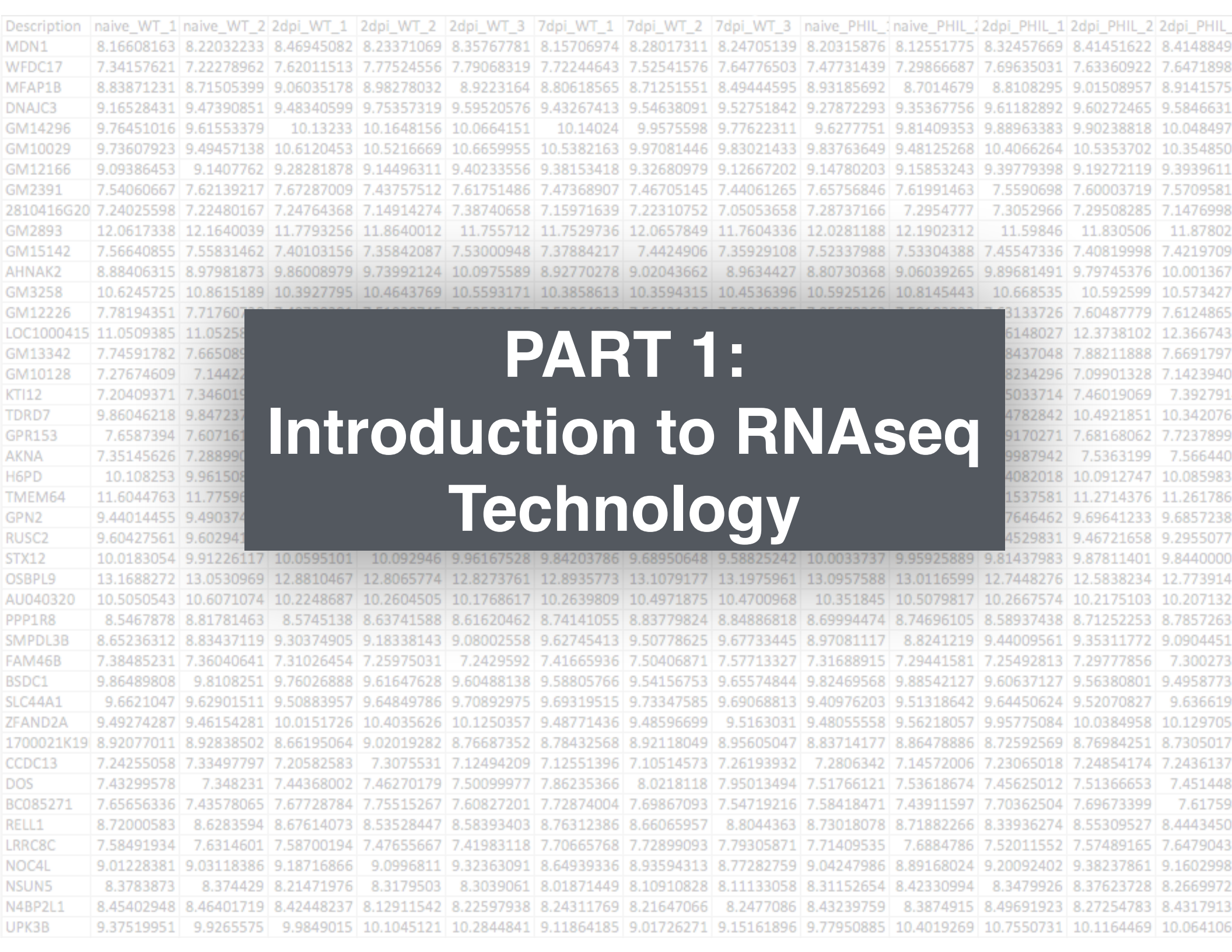

PART 1:
Introduction to RNAseq Technology

Technology is fickle

illumina®

Illumina, Inc.
5200 Illumina Way
San Diego, CA 92122 USA
tel 858.202.4500
fax 858.202.4545
www.illumina.com

24 March 2015

Product Obsolescence Notification

Dear Valued Customer,

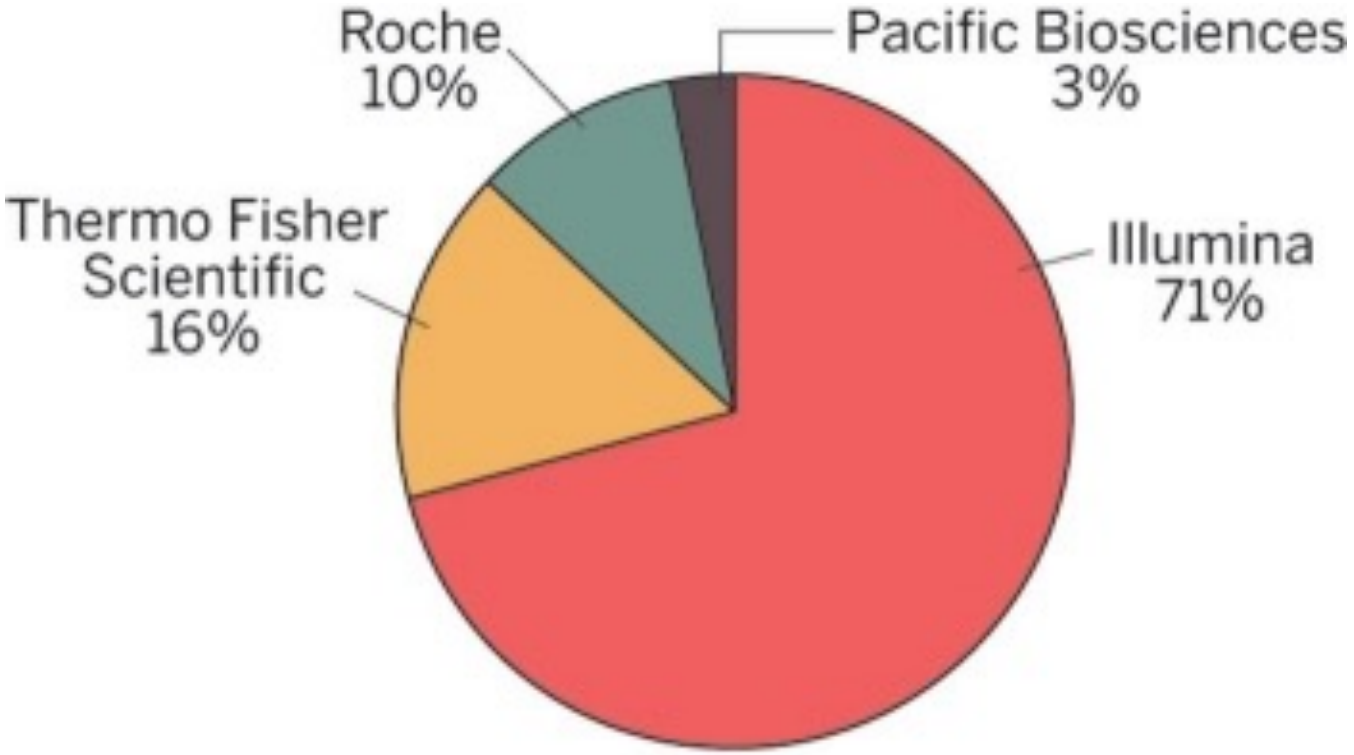
Due to unexpected demand for Illumina's Mouse Gene Expression BeadArray™, the inventory of beads for this product has been completely consumed. We are unable to synthesize exact replacement beadpools due to the nature and complexity of the product. As a result we are forced to discontinue and no longer accept orders for the products listed below.

You are receiving this notice because our records indicate that you have ordered one or more of these products or have orders pending.

Products Affected

Product Description	Catalog Number
MouseRef-8 v2.0 BeadChip (16 Samples)	BD-202-0202
MouseRef-8 v2.0 BeadChip (48 Samples)	BD-202-0602
MouseWG-6 v2.0 BeadChip (12 Samples)	BD-201-0202
MouseWG-6 v2.0 BeadChip (36 Samples)	BD-201-0602

The sequencing landscape is dominated by Illumina



World market in 2013 = $1.3 billion

The year of sequencing

In 2007, the next-generation sequencing technologies have come into their own with an impressive array of successful applications. Kelly Rae Chi reports.

CREATING THE GENOME ANALYZER

When John West started as CEO of Solexa Ltd. in August of 2004, the longest stretch of DNA that the company could sequence was only six bases long. "That was a little bit intimidating," recalls West, now the vice president and general manager of Illumina's DNA sequencing business unit following the acquisition of Solexa Ltd. by Illumina. "The problem was we never had a commercial platform to be able to sequence it on."

The next-generation sequencing platform they developed and commercialized by June 2006, West says, has been a huge success. Since the commercial release of the platform they have sold 100 instruments and increased the scale of what they have tackled using the technology—from a 5,300-base-pair viral genome to a 150-million-base-pair human X chromosome. But the machine was a challenge to develop. The developers had to bring together key elements of chemistry, people and technology to make it work.

By the time Solexa Ltd. announced its plans to merge with Lynx Therapeutics in August 2004, a lot of the core sequencing-by-synthesis chemistry and molecu biology had already been done, West says. The early chemi work, done by Solexa Ltd. in the United Kingdom, led to t creation of a reversible terminator nucleotide and a polym that would incorporate it. Solexa Ltd. and Lynx Therapeut had bought cluster technology for solid phase amplificatic together from the Swiss company Manteia SA, and did som instrumentation design.

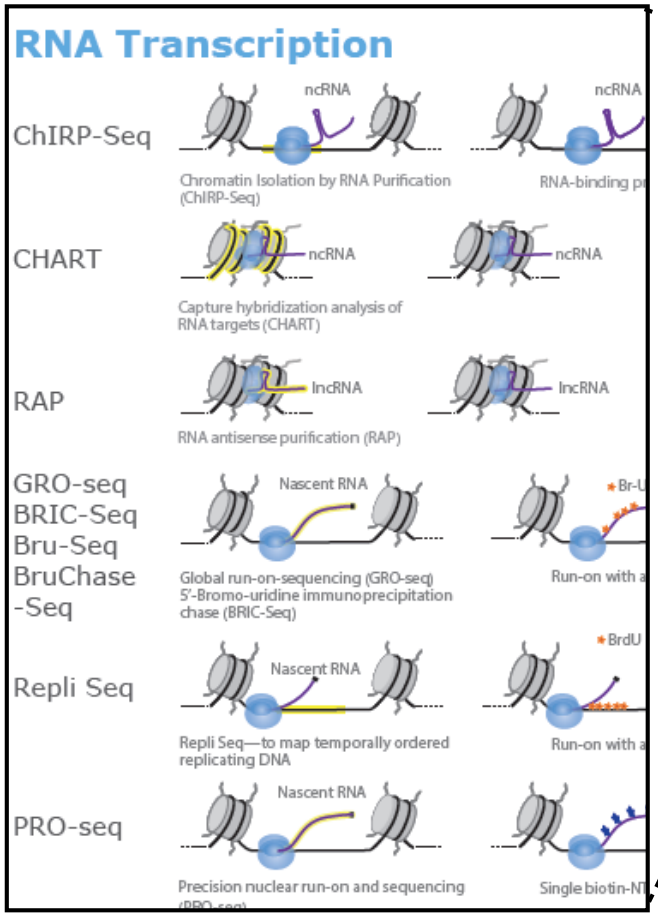
Still, the company was in a state of flux in late 2004 and needed to figure out how to combine the chemistry and technology into a complete system. The researchers needed to meet and brainstorm, operating on an eight-hour time difference between the two branches, one in the UK and one in the US. Ligation chemistry developed by Lynx Therapeutics was an option. There were differing opinions about which chemistry, ligation or polymerase would work best, but ultimately they took a gamble on the Solexa polymerase because it could

Illumina's (Solexa)

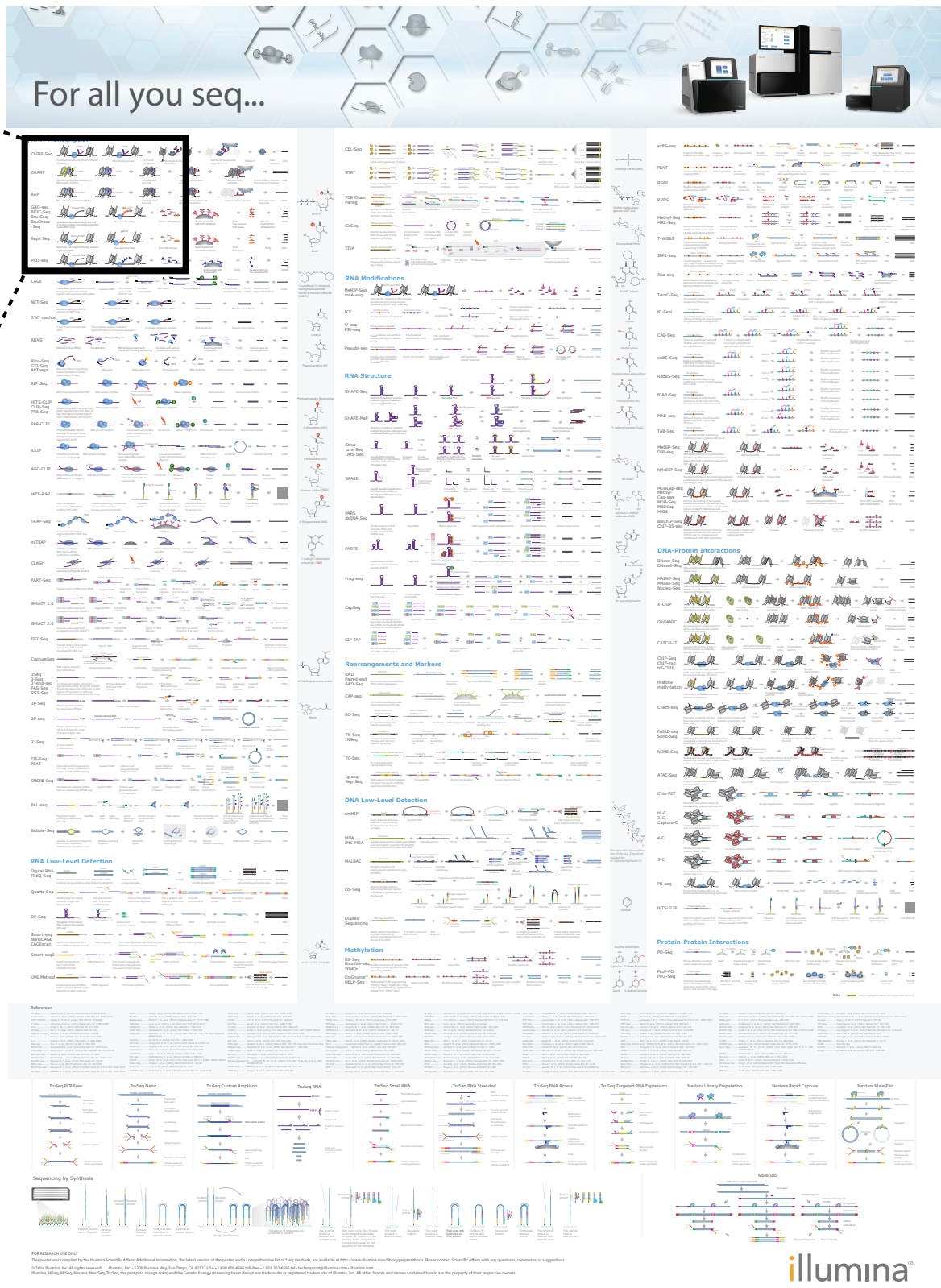
The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing

Ugrappa Nagalakshmi,[1]* Zhong Wang,[1]* Karl Waern,[1] Chong Shou,[2] Debasish Raha,[1] Mark Gerstein,[2,3] Michael Snyder[1,2,3]†

The identification of untranslated regions, introns, and coding regions within an organism remains challenging. We developed a quantitative sequencing-based method called RNA-Seq for mapping transcribed regions, in which complementary DNA fragments are subjected to high-throughput sequencing and mapped to the genome. We applied RNA-Seq to generate a high-resolution transcriptome map of the yeast genome and demonstrated that most (74.5%) of the nonrepetitive sequence of the yeast genome is transcribed. We confirmed many known and predicted introns and demonstrated that others are not actively used. Alternative initiation codons and upstream open reading frames also were identified for many yeast genes. We also found unexpected 3′-end heterogeneity and the presence of many overlapping genes. These results indicate that the yeast transcriptome is more complex than previously appreciated.



http://bit.ly/1NBxOJH



Illumina RNAseq

MiSeq

14-20 million reads
250bp paired-end
24 hr run time

- well-suited for a lab, but not sufficient for transcriptomics
- Best suited for:
  - ✓ amplicon sequencing (16S)
  - ✓ bacterial genome seq
  - ✓ QC of RNAseq libraries

NextSeq

400 million reads
150bp paired-end
24 hr run time

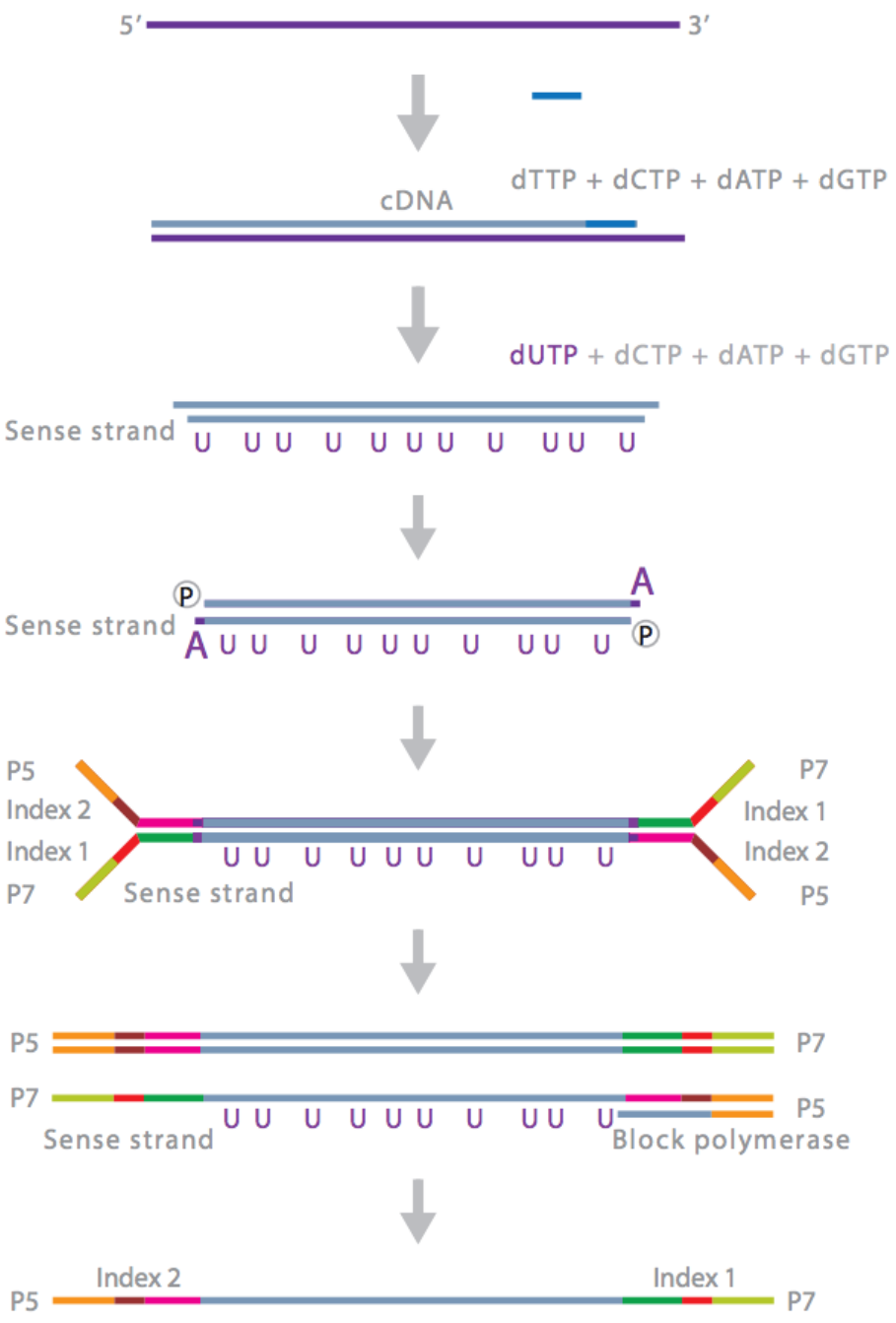
HiSeq

0.3 - 2 billion reads
150-250bp paired-end
5 day run time

- best fit for a large core facility
- requires a dedicated technician
- complicated set-up
- Best suited for:
  - ✓ RNAseq and DNAseq
  - ✓ metagenomics

All sequencing starts with a ‘library prep’

