UCSF cancer center

Integrated Data Systems for Interpreting Genome-Focused Data in Cancer

http://jainlab.ucsf.edu

Ajay N. Jain, PhD

Director of Informatics, UCSF Cancer Center

Associate Professor, Cancer Research Institute
and Dept. of Laboratory Medicine

ajain@cc.ucsf.edu

UCSF cancer center

What does my lab care about?

Predictive quantitative modeling across many scales that are relevant in biological and medical research.

- Molecular interactions
  - Protein ligand: flexible molecular docking; binding site models
  - Protein DNA: transcription factor binding site models

Lawrence Hon
BMI 3rd year

- Patterns of variation in measurable molecular species
  - Relationship to pathway structure
  - Induction of physical relationships

Barbara Novak
BMI 3rd year

  - Relationship to observable phenotype

Jane Fridlyand, PhD
Post-Doc

Taku Tokuyasu, PhD
Post-Doc

  - Relationship to clinical outcome
  - Integrative analysis combining experimental data with annotation information

Chris Kingsley
BMI 4th year

UCSF cancer center

Biology and medicine have become quantitative molecular sciences

Identification of all molecular species in a cell

- DNA sequence
- Genes, gene products
- Smaller organic molecules

3,000,000,000 DNA bases

~60,000 genes, gene products (plus variants)

Sugars, vitamins, membrane lipids, metabolic intermediates, cofactors…

Characterization of each species

- Concentration: over time, under different conditions
- Three-dimensional structure

DNA microarray technology: expression, DNA copy number

High-throughput methylomics

High-throughput proteome characterization

X-ray crystallography, NMR

Interactions between species

- Non-covalent interactions
- Covalent interactions
- Enzymatic transformations

Low-throughput methods: enzyme assays, Biacore, mass-spec approaches, hard biochemistry, clinical trials

Over the next 10-15 years, the bottlenecks will cease to be in the technology of measurement and characterization.

The bottlenecks will lie in analysis and interpretation, requiring individuals cross-trained in many disciplines.

UCSF cancer center

Complex quantitative relationships in heterogeneous data

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

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 very difficult 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

Distribution of correlations is nearly normal under the null hypothesis

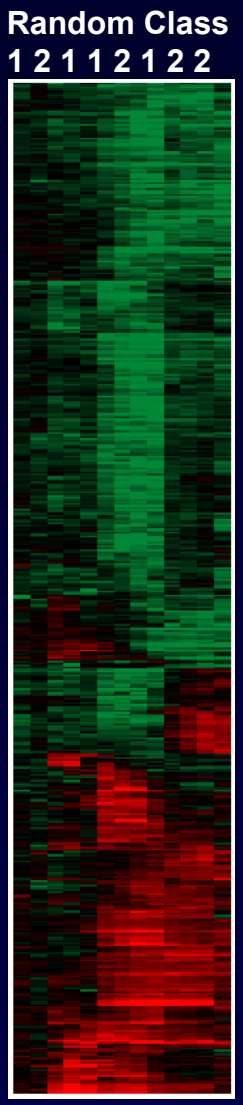
Consider some real gene expression data: 9712 genes in several conditions

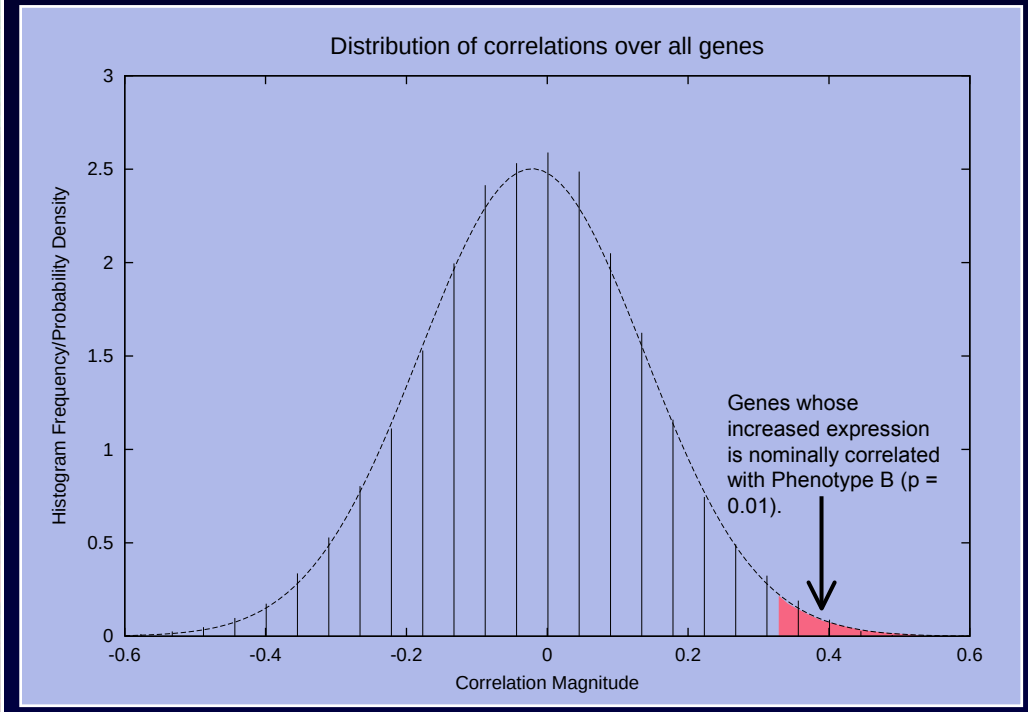
Assign random “outcomes” to it

What is the apparent correlation between relative gene expression level and class?

- Compute correlation of each gene’s expression to class membership
- Select those genes that exceed the p = 0.01 threshold for correlation

Conclude that 100 genes are correlated with class





The distribution of correlation of gene expression to class membership is close to normally distributed, even though the correlation is, by definition, random.

Iyer et al., Science 283: 83-87, 1999
Temporal response of fibroblasts to serum.

UCSF cancer center

What kinds of systematic questions do we have?

Questions about the samples

- Can we predict useful information about our samples?
  - Likelihood of recurrence
  - Survival
  - Histology when correlated with prognosis, particularly if histological assignment is tricky
- These questions give rise to new ones
  - Which sets of variables are causally related to the distinction we can predict?
  - The credit-assignment problem is generally very difficult

Questions about the variables

- Are any variables significantly correlated with an important distinction (accounting for the large number of measurements)?
- Are any of the variables related to each other based on their pattern over the samples?
  - Can we infer pathway commonality from genes whose temporal expression is similar?
  - Can we infer pathway linkage from genes that are positively or negatively correlated across samples?

UCSF cancer center

Practical Approach: Three simple things

Compute correlation for each measured variable; assess significance by permutation analysis

Look at clustering, assisted or not by variable selection, to visualize relationship of data to outcome

Use formal pattern classification (e.g. K nearest neighbors) to see if data can predict outcome; assess significance by cross-validation or blind testing

UCSF cancer center

Expression arrays: Lymphoblastic versus myeloid leukemia

Lander data

- 6817 unique genes
- Acute Lymphoblastic Leukemia and Acute Myeloid Leukemia (ALL and AML) samples
- RNA quantified by Affymax oligo-technology
- 38 training cases (27 ALL, 11 AML)
- 34 testing cases (20/14)

Science paper showed

- Able to predict ALL vs AML
- Able to cluster into the classes
- Methods were somewhat complex

REPORTS

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

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

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

SCIENCE VOL 286 15 OCTOBER 1999

UCSF cancer center

Permutation analysis: Initial correlation

Sample	Data									Outcome
1	0.99	0.98	0.98	0.97	0.97	0.95	0.95	0.95	0.96	1
2	1.15	1.11	1.07	1.04	1.01	0.99	0.98	0.96	0.96	1
3	1.11	1.14	1.22	1.3	1.37	1.39	1.39	1.39	1.37	1
4	1	1.01	1.01	0.99	0.96	0.93	0.91	0.89	0.88	1
5	1.04	1.01	0.97	0.94	0.93	0.92	0.9	0.9	0.91	1
6	1.17	1.25	1.32	1.38	1.43	1.46	1.5	1.53	1.55	0
7	1.12	1.16	1.2	1.26	1.34	1.42	1.49	1.54	1.53	0
8	0.96	0.97	0.97	0.97	0.96	0.96	0.97	0.98	0.98	0
9	1.03	1.04	1.05	1.06	1.07	1.09	1.1	1.12	1.17	0
10	1.16	1.19	1.21	1.23	1.25	1.25	1.26	1.27	1.28	0
	0.16	0.24	0.18	0.27	0.27	0.27	0.38	0.38	0.42	

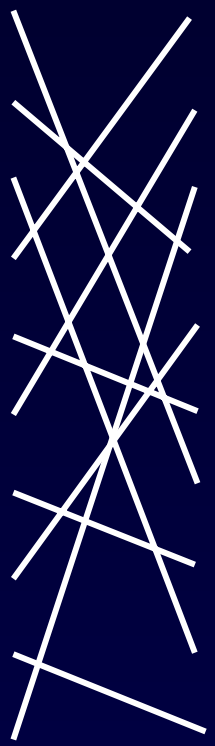
Correlation for each locus

Maximum magnitude correlation

UCSF cancer center

Permutation 1: Bogus correlation

Sample	Data								Outcome
1	0.99	0.98	0.98	0.97	0.97	0.95	0.95	0.95	0.96	1
2	1.15	1.11	1.07	1.04	1.01	0.99	0.98	0.96	0.96	1
3	1.11	1.14	1.22	1.3	1.37	1.39	1.39	1.39	1.37	1
4	1	1.01	1.01	0.99	0.96	0.93	0.91	0.89	0.88	1
5	1.04	1.01	0.97	0.94	0.93	0.92	0.9	0.9	0.91	1
6	1.17	1.25	1.32	1.38	1.43	1.46	1.5	1.53	1.55	0
7	1.12	1.16	1.2	1.26	1.34	1.42	1.49	1.54	1.53	0
8	0.96	0.97	0.97	0.97	0.96	0.96	0.97	0.98	0.98	0
9	1.03	1.04	1.05	1.06	1.07	1.09	1.1	1.12	1.17	0
10	1.16	1.19	1.21	1.23	1.25	1.25	1.26	1.27	1.28	0
	0.15	0.09	0.09	0.04	0.02	0.02	0.02	0.07	0.04	



Correlation for each locus

Maximum magnitude correlation

UCSF cancer center

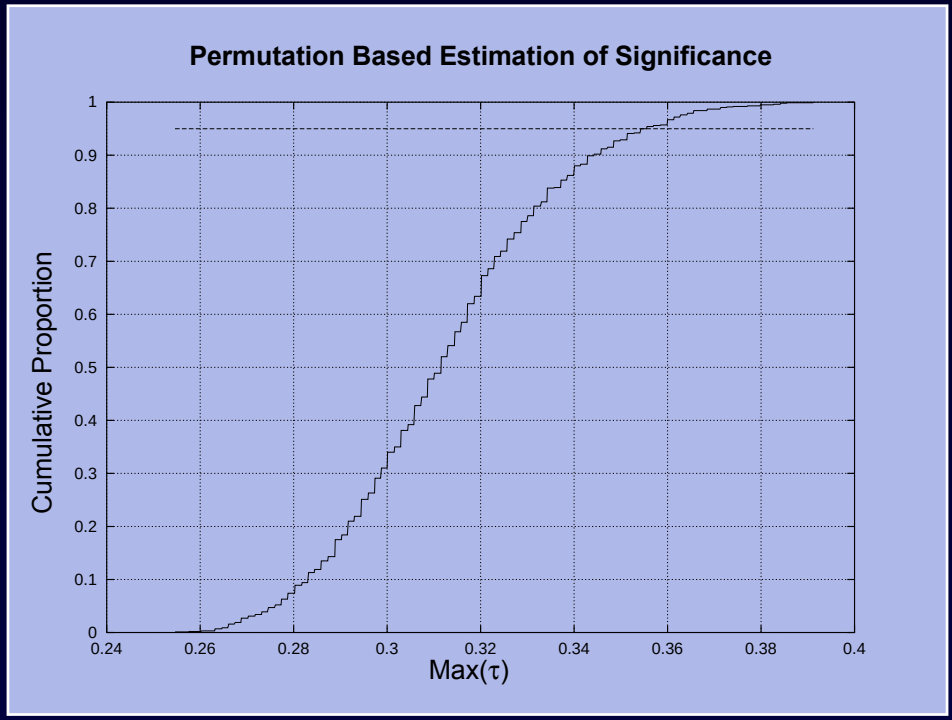
Repeated permutation yields a cumulative distribution

Unadjusted critical value

- $\tau = 0.17$
- Yields 1751 genes as "significant"
- Less than half confirmed on the test set

Adjusted critical value

- $\tau = 0.354$
- 51 genes significant
- 90% of these are confirmed on the test set



From the cumulative distribution, we observe that $\tau = 0.354$ corresponds to $p = 0.05$.

UCSF cancer center

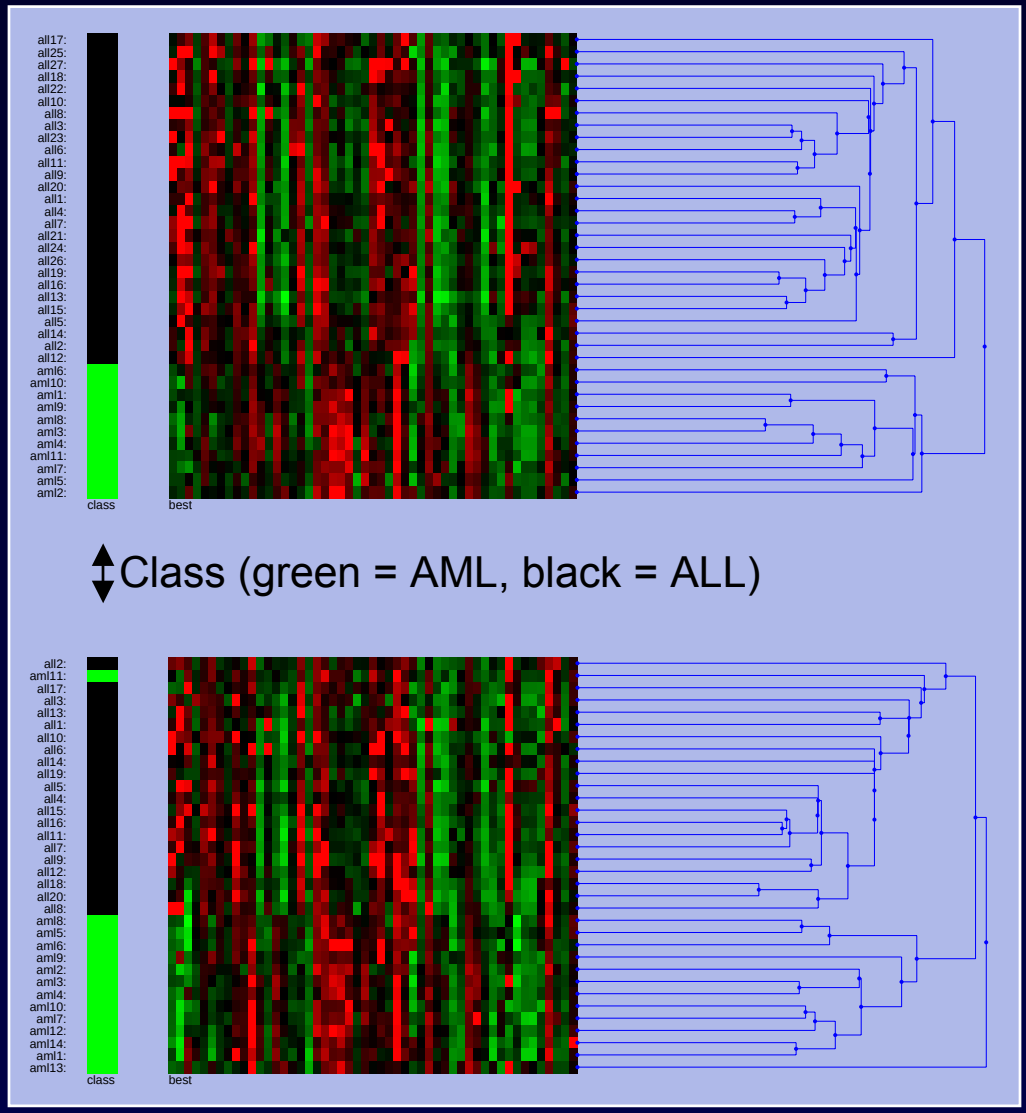
We cluster the data using the 51 genes selected on the training set

