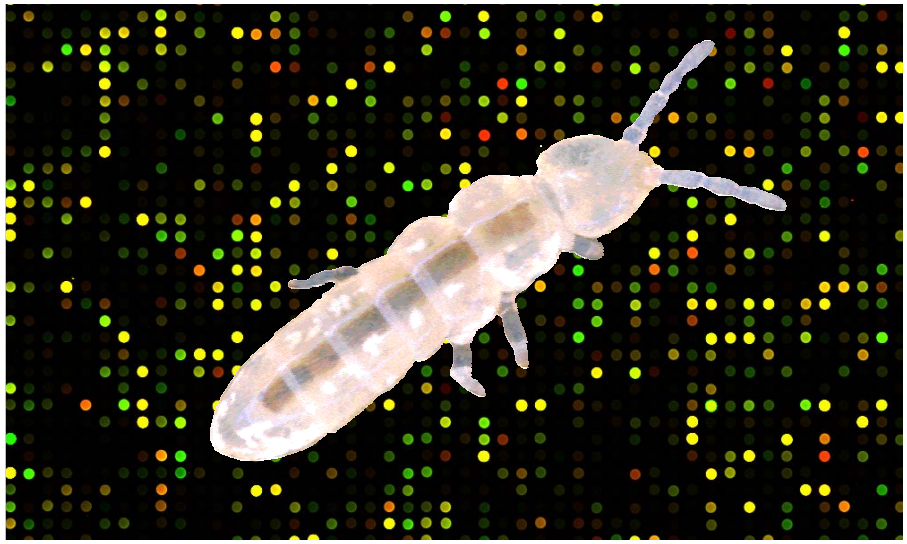
Soil invertebrates as a genomic model to study pollutants in the field



Dick Roelofs, Martijn Timmermans, Muriel de Boer,

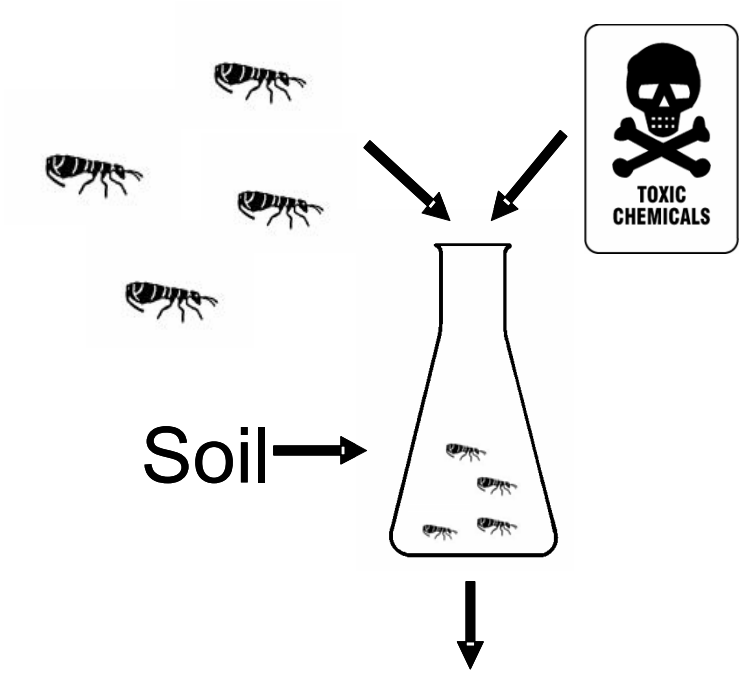
Ben Nota, Tjalf de Boer, Janine Mariën, Nico van Straalen

ecogenomics

vrije Universiteit amsterdam

Folsomia candida in soil quality testing

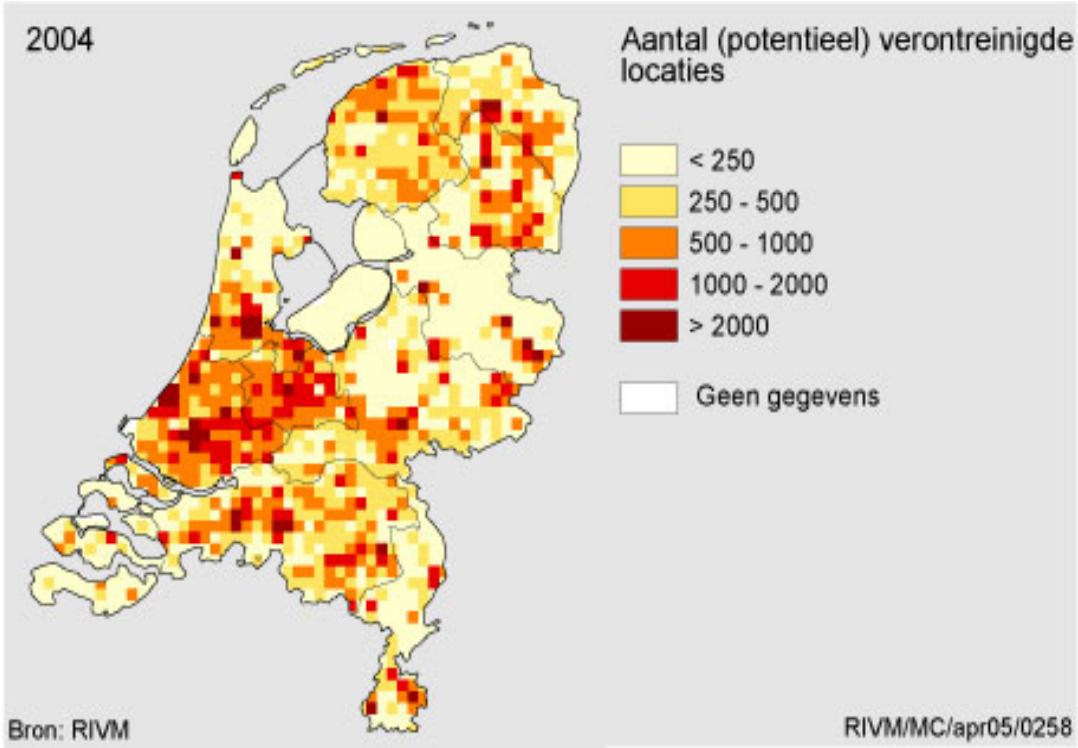
A test organism for more than 40 years for estimating the effects of pesticides and environmental pollutants.

