



GPProcessor 2.0a

A Free Software for Processing Microarray GenePix GPR Files

User's Manual

By Zhong Guan, Ph. D.

Hongyu Zhao's Lab of Statistical Genetics

Yale University School of Medicine

Check for update at

<http://bioinformatics.med.yale.edu/softwarelist.html>

Microarray is one of the most fascinating technologies in the millennium. It is a novel invention to exploit DNA sequence data and produce useful information of gene expression levels for the entire genomes. Besides its enormous scientific potential to help people understand gene regulation and interactions, microarrays have promising applications in pharmaceutical and clinical research. By comparing expression levels of genes in normal and disease cells, microarrays can be used to identify disease genes and targets for therapeutic drugs.

It is commonly recognized that single microarray is subject to variations. Replicated microarray data may provide us with more reliable information about gene expression levels. **GenePixTM Pro** of **Axon Instruments, Inc.** is one of the most widely-used DNA gene expression microarray image analysis software. **GPProcessor** aims at pooling and processing replicated **GenePixTM Pro** output GPR (GenePix Results format-*.gpr) files with the analysis of variance (ANOVA) and T-test methods. The users can also choose to calculate the adjusted p -values and q -values, the false discovery rate (FDR) version of p -values. Furthermore, **GPProcessor** also generates

output files which can be used as input of other popular microarray data analysis software, such as University. [Cluster of Eisen's Lab](#) and [SAM](#) of Tibshirani's Lab at Stanford University.

1 Features and Usage

1.1 Data files input:

- (1) Save *GProcessor.exe* in the folder containing your gpr files. Double click on the icon of the program. *GProcessor* will automatically find the gpr files (see Figure 1).
- (2) Enter the number of replicated spots of each gene in each slide. In some arrays, there may be equal number of consecutive spots represent the same gene. For example, in the following portion of a gpr file (see Figure 2), each gene is spotted consecutively twice. Therefore the number of replicate spots in this array is 2.
- (3) Choosing a **ratio formulae**:
In **GenePix** gpr files, there are 5 different forms of ratio calculations for user to select, *i.e.*, *Ratio of Medians*, *Ratio of Means*, *Median of Ratios*, *Mean of Ratios* and *Regression Ratio*. In order to perform dye effect adjustments (Lowess normalization) or ANOVA, you have to choose either *Ratio of Medians* or *Ratio of Means*.
- (4) Once the gpr files to be processed are selected in the list box "Selected GPR Files", just push the "Load Files" (Alt+L) button. *GProcessor* accept all all version GPR files. For GenePix 4.x gpr files which may contain up to four colors, *GProcessor* will automatically detect the number of colors and the ratio formulations.

1.2 Changing Ratio formula:

In some microarray experiments, dye exchange was conducted. In this case you can define the *positive* dye configuration, then the dye configuration after exchange can be called *negative*. For those slides with *negative* dye configuration, a user can click *Option* button of **GenePixTM Pro** to change the ratio formula. For a *positive* slides select W_1/W_2 ; or select W_2/W_1 for a *positive*

