Bioconductor in Academic and Translational Research

Dr. Martin Morgan, PhD
Roswell Park Cancer Institute

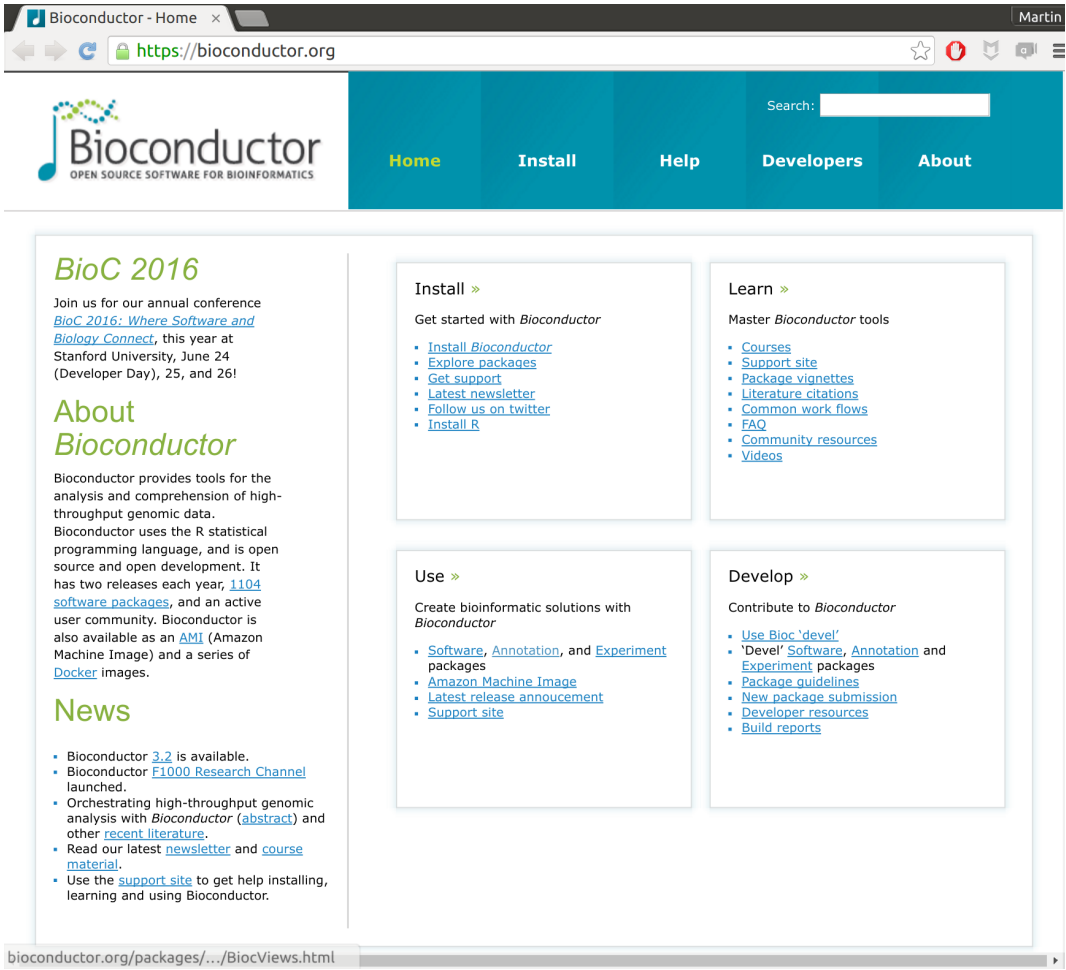
13 March 2016

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

ROSWELL PARK
CANCER INSTITUTE

Bioconductor

- Statistical **analysis** and **comprehension** of high-throughput genomic data
  - Sequencing, microarrays (expression, methylation, copy number, ...), flow cytometry, proteomics, ...
- Open-source, open-development, widely used
  - 1100 'packages' from core team and international contributors
  - 14,000 full-text citations in the literature
  - 250,000 unique IP downloads / year
- Established 2002
  - **Academic** motivation – rigorous statistical analysis of microarray expression data



