“OMICs” – High throughput (non biased) Datenerzeugung

1. Genome und Genomanalyse (Genomics, Metagenomics)
2. Transkriptomanalyse (Transcriptomics)

---

3. Proteomanalyse (Proteomics)
4. Andere Analysemethoden (Metabolomics)

Voraussetzung für „omics“

- Vorhandensein von möglichst der kompletten Genomsequenz

Relevanz des Zielorganismus

- als Modellorganismus (Beispiele: *E. coli*, Hefe, *Synechocystis*, Zebrafisch, *Caenorhabditis elegans*, *Drosophila melanogaster*, Ratte, Maus, *Arabidopsis thaliana*, *Physcomitrella patens*)

- aus ökonomischen und humanitären Aspekten (Beispiele: pathogene Mikroben, Mensch, einige Kulturpflanzen)

Humangenomprojekt – HUGO vs. Craig Venter

Sequenzierungsstrategien - Metagenomics

Shot-gun Sequenzierung von Zufallsfragmenten aus Umweltproben bzw. von Lebensgemeinschaften
bis hin zur Zusammenstellung kompletter bakterieller Genome

THE METAGENOMICS PROCESS

Extract all DNA from microbial community in sampled environment

DETERMINE WHAT THE GENES ARE
(Sequence-based metagenomics)

- Identify genes and metabolic pathways
- Compare to other communities
- and more…

DETERMINE WHAT THE GENES DO
(Function-based metagenomics)

- Screen to identify functions of interest, such as vitamin or antibiotic production
- Find the genes that code for functions of interest
- and more…

PANGENOME TERMS

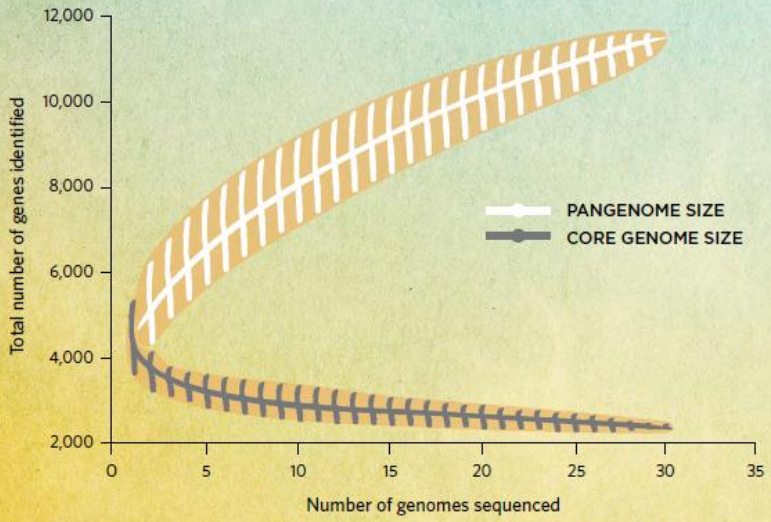
Core genes: Genes found in every individual sequenced; often involved in housekeeping functions and gene regulation

Variable genes: Genes found only in some individuals or strains of a species; often involved in adaptation to particular niches or new functions

Unique genes: Genes found only in one individual or strain

PARTIONING THE GENOME

From the sequence of a single genome, it's impossible to determine which genes are shared by all members of a species and which are possessed by only some. However, just one additional sequence offers the opportunity to distinguish shared and variable content. As more genomes are sequenced, more genes are discovered and some genes that were believed to be ubiquitous are found to be lacking from certain individuals. As a result, the estimated size of a species's core genome—the set of genes shared by all members of a species—generally decreases, and the size of the pangenome—the set of all distinct genes in the species—increases.

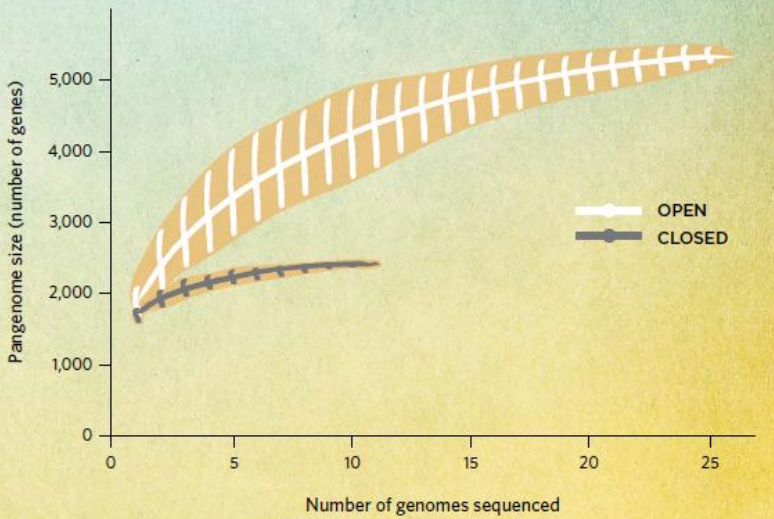


HTTP://PANGP.BIG.AC.CN

GENOME BIOL. 8:R71, 2007.

OPEN AND CLOSED

Provided genomes are sampled randomly from a population, the number of genes in the pangenome can be estimated by plotting the number of genes discovered with every new sequence. If this plot, known as a rarefaction curve, is asymptotic—i.e., after a few sequences, no more novel gene content is discovered—then the pangenome is said to be "closed" (gray). The slow-evolving bacterium *Bacillus anthracis,* for example, has a closed pangenome comprising approximately 2,900 core genes and just 85 variable genes. If the number of new genes shows no sign of plateauing, the pangenome is said to be "open," meaning that it is, at least theoretically, infinite (white). Many human pathogens, including *E. coli* and *Streptococcus agalactiae* have open pangenomes.



z.B. *Prochlorococcus marinus*, globale Population - 3×10^{27}

45 Genome: ca. 1.000 Gene im core genome, mind. 80.000 im pangenome

Bioinformatik zur Auswertung

1. **Vorhersage möglicher kodierender Strukturen (gene prediction)**

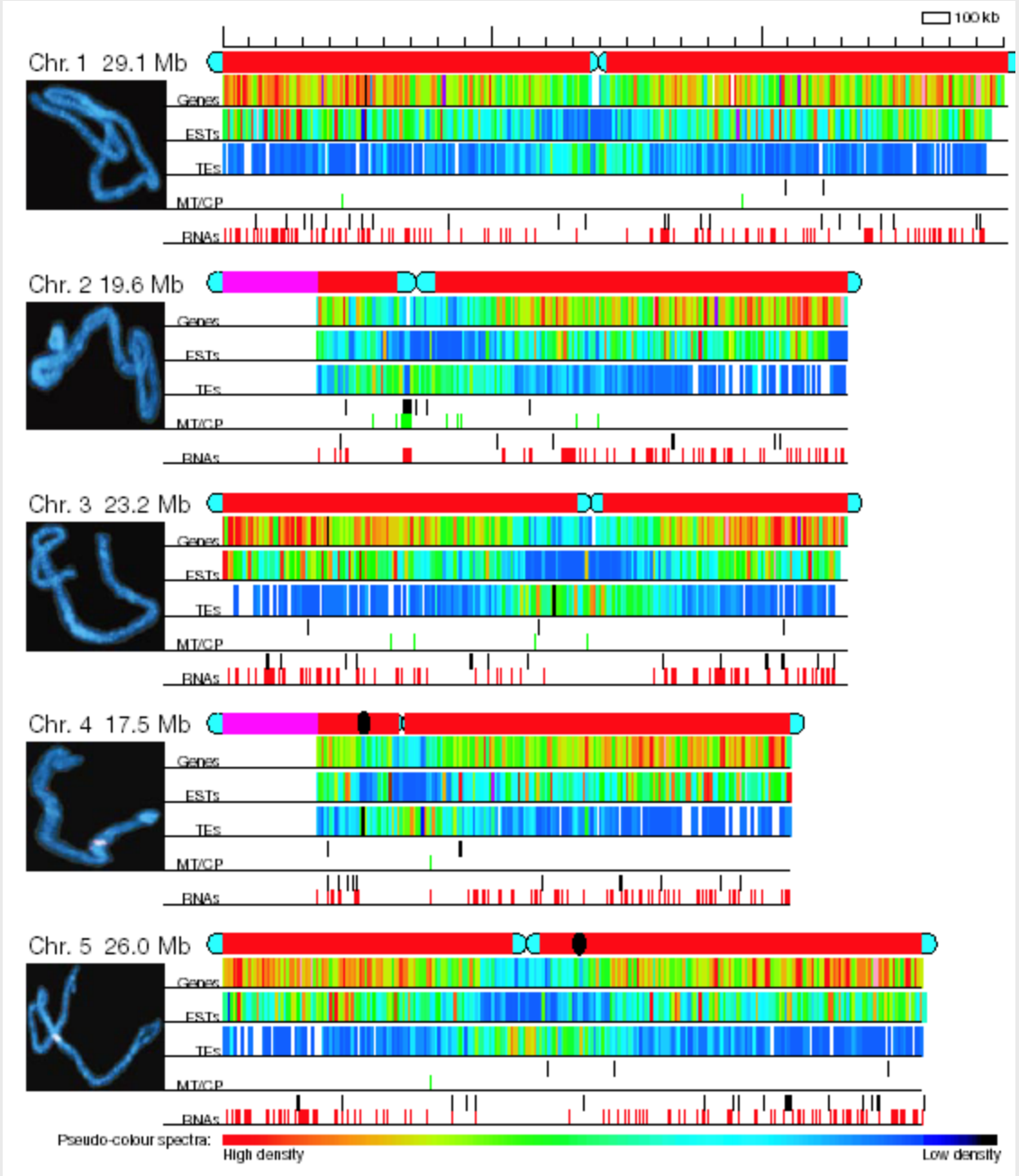
- GENSCAN
- GRAIL
- NETPLANTGENE

2. **Vergleich mit bekannten Genen (BLAST-Searches)**

- basic local alignment search tool
- verschiedene Arten: blastn, blastp, blastx, heute auch mit Clusteranalyse

http://www.ncbi.nlm.nih.gov/BLAST/

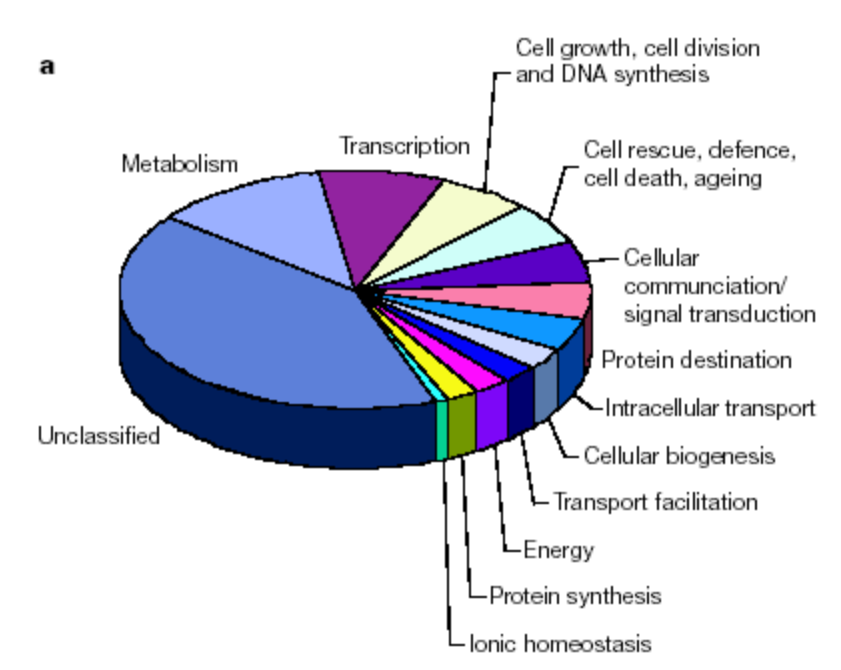
http://www.expasy.ch/


100 kb
Chr. 1 29.1 Mb
Genes
ESTs
TEs
MT/CP
RNAs
Chr. 2 19.6 Mb
Genes
ESTs
TEs
MT/CP
RNAs
Chr. 3 23.2 Mb
Genes
ESTs
TEs
MT/CP
RNAs
Chr. 4 17.5 Mb
Genes
ESTs
TEs
MT/CP
RNAs
Chr. 5 26.0 Mb
Genes
ESTs
TEs
MT/CP
RNAs
Pseudo-colour spectra:
High density
Low density

Probleme der bioinformatischen Vorhersagen

Trefferquote der Vorhersage kodierender Regionen mit 95 % zu gering

- Vorhersage über die Ähnlichkeit zu bekannten Genen zu ungenau
    - noch unbekannte Gene können so nicht gefunden werden
    - Gene werden falsch annotiert (Annotationsproblem)
- Am besten interaktive Datenbanken nutzen!!



2. Transkriptomanalyse - Transcriptomics

Ziel: Untersuchung möglichst vieler (aller) Transkripte gleichzeitig in einem Zielorganismus

Bei Prokaryoten realistisch aber auch hier Variabilität je nach Wachstumsbedingungen!

