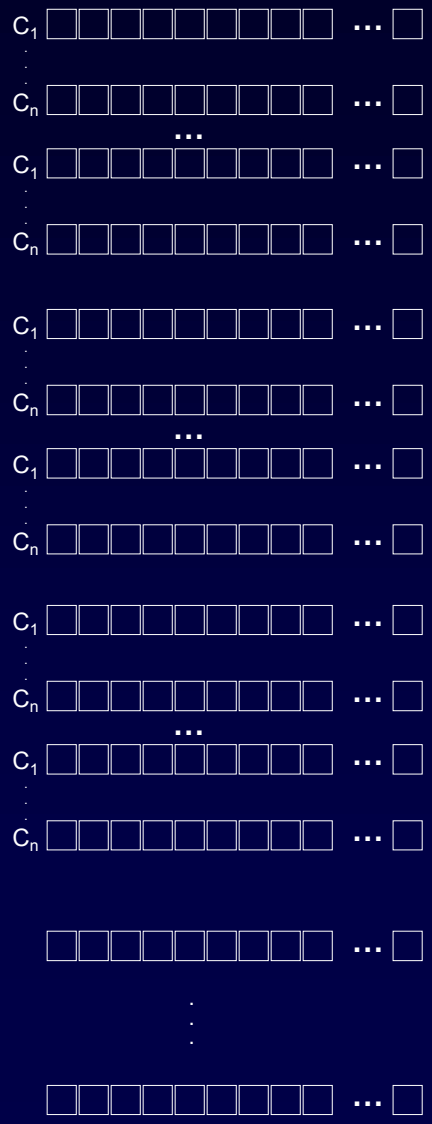
Gene expression levels over multiple conditions.

DNA

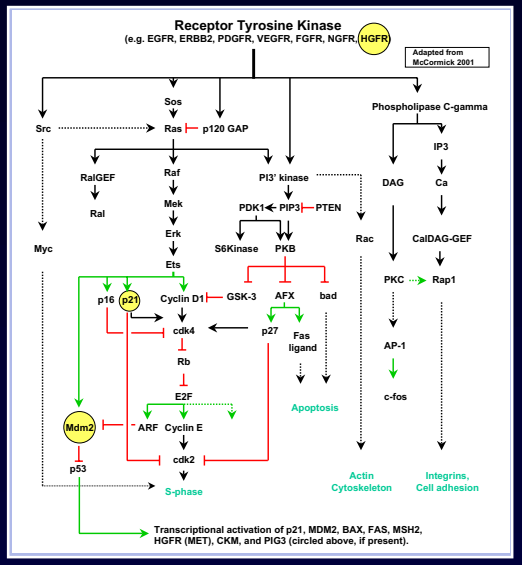
L_1 … L_n

DNA copy number over the entire genome.

Cell Lines



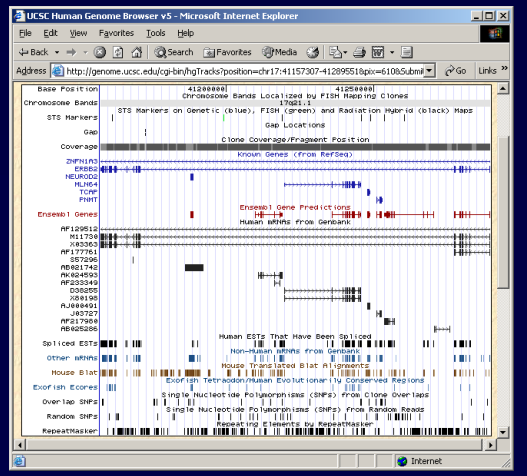
Pathway Structure →



ERBB2:
EC Number: 2.7.1.112
oncogenesis
cell proliferation
Neu/ErbB-2 receptor
protein phosphorylation
protein dephosphorylation
cell growth and maintenance
receptor signaling tyrosine kinase

← Gene Annotations

Genomic Mapping + Context →



UCSF cancer center

Biological Pathways: The receptor tyrosine kinase signaling pathways

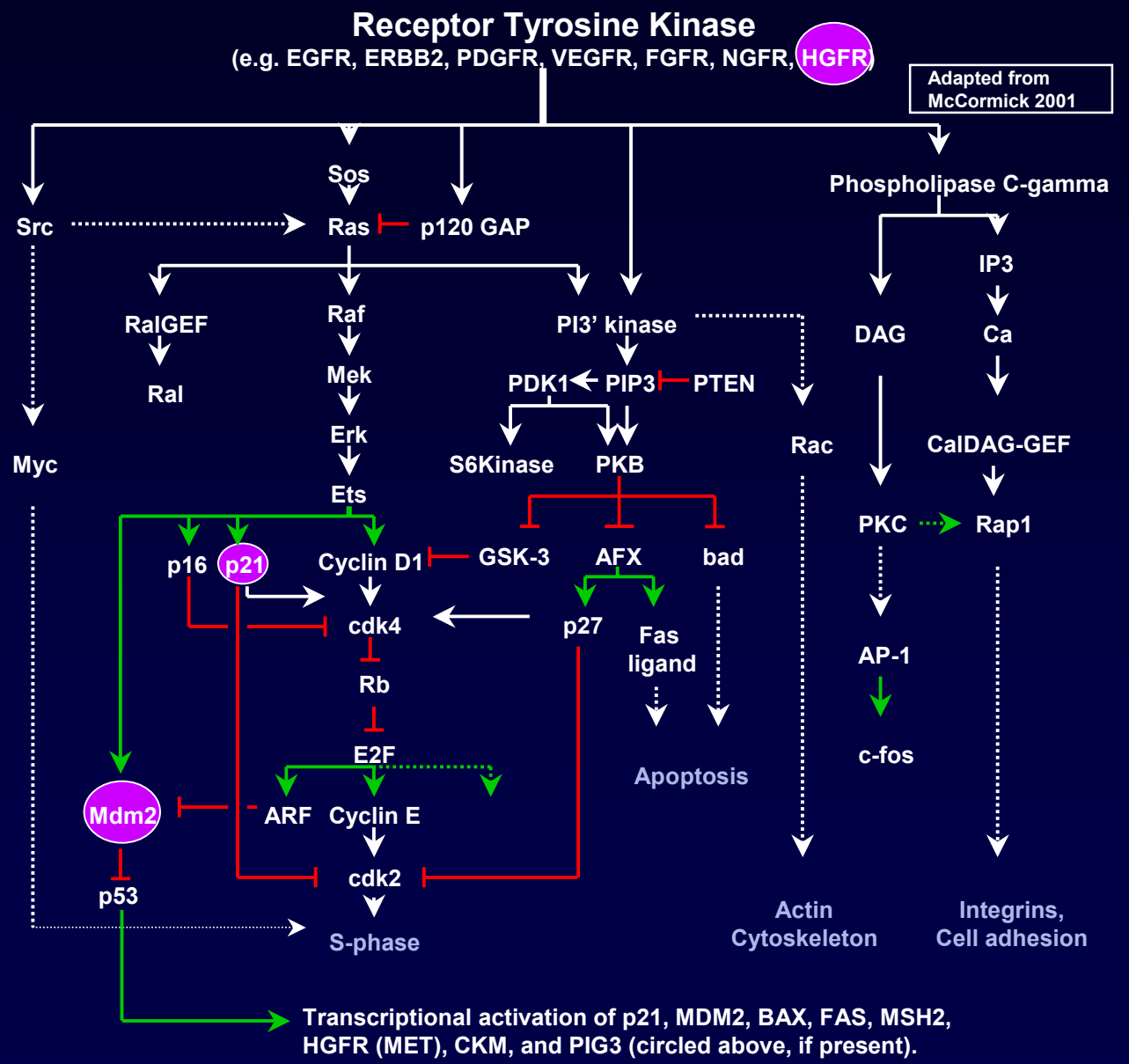
RTK signaling:

- Cell cycle control
- Apoptosis
- Cell adhesion

Therapeutics:

- Herceptin
- PI3-kinase inhibitors
- Onyx015

Pathway derived through hard molecular biology and biochemistry



UCSF cancer center

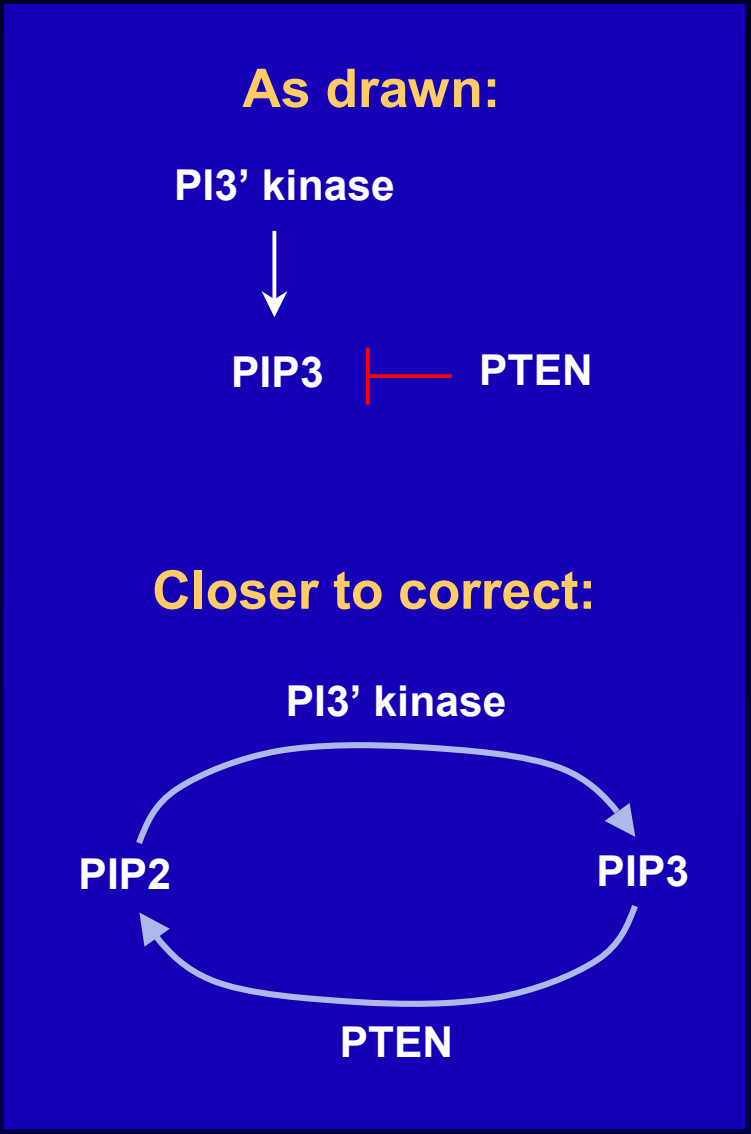
Pathways in human cell biology are complex and variably understood

