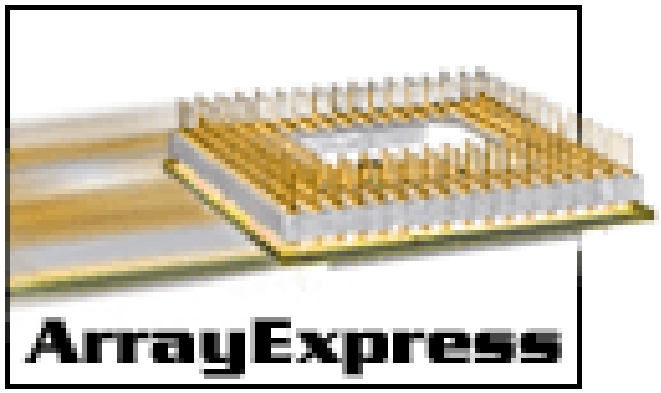
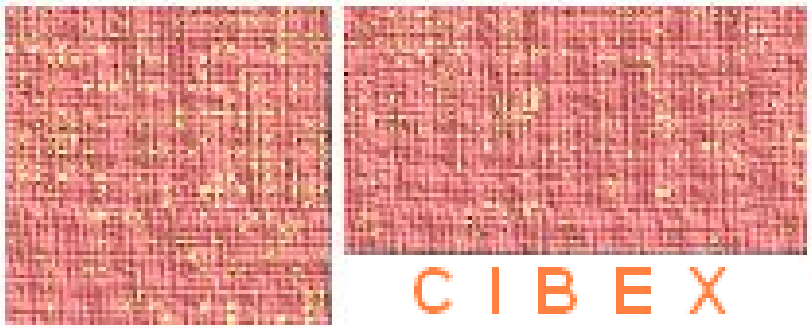
ArrayExpress vs NCBI GEO

Presentation by *Warren Cheung*

- National Center for Biotechnology Information (NCBI) Gene Expression Omnibus — North America
- established 2000 — older
- stores a variety of "high-thoroughput molecular abundance data"
- Aim: robust, *versatile* data repository
- `http://www.ncbi.nlm.nih.gov/geo`



- European Bioinformatics Institute (EBI) — Europe
- established 2002 — newer
- stores microarray data only
- Aim: data *repository* as well as well-annotated data *warehouse*
- `http://www.ebi.ac.uk/arrayexpress`



- Center for Information Biology gene EXpression database
- `http://cibex.nig.ac.jp/`
- National Institute for Genetics — Japan
- WIP?

