

Technical Training



Week 3

Presented by: Annie, Yves, Erin, Goral

What is Bioconductor?

- <https://www.bioconductor.org>
- Open source and Open development
- Free
- Software project for analysis of genomic data - and related tools, resources/datasets
- Current version 3.11, consists of 1903 packages

Example of packages can you find in Bioconductor?

- Annotation packages
- Experimental data packages

The screenshot shows the Bioconductor website homepage. At the top left is the Bioconductor logo with the tagline "OPEN SOURCE SOFTWARE FOR BIOINFORMATICS". To the right is a search bar and a navigation menu with links for Home, Install, Help, Developers, and About. The main content area is divided into several sections: "About Bioconductor" which describes the project's goals and user community; "News" which lists recent events and announcements; "BioC 2020" which provides information about the upcoming conference; "Install" which guides users on how to get started; "Learn" which offers resources for mastering the tools; "Use" which encourages creating solutions with the software; and "Develop" which provides resources for contributing to the project. A prominent "Bioconductor is hiring!" announcement is also visible at the bottom of the main content area.

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

[Home](#) [Install](#) [Help](#) [Developers](#) [About](#)

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and [Docker](#) images.

News

- See our [google calendar](#) for events, conferences, meetings, forums, etc. Add your event with email to events at bioconductor.org.
- [Bioconductor 3.11](#) is available.
- Nominate an outstanding community member for a *Bioconductor Award*! See the [support site](#) for more information.
- Registration open for [BioC2020](#).
- Core team **job opportunities** available, contact [Martin.Morgan at RoswellPark.org](mailto:Martin.Morgan@RoswellPark.org)
- [Bioconductor F1000 Research Channel](#) is available.
- Orchestrating single-cell analysis with *Bioconductor* ([abstract](#); [website](#)) and other [recent literature](#).

BioC 2020

Get the latest updates on the [BioC 2020 Conference!](#)

- BioC 2020 is going virtual July 27 - July 31. Please see the [Registration Page](#) for more information.
- Nominate an outstanding *Bioconductor* community member for a *Bioconductor Award*! See [posting](#) for more information.
- Call for birds-of-feather, hack-a-thon, and how-to sections. Please see [posting](#) for more information.
- Registration is now open. [Register today](#).

Install »

- Discover [1903 software packages](#) available in *Bioconductor* release 3.11.

Get started with *Bioconductor*

- [Install Bioconductor](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn »

Master *Bioconductor* tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use »

Create bioinformatic solutions with *Bioconductor*

- [Software](#), [Annotation](#), and [Experiment packages](#)
- [Docker](#) and [Amazon](#) machine images
- [Latest release announcement](#)
- Use *Bioconductor* in the [AnVIL](#). See our

Develop »

Contribute to *Bioconductor*

- [Developer resources](#)
- [Use BioC 'devel'](#)
- ['Devel' packages](#)
- [Package guidelines](#)
- [New package submission](#)
- [Git source control!](#)

Bioconductor is hiring!

Bioconductor is hiring for [full-time positions](#) on the *Bioconductor* Core Team! Individual projects are flexible and offer unique opportunities to

