CaBIG Kickoff Meeting:
UCSF Cancer Center
Integrative Cancer Research Workspace

Ajay N. Jain, PhD

ajain@cc.ucsf.edu

http://jainlab.ucsf.edu

Overview

UCSF Comprehensive Cancer Center

- Over 300 member faculty
- 3 NCI SPORE Grants
  - Breast (Gray)
  - Prostate (Shuman)
  - Brain (Berger)
- Mouse Models of Human Cancer Consortium
  - Albertson
  - Ballmain
  - Shannon
- Broad and deep portfolio of NCI funded basic and translational research

We have a growing number of investigators studying cancer from a *systems biology* perspective

Consequently, we have both *contributions* and *needs* relevant to the Grid

UCSF-CC and CaBIG

Systems biology in multiple organ systems

- Breast
- Ovary
- Pancreas
- Prostate

Heterogeneous data

- Array-based CGH
- Array-based expression
- ESP: a novel technology for profiling the cancer genome
- Phenotype
- Protein

Dry-side investigators have contributions

- Tools for analyzing complex data sets
- Tools for data generation (e.g. UCSF Spot)
- Tools for data management (e.g. array data management)
- Willing guinea-pigs for informatics infrastructure

Wet-side investigators have contributions

- Data of multiple types
- Technology (e.g. array-CGH, ESP)
- Willing guinea-pigs for deployed tool

UCSF-CC and CaBIG

Systems biology in multiple organ systems

- Breast
- Ovary
- Pancreas
- Prostate

Heterogeneous data

- Array-based CGH
- Array-based expression
- ESP: a novel technology for profiling the cancer genome
- Phenotype
- Protein

Dry-side investigators have needs as well

- Object name space resolution (genes, gene products…)
- Stable annotation-space with linkable APIs
- Community in which to share and collaborate

Wet-side investigators have sophisticated needs

- Data sharing
- Data visualization
- Integration with annotation-space (pathways, Gene Ontology, etc…)

Magellan: Quantitative Analysis of Heterogeneous Data Sets

Synergistic Combination of Data and Annotations

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.

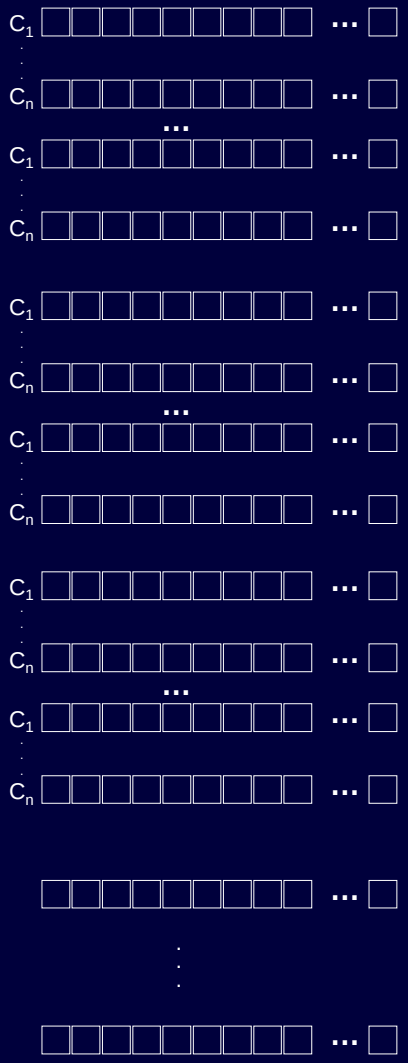
RNA

G_1 … G_n Gene expression levels over multiple conditions.

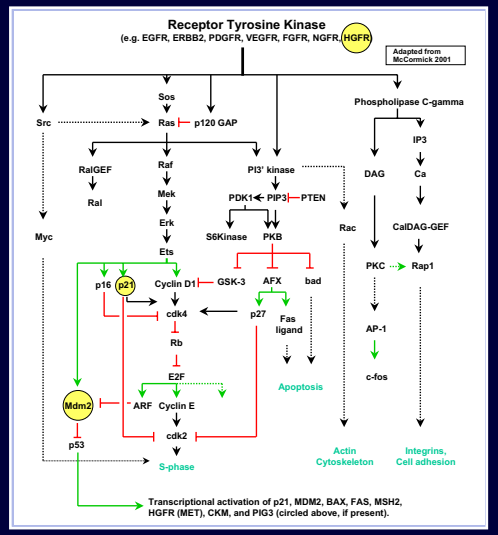
DNA

L_1 … L_n DNA copy number over the entire genome.

Cell Lines



Pathway Structure →



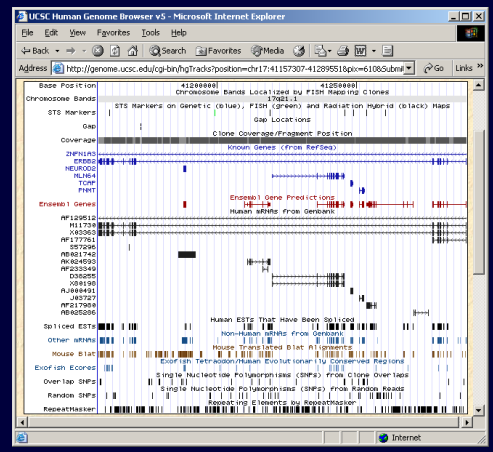
ERBB2:

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 Breast SPORE
Breast tumor-derived cell lines

RNA

~40 tumor-derived cell lines

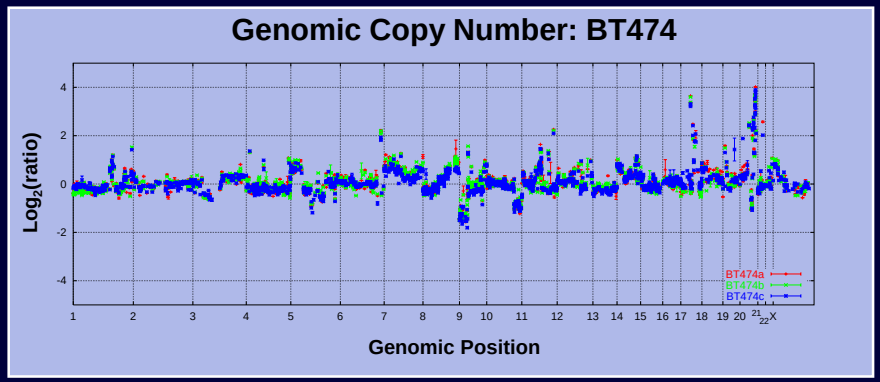
G_1 Gene expression levels in a single condition, ~6000 cDNAs

G_n (Ross et al. 2000)

DNA

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

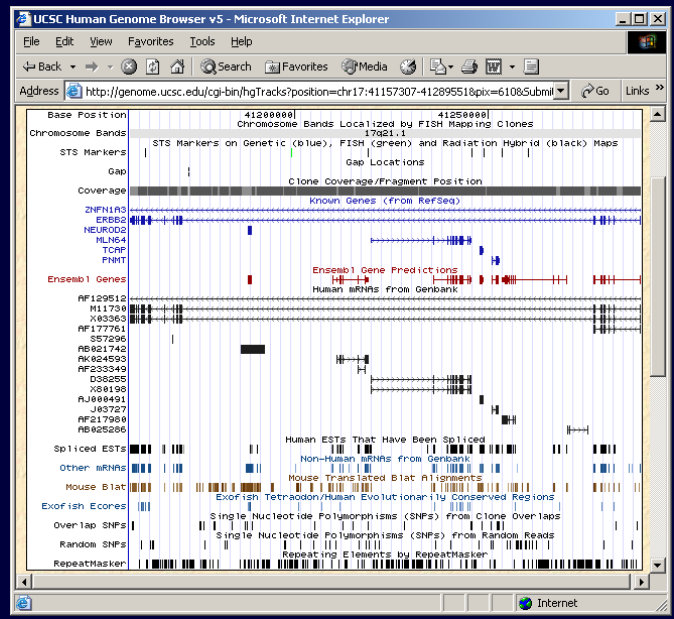
L_n



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



We relate different data types via identifiers and annotations

CGH data

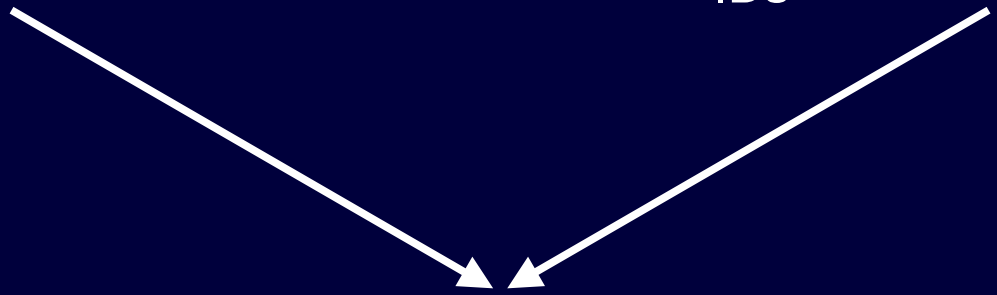
- BAC clones
- STS IDs

- 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 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.

