Genetics and population analysis

KGraph: a system for visualizing and evaluating complex genetic associations

Reagan J. Kelly¹,²,*, Douglas M. Jacobsen¹, Yan V. Sun¹, Jennifer A. Smith¹ and Sharon L. R. Kardia¹,²

¹Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI 48104, USA and ²Bioinformatics Program, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA

Received on August 15, 2006; accepted on September 30, 2006
Advance Access publication October 10, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Summary: The KGraph is a data visualization system that has been developed to display the complex relationships between the univariate and bivariate associations among an outcome of interest, a set of covariates, and a set of genetic factors, such as single nucleotide polymorphisms (SNPs). It allows for easy viewing and interpretation of genetic associations, correlations among covariates and SNPs, and information about the replication and cross-validation of the associations. The KGraph allows the user to more easily investigate multicollinearity and confounding through visualization of the multidimensional correlation structure underlying genetic associations. It emphasizes gene–environment and gene–gene interaction, both important components of any genetic system that are often overlooked in association frameworks.

Availability: http://www.epidkardia.sph.umich.edu/software/kgrapher
Contact: reagank@umich.edu
Supplementary information: A description of system requirements and a full user manual are available at http://www.epidkardia.sph.umich.edu/software/kgrapher

INTRODUCTION

The field of common complex disease studies is steadily moving beyond the single gene paradigm that has dominated human genetics for nearly a hundred years (Newton-Cheh and Hirschhorn, 2005). Single genetic factors with large, independent effects on complex phenotypes are relatively rare. For decades, genetic researchers have known that the genetic architecture of common chronic diseases, such as heart disease, hypertension, and diabetes involves many genes acting either additively or non-additively with environmental factors (Sing et al., 1996). Most genetic factors are expected to have small to moderate effects on disease outcomes that are also dependent upon environmental or genetic context (Kardia, 2000). Only a thorough investigation of this gene–gene and gene–environment interaction will allow scientists to fully understand the complexity of these genetic systems and the effect that interindividual variations in the genome, such as single nucleotide polymorphisms (SNPs) and other polymorphisms (e.g. insertion/deletions), have on the natural history of disease. Furthermore, there is a need to incorporate this information into a more thorough understanding of the underlying correlation structure among covariates, among SNPs, and between SNPs and covariates that are predictors of disease risk. Many genetic associations with disease are expected to occur through their effects on key risk factors (e.g. genetic factors associated with plasma cholesterol levels will also be likely associated with heart disease) or as a consequence of linkage disequilibrium (LD). Currently, there are no methods for integrating the information from the multitude of statistical associations underlying a typical genetic association study with disease risk.

In order for the field to move forward, systems and tools must be developed that realistically portray the genetic architecture underlying these complex traits by simultaneously displaying multiple predictors’ main effects, interaction effects, and the underlying correlation structure among predictors. Given the difficulty of identifying genetic factors with replicable effects across studies, the ability to display meta-data about whether the statistical associations cross-validate or are replicated in additional samples also needs to be integrated into these tools. In accordance with Tufte’s principles of graphical excellence (Tufte, 1983), we have developed such a data visualization tool, the KGraph, and the associated utility for generating them, KGrapher, which we describe here.

ANATOMY OF A KGRAPHER

The KGraph has eight graphical regions, each of which displays the results from a single type of statistical analysis. These regions have been specifically arranged into two major sections so that they show (1) the underlying correlation among genetic factors and covariates and (2) the inter-relationships among these genetic and covariate associations with the outcome. The inner section displays this underlying correlation structure in the form of SNP–SNP LD, SNP–covariate association, and covariate–covariate correlation. The outer section displays associations with the outcome of interest (i.e. single covariate association, single SNP association, covariate–covariate interaction, SNP–covariate interaction, and SNP–SNP interaction).

The correlation structure displayed by the inner section provides information about collinearity and potential confounding among covariates, among SNPs, and among SNPs and covariates. The first region, labeled 1 in Figure 1, displays the association between the SNPs and covariates. Cells representing significant results are colored light green, while those representing highly significant

*To whom correspondence should be addressed.
results are dark green. Region 2 displays covariate–covariate correlations with moderate correlations colored light gray and strong correlations colored dark gray. LD among SNPs is displayed in Region 3, with strong disequilibrium shaded dark red and moderate disequilibrium shaded light red.

The outer section displays the associations between the outcome of interest and the covariates and SNPs being examined. For consistency, cells in all regions of the outer section representing significant associations are colored light blue and cells representing highly significant associations are dark blue. Region 4 displays the association between the covariates and the outcome of interest, and region 5 displays the association between the SNPs and the outcome. The remaining three sections display the results from testing for first-order interactions between pairs of covariates (Region 6), between pairs of SNPs and covariates (Region 7), or between pairs of SNPs (Region 8).

Because cross-validation and replication are now standard tools used to differentiate between true and false positive results in genetic studies (Manly, 2005; Molinaro et al., 2005), the KGraph can represent both cross-validation (indicated by a small horizontal bar within the cell) and/or replication results (displayed by dividing a cell with a diagonal line, with each half-cell representing one sample).

**IMPLEMENTATION**

The KGrapher utility provides a ‘wizard interface’ that prompts the user for the files and criteria needed to plot each region. The input files are the results from the statistical analyses. Criteria include the names of the SNPs and covariates, the significance measurements desired by the user (e.g. ANOVA P-value, Pearson’s correlation coefficient, etc.), the alpha levels used to designate significant and highly significant results, and the cross-validation measurement and cutoff values. KGrapher is a full-featured KGraph viewer that uses both an overlay and a tooltip system to provide the user with specific information about individual tests and patterns in tests involving the same predictors. It also allows the user to generate a number of KGraphs simultaneously using a batch creation mode.

KGrapher accepts both tab delimited (.txt) and comma separated (.csv) files for data input. It uses a dedicated file format (.kgr) that contains both the input data and the configuration settings for a given KGraph. Users are able to easily save KGraphs, print
them, and share them with other researchers. Additionally, KGrapher allows users to export KGraphs as high-resolution JPEG (.jpg) files for publication or display on the Internet.

ACKNOWLEDGEMENTS
The authors would like to thank Jian Chu, Guo Li, Kristin Meyers, and Todd Greene for their helpful comments in both the development of the KGraph format and on this manuscript. This work was supported by National Institute of Health grant HL54457.

Conflict of Interest: none declared.

REFERENCES