Bayesian screening for feature selection

A. Lawrence Gould
Richard Baumgartner
Amanda Zhao

goulda@merck.com

OVERVIEW

I. Introduction

The software has been written with the aims of simplicity and ease of use for most users. There should seldom if ever be a need for a user to write code to carry out the computations. At most, a user will need to do simple manipulations or perhaps copy output to a spreadsheet for simple formatting in order to incorporate the result in a document.

The software is designed to be used in the RStudio environment, although can be used in without RStudio. The Supplemental Material contains the following:

```
                app.r = R source file with text representations of all functions
               .RData = R data file used by RStudio
        NrmScrn.Rproj = RStudio project identification
         testData.rda = Test data sets used in the paper
           readme.txt = Brief use instructions
AnnotatedSession.pdf    This document
```

The first step is to copy the first three items to a convenient directory (ideally project-specific). The **testData.rda** file provides a way to reproduce the calculations in the paper, and is not necessary for execution of the programs.

The directory needs to contain files providing the project-specific data that will be used for the analysis. A typical data file is assumed to be a list or data frame containing at least the following

1. Control group feature measurements (could be the means of n replicates)
2. Test group feature measurements (likewise)
3. A pooled within-group sum of squares for each feature
4. A multiplier for the precision of the control group measurements, e.g., if each control group measurement is the mean of n replicates, then the multiplier is n, and likewise for the test group
5. The degrees of freedom associated with the pooled within-group sum of squares
6. An indicator of whether the feature expression value is known to differ between the control and test groups (this is useful for assessing diagnostic properties based on simulated data, and is not otherwise used for analyzing real data)

These variables can have any names. However, the data files are assumed to adhere to a naming convention of the form **root_dta** to simplify keeping track of various output files.

A shiny interface is used to provide input to a driver function that performs specific calculations and writes summary lists to the workspace. A typical example of the use of the interface is provided

below. The function **Plots_and_Tables.fn** provides a variety of summary statistics and graphics based on the result of carrying out the screening calculations.

The workspace that is supplied initially contains just one function:

```
runNrmScrn.fn <- function()
{
  source('app.r')
  shinyApp(ui, server)
}
```

The first time the software is executed, type

```
runNrmScrn.fn()
```

at the > prompt. Multiple executions using the shiny interface can be carried out after the first execution by typing

```
shinyApp(ui, server)
```

at the prompt because the source material will have been entered at the first execution.

There are, in addition, 3 special purpose functions that might be used rarely

gen_test.fn Generates a simulated dataset for evaluating diagnostic properties

Table_A2_1.fn, Table_A2_2.fn Produce the tables in Appendix 2 of the paper

The **testData.rda** file contains five datasets, of which **ALLAML_dta** will be used to illustrate the use of the software.

```
> dim(as.data.frame(ALLAML_dta))
[1] 7128    6
> as.data.frame(ALLAML_dta)[1:5,]
        Y_ALL      Y_AML aL aM         u  m
1 -1.06756038 -0.9284565 47 25  5.860731 70
2 -1.17759359 -1.1724987 47 25  4.161429 70
3 -0.62480419 -0.5385304 47 25 16.813458 70
4  0.09199291  0.2313916 47 25  6.404681 70
5 -1.39577279 -1.4113315 47 25  4.979795 70
```

The calculations will be carried out using the dataset twice: once with **Y_ALL** and **aL** identified with the test group and **Y_AML** and **aM** identified as the control group, and once with the identifications reversed.

II Annotated Session Log

```
> sessionInfo()

R version 3.6.0 (2019-04-26)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 17763)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252  LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252 LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

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

loaded via a namespace (and not attached):
[1] compiler_3.6.0 tools_3.6.0    packrat_0.5.0

> load("testData.rda")

> ls()

[1] "ALLAML_dta"     "runNrmScrn.fn"  "test.005_4_dta" "test.005_6_dta"
"test.25_4_dta"  "test.25_6_dta"

> set.seed(18)

> runNrmScrn.fn()

Listening on http://127.0.0.1:5957
```

This statement produces a graphical input panel, here filled in with an example.

