DWD “Distance Weighted Discrimination” and SSP “Single Sample Predictor”

Users guide and technical document

1 Introduction

Distance Weighted Discrimination (DWD) can perform systematic bias adjustment in microarray data (https://genome.unc.edu/pubsup/dwd/). This document will focus on the following areas: Installation and Running of the software.

2 Summary of Changes

The following are changes since the initial release of DWDSSP 1.0.

2.1 Changes in DWDSSP 1.21

- Graphic User Interface was changed
- The actual numbers of genes and samples that are used for DWD adjustment and SSP prediction are displayed.
-

-

-

….

.

3 How to Obtain DWD-SSP

The current version of DWD-SSP is available and can be downloaded at https://genome.unc.edu/pubsup/dwd/

4 System Requirements

- The current version of DWD-SSP has only been tested in Microsoft Windows XP and 2000.
- Java Runtime Environment (JRE) 1.4 or above is required, you could download it from http://java.sun.com/j2se/1.4.2/download.html. It is absolutely necessary to verify that your default JRE is the one you have downloaded, rather than something else. You could start DOS and type “java –version” to verify that.

- At least 256M RAM is required, the more RAM you have, the better the performance of the software

5 How to Install

You need to use a program like WinRAR orWinZip to unzip the DWD.zip file you just downloaded, and extract the contents to your C: drive. After that you will have a fold "DWD" in your C: drive.

The software should be directly installed under c: drive, or else you would have some configuration work to do. There are three subdirectories under the DWD directory: "lib", "bin" and "data". There are also some important files in the DWD directory: README.txt, "RunDWD.bat", "DWD.jar" and other six jar files.

The lib subdirectory contains all libraries and converted c executable.

The data subdirectory can house all data files, including your original ones and output from DWD.

Note: It is needed to put your input files into this folder to avoid any errors.

6 How to Uninstall

Simply delete the "DWD" folder in your C: drive.

7 Where to find this documentation

This manual for DWD-SSP is part of the DWD.zip.

8 Data Formats

The current version of DWD can take Stanford-like text delimited file and MAGE-ML format file as input.

- For Stanford-like text delimited file. The first column of the Stanford-like text delimited file is identification, the second column contains some annotation, and data start from the third column. The gene identification must be unique (duplicates are not allowed right now). The first row contains the information of samples; the second row of the data file has information about the response measurement (target variable); the rest rows store the gene expression data, every gene per row. The target variables are real number (1,2, 3, …) which represent different groups of your data; for example different platform of the data, different subtypes of the disease, different response to a medical treatment, etc. This kind

of data sets has been extensively tested. For DWD adjustment purpose only, the second line is not necessay, but it is abosulutely needed for SSP. Two sample files are also included in the DWDdata subdirectory. The software can also detect the missing elements and non-digit values where a digit value is supposed to be and report each error one by one each time. The first two rows can not have any missing values. Other lines with missing values will not be used and saved into a file named as NotLoadLines in the data folder. Missing values can be imputed. see http://bioinformatics.oupjournals.org/cgi/reprint/17/6/520.pdf for a good introduction to the main ideas of, and methods typically used for, imputation, and see http://www.scripps.edu/researchservices/dna_array/new/Data_Analysis_SAM.htm for imputation software. When you edit the file with MS Excel, please delete any blank lines. We suggest using TextPad editor to modify your files. Here is the website to download TextPad: http://www.textpad.com/download/

- For MAGE-ML format file. We have tested files generated from Agilent and Affymatrix, either with internal data or external data files. At this time, validation with MAGE-ML.dtd does not work. Validation itself works, but it massed up with the reading internal data generated with Agilent software. The work-around is to delete those lines related to MAGE-ML.dtd in the .xml input files. This is very important.

10 How to Run DWD-SSP

- Running of DWD

There are two methods to start to run DWD.
Method 1: Start DOS command line window, change the directory to DWD, type RunDWD.bat as shown in the Fig. 1. Using this method, the error message can be retained after you close DWD.

