

# TMA-Combiner, a Software Tool for TMA Replicate Cores

Contains documentation for the following:

The TMA-Combiner  
Version 1.00  
Usage Walkthrough

Stainfinder  
Version 1.1  
Usage Walkthrough

# Table of Contents

<b>Terms and Conditions for Use</b> .....	<b>3</b>
<b>System Requirements</b> .....	<b>5</b>
<b>Walkthrough</b> .....	<b>6</b>
TMA-Combiner Walkthrough .....	7
Stainfinder Version 1.1 .....	18
<b>Appendix</b> .....	<b>21</b>
PCL File Format .....	21
Quantitative Scoring Systems .....	24
Revision History .....	27
Frequently Asked Questions (FAQ) .....	28

## **Terms and Conditions for Use**

Software is being provided free of charge, provided that:

1. User is affiliated with an academic or non-profit organization
2. User may not distribute any modified versions of the software or any of its accompanying documentation without prior consultation with the authors
3. User provides proper citation or acknowledgment if any of the software programs documented in this file are used in a paper that has been accepted for publication

Because the software is being provided free of charge, the authors do not claim responsibility for any damages, monetary or otherwise, stemming from misuse of this software.

Commercial interests must contact the authors for proper licensing arrangements prior to use.

Please refer to the TMA-Combiner website, <http://genome-www.stanford.edu/TMA/combiner>, for correspondence information of the authors.

## Overview

The TMA-Combiner is a TMA dataset combination program designed to be used with the TMA-Deconvoluter. It serves the two following main functions:

1. To combine replicate cores within a body of TMA data, residing either in a single TMA or in multiple TMAs, and
2. To combine multiple TMA together into a single file for analysis.

If you are not familiar with the TMA-Deconvoluter or the general process for high-throughput analysis of tissue microarrays (TMAs), please refer to our prior publication, “Software Tools for High-throughput Analysis and Archiving of Immunohistochemistry Staining Data Obtained with Tissue Microarrays” (Am. J. Path Nov 2002, Vol. 161, No. 5, pp. 1557-1565), its accompanying web supplement, <http://genome-www.stanford.edu/TMA/>, and documentation of the TMA-Deconvoluter. The remainder of this documentation will assume that the user is familiar with the TMA-Deconvoluter and its accompanying process for acquiring and analyzing TMA data.

For those of you who are familiar with the TMA-Deconvoluter, the TMA-Combiner integrates easily into our TMA data processing method. It accepts one or more text tab-delimited files in PCL format, outputted by the TMA-Deconvoluter, where each file represents the deconvoluted scores of a single TMA. For each combination batch, the TMA-Combiner outputs a single text tab-delimited file in PCL format, which you can use for subsequent analysis (e.g. Cluster, TreeView, Stainfinder), just as you would with the output files from the TMA-Deconvoluter.

In this document, you will find instructions on how to modify your existing system of TMA data management to take advantage of the dataset combination ability provided for by the TMA-Combiner. You will find descriptions on the requirements and operation of the TMA-Combiner in this document, along with a description on the updates made to Stainfinder to handle combined datasets.

For further details on how you can adapt the TMA-Combiner to your existing high-throughput TMA data management system, please refer to the corresponding paper:

Liu et al., “TMA-Combiner, a Simple Software Tool to Permit Analysis of Replicate Cores on Tissue Microarrays,” (2005), *in review*.

The instructions and documentation you will find here will largely overlap the contents of the TMA-Combiner website, <http://genome-www.stanford.edu/TMA/combiner>. The only differences will involve features not documented in the website that relate to the finer points of operating the TMA-Combiner.

Please note that important feature additions, changes, bug fixes, and any other changes affecting this TMA data management system will be updated on the website more frequently than in this documentation.

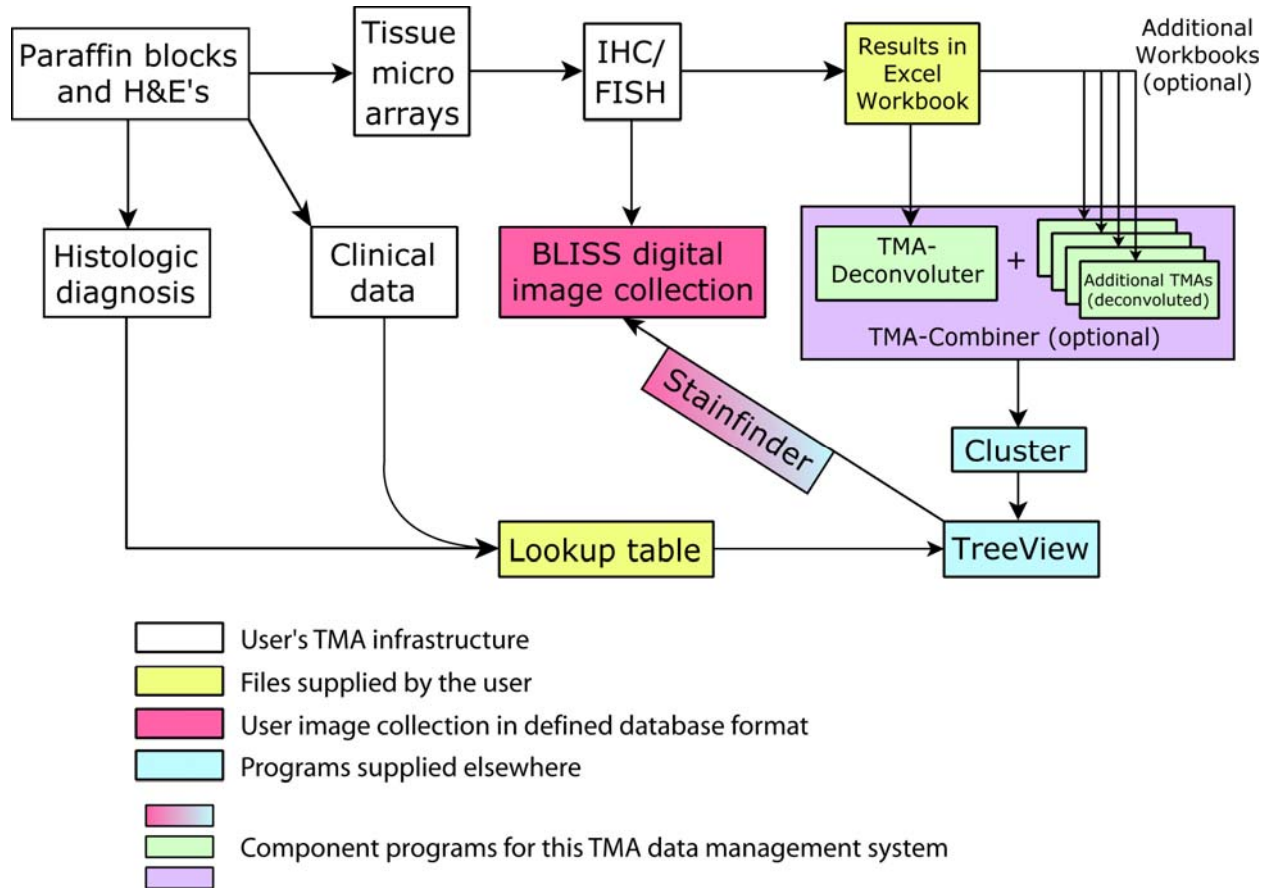
## System Requirements

This is a quick summary of the items you will need to use the TMA-Combiner. Please note that this assumes prior experience with the TMA-Deconvoluter; if you are new, please begin at the TMA website. You will be needing the following:

- “Deconvoluted” PCL-format text tab-delimited files. Files in other formats may also work but may result in suboptimal operation. Please read the file format requirements before proceeding.
- The TreeView and Cluster programs. Please refer to the TMA system requirements for additional details.
- Optional: An on-line database of stored images. As with the TMA-Deconvoluter, the TMA-Combiner is designed to work with Stainfinder and an on-line database of stored images. Please refer to the TMA website for details.
- A Windows PC running Microsoft Excel 2000 or later. This is necessary to run the TMA-Combiner program. *Note: the Mac OS is NOT supported*, because Active X controls are used, which are not available in Macintosh versions of Microsoft Office. Furthermore, **ANY older version of Office (including Office 97) will not work either.**

Once you have all of these components assembled, proceed to the walkthrough section.