Figure 1: *GP*rocessor interface

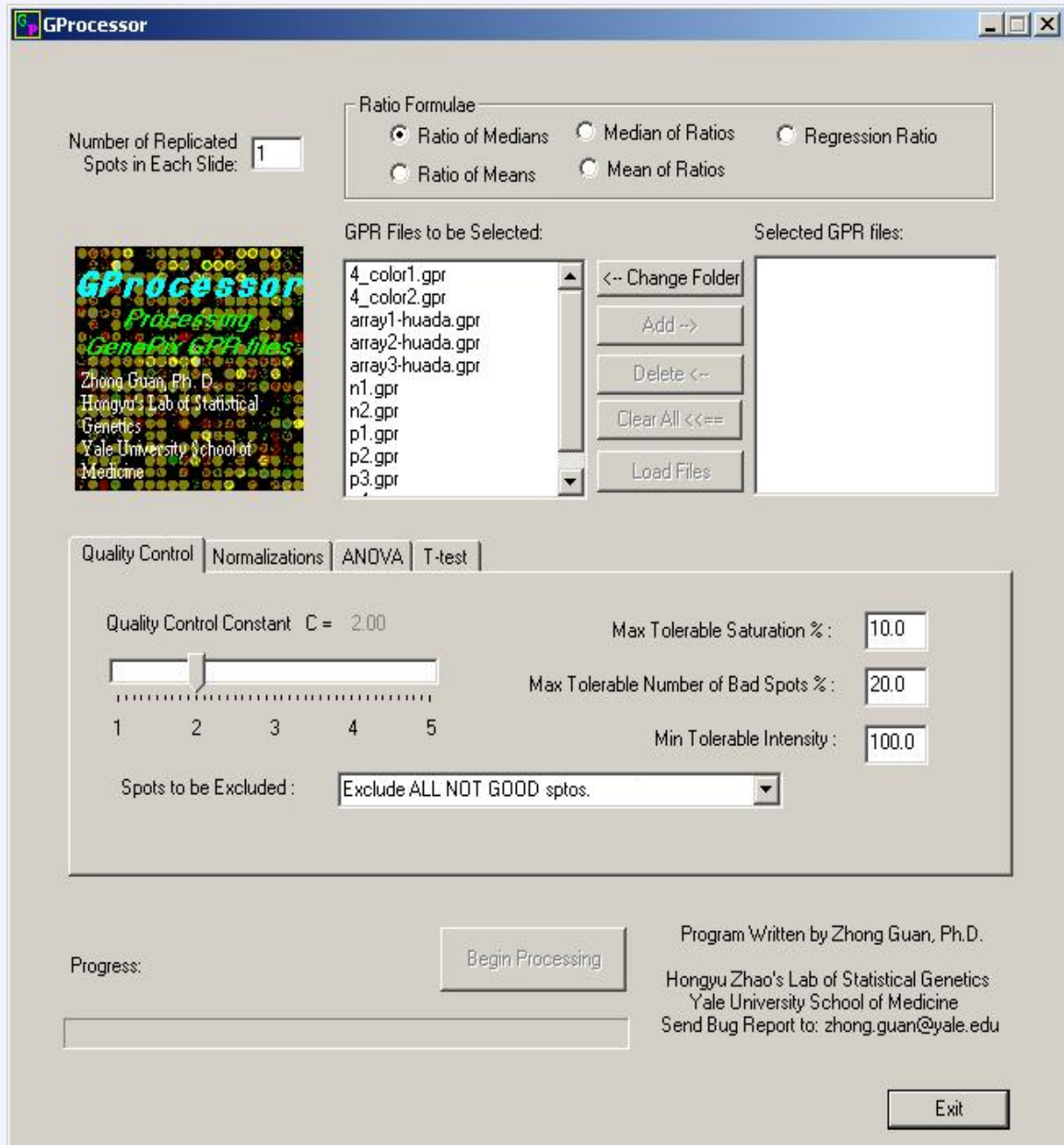


Figure 2: Portion of a Typical GPR file with Number of Replicated Spots Equal to 2.

	A	B	C	D	E	F	G	H	I	J
28	ScanRegion=276,1172,2104,5116									
29	Supplier=									
30	Block	Column	Row	Name	ID	X	Y	Dia.	F635 Medi	F635 Mear
31	1	1	1	ZF-12	Mm.11663	3810	12420	90	7811	7575
32	1	2	1	ZF-12	Mm.11663	3990	12390	90	13362	14203
33	1	3	1	Nmbr	Mm.57042	4160	12400	120	457	490
34	1	4	1	Nmbr	Mm.57042	4340	12400	120	472	494
35	1	5	1	Gjb2	Mm.34118	4520	12400	120	621	642
36	1	6	1	Gjb2	Mm.34118	4700	12400	120	702	765
37	1	7	1	Hoxd9	Mm.24420	4880	12390	80	3007	3160
38	1	8	1	Hoxd9	Mm.24420	5060	12380	70	2954	2957
39	1	9	1	Tif1b	Mm.15701	5240	12420	100	11976	11129
40	1	10	1	Tif1b	Mm.15701	5420	12400	90	13561	13889
41	1	11	1	Col5a1	Mm.7281	5590	12400	80	1479	1504
42	1	12	1	Col5a1	Mm.7281	5790	12380	80	1313	1327
43	1	13	1	Xrcc1	Mm.4347	5950	12390	120	845	842
44	1	14	1	Xrcc1	Mm.4347	6110	12400	150	868	879

