H3ABioNet

H3ABioNet

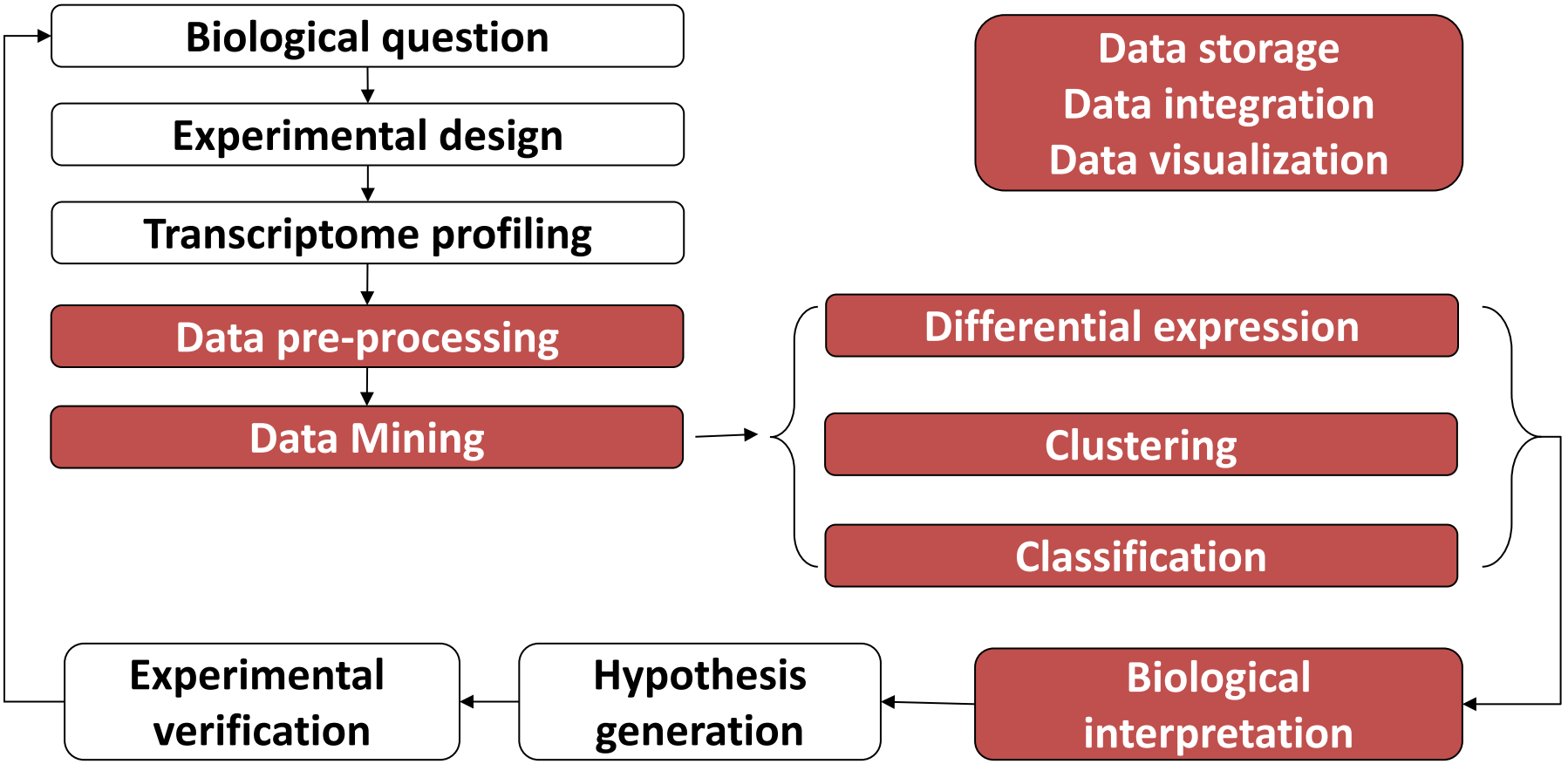
Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics Online Course: IBT

Gene Expression

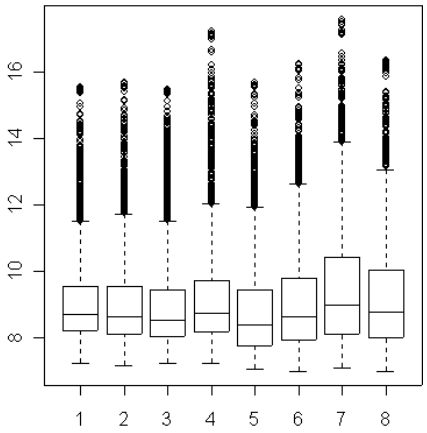
Transcriptome Informatics

Bioinformatics tasks

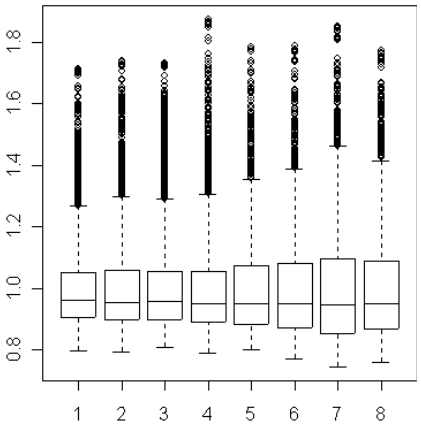


Normalization (among arrays)

