【S-3】

Development of Toxicogenomics Knowledge Base by Integrating High-throughput Genomic Data

Ju Han Kim, M.D., Ph.D., M.S.

Seoul National University Biomedical Informatics (SNUBI)

Bioinformatics is a rapidly emerging field of biomedical research. A flood of large-scale genomic and postgenomic data means that many of the challenges in biomedical research are now challenges in computational sciences. Postgenome informatics, powered by high throughput technologies and genomic-scale databases, is likely to transform our biomedical understanding forever much the same way that biochemistry did a generation ago. In this talk, I will describe how these technologies will impact toxicology research, introducing recent advances in databasing gene expression profiles with the emphasis on the necessity of tight integration of private and public databases and intelligent analysis toolkits. I will introduce some of our research efforts for toxicogenomics knowledge base. Xperanto (Expressionist's Esperanto in XML) integrates major data models for DNA microarray, tissue microarray and array CGH data with extended clinical and histo-pathological information models and supports analysis tools in an effort to establish a comprehensive knowledge base for toxicogenomics research. Each step will be given with real examples from ongoing research activities in the context of clinical relevance.

Development of Toxicogenomics Knowledgebase by Integrating High-throughput Genomic Technologies

Pharmaceutical side effects and Drug-drug interactions

- One in five new drugs have some serious side effects (by Lasser KE, *et al.* JAMA 2002 287:2215)
- Need ways to predict potential interactions and side effects
  - impact of environmental elements that have the potential for biological damage
  - the relationship of the activities of toxins to genetic makeup
- Previous tools for measuring toxic events
  - in vivo animal models such as rat, mouse, and dog
  - in vitro assays
    - Ames testing
    - micronucleus assays
    - unscheduled DNA synthesis measurement

2

New way to measure toxic events through gene expression profiles

- Human Genome Project + Microarray technology
  - interrogate the expression of tens of thousands of genes simultaneously in response to a toxic agent
  - investigate effects of human genetic variation on drug toxicity and efficacy

3

cDNA microarray schema

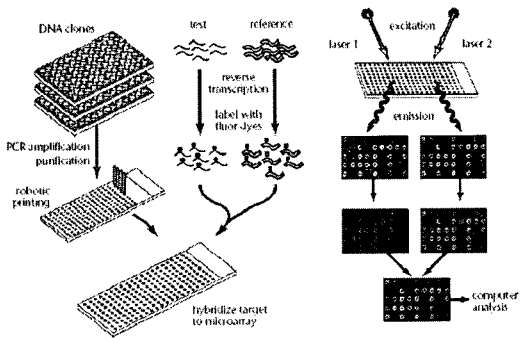


Figure from *Nature Genetics* (1999) 21:11

4

Gene expression profiling of chemotherapeutic drugs

Cluster analysis of temporal changes in gene expression of MCF-7 cells treated with doxorubicin

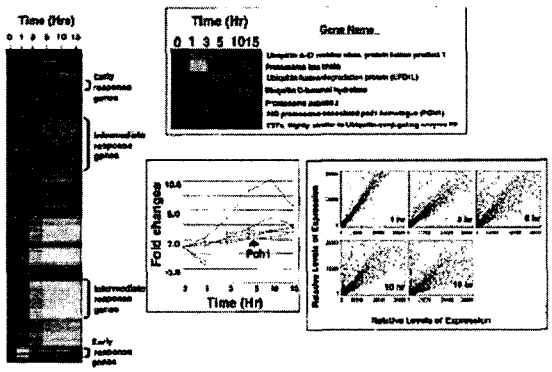


Figure from *Pharmaceutical Research* (2002) 19:1776

5

DNA arrays for toxicogenomics

DNA arrays allow quantitative measurements of gene expression to be made for tens of thousands of genes in parallel

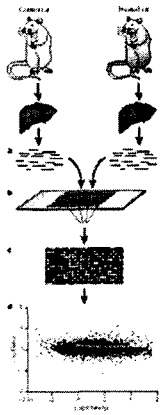


Figure from *Nature Reviews* (2001) 1:84

6

Gene expression profiling of toxic compounds

A compendium showing gene-expression change induced by 48 different hepatotoxic compounds in the rats

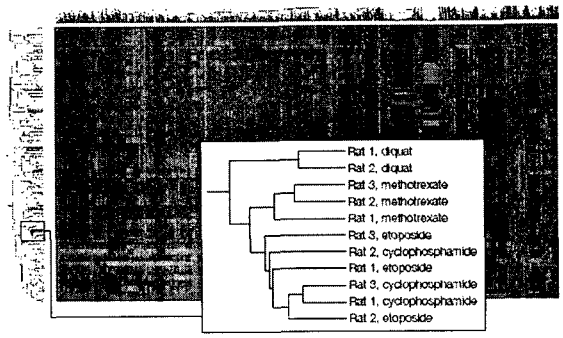


Figure from *Nature Reviews* (2001) 1:84

7

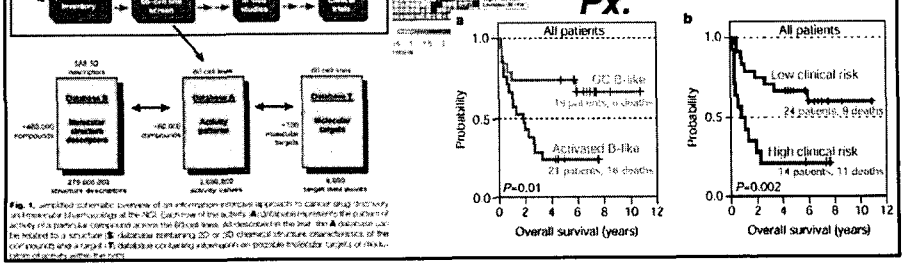
Clinical relevance of Biochip informatics

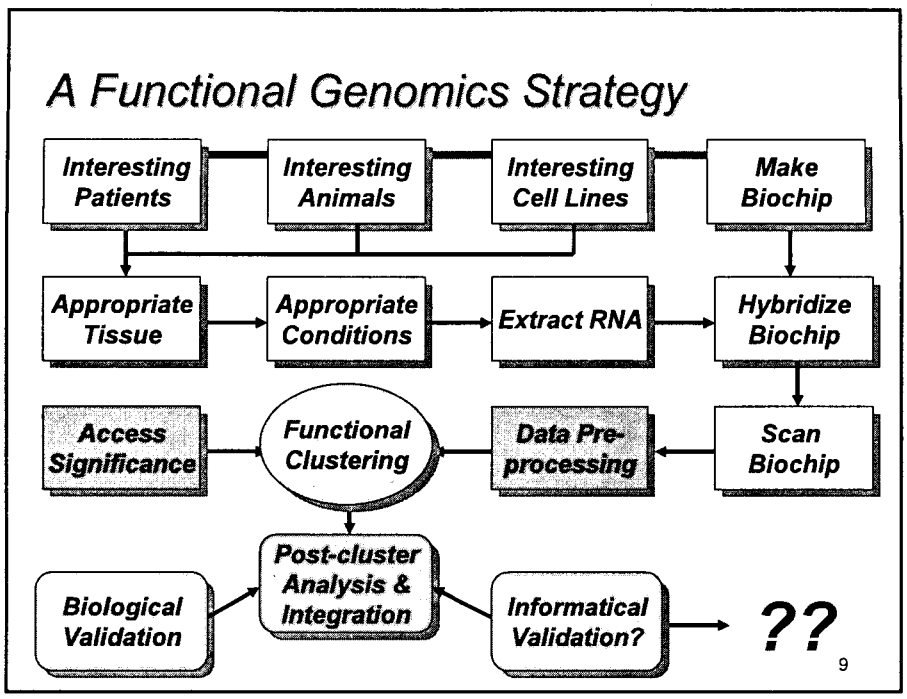
Dx.

