

Big Data Mining in Healthcare

REPORT

arrayExpress DataSet conversion for
R/Bioconductor

Submitted by:

Rakesh Rawat - MT17046

Kuldeep Singh - MT17022

Abhishek Aggarwal - MT17141

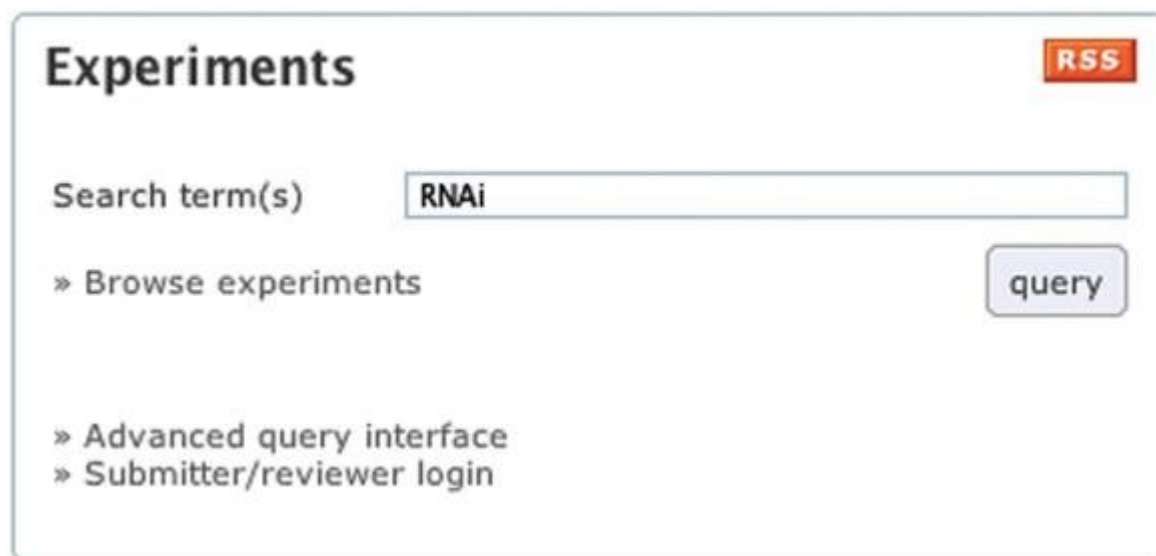
1. ArrayExpress :

ArrayExpress is a public database for storing and providing access to high throughput functional genomics data. ArrayExpress consists of two components specialized for distinct purposes—the ArrayExpress Repository of publicly available archived experimental data and the ArrayExpress Data Warehouse of gene expression profiles.

1.1 ArrayExpress Repository :

ArrayExpress is one of the three databases recommended by the MGED society for depositions of publication related microarray data the other two being Gene Expression Omnibus and CiBEX. ArrayExpress provides the means to store pre-publication data confidentially whilst allowing access to authorized users such as journal editors and referees. The data are made publicly available upon publication of the paper to which they relate.

A new ArrayExpress experiment browse and query interface was released in 2006. It allows the user to browse the entire content of the database in a summary view or query public datasets using free text and displays the query results in a summary view of up to 500 experiments per page which can be sorted by name, accession number and load date, and filtered by array design, species, date or availability of raw and processed data.



The image shows a web interface titled "Experiments" in a large, bold, black font. In the top right corner, there is a red rectangular button with the text "RSS" in white. Below the title, there is a search section with the label "Search term(s)" in a grey font. To the right of this label is a text input field containing the text "RNAi". Below the search section, there are three links in a grey font: "» Browse experiments", "» Advanced query interface", and "» Submitter/reviewer login". To the right of these links is a grey rectangular button with the text "query" in a darker grey font.

Figure 1: ArrayExpress experiment query form. Queries on experiment properties: organisms, author's names, array types or accession numbers are supported.

