A Data Warehouse-based Gene Expression Analysis Platform

T. Kirsten, H.-H. Do, E. Rahm
University of Leipzig, Germany
www.izbi.de, dbs.uni-leipzig.de

Current Activities and Selected Publications (1)

- DILS 2004
  - Rahm: Data Integration in the Life Sciences. Springer-Verlag, LNBI 2994, 2004

- GenMapper
  - Do, Rahm: Flexible Integration of Molecular-biological Annotation Data: The GenMapper Approach. Proc. EDBT 2004, Heraklion, Greece, March 2004
  - Joint work with MPI EVA
- GeWare
  - Do, Kirsten, Rahm: Comparative Evaluation of Microarray-based Gene Expression Databases, Proc. 10th Conf. on Database Systems for Business, Technology, and the Web, 2003
  - Kirsten, Do, Rahm: A Multidimensional Data Warehouse for Gene Expression Analysis. Poster/Abstract, Proc. German Conference on Bioinformatics (GCB), Munich, October 2003
  - The IZBI Gene Expression Analysis Platform, Internal Status Report, IZBI, 2003

Current Activities and Selected Publications (2)

- GenBank Management
  - Joint work with G. Fritzsch (AG4)
- Oligo Sequence Sensitivity Analysis
  - Project involvement (coordination and main analysis by H. Binder)
  - Binder et al: The effect of base composition on the sensitivity of microarray oligonucleotide probes. In submission
  - Binder et al: Interactions in oligonucleotide duplexes upon microarray hybridization. In submission

Outline

- Motivation
- GeWare architecture
- Annotation integration
- Analysis support
- Conclusions

Gene Expression Data

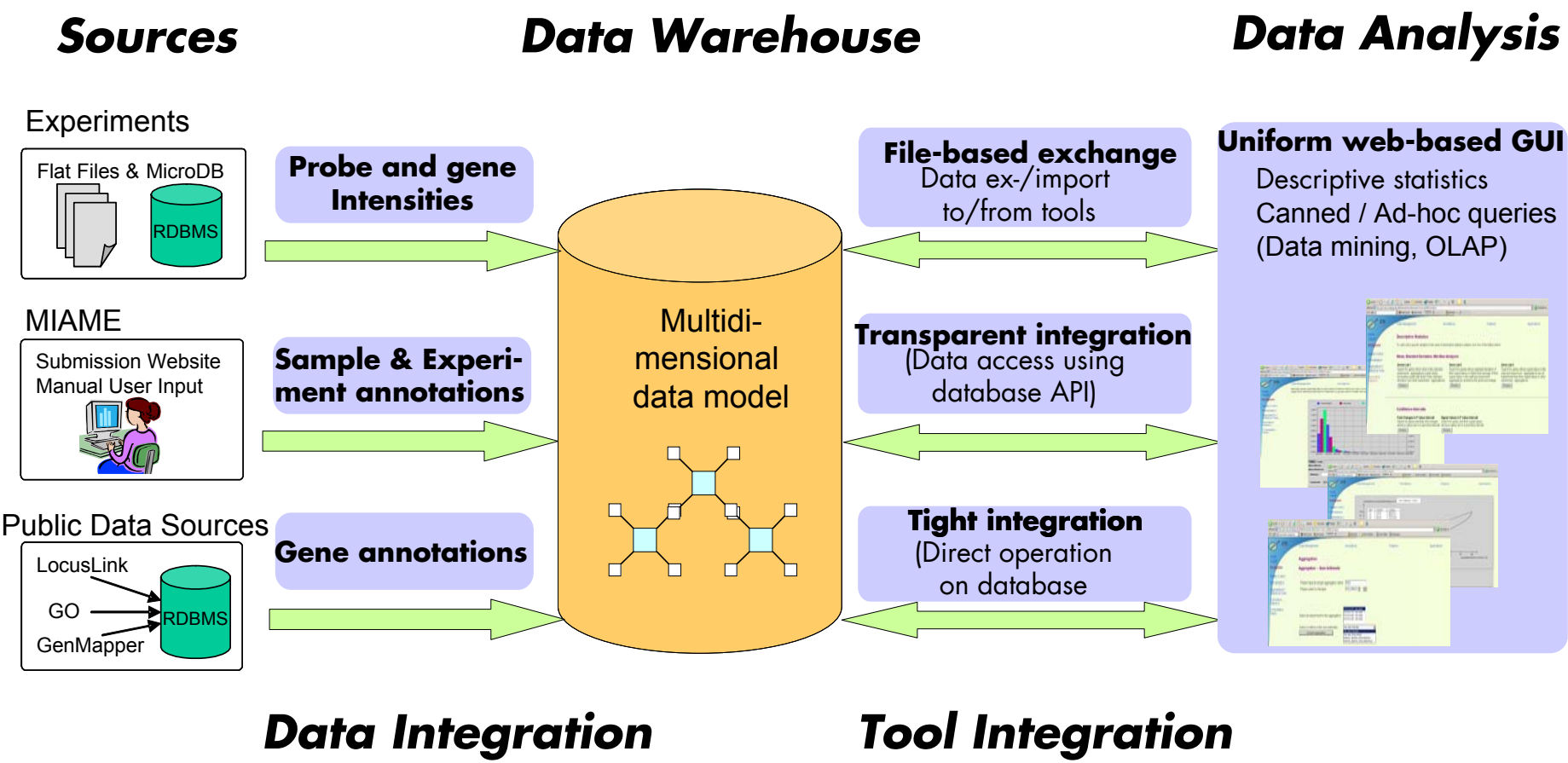
- Microarrays to measure expression of thousands of genes at the same time
- Various kinds of data with different characteristics and requirements