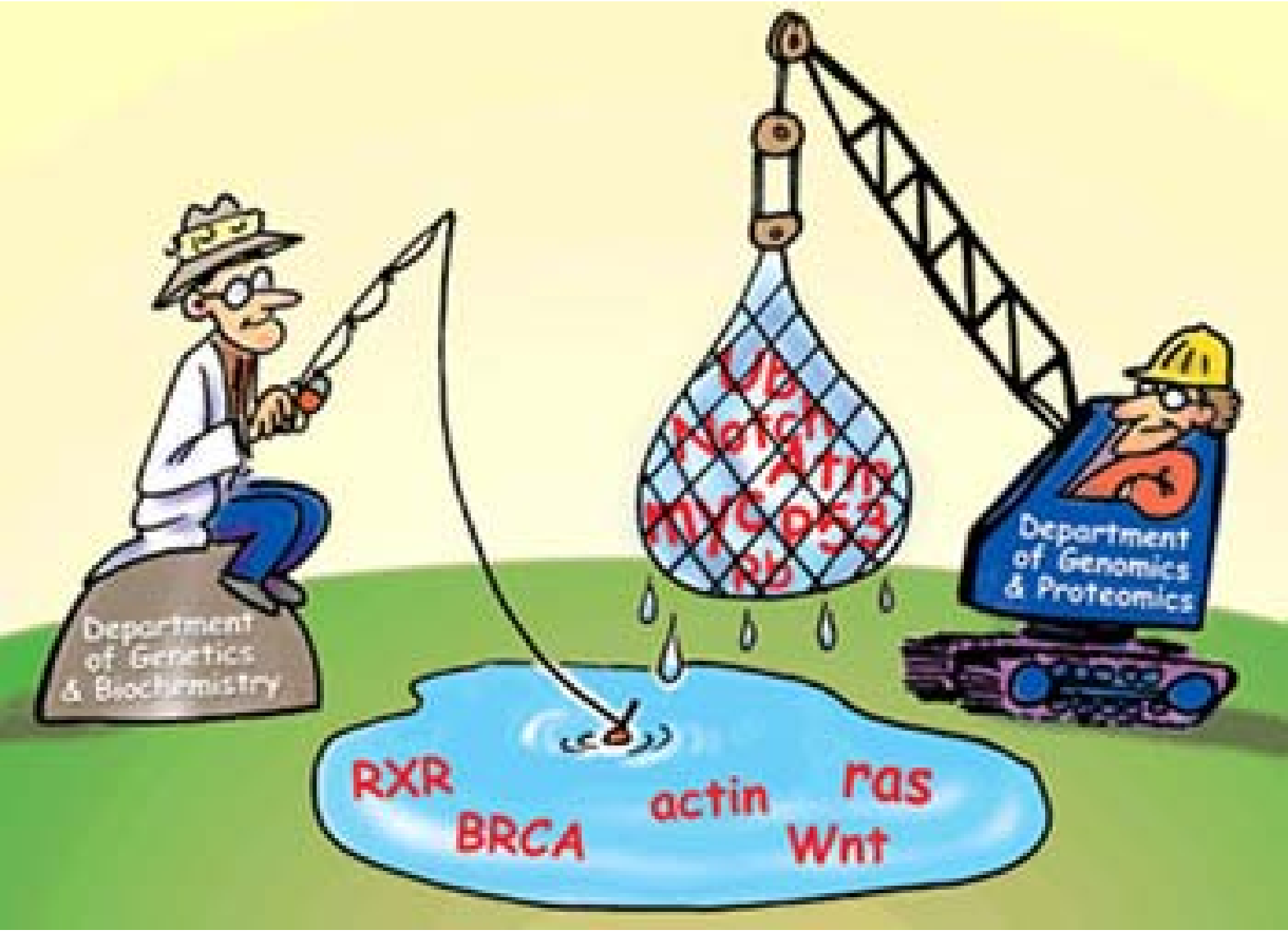


F. candida in soil quality testing

Disadvantages:

- Labor intensive
- Long duration (28 days)
- No specific information on mode of action of a toxicant

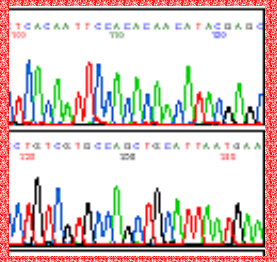


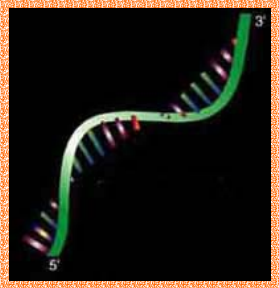

Department of Genetics & Biochemistry
Department of Genomics & Proteomics
Notch
Myc
p53
Rb
RXR
BRCA
actin
ras
Wnt

Transcriptomics

Genome

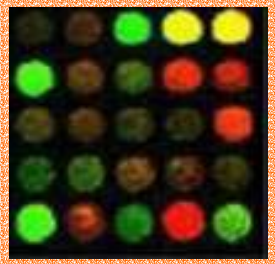
(Genes)

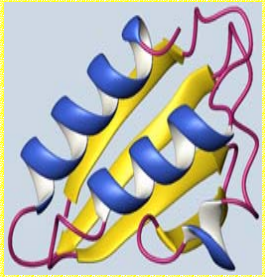




Transcriptome

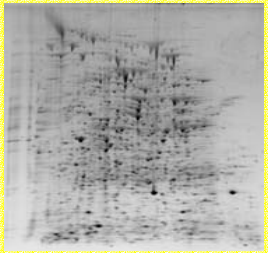
(mRNAs, Transcripts)

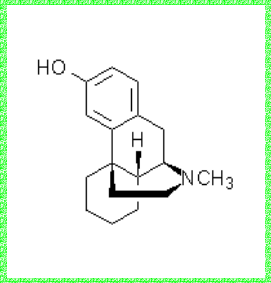




Proteome

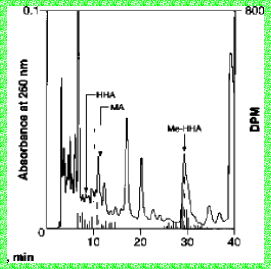
(Proteins)





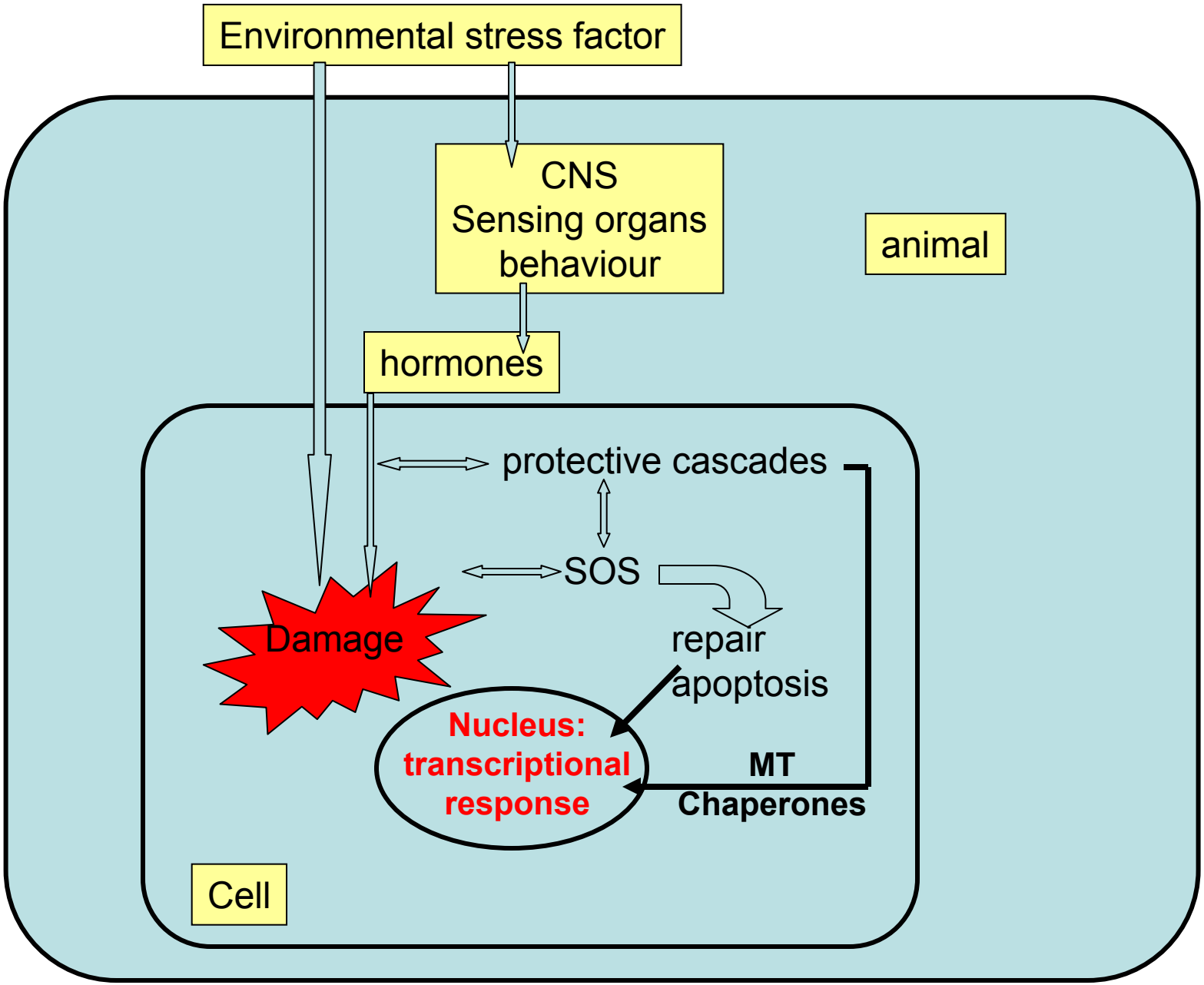
Metabolome

(Metabolites)



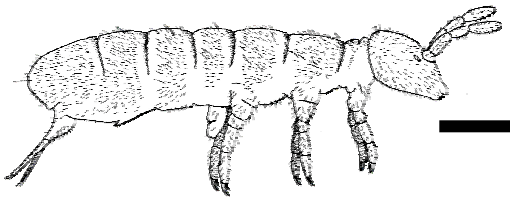
GENOTYPE

PHENOTYPE



Roelofs et al. Functional Ecology 2007

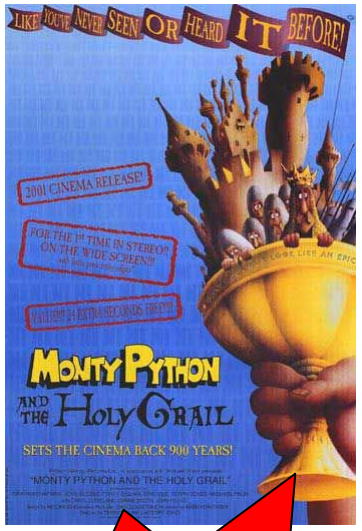
Overview



EST database:
Collembase.org

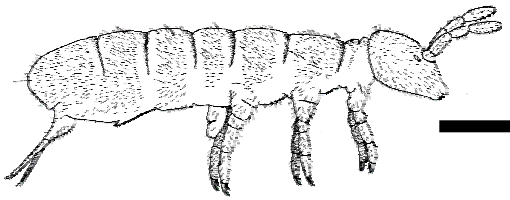
Microarray design &
Transcriptional profiling

