


Gene expression microarray data analysis in plant – pathogen interaction studies



Maruša Pompe-Novak, Špela Baebler, Hana Krečič-Stres,
Ana Rotter, Kristina Gruden, Maja Ravnikar

National Institute of Biology, Večna pot 111, 1000 Ljubljana, Slovenia

National Institute of Biology



Department of Plant Physiology and Biotechnology



The main research areas

- Research of plant – pathogen and plant – pest interactions
- Development of high through-put molecular methods
- Detection of GMOs
- Detection of plant pathogens (also on-site detection) by
 - qPCR
 - PCR
 - ELISA
 - FISH
 - test plants and
 - other methods

also in a connection with bio-safety issues

Plant – pathogen interactions:

- Identification of differentially expressed genes:
 - potato – PVY^{NTN} and PVY^N
 - grapevine – Bois noir and Flavescence doree
 - colorado beetle – potato – PVY^{NTN}
- Biochemical compound analysis and their physiological role:
 - plant hormones (JA, SA...)
 - peroxidases
- Ultrastructural changes
- Localization and immuno-localization of differentially expressed molecules
- Genetic diversity of viruses

Gene expression microarray data analysis:

- Data analysis used in gene expression microarray studies of:
 - Potato – PVY^{NTN} and PVY^N
 - Grapevine – Bois noir and Flavescence doree
 - Colorado beetle – potato – PVY^{NTN}
- The main stress on data visualizing tools (MapMan)

Sets of microarrays:

- ❑ self-made 400 clones potato cDNA microarrays spotted at Plant Research International in Wageningen in The Netherlands
- ❑ self-made 4000 clones potato cDNA microarrays spotted at Rikilt in Wageningen in The Netherlands
- ❑ self-made 400 clones colorado beetle cDNA microarrays spotted at Plant Research International in Wageningen in The Netherlands
- ❑ TIGR 10000 clones potato microarrays provided by The Institute for Genomic Research from Maryland, USA.
- ❑ grapevine oligo microarrays from Genoplant

Sets of microarrays:

- Differed in:
 - nature of the probe (cDNA or oligomers)
 - organism (potato, grapevine, colorado potato beetle)
 - number of spotted probes (from 400 to 10,000)
 - manner of production (self-made or pre-spotted)

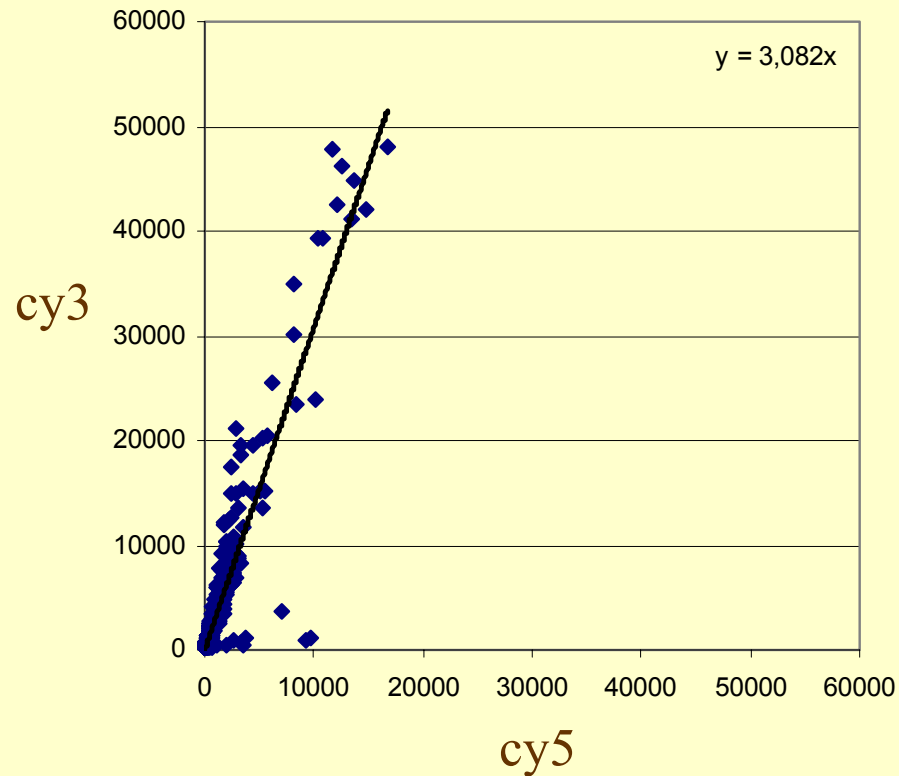
- The most suitable data analysis procedure for one set of microarrays was not also the most suitable procedure for the other sets, especially if differed in the number of probes spotted

Data analysis of 400 clones microarrays:

- ❑ Self-made 400 clones potato cDNA microarrays spotted at Plant Research International in Wageningen in The Netherlands
- ❑ used in gene expression microarray studies of potato – PVY^{NTN}
- ❑ Data analysis performed mainly by Microsoft Excell

- ❑ Procedure published in:
 - POMPE NOVAK, Maruša, GRUDEN, Kristina, BAEBLER, Špela, KREČIČ STRES, Hana, KOVAČ, Maja, JONGSMA, Maarten Anthonie, RAVNIKAR, Maja. 2006. Potato virus Y induced changes in the gene expression of potato (*Solanum tuberosum* L.). *Physiol. mol. plant pathol.* 67:237–247.