# Walkthrough



The diagram above provides an overview of the TMA data management system and where the various system components lie in relation to each other. You may note that this diagram has been updated from the initial publication of the TMA-Deconvoluter with the incorporation of the TMA-Combiner. The TMA-Combiner remains a fully optional component of this system, but its use integrates easily and is crucial for combining core replicates and multiple TMAs together.

Once you have produced the necessary deconvoluted TMA files (as outlined in green in the diagram above), proceed to the walkthrough for using the TMA-Combiner to generate a single combined TMA dataset file for subsequent data analysis.

Afterwards, proceed to the updated description of Stainfinder to see what modifications you may need to make to enable Stainfinder to correctly display combined TMA datasets, when viewed under TreeView with an on-line Bliss digital image collection.

## **TMA-Combiner Walkthrough**

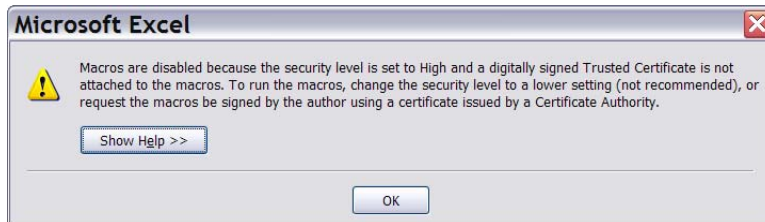
*Note: if this is your first time using the TMA-Combiner, it is recommended that you proceed to the Downloads section of the TMA-Combiner website, <http://genome-www.stanford.edu/TMA/combiner/download.shtml> and download the Demo Suite files. The examples presented in this walkthrough employ those files, and you should obtain the same results in using these files and with the other necessary programs available elsewhere.*

What you will need for this walkthrough:

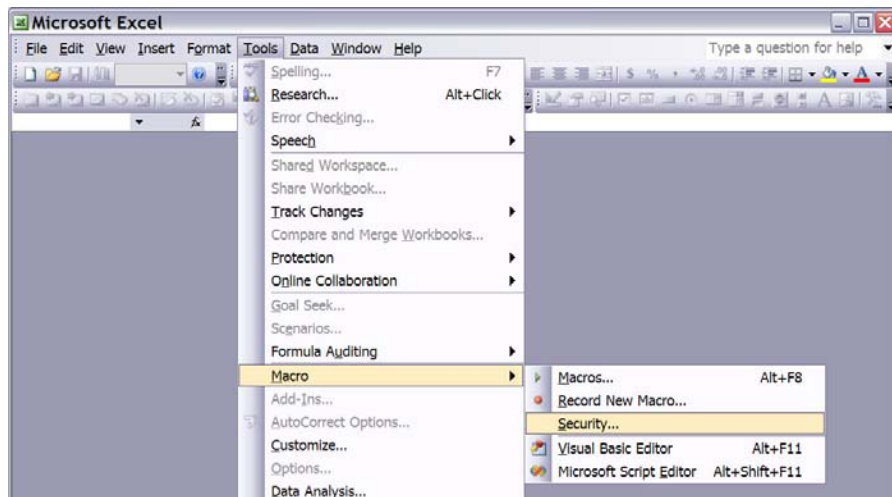
- Deconvoluted text tab-delimited files containing your TMA data
- The TMA-Combiner
- The Cluster program
- The TreeView program
- Microsoft Excel for Windows with Visual Basic support installed
- A Windows PC

Begin by storing a copy of the TMA-Combiner and your deconvoluted TMA files in the same folder. (If you decide to put the files elsewhere, you will need to know how to specify the appropriate path information in the entry fields for those files).

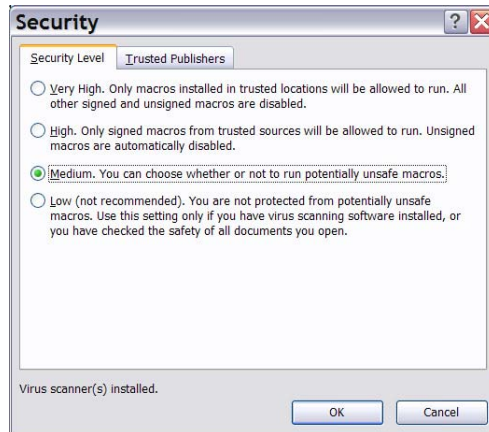
Note that Office XP, 2003, and later usually defaults the security level to “high”. In that situation, you may get this dialog box:



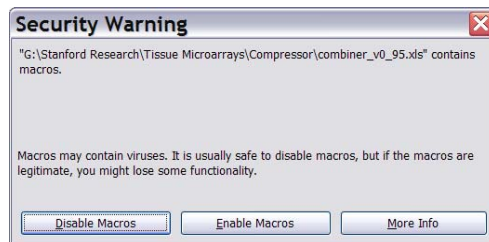
To fix this, go to the Tools menu, select Macros, and then select Security, as shown below.



You will get this dialog box:

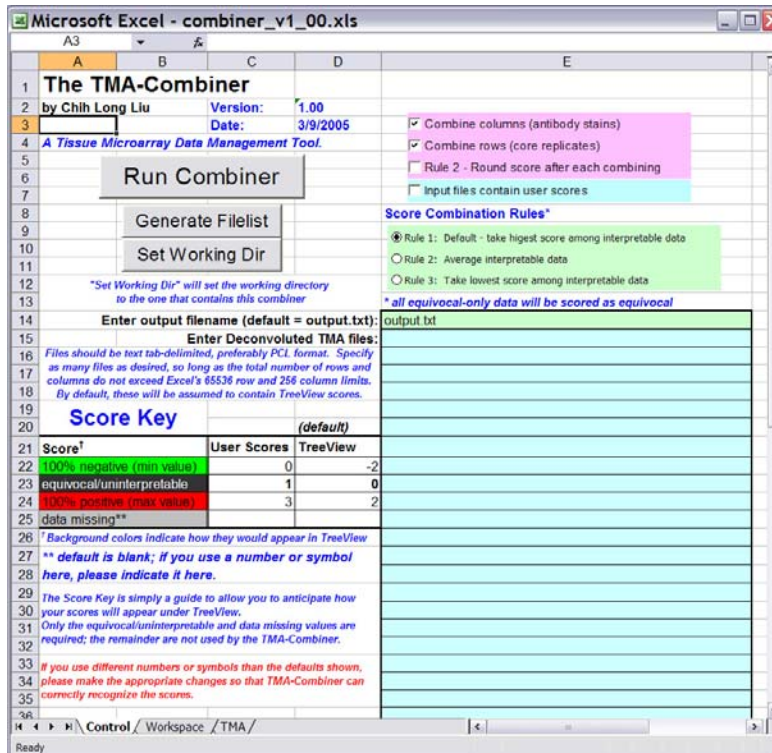


Please ensure that “Medium” is selected. If you have “High” or “Very High” selected, the TMA-Combiner will not run. Once you’ve clicked “OK”, you may now open the TMA-Combiner. A dialog box might pop up before the TMA-Combiner opens, asking you if you want to enable macros (as shown below). Click “Enable Macros”. If you leave them disabled, the TMA-Combiner will not run.



