MetageneProjection Documentation

Description: Project one or more data sets onto the metagene representation of another data set, using the metagene projection methodology described by Tamayo et al (2007)

Author: Pablo Tamayo (Broad Institute) tamayo@broadinstitute.org with contributions from Marc-Danie Nazaire, gp-help@broadinstitute.org

Summary

A major challenge in the interpretation of DNA microarray data is the noise inherent in analyzing tens of thousands of genes across a small number (tens or hundreds) of samples. One way to address this challenge is to describe all of the genes in terms of a small number of *metagenes*, where the expression patterns of the metagenes characterize the major, invariant expression patterns in the data. Analyzing the smaller set of metagenes reduces noise and emphasizes relevant biological correlations. Metagene projection allows for data sets to be compared in this lower-dimension metagene space. The greater robustness to noise in this space enables cross-platform and cross-species analysis.

This module implements the metagene projection methodology described by Tamayo et al (2007). The methodology identifies metagenes using nonnegative matrix factorization (NMF), as described by Brunet et al (2004).

References

1. Tamayo P, Scanfeld D, Ebert BL, Gillette MA, Roberts CWN, Mesirov JP. Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *PNAS*. 2007;104:5959-5964. http://www.pnas.org/cgi/content/abstract/0701068104v1
2. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *PNAS.* 2004;101:4164-4169. http://www.pnas.org/cgi/content/abstract/101/12/4164
3. Rifkin R, Mukherjee S, Tamayo P, Ramaswamy S, Yeang C-H, Angelo M, Reich M, Poggio T, Lander ES, Golub TR, Mesirov JP. An Analytical Method for Multiclass Molecular Cancer Classification. *SIAM Review*. 2003;45:706-723. http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=68
4. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*. 2005;102:15545-15550. http://www.pnas.org/content/102/43/15545.abstract

Model and Test Data Sets

The MetageneProjection module accepts a model data set and, optionally, one or more test data sets. Each data set consists of a gene expression (.gct) data file and a phenotype class (.cls) file. See *GenePattern File Formats* for full descriptions of the .gct and .cls file formats.

The gene expression data files must use a common set of gene names or probe identifiers. This allows the analysis to match the genes in the test data to the corresponding genes in the model data. For example, a model data file produced using the HG_U133A chip uses one set of probe identifiers and a test data file produced using the HG_U95Av2 chip uses a different set of probe identifiers. To analyze the two data files, translate the probe identifiers in both to a common set of gene names or translate the probe identifiers in one to the probe identifiers of the other. Only gene names or probe identifiers that exist in both the model and test data sets will be used in this analysis.

Analysis

Before using this module to apply the metagene projection method, read Tamayo et al (2007) for a full description of the method and three examples of its use.

In brief, the MetageneProjection module does the following (figures are based on the first example in the paper):

1. Uses NMF to *project* gene expression data from a model data set onto a small number of metagenes. Two output files describe the projected model data:
    - *.model.W.gct is a gene-by-metagene matrix that describes how much each gene contributes to each of the metagenes.
    - *.model.H.gct is a metagene-by-sample matrix that represents the metagene expression levels for each sample.
2. Matches the genes in each test data set to the corresponding genes in the model data set and projects the gene expression data from the test data sets onto the metagenes. An output data set contains the projected model and test data:
    - *.all.H.gct is a metagene-by-sample matrix that represents the metagene expression levels for all samples in all model and test data sets.
    - *.all.H.cls is a matching class file created by merging the .cls files for all model and test data sets.

3. Analyzes the projected data sets and generates the following analysis results:
    - Heat map of the projected data sets (*.output.heatmap.sorted.jpeg, Figure 1). The heat map shows the metagene expression levels for each sample. In this example, the metagenes clearly correspond to the same leukemia subtypes in the model (ALL-B, ALL-T, AML) and test (ALL-B2, ALL-T2, AML2) data.
    - 2D biplot of the projected data sets (*.output.2D.proj.jpeg, Figure 2). This is a plot of the model and test metagene expression levels in the output file *output.model.all.H.gct. Each axis corresponds to a metagene. In this example, the plot clearly shows a correlation between metagene expression levels and leukemia subtypes in both the model and test data sets.
    - Hierarchical tree based on clustering of the projected data sets (*.htree.jpeg, Figure 3). In this example, the tree shows that the model and test data cluster by leukemia subtype. For comparison purposes, the module also generates a hierarchical tree based on the original (pre-projected) data sets.
    - Classification prediction results based on applying the support vector machines (SVM) algorithm[3] to the projected data sets (*.pred.gct, Figure 4). In this example, a heat map of the prediction results in *.all.H.gct indicates that the metagenes are reliable predictors of the leukemia subtypes.