Procedure

- Selection of genes based on correlation
- Single-linkage hierarchical clustering
- Distance = (1-τ)

Results

- Training set
  - All samples segregate by class
  - Unsurprising (we selected the genes to do this)
- Testing set
  - Nearly all samples segregate by class
  - Suggests that KNN should work



UCSF cancer center

K nearest neighbors

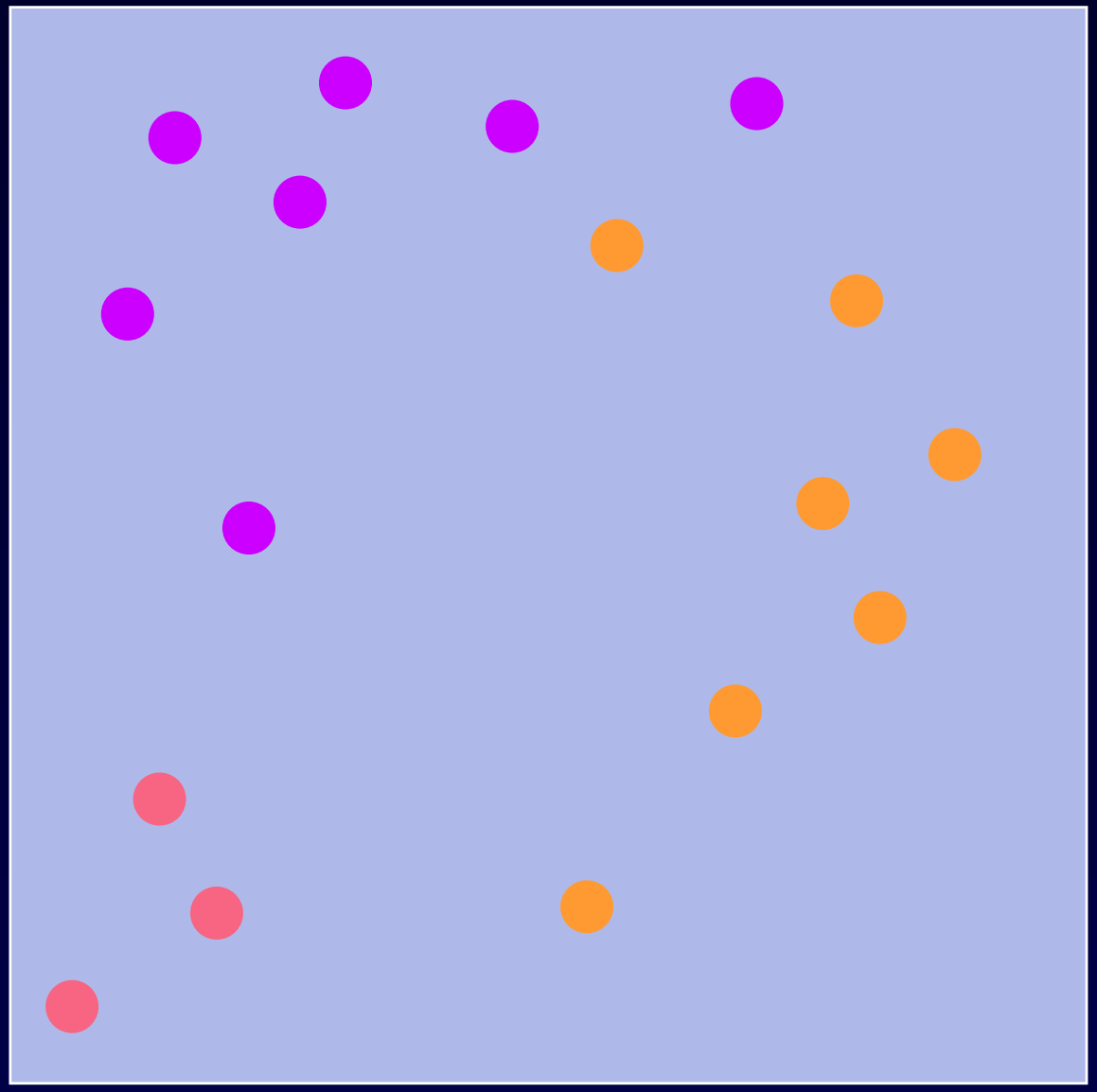
Data are represented as high-dimensional vectors

