Mass Spec and MicroArrays

Applications in Proteomics and Systems Biology

Proteomics: *From Technology Development to Biomarker Applications*

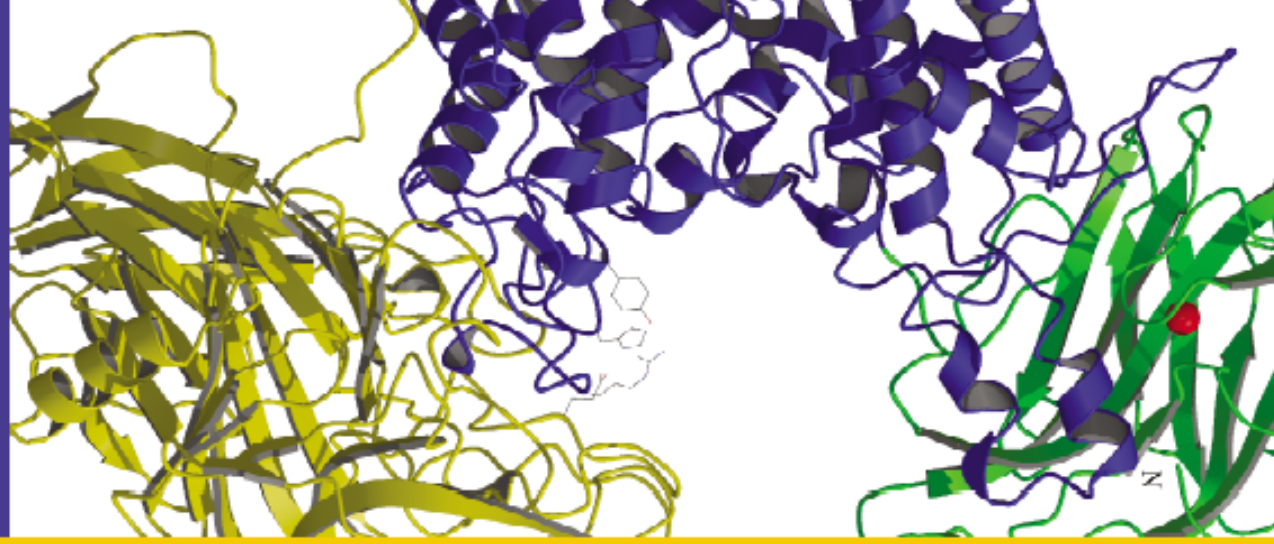
CH370 - Hackert

HUPO

Human Proteome Organisation

Long Beach Convention Center, California, USA

Saturday October 28th through Wednesday November 1st, 2006



HUPO 5TH ANNUAL WORLD CONGRESS, LONG BEACH 2006

TRANSLATING PROTEOMICS FROM BENCH TO BEDSIDE

oteomics Education, an Important Challenge for the Scientific mmunity: Report on the Activities of the EuPA Education mmittee

PA Tutorial Program (preliminary draft)
ndamentals and Core Techniques

in Chemistry	
	Amino acid chemistry/functionality
	PTM natural chemical/enzymatic modifications
	PTM un-natural chemical/enzymatic modifications
	Protein function families: E.C: GO classification
	X-ray principles
	NMR principles
	Protein substructure principles
	Protein structure families
	Membrane protein structure/function
	Extracellular protein structure/function
in-protein Interaction	
	Protein complex isolation & examples
	MS-TAP approach to complexes
	Two-hybrid approach
	Biacore, microcalorimetry & CD, FT, ...
RNA Techniques	
	DNA cloning & sequencing
	RNA structure determination
	Microarray formats
	SAGE
	SNP, methylation, CGH analysis
ration Science	
	Affinity chromatography
	Free flow electrophoresis
	CZE
	Centrifugation
	HPLC
	2D-PAGE
in Expression	
	Antibody generation and use
	Phage display
	Protein arrays
	Tissue arrays

European Proteomics Association (EuPA)

MS Basics	
	MALDI ionisation
	ESI ionisation
	TOF
	Quads
	Ion-trap, linear & 3D
	FT-ICR, Orbitrap
	Detectors
	Scan modes
Metabolomics	
	GC-MS approaches & derivatisation che
	ESI-MS approaches & derivatisation che
	NMR approaches
	Pathway analysis & modelling EcoCYC
Applied Technologies	
	Microfluidics
	Automation
	Fluorescent labeling, DNA sequencing, r
Bioinformatics/Systems Biology	
	Sequence homology searching
	Protein id by MALDI
	Protein id by MS/MS
	ID verification principles, Prophet, *etc.*
	Array analysis
	Database structure
	Relevant stat applications
	Advanced data mining techniques

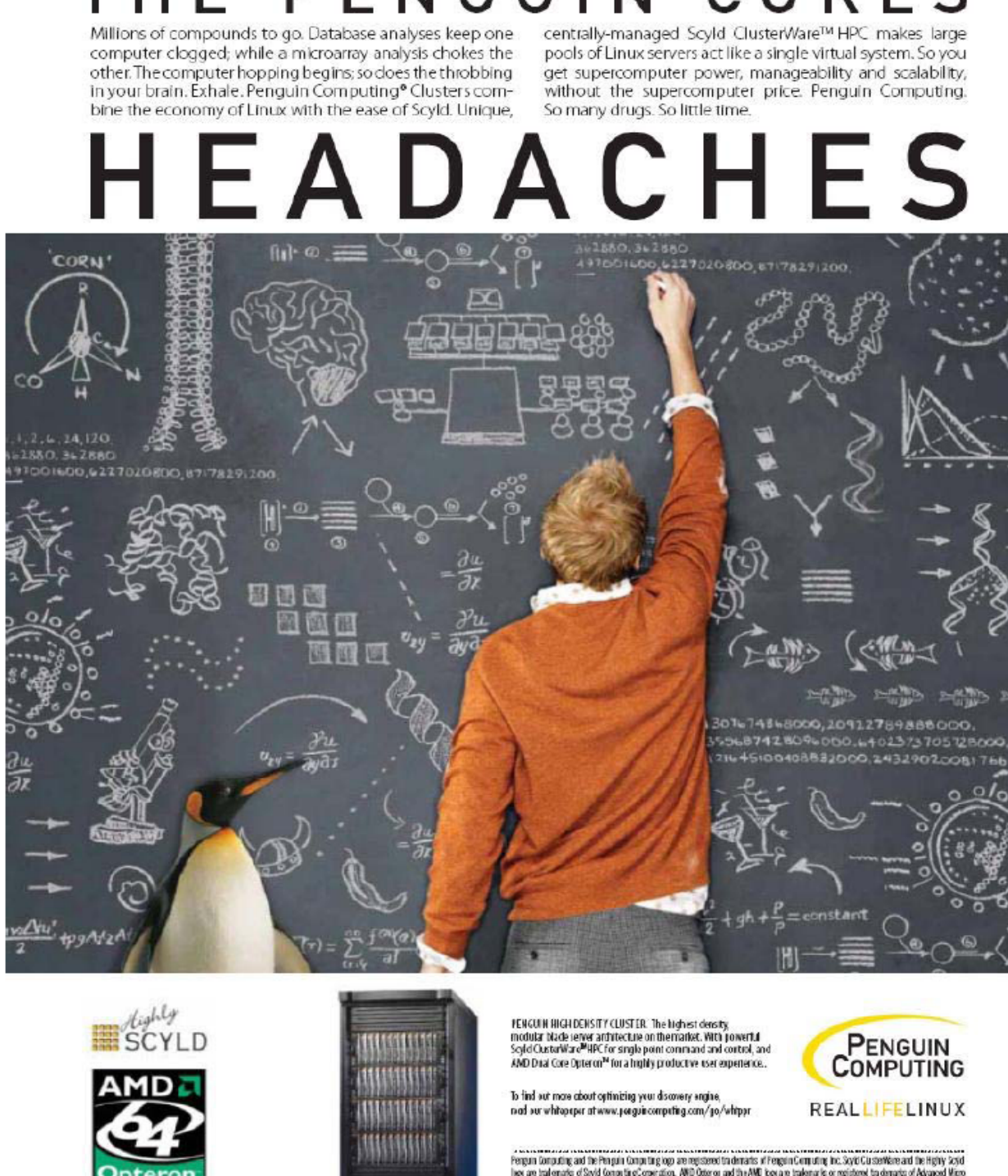
Genomics

Proteomics

Interactomics

Systems Biology –

None of these fields of research would be possible without **Bioinformatics,** which would not be possible with lots of **computing power!**



Mass Spec and MicroArrays / Applications

enome – the genome of an organism is its whole hereditary informati
coded in its DNA (or, RNA for some viruses) and includes both the
ding (genes) and non-coding sequences of the DNA.

oteome – Proteomics is often considered the next step in the study o
logical systems, after genomics. It is much more complicated than
nomics, mostly because while an organism's genome is rather
nstant, a proteome differs from cell to cell and constantly changes
ough its biochemical interactions with the genome and the
vironment.

teractome – whole set of molecular interactions in cells, in the conte
proteomics, it refers to protein-protein interaction network(PPI), or
otein network (PN).

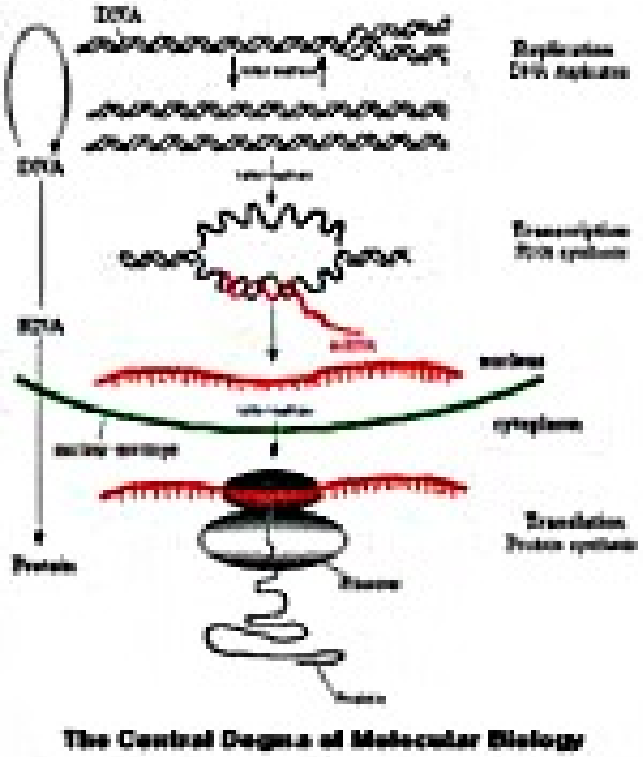
stems Biology - seeks to understand how biological systems functio
studying the relationships and interactions between various parts of
logical system (e.g. metabolic pathways, organelles, cells,
ysiological systems, organisms etc.), it is hoped that eventually a

The Proteome

All an organism's cells carry the same Genome, and it is Static. Genomes do not describe function. They are a parts list.

Different cells express different proteins. The type and quantity of this expression changes.

The Proteome is Dynamic. It is the total of all proteins expressed by a particular *cell* at a given *time*, under specific *conditions*.



The Central Dogma of Molecular Biology

A Proteome cannot be studied the way a Genome is sequenced. There has to be a specific biological question behind an experiment. The questions may be either very broad or strictly defined.

NATURE | VOL 422 | 13 MARCH 2003 | www.nature.com/nature

Mass spectrometry-based proteomics

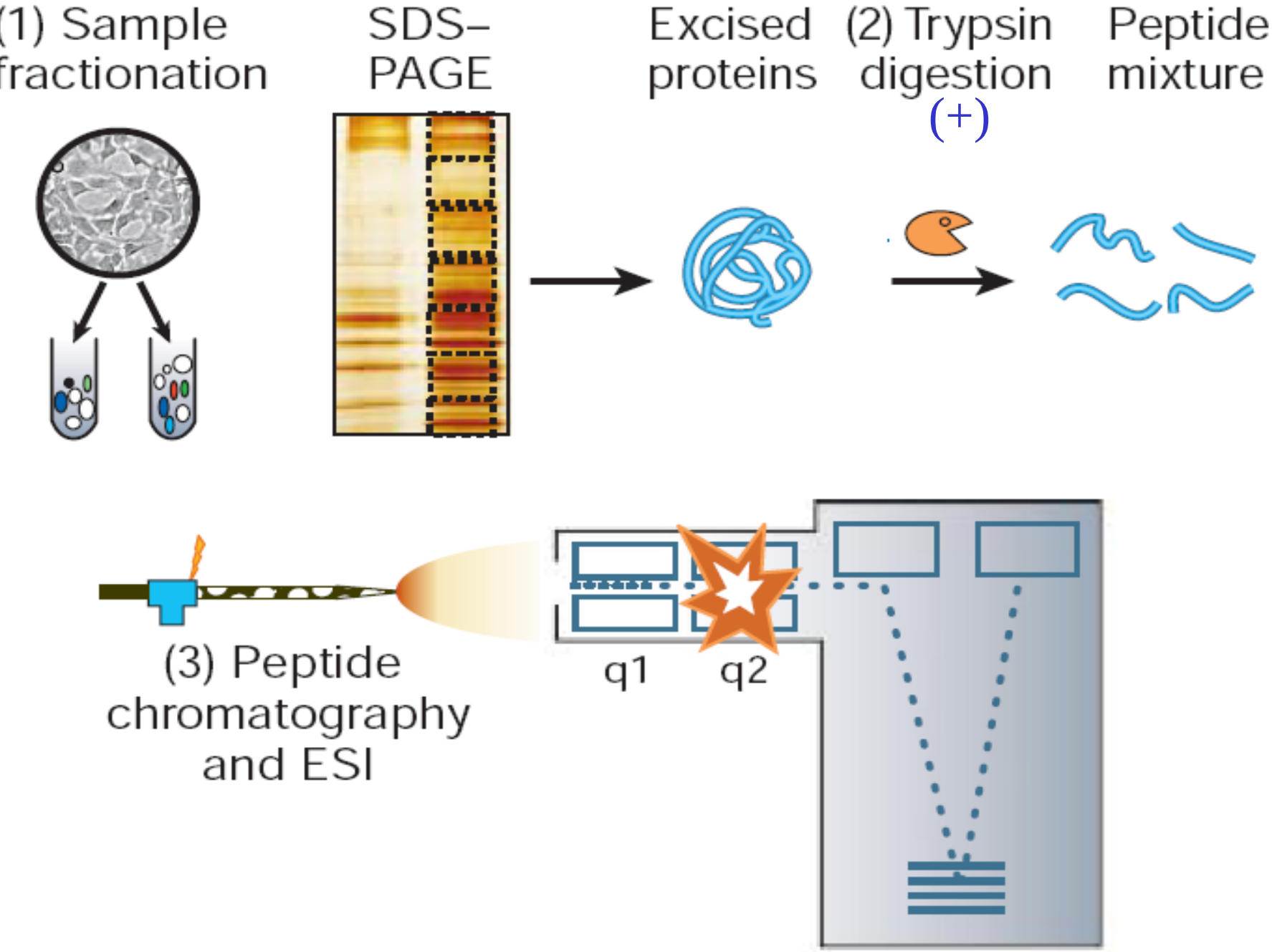
di Aebersold* & Matthias Mann†

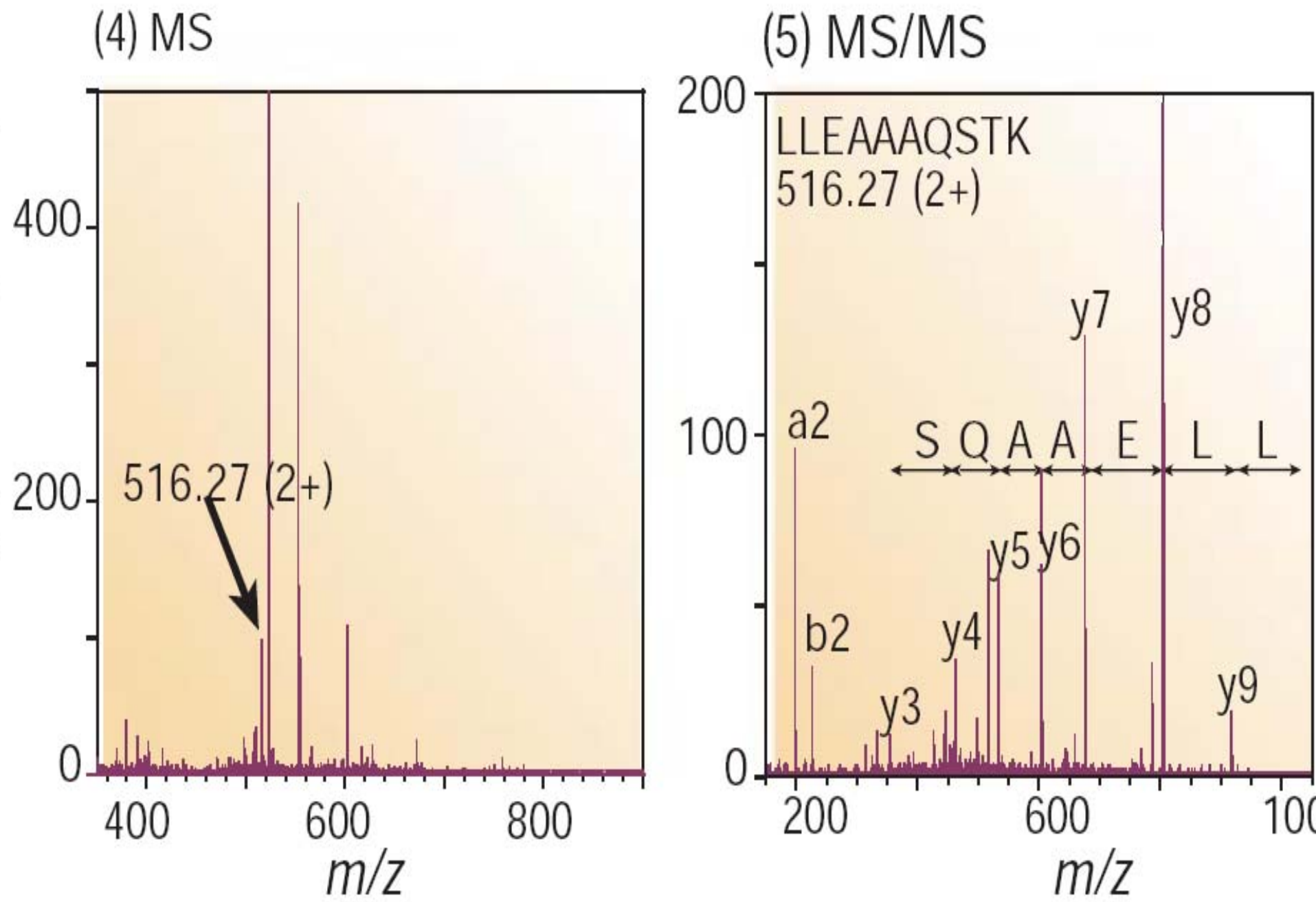
titute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103-8904, USA (e-mail: raebersold@systemsbiology.org)
nter for Experimental BioInformatics (CEBI), Department of Biochemistry and Molecular Biology, University of Southern Denmark,
npusvej 55, DK-5230 Odense M, Denmark (e-mail: mann@bmb.sdu.dk)

cent successes illustrate the role of mass spectrometry-based proteomics as an indispensable tool for lecular and cellular biology and for the emerging field of systems biology. These include the study of tein–protein interactions via affinity-based isolations on a small and proteome-wide scale, the mapping nerous organelles, the concurrent description of the malaria parasite genome and proteome, and the neration of quantitative protein profiles from diverse species. The ability of mass spectrometry to identi , increasingly, to precisely quantify thousands of proteins from complex samples can be expected to pact broadly on biology and medicine.

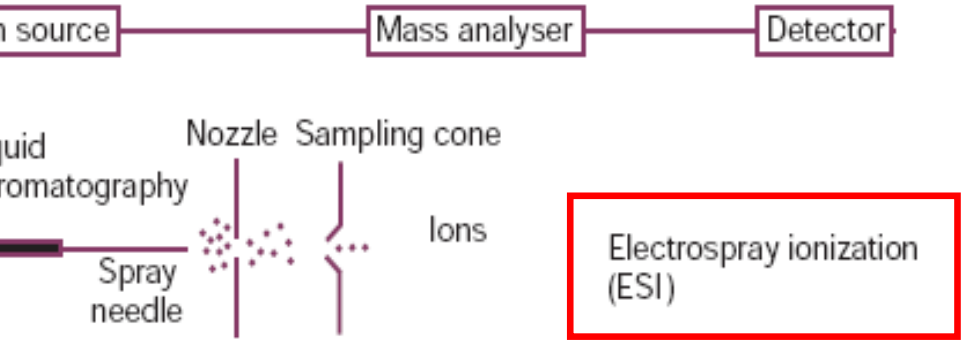
Note: HT Proteomics is restricted to those species where a sequence database exists!