Executing this last statement produces an output file, **`ALLT_AMLC_out`** and loads additional objects into the workspace:

```
> ls()
```

```
 [1]  "ALLAML_dta"           "ALLT_AMLC_out"     "Elapsed.time.fn"
 [4]  "FDRMDRplots.fn"       "gen_test.fn"       "NormBayesScreenMenu.fn"
 [7]  "Plots_and_Tables.fn"  "runNrmScrn.fn"     "server"
[10]  "Setup.fn"             "string_to_num.fn"  "Table_A2_1.fn"
[13]  "Table_A2_2.fn"        "test.005_4_dta"    "test.005_6_dta"
[16]  "test.25_4_dta"        "test.25_6_dta"     "ui"
[19]  "vec2mat.fn"
```

The Setup.fn function, which is invoked automatically, brings additional packages into the enviroment,

```
> Setup.fn()
```

```
                              . . .
 other attached packages:
 [1] shinyBS_0.61 bsplus_0.1.1 htmltools_0.4.0 shiny_1.4.0    SetTest_0.2.0 abind_1.4-5     sjmisc_2.8.2
 [8] scales_1.1.0 ashr_2.2-39  qvalue_2.18.0   gridExtra_2.3 ggplot2_3.3.0 lattice_0.20-38 stringr_1.4.0
```

```
loaded via a namespace (and not attached):
```

[1]	Rcpp_1.0.3	lubridate_1.7.4	assertthat_0.2.1	packrat_0.5.0
[5]	digest_0.6.23	foreach_1.4.7	mime_0.8	truncnorm_1.0-8
[9]	R6_2.4.1	plyr_1.8.5	pillar_1.4.3	rlang_0.4.5
[13]	pscl_1.5.2	rstudioapi_0.10	Matrix_1.2-17	labeling_0.3
[17]	splines_3.6.0	munsell_0.5.0	mixsqp_0.2-2	compiler_3.6.0
[21]	httpuv_1.5.2	pkgconfig_2.0.3	SQUAREM_2017.10-1	insight_0.7.1
[25]	tidyselect_1.0.0	tibble_3.0.0	codetools_0.2-16	fansi_0.4.0
[29]	crayon_1.3.4	dplyr_0.8.5	withr_2.1.2	later_1.0.0
[33]	MASS_7.3-51.4	grid_3.6.0	jsonlite_1.6	xtable_1.8-4
[37]	gtable_0.3.0	lifecycle_0.2.0	magrittr_1.5	cli_2.0.0
[41]	stringi_1.4.3	farver_2.0.1	reshape2_1.4.3	promises_1.1.0
[45]	doParallel_1.0.15	ellipsis_0.3.0	vctrs_0.2.4	sjlabelled_1.1.1
[49]	iterators_1.0.12	tools_3.6.0	forcats_0.4.0	glue_1.3.1
[53]	purrr_0.3.3	hms_0.5.2	rsconnect_0.8.16	parallel_3.6.0
[57]	fastmap_1.0.1	colorspace_1.4-1	haven_2.2.0	

