DDG-GUI Manual
Version 1.0(beta)

Diana HP Low[1], Efthimios Motakis[2] & Vladimir A. Kuznetsov[3]
Genome and Gene Expression Data Analysis Division
Bioinformatics Institute
Agency for Science and Technology Research (A*STAR), Singapore
dianal@bii.a-star.edu.sg[1]
efthimiosm@bii.a-star.edu.sg[2]
vladimirk@bii.a-star.edu.sg[3]

Updated: 22 November 2011

Introduction

Data-Driven grouping (DDg) a statistically-based feature (or variable) selection and patient grouping scheme which stratifies patients into two or more classes based on extreme multivariate statistics of the semi-parametric Cox proportional hazard regression model. DDg selects features whose critical cut-off values maximizes the separation between patient risk groups. Such groups are produced by individual survival significant variables (SSV, e.g. microarray genes), the synergistic survival significant paired variables (SSS-PV) and the voting multivariable survival significant (VSS) features.

This document provides instruction on how to use the DDG-GUI package implemented in R. This package is used to stratify patients into two or three risk groups via identification of survival significant genes based on microarray intensity data and the relevant patient survival information.

Installing DDG-GUI and dependencies

DDG-GUI can be downloaded from `http://web.bii.a-star.edu.sg/~dianal/DDG/`

Before running `DDG-GUI`, you will need the free `R` programming language which can be downloaded from `http://cran.r-project.org/bin/windows/base/`. Make sure you have the latest R version installed.

The `R survival` package can be installed from the `R` command prompt by typing:

```
> install.packages("survival")
```

`DDG-GUI` runs on top of R.

For Windows users : Simply double-click `DDG_v1.0b.Rdata` to launch the package.

For Linux users : Load the Rdata file within the R console and load the DDG_gui function.

```
> setwd("your_working_directory")
> load("DDG_v1.0b.Rdata")
> DDG_gui()
```

Using DDG-GUI

Step 1: Prepare your input files. You will need a matrix of expression intensities from a microarray experiment, the corresponding gene annotations for the microarray, and the clinical (survival) information for the samples in the microarray. You can download sample input files from:

`http://web.bii.a-star.edu.sg/~dianal/DDG/sample_input.zip`

The structure of the input files should be as follows:

Gene annotation : 2 column tab-separated text file with array IDs in the 1st column and the gene name in the 2nd column.

```
affy_id       gene
"A.1007_s_at" "DDR1"
"A.1053_at"   "RFC2"
"A.117_at"    "HSPA6"
"A.121_at"    "PAX8"
"A.1255_g_at" "GUCA1A"
"A.1294_at"   "UBE1L"
"A.1316_at"   "THRA"
"A.1320_at"   "PTPN21"
"A.1405_i_at" "CCL5"
"A.1431_at"   "CYP2E1"
```

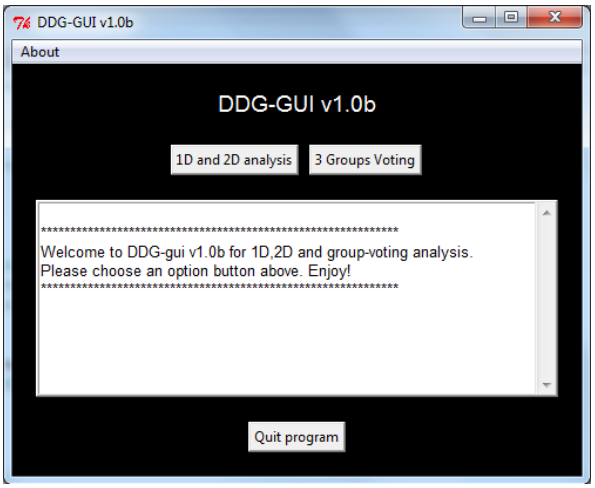
Clinical data : 3 column tab-seperated text file with patient ID, time and event information

```
Patient Time Event
 X003WI 8.48     0
 X005JO 5.55     0
 X010BJ 5.75     0
 X011DA 4.51     0
 X014ER 1.26     1
 X015HE 2.79     1
 X018GU  8.4     0
 X019ER 8.32     0
 X024BJ 6.13     0
 X026NA 7.75     0
```

