High-throughput sequence analysis with *R* and *Bioconductor*

Martin Morgan[*], Marc Carlson[†], Valerie Obenchain[‡], Dan Tenenbaum[§], Hervé Pagès[¶]

August 17, 2011

Contents

[*]mtmorgan@fhcrc.org
[†]mcarlson@fhcrc.org
[‡]vobencha@fhcrc.org
[§]dtenenba@fhcrc.org
[¶]hpages@fhcrc.org

1 Introduction

1.1 *Bioconductor*

Bioconductor is a collection of *R* packages for the analysis and comprehension of high-throughput genomic data. *Bioconductor* started more than 10 years ago. It gained credibility for its statistically rigorous approach to microarray pre-preprocessing and designed experiments, and integrative and reproducible approaches to bioinformatic tasks. There are now more than 500 *Bioconductor* packages for expression and other microarrays, sequence analysis, flow cytometry, imaging, and other domains. The *Bioconductor* web site provides installation, package repository, help, and other documentation.

1.2 High-throughput sequence analysis

Recent technological developments introduce high-throughput sequencing approaches. A variety of experimental protocols and analysis work flows address gene expression, regulation, and encoding of genic variants. Experimental protocols produce a large number (millions per sample) of short (e.g., 35-100, single or paired-end) nucleotide sequences. These are aligned to a reference or other genome. Analysis work flows use the alignments to infer levels of gene expression (RNA-seq), binding of regulatory elements to genomic locations (ChIP-seq), or prevalence of structural variants (e.g., SNPs, short indels, large-scale genomic rearrangements). Sample sizes range from minimal replication (e.g,. 2 samples per treatment group) to thousands of individuals.

1.3 This workshop

This workshop introduces use of *R* and *Bioconductor* for analysis of high-throughput sequence data. The workshop is structured as a series of short remarks followed by group exercises. The exercises explore the diversity of tasks for which *R* / *Bioconductor* are appropriate for, but are far from comprehensive.

The goals of the workshop are to: (1) develop familiarity with *R* / *Bioconductor* software for high-throughput analysis; (2) expose key statistical issues in the analysis of sequence data; and (3) provide inspiration and a framework for further independent exploration.

2 *R* and *Bioconductor*

R is an open-source statistical programming language. It is used to manipulate data, to perform statistical analyses, and to present graphical and other results. *R* consists of a core language, additional 'packages' distributed with the *R* language, and a very large number of packages contributed by the broader community. Packages add specific functionality to an *R* installation. *R* has become the primary language of academic statistical analyses, and is widely used in diverse areas of research, government, and industry.

R has several unique features. It has a surprisingly 'old school' interface: users type commands into a console; scripts in plain text represent work flows; tools other than *R* are used for editing and other tasks. *R* is a flexible programming language, so while one person might use functions provided by *R* to accomplish advanced analytic tasks, another might implement their own functions for novel data types. As a programming language, *R* adopts syntax and grammar that differ from many other languages: objects in *R* are 'vectors', and functions are 'vectorized' to operate on all elements of the object; *R* objects have 'copy on change' and 'pass by value' semantics, reducing unexpected consequences for users at the expense of less efficient memory use; common paradigms in other languages, such as the 'for' loop, are encountered much less commonly in *R*. Many authors contribute to *R* so there can be a frustrating inconsistency of documentation and interface. *R* grew up in the academic community, so authors have not shied away from trying new approaches. Of course common statistical analyses, especially exploratory, are very well-developed.

2.1 Statistical programming

Many academic and commercial software products are available; why would one use *R* and *Bioconductor*? One answer is to ask what demands high-throughput genomic data place on the effectiveness of computational biology software.

Effective computational biology software High-throughput questions make use of large data sets. This applies both to the primary data (microarray expression values, sequenced reads, etc.) and also to the annotations on those data (coordinates of genes and features such as exons or regulatory regions; participation in biological pathways, etc.). Large data sets place demands on our tools that preclude some standard approaches, such as spread sheets. Likewise, intricate relationships between data and annotation, and the diversity of research questions, require flexibility typical of a programming language rather than a narrowly-enabled graphical user interface.

Analysis of high-throughput data is necessarily statistical. The volume of data requires that it be appropriately summarized before any sort of comprehension is possible. The data are produced by advanced technologies, and these introduce artifacts (e.g., probe-specific bias in microarrays; sequence or base calling bias in RNA-seq experiments) that need to be accommodated to avoid incorrect or inefficient inference. Data sets typically derive from designed experiments, requiring a statistical approach both to account for the design, and to correctly address the large number of observed values (e.g., gene expression or sequence tag counts) and small number of samples accessible in typical experiments.

Research needs to be reproducible. Reproducibility is both an ideal of the scientific method, and a pragmatic requirement. The latter comes from the long-term and multi-participant nature of contemporary science. An analysis will be performed for the initial experiment, revisited during manuscript preparation, and revisited during reviews or in determining next steps. Likewise, analyses typically involve a team of individuals with diverse domains of expertise. Effective collaborations result when it is easy to reproduce, perhaps with minor modifications, an existing result, and when sophisticated statistical or bioinformatic analyses can be effectively conveyed to other group members.

Science moves very quickly. This is driven by the novel questions that are the hallmark of discovery, and by technological innovation and accessibility. This places significant burdens on software, which must also move quickly. Effective software cannot be too polished, because that requires that the correct analyses are 'known' and that significant resources of time and money have been invested in developing the software; this implies software that is tracking the trailing edge of innovation. On the other hand, leading-edge software cannot be too idiosyncratic; it must be usable by a wider audience than the creator of the software, and fit in with other software relevant to the analysis.

Effective software must be accessible. Affordability is one aspect of accessibility. Another is transparent implementation, where the novel software is sufficiently documented and source code accessible enough for the assumptions, approaches, practical implementation decisions, and inevitable coding errors to be assessed by other skilled practitioners. A final aspect of affordability is that the software is actually usable. This is achieved through adequate documentation, support forums, and training opportunities.

***Bioconductor* as effective computational biology software** What features of *R* and *Bioconductor* contribute to its effectiveness as a software tool?

Bioconductor is well suited to handle extensive data and annotation. *Bioconductor* 'classes' represent high-throughput data and their annotation in an integrated way. *Bioconductor* methods use advanced programming techniques or *R* resources (such as transparent data base or network access) to minimize memory requirements and integrate with diverse resources. Classes and methods coordinate complicated data sets with extensive annotation. Nonetheless, the basic model for object manipulation in *R* involves vectorized in-memory representations. For this reason, particular programming paradigms (e.g., block processing of data streams; explicit parallelism) or hardware resources (e.g., large-memory computers) are sometimes required when dealing with extensive data.

R is ideally suited to addressing the statistical challenges of high-throughput data. Three examples include the development of the 'RMA' and other normalization algorithm for microarray pre-processing, use of moderated t-statistics for assessing microarray differential expression, and development of approaches to estimating dispersion read counts necessary for appropriate analysis of RNAseq designed experiments.

Many of the 'old school' aspects of *R* and *Bioconductor* facilitate reproducible research. An analysis is often represented as a text-based script. Reproducing the analysis involves re-running the script; adjusting how the analysis is performed involves simple text-editing tasks. Beyond this, *R* has the notion of

a 'vignette', which represents an analysis as a LaTeX document with embedded *R* commands. The *R* commands are evaluated when the document is built, thus reproducing the analysis. The use of LaTeX means that the symbolic manipulations in the script are augmented with textual explanations and justifications for the approach taken; these include graphical and tabular summaries at appropriate places in the analysis. *R* includes facilities for reporting the exact version of *R* and associated packages used in an analysis so that, if needed, discrepancies between software versions can be tracked down and their importance evaluated. While users often think of *R* packages as providing new functionality, packages are also used to enhance reproducibility by encapsulating a single analysis. The package can contain data sets, vignette(s) describing the analysis, *R* functions that might have been written, scripts for key data processing stages, and documentation (via standard *R* help mechanisms) of what the functions, data, and packages are about.

The *Bioconductor* project adopts practices that facilitate reproducibility. Versions of *R* and *Bioconductor* are released twice each year. Each *Bioconductor* release is the result of development, in a separate branch, during the previous six months. The release is built daily against the corresponding version of *R* on Linux, Mac, and Windows platforms, with an extensive suite of tests performed. The `biocLite` function ensures that each release of *R* uses the corresponding *Bioconductor* packages. The user thus has access to stable and tested package versions. *R* and *Bioconductor* are effective tools for reproducible research.

R and *Bioconductor* exist on the leading portion of the software life cycle. Contributors are primarily from academic institutions, and are directly involved in novel research activities. New developments are made available in a familiar format, i.e., the *R* language, packaging, and build systems. The rich set of facilities in *R* (e.g., for advanced statistical analysis or visualization) and the extensive resources in *Bioconductor* (e.g., for annotation using third-party data such as Biomart or the UCSC genome browser tracks) mean that innovations can be directly incorporated into existing work flows. The 'development' branches of *R* and *Bioconductor* provide an environment where contributors can explore new approaches without alienating their user base.

R and *Bioconductor* also fair well in terms of accessibility. The software is freely available. The source code is easily and fully accessible for critical evaluation. The *R* packaging and check system requires that all functions are documented. *Bioconductor* requires that each package contain vignettes to illustrate the use of the software. There are very active *R* and *Bioconductor* mailing lists for immediate support, and regular training and conference activities for professional development.

2.2 *R* data types

Opening an *R* session results in a prompt. The user types instructions at the prompt. Here's an example:

```
> ## assign values 5, 4, 3, 2, 1 to variable 'x'
> x <- c(5, 4, 3, 2, 1)
> x

[1] 5 4 3 2 1
```

The first line starts with a `#` to represent a comment; the line is ignored by *R*. The next line creates a variable `x`. The variable is assigned (using `<-`, we could have used `=` almost interchangeably) a value. The value assigned is the result of a call to the `c` function. That it is a function call is indicated by the symbol named followed by parentheses, `c()`. The `c` function takes zero or more arguments, and returns a vector. The vector is the value assigned to `x`. *R* responds to this line with a new prompt, ready for the next input. The next line asks *R* to display the value of the variable `x`. *R* responds by printing `[1]` to indicate that the subsequent number is the first element of the vector. It then prints the value of `x`.

R has many features to aid common operations. Entering sequences is a very common operation, and expressions of the form `2:4` create a sequence from `2` to `4`. Subsetting one vector by another is enabled with `[`. Here we create a sequence from 2 to 4, and use the sequence as an index to select the second, third, and fourth elements of `x`

```
> x[2:4]

[1] 4 3 2
```

R functions operate on variables. Functions are usually vectorized, acting on all elements of their argument and obviating the need for explicit iteration. Functions can generate warnings when performing suspect operations, or errors if evaluation cannot proceed; try `log(0)` or `log(-1)`.

```
> log(x)

[1] 1.61 1.39 1.10 0.69 0.00
```

Essential data types *R* has a number of standard data types, to represent `integer`, `numeric` (floating point), `complex`, `character`, `logical` (boolean), and `raw` (byte) data. It is possible to convert between data types, and to discover the type or mode of a variable.

```
> c(1.1, 1.2, 1.3)          # numeric

[1] 1.1 1.2 1.3

> c(FALSE, TRUE, FALSE)     # logical

[1] FALSE  TRUE FALSE

> c("foo", "bar", "baz")    # character, single or double quote ok

[1] "foo" "bar" "baz"

> as.character(x)           # convert 'x' to character

[1] "5" "4" "3" "2" "1"

> typeof(x)                 # the number 5 is numeric, not integer

[1] "double"
```

```
> typeof(2L)                # append 'L' to force integer

[1] "integer"

> typeof(2:4)               # ':' produces a sequence of integers

[1] "integer"
```

R includes data types particularly useful for statistical analysis, including `factor` to represent categories and `NA` (used in any vector) to represent missing values.

```
> sex <- factor(c("Male", "Female", NA), levels=c("Female", "Male"))
> sex

[1] Male   Female <NA>
Levels: Female Male
```

Lists, data frames, and matrices All of the vectors mentioned so far are homogenous, consisting of a single type of element. A `list` can contain a collection of different types of elements and, like all vectors, these elements can be named to create a key-value association.

```
> lst <- list(a=1:3, b=c("foo", "bar"), c=sex)
> lst

$a
[1] 1 2 3

$b
[1] "foo" "bar"

$c
[1] Male   Female <NA>
Levels: Female Male
```

Lists can be subset like other vectors to get another list, or subset with `[[` to retrieve the actual list element; as with other vectors, subsetting can use names

```
> lst[c(3, 1)]              # another list

$c
[1] Male   Female <NA>
Levels: Female Male

$a
[1] 1 2 3

> lst[["a"]]                # the element itself, by name

[1] 1 2 3
```

A `data.frame` is a list of equal-length vectors, representing a rectangular data structure not unlike a spread sheet. Each column of the data frame is a vector, so data types must be homogenous within a column. A `data.frame` can be subset by row or column, and columns can be accessed with `$` or `[[`.

```
> df <- data.frame(age=c(27L, 32L, 19L),
+                  sex=factor(c("Male", "Female", "Male")))
> df
  age    sex
1  27   Male
2  32 Female
3  19   Male
> df[c(1, 3),]
  age  sex
1  27 Male
3  19 Male
> df[df$age > 20,]
  age    sex
1  27   Male
2  32 Female
```

A `matrix` is also a rectangular data structure, but subject to the constraint that all elements are the same type. A matrix is created by taking a vector, and specifying the number of rows or columns the vector is to represent. On subsetting, *R* coerces a single column `data.frame` or single row or column `matrix` to a vector if possible; use `drop=FALSE` to stop this behavior.

```
> m <- matrix(1:12, nrow=3)
> m
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
> m[c(1, 3), c(2, 4)]
     [,1] [,2]
[1,]    4   10
[2,]    6   12
> m[, 3]
[1] 7 8 9
> m[, 3, drop=FALSE]
     [,1]
[1,]    7
[2,]    8
[3,]    9
```

An `array` is a data structure for representing homogenous, rectangular data in higher dimensions.

S3 and S4 classes More complicated data structures are represented using the 'S3' or 'S4' object system. Objects are often created by functions (`lm`, below), with parts of the object extracted or assigned using *accessor* functions. The following generates 1000 random normal deviates as `x`, and uses these to create another 1000 deviates `y` that are linearly related to `x` but with some error. We fit a linear regression using a 'formula' to describe the relationship between variables, summarize the results in a familiar ANOVA table, and access `fit` (an S3 object) for the residuals of the regression, using these as input first to the `var` (variance) and then `sqrt` (square-root) functions. Objects can be interogated for their class.

```
> x <- rnorm(1000, sd=1)
> y <- x + rnorm(1000, sd=.5)
> fit <- lm(y ~ x)       # formula describes linear regression
> fit                    # an 'S3' object

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    0.00863      0.99886

> anova(fit)

Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value Pr(>F)
x           1    940     940    3726 <2e-16 ***
Residuals 998    252       0
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

> sqrt(var(resid(fit)))  # residuals accessor and subsequent transforms

[1] 0.5

> class(fit)

[1] "lm"
```

Many *Bioconductor* packages implement S4 objects to represent data. S3 and S4 systems are quite different from a programmer's perspective, but fairly similar from a user's perspective: both systems encapsulate complicated data structures, and allow for methods specialized to different data types; accessors are used to extract information from the objects.

