Tools for managing and analyzing microarray data

André Koschmieder, Karin Zimmermann, Silke Trißl, Thomas Stoltmann and Ulf Leser

Submitted: 17th December 2010; Received (in revised form): 11th February 2011

Abstract
The microarray-based analysis of gene expression has become a workhorse for biomedical research. Managing the amount and diversity of data that such experiments produce is a task that must be supported by appropriate software tools, which led to the creation of literally hundreds of systems. In consequence, choosing the right tool for a given project is difficult even for the expert. We report on the results of a survey encompassing 78 of such tools, of which 22 were inspected in detail and seven were tested hands-on. We report on our experiences with a focus on completeness of functionality, ease-of-use, and necessary effort for installation and maintenance. Thereby, our survey provides a valuable guideline for any project considering the use of a microarray data management system. It reveals which tasks are covered by mature tools and also shows that important requirements, especially in the area of integrated analysis of different experimental data, are not yet met satisfyingly by existing systems.

Keywords: microarray; statistical data analysis; bioinformatics; affymetrix; differential gene expression; data management; databases

INTRODUCTION
Microarrays in their various instances have become a fundamental experimental technique for biomedical research. Their most prevalent form are gene expression arrays, which measure the expression levels of thousands of genes in a given sample on a single array [1–3]. These so-called gene chips have become a workhorse in biomedical research and are used for such diverse questions as revealing regulatory relationships between genes [4–6], definition and identification of diseases and disease subtypes [7], identification of functional gene modules [6, 8, 9] or alternative splicing [10–12]. A simple PubMed search with the keywords ‘gene and expression and (array or chip)’ reveals that almost 40 000 papers have been devoted to this topic and the international databases for gene expression data sets, ArrayExpress [13], GEO [14] and CIBEX [15], already host the data of more than 400 000 hybridizations.

Expression arrays are just one form of microarrays, and many related technologies have been developed over the last years (Supplementary Data, Section S1.2). They all have in common that short, single-stranded pieces of DNA are fixed to a solid carrier, the chip. On a gene array, each spot on the chip is derived from the sequence of a specific gene [16], whereas on an exon array, each spot represents an exon of a gene [17]. The most complete coverage over a genome provides a tiling array [18], where each spot stands for a stretch of the genome that

Corresponding author. André Koschmieder, Department of Computer Science, Humboldt-Universität zu Berlin, Rudower Chaussee 25, 12489 Berlin, Germany. Tel: +49-30-2093-5486; Fax: +49-30-2093-3045; E-mail: koschmie@informatik.hu-berlin.de

André Koschmieder has finished his Master in Computer Science in 2009 and is now working as a PhD student in biomedical databases in the group of Prof. Leser, Humboldt-Universität zu Berlin, Germany.

Karin Zimmermann has finished her Master in Bioinformatics in 2008 and is now working as a PhD student in high-throughput data analysis in the group of Prof. Leser, Humboldt-Universität zu Berlin, Germany.

Silke Trißl holds a Diploma in Biotechnology and a MSc in Bioinformatics. She is about to finish her PhD at Humboldt-Universität and is currently working at Leibniz Institute for Farm Animal Biology, Dummerstorf, Germany.

Thomas Stoltmann is currently studying Computer Science and is working as a research assistant in the group of Prof. Leser, Humboldt-Universität zu Berlin, Germany.

Ulf Leser holds the Chair of Knowledge Management in Bioinformatics at the department for Computer Science of Humboldt-Universität zu Berlin, Germany.
might be coding or non-coding. While exon arrays and tiling arrays are becoming increasingly important, the bulk of microarray data is still produced using gene arrays.

A single chip from the probably most popular platform, an Affymetrix Gene Chip Human Genome U133 Plus 2.0 [19], produces expression values for roughly 1.3 million spots, which usually are aggregated into values for \(~39,000\) genes. The workflow to produce those gene expression values is complex. It includes image analysis, spot detection, background correction, normalization and various quality control techniques [20]. It is not rare that projects use dozens to hundreds of such arrays, often in duplicate or even triplicate form to enhance the robustness of the results [21]. Before analyzing the results in more depth, values for multiple copies of the same probe must be aggregated appropriately across replicates [22]. Also, many forms of data analysis require the availability of other types of data, such as pathways, transcription factor binding sites or chromosomal locations [23]. Such data must be imported from other databases, linked to the microarray results and updated continuously. Only after all these technical steps have been performed, questions concerned with the real goal of a study can be addressed, such as the identification of differentially expressed gene sets [24] or the revelation of regulatory networks [4].