Bioinformatic
analysis



Testing field-derived soil samples

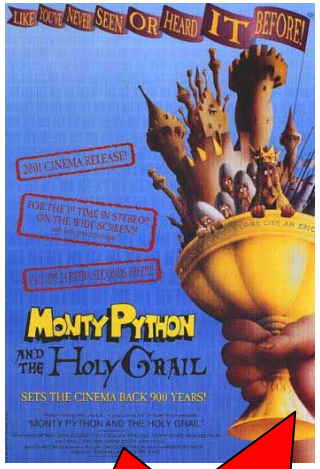
Overview



EST database:
Collembase.org

Microarray design &
Transcriptional profiling

Bioinformatic
analysis



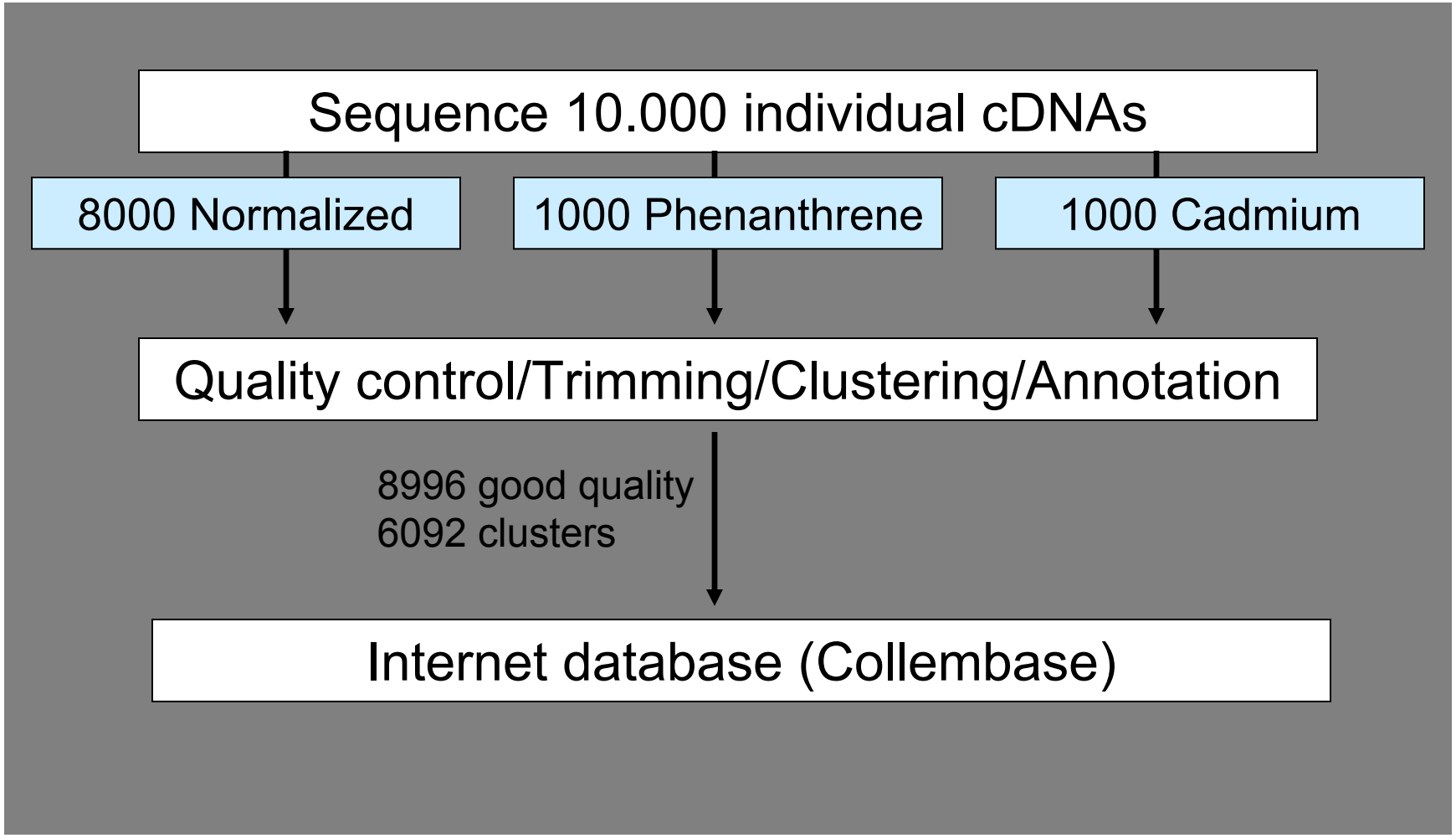
Testing field-derived soil samples

Sequence information *Folsomia candida*

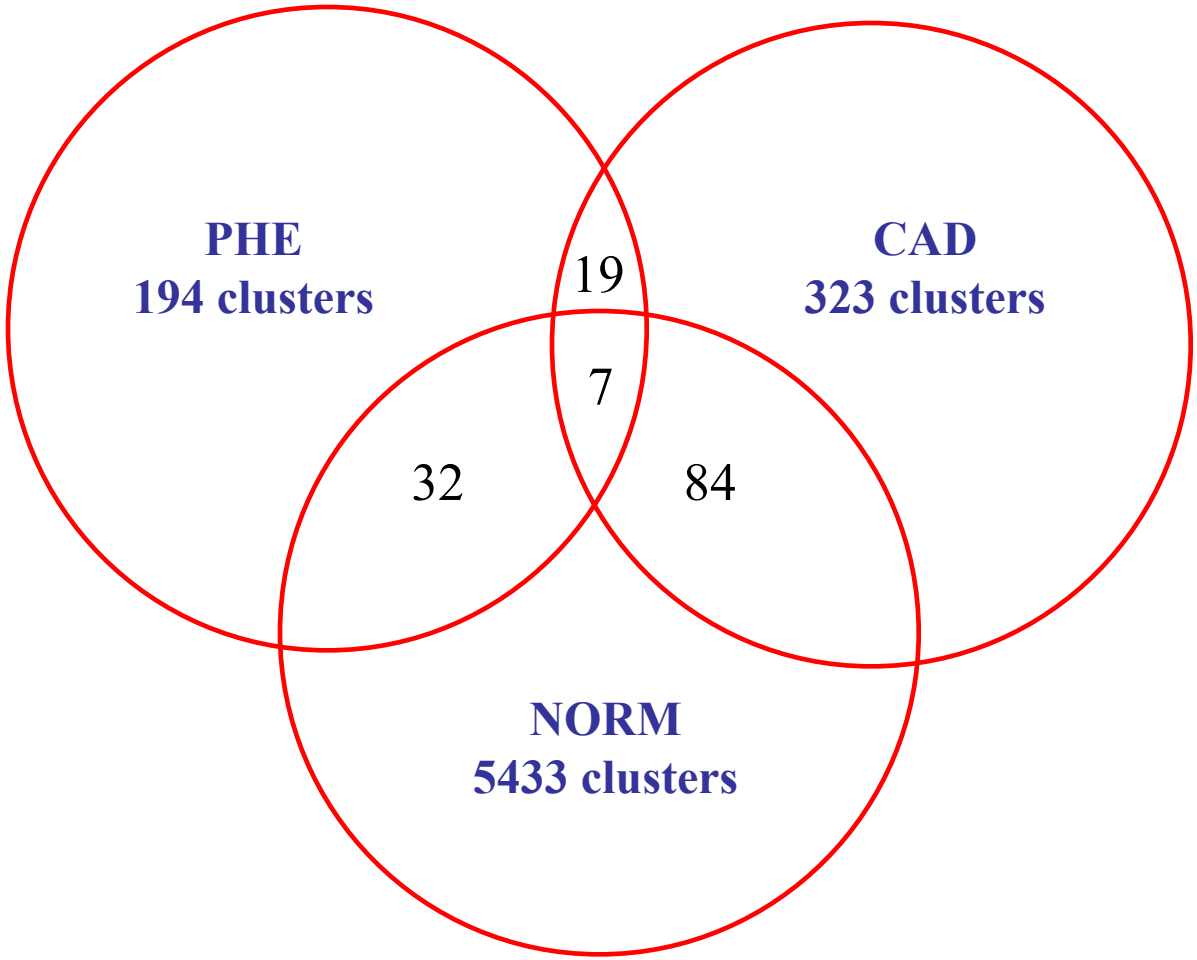
Gene discovery in a non-genomic model organisms

- Challenge: most ecologically relevant organisms are not supported by a genomic database
- Cost effective identification of ecologically important genes
- Important tool: cDNA synthesis and cloning

Expressed Sequence Tags



Overlap libraries



Collembase - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www.collembase.org/search.html Go Links

home search blast order news

www.Collembase.org

Search Collembase

Retrieve a specific cluster

Go

Enter a PartiGene-style cluster ID (e.g. Fol00451 / Orc00101)

Search by BLAST annotation

Select organism: Folsomia candida

Text Minimum e value

Search database Reset

To view all clusters for an organism just leave the text box blank

Search options

On this page it is only possible to search for blast annotations and/or for cluster IDs. To search for sequence similarity (BLAST) go to our Collembase BLAST page.

