Integrated Microarray Database System

NHLBI-MGH-PGA

Desired Features for Database

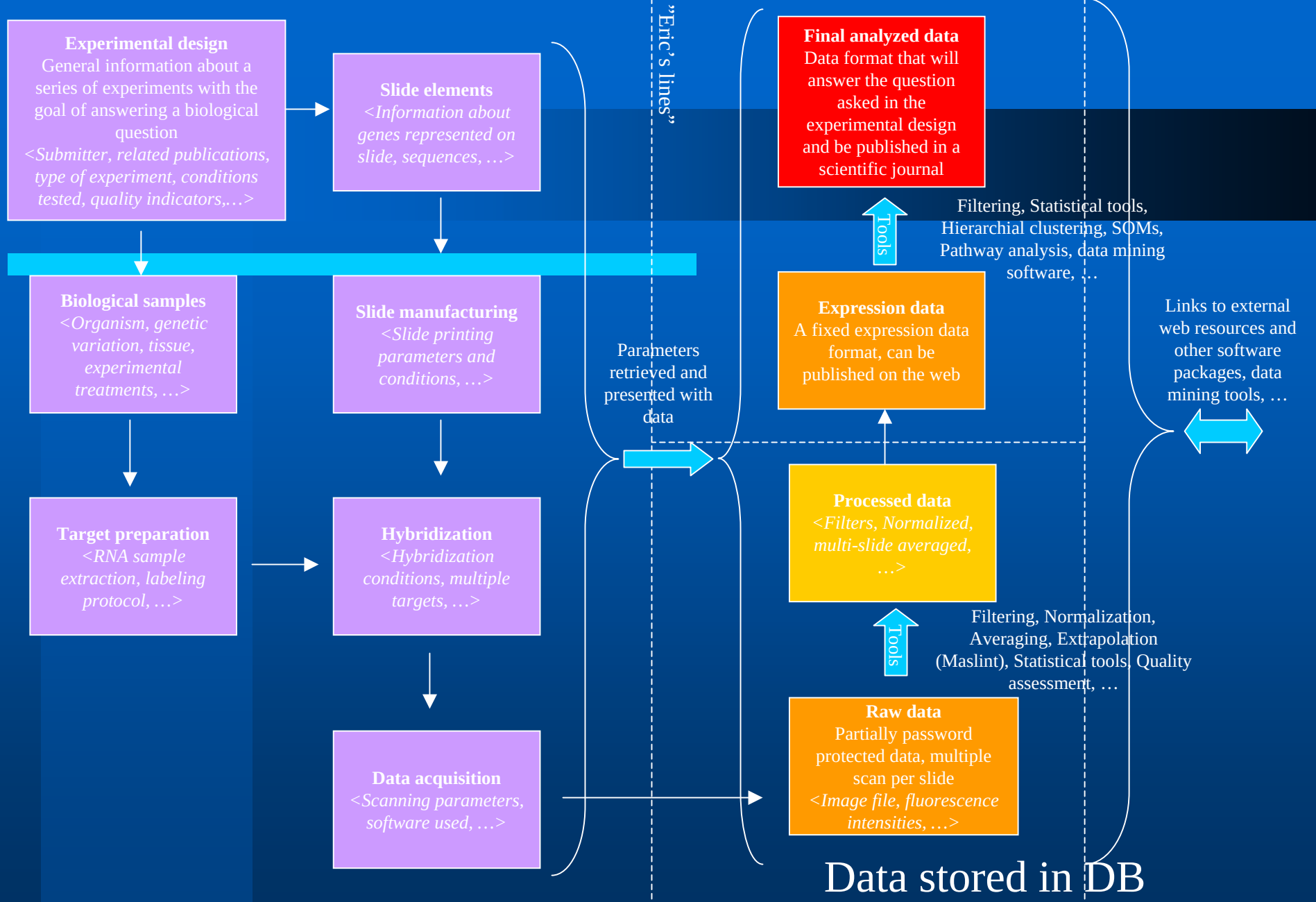
- Ability to accept data from MGH Core Facility and Core Facilities of remote collaborators
- Ability to store both spotted array data and Affymetrix data
- Web-accessibility
- Flexibility to accommodate various types of experiments and the descriptions of those experiments
- Tools for analyzing data and exporting data as tab-delimited files and XML (GEML)

