V. 1.1 July 19, 2010. X_Dim and Y_Dim were moved to a config file to avoid a requirement for recompilation.

Fodor lab Software suite for Microarray analysis. V 1.1. December 13, 2008.

Disclaimer:

This software described herein is distributed as is. The authors take no responsibility for any use or misuse. We ask that any work benefiting from the use of this package cite these papers:

(1) Determining gene expression on a single pair of microarrays. Reid R.W. and Fodor A.A. BMC Bioinformatics 2008, 9:489

(2) Fodor AA, Tickle TL, Richardson C. Towards the uniform distribution of null p-values on microarrays. Genome Biol. 2007, 8(5):R69

If you have question, concerns, complaints or find bugs you should contact Anthony Fodor. My e-mail is anthony.fodor@gmail.com. New versions of this software (and news as to what I may or may not be up to) may from time to time be released at http://www.afodor.net.

Contents:

This document has two parts. In part 1, directions for comparing n=1 sample sized experiments using PINC are given. In part 2, scheme 4 is used to perform an analysis with at least n=3 in each condition.

Comparing two single .cel files with n=1 in each condition.

This section describes how to perform a pairwise comparison between Affymetrix .cel files with n=1 in each condition using PINC (ref. 1 above). Although PINC is written in Java (with a dependency on R) and will work across platforms, we performed quantile-quantile normalization with DCHIP, which is Windows only. If you are not using Windows and wish to normalize your CEL files prior to proceesing with PINC, you will need to find an alternative way to normalize your CEL files (try the Bioconductor package in R).

The terminology in this manual assumes one (or more) "baseline" CEL files and one (or more) "experiment" CEL files. Fold change for each gene is calculated as the ratio of the intensities of the experiment/baseline. P-values are evaluated with regard to the null hypothesis that the intensity for each gene in the baseline condition is the same at the intensity for each gene in the experiment condition.

You will need:

-Affymetrix CEL files from your experiment. (Note that this code does not work on exon junction arrays; we may add that in the future if someone asks us to!)

- DCHIP (http://biosun1.harvard.edu/complab/dchip/) or an alternative to quantile-quantile normalize your CEL files. You may of course use an alternative to quantile-quantile normalization.

-The CDF file for your Affymetrix Microarray (http://www.affymetrix.com/support/technical/libraryfilesmain.affx).

-A recent (1.5 or greater) version of the Java runtime environment (available at java.sun.com)

-R (http://www.r-project.org/) installed with Bioconductor (http://www.bioconductor.org/). Be sure to have Bioconductor installed or you will get an Exception that the results from R were not found when you run PINC.

(1) Normalize your CEL files with DCHIP.
First make a backup of any CEL data, just in case something goes horribly awry. Click analysis, open group and choose the appropriate folder that contains your CEL files. Click on the "Other Information" tab and choose the appropriate location of the CDF files on your computer (DCHIP needs these). Press OK.

To normalize, choose analysis -> Normalize & Model. Make sure "normalize" is checked. There are plenty of other DCHIP options for you to choose but those are beyond the scope of these instructions.

After normalization, choose Image -> Export CEL. Make sure to add some suffix to the name so that your original CEL files are not erased! These new CEL files are what we will be feeding into PINC.

(2) Move the normalized baseline and experiment CEL files into separate directories. PINC requires that the experiment and baseline cel files be separated into different folders. Create a folder for baseline CELs and another for Experiment CELs and add whichever cells you want to have compared. We will tell PINC where these folders are in the next step. Since PINC does 1 to 1 comparisons, every cel added to a folder will be compared to all cels in the other folder. That is to say if you add 3 cels to each folder, you will end up with 3 X 3 = 9 different comparisons. **Make sure the CEL files that you want analyzed end in .CEL. All files that do not end in .CEL will be ignored.**