Database information

Quality check and analysis of the ESTs was done using the PartiGene pipeline. ESts were blasted against the Genbank non-redundant protein database. Blast results were stored in a PostgreSQL database.

Links

- Apterygota meeting 2006
- Collembola.org
- EIS
- Animal Ecology
- Steve Hopkin's Collembola site

Timmermans et al. BMC Genomics 2007

2 Sequences

Folsomia candida Cluster: Fol00643

Blast information for cluster Fol00643

Download all 2 seqs

Constituent sequences are:

FOLSXXT01672 /
FOLSXXT02565

Top 10 hits for blastx vs **nr** View all nr blastx results

```
Contig_01_RH17228p [Drosophila melanogaster] ref|NP_477176.1| CG 2e-51
Contig_01_ENSANGP00000020091 [Anopheles gambiae str. PEST] gb|EA 2e-51
Contig_01_cytochrome c 2e-51
Contig_01_Cytochrome c prf||1211285B cytochrome c 4e-51
Contig_01_Cytochrome c 5e-51
```

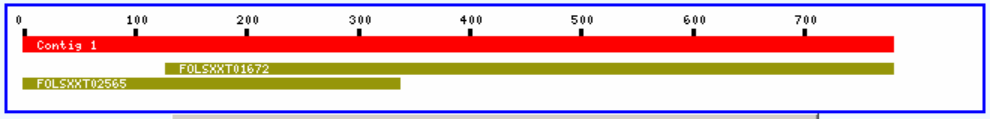
Number of contigs : 1

The following Blast searches are also available:
View **nr** results

Fol00643 - contig : 1

Length of Contig : 780

FOLSXXT01672 / FOLSXXT02565 /



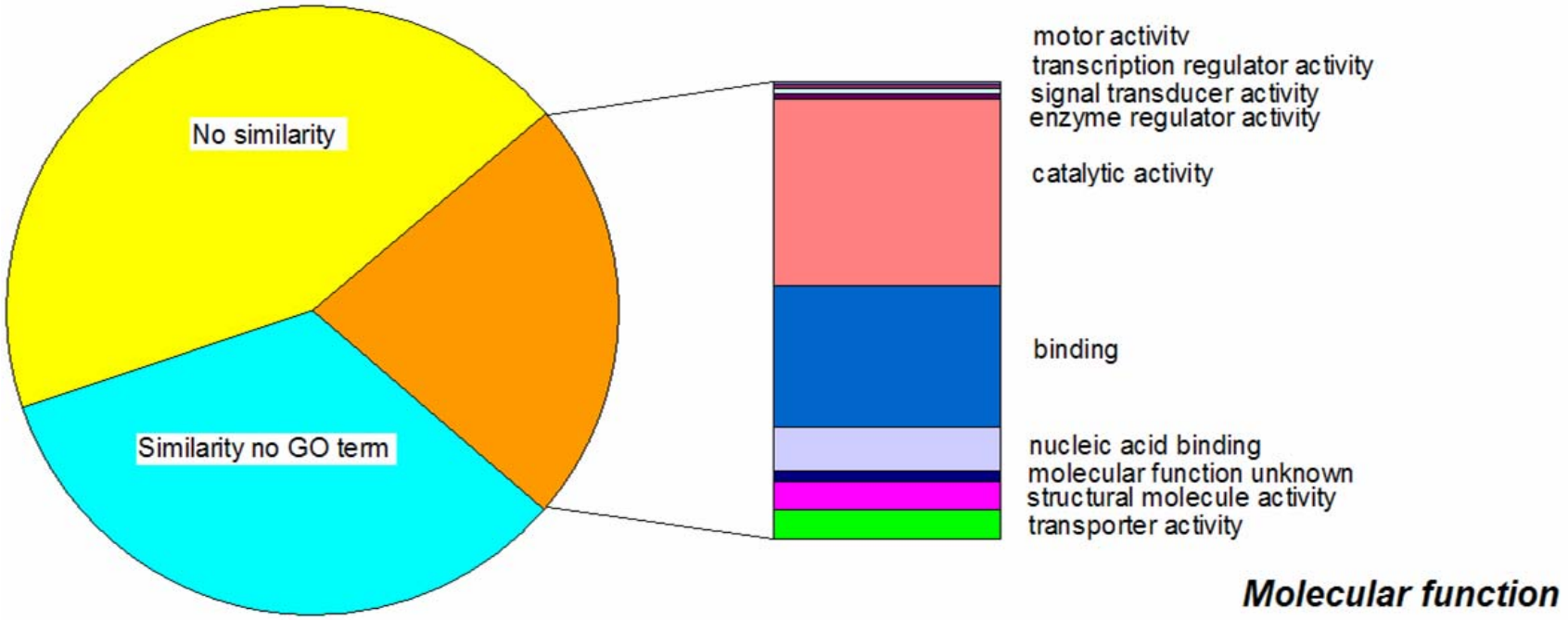
Download sequences associated with this contig

Sequence :

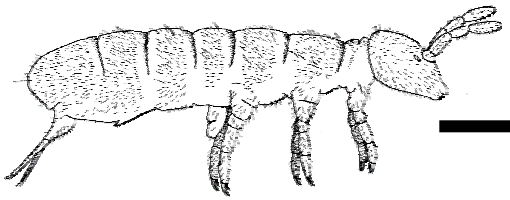
```
TTTTTTTTTTTTTTCGGTATTTCAGTATGTTCAACAAAATCAGTTTTTATTTATTTTTAAAAGTATACTGTATTAAACT
GAATCCTGCAGTTCTAATGTTTTTTTCTACTGACAAATTAAGTTATATATAGTATGCACTTATGCAATTCATAAATAATA
CACTTGAGGAGGATAACCAATTTTAAACTAAAGCTTTGTCCAGATTTCTTTGTAACATCAAAACCAATAAATCTTTGTGA
ACACAGTTCAAATCAAAAAAATGTCTACTCGTATAGTCTACGTTAATAATGTTACTGCATGAAGAACATCTTTTGATGGT
TGTCATTCACTTTCTAACATTCCTTCTTTCATTAAGTTAACATTTATATCACTAATCATTTTTTACTTTGTAGCAGAGCG
TAGATAGGCGATAAGGTCTGCACGTTCGTTGGGCTTCTTTAAACCAACAAACACCATCTTAGTACCAGGAATATATTTCT
```

Expressed sequence tag and gene function

Annotation	"BLAST Hit (e-value < 0.0001)	
D. melanogaster	32 %	
C. elegans	25 %	~180 yeast clusters removed
M. musculus	31 %	15 human sequences removed
All nr Genbank CDS	44 %	



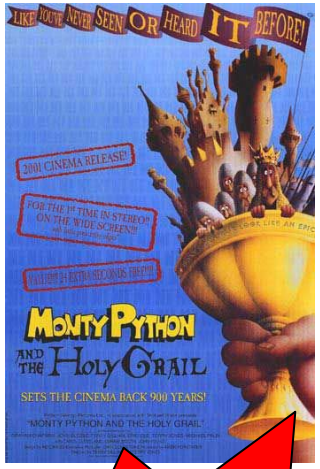
Overview



EST database:
Collembase.org

Microarray design &
Transcriptional profiling

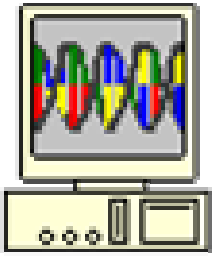
Bioinformatic
analysis

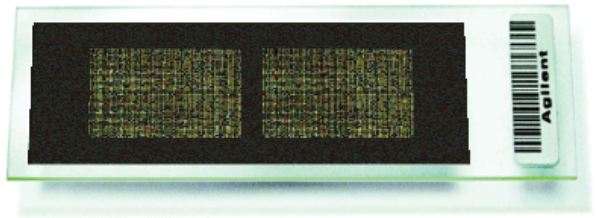


Testing field-derived soil samples

Microarray (Agilent platform)