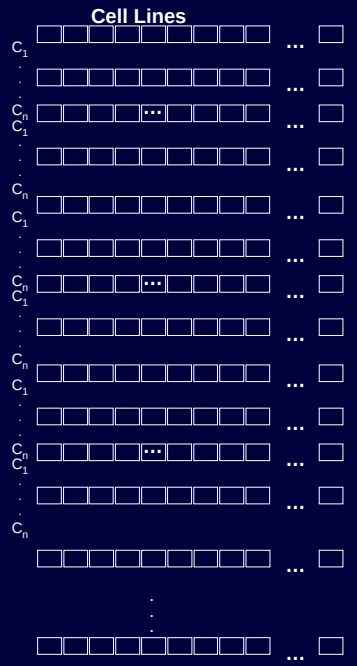
RNA

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

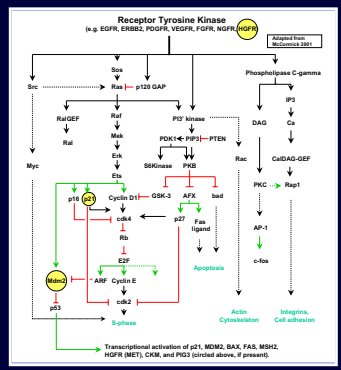
DNA

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

Cell Lines



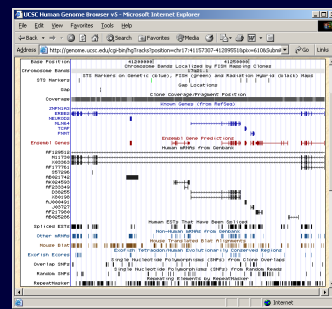
Pathway Structure



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 →



Magellan: Chris Kingsley

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

Map both Affy probes 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

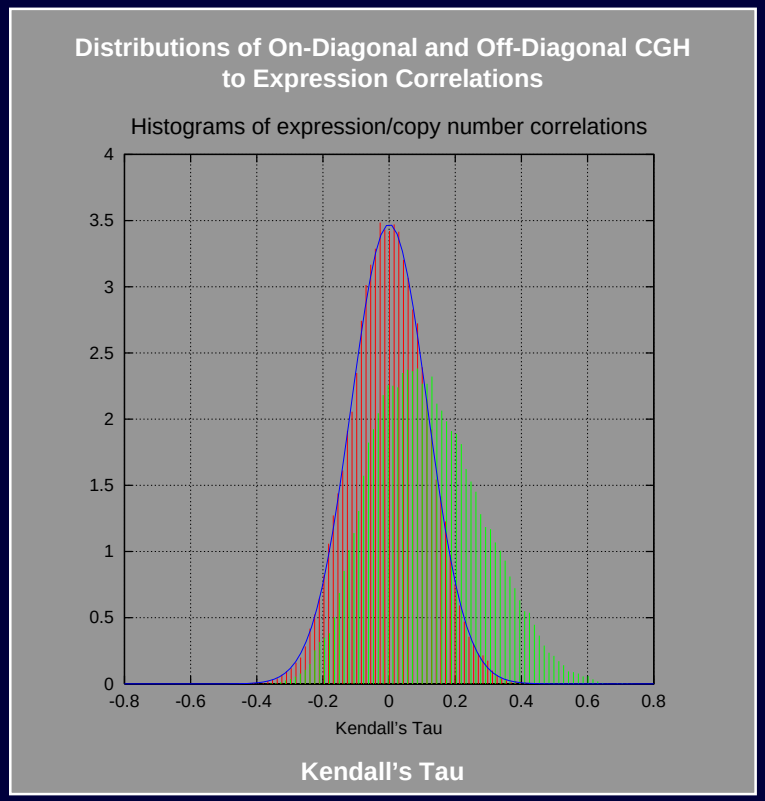
- Consider the set of genes that map to a particular genomic position
- Consider the set of BACs that map to the same place
- Are those genes' expression correlated with copy number at those loci?

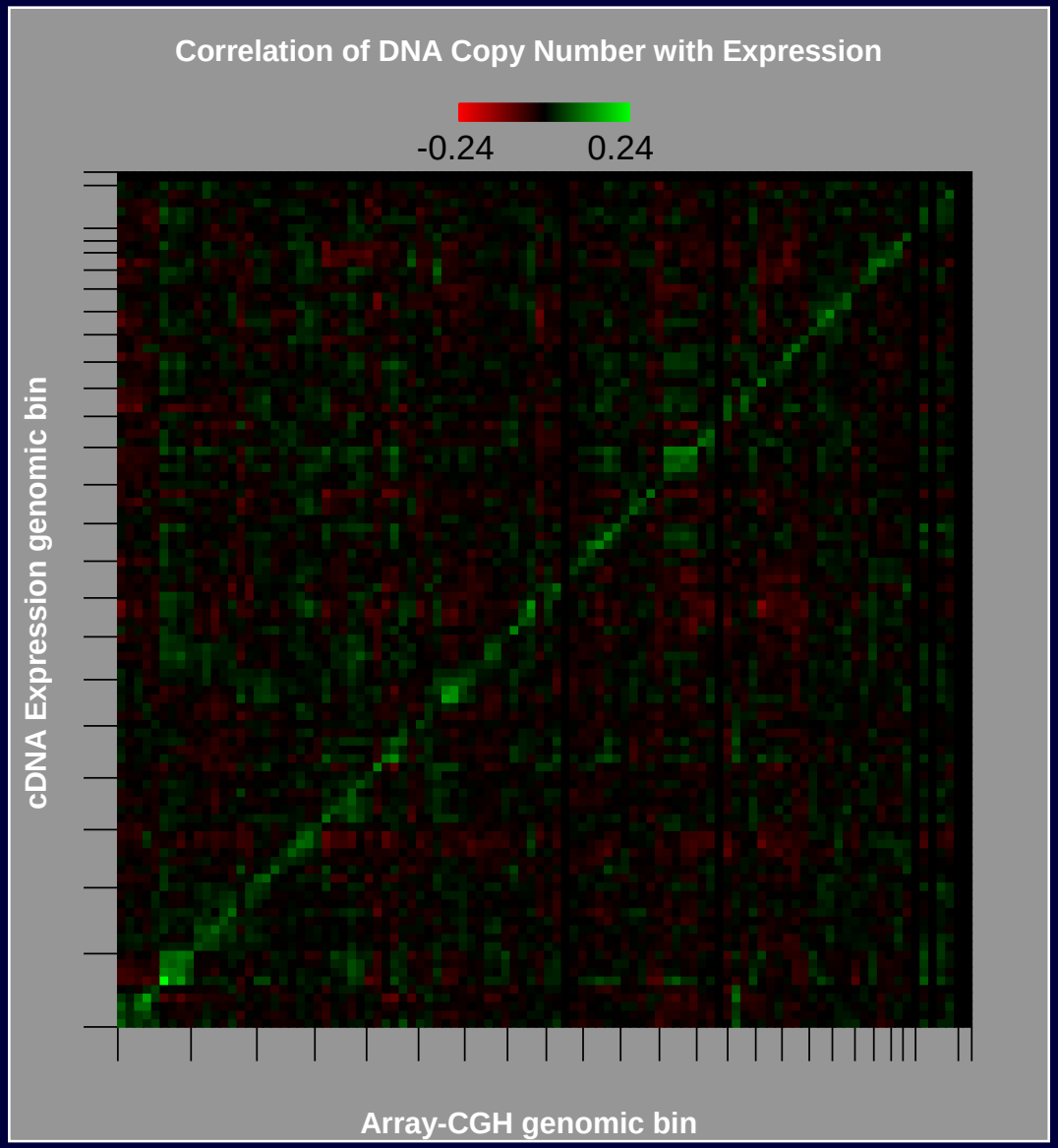
We can do statistics on the populations of pairwise correlations

- 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)?

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





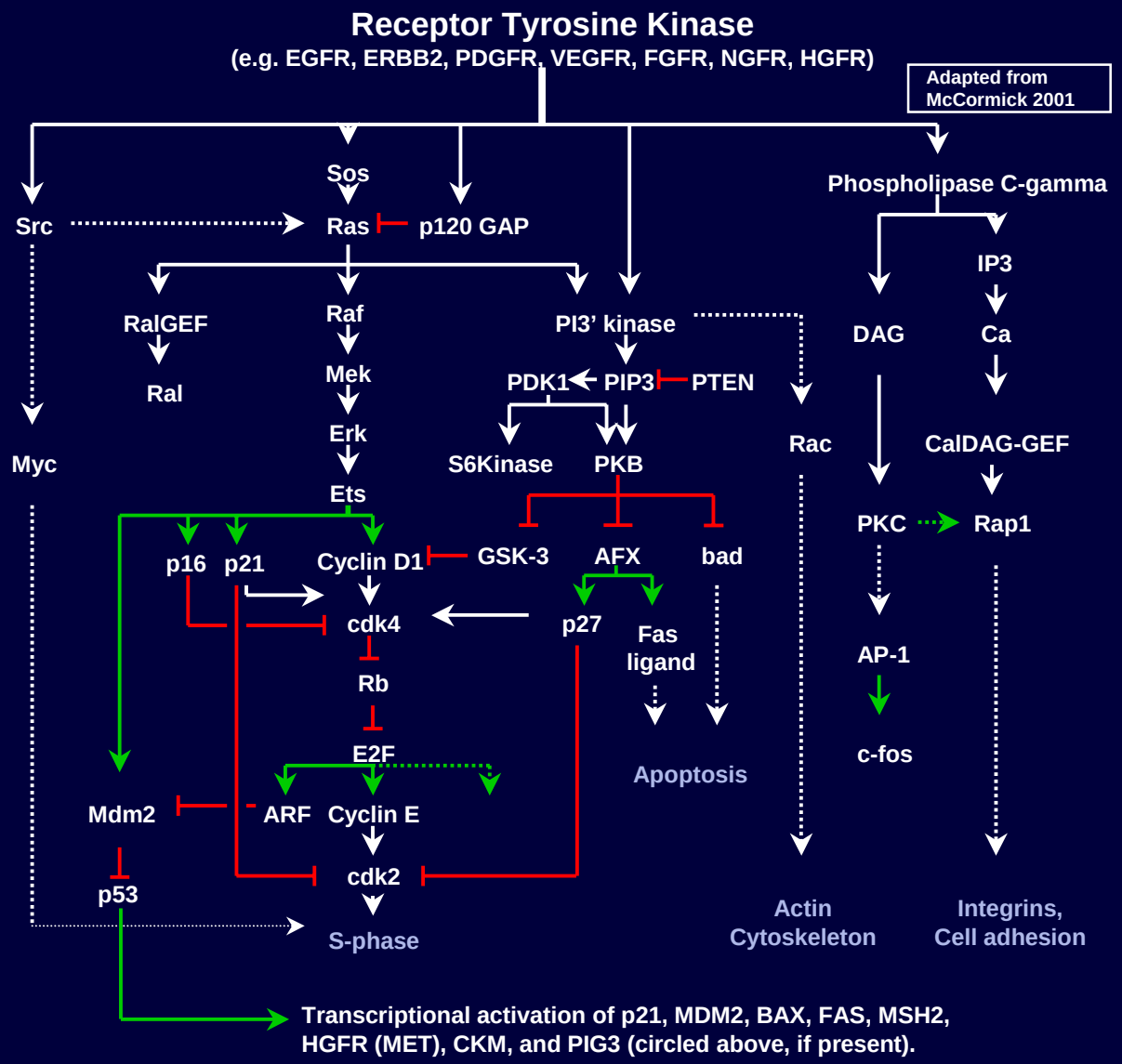
QPACA: Quantitative Pathway Analysis in Cancer

Synergistic Combination of Data and Pathway Information

Pathways may help us answer more complicated questions

Is sensitivity to Herceptin mediated by downstream abnormalities?

