::: Accessing to Raw Data and Microarray Databases
::: Accessing raw data and Background Correction

We will focus this part of the course on

✓ **Agilent raw data**: how txt files obtained from Agilent software (FE);

✓ **Microarray databases**: EBI-ArrayExpress and NCBI-GEO;

• **Downloading raw data from NCBI-GEO**: Practice 1.

• **Background subtraction**: Goal and methods;

• **Extracting information from txt files with spreadsheets**: Practice 2.
COMPONENTS OF ANY MICROARRAY EXPERIMENT

- Sample source
- Sample treatments
- RNA extraction protocol
- Labeling protocol

Experiment

- Hybridization protocol

Sample → Hybridization ← Array

- Image
- Scanning protocol
- Software specifications

Raw data

- Array design information
- Location of each element
- Description of each element

- Quantification matrix
- Software specifications

Normalized data

- Control array elements
- Normalization method

Accessing to raw data

Gabriella Rustici, PhD
Microarray Informatics Team
European Bioinformatics Institute
From scanned images to “raw data” text files

Data Workflow

- Design File
- GAL File
- Text File

- Tiff Image

- Feature Extraction Software
  - Grids & Find Spots
  - Rejects Outlier Pixels
  - Flags Feature & Blqgd Outliers
  - Computes Blqgd & Signal Biases
  - Corrects Dye Biases
  - Calculates Log Ratios & Errors

- Agilent .txt files (raw data)

- Agilent GE2, Non-Agilent GE1, Agilent CGH, Agilent Location Analysis, Agilent MicroRNA (v9.5)

- XDR Mode

- Agilent Microarray Scanner

- QC Report
- Grid
- Shape
- Text
- JPEG
- MAGE-ML

Feature Extraction Result Files

Accessing to raw data

UBio
Gene expression data matrix

Samples
Sample annotation

Genes
Gene annotations

Gene expression levels

Accessing to raw data
::: Accessing raw data from Public Databases: GEO

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

Does NCBI do?

NCBI was established in 1988 as a national resource for molecular biology information. NCBI conducts research in molecular biology, develops software for analyzing genome data, and curates biomedical information. For more understanding of molecular biology affecting human health and disease.

More about NCBI...

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources
- Influenza Virus Resource

Protein Clusters

The new Protein Clusters database contains Reference Sequence (RefSeq) protein records that are grouped and annotated by sequence and functional similarity. Source sequences come from the complete genomes of prokaryotes, plasmids, and organelles.

Read more about Protein Clusters.
Gene Expression Omnibus: a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.

Accessing to raw data

::: GEO Datasets and GEO Profiles
### GEO Database Organization

**Platform**
- Platform records are supplied by submitters.
- A Platform record defines the list of elements that may be detected and quantified in an experiment (e.g., cDNAs, oligonucleotide probesets). Each Platform record is assigned a unique and stable GEO accession number (GPLxxx).
- A Platform may reference many Samples that have been submitted by multiple submitters.
- **Example Platform record**

<table>
<thead>
<tr>
<th>Platform</th>
<th>Text description of the array</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Text tab-delimited table of the array template</td>
</tr>
</tbody>
</table>

**Sample**
- Sample records are supplied by submitters.
- A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it. Each Sample record is assigned a unique and stable GEO accession number (GSMxxx).
- A Sample entity must reference only one Platform and may be included in multiple Series.
- **Example Sample record**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Text description of a biological sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Text tab-delimited table of processed hybridization result (may optionally include raw data columns)</td>
</tr>
<tr>
<td></td>
<td>Original raw data file</td>
</tr>
</tbody>
</table>

**Series**
- Series records are supplied by submitters.
- A Series record links together a group of related Samples and provides a focal point and description of the whole study. Series records may also contain tables describing extracted data, summary conclusions, or analyses. Each Series record is assigned a unique and stable GEO accession number (GSExxx).
- **Example Series record**

<table>
<thead>
<tr>
<th>Series</th>
<th>Text description of the overall experiment</th>
</tr>
</thead>
</table>

---

**Datasets**

- **Accessing to raw data**

---
### GEO Datasets and GEO Series

#### Summary

The GEO Datasets and GEO Series page provides access to gene expression data, allowing users to explore and analyze datasets related to various biological and medical research topics. The datasets are organized in a table format, with columns for **Disease, Year, and Country**. Each dataset entry includes links to more detailed information and downloadable files.

#### Key Features

- **Gene Expression Data**: Access to gene expression datasets from various experiments and conditions.
- **Search Functionality**: Ability to search for datasets by keywords, enabling easy filtering.
- **Download Options**: Options to download datasets in various formats, facilitating further analysis and integration into research workflows.