Functions *R* functions accept arguments, and return values. Arguments can be required or optional. Some functions may take variable numbers of arguments, e.g., the columns in a `data.frame`

```
> y <- 5:1
> log(y)

[1] 1.61 1.39 1.10 0.69 0.00

> args(log)        # arguments 'x' and 'base'; see ?log

function (x, base = exp(1))
NULL

> log(y, base=2)   # 'base' is optional, with default value

[1] 2.3 2.0 1.6 1.0 0.0

> try(log())       # 'x' required; 'try' continues even on error
> args(data.frame) # ... represents variable number of arguments

function (..., row.names = NULL, check.rows = FALSE, check.names = TRUE,
    stringsAsFactors = default.stringsAsFactors())
NULL
```

Arguments can be matched by name (highest priority)or position. If an argument appears after `...`, it must be named.

```
> log(base=2, y)   # match argument 'base' by name, 'x' by position

[1] 2.3 2.0 1.6 1.0 0.0
```

A function such as `anova` is a *generic* that provides an overall signature but dispatches the actual work to the *method* corresponding to the class(es) of the arguments used to invoke the generic. A generic may have fewer arguments than a method, as with the S3 function `anova` and its method `anova.glm`.

```
> args(anova)

function (object, ...)
NULL

> args(anova.glm)

function (object, ..., dispersion = NULL, test = NULL)
NULL
```

The `...` argument in the `anova` generic means that additional arguments are possible; the `anova` generic hands these arguments to the method it dispatches to.

2.3 Packages

Packages provide functionality beyond that available in base *R*. There are over 3000 packages in CRAN (comprehensive *R* archive network) and more than 500 *Bioconductor* packages. Packages are contributed by diverse members of the community; they vary in quality (many are excellent) and sometimes contain idiosyncratic aspects to their implementation.

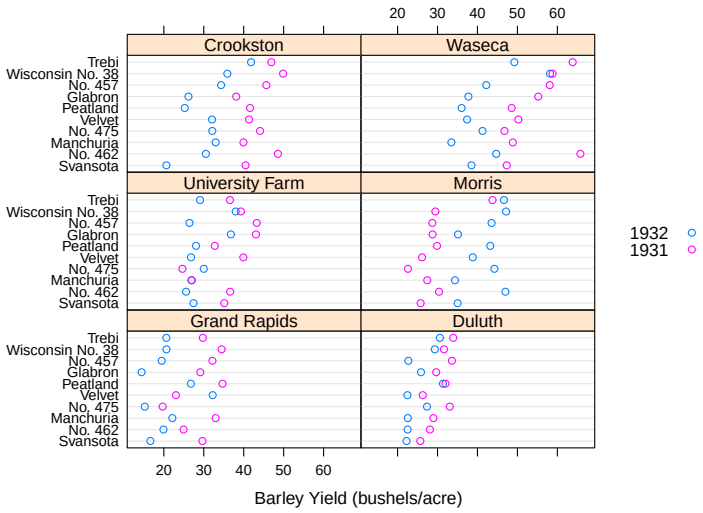

Figure 1: Variety yield conditional on site and grouped by year, for the `barley` data set.

The *lattice* package is distributed with *R* but not loaded by default. It provides a very expressive way to visualize data. The following example plots yield for a number of barley varieties, conditioned on site and grouped by year. Figure 1 is read from the lower left corner. Note the common scales, efficient use of space, and not-too-pleasing default color palette. The Waseca sample appears to be mis-labelled for 'year', an apparent error in the original data. Find out about the built-in data set used in this example with `?barley`.

```
> library(lattice)
> dotplot(variety ~ yield | site, data = barley, groups = year,
+         key = simpleKey(levels(barley$year), space = "right"),
+         xlab = "Barley Yield (bushels/acre) ",
+         aspect=0.5, layout = c(2,3), ylab=NULL)
```

New packages can be added to an *R* installation using `install.packages`. A package is installed only once per *R* installation, but needs to be loaded (with `library`) in each session in which it is used. Loading a package also loads any package that it depends on. Packages loaded in the current session are displayed with `search`. The ordering of packages returned by `search` represents the order in which the global environment (where commands entered at the prompt are evaluated) and attached packages are searched for symbols; it is possible for a package earlier in the search path to mask symbols later in the search path; these can be disambiguated using `::`.

```
> search()

 [1] ".GlobalEnv"
 [2] "package:SeattleIntro2011"
 [3] "package:TxDb.Hsapiens.UCSC.hg19.knownGene"
 [4] "package:genefilter"
 [5] "package:BSgenome.Dmelanogaster.UCSC.dm3"
 [6] "package:org.Dm.eg.db"
 [7] "package:RSQLite"
 [8] "package:DBI"
 [9] "package:chipseq"
[10] "package:BSgenome"
[11] "package:goseq"
[12] "package:geneLenDataBase"
[13] "package:BiasedUrn"
[14] "package:ShortRead"
[15] "package:latticeExtra"
[16] "package:RColorBrewer"
[17] "package:Rsamtools"
[18] "package:lattice"
[19] "package:Biostrings"
[20] "package:SeattleIntro2011Data"
[21] "package:edgeR"
[22] "package:GenomicFeatures"
[23] "package:AnnotationDbi"
[24] "package:Biobase"
[25] "package:GenomicRanges"
[26] "package:IRanges"
[27] "package:stats"
[28] "package:graphics"
[29] "package:grDevices"
[30] "package:utils"
[31] "package:datasets"
[32] "package:methods"
[33] "Autoloads"
[34] "package:base"

> base::log(1:3)

[1] 0.00 0.69 1.10
```

Exercise 1
Use the `library` function to load the SeattleIntro2011 package. Use the `sessionInfo` function to verify that you are using R version 2.14.0 and current packages, similar to those reported here. What other packages were loaded along with SeattleIntro2011?

Solution:

```
> library(SeattleIntro2011)
> sessionInfo()
```

```
R version 2.14.0 alpha (2011-10-04 r57169)
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

locale:
[1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] SeattleIntro2011_0.0.22
 [2] TxDb.Hsapiens.UCSC.hg19.knownGene_2.6.2
 [3] genefilter_1.35.0
 [4] BSgenome.Dmelanogaster.UCSC.dm3_1.3.17
 [5] org.Dm.eg.db_2.6.1
 [6] RSQLite_0.9-4
 [7] DBI_0.2-5
 [8] chipseq_1.3.4
 [9] BSgenome_1.21.3
[10] goseq_1.5.0
[11] geneLenDataBase_0.99.7
[12] BiasedUrn_1.04
[13] ShortRead_1.11.42
[14] latticeExtra_0.6-18
[15] RColorBrewer_1.0-5
[16] Rsamtools_1.5.75
[17] lattice_0.19-33
[18] Biostrings_2.21.9
[19] SeattleIntro2011Data_0.0.3
[20] edgeR_2.3.52
[21] GenomicFeatures_1.5.27
[22] AnnotationDbi_1.15.38
[23] Biobase_2.13.11
[24] GenomicRanges_1.5.49
[25] IRanges_1.11.26

loaded via a namespace (and not attached):
 [1] Matrix_1.0-0        RCurl_1.6-10        XML_3.4-3
 [4] annotate_1.31.1     biomaRt_2.9.2       grid_2.14.0
 [7] hwriter_1.3         limma_3.9.20        mgcv_1.7-6
[10] nlme_3.1-102        rtracklayer_1.13.17 splines_2.14.0
[13] survival_2.36-10    tools_2.14.0        xtable_1.5-6
[16] zlibbioc_0.1.7
```

2.4 Help

Find help using the *R* help system. Start a web browser with

```
> help.start()
```

The 'Search Engine and Keywords' link is helpful in day-to-day use.

Manual pages Use manual pages to find detailed descriptions of the arguments and return values of functions, and the structure and methods of classes. Find help within an *R* session as

```
> ?data.frame
> ?lm
> ?anova                  # a generic function
> ?anova.lm               # an S3 method, specialized for 'lm' objects
```

S3 methods can be queried interactively. For S3,

```
> methods(anova)
```

```
 [1] anova.MAList*      anova.coxph*       anova.coxphlist*   anova.gam*
 [5] anova.glm          anova.glmlist      anova.gls*         anova.lm
 [9] anova.lme*         anova.loess*       anova.mlm          anova.nls*
[13] anova.survreg*     anova.survreglist*

   Non-visible functions are asterisked
```

```
> methods(class="lm")
```

```
 [1] add1.lm*           alias.lm*          anova.lm           case.names.lm*
 [5] confint.lm*        cooks.distance.lm* deviance.lm*       dfbeta.lm*
 [9] dfbetas.lm*        drop1.lm*          dummy.coef.lm*     effects.lm*
[13] extractAIC.lm*     family.lm*         formula.lm*        hatvalues.lm
[17] influence.lm*      kappa.lm           labels.lm*         logLik.lm*
[21] model.frame.lm     model.matrix.lm    nobs.lm*           plot.lm
[25] predict.lm         print.lm           proj.lm*           qqnorm.lm*
[29] qr.lm*             residuals.lm       rstandard.lm       rstudent.lm
[33] simulate.lm*       summary.lm         variable.names.lm* vcov.lm*

   Non-visible functions are asterisked
```

It is often useful to view a method definition, either by typing the method name at the command line or, for 'non-visible' methods, using `getAnywhere`:

```
> anova.lm
> getAnywhere(anova.loess)
```

For instance, the source code of a function is printed if the function is invoked without parentheses. Here we discover that the function `head` (which returns the first 6 elements of anything) defined in the *utils* package, is an S3 generic (indicated by `UseMethod`) and has several methods. We use `head` to look at the first six lines of the `head` method specialized for `matrix` objects.

```
> utils::head
```

```
function (x, ...)
UseMethod("head")
<bytecode: 0x2a382b4>
<environment: namespace:utils>
```

```
> methods(head)

[1] head.data.frame* head.default*    head.ftable*     head.function*
[5] head.matrix      head.table*

   Non-visible functions are asterisked

> head(head.matrix)

1 function (x, n = 6L, ...)
2 {
3     stopifnot(length(n) == 1L)
4     n <- if (n < 0L)
5         max(nrow(x) + n, 0L)
6     else min(n, nrow(x))
```

S4 classes and generics are queried in a similar way to S3 classes and generics, but with different syntax, as for the `complement` generic in the *Biostrings* package:

```
> showMethods(complement)

Function: complement (package Biostrings)
x="DNAString"
x="DNAStringSet"
x="MaskedDNAString"
x="MaskedRNAString"
x="RNAString"
x="RNAStringSet"
x="XStringViews"
```

Methods defined on the `DNAStringSet` class of *Biostrings* can be found with

```
> showMethods(class="DNAStringSet", where=getNamespace("Biostrings"))
```

Obtaining help on S4 classes and methods requires syntax such as

```
> class ? DNAStringSet
> method ? "complement,DNAStringSet"
```

The specification of method and class in the latter must not contain a space after the comma. The definition of a method can be retrieved as

```
> selectMethod(complement, "DNAStringSet")
```

Vignettes Vignettes, especially in *Bioconductor* packages, provide an extensive narrative describing overall package functionality. Use

```
> browseVignettes("SeattleIntro2011")
```

to see, in your web browser, vignettes available in the *SeattleIntro2011* package. Vignettes usually consist of text with embedded *R* code, a form of literate programming. The vignette can be read as a PDF document, while the *R* source code is present as a script file ending with extension `.R`. The script file can be sourced or copied into an *R* session to evaluate exactly the commands used in the vignette.

2.5 The *Bioconductor* web site

The *Bioconductor* web site is at `bioconductor.org`. Features include:

- Brief introductory work flows.
- A manifest of all *Bioconductor* packages arranged alphabetically or as BiocViews.
- Annotation (data bases of relevant genomic information, e.g., Entrez gene ids in model organisms, KEGG pathways) and experiment data (containing relatively comprehensive data sets and their analysis) packages.
- Access to the mailing lists, including searchable archives, as the primary source of help.
- Course and conference information, including extensive reference material.
- General information about the project.
- Information for package developers, including guidelines for creating and submitting new packages.

Exercise 2
Scavenger hunt. Spend five minutes tracking down the following information.

a. *The package containing the* `library` *function.*

b. *The author of the* `alphabetFrequency` *function, defined in the Biostrings package.*

c. *A description of the GappedAlignments class.*

d. *The number of vignettes in the GenomicRanges package.*

e. *From the Bioconductor web site, instructions for installing or updating Bioconductor packages.*

f. *A list of all packages in the current release of Bioconductor.*

g. *The URL of the Bioconductor mailing list subscription page.*

Solution: Possible solutions are found with the following *R* commands

```
> ?library
> library(Biostrings)
> ?alphabetFrequency
> class?GappedAlignments
> browseVignettes("GenomicRanges")
```

and by visiting the *Bioconductor* web site, e.g., `http://bioconductor.org/install/` (installation instructions), `http://bioconductor.org/packages/release/bioc/` (current software packages), and `http://bioconductor.org/help/mailing-list/` (mailing lists).

2.6 Resources

Dalgaard [4] provides an introduction to statistical analysis with *R*. Matloff [12] introduces *R* programming concepts. Chambers [3] provides more advanced insights into *R*. Gentleman [5] emphasizes use of *R* for bioinformatic programming tasks. The *R* web site enumerates additional publications from the user community.

3 Ranges and strings

This section introduces two essential ways in which sequence data are manipulated. Ranges describe both aligned reads and features of interest on the genome. Sets of DNA strings represent the reads themselves and the nucleotide sequence of reference genomes.

3.1 Reads and genomic ranges

Next-generation sequencing data consists of a large number of short reads. These are, typically, aligned to a reference genome. Basic operations are performed on the alignment, asking e.g., how many reads are aligned in a genomic range defined by nucleotide coordinates (e.g., in the exons of a gene), or how many nucleotides from all the aligned reads cover a set of genomic coordinates? How is this type of data, the aligned reads and the reference genome, to be represented in *R* in a way that allows for effective computation?

The *IRanges*, *GenomicRanges*, and *GenomicFeatures* *Bioconductor* packages provide the essential infrastructure for these operations; we start with the *GRanges* class, defined in *GenomicRanges*.

GRanges Instances of *GRanges* are used to specify genomic coordinates. Suppose we wished to represent two *D. melanogaster* genes. The first is located on the positive strand of chromosome 3R, from position 19967117 to 19973212. The second is on the minus strand of the X chromosome, with 'left-most' base at 18962306, and right-most base at 18962925. The coordinates are *1-based* (i.e., the first nucleotide on a chromosome is numbered 1, rather than 0), *left-most* (i.e., reads on the minus strand are defined to 'start' at the left-most coordinate, rather than the 5' coordinate), and *closed* (the start and end coordinates are included in the range; a range with identical start and end coordinates has width 1, a 0-width range is represented by the special construct where the end coordinate is one less than the start coordinate).

