Available online at www.sciencedirect.com
ScienceDirect
PHYTOCHEMISTRY
Phytochemistry 68 (2007) 19–32
www.elsevier.com/locate/phytochem

Review

Development of bioinformatic tools to support EST-sequencing, *in silico*- and microarray-based transcriptome profiling in mycorrhizal symbioses

Helge Küster [a,b,*], Anke Becker [a], Christian Firnhaber [a], Natalija Hohnjec [a,b], Katja Manthey [j], Andreas M. Perlick [j], Thomas Bekel [c], Michael Dondrup [b,c], Kolja Henckel [c], Alexander Goesmann [c], Folker Meyer [c], Daniel Wipf [d], Natalia Requena [e], Ulrich Hildebrandt [f], Rüdiger Hampp [g], Uwe Nehls [g], Franziska Krajinski [h], Philipp Franken [i], Alfred Pühler [j]

[a] *Institute for Genome Research, Center for Biotechnology (CeBiTec), Universität Bielefeld, Universitätsstrasse 25, 33594 Bielefeld, Germany*
[b] *International Graduate School in Bioinformatics and Genome Research, Center for Biotechnology (CeBiTec), Universität Bielefeld, Universitätsstrasse 25, 33594 Bielefeld, Germany*
[c] *Bioinformatics Resource Facility, Center for Biotechnology (CeBiTec), Universität Bielefeld, Universitätsstrasse 25, 33594 Bielefeld, Germany*
[d] *Institut für Zelluläre und Molekulare Botanik, Rheinische Friedrich-Wilhelms-Universität Bonn, Kirschallee 1, 53115 Bonn, Germany*
[e] *University of Karlsruhe, Institute of Applied Biosciences, Fungal-Plant Interactions Group, Hertzstrasse 16, 76187 Karlsruhe, Germany*
[f] *Julius-von-Sachs-Institut für Biowissenschaften, Lehrstuhl für Botanik II - Ökophysiologie und Vegetationsökologie, Universität Würzburg, Julius-von-Sachs-Platz 3, 97082 Würzburg, Germany*
[g] *Eberhard-Karls-Universität Tübingen, Botanisches Institut, Physiologische Ökologie der Pflanzen, Auf der Morgenstelle 1, 72076 Tübingen, Germany*
[h] *Institut für Pflanzengenetik, Abteilung II Pflanzenbiotechnologie, Universität Hannover, Herrenhäuser Straße 2, 30419 Hannover, Germany*
[i] *Institut für Gemüse- und Zierpflanzenbau, Abteilung Pflanzenernährung, Theodor-Echtermeyer-Weg 1, 14979 Grossbeeren, Germany*
[j] *Lehrstuhl für Genetik, Fakultät für Biologie, Universität Bielefeld, Universitätsstrasse 25, 33615 Bielefeld, Germany*

Received 27 April 2006; received in revised form 9 June 2006
Available online 1 November 2006

Abstract

The great majority of terrestrial plants enters a beneficial arbuscular mycorrhiza (AM) or ectomycorrhiza (ECM) symbiosis with soil fungi. In the SPP 1084 ''MolMyk: Molecular Basics of Mycorrhizal Symbioses'', high-throughput EST-sequencing was performed to obtain snapshots of the plant and fungal transcriptome in mycorrhizal roots and in extraradical hyphae. To focus activities, the interactions between *Medicago truncatula* and *Glomus intraradices* as well as *Populus tremula* and *Amanita muscaria* were selected as models for AM and ECM symbioses, respectively. Together, almost, 20.000 expressed sequence tags (ESTs) were generated from different random and suppressive subtractive hybridization (SSH) cDNA libraries, providing a comprehensive overview of the mycorrhizal transcriptome. To automatically cluster and annotate EST-sequences, the BioMake and SAMS software tools were developed. In connection with the eNorthern software SteN, plant genes with a predicted mycorrhiza-induced expression were identified. To support experimental transcriptome profiling, macro- and microarray tools have been constructed for the two model mycorrhizae, based either on PCR-amplified cDNAs or 70mer oligonucleotides. These arrays were used to profile the transcriptome of AM and ECM roots under different conditions, and the data obtained were uploaded to the ArrayLIMS and EMMA databases that are designed to store and evaluate expression profiles from DNA arrays. Together, the EST- and transcriptome databases can be mined to identify candidate genes for targeted functional studies.

* Corresponding author. Address: Institute for Genome Research, Center for Biotechnology (CeBiTec), Universität Bielefeld, Universitätsstrasse 25, 33594 Bielefeld, Germany. Tel.: +49 0 521 106 4819; fax: +49 0 521 106 5626.
E-mail address: Helge.Kuester@Genetik.Uni-Bielefeld.DE (H. Küster).

0031-9422/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2006.09.026

Keywords: *Glomus intraradices*; *Medicago truncatula*; *Populus tremula*; *Arbuscular mycorrhiza*; Bioinformatics; Ectomycorrhiza; Ectomycorrhizal fungi; EST-Sequencing; EST-Clustering; *in silico* Profiling; Macroarrays; Microarrays; Transcriptome database

Contents

1. Introduction

Mycorrhizal symbioses between higher plants and soil fungi are the most widespread beneficial plant-microbe interactions on earth. The two most common mycorrhizae are referred to as arbuscular mycorrhiza (AM) and ectomycorrhiza (ECM), respectively. It can be estimated that appr. 90% of all families of land plants establish mycorrhizal interactions (Smith and Read, 1997).

From a functional point of view, mycorrhizae are characterized by the transfer of limiting nutrients, in particular phosphorus and nitrogen, from the fungal hyphae to the plant. In exchange for this improved nutrient uptake, plants deliver carbon compounds to the symbiotic fungi, consuming up to, 20% of the plant photosynthate in case of AM (Harrison, 2005). Apart from the direct beneficial effects resulting from an improved nutrition, indirect benefits of an efficient mycorrhization can be seen in an enhanced plant resistance against abiotic and biotic stress (Smith and Read, 1997).

Molecular research in particular on AM symbioses has for a long time been hampered by the obligate biotrophic lifestyle of glomeromycotain fungi and by the asynchronous development of infection and colonization (Gianinazzi-Pearson and Brechenmacher, 2004), with comparably few molecular data being available only a couple of years ago (Franken and Requena, 2001). With the emergence of high-throughput genomics technologies, discovery science experiments relying on the identification of differentially expressed genes by untargeted transcriptome profiling became possible. Since these untargeted approaches are not impaired by the growth properties of mycorrhizal fungi, their application was particularly promising for research on AM and ECM interactions.

In the past years, different expression profiling strategies were pursued to identify genes activated during AM and ECM interactions (e.g. Journet et al., 2002; Liu et al., 2003; Hohnjec et al., 2005; Moreau et al., 2005). These strategies based on (1) high-throughput EST-sequencing, EST-clustering and annotation, (2) *in silico* profiling of mycorrhiza-related gene expression and (3) transcriptome profiling relying on microarrays. The SPP 1084 ''MolMyk: Molecular Basics of Mycorrhizal Symbioses'' was established to support such genomics experiments by sharing up-to-date bioinformatics and functional genomics technologies.

Genomics approaches greatly benefit from the identification of suitable model systems supported by a large community of researchers. In case of AM, two legumes were selected that proved to be excellent model systems: *Medicago truncatula* (Barker et al., 1990) and *Lotus japonicus* (Handberg and Stougaard, 1992). In addition to well-advanced genome sequencing projects and the existence of comprehensive mutant collections (Tadege et al., 2005; Udvardi et al., 2005; Young et al., 2005; Town, 2006; Sato and Tabata, 2006), more than 330,000 ESTs were generated in the two legumes (Quackenbush et al., 2000), and several expression profiling tools have been developed in the legume community (Journet et al., 2002; Colebatch et al., 2002, 2004; Kouchi et al., 2004; Küster et al., 2004; Hohnjec et al., 2005; Lohar et al., 2005).

A fringe benefit of selecting legumes as AM models can be seen in the fact that this plant family is able to enter a second important plant-microbe symbiosis, the development of nitrogen-fixing root nodules (Brewin, 1991; Schultze and Kondorosi, 1998). Interestingly, the existence of common signaling cascades allows the conclusion that nodule symbioses modified ancient signal recognition and signal transduction pathways that initially evolved to facilitate a successful colonization of plant tissues by AM fungi (Kistner and Parniske, 2002).

Complementing genomics acitivities in legume macrosymbionts, a genome project has been launched for *Glomus intraradices* (http://darwin.nmsu.edu/~fungi/), a model fungus from the *Glomeromycota* (Schüssler et al., 2001) that is characterized by a haploid genome of only ~15 Mb with few repetitive sequences (Martin et al., 2004; Lammers et al., 2004). The availability of genome sequences for two model legumes and a model fungus is expected to further advance our understanding of AM symbioses in the forthcoming years, although the lack of an efficient transformation system for the fungal symbiont will remain a major drawback.

From a genomics point of view, *Populus spec.* (poplar) is probably the best model currently available for ectomycorrhizal symbioses. The *Populus trichocarpa* genome sequence was reported recently (http://genome.jgi-psf.org/Poptr1/Poptr1.home.html), deep EST-sequencing covering different organs and growth conditions was pursued (Sterky et al., 2004; Sterck et al., 2005) and expression profiling tools have been developed (Moreau et al., 2005). With respect to ectomycorrhizal fungi, *Laccaria bicolor* has recently been chosen for a genome sequencing project (Martin et al., 2004; Lammers et al., 2004), complementing the collection of genomes with importance for mycorrhizal symbioses. In contrast to AM, the ECM symbiosis evolved multiple times (Kretzer and Bruns, 1999; Graham and Miller, 2005), resulting in more heterogeneous characteristics (Martin et al., 2001; Smith and Read, 1997). Consequently, only one model system will not be sufficient to address all topics related to ectomycorrhiza.