(3) In the directory in which you unzipped the Java code (the same directory as this PDF document), there is a file called AffyAnalysis.properties. The contents of this file will look something like this:

```
# Required for all analyses
PINC_DIRECTORY=C:\\DirectoryWhereYouUnzippedPinc
R_DIRECTORY=C:\\Program Files\\R\\R-2.2.1\\bin
AFFY_ANALYSIS_OUTPUT_DIR=C:\\PINC_TEST\\output


#The size of the array to initialize to hold your data
#If these are set to small, you will get a
# java.lang.ArrayIndexOutOfBoundsException
#If they are set to large, you get an
#out of memory exception
#These must be set to integers or you get a cast exception
X_DIM=1165
Y_DIM=1165

# required only for PINC (low sample size comparisons at
the Probe level)
BASELINE_CELS_DIRECTORY=F:\\PINC_TEST\\BASELINE
EXPERIMENT_CELS_DIRECTORY=F:\\PINC_TEST\\EXPERIMENT
EXPECTED_NUMBER_PROBES=11
CDF_FILE=C:\\LatinSquare\\HG-U133A_tag.CDF
```

Notice the use of doubleback slashes for Windows users (Unix and OS-X users can use a single forward slash ‘/’).

The “PINC_DIRECTORY” is the directory where the code (and this pdf) were unzipped.

The "CDF_FILE" points to the CDF file associated with your CEL files. (Note that if the CDF files have x or y coordinates that are more than 1,165, you will need to set X_DIM and Y_DIM appropriately).

The "BASELINE_CELS_DIRECTORY" and "EXPERIMENT_CELS_DIRECTORY" are where you put your CEL files in the previous step.

"R_DIRECTORY" should point to the directory that holds the R executable file.

"AFFY_ANALYSIS_OUTPUT_DIR" should be an existing directory where you would like the output of the PINC analysis to be run.

"EXPECTED_NUMBER_PROBES" is the minimal number of probes that a probeset has to have to be included in the analysis. If a probeset has fewer than this number of probes, it is removed from the analysis. If it has more than this number of probes, only the first EXPECTED_NUMBER_PROBES are used. For the HG-U133A arrays, this number is set to 11 since nearly every probeset has exactly 11 probes. This number must be an integer or you will get a NumberFormatException. This parameter is required because PINC assumes that all the probesets on the array form a single normal distribution and hence will not do well with unequal sample sizes.

(4) Open up a DOS box (or a UNIX box). On Windows this can be done by clicking on the start menu and typing "cmd". Navigate to the directory where this PDF is installed (e.g. by typing "C:\\DIRECTORY_WHERE_UNZIPPED"). When in that directory type:

java –mx512m launchers.PincLauncher

Where mx is as much memory (in megabytes) as you can spare (hopefully at least 512 MB). (Remember that in 32 bit OS systems like Windows XP, Java can only use about 1.4 Gigs of memory).

See the discussion of the sample data below for a description of what is produced by this command.

<u>Running PINC on two sample CEL files from the Latin Square experiment</u>

In the "sampleData" directory, you will find two CEL files "E1_R1.cel" (in the "BASELINE" directory) and "E2_R1.cel" (in the "EXPERIMENT" directory). These represent two Latin Square experiments (12_13_02_U133A_Mer_Latin_Square_Expt1_R1.CEL and 12_13_02_U133A_Mer_Latin_Square_Expt2_R1.CEL) after normalization with DCHIP. If you wish to test your PINC installation, move E1_R1.cel to the "BASELINE_CELS_DIRECTORY" specified in your AffyAnalysis.properties and move E2_R1.cel to the "EXPERIMENT_CELS_DIRECTORY". (Make sure these are the only CEL files in those directories). Then run the PincLauncher as described above.

You will need to download the Affymetrix Latin Square CDF file from here (http://www.affymetrix.com/redirect/taf.jsp?source=dc2&dest=/Download/data/HG-U133A_tag_CDF.zip) and point the CDF_FILE in your properties file to the location of the CDF file.

If the run is successful, you should see the following in the directory specified by AFFY_ANALYSIS_OUTPUT_DIR.

`- DataForROkToDelete.txt`
This file holds the data formatted for cyberT. Each row holds 11 ratios, the ratio for each probe of the unlogged experiment value over the unlogged baseline. The last (12^{th}) column is an estimated expression for each gene used to rank the genes during the calculation of the cyber-T paired statistic (see the function bayesT.pair in the file PINC_DIRECTORY\ rScripts for more details).

`-rCommand.txt`
Contains the commands that are actually fed to R. You can open up an R shell and feed these in manually if you want to see what R does.

`-rResultsOkToDelete.txt`
Contains the results of the cyber-T algorithm. You can use this file directly in your analyses, although the gene names have been stripped and will be put back in by the Java layer.

