USE R 2016 CONFERENCE

shinyGEO: a web application for analyzing Gene Expression Omnibus (GEO) datasets using shiny

Jasmine Dumas[1], Michael A. Gargano[2], Garrett M. Dancik[2]

[1]DePaul University, Chicago, Illinois, USA, 2015 (Johns Hopkins Engineering For Professionals, 2016)
[2]Department of Mathematics and Computer Science, Eastern Connecticut State University, Willimantic, CT, USA

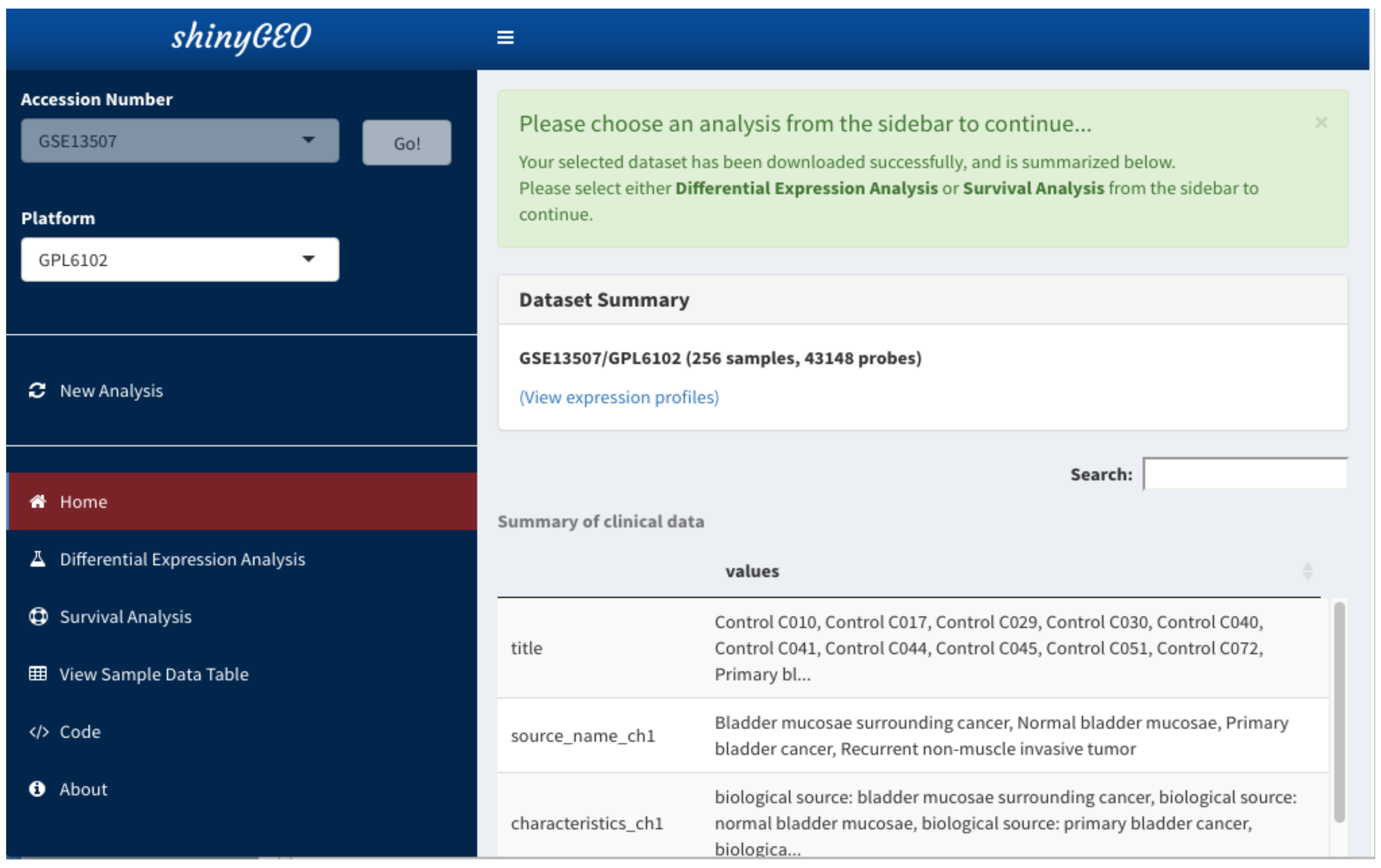
Introduction

Identifying associations between patient gene expression profiles and clinical data provides insight into the biological processes associated with health and disease. The Gene Expression Omnibus (GEO) is a public repository of gene expression and sequence-based datasets, and currently includes >42,000 datasets with gene expression profiles obtained by microarray. Although GEO has its own analysis tool (GEO2R) for identifying differentially expressed genes, the tool is not designed for survival analysis and does not generate publication-ready graphics. In this work, we describe a web-based, easy-to-use tool for biomarker analysis in GEO datasets, called *shinyGEO*.

Methods

The tool is developed using *shiny*, a web application framework for R. Specifically, *shinyGEO* allows a user to download the expression and clinical data from a GEO dataset, to modify the dataset correcting for spelling and misaligned data frame columns, to select a gene of interest, and to perform a survival or differential expression analysis using the available data. The tool uses the Bioconductor package *GEOquery* to retrieve the GEO dataset, while survival and differential expression analyses are carried out using the *survival* and *stats* packages, respectively. For both analyses, *shinyGEO* produces publication-ready graphics using *ggplot2* and generates the corresponding R code to ensure that all analyses are reproducible.

Results



Please choose an analysis from the sidebar to continue...

Your selected dataset has been downloaded successfully, and is summarized below.
Please select either **Differential Expression Analysis** or **Survival Analysis** from the sidebar to continue.

