Notes for Microarray Processing Macro

Written by Lin Liu, Boons Group, CCRC, UGA

liulin.cn@gmail.com

This is not a step-by-step manual. This is a combination of notes regarding various aspects of the macro, which should be useful from time to time during every day uses.

We have been using this macro routinely for more than 5 years since 2016 and it is working quite nicely. The data obtained using this macro has been published in journals such as Nature Chemistry, JACS, ACS Central Science, and PNAS.

History:

The macro is the result of my efforts to automate the processing of our microarray data in Boons' group, as a chemist and an amateur programmer. It started by writing Margreet an automated worksheet that could turn a table 90 degrees. Then I decided to fully automate the whole process by writing a proper macro in VBA. The major part was written during a timeframe between late 2015 and early 2016, and it is still being constantly updated.

Developing Environment:

It was written using VBA and tested under **Microsoft Excel 2013/2016**. Most functions will work on Mac, with a few minor exceptions. It is not compatible with Microsoft Excel 2010.

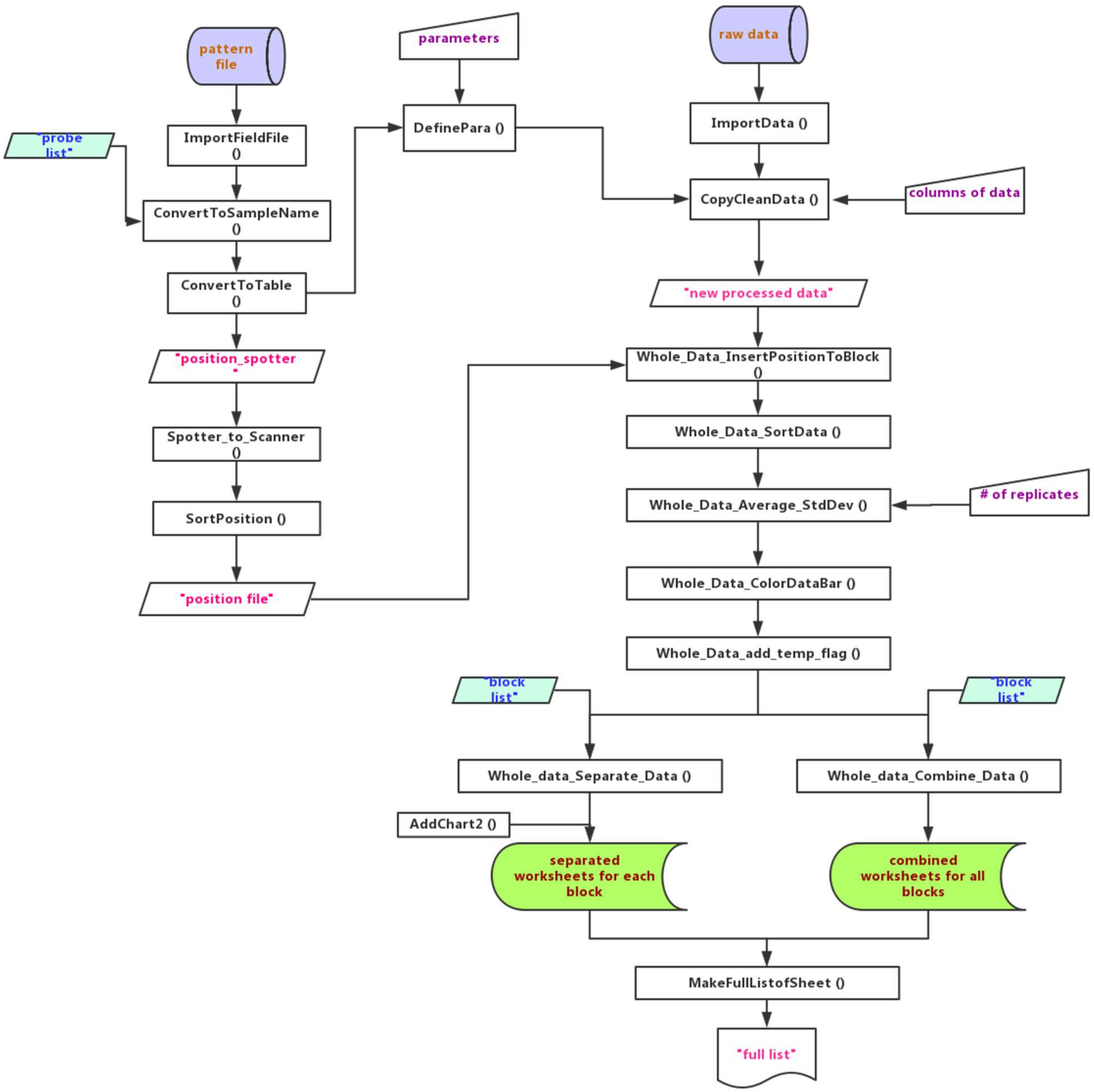
Designed Applications:

It was developed to be used with a Scienion S3 printer and a GenePix 4000B scanner. However, other scanner data could be used after minor modifications. GAL file-based data files can also be incorporated. Please email me for a new version of the program to process GAL files.

