The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year university library subscription	$125	1-year university library subscription	$165
2-year university library subscription	$215	2-year university library subscription	$295
3-year university library subscription	$315	3-year university library subscription	$435
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2013)
13, Number 2, pp. 323–328

Generating Manhattan plots in Stata

Daniel E. Cook
University of Iowa
Iowa City, IA
daniel-e-cook@uiowa.edu

Kelli R. Ryckman
University of Iowa
Iowa City, IA
kelli-ryckman@uiowa.edu

Jeffrey C. Murray
University of Iowa
Iowa City, IA
jeff-murray@uiowa.edu

Abstract. Genome-wide association studies hold the potential for discovering genetic causes for a wide range of diseases, traits, and behaviors. However, the incredible amount of data handling, advanced statistics, and visualization have made conducting these studies difficult for researchers. Here we provide a tool, `manhattan`, for helping investigators easily visualize genome-wide association studies data in Stata.

Keywords: st0295, manhattan, Manhattan plots, genome-wide association studies, single nucleotide polymorphisms

1 Introduction

The number of published genome-wide association studies (GWAS) has seen a staggering level of growth from 453 in 2007 to 2,137 in 2010 (Hindorff et al. 2011). These studies aim to identify the genetic cause for a wide range of diseases, including Alzheimer's (Harold et al. 2009), cancer (Hunter et al. 2007), and diabetes (Hayes et al. 2007), and to elucidate variability in traits, behavior, and other phenotypes. This is accomplished by looking at hundreds of thousands to millions of single nucleotide polymorphisms and other genetic features across upward of 10,000 individual genomes (Corvin, Craddock, and Sullivan 2010). These studies generate enormous amounts of data, which present challenges for researchers in handling data, conducting statistics, and visualizing data (Buckingham 2008).

One method of visualizing GWAS data is through the use of Manhattan plots, so called because of their resemblance to the Manhattan skyline. Manhattan plots are scatterplots, but they are graphed in a characteristic way. To create a Manhattan plot, you need to calculate p-values, which are generated through one of a variety of statistical tests. However, because of the large number of hypotheses being tested in a GWAS, local significance levels typically fall below $p = 10^{-5}$ (Ziegler, König, and Thompson 2008). Resulting p-values associated with each marker are $-\log_{10}$ transformed and plotted on the y axis against their chromosomal position on the x axis. Chromosomes lie end to end on the x axis and often include the 22 autosomal chromosomes and the X, Y, and mitochondrial chromosomes.

Manhattan plots are useful for a variety of reasons. They allow investigators to visualize hundreds of thousands to millions of p-values across an entire genome and to quickly identify potential genetic features associated with phenotypes. They also enable

investigators to identify clusters of genetic features, which associate because of linkage disequilibrium. They can be used diagnostically—to ensure GWAS data are coded and formatted appropriately. Finally, they offer an easily interpretable graphical format to present signals with formal levels of significance. For these reasons, Manhattan plots are a common feature of GWAS publications.

While Manhattan plots are in essence scatterplots, formatting GWAS datasets for their generation can be difficult and time consuming. To help researchers in this process, we have developed a program executed through a new command, `manhattan`, that formats data appropriately for plotting and allows for annotation and customization options of Manhattan plots.

2 Data formatting

Following data cleaning and statistical tests, researchers are typically left with a dataset consisting of, at a minimum, a list of genetic features (string), p-values (real), chromosomes (integer), and their base pair location on a chromosome (integer). Using the `manhattan` command, a user specifies these variables. `manhattan` uses temporary variables to manipulate data into a format necessary for plotting. The program first identifies the number of chromosomes present and generates base pair locations relative to their distance from the beginning of the first chromosome as if they were laid end to end in numerical order. The format in which p-values are specified is detected and, if need be, log transformed. `manhattan` then calculates the median base pair location of each chromosome as locations to place labels. Labels are generated by using chromosome numbers except for the sex chromosomes and mitochondrial chromosomes, which define chromosomes 23, 24, and 25 with the X, Y, and M labels, respectively.

Once data have been reformatted in `manhattan`, plots are generated. Additional options may require additional data manipulation. These options include `spacing()`, `bonferroni()`, and `mlabel()`.

3 The manhattan command

3.1 Syntax

`manhattan` *chromosome base-pair pvalue* [*if*] [, *options*]

options are listed in section 3.2.