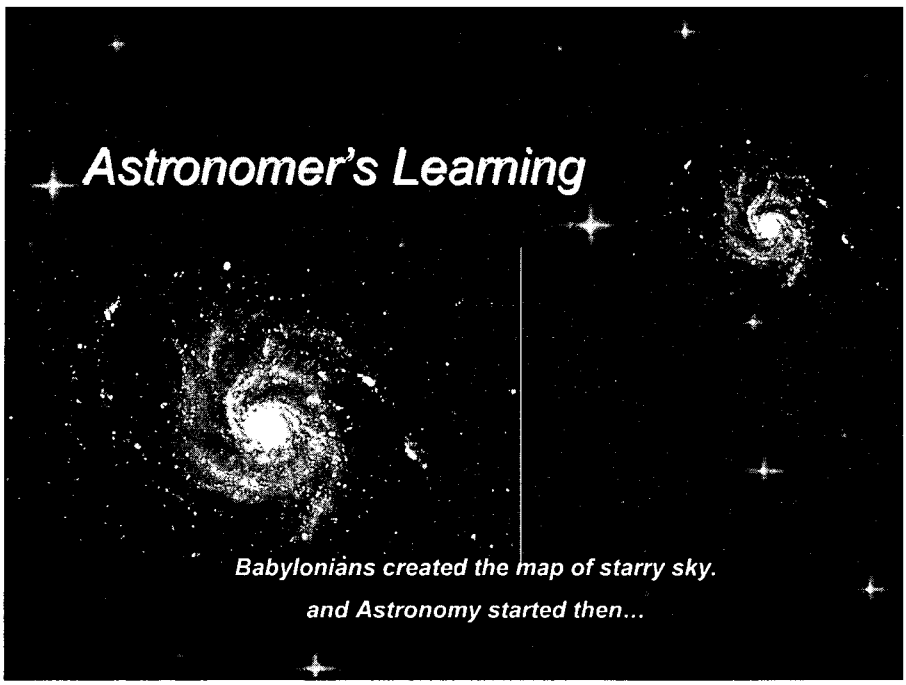
Discovery

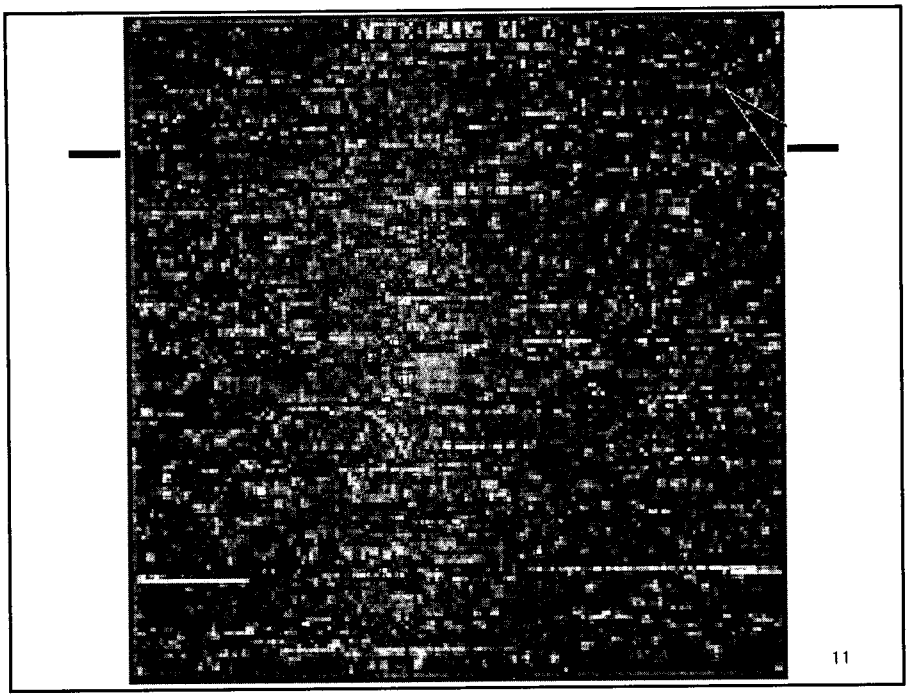
Tx.

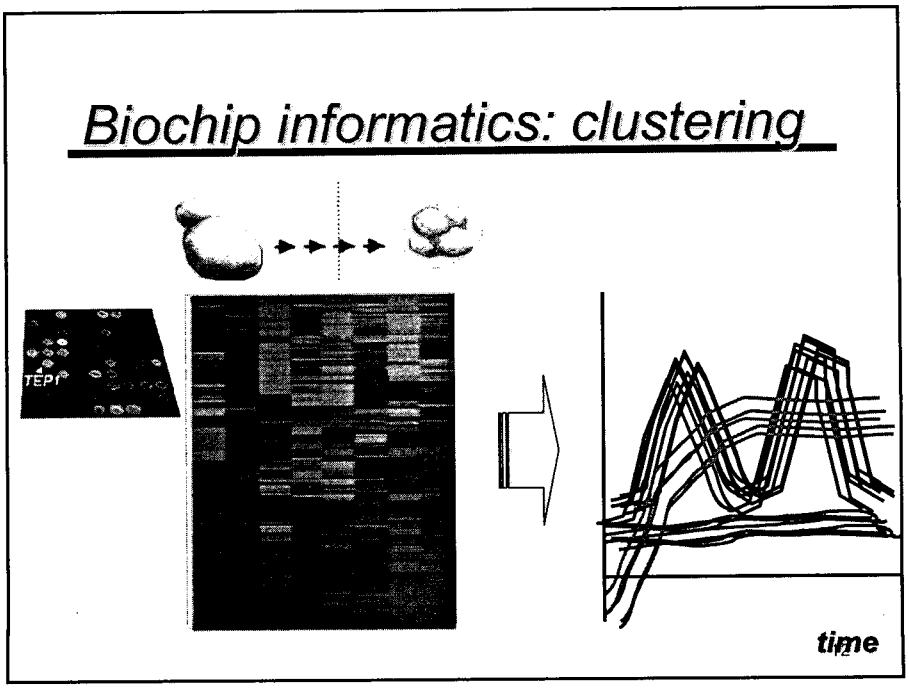
Px.