Data		Source	Type	Characteristics	Usage
Image Data		Array scan	binary	large files	Generation of expression data
Expression Data		Image analysis	number	fast growing volume	Visualization, statistical and cluster analysis
Annotation Data	Gene	External public sources	text	regularly updated	Interpreting / Relating / Inferring gene functions
	Experiment	User input		user-specified, often free text	

Goals

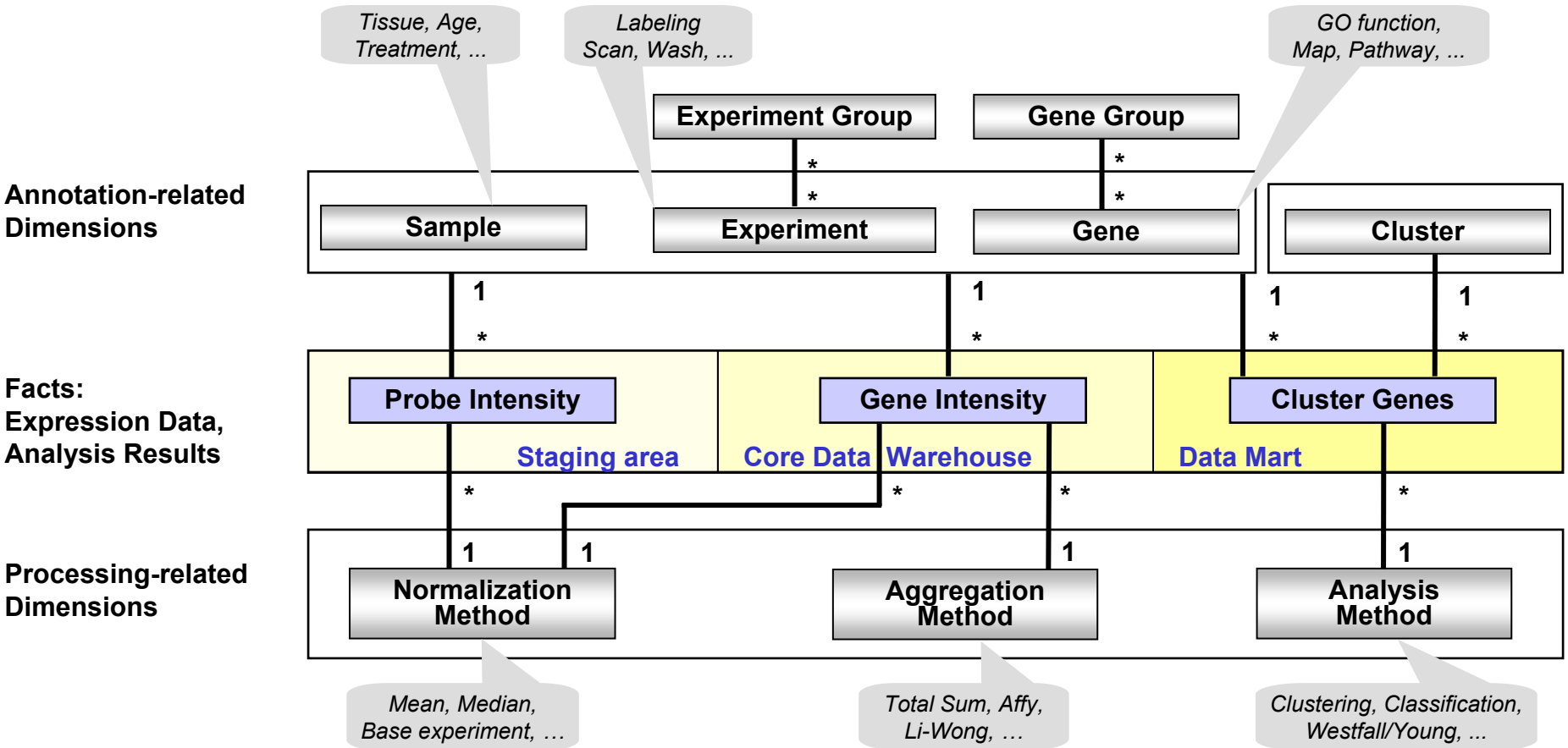
- Central data management and analysis platform
- Data Warehouse approach
  - Expression data import, e.g. from Affymetrix system
  - Fact tables to store both raw and derived data
  - Uniform specification of experiment annotations
  - Integration of gene annotations from public sources
  - Integration of analysis and data mining algorithms/tools

