

An Introduction to R and Bioconductor

Robert Gentleman and Seth Falcon
Program in Computational Biology
Fred Hutchinson Cancer Research Center

© Copyright 2006, all rights reserved

Overview

- **biology is a computational science**
- **problems of data analysis, data generation, reproducibility require computational support and computational solutions**
- **we value code reuse**
 - many of the tasks have already been solved
 - if we use those solutions we can put effort into new research
- **well designed, self-describing data structures help us deal with complex data**

Goals

- Provide access to powerful statistical and graphical methods for the analysis of genomic data.
- Facilitate the integration of biological metadata (**GenBank, GO, Entrez Gene, PubMed**) in the analysis of experimental data.
- Allow the rapid development of extensible, interoperable, and scalable software.
- Promote high-quality documentation and reproducible research.
- Provide training in computational and statistical methods.

Bioconductor

- Bioconductor is an **open source** and **open development** software project for the analysis of biomedical and genomic data.
- The project was started in the Fall of 2001 and includes core developers in the US, Europe, and Australia.
- **R** and the **R package system** are used to design and distribute software.
- A goal of the project is to develop software modules that are integrated and which make use of available web services to provide comprehensive software solutions to relevant problems.
- **ArrayAnalyzer**: Commercial port of Bioconductor packages in S-Plus.

Why are we Open Source

- so that you can find out what algorithm is being used, and how it is being used
- so that you can modify these algorithms to try out new ideas or to accommodate local conditions or needs
- so that they can be used as components (potentially modified)

Bioconductor packages

Release 2.0, May, 2007

214 Packages!

- **General infrastructure:**
Biobase, DynDoc, tkWidgets, widgetTools, BioStrings, multtest
- **Annotation:**
annotate, annaffy, biomaRt, AnnBuilder → data packages.
- **Graphics:**
geneplotter, hexbin
- **Pre-processing Affymetrix oligonucleotide chip data:**
affy, affycomp, affydata, makecdfenv, vsn, gcrma
- **Pre-processing two-color spotted DNA microarray data:**
marray, vsn, arrayMagic, arrayQuality
- **Differential gene expression:**
eddi, genefilter, limma, ROC, siggenes, EBArrays, factDesign
- **GSEA/Hypergeometric Testing**
Category, G0stats, topGO
- **Graphs and networks:**
graph, RBGL, Rgraphviz
- **Flow Cytometry:**
prada, flowCore, flowViz, flowUtils
- **Protein Interactions:**
ppiData, ppiStats, ScISI, Rintact
- **Other data:**
SAGElyzer, DNACopy, PROcess, aCGH

Component software

- **most interesting problems will require the coordinated application of many different techniques**
- **thus we need integrated interoperable software**
- **web services are one tool**
- **well designed software modules are another**
- **you should design your piece to be a cog in a big machine**

Data complexity

- Dimensionality.
- Dynamic/evolving data: e.g., gene annotation, sequence, literature.
- Multiple data sources and locations: in-house, WWW.
- Multiple data types: numeric, textual, graphical.

No longer $X_{n \times p}$!

We distinguish between biological metadata and experimental metadata.

Experimental metadata

- **Gene expression measures**
 - scanned images, i.e., raw data;
 - image quantitation data, i.e., output from image analysis;
 - normalized expression measures,
 - Reliability/quality information for the expression measures.
- **Information on the probe sequences printed on the arrays (array layout).**
- **Information on the target samples hybridized to the arrays.**
- **See Minimum Information About a Microarray Experiment (MIAME) standards and the `MAGEML` package.**