Fig 1

Method 2: In window explorer, double click RunDWD.bat.

The GUI is split into four parts as shown in the Fig. 2.

The upper left panel displays the results of the input and output files.
The lower left panel displays the results of DWD type and Mean Adjust Type.
The upper right panel is used to enter all parameters.

The lower right panel is displays the running status.

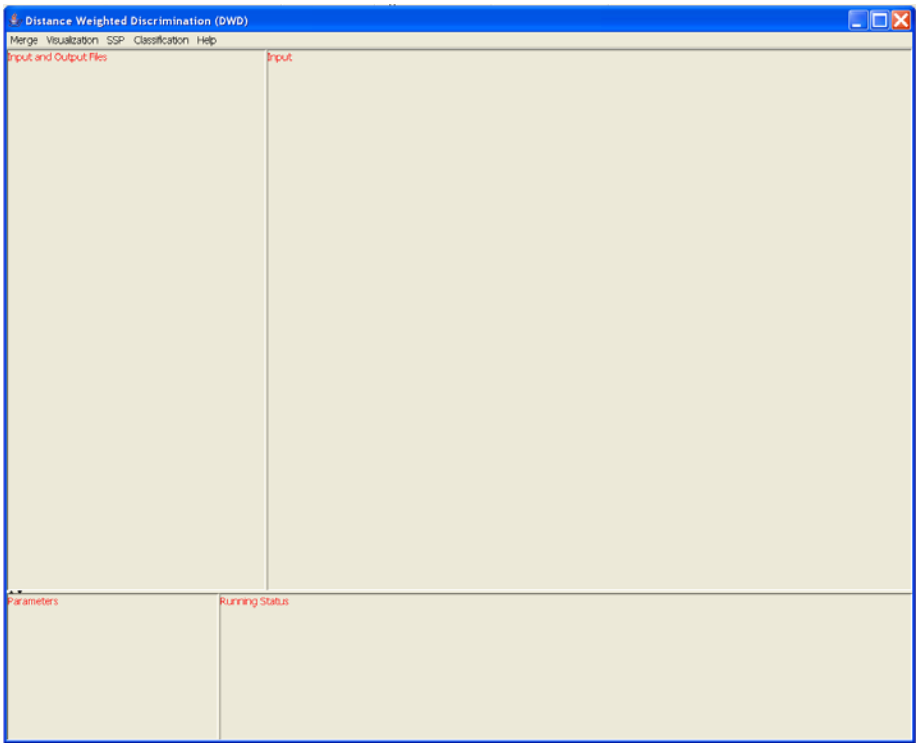


Fig. 2

For DWD adjustment, click “Merge->Stanford Text Files”, the upper right panel displays all requirements you need to fill as shown in the Fig.3.

