USC Department Of Translational Genomics
Keck Medicine of USC

TRGN 527: Applied Data Science and Bioinformatics

UNIT III. Supervised Statistical Tests

Week 6 - Lecture 1 – Case Study

Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S. | **Assistant Professor**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

David W. Craig, Ph.D. | **Professor and Vice Chair**

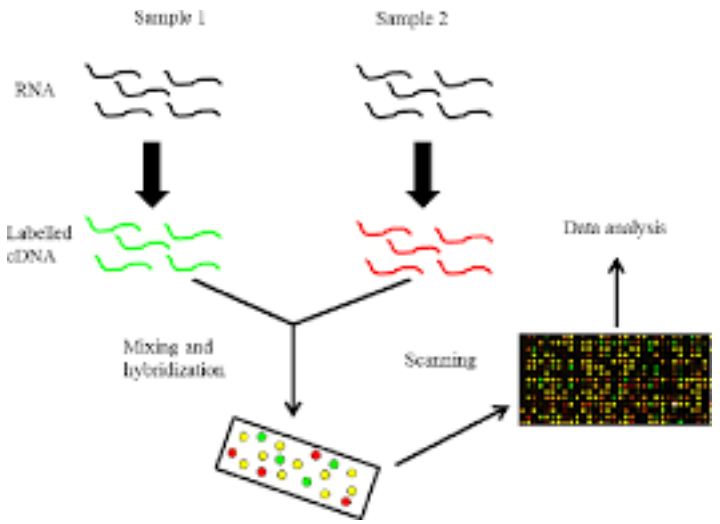
Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Co-Director, Institute of Translational Genomics

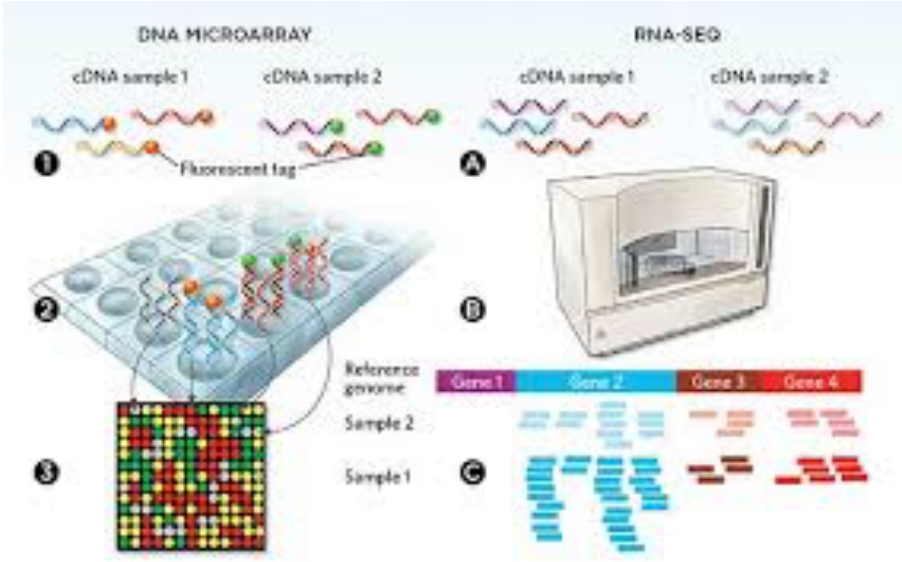
Topics

- Gene expression microarrays, Microarray analysis



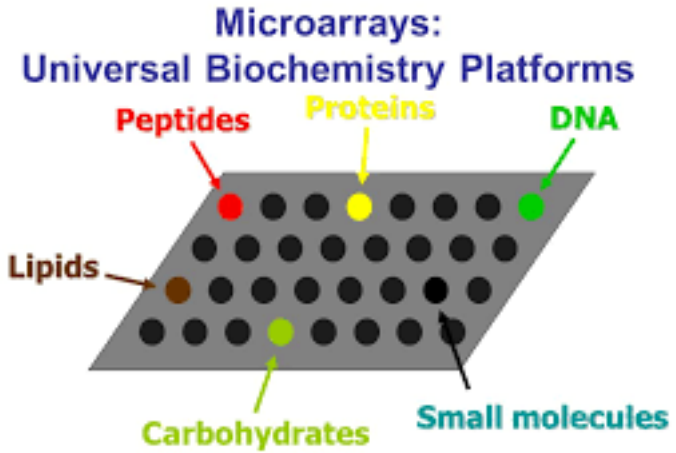
Gene expression microarrays

- Microarray experiments have dominated the genomic and transcriptomics for a long decade.
- Tremendous work has been dedicated to the perfection of algorithms and statistical evaluation of microarray data.
- Gene expression microarray data is still generated where the instruments are readily available.
- A vast microarray data is available from databases such as Gene Expression Omnibus (GEO) or array express.
  - These are excellent sources for meta-analysis, for data mining purposes or as reference datasets for NGS experiments.