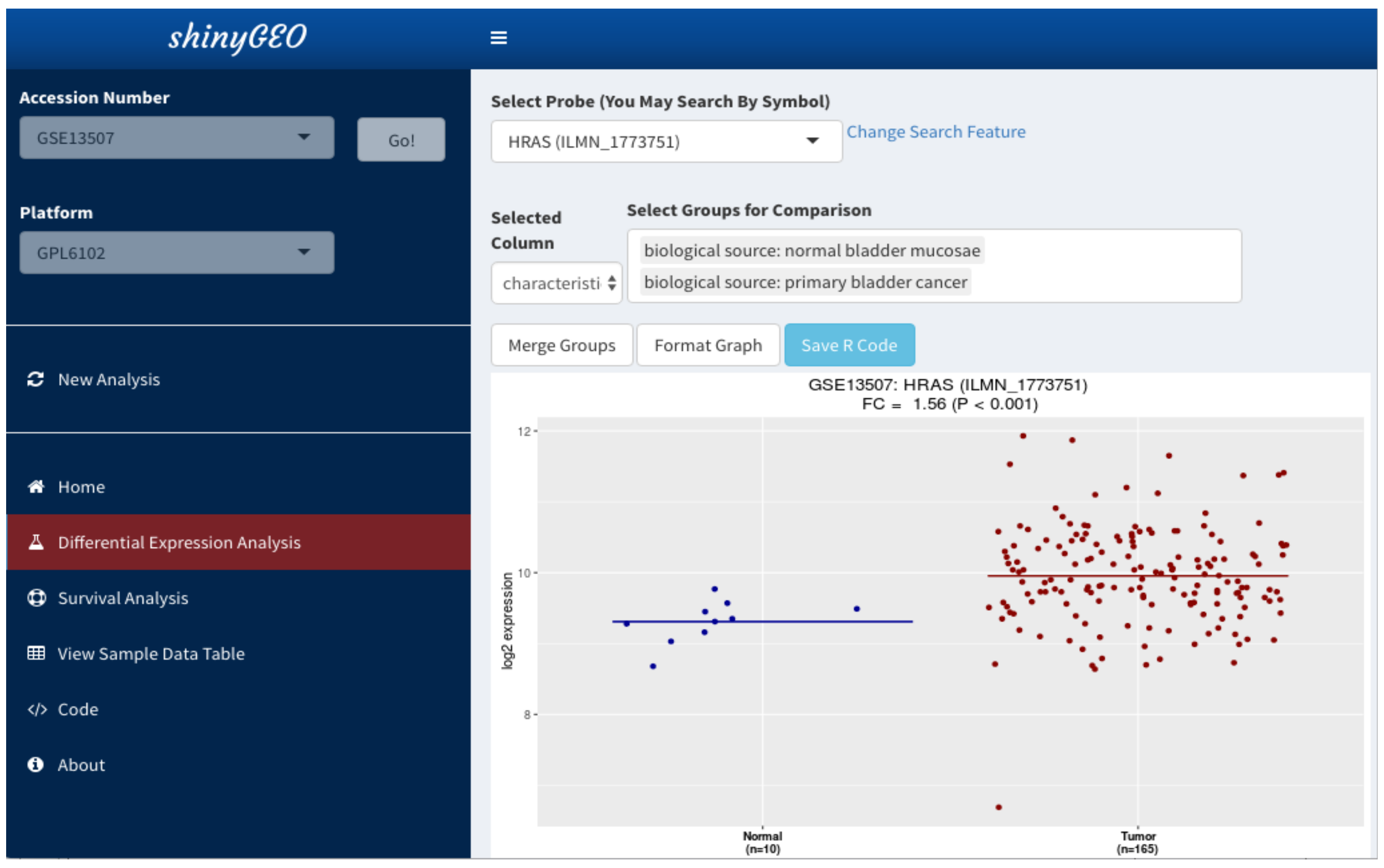
Dataset Summary

GSE13507/GPL6102 (256 samples, 43148 probes)
(View expression profiles)

Summary of clinical data

	values
title	Control C010, Control C017, Control C029, Control C030, Control C040, Control C041, Control C044, Control C045, Control C051, Control C072, Primary bl...
source_name_ch1	Bladder mucosae surrounding cancer, Normal bladder mucosae, Primary bladder cancer, Recurrent non-muscle invasive tumor
characteristics_ch1	biological source: bladder mucosae surrounding cancer, biological source: normal bladder mucosae, biological source: primary bladder cancer, biologica...

Figure 1. Screenshot of *shinyGEO* following dataset selection. The bladder cancer dataset GSE13507 has been selected and downloaded. From the homepage, the main panel summarizes the dataset. From the sidebar, the user can choose to carry out a differential expression analysis or survival analysis, view the sample data table, view the *R* code used to generate the analyses, or get more information about *shinyGEO*.