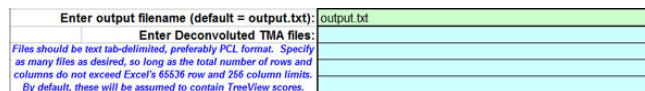
After opening the TMA-Combiner, you should be greeted with the screen below:





Excel normally sets the current working directory based on your default settings, which may not necessarily be the directory containing the TMA-Combiner and the other files that you placed in that folder. **Click on “Set Working Dir”** to set the working directory to be the container for your TMA files. You may see a dialog box indicating the current working directory and the new working directory, and it will request whether this is correct. Click “Yes” or “No” to continue.

### Specifying the filenames

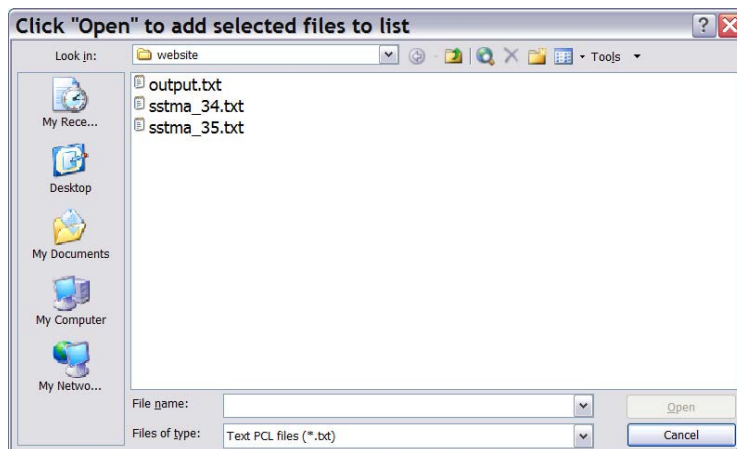


Specify the names of your deconvoluted, PCL-format text tab-delimited files in the light blue worksheet cells. Please refer to the Appendix for information on this file format. As mentioned above, you can specify as many files as you want, so long as the total number of rows and columns do not exceed Excel’s 65536 row and 256 column limits. If you have more files than the number of light blue worksheet cells available, then that is fine; all files will be processed even if they are not in a light blue worksheet cell, so long as the file list is contiguous (i.e. no gaps) from the very first cell located next to “Enter Deconvoluted TMA files” as shown above.

**Note:** ALL files you specify will be combined into a SINGLE output file. If you actually wanted to combine your TMAs into more than one separate combined dataset, you will need to process each combined dataset separately. Furthermore, if portions of your TMA dataset to be combined need different score combination rules applied, you should combine those portions separately with the appropriate rule before combining the combined files together. Only then can you

assemble the entire dataset with the TMA-Combiner (at which point you can select any rule; it won't make a difference, at this stage, in the final combined result).

You may type these in manually. Alternatively, particularly if you are going to process multiple files, you can **click on “Generate Filelist”**. This allows you to graphically select the desired files in your folder, instead of having to type in all the file names of the list of files you wish to process. An **Open file** dialog box will appear below:



Select files by clicking and dragging. To select multiple files that are not listed next to each other, hold down the Ctrl key while selecting the desired files, either with the mouse or by using the arrow keys and the spacebar. **Click “Open”** to add those files to your list.

### **Selecting Score Combination Rules**

This section represents what is perhaps the most important part of the TMA-Combiner. At first glance, one may think that combining scores by simple averaging would be sufficient. However, when it comes to combining stained and scored IHC replicate cores, the situation is not that simple! One needs to take a step back and carefully consider the significance of what the combination process is doing to these IHC-stained replicate cores. Issues such as core sampling biases, antibody staining properties, and number of replicate cores weigh considerably in this process.

We discuss the importance of these considerations and why we constructed the score combination rules shown below, in our publication, so we won't repeat it here. We will, however, go into greater detail on the mechanics of the score combination process, particularly on how each rule determines the outcome score from the input score pool of replicate cores.

First off, it is important to establish the scoring system that is introduced earlier in the TMA-Deconvoluter walkthrough. At this time, the TMA-Combiner can handle only the 5-member scoring system in use in the TMA-Deconvoluter. The TMA-Combiner can handle quantitative scoring systems, which we describe in greater detail here. However, we use a discrete integer scoring system, so the TMA-Combiner defaults to that, as shown in the score key below:

Score Key		
		(default)
Score <sup>†</sup>	User Scores	TreeView
100% negative (min value)	0	-2
equivocal/uninterpretable	1	0
100% positive (max value)	3	2
data missing**		

<sup>†</sup> Background colors indicate how they would appear in TreeView

Note that scores are either user scores or TreeView scores. This is dependent on the type of output selected for the TMA-Deconvoluter - PCL files use TreeView-compatible scores, whereas K-M files use unconverted, user scores. User scores might also be present if the user generated the input files from other sources or with methods other than the TMA-Deconvoluter.

In order for the TMA-Combiner to apply the score combination rules correctly, the user will need to indicate whether or not user scores are present in the input files. Since the PCL file format is the native format for the TMA-Combiner, the TMA-Combiner will assume by default that the input files contain TreeView-compatible scores. Thus, if the user is combining scores based on the user's scoring system, the user should select the option below:

Input files contain unconverted scores

\*\*In the case of missing datapoints, while the default is a blank cell, the user may use a number or symbol (such as "X") instead. This should be indicated in the Score Key, like in the example below:

Score Key		
		(default)
Score <sup>†</sup>	User Scores	TreeView
100% negative (min value)	0	-2
equivocal/uninterpretable	1	0
100% positive (max value)	3	2
data missing**	X	

One other, **very important** consideration with scores – if the user is planning to combine multiple TMAs together into a single file, the user should ensure that *every file in the entire dataset consists entirely of user scores or of TreeView-compatible scores (i.e. the constituent files comprising the whole dataset to be combined should contain only one type of score, not both)*. Otherwise, the output may contain errors, and/or the Combiner may crash. The easiest way to fix this is to use the TMA-Deconvoluter's Score Conversion Utility to convert the scores. For more information on this, consult the TMA-Deconvoluter documentation.

The Score Combination Rules appear in the TMA-Combiner as a number of options, as shown below:

**Score Combination Rules\***

Rule 1: Default - take highest score among interpretable data

Rule 2: Average interpretable data

Rule 3: Take lowest score among interpretable data

\* all equivocal-only data will be scored as equivocal

Note - in the examples shown in the tables below, the scoring system shown is the discrete integer system we currently use. However, the cases as illustrated below also apply to

quantitative scoring systems. If  $n$  = minimum score,  $p$  = maximum score,  $n_i$  = intermediate negative score,  $p_i$  = intermediate positive score, and  $u$  = equivocal/uninterpretable score, setting  $-2 = n$ ,  $-2 < n_i < 0$ ,  $0 = u$ ,  $1 = p_i$ , and  $2 = p$  in the tables below should yield the general case for a quantitative scoring system.

**Rule 1 – take highest score among interpretable data.**

This rule consists of taking the highest score among interpretable data. It is set as the default rule because it corresponds to the standard clinical practice of diagnostic IHC used for most antibodies.

