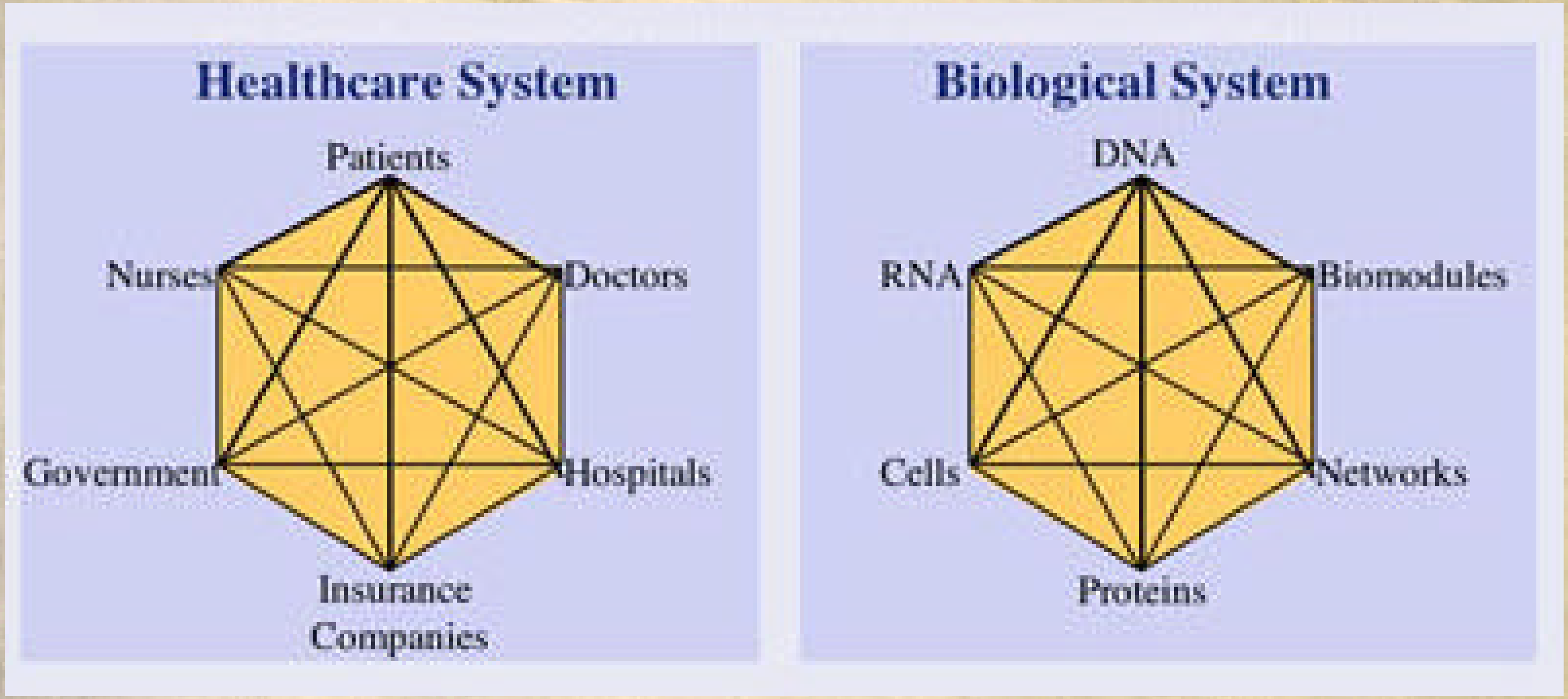
Bioinformatics in the Health Sciences:

Towards tailored medicine?

Brendan McConkey
Department of Biology
University of Waterloo

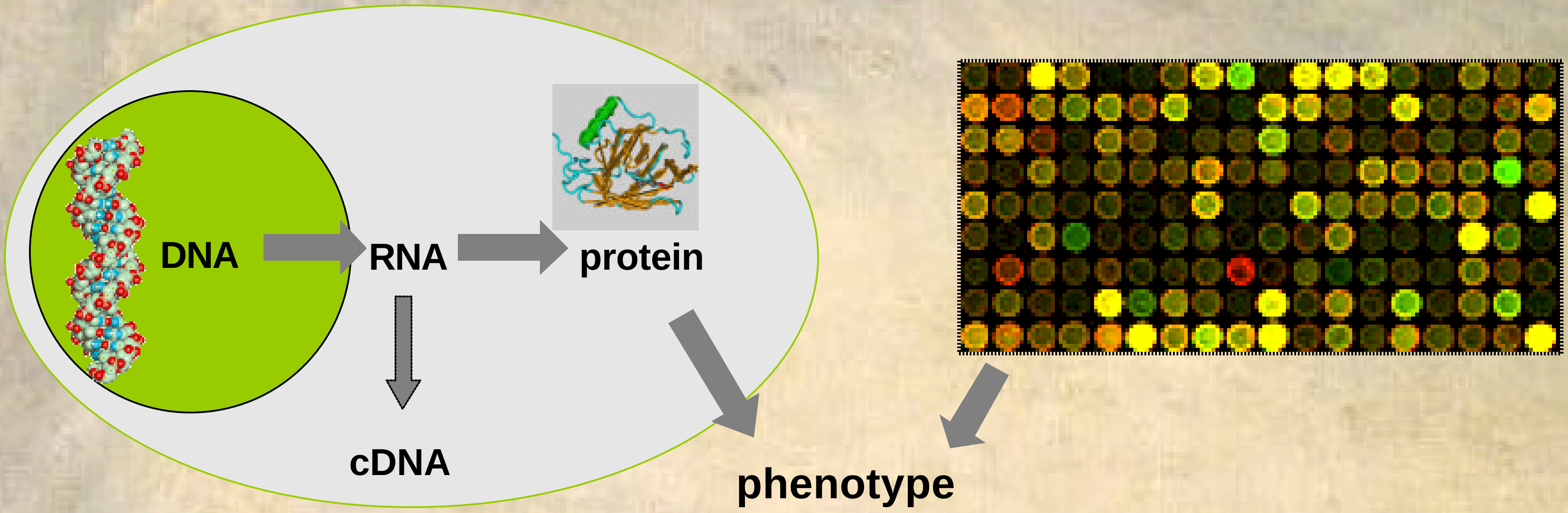
Outline:

- health informatics and bioinformatics
- systems biology approaches
- bioinformatics technologies in health sciences
- biomarkers and diagnosis
- combinatoric therapeutics



www.systemsbiology.org

Bioinformatics and Health Informatics are both concerned with data management, and interactions between components of the system



adapted from Pevsner, 2003

hypothesis driven research

focus on one gene
or one protein

↓

determine relations
between genes/proteins

↓

integrate components,
describe effect on phenotype

↔

discovery based research

enumerate parts of a system
(gene or protein expression)

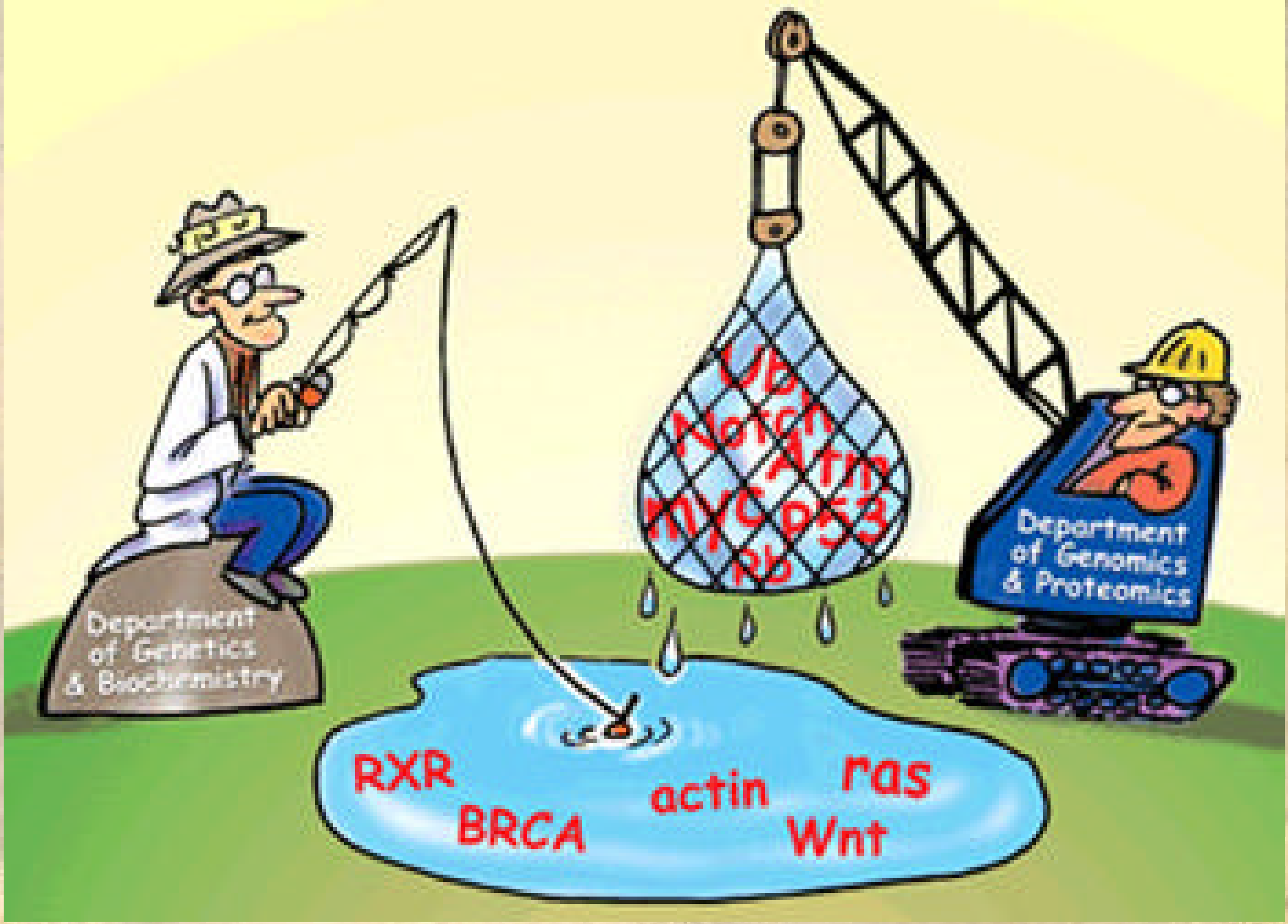
↓

identify patterns in data,
relations between components

↓

relate patterns to phenotype

The optimistic view of genomics and proteomics:


Department of Genetics & Biochemistry
Ub
Notch
Atm
myc
p53
Rb
Department of Genomics & Proteomics
RXR
BRCA
actin
ras
Wnt

Science (2001) **291** :1221

... but can you tell a fish from a rubber boot?

The promise of bioinformatics in the health sciences

- genetic profiling
  - identification of genetic predispositions

- prognosis and treatment
  - prediction of response to treatment
  - tailoring treatment by tissue subtype

- diagnostics - early detection of disease
  - serum protein biomarkers

- identification of novel drug targets

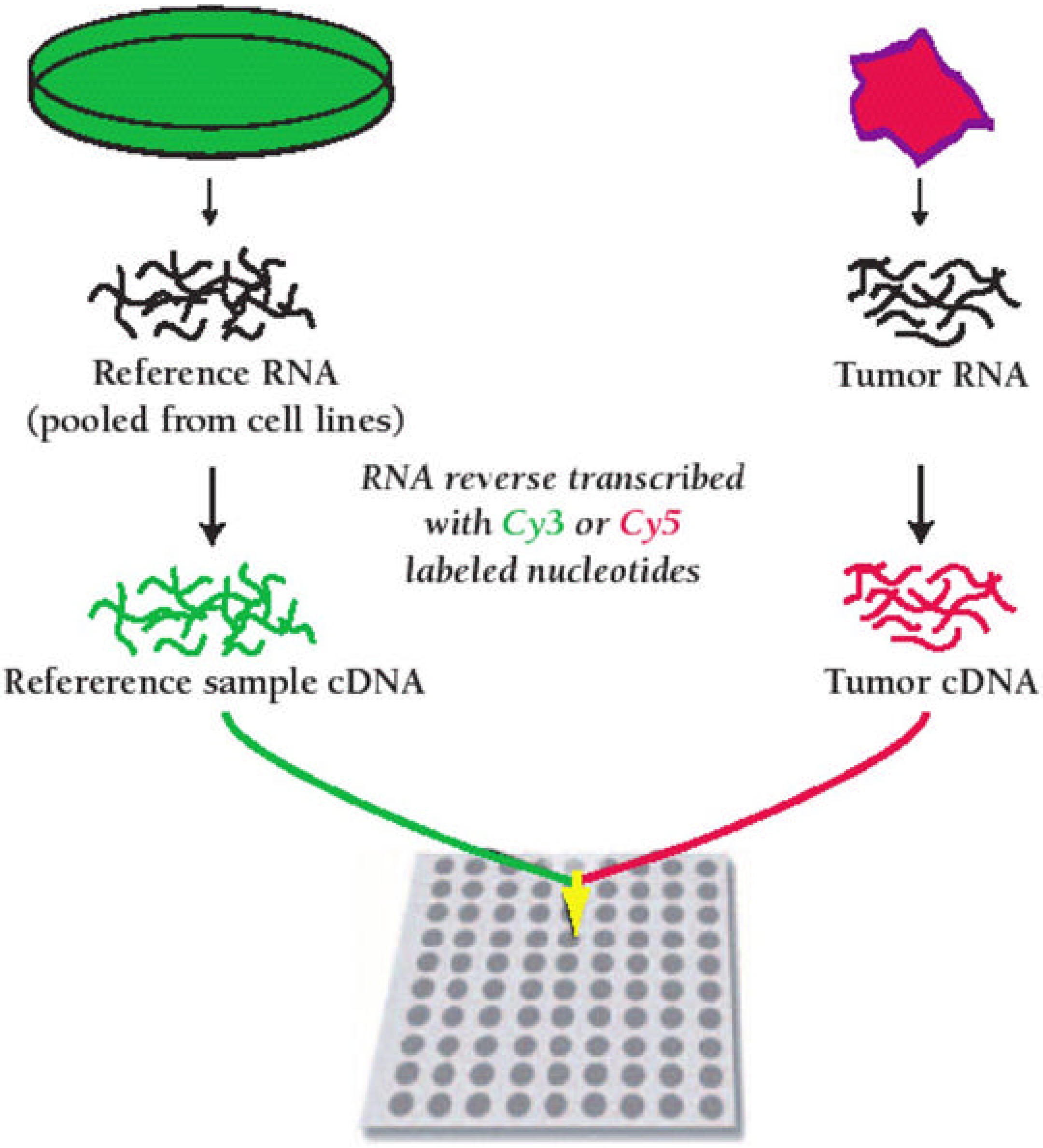
- application to multi-factor disease

The down side

- highly dimensional data sets
  - e.g. two treatments, >10,000 genes
- data analysis - what does it all mean?
- ‘black-box’ approaches
- isolating cause and effect
- data management
- currently, cost is often high
  - expensive equipment and/or consumables