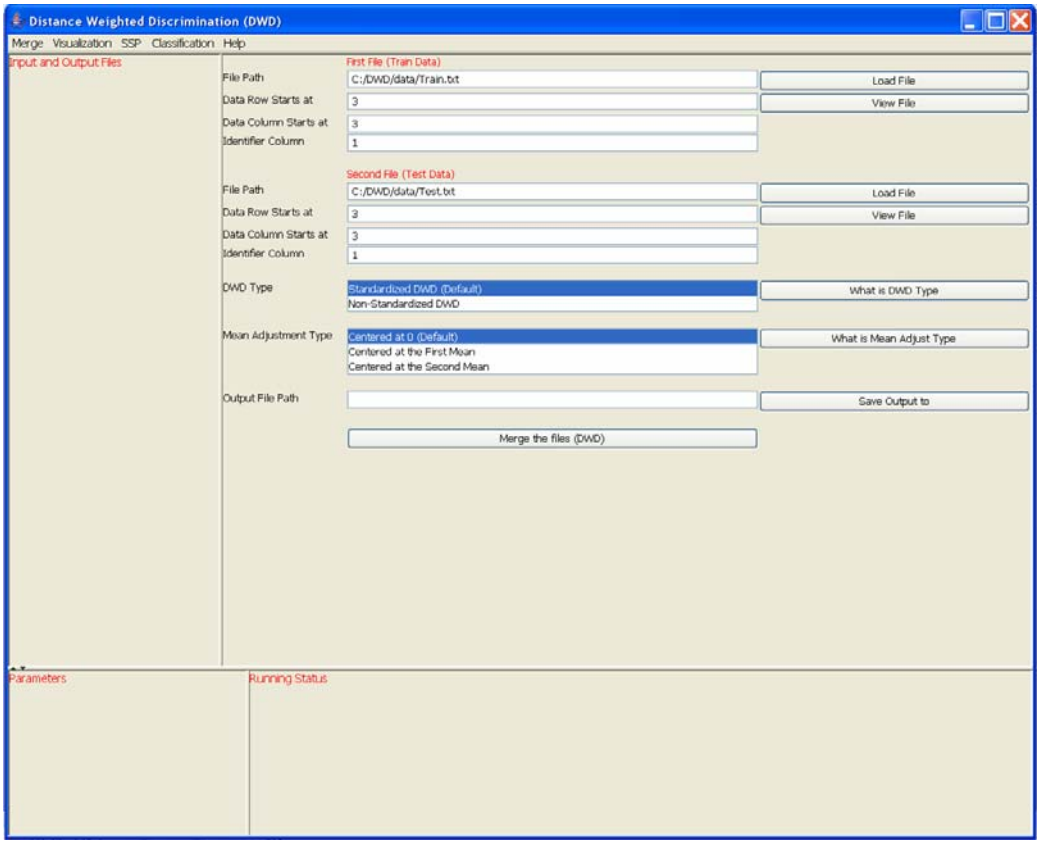
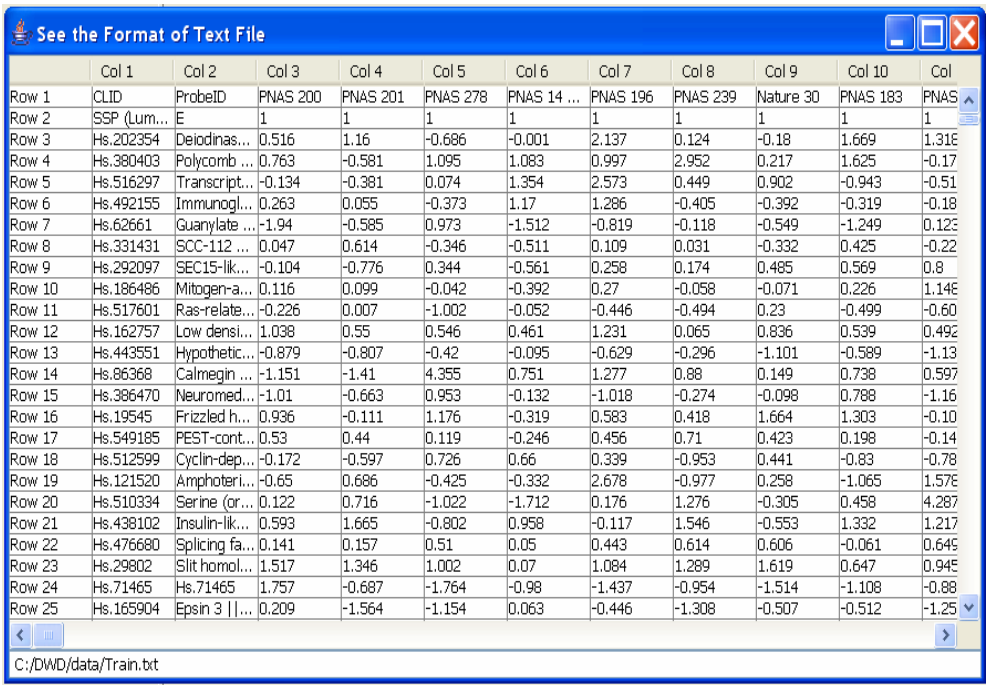


Fig. 3

Load your two files you want to merge, and give a file name for the output files (merge the two files together), select the "DWD Type" and "Mean Adjustment Type", then click "Merge the files (DWD)", and the DWD merge will start. Current version can only merge two files together at a time.