E2_R1_VS_E1_R1__CyberTPaired.txt

	A	B	C	D	E	F	G	H	I	J	K
1	ProbeSetID_E2	isTruePositive_	verage_E2_F	Score_E2_R1	UnnormalizedF	pValue_E2_R1	expectedPVal	expectedS	unNormali	BH_FDR_E2_	BY_FDR_E2_R
2	AFFX-r2-TagA_	TRUE	11.078095	15.166518	1.6301014	0	4.50E-05	0	0	0	0
3	AFFX-LysX-3_a	TRUE	8.948884	-47.564182	0.01623785	0	9.00E-05	0	0	0	0
4	AFFX-PheX-3_a	TRUE	9.380443	-37.8344	0.01891104	0	1.35E-04	0	0	0	0
5	AFFX-r2-TagE_	TRUE	12.2379265	18.887886	1.6916716	0	1.80E-04	0	0	0	0
6	AFFX-r2-TagD_	TRUE	12.186148	13.982314	1.4731327	0	2.25E-04	0	0	0	0
7	AFFX-r2-TagC_	TRUE	12.15828	14.866628	1.5259037	0	2.70E-04	0	0	0	0
8	AFFX-ThrX-3_a	TRUE	9.450646	-47.79646	0.010555619	0	3.15E-04	0	0	0	0
9	AFFX-r2-TagB_	TRUE	10.811989	13.230654	1.7112937	1.11E-16	3.60E-04	0	0	3.08E-13	3.26E-12
10	203471_s_at	TRUE	9.932429	13.1834	1.6486976	1.11E-16	4.05E-04	0	0	2.74E-13	2.90E-12
11	AFFX-r2-TagG_	TRUE	13.026711	12.371444	1.4069917	1.33E-15	4.50E-04	0	0	2.96E-12	3.13E-11
12	AFFX-r2-TagH_	TRUE	12.984244	12.3418665	1.4023675	1.44E-15	4.95E-04	0	0	2.92E-12	3.09E-11
13	204951_at	TRUE	9.84414	12.102198	1.565316	2.78E-15	5.40E-04	0	0	5.14E-12	5.44E-11
14	AFFX-r2-TagF_	TRUE	12.3894615	11.773339	1.4247434	6.88E-15	5.85E-04	0	0	1.18E-11	1.25E-10
15	207540_s_at	TRUE	8.936034	11.356586	1.5820745	2.20E-14	6.30E-04	0	0	3.49E-11	3.69E-10
16	AFFX-DapX-3_a	TRUE	12.063095	9.643965	1.4017781	3.27E-12	6.75E-04	0	0	4.85E-09	5.14E-08
17	205267_at	TRUE	7.857175	9.506161	1.4872504	4.98E-12	7.20E-04	0	0	6.91E-09	7.32E-08
18	200665_s_at	TRUE	8.109208	8.017472	1.4154257	5.29E-10	7.65E-04	0	0	6.91E-07	7.32E-06
19	204430_s_at	TRUE	8.845409	7.753109	1.3303845	1.24E-09	8.10E-04	0	0	1.53E-06	1.62E-05
20	205790_at	TRUE	8.299161	7.0711207	1.322216	1.15E-08	8.55E-04	0	0	1.35E-05	1.42E-04
21	204417_at	TRUE	7.3787155	6.516983	1.2913531	7.16E-08	9.00E-04	0	0	7.96E-05	8.43E-04
22	209666_s_at	FALSE	6.8761783	-5.876397	0.7950773	5.98E-07	9.45E-04	0	0	6.33E-04	0.006699823
23	203310_at	FALSE	6.9475894	-5.770949	0.7878542	8.48E-07	9.90E-04	0	0	8.56E-04	0.009064982
24	201257_x_at	FALSE	12.422067	5.438279	1.1438775	2.54E-06	0.00103501	0	0	0.002454095	0.025979236
25	209142_s_at	FALSE	8.399798	-5.217	0.8085276	5.25E-06	0.001080011	0	0	0.004859747	0.051445649
26	201843_s_at	FALSE	8.573159	-5.070944	0.7849626	8.45E-06	0.001125011	0	0	0.00751369	0.079540499
27	205398_s_at	TRUE	8.30159	5.0184197	1.2129312	1.00E-05	0.001170012	0	0	0.008570379	0.090726689
28	205091_x_at	FALSE	7.7478237	-4.9959083	0.82494354	1.08E-05	0.001215012	0	0	0.008878909	0.093992803
29	208929_x_at	FALSE	11.944964	4.99534	1.1548386	1.08E-05	0.001260013	0	0	0.008577613	0.090803267
30	207968_s_at	TRUE	8.733382	4.9948206	1.320855	1.08E-05	0.001305013	0	0	0.008295806	0.08782003
31	222309_at	FALSE	6.8779774	-4.9916196	0.8069594	1.09E-05	0.001350014	0	0	0.008103038	0.085779378
32	209023_s_at	FALSE	8.016117	-4.9615064	0.7775805	1.21E-05	0.001395014	0	0	0.008646272	0.091530092
33	218381_s_at	FALSE	8.088617	-4.957934	0.8085749	1.22E-05	0.001440014	0	0	0.008473643	0.089702632
34	210360_s_at	FALSE	6.2519116	4.947371	1.1915953	1.26E-05	0.001485015	0	0	0.008503012	0.09001353
35	218049_s_at	FALSE	8.614902	-4.909383	0.83911896	1.43E-05	0.001530015	0	0	0.009332995	0.098799796
36	221607_x_at	FALSE	12.442977	4.8868966	1.1284773	1.54E-05	0.001575016	0	0	0.009750093	0.103215232
37	211073_x_at	FALSE	12.0335245	4.863966	1.1238142	1.65E-05	0.001620016	0	0	0.010208101	0.10806374
38	212661_x_at	FALSE	13.2198	4.823713	1.1274439	1.88E-05	0.001665017	0	0	0.01130944	0.119722593
39	218663_at	FALSE	7.998488	-4.8172092	0.8377902	1.92E-05	0.001710017	0	0	0.011245057	0.119041029
40	203983_at	FALSE	7.6688538	-4.8163557	0.83092684	1.93E-05	0.001755018	0	0	0.010986893	0.116308088
41	201624_at	FALSE	7.3692594	-4.786782	0.8236693	2.12E-05	0.001800018	0	0	0.011782402	0.124729406
42	209799_at	FALSE	7.425412	-4.7506666	0.82683057	2.38E-05	0.001845018	0	0	0.012910222	0.136668597