Bioconductor basics

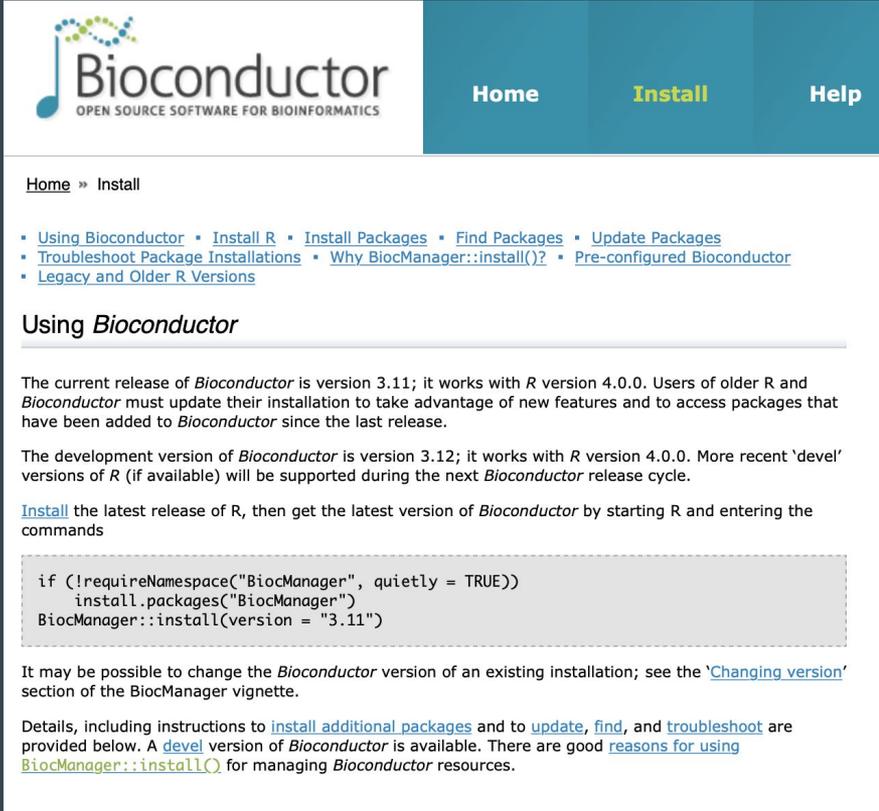
- Install Bioconductor
- Install the following libraries
 - affy
 - affyPLM
 - sva
 - AnnotationDbi
 - hgu133plus2.db
 - simpleaffy
 - arrayQualityMetrics
 - affyQCReport
 - gcrma

Using the code:

```
BiocManager::install('affy')
```

Load the library:

```
library(package_name)
```



The screenshot shows the Bioconductor website's 'Install' page. At the top, there is a navigation bar with 'Home', 'Install', and 'Help' links. The 'Install' link is highlighted in yellow. Below the navigation bar, the page title is 'Home » Install'. There is a list of links: 'Using Bioconductor', 'Install R', 'Install Packages', 'Find Packages', 'Update Packages', 'Troubleshoot Package Installations', 'Why BiocManager::install()?', 'Pre-configured Bioconductor', and 'Legacy and Older R Versions'. The main heading is 'Using *Bioconductor*'. The text explains that the current release is version 3.11, which works with R version 4.0.0. It also mentions that users of older R and Bioconductor must update their installation to take advantage of new features and to access packages that have been added since the last release. The development version is version 3.12, which works with R version 4.0.0. More recent 'devel' versions of R (if available) will be supported during the next Bioconductor release cycle. The text then instructs to 'Install' the latest release of R, then get the latest version of Bioconductor by starting R and entering the commands:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install(version = "3.11")
```

It may be possible to change the *Bioconductor* version of an existing installation; see the '[Changing version](#)' section of the BiocManager vignette.

Details, including instructions to [install additional packages](#) and to [update](#), [find](#), and [troubleshoot](#) are provided below. A [devel](#) version of *Bioconductor* is available. There are good [reasons for using BiocManager::install\(\)](#) for managing *Bioconductor* resources.

Reading in the Data

- 1) Download the Raw.tar file from GEO website
- 2) Move the downloaded file to your working directory
- 3) Load affy package :

```
library(affy)
```

- 4) Read the .CEL.gz files directly from the folder containing them

```
raw <- ReadAffy(celfile.path = "~/week_3/GSE32323_RAW")
```

- 5) Take a look at your data