System Architecture

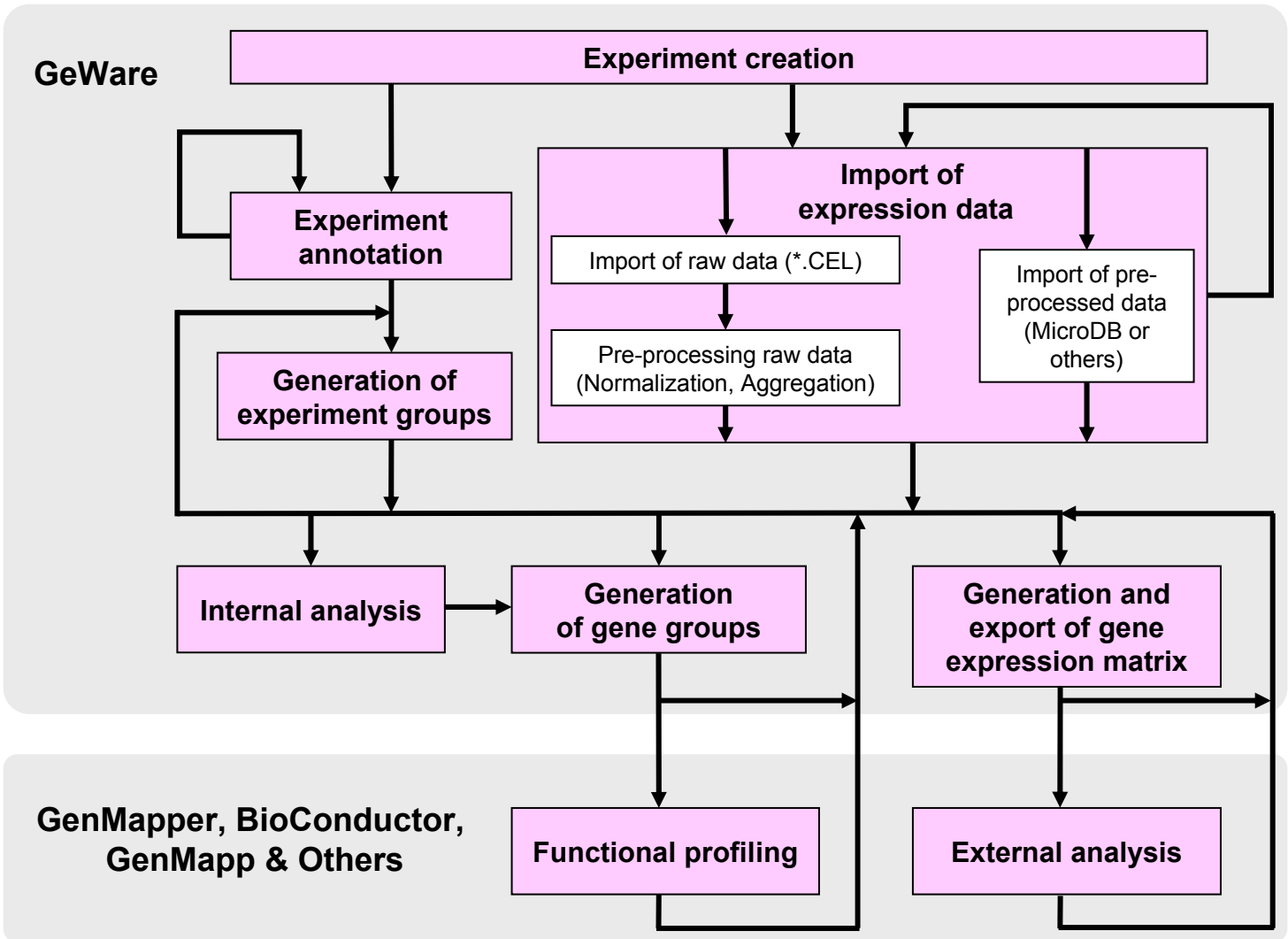


Data Warehouse Model

- Multidimensional data model (star schema)



Analysis Workflow



Experiment Annotation (1)

- Goal: Uniform and comprehensive annotation
- Controlled annotation vocabularies
  - Sets of predefined terms
- Annotation templates
  - Collections of annotation categories for which the annotation values has to be captured
  - Hierarchical arrangement of categories
  - Definition of MIAME compliant templates (Human biopsy, Human cell line, …) in cooperation with biologists
- MAGE-ML export (data exchange)

Experiment Annotation (2)

- Template specification
  - Easy specification and adaptation
  - Automatically generated web GUI

Category Definitions

Page: *culture conditions*

Please note: Select boxes, check boxes and radios have to possess a vocabulary.
All other types don't have a predefined vocabulary.