E-TABM-65	Comparative genomic hybridization of cell lines from 9 different cancer tissue of origin types (Breast, Central Nervous System, Colon, Leukemia , Melanoma, Non-Small Cell Lung, Ovarian, Prostate, Renal) from NCI-60 panel	60	Human samples	2005-11-23		
Accession number:	+ E-TABM-65 [open advanced view]					
Title:	Comparative genomic hybridization of cell lines from 9 different cancer tissue of origin types (Breast, Central Nervous System, Colon, Leukemia , Melanoma, Non-Small Cell Lung, Ovarian, Prostate, Renal) from NCI-60 panel					
Data:	+ View detailed data retrieval page... + FTP server direct link...					
Array:	UCSF_Gray_R_sagepnt Onco6AC DS00 (A-MEXP-385) + A-MEXP-385					
Experiment design:	+ .png + .svg + .xls					
Hybridizations:	+ .xls					
Protocols:	+ Experimental protocols					
Contact:	Kimberly Busssey					
Publication:	Kimberly J. Busssey; Kwei Chin; Samir Lababidi; Mark Raimann; William Reinhold; Wen-Lin Kuo; Foad Gwadry; Ajay; Hoonin Kuroos-Mehr; Jane Fridlyand; Ajay Jain; Colin Collins; Satoshi Nishizuka; Giovanni Tononi; Anna Rozhnika; Kristen Gehlhaus; Jan Kirochi; Dominic Scudiero; Joe Gray; John Weinstein. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 Cell Line Panel. Mol Cancer Ther.					
Design type(s):	comparative genomic hybridization					
Description:	Array comparative genomic hybridization characterization and comparison of cell lines from 9 different cancer tissue of origin types (Breast, Central Nervous System, Colon, Leukemia , Melanoma, Non-Small Cell Lung, Ovarian, Prostate, Renal) from NCI-60 panel.					

Figure 2: A detailed view of an experiment from the repository. Strings matching the query terms are highlighted in yellow.

1.2 ArrayExpress Warehouse :

The ArrayExpress Data Warehouse currently contains 1 500 000 gene expression profiles from >2500 hybridizations. It is currently populated with gene expression data from ArrayExpress. Selection for inclusion is on the basis of MIAME compliance, a curator's assessment of quality of annotation of the samples and arrays.

The query interface to the ArrayExpress Data Warehouse allows the user to find gene expression profiles by gene name or identifier, database accession number (e.g. UniProt) or other annotation such as Gene Ontology terms. If a query returns more than one gene (for instance, a query for BRCA will match BRCA1 and BRAC2), then a list of matches is provided and the user can view the properties of these genes before selecting appropriate ones.

Expression Profiles

Gene(s)

Species

Description

Figure 3: Gene expression profiles query form. Queries on gene properties: names, accession numbers, synonyms and gene ontology terms are supported.

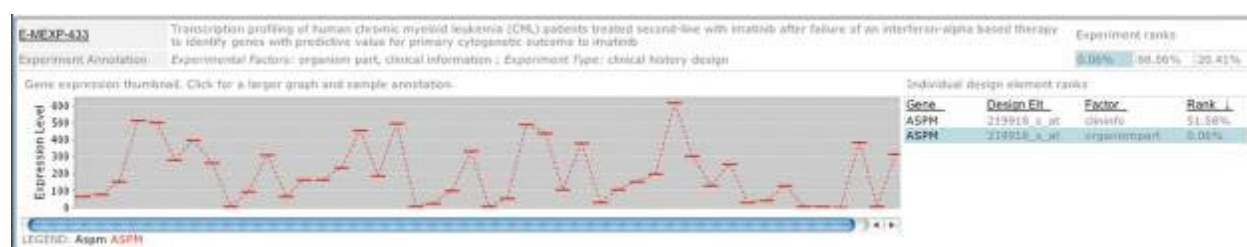


Figure 4: A view of gene expression for gene 'aspm' within a single experiment in the data warehouse.