Examples:

- microarrays in the diagnosis and treatment of breast cancer
- biomarkers of disease
  - serum analysis
- identification of drug targets
- cell cycle modeling - controls on cell division



Reference RNA
(pooled from cell lines)

Tumor RNA

RNA reverse transcribed
with Cy3 or Cy5
labeled nucleotides

Refererence sample cDNA

Tumor cDNA

cDNAs mixed and hybridized to microarray

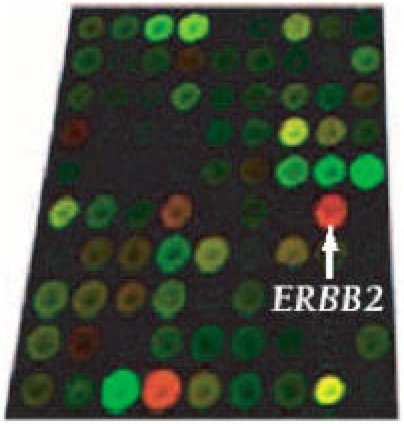


image of scanned microarray

Jeffrey SS, Fero MJ, Borresen-Dale AL, Botstein D. Expression array technology in the diagnosis and treatment of breast cancer. Mol Interv. 2002 Apr;2(2):101-9.

Microarray setup

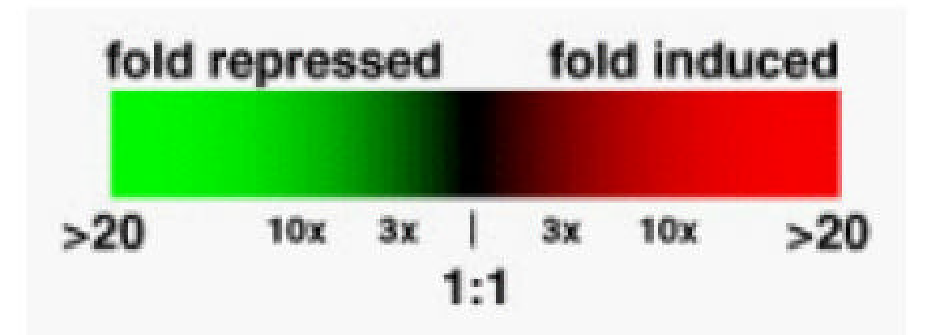
Stanford (Brown lab) microarrays

- unique Expressed Sequence Tag clusters for human cDNA
- >20000 ESTs represented
- typical distribution:
  - 40% annotated genes (UniGene)
  - 10% partly annotated
  - 50% little or no annotation
- requires 2-4 ug mRNA
- RNA sample often amplified
- standard reference RNAs available

Microarray data handling

- each data point (spot on microarray) represents a change in expression level versus a reference sample
  (cDNA, sample#, ? expression)

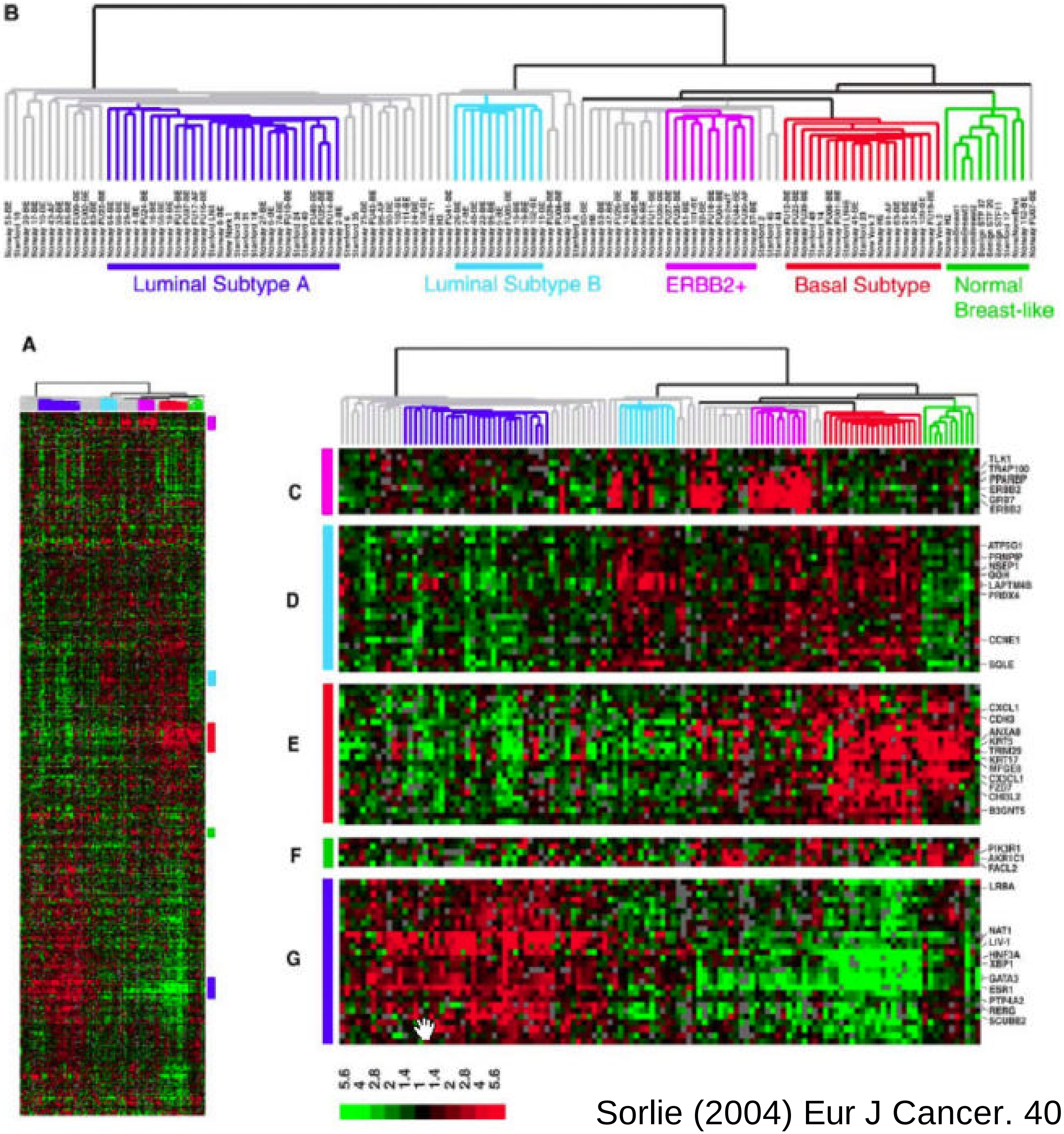
- changes in expression ratios can be represented on a colour scale, to enable visualization of large data sets



after Campbell and Heyer, 2002

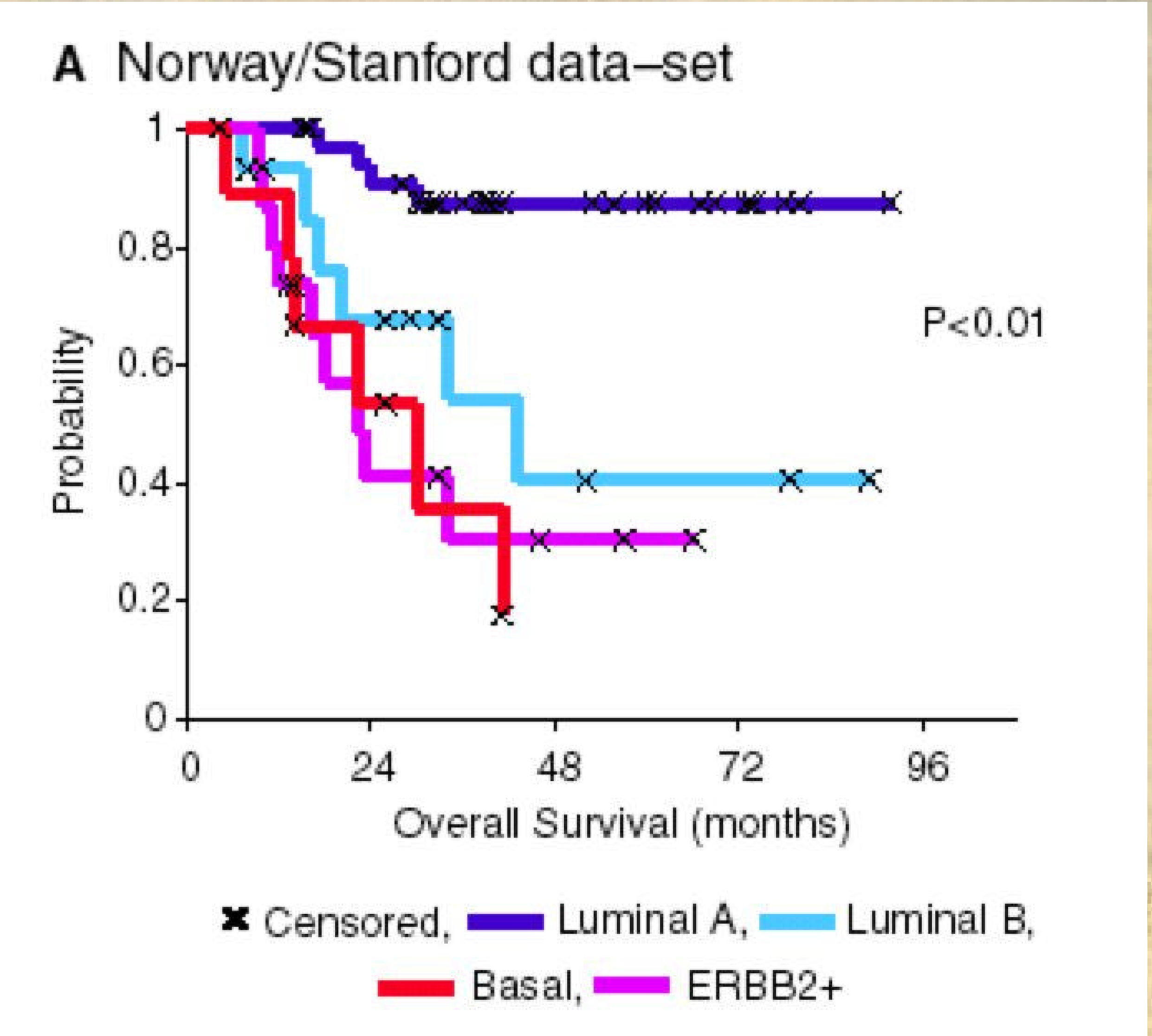
- data points may be clustered by sample similarity and by expression similarity

- sample data set: 115 breast tumor tissues + 7 non-malignant tissues



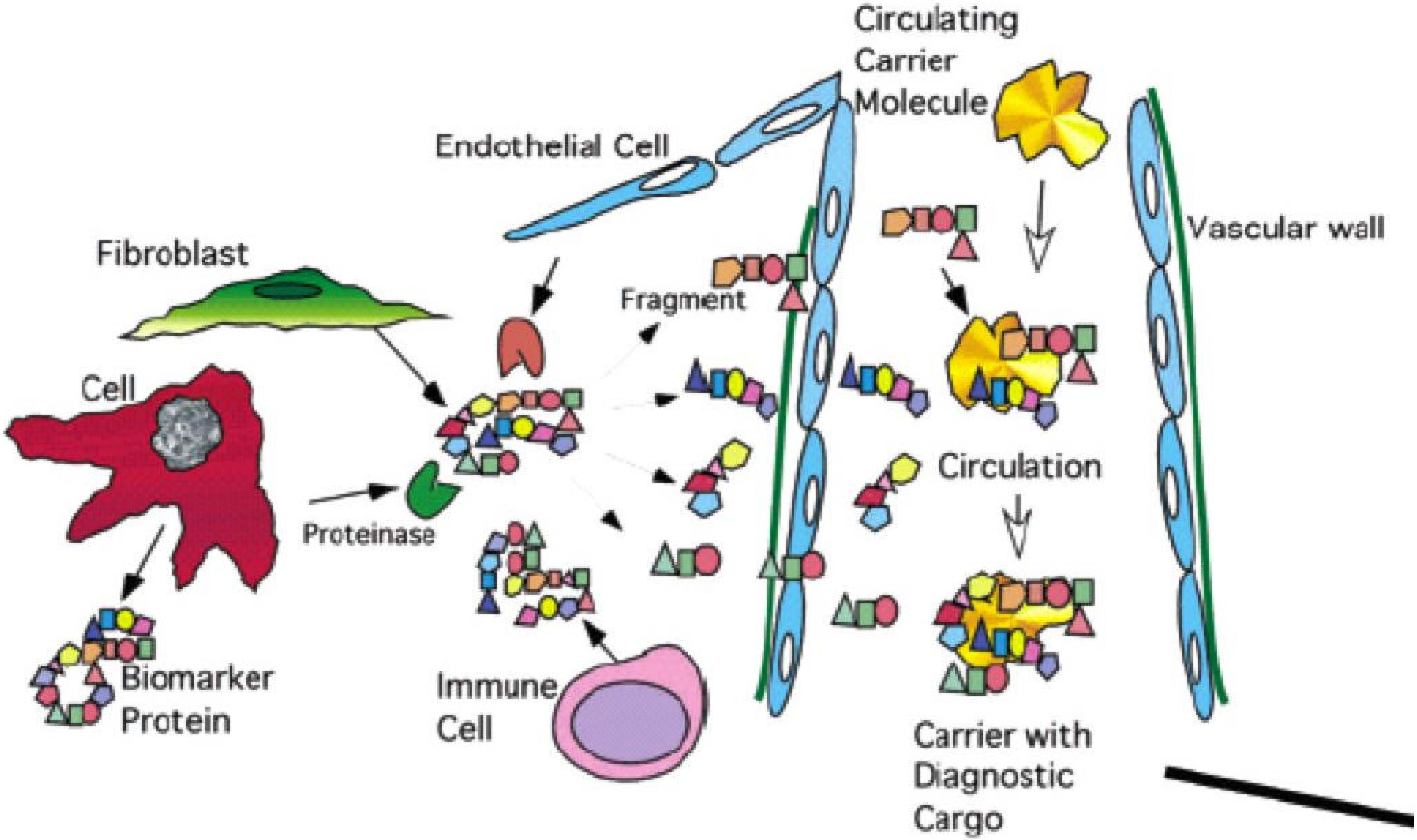
Sorlie (2004) Eur J Cancer. 40(18):2667-75.