Problem 1: symbols are overloaded

Problem 2: shorthand is used

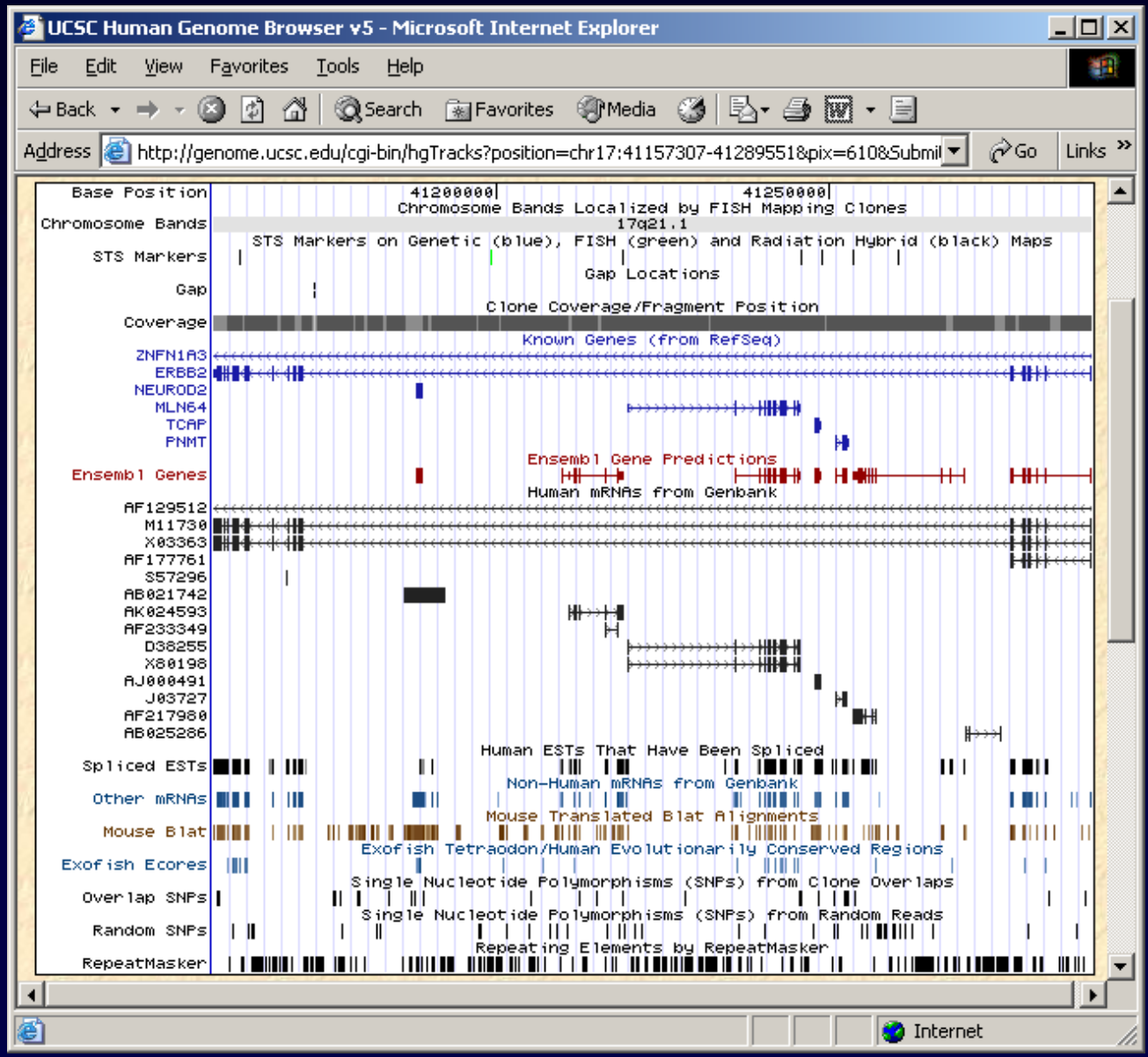
Problem 3: knowledge is incomplete

But we must still try to represent this information



UCSF cancer center

The human genome contains a great deal of information



UCSF cancer center

Biological Data Analysis in the New World: General Statistical Problem

Large number of measurements

- ~3000 for genome-wide array-CGH
- >30,000 for expression arrays

Small number of samples

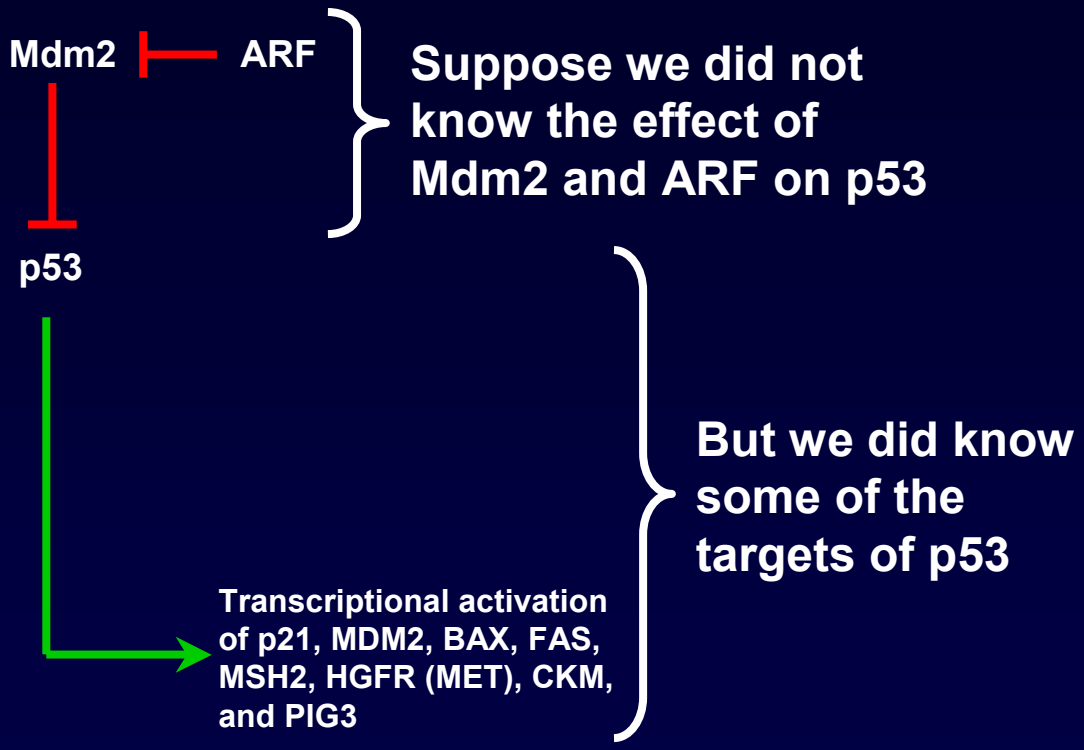
- Typically 5 to 50 cell lines, time points, or tissue samples
- Ratio of measurements to samples can be as bad as 10^3

It is often impossible to make a rigorous quantitative conclusion in such cases.

Explicit use of orthogonal knowledge sources to constrain your questions makes it possible to derive quantitative conclusions.

UCSF cancer center

Array-based data can potentially accelerate the derivation of biological pathways



Suppose we did not know the effect of Mdm2 and ARF on p53

But we did know some of the targets of p53

We have a collection of cell lines and collect the following information

- P53 mutational status
- Temporal expression data

Why are the targets of p53 non-responsive to wildtype p53 in some cell lines?

Look for genes whose expression is positively or negatively correlated with expression of p53 targets in the subset of cell lines with WT p53.

Find that mdm2 is negatively correlated with p53 targets and ARF is positively correlated

In order to do this, we need to integrate experimental data with systematized biological knowledge.

UCSF cancer center

The NCI60 cell-lines as a complex system: Each cell line is a (large) perturbation

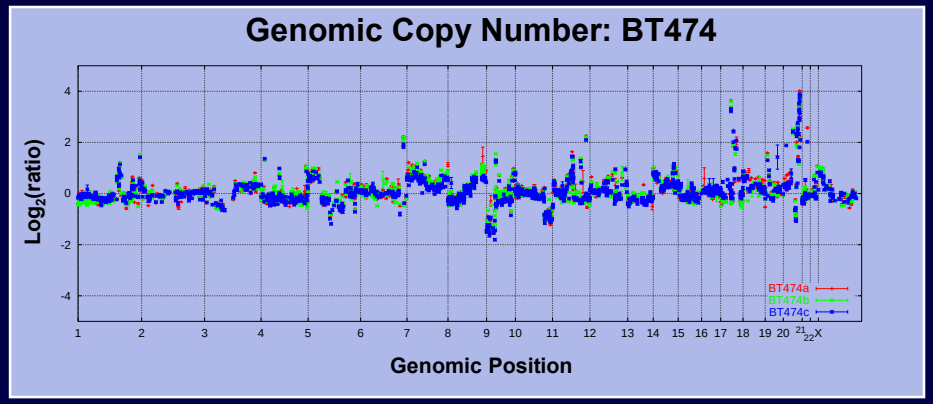
RNA

60 tumor-derived cell lines

G_1 Gene expression levels
⋮ in a single condition,
G_n ~6000 cDNAs (Ross et al. 2000)

DNA

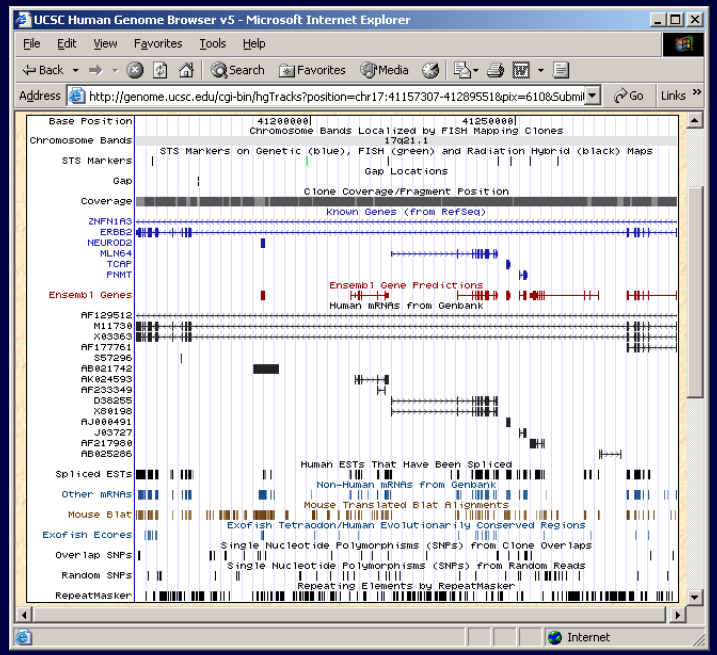
L_1 DNA copy number over
⋮ the entire genome,
L_n ~500 BACs (Gray Lab)



Gene Annotations

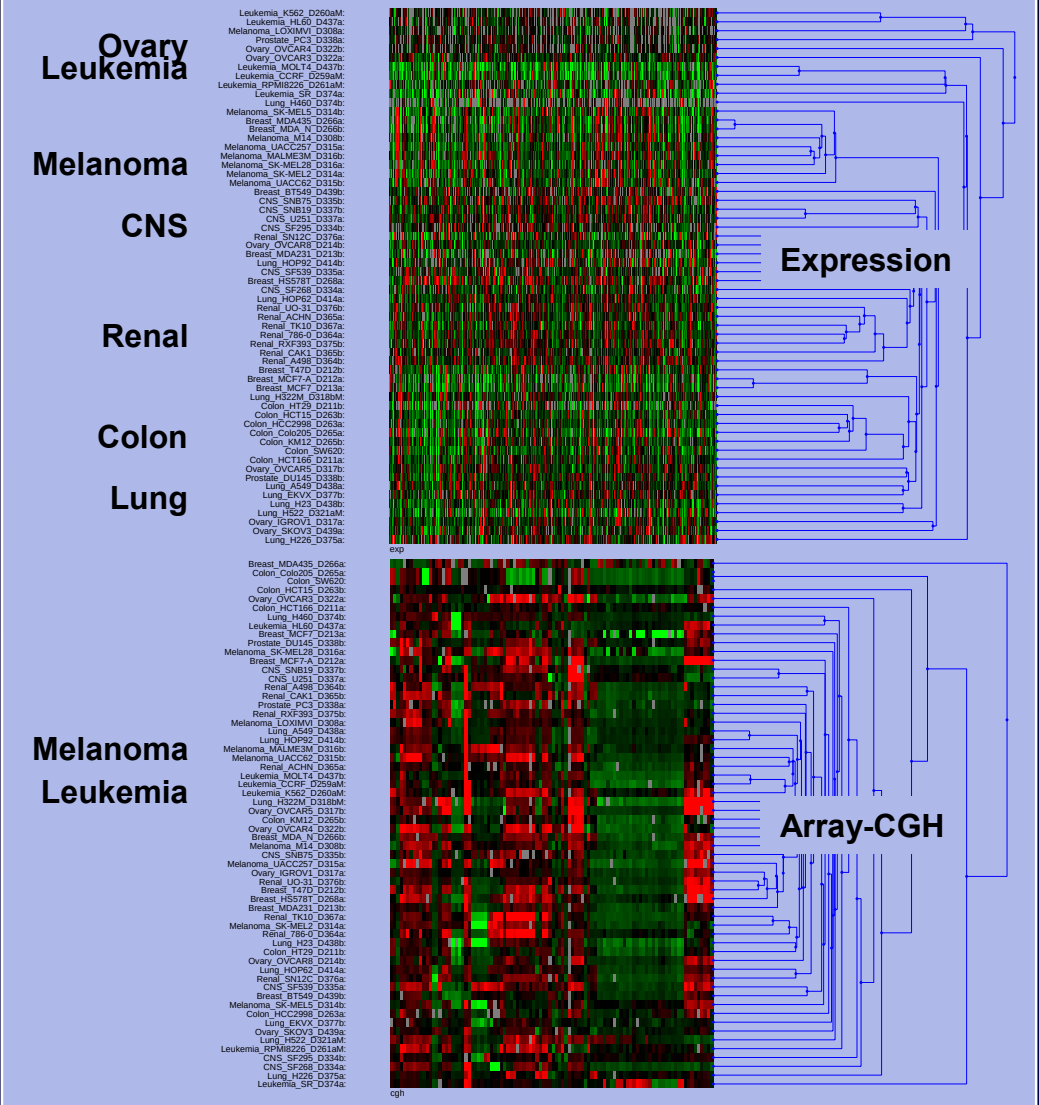
ERBB2:
EC Number: 2.7.1.112
oncogenesis
cell proliferation
Neu/ErbB-2 receptor
protein phosphorylation
protein dephosphorylation
cell growth and maintenance
receptor signaling tyrosine kinase