Database Users

- MGH researchers (able to submit data)
- Collaborators (able to submit data through MGH collaborator)
- Scientific community (able to access published data through the web interface)

Types of Tools for Database

- **Tools for visualization of the array image (TIFF or proxy GIF file) as a clickable image map**
  - **Browse individual spots**
  - **Evaluate the placement of the grid used during data acquisition**
  - **Change the flag status of any of the spots**
- **Normalization tools**
- **Clustering analysis tools**



Parameters stored in DB

Each box contains a set of tables

Data stored in DB

Data to be manipulated by tools to different levels (not all data will end in a publication). Data has to be viewed and monitored in the process to determine the necessity to continue the analysis and filter out data points. Experimental parameters and external web resources may need to be called upon in the process.

Background:
Related Software and Other Implementations

- **Stanford Microarray Database**
- **Express DB**
- **Array Express/Expression Profiler**
- **MaxD**

Stanford Microarray Database

- **Strengths**
  - **Open source system**
  - **Supports spotted microarrays**
  - **Sophisticated data normalization tools**
- **Weaknesses**
  - **Affymetrix data format not supported**
  - **RDBMS is Oracle, with Oracle-specific functions in the source code**

Express DB

- **Strengths**
  - **Supports both spotted microarrays and Affymetrix data**
- **Weaknesses**
  - **RDBMS is Sybase 11**
  - **Used as a demonstration system with Saccharomyces, but not yet adapted for other organisms**

Array Express/Expression Profiler

- **Strengths**
  - **Supports both spotted microarrays and Affymetrix data**
  - **Implements the MIAME data specification**
- **Weaknesses**
  - **No storage of raw luminosity data**
  - **RDBMS is Oracle**
  - **More tables would need to be added to contain data pertaining to sample preparation, hybridization and other experimental details**

MaxD

- **Strengths**
  - **Implementation of Array Express table structure suitable for SQL92-complaint databases, thus supporting MySQL**
  - **Java based software with source code available for download on the web**
  - **Strengths of Array Express**
- **Weaknesses**
  - **Weaknesses of Array Express**
  - **Not open source**

Formats of Data Input

- **Automatically entered when spotted arrays are scanned by the core facility**
  - **Array ID, chip layout, spot intensities, software used by the Arrayer**
- **Directly entered by users**
  - **Experiment names, hybridization conditions, procedures**
- **Imported from flat files**
  - **Spot layout of chips, normalization intensities generated by third party software packages (Affymetrix)**

Critical Data to Be Stored

- Description of each experiment
- Information about the submitter
- Description of the hybridization
- Description of the array design
- Description of experiment info related to Affymetrix chips or the core Axon Arrayer
- Description of the sample and target

Critical Data to Be Stored: Experiment

- Unique experiment ID
- Human-readable experiment name
- Classification of experiment type
- Free text description of experiment
- Date of entry
- References to publications
- Submitter ID

Critical Data to Be Stored: Submitter

- **Submitter ID**
- **Submitter's name**
- **Institution**
- **Laboratory**
- **Principal Investigator**
- **Grant**
- **Email address**
- **Postal address**
- **Phone number**

Critical Data to Be Stored: Hybridization

- **Hybridization ID**
- **Reference to the associated experiment and arrays**
- **Free text description of a particular hybridization**
- **Hybridization protocol**
- **Ordinal number for a particular hybridization if the hybridization is part of a sequential set of hybridizations**

Critical Data to Be Stored: Array Design

- Array Design ID
- Human-readable name of the chip design
- Indication of the type of probe used (i.e., spotted vs. synthesized, cDNA vs. oligos)
- Size of array (number of rows and columns and total spots)
- Kind of chip used (e.g., glass, nylon)
- Type of Array (Affymetrix or Axon)
- Supplier who produced the slide (company, individual)
- Protocol to create the chip or provider information if purchased

Critical Data to Be Stored: Affymetrix

- **Name of chip**
- **Sample applied to chip**
- **Probe used with chip**
- **Experimental information found in Affymetrix .EXP files**