Raw data:

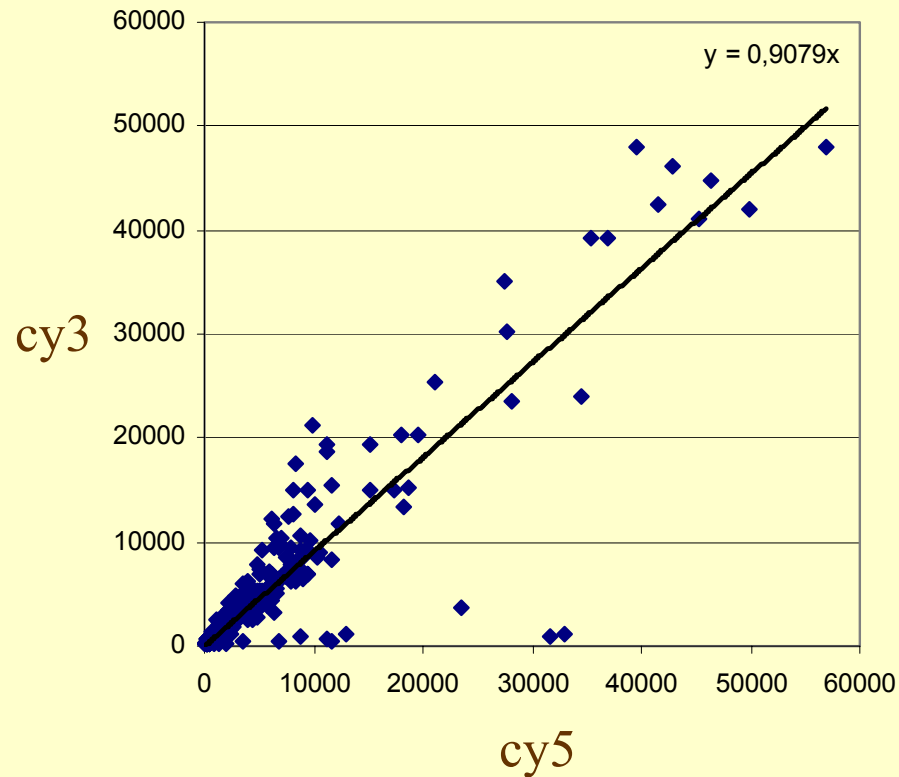


- Unequal brightness of cy3 and cy5 dyes

Normalization:

- group of normalization genes
 - **spike-in controls (luciferase...)**
 - internal controls (house keeping genes, 18S rRNA...)
- trendline
- by equalizing the distribution (Xpression, InforMax)
 - by setting average ratio to 1
 - by setting average to the reference

Normalized data:



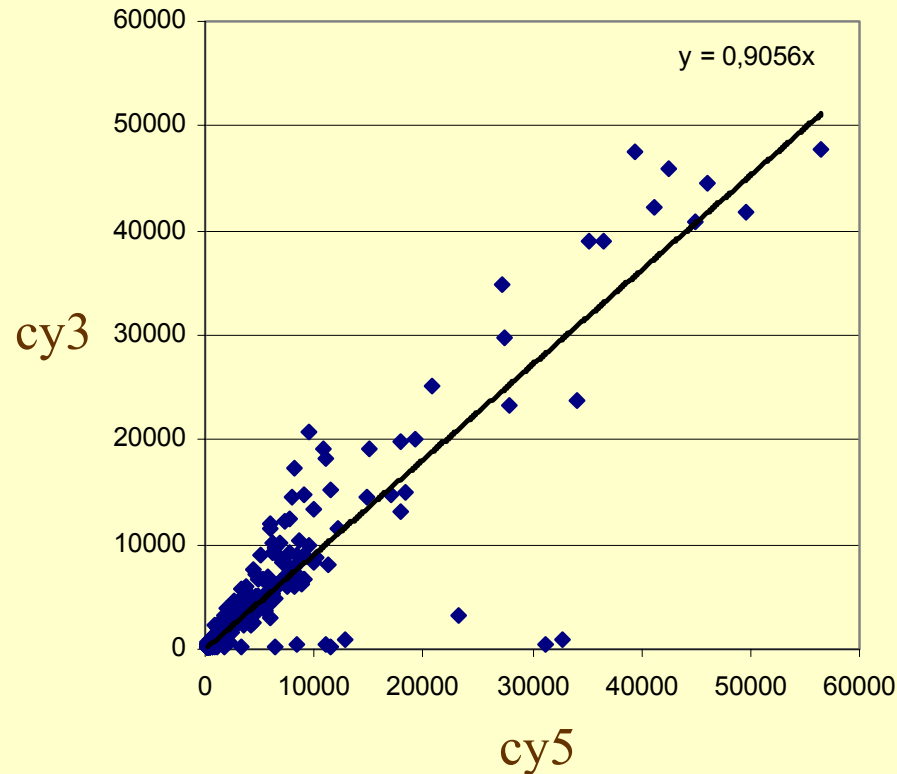
$$cy3_{\text{normalized}} = cy3_{\text{raw}}$$

$$cy5_{\text{normalized}} = cy5_{\text{raw}} * (\text{average luciferase intensity cy3} / \text{average luciferase intensity cy5})$$