A complete definition of these genes as *GRanges* is:

```
> genes <- GRanges(seqnames=c("3R", "X"),
+                  ranges=IRanges(
+                      start=c(19967117, 18962306),
+                      end=c(19973212, 18962925)),
+                  strand=c("+", "-"),
+                  seqlengths=c(`3R`=27905053L, `X`=22422827L))
```

The components of a *GRanges* object are defined as vectors, e.g., of `seqnames`, much as one would define a *data.frame*. The start and end coordinates are grouped into an *IRanges* instance. The optional `seqlengths` argument specifies the maximum size of each sequence, in this case the lengths of chromosomes 3R and X in *D. melanogaster*. This data is displayed as

```
> genes
```

```
GRanges with 2 ranges and 0 elementMetadata values:
      seqnames               ranges strand
         <Rle>            <IRanges>  <Rle>
```

```
  [1]       3R [19967117, 19973212]      +
  [2]        X [18962306, 18962925]      -
  ---
  seqlengths:
         3R        X
   27905053 22422827
```

For the curious, the gene coordinates and sequence lengths are derived from the *org.Dm.eg.db* for genes with Flybase identifiers FBgn0039155 and FBgn0085359, using the annotation facilities described in section 7.

The *GRanges* class has many useful methods defined on it. Consult the help page

```
> ?GRanges
```

and package vignettes (especially 'An Introduction to *GenomicRanges*')

```
> browseVignettes("GenomicRanges")
```

for a comprehensive introduction. A *GRanges* instance can be subset, with accessors for getting and updating information.

```
> genes[2]
```

```
GRanges with 1 range and 0 elementMetadata values:
      seqnames               ranges strand
         <Rle>            <IRanges>  <Rle>
  [1]        X [18962306, 18962925]      -
  ---
  seqlengths:
         3R        X
   27905053 22422827
```

```
> strand(genes)
```

```
'factor' Rle of length 2 with 2 runs
  Lengths: 1 1
  Values : + -
Levels(3): + - *
```

```
> width(genes)
```

```
[1] 6096  620
```

```
> length(genes)
```

```
[1] 2
```

```
> names(genes) <- c("FBgn0039155", "FBgn0085359")
> genes # now with names
```

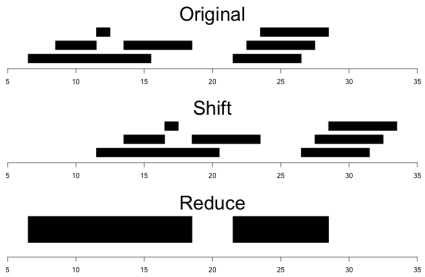

Figure 2: Ranges

```
GRanges with 2 ranges and 0 elementMetadata values:
             seqnames              ranges strand
                <Rle>           <IRanges>  <Rle>
  FBgn0039155       3R [19967117, 19973212]      +
  FBgn0085359        X [18962306, 18962925]      -
  ---
  seqlengths:
         3R        X
   27905053 22422827
```

`strand` returns the strand information in a compact representation called a *run-length encoding*, this is introduced in greater detail below. The 'names' could have been specified when the instance was constructed; once named, the *GRanges* instance can be subset by name like a regular vector.

As the `GRanges` function suggests, the *GRanges* class extends the *IRanges* class by adding information about `seqname`, `strand`, and other information particularly relevant to representing ranges that are on genomes. The *IRanges* class and related data structures (e.g., *RangedData*) are meant as a more general description of ranges defined in an arbitrary space. Many methods implemented on the *GRanges* class are 'aware' of the consequences of genomic location, for instance treating ranges on the minus strand differently (reflecting the 5' orientation imposed by DNA) from ranges on the plus strand.

Operations on ranges The *GRanges* class has many useful methods from the *IRanges* class; some of these methods are illustrated here. We use *IRanges* to illustrate these operations to avoid complexities associated with strand and seqname, but the operations are comparable on *GRanges*. We begin with a simple set of ranges:

```
> ir <- IRanges(start=c(7, 9, 12, 14, 22:24),
+               end=c(15, 11, 12, 18, 26, 27, 28))
```

These are illustrated in the upper panel of Figure 2.

Methods on ranges can be grouped as follows:

Intra-range methods act on each range independently. These include `flank`, `narrow`, `reflect`, `resize`, `restrict`, and `shift`, among others. An illustration is `shift`, which translates each range by the amount specified by the `shift` argument. Positive values shift to the right, negative to the left;

`shift` can be a vector, with each element of the vector shifting the corresponding element of the *IRanges* instance. Here we shift all ranges to the right by 5, with the result illustrated in the middle panel of Figure 2

```
> shift(ir, 5)
```

```
IRanges of length 7
    start end width
[1]    12  20     9
[2]    14  16     3
[3]    17  17     1
[4]    19  23     5
[5]    27  31     5
[6]    28  32     5
[7]    29  33     5
```

Inter-range methods act on the collection of ranges as a whole. These include `disjoin`, `reduce`, `gaps`, and `range`. An illustration is `reduce`, which reduces overlapping ranges into a single range, as illustrated in the lower panel of Figure 2.

```
> reduce(ir)
```

```
IRanges of length 2
    start end width
[1]     7  18    12
[2]    22  28     7
```

`coverage` is an inter-range operation that calculates how many ranges overlap individual positions. Rather than returning ranges, `coverage` returns a compressed representation of an integer vector representing a (run-length encoding)

```
> coverage(ir)
```

```
'integer' Rle of length 28 with 12 runs
  Lengths: 6 2 4 1 2 3 3 1 1 3 1 1
  Values : 0 1 2 1 2 1 0 1 2 3 2 1
```

The run-length encoding can be interpreted as 'a run of length 6 of nucleotides covered by 0 ranges, followed by a run of length 2 of nucleotides covered by 1 range...'.

Between methods act on two (or sometimes more) *IRanges* instances. These include `intersect`, `setdiff`, `union`, `pintersect`, `psetdiff`, and `punion`.

`countOverlaps` and `findOverlaps` also operate on two sets of ranges. `countOverlaps` takes its first argument (the `query`) and determines how many of the ranges in the second argument (the `subject`) each overlaps. The result is an integer vector with one element for each member of `query`. `findOverlaps` performs a similar operation but returns a more general matrix-like structure that identifies each pair of query / subject overlaps. Both arguments allow some flexibility in the definition of 'overlap'.

`elementMetadata` **and** `metadata` The *GRanges* class (actually, most of the data structures defined or extending those in the *IRanges* package) has two additional very useful data components. The `elementMetadata` function (or its synonym `values`) allows information on each range to be stored and manipulated (e.g., subset) along with the *GRanges* instance. The element metadata is represented as a *DataFrame*, defined in *IRanges* and acting like a standard *R data.frame* but with the ability to hold more complicated data structures as columns (and with element metadata of its own, providing an enhanced alternative to the *Biobase* class *AnnotatedDataFrame*).

```
> elementMetadata(genes) <-
+     DataFrame(EntrezId=c("42865", "2768869"),
+               Symbol=c("kal-1", "CG34330"))
```

`metadata` allows addition of information to the entire object. The information is in the form of a list; any data can be provided.

```
> metadata(genes) <-
+     list(CreatedBy="A. User", Date=date())
```

The `GRanges` class is extremely useful for representing simple ranges. Some next-generation sequence data and genomic features are more hierarchically structured. A gene may be represented by several exons within it. An aligned read may be represented by discontinuous ranges of alignment to a reference. The *GRangesList* class represents this type of information. It is a list-like data structure, which each element of the list itself a *GRanges* instance. The gene FBgn0039155 contains several exons, and can be represented as a length 1 list, where the element of the list contains a *GRanges* object with 7 elements:

```
GRangesList of length 1:
$FBgn0039155
GRanges with 7 ranges and 2 elementMetadata values:
      seqnames              ranges strand |   exon_id   exon_name
         <Rle>           <IRanges>  <Rle> | <integer> <character>
  [1]    chr3R [19967117, 19967382]      + |     64137        <NA>
  [2]    chr3R [19970915, 19971592]      + |     64138        <NA>
  [3]    chr3R [19971652, 19971770]      + |     64139        <NA>
  [4]    chr3R [19971831, 19972024]      + |     64140        <NA>
  [5]    chr3R [19972088, 19972461]      + |     64141        <NA>
  [6]    chr3R [19972523, 19972589]      + |     64142        <NA>
  [7]    chr3R [19972918, 19973212]      + |     64143        <NA>

---
seqlengths:
    chr3R
 27905053
```

The *GRangesList* object has methods one would expect for lists (e.g., `length`, subsetting). Many of the methods introduced for working with *IRanges* are also available, with the method applied element-wise.

The *GenomicFeatures* package Many public resources provide annotations about genomic features. For instance, the UCSC genome browser maintains the 'knownGene' track of established exons, transcripts, and coding sequences of many model organisms. The *GenomicFeatures* package provides a way to retrieve, save, and query these resources. The underlying representation is as sqlite data bases, but the data are available in *R* as *GRangesList* objects. The following exercise explores the *GenomicFeatures* package and some of the functionality for the *IRanges* family introduced above.

Exercise 3
Use the helper function `bigdata` *and* `list.files` *to identify the path to a data base created by* `makeTranscriptDbFromUCSC`*.*

Load the saved TranscriptDb object using `loadFeatures`*.*

Extract all exon coordinates, organized by gene, using `exonsBy`*. What is the class of this object? How many elements are in the object? What does each element correspond to? And the elements of each element? Use* `elementLengths` *and* `table` *to summarize the number of exons in each gene, for instance, how many single-exon genes are there?*

Select just those elements corresponding to flybase gene ids FBgn0002183, FBgn0003360, FBgn0025111, and FBgn0036449. Use `reduce` *to simplify gene models, so that exons that overlap are considered 'the same'.*

Solution:

```
> txdbFile <- list.files(bigdata(), "sqlite", full=TRUE)
> txdb <- loadFeatures(txdbFile)
> ex0 <- exonsBy(txdb, "gene")
> head(table(elementLengths(ex0)))

   1    2    3    4    5    6
3182 2608 2070 1628 1133  886

> ids <- c("FBgn0002183", "FBgn0003360", "FBgn0025111", "FBgn0036449")
> ex <- reduce(ex0[ids])
```

Exercise 4
(Independent) Create a TranscriptDb instance from UCSC, using `makeTranscriptDbFromUCSC`*.*

Solution:

```
> txdb <- makeTranscriptDbFromUCSC("dm3", "ensGene")
> saveFeatures(txdb, "my.dm3.ensGene.txdb.sqlite")
```

3.2 Working with strings

Underlying the ranges of alignments and features are DNA sequences. The *Biostrings* package provides tools for working with this data. The essential data structures are *DNAString* and *DNAStringSet*, for working with one or multiple DNA sequences. The *Biostrings* package contains additional classes for representing amino acid and general biological strings. The *BSgenome* and related packages (e.g., *BSgenome.Dmelanogaster.UCSC.dm3*) are used to represent whole-genome sequences. The following exercise explores these packages.

Exercise 5
The objective of this exercise is to calculate the GC content of the exons of a single gene, whose coordinates are specified by the `ex` *object of the previous exercise.*

Load the BSgenome.Dmelanogaster.UCSC.dm3 data package, containing the UCSC representation of D. melanogaster *genome assembly dm3.*

Extract the sequence name of the first gene of `ex`*. Use this to load the appropriate* D. melanogaster *chromosome.*

Use `Views` *to create views on to the chromosome that span the start and end coordinates of all exons.*

The SeattleIntro2011 package defines a helper function `gcFunction` *(developed in a later exercise) to calculate GC content. Use this to calculate the GC content in each of the exons.*

Solution:

```
> library(BSgenome.Dmelanogaster.UCSC.dm3)
> nm <- as.character(unique(seqnames(ex[[1]])))
> chr <- Dmelanogaster[[nm]]
> v <- Views(chr, start=start(ex[[1]]), end=end(ex[[1]]))
```

Here is the helper function, available in the *SeattleIntro2011* package, to calculate GC content:

```
> gcFunction

function (x)
{
    alf <- alphabetFrequency(x, as.prob = TRUE)
    rowSums(alf[, c("G", "C")])
}
<environment: namespace:SeattleIntro2011>
```

The subject GC content is

```
> subjectGC <- gcFunction(v)
```

4 Exploring sequence data: short reads and alignments

The following sections introduce core tools for working with high-throughput sequence data. This section focus on the reads and alignments that are the raw material for analysis. Section 5 addresses statistical approaches to assessing differential representation in RNA-seq experiments. Section 6 outlines ChIP-seq analysis.

4.1 The *pasilla* data set

As a running example, we use the *pasilla* data set, derived from [2]. The authors investigate conservation of RNA regulation between *D. melanogaster* and mammals. Part of their study used RNAi and RNA-seq to identify exons regulated by Pasilla (*ps*), the *D. melanogaster* ortholog of mammalian NOVA1 and NOVA2. Briefly, their experiment compared gene expression as measured by RNAseq in S2-DRSC cells cultured with, or without, a 444bp dsRNA fragment corresponding to the *ps* mRNA sequence. Their assessment investigated differential exon use, but our worked example will focus on gene-level differences.

In this section we look at a subset of the *ps* data, corresponding to reads obtained from lanes of their RNA-seq experiment, and to the same reads aligned to a *D. melanogaster* reference genome. Reads were obtained from GEO and the Short Read Archive (SRA); reads were aligned to *D. melanogaster* reference genome dm3 as described in the *pasilla* experiment data package.

4.2 Short reads

Sequencer technologies The Illumina GAII and HiSeq technologies generate sequences by measuring incorporation of florescent nucleotides over successive PCR cycles. These sequencers produce output in a variety of formats, but *FASTQ* is ubiquitous. Each read is represented by a record of four components:

```
@SRR031724.1 HWI-EAS299_4_30M2BAAXX:5:1:1513:1024 length=37
GTTTTGTCCAAGTTCTGGTAGCTGAATCCTGGGGCGC
+SRR031724.1 HWI-EAS299_4_30M2BAAXX:5:1:1513:1024 length=37
IIIIIIIIIIIIIIIIIIIIIIIIIIII+HIIII<IE
```

The first and third lines (beginning with `@` and `+` respectively) are unique identifiers. In the sample above the identifier produced by the sequencer typically includes a machine id followed by colon-separated information on the lane, tile, x, and y coordinate of the read. The example illustrated here also includes the SRA accession number, added when the data was submitted to the archive. The machine identifier could potentially be used to extract information about batch effects. The spatial coordinates (lane, tile, x, y) are often used to identify optical duplicates; spatial coordinates can also be used during quality assessment to identify artifacts of sequencing, e.g., uneven amplification across the flow cell, though these spatial effects are rarely pursued.

The second and fourth lines of the FASTQ record are the nucleotides and qualities of each cycle in the read. This information is given in 5' to 3' orientation as seen by the sequencer. A letter `N` is used to signify bases that the

sequencer was not able to call. The fourth line of the FASTQ record encodes the quality (confidence) of the corresponding base call. The quality score is encoded following one of several conventions, with the general notion being that letters later in the visible ASCII alphabet

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNO
PQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

are of lower quality; this is developed further below. Both the sequence and quality scores may span multiple lines.

Technologies other than Illumina use different formats to represent sequences. Roche 454 sequence data is generated by 'flowing' labeled nucleotides over samples, with greater intensity corresponding to longer runs of A, C, G, or T. This data is represented as a series of 'flow grams' (a kind of run-length encoding of the read) in Standard Flowgram Format (SFF). The *Bioconductor* package *R453Plus1Toolbox* has facilities for parsing SFF files, but after quality control steps the data are frequently represented (with some loss of information) as FASTQ. SOLiD technologies produce sequence data using a 'color space' model. This data is not easily read in to *R*, and much of the error-correcting benefit of the color space model is lost when converted to FASTQ; SOLiD sequences are not well-handled by *Bioconductor* packages.