Name	Position	Type	Vocabulary	Mandatory	
		check box		yes no	New
medium	1	heading 2		yes no	Save
medium type	2	select box	medium type	yes no	Save
serum	3	radio button	serum	yes no	Save
antibiotics	4	check box	antibiotics	yes no	Save
atmosphere	5	heading 2		yes no	Save
oxygen	6	input field		yes no	Save
carbondioxide	7	input field		yes no	Save

Experiment Annotation Specification

Template: *Cell biological microarray experiments 1*

Submission: *Experiment1*

culture conditions

<< previous page next page >>

back to the start page of the submission

medium

medium type* DME

serum* fetal calf serum / horse serum

antibiotics penicillin / streptomycin

atmosphere

oxygen 12 %

carbondioxide

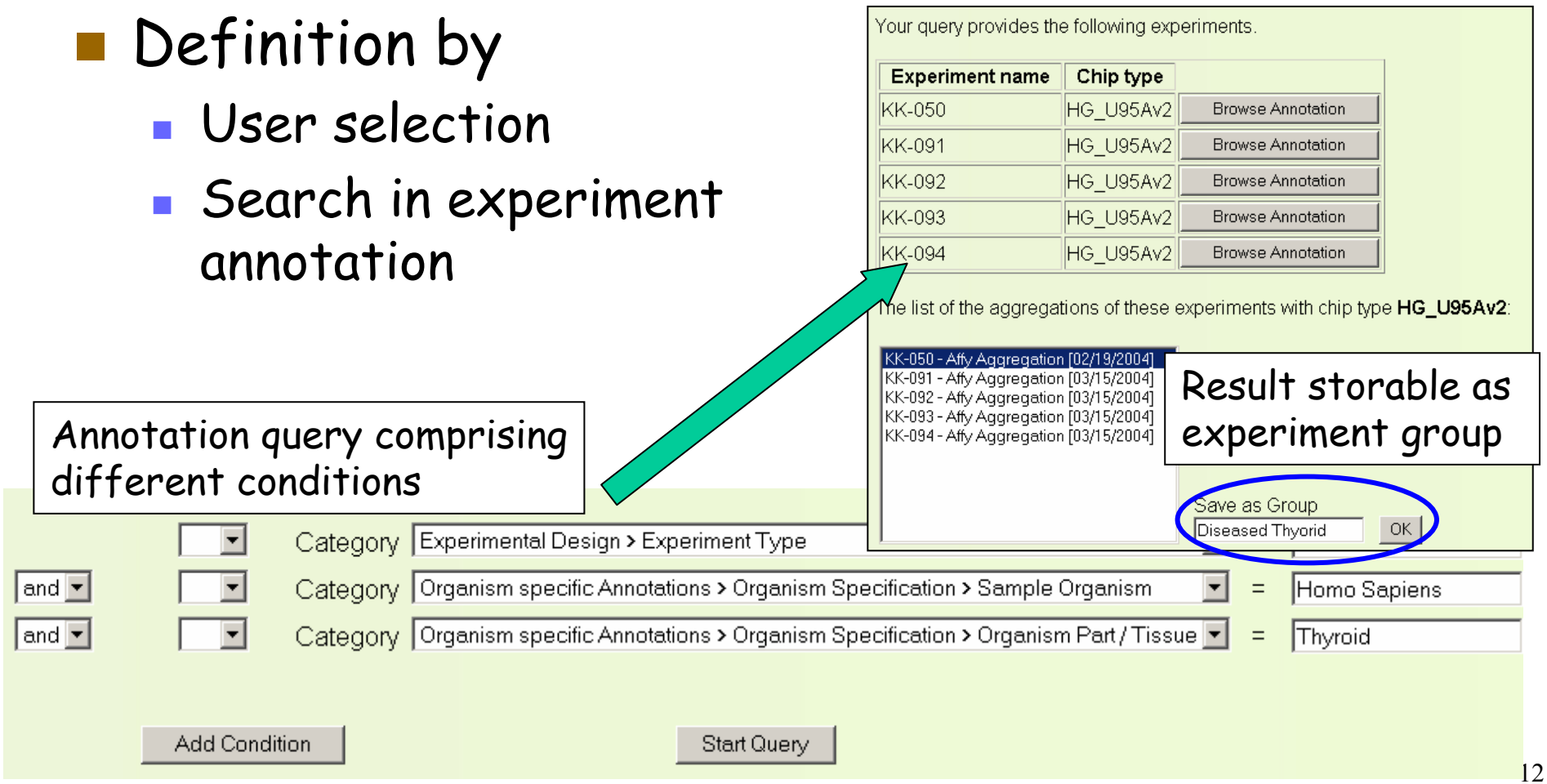
temperature 20°C

Store

End of submission

Experiment Groups

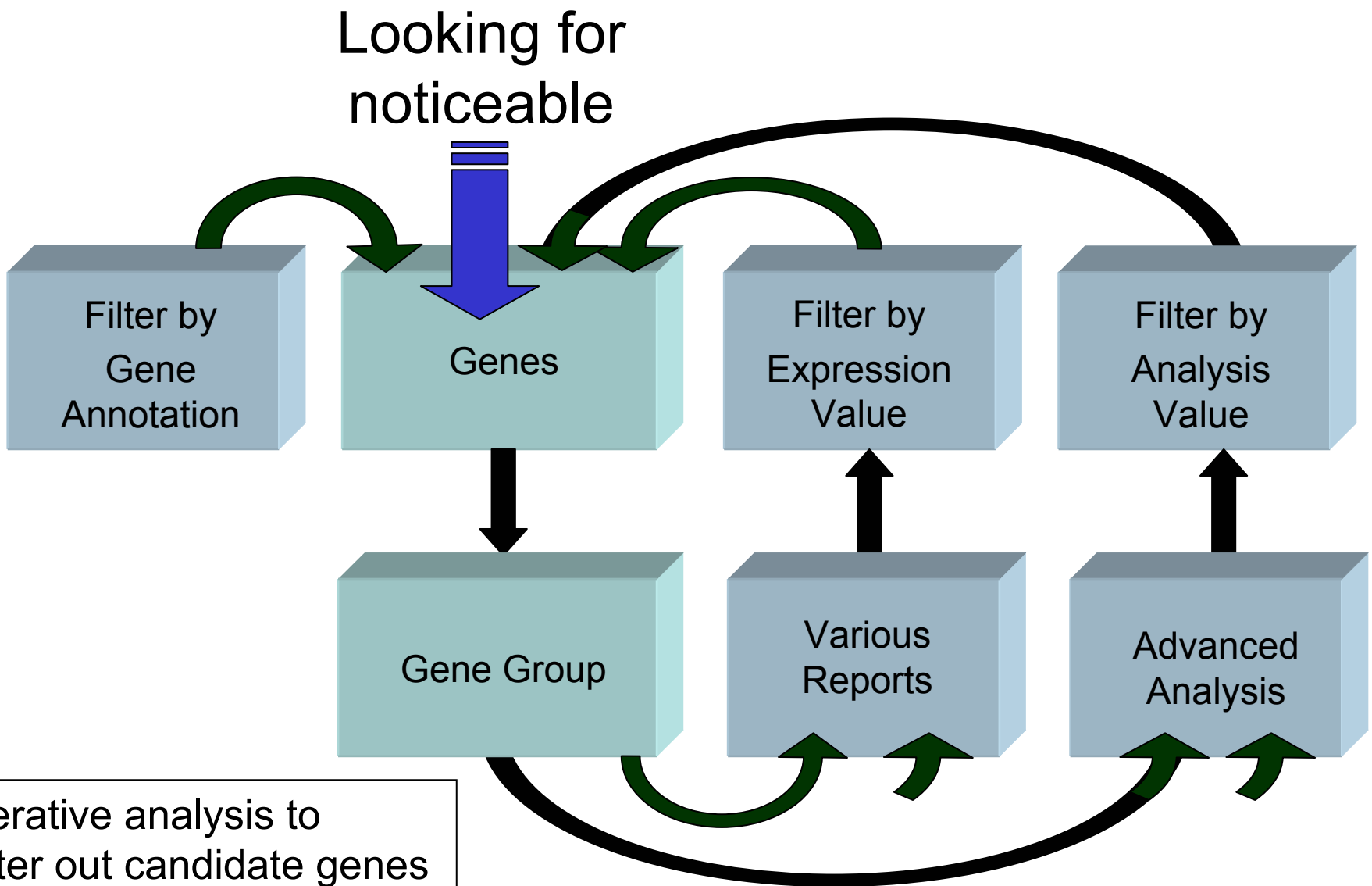
- Collections of experiments with common patterns
- Input for reporting and further analysis
- Definition by
  - User selection
  - Search in experiment annotation



Gene Annotation Integration

- Materialized integrated gene annotations
  - Source: Affymetrix Netaffx
  - Various annotation attributes (unigene, locuslink, map location, gene symbol …)
  - Directly associated with the gene dimension
- Application
  - Gene group generation
  - Direct access in expression analysis
- Future work: More annotations from different public sources