2. MIAME (Minimum Information About a Microarray Experiment):

The MIAME guidelines outline the minimum information that should be included when describing a microarray experiment. Many journals and funding agencies require microarray data to comply with MIAME.

The six most critical elements contributing towards MIAME are -

- The raw data for each hybridization (e.g., CEL or GPR files)
- The final processed (normalized) data for the set of hybridizations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
- The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
- The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridizations are technical, which are biological replicates)
- Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
- The essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data)

3. MAGE-TAB (MicroArray Gene Expression Tabular) :

In order to provide a common platform for sharing characterization data within the research community, the Microarray Gene Expression Data (MGED) Society developed the Minimum Information About a Microarray Experiment (MIAME) standard. MIAME describes the data and accompanying metadata that investigators must provide so that the experiment can be reproduced and the results can be interpreted in light of the experimental conditions. MAGE-TAB (MicroArray Gene Expression Tabular) uses simple spreadsheet-based format for representing primary data and associated metadata. MAGE-TAB specification is based on the Microarray and Gene Expression Object Model (MAGE-OM).

MAGE-TAB Files : To capture experiment details and the relationships between related data files (i.e. data files from different stages of sample data as protocols are continuously applied to it) TCGA uses the MAGE-TAB standard. MAGE-TAB files are tab-delimited text files that model data in the form of columns and rows and is able to capture complex experimental relationships such as an entire study using multiple assays.

MAGE-TAB format uses three different types of files to capture information about an experiment

File Type	File Extension	Platform	Description
Array Design Format (ADF)	.adf	mage-tab	Defines each array type used. An ADF file describes the design of an array, e.g., what sequence is located at each position on an array and what the annotation of this sequence is. An ADF may exist in the MAGE-TAB archive or through the Data Portal on the Platform Design page.
Investigation Description Format (IDF)	.idf	mage-tab	Provides general information about the investigation, including its name, a brief description, the investigator's contact details, bibliographic references, and free text descriptions of the protocols used in the investigation.
Sample and Data Relationship Format (SDRF)	.sdrf	mage-tab	Describes the relationships between samples, arrays, data, and other objects used or produced in the investigation, and providing all MIAME information that is not provided elsewhere. In TCGA SDRF files, a row represents an analyzed element (often an aliquot) in its most basic electronic form (i.e. raw data file) and the production of higher-level data files (Level 2 and 3) as protocols (e.g. normalization) are applied to the file and its derivatives. These protocols correspond to those listed in the IDF.

The most important concept behind the SDRF is the investigation design graph, which is a directed acyclic graph (DAG), where nodes correspond to biomaterials (e.g., samples, RNA extracts, labeled cDNA, etc.) or data objects (e.g., raw or normalized data files), and arcs correspond to the relationships

between these objects. Biomaterials have properties, some of which can be experimental factors. Attributes can be attached to nodes and to arcs to describe biomaterial or data properties, e.g., sample descriptions attached to sample nodes, protocol references attached to edges, raw data-files attached to assays. Attributes can be pointers to some longer descriptions or external objects, e.g., protocols described in the IDF file.

The investigation design graph could be encoded in various ways, for instance using the graph mark-up language GML.

Here we use a tabular format for the following reasons:

1. The observation that large investigation designs typically have a regular structure, i.e., the same subgraph is repeated many times (possibly with well defined modifications); moreover, the replicated structure is simple. This observation was supported by analysis of the structure of over 1,000 different investigations in the ArrayExpress database.
2. The degree of nodes in these graphs (i.e., the number of incoming and outgoing edges for a node), is small (most often 1 to 4), except for a few specific nodes which are related 'reference' samples or extracts (e.g., 'Reference LE' in Figure 24), or common source nodes (e.g., Figure 37).
3. The observation that DAGs which correspond to commonly used investigation designs have a property that their nodes can be grouped in consecutive layers, i.e., the source nodes (the nodes in the DAG which do not have entering edges) are in layer 1, the nodes that are connected to source nodes by an edge are in layer 2, etc. Furthermore, the grouping can be done so that each layer only contains objects of the same type, e.g., for the graph in Figure 8(a), we have source layer 1, sample layer 2, extract layer 3, labeled extract layer 4 and assay layer 5.

