R on the Grid

BeSTGRID Technical Working Group

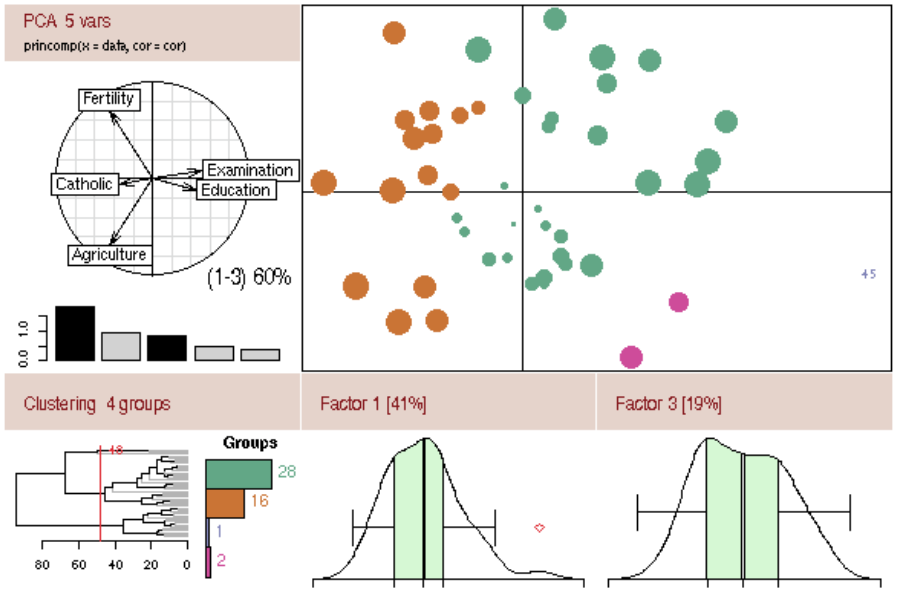
Mik Black
Department of Biochemistry
University of Otago

What is R?

Introduction to R

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

The R Project for Statistical Computing



http://www.r-project.org/

Data Analysts Captivated by R's Power

Stuart Isett for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

By ASHLEE VANCE
Published: January 6, 2009

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

http://www.nytimes.com

UNIVERSITY of OTAGO

Te Whare Wānanga o Otāgo

January 8, 2009, 1:52 PM

R You Ready for R?

By ASHLEE VANCE

Statistics professor Robert Gentleman who helped developed the R programming language. (Credit: Stuart Isett for The New York Times)

There seems to be a cathaRsis taking place.