KNN requires

- Distance metric
- Choice of K
- Choice of which elements belong in the vectors

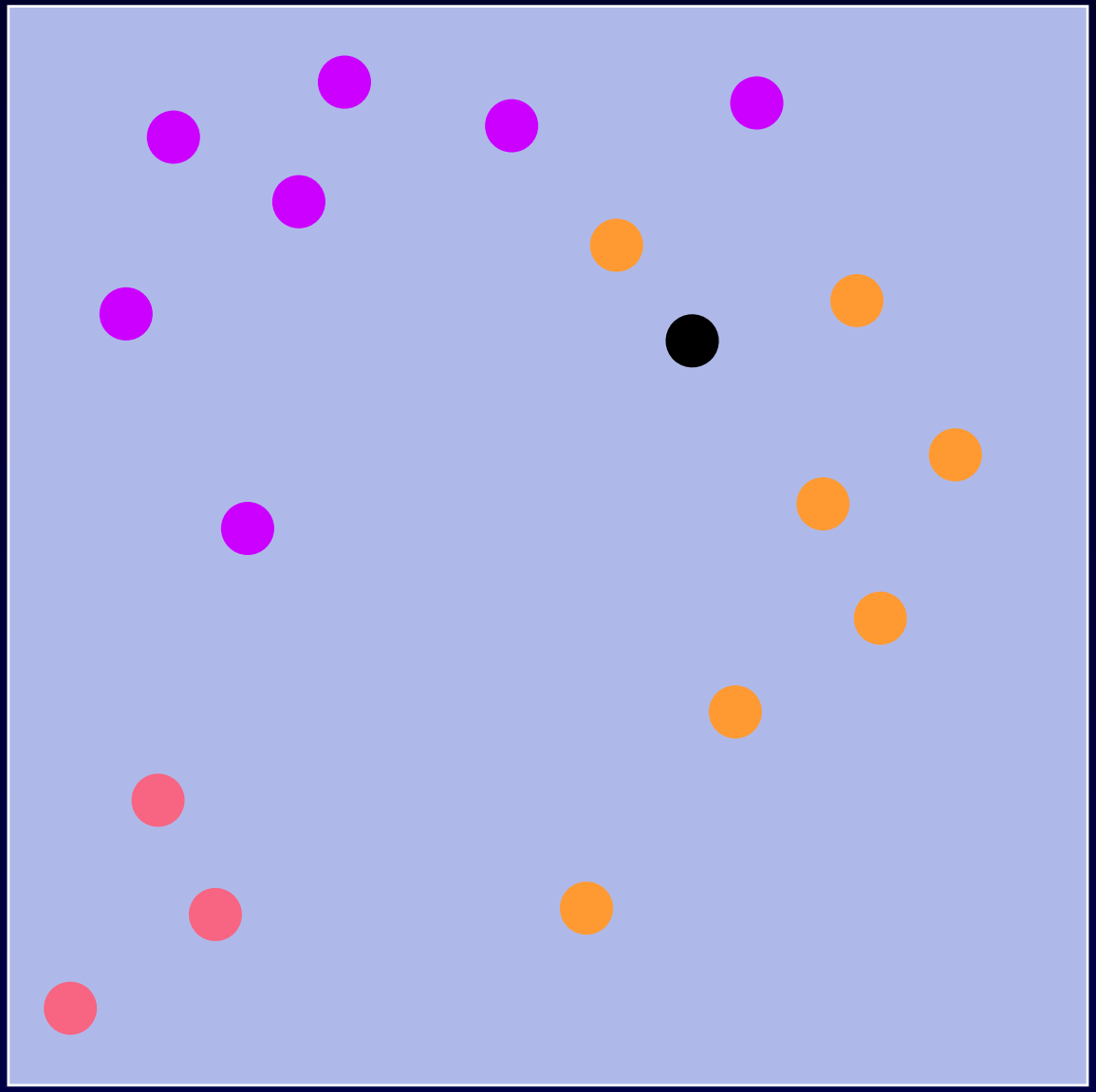
Given a new example

- Compute distances to each known example
- Choose class of most popular



K nearest neighbors

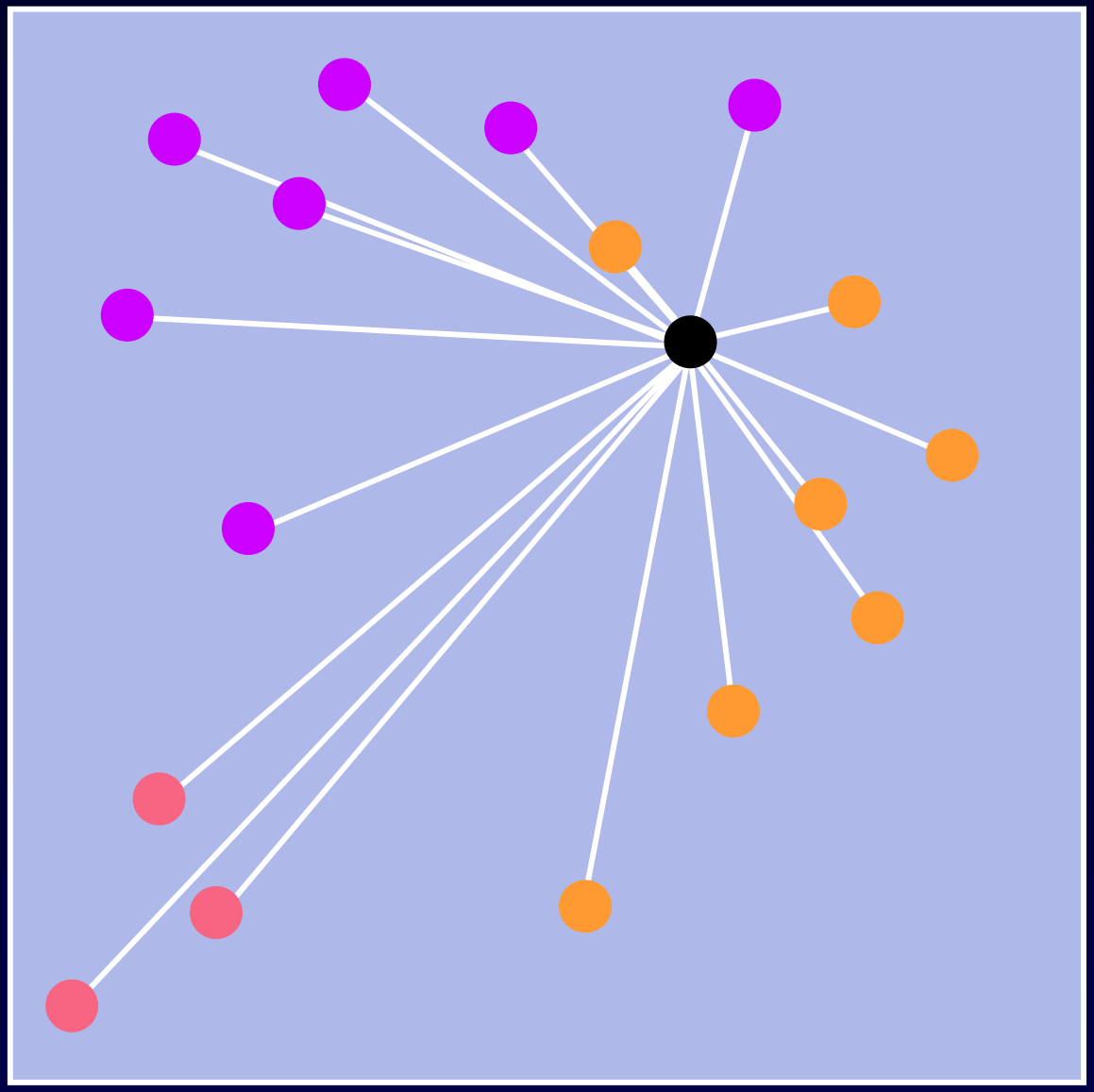
New item



K nearest neighbors

New item

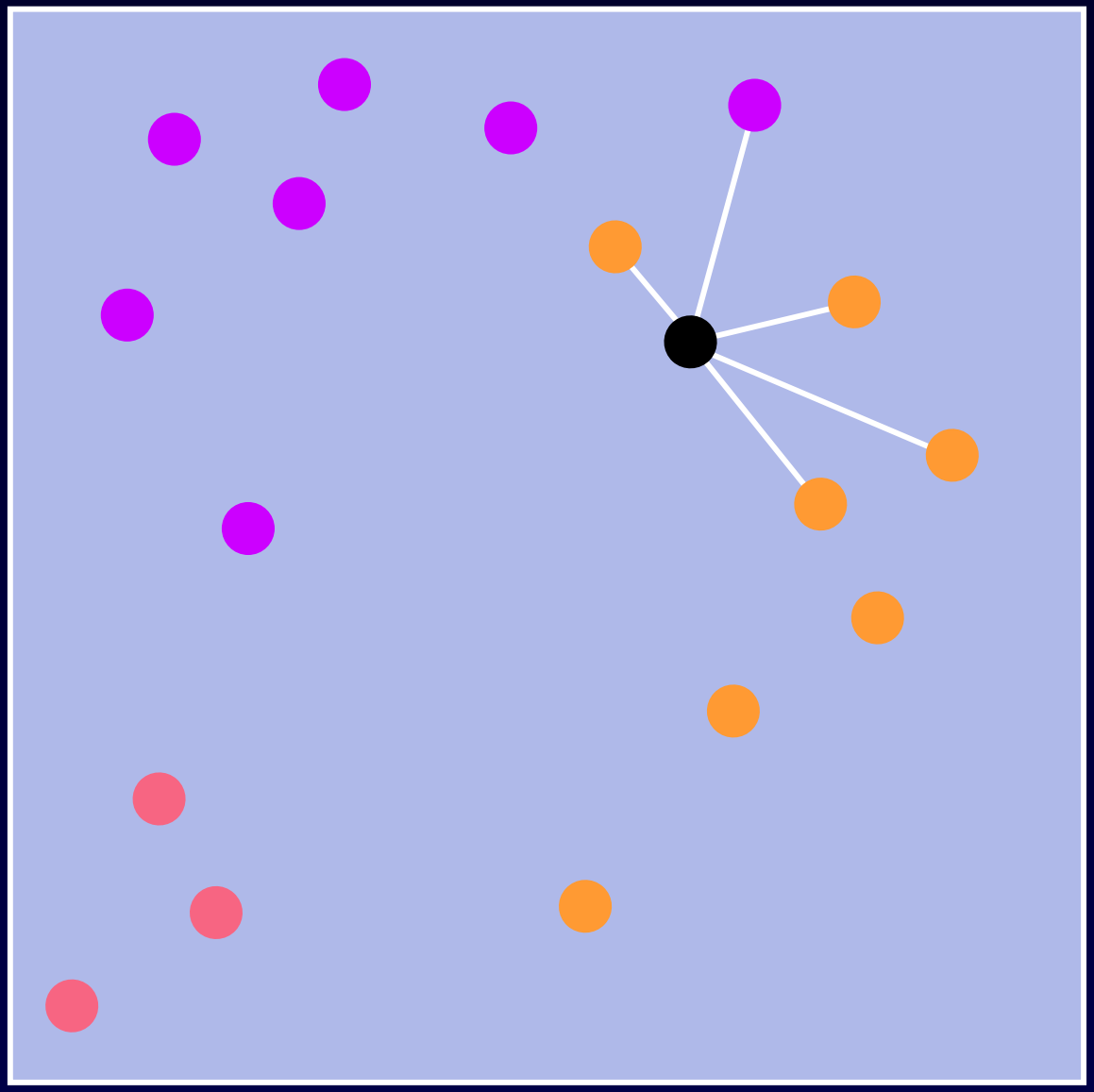
- Compute distances



K nearest neighbors

New item

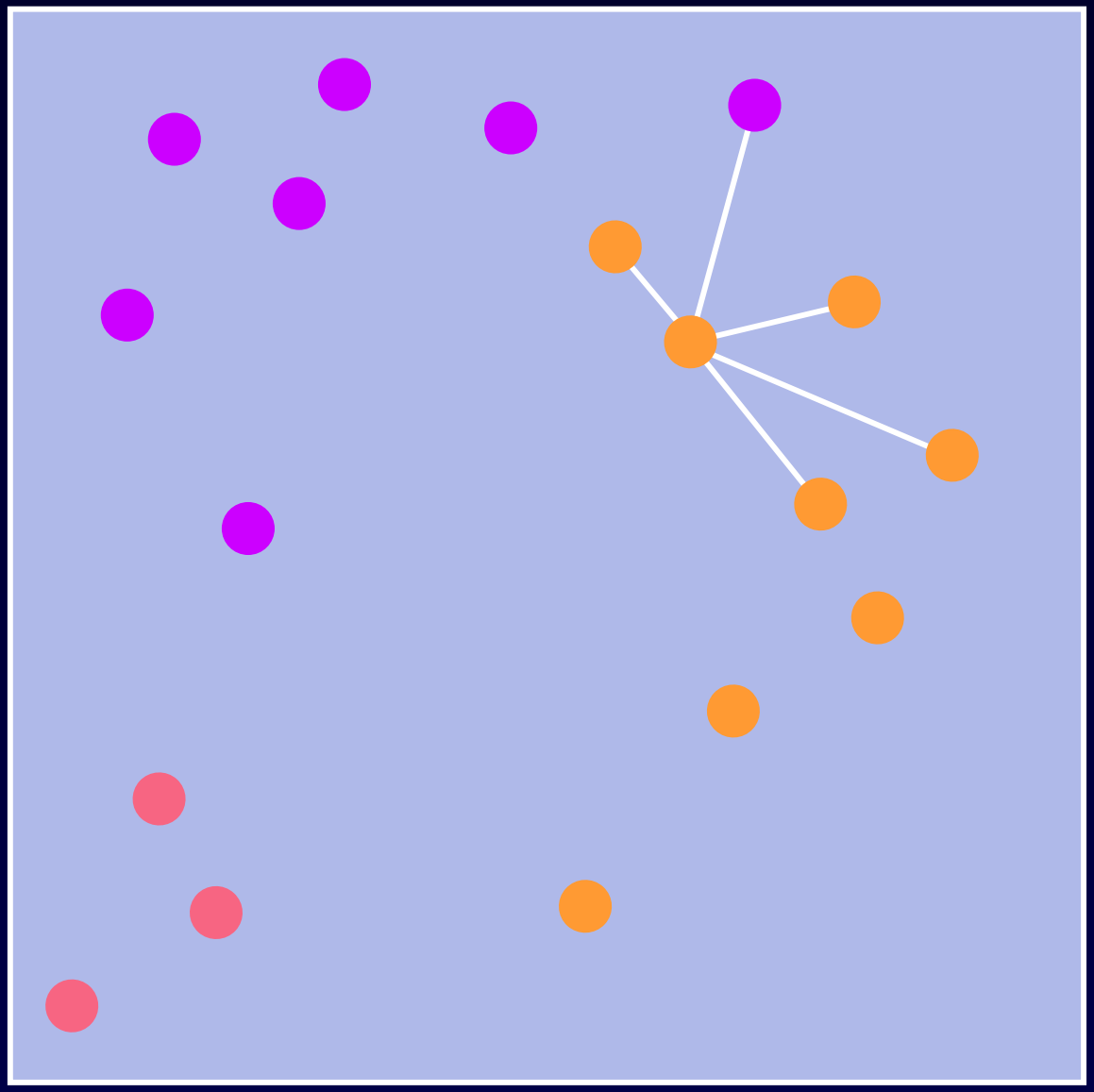
- Compute distances
- Pick K best distances



K nearest neighbors

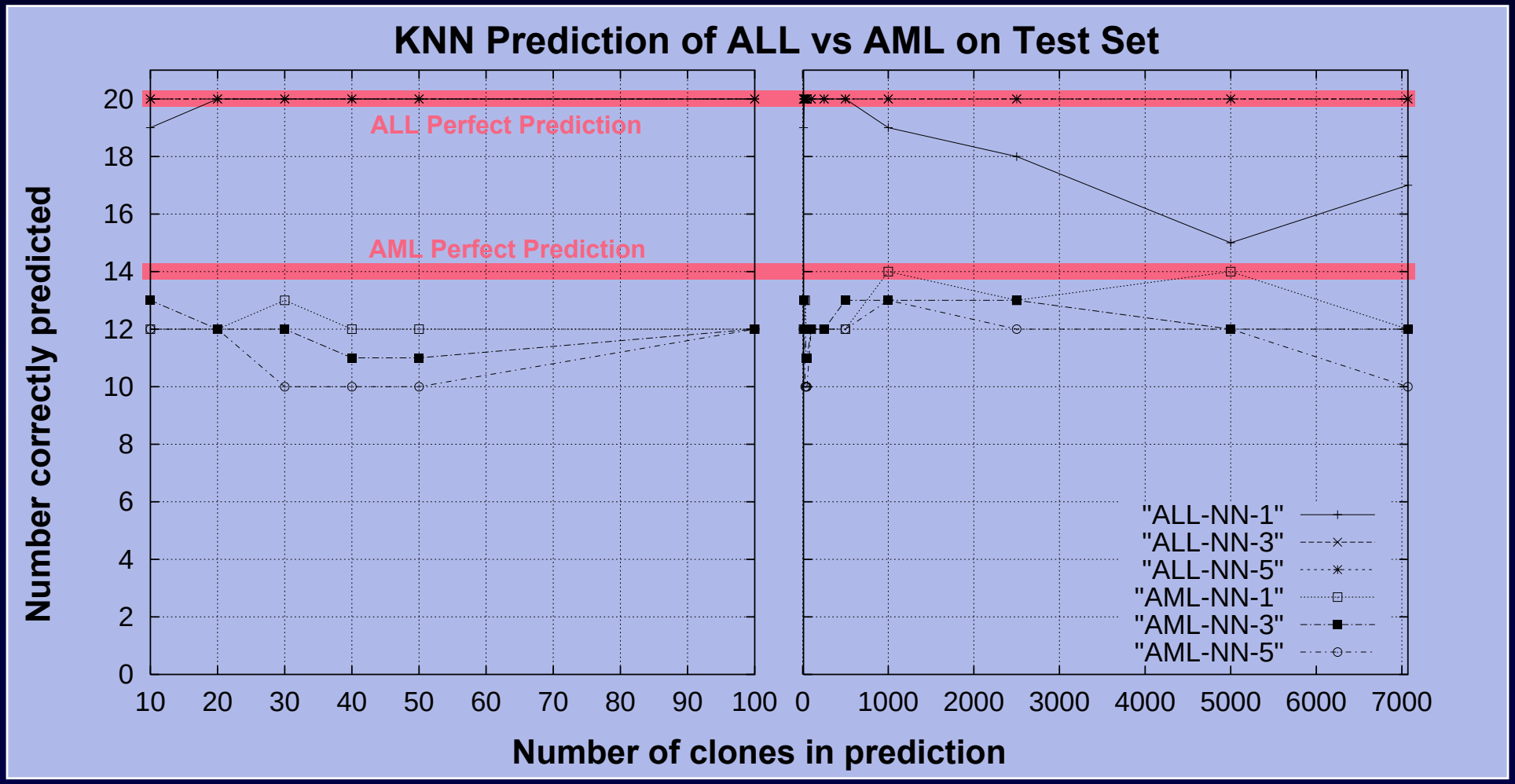
New item

- Compute distances
- Pick K best distances
- Assign class to new example



UCSF cancer center

KNN applied to the AML/ALL set yields robust results: 90-100% predictive accuracy



Multiple values of K (1, 3, 5)
Multiple sizes of the selected gene subset (10 to 7071)

Everything works on this data set.

UCSF cancer center

Expression data: The three simple methods often work

BIOINFORMATICS

Vol. 18 no. 0 2002
Pages 1–10

Deriving quantitative conclusions from microarray expression data

Adam B. Olshen [1,2] and Ajay N. Jain [1,3]

[1]Comprehensive Cancer Center, Cancer Research Institute, and Department of Laboratory Medicine, University of California, San Francisco, CA 94143-0128, USA

Received on July 21, 2001; revised on December 10, 2001; January 19, 2002; accepted on January 24, 2002

- Golub/Lander set (AML/ALL): Confirms authors' observations,
  Occasionally exceeds performance
- Perou set (Breast tumors): Additional quantitative support for authors' conclusions,
  Possible to find mRNA to protein links for ERBB2 and ER without prior knowledge
- Alizadeh set (Lymphoma): Cannot find direct link between survival and genes,
  Strongly supports the observation of two subtypes of DLBCL,
  Open question about the importance of particular genes to the subtypes

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*‖

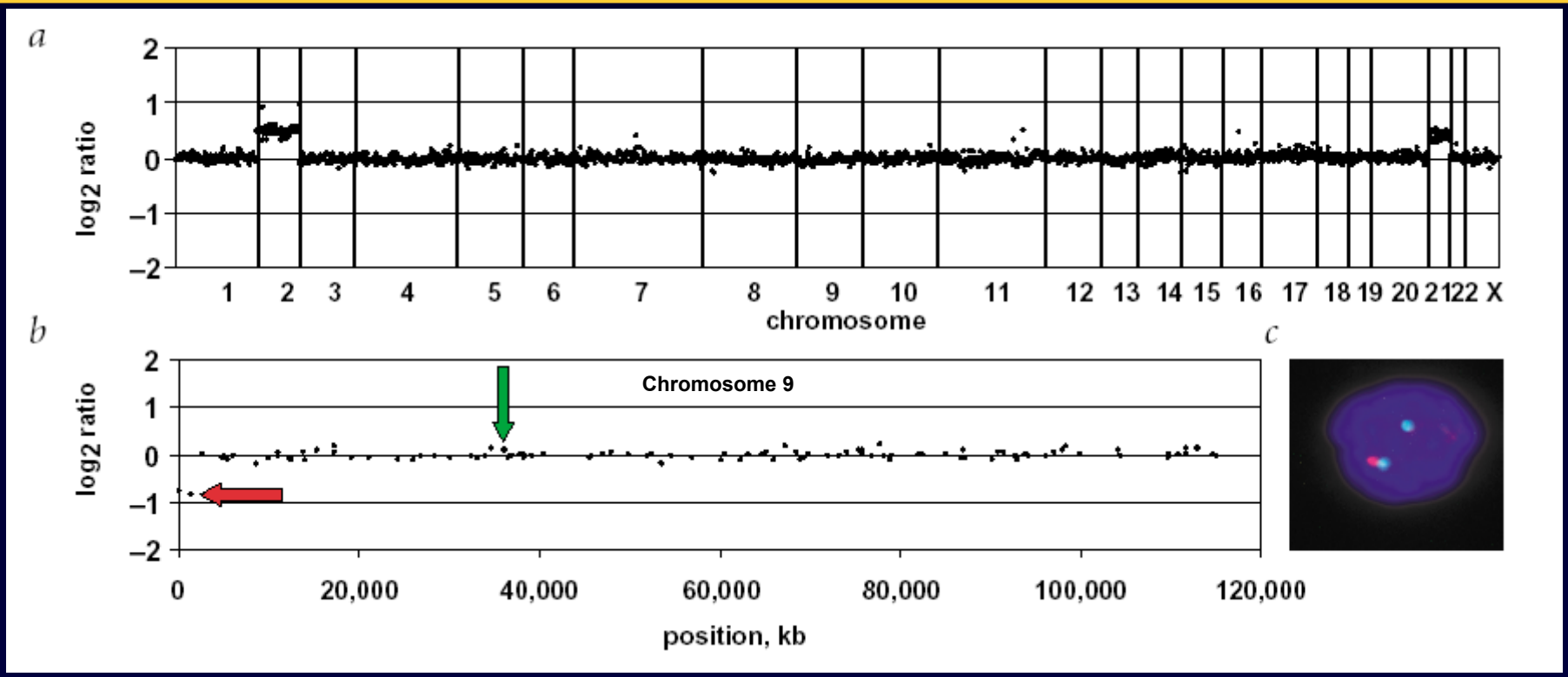
*UCSF Cancer Center, University of California, San Francisco, Box 0128, San Francisco, CA 94143-0128; ‡Departments of Genetics and Oncology, Institute for Cancer Research, Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway; §Department of Oncology, Haukeland Hospital, 5021 Haukeland Sykehus, Norway;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

Genome wide array-based CGH: 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.

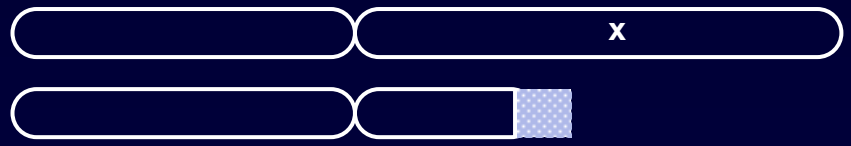
Bottom left shows copy number for cell strain GM03563 on chromosome 9. Bottom right: single copy deletion verified by FISH.

