

Biostatistics and Computational Biology

Some selected examples . . . and a bit of R and Bioconductor

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

December 7, 2009

Bioinformatics and computational biology

Wikipedia:

Bioinformatics and computational biology involve the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve biological problems usually on the molecular level.

. . .

Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution.

Wikipedia:

The terms bioinformatics and computational biology are often used interchangeably. However [bioinformatics](#) more properly refers to the creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems inspired from the management and analysis of biological data. [Computational biology](#), on the other hand, refers to hypothesis-driven investigation of a specific biological problem using computers, carried out with experimental or simulated data, with the primary goal of discovery and the advancement of biological knowledge.

Bioinformatics and computational biology

NIH definition of Bioinformatics and Computational Biology:

Bioinformatics and computational biology are rooted in life sciences as well as computer and information sciences and technologies. Both of these interdisciplinary approaches draw from specific disciplines such as mathematics, physics, computer science and engineering, biology, and behavioral science.

...

Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface.

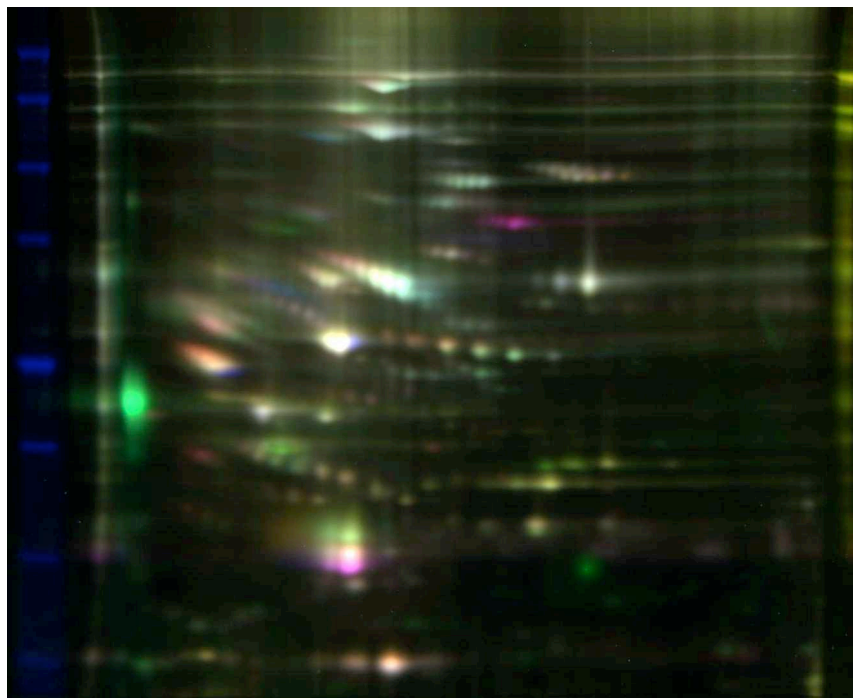
NIH definition of Bioinformatics and Computational Biology:

The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

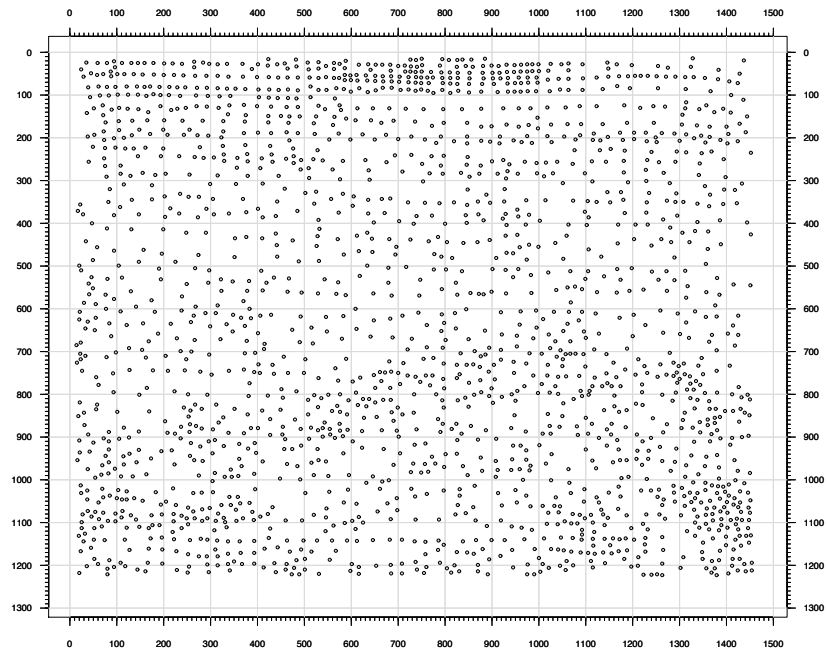
Bioinformatics: *Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.*

Computational Biology: *The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.*