4. Similar tabular formats have been used successfully in the biosciences and are familiar to many practitioners.

Once a DAG of a regular structure has been represented in such a layered fashion, it is natural to encode it as a tab-delimited

file (a 'spreadsheet' in the broad sense of this word). Each column in the spreadsheet corresponds to a layer in the DAG, while

each row corresponds to a path in the graph from one of the source nodes, to one of the 'sink' nodes.

An Introduction to Bioconductor ExpressionSet Class

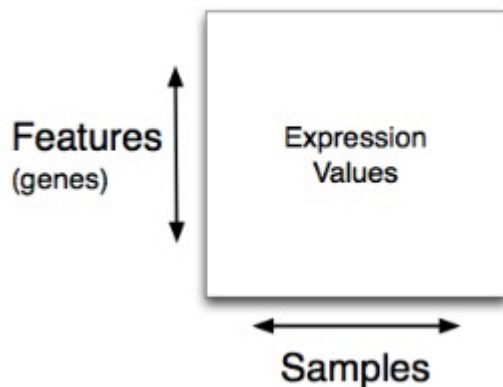
- Installing and loading of biobase package
- Expression set from CEL and other files
 - Readaffy for CEL
 - If resulting object not in expressionSet, use library(convert) to convert into expressionset object.
- Expression Set from scratch
 - Assay data
 - Matrix of expression values(F X S)
 - F: Features on chip
 - S: # Samples
 - Read.table - command for loading content of file
 - First two lines create file path,eliminate them
 - Phenotypic data
 - Information about sample(Covariates)
 - (S X V) where V is # covariates
 - Covariates may include numeric or factors(use colClasses)
 - #rows(phenotype data) = #cols(Expression data)
 - **AnnotatedDataFrame** : Conveniently stores and manipulates phenotype data
 - Annotations and feature data

- Metadata for features is important
- Annotate and annotationDbi package - Data manipulation tools for metadata packages
- Experiment description
 - Basic description about experiment (investigator, lab) recorded by creating MIAME object.
- Assembling ExpressionSet
 - Assemble all created objects before into one expressionSet
- For basic operations on eSet, refer

<https://www.bioconductor.org/packages/3.7/bioc/vignettes/Biobase/inst/doc/BiobaseDevelopment.pdf>

4. **Bioconductor Class :**

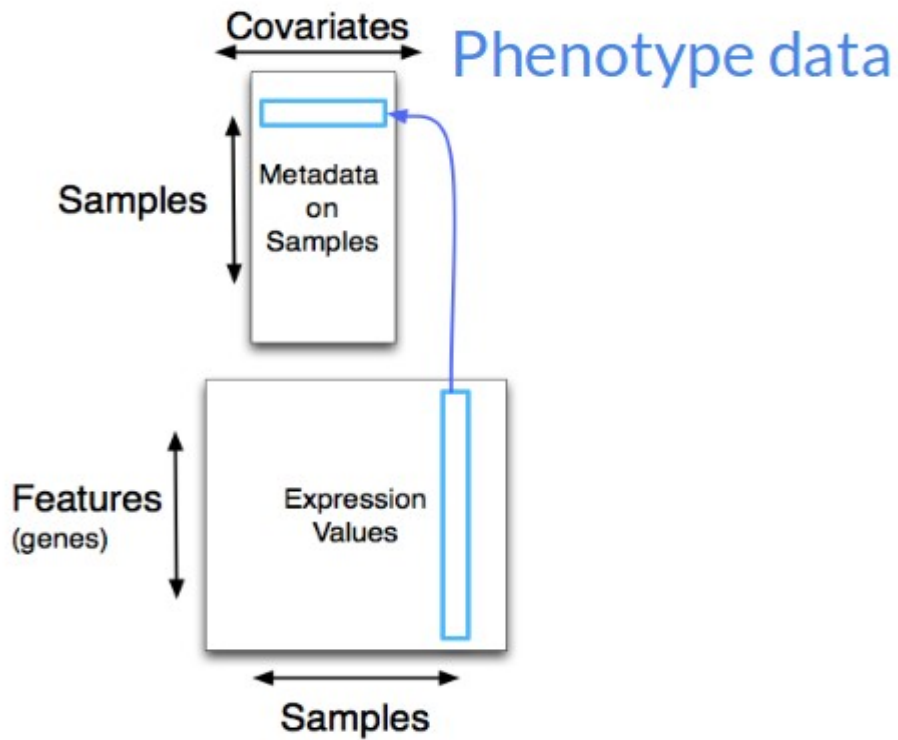
MAGE-TAB -----> eSet -----(One color)-----ExpressionSet
 -----(2 color)-----N-channel set
 -----(Affymetrix Genechips)-----Affybatch



