Introduction to R/Bioconductor

Dr. Mikhail Dozmorov and Cory Giles
Oklahoma Medical Research Foundation
Oklahoma City, OK, USA

Welcome!

- Day 1:
  - 1:30-3:30 | **Introduction to R/BioC (CG)**
  - 4:00-5:00 | **Microarray data retrieval and preprocessing (MD)**
- Day 2:
  - 11:00-12:30 | **Differential expression (MD)**
  - 1:30-3:30 | **Prediction analysis (MD)**
  - 4:00-5:00 | **Annotation & enrichment (CG)**

Welcome!

- Day 3:
  - 9:30-10:30 | **Clustering and Visualization (MD)**
  - 11:00-12:30 | **Network analysis (CG)**
  - 1:30-3:30 | **Array CGH analysis & intermediate R (CG)**

Preliminaries

- Install R
- Install RStudio or other IDE (optional)
- Download course code
  - http://github.com/gilesc/ci-workshop
- Install necessary Bioconductor packages
- Brief RStudio tour

What is R?

- A programming language
- A data analysis and numerical computing environment
- A platform for new statistical techniques
- Widely used for microarray and sequencing analysis
- A FOSS language clone of commercial S (GPL)

Brief History of R

- Heavily inspired by Abelson & Sussman's “Structure and Interpretation of Computer Programs” (a seminal Lisp book)
- Lisp influences melded w/ commercial S
- First release 1993, GPLed 1995
- Gained popularity among bioinformaticians c. 2000-2003 for microarray analysis
  - Rise of Bioconductor around this time
- Today, used for wide variety of bioinformatics tasks, usually on the analysis side (not strong in, e.g., web or end user applications)

R vs. other programming languages

- **Algol family syntax** (C/C++/Java)
- **Vectorized computation** (Matlab)
- **Dynamically/weakly typed** (Python/Perl/Ruby)
- **Strong library system** (Python/Perl)
- **Excellent visualization and symbolic mathematics tools** (Mathematica)
- **Functional style** (Lisp, OCaml, Haskell)
- **Interpreted, with powerful REPL**: (Lisp, Python)
- Weaknesses: slow (for non-numeric tasks), idiosyncratic. Slowness often overcome with C or Fortran FFI.

What is Bioconductor?

- **CRAN**, (similar to and named after Perl's CPAN) is R's repository for generic statistical and numeric programming packages.
- **Bioconductor** is a smaller repository for biology-specific R packages.
- You can download and install (often precompiled) R code for your platform of choice from both of these repositories.

Session 1: A whirlwhind tour of R & Bioconductor

- Data types:
  - Vectors, lists, data frames, tables, factors
- Getting around in the R environment:
  - Help and documentation system
  - Basic functional programming
  - I/O and string manipulation
- Using R repositories
- Basic statistics and Bioconductor ExpressionSet