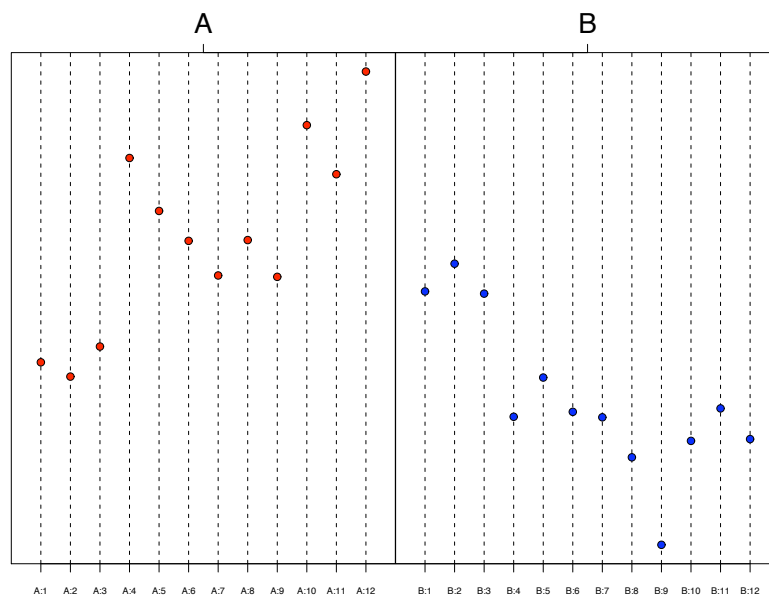
2D gel electrophoresis



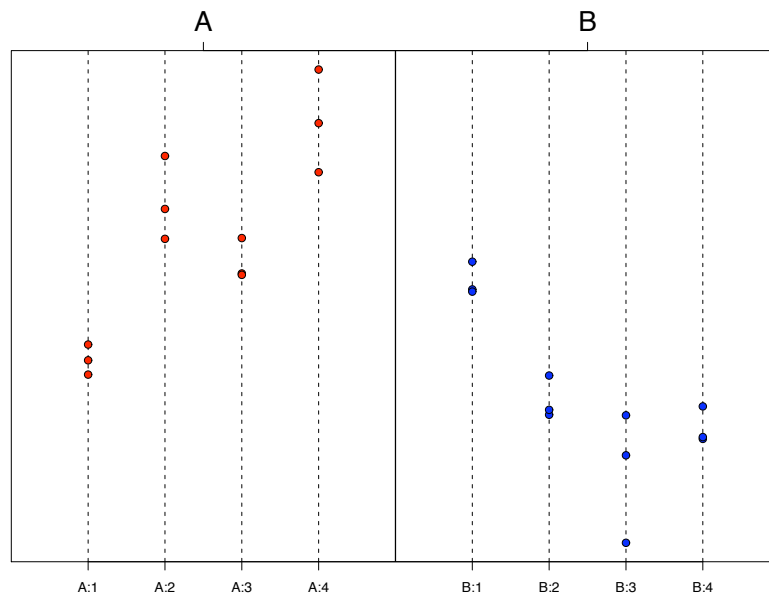
2D gel electrophoresis



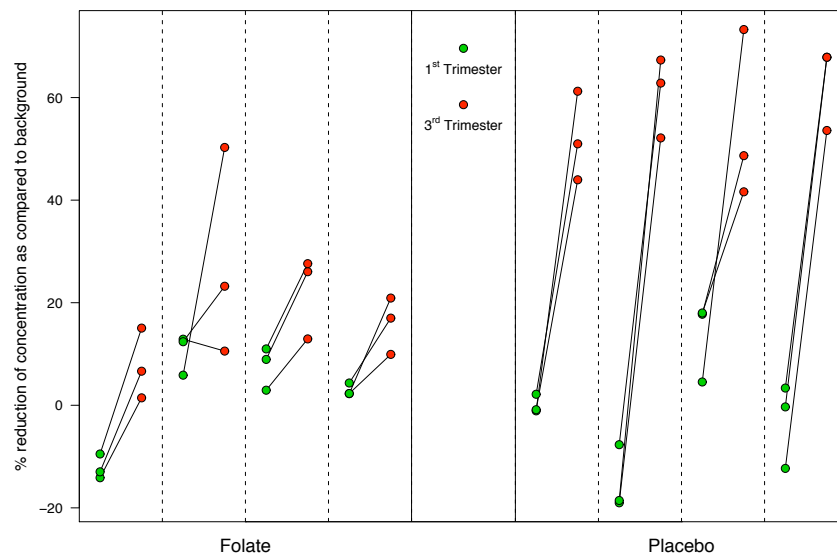
2D gel electrophoresis



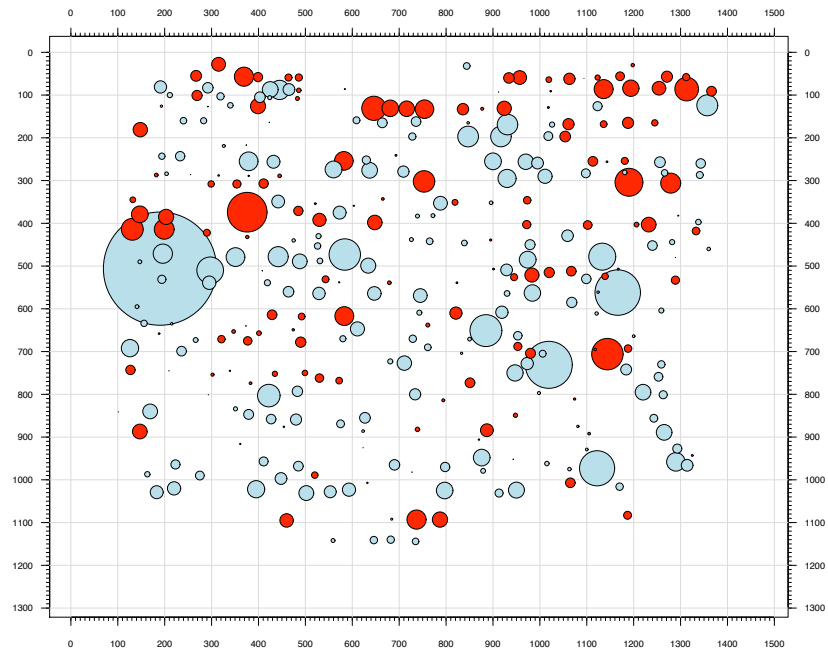
2D gel electrophoresis



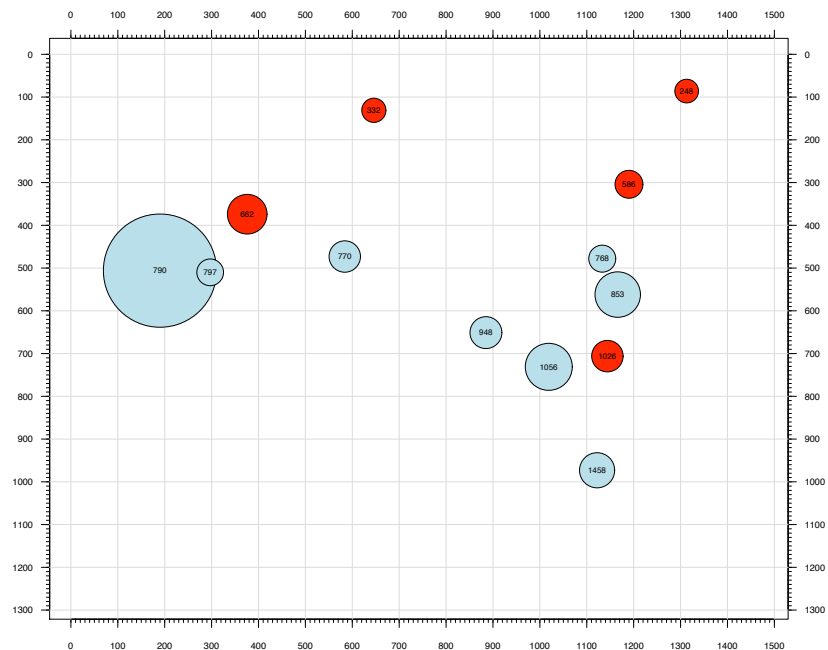
2D gel electrophoresis