The complete microarray analysis process hence is complex, encompasses multiple processing steps, depends on various programs and databases and deals with large quantities of heterogeneous data; all of which calls for proper support by software tools. Particularly, when studies involve different laboratories, run over a longer period of time, or are carried out by a changing crew of people, a proper data management is of high importance to ensure optimal data quality and to reduce superfluous or redundant work. Data management must at least address storage, association of data sets to experiments and description of experimental data with proper metadata. Furthermore, the data management system should be combined with tools for data analysis that allow for a seamless application of the methods of choice on the data sets of choice.

These requirements, together with the enormous popularity of the microarray technique, have lead to the development of a flood of tools that, in one way or the other, support the handling of microarray data. In this paper, we review the capabilities of 78 publically available and free such tools. Note that commercial tools, such as Gene Expressionist (http://www.genedata.com/products/expressionist/), Partek (http://www.partek.com/), GeneSpring GX (http://www.chem.agilent.com/en-US/Products/software/lifesciencesinformaticsgenespringgx), or Geospiza (http://www.geospiza.com/) are not part of this survey.

We first performed a literature-based review of 78 tools using a catalog of 55 criteria classified into ten categories. We then focused on a particular scenario which we consider as typical for a large fraction of microarray applications: a publicly funded, collaborative study encompassing different laboratories with the goal to gain new insights into the mechanism of a particular disease. Such a project is not primarily interested in developing new techniques for microarray analysis, but rather in applying common methods and techniques on their newly obtained data. It mostly consists of biologists or medical doctors and therefore requires software that is simple to install, maintain and use. From this scenario, we derived a set of four criteria that we used to filter the list of 78 tools to a short list of 22 tools. These criteria were: (i) since projects usually do not want to give their data out-of-hand before publication, we required that a tool must be available for local installation. (ii) The tool must support Affymetrix Gene Chips, as these are the most frequently used platform. (iii) The tool must be able to perform the most common analysis tasks, especially pre-processing and normalization, differential expression, clustering and some form of functional analysis. (iv) The tool must offer a user-friendly interface to the management and analysis of the data it handles.

We closely inspected all 22 tools passing this filter and describe their functionality in detail in this review. We particularly assessed each tool with respect to its level of support for data analysis, pre-processing, data privacy, definition of customized analysis pipelines, experimental data types, integration with non-array data and compliance to international standards. We decided to disregard tools from our list that depend on commercial components (databases, operating systems, etc.), as publicly funded projects often depend on an open-source infrastructure. This reduced the list down to 14 tools. In two cases, two of the remaining tools are supposed to work together closely, so we viewed these tools as one combined tool. This reduced the number of tested tools to 12.
As the third and final step, we defined a series of tasks that are typical in microarray projects to test the tools hands-on. Of the 12 tools on our short-list, we were able to obtain source code or installation binaries for 10 and managed to install seven: BioArray Software Environment (BASE) [25], Chipster [26], EzArray [27], Gene Expression Open Source System (Geoss)/GeneX [28], Microarray Analysis and Retrieval System (MARS) [29] and comprehensive R based microarray web (CARMAWeb) [30] (combined), Mayday [31] and TM4 [32]. For each, we loaded the same real-life data sets, performed the same type of preprocessing, and carried out differential analysis, hierarchical clustering and functional analysis. We judged each tool by properties such as completeness of functionality, documentation and perceived level of comfort in handling. To minimize the influence of subjective preferences, we had each tool tested by two bioinformaticians and report their consolidated opinions.

Despite their high importance, microarray tools have rarely been surveyed before. General problems of managing and sharing large amounts of genomic high-throughput data sets have recently been surveyed by Stein [33] and Schadt et al. [34], but both works do not cover specific systems. Gardiner-Garden and Littlejohn [35] describe 13 tools for storage of microarray data using a small set of criteria and purely based on literature review. Sherlock and Ball [36] discuss six tools, but only consider their abilities in terms of storage, disregarding data analysis. Do et al. [37] compare four tools with a focus on the underlying computational infrastructure. None of these reviews described hands-on experiences. Furthermore, the information in these papers (published between 2001 and 2005) are already outdated to a large degree, as many tools continuously improve their functionality.

**MATERIALS AND METHODS**

**Scope and method of evaluation**

We survey royalty-free tools for the management and analysis of data produced by microarray experiments. We focus on gene expression arrays, but partly also consider other types of microarray experiments, such as ChIP-Chip or tissue microarrays (TMAs). Tools that implement only specific functions and do not support most of the standard workflow for microarray data are disregarded. Also, we do not consider tools whose usage requires programming skills.