Gene Group Generation and Usage



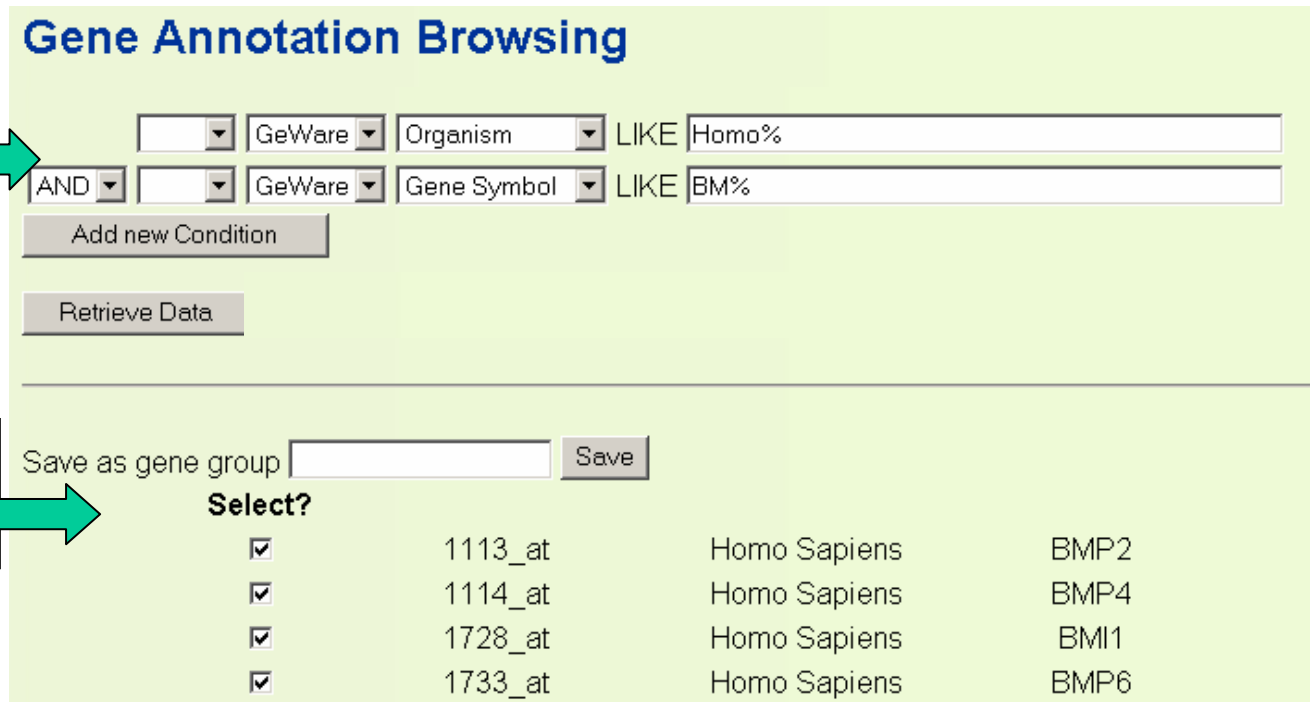
Iterative analysis to filter out candidate genes

Gene Annotation Filter

- Application of different search types (exact / fuzzy matching)
- Combination of filter conditions using boolean operators (and, or, not)