Generic Mass Spectrometry-based Proteomics

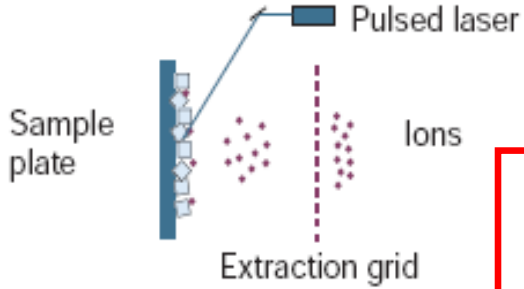



(4) MS
400
200
0
516.27 (2+)
400
600
800
m/z
(5) MS/MS
200
100
0
LLEAAAQSTK
516.27 (2+)
y7
y8
a2
S Q A A E L L
y5
y6
b2
y4
y3
y9
200
600
m/z

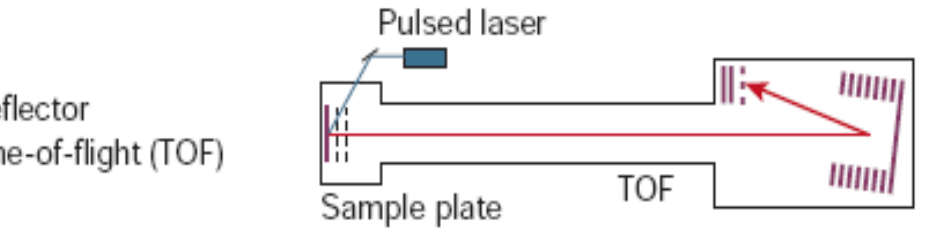
Common MSs Used in Proteomics Research

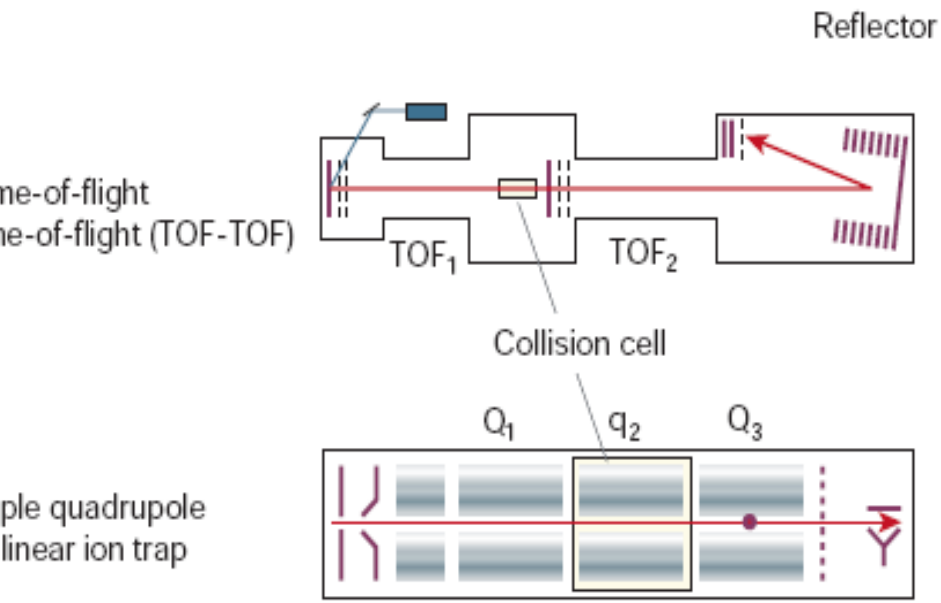


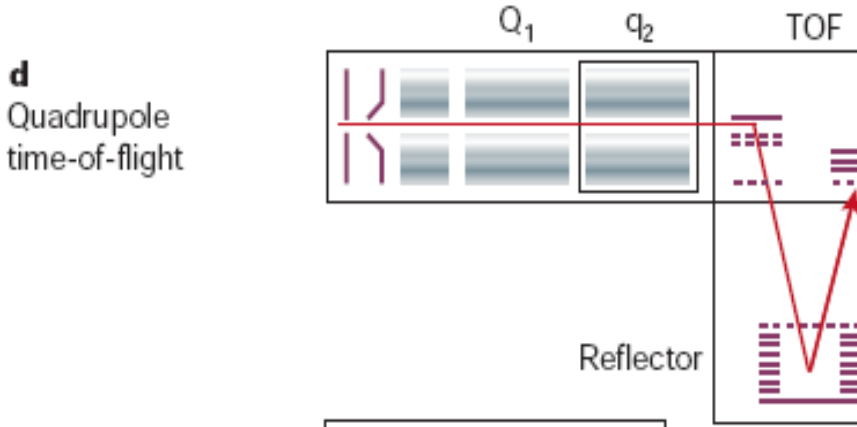
Electrospray ionization (ESI)

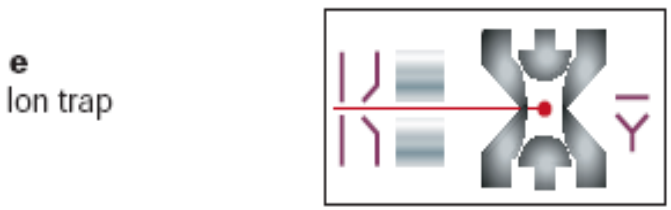


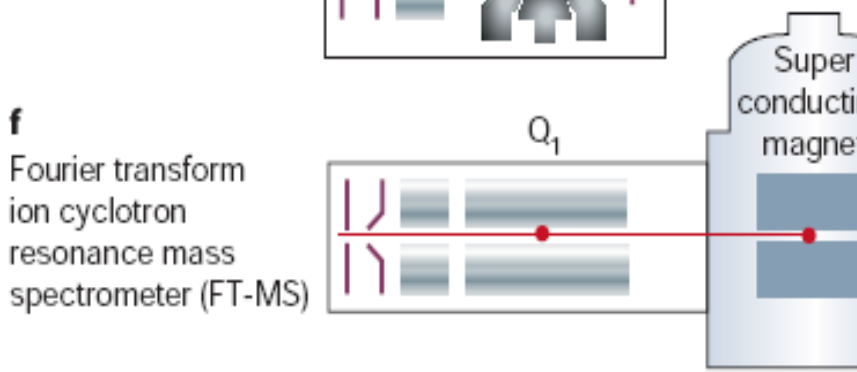
Matrix-assisted laser desorption/ionization (MALDI)



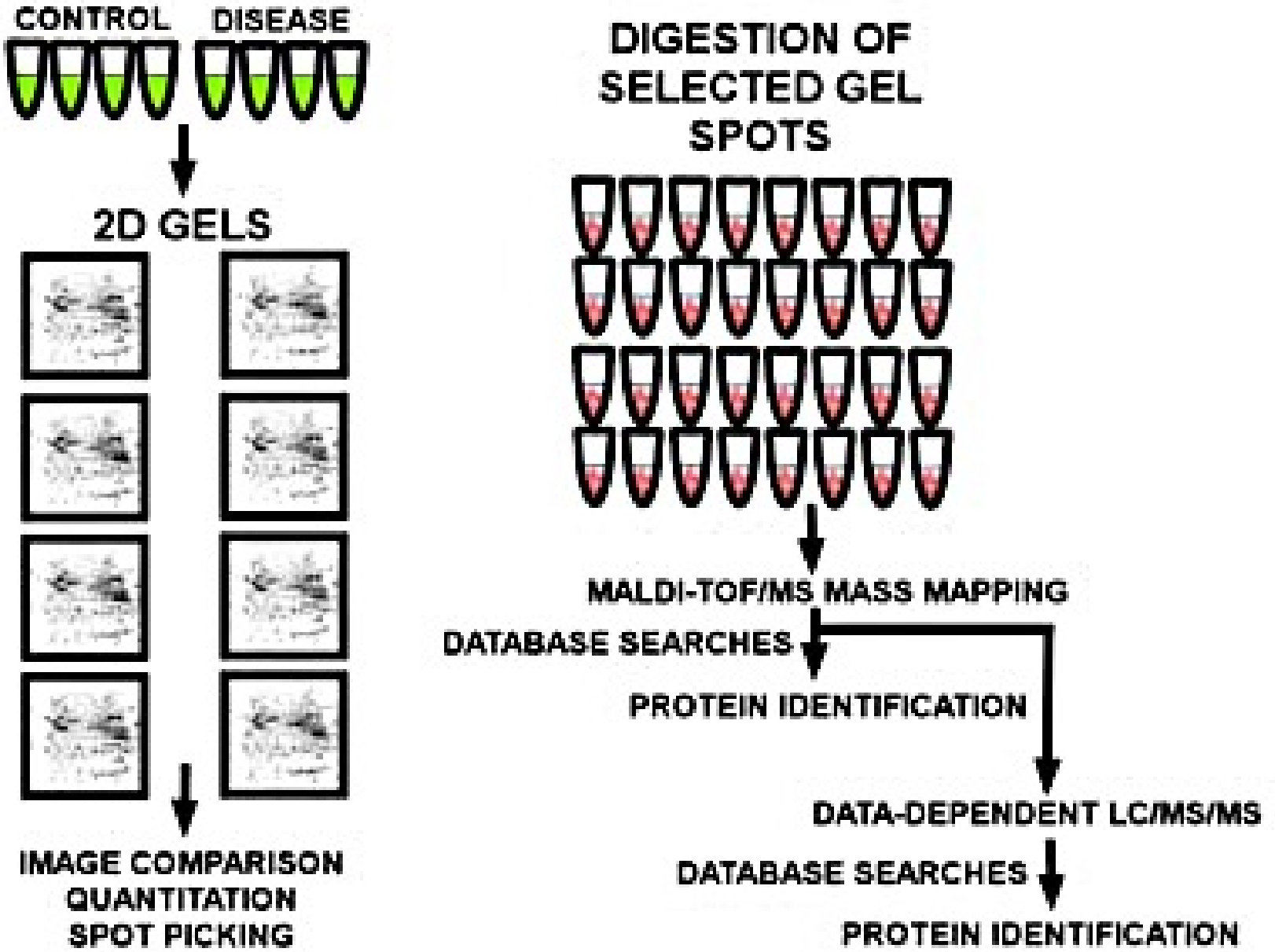








Differential Expression Proteomics



Two Dimensional Gel Electrophoresis

Isoelectric focusing is performed on precast gel strips using commercial instruments. Many pH ranges are available. Multiple strips can be run in parallel.

An immobilized pH gradient is created in a polyacrylamide gel strip by incorporating a gradient of acidic and basic buffering groups when the gel is cast.

Resolution is determined by the slope of the pH gradient and the field strength.

Loading capacity depends on gel size and thickness.

In 2D IEF/PAGE, the gel strip from IEF is loaded into a single large well.

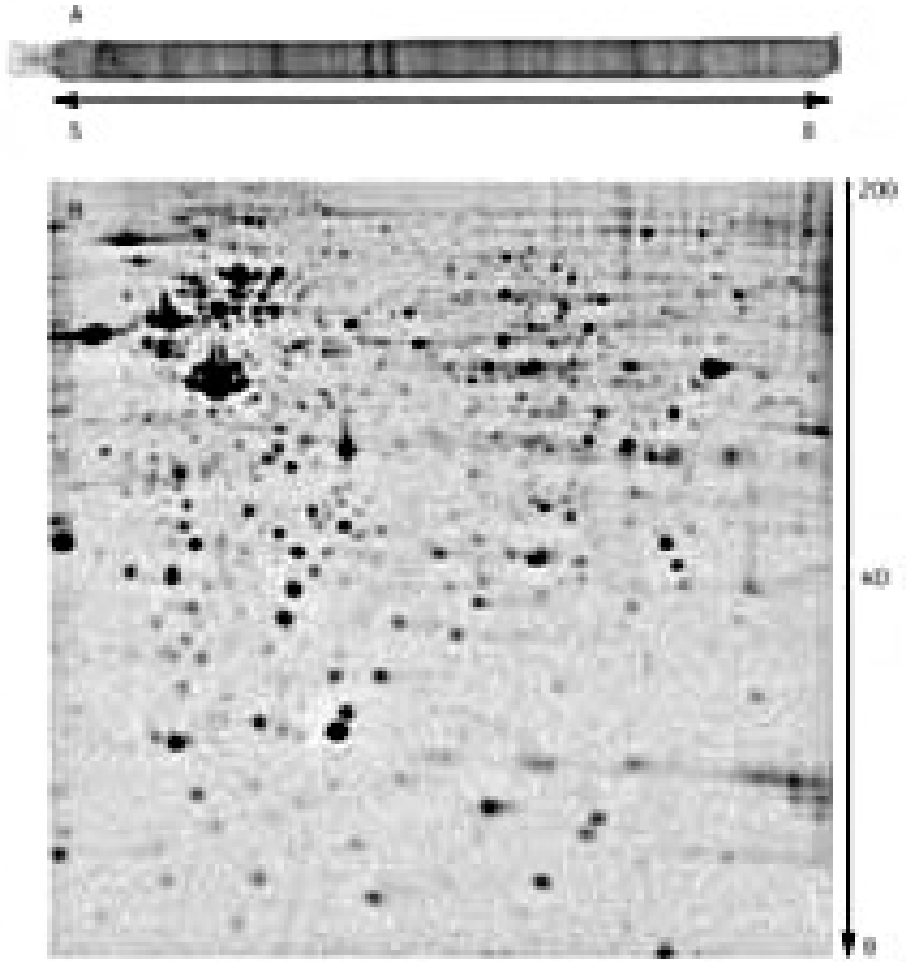


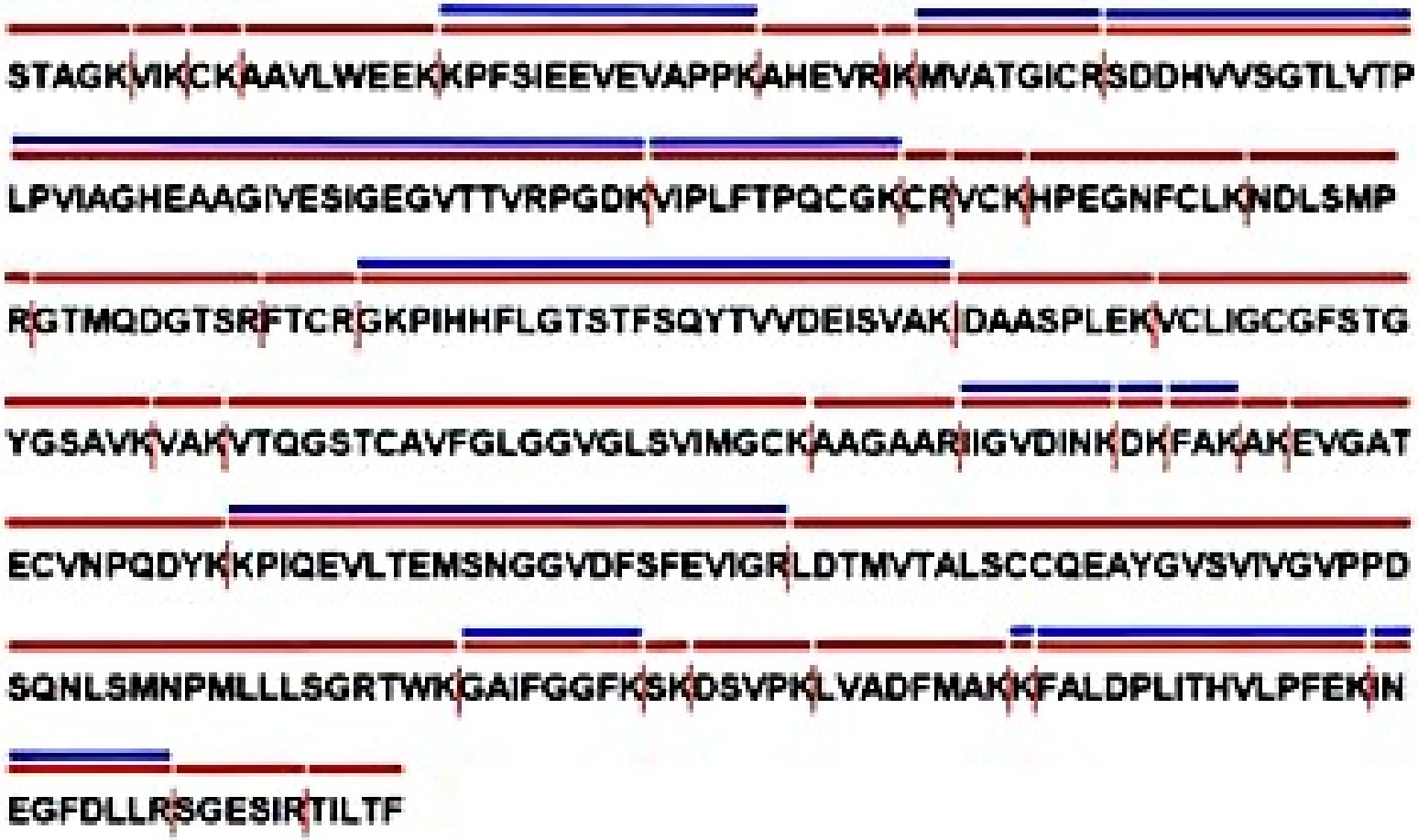
Fig. 1. Principle of 2-D electrophoresis. A, pre-B lymphoma cell extract (1 mg) was separated by IEF on a ReadyStrip pH 5–8 IPG strip, and stained with Bio-Safe™ Coomassie stain. B, Equilibrated strip was run in the second dimension by SDS-PAGE (12% acrylamide). The gel was stained with Coomassie Blue.

Figure from BioRad Product Literature

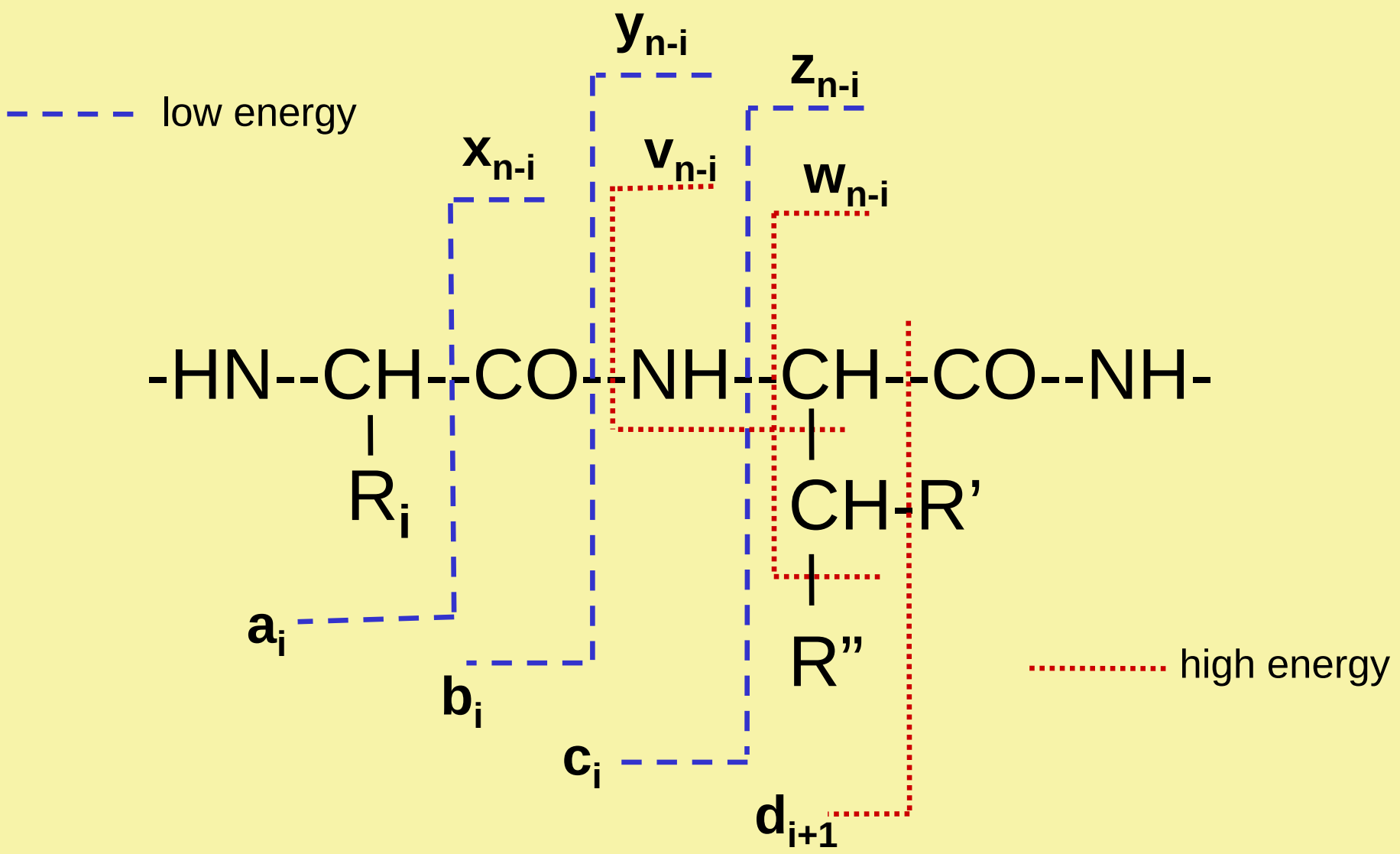
With the new genomic data bases of model species, such as *Escherichia coli, Saccharomyces cerevisae,* mouse, and human, the sequences of many/most proteins of biological interest will in principle be known, and the problem of characterizing a protein primary structure will be reduced to identifying it in the data base.

Within the past few years research groups have demonstrated how MS can be used for identification of proteins in sequence data bases. One approach is to **cleave the protein with a sequence-specific proteolytic enzyme**, **measure molecular weight** values for the resulting peptide mixture by mass spectrometry, and **search a sequence data base for proteins that should yield these values**. **Search algorithms** can utilize low resolution tandem mass spectra of selected peptides (<3 kDa) from the protein degradation. Yates and coworkers compared the MS/MS sequence data to the sequences predicted for each of the peptides that would be generated from each protein in the data base. **In the PEPTIDESEARCH sequence tag approach of Mann and Wilm, a partial sequence of 2–3 amino acids is assigned from the fragment mass differences in the MS/MS spectrum.** This partial sequence and its mass distance from each end of the peptide (based on the masses of the fragment and molecular ions) are used for the data base search. Often, **a single sequence tag**

Tryptic Digest of ADH: Expected Peptides vs. Those Detected



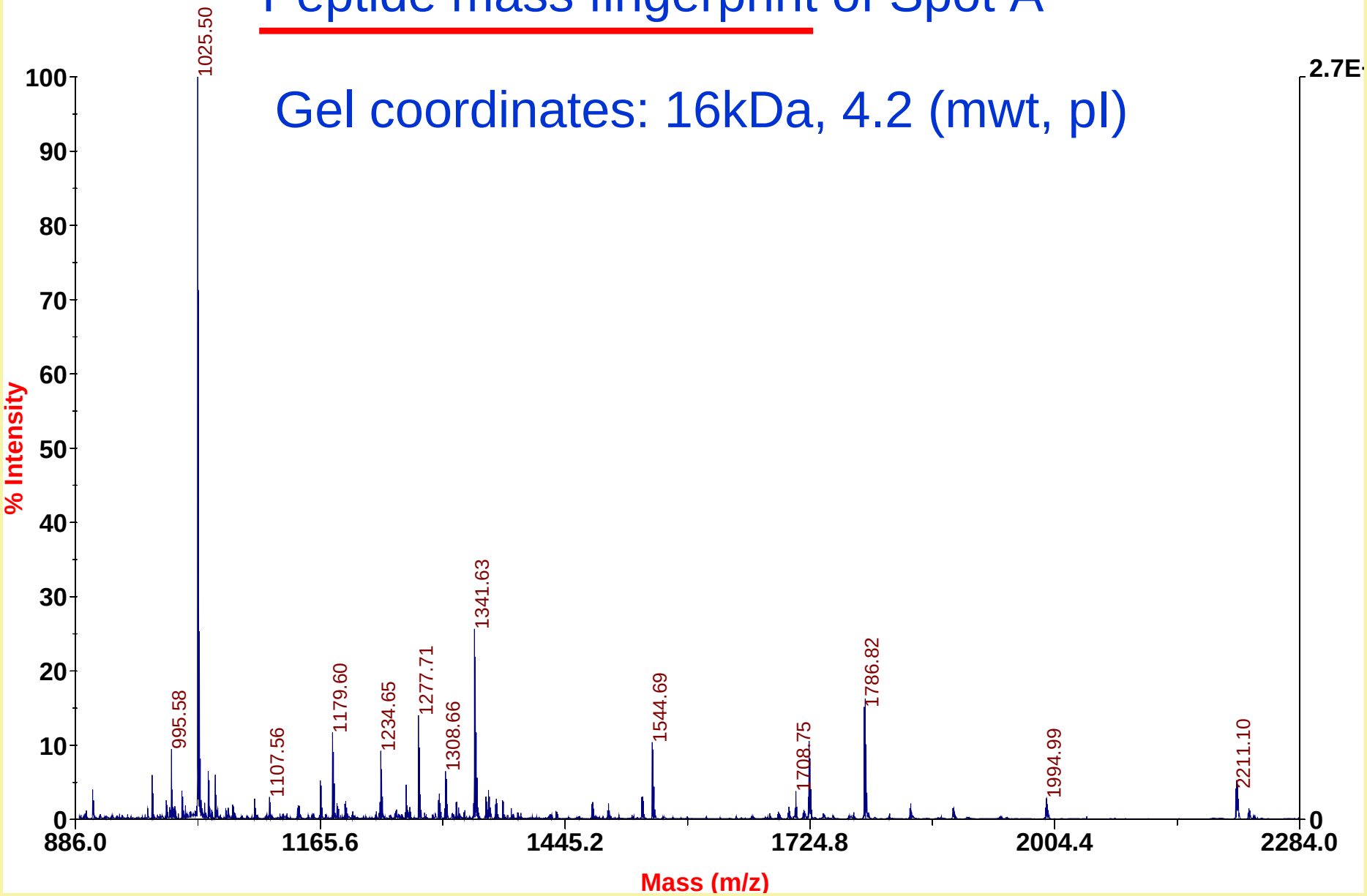
Cleavages Observed in MS/MS of Peptides



CID (Collision InDuced) Spectra – adds **sequence data** to **mass mapping** for improved database identification!