Tumor subtype is highly correlated to survival



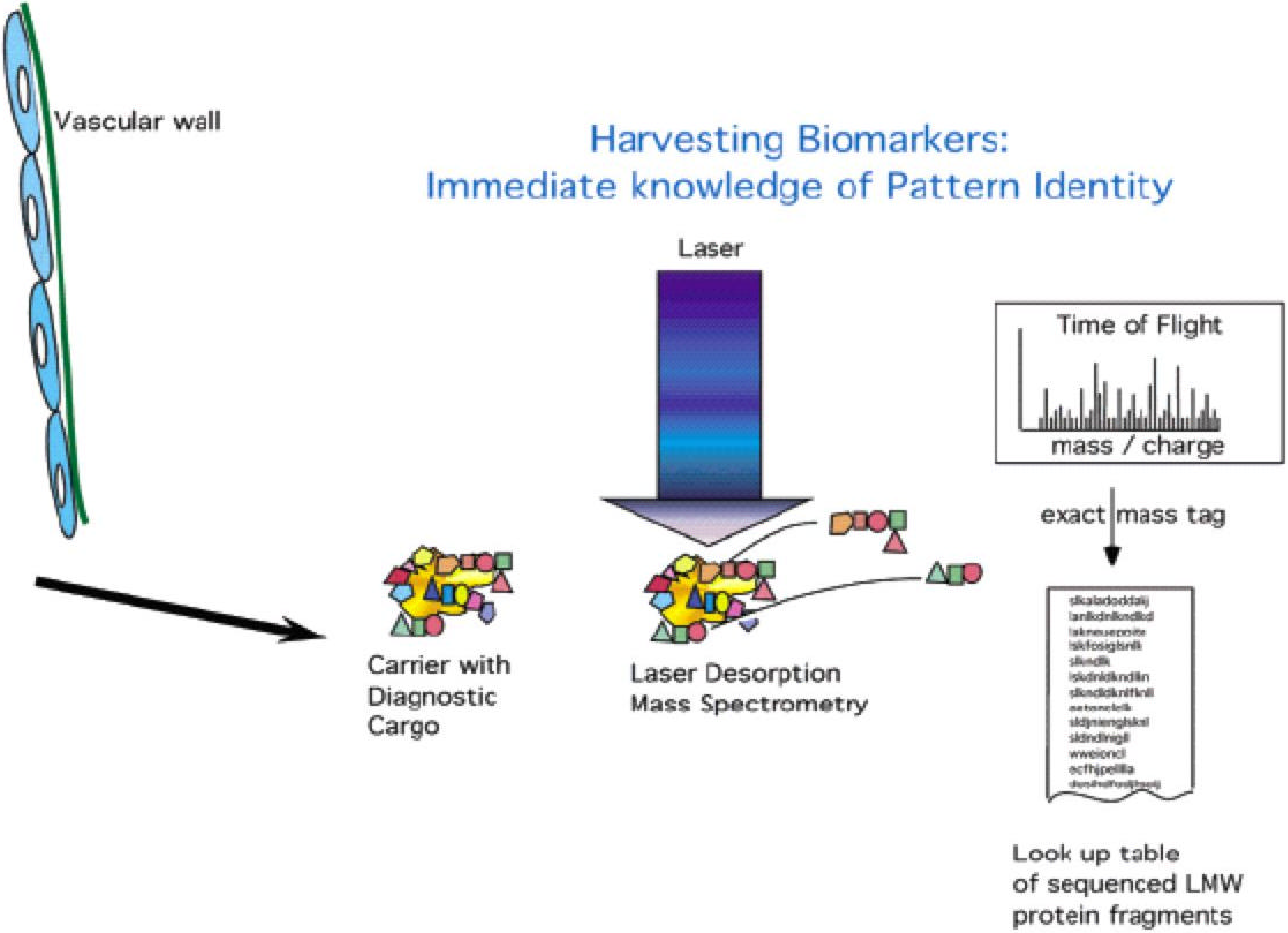
Sorlie (2004) Eur J Cancer. 40(18):2667-75.

Biomarker Amplification and Harvesting by Carrier Molecules



After Petricoin (2004) J Proteome Res.

Harvesting Biomarkers: Immediate knowledge of Pattern Identity



Vascular wall

Laser

Time of Flight

mass / charge

exact mass tag

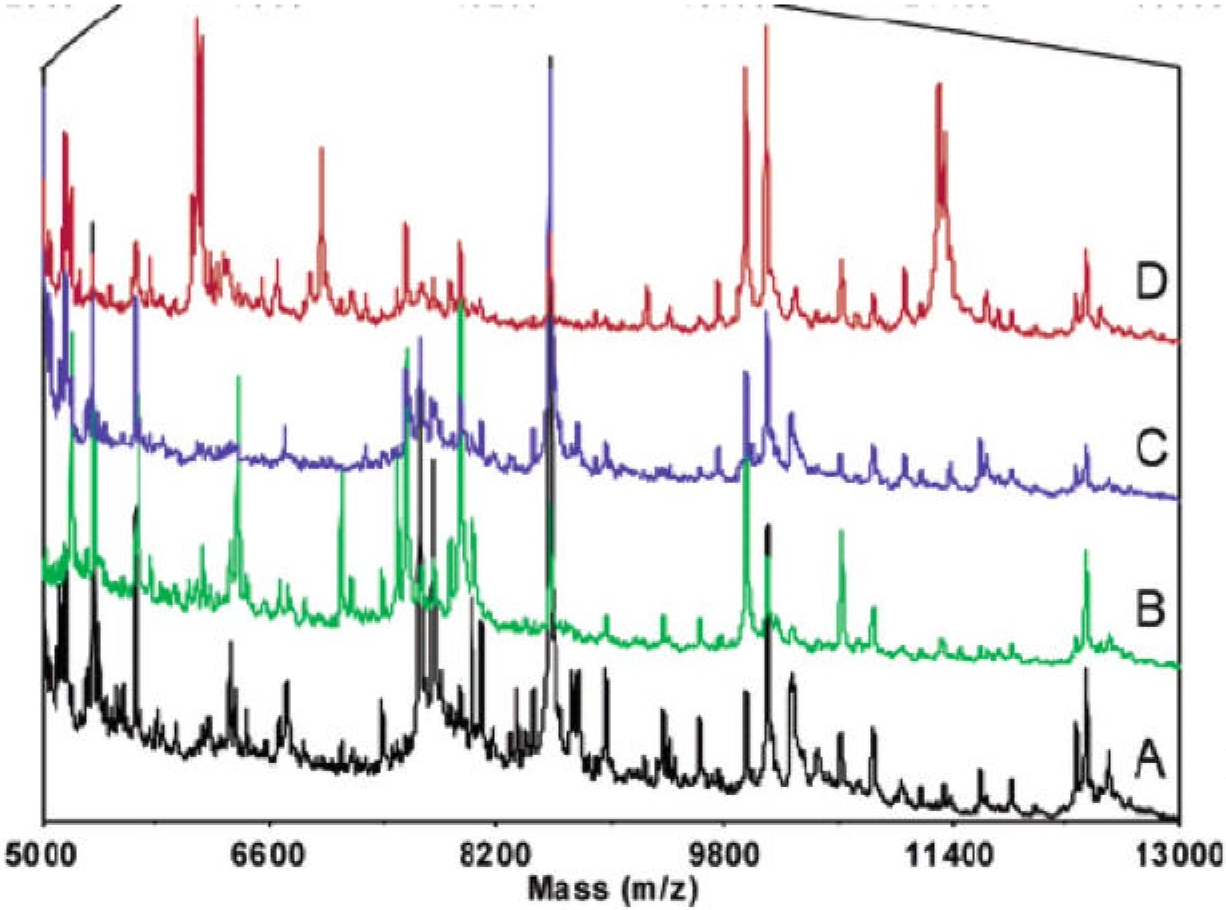
Carrier with
Diagnostic
Cargo

Laser Desorption
Mass Spectrometry

Look up table
of sequenced LMW
protein fragments

After Petricoin (2004) J Proteome Res.

Mass spectrometry of brain tumor biopsy samples



glioblastoma

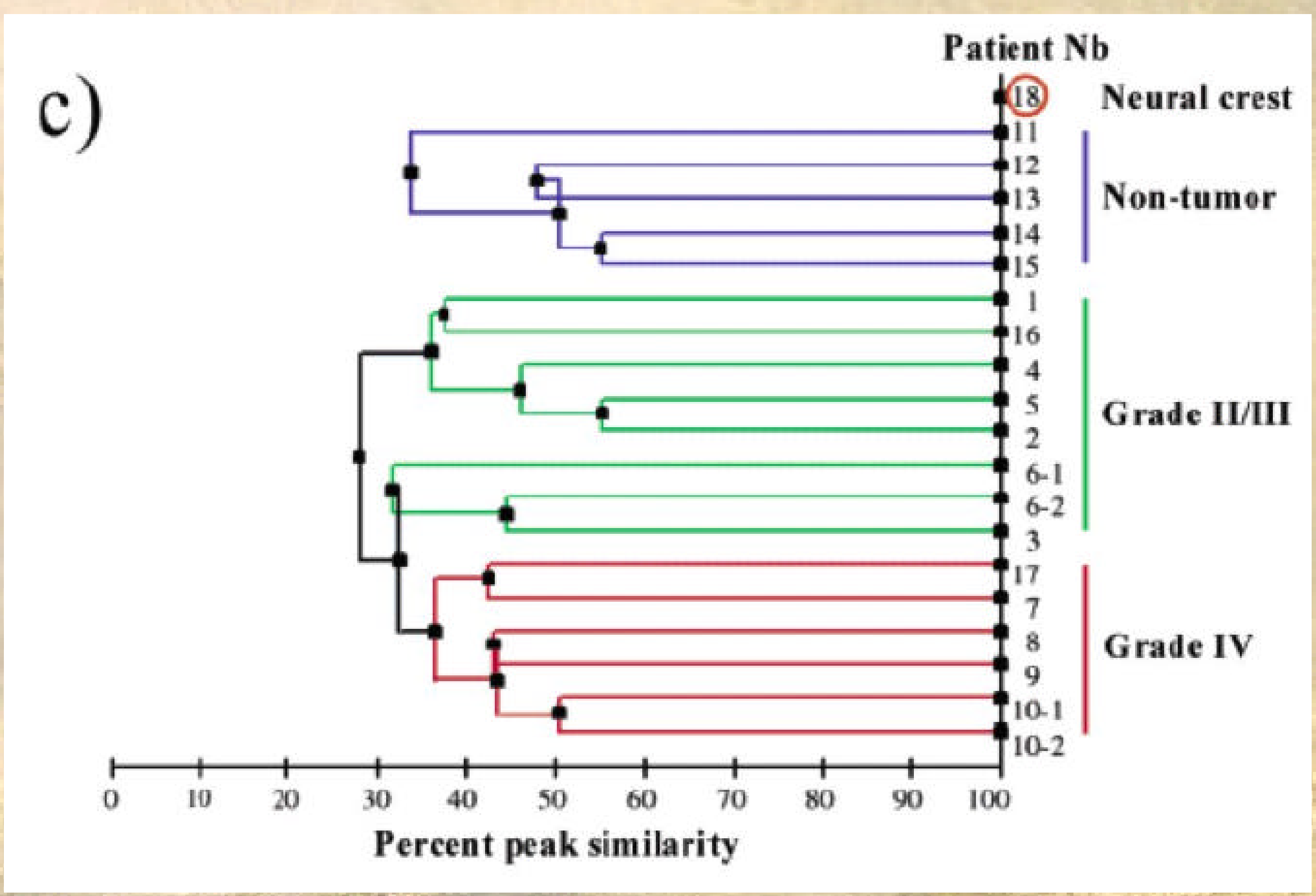
grade III astrocytoma

grade II astrocytoma

normal cells

Mass spectrometry of brain tumor biopsy samples

-classification from mass spectrometry pattern:

