University of Zurich[UZH]

URPP Systems Biology / Functional Genomics

Swiss Institute of Bioinformatics

Galaxy and edgeR

Mark D. Robinson

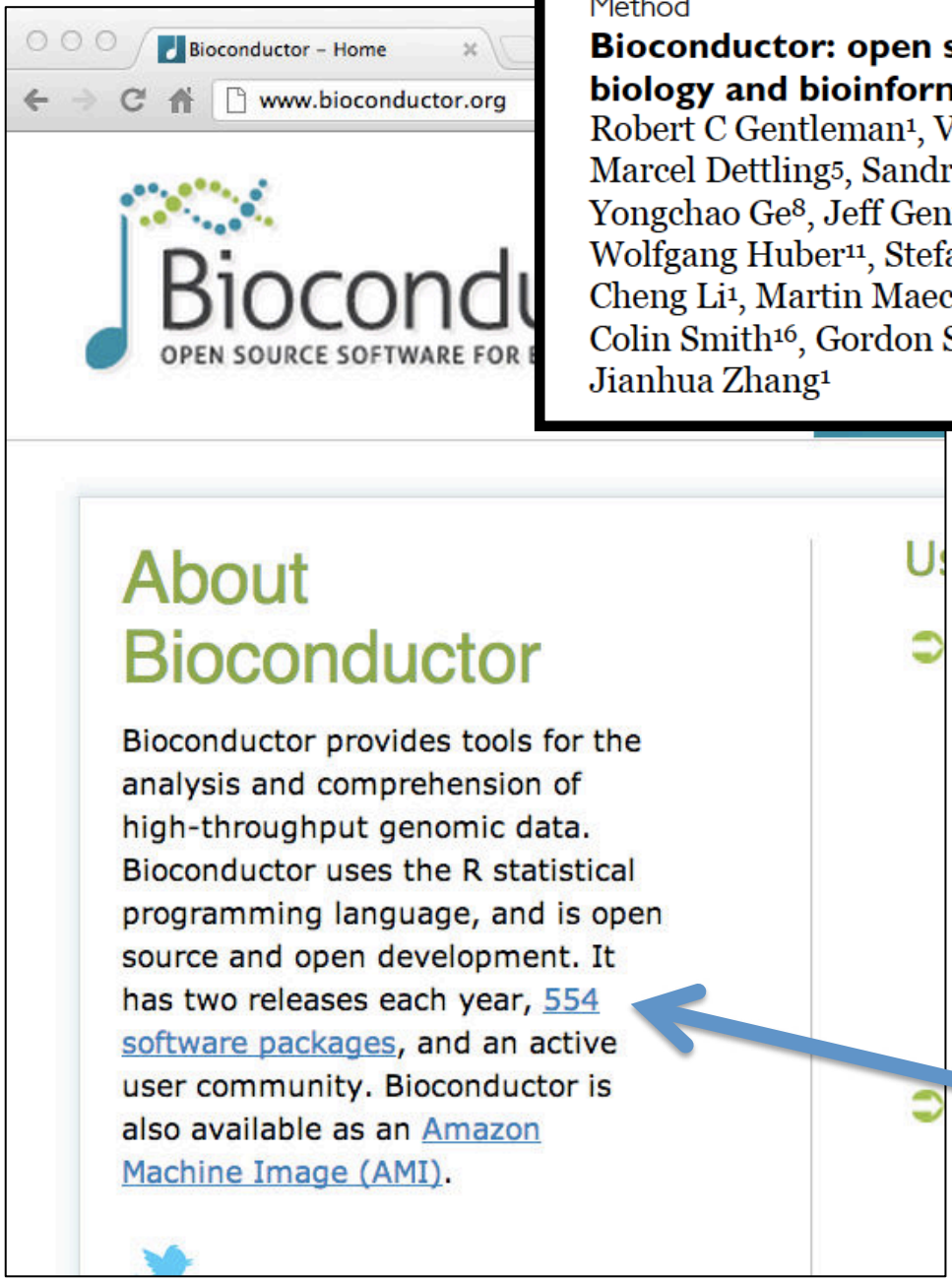
University of Zurich

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**



Method

Open Access

Bioconductor: open software development for computational biology and bioinformatics

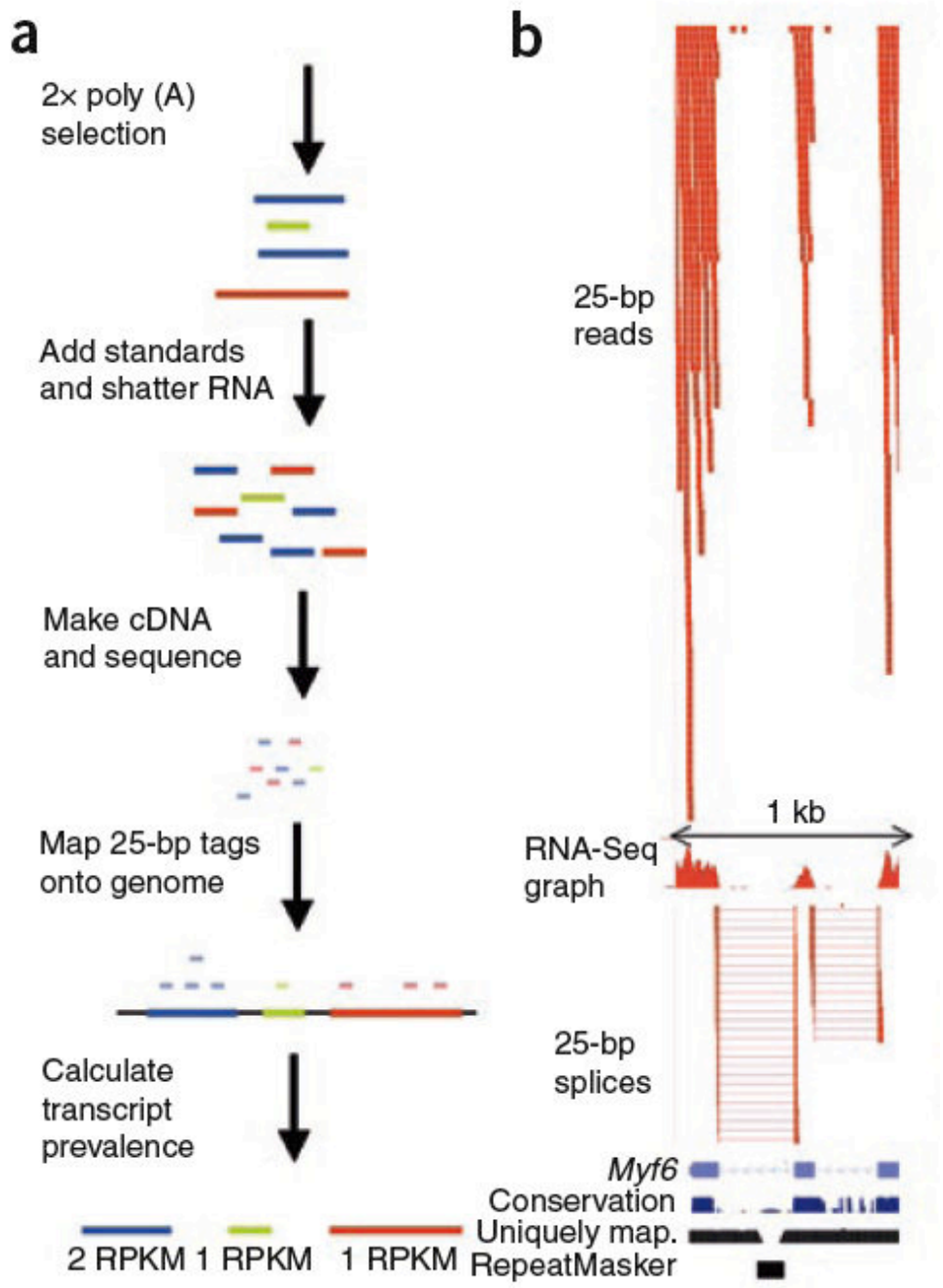
Robert C Gentleman[1], Vincent J Carey[2], Douglas M Bates[3], Ben Bolstad[4], Marcel Dettling[5], Sandrine Dudoit[4], Byron Ellis[6], Laurent Gautier[7], Yongchao Ge[8], Jeff Gentry[1], Kurt Hornik[9], Torsten Hothorn[10], Wolfgang Huber[11], Stefano Iacus[12], Rafael Irizarry[13], Friedrich Leisch[9], Cheng Li[1], Martin Maechler[5], Anthony J Rossini[14], Gunther Sawitzki[15], Colin Smith[16], Gordon Smyth[17], Luke Tierney[18], Jean YH Yang[19] and Jianhua Zhang[1]