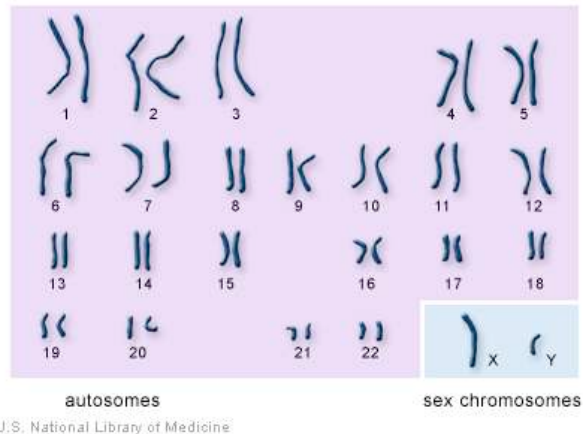
2D gel electrophoresis



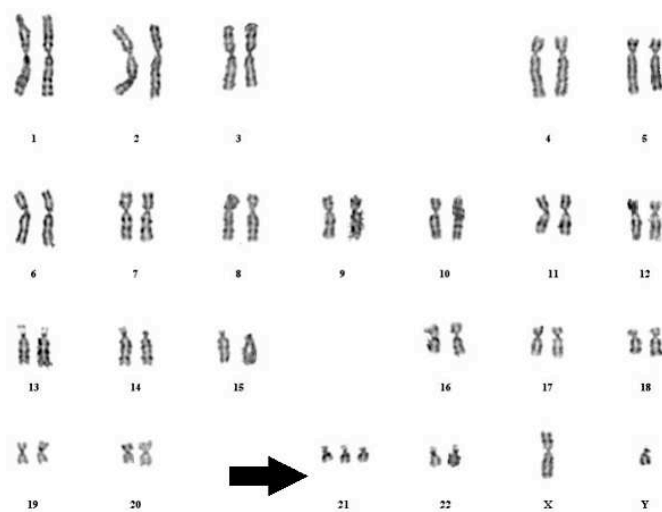
2D gel electrophoresis



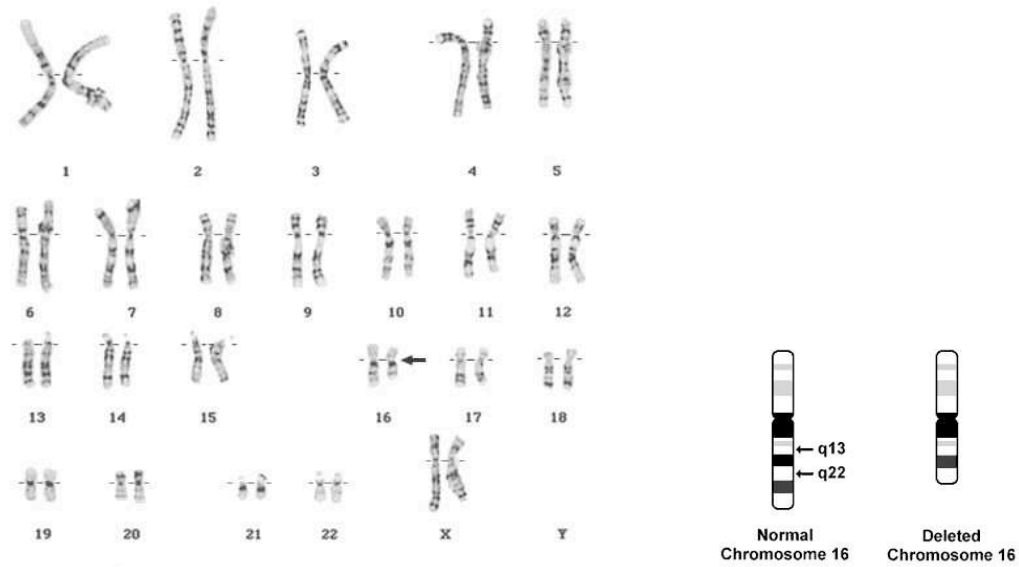
Karyotypes



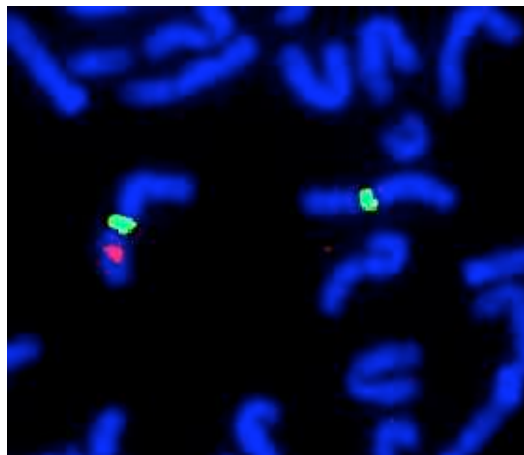
Trisomy



Karyotypes

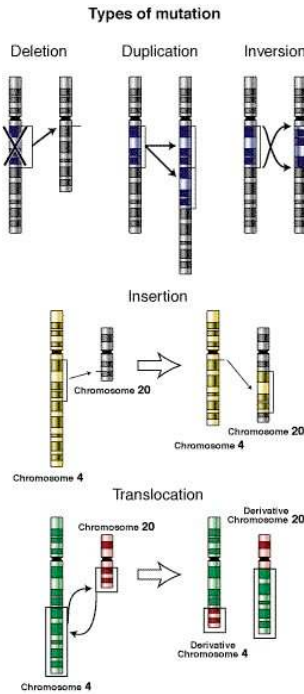


FISH