Can we use DNA copy number, expression, and phenotypic data to answer this?



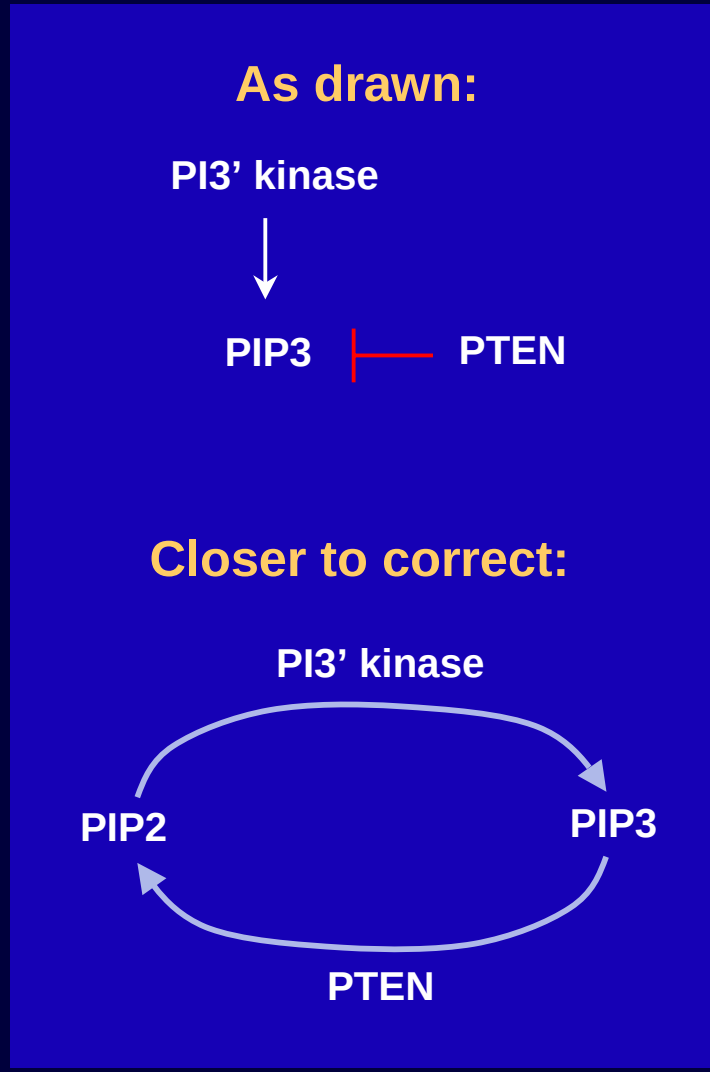
Pathways in human cell biology are complex and variably understood

Problem 1: symbols are overloaded

Problem 2: shorthand is used

Problem 3: knowledge is incomplete

But we must still try to represent this information



We can accomplish this by explicit representation of pathway structure

Informal representations of biochemical pathways must be formalized

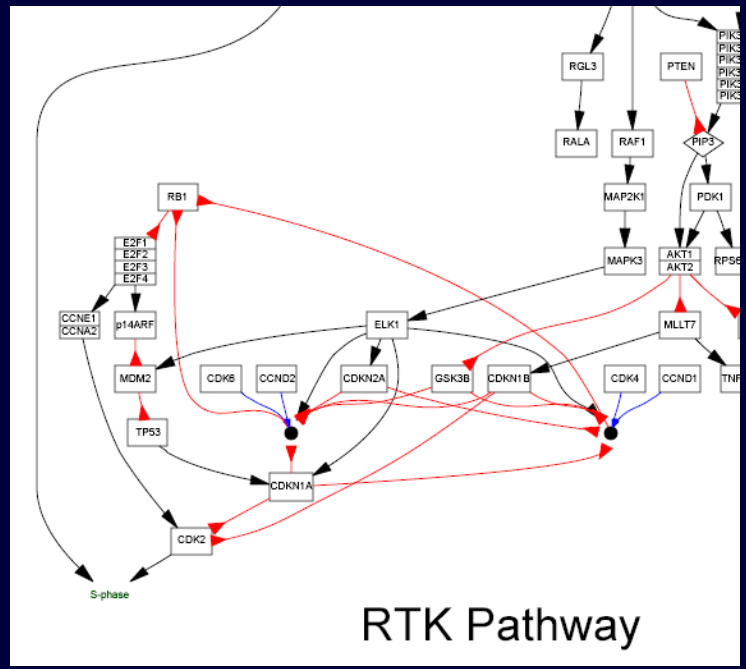
Mappings from experimental data space to pathway space are required

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

Pathway Knowledge

Structured pathway language

```
pathway_name = "RTK Pathway" #element section
pathway_elements {
cdk4[locusid=1019;type=gene_product]
ccnd1[locusid=595;type=gene_product]
rb1[locusid=5925;type=gene_product]
cdk2[locusid=1017;type=gene_product]
akt2[locusid=208;type=gene_product]
akt1[locusid=207;type=gene_product]
ccne1[locusid=898;type=gene_product]
ccna2[locusid=890;type=gene_product]
cdk4_compound[members=cdk4,ccnd1;type=compoun
d]
cdk6_compound[members=cdk6,ccnd2;type=compoun
d] } #pathway sections pathway { elk1 -> mdm2 elk1 -
> cdkn2a elk1 -> cdkn1a elk1 -> cdk4_compound elk1
-> cdk6_compound cdk6_compound -| rb1
cdk4_compound -| rb1 rb1 -| e2f_family e2f_family ->
p14arf e2f_family -> ccne_family ccne_family -> cdk2
p14arf -| mdm2 mdm2 -| tp53 tp53 -> cdkn1a cdkn1a -|
cdk2 cdkn1a -| cdk6_compound cdkn1a -|
cdk4_compound cdkn2a -| cdk6_compound cdkn2a -|
cdk4_compound cdkn1b -| cdk2 cdkn1b -|
cdk6_compound cdkn1b -| cdk4_compound pip3 ->
akt_family akt_family -| gsk3b gsk3b -|
cdk4_compound gsk3b -| cdk6_compound cdk2 -> s-
phase }
```



