Introduction to Microarray Data Analysis using R/Bioconductor

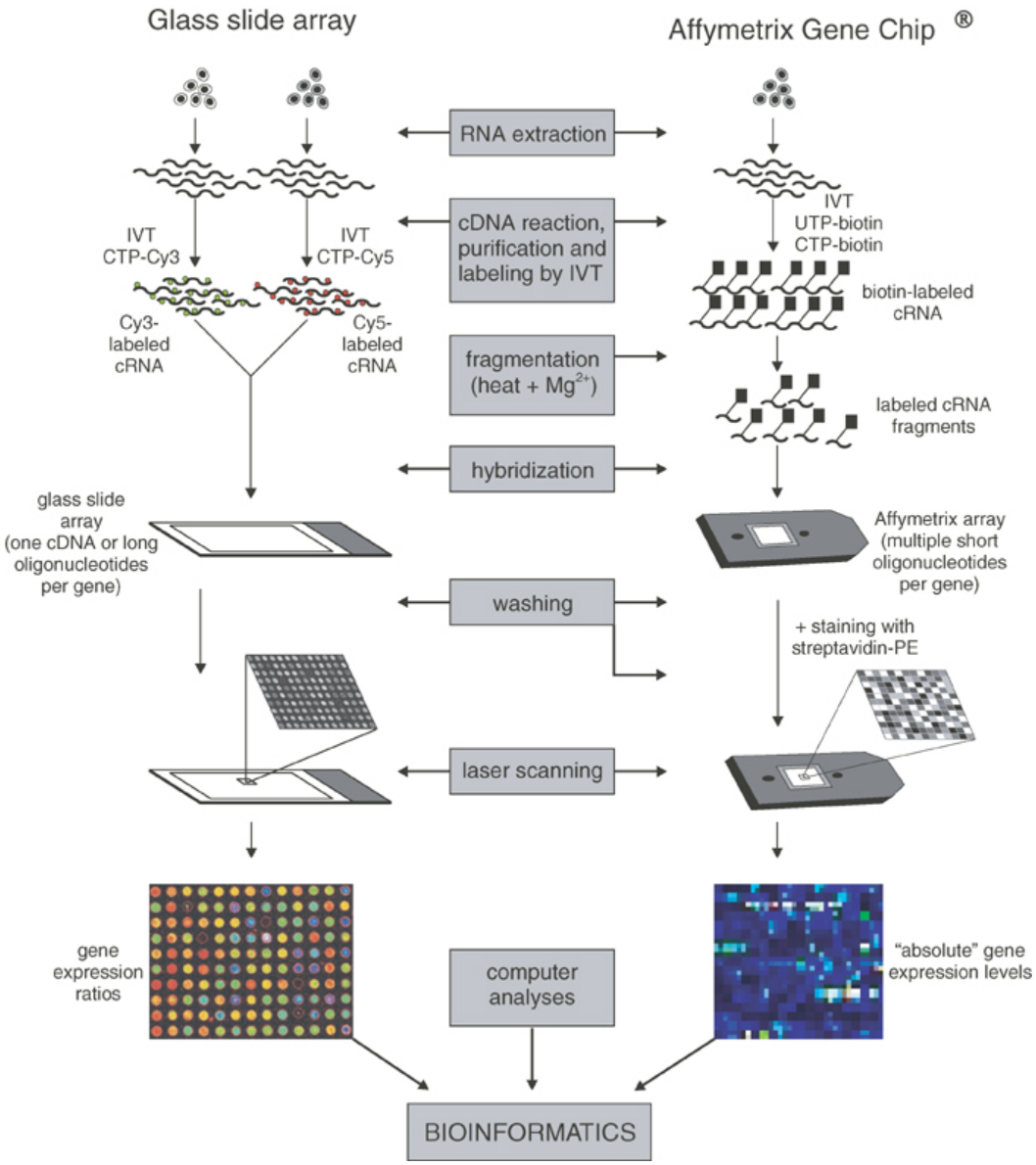
Fathi Elloumi, PhD

Fathi.Elloumi@nih.gov

Outline

- Microarray analysis workflow overview
- Affymetrix arrays
  - Processing & Normalization
  - Bioconductor packages
- Use case using TCGA data
  - Normalization and QC with SimpleAffy
- Exploratory analysis and visualization
  - PCA, Clustering & Heatmaps
  - DEG and Annotations
  - Survival analysis/ KM curves

Data preparation & generation



Microarray data analysis workflow

- Raw data Quality Control
- Normalization

- Filtering
- Estimate missing Values
- Differential gene inference
- Clustering
- Classification/prediction
- Annotation
- GO analysis
- Pathway analysis
- Survival analysis

Array Software

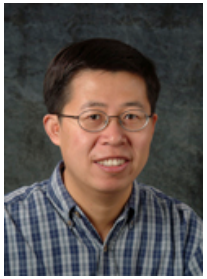
Affymetrix
Expression Console

Third Party Software
Open Source Software

R/Bioconductor (affy, limma)
Partek, GeneSpring

Commonly Used Commercial Platforms at LMT

Manufacturer	Types	Species	Released	# Probe Sets	# Probes per transcript	Input RNA	Cost	Usage
Affymetrix Genechips (25 bp)	Gene ST Array	Human, Mouse, Rat	2007	>19K (Gene 1.0 ST) >48K (Gene 2.0 ST)	21	100 ng	$218* ($326)	Pathway analysis, ease of interpretation
	PrimeView Array	Human	2012	>49K	9-11	100 ng	$168* ($218)	Pathway analysis, ease of interpretation
	3' IVT Arrays	Human Most sp.	2003	HG-U133 Plus 2.0 Array: >54K HG-U133A 2.0 Array: >22K	9-11	10-100 ng	$235* ($350)	Gene signatures, Pathway analysis
	Exon 1.0 ST Array	Human, Mouse, Rat	2007	>1.4M exon clusters	40	100 ng	$300* ($450)	Exon level differential expression
	Human Transcriptome Array (HTA) 2.0	Human	2013	6.7M probes	150	100 ng	($360)	Alternative splicing discovery



Xiaolin Wu

** Subsidized at 33% by OSTR*

Laboratory of Molecular Technology (LMT)
Advanced Technology Program, Frederick
http://ncifrederick.cancer.gov/atp/

Outline

- Microarray analysis workflow overview
- Affymetrix arrays
  - Processing & Normalization
  - Bioconductor packages
- Use case using TCGA data
  - Normalization and QC with SimpleAffy
- Exploratory analysis and visualization
  - PCA, Clustering/Heatmaps
  - DEG and Annotations
  - Survival analysis/ KM curves

Affymetrix arrays

Figure 1: GeneChip® Human Genome U133 Arrays shown in cartridge and plate formats.