Microarray analysis: probes and samples

- Microarrays were the first high-throughput methods to measure the expression of many genes in parallel.
- They are excellent examples of the accelerating development of experimental methods in molecular biology.
- The first microarray-based experiment was pubished in 1995, with the parallel measurement of only 45 Arabidopsis genes (Schena et al. 1995).
- The company Affymetrix extablished by researchers was the one developing the technology to a commercial product with high success.
- Bioconductor offers the best support for Affymetrix data.
- The bioinformatics skills of analyzing microarray data are still essential in high-throughput transcriptomics.



Experimental Background

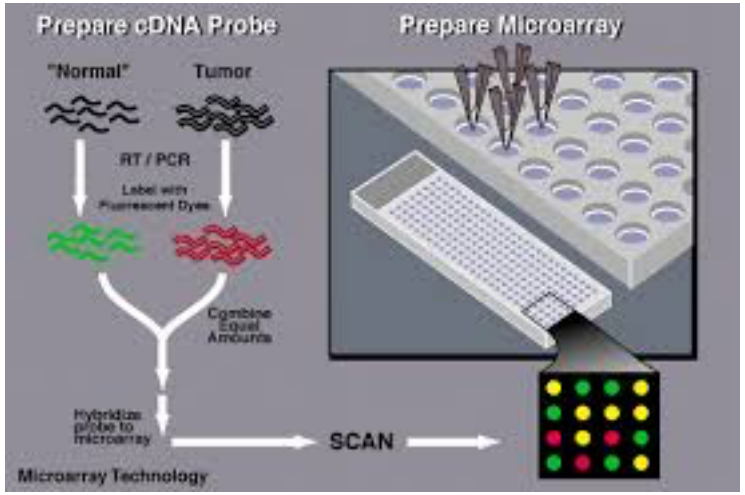
- The measuring of gene expression with microarray technologies requires platform-specific experimental design.
- Most often, RNA is isolated from the biological samples and reverse the transcribed to cDNA.
- There are two, rather different technologies involved here.
  - In classic two channel (or two-color) experiments, the cDNA isolated from two samples are labeled with different fluorescent cyanine dyes, usually by Cy3 (green) and Cy5 (red). The labeled samples are mixed and hybridized to the same array.
    - The data coming from here are Cy3/Cy5 expression ratio.
  - The more modern single channel (or single color) technology is used mostly by Affymetrix arrays and Illumina Bead Chips.
    - Here internal controls assure the specificity of hybridization, and a single dye is used to measure the expression levels.
    - The data coming from these experiments give the absolute expression levels that are more suitable for meta-analyses.
- The primary data coming from the instrument is an image.
- If there are internal control probes (like the MM probes), they have to be combined to provide a single expression or CY3/Cy5 value for each probe.
- The statistical comparison of the differences among the samples from distinct experimental conditions can be performed.
- The goal here is to find differential gene expression among conditions, such as genes over or under expression in one sample group.

Archiving and Publishing microarray data

- Most international research journals do not accept studies using microarray experiments without depositing the relevant data to pubic databases.
- During the past decade, two databases emerged as trusted repositories for these datasets.
  - GEO, established in 2000, stores high-throughput molecular abundance data, including, but not limited to, microarray data.
  - ArrayExpress at the European Bioinformatics Institute (EBI) focuses originally on microarray data, although they also accept RNA-seq recently.
    - This dataset was established in 2002 and if contains fewer data entries than GEO.