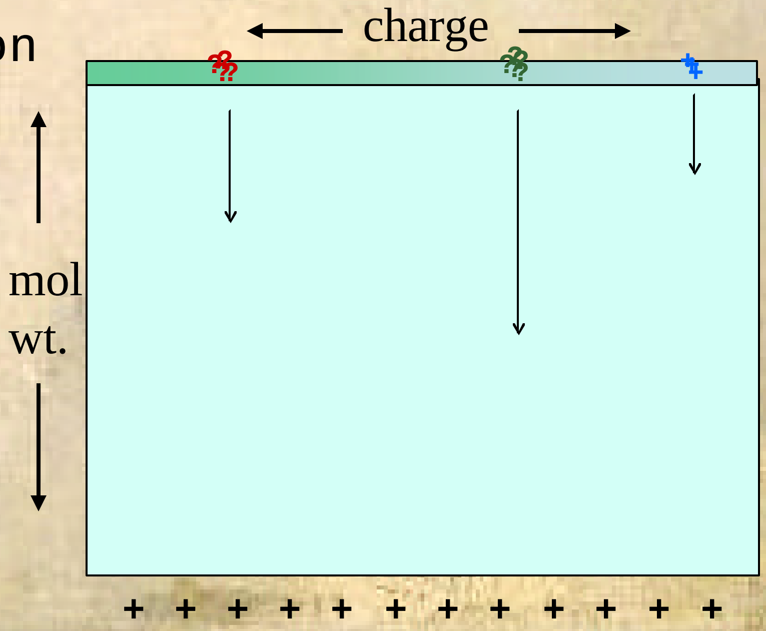


Proteomics using 2-D electrophoresis

- separation in first dimension by charge, second dimension by molecular weight
- 1st dimension uses an immobilized pH gradient (e.g. pH 3-10)
- high voltage is gradually applied across gel, causing proteins to migrate to the location where they have zero net charge (Isoelectric Focussing - IEF)
- SDS-PAGE used for second dimension

  separation by M.W.
- proteins are visualized by staining -

  silver stain, fluorescent dyes, …



Differential In-gel Electrophoresis

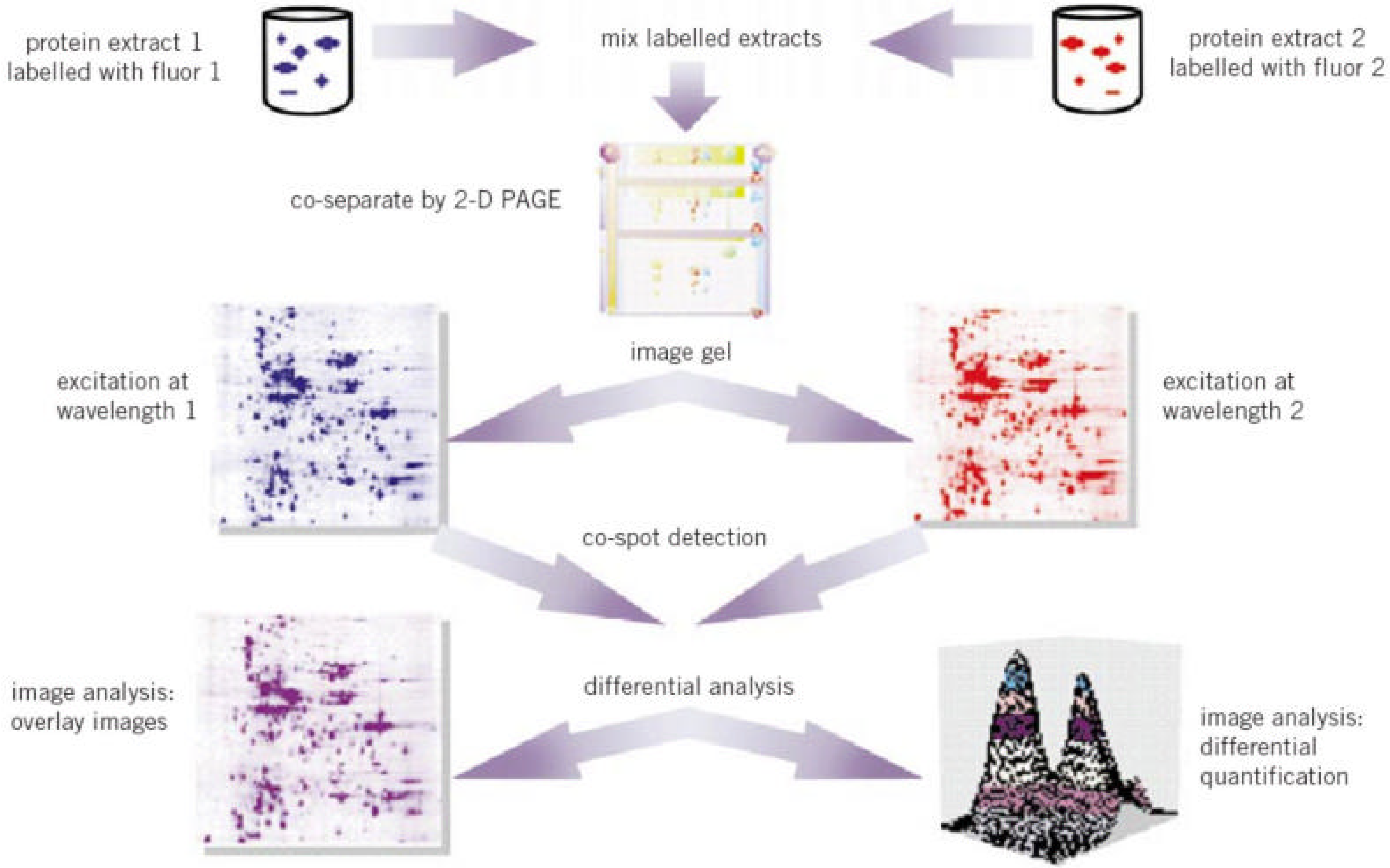
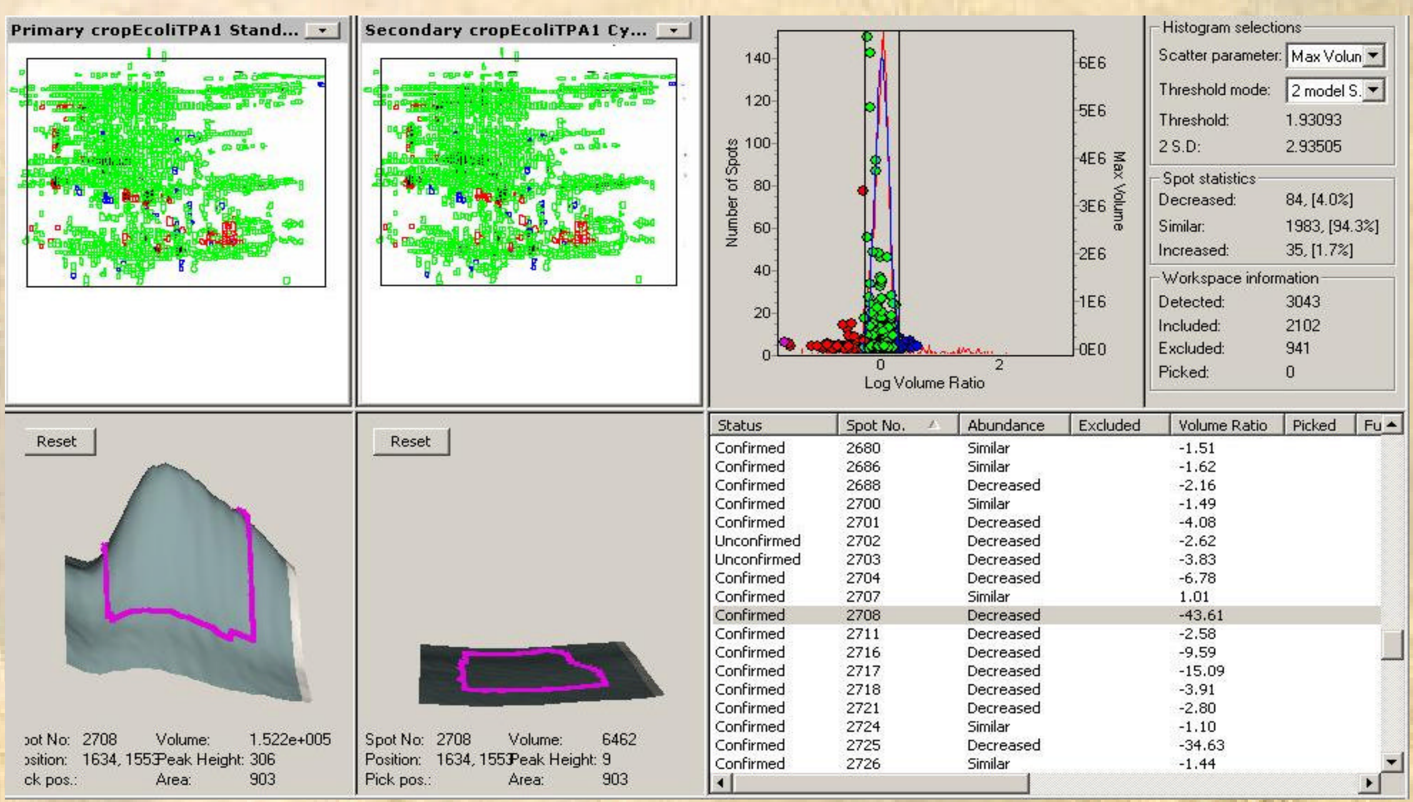


Fig 1. Outline of the 2-D DIGE technology. Protein samples are labelled with different fluorescent dyes, mixed, and co-separated by 2-D electrophoresis. Spots in the gel are visualized in the CCD-based imager and quantitatively analysed using 2-D analysis software. Spots showing quantitatively significant differences are then picked, digested, and analysed by mass spectrometry.

Statistical analysis using DeCyder Software



Histogram selections

Scatter parameter: Max Volun

Threshold mode: 2 model S.

Threshold: 1.93093

2 S.D: 2.93505

Spot statistics

Decreased: 84, [4.0%]

Similar: 1983, [94.3%]

Increased: 35, [1.7%]

Workspace information

Detected: 3043

Included: 2102

Excluded: 941

Picked: 0

ɔot No: 2708 Volume: 1.522e+005

ɔsition: 1634, 1553 Peak Height: 306

ck pos.: Area: 903

Spot No: 2708 Volume: 6462

Position: 1634, 1553 Peak Height: 9

Pick pos.: Area: 903

Status	Spot No.	Abundance	Excluded	Volume Ratio	Picked	Fu
Confirmed	2680	Similar		-1.51		
Confirmed	2686	Similar		-1.62		
Confirmed	2688	Decreased		-2.16		
Confirmed	2700	Similar		-1.49		
Confirmed	2701	Decreased		-4.08		
Unconfirmed	2702	Decreased		-2.62		
Unconfirmed	2703	Decreased		-3.83		
Confirmed	2704	Decreased		-6.78		
Confirmed	2707	Similar		1.01		
Confirmed	2708	Decreased		-43.61		
Confirmed	2711	Decreased		-2.58		
Confirmed	2716	Decreased		-9.59		
Confirmed	2717	Decreased		-15.09		
Confirmed	2718	Decreased		-3.91		
Confirmed	2721	Decreased		-2.80		
Confirmed	2724	Similar		-1.10		
Confirmed	2725	Decreased		-34.63		
Confirmed	2726	Similar		-1.44		

Non-Induced Culture

Advantages of DIGE

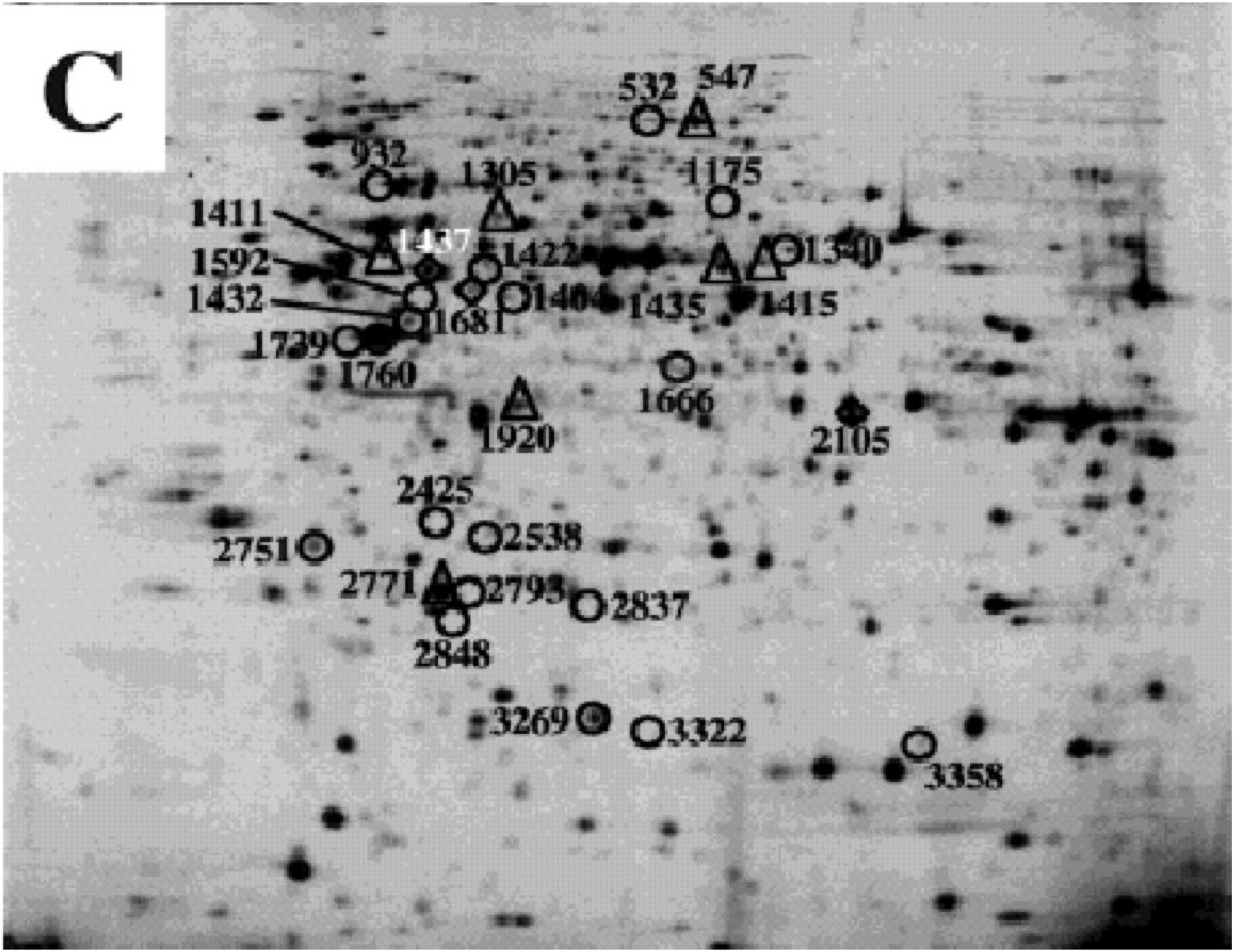
- multiplexed samples in the same gel
- Spot matching of paired samples vs. internal standard
- Automated spot matching
- differential analysis and statistical significance estimates
- Multiple gels analyzed vs. same internal standard
- large reduction in variation when comparing between samples