Expression Matrix

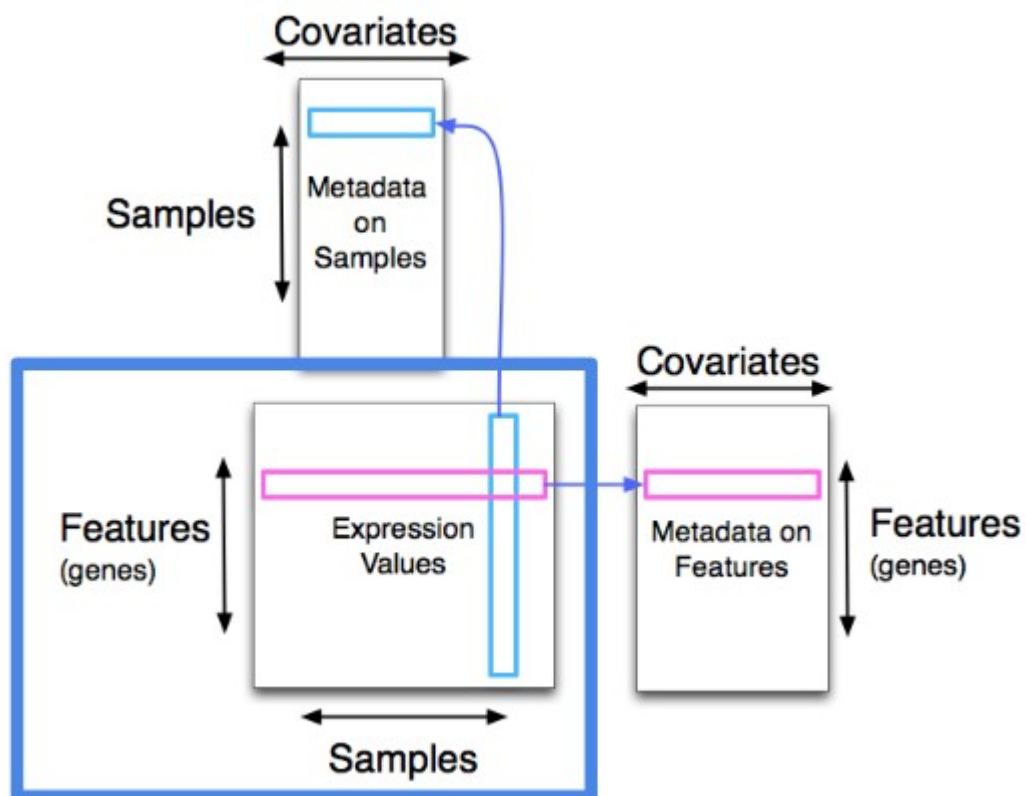
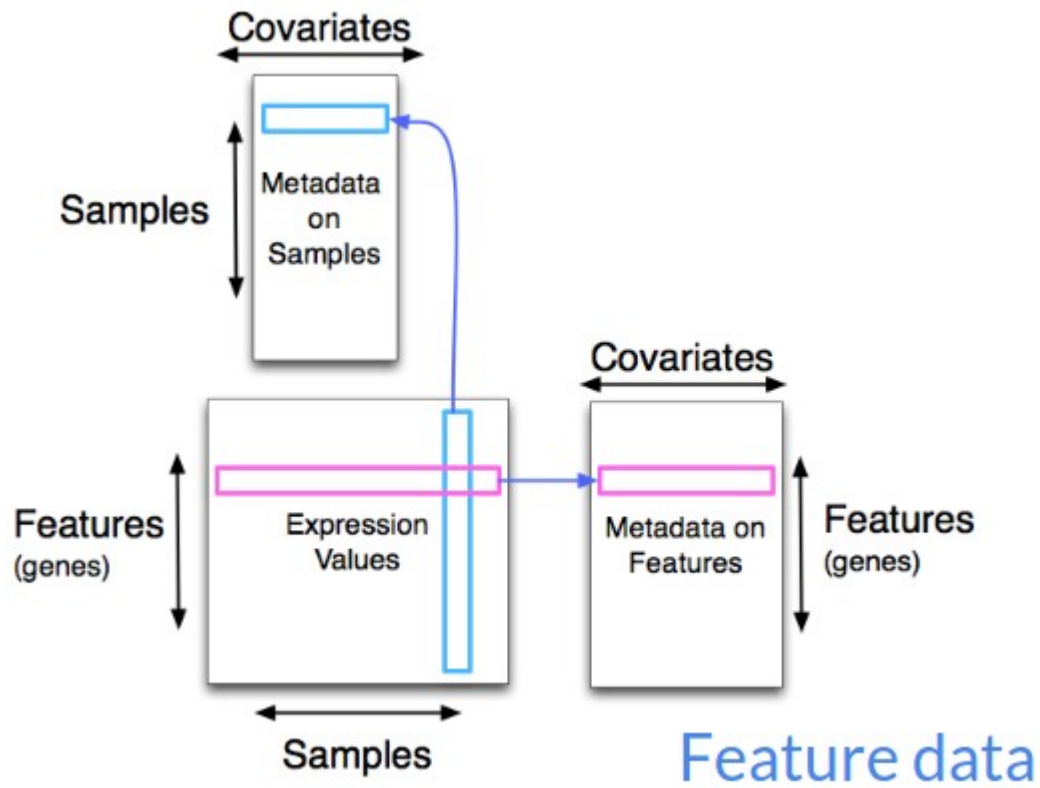
Reference:

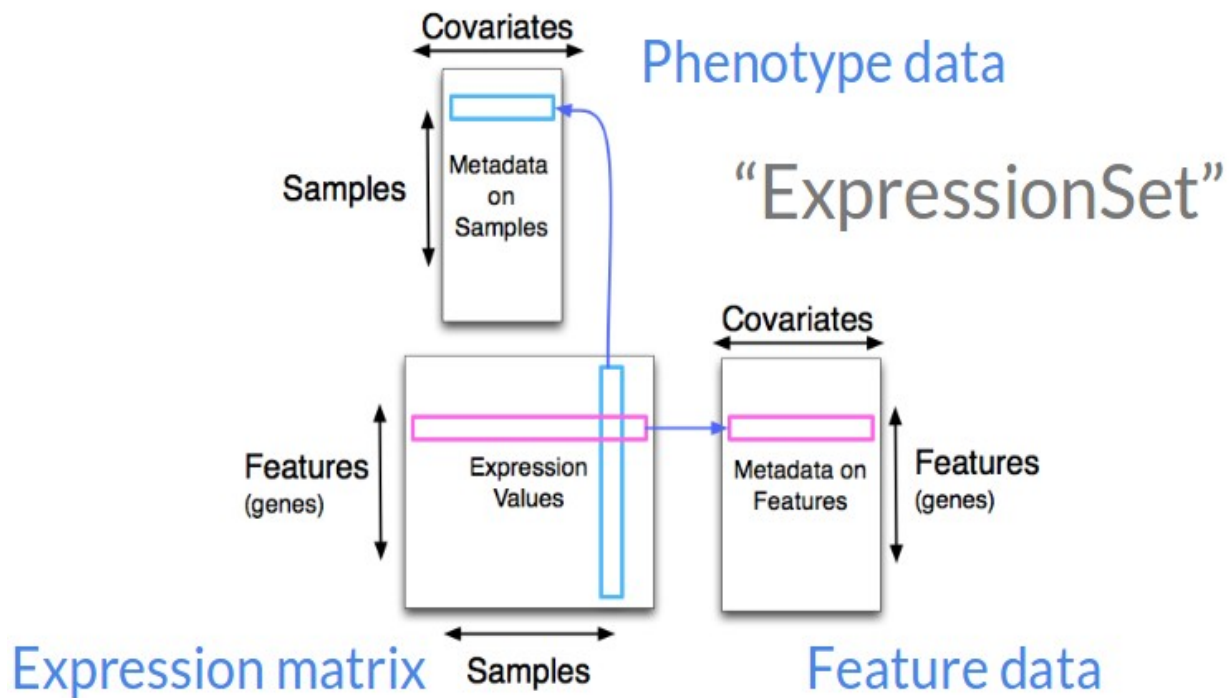
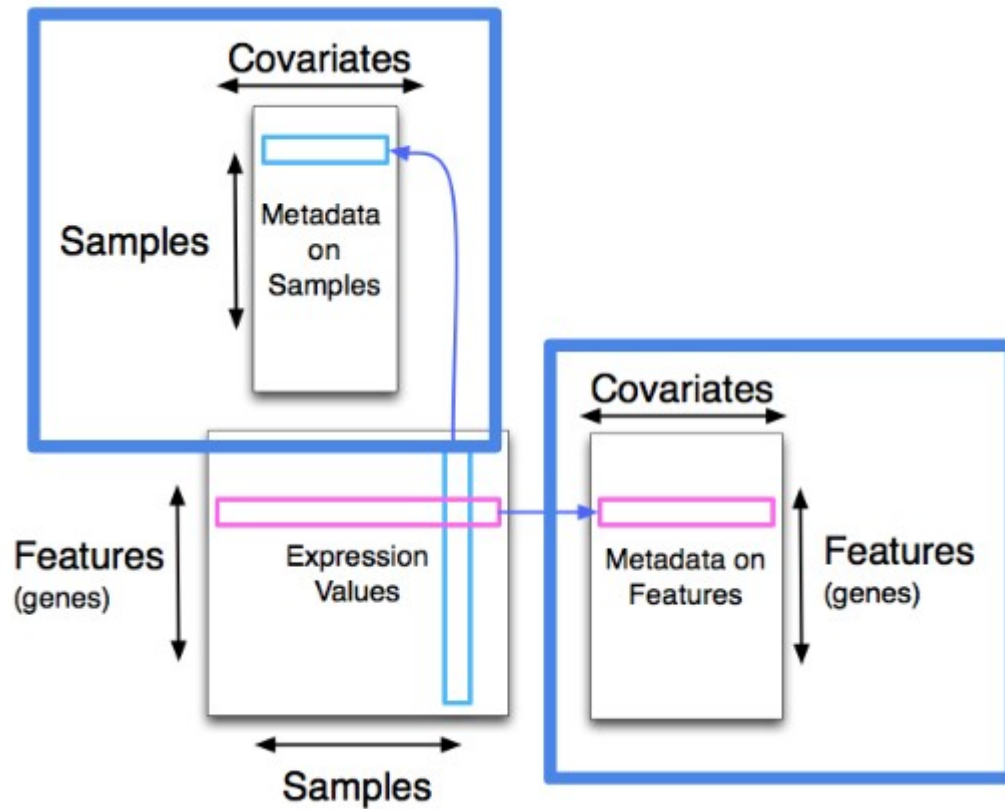
http://kasperdanielhansen.github.io/genbioconductor/pdf/BiocIntro_ExpressionSet_Overview.pdf