Fig. 1. The results of they Cyber-T paired algorithm (the top part of the file "E1_R1_VS_E2_R1__CyberTPaired.txt").

Within the directory "E1_R1_VS_E2_R1" you will see two files that hold the results of the PINC algorithm. The first of these, "E1_R1_VS_E2_R1__CyberTPaired.txt" holds the results from the CyberTPaired algorithm (i.e. the p-values from PINC have not yet been applied). The following columns are defined in this file (see Fig. 1):

`isTruePositive` – Is true if the ProbeId is a spiked-in probe from the Latin Square U133A set. (If you want to change this to reflect some other dataset you will need to change the method dataSetDescribers.PincProbe.getSpikeDetails() and recompile. Otherwise, just ignore this column).

`average` – This is the average value of the log expression for all 22 probes.

`score` – This is the paired cyberT score.

`approximateFoldChange` – This is the fold change. Kind of. It is the average of all the log scores of the experiment minus the average of the log scores of the controls. That is it is:

$$2^{\sum \log(E)/n - \sum \log(B)/n}$$

where E and B are the 11 values for the Expression and Baseline and n is the number of probes (11). This will be approximately equal to the usual meaning of fold change for two conditions with equal sample size:

$$\sum E \Big/ \sum B$$

If anyone needs the software to do the real (unlogged) fold change, let me know (anthony.fodor@gmail.com) and I will fix it.

`pValue` – This is the pValue from the paired cyberT test.

`expectedPValue` – This is the expected pValue if all the genes were the same on both chips (i.e. if for every gene the null hypothesis of no differential expression were true).

`BH_FDR` – Benjamini and Hochberg false discovery rate (see methods sections in our papers). The values in this column reflect the lowest threshold that would include the gene in this row. For example, a BH - FDR cutoff of 10% would include 249 genes in this file reflecting our argument that our pipeline using the paired cyber-T test produces p-values that are too small.