TruSeq RNA Stranded



RNA
Random primer

Create cDNA

Create second strand cDNA

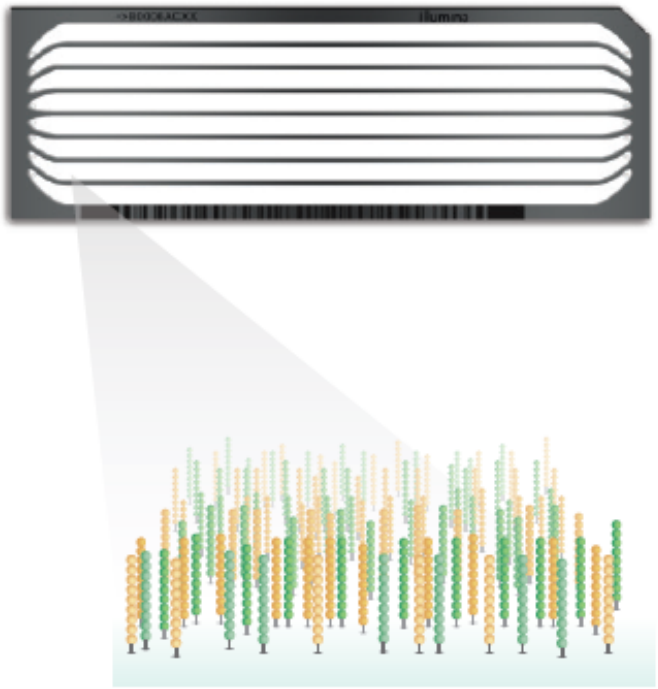
End repair
Phosphorylate
A-overhang

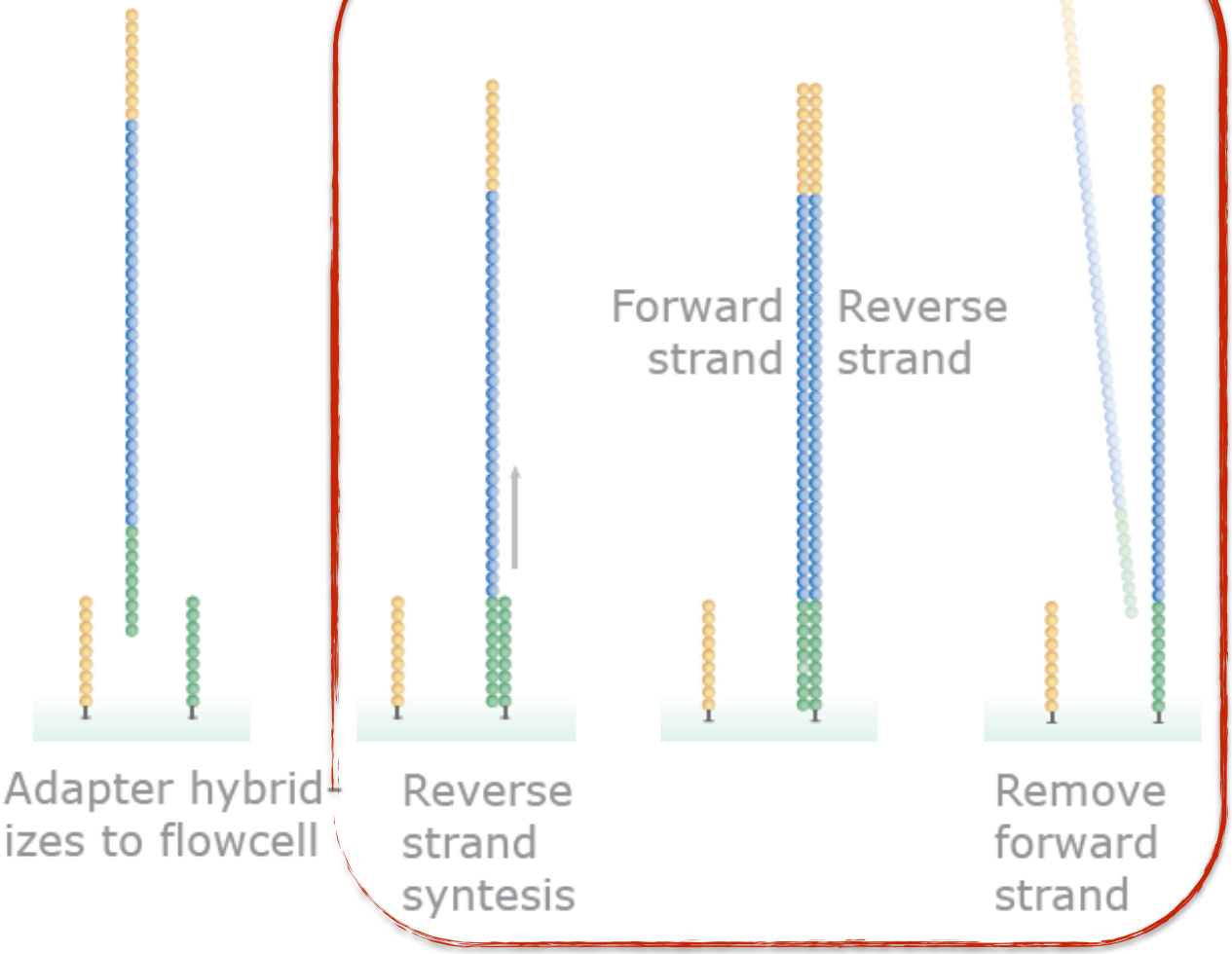
Adaptor ligation

Denature and amplify

Product ready for cluster generation

Sequencing by Synthesis

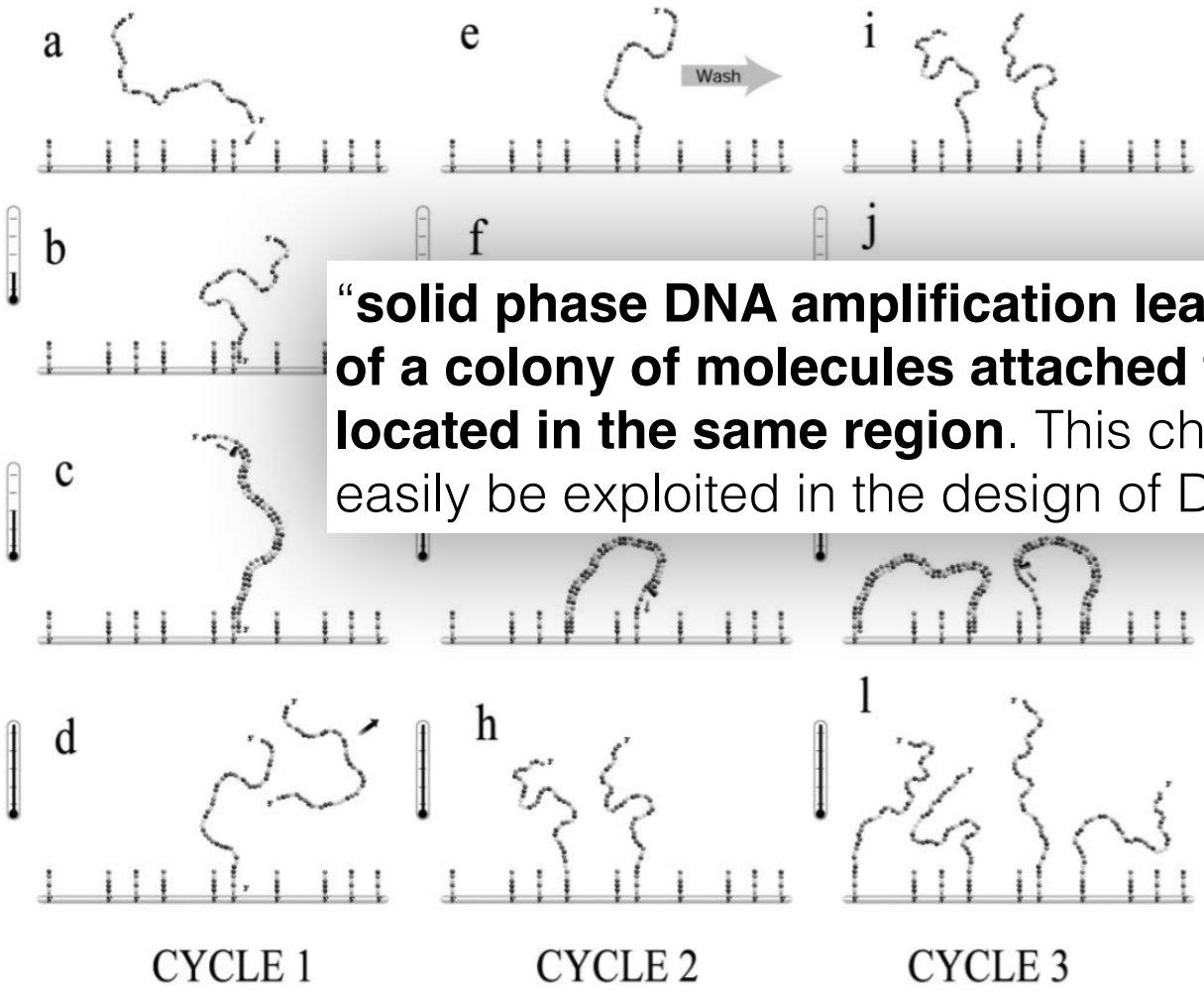




Biophysical Journal Volume 85 October 2003 2075–2086

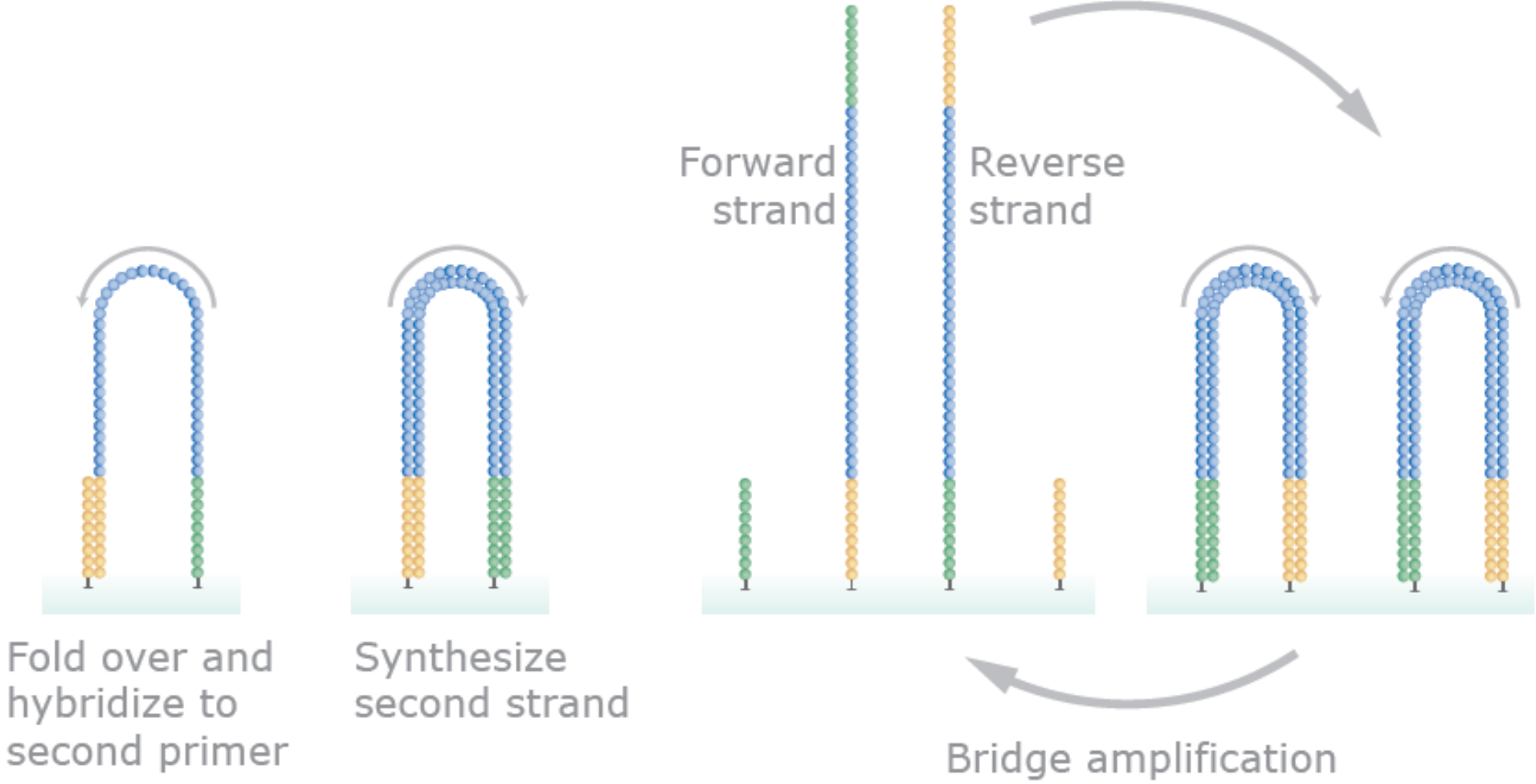
Solid Phase DNA Amplification: A Simple Monte Carlo Lattice Model

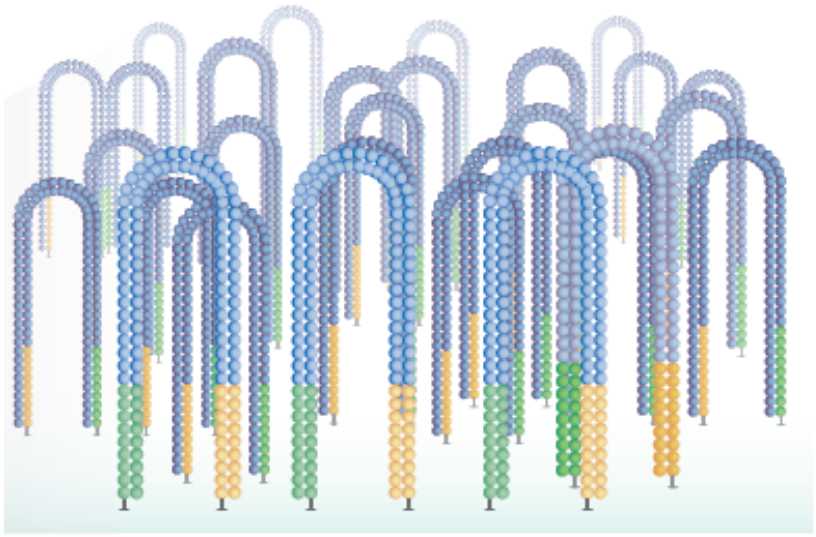
Jean-Francois Mercier,* Gary W. Slater,* and Pascal Mayer†
*Department of Physics, University of Ottawa, Ottawa, Ontario, Canada; and †Manteia Predictive Medicine S.A., Coinsins, Switzerland



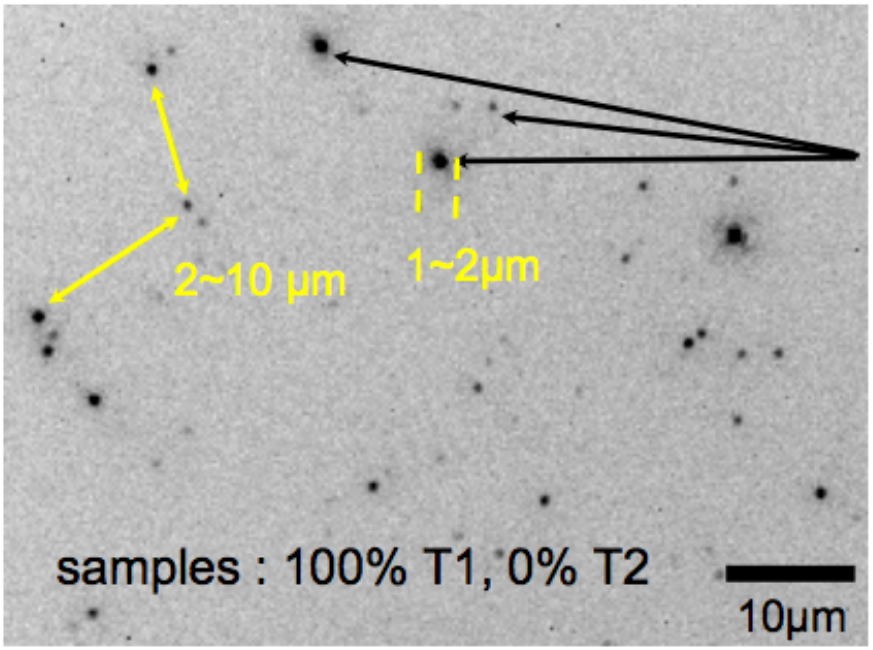
"**solid phase DNA amplification leads to the growth of a colony of molecules attached to the surface and located in the same region**. This characteristic could easily be exploited in the design of DNA microarrays."

Sequencing by Synthesis (SBS)

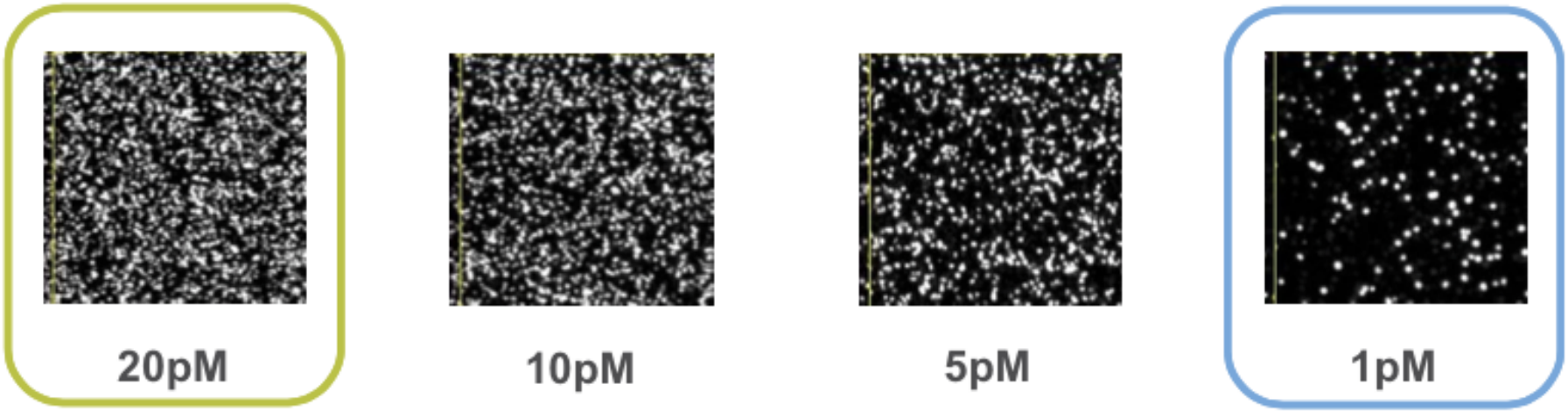




Thousands of molecules are amplified in parallel

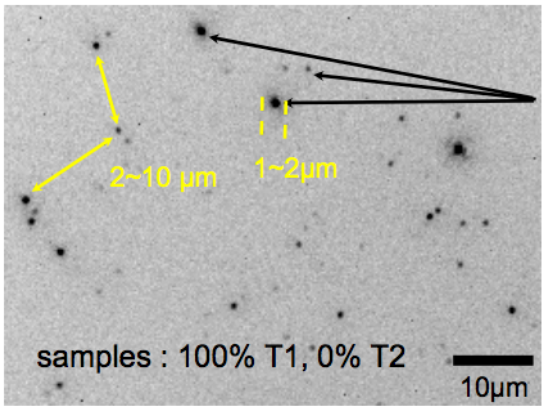


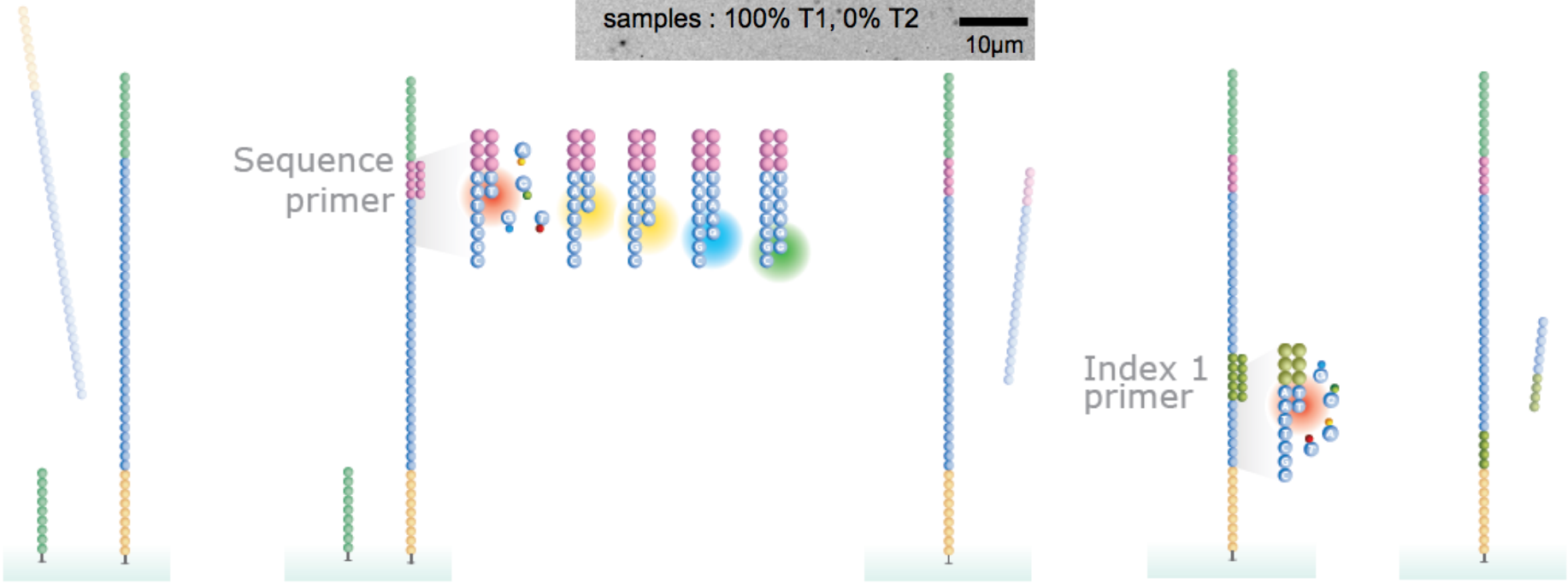
Mayer et al., Presentation 1998



**HiSeq: 600-800K clusters/mm²*
**NextSeq: 130-160K clusters/mm²*

Sequencing by Synthesis (SBS)