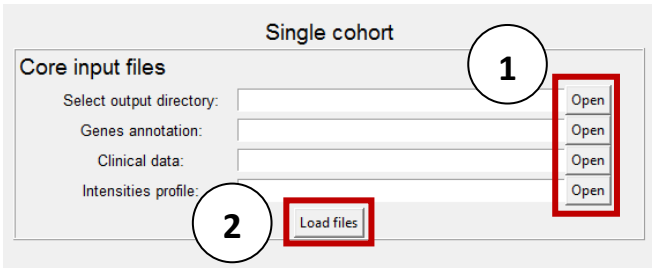
Microarray intensity file : matrix of tab-separated intensity values with array IDs in rows and patients in columns.

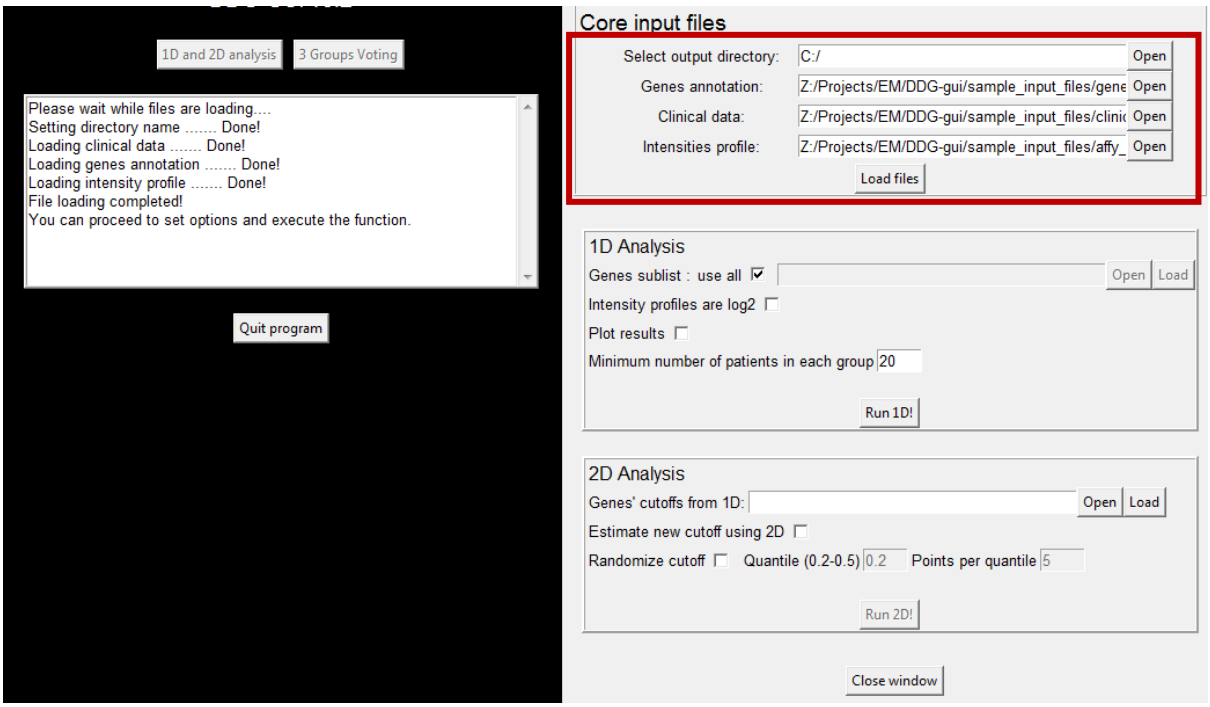
```
           X    X003WI    X005JO    X010BJ    X011DA
A.1007_s_at 10.559447 11.116684 10.619543 11.023815
  A.1053_at  6.145723    6.8313  6.982069  6.850756
   A.117_at  7.405281  7.345258  7.053148  7.445947
   A.121_at  8.988401  9.347112    9.0983  9.374462
A.1255_g_at  4.360498  4.474786  4.369169  4.289754
  A.1294_at  7.956301  8.299001  7.838204  8.625779
  A.1316_at  5.838735  6.043306  5.841217  5.919682
  A.1320_at  5.134566  5.277698  5.366259  5.274535
A.1405_i_at  6.816194  7.024171  5.599198  6.273459
  A.1431_at  4.225934  4.399599  4.253313  4.233773
```

Step 2: Start `DDG-GUI`. To start `DDG-GUI`, simply double click the `DDG_v1.0b.Rdata` file. R will automatically launch and you will be presented with the interface window (Figure 1). Click the **1D and 2D analysis** button to begin analysis.