      Note: The MetageneProjection module writes the SVM prediction results to a .gct file. To view the results as a heat map, use the HeatMapViewer module.

After reviewing the analysis results, consider the following options for further research:

- Run additional analyses on the data set that contains the projected model and test data (step 2 above).
- For metagenes of interest, use the W matrix (step 1 above) to focus on the genes that comprise the metagene. For example, use the Molecular Signatures Database (MSigDB, http://www.broadinstitute.org/gsea/msigdb/) page to annotate the set of genes that comprise, or show enrichment in, the metagene (Subramanian et al, 2005).

GenePattern

Figure 1. Heat map

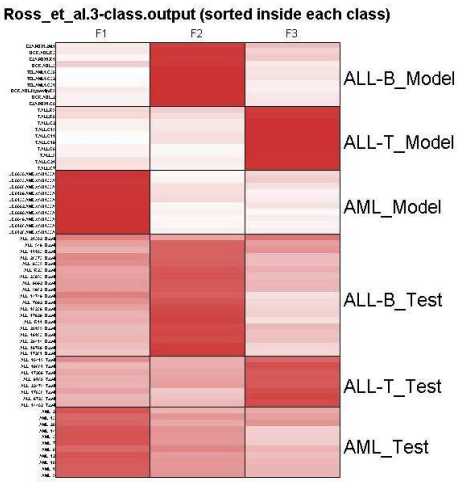


Figure 2. Biplot

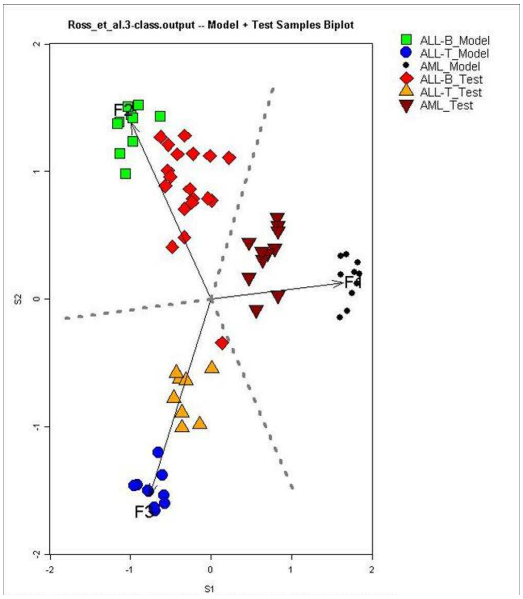


Figure 3. Hierarchical tree

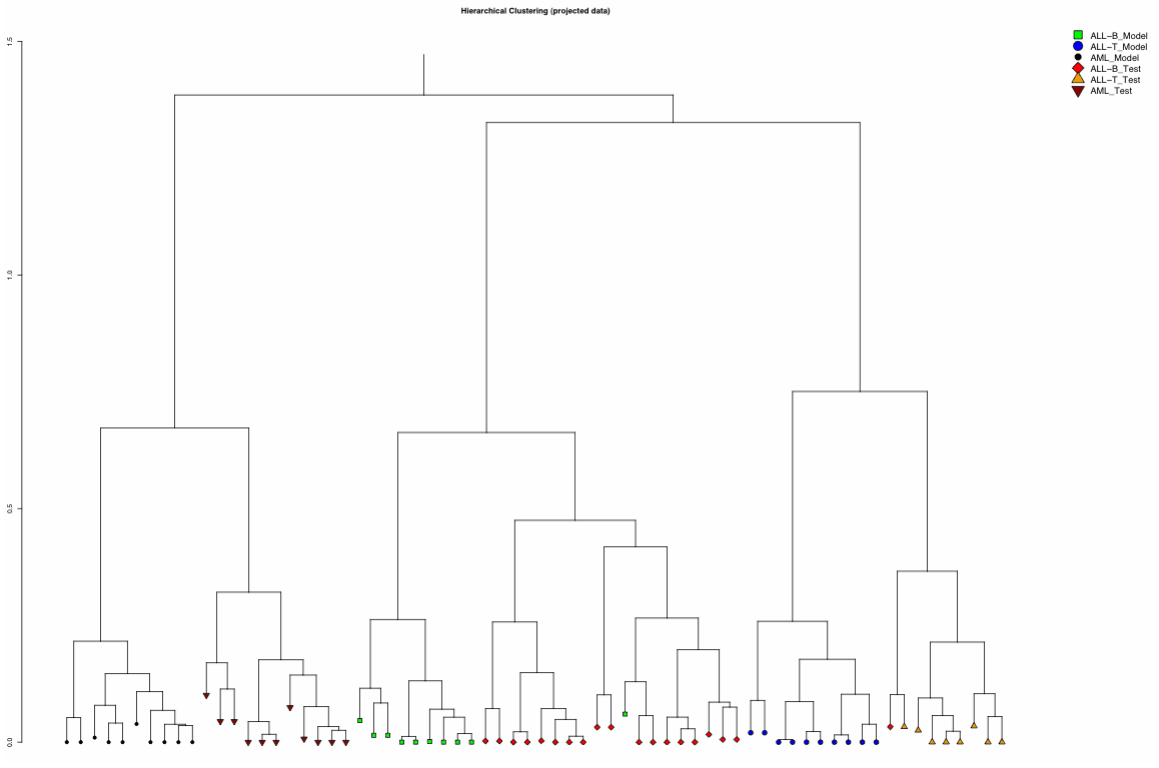
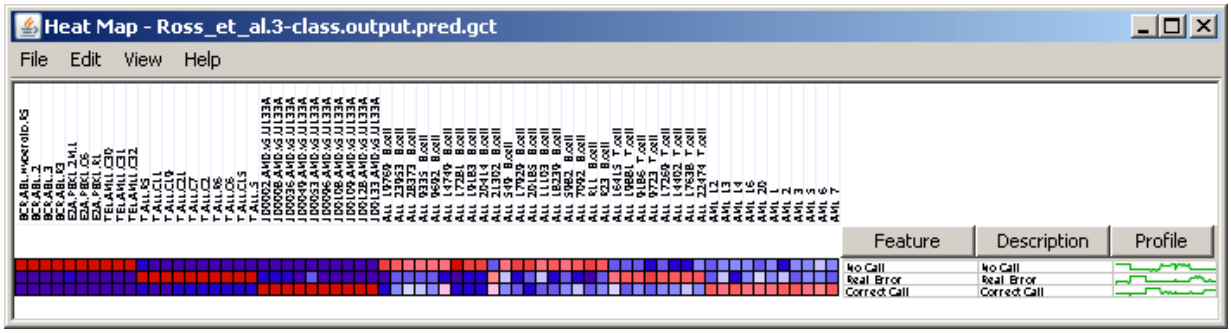


Figure 4. SVM prediction results



Parameters

Name	Description
model gct file	The model .gct data file
model cls file	The model .cls data file
model preprocessing file	The preprocessing parameters for the model data set in a .txt file, as described in Preprocessing parameter file Example: gct.file = "all_aml_train.gct" cls.file = "all_aml_train.cls" column.subset = "ALL" column.sel.type = "samples" thres = 20 ceil = 100000 fold = 5 delta = 500 norm = 6
test gct file	The test .gct or zip of .gct data files
test cls file	The test .cls or zip of .cls data files
test preprocessing file	A .txt file that contains a set of preprocessing parameters for each test data set, as described in Preprocessing parameter file
num characters	Number of characters to use for phenotype labels during classification prediction (default: 9) For example, if the *model cls file* contains the three class labels ALL-B, ALL-T and AML and *num characters* = 3, classification prediction is based on two classes: ALL and AML. If *num characters* = 9, classification prediction is based on all three classes. See the *.pred.txt output file for complete classification prediction results.
identifier	Prefix to prepend to all output file names