Peptide mass fingerprint of Spot A

Gel coordinates: 16kDa, 4.2 (mwt, pI)



de-mass fingerprinting tool from the UCSF Mass Spectrometry Facility that tries to fit a user's mass spectrometry data to a protein sequence in an existing database and thus sug
of the user's protein. The MS input data should be generated by analyzing the peptides produced by the enzymatic digestion of a user's protein.

nProspector Home	MS-Tag	MS-Seq	MS-Edman	MS-Fit at UCSF (San Francisco)
MS-Digest	MS-Product	MS-Comp	DB-Stat	MS-Isotope

Start Search

ase: SwissProt.012601 Instrument: MALDI-TOF

Frame translation: 3

Hits: From: msfit Filename: lastres

Hits to file: Filename: lastres

s: All

f Protein: (from 1000 Da to 150000 Da) All

n pI: (from 3.0 to 10.0) All

: Trypsin

ax. # of missed cleavages: 1

nes modified by: unmodified

inus: Hydrogen C terminus: Free Acid

e ID (comment): Unknown A

Reported Hits: 15

ole
fications
(default)

Peptide N-terminal Gln to pyroGlu
Oxidation of M
Protein N-terminus Acetylated
Acrylamide Modified Cys

User Defined Modification 1:

Phosphorylation of S, T and Y

OR

logy Mode (select any mode but identity)

h mode: identity

matches with NO AA substitutions: 1

Peptide masses are: monoisotopic

Min. # peptides required to match: 3

Report MOWSE Scores: Pfactor: 0.4

Peptide Masses

mass tolerance: +/- 15 ppm

Mass (m/z) Charge (z)

905.6874
973.5183
989.6093
995.5787
1007.4948
1024.4374
1025.4959
1025.7433
1037.5184
1045.5657
1090.5471
1106.5649
1139.5205
1164.5909
1165.5664
1179.6002
1184.5958
1193.6111
1233.5911
1234.6510
1263.6858
1267.7091
1277.7065

Mass accuracy tolerance = 15 ppm

This means that th mass is within 0.015 Da at m/z 1000

op on your browser if you wish to abort this MS-Fit search prematurely.

ID (comment): **Unknown A**

se searched: **SwissProt.012601**

ular weight search (**1000 - 150000 Da**) selects **90539** entries.

range: **92236** entries.

ned molecular weight and pI searches select **90539** entries.

search selects **858** entries (results displayed for top **15** matches).

ered modifications: | **Peptide N-terminal Gln to pyroGlu** | **Oxidation of M** | **Protein N-terminus Acetylated** | **Acrylamide Modified Cys** |

Peptides Match	Peptide Mass Tolerance (+/-)	Peptide Masses are	Digest Used	Max. # Missed Cleavages	Cysteines Modified by	Peptide N terminus	Peptide C terminus	Input # Peptide Masses
3	**15.000 ppm**	**monoisotopic**	**Trypsin**	**1**	**unmodified**	**Hydrogen (H)**	**Free Acid (O H)**	**46**

Result Summary

MOWSE Score	# (%) Masses Matched	Protein MW (Da)/pI	Species	SwissProt.012601 Accession #	Protein Name
.86e+005	9/46 (19%)	16930.2 / 4.56	HUMAN	P16475	MYOSIN LIGHT CHAIN ALKALI, NON-MUSCLE ISOFORM (MLC3NM) (LC17A) (LC17-NM)
.86e+005	9/46 (19%)	16961.2 / 4.46	HUMAN	P24572	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (MLC3SM) (LC17B) (LC17-GI)
.86e+005	9/46 (19%)	16975.3 / 4.46	RAT	Q64119	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (MLC3SM)
.77e+004	7/46 (15%)	15730.9 / 4.80	MOUSE	Q60605	MYOSIN LIGHT CHAIN ALKALI, NON-MUSCLE ISOFORM (MLC3NM)
.41e+004	7/46 (15%)	66018.0 / 8.16	HUMAN	P04264	KERATIN, TYPE II CYTOSKELETAL 1 (CYTOKERATIN 1) (K1) (CK 1) (67 KDA CYTOKERATIN) (HAIR A PROTEIN)
.19e+003	4/46 (8%)	15282.4 / 6.10	STRPU	P32006	PROFILIN
420	5/46 (10%)	16983.3 / 4.63	CHICK	P08296	MYOSIN LIGHT CHAIN ALKALI, NON-MUSCLE ISOFORM (FIBROBLAST) (G2 CATALYTIC) (LC17-NM)
419	5/46 (10%)	16987.4 / 4.52	CHICK	P02607	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (GIZZARD) (G2 CATALYTIC) (LC17-GI)
391	4/46 (8%)	38545.3 / 8.59	XENLA	P27006	ANNEXIN II TYPE I (LIPOCORTIN II) (CALPACTIN I HEAVY CHAIN) (CHROMOBINDIN 8) (P36) (PROTEI (PLACENTAL ANTICOAGULANT PROTEIN IV) (PAP-IV)
286	5/46 (10%)	22156.3 / 5.03	RAT	P16409	MYOSIN LIGHT CHAIN 1, SLOW-TWITCH MUSCLE B/VENTRICULAR ISOFORM
262	3/46 (6%)	19590.2 / 9.34	BGMV	P05174	AL2 PROTEIN (19.6 KD PROTEIN)
220	5/46 (10%)	21932.2 / 5.03	HUMAN	P08590	MYOSIN LIGHT CHAIN 1, SLOW-TWITCH MUSCLE B/VENTRICULAR ISOFORM (MLC1SB) (ALKALI)
211	3/46 (6%)	16990.5 / 6.92	ECOLI	P37052	HYPOTHETICAL 17.0 KDA PROTEIN IN HNR-PURU INTERGENIC REGION
202	3/46 (6%)	17947.3 / 5.24	ARATH	P25855	GLYCINE CLEAVAGE SYSTEM H PROTEIN 1, MITOCHONDRIAL PRECURSOR

matches (19%). 16930.2 Da, pI = 4.56. Acc. # P16475. HUMAN. MYOSIN LIGHT CHAIN ALKALI, NON-MUSCLE ISOFORM (MLC3NM) (LC17A) (LC17-NM).

ted	MH^+ matched	Delta ppm	start	end	Peptide Sequence (Click for Fragment Ions)	Modifications
787	995.5890	-10.3014	111	119	(R)HVLVTLGEK(M)	
959	1025.5056	-9.4785	14	21	(K)EAFQLFDR(T)	
911	1233.5898	1.0857	99	110	(K)EGNGTVMGAEIR(H)	
187	1354.7331	-10.5955	38	50	(R)ALGQNPTNAEVLK(V)	
928	1544.6869	3.8248	82	94	(K)DQGTYEDYVEGLR(V)	
598	1722.8485	6.5620	95	110	(R)VFDKEGNGTVMGAEIR(H)	
229	1786.8248	-1.0535	80	94	(K)NKDQGTYEDYVEGLR(V)	
274	1888.0043	12.2526	64	79	(K)VLDFEHFLPMLQTVAK(N)	
294	2226.1552	-11.6082	99	119	(K)EGNGTVMGAEIRHVLVTLGEK(M)	1Met-ox

atched masses: 905.6874 973.5183 989.6093 1007.4948 1024.4374 1025.7433 1037.5184 1045.5657 1090.5471 1106.5649 1139.5205 1164.5909 1165.5664 1179.6002 1184.5
111 1234.6510 1263.6858 1267.7091 1277.7065 1300.5432 1307.6644 1308.6596 1340.6612 1341.6288 1357.6707 1373.6434 1475.7257 1493.7172 1532.6160 1699.8525 170
276 1723.8256 1838.9438 1993.9497 2211.1041

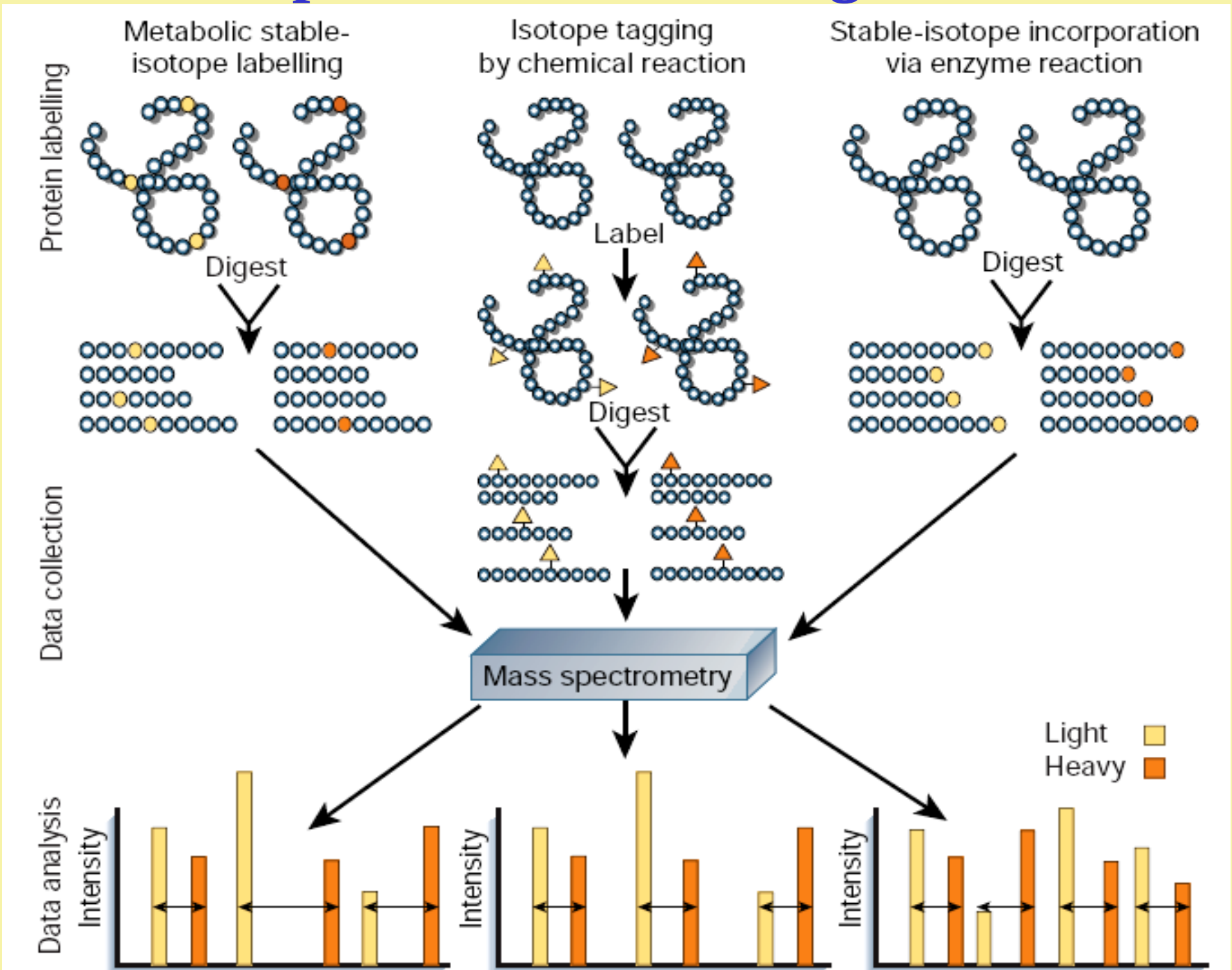
tched peptides cover **50%** (77/151 AA's) of the protein.
ge Map for This Hit (MS-Digest index #): 11572

matches (19%). 16961.2 Da, pI = 4.46. Acc. # P24572. HUMAN. MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (MLC3SM) (LC17B) (LC17-GI).

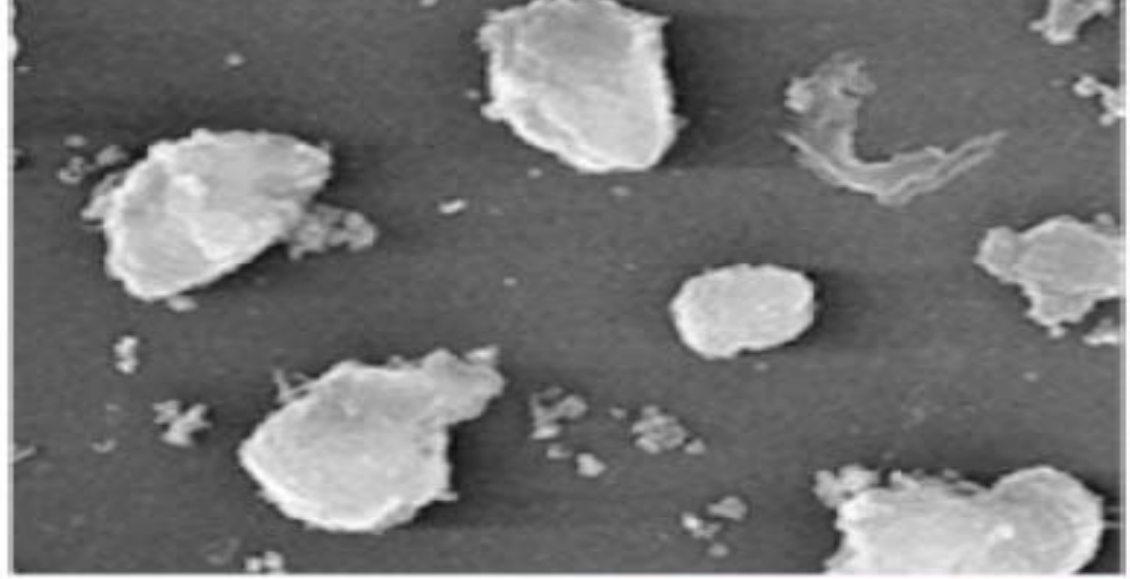
ted	MH^+ matched	Delta ppm	start	end	Peptide Sequence (Click for Fragment Ions)	Modifications
787	995.5890	-10.3014	111	119	(R)HVLVTLGEK(M)	
959	1025.5056	-9.4785	14	21	(K)EAFQLFDR(T)	
911	1233.5898	1.0857	99	110	(K)EGNGTVMGAEIR(H)	
187	1354.7331	-10.5955	38	50	(R)ALGQNPTNAEVLK(V)	
928	1544.6869	3.8248	82	94	(K)DQGTYEDYVEGLR(V)	
598	1722.8485	6.5620	95	110	(R)VFDKEGNGTVMGAEIR(H)	
229	1786.8248	-1.0535	80	94	(K)NKDQGTYEDYVEGLR(V)	
274	1888.0043	12.2526	64	79	(K)VLDFEHFLPMLQTVAK(N)	
294	2226.1552	-11.6082	99	119	(K)EGNGTVMGAEIRHVLVTLGEK(M)	1Met-ox

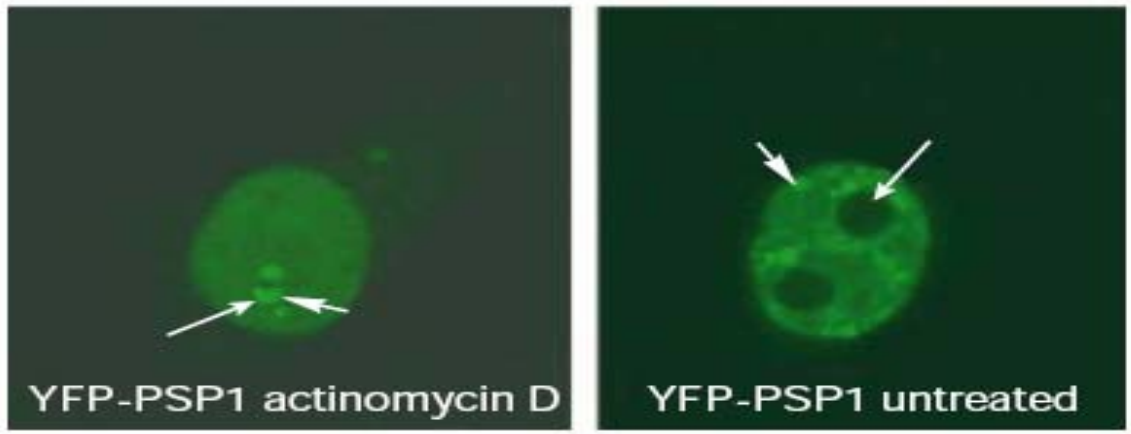
atched masses: 905.6874 973.5183 989.6093 1007.4948 1024.4374 1025.7433 1037.5184 1045.5657 1090.5471 1106.5649 1139.5205 1164.5909 1165.5664 1179.6002 1184.5

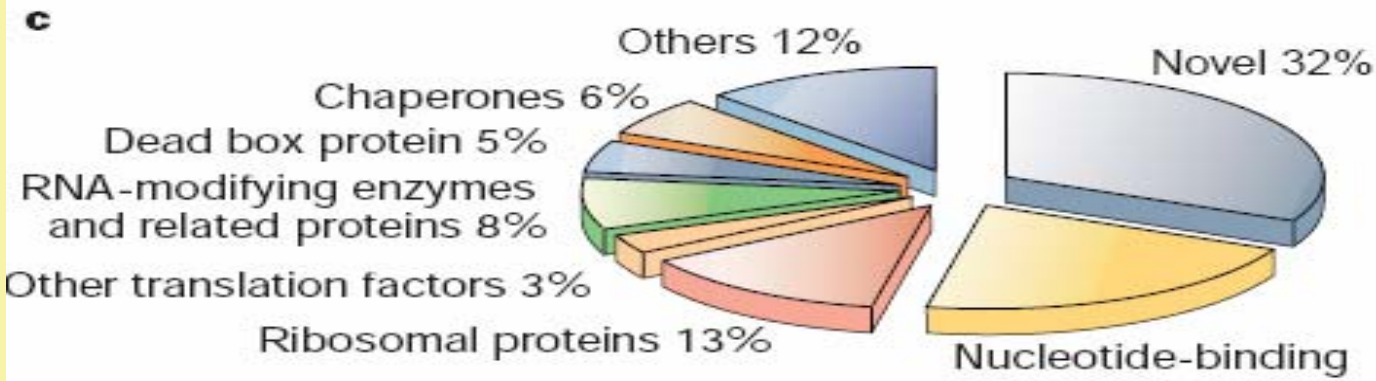
Stable-isotope Protein Labelling for Proteomics



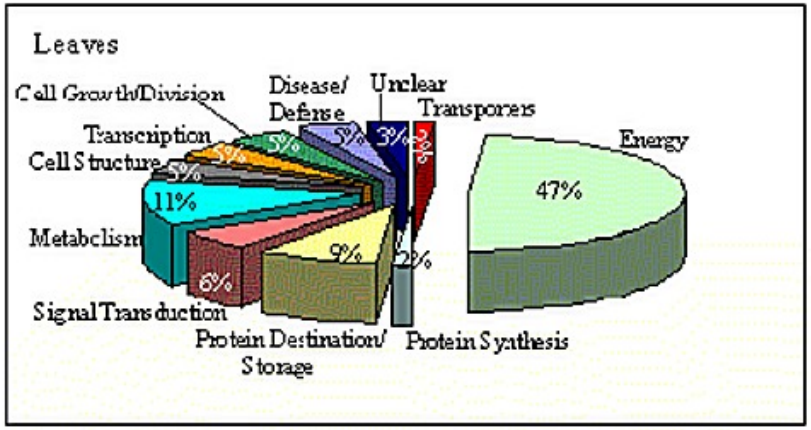
Organellar Proteomics: Combined MS and Imaging Methods

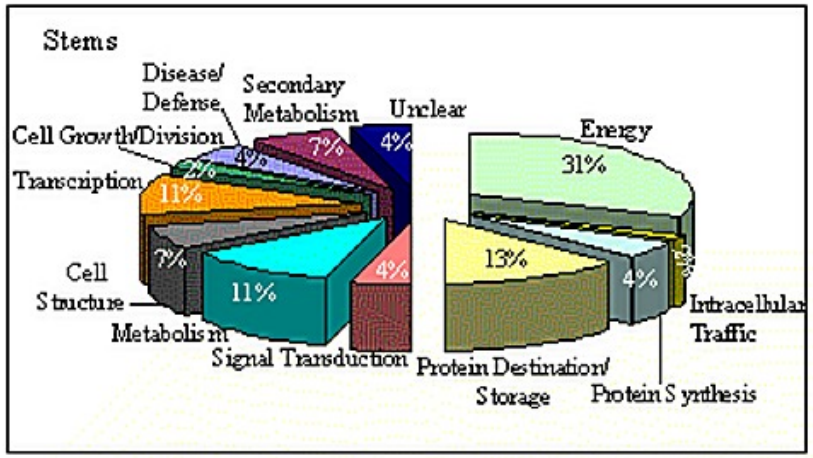


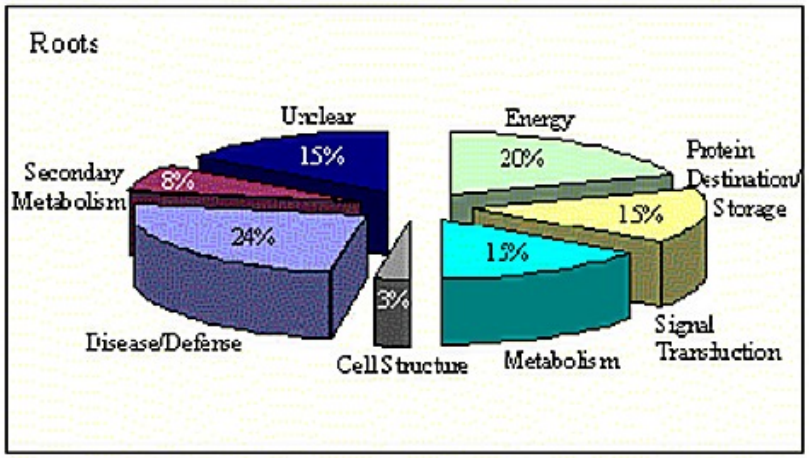


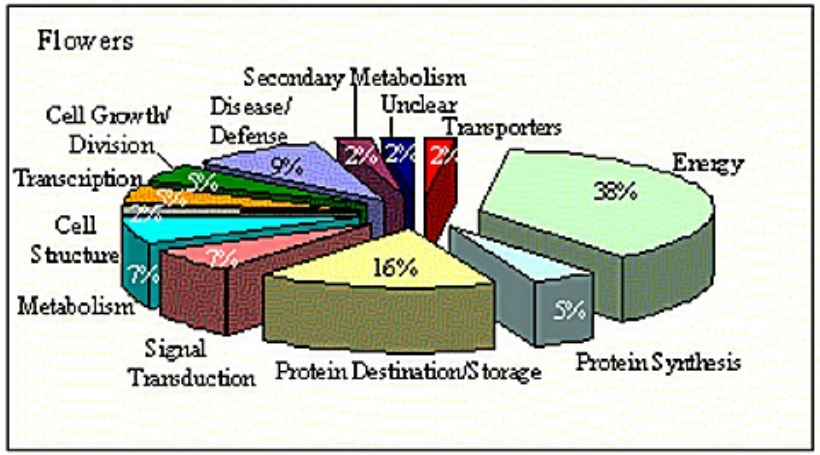


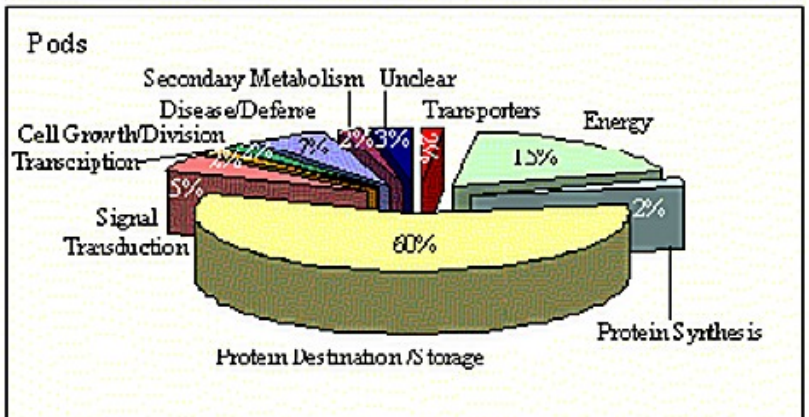
Summary of the functions of various proteins identified in specific tissues of *M. truncatula*.