To provide molecular information on AM and ECM, the SPP 1084 "MolMyk: Molecular Basics of Mycorrhizal Symbioses" from the beginning pursued transcriptomics projects that focused on the interactions of *M. truncatula* with *G. intraradices* as a model for AM and of *P. tremula* with *Amanita muscaria* (fly agaric) as a model for ECM. In the course of these efforts, random and suppressive subtractive hybridization (SSH) cDNA libraries constructed by different partner laboratories were subjected to high-throughput EST-sequencing.

Together, almost, 20000 ESTs were obtained, providing a comprehensive overview of the mycorrhizal transcriptome under different conditions. Based on the transcript sequences, macro- and microarrays were constructed for the models, and these tools were used to profile the transcriptome of AM and ECM interactions (Küster et al., 2004; Manthey et al., 2004; Hohnjec et al., 2005). The sequence and expression data were processed and evaluated using a central web-based software (Dondrup et al., 2003), thus facilitating a selection of candidate genes for functional studies in the different partner laboratories. We here give an overview on the integration of EST-sequencing, *in silico*- and microarray-based transcriptome profiling approaches in the SPP; where appropriate referencing to more detailed analyses in other contributions from the present issue. Together with targeted functional studies of different mycorrhiza-related genes in the partner laboratories, the transcriptome profiling data obtained provide a molecular framework for the initiation, development and function of mycorrhizal interactions.

2. High-throughput EST-sequencing delivers snapshots of the plant transcriptome in arbuscular mycorrhizal and ectomycorrhizal roots

At the start of the SPP 1084 "Molmyk: Molecular Basics of Mycorrhizal Symbioses", information on genes expressed during the *M. truncatula*/*Glomus spec.* AM was quite limited (Franken and Requena, 2001). To overcome this deficit, we constructed a random cDNA library designated MtAMP, based on pooled tissue samples containing all major developmental stages of the *M. truncatula* and *G. intraradices* AM (Frenzel et al., 2005). This library was used to generate 3448 ESTs that exclusively covered 5′ regions of the directionally cloned cDNAs (Frenzel et al., 2005; cf. also the review by Krajinski et al. in the present issue). In parallel efforts from the Samuel Roberts Noble Foundation (Ardmore, USA) and the CNRS-INRA (Dijon, France), additional comprehensive EST-sequencing efforts were undertaken: First, 8567 ESTs from the random MtBC cDNA library covering the *M. truncatula*/*G. intraradices* AM were obtained by Journet et al. (2002) and second, 9030 ESTs were generated from the random MHAM and MHAM2 cDNA libraries representing the *M. truncatula*/*G. versiforme* AM (Liu et al., 2003).

Since random libraries tend to contain a high proportion of strongly expressed genes making the cloning of rare transcripts more difficult (Diatchenko et al., 1996), the suppressive-subtractive hybridization (SSH) cDNA library MtGIM was generated from *M. truncatula*/*G. intraradices* AM roots, subtracting sequences expressed in uninfected roots, phosphate-treated roots, *Sinorhizobium meliloti*-infected roots and *Aphanomyces euteiches*-infected roots. From this library, 1686 ESTs were obtained that proved to contain a high proportion of novel mycorrhiza-induced genes (Wulf et al., 2003; Frenzel et al., 2005; cf. also the review by Krajinski et al. in the present issue).

Since all ESTs were deposited in the GenBank dbEST collection (http://www.ncbi.nlm.nih.gov/dbEST/), they were also included in the TIGR *M. truncatula* Gene Index (http://www.tigr.org; Quackenbush et al., 2000). In the latest MtGI release 8 of January, 2005, 22731 from a total of 2,26,923 ESTs were generated from the AM libraries mentioned above, making the *M. truncatula*/*Glomus* interaction the best-studied AM model on the level of EST-sequencing (Hohnjec et al., 2006).

With respect to ectomycorrhiza, the model interaction of *Populus tremula* and *Amanita muscaria* has been studied by generating 6669 ESTs from a random cDNA library of poplar ECM roots (Table 1, Grunze et al., 2004; cf. also the review by Nehls et al. in the present issue). All ESTs were deposited in the GenBank dbEST

database, where they constitute the most comprehensive public collection of ECM transcript sequences; complementing the 2,61,885 non-symbiotic ESTs assembled in the TIGR Poplar Gene Index (http://www.tigr.org; Quackenbush et al., 2000).

3. Untargeted EST-sequencing specifies genes expressed in arbuscular mycorrhizal and ectomycorrhizal fungi

Since AM roots contain polyadenylated transcripts both from the plant and the mycorrhizal fungus, a certain proportion of these ESTs is of fungal origin. While *in silico* methods to *ab initio* predict the origin of the sequences proved to be difficult (Hraber and Weller, 2001), analyses based on comparisons to ESTs from non-mycorrhizal tissues as well as on PCR-amplifications from genomic DNA indicated that the proportion of fungal sequences in ESTs from AM roots is below 10% (Liu et al., 2003; Hohnjec et al., 2005). Consequently, ESTs generated from AM roots are primarily useful for identifying plant genes relevant for the establishment and function of the AM symbiosis, and ESTs obtained from fungal hyphae are required to derive insights into mycorrhiza-related processes occuring in the microsymbiont.

In case of the AM fungi *G. mosseae* and *G. intraradices*, several cDNA libraries were characterized by EST-sequencing. For *G. mosseae*, 658 ESTs were generated from an SSH library constructed from appressoria, in order to study fungal gene expression in the earliest stages of the symbiosis (Table 1, Breuninger and Requena, 2004; cf. also the review by Requena et al. in the present issue). With respect to *G. intraradices*, a total of, 2059 ESTs were obtained from different SSH libraries representing extraradical mycelia grown in the presence of different zinc levels (Ouziad et al., 2005), NaCl concentrations, and varying nitrogen sources (see Table 1 for details). In addition, the response to *Paenibacillus validus* in co-culture with asymbiotic *G. intraradices* mycelia was analysed (Table 1, Hildebrandt et al., 2006). Finally, two *G. intraradices* SSH libraries from germinated spores and extraradical mycelia were sequenced, together yielding 3708 ESTs (Table 1).

In total, 5602 *G. intraradices* ESTs larger than 50 bp were obtained, adding to another 3552 entries in the current GenBank dbEST release. EST-clustering (see below) revealed the presence of almost 3500 unique sequences,

Table 1
EST-sequencing efforts in the SPP 1084 "MolMyk: Molecular Basics of Mycorrhizal Symbioses"

Interaction/species	Library	ESTs	Tissues represented	References
P. tremula/ *A. muscaria*	Random	6669	*P. tremula* – *A. muscaria* ectomycorrhizal roots	Grunze et al. (2004)
M. truncatula/ *G. intraradices*	Random	3448	*M. truncatula* mycorrhizal roots harvested 5 weeks post inoculation with *G. intraradices*	Frenzel et al. (2005)
M. truncatula/ *G. intraradices*	SSH	1686	*M. truncatula* roots mycorrhized with *G. intraradices* subtracted by a mixture of cDNAs from non-infected roots, *S. meliloti*-infected roots, *A. euteiches*-infected roots and non-infected roots grown on 5 mM phosphate	Wulf et al. (2003), Frenzel et al. (2005)
A. muscaria	SSH	582	Fungal hyphae grown on 1% (w/v) (55 mM) glucose at low [0.025 g $(NH_4)_2HPO_4$/l] nitrogen supply; subtracted with fungal hyphae after 2 h in the absence of glucose and in the presence of high [0.25 g $(NH_4)_2HPO_4$/l] nitrogen supply	Rüdiger Hampp, Universität Tübingen
H. cylindrosporum	Random	459	Fungal hyphae grown on different nitrogen sources	Wipf et al. (2003)
G. mosseae	SSH	658	Appressorium stages subtracted with germinated sporocarps	Breuninger and Requena (2004)
G. intraradices	SSH	529	Extraradical mycelium grown with 100 μM Zn subtracted with extraradical mycelium grown in the presence of 1 μM Zn	Ouziad et al. (2005)
G. intraradices	SSH	742	Extraradical mycelium grown with 3.2 mM nitrate subtracted with extraradical mycelium grown in the presence of 3.2 mM glutamate	Ulrich Hildebrandt, Sandra Knolle, and Hermann Bothe; Universität zu Köln
G. intraradices	SSH	418	Mycelium grown asymbiotically in co-culture with *Paenibacillus validus* subtracted with asymbiotic mycelium grown in the absence of *P. validus*	Hildebrandt et al. (2006)
G. intraradices	SSH	370	Extraradical mycelium grown for 48 h in liquid M-medium containing 0.7% (w/v) NaCl subtracted with extraradical mycelium grown in liquid M-medium without NaCl	Ulrich Hildebrandt, Sandra Knolle, and Hermann Bothe; Universität zu Köln
G. intraradices	SSH	1822	Germinating spores subtracted with 25S und 18 S rDNA	Astrid Waschke and Philipp Franken, IGZ Grossbeeren
G. intraradices	SSH	1886	Extraradical mycelium subtracted with 25S und 18S rDNA	Astrid Waschke and Philipp Franken, IGZ Grossbeeren

Species studied by EST-sequencing are given in addition to the tissues represented. The type of library and the number of vector-clipped and phred13-processed ESTs is indicated.

ideally representing expressed *G. intraradices* genes. In addition to identifying genes activated in AM fungi, these clusters provide a valuable resource for gene prediction, once the *G. intraradices* genome sequence becomes available (Martin et al., 2004; Lammers et al., 2004).