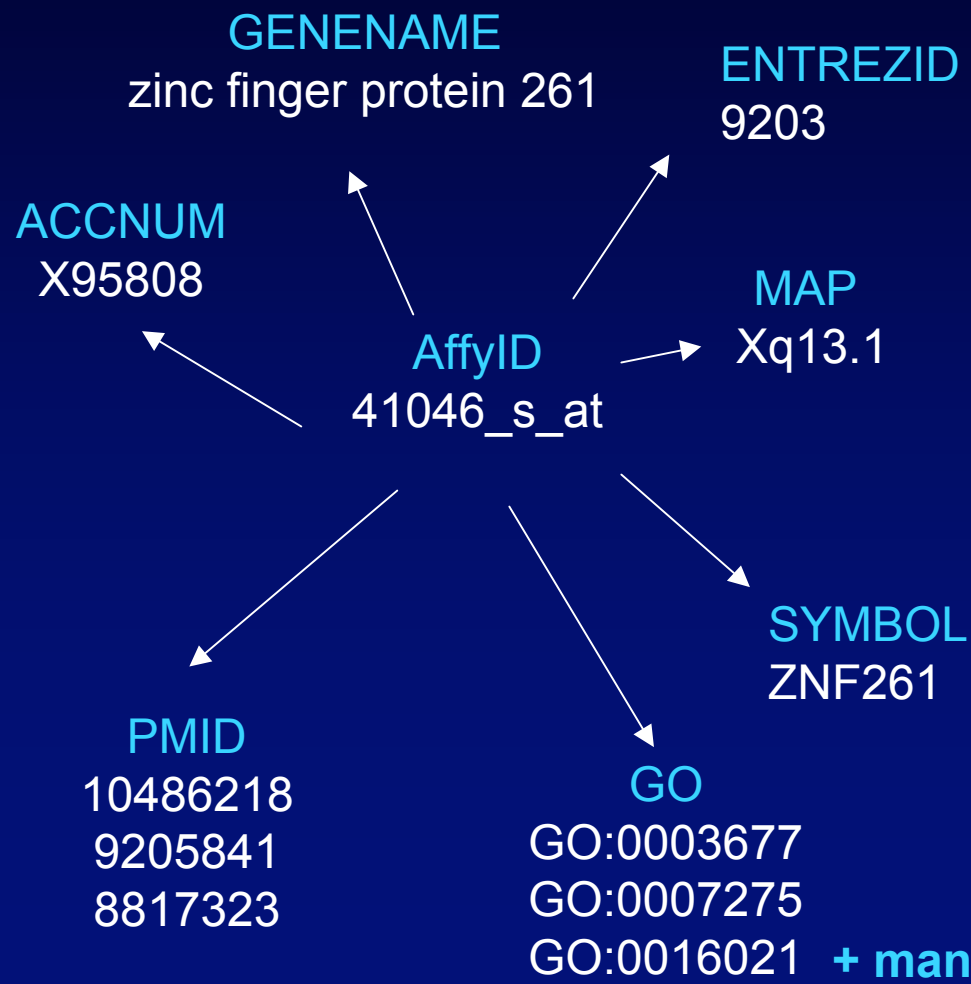
Biological metadata

- **Biological attributes that can be applied to the experimental data.**
- **E.g. for genes**
 - chromosomal location;
 - gene annotation (Entrez Gene, GO);
 - relevant literature (PubMed).
- **Biological metadata sets are large, evolving rapidly, and typically distributed via the WWW.**
- **Tools:** `annotate`, `annaffy`, `biomaRt`, and `AnnBuilder` **packages, and annotation data packages.**

Annotation packages

annotate, annafy, biomaRt, and AnnBuilder

Metadata package `hgu95av2` mappings between different gene IDs for this chip.



- Assemble and process genomic annotation data from public repositories.
- Build annotation data packages.
- Associate experimental data in real time to biological metadata from web databases such as GenBank, GO, KEGG, Entrez Gene, and PubMed.
- Process and store query results: e.g., search PubMed abstracts.
- Generate HTML reports of analyses.

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 from the R `tools` package.

Short Courses/Conferences

- we have given many short courses
 - see bioconductor.org for more details on upcoming courses
- **BioC2007 - Seattle, Aug 6th-8th**
- **BioC Training: Chicago, early Oct**