MIAME compliance

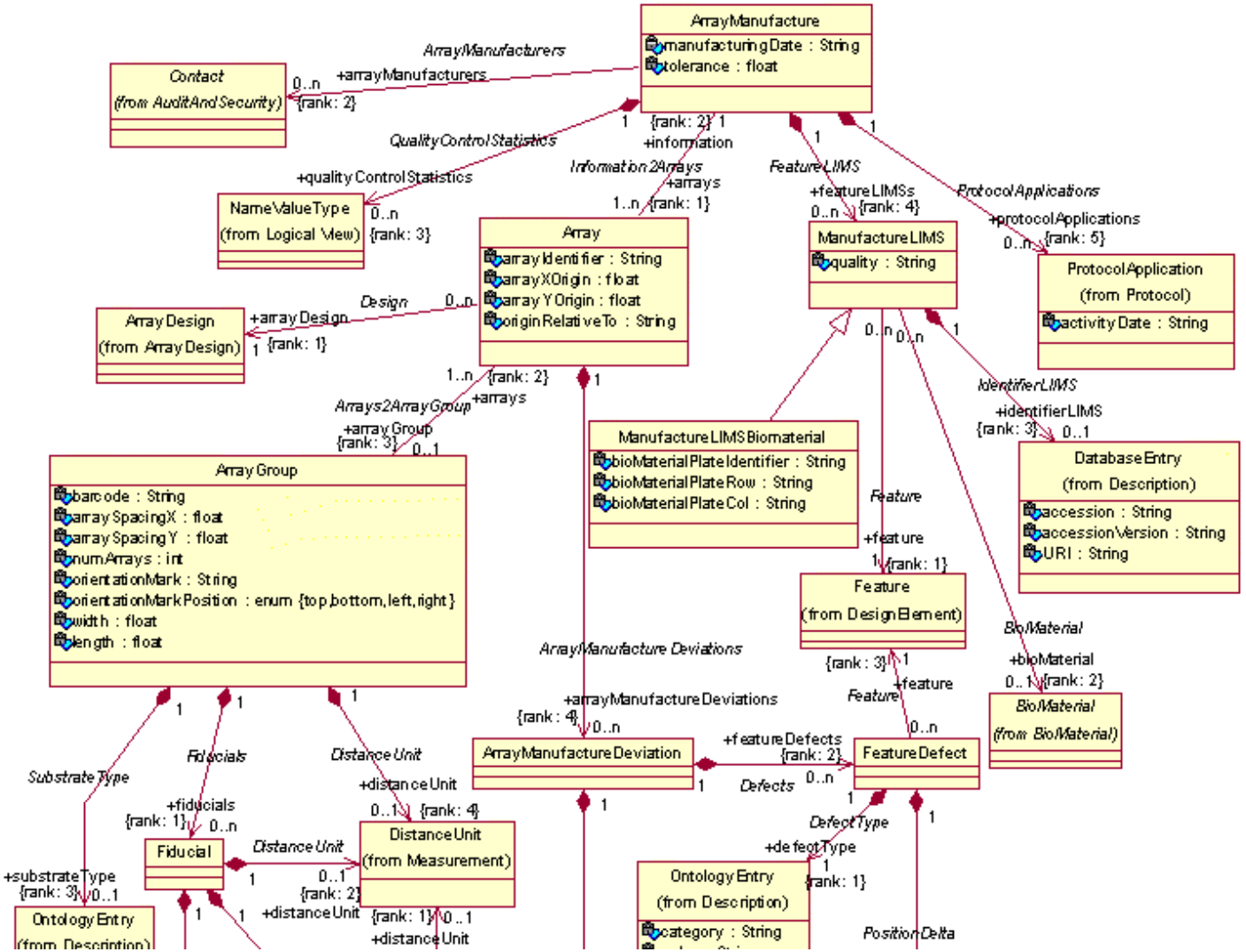
- Minimum Information About a Microarray Experiment
- standards/recommendations from the Microarray Gene Expression Data society (`http://www.mged.org`)
- Checklist of information describing the experiment, with sufficient detail to replicate the experiment
- Goal is to make submission of microarray data a part of a publication submission process, like sequence submission

Expression Data

- ArrayExpress: microarray data only
- NCBI GEO: "high-thoroughput molecular abundance data"
  - microarray data
  - SAGE data
  - mass spectrometry peptide profiling