Pathway language is simple, intuitive, and extensible

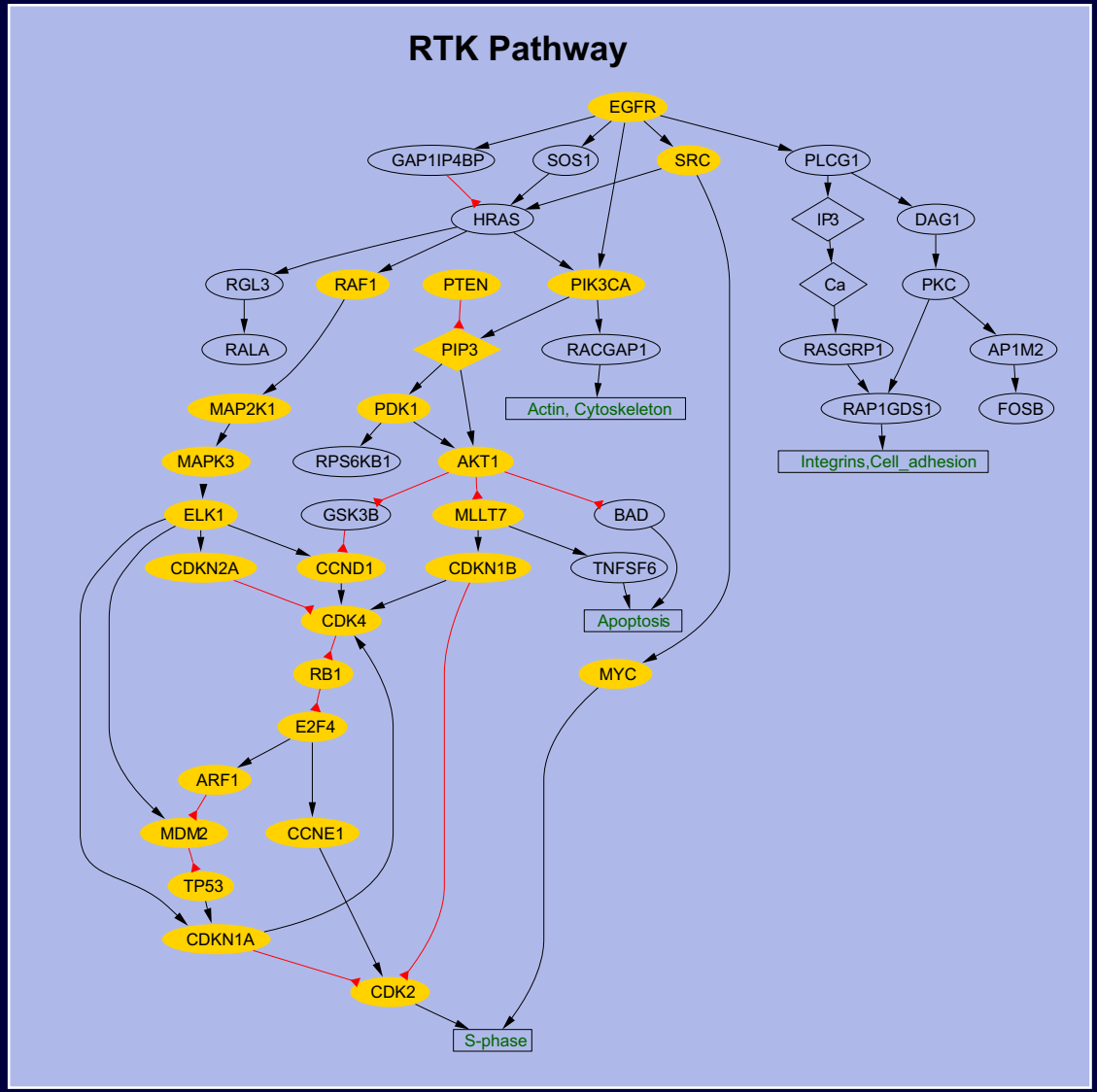
```
pathway_name = "RTK Pathway"
pathway_elements {
    cdk4[locusid=1019;type=gene_product]
    ccnd1[locusid=595;type=gene_product]
    rb1[locusid=5925;type=gene_product]
    cdk2[locusid=1017;type=gene_product]
    akt2[locusid=208;type=gene_product]
    akt1[locusid=207;type=gene_product]
    ccne1[locusid=898;type=gene_product]
    ccna2[locusid=890;type=gene_product]
    gsk3b[locusid=2932;type=gene_product]
    cdkn1b[locusid=1027;type=gene_product]
    cdk6[locusid=1021;type=gene_product]
    ccnd2[locusid=894;type=gene_product]
    cdkn1a[locusid=1026;type=gene_product]
    s-phase[type=process]
    pip3[type=molecule]
    akt_family[members=akt1,akt2;type=alt]
    cdk4_compound[members=cdk4,ccnd1;type=c
    ompound]
    cdk6_compound[members=cdk6,ccnd2;type=c
    ompound]
}
```

```
#pathway sections
pathway {
    elk1 -> mdm2
    elk1 -> cdkn2a
    elk1 -> cdkn1a
    elk1 -> cdk4_compound
    elk1 -> cdk6_compound
    cdk6_compound -| rb1
    cdk4_compound -| rb1
    rb1 -| e2f_family
    e2f_family -> p14arf
    e2f_family -> ccne_family
    ccne_family -> cdk2
    p14arf -| mdm2
    mdm2 -| tp53
    tp53 -> cdkn1a
    pip3 -> akt_family
    akt_family -| gsk3b
    gsk3b -| cdk4_compound
    gsk3b -| cdk6_compound
    cdk2 -> s-phase
}
```

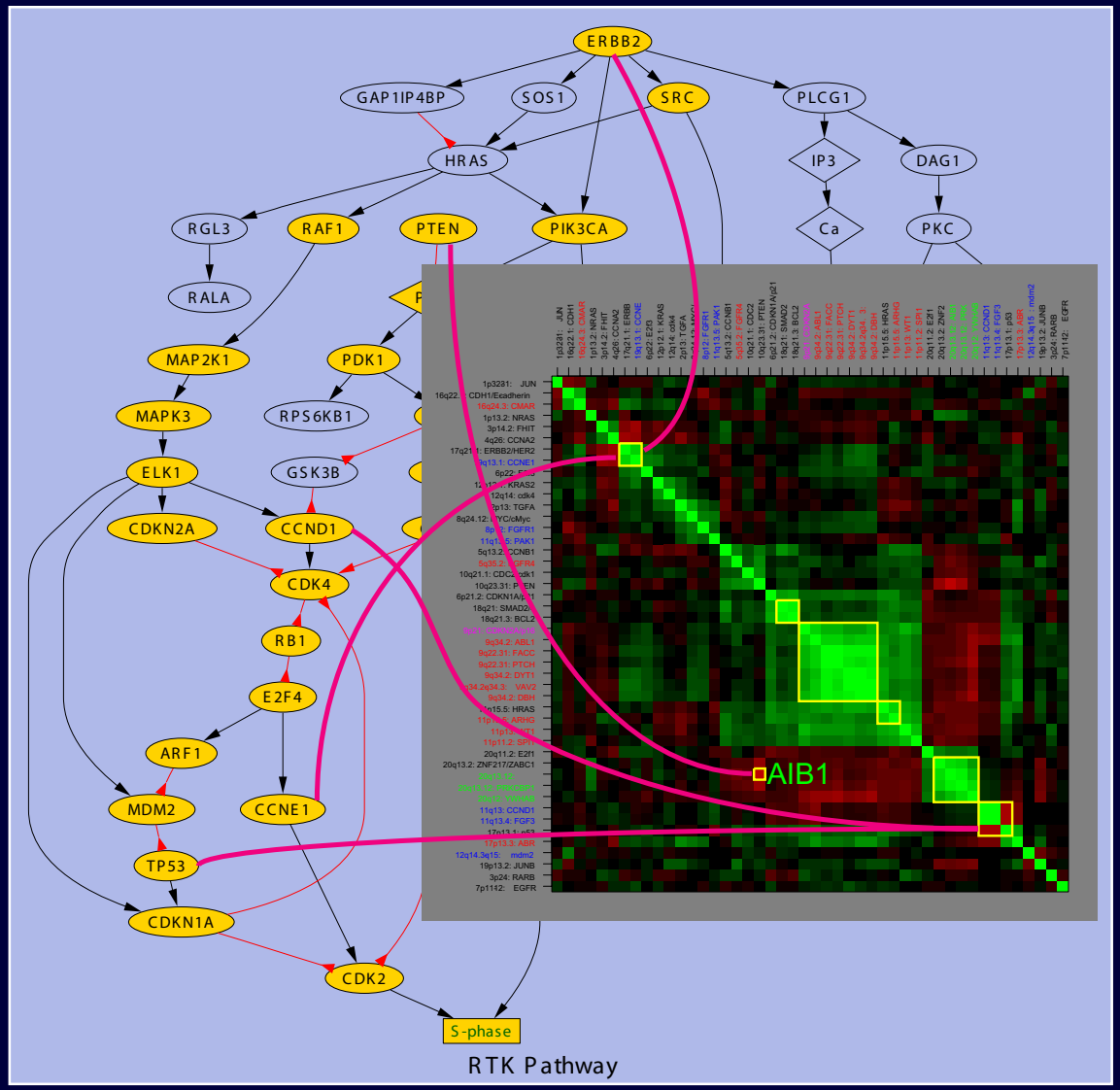
We can use the pathway to restrict our loci for comparison

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.



Example: Bladder tumor data (Waldman Lab)
Doe we see linkage between genomic loci?



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

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

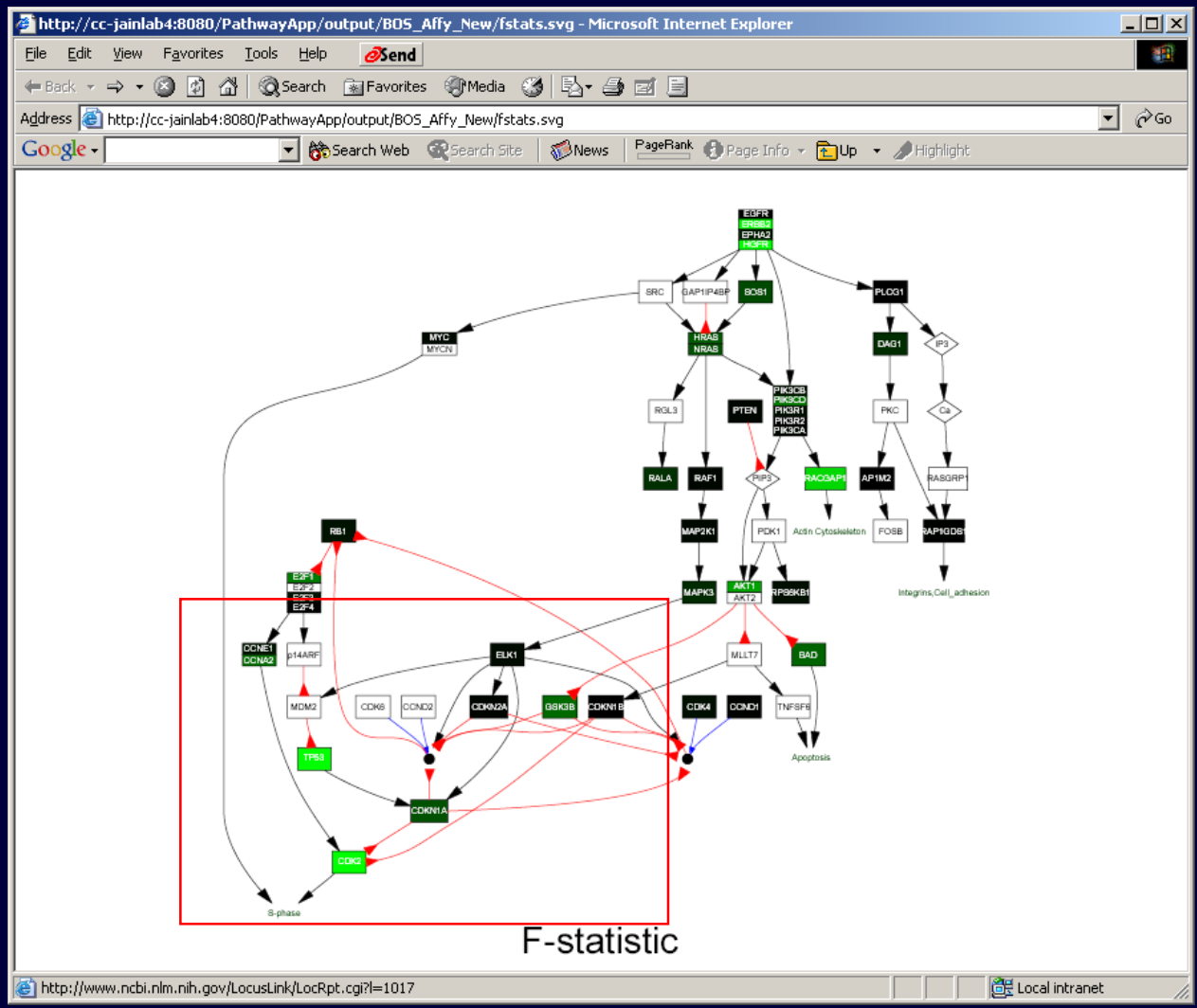
Loss of p53 (17p13.3) and gain of CCND1 (11q13)