UCSF cancer center

Characteristics of Copy Number Changes in Cancer

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

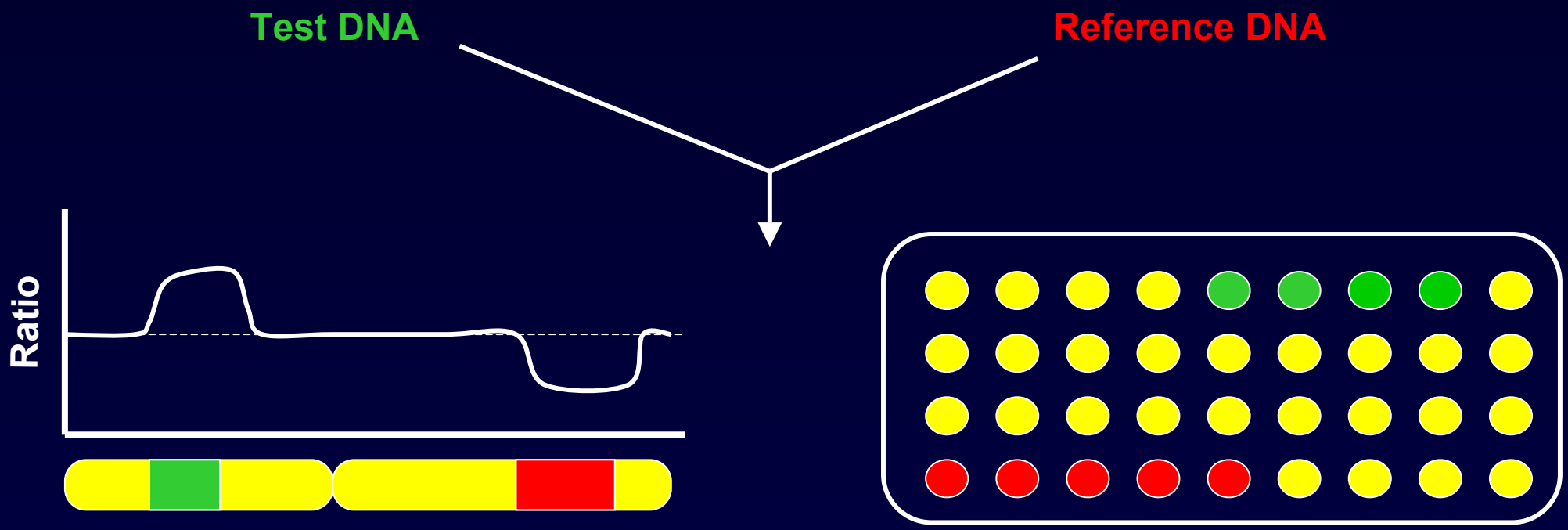
Mutation plus terminal deletion of tumor a suppressor gene



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

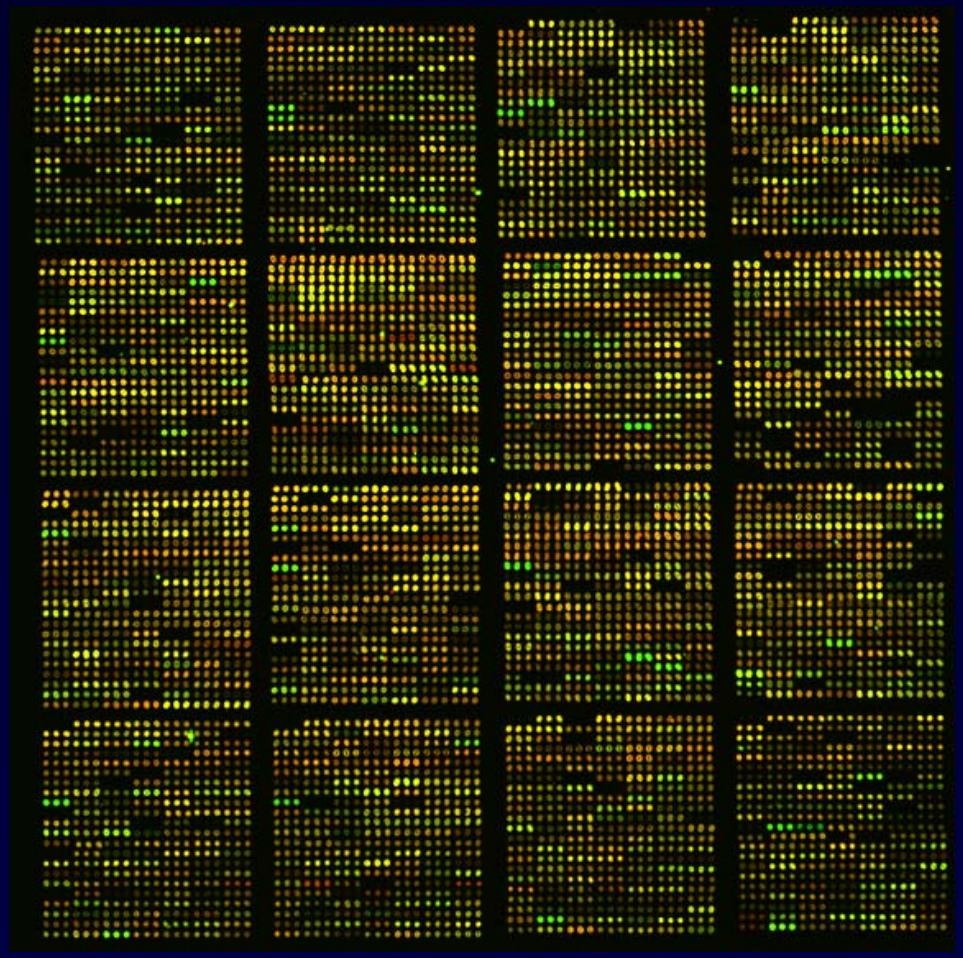
Genome Scanning Array with ~ 1.4 Mb Resolution

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

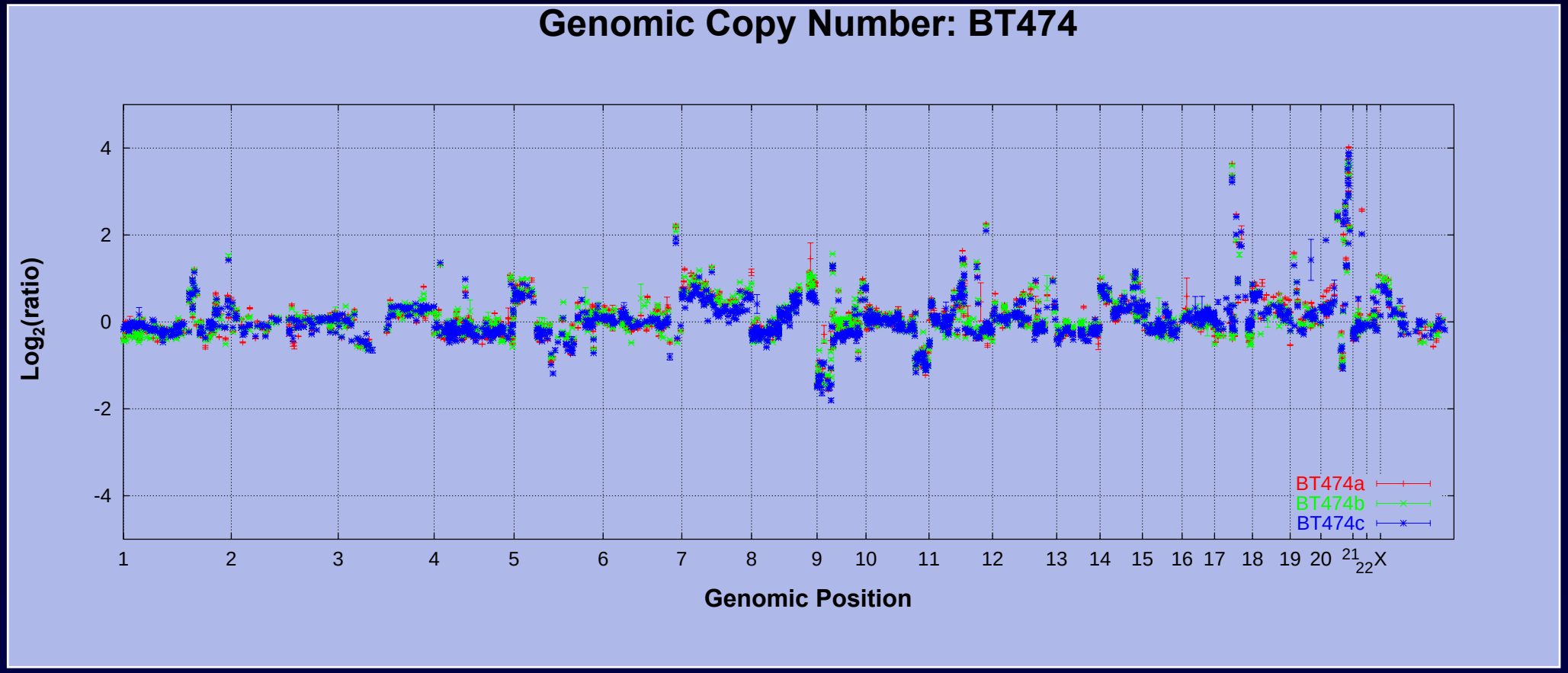
Triplicate spots

130 μm centers

864 well plates

UCSF cancer center

We see very complex genomic patterns



UCSF cancer center

CGH Analysis: Visualization and Correlation with Outcome

Data (J. Gray, K. Chin)

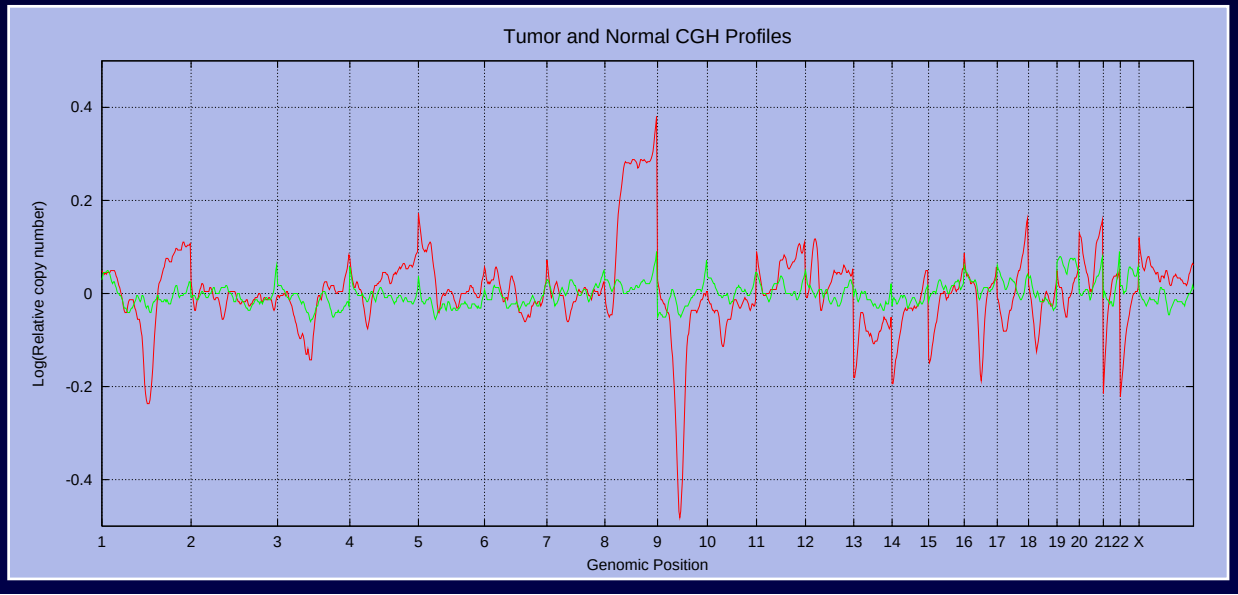
- 60 CGH profiles
  - 1225 “observables”
  - 52 tumor profiles
  - 8 normal profiles
- Patient information
  - Age of onset
  - Overall survival
  - Disease free survival
  - Alive or dead
- Tumor status
  - Size/Stage
  - Estrogen receptor
  - Progesterone receptor
  - p53

Is there a statistically significant correlation between CGH profile similarity and outcome (e.g. survival)?

If so, is the correlation degenerate?

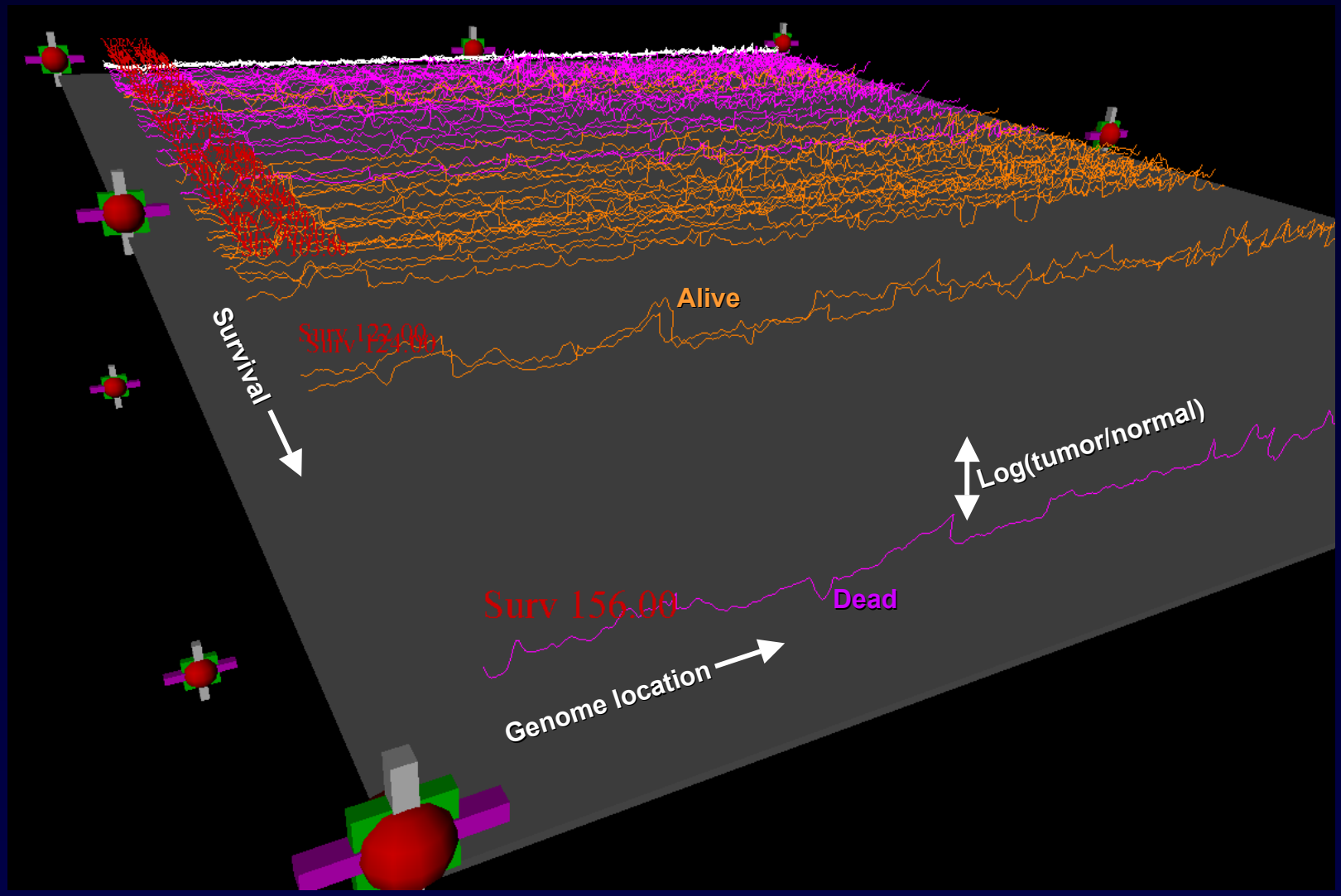
- How can we visualize the data in order to assess this?
- How can we test hypotheses generated from this exercise?

What are the genomic variations that are directly correlated with outcome?



