PedWiz: a web-based tool for pedigree informatics

Yeunjoo E. Song and Robert C. Elston*

Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

A novel web-based tool PedWiz that pipelines the informatics process for pedigree data is introduced. PedWiz is designed to assist researchers in the analysis of pedigree data. It provides a convenient tool for pedigree informatics: descriptive statistics, relative pairs, genetic similarity coefficients, the variance-covariance matrix for three estimated coefficients of allele identical-by-descent sharing as well as mean allele sharing, a plot of the pedigree structures, and a visualization of the identity coefficients. With a renewed interest in linkage and other family based methods, PedWiz will be a valuable tool for the analysis of family data.

Keywords: pedigree, informatics, genetic similarity, identity-by-descent, relative pairs, family data

INTRODUCTION

When a researcher has collected or is provided with a set of nuclear family or extended pedigree data for genetic analysis, the first thing that needs to be done is to find out what information is available on the family or families before proceeding in the analysis of phenotype and/or genotype data to study the characteristics of a certain disease or trait, i.e., pedigree informatics. This can include descriptive statistics, visualization of family data, the degree of genetic relatedness among members of a family, and so on.

Descriptive statistics summarize and provide basic information on the family data, as done in the PEDINFO program in S.A.G.E. (2012). The visualization of family data is a fundamental task for both family studies and genetic counseling. There are many computer programs available that provide the graphical representation of pedigree data, including the R packages kinship (Zhao, 2006) and pedigantry (Morrissey, 2010). The concept of genetic relatedness is essential in modern genetic analysis, and the applications of kinship and condensed identity coefficients are everywhere in analyses that have a genetic component. In human genetics, they are used in genotype prediction, calculation of genetic risk ratios for binary disease-status, calculations of correlations between relatives, and robust linkage analysis. Robust linkage analysis, a powerful approach to map disease genes, is based on comparing the genetic marker profiles, i.e., allele identical-by-descent (IBD) sharing, of pairs of relatives. There are many software programs that calculate kinship and inbreeding coefficients, but not many for the nine condensed coefficients of IBD sharing.

A brief survey of available R packages with their relevant components of pedigree informatics is shown in Table 1. As can be seen, there is no program that provides all the different genetic similarity measurements together with the variance-covariance matrix of the estimated coefficients of IBD. Abney (2009)’s graphical algorithm for the computation of the generalized kinship coefficients is implemented in idkocy2 (written in C++) and implemented as the R package identity), and this is the only currently available program that outputs the nine condensed coefficients of IBD. The R package ibdreg by Schaid et al. (2007) has two functions, sim.ibd.var and exact.ibd.var, to calculate the variance-covariance of mean allele sharing, but not the variance-covariance of the individual coefficients of IBD. An essential part of score tests is the choice of the denominator variance, and some of these tests for genetic linkage require the variance-covariance of allele IBD sharing statistics under the null, i.e., of the coefficients of IBD. It would be useful to make available the variance-covariance matrix of these coefficients for a pedigree independent of the choice of test statistics, so that it can be used for different choices of test statistics. Currently, no such tools are available.

PedWiz (Pedigree Informatics Wizard) is designed to fulfill this need as a web-based tool for pedigree informatics, to assist researchers in the analysis of pedigree data. It provides a convenient “one-stop-shop” for pedigree informatics. It provides all the genetic similarity coefficients mentioned above, including the nine condensed coefficients of IBD and the variance-covariance matrix of the one-locus three marginal coefficients of allele IBD sharing, as well as other pedigree descriptive statistics. Additionally, it provides a plot of the pedigree structure and a visualization of the identity coefficients, something that no other program provides. PedWiz is an automated pipeline for extracting pedigree informatics before conducting specialized analyses of phenotype and/or genotype data.

MATERIAL AND METHODS

IMPLEMENTATION

The web interface of PedWiz is implemented using a combination of XHTML, (eXensible HyperText Markup Language), CSS (Cascading Style Sheets), and PHP (Hypertext Preprocessor) on an Apache web server. The interactivity is provided by javascript and Ajax technologies. Custom Python modules handle the overall
Table 1 | R packages available for pedigree informatics.

| Name            | Plot | Stat   | F | Δ | VC(2Δ) | VC(Δ) | Simulation |
|-----------------|------|--------|---|---|---------|-------|------------|
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |
| adegenet        |      |        |   |   |         |       |            |

Plot: pedigree plot; stat: descriptive statistics; F: inbreeding coefficient; Δ: kinship coefficient; VC(2Δ): variance-covariance matrix of mean allele sharing; VC(Δ): variance-covariance matrix of IBD coefficients.

This tool is used to visualize a pedigree. PedWiz utilizes the R package pedigree to generate the plot (Zhao, 2005). As in a typical pedigree diagram, males and females are shown as squares and circles, respectively. The resulting pedigree plot is reported to the user as a pdf file on the website.