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 can mark cell lines based on CGH observations: Amplified ERBB2 (9/40 cell lines)

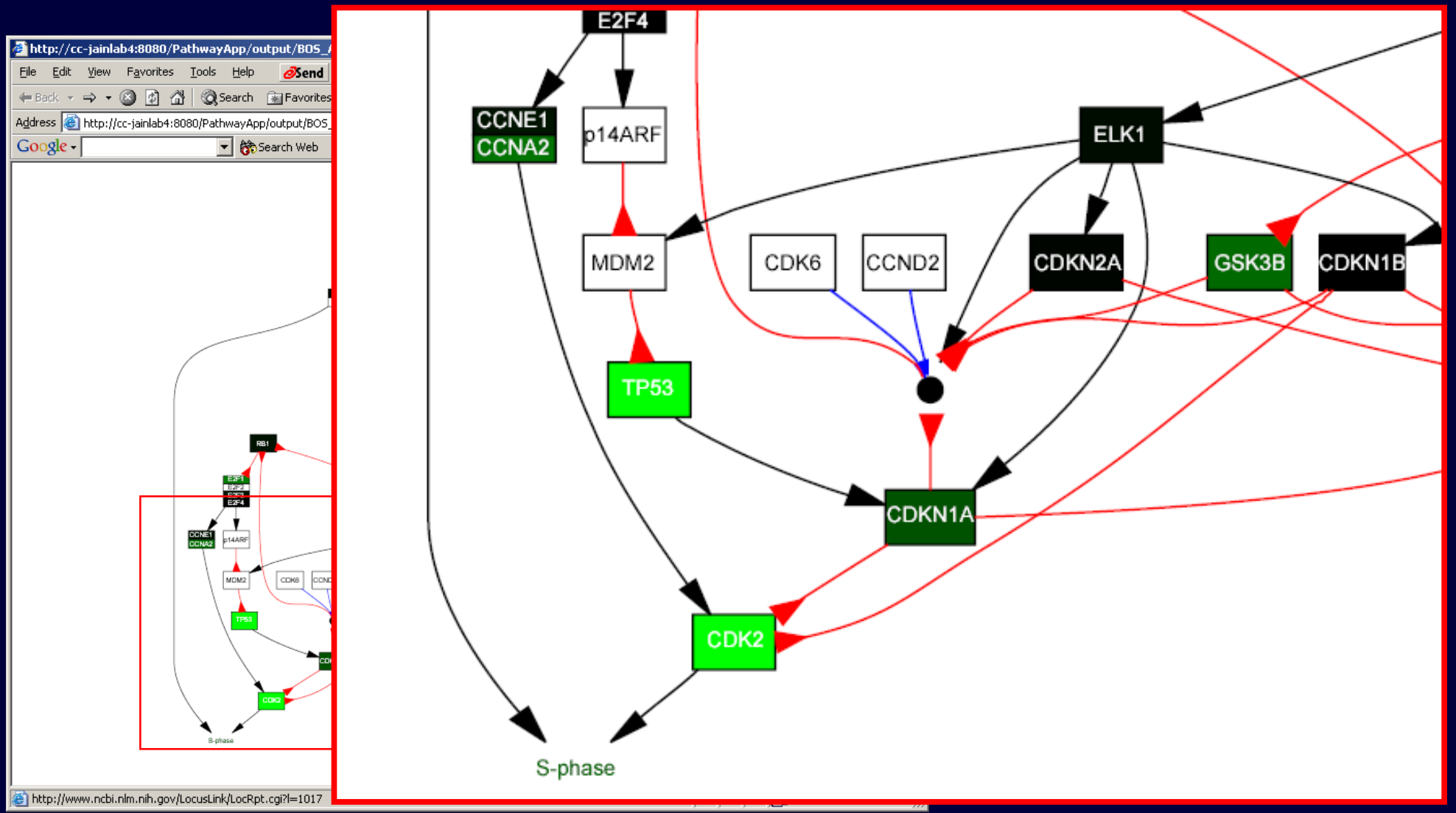


Pathway structure generated from curated representation

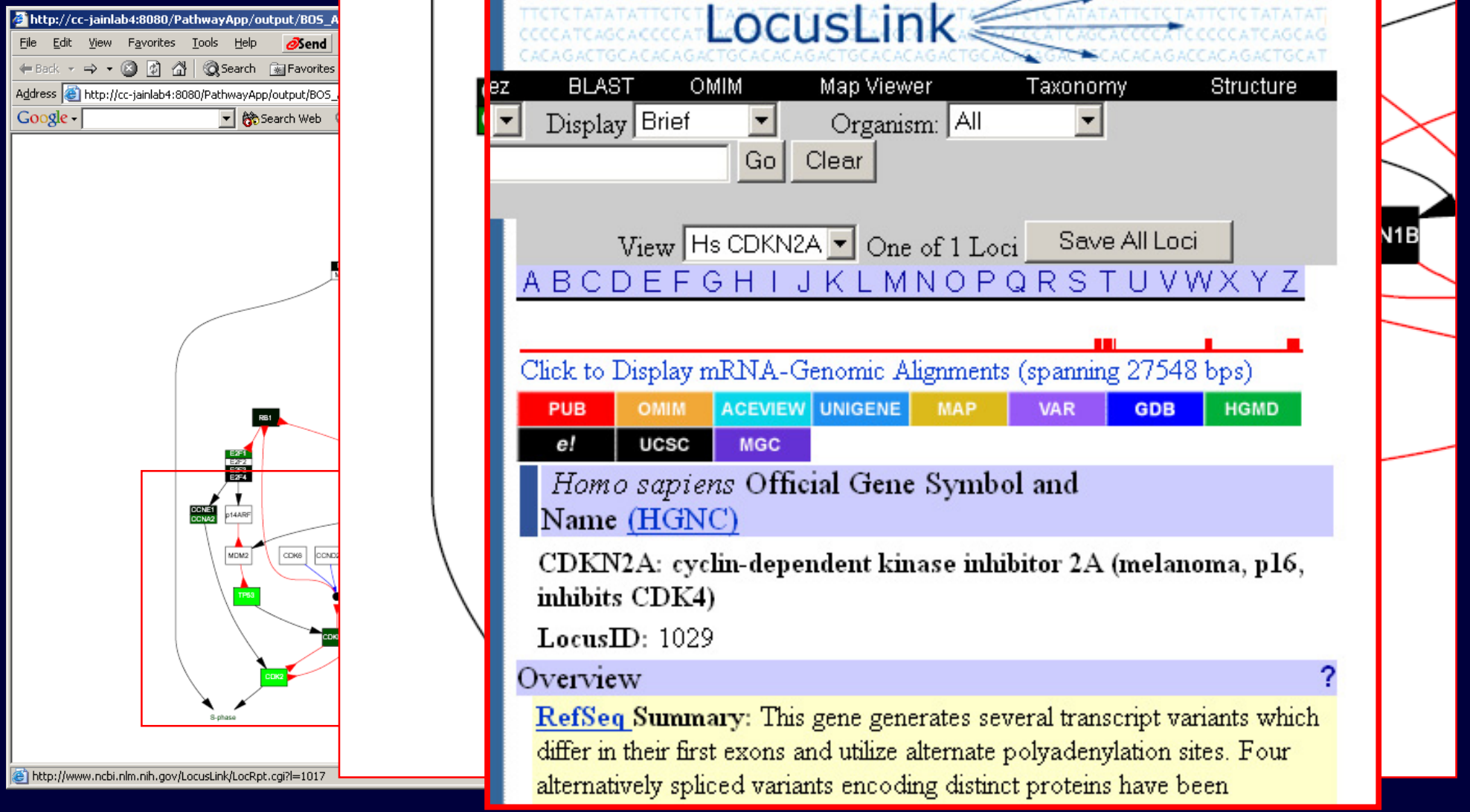
Pathway members colored by F-statistic based on class labeling