#### Example Dataset Entries

1. **GSE10010**:
   - **Disease**: Colon cancer
   - **Year**: 2005
   - **Country**: USA
   - **Summary**: Interleukin-22 (IL-22) treatment of colon cancer cells was found to induce a protective response against viral infection. This dataset includes gene expression data for cells treated with IL-22 and control conditions.

2. **GSE51983**:
   - **Disease**: Breast cancer
   - **Year**: 2014
   - **Country**: USA
   - **Summary**: The dataset includes gene expression profiles from breast cancer patients treated with different therapeutic regimens, providing insights into the molecular mechanisms underlying cancer progression.

3. **GSE2872**:
   - **Disease**: Whole Mouse Genome Microarray
   - **Year**: 2006
   - **Country**: USA
   - **Summary**: This dataset provides gene expression profiles across various mouse tissues, useful for understanding gene expression patterns and regulatory networks in mouse models.

### Table of Datasets

<table>
<thead>
<tr>
<th>Disease</th>
<th>Year</th>
<th>Country</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon cancer</td>
<td>2005</td>
<td>USA</td>
<td>Interleukin-22 (IL-22) treatment of colon cancer cells induces a protective response against viral infection.</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>2014</td>
<td>USA</td>
<td>Gene expression profiles from breast cancer patients treated with different therapeutic regimens.</td>
</tr>
<tr>
<td>Whole Mouse Genome Microarray</td>
<td>2006</td>
<td>USA</td>
<td>Gene expression profiles across various mouse tissues.</td>
</tr>
</tbody>
</table>
What is the difference between a Series and a DataSet?

A GEO Series (GSExxx) is an original submitter-supplied record that summarizes an experiment.

These data are reassembled by GEO staff into GEO Datasets (GDSxxx).

A DataSet represents a collection of biologically- and statistically-comparable Samples processed using the same Platform. Information reflecting experimental variables is provided through DataSet subsets.

Both Series and DataSets are searchable using the Entrez GEO DataSets interface, but only DataSets form the basis of GEO's advanced data display and analysis tools including gene expression profile charts and DataSet clusters. Not all submitted data are suitable for DataSet assembly and we are experiencing a backlog in DataSet creation, so not all Series have a corresponding DataSet record(s).
**Series (GSE)**

- **Title**: IL-22 plays an indispensable role in early host defense against attaching and effacing (A/E) bacterial pathogens
- **Organism(s)**: *Mus musculus*
- **Experiment type**: Expression profiling by array
- **Summary**: Infection by attaching and effacing (A/E) pathogens poses a serious threat to public health, as was highlighted by the recent outbreak of *E. coli* O157:H7 infection in the United States. Here, by using a murine A/E pathogen, Citrobacter rodentium, we demonstrate that C. rodentium infection is lethal to IL-22(-/-) mice within two weeks. IL-22, in the early phase of infection, is indispensable for preventing the invasion of bacteria through the intestinal epithelium, and thereby preventing systemic spread and mortality. We also show that IL-23 is required for the early induction of IL-22 during C. rodentium infection. Finally, our data suggest that IL-22 exerts its function by triggering the innate immune responses of colonic epithelial cells, especially though the induction of anti-microbial proteins, RegIIIγ and RegIIIγ.
- **Keywords**: treatment comparison, control or IL-22-treated mouse colon in triplicate.

**Dataset (GDS)**

- **Title**: Interleukin-22 effect on colon cultures
- **Summary**: Analysis of C57BL/6j colon cultures treated with 10 ng/ml of interleukin-22 (IL-22) for 24 hours. IL-22, a member of the IL-10 family of cytokines, can induce a marked antimicrobial response in vitro. Results provide insight into the molecular basis of IL-22 induced host defense mechanisms.
- **Platforms**: GPL2872 Agilent-0126H Whole Mouse Genome G4122A (Feature Number version)
- **Samples**: GDS252890 colon_control_rep1
  GDS252931 colon_control_rep2
  GDS252932 colon_control_rep3

**Download family**

<table>
<thead>
<tr>
<th>Format</th>
<th>SOFT</th>
<th>MINML</th>
<th>R</th>
<th>TXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplementary file</td>
<td>SOFT</td>
<td>MINML</td>
<td></td>
<td>TXT</td>
</tr>
<tr>
<td>GSE10010_RAW.tar</td>
<td>74.3 MB</td>
<td>(ftp/http)</td>
<td>TARGA</td>
<td>(of TXT)</td>
</tr>
</tbody>
</table>