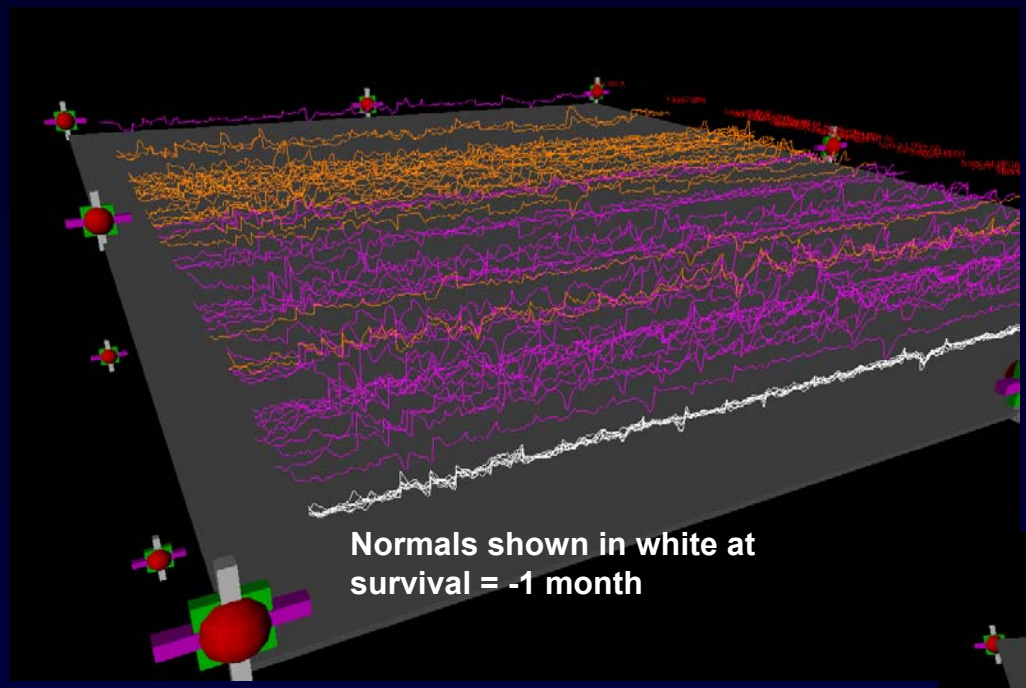
UCSF cancer center

We can visualize complex profile data using 3D virtual worlds

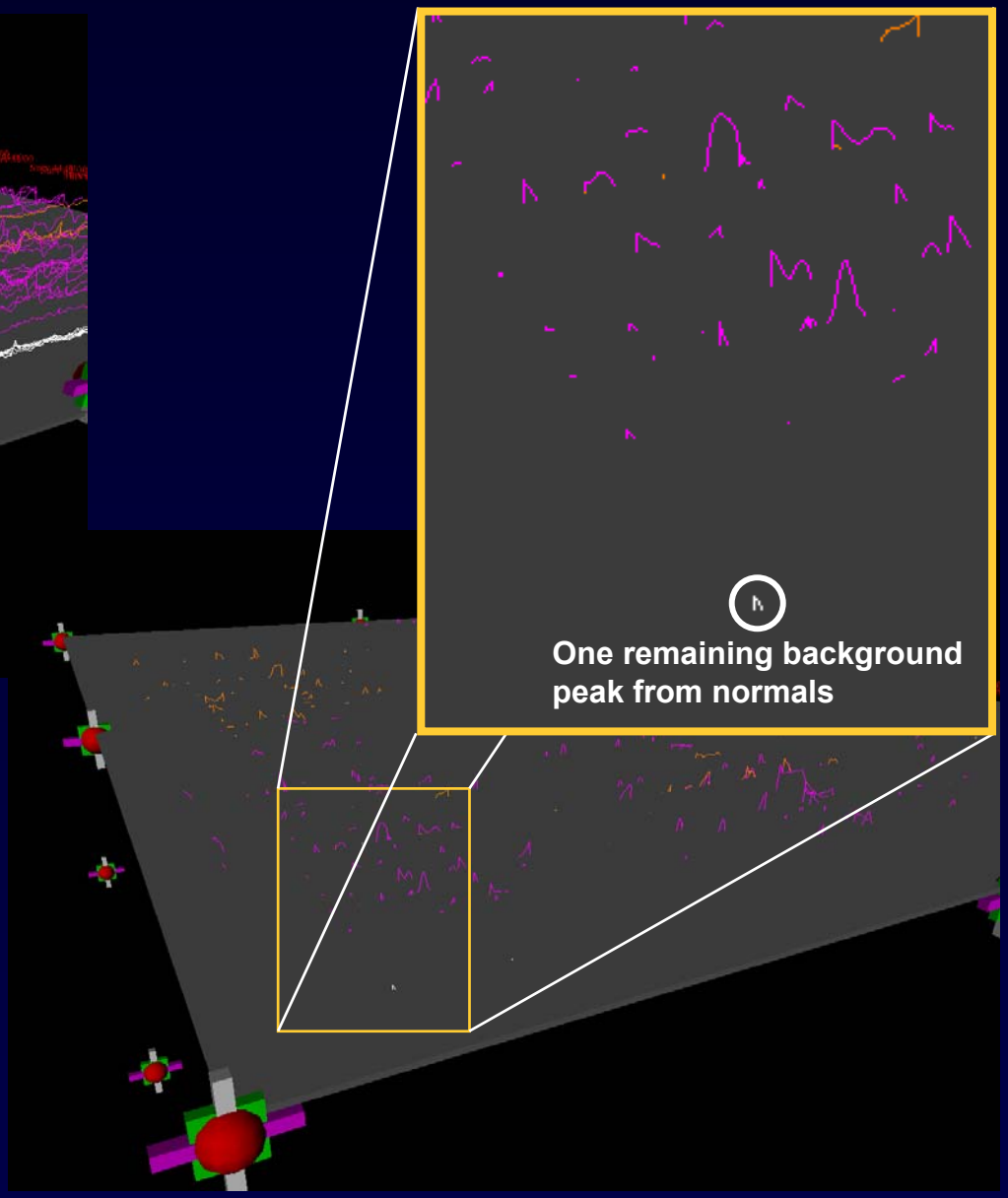


UCSF cancer center

By sliding the opaque XZ plane, we can select peaks above background



Normals shown in white at survival = -1 month



One remaining background peak from normals

UCSF cancer center

General Principle: Reduce the number of observations

Any method we can use to subselect a smaller set of observations from the larger set helps us, provided:

- The subselection method must be orthogonal to the correlation being studied
  - If we're trying to link copy number to survival, we can't systematically employ the survival outcomes in making our subselection
- Ideally, the method should have some compelling intuitive support based on the data
- Restricting observations based on frequency/magnitude is a generally useful technique: it tends to eliminate noise

UCSF cancer center

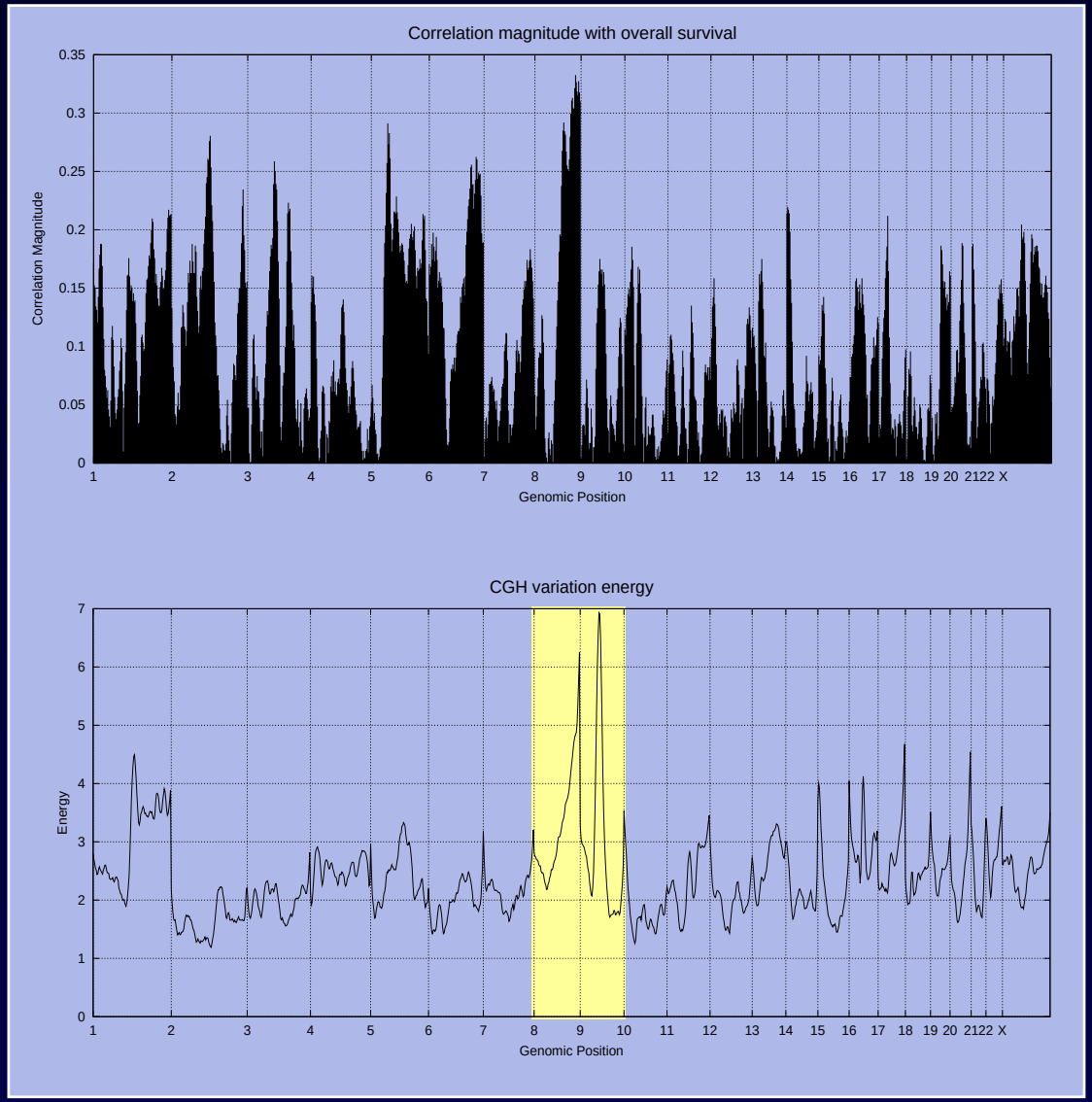
Chromosomes 8 and 9 have substantial variations and show some correlation

Compute the direct correlation for each of 1225 loci

- Strongest correlation at 8q24
- Many other peaks

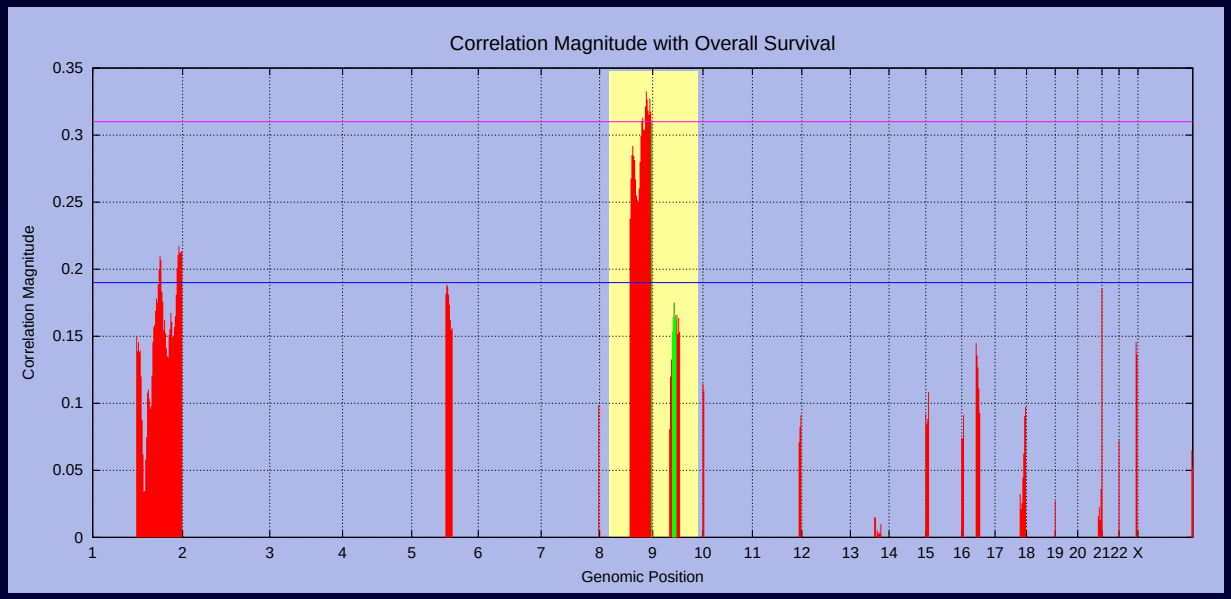
Compute the degree of aberration at each locus over the tumor samples

- We can focus our analysis on these loci
- Reduces the data bandwidth



UCSF cancer center

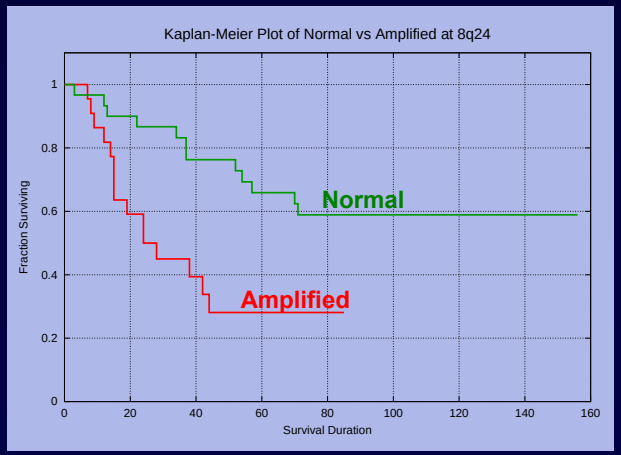
Both 8q24 and 9q13 are significantly correlated with survival

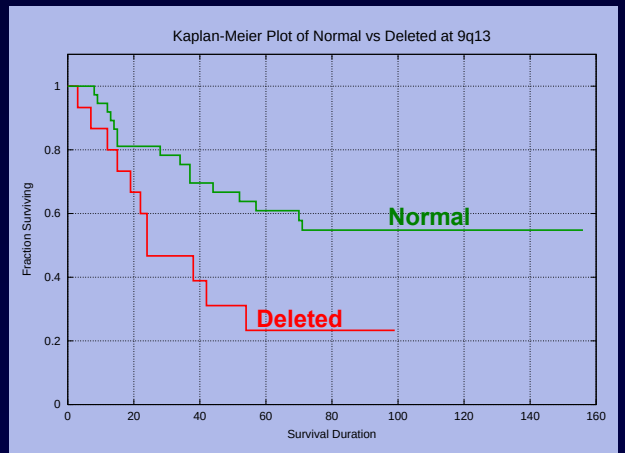


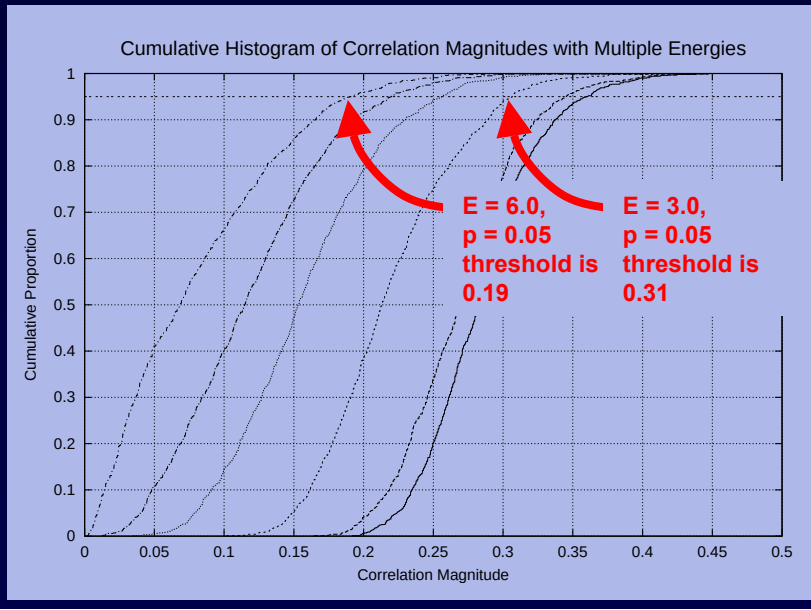
Permutation analysis for significance:

Do 10,000 times:

- Scramble tumor to CGH mapping
- Compute “correlation”
- Record magnitude of maximum correlation
- Plot CDF of results
- Set threshold to eliminate 95% of the distribution