ERBB2, HGFR, TP53, CDK2, RACGAP1, + others are very different in the two cases

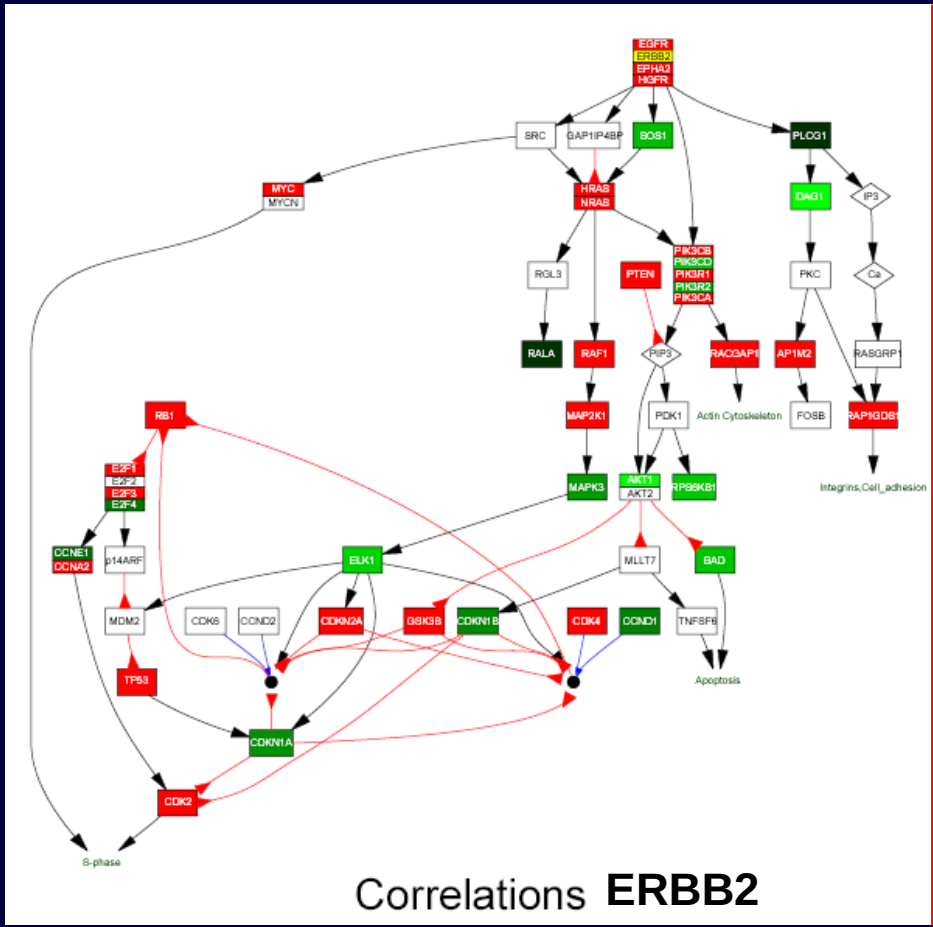
The individual nodes can be linked to external annotations



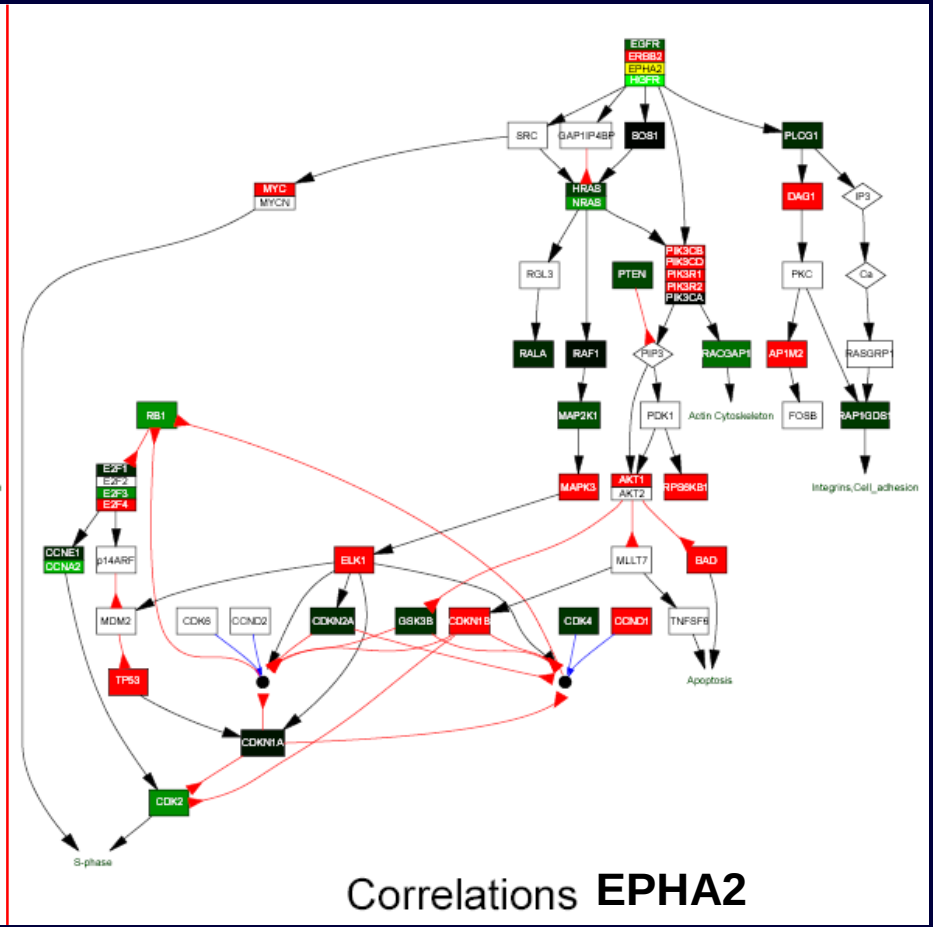
The individual nodes can be linked to external annotations