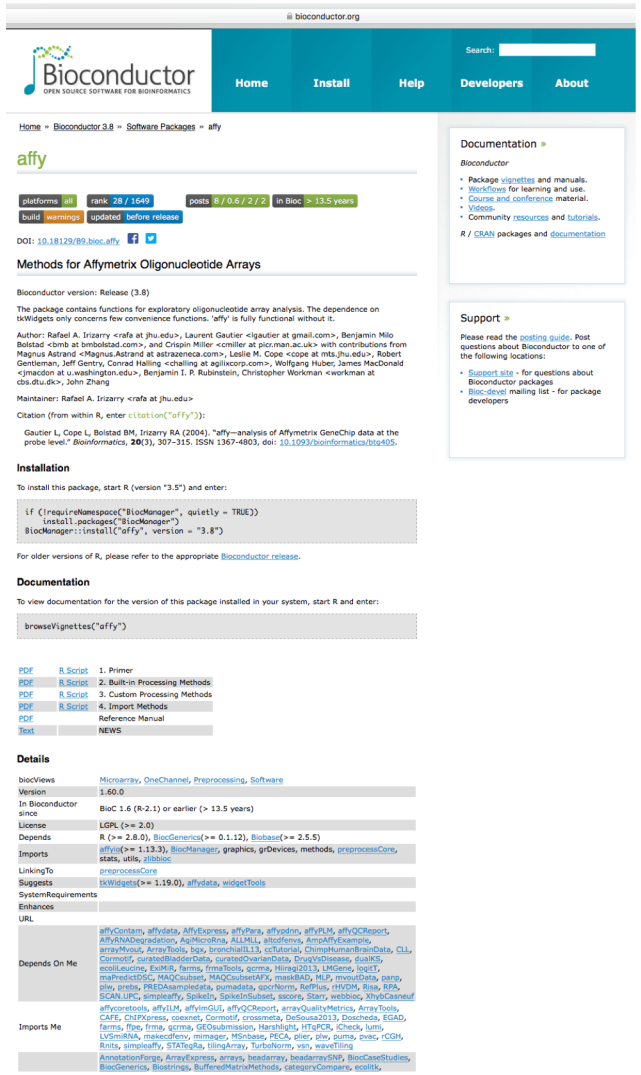
Data preprocessing

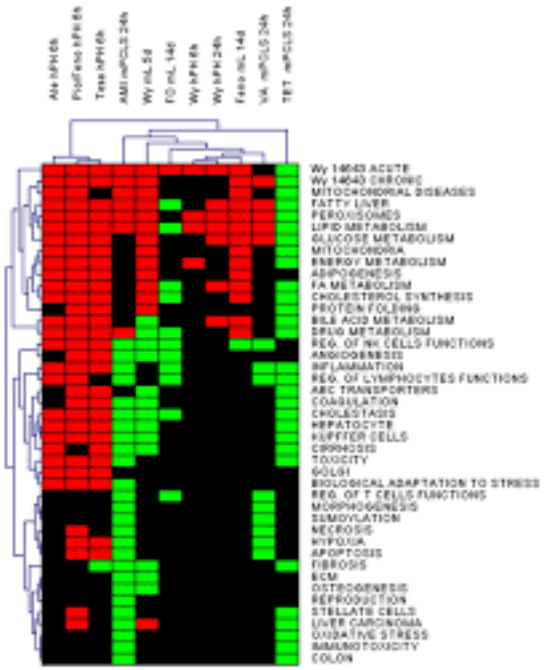
- Raw data coming from microarray reader instruments need some initial treatment to be ready for statistical analysis on probe or gene level.
- These data preprocessing steps are essential for making the measurements coming from different arrays comparable.
- An incredible amount of bioinformatics research was invested into microarray data analysis for making it usable in different experimental scenarios.
- Databases offer microarray data in multiple formats.
  - Raw data are available as simple matrices in a text file for the Agilent and Illumina platforms or as binary CEL files for Affymetrix platforms.
  - Normalized data can be accessed in many formats, including text files or SOFT formatted files also holding some annotation information.



Data preprocessing

- Bioconductor offers several packages to access raw data from the major platforms.
- For Affymetrix data:
  - affy (Gautier et al. 2004)
  - Affycoretools (MacDonald 2008)
- For Illumina data:
  - lumi (Du, Kibbe and Lin 2008)
  - Beadarray (Dunning et al. 2010)
    - Can be applied while there is agilp (Chain et al. 2010) for Agilent.





Accessing data from CEL files

- We will analyze raw Affymetrix array data from GEO.
- First download the raw CEL files (Microarray samples zip file from Blackboard).
- Gzip is a file compressor that prepares smaller files than zip.
- However it is not necessary to un-compress the file, for this learning-analysis, it is recommended to un-compress the file and copy the samples to your R-working-directory.

Rmarkdown_TRGN599_Week_6_Lecture_1_Case_Study_2_Part_1

Enrique I. Velazquez Villarreal, MD, PhD, MPH, MS

2/7/2019

Accessing data from CEL files

```
getwd()
```

```
## [1] "/Users/enriquevelazquez/Documents/R_working_directory"
```

```
# Where the CEL files are - Type the line below directly in your R Console.
#setwd('/Users/enriquevelazquez/Documents/R_working_directory/TRGN599_CelFiles/Microarray_samples')
```

Accessing data from CEL files

Installing Affy package ()

```
# if (!requireNamespace("BiocManager", quietly = TRUE))
#     install.packages("BiocManager")
# BiocManager::install("affy", version = "3.8")
```

Using Affy Package

```
library(affy)
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

Accessing data from CEL files

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind,
##     colMeans, colnames, colSums, dirname, do.call, duplicated,
##     eval, evalq, Filter, Find, get, grep, grepl, intersect,
##     is.unsorted, lapply, lengths, Map, mapply, match, mget, order,
##     paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,
##     Reduce, rowMeans, rownames, rowSums, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which, which.max,
##     which.min
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
TRGN599_data<-ReadAffy()

#Installing required package for Affy
TRGN599_data
```