Häufig angewandte Verfahren:

a) Electronic Northerns

b) DNA-arrays

1. cDNA-Arrays
2. Oligonukleotid-Chips (Affymetrix-Arrays)

c) RNAseq – high throughput cDNA Sequencing (Solexa, 454)

cDNA- bzw. EST-Datenbanken

- experimenteller Ansatz zur Feststellung und Verifikation kodierender Bereiche
- werden fast ausschließlich bei eukaryotischen Organismen eingesetzt
- z.B. Abgleich mit Genomsequenz zur ORF-Verifizierung

Anwendungen:

1. ESTs (Expressed sequence tags)
2. Vollständige Sequenzierung von cDNA-Klonen (FL-cDNAs)

Varianten und Probleme bei der Analyse von cDNA-Datenbanken

- alternatives **Splicing**

    - häufig in Tieren, seltener in Pflanzen

- seltene **Transkripte**

    - Expression abhängig von äußeren Faktoren oder Entwicklungsstadien

    - Herstellung spezifischer cDNA-Banken

- transkribierte **Pseudogene**

    - führen zum Ausschluss dieser Regionen aus der primären Analyse

- miRNAs, siRNAs – **fehlen meist (methodische Probleme)**

    - Funktionen bisher weitgehend unbekannt

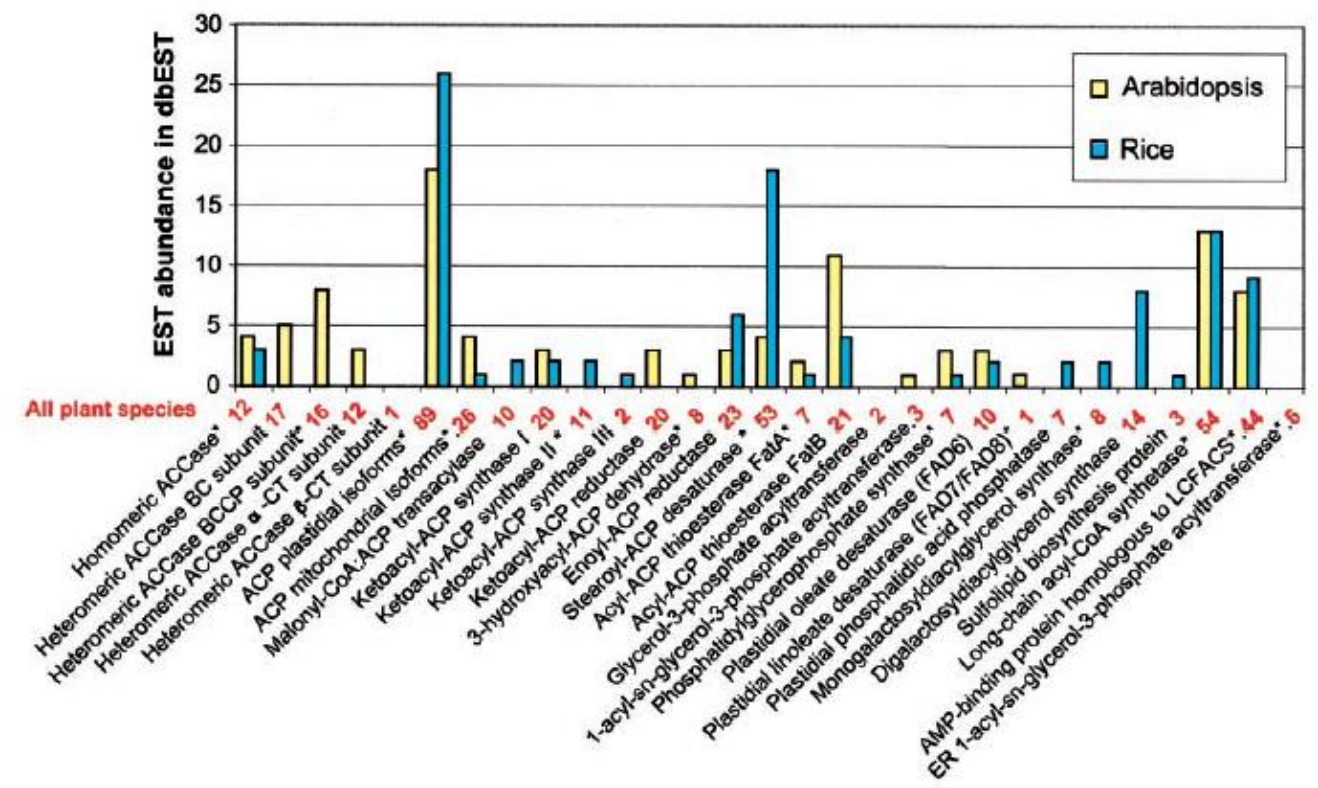
    - beteiligt an verschiedensten Prozessen durch Repression von Zielgenen

2.1. Electronic Northerns

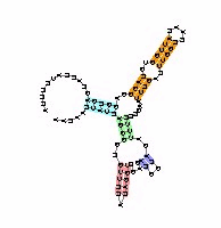
- Ermittlung der relativen Transkriptionsrate von Genen durch ihre Anzahl von ESTs

- Methode abhängig von Größe der EST-Datenbanken

- nur eingeschränkt einsetzbar und aussagefähig



2.2 Transcript”omics” – DNA-Microarray



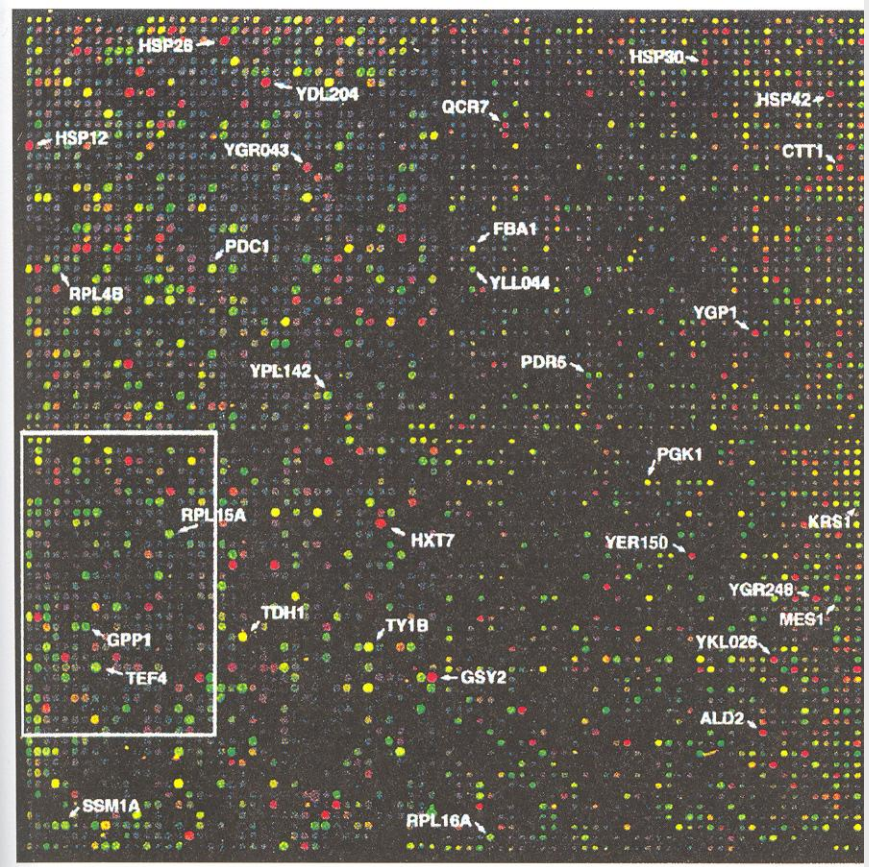
DNA-Stücke bekannter Gene werden auf einen Träger immobilisiert und durch Hybridisierung mit markierter cDNA quantitativ nachgewiesen

2.2. (c)DNA-Arrays

- DNA-Arrays bestehen aus einem inerten Trägermaterial, das mit mehreren tausend (c)DNAs beschickt ist.

- Bei (c)DNA-Arrayanalysen werden die beiden zu untersuchenden RNAs markiert und abhängig von der Markierung auf einen oder zwei Arrays hybridisiert („umgekehrter Northern“).

- Markierung erfolgt entweder radioaktiv oder mit Fluoreszenzfarbstoffen



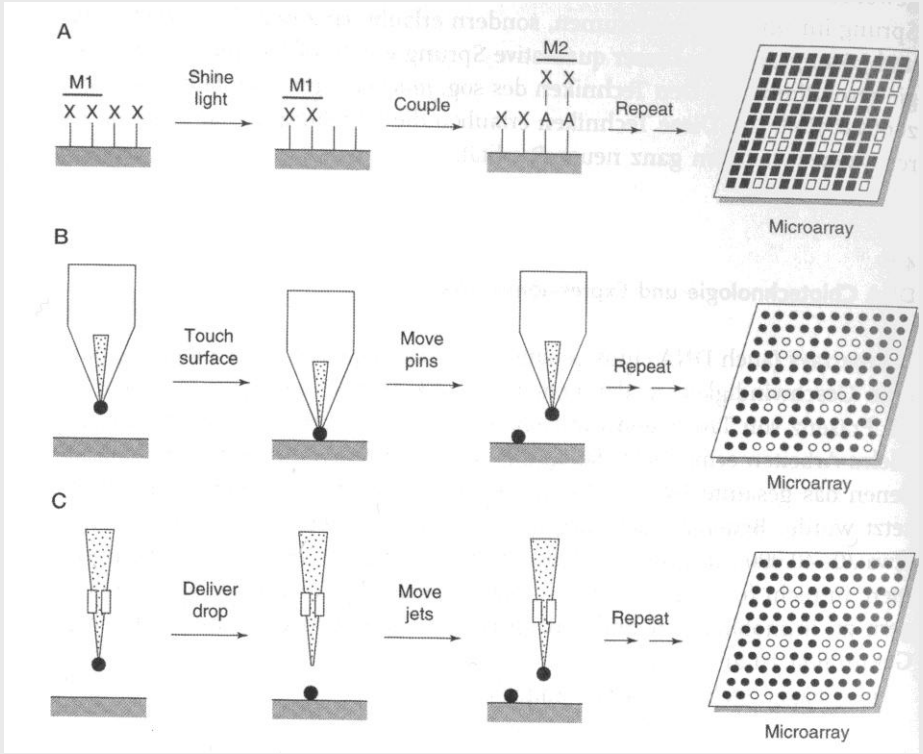
Beispiel für einen cDNA-Array von Hefe

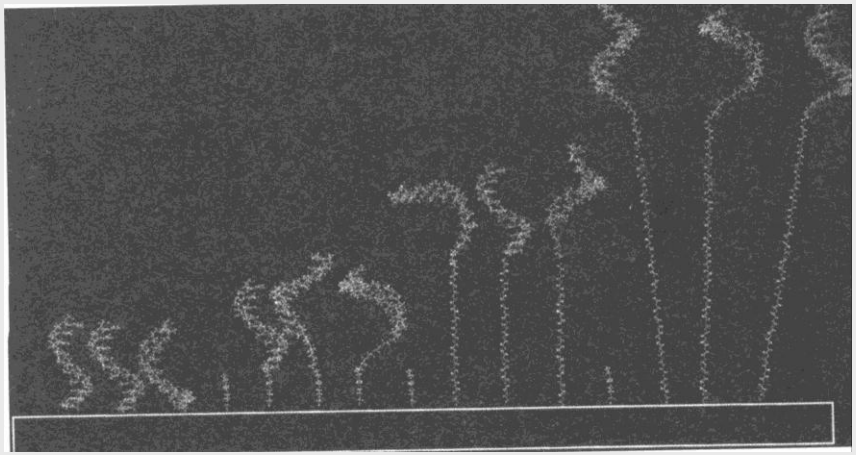
Methoden der Arrayherstellung

a) Photolithographie

b) Microspotting

c) Microspraying





- Verbesserung der Zugänglichkeit der Proben durch Spacermoleküle

Herstellung von (c)DNA-Arrays

- zunächst werden Gene des Interesses ausgewählt

- spezifische Sequenzen werden amplifiziert (PCR)

- die Produkte werden aufgereinigt und auf Membranen oder Glas (Slides) aufgebracht und immobilisiert

Verfahren der Analyse

- mRNAs von zwei verschiedenen Zuständen eines Organismus werden mit zwei verschiedenen Farbstoffen markiert

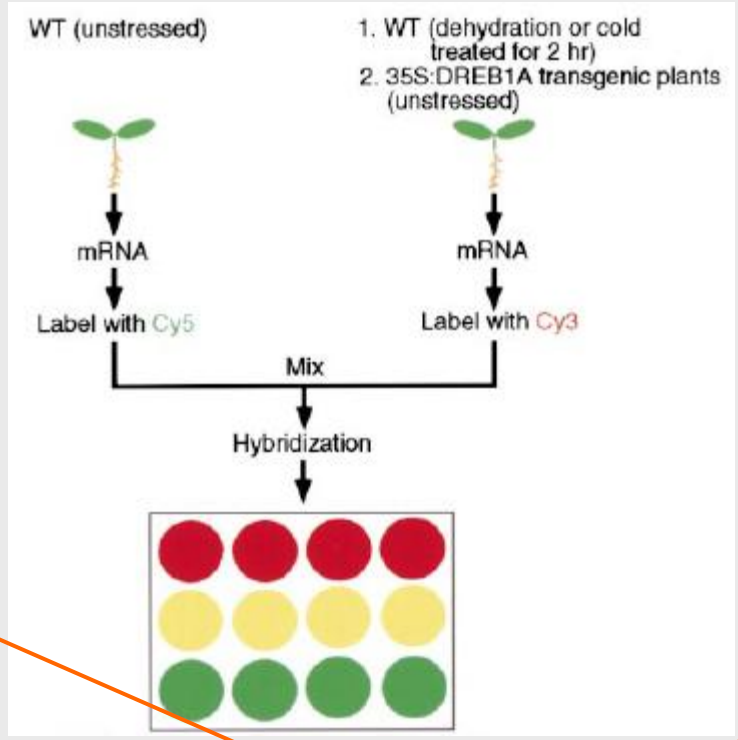
- diese Proben werden anschließend mit den Arrays hybridisiert

- eine Falschfarbendarstellung der Images der Arrays zeigt dann die Veränderung der Transkription der einzelnen Gene

Prinzip der Auswertung von cDNA-Arrays

Doppelspotting zur Absicherung

keine Expression



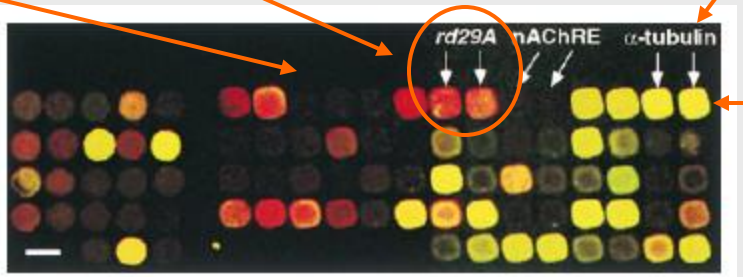


Figure 1. cDNA Microarray Analysis of Gene Expression under Cold Stress.