UCSF cancer center

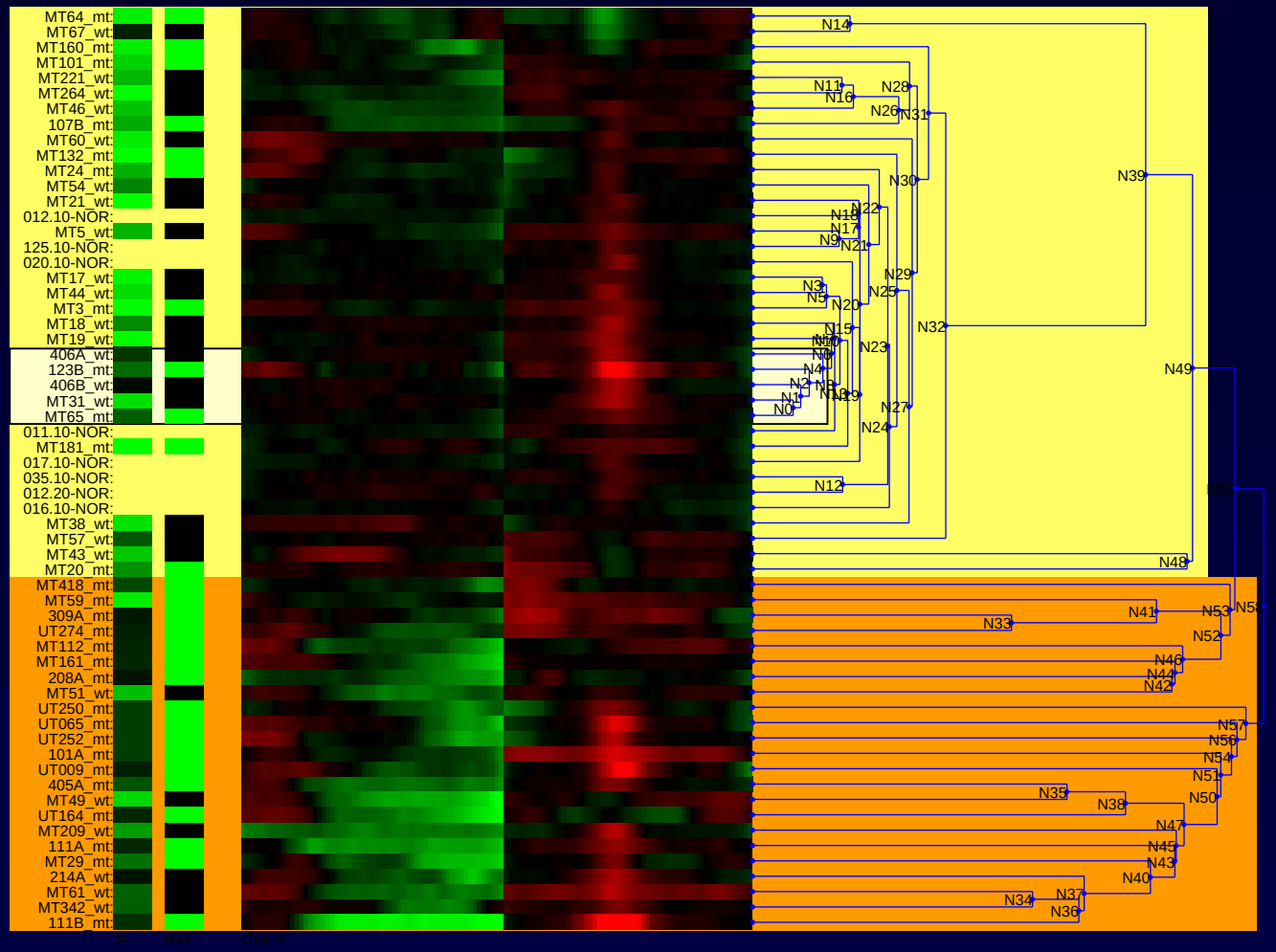
Clustering based on chromosomes 8 and 9 reveal patterns of survival and tumor phenotype

Cluster profiles based on Chr 8,9

- Display raw data
- Display survival, p53 status

Cluster enrichment is statistically significant

- Orange block
  - Surv < 35 months
  - p53 often mutant
- Yellow block
  - Surv > 75 months
  - p53 often wt



p53 status (green = mut, black = wt)

Survival (black = low, green = high)

UCSF cancer center

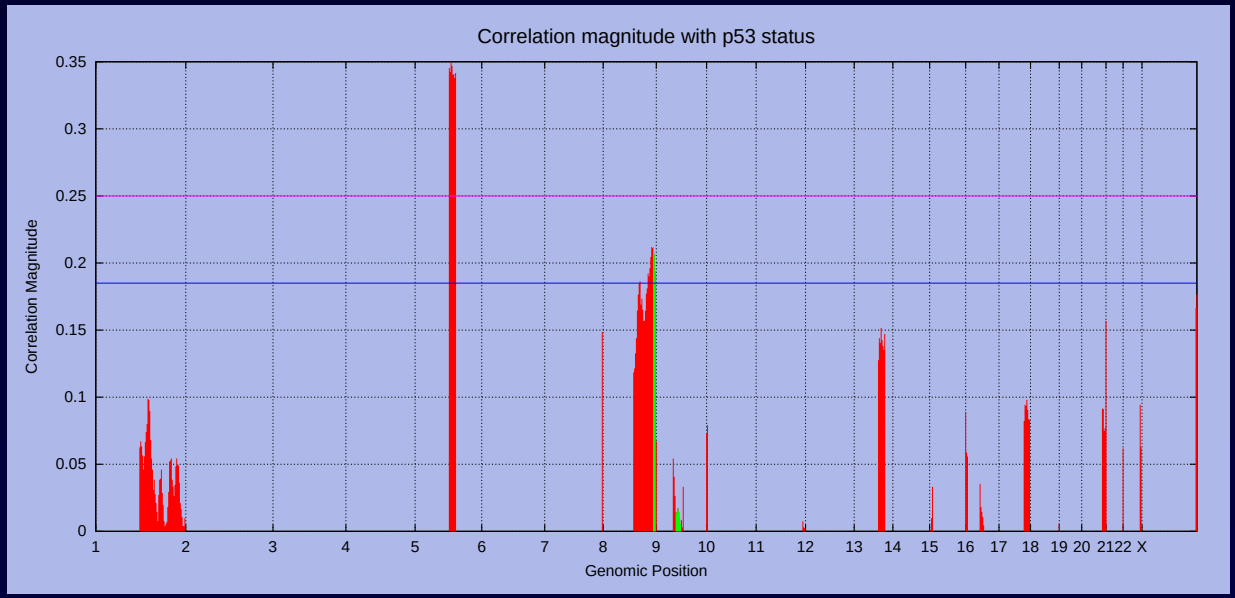
Correlations between p53 mutational status and genomic copy number suggest pathway linkage

Many possibilities to explain the correlation

Speculation:

- 8q24
  - C-myc transactivates p53
  - Perhaps there is a specific benefit to loss of p53 the 8q24 amplified cases
- 5q13-21
  - Is there some gene here that interacts with the p53 pathway?

Deletion at 5q11-31 and amplification at 8q24 are correlated with mutant p53



UCSF cancer center

Summary so far…

Modern measurement technologies yield data sets with many more measurements than samples

We can address the statistical concerns with such data

- By using permutation-based statistics
- By employing pattern classification methods coupled with appropriate model testing (blind testing or cross-validation)
- By *reducing* the number of variables *based on orthogonal criteria*

How can we address aspects of discovery biology with computational modeling?

UCSF cancer center

Example: NCI60 Data
Collaborators: Joe Gray, John Weinstein

RNA

60 tumor-derived cell lines

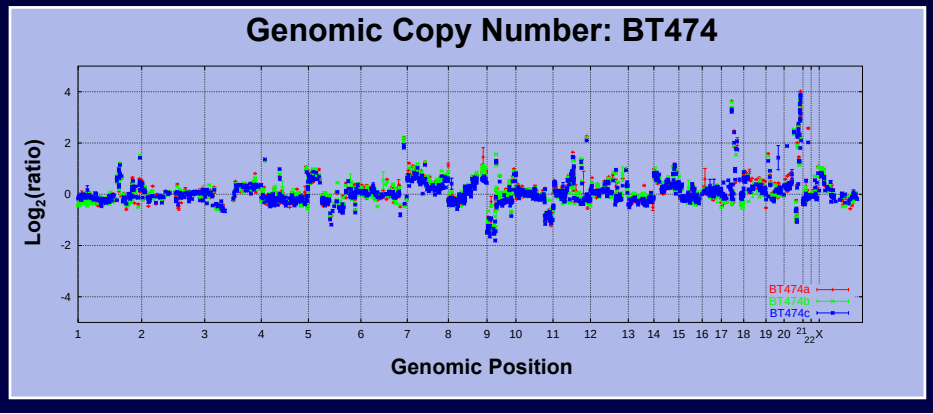
G_1 Gene expression levels in a single condition, ~6000 cDNAs

G_n (Ross et al. 2000)

DNA

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

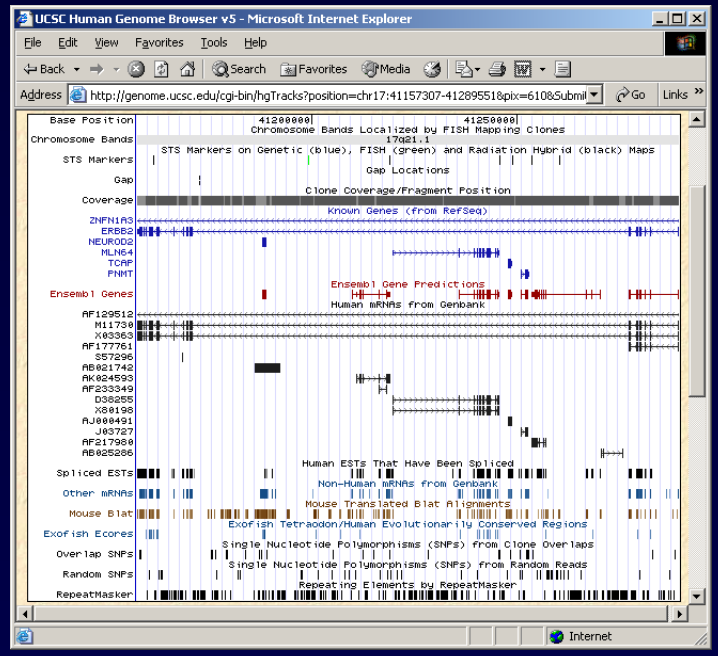
L_n



Gene Annotations

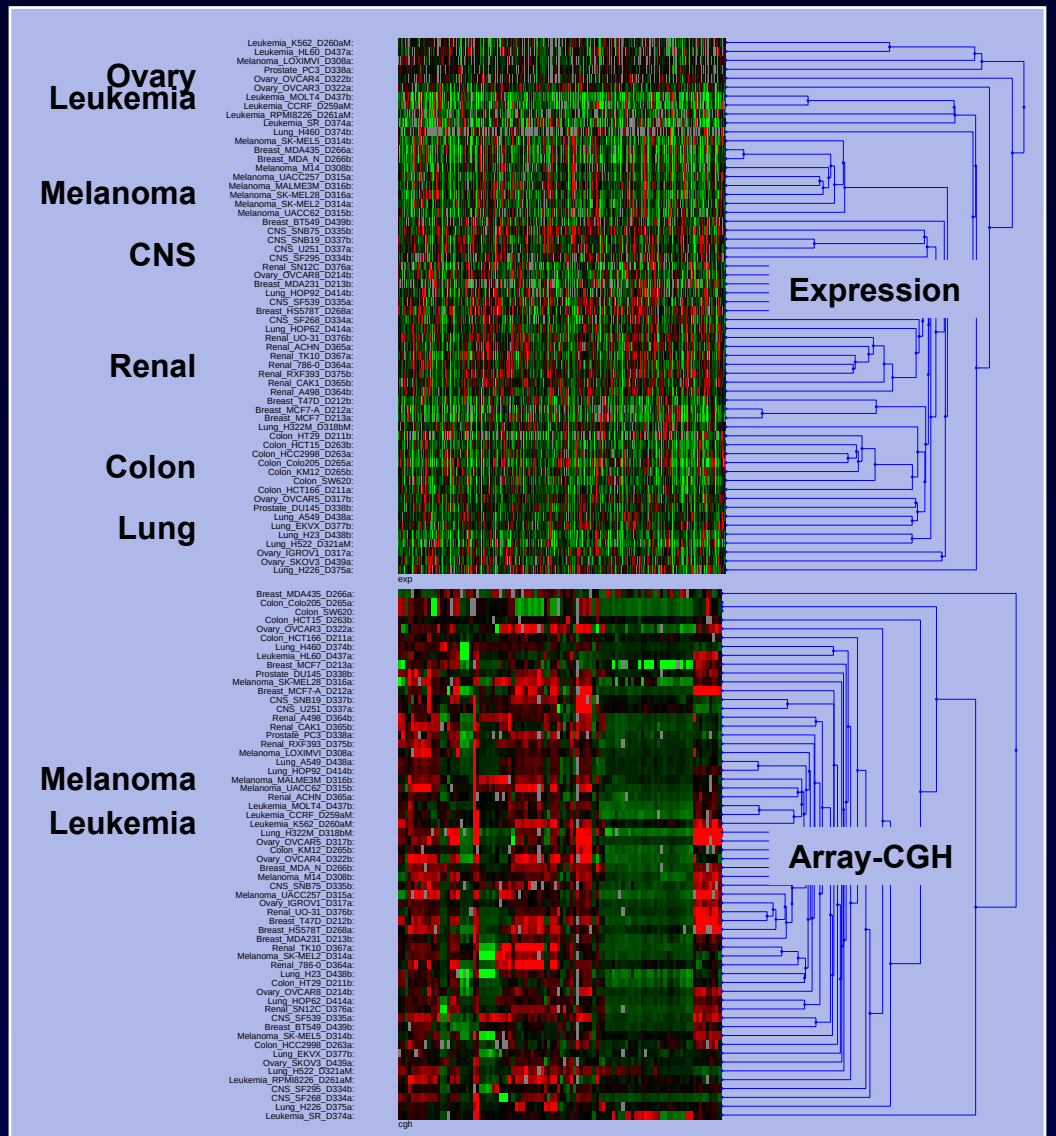
ERBB2:
EC Number: 2.7.1.112
oncogenesis
cell proliferation
Neu/ErbB-2 receptor
protein phosphorylation
protein dephosphorylation
cell growth and maintenance
receptor signaling tyrosine kinase

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 on a genome-wide scale?

We expect to see a gene dosage effect.

How do we expose it?