… the genetics community is fortunately familiar with the ... principles of stewardship of modular software embodied in the *Bioconductor* suite (http://www.bioconductor.org/). The journal has sufficient experience with these resources to **endorse their use** by authors. Nature Genetics 46, 1 (2014) doi:10.1038/ng.2869

File Edit Code View Plots Session Build Debug Tools Help

Go to file/function

Project: (None)

Console | R Markdown

~/a/bioC/Docs/Talks/CI4CC/

```
> dds = DESeq(DESeqDataSet(airway, ~ cell + dex))
estimating size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
> plotMA(dds, alpha=.01)
```

Report.Rmd

Knit HTML | Run | Chunks

```
---
title: "Bioinformatics Report"
author: "Martin Morgan"
date: "March 12, 2016"
output: html_document
---

# Introduction

It is very easy to create advanced reports in a fully reproducible way. Reports can be distributed as PDF, HTML, ...

Reports can include figures, tables, analytic results, interactive applets., etc.

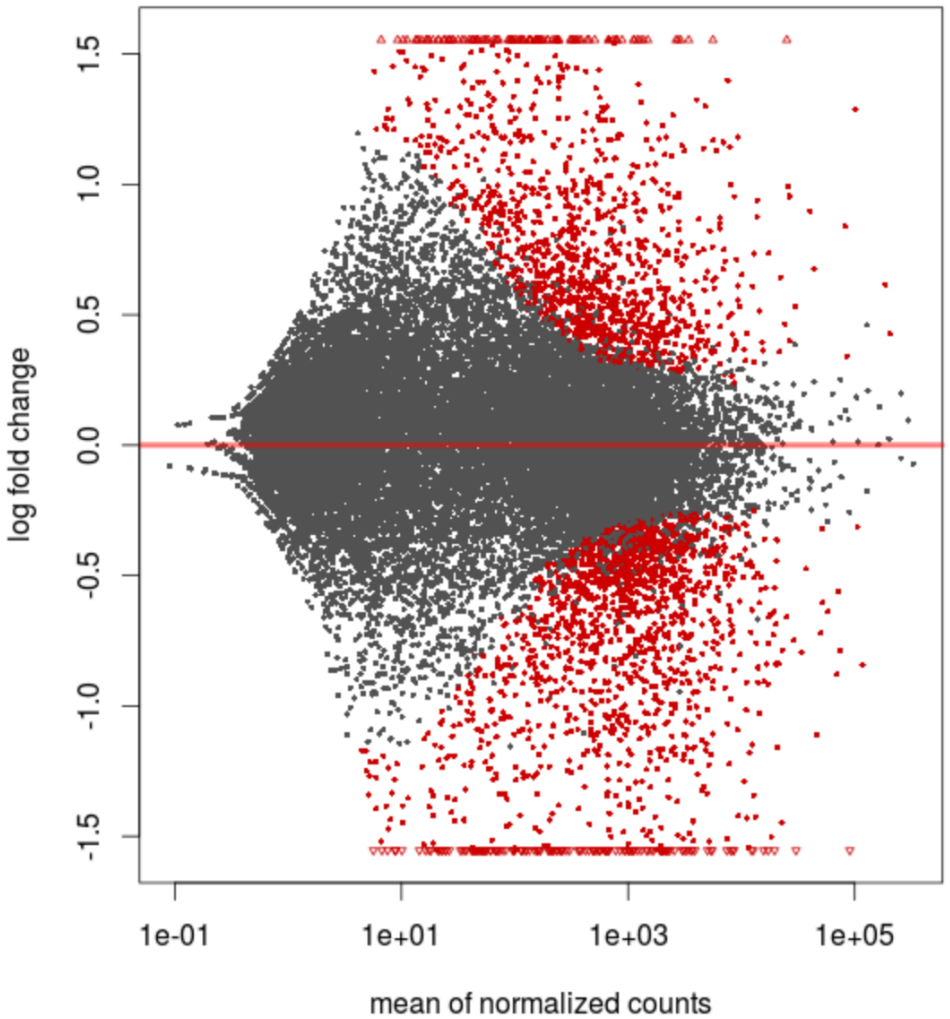
```{r, warning=FALSE, message=FALSE, echo=FALSE}
## R code evaluated when report produced
library(DESeq2)
library(airway)      # example data set
data(airway)
dds = DESeq(DESeqDataSet(airway, ~ cell + dex))
plotMA(dds, alpha=.01)
```
```