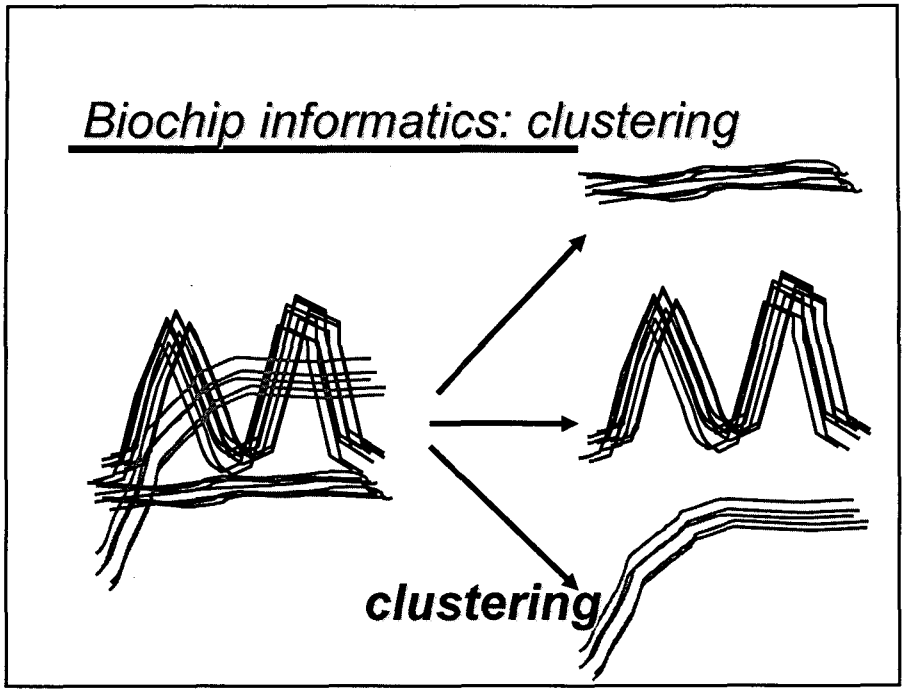


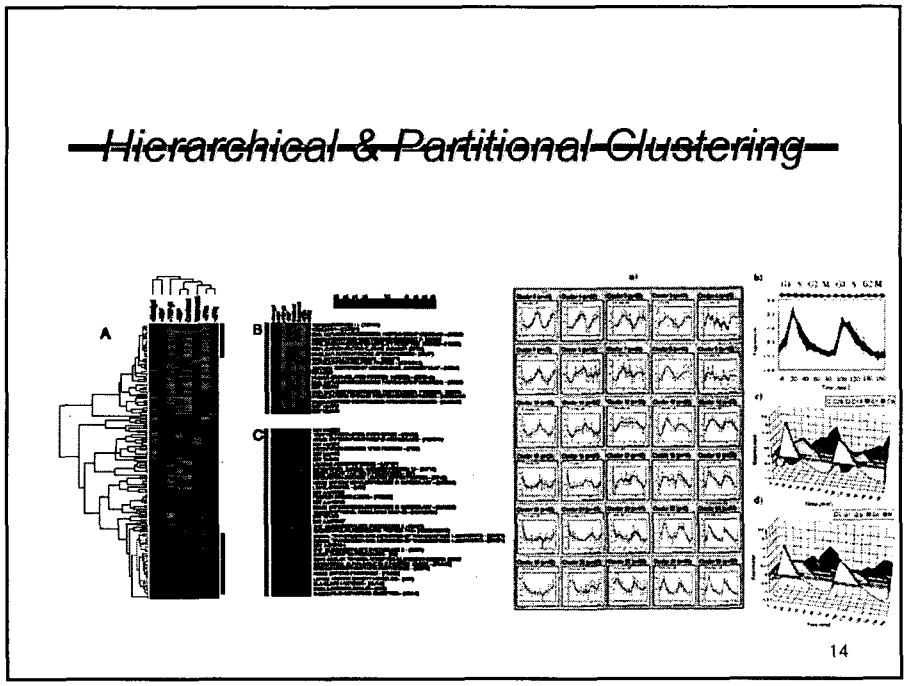

A Functional Genomics Strategy
Interesting Patients
Interesting Animals
Interesting Cell Lines
Make Biochip
Appropriate Tissue
Appropriate Conditions
Extract RNA
Hybridize Biochip
Access Significance
Functional Clustering
Data Pre-processing
Scan Biochip
Post-cluster Analysis & Integration
Biological Validation
Informatical Validation?
??
9


Astronomer's Learning
Babylonians created the map of starry sky.
and Astronomy started then...

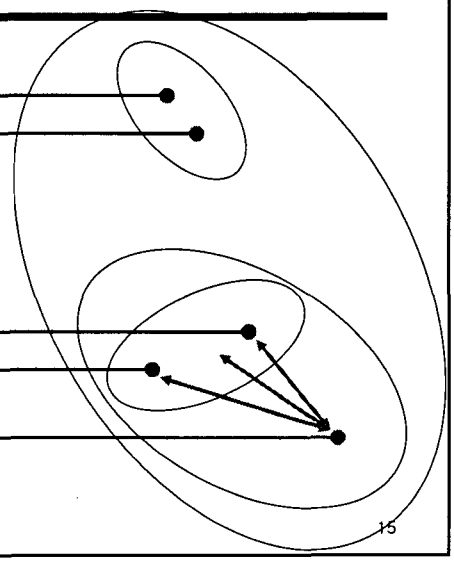

11


Biochip informatics: clustering
TEPI
time


Biochip informatics: clustering
clustering


Hierarchical & Partitional Clustering
A
B
C
14

Hierarchical clustering in Genomics

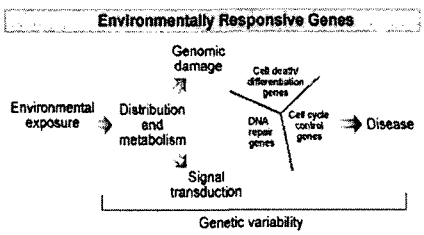


- *single-linkage (nearest neighbor)*
- *complete-linkage (farthest neighbor)*
- *weighed pair-group average*
- *unweighed pair-group average*
- *weighted pair-group centroid*
- *unweighted pair-group centroid*
- *Ward's method: min. sum of squares*

15

Toxicogenomics

- the application of genomic tools to the study of biological responses to toxic substances



16

nature

19 April 2001 Volume 410 Issue no 6831

Free and public expression

After a slow start, progress towards developing public repositories for gene expression data is poised to accelerate. For the many biologists working with DNA microarrays, that should be welcome news.

With a single format for gene expression data, databases should be able to 'talk' to one another and exchange data. The existence of a standard language should also spur development of software tools to query the databases, and to manage and display gene expression data.

17

nature

25 September 2002 Volume 419 Issue no 6905

Microarray standards at last

Not a moment too soon, the microarray community has issued guidelines that will make their data much more useful and accessible. *Nature* and the Nature research journals will respond accordingly.

You read a paper with a fascinating conclusion about the expression of several genes. You decide to use some of the same experiments on your system of choice. But when you wade through hundreds of pages of supplementary information, you find that crucial details needed for replication are missing.

Welcome to the exciting but frustrating world of DNA microarray research. Microarrays are plastic or glass chips spotted with tiny amounts of thousands of probes, used to query the activity levels of that many genes in any tissue or organism at one time. Variables in every step of the experiment often make cross-paper comparison virtually impossible. Microarray papers also pose a considerable strain on the refereeing process: the vast amounts of data mean that critical review is a monumental task.

Yet referees sometimes feel they are not given enough details, leading cautious reviewers to think that they must reanalyse the primary data set. In other cases, the primary data provided are in proprietary software and so are impossible to comment on. Many journals allowed authors to put the huge data files on their own websites for the review process, until it became clear that unscrupulous authors compromised the anonymity of referees by tracking who had visited the website.

For authors, the proposal provides a checklist of variables that should be included in every microarray publication, at http://www.mged.org/Workgroups/MIAME/miame_checklist.html. This checklist, with all variables completed, would be supplied as supplementary information at the time of submission. The MGED group suggests that journals require submission of microarray data to either of two databases emerging as the main public repositories: GEO (www.ncbi.nlm.nih.gov/geo/) or ArrayExpress (www.ebi.ac.uk/arrayexpress).

Harried editors can rejoice that, at last, the community is taming the unruly beast that is microarray information. Therefore all submissions to *Nature* and the Nature family of journals received on or after 1 December containing new microarray experiments must include the mailing of five compact disks to the editor. These disks should include necessary information compliant with the MIAME standard. The information must be supplied in a format that could be read by widely available software packages. Data integral to the paper's conclusions should be submitted to the ArrayExpress or GEO databases, with accession numbers where available, supplied at or before acceptance for publication.

How much data should authors provide to the community?

Microarray standards are needed to facilitate moving data around

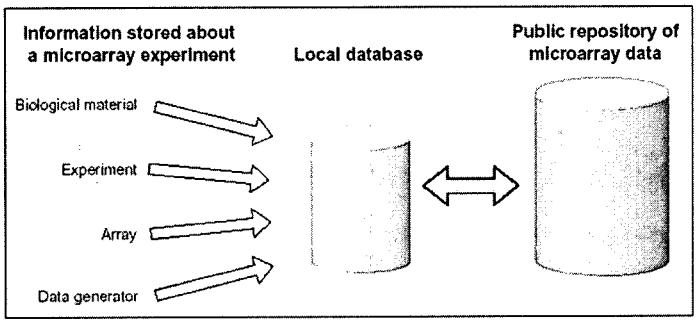
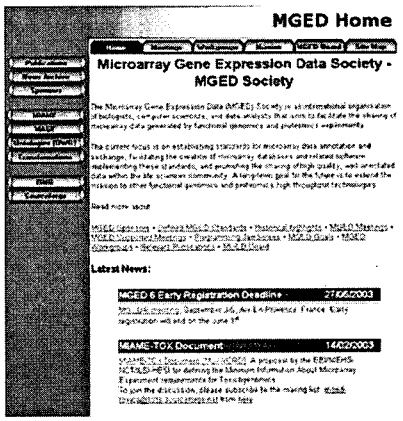