Below is a table indicating a number of different possible cases, each having four replicate cores. Each number shown is a “converted” TreeView-compatible score, with its corresponding background color matched to how it would appear in the TreeView heatmap.

Rule 1	Score 1	Score 2	Score 3	Score 4	Combined Score
Case 1	2	2	2	2	2
Case 2	-2	-2	-2	2	2
Case 3	-2	-2	1	-2	1
Case 4	-2	0	-2	-2	-2
Case 5	0	0	0	-2	-2
Case 6	0	0	0	0	0
Case 7	missing	missing	missing	missing	missing
Case 8	missing	missing	missing	2	2

As one would expect with Case 1, combining four replicates with identical scores results in the same score. This is also true for Case 6 (which will be true for all three rules). Also true for all three rules is Case 7, where missing data for all four replicates results in a missing data score. Case 8 represents a very frequent occurrence when replicate scores from multiple TMA datasets are combined. This results when that particular replicate and antibody stain is unique to one of the TMAs being combined.

In Cases 2 and 3, the effects of Rule 1 are clearly shown here, where the high scores of 2 and 1, respectively, are used for the combined score. One can also observe that, for Cases 4 and 5, the equivocal scores are being eliminated from consideration before the combined score is calculated – even though the equivocal score is numerically higher than a negative stain score, the “highest” score is still taken to be the negative stain score.

**Rule 2 – average interpretable data.**

This rule is more appropriate in cases where IHC staining is known to be quantitative. Below is another table, here illustrating the properties of Rule 2:

Rule 2	Score 1	Score 2	Score 3	Score 4	Combined Score
Case 1	2	2	2	2	2
Case 2	-2	-2	-2	2	-1
Case 3	-2	1	1	2	0.5
Case 4	-2	0	1	1	0
Case 5	2	0	0	-2	0
Case 6	0	0	0	-2	-2

Here, in Cases 1-5, the scores are clearly the arithmetic mean of the replicate scores. The biological interpretation can be clearly different here than in Rule 1, particularly for Cases 4 and 5 – if a number of replicate cores taken from a given biopsy produce an average score of 0, this would seem to suggest that the heterogeneity in the scores of the sampled cores would make it difficult to interpret the overall score of the biopsy as a whole. Case 6, here, illustrates that the Rule 2 averaging will not count equivocal scores towards the average (if it did, the combined score would be -0.5).

**Rule 3 – take lowest score among interpretable data.**

This rule consists of taking the lowest score among interpretable data. This rule may seem counterintuitive at first, but its use becomes more apparent when one considers IHC staining for antibodies and tissue types that are susceptible to false positives or which only strong positive staining correlates with a biologically significant outcome. Below is a table illustrating properties of Rule 3:

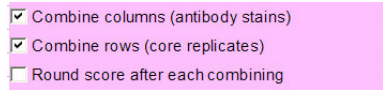
Rule 3	Score 1	Score 2	Score 3	Score 4	Combined Score
Case 1	2	2	2	2	2
Case 2	2	2	1	2	1
Case 3	-2	2	2	2	-2
Case 4	2	0	1	1	1
Case 5	0	0	0	2	2

Note that only in Cases 1 and 5 is a strong positive score of 2 obtained; in Cases 2 and 3, the lowest replicate score is used to represent the entire case. In Cases 4 and 5, the equivocal scores are again not taken into consideration, even though it presents the numerically lowest score in their respective cases.

*Note: the current version of the TMA-Combiner can only apply one score combination rule per processing run. If your TMA datasets to be combined contain IHC stains that require different score combination rules to be applied, you will need to separate your TMA datasets accordingly and process each one separately.*

**Other Combination Options**

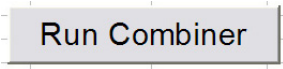
Lastly, there are some score combination options as shown below:



By default, both rows (core replicates) and columns (antibody stains) are selected. However, if desired, it is possible to combine only the rows or only the columns. If you try to deselect both options, however, the TMA-Combiner will catch this and abort the run.

Scores can also be rounded after each combining. This applies only if both the *combine columns* and the *combine rows* options are checked. The combination order is rows, then columns. Score rounding only matters in the case of Rule 2, and rounding is performed to the nearest integer. Since TreeView can handle non-integral values, this option is unselected by default but is available for users may desire to use this feature.

**Run the TMA-Combiner**



You are now ready to run the TMA-Combiner. Click on the button as shown above.

**The Output File**

After the TMA-Combiner has completed its operation, you will find yourself back at the main screen of the TMA-Combiner (at the “Control”) worksheet. The output file should be located in the same working directory. When you open it, you will find that the file is still in PCL format, such as in the example below:

	A	B	C	D	E
1	UID	NAME	GWHEIGHT	bc12	cam5.2
2	EWHEIGHT				
3	1_19_1_1_14_1040_1.jpg*1_19_2_1_14_1040_2.jpg*11cam5.2 pankeratinlck7 em(2)1	SS   monophasic   gr 3   media	1	2	-2
4	1_19_3_1_14_1040_3.jpg*1_19_4_1_14_1040_4.jpg*12cam5.2 pankeratinlck7 em(2)2	SS   monophasic   gr 2   thigh	1	1	1
5	1_19_5_1_14_1040_5.jpg*1_19_5_2_14_1040_17.jpg*13cam5.2 pankeratinlck7 ei(2)3	SS   biphasic   gr 3   chest wall	1	2	-2
6	1_19_3_2_14_1040_19.jpg*1_19_4_2_14_1040_18.jpg*14cam5.2 pankeratinlck7 h(2)4	SS   monophasic   gr 3   proime	1	2	1
7	1_19_1_2_14_1040_21.jpg*1_19_2_2_14_1040_20.jpg*15cam5.2 pankeratinlck7 h(2)5	SS   monophasic   gr 2   should	1	2	1
8	1_19_1_3_14_1040_23.jpg*1_19_2_3_14_1040_24.jpg*16cam5.2 pankeratinlck7 h(2)6	SS   monophasic   gr 2   proime	1	1	-2
9	1_19_3_3_14_1040_25.jpg*1_19_4_3_14_1040_26.jpg*17cam5.2 pankeratinlck7 h(2)7	SS   biphasic   gr 3   chest wall	1	2	2
10	1_19_5_3_14_1040_27.jpg*1_19_5_4_14_1040_39.jpg*18cam5.2 pankeratinlck7 h(2)8	SS   monophasic   gr 2   proime	1	-2	-2
11	1_19_3_4_14_1040_41.jpg*1_19_4_4_14_1040_40.jpg*19cam5.2 pankeratinlck7 h(2)9	SS   monophasic   gr 2   lung	1	2	-2
12	1_19_1_4_14_1040_43.jpg*1_19_2_4_14_1040_42.jpg*10cam5.2 pankeratinlck7(2)10	SS   monophasic   gr 3   pelvis	1	2	-2
13	1_19_1_6_14_1040_55.jpg*1_19_2_6_14_1040_54.jpg*11cam5.2 pankeratinlck7(2)11	SS   monophasic   gr 2   perip	1	2	-2
14	1_19_3_6_14_1040_53.jpg*1_19_4_6_14_1040_52.jpg*12cam5.2 pankeratinlck7(2)12	SS   monophasic   gr 2   supe	1	2	-2
15	1_19_5_6_14_1040_51.jpg*1_19_5_7_14_1040_61.jpg*13cam5.2 pankeratinlck7(2)13	SS   biphasic   gr 2   proimal t	1	1	2
16	1_19_3_7_14_1040_59.jpg*1_19_4_7_14_1040_60.jpg*14cam5.2 pankeratinlck7(2)14	SS   monophasic   gr 2   abdo	1	2	2
17	1_19_1_7_14_1040_57.jpg*1_19_2_7_14_1040_58.jpg*15cam5.2 pankeratinlck7(2)15	SS   monophasic   gr 2   foree	1	2	2
18	1_19_1_8_14_1040_77.jpg*1_19_2_8_14_1040_76.jpg*16cam5.2 pankeratinlck7(2)16	SS   monophasic   gr 3   shou	1	1	2
19	1_19_3_8_14_1040_75.jpg*1_19_4_8_14_1040_74.jpg*17cam5.2 pankeratinlck7(2)17	SS   monophasic   gr 2   lumbi	1	1	-2
20	1_19_5_8_14_1040_73.jpg*1_19_5_9_14_1040_83.jpg*18cam5.2 pankeratinlck7(2)18	SS   monophasic   gr 2   proim	1	-2	-2