Structure of the Macro:

The macro is divided into two parts. Part 1, on worksheet “convert sample”, handles the transformation of the pattern file from printer (.fld file) to a format that can be used to process the data from scanner. Part 2, controlled from worksheet “parameter”, will import the raw data and process it. The flowchart of the macro is here:

Flow Chart for Microarray Processing Macro



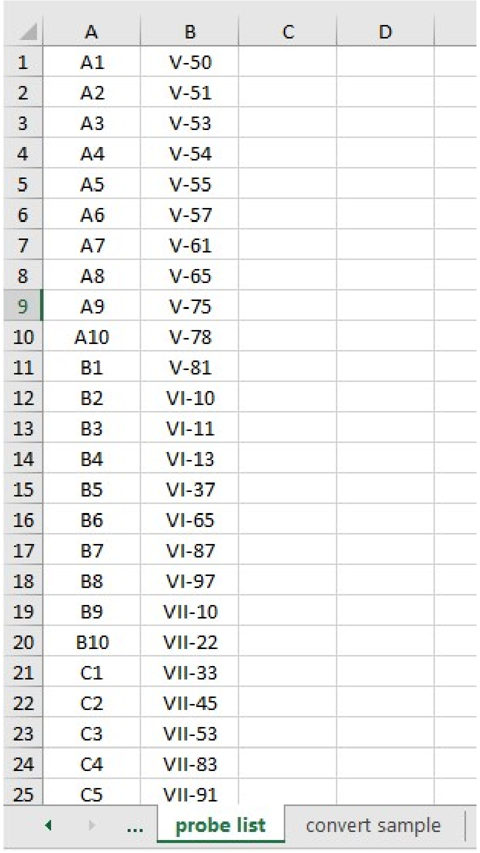
Name for the worksheets:

All the worksheets' names are case-sensitive. Some of the worksheets could have been named in a better way to reflect their true function, however, the original names have been buried deep down in the legacy codes, so do not change them manually.

Probe list:

a worksheet named "**probe list**", to correlate the well number used on the printer with actual compound names.

An example of probe list looks like this:



	A	B	C	D
1	A1	V-50		
2	A2	V-51		
3	A3	V-53		
4	A4	V-54		
5	A5	V-55		
6	A6	V-57		
7	A7	V-61		
8	A8	V-65		
9	A9	V-75		
10	A10	V-78		
11	B1	V-81		
12	B2	VI-10		
13	B3	VI-11		
14	B4	VI-13		
15	B5	VI-37		
16	B6	VI-65		
17	B7	VI-87		
18	B8	VI-97		
19	B9	VII-10		
20	B10	VII-22		
21	C1	VII-33		
22	C2	VII-45		
23	C3	VII-53		
24	C4	VII-83		
25	C5	VII-91		

Column A: Well number

Column B: compound names

Do not put any information in other columns. Anything in Columns C and D will be overwritten.

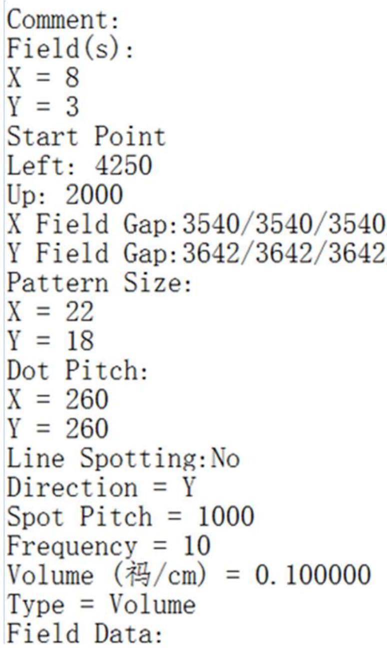
(If the program cannot find a match of a certain probe name defined in the .fld file in this list, it will be ignored)

Field file:

Pattern file from the printer (.fld file)

1. Make sure you have a saved pattern file for each printing.
2. Under no circumstances should you manually edit the .fld file.
3. Only **the first block** in the field file will be read and transformed. This implies all the blocks on the slide have the same pattern.
4. ~~The **plate information from the printer is removed** before further processing. This means that if you are using multiple plates, you should not have samples in the same wells on different plates. If later you have decided that we have more than 96 or 384 compounds and need to use the same well numbers on different plates, I will modify the program to accommodate it.~~
   Now the macro will accommodate multiple plates. Mark the "Multiplate(Y/N)" in worksheet "convert sample" as "Y", then use 1A1, 1A2, …., 2A1, 2A2….., etc in the probe list.