Critical Data to Be Stored: Axon Arrayer

- **Description of information from core Axon Arrayer that is also stored in the core microarray database**

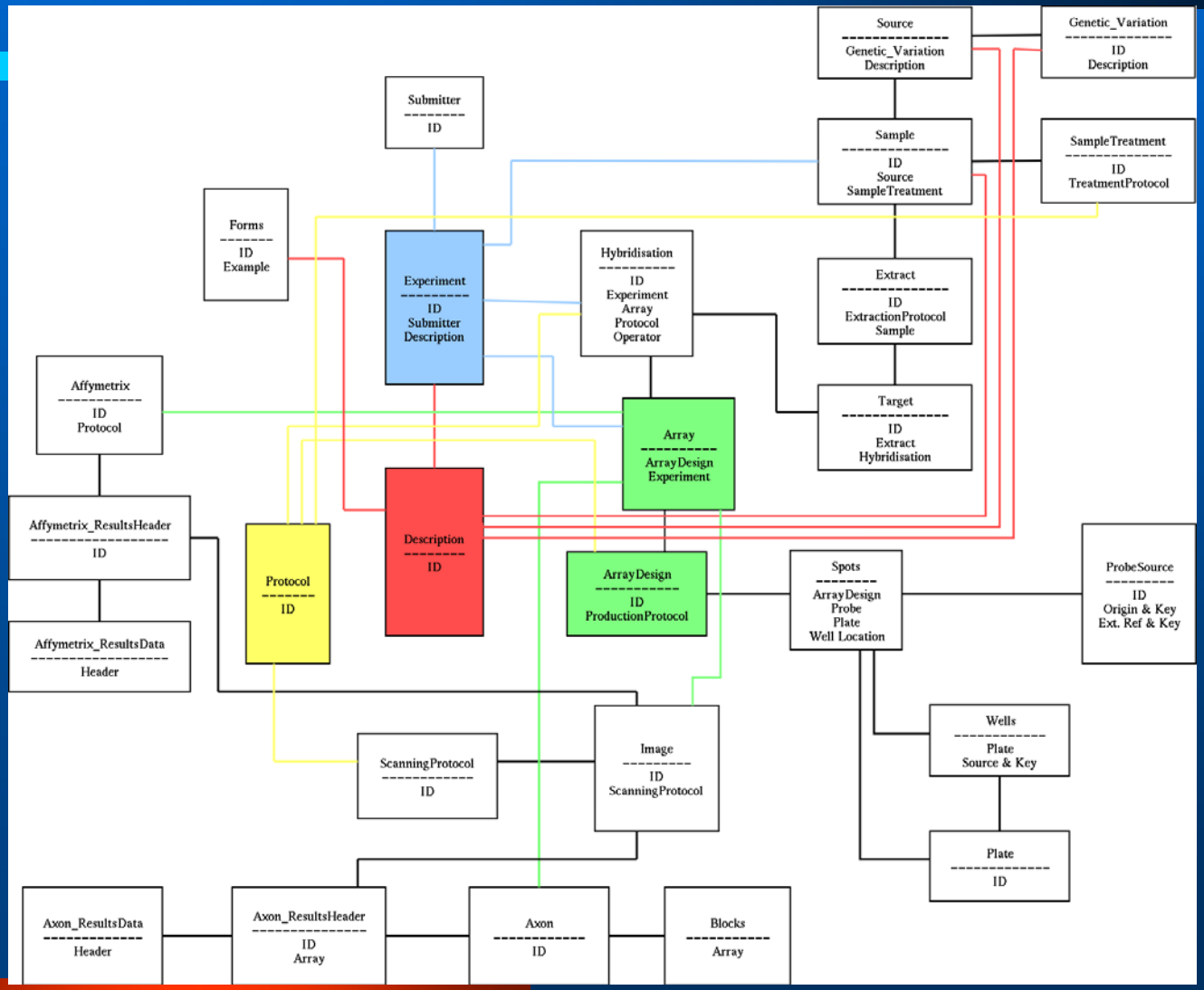
Critical Data to Be Stored: Sample

- **Description of the sample used to make the target that is applied to the chip**
- **Description of the source of the sample (which may include the following information as applicable to a given sample: ID, genus, species, strain, ecotype, organism, organ, tissue, cell type, cell line, cell culture, developmental stage, sex, genetic variation)**