You will immediately notice that if your aggregate dataset contained replicate cores, there is now only one unique entry to represent each combined set of cores, and that such an entry will have a number in parenthesis preceding the rest of the entry. In the “NAME” column, the core descriptor will appear like the example below:

```
(2)10 | SS | monophasic | gr 3 | pelvis |
```

where the original descriptor in your input dataset would have been as follows:

```
10 | SS | monophasic | gr 3 | pelvis |
```

The (2) indicates that the case 106 entry in the combined dataset contains the combined score derived from up to 2 replicate cores. This is also true for antibodies – in the example above if *bcl2* had been combined from two different bcl2 scores, the column header would appear as (2)bcl2.

Core and antibody replicates that occur only once in the aggregate dataset, however, will appear in the combined output file unchanged, and they will not have a preceding number in parenthesis. If you notice some antibody replicates that unexpectedly remained uncombined, you may want to go back and double-check to ensure that your files fulfill the file format requirements.

The other major change you will notice is the format of the entries in the UID column. Here, the jpg filenames corresponding to the three replicate cores are present here, as well as all antibody names for which a score was available in the uncombined dataset. The corresponding UID entry for the example case 106 is shown below:

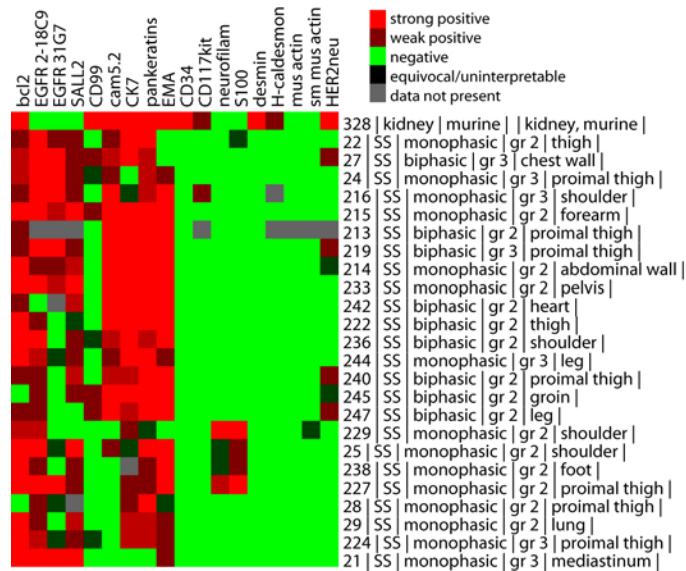
```
1_19_1_4_14_1040_43.jpg*1_19_2_4_14_1040_42.jpg!10!cam5.2!pankeratin!ck7!ema!  
o13(cd99)!...
```

*(not all of the UID entry is shown, for sake of brevity; the important section that has changed from the uncombined dataset is shown above in red text)*

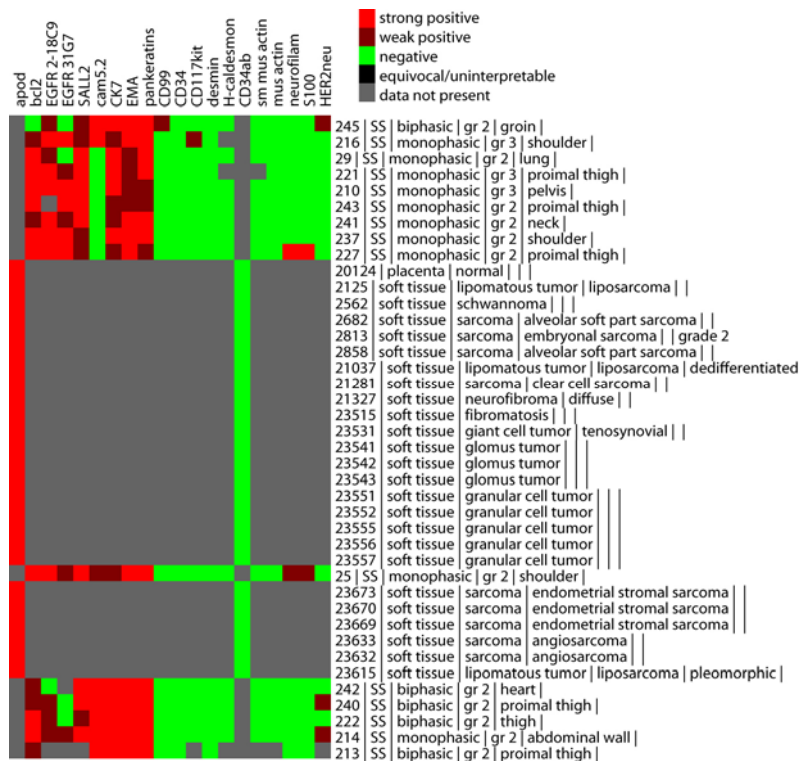
The significance of the modifications in the formatting of the UID column information will be covered in more detail in the Stainfinder update section. For those of you who do not use Stainfinder, you may safely disregard the information in this column.

## **Cluster and TreeView**

For clustering analysis and visualization under TreeView, you may treat this output file as you would with any other PCL-format output file generated by the TMA-Deconvoluter. The combined datasets should not appear significantly different from the uncombined data, save for the differences in the data deriving from the score combination process. There are two exceptions – one when Rule 2 score combining is performed – in this case, you may notice additional, intermediate color gradations between the colors shown in the standard score key, such as in the example below. This is to be expected and should be interpreted at face value.



The other exception occurs when you combine multiple TMA datasets together when there is very little overlap. Under such a situation, the heatmap in TreeView will display large regions of gray, indicating the lack of overlap. An example of this is shown below.



If you do not know how to use Cluster and TreeView, please refer to the TMA website and documentation.



You may now browse your dataset for additional analysis. If you use Stainfinder and wish to use it to view combined TMA datasets, proceed to the Stainfinder update section, which begins on the next page.

## **Stainfinder Version 1.1**

### **Updates to Stainfinder - Version 1.1**

If this is the first time you've set up Stainfinder, please go to the Stainfinder Walkthrough first, which you will find in the Deconvoluter Documentation.

Since the initial publication of our first paper, a number of updates have been made to the functionality of Stainfinder. They include the following:

- An ability to display replicate cores for TMA datasets containing combined scores
- An ability to compartmentalize multiple datasets into their own spaces. More on this below.
- No case sensitivity in matching antibody names passed in from the URL to the directories listed on the pathologist's TMA image database. However, the code can be easily modified to restore case sensitivity.
- A "check all checkboxes" option for displaying antibody stains to display

This updated code is available in the Downloads section.

