

# DR-Integrator User Manual

## V1.0 (November 2009)

### Table of Contents

#### **1. Installation**

- 1.1 Installing R
- 1.2 Installing R Packages & R(D)COM Server
- 1.3 Installing the Excel Plug-In

#### **2. Data formats**

- 2.1 Inputs
- 2.2 Outputs

#### **3. Usage**

- 3.1 Guidelines
- 3.2 Preprocessing
- 3.3 DR-Correlate
- 3.4 DR-SAM

## 1. Installation

### 1.1 Installing R

DR-Integrator works as a plug-in to Microsoft Excel (Windows XP/Vista version), and relies on a code base written in the statistical programming language R. Prior to installing DR-Integrator, download and install the latest version of R, which is freely available at <http://www.r-project.org/>. Be sure to check the box to “Save the R version in the registry” during installation.

NOTE: If an older version of R or the R(D)COM Server is installed, it is recommended to uninstall the older version before installing the newest versions.

### 1.2 Installing R Packages & R(D)COM Server

After installing R, double-click the DRI-Installer.msi installation file (available on the Pollack Lab website, <http://pollacklab.stanford.edu/>) to begin setup. Note the location to which the DR-Integrator program files are saved. Three R packages must be installed to run DR-Integrator: rscproxy, impute, and DRI. The rscproxy and impute packages can be installed by opening the R interpreter, and from the Packages menu, click “Install package(s).” Select a nearby mirror site, and then select the package ‘rscproxy’ from the list, and click OK; repeat for the impute package. To install the DRI package, click “Install package(s) from local zip files” from the Packages menu. Navigate to the folder in which the DR-Integrator program files were saved, and in the R folder, double-click the DRI zip file. The R interpreter can be closed at this point.

In order for Microsoft Excel to have access to the R interpreter, the R (D)COM Server must also be installed, which can be downloaded from <http://cran.cnr.Berkeley.edu/other-software.html> (or the equivalent page from any of the CRAN mirror sites). Double-click the installation file and follow the default instructions to install. You can check that the R (D)COM Server is working by running the “Server 01 – Basic Test” from the Start Menu (Programs → R → (D)COM Server → Server 01 – Basic Test). No error should be generated when ‘Start R’ button is clicked.

### 1.3 Installing the Excel Plug-In

For Excel 2003 and earlier:

Open Microsoft Excel to load the DR-Integrator plug-in. From the Tools menu, select Add-Ins and then click the Browse button. Navigate to the folder in which the DR-Integrator program files were saved, and in the Add-In folder, double-click the DR-Integrator.xla file. Make sure the box next to “DR-Integrator” is checked under the list of Add-Ins Available, and click OK.

For Excel 2007:

Open Microsoft Excel to load the DR-Integrator plug-in. Click the Office Button (top left corner) and select Excel Options. From the Options menu, select Add-Ins. On the bottom of the Add-Ins options, select Excel Add-Ins from the Manage drop-down menu and click Go. Click the Browse button to navigate to the folder in which the DR-Integrator program files were saved. In the Add-In folder, double-click the DR-Integrator.xla file. Make sure the box next to “DR-Integrator” is checked under the list of Add-Ins Available, and click OK.

## 2. Data formats

### 2.1 Inputs

The input file data format for DR-Integrator is a tab-delimited text file with 4 columns of annotations (Gene ID, Gene Name, Chromosome, Nucleotide Position) for the copy number arrays, followed by the DNA copy number data and then 4 columns of annotations for the expression arrays followed by the gene expression data. To perform DR-Integrator analyses, the DNA copy number data should not contain any missing values. As such, it is recommended to perform a smoothing step (e.g. moving window average), or call gains/losses with a suitable algorithm (e.g. Fused Lasso [1], CBS [2], etc), before analysis. DR-Integrator can perform these preprocessing steps for you if desired (see section 3.2 of this manual for more details). Missing values in the gene expression data are imputed by the k-nearest neighbors method. Finally, every gene should have both DNA copy number values and gene expression values for each sample. If not done prior to analysis, DR-Integrator can also perform a mapping between the copy number and gene expression probes to assign every gene a copy number and expression value (see section 3.2 of this manual for more details).

**IMPORTANT:** The ordering of samples (columns) of copy number data should be the same as the ordering of samples (columns) of gene expression data.