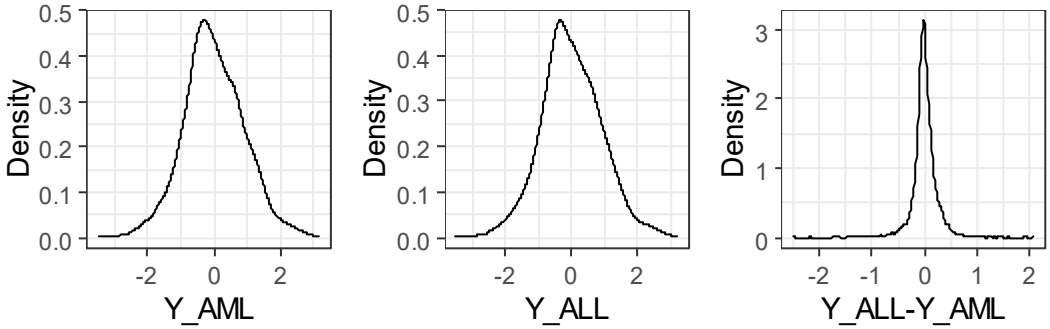
The output object **ALLT_AMLC_out** has a number of components

```
> attributes(ALLT_AMLC_out)
$names
 [1] "call"          "date"         "dta"          "d"        "dcrit"
 [6] "diffcdf"       "wcrit"        "ind_crit"     "F_D"      "p_gam0"
[11] "data_plots"    "pi0_den_num"  "pi0_den_plts" "thetaC"   "p_gam0_tabs"
[16] "zeta1"         "zeta2"        "phi0"         "phi1"     "Elapsed.Time"
```

The function **Plots_and_Tables.fn"** produces graphical and tabular summaries of the computational results

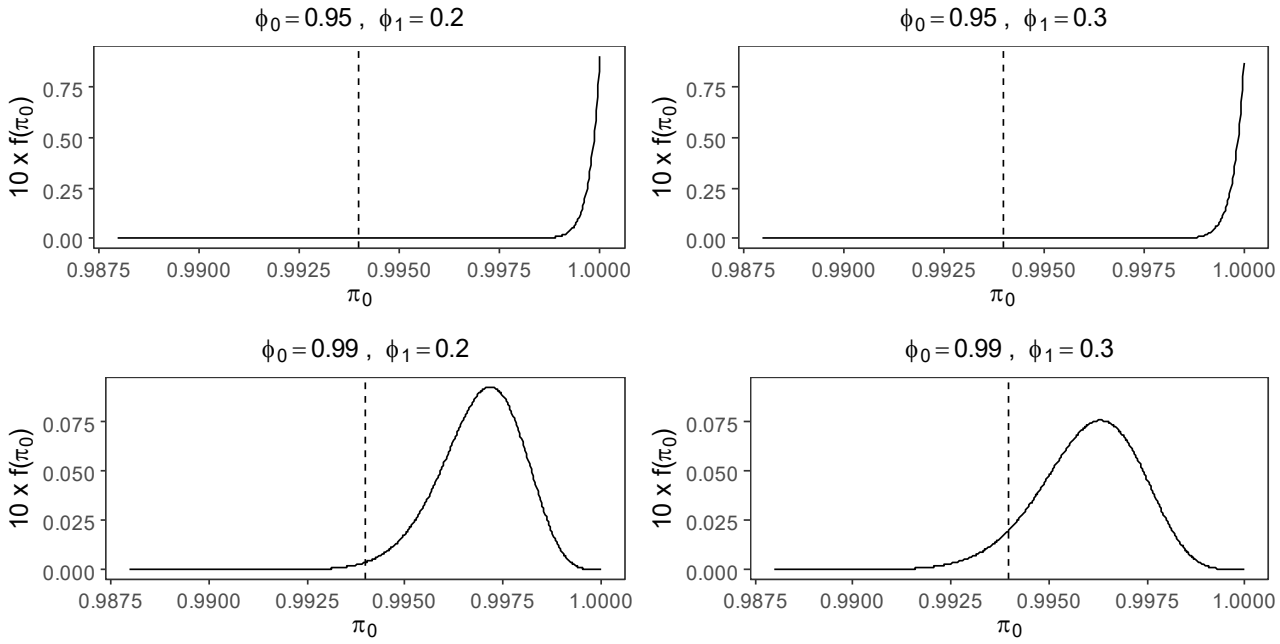
```
> Plots_and_Tables.fn()

What is the name of the output from NormBayesScreenMenu.fn
 (typically something like foo_out)?

ALLT_AMLC_out

What do you want to do? (Enter a number or 0 to exit)

1: Plot densities of T and C responses and their difference
2: Plot posterior densities of pi0 for each dcrit value
3: Get table of pSP for features with pSP < wcrit
4: Get table of pSP for all features (a big matrix)
5: Get table of posterior CDF of T-C feature differences
6: Get table with pfdr q statistics and flags added
7: Get table of features selected by BJ and HC criteria
8: Get table of features selected by ashr (Emp Bayes)

Selection: 1
```