The descriptive statistics tool
This tool is used to report all relative pairs existing in a pedigree. PedWiz uses an internal C++ program that finds all existing relative pairs by traversing the pedigree structure recursively as done in the FCOR program in S.A.G.E. (2012). The results are reported to the user on the website as a text file containing the relative pair matrix and the list of relative pairs for each relative type.

The genetic similarity tool
This tool is used to provide the various genetic similarity coefficients. PedWiz uses an internal C++ program to perform this task. The results include two matrices; one is the matrix of kinship/inbreeding coefficients (inbreeding coefficients on the diagonal and kinship coefficients off the diagonal), and the other is the matrix of nine condensed coefficients of IBD. The coefficients of relationship, which are twice the kinship coefficients, can be easily calculated from the kinship/inbreeding coefficients. The resulting matrices are reported to the user on the website as a text file.

The visualization of genetic similarity tool
This tool is used to visualize the two matrices generated by the genetic similarity tool. PedWiz uses a custom R script to represent a matrix graphically as a heat map. The resulting heat maps are reported to the user as a pdf file on the website.

The variance-covariance of genetic similarity tool
This tool is used to find the variance-covariance matrix of the coefficients reported by the genetic similarity tool. PedWiz uses an internal C++ program to perform this task. The variance-covariance matrix of kinship coefficients is estimated by an exact method given by Chen and Abecasis (2006). The variance-covariance matrix of IBD coefficients is estimated by a simulation method, given a pedigree structure (MacCluer et al., 1986), based on 500 simulation replicates. The simulation method approximates the distribution of IBD states by gene dropping, so it can be used regardless of pedigree size and structure. The results are reported to the user on the website as a text file.

RESULTS
We developed a novel web-based tool that pipelines the informatics process for pedigree data. PedWiz may be accessed at http://darwin.cwru.edu/~song/pedwiz. Here we present an application example using pedigree data from the Madeline 2.0 website (Trager et al., 2007). These pedigree data contain a consanguineous marriage between cousins. The user inputs configuration information and the location of the pedigree file through the interface on the website as shown in Figure 2.

After configuration information and the location of the pedigree file have been submitted by the user, PedWiz produces a table with the descriptive statistics for each pedigree on the website as shown in Figure 3. All results are accessed through a set of buttons under the descriptive statistics table for each pedigree. The user uses a radio button to select a pedigree for an analysis pipeline.

The complete process utilizes many internal Python scripts (which are not detailed here) to create junctions between the programs (format compatibility) and to create the necessary R scripts.

The descriptive statistics tool
This tool is used to calculate the descriptive statistics for each pedigree contained in the user-submitted pedigree file. PedWiz utilizes the existing C++ program PEDINFO of the S.A.G.E. package (v6.3 with pedgenet = true option). PEDINFO provides many useful descriptive statistics on pedigree data including means, standard deviations; family, sibship and pedigree sizes; and counts of each type of relative pair. The results are parsed and reported to the user by PedWiz as a table on the website. From here, the user selects a pedigree to proceed with other tools.

The pedigree plot tool
This tool is used to visualize a pedigree. PedWiz utilizes the R package kinship to generate the plot (Zhao, 2005). As in a typical pedigree diagram, males and females are shown as squares and circles, respectively. The resulting pedigree plot is reported to the user as a pdf file on the website.

The relative pairs tool
This tool is used to report all relative pairs existing in a pedigree. PedWiz uses an internal C++ program that finds all existing relative pairs by traversing the pedigree structure recursively as done in the FCOR program in S.A.G.E. (2012). The results are reported to the user on the website as a text file containing the relative pair matrix and the list of relative pairs for each relative type.

The genetic similarity tool
This tool is used to provide the various genetic similarity coefficients. PedWiz uses an internal C++ program to perform this task. The results include two matrices; one is the matrix of kinship/inbreeding coefficients (inbreeding coefficients on the diagonal and kinship coefficients off the diagonal), and the other is the matrix of nine condensed coefficients of IBD. The coefficients of relationship, which are twice the kinship coefficients, can be easily calculated from the kinship/inbreeding coefficients. The resulting matrices are reported to the user on the website as a text file.

The visualization of genetic similarity tool
This tool is used to visualize the two matrices generated by the genetic similarity tool. PedWiz uses a custom R script to represent a matrix graphically as a heat map. The resulting heat maps are reported to the user as a pdf file on the website.

The variance-covariance of genetic similarity tool
This tool is used to find the variance-covariance matrix of the coefficients reported by the genetic similarity tool. PedWiz uses an internal C++ program to perform this task. The variance-covariance matrix of kinship coefficients is estimated by an exact method given by Chen and Abecasis (2006). The variance-covariance matrix of IBD coefficients is estimated by a simulation method, given a pedigree structure (MacCluer et al., 1986), based on 500 simulation replicates. The simulation method approximates the distribution of IBD states by gene dropping, so it can be used regardless of pedigree size and structure. The results are reported to the user on the website as a text file.
This selection information is reflected under the table (shown in the green eclipse). The resulting output from each tool for the example pedigree is shown also.