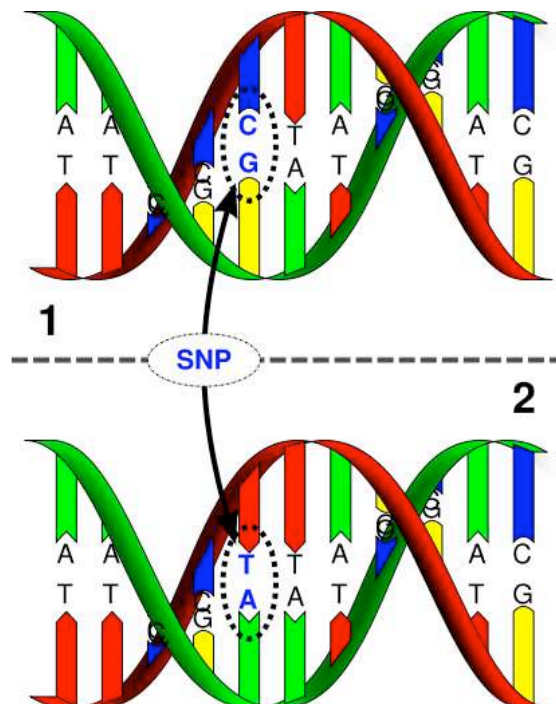


Courtesy of the Pevsner Laboratory

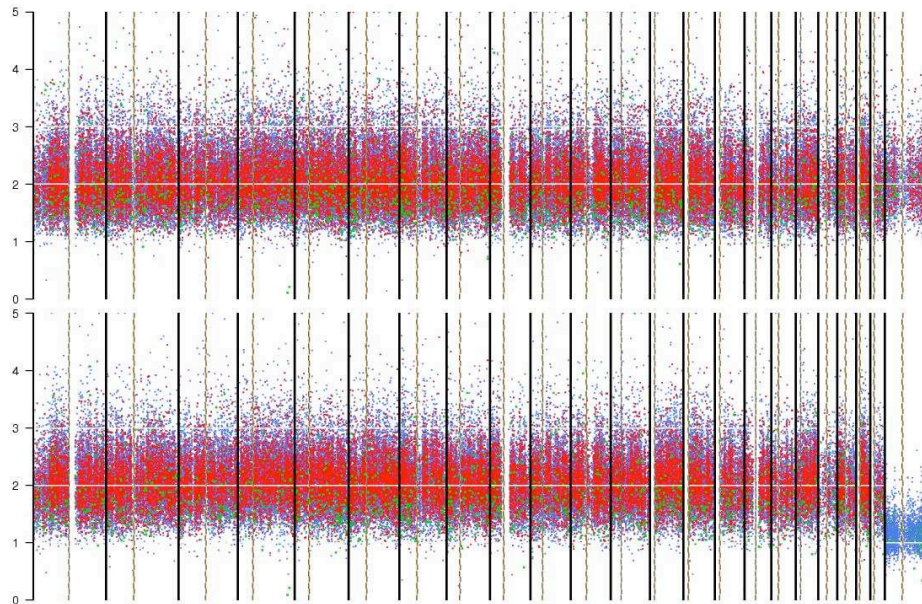
DNA changes



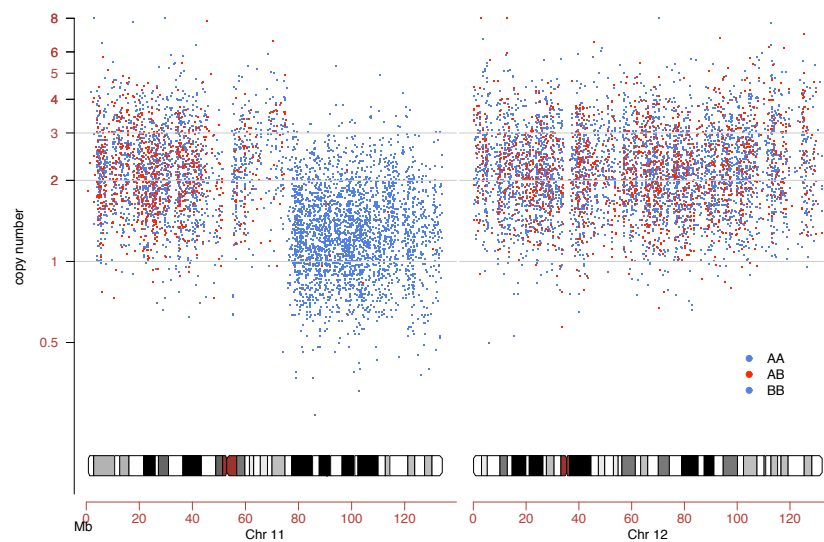
Single nucleotide polymorphisms



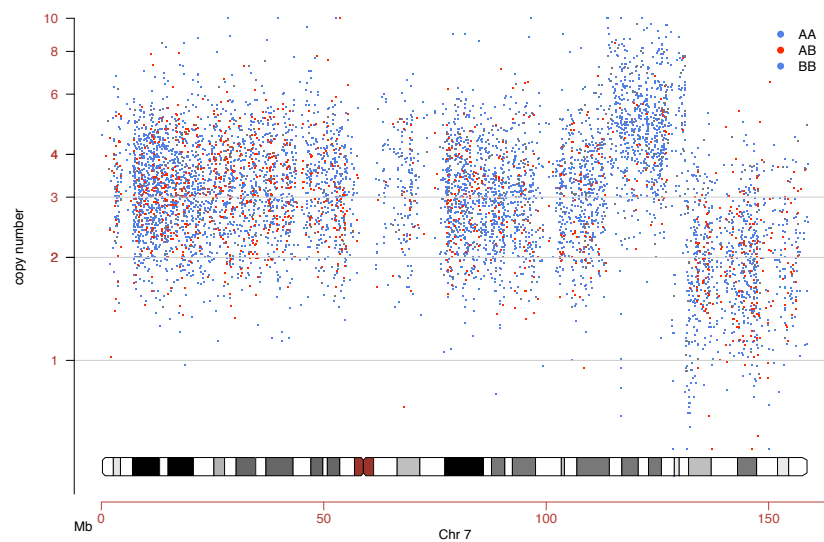
SNP chip data



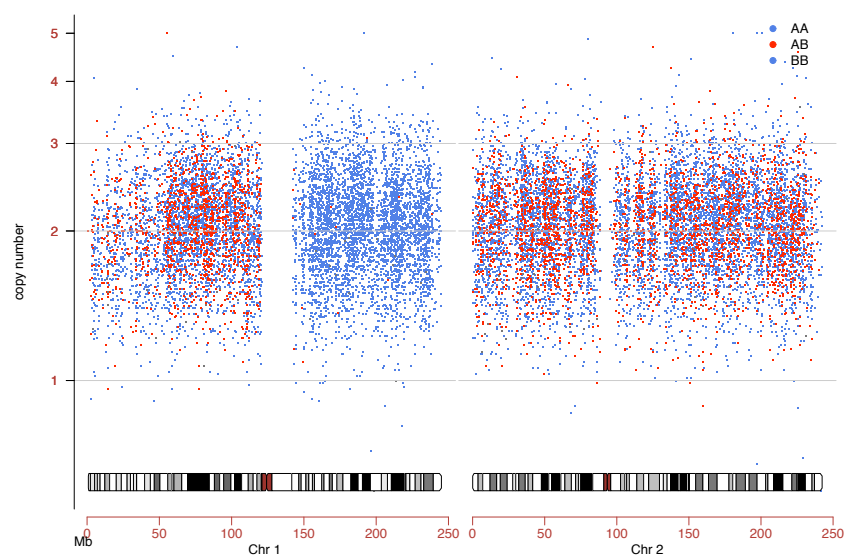
Deletion



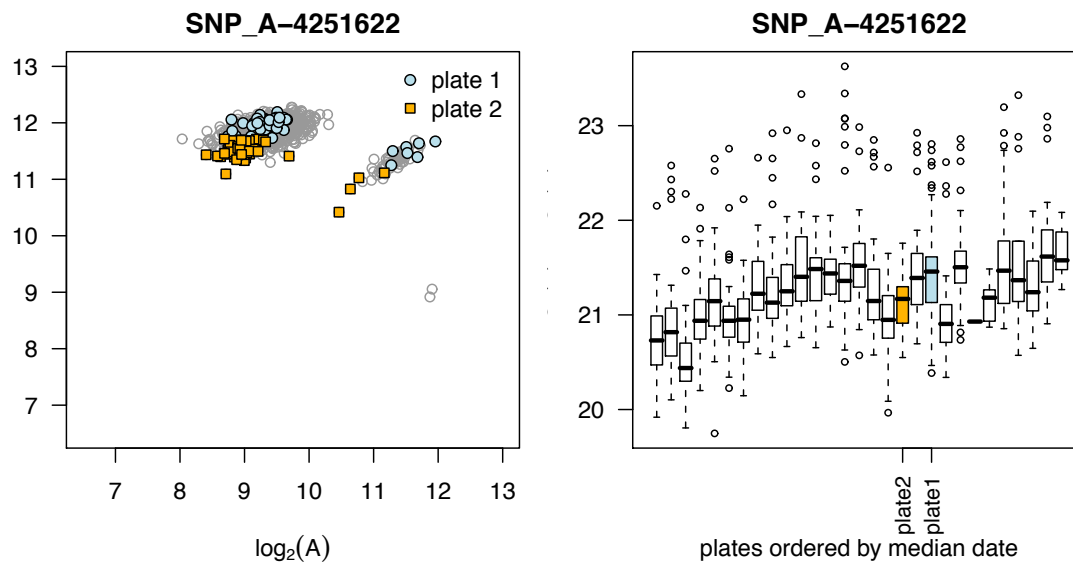