3.2 Options

options	Description
Plot options	
`title(`*string*`)`	display a title
`caption(`*string*`)`	display a caption
`xlabel(`*string*`)`	set x label; default is `xlabel(Chromosome)`
`width(`#`)`	set width of plot; default is `width(15)`
`height(`#`)`	set height of plot; default is `height(5)`
Chromosome options	
`x(`#`)`	specify chromosome number to be labeled as X; default is `x(23)`
`y(`#`)`	specify chromosome number to be labeled as Y; default is `y(24)`
`mito(`#`)`	specify chromosome number to be labeled as M; default is `mito(25)`
Graph options	
`bonferroni(h|v|n)`	draw a line at Bonferroni significance level; label line with horizontal (`h`), vertical (`v`), or no (`n`) labels
`mlabel(`*var*`)`	set a variable to use for labeling markers
`mthreshold(`#`|b)`	set a $-\log(p\text{-value})$ above which markers will be labeled, or use `b` to set your threshold at the Bonferroni significance level
`yline(`#`)`	set $\log(p\text{-value})$ at which to draw a line
`labelyline(h|v)`	label line specified with `yline()` by using horizontal labels (`h`) or vertical labels (`v`)
`addmargin`	add a margin to the left and right of the plot, leaving room for labels
Style options	
`color1(`*color*`)`	set first color of markers
`color2(`*color*`)`	set second color of markers
`linecolor(`*color*`)`	set the color of Bonferroni line and label or y line and label

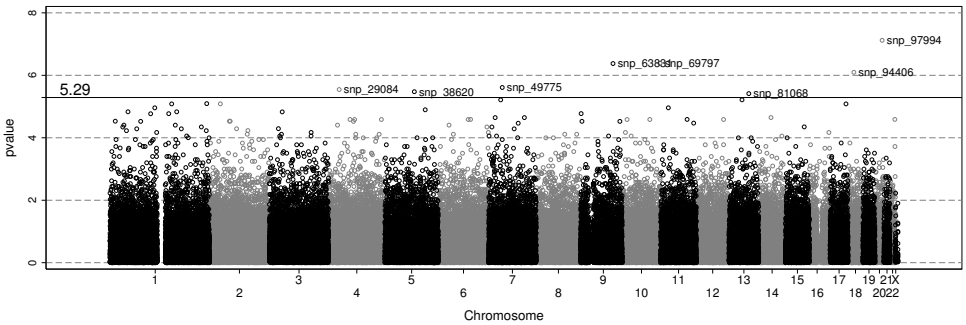
4 Examples

The following examples were created using `manhattan_gwas.dta`, which is available as an ancillary file within the `manhattan` package. All the p-values were generated randomly; therefore, all genetic elements are in linkage equilibrium and are not linked.

4.1 Example 1

Below you will find a typical Manhattan plot generated with `manhattan`. Several options were specified in the generation of this plot. First, `bonferroni(h)` is used to specify that a line be drawn at the Bonferroni level of significance. The `h` indicates that the label should be placed horizontally, on the line. Next, `mlabel(snp)` is used to indicate that markers should be labeled with the variable `snp`, which contains the names of each marker. Additionally, `mthreshold(b)` is used to set a value at which to begin labeling markers. In this case, `b` is used to indicate that markers should be labeled at $-\log_{10}$ (p-values) greater than the Bonferroni significance level. Finally, `addmargin` is used to add space on either side of the plot to prevent labels from running off the plot.

```
. manhattan chr bp pvalue, bonferroni(h) mlabel(snp) mthreshold(b) addmargin
p-values log transformed.
Bonferroni Correction -log10(p) = 5.2891339
Label threshold set to Bonferroni value.
97298
```

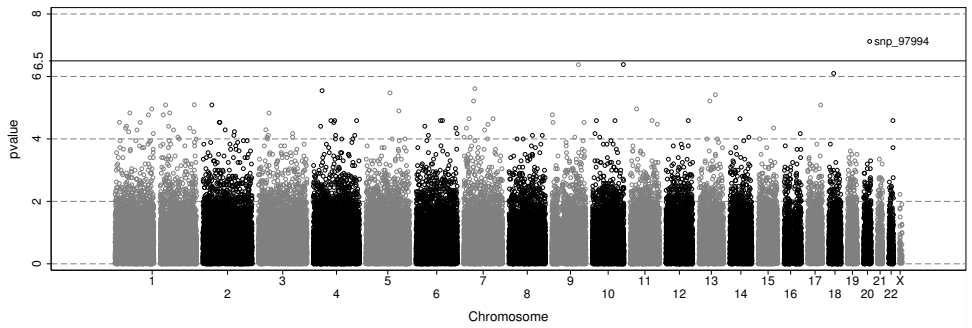


4.2 Example 2

Here `yline(6.5)` is used to draw a horizontal line at $\log_{10}(6.5)$, and `labelyline(v)` adds an axis label for the value of this line. Additionally, the variable used for marker labels is identified using `mlabel(snp)`, and a threshold at which to begin adding labels to markers is given as the same value as the horizontal line by using `mthreshold(6.5)`. Spacing is added between chromosomes with `spacing(1)` to keep labels on the x axis from running into one another. Finally, a margin is added on either side of the plot by using `addmargin`, because some of the marker labels would otherwise fall off the plot.

The colors of the markers are changed with `color1(black)` and `color2(gray)`. The color of the line plotted on the y axis by using `yline(v)` has been changed to black by using `linecolor(black)`.

```
. manhattan chr bp pvalue, yline(6.5) labelyline(v) mlabel(snp) mthreshold(6.5)
> spacing(1) addmargin color1(black) color2(gray) linecolor(black)
p-values log transformed.
97298
```



5 Conclusions

As the number of GWAS publications continues to grow, easier tools are needed for investigators to manipulate, perform statistics on, and visualize data. `manhattan` aims to provide an easier, more standard method by which to visualize GWAS data in Stata. We welcome help in the development of `manhattan` by users and hope to improve `manhattan` in response to user suggestions and comments.

6 Acknowledgments

This work was supported by the March of Dimes (1-FY05-126 and 6-FY08-260), the National Institutes of Health (R01 HD-52953, R01 HD-57192), and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (K99 HD-065786). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development.

7 References

Buckingham, S. D. 2008. Scientific software: Seeing the SNPs between us. *Nature Methods* 5: 903–908.

Corvin, A., N. Craddock, and P. F. Sullivan. 2010. Genome-wide association studies: A primer. *Psychological medicine* 40: 1063–1077.

Harold, D., R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, M. L. Hamshere, J. Singh Pahwa, V. Moskvina, K. Dowzell, A. Williams, N. Jones, C. Thomas, A. Stretton,

A. R. Morgan, S. Lovestone, J. Powell, P. Proitsi, M. K. Lupton, C. Brayne, D. C. Rubinsztein, M. Gill, B. Lawlor, A. Lynch, K. Morgan, K. S. Brown, P. A. Passmore, D. Craig, B. McGuinness, S. Todd, C. Holmes, D. Mann, A. D. Smith, S. Love, P. G. Kehoe, J. Hardy, S. Mead, N. Fox, M. Rossor, J. Collinge, W. Maier, F. Jessen, B. Schürmann, H. van den Bussche, I. Heuser, J. Kornhuber, J. Wiltfang, M. Dichgans, L. Frölich, H. Hampel, M. Hüll, D. Rujescu, A. M. Goate, J. S. K. Kauwe, C. Cruchaga, P. Nowotny, J. C. Morris, K. Mayo, K. Sleegers, K. Bettens, S. Engelborghs, P. P. De Deyn, C. Van Broeckhoven, G. Livingston, N. J. Bass, H. Gurling, A. McQuillin, R. Gwilliam, P. Deloukas, A. Al-Chalabi, C. E. Shaw, M. Tsolaki, A. B. Singleton, R. Guerreiro, T. W. Mühleisen, M. M. Nöthen, S. Moebus, K.-H. Jöckel, N. Klopp, H.-E. Wichmann, M. M. Carrasquillo, V. S. Pankratz, S. G. Younkin, P. A. Holmans, M. O'Donovan, M. J. Owen, and J. Williams. 2009. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature Genetics* 41: 1088–1093.

Hayes, M. G., A. Pluzhnikov, K. Miyake, Y. Sun, M. C. Y. Ng, C. A. Roe, J. E. Below, R. I. Nicolae, A. Konkashbaev, G. I. Bell, N. J. Cox, and C. L. Hanis. 2007. Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. *Diabetes* 56: 3033–3044.

Hindorff, L. A., J. MacArthur, A. Wise, H. A. Junkins, P. N. Hall, A. K. Klemm, and T. A. Manolio. 2011. A catalog of published genome-wide association studies. http://www.genome.gov/gwastudies/.

Hunter, D. J., P. Kraft, K. B. Jacobs, D. G. Cox, M. Yeager, S. E. Hankinson, S. Wacholder, Z. Wang, R. Welch, A. Hutchinson, J. Wang, K. Yu, N. Chatterjee, N. Orr, W. C. Willett, G. A. Colditz, R. G. Ziegler, C. D. Berg, S. S. Buys, C. A. McCarty, H. S. Feigelson, E. E. Calle, M. J. Thun, R. B. Hayes, M. Tucker, D. S. Gerhard, J. F. Fraumeni, Jr., R. N. Hoover, G. Thomas, and S. J. Chanock. 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics* 39: 870–874.

Ziegler, A., I. R. König, and J. R. Thompson. 2008. Biostatistical aspects of genome-wide association studies. *Biometrical Journal* 50: 8–28.

About the authors

Daniel E. Cook is a research assistant in the Department of Pediatrics at the University of Iowa. His research focuses on genetics and bioinformatics approaches.

Kelli K. Ryckman is an Associate Research Scientist in the Department of Pediatrics at the University of Iowa. Her research focuses on the genetics and metabolomics of maternal and fetal complications in pregnancy.

Jeffrey C. Murray received his medical degree from Tufts Medical School in Boston in 1978. He has been conducting research at the University of Iowa since 1984. The Murray laboratory is focused on identifying genetic and environmental causes of complex diseases, specifically premature birth and birth defects such as a cleft lip and palate.