

# Bioconductor overview

Mikhail Dozmorov

Fall 2018

# Bioconductor Project

- The Bioconductor project started in 2001

**Goal:** make it easier to conduct reproducible consistent analysis of data from new high-throughput biological technologies.

- Core maintainers of the Bioconductor website located at Fred Hutchinson Cancer Research Center.
- Updated version released biannually coinciding with the release of R.
- Many contributed software packages.

# Bioconductor packages

- Bioconductor software consists of R add-on packages.
- An R package is a structured collection of code (R, C, or other), documentation, and/or data for performing specific types of analyses.
- E.g. **affy**, **limma**, **sva** packages provide implementations of specialized statistical and graphical methods.

# Goals of the Bioconductor Project

- Provide access to statistical and graphical tools for analysis of high-dimensional biological data.
- ① Microarray analysis.
- ② High-throughput 'omics' data.

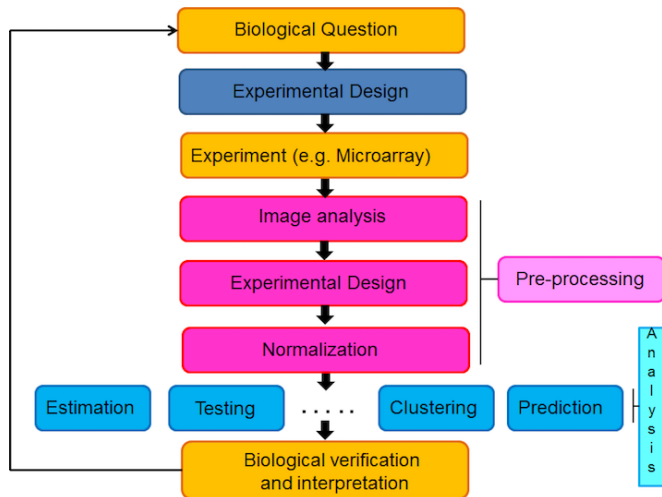
# Goals of the Bioconductor Project

- Include comprehensive **documentation** describing and providing examples for packages.
- Packages have associated **vignettes** that provide examples of how to use functions.
- Have additional tools to work with publically available databases and other meta-data.

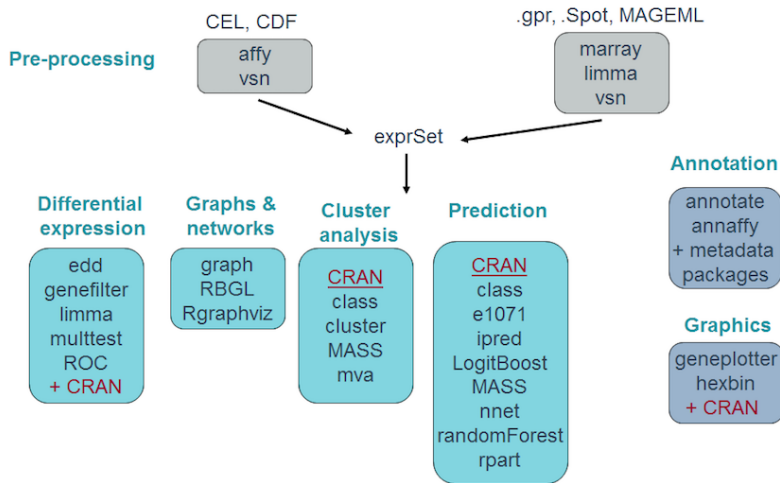
# Vignettes

- Bioconductor has adopted a new documentation paradigm, the vignette.
- A vignette is an executable document consisting of a collection of documentation text and code chunks.
- Vignettes form dynamic, integrated, and reproducible statistical documents that can be automatically updated if either data or analyses are changed.
- Vignettes can be generated using the `sweave` function (or, `roxygen2` package).

# Microarray data analysis



# Microarray data analysis





# Bioconductor website

Lets take a look at the website. . .

<http://bioconductor.org/>

# marrayRaw class

Pre-normalization intensity data for a batch of arrays

`maRf`

`maGf`

Matrix of red and green foreground intensities

`maRb`

`maGb`

Matrix of red and green background intensities

`maW`

Matrix of spot quality weights

`maLayout`

Array layout parameters - `marrayLayout`

`maGnames`

Description of spotted probe sequences  
- `marrayInfo`

`maTargets`

Description of target samples - `marrayInfo`

`maNotes`

Any notes

# AffyBatch class

Probe-level intensity data for a batch of arrays (same CDF)

<code>cdfName</code>	Name of <b>CDF</b> file for arrays in the batch	
<code>nrow</code>	<code>ncol</code>	Dimensions of the array
<code>exprs</code>	<code>se.exprs</code>	Matrices of probe-level intensities and SEs rows → probe cells, columns → arrays.
<code>phenoData</code>	Sample level covariates, instance of class <code>phenoData</code>	
<code>annotation</code>	Name of annotation data	
<code>description</code>	MIAME information	
<code>notes</code>	Any notes	

# ExpressionSet class

Processed Affymetrix or spotted array data

`exprs`

Matrix of expression measures, genes x samples

`se.exprs`

Matrix of SEs for expression measures, genes x samples

`phenoData`

Sample level covariates, instance of class `phenoData`

`annotation`

Name of annotation data

`description`

MIAME information

`notes`

Any notes

## Minimum Information About a Microarray Experiment (MIAME)

The six most critical elements contributing towards MIAME are:

- 1 The raw data for each hybridization (e.g., CEL or GPR files).
- 2 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).
- 3 The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)

<http://fged.org/projects/miame/>

## Minimum Information About a Microarray Experiment (MIAME)

- ④ 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)
- ⑤ Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
- ⑥ The essential laboratory and data processing protocols (e.g., what normalization method has been used to obtain the final processed data)

<http://fged.org/projects/miame/>

# Pre-processing packages

- 1 marray: Spotted DNA microarrays.
  - 2 affy: Affymetrix oligonucleotide chips.
  - 3 limma: all, from spotted arrays to Affy to RNA-seq.
- Reading in intensity data, diagnostic plots, normalization, computation of expression measures.
  - The packages start with very different data structures, but produce similar objects of class `ExpressionSet`.
  - One can then use other Bioconductor and CRAN packages for exploratory data analysis and visualization, differential expression detection.

# Pre-processing packages

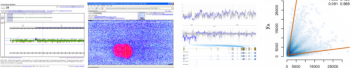
The `aroma.affymetrix` package is an R package for analyzing small to extremely large Affymetrix data sets. It allows you to analyze any number of arrays of various chip types, e.g. 10,000s of expression arrays, SNP chips, exon arrays and so on. Allows for alternative splicing analysis, copy number variations, among other options.

[Aroma](#) [Get Started](#) [Documentation](#) [Features](#) [Resources](#) [Forum](#) [FAQ](#) [Developers' Corner](#) [About](#)

News and recent updates:

- Sept 13, 2017: `aroma.core 3.1.1` released (fixes minor incompatibility with new `future 1.6.1`).
- June 3, 2017: `future 1.5.0` and `future.batchtools 0.5.0` (provides parallel and compute-cluster processing)
- March 23, 2017: `aroma.affymetrix 3.1.0` released.
- Nov 11, 2016: `PSCBS 0.62.0` released.
- Jan 11, 2016: The Aroma Project turns 10 years!
- Oct 28, 2015: `aroma.cn 1.6.1` released.

## An open-source R framework for your microarray analysis



<http://www.aroma-project.org/>