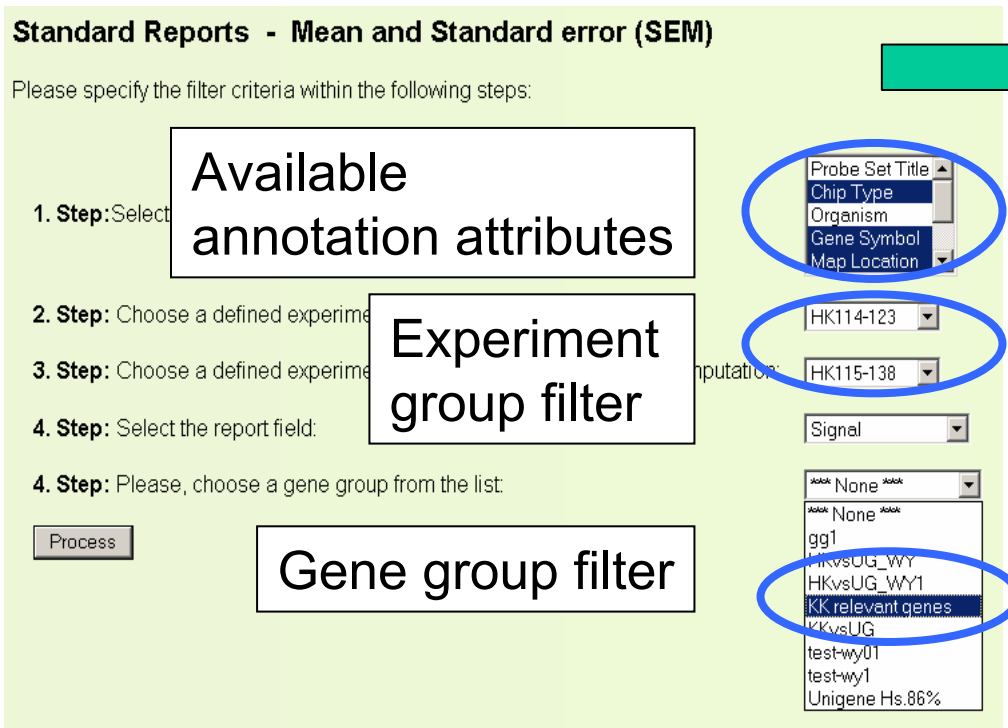
Gene annotation conditions

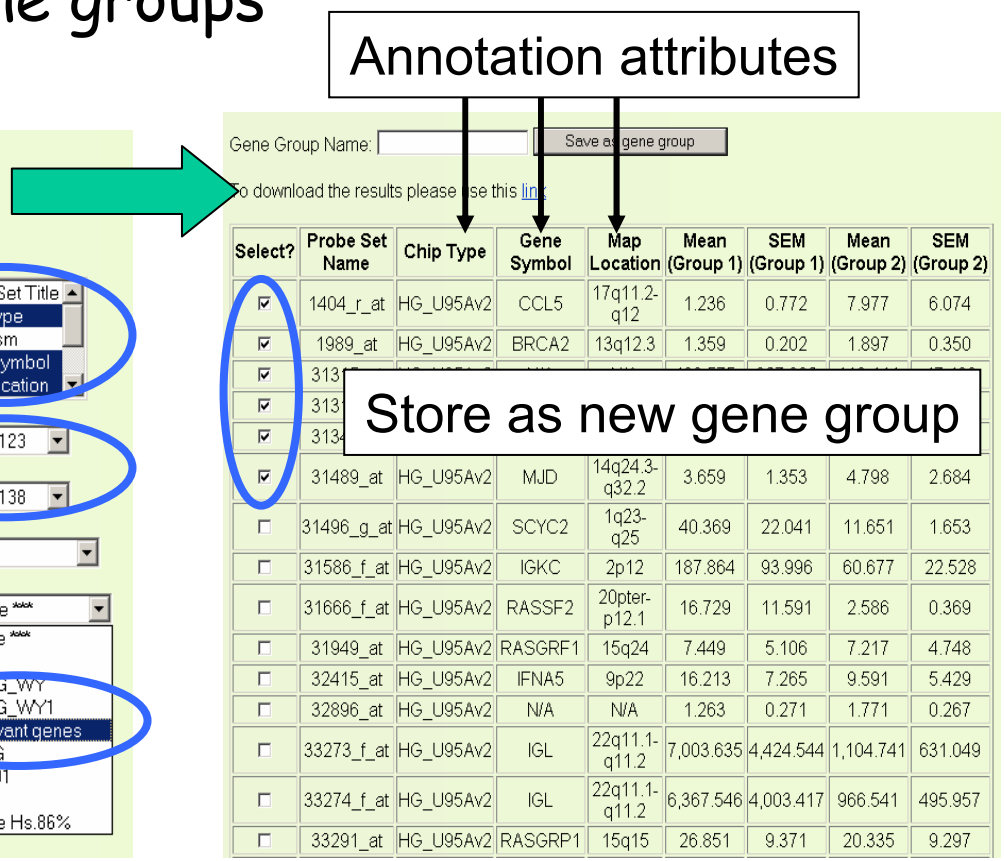
Query result storable as gene group



Expression Value Reporting and Filter

- Several statistical reports used for analysis entry and outlier detection
  - Using experiment and gene groups to filter
  - Generation of new gene groups
  - Downloadable results





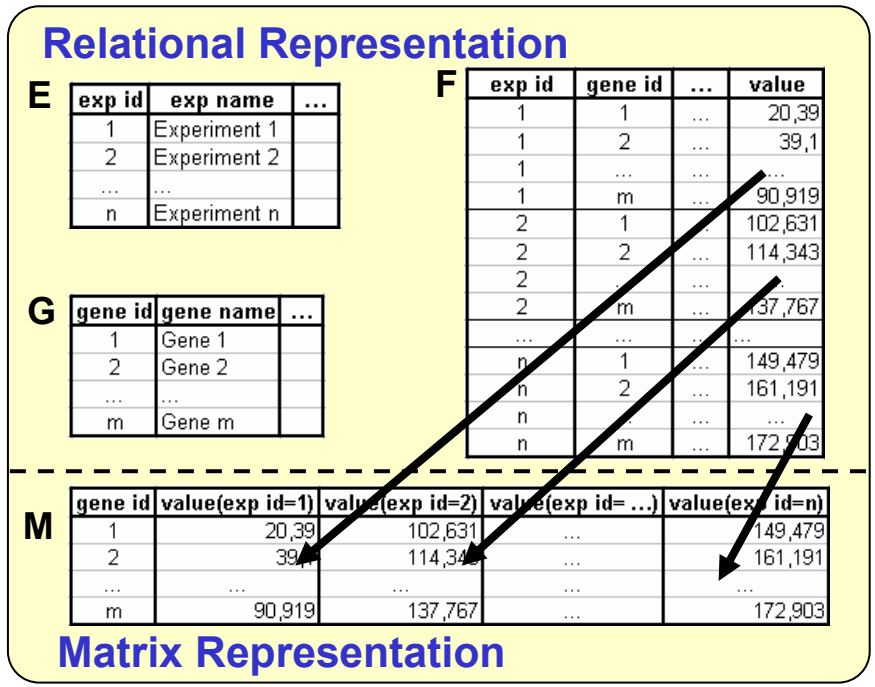
Select?	Probe Set Name	Chip Type	Gene Symbol	Map Location	Mean (Group 1)	SEM (Group 1)	Mean (Group 2)	SEM (Group 2)
☑	1404_r_at	HG_U95Av2	CCL5	17q11.2-q12	1.236	0.772	7.977	6.074
☑	1989_at	HG_U95Av2	BRCA2	13q12.3	1.359	0.202	1.897	0.350
☑	313							
☑	313							
☑	313							
☑	31489_at	HG_U95Av2	MJD	14q24.3-q32.2	3.659	1.353	4.798	2.684
☐	31496_g_at	HG_U95Av2	SCYC2	1q23-q25	40.369	22.041	11.651	1.653
☐	31586_f_at	HG_U95Av2	IGKC	2p12	187.864	93.996	60.677	22.528
☐	31666_f_at	HG_U95Av2	RASSF2	20pter-p12.1	16.729	11.591	2.586	0.369
☐	31949_at	HG_U95Av2	RASGRF1	15q24	7.449	5.106	7.217	4.748
☐	32415_at	HG_U95Av2	IFNA5	9p22	16.213	7.265	9.591	5.429
☐	32896_at	HG_U95Av2	N/A	N/A	1.263	0.271	1.771	0.267
☐	33273_f_at	HG_U95Av2	IGL	22q11.1-q11.2	7,003.635	4,424.544	1,104.741	631.049
☐	33274_f_at	HG_U95Av2	IGL	22q11.1-q11.2	6,367.546	4,003.417	966.541	495.957
☐	33291_at	HG_U95Av2	RASGRP1	15q15	26.851	9.371	20.335	9.297