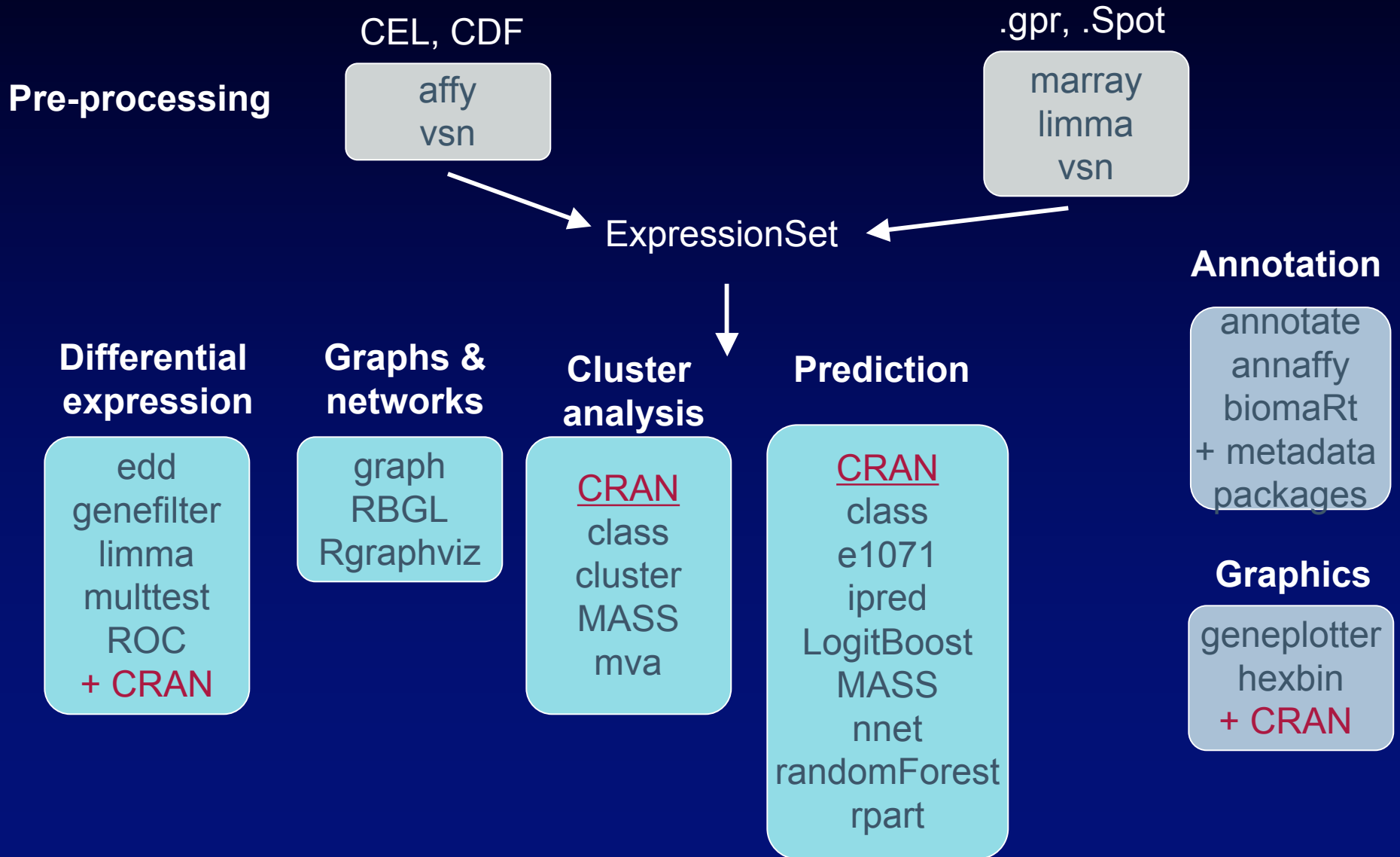
Bioconductor Software

- we concentrate our development on a few important aspects
- **Biobase**: core classes and definitions that allow for succinct description and handling of the data
- **annotate**: generic functions for annotation that can be specialized
- **genefilter**: fast filtering via virtually every mechanism
- **graph/Rgraphviz/RBGL**: code for handling graphs and networks

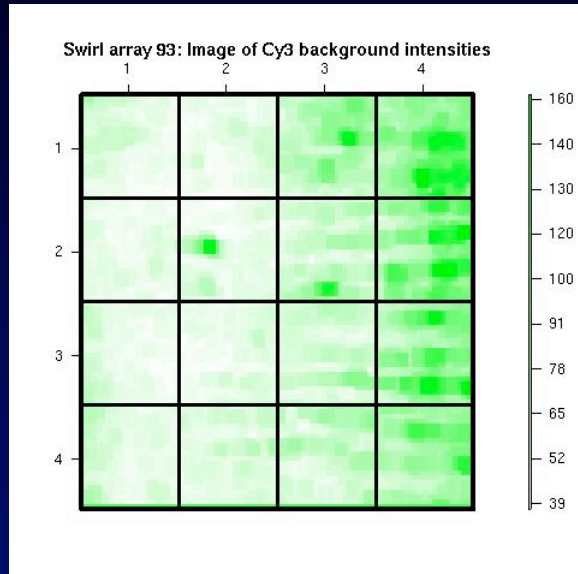
Biobase:ExpressionSet

- software should help organize and manipulate your data
- this was the intention of the original exprSet class
- the data need to be assembled correctly once, and then they can be processed, subset etc without worrying about them
- exprSet was too limited (and too oriented to single channel arrays)
- we developed the new ExpressionSet class