Affymetrix Microarray Probe Design

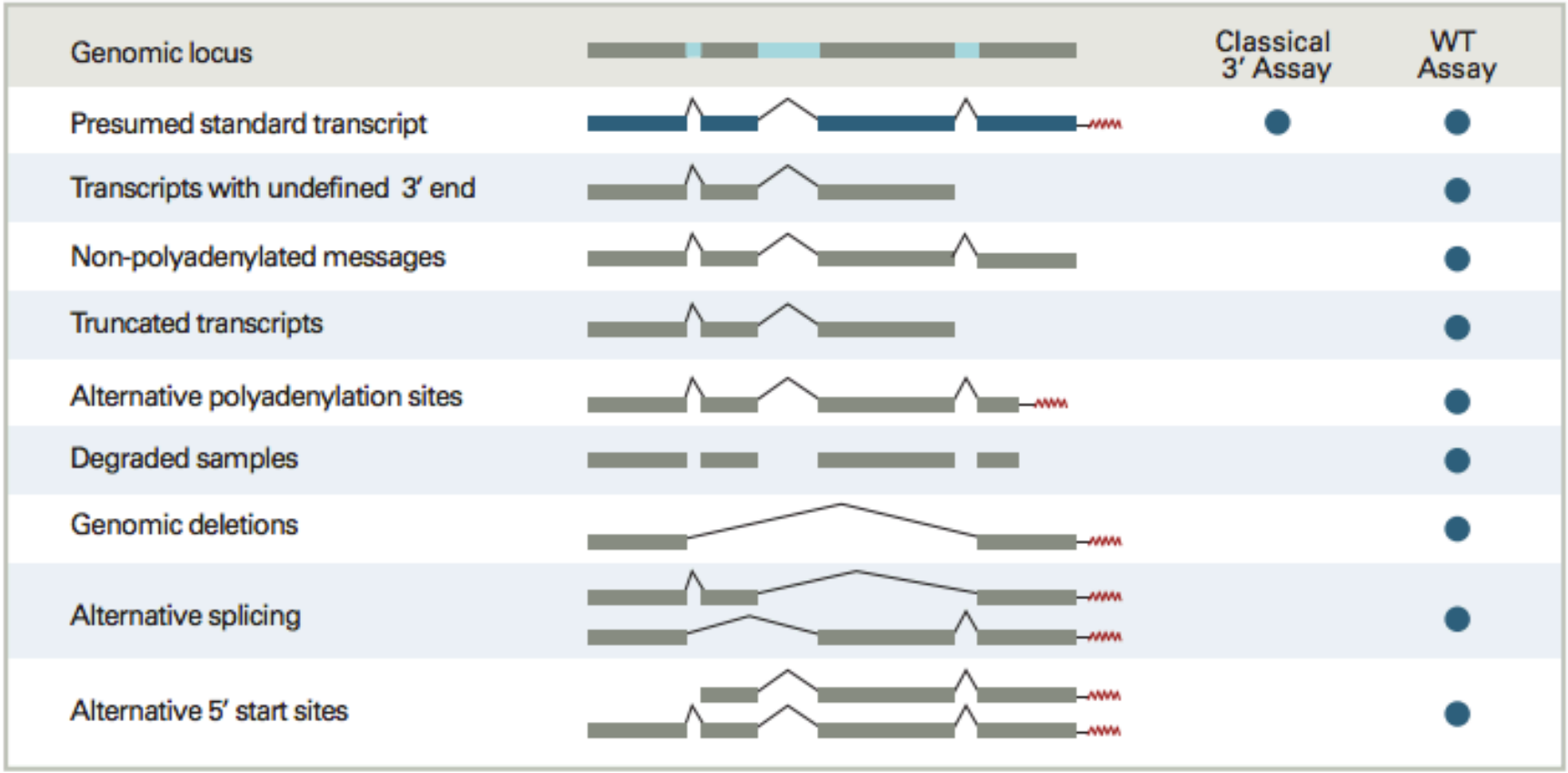
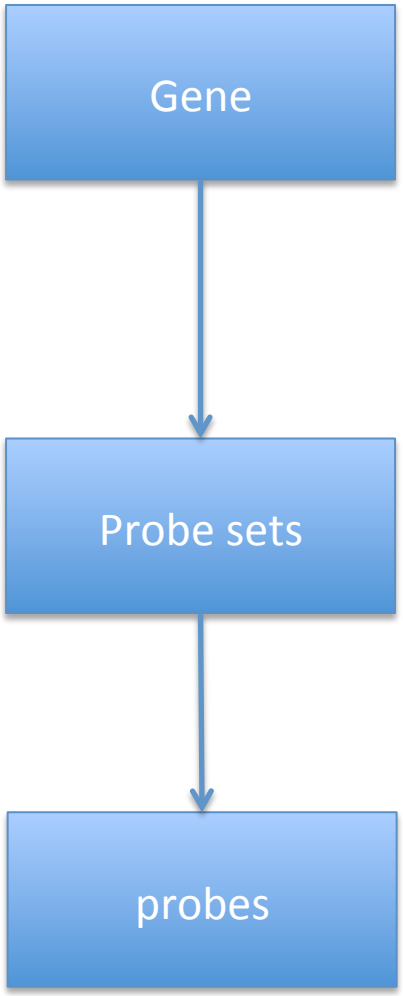


Figure 1: Types of transcripts captured by a whole-transcript assay. Most of these cannot be detected with the classical 3′ assay.

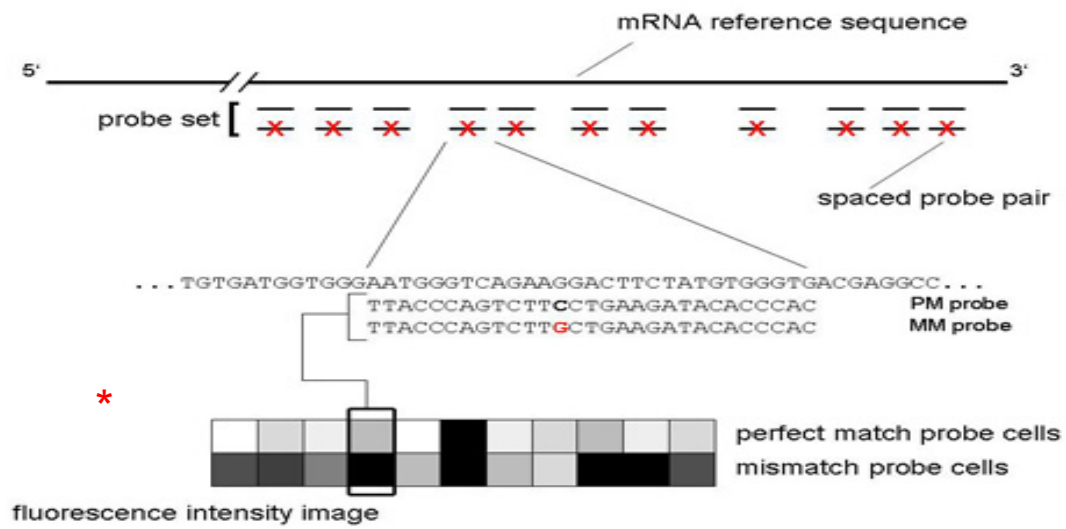
Critical Specifications for GeneChip® Human Genome Products

	Cartridge Format		Plate Format	
	Human Genome U133 Plus 2.0 Array	**Human Genome U133A 2.0 Array**	**Human Genome U133 A Array Plate**	**Human Genome U133 B Array Plate**
Number of transcripts	~47,400	~18,400	~18,400	~20,600
Number of genes	>38,500	>14,500	>14,500	>18,500
Number of probe sets	>54,000	>22,000	>22,000	>22,000
Feature size	11 µm	11 µm	8 µm	8 µm
Oligonucleotide probe length	25-mer	25-mer	25-mer	25-mer
Probe pairs/sequence	11	11	11	11
Control sequences included:				
Hybridization controls	*bioB, bioC, bioD, cre*	*bioB, bioC, bioD, cre*	*bioB, bioC, bioD, cre*	*bioB, bioC, bioD, cre*
Poly-A controls	*dap, lys, phe, thr*	*dap, lys, phe, thr*	*dap, lys, phe, thr*	*dap, lys, phe, thr*
Normalization control set	100 probe sets	100 probe sets	100 probe sets	100 probe sets
Housekeeping/Control genes	GAPDH, beta-Actin, ISGF-3 (STAT1)	GAPDH, beta-Actin, ISGF-3 (STAT1)	GAPDH, beta-Actin, ISGF-3 (STAT1)	GAPDH, beta-Actin, ISGF-3 (STAT1)
Detection sensitivity	1:100,000*	1:100,000*	1:100,000*	1:100,000*

*As measured by detection of pre-labeled transcripts derived from human cDNA clones in a complex human background.



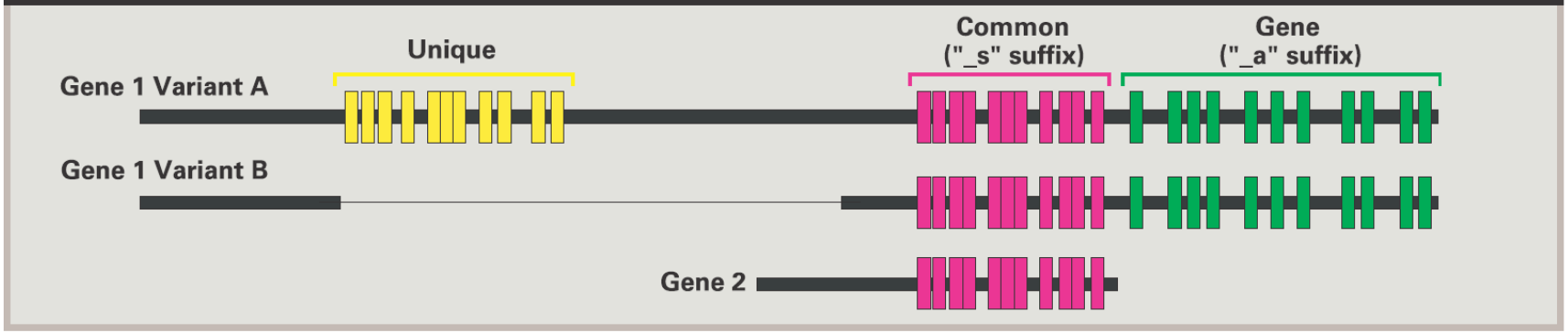
Affymetrix Microarray Probe Design (3' IVT)