Tubulin als interne Kontrolle

Auslöschung des Signals durch gleich starke Expression in beiden Zuständen

Vor- und Nachteile von cDNA-Arrays

- begrenzte Anzahl von cDNAs werden benötigt (keine Kenntnis des kompletten Genoms vonnöten)

- können auch in Eigenregie hergestellt werden

- Einsatz für Spezialanwendungen möglich („custom chips“

z.B. von Agilent, „Phylochips“, „Tiling arrays“)

- Risiko der Kreuzhybridisierung

- Standardisierung und Vergleich mit anderen Chipergebnissen sehr erschwert

- **Ist immer spezifisch für einen Organismus**

Affymetrix-Arrays

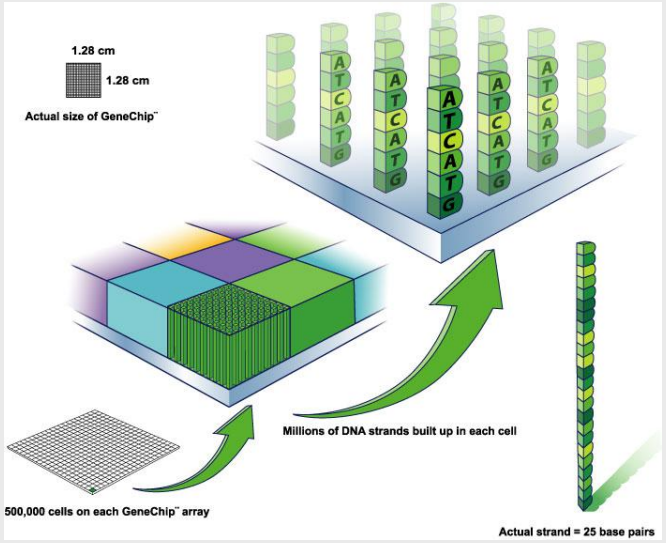
- Pro Gen existieren mehrere Sonden

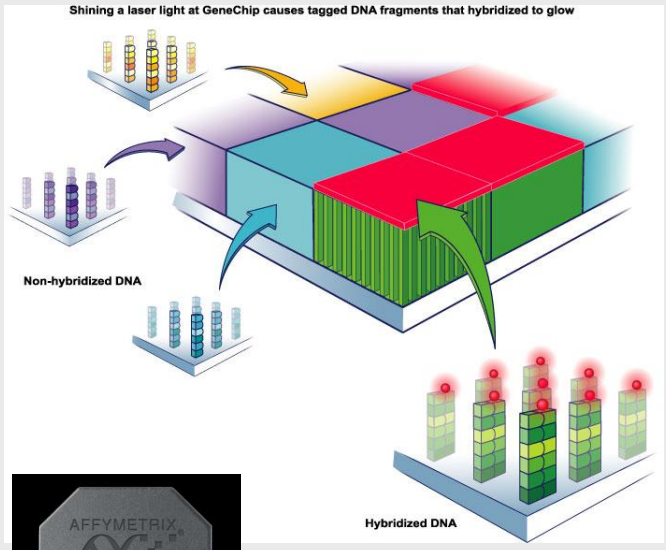
- Sonden bestehen aus maximal 25 bp

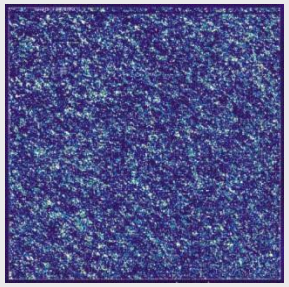
- Arrays sind extrem spezifisch, selbst Unterschiede von nur einem bp werden erkannt

- Transkripte werden in cDNA umgeschrieben, fragmentiert, markiert, und danach hybridisiert

- Nach Hybridisierung erfolgt Scanning und Auswertung







Vor- und Nachteile von Affymetrix-Arrays

- extrem spezifisch

- Gesamtgenom-Chips für eine Reihe von Organismen erhältlich

- sehr gut standardisierbar

- sehr teuer in Anschaffung und Anwendung

- **Ist immer spezifisch für einen Organismus**

Clusteranalysen

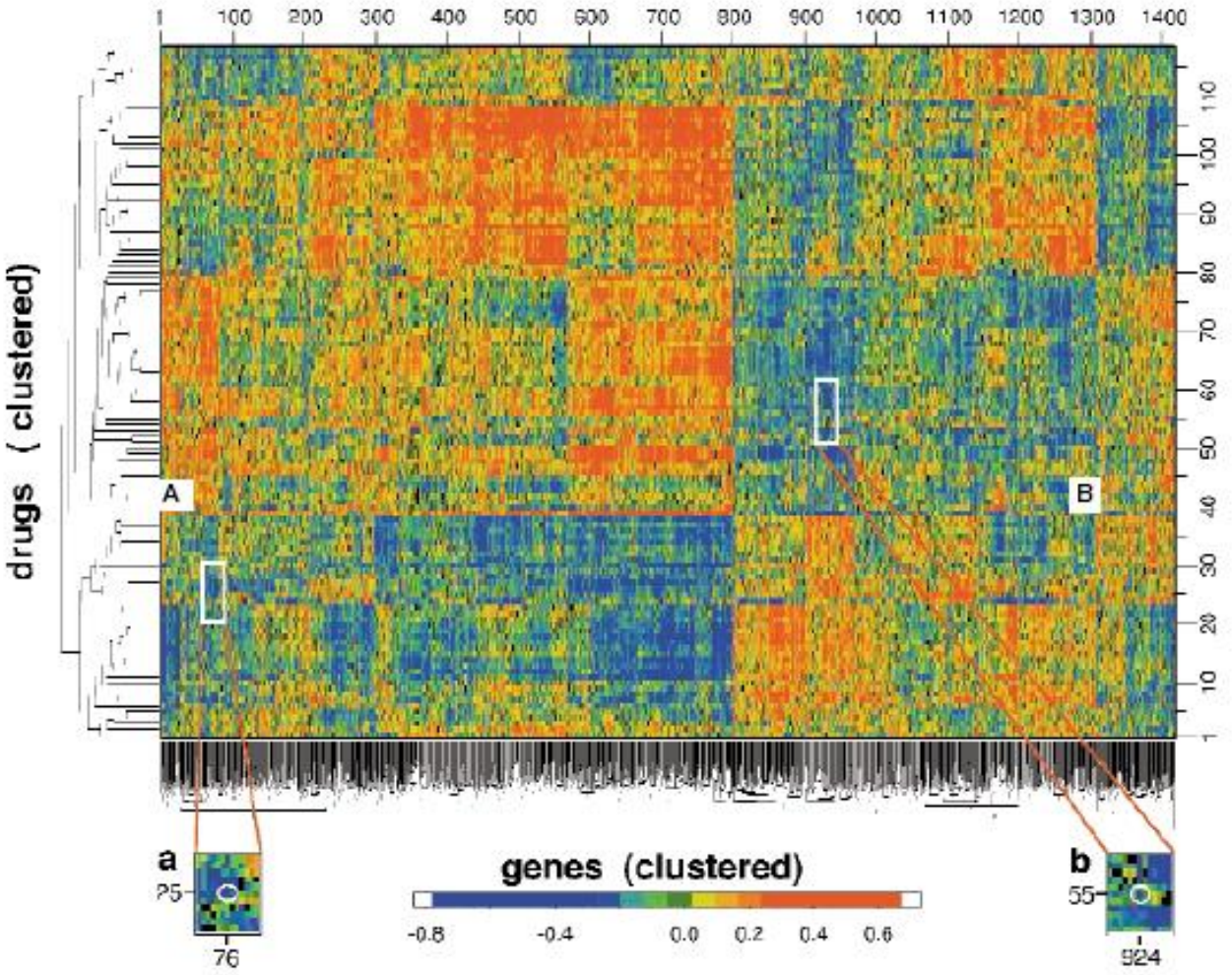
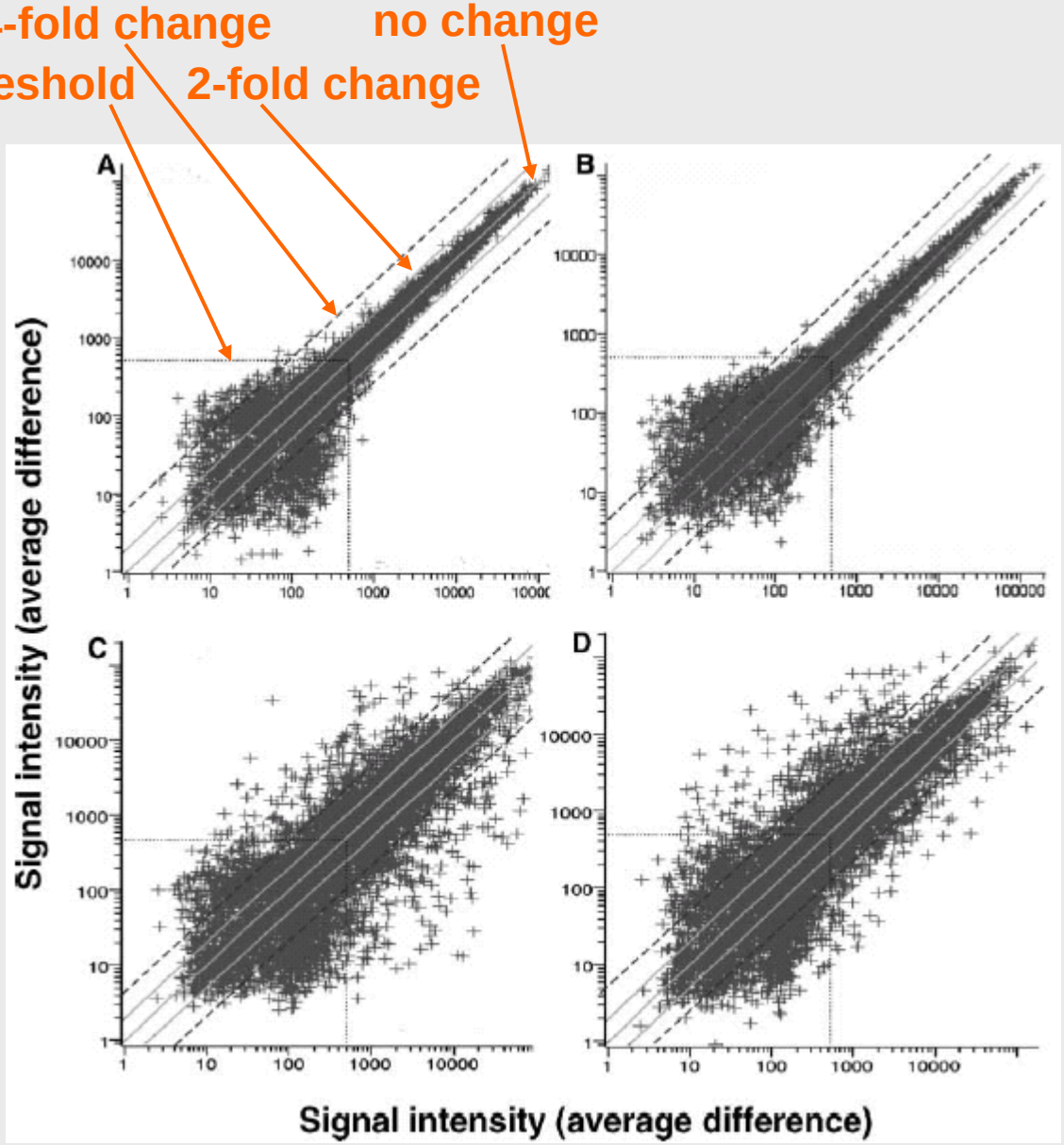


Fig. 4 CIM relating activity patterns of 118 tested compounds to the expression patterns of 1,376 genes in the 60 cell lines. Included, in addition to the gene expression levels, are data for 40 molecular targets assessed one at a time in the cells. A red point (high positive Pearson correlation coefficient) indicates that the agent tends to be more active (in the two-day SRB assay) against cell lines that express more of the gene; a blue point (high negative correlation) indicates the opposite tendency. Genes were cluster-ordered on the basis of their correlations with drugs (mean-subtracted, average-linkage clustered with correlation metric); drugs were clustered on the basis of their correlations with genes (mean-subtracted, average-linkage clustered with correlation metric). The drug cluster tree is the same as that in Fig. 3*b*, which can be consulted to identify individual drugs. A larger version of this $A \cdot T^T$ clustered correlation (ClusCorr) CIM (with the drug and gene names and the cluster trees; refs 8,28,38) is available (http://discover.nci.nih.gov). Inset A shows a magnified view of the region around the point (white circle) representing the correlation between *DPYD* (76) and 5-FU (25). Inset B is an analogous magnified view for *ASNS* (924) and the drug L-asparaginase (55).