```
df <- as.data.frame(exprs(raw))
```

NOT RECOMMEND TO VIEW THE WHOLE DATAFRAME

If you want to take a look at your data do: `View(head(df))`

NCBI Gene Expression Omnibus

COVID-19 is an emerging, rapidly evolving situation. Get the latest public health information from CDC: <https://www.coronavirus.gov>.

Get the latest research from NIH: <https://www.nih.gov/coronavirus>. Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

NCBI > GEO > Accession Display

Scope: Self Format: HTML Amount: Quick GEO

accession: GSE32323

Series GSE32323 Query DataSets for GSE32323

Status Public on Mar 20, 2012

Title Screening for Epigenetically Masked Genes in Colorectal Cancer using 5-aza-2'-deoxycytidine treatment, Microarray and Gene Expression Profile

Platforms (1) [GPL570](#) [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples (44) [GSM800742](#) patient 006b, normal, homogenized

[GSM800743](#) patient 011b, normal, homogenized

[GSM800744](#) patient 024b, normal, homogenized

Relations

BioProject [PRJNA147273](#)

Analyze with GEO2R

Download family	Format
SOFT formatted family file(s)	SOFT @
MINIML formatted family file(s)	MINIML @
Series Matrix File(s)	TXT @

Supplementary file	Size	Download	File type/resource
GSE32323_RAW.tar	211.3 Mb	(http) (custom)	TAR (of CEL)

Raw data provided as supplementary file

Processed data included within Sample table