- Each probe set is represented by 11 probes
- Probe pairs are designed from the 3' end of the gene
- Probe pair consists of PM (perfect match) and MM (mismatch) probes
- MM probe has an altered 13th base in 25 bases of probe sequence

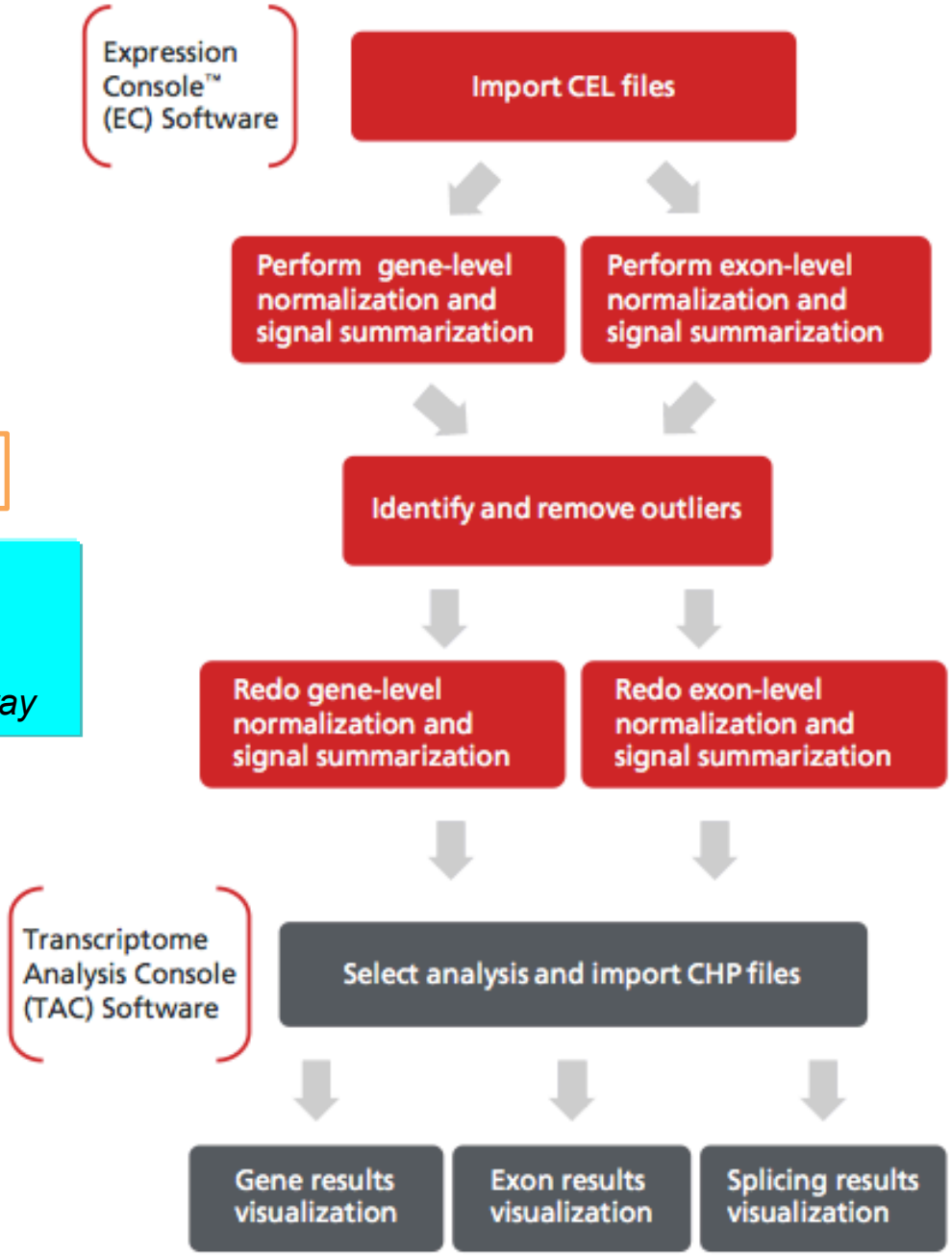
HG-U133 Plus 2.0 Array

Figure 3. Different probe set types are indicated by suffices to the probe set name. Unique probe sets are predicted to perfectly match only a single transcript. Gene probe sets, with an "_a" suffix, are predicted to only perfectly match transcripts from the same gene. Common probe sets, with a "_s" suffix, are predicted to perfectly match multiple transcripts, which may be from different genes. Probe sets that have a "_x" suffix are not shown here but are described in the text.



Probe Set ID	Property
12345_at	unique
12345_a_at	same gene family
12345_s_at	cross-hyb with another gene
12345_x_at	has at least 1 probe that cross-hyb with another probeset

Affymetrix EC workflow



GENE LEVEL

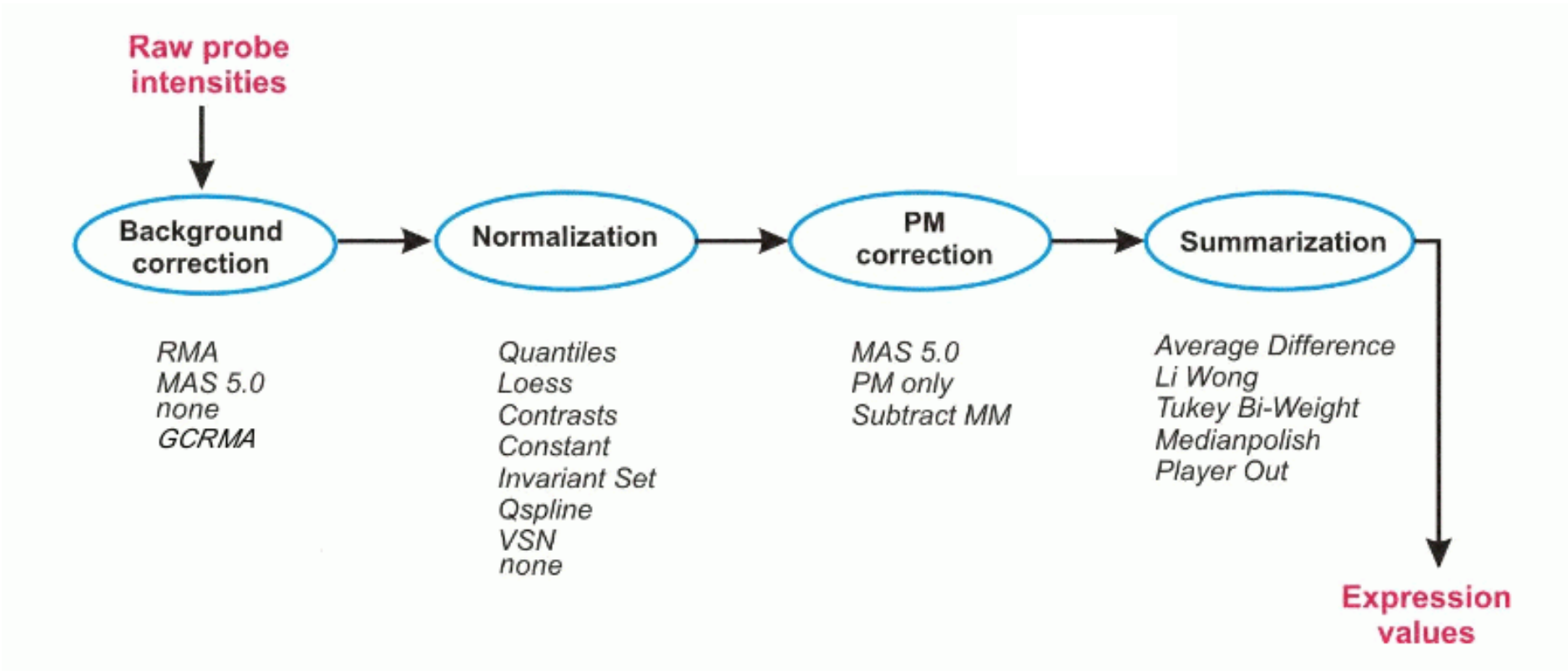
3' IVT, Gene ST, Exon ST, miRNA, Human Transcriptome Array

Main Affymetrix Files

Affy Console

File extension	Description	File type
DAT	pixel intensity file, scanned image	binary
CEL	Cell intensity file (created from a *.dat file)	(v.4) binary format (v.3) text format
ARR	Sample file	XML
AUDIT	Processing information (fluidics, attributes, barcode, library, etc.)	XML
CHP	Chip file containing expression data generated by analyzing *.cel file with different algorithms (e.g. mas5 or rma)	binary
RPT	Data quality information about the chip	text
Downloaded from Affymetrix Website		
CDF BGP, PGF, CLF	**C**hip **d**escription **f**ile (library file installed by *.exe file) Several library files for Whole Transcript type arrays (Gene, Exon)	text
CSV	Gene or transcript level annotation flat file provided by Affymetrix	text