[4341]

RNA-seq

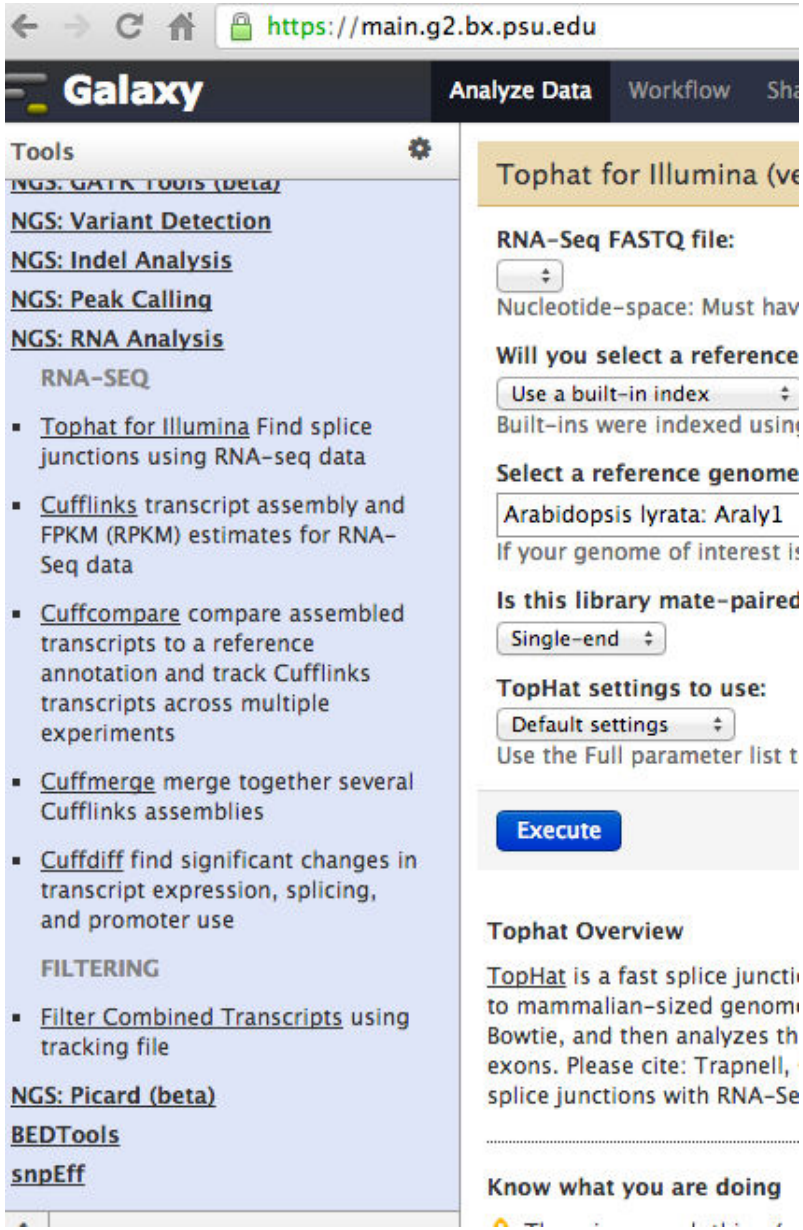
differential expression analysis

differential isoform analysis



Mortazavi et al., Nature Methods, 2008

Current RNA-seq workflows in Galaxy

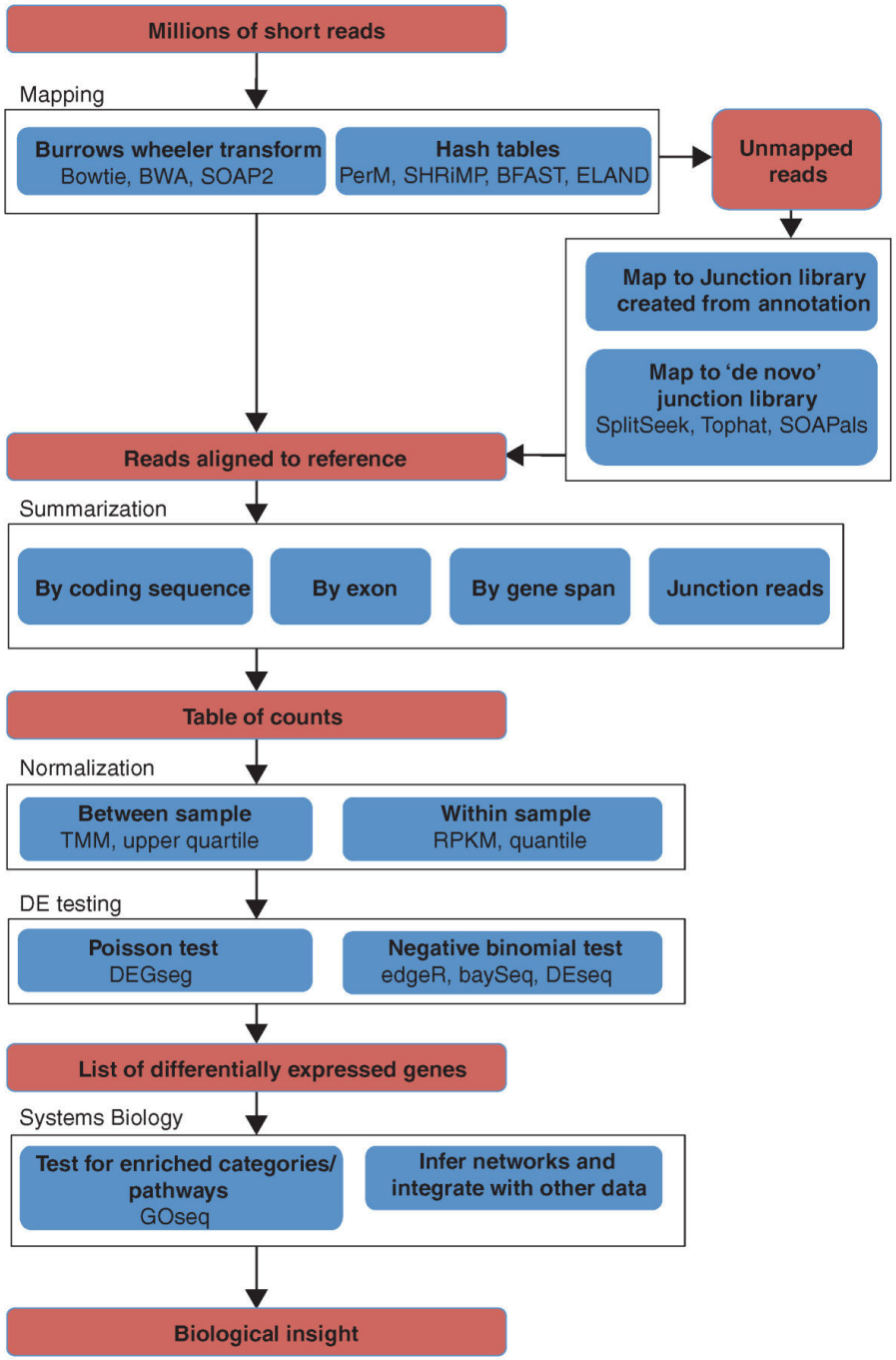


Bowtie, tophat – mapping reads

edgeR is an alternative to 'cuffdiff'