***Step 3*:** Load input files. To load your input files, click **Open** and choose the corresponding files to load. After this has been done, click **Load files** to read the files into the R environment. You can monitor the loading progress on the left text box.



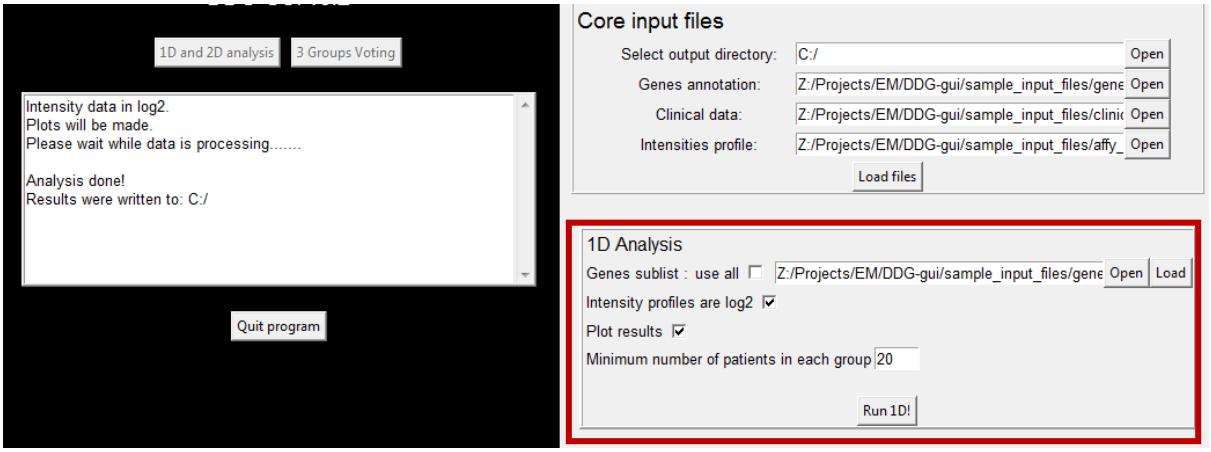


***Step 4*:** Running the 1D analysis. To find survival significant genes, we now run the 1D analysis.

To find genes over the entire gene set, simply leave the **use all** tick-box checked. However, iterating over the full set of genes is time consuming and if you have identified a subset of genes to analyze, un-check the tick box and select a smaller set of genes (Remember to **Load** the file after opening!).

Also indicate if the microarray intensity values provided are in log2 (the sample intensity file provided are in log2). You can also choose to plot the survival curves for all the genes chosen for analysis. The last option allows you to specify the minimum number of samples allowed in a group for it to be considered survival significant.

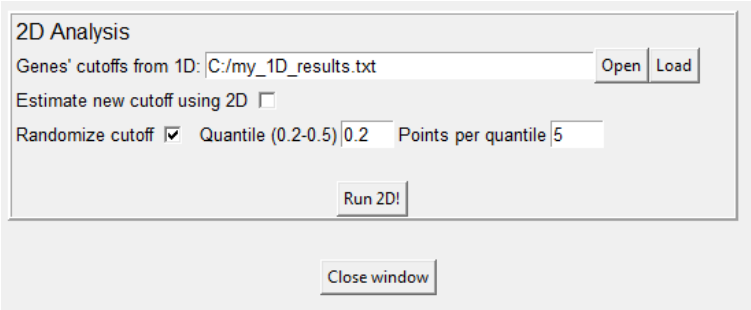
Click **Run 1D!** when you are done. Progress again will be indicated on the left text box and more detailed background information can be seen on the R command line. Results files will be discussed in Step 6.