**DISCUSSION**

We developed a novel web-based tool PedWiz that pipelines the informatics process for pedigree data. PedWiz is designed to assist researchers in the analysis of pedigree data. It provides a convenient tool for pedigree informatics: descriptive statistics, relative pairs, genetic similarity coefficients, the variance-covariance matrix of three coefficients of allele IBD sharing, as well as mean allele sharing, a plot of the pedigree structure, and visualization of identity coefficients. PedWiz is an automated pipeline for extracting pedigree informatics before conducting specialized analysis of phenotype and/or genotype data.

Emerging availability of whole genome sequence data has led to a renewed interest in linkage and other family based methods for selecting information in the green eclipse. The resulting output from each tool for the example pedigree is shown also.

Emerging availability of whole genome sequence data has led to a renewed interest in linkage and other family based methods. Many researchers have been emphasizing the importance and advantages of family studies all along (Clergy-Darpoux and Elston, 2007; Stein and Elston, 2009), especially to interpret next generation sequence data (Bailey-Wilson and Wilson, 2011; Wijsman, 2012). Family study designs provide not only the enrichment of genetic loci containing rare variants, but also methods to control for genetic heterogeneity and population stratification. PedWiz is a valuable tool for initial analysis of those family data.

Additionally, the results from each tool in PedWiz will be useful for later analysis of phenotype and/or genotype data. As stated before, an essential part of score tests is the choice of the denominator variance, and some of these tests for genetic linkage require the variance-covariance of the coefficients of IBD. No software tools are currently available to provide this information independent of the choice of test statistics. The variance-covariance of the genetic similarity tool of PedWiz provides this need, so that it...
Song and Elston PedWiz: a web-based tool for pedigree informatics

**FIGURE 2** Starting PedWiz. This figure illustrates the user interface to start PedWiz.

The genetic similarity tool of PedWiz can be used for different choices of test statistics. The information from the genetic similarity tool of PedWiz can be used for weighting pedigrees of different sizes. Another potential use of this tool is for selecting families with the most information in terms of genetic relatedness that would best suit a phenotype/genotype analysis of choice. Selecting families with multiple affected subjects, or families with extreme values, is known to provide improved ability to measure, and detect, the effects of rare variants (Ionita-Laza and Ottman, 2011; Wijsman, 2012). The strategy of selecting “large linked families” for initial screening has long been a successful strategy (Bowden et al., 2010). To be successful with this approach, selecting families with a real linkage signal in specific regions is essential. This new tool will be useful for selecting such families when used together with phenotype/genotype information.

With a modular design, each analysis tool within PedWiz is independent of the others, so it is very easy to extend and add more tools. Planned additions in the near future are simulation and pedigree split tools, shown in Figure 1 with dotted lines. PedWiz is currently specialized to deal with the information contained within pedigree structures only. Therefore, it is very fast and safe with regard to data transfer over the web. However, it is always possible to add more pipeline modules that could process the information from phenotype and/or genotype data. Good candidates for this addition would be simulation conditional on given phenotype and/or genotype data, and imputation. Another extension that could be added on is the inclusion of a backend database to save data and results for reuse.

The genetic similarity tool of PedWiz is specifically designed to provide the information on within-pedigree relatedness. As a reviewer pointed out, a tool that addresses between-pedigree relatedness, especially for pedigrees from a relatively isolated population like the Hutterites, would be a useful addition to
Cryptic relatedness among unrelated individuals can be estimated by incorporating a number of dense markers across different chromosomes (Weir et al., 2006; Bink et al., 2008; Astle and Balding, 2009; Sillanpää, 2011). There are many software tools available to estimate the genome-average relatedness, for example, SPAGeDi (Hardy and Vekemans, 2002), PLINK (Purcell et al., 2007), FEST (Skare et al., 2009), CoCoa (Maenhout et al., 2009), CrypticIBDcheck (Nembot-Simo et al., 2013). Adding this to PedWiz would require an extension to process information from phenotype and/or genotype data, as mentioned above.

ACKNOWLEDGMENTS
Some of the results of this paper were obtained by using the program package S.A.G.E., which was supported by a U.S. Public Health Service Resource Grant (RR03655) from the National Center for Research Resources.