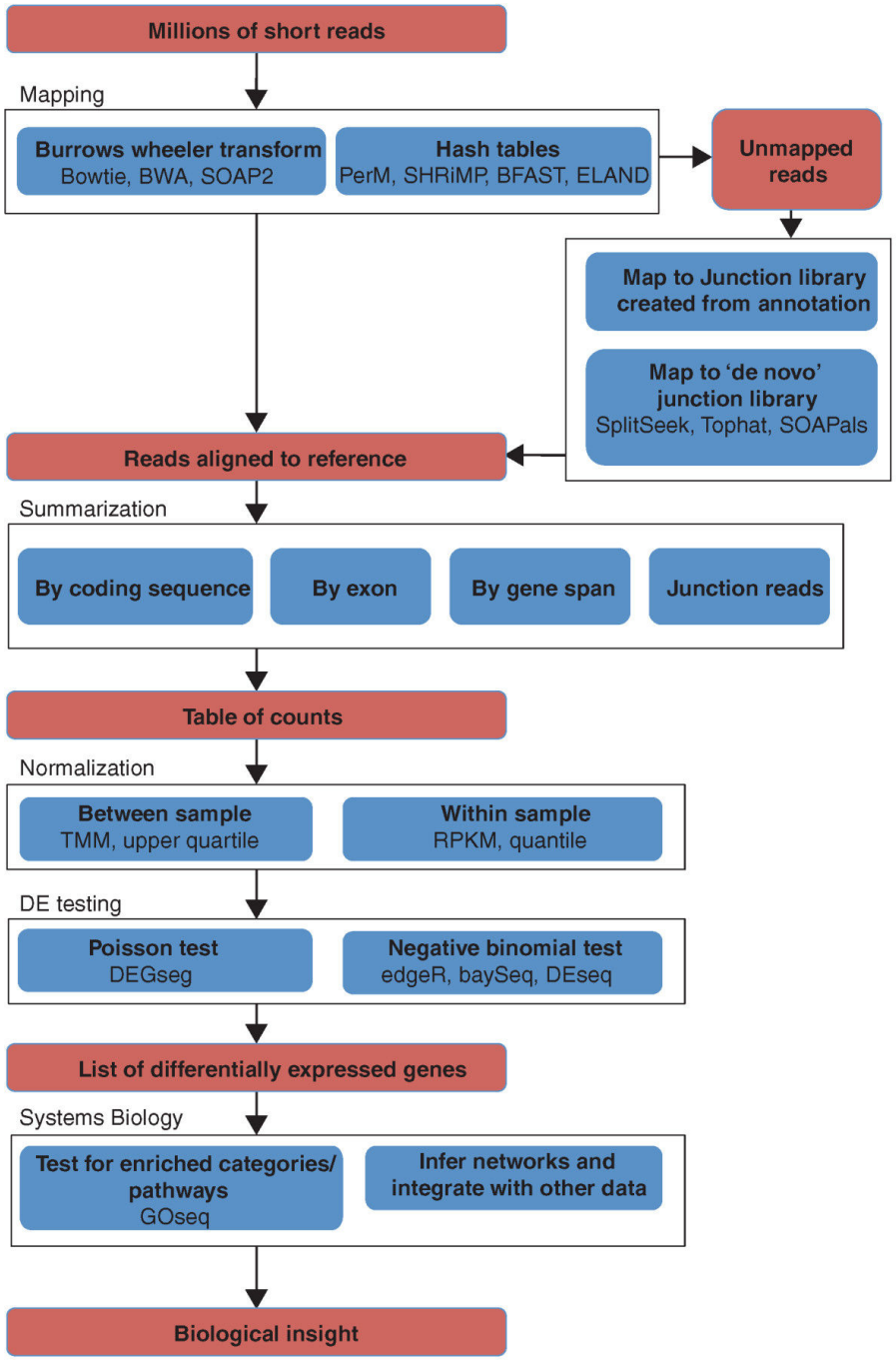
Potential disadvantages of cuffdiff:

- only allows 2-sample comparisons
- Can be problematic at low depth / short read (our obs.)
- complicated



edgeR workflow

(similar to DESeq, baySeq, …)



edgeR workflow

Counting:

- htseq-count (python)
- various tools within Bioconductor

What a standard edgeR analysis might look like

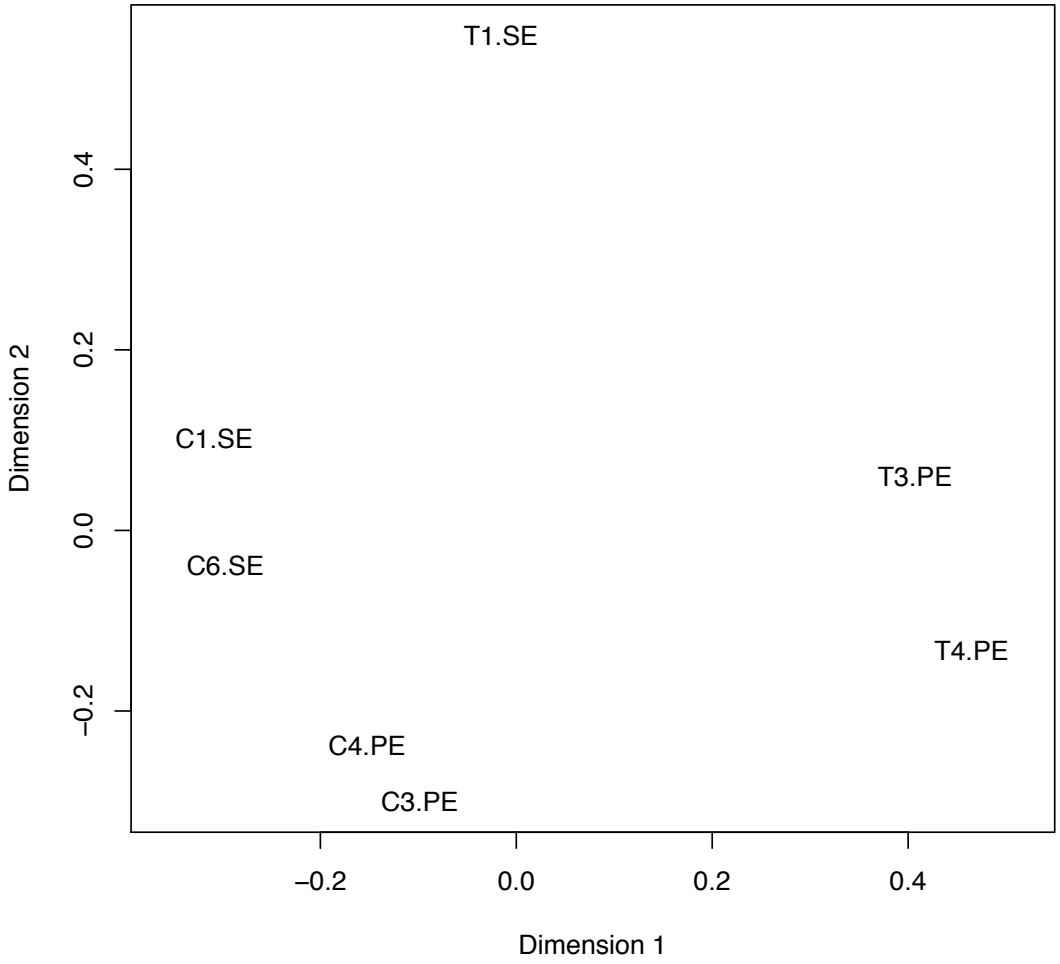
```
library(edgeR)
d <- DGEList(counts=D, group=grp)
d <- calcNormFactors(d)
d <- estimateTagwiseDisp(d)
```