k projection	Number of metagenes in projection (default: 3) If you are unfamiliar with the model data set, use the NMFConsensus module to compare alternate values of *k* Brunet et al (2004).
algorithm	Algorithm for Metagene Projection: 1) NMF with divergence (default): Non-Negative Matrix Factorization using the divergence cost 2) NMF: Non-Negative Matrix Factorization using the Euclidean cost 3) NSNMF with divergence: Non-smooth NMF (Carmona-Saez P, et al. *BMC Bioinformatics*. 2006;7:78.) 4) SNMF: Sparse NMF decomposition (Gao Y, Church G. *Bioinformatics.* 2005;21:3970-3975. Adapted from an original C++ program kindly provided by Yuan Gao) 5) PCA: Principal Components via SVD Use the default value. Expert users interested in the alternate algorithms should examine the MetageneProjection code.
number of iterations	Number of algorithm iterations (default: 2000)
seed	Random seed to initialize metagene matrices (default:1234) **(Note: results may only be reproducible on the same platform)**
post projection normalization	Whether to normalize (i.e., scale points to unit hypersphere) the projected data sets (default: yes) Normalization can amplify weak signals but can also amplify noise. Use normalization if you expect weak but known phenotypes; for example, when analyzing the same tissue from another species. Do not apply normalization if you are analyzing unknown data.
heatmap row norm	Whether to row-normalize (standardize) the rows in the heatmap (default: no)
heatmap color scheme	Color scheme options for heatmap: • reddish color map (default) • vintage pinkogram • scale of grays • high resolution pinkogram
confidence threshold	Confidence threshold (Brier score) to separate calls from no-calls (default: 0.3)

phenotype plotting colors	Text file containing plot color mapping for the samples. The file contains one color per line. Samples are assigned colors in the order they appear in the model and then test data sets. The possible color options are listed here.
phenotype plotting symbols	Text file containing plot symbol mapping for the samples. The file contains one symbol per line. Samples are assigned symbols in the order they appear in the model and then test data sets. The symbol options are: • square (■) • circle (●) • diamond (♦) • triangle (▲) • reverse_triangle (▼)
symbol scaling	Graphical scaling for symbols in plots and plot legends (default: 1) For a large data set, select a smaller value such as 0.8 to make the symbols and fonts slightly smaller. For a small data set, select a larger value such as 1.2 to increase the size of the symbols and fonts.
kernel	Kernel function for SVM: "radial" or "linear" (default: "radial") Used for *model set refinement*. Expert users interested in modifying SVM settings should first examine the MetageneProjection code to see how this parameter is used.
cost	Cost parameter for SVM (default: 1) Used for *model set refinement*. Expert users interested in modifying SVM settings should first examine the MetageneProjection code to see how this parameter is used.
gamma	Gamma coefficient for radial base function kernel for SVM (default: 0.05) Used for *model set refinement*. Expert users interested in modifying SVM settings should first examine the MetageneProjection code to see how this parameter is used.
theta	Smoothing parameter for the NSNMF with divergence algorithm (default 0) Use the default value. Expert users interested in the alternate algorithms should first examine the MetageneProjection code to see how this parameter is used.

lambda	Sparse parameter used for the SNMF algorithm (default 0) Use the default value. Expert users interested in the alternate algorithms should first examine the MetageneProjection code to see how this parameter is used.
model set refinement	Whether to use support vector machine (SVM) classification to trim outliers after projecting the model data set into metagene space (default: yes) Outliers may provide biologically relevant information. If you are analyzing unfamiliar data, run the analysis twice: first trimming outliers and then leaving them. Compare the results to examine the impact (if any) of the outliers.

Preprocessing Parameter File

This is a text (.txt) file that defines the preprocessing parameters for one or more data sets. The parameters file specified for the *model preprocessing file* parameter defines preprocessing parameters for the model data set. The parameters file specified for the *test preprocessing file* parameter defines a set of preprocessing parameters for each test data set. In general, you should specify the same preprocessing parameters for all data sets.

Example parameters file for two test sets:

```
gct.file="Valk_et_al.gct"
cls.file="Valk_et_al.cls"
column.subset="ALL"
column.sel.type="samples"
thres=20
ceil=100000
fold=5
delta=500
norm=6
gct.file="Ross_et_al.gct"
cls.file="Ross_et_al.cls"
column.subset="ALL"
column.sel.type="samples"
thres=20
ceil=100000
fold=5
delta=500
norm=6
```

The parameters in the parameters file are:

gct.file	Name of the .gct file
cls.file	Name of the corresponding .cls file
column.sel.type	Select a subset of the data set. Use this option to choose what to base the selection on ("samples" or "phenotypes"). Use the *column.subset* parameter to select which samples or phenotypes to include.

column.subset	Select a subset of the data set. Use the *column.sel.type* parameter to choose what to base the selection on ("samples" or "phenotypes"). Use this parameter to select which samples or phenotypes to include: • "ALL" (default) – all of them; for example, column.subset = "ALL" • seq(begin, end) – a sequence of sample number; for example, column.subset = seq(1,8) • c(num 1, num 2, num 3) – a list of non-consecutive sample numbers; for example, column.subset = c(1, 4, 10, 15)
thres	Threshold to apply to data set before projection
ceil	Ceiling to apply to data set before projection
fold	Fold change (max/min) for variation filter before projection
delta	Absolute difference (max - min) for variation filter before projection
norm	A number indicating the type of normalization to perform before projection. Use a value of 6 for column rank normalization. Expert users interested in other normalization options should examine the MetageneProjection code.