We used extensive PubMed searches to obtain a comprehensive list of tools fitting to this scope. Starting from this initial list, we checked systems referenced in the publications, papers citing our matching papers, and followed PubMed recommendations for related articles. Furthermore, we used Internet searches to find tools which were published in venues not indexed in Medline. Overall, we looked into several hundred papers, 150 of which turned out to be relevant. These 150 papers describe 78 different systems. Reasons for this discrepancy are, for instance, tools which have different versions that were published separately or papers merely reporting on the usage of a tool and not on its functionality.

Each of the 150 articles was read by at least two bioinformaticians (reviewers in the following). Each reviewer filled a spreadsheet with 55 predefined evaluation criteria (see ‘Evaluation criteria’ section). Whenever available, they also consulted the documentation of the systems to get more current information than the published one. The complete spreadsheet of all tools and all reviewers can be found in the Supplementary Data, Section S7.

On top of this paper-based evaluation, we performed a hands-on test of selected tools. Since it appeared infeasible to install 78 tools, we defined four criteria as mandatory for a tool to be tested life. These criteria (see ‘Selection of tools for local testing’ section) were derived from a scenario we consider typical for many settings in which sizable quantities of microarray data are produced and analyzed. We tried to install all 12 tools fulfilling these criteria, which was successful for seven systems. Each of these systems was subjected to a series of tasks that we consider as fundamental for microarray analysis (Supplementary Data, Section S2). While performing these tasks, we gathered assessments on functionality, performance, usability and ease-of-handling for each tool. Finally, we aggregated the results of the hands-on and the paper-based evaluation to derive suggestions on which tools appear most recommendable for which kind of applications.

**Evaluation criteria**

For the evaluation, we used a catalog of 55 criteria divided into ten categories. These criteria are based on our own experiences, augmented with properties of systems as described in the papers. Table 1 lists the different categories and the number of criteria they include. A complete list of descriptions for all
evaluation criteria can be found in the Supplementary Table S3.

Selection of tools for local testing
We used four criteria to choose tools for local installation and hands-on testing. These are: (i) since projects usually do not want to give their data out-of-hand before publication, a tool must be available for local installation, either as source code or binaries. (ii) The tool must (at least) support analysis of Affymetrix Gene Chips, as these are the most frequently used techniques to-date. [GEO (Gene Expression Omnibus) currently contains about 400 000 samples measuring RNA expression, about 200 000 of these samples are hybridized on Affymetrix arrays] (iii) The tool must support at least the most common analysis tasks, i.e. quality control, normalization, differential analysis, clustering and functional analysis. (iv) The tool must ship with a user-friendly interface that offers full access to all functionality of the system, i.e. users do not need to use any other interface for their daily work. This interface can be web-based or have the form of a graphical stand-alone application. From the 78 tools, 22 fulfilled these four criteria. However, some of these tools, although freely available, actually require the installation of commercial components, mostly of a commercial database system. We removed those tools from further consideration. We also view systems that are supposed to work together as one system. Thus, the final number of tools we tried to install was 12 (see ‘Overview of selected tools’ section).

We successfully installed seven systems. The process of installation we followed is described in Supplementary Data, Section S3. The reasons for failed installations were manifold, such as nonavailability of code (although stated otherwise in the publication), dependence on heavily outdated libraries or unresolvable installation errors. Detailed information on tools that could not be installed can be found in Supplementary Data, Section S4.

RESULTS
Over 70 different tools for management and analysis of microarray data have been developed by many groups around the world. We performed a literature-based analysis of all tools we could find references for and evaluated each tool with respect to a list of 55 predefined criteria. The complete list of all considered systems together with their evaluation can be found in the Supplementary Table S2.

Overall, we found the individual capabilities of the tools differing to a vast degree. Many tools...
have a strong focus on some specific functionality and support other tasks only marginally. Of the 78 systems, 29 focus strictly on data analysis, while 17 only address data storage (Figure 1). While most of the tools focus on one experimental technology only, i.e. two-color, Affymetrix GeneChips or tissue microarrays (TMA), 26 systems can handle at least the two most common formats; Affymetrix GeneChips and two-color arrays (Figure 2). Only three tools can handle all types of classical microarrays, namely ArrayExpress and NCBI GEO for storage and Bioconductor for analysis. Tools for managing TMA are strictly separated from the others, which means that we could not find a single tool that was able to handle TMA and any other kind of microarray data together.

**Overview of selected tools**

From the 78 tools subjected to the literature-based evaluation, we selected 22 for closer inspection (Table 2). Of these, in four cases, two tools are supposed to work together closely, so we considered them as one system. Seven of these systems could be installed successfully (see ‘Systems one-by-one’ section). In this section, we describe each of these systems in detail. Information on the tools we could not install are given in Supplementary Data, Section S4.