Or, start with table of metadata:

```
> samples
  rep condition libtype shortname                    countfile
1  T1         T      SE     T1.SE S2_DRSC_CG8144_RNAi-1.count
2  T3         T      PE     T3.PE S2_DRSC_CG8144_RNAi-3.count
3  T4         T      PE     T4.PE S2_DRSC_CG8144_RNAi-4.count
4  C1         C      SE     C1.SE   S2_DRSC_Untreated-1.count
5  C6         C      SE     C6.SE   S2_DRSC_Untreated-6.count
6  C3         C      PE     C3.PE   S2_DRSC_Untreated-3.count
7  C4         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)])

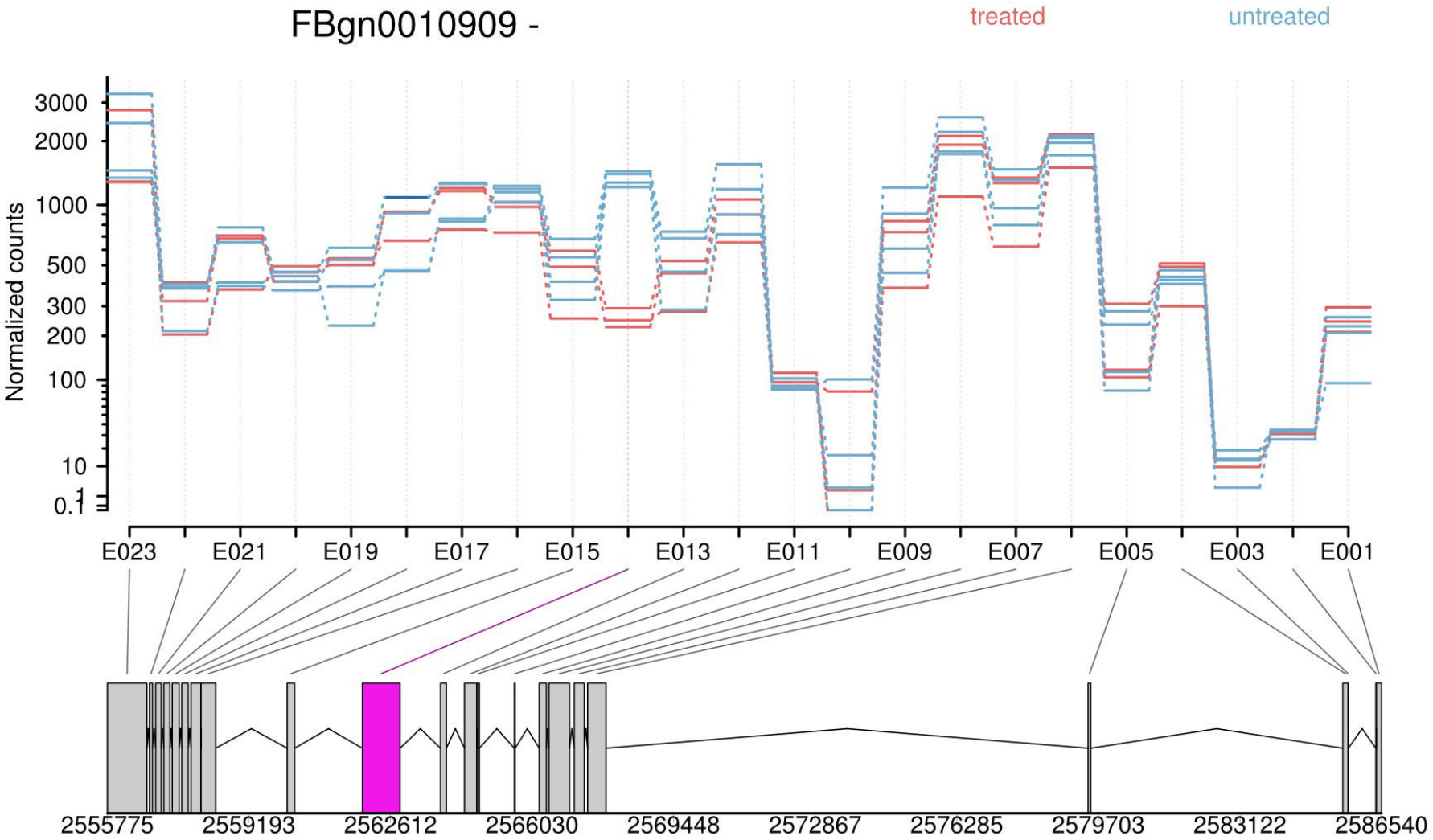


What a standard edgeR analysis might look like

```
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 11.339722 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)



edgeR integrated into Savant

Savant Genome Browser 2: visualization and analysis for population-scale genomics

Marc Fiume[1], Eric J. M. Smith[1], Andrew Brook[1], Dario Strbenac[2], Brian Turner[3], Aziz M. Mezlini[4], Mark D. Robinson[5], Shoshana J. Wodak[2,6] and Michael Brudno[1,4,*]

[1]Department of Computer Science, University of Toronto, Ontario, Canada M5S 2E4, [2]Epigenetics Laboratory Cancer Research Program, Garvan Institute of Medical Research, Sydney, New South Wales 2011, Australia, [3]Molecular Structure and Function Program, [4]Centre for Computational Medicine, Hospital for Sick Children, Toronto, Ontario, Canada M5G 1L7, [5]Institute of Molecular Life Sciences, University of Zurich, 8057 Zürich, Switzerland and [6]Department of Biochemistry, University of Toronto, Ontario, Canada M5S 1A8