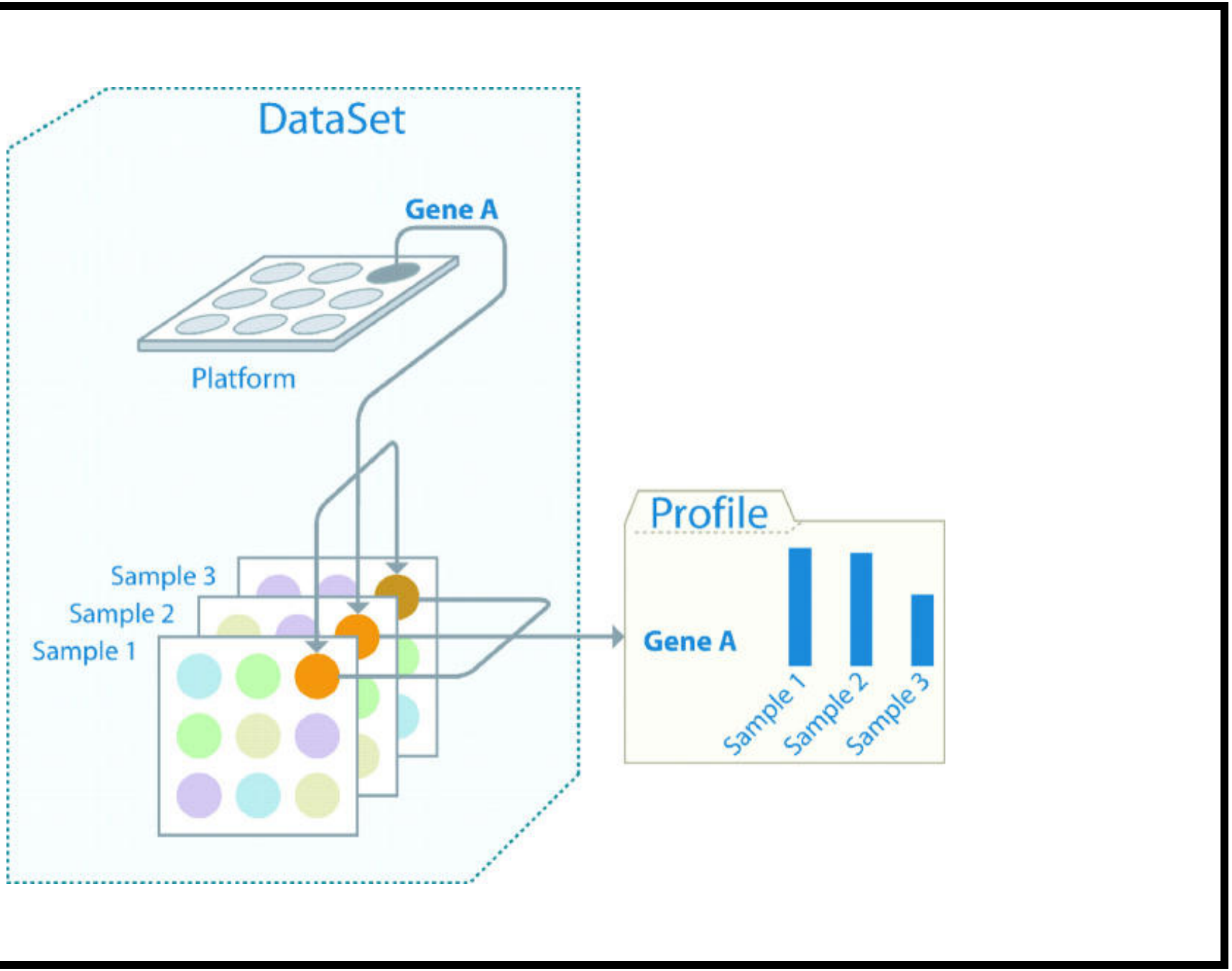
Data Organisation — ArrayExpress

- Experiment refers to set of hybridizations (linked to publication)
- accession numbers assigned to each experiment
- database schema translated from MAGE-OM to Oracle


ArrayManufacture
manufacturingDate : String
tolerance : float
ArrayManufacturers
Contact
(from AuditAndSecurity)
0..n
+arrayManufacturers
{rank: 2}
1
{rank: 2} 1
+information
QualityControlStatistics
+qualityControlStatistics
NameValueType
(from Logical View)
0..n
{rank: 3}
Information2Arrays
+arrays
1..n {rank: 1}
Array
arrayIdentifier : String
arrayXOrigin : float
arrayYOrigin : float
originRelativeTo : String
FeatureLIMS
+featureLIMSs
0..n {rank: 4}
ManufactureLIMS
quality : String
ProtocolApplications
+protocolApplications
0..n {rank: 5}
ProtocolApplication
(from Protocol)
activityDate : String
Design
0..n
+arrayDesign
1 {rank: 1}
Array Design
(from Array Design)
1..n {rank: 2}
+arrays
Arrays2ArrayGroup
+arrayGroup
{rank: 3} 0..1
Array Group
barcode : String
arraySpacingX : float
arraySpacingY : float
numArrays : int
orientationMark : String
orientationMarkPosition : enum {top,bottom,left,right}
width : float
length : float
ManufactureLIMSBiomaterial
bioMaterialPlateIdentifier : String
bioMaterialPlateRow : String
bioMaterialPlateCol : String
IdentifierLIMS
+identifierLIMS
{rank: 3} 0..1
DatabaseEntry
(from Description)
accession : String
accessionVersion : String
URI : String
Feature
+feature
1 {rank: 1}
Feature
(from DesignElement)
BioMaterial
+bioMaterial
0..1 {rank: 2}
BioMaterial
(from BioMaterial)
ArrayManufacture Deviations
+arrayManufactureDeviations
{rank: 4} 0..n
ArrayManufactureDeviation
+featureDefects
{rank: 2}
Defects
0..n
FeatureDefect
DefectType
+defectType
1 {rank: 1}
Ontology Entry
(from Description)
category : String
PositionDelta
Substrate Type
+substrateType
{rank: 3} 0..1
Ontology Entry
(from Description)
Fiducials
+fiducials
{rank: 1} 0..n
Fiducial
Distance Unit
+distanceUnit
0..1 {rank: 4}
Distance Unit
(from Measurement)
0..1
{rank: 2}
+distanceUnit
{rank: 1} 0..1
+distanceUnit