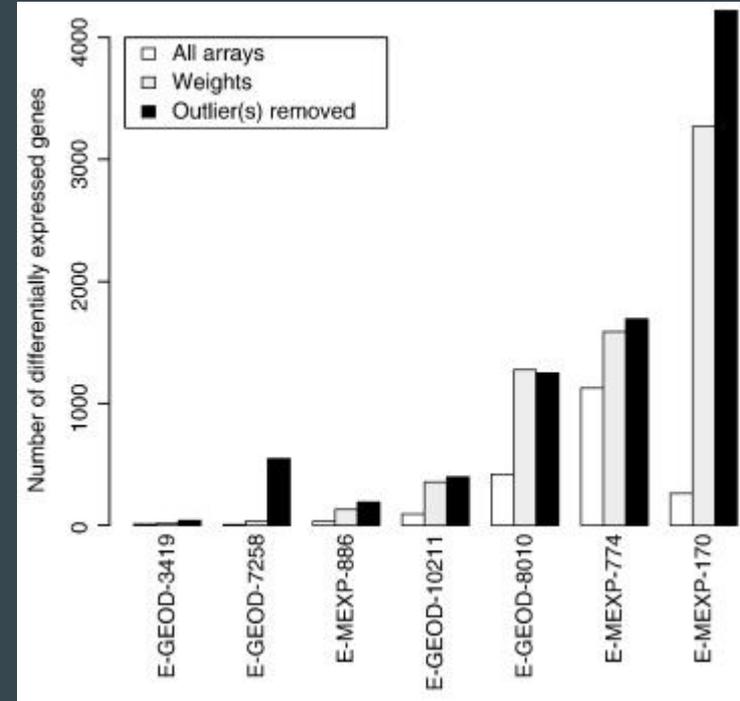
RNA Degradation, Hybridization, and Spike-in

- **Degradation of RNA** itself is a physiological process during the cell cycle to regulate RNA-dependent mechanisms.
 - Therefore, RNA degradation over time may affect the integrity of the samples and data
- **RNA Spike-in:** RNA transcript of known sequence and quantity used to calibrate measurements in RNA hybridization assays,
 - A **hybridization assay** comprises any form of quantifiable hybridization *i.e.* the quantitative annealing of two complementary strands of nucleic acids, known as nucleic acid hybridization.
 - **Spike-in controls** are needed in all types of genome-wide profiling analyses by microarray or sequencing where changes in absolute amounts of the total signal are suspected to occur between different experimental conditions.
- **Why is this relevant?** Such high-throughput methods can be error prone, and known controls are necessary to detect and correct for levels of error. RNA spike-in controls can provide a measure of sensitivity and specificity of an RNA-Seq experiment.

Why is this important?

- Being able to **identify outlier data**
- Help researchers identify the potential **causes of outlier data**
- Decide whether a sample needs to be repeated or removed from the set
- Improving the **signal-to-noise ratio** to improve power of test:
 - Signal-to-noise ratio (SNR) is a way to determine whether or not a peak should be considered when evaluating the data.

****Note**** : systematic variation may be imparted by the experimental process itself and can be difficult to remove through standard normalization - may not be detected.

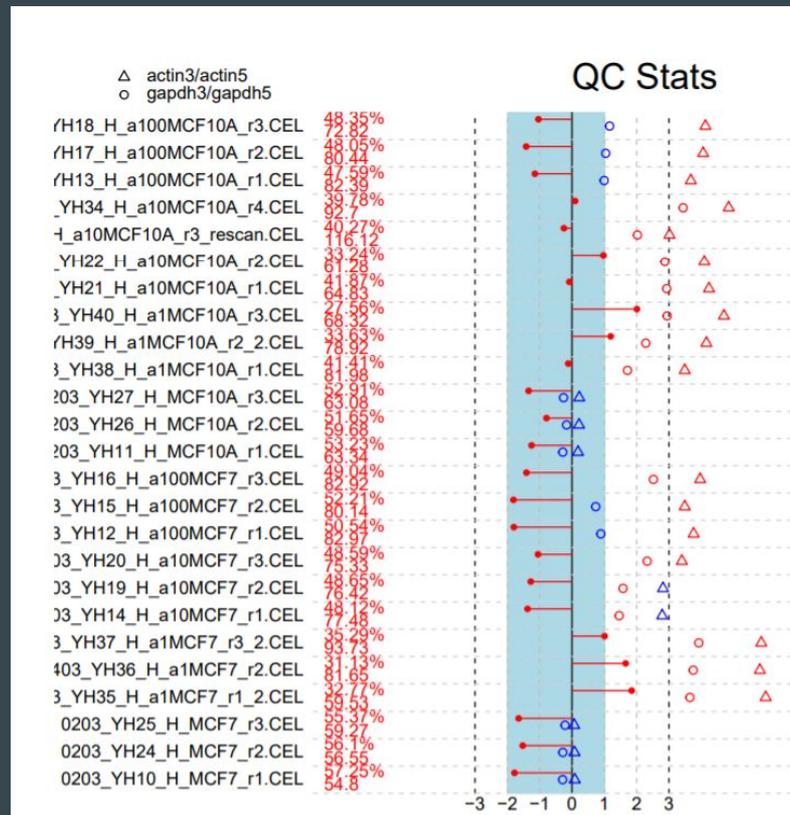


simpleAffy

“Provides high level functions for **reading Affy .CEL files**, **phenotypic data**, and then computing simple things with it, such as **t-tests**, **fold changes** and the like. Also has some basic scatter plot functions and mechanisms for generating high resolution journal figures”

- A **CEL file** is a data **file** created by Affymetrix DNA microarray image analysis software. It contains the data extracted from "probes" on an Affymetrix GeneChip and can store thousands of data points, which may make it large in **file size**

QC Reports: The function qc produces an object of class 'QCStats' containing QC metrics for each array in a project



Affy QC Report

affyQCReport Package Link: [affyQCReport](#)

Example Report: [affyQCExReport](#)

Installation

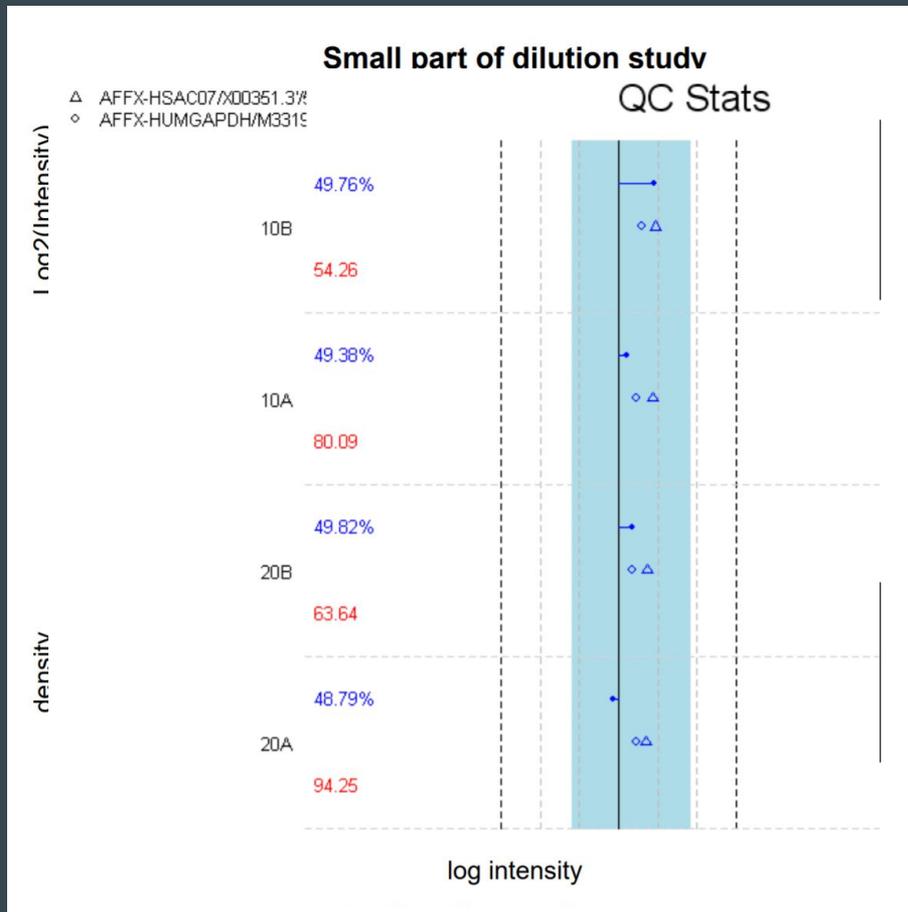
To install this package, start R (version "4.0") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

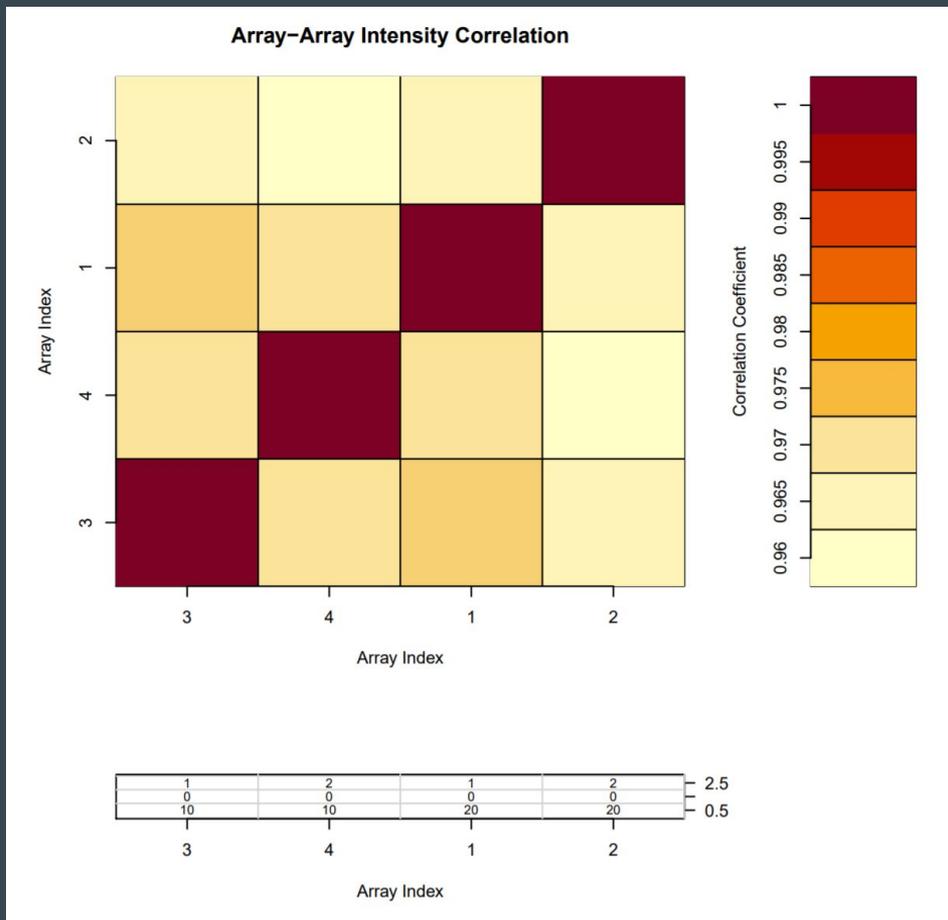
BiocManager::install("affyQCReport")
```