If you do not know the format of the file, you can click the View File button besides the file as shown in the Fig. 4.



	Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9	Col 10	Col
Row 1	CLID	ProbeID	PNAS 200	PNAS 201	PNAS 278	PNAS 14 ...	PNAS 196	PNAS 239	Nature 30	PNAS 183	PNAS
Row 2	SSP (Lum...	E	1	1	1	1	1	1	1	1	1
Row 3	Hs.202354	Deiodinas...	0.516	1.16	-0.686	-0.001	2.137	0.124	-0.18	1.669	1.318
Row 4	Hs.380403	Polycomb ...	0.763	-0.581	1.095	1.083	0.997	2.952	0.217	1.625	-0.17
Row 5	Hs.516297	Transcript...	-0.134	-0.381	0.074	1.354	2.573	0.449	0.902	-0.943	-0.51
Row 6	Hs.492155	Immunogl...	0.263	0.055	-0.373	1.17	1.286	-0.405	-0.392	-0.319	-0.18
Row 7	Hs.62661	Guanylate ...	-1.94	-0.585	0.973	-1.512	-0.819	-0.118	-0.549	-1.249	0.123
Row 8	Hs.331431	SCC-112 ...	0.047	0.614	-0.346	-0.511	0.109	0.031	-0.332	0.425	-0.22
Row 9	Hs.292097	SEC15-lik...	-0.104	-0.776	0.344	-0.561	0.258	0.174	0.485	0.569	0.8
Row 10	Hs.186486	Mitogen-a...	0.116	0.099	-0.042	-0.392	0.27	-0.058	-0.071	0.226	1.148
Row 11	Hs.517601	Ras-relate...	-0.226	0.007	-1.002	-0.052	-0.446	-0.494	0.23	-0.499	-0.60
Row 12	Hs.162757	Low densi...	1.038	0.55	0.546	0.461	1.231	0.065	0.836	0.539	0.492
Row 13	Hs.443551	Hypothetic...	-0.879	-0.807	-0.42	-0.095	-0.629	-0.296	-1.101	-0.589	-1.13
Row 14	Hs.86368	Calmegin ...	-1.151	-1.41	4.355	0.751	1.277	0.88	0.149	0.738	0.597
Row 15	Hs.386470	Neuromed...	-1.01	-0.663	0.953	-0.132	-1.018	-0.274	-0.098	0.788	-1.16
Row 16	Hs.19545	Frizzled h...	0.936	-0.111	1.176	-0.319	0.583	0.418	1.664	1.303	-0.10
Row 17	Hs.549185	PEST-cont...	0.53	0.44	0.119	-0.246	0.456	0.71	0.423	0.198	-0.14
Row 18	Hs.512599	Cyclin-dep...	-0.172	-0.597	0.726	0.66	0.339	-0.953	0.441	-0.83	-0.78
Row 19	Hs.121520	Amphoteri...	-0.65	0.686	-0.425	-0.332	2.678	-0.977	0.258	-1.065	1.578
Row 20	Hs.510334	Serine (or...	0.122	0.716	-1.022	-1.712	0.176	1.276	-0.305	0.458	4.287
Row 21	Hs.438102	Insulin-lik...	0.593	1.665	-0.802	0.958	-0.117	1.546	-0.553	1.332	1.217
Row 22	Hs.476680	Splicing fa...	0.141	0.157	0.51	0.05	0.443	0.614	0.606	-0.061	0.649
Row 23	Hs.29802	Slit homol...	1.517	1.346	1.002	0.07	1.084	1.289	1.619	0.647	0.945
Row 24	Hs.71465	Hs.71465	1.757	-0.687	-1.764	-0.98	-1.437	-0.954	-1.514	-1.108	-0.88
Row 25	Hs.165904	Epsin 3 ||...	0.209	-1.564	-1.154	0.063	-0.446	-1.308	-0.507	-0.512	-1.25

Fig. 4.