array. After changing the formula, click *Analysis* button of **GenePix™ Pro 3.0** to extract data from these *negative* slides, the data should be exported as *.gpr* files. The *.gpr* files obtained in this way from *negative* slides, together with those *.gpr* files from *positive* slides, can be used to **GPProcessor** for pooling purposes. If you do not have the original GPS (GenePix Settings format-*.gps) file or you do not want to modify the GPR files, you may have to manually input the *RatioFormulations*. You can do so in dialogs when prompted.

1.3 Quality Control and Outlier Detection(see Figure 3):

- (1) **Quality Control Constant**: For a given spot, if the ratio of intensity over background is less than constant C (range from 1.0 to 5.0) for all the channels, then this spot will also not be accounted for by **GPProcessor**.
- (2) **Max Tolerable Saturation (%)**: The saturations are given in the columns such as "F635 % Sat. " in gpr files. High percentage of saturation will

Figure 3: Parameters Input for Quality Control

Quality Control | Normalizations | ANOVA | T-test

Quality Control Constant C = 2.00

Max Tolerable Saturation % : 10.0

Max Tolerable Number of Bad Spots % : 20.0

Min Tolerable Intensity : 100.0

Spots to be Excluded : Exclude ALL NOT GOOD spots.

mask the true expression level. Input the highest tolerable percentage of saturation. Default is 10%.

- (3) **Max Tolerable Number of Bad Spots (%)** : In some experiments, you may have many, say more than 10, replicated microarrays, and you may want to keep those genes with small proportion, say 20%, of bad spots in all these arrays. Input the maximum tolerable proportion of Bad Spots. Default is 20%.
- (4) **Min Tolerable Intensity** : Very weak feature intensity(after background subtraction) may also mess up the analysis results. Input the lowest tolerable intensity value. Default is 100.
- (5) **Spots to be Excluded**: Finally, select what kinds of spots you want to exclude in the analysis. In gpr files, the flags 100, -100, -75 -50 and 0 indicate *Good*, *Bad*, *Absent*, *Not Found* and un-flagged features (spots) respectively. Any un-flagged spot which did not pass the above quality filters 1, 2 and 4 will be flagged -25. Spot with negative flag is called "Not Good Spot".
- (6) A very simple outlier detecting algorithm was incorporated in T-test of **GPProcessor**. Those spots which leads the large difference between the mean of ratios and the median of ratios are defined as outliers and

Figure 4: Parameters Input for Normalization

The image shows a software interface with four tabs: 'Quality Control', 'Normalizations', 'ANOVA', and 'T-test'. The 'Normalizations' tab is active. Inside the main window, there is a checked checkbox labeled 'Dye Effect Adjustment (Lowess Normalization)'. Below this, there are three input fields: 'f = 0.66', 'nsteps = 3', and 'Delta = 0.01 of Data Range'.

removed from the calculation. The number of outliers for each gene is listed in the output file *T-test.txt* for T-test results.

1.4 Lowess fit Normalization(see Figure 4)

If you have chosen one of the first two ratio types, texti.e. *Ratio of Medians*, *Ratio of Means*, you have the option to normalize the channel intensities and the ratios by using the Lowess fit method. This procedure works for gpr files with any number of colors. You may also want to change some parameters used for Lowess Fit Normalization:

- (1) f is used to specify the amount of smoothing; f is the fraction of points used to compute each fitted value; the larger the value of f , the smoother the fitted values become; choosing f between 0.2 and 0.8 should serve most purposes. If you have no idea which value to use, try $f = 0.5$.
- (2) $nsteps$ is the number of iterations in the robust fit; if $nsteps = 0$, the non-robust fit is returned; setting $nsteps$ equal to 2 should serve most purposes. Since microarray data usually has very huge size, the default value for $nsteps$ is set to be 3.