- QC Report for an AffyBatch Object. The QC Report is intended to allow the user to **quickly assess the quality of a set of arrays** in an AffyBatch object
- The report generated consists of 6 pages:
 - 1) **First page:** list of the sample names and an index number that is used to identify each array in later plots.
 - 2) **Second page:** consists of two plots made using the affy package
 - 3) **Third page** is the QC plot generated with the simpleaffy package, which shows the 3' : 5' ratios for spiked-in and control genes specific to the array type.
 - 4) **Fourth page:** boxplots of the intensities of the positive and negative control elements on the outer edges of the Affymetrix arrays.
 - 5) **Fifth Page:** is a plot of the **center of intensity” (COI)** for the positive and negative border elements.
 - 6) **Sixth page:** is a heat map of the array-array **Spearman rank correlation coefficients** of the array intensities.

Affy QC Report [Cont.]



Affy QC Report [Cont.]



affyPLM + Relative Log Expression (RLE)

affyPLM Package Link: [affyPLM](#)

- This package creates a **QC Report for an AffyBatch Object**. The QC Report is intended to allow the user to quickly assess the quality of a set of arrays in an AffyBatch object
- The central focus of this package is on implementing methods for fitting **probe-level models and tools** (i.e. PLM-based Quality Assessment Tools) using these models.
- **Relative Log Expression**, also known as RLE, is a type of quality assessment tool.
 - RLE Values are computed for each probeset by comparing the expression value on each against the median expression value for that probeset across all arrays.
 - Assuming that most genes are not changing in expression across arrays means **ideally most of these RLE values will be near 0.**

Installation

To install this package, start R (version "4.0") and enter:

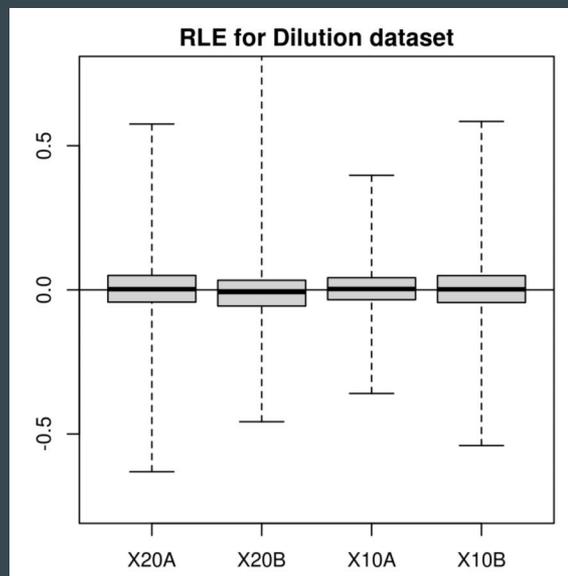
```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("affyPLM")
```

affyPLM + Relative Log Expression (RLE) [Cont.]