Critical Data to Be Stored: Target

- **Description extract used to make the target**
- **Description of the extraction protocol**
- **Description of the labeling method (if any)**

Database Schema for Integrated Microarray Database System



I. Submitter Information:

Summitter Name: (blank text field to type in name of person who is submitting the experiment (not the data entry person, if different)

Organization: MGH, other

Laboratory: Ausubel, Freeman, Pier, Seed, other

***Grant:** PGA, other

***Grant Number:**

PI of Grant: Ausubel, Freeman, Pier, Seed, other

Email: submitter@institution.edu

Address: Lipid Metabolism Unit, Massachusetts General Hospital, 32 Fruit Street, GRJ 1328, Boston, MA 02114 (blank text field)

Phone: (xxx) xxx-xxxx (blank text field)

Experiment name: name of experiment (blank text field)

Abstract: one line description of experiment (blank text field)

II. Taxonomy:

Organism: Mouse (pull-down choices)
Genus: *Mus* (pull-down choices)
Species: *musculus* (pull-down choices)
Genotype: wild type, mutant, transgenic (pull-down choices)
Strain:
Organ/Tissue: lungs, liver (text field)
Cell type: text field
Cell line: text field
Cell culture: text field
Developmental Stage: text field
Sex: Male, Female, hermaphrodite
Genetic Variation: link to supplemental database if needed
Free Text:
Mutant Name: *tlr4* (free text)
***Name of mutated gene:** *toll-like receptor 4* (free text)
Gene abbreviation: *tlr4* (free text)
Allele name: free text
Dominance: dominant, recessive, semi-dominant, other (pull-down choices)
Mutant type: gain of function, loss of function, null, overexpressor, suppressor, unknown, other (pull-down choices)
Description: free text

III. Sample Treatment:

Sample Description: free text
***Is this experiment a time course?** Yes or No (radio buttons)
Hours after treatment: 2, 4, other (free text)
Temperature:
Type of Treatment: pathogen, hormone, chemical, serum, growth-factor, other (pull-down choices)
Compound: name of chemical, hormone, pathogen, etc. (free text)
***Dose:** free text
***Concentration:** free text
Treatment Protocol: free text
RNA extraction method: free text
Amount of RNA obtained: free text
Hybridization: free text
Number of Hybridization: (if more than one hybridization per chip) free text of a number
Hybridization protocol: free text
Labeling method for target: free text
Labeling protocol: free text
Amount of sample used to make target: free text
Supplemental Database: (pull-down choice) plant

Example Queries

1) **List all experiments performed by a single user.**
2) **Retrieve all experiments entered into the database since October 31, 2001.**
3) **Retrieve normalized data for two arrays in an experiment and graph the luminosity values on a log-log scatter plot.**

Example Queries

4) **List all experiments from a particular lab, or operator.**
5) **List all experiments using a particular protocol.**
6) **List all experiments performed on an extract from a particular tissue type.**

Example Queries

7) **Which genes are expressed in response to pathogen A, but not pathogen B in a given host?**
8) **Compare the results of multiple treatments and produce a Venn diagram showing sets of genes induced or repressed by these different treatments or pathogens.**
9) **Calculate distance matrices to analyze the extent of differences between treatments, time points or mutants.**

Tools

- **Cluster (Stanford): clustering on large datasets (hierarchical, SOMs, kmeans, PCA)**
- **TreeView (Stanford): view cluster output**
- **EPCLUST (EBI): hierarchical clustering of gene expression datasets**

IMDS Development Team

- Harry Bjorkbacka (End User/Feature Consultant)
- Cheri Chen (End User)
- Lance Davidow (Developer/User)
- Julia Dewdney (End User/Feature Consultant)
- Chen Liu (Developer)
- Christina Powell (Developer/End User)
- Sean Quinlan (Database/Program Developer)
- Jonathan M. Urbach (Program Developer)
- Eric VanHelene (Manager)