Microarray data analysis

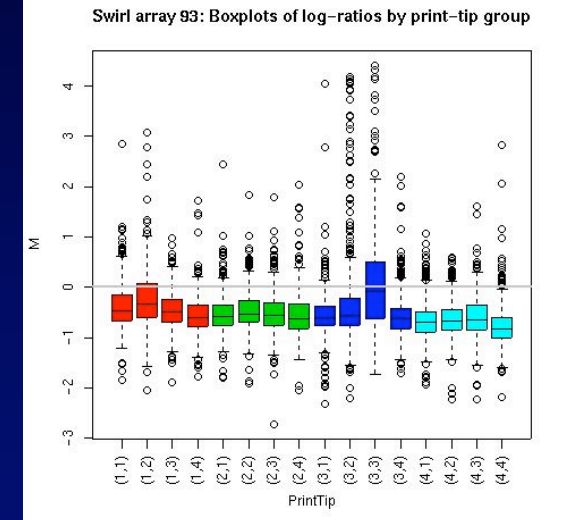
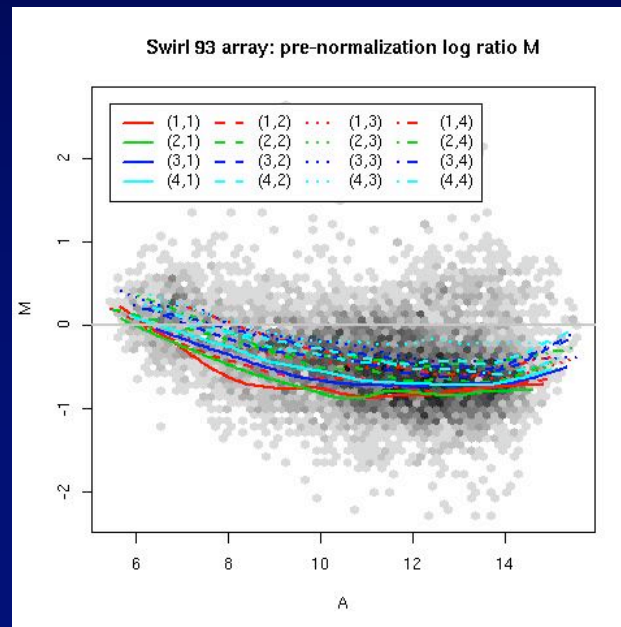


array packages



Pre-processing two-color spotted array data:

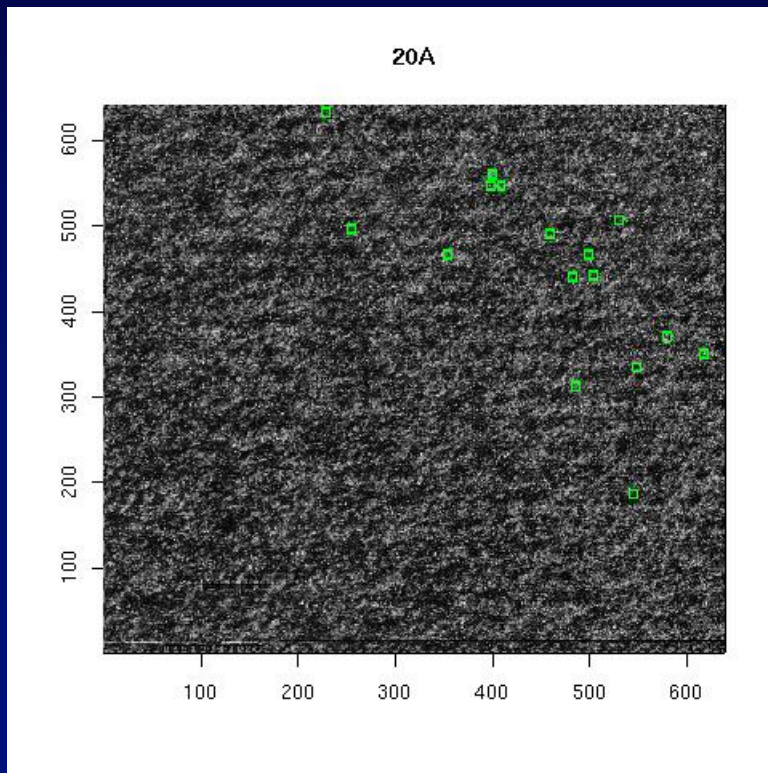
- diagnostic plots,
- robust adaptive normalization (loess).



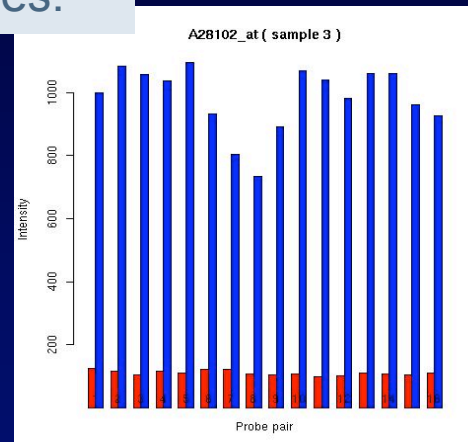
affy package

Pre-processing oligonucleotide chip data:

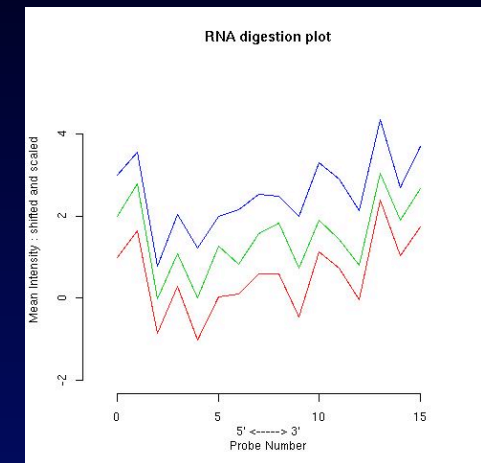
- diagnostic plots,
- background correction,
- probe-level normalization,
- computation of expression measures.



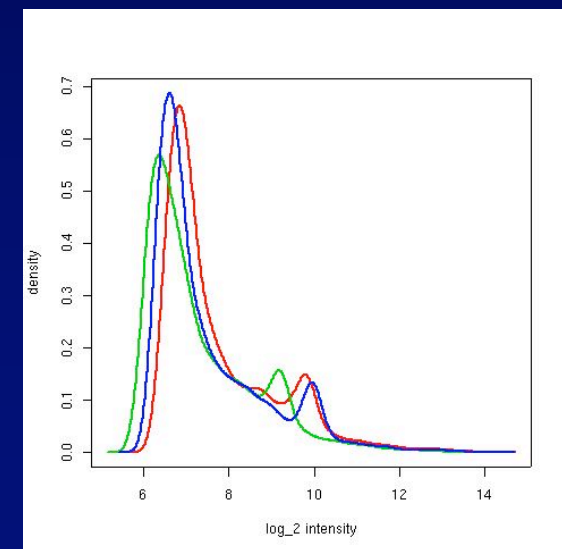