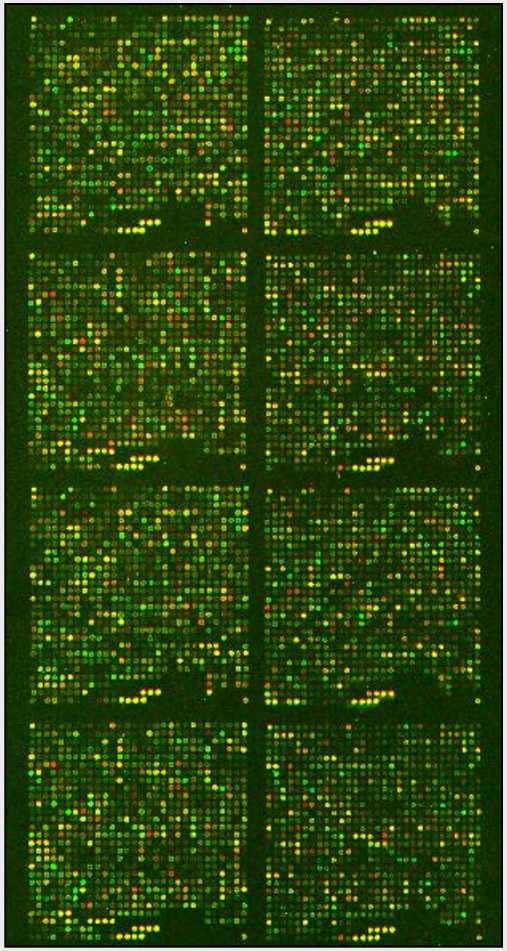
Probleme bei der Analyse von Arraydaten

- Relevanz der Daten (Signal-Noise-Ratio)
    - teilweise experimentelles „Rauschen“ stärker als tatsächliche Unterschiede
    - Verlust von Informationen als Folge (z. B. Gene mit zu geringer Expression werden nicht gemessen)
- Normalisierung der Daten (Kalibrierung der Arrays)
- Auswertung der Daten (z. B. Clusteranalyse)
- **Probenkultivierung und –Aufarbeitung!!! RNA und Anzuchtqualität**

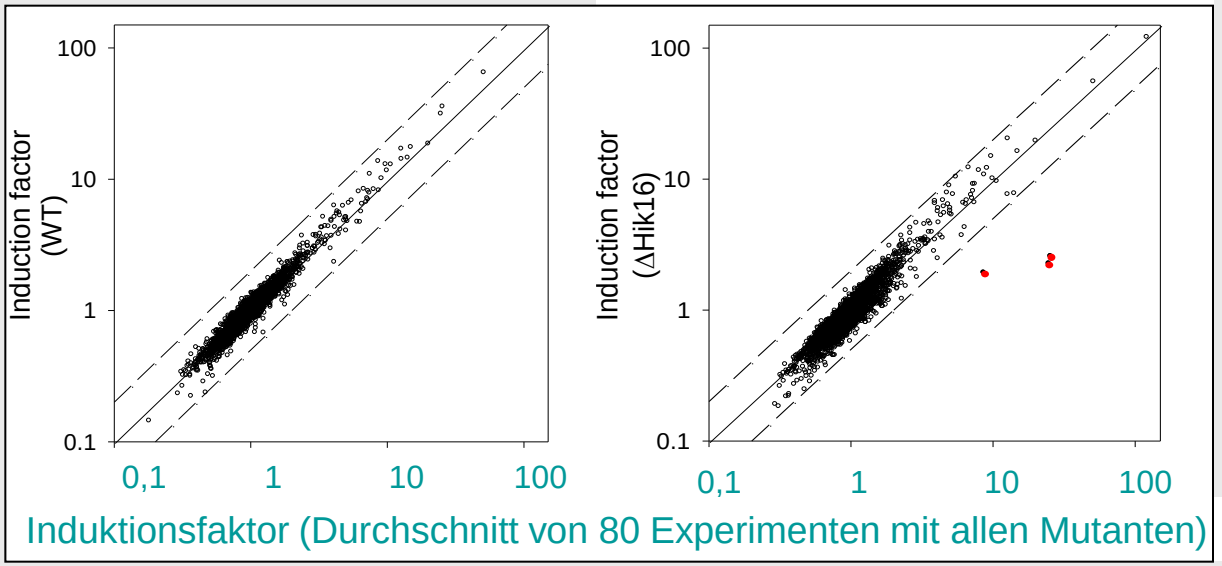
Signal-Noise-Ratio



Untersuchung der salzregulierten Genexpression bei allen 43 *hik*-Mutanten von *Synechocystis* durch Transcriptomics

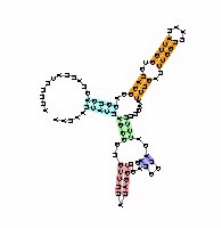


DNA-Microarray von TaKaRa mit 3079 *Synechocystis* Genen



Mit dem DNA-Microarray wurde gezeigt, dass in der Mutante Hik16 drei Gene (rot) im Gegensatz zum WT nicht mehr salzreguliert sind!

2.3 Transcript”omics” - RNAseq



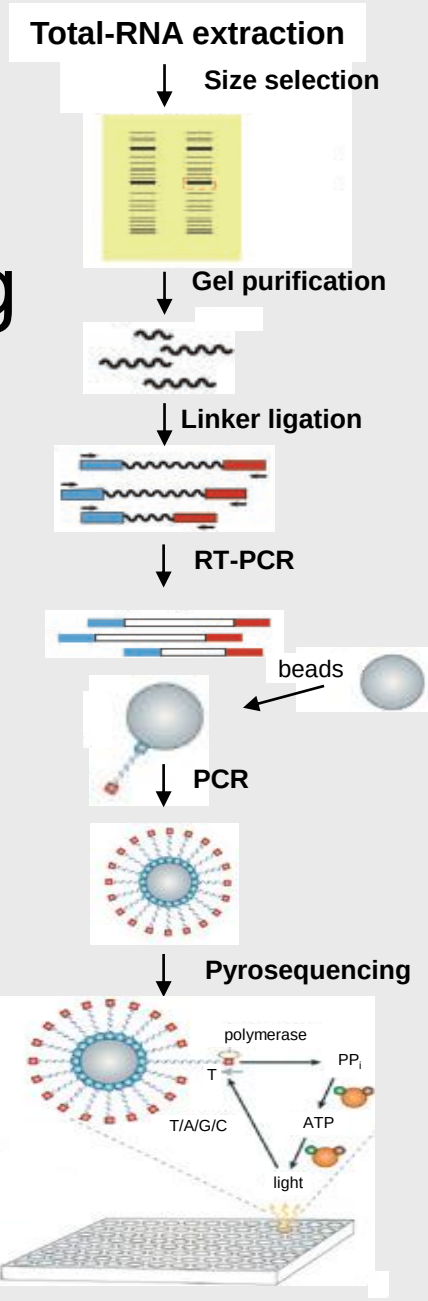
Isolation and sequence analysis of all (small) RNAs („RNAseq, dRNA-seq“, 454, Illumina, etc.)

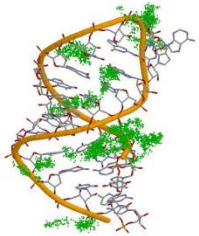
Universally applicable (not organism-specific)

Annotation quality of genome is less important, it improves it!

Direct identification of ncRNAs: "*Next Generation*" sequencing („transcriptomics")

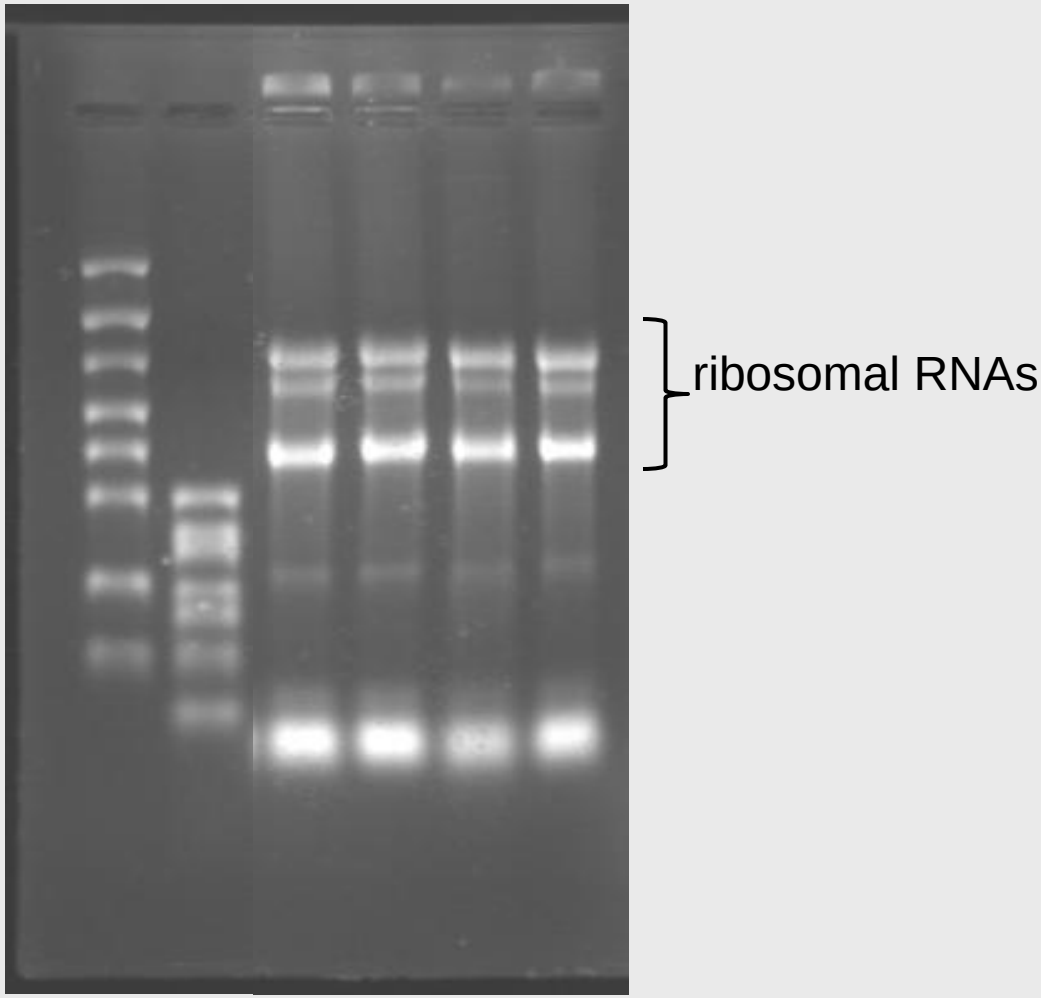
e. g. 454, Solexa HiSeq2000
deep sequencing



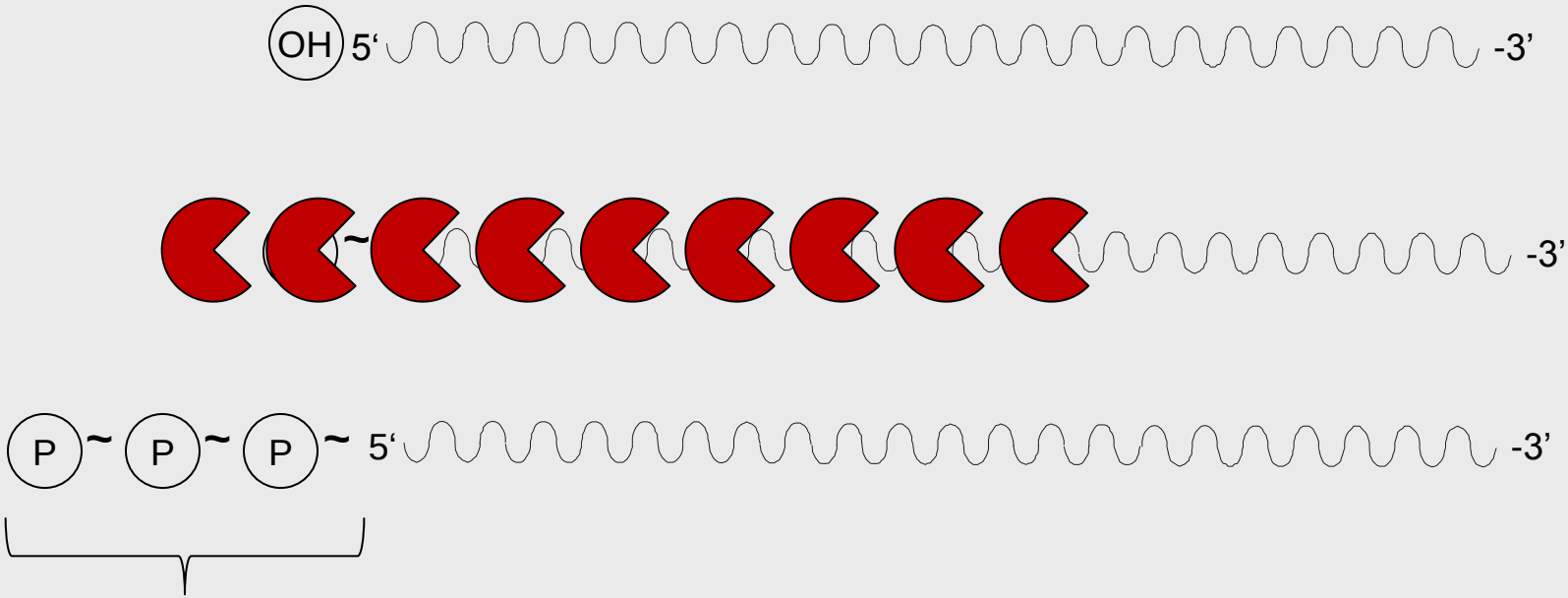


From: Hüttenhofer & Vogel NAR 2006,
Medini et al. Nature 2008.