Amplification



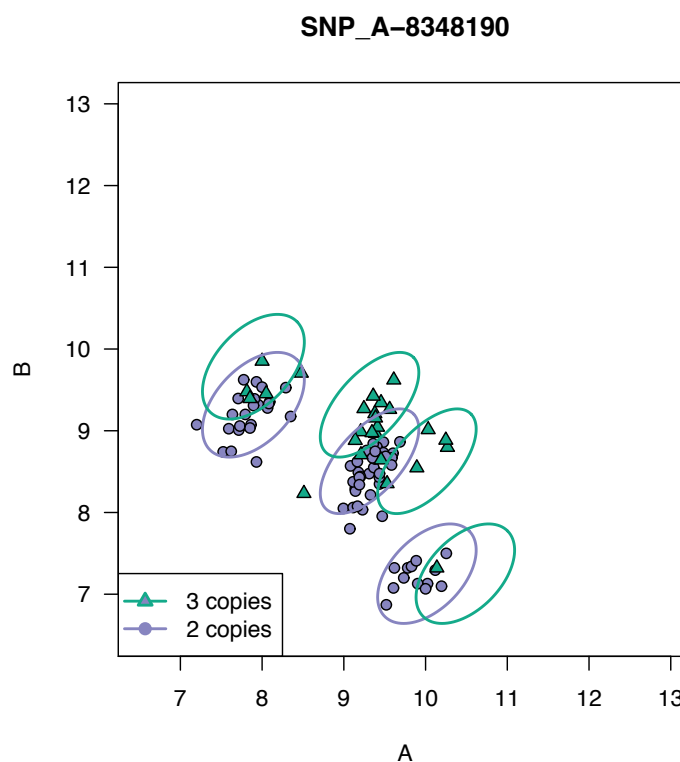
Uniparental isodisomy



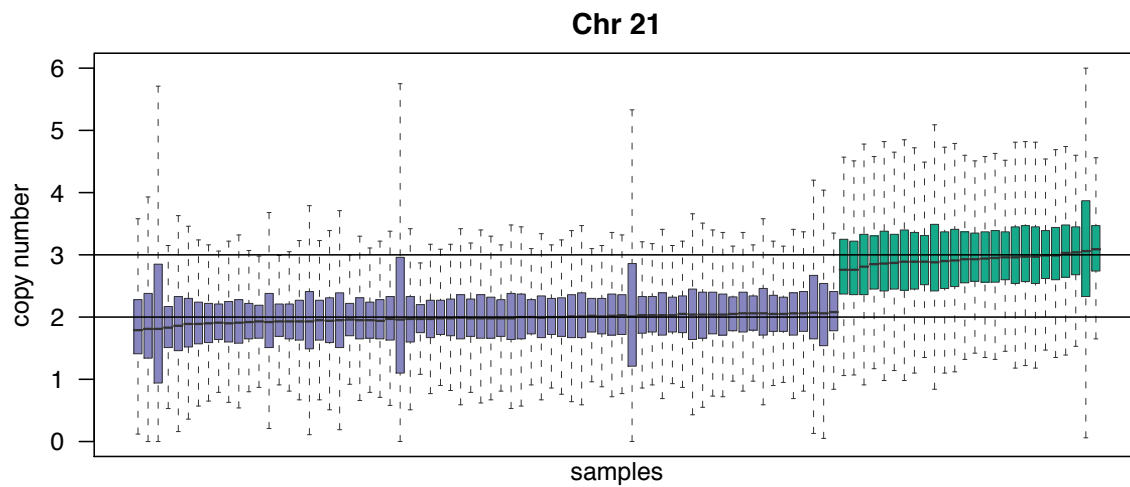
A versus B plots



A versus B plots

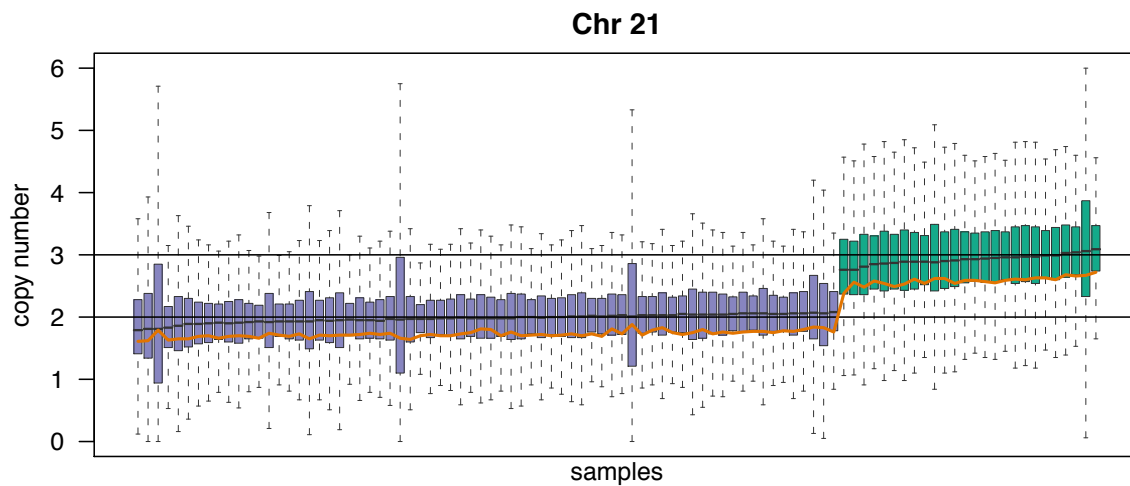


Trisomy 21



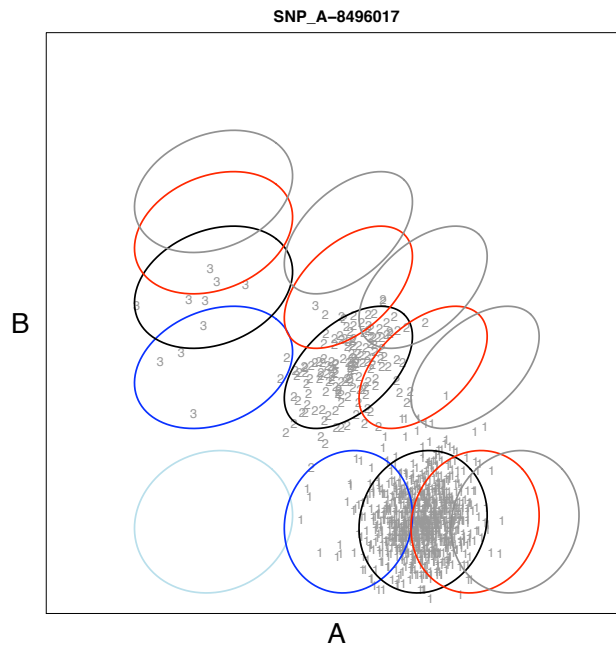
Samples from Aravinda Chakravarti and Betty Doan