image



barplot.ProbeSet

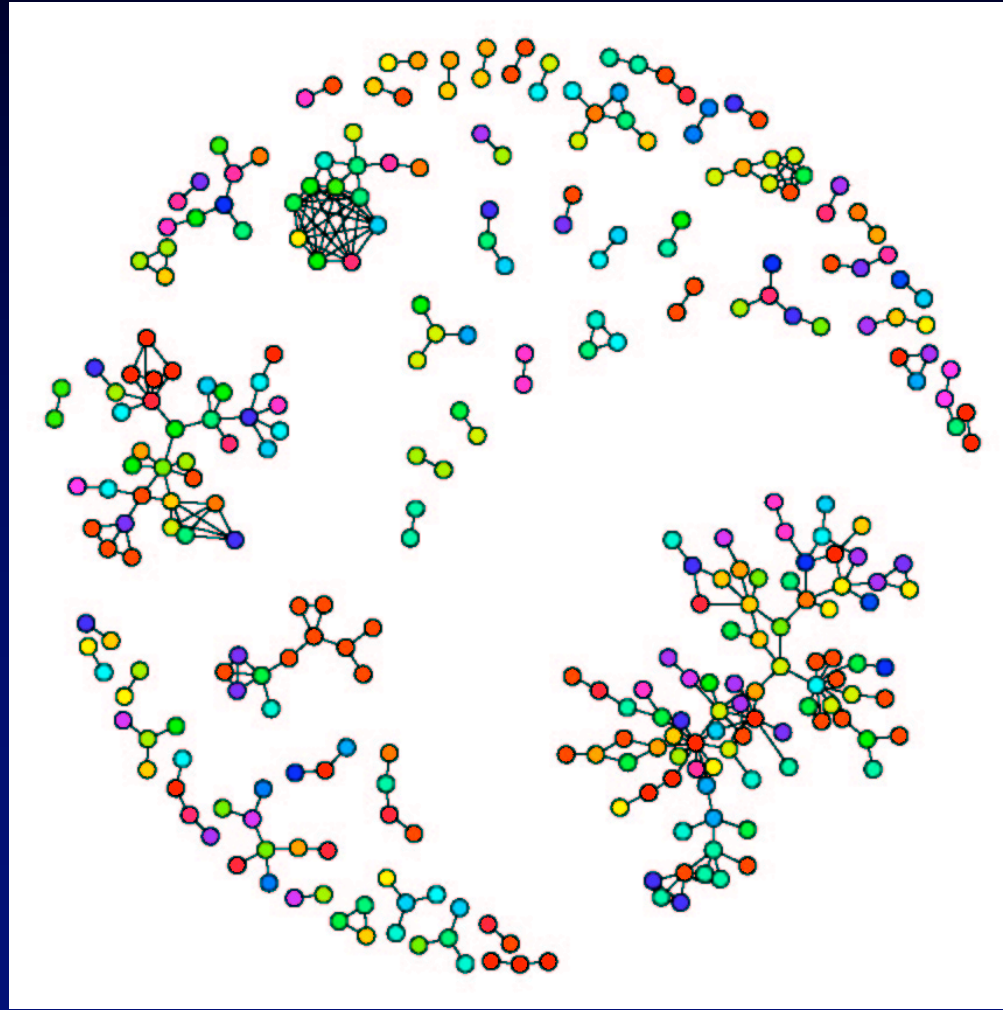


plotAffyRNAdeg



plotDensity

graph and Rgraphviz



apComplex

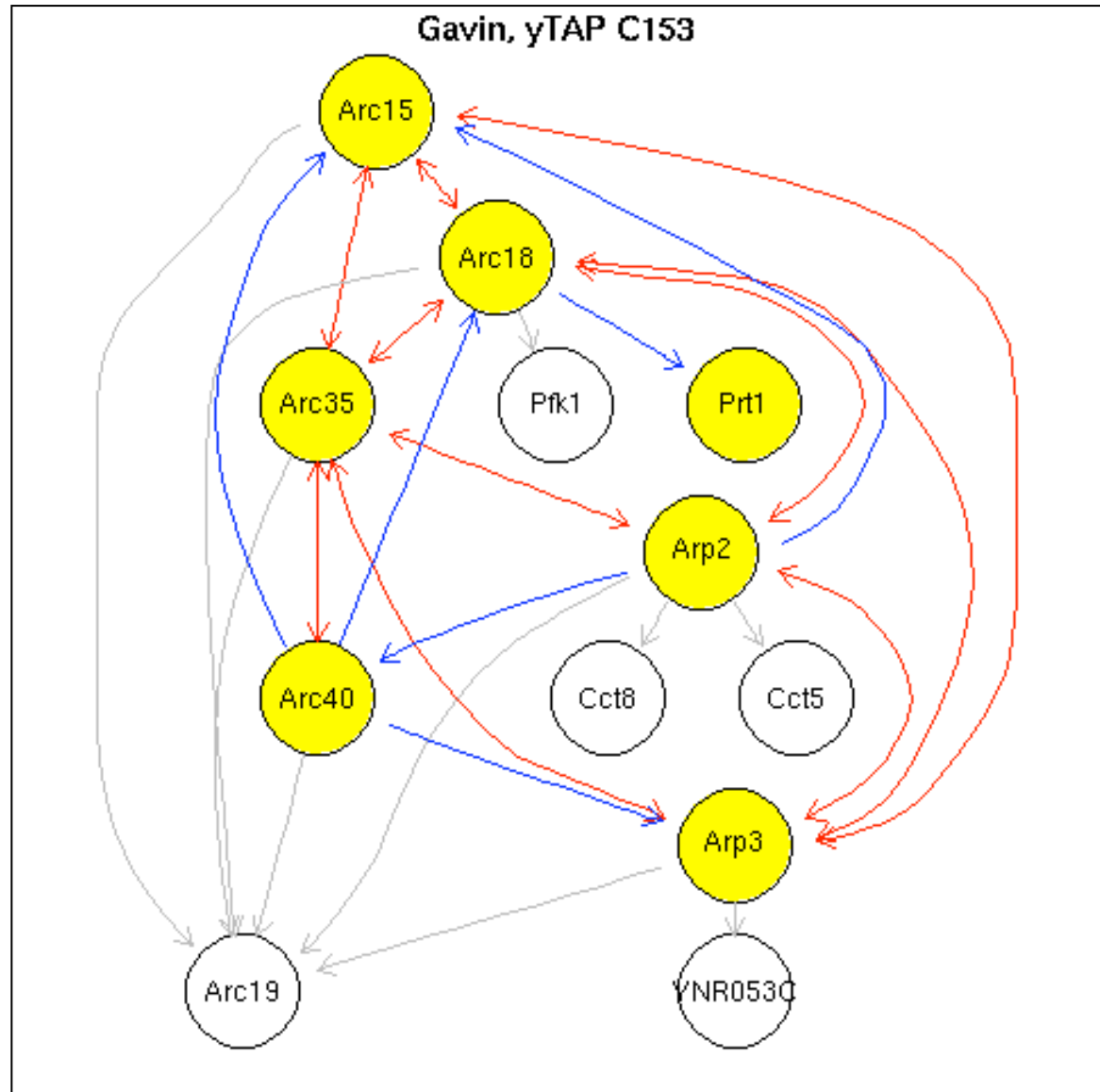
Arp2/3

Arp2/3 complex:

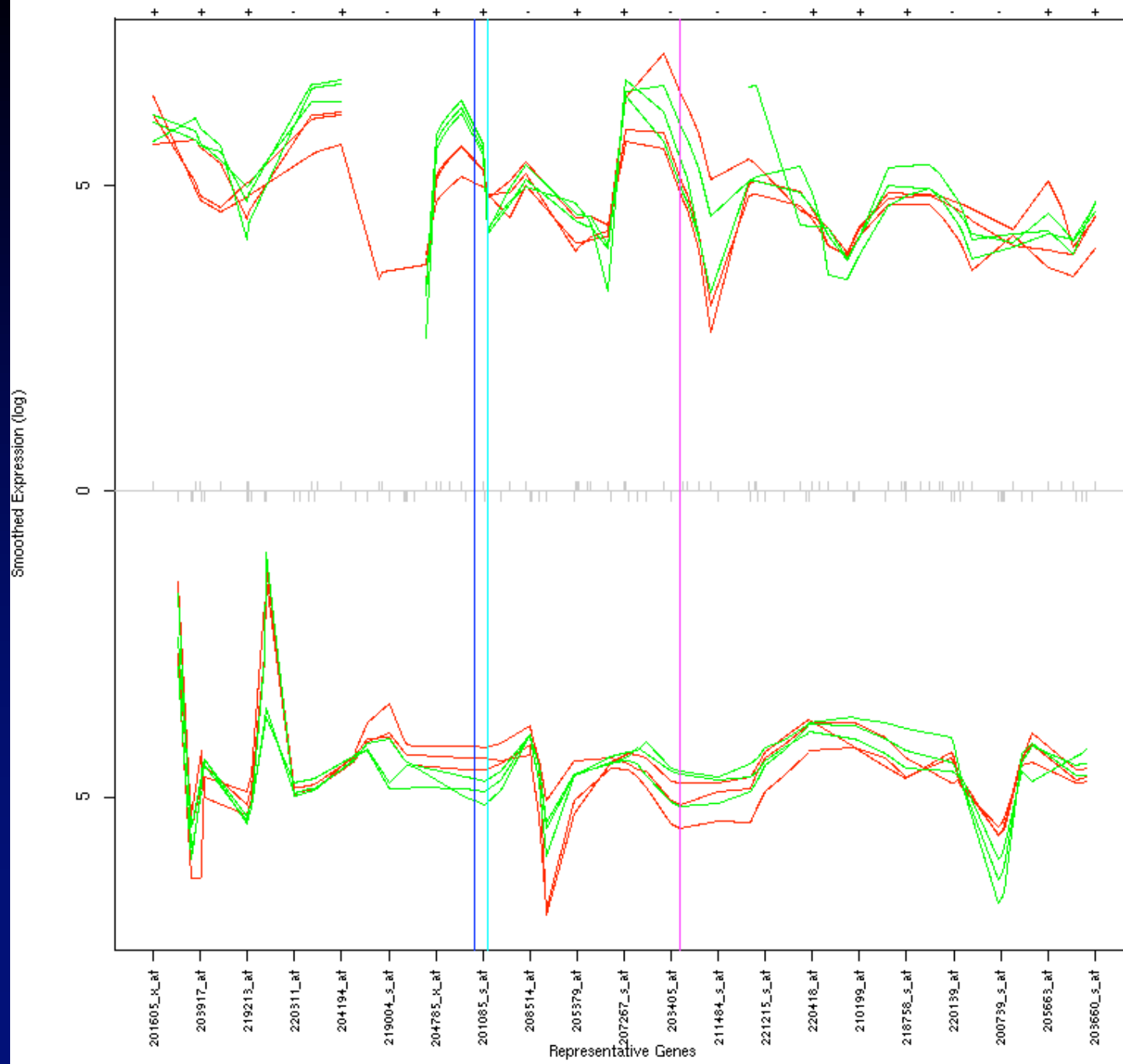
- Arp2
- Arp3
- Arc15
- Arc18
- Arc19
- Arc35
- Arc40

'The Arp2/3 complex is a stable multiprotein assembly required for the nucleation of actin filaments in all eukaryotic cells and consists of seven proteins in human and yeast.'