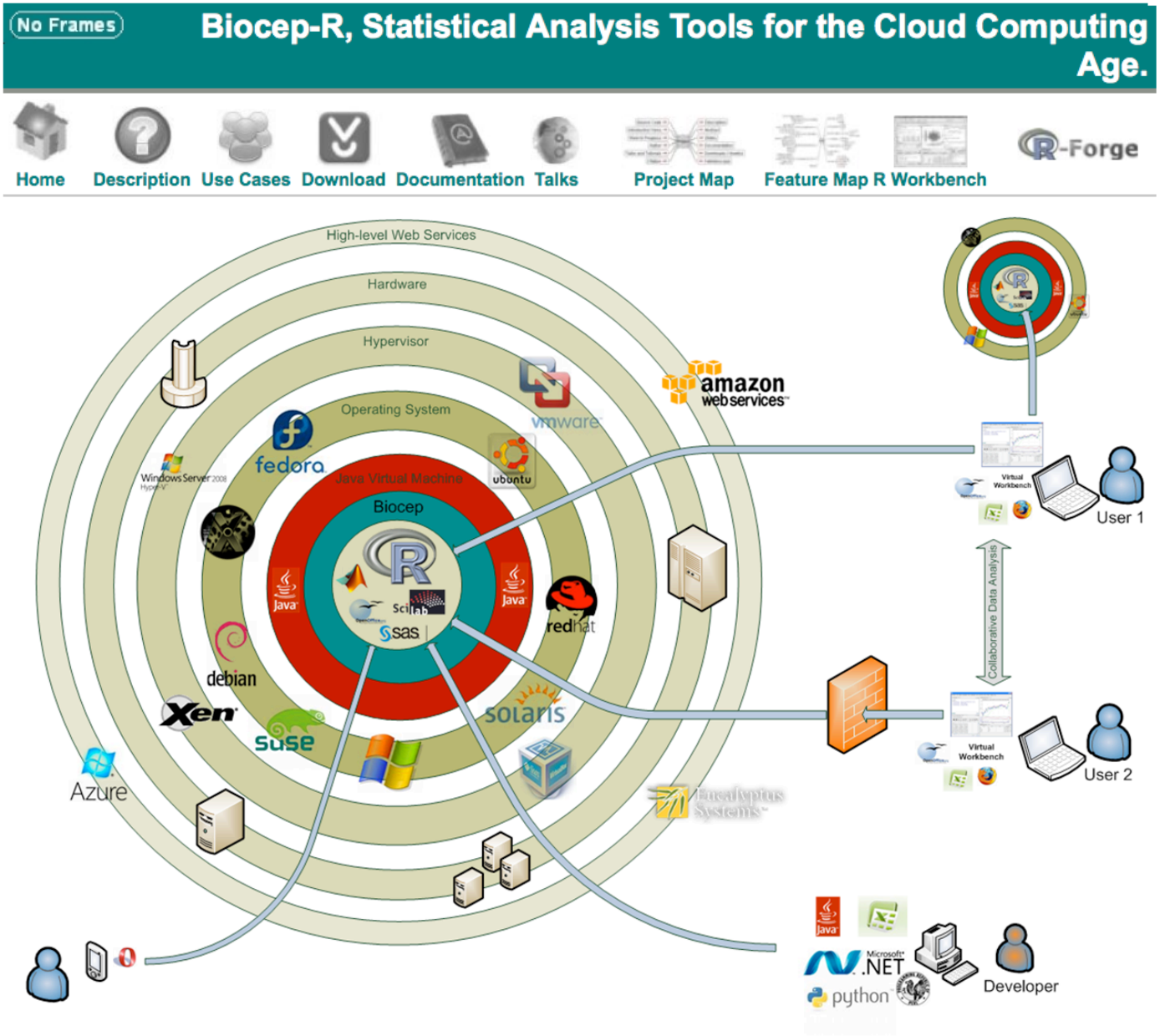
My story published Tuesday on the R programming language has generated a flood of reader e-mail messages. The story covers the software's broad usage and vibrant developer community in detail, but, in short, R helps people deal with large volumes of data in a wide variety of industries, including pharmaceuticals, finance and oil and gas.

ENABLED BY KAREN

http://www.nytimes.com

R for genomics - *Bioconductor*

http://www.bioconductor.org



http://biocep-distrib.r-forge.r-project.org/

No Frames

Biocep-R, Statistical Analysis Tools for the Cloud Computing Age.

Home | Description | Use Cases | Download | Documentation | Talks | Project Map | Feature Map | R Workbench | R-Forge

Pluggability, reusability

Biocep is a general unified open source Java solution for integrating and virtualizing the access to R engines/servers. It aims to become a federative user-friendly computational e-platform for research, finance and education. The Biocep virtual workbench provides a framework enabling the connection of all the elements of a computational environment:

- 1. The computational resource (whether it is a local machine, a cluster, a grid or a cloud server) via a simple URL.
- 2. The computational components via the import of R packages.
- 3. The GUIs via the import of plugins from repositories or the design of new views with a drag-and-drop GUI editor.

http://biocep-distrib.r-forge.r-project.org/

Cancer informatics: NCI & caBIG

- In the US the National Cancer Institute (NCI) has funded the Cancer Biomedical Informatics Grid (caBIG).
  - Broad goal of providing a "truly collaborative information network".
  - Grid-based tools for storage, sharing coordination and analysis of many types of biomedical research data.

UNIVERSITY of OTAGO

Te Whare Wānanga o Otāgo

http://www.broad.mit.edu/cancer/software/genepattern/

GenePattern

Home | Features | Analyses | Download | Documentation | Resources | About | Contact

Search

BROAD INSTITUTE

About

GenePattern combines a powerful **scientific workflow platform** with more than 90 **genomic analysis tools**. An intuitive web interface makes it **easy to use**.

- Case Studies
- Newsletters
- Related Publications
- Collaborations
- Public Datasets

Current version

Release: **3.1.1**
Release date: 7/21/2008

- Hardware requirements
- Supported operating systems

Funding

Funded by the National Cancer Institute, National Institute of Health, and NIGMS.