Apart from *Glomus* spec., ectomycorrhizal fungi were studied by EST-sequencing. Here, an analysis of the poplar ECM library already revealed the presence of appr. 68% fungal ESTs. However, even though more fungal ESTs were present in this library, the number of fungal and plant TCs proved to be similar after clustering (1388 poplar and 1546 *Amanita* TCs), indicating a much higher redundancy in the fungal EST population (U. Nehls, Tübingen University, personal communication). In addition to the poplar library, an SSH cDNA-library from *A. muscaria* hyphae initially grown on glucose at low nitrogen supply (a situation as found in mycorrhizae) that subsequently were transferred to no glucose at high nitrogen supply (simulating conditions of soils supplied with nitrogen) was constructed and yielded 582 ESTs (Table 1, R. Hampp, Tübingen University, personal communication). In case of *Hebeloma cylindrosporum*, a random cDNA library from hyphae grown on different nitrogen sources was used to generate 459 ESTs (Table 1, Wipf et al., 2003; cf. also the review by Müller et al. in the present issue).

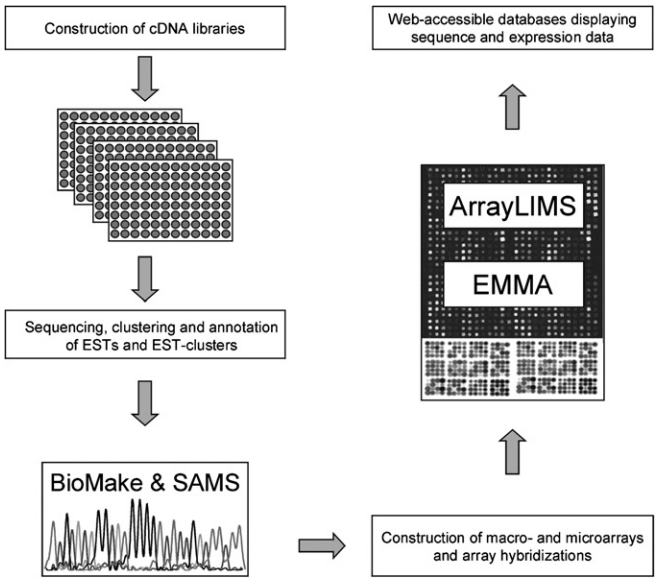


Fig. 1. Organization of expression profiling experiments in the SPP 1084 "MolMyk: Molecular Basics of Mycorrhizal Symbioses". Different random and suppressive-subtractive hybridization (SSH) cDNA libraries constructed by different project participants (Table 1) were sequenced. The web-based EST-clustering and cluster annotation software tools BioMake and SAMS were developed to derive annotations and to select EST-clusters for the construction of macro- and microarrays. Data obtained from array hybridizations were entered into the web-based ArrayLIMS and EMMA databases for storage of hybridization data and expression analysis. Sequence and expression data can be mined to select genes for targeted functional studies. The databases mentioned can be accessed via http://www.cebitec.uni-bielefeld.de/groups/brf/software.

Taken together, our untargeted EST-sequencing efforts in the mycorrhizal macro- and microsymbionts delivered a collection of almost, 20,000 mycorrhiza-related ESTs. This global overview of the transcriptome forms the basis for molecular studies targeted at functional analyses of selected genes in AM and ECM interactions (Fig. 1). Results from these experiments are reported in different reviews from the present issue.

4. Development of software tools for automatic clustering and annotation of ESTs

Untargeted EST-sequencing delivers thousands of raw sequence files that have to be processed prior to a functional analysis of the encoded gene products. To ease EST-processing, the BioMake and SAMS ("Sequence Analysis and Management System") tools were developed. These tools can be conFigured to restrict access to particular users or groups of users. Either software can be queried using standard web browsers, providing access independent of time and place, eliminating the need for local software installation and facilitating a constant central upgrade as well as an integration of novel analysis pipelines.

The BioMake tool was originally designed and implemented for an automatic, first-look analysis of ESTs and shotgun sequencing libraries. It is a simple, easy to install and to maintain open source system that automatically performs an analysis of raw sequence data as they come from the sequencing robots, subsequently presenting the results in a user-friendly web interface.

In BioMake, a single command will first generate automatic analyses consisting of three major steps. Starting with raw sequence data that are typically in scf-format (Dear and Staden, 1992), a normalization step is performed. Afterwards, every single sequence read is analyzed with respect to sequence quality. This process of removing the low quality regions is referred to as quality-clipping. Both the quality-clipping and the subsequent base calling steps are performed using the phred software (Ewing et al., 1998a,b). After removing low quality regions, vector and cloning adaptor contents of the sequence reads need to be identified, a process referred to as vector-clipping. This step is based on a BioMake pipeline applying a locally developed tool performing sequence comparisons to a database of standard vector and adaptor sequences enriched with user-contributed regions.

Once the data are free of low quality sequences and the vector contents was removed, the resulting cleaned sequence files are ready for analysis. Here, BioMake runs a variety of BLAST (Altschul et al., 1997) tools against different user-defined databases and present the results for each individual sequence in a web front-end.

Since the EST collections BioMake was applied to grew constantly, a new system had been developed to handle more comprehensive data sets. This software – the Sequence Analysis and Management System (SAMS;

T. Bekel, Bioinformatics Resource Facility, CeBiTec, Bielefeld University) – was originally implemented on the basis of the GenDB genome annotation system (Meyer et al., 2003) to facilitate quality control of data obtained during high-throughput phases of genome sequencing. In SAMS, sequence libraries, DNA and protein sequences, bioinformatics tools, tool results and annotations are modelled in an object-oriented approach. The software was implemented in the Perl (http://www.perl.org) programming language, and a relational MySQL database management system (http://www.mysql.com) was chosen as storage component.

The EST analysis pipeline implemented in SAMS is an improved and enhanced adaptation of the BioMake pipeline described above. Within the web-based sequence import, user-defined quality values can be set as well as library affiliations and parameters that govern vector-clipping. For quality-clipping, the phred13 standard was applied that proved to be both sufficient and efficient for EST-assemblies (Quackenbush et al., 2000). Since the clipping modules are implemented as tools applied on single sequences, the processing of EST reads can be distributed on a compute cluster where many compute nodes work in parallel.

Before EST sequences are annotated in SAMS, a clustering step is usually performed to remove redundancies by grouping the ESTs into clusters that are subsequently assembled into into TCs (Tentative Consensus sequences). Although SAMS provides user-configurable parameters to adjust the outcome of the clustering process, the standard TIGR (The Institute for Genome Research, http://www.tigr.org/) parameter set (Liang et al., 2000) is applied by default. Using these criteria, ESTs are clustered into one group if the following similarity conditions are fulfilled. First, two sequence reads must show an alignment of not less than 40 base pairs with at least 95% identity in a pairwise comparison. Second, flanking unmatched overhangs next to the alignment must not exceed a length of, 20 bp. After clustering, each cluster is assembled using CAP3 (Huang and Madan, 1999), producing TCs as sequence contigs and in addition a certain number of singlet ESTs. Together, TCs and singlet ESTs represent a nearly non-redundant collection of sequence data that form the basis for annotation and subsequent functional analyses.

Similar to the analysis of potential coding regions predicted on a bacterial genome sequence, EST and TC sequences are automatically processed and annotated in SAMS. For annotation, the Metanor automatic function prediction tool uses a combination of standard bioinformatics tools such as BLAST (Altschul et al., 1997)-based sequence comparisons, HMMer (Eddy, 1998)-based hidden Markov model scans and InterPro (Apweiler et al., 2001)-based searches for functional domains. These tools are run automatically on each sequence, leading to a consistent annotation assigning gene names, gene products, descriptions, functional categories, gene ontology (Ashburner et al., 2000) numbers, and other functional attributes to the TC and EST sequences.

5. Development of software tools for in silico expression profiling

The tissue origin of ESTs that were clustered in a particular TC can deliver information on the expression of the corresponding gene, an approach referred to as eNorthern, *in silico* or digital expression profiling (Alba et al., 2004). In addition to simply scoring EST-frequencies, mathematical approaches were proposed to associate a test statistic (*R*-value) with predicted differential expressions (Stekel et al., 2000). Such *in silico* studies were widely used in the last couple of years, and several hundred genes were predicted as up-regulated in AM and ECM (Journet et al., 2002; Manthey et al., 2004; Frenzel et al., 2005; Moreau et al., 2005; cf. also the reviews by Küster et al. and Krajinski et al. in the present issue).

Different eNorthern tools relying on the *R*-value test statistics are implemented in public databases, e.g. the TIGR Gene Indices for *M. truncatula* and poplar (http://www.tigr.org/, Quackenbush et al., 2000), as well as the MENS (Medicago EST Navigation System; Journet et al., 2002) and MtDB databases (Lamblin et al., 2003), both storing *M. truncatula* EST-clusters. To allow *in silico* analyses via the bioinformatics platform developed in the frame of the SPP 1084 "Molmyk: Molecular Basics of Mycorrhizal Symbioses", the SteN ("Statistical electronic Northern Blot") software was implemented as a module in SAMS, that way allowing the calculation of *R*-values according to Stekel et al. (2000) on the basis of ESTs generated in the project.

