DNA Chip Database for the Korean Functional Genomics Project

Sangsoo Kim, Ph.D.
Korea Research Institute of Bioscience & Biotechnology
E-mail: sskimb@mail.kribb.re.kr

Abstract

The Korean Functional Genomics Project focuses on stomach and liver cancers. Specimens collected by six hospital teams are used in DNA microarray experiments. Experimental conditions, spot measurement data, and the associated clinical information are stored in a relational database. Microarray database schema was developed based on EBI's ArrayExpress. A diagrammatic representation of the schema is used to help navigate over many tables in the database. Field description, table-to-table relationship, and other database features are also stored in the database and these are used by a PERL interface program to generate web-based input forms on the fly. As such, it is rather simple to modify the database definition and implement controlled vocabularies. This PERL program is a general-purpose utility which can be used for inputting and updating data in relational databases. It supports file upload and user-supplied filters of uploaded data. Joining related tables is implemented using JavaScripts, allowing this step to be deferred to a later stage. This feature alleviates the pain of inputting data into a multi-table database and promotes collaborative data input among several teams. Pathological finding, clinical laboratory parameters, demographical information, and environmental factors are also collected and stored in a separate database. The same PERL program facilitated developing this database and its user-interface.

CV

2000년-현재: 생명공학연구소 유전체연구센터 책임연구원, 과학기술부 21세기 프론티어사업 "인간유전체기능연구사업단" Bioinformatics 책임자

1999년-2000년: ㈜LG화학 책임연구원, Bioinformatics 팀장

1995년-1998년: ㈜LG화학 책임연구원, Biopharmaceutical Design 팀장

1988년-1995년: ㈜LG화학 선임연구원, Biopharmaceutical Design 팀장

1986년-1988년: Purdue University, Postdoctoral Res. Ass., Dept. of Bio. Sci.

1983년-1986년: Iowa State University, Ph.D. in Physical Chemistry

1981년-1983년: Seoul National University, M.S. in Physical Chemistry

1977년-1981년: Seoul National University, B.S. in Chemistry

cDNA Microarray Database for Korean Functional Genomics Project

Oct. 2001

Korea Research Institute of
Bioscience & Biotechnology
Sangsoo Kim

Scope and Objectives of Korean Function Genomics

- Establishment of basic infrastructure for Functional Genomics
  - HT sequencing, DNA chip, Proteomics, bfx
- Discovery of gene markers associated with diseases common in Korean population
  - Diagnostic and prognostic markers of stomach or liver cancer

Phase I: Infrastructure

1st year	2nd year
• Tissue collection	• Cell-based assay
• **HT sequencing**	• Model organisms
• Korean-type SNP	• KO/tg mice
• **DNA microarray**	• Structural genomics
• Proteomics	• Chemical genomics
• **Bioinformatics**	• Protein interaction

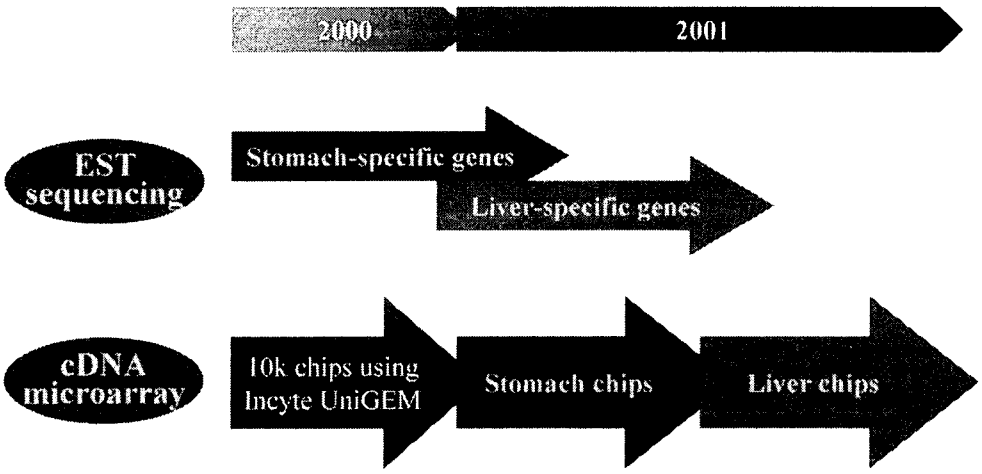
Bioinformatics Mission

v Data Archiving
  – Scalable, Secure, Flexible, Affordable
v Analysis Support
  – State-of-the-art algorithms, Web-GUI
• Sharing Information
  – Data integration, Performance

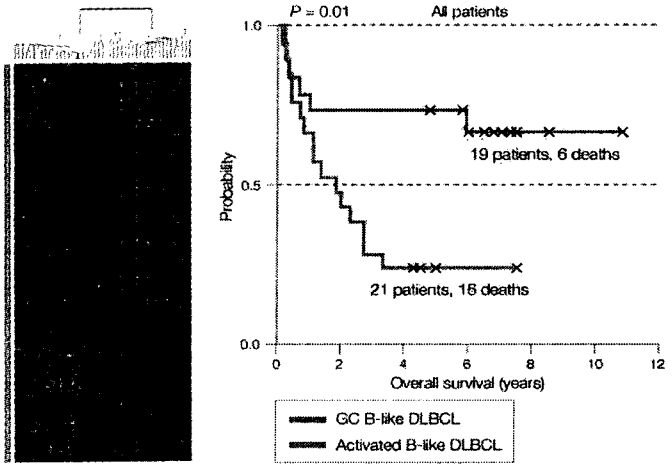
Data Archiving

- v HT sequencing (1 team)
  - EST sequencing of cancer cDNA libraries
- v DNA microarray (2 teams)
  - 2.5k chips (now), 10k (end of this year) chips
- v Tissue collection (6 teams)
  - 3 teams for stomach and liver cancers, respectively
- • Korean-type SNPs (5 teams)
  - Scoring known markers, Discovering cSNPs
- • Proteomics (2 teams)
  - 2-D Gel images and processed data