Figure 5: Parameters Input for ANOVA

- (3) δ is a nonnegative parameter which may be used to save computations; if data size is less than 100, set δ equal to 0.0; Otherwise the default δ value is set to be 1%=0.01 of the range of the abscissa points on the scatter plot.

1.5 ANOVA Procedure(see Figure 5)

Again if you choose either *Ratio of Medians* or *Ratio of Means*, the ANOVA procedure is another option for the you. Assuming that there is no dye effect and the dye effect has been adjusted by performing the Lowess normalization. You can try and select any one of the following models:

$$\text{Model 1: } \log(y_{ikg}) = \mu + A_i + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \epsilon_{ikg};$$

$$\text{Model 2: } \log(y_{ikg}) = \mu + A_i + V_k + G_g + (VG)_{kg} + \epsilon_{ikg};$$

where y is the background-subtracted intensity, μ is the overall mean, A_i is the effect of i -th array, V_k the effect of k -th variety (treatment), G_g the effect of g -th gene, $(AG)_{ig}, (VG)_{kg}$ are corresponding interactions, and ϵ_{ikg} is the random error.

If there are dye effects, Lowess Fit Normalization should be performed in order to do ANOVA. For the selected model, the ANOVA results will be

saved in a file *AnovaResult.txt*. This file contains the ANOVA table which gives the F -ratios and the corresponding P-values together with the multiple R^2 and the estimated ratio fold changes together with the P-values. If you have checked the corresponding options, this file also give you the Adjusted p-values, FDR's and the q-values. The FDR for each gene is the FDR calculated for all the hypotheses if we use the p-value(not adjusted) of this gene as the significance level.

1.6 T-test:

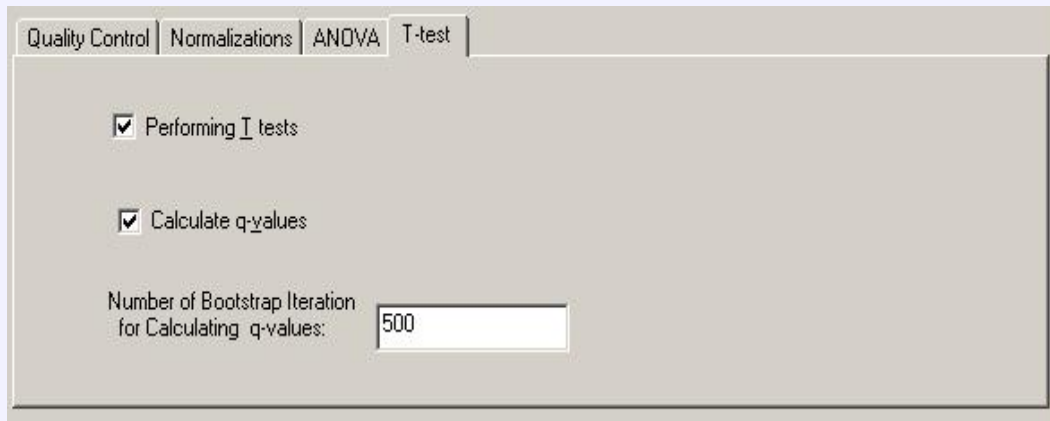
If you choose to perform T-test, you have the option to calculate the q-values of the multiple hypotheses test (see Figure 6). **GPProcessor** calculates the *Mean of Ratios*, *Median of Ratios*, *the Mean of log2 ratio* for each gene using all good replicated spots from different arrays. The values of the mean of ratios, the mean of log2-ratios and the median of ratios are given in the output file *T-test.txt*. CV's (coefficient of variation defined as the standard deviation divided by the mean) for all useful replicated spots was calculated and listed as a column in this output file. Users can use this information to check the quality of their spots or even their slides. For each gene, the *P-value* of *T-statistic* is calculated based on all the good replicated spots. This P-value serves only as a reference of significance level of differential expression. In *T-test.txt* you can also find the Gene Number, Gene Name, ID, the number of spots actually used and the number of outliers.