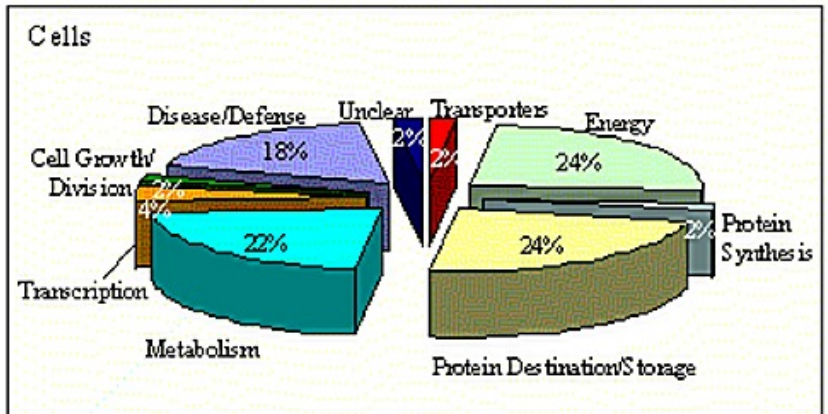












A Mammalian Organelle Map by Protein Correlation Profiling

Leonard J. Foster,[1,2] Carmen L. de Hoog,[1,2] Yanling Zhang,[3,4] Yong Zhang,[3,4] Xiaohui Xie,[5] Vamsi K. Mootha,[5,6] and Matthias Mann[1,3,*]

[1] Center for Experimental BioInformatics (CEBI), Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

[2] Centre for Proteomics, Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

[3] Department of Proteomics and Signal Transduction, Max-Planck Institute for Biochemistry, Martinsried, Germany D-82152

[4] Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China

[5] Broad Institute of Harvard and MIT, Cambridge, MA 02139, USA

[6] Department of Systems Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115, USA

*Contact: mmann@biochem.mpg.de

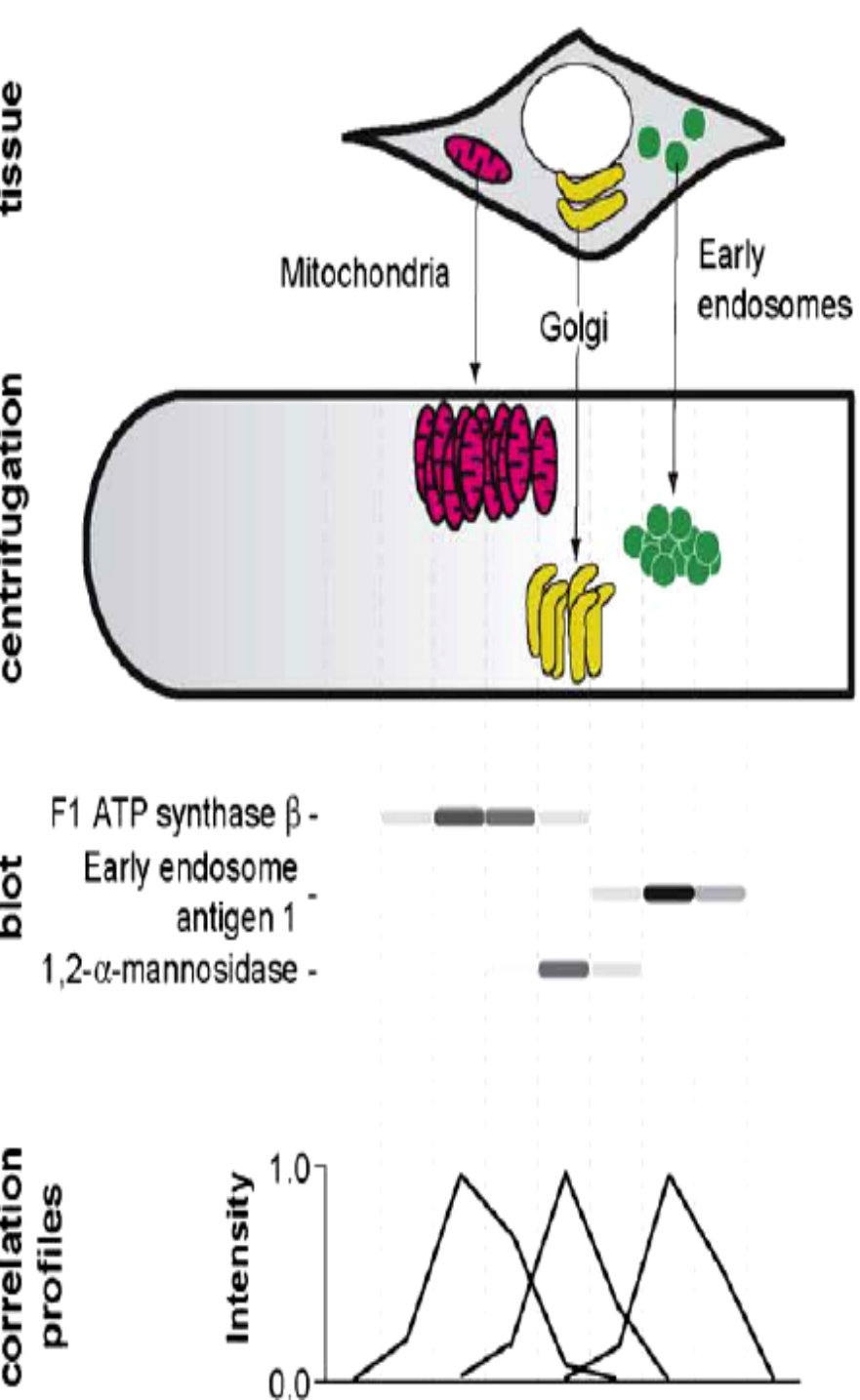
DOI 10.1016/j.cell.2006.03.022

SUMMARY

Protein localization to membrane-enclosed organelles is a central feature of cellular organization. Using protein correlation profiling, we have mapped 1,404 proteins to ten subcellular locations in mouse liver, and these correspond with enzymatic assays, marker protein profiles, and confocal microscopy. These localizations allowed assessment of the specificity in published organellar proteomic inventories and demonstrate multiple locations for 39% of all organellar proteins. Integration of proteomic and genomic data enabled us to identify networks of coexpressed genes, *cis*-regulatory motifs, and putative transcriptional regulators involved in organelle biogenesis. Our analysis ties bio-

microscopic examination of an organelle, certain proteins or enzyme activities that appear to localize exclusively to that organelle are considered markers, essentially defining that compartment.

Recently, proteomics (de Hoog and Mann, 2004) has been applied to study organelle composition. The genetic tractability of *Saccharomyces cerevisiae* has allowed a large fraction of yeast ORFs to be tagged for localization studies (Ross-Macdonald et al., 1999; Kumar et al., 2002; Huh et al., 2003), but such an approach is more challenging in mammalian systems due, in part, to artifacts from overexpression (Simpson et al., 2000). Mass spectrometry-based proteomics (Aebersold and Mann, 2003) is often employed to characterize the protein composition of organelle-enriched fractions. Indeed, protein catalogs are now available for virtually all cytoplasmic organelles as well as most of the major nuclear ones (reviewed in Yates et al., 2005). However, due to the high sensitivity of mass spectrometers and the difficulties inherent in pu-



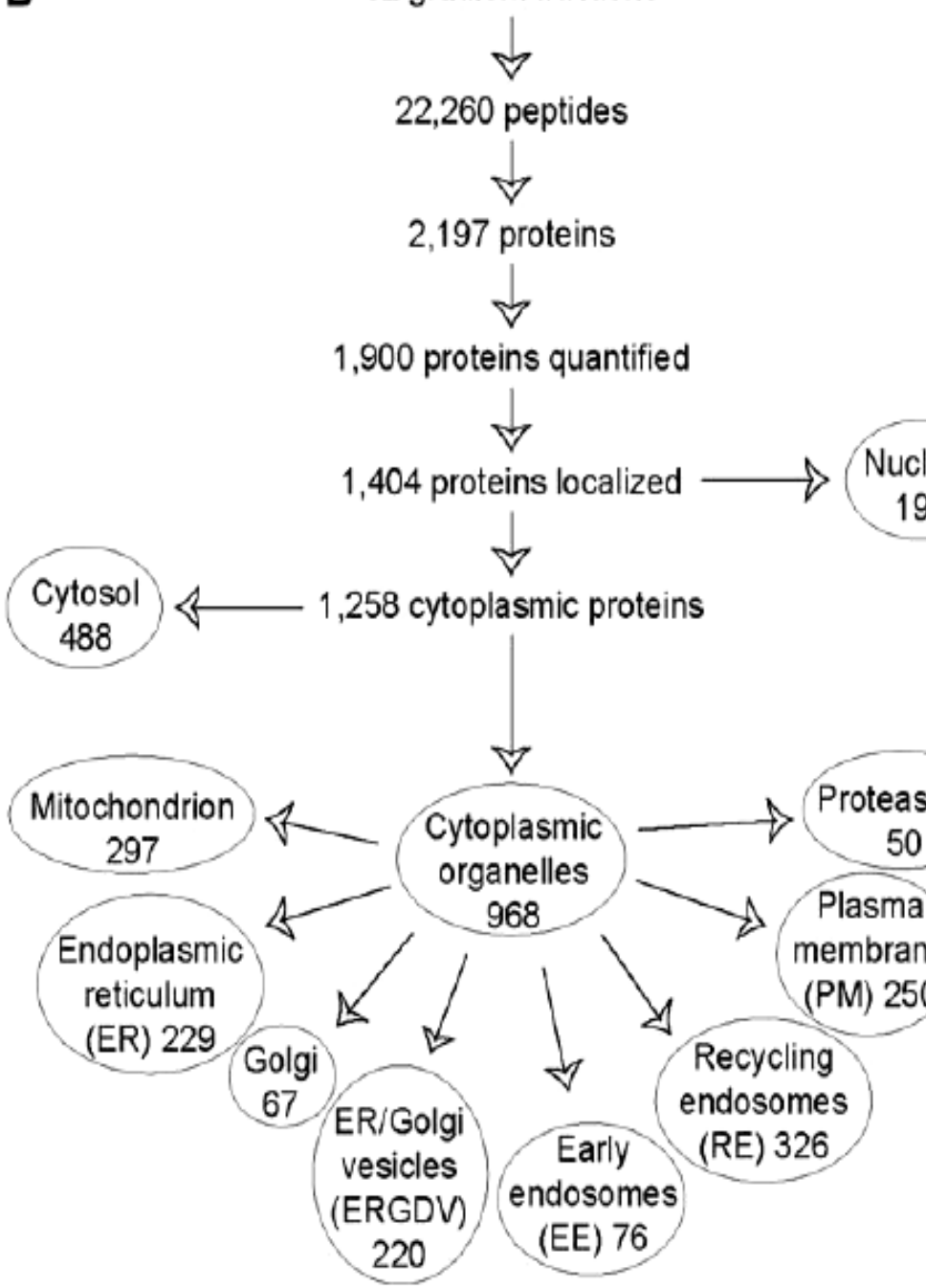
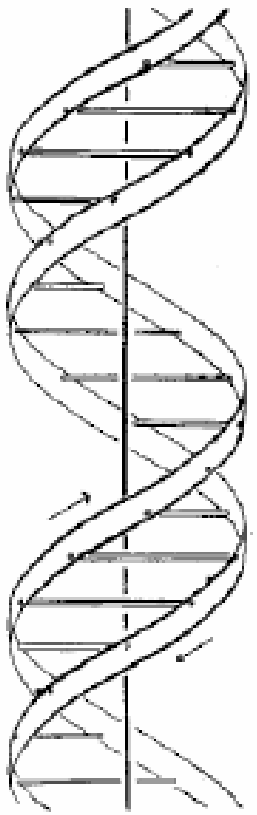


Figure 1. Organelle Profiling with Gradient Centrifugatio

The Birth of Molecular Biology: DNA Structure

nside, linked together by hydrogen bonds. This tructure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate di-ester groups joining β-D-deoxy-ribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There

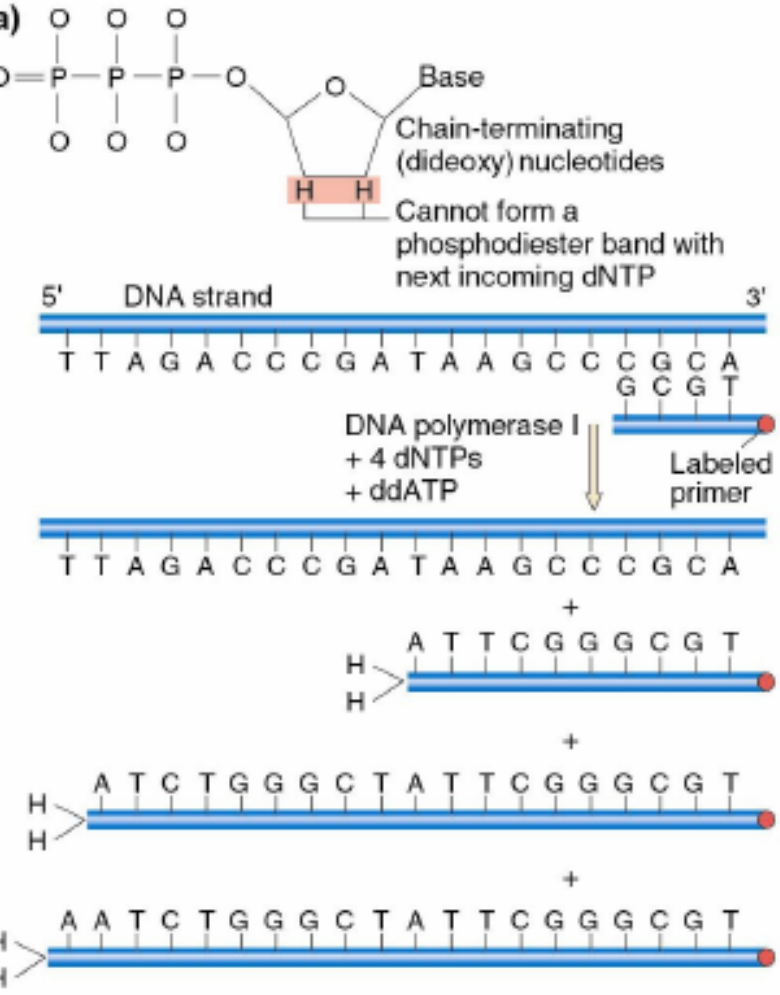


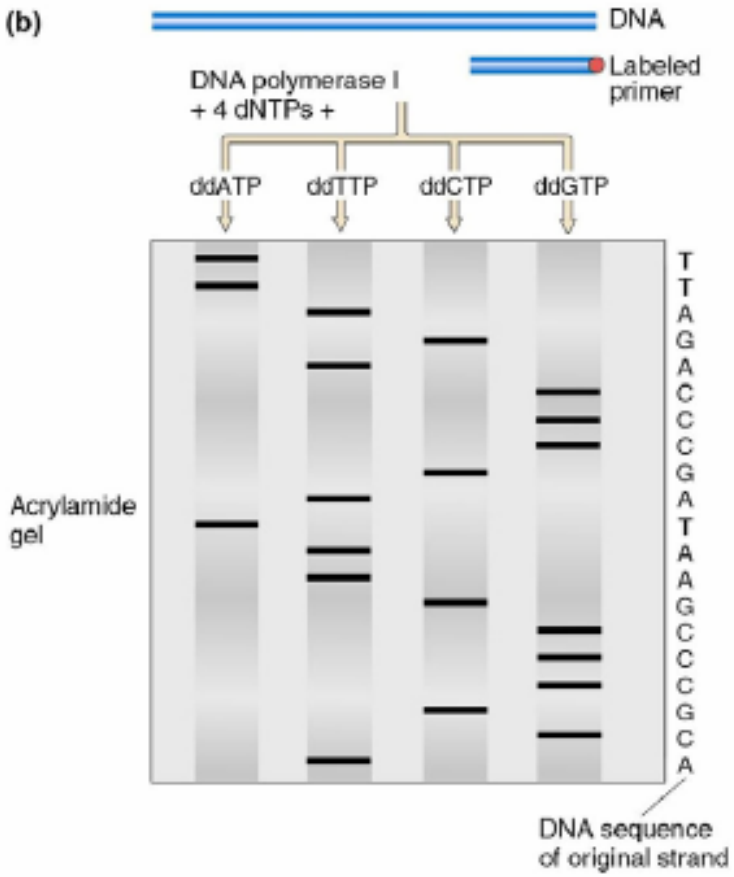
This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