Genomic Mapping + Context



UCSF cancer center

The signals are clearly different in the CGH and expression data



The dominant signal in the expression data is tissue of origin.
The dominant signal in the copy number data is not.

Is gene expression related to DNA copy number?

Are genes within similar pathways expressed similarly?

Are genes that are co-expressed similar in their regulatory sequences?

UCSF cancer center

We expect that there will be effects of gene dosage on expression

Direct effect

- A gene that is on a genomic region represented 8 times in the genome of a particular cell line should have higher expression than in a cell line with normal copy number
- A gene that is homozygously deleted should not be expressed

Indirect effects

- A genomic locus that is amplified as a result of selective pressure might have specific effects on many genes' expression

Cannot find this using direct correlation and permutation analysis

- 1.5 million correlations
- Nothing is nominally significant

We can show that there is a relationship between gene expression and copy number by considering sample/sample distances (Mantel statistic)

- Compute 60x60 sample/sample distances using expression vectors
- Compute 60x60 sample/sample distances using copy number vectors
- There is a correlation, but it is not stunning

UCSF cancer center

How can we exploit annotation information to look at the direct effects? **Subsets!**

We can look at the direct effect by considering **subsets**

- Consider the set of genes that map to a particular genomic position
- Consider the set of BACs that map to the same place
- Are those genes' expression correlated with copy number at those loci?

Comparing subsets completes the argument

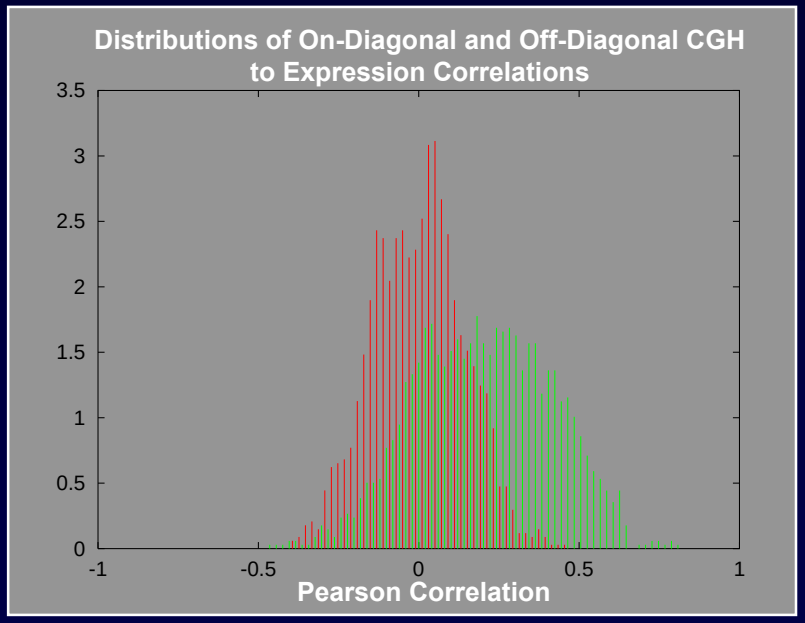
- Consider the set of gene/locus pairs that map within 1 Mb of one another
- Consider the set of gene/locus pairs that map greater than 50 Mb apart
- Are the correlations from (1) higher than from (2)?

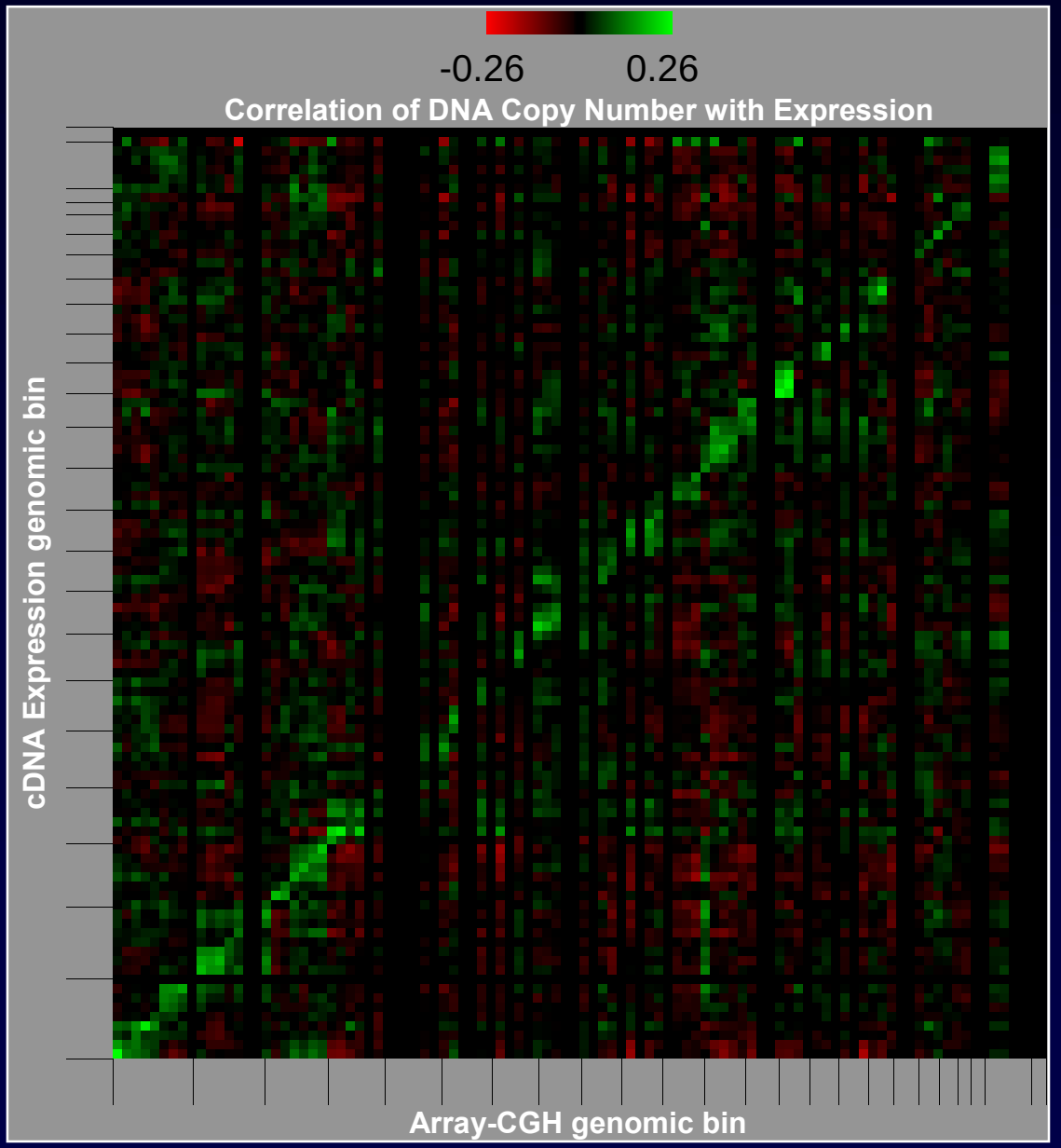
In order to do this, one must accurately map both cDNAs and BACs to genomic sequence.

