The GeWare data warehouse platform for the analysis of molecular-biological and clinical data

Erhard Rahm[†‡], Toralf Kirsten[†], Jörg Lange[†]

[‡] Dept. of Computer Sciences, University of Leipzig
[†] Interdisciplinary Center for Bioinformatics, University of Leipzig

Abstract

We introduce the GeWare data warehouse platform for the integrated analysis of clinical information, microarray data and annotations within large biomedical research studies. Clinical data is obtained from a commercial study management system while publicly available data is integrated using a mediator approach. The platform utilizes a generic approach to manage different types of annotations. We outline the overall architecture of the platform, its implementation as well as the main processing and analysis workflows.

1 Introduction

Biomedical research projects often have to integrate and analyze a variety of data, in particular high-volume genome-wide expression data and different types of annotation data. Many tailored data management solutions have been developed, in particular to manage microarray-based gene expression data. Such database efforts include ArrayDB [ERP+98], GeneX [MSZ+01], M-CHIPS [FHB+02], RAD2 [SPM+01], and SMD [SHB+01]. However, these approaches often provide only limited support to integrate experimental, genetic or ontological annotations to help analyze and explain the expression behavior [DKR03]. Moreover, the systems mostly have no support to integrate clinical data as necessary for biomedical research studies, e.g. to investigate the genotype-phenotype interrelationships for diseases and their therapies.

To overcome the limitations of previous approaches and to support the analysis needs for large collaborative biomedical research projects, we have designed and implemented an integrated data management and analysis platform. The core of the platform is the *GeWare* (Genetic Warehouse) data warehouse. The key aspects of our approach are the following:

- We follow a data warehouse approach [JLVV03] to physically integrate all relevant data in a central store, in particular experimental data from microarrays [SSD+95] and Matrix-CGH arrays [SLS+97], clinical information and annotations. The data warehouse makes all data directly accessible, thus supporting both good performance and extensive analysis capabilities. A flexible multidimensional data warehouse schema is employed so that the voluminous array data can be analyzed from different perspectives and many conditions.
- Different types of annotations are supported, in particular molecular-biological, experimental and clinical annotations. Annotations are internally managed in a generic format to deal with the high degree of heterogeneity and to be open for new types of annotations. Experimental annotations can be manually specified in a consistent way based on controlled vocabularies. Publicly available molecular-biological annotation data can be integrated automatically using a mediator-based approach [KDKR05].
- Clinical data is automatically imported from a commercial study management system managing patient-oriented clinical, pathological and genetic findings in an anonymized way. We use special mapping tables based on patient identifiers to interrelate clinical in-

formation with the corresponding expression/mutation data and their experimental annotations [KLR06].

- *GeWare* provides different algorithms for pre-processing and analyzing expression data, e.g. to identify lists of interesting genes/clones. The analysis methods are typically applied to selected groups of patients or genes of interest. Users can run predefined analysis workflows or interactively analyze the data, and store analysis results for later investigation.

The *GeWare*-based platform is fully operational and being used in several research projects in Leipzig. For example, it is the central data management platform for two large collaborative cancer research studies in Germany analyzing molecular mechanism of malignant lymphoma[1] and glioma[2] [KLR06]. First results are described in [HBB+06]. Furthermore, *GeWare* is used for gene expression analysis of thyroid nodules [EKB+05] and the analysis of factors influencing the specific binding of sequences on microarrays [BKH+04, BKLS04, BPK05].

In the next section we give an overview of the platform architecture and the supported workflows. Section 3 describes the multidimensional data model of *GeWare*. In Section 4, we present the mechanisms to enforce uniform experiment annotation. Section 5 sketches the integration of external clinical data. Section 6 illustrates different analysis capabilities supported by *GeWare*. Section 7 concludes the paper.

2 System Overview

2.1 Architecture

Figure 1 shows the overall architecture of *GeWare*. Data is imported from several sources and transformed within a so-called staging area before it is integrated and stored in the central warehouse database for analysis. So-called data marts [JLVV03] support special analysis needs. They are either derived from the warehouse or are used to store the results of a particular analysis method for later reuse. All administration and analysis functions of *GeWare* are accessible via web interfaces.

The following kinds of data are integrated in the warehouse:

- *Expression data*: Influenced by requirements of joint research projects, we currently focus on expression data produced by Affymetrix microarrays. These oligonucleotide arrays use short sequences of 25 base pairs as *probes*. Each gene is represented by a so-called *probeset* consisting of 11-20 probes. A microarray measures the expression level for thousands of such probes at the same time from which the corresponding probeset intensities are derived. *GeWare* has import interfaces for probe-level and probeset-level intensity values. While probe intensities are typically stored within so-called CEL files the probeset intensities are usually represented by expression matrices and stored within flat files. Obviously, probe intensities are much more voluminous but allow us to apply different normalization methods to clean the data as well as different aggregation methods to combine the intensities of the probes to probeset intensities. When importing expression data from expression matrices in flat files the user needs to specify the corresponding normalization and aggregation methods.