Designed to scan EST-collections for differential gene expression, the SteN software uses a relational mySQL database with an object-oriented O2DBI2 layer to store information on cDNA libraries, the ESTs generated and the TCs clustered. Once TCs have been assembled from a particular EST-project, the SteN database is updated automatically.

The web-based SteN interface allows an easy setup of queries by selecting one of the three states "USE", "DON'T USE" and "IGNORE" for each library in an EST project. "USE" means that a TC has to have at least one EST from this library, "DON'T USE" specifies that a TC is not allowed to have an EST from this library, and "IGNORE" means that the library is not considered in the query. Additionally, the logic operators "AND" or "OR" can be used to combine the selected libraries. That way, a typical query could read "Search for all TCs, which have ESTs from library_A OR library_B, but not from library_C. Library_D should be ignored."

The result of the query is a list with all TCs matching the selected search criteria. In addition, an *R*-value is calculated for each TC, as defined by Stekel et al. (2000). To interpret *R*-values, the option to ignore libraries in the query is essential. The *R*-value calculation is based on the total number of all ESTs in the query, and these are only taken from the "USE" and "DON'T USE" libraries, but not from libraries set to "IGNORED". This e.g. allows

to disqualify artificial clone collection such as "unigene sets" or other curated sets not directly representative for a particular tissue. Also, non-random libraries such as SSH libraries can be ignored, since they usually result in disproportional statistical parameters due to their biased mode of construction.

Based on the integration with SAMS, the SteN output Table includes references to the respective TC, information on the composition of ESTs from the different libraries, as well as the gene name, gene product, EC number, and functional categories. Using the calculated *R*-values, a graph is displayed showing *R*-values in relation to the number of TCs. In addition, average, median, minimum and maximum *R*-values are given.

SteN is able to execute more than one query at a time, allowing to combine the results later by using the logic operators "AND" or "OR". The result of this combination is stated in an extra list containing the additional information mentioned above. The results from SteN queries can be sorted and filtered as well as exported to spreadsheet software for documentation and further use.

Recently, SteN was applied to predict all AM-specific genes of *M. truncatula* (Hohnjec et al., 2006). This study identified 657 TCs of the TIGR *M. truncatula* Gene Index release 8 that were exclusively composed of ESTs from the five AM cDNA libraries mentioned above, including the MtAMP and MtGIM libraries constructed in the frame of the SPP (Wulf et al., 2003; Frenzel et al., 2005). The collection of AM-specific genes comprised several marker genes, where targeted analyses had demonstrated a specific expression in arbuscule-containing cells (e.g. Wulf et al., 2003; Frenzel et al., 2005; Hohnjec et al., 2005; Valot et al., 2006). As expected (Journet et al., 2002; Manthey et al., 2004), the overlap between *in silico* and experimental expression data dropped with decreasing EST counts per TC and lower *R*-values (Hohnjec et al., 2006). Nevertheless, *in silico* studies based on deep EST-sequencing can obviously identify candidate genes expressed at a low level (Manthey et al., 2004) or in a limited number of cells, making analyses on the level of ESTs complementary to microarray-based approaches.

6. Development of macro- and microarray tools to study arbuscular mycorrhizal and ectomycorrhizal symbioses

Expression profiling tools can be separated into spotted DNA arrays constructed from either PCR-products or 70mer oligonucleotides and DNA chips generated by the on-chip synthesis of gene-specific 25-base oligonucleotides (Becker, 2004). In case of spotted DNA arrays, macroarrays and microarrays have to be differentiated. Whereas macroarrays are obtained by gridding comparably high amounts (10–50 ng/spot) of DNA probes on nylon membranes, microarrays can be printed with a 50-fold less probe amount, thus allowing to generate substantially more arrays from the same initial quantum of probes. In

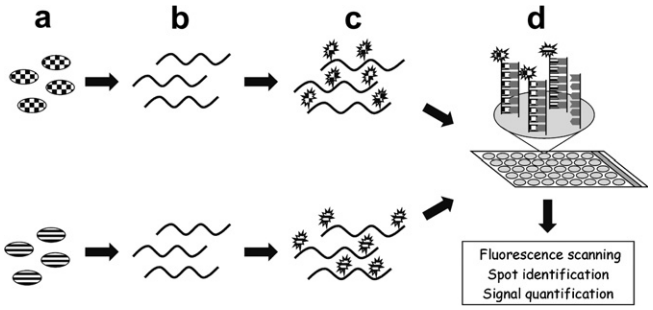


Fig. 2. Concept of a microarray hybridization using two combined fluorescently labeled targets representing an experimental and a control condition. (a) Biological samples in an experimental (dotted) and a control (striped) condition. (b) Pools of total RNA are isolated from the two samples. (c) Labeled targets derived from the two RNA populations. Two fluorescent dyes with different spectral properties are used for labeling. (d) Hybridizing the combined labeled targets to a microarray carrying gene-specific probes (grey circles) allows to determine ratios between fluorescence intensities of the two labels. This ratio is proportional to differences in the amount of corresponding transcripts in the original samples. Modified from Becker (2004).

contrast to macroarrays that are hybridized successively with radioactively labeled nucleic acids (designated "target") representing one biological condition, microarrays are hybridized simultaneously with combined fluorescently-labeled targets from one experimental and one control condition (Fig. 2). Thus, microarray approaches avoid array-to-array comparisons and allow to determine expression ratios in a single experiment (Quackenbush, 2002).

Prerequisite to the construction of spotted DNA arrays, a sufficient amount of sequence information derived from genomic or transcript sequences has to be available (Alba et al., 2004). Whereas PCR-products for prokaryotic arrays are obtained using gene-specific primer pairs designed against open reading frames identified in the genome sequence (Becker, 2004), eukaryotic PCR-based expression profiling tools rely on comprehensive collections of cDNAs (complementary DNAs). These are derived from the reverse transcription of mRNA and the subsequent cloning of synthesized cDNA copies (Alba et al., 2004). Subsequent to cDNA sequencing projects resulting in the generation of several thousand ESTs, different clustering algorithms are used to assemble overlapping EST consensus sequences designated either "EST-clusters" (Journet et al., 2002) or "Tentative Consensus sequences" (TCs, Quackenbush et al., 2000). That way, a reduction in the complexity of EST-collections is reached and near full-length transcript sequences mostly representing the coding sequence of genes can be assembled.

Subsequent to EST-clustering, particular cluster-representative ESTs have to be selected for the construction of cDNA-based arrays. Here, maximal gene-specificity is best achieved by choosing cDNAs located as far as possible in the 3′ region of the EST-cluster to minimize cross-hybridization with different members of multi-gene families (Journet et al., 2002; Küster et al., 2004). Once this selection is made, cluster-representative cDNAs have to be rearrayed

from the original cDNA libraries to obtain microplate sets suitable for robotic PCR-amplification. Since plasmids with a universal cloning site are usually used for the construction of cDNA-libraries, universal primers can be applied to PCR-amplify several thousand cDNAs. During the SPP 1084 "Molmyk: Molecular Basics of Mycorrhizal Symbioses", we initially started with the design and construction of macroarray that in the course of the project were substituted by cDNA- and 70 mer oligonucleotide microarrays (Küster et al., 2004).

In *M. truncatula*, we first constructed Mt6k-RIT (*Medicago truncatula* 6k Root Interaction Transcriptome) macro- and microarrays, based on a collection of 6359 EST-clusters representing 21,473 5′ and 3′ ESTs generated from cDNA libraries representing nitrogen-starved roots (MtBA), young root nodules (MtBB) and AM roots (MtBC), all described in detail by Journet et al. (2002). For each cluster, one representative was selected from the MtC libraries, with a preference for clones that were sequenced from both ends. Amongst those, the cDNA clone with the smallest possible size and the shortest overlap with predicted open reading frames was picked to facilitate PCR-amplification and to give more relative importance to the 3' UTR; in order to better discriminate between members of multigene families. In addition to 5648 gene-specific cDNAs that were PCR-amplified, transcript sequences representing different nodule- or mycorrhiza-related marker genes as well as several control genes were included (Brechenmacher et al., 2004; Küster et al., 2004; El Yahyaoui et al., 2004).

Based on 2516 ESTs generated from two random cDNA libraries deriving of developing flowers and pods, 1776 *M. truncatula* cDNAs representing non-redundant EST-clusters were PCR-amplified and subsequently used to upgrade the Mt6k-RIT collection to an Mt8k cDNA set. Due to a certain overlap between both sets of PCR-products, appr. 6300 different *M. truncatula* genes are represented by the Mt8k probe set (Firnhaber et al., 2005).

Mt6k-RIT and Mt8k cDNA arrays were shared between different participants of the SPP 1084 "Molmyk: Molecular Basics of Mycorrhizal Symbioses" as well as the European Union genome project "MEDICAGO: Integrated structural, functional and comparative genomics of the model legume *Medicago truncatula*", and were used to obtain expression profiles during different AM as well as during nodulation conditions (El Yahyaoui et al., 2004; Küster et al., 2004; Manthey et al., 2004; cf. also the reviews by Franken et al., Hause et al., Krajinski et al., and Küster et al. in the present issue). In the frame of other genomics projects focusing on the model legume *M. truncatula*, alternative expression profiling tools were generated, including a 2.5 k cDNA macroarray exclusively representing transcript sequences from AM roots (Liu et al., 2003) and a 6 k microarray based on a unigene cDNA collection (Lohar et al., 2005).