Short reads in *R* FASTQ files can be read in to *R* using the `readFastq` function from the *ShortRead* package. Use this function by providing the path to a FASTQ file. There are sample data files available in the *SeattleIntro2011Data* package, each consisting of 1 million reads from a lane of the Pasilla data set.

```
> fastqDir <- file.path(bigdata(), "fastq")
> fastqFiles <- list.files(fastqDir, full=TRUE)
> fq <- readFastq(fastqFiles[1], withIds=TRUE)
> fq

class: ShortReadQ
length: 1000000 reads; width: 37 cycles
```

The data are represented as an object of class *ShortReadQ* ('short read and *q*uality').

```
> head(sread(fq), 3)

  A DNAStringSet instance of length 3
    width seq
[1]    37 GTTTTGTCCAAGTTCTGGTAGCTGAATCCTGGGGCGC
[2]    37 GTTGTCGCATTCCTTACTCTCATTCGGGAATTCTGTT
[3]    37 GAATTTTTTGAGAGCGAAATGATAGCCGATGCCCTGA

> head(quality(fq), 3)

class: FastqQuality
quality:
  A BStringSet instance of length 3
    width seq
```

```
[1]    37 IIIIIIIIIIIIIIIIIIIIIIIIIII+HIIII<IE
[2]    37 IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
[3]    37 IIIIIIIIIIIIIIIIIIIII'IIIIIGBIIII2I+
```

```
> head(id(fq), 3)
```

```
  A BStringSet instance of length 3
    width seq
[1]    58 SRR031724.1 HWI-EAS299_4_30M2BAAXX:5:1:1513:1024 length=37
[2]    57 SRR031724.2 HWI-EAS299_4_30M2BAAXX:5:1:937:1157 length=37
[3]    58 SRR031724.4 HWI-EAS299_4_30M2BAAXX:5:1:1443:1122 length=37
```

The *ShortReadQ* class illustrates *class inheritance*. It extends the *ShortRead* class

```
> getClass("ShortReadQ")
```

```
Class "ShortReadQ" [package "ShortRead"]

Slots:

Name:       quality        sread           id
Class: QualityScore DNAStringSet   BStringSet

Extends:
Class "ShortRead", directly
Class ".ShortReadBase", by class "ShortRead", distance 2

Known Subclasses: "AlignedRead"
```

Methods defined on *ShortRead* are available for *ShortReadQ*.

```
> showMethods(class="ShortRead", where=getNamespace("ShortRead"))
```

For instance, the `width` can be used to demonstrate that all reads consist of 37 nucleotides.

```
> table(width(fq))
```

```
     37
1000000
```

The `alphabetByCycle` function summarizes use of nucleotides at each cycle in a (equal width) *ShortReadQ* or *DNAStringSet* instance.

```
> abc <- alphabetByCycle(sread(fq))
> abc[1:4, 1:8]
```

```
        cycle
alphabet   [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]
       A  78194 153156 200468 230120 283083 322913 162766 220205
       C 439302 265338 362839 251434 203787 220855 253245 287010
       G 397671 270342 258739 356003 301640 247090 227811 246684
       T  84833 311164 177954 162443 211490 209142 356178 246101
```

FASTQ files are getting larger. A very common reason for looking at data at this early stage in the processing pipeline is to explore sequence quality. In these circumstances it is often not necessary to parse the entire FASTQ file. Instead create a representative sample

```
> sampler <- FastqSampler(fastqFiles[1], 1000000)
> yield(sampler) # sample of 1000000 reads

class: ShortReadQ
length: 1000000 reads; width: 37 cycles
```

A second common scenario is to pre-process reads, e.g., trimming low-quality tails, adapter sequences, or artifacts of sample preparation. The *FastqStreamer* class can be used to 'stream' over the fastq files in chunks, processing each chunk independently.

ShortRead contains facilities for quality assessment of FASTQ files. Here we generate a report from a sample of 1 million reads from each of our files and display it in a web browser

```
> qas0 <- Map(function(fl, nm) {
+     fq <- FastqSampler(fl)
+     qa(yield(fq), nm)
+ }, fastqFiles,
+    sub("_subset.fastq", "", basename(fastqFiles)))
> qas <- do.call(rbind, qas0)
> rpt <- report(qas, dest=tempfile())
> browseURL(rpt)
```

A report from a larger subset of the experiment is available

```
> rpt <- system.file("GSM461176_81_qa_report", package="SeattleIntro2011")
> browseURL(rpt)
```

Exercise 6

Use the helper function `bigdata` *(defined in the SeattleIntro2011 package) and the* `file.path` *and* `list.files` *functions to locate two fastq files from [2] (the files were obtained as described in the appendix and pasilla experiment data package.*

Input one of the fastq files using `readFastq` *from the ShortRead package.*

Use `alphabetFrequency` *to summarize the GC content of all reads (hint: use the* `sread` *accessor to extract the reads, and the* `collapse=TRUE` *argument to the* `alphabetFrequency` *function). Using the helper function* `gcFunction` *from the SeattleIntro2011 package, draw a histogram of the distribution of GC frequencies across reads.*

Use `alphabetByCycle` *to summarize the frequency of each nucleotide, at each cycle. Plot the results using* `matplot`*, from the graphics package.*

As an advanced exercise, and if on Mac or Linux, use the parallel package and `mclapply` *to read and summarize the GC content of reads in two files in parallel.*

Solution: Discovery:

```
> list.files(bigdata())
```

```
[1] "bam"                       "dm3.ensGene.txdb.sqlite"
[3] "fastq"
```

```
> fls <- list.files(file.path(bigdata(), "fastq"), full=TRUE)
```

Input:

```
> fq <- readFastq(fls[1])
```

GC content:

```
> alf0 <- alphabetFrequency(sread(fq), as.prob=TRUE, collapse=TRUE)
> sum(alf0[c("G", "C")])
```

```
[1] 0.55
```

A histogram of the GC content of individual reads is obtained with

```
> gc <- gcFunction(sread(fq))
> hist(gc)
```

Alphabet by cycle:

```
> abc <- alphabetByCycle(sread(fq))
> matplot(t(abc[c("A", "C", "G", "T"),]), type="l")
```

Advanced (Mac, Linux only): processing on multiple cores.

```
> library(parallel)
> gc0 <- mclapply(fls, function(fl) {
+   fq <- readFastq(fl)
+   gc <- gcFunction(sread(fq))
+   table(cut(gc, seq(0, 1, .05)))
+ })
> ## simplify list of length 2 to 2-D array
> gc <- simplify2array(gc0)
> matplot(gc, type="s")
```

Exercise 7
Use `quality` *to extract the quality scores of the short reads. Interpret the encoding qualitatively.*

Convert the quality scores to a numeric matrix, using `as`*. Inspect the numeric matrix (e.g., using* `dim`*) and understand what it represents.*

Use `colMeans` *to summarize the average quality score by cycle. Use* `plot` *to visualize this*

Solution:

```
> head(quality(fq))
```

```
class: FastqQuality
quality:
  A BStringSet instance of length 6
    width seq
[1]    37 IIIIIIIIIIIIIIIIIIIIIIIIIIIII+HIIII<IE
[2]    37 IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
[3]    37 IIIIIIIIIIIIIIIIIIIII'IIIIIGBIIII2I+
[4]    37 IIIIIIIIIIIIIIIIIIIII,II*E,&4HI++B
[5]    37 IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII&.$
[6]    37 III.IIIIIIIIIIIIIIIII%IIE(-EIH<IIII
```

```
> qual <- as(quality(fq), "matrix")
> dim(qual)
```

```
[1] 1000000      37
```

```
> plot(colMeans(qual), type="b")
```

4.3 Alignments

Most down-stream analysis of short read sequences is based on reads aligned to reference genomes. There are many aligners available, including BWA [11, 10], Bowtie [9], GSNAP, and Illumina's ELAND; merits of these are discussed in the literature. There are also alignment algorithms implemented in *Bioconductor* (e.g., `matchPDict` in the *Biostrings* package, and the *Rsubread* package); `matchPDict` is particularly useful for flexible alignment of moderately sized subsets of data.

Alignment formats Most main-stream aligners produce output in SAM (text-based) or BAM format. A SAM file is a text file, with one line per aligned read, and fields separated by tabs. Here is an example, split into fields.

```
> fl <- system.file("extdata", "ex1.sam", package="Rsamtools")
> strsplit(readLines(fl, 1), "\t")[[1]]
```

```
 [1] "B7_591:4:96:693:509"
 [2] "73"
 [3] "seq1"
 [4] "1"
 [5] "99"
 [6] "36M"
 [7] "*"
 [8] "0"
 [9] "0"
[10] "CACTAGTGGCTCATTGTAAATGTGTGGTTTAACTCG"
[11] "<<<<<<<<<<<<<<;<<<<<<<<<5<<<<<;:<;7"
[12] "MF:i:18"
[13] "Aq:i:73"
[14] "NM:i:0"
```

Table 1: Fields in a SAM record. From http://samtools.sourceforge.net/samtools.shtml

Field	Name	Value
1	QNAME	Query (read) NAME
2	FLAG	Bitwise FLAG, e.g., strand of alignment
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	Extended CIGAR string
7	MRNM	Mate Reference sequence NaMe
8	MPOS	1-based Mate POSistion
9	ISIZE	Inferred insert SIZE
10	SEQ	Query SEQuence on the reference strand
11	QUAL	Query QUALity
12+	OPT	OPTional fields, format TAG:VTYPE:VALUE

```
[15] "UQ:i:0"
[16] "H0:i:1"
[17] "H1:i:0"
```

Fields in a SAM file are summarized in Table 1. We recognize from the FASTQ file the identifier string, read sequence and quality. The alignment is to a chromosome 'seq1' starting at position 1. The strand of alignment is encoded in the 'flag', field. The alignment record also includes a measure of mapping quality, and a CIGAR string describing the nature of the alignment. In this case, the CIGAR is 35M, indicating that the alignment consisted of 35 `M`atches or mismatches, with no indels or gaps; indels are represented by `I` and `D`; gaps (e.g., from alignments spanning introns) by `N`.

BAM files encode the same information as SAM files, but in a format that is more efficiently parsed by software; BAM files are the primary way in which aligned reads are imported in to *R*.

Aligned reads in *R* The `readGappedAlignments` function from the *GenomicRanges* package reads essential information from a BAM file in to *R*. The result is an instance of the *GappedAlignments* class. The *GappedAlignments* class has been designed to allow useful manipulation of many reads (e.g., 20 million) under moderate memory requirements (e.g., 4 GB).

```
> alnFile <- system.file("extdata", "ex1.bam", package="Rsamtools")
> aln <- readGappedAlignments(alnFile)
> head(aln, 3)

GappedAlignments with 3 alignments and 0 elementMetadata values:
      seqnames strand       cigar    qwidth     start       end     width
         <Rle>  <Rle> <character> <integer> <integer> <integer> <integer>
  [1]     seq1      +         36M        36         1        36        36
  [2]     seq1      +         35M        35         3        37        35
  [3]     seq1      +         35M        35         5        39        35
          ngap
```

```
    <integer>
[1]         0
[2]         0
[3]         0
---
seqlengths:
 seq1 seq2
 1575 1584
```

The `readGappedAlignments` function takes an additional parameter, `param`, allowing the user to specify regions of the BAM file (e.g., known gene coordinates) from which to extract alignments.

A *GappedAlignments* instance is like a data frame, but with accessors as suggested by the column names. It is easy to query for, e.g., the distribution of reads aligning to each strand, the width of reads, or the cigar strings

```
> table(strand(aln))

   +    -
1647 1624

> table(width(aln))

  30   31   32   33   34   35   36   38   40
   2   21    1    8   37 2804  285    1  112

> head(sort(table(cigar(aln)), decreasing=TRUE))

     35M      36M      40M      34M      33M 14M4I17M
    2804      283      112       37         6        4
```

Exercise 8

Use `bigdata`, `file.path` and `list.files` to obtain file paths to the BAM files. These are a subset of the aligned reads, overlapping just four genes.

Input the aligned reads from one file using `readGappedAlignments`. Explore the reads, e.g., using `table` or `xtabs` to summarize which chromosome and strand the subset of reads is from.

The object `ex` created earlier contains coordinates of four genes. Use `countOverlaps` to first determine the number of genes an individual read aligns to, and then the number of uniquely aligning reads overlapping each gene. Since the RNAseq protocol was not strand-sensitive, set the strand of `aln` to `*`.

Write the sequence of steps required to calculate counts as a simple function, and calculate counts on each file. On Mac or Linux, can you easily parallelize this operation?

Solution: We discover the location of files using standard *R* commands:

```
> fls <- list.files(file.path(bigdata(), "bam"), ".bam$", full=TRUE)
> names(fls) <- sub("_.*", "", basename(fls))
```

Use `readGappedAlignments` to input data from one of the files, and standard *R* commands to explore the data.

```
> ## input
> aln <- readGappedAlignments(fls[1])
> xtabs(~rname + strand, as.data.frame(aln))
```

```
       strand
rname      +    -
  chr3L 5402 5974
  chrX  2278 2283
```

To count overlaps in regions defined in a previous exercise, load the regions.

```
> data(ex)              # from an earlier exercise
```

Many RNA-seq protocols are not strand aware, i.e., reads align to the plus or minus strand regardless of the strand on which the corresponding gene is encoded. Adjust the strand of the aligned reads to indicate that strand is not known.

```
> strand(aln) <- "*"   # protocol not strand-aware
```

For simplicity, we are interested in reads that align to only a single gene. Count the number of genes a read aligns to...

```
> hits <- countOverlaps(aln, ex)
> table(hits)
```

```
hits
    0     1     2
  772 15026   139
```

and reverse the operation to count the number of times each region of interest aligns to a uniquely overlapping alignment.

```
> cnt <- countOverlaps(ex, aln[hits==1])
```

A simple function for counting reads is

```
> counter <-
+     function(filePath, range)
+ {
+     aln <- readGappedAlignments(filePath)
+     strand(aln) <- "*"
+     hits <- countOverlaps(aln, range)
+     cnt <- countOverlaps(range, aln[hits==1])
+     names(cnt) <- names(range)
+     cnt
+ }
```

This can be applied to all files using `sapply`

```
> counts <- sapply(fls, counter, ex)
```

The counts in one BAM file are independent of counts in another BAM file. This encourages us to count reads in each BAM file in parallel, decreasing the length of time required to execute our program. On Linux and Mac OS, a straight-forward way to parallelize this operation is:

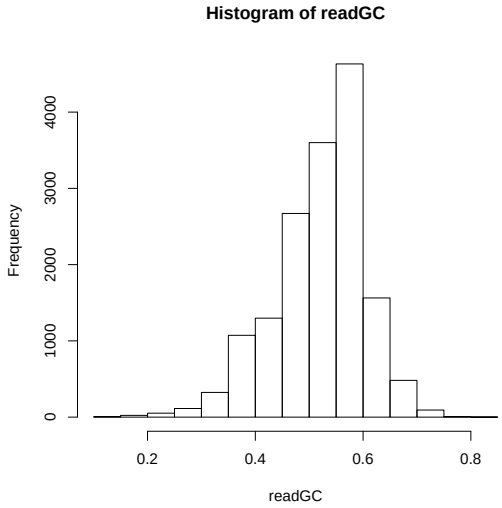

Figure 3: GC content in aligned reads

```
> if (require(parallel))
+     simplify2array(mclapply(fls, counter, ex))
```

The *GappedAlignments* class inputs only some of the fields of a BAM file, and may not be appropriate for all uses. In these cases the `scanBam` function in *Rsamtools* provides greater flexibility. The idea is to view BAM files as a kind of data base. Particular regions of interest can be selected, and the information in the selection restricted to particular fields. These operations are determined by the values of a *ScanBamParam* object, passed as the named `param` argument to `scanBam`.