Download | Online | Quickstart

Analyses

- **Gene Expression Analysis**: Standard and novel clustering, prediction, and marker selection methodologies.
- **Proteomics**: Peak detection, noise subtraction, peak matching, and more for advanced analysis of MALDI, SELDI, and LC-MS data.
- **SNP Analysis**: Analyze SNP microarrays using normalization, copy number estimation, smoothing, LOH determination, and visualization.
- **Data Format Conversion**: Import and export data, normalize and filter data, convert gene identifiers, and more.

Features

- **Public Server**: Run analyses using the GenePattern public server.
- **Analysis Pipelines**: Create analysis workflows by chaining tasks together.
- **Reproducible Research**: Ensure that all versions of an analysis are available and its results are reproducible.
- **Programming Environment**: Call any GenePattern module from Java, MATLAB, or R.

Awards

Bio-IT World Best Practices AWARD

GenePattern is a winner of the Editor's Choice award for the 2005 BioIT World Best Practices competition.

Citation

- **Cite Genepattern**: Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP (2006) GenePattern 2.0 Nature Genetics 38 no. 5 (2006): pp500-501 doi:10.1038/ng0506-500.

News

8/11/2008: The GenePattern Team is awarded founding membership in the caBIG™ **Molecular Analysis Tools Knowledge Center**. The Knowledge Center promotes the adoption of Cancer Biomedical Informatics Grid (caBIG™) technologies aiming to facilitate the discovery of the next generation of cancer diagnostics and therapeutics.

- Press Releases

Reich et al. (2006) GenePattern 2.0., *Nature Genetics*, 38, 500-501.

ENABLED BY KAREN

UNIVERSITY of OTAGO

Te Whare Wānanga o Otāgo

ENABLED BY KAREN

UNIVERSITY of OTAGO

Te Whare Wānanga o Otāgo

Home | News | Contact | caArray | GenePattern | Galaxy | Community

Integrated Genomics
For Health & Disease.

http://bioanalysis.otago.ac.nz

caArray
NZ Array Data Management System

GenePattern
GenePattern

Galaxy
mRNA Analysis Otago

Genomic Community

Transterm
A database of mRNA sequences and elements

Hepatitis B Virus
Hepatitis B Virus Regulatory Sequence Database

Disease Associated 3' UTR variants

Integrated Genomics Resources for Health and Disease

We provide tools to securely manage, analyse and visualise microarray data within New Zealand. Our service allows easy access to bioinformatic tools and the means for you to collaborate with other researchers. See the proposal abstract for more information.

caArray

caArray is available as a New Zealand based data management system for microarray data. This is an installation of software originally developed at the National Cancer Institute as part of the caBig project.

- Securely store your array data in New Zealand.
- Collaborate with other researchers and bioinformaticians.
- Ensure important experimental data are maintained.
- Process data from a provider such as the Otago Genomics Facility.

GenePattern

The NZ installation of GenePattern, provides local access for analysing and visualising expression data. GenePattern is a widely implemented, used and cited platform from the Broad Institute (MIT) with over 4000 registered users worldwide.

It combines a powerful scientific workflow platform with more than 90 computational and visualization tools for the analysis of genomic data.

The focus of this platform is on

- Gene Expression Analysis: Standard and novel clustering, prediction, and marker selection methodologies.
- SNP Analysis: Analyze SNP microarrays using normalization, copy number estimation, smoothing, LOH determination, and visualization.

It also has powerful features for Proteomics and Data Format Conversion

See the GenePattern project homepage for more details.

Reference
Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP (2006) GenePattern 2.0 Nature Genetics 38 no. 5 (2006): pp500-501 doi:10.1038/ng0506-500.

Galaxy : mRNA Analysis Otago

The NZ installation of Galaxy is a tool for the analysis and visualisation of sequence data.

The Galaxy tool provides easy access to UCSC and other data sources. Queries of these sources may be coherently combined. The results may then be analysed right there using popular tools such as Emboss.

Genomics Community

A wiki site is available to facilitate the interchange of ideas within the Genomics Community.