Trisomy 21

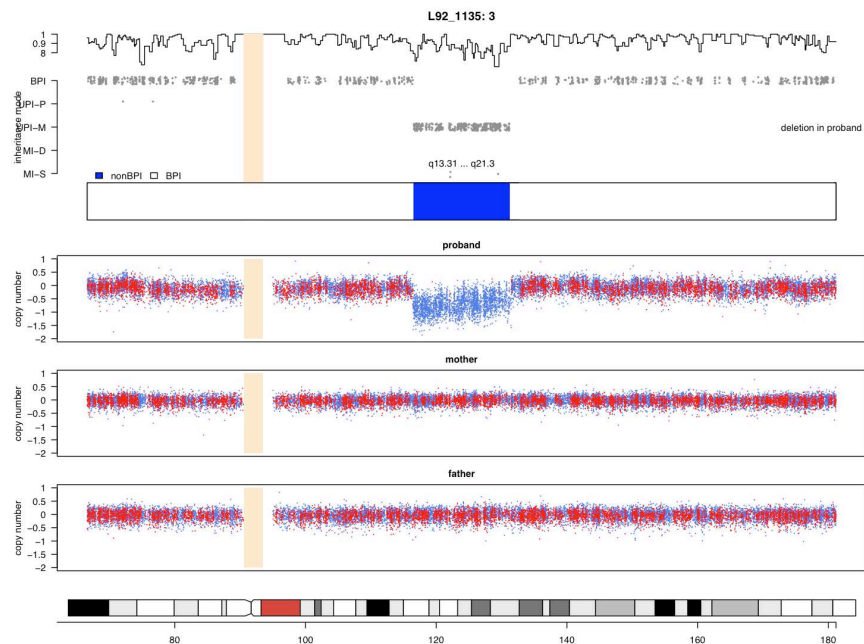


Samples from Aravinda Chakravarti and Betty Doan

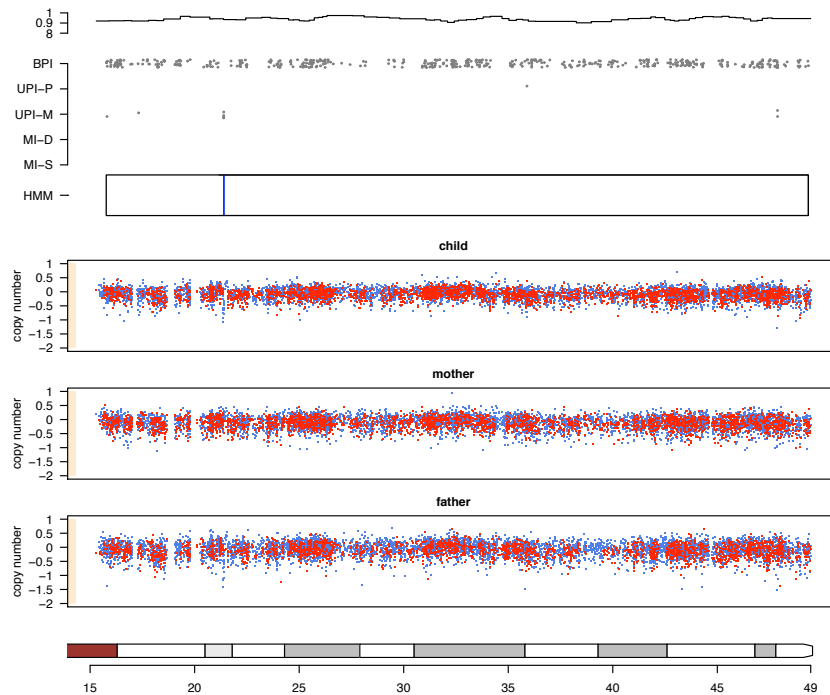
Prediction regions for copy number



De novo deletion



De novo deletion



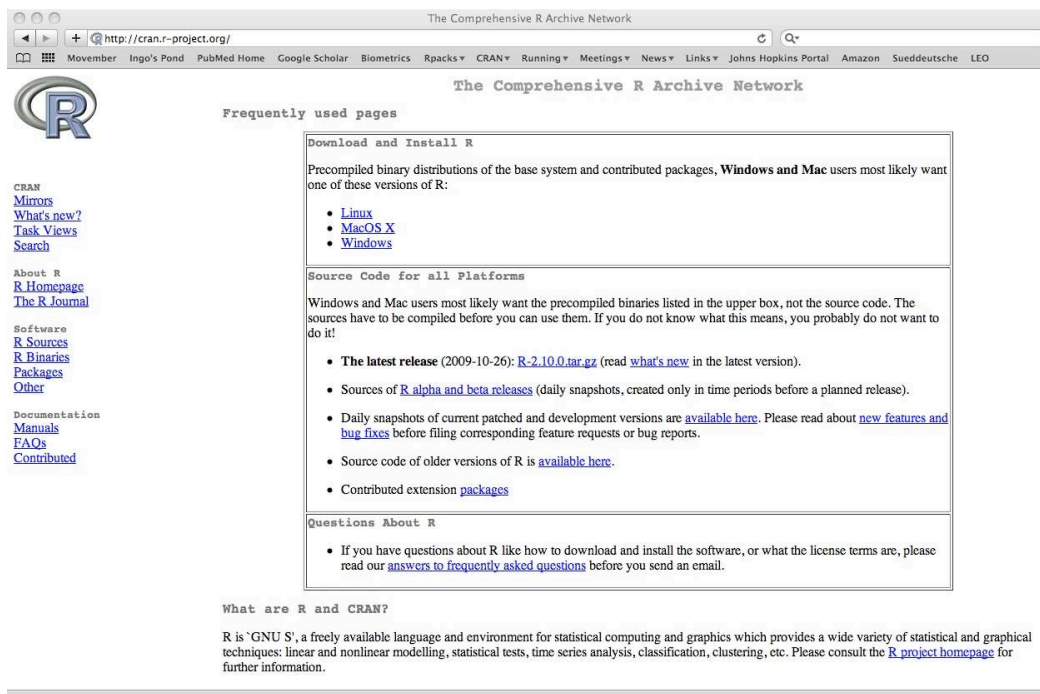
The statistical environment R

- R is an environment for data analysis and visualization.
- R is both open source and open development.
- You can look at the source code and propose changes.
- R is not in the public domain.
- You are given a license to run the software (currently GPL).

The R software

- R is mainly written in C.
- R is available for many platforms:
 - Unix of many flavors, including Linux, Solaris, FreeBSD.
 - Windows 95 and later.
 - MacOS X.
- Binaries and source code are available from `www.r-project.org`.
- R “talks” to data bases, programming languages, and other statistical packages.
- R should be source code compatible with most of the Splus code written.

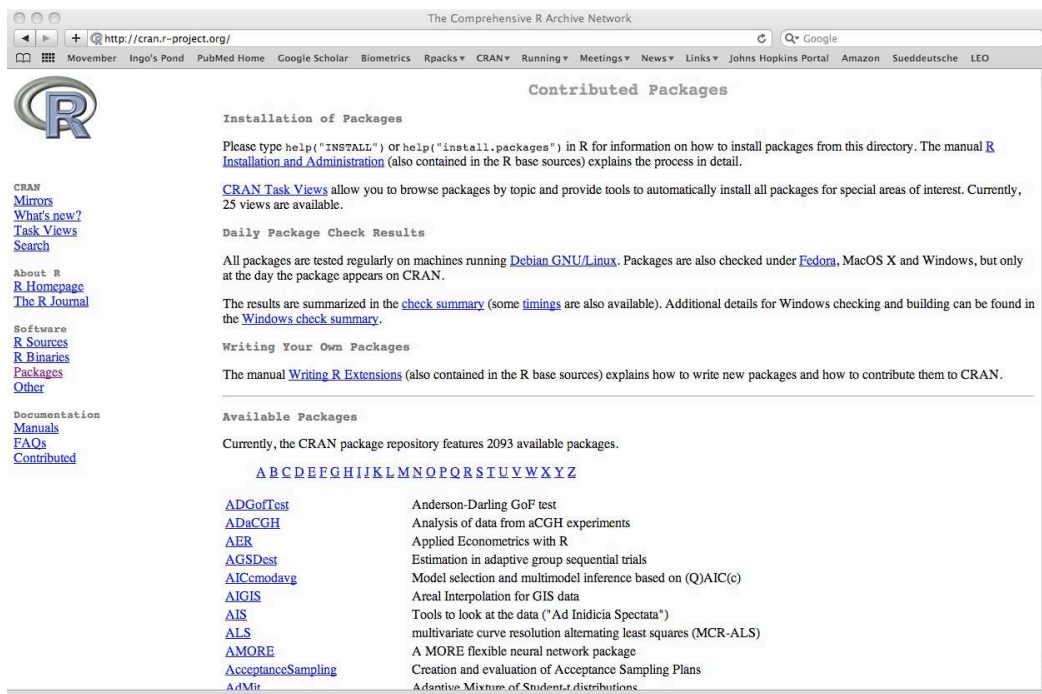
CRAN



The R package system

- Packages are self-contained units of code with documentation.
- The packages are simple to obtain and to understand, and can easily be updated.
- You can write your own packages!
- All functions must have examples to run.
- There are automatic testing features built in.