Received March 7, 2012; Revised April 23, 2012; Accepted April 24, 2012

edgeR plugin

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`:

```
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

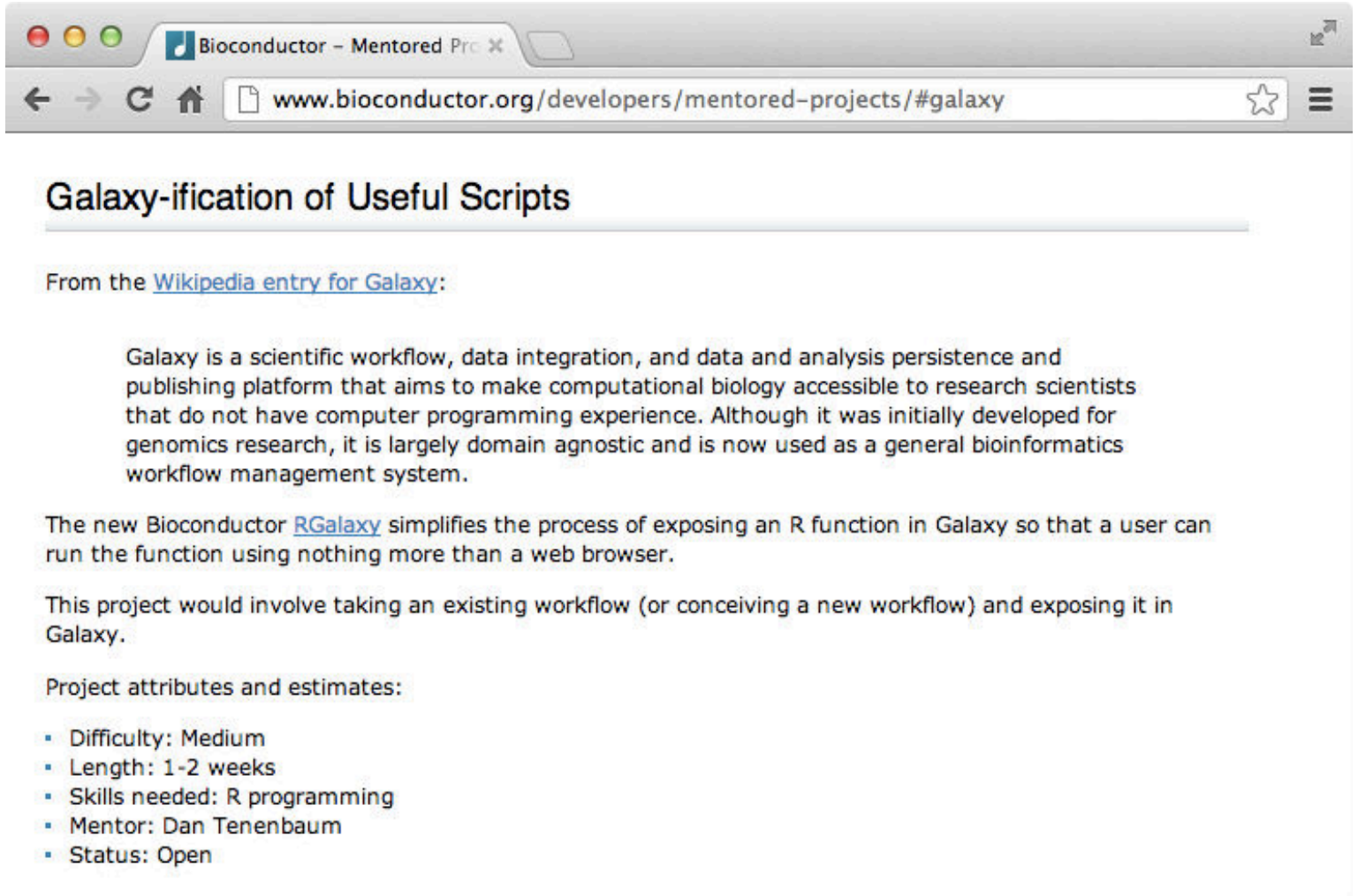
6 Adding a function to Galaxy

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

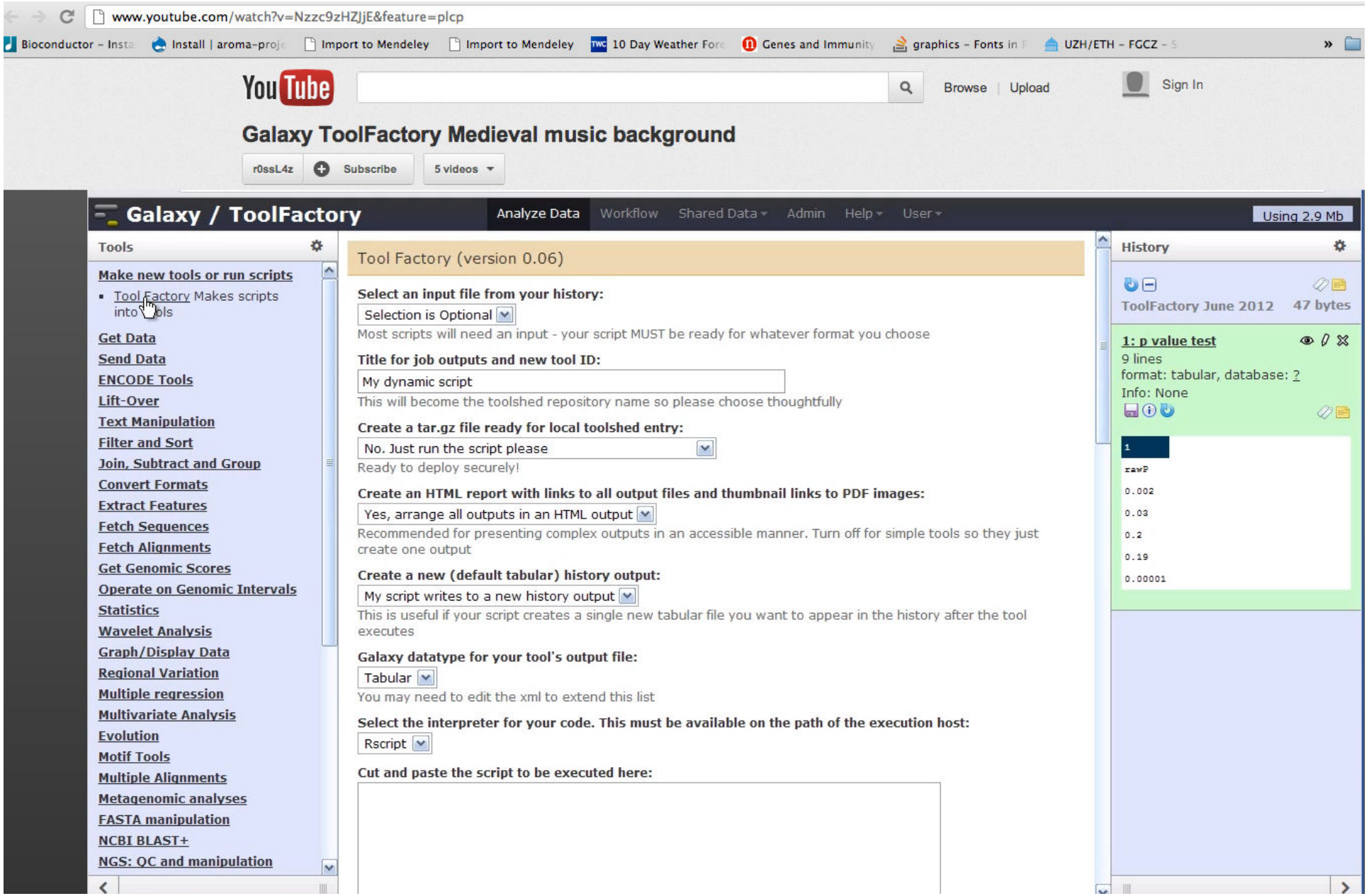