Data Organisation — GEO

- Basic Entities:
  - Platform: elements that will be assayed
  - Sample: reference to Platform, measurement in an experiment
  - Series: related samples in an experiment, includes analysis/extracted summary
- Curated: DataSets are grouping of samples for platform. This is further subdivided into GEO Profiles, for each gene in the dataset.
- each has accession numbers with distinguishing prefix
- raw data is available via FTP


DataSet
Gene A
Platform
Sample 3
Sample 2
Sample 1
Profile
Gene A
Sample 1
Sample 2
Sample 3

SOFT File format

```
^SERIES = GSE983
!Series_title = Gene Expression-Based High Throughput Screening:
!Series_geo_accession = GSE983
!Series_status = Public on Jan 30 2004
!Series_submission_date = Jan 21 2004
!Series_pubmed_id = 14770183
!Series_web_link = http://www.broad.mit.edu/cancer/GE-HTS_leuk
!Series_summary = We developed a general approach to small molecu
!Series_summary =

!Series_summary = This data set contains 3 primary patient AML sa
!Series_summary = Keywords = AML
!Series_summary = Keywords = neutrophil
```

Curation — ArrayExpress

- repository: basic check — author submits, curation team checks before adding
- repository: external databases can also be linked to ArrayExpress and monitored by curation team
- warehouse: add biological annotation, including current version in sequence database, gene annotations, gene names added

Curation — GEO

- basic entities: basic curation check
- GEO Datasets: curator assembled. Extracts values from samples, and then grouped into subsets
- GEO Profiles: Sample data from the Datasets are indexed by gene

MAGE-ML

- Microarray Gene Expression - Mark up Language
- automatically derived from UML specification MAGE-OM (Object Model)

Submission

- ArrayExpress
  - online via MIAMEExpress software
  - MAGE-ML pipeline to external database
- GEO
  - web-based forms
  - Simple Omnibus Format (batch file format)
  - MAGE-ML FTP

Data Availability

- Both allow for data download via FTP
- Both allow for searching via web forms
- ArrayExpress provides scripts to set up local version of the repository
- GEO provides some additional data analysis tools online

The Numbers

- NCBI GEO:
  - Platforms: 1652
  - Samples: 50680
  - Series: 2386
  - Datasets: 962
- ArrayExpress:
  - Experiments: 910
  - Arrays: 594
  - Protocols: 4730
  - Hybridizations: 25870

The Numbers — Round II

- CIBEX
  - Experiments : 3
  - Arrays : 5
  - Hybridizations : 448
- Genbank: Over 40 million sequence records

Conclusion

- GEO: store data
- ArrayExpress: access data

References

- *NCBI GEO: mining millions of expression profiles—database and tools* by Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D562-6.
- *ArrayExpress—public repository for microarray gene expression data at the EBI.* H. Parkinson, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. Garcia Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone, and A. Brazma. Nucl. Acids Res., 2005, 33: D553-D555; doi:10.1093/nar/gki056
- *Navigating public microarray databases.* Christopher J. Penkett and Jürg Bähler. Preprint from Comp. Funct. Genom. 2004; 5:471-479., 2004.