`BY_FDR` – The more conservative Benjamini and Yekutieli false discovery rate (see methods sections in our papers). A BY –FDR cutoff of 10% includes 35 genes.

The other file in the E2_R1_VS_E1_R1 directory ("E2_R1_VS_E1_R1_PINC_CyberTPaired.txt") has the same format but includes the PINC procedure. The PINC procedure involves two post-processing steps after Cyber-T paired. The first is "Statistics level normalization". Fig. 2 shows that the paired Cyber-T score (the `unnormalizedScore` column in the file) has small drifts from the expected score of zero as a function of the average probe expression (the `average` column in the file). To correct for this, we fit a Loess regression line (red line Fig. 2; the `expectedScore` column in the file) and subtract the value of that regression line from each Cyber-T score (see the methods sections in our papers). The `score` column in "E2_R1_VS_E1_R1_PINC_CyberTPaired.txt" reflects this new score. Next, we assume that all paired Cyber-T scores corrected in this way form a single normal background distribution and assign p-values for each gene based on this distribution. The `pValue` column in "E2_R1_VS_E1_R1_PINC_CyberTPaired.txt" is this corrected pValue. We see that with these new p values, both BH and BY FDR nearly perfectly find the correct cutoff for the Latin Square file (almost all true positives above the cutoff), which is the point of our PINC paper.

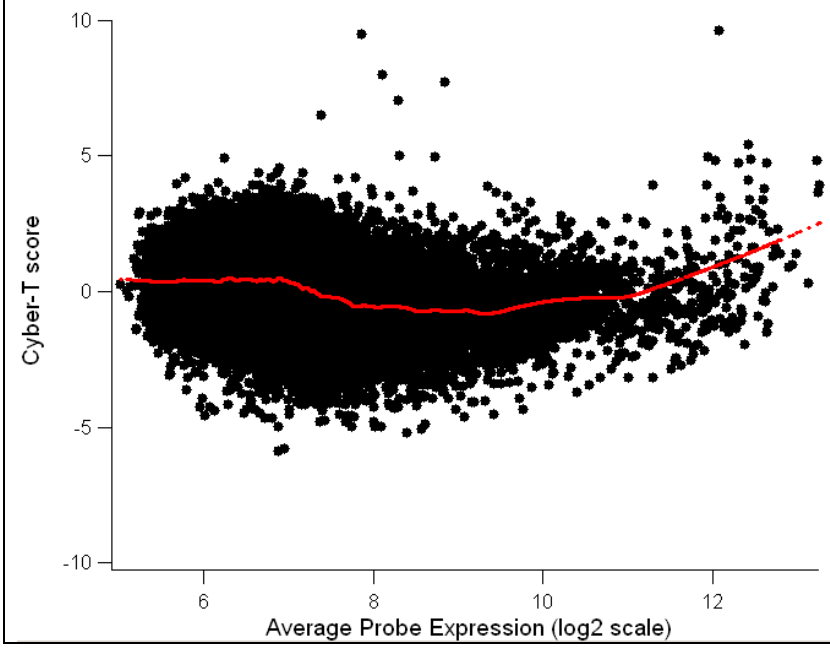

Fig. 2. The paired Cyber-T score as a function of the average

Comparing two conditions with at least n=3 in each condition (Scheme 4).

The terminology in this manual assumes one (or more) "baseline" CEL files and one (or more) "experiment" CEL files. Fold change for each gene is calculated as the ratio of the intensities of the experiment/baseline. P-values are evaluated with regard to the null hypothesis that the intensity for each gene in the baseline condition is the same at the intensity for each gene in the experiment condition.

You will need:

-A spreadsheet with your data in the form….

SomeLabel	Experiment1	Experiment2	Experiment3	More Experiments→
Gene1				
Gene2				
Gene3				
Gene4				
More Genes…				

The data should be normalized and log-transformed (our software assumes a base-2 log transform). The easiest way to obtain this spreadsheet from your CEL files is to use a program such as RMA Express (http://rmaexpress.bmbolstad.com/) or the equivalent RMA algorithm in the Bioconductor package in R. An example of a quantile-quantile normalized data file in this format for the 42 experiments of the U133A LatinSquare data set is in the "SampleData" directory of this distribution with the filename ("RMA_LATIN_SQUARE_RESULTS.txt").