Figure from *Nature Genetics* (2002) 32: 460

19

MGED standards



http://www.mged.org

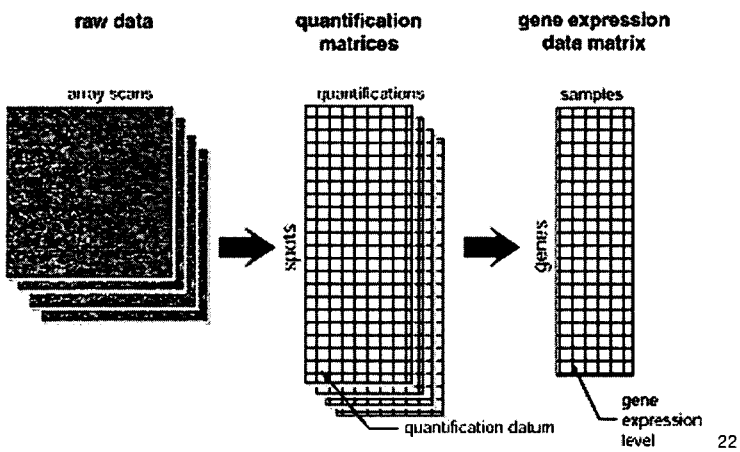
20

Microarray Standards

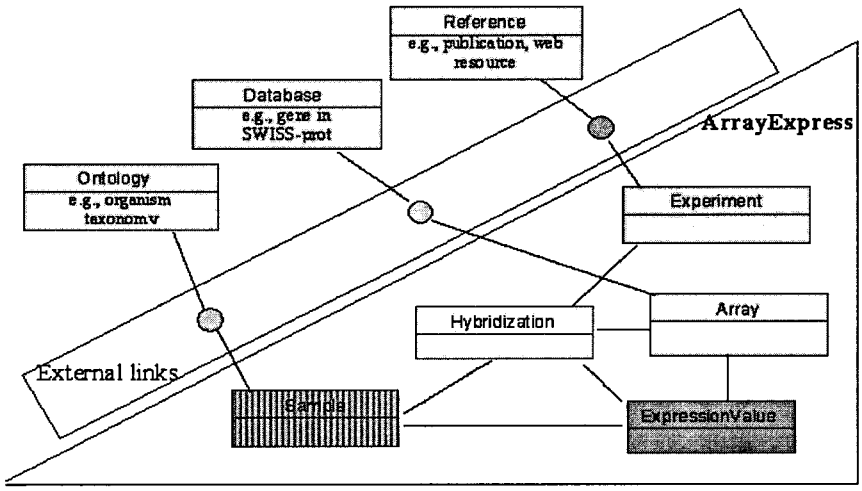
- **MIAME**
  - *Minimum Information About a Microarray Experiment*
  - *Experimental Design, Array Design, Hybridization, Samples, Measurements and Normalization*
- **MAGE-ML**
  - *XML Implementation of the MIAME Standard*
  - *De Facto Widespread Industry Support*
- **MAGE-OM**
  - *Object Model as a framework for developing MAGE*
  - *OMG specifications are developed in UML*

21

Three levels of microarray gene expression data processing



22



23

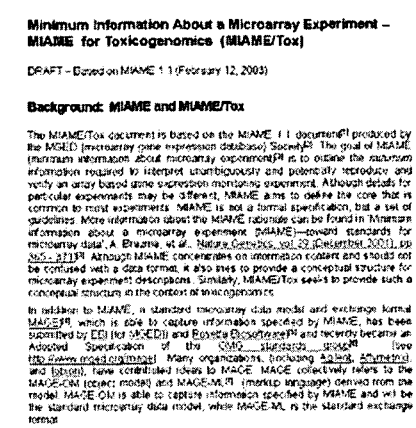
Goals of MGED Efforts

- Elucidate information needed for experiments
  - MIAME
- Provide a means to share this information
  - MAGE
- Provide a common language for experiments
  - MGED Ontology
- Provide standard operating procedures for analysis
  - Data transformation

24

Standards will aid data integration in toxicogenomics efforts

HESI, NIEHS, FDA, EBI, academics
- new MIAME-Tox



Minimum Information About a Microarray Experiment – MIAME for Toxicogenomics (MIAME/Tox)

DRAFT - Based on MIAME 1.1 (February 12, 2003)

Background: MIAME and MIAME/Tox

The MIAME/Tox document is based on the MIAME 1.1 document[1] produced by the MGED (microarray gene expression database) Society[2]. The goal of MIAME (minimum information about microarray experiment)[3] is to outline the minimum information required to interpret unambiguously and potentially reproduce and verify an array based gene expression monitoring experiment. Although details for particular experiments may be different, MIAME aims to define the core that is common to most experiments. MIAME is not a formal specification, but a set of guidelines. More information about the MIAME rationale can be found in 'Minimum information about a microarray experiment (MIAME)—toward standards for microarray data', A. Brazma, et al., Nature Genetics, vol. 29 (December 2001), pp. 365 - 371[4]. Although MIAME concentrates on information content and should not be confused with a data format, it also tries to provide a conceptual structure for microarray experiment descriptions. Similarly, MIAME/Tox seeks to provide such a conceptual structure in the context of toxicogenomics.

In addition to MIAME, a standard microarray data model and exchange format MAGE[5], which is able to capture information specified by MIAME, has been submitted by EBI (for MGED)) and Rosetta Biosoftware[6] and recently became an Adopted Specification of the OMG standards group[7] (see http://www.mged.org/magel). Many organizations, including Agilent, Affymetrix and Iobion, have contributed ideas to MAGE. MAGE (collectively refers to the MAGE-OM (object model) and MAGE-ML (markup language) derived from the model. MAGE-OM is able to capture information specified by MIAME and will be the standard microarray data model, while MAGE-ML is the standard exchange format.

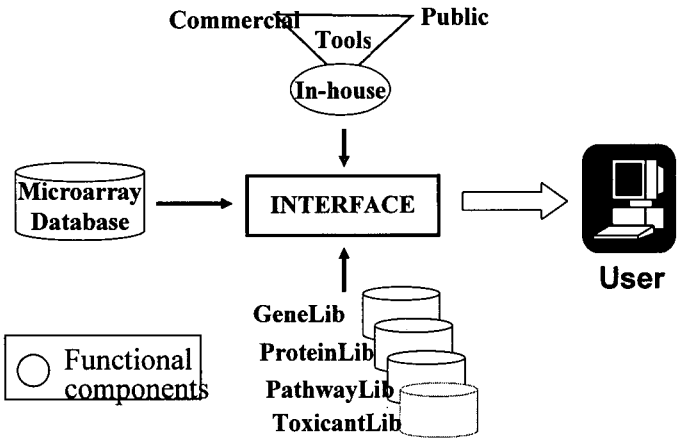
25

Minimum information to be recorded about toxicogenomics experiments

- Experimental design parameters, animal husbandry information or cell line and culture information, exposure parameters, dosing regimen, dose groups, and in-life observations.
- Microarray data, specifying the number and details of replicate array bioassays associated with particular samples, and including PCR transcript analysis if available.
- Numerical biological endpoint data, including necropsy weights or cell counts and doubling times, clinical chemistry and enzyme assays, hematology, urinalysis, other.
- Textual endpoint information such as gross observations, pathology and microscopy findings.

26

ArrayTrack



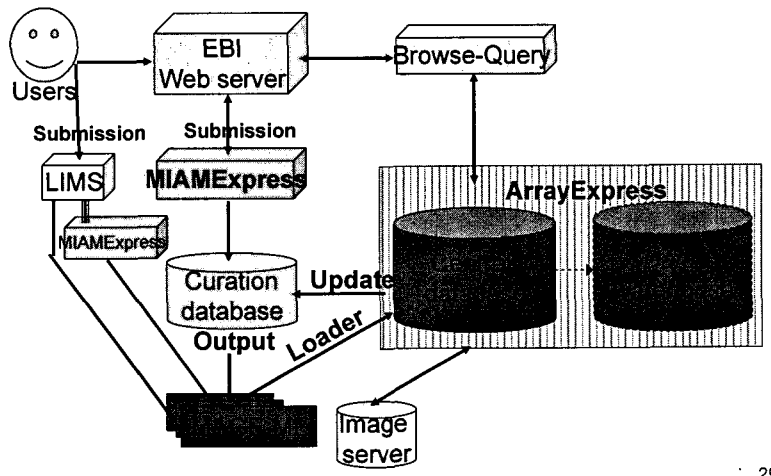
27

dbZach

- Platform independent
- Minimum Information About Microarray Experiment (MIAME) Supportive
- Database (Oracle 9*i*)
  - Data Modeling
  - Modular Subsystem Architecture
  - Data Management
- Tools
  - Graphical
  - User-friendly
  - Java 2
  - Data Mining
  - Data Visualization

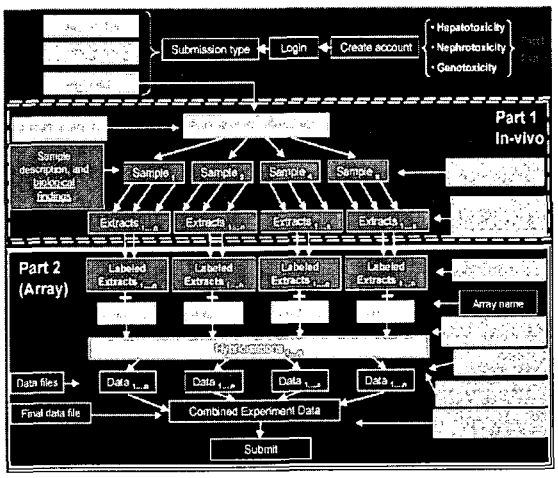
28

ArrayExpress – data flow



29

Toxicogenomics specific MIAMExpress (ILSI- HESI) - in development



30

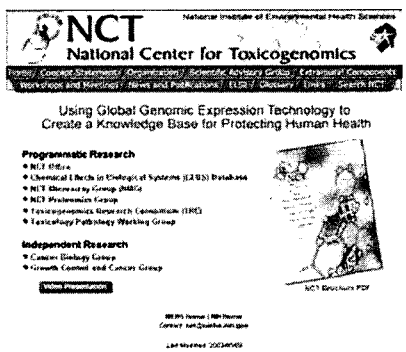
toxicogenomics project by ILSI

- International Life Sciences Institute (ILSI/HESI) toxicogenomics database
- cross-platform gene expression data on the effects of various toxic compounds