UCSF cancer center

We relate different data types via identifiers and annotations

CGH data

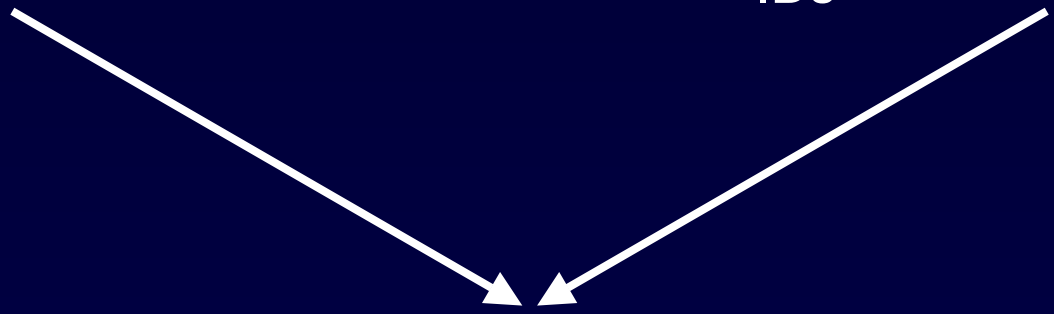
- BAC clones
- STS IDs

- Derive genomic mapping based on sequences from STS IDs

Expression data

- Affy probes
- Genbank IDs, Unigene IDs

- Derive genomic mapping based on sequences from Genbank IDs

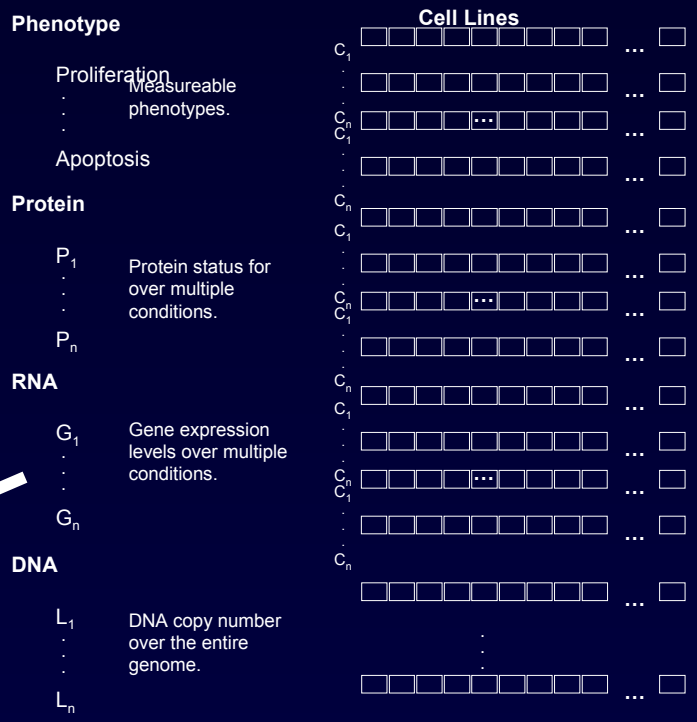


From these mappings, we can relate DNA copy number data with mRNA expression data

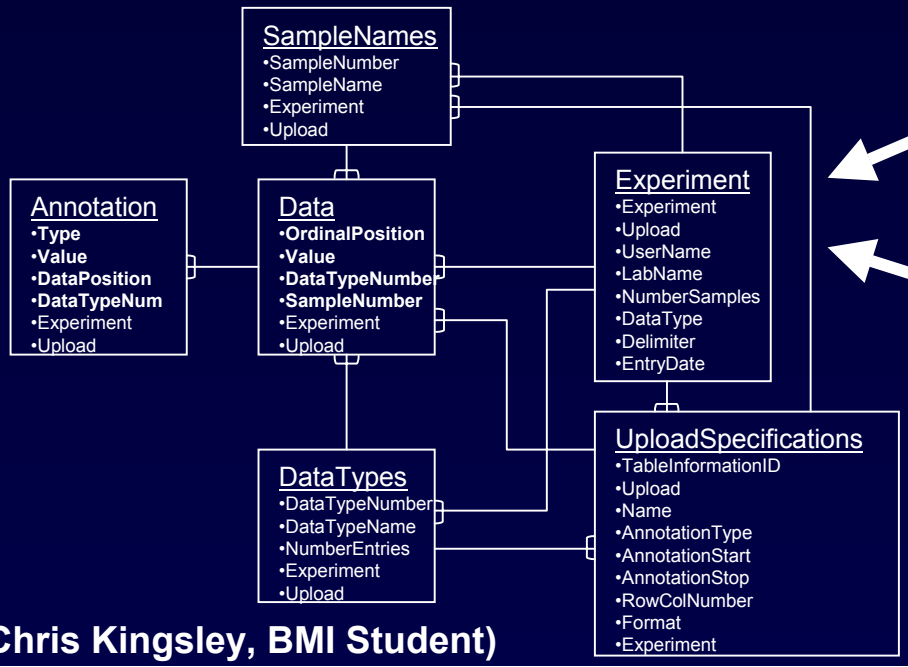
We can accomplish this using a general database back-end coupled with rich operations for subsetting

If we have data of multiple types, of large scope, we can use annotation information to help answer biological questions.

It is possible to do so with a very small number of primitive operations and with a simple but general interface and database architecture.

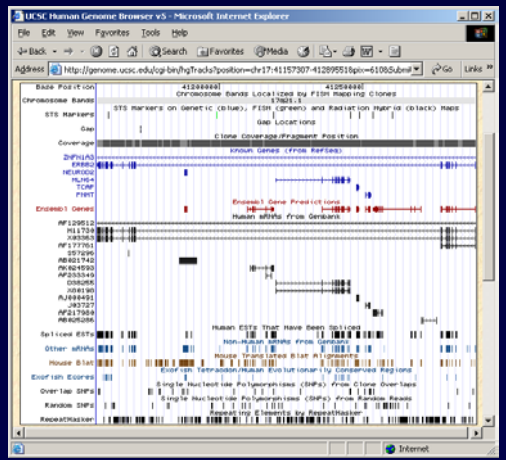


Experimental Data →



(Chris Kingsley, BMI Student)

Genomic Mapping + Context →



UCSF cancer center

Relating the data via genomic mapping allows us to directly compare expression and copy number

Map both cDNAs and BACs to genomic sequence

- Easy for 1 or 2 sequences
- Hard for several thousand (human genome is 3 gigabases)

We can look at the direct effect by binning the data

- 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?

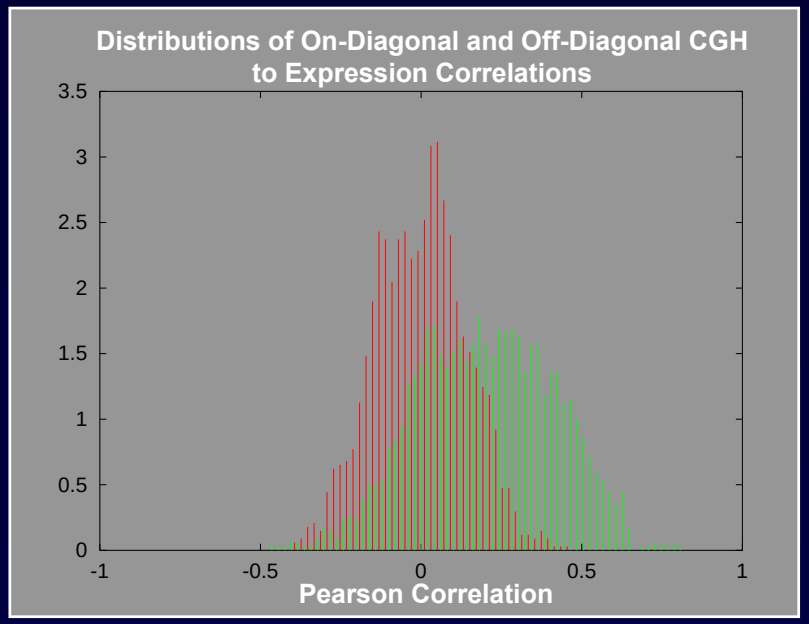
We can do statistics on the populations of pairwise correlations

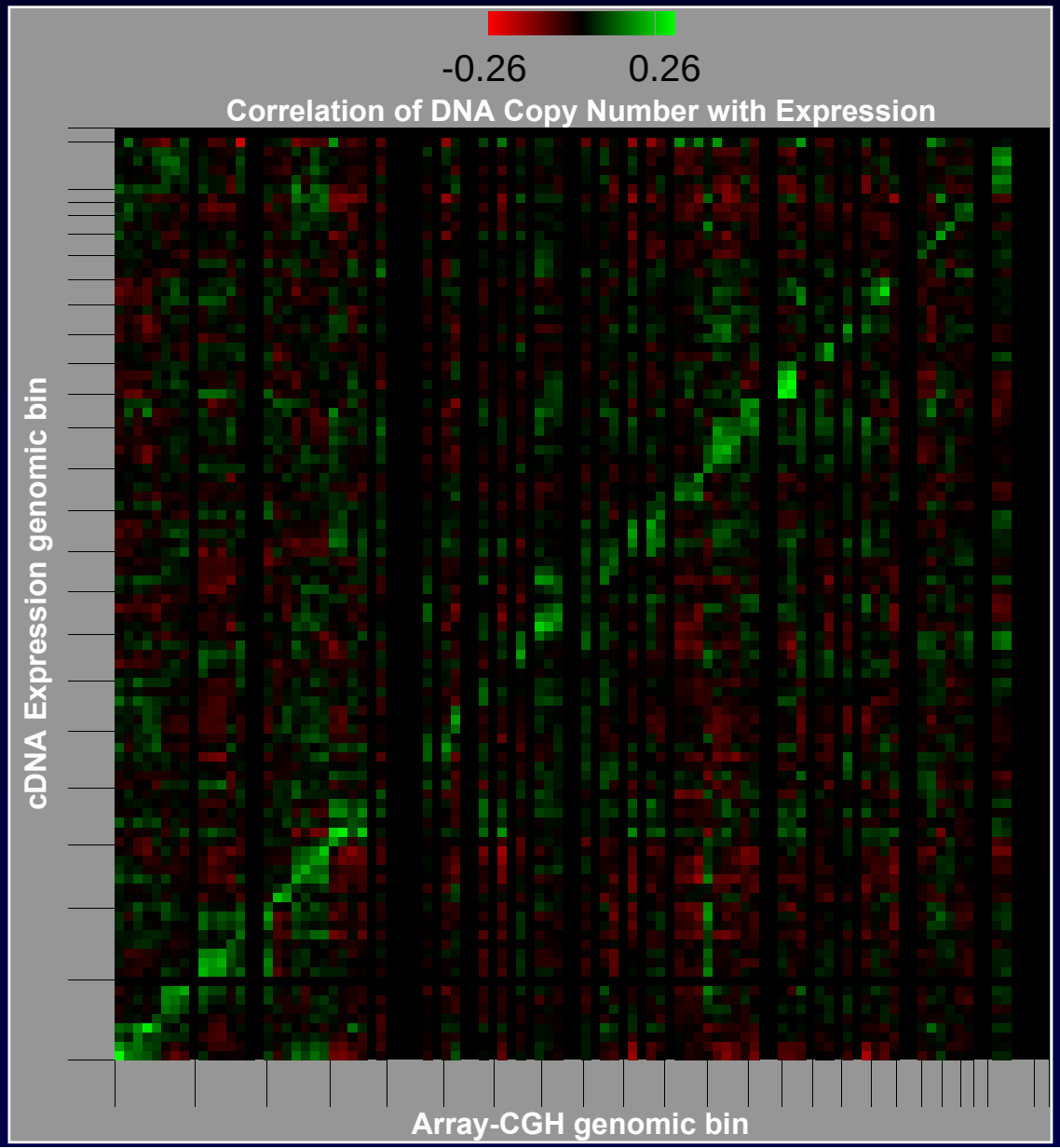
- Consider the set of gene/locus pairs that map within 1 Mb of one another
- Consider the set of gene/locus pairs that map greater than 50 Mb apart
- Are the correlations from (1) higher than from (2)?

UCSF cancer center

NCI60 data: 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

Biological Pathways in Data Analysis: The receptor tyrosine kinase signaling pathways

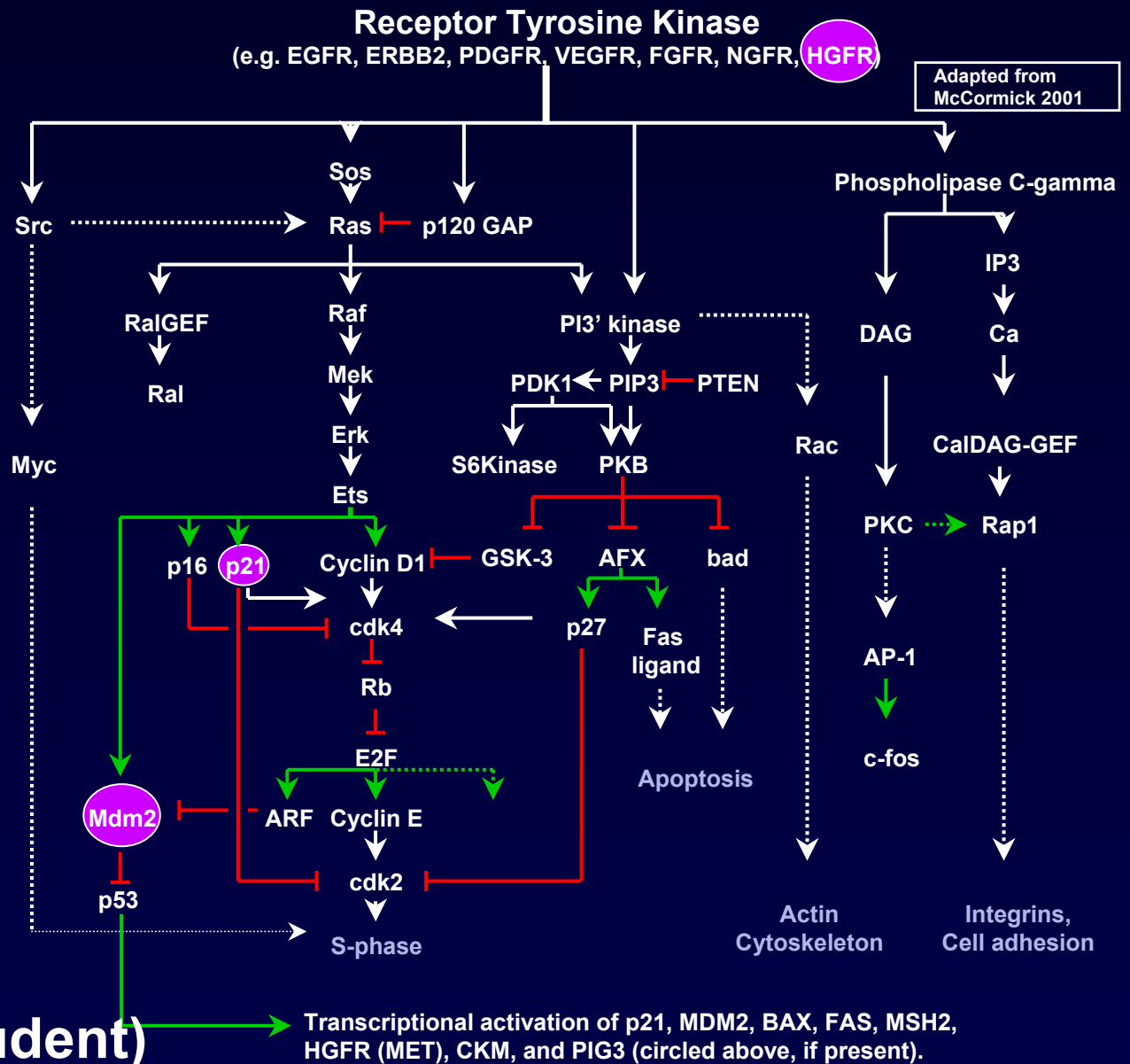
RTK signaling:

- Cell cycle control
- Apoptosis
- Cell adhesion

Therapeutics:

- Herceptin
- PI3-kinase inhibitors
- Onyx015

Pathway derived through hard molecular biology and biochemistry



(Barbara Novak, BMI Student)

UCSF cancer center

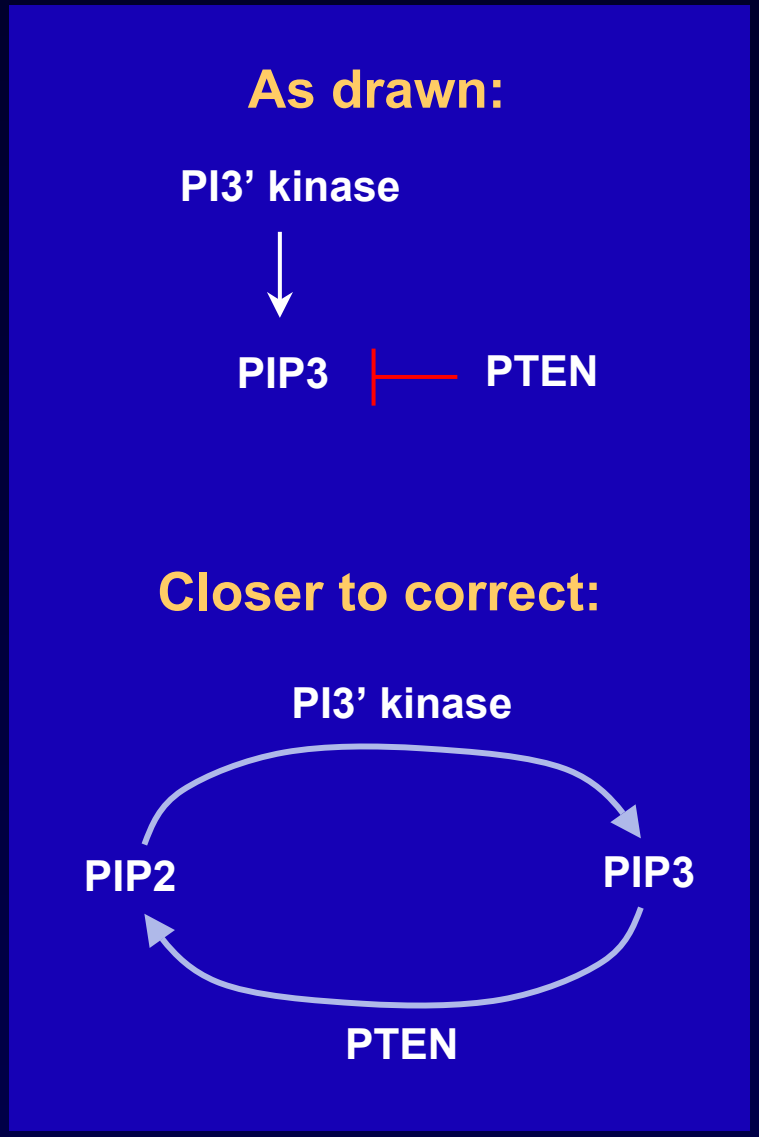
Pathways in human cell biology are complex and variably understood

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

How can we exploit and augment complex biological pathways?