Gene Expression Matrix Management (1)

- Gene expression matrix (GEM)
  - Genes as row, experiments as column label
  - "Standard" input format for many analysis tools
- Requirements
  - Support for different matrix types (absolute / relative values, nested, …)
  - Input for advanced analysis, reporting and export in GeWare
  - Problem: How to manage GEM in relational databases?
    - Complexity / size limitations of resulting SQL statements
    - Performance aspects

Gene Expression Matrix Management (2)

- Schema
  - G (gene id, gene name, ...)
  - E (exp id, exp name, ...)
  - F (gene id, exp id, value, ...)
  - M (gene id, value (exp id 1) ... value (exp id n))



Need a mapping: F → M

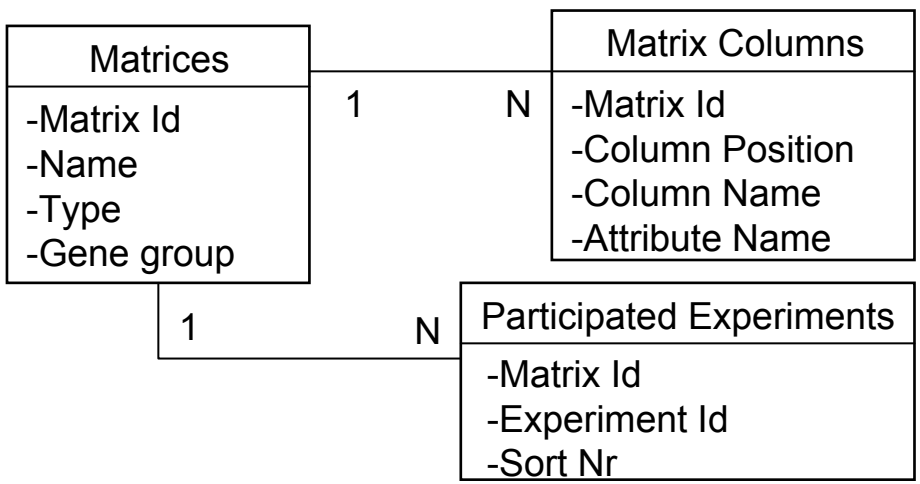
- Virtual mapping (view)
- Materialized mapping (mat. view, table)

Example: Virtual Mapping:

```
CREATE VIEW F_M_Mapping AS
SELECT    G.gene id, F1.value, F2.value …Fn.value
FROM      G, F as F1, F as F2 … F as Fn
WHERE     G.gene id = F1.gene id
AND       G.gene id = F2.gene id
AND       G.gene id = …
AND       G.gene id = Fn.gene id
AND       F1.exp id = 1
AND       F2.exp id = 2
AND       …
AND       Fn.exp id = n
```

Gene Expression Matrix Management (3)

- GEM management in GeWare
  - Materialized representation of GEM due to
    - Database limitations (query size)
    - Expected less performance using views
  - Flexible generation of different GEM types
  - Application of first class objects and high level operations, e.g.
    - generateMatrix (Experiment Group, Gene Group)
    - generateMatrix (Experiment Pairs, Gene Group)
  - Matrix visualization
  - Generic GEM metadata management



Analysis Coupling

- Tight integration
  - Various predefined canned queries for analysis entry and outlier detection
  - Concentration ratio (Lorenz curve, Gini-Coefficient)
  - Sequence specific database functions (UDF)
- Transparent integration (database API)
  - Oligo sequence sensitivity analysis
  - OLAP
- File-based exchange
  - Application of R / BioConductor for
    - Intensity transformations (MAS5, RMA, LiWong R/F)
    - Advanced analysis (Westphal/Young univariate beta test with resampling strategy, …)
  - Import of analysis results for further analysis

Conclusions / Future Work

- GeWare
  - Management of a high volume of expression data
  - Flexible experiment annotation
  - Storing experiment and gene groups
  - Management of different types of expression matrices
  - Different kinds of analysis, export
- Future work
  - Coupling with advanced analysis/ data mining routines
  - Visualization extension

Special Thanks

- Database group / IZBI
  - Hans Binder
  - Martin Beck
  - Guido Fritzsch
- IZKF/Medical Dept. University of Leipzig
  - Friedemann Horn
  - Knut Krohn
  - Markus Eszlinger
- MPI for Evolutionary Anthropology
  - Philipp Khaitovich
  - Wolfgang "Wolfi" Enard
  - Björn Mützel
  - Svante Pääbo