Background subtraction:

- **group of background spots (yeast DNA...)**
- background around spots
 - global background – average over the whole microarray area
 - local background – background around individual spots
- subtracted before normalization
- **subtracted after normalization**
- channel specific subtraction
- **subtraction of maximal background**

Subtracted background:

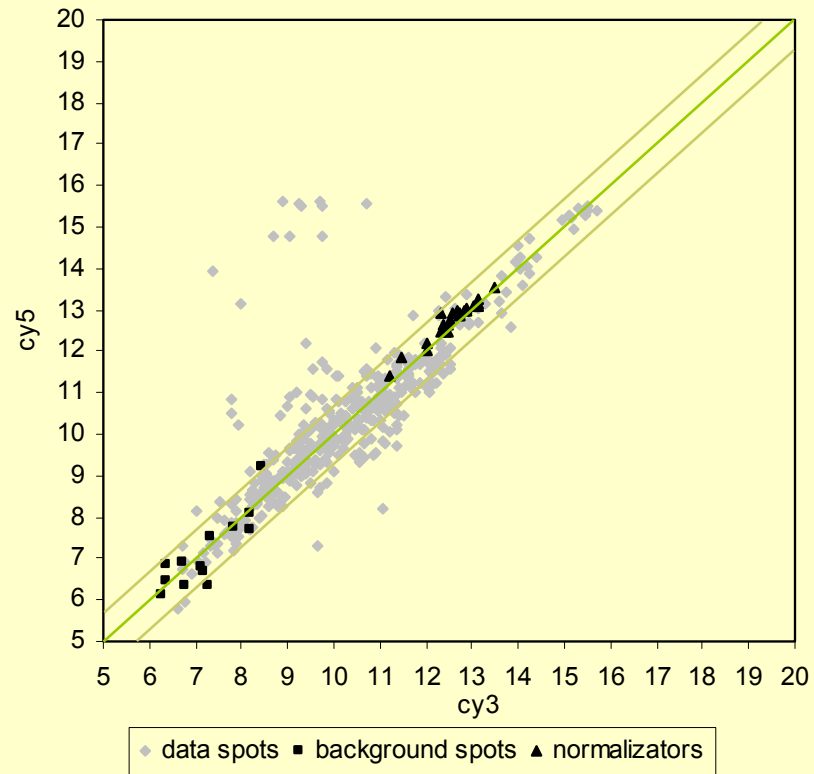


$\text{background} = \text{MAX}(\text{average yeast gene intensity cy3}; \text{average yeast gene intensity cy5})$

$\text{cy3}_{\text{sub.back.}} = \text{IF}(\text{cy3}_{\text{norm.}} - \text{background} > \text{background}; \text{cy3}_{\text{norm.}} - \text{background}; \text{background})$

$\text{cy5}_{\text{sub.back.}} = \text{IF}(\text{cy5}_{\text{norm.}} - \text{background} > \text{background}; \text{cy5}_{\text{norm.}} - \text{background}; \text{background})$

Log₂ of processed data:



Statistical significance:

- **cv – Coefficient of Variability**

cv = standard deviation / average

$$cv = \sqrt{cv_1^2 + cv_2^2 + \dots + cv_n^2 + cv_{\text{between}}^2}$$

- **Student's T-test or Mann-Whitney U-test**

p < 0.01 ***

p < 0.1 **

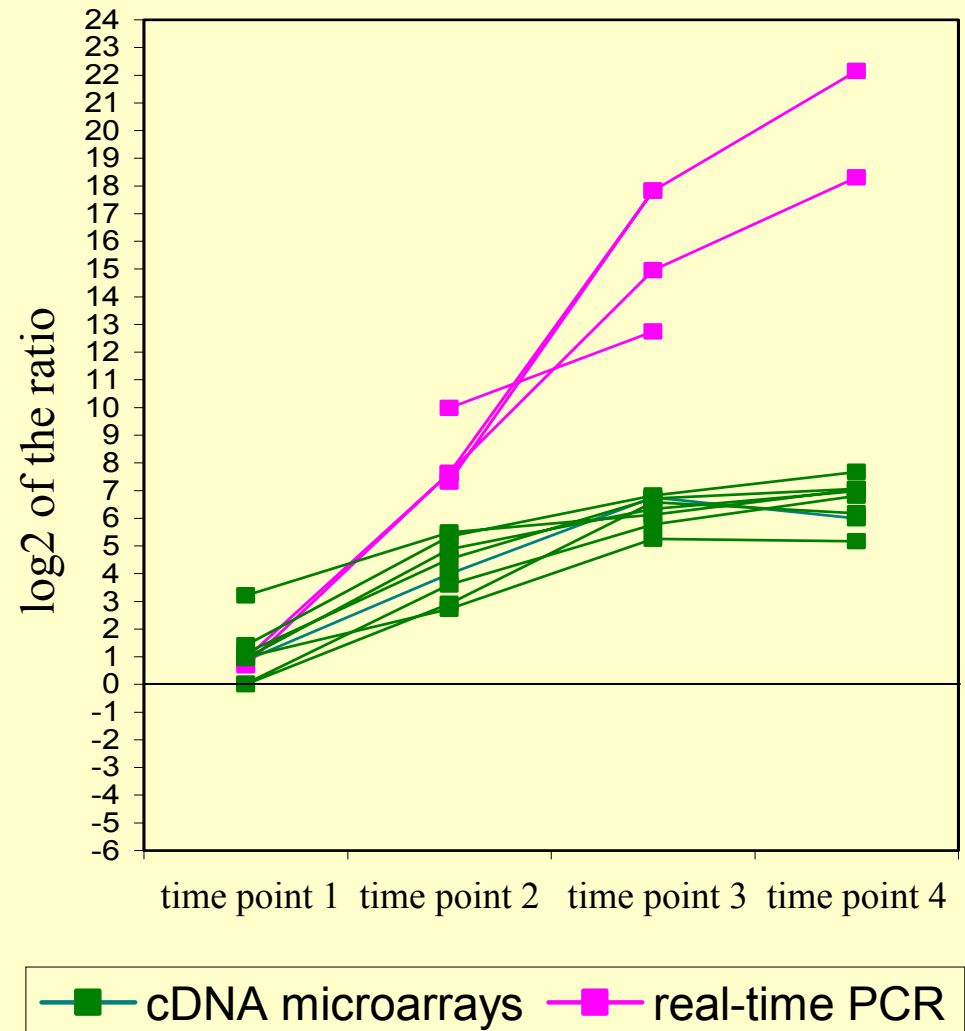
p < 0.5 *

Data visualization

clone number	name of the most similar sequence form database	log2 of the ratio between gene expression in infected and healthy plants							
		upper non-inoculated leaves 7 days after inoculation		lower inoculated leaves 7 days after inoculation		lower non-inoculated leaves 14 days after inoculation		lower leaves of secondary infected plants	
		average	p	average	p	average	p	average	p
pot 190	heat shock protein 70 [2/2]	-0,2697	**	-0,7520	*	-1,2264	***	-0,8117	***
pot 015	glycine-rich RNA binding protein [2/2]	-0,5289	*	-0,7681	***	-0,5416		0,6403	*
pot 258	ABC transporter protein 1	0,3000		0,5449	**	1,3379	*	1,4043	**
pot 197	auxin repressed protein [1/1]	-0,6203	***	-1,1433	***			-0,1355	
pot 315	beta(1,3)glucanase	-0,2227	**	0,3671	*	0,9394		0,6021	*
pot 282	catalase 1 [1b/1]	0,1139		-0,1861		-1,7565	***	-0,2630	*
pot 031	unknown function [24/67]	-0,9003	**	-0,7164	*	-1,6731	***	0,2716	
pot 270	unknown function [43/67]	1,4043	***	1,2629	**			0,9445	

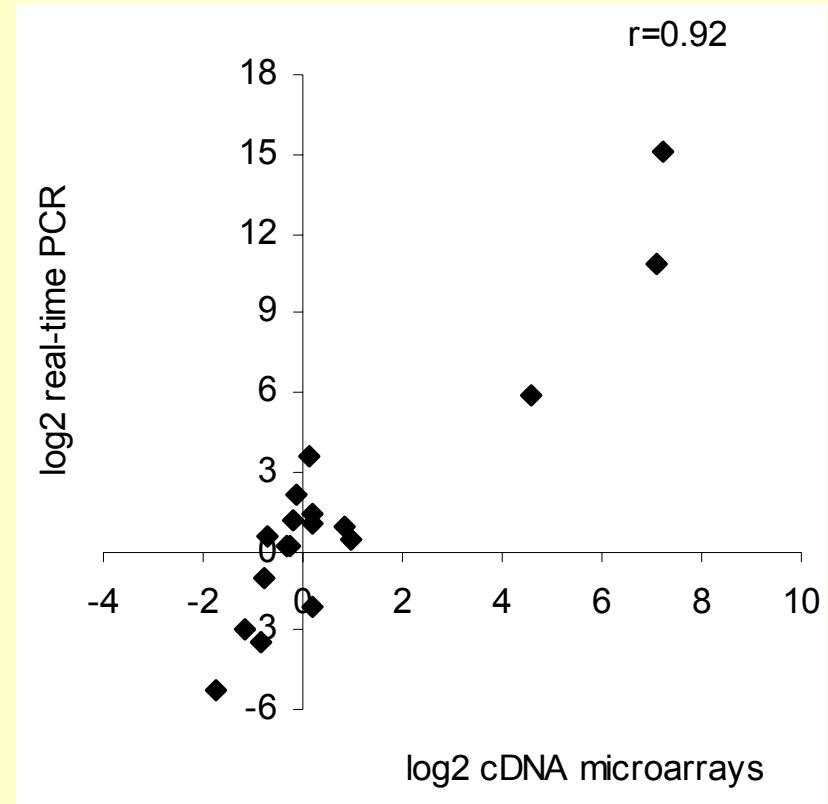
Result verification by RT-qPCR

- In general, RT-qPCR results showed stronger response in gene expression than cDNA microarrays, so only trends in gene expression could be compared