Processing Affymetrix raw data



Background correction

Remove local artifacts and “noise” (caused by autofluorescence of the array surface and non-specific binding)

MM = mismatch probes – in theory can be background correction for PM

Normalization

Correction for differences in overall chip brightness and systematic biases in raw data in order to improve comparability in gene expression data (across arrays)

Overall assumption: Most genes are not changed – can only use this when the number of genes being measured is sufficiently high

Type of Systematic Error	Normalization Methods
Total Signal	Global normalization (median, mean, trimmed mean, etc.)
Distribution	Quantile normalization, Z-normalization
Skewing e.g. Dye bias (2-channel) or between multiple single-channel arrays	Loess normalization, cyclic Loess

NB. Log transformation of expression data is also typically applied to make distribution more normal

Different Summarization Methods

MAS 5.0

- Tukey's biweight algorithm (weighted mean): robust average of log(PM-MM) using one step Tukey's biweight estimate, where outliers are penalized with low weights

RMA

- Multichip linear model is fit to data from each probeset using Tukey's medianpolish

dChip

- model-based expression values are weighted average of PM-MM (or PM) values, with larger weights (ϕ' s) given to sensitive (responding) probes, and non-responsive probes with small ϕ' s are down-weighted or ignored in the MBEI

Commonly used algorithms for Affymetrix data

	Algorithm Name	Background correction	Normalization	PM Correction	Summarization	Reference
RMA	Robust Multi-array Average	Model-based	Quantile	PM only	Median Polish Returned values are log2 scales	Irizarry et al (Nucleic Acids Res, 2003)
GCRMA	GC-content corrected RMA	GC-content	Quantile	PM only	Median Polish	Wu et al (J Am Stat Assoc, 2004)
MAS 5	Microarray Suite 5.0	Local (4x4)	Trimmed mean	Ideal Mismatch	Tukey's Biweight	Affymetrix (Statistical Algorithms Description 2002)
PLIER	Probe Logarithmic Intensity Error	calculated from "feature responses"	None	PM-MM or PM-B	"Inconsistent features" downweighted by Geman-McClure function	Affymetrix (Technical report, 2005)
dChip	DNA-Chip Analyzer	Local (10x10)	Invariant Set	PM only	Model-Based Expression Values (MBEI)	Li and Wong (Genome Biol, 2001)

Bioconductor packages for Affymetrix data analysis (3' IVT)

Package	Main tasks
Affy	QC, Normalization
SimpleAffy	QC, Normalization, ttest
AffyPLM	Normalization, MAplot
arrayQualityMetrics	HTML QC report
Limma	Statistical test for differential expression analysis
Htggu1331a.db	Gene annotation
Gplots/ggplots2	Plots, heatmaps
survival	Survival analysis (KM curves) and Cox Model
…	

Main expression classes

- ExpressionSet: combine several different sources of information into a single convenient structure (expression data, phenotype, annotations and metadata) (Biobase package)

- AffyBatch: This is a class representation for Affymetrix GeneChip probe level data (Affy package)

Basic steps

	input	Output
Create an expression dataset from raw data	CEL, CDF, sample and experiment information	Object of class AffyBatch
Normalize data	AffyBatch object, normalization method	Object of class ExpressionSet
QC	AffyBatch/ExpressionSet, QC method	Metrics and plots
Clustering, DEG,…	ExpressionSet/matrix, task_method	Metrics and plots

Outline

- Microarray analysis workflow overview
- Affymetrix arrays
  - pre-processing and Normalization
  - Bioconductor packages
- Use case using TCGA data
  - Normalization and QC with SimpleAffy
- Exploratory analysis and visualization
  - PCA, Clustering/Heatmaps
  - DEG and Annotations
  - Survival analysis/ KM curves

https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm

tcga data portal

Home | Download Data | Tools | About the Data | Publication Guidelines

Home > Download Data > Data Matrix

In This Section

Data Matrix

The Data Matrix only provides the latest revision of each archive; older revisions are available through bulk download or HTTP access. Also, it does not allow for querying across multiple disease studies.

Select initial matrix filter settings. To view all data, click here or click "Apply" without choosing any settings. (Note: unfiltered matrix is large and can take some time to load.)

Filter Settings

Select a disease: GBM – Glioblastoma multiforme

Data Type:
- Clinical
- DNA Methylation
- Expression-Exon
- Expression-Genes

Batch Number:
- Batch 5
- Batch 6
- Batch 7
- Batch 8
- Batch 10

Data Level:
- Level 1
- Level 2
- Level 3

Availability:
- Available
- Pending
- Not Available

Preservation: Frozen Help

Center/Platform:
- All
- BCGSC (IlluminaHiSeq_miRNASeq)
- BCM (ABI)
- BI (ABI)

Sample:
ID Matches:
TCGA- -- -- Remove

Add Row

Paste Sample List:

Upload Sample List:
Browse... No file selected.

Access Tier:
- All
- Protected
- Public

Tumor/Normal:
- Tumor - matched
- Tumor - unmatched
- Normal - matched
- Organ-Specific Control
- Cell Line Control

Submitted Since (Date): mm/dd/yyyy

Submitted Up To (Date): mm/dd/yyyy

Only show samples with data available for all columns

Get web service URL for this filter

Apply Clear

ps://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm?mode=ApplyFilter&diseaseType=GBM

tcga data portal

Data Matrix Datasets

The Data Matrix only provides the latest revision of each archive; older revisions are available through bulk download or HTTP access. Also, it does not allow for querying across multiple disease studies.

GBM Data Matrix **Options:** Reset Matrix | Edit Filter | Remove Filter

Help

Preservation: Frozen

Color Cells By: Tumor/Normal **Scroll Size:** Standard

Build Archive

Legend:

- TN Tumor, Matched Normal
- T Tumor, no Matched Normal
- NT Normal, Matched Tumor
- N Organ-Specific Control
- C Cell Line Control
- No Available Dataset

*Protected data

Batch/Sample	Level	Clinical		Exp-Gene					
		XML	Biotab	BI HT_HG-U133A			UNC AgilentG4502A_07		
				1	2	3	1	2	3
Batch 8	TCGA-06-0192-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-06-0216-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-06-0649-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-06-0686-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-06-0743-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-06-0744-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-06-0745-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-06-0747-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-06-0749-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-06-0750-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-12-0653-01	TN	TN				TN	TN	TN
	TCGA-12-0654-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-12-0656-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-12-0657-01	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-12-0662-01	TN	TN				TN	TN	TN

ps://tcga-data.nci.nih.gov/tcga/dataAccessDownload.htm

tcga data portal

Data Download

By downloading, analyzing, and/or utilizing TCGA data for publication purposes, the user accepts the data use restrictions and requirements as outlined in the TCGA Publication Guidelines. See http://cancergenome.nih.gov/abouttcga/policies/publicationguidelines for additional information.

Enter E-mail Address:

Re-Enter E-mail Address:

Download | Cancel

Estimated Uncompressed Size: 180.379 MB

70 Gb maximum allowable size

Archive Options: ☐ Use Compression ***(Selecting this option may greatly increase the wait time for your download to be available)***

☐ Flatten Directory Structure

Please enter and confirm your e-mail address. Upon selecting "Download", your files will be tar'd and gzip'd. When completed, an e-mail will be sent to you with a link to your file. This file will remain on the server for 24 hours. A link to the file will also appear in the browser window.