Supplementary Table S1 summarizes the system requirements of the successfully installed tools. The list encompasses web-based systems as well as stand-alone applications. The source code is publicly available for all of them. Table 3 shows the capabilities of the tools with respect to data management and experimental techniques covered. Data management includes the capability of storing experiment data in a database and to control access for certain users or groups of users (system-specific details are given below). Such functionality is provided by BASE, MARS, EzArray and GeneX. TM4 only supports storage of two-color data. CARMAWeb provides no storage but uses a database to store the current session and intermediate results. All tools support Affymetrix chips; some exclusively those, while others cover a wide range of different techniques. Support for standards is generally weak; only one tool supports importing MAGE-TAB and few are able to export into a standard format. None of the tools can directly import meta-information for experiments from GEO. Note that all tools are capable of importing CEL Files.

**Systems one-by-one**

After installing the systems, we performed a series of typical tasks for microarray data sets, such as upload, normalization, quality control, clustering, etc. (Supplementary Data, Section S2). The analysis results of the different tools are difficult to compare as not all tools allow for a fine-granular selection of the parameters and methods applied. Systems differ, especially in the particular tests applied (plain *t*-test versus Welch’s *t*-test) and in the linkage criteria for the hierarchical clustering (single-link versus complete-link). Overall, we found the results of all analyses to be about equal. Here, we report on ease-of-use, completeness of offered functionality and specific strengths and weaknesses of all seven tools. One screenshot for each tool can be found in the Supplementary Data.
**BASE**

The BASE [25] is a MIAME compliant web-based Laboratory Information Management System (LIMS) for microarray data. BASE focuses on data management. Many types of microarray technologies, single- or two-channel data, can be stored, as plug-ins allow to load array designs and scanned data. Normalization and certain forms of quality assessment can directly be called, while further analysis methods are provided by MeV, a part of TM4 (see ‘TM4’ section). As BASE supports a plug-in to export data to MeV, data transfer is seamless. An online documentation is available, but unfinished.

**Storage.** BASE uses a PostgreSQL or MySQL database for storage. A sophisticated user management and access control system featuring users and groups is provided. Data can be exported in the MAGE-TAB format.

**Analysis.** Affymetrix data can be normalized using an RMA plug-in, which does also allow for quality...
control plots. Two-color data can also be analyzed, but the preprocessing methods are limited. MeV provides advanced data analysis (see ‘TM4’ section).

**Usability.** Managing data in BASE is performed using various web forms. The interface is comprehensive but can be overwhelming for novice users. Detailed metadata can be stored with each experiment, but adding such data involves several input masks, an often repetitive and time-consuming task.

**Strengths/weaknesses.** BASE provides a sophisticated user and data management while it has limited...
analysis capabilities. To analyze data, an external tool (MeV) has to be used. Notably, it is the only tool that can read and write the popular MAGE-TAB format.

**Chipster**

Chipster [26] is a Java Web Start application supporting various technologies including Affymetrix (GeneChips, ExonChips and SNP), Illumina and Agilent. Chipster provides the most complete set of functionalities among the eight installed tools in all classes of data analysis we looked at. Online documentation is available as a user manual, tutorials and technical references.

**Storage.** Chipster does not support storing data in a database but can only read data from the local computer. No user management is provided; restricted access rights have to be enforced by the operating system. Though Chipster cannot store data itself, the import from GEO in the SOFT format is supported.

**Analysis.** Chipster offers all standard methods like different techniques for normalization, preprocessing, statistics, clustering and visualization. Additionally, it features many more sophisticated methods like KEGG/GO enrichment or promoter analysis.

**Usability.** Chipster has a well-arranged user interface. Different analysis tools are sorted into color-coded categories greatly helping orientation. All steps of an analysis session can be visualized as a graph, where data sets and analysis methods are nodes connected by edges showing the flow of data.

**Strengths/weaknesses.** Chipster is based on a client-server architecture. Data is imported and visualized at the client side (i.e. on the local machine), while all processing is performed on the server side using R. An advantage is that the local computer does not need much computational power, but it also implies that all data needs to be transferred between client and server for each analysis step. This can be very time-consuming if the bandwidth is limited or data sets are very large (as, for instance, in the case of exon arrays). The outstanding strengths of Chipster are its intuitive user interface and the completeness of the offered analysis methods as well as the possibility of data import from public databases like GEO and ArrayExpress.

**EzArray**

EzArray [27] is a web-based tool that supports only Affymetrix data. MIAME compliance is not ensured. A key feature is its ability to read data sets directly from NCBI GEO. Analysis, including normalization, is conducted using R/Bioconductor, for which EzArray provides the scripts as well as the output of the analyses. An online help is available.

**Storage.** EzArray uses a MySQL database to store the experiments. However, there is no access control, which means that all users can see the data of all other users.