UCSF cancer center

Subsets based on mapping: Genome-wide gene expression, on average, correlates with genomic copy number

The close-mapping pairs have significantly higher correlations than the distant-mapping pairs





UCSF cancer center

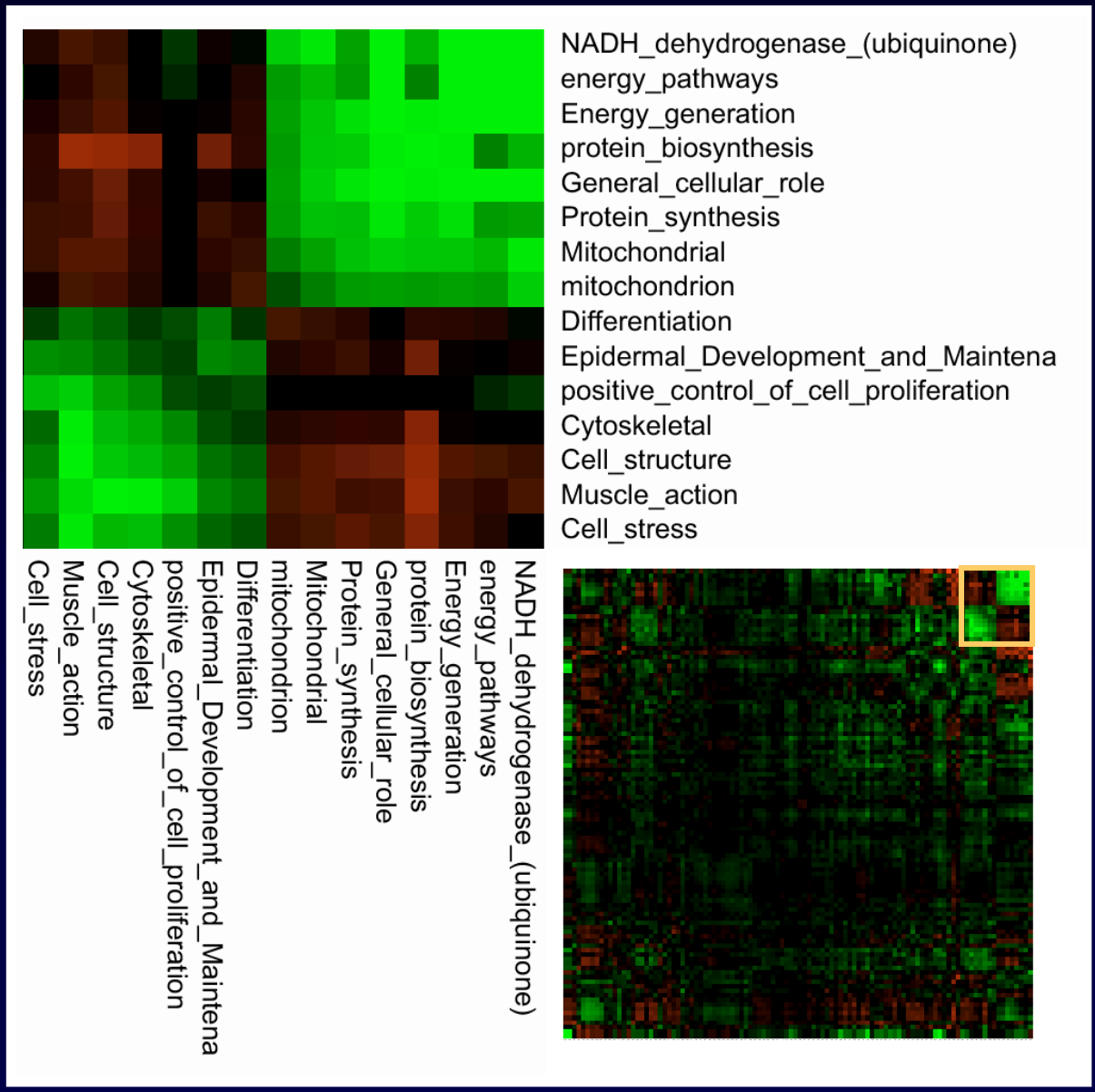
Subsets based on gene function: Gene Ontology information as a surrogate for pathway reveals enrichment of co-expression

Map cDNAs to curated NCBI RefSeqs

Use Gene Ontology and other controlled vocabularies as annotations of the genes

Expression of **sets** of genes that share annotations is correlated

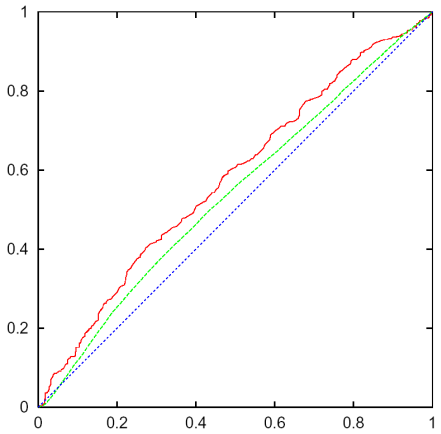
- Particularly in basic cellular metabolic systems
- Note: we have eliminated all identity correlations. Each bin at right contains the average of at least 50 non-degenerate correlations.



UCSF cancer center

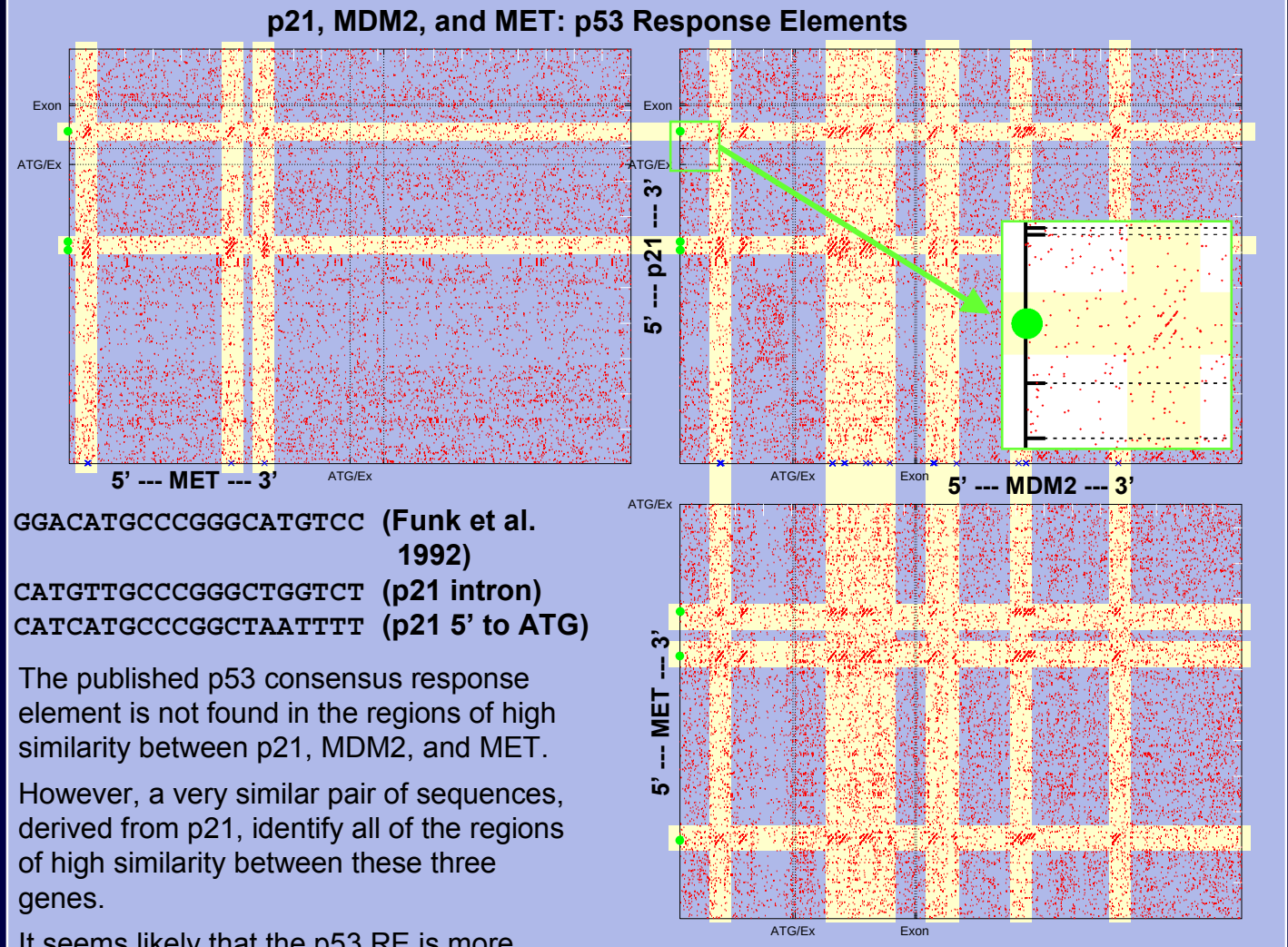
Subsets based on gene regulatory sequence similarity:

Genes pairs that are co-expressed have more similar regulatory regions than other pairs



ROC curve depicting separation of compositional similarity of co-expressed gene pairs from non-co-expressed gene pairs (green)

ROC curve depicting separation of expression correlation for gene pairs with high compositional similarity versus low similarity (red)



GGACATGCCCGGGCATGTCC **(Funk et al. 1992)**

CATGTTGCCCGGGCTGGTCT **(p21 intron)**

CATCATGCCCGGCTAATTTT **(p21 5' to ATG)**

The published p53 consensus response element is not found in the regions of high similarity between p21, MDM2, and MET.

However, a very similar pair of sequences, derived from p21, identify all of the regions of high similarity between these three genes.

It seems likely that the p53 RE is more complex than has been proposed. We are exploring the specificity of these sequences.

UCSF cancer center

We can answer simply constructed questions about biology

- Do specific DNA copy number abnormalities bear on patient outcome or tumor phenotype? **Yes**
- Is gene expression, genome wide, on average, quantitatively related to genomic copy number? **Yes**
- Are the regulatory regions of gene pairs that share patterns of expression more similar than those that don't? **Yes**
- Are the patterns of expression of gene pairs that have high similarity in regulatory regions more concordant than those with low similarity? **Yes**

UCSF cancer center

We want to be able to answer complicated questions about biology

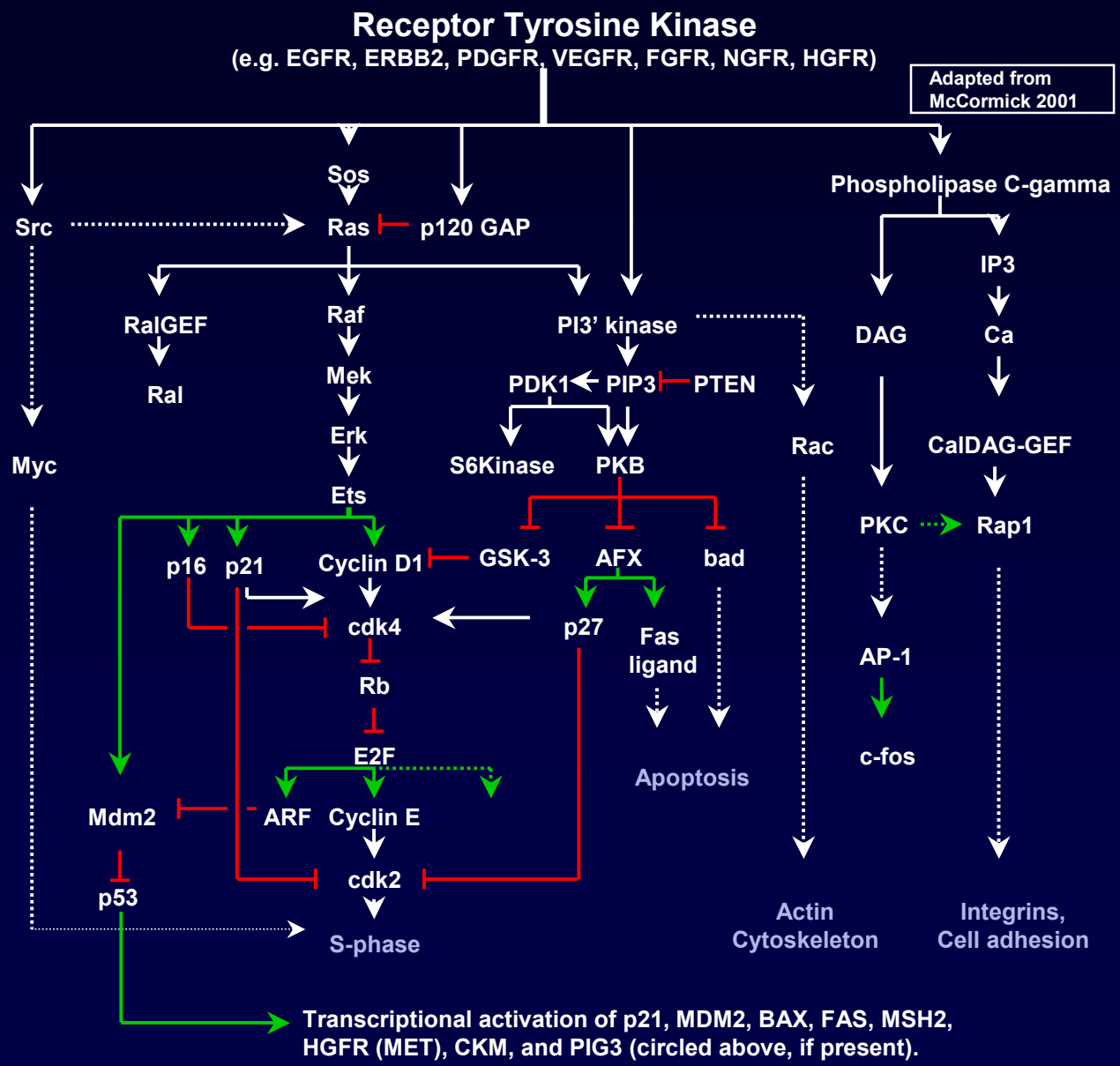
Are the genome copy number patterns of genes that impinge on S-phase checkpoint control quantitatively related?

Are other genes related in their pattern of aberration?

Can the context of the RTK pathway help in the analysis?

Bladder tumor data

- Waldman Lab (Joris Veltman)
- 41 tumors (9 Ta, 7 T1, 25 T2-4)
- ArrayCGH, both high-resolution (2000 clones) and oncogene focused arrays (500 clones).