Collembase



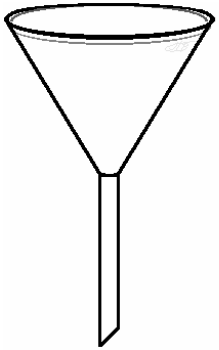


5064

6092

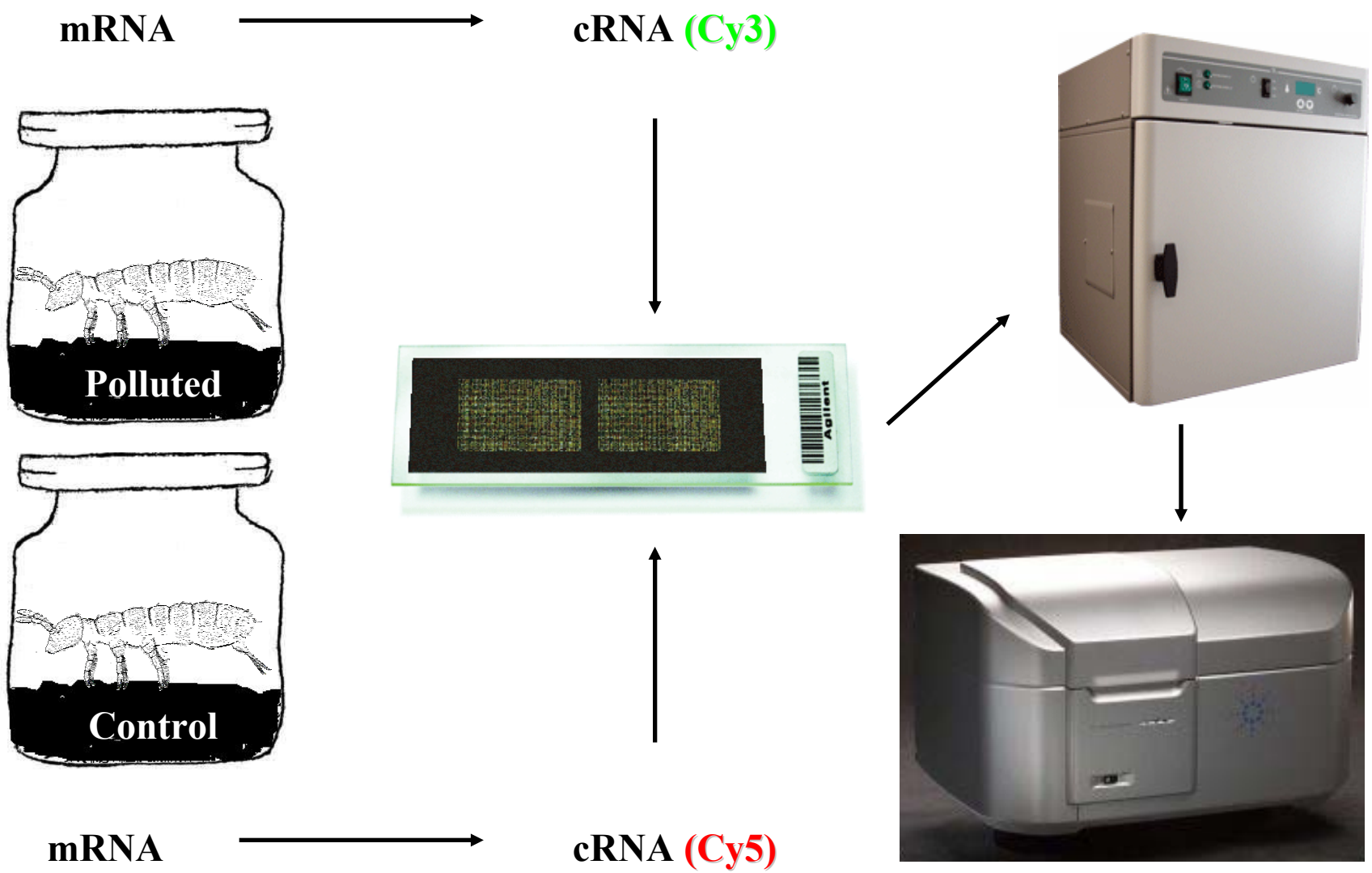
OligoArray 2.0

5702

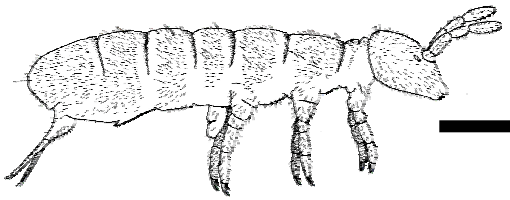


5300

Toxicant exposure



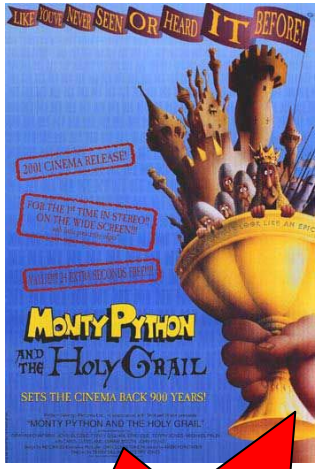
Overview



EST database:
Collembase.org

Microarray design &
Transcriptional profiling

Bioinformatic
analysis

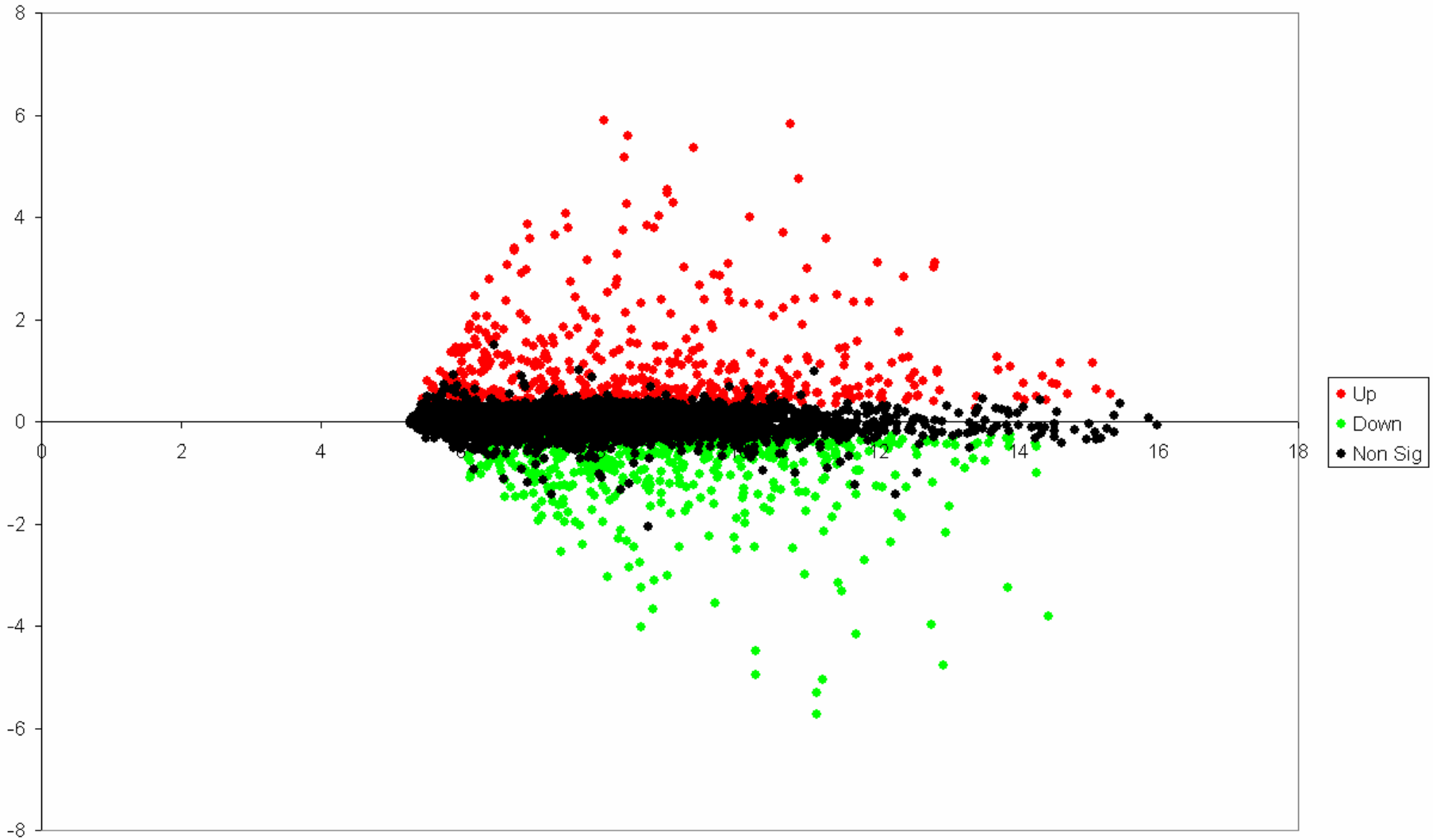


Testing field-derived soil samples

Data analysis (limma)