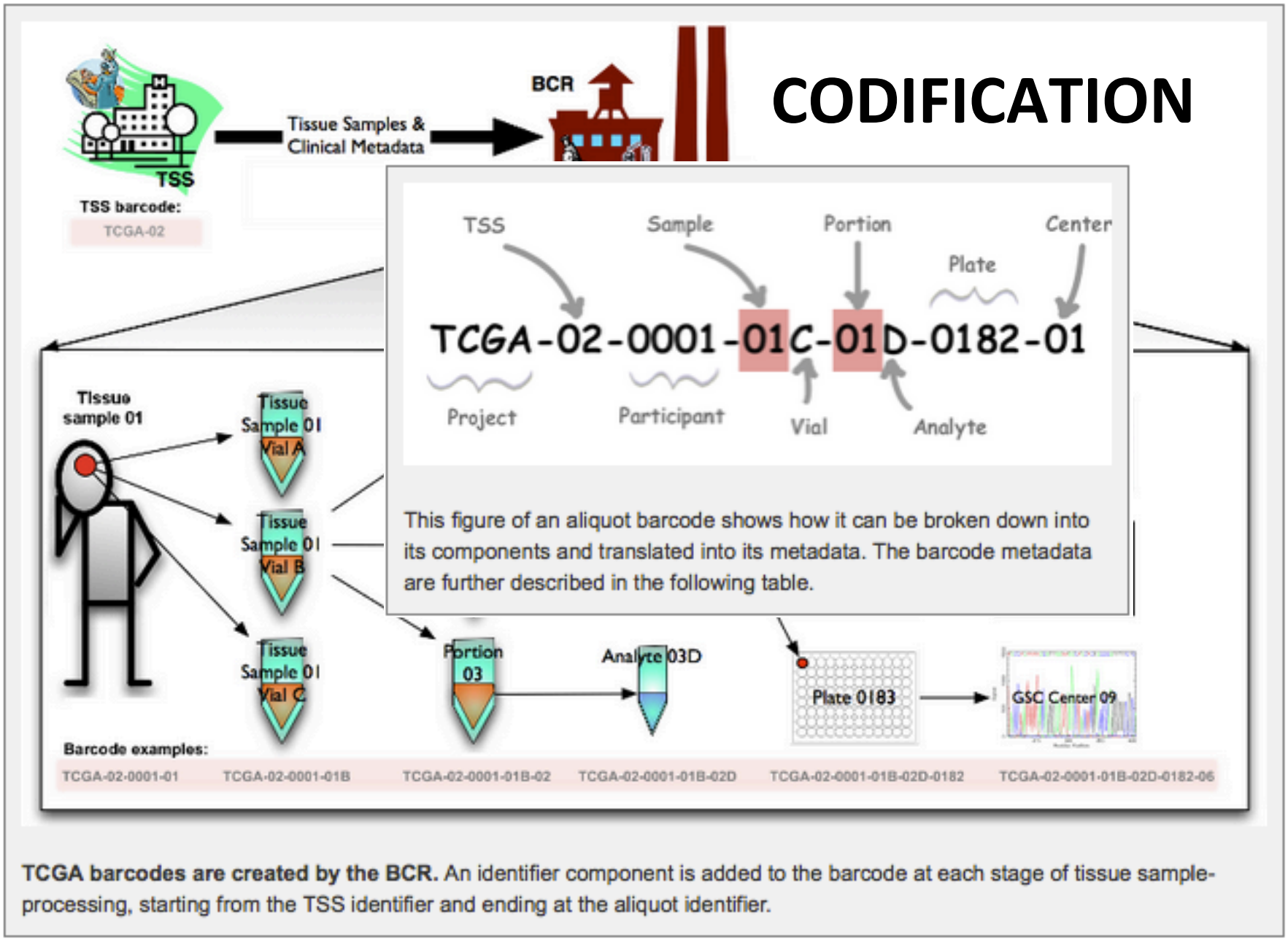
IMPORTANT: Data downloaders are urged to use the data annotation search interface (https://tcga-data.nci.nih.gov/annotations/) to query the case, sample, and aliquot identifiers in their download to obtain the latest information associated with their data.

Select files to include in your archive:

- ☑ METADATA
  - ☑ BI (HT_HG-U133A)
    - ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.idf.txt (3.705 KiB)
    - ☑ selected_samples::broad.mit.edu_GBM.HT_HG-U133A.sdrf.txt (551.172 KiB)
- ☑ Expression-Genes
  - ☑ BI (HT_HG-U133A)
    - ☑ Level 1
      - ☑ TCGA-06-0192-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_A01_298034.CEL (5.28 MiB)
      - ☑ TCGA-06-0686-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_A02_298028.CEL (5.28 MiB)
      - ☑ TCGA-06-0744-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_A03_297980.CEL (5.28 MiB)
      - ☑ TCGA-12-0775-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_A04_298010.CEL (5.28 MiB)
      - ☑ TCGA-06-0216-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_A05_298092.CEL (5.28 MiB)
      - ☑ TCGA-12-0688-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_A06_298074.CEL (5.28 MiB)
      - ☑ TCGA-06-0745-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_A07_297924.CEL (5.28 MiB)
      - ☑ TCGA-12-0776-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_A08_297946.CEL (5.28 MiB)
      - ☑ TCGA-06-0649-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_A09_298056.CEL (5.28 MiB)
      - ☑ TCGA-06-0680-11::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_A10_298048.CEL (5.28 MiB)
      - ☑ TCGA-12-0692-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_B02_297952.CEL (5.28 MiB)
      - ☑ TCGA-06-0749-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_B03_298026.CEL (5.28 MiB)
      - ☑ TCGA-12-0780-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_B04_298060.CEL (5.28 MiB)
      - ☑ TCGA-12-0654-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_B05_298096.CEL (5.28 MiB)
      - ☑ TCGA-12-0703-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_B06_298078.CEL (5.28 MiB)
      - ☑ TCGA-06-0750-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_B07_297984.CEL (5.28 MiB)
      - ☑ TCGA-12-0656-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_B09_297962.CEL (5.28 MiB)
      - ☑ TCGA-12-0707-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_B10_297936.CEL (5.28 MiB)
      - ☑ TCGA-15-0742-01::BONES_p_TCGA_Batch8_9_RNA_HT_HG-U133A_96-HTA_C01_298088.CEL (5.28 MiB)

You will receive a tar file !

CODIFICATION



TCGA barcodes are created by the BCR. An identifier component is added to the barcode at each stage of tissue sample-processing, starting from the TSS identifier and ending at the aliquot identifier.

A **Biospecimen Core Resource (BCR)** is a TCGA center where samples are carefully catalogued, processed, quality-checked and stored along with participant clinical information.

Lab1

- Create an experiment using TCGA GBM data
  - Location of CEL files
  - Sample information
- Main features and methods of an affyBatch object
- Normalize the data
- QC data
  - SimpleAffy

SimpleAffy QC metrics

Detect issues with RNA extraction, labelling, scanning

1. Average background: should be similar across all chips
2. Scale factor: the assumption is that gene expression does not change for the majority of genes => trimmed mean intensity should be the same. Affymetrix recommend that their scale factors should be within 3-fold change of one another
3. Number of genes called present: Probesets are flagged Marginal or Absent when the PM values for that probeset are not considered to be significantly above the MM probes
4. 3' to 5' ratios of actin and GAPDH: measure the quality of the RNA hybridized to the chip. For affy "standard protocol"
   1. GAPDH 3':5' ratio should be around 1 (default 1.25)
   2. ACTIN 3':5' ratio should be less than 3
5. Values for spike-in controls transcrips (hybridization controls): BioB, BioC, BioD and CreX should be present (especially BioB)
6. Uses ordered probes in all probeset to detect possible RNA degradation.