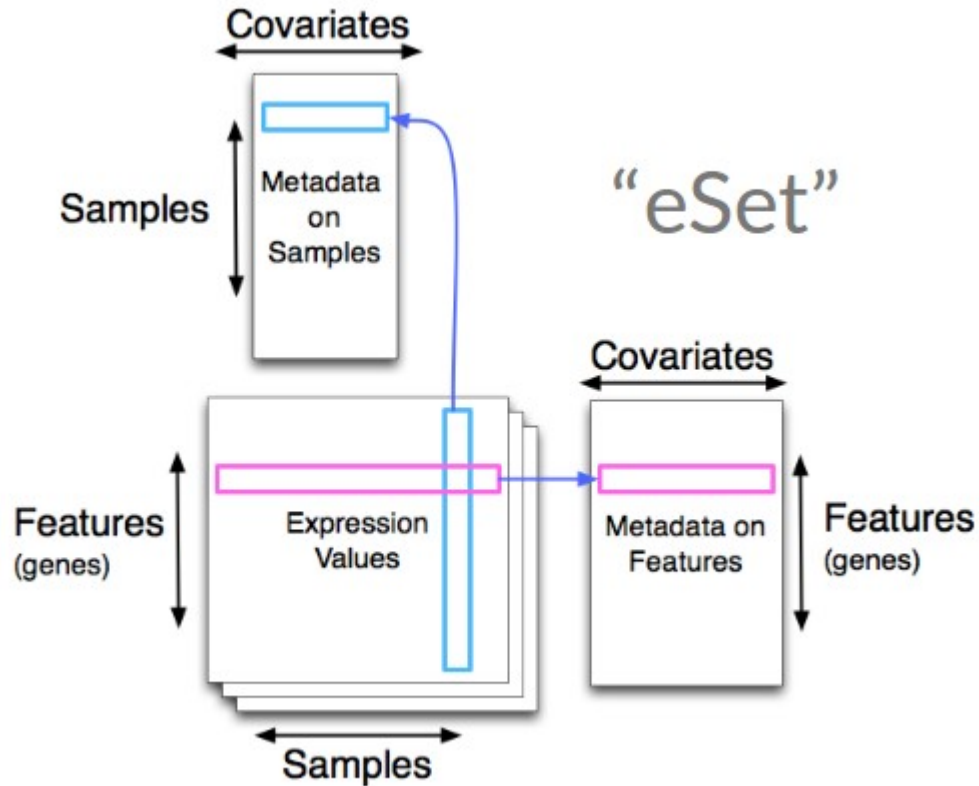


Reference:

http://kasperdanielhansen.github.io/genbioconductor/pdf/BiocIntro_ExpressionSet_Overview.pdf







References:

- <https://academic.oup.com/bioinformatics/article/25/16/2092/204750>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1716725/>
- <https://www.ncbi.nlm.nih.gov/geo/info/MIAME.html>
- <https://wiki.nci.nih.gov/display/TCGA/MAGE-TAB>
- <https://www.bioconductor.org/packages/3.7/bioc/vignettes/Biobase/inst/doc/BiobaseDevelopment.pdf>
- <https://www.bioconductor.org/packages/3.7/bioc/vignettes/Biobase/inst/doc/ExpressionSetIntroduction.pdf>
- <https://www.bioconductor.org/packages/3.7/bioc/vignettes/affy/inst/doc/affy.pdf>
- <http://kasperdanielhansen.github.io/genbioconductor/>