**Analysis.** EzArray has a broad range of methods for data normalization and quality assessment, but lacks clustering or functional analysis. The tool also allows to add various annotations (i.e. different gene identifiers) using Bioconductor packages for Affymetrix.

**Usability.** EzArray intuitively structures its analysis methods in three groups: preprocessing, statistical testing and annotation. This enables an almost self-explanatory usage. Analysis methods can be customized using various parameters, which may either be filled in directly (for the expert) or left unchanged (for the less experienced user).

**Strengths/weaknesses.** The clear strength of EzArray is its ready-to-use predefined workflow (for Affymetrix data). If no parameter changes are desired, an entire analysis can be performed with one click. Weaknesses of EzArray are the lack of support for access restrictions and its lack of several types of analysis methods.

**Geoss**

The Geoss (formerly called GeneX) [28] is a web-based, long-existing storage and analysis tool for Affymetrix GeneChip data. A tutorial is available, but we found it hard to use because it lacks a clear structure.

**Storage.** Geoss uses a PostgreSQL database to store the experiments. It provides user and group management to control access.

**Analysis.** Geoss supports customizable analysis trees. However, in our installation, attempts to perform any analysis task failed as we were not able to select data.
**Usability.** The user interface is kept very simple, but lacks uniformity between different pages. The tool does not provide a bulk upload feature, i.e. uploading all experiments at once, thus adding experiments can be time-consuming.

**Strengths/weaknesses.** The unintuitive usage of the tool is a clear downside. Although user and group management is provided in principle, we observed that all users seem to have administration rights by default, thus breaking an access policy. In addition to its very limited set of analysis features, our practical experiences with the tool raise concerns about its maturity.

**MARS**
The MARS is a web-based system to store and normalize Affymetrix and two-color experiments along with MIAME compliant metadata. A documentation containing many useful screenshots is available though unfinished.

**Storage.** MARS uses a PostgreSQL or Oracle database. The system supports a sophisticated user management system and can export data in MAGE-ML format.

**Analysis.** MARS does not have analysis capabilities, but integrates with CARMAWeb. Normalized data can also be exported to other analysis tools like the Pathway Explorer [45].

**Usability.** Users can import and manage their data using a set of well-arranged web forms. MIAME-compatibility is enforced by defining required fields such as the name of a biological sample, sample type and sample origin. This is important when planning to submit the data to a public repository.

**Strengths/weaknesses.** Uploading the HG-U133-Plus2.0 array design from Affymetrix failed for lack of main memory (we assigned 2 GB), indicating that the system has not yet been adapted to modern high-density arrays. Smaller design files are loaded without problem. A weakness of MARS is that it requires to install and maintain a second system for analyzing data. A strength is the advanced user management and the support for bulk upload of data files.

**CARMAWeb**
The CARMAWeb frontend [30] is a web-based tool dedicated to data analysis. It can be used together with MARS or as a stand-alone application. Affymetrix and two-color data are supported. A comprehensive documentation is available.

**Storage.** CARMAWeb uses a PostgreSQL database to store uploaded experiments, but no user management is supported in this configuration. Data that is accessed through MARS will only be visible to people with proper access rights on the MARS server.

**Analysis.** CARMAWeb normalization and analysis methods are incorporated from R/Bioconductor. The choice of methods for differential testing is limited. CARMAWeb lacks clustering algorithms, but links to another system (GenesisWeb [46] for this purpose. Functional analysis using GO terms requires providing an Entrez-to-GO mapping.

**Usability.** CARMAWeb comes with an intuitive, wizard-like frontend. The input format of the two-color data was automatically detected. At the end of an analysis, users can download a PDF document that contains information on the data sets and R commands used to create the result, as well as the result of the analysis itself.

**Strengths/weaknesses.** The strength of CARMAWeb is the guided, straightforward usage and its easy data handling. Analysis capabilities are limited and clustering is only accessible by an external tool which requires its independent installation, which is a downside for inexperienced users.

**Mayday**
Mayday [31], a workbench for Microarray Data Analysis and visualization, is a Java Web Start application focusing on data analysis. Affymetrix cel-files can be normalized by the system, but all other types of data need to be normalized externally before import. A comprehensive documentation is available including an introduction to all core features and how-to guides for many tasks.

**Storage.** Mayday has no dedicated storage, but data must be read from the file system. A nice feature is its ability to store and resume analysis sessions.