The reverse strand is cleaved and washed away

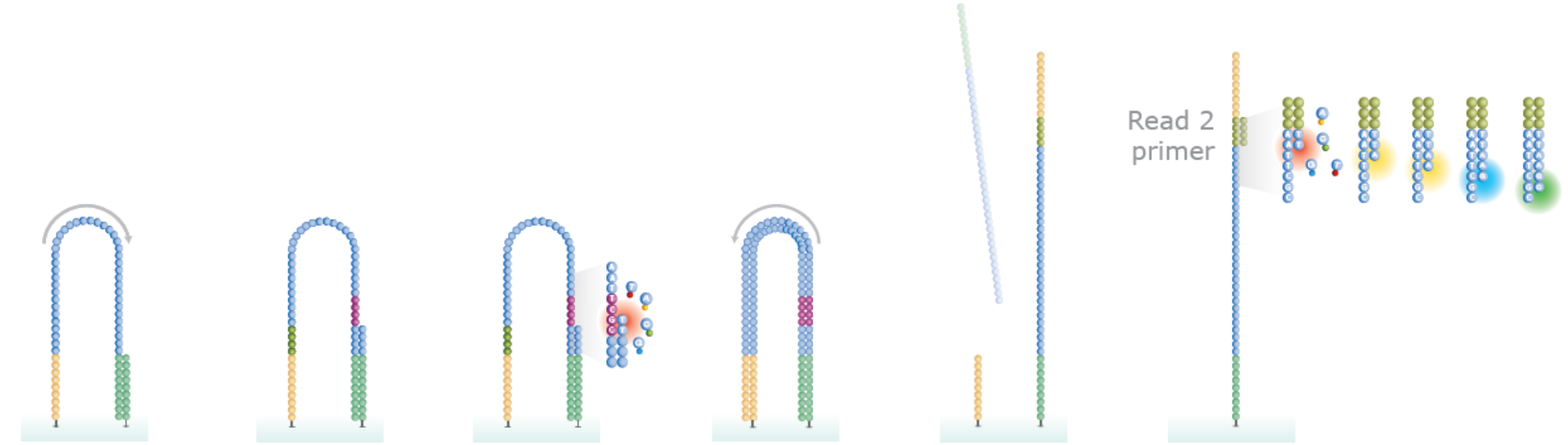
With each cycle, four fluores-cently tagged nucleotides compete for addition to the growing chain. Only one is incorporated based on the sequence of the template.

The read product is washed away

Sequence Index1

The read product is washed away

Sequencing by Synthesis (SBS)



https://www.youtube.com/watch?v=HMyCqWhwB8E

experimental design

1. What is the question that you want to address with sequencing?
2. Do you have independent evaluation that your experiment worked?
3. What are the appropriate controls?
4. What are the questions that will drive your data analysis?
5. How many replicates are needed?
6. How deep should I sequence (reads)?
7. How much will it cost?

Deeper sequencing or more replicates?

BIOINFORMATICS **DISCOVERY NOTE**

Vol. 30 no. 3 2014, pages 301–304
doi:10.1093/bioinformatics/btt688

Gene expression

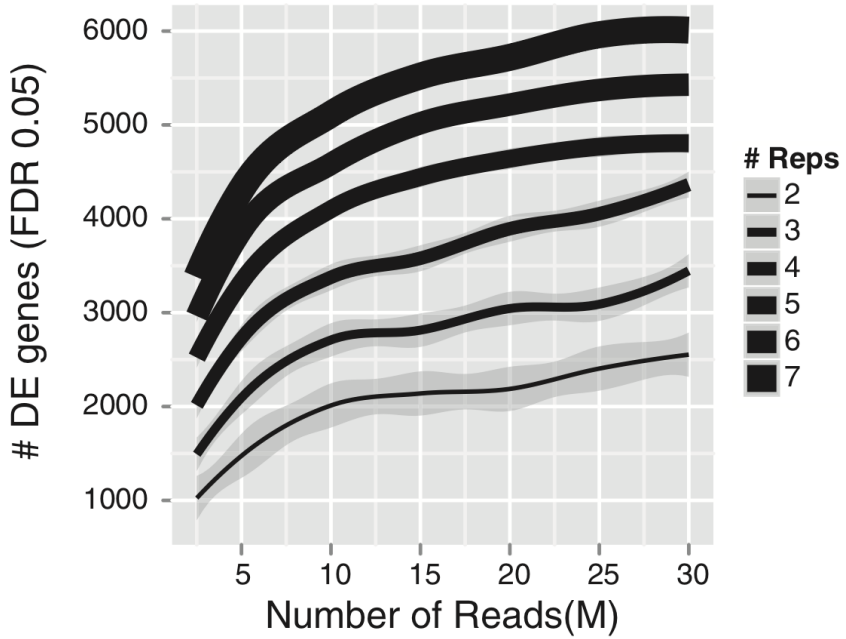
Advance Access publication December 6, 2013

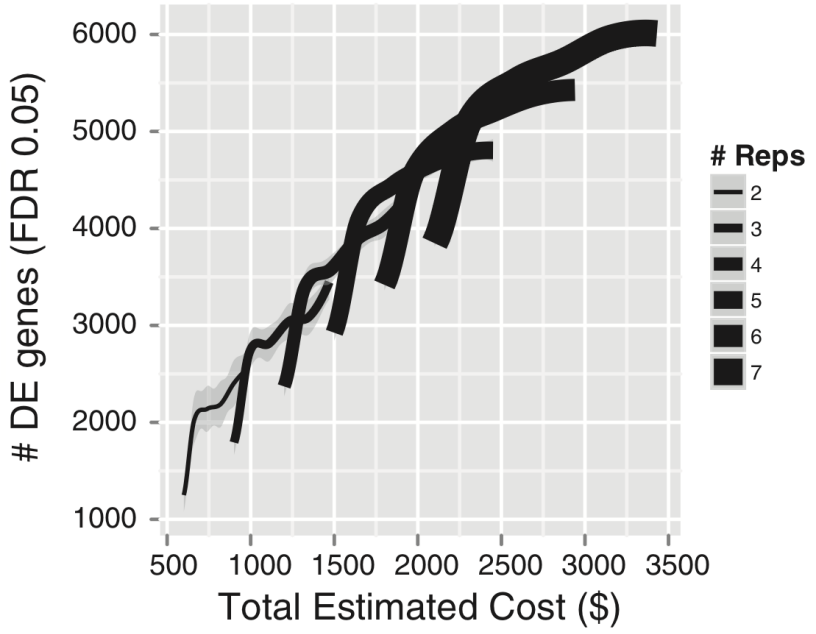
RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

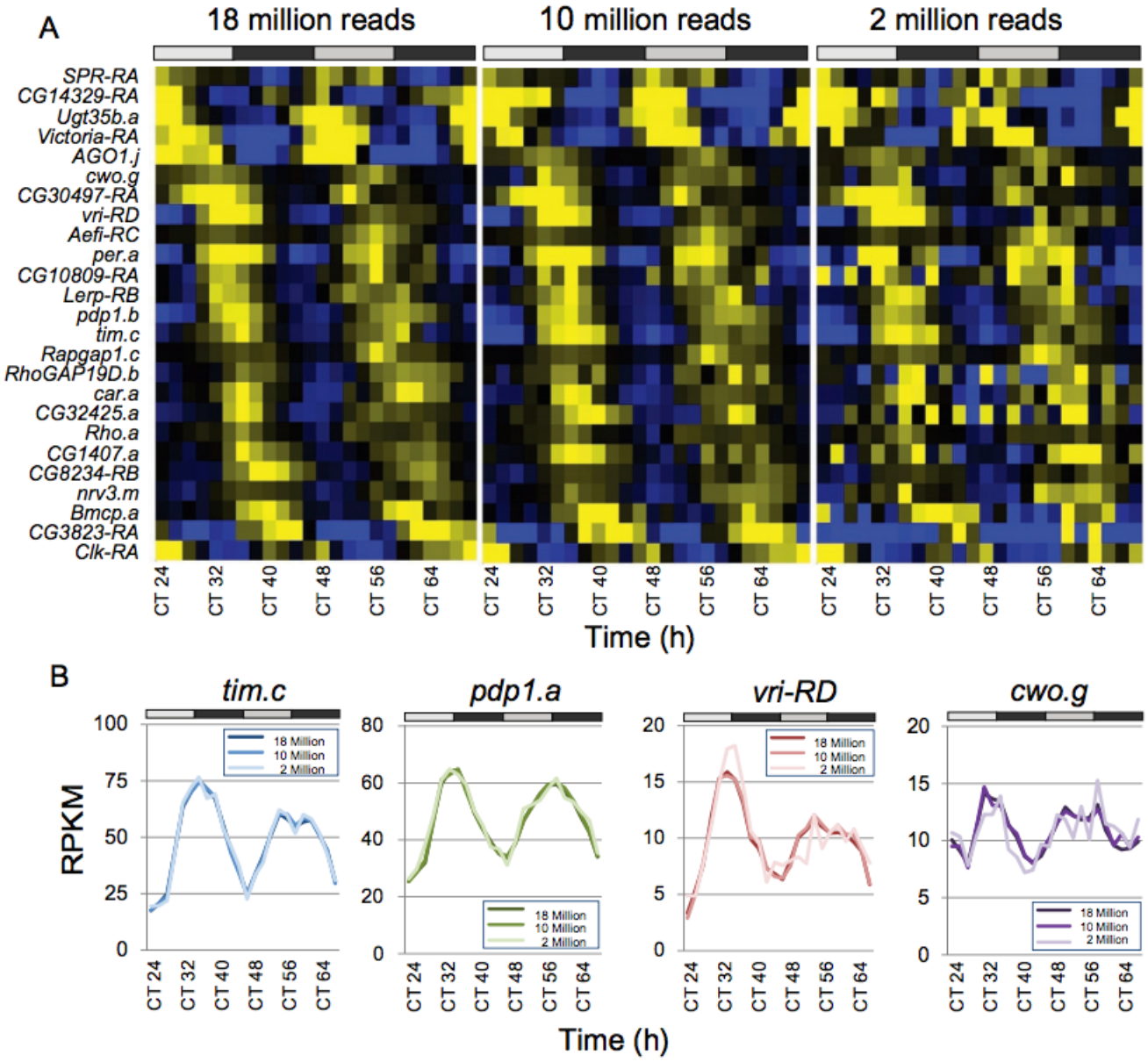
[1]Institute of Genomics and Systems Biology, [2]Committee on Development, Regeneration, and Stem Cell Biology and [3]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

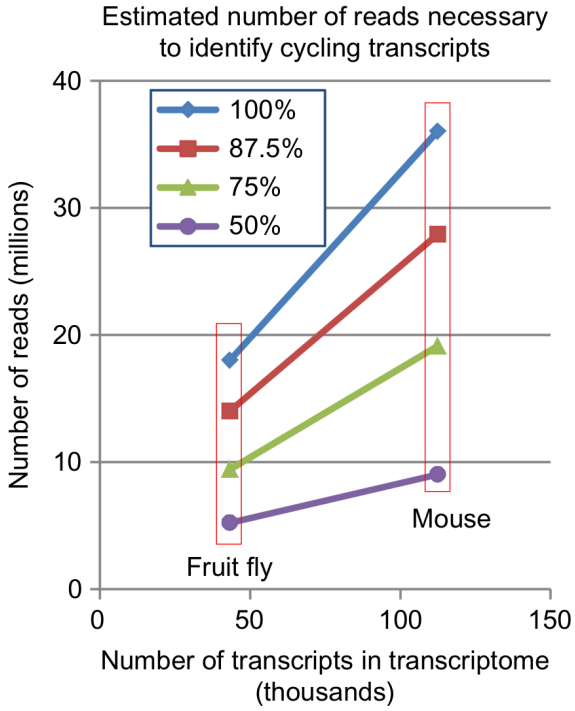
Associate Editor: Janet Kelso





low read depth can still reveal biology





Li et al., Methods Enzymology, 2015

experimental design

1. What is the question that you want to address with sequencing?
2. Do you have independent evaluation that your experiment worked?
3. What are the appropriate controls?
4. What are the questions that will drive your data analysis?
5. How many replicates are needed?
6. How deep should I sequence (reads)?
7. How much will it cost?

Library Preparation

Sequencing

*mid output
130M clusters

*high output
400M clusters

TruSeq kit	cat#	cost
stranded total RNA LT (w/ RiboZero)	RS-122-2201	$5,520
stranded mRNA LT	RS-122-2101	$2,640

*all kits process 48 samples

cycles	cat#	cost
300	FC-404-2004	$4,120
150	FC-404-2002	$2,575
75	FC-404-2005	$1,340

total transcriptome, 150 bp PE

Library Preparation

Sequencing

**mid output*
130M clusters

**high output*
400M clusters

TruSeq kit	cat#	cost
stranded total RNA LT (w/ RiboZero)	RS-122-2201	$5,520
stranded mRNA LT	RS-122-2101	$2,640

**all kits process 48 samples*

cycles	cat#	cost
300	FC-404-2004	$4,120
150	FC-404-2002	$2,575
75	FC-404-2005	$1,340

12 samples
library prep = $115/sample
sequencing = $343/sample
data output = 30M reads/sample

mRNAseq, 150 bp SE

Library Preparation

Sequencing

**mid output*
130M clusters

**high output*
400M clusters

TruSeq kit	cat#	cost
stranded total RNA LT (w/ RiboZero)	RS-122-2201	$5,520
stranded mRNA LT	RS-122-2101	$2,640

**all kits process 48 samples*

cycles	cat#	cost
300	FC-404-2004	$4,120
150	FC-404-2002	$2,575
75	FC-404-2005	$1,340

12 samples
library prep = $55/sample
sequencing = $215/sample
data output = 30M reads/sample

mRNAseq, 75 bp SE

Library Preparation

Sequencing

**mid output*
130M clusters

**high output*
400M clusters

TruSeq kit	cat#	cost
stranded total RNA LT (w/ RiboZero)	RS-122-2201	$5,520
stranded mRNA LT	RS-122-2101	$2,640

**all kits process 48 samples*

cycles	cat#	cost
300	FC-404-2004	$4,120
150	FC-404-2002	$2,575
75	FC-404-2005	$1,340

12 samples
library prep = $55/sample
sequencing = $112/sample
data output = 30M reads/sample

mRNAseq, 75 bp SE (24 samples)

Library Preparation

Sequencing

**mid output*
130M clusters

**high output*
400M clusters

TruSeq kit	cat#	cost
stranded total RNA LT (w/ RiboZero)	RS-122-2201	$5,520
stranded mRNA LT	RS-122-2101	$2,640

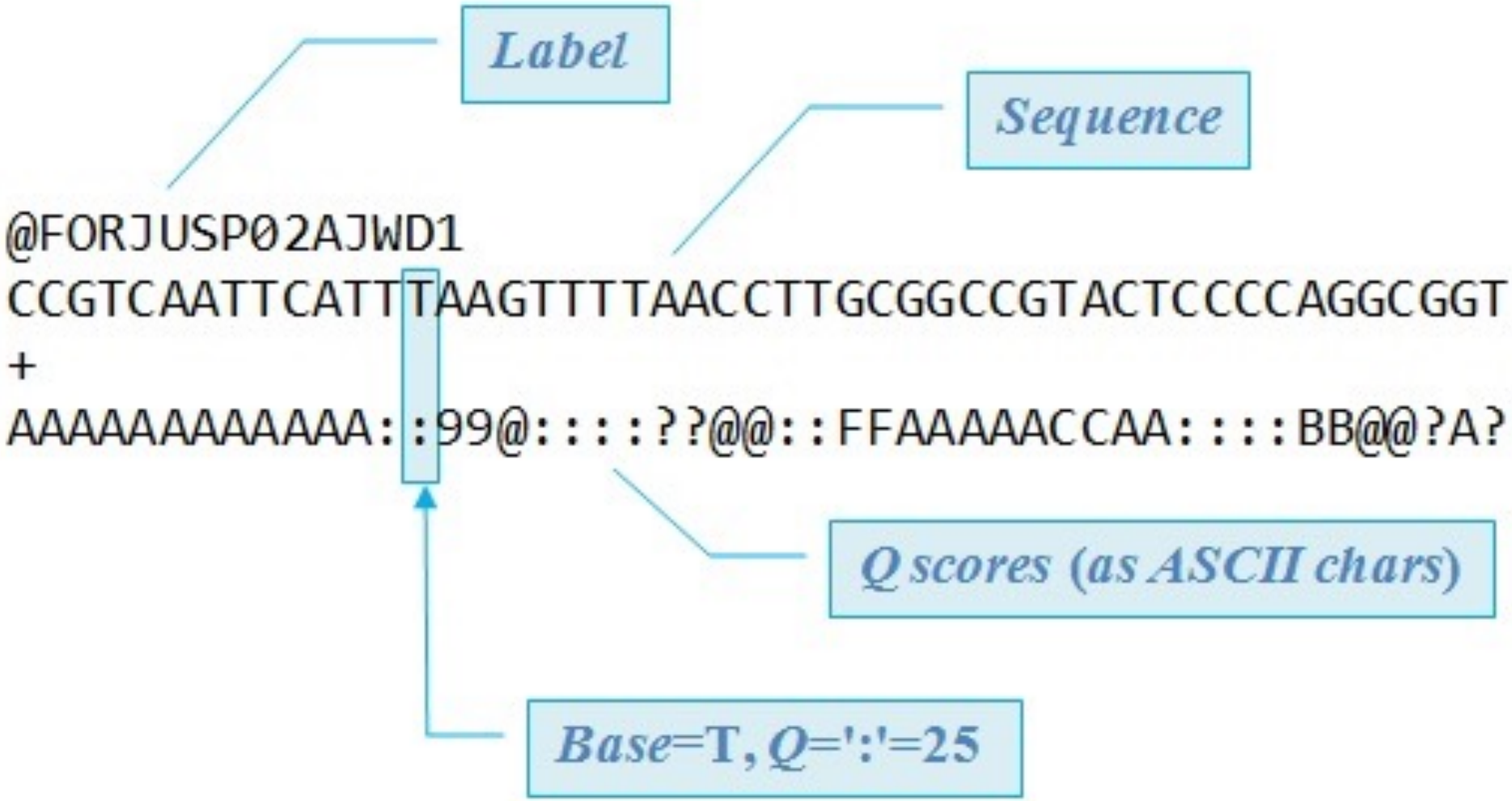
**all kits process 48 samples*

cycles	cat#	cost
300	FC-404-2004	$4,120
150	FC-404-2002	$2,575
75	FC-404-2005	$1,340

24 samples
library prep = $55/sample
sequencing = $56/sample
data output = 15M reads/sample

A word about raw data

RNAseq (.fastq.gz)



5-10Gb/sample


PART 2:
Tools of the trade

Workshop website

Toxoplasma Pre-conference Workshop

Introduction

This is the GitHub repository for the Toxoplasma Pre-Conference Workshop on data analysis and career perspectives. Goals of this workshop include:

- to interact with scientists who have pursued careers outside of the traditional academic tenure-track
- to gain an understanding of how RNAseq data is generated and how the raw data is processed (for practical reasons, we will discuss but not demonstrate this)
- to learn to use the R programming language to manage and explore the data (hands-on with example dataset)
- to learn to use ToxoDB to interpret RNAseq data in the context of a genome browser (also hands-on)

The workshop will have two main components: 1) career perspectives from scientists outside of the traditional tenure-track. 2) hands-on workshop on analysing RNAseq data. The latter will be broken a morning and afternoon workshop. The morning workshop will cover RNAseq analysis in R/bioconductor (Dan Beiting); and the afternoon will focus on using the genome browser to view RNAseq data (Omar Harb). We encourage all attendees to bring their laptops to allow participation in the hands-on exercises.