```
> galaxy(functionToGalaxify,
+     manpage="functionToGalaxify",
+     version="0.99.0",
+     name="Add",
+     package="RGalaxy",
+     galaxyConfig=GalaxyConfig(getwd(), "mytool", "Test Section",
+         "testSectionId")
+     )
```

Bioconductor and Galaxy, generally

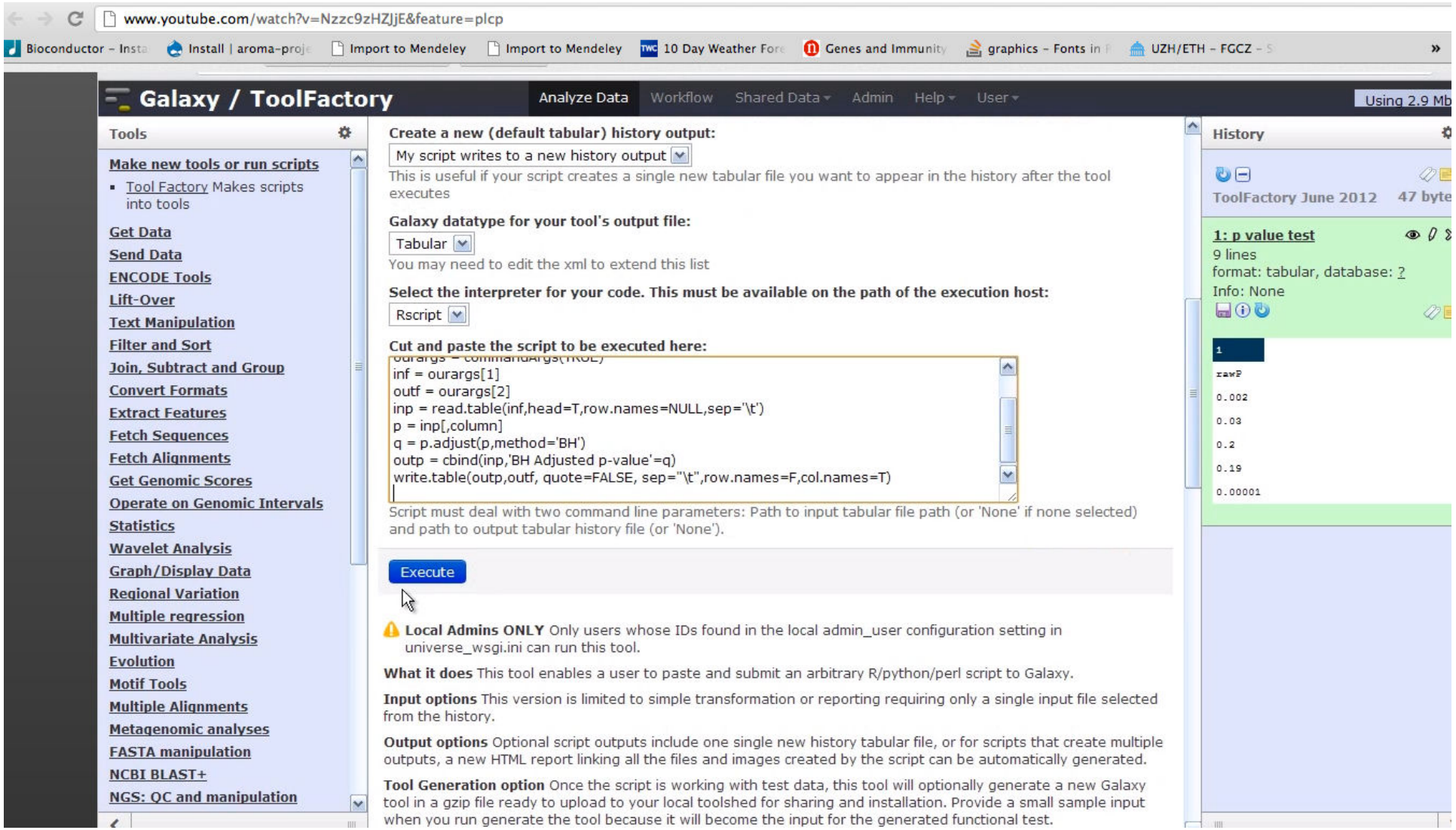
- Mentored projects at BioC