31

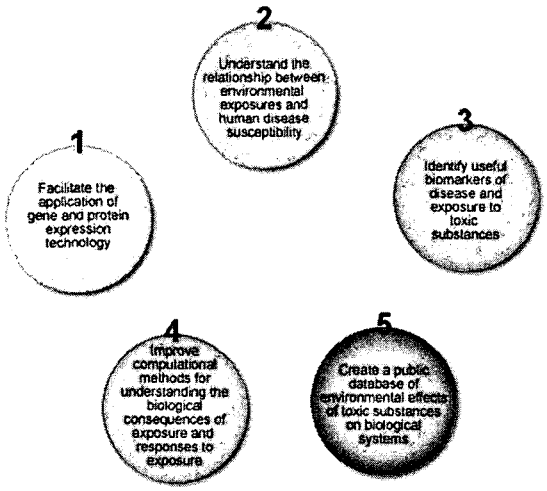
Government efforts: National Center for Toxicogenomics

- Goal
  - support research in the field of toxicogenomics
  - compile, analyze, and publish the resulting data
  - identify genes that are regulated in response to toxicans
  - develop a "Toxchip" to monitor changes



32

Goals of National Center for Toxicogenomics



33

Toxicogenomics Infrastructure

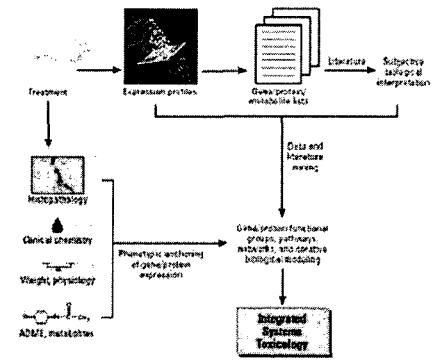
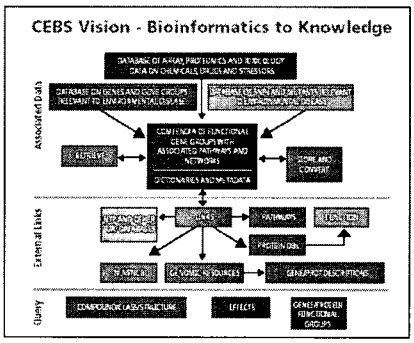


Figure from *EHP Toxicogenomics* (2003) 111:15

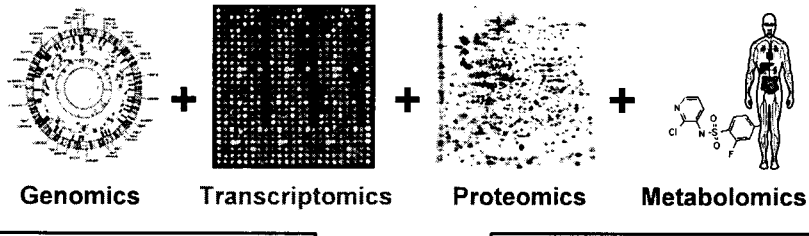
34

Chemical Effects in Biological Systems (CEBS)



35

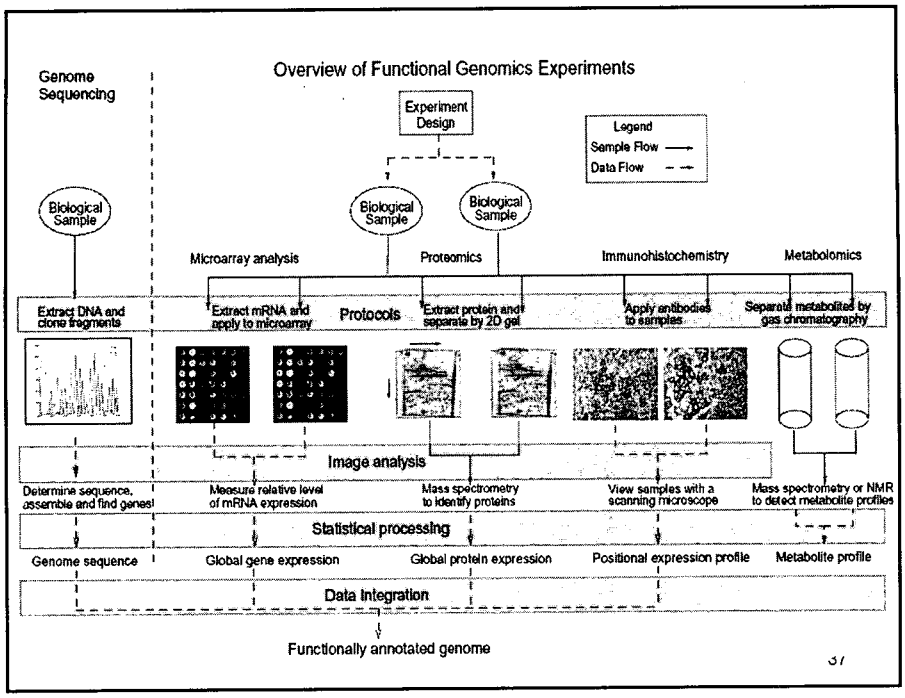
Proteomics in the "-omics" world.....

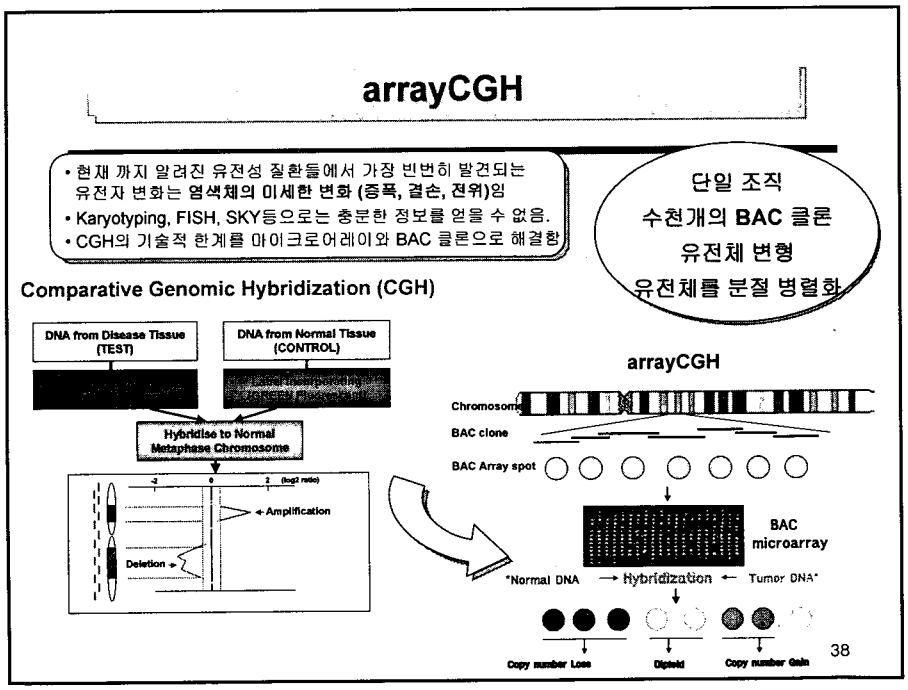


Biological Knowledge

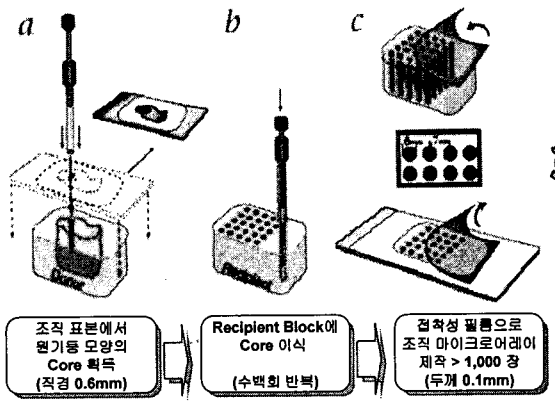
Disease Genes, Drug Targets, Mechanisms for Disease Causation