Result verification by RT-qPCR

- Comparability of results from cDNA microarrays and RT-qPCR depended on the position of qPCR probe
- Results were more comparable if the amplicon was designed in the conserved region
- Anyway, high correlation coefficients were achieved



Data analysis of 10000 clones microarrays:

- TIGR 10000 clones potato microarrays provided by The Institute for Genomic Research from Maryland, USA.
- Used in gene expression microarray studies of potato – PVY
- Data analysis was performed mainly by:
 - ArrayPro
 - R
 - MEV (TM4)
 - MapMan
- Procedure published in:
 - BAEBLER, Špela, HREN, Matjaž, KOGOVŠEK, Polona, KREČIČ STRES, Hana, CURK, Tomaž, JUVAN, Peter, ZUPAN, Blaž, POMPE NOVAK, Maruša, GRUDEN, Kristina. 2005. *Laboratory and computer practice protocols*. Ljubljana: Department of Plant Physiology and Biotechnology, National Institute of Biology.

ArrayPro:

Ignored spots were determined:

- non-validated spots
- missing spots
- not uniform spots (smeared, doughnut shape...)

$$\text{Raw_Int_TM_cy3} / \text{Raw_Int_SD_cy3} < 1 \text{ OR}$$

$$\text{Raw_Int_TM_cy5} / \text{Raw_Int_SD_cy5} < 1$$

- spots with not uniform background

$$\text{Raw_Int_TM_cy3} / \text{Back_Int_SD_cy3} < 3 \text{ OR}$$

$$\text{Raw_Int_TM_cy5} / \text{Back_Int_SD_cy5} < 3$$

- spots with low signal intensity

$$\text{Raw_intensity_TM_cy3} - \text{Background_TM_cy3} < 1.5 *$$

$$\text{Background_TM_cy3} \text{ AND}$$

$$\text{Raw_intensity_TM_cy5} - \text{Background_TM_cy5} < 1.5 *$$

$$\text{Background_TM_cy5}$$

R:

- Cross-channel normalization by Local regression (Loess)

- Background subtraction

Net_Int

Raw_intensity_TM_cy3 - Background_TM_cy3

Raw_intensity_TM_cy5 - Background_TM_cy5

- $M = \log_2(\text{Net_Int_cy3}/\text{Net_Int_cy5})$

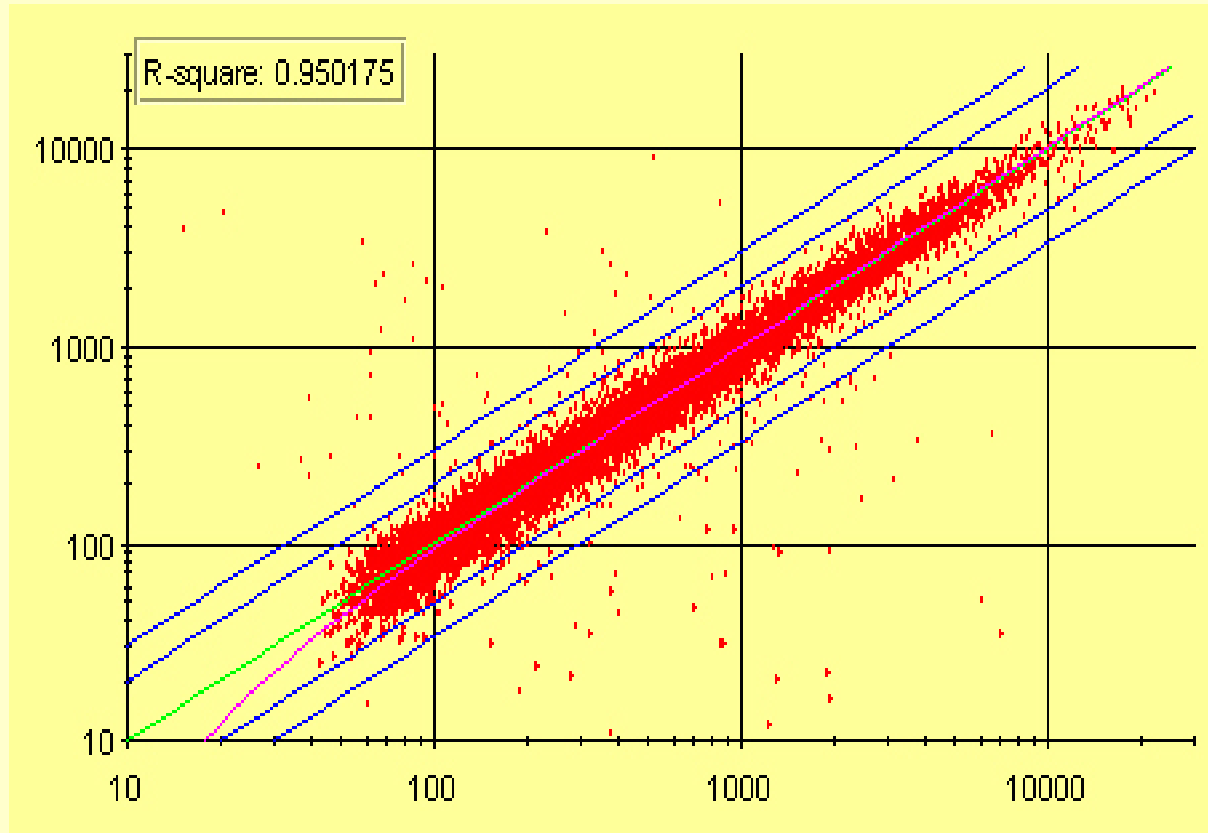
- $A = \log_2(\text{Net_Int_cy3} * \text{Net_Int_cy5} / 2)$