#### **Displaying replicate cores**

Combined TMA datasets will display a single combined score for the appropriate cores, when viewed under TreeView or under the CaseExplorer datasets published in the TMA Portal. Stainfinder 1.1, however, is capable of displaying all of the replicate cores used to produce the combined score, in a properly configured system. Please visit the explore page (<http://genome-www.stanford.edu/TMA/combiner/explore.shtml>), where you can browse a number of datasets published by our consortium, that have been configured in this manner. Typically, two replicate cores are displayed for each combined score, though you may on occasion run across cases that display more than two replicate cores.

#### **Multiple Datasets**

Pathology labs wishing to publish datasets on their own servers using Stainfinder, whether publicly or on their own intranet, previously had to install and configure a separate Stainfinder instance for each dataset. In this updated version of Stainfinder, a single copy of Stainfinder is now sufficient, and a minor modification can be made in the URL to access the dataset. See the configuration section below to see how Stainfinder 1.1 should be configured.

#### **Configuration**

*Note: these instructions assume that you have prior experience in configuring Stainfinder and will only contain instructions on modifications that need to be done on an existing Stainfinder installation. These instructions also assume that you are reasonably comfortable with UNIX/Linux command line syntax and that you've had some exposure to PERL and programming in CGI. If these sound like alien things to you, it is strongly recommended that you*

*seek assistance of an experienced UNIX/Linux systems administrator to assist you in configuring Stainfinder 1.1. Furthermore, if this is the first time you have attempted to configure Stainfinder, please refer to the Stainfinder walkthrough on the original TMA website.*

When you are ready, proceed to the next section to configure Stainfinder 1.1.

### **Configuring Stainfinder 1.1 to work with Multiple Datasets**

In Version 1.1 of Stainfinder, you may notice that the code referring to server path names have changed. This is to accommodate Stainfinder's ability to access multiple datasets that are each stored in their own directory. For example, if your image database was located in your server under the path `/share/tissue/archive/scans/`, and is accessible online via `http://www.myserver.edu/scans/`, you would have configured Stainfinder 1.0 as follows:

```
my $rootDir = `/share/tissue/archive/`;
my $webDir = `/scans/`;
my $mainDir = `/share/tissue/archive/scans/`;
```

So a user accessing a server running Stainfinder 1.0 would probably click on a case descriptor in TreeView or on the “SF” button in CaseExplorer, a new browser window is opened with a URL that may look like this:

```
http://www.myserver.edu/cgi-bin/Stainfinder.pl?uniqid=
4_35_4_5_35_1128_436.jpg*4_35_5_5_35_1128_437.jpg!1326!apod!CD34
```

*(the URL is actually one single line; it's wrapped around to two lines here for ease of display)*

In Stainfinder 1.1, however, you will notice that `$rootDir` has gone away, and that there is now a new variable `$dirname`.

```
my $mainDir = `/share/tissue/archive/scans/`.$dirname.'/';
my $webDir = `/scans/`.$dirname.'/';
```

This variable gets passed in from the URL as the first item in the URL. In such a situation, a user accessing the “myTMAdataset” TMA dataset on a server running Stainfinder 1.1 would in the process open a new browser window with a URL that may look like this:

```
http://www.myserver.edu/cgi-bin/Stainfinder.pl?uniqid=
myTMAdataset!4_35_4_5_35_1128_436.jpg*4_35_5_5_35_1128_437.jpg!1326!apod!CD34
```

*(the URL is actually one single line; it's wrapped around to two lines here for ease of display; the change is highlighted in **red bold text** for clarity)*

What happens is that “myTMAdataset” is passed into the Stainfinder program and stored into the variable `$dirname`. Stainfinder will now look for images related to “myTMAdataset” in the following places:

```
my $mainDir = `/share/tissue/archive/scans/myTMAdataset/'
my $webdir = `/scans/myTMAdataset/'
```

The best place to configure your TMA Dataset to work with this feature of Stainfinder 1.1 is to append the dataset name to each value in the UID column such that it begins with the name of the dataset, followed by an exclamation point. For example, if a given cell in the UID column looked like this:

```
4_35_4_5_35_1128_436.jpg*4_35_5_5_35_1128_437.jpg!1326!apod!CD34
```

then you will want to change it to this:

```
myTMAdataset!4_35_4_5_35_1128_436.jpg*4_35_5_5_35_1128_437.jpg!1326!apod!CD34
```

This can be accomplished very simply by using an Excel formula like the following: “=“myTMAdataset!”&A4” (where A4 or its equivalent is the address that should point to the first cell in the UID column containing the dataset’s unique identifiers). In the combined PCL text file, copy and paste the formula for the entire dataset, and paste special as values back into the UID column. **Don’t forget that the name of the dataset has to match the name of the directory in which the images are stored. Also don’t forget that if your server is running UNIX/Linux, names are case sensitive.**

### **Configuring Stainfinder 1.1 to display replicate cores**

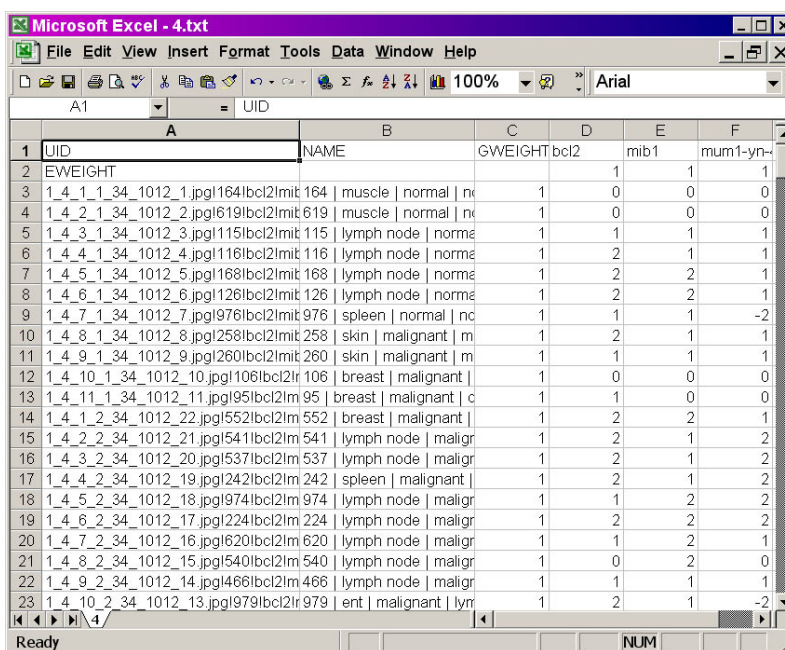
Configuring your image database to allow Stainfinder 1.1 to display replicate cores is actually quite simple. You will notice that the example URL as shown above contains some extra information compared to an uncombined dataset, which in this case is an additional JPG filename. Each of the JPG filenames refer to the individual core replicate images. These don’t have to be stored within the same subdirectories, but they do have to be present somewhere within the directory tree that contains all the images for the TMA dataset. No additional action is required - as long as your directory and JPG filenames conform to the Bliss image system [nomenclature](#), Stainfinder 1.1 should be able to find and retrieve those images.

# Appendix

## PCL File Format

The standard file format for the TMA-Combiner is the PCL format. The PCL format is the Pre CLuster format as described in the TMA-Deconvoluter walkthrough. It is one of the two output formats of the TMA-Deconvoluter. Since the TMA-Combiner is designed to work with the TMA-Deconvoluter, this should not present a problem to most users.