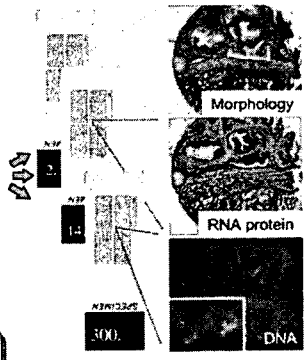
36


Overview of Functional Genomics Experiments
Genome Sequencing
Experiment Design
Legend
Sample Flow
Data Flow
Biological Sample
Biological Sample
Biological Sample
Microarray analysis
Proteomics
Immunohistochemistry
Metabolomics
Extract DNA and clone fragments
Extract mRNA and apply to microarray
Protocols
Extract protein and separate by 2D gel
Apply antibodies to samples
Separate metabolites by gas chromatography
Image analysis
Determine sequence, assemble and find genes!
Measure relative level of mRNA expression
Mass spectrometry to identify proteins
View samples with a scanning microscope
Mass spectrometry or NMR to detect metabolite profiles
Statistical processing
Genome sequence
Global gene expression
Global protein expression
Positional expression profile
Metabolite profile
Data Integration
Functionally annotated genome
37


arrayCGH
• 현재 까지 알려진 유전성 질환들에서 가장 빈번히 발견되는 유전자 변화는 염색체의 미세한 변화 (증폭, 결손, 전위)임
• Karyotyping, FISH, SKY등으로는 충분한 정보를 얻을 수 없음.
• CGH의 기술적 한계를 마이크로어레이와 BAC 클론으로 해결함
단일 조직
수천개의 BAC 클론
유전체 변형
유전체를 분절 병렬화
Comparative Genomic Hybridization (CGH)
DNA from Disease Tissue (TEST)
DNA from Normal Tissue (CONTROL)
Hybridise to Normal Metaphase Chromosome
Amplification
Deletion
arrayCGH
Chromosome
BAC clone
BAC Array spot
BAC microarray
*Normal DNA
Hybridization
Tumor DNA*
Copy number Loss
Diploid
Copy number Gain
38

Tissue microarray





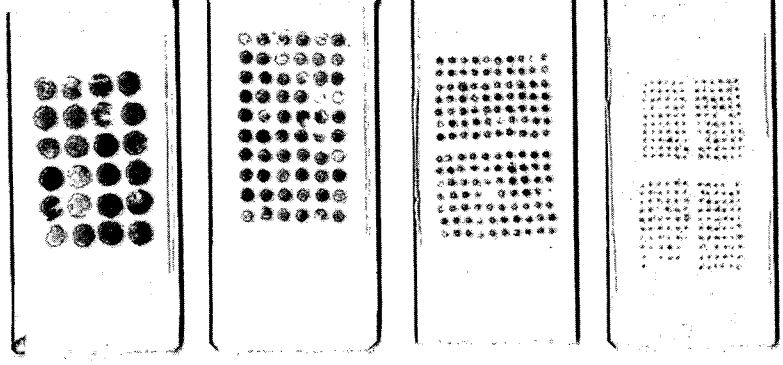
조직 표본에서 원기둥 모양의 Core 획득 (직경 0.6mm) → Recipient Block에 Core 이식 (수백회 반복) → 접착성 필름으로 조직 마이크로어레이 제작 > 1,000 장 (두께 0.1mm)

Target	Probe	방법
단백질	항체	면역조직화학법
DNA	DNA, fluorescent labeled	FISH
RNA	RNA, biotin labeled	in situ hybridization
Virus	primers	in situ PCR
DNA	primers	methylation specific in situ PCR

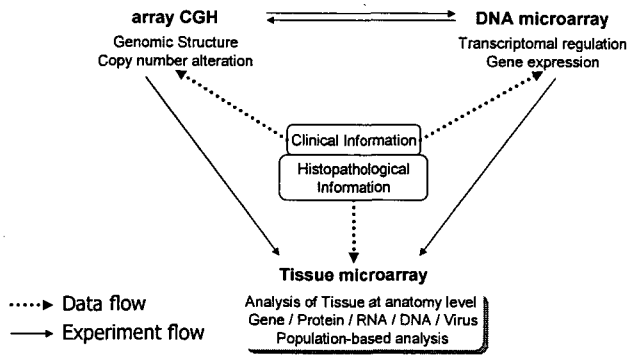
복수 조직
조직정보 포함
단일 검출 (다양한)
병리조직을 병렬화

Tissue microarray

4.0 mm	2.0 mm	1.0 mm	0.6 mm
24 specimens	60 specimens	140 specimens	240 specimens
SuperBioChips	*SuperBioChips*	*SuperBioChips*	*SuperBioChips*



Toxicogenomics research using high throughput technologies



41

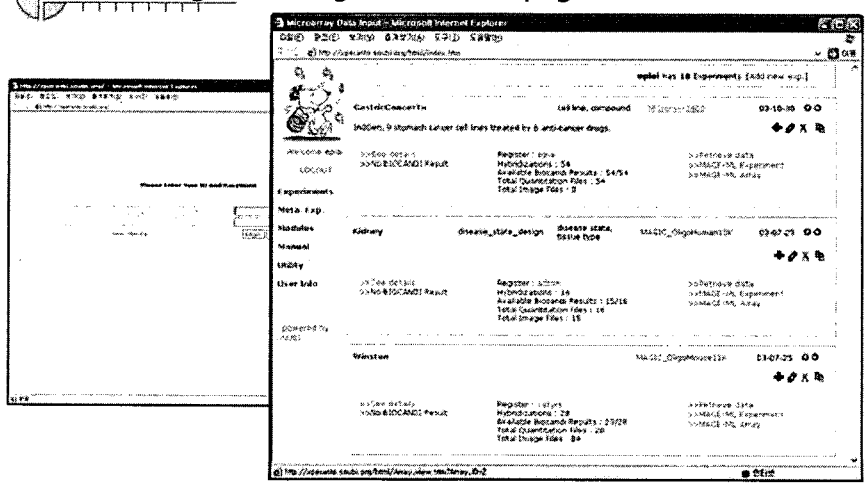
Xperanto: *Expressionist's Esperanto in XML*

- **MGED MIAME, MAGE-OM Standards**
- **Linux, MySQL, XML : open source**
- **Oligo and cDNA microarray**
- **Histopathology and Laboratory Data Model**
- **Batch upload and error check**
- **Diagnostic plots**
- **Visualization : genome browser, pathway browser)**
- **Automated gene / protein annotation**
- **Multi-user and group Environment**