SimpleAffy QC metrics

Detect issues with RNA extraction, labelling, scanning

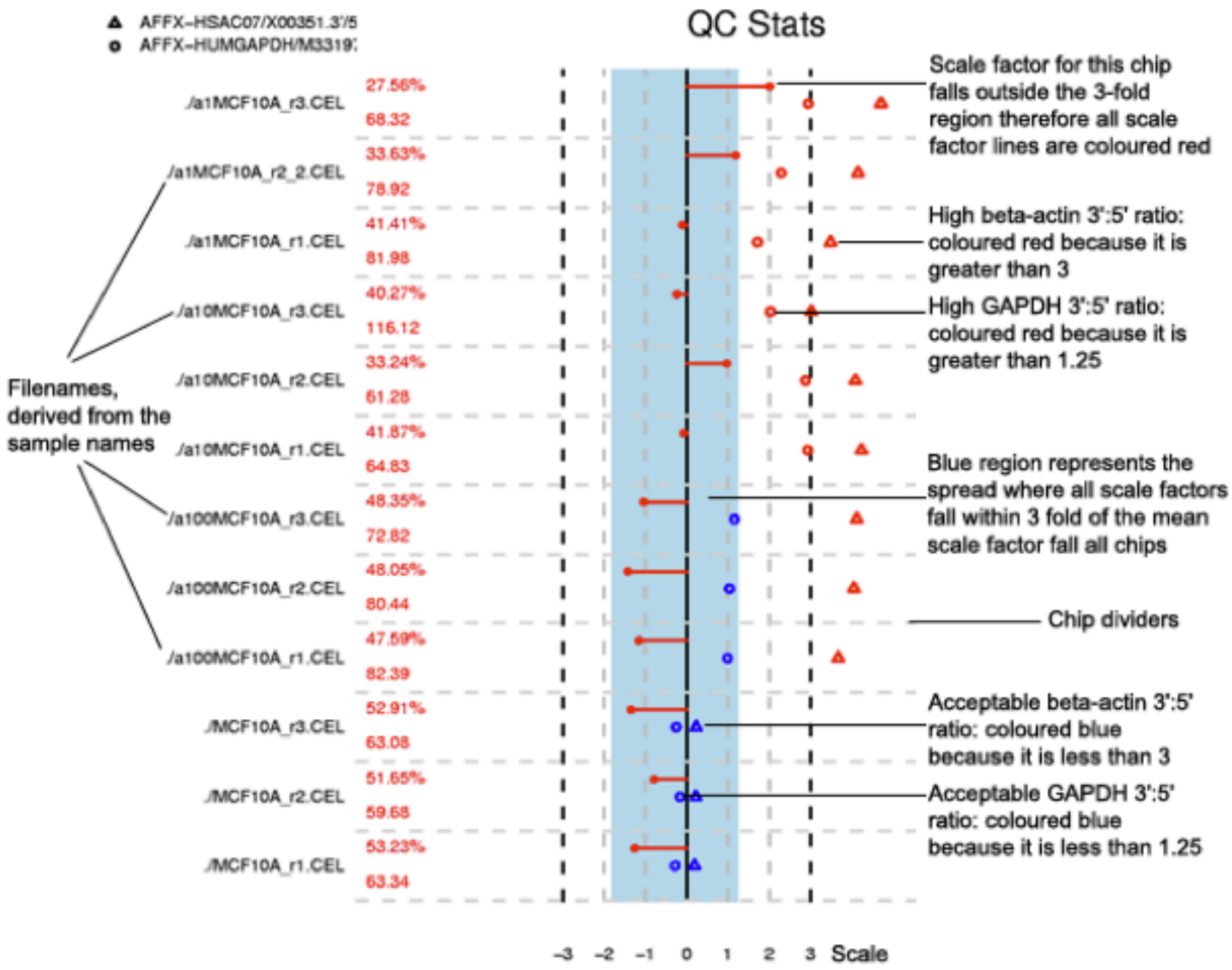
7. Probe-sets homogeneity

- **NUSE plot** (package affyPLM)

The Normalized Unscaled Standard Error (NUSE) is the individual probe error fitting the Probe-Level Model (the PLM models expression measures using a M-estimator robust regression). The NUSE values are standardized at the probe-set level across the arrays: median values for each probe-set are set to 1. The boxplots allow checking (1) if all distributions are centered near 1 – typically an array with a boxplot centered around 1.1 shows bad quality and (2) if one array has globally higher spread of NUSE distribution than others, which may also be a sign of low quality.

- **RLE plot**

The Relative Log Expression (RLE) values are computed by calculating for each probe-set the ratio between the expression of a probe-set and the median expression of this probe-set across all arrays of the experiment. It is assumed that most probe-sets are not changed across the arrays, so it is expected that these ratios are around 0 on a log scale. The boxplots presenting the distribution of these log-ratios should then be centered near 0 and have similar spread. Other behavior would be a sign of low quality.

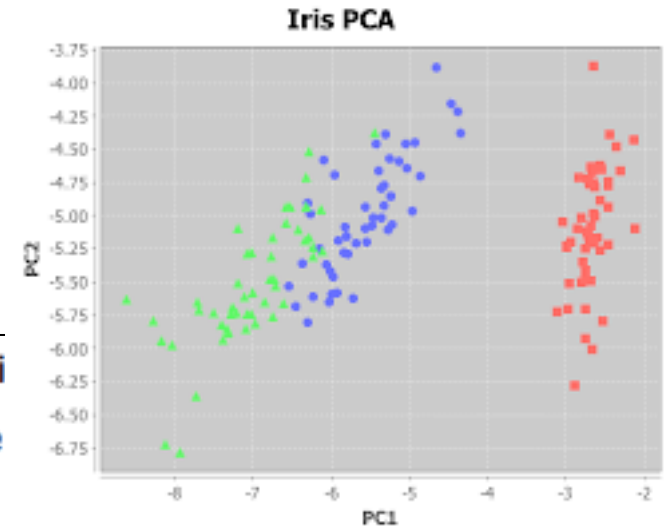

QC Stats
AFFX-HSAC07/X00351.3'/5
AFFX-HUMGAPDH/M33197
./a1MCF10A_r3.CEL
./a1MCF10A_r2_2.CEL
./a1MCF10A_r1.CEL
./a10MCF10A_r3.CEL
./a10MCF10A_r2.CEL
./a10MCF10A_r1.CEL
./a100MCF10A_r3.CEL
./a100MCF10A_r2.CEL
./a100MCF10A_r1.CEL
./MCF10A_r3.CEL
./MCF10A_r2.CEL
./MCF10A_r1.CEL
27.56%
68.32
33.63%
78.92
41.41%
81.98
40.27%
116.12
33.24%
61.28
41.87%
64.83
48.35%
72.82
48.05%
80.44
47.59%
82.39
52.91%
63.08
51.65%
59.68
53.23%
63.34
-3 -2 -1 0 1 2 3 Scale
Filenames, derived from the sample names
Scale factor for this chip falls outside the 3-fold region therefore all scale factor lines are coloured red
High beta-actin 3':5' ratio: coloured red because it is greater than 3
High GAPDH 3':5' ratio: coloured red because it is greater than 1.25
Blue region represents the spread where all scale factors fall within 3 fold of the mean scale factor fall all chips
Chip dividers
Acceptable beta-actin 3':5' ratio: coloured blue because it is less than 3
Acceptable GAPDH 3':5' ratio: coloured blue because it is less than 1.25

Outline

- Microarray analysis workflow overview
- Affymetrix arrays
  - Processing & Normalization
  - Bioconductor packages
- Use case using TCGA data
  - Normalization and QC with SimpleAffy
- Exploratory analysis and visualization
  - PCA, Clustering & Heatmaps
  - DEG and Annotations
  - Survival analysis/ KM curves

Principal Component Analysis

- Method for dimension reduction to identify patterns (thousands of genes = thousands of dimensions)



What is a "good" subspace?

Let's assume that our goal is to reduce the dimensions of a *d*-dimensi
k-dimensional subspace (where $k < d$). So, how do we know what size
know if we have a feature space that represents our data "well"?

Later, we will compute eigenvectors (the components) from our data set and collect them in a so-called scatter-matrix (or alternatively calculate them from the covariance matrix). Each of those eigenvectors is associated with an eigenvalue, which tell us about the "length" or "magnitude" of the eigenvectors. If we observe that all the eigenvalues are of very similar magnitude, this is a good indicator that our data is already in a "good" subspace. Or if some of the eigenvalues are much much higher than others, we might be interested in keeping only those eigenvectors with the much larger eigenvalues, since they contain more information about our data distribution. Vice versa, eigenvalues that are close to 0 are less informative and we might consider in dropping those when we construct the new feature subspace.

Eigenvalue: describes <u>the total variance</u> in an eigenvector.
<u>The eigenvector with the largest eigenvalue</u> is the first principal component. The second largest eigenvalue will be the direction of the second largest variance.

Cluster Analysis

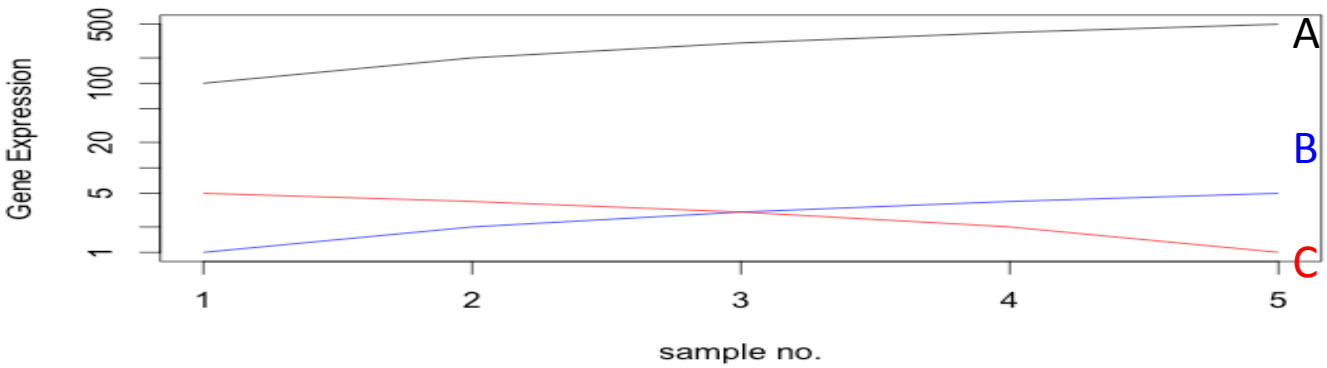
- Grouping together similar entities based on a distance metric