Exercise 9
Consult the help page for `ScanBamParam`*, and construct an object that restricts the information returned by a* `scanBam` *query to the aligned read DNA sequence. Your solution will use the* `what` *parameter to the* `ScanBamParam` *function.*

Use the `ScanBamParam` *object to query a BAM file, and calculate the GC content of all aligned reads. Summarize the GC content as a histogram (Figure 3).*

Solution:

```
> param <- ScanBamParam(what="seq")
> seqs <- scanBam(fls[[1]], param=param)
> readGC <- gcFunction(seqs[[1]][["seq"]])
> hist(readGC)
```

5 RNA-seq

5.1 Varieties of RNA-seq

RNA-seq experiments typically ask about differences in representation of genes or other features across experimental groups. The analysis of designed experiments is of course statistical, and hence an ideal task for *R*. The overall structure of the analysis, with tens of thousands of features and tens of samples, is also reminiscent of microarray analysis; one might hope that insights from the microarray domain will apply, at least conceptually, to the analysis of RNA-seq experiments.

The most straight-forward RNA-seq experiments quantify abundance of known gene models. The known models are derived from reference databases, reflecting the accumulated wisdom of the community responsible for the data. The 'knownGenes' track of the UCSC genome browser represents one source of data. It contains, for each gene, the transcripts and exons that are thought through experimental or computational approaches to exist. The *GenomicFeatures* package allows ready access to this information, as we have seen. The data base of known genes is coupled with high throughput sequence data by counting or otherwise estimating the number of reads associated with each gene.

A more ambitious approach to RNA-seq attempts to identify novel transcripts. This requires that sequenced reads be assembled into contigs that, presumably, correspond to expressed transcripts that are then located in the genome. Regions identified in this way may correspond to known transcripts, to novel organization of known exons (e.g., through alternative splicing), or to completely novel constructs. We will not address the identification of completely novel transcripts here, but note that, having quantified transcript abundances in several samples, one is still interested in the analysis of designed experiments – do transcript abundances, novel or otherwise, differ between experimental groups?

Bioconductor packages play a role in several stages of an RNA-seq analysis. The *GenomicRanges* infrastructure we have already been exposed to can be effectively employed to quantify known exon or transcript abundances. Quantified abundances are in essence a matrix of counts, with rows representing features and columns samples. The *edgeR* [16] and *DESeq* [1] packages facilitate analysis of this data in the context of designed experiments, and are appropriate when the questions of interest involve between-sample comparisons of relative abundance. The *DEXSeq* package extends the approach in *edgeR* and *DESeq* to ask about within-gene, between group differences in exon use, i.e., for a given gene, do groups differ in their exon use?

5.2 Data preparation

Counting reads aligning to genes An essential step is to arrive at some measure of gene representation amongst the aligned reads. A straight-forward and commonly used approach is to count the number of times a read overlaps exons. Nuance arises when a read only partly overlaps an exon; when two exons overlap (and hence a read appears to be 'double counted'); when reads are aligned with gaps, and the gaps are inconsistent with known exon boundaries; etc. The `summarizeOverlaps` function in the *GenomicRanges* package provides

facilities for implementing different count strategies, using the argument `mode` to determine the counting strategy. The result of `summarizeOverlaps` can easily be used in subsequent steps of an RNA-seq analysis.

Software other than *R* can also be used to summarize count data. An important point is that the desired input is often raw count data, rather than normalized (e.g., reads per kilobase of (gene) model per million mapped reads) values. This is because counts allow information about uncertainty of estimates to propagate to later stages in the analysis.

Object creation and filtering The following exercise illustrates key steps in data preparation.

Exercise 10
The SeattleIntro2011Data package contains a data set `counts` *with pre-computed count data. Use the* `data` *command to load it. Create a variable* `grp` *to define the groups associated with each column, using the column names as a proxy for more authoritative metadata.*

Create a `DGEList` *object (defined in the edgeR package) from the count matrix and group information. Calculate relative library sizes using the* `calcNormFactors` *function.*

A lesson from the microarray world is to discard genes that cannot be informative (e.g., because of lack of variation). Filter reads to remove those that are represented at less than 1 per million mapped reads, in fewer than 2 samples.

Use `plotMDS` *on the filtered reads to perform multi-dimensional scaling. Interpret the resulting plot.*

Solution: Here we load the data (a matrix of counts) and create treatment group names from the column names of the counts matrix.

```
> data(counts)
> dim(counts)
```

```
[1] 14470     7
```

```
> grp <- factor(sub("[1-4].*", "", colnames(counts)),
+               levels=c("untreated", "treated"))
```

We use the *edgeR* package, creating a *DGEList* object from the count and group data. The `calcNormFactors` function estimates relative library sizes for use as offsets in the generalized linear model.

```
> library(edgeR)
> dge <- DGEList(counts, group=grp)
> dge <- calcNormFactors(dge)
```

To filter reads, we scale the counts by the library sizes and express the results on a per-million read scale. We require that the gene be represented at a frequency of at least 1 read per million mapped in two or more of each sample, and use this criterion to subset the *DGEList* instance.

```
> m <- 1e6 * t(t(dge$counts) / dge$samples$lib.size)
> ridx <- rowSums(m > 1) >= 2
> table(ridx)   # number filtered / retained $
```

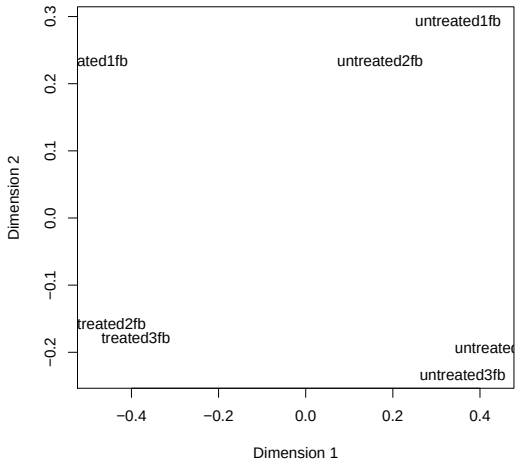

Figure 4: MDS plot of lanes from the Pasilla data set.

```
ridx
FALSE  TRUE
 6476  7994
```

```
> dge <- dge[ridx,]
```

Multi-dimensional scaling takes data in high dimensional space (in our case, the dimension is equal to the number of genes in the filtered *DGEList* instance) and reduces it to fewer (e.g., 2) dimensions, allowing easier assessment. The plot is shown in Figure 4; that the samples separate into distinct groups provides some reassurance that the data differ according to treatment. Nonetheless, there appears to be considerable heterogeneity within groups. Any guess, perhaps from looking at the quality report generated early, what the within-group differences are due to?

```
> plotMDS.DGEList(dge)
```

```
Using grid search to estimate tagwise dispersion.
```

5.3 Differential representation

RNA-seq differential representation experiments, like classical microarray experiments, consist of a single statistical design (e.g, comparing expression of samples assigned to 'Treatment' versus 'Control' groups) applied to each feature for which there are aligned reads. While one could naively perform simple tests (e.g., t-tests) on all features, it is much more informative to identify important aspects of RNAseq experiments, and to take a flexible route through this part of the work flow. Key steps involve formulation of a model matrix to capture the experimental design, estimation of a test static to describe differences

between groups, and calculation of a P value or other measure as a statement of statistical significance.

Experimental design In R, an experimental design is specified with the `model.matrix` function. The function takes as its first argument a `formula` describing the independent variables and their relationship, and as a second argument a `data.frame` containing the (phenotypic) data that the formula describes. A simple formula might read `~ 1 + grp`, which says that the response is a linear function involving an intercept (1) plus a term encoded in the variable `grp`. If (as in our case) `grp` is a factor, then the first coefficient (column) of the model matrix corresponds to the first level of `grp`, and subsequent terms correspond to *deviations* of each level from the first. If `grp` were *numeric* rather than *factor*, the formula would represent linear regressions with an intercept. Formulas are very flexible, allowing representation of models with one, two, or more factors as main effects, models with or without interaction, and with nested effects.

Exercise 11
To be more concrete, use the `model.matrix` function and a formula involving `grp` to create the model matrix for our experiment.

Solution: Here is the experimental design; it's worth discussing with your neighbor the interpretation of the `design` instance.

```
> (design <- model.matrix( ~ grp ))

  (Intercept) grptreated
1           1          1
2           1          1
3           1          1
4           1          0
5           1          0
6           1          0
7           1          0
attr(,"assign")
[1] 0 1
attr(,"contrasts")
attr(,"contrasts")$grp
[1] "contr.treatment"
```

The coefficient (column) labeled 'Intercept' corresponds to the first level of `grp`, i.e., 'untreated'. The coefficient 'grptreated' represents the deviation of the treated group from untreated. Eventually, we will test whether the second coefficient is significantly different from zero, i.e., whether samples with a '1' in the second column are, on average different from samples with a '0'. On the one hand, use of `model.matrix` to specify experimental design implies that the user is comfortable with something more than elementary statistical concepts, while on the other it provides great flexibility in the type of experiment that can be analyzed.

Negative binomial error RNA-seq count data are often described by a negative binomial model. This model includes a 'dispersion' parameter that describes biological variation beyond the expectation under a Poisson model. The simplest approach estimates a dispersion parameter from all the data. The estimate needs to be conducted in the context of the experimental design, so that variability between experimental factors is not mistaken for variability in counts. The square root of the estimated dispersion represents the coefficient of variation between biological samples.

```
> dge <- estimateCommonDisp(dge, design)
> sqrt(dge$common.dispersion)
```

```
[1] 0.78
```

This approach assumes that a common dispersion parameter is shared by all genes. A different approach, appropriate when there are more samples in the study, is to estimate a dispersion parameter that is specific to each tag (using `estimateTagwiseDisp` in the *edgeR* package). As another alternative, Anders and Huber [1] note that dispersion increases as the mean number of reads per gene decreases. One can estimate the relationship between dispersion and mean using `estimateGLMTrendedDisp` in *edgeR*, using a fitted relationship across all genes to estimate the dispersion of individual genes. Because in our case sample sizes (biological replicates) are small, gene-wise estimates of dispersion are likely imprecise. One approach is to moderate these estimates by calculating a weighted average of the gene-specific and common dispersion; `estimateGLMTagwiseDisp` performs this calculation, requiring that the user provide the weight *a priori*.

Differential representation The final steps in estimating differential representation are to fit the full model; to perform the likelihood ratio test comparing the full model to a model in which one of the coefficients has been removed; and to summarize, from the likelihood ratio calculation, genes that are most differentially represented. The result is a 'top table' whose row names are the Flybase gene ids used to label the elements of the `ex` *GRangesList*.

Exercise 12
Use `glmFit` to fit the general linear model. This function requires the input data `dge`, the experimental design `design`, and as estimate of dispersion.

Use `glmLRT` to form the likelihood ratio test. This requires the original data `dge` and the fitted model from the previous part of this question. Which coefficient of the design matrix do you wish to test?

Finally, create a 'top table' of differentially represented genes using `topTags`.

Solution: Here we fit a glm to our data and experimental design, using the common dispersion estimate.

```
> fit <- glmFit(dge, design, dispersion=dge$common.dispersion)
```

The fit can be used to calculate a likelihood ratio test, comparing the full model to a reduced version with the second coefficient removed. The second coefficient captures the difference between treated and untreated groups, and the likelihood ratio test asks whether this term contributes meaningfully to the overall fit.

```
> lrTest <- glmLRT(dge, fit, coef=2)
```

Here `topTags` function summarizes results across the experiment.

```
> (tt <- topTags(lrTest))
```

```
Coefficient:  grptreated
            logConc logFC LR P.Value   FDR
FBgn0039155   -9.6  -4.7 20 9.2e-06 0.053
FBgn0085359  -12.3  -4.8 19 1.4e-05 0.053
FBgn0024288  -12.4  -4.7 18 2.0e-05 0.053
FBgn0039827  -10.6  -4.2 17 4.1e-05 0.082
FBgn0034434  -11.4  -4.0 15 9.8e-05 0.156
FBgn0033764  -12.1   3.5 14 1.5e-04 0.201
FBgn0034736  -11.0  -3.5 12 4.3e-04 0.496
FBgn0033065  -13.0   3.2 12 5.4e-04 0.533
FBgn0037290  -12.0   3.1 12 6.6e-04 0.533
FBgn0035189  -11.0   3.0 12 6.7e-04 0.533
```

As a 'sanity check', summarize the original data for the first several probes

```
> sapply(rownames(tt$table)[1:4],
+        function(x) tapply(counts[x,], grp, mean))
```

```
           FBgn0039155 FBgn0085359 FBgn0024288 FBgn0039827
untreated        1576       118.2       102.5         554
treated            64         4.7         4.3          31
```

6 ChIP-seq

6.1 Varieties of ChIP-seq

ChIP-seq experiments combine chromosome immunopreciptation (ChIP) with sequence analysis. The idea is that the ChIP protocol enriches genomic DNA for regions of interest, e.g., sties to which transcription factors are bound. The regions of interest are then subject to high throughput sequencing, the reads aligned to a reference genome, and the location of mapped reads ('peaks') interpretted as indicators of the ChIP'ed regions. Reviews include those by Park and collegues [14, 6]; there is a large collection of peak-calling software, some features of which are summarized in Pepke et al. [15].

Initial stages in a ChIP-seq analysis differ from RNA-seq in several important ways. The ChIP protocol is more complicated and idiosyncratic than RNA-seq protocols, and the targets of ChIP more variable in terms of sequence and other characteristics. Both RNA- and ChIP-seq involve aligning reads to a reference genome, ChIP-seq requires that the aligned reads be processed to identify peaks, rather than simply counted in known gene regions.

Many early ChIP-seq studies focused on characterizing one or a suite of transcription binding sites across a small number of samples from one or two groups. The main challenge initially was to develop efficient peak-calling software, often tailored to the characteristics of the peaks of interest (narrow and well-defined, e.g., CTCF, vs. broad histone marks). More comprehensive studies, e.g., in *Drosophila* [7] drawn from multiple samples, e.g., in the ENCODE project [13]. Decreasing sequence costs and better experimental and data analytic protocols mean that these larger-scale studies are increasingly accessible to individual investigators. Peak-calling in this study represents an initial step, but significant analytic challenges are to interpret analyses derived from multiple samples.

Bioconductor packages play a role in several stages of a ChIP-seq analysis. The *ShortRead* package can provide a quality assessment report of reads. Following alignment, the *chipseq* package can be used, in conjunction with *ShortRead* and *GenomicRanges*, to identify enriched regions in a statistically informed and flexible way. The *ChIPpeakAnno* package assists in annotating peaks in terms of known genes and other genomic features. Pattern matching in *Biostrings*, and specialized packages such as *MotIV* can assist in motif identification. Additional packages summarized in the 'ChIPseq' BiocViews term provide diverse approaches to peak identification and analysis.