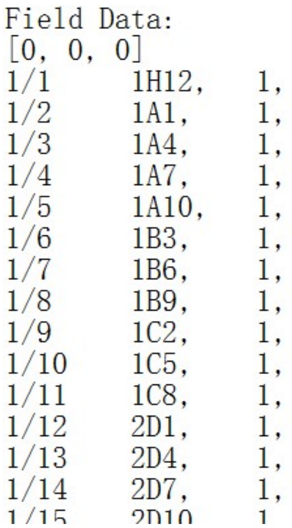
Here is a typical file header of a field file:



```
Comment:
Field(s):
X = 8
Y = 3
Start Point
Left: 4250
Up: 2000
X Field Gap:3540/3540/3540/3540/3540/3540/3540/3540/
Y Field Gap:3642/3642/3642/
Pattern Size:
X = 22
Y = 18
Dot Pitch:
X = 260
Y = 260
Line Spotting:No
Direction = Y
Spot Pitch = 1000
Frequency = 10
Volume (祃/cm) = 0. 100000
Type = Volume
Field Data:
```

The macro will read the X, Y for fields, and the X, Y for pattern. The values will be transferred to “Parameter”.

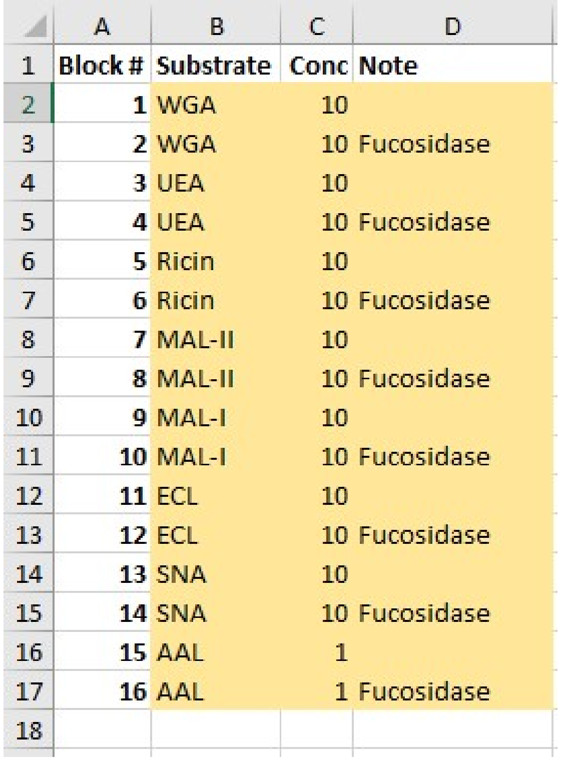
Here is a part of the block pattern data from the field file.



```
Field Data:
[0, 0, 0]
1/1      1H12,    1,
1/2      1A1,     1,
1/3      1A4,     1,
1/4      1A7,     1,
1/5      1A10,    1,
1/6      1B3,     1,
1/7      1B6,     1,
1/8      1B9,     1,
1/9      1C2,     1,
1/10     1C5,     1,
1/11     1C8,     1,
1/12     2D1,     1,
1/13     2D4,     1,
1/14     2D7,     1,
1/15     2D10     1
```

~~The plate information, like in 1H12, 2D1, will be removed. They will be H12 and D1.~~ As stated above, now the plate information could be preserved.

Block list: a worksheet named "block list" for each screening.



	A	B	C	D
1	Block #	Substrate	Conc	Note
2	1	WGA	10	
3	2	WGA	10	Fucosidase
4	3	UEA	10	
5	4	UEA	10	Fucosidase
6	5	Ricin	10	
7	6	Ricin	10	Fucosidase
8	7	MAL-II	10	
9	8	MAL-II	10	Fucosidase
10	9	MAL-I	10	
11	10	MAL-I	10	Fucosidase
12	11	ECL	10	
13	12	ECL	10	Fucosidase
14	13	SNA	10	
15	14	SNA	10	Fucosidase
16	15	AAL	1	
17	16	AAL	1	Fucosidase
18				

The final title for each block will be B+C+D+block #.

Position spotter:

The worksheet "**position_spotter**" is the output of the "**convert sample**" using "**probe list**" and the **filed file**. This is the worksheets that the data processing part will rely on.

You only need to do the "**convert sample**" part one time for each printing, then save the file as a template for all screening using the slides from this printing.

Any samples removed from worksheet "**position_spotter**" will not be processed. If you delete certain compounds from this worksheet, the compounds won't show up in the final results.

Make sure you have the correct **number of replicates.** It can be checked using “sample info” function.