<u>Distance Metrics</u>:

1. Euclidean Distance (formula)

2. Correlation Distance – based on Pearson's correlation coefficient (r)

   Distance = 1-r (1-Pearson correlation)
   Maximum distance = 1.0 (range: 0-1)



3. Others: binary, maximum, canberra, minkowski, manhattan, mahanalobis

Cluster methods

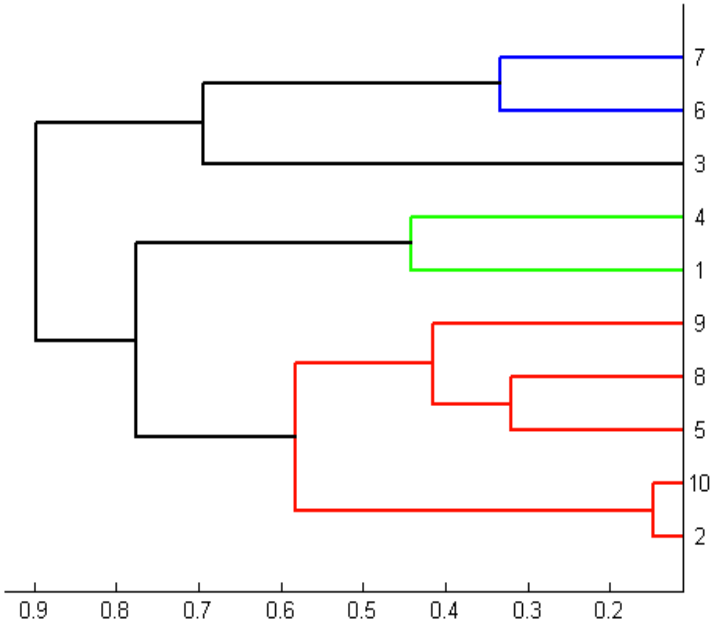
1. Hierarchical method
   - Hierarchical clustering methods produce a tree or dendrogram.
   - The tree can be built in two distinct ways
     - bottom-up: agglomerative clustering: simple and precise at bottom of the tree
     - top-down: divisive clustering.

2. Partitioning method
   - Partition the data into a pre-specified number k of mutually exclusive and exhaustive groups
   - Iteratively reallocate the observations to clusters until some criterion is met, e.g. minimize within cluster sums of squares

Hierarchical Clustering

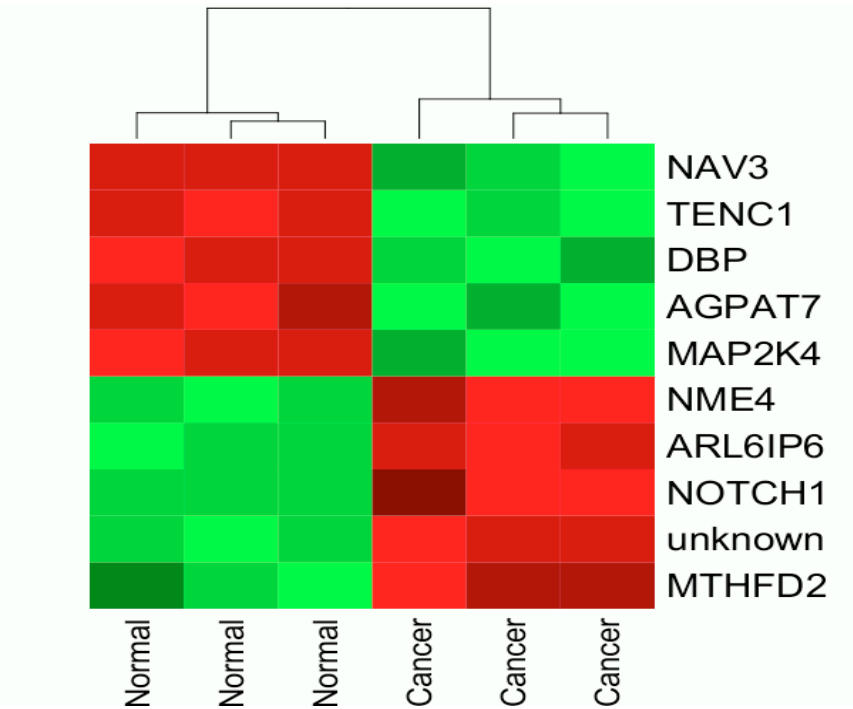
Dendrogram/tree

- branching diagram representing a hierarchy of categories based on degree of similarity



root branches leaves

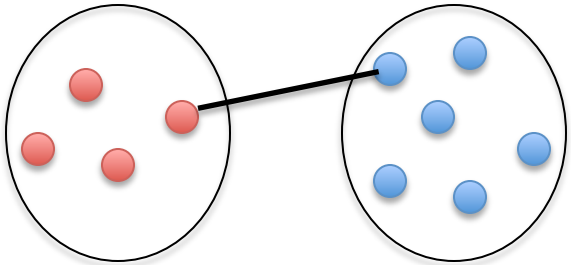
Heatmap



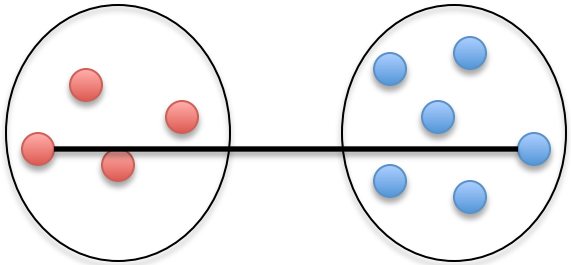
Hierarchical Clustering

<u>Agglomerative clustering methods</u>

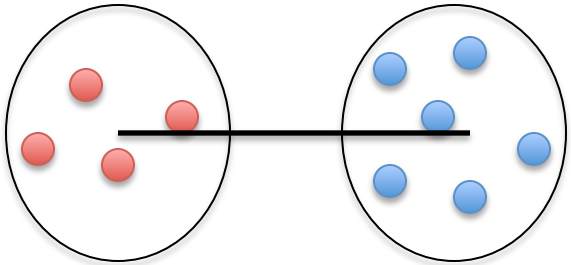
1. Single Linkage



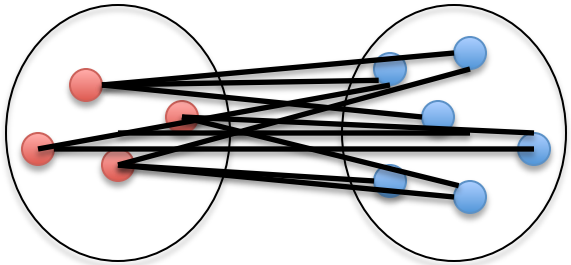
2. Complete Linkage



3. Centroid Linkage



4. Average Linkage



Lab2

- PCA, clustering and heatmaps using:
  - ArrayQualityMetrics (before/after normalization)
  - SimpleAffy
  - Gplots
  - Stats