produces

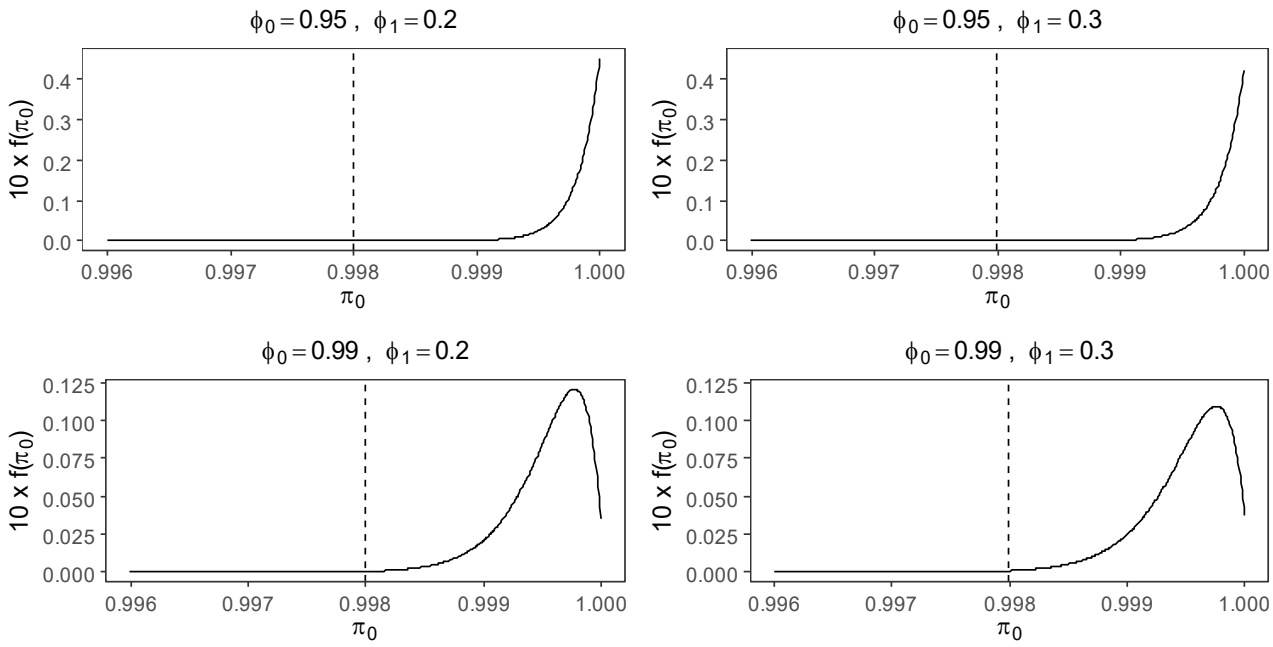


The menu will reappear, and another selection can be made:

```
Selection: 2

There are 4 density plots for each dcrit value.
In how many rows do you want them arranged?
Enter the e.g., 2 if there are 6 plots (<CR> -> default value = 1):
2

Choose one or more dcrit values (0 for all)

1: dcrit=0.5
2: dcrit=0.6

Selection: 0

[1] "Plot for dcrit = 0.5"
[1] "Press <CR> when you are ready for the next plot"
```




```
[1] "Plot for dcrit = 0.6"
```



Back to the output selection menu,

Selection: **3**

Choose a dcrit value

1: dcrit=0.5
2: dcrit=0.6

Selection: **1**

Choose a wcrit value

1: wcrit=0.2
2: wcrit=0.3
3: wcrit=0.4

Selection: **2**

writes the object **ALLT_AMLC_pgam0tab_1_2** to the workspace. It looks like this

```
> ALLT_AMLC_pgam0tab_1_2
```

```
$date
[1] "Wed Apr 29 14:32:17 2020"

 $dcrit
 [1] 0.5

 $wcrit
 [1] 0.3

 $result
```