What you will need to participate

There are four main things you will need in order to particpate in the hands-on aspects of the workshop

1. Your laptop computer

- Everyone is encouraged to bring their own internet-enabled laptop equiped with either the recent Mac or Windows OS.

2. A "toolbox" of free software

- Download and install the appropriate version of the R Programming Language for your operating system
- Download and install the graphical user interface for R, called RStudio
- Download a text editor. I use TextWrangler (Mac only) and Sublime (Mac and PC)

3. Sample dataset

- We will be analyzing RNAseq data from this 2014 PLOS One published by Kami Kim's lab.
- The original raw data is available on the Short Read Archive here, but you **do not** need to download this (files are huge). Instead, please download the 'digital gene expression list' (DGEList) from here
- In addition to the data, please download this simple text file that describes the design of the study

4. An analysis script

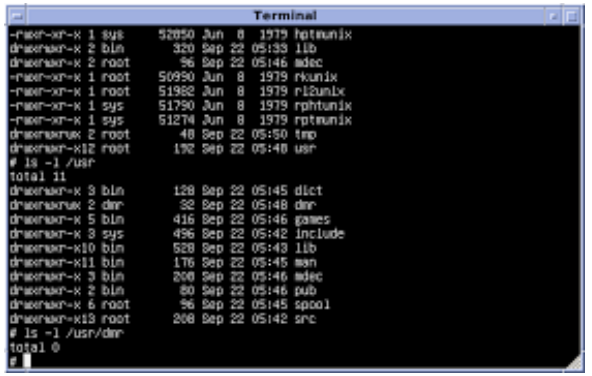
- the R script we'll use in the workshop can be downloaded here.

Text editor	TextWrangler; Sublime	Emacs; sublime
R/bioconductor	R	R
GUI for R	RStudio	RStudio
Terminal	terminal	terminal
Network analysis	Cytoscape	Cytoscape
Graphing	R	R


Perl


MySQL®

Unix (shell)


Terminal
-rwxr-xr-x 1 sys 52050 Jun 8 1979 hptmunix
drwxrwxr-x 2 bin 320 Sep 22 05:33 lib
drwxrwxr-x 2 root 96 Sep 22 05:46 adec
-rwxr-xr-x 1 root 50990 Jun 8 1979 rkunix
-rwxr-xr-x 1 root 51982 Jun 8 1979 rl2unix
-rwxr-xr-x 1 sys 51790 Jun 8 1979 rphtunix
-rwxr-xr-x 1 sys 51274 Jun 8 1979 rptmunix
drwxrwxrwx 2 root 48 Sep 22 05:50 tmp
drwxrwxr-x12 root 192 Sep 22 05:48 usr
ls -l /usr
total 11
drwxrwxr-x 3 bin 128 Sep 22 05:45 dict
drwxrwxrwx 2 dmr 32 Sep 22 05:48 dmr
drwxrwxr-x 5 bin 416 Sep 22 05:46 games
drwxrwxr-x 3 sys 496 Sep 22 05:42 include
drwxrwxr-x10 bin 528 Sep 22 05:43 lib
drwxrwxr-x11 bin 176 Sep 22 05:45 man
drwxrwxr-x 3 bin 208 Sep 22 05:46 adec
drwxrwxr-x 2 bin 80 Sep 22 05:46 pub
drwxrwxr-x 6 root 96 Sep 22 05:45 spool
drwxrwxr-x13 root 208 Sep 22 05:42 src
ls -l /usr/dmr
total 0
#


Ruby


R


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS


python™

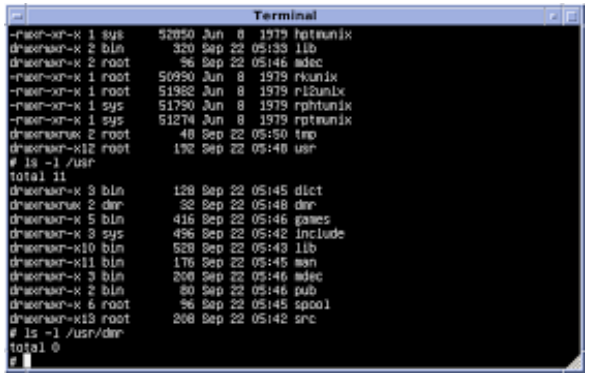

C++


Java™

Perl

MySQL

Unix (shell)


Terminal
-rwxr-xr-x 1 sys 52050 Jun 8 1979 hptmunix
drwxrwxr-x 2 bin 320 Sep 22 05:33 lib
drwxrwxr-x 2 root 96 Sep 22 05:46 wdec
-rwxr-xr-x 1 root 50950 Jun 8 1979 rkunix
-rwxr-xr-x 1 root 51982 Jun 8 1979 rl2unix
-rwxr-xr-x 1 sys 51790 Jun 8 1979 rphtunix
-rwxr-xr-x 1 sys 51274 Jun 8 1979 rptmunix
drwxrwxrwx 2 root 48 Sep 22 05:50 tmp
drwxrwxr-x12 root 192 Sep 22 05:40 usr
ls -l /usr
total 11
drwxrwxr-x 3 bin 128 Sep 22 05:45 dict
drwxrwxrwx 2 dmr 32 Sep 22 05:48 dmr
drwxrwxr-x 5 bin 416 Sep 22 05:46 games
drwxrwxr-x 3 sys 496 Sep 22 05:42 include
drwxrwxr-x10 bin 528 Sep 22 05:43 lib
drwxrwxr-x11 bin 176 Sep 22 05:45 man
drwxrwxr-x 3 bin 208 Sep 22 05:46 wdec
drwxrwxr-x 2 bin 80 Sep 22 05:46 pub
drwxrwxr-x 6 root 96 Sep 22 05:45 spool
drwxrwxr-x13 root 208 Sep 22 05:42 src
ls -l /usr/dmr
total 0
#


R

Ruby

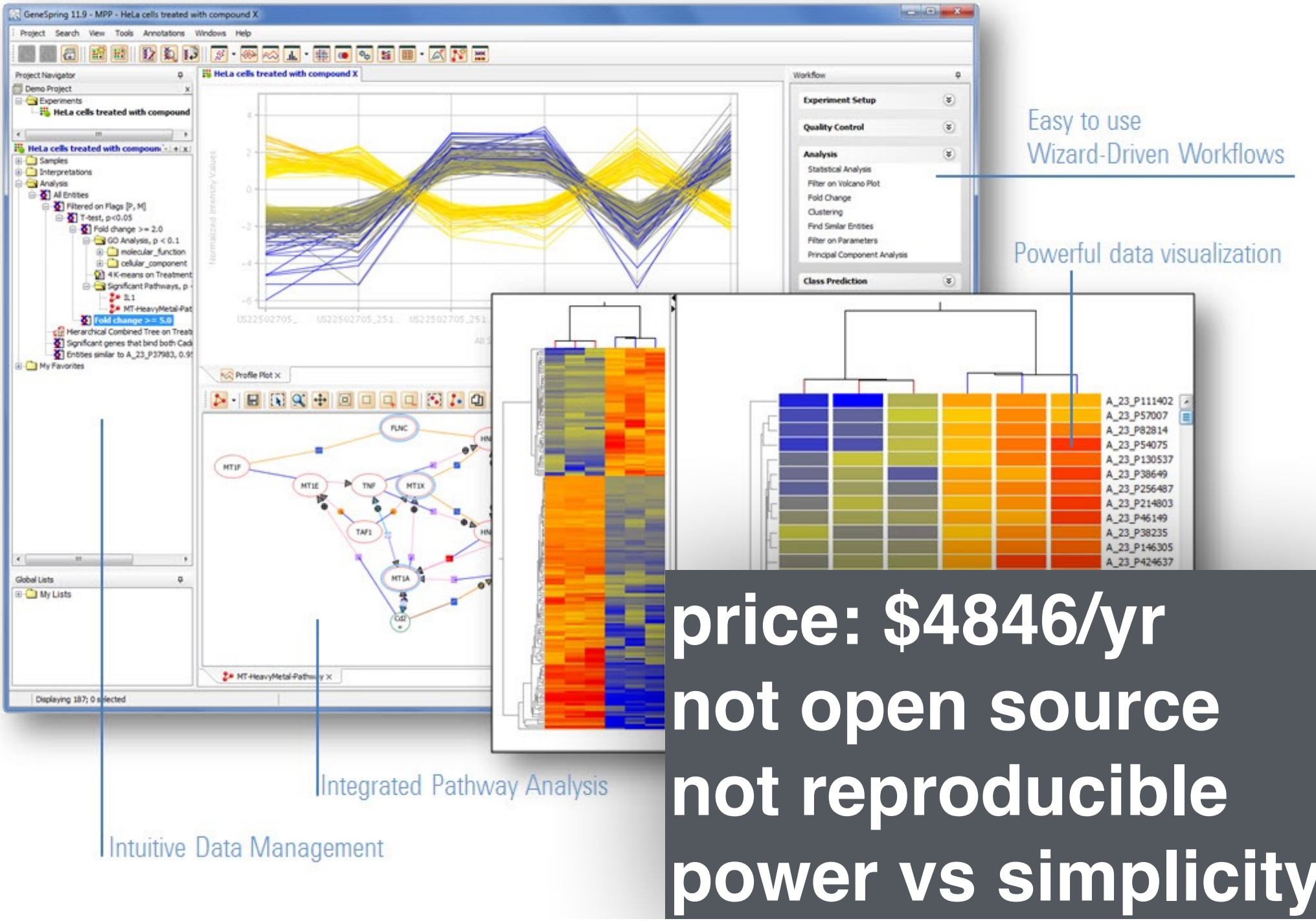

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

C++

Java

python

Commerical Solutions



Open R

What you need to participate…

There are four main things you will need in order to particpate in the hands-on aspects of the workshop

1. Your laptop computer

- Everyone is encouraged to bring their own internet-enabled laptop equiped with either the recent Mac or Windows OS.

2. A "toolbox" of free software

- Download and install the appropriate version of the R Programming Language for your operating system
- Download and install the graphical user interface for R, called RStudio
- Download a text editor. I use TextWrangler (Mac only) and Sublime (Mac and PC)

3. Sample dataset

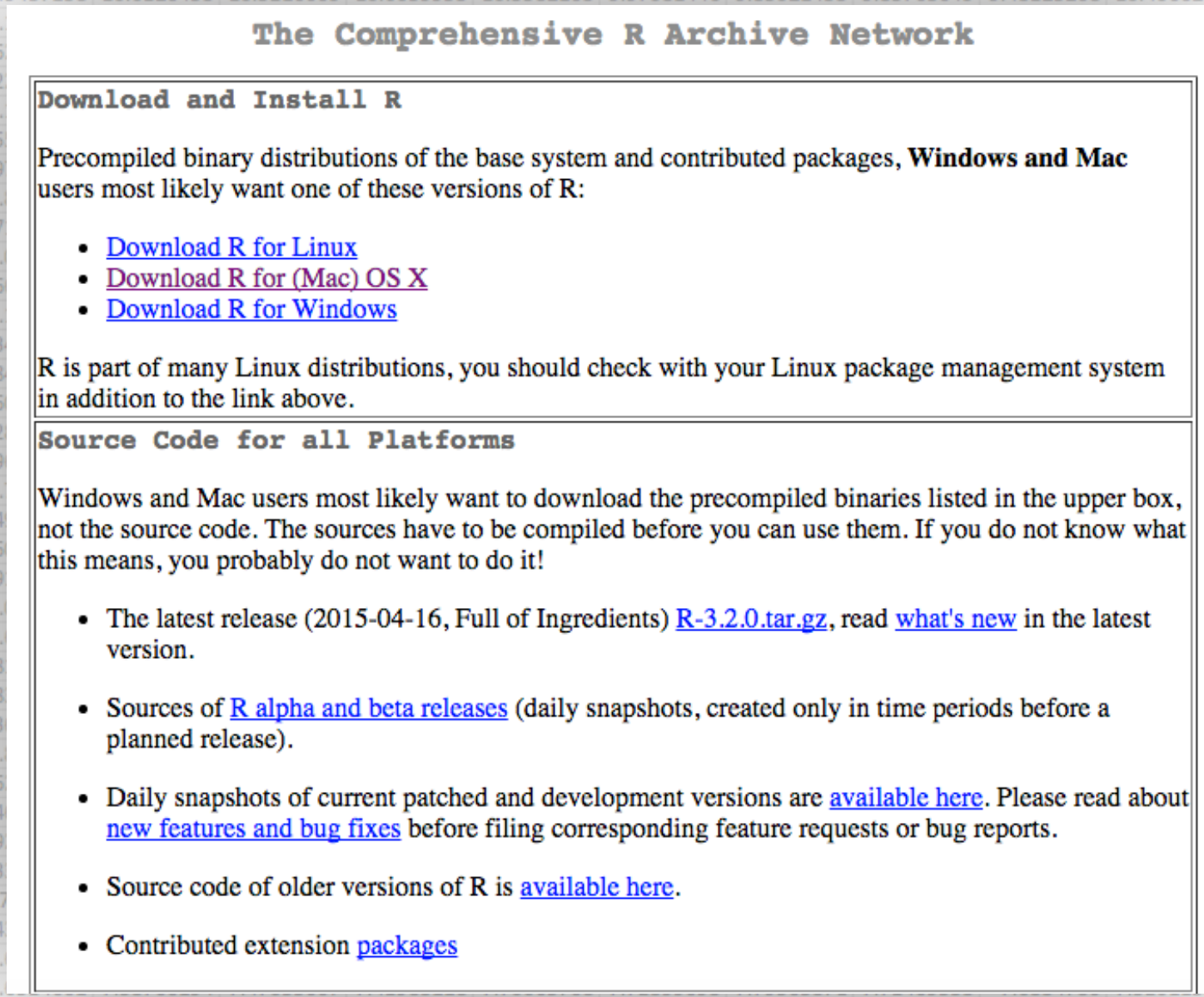
- We will be analyzing RNAseq data from this 2014 PLOS One published by Kami Kim's lab.
- The original raw data is available on the Short Read Archive here, but you **do not** need to download this (files are huge). Instead, please download the 'digital gene expression list' (DGEList) from here
- In addition to the data, please download this simple text file that describes the design of the study

4. An analysis script

- the R script we'll use in the workshop can be downloaded here.

ToxoGenomicsWorkshop.github.io

...you need R



The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2015-04-16, Full of Ingredients) R-3.2.0.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

R Console

~

Help Search

```
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.65 (6833) x86_64-apple-darwin13.4.0]

[Workspace restored from /Users/danielbeiting/.RData]
[History restored from /Users/danielbeiting/.Rhistory]

>
```

Why R/bioconductor?

A powerful and flexible tool for bioinformatics

Pros

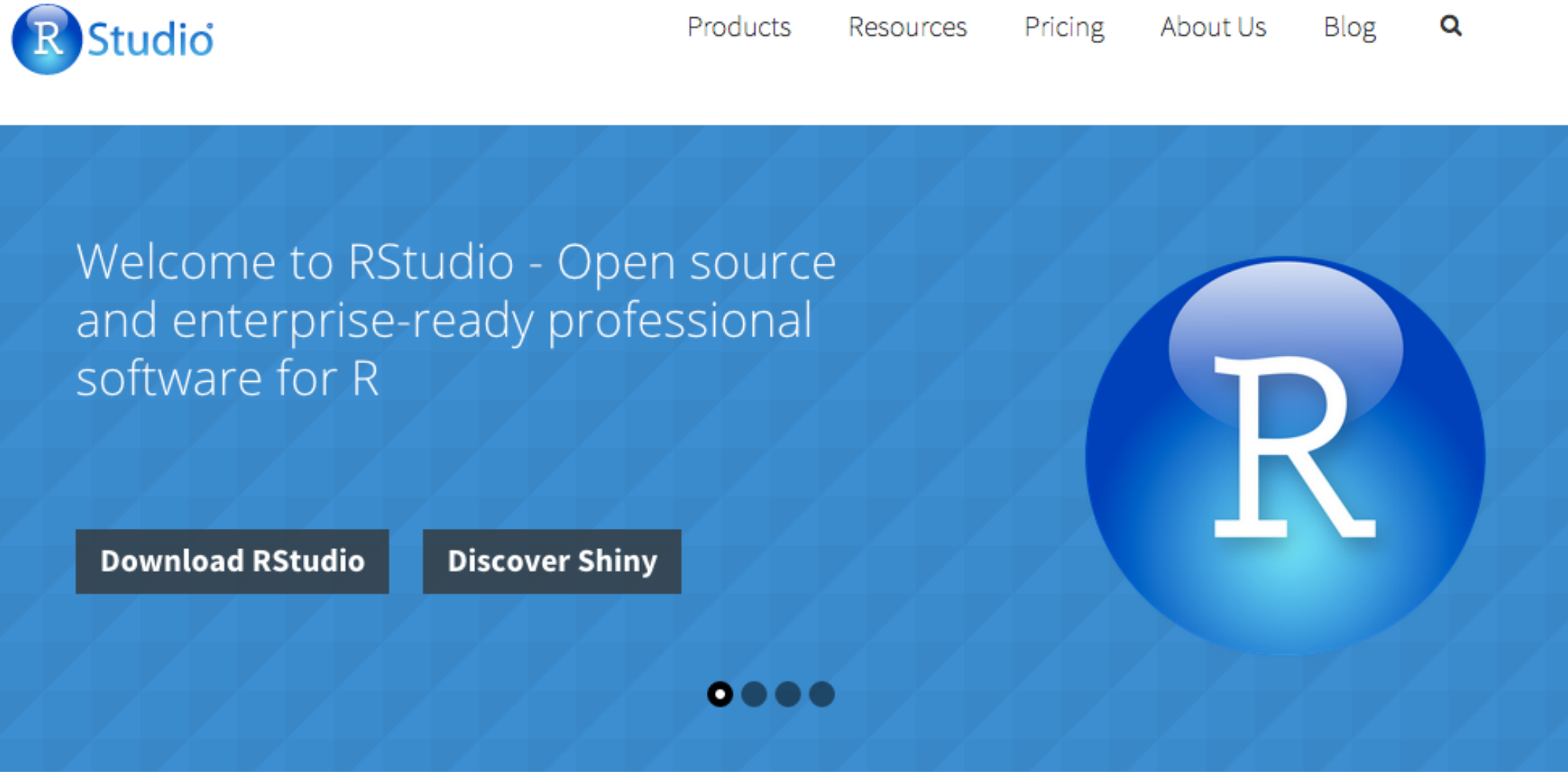
- R supports missing values - handles real-world data
- R's package system - simple method for handling user-contributed functions so that others understand them (standardized, modular, and extensible)
- used for over a decade to analyze genomic data
- often don't need to write code *de novo*
- huge support community
- publication quality graphics
- not just arrays - ChIPseq, RNAseq, genetic screens, interaction networks

Why R/bioconductor?

A powerful and flexible tool for bioinformatics

Cons

- steep learning curve
- no single ‘right way’ to do anything
- not as fast as C++
- historically, R was lacking in terms of user interface


RStudio
Products
Resources
Pricing
About Us
Blog
Welcome to RStudio - Open source and enterprise-ready professional software for R
Download RStudio
Discover Shiny


RStudio®

What you need to participate…

There are four main things you will need in order to particpate in the hands-on aspects of the workshop

1. Your laptop computer

- Everyone is encouraged to bring their own internet-enabled laptop equiped with either the recent Mac or Windows OS.