-A recent (1.5 or greater) version of the Java runtime environment (available at java.sun.com)

-R (http://www.r-project.org/) installed with Bioconductor (http://www.bioconductor.org/).

1) In the directory in which you unzipped the Java code (the same directory as this PDF document), there is a file called AffyAnalysis.properties. The contents of this file will look something like this:

```
# Required for all analyses
PINC_DIRECTORY=C:\\someDirectoryWhereYouUnzippedPinc
R_DIRECTORY=C:\\Program Files\\R\\R-2.2.1\\bin
AFFY_ANALYSIS_OUTPUT_DIR=C:\\PINC_TEST\\output
```

```
#required only for Scheme 4 (comparisons at the ProbeSet
level)
BASELINE_CHIP_NAMES=12_13_02_U133A_Mer_Latin_Square_Expt1_R
1.CEL,12_13_02_U133A_Mer_Latin_Square_Expt1_R2.CEL,12_13_02
_U133A_Mer_Latin_Square_Expt1_R3.CEL

EXPERIMENT_CHIP_NAMES=12_13_02_U133A_Mer_Latin_Square_Expt2
_R1.CEL,12_13_02_U133A_Mer_Latin_Square_Expt2_R2.CEL,12_13_
02_U133A_Mer_Latin_Square_Expt2_R3.CEL

PATH_TO_LOG_2_DATA=C:\\LatinSquare\\RMA_LATIN_SQUARE_RESULT
S.txt
```

Notice the use of doubleback spaces for Windows users (Unix and OS-X users can use a single forward slash '/').

The "PINC_DIRECTORY" is the directory where the code (and this pdf) were unzipped.

"R_DIRECTORY" should point to the directory that holds the R executable file.

"AFFY_ANALYSIS_OUTPUT_DIR" should be an existing directory where you would like the output of the Scheme 4 analysis to be run.

"PATH_TO_LOG_2_DATA" – The full path to the spreadsheet that has your expression data (described on the previous page)

"BASELINE_CHIP_NAMES" – A comma separated list of the headings in your spreadsheet that correspond to the chips in your baseline set.

"EXPERIMENT_CHIP_NAMES" – A comma separated list of the headings in your spreadsheet that correspond to the chips in your experiment set.

(4) Open up a DOS box (or a UNIX box). On Windows this can be done by clicking on the start menu and typing "cmd". Navigate to the directory where this PDF is installed (e.g. by typing "C:\\DIRECTORY_WHERE_UNZIPPED"). When in that directory type:

java –mx512m launchers.Scheme4Launcher

Where mx is as much memory (in megabytes) as you can spare (hopefully at least 512MB). (Remember that in 32 bit OS systems like Windows XP, Java can only use about 1.4 Gigs of memory).

See the discussion of the sample data below for a description of what is produced by this command.

<u>Running Scheme 4 comparing Experiments 1 and 2 (both n=3) of the Latin Square data set</u>

In the SampleData directory as part of this distribution, there is a file called "RMA_LATIN_SQUARE_RESULTS.txt". With the PATH_TO_LOG_2_DATA in the config file pointing to this file, running launchers.Scheme4Launcher will produce the following results files to AFFY_ANALYSIS_OUTPUT_DIR.

-DataForROkToDelete.txt
-rCommand.txt
-rResultsOkToDelete.txt

(See the discussion of PINC above for the contents of these files. Scheme 4 uses CyberT rather than CyberT Paired, but otherwise the format of these files is identical between Scheme 4 and PINC).

In the folder "RMA_LATIN_SQUARE_RESULTS" in your AFFY_ANALYSIS_OUTPUT_DIR you will see the files

-RMA_LATIN_SQUARE_RESULTS_CyberT.txt
-RMA_LATIN_SQUARE_RESULTS_Scheme4_CyberT.txt.

File "RMA_LATIN_SQUARE_RESULTS_CyberT.txt" holds the results of the CyberT algorithm and the file "RMA_LATIN_SQUARE_RESULTS_Scheme4_CyberT.txt" holds the results of scheme 4, which applies statistical level normalization and then calculates p values based on the assumption of a single normal distribution. See the discussion of PINC above for details on the format of these files. PINC and Scheme 4 generate identical results except that PINC applies its transformation to the cyber-T paired test at the probe level while scheme 4 applies its transformation of p-values to the cyber-T test applied at the probeset level.