**Analysis.** Mayday features a wide range of analysis methods, ranging from various statistical methods over many possibilities to filter probes—either based on statistical tests or based on textual
values—to different cluster algorithms. The tool also offers a rich selection of visualization methods including a Pathway Viewer. Plots can be saved in different formats (such as SVG and JPG). A powerful feature is the fact that all plots displayed during an analysis process work together interactively, i.e. a gene highlighted in one plot will also be highlighted in all corresponding plots. Mayday furthermore, offers a number of features not listed in Table 4, especially for differential analysis (SVM, Max Minority, Sum of Variances, Information Gain, Quartet Mining, Pearson Correlation, Twoing Rule, Gini Index, Relational Ranking, etc.).

**Usability.** The functionally-rich interface, that also allows to set various parameters for every method, is clearly geared towards the expert and can be demanding for less experienced or occasional users.

**Strengths/weaknesses.** Analyses in Mayday run comparably fast, as all available analysis methods have been re-implemented in Java, which saves the time-consuming calls to R performed by most other tools. A further advantage is the fact that Mayday is under active development, resulting in quick response to user requests (at least to ours). Unlike other tools, Mayday does not require a registration when using the web-start version. On the downside, Mayday does not offer any data management and its support for standards is weak.

**TM4**

TM4 [32] actually is a suite of different Java-based stand-alone tools and not a single, closed system. ‘Spotfinder’ analyzes scanned images and produces numerical values. ‘Madam’ provides an interface to store two-color microarray data in a relational database. ‘ExpressionConverter’ converts Agilent and GenePix formatted data to the MEV format used by the other components. ‘MIDAS’ normalizes two-color data. Finally, ‘MeV’ is a tool to analyze two-color or Affymetrix data. Note that Affymetrix data cannot be normalized. We focus our evaluation on ‘Madam’ and ‘MeV’. For both, an extensive documentation is available.

**Storage.** ‘Madam’ uses a MySQL database to store two-color data. User management is provided by the database management system. It enforces annotation of each sample and each array to yield MIAME compliance and exports data in MAGE-ML.

**Analysis.** ‘MeV’ only allows to analyze already normalized data, but can be used together with BASE which offers these features, though only in a limited form (see above). Support for quality assessment is virtually nonexisting. Its capabilities range from statistical analyses using various tests to multiple cluster analyses. The results of all analysis steps are displayed in a well-structured way.

**Usability.** The different tools of TM4, though individually convenient in use, lack a homogeneous, seamless user interface. Data from one tool has to be stored on disk before it can be used by another application of TM4. Madam comes with a simple and intuitive user interface for managing two-color microarray experiments. MeV has a well-structured interface supporting various analysis methods.

**Strengths/weaknesses.** The loosely-coupled nature of the TM4 tools makes working with them rather cumbersome. A big asset of TM4 is its exceptionally wide range of methods for all analysis steps after data normalization. ‘MeV’ only allows to analyze already normalized data, which makes it necessary to preprocess Affymetrix data with a different tool (such as BASE).

**DISCUSSION**

While searching for available tools, installing systems and using them to analyze data, we gathered a good overview of the current status of microarray data analysis tools. In this section, we forward our impressions and give some advice on how to find the most fitting tool for one’s own microarray project.

**Number and availability of systems**

Overall, we found more than 75 published tools for managing and analyzing microarray data. Given that there is an additional number of commercial tools, tools for which no paper has been published yet and tools for which we may not have found a publication, one can safely state there is a great need for such tools. On the other hand, the great number of tools raises the suspicion that work has been duplicated. Apparently, very often people decided to build a new system instead of reusing or adapting an existing one. There are two possible explanations for this (on top of the well-known ‘not invented here’ syndrome): First, developers might not be aware of the existence of proper tools. Second, requirements in specific projects might be so specific that reuse is
impossible, or that adaptation would be more costly than re-implementation. Probably, both explanations are true to a certain degree. The general lack of comparative works on microarray databases certainly has led to a rather low level of awareness of such systems. Also, many tools do have specific, often exclusive, features which might have been the reason for their development in the first place.

We were also surprised how many of the systems are still available online. Almost all tools covered by this review are presented on a functioning website, even though these might not have been updated for years. Additionally, for many tools the source code is available (Supplementary Table S2).

Installation

Installing the selected tools locally was often a difficult task. Although these impressions are slightly subjective as they depend on the experience of the person installing the systems, we think that describing the encountered difficulties is helpful for judging the complexity of the respective installation processes. All installations were performed by an experienced computer scientist.

Although installation manuals are available for all tools, we often found them incomplete, missing important steps or not listing all required libraries, especially the latter was a recurring problem. Often, installations required highly specific, yet, outdated versions of certain libraries, especially for R and Bioconductor (for instance, GeneX/Geoss). The reason thereof is that R and Bioconductor are in constant development, while many of the reviewed systems are not developed further any more and therefore are not adjusted to newer versions of the libraries they depend upon. Thus, before trying to install a system, one should make sure that all libraries in the required versions are within reach.