2. A "toolbox" of free software

- Download and install the appropriate version of the R Programming Language for your operating system
- Download and install the graphical user interface for R, called RStudio
- Download a text editor. I use TextWrangler (Mac only) and Sublime (Mac and PC)

3. Sample dataset

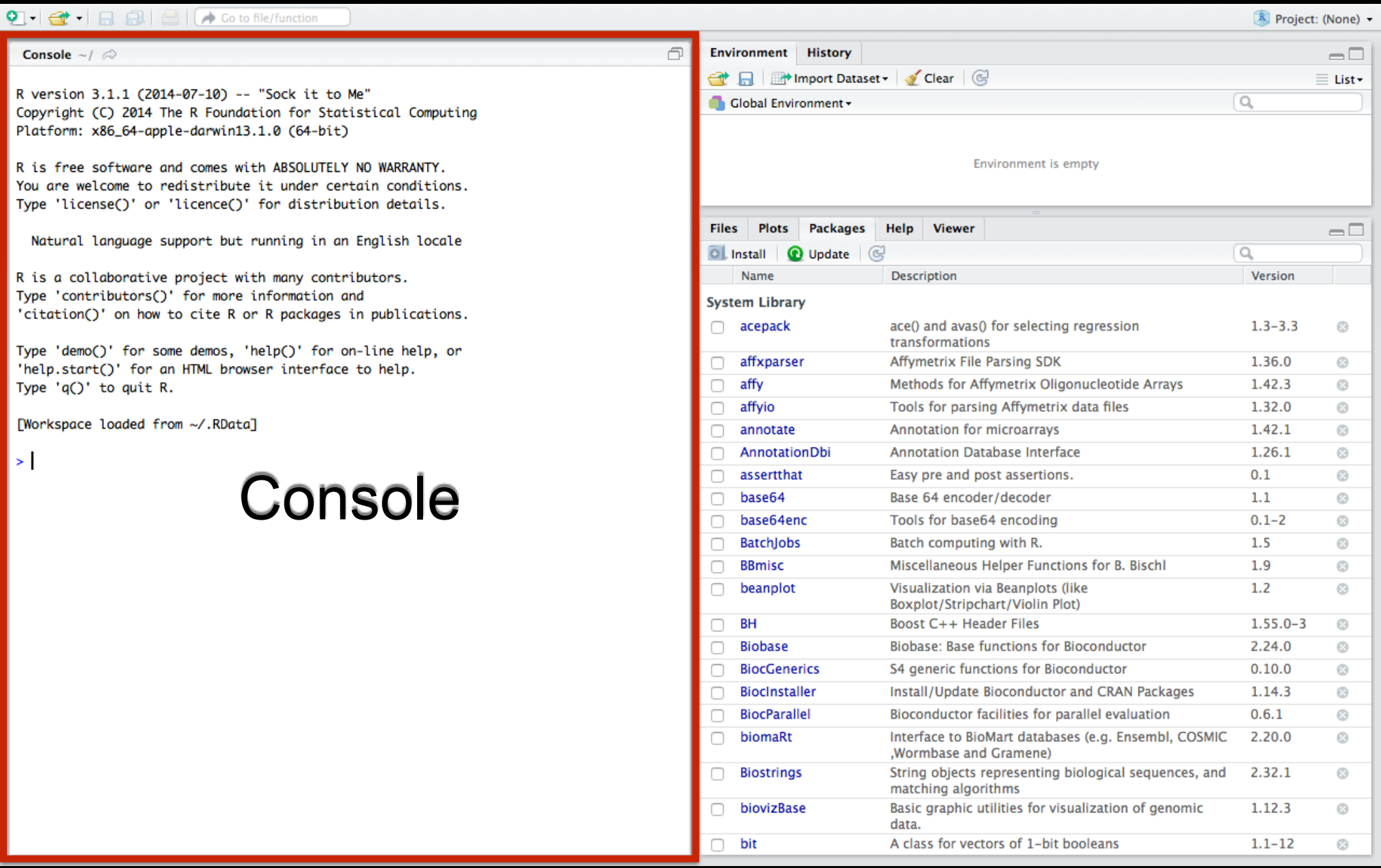
- We will be analyzing RNAseq data from this 2014 PLOS One published by Kami Kim's lab.
- The original raw data is available on the Short Read Archive here, but you **do not** need to download this (files are huge). Instead, please download the 'digital gene expression list' (DGEList) from here
- In addition to the data, please download this simple text file that describes the design of the study

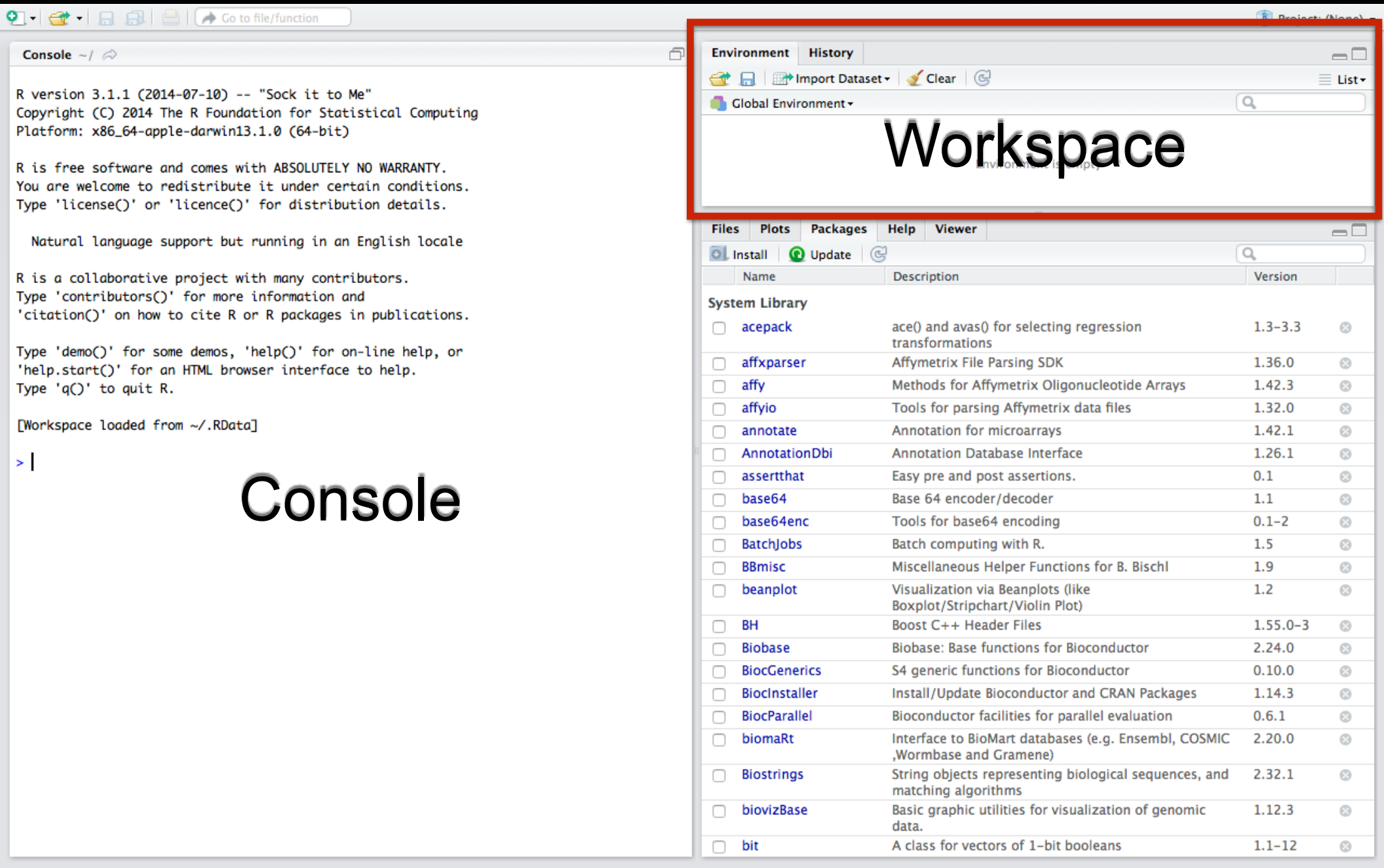
4. An analysis script

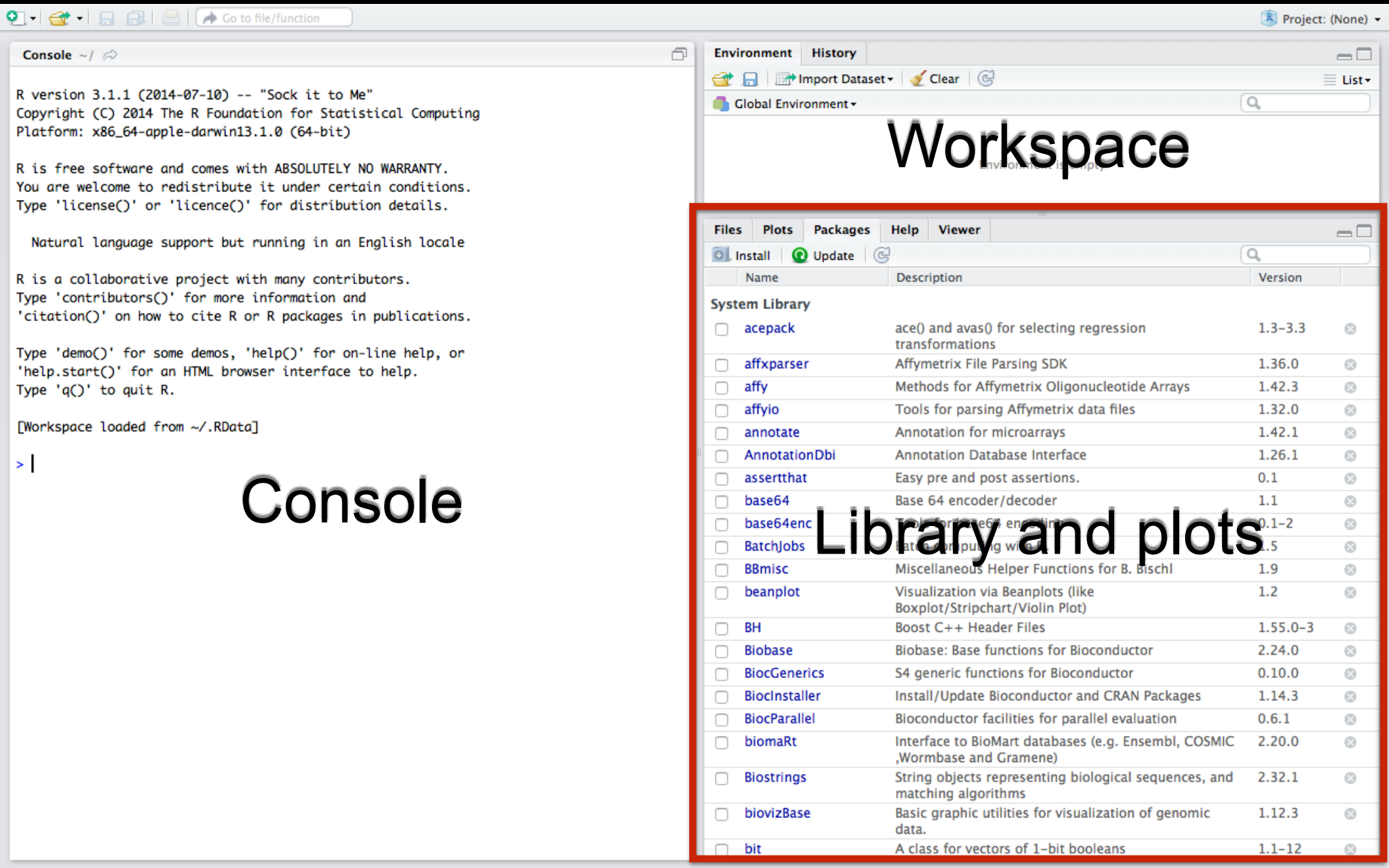
- the R script we'll use in the workshop can be downloaded here.

ToxoGenomicsWorkshop.github.io

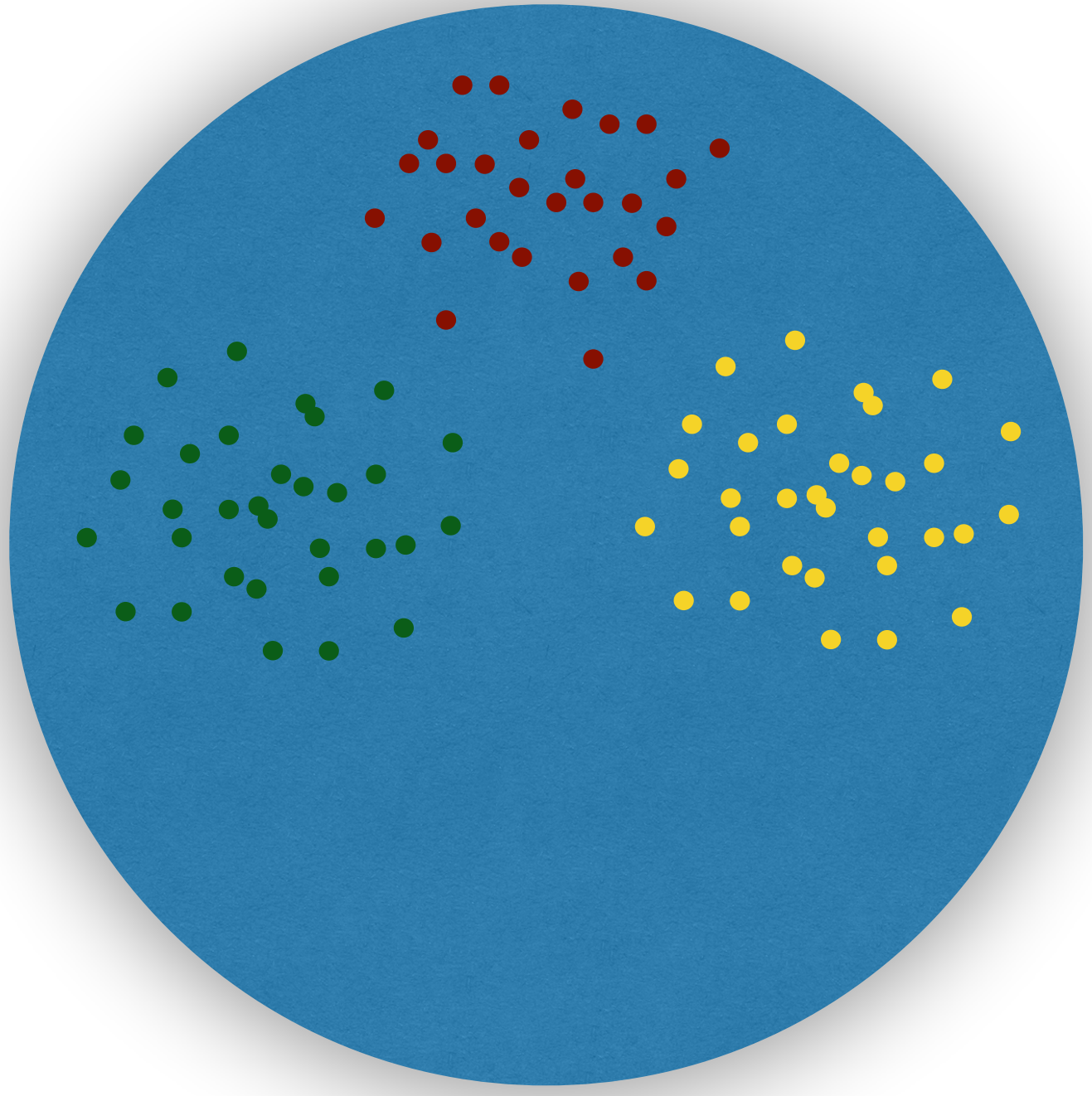
Open RStudio


Go to file/function
Project: (None)
Console ~/
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.1.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Console
Environment
History
Import Dataset
Clear
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
System Library
acepack
ace() and avas() for selecting regression transformations
1.3-3.3
affxparser
Affymetrix File Parsing SDK
1.36.0
affy
Methods for Affymetrix Oligonucleotide Arrays
1.42.3
affyio
Tools for parsing Affymetrix data files
1.32.0
annotate
Annotation for microarrays
1.42.1
AnnotationDbi
Annotation Database Interface
1.26.1
assertthat
Easy pre and post assertions.
0.1
base64
Base 64 encoder/decoder
1.1
base64enc
Tools for base64 encoding
0.1-2
BatchJobs
Batch computing with R.
1.5
BBmisc
Miscellaneous Helper Functions for B. Bischl
1.9
beanplot
Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot)
1.2
BH
Boost C++ Header Files
1.55.0-3
Biobase
Biobase: Base functions for Bioconductor
2.24.0
BiocGenerics
S4 generic functions for Bioconductor
0.10.0
BiocInstaller
Install/Update Bioconductor and CRAN Packages
1.14.3
BiocParallel
Bioconductor facilities for parallel evaluation
0.6.1
biomaRt
Interface to BioMart databases (e.g. Ensembl, COSMIC ,Wormbase and Gramene)
2.20.0
Biostrings
String objects representing biological sequences, and matching algorithms
2.32.1
biovizBase
Basic graphic utilities for visualization of genomic data.
1.12.3
bit
A class for vectors of 1-bit booleans
1.1-12


Go to file/function
Console ~/
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.1.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Console
Environment
History
Import Dataset
Clear
List
Global Environment
Workspace
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
System Library
acepack
ace() and avas() for selecting regression transformations
1.3-3.3
affxparser
Affymetrix File Parsing SDK
1.36.0
affy
Methods for Affymetrix Oligonucleotide Arrays
1.42.3
affyio
Tools for parsing Affymetrix data files
1.32.0
annotate
Annotation for microarrays
1.42.1
AnnotationDbi
Annotation Database Interface
1.26.1
assertthat
Easy pre and post assertions.
0.1
base64
Base 64 encoder/decoder
1.1
base64enc
Tools for base64 encoding
0.1-2
BatchJobs
Batch computing with R.
1.5
BBmisc
Miscellaneous Helper Functions for B. Bischl
1.9
beanplot
Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot)
1.2
BH
Boost C++ Header Files
1.55.0-3
Biobase
Biobase: Base functions for Bioconductor
2.24.0
BiocGenerics
S4 generic functions for Bioconductor
0.10.0
BiocInstaller
Install/Update Bioconductor and CRAN Packages
1.14.3
BiocParallel
Bioconductor facilities for parallel evaluation
0.6.1
biomaRt
Interface to BioMart databases (e.g. Ensembl, COSMIC ,Wormbase and Gramene)
2.20.0
Biostrings
String objects representing biological sequences, and matching algorithms
2.32.1
biovizBase
Basic graphic utilities for visualization of genomic data.
1.12.3
bit
A class for vectors of 1-bit booleans
1.1-12


Go to file/function
Project: (None)
Console ~/
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.1.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
>
Console
Environment
History
Import Dataset
Clear
List
Global Environment
Workspace
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
System Library
acepack
ace() and avas() for selecting regression transformations
1.3-3.3
affxparser
Affymetrix File Parsing SDK
1.36.0
affy
Methods for Affymetrix Oligonucleotide Arrays
1.42.3
affyio
Tools for parsing Affymetrix data files
1.32.0
annotate
Annotation for microarrays
1.42.1
AnnotationDbi
Annotation Database Interface
1.26.1
assertthat
Easy pre and post assertions.
0.1
base64
Base 64 encoder/decoder
1.1
base64enc
0.1-2
BatchJobs
1.5
BBmisc
Miscellaneous Helper Functions for B. Bischl
1.9
beanplot
Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot)
1.2
BH
Boost C++ Header Files
1.55.0-3
Biobase
Biobase: Base functions for Bioconductor
2.24.0
BiocGenerics
S4 generic functions for Bioconductor
0.10.0
BiocInstaller
Install/Update Bioconductor and CRAN Packages
1.14.3
BiocParallel
Bioconductor facilities for parallel evaluation
0.6.1
biomaRt
Interface to BioMart databases (e.g. Ensembl, COSMIC ,Wormbase and Gramene)
2.20.0
Biostrings
String objects representing biological sequences, and matching algorithms
2.32.1
biovizBase
Basic graphic utilities for visualization of genomic data.
1.12.3
bit
A class for vectors of 1-bit booleans
1.1-12
Library and plots