	Ftr_#	YT	YC	Diff	seDiff	zDiff	pval	p0=0.95 p1=0.2	p0=0.95 p1=0.3	p0=0.99 p1=0.2	p0=0.99 p1=0.3
1	804	1.82	1.24	0.57	0.05	10.76	0	0.99	0.98	0.18	0.16
2	1144	0.96	0.05	0.91	0.1	9.31	0	0.98	0.99	0.11	0.12

3	2354	1.81	1.07	0.74	0.07	9.9	0	0.98	0.98	0.09	0.1
4	4328	1.81	1.13	0.68	0.08	8.99	0	0.99	0.99	0.25	0.24
5	6281	0.72	0.06	0.65	0.08	8.63	0	0.99	0.99	0.19	0.15
6	6854	1.31	0.43	0.88	0.07	12.26	0	0.76	0.8	0	0

The entries in the last 4 columns are the posterior probabilities that the same process generated the control (AML) and test (ALL) feature values.

```
Continuing with the Plots_and_Tables menu,

Selection: 5

Choose a dcrit value
```

Posterior CDF of T-C feature differences

```
1: dcrit=0.5
2: dcrit=0.6

Selection: 1

Choose a wcrit value

1: wcrit=0.2
2: wcrit=0.3
3: wcrit=0.4

Selection: 2

Choose a phi0, phi1 combination

1: phi0, phi1 = 0.95,0.2
2: phi0, phi1 = 0.95,0.3
3: phi0, phi1 = 0.99,0.2
4: phi0, phi1 = 0.99,0.3

Selection: 3

What do you want to do? (Enter a number or 0 to exit)

Selection: 6
```

pfdr q statistics and flags

```
Current qcrit values:

1e-05 2e-05 3e-05 4e-05 5e-05 6e-05 7e-05 8e-05 9e-05 1e-04 2e-04 3e-04 4e-04
5e-04 6e-04 7e-04 8e-04 9e-04 0.001 0.0011 0.0012 0.0013 0.0014 0.0015 0.0016
0.0017 0.0018 0.0019 0.002 0.0021 0.0022 0.0023 0.0024 0.0025 0.0026 0.0027
0.0028 0.0029 0.003 0.0031 0.0032 0.0033 0.0034 0.0035 0.0036 0.0037 0.0038
0.0039 0.004 0.0041 0.0042 0.0043 0.0044 0.0045 0.0046 0.0047 0.0048 0.0049
0.005 0.006 0.007 0.008 0.009 0.01

Do you want to change the values?

1: Yes
2: No

Selection: 2

1-sided or 2-sided test?

1: 1-sided
2: 2-sided
```

```
Selection: 2

What do you want to do? (Enter a number or 0 to exit)

Selection: 7                    Features selected by HC and BJ (Higher Criticism)

What do you want to do?
(Enter a number or 0 to exit)

Selection: 8                    Features selected by Empirical Bayes

Current lfsr bound = 1e-8

Do you want to change the values?

1: Yes
2: No

Selection: 2

What do you want to do? (Enter a number or 0 to exit)

Selection: 0
```