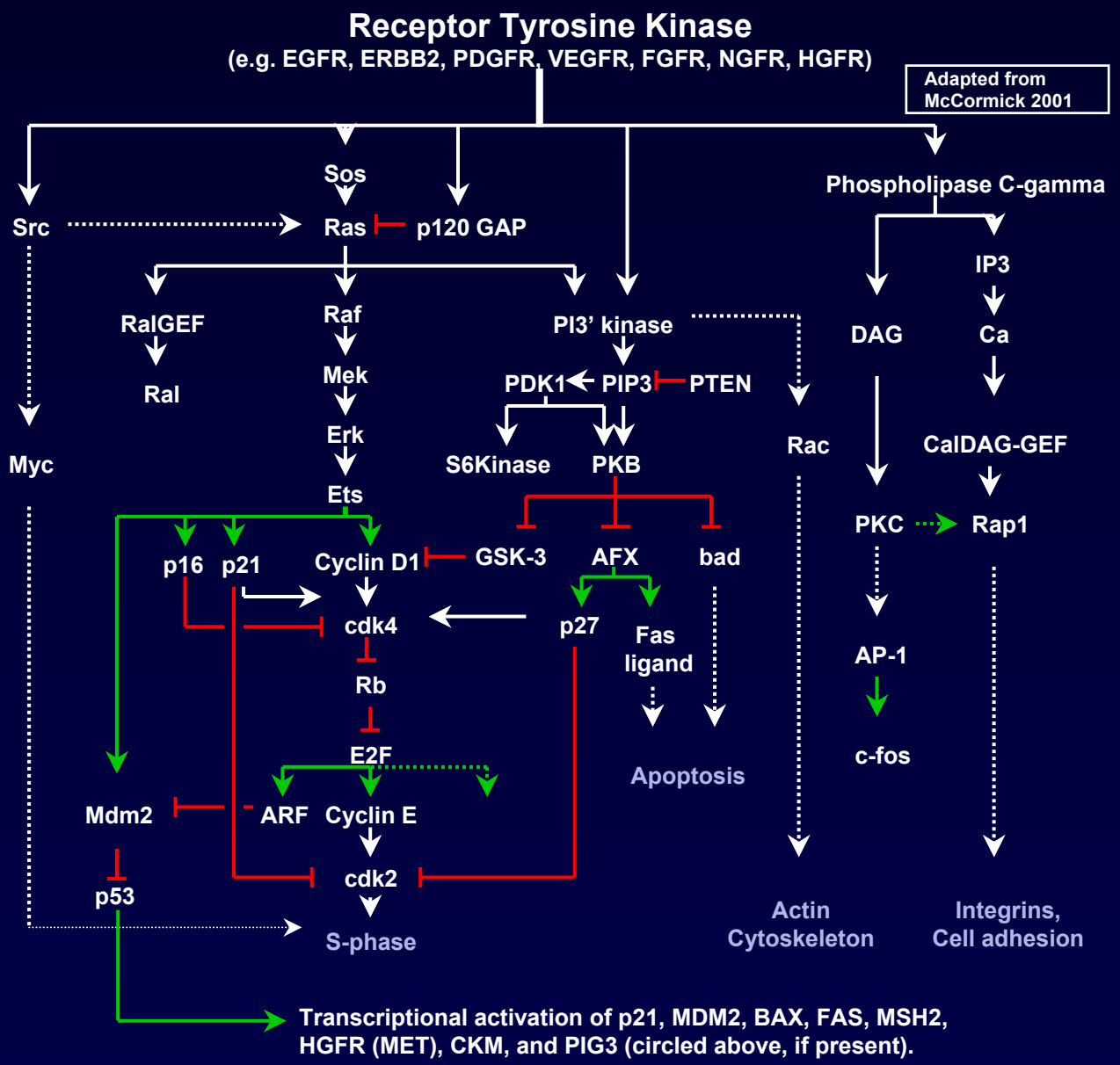
Are the genome copy number patterns of genes that impinge on S-phase checkpoint control quantitatively related?

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).



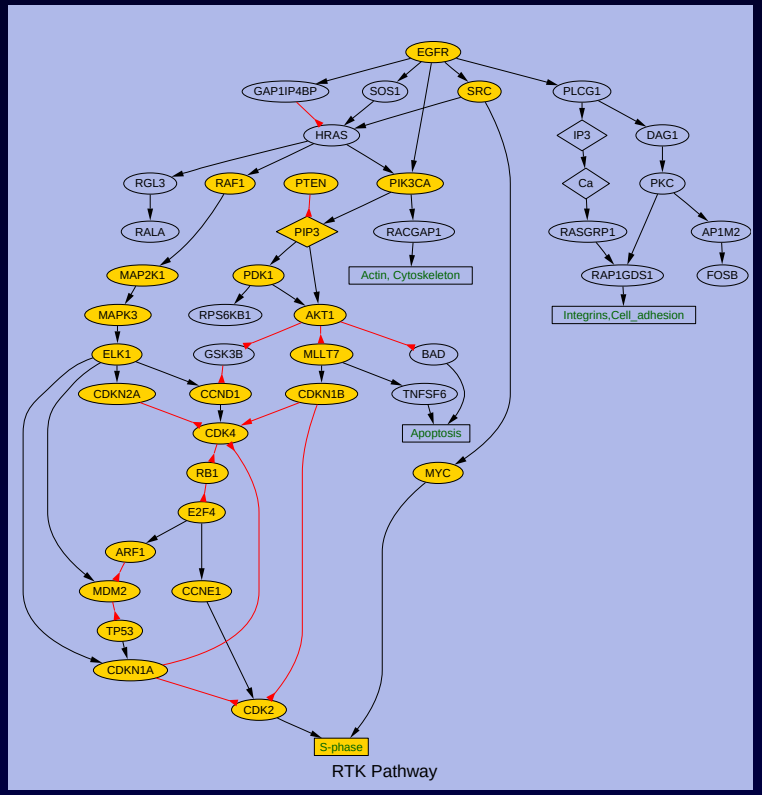
UCSF cancer center

We can accomplish this by explicit representation of pathway structure

Informal representations of biochemical pathways must be formalized

Mappings from experimental data space to pathway space are required

Direct questions involving pathway arms and gene product sets are then easily asked



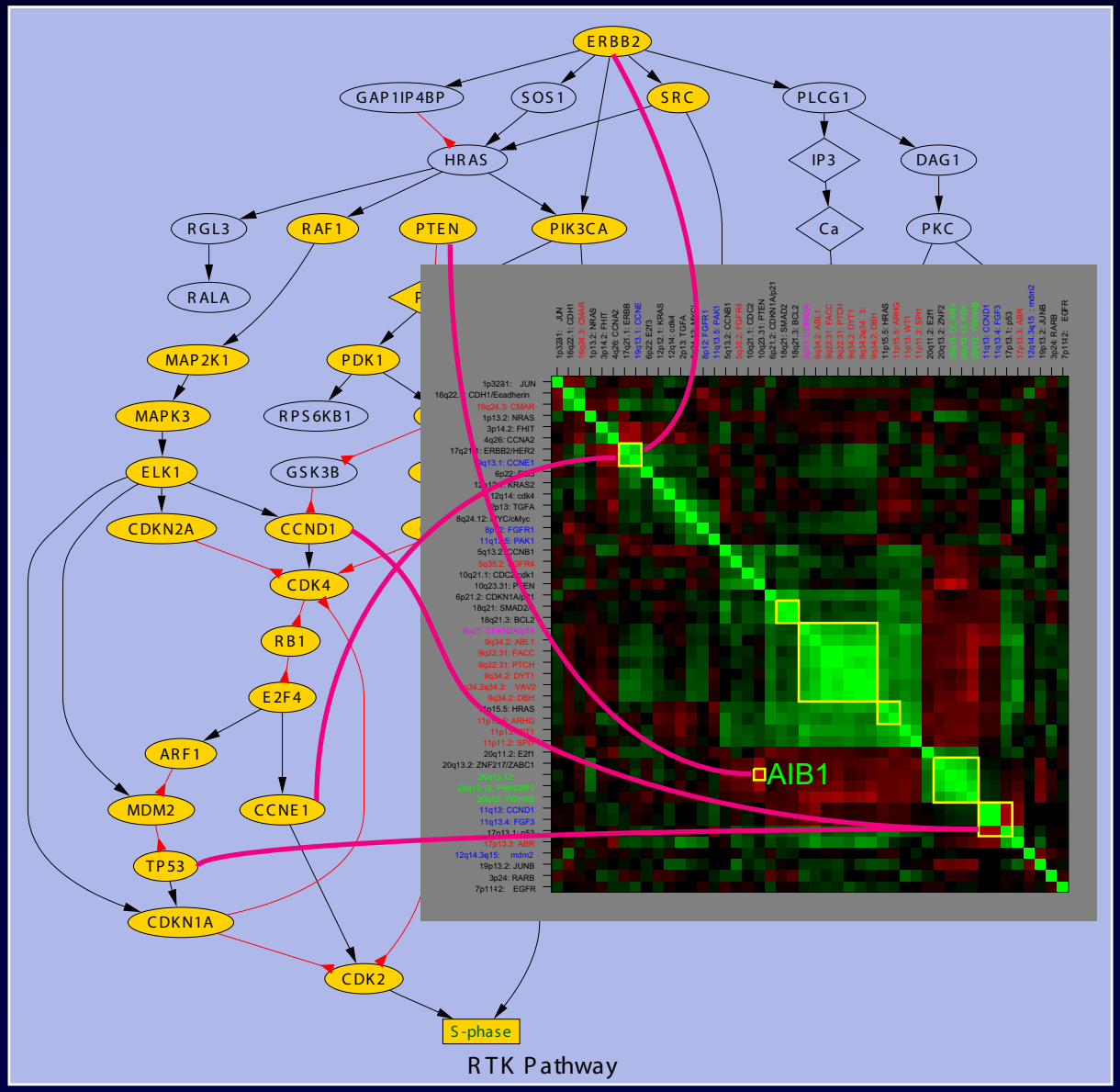
Pathway Knowledge

XML representation

```
<pathway>
  <att name="label" value="RTK Pathway"/>
  <!--ELEMENTS and ASSEMBLIES-->
  <element id="e2" type="gene product">
    <att name="name" value="SRC"/>
    <att name="locusid" value="6714"/>
  </element>
  <assembly id="a2" type="alias“ objects="e2">
  </assembly>
  ...
  <!--NODES-->
  <node id="n1" assembly="a1"></node>
  ...
  <!--EDGES-->
  <edge from="n1"  to="n2"  type="1"  weight="1">
  </edge>
  ...
</pathway>
```

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)

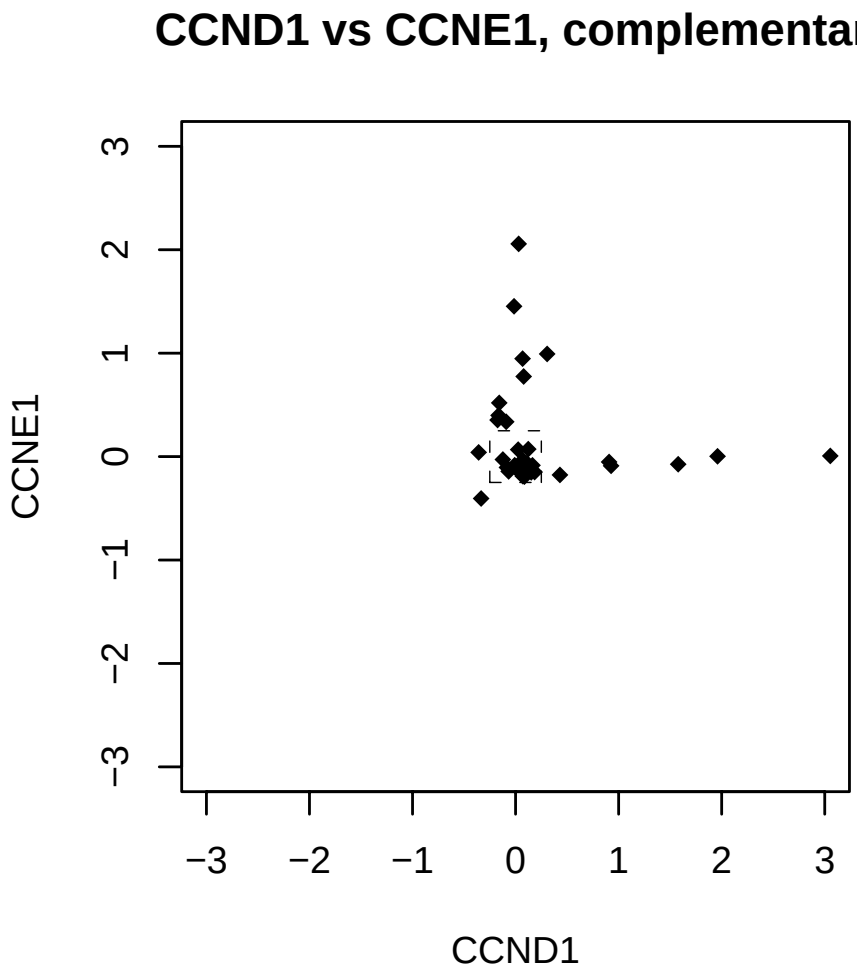
Loss of p53 (17p13.3) and gain of FGF3 (11q13)

Bladder tumor data

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

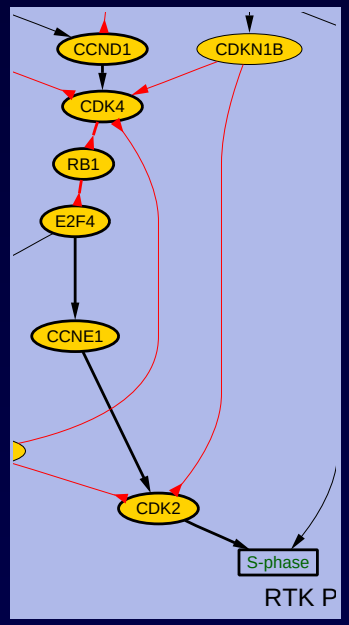
UCSF cancer center

Some gene pairs complement each other



Cyclin D1 and Cyclin E1

- Change in Cyclin D1 precludes change in Cyclin E1 and vice-versa
- Both genes have the same overall effect on S-phase checkpoint control



Bladder tumor data

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

UCSF cancer center

The intersection of fields makes all of this possible

Fields that collide

- Basic biology
- Clinical medicine
- Measurement technology
- Biochemistry
- Chemistry
- Biophysics
- Mathematics
- Statistics
- Computer science

What can we do?

- Build predictive models of biologically or clinically important phenomena
- Use the models to make hypotheses about things that might be true

What genes form which pathways?

Which genomic loci are causally involved in cancer progression?

Which molecules are active?

UCSF cancer center

Acknowledgements

Experimental collaborators

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

- John Weinstein (NCI)
- Gordon Mills (MD Anderson)

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