	A	B	C	D	E	F	G	H	I	J
1	CLID	NAME	Chromosome	Nuc Position	CACO2	COLO205	COLO320	COLO741	DLD1	HCA7
2					1	1	1	1	2	2
3	IMAGE:322807	ESTs	1	736116	0	-0.30554	0	0	-0.27301	0
4	IMAGE:1659132	LOC64383	1	756847	0	-0.30554	0	0	-0.27301	0
5	IMAGE:128826	SAMD11	1	868801	0	-0.30554	0	0	-0.27301	0
6	IMAGE:366353	NOC2L	1	869458	0	-0.30554	0	0	-0.27301	0
7	IMAGE:190915	KLHL17	1	885829	0	-0.30554	0	0	-0.27301	0
8	IMAGE:742132	ISG15	1	938807	0	-0.30554	0	0	-0.27301	0
9	IMAGE:810801	AGRN	1	945365	0	-0.30554	0	0	-0.27301	0
10	IMAGE:741879	B3GALT6	1	1159733	0	-0.30554	0	0	-0.27301	0
11	IMAGE:2021882	SCNN1D	1	1205830	0	-0.30554	0	0	-0.27301	0
12	IMAGE:152617	CENTB5	1	1225261	0	-0.30554	0	0	-0.27301	0
13	IMAGE:1612722	CPSF3L	1	1236844	0	-0.30554	0	0	-0.27301	0
14	IMAGE:878605	LOC72788	1	1277933	0	0	0	0	-0.27301	0
15	IMAGE:506623	CCNL2	1	1317536	0	0	0	0	0	0
16	IMAGE:505344	LOC14841	1	1325048	0	0	0	0	0	0
17	IMAGE:1925973	VWA1	1	1364725	0	0	0	0	0	0

For DR-SAM, specify (with 1's and 2's) the class membership of each sample in the row under the sample names. Class labels are not required to run DR-Correlate. A sample file "DRI-TestFile.xls" is provided for reference.

### 2.2 Outputs

DR-Integrator output is displayed in a new worksheet called DRI-Report within the working Excel file. The output is separated into positive and negative correlations (DR-Correlate) or

differences in DNA/RNA (DR-SAM). The genes are ranked by their significance, and the score and estimated false discovery rate are provided for each gene.

	A	B	C	D	E	F
1	<b>Positive Correlations</b>					
2	<b>Rank</b>	<b>Row #</b>	<b>Gene/Clone ID</b>	<b>Gene/Clone Name</b>	<b>Correlation Score</b>	<b>FDR</b>
3	1	9529	IMAGE:1473298	FGFR1OP2	0.746	0
4	2	10390	IMAGE:1688562	C13orf24	0.740	0.014045
5	3	10704	IMAGE:43295	NAT12	0.737	0.013758
6	4	12861	IMAGE:711450	THOC1	0.735	0.013483
7	5	10274	IMAGE:1669223	C13orf23	0.735	0.013219
8	6	10450	IMAGE:753917	TM9SF2	0.735	0.012964
9	7	10184	IMAGE:447916	PSPC1	0.734	0.01272
10	8	6829	IMAGE:788745	DCTN6	0.734	0.012484
11	9	8489	IMAGE:788203	KIAA0157	0.734	0.012257
12	10	9143	IMAGE:308163	YAP1	0.734	0.012038
13	11	10427	IMAGE:428486	DNAJC3	0.730	0.011827
14	12	6749	IMAGE:1690862	VPS37A	0.729	0.011623
15	13	5386	IMAGE:841022	BTBD9	0.729	0.011426
16	14	7608	IMAGE:502062	STX17	0.729	0.011236
17	15	14537	IMAGE:33616	CABIN1	0.727	0.011052
18	16	10237	IMAGE:23121	POMP	0.726	0.010873

Additionally, an optional heatmap representation of significant DR-Correlate genes can be generated.

### 3. Usage

#### 3.1 Guidelines

The integration of copy number and gene expression data raises several notable issues that should be considered when performing an integrative analysis. The first is that while several bona fide oncogenes and tumor suppressor genes exhibit significant correlations in their copy number and expression profiles across a sample set, this is not always the case, as copy number alteration is not the only mechanism by which cancer genes acquire abnormal expression patterns. Another issue is that some cancer genes exhibit copy number alterations and resultant expression changes at relatively small frequencies and thus such genes may not achieve statistical significance when using a Pearson's or Spearman's correlation. As such, we recommend exploring a dataset with several different statistical metrics (e.g., the extremes t-test of DR-Correlate is more sensitive to "outlier" alterations occurring at low frequency when used with a small percentile cutoff). Importantly, the statistical techniques implemented in DR-Integrator are designed to detect genes whose expression patterns can be explained in large part by gene copy number (and not other mechanisms). Thus, DR-Integrator implements a focused (and not comprehensive) set of analysis tools that represent one of several possible types of analyses that one can perform when paired copy number and expression data are available.