[1] http://www.lymphome.de/en/Projects/MMML/index.jsp

[2] http://www.gliomnetzwerk.de

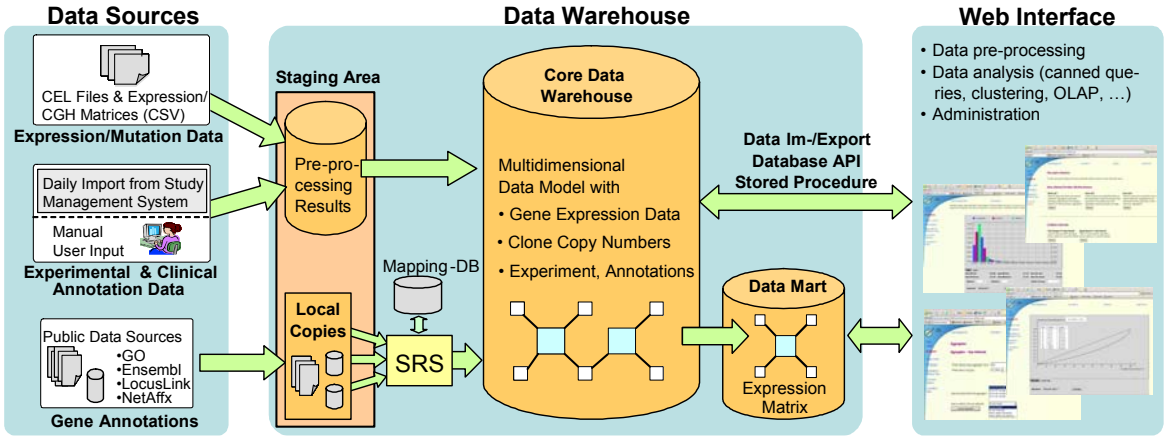


Figure 1: Overall architecture of GeWare

- *Mutation data*: This data focuses on the genetic diversity. Usually, genes are located at fixed positions on a chromosome. However, individual mutations (insertions, deletions, moves) of sequence regions can have a significant impact on the expression and, thus, on associated gene functions. This holds particularly for large block-wise mutations, such as copies and movements of DNA sequences across different chromosomes. Matrix-based comparative genomic hybridization (Matrix-CGH) arrays are a currently used technique to measure such genetic imbalances [SLS+97]. A Matrix-CGH array comprises thousands of so-called *clone*s, i.e. selected DNA sequences from specific chromosome regions, for which the occurrence (copy) number is comparatively measured, e.g. for tumor vs. healthy tissue samples. Sequences with a high (zero) copy number refer to recurrent (unavailable or heavily mutated) chromosome regions such that higher (lower) expression values of genes within these regions can be explained. Like Affymetrix microarrays for determining the gene expression, this technique operates genome-wide and, thus, generates a huge amount of data. *GeWare* has interfaces to import files comprising Matrix-CGH mutation data.
- *Experiment & clinical annotations*: Experiment annotations can be specified by users via web interfaces together with the import of new expression data. *GeWare* allows the specification of experiment-specific annotation templates to indicate which descriptions should be provided and which controlled vocabularies be used. Based on these templates the web pages for user annotation are automatically generated. Relevant portions of clinical data are imported automatically from a study management system, eRN in our projects, which holds patient data already in anonymized form.
- *Gene annotations*: *GeWare* integrates publicly available annotation data describing biological objects, such as genes, clones, and proteins, which are necessary to interpret analysis results. For this purpose, we apply the mediator-based integration approach outlined in [KDKR05]. We utilize the commercial system SRS to quickly obtain data from sources such as NetAffx [LLS+03], Ensembl [BAB+04], and GeneOntology [HCI+04].

Imported data first has to be cleaned and transformed for integration. The intermediate results of these pre-processing steps are stored in a dedicated staging area. For expression data, several methods for normalization and aggregation are supported to pre-process probe-level expression data, since there are not yet generally accepted approaches for these tasks.