The wonderful world of R



Package type
- statistics
- graphing
- modeling

~6500 packages

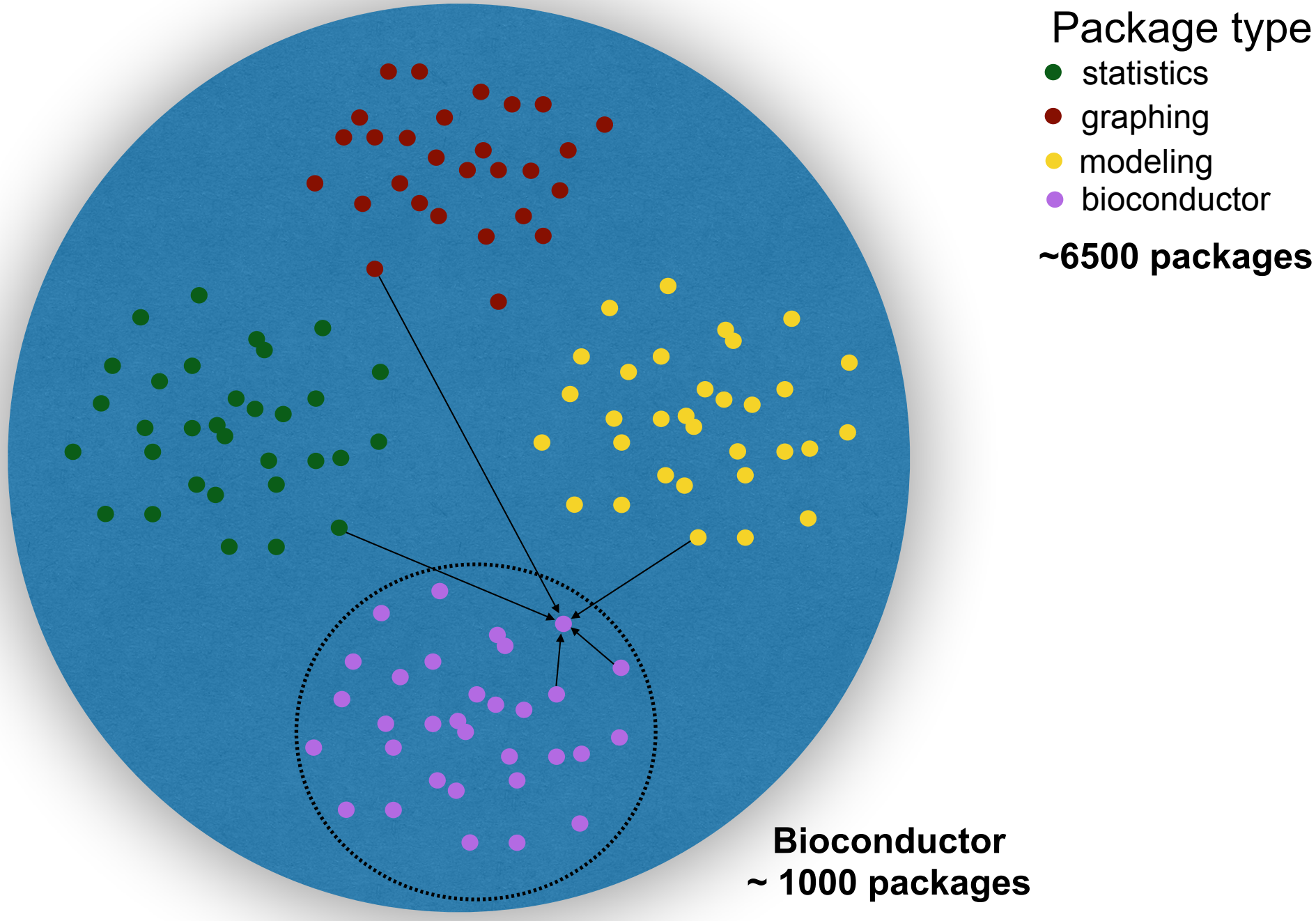
Orchestrating high-throughput genomic analysis with Bioconductor

Wolfgang Huber[1], Vincent J Carey[2,3], Robert Gentleman[4], Simon Anders[1], Marc Carlson[5], Benilton S Carvalho[6], Hector Corrada Bravo[7], Sean Davis[8], Laurent Gatto[9], Thomas Girke[10], Raphael Gottardo[11], Florian Hahne[12], Kasper D Hansen[13,14], Rafael A Irizarry[3,15], Michael Lawrence[4], Michael I Love[3,15], James MacDonald[16], Valerie Obenchain[5], Andrzej K Oleś[1], Hervé Pagès[5], Alejandro Reyes[1], Paul Shannon[5], Gordon K Smyth[17,18], Dan Tenenbaum[5], Levi Waldron[19] & Martin Morgan[5]

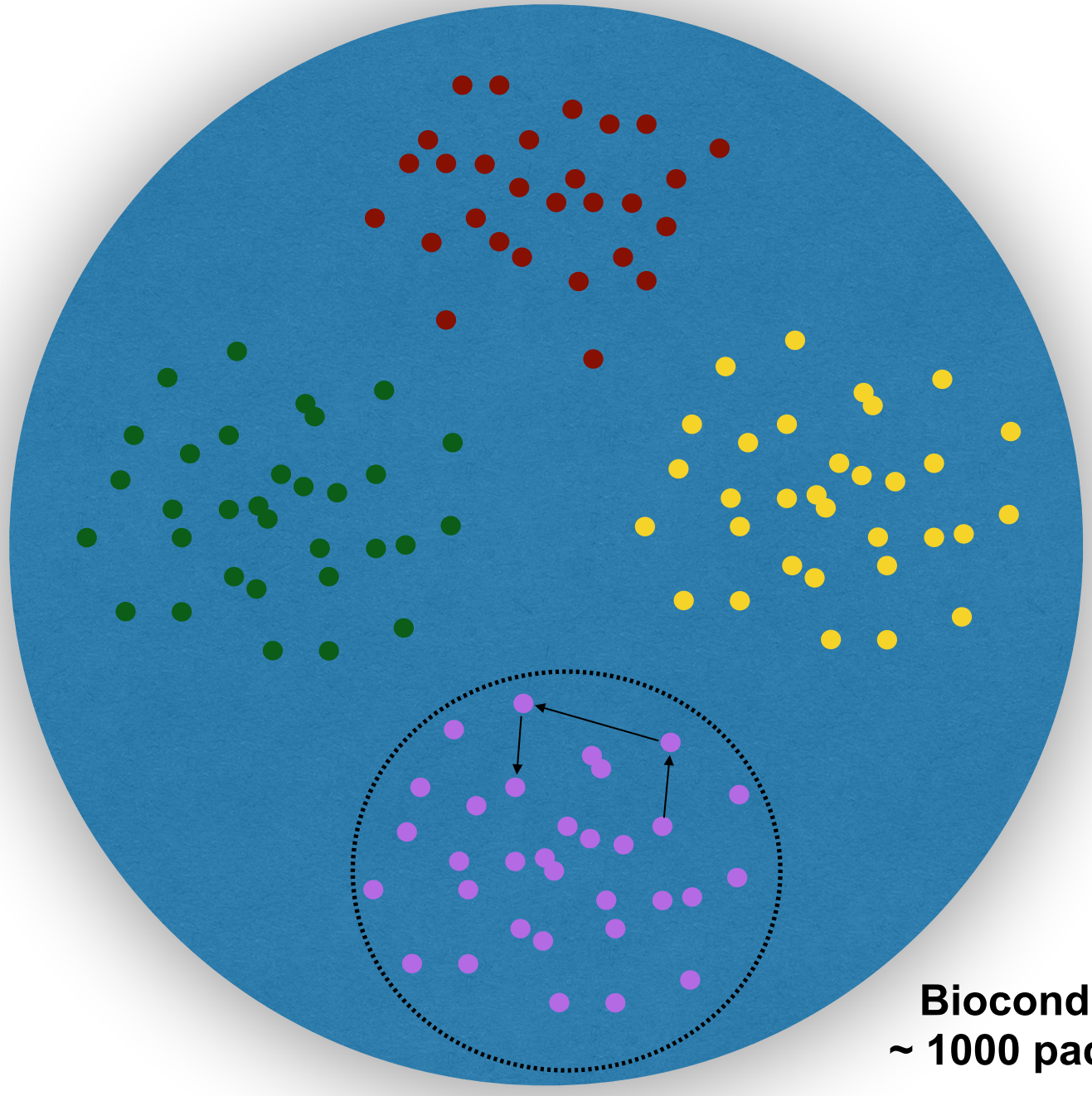
NATURE METHODS | VOL.12 NO.2 | FEBRUARY 2015 | 115

"We have embraced R for its scientific and statistical computing capabilities, for its graphics facilities and for the convenience of an interpreted language. R also interfaces with low-level languages including C and C++ for computationally intensive operations, Java for integration with enterprise software and JavaScript for interactive web-based applications and reports."

Packages are modular and can work together



Workflows utilize multiple packages to step through a complex process

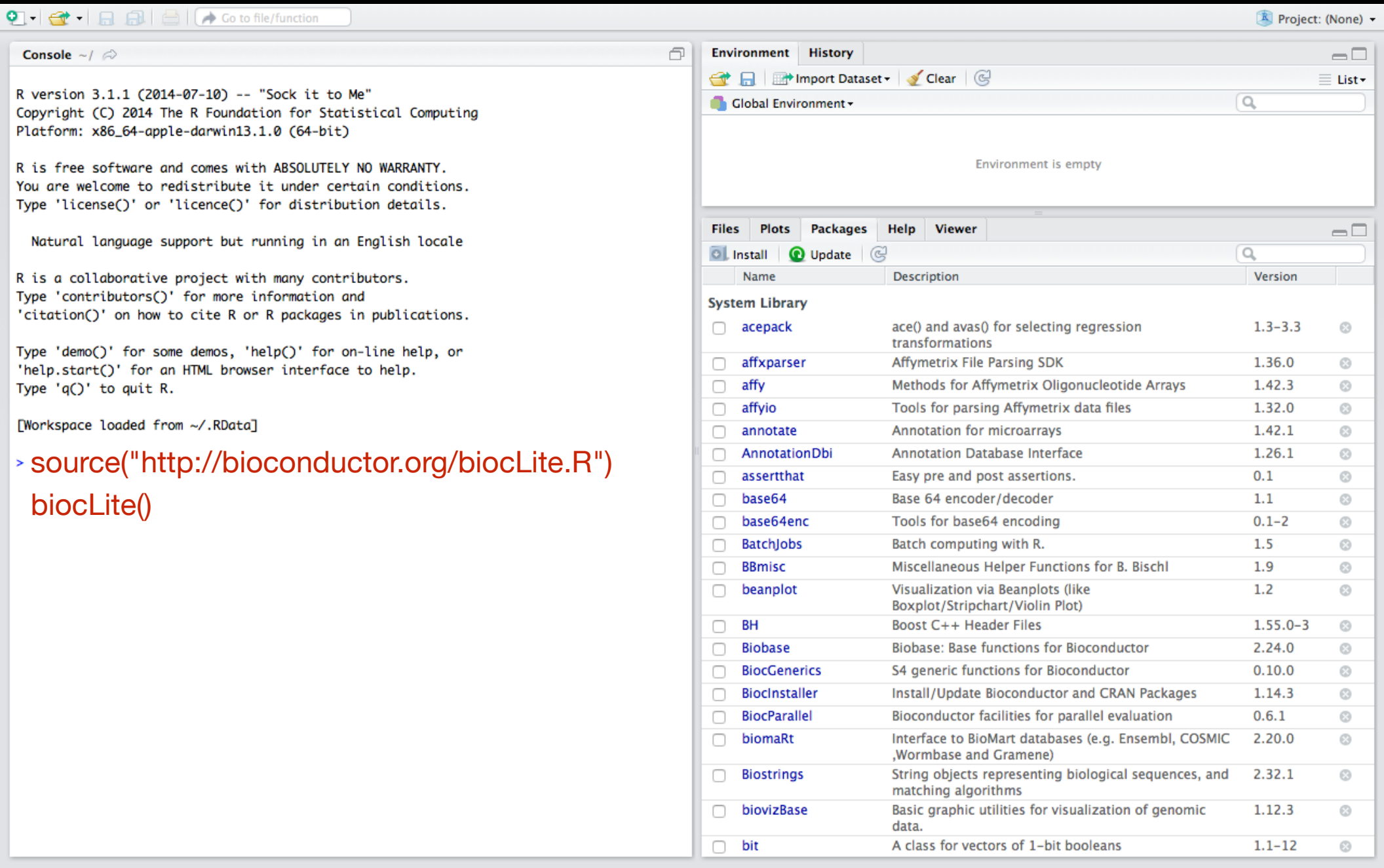


Package type
- statistics
- graphing
- modeling
- bioconductor

~6500 packages

Bioconductor
~ 1000 packages

how do I actually *get* bioconductor?


Go to file/function
Project: (None)
Console ~/
R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.1.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Workspace loaded from ~/.RData]
> source("http://bioconductor.org/biocLite.R")
biocLite()
Environment
History
Import Dataset
Clear
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
System Library
acepack
ace() and avas() for selecting regression transformations
1.3-3.3
affxparser
Affymetrix File Parsing SDK
1.36.0
affy
Methods for Affymetrix Oligonucleotide Arrays
1.42.3
affyio
Tools for parsing Affymetrix data files
1.32.0
annotate
Annotation for microarrays
1.42.1
AnnotationDbi
Annotation Database Interface
1.26.1
assertthat
Easy pre and post assertions.
0.1
base64
Base 64 encoder/decoder
1.1
base64enc
Tools for base64 encoding
0.1-2
BatchJobs
Batch computing with R.
1.5
BBmisc
Miscellaneous Helper Functions for B. Bischl
1.9
beanplot
Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot)
1.2
BH
Boost C++ Header Files
1.55.0-3
Biobase
Biobase: Base functions for Bioconductor
2.24.0
BiocGenerics
S4 generic functions for Bioconductor
0.10.0
BiocInstaller
Install/Update Bioconductor and CRAN Packages
1.14.3
BiocParallel
Bioconductor facilities for parallel evaluation
0.6.1
biomaRt
Interface to BioMart databases (e.g. Ensembl, COSMIC ,Wormbase and Gramene)
2.20.0
Biostrings
String objects representing biological sequences, and matching algorithms
2.32.1
biovizBase
Basic graphic utilities for visualization of genomic data.
1.12.3
bit
A class for vectors of 1-bit booleans
1.1-12

Getting help

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

Home » Help

Bioconductor Release »

Packages in the stable, semi-annual release:

- Software
- Annotation Data (Genome, Array, etc.)
- Experiment Data
- Latest Release Announcement

Bioconductor is also available as an Amazon Machine Image and a series of Docker images.

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

Courses & Conferences »

Bioconductor courses and conferences explore R programming and the analysis of genomic data. View catalog of a recent course material.

- 2015 • 2014 • 2013 • 2012 • 2011
- 2010 • 2009 • 2008 • 2007 • 2006
- 2005 • 2004 • 2003 • 2002

Workflows »

Common Bioconductor workflows include:

- Oligonucleotide Arrays
- High-throughput Sequencing
- Counting Reads for Differential Expression (parathyroideSE vignette)
- Annotation
- Annotating Variants
- Annotating Ranges
- Flow Cytometry and other assays
- Candidate Binding Sites for Known Transcription Factors
- Cloud-enabled cis-eQTL search and annotation
- RNA-Seq workflow: gene-level exploratory analysis and differential expression
- Changing genomic coordinate systems with rtracklayer::liftOver
- Mass spectrometry and proteomics data analysis

Community Help Resources »

Community-contributed resources for learning Bioconductor:

- Book and lab from MOOC: PH525x Data Analysis for Genomics.
- Video Lectures by Rafael Irizarry and guests
- R/Bioconductor Notes by Thomas Girke.
- Using Bioconductor to Analyze your 23andme Data by Vince Buffalo.
- Bioconductor Code Search by Itoshi NIKAIDO.
- R/Bioconductor Blog By Sean Davis.

Newsletters »

- April 2014
- July 2014
- October 2014
- January 2015

http://www.bioconductor.org/help/

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

Home » Help

Bioconductor Release »

Packages in the stable, semi-annual release:

- Software
- Annotation Data (Genome, Array, etc.)
- Experiment Data
- Latest Release Announcement

Bioconductor is also available as an Amazon Machine Image and a series of Docker images.

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

Courses & Conferences »

Bioconductor courses and conferences explore R programming and the analysis of genomic data. View catalog of a recent course material.

- 2015 • 2014 • 2013 • 2012 • 2011
- 2010 • 2009 • 2008 • 2007 • 2006
- 2005 • 2004 • 2003 • 2002

Workflows »

Common Bioconductor workflows include:

- Oligonucleotide Arrays
- High-throughput Sequencing
- Counting Reads for Differential Expression (parathyroideSE vignette)
- Annotation
- Annotating Variants
- Annotating Ranges
- Flow Cytometry and other assays
- Candidate Binding Sites for Known Transcription Factors
- Cloud-enabled cis-eQTL search and annotation
- RNA-Seq workflow: gene-level exploratory analysis and differential expression
- Changing genomic coordinate systems with rtracklayer::liftOver
- Mass spectrometry and proteomics data analysis

Community Help Resources »

Community-contributed resources for learning Bioconductor:

- Book and lab from MOOC: PH525x Data Analysis for Genomics.
- Video Lectures by Rafael Irizarry and guests
- R/Bioconductor Notes by Thomas Girke.
- Using Bioconductor to Analyze your 23andme Data by Vince Buffalo.
- Bioconductor Code Search by Itoshi NIKAIDO.
- R/Bioconductor Blog By Sean Davis.

Newsletters »

- April 2014
- July 2014
- October 2014
- January 2015

http://www.bioconductor.org/help/

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

Home » Help

Bioconductor Release »

Packages in the stable, semi-annual release:

- Software
- Annotation Data (Genome, Array, etc.)
- Experiment Data
- Latest Release Announcement

Bioconductor is also available as an Amazon Machine Image and a series of Docker images.

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

Courses & Conferences »

Bioconductor courses and conferences explore R programming and the analysis of genomic data. View catalog of a recent course material.

- 2015 • 2014 • 2013 • 2012 • 2011
- 2010 • 2009 • 2008 • 2007 • 2006
- 2005 • 2004 • 2003 • 2002

Workflows »

Common Bioconductor workflows include:

- Oligonucleotide Arrays
- High-throughput Sequencing
- Counting Reads for Differential Expression (parathyroideSE vignette)
- Annotation
- Annotating Variants
- Annotating Ranges
- Flow Cytometry and other assays
- Candidate Binding Sites for Known Transcription Factors
- Cloud-enabled cis-eQTL search and annotation
- RNA-Seq workflow: gene-level exploratory analysis and differential expression
- Changing genomic coordinate systems with rtracklayer::liftOver
- Mass spectrometry and proteomics data analysis

Community Help Resources »

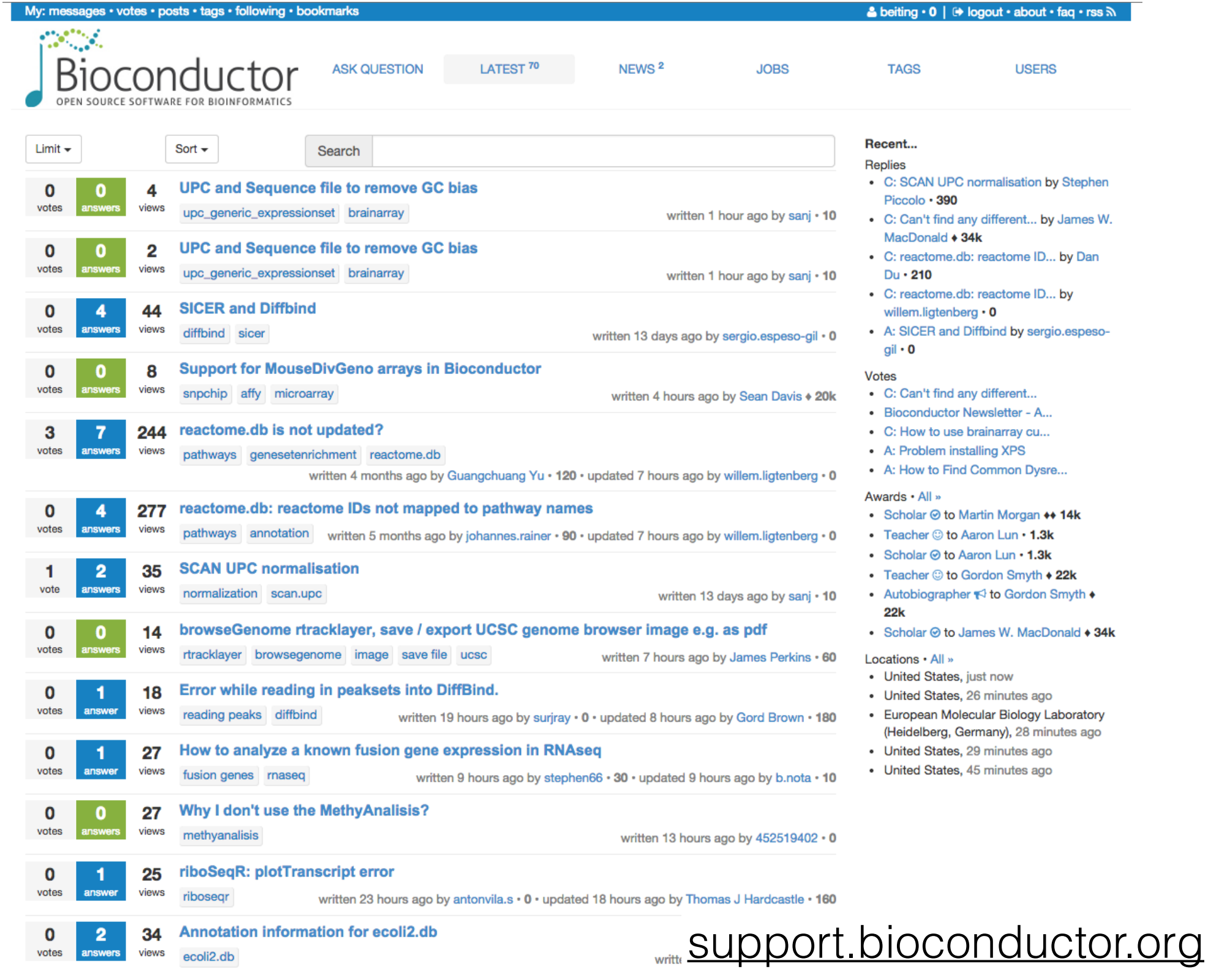
Community-contributed resources for learning Bioconductor:

- Book and lab from MOOC: PH525x Data Analysis for Genomics.
- Video Lectures by Rafael Irizarry and guests
- R/Bioconductor Notes by Thomas Girke.
- Using Bioconductor to Analyze your 23andme Data by Vince Buffalo.
- Bioconductor Code Search by Itoshi NIKAIDO.
- R/Bioconductor Blog By Sean Davis.