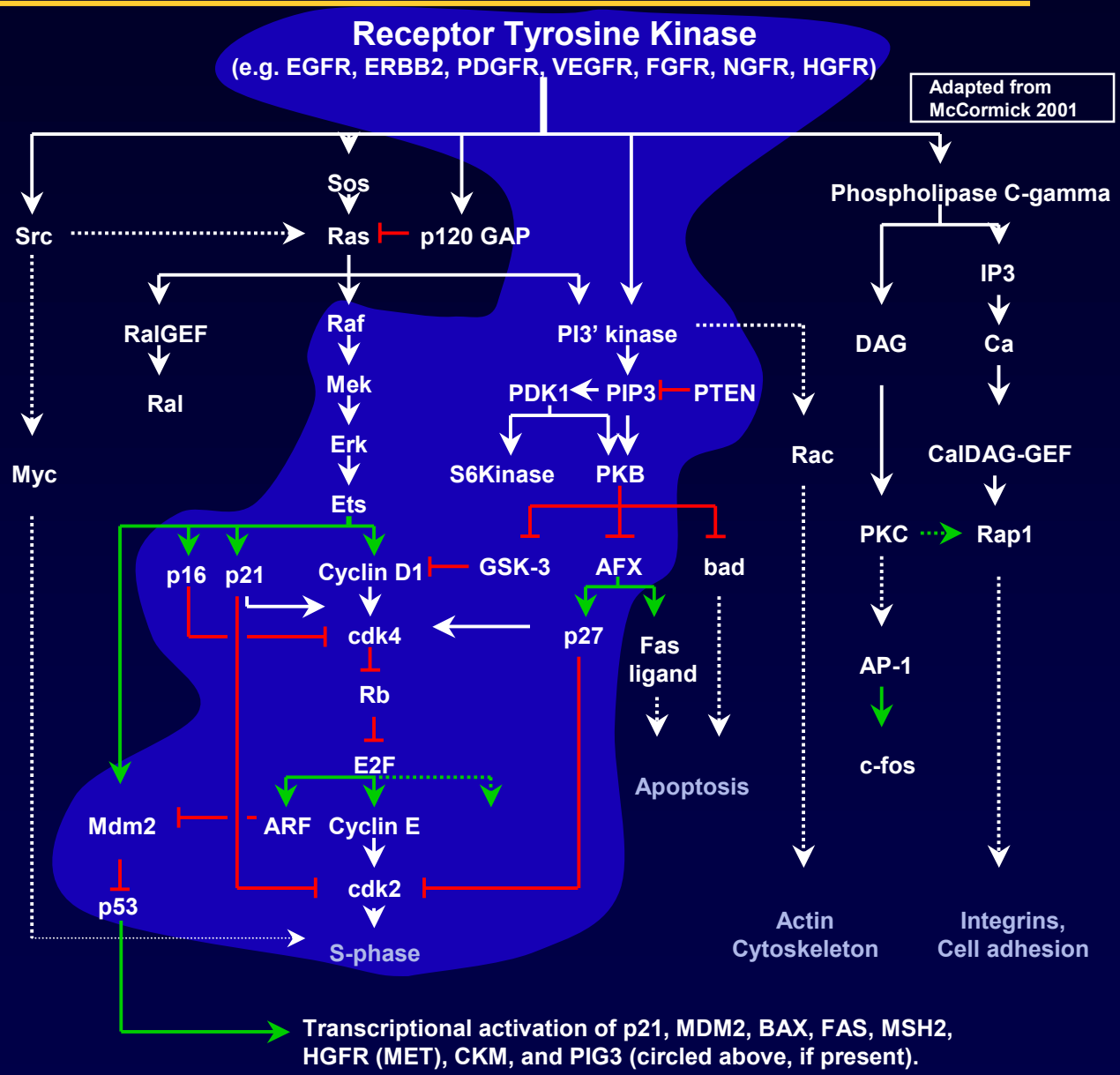


UCSF cancer center

We want to be able to answer complicated questions about biology

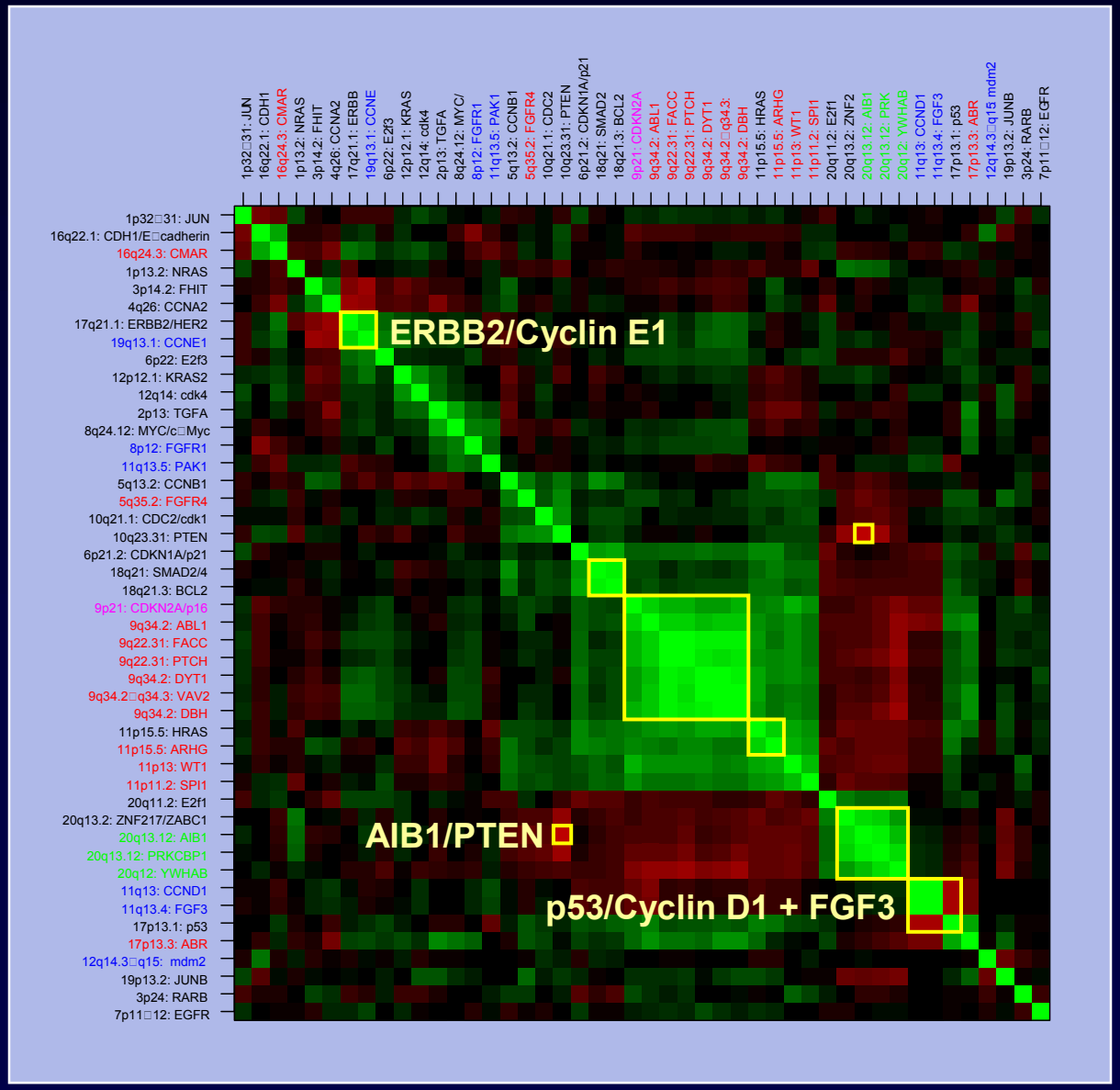
Gene/gene relationships in S-phase checkpoint control:

- Set 1: all genomic clones in the array-CGH experiment that are connected within 7 steps to S-phase control
- Set 2: all genomic clones that are frequently amplified or deleted
- Compute the correlation between copy number patterns of all gene pairs.
- Significance quantified by permutation analysis.



UCSF cancer center

Some gene pairs correlate



Gain of ERBB2 (17q12) and gain of CCNE (19q13.11)

Gain of AIB1 (20q13) and loss of PTEN (10q23)

Loss of p53 (17p13.3) and gain of CCND1 (11q13)

Loss of p53 (17p13.3) and gain of FGF3 (11q13)

Bladder tumor data

- Waldman Lab (Joris Veltman)
- ArrayCGH, both high-resolution (2000 clones) and oncogene focused arrays (500 clones).