Outline

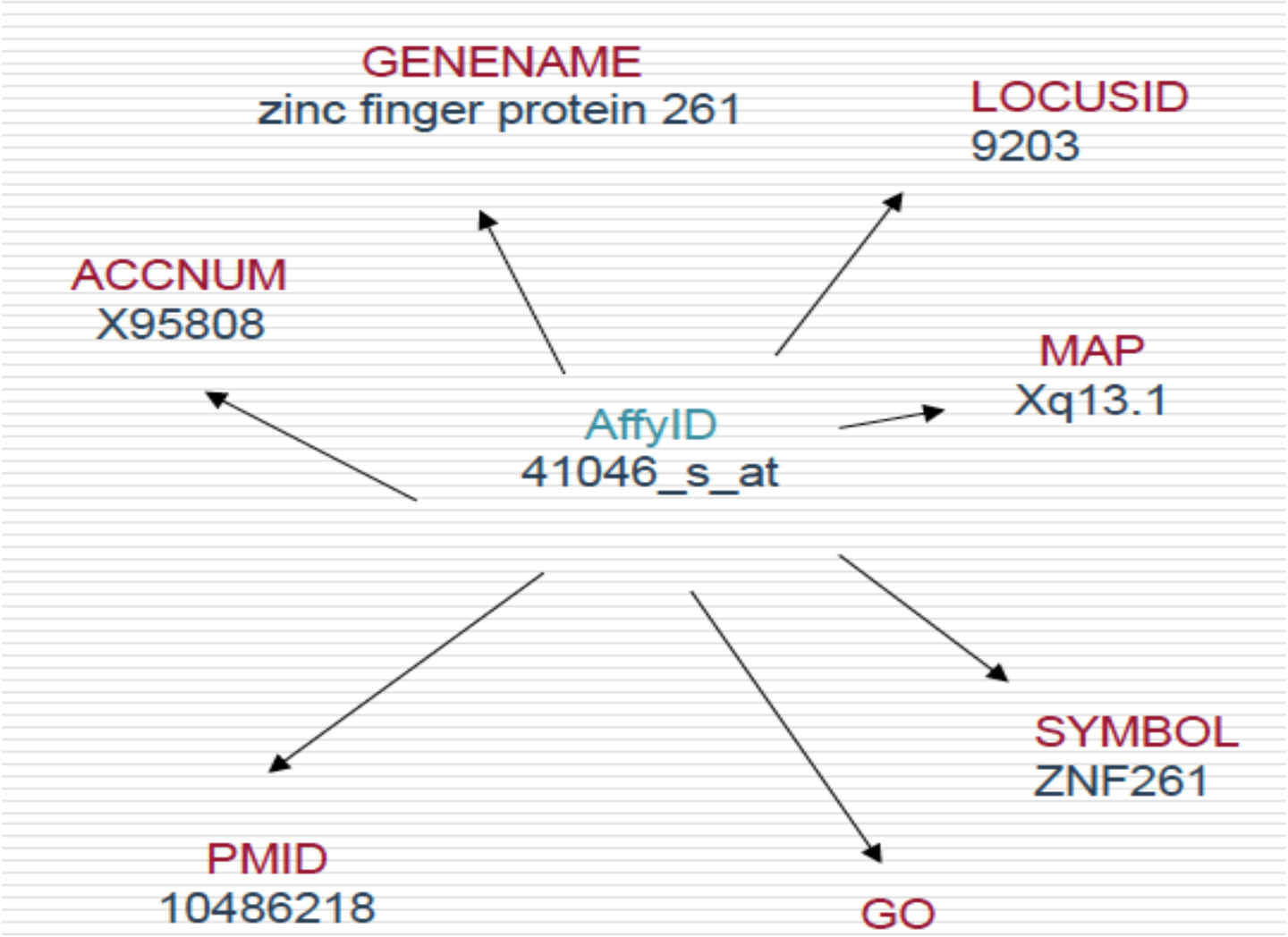
- Microarray analysis workflow overview
- Affymetrix arrays
  - Processing & Normalization
  - Bioconductor packages
- Use case using TCGA data
  - Normalization and QC with SimpleAffy
- Exploratory analysis and visualization
  - PCA, Clustering & Heatmaps
  - DEG and Annotations
  - Survival analysis/ KM curves

Typical question

- What are the genes that are differentially expressed between two or more groups?
  - do statistical test:
    - T-test
    - Empirical Bayes (moderated t-test)
    - Significance Analysis of Microarrys (SAM)
    - Anova (> 2 groups)
    - …
  - adjust for multiple testing (FDR….)

ANNOTATION

Annotate,
hthgu133a.db,
…



Lab3

- DEG using:
  - SimpleAffy
  - Limma
- Annotation with hthgu133a.db
- Clustering genes and samples:
  - SimpleAffy
  - Gplots

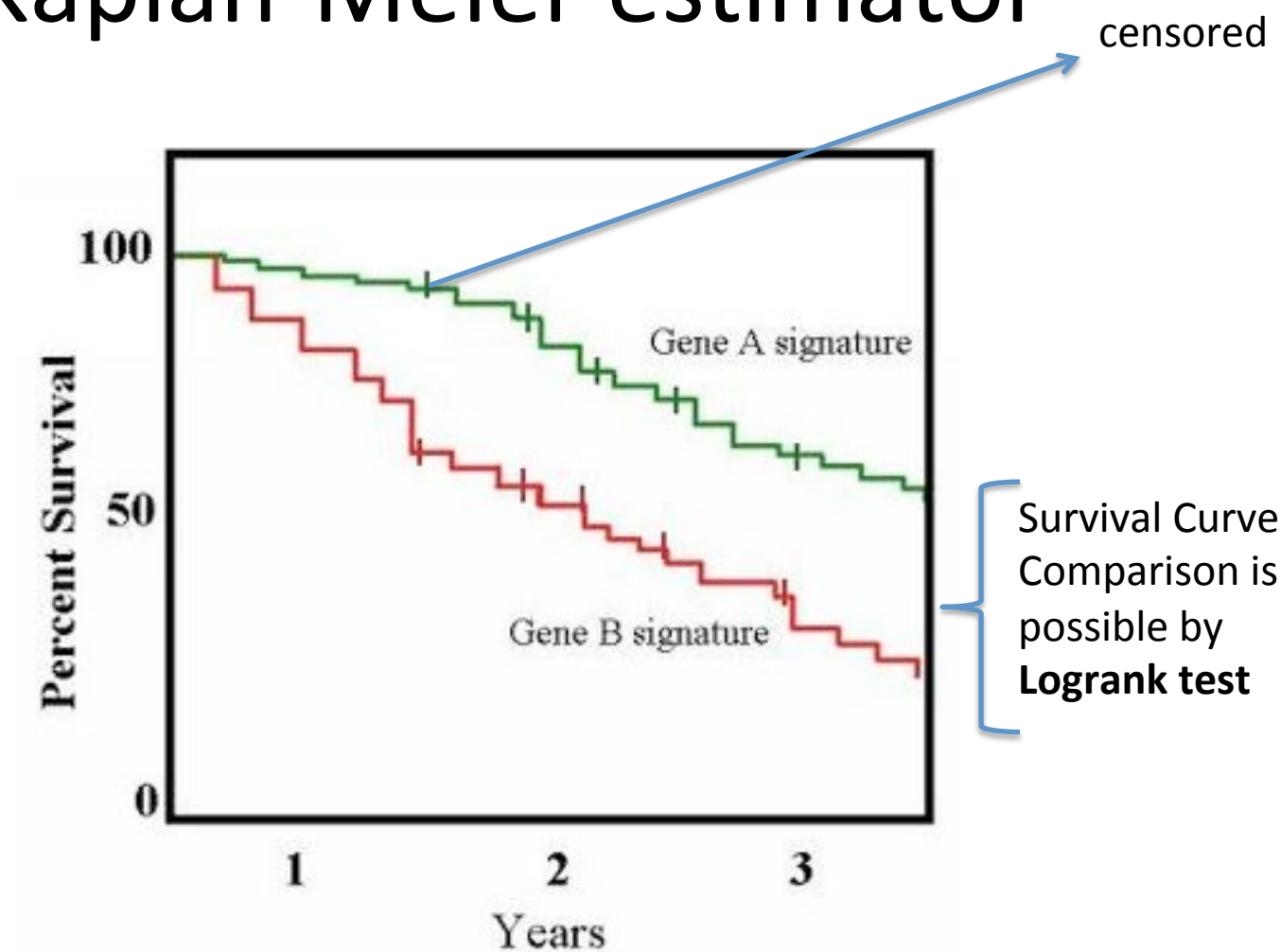
Outline

- Microarray analysis workflow overview
- Affymetrix arrays
  - Processing & Normalization
  - Bioconductor packages
- Use case using TCGA data
  - Normalization and QC with SimpleAffy
- Exploratory analysis and visualization
  - PCA, Clustering & Heatmaps
  - DEG and Annotations
  - Survival analysis/ KM curves

Survival Analysis

- **Survival analysis** is a branch of statistics which deals with analysis of time duration to until one or more events happen, such as death in biological organisms and failure in mechanical systems [wikipedia]
- The object of primary interest is the **survival function**, conventionally denoted *S*, which is defined as $S(t)=Pr(T>t)$, probability that the time of death T is later than some specified time t
- $S(0)=1$, $S(+\infty)=0$ and $S(t)$ is a decreasing function

Kaplan-Meier estimator



S(t) is estimated using a step function in which the estimated survival probabilities are constant between adjacent death times and only decrease at each death.

Censored observations

- The event time cannot always be measured due to:
  - Study end or dropout: the patient leaves the study before the event occurs or the study ends before the event has occurred (right censoring)
  - Event already occurred before study enrolment (left censoring)
  - ….

Survival in R

1. Create a survival object: Surv
   - Ex: Right-censored lifetimes: 26, 42, 71, 80+, 80+
   - Time=c(26,42,71,80,80)
   - Events=c(1,1,1,0,0)
   - Survobj=Surv(times,events)
   - Survobj => **26 42 71 80+ 80+**
2. Estimate the survival function: survfit
3. Test for difference in lifetime distributions: survdiff

Lab4

- Survival analysis:
  - Gender effect
  - Gene expression effect

Exercise

- Cluster only tumor samples using Pearson correlation
- Cut the tree in 4 clusters and take the 2 largest ones to look at DEG
- Generate KM curves for the 2 clusters