The *GeWare* data warehouse is organized according to the multidimensional data model described in Section 3. For its implementation we use the relational database management system DB2 of IBM (on a high-end Linux server) supporting very high data volumes and a mul-

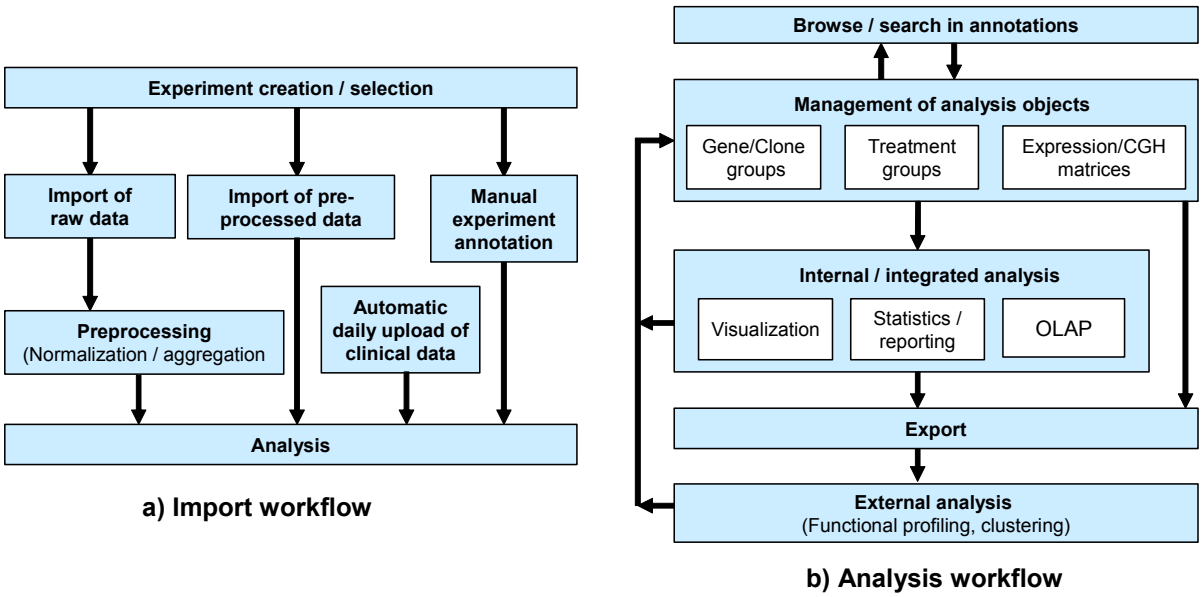


Figure 2: Operational workflows in GeWare

titude of performance tuning options such as indexing, materialized views, data partitioning, etc. Specific portions of the warehouse can be redundantly stored in data marts to improve performance for special analysis tasks. For example, expression (CGH) matrices can be extracted for relevant gene (clone) groups and treatment groups and saved in corresponding data marts so that they can quickly be reused and visualized without recalculation.

Data access in *GeWare* is authenticated through a sophisticated concept of user/user group and right management. In particular, access rights can be granted/revoked not only for the data (expression/mutation, annotations and clinical data), but also for the functions on the data, such as import, export, query etc. According to the user profile, the web interface is automatically generated to only cover the allowed functions.

2.2 Workflows

Figure 2a and b show common workflows supported by *GeWare* for data import and analysis, respectively. To import new data, an experiment first needs to be created in the data warehouse. Experiments serve as container objects to hold all data generated from microarrays, i.e. probe intensities, probeset/gene intensities, and Matrix-CGH arrays as well as relevant experiment annotations and clinical information. While mutation data are imported in pre-processed form, expression data can be imported in both, raw and pre-processed form. *GeWare* also supports batch import of expression data for many microarrays at the same time. Raw expression data further has to be normalized and aggregated using an integrated method. Independently from the import process of experimental data, each chip can be fully described by the experimenter using a selected annotation template. Moreover, clinical data is automatically imported for pre-defined experiments and associated with an existing annotation template.

GeWare supports several forms of data analysis, such as visualization, pre-defined reports and statistical algorithms. Flexible combination and integration of the different methods is achieved by means of the uniform exchange of gene/clone groups, treatment groups (i.e. groups of chips), and gene expression/CGH matrices, which are centrally created and man-