Below is an example of the PCL format:



UID	NAME	GWEIGHT	bcl2	mib1	mum1-yn
1	EWWEIGHT			1	1
2	1_4_1_1_34_1012_1.jpg 164 bcl2 mit 164   muscle   normal   n	1	0	0	0
3	1_4_2_1_34_1012_2.jpg 619 bcl2 mit 619   muscle   normal   n	1	0	0	0
4	1_4_3_1_34_1012_3.jpg 115 bcl2 mit 115   lymph node   norm	1	1	1	1
5	1_4_4_1_34_1012_4.jpg 116 bcl2 mit 116   lymph node   norm	1	2	1	1
6	1_4_5_1_34_1012_5.jpg 168 bcl2 mit 168   lymph node   norm	1	2	2	1
7	1_4_6_1_34_1012_6.jpg 126 bcl2 mit 126   lymph node   norm	1	2	2	1
8	1_4_7_1_34_1012_7.jpg 976 bcl2 mit 976   spleen   normal   nc	1	1	1	-2
9	1_4_8_1_34_1012_8.jpg 258 bcl2 mit 258   skin   malignant   m	1	2	1	1
10	1_4_9_1_34_1012_9.jpg 260 bcl2 mit 260   skin   malignant   m	1	1	1	1
11	1_4_10_1_34_1012_10.jpg 106 bcl2 lr 106   breast   malignant	1	0	0	0
12	1_4_11_1_34_1012_11.jpg 95 bcl2 lr 95   breast   malignant   c	1	1	0	0
13	1_4_1_2_34_1012_22.jpg 552 bcl2 lr 552   breast   malignant	1	2	2	1
14	1_4_2_2_34_1012_21.jpg 541 bcl2 lr 541   lymph node   maligr	1	2	1	2
15	1_4_3_2_34_1012_20.jpg 537 bcl2 lr 537   lymph node   maligr	1	2	1	2
16	1_4_4_2_34_1012_19.jpg 242 bcl2 lr 242   spleen   malignant	1	2	1	2
17	1_4_5_2_34_1012_18.jpg 974 bcl2 lr 974   lymph node   maligr	1	1	2	2
18	1_4_6_2_34_1012_17.jpg 224 bcl2 lr 224   lymph node   maligr	1	2	2	2
19	1_4_7_2_34_1012_16.jpg 620 bcl2 lr 620   lymph node   maligr	1	1	2	1
20	1_4_8_2_34_1012_15.jpg 540 bcl2 lr 540   lymph node   maligr	1	0	2	0
21	1_4_9_2_34_1012_14.jpg 466 bcl2 lr 466   lymph node   maligr	1	1	1	1
22	1_4_10_2_34_1012_13.jpg 979 bcl2 lr 979   ent   malignant   lym	1	2	1	-2

An example of the PCL file format, outputted by the TMA-Deconvoluter, ready for use with the TMA-Combiner.

- Column A: **UID** (for Unique IDentifier; required). If you use Stainfinder, this column contains the image filename and antibody stains that are passed into the Stainfinder program. The way this is done can be found [here](#) under the Stainfinder walkthrough.
- Column B: **NAME** (required). This is the most important column in the file, since the TMA-Combiner uses this as the basis for identifying replicates, for subsequent “compression”. Each cell contains a case number followed by various descriptors, all of them each separated by a “pipe” (“|”) delimiter. For example:

1208 | breast | malignant | carcinoma | ductal | invasive

Your NAME column must contain the case number as the very first item (1208 in this example), and your NAME column must use the “pipe” (“|”) character as the delimiter. Again, this is the standard format used in the TMA-Deconvoluter output files, so this should not pose any problems for most users.

- Column C: **GWEIGHT** (optional but highly recommended). This is the “GWEIGHT” column used by the Cluster program for providing the option of weighting cases differently (see TMA website and Cluster manual for details). The PCL file format incorporates this column by default; if it is absent in the input file, it will be inserted by the TMA-Combiner.
- Column D, etc.: **Antibodies**. Row 1 contains the name of the target protein for the antibody stain. If different TMAs contain the same antibody, and/or if a given TMA contains replicates or multiple score sets (e.g. by different pathologists), the name of the target protein should be separated with an underscore (“\_”) from the initials of the scorer or other unique identifying information. This is very important because TMA-Combiner will use that as the basis for determining what columns are to be combined, and any names that are not identical will be treated as different entities that will not be combined. For example:

Column	Before	After
D	bcl2	bcl2
E	mib1	mib1
F	er_mv-10-00	(2)er
G	er_lt-03-01	--
H	ER	ER
I	mib2_yv-10_03	mib2

Note that Column H will **NOT** be combined with Columns F and G, because the name matching is case sensitive. Furthermore, any annotations after the **leftmost** (“\_”) will be truncated (such as for Column I), even if the column is not combined with any other columns in the final dataset.

- Row 2: **EWEIGHT** (optional but highly recommended). This is the “EWEIGHT” row used by the Cluster program for providing the option of weighting antibodies differently (see TMA website and Cluster manual for details). The PCL file format incorporates this row by default; if it is absent in the input file, it will be inserted by the TMA-Combiner.

*Note: the TMA-Combiner will output files in PCL format, regardless of the format of the input files.*

### **Other formats**

While the PCL file format is the native format of the TMA-Combiner, it will recognize two other formats, for ease of convenience to the user. *Note: if problems arise, the user is requested to use the TMA-Deconvoluter to output into the PCL format.*

- **The K-M format.** A detailed description of this format is present [here](#). An example of this format is shown below.

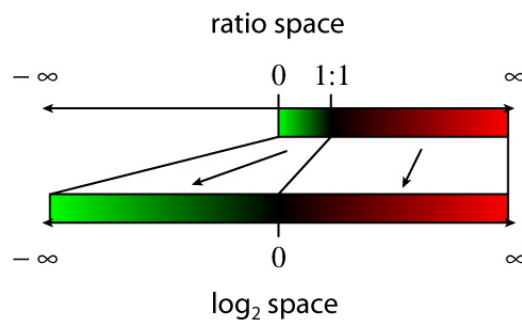
	A	B	C	D	E	F	G	H
1	s_a_c_r	Scan Filename for 1st Ab	FP #	bcl2	mib1	mum1-yr-	mum1-yr-	Description
2	1_4_1_1	1_4_1_1_34_1012_1.jpg	164		1	1	1	164   muscle   n
3	1_4_2_1	1_4_2_1_34_1012_2.jpg	619					619   muscle   n
4	1_4_3_1	1_4_3_1_34_1012_3.jpg	115	2	2	2	2	115   lymph nod
5	1_4_4_1	1_4_4_1_34_1012_4.jpg	116	3	2	2	2	116   lymph nod
6	1_4_5_1	1_4_5_1_34_1012_5.jpg	168	3	3	2	2	168   lymph nod
7	1_4_6_1	1_4_6_1_34_1012_6.jpg	126	3	3	2	2	126   lymph nod
8	1_4_7_1	1_4_7_1_34_1012_7.jpg	976	2	2	0	0	976   spleen   ne
9	1_4_8_1	1_4_8_1_34_1012_8.jpg	258	3	2	2	2	258   skin   mali
10	1_4_9_1	1_4_9_1_34_1012_9.jpg	260	2	2	2	2	260   skin   mali
11	1_4_10_1	1_4_10_1_34_1012_10.jp	106					106   breast   m
12	1_4_11_1	1_4_11_1_34_1012_11.jp	95	2				95   breast   ma
13	1_4_1_2	1_4_1_2_34_1012_22.jpg	552	3	3	2	2	552   breast   m
14	1_4_2_2	1_4_2_2_34_1012_21.jpg	541	3	2	3	3	541   lymph nod
15	1_4_3_2	1_4_3_2_34_1012_20.jpg	537	3	2	3	3	537   lymph nod
16	1_4_4_2	1_4_4_2_34_1012_19.jpg	242	3	2	3	3	242   spleen   m
17	1_4_5_2	1_4_5_2_34_1012_18.jpg	974	2	3	3	3	974   lymph nod
18	1_4_6_2	1_4_6_2_34_1012_17.jpg	224	3	3	3	3	224   lymph nod
19	1_4_7_2	1_4_7_2_34_1012_16.jpg	620	2	3	2	2	620   lymph nod
20	1_4_8_2	1_4_8_2_34_1012_15.jpg	540		3			540   lymph nod
21	1_4_9_2	1_4_9_2_34_1012_14.jpg	466	2	2	2	2	466   lymph nod
22	1_4_10_2	1_4_10_2_34_1012_13.jp	979	3	2	0	0	979   ent   malig
23	1_4_11_2	1_4_11_2_34_1012_12.jp	535		0			535   lymph nod