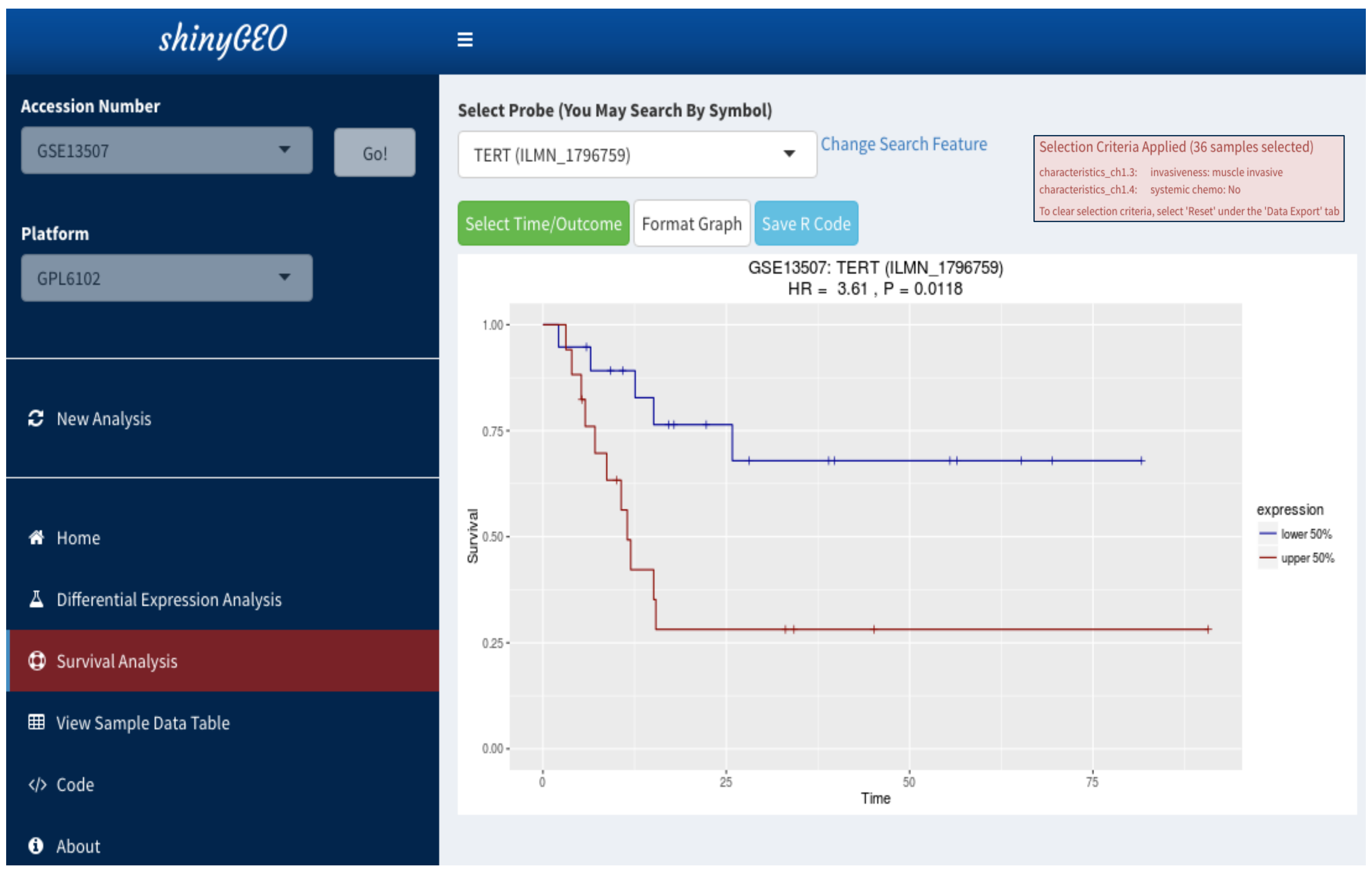


Select Probe (You May Search By Symbol): HRAS (ILMN_1773751)

Selected Column: characteristi

Select Groups for Comparison: biological source: normal bladder mucosae; biological source: primary bladder cancer