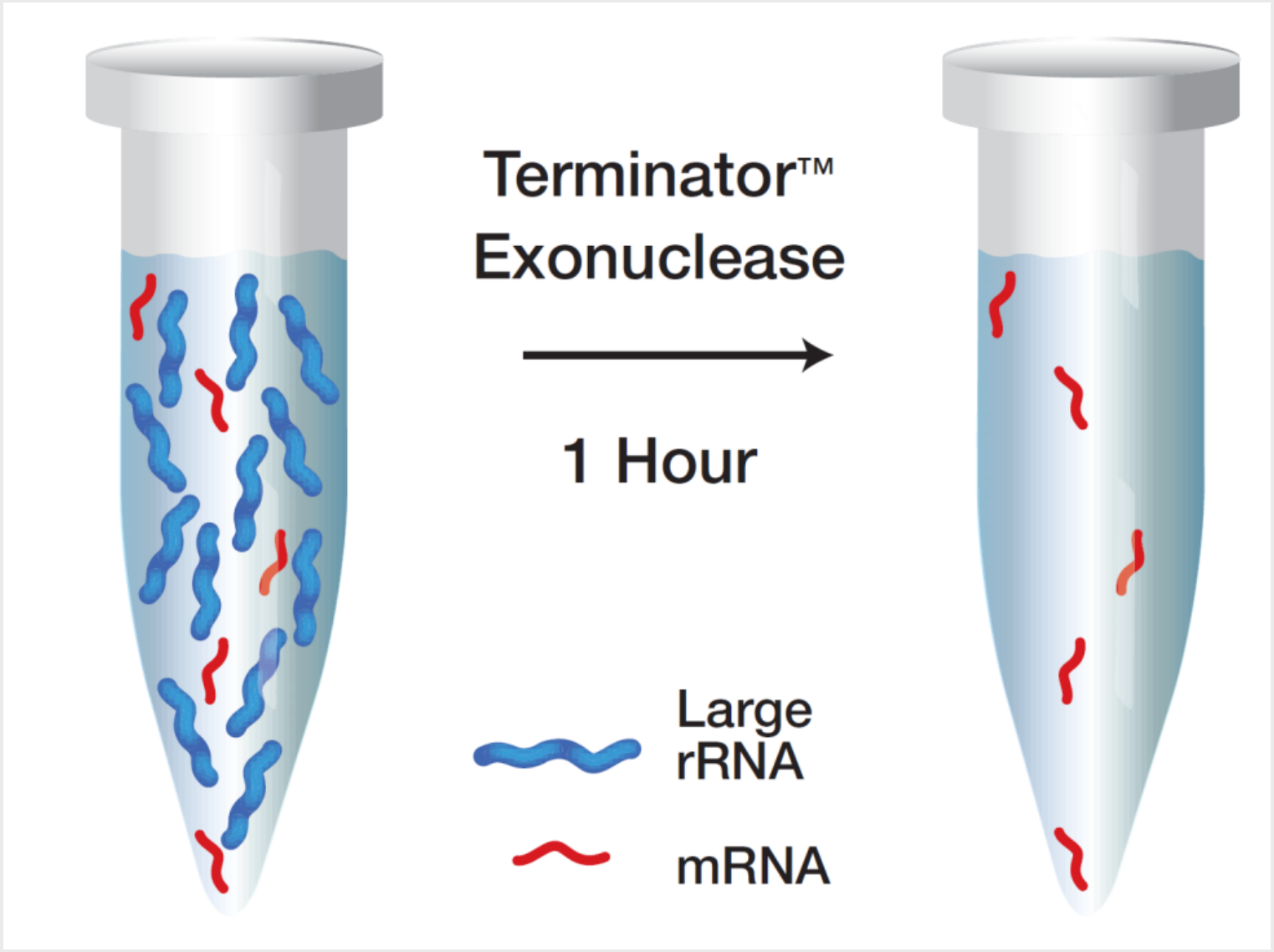
RNAseq mainly sequences ribosomal RNA



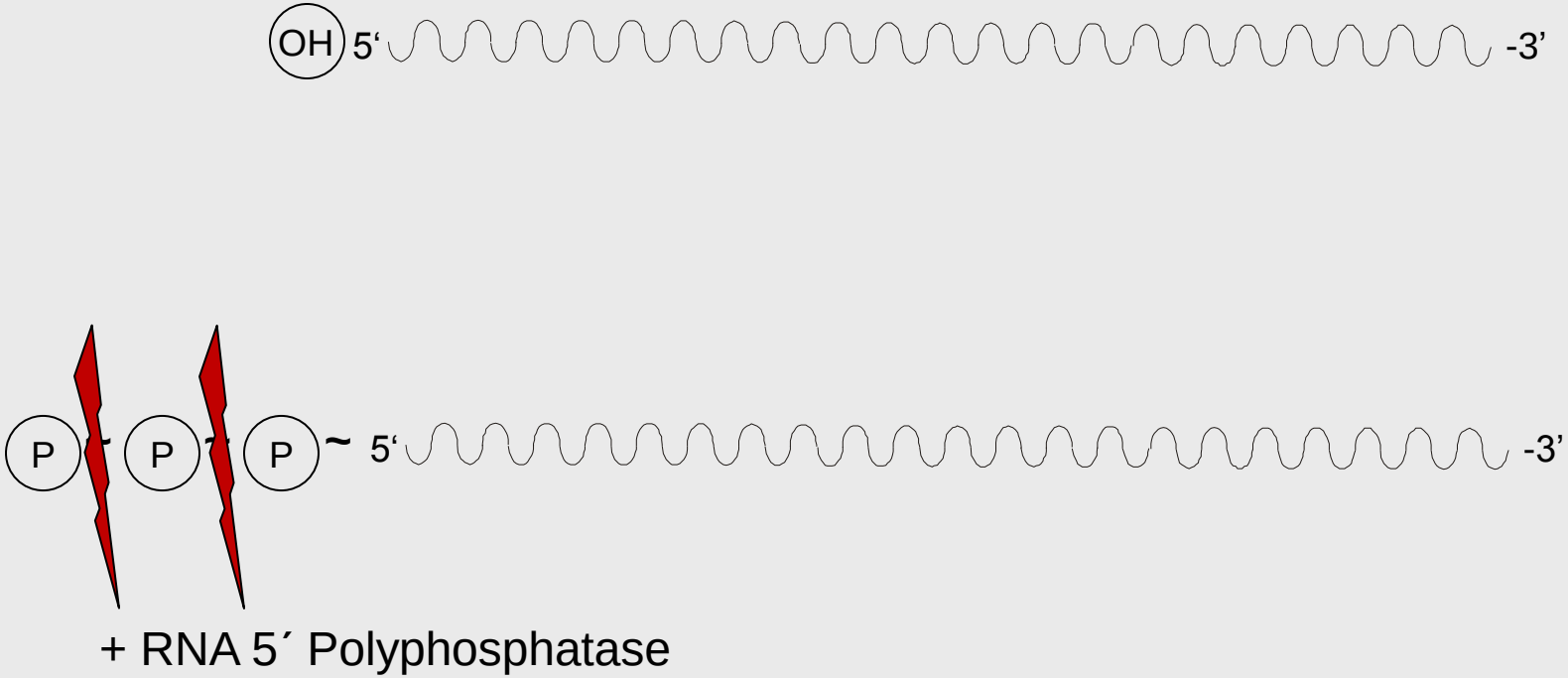
Differential RNA-seq (dRNA-seq) PRINCIPLE



from initiation of transcription („fresh“ transcripts)

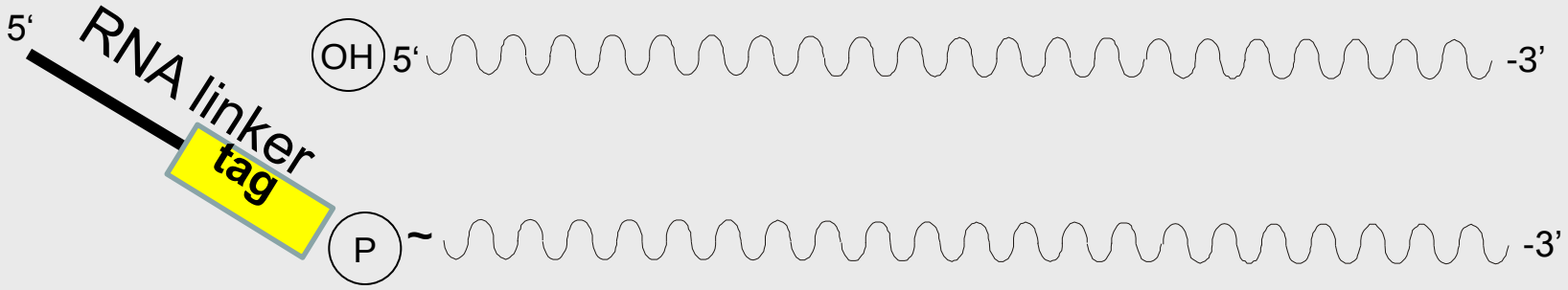

Terminator™
Exonuclease
1 Hour
Large
rRNA
mRNA

Differential RNA-seq (dRNA-seq) PRINCIPLE



Mitschke et al. (2011b) Proc. Natl. Acad. Sci. USA, *108*, 20130-20135.

Differential RNA-seq (dRNA-seq) PRINCIPLE

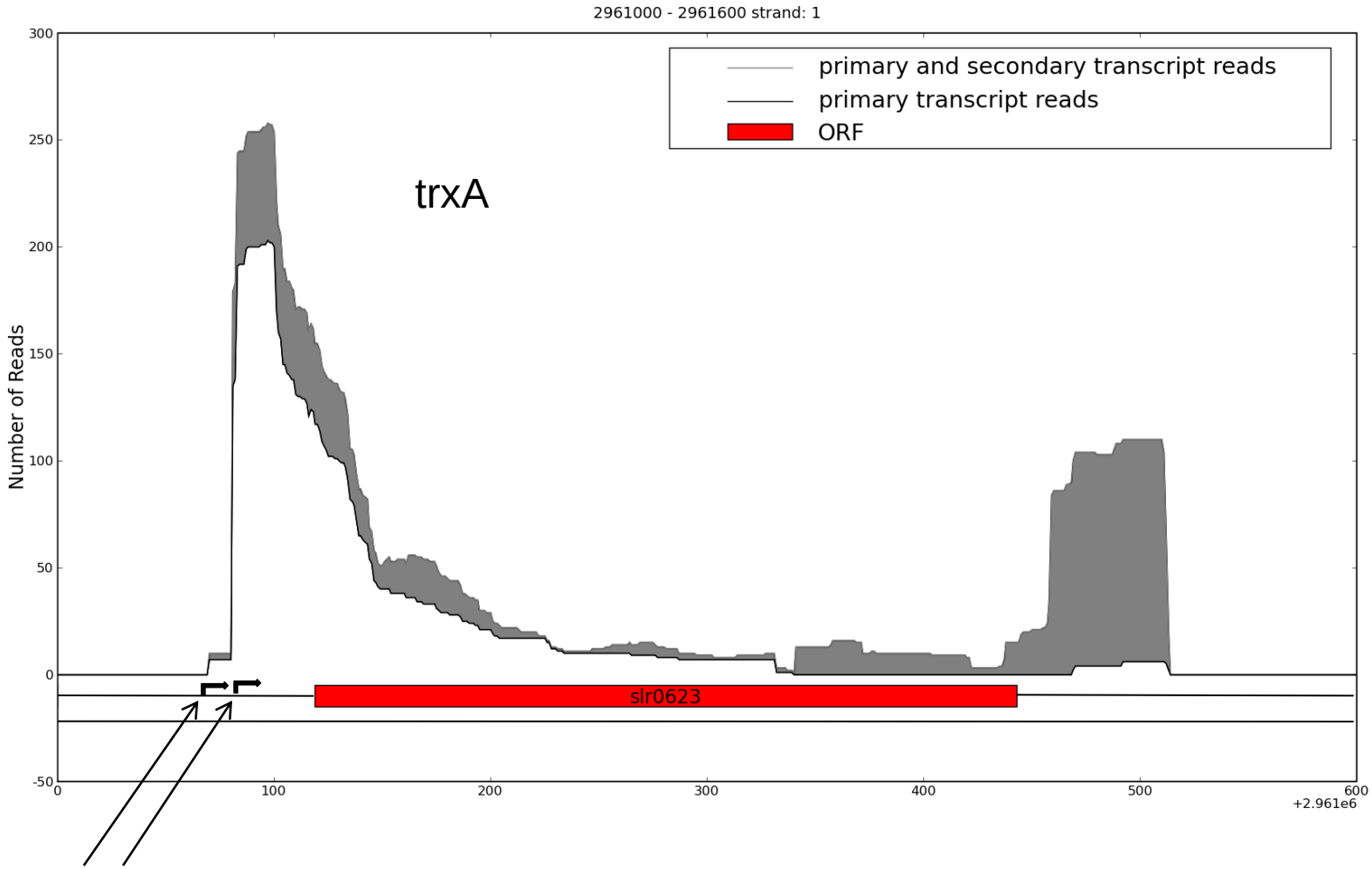


- reverse transcription
- fragmentation
- size fractionation
- adding of Illumina adaptors with 4 nt tags

→ Illumina sequencing

Mitschke et al. (2011b) Proc. Natl. Acad. Sci. USA, *108*, 20130-20135.