Position file:

The program could also start using a "**position file**" worksheet, instead of using the “**position_spotter**” worksheet. In this case, remove the **Application.Run "Spotter_to_Scanner"** line in **Sub Whole_Data_Process_Whole()**. This is the case in the special version I prepared for Margreet to be used with the old heparan sulfate array.

Raw data:

Data requirements: The header ends at row 32. The data, including the column titles, starts at row 33.

Format requirements: A: Flags D: Block E: Column F: Row. I might go back to the codes later when time permits to make changes so it won’t require this.

On GenePix 4000B, usually there will be no need to play with the data format as long as a few spots have been flagged. The exported data file will have the correct layout.

On other scanners, it is necessary to pay attention to the data structure and make appropriate changes.

Added: Column for Flag has to have data. It cannot be empty. All 0 is fine.

Requirements for the raw data

The raw data must have **the same layout as indicated in the spotter layout**. If the data was obtained in a format different than the printer, say acquired data in 18*20, but the position has a 18*22 layout, the data processing will give wrong results. Currently there is no checking for this mismatch.

The position-spotter could be edited manually to fit different needs, if less amount of spots were used and only a fraction of spots were scanned.

Which data to use to represent the strength: Total, Mean, or Median?

In the macro, the **total intensity of the spot** is the preferred data to represent the strength of the signal.

If you would like to use the **Mean or Median as the data**, it could be done easily by entering another letter for the desired data column in "**parameter**". In this case, the background removal process has to be changed. Contact me for details. However, **choosing Mean or Median is not recommended.** This would require a different way for the scanner to pick up the spot.

CFG chooses Median as the data. **We have chosen to go with the total intensity** for several reasons. I have more information in my discussion regarding this issue on my lab notebook.

I believe our method suits our needs and is accurate, provides higher quality data.

Background removal:

By default, the macro will remove background from each spot using the **median background level**. Refer to the scanner manual for how the scanner would determine background.

New data(bkg removed) = old data – bkg mean * spot area.

This process is done in macro Copy_Clean_Data (). If the new data (after removal of background) <0, it is assigned as 1.

There is another option called remove blank (). The value from a sample named 'blank' is used and subtracted from data. Do not use this one unless there is a specific reason.

Some journals, including Nature series journals, are requesting the authors to show the individual data point on the graph. It is easier to leave the negative data as it is in this case. Email me for a newer version of the program for this.

Data averaging method:

Currently we are using the method CFG employs. The highest and lowest value for a probe would be dropped, and the rest of data would be averaged. If you have only 2 replicates, although that's not a recommended pattern, the code would still handle them by simply averaging the two without dropping any value.

Manual removal of bad data:

It could be done in the two step processing. Go through the "**new processed data**" after the **first step processing**, delete or mask any data which is apparently erroneous, then proceed to **second step processing**.

Any data flagged bad in the scanner (Genepix 4000B) will be shown in red in "**new processed data**". The code of flag for bad is **-100**. Other flags are not handled right now.

VBA codes:

I put a lot of comments in the codes. Feel free to read through the code with the help of the comments.

Technical details regarding the algorithm and function:

I have more detailed information in my notes. Email me if need anything.

Error Messages:

If you get a pop-up window saying " Worksheet xxxx already exist", manually delete that worksheet. I program it not to overwrite any existing worksheets.

Some other error messages are from the slight differences in different version of VBA. Please use Excel 2013/2016 if possible.

For other messages, click on Debug, and look into the codes. Excel's error message are painfully unhelpful, but look at the line where the error occurs. Most if not all would be very minor issues that could be fixed easily.

Mac/PC compatibility:

I have spent a lot of time to make this macro Mac compatible. The following is a list of compatibility issues.

1. The Mac version of VBA doesn't support GetOpenFile (filefilter).
2. The Mac version cannot handle the AddChart2 function correctly. Use AddChart instead.
3. The import file process in Mac wouldn't handle :

   .PreserveFormarting,
   .RefreshPeriod,
   .TextFilePlatform,
   .TextFileTrailingMinusNumbers.

On the above cases, a script to determine if the running environment is PC or Mac is used. If it is running on a Mac, appropriate codes would be selected to run.

4. The Mac version of VBA doesn't support **dictionary class**, since it doesn't have "Microsoft Scripting Runtime". So the **"Sample Info", "Color Sample Table", and "Color Scanner"** button will not run on Mac. A workaround would be create your own dictionary class with Collection, and import the .cls file to every Mac that has to run this macro. Or just ignore these functions on Mac. See https://sysmod.wordpress.com/2011/11/02/dictionary-class-in-vba-instead-of-scripting-dictionary/

Recent additions:

Ratio function added as requested.

Ability to work with multiple plates were added as needed.

20210615: added a function to clean the combined process data, to provide a worksheet, which can be directly pasted to Prism.