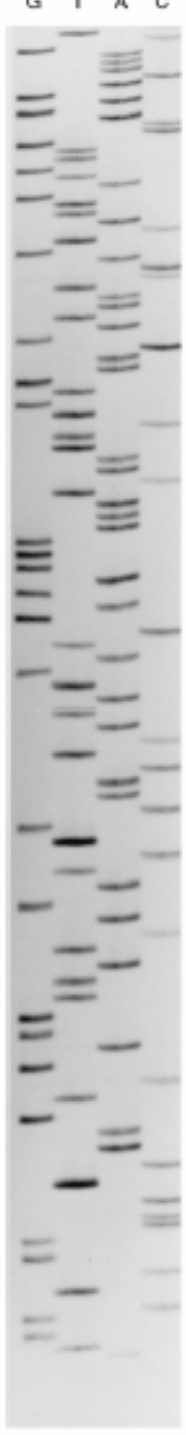
Nature – 1953

Nature – 2001

Dideoxy sequencing







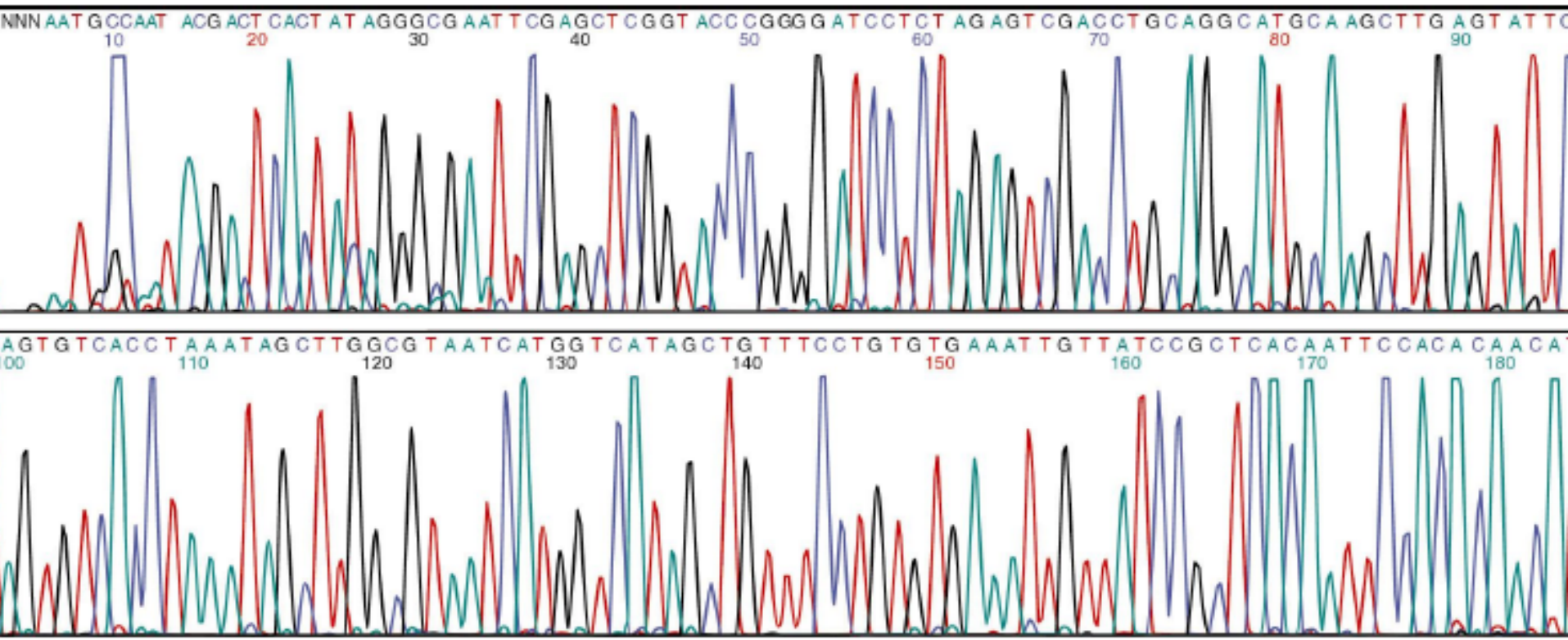
Automated dye-terminator sequencing

4-fluorescently labelled dideoxy dye terminators

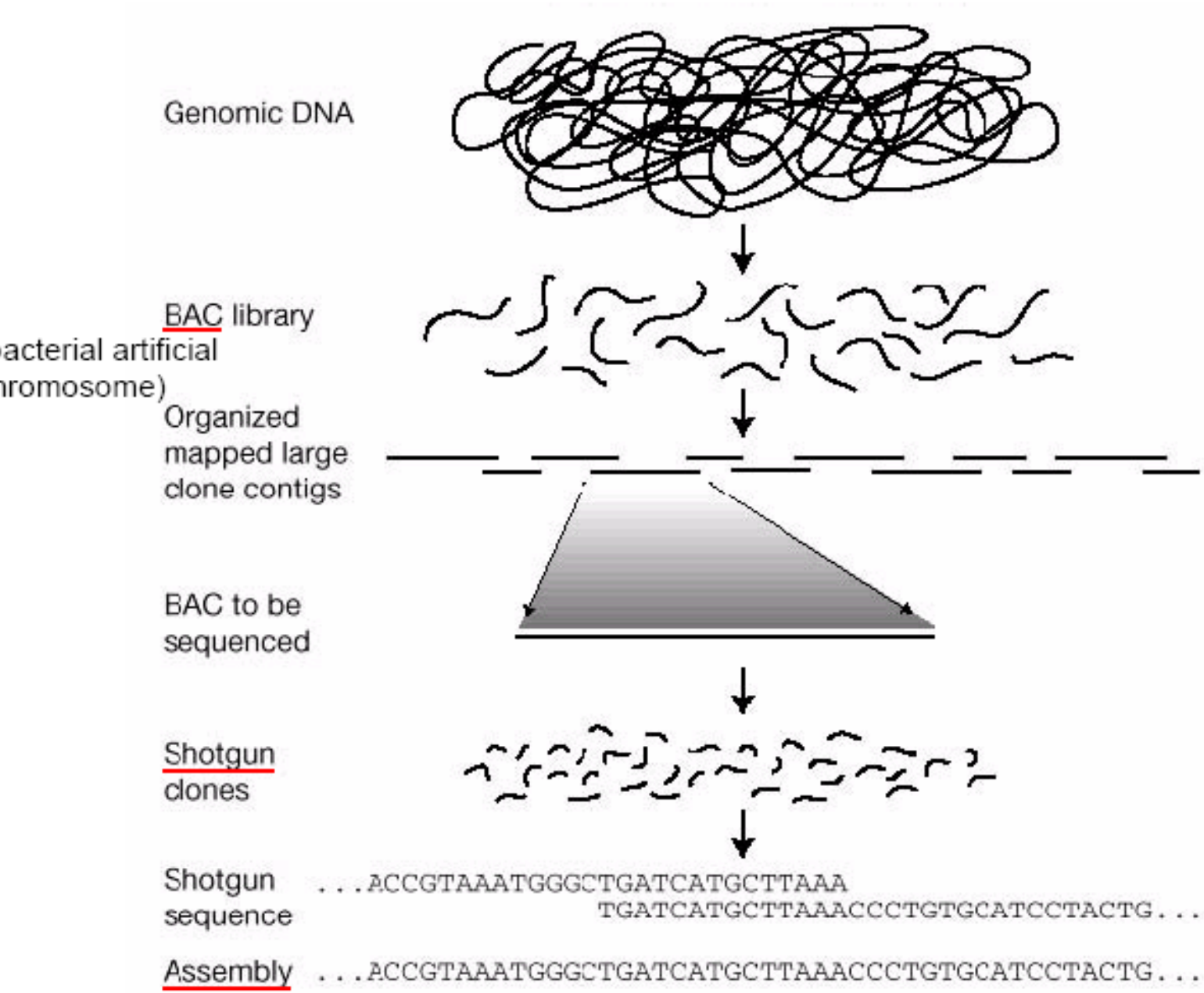
ddATP
ddGTP
ddCTP
ddTTP

pool and load in a single well or capillary

- scan with laser + detector specific for each dye
- automated base calling
- very long reads (~ 1000 bases)/run



Physical mapping and sequencing of the human genome



Nature (2001) **409** p. 860-921

im Kent is a research scientist at UC Santa Cruz.

he human genome project was ultimately a race between Celera enomics and the public effort, with the final push being a bioinformatics roblem to put all of the sequence reads together into a draft genome equence. **Jim Kent was a grad student at UCSC**, who worked for weeks eveloping the algorithm to put all of this together, **beating Celera by 3 ays** to an assembled human genome sequence.

is efforts ensured that the human genome data remained in the public omain and were not patented into private intellectual property.

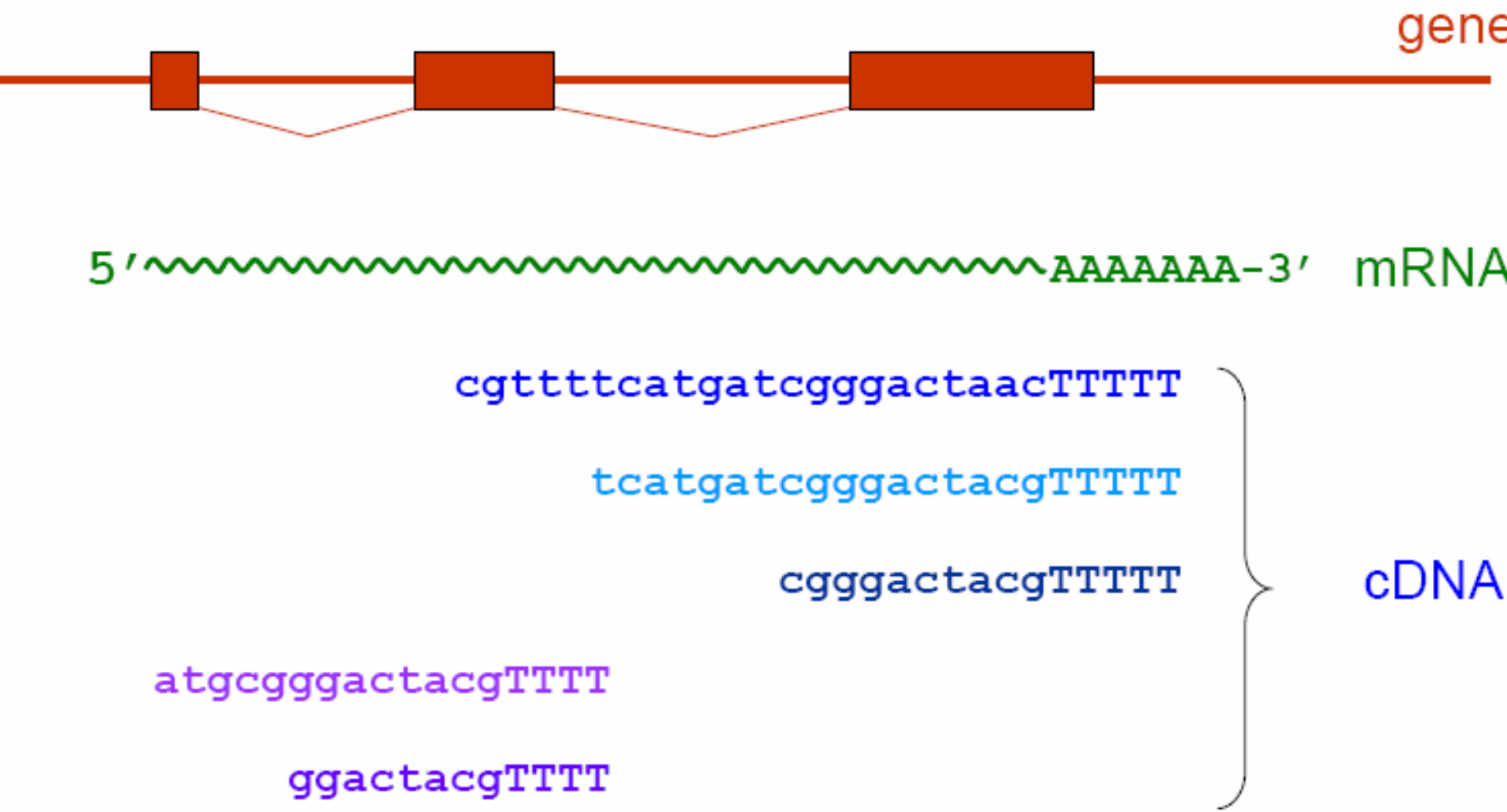
ent built a grid of cheap, commodity PC's running the Linux operating ystem and other Freeware to beat Celera's, what was thought of then as he, world's most powerful civilian computer. In **June 2000**, thanks to the ork done by Kent and several others, the Human Genome Project was ble to publish its data in the Public Domain just hours ahead of Celera.

ent went on to write BLAT and the UCSC Human Genome Browser to help nalyze important genome data, receiving his PhD in biology in 2002. Today t UCSC he works primarily on web tools to help understand the human enome. He helps maintain and upgrade the browser, and has worked on

Finding genes in genomes

- compare to EST or cDNA sequence
- look for open reading frames
- similarity to other genes and proteins
- Gene prediction algorithms (identifying splice sites, coding sequence bias, etc.)

Genes can also be identified by sequencing cDNAs at random. The sequenced cDNAs are called ESTs (expressed sequence tags)



The BIG QUESTION:

Why do we have so few genes?

ecies	Genome size	Number of ger
man (*Homo sapiens*)	2.9 billion base pairs	25,000 - 30,0
it fly (*Drosophila melanogaster*)	120 million base pairs	13,6
rm (*Caenorhabditis elegans*)	97 million base pairs	19,0
dding yeast (*Saccharomyces cerevisiae*)	12 million base pairs	6,0
coli	4.1 million base pairs	4,8

Genomics vs. Proteomics

With the completion of a rough draft of the human genome many researchers are looking at how genes and proteins interact to form other proteins. A surprising finding of the Human Genome Project is that there are far fewer protein-coding genes in the human genome than proteins in the human proteome (20,000 to 25,000 genes vs. about 1,000,000 proteins). The human body may contain more than 2 million proteins, each having different functions. The protein diversity is thought to be due to alternative splicing and post-translational modification of proteins. The discrepancy implies that protein diversity cannot be fully characterized by gene expression analysis, thus proteomics is useful for characterizing cells and tissues.

Functional genomics and proteomics

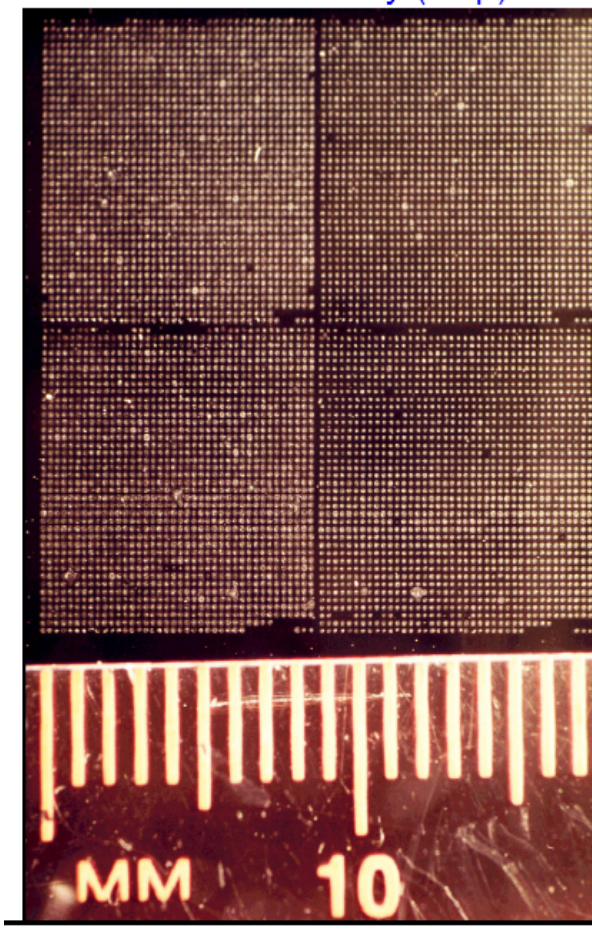
- Identify genes and proteins encoded in the genome (Gene finding)
- Measure gene expression on a genome-wide scale (microarrays)
- Identify protein function
  30-50% of the genes in a genome are of unknown function
- Identify protein interactions, biochemical pathways, gene interaction networks inside cells

Methods of making microarrays

- Robotic spotting
  - using a printing tip
  - using inkjets

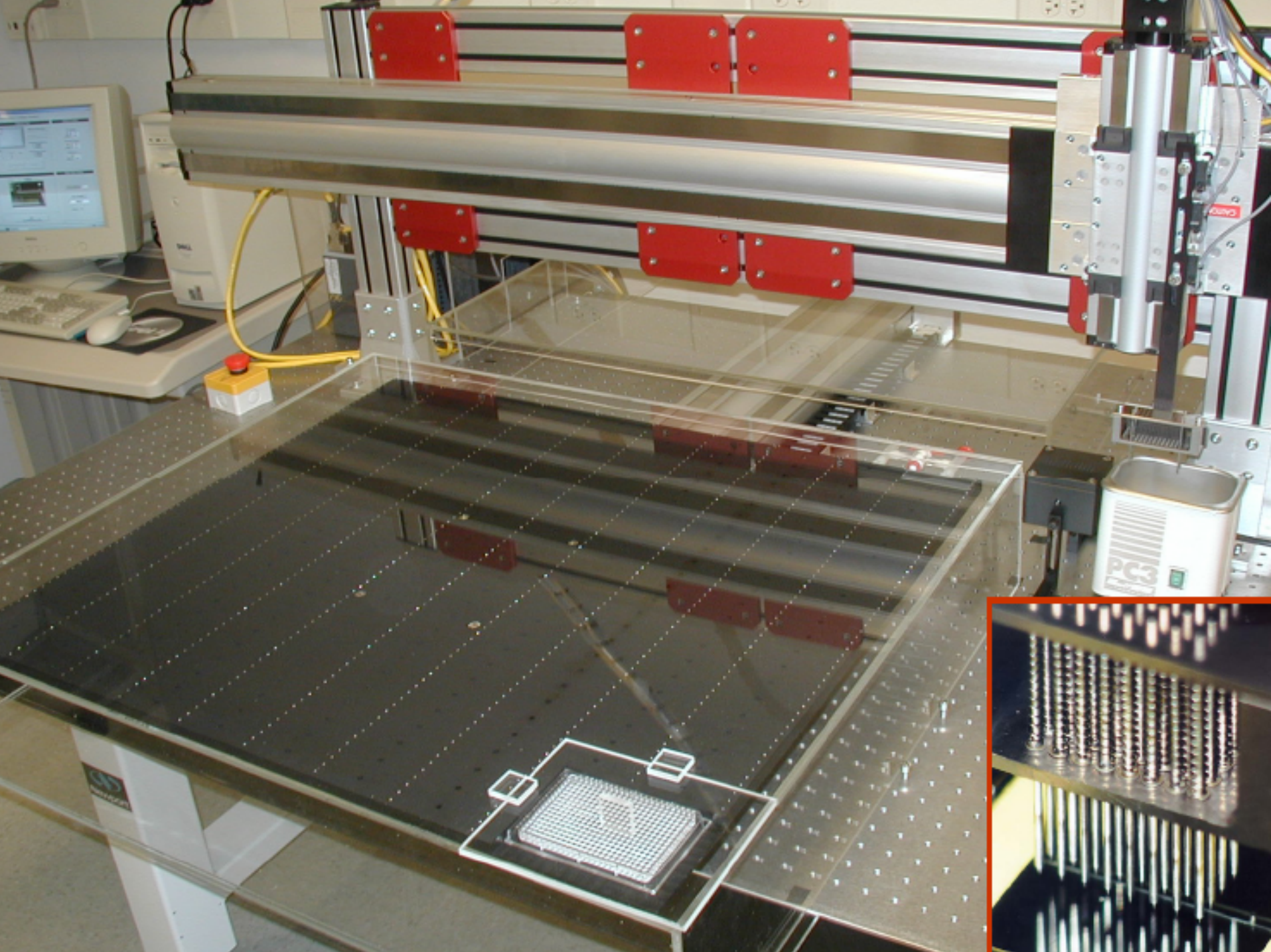
- Synthesis of oligonucleotides
  - photolithography (Affymetrix)
  - using inkjets
  - Digital Light Processor (DLP) or Digital Micromirror Device (DMD)

DNA microarray (chip)

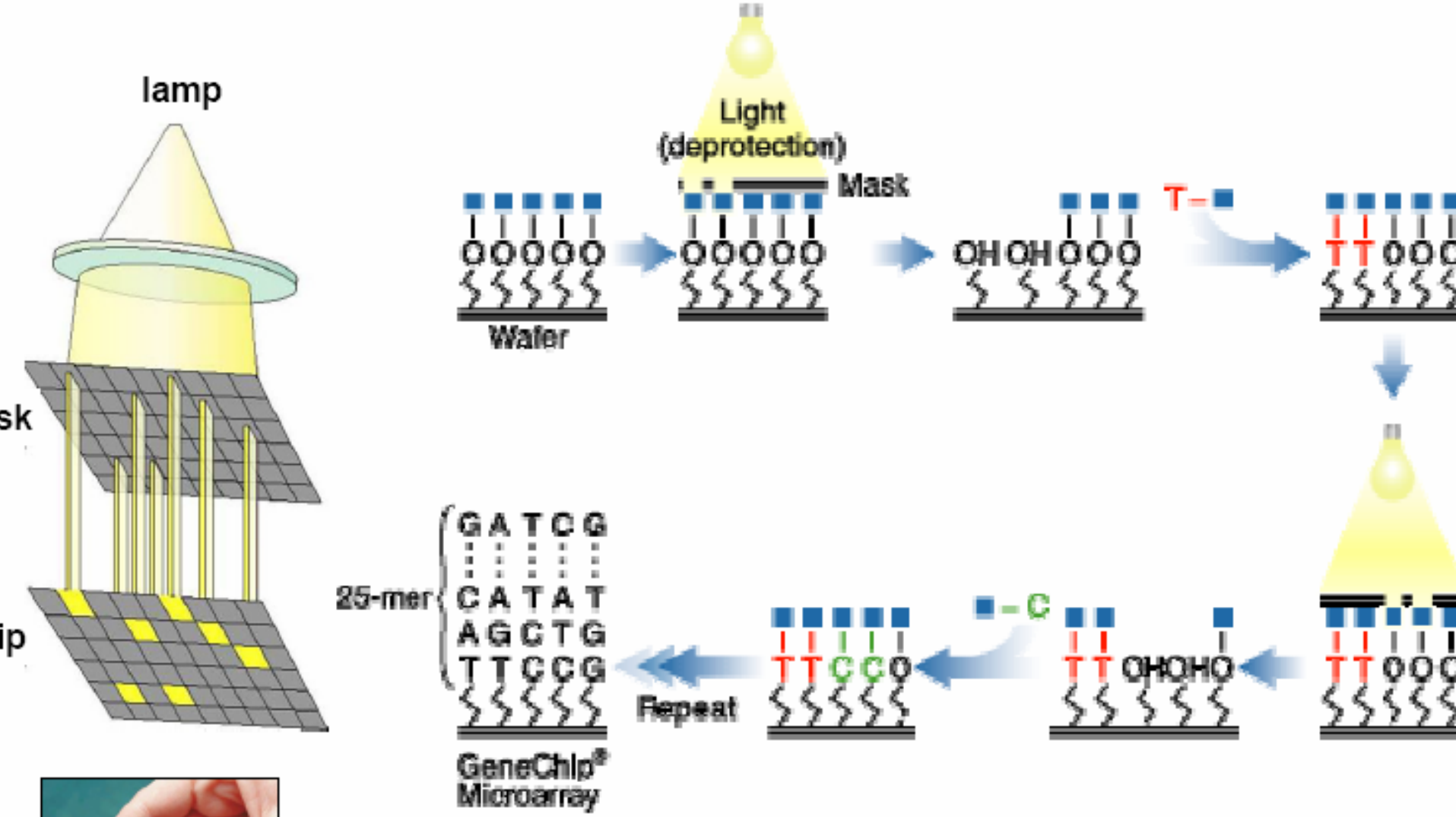


Microarrays can be used to study gene expression, DNA-protein interactions, mutations, protein-protein interactions, etc., all on a genome-wide scale

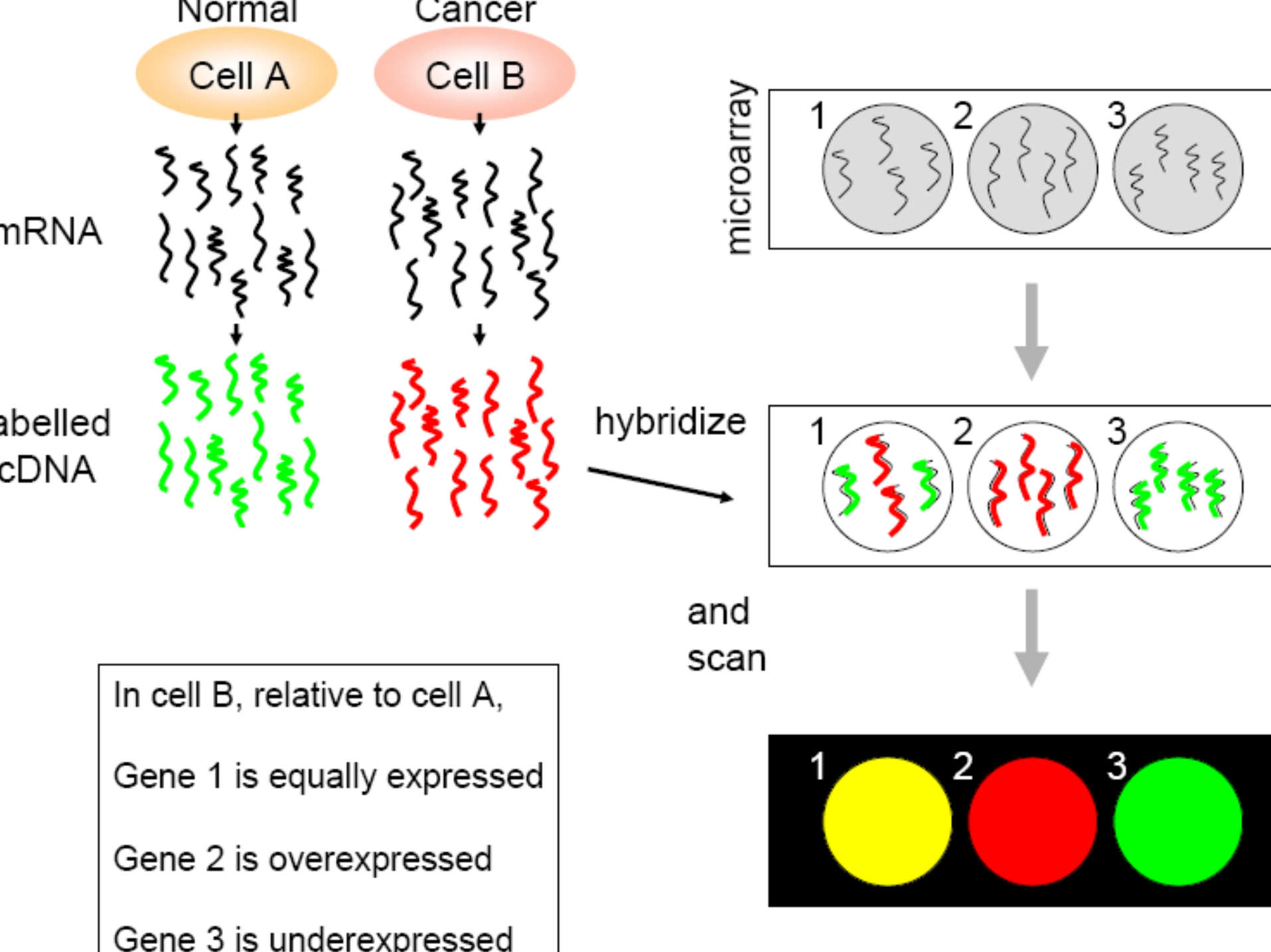
Note: Thanks to Prof. Vishy Iyer for many of these slides on microarrrays