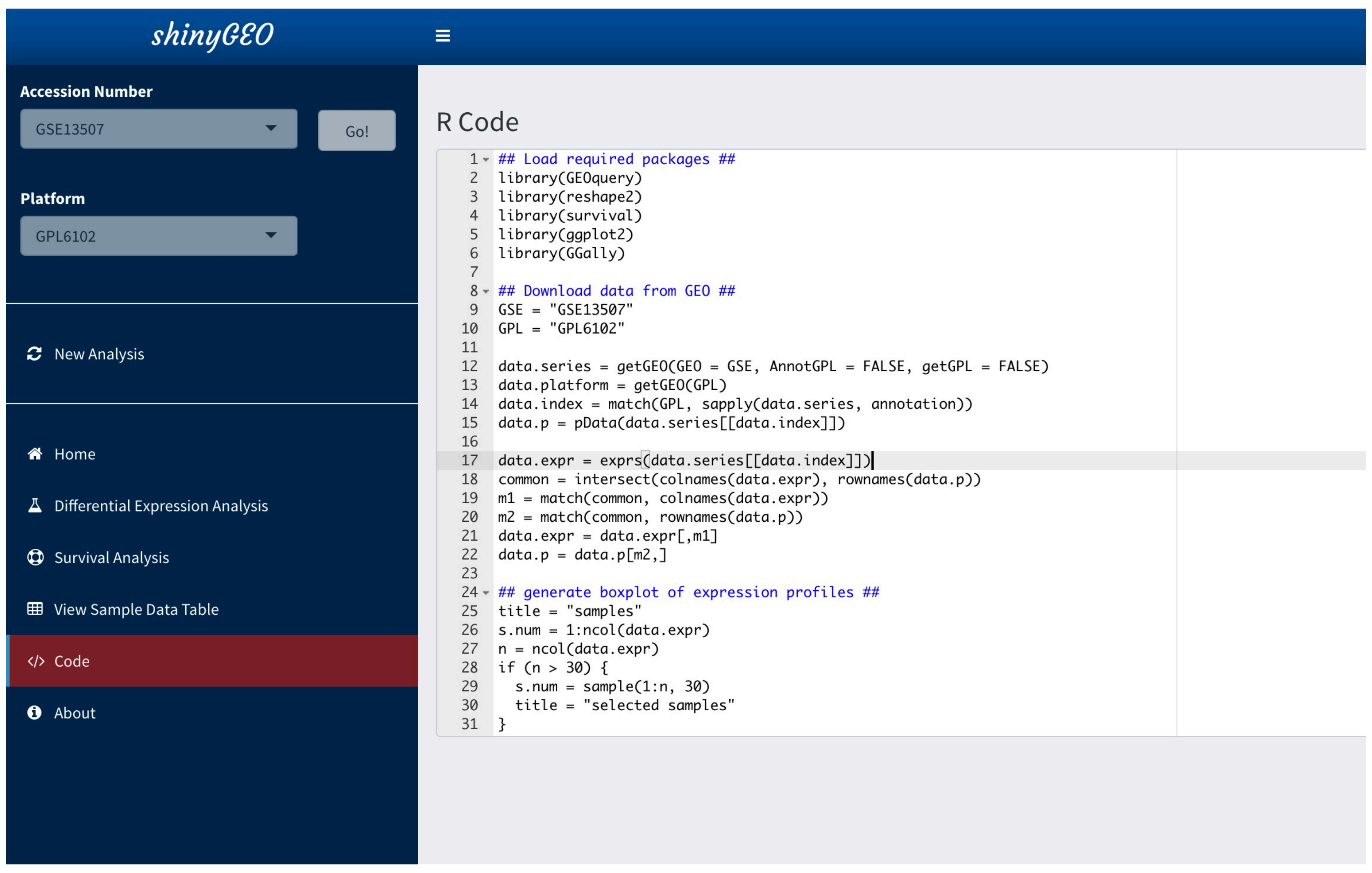
Figure 2. Screenshot of *shinyGEO* following differential expression analysis. Results indicate that the oncogene HRAS is significantly up-regulated in bladder tumors (FC = 1.56, P < 0.001) in this cohort. For differential expression analyses, the user selects the probe and the groups to compare. Optionally, the user can format the graph and save the *R* code used to generate the analysis.



Select Probe (You May Search By Symbol): TERT (ILMN_1796759)

Selection Criteria Applied (36 samples selected)
characteristics_ch1.3: invasiveness: muscle invasive
characteristics_ch1.4: systemic chemo: No
To clear selection criteria, select 'Reset' under the 'Data Export' tab

Figure 3. Screenshot of *shinyGEO* following survival analysis. In this analysis, we select patients with muscle-invasive tumors who did not receive chemotherapy (inset). Results indicate that high TERT expression is significantly associated with poor disease- specific survival (DSS) in these patients (HR = 3.61, P = 0.0118), reproducing the analysis of Borah et al. (Science 347, 1006–1010 (2015)). For survival analyses, the user selects the probe and columns containing time and outcome information.