Park et al. Genomics and Informatics 2005

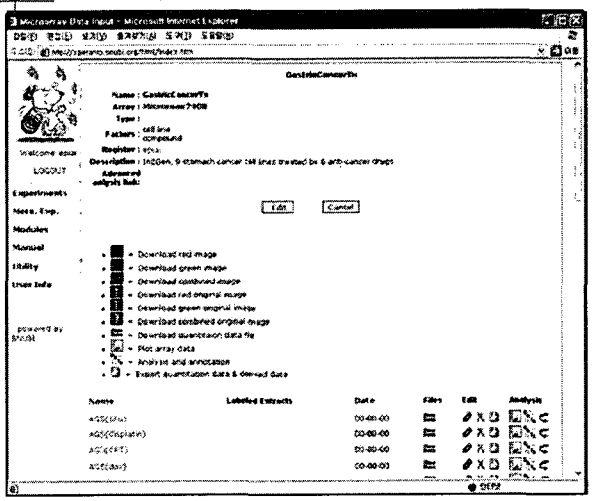
Core module

Data management: log-in and main page



Core module

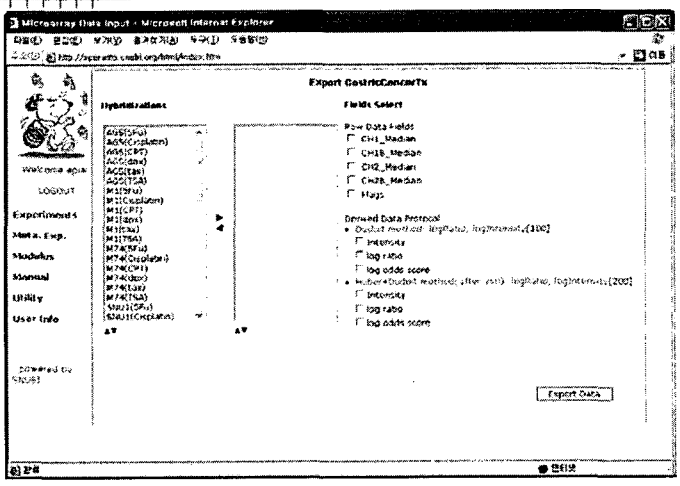
Data management: experiment



45

Core module

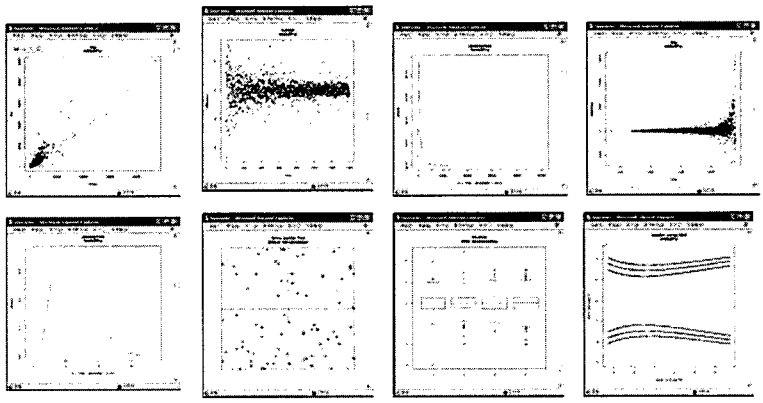
Data management: data export



46

Core module

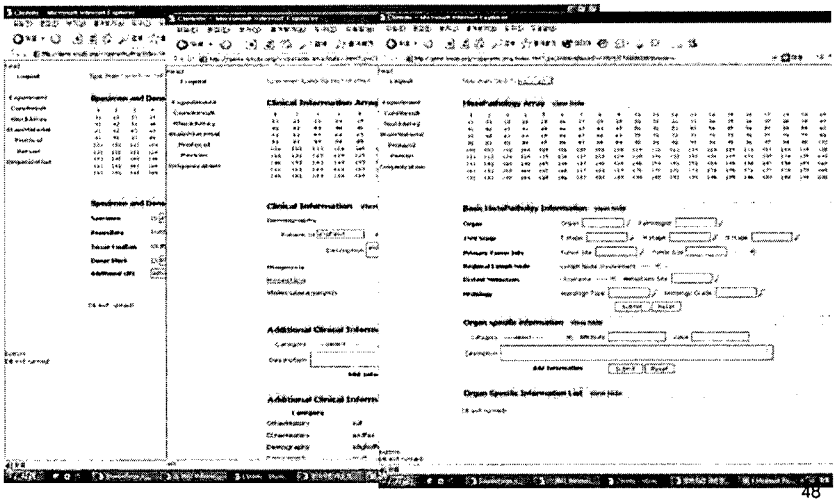
Data Analysis: plots in normalization process