New data provided as supplementary file Processed data included within Sample table.
## Platform GPL2872

<table>
<thead>
<tr>
<th>Status</th>
<th>Public on Sep 24, 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Agilent-012694 Whole Mouse Genome G4122A (Feature Number version)</td>
</tr>
<tr>
<td>Technology type</td>
<td>in situ oligonucleotide</td>
</tr>
<tr>
<td>Distribution</td>
<td>commercial</td>
</tr>
<tr>
<td>Organism(s)</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>Manufacturer</td>
<td>Agilent Technologies</td>
</tr>
<tr>
<td>Catalog number</td>
<td>G4122A</td>
</tr>
</tbody>
</table>

### Description

Designed to truly represent all known genes in the mouse genome and their resulting transcripts, Agilent's Whole Mouse Genome Oligo microarray is comprised of 41,534 50-mer oligonucleotide probes representing over 41,000 mouse genes and transcripts. Content for this microarray was generated from leading public databases, including RefSeq, RIKEN, NIA, Ensembl, RIKEN, UCSC GoldenPath, and Unigene.

Arrays of this design have barcodes that begin with 16012694 or 2512694.

**Orientation:**

Features are numbered numbered Left-to-Right, Top-to-Bottom as scanned by an Agilent scanner (barcode on the left, DNA on the back surface, scanned through the glass), matching the FeatureNum output from Agilent's Feature Extraction software.

The ID column represents the Agilent Feature Extraction feature number.

Rows and columns are numbered as scanned by an Axon Scanner (barcode on the bottom, DNA on the front surface).

To match data scanned on an Axon scanner, use the RefNumber column contained in the Agilent-provided GAL file as the ID.REF column in sample submissions.

*** A different version of this platform with the Agilent Probe names in the ID column is assigned accession number GPL7042.
Probe annotation is found here:

<table>
<thead>
<tr>
<th>Data table header descriptions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Agilent feature number</td>
</tr>
<tr>
<td>COL</td>
<td>Column</td>
</tr>
<tr>
<td>ROW</td>
<td>Row</td>
</tr>
<tr>
<td>NAME</td>
<td>Name</td>
</tr>
<tr>
<td>SPOT_ID</td>
<td>Spot identifier</td>
</tr>
<tr>
<td>CONTROL_TYPE</td>
<td>Control type</td>
</tr>
<tr>
<td>REFSEQ</td>
<td>RefSeqAccession</td>
</tr>
<tr>
<td>GB_ACC</td>
<td>GenBankAccession</td>
</tr>
<tr>
<td>GENE</td>
<td>Entrez Gene ID</td>
</tr>
<tr>
<td>GENE_SYMBOL</td>
<td>Gene Symbol</td>
</tr>
<tr>
<td>GENE_NAME</td>
<td>Gene Name</td>
</tr>
<tr>
<td>UNIGENE_ID</td>
<td>UniGeneID</td>
</tr>
<tr>
<td>ENSEMBL_ID</td>
<td>EnsemblID</td>
</tr>
<tr>
<td>TIGR_ID</td>
<td>TIGRID</td>
</tr>
<tr>
<td>ACCESSION_STRING</td>
<td>Accession String</td>
</tr>
<tr>
<td>CHROMOSOMAL_LOCATION</td>
<td>Chromosomal Location</td>
</tr>
<tr>
<td>CYTOSBAND</td>
<td>Cyto band</td>
</tr>
<tr>
<td>DESCRIPTION</td>
<td>Description</td>
</tr>
<tr>
<td>GO_ID</td>
<td>GoIDs</td>
</tr>
<tr>
<td>SEQUENCE</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data table</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>COL</td>
</tr>
<tr>
<td>1</td>
<td>103</td>
</tr>
<tr>
<td>2</td>
<td>103</td>
</tr>
<tr>
<td>3</td>
<td>103</td>
</tr>
<tr>
<td>4</td>
<td>103</td>
</tr>
<tr>
<td>5</td>
<td>103</td>
</tr>
<tr>
<td>6</td>
<td>103</td>
</tr>
<tr>
<td>7</td>
<td>103</td>
</tr>
<tr>
<td>8</td>
<td>103</td>
</tr>
<tr>
<td>9</td>
<td>103</td>
</tr>
<tr>
<td>10</td>
<td>103</td>
</tr>
<tr>
<td>11</td>
<td>103</td>
</tr>
<tr>
<td>12</td>
<td>103</td>
</tr>
<tr>
<td>13</td>
<td>103</td>
</tr>
<tr>
<td>14</td>
<td>103</td>
</tr>
</tbody>
</table>