Deep sequencing of RNA: *Synechocystis* PCC6803

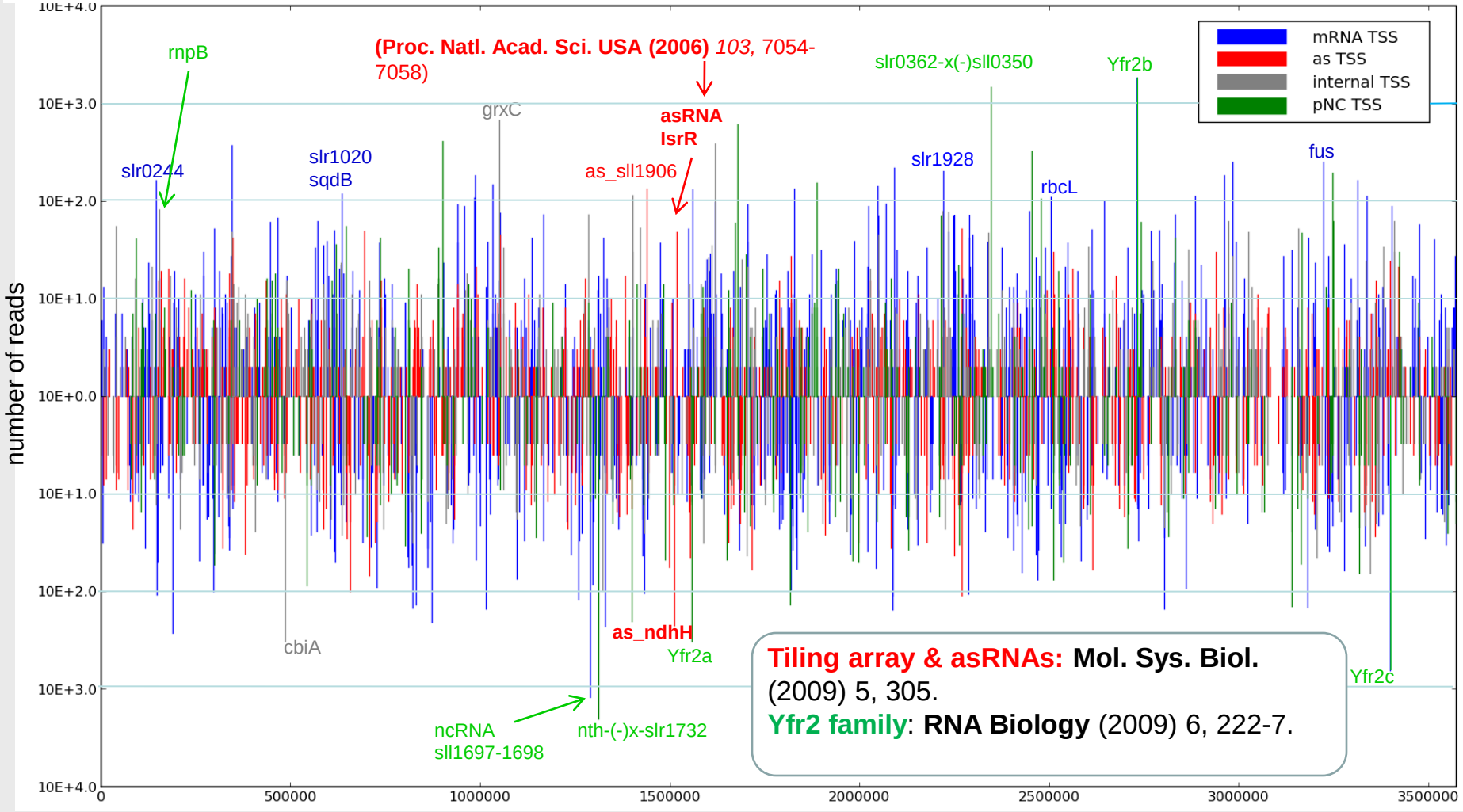


Two TSS previously described

(Navarro et al. (2000) *Electron transport controls transcription of the thioredoxin gene (trxA) in the cyanobacterium* Synechocystis *sp. PCC 6803*. Plant Molecular Biology 43: 23–32)

Distribution of >3000 TSS along the genome of *Synechocystis* measured by dRNA-seq

(Mitschke, Georg, *et al.* (2011) **Proc. Natl. Acad. Sci. USA**, 108, 2124-2129)



genome coordinate

386 TSS for ncRNA

>1000 TSS in antisense orientation

65% of all promoters give raise to non-coding RNA.....in a genome that is 87% protein-coding!

3. Proteomics

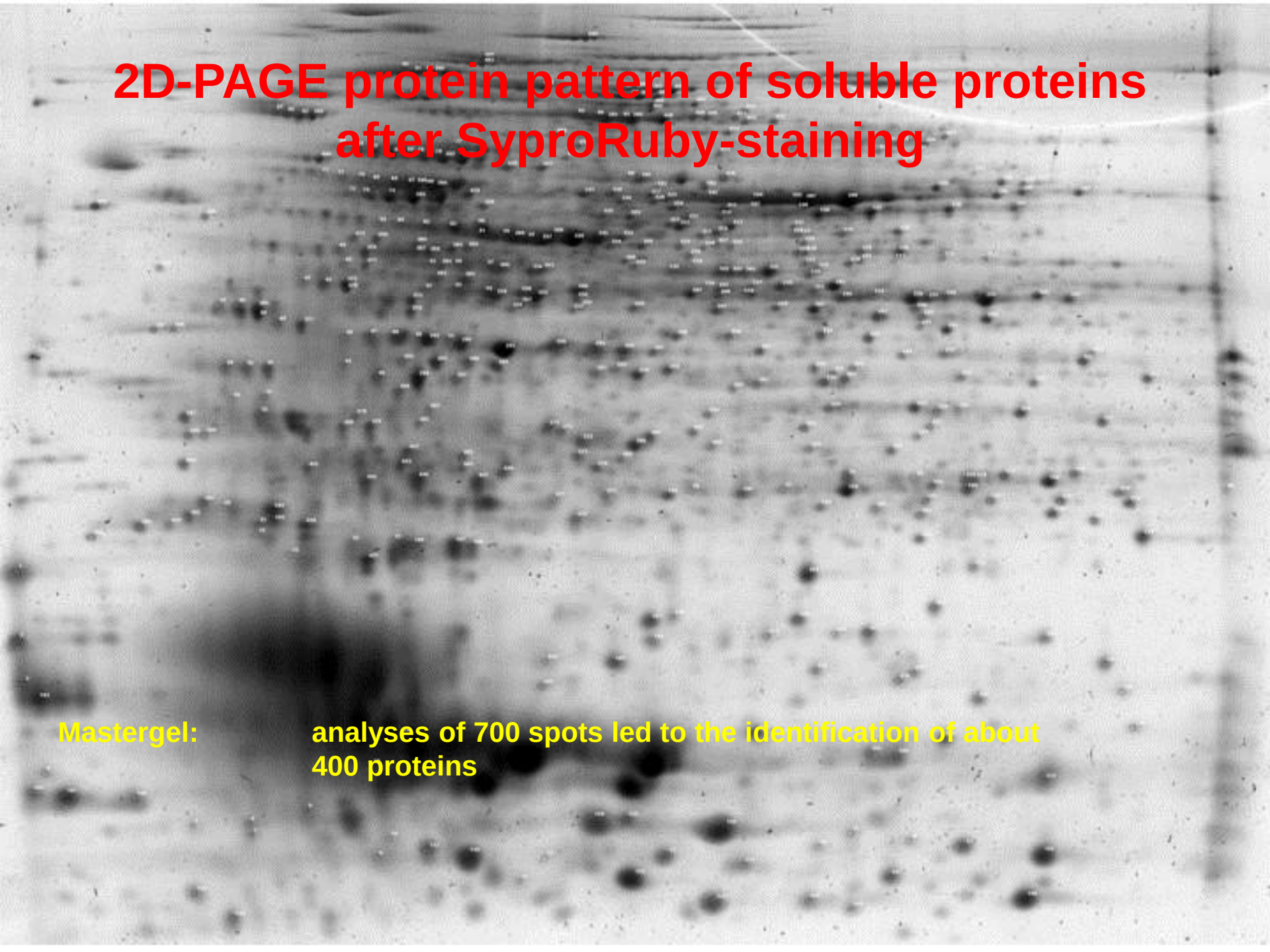
Ziel: Untersuchung möglichst vieler (aller) Proteine und deren Modifizierungen gleichzeitig in einem Zielorganismus

Heute nicht mit allen Proteinen realisierbar!

4.1 Analyse ausgewählter Proteinfraktionen

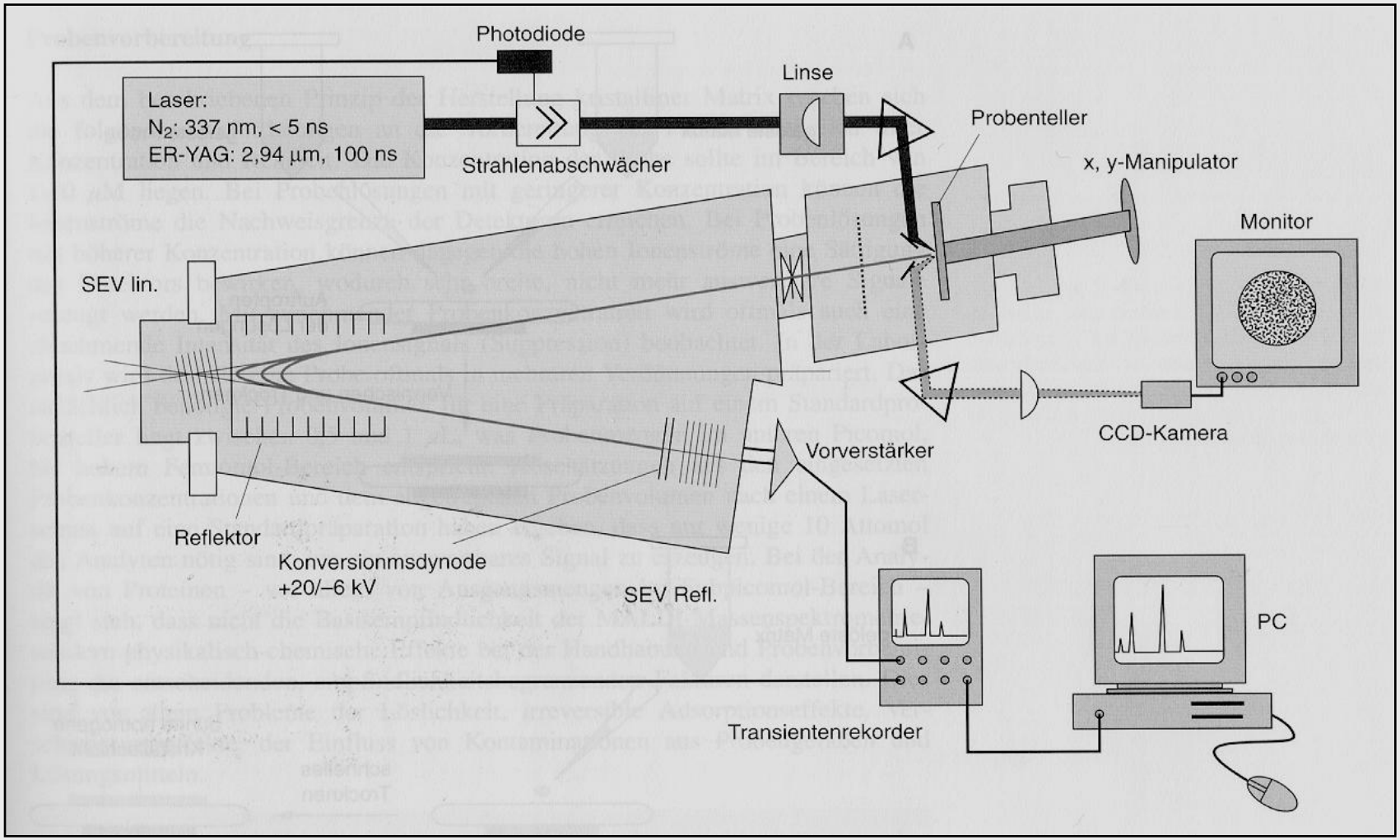
4.2 Vergleich von Proteinmustern

2D-PAGE protein pattern of soluble proteins after SyproRuby-staining



Mastergel: analyses of 700 spots led to the identification of about 400 proteins

4.1 Nachweis der Proteine durch Massenspektroskopie (MALDI-TOF)



Anwendungen von MALDI-TOF

Proteomics: Quantitative und Qualitative Analyse **aller** Proteine einer Zelle/Gewebe/Organ/Organismus

Proteinidentifizierung durch „Peptide-Mass-Fingerprinting“

Voraussetzung: Proteinsequenz ist in der Datenbank

- Erzeugen eines zu mind. 90 % reinen Proteins (klassische Proteinreinigung, einzelner Spot nach 2D-PAGE, 2D-nano LC Chromatographie)

-Definierte Fragmentierung des Proteins i.d.R. durch Trypsin-Verdau

-Schonende Ionisierung durch „Matrix-assisted-laser-dissorption“ MALDI

-Bestimmung der Massen der Fragmente durch MALDI-TOF

-Abgleich der ermittelten Bruchstückmassen mit Bruchstücken bekannter Proteine in der Datenbank

-Übereinstimmung wird durch statistische Werte angegeben (MASCOT etc.)