#### 3.2 Preprocessing

Beyond the standard normalization of copy number and expression data, additional preprocessing is required before a DR-Integrator analysis is carried out. The DNA/RNA dataset must have no missing values before either DR-Correlate or DR-SAM analysis. Missing values in the expression data set are automatically imputed by k-nearest neighbors; thus, we recommend filtering out poorly-measured genes before analysis to prevent inaccurate imputation. To eliminate missing values in the copy number data, and to reduce noise, we recommend performing some form of smoothing or segmentation on the copy number data. The option to smooth the data by averaging over a moving window (size is user-defined; default=5) along the chromosomes is provided in the dialog box of either DR-Integrator analysis. There is also the option to run Fused Lasso on the copy number data to call gains and losses (with a user-defined false discovery rate) to segment the data before subsequent analysis is performed (see ref. 1 for more details on the Fused Lasso method). The user can also segment the copy number data externally with an algorithm of his/her choice and use the segmented data directly in a DR-Integrator analysis. An error message will appear if the copy number data has any missing values after preprocessing.

The final preprocessing step is merging the DNA copy number data with the gene expression data. Each gene in a DR-Integrator analysis should have paired measurements (i.e., copy number values and expression values). While some microarray platforms can be used for both copy number and expression profiling, in most cases different platforms are employed for copy number and gene expression analysis that contain distinct probes interrogating a non-identical set of genes. If not done prior to analysis, DR-Integrator can perform a mapping between the copy number probes and gene expression probes. The merging procedure assigns for each gene with an expression value a copy number value by averaging the values of the 5' and 3' probes nearest the genomic coordinate of the expression probe. If a copy number probe interrogates the same genomic coordinate as the expression probe, the copy number probe data will be applied directly (no averaging is done). The result of this procedure is a new copy number data matrix with the same number of rows (genes) as the gene expression data matrix.

The preprocessed data can be outputted to a new worksheet so that subsequent analyses can avoid repeating the preprocessing steps.

### *3.3 DR-Correlate*

To begin an analysis, open the input data file in Excel and click & drag to select all the cells containing gene annotations and the copy number/gene expression data. Click the DR-Integrator button on the toolbar to begin an analysis (if you do not see the DR-Integrator button, you may need to add the plug-in again – see instructions under the installation section above).

DR-Correlate performs an analysis to identify all genes with statistically significant correlations between their DNA copy number and gene expression levels. Three options for the statistic to measure correlation are implemented:

- (1) Pearson's correlation
- (2) Spearman's rank correlation
- (3) an "extremes" *t*-test

For Pearson's and Spearman's correlations, the respective correlation coefficient is computed for each gene. For the extremes *t*-test, a modified Student's *t*-test is computed for each gene, comparing gene expression levels of samples comprising the lowest and the highest quantiles with respect to DNA copy number. In other words, for each gene the samples are rank-ordered by DNA copy number and samples below the lowest quantile and above the highest quantile form two groups whose means of gene expression are compared with a modified *t*-test. The percentile cutoff defining the two extreme quantile groups is user-adjustable.

To run DR-Correlate, click on the DR-Correlate tab of the DR-Integrator dialog box. Fill in the fields with the row/column numbers corresponding to the listed annotations and data inputs. A false discovery rate (FDR) is estimated from random permutations of the data. Specify the number of permutations to perform and the FDR cutoff to apply, and click OK to run.

### 3.4 DR-SAM

To begin an analysis, open the input data file in Excel and click & drag to select all the cells containing gene annotations and the copy number/gene expression data. Click the DR-Integrator button on the toolbar to begin an analysis (if you do not see the DR-Integrator button, you may need to add the plug-in again – see instructions under the installation section above).

DR-SAM performs a supervised analysis to identify genes with statistically significant differences in both DNA copy number and gene expression between different classes (e.g., tumor subtype-A vs. tumor subtype-B). The goal of this analysis is to identify genetic differences (copy number alterations) that underlie gene expression differences between two groups of interest. DR-SAM implements a modified Student's *t*-test to generate for each gene two *t*-scores

assessing differences in DNA copy number and differences in gene expression. A final score is computed by summing the copy number  $t$ -score and gene expression  $t$ -score, and weighting the sum to favor genes with strong differences in both DNA copy number and gene-expression between the two classes.

To run DR-SAM, click on the DR-SAM tab of the DR-Integrator dialog box. Fill in the fields with the row/column numbers corresponding to the listed annotations and data inputs. A false discovery rate (FDR) is estimated from random permutations of the data. Specify the number of permutations to perform and the FDR cutoff to apply, and click OK to run.

## References

1. Tibshirani, R. and Wang, P. (2008) Spatial smoothing and hot spot detection for CGH data using the fused lasso, *Biostatistics*, **9**, 18-29.
2. Olshen, A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data, *Biostatistics*, **5**, 557-572.