Limitations of 2-DE

- No information on location within cell unless pre-fractionation of sample into cellular components
- Low abundance proteins difficult to see but may be overcome by pre-fractionation at cellular level, HPLC, differential solubilization
- No kinetic information
- Denaturing disassociates protein subunits, cofactors, substrates

Proteomics using 2-D electrophoresis

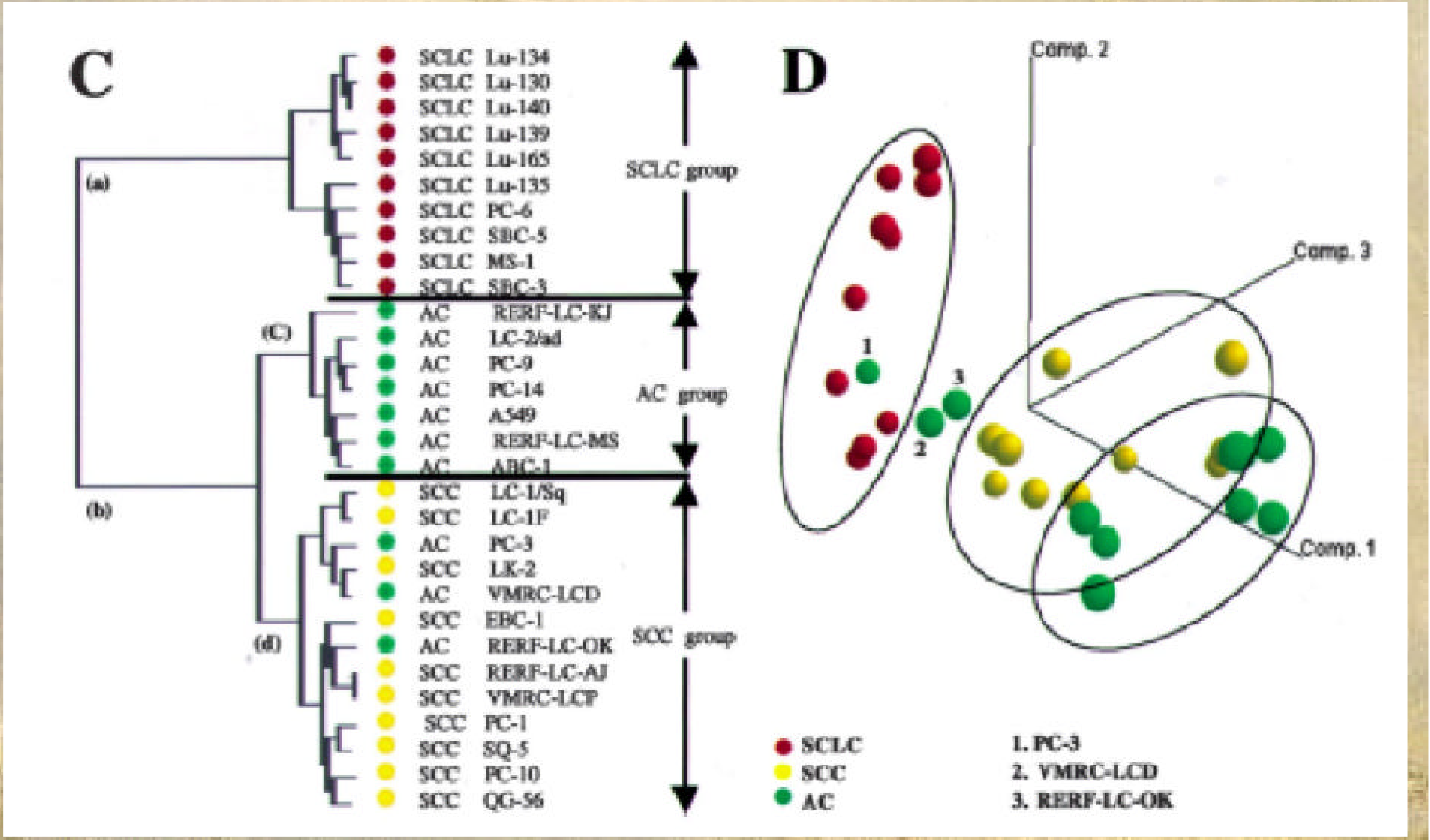
- protein expression patterns may be able to distinguish between cancer types
- determine origin of metastatic tumours
- example: classification of lung cancer subtypes using 2D-DIGE



Spots uniformly up-regulated (triangles) or down-regulated (circles) in lung small cell carcinoma. Seike et al 2004, *Proteomics* 4: 2776.

Hierarchical cluster analysis of 71 protein spots on DIGE

SCLC, small cell lung carcinoma; SCC, squamous cell carcinoma; AC, adenocarcinoma.



Seike et al 2004, *Proteomics* 4: 2776.

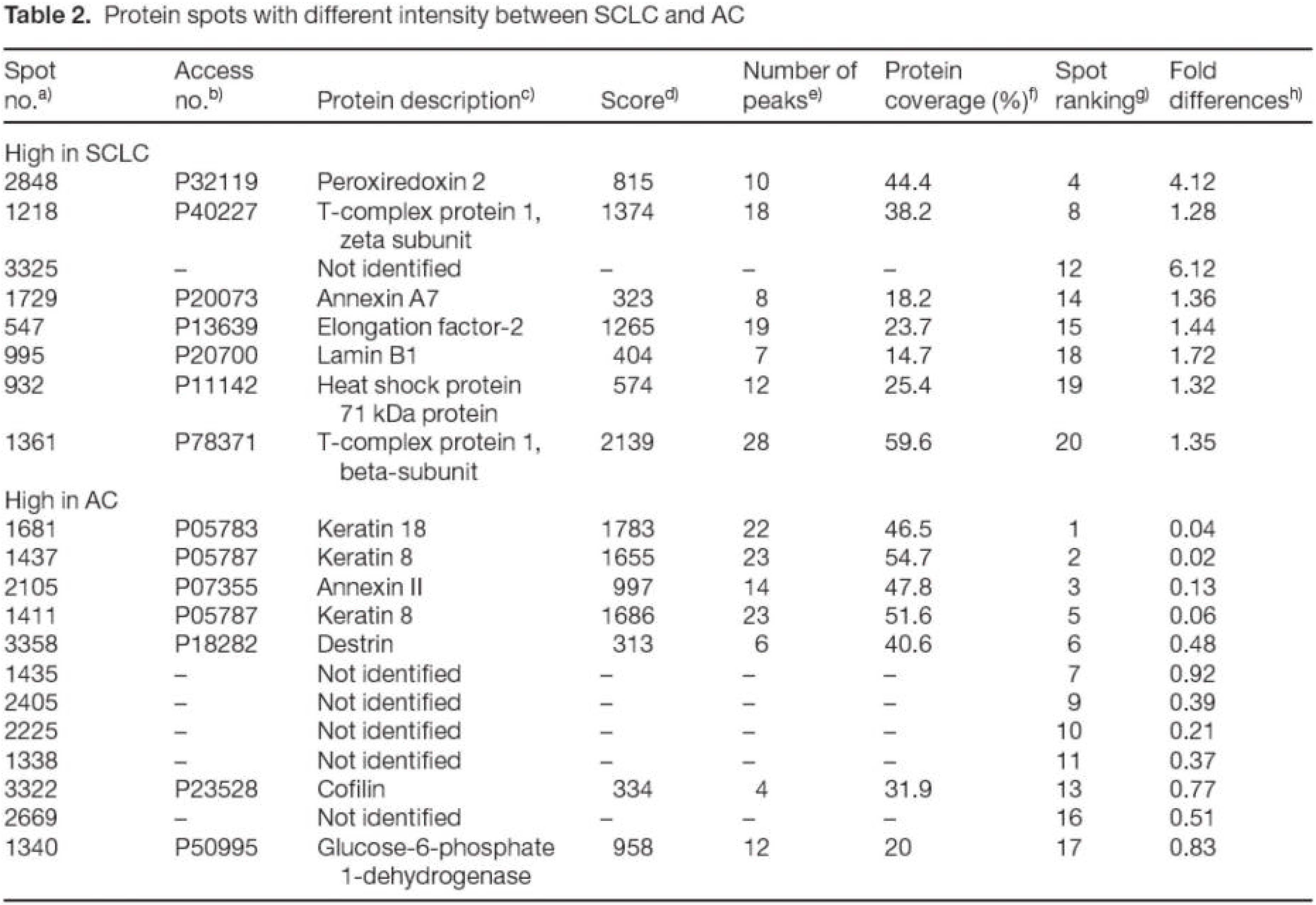


Table 2. Protein spots with different intensity between SCLC and AC

Spot no.[a)]	Access no.[b)]	Protein description[c)]	Score[d)]	Number of peaks[e)]	Protein coverage (%)[f)]	Spot ranking[g)]	Fold differences[h)]
High in SCLC							
2848	P32119	Peroxiredoxin 2	815	10	44.4	4	4.12
1218	P40227	T-complex protein 1, zeta subunit	1374	18	38.2	8	1.28
3325	–	Not identified	–	–	–	12	6.12
1729	P20073	Annexin A7	323	8	18.2	14	1.36
547	P13639	Elongation factor-2	1265	19	23.7	15	1.44
995	P20700	Lamin B1	404	7	14.7	18	1.72
932	P11142	Heat shock protein 71 kDa protein	574	12	25.4	19	1.32
1361	P78371	T-complex protein 1, beta-subunit	2139	28	59.6	20	1.35
High in AC							
1681	P05783	Keratin 18	1783	22	46.5	1	0.04
1437	P05787	Keratin 8	1655	23	54.7	2	0.02
2105	P07355	Annexin II	997	14	47.8	3	0.13
1411	P05787	Keratin 8	1686	23	51.6	5	0.06
3358	P18282	Destrin	313	6	40.6	6	0.48
1435	–	Not identified	–	–	–	7	0.92
2405	–	Not identified	–	–	–	9	0.39
2225	–	Not identified	–	–	–	10	0.21
1338	–	Not identified	–	–	–	11	0.37
3322	P23528	Cofilin	334	4	31.9	13	0.77
2669	–	Not identified	–	–	–	16	0.51
1340	P50995	Glucose-6-phosphate 1-dehydrogenase	958	12	20	17	0.83

Seike et al 2004, *Proteomics* 4: 2776.