Generally speaking, installation of those systems that support storage and analysis of microarray data should only be attempted by an experienced engineer. These systems need to have a database installed (often in a specific version) which needs to be configured properly. The only such tool that was easy to install is BASE. Analysis-only tools were generally easier to install, especially those that are completely Java-based, i.e. TM4 and Mayday posed no problems at all. The other analysis tools need additional software, mostly R/Bioconductor, with the problems mentioned above. An exception is Chipster; when only the client is used, no installation at all is required, but this also means that data is shipped to the Chipster server for analysis (see ‘Systems one-by-one’ section).

System architecture

The installed tools differ in their principal software architecture, which makes them more or less suitable for certain types of applications. Pure file-based systems, i.e. Chipster and Mayday, inherently lack an own user management and also cannot be used to share data across machines. A client-server-based system such as Chipster has the advantage that complex computations can be run on dedicated machines serving many clients, but suffers from the need to frequently transport data between clients and the server. Thus, such a system is best suited for a closed organization where connections are fast and access rights and data sharing can be dealt with on the operating system level. In contrast, systems managing their data in a database usually have a good level of support for data privacy, simply by using the user management of the underlying database. Only EzArray and CARMAWeb (when used without MARS) fail to properly exploit this option. However, in those systems data analysis is carried out at the client-side, which might, depending on the amount of data being analyzed and the complexity of the analysis, require rather strong machines.

User friendliness

The tools installed vary greatly in their user friendliness. Please be aware that user friendliness is a mostly subjective criterion; the following only reflects our own impressions.

Four of the seven tested tools, namely BASE, MARS, GeneX and TM4 (Madam), come with a dedicated storage system. In those systems, data must first be uploaded before it can be analyzed. On the other hand, only those tools allow to attach metadata to the experiments. For both tasks, all four mostly work with web forms, grouped by biosource, sample, extract, labeled extract, probe, array, hybridization and scan. Madam and MARS enforce MIAME compatibility by requiring certain fields to be filled. In contrast, BASE and GeneX allow to add annotations, but do not enforce any particular level of completeness. This is advantageous in projects where standard compatibility is not an issue, as the amount of work necessary to give MIAME-compatible metadata can be considerable. Madam, MARS and BASE present well arranged
menus and intuitively structured forms; especially Madam and MARS are very easy to handle. In contrast, we found the interface of GeneX less intuitive. BASE is the only system that allows to search the set of managed experiments using various criteria, an important feature for projects that create masses of data.

The tools follow different paradigms in terms of supporting users during data analysis. EzArray is particularly easy to use as it provides completely predefined analysis workflows, which is specifically appealing for infrequent users. Also, CARMAWeb gives helpful guidance throughout the analysis process. However, simplicity comes at a price: Both tools offer only a rather limited number of analysis methods. In contrast, MeV (TM4), Mayday and Chipster offer a full range of methods, but also require more knowledge and more training from a user. MeV and Chipster present all methods in well-arranged menus, while Mayday tends to have overloaded and complex interface elements. Further, each tool has its own strong points. Chipster provides short explanations on parameter settings and the basic idea of the method chosen, which is invaluable for more novice users. Mayday offers the MPF designer to define and refine analysis workflows which is a very handy feature for performing repetitive tasks. All three tools present the analysis workflow and its result in appealing graphical overviews; further, all three allow to store intermediate data sets and analysis results on disk for later reuse, a feature that is not included in any of the storage-oriented systems.

Privacy of data
In any project of nontrivial size, data privacy is an important issue. Researchers often do not want their experimental results to be visible to other people before they are sure about their quality and the most important results. Also, one often wishes to define access rights on an individual basis, granting access only to selected people or groups of colleagues.

Data privacy is a feature that is only offered by the storage-oriented systems. BASE provides the best possibilities in this respect; even analysis plug-ins have to be made explicitly available to other users, offering the possibility to have private analysis methods. MARS is the only tool that distinguishes between read and read-write access. TM4 has a privacy management that is only user based i.e. access rights cannot be granted to groups of users. All tools that access data primarily from file systems pay less attention to privacy; in those cases, access rights have to be set using the operating systems. CARMAWeb protects user data and analysis results by passwords, but, like TM4, does not provide group management. EzArray requires a user login, but all users can see and alter all data once logged-in. Standalone tools like MeV (TM4), Mayday and Chipster do not provide any user management at all.