We see divergent behavior in genes relative to ERBB2 and EPHA2



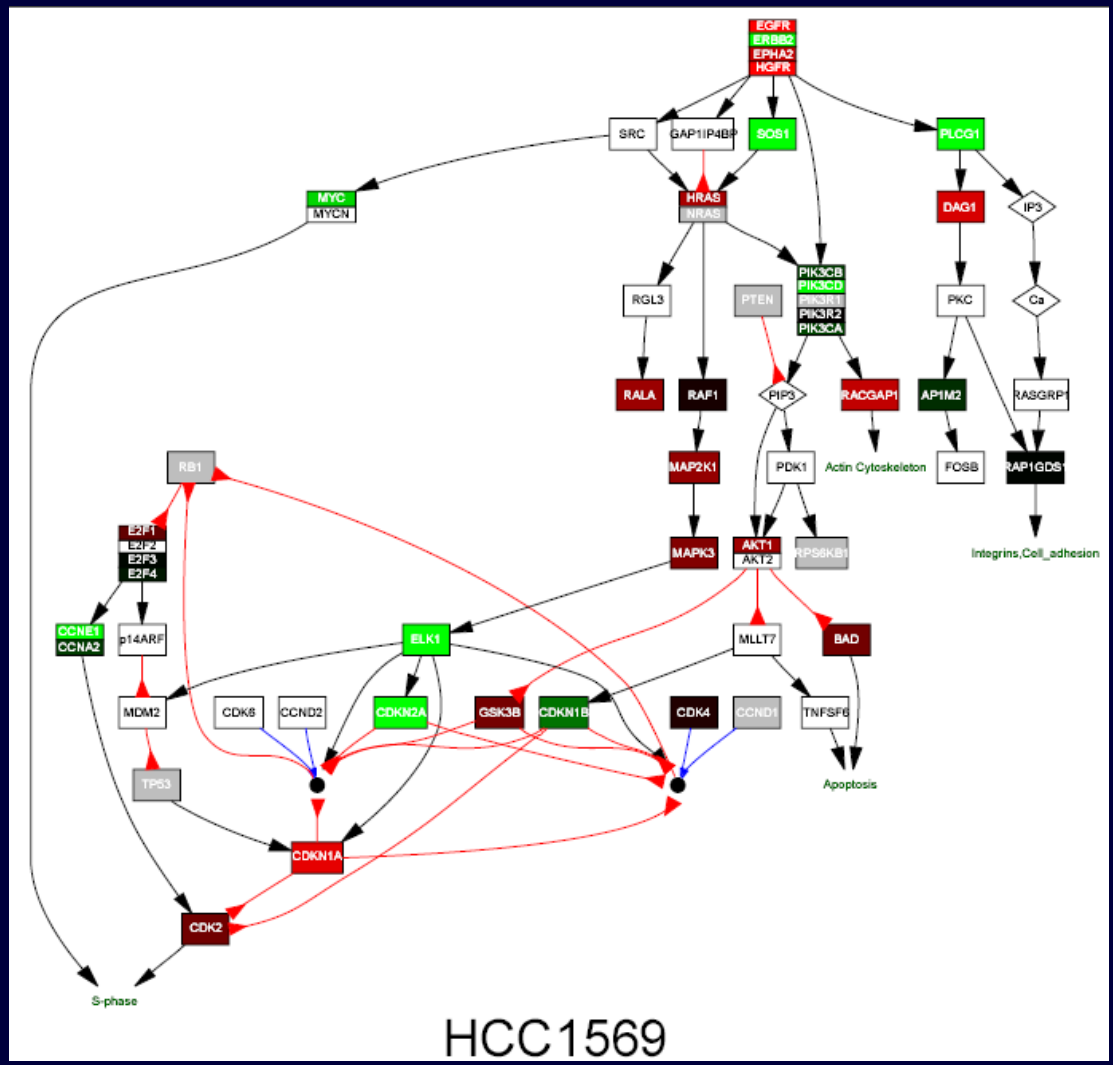
Correlations **ERBB2**



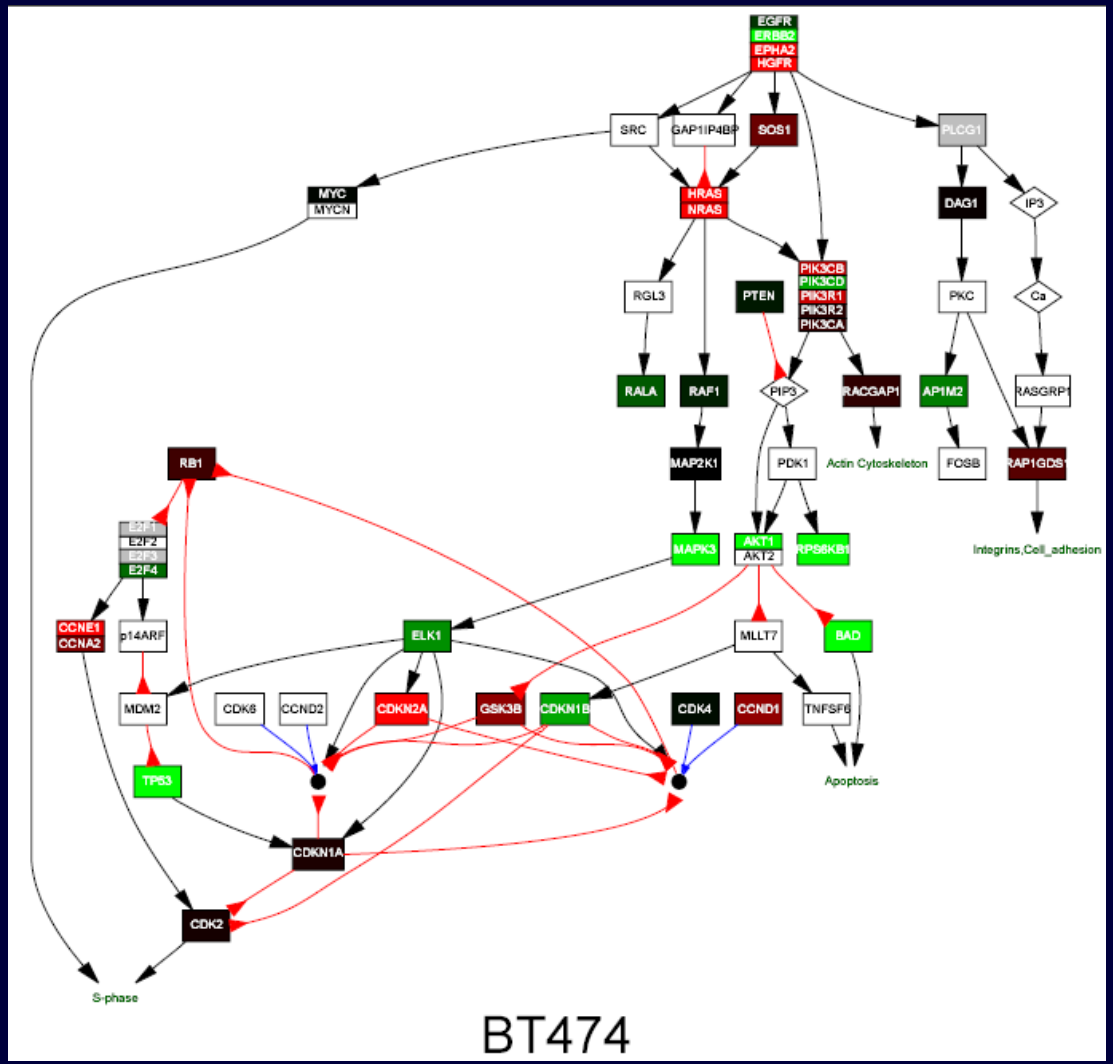
Correlations **EPHA2**

The correlations of genes to ERBB2 (left) and EPHA2 (right) are opposite in polarity in many cases. There appears to be a significant switch around EPHA2/ERBB2/ERBB3. Data and biological observations: J. Yeh, L. Timmerman, R. Neve, J. Gray, F. McCormick, J. Gray.

Within ERBB2 amplified cell lines, we see different patterns of gene expression



Within ERBB2 amplified cell lines, we see different patterns of gene expression



BT474

Data Analysis Conclusions: Magellan and QPACA

Array-oriented data is complex

- Variables/samples is unfavorable
- Quantitative conclusions can be elusive
- Using other knowledge can make such conclusions possible

Data integration is possible

- Requires extensive name-space cross-referencing
- One can then relate experimental data types (e.g. CGH and expression) to annotations (e.g. genomic mapping)

Annotation information is useful

- Genomic mapping
- Gene function
- Pathway structure

Goal: full integration of multiple experimental and annotation data types

- Move from one data type to another
- Restrict questions to those that are biologically focused and motivated
- Obtain quantitatively supportable hypotheses

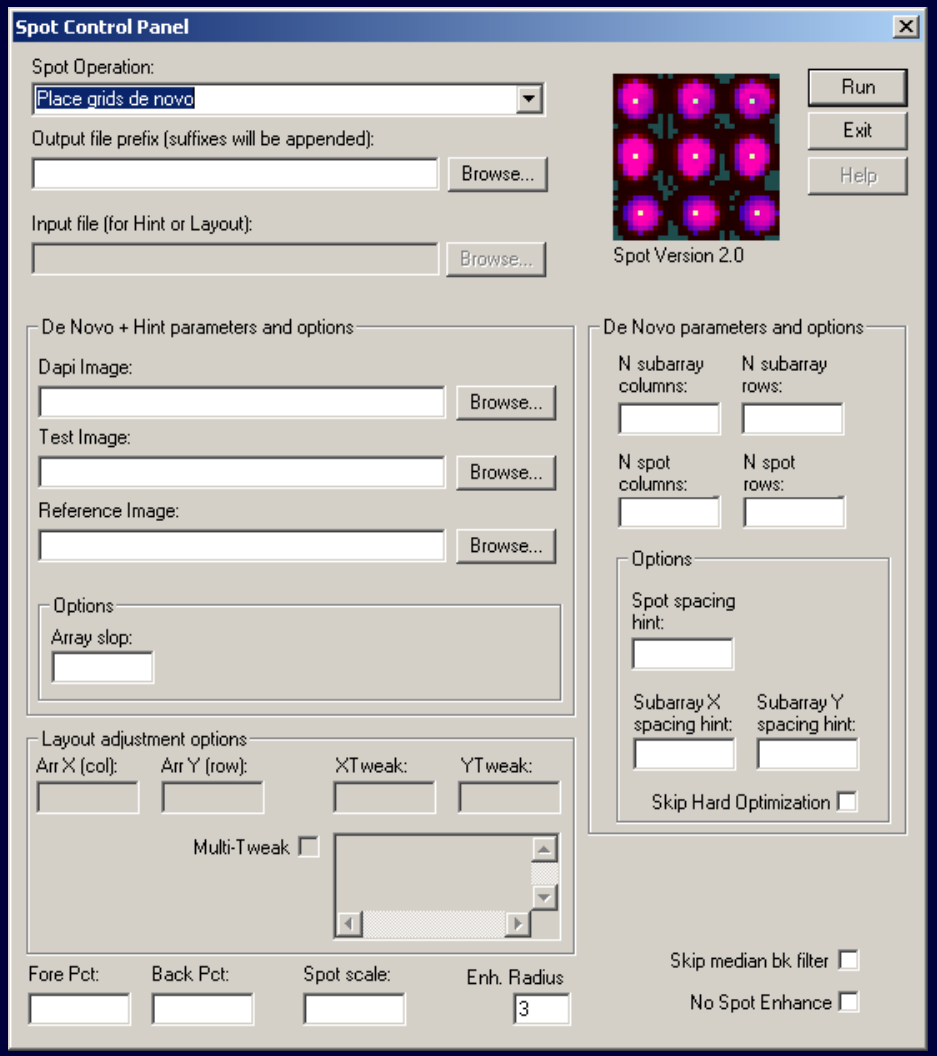
Data Generation and Storage

Spot is a fully automatic array quantification program

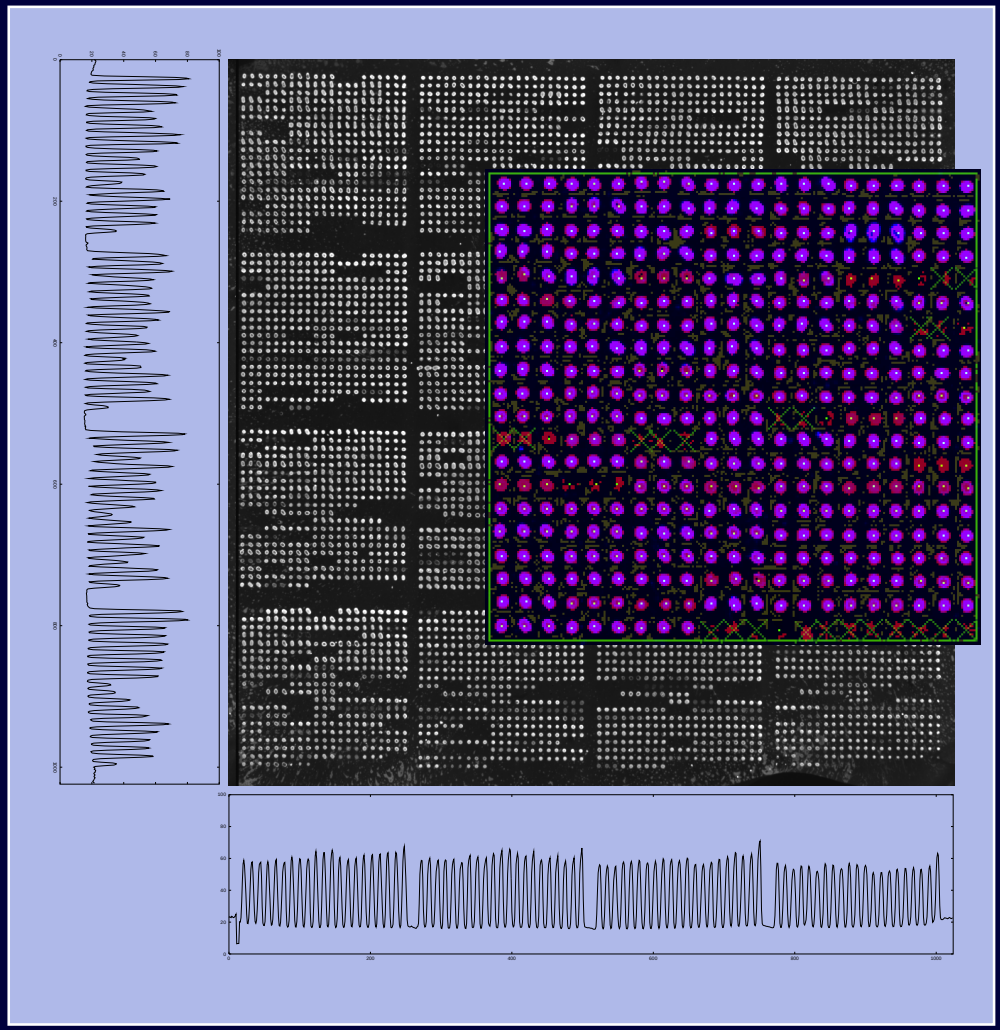
Spot goals

- Accurate
- Reproducible
- Fully automatic
- Fast
- Batch operation
- Adequate user interface
- Supportive of systematic post-processing (e.g. by generating extensive statistics on spots)