Tissue-specific UniGene Chips



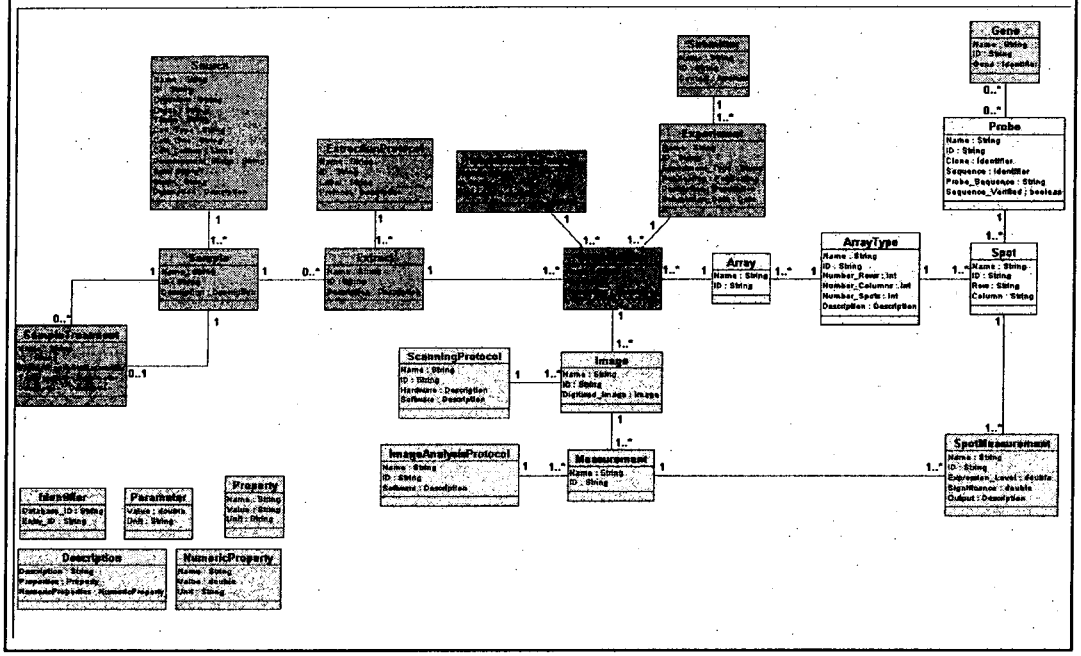
Discovery of diffuse large B-cell lymphoma (DLBCL) clinical subtypes by DNA array profiling

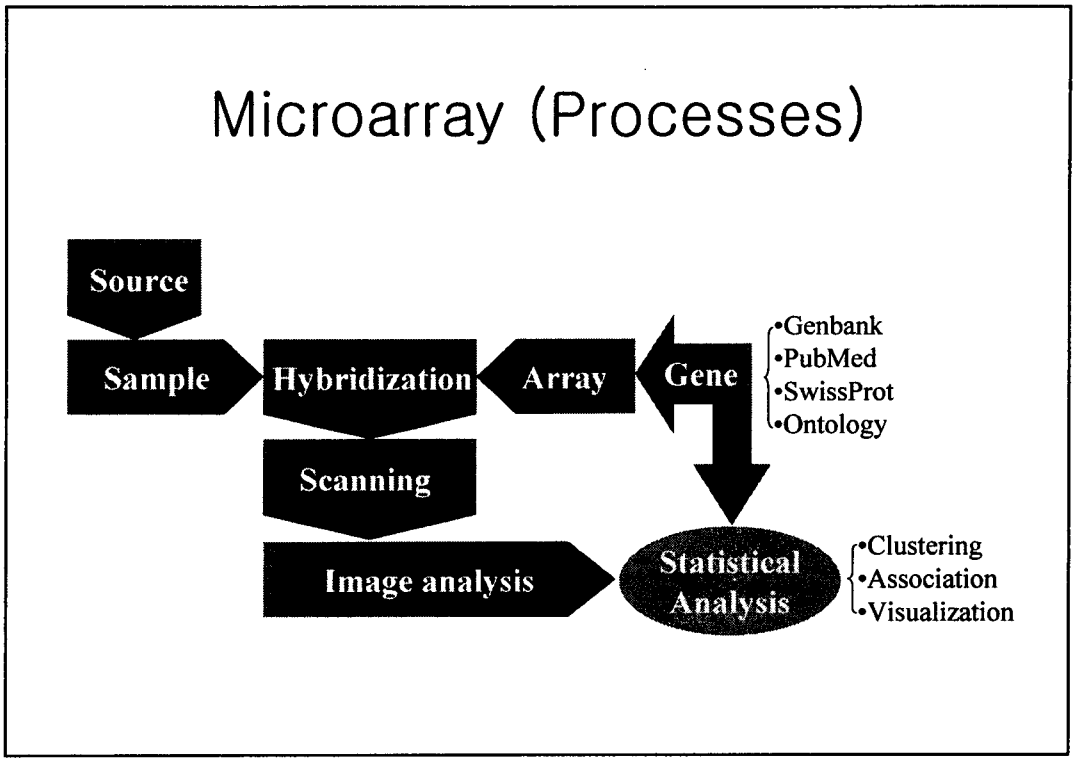


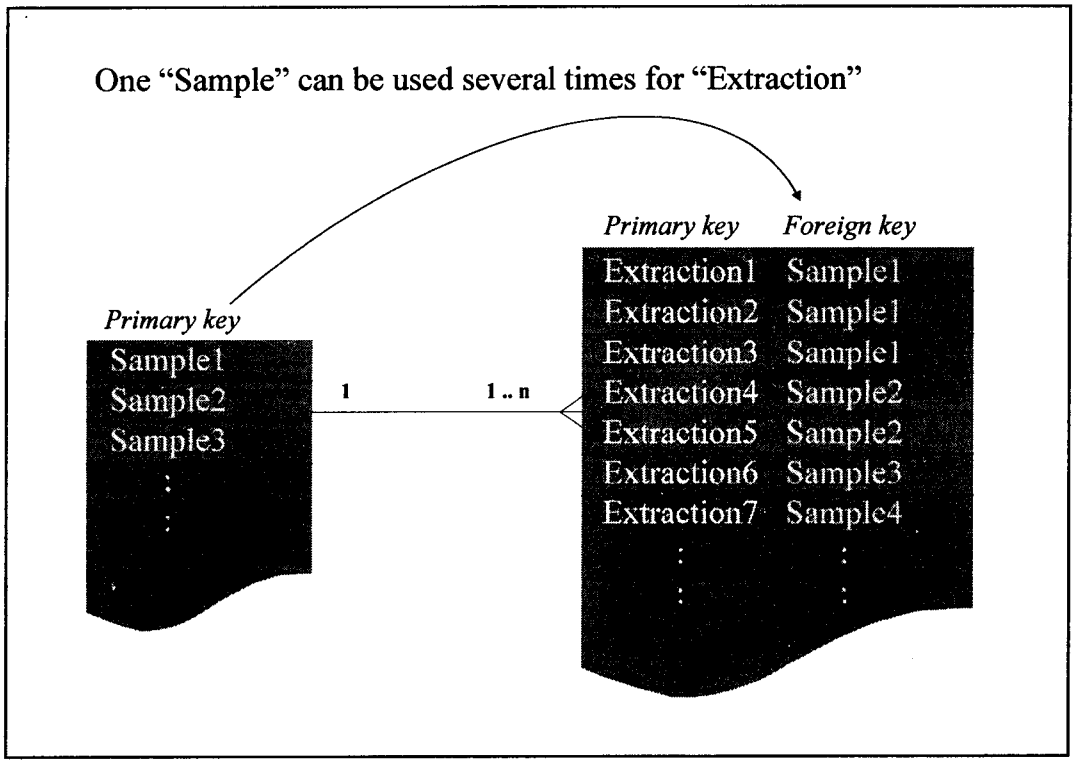
Nature Reviews Genetics 1: 48-56 (2000)
MOLECULAR PROFILING OF HUMAN CANCER