Differential expression of proteins in cellular senescence

Preliminary work has been done to characterize proteins potentially involved in the human cell aging process

Experimental setup

- human fibroblast cell lines
- senescence induced using *RAS* oncogene
- control and induced cells compared using DIGE
- differentially expressed proteins identified by mass spec

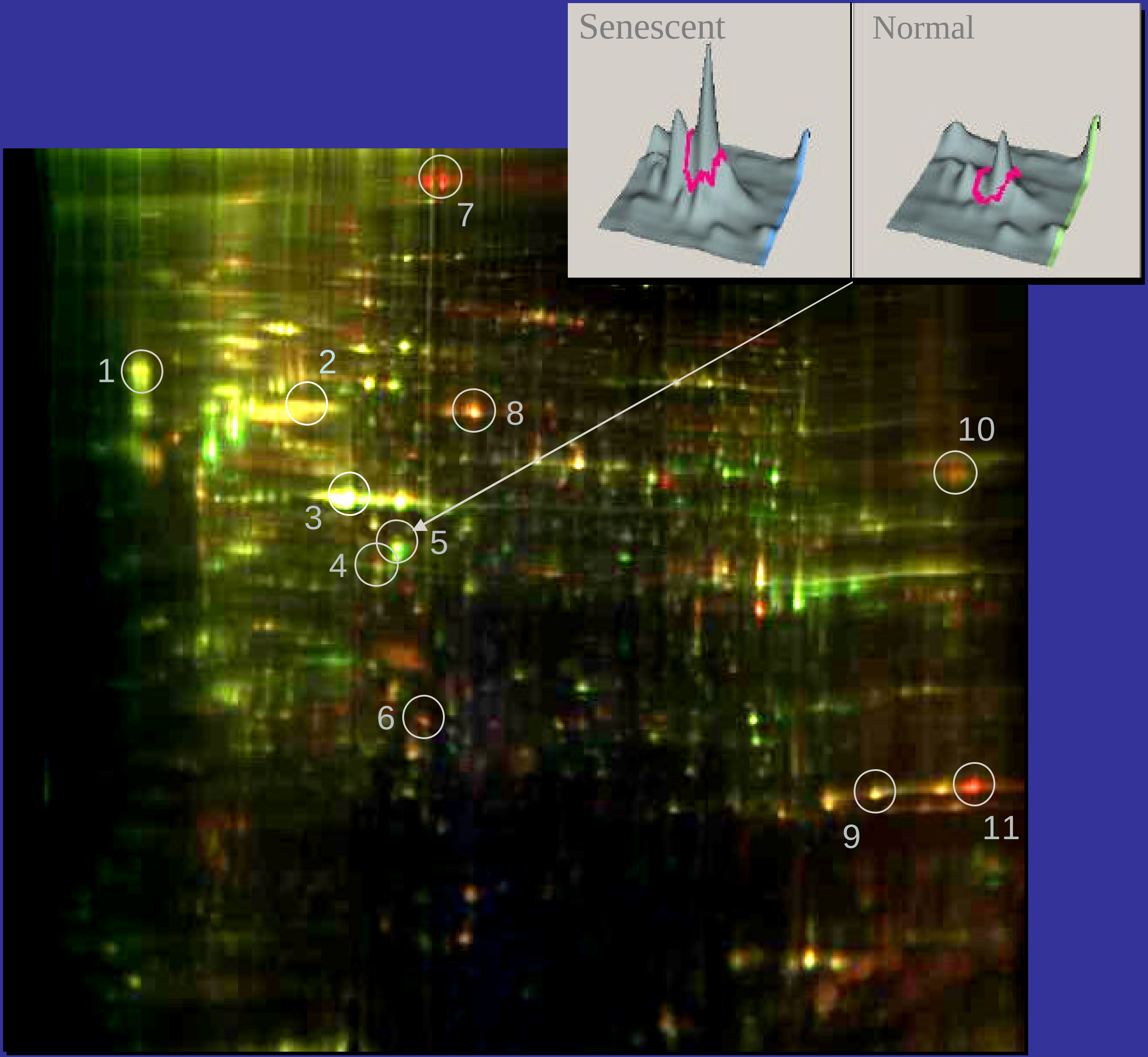
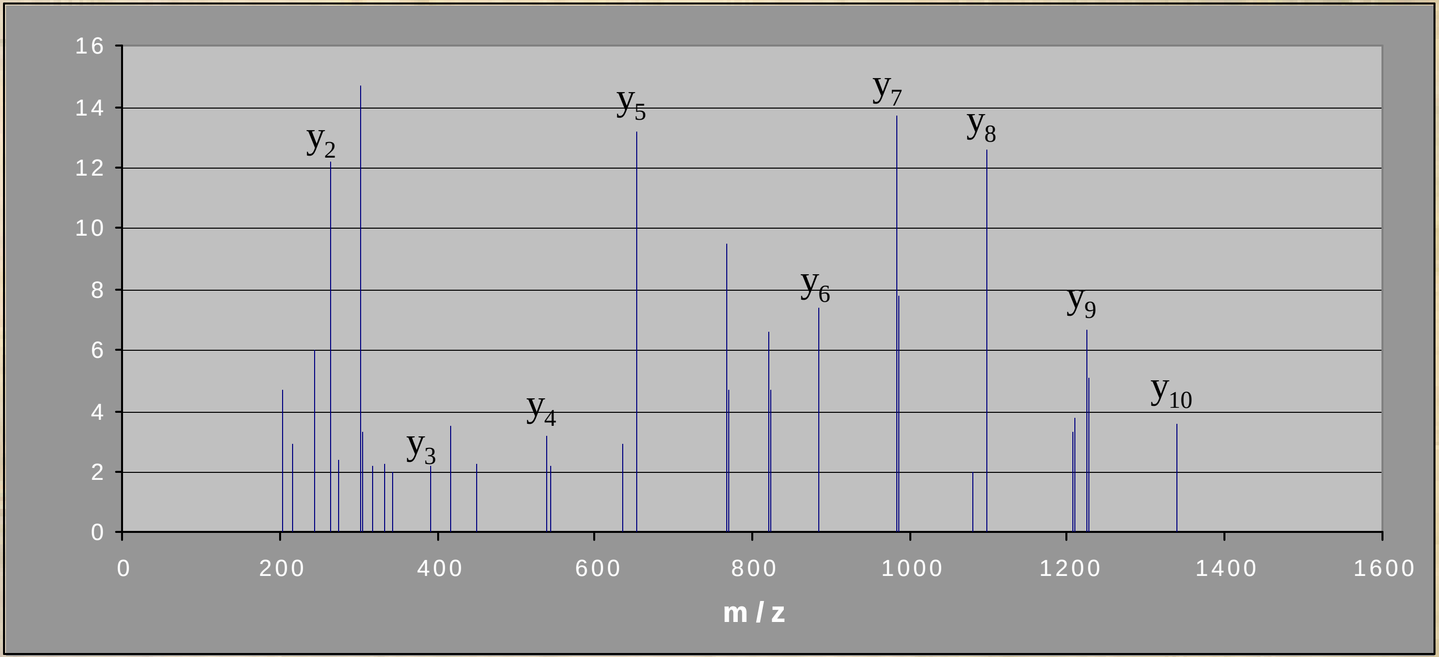


Image of a gel bearing both CyDye-labeled normal (red) and induced-senescent (green) fibroblast proteins. A volume map is shown for spot 5.

Fragmentation spectrum of a peptide obtained from a tryptic digest of disulfide isomerase ER-60. Interpretation of this spectrum yielded the sequence LELTDDNFESR.



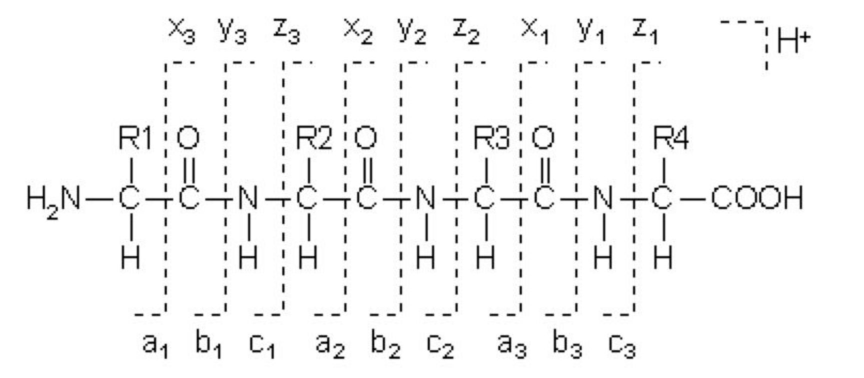
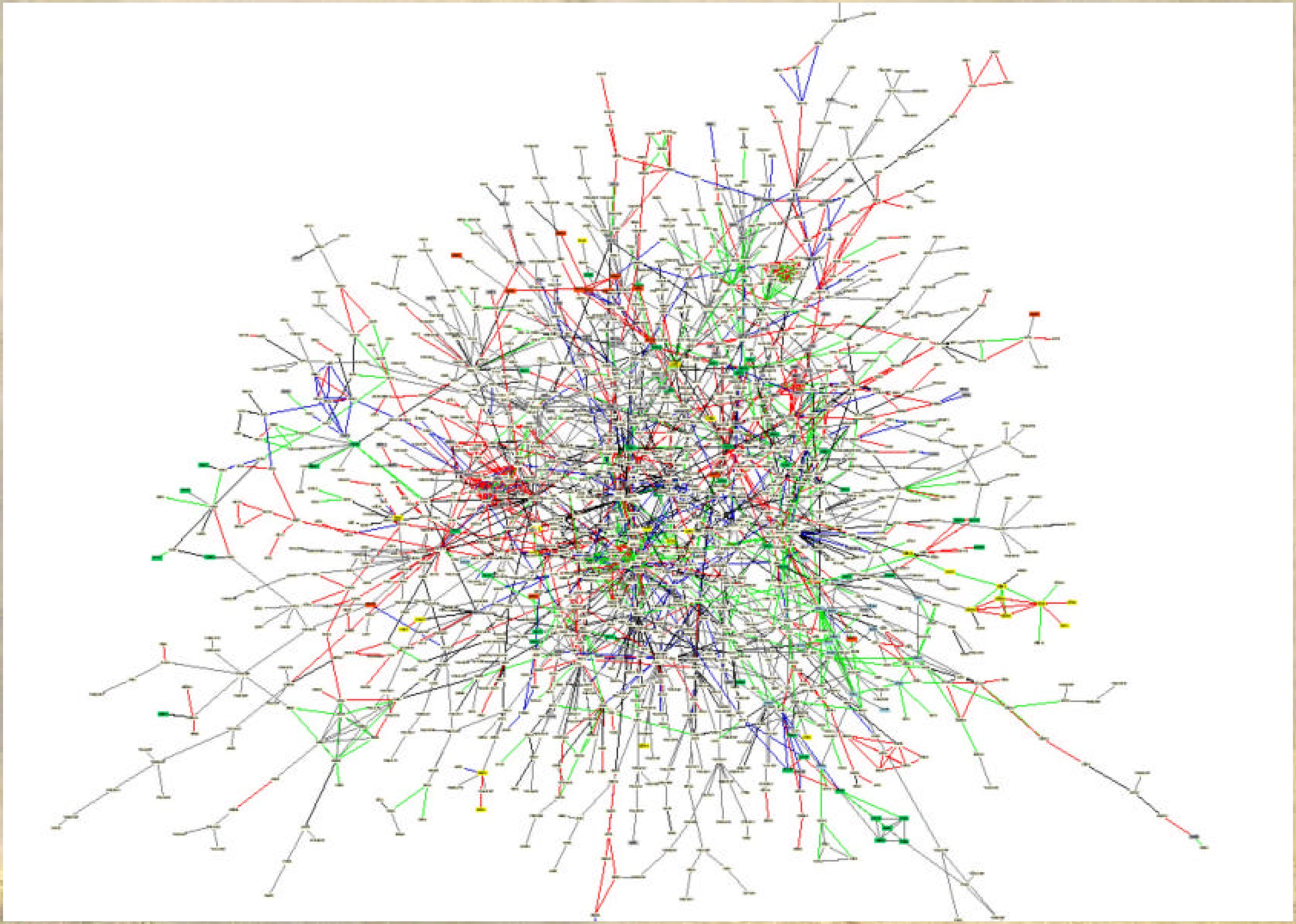
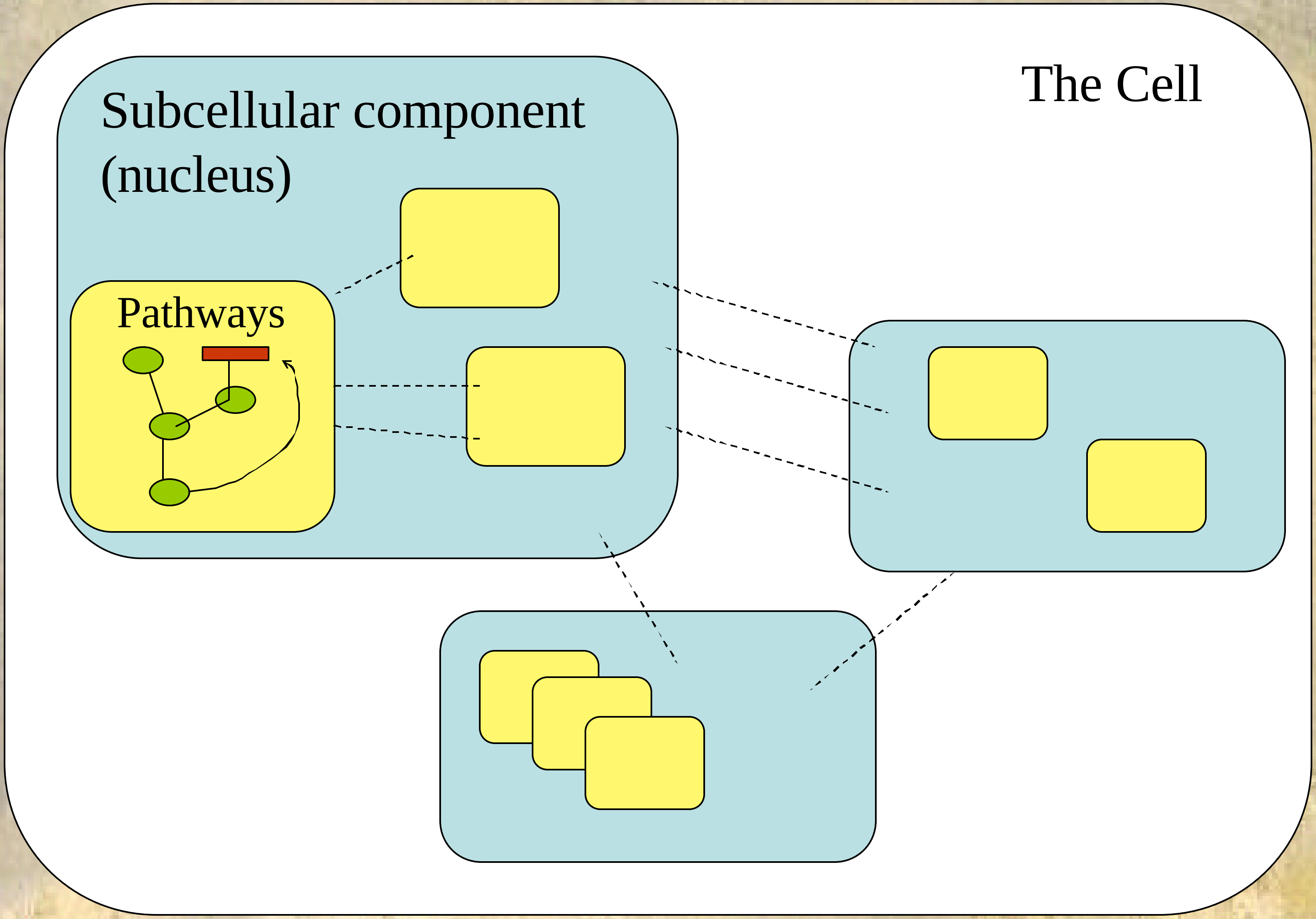

x3 y3 z3 x2 y2 z2 x1 y1 z1
H+
R1 O R2 O R3 O R4
H2N—C—C—N—C—C—N—C—C—N—C—COOH
H H H H H H H
a1 b1 c1 a2 b2 c2 a3 b3 c3