You also can find more details about DWD Type and Mean Adjustment Type by clicking the corresponding buttons as shown in the Fig. 5 and Fig. 6.

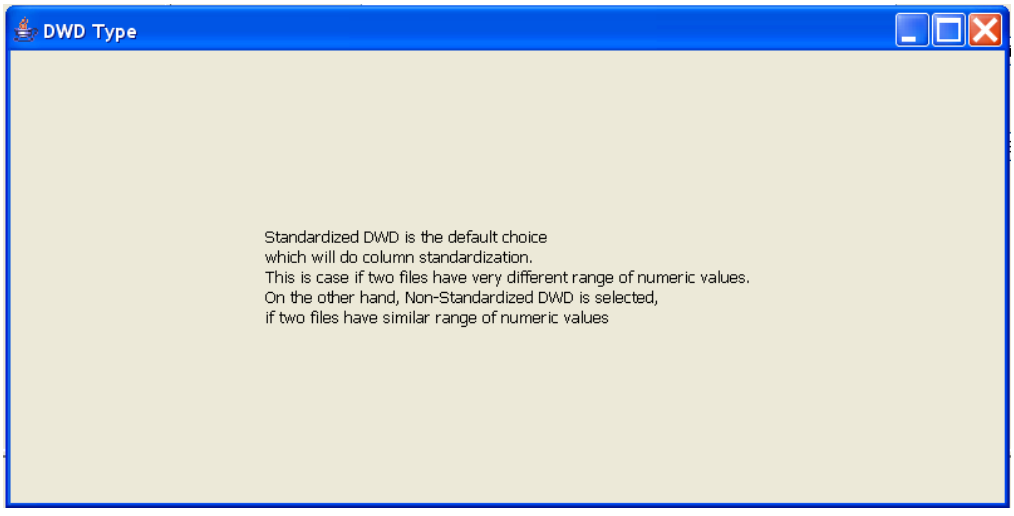


Fig. 5

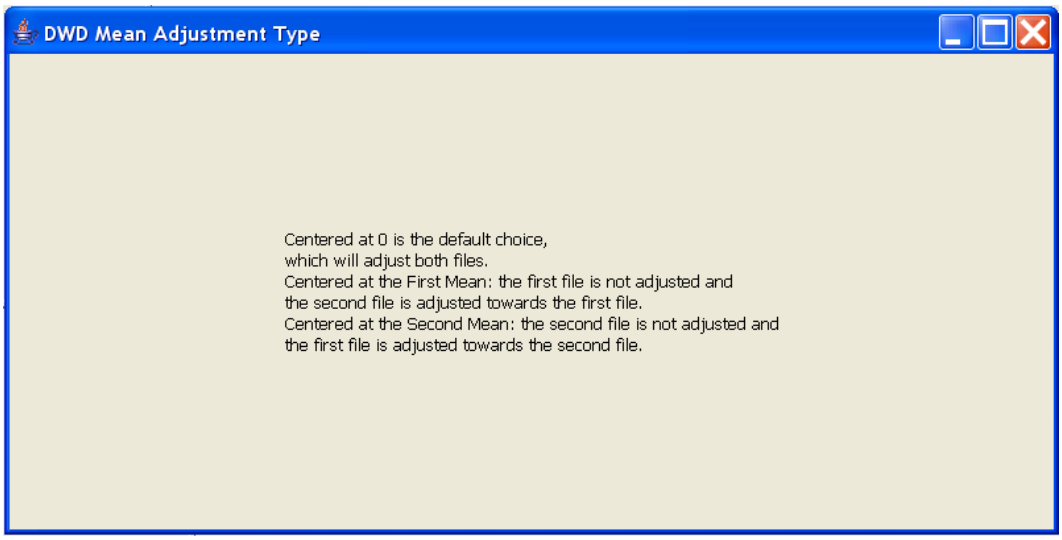


Fig. 6

Here are more details for those two parameters.