***Step 5*:** Running the 2D analysis. The 2D analysis allows you to find significant gene pairs. The algorithm allows you to use the genes identified from the 1D analysis and continue on with the 2D analysis. The gene list supplied will be permuted and the resulting pairs will be used to calculate the gene-pair p-values. Here, you have the option for the analysis to use the previous gene expression cut-offs for survival from the 1D analysis, or to re-estimate the cut-offs based on the gene-pairs.

For data analysis that requires 2D cut-offs re-estimation the algorithm searches for all possible cut-off pairs which can be time consuming (depending on the number of patients' samples). To reduce processing time and get a good approximation of the final results, we introduce three additional parameters: "randomize cut-off", "Quantile" and "Points per quantile".

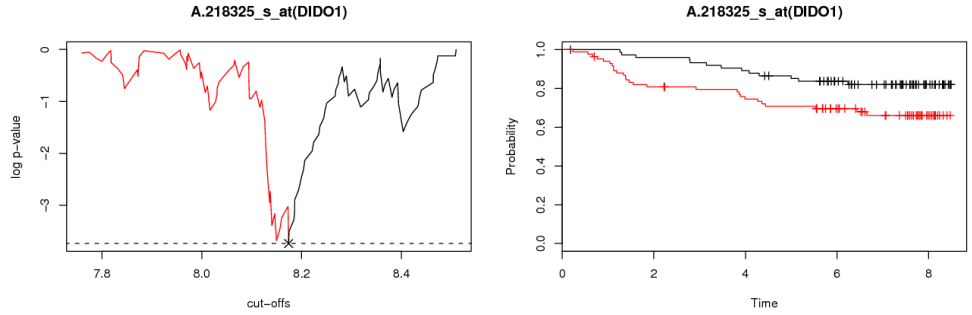
Cut-off randomization under cut-off re-estimation using 2D: if cut-off randomization is selected the algorithm finds the ith quantiles of each probeset's data (variable "Quantile" ranging from 10% to 50%) and within each quantile it samples at random a selected number of values (variable "Points per quantile"). These values are then used as trial cut-offs. Suggested values for our Stockholm (159 patients) and Uppsala (251 patients) cohorts are: Quantile (0.1-0.5) = 0.2 and Points per quantile = 5.



***Step 6*:** Analyzing the results.

1D analysis

affyID	genename	1D cutoff	1D pvalue	FDR adjusted pvalue	design	mean intensity (low-risk)	mean intensity (high-risk)	fold change	wilcoxon pvalue	zph test	number of low-risks	number of high-risks	X003WI	X005JO	X010BJ	X011DA
A.218324_s_at	SPATS2	6.924979	0.198711039	0.198711039	2	6.484303055	7.147331484	1.583402933	6.50E-18	0.038764696	128	31	1	1	2	1
A.218325_s_at	DIDO1	8.173964	0.023775912	0.055231849	1	8.488242693	7.909202512	1.493855063	1.66E-27	0.107843616	75	84	1	2	2	1
A.218326_s_at	LGR4	5.103521	0.165736915	0.198711039	2	4.77073564	5.634329743	1.819565648	5.15E-24	0.591337043	50	109	2	1	2	2
A.218327_s_at	SNAP29	8.460486	0.003643943	0.03643943	2	8.123124079	8.695206031	1.486667432	2.61E-18	0.002305509	127	32	2	1	1	2
A.218328_at	COQ4	7.982375	0.166711531	0.198711039	1	8.242606768	7.646810845	1.511306131	2.54E-25	0.190522322	56	103	2	2	2	1
A.218329_at	PRDM4	7.891556	0.1738874	0.198711039	1	8.041425094	7.614160496	1.344681593	2.61E-18	0.570864724	32	127	1	2	1	2
A.218330_s_at	NAV2	6.352476	0.012093942	0.040313142	1	7.307448057	5.793709378	2.855490686	3.65E-20	0.352873217	122	37	1	1	1	1
A.218332_at	BEX1	5.696361	0.008327271	0.040313142	1	6.665150978	5.613621542	2.072726031	6.60E-15	0.407682137	135	24	1	1	1	1
A.218333_at	DERL2	7.882752	0.189713928	0.198711039	2	7.676968936	8.259595036	1.497572761	2.99E-23	0.100359608	47	112	1	2	1	2
A.218334_at	THOC7	8.405132	0.027615925	0.055231849	2	8.077734128	8.94186195	1.820238913	7.57E-21	0.688656708	39	120	2	2	2	2