Academic investigators can get Spot from the Jain Lab: http://jainlab.ucsf.edu



Spot requires little information: images, geometry (e.g. 4x4, 21x20), output path



Spot Control Panel

Spot Operation:
Place grids de novo

Output file prefix (suffixes will be appended):
Output-file
Browse...

Input file (for Hint or Layout):
Browse...

Run
Exit
Help

Spot Version 1.2.3

De Novo + Hint parameters and options

Dapi Image:
Dapi-image.tif
Browse...

Test Image:
Cy3-image.tif
Browse...

Reference Image:
Cy5-image.tif
Browse...

Options
Array slop:

Layout adjustment options
Arr X (col):
Arr Y (row):
XTweak:
YTweak:
Multi-Tweak

Fore Pct:
Back Pct:
Spot scale:
Enh. Radius
3

De Novo parameters and options
N subarray columns: 4
N subarray rows: 4
N spot columns: 21
N spot rows: 20

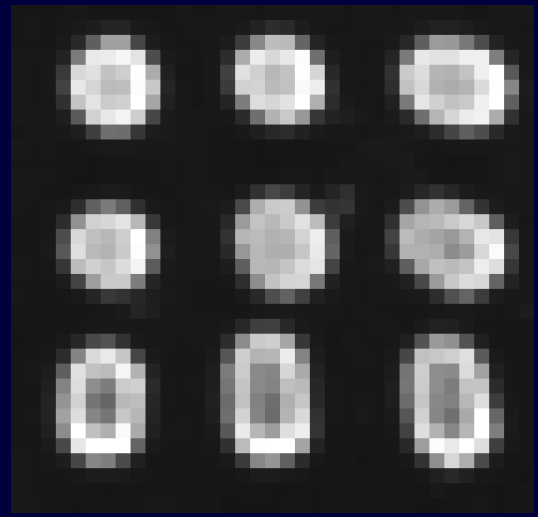
Options
Spot spacing hint:
Subarray X spacing hint:
Subarray Y spacing hint:
Skip Hard Optimization

Normalize background
No Spot Enhance

Spot segments targets explicitly

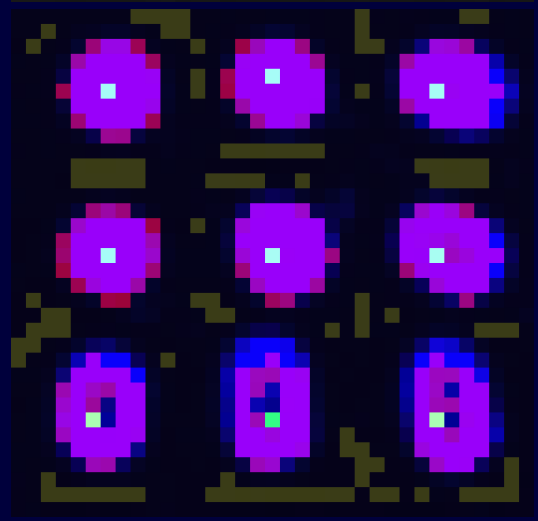
Dapi (or composite cy3/cy5) image for segmentation

- Grid placement is automatic, allowing for deviations from rectilinearity
- Histogram-based thresholds plus a geometric constraint are used for pixel identification



Segmentation summary image for review

- Blue channel: original Dapi
- Red: pixels identified as “spot”
- Dim yellow: pixels identified as background



Spot produces statistics for each spot
It is the de-facto standard for array-CGH quantification

SpotNum	1	2	3	Unique integer for each spot.
Image-x	28	41	52	X image coordinate of spot center.
Image-y	25	25	25	Y image coordinate of spot center.
Arr-colx	1	1	1	Subgrid column (1-based).
Arr-rowy	1	1	1	Subgrid row.
Spot-colx	1	2	3	Spot column (within subgrid).
Spot-rowy	1	1	1	Spot row.
Flag	0	0	0	Flag: 0 if spot is OK.
nfore	23	24	22	Number of foreground pixels.
nback	10	11	7	Number of background pixels.
DapiFore	644.609	644.292	618.818	Dapi (or composite) channel foreground mean.
DapiBack	335	336.091	336.714	Dapi background mean.
Dapi	309.609	308.201	282.104	Dapi signal (fore - back).
TestFore	1635.043	1674.083	1706.409	Test foreground mean.
TestForeSD	170.962	198.861	145.838	Test foreground standard deviation.
TestBack	481.3	499	501.571	Test background mean.
TestBackSD	11.973	17.39	15.252	Test background standard deviation.
Test	1153.743	1175.083	1204.838	Test signal (fore - back).
RefFore	751	779.75	778.727	Reference
RefForeSD	48.677	58.303	51.374	
RefBack	420.8	429.818	433.429	
RefBackSD	9.028	8.875	4.276	
Ref	330.2	349.932	345.299	
Log2Rat	1.805	1.748	1.803	log2(RawRat)
RawRat	3.494	3.358	3.489	Test/Ref
SpotCorr	0.678	0.804	0.593	Pearson's correlation of test to ref foreground pixels.
TestZstat	32.184	28.71	38.1	Z-statistic of difference between test fore and back.
RefZstat	31.317	28.687	31.188	Z-statistic of difference between ref fore and back.
MinF-B	309.609	308.201	282.104	Minimum (fore-back) for all three channels.
Slope	2.382	2.742	1.683	Slope of line fit to test/ref fore pixels.
Log2Slope	1.252	1.455	0.751	log2(Slope)
TestIntercept	0.318	0.184	0.518	Normalized intercept.
RefIntercept	-0.467	-0.225	-1.074	Normalized intercept.
OriginDist	0.565	0.29	1.192	Distance of intercept to origin.
MeanLog2Rat	1.804	1.746	1.809	Mean of pixel by pixel log2(ratio)
MeanRat	3.492	3.353	3.504	Corresponding non-logged ratio to MeanLog2Rat
MedianRatio	3.525	3.258	3.354	Median pixel by pixel ratio.
Log2MedRat	1.818	1.704	1.746	log2(MedianRatio)

Important columns

- SpotNum
- Log2Rat
- Nfore
- SpotCorr
- TestZstat
- RefZstat

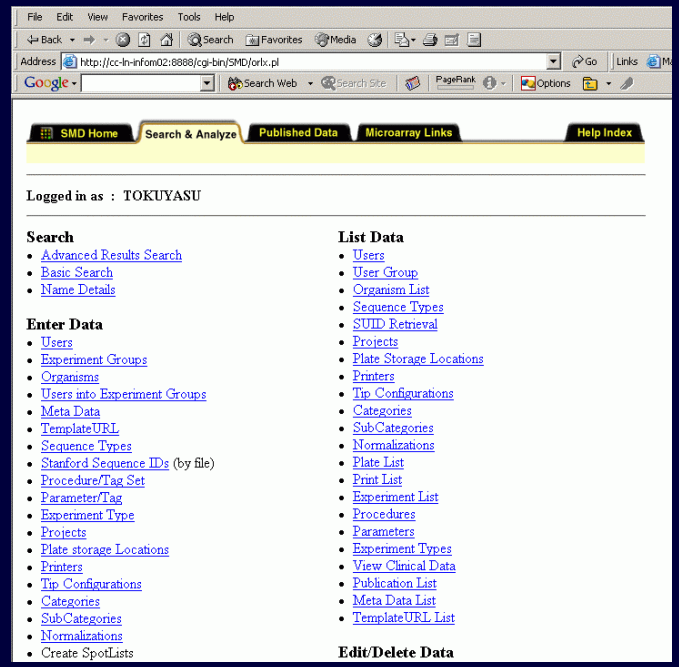
A. N. Jain, T. A. Tokuyasu, A. M. Snijders, R. Segraves, D. G. Albertson, and D. Pinkel. Fully automatic quantification of microarray image data. Genome Research 12: 325-332, 2002.

We are also addressing the data management issue: CC Informatics Core

Systems biology experiments with local custom data sources require scalable and extensible data systems

We are extending the Stanford SMD platform

- Linux implementation (versus Sun Solaris $$$$)
- Oracle back-end
- Augmenting for additional data types
  - cDNA expression (made for this)
  - Affymetrix expression
  - Custom oligo-based expression
  - Array-based CGH
  - ESP breakpoints
  - Protein measurements
  - Numerical phenotypic data
- We may move to LAD (Longhorn w/Postgresql)



UCSF-CC and CaBIG

- Tools under development: Magellan and QPACA
- Completed tools: Spot and SPROC
- Data management: SMD-squared
- Data to share: array-CGH, ESP, expression…
- There’s also a bunch of stuff we want!