-Kontrolle der erhaltenen Ergebnisse ist wichtig

Nachteil der 2D-PAGE

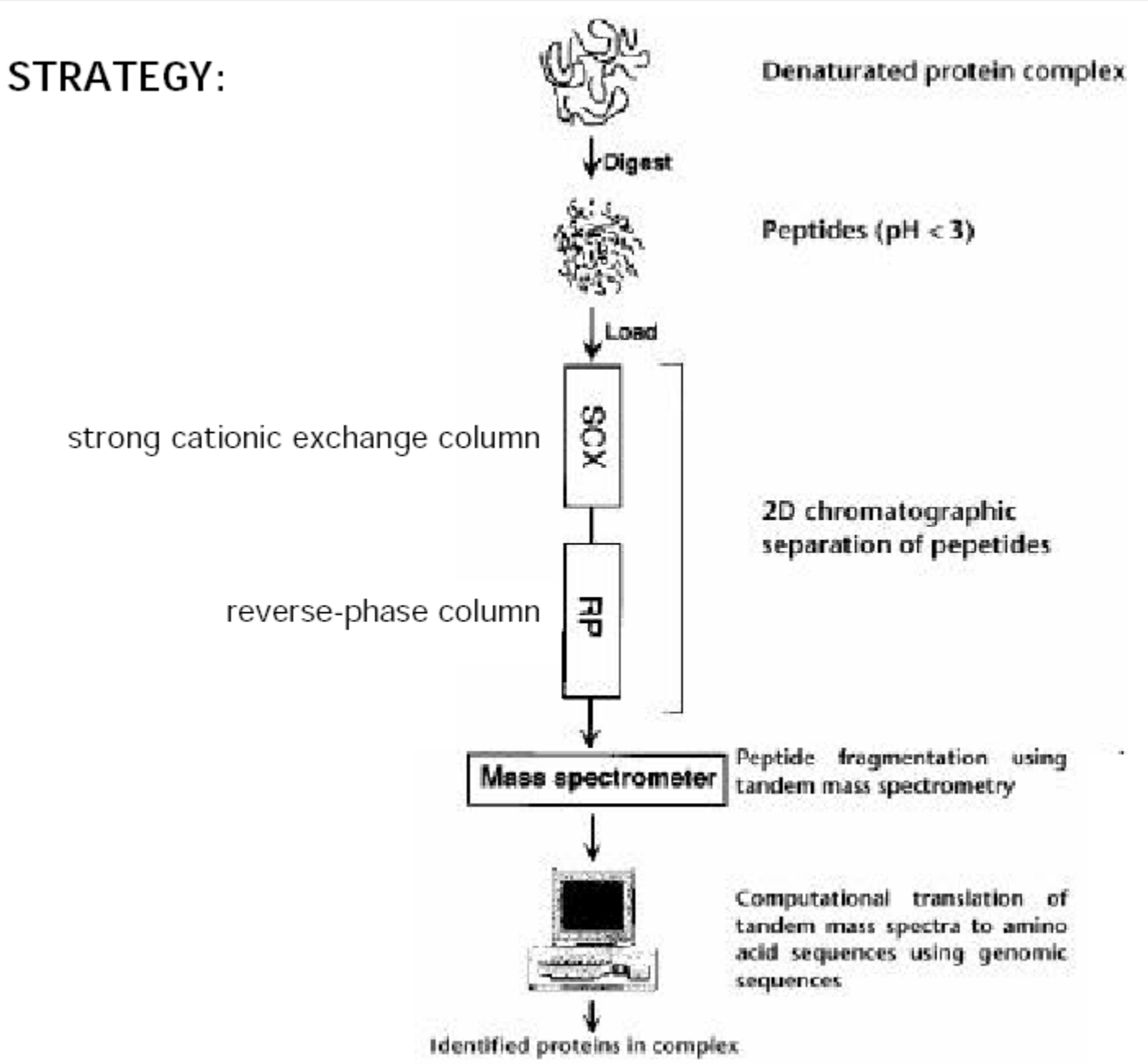
Proteine werden diskriminiert.
Membran- und basische Proteine.
Detektion gering exprimierter Proteine.
Quantifizierung eingeschränkt

Ausweg: Gelfreie Proteomics
Analyse von Proteingemischen – MS/MS
Meist QTOF
MudPIT

Quantitative Proteomics - Proteinmarkierungen – ICAT

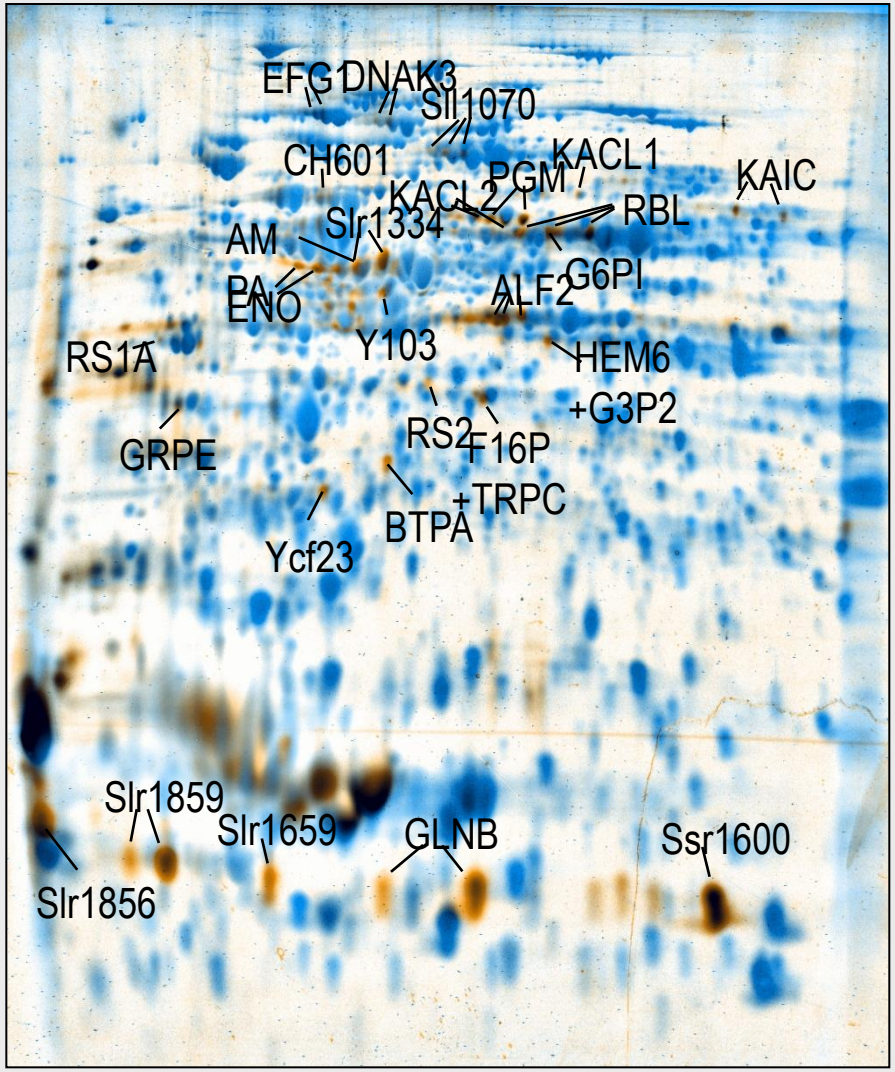
Vorteile: man sieht etwas
Proteinmodifizierungen gut abgebildet