17:21 | Chunk 1 | R Markdown

Environment | History

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Bioinformatics Report

file:///home/mtmorgan/a/bioC/Docs/Talks/CI4CC/Report.html

Bioinformatics Report

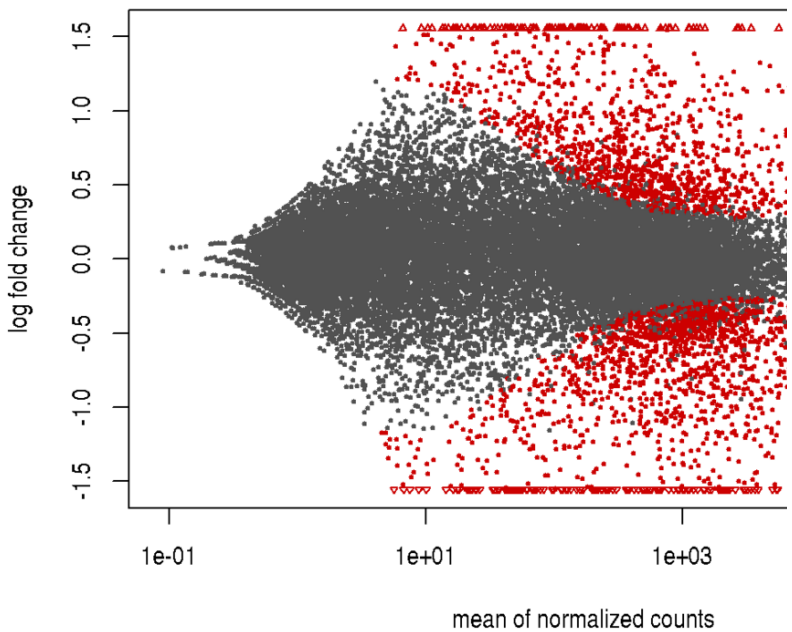
Martin Morgan

March 12, 2016

Introduction

It is very easy to create advanced reports in a fully reproducible way. Reports

Reports can include figures, tables, analytic results, interactive applets., etc.



SEPA: Single-Cell Gene Exp...

127.0.0.1:6302

Search

SEPA: Single-Cell Gene Expression Pattern Analysis

Youtube short video demo

Main Menu

- Reading in Gene Expression Profile
- Analysis for True Experimental Time
- Analysis for Pseudo Temporal Cell Ordering
- Pattern Comparison
- About

Select Step

- Step 1: Input Gene Expression Data
- Step 2: Choose species and identifier
- Step 3: Take logarithm (optional)
- Step 4: Filter genes and cells (optional)

Step 1: Input Gene Expression Data (See Instructions on the right!)

Choose File

Browse... No file selected.

Show instructions

Instructions:

Single cell data should be prepared in a matrix-like data format. Each row corresponds to a gene/feature and each column corresponds to a single cell.

Notice that each row should have the same number of entries, especially for the header (first row, see example below)

Remove all the single or double quote in the expression file

Please make sure the data is correctly read in before any further analysis is conducted. Adjust the options on the left to read in different file formats.

A typical example of tab-delimited file format:

```
Gene    T0_A01  T0_B01  T1_G02  T1_G01
SOX2    0.455   0.543   0.000   2.188
PAT1    0.231   2.792   1.222   0.000
```