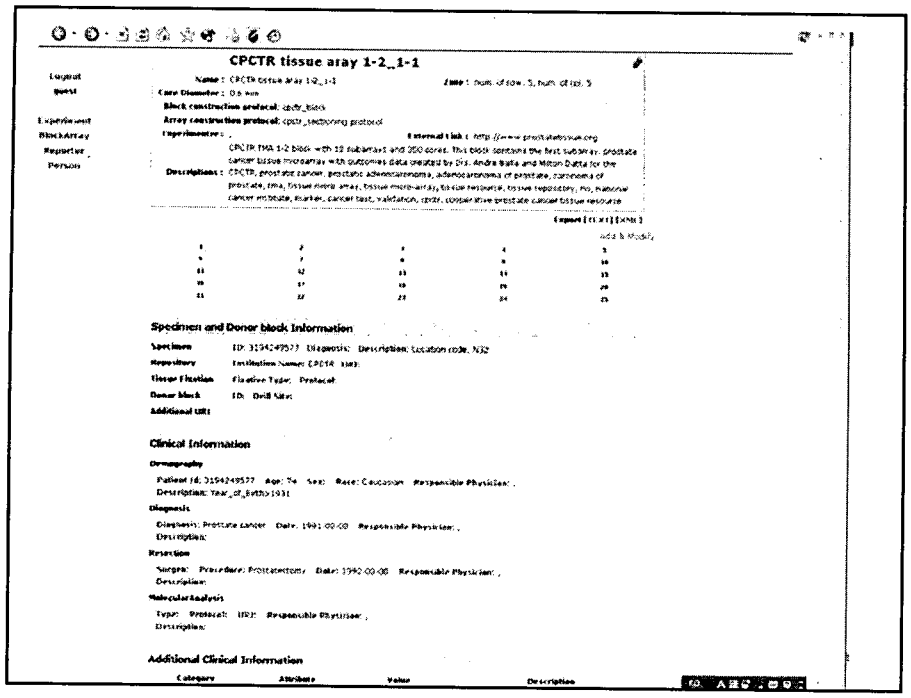
47

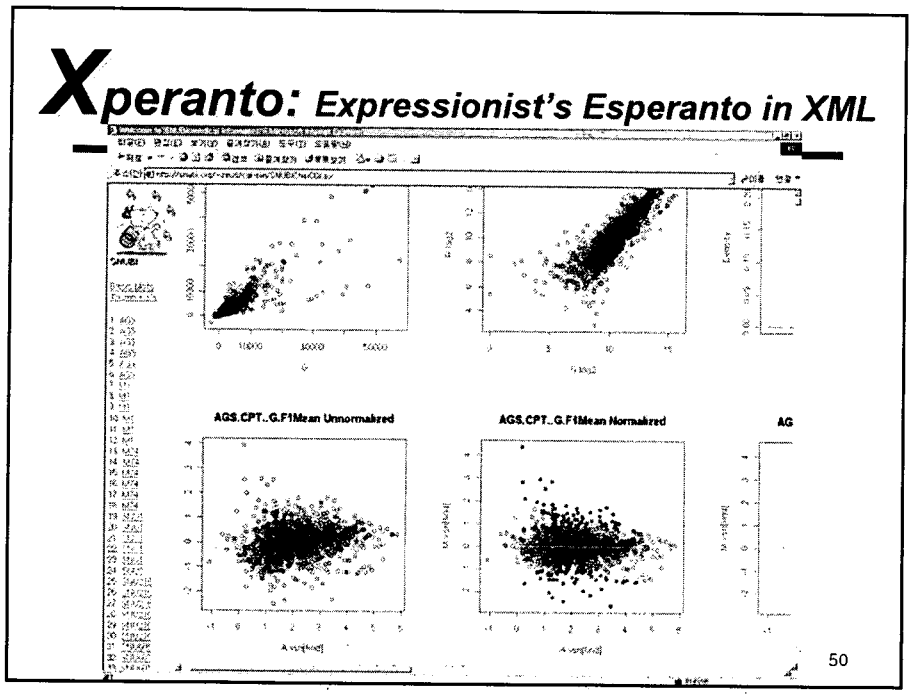
Web interface

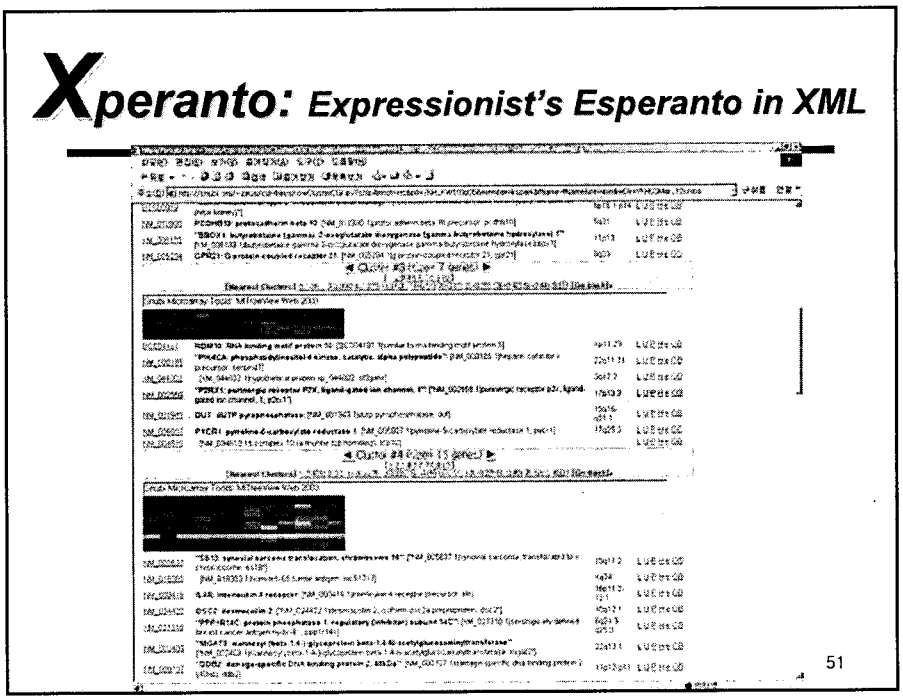
Submit Data

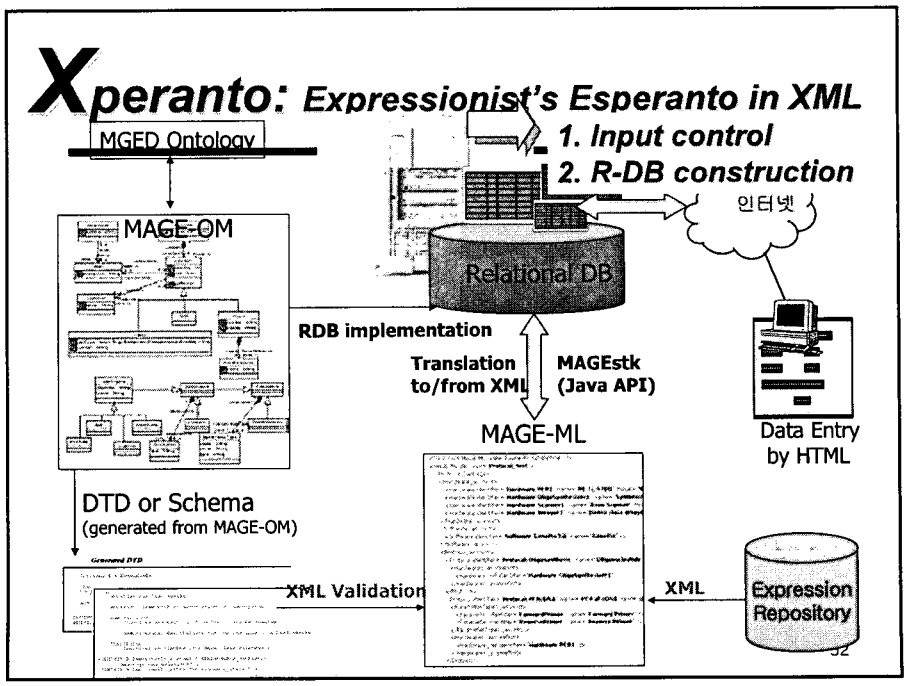


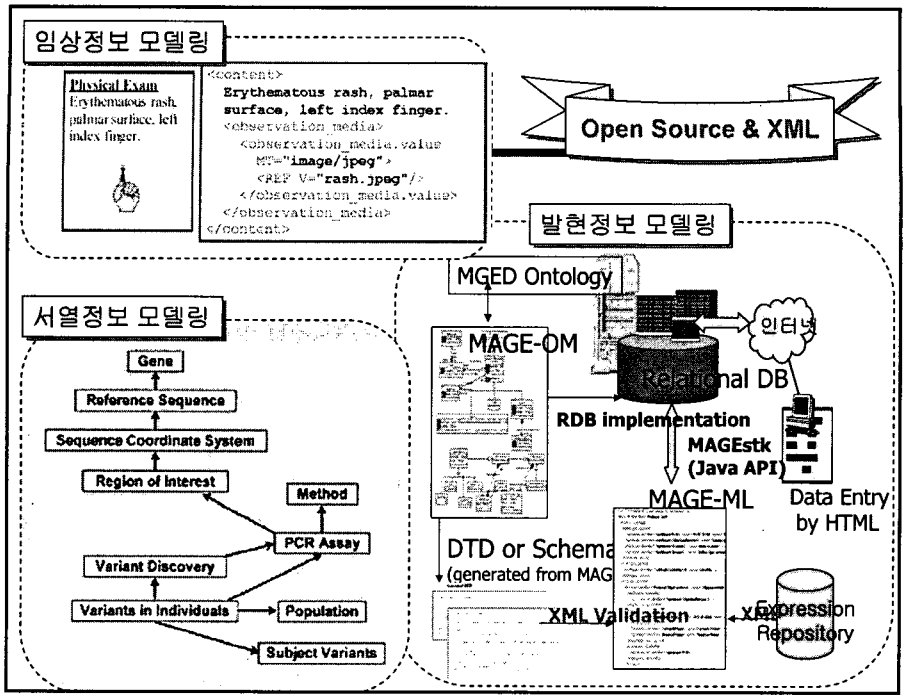
48

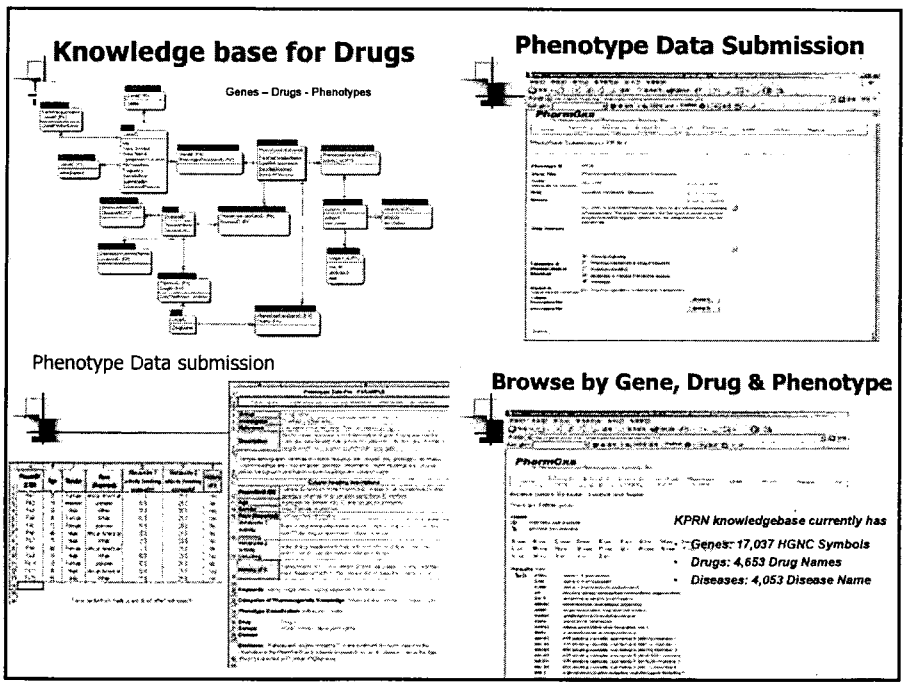

CPCTR tissue array 1-2_1-1
Specimen and Donor block Information
Clinical Information
Additional Clinical Information


Xperanto: Expressionist's Esperanto in XML
AGS.CPT..G.F1Mean Unnormalized
AGS.CPT..G.F1Mean Normalized
AG
50


Xperanto: Expressionist's Esperanto in XML
51


Xperanto: Expressionist's Esperanto in XML
MGED Ontology
1. Input control
2. R-DB construction
인터넷
MAGE-OM
Relational DB
RDB implementation
Translation to/from XML
MAGEstk (Java API)
MAGE-ML
Data Entry by HTML
DTD or Schema
(generated from MAGE-OM)
XML Validation
XML
Expression Repository
52


임상정보 모델링
Physical Exam
Erythematous rash, palmar surface, left index finger.
<content>
Erythematous rash, palmar surface, left index finger.
<observation_media>
<observation_media.value MT="image/jpeg">
<REF V="rash.jpeg"/>
</observation_media.value>
</observation_media>
</content>
Open Source & XML
발현정보 모델링
MGED Ontology
MAGE-OM
인터넷
Relational DB
RDB implementation
MAGEstk
(Java API)
MAGE-ML
Data Entry
by HTML
DTD or Schema
(generated from MAG
XML Validation
Expression
Repository
서열정보 모델링
Gene
Reference Sequence
Sequence Coordinate System
Region of Interest
Method
PCR Assay
Variant Discovery
Variants in Individuals
Population
Subject Variants


Knowledge base for Drugs
Genes – Drugs - Phenotypes
Phenotype Data Submission
Phenotype Data submission
Browse by Gene, Drug & Phenotype
KPRN knowledgebase currently has
Genes: 17,037 HGNC Symbols
Drugs: 4,653 Drug Names
Diseases: 4,053 Disease Name


Thank you!
http://www.snubi.org/