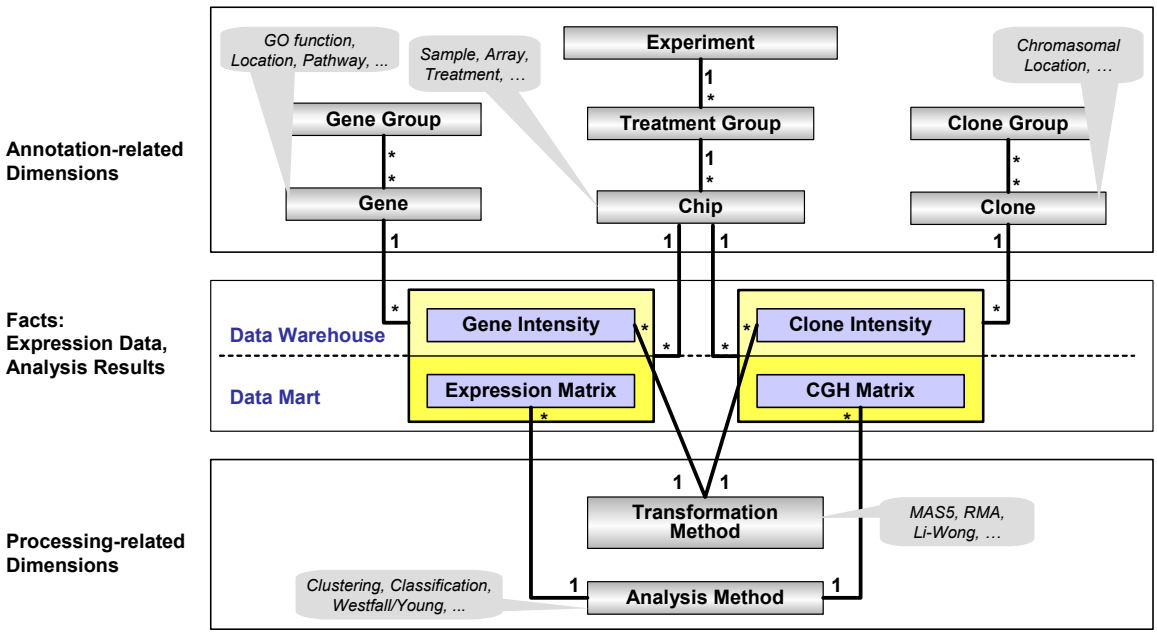


Figure 3: High-level data warehouse schema

aged by *GeWare*. Groups of interesting chips[3] and genes may be formed by manual user specification, by querying available annotations, or using some analysis algorithms. Focused analysis is possible with expression (CGH) matrices generated for the relevant treatment and gene (clone) groups. The results, typically sub-groups of relevant genes and clones, respectively, can in turn be saved for further queries and analysis. *GeWare* also supports the export of the pre-computed analysis results, e.g. gene/clone groups and expression/CGH matrices, to perform analysis in external tools, such as for functional profiling and clustering.

3 Data Warehouse Model

Figure 3 shows a high-level view on the multidimensional data warehouse schema of *GeWare*. It is built of *dimension* and *fact* tables. Facts are numeric and additive data, while dimensions provide information on the meaning of facts or how they have been determined.

Currently, our schema includes two main fact tables, *Gene Intensity* and *Clone Intensity*, representing expression intensity values at the gene (probeset) level and copy numbers for clones. The gene fact table is used to store the probeset intensities imported from flat-file expression matrices or as the results of other aggregation methods that were applied to raw expression data stored in CEL files. Additional fact tables are kept for data marts, e.g. *Expression Matrices* and *CGH Matrices*, to store the intensities of those genes (clones) participating in gene (clone) groups determined by a specific analysis method, such as clustering.

The dimensions can be grouped into annotation- and processing-related dimensions, which are shown in Figure 3 together with some illustrating examples. Annotation-related dimensions include tables for g*enes*, *chips*, and *clones* and their groupings. The experiment dimension holds the user-specified conditions of the experiments, while the gene (clone) dimension provides the facts currently known about the genes (clones), such as their functions. Processing-related dimensions specify *transformation* and *analysis methods* describing the computational methods and their parameters used to compute gene/clone intensities and to determine gene (clone) groups for expression (CGH) matrices, respectively.

[3] Both, gene expression microarrays and Matrix-CGH arrays, are commonly called chips.

Typically, dimensions are organized in generalization / specialization hierarchies and, thus, providing different levels of abstraction for analysis. For example, the experiment dimension is organized into three levels, experiment, treatment group and chip. Experiment is the most abstract level describing a biological experiment encompassing many chips which can be grouped into so-called treatment groups. Each of the treatment groups may include chips for a specific experimental condition, e.g. replicates of diseased tissue probes from a specific organism.

The sketched multidimensional data model supports a high flexibility for gene expression analysis. While current approaches typically evaluate a complete data matrix, e.g. gene expression matrix and CGH matrix, containing the intensity values for all measured genes/ clones and several/all chips, we now can focus on individual or comparative analysis to an arbitrary subset of intensity values determined by specific annotation values of interest. The selection may be based on a value at a specific level of a single dimension or any combination for several dimensions (e.g. compare different analysis methods for a given treatment group and a gene group with a particular GeneOntology function). Moreover, the data model is easily extensible. Within each dimension, new processing methods or annotations can be added without affecting the existing data organization. New data marts and fact tables can also be added and associated to the existing or new dimensions.