R Code

```
## Load required packages ##
library(GEOquery)
library(reshape2)
library(survival)
library(ggplot2)
library(GGally)

## Download data from GEO ##
GSE = "GSE13507"
GPL = "GPL6102"

data.series = getGEO(GEO = GSE, AnnotGPL = FALSE, getGPL = FALSE)
data.platform = getGEO(GPL)
data.index = match(GPL, sapply(data.series, annotation))
data.p = pData(data.series[[data.index]])

data.expr = exprs(data.series[[data.index]])
common = intersect(colnames(data.expr), rownames(data.p))
m1 = match(common, colnames(data.expr))
m2 = match(common, rownames(data.p))
data.expr = data.expr[,m1]
data.p = data.p[m2,]

## generate boxplot of expression profiles ##
title = "samples"
s.num = 1:ncol(data.expr)
n = ncol(data.expr)
if (n > 30) {
  s.num = sample(1:n, 30)
  title = "selected samples"
}
```

Figure 4. Screenshot of *shinyGEO* following view of reproducbile *R* code. R code is automatically generated following dataset download. The user can optionally save *R* code following differential expression or survival analysis.

Availability: http://gdancik.github.io/shinyGEO/

Acknowledgements: This work was supported, in part, by Google Summer of Code 2015 funding to JD.