Our attention here is limited to identifying and annotating peaks. We start with a work flow using third-party tools, then re-iterate key components in *Bioconductor*.

6.2 Data preparation

In this section we use data from the ENCODE project to illustrate a typical ChIP-seq work flow. The data is from GEO accession GSE30263, representing ENCODE CTCF binding sites. CTCF is a zinc finger transcription factor. It is a sequence specific DNA binding protein that functions as an insulator, blocking enhancer activity, and possibly the spread of chromatin structure. The original analysis involved Illumina ChIP-seq and matching 'input' lanes of 1 or 2 replicates from many cell lines. The accession includes BAM files of aligned

reads, in addition to tertiary files summarizing identified peaks. We focus on 15 cell lines aligned to hg19.

The main computational stages in the original work flow invovled alignment using Bowtie, followed by peak identification using an algorithm ('HotSpots', [17]) originally developed for lower-throughput methodologies.

Initial quality assessment

Exercise 13
The SeattleIntro2011Data package contains a quality assessment report generated from the BAM files. View this report. Are there indications of batch or other systematic effects in the data?

Solution: Here we visit the QA report.

```
> rpt <- system.file("GSE30263_qa_report",
+            package="SeattleIntro2011")
> if (interactive())
+     browseURL(rpt)
```

Samples 1-15 correspond to replicate 1, 16-26 to replicate 2, and 27 through 41 the 'input' samples. Notice that overall nucleotide frequencies fall into three distinct groups, and that samples 1-11 differ from the other input samples. The 'Depth of Coverage' portion of the report is particularly relevant for an early assessment of ChIP-seq experiments.

Peak calling with MACS We chose to perform an initial analysis with MACS [18]. MACS is one of the earlier peak calling implementations; it is well-described, based on reasonable principles, and relatively widely used. MACS uses information about tags aligned to the plus and minus strand, allows for Poisson-distributed local biases in peak density, and incorporates an appropriately scaled input lane when available. We used MACS version 1.4.1 20110627, with the following command line invokation

```
macs14 -t cellLineCTCP.bam -c cellLineControl.bam -n cellLine
```

The output from MACS include `.bed` files with the locations of all identified summits, and tab-delimited files (labelled `.xls`) with the genomic coordinates (start, end) of each peak. These commands were evaluated for all replicates of all cell lines aligned to hg19.

We collated the output files with a goal of enumerating all peaks from all files, but collapsing the coordinates of sufficiently similar peaks to a common location. To do this, we created ranges of width 40bp, centered on each peak. We identified overlapping ranges, over all samples, and collapsed these into a single synthetic peak with width equal to the bounds of the overlapping ranges. We then re-organized the information on called peaks in to a matrix. Rows of the matrix represent distinct peaks. Columns of the matrix represent samples. Entries in the matrix are the number of reads supporting the corresponding peak and column, from the MACS estimate. The data is represented in a *SummarizedExperiment* object; the script performing these operations is available; here we load the data as an *R* object `stam` (an abbreviation for the lab generating the data).

```
> scrpt <- system.file("script", "chipseq_make_stam.R",
+                      package="SeattleIntro2011Data")
> stamFile <-
+     system.file("data", "stam.Rda", package="SeattleIntro2011Data")
> load(stamFile)
```

Data exploration

Exercise 14
Explore `stam`*. Tabulate the number of peaks represented 1, 2, . . . 26 times. We expect replicates to have similar patterns of peak representation; do they?*

Solution: Load the data and display the *SummarizedExperiment* instance. The `colData` summarizes information about each sample, the `rowData` about each peak. Use `xtabs` to summarize `Replicate` and `CellLine` representaiton within `colData(stam)`.

```
> load(stamFile)
> stam

class: SummarizedExperiment
dim: 222354 26
assays(3): counts pvalues summitPerPeak
rownames: NULL
rowData values names(0):
colnames(26): Ag04449_1 Ag04450_1 ... Hpaf_2 Hpf_2
colData names(5): SummitFile PeaksFile Genome CellLine Replicate

> head(colData(stam), 3)

DataFrame with 3 rows and 5 columns
                  SummitFile         PeaksFile      Genome    CellLine
                 <character>       <character> <character> <character>
Ag04449_1 Ag04449_summits.bed Ag04449_peaks.xls        hg19     Ag04449
Ag04450_1 Ag04450_summits.bed Ag04450_peaks.xls        hg19     Ag04450
Ag09309_1 Ag09309_summits.bed Ag09309_peaks.xls        hg19     Ag09309
          Replicate
           <factor>
Ag04449_1         1
Ag04450_1         1
Ag09309_1         1

> head(rowData(stam), 3)

GRanges with 3 ranges and 0 elementMetadata values:
      seqnames           ranges strand
         <Rle>        <IRanges>  <Rle>
  [1]     chr1 [ 16223,  16279]      *
  [2]     chr1 [ 91510,  91546]      *
  [3]     chr1 [104970, 105018]      *
  ---
  seqlengths:
```

```
    chr1 chr10 chr11 chr12 chr13 chr14 ...  chr7  chr8  chr9  chrX  chrY  chrM
      NA    NA    NA    NA    NA    NA ...    NA    NA    NA    NA    NA    NA
```

```
> xtabs(~Replicate + CellLine, colData(stam))
```

```
         CellLine
Replicate Ag04449 Ag04450 Ag09309 Ag09319 Ag10803 Aoaf Hasp Hbmec Hcfaa Hcpe
        1       1       1       1       1       1    1    1     1     1    1
        2       1       0       1       1       1    1    0     1     0    1
         CellLine
Replicate Hee Hmf Hpaf Hpf Hrpe
        1   1   1    1   1    1
        2   1   1    1   1    0
```

Extract the `counts` matrix from the assays. This is a standard *R matrix*. Test which matrix elements are non-zero, tally these by row, and summarize the tallies. This is the number of times a peak is detected, across each of the samles

```
> m <- assays(stam)[["counts"]] > 0 # peak detected...
> table(rowSums(m))
```

```
     1      2      3      4      5      6      7      8      9     10     11
117773  19783   8386   4934   3653   3053   2611   2422   2125   1976   1912
    12     13     14     15     16     17     18     19     20     21     22
  1820   1702   1723   1607   1595   1609   1571   1611   1586   1699   1931
    23     24     25     26
  2251   3008   4776  25237
```

To explore similarity between replicates, extract the matrix of counts. Select just those peaks that are common to all samples, and from these use the `varFilter` function from the *genefilter* package to select the 5% most variable peaks; we choose these as they contain the most statistical inforamtion. Note that we select variable peaks independent of replciate; the scaling is very rough. The results are in Figure 5.

```
> ## Cell lines with 2 replicates, peaks in all samples
> cidx = with(colData(stam), CellLine %in% CellLine[Replicate==2])
> m <- assays(stam)[["counts"]][,cidx]
> ridx <- rowSums(m > 0) == ncol(m)
> m <- m[ridx,]
> ## filter -- 5% most variable
> library(genefilter)
> m1 <- varFilter(m, var.cutoff=.95)
> colscale <-mean(colSums(m1)) / colSums(m1)[col(m1)]
> m1 <-  m1 * colscale
> heatmap(m1, Rowv=NA, labRow=NA)
```

6.3 Peak calling with *R / Bioconductor*

The following illustrates basic ChIP-seq work flow components in *R*. It is likely that these would be used either in an exploratory way, or as foundations for developing work flows tailored to particular ChIP experiments.

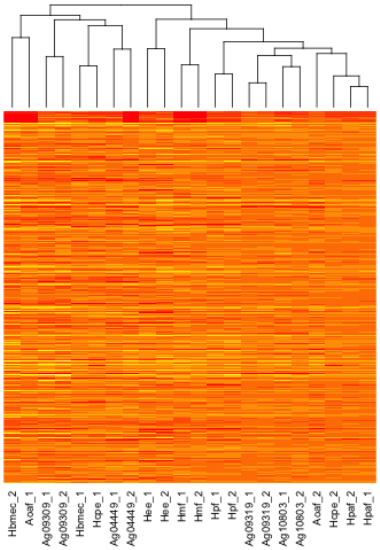

Figure 5: Replicates clustered on 5% most variable peaks

Data input and pre-processing The work flow starts with data input. We suppose an available `bamFile`, with a `ScanBamParam` object `param` defined to select regions we are interested in.

```
> aln <- readGappedAlignments(bamFile, param=param)
> seqlevels(aln) <-  names(bamWhich(param))
> aln <- as(aln, "GRanges")
```

We use `readGappedAlignments` followed by coercion to a *GRanges* object as a convenient way to retrieve a minimal amount of data from the BAM file, and to manage reads whose alignments include indels; `Rsamtools::scanBam`. is a more flexible alternative. The `seqlevels` are adjusted to contain just the levels we are interested in, rather than all levels in the BAM file (the default returned by `readGappedAlignments`).

Sequence work flows typically filter reads to remove those that are optical duplicates or otherwise flagged as invalid by the manufacturer. Many work flows do not handle reads aligning to multiple locations in the genome. ChIP-seq experiments often eliminate reads that are duplicated in the sense that more than one read aligns to the same chromosome, strand, and start position; this acknowledges artifacts of sample preparation. These filters are handled by different stages in a typical work flow – flagging optical duplicates and otherwise suspect reads by the manufacturer or upstream software (illustrated in an exercise, below); discarding multiply aligning reads by the aligner (in our case, using the |-m| and |-n| options in *Bowtie*); and discarding duplicates as a pre-processing step. Simple alignment de-duplication is

```
> aln <- aln[!duplicated(aln)]
```

It is common to estimate fragment length (e.g., via the 'correlation' method [8], implemented in the *chipseq* package) and extend the 5' tags by the estimated length.

```
> fraglen <- estimate.mean.fraglen(aln, method="correlation")
> aln <- resize(aln, width=fraglen)
```

The end result can be summarized as a 'coverage vector' describing the number of (extended) reads at each location in the genome; a run length encoding is an efficient representation of this.

```
> coverage(aln)
```

These pre-processing (this might be a misnomer) steps can be summarized as a simple work-flow.

```
> chipPreprocess <- function(bamFile, param) {
+     aln <- readGappedAlignments(bamFile, param = param)
+     seqlevels(aln) <- names(bamWhich(param))
+     aln <- as(aln, "GRanges")
+     aln <- aln[!duplicated(aln)]
+     fraglen <- estimate.mean.fraglen(aln, method = "correlation")
+     aln <- resize(aln, width = fraglen)
+     coverage(aln)
+ }
```

Peak identification The coverage vector is a very useful representation of the data, and numerous peak discovery algorithms can be implemented on top if it. The *chipseq* package implements a straight-forward approach. The first step uses the distribution of singleton and doubleton islands to estimate a background Poisson noise distribution, and hence to identify a threshold island elevation above which peaks can be called at a specified false discovery rate.

```
> cutoff <- round(peakCutoff(cvg, fdr.cutoff=0.001))
```

Peaks are easily identified using `slice`

```
> slice(cvg, lower = cutoff)
```

resulting in a peak-finding work flow

```
> findPeaks <- function(cvg) {
+     cutoff <- round(peakCutoff(cvg, fdr.cutoff = 0.001))
+     slice(cvg, lower = cutoff)
+ }
```

Exercise 15
Walk through the work flow, from BAM file to called peaks, using the provided BAM files. These are from the Ag09319 cell line, CTCF replicate 1 and input lanes, filtered to include only reads from chromosome 6. Compare peaks found in the ChIP and Input lanes, and in the MACS analysis. It is possible to pick up the analysis after pre-processing by loading the `cvgs` *object. It can be very helpful to explore the data along the way; see the chipseq vignette for ideas.*

Solution: Specify the location of the BAM files, and the location where the coverage vectors will be saved. Storing the coverage vectors represents a check-pointing strategy, making it easy to resume an analysis if interrupted.

```
> library(GenomicRanges)
> bamDir <- character()                    # TODO: specify location
> bamFiles <- c(ChIP=file.path(bamDir,
+                 "wgEncodeUwTfbsAg09319CtcfStdAlnRep1.bam"),
+               Input=file.path(bamDir,
+                 "wgEncodeUwTfbsAg09319InputStdAlnRep1.bam"))
> stopifnot(all(file.exists(bamFiles)))
> cvgsSaveFile <- character()              #TODO: specify location
```

Create a `ScanBamParam` object specifying the regions of interest and other restrictions on reads to be input.

```
> chr6len <- scanBamHeader(bamFiles)[[1]][["targets"]][["chr6"]]
> param <- ScanBamParam(which=GRanges("chr6", IRanges(1, chr6len)),
+                       what=character(),
+                       flag=scanBamFlag(isDuplicate=FALSE,
+                         isValidVendorRead=TRUE))
```

Process each BAM file using `lapply`, and save the result.

```
> cvgs <- lapply(bamFiles, chipPreprocess, param)
> save(cvgs, cvgsSaveFile)
```

Load the saved coverage file, and find peaks using the simple approach outlined above.

```
> library(chipseq)
> cvgsFile <- system.file("data", "chipseq_chr6_cvgs.Rda",
+                        package="SeattleIntro2011Data")
> stopifnot(file.exists(cvgsFile))
> load(cvgsFile)                           # previously saved
> peaks <- lapply(cvgs, findPeaks)
```

Compare the peaks using *GRanges* commands (e.g. convert the peaks to `IRanges` instances and use `countOverlaps` to identify peaks in common between the ChIP and Input lanes), and the `diffPeakSummary` function from the *chipseq* package. Compare the peaks to those found in [].

```
> chip <- as(peaks[["ChIP"]][["chr6"]], "IRanges")
> inpt <- as(peaks[["Input"]][["chr6"]], "IRanges")
> table(countOverlaps(inpt, chip))

  0   1   2
635  19   3
```

```
> stamFile <- system.file("data", "stam.Rda",
+                 package="SeattleIntro2011Data")
> load(stamFile)
> stam0 <- stam[,"Ag09319_1"]
> idx <- seqnames(rowData(stam0)) == "chr6" &
+            assays(stam0)[["counts"]] != 0
> rng <- ranges(rowData(stam0))[as.logical(idx)]
> table(countOverlaps(chip, rng))

   0    1    2
2218 3339    1
```

6.4 Annotation

Exercise 16

Annotating ChIP peaks is straight-forward. Load the ENCODE summary data, select the peaks found in all samples, and use the center of these peaks as a proxy for the true ChIP binding site. Use the transcript db data base for the UCSC Known Genes track of hg19 as a source for transcripts and transcription start sites (TSS). Use `nearest` *to identify the TSS that is nearest each peak, and calculate the distance between the peak and TSS; measure distance taking account of the strand of the transcript, so that peaks 5' of the TSS have negative distance. Summarize the locations of the peaks relative to the TSS.*

Solution: Read in the ENCODE ChIP peaks for all cell lines.

```
> stamFile <-
+     system.file("data", "stam.Rda", package="SeattleIntro2011Data")
> load(stamFile)
```

Identify the rows of `stam` that have non-zero counts for all cell lines:

```
> ridx <- rowSums(assays(stam)[["counts"]] > 0) == ncol(stam)
```

Select the center of the ranges of these peaks, as a proxy for the ChIP binding site:

```
> peak <- resize(rowData(stam)[ridx], width=1, fix="center")
```

Obtain the TSS from the *TxDb.Hsapiens.UCSC.hg19.knownGene* using the `transcripts` function to extract coordinates of each transcript, and `resize` to a width of 1 for the TSS; does this do the right thing for transcripts on the plus and on the minus strand?

```
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)
> tx <- transcripts(TxDb.Hsapiens.UCSC.hg19.knownGene)
> tss <- resize(tx, width=1)
```

The `nearest` function returns the index of the nearest `subject` to each `query` element; the distance between peak and nearest TSS is thus

```
> idx <- nearest(peak, tss)
> dist <- (start(peak) - start(tss)[idx]) *
+     as.numeric(ifelse(strand(tss)[idx] == "+", 1, -1))
```

Here we summarize the distances as a simple table and a histogram; the histogram is in Figure 6.

```
> table(sign(dist))

   -1     0     1
12239     3 12995
```

```
> plt <- densityplot(log10(abs(dist)) * sign(dist), plot.points=FALSE,
+           main="CTCF Ag09319_1", xlab="log10 Distance to Nearest TSS")
```

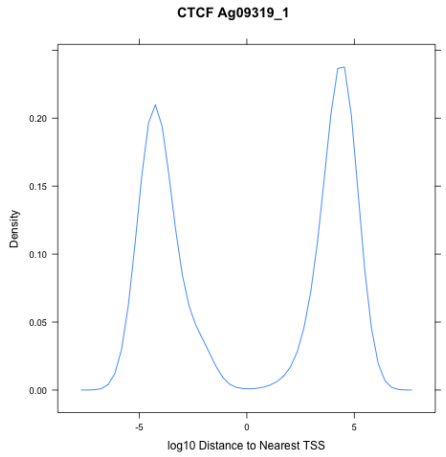

Figure 6: Distance to nearest TSS amongst conserved peaks

Exercise 17
As an additional exercise, extract the sequences of all conserved peaks on 'chr6'. Do this using the BSgenome.Hsapiens.UCSC.hg19 package and `getSeq` *function. What strategies are available for motif discovery?*

Solution: The following code requires an additional package, and is not evaluated.

```
> library(BSgenome.Hsapiens.UCSC.hg19)
> pk6 <- peak[seqnames(peak) == "chr6"]
> seqs <- getSeq(Hsapiens, resize(pk6, 20, "center"))
```

6.5 Additional *Bioconductor* resources

There are many additional *Bioconductor* resources for working with ChIP-seq data, including advanced methods for peak identification, annotation, and motif discovery. These are summarized as the ChIP-seq BiocViews term.

7 Annotation

7.1 Major types of annotation in *Bioconductor*

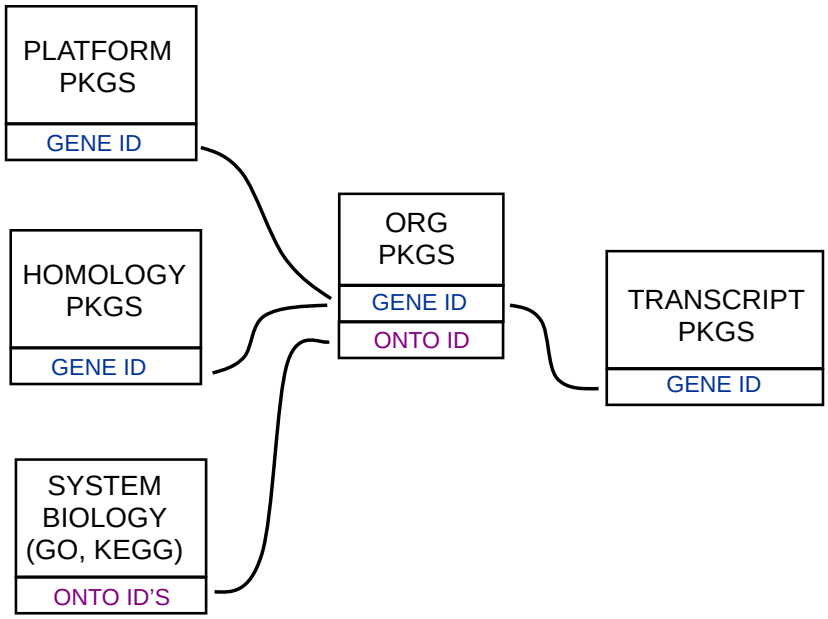

Figure 7: Annotation Packages: the big picture

Gene centric *AnnotationDbi* packages:

- Organism level: e.g. *org.Mm.eg.db*.
- Platform level: e.g. *hgu133plus2.db*, *hgu133plus2.probes*, *hgu133plus2.cdf*.
- Homology level: e.g. *hom.Dm.inp.db*.
- System-biology level: *GO.db* or *KEGG.db*.

Genome centric *GenomicFeatures* packages:

- Transcriptome level: e.g. *TxDb.Hsapiens.UCSC.hg19.knownGene*
- Generic genome features: Can generate via *GenomicFeatures*

biomaRt:

- Query web-based 'biomart' resource for genes, sequence, SNPs, and etc.

7.2 Organism level packages

An organism level package (aka "org package") is organized around a central gene identifier (e.g. Entrez Gene id) and contains a collection of mappings between this central id and other kinds of ids (e.g. GenBank or Uniprot accession number, RefSeq id, etc.). The name of an org package is always of the form *org.<Ab>.<efg>.db* (e.g. *org.Sc.sgd.db*) where *<Ab>* is a 2-letter abbreviation of the organism (e.g. *Sc* for *Saccharomyces cerevisiae*) and *<efg>* is an abbreviation (in lower-case) describing the type of central gene id (e.g. *sgd* for gene ids assigned by the Saccharomyces Genome Database people or *eg* for Entrez Gene ids).

The document of reference for using org packages is the *How to use the ".db" annotation packages* vignette in the *AnnotationDbi* package (org packages are only one type of ".db" annotation packages).

Like almost all annotation packages in *Bioconductor*, the ".db" annotation packages are updated every 6 months (i.e. at every new *Bioconductor* release).

Exercise 18

What is the name of the org package for Drosophilla*? Load it.*

Use `ls("package:<pkgname>")` *to display the list of all symbols defined in this package. Explore a few of the symbols by looking at their man page, at their class, and by extracting a few samples from them with* `sample(map, 5)`.

Turn a map into a data frame with `toTable` *(use* `head` *to only display the first rows). What are the left keys? What are the right keys?*

Most maps can be reversed with `revmap`*. Reverse a map and extract a few samples from the reversed map.*

Note that reversing a map does NOT switch the left and right keys. You can check this with the `Lkeys` *and* `Rkeys` *accessors.*

Solution:

```
> library(org.Dm.eg.db)
> ls('package:org.Dm.eg.db')
```

```
 [1] "org.Dm.eg"                "org.Dm.eg.db"
 [3] "org.Dm.egACCNUM"          "org.Dm.egACCNUM2EG"
 [5] "org.Dm.egALIAS2EG"        "org.Dm.egCHR"
 [7] "org.Dm.egCHRLENGTHS"      "org.Dm.egCHRLOC"
 [9] "org.Dm.egCHRLOCEND"       "org.Dm.egENSEMBL"
[11] "org.Dm.egENSEMBL2EG"      "org.Dm.egENSEMBLPROT"
[13] "org.Dm.egENSEMBLPROT2EG"  "org.Dm.egENSEMBLTRANS"
[15] "org.Dm.egENSEMBLTRANS2EG" "org.Dm.egENZYME"
[17] "org.Dm.egENZYME2EG"       "org.Dm.egFLYBASE"
[19] "org.Dm.egFLYBASE2EG"      "org.Dm.egFLYBASECG"
[21] "org.Dm.egFLYBASECG2EG"    "org.Dm.egFLYBASEPROT"
[23] "org.Dm.egFLYBASEPROT2EG"  "org.Dm.egGENENAME"
[25] "org.Dm.egGO"              "org.Dm.egGO2ALLEGS"
[27] "org.Dm.egGO2EG"           "org.Dm.egMAP"
[29] "org.Dm.egMAP2EG"          "org.Dm.egMAPCOUNTS"
[31] "org.Dm.egORGANISM"        "org.Dm.egPATH"
[33] "org.Dm.egPATH2EG"         "org.Dm.egPMID"
```

```
[35] "org.Dm.egPMID2EG"         "org.Dm.egREFSEQ"
[37] "org.Dm.egREFSEQ2EG"       "org.Dm.egSYMBOL"
[39] "org.Dm.egSYMBOL2EG"       "org.Dm.egUNIGENE"
[41] "org.Dm.egUNIGENE2EG"      "org.Dm.egUNIPROT"
[43] "org.Dm.eg_dbInfo"         "org.Dm.eg_dbconn"
[45] "org.Dm.eg_dbfile"         "org.Dm.eg_dbschema"
```

```
> org.Dm.egUNIPROT
```

```
UNIPROT map for Fly (object of class "AnnDbBimap")
```

```
> class(org.Dm.egUNIPROT)
```

```
[1] "AnnDbBimap"
attr(,"package")
[1] "AnnotationDbi"
```

```
> sample(org.Dm.egUNIPROT, 5)
```

```
$`37966`
[1] "Q9W106"

$`38813`
[1] "Q9VS51"

$`8674073`
[1] "E1JHX0"

$`35394`
[1] "Q0E8N6" "Q9VIE2"

$`248335`
[1] NA
```

```
> head(toTable(org.Dm.egUNIPROT))
```

```
  gene_id uniprot_id
1   30970     Q8IRZ0
2   30970     Q95RP8
3   30971     Q95RU8
4   30972     Q9W5H1
5   30973     P39205
6   30975     Q24312
```

The left keys are the Entrez Gene ids and the right keys the Uniprot accession numbers. Note that for all the maps in an org package the left key is always the central gene id.

```
> revmap(org.Dm.egUNIGENE)
```

```
revmap(UNIGENE) map for Fly (object of class "AnnDbBimap")
```

```
> sample(revmap(org.Dm.egUNIGENE), 5)
```

```
$Dm.3278
[1] "40453"

$Dm.3947
[1] "33540"

$Dm.27050
[1] "38453"

$Dm.26756
[1] "34094"

$Dm.5661
[1] "43545"
```

```
> identical(Lkeys(org.Dm.egUNIGENE), Lkeys(revmap(org.Dm.egUNIGENE)))
```

```
[1] TRUE
```

Exercise 19

For convenience, `lrTest`, the DGEGLM object obtained in the previous section with `glmLRT`, has been included to the SeattleIntro2011Data package. Load it and create again the 'top table' of differentially represented genes with `topTags`.

Extract the Flybase gene ids from this table and map them to their corresponding Entrez Gene id (create a named character vector with names the Flybase gene ids and values the Entrez Gene ids).

Finally, add 2 columns to the `table` component of the TopTags object created previously: one for the Entrez Gene ids and one for their corresponding gene symbols.

Solution:

```
> library(org.Dm.eg.db)
> data(lrTest)
> tt <- topTags(lrTest)
```

```
> fbids <- rownames(tt$table)
> egids <- unlist(mget(fbids, revmap(org.Dm.egFLYBASE), ifnotfound=NA))
> egids
```

```
FBgn0039155 FBgn0039827 FBgn0034434 FBgn0034736 FBgn0035189 FBgn0085359
    "42865"     "43689"     "37219"     "37572"     "38124"   "2768869"
FBgn0033764 FBgn0000071 FBgn0024288 FBgn0037290
         NA     "40831"     "45039"     "40613"
```

Because `unlist` mangles names when the list has duplicated names, a better way to do this is:

```
> fbids <- rownames(tt$table)
> map <- org.Dm.egFLYBASE
> fbids <- intersect(mappedRkeys(map), fbids)
> egids <- as.character(revmap(map)[fbids])
> egids
```

```
FBgn0034434 FBgn0034736 FBgn0035189 FBgn0037290 FBgn0000071 FBgn0039155
    "37219"     "37572"     "38124"     "40613"     "40831"     "42865"
FBgn0039827 FBgn0024288 FBgn0085359
    "43689"     "45039"   "2768869"
```

To add the 2 columns to `tt$table`, we proceed in 3 steps: (1) merge the 2 mappings in a single data frame `anno0`, (2) align the rows in `anno0` with the rows in `tt$table` (by redordering them), and (3) `cbind` `tt$table` with the 2 new columns:

```
> eg2fb <- toTable(org.Dm.egFLYBASE[egids])
> eg2sym <- toTable(org.Dm.egSYMBOL[egids])
> (anno0 <- merge(eg2fb, eg2sym))

  gene_id  flybase_id  symbol
1 2768869 FBgn0085359 CG34330
2   37219 FBgn0034434    Rgk1
3   37572 FBgn0034736  CG6018
4   38124 FBgn0035189  CG9119
5   40613 FBgn0037290  CG1124
6   40831 FBgn0000071     Ama
7   42865 FBgn0039155   kal-1
8   43689 FBgn0039827  CG1544
9   45039 FBgn0024288 Sox100B

> (anno0 <- anno0[match(rownames(tt$table), anno0$flybase_id), ])

   gene_id  flybase_id  symbol
7    42865 FBgn0039155   kal-1
8    43689 FBgn0039827  CG1544
2    37219 FBgn0034434    Rgk1
3    37572 FBgn0034736  CG6018
4    38124 FBgn0035189  CG9119
1  2768869 FBgn0085359 CG34330
NA    <NA>        <NA>    <NA>
6    40831 FBgn0000071     Ama
9    45039 FBgn0024288 Sox100B
5    40613 FBgn0037290  CG1124

> anno <- cbind(tt$table, anno0[ , c("gene_id", "symbol")])
```

7.3 AnnotationDb objects and select

In the most recent version of AnnotationDbi, a new set of operations have been added that allow a simpler way of extracting identifier based annotations. All the annotation packages that support these new methods expose an object named exactly the same way as the package itself. These objects are collectively called *AnntoationDb* objects, with more specific class names such as *OrgDb* or *ChipDb* objects. The methods that can be applied to these objects are `cols`, `keys`, `keytypes` and `select`.

Exercise 20
Display the OrgDb object for the org.Dm.eg.db package.

Use the `cols` method to discover which sorts of annotations can be extracted from it. Is this the same as the result from the `keytypes` method?

use the `keys` method to extract some uniprot identifiers and then pass those keys in to the `select` method in such a way that you extract the gene symbol and KEGG pathway information for each.