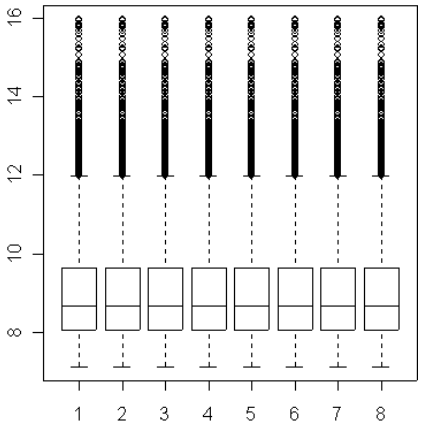
- Adjust the arrays using “housekeeping genes” (not recommended)
- Multiply each array by a constant to make the median intensity the same for each individual array (Global normalization)
- Match the percentiles of each array (Quantile normalization)



Without normalization



Global normalization



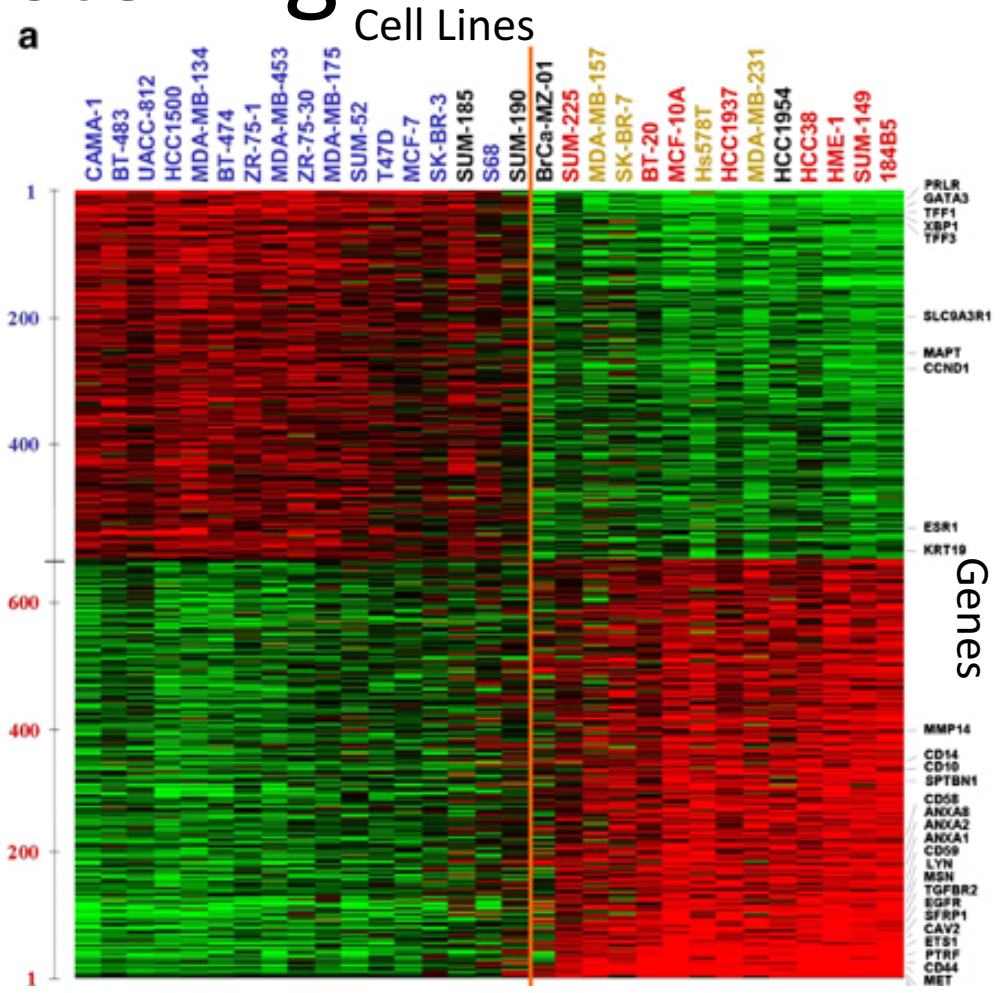
Quantile normalization

Three major goals of gene expression studies

1. Class comparison (what genes differentiate?)
   - Differential expression analysis
   - Input: gene expression data, class label of the samples
   - Output: differentially expressed genes
2. Class detection (which samples are similar?)
   - Clustering analysis
   - Input: gene expression data
   - Output: groups of similar samples or genes
3. Class prediction (which genes predict outcome?)
   - Machine learning techniques
   - Input: training set (expression plus class labels)
   - Output: prediction model with test set evaluation

Biclustering

"a method that simultaneously clusters genes and conditions, finding distinctive "checkerboard" patterns in matrices of gene expression data, if they exist."