- Normalization of intensities (Lowess)
- Linear Models for Microarray Data
- Triplicate spots and four biological replicates
- *p* - value adjustment for multiple testing
  - Benjamini & Hochberg

Genes affected by cadmium



(adjusted $p < 0.05$)

Cadmium

Significant genes (adjusted $p < 0.05$)

Up: 513

15 Antibiotic proteins

11 Transporters (ABC, cation)

4 Glutathione-S-transferases

4 Myosin

1 Heat shock protein (Hsp70)

2 ATP synthases (subunits)

Down: 498

20 Ribosomal proteins

4 Proteasome

3 t-RNA synthetases

2 RNA polymerases

4 DNAJ

3 Fatty acid desaturases ($\Delta 5$, $\Delta 9$)

Gene Ontology (GO)

- Genes are annotated in GO terms
  - Molecular Function
  - Biological Process
  - Cellular Component

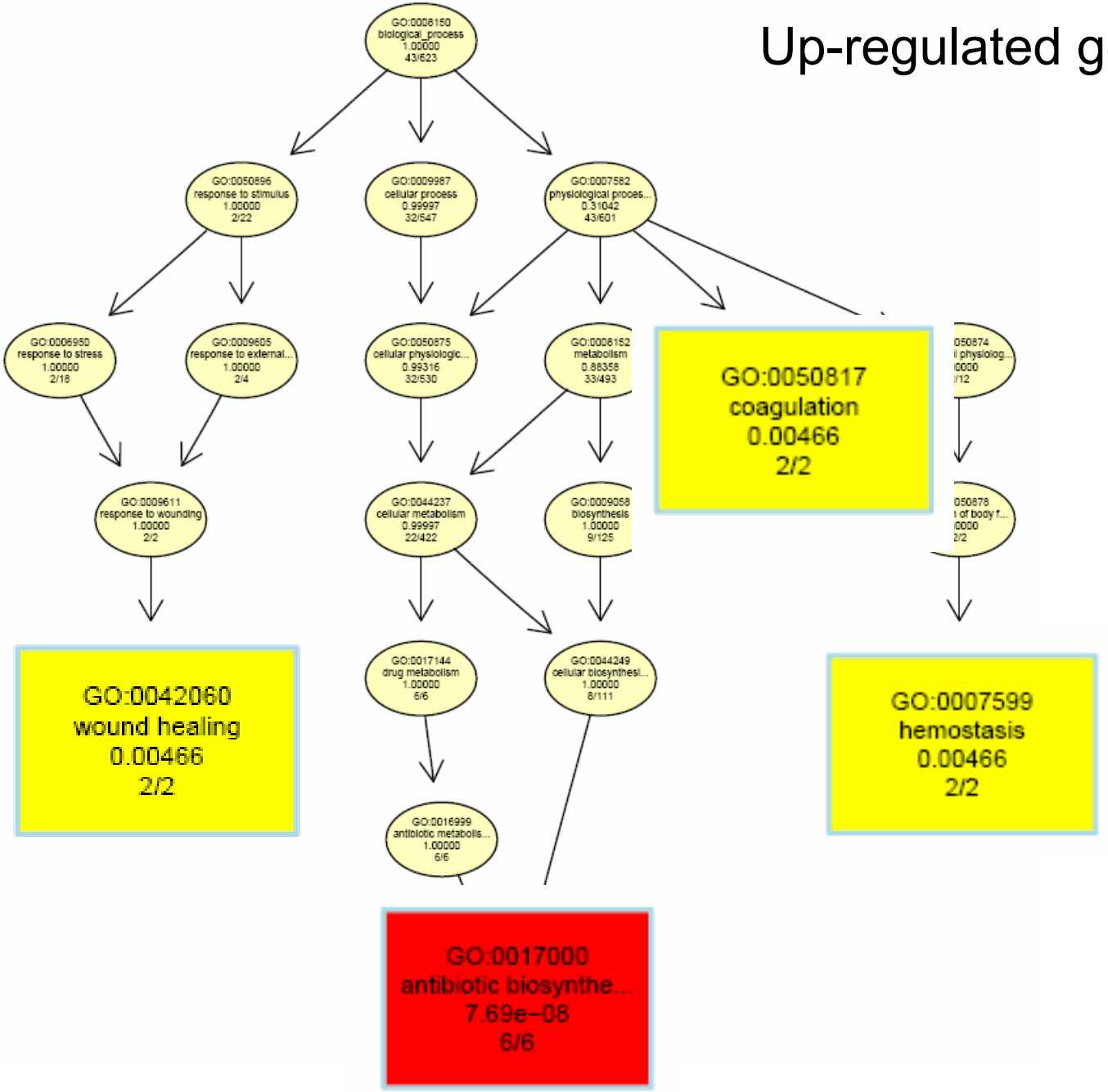
- Gene set enrichment
  - Fisher's exact test: topGO (Alexa, *et al.* 2006)
  - Example: Down regulation of Carbohydrate metabolism (GO:0005957):

    - 623 genes with GO terms on *F. candida* chip
    - 89 genes with GO terms are significantly down regulated upon Cd exposure
      - 89/623 = 1/7
    - 48 annotated genes involved in carbohydrate metabolism on chip:
      - Expected: 48 X 1/7 = 6.8
      - Observed: 13
    - Fisher's exact test p = 0.013 : Carbohydrate metabolism is significantly affected by Cd

GO results Cadmium for Biological process

GO.ID	Term	Annotated	Significant
GO:0017000	antibiotic biosynthesis	6	6
GO:0007599	hemostasis	2	2
GO:0042060	wound healing	2	2
GO:0050817	coagulation	2	2
GO:0006412	protein biosynthesis	68	19
GO:0019538	Carbohydrate metabolism	48	13
GO:0006094	Gluconeogenesis	2	2
GO:0006636	Fatty acid desaturation	3	2
GO:0043087	Regulation GTPase activity	3	2