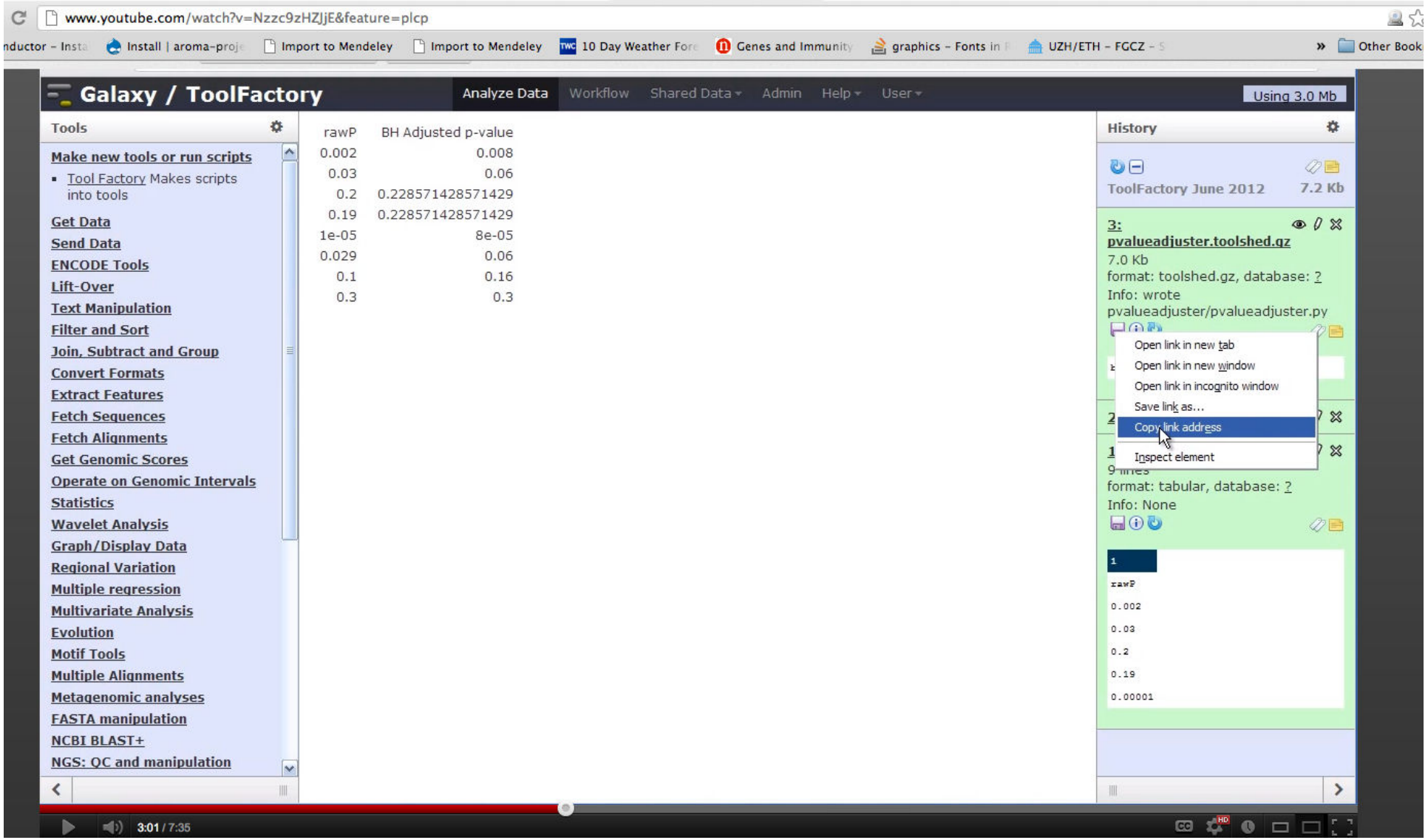
Ross Lazarus (BakerIDI, Melbourne)



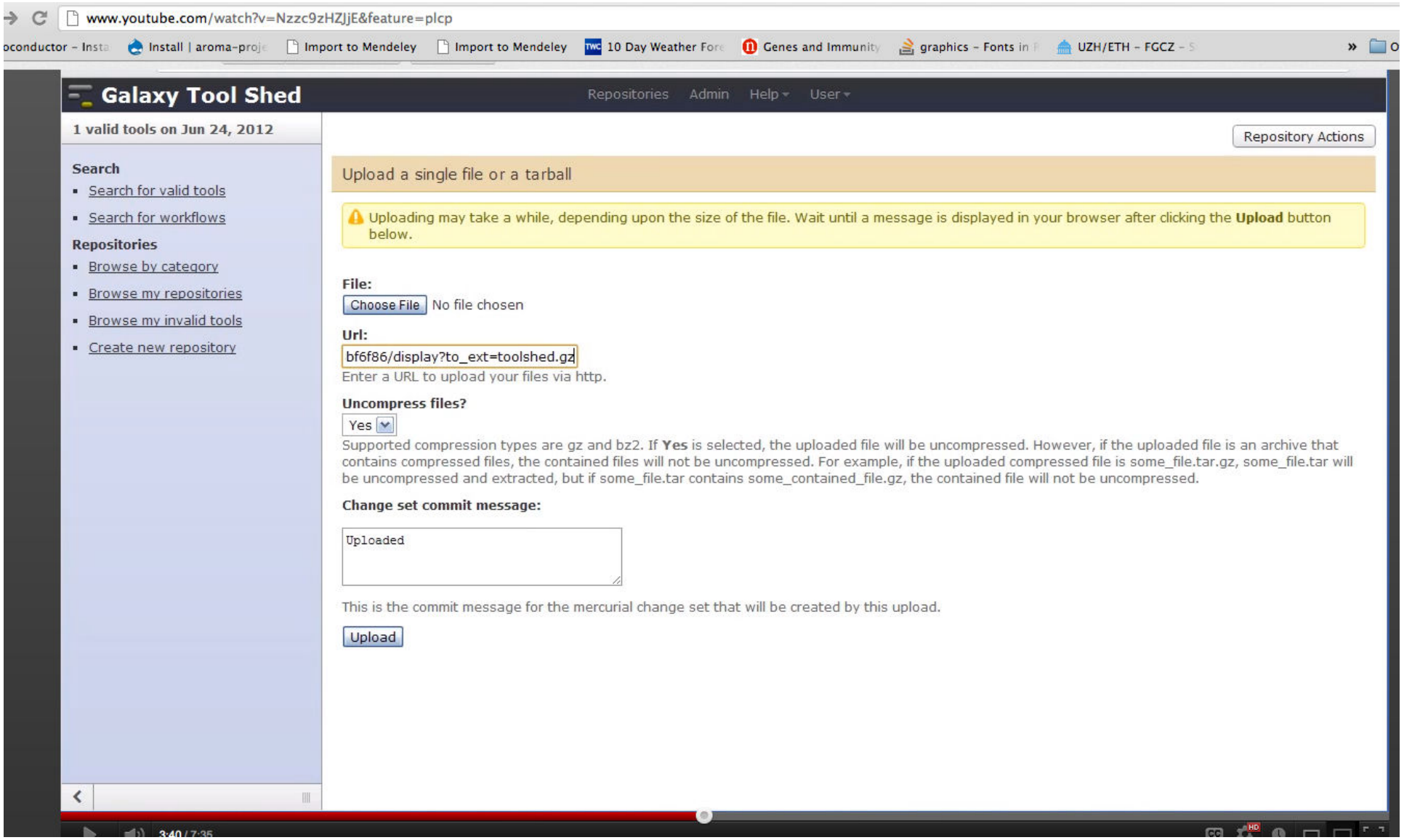
Ross Lazarus (BakerIDI, Melbourne)



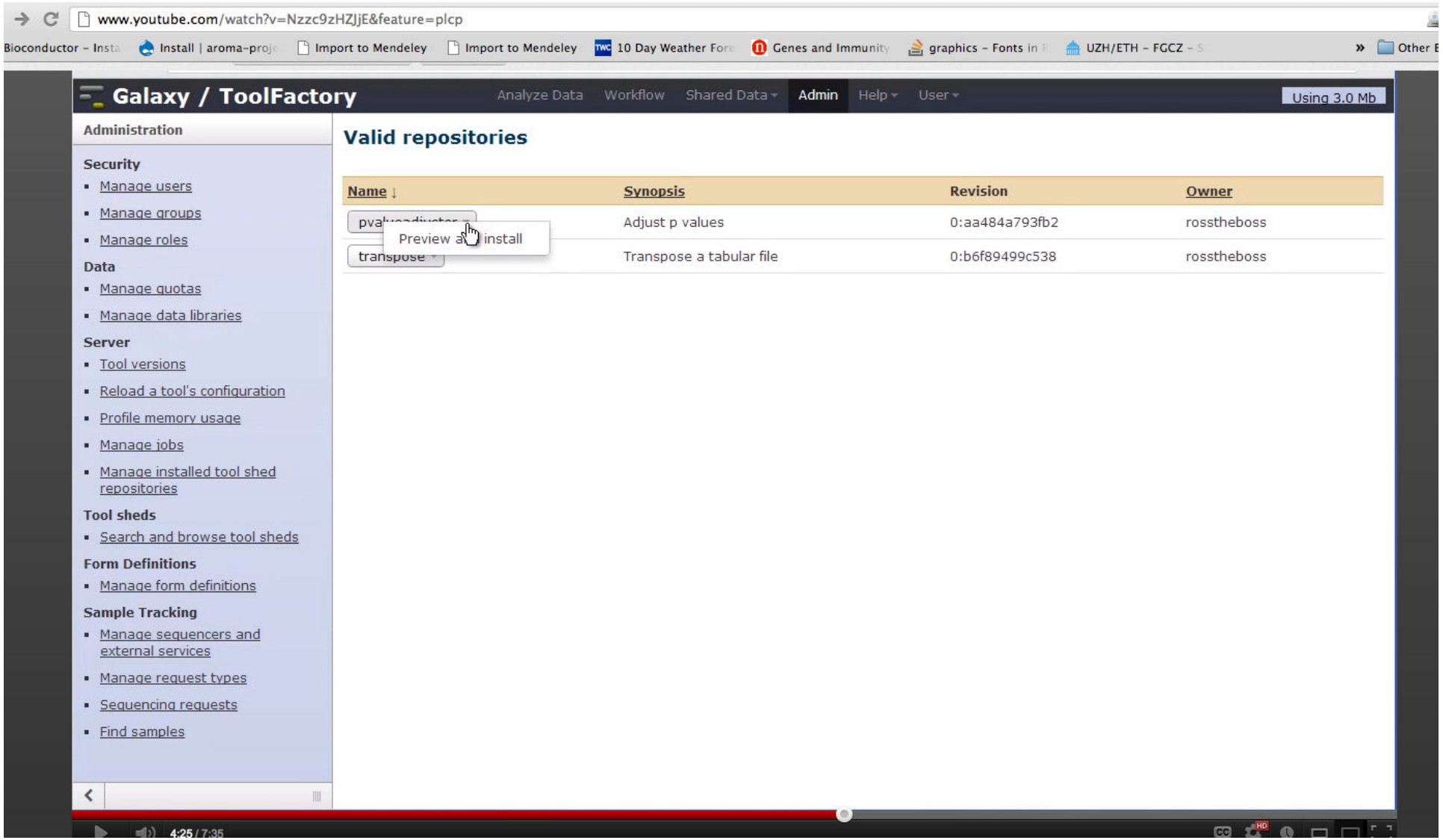
Ross Lazarus (BakerIDI, Melbourne)



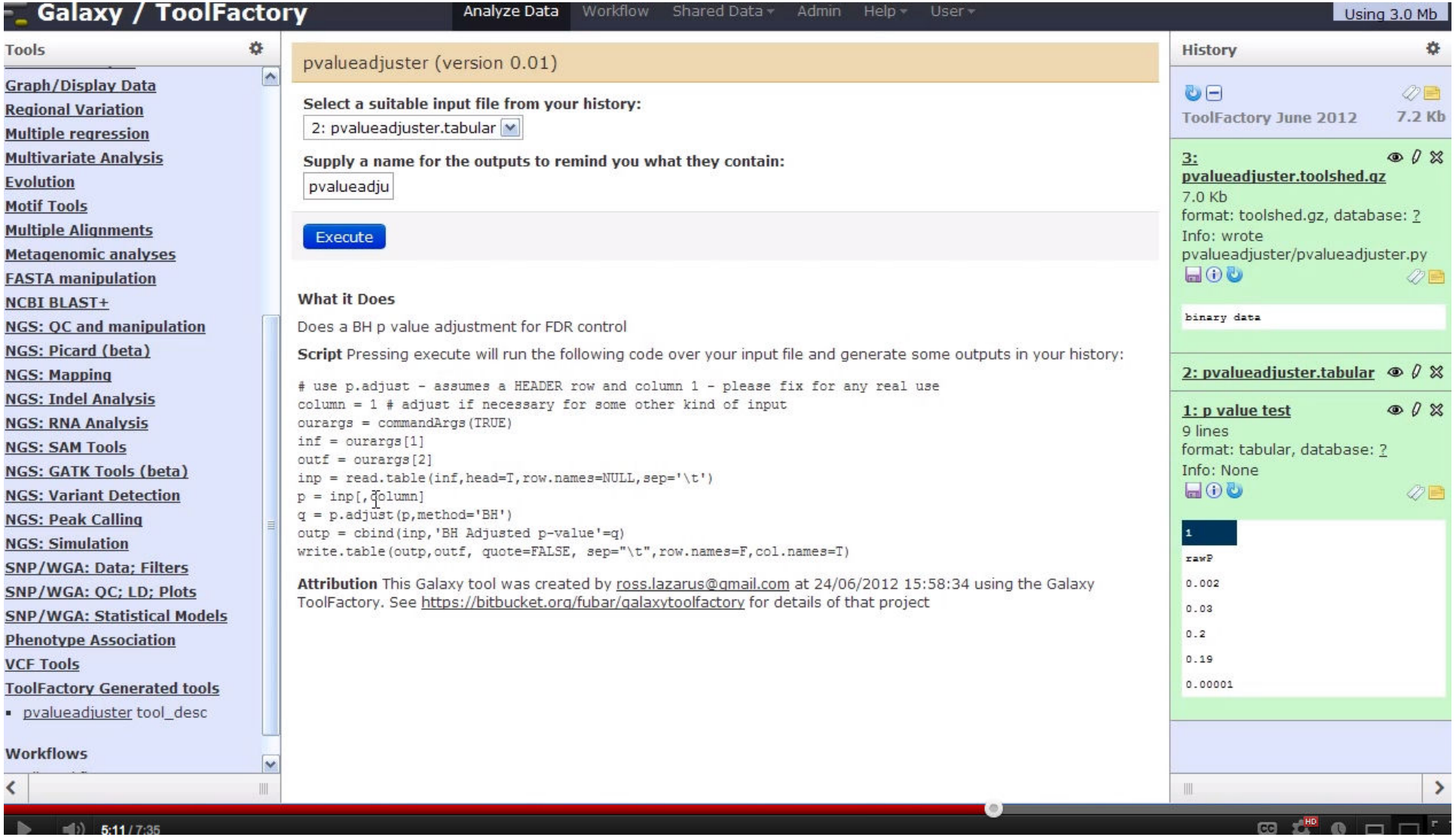
Ross Lazarus (BakerIDI, Melbourne)



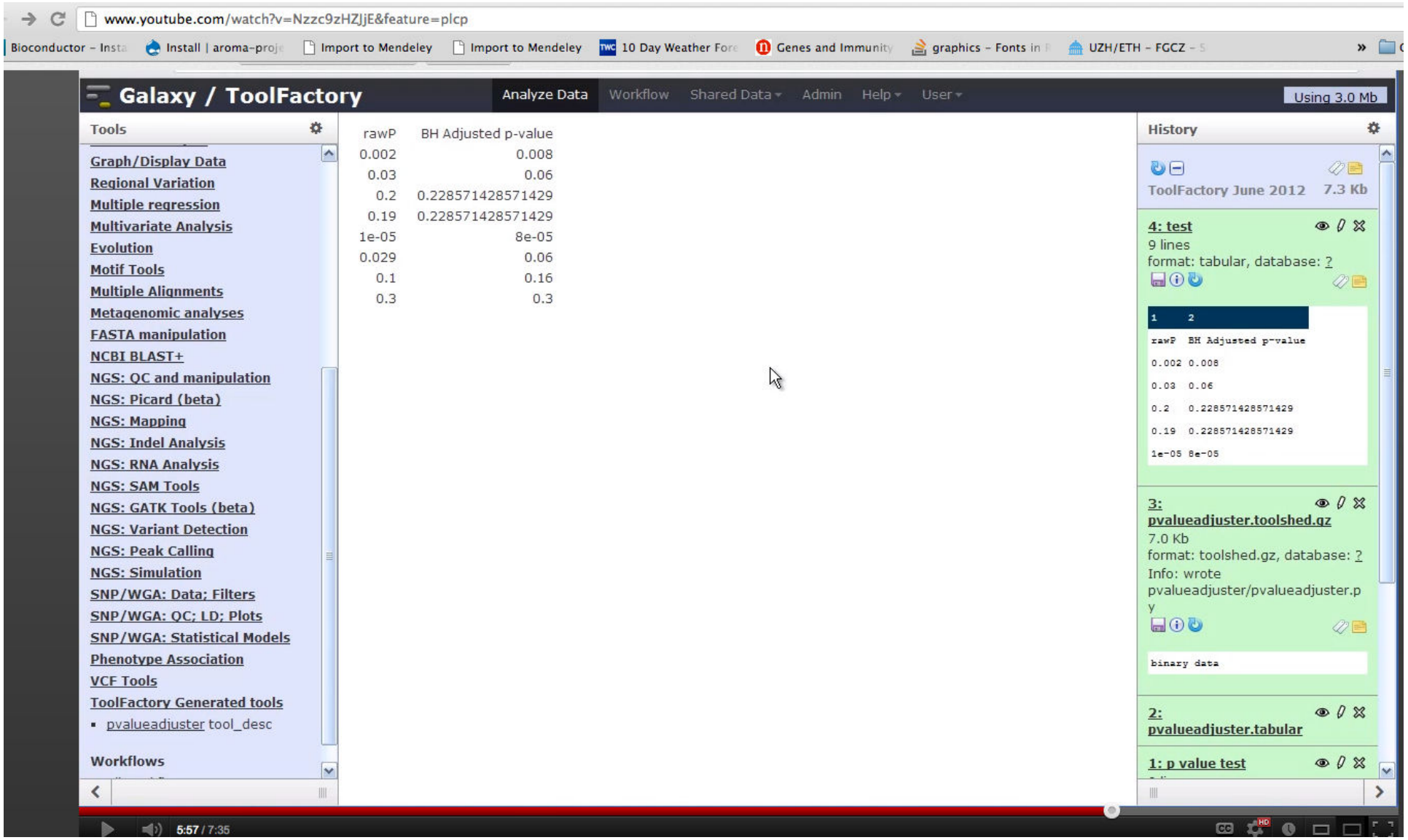
Ross Lazarus (BakerIDI, Melbourne)