DWD Type:

Standardized DWD: This option should be selected when the two data sets are not comparable in terms of scale (i.e. range of the expression values), such as occurs when merging data from different platforms, such as Affymetrix and Agilent.

Non-Standardized DWD: This option is useful when the scale (of the expression values) of both data sets are similar, e.g. for merging two data sets within the same lab, but from two fabrication batches.

Mean Adjust Type:

Center at 0: Adjust the data so that both data sets have mean 0, when projected onto the DWD direction vector. This should be selected when both data sets measure differential expression (e.g. Agilent or cDNA), or when the resulting merged data will be thought of as differential (e.g. an Agilent and an Affymetrix data set are being merged).

Center at the First Mean: Adjust data to have the same mean as the first data set, when projected onto the DWD direction vector. This should be used for platforms involving absolute expression values, such as Affymetrix, and when the first data file has more samples or is considered as the standard one. For example, fix the training dataset and adjust the testing dataset only

Center at the Second Mean: Adjust data to have the same mean as the first data set, when projected onto the DWD direction vector. Conditions for use of this are the same as above, expect the second data set is considered to be the standard one.

- Running of SSP after DWD

In our protocol, we run a DWD for training and testing dataset if the two datasets come from different platforms or different batches. Clicking SSP will bring up the GUI for SSP as shown in the Fig. 7.