This last option exits the Plots_and_Tables.fn function. Here's what the workspace contains now:

```
 > ls()
 [1] "ALLAML_dta"               "ALLT_AMLC_ash"      "ALLT_AMLC_BJ"
 [4] "ALLT_AMLC_diffcdf_1_2_3" "ALLT_AMLC_HC"       "ALLT_AMLC_out"
 [7] "ALLT_AMLC_pgam0tab_1_2"  "ALLT_AMLC_qvf_2"    "Elapsed.time.fn"
[10] "FDRMDRplots.fn"          "gen_test.fn"        "NormBayesScreenMenu.fn"
[13] "Plots_and_Tables.fn"     "runNrmScrn.fn"      "server"
[16] "Setup.fn"                "string_to_num.fn"   "Table_A2_1.fn"
[19] "Table_A2_2.fn"           "test.005_4_dta"     "test.005_6_dta"
[22] "test.25_4_dta"           "test.25_6_dta"      "ui"
[25] "vec2mat.fn"

> ALLT_AMLC_diffcdf_1_2_3

$date
[1] "Wed Apr 29 14:32:17 2020"

$dcrit
[1] 0.5

$wcrit
[1] 0.3

$phi0phi1
[1] "phi0, phi1 = 0.99,0.2"

$result
     Ftr_# Diff  seDiff zDiff  pval pgam0 d=0.4 d=0.45 d=0.5 d=0.55 d=0.6 d=0.65 d=0.7
[1,] 804   0.574 0.053  10.76  0    0.181 1     1      1     0.998  0.982 0.903  0.683
[2,] 1144  0.908 0.098  9.309  0    0.107 1     1      1     1      0.999 0.996  0.986
[3,] 2354  0.737 0.074  9.903  0    0.092 1     1      1     1      1     0.997  0.985
[4,] 4328  0.681 0.076  8.99   0    0.246 1     1      1     0.999  0.995 0.977  0.917
[5,] 6281  0.652 0.076  8.626  0    0.186 1     0.997  0.984 0.938  0.817 0.603  0.345
[6,] 6854  0.876 0.071  12.261 0    0     1     1      1     1      1     1      0.999
```

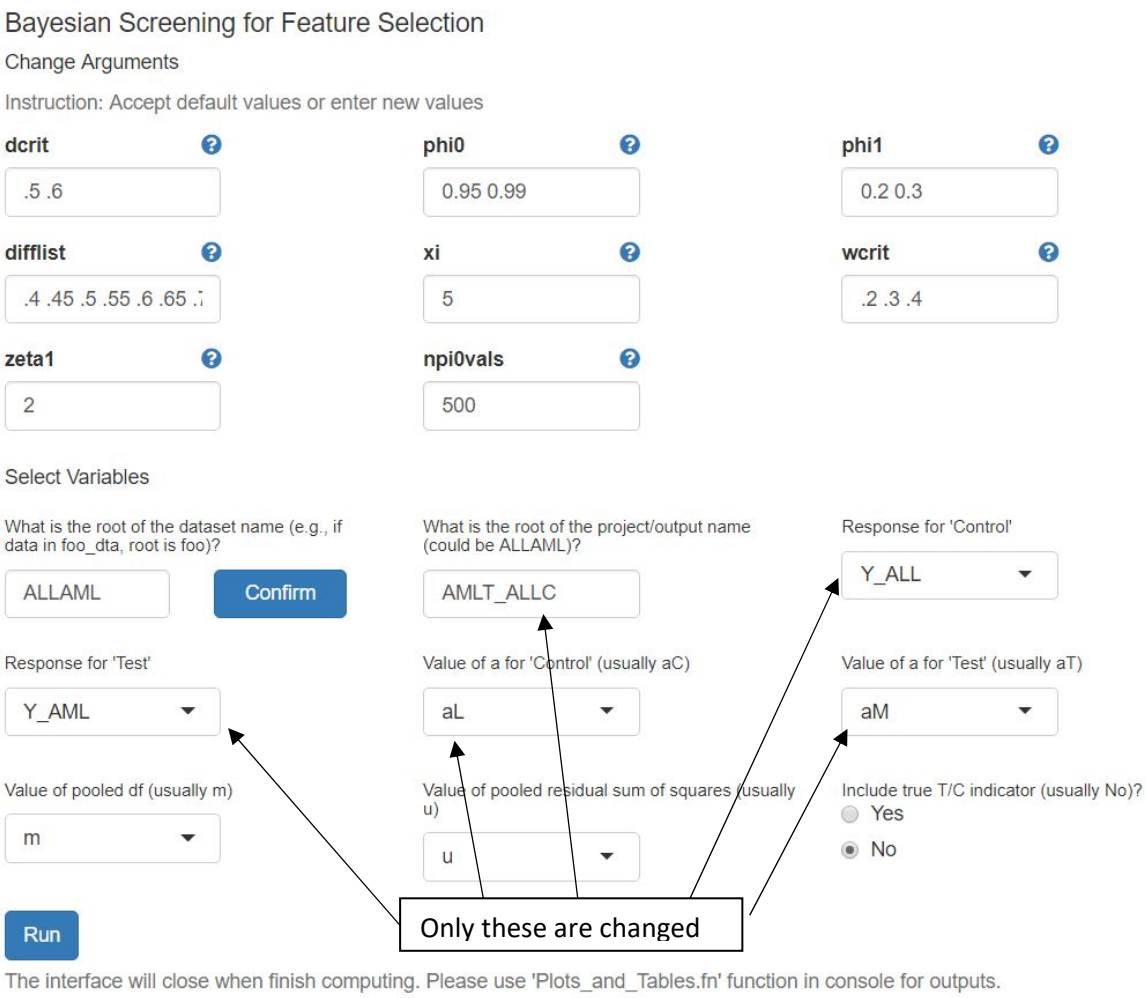
Displays of ALLT_AMLC_ash" "ALLT_AMLC_BJ", and ALLT_AMLC_qvf_2 are omitted because these