In recent years, cDNA probes have increasingly been replaced by 70mer oligonucleotides. These can be designed to be as gene-specific as possible and the construction of microarrays based on 70mer long oligonucleotides does not require error-prone rearrying and cDNA amplification steps (Alba et al., 2004). During the 1st International Conference on Legume Genomics and Genetics (St. Paul, Minnesota, 2002), the *Medicago* community decided to commission the synthesis of a 16 k 70mer oligonucleotide collection representing all TCs from the TIGR *M. truncatula* Gene Index version 5. This collection of probes was recently extended by 384 probes targeted against transcription factors and other regulators. These tools are referred to as Mt16kOLI1 (Hohnjec et al., 2005) and Mt16kOLI1Plus (Thompson et al., 2005), respectively. Based on the estimated gene count in *M. truncatula* (Town, 2006), it can be estimated that appr. 35% of all *M. truncatula* genes are represented on these microarrays. Based on Mt16kOLI1, Hohnjec et al. (2005) specified, 201 *M. truncatula* genes co-induced in the interaction with different AM fungi. Some of the marker genes identified were also found to be expressed in root

Table 2
Expression profiling tools generated in the SPP 1084 "MolMyk: Molecular Basics of Mycorrhizal Symbioses"

Species	Array name	Array type	Identifier	Probes on the array	References
M. truncatula	Mt6kRIT	cDNA macroarray	A-MEXP-80	Probes representing 5648 EST-clusters from *M. truncatula* root nodules, AM roots, and uninfected roots	Küster et al. (2004)
M. truncatula	Mt6kRIT	cDNA microarray	A-MEXP-81	Probes representing 5648 EST-clusters from *M. truncatula* root nodules, AM roots, and uninfected roots	Küster et al. (2004)
M. truncatula	Mt8k	cDNA microarray	A-MEXP-84	Mt6kRIT probe set plus 1776 probes representing EST-clusters from *M. truncatula* flowers and pods	Firnhaber et al. (2005)
M. truncatula	Mt16kOLI1	70mer oligonucleotide microarray	A-MEXP-85	Probes representing 16.086 tentative consensus sequences of the TIGR *M. truncatula* Gene Index version 5	Hohnjec et al. (2005)
M. truncatula	Mt16kOLI1Plus	70mer oligonucleotide microarray	A-MEXP-138	Mt16kOLI1 probe set plus 384 probes primarily representing transcription factors	Thompson et al. (2005)
P. tremula/ *A. muscaria*	Pt2.4kOLI1	70mer oligonucleotide microarray	A-MEXP-202	Probes representing 2350 EST-clusters from poplar ECM	Uwe Nehls, Universität Tübingen

Relevant features of the macro- and microarrays are indicated, together with their identifiers in the ArrayExpress (http://www.ebi.ac.uk/arrayexpress) database. Regular updates of array-related information are available under http://www.cebitec.uni-bielefeld.de/groups/glp/microarrays/.

nodules, adding to the observed transcriptional overlap during nodulation and mycorrhization (cf. also the review by Küster et al. in the present issue).

Due to the increasing use of long oligonucleotide microarrays, an expression profiling tool to study the poplar ectomycorrhizal symbiosis was exclusively constructed from a collection of 70mer probes. These were designed against 2.4 k EST-clusters from the*P. tremula* – *A. muscaria* ECM cDNA library (Table 1). The corresponding microarray is referred to as Pt2.4kOLI1 and was used to profile gene expression in different poplar tissues including ECM roots (cf. also the review by Nehls et al. in the present issue).

To allow deposition of microarray-based expression data in public databases, array design files for all macro- and microarray tools constructed in the frame of the SPP were submitted to the ArrayExpress (http://www.ebi.ac.uk/arrayexpress) database (Table 2). Regular updates of array-related information, such as improved annotation files, are available under http://www.cebitec.uni-bielefeld.de/groups/glp/microarrays/.

As for other plant species (Zhu, 2003), the *M. truncatula* and poplar field in the meantime moved from the project-driven construction of cDNA-macroarrays, cDNA-microarrays, and 70mer oligonucleotide microarrays to commercial Affymetrix chips, with ~61k *Medicago* and Poplar GeneChip® Genome Arrays being available (http://www.affymetrix.com). Due to a significant increase in the number of probes, the future application of these tools for transcriptome profiling in mycorrhizal roots can be expected to widen our insights into molecular processes relevant for AM and ECM.

7. Development of web-accessible databases for storage, evaluation and mining of expression profiles

One goal of the SPP 1084 "MolMyk: Molecular Basics of Mycorrhizal Symbioses" is to deliver integrated databases that store the microarray-based expression data obtained in the course of the project, together with relevant information on the experimental conditions profiled and the protocols used to obtain transcriptome profiles. Data integration can only be achieved via software tools fulfilling two requirements: (1) the possibility to upload raw data from microarray hybridizations together with a description of experimental conditions into a central data repository and (2) the possibility to combine these data sets in order to relate expression profiles obtained in different experiments, e.g. by hierarchical clustering (Slonim, 2002; Zhu, 2003).

Two databases were set up to fulfill these requirements: an ArrayLIMS (Laboratory Information Management System) for the storage of data on experimental conditions and for storing raw data, and the EMMA (EST Meets Micro Arrays) software for data evaluation and data mining. Applying the same access policy described for the BioMake and SAMS EST-databases, ArrayLIMS and EMMA can be queried using standard web browsers (Dondrup et al., 2003).

The ArrayLIMS system requires the user to deliver in a step by step process all information that is required by others to follow the setup of a microarray experiment. This approach is essential in order to comply to the MIAME (Minimal Information ON A Microarray Experiment; Brazma et al., 2001) standard that is prerequisite to a publication of expression data. In ArrayLIMS, the basic information to be entered relates to the biological conditions profiled, the material harvested and the exact growth conditions used. Based on these entries, next steps call for relevant features of the targets isolated from the biological material profiled, the target labeling, and the microarray hybridizations performed. In the ArrayLIMS database, identifiers for the microarrays to be used in a project are stored and any information entered can be linked to those identifiers. Finally, the microarray identifiers are linked to experimental raw data: tiff-files from microarray scanners and data-files derived from image processing. Here, the ArrayLIMS system is compatible with all commercial microarray scanners and with the output formats of major commercial image processing software. As a consequence of multiple uploads from different project participants, a central database storing information on all microarray hybridizations performed in the frame of the project is generated.

Subsequent to ArrayLIMS imports, microarray data can be used to set up experiments in EMMA (Dondrup et al., 2003), a web-based software designed to facilitate data analysis according to accepted standards for cDNA and 70mer oligonucleotide microarrays (Quackenbush, 2002; Leung and Cavalieri, 2003). The EMMA system represents an effort to build a microarray analysis software that simplifies complex analysis steps and queries of large datasets, while adhering to standards for data interchange and optimizing the inter-operation with other software. EMMA is written mainly in Perl while some portions are programmed in R (Ihaka and Gentleman, 1996) or Java. The data repository is implemented using O2DBI, an object-oriented code generator that significantly simplifies the creation of complex database applications. The application programmer's interface (API) of EMMA is implemented and documented by automatically generated Perl modules created by O2DBI that also contain additional functions manually added (Dondrup et al., 2003).

Fig. 3 illustrates key features of the architecture of the current EMMA release. This release includes different pipelines for data normalization (Yang et al., 2002), such as global mean normalization, lowess (locally weighted scatter plot smooth; Cleveland et al., 1992) normalization, and different print-tip based normalizations (Dondrup et al., 2003). Normalized data can be used to run Student's *t*-tests and LIMMA significance tests (Dudoit et al., 2002; Smyth et al., 2003) to identify differentially expressed genes, providing one-sided and two-sided pipelines for either query.

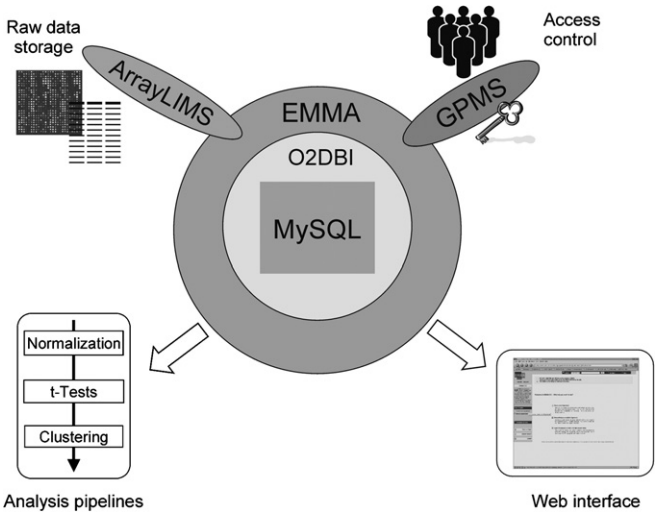


Fig. 3. Architecture of the ArrayLIMS/EMMA database. Raw data from microarray scanners and image processing software are stored in the ArrayLIMS database. From here, data can be imported into the EMMA software that builds on a MySQL database and an O2DBI object to database interface. Access to the databases is controlled by a General Password Management System (GPMS). EMMA provides different pipelines for data analysis, such as normalization, statistical tests and hierarchical clustering. Results are displayed as web frontends and can thus be queried and downloaded using standard web browsers.