Academic development and use

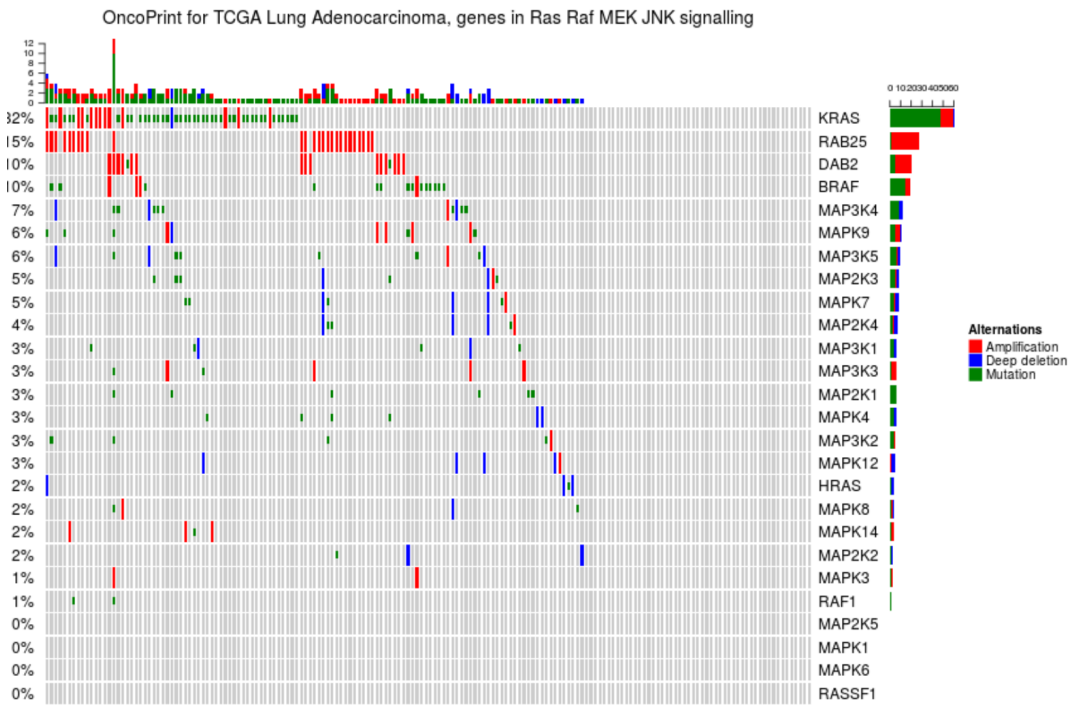
- Contributed packages
  - Often from independently funded projects
  - Reviewed for technical soundness
  - Diversity of quality, motivation
- Best practices
  - Version control
  - Unit tests, coverage statistics
  - Nightly builds
  - 6-month release schedule

- Package discovery
  - Software ontology
  - 'Landing' pages
- Use
  - Vignettes
  - Runnable examples
  - Active support site
  - Courses & conferences

https://bioconductor.org
https://support.bioconductor.org

Translational / Clinical Integration

- Very easy to 'wrap' computations into...
  - Static reports
  - Dynamic applications
- Many packages / groups do this
  - No **standardization** on presentation of results
  - Limited formal **assessment** of utility to translational community
  - Demanding **skill sets** required for successful analysis / development / presentation



ComplexHeatmap::oncoPrint()

Paths to Commercial Participation

- Third-party use / reimplementation
  - **Heterogeneity of licenses**
- SBIR & other funding collaborations
  - **Infrastructure** to benefit the project
- Contracts
  - Strong **influence** on development direction
- Sponsorship, e.g., of our annual conference
  - Fostering **community** and developing **human resources**

Acknowledgments

- Core team (current & recent)
  - Valerie Obenchain, Herve Pages, Dan Tenenbaum, Brian Long, Jim Hester, Jim Java, Sonali Arora, Nate Hayden, Paul Shannon, Marc Carlson
- Technical advisory board
  - Vincent Carey, Wolfgang Huber, Robert Gentleman, Rafael Irizzary, Levi Waldron, Michael Lawrence, Sean Davis
- Scientific advisory board
  - Simon Tavare (CRUK), Paul Flicek (EMBL/EBI), Simon Urbanek (AT&T), Vincent Carey (Brigham & Women's), Wolfgang Huber (EBI), Robert Gentleman (23andMe)

NIH National Human Genome Research Institute
Advancing human health through genomics research

NIH NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research

SOUND

ROSWELL PARK CANCER INSTITUTE

FRED HUTCH
40 YEARS OF CURES 1975–2015

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS