Galaxy and edgeR

Mark D. Robinson
University of Zurich
My work

• Research:
  – 50%: statistical method development for genomics
  – 35%: applying statistical methods for/with collaborators
  – 15%: tool development (Bioconductor)

• Software:
  – 85% R/Bioconductor
  – 5% browsers – UCSC/IGV
  – 5% command line
  – 5% other – perl/python/Galaxy

• Galaxy: Interfaces to R/Bioconductor
About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 554 software packages, and an active user community. Bioconductor is also available as an Amazon Machine Image (AMI).
RNA-seq

differential expression analysis

differential isoform analysis

Mortazavi et al., Nature Methods, 2008
Current RNA-seq workflows in Galaxy

Bowtie, tophat – mapping reads

edgeR is an alternative to ‘cuffdiff’

Potential disadvantages of cuffdiff:
- only allows 2-sample comparisons
- Can be problematic at low depth / short read (our obs.)
- complicated
edgeR workflow
(similar to DESeq, baySeq, ...)

- Millions of short reads
  - Mapping
    - Burrows wheeler transform
      - Bowtie, BWA, SOAP2
    - Hash tables
      - PerM, SHRIMP, BFAST, ELAND
    - Unmapped reads
      - Map to Junction library created from annotation
      - Map to 'de novo' junction library
        - SplitSeek, Tophat, SOAPals
  - Reads aligned to reference
    - Summarization
      - By coding sequence
      - By exon
      - By gene span
      - Junction reads
      - Table of counts
        - Normalization
          - Between sample
            - TMM, upper quartile
          - Within sample
            - RPKM, quantile
  - DE testing
    - Poisson test
      - DEGseg
    - Negative binomial test
      - edgeR, baySeq, DESeq
  - List of differentially expressed genes
    - Systems Biology
      - Test for enriched categories/pathways
        - GOseq
      - Infer networks and integrate with other data
      - Biological insight
edgeR workflow

Counting:
- `htseq-count` (python)
- various tools within Bioconductor
What a standard edgeR analysis might look like

```r
library(edgeR)
d <- DGEList(counts=D, group=grp)
d <- calcNormFactors(d)
d <- estimateTagwiseDisp(d)
```

Or, start with table of metadata:

```
> samples
   rep condition libtype shortname countfile
  1   1      T      SE    T1.SE   S2_DRSC.CG8144.RNAi-1.count
  2   2      T      PE    T3.PE   S2_DRSC.CG8144.RNAi-3.count
  3   3      T      PE    T4.PE   S2_DRSC.CG8144.RNAi-4.count
  4   4      C      SE    C1.SE   S2_DRSC.Untreated-1.count
  5   5      C      SE    C6.SE   S2_DRSC.Untreated-6.count
  6   6      C      PE    C3.PE   S2_DRSC.Untreated-3.count
  7   7      C      PE    C4.PE   S2_DRSC.Untreated-4.count
```
What a standard edgeR analysis might look like

plotMDS(d, col=c("blue","orange")[factor(grp)])
What a standard edgeR analysis might look like

de <- exactTest(d)
tt <- topTags(de, n=nrow(d))
rn.de <- rownames(tt)[tt$table$FDR < .05][1:500]
head(tt)

> head(tt$table)
    logFC logCPM     PValue           FDR
ENSG00000151224  7.714164 11.404623 4.599520e-167 7.907494e-163
ENSG00000134339  7.428391 13.098576 6.366207e-164 5.472392e-160
ENSG00000173432  7.335286 13.061117 3.179161e-163 1.821871e-159
ENSG00000138115  7.431459 13.097222 2.968370e-159 1.275806e-155
ENSG00000141485  8.123902  9.329354 3.578253e-158 1.230347e-154
ENSG00000160868  7.239692 11.218554 3.996155e-154 1.128469e-150
DEXSeq – exon counts – differential isoform usage (edgeR has a similar mode)
The analysis of quantitative HTS data (e.g. from RNA-seq or ChIP-seq) relies on statistical procedures that highlight differential regions. For example, the density of mapped reads in a particular genomic region may represent enrichment level of a protein-DNA interaction (ChIP-seq), or gene expression level (RNA-seq). The edgeR plugin is a wrapper for software written in the R statistical programming language for the detection of significantly differentially enriched regions or expressed genes, relative to observed biological variation, directly within Savant (12). The plugin computes on multiple BAM tracks, some designated as Case and others as Control, and provides a table of ranked results, including the region locations, log-fold-changes, P-values and estimated false discovery rates of the change between conditions.

General, can be applied to other types of data (e.g. ChIP-seq)
Bioconductor and Galaxy, generally

• RGalaxy package

The following example illustrates these best practices (this function is in the RGalaxy package under the name functionToGalaxify:

```r
function (inputfile1 = GalaxyInputFile(), inputfile2 = GalaxyInputFile(),
plotTitle = character(), plotSubTitle = "My subtitle", outputfile1 = GalaxyOutput("mydata", "csv"), outputfile2 = GalaxyOutput("myplot", "pdf"))
{
  data1 <- tryCatch({
    as.matrix(read.delim(inputfile1, row.names = 1))
  }, error = function(err) {
    stop("failed to read first data file: ", conditionMessage(err))
  })
  data2 <- tryCatch({
    as.matrix(read.delim(inputfile2, row.names = 1))
  }, error = function(err) {
    stop("failed to read second data file: ", conditionMessage(err))
  })
  data3 <- data1 + data2
  write.csv(data3, file = outputfile1)
  pdf(outputfile2)
  if (missing(plotTitle))
    plotTitle <- ""
  plot(data3, main = plotTitle, sub = plotSubTitle)
  dev.off()
}
<environment: namespace:RGalaxy>
```
Bioconductor and Galaxy, generally

• RGalaxy package

6 Adding a function to Galaxy

Now that you have written a function that follows the best practices described above, you can make it available to Galaxy as follows:

```r
> galaxy(functionToGalaxify,
+    manpage="functionToGalaxify",
+    version="0.99.0",
+    name="Add",
+    package="RGalaxy",
+    galaxyConfig=GalaxyConfig(getwd(), "mytool", "Test Section",
+                                "testSectionId")
+    )
```
Bioconductor and Galaxy, generally

- Mentored projects at BioC

Galaxy-ification of Useful Scripts

From the Wikipedia entry for Galaxy:

Galaxy is a scientific workflow, data integration, and data and analysis persistence and publishing platform that aims to make computational biology accessible to research scientists that do not have computer programming experience. Although it was initially developed for genomics research, it is largely domain agnostic and is now used as a general bioinformatics workflow management system.

The new Bioconductor RGalaxy simplifies the process of exposing an R function in Galaxy so that a user can run the function using nothing more than a web browser.

This project would involve taking an existing workflow (or conceiving a new workflow) and exposing it in Galaxy.

Project attributes and estimates:

- Difficulty: Medium
- Length: 1-2 weeks
- Skills needed: R programming
- Mentor: Dan Tenenbaum
- Status: Open
Ross Lazarus (BakerIDI, Melbourne)
Ross Lazarus (BakerIDI, Melbourne)

<table>
<thead>
<tr>
<th>Tools</th>
<th>rawP</th>
<th>BH Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make new tools or run scripts</td>
<td>0.002</td>
<td>0.008</td>
</tr>
<tr>
<td>Tool Factory</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td>Get Data</td>
<td>0.2</td>
<td>0.008</td>
</tr>
<tr>
<td>Send Data</td>
<td>0.19</td>
<td>0.029</td>
</tr>
<tr>
<td>ENCODE Tools</td>
<td>0.029</td>
<td>0.06</td>
</tr>
<tr>
<td>Lift-Over</td>
<td>0.1</td>
<td>0.16</td>
</tr>
<tr>
<td>Text Manipulation</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Filter and Sort</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Join, Subtract and Group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convert Formats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extract Features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fetch Sequences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fetch Alignments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Get Genomic Scores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operate on Genomic Intervals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelet Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graph/Display Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regional Variation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multivariate Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evolution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motif Tools</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Alignments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metagenomic analyses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FASTA manipulation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCBI BLAST+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NGSQC and manipulation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**History**

- ToolFactory June 2012
- 7.2 Kb
- format: toolshed.g2, database: 2
- Info: wrote `pvalueadjuster/toolshed.py`
Ross Lazarus (BakerIDI, Melbourne)
Ross Lazarus (BakerIDI, Melbourne)
Ross Lazarus (BakerIDI, Melbourne)
Ross Lazarus (BakerIDI, Melbourne)
Some challenges in making R/Bioconductor tools available (to Galaxy)

• How to allow facilitate general design matrices?

```r
> design <- model.matrix(~ libtype + condition, samples)
> design

       (Intercept) libtypeSE conditionT
S2_DRSC.CG8144_RNAi-1.bam   1     1         1
S2_DRSC.CG8144_RNAi-3.bam   1     0         1
S2_DRSC.CG8144_RNAi-4.bam   1     0         1
S2_DRSC_Untreated-1.bam    1     1         0
S2_DRSC_Untreated-6.bam    1     1         0
S2_DRSC_Untreated-3.bam    1     0         0
S2_DRSC_Untreated-4.bam    1     0         0
```
Some challenges in making R/Bioconductor tools available (to Galaxy)

- How to specify metadata?
- How to adequately give all the plotting options?
- Versioning
- Making reproducible analyses available
- Only polished analyses made available?
- Citations
- Best practices – multiple best practices?
Copy-number-aware differential analysis of quantitative DNA sequencing data

Mark D. Robinson [1,2,3,*], Dario Strbenac [3], Clare Stirzaker [3,6], Aaron L. Statham [3], Jenny Song [3], Terence P. Speed [4,5], Susan J. Clark [3,6]

[1] Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190 CH-8057 Zurich, Switzerland
[2] Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[3] Epigenetics Laboratory, Cancer Research Program, Garvan Institute of Medical Research, Sydney 2010, New South Wales, Australia
[4] Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Melbourne 3052, Victoria, Australia
[5] Department of Medical Biology, University of Melbourne, 3050, Victoria, Australia
[6] St Vincent's Clinical School, University of NSW, Sydney, NSW, Australia
[*] Corresponding author: mark.robinson@imls.uzh.ch.

Summary

Developments in microarray and high throughput sequencing (HTS) technologies have resulted in a rapid expansion of research into epigenomic changes that occur in normal development and in the progression of disease, such as cancer. Not surprisingly, copy number variation (CNV) has a direct effect on HTS read densities and can therefore bias differential detection results. We have developed a flexible approach called ABCD-DNA (Affinity Based Copy-number-aware Differential quantitative DNA sequencing analyses) that integrates CNV and other systematic factors directly into the differential enrichment engine.

Supplementary Data (semi-processed), R Code for all Figures and analyses:

- Archive with all data/code [ABCD-DNA_Supplement_RCCode_Data.tar.gz, 649 MB]
- Sweaved PDF document with annotated ABCD-DNA analysis [ABCD-DNA.pdf]
- [Track] file (use with UCSC Genome Browser - hg18) for ABCD-DNA (and other algorithm) calls, Illumina 450k array data
Bioconductor meeting in Zurich – December 13-14

Overview
This workshop is aimed at Bioconductor contributors and bioinformaticians who wish to contribute packages to the Bioconductor project. The aim of the meeting is to foster the exchange of technical expertise, to keep contributors up to speed with the latest developments and to coordinate related efforts.

Topics this year will include:
- New developments in Bioconductor core packages and infrastructure
- Infrastructure and methods for understanding gene regulation
- Next Generation Data handling and analysis
- Reproducible research and software engineering
- Tools for the management and analysis of proteomics datasets

Time and location
December 13-14th 2012
University of Zürich, Irchel Campus, room Y55 06/08
Invited speakers

The workshop is supported by

- University of Zurich
- Swiss Institute of Bioinformatics
- SystemsX.ch
- The Swiss Initiative in Systems Biology
- FNS SNF
- Schweizerischer Nationalfonds
- Fondo Nazionale Svizzero
- Swiss National Science Foundation
- Agilent Technologies