Higher order analyses are supported by hierarchical clustering pipelines that display results in Java-applets providing several visualisation options to browse clustered datasets. All transformed datasets (normalized data, significance tests, hierarchical clusterings) are stored in the EMMA project database and can be downloaded for further analyses. The most recent features of ArrayLIMS and EMMA including constantly updated manuals can be retrieved from http://www.cebitec.uni-bielefeld.de/groups/brf/software.

Since the ArrayLIMS/EMMA software is also used in other genomics projects on the model legume *M. truncatula*, e.g. the European Union project GRAIN LEGUMES (http://www.eugrainlegumes.org) as well as previous European Union projects, a cross-project integration of expression data is possible. To facilitate submission of expression data to public databases, e.g. ArrayExpress (http://www.ebi.ac.uk/arrayexpress) and to allow a general exchange with other projects that use a different expression analysis software, EMMA was designed to comply to the MAGE-ML (Micro Array and Gene Expression Markup Language; Spellman et al., 2002) standard. This MAGE-compatibility lays the foundation for an integration of expression data from different transcriptomics projects.

An essential feature of the ArrayLIMS/EMMA software is a password management system that allows to fine-tune different levels of access to expression data. In the frame of the SPP, we employed an open access policy that allows project members to query all expression data stored in the central ArrayLIMS repository. This philosophy of data accessibility is prerequisite to an application of the advanced data mining strategies implemented in EMMA. Whereas global data mining approaches require the setup, configuration and detailed analysis of new groups of microarray hybridizations that were imported from ArrayLIMS, focused data mining allows to quickly solve more targeted questions. In the most simple query, these can relate to the retrieval of expression information for a gene of interest in specific or all conditions profiled, whereas more advanced queries require to build data mining pipelines using logic operators. These pipeline allow to return expression information for all genes annotated e.g. as transcription factor in selected or all experiments stored in the database.

8. Towards an integration of ESTs, *in silico* and experimental transcriptome profiles

The software systems BioMake, SAMS, SteN, ArrayLIMS, and EMMA were designed as specialized components for separate scopes; e.g. SAMS being used for EST-clustering and -annotation, and EMMA (Dondrup et al., 2003) being applied for microarray data storage as well as expression analysis. So far, each tool can be used either separately or in combination with one of the others. At the same time, the modular design allows to re-use each component for similar tasks in other applications, e.g. by combining them with existing tools such as the GenDB genome annotation system (Meyer et al., 2003) or the ProDB software (Wilke et al., 2003) designed to evaluate proteomics data. The architecture of these systems allows to directly link individual data sets on the programmer's level via the BRIDGE layer (Goesmann et al., 2003, 2005). That way, we can link results from microarray analyses to the corresponding TC sequences, allowing users to navigate from EMMA to SAMS and *vice versa*. Since this data connection is realized on the level of individual objects, an integration of data creation, data access, and data manipulation is achieved.

In order to further elaborate the bioinformatics analysis tools in the area of mycorrhizal symbioses, we are currently working on several extensions of the web-based user frontends. While continuously improving the functionality of each component by implementing novel features, our primary goal is to develop advanced data mining and visualization techniques for heterogenous data sets (Goesmann et al., 2005). Based on the BRIDGE integration we are e.g. working on the analysis of global transcriptional regulation by visualizing expression profiles for general functional categories of genes derived e.g. from KEGG metabolic pathways or the Clusters of Orthologous Groups (COG) of proteins. Moreover, we will correlate gene expression data from microarray experiments with *in silico* expression profiles provided by SAMS and SteN in order to verify hypotheses about AM and ECM symbioses and propose candidate genes for phenotypical studies in plant mutants (Fig. 4). As a long-term goal, we propose

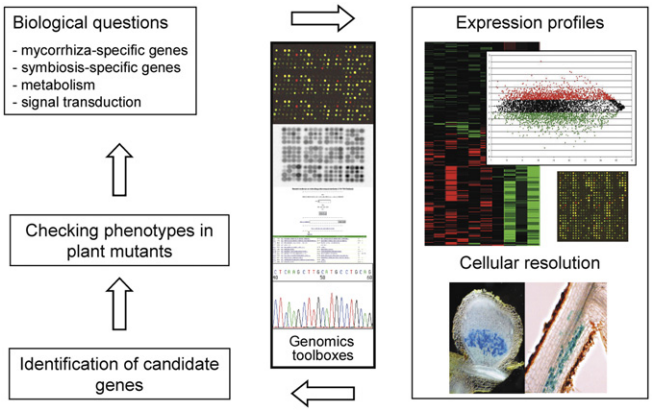


Fig. 4. Discovery science: Application of untargeted expression profiling strategies to identify candidate genes for functional analyses. Genomics toolboxes represent transcriptome profiling resources, such as EST-databases, *in silico* expression profiles, and microarray-based expression data. Mining these resources to solve questions related to mycorrhizal interactions, together with targeted analyses e.g. localizing gene expression to specific cells, allows to identify candidate genes with possible relevance for the conditions investigated. These can be checked phenotypically in mutants of the genes concerned, usually generating additional questions that might require to query genomics resources again.

to establish a bioinformatics platform for mycorrhizal research that also integrates proteome and metabolome data as well as tools for comparative genomics, once whole-genome sequences for the relevant mycorrhizal symbionts are available.

9. Concluding remarks

The upcoming genome sequences of AM and ECM macro- and microsymbionts hallmark a new era in molecular research on mycorrhizae. Our efforts in EST-sequencing can immediately be related to genomic sequences, not only delivering valuable data sets for genome annotation but most importantly providing a snapshot of the transcriptionally active part of the genomes during mycorrhizal interactions. These transcriptional snapshots provide the basis for functional analyses in plant mutants (Tadege et al., 2005) or by applying RNAi (Isayenkov et al., 2005; Ivashuta et al., 2005) approaches.

With respect to transcriptome profiling, the generation of expression profiles from a sufficient amount of plant material became a routine technology, providing transcriptional snapshots of different mycorrhizal interactions. Since expression data obtained from pooled tissue samples cannot be related to specific symbiotic structures, e.g. the arbuscules or the areas of hyphal penetration, transcriptome profiling experiments in mycorrhizal interactions up to now suffered from a significant loss of information. Recently, robust single cell technologies in connection with efficient target amplification protocols were developed also for plant species (Kehr, 2003; Kerk et al., 2003). It can be expected that their future application to study mycorrhizal symbioses will open completely novel perspectives in our molecular understanding of the cellular basis of these beneficial plant-microbe interactions.

Acknowledgements

The transcriptomics work was supported by the Deutsche Forschungsgemeinschaft in the SPP 1084 "Mykorrhiza" projects Pu-28/22-1, KU1478/1-2, and Pu-28/25-3 as well as by the European Union projects MEDICAGO (QLG-CT2000-00676) and GRAIN LEGUMES (FOOD-CT-2004-506223). Helge Küster, Michael Dondrup and Natalija Hohnjec received financial support from the International Graduate School in Bioinformatics and Genome Research. We are thankful to Vivienne Gianinazzi-Pearson and Diederik van Tuinen (both INRA, Dijon, France) as well as to Jerôme Gouzy and Pascal Gamas (all CNRS-INRA, Toulouse, France) for long-lasting, fruitful collaborations in the field of *M. truncatula* transcriptomics. Greg May (Samuel Roberts Noble Foundation, Ardmore, Oklahoma, USA) and Chris Town (TIGR, Rockville, Maryland, USA) are acknowledged for the cooperation during the setup of a *M. truncatula* 16 k 70mer oligonucleotide set. We are grateful to Manuela Meyer, Eva Schulte-Berndt, and Tanja Thias (Institute for Genome Research, Center for Biotechnology, Universität Bielefeld) for their excellent technical support throughout the project.

References

Alba, R., Zhangjun, F., Payton, P., Liu, Y., Moore, S.L., Debbie, P., Cohn, J., D'Ascenzo, M., Gordon, J.S., Rose, J.K., Martin, G., Tanksley, S.D., Bouzayen, M., Jahn, M.M., Giovannoni, J., 2004. ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. Plant J. 39, 697–714.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25, 3389–3402.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D.R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J.A., Zdobnov, E.M., 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucl. Acids Res. 29, 37–40.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat. Genet. 25, 25–29.

Barker, D.G., Bianchi, S., Blondon, F., Dattée, Y., Duc, G., Essad, S., Flament, P., Gallusci, P., Génier, G., Guy, P., Muel, X., Tourneur, J., Dénarié, J., Huguet, T., 1990. *Medicago truncatula*, a model plant for studying the molecular genetics of the *Rhizobium*-legume symbiosis. Plant Mol. Biol. Rep. 8, 40–49.

Becker, A., 2004. Design of microarrays for genome-wide expression profiling. In: Akkermans, A., de Bruijn, F.J., Kowaltchuk, G., van Elsas, J. (Eds.), Molecular Microbial Ecology Manual, second ed.

Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 1543–1562.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M., 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet. 29, 365–371.

Brechenmacher, L., Weidmann, S., van Tuinen, D., Chatagnier, O., Gianinazzi, S., Franken, P., Gianinazzi-Pearson, V., 2004. Expression profiling of up-regulated plant and fungal genes in early and late stages of *Medicago truncatula–Glomus mosseae* interactions. Mycorrhiza 14, 253–262.

