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
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
Transcriptomics

- **Genome** (Genes)
- **Transcriptome** (mRNAs, Transcripts)
- **Proteome** (Proteins)
- **Metabolome** (Metabolites)

**GENOTYPE**

**PHENOTYPE**
Environmental stress factor

CNS Sensing organs behaviour

hormones

Damage

protective cascades

SOS

repair apoptosis

MT Chaperones

Nucleus: transcriptional response

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

Sequence 10,000 individual cDNAs

8000 Normalized 1000 Phenanthrene 1000 Cadmium

Quality control/Trimming/Clustering/Annotation

8996 good quality 6092 clusters

Internet database (Collembase)
Overlap libraries

- PHE: 194 clusters
- CAD: 323 clusters
- NORM: 5433 clusters

Clusters overlap as follows:
- 19 clusters in PHE and CAD
- 7 clusters in PHE and NORM
- 32 clusters in CAD and NORM
- 84 clusters in all three
www.Collembase.org

Search Collembase

Retrieve a specific cluster

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

Search by BLAST annotation

Select organism: Foliolea candida

Text: [Input Field]

Minimum e value: [Input Field]

Search database

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

Top 10 hits for blastx vs nr

Number of contigs: 1

The following Blast searches are also available:

Fol00643 - contig:
1
Length of Contig: 780

Download sequences associated with this contig

Sequence:

TTTTTTTTTTTTTCTC9TAATTCGATGTTCAGCCAAAAAGCAATTTTTATTATTTTTAAAAAGATAGATTAAACT

Download all 2 seqs

Consistent sequences are:
FOLSXXT01672 / FOLSXXT02565
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

OligoArray 2.0

5702

5064

5300
Toxicant exposure

mRNA → cRNA (Cy3)

mRNA (Polluted) → cRNA (Cy5)

mRNA (Control) → cRNA (Cy5)
Overview

EST database: Collembase.org

Microarray design & Transcriptional profiling

Testing field-derived soil samples

Bioinformatic analysis
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

<table>
<thead>
<tr>
<th>GO.ID</th>
<th>Term</th>
<th>Annotated</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0017000</td>
<td>antibiotic biosynthesis</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>GO:0007599</td>
<td>hemostasis</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GO:0042060</td>
<td>wound healing</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GO:0050817</td>
<td>coagulation</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GO:0006412</td>
<td>protein biosynthesis</td>
<td>68</td>
<td>19</td>
</tr>
<tr>
<td>GO:0019538</td>
<td>Carbohydrate metabolism</td>
<td>48</td>
<td>13</td>
</tr>
<tr>
<td>GO:0006094</td>
<td>Gluconeogenesis</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GO:0006636</td>
<td>Fatty acid desaturation</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>GO:0043087</td>
<td>Regulation GTPase activity</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>
Up-regulated genes

- GO:0050817 coagulation 0.00466 2/2
- GO:0042060 wound healing 0.00466 2/2
- GO:0007599 hemostasis 0.00466 2/2
- GO:0017000 antibiotic biosynthesis 7.69e-08 6/6
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

Linearly nonseparable data in 2 dimensional space

becomes linearly separable in 4 dimensional space (kernel function)

Expression space

Noble, Nature biotechnology 2006
Fig. 3. (A) Prediction strengths. The scatter-plots 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, accessions numbers, and raw expression values, see www.genome.wustl.edu/MFR.
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)
Yes, you look pale; I feel much more colorful in clean soil.

Brrrr.. Living in polluted soil is rough.