- Replicated spots average

- Filtration

$-0.2 < M < 0.2$

Processed data:



Microarray Experiment Viewer (MEV):

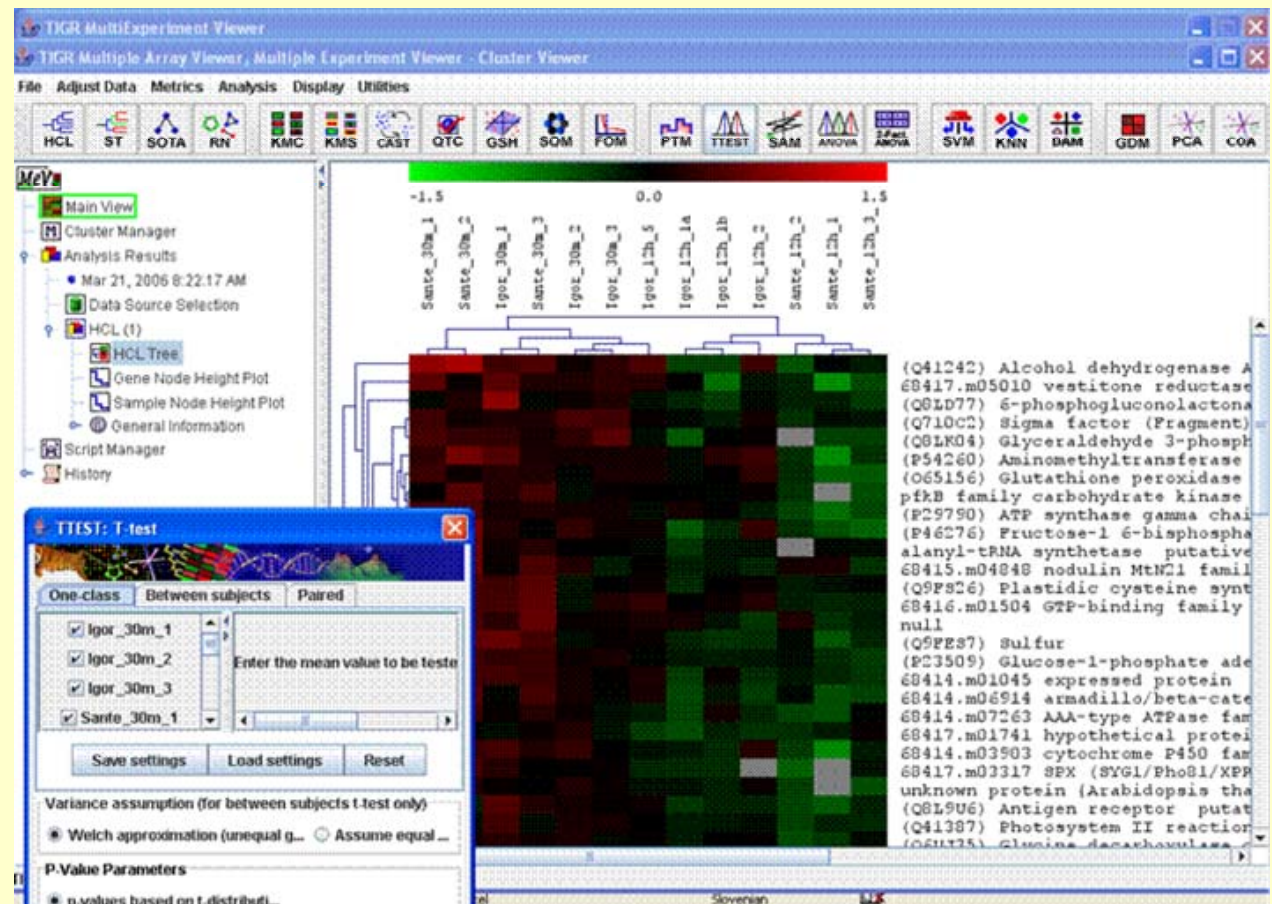
□ Program package TM4

□ T-test

□ ANOVA

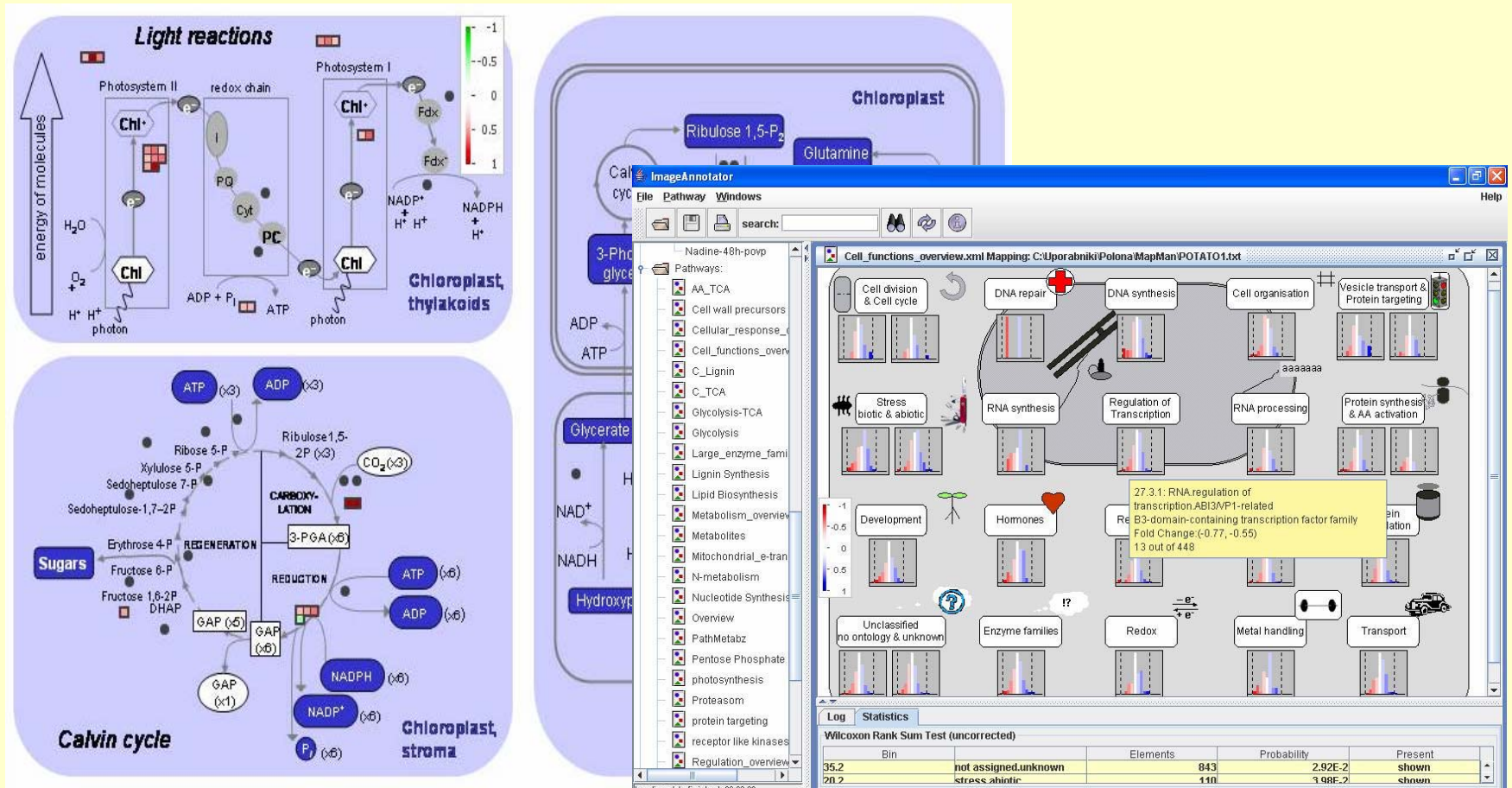
□ clustering

□ heat map



DNA microarrays – data visualization

MapMan (MPI-MPP, Potsdam, Germany):

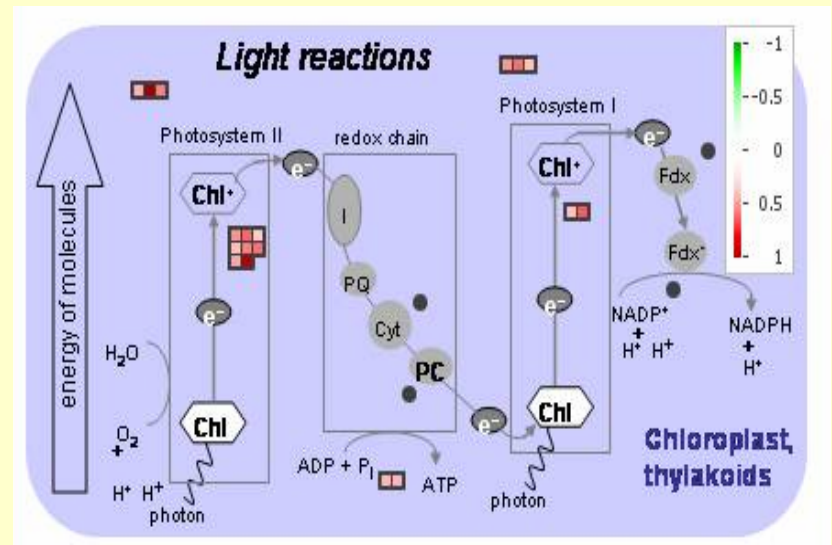


MapMan:

- ❑ The use of MapMan offers the possibility to paint out microarray profiling experiments onto diagrams of metabolic pathways or processes, and to visualize the responses of gene expression in a biological context
- ❑ MapMan is supported by a plant specific ontology
- ❑ The principle of the MapMan ontology is a hierarchical BIN-based structure. Each BIN comprises items of similar biological function, and can be further split into subBINs corresponding to submodes of the biological function.