Integration of additional data
Microarrays nowadays are rarely analyzed in isolation, but together with additional data, such as functional annotation of genes, arrangement of genes in pathways or positions of genes in a genome [38, 39]. Therefore, most of the tools support some integration of external data, either by directly accessing a remote web site or by importing files. But the type and extend of such data varies considerably. MARS, CARMAWeb and TM4 only allow to assign GO terms to probes. Only EzArray supports the direct access to the GEO microarray repository. GeneX, Chipster and Mayday all support annotation of genes with a comprehensive set of references to sequence, pathway, function or proteomics databases. However, most of these link-outs are only used for offering the user comfortable access to the data provided in these sources. Using them for an integrated analysis is less common. The most frequent form, gene set enrichment analysis for pathways or GO terms, is offered by Mayday, Chipster, TMA and CARMAWeb (only GO terms).

Suitability of tools for specific tasks
The tools we selected for closer review have strong points in different areas. There is no single tool that fulfills all possible requirements and that would be the method of choice for all kinds of projects. Instead, the choice of a tool must depend on the data (diversity and amount) that is to be analyzed, the level of experience of users that will use the tool, the computational infrastructure available, the types of analyses that are to be performed, etc. Here, we give an overview of which tools we suggest to have a closer look at under what circumstances.

For geographically distributed projects, a central data repository greatly eases data sharing and the performance of joint analyses. Such a repository is also indispensable for tracking project progress and reporting and can be a valuable aid in ensuring
common standards in terms of data annotation and treatment. For such cases, BASE and MARS could be the tool in favor. Both feature an elaborated user and privacy management and hook to powerful tools for data analysis. MARS also enforces MIAME compatibility. If the analysis capabilities of those tools are not sufficient, data can be downloaded and saved locally and then analyzed using one of the standalone tools. However, this requires separate installations.

For smaller projects which can manage their data in a file system, analysis-only tools could be preferable. These tools are easier to install, maintain and often also to use. While Mayday features a wide range of analysis features, it is not easily understandable for the average users. EzArray, on the other hand, is an easy-to-use, web-based analysis tool that can be recommended also for the less experienced users, but lacks some desirable functionality. Chipster has the greatest coverage of analysis features and an additional advantage if many compute-intensive tasks are required, as those can be performed on a backend computational server.

There are a number of features that are distinctive for a single system and that may be decisive in choosing a tool. For instance, EzArray can directly import public data from GEO and provides a ‘one-click’ workflow for a standard analysis. CARMAWeb has the most comprehensive support for dual-channel technologies, while Chipster excels in one-channel technologies. It also provides the best integration with external probe annotation. The preview feature in Mayday is very handy for experienced users and its support for predefined workflows greatly simplifies repetitive tasks.

To check the suitability of the different tools for dealing with larger data sets, we also tried to test all tools with an experiment containing 100 samples of Affymetrix HGU-133 Plus 2 arrays. It was no problem to load and analyze the data in the analysis-only tools (Chipster, Mayday and TM4). Differential analyses and clustering of this data set was completed within a few minutes in each of these tools. For the storage systems, this proved to be more complicated. The data set could not be imported into Geoss (missing bulk upload) and MARS (the HGU-133Plus2 array design cannot be imported). BASE and EzArray can handle the data, although importing it requires many steps and takes a lot of time. The analyses in EzArray are comparable to the analysis-only systems, while BASE itself does not support doing analyses.

CONCLUSION
We surveyed 78 tools for managing and analyzing microarray data, 22 of which were subjected to a close inspection. Seven systems were tested in a hands-on setting using a standardized analysis workflow. Our study reveals that systems differ largely in many aspects, such as completeness of functionality, ease-of-use, support for data management and/or data analysis, integration of external data, maturity of the package, etc. However, our comparative evaluation identified classes of systems that are appropriate for certain types of projects, thus providing valuable information for any project that aims at using a microarray data management system.

Despite all heterogeneity, our study also shows that certain requirements of current research are not covered by any of the systems. Most importantly, almost all systems focus exclusively on gene microarray data. Only very few systems can also manage other types of experimental data; even closely related techniques, like exon arrays [40], tissue microarrays [41] or RNA-seq [42, 43], are beyond the scope of all but a few tools. However, many current research questions require a much more complete view on cellular processes than deducible from gene expression data only [44]; important additional information include proteomic or metabolomic studies, results of RT-PCR or immunohistochemistry, protein–protein interactions and spatial information about cellular events and entities. Projects also encompassing such data currently are forced to run several systems in parallel, which is not only more costly in terms of installation and maintenance and less comfortable for users, but also makes integrated data analysis much more complex. We therefore believe that the ability to deal with a multitude of experimental techniques is a function that must be a focus of future system developments.

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements
We would like to thank all authors of the tools we reviewed for their generous provision of open source software and documentation. We already now want to apologize for any errors in our evaluation.
FUNDING
This work was funded by the Deutsche Forschungsgemeinschaft (DFG) as part of the Transregio TRR-54.

References