Ross Lazarus (BakerIDI, Melbourne)



Ross Lazarus (BakerIDI, Melbourne)



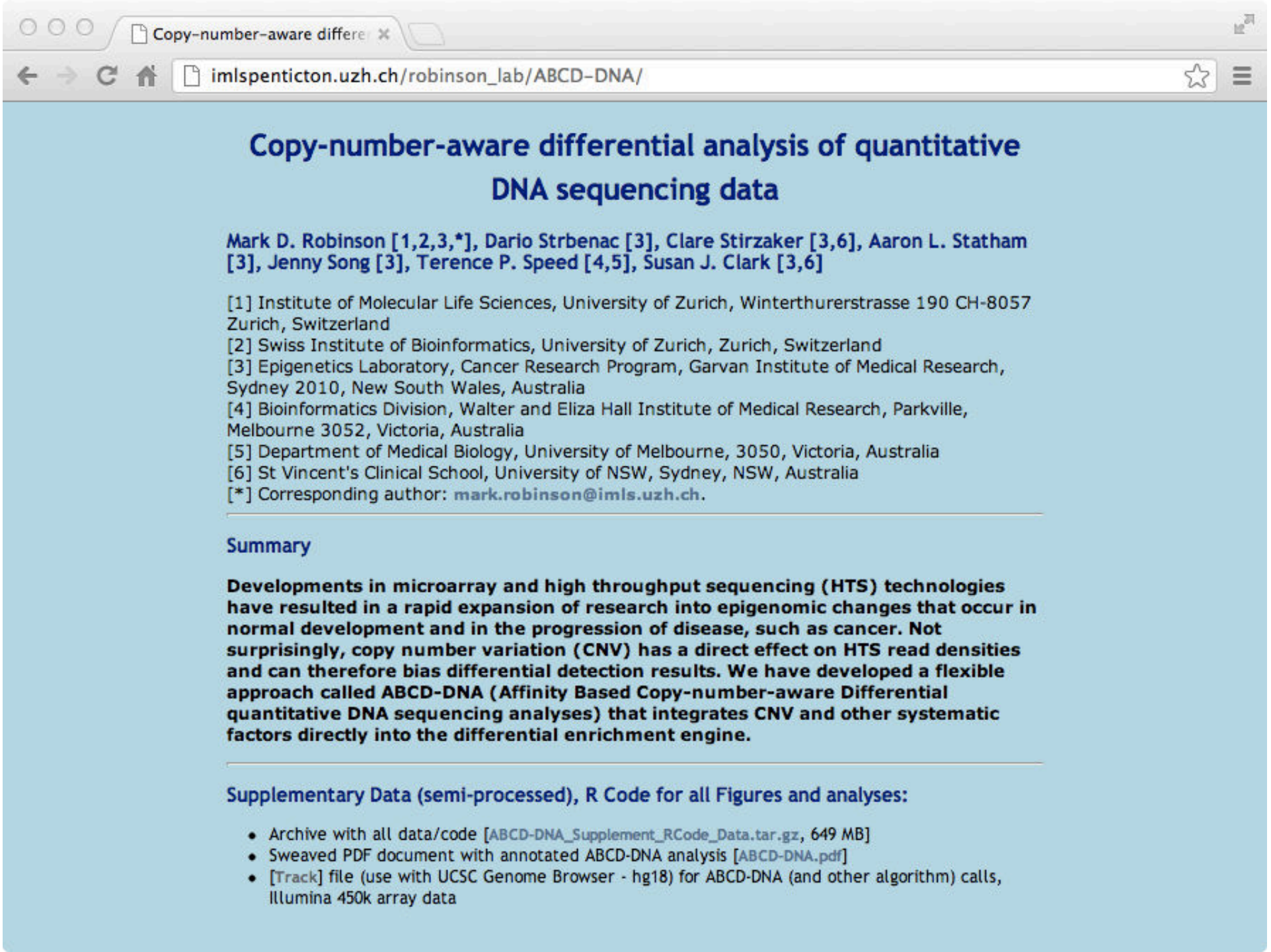
Some challenges in making R/ Bioconductor tools available (to Galaxy)

- How to allow facilitate general design matrices?

```
> 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_RCode_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