CRAN packages



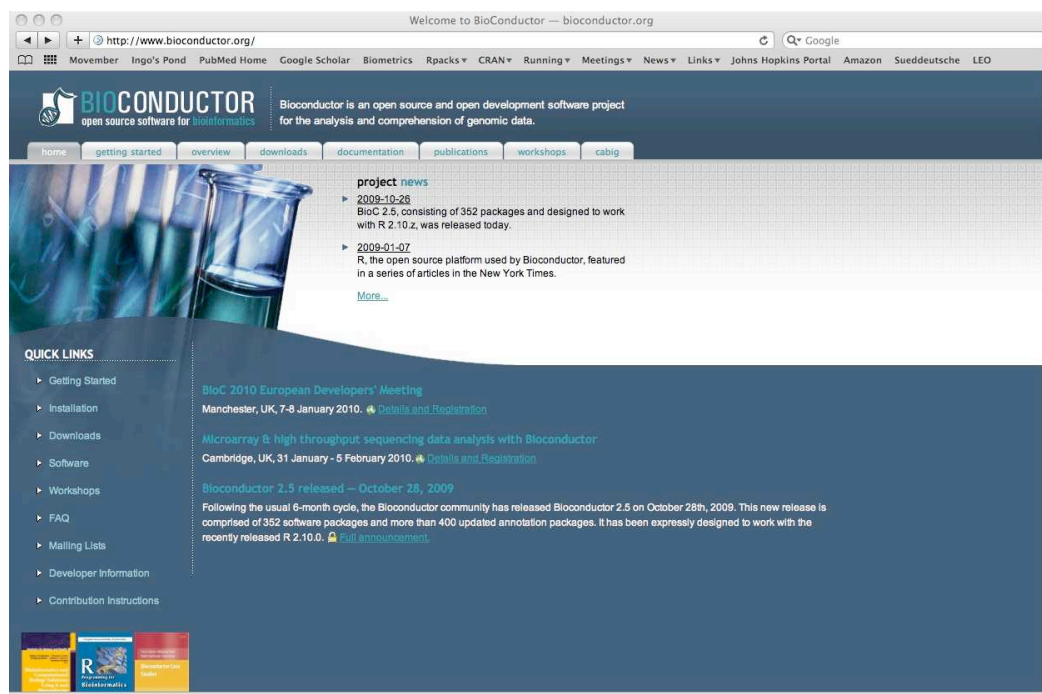
Advantages

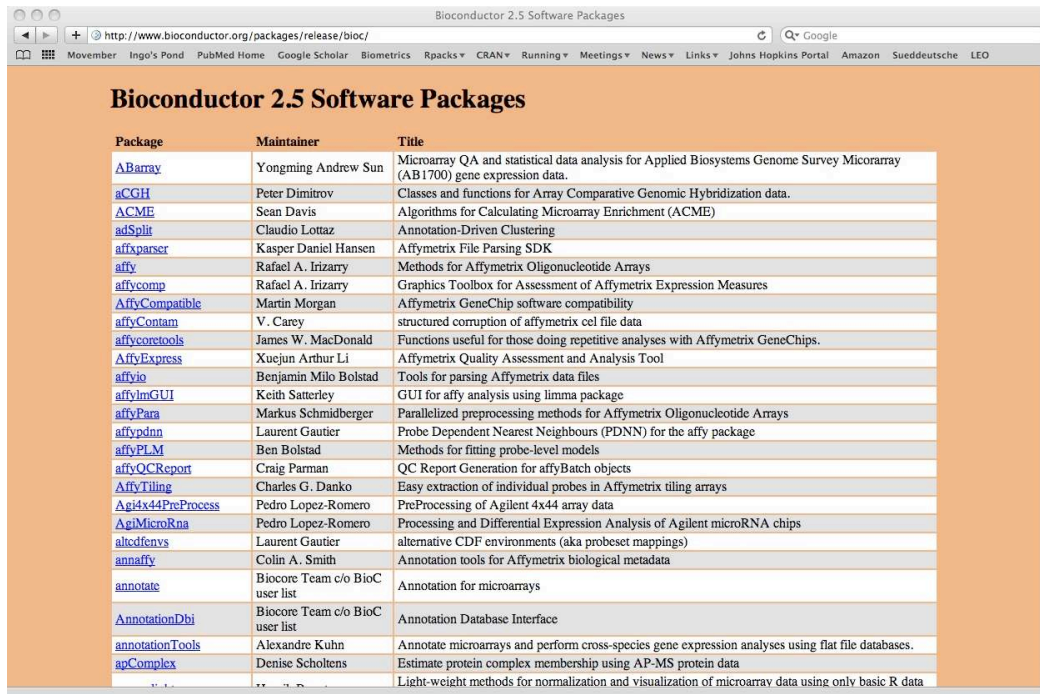
- Free
- Available for all major platforms
- Comprehensive
- Powerful graphics
- Well-designed programming language
- Unlimited extensibility
- Widely used by statisticians
- Increasingly used for genomic analyses (Bioconductor)

Disadvantages

- No dedicated support
- Complex syntax
- Not point-and-click
- Some simple tasks are rather hard

Bioconductor





Bioconductor 2.5 Software Packages

Package	Maintainer	Title
ABarray	Yongming Andrew Sun	Microarray QA and statistical data analysis for Applied Biosystems Genome Survey Micorarray (AB1700) gene expression data.
aCGH	Peter Dimitrov	Classes and functions for Array Comparative Genomic Hybridization data.
ACME	Sean Davis	Algorithms for Calculating Microarray Enrichment (ACME)
adSplit	Claudio Lottaz	Annotation-Driven Clustering
affyparser	Kasper Daniel Hansen	Affymetrix File Parsing SDK
affy	Rafael A. Irizarry	Methods for Affymetrix Oligonucleotide Arrays
affycomp	Rafael A. Irizarry	Graphics Toolbox for Assessment of Affymetrix Expression Measures
AffyCompatible	Martin Morgan	Affymetrix GeneChip software compatibility
affyContam	V. Carey	structured corruption of affymetrix cel file data
affycoretools	James W. MacDonald	Functions useful for those doing repetitive analyses with Affymetrix GeneChips.
AffyExpress	Xuejun Arthur Li	Affymetrix Quality Assessment and Analysis Tool
affyio	Benjamin Milo Bolstad	Tools for parsing Affymetrix data files
affymGUI	Keith Satterley	GUI for affy analysis using limma package
affyPara	Markus Schmidberger	Parallelized preprocessing methods for Affymetrix Oligonucleotide Arrays
affypdnn	Laurent Gautier	Probe Dependent Nearest Neighbours (PDNN) for the affy package
affyPLM	Ben Bolstad	Methods for fitting probe-level models
affyQCReport	Craig Parman	QC Report Generation for affyBatch objects
AffyTiling	Charles G. Danko	Easy extraction of individual probes in Affymetrix tiling arrays
Agi4x44PreProcess	Pedro Lopez-Romero	PreProcessing of Agilent 4x44 array data
AgiMicroRna	Pedro Lopez-Romero	Processing and Differential Expression Analysis of Agilent microRNA chips
alcdenvs	Laurent Gautier	alternative CDF environments (aka probeset mappings)
annaffy	Colin A. Smith	Annotation tools for Affymetrix biological metadata
annotate	Biocore Team c/o BioC user list	Annotation for microarrays
AnnotationDbi	Biocore Team c/o BioC user list	Annotation Database Interface
annotationTools	Alexandre Kuhn	Annotate microarrays and perform cross-species gene expression analyses using flat file databases.
apComplex	Denise Scholtens	Estimate protein complex membership using AP-MS protein data
		Lieght-weight methods for normalization and visualization of microarray data using only basic R data

JHSPH Biostatistics classes

3 **140.615** Biostatistics for Laboratory Scientists I
MWF 10:30 – 11.20 (Ingo Ruczinski)

140.644 Practical Machine Learning
MW 1:30 – 2.50 (Rafael Irizarry)

4 **140.616** Biostatistics for Laboratory Scientists II
MWF 10:30 – 11.20 (Ingo Ruczinski)

140.688 Statistics for Genomics
MW 10:30 – 11.50 (Jeff Leek)

- What is statistics?
 - Data exploration and analysis.
 - Quantification of evidence and uncertainty.
 - Inductive inference with probability.
- What is probability?
 - A branch of mathematics concerning the study of random processes.

Diagnostics

		DISEASE	
		+	−
TEST	+	TP	FP
	−	FN	TN

Diagnostics

		DISEASE	
		+	-
TEST	+	TP	FP
	-	FN	TN

Sensitivity	→	$\Pr(\text{positive test} \mid \text{disease})$
Specificity	→	$\Pr(\text{negative test} \mid \text{no disease})$
Positive Predictive Value	→	$\Pr(\text{disease} \mid \text{positive test})$
Negative Predictive Value	→	$\Pr(\text{no disease} \mid \text{negative test})$
Accuracy	→	$\Pr(\text{correct outcome})$

Diagnostics

Assume that some disease has a 0.1% prevalence in the population. Assume we have a test kit for that disease that works with 99% sensitivity and 99% specificity. What is the probability of a person having the disease, **given the test result is positive**, if we randomly select a subject from the general population?