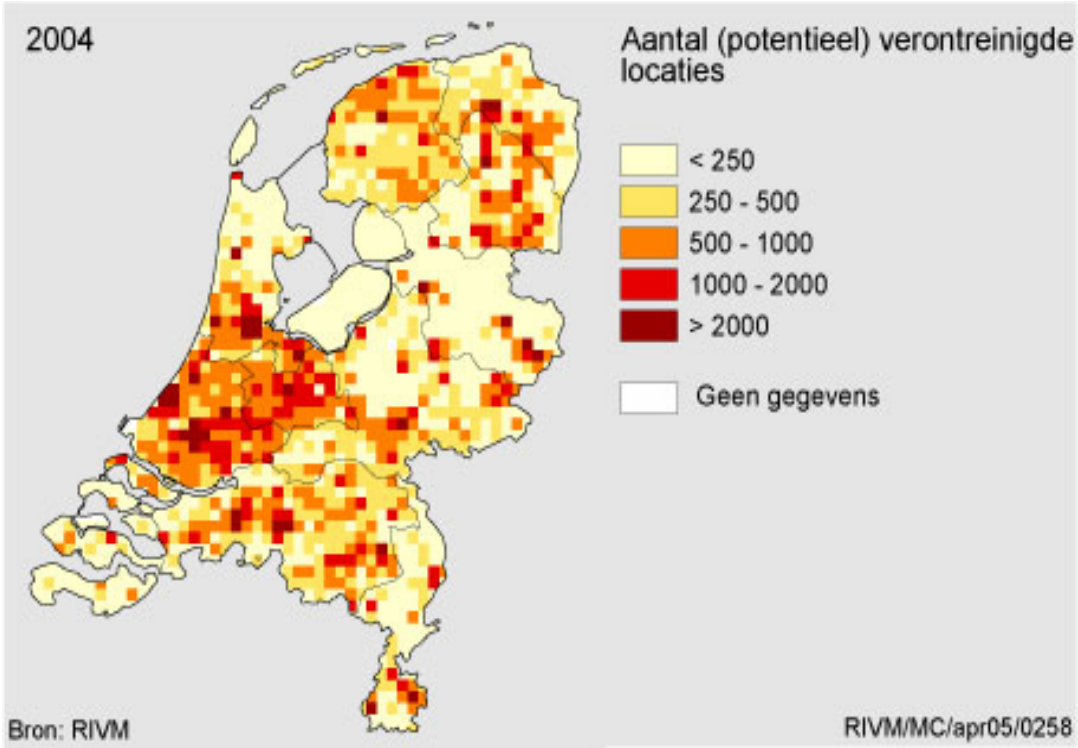
Up-regulated genes

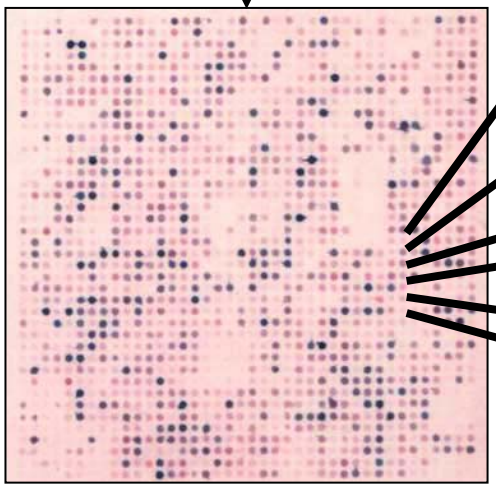


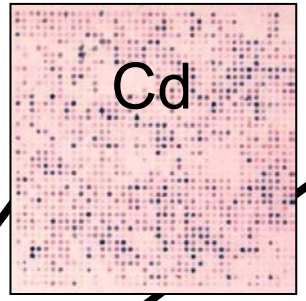
F. candida in soil quality testing

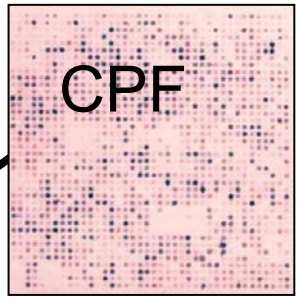
Disadvantages:

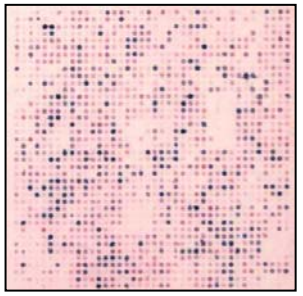
- Labor intensive
- Long duration (28 days)
- No specific information on mode of action of a toxicant

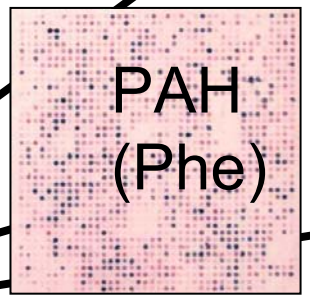


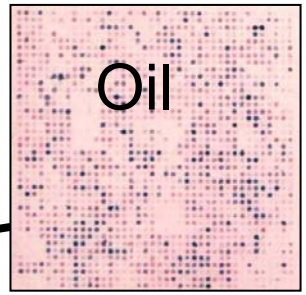


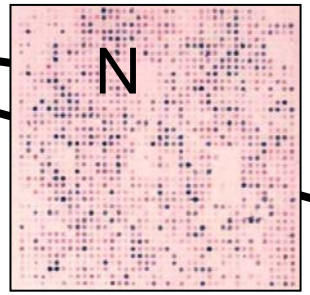


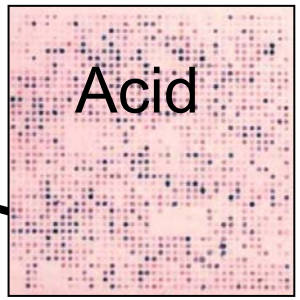












Gene expression profile of test organism exposed to suspect soil

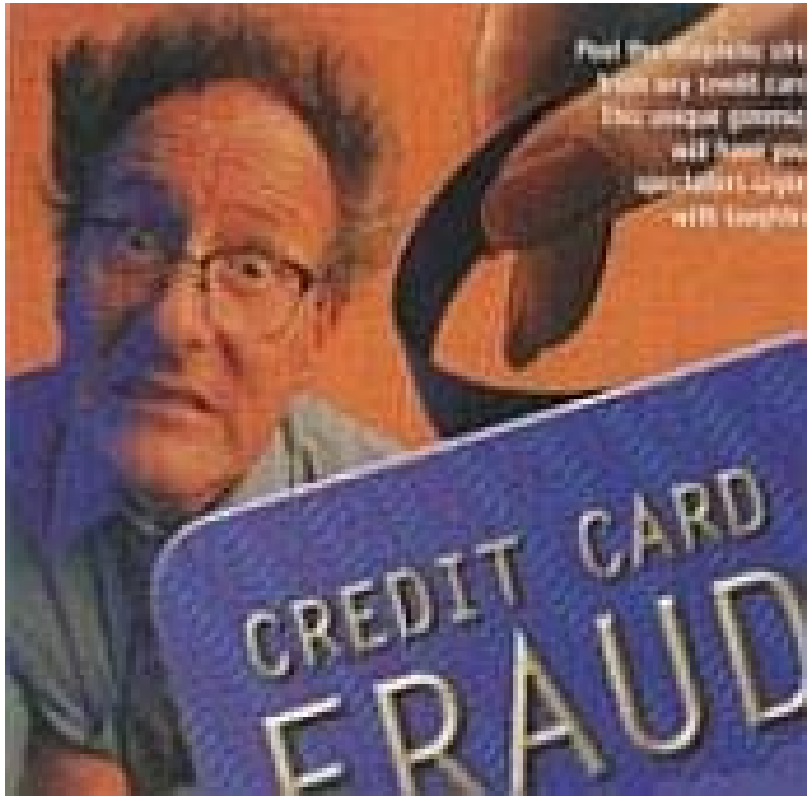
Comparison to database of reference expression profiles

Diagnosis: classification pollution type, risk assessment, advice on measures

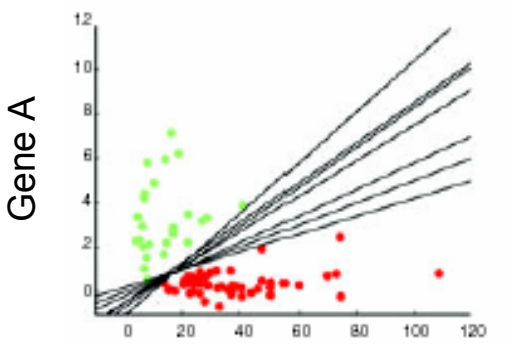
Van Straalen & Roelofs 2006

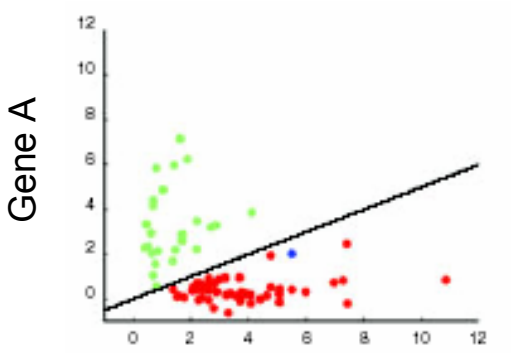
Class prediction with gene expression profiles: Support Vector Machine

- Algorithm that learns by example to assign labels to objects:
  - Recognize fraudulent credit card activities
  - Recognize hand written characters
- Cancer diagnostics: automatic classification of microarray data (prognosis/diagnosis)

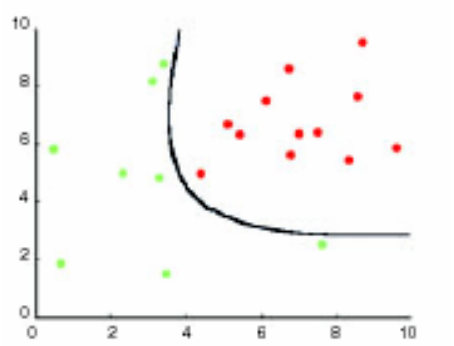


SVM at work: the separating hyperplane





● group 1 ● group 2 ● unknown



becomes linearly separable in 4 dimensional space (kernel function)