- The user can create a boxplot of RLE Values (for each array) using:

```
> RLE(Pset, main="RLE for Dilution dataset")
```



affyPLM + Relative Log Expression (RLE) [Cont.]

- The user can also compute a summary of statistics using:

```
> RLE(Pset, type="stats")
```

```
                20A                20B
median 0.002359773 -0.007542337
IQR    0.092560253  0.089580148
                10A                10B
median 0.003051789  0.001902185
IQR    0.076335304  0.093406833
```

Normalized Unscaled Standard Error

- Useful for understanding how departures in quality of the data affect the expression estimates
- Want to combine residuals into estimated standard errors of expression estimates and summarize those a chip level

$$\text{unscaled SE } (c_{kj}) = 1/\sum_i w_{kij}$$

- For each chip, you get a vector of unscaled standard errors of estimated expression, one for each probe set
- To remove the source of heterogeneity, we can normalize the unscaled standard errors by dividing the average value across the chips.

Different types of normalization methods

MAS5	RMA	GCRMA
Normalizes each array independently and sequentially.	RMA = robust multi-array, uses a multi-chip model. Generally, more commonly used today.	Gene chip RMA
Uses data from mismatch probes to calculate a “robust average”, by subtracting mismatch probe values from match probe values. <u>However, subtracting mismatch data will result in loss of interesting signal in many probes and may cause noise at low intensity levels</u>	Doesn't use mismatch probes, because their intensities (positive values) are generally higher than match probes. Therefore they are unreliable as indicators of non-specific binding. With this method, normalizing at probe level avoids the loss of information	Probe affinity calculated using position dependent base effects. Mismatch data is based on probe affinity, and then subtracted from perfect match. In this way, no MM data is lost
Weights each probe intensity based on the distance from the mean (weighted mean uses one-step Tukey Biweight Estimate)	Data is a combination of background and signal. Assumes a strictly positive distribution for signal.	Combines intensity values from the probes in the probe set to get a single intensity value for each gene
Robust average means that it is insensitive to any small changes made from assumptions	Normalizes across all arrays to make all distributions the same. This protects against outliers	

Advantages & Disadvantages of RMA/GC-RMA vs. MAS5

Advantages	Disadvantages
Gives less false positives	May hide real changes, especially at low expression levels (i.e. false negatives)
Sees less variance at lower expression levels	Makes quality control after normalization more difficult
Provides more consistent fold change estimates	Normalization assumes equal distribution which may hide biological changes.
Inclusion of adjusted MM data in GC-RMA reduces noise, and retains MM data	Less precise than MAS5

Ideal solution - use standard MAS5.0 techniques for quality control. Then go back and perform probe level normalization on quality controlled genes

Principal Component Analysis (PCA)

- PCA is used for [exploratory data analysis](#), allowing you to better visualize the [variation](#) within your dataset of many variables
- Particularly useful with [wide datasets](#), where you have many variables in each sample

```
mtcars.pca <- prcomp(mtcars[,c(1:7,10,11)], center = TRUE, scale. = TRUE)
```

```
summary(mtcars.pca)
```

```
## Importance of components:
```

##	PC1	PC2	PC3	PC4	PC5	PC6
## Standard deviation	2.3782	1.4429	0.71008	0.51481	0.42797	0.35184
## Proportion of Variance	0.6284	0.2313	0.05602	0.02945	0.02035	0.01375
## Cumulative Proportion	0.6284	0.8598	0.91581	0.94525	0.96560	0.97936

##	PC7	PC8	PC9
## Standard deviation	0.32413	0.2419	0.14896
## Proportion of Variance	0.01167	0.0065	0.00247
## Cumulative Proportion	0.99103	0.9975	1.00000