MapMan:

- ❑ The MapMan ontology and software were developed for Arabidopsis ATH1 arrays (Thimm et al. 2004)
- ❑ Our recent very important goal, in agreement with our collaborators from Max-Planck-Institute of Molecular Plant Physiology (Potsdam-Golm, Germany), is to extend this software platform to allow it to be applied to potato and grapevine plants



MapMan:

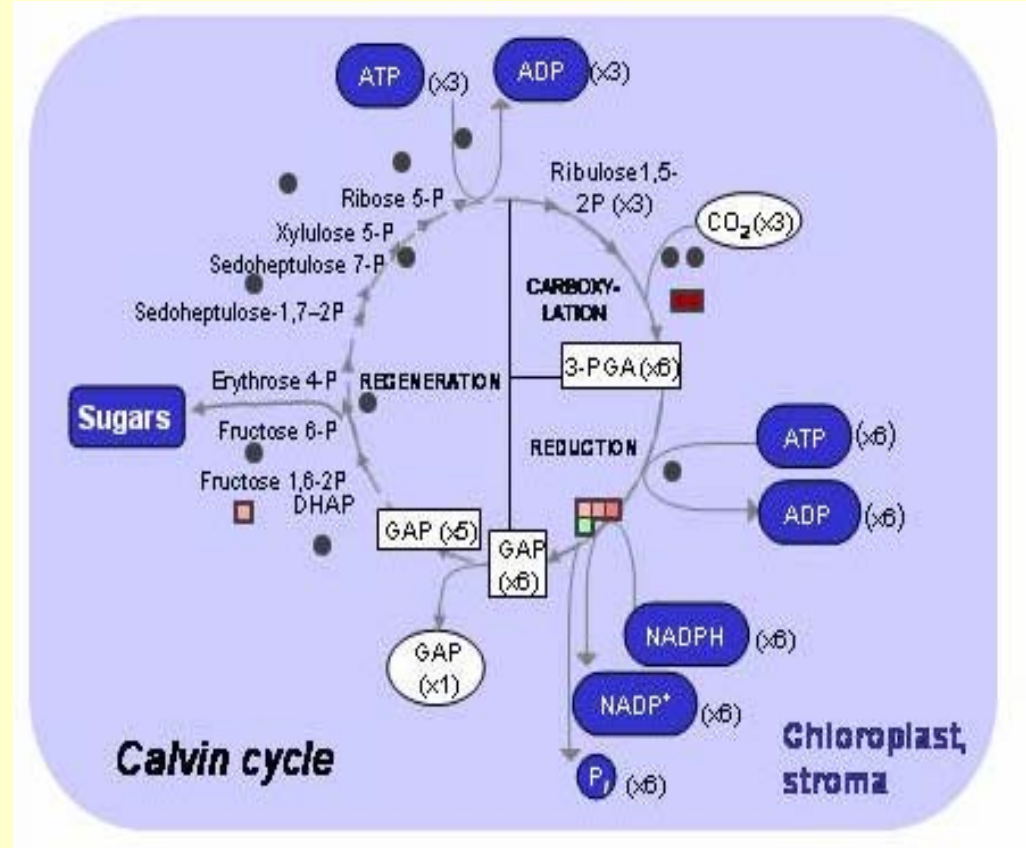
- The problems in conversion of MapMan for potato 10,000 clones array are:
 - Differences in gene sequences of potato and Arabidopsis
 - Some specific differences in the metabolism of potato and Arabidopsis
 - Whereas the full genome sequence is available for Arabidopsis, potato tentative consensus sequences (TC) are based on clustering ESTs
 - Further, the Arabidopsis ATH1 array is based on oligomers of known composition, whereas the TIGR potato 10k array is cDNA based

MapMan:

- ❑ In order to transfer the MapMan BIN system to the potato TC system represented on the TIGR microarrays, the sequences and the annotations of the potato TCs were downloaded from the StGI (Solanum tuberosum gene indices) database
- ❑ The sequences were then BLASTed against the Arabidopsis proteome, release TIGR 5, which is the basis for the original MapMan BIN structure
- ❑ In this way every potato TC was assigned to the best matching Arabidopsis gene or to no gene at all
- ❑ TC that did not fall into either BIN, are being assigned to corresponding BINs manually, on the basis of its sequence (NCBI blast), protein domain information from the StGI and TIGR annotation

MapMan:

- Special consideration will be made for the genes that are involved in different processes
- In this way the original BIN structure can be modified



Data mining

- Data mining
 - Extraction of useful and understandable patterns from large volumes of heterogeneous data
- Different transgenic lines of potato were analyzed by data mining
 - In order to find differences in gene expression level characteristic for resistant plants that differentiate them from sensitive ones
 - 12 genes that determined sensitivity of plants and 16 genes that determined resistance of analyzed plants were found

National Institute of Biology, Ljubljana, Slovenia
Max-Planck-Institute of Molecular Plant Physiology,
Potsdam-Golm, Germany
Plant Research International, Wageningen, The Netherlands
Jožef Stefan Institute, Ljubljana, Slovenia
Rikilt Institute of Food Safety, Wageningen, The Netherlands
Central Science Laboratory, York, Great Britain
University of Bristol, Bristol, Great Britain
University of Poitiers, Poitier, France
INRA, Montpellier, France
University of Udine, Udine, Italy

Thank you!