Text file output:

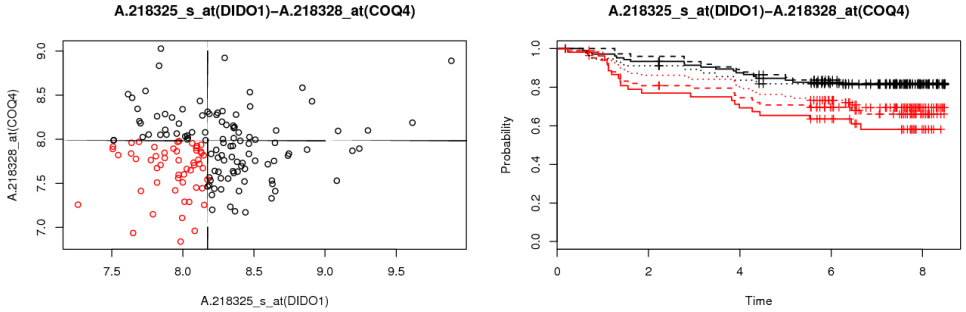
- "affyID" = Affymetrix ID
- "genename" = RefSeq gene name
- "1D cut-off" = 1D cut-off of microarray expression intensity
- "1D pvalue" = 1D Wald p-value estimated by 1D DDg
- "FDR adjusted" = FDR adjusted Wald p-value
- "design" = if 1, low expression is associated with low-risk patients and high-expression with high-risk patients; if 2, the opposite association holds
- "mean intensity (low-risk)" = mean of the intensities of the low-risk group
- "mean intensity (high-risk)" = mean of the intensities of the high-risk group
- "fold change" = 2^(|mean intensity (low-risk) - mean intensity (high-risk)|)
- "wilcoxon pvalue" = wilcoxon test for the equality of the medians between low-risk and high-risk groups
- "zph test" = pvalues of the test for proportional hazards (low pvalues indicate violation of the assumption)
- "number of low-risks " = number of low-risk patients
- "number of high-risks " = number of high-risk patients
- the rest of the columns are the patient identifiers and their grouping for each probeset. 1 is associated with low-risk group and 2 with high-risk group.

Graphs:

- Left: Patients grouping using 1D cut-off; Right: Kaplan Meier survival curves
- For both: black lines = low-risk patients, red lines = high risk patients
- For this example, the high-risk patients have lower expression for the gene