Breuninger, M., Requena, N., 2004. Recognition events in AM symbiosis: analysis of fungal gene expression at the early appressorium stage. Fungal Genet. Biol. 41, 794–804.

Brewin, N.J., 1991. Development of the legume root nodule. Ann. Rev. Cell Biol. 7, 191–226.

Cleveland, W.S., Grosse, E., Shyu, W.M., 1992. Local regression models. In: Chambers, J.M., Hastie, T. (Eds.), Statistical Models. S. Wadsworth and Brooks/Cole, Pacific Grove, California, USA, pp. 309–376.

Colebatch, G., Kloska, S., Trevaskis, B., Freund, S., Altmann, T., Udvardi, M.K., 2002. Novel aspects of symbiotic nitrogen fixation uncovered by transcript profiling with cDNA arrays. Mol. Plant-Microbe Interact. 15, 411–420.

Colebatch, G., Desbrosses, G., Ott, T., Krusell, L., Montanari, O., Kloska, S., Kopka, J., Udvardi, M.K., 2004. Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*. Plant J. 39, 487–512.

Dear, S., Staden, R., 1992. A standard file format for data from DNA sequencing instruments. DNA Sequence 3, 107–110.

Diatchenko, L., Lau, Y.F., Campbell, A.P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E.D., Siebert, P.D., 1996. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc. Natl. Acad. Sci. USA 93, 6025–6030.

Dondrup, M., Goesmann, A., Bartels, D., Kalinowski, J., Krause, L., Linke, B., Rupp, O., Szyrba, A., Pühler, A., Meyer, F., 2003. EMMA: a platform for consistent storage and efficient analysis of microarray data. J. Biotechnol. 106, 135–146.

Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P., 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat. Sinica 12, 111–139.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763.

El Yahyaoui, F., Küster, H., Ben Amor, B., Hohnjec, N., Pühler, A., Becker, A., Gouzy, J., Vernié, T., Gough, C., Niebel, A., Godiard, L., Gamas, P., 2004. Expression profiling in *Medicago truncatula* identifies more than 750 genes differentially expressed during nodulation, including many potential regulators of the symbiotic program. Plant Physiol. 136, 3159–3176.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998a. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8, 175–185.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998b. Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res. 8, 186–194.

Firnhaber, C., Pühler, A., Küster, H., 2005. EST sequencing and time course microarray hybridizations identify more than 700 *Medicago truncatula* genes with developmental expression regulation in flowers and pods. Planta 222, 269–283.

Franken, P., Requena, N., 2001. Analysis of gene expression in arbuscular mycorrhiza: new approaches and challenges. New Phytol. 150, 431–439.

Frenzel, A., Manthey, K., Perlick, A.M., Meyer, F., Pühler, A., Krajinski, F., Küster, H., 2005. Combined transcriptome profiling reveals a novel family of arbuscular mycorrhizal-specific *Medicago truncatula* lectin genes. Mol. Plant-Microbe Interact. 18, 771–782.

Gianinazzi-Pearson, V., Brechenmacher, L., 2004. Functional genomics of arbuscular mycorrhiza: decoding the symbiotic cell programme. Can. J. Botany 82, 1228–1234.

Goesmann, A., Linke, B., Rupp, O., Krause, L., Bartels, D., Dondrup, M., McHardy, A.C., Wilke, A., Puhler, A., Meyer, F., 2003. Building a BRIDGE for the integration of heterogeneous data from functional genomics into a platform for systems biology. J. Biotechnol. 106, 157–167.

Goesmann, A., Linke, B., Bartels, D., Dondrup, M., Krause, L., Neuweger, H., Oehm, S., Paczian, T., Wilke, A., Meyer, F., 2005. BRIGEP-the BRIDGE-based genome-transcriptome-proteome browser. Nucl. Acids Res. 33, W710–W716.

Graham, J.H., Miller, M., 2005. Mycorrhizas: gene to function. Plant Soil 274, 79–100.

Grunze, N., Willmann, M., Nehls, U., 2004. The impact of ectomycorrhiza formation on monosaccharide transporter gene expression in poplar roots. New Phytol. 164, 147–155.

Handberg, K., Stougaard, J., 1992. *Lotus japonicus*, an autogamous, diploid legume species for classical and molecular genetics. Plant J. 2, 487–496.

Harrison, M.J., 2005. Signaling in the arbuscular mycorrhizal symbiosis. Ann. Rev. Microbiol. 59, 19–42.

Hildebrandt, U., Ouziad, F., Marner, F.J., Bothe, H., 2006. The bacterium *Paenibacillus validus* stimulates growth of the arbuscular mycorrhizal fungus *Glomus intraradices* up to the formation of fertile spores. FEMS Microbiol. Lett. 254, 258–267.

Hohnjec, N., Vieweg, M.F., Pühler, A., Becker, A., Küster, H., 2005. Overlaps in the transcriptional profiles of *Medicago truncatula* roots inoculated with two different *Glomus* fungi provide insights into the genetic program activated during arbuscular mycorrhiza. Plant Physiol. 137, 1283–1301.

Hohnjec, N., Henckel, K., Bekel, T., Gouzy, J., Dondrup, M., Goesmann, A., Küster, H., 2006. Transcriptional snapshots provide insights into the molecular basis of arbuscular mycorrhiza in the model legume *Medicago truncatula*. Funct. Plant Biol. 33, 737–748.

Hraber, P.T., Weller, J.W., 2001. On the species of origin: diagnosing the source of symbiotic transcripts. Genome Biol. 2, research0037.1–research0037.14.

Huang, X., Madan, A., 1999. CAP3: a DNA sequence assembly program. Genome Res. 9, 868–877.

Ihaka, R., Gentleman, R., 1996. *R*: a language for data analysis and graphics. J. Comp. Graph. Stat. 5, 299–314.

Isayenkov, S., Mrosk, C., Stenzel, I., Strack, D., Hause, B., 2005. Suppression of allene oxide cyclase in hairy roots of *M. truncatula* reduces jasmonate levels and the degree of mycorrhization with *Glomus intraradices*. Plant Physiol. 139, 1401–1410.

Ivashuta, S., Liu, J., Lohar, D.P., Haridas, S., Bucciarelli, B., VandenBosch, K.A., Vance, C.P., Harrison, M.J., Gantt, J.S., 2005. RNA interference identifies a calcium-dependent protein kinase involved in *Medicago truncatula* root development. Plant Cell 17, 2911–2921.

Journet, E.P., van Tuinen, D., Gouzy, J., Crespeau, H., Carreau, V., Farmer, M.J., Niebel, A., Schiex, T., Jaillon, O., Chatagnier, O., Godiard, L., Micheli, F., Kahn, D., Gianinazzi-Pearson, V., Gamas, P., 2002. Exploring root symbiotic programs in the model legume *Medicago truncatula* using EST analysis. Nucl. Acids Res. 30, 5579–5592.

Kehr, J., 2003. Single cell technology. Curr. Opin. Plant Biol. 6, 617–621.

Kerk, N.M., Ceserani, T., Tausta, S.L., Sussex, I.M., Nelson, T.M., 2003. Laser capture microdissection of cells from plant tissues. Plant Physiol. 132, 27–35.

Kistner, C., Parniske, M., 2002. Evolution of signal transduction intracellular symbiosis. Trends Plant Sci. 7, 511–518.

Kouchi, H., Shimomura, K., Hata, S., Hirota, A., Wu, G.-J., Kumagai, H., Tajima, S., Suganuma, N., Suzuki, A., Aiko, T., Hayashi, M., Yokoyama, T., Ohyama, T., Asamizu, E., Kuwata, C., Shibata, D., Tabata, S., 2004. Large-scale analysis of gene expression profiles

during early stages of root nodule formation in a model legume, *Lotus japonicus*. DNA Res. 11, 263–274.

Kretzer, A.M., Bruns, T.D., 1999. Use of atp6 in fungal phylogenetics: an example from the Boletales. Mol. Phylogenet. Evol. 13, 483–492.

Küster, H., Hohnjec, N., Krajinski, F., El Yahyaoui, F., Manthey, K., Gouzy, J., Dondrup, M., Meyer, F., Kalinowski, J., Brechenmacher, L., van Tuinen, D., Gianinazzi-Pearson, V., Pühler, A., Gamas, P., Becker, A., 2004. Construction and validation of cDNA-based Mt6k-RIT macro- and microarrays to explore root endosymbioses in the model legume *Medicago truncatula*. J. Biotechnol. 108, 95–113.

Lamblin, A.F., Crow, J.A., Johnson, J.E., Silverstein, K.A., Kunau, T.M., Kilian, A., Benz, D., Stromvik, M., Endre, G., VandenBosch, K.A., Cook, D.R., Young, N.D., Retzel, E.F., 2003. MtDB: a database for personalized data mining of the model legume *Medicago truncatula* transcriptome. Nucl. Acids Res. 31, 196–201.

Lammers, P., Tuskan, G.A., DiFazio, S.P., Podila, G.K., Martin, F., 2004. Mycorrhizal symbionts of *Populus* to be sequenced by the United States Department of Energy's Joint Genome Institute. Mycorrhiza 14, 63–64.

Leung, Y.F., Cavalieri, D., 2003. Fundamentals of cDNA microarray data analysis. Trends Genet. 19, 649–659.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., Quackenbush, J., 2000. An optimized protocol for analysis of EST sequences. Nucl. Acids Res. 18, 3657–3665.