REFERENCES
Abney, M. (2009). A graphical algorithm for fast computation of identity coefficients and generalized kinship coefficients. Bioinformatics 25, 1561–1563. doi: 10.1093/bioinformatics/btp185

Astle, W., and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. Stat. Sci. 24, 451–471. doi: 10.1214/09-STS307

Baker-Wilson, J. E., and Wilson, A. E. (2011). Linkage analysis in the next generation sequencing era. Hum. Hered. 72, 226–236. doi: 10.1159/000334381

Bink, M. C., Anderson, A. D., van de Weg, W. E., and Thompson, E. A. (2008). Comparison of marker-based pairwise relatedness estimators on a pedigreed plant population. Theor. Appl. Genet. 117, 843–855. doi: 10.1007/s00122-008-0824-1

Bowden, D. W., An, S. S., Palmer, N. D., Brown, W. M., Norris, J. M., Haffner, S. M., et al. (2010). Molecular basis of a linkage peak: exome sequencing.
Song and Elston

PedWiz: a web-based tool for pedigree informatics

and family-based analyses identify a rare genetic variant in the ADIPQ gene in the IRAS Family Study. Hum. Mol. Genet. 19, 4112–4120. doi: 10.1093/hmg/ddq327

Chen, W. M., and Abecasis, G. R. (2006). Estimating the power of variance-component linkage analyses in large pedigrees. Genet. Epidemiol. 30, 472–484. doi: 10.1002/gepi.20560

Clerget-Darpoux, F., and Elston, R. C. (2007). Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. Hum. Hered. 64, 91–96. doi: 10.1159/000101960

Husby, O. I., and Vokrouhlicky, D. (2002). SPHGen: a versatile computer program to analyze spatial genetic structure at the individual or population levels. Mol. Ecol. Note 2, 618–620. doi: 10.1046/j.1471-2286.2002.0070X.a

Ionita-Laus, I., and Ottman, R. (2011). Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. Genetics 189, 1061–1068. doi: 10.1534/genetics.111.133813

MaxClair, J. W., Vandenberg, J. L., Read, B., and Byers, O. A. (1996). Pedigree analysis by computer simulation. Zoolog. J. Linn. Soc. 5, 149–160. doi: 10.1111/j.1469-2130.1996.tb01219.x

Muenholtz, S., De Barros, B., and Hagen, G. (2009). CoCoa: A software tool for estimating the coefficient of coancestry from multilocus genotype data. Bioinformatics 25, 2753–2754. doi: 10.1093/bioinformatics/btp499

Münterey, N. E. (2010). Pedimice: an R package for pedigree-based genetic simulations and pedigree manipulation, characterization and viewing. Mol. Ecol. Resour. 10, 731–739. doi: 10.1111/j.1755-0998.2009.02817.x

Nemeth-Simo, A., Graham, J., and McNeely, B. (2013). CryptiBID-check: an R package for checking cryptic relatedness in nominally unrelated individuals. Source Code Biol. Med. 8, 5. doi: 10.1186/1755-0714-8-5

Oni, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. Nat. Rev. Genet. 12, 465–474. doi: 10.1038/nrg2999

Patilii, K., Niu, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Benidst, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. doi: 10.1086/510793

S.A.G.E. (2012). Statistical Analysis and Manipulation, Characterization and View Pedigrees. Front. Genet. 3, 10.3389/fgene.2012.00305.x

Snedaker, S. A. (2009). CoCoa: a software computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Note 9, 149–160. doi: 10.1111/j.1471-2286.2009.01983.x

Stein, C. M., and Elston, R. C. (2009). Finding genes underlying human disease. Clin. Genet. 75, 101–109. doi: 10.1111/j.1399-0004.2008.01983.x

Stiger, E. H., Khair, R., Marrs, A., Siden, L., Bransh, K. E. H., Swaroop, A., et al. (2007). Madeline 2.0: a new program for local and web-based linkage analysis. Bioinformatics 23, 1854–1856. doi: 10.1093/bioinformatics/btm242

Wian, D., Anderson, A., and Hopper, A. (2000). Genetic relatedness analysis: modern data and new challenges. Nat. Rev. Genet. 7, 771–780. doi: 10.1038/nrg1603

Zhao, J. H. (2005). Mixed-effects Cox models of alcohol dependence in extended pedigrees. BMC Genet. 6,227. doi: 10.1186/1471-2156-6-227

Zhao, J. H. (2006). Pedigree drawing with R and graphviz. Bioinformatics 22, 1015–1014. doi: 10.1093/bioinformatics/btl058

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

Received: 09 Jul 2013; accepted: 04 September 2013; published online: 25 September 2013.

Clusser: Sergey ZE and Elston BC (2013) PedWiz: a web-based tool for pedigree informatics. Front. Genet. 4:189. doi: 10.3389/fgene.2013.00189

This article was submitted to Applied Genetic Epidemiology, a section of the Journal Frontiers in Genetics.

Copyright © 2013 Song and Elston. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction in any medium is permitted which does not comply with these terms.

"fgene-04-00189" — 2013/9/23 — 20:54 — page 6 — #6