Nature Reviews | Genetics

EBI Microarray Database Schema (www.ebi.ac.uk/arrayexpress)

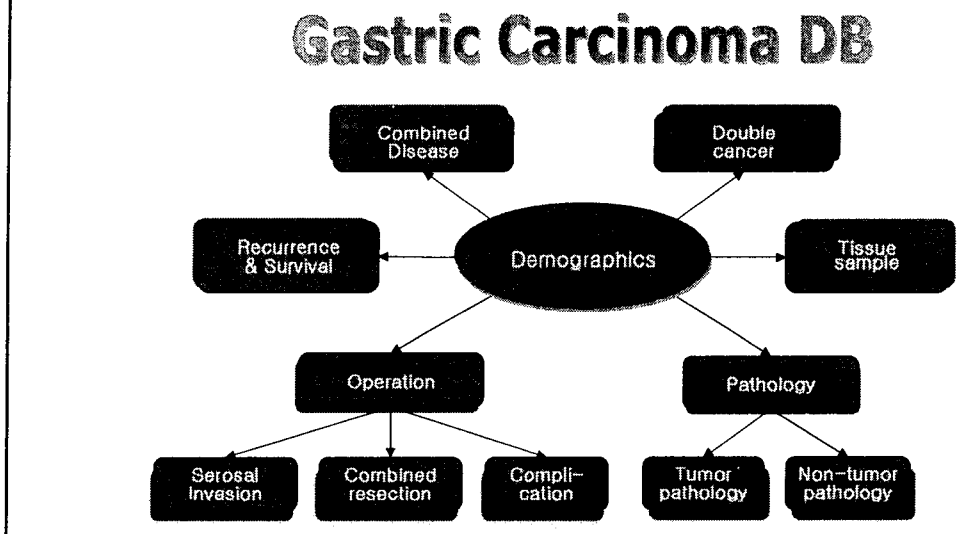


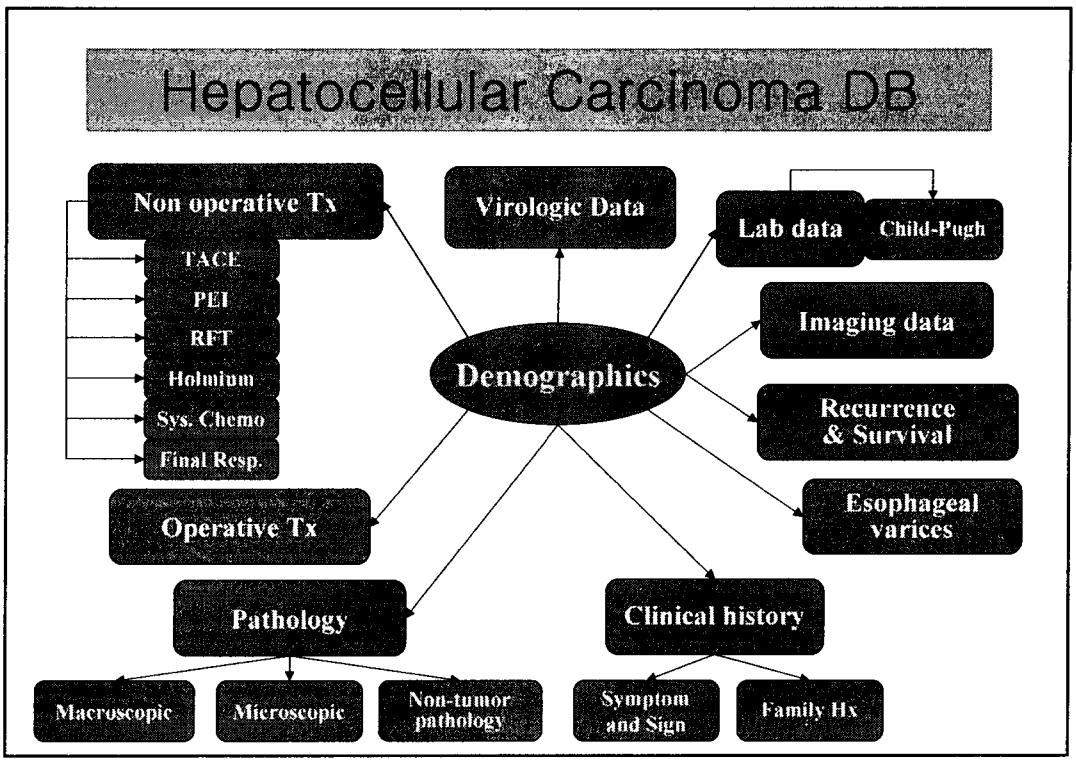

Microarray (Processes)
Source
Sample
Hybridization
Array
Gene
•Genbank
•PubMed
•SwissProt
•Ontology
Scanning
Image analysis
Statistical Analysis
•Clustering
•Association
•Visualization

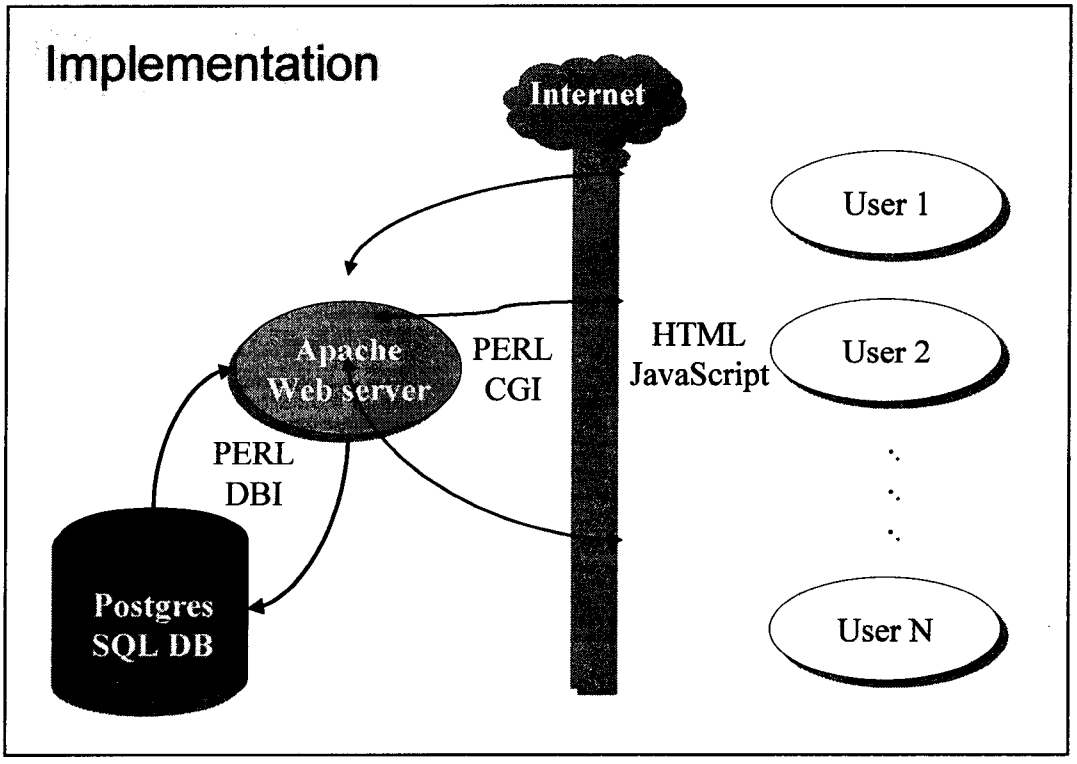

One "Sample" can be used several times for "Extraction"
Primary key
Foreign key
Primary key
Sample1
Sample2
Sample3
1
1 .. n
Extraction1 Sample1
Extraction2 Sample1
Extraction3 Sample1
Extraction4 Sample2
Extraction5 Sample2
Extraction6 Sample3
Extraction7 Sample4