Liu, J., Blaylock, L.A., Endre, G., Cho, J., Town, C.D., VandenBosch, K.A., Harrison, M.J., 2003. Transcript profiling coupled with spatial expression analyses reveals genes involved in distinct developmental stages of an arbuscular mycorrhizal symbiosis. Plant Cell 15, 2106–2123.

Lohar, D.P., Sharopova, N., Endre, S., Peñuela, S., Samac, D., Town, C., Silverstein, K.A.T., VandenBosch, K.A., 2005. Transcript analysis of early nodulation events in *Medicago truncatula*. Plant Physiol. 140, 221–234.

Manthey, K., Krajinski, F., Hohnjec, N., Firnhaber, C., Pühler, A., Perlick, A.M., Küster, H., 2004. Transcriptome profiling in root nodules and arbuscular mycorrhiza identifies a collection of novel genes induced during *Medicago truncatula* root endosymbioses. Mol. Plant-Microbe Interact. 17, 1063–1077.

Martin, F., Duplessis, S., Ditengou, F., Lagrange, H., Voiblet, C., Lapeyrie, F., 2001. Developmental cross talking in the ectomycorrhizal symbiosis: signals and communication genes. New Phytol. 151, 145–154.

Martin, F., Tuskan, G.A., DiFazio, S.P., Lammers, P., Newcombe, G., Podila, G.K., 2004. Symbiotic sequencing of the *Populus* mesocosm. New Phytol. 161, 330–335.

Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., Pühler, A., 2003. GenDB: an open source genome annotation system for prokaryote genomes. Nucl. Acids Res. 31, 2187–2195.

Moreau, C., Aksenov, N., Lorenzo, M.G., Segerman, B., Funk, C., Nilsson, P., Jansson, S., Tuominen, H., 2005. A genomic approach to investigate developmental cell death in woody tissues of *Populus* trees. Genome Biol. 6, R34.

Ouziad, F., Hildebrandt, U., Schmelzer, E., Bothe, H., 2005. Differential gene expressions in arbuscular mycorrhizal-colonized tomato grown under heavy metal stress. J. Plant Physiol. 162, 634–649.

Quackenbush, J., 2002. Microarray data normalization and transformation. Nat. Genet. 32, S496–S501.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvici, B., Pertea, G., Sultana, R., White, J., 2000. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucl. Acids Res. 29, 159–164.

Sato, S., Tabata, S., 2006. *Lotus japonicus* as a platform for legume research. Curr. Opin. Plant Biol. 9, 128–132.

Schultze, M., Kondorosi, A., 1998. Regulation of symbiotic root nodule development. Ann. Rev. Genet. 32, 33–57.

Schüssler, A., Schwarzott, D., Walker, C., 2001. A new fungal phylum, the Glomeromycota: phylogeny and evolution. Mycol. Res. 105, 1413–1421.

Slonim, D.K., 2002. From patterns to pathways: gene expression data analysis comes of age. Nat. Genet. 32, S502–S508.

Smith, S.E., Read, D.J., 1997. Mycorrhizal Symbiosis. Academic Press, London.

Smyth, G.K., Yang, Y.-H., Speed, T.P., 2003. Statistical issues in microarray data analysis. In: Brownstein, M.J., Khodursky, A.B. (Eds.), Functional Genomics: Methods and Protocols, Methods in Molecular Biology, vol. 224. Humana Press, Totowa, New Jersey, USA, pp. 111–136.

Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W.L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B.J., Robinson, A., Bassett, D., Stoeckert Jr., C.J., Brazma, A., 2002. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. 3, research0046.1-0046.9.

Stekel, D.J., Git, Y., Falciani, F., 2000. The comparison of gene expression from multiple cDNA libraries. Genome Res. 10, 2055–2061.

Sterck, L., Rombauts, S., Jansson, S., Sterky, F., Rouze, P., Van de Peer, Y., 2005. EST data suggest that poplar is an ancient polyploid. New Phytol. 167, 165–170.

Sterky, F., Bhalerao, R.R., Unneberg, P., Segerman, B., Nilsson, P., Brunner, A.M., Charbonnel-Campaa, L., Lindvall, J.J., Tandre, K., Strauss, S.H., Sundberg, B., Gustafsson, P., Uhlen, M., Bhalerao, R.P., Nilsson, O., Sandberg, G., Karlsson, J., Lundeberg, J., Jansson, S., 2004. A *Populus* EST resource for plant functional genomics. Proc. Natl. Acad. Sci. USA 101, 13951–13956.

Tadege, M., Ratet, P., Mysore, K.S., 2005. Insertional mutagenesis: a Swiss army knife for functional genomics of *Medicago truncatula*. Trends Plant Sci. 10, 229–235.

Thompson, R., Ratet, P., Küster, H., 2005. Identification of gene functions by applying TILLING and insertional mutagenesis strategies on microarray-based expression data. Grain Legumes 41, 20–22.

Town, C.D., 2006. Annotating the genome of *Medicago truncatula*. Curr. Opin. Plant Biol. 9, 122–127.

Udvardi, M.K., Tabata, S., Parniske, M., Stougaard, J., 2005. *Lotus japonicus*: legume research in the fast lane. Trends Plant Sci. 10, 222–228.

Valot, B., Negroni, L., Zivy, M., Gianinazzi, S., Dumas-Gaudot, E., 2006. A mass spectrometric approach to identify arbuscular mycorrhiza-related proteins in root plasma membrane fractions. Proteomics, doi:10.1002/pmic.200500403.

Wilke, A., Rückert, C., Bartels, D., Dondrup, M., Goesmann, M., Hüser, A.T., Kespohl, S., Linke, B., Mahne, M., McHardy, A.C., Pühler, A., Meyer, F., 2003. Bioinformatics support for high-throughput proteomics. J. Biotechnol. 106, 147–156.

Wipf, D., Benjdia, M., Rikirsch, E., Zimmermann, S., Tegeder, M., Frommer, W.B., 2003. An expression cDNA library for suppression cloning in yeast mutants, complementation of a yeast his4 mutant and EST analysis from the symbiotic basidiomycete *Hebeloma cylindrosporum*. Genome 46, 177–181.

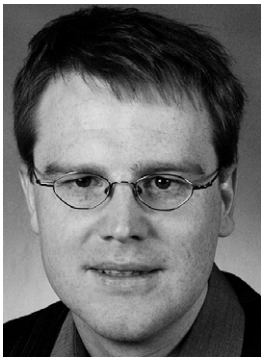
Wulf, A., Manthey, K., Doll, J., Perlick, A.M., Linke, B., Bekel, T., Meyer, F., Franken, P., Küster, H., Krajinski, F., 2003. Transcriptional changes in response to arbuscular mycorrhiza development in the model plant *Medicago truncatula*. Mol. Plant-Microbe Interact. 16, 306–314.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., Speed, T.P., 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucl. Acids Res. 30, e15.

Young, N.D., Cannon, S.B., Sato, S., Kim, D., Cook, D.R., Town, C.D., Roe, B.A., Tabata, S., 2005. Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiol. 137, 1174–1181.

Zhu, T., 2003. Global analysis of gene expression using GeneChip microarrays. Curr. Opin. Plant Biol. 6, 418–425.

Helge Küster heads the Junior Group "Genomics of Legume Plants" at the Institute for Genome Research, Center for Biotechnology, Bielefeld University. He obtained his Ph.D. in Plant Genetics from Bielefeld University in 1995, at that time investigating genes expressed in *Vicia faba* root nodules. After switching his research interests towards the development and application of transcriptomics tools to study arbuscular mycorrhizal interactions, he received his Habilitation at the Faculty of Biology, Bielefeld University in 2004. Currently, H. Küster serves as "Head of the Scientific Committee" of the European Union Integrated Project "Grain Legumes", with a focus on integrating transcriptomics activities in different legume projects. His group primarily deals with the identification and functional analysis of genes differentially regulated during root endosymbioses by combining transcriptome profiling approaches with studies on promoter activities and phenotypical analyses in transgenic plants.



Folker Meyer heads the Bioinformatics Resource Facility at the Center for Biotechnology, Bielefeld University. Originally a computer scientist, he obtained his Ph.D. in Bioinformatics from Bielefeld University in 2002, working on a novel software system for the annotation of bacterial genomes. He later on extended his interests to the analysis of more diverse genomics and post-genomics data, and in the past few years comparative genomics has become his main area of interest. Currently, F. Meyer serves as "Head of the Bioinformatics Committee" of the European Union Network of Excellence "Marine Genomics Europe", which aims at bringing genomics technologies to the marine sciences. His group provides software solutions for various areas of biology, such as the study of microarray expression data, high throughput proteomics, metabolomics and sequence analysis, always interacting closely with partners from biological sciences.

Alfred Pühler originally studied Physics, received his Ph.D. degree in Microbiology and habilitated in Genetics. Since 1979, he is Head of the Chair of Genetics at Bielefeld University. The chair has a strong focus on biotechnology, in particular on the amino acid producer *Corynebacterium glutamicum* and on the symbiotically nitrogen-fixing soil bacterium *Sinorhizobium meliloti*. A. Pühler is the coordinator of a German Research Network dealing with Bacterial Genome Research. He also coordinates the Focus Programme SPP 1084 of the German Research Foundation entitled "Molecular Basics of Mycorrhizal Symbioses". A. Pühler is member in three Academies of Sciences. During the last six years, he served as a member of the Science Council installed by the Federal President of Germany.