4 Specification of consistent Experiment Annotation

Depending on the biological focus, microarray experiments can be conducted and documented in different ways. For example, annotating the time points in a time-series experiment is not necessary for an experiment comparing normal and diseased tissues. Such different experiment designs make it difficult to achieve a single unique schema for experiment annotation. Addressing this problem, the MIAME standard [BHQ+01] gives a recommendation about the minimal information to be captured about a microarray experiment. While specify-

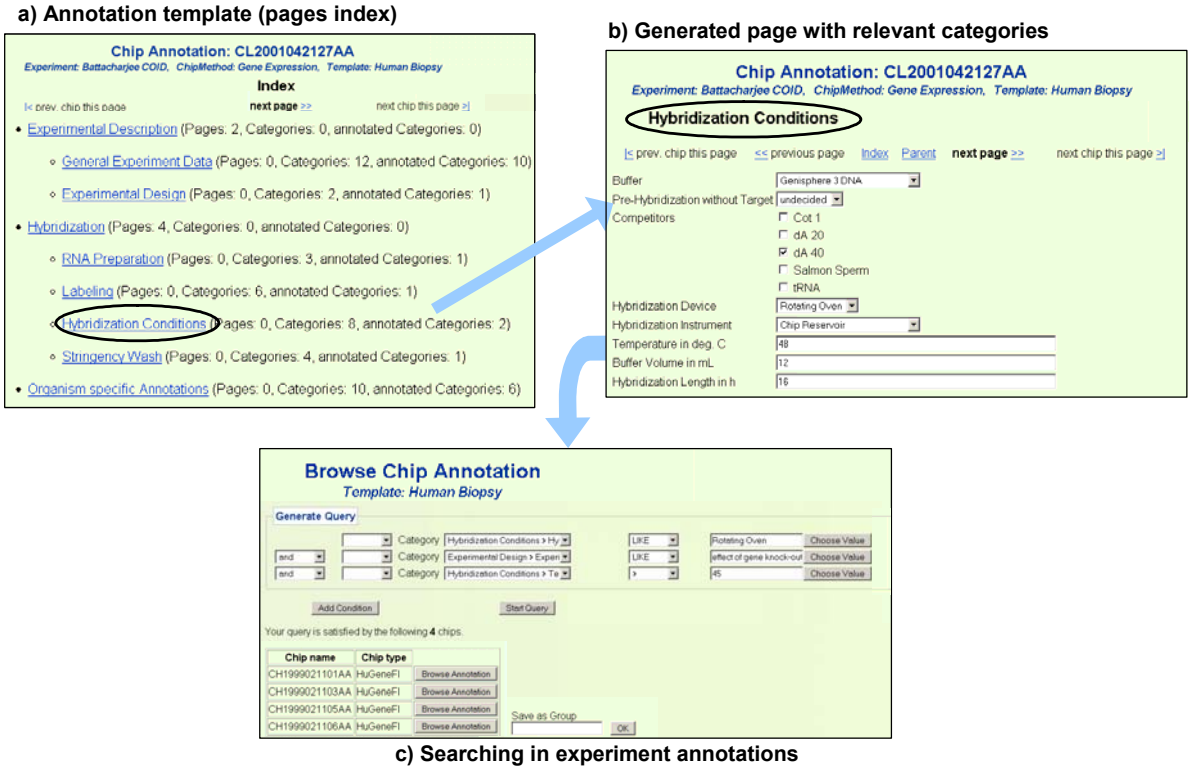


Figure 4: Capturing experiment annotation

ing the categories to be filled in, such as for array design, sample description, hybridization procedure, etc., MIAME leaves open how the values are to be filled in for those categories. This can easily lead to conflicting values, such as "Homo sapiens" and "Human" for an *Organism* category.

To achieve a consistent experiment annotation we support so-called *annotation templates* to prescribe the categories to be annotated and *controlled vocabularies* to constrain the values for the categories. An annotation template is typically hierarchically organized and consists of *pages*, which group related *categories* together, such as for array design, description of samples etc. The definition of an annotation template consists of constructing such pages, specifying relationships between the pages, and eventually defining categories for the single pages. A category is either of type “free text” or, preferably, controlled. In the latter case, an existing controlled vocabulary has to be chosen to provide corresponding choices for user input. It is possible to copy and modify an existing template to speed up the construction of a new, similar template. Currently, a template consisting of MIAME categories is provided as a basis to construct new project-specific templates. *GeWare* supports user-defined controlled vocabularies as well as existing standard ontologies, such as NCBI taxonomy. Once defined or imported, a controlled vocabulary can be shared by different templates.

From the definition of the annotation template, *GeWare* automatically generates web pages for user input. To annotate an experiment and all its containing chips, the user simply walks through the generated pages to provide values for the corresponding categories. Figure 4a shows the index of all pages defined in an existing template, *Human Biopsy*. A page is displayed in Figure 4b and contains categories to hybridization conditions of an in-vivo experiment. The values for vocabulary-based categories can be chosen from corresponding select and check boxes.