Figure Label	Accession Number	Protein	Average Ratio Senescent: Normal
1	P27797	Calreticulin (CRTC) - lectin, calcium binding chaperone	1.01
2	multiple	Vimentin + α-tubulin + tubulin α-chain 1 - filament protein	1.15
3	***P02571***	***g-actin (possibly with b-actin)*** ***- cytoskeleton***	***2.51***
4	P05218	tubulin β-5 chain - microtubules	0.88
5	***Q8WU19***	***K-ALPHA-1 protein*** ***- microtubules, cytoskeleton***	***2.70***
6	P04792	Heat shock protein 27 kDa - stress response, actin organization	0.62
7	P02452	Collagen α 1(I) chain precursor - fibrillar forming, structural protein	0.06
8	P30101	Probable protein disulfide isomerase ER-60 - protein folding	0.50
9	Q06830	Peroxiredoxin precursor - redox regulation, signal cascades via H2O2(?)	0.62
10	P29043	Heat shock protein Hsp 47 precursor - collagen binding	0.35
11	*Q01995*	*Transgelin **contradicts RNA expression* *- linked to replicative senescence*	*0.15*

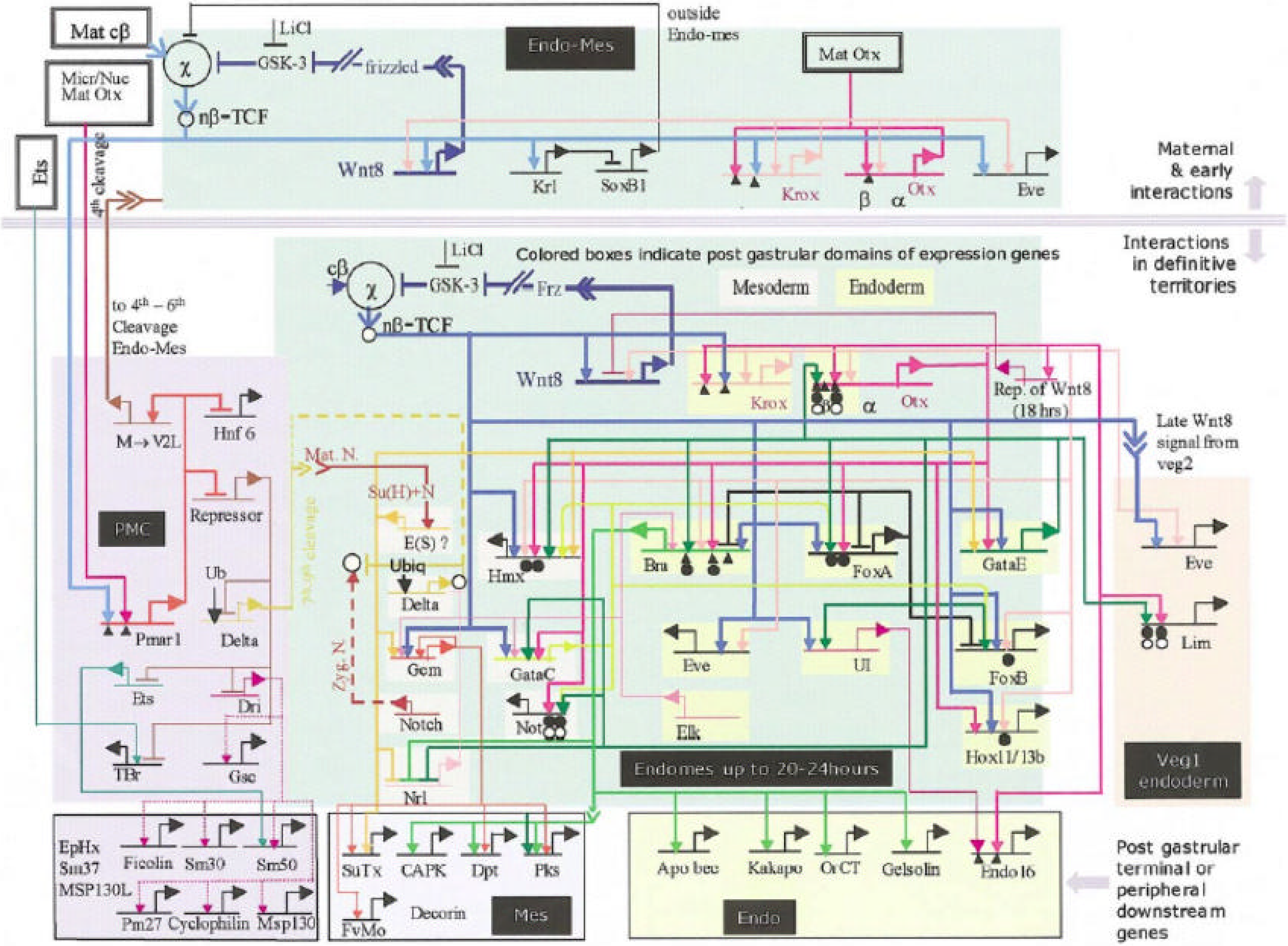
A messy way to look at interaction data:



A better way - as a series of nested, interacting organizational units:



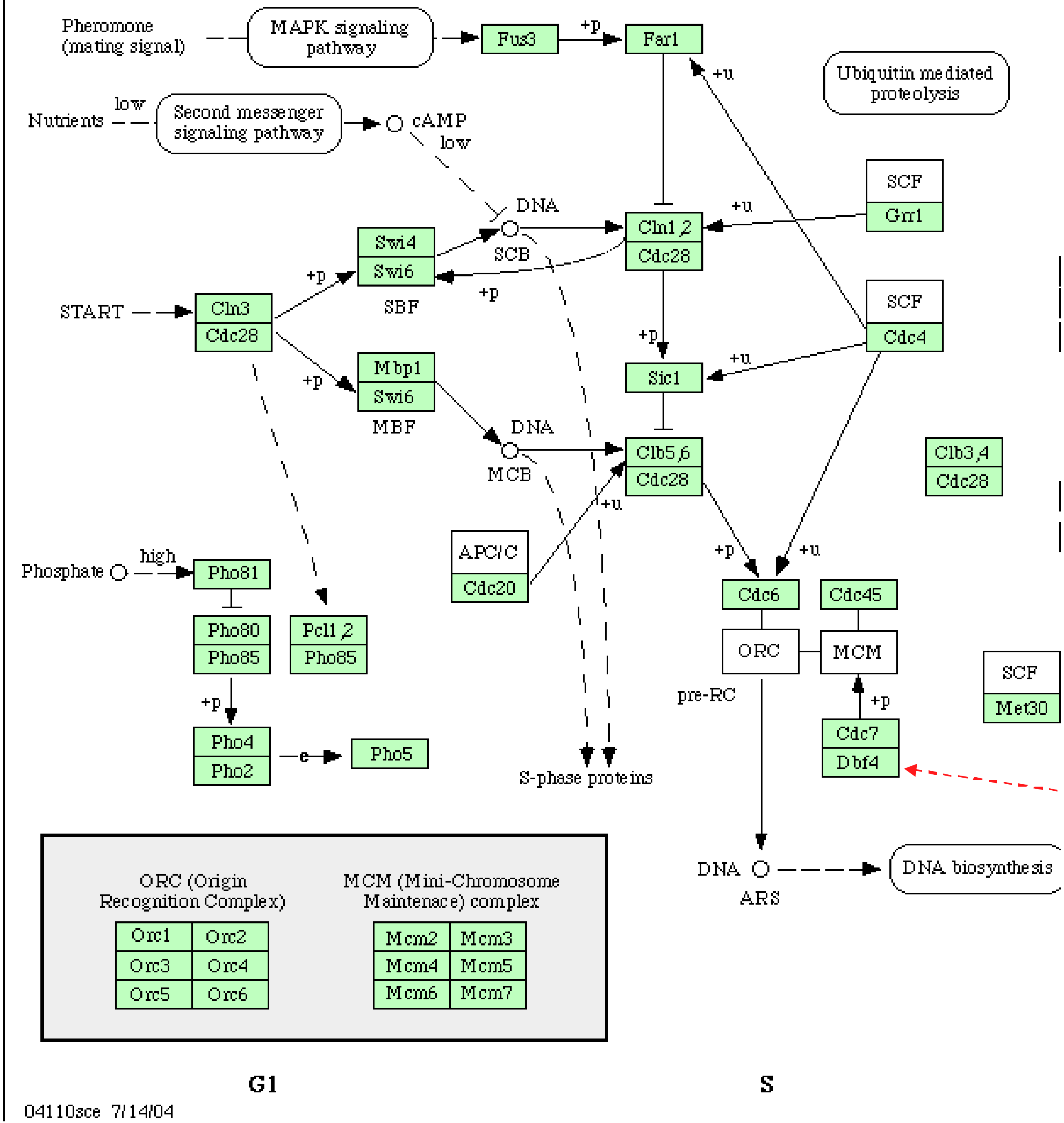
Preliminary Regulatory Network in the Sea urchin for endomesodermal development

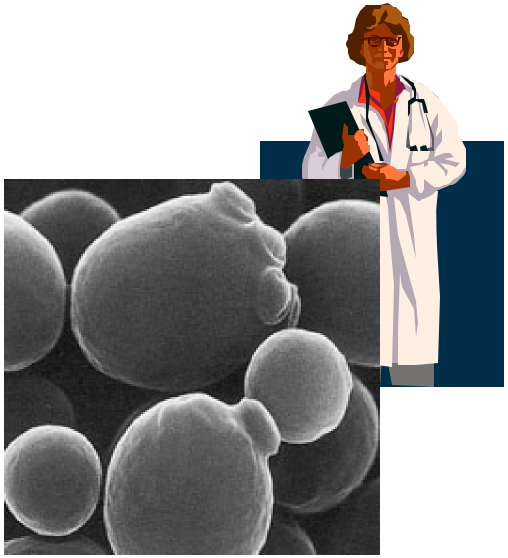


Davidson et al, (2003) Dev. Biol.

Cell cycle modeling in yeast

- focus is the genes, proteins and interactions involved in the initiation of DNA replication (G1 --> S transition)
- DIGE used to isolate, quantify and identify differentially expressed proteins and protein isoforms
- protein levels and interactions modeled using differential equation model (Chen/Tyson/Novak)
- cell cycle model used to predict effect of perturbations on system
- perturbations to system created by genetic manipulation
- changes in protein level/type used to refine model



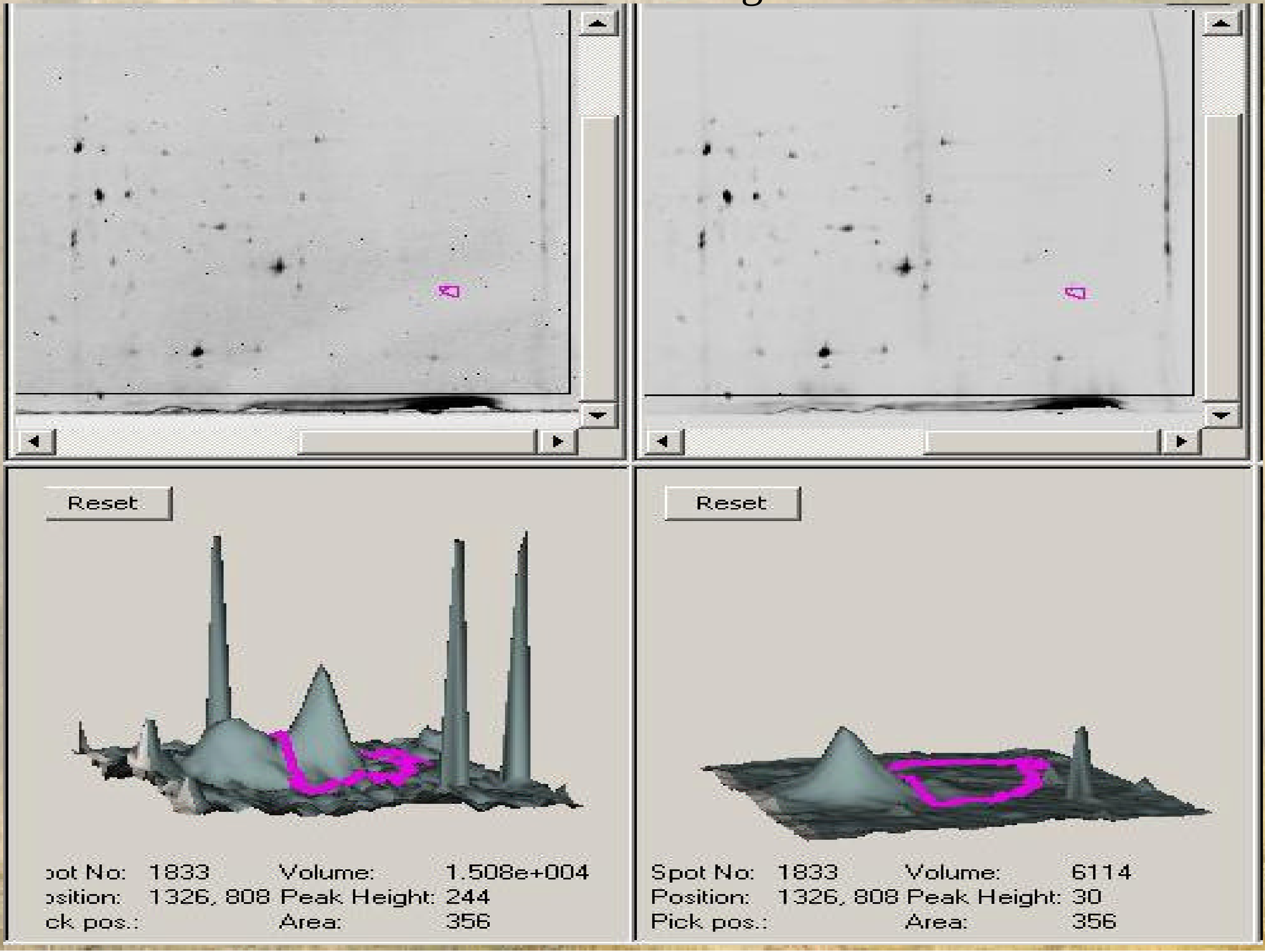


?

