The R Genetics Project
Bioconductor for Genetics

Gregory Warnes, Scott Chasalow, Giovanni Montana, Michael O'Connell, David Henderson, Nitin Jain, Weiliang Qiu, Junsheng Cheng, Ross Lazarus

Abstract

The R Genetics Project is a collaborative effort to develop a complete set of tools for storing, accessing, manipulating, and analyzing genetics data, from small candidate gene studies consisting of a few genetic markers to large whole genome studies containing hundreds of thousands of markers. The initial goal is to provide a foundation of efficient data structures and easy to use manipulation functions. We intend this foundation to allow methods developers to quickly and easily develop packages implementing their own techniques, while maintaining interoperability. This will reduce the burden on both method developers and applied data analysis, who must currently move data between numerous packages and data formats.

The foundation R Genetics packages, GeneticsBase, has reached sufficient maturity for introduction to the R community. This talk will describe the R Genetics project, provide an outline of the data structures and features within GeneticsBase We will give a brief demo of some of these features, as well as mentioning several additional packages which are building upon this common base, including FBAT, GeneticsDesign and GeneticsPed.

Current Support for Genetics in R

- Multiple packages
- Incompatible data structures and file formats
- Designed for small genetics data sets (< 100 markers)
- Overlapping functionality
- No collaboration among package developers

The R Genetics Project

- *Goals:*
  - Common Data Structures, and Tools for
    - Management
    - Visualization
    - Reporting
    - Annotation
    - Exchange
  - Efficient support *LARGE* data sets (whole Genome SNP studies 100K-500K markers)
  - Logical organization of functionality among packages
  - → Encourage Collaboration among methods developers
  - → Rich set of tools in a single common environment

People

- Scott D Chasalow, Bristol-Myers Squibb
- Junsheng Cheng, Univ. Chicago
- Nitin Jain. Smith Hanley (Pfizer)
- David Henderson, Insightful
- Nicolaus Lewin Koh, Lilly
- Ross Lazarus, Channing Laboratory (NIH RO1)
- Giovanni Montana, Bristol-Myers Squibb
- Michael O'Connell, Insightful
- Gregory Warnes, University of Rochester (NIH RO1)

Component Packages:

- GeneticsBase: foundation package
  - Efficient data structures modeled on BioBase's eSet
  - Manipulation functions ([, [[, [<-, etc)
  - I/O for standard file formats
  - Basic summaries (Gene & Marker frequency tables, LD, HWE, …)
- GeneticsDesign: power and sample-size
- GeneticsPed: pedigree data
  - Data structures and functions
  - genetic relationship measures
  - relationship and inbreeding coefficients
- fbat: "Family Based Association Studies" and related pedigree-based methods

More Information

- R Genetics Information Web Site
  - http://r-genetics.org
- Sourceforge Project Site (source code!)
  - http://www.sourceforge.net/projects/r-genetics
- R Genetics Discussion Group
  - R-genetics-talk@lists.sourceforge.net
- Contact Me
  - greg@warnes.net