Affymetrix GeneChip



courtesy: *www.affymetrix.com*



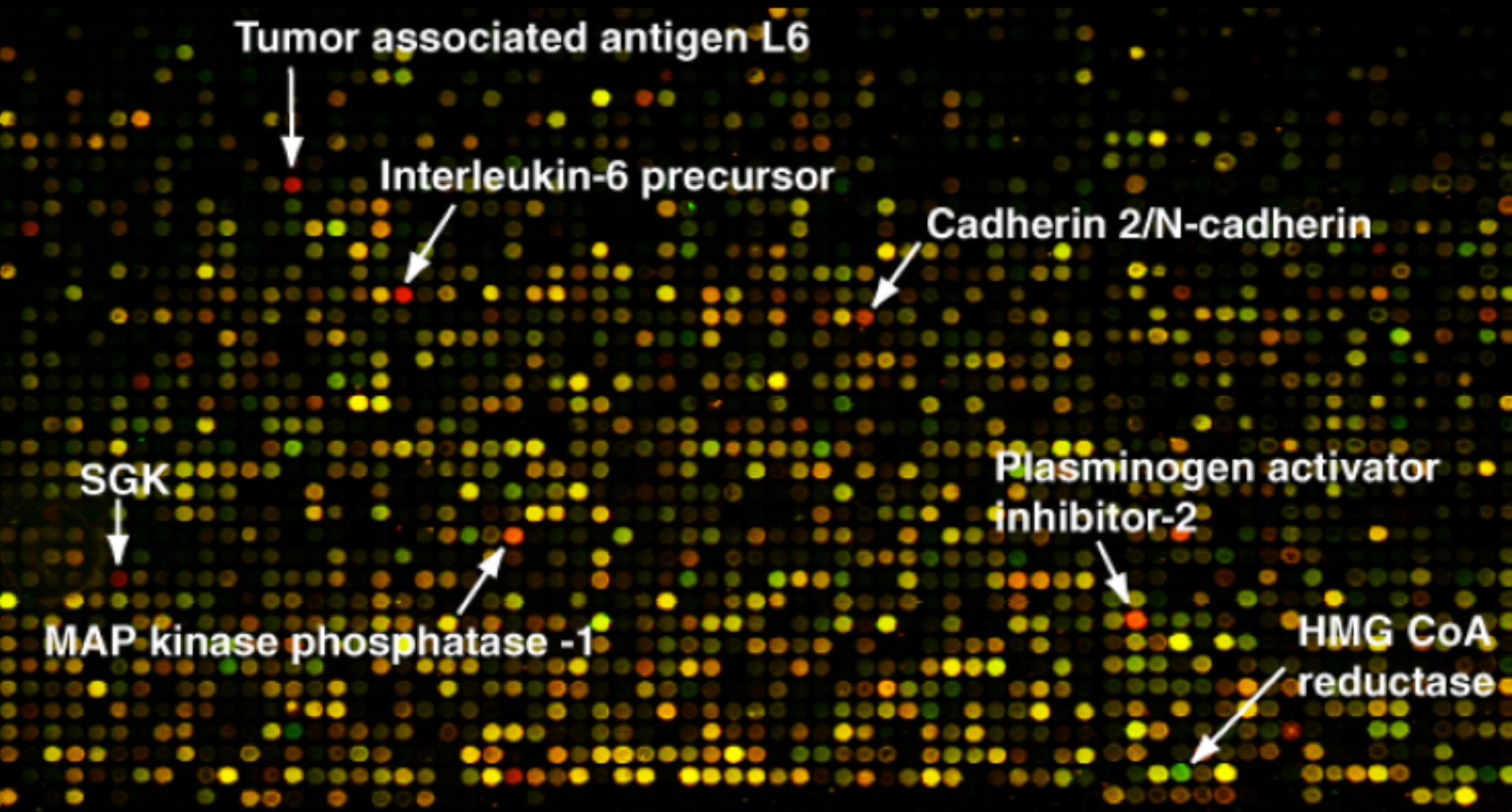
In cell B, relative to cell A,

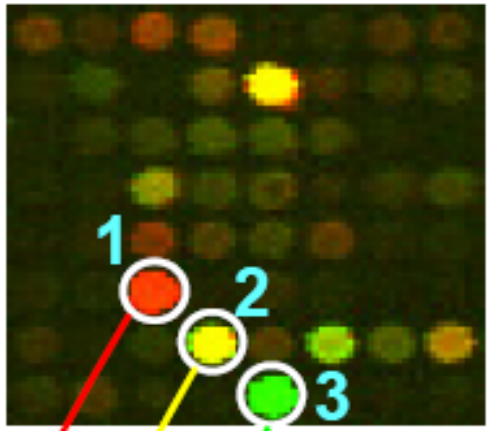
Gene 1 is equally expressed

Gene 2 is overexpressed

Gene 3 is underexpressed

DNA microarray after hybridization of fluorescent probes





Original microarray image

Colour representation of differential gene expression

Green	Red	Red / Green
200	10000	50.00
4800	4800	1.00
9000	300	0.03

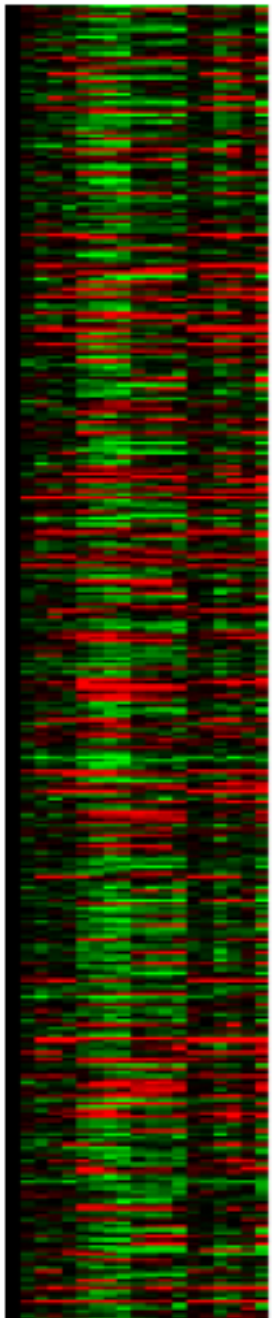
Gene 1
Gene 2
Gene 3

- Large amounts of data can be displayed in this manner
- Gene expression data can be computationally analyzed and organized to reveal patterns

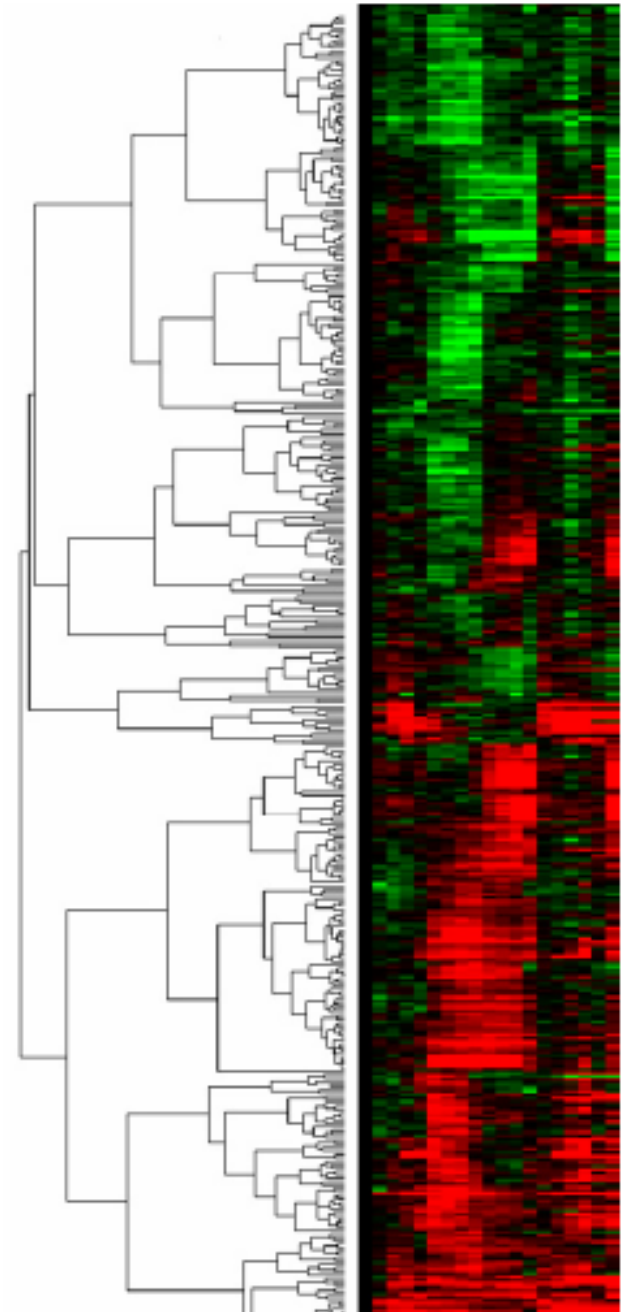
Experiments

riginal data

Genes

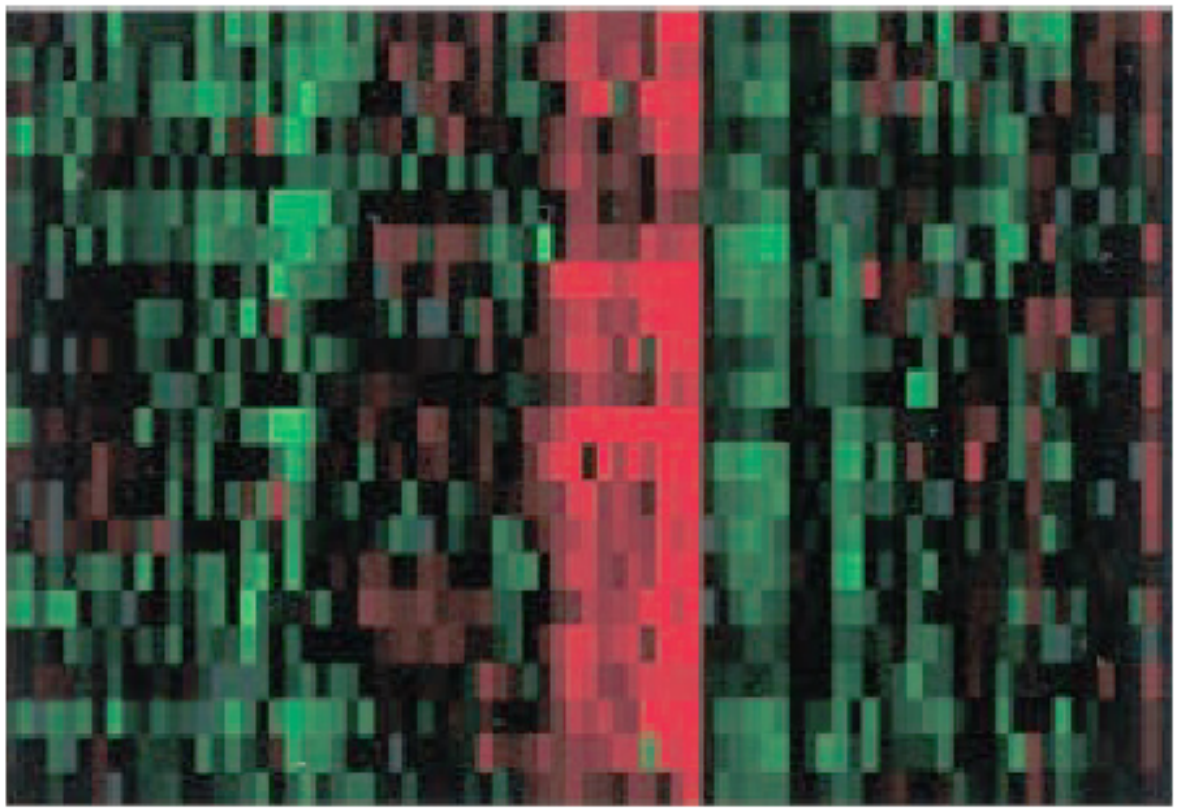


Data after hierarchical clustering



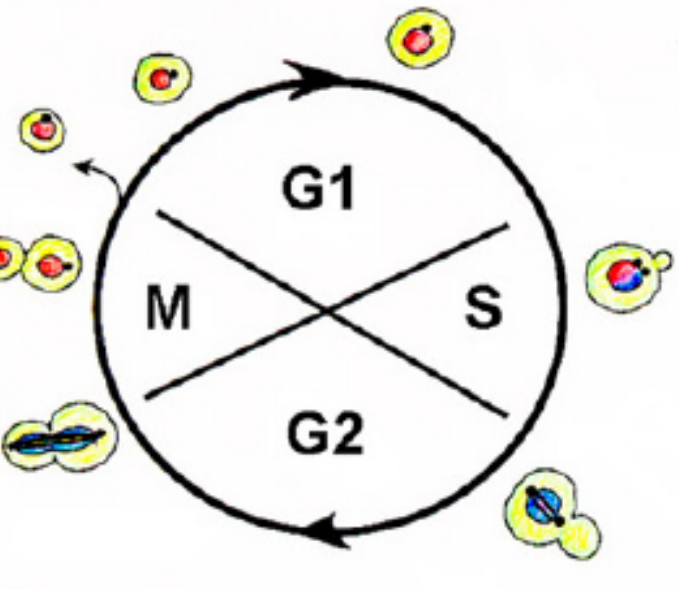
Functionally related genes are often co-expressed

A cluster of co-expressed genes

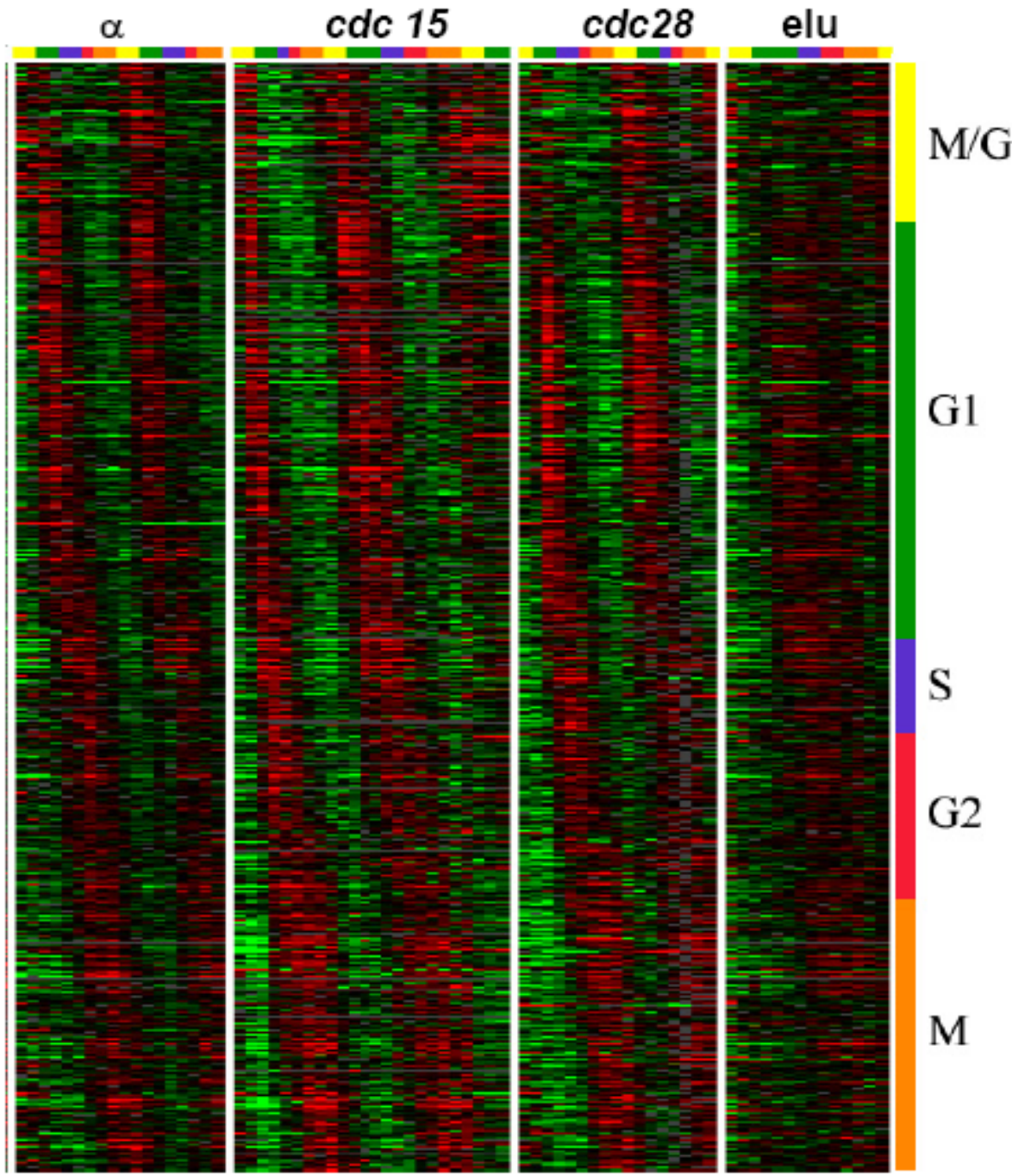


Thus, unknown Gene X may also be a ribosomal protein

S. cerevisiae mitotic cell-cycle



Brenda Andrews lab, University of Toronto



Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

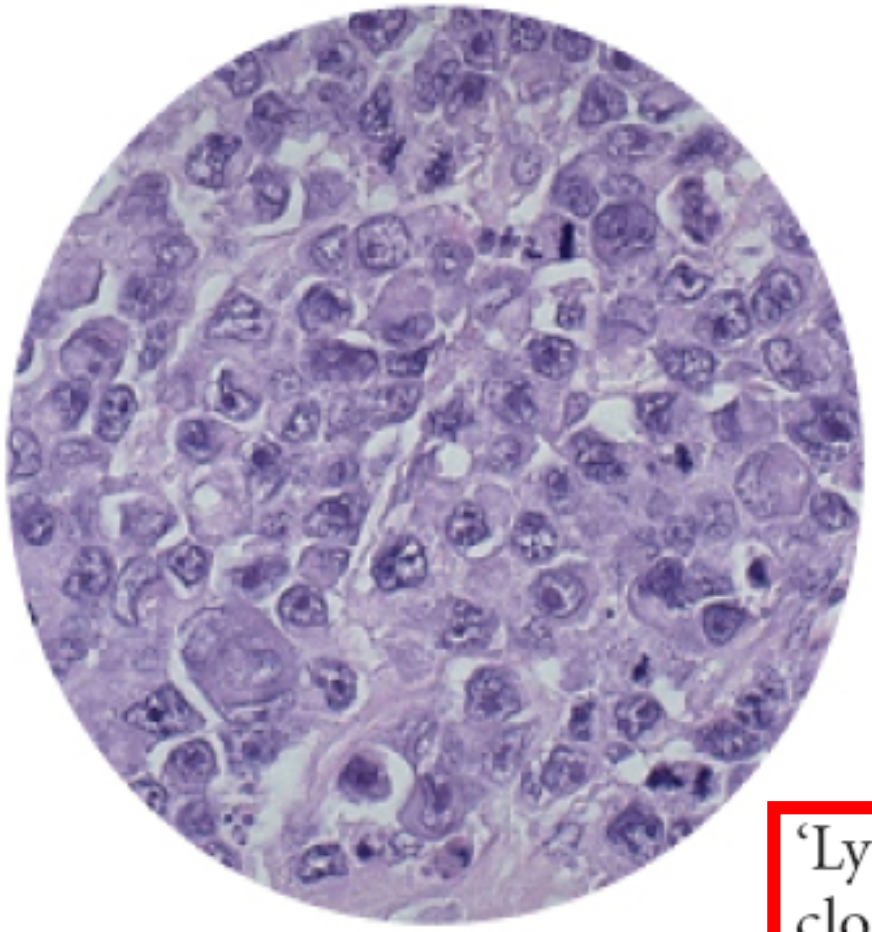
h A. Alizadeh[1,2], Michael B. Eisen[2,3,4], R. Eric Davis[5], Chi Ma[5], Izidore S. Lossos[6], Andreas Rosenwald[5], Jennifer C. Boldrick[1], jeer Sabet[5], Truc Tran[5], Xin Yu[5], John I. Powell[7], Liming Yang[7], Gerald E. Marti[8], Troy Moore[9], James Hudson Jr[9], Lisheng Lu[10], vid B. Lewis[10], Robert Tibshirani[11], Gavin Sherlock[4], Wing C. Chan[12], Timothy C. Greiner[12], Dennis D. Weisenburger[12], mes O. Armitage[13], Roger Warnke[14], Ronald Levy[6], Wyndham Wilson[15], Michael R. Grever[16], John C. Byrd[17], David Botstein[4], brick O. Brown[1,18] & Louis M. Staudt[5]

NATURE | VOL 403 | 3 FEBRUARY 2000 | www.nature.com

use large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma, is clinically heterogeneous: 4
patients respond well to current therapy and have prolonged survival, whereas the remainder succumb to the disease. W
posed that this variability in natural history reflects unrecognized molecular heterogeneity in the tumours. Using DNA
roarrays, we have conducted a systematic characterization of gene expression in B-cell malignancies. Here we show that th
iversity in gene expression among the tumours of DLBCL patients, apparently reflecting the variation in tumour prolifera
, host response and differentiation state of the tumour. We identified two molecularly distinct forms of DLBCL which had g
ression patterns indicative of different stages of B-cell differentiation. One type expressed genes characteristic of germi
tre B cells ('germinal centre B-like DLBCL'); the second type expressed genes normally induced during *in vitro* activation
ipheral blood B cells ('activated B-like DLBCL'). Patients with germinal centre B-like DLBCL had a significantly better ove
vival than those with activated B-like DLBCL. The molecular classification of tumours on the basis of gene expression can t
ntify previously undetected and clinically significant subtypes of cancer.

pite the variety of clinical, morphological and molecular param-
s used to classify human malignancies today, patients receiving
same diagnosis can have markedly different clinical courses and
tment responses. The history of cancer diagnosis has been
ctuated by reassortments and subdivisions of diagnostic cate-
es. There is little doubt that our current taxonomy of cancer still
ps together molecularly distinct diseases with distinct clinical
notypes. Molecular heterogeneity within individual cancer
gnostic categories is already evident in the variable presence of
omosomal translocations, deletions of tumour suppressor genes
numerical chromosomal abnormalities. The classification of
nan cancer is likely to become increasingly more informative
clinically useful as more detailed molecular analyses of the
ours are conducted.