By using a single template, experiments can be uniformly annotated, making it easy to identify related chips within an experiment by searching in their annotations. As indicated in Figure 4c, the terms from controlled vocabularies are shown for the corresponding categories so that the user can specify search criteria. The criteria are combined using the logical operators AND, OR, and NOT. The identified chips can then be saved as a treatment group for later analysis.

5 Integration of clinical Data

While experiment annotations describe the experimental procedure and conditions, clinical data describes the state of a patient or her tissue probes at a specific point in time. Typically there are different types of clinical findings. Such findings are generated whenever a patient visits a doctor for a normal check-up or in case of an emergency. Pathological findings are created by pathologists and report about the state and abnormalities of a patients' extracted tissue sample. Collecting and integration such findings are of high interest in a clinical trial to address medical research questions. Study management systems are established tools to maintain clinical findings. They allow authorized users across different organizations and institutions to consistently and autonomously specify relevant portions of the produced clinical and pathological findings together with patient-related personal data, such as age, sex, material status.

In Leipzig, we use the commercial study management system eRN[4] to centrally collect all relevant clinical data for a certain set of trials. In this system, the data is stored in anonymized

[4] http://www.ert.com

form, i.e. without patient-identifying data, such as personal id card number, social insurance number or the patients' name. All patient data is associated with an automatically generated technical identifier, the patient id. Since the study management system does not store voluminous chip data, such as for microarray-based gene expression and array-CGH-based mutation data, the *GeWare* platform daily imports a relevant portion of clinical data from eRN. To relate clinical data from the study management system with expression and mutation data in *GeWare*, we use a manually managed mapping table in which patient ids are associated to chip ids which are used for expression and mutation data in *GeWare*. The clinical data is described in a pre-defined annotation template. Once imported, *GeWare* users can utilize the clinical data, e.g. to group and analyze chips sharing similar clinical and experimental conditions (see also Figure 4c). Moreover, the integration of clinical data in *GeWare* permits an over-spanning analysis of expression and mutation data together with all relevant annotations and clinical data as illustrated in the next section.

6 Expression Analysis

To illustrate the usefulness of the platform, we present a selection of analysis methods supported by *GeWare*. We first focus on the pre-processing of raw gene expression data. We then discuss routines for statistical analysis, reporting and visualization.

6.1 Pre-processing

Pre-processing consists of two main tasks, 1) normalization to clean noise signals from raw probe intensities and to make the probe intensities from different chips comparable, and 2) aggregation to produce gene/probeset intensities from normalized probe intensities. To limit the implementation effort, *GeWare* utilizes the open-source package BioConductor [GCB+04], and thus has access to a library of different methods commonly used in each pre-processing step. *GeWare* allows tailoring a pre-processing strategy by combining different methods per step. BioConductor also offers more sophisticated algorithms covering both steps. Such algorithms including Speed's RMA [IBC+03], variants of Li/Wong's model [LW01] and Affymetrix MAS5 [Aff02] can also be specified in *GeWare*.

6.2 Reporting and statistical analysis

GeWare provides various functions for interactive analysis of gene/clone intensity values. These include different visualization forms and reports to detect groups of genes (clones) with specific expression (mutation) patterns.

Expression/mutation data can be visualized in different ways. Figure 5a shows an expression heat map, i.e. a colored expression matrix labeled by genes as rows (right hand side) and chips/patients as columns (on the bottom). Both, genes and chips, are hierarchically clustered according to their expression data. The clustering results are represented by two dendrograms on top (genes) and on the left hand side (chips/patients) of the matrix. The colors symbolize over-expression (red), under-expression (green) and no-change (black) in comparison to the average expression level. In addition, a classification of the chip data by pre-defined clinical classifiers is visualized by a colored band on top of the heatmap. In the shown example, the cancer stage from the clinical diagnosis is used as a classifier. Users can thus visually determine the correlation between the clinical classifier and clustered expression data.

Figure 5b shows a line chart illustrating the expression profile of genes over multiple chips/patients to help recognize co-expressed genes, i.e. genes with similar expression profile, possibly on different levels. Another often used visualization technique is the M/A plot (Fig-

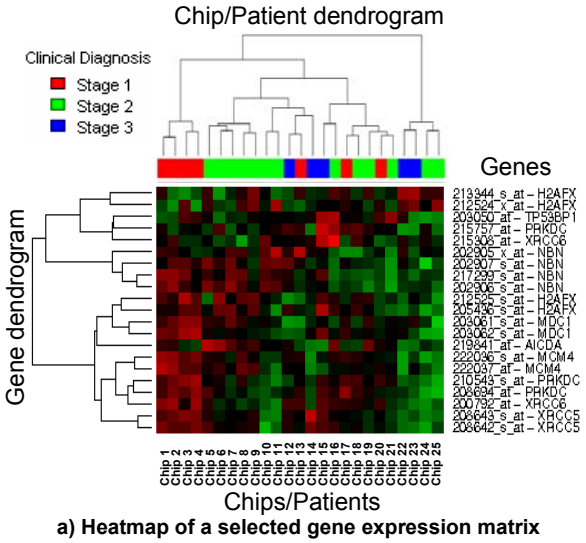
ure 5c) comparing the expression values of two selected chips. These plots are useful to visually identify differentially expressed genes.