MudPIT – multidimensional protein identification technology

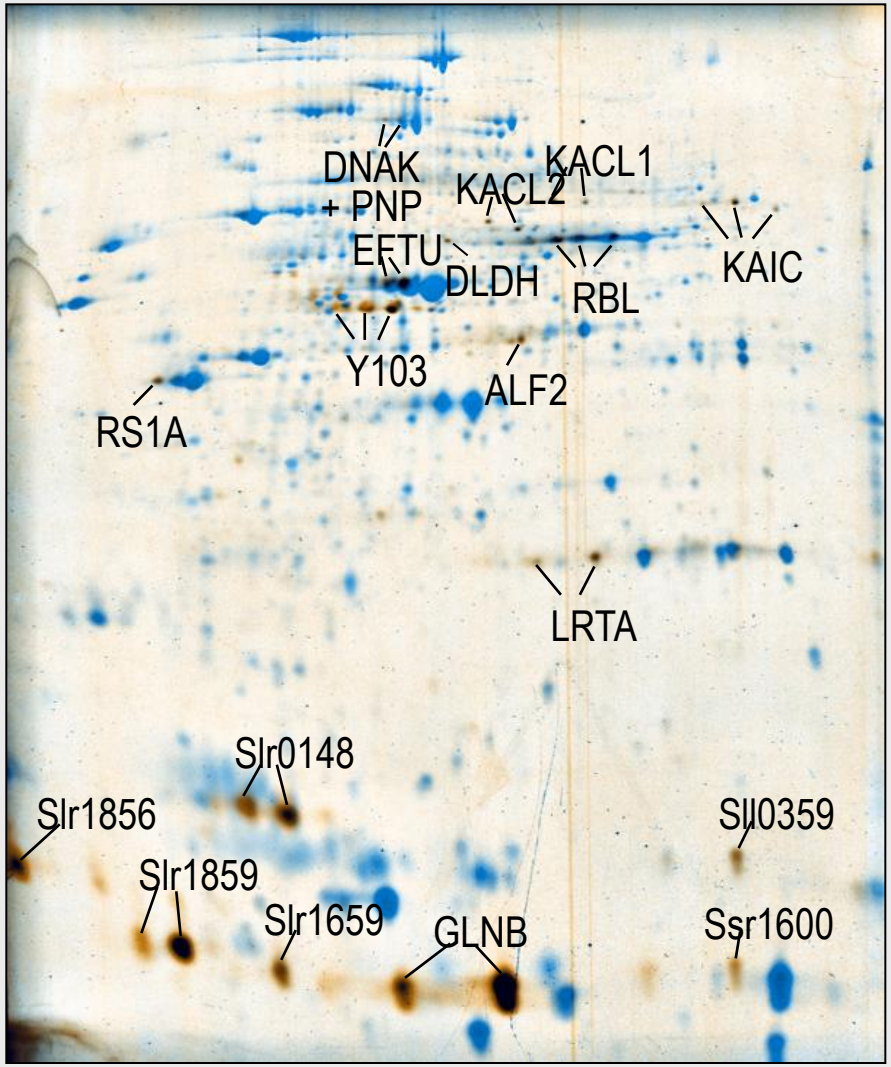


4.2 2DE-Nachweis von Proteinphosphorylierungen

Proteingesamtextrakt



Angereicherte Phosphoproteine



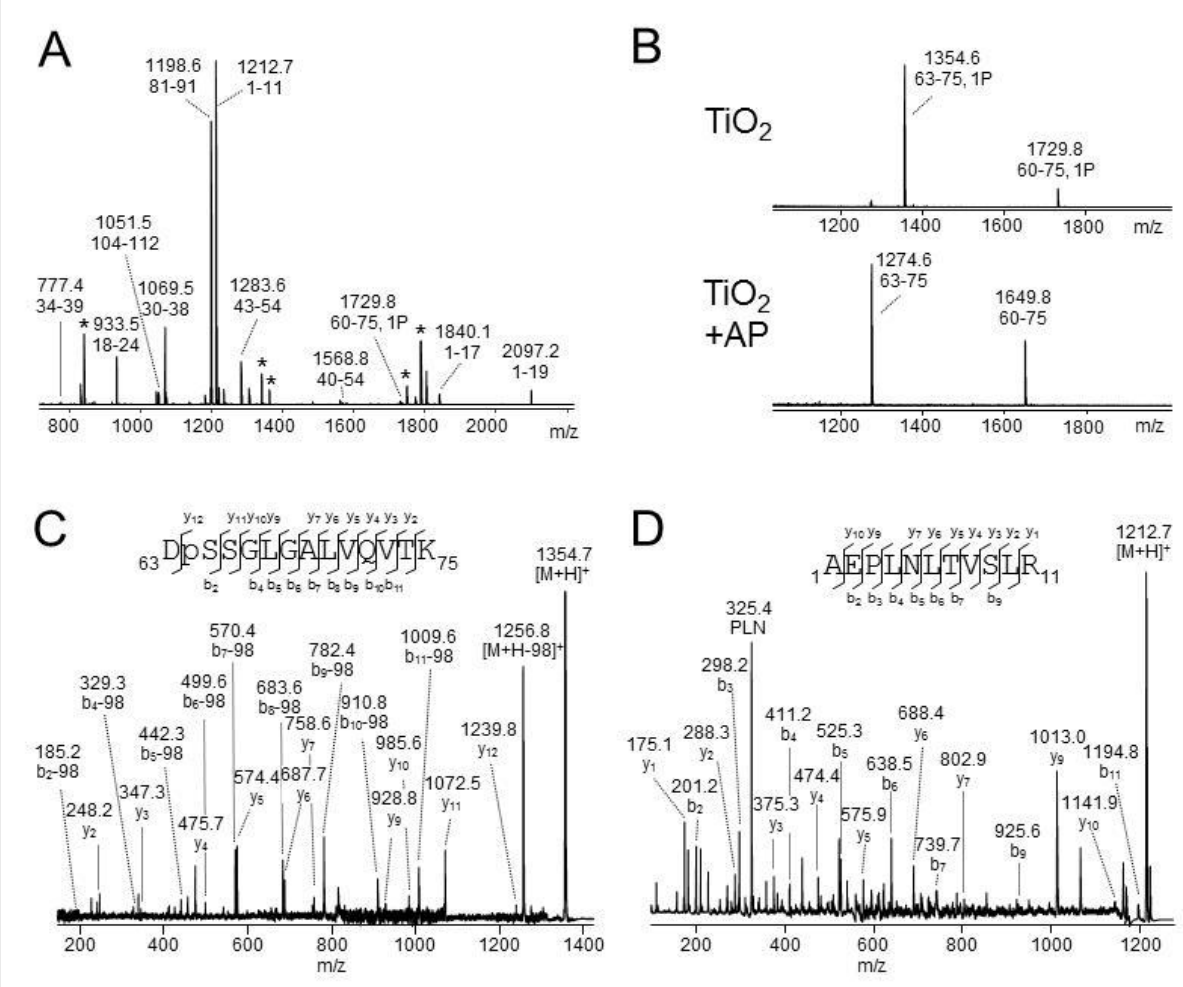
Proteinspots – blau; Phosphoproteinspots - gelb

4.2 Nachweis von Proteinphosphorylierungen

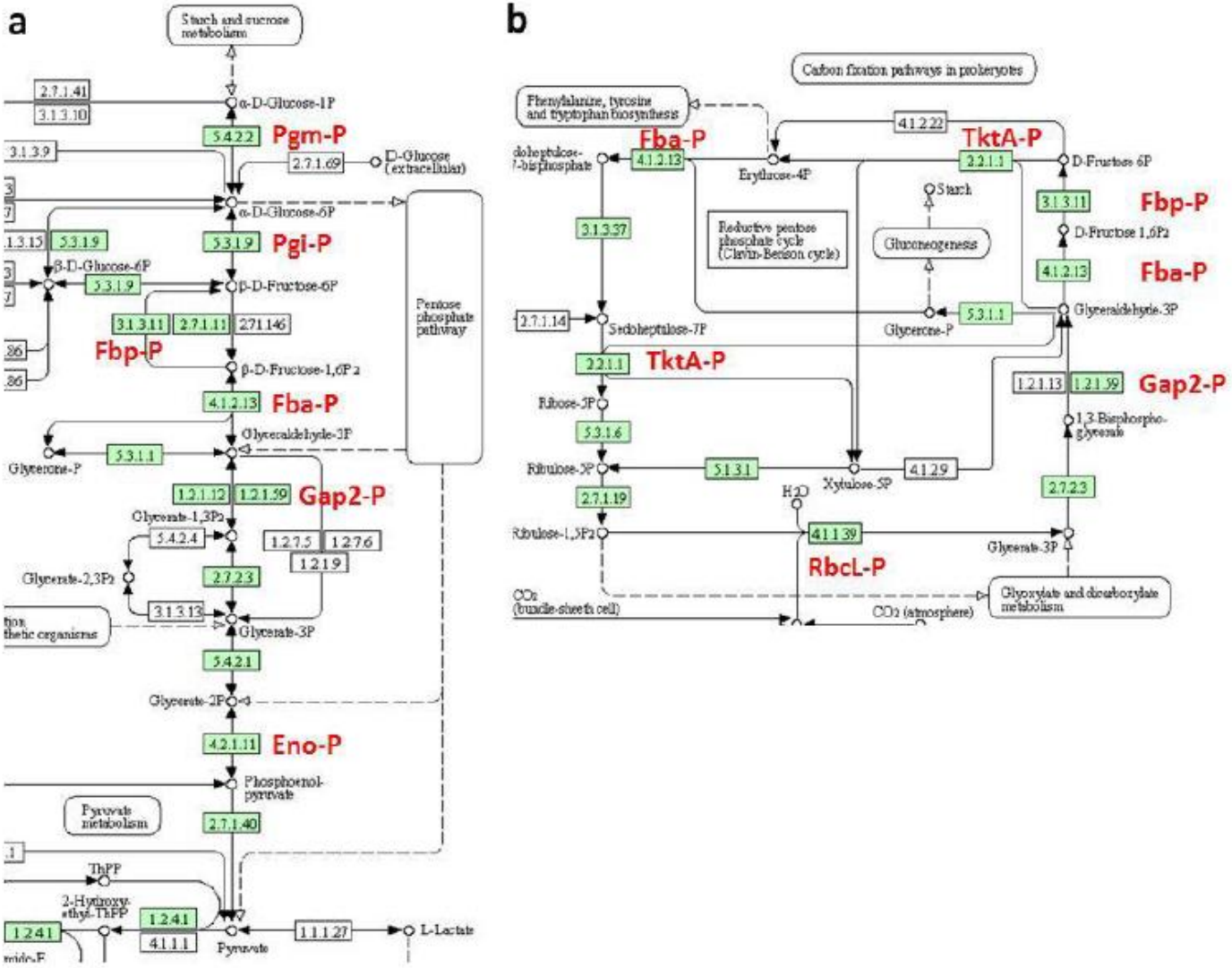
Verifizierung phosphorylierter Proteine durch MALDI-TOF

Gesamtspektrum eines trypsinierten Proteins

Angereicherte Phosphopeptide



Supplemental Figure S4: Glycolysis (a) and Calvin-Benson cycle (b) schemes copied from the KEGG pathway data base (http://www.genome.jp/kegg/pathway.html). Enzymes annotated in the genome of *Synechocystis* sp. PCC 6803 are shown in green. Enzymes found to be phosphorylated in the present study are shown in red with an added –P (see Table 1 for Abbreviations).



4. Metabolomics

Ziel: Untersuchung möglichst vieler (aller) Metabolite und deren Syntheseleistungen gleichzeitig in einem Zielorganismus

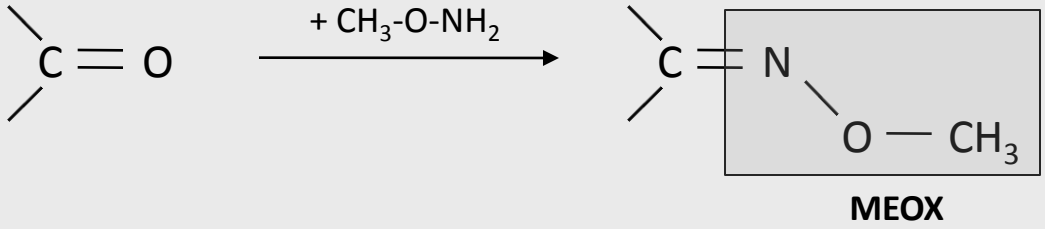
Heute nicht mit allen Metaboliten realisierbar!
Aber weitgehend unabhängig vom Organismus einsetzbar!

Metabolomic analysis

Sampling → Extraction → **Derivatisation** → Analysis → Data evaluation

Methoxyamination

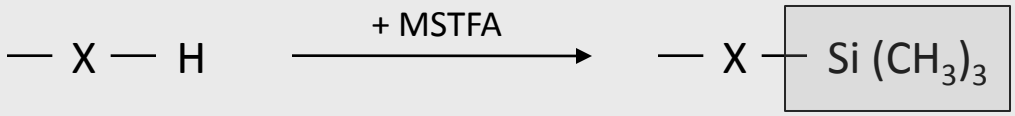
Stabilisation of carbonyl moieties



MEOX

Trimethylsilyation

Silylation of acidic proton in functional groups



TMS

Metabolomic analysis

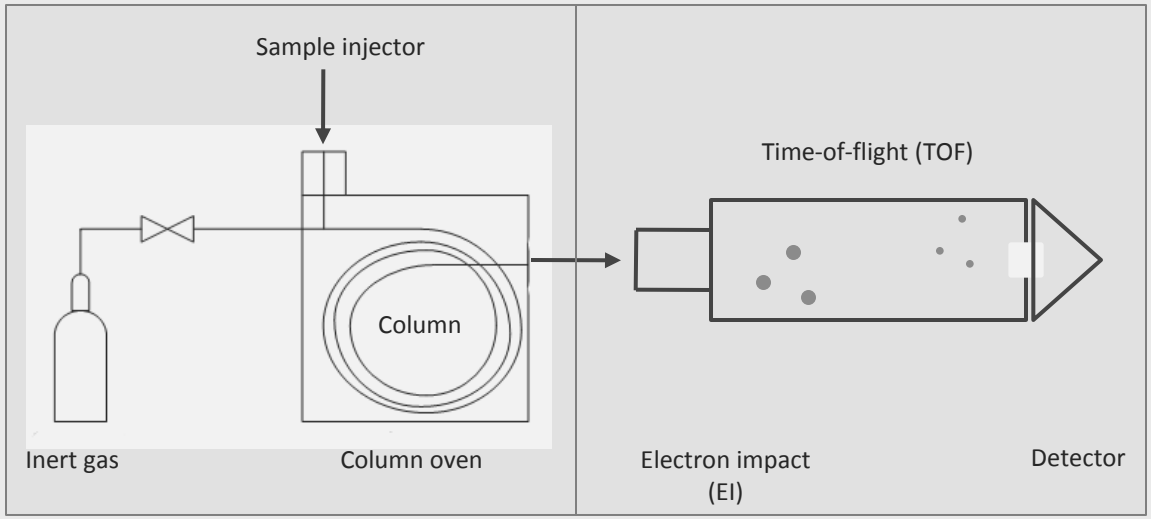
Sampling → Extraction → Derivatisation → **Analysis** → Data evaluation

Gas chromatography

Separation of the metabolite derivates

Mass spectrometry

Identification of the metabolite fragmentations

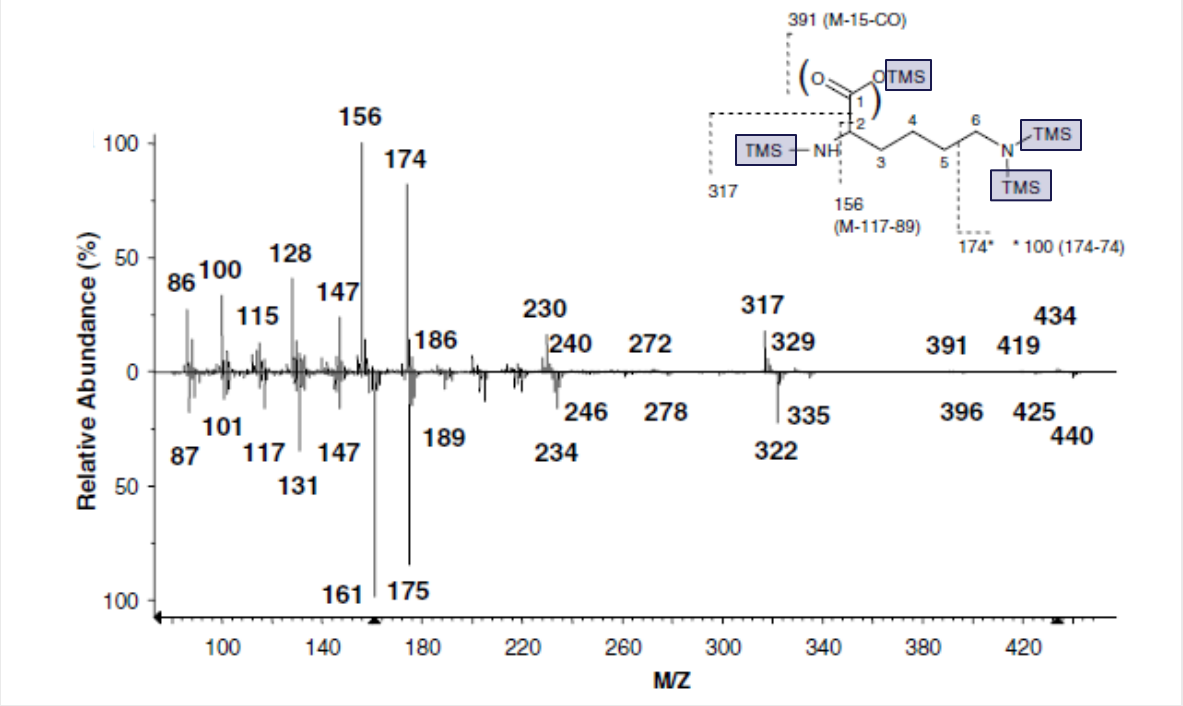


Metabolomic analysis

Sampling → Extraction → Derivatisation → Analysis → **Data evaluation**

Mass spectra

Lysine (4TMS)



Huege *et al.* (2007)

Metabolic profiling