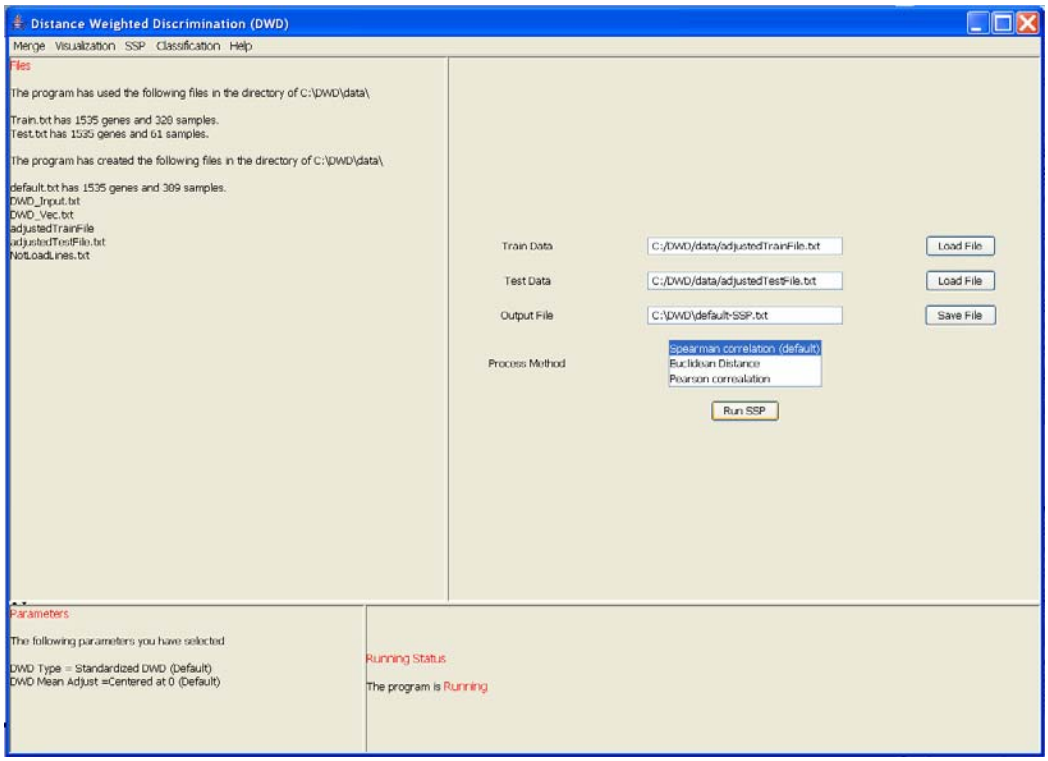


Fig. 7

. Output

A. For Stanford-like text delimited file.

The output files include: DWD_input.txt, DWD_Vec.txt, and DWD_Non_Std_Output.txt/ DWD_Std_Output.txt (if you use default output). DWD_Non_Std_Output.txt/ DWD_Std_Output.txt is the final output corresponding to the two DWD types. But other files (DWD_input.txt and DWD_Vec.txt) will also be used in the visual diagnostics analysis. Please do not delete them. They are automatically overwritten from one run of DWD to another.

B. For MAGE-ML format file.

In addition to the same files generated as above, there are two extra output files, when MAGE-ML format files are used as input. They are ExternalAdjustedDataFile.txt and DWD_Non_Std_Output.xml/ DWD_Std_Output.xml (if you use default output). The adjusted data will be stored in an external text file named ExternalAdjustedDataFile.txt. The text file,

DWD_Non_Std_Output.txt/ DWD_Std_Output.txt (if you use default output) will be used for the visualization.