Diagnostics

		DISEASE	
		+	-
TEST	+	99	999
	-	1	98901

Diagnostics

		DISEASE	
		+	-
TEST	+	99	999
	-	1	98901

Sensitivity	→	$99 / (99+1) = 99\%$
Specificity	→	$98901 / (999+98901) = 99\%$
Positive Predictive Value	→	$99 / (99+999) \approx 9\%$
Negative Predictive Value	→	$98901 / (1+98901) > 99.9\%$
Accuracy	→	$(99+98901) / 100000 = 99\%$

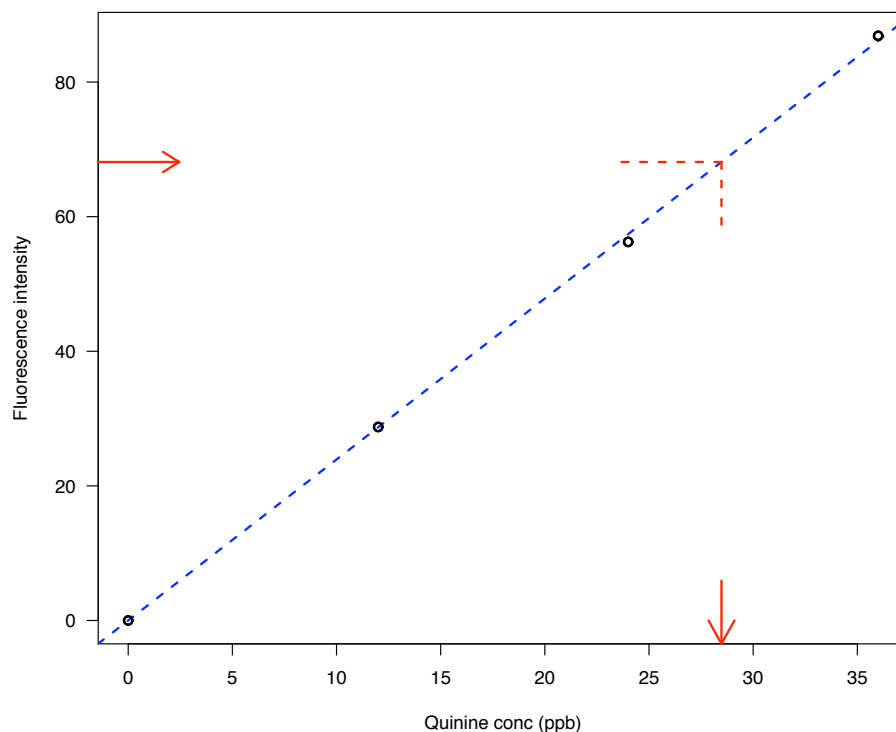
Calibration

Goal: Determine, by fluorescence, the concentration of quinine in a sample of tonic water.

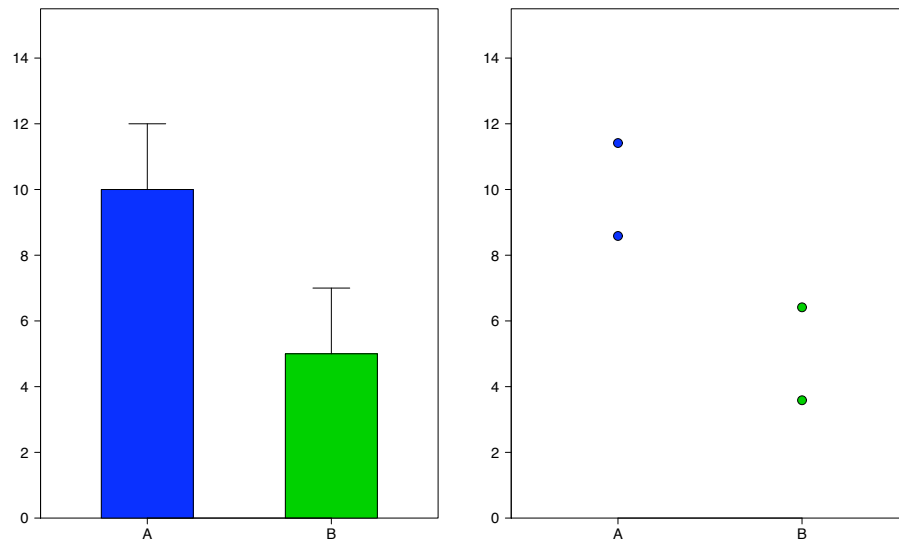
- 1 Obtain a stock solution with known concentration of quinine.
- 2 Create several dilutions of the stock.
- 3 Measure fluorescence intensity of each such standard.
- 4 Measure fluorescence intensity of the unknown.
- 5 Fit a line to the results for the standards.
- 6 Use line to estimate quinine concentration in the unknown.

Question: How precise is the resulting estimate?

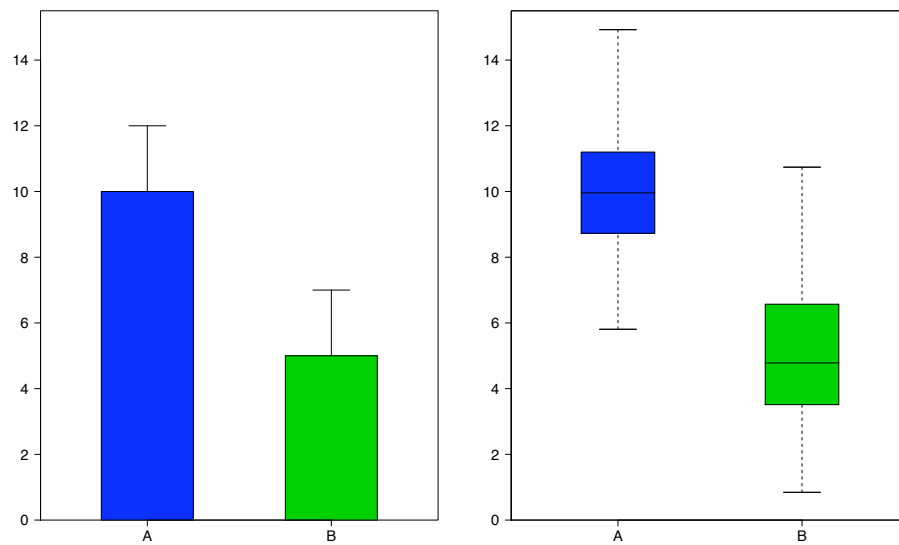
Calibration



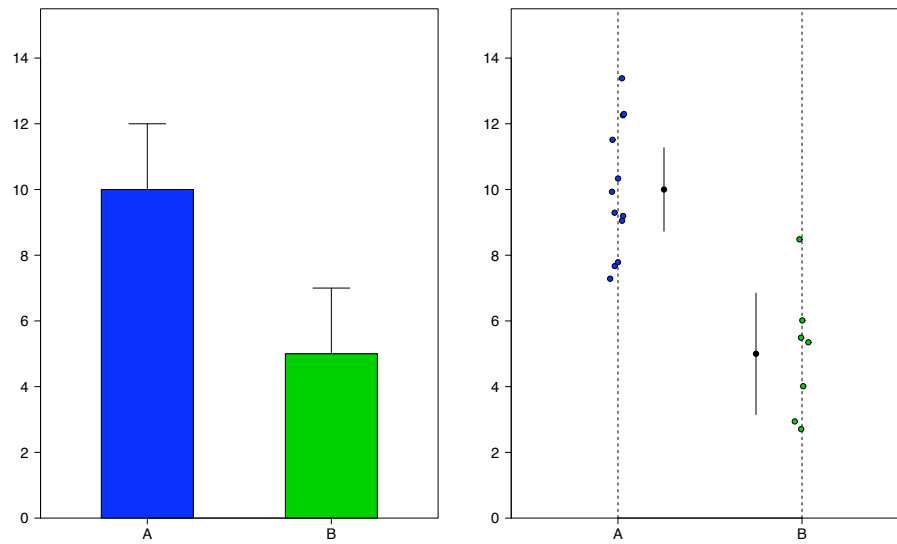
Summarizing data



Summarizing data



Summarizing data



<http://biostat.jhsph.edu/~iruczins/>
ingo@jhu.edu