Linearly nonseparable data in 2 dimensional space

Noble, Nature biotechnology 2006

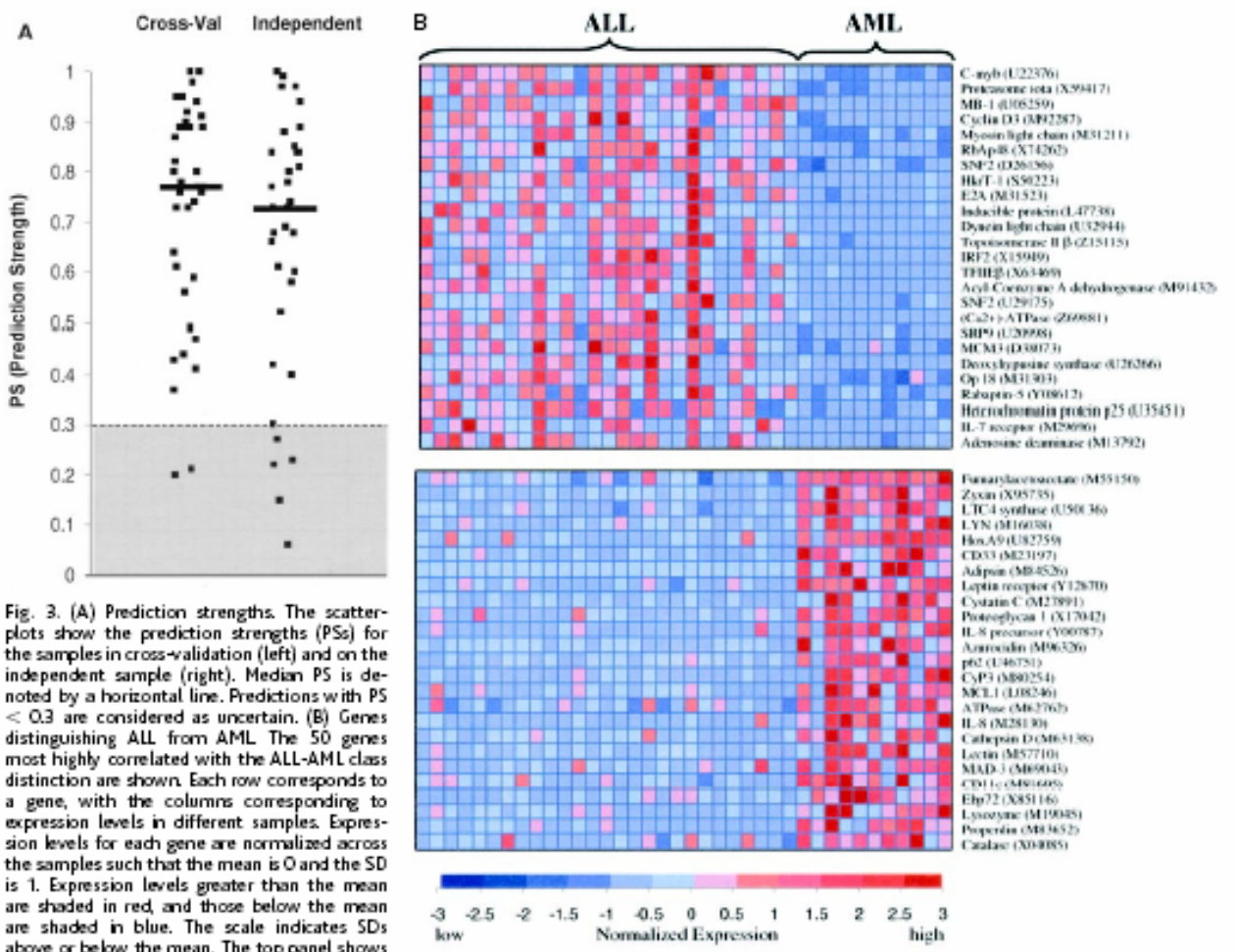
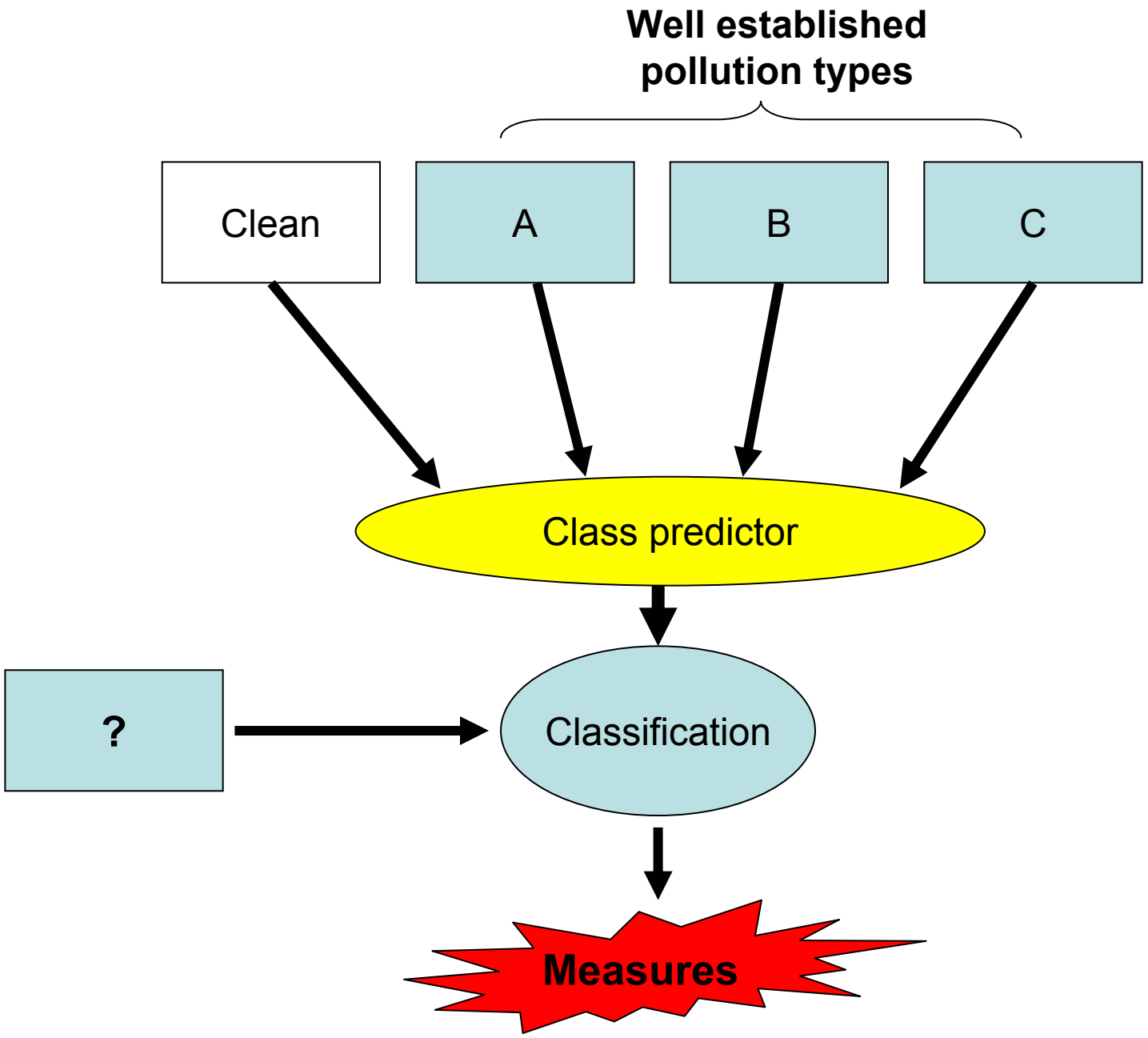


Fig. 3. (A) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (B) Genes distinguishing ALL from AML. The 50 genes most highly correlated with the ALL-AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and the SD is 1. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue. The scale indicates SDs above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML. Although these genes as a group appear correlated with class, no single gene is uniformly expressed across the class, illustrating the value of a multigene prediction method. For a complete list of gene names, accession numbers, and raw expression values, see www.genome.wi.mit.edu/MPR.

Golub et al. Science 1999

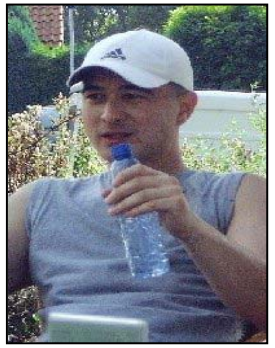

Well established
pollution types
Clean
A
B
C
Class predictor
?
Classification
Measures

Concluding remarks

- Collembase: a free accessible Genomic database
  - DNA sequence information
  - Future: Gene expression information
- Single toxicant exposure: mode of action
- Rearrange gene expression data sets in training groups to train a class predictor: diagnosis of unknown samples (fast, accurate and informative)

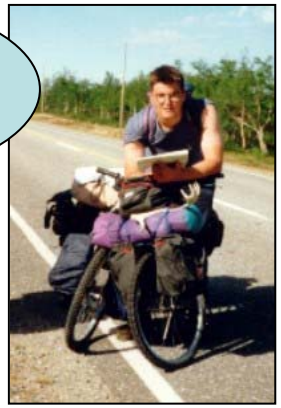
Martijn Timmermans

Nico van Straalen



Ben Nota

Janine Mariën



Thierry Janssens

Tjalf de Boer

Muriel de Boer