11 Interpretation of DWD-SSP output

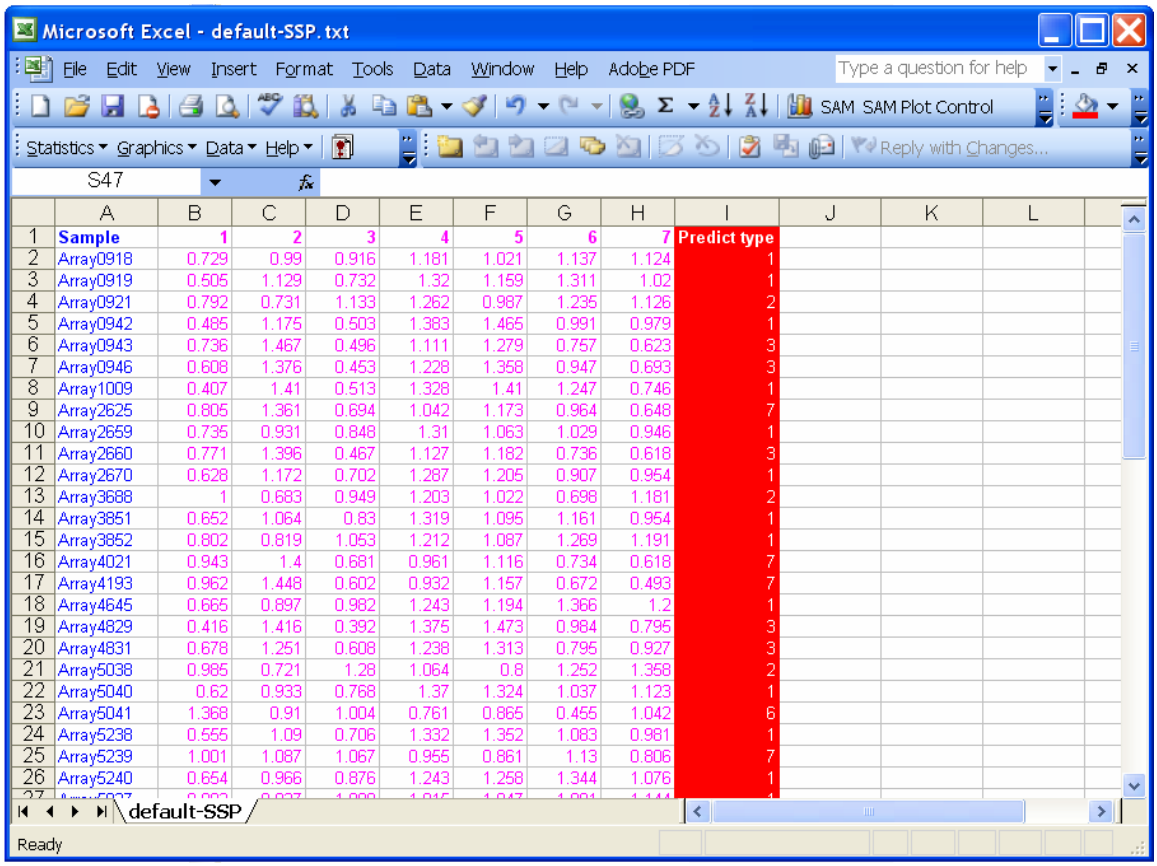
11.1 Output of DWD

The output files of DWD are in the same format of the input files, tab delimited txt files in Stanford microarray data format. But the output files have been DWD adjusted to remove the batch bias or platform bias.

11.2 Output of SSP

The output files of SSP are also in tab delimited txt format as displayed in the Fig. 8.. The first row (in bold) tells the target variables (for example, the subtypes the cancer, the drug response group, etc). The first column (in blue) is the sample name. The last column (in white color and red background) is the predicted groups (predicted target variable). The real numbers in the middle (in pink) are the distance of each sample to the centroids of all the groups.

Right now, the SSP software is using three distance functions: Euclidean Distance, (1 - Pearson Correlation) and (1 - Spearman Correlation).



	A	B	C	D	E	F	G	H	I	J	K	L
1	**Sample**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Predict type**			
2	Array0918	0.729	0.99	0.916	1.181	1.021	1.137	1.124	1			
3	Array0919	0.505	1.129	0.732	1.32	1.159	1.311	1.02	1			
4	Array0921	0.792	0.731	1.133	1.262	0.987	1.235	1.126	2			
5	Array0942	0.485	1.175	0.503	1.383	1.465	0.991	0.979	1			
6	Array0943	0.736	1.467	0.496	1.111	1.279	0.757	0.623	3			
7	Array0946	0.608	1.376	0.453	1.228	1.358	0.947	0.693	3			
8	Array1009	0.407	1.41	0.513	1.328	1.41	1.247	0.746	1			
9	Array2625	0.805	1.361	0.694	1.042	1.173	0.964	0.648	7			
10	Array2659	0.735	0.931	0.848	1.31	1.063	1.029	0.946	1			
11	Array2660	0.771	1.396	0.467	1.127	1.182	0.736	0.618	3			
12	Array2670	0.628	1.172	0.702	1.287	1.205	0.907	0.954	1			
13	Array3688	1	0.683	0.949	1.203	1.022	0.698	1.181	2			
14	Array3851	0.652	1.064	0.83	1.319	1.095	1.161	0.954	1			
15	Array3852	0.802	0.819	1.053	1.212	1.087	1.269	1.191	1			
16	Array4021	0.943	1.4	0.681	0.961	1.116	0.734	0.618	7			
17	Array4193	0.962	1.448	0.602	0.932	1.157	0.672	0.493	7			
18	Array4645	0.665	0.897	0.982	1.243	1.194	1.366	1.2	1			
19	Array4829	0.416	1.416	0.392	1.375	1.473	0.984	0.795	3			
20	Array4831	0.678	1.251	0.608	1.238	1.313	0.795	0.927	3			
21	Array5038	0.985	0.721	1.28	1.064	0.8	1.252	1.358	2			
22	Array5040	0.62	0.933	0.768	1.37	1.324	1.037	1.123	1			
23	Array5041	1.368	0.91	1.004	0.761	0.865	0.455	1.042	6			
24	Array5238	0.555	1.09	0.706	1.332	1.352	1.083	0.981	1			
25	Array5239	1.001	1.087	1.067	0.955	0.861	1.13	0.806	7			
26	Array5240	0.654	0.966	0.876	1.243	1.258	1.344	1.076	1			

Fig. 8

12 References

The methods of DWD and SSP have been described in detail at:

Hu Z. et al. **The molecular portraits of breast tumors are conserved across microarray platforms**. *BMC Genomics* 2006; 7:96-96.

Benito M. et al. **Adjustment of systematic microarray data biases.** *Bioinformatics* 2004, **20**(1)**:**105-114.