Furthermore, *GeWare* has built-in statistical approaches to identify interesting genes from expression data, such as outlier genes and differentially expressed genes. Outliers are genes showing extreme intensity values in one chip or treatment group compared to other chips or treatment groups. Figure 6d shows part of an outlier detection report for two treatment groups based on standard deviation: outlier genes are displayed along the selected annotations, i.e. gene symbol and map location. For the determination of differently expressed genes more complex statistical tests like the Westfall/Young approach [WY93] are supported.

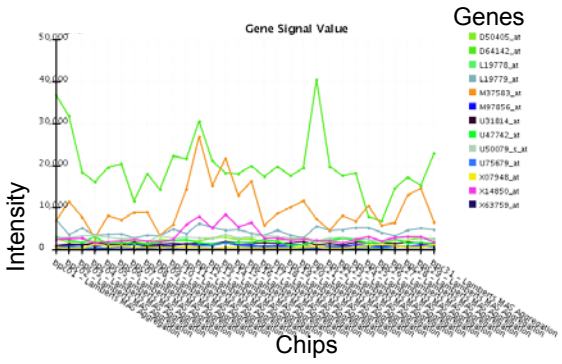
Similar to searching in gene (clone) annotations, the reports produce groups of interesting genes (clones). The gene/clone groups can be used as input in further analysis steps. Hence, iterative analysis is possible by continuing with pre-computed results, e.g. to successively filter the genes and clones of interest. Furthermore, gene/clone groups and expression/CGH matrices can be exported for analysis in stand-alone tools. While expression/CGH matrices are the typical input format of existing clustering tools, gene groups can be used to perform functional profiling in tools such as FUNC [MDK+03] and GenMapper [DR04].

7 Conclusions

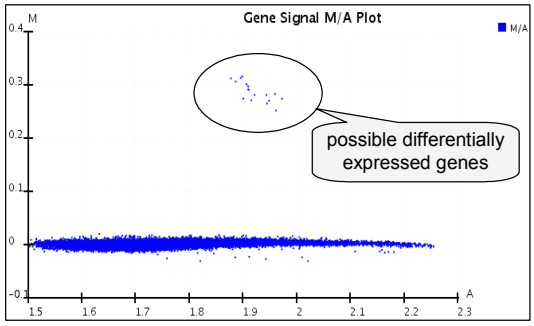
We presented a data warehouse platform for the integrated analysis of high-volume experimental chip data and different types of annotations including clinical information. The *Ge-*



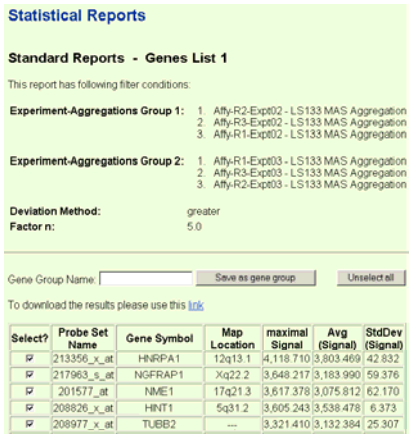
a) Heatmap of a selected gene expression matrix



b) Expression profiling: Probeset intensities of a selected treatment group



c) M/A plot to visualize differentially expressed genes



d) Report for outlier genes

Figure 5: Interactive analysis and of gene expression data

Ware data warehouse stores all relevant data in a multidimensional data model. Consistent experiment annotation is achieved by means of pre-defined annotation templates and controlled vocabularies. Relevant portions of clinical information are imported from a commercial study management system holding anonymized patient-related data. Furthermore, gene/clone annotations are integrated from selected public sources. Different types of statistical and visual analysis are supported, e.g. for pre-processing (normalizing) expression data, for clustering genes with similar expression behavior, to correlate expression data with clinical status indicators, and to determine differentially expressed genes and outlier genes. The analysis results are uniformly stored in treatment and gene/clone groups or expression/CGH matrices, which can be exchanged between different analysis steps. The *Ge*Ware-based platform is operational and currently supports several research projects in Leipzig as well as two large Germany-wide clinical trials. In future work we will add more data sources and analysis capabilities, in particular additional chip types and ontology-based analysis.

Acknowledgments: We thank Hans Binder (IZBI, University of Leipzig), Friedemann Horn, Knut Krohn, Markus Eszlinger, Hilmar Berger and Markus Löffler (Medical Department, University of Leipzig) for continuous cooperation and helpful comments on *GeWare*. This work is supported by DFG grant BIZ 6/1-3.