The challenge of cancer diagnosis



Diffuse large B-cell lymphoma is the most common subtype of non-Hodgkin's lymphoma. With current treatments, long term survival can be achieved in only 40 of patients. There are no reliable indicat — morphological, clinical, immunohistochemical or genetic — that can be used to recognize subclasses of **DLBCL** and point to a differential therapeutic approach to patients.

'Lymphochip', a microarray carrying 18,0 clones of complementary DNA designed monitor genes involved in normal a abnormal lymphocyte development.

What type of cancer?

What is the underlying molecular basis?

What is the optimal treatment?

ox 1: Gene-expression profiling with microarrays

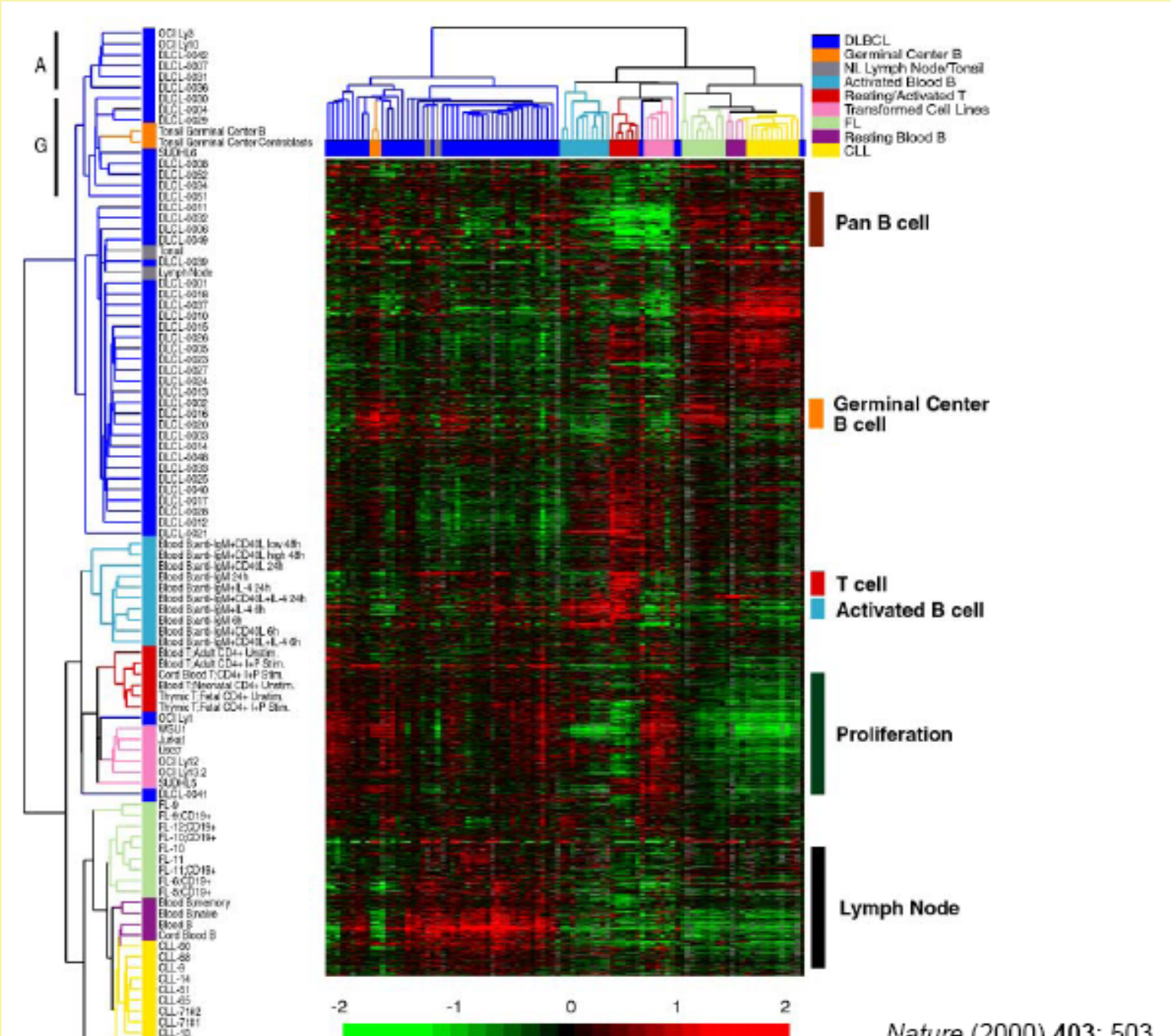
agine a 1-cm^2 chessboard. tead of 64 squares, it has ousands, each containing A from a specific gene. This a DNA microarray. The tivity of each gene on the croarray can be compared two populations of cells (A d B).

When a gene is expressed makes a transcript, and the ole population of these oducts from a cell can be tagged with a fluorescent dye (say, red for the A cells, green for the B cells). The microarray is bathed in a mixture of the red and green transcripts. Those that originate from a specific gene will bind to that gene on the microarray, turning red, green or somewhere in between, depending on the relative numbers of transcripts in the two cell types.

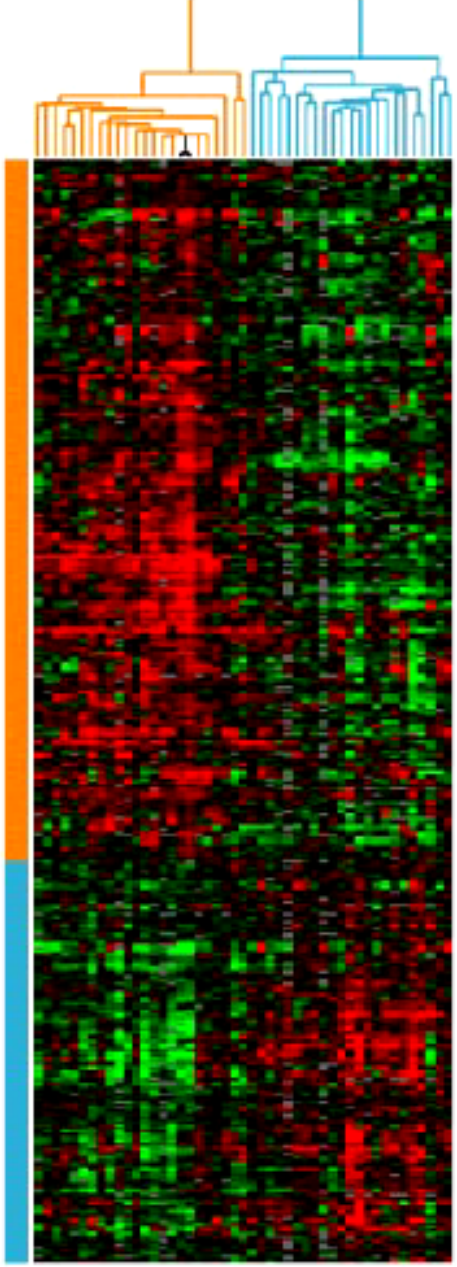
So the microarray provides a snapshot of gene activity fo thousands of genes. Data from many experiments can be compared and genes that have consistent patterns of activity can be grouped or clustered. In this way, genes that characterize a particular cell state, such as malignanc can be identified — so providing new information about the biology of the cell state.

Mark Patters

Hierarchical clustering of gene expression data (as ratios)

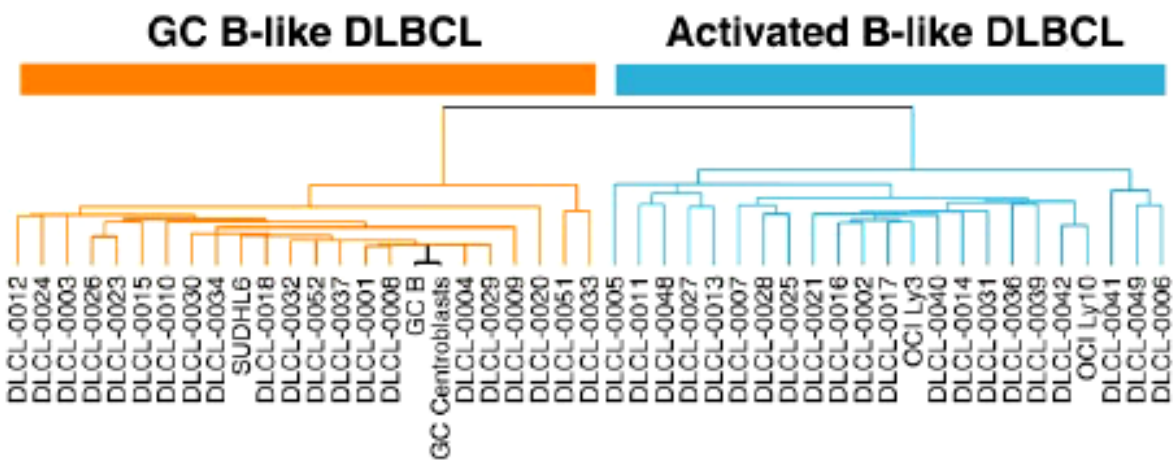


Nature (2000) **403**: 503

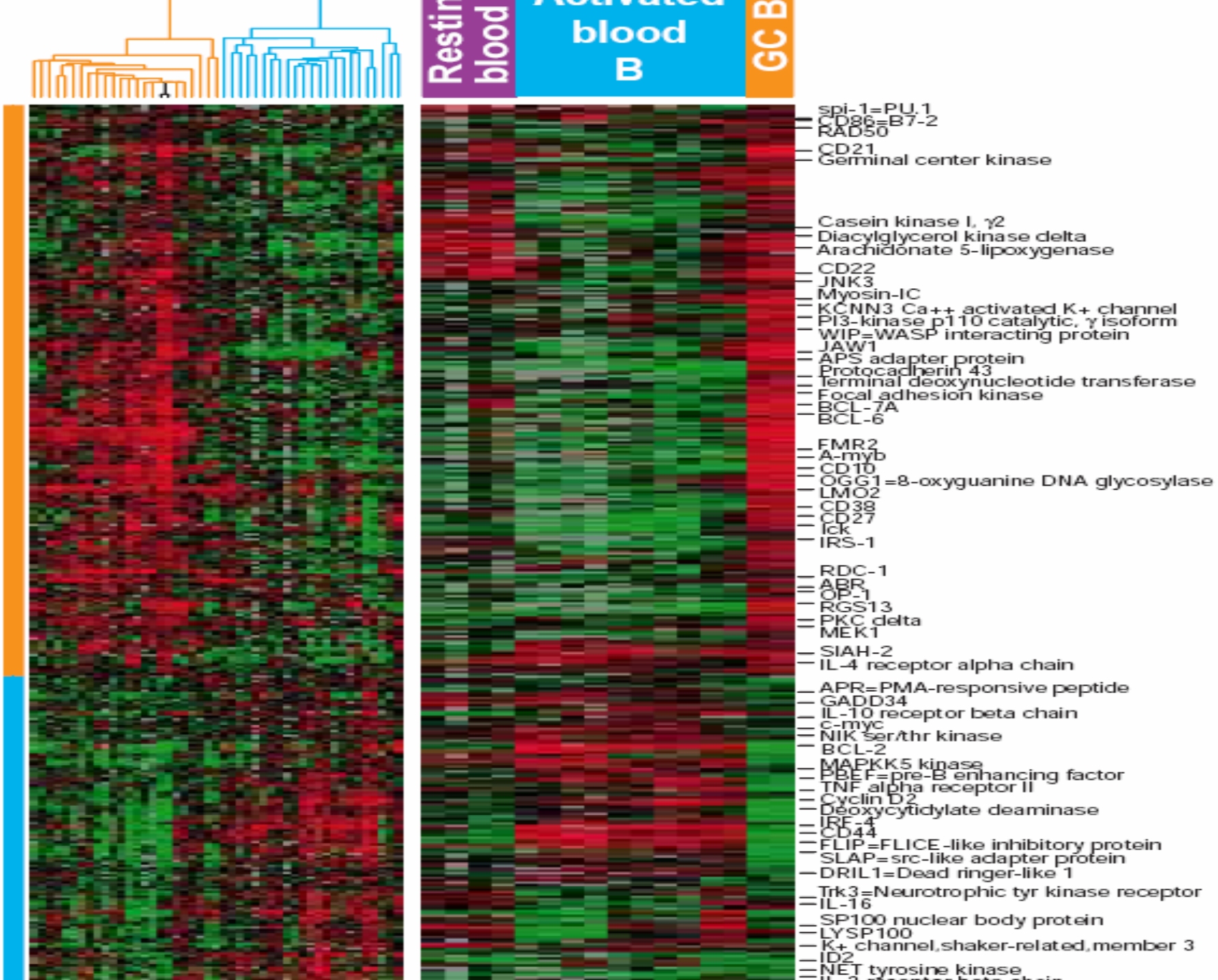


Clustering of tumour samples from cancer patients can be used for molecular classification of cancers. This may be useful for diagnosis and treatment

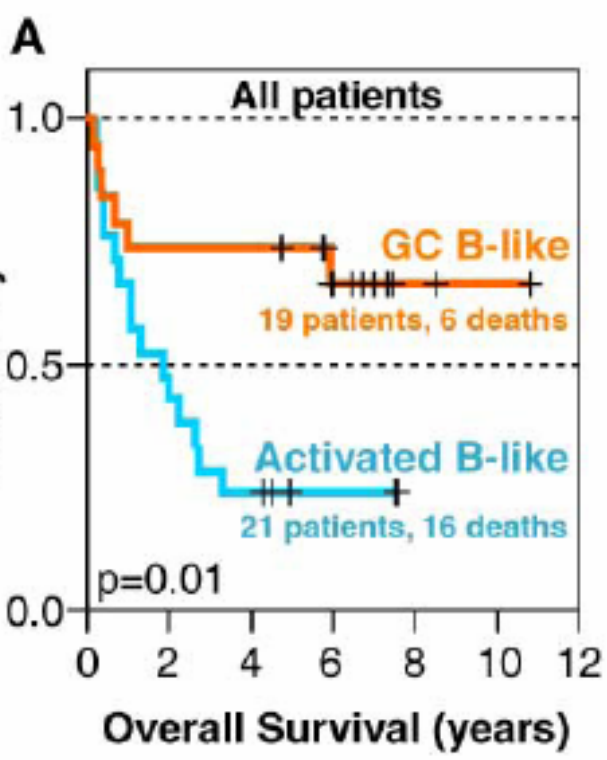
Subtypes of Diffuse Large B-Cell Lymphoma (DLBCL)

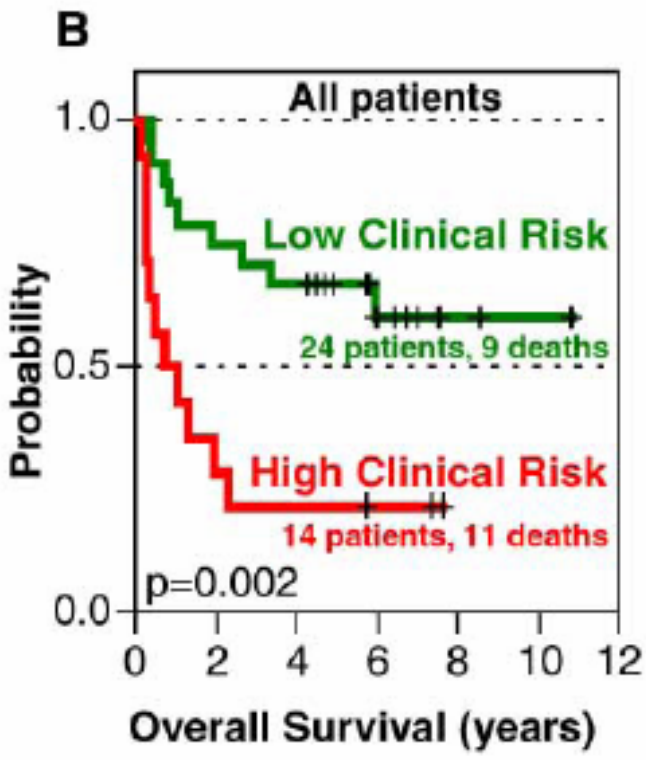


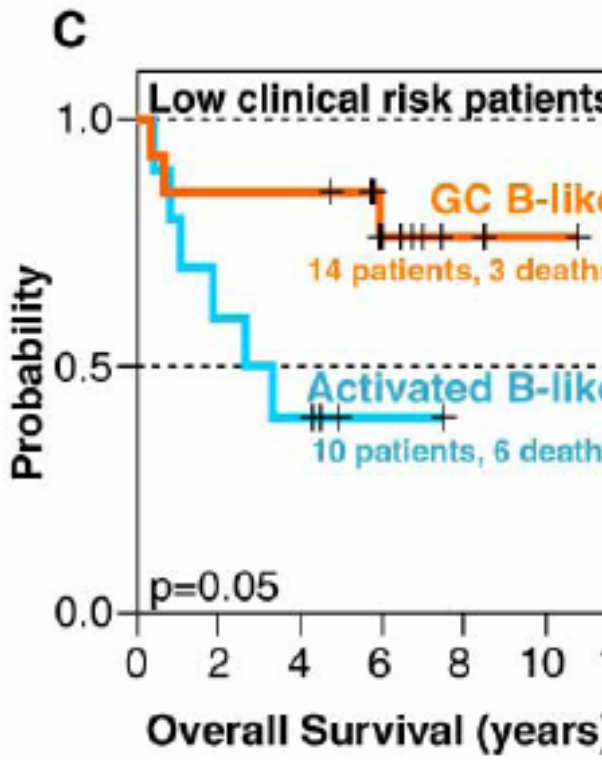
Nature (2000) **403**: 503


Resting blood B
Activated blood B
GC B
spi-1=PU.1
CD86=B7-2
RAD50
CD21
Germinal center kinase
Casein kinase I, γ2
Diacylglycerol kinase delta
Arachidonate 5-lipoxygenase
CD22
JNK3
Myosin-IC
KCNN3 Ca++ activated K+ channel
PI3-kinase p110 catalytic, γ isoform
WIP=WASP interacting protein
JAW1
APS adapter protein
Protocadherin 43
Terminal deoxynucleotide transferase
Focal adhesion kinase
BCL-7A
BCL-6
FMR2
A-myb
CD10
OGG1=8-oxyguanine DNA glycosylase
LMO2
CD38
CD27
lck
IRS-1
RDC-1
ABR
OP-1
RGS13
PKC delta
MEK1
SIAH-2
IL-4 receptor alpha chain
APR=PMA-responsive peptide
GADD34
IL-10 receptor beta chain
c-myc
NIK ser/thr kinase
BCL-2
MAPKK5 kinase
PBEF=pre-B enhancing factor
TNF alpha receptor II
Cyclin D2
Deoxycytidylate deaminase
IRF-4
CD44
FLIP=FLICE-like inhibitory protein
SLAP=src-like adapter protein
DRIL1=Dead ringer-like 1
Trk3=Neurotrophic tyr kinase receptor
IL-16
SP100 nuclear body protein
LYSP100
K+ channel,shaker-related,member 3
ID2
NET tyrosine kinase

sing "clustering analysis," Alizadeh *et al.* could separate DLBCL into
wo categories, which had marked differences in overall survival of the
atients concerned. The gene expression signatures of these subgrou
orresponded to distinct stages in the differentiation of B cells, the typ
f lymphocyte that makes antibodies.