TelstraClear

UNIVERSITY of OTAGO

Te Whare Wānanga o Otāgo

Home | News | Contact | caArray | GenePattern | Galaxy | Community

Integrated Genomics
For Health & Disease.

http://bioanalysis.otago.ac.nz

caArray
NZ Array Data Management System

GenePattern
GenePattern

Galaxy
mRNA Analysis Otago

Genomic Community

Transterm
A database of mRNA sequences and elements

Hepatitis B Virus
Hepatitis B Virus Regulatory Sequence Database

Disease Associated 3' UTR variants

Integrated Genomics Resources for Health and Disease

We provide tools to securely manage, analyse and visualise microarray data within New Zealand. Our service allows easy access to bioinformatic tools and the means for you to collaborate with other researchers. See the proposal abstract for more information.

caArray

caArray is available as a New Zealand based data management system for microarray data. This is an installation of software originally developed at the National Cancer Institute as part of the caBig project.

- Securely store your array data in New Zealand.
- Collaborate with other researchers and bioinformaticians.
- Ensure important experimental data are maintained.
- Process data from a provider such as the Otago Genomics Facility.

GenePattern

The NZ installation of GenePattern, provides local access for analysing and visualising expression data. GenePattern is a widely implemented, used and cited platform from the Broad Institute (MIT) with over 4000 registered users worldwide.

It combines a powerful scientific workflow platform with more than 90 computational and visualization tools for the analysis of genomic data.

The focus of this platform is on

- Gene Expression Analysis: Standard and novel clustering, prediction, and marker selection methodologies.
- SNP Analysis: Analyze SNP microarrays using normalization, copy number estimation, smoothing, LOH determination, and visualization.

It also has powerful features for Proteomics and Data Format Conversion

See the GenePattern project homepage for more details.

Reference
Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP (2006) GenePattern 2.0 Nature Genetics 38 no. 5 (2006): pp500-501 doi:10.1038/ng0506-500.

Galaxy : mRNA Analysis Otago

The NZ installation of Galaxy is a tool for the analysis and visualisation of sequence data.

The Galaxy tool provides easy access to UCSC and other data sources. Queries of these sources may be coherently combined. The results may then be analysed right there using popular tools such as Emboss.

Genomics Community

A wiki site is available to facilitate the interchange of ideas within the Genomics Community.

TelstraClear

GenePattern
GenePattern

ENABLED BY KAREN

UNIVERSITY of OTAGO

Te Whare Wānanga o Otāgo

http://bioanalysis.otago.ac.nz/gp/pages/index.jsf

GenePattern

My Settings | Sign Out Mik Black

Modules & Pipelines | Suites | Job Results | Resources | Downloads | Help

Modules & Pipelines

category suite all

open all | close all

- **Recently Used**
  - ComparativeMarkerSelectionViewer
  - Rtest2.5
  - SVGtest
  - SVGtest2.5
- **Annotation**
  - GeneCruiser
- **Clustering**
  - ConsensusClustering
  - HierarchicalClustering
  - HierarchicalClusteringGrid
  - KMeansClustering
  - NMFConsensus
  - SOMClustering
  - SubMap
- **Gene List Selection**
  - ClassNeighbors
  - ComparativeMarkerSelection
  - ExtractComparativeMarkerResults
  - GeneNeighbors
  - GSEA
  - SelectFeaturesColumns
  - SelectFeaturesRows
- **Gene Set Analysis**
  - PCOT2
- **Image Creators**
  - HeatMapImage
  - HierarchicalClusteringImage
- **MikModules**
  - Gene.Boxplot
  - InstallBiocPkg
  - IntrinsicSubtypeSSP
  - Limma
  - limma.prealpha
  - PAMR
  - Rtest
  - Rtest2.5
  - snowtest
  - SVGtest
  - SVGtest2.5
- **Missing Value Imputation**
  - ImputeMissingValues.KNN

Welcome to GenePattern

Analyzing genomic data in GenePattern

comments/suggestions

To run an analysis: Select a module from the list at the left, enter values for the parameters and click *Run*. When the analysis completes, the result files are displayed in the Recent Jobs pane at the right.

To view analysis results: Click the arrow icon next to a result file. A menu lists the modules most often used to view or analyze this result. Select a module, enter values for the parameters and click *Run*.

what do you want to do?

Protocols for running common analyses in GenePattern:

Differential Expression Analysis
Find genes that are significantly differentially expressed between classes of samples.