Solution:

```
> org.Dm.eg.db
```

```
OrgDb object:
| DBSCHEMAVERSION: 2.1
| Db type: OrgDb
| package: AnnotationDbi
| DBSCHEMA: FLY_DB
| ORGANISM: Drosophila melanogaster
| SPECIES: Fly
| EGSOURCEDATE: 2011-Sep14
| EGSOURCENAME: Entrez Gene
| EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| CENTRALID: EG
| TAXID: 7227
| GOSOURCENAME: Gene Ontology
| GOSOURCEURL: ftp://ftp.geneontology.org/pub/go/godatabase/archive/latest-lite/
| GOSOURCEDATE: 20110910
| GOEGSOURCEDATE: 2011-Sep14
| GOEGSOURCENAME: Entrez Gene
| GOEGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
| KEGGSOURCENAME: KEGG GENOME
| KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
| KEGGSOURCEDATE: 2011-Mar15
| GPSOURCENAME: UCSC Genome Bioinformatics (Drosophila melanogaster)
| GPSOURCEURL: ftp://hgdownload.cse.ucsc.edu/goldenPath/dm3
| GPSOURCEDATE: 2009-Jul5
| FBSOURCENAME: Flybase
| FBSOURCEURL: ftp://ftp.flybase.net/releases/current/precomputed_files/genes/
| ENSOURCEDATE: 2011-Jun30
| ENSOURCENAME: Ensembl
| ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
| FBSOURCEDATE: 2011-Sept21
```

```
> cols(org.Dm.eg.db)
```

```
 [1] "ACCNUM"       "ALIAS2EG"     "CHR"          "ENZYME"       "GENENAME"
 [6] "MAP"          "PATH"         "PMID"         "REFSEQ"       "SYMBOL"
[11] "UNIGENE"      "CHRLOC"       "CHRLOCEND"    "FLYBASE"      "FLYBASECG"
[16] "FLYBASEPROT"  "UNIPROT"      "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
[21] "GO"
```

```
> keytypes(org.Dm.eg.db)
```

```
 [1] "ACCNUM"       "ALIAS2EG"     "CHR"          "ENZYME"       "GENENAME"
 [6] "MAP"          "PATH"         "PMID"         "REFSEQ"       "SYMBOL"
[11] "UNIGENE"      "CHRLOC"       "CHRLOCEND"    "FLYBASE"      "FLYBASECG"
[16] "FLYBASEPROT"  "UNIPROT"      "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
[21] "GO"
```

```
> uniKeys <- head(keys(org.Dm.eg.db, keytype="UNIPROT"))
> cols <- c("SYMBOL","PATH")
> select(org.Dm.eg.db,keys=uniKeys,cols=cols,keytype="UNIPROT")
```

```
      gene_id    symbol path_id uniprot_id
13571 3771890       eys    <NA>     A0A1F4
18623   40838        dj    <NA>     A0AMH4
24284   46058 Prosbeta1   03050     A0AQH0
19213   41286   Takr86C   04080     A0AVU4
21126   42812   CG10375    <NA>     A0AVU6
28868 5740176   CG34340    <NA>     A0AVV3
```

Exercise 21
Look again at the topTable data that we annotated in the the previous section.

How could you have extracted this data if you had instead used the OrgDb object along with the `select` *method?*

Solution:

```
> fbids <- rownames(tt$table)
> cols <- "SYMBOL"
> annots <- select(org.Dm.eg.db, keys=fbids, cols=cols, keytype="FLYBASE")
> tt <- cbind(flybase_id=rownames(as.data.frame(tt)), as.data.frame(tt))
> merge(tt, annots,by.x="flybase_id",by.y="flybase_id")
```

```
   flybase_id logConc logFC  LR P.Value adj.P.Val gene_id  symbol
1 FBgn0000071   -10.6   2.8 183 1.1e-41   1.1e-38   40831     Ama
2 FBgn0024288   -12.4  -4.7 179 7.1e-41   6.3e-38   45039 Sox100B
3 FBgn0034434   -11.4  -4.0 222 2.8e-50   7.3e-47   37219    Rgk1
4 FBgn0034736   -11.0  -3.5 207 6.9e-47   1.4e-43   37572  CG6018
5 FBgn0035189   -11.0   3.1 204 2.6e-46   4.2e-43   38124  CG9119
6 FBgn0037290   -12.0   3.1 159 1.9e-36   1.5e-33   40613  CG1124
7 FBgn0039155    -9.6  -4.7 378 3.3e-84   2.6e-80   42865   kal-1
8 FBgn0039827   -10.6  -4.3 291 2.9e-65   1.1e-61   43689  CG1544
9 FBgn0085359   -12.3  -4.7 191 2.4e-43   3.2e-40 2768869 CG34330
```

7.4 Using biomaRt

The *biomaRt* package offers access to several different resources referred to as 'marts'. Each mart allows access to multiple datasets; there is a standard method `getBM` for retrieving data from each of these datasets.

Exercise 22
Load the biomaRt package and list the available marts. Now choose the ensembl mart and list the datasets for that mart. Set up a mart that uses the ensembl "mart" and the hsapiens_gene_ensembl dataset.

biomaRt datasets can be accessed via `getBM` *which itself takes filters, values and attributes as arguments. Use the appropriate functions to list the optional values for these arguments.*

Now call `getBM` *using appropriate arguments of your choosing.*

Solution:

```
> library(biomaRt)
> ## list the marts
> head(listMarts())
> ## list the datasets for a mart
> head(listDatasets(useMart("ensembl")))
> ## set up the fully qualified mart object
> ensembl <- useMart("ensembl", dataset = "hsapiens_gene_ensembl")
> ## list filters
> head(listFilters(ensembl))
> myFilter <- "chromosome_name"
> ## list values that be returned
> head(filterOptions(myFilter, ensembl))
> myValues <- c("21", "22")
> ## list attributes
> head(listAttributes(ensembl))
> myAttributes <- c("ensembl_gene_id","chromosome_name")
> ## assemble and query the mart
> res <- getBM(attributes =  myAttributes,
+              filters =  myFilter,
+              values =  myValues,
+              mart = ensembl)
```

Use `head(res)` to see the results.

A Data retrieval

A.1 RNA-seq data retrieval

The following script was used to retrieve a portion of the Pasilla data set from the short read archive. The data is very large; extraction relies on installation of the SRA SDK, available from the Short Read Archive.

```
> library(RCurl)
> srasdk <- "/home/mtmorgan/bin/sra_sdk-2.0.1" # local installation
> sra <- "ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByExpt/sra"
> expt <- "SRX/SRX014/SRX014458/"
> url <- sprintf("%s/%s", sra, expt)
> acc <- strsplit(getURL(url, ftplistonly=TRUE), "\n")[[1]]
> urls <- sprintf("%s%s/%s.sra", url, acc, acc)
> for (fl in urls)
+     system(sprintf("wget %s",  fl), wait=FALSE, ignore.stdout=TRUE)
> app <- sprintf("%s/bin64/fastq-dump", srasdk)
> for (fl in file.path(wd, basename(urls)))
+      system(sprintf("%s %s", app, fl), wait=FALSE)
```

A.2 ChIP-seq data retrieval and MACS analysis

References

[1] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[2] A. N. Brooks, L. Yang, M. O. Duff, K. D. Hansen, J. W. Park, S. Dudoit, S. E. Brenner, and B. R. Graveley. Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Research*, pages 193–202, 2011.

[3] J. M. Chambers. *Software for Data Analysis: Programming with R.* Springer, New York, 2008.

[4] P. Dalgaard. *Introductory Statistics with R.* Springer, 2nd edition, 2008.

[5] R. Gentleman. *R Programming for Bioinformatics.* Computer Science & Data Analysis. Chapman & Hall/CRC, Boca Raton, FL, 2008.

[6] J. W. Ho, E. Bishop, P. V. Karchenko, N. Negre, K. P. White, and P. J. Park. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics*, 12:134, 2011. [PubMed Central:PMC3053263] [DOI:10.1186/1471-2164-12-134] [PubMed:21356108].

[7] P. V. Kharchenko, A. A. Alekseyenko, Y. B. Schwartz, A. Minoda, N. C. Riddle, J. Ernst, P. J. Sabo, E. Larschan, A. A. Gorchakov, T. Gu, D. Linder-Basso, A. Plachetka, G. Shanower, M. Y. Tolstorukov, L. J. Luquette, R. Xi, Y. L. Jung, R. W. Park, E. P. Bishop, T. K. Canfield, R. Sandstrom, R. E. Thurman, D. M. MacAlpine, J. A. Stamatoyannopoulos, M. Kellis, S. C. Elgin, M. I. Kuroda, V. Pirrotta, G. H. Karpen, and P. J. Park. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature*, 471:480–485, Mar 2011. [PubMed Central:PMC3109908] [DOI:10.1038/nature09725] [PubMed:21179089].

[8] P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, 26:1351–1359, Dec 2008. [PubMed Central:PMC2597701] [DOI:10.1038/nbt.1508] [PubMed:19029915].

[9] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10:R25, 2009.

[10] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–1760, Jul 2009.

[11] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26:589–595, Mar 2010.

[12] N. Matloff. *The Art of R Programming.* No Starch Pess, 2011.

[13] R. M. Myers, J. Stamatoyannopoulos, M. Snyder, I. Dunham, R. C. Hardison, B. E. Bernstein, T. R. Gingeras, W. J. Kent, E. Birney, B. Wold, G. E. Crawford, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, T. S. Mikkelsen, P. Kheradpour, X. Zhang, L. Wang, R. Issner, M. J. Coyne,

T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. F. Lin, M. Kellis, T. R. Gingeras, C. A. Davis, P. Kapranov, A. Dobin, C. Zaleski, F. Schlesinger, S. Chakrabortty, S. Jha, W. Lin, J. Drenkow, H. Wang, K. Bell, H. Gao, I. Bell, E. Dumais, J. Dumais, S. E. Antonarakis, C. Borel, R. Guigo, S. Djebali, P. Ribeca, M. Sammeth, T. Alioto, A. Merkel, H. Tilgner, P. Carninci, Y. Hayashizaki, T. Lassmann, H. Takahashi, R. F. Abdelhamid, G. Hannon, K. Fejes-Toth, J. Preall, A. Gordon, V. Sotirova, A. Reymond, C. Howald, E. A. Graison, J. Chrast, Y. Ruan, X. Ruan, A. Shahab, W. Ting Poh, C. L. Wei, G. E. Crawford, T. S. Furey, A. P. Boyle, N. C. Sheffield, L. Song, Y. Shibata, T. Vales, D. Winter, Z. Zhang, D. London, T. Wang, E. Birney, V. R. Iyer, B. K. Lee, R. M. McDaniell, Z. Liu, A. Battenhouse, A. A. Bhinge, J. D. Lieb, L. L. Grasfeder, K. A. Showers, P. G. Giresi, S. K. Kim, C. Shestak, R. M. Myers, F. Pauli, T. E. Reddy, J. Gertz, E. Christopher, P. Jain, R. O. Sprouse, A. Bansal, B. Pusey, M. A. Muratet, K. E. Varley, K. M. Bowling, K. M. Newberry, A. S. Nesmith, J. A. Dilocker, S. L. Parker, L. L. Waite, K. Thibeault, K. Roberts, D. M. Absher, A. Mortazavi, B. Williams, G. Marinov, S. Pepke, B. King, K. McCue, A. Kirilusha, G. DeSalvo, K. Fisher-Aylor, H. Amrhein, J. Vielmetter, G. Sherlock, A. Sidow, S. Batzoglou, R. Rauch, A. Kundaje, M. Libbrecht, E. H. Margulies, S. C. Parker, L. Elnitski, E. D. Green, F. Kokocinski, A. Frankish, T. Hunt, G. Despacio, M. Kay, G. Mukherjee, A. Bignell, G. Saunders, V. Boychenko, M. J. Van Baren, R. H. Brown, E. Khurana, S. Balasubramanian, Z. Zhang, H. Lam, P. Cayting, R. Robilotto, Z. Lu, R. Guigo, A. Tanzer, D. G. Knowles, M. Mariotti, W. James Kent, D. Haussler, R. Harte, M. Diekhans, M. Kellis, M. Lin, P. Kheradpour, J. Ernst, A. Reymond, C. Howald, E. A. Graison, J. Chrast, M. Tress, J. Manuel, M. Snyder, S. G. Landt, D. Raha, M. Shi, G. Euskirchen, F. Grubert, M. Kasowski, J. Lian, P. Cayting, P. Lacroute, H. Monahan, D. Patacsil, T. Slifer, X. Yang, A. Charos, B. Reed, L. Wu, R. K. Auerbach, L. Habegger, M. Hariharan, J. Rozowsky, A. Abyzov, S. M. Weissman, K. Struhl, N. Lamarre, M. Lindahl-Allen, B. Miotto, Z. Moqtaderi, J. D. Fleming, P. Newburger, P. J. Farnham, S. Frietze, H. O'Geen, X. Xu, K. R. Blahnik, A. R. Cao, S. Iyengar, J. A. Stamatoyannopoulos, R. Kaul, R. E. Thurman, H. Wang, P. A. Navas, R. Sandstrom, P. J. Sabo, M. Weaver, T. Canfield, K. Lee, S. Neph, A. Reynolds, A. Johnson, E. Rynes, E. Giste, J. Neri, T. Frum, E. M. Johnson, E. D. Nguyen, A. K. Ebersol, M. E. Sanchez, H. H. Sheffer, D. Lotakis, E. Haugen, R. Humbert, T. Kutyavin, T. Shafer, J. Dekker, B. R. Lajoie, A. Sanyal, W. James Kent, K. R. Rosenbloom, T. R. Dreszer, B. J. Raney, G. P. Barber, L. R. Meyer, C. A. Sloan, V. S. Malladi, M. S. Cline, K. Learned, V. K. Swing, A. S. Zweig, B. Rhead, P. A. Fujita, K. Roskin, D. Karolchik, R. M. Kuhn, D. Haussler, E. Birney, I. Dunham, S. P. Wilder, D. Keefe, D. Sobral, J. Herrero, K. Beal, M. Lukk, A. Brazma, J. M. Vaquerizas, N. M. Luscombe, P. J. Bickel, N. Boley, J. B. Brown, Q. Li, H. Huang, L. Habegger, A. Sboner, J. Rozowsky, R. K. Auerbach, K. Y. Yip, C. Cheng, K. K. Yan, N. Bhardwaj, J. Wang, L. Lochovsky, J. Jee, T. Gibson, J. Leng, J. Du, R. C. Hardison, R. S. Harris, W. Miller, D. Haussler, K. Roskin, B. Suh, T. Wang, B. Paten, W. S. Noble, M. M. Hoffman, O. J. Buske, Z. Weng, X. Dong, J. Wang, H. Xi, S. A. Tenenbaum, F. Doyle, L. O. Penalva, S. Chittur, T. D. Tullius, S. C.

Parker, K. P. White, S. Karmakar, A. Victorsen, N. Jameel, R. L. Grossman, M. Snyder, S. G. Landt, X. Yang, D. Patacsil, T. Slifer, J. Dekker, B. R. Lajoie, A. Sanyal, Z. Weng, T. W. Whitfield, J. Wang, P. J. Collins, N. D. Trinklein, E. C. Partridge, R. M. Myers, M. C. Giddings, X. Chen, J. Khatun, C. Maier, Y. Yu, H. Gunawardena, B. Risk, E. A. Feingold, R. F. Lowdon, L. A. Dillon, and P. J. Good. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, 9:e1001046, Apr 2011. [PubMed Central:PMC3079585] [DOI:10.1371/journal.pbio.1001046] [PubMed:21526222].

[14] P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, 10:669–680, Oct 2009. [PubMed Central:PMC3191340] [DOI:10.1038/nrg2641] [PubMed:19736561].

[15] S. Pepke, B. Wold, and A. Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, 6:22–32, Nov 2009. [DOI:10.1038/nmeth.1371] [PubMed:19844228].

[16] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, Jan 2010.

[17] P. J. Sabo, M. Hawrylycz, J. C. Wallace, R. Humbert, M. Yu, A. Shafer, J. Kawamoto, R. Hall, J. Mack, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl. Acad. Sci. U.S.A.*, 101:16837–16842, Nov 2004.

[18] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9:R137, 2008. [PubMed Central:PMC2592715] [DOI:10.1186/gb-2008-9-9-r137] [PubMed:18798982].