1.7 Output Files

Except the output files *AnovaResult.txt* and *T-test.txt* mentioned above. There are several other output files produced by **GPProcessor**. They are *Cluster.txt*, *CorrCoef.xls*, *MergeData.txt*, *SamData1Class.txt* and *SamData-Paired.txt*

- (1) **Merge of Normalized Data** The ratios of the selected type for all the genes are given in the file *MergeData.txt*. If the ratio type is *Ratio of Medians* or *Ratio of Means*, then the ratios are normalized by Lowess fit method, otherwise, the ratios are normalized by the simple normalization method using the **GenePix** normalization factor according to the selected ratio.

Figure 6: Parameters Input for T-test



Quality Control | Normalizations | ANOVA | T-test

Performing T tests

Calculate q-values

Number of Bootstrap Iteration for Calculating q-values:

(2) **Correlation Coefficients:**

In output file *CorrCoef.xls*, different kinds of correlation coefficients are summarized in table formats. These include the correlation coefficients both for intensity and for ratio measurements across different data sets or across different slides. Users can use these information to examine and compare their data.

(3) **Input Data Files for SAM and Cluster Analysis**

Cluster.txt is the input file in the format of cluster analysis software [Cluster of Eisen's Lab](#). *SamData1Class.txt* and *SamDataPaired.txt* are One-Class and Paired input files for [SAM](#) of Tibshirani's Lab at Stanford University.

All these output data files will be saved in the same folder which contains the processed gpr files. Although the program will prompt you to rename or move the existing output data files which you do not want to be overwritten, to speed up the processing, it is recommended that you do this before running [GPRocessor](#).

1.8 Different version GPR files

We noticed that the header format of a GPR file of **GenePix™ Pro 3.0.0.x-3.0.5.x** is different from that of **GenePix™ Pro 3.0.6.x**. This version (2.0a) of *GPProcessor* can automatically detect the different versions of GPR files including the latest version 4.x containing up to 4 colors and treat them differently. So the users can use all the GPR files available as input of *GPProcessor*. Of course, the user is responsible for making sure that all the GPR files of different versions use the same gene list.

2 Remark

- Always use **original** GPR files for *GPProcessor*. Although the current version (2.0a) also works for GPR file which have been opened and re-saved by programs like **MS Excel**, we recommend not to open and re-save the GPR file with **MS Excel** before using them for *GPProcessor*, since **Microsoft® Excel** will change the format of GPR file.

3 Release and Version

The current version is *GPProcessor* 2.0a. *GPProcessor* was designed for the convenience of **YMD**(Yale Microarray Database) users to process replicated microarray data sets obtained from **GenePix™ Pro** image analysis software. No warranty is expressed or implied.

The *GPProcessor* 2.0a was released as a package in zip format which can be downloaded from <http://bioinformatics.med.yale.edu/softwarelist.html>. It is user's responsibility to check this web site for update. The package includes a Windows application *GProcessor.exe*, a *User's Manual* in PDF format, two replicated data sets *p1.gpr* and *p2.gpr*. User can use these two files to test the program, in each data set there are 1 replicated spots.

GPProcessor 2.0a was tested for original GPR files generated by **GenePix™ Pro** of all versions. If you have GPR files which cannot be zhong.guan@yale.edu and attach your GPR files for the test purpose.

4 Contacts

Please send bugs report, comments, suggestions and critics to zhong.guan@yale.edu, hongyu.zhao@yale.edu, kenneth.williams@yale.edu and janet.hager@yale.edu.

References

1. Yang, Y. H. *et al.*, Normalization for cDNA Microarray Data SPIE BiOS 2001, San Jose, California, January 2001.
2. Dudoit, S. *et al.*, Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments; Technical Report #578, Aug. 2000, Dept. of Statistics, UC Berkeley
3. Tseng, G. C. *et al.*, Issues in cDNA Microarray Analysis: Quality Filtering, Channel Normalization, Models of Variation and Assessment of Gene Effects. *Neucleic Acids Research*, 2001, Vol.29, No 12, 2549-1557.
4. Kerr, M. K. *et al.*, Analysis of Variance for Gene Expression Microarray Data, *J. Comput. Biol.*, 7, 819-837.

Last Update

February 3, 2003.