*K-M file format example. This is the other file output format of the TMA-Deconvoluter.*

- **A simple text tab-delimited format.** This would be equivalent to the PCL format, except that the UID and GWEIGHT columns and the EWEIGHT row are missing.
- value in cell L25 will be used as the name of the antibody in the output file. Within each sector, scoring data is provided. There is a limit of 253 scoring worksheets per workbook – this corresponds to the 256-column limitation in Microsoft Excel and the three columns required for the format of the pre-cluster file for the Cluster program.

## Quantitative Scoring Systems

The score conversion process is meant to convert the range of scores from the user's scoring system into a range that is compatible with TreeView. TreeView was originally designed for displaying 2-color ratiometric DNA microarray data, where one color (green) represents a green-labeled DNA probe, and the other color (red) represents a red-labeled probe. When ratios are transformed into  $\log_2$  space, the space between 0 and 1 (representing a higher abundance ratio of the green probe vs red probe) is transformed into  $-\infty$  to 0, while the space between 1 and infinity (representing a higher abundance ratio of the red probe vs green probe) is transformed into the space between 0 and infinity. The value of 0 represents the 1:1 ratio at which the abundances of the green and red probes are equal and is shown as black. This is illustrated below.



For those of you familiar with 2-color ratiometric DNA microarray data, the above explanation is a very simplified one and doesn't take into account all the possible variations done in processing the data. However, such variations aren't relevant for our purposes here, so the explanation here should suffice.

The most important aspect of this transformation is that the dynamic range for negative vs positive values is symmetric around 0 (represented by black).

Thus, when we adapted our current scoring system for TreeView visualization, we designed it as shown in the table below:

Score	Description	Treeview score	Appearance under Treeview
	Missing datapoint		Grey
0	Negative	-2	Green
1	equivocal/uninterpretable	0	Black
2	weak	1	Dark Red
3	strong	2	Red

This meant that we needed to map our scoring system into the number space used by TreeView and to make it symmetric around zero. This thus accounts for the default mappings of the score conversion process as shown in the score key above.



With the release of the TMA-Combiner, which can handle quantitative scoring systems, we wanted to extend the ability of the TMA-Deconvoluter to handle these scoring systems. However, to keep such systems compatible with TreeView, for those who wish to cluster their data, we devised a score transformation-scaling function that would achieve the objective of setting 100% negative and 100% positive scores on the two extremes of the range, with the equivocal/uninterpretable value being centered exactly in between these two ranges.

For those who use a continuous scoring system (no discrete integers), this allows one to set a “equivocal/uninterpretable” threshold that would then center the rest of the scoring range.

### The Score Conversion Process

The actual formula as implemented in the TMA-Deconvoluter is then as follows:

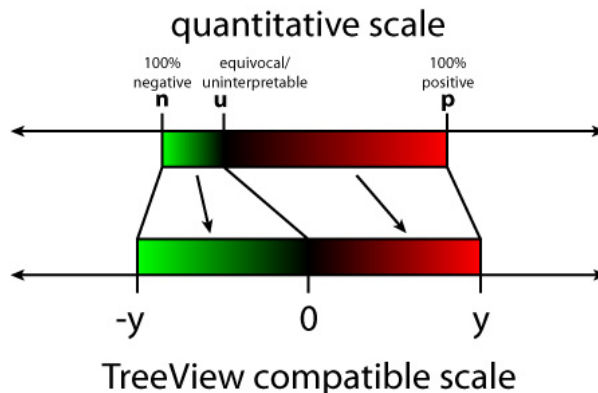
**For the following:**

- n = 100% negative, taken to be the minimum limit of the range in the user’s scoring system
- u = uninterpretable/equivocal, taken to be an intermediate value between n and p
- p = 100% positive, taken to be the maximum limit of the range in the user’s scoring system
- s = score to be scaled/transformed
- t = transformed/scaled score
- y = TreeView score range limits

```

if s < u then
    t = -y * (s - u) / (u - n)
else
    t = y * (s - u) / (p - u)
end if
    
```

The end result is a scoring system with range -y to y, with u centered at 0. This is illustrated in the diagram below.



The scaling performed is proportional and linear. You may note that the TreeView scores shown in the discrete score conversion table used with our scoring system, as shown in the table earlier in this section, is still consistent with this score conversion formula for  $y = 2$ .

This scoring conversion process should be flexible for most conceivable scoring systems currently used by pathologists. It has the following requirements:

- a minimum value that represents the highest confidence of negative staining or lowest degree of staining
- a **single** intermediate value that represents the lowest staining confidence or otherwise an uninterpretable/equivocal stain
- a maximum value that represents the highest confidence of positive staining or highest degree of staining
- scoring system is linearly scalable

There is currently no planned support for non-linear scaling.

## **Revision History**

Revision history of the TMA-Combiner.

Current version is **1.00** (3-9-05).

*Earliest supported version is **1.00**. If you are encountering problems, ensure that your current version is this version or higher.*

### **Combiner Version 1.00, 3/9/05**

- **Initial Release Version.** This is the initial release version of the TMA-Combiner.

## **Frequently Asked Questions (FAQ)**

**Q:** What exactly is the TMA-Combiner? (2-5-05)

**A:** The TMA-Combiner is a TMA dataset combination program designed to be used with the TMA-Deconvoluter. It serves the two following main functions:

1. To combine replicate cores within a body of TMA data, residing either in a single TMA or in multiple TMAs, and
2. To combine multiple TMA together into a single file for analysis.

The TMA-Combiner should be particularly useful for laboratories that have large numbers of TMA datasets and/or work with TMAs that contain large numbers of replicate cores.

**Q:** Why are there three separate score combination rules? Can't I just average my TMA replicate cores? (2-5-05)

**A:** You can (Rule 2), but averaging might not be the best way to combine your replicate data. We discuss this at length [here](#) and in our publication. In short, replicate cores are only representatives of an entire biopsy, and not all IHC antibodies stain quantitatively (which is the assumption you are making should you decide to perform simple averaging with your data). Furthermore, if you were to average, you need to remove equivocal cores from the average to avoid skewing downwards your average score. The score combination rules we provide account for these additional considerations.

**Q:** I have a scoring system that is a quantitative or semi-quantitative, continuous range system (i.e. not discrete numerals). How do we adapt the TMA-Combiner to our scoring system? (2-26-05)

**A:** The TMA-Combiner can handle continuous range scoring systems. You can find more information on how this is handled in the "Quantitative Scoring Systems" section under the Appendix, and the walkthrough has been updated accordingly.