References

[Aff02] Affymetrix: Statistical algorithms description document. White Paper, (http://www.affymetrix.com), 2002.

[BKH+04] H. Binder, T. Kirsten, I. Hofacker et al: Interactions in oligonucleotide hybrid duplexes on microarrays. Journal Physical Chemistry in the Biology, 108(46):18015-18025, 2004.

[BKLS04] H. Binder, T. Kirsten, M. Löffler, P. Stadler: Sensitivity of microarray oligonucleotide probes: Variability and effects of base composition. Journal Physical Chemisty in the Biology, 108(46):18003-18014, 2004.

[BPK05] H. Binder, S. Preibisch, T. Kirsten: Base pair interactions and hybridization isotherms of matched and mismatched oligonucleotide probes on microarrays. Langmuir 21:9287-9302, 2005.

[BAB+04] E. Birney, T. Andrews, P. Bevan et al: An overview of Ensembl. Genome Research, 14: 925-928, 2004

[BHQ+01] A. Brazma, P.Hingamp, J. Quackenbush et al: Minimum information about a microarray experiment (MIAME) - Toward standards for microarray data. Nature Genetics 29:365-371, 2001

[DKR03] H.-H. Do, T. Kirsten, E. Rahm: Comparative evaluation of microarray-based gene expression databases. Proc. 10th Conf. Database Systems for Business, Technology and Web (BTW), 2003

[DR04] Do, H.H., Rahm, E.: Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach. Proc. EDBT 2004, Springer LNCS, 2004

[EKB+05] M. Eszlinger, K. Krohn, K. Berger et al: Gene Expression Analysis Reveals Evidence for Increased Expression of Cell Cycle-Associated Genes and Gq-Protein-Protein Kinase C Signaling in Cold Thyroid Nodules. The Journal of Clinical Endocrinology & Metabolism 90(2): 1163-1170, 2005.

[ERP+98] O. Ermolaeva, M. Adams, O. White et al: Data management and analysis for gene expression arrays. Nature Genetics 20, 1998.

[FHB+02] K. Fellenberg, N. Hauser, B. Brors et al: Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis. Bioinformatics 18(3):423-433, 2002.

[GCB+04] R. Gentleman, V. Carey, D. Bates et al: Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5(10): R80, 2004.

[HBB+06] M. Hummel, S. Bentik, H. Berger et al: A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. The New England Journal of Medicine 354(23): 2419-2430, 2006.

[HCI+04] M. Harris, J. Clark, A. Ireland et al: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32(Database Issue):258-261, 2004.

[IBC+03] R. Irizarry, B. Bolstad, F. Collin et al: Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research 31(4):E15, 2003.

[JLVV03] M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis (Eds): Fundamentals of data warehouses, Springer, 2nd ed., 2003.

[KDKR05] T. Kirsten, H.-H. Do, C. Körner, E. Rahm: Hybrid integration of molecular-biological annotation data. Proc. Intl. Workshop on Data Integration in the Life Sciences, 2005.

[KLR06] T. Kirsten, J. Lange, E. Rahm: An integrated platform for analyzing molecular-biological data within clinical studies. Proc. EDBT-Workshop Information Integration in Healthcare Applications, Springer-Verlag, LNCS 4254, 2006.

[LLS+03] G. Liu, A. Loraine, R. Shieta et al: NetAffx - Affymetrix probesets and annotations. Nucleic Acid Research, 31(1):82-86, 2003.

[LW01] C. Li, W. Wong: Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, Proc. National Academy of Sciences, 98(1):31-36, 2001.

[MDK+03] B. Mützel, H.-H. Do, P. Khaitovich et al: Functional Profiling of Genes Differently Expressed in the Brains of Humans and Chimpanzees. Proc 2nd Biotechnology Day, 264-265, Leipzig, 2003.

[MSZ+01] H. Mangalam, J. Stewart, J. Zhou et al: GeneX: An open source gene expression database and integrated tool set. IBM Systems Journal 40(1):552-569, 2001.

[SHB+01] G. Sherlock, T. Hernandez-Boussard, A. Kasarskis et al.: The Stanford microarray database. Nucleic Acids Research 29(1):152-155, 2001.

[SLS+97] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer et al.: Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. Genes Chromosomes Cancer 20(4):399-407, 1997.

[SPM+01] C. Stoeckert, A. Pizzarro, E. Manduchi et al: A relational schema for both array-based and SAGE gene expression experiments. Bioinformatics 17(4):300-308, 2001.

[SSD+95] M. Shena, D. Shalon, R.W. Davis et al.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-470, 1995.

[WY93] P. Westfall, S. Young: Resampling-based multiple testing: Examples and methods for p-value adjustment. Wiley Interscience, 1993.