Clinical Tissue Database

- Epidemiological parameters
  - sex, age, education, lifestyle, job, home
  - family history
  - prior diseases, hepatitis
  - dietary habit, smoking, alcohol, occ. hazard
  - physical exam., clinical lab.
  - lifespan
- Pathological parameters
  - surgery, tumor size, stage, type, invasion, lymphocyte infiltration




Hepatocellular Carcinoma DB
Non operative Tx
TACE
PEI
RFT
Holmium
Sys. Chemo
Final Resp.
Virologic Data
Lab data
Child-Pugh
Imaging data
Demographics
Recurrence & Survival
Esophageal varices
Operative Tx
Pathology
Clinical history
Macroscopic
Microscopic
Non-tumor pathology
Symptom and Sign
Family Hx


Implementation
Internet
User 1
User 2
User N
Apache Web server
PERL CGI
HTML JavaScript
PERL DBI
Postgres SQL DB

Database Features

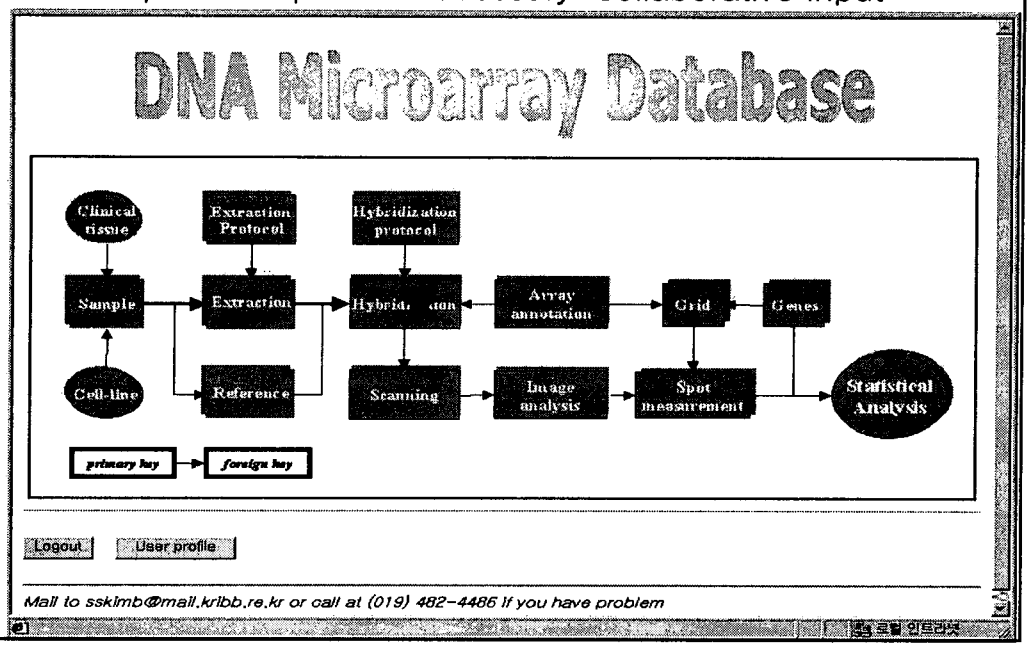
- User authentication
  - cookie-based
- Navigation via 'map diagram'
  - Sequential input not necessary; cooperative input
- Table definitions in DB
  - Flexible design change; controlled vocabulary
- Table-table relationships in DB
  - Copy values from upstream to downstream tables
- File-upload for long tables with filters
- Link-out to external URLs
- General-purpose DB interface

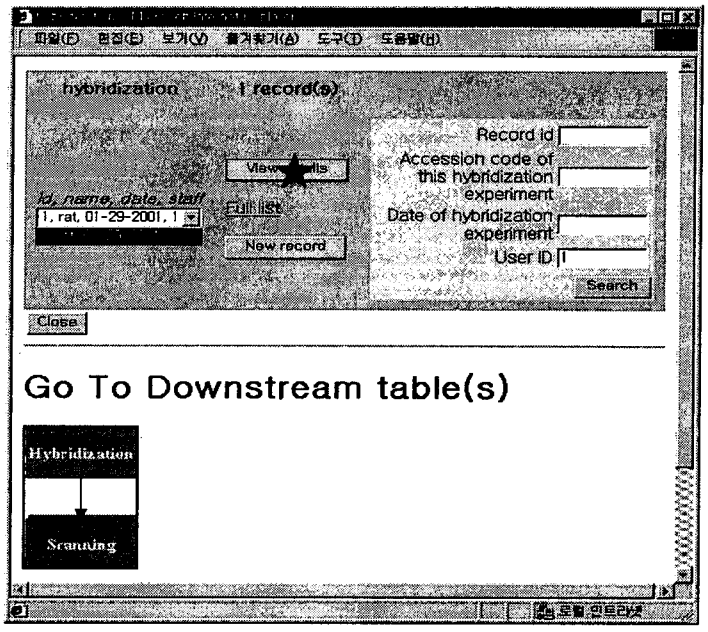
Table definitions stored in DB – easy design change, controlle

TABLE	FIELD	TYPE	DESCRIPTION
demograp	hospital_code	character varying(10)	Hospital Code
demograp	ch c	character varying(10)	Ch
demograp	social_no	character varying(14)	Social No
demograp	name c	character varying(10)	Name
demograp	age c	int2	Age
demograp	sex c	character	Sex { M, F }
demograp	date c	date	Enr b
demograp	staff c	int2	user ID
symp	name	character varying(10)	patient name
symp	date	date	Enr b
symp	staff	int2	user ID
symp	abdominal_pain	int2	Abdominal pain { 1 (Yes), 2 (No) }
symp	abdominal_distension	int2	Abdominal distension { 1 (Yes), 2 (No) }
symp	weight_loss	int2	weight loss { 1 (Yes), 2 (No) }
symp	weakness	int2	weakness { 1 (Yes), 2 (No) }
symp	jaundice	int2	Jaundice { 1 (Yes), 2 (No) }
symp	vomiting	int2	Vomiting { 1 (Yes), 2 (No) }
symp	palpable_mass	int2	Palpable mass { 1 (Yes), 2 (No) }
symp	fever	int2	Fever { 1 (Yes), 2 (No) }
symp	other	character varying(32)	other
symp	no_symptom	int2	No symptom { 1 (Yes), 2 (No) }
family_hx	name	character varying(10)	patient name
family_hx	date	date	Enr b
family_hx	staff	int2	user ID
family_hx	hx_of_chronic_liver_diseases	int2	Hx of chronic liver diseases { 1 (Yes), 2 (No) }
family_hx	hcc_hx	int2	HCC Hx { 1 (parents), 2 (sibling), 3 (offspring) }
family_hx	hbv_hx	int2	HBV Hx { 1 (parents), 2 (sibling), 3 (offspring) }
family_hx	hcv_hx	int2	HCV Hx { 1 (parents), 2 (sibling), 3 (offspring) }
clinical_history	name	character varying(10)	patient name
clinical_history	date	date	Enr b
clinical_history	staff	int2	user ID
clinical_history	duration_of_hbv_hx	float4	Duration of HBV hx (years), 0 for none
clinical_history	duration_of_hcv_hx	float4	Duration of HCV hx (years), 0 for none
clinical_history	duration_of_cirrhosis	float4	Duration of cirrhosis (years), 0 for none
clinical_history	transfusion_hx	float4	Transfusion Hx (years), 0 for none