Clustering
Group genes and/or samples by similar expression profiles.

Prediction
Create a model, also referred to as a classifier or class predictor, that correctly classifies unlabeled samples into known classes.

SNP Copy Number and Loss of Heterozygosity Estimation
Compute SNP copy number (CN) and loss of heterozygosity (LOH) based on Affymetrix SNP chip data for paired target/normal samples.

Recent Jobs

3414 SVGtest2.5
May 26 10:45:48 AM
stdout.txt
stderr.txt

3413 Rtest2.5
May 26 10:44:15 AM
Rtest-output.txt
stdout.txt
stderr.txt

3412 SVGtest
May 26 10:42:14 AM
SVGtest-output.svg
stdout.txt
stderr.txt

3410 Limma
May 26 10:31:14 AM
breast-cohort1-mas5-noqnorm-test.online-ERneg_vs_ERpos.odf
breast-cohort1-mas5-noqnorm-test.online_volcanoplot-ERneg_vs_ERpos.svg
breast-cohort1-mas5-noqnorm-test.online-ERpos.odf
stdout.txt
stderr.txt

ENABLED BY KAREN

UNIVERSITY of OTAGO

Te Whare Wānanga o Otāgo

Show parameter descriptions

HierarchicalClusteringGrid version 38

* required field

Run | Reset | properties | export | edit | help

input filename*	Browse... Specify URL / Upload File input data file name - .gct, .res, .odf type = Dataset
grid myproxy username*	MyProxy username
grid myproxy password*	MyProxy password
grid submit location*	Otago Maggie Submit location
grid debug	Quiet Turn debugging on
grid java_flags	2G Java options
grid walltime*	86400 Wall clock time limit (in seconds)
grid cpus*	Serial task - can only use 1 CPU Number of CPUs to use
grid vo*	BeSTGRID Virtual Organization (run job as)

THANKS VLAD!

Run | Reset | properties | export | edit | help

ENABLED BY KAREN

UNIVERSITY of OTAGO

Te Whare Wānanga o Otāgo

Journal of Statistical Software

August 2009, Volume 31, Issue 1. *http://www.jstatsoft.org/*

State of the Art in Parallel Computing with R

Markus Schmidberger
Ludwig-Maximilians-Universität München

Martin Morgan
Fred Hutchinson Cancer Research Center

Dirk Eddelbuettel
Debian Project

Hao Yu
University of Western Ontario

Luke Tierney
University of Iowa

Ulrich Mansmann
Ludwig-Maximilians-Universität München

ENABLED BY KAREN

Cluster computing with R

- Technologies: MPI, PVM, NWS, sockets
- R packages: Rmpi, rpvm, rnws.
- SNOW (simple network of workstations) package offers cluster computing within R via these interfaces.

Journal of Statistical Software
August 2009, Volume 31, Issue 1. *http://www.jstatsoft.org/*

Grid computing with R

- gridR: "The server side implementation of gridR uses several external software components: Globus Toolkit 4 grid middleware, … a GRMS-Server installation from the Gridge toolkit"
- multiR: "If you wish to use multiR, please email us and we will talk to you about requirements. multiR is free to use but does take some configuring." http://www.ncess.ac.uk/tools/multir/
- Biocep-R: "Java solution for integrating and virtualizing the access to servers with R"

Journal of Statistical Software

August 2009, Volume 31, Issue 1. *http://www.jstatsoft.org/*

SNOW example

```
library(Rmpi)
library(snow)
cl.print<-function(i,x) print(x[[i]])
aa<-list(a="apple",b="banana")
cl <- makeCluster(spec=2)
do.call("rbind", clusterCall(cl,
  function(cl) Sys.info()["nodename"]))
clusterEvalQ(cl, sessionInfo())
bb<-clusterApplyLB(cl,2:1,cl.print,aa)
print(bb)
stopCluster(cl)
```

Using parallel R on the Grid

- Submission of multiple R jobs to the grid to run independently in parallel (e.g., OGRE).
  - Embarrassingly parallel tasks
- Access to grid from within R code (e.g., SNOW).
  - Allows mix of parallel and non-parallel code.
- Access to grid via R from application (e.g., GenePattern module utilising SNOW).
  - Provides canned grid access for specific applications
  - Ideal for popular tasks, and for non-expert users.