Winter, et al (1997). *Curr Biol.*
Higgs and Pollard (2001). *Annu Rev Biochem.*



Chromosome 21



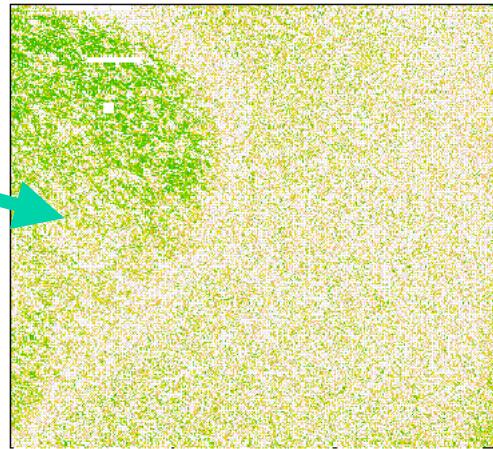
Quality Assessment using residuals from RMA

- Probe level models quantities useful for assessing chip quality
 - Weights
 - Residuals
 - Standard Errors
- Expression values relative to median chip

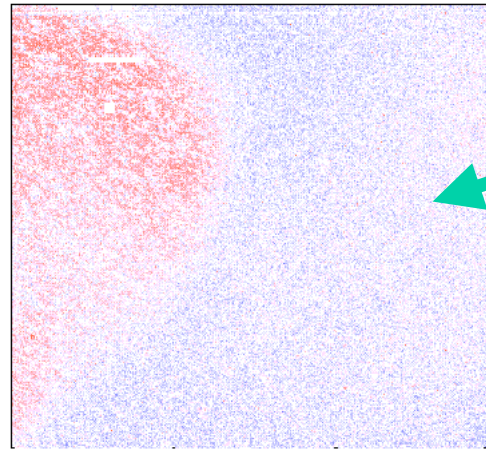
Available from the [affyPLM](#) package

Pseudo-chip images

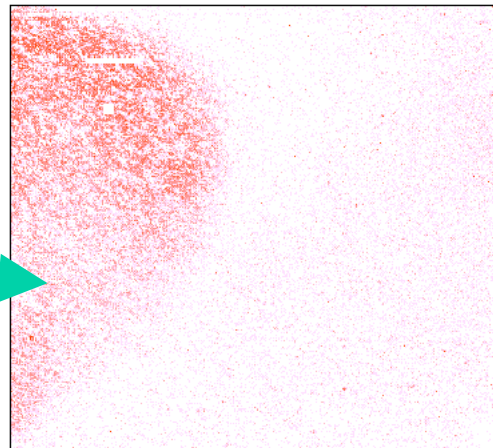
Weights



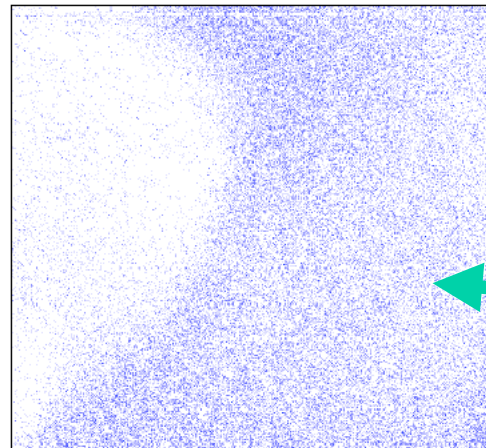
Residuals



Positive
Residuals



Negative
Residuals



Machine Learning

- A new machine learning package
MLInterfaces
- goal is to provide uniform calling sequences and return values for all machine learning algorithms
- we have postpended a B (e.g. knnB)
- return values are of class **classifOutput**
- see the **MLInterfaces** vignette for more details

Publications

- **Bioconductor: Open software development for computational biology and bioinformatics, *Genome Biology* 2004, 5:R80, <http://genomebiology.com/2004/5/10/R80>**
- **The Analysis of Gene Expression Data: Methods and Software, Springer, 2003, G. Parmigiani, E. S. Garrett, R. A. Irizarry and S. L. Zeger eds.**
- **Bioinformatics and Computational Biology Solutions using R and Bioconductor, Springer, 2005, R. Gentleman, V. Carey, W. Huber, R. Irizarry, S. Dudoit eds.**

References

- **R** www.r-project.org, cran.r-project.org
 - software (CRAN);
 - documentation;
 - newsletter: R News;
 - mailing list.
- **Bioconductor** www.bioconductor.org
 - software, data, and documentation (vignettes);
 - training materials from short courses;
 - mailing list (please read the posting guide)

Acknowledgments

- **Bioconductor core team:**
- Ben Bolstad, **UC Berkeley**
- Vince Carey, **Channing Laboratory, Harvard**
- Sandrine Dudoit, **Biostatistics, UC Berkeley**
- Seth Falcon, **FHCRC**
- Robert Gentleman, **FHCRC**
- Wolfgang Huber, **European Bioinformatics Institute**
- Rafael Irizarry, **Biostatistics, Johns Hopkins**
- Li Long, **ISB, Laussane**
- Jim MacDonald, **Michigan**
- Crispin Miller, **PICR**
- Martin Morgan, **FHCRC**
- Herve Pages, **FHCRC**
- Gordon Smyth, **WEHI**
- Yee Hwa (Jean) Yang, **Sydney**