Newsletters »

- April 2014
- July 2014
- October 2014
- January 2015

http://www.bioconductor.org/help/


My: messages • votes • posts • tags • following • bookmarks
beiting • 0 | logout • about • faq • rss
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
ASK QUESTION
LATEST 70
NEWS 2
JOBS
TAGS
USERS
Limit
Sort
Search
0 votes 0 answers 4 views
UPC and Sequence file to remove GC bias
upc_generic_expressionset brainarray
written 1 hour ago by sanj • 10
0 votes 0 answers 2 views
UPC and Sequence file to remove GC bias
upc_generic_expressionset brainarray
written 1 hour ago by sanj • 10
0 votes 4 answers 44 views
SICER and Diffbind
diffbind sicer
written 13 days ago by sergio.espeso-gil • 0
0 votes 0 answers 8 views
Support for MouseDivGeno arrays in Bioconductor
snpchip affy microarray
written 4 hours ago by Sean Davis ♦ 20k
3 votes 7 answers 244 views
reactome.db is not updated?
pathways genesetenrichment reactome.db
written 4 months ago by Guangchuang Yu • 120 • updated 7 hours ago by willem.ligtenberg • 0
0 votes 4 answers 277 views
reactome.db: reactome IDs not mapped to pathway names
pathways annotation
written 5 months ago by johannes.rainer • 90 • updated 7 hours ago by willem.ligtenberg • 0
1 vote 2 answers 35 views
SCAN UPC normalisation
normalization scan.upc
written 13 days ago by sanj • 10
0 votes 0 answers 14 views
browseGenome rtracklayer, save / export UCSC genome browser image e.g. as pdf
rtracklayer browsegenome image save file ucsc
written 7 hours ago by James Perkins • 60
0 votes 1 answer 18 views
Error while reading in peaksets into DiffBind.
reading peaks diffbind
written 19 hours ago by surjray • 0 • updated 8 hours ago by Gord Brown • 180
0 votes 1 answer 27 views
How to analyze a known fusion gene expression in RNAseq
fusion genes rnaseq
written 9 hours ago by stephen66 • 30 • updated 9 hours ago by b.nota • 10
0 votes 0 answers 27 views
Why I don't use the MethyAnalisis?
methyanalisis
written 13 hours ago by 452519402 • 0
0 votes 1 answer 25 views
riboSeqR: plotTranscript error
riboseqr
written 23 hours ago by antonvila.s • 0 • updated 18 hours ago by Thomas J Hardcastle • 160
0 votes 2 answers 34 views
Annotation information for ecoli2.db
ecoli2.db
Recent...
Replies
C: SCAN UPC normalisation by Stephen Piccolo • 390
C: Can't find any different... by James W. MacDonald ♦ 34k
C: reactome.db: reactome ID... by Dan Du • 210
C: reactome.db: reactome ID... by willem.ligtenberg • 0
A: SICER and Diffbind by sergio.espeso-gil • 0
Votes
C: Can't find any different...
Bioconductor Newsletter - A...
C: How to use brainarray cu...
A: Problem installing XPS
A: How to Find Common Dysre...
Awards • All »
Scholar to Martin Morgan ♦♦ 14k
Teacher to Aaron Lun • 1.3k
Scholar to Aaron Lun • 1.3k
Teacher to Gordon Smyth ♦ 22k
Autobiographer to Gordon Smyth ♦ 22k
Scholar to James W. MacDonald ♦ 34k
Locations • All »
United States, just now
United States, 26 minutes ago
European Molecular Biology Laboratory (Heidelberg, Germany), 28 minutes ago
United States, 29 minutes ago
United States, 45 minutes ago
support.bioconductor.org

limma

Linear Models for Microarray Data

Bioconductor version: Release (3.0)

Data analysis, linear models and differential expression for microarray data.

Author: Gordon Smyth [cre,aut], Matthew Ritchie [ctb], Jeremy Silver [ctb], James Wettenhall [ctb], Natalie Thorne [ctb], Davis McCarthy [ctb], Di Wu [ctb], Yifang Hu [ctb], Wei Shi [ctb], Belinda Phipson [ctb], Alicia Oshlack [ctb], Carolyn de Graaf [ctb], Mette Langaas [ctb], Egil Ferkingstad [ctb], Marcus Davy [ctb], Francois Pepin [ctb], Dongseok Choi [ctb]

Maintainer: Gordon Smyth <smyth at wehi.edu.au>

Citation (from within R, enter `citation("limma")`):

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W and Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research*, **43**, pp. doi: 10.1093/nar/gkv007.

Installation

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("limma")
```

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("limma")
```

PDF	R Script	Limma One Page Introduction
PDF		usersguide.pdf
PDF		Reference Manual
Text		NEWS

Workflows »

Common Bioconductor workflows include:

- Oligonucleotide Arrays
- High-throughput Sequencing
- Counting Reads for Differential Expression (parathyroideSE vignette)
- Annotation
- Annotating Variants
- Annotating Ranges
- Flow Cytometry and other assays
- Candidate Binding Sites for Known Transcription Factors
- Cloud-enabled cis-eQTL search and annotation
- RNA-Seq workflow: gene-level exploratory analysis and differential expression
- Changing genomic coordinate systems with rtracklayer::liftOver
- Mass spectrometry and proteomics data analysis

Mailing Lists »

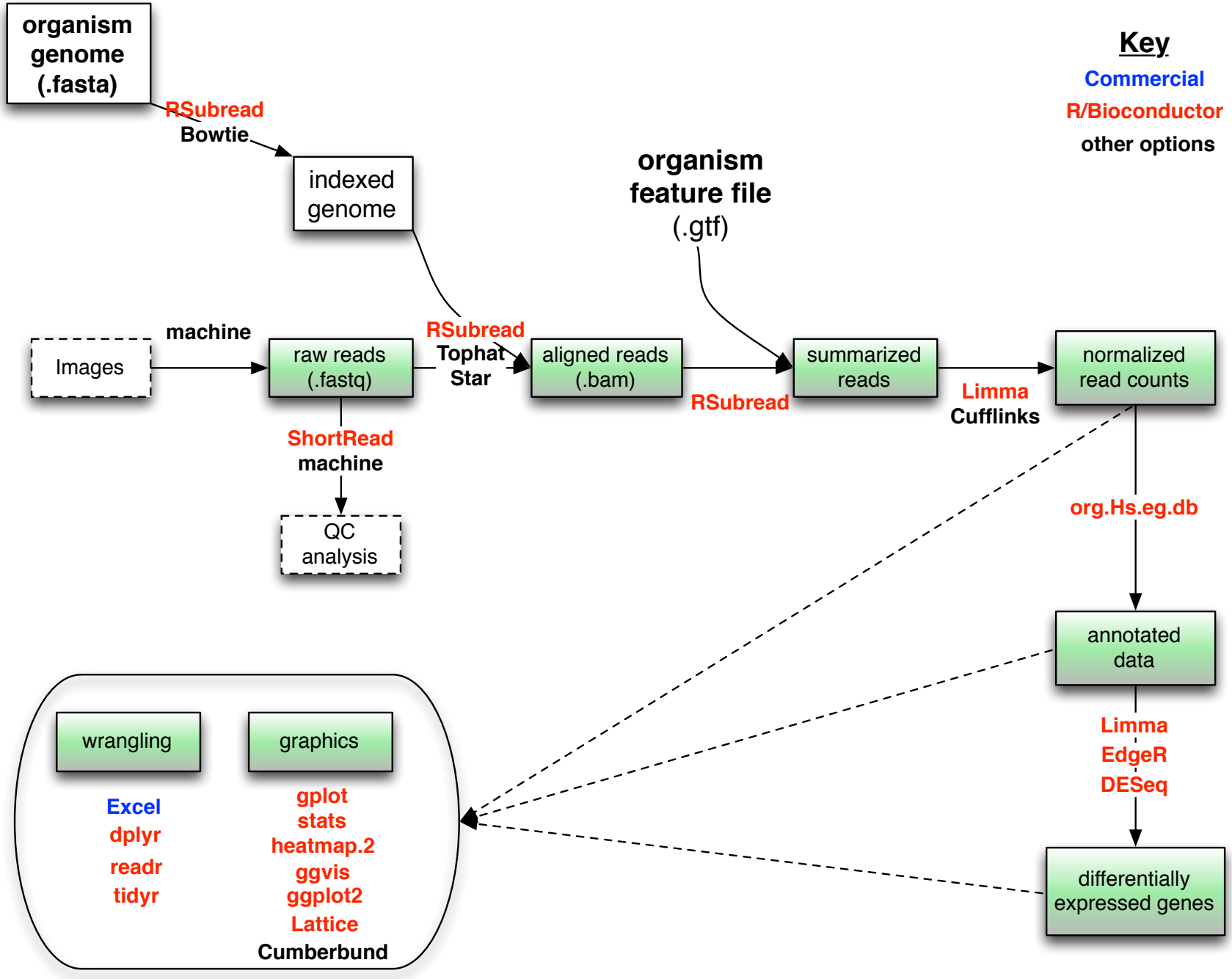
Post questions about Bioconductor packages to our mailing lists. Read the posting guide before posting!

- bioconductor
- bioc-devel

help for individual packages

How is R/Bioconductor used to analyze RNAseq data?

RNAseq analysis workflow



RNAseq: Subread aligner + FeatureCounts

The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote

Yang Liao[1,2], Gordon K. Smyth[1,3] and Wei Shi[1,2,*]

Nucleic Acids Research, 2013, Vol. 41, No. 10 ***e108***
doi:10.1093/nar/gkt214

[1]Division of Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, [2]Department of Computing and Information Systems, The University of Melbourne, Parkville, Victoria 3010, Australia and [3]Department of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia

BIOINFORMATICS **ORIGINAL PAPER**

Vol. 30 no. 7 2014, pages 923–930
doi:10.1093/bioinformatics/btt656

Sequence analysis

Advance Access publication November 13, 2013

featureCounts: an efficient general purpose program for assigning sequence reads to genomic features

Yang Liao[1,2], Gordon K. Smyth[1,3] and Wei Shi[1,2,*]

[1]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, [2]Department of Computing and Information Systems and [3]Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia

Associate Editor: Martin Bishop

What you need to participate…

There are four main things you will need in order to particpate in the hands-on aspects of the workshop

1. Your laptop computer

- Everyone is encouraged to bring their own internet-enabled laptop equiped with either the recent Mac or Windows OS.

2. A "toolbox" of free software

- Download and install the appropriate version of the R Programming Language for your operating system
- Download and install the graphical user interface for R, called RStudio
- Download a text editor. I use TextWrangler (Mac only) and Sublime (Mac and PC)

3. Sample dataset

- We will be analyzing RNAseq data from this 2014 PLOS One published by Kami Kim's lab.
- The original raw data is available on the Short Read Archive here, but you **do not** need to download this (files are huge). Instead, please download the 'digital gene expression list' (DGEList) from here
- In addition to the data, please download this simple text file that describes the design of the study

4. An analysis script

- the R script we'll use in the workshop can be downloaded here.

ToxoGenomicsWorkshop.github.io

OPEN ACCESS Freely available online

PLOS ONE

Distinct Strains of *Toxoplasma gondii* Feature Divergent Transcriptomes Regardless of Developmental Stage

Matthew McKnight Croken[1], Yanfen Ma[2], Lye Meng Markillie[3], Ronald C. Taylor[4], Galya Orr[3], Louis M. Weiss[2,5]*, Kami Kim[1,2,5]*

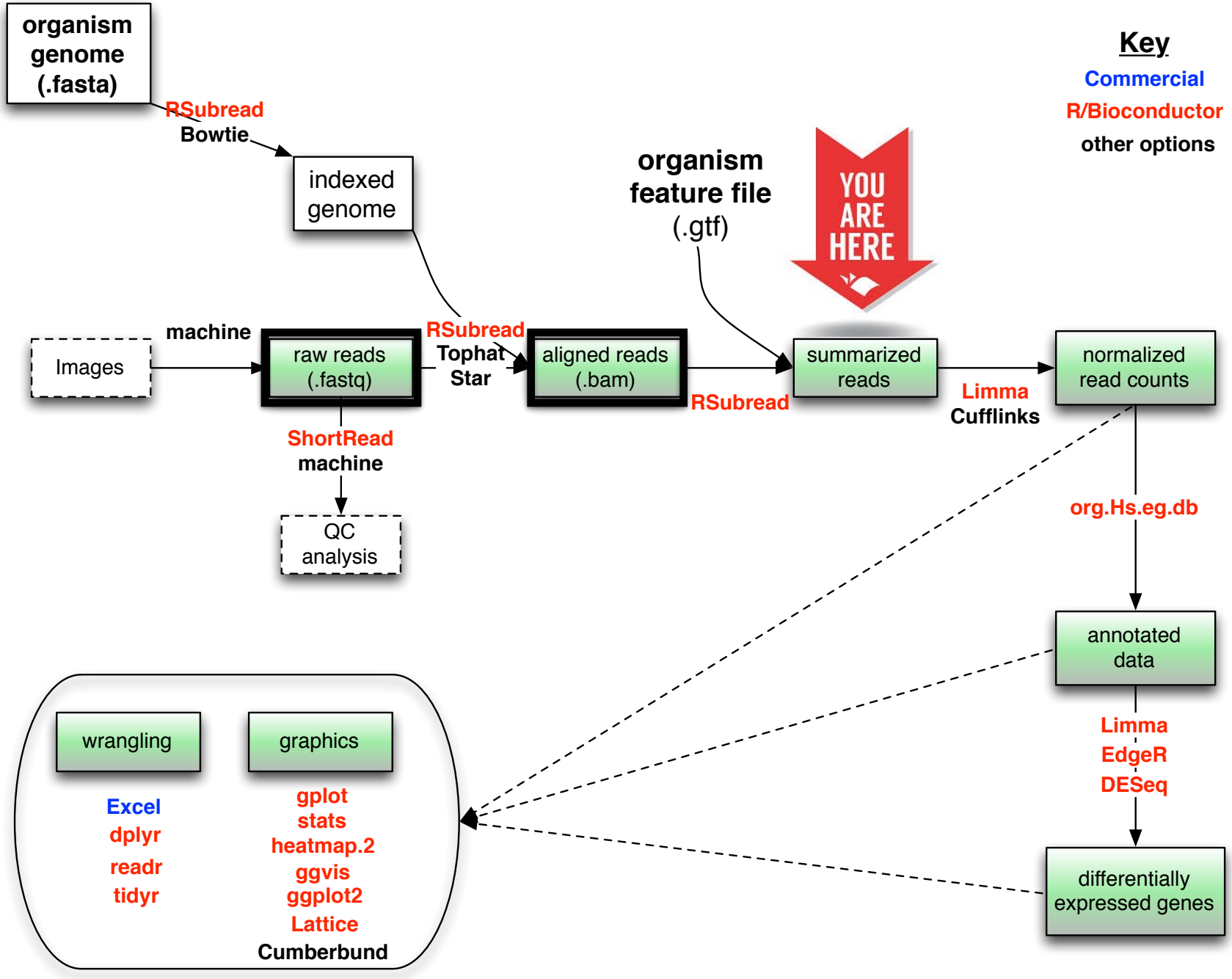
s of America, **2** Department of Pathology, Albert
, Pacific Northwest National Laboratory, Richland,
, Pacific Northwest National Laboratory, Richland,
rk, United States of America

studyDesign.txt

studyDesign.txt

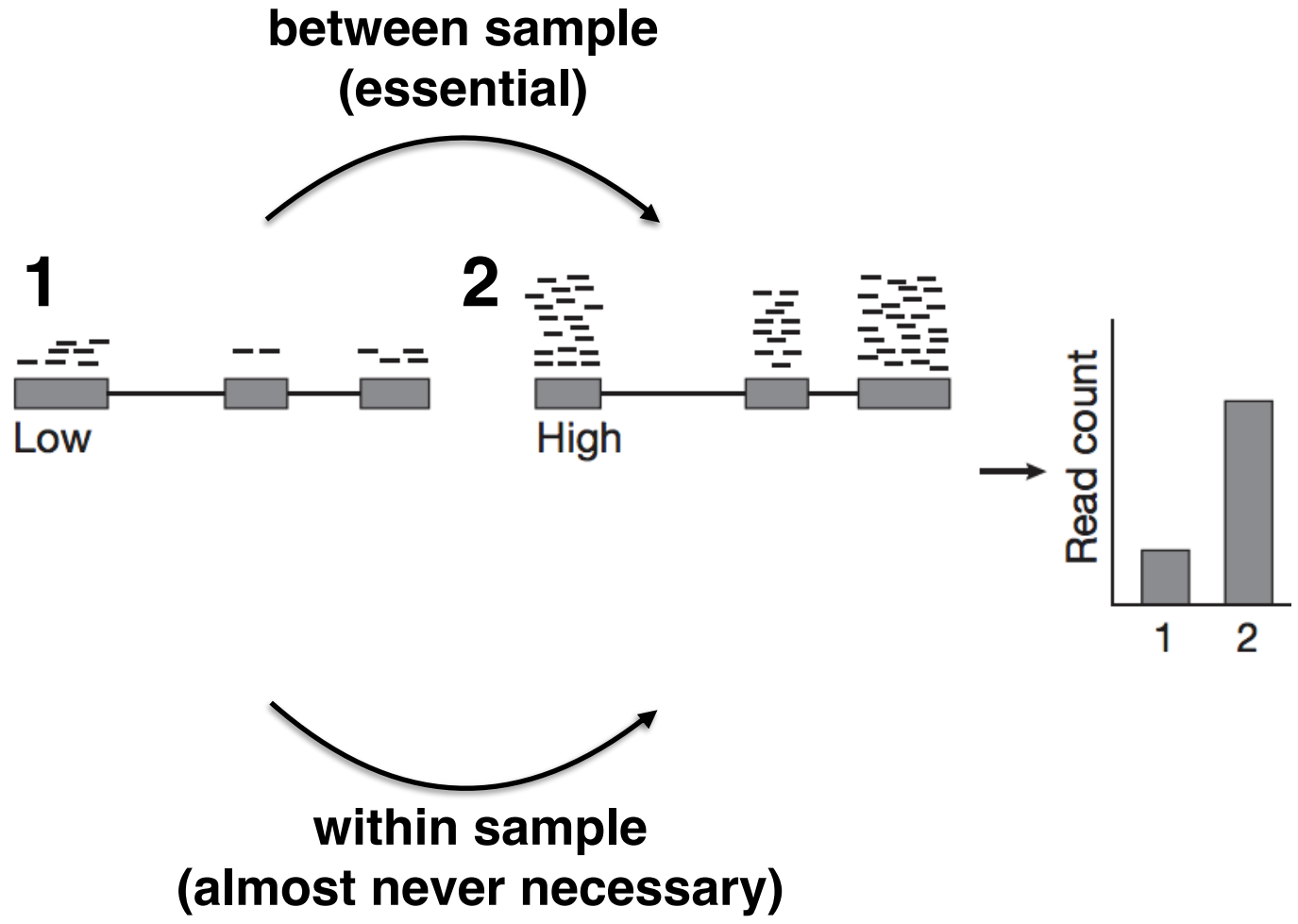
	fastq	strain	culture	stage	replicate	output
2	SRR1542919.fastq.gz	RH	pH7	tachy	rep1	RH.pH7.rep1.bam
3	SRR1542920.fastq.gz	RH	pH7	tachy	rep2	RH.pH7.rep2.bam
4	SRR1542921.fastq.gz	RH	pH7	tachy	rep3	RH.pH7.rep3.bam
5	SRR1542922.fastq.gz	RH	pH8	brady	rep1	RH.pH8.rep1.bam
6	SRR1542923.fastq.gz	RH	pH8	brady	rep2	RH.pH8.rep2.bam
7	SRR1542924.fastq.gz	RH	pH8	brady	rep3	RH.pH8.rep3.bam
8	SRR1542925.fastq.gz	PLK	pH7	tachy	rep1	PLK.pH7.rep1.bam
9	SRR1542926.fastq.gz	PLK	pH7	tachy	rep2	PLK.pH7.rep2.bam
10	SRR1542927.fastq.gz	PLK	pH7	tachy	rep3	PLK.pH7.rep3.bam
11	SRR1542928.fastq.gz	PLK	pH8	brady	rep1	PLK.pH8.rep1.bam
12	SRR1542929.fastq.gz	PLK	pH8	brady	rep2	PLK.pH8.rep2.bam
13	SRR1542930.fastq.gz	PLK	pH8	brady	rep3	PLK.pH8.rep3.bam
14	SRR1542931.fastq.gz	CTG	pH7	tachy	rep1	CTG.pH7.rep1.bam
15	SRR1542932.fastq.gz	CTG	pH7	tachy	rep2	CTG.pH7.rep2.bam
16	SRR1542933.fastq.gz	CTG	pH7	tachy	rep3	CTG.pH7.rep3.bam
17	SRR1542934.fastq.gz	CTG	pH8	brady	rep1	CTG.pH8.rep1.bam
18	SRR1542935.fastq.gz	CTG	pH8	brady	rep2	CTG.pH8.rep2.bam
19	SRR1542936.fastq.gz	CTG	pH8	brady	rep3	CTG.pH8.rep3.bam
20						