Navigation via 'map diagram'

- Sequential input not necessary; collaborative input





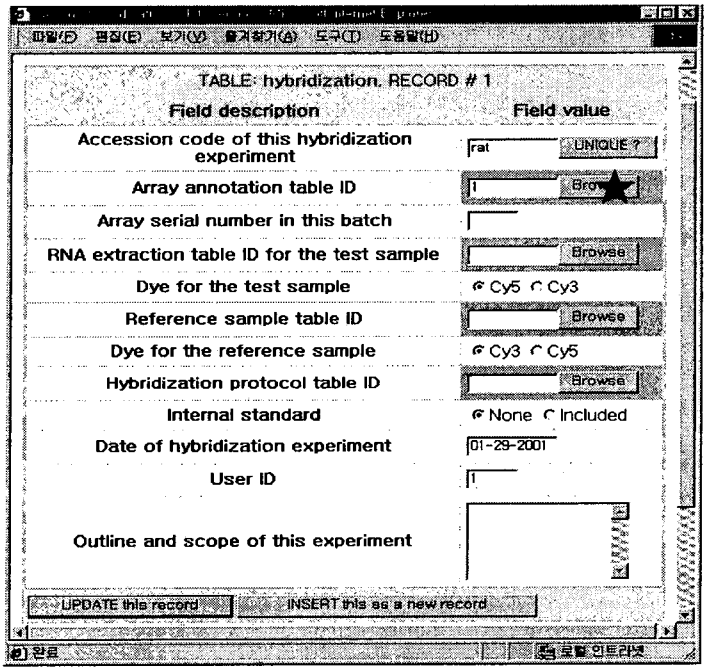
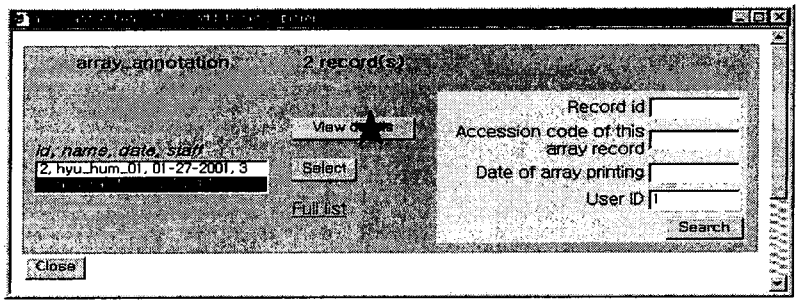
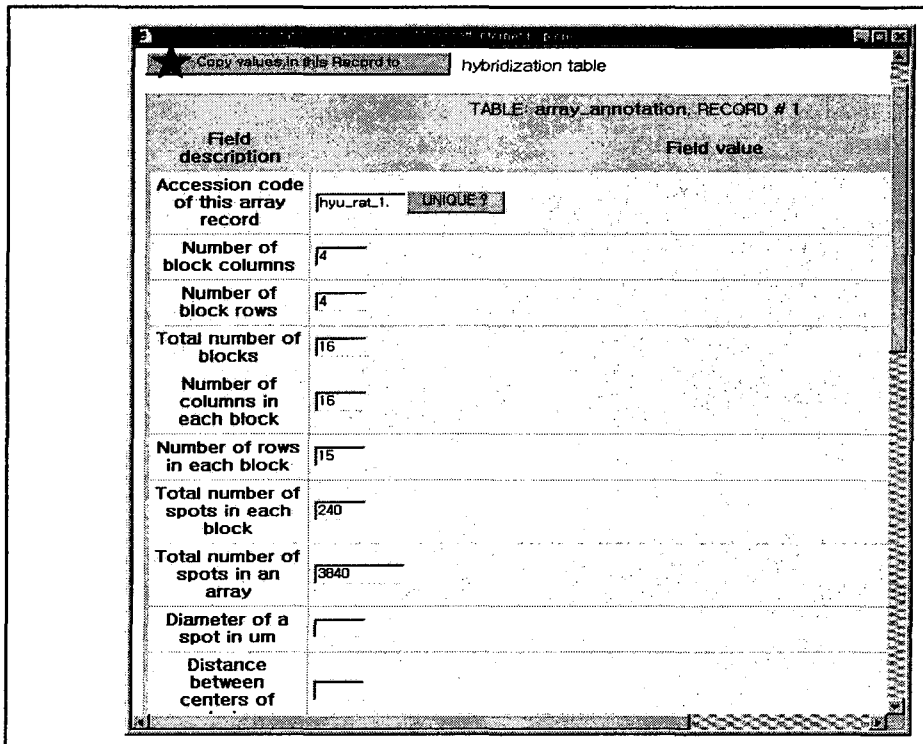


TABLE: hybridization, RECORD # 1
Field description
Field value
Accession code of this hybridization experiment
rat
UNIQUE ?
Array annotation table ID
Array serial number in this batch
RNA extraction table ID for the test sample
Browse
Dye for the test sample
Cy5 Cy3
Reference sample table ID
Dye for the reference sample
Cy3 Cy5
Hybridization protocol table ID
Internal standard
None Included
Date of hybridization experiment
01-29-2001
User ID
Outline and scope of this experiment
UPDATE this record
INSERT this as a new record


array_annotation
2 record(s)
id, name, date, staff
2, hyu_hum_01, 01-27-2001, 3
Select
Full list
Record id
Accession code of this array record
Date of array printing
User ID
Search
Close


Copy values in this Record to
hybridization table
TABLE: array_annotation, RECORD # 1
Field description
Field value
Accession code of this array record
hyu_rat_1.
UNIQUE ?
Number of block columns
4
Number of block rows
4
Total number of blocks
16
Number of columns in each block
16
Number of rows in each block
15
Total number of spots in each block
240
Total number of spots in an array
3840
Diameter of a spot in um
Distance between centers of