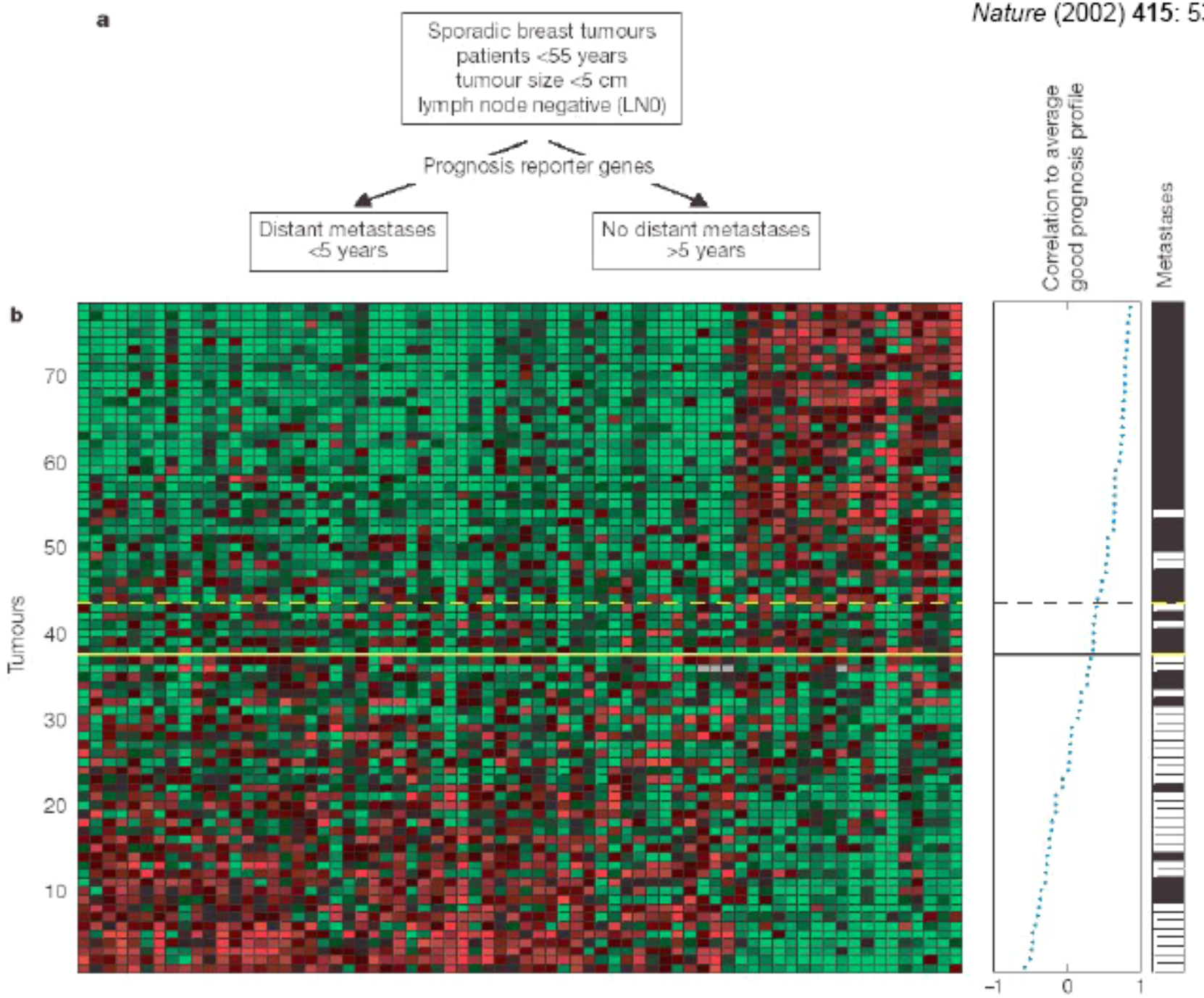






Identifying a gene expression signature for breast cancer metastasis

Nature (2002) **415**: 530



l 440|30 March 2006|doi:10.1038/nature04532

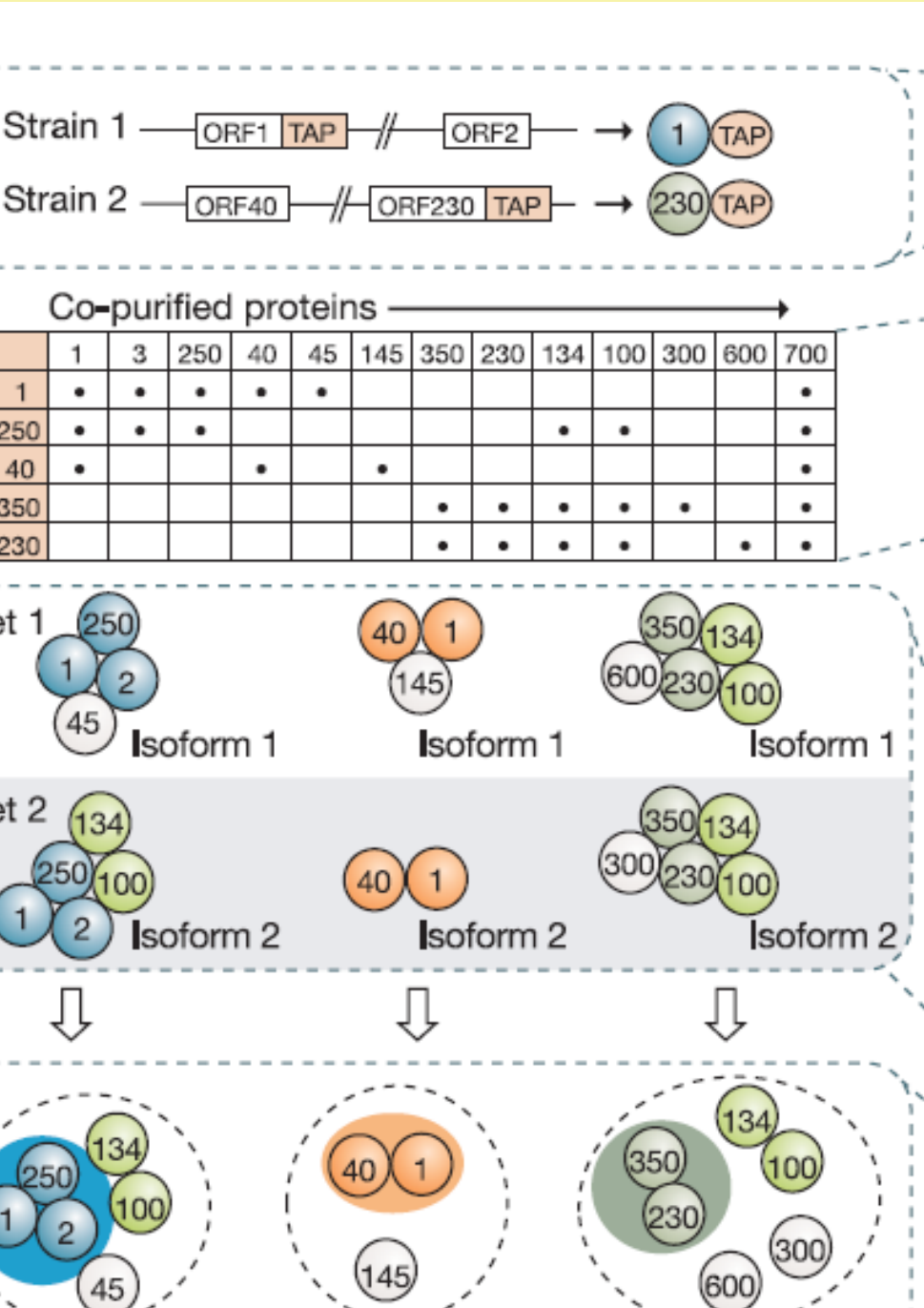
The Interactome

Proteome survey reveals modularity of he yeast cell machinery

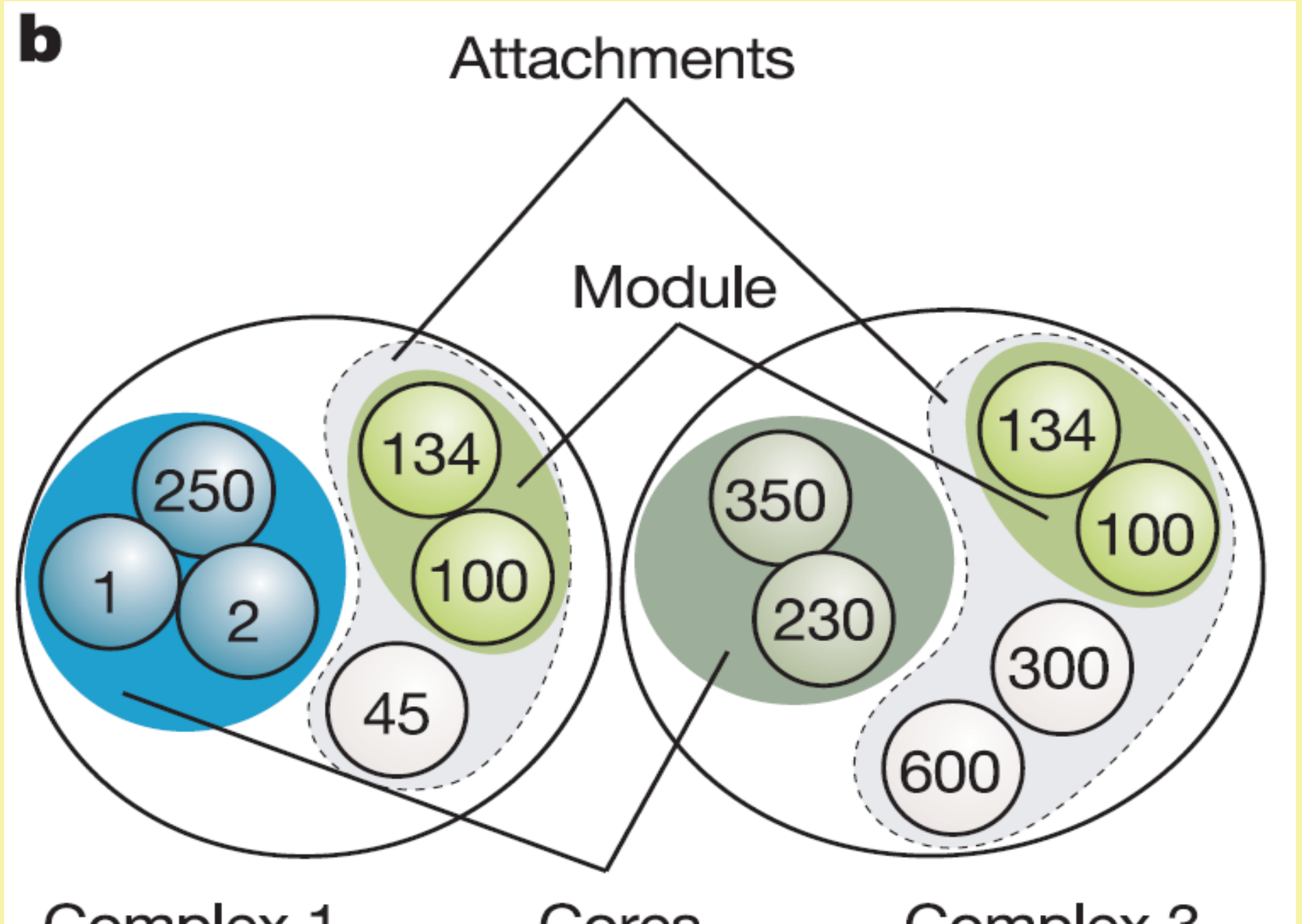
nne-Claude Gavin[1]*†, Patrick Aloy[2]*, Paola Grandi[1], Roland Krause[1,3], Markus Boesche[1], Martina Marzioch[1], ristina Rau[1], Lars Juhl Jensen[2], Sonja Bastuck[1], Birgit Dümpelfeld[1], Angela Edelmann[1], Marie-Anne Heurtier erena Hoffman[1], Christian Hoefert[1], Karin Klein[1], Manuela Hudak[1], Anne-Marie Michon[1], algorzata Schelder[1], Markus Schirle[1], Marita Remor[1], Tatjana Rudi[1], Sean Hooper[2], Andreas Bauer[1], ewis Bouwmeester[1], Georg Casari[1], Gerard Drewes[1], Gitte Neubauer[1], Jens M. Rick[1], Bernhard Kuster[1], eer Bork[2], Robert B. Russell[2] & Giulio Superti-Furga[1,4]

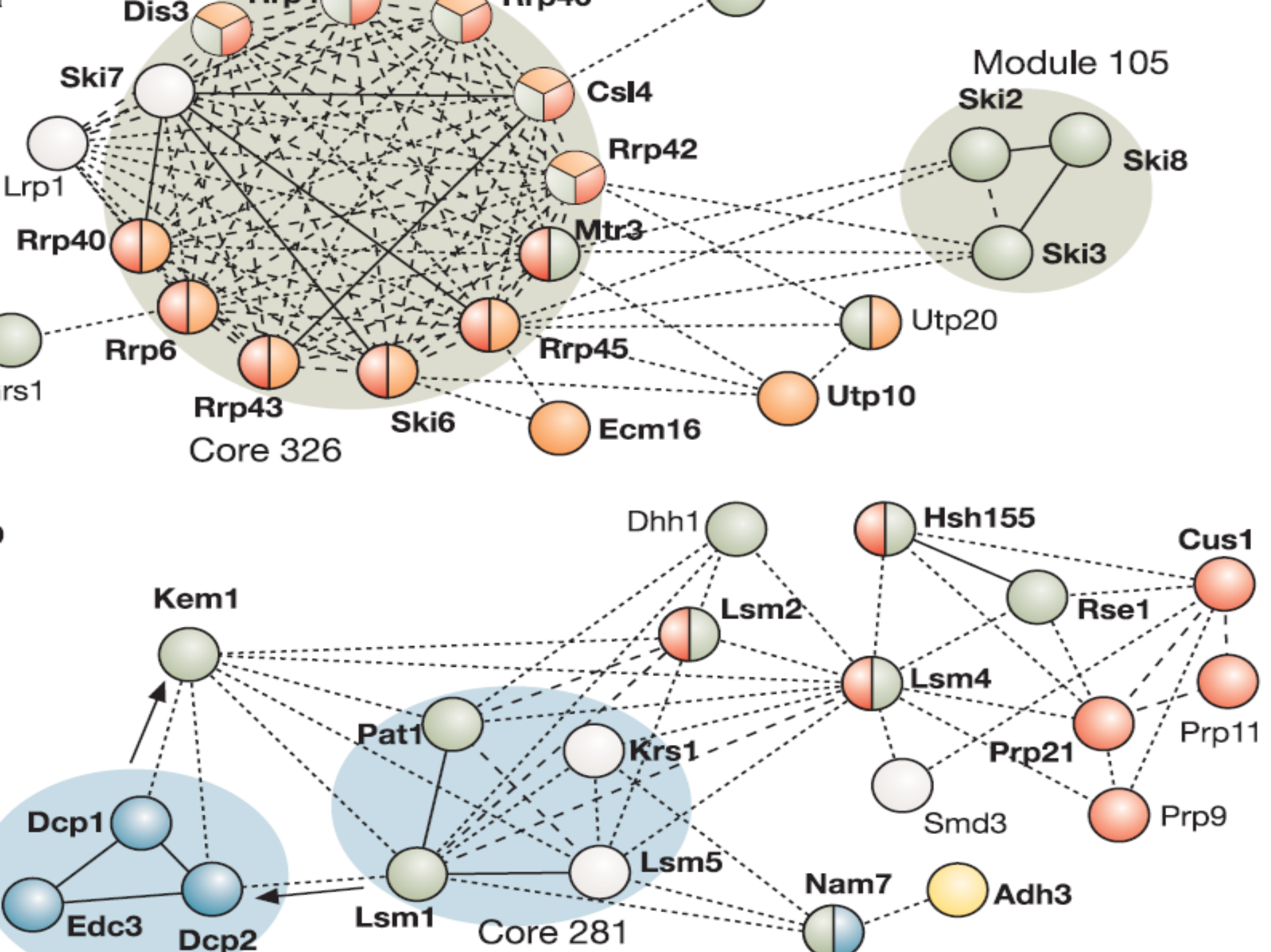
otein complexes are key molecular entities that integrate multiple gene products to perform cellular functions. Here e report the first genome-wide screen for complexes in an organism, budding yeast, using affinity purification and ma ectrometry. Through systematic tagging of open reading frames (ORFs), the majority of complexes were purified veral times, suggesting screen saturation. The richness of the data set enabled a *de novo* characterization of the mposition and organization of the cellular machinery. The ensemble of cellular proteins partitions into 491 complexes which 257 are novel, that differentially combine with additional attachment proteins or protein modules to enable a versification of potential functions. Support for this modular organization of the proteome comes from integration wi ailable data on expression, localization, function, evolutionary conservation, protein structure and binary interactions is study provides the largest collection of physically determined eukaryotic cellular machines so far and a platform fo

Architecture and Modularity of Complexes

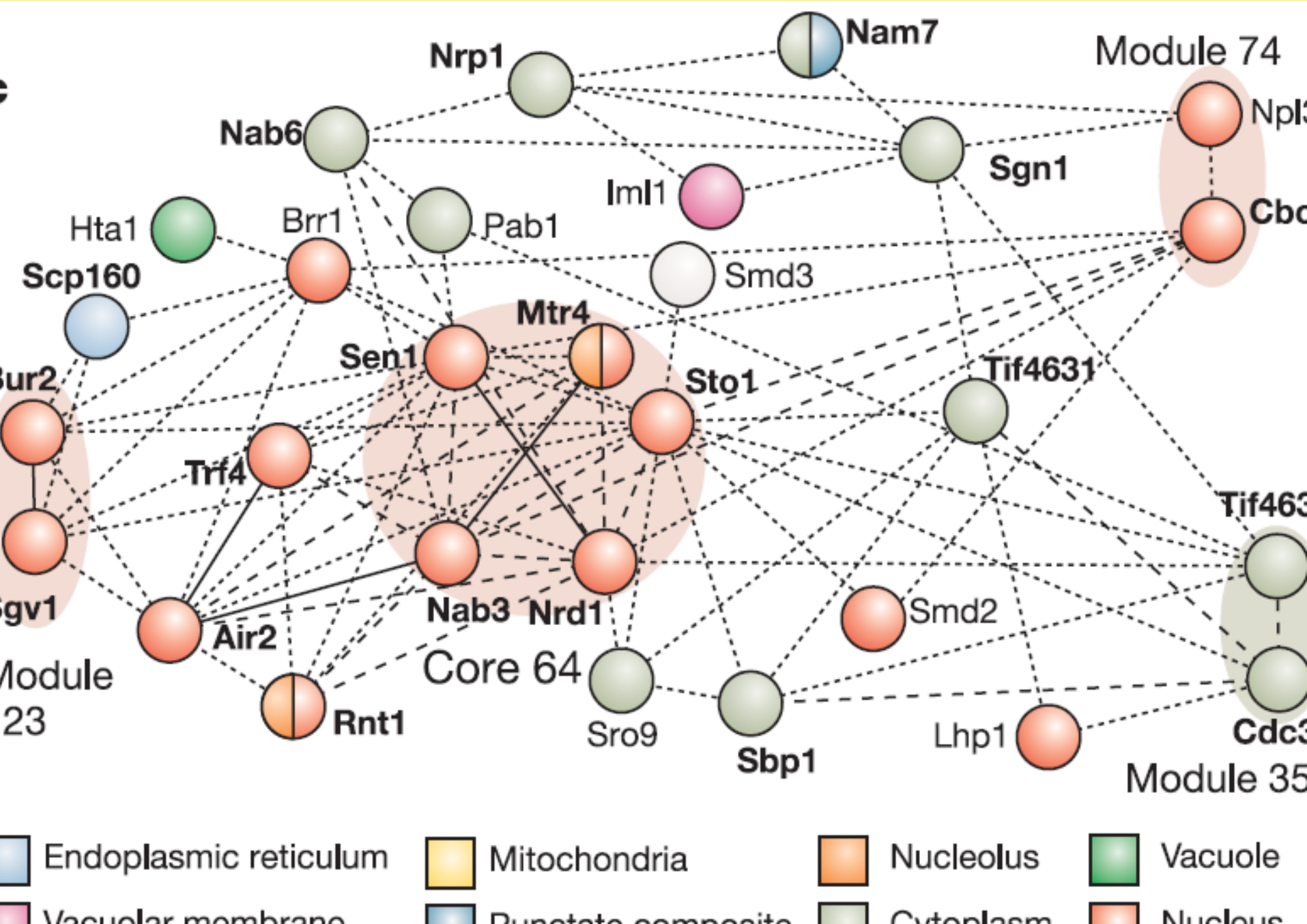


Strategy		External dataset
Yeast genome	6,466 ORFs	MIPS/SGD 2002 genome annotation
TAP cassette integration	5,474 (85%)	
TAP fusion expression	3,206 (59%)	
Overall purifications	3,206	Proteome size: Ghaemmaghami *et al.*[20] Huh *et al.*[20] Washburn *et al.*[21]
Successful purifications	1,993 (62%)	
MS protein identifications	2,760 (58%)	
'Social affinity' scoring (log odds)	1,159,003 pairs	
Iterative clustering	1,784 sets of complexes	Set of known complexes: Aloy *et al.*[14]
Selection (accuracy/coverage ≥ 70%)	92 sets of complexes	
Complexes	491	Set of known complexes: MIPS (217) PubMed (62)
Cores	478	
Modules	147	
Isoforms	5,488	

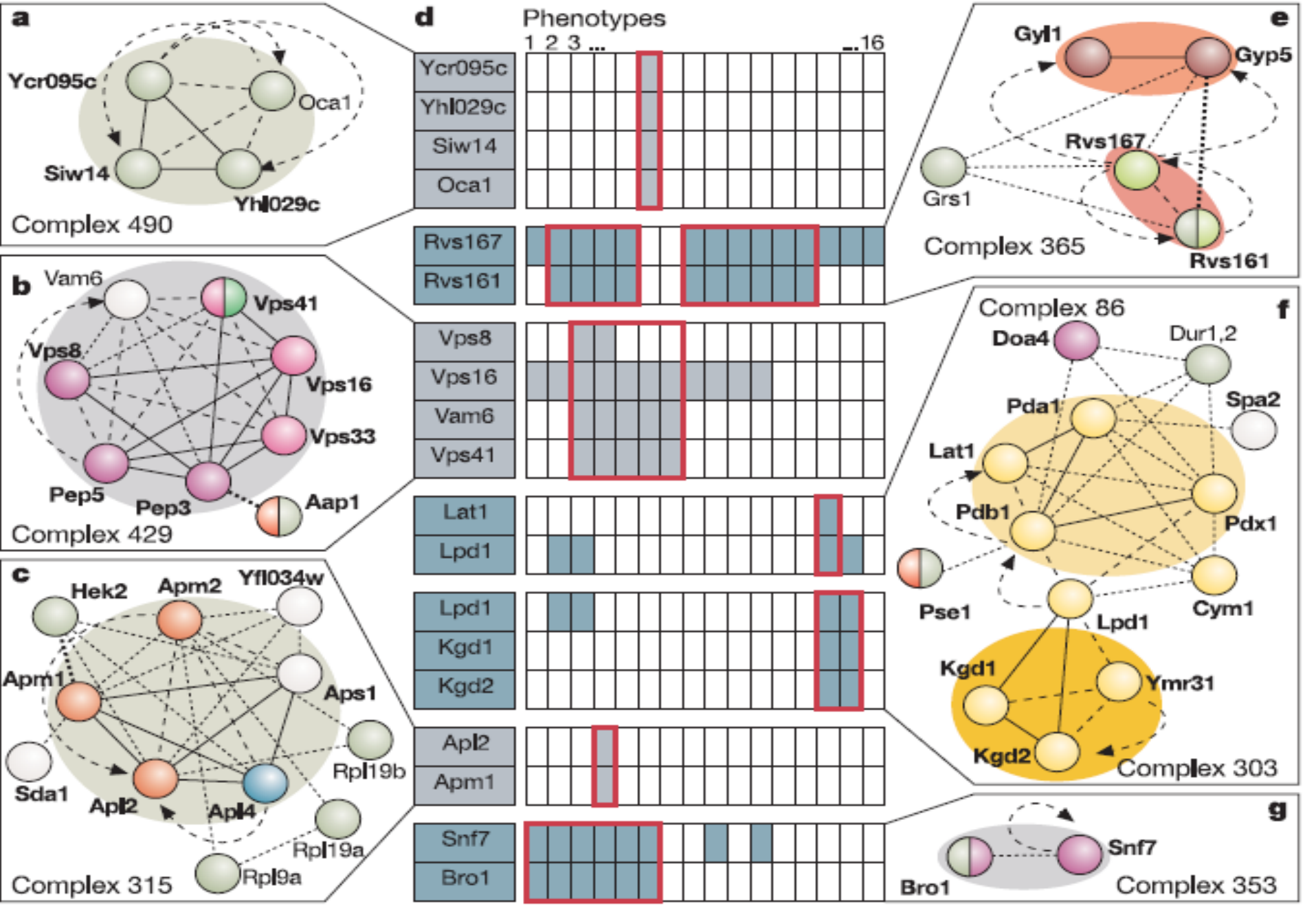

b
Attachments
Module
250
1
2
134
100
45
350
230
134
100
300
600
Complex 1
Cores
Complex 2


Dis3
Ski7
Csl4
Module 105
Ski2
Ski8
Lrp1
Rrp42
Rrp40
Mtr3
Ski3
Utp20
Rrp6
Rrp45
Utp10
Rrp43
Ski6
Ecm16
Core 326
Dhh1
Hsh155
Cus1
Kem1
Lsm2
Rse1
Lsm4
Pat1
Krs1
Prp21
Prp11
Smd3
Prp9
Dcp1
Lsm5
Nam7
Adh3
Edc3
Dcp2
Lsm1
Core 281

Architecture and Modularity of Complexes



Phenotype Data Mapped to Complexes



Systems Biology Approach