Hct1

http://www.genome.jp/kegg/pathway.html

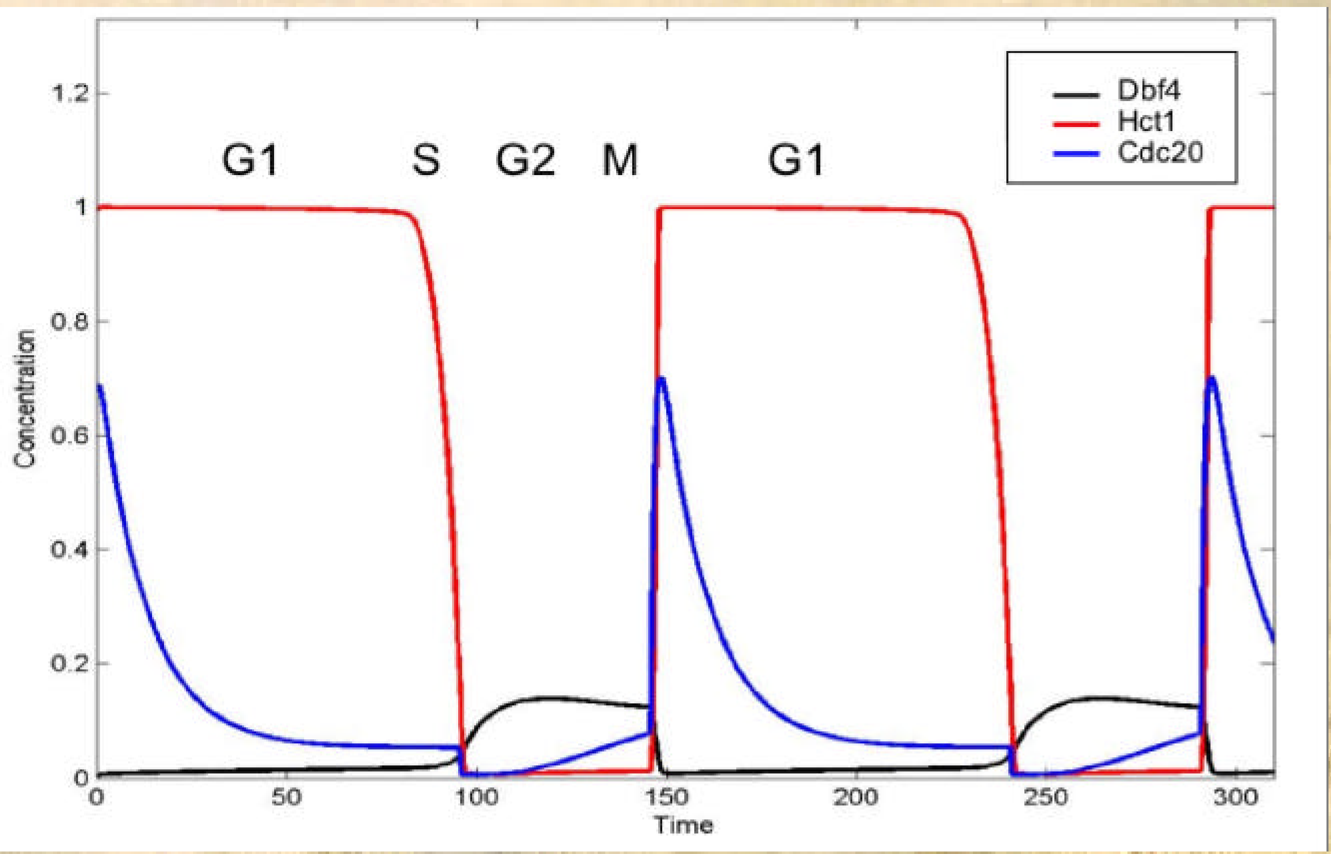
Tentative identification of hct1 on DIGE gels - wt vs. Hct1 knockout



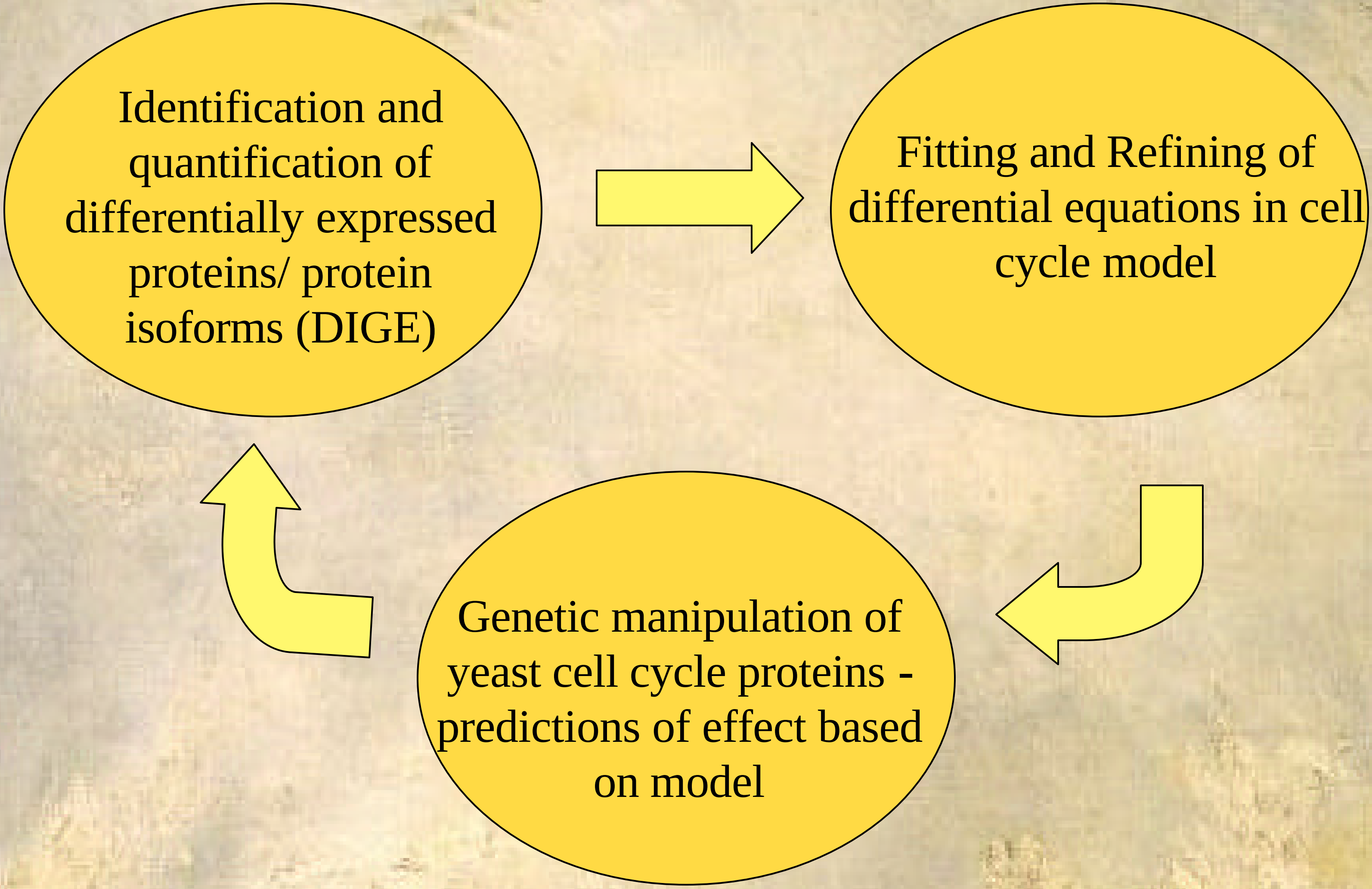
```
pot No:    1833        Volume:        1.508e+004
osition:   1326, 808   Peak Height:   244
ck pos.:               Area:          356
```

```
Spot No:   1833        Volume:        6114
Position:  1326, 808   Peak Height:   30
Pick pos.:             Area:          356
```

- model of three selected cell cycle proteins: Dbf4, Cdc20, and Hct1
- Hct1 and Cdc20 are from Chen/Tyson/Novak model
- Dbf4 was modeled as dependent on Hct1



REFINEMENT OF CELL CYCLE MODEL



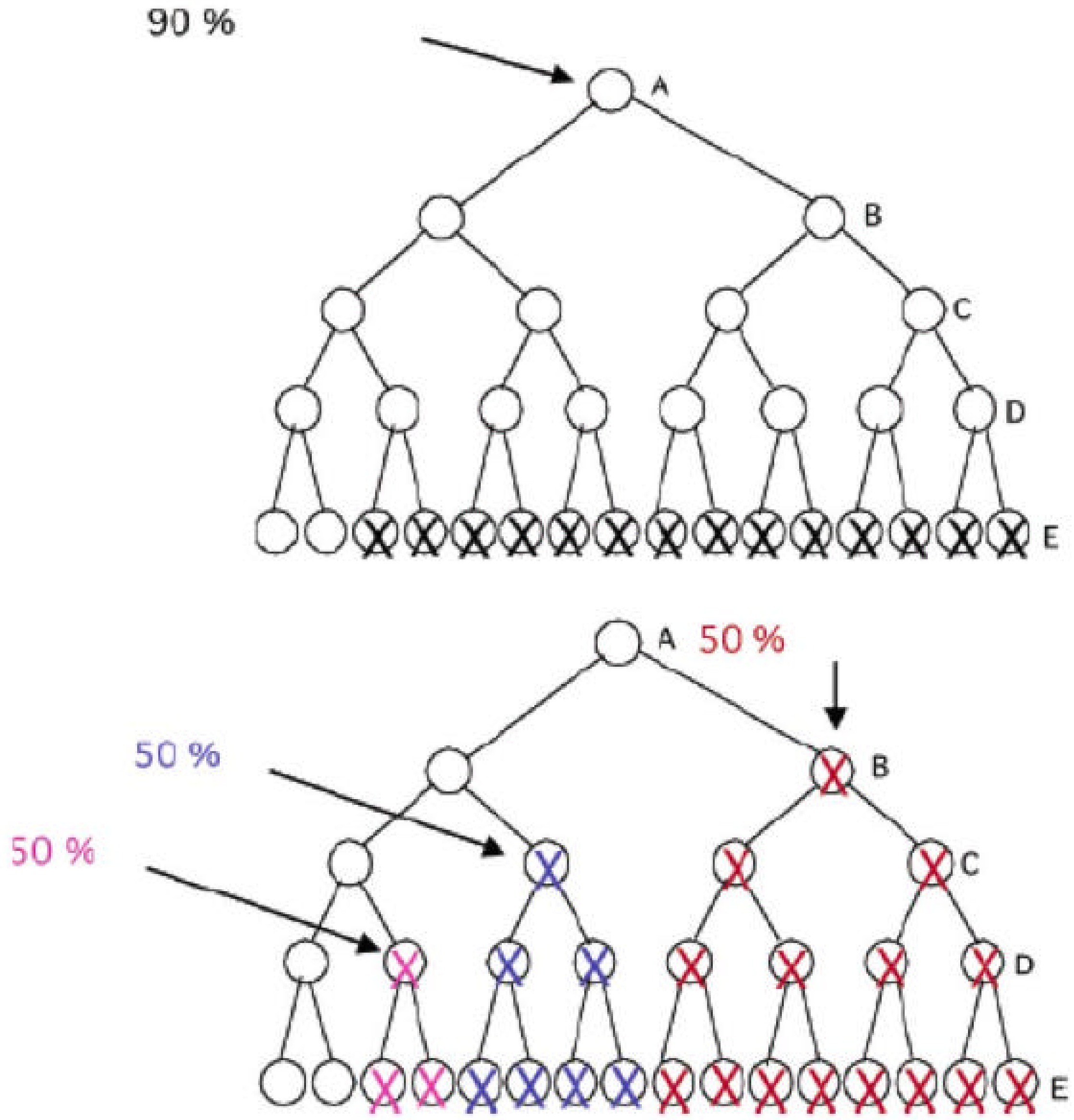
Identification and quantification of differentially expressed proteins/ protein isoforms (DIGE)

Fitting and Refining of differential equations in cell cycle model

Genetic manipulation of yeast cell cycle proteins - predictions of effect based on model

Eventual goal - an accurate description of the gene/protein network involved in the initiation of DNA replication

Combinatorial Therapeutics: reduced toxicity, increased efficacy



Conclusions

- bioinformatics technologies (microarrays, 2D-DIGE, mass spectrometry, ...) have wide potential application in diagnostics and treatment
- still in a relatively early phase of development- not practical for medical applications yet
- potential for identifying novel drug targets or multiple interacting targets
- remains to be seen whether utility of these methods outweigh instrumentation costs and required expertise

Acknowledgements

Thanks to...

Collaborators:

Gerardo Ferbeyre

Bernard Duncker
Darah Christie

Brian Ingalls

Bernie Glick

Lab members:

Andrea Spires

Zhenyu Cheng

Carl White

Esther Mak

Gabriel Renaud

Andrew Doxey

Rasmus Jostrup