2D analysis

affyID1	affyID2	genename1	genename2	1D pvalue1	1D pvalue2	1D cutoff1	1D cutoff2	design	2D p-value	FDR adjusted pvalue	zph test	number of low-risks	number of high-risks	X003WI	X005JO	X010BJ	X011DA
A.218324_s_at	A.218325_s_at	SPATS2	DIDO1	0.198711039	0.023775912	6.924979	8.173964	4.2	0.007038052	0.026392694	0.016733299	60	99	1	2	2	1
A.218324_s_at	A.218326_s_at	SPATS2	LGR4	0.198711039	0.165736915	6.924979	5.103521	5.2	0.029875863	0.058452776	0.052447938	135	24	1	1	2	1
A.218324_s_at	A.218327_s_at	SPATS2	SNAP29	0.198711039	0.003643943	6.924979	8.460486	2.2	0.002332733	0.019393507	0.193568626	102	57	2	1	2	2
A.218324_s_at	A.218328_at	SPATS2	COQ4	0.198711039	0.166711531	6.924979	7.982375	4.2	0.040003847	0.072006925	0.903528074	42	117	2	2	2	1
A.218324_s_at	A.218329_at	SPATS2	PRDM4	0.198711039	0.1738874	6.924979	7.891556	3.1	0.115614566	0.136911986	0.094190903	137	22	1	1	1	1
A.218324_s_at	A.218330_s_at	SPATS2	NAV2	0.198711039	0.012093942	6.924979	6.352476	4.2	0.02114068	0.04694419	0.035518426	97	62	1	1	2	1
A.218324_s_at	A.218332_at	SPATS2	BEX1	0.198711039	0.008327271	6.924979	5.696361	4.2	0.022930697	0.04694419	0.04967913	113	46	1	1	2	1
A.218324_s_at	A.218333_at	SPATS2	DERL2	0.198711039	0.189713928	6.924979	7.882752	5.2	0.169252402	0.181341859	0.024423704	133	26	1	1	1	1
A.218324_s_at	A.218334_at	SPATS2	THOC7	0.198711039	0.027615925	6.924979	8.405132	5.2	0.067776717	0.094017507	0.048554228	134	25	1	1	2	1
A.218325_s_at	A.218326_s_at	DIDO1	LGR4	0.023775912	0.165736915	8.173964	5.103521	3.2	0.047876438	0.076966525	0.236081552	23	136	2	2	2	2
A.218325_s_at	A.218327_s_at	DIDO1	SNAP29	0.023775912	0.003643943	8.173964	8.460486	3.2	0.008279261	0.028658981	0.620380585	61	98	2	2	2	2
A.218325_s_at	A.218328_at	DIDO1	COQ4	0.023775912	0.166711531	8.173964	7.982375	2.1	0.003016768	0.019393507	0.69270115	106	53	1	2	2	1



The text file output is largely similar to the 1D case above, except it now refers to 2D values as well. The column "design" is explained by the figure and description below and is represented in the graphical plot.

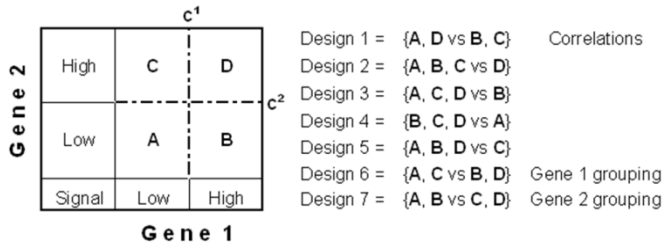


Figure 1. Schematic representation of 2-dimensional grouping (genes 1 and 2 with respective cut-offs c^1 and c^2).

1.1 = Design 1 with A,D containing low-risk patients and B,C containing high-risk patients
1.2 = Design 1 with A,D containing high-risk patients and B,C containing low-risk patients

2.1 = Design 2 with D containing low-risk patients and A,B,C containing high-risk patients
2.2 = Design 2 with D containing high-risk patients and A,B,C containing low-risk patients

3.1 = Design 3 with B containing high-risk patients and A,C,D containing low-risk patients
3.2 = Design 3 with B containing low-risk patients and A,C,D containing high-risk patients

4.1 = Design 4 with A containing high-risk patients and B,C,D containing low-risk patients
4.2 = Design 4 with A containing low-risk patients and B,C,D containing high-risk patients

5.1 = Design 5 with C containing high-risk patients and A,B,D containing low-risk patients
5.2 = Design 5 with C containing low-risk patients and A,B,D containing high-risk patients

Graphs:

Left: patients grouping using 2D cut-offs (black horizontal and vertical lines). Black dots = low-risk patients; Red dots = high-risk patients. Right: Kaplan-Meier survival curves. Black lines = low-risk group; Red lines = high-risk group. Dashed lines refer to gene on the x-axis; Dotted lines refer gene on the y-axis.

References

Cox, D.R. 1972. Regression models and life-tables. Journal of the Royal Statistical Society B 34, 187-220.

Motakis, E., Ivshina, A.V. and Kuznetsov, V.A. 2009. Data-driven approach to predict survival of cancer patients. IEEE Engineering in Medicine and Biology 28, 58-66.

Motakis, E., Low, D.H.P. and Kuznetsov, V.A. 2011. Gene selection of breast cancer survival biomarkers via data-driven patients grouping and network analysis: a human genome study (under submission).