Displays of **ALLT_AMLC_ash**, **ALLT_AMLC_BJ**, and **ALLT_AMLC_qvf_2** are omitted because the include large arrays of statistic values corresponding to each feature that can be summarized using standard R function. However, ALLT_AMLC_HC is of manageable size,

```
> ALLT_AMLC_HC
```

	Feat	YC	YT	aC	aT	m	u	Diff	seDiff	zDiff	pval
804	804	1.242	1.816	25	47	70	3.253	0.574	0.053	10.760	0.00E+00
1144	1144	0.047	0.955	25	47	70	10.878	0.908	0.098	9.309	0.00E+00
1685	1685	-0.046	1.564	25	47	70	35.272	1.609	0.176	9.159	0.00E+00
2354	2354	1.070	1.807	25	47	70	6.319	0.737	0.074	9.903	0.00E+00
2642	2642	0.441	1.650	25	47	70	25.339	1.209	0.149	8.118	2.22E-16
4211	4211	1.389	1.779	25	47	70	2.156	0.390	0.043	8.988	0.00E+00
4328	4328	1.127	1.808	25	47	70	6.551	0.681	0.076	8.990	0.00E+00
5501	5501	1.376	1.814	25	47	70	2.441	0.439	0.046	9.495	0.00E+00
6281	6281	0.064	0.716	25	47	70	6.525	0.652	0.076	8.626	0.00E+00
6854	6854	0.431	1.307	25	47	70	5.829	0.876	0.071	12.261	0.00E+00

All of the calculations displayed above can be carried out easily by interchanging ALL and AML in the original call of **runNrmScrn.fn**. The shiny panel is filled in only slightly differently,


Bayesian Screening for Feature Selection
Change Arguments
Instruction: Accept default values or enter new values
dcrit
.5 .6
phi0
0.95 0.99
phi1
0.2 0.3
difflist
.4 .45 .5 .55 .6 .65 .7
xi
5
wcrit
.2 .3 .4
zeta1
2
npi0vals
500
Select Variables
What is the root of the dataset name (e.g., if data in foo_dta, root is foo)?
ALLAML
Confirm
What is the root of the project/output name (could be ALLAML)?
AMLT_ALLC
Response for 'Control'
Y_ALL
Response for 'Test'
Y_AML
Value of a for 'Control' (usually aC)
aL
Value of a for 'Test' (usually aT)
aM
Value of pooled df (usually m)
m
Value of pooled residual sum of squares (usually u)
u
Include true T/C indicator (usually No)?
Yes
No
Only these are changed
Run
The interface will close when finish computing. Please use 'Plots_and_Tables.fn' function in console for outputs.