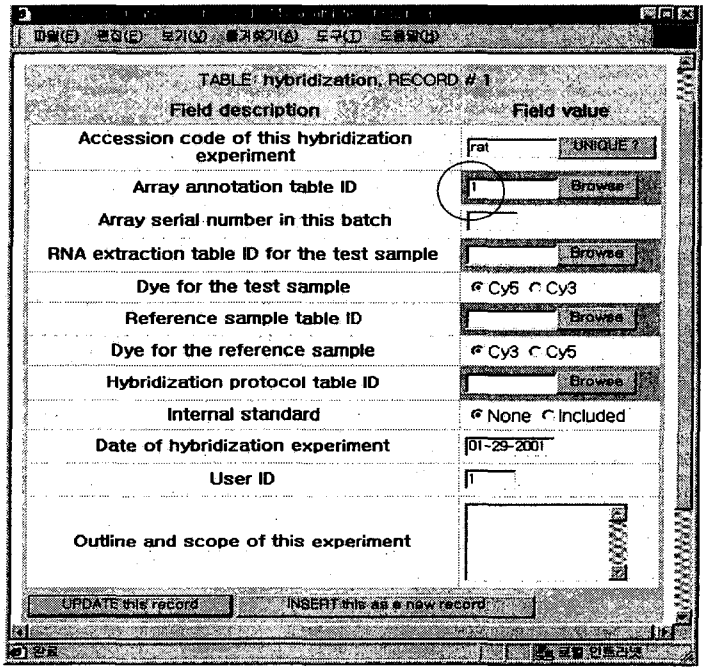
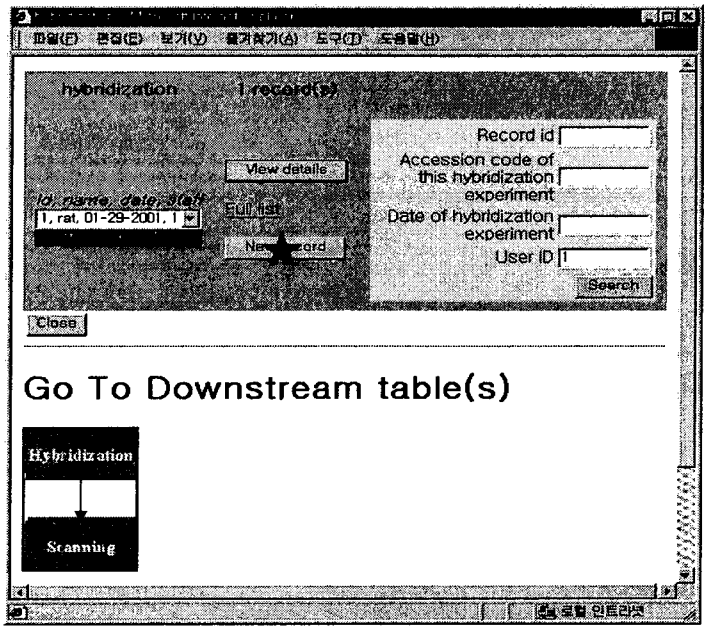
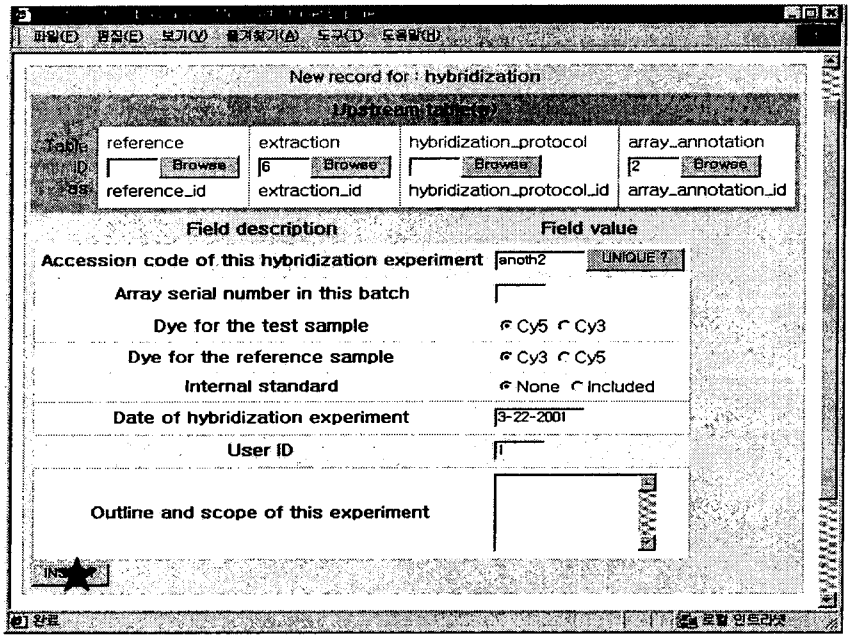
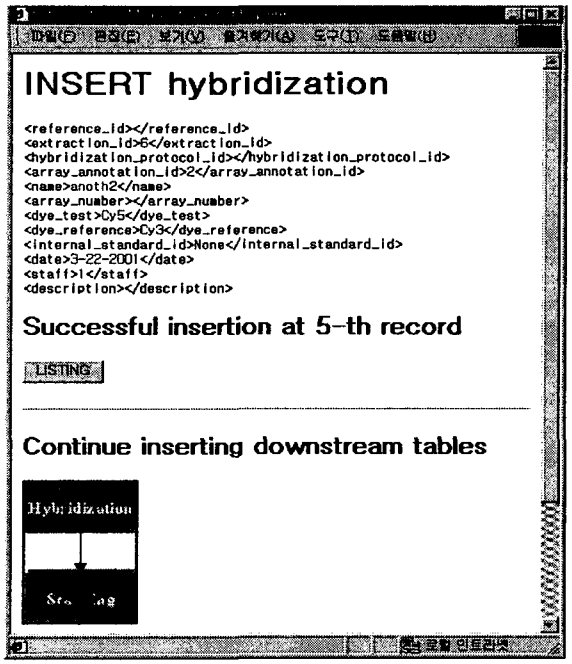
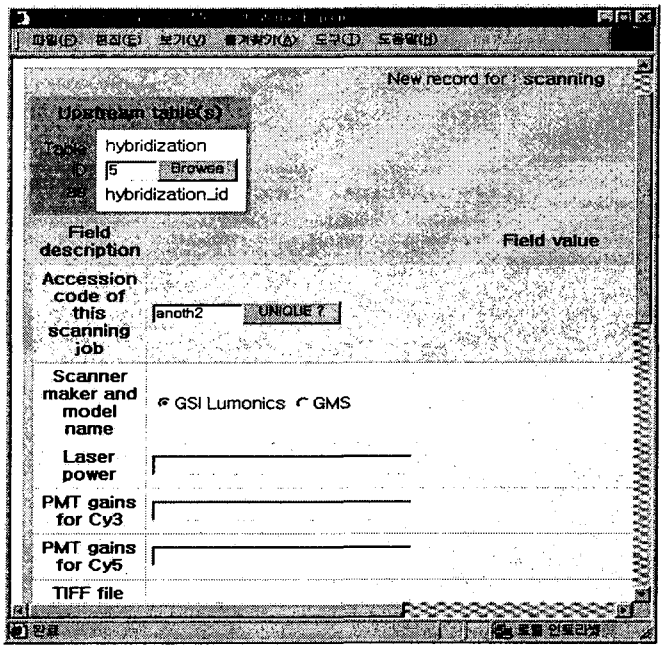


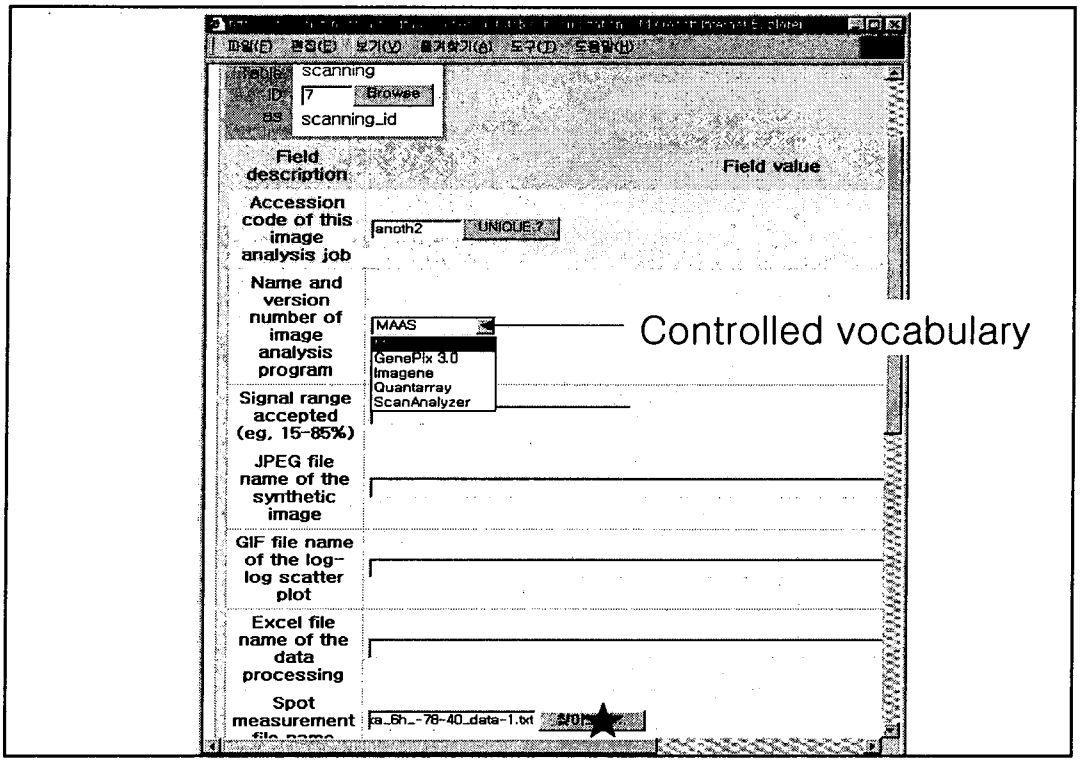
TABLE: hybridization, RECORD # 1
Field description
Field value
Accession code of this hybridization experiment
rat
UNIQUE ?
Array annotation table ID
1
Browse
Array serial number in this batch
RNA extraction table ID for the test sample
Browse
Dye for the test sample
Cy5 Cy3
Reference sample table ID
Browse
Dye for the reference sample
Cy3 Cy5
Hybridization protocol table ID
Browse
Internal standard
None Included
Date of hybridization experiment
01-29-2001
User ID
1
Outline and scope of this experiment
UPDATE this record
INSERT this as a new record

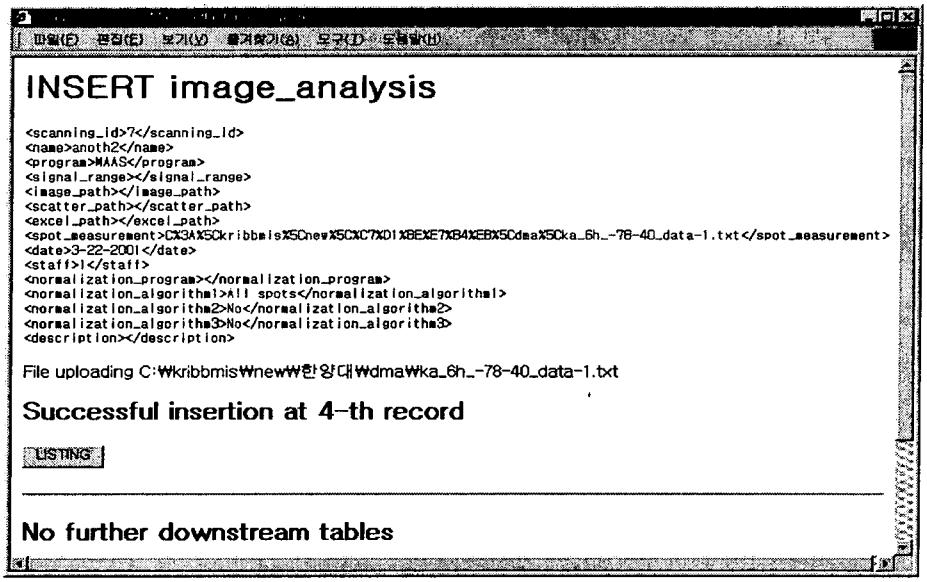

hybridization
1 record(s)
View details
Record id
Accession code of this hybridization experiment
Date of hybridization experiment
User ID
Search
Close
Go To Downstream table(s)
Hybridization
Scanning


New record for : hybridization
reference
extraction
hybridization_protocol
array_annotation
Browse
reference_id
extraction_id
hybridization_protocol_id
array_annotation_id
Field description
Field value
Accession code of this hybridization experiment
anoth2
UNIQUE ?
Array serial number in this batch
Dye for the test sample
Cy5 Cy3
Dye for the reference sample
Cy3 Cy5
Internal standard
None Included
Date of hybridization experiment
3-22-2001
User ID
Outline and scope of this experiment

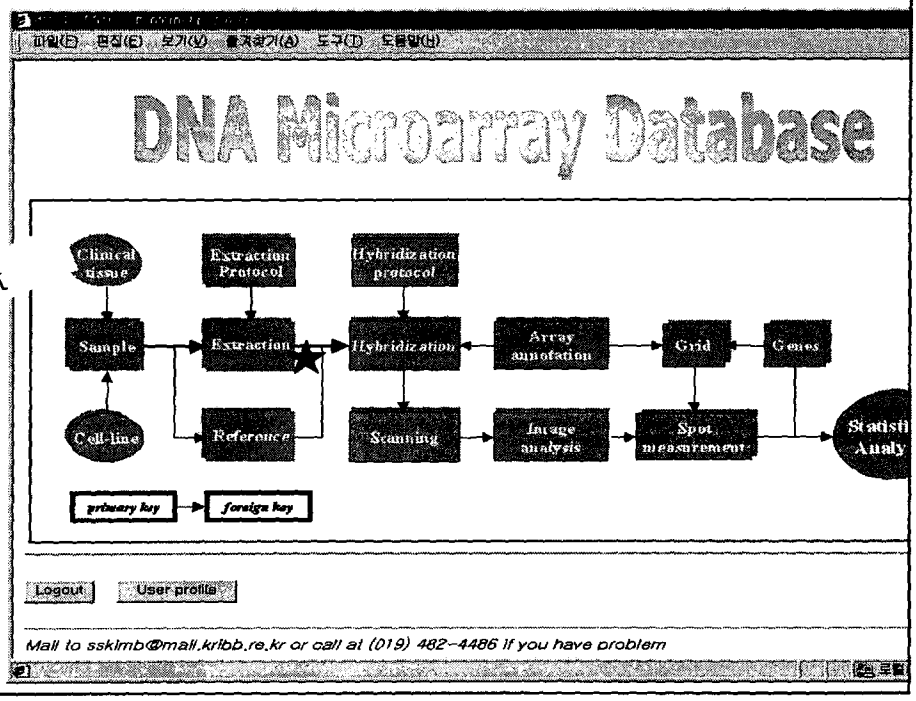

INSERT hybridization
<reference_id></reference_id>
<extraction_id>6</extraction_id>
<hybridization_protocol_id></hybridization_protocol_id>
<array_annotation_id>2</array_annotation_id>
<name>anoth2</name>
<array_number></array_number>
<dye_test>Cy5</dye_test>
<dye_reference>Cy3</dye_reference>
<internal_standard_id>None</internal_standard_id>
<date>3-22-2001</date>
<staff>1</staff>
<description></description>
Successful insertion at 5-th record
LISTING
Continue inserting downstream tables

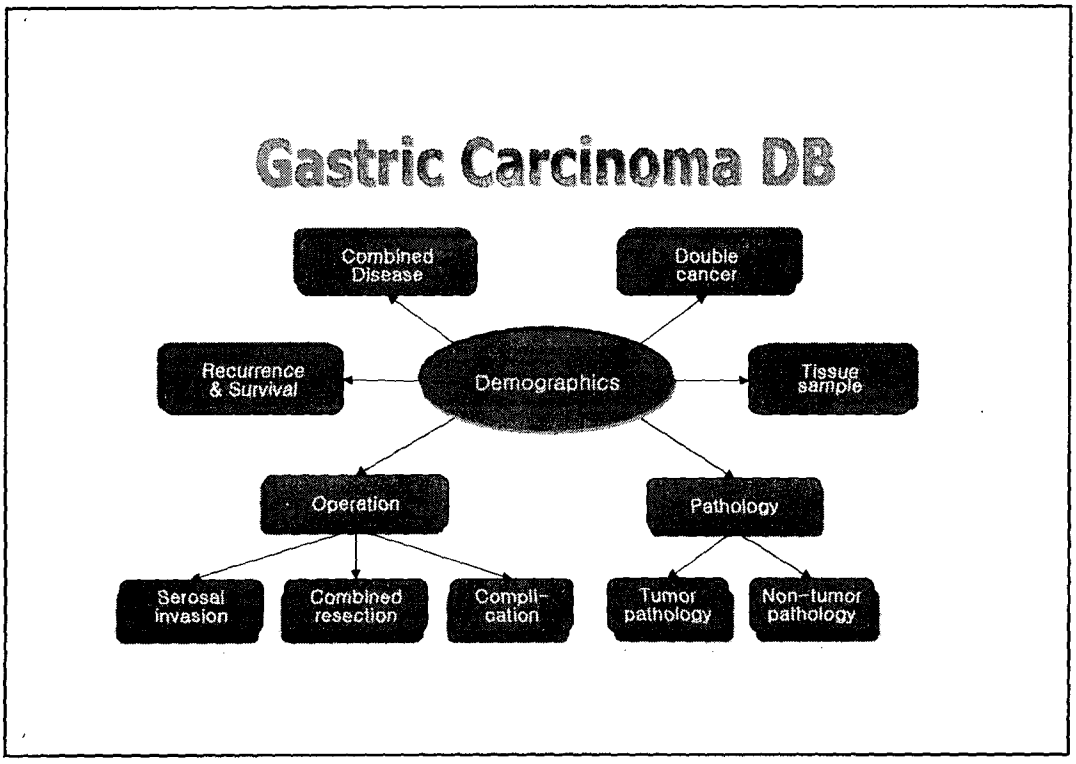

New record for : scanning
hybridization
5
Browse
hybridization_id
Field description
Field value
Accession code of this scanning job
anoth2
UNIQUE ?
Scanner maker and model name
GSI Lumonics
GMS
Laser power
PMT gains for Cy3
PMT gains for Cy5
TIFF file


scanning
7
Browse
scanning_id
Field description
Field value
Accession code of this image analysis job
anoth2
UNIQUE ?
Name and version number of image analysis program
MAAS
GenePix 3.0
Imagene
Quantarray
ScanAnalyzer
Controlled vocabulary
Signal range accepted (eg. 15-85%)
JPEG file name of the synthetic image
GIF file name of the log-log scatter plot
Excel file name of the data processing
Spot measurement
ka_6h_-78-40_data-1.txt


INSERT image_analysis
<scanning_id>7</scanning_id>
<name>anoth2</name>
<program>MAAS</program>
<signal_range></signal_range>
<image_path></image_path>
<scatter_path></scatter_path>
<excel_path></excel_path>
<spot_measurement>C%3A%5Ckribbmis%5Cnew%5C%C7%D1%BE%E7%B4%EB%5Cdma%5Cka_6h_-78-40_data-1.txt</spot_measurement>
<date>3-22-2001</date>
<staff>1</staff>
<normalization_program></normalization_program>
<normalization_algorithm1>All spots</normalization_algorithm1>
<normalization_algorithm2>No</normalization_algorithm2>
<normalization_algorithm3>No</normalization_algorithm3>
<description></description>
File uploading C:₩kribbmis₩new₩한양대₩dma₩ka_6h_-78-40_data-1.txt
Successful insertion at 4-th record
LISTING
No further downstream tables

Link-out


DNA Microarray Database
Clinical tissue
Extraction Protocol
Hybridization protocol
Sample
Extraction
Hybridization
Array amnofation
Grid
Genes
Cell-line
Reference
Scanning
Image analysis
Spot measurement
Statisti Analy
primary key
foreign key
Logout
User profile
Mail to sskimb@mail.kribb.re.kr or call at (019) 482-4486 if you have problem


Gastric Carcinoma DB
Combined Disease
Double cancer
Recurrence & Survival
Demographics
Tissue sample
Operation
Pathology
Serosal invasion
Combined resection
Compli-cation
Tumor pathology
Non-tumor pathology

Hierarchical Clustering of Stomach Cancer Samples
- 20 normal+tumor pairs
- courtesy of Inchul Lee, Asan Medical Center

Normal Cancer+*normal*