Y Cheng and GM Church. *ICISMB* (2000) 10.1038/sj.onc.1209254

Y Kluger et al. *Genome Research* (2003) 13:703-716

Statistical analysis: hypothesis testing

Samples

Genes

probe_set_id	HNE0_1	HNE0_2	HNE0_3	HNE60_1	HNE60_2	HNE60_3
1007_s_at	8.6888	8.5025	8.5471	8.5412	8.5624	8.3073
1053_at	9.1558	9.1835	9.4294	9.2111	9.1204	9.2494
117_at	7.0700	7.0034	6.9047	9.0414	8.6382	9.2663
121_at	9.7174	9.7440	9.6120	9.7581	9.7422	9.7345
1255_g_at	4.2801	4.4669	4.2360	4.3700	4.4573	4.2979
1294_at	6.3556	6.2381	6.2053	6.4290	6.5074	6.2771
1316_at	6.5759	6.5330	6.4709	6.6636	6.6438	6.4688
1320_at	6.5497	6.5388	6.5410	6.6605	6.5987	6.7236
1405_i_at	4.3260	4.4640	4.1438	4.3462	4.3876	4.6849
1431_at	5.2191	5.2070	5.2657	5.2823	5.2522	5.1808
1438_at	7.0155	6.9359	6.9241	7.0248	7.0142	7.0971
1487_at	8.6361	8.4879	8.4498	8.4470	8.5311	8.4225
1494_f_at	7.3296	7.3901	7.0886	7.2648	7.6058	7.2949
1552256_a_at	10.6245	10.5235	10.6522	10.4205	10.2344	10.3144
1552257_a_at	10.3224	10.1749	10.1992	10.2464	10.2191	10.2405

Case Control

T-test evaluates probability of observations under the null hypothesis.

Null hypothesis $H_0 : \mu_1 = \mu_2$ Alternative hypothesis $H_1 : \mu_1 \neq \mu_2$

Limma and pooled variance

- The Bioconductor "Linear Models for Microarray Data" package includes the empirical Bayes moderated t-statistic test.
- If you have few replicates and many gene measurements, evaluate variance *among all genes of this sample* rather than among replicates for this gene!

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

https://bioconductor.org/packages/release/bioc/html/limma.html

Multiple testing correction

- Thousands of genes → thousands of tests
- Protect against *any* errors: Bonferroni

$$AdjustedThreshold = \frac{.05}{NumTrials}$$

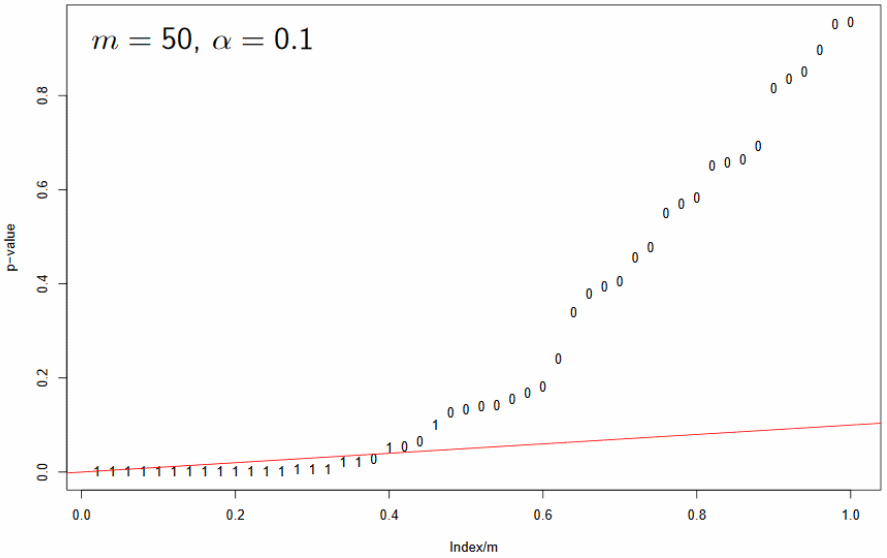
- Limit the *rate* of false positives: Benjamini-Hochberg

First, sort p-values.

$$GeneOneThreshold = (1/NumTrials) * 0.05$$

$$GeneTwoThreshold = (2/NumTrials) * 0.05$$

$$GeneThreeThreshold = (3/NumTrials) * 0.05$$



Gene Expression Omnibus

https://www.ncbi.nlm.nih.gov/geo/

- The NCBI has collected > 4000 data sets for gene expression at the GEO repository.
- The number of diseases and conditions explored is extremely diverse.
- Each set is accompanied by metadata and statistical code to assist interpretation.

Takeaway Messages

- Gene expression analysis needs contributions from both bioinformatics and biostatistics.
- Bioconductor and GEO are incredibly useful to investigators in gene expression.
- As with all science, a goal for any experiment is to refine the next question. Some have called “OMICs” technologies *hypothesis generation experiments*.