We have adopted a very flexible data model: Entity attribute value

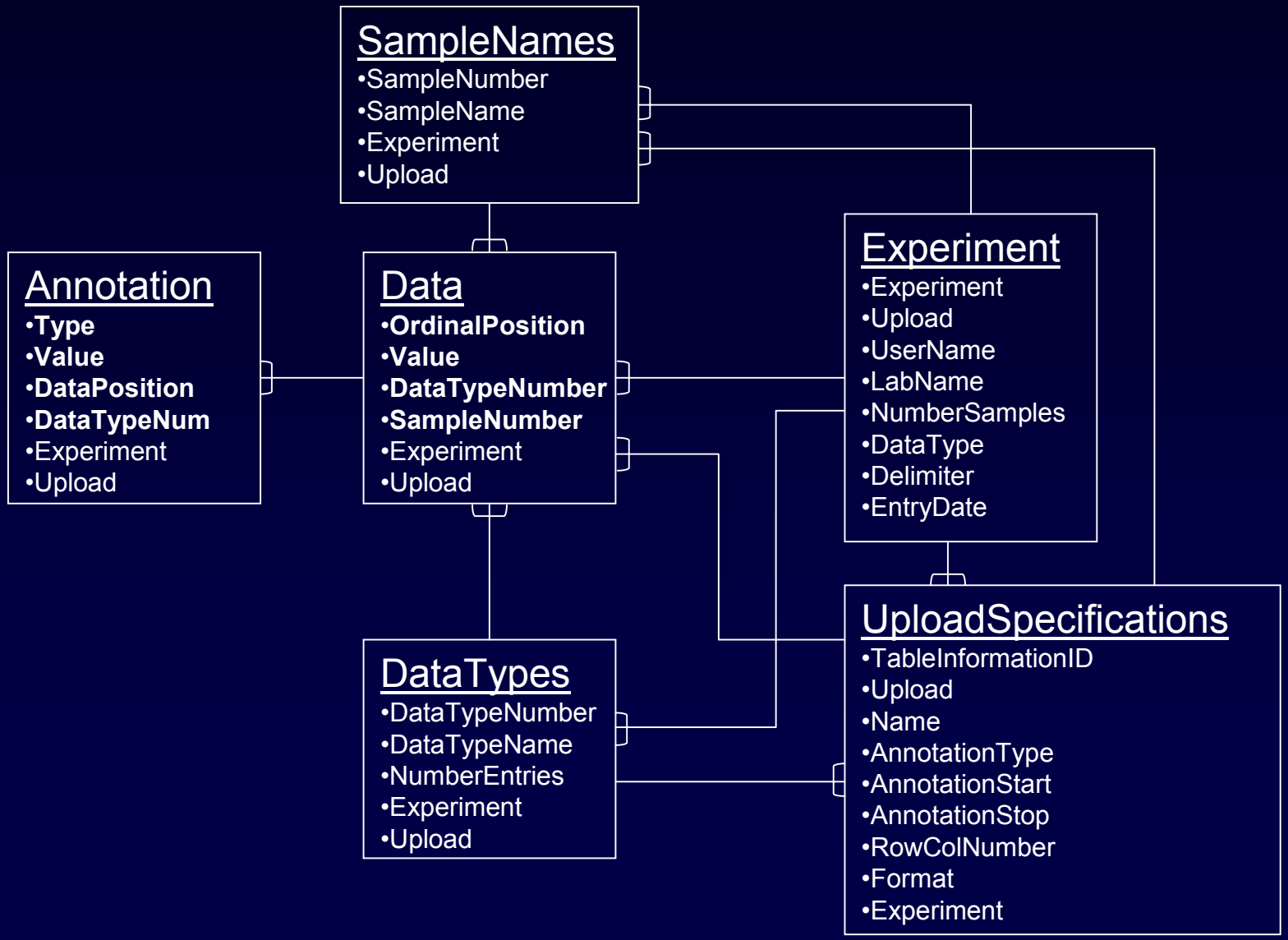
UCSF cancer center

User can define essentially any data type

No restriction on quantified biological data

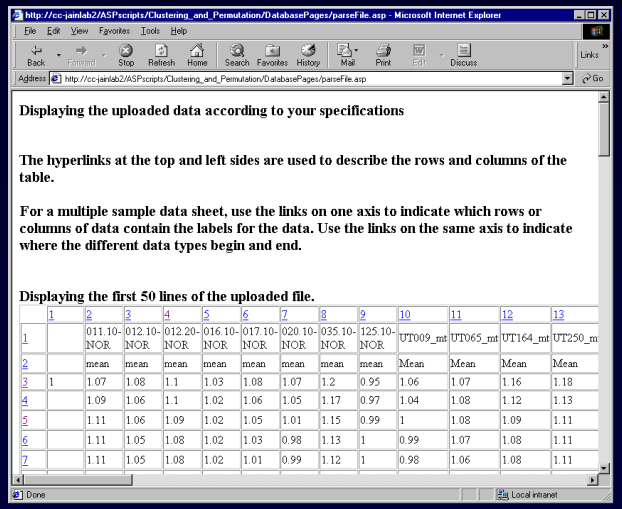
Same basic model can be used for representing free-form user annotation data

However, for the highly structured curated information, we are adopting more specialized models



UCSF cancer center

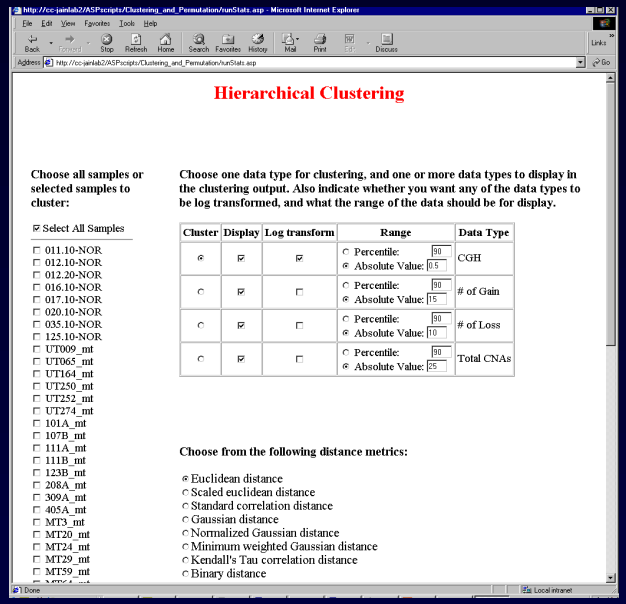
We are building a web-based data system that embeds these analysis and visualization tools



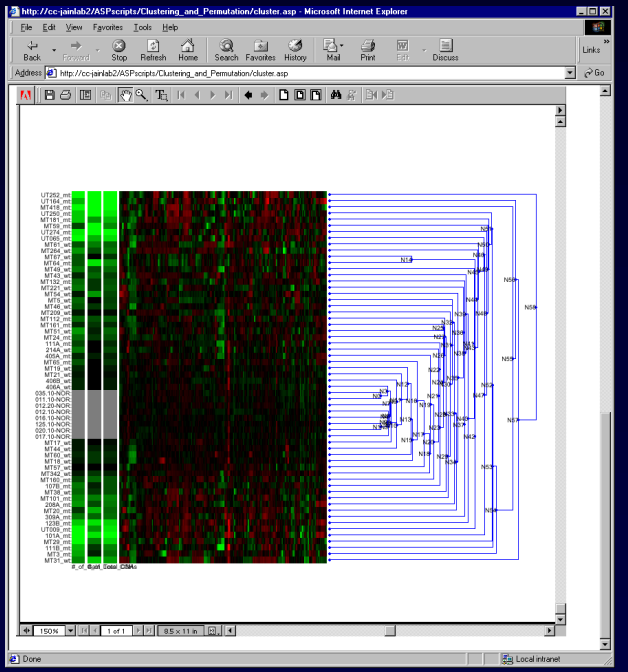
Choose a file to upload to the Web server.

Identify the data types and their positions within the file.

Data can be from *any* experiment of any type. The only restriction is that it can be represented as tab-delimited text for import to the system.



Several methods of analysis are available for selection. We have displayed hierarchical clustering.
The user can select samples, distance metrics, and additional information to display.



The system generates a PDF file that is displayed to the user's browser.

Conclusions

Biology's shift towards being a quantitative molecular science requires fundamentally new analytical methods

The problems inherent in array-based data are not insurmountable: it is often possible to derive quantitatively supportable conclusions

Integration of experimental data with systematically represented biological knowledge (annotations) can reveal relationships otherwise impossible to see by supporting definition of **meaningful subsets**

UCSF cancer center

Acknowledgements

Experimental collaborators

- Albertson Lab
- Collins Lab
- Gray Lab
- Pinkel Lab
- Waldman Lab

Jain Lab

- Jane Fridlyand, PhD
- Lawrence Hon
- Chris Kingsley
- Barbara Novak

UCSF Biological and Medical Informatics (BMI) PhD Students

- Taku Tokuyasu, PhD
- Adam Olshen, PhD
  [Now faculty at Sloan-Kettering]

Ajay N. Jain, PhD ajain@cc.ucsf.edu http://jainlab.ucsf.edu