RNAseq analysis workflow

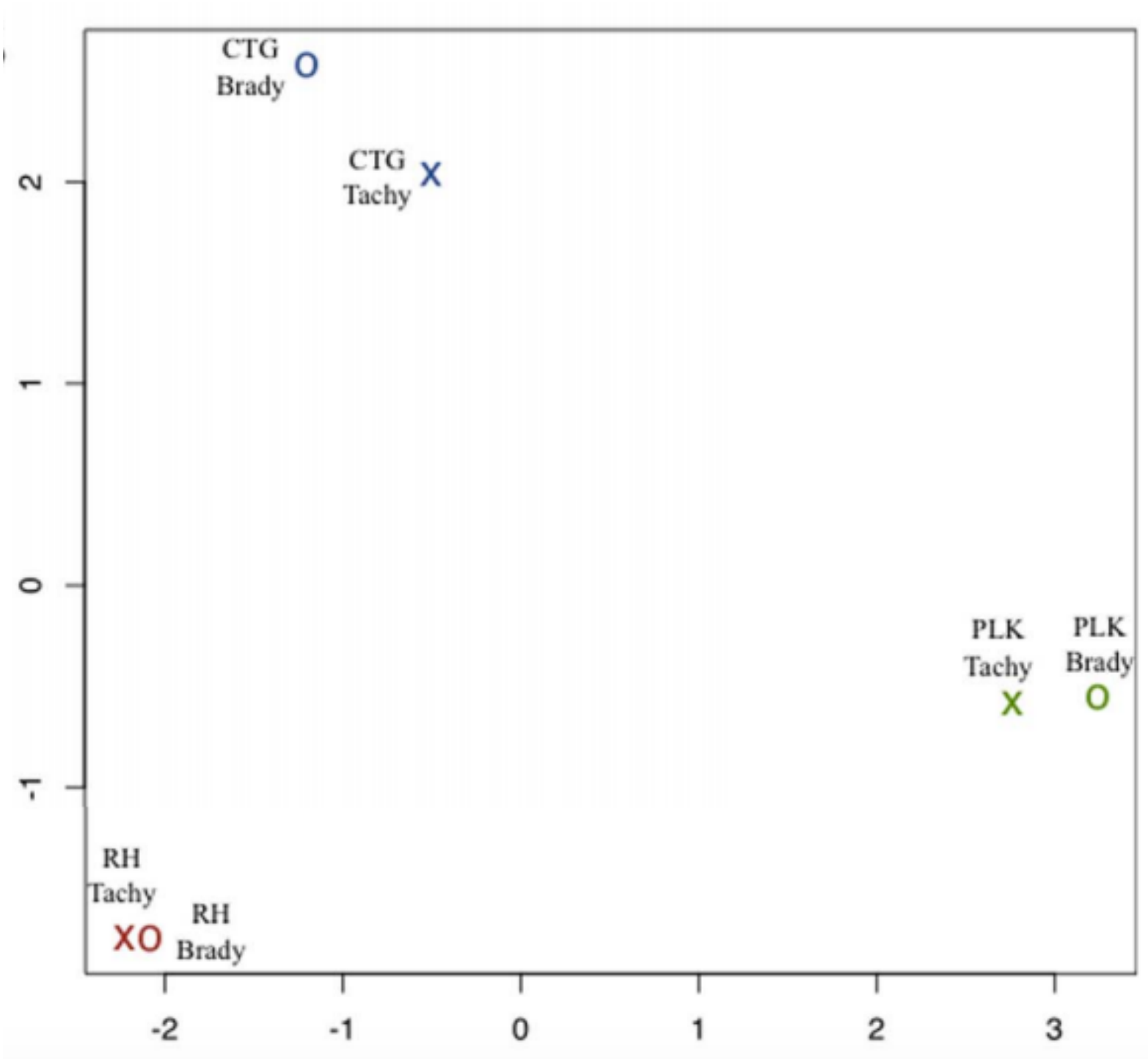


Demo analysis

Normalizing RNAseq data



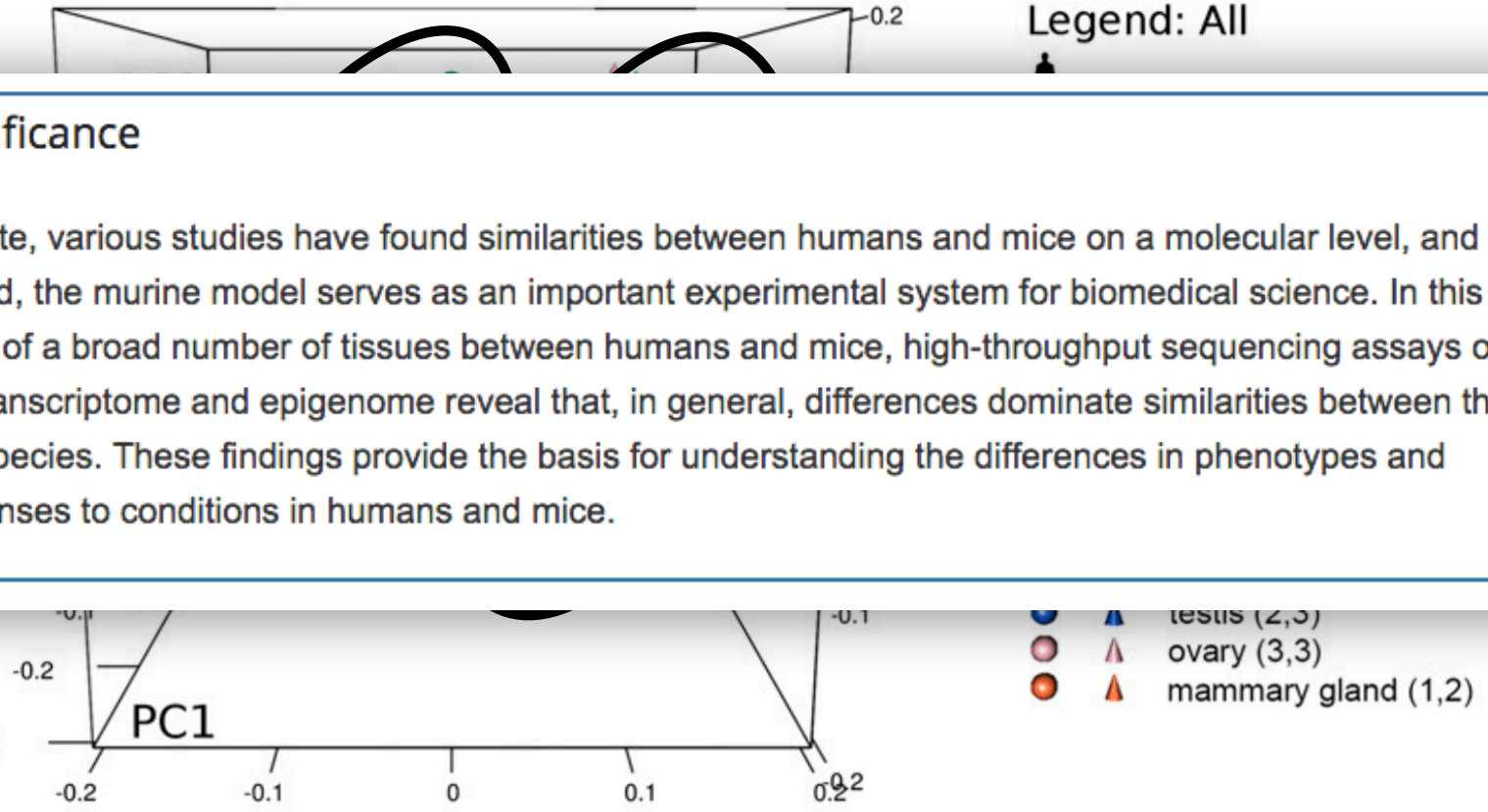
modified from Garber et al., Nature Methods, 2011


CTG
Brady
CTG
Tachy
PLK
Tachy
PLK
Brady
RH
Tachy
RH
Brady
2
1
0
-1
-2
-1
0
1
2
3

Comparison of the transcriptional landscapes between human and mouse tissues

Shin Lin[a,b,1], Yiing Lin[c,1], Joseph R. Nery[d], Mark A. Urich[d], Alessandra Breschi[e,f], Carrie A. Davis[g], Alexander Dobin[g], Christopher Zaleski[g], Michael A. Beer[h], William C. Chapman[c], Thomas R. Gingeras[g,i], Joseph R. Ecker[d,j,2], and Michael P. Snyder[a,2]

[a]Department of Genetics, Stanford University, Stanford, CA 94305; [b]Division of Cardiovascular Medicine, Stanford University, Stanford, CA 94305; [c]Department of Surgery, Washington University School of Medicine, St. Louis, MO 63110; [d]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037; [e]Centre for Genomic Regulation and UPF, Catalonia, 08003 Barcelona, Spain; [f]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Spain; [g]Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11742; [h]McKusick-Nathans Institute of Genetic Medicine and the Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205; [i]Affymetrix, Inc., Santa Clara, CA 95051; and [j]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037



Significance

To date, various studies have found similarities between humans and mice on a molecular level, and indeed, the murine model serves as an important experimental system for biomedical science. In this study of a broad number of tissues between humans and mice, high-throughput sequencing assays on the transcriptome and epigenome reveal that, in general, differences dominate similarities between the two species. These findings provide the basis for understanding the differences in phenotypes and responses to conditions in humans and mice.

RESEARCH ARTICLE

A reanalysis of mouse ENCODE comparative gene expression data [v1; ref status: awaiting peer review, http://f1000r.es/5ez]

Yoav Gilad, Orna Mizrahi-Man

Department of Human Genetics, University of Chicago, Chicago, IL, 60637, USA

"The implications of the observation that human and mouse gene expression data may be clustering by species more than by tissues can be profound. To a large degree, modern biology is built upon the empirical observation that homologous gene regulatory networks establish the identities of homologous cell-types, tissues, and organs across species – the results of Lin *et al.*, if true, challenge these observations and the biological basis of homology. From a more practical perspective, the mouse is arguably the most important animal model for biomedical research. If gene regulation in any mouse tissue is markedly more representative of a general mouse regulatory network than the regulatory network of a corresponding human tissue, this would call into question the utility of the mouse, and perhaps any other non-human animal, as a useful model system for biomedical research."

@<machine_id>:<runnumber>:<flowcellID>:<lane>:<tile>:<x-pos>: <y-pos> <read>:<is filtered>:<control number>:<index sequence>

@D4LHBFN1:276:C2HKJACXX:4:1101:3448:12374 1:N:0:AGTTCC

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

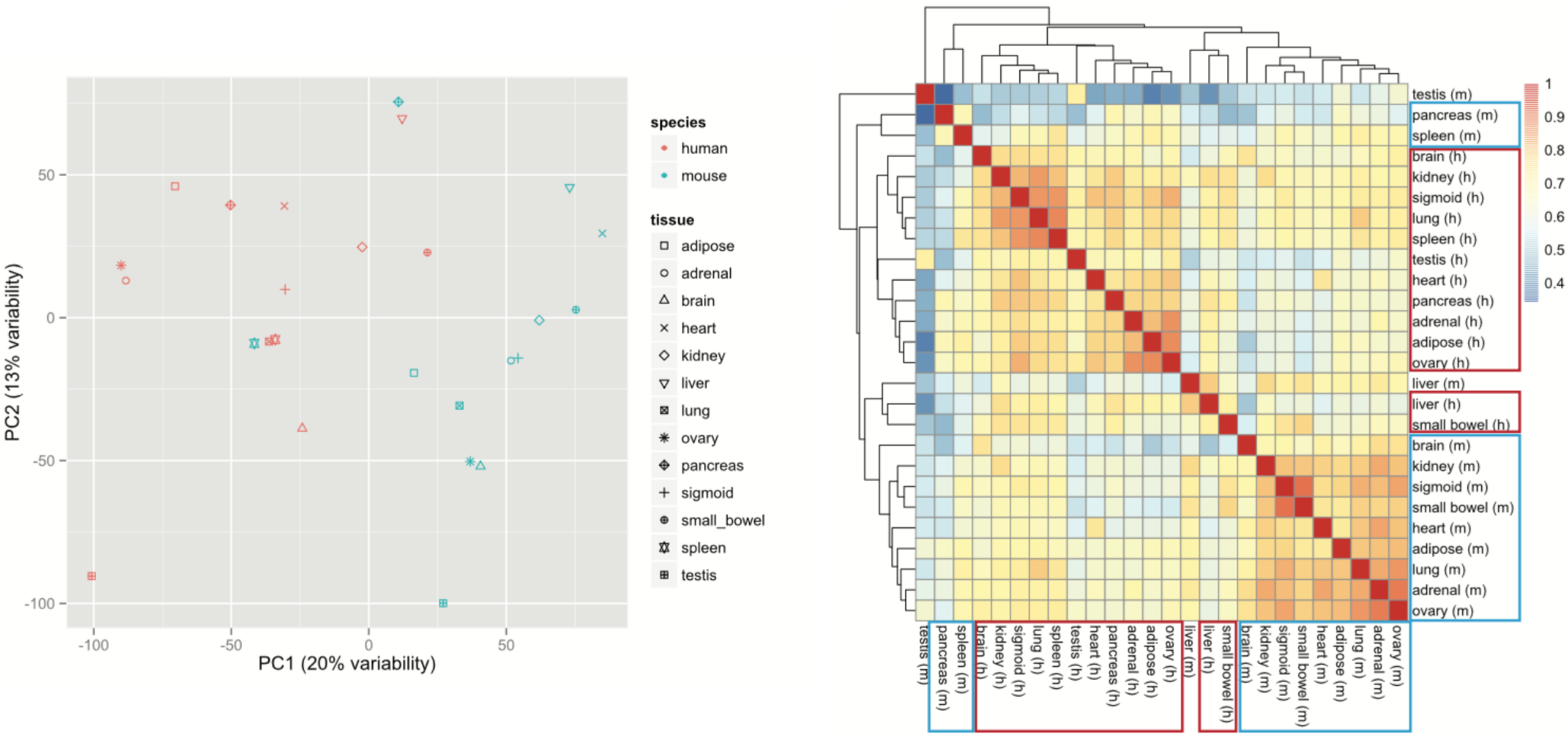


Figure 2. Recapitulating the patterns reported by the mouse ENCODE papers. a. Two-dimensional plots of principal components calculated by performing PCA of the transposed log-transformed FPKM values (from 14,744 orthologous gene pairs) for the 26 samples, after removal of invariant columns (genes). **b**. Heatmap based on pairwise Pearson correlation of expression data used in panel **a**. We used Euclidean distance and complete linkage as distance measure and clustering method, respectively.

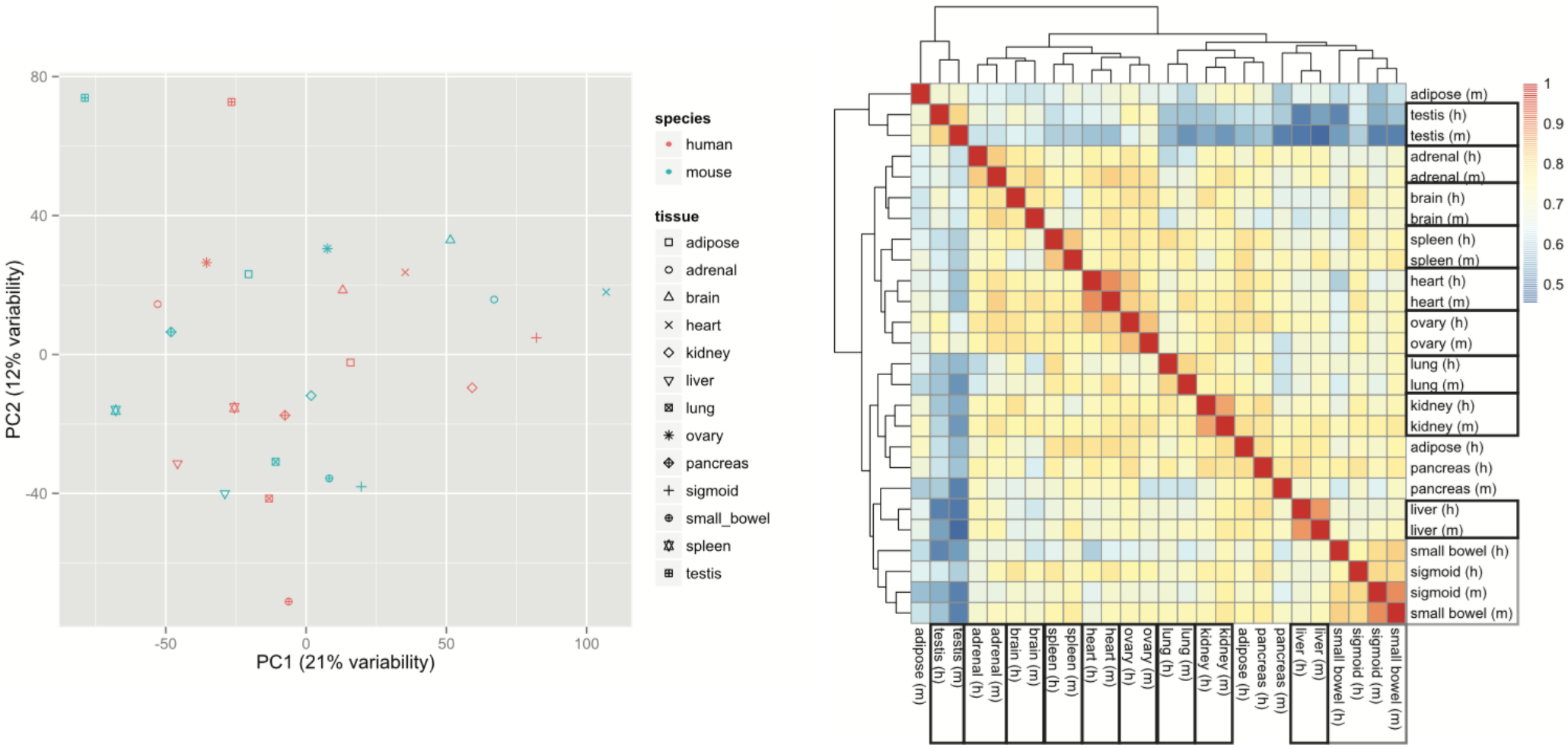


Figure 3. Clustering of data once batch effects are accounted for. a. Two-dimensional plots of principal components calculated by applying PCA to the transposed matrix of batch-corrected log-transformed normalized fragment counts (from 10,309 orthologous gene pairs that remained after the exclusion steps described in the results) for the 26 samples, after removal of invariant columns (genes). **b**. Heatmap based on pairwise Pearson correlation of the expression data used in panel **a**. We used Euclidean distance and complete linkage as distance measure and clustering method, respectively.

Discuss this Article

Reader Comment 20 May 2015

Shin Lin, Department of Genetics, Stanford University, USA

"There remains the issue of our study design with respect to confounding of lane effect and species. It should be noted that our study design minimized library preparation and primer index effect. Although our experience is consistent with [reports] showing that lane effect is not a large contributor to variance, we recognize it is better to have data. Thus, we are sequencing under a new pattern of pooled libraries, and soon, we will post the results for the community." - *authors of PNAS study*

Author Response 20 May 2015

Yoav Gilad, Human Genetics, University of Chicago, USA

Why are you calling these 'lane effects'? The samples were sequenced on different instruments (different sequencers) and by default - different flow cells. Can you explicitly acknowledge that fact please?

Data wrangling with the dplyr package

4 essential 'verbs' encompass
most of the functionality of dplyr

1. **Mutate -** add new columns as a function of existing data
2. **Filter -** picks rows based on rules you apply
3. **Select -** picks columns based on rules you apply
4. **Arrange -** reorders rows by sorting based on column(s)