Output files

Standard output:

1) Stdout:	the "stdout" text output from running the program.
2) Stderr:	the "stderr" error report from running the program if errors occurred

Main output:

1) <identifier>.<date>_<time>.params.txt	File containing the parameters used in the run and the data answer time
2) <identifier>.model_dataset.H.gct	projected model data set
3) <identifier>.all.H.cls	projection of model + test data sets (.cls phenotypes)
4) <identifier>.all.H.gct	projection of model + test data sets (.gct data set)
5) <identifier>.heatmap.jpeg	heat map of projection
6) <identifier>.heatmap.sorted.jpeg	heat map of projection sorted inside each phenotype
7) <identifier>.2D.proj.jpeg	2D biplot of projected model and test data sets
8) <identifier>.H.htree.jpeg	hierarchical tree built on the projected model and test data sets

9) <identifier>.pred.gct	projection-based SVM prediction results (.gct data set)
10) <identifier>.pred.txt	projection-based SVM prediction results (.txt file)
11) <identifier>.H.mem.txt	listing of the samples assigned to each metagene cluster (.txt file) This file lists the samples as they appear in the model and test data sets and also sorts the samples by cluster membership. It can be used to create data sets based on cluster membership.

Additional output:

1) <identifier>.model.H.gct	H matrix from the NMF decomposition
2) <identifier>.model.W.gct	W matrix from the NMF decomposition
3) <identifier>.model_set.2.cls	model data set after pre-preprocessing and refinement (.cls phenotypes)
4) <identifier>.model_set.2.gct	model data set after pre-preprocessing and refinement (.gct file)
5) <identifier>.model_set.1.cls	model data set after pre-preprocessing and before refinement (.cls phenotypes)
6) <identifier>.model_set.1.gct	model data set after pre-preprocessing and before refinement (.gct files)
7) <identifier>.model_set.0.cls	model data set before pre-preprocessing but containing samples after refinement (.cls phenotypes)
8) <identifier>.model_set.0.gct	model data set before pre-preprocessing but containing samples after refinement (.gct file)
9) <identifier>.htree.jpeg	hierarchical tree on original pre-projection data set
10) <identifier>.all.cls	consolidated model + test data set in the space of common genes (.cls phenotypes)
11) <identifier>.all.gct	consolidated model + test data set in the space of common genes (.gct data set)
12) <identifier>.prelim.pred.txt	preliminary projection-based SVM prediction results (used in refinement)

Platform dependencies

Task type: Projection
CPU type: any
OS: any
Language: R

Plot Colors

[1]	white	aliceblue	antiquewhite	antiquewhite1	antiquewhite2
[6]	antiquewhite3	antiquewhite4	aquamarine	aquamarine1	aquamarine2
[11]	aquamarine3	aquamarine4	azure	azure1	azure2
[16]	azure3	azure4	beige	bisque	bisque1
[21]	bisque2	bisque3	bisque4	black	blanchedalmond
[26]	blue	blue1	blue2	blue3	blue4
[31]	blueviolet	brown	brown1	brown2	brown3
[36]	brown4	burlywood	burlywood1	burlywood2	burlywood3
[41]	burlywood4	cadetblue	cadetblue1	cadetblue2	cadetblue3
[46]	cadetblue4	chartreuse	chartreuse1	chartreuse2	chartreuse3
[51]	chartreuse4	chocolate	chocolate1	chocolate2	chocolate3
[56]	chocolate4	coral	coral1	coral2	coral3
[61]	coral4	cornflowerblue	cornsilk	cornsilk1	cornsilk2
[66]	cornsilk3	cornsilk4	cyan	cyan1	cyan2
[71]	cyan3	cyan4	darkblue	darkcyan	darkgoldenrod
[76]	darkgoldenrod1	darkgoldenrod2	darkgoldenrod3	darkgoldenrod4	darkgray
[81]	darkgreen	darkgrey	darkkhaki	darkmagenta	darkolivegreen
[86]	darkolivegreen1	darkolivegreen2	darkolivegreen3	darkolivegreen4	darkorange
[91]	darkorange1	darkorange2	darkorange3	darkorange4	darkorchid
[96]	darkorchid1	darkorchid2	darkorchid3	darkorchid4	darkred
[101]	darksalmon	darkseagreen	darkseagreen1	darkseagreen2	darkseagreen3
[106]	darkseagreen4	darkslateblue	darkslategray	darkslategray1	darkslategray2
[111]	darkslategray3	darkslategray4	darkslategrey	darkturquoise	darkviolet
[116]	deeppink	deeppink1	deeppink2	deeppink3	deeppink4
[121]	deepskyblue	deepskyblue1	deepskyblue2	deepskyblue3	deepskyblue4
[126]	dimgray	dimgrey	dodgerblue	dodgerblue1	dodgerblue2
[131]	dodgerblue3	dodgerblue4	firebrick	firebrick1	firebrick2
[136]	firebrick3	firebrick4	floralwhite	forestgreen	gainsboro
[141]	ghostwhite	gold	gold1	gold2	gold3
[146]	gold4	goldenrod	goldenrod1	goldenrod2	goldenrod3
[151]	goldenrod4	gray	gray0	gray1	gray2
[156]	gray3	gray4	gray5	gray6	gray7

[161]	gray8	gray9	gray10	gray11	gray12
[166]	gray13	gray14	gray15	gray16	gray17
[171]	gray18	gray19	gray20	gray21	gray22
[176]	gray23	gray24	gray25	gray26	gray27
[181]	gray28	gray29	gray30	gray31	gray32
[186]	gray33	gray34	gray35	gray36	gray37
[191]	gray38	gray39	gray40	gray41	gray42
[196]	gray43	gray44	gray45	gray46	gray47
[201]	gray48	gray49	gray50	gray51	gray52
[206]	gray53	gray54	gray55	gray56	gray57
[211]	gray58	gray59	gray60	gray61	gray62
[216]	gray63	gray64	gray65	gray66	gray67
[221]	gray68	gray69	gray70	gray71	gray72
[226]	gray73	gray74	gray75	gray76	gray77
[231]	gray78	gray79	gray80	gray81	gray82
[236]	gray83	gray84	gray85	gray86	gray87
[241]	gray88	gray89	gray90	gray91	gray92
[246]	gray93	gray94	gray95	gray96	gray97
[251]	gray98	gray99	gray100	green	green1
[256]	green2	green3	green4	greenyellow	grey
[261]	grey0	grey1	grey2	grey3	grey4
[266]	grey5	grey6	grey7	grey8	grey9
[271]	grey10	grey11	grey12	grey13	grey14
[276]	grey15	grey16	grey17	grey18	grey19
[281]	grey20	grey21	grey22	grey23	grey24
[286]	grey25	grey26	grey27	grey28	grey29
[291]	grey30	grey31	grey32	grey33	grey34
[296]	grey35	grey36	grey37	grey38	grey39
[301]	grey40	grey41	grey42	grey43	grey44
[306]	grey45	grey46	grey47	grey48	grey49
[311]	grey50	grey51	grey52	grey53	grey54
[316]	grey55	grey56	grey57	grey58	grey59
[321]	grey60	grey61	grey62	grey63	grey64
[326]	grey65	grey66	grey67	grey68	grey69
[331]	grey70	grey71	grey72	grey73	grey74

[336]	grey75	grey76	grey77	grey78	grey79
[341]	grey80	grey81	grey82	grey83	grey84
[346]	grey85	grey86	grey87	grey88	grey89
[351]	grey90	grey91	grey92	grey93	grey94
[356]	grey95	grey96	grey97	grey98	grey99
[361]	grey100	honeydew	honeydew1	honeydew2	honeydew3
[366]	honeydew4	hotpink	hotpink1	hotpink2	hotpink3
[371]	hotpink4	indianred	indianred1	indianred2	indianred3
[376]	indianred4	ivory	ivory1	ivory2	ivory3
[381]	ivory4	khaki	khaki1	khaki2	khaki3
[386]	khaki4	lavender	lavenderblush	lavenderblush1	lavenderblush2
[391]	lavenderblush3	lavenderblush4	lawngreen	lemonchiffon	lemonchiffon1
[396]	lemonchiffon2	lemonchiffon3	lemonchiffon4	lightblue	lightblue1
[401]	lightblue2	lightblue3	lightblue4	lightcoral	lightcyan
[406]	lightcyan1	lightcyan2	lightcyan3	lightcyan4	lightgoldenrod
[411]	lightgoldenrod1	lightgoldenrod2	lightgoldenrod3	lightgoldenrod4	lightgoldenrodyellow
[416]	lightgray	lightgreen	lightgrey	lightpink	lightpink1
[421]	lightpink2	lightpink3	lightpink4	lightsalmon	lightsalmon1
[426]	lightsalmon2	lightsalmon3	lightsalmon4	lightseagreen	lightskyblue
[431]	lightskyblue1	lightskyblue2	lightskyblue3	lightskyblue4	lightslateblue
[436]	lightslategray	lightslategrey	lightsteelblue	lightsteelblue1	lightsteelblue2
[441]	lightsteelblue3	lightsteelblue4	lightyellow	lightyellow1	lightyellow2
[446]	lightyellow3	lightyellow4	limegreen	linen	magenta
[451]	magenta1	magenta2	magenta3	magenta4	maroon
[456]	maroon1	maroon2	maroon3	maroon4	mediumaquamarine
[461]	mediumblue	mediumorchid	mediumorchid1	mediumorchid2	mediumorchid3
[466]	mediumorchid4	mediumpurple	mediumpurple1	mediumpurple2	mediumpurple3
[471]	mediumpurple4	mediumseagreen	mediumslateblue	mediumspringgreen	mediumturquoise
[476]	mediumvioletred	midnightblue	mintcream	mistyrose	mistyrose1
[481]	mistyrose2	mistyrose3	mistyrose4	moccasin	navajowhite
[486]	navajowhite1	navajowhite2	navajowhite3	navajowhite4	navy
[491]	navyblue	oldlace	olivedrab	olivedrab1	olivedrab2
[496]	olivedrab3	olivedrab4	orange	orange1	orange2
[501]	orange3	orange4	orangered	orangered1	orangered2
[506]	orangered3	orangered4	orchid	orchid1	orchid2

[511]	orchid3	orchid4	palegoldenrod	palegreen	palegreen1
[516]	palegreen2	palegreen3	palegreen4	paleturquoise	paleturquoise1
[521]	paleturquoise2	paleturquoise3	paleturquoise4	palevioletred	palevioletred1
[526]	palevioletred2	palevioletred3	palevioletred4	papayawhip	peachpuff
[531]	peachpuff1	peachpuff2	peachpuff3	peachpuff4	peru
[536]	pink	pink1	pink2	pink3	pink4
[541]	plum	plum1	plum2	plum3	plum4
[546]	powderblue	purple	purple1	purple2	purple3
[551]	purple4	red	red1	red2	red3
[556]	red4	rosybrown	rosybrown1	rosybrown2	rosybrown3
[561]	rosybrown4	royalblue	royalblue1	royalblue2	royalblue3
[566]	royalblue4	saddlebrown	salmon	salmon1	salmon2
[571]	salmon3	salmon4	sandybrown	seagreen	seagreen1
[576]	seagreen2	seagreen3	seagreen4	seashell	seashell1
[581]	seashell2	seashell3	seashell4	sienna	sienna1
[586]	sienna2	sienna3	sienna4	skyblue	skyblue1
[591]	skyblue2	skyblue3	skyblue4	slateblue	slateblue1
[596]	slateblue2	slateblue3	slateblue4	slategray	slategray1
[601]	slategray2	slategray3	slategray4	slategrey	snow
[606]	snow1	snow2	snow3	snow4	springgreen
[611]	springgreen1	springgreen2	springgreen3	springgreen4	steelblue
[616]	steelblue1	steelblue2	steelblue3	steelblue4	tan
[621]	tan1	tan2	tan3	tan4	thistle
[626]	thistle1	thistle2	thistle3	thistle4	tomato
[631]	tomato1	tomato2	tomato3	tomato4	turquoise
[636]	turquoise1	turquoise2	turquoise3	turquoise4	violet
[641]	violetred	violetred1	violetred2	violetred3	violetred4
[646]	wheat	wheat1	wheat2	wheat3	wheat4
[651]	whitesmoke	yellow	yellow1	yellow2	yellow3
[656]	yellow4	yellowgreen			