BioData Mining

SHORT REPORT Open Access

PGxClean: a quality control GUI for the Affymetrix DMET chip and other candidate gene studies with non-biallelic alleles

Daniel Rotroff[1], John Jack[1], Nathan Campbell[2], Scott Clark[2] and Alison A Motsinger-Reif[1,3*]

* Correspondence: alison_motsinger@ncsu.edu
[1]Bioinformatics Research Center, Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
[3]NCSU Department of Statistics, 2311 Stinson Drive, Campus Box 7566, Raleigh, NC 27695, USA
Full list of author information is available at the end of the article

Abstract

Background: PGxClean is a new web application that performs quality control analyses for data produced by the Affymetrix DMET chip or other candidate gene technologies. Importantly, the software does not assume that variants are biallelic single-nucleotide polymorphisms, but can be used on the variety of variant characteristics included on the DMET chip. Once quality control analyses has been completed, the associated PGxClean-Viz web application performs principal component analyses and provides tools for characterizing and visualizing population structure.

Findings: The PGxClean web application accepts genotype data from the Affymetrix DMET chip or the PLINK PED format with genotypes annotated as (A,C,G,T or 1,2,3,4). Options for removing missing data and calculating genotype and allele frequencies are offered. Data can be subdivided by cohort characteristics, such as family ID, sex, phenotype, or case–control status. Once the data has been processed through the PGxClean web application, the output files can be entered into the PGxClean-Viz web application for performing principal component analysis to visualize population substructure.

Conclusions: The PGxClean software provides rapid quality-control processing, data analysis, and data visualization for the Affymetrix DMET chip or other candidate gene technologies while improving on common analysis platforms by not assuming that variants are biallelic. The web application is available at www.pgxclean.com.

Keywords: SNP, Bioinformatics, Data visualization, Genomics

Findings

While current single nucleotide polymorphism (SNP) chip technologies produce generally very high quality data, it is still critical that quality control and data cleaning steps are components of any genetic study analysis plan. There are several quality control (QC) steps that have become "best practices" in data cleaning in SNP data [1]. These steps have been integrated into commonly used software packages, such as PLINK [2]. These software packages have generally been developed for genome-wide SNP chips, that are designed to genotype bi-allelic SNPs or copy number variants.

While current knowledge supports that biallelic SNPs are the most common in the genome, we know there are a number of genes with multi-allelic genotypes, complex

haplotype/diplotype structures, etc. that are not readily genotyped on standard genome-wide chips [3]. In the field of pharmacogenomics, this is of particular importance because many of the established associations are in genes that do not follow the typical biallelic assumptions and are not well covered on standard chips [4,5]. In response to this, both Affymetrix (www.affymetrix.com) and Illumina (www.illumina.com) have developed specific genotyping arrays for pharmacogenomics genes.

As these chips are growing in popularity, it drives the need for quality control tools that properly "clean" and process data without requiring that the genotypes are biallelic. In the current study, we introduce PGxClean, a web-based software application with a graphical user interface (GUI) that was designed to perform basic QC and publication quality figures for typical quality control procedures. The software was designed for the output format of the Affymetrix DMET Plus chip, but can also be used with the commonly used PLINK "PED" format [2] to make it readily compatible with other data (both from the Illumina pharmacogenomics chip, and data collected on other platforms).

PGxClean is implemented in the freely available R-Shiny software [6,7], and several available packages, as described in the documentation found at http://cran.us.r-project.org. The source code is available for download at http://www4.stat.ncsu.edu/~motsinger. The PGxClean website can be accessed at www.pgxclean.com. Example data files are available for download on the homepage (both DMET and PED formats) to help format your own files or for experimenting with the website functionality. If a PED file is uploaded, an accompanying MAP file must also be uploaded to provide the appropriate column headers. Additional details about the MAP files are available under the 'Documentation' tab in the

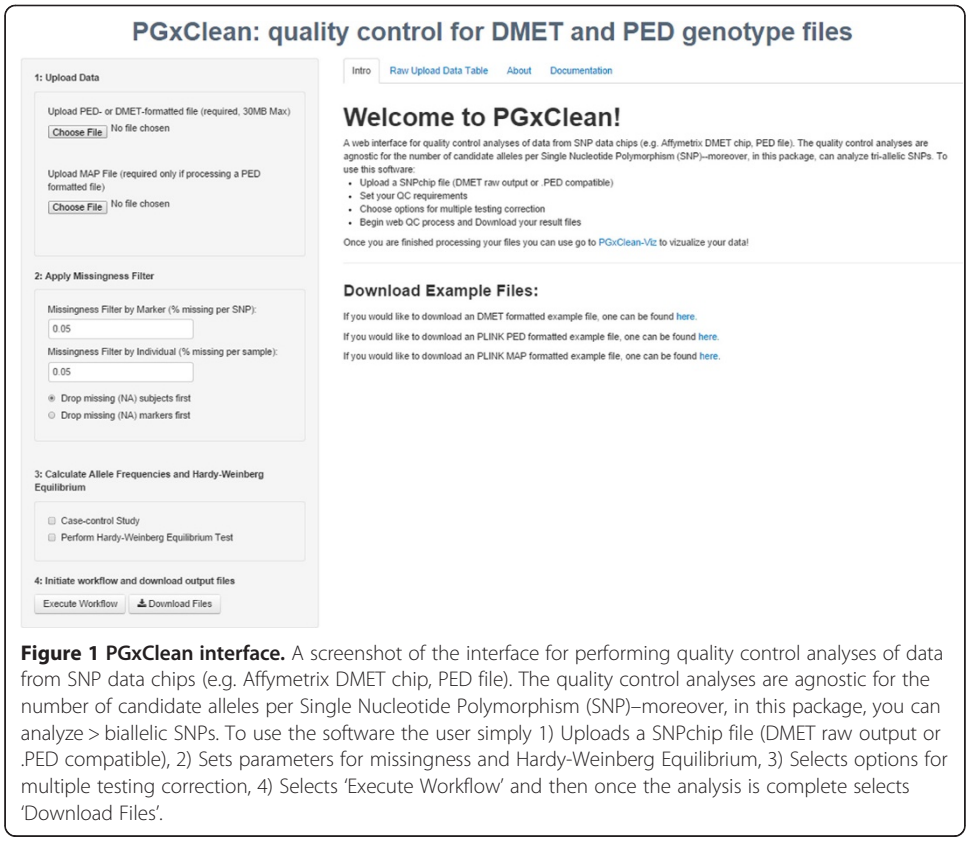

Figure 1 PGxClean interface. A screenshot of the interface for performing quality control analyses of data from SNP data chips (e.g. Affymetrix DMET chip, PED file). The quality control analyses are agnostic for the number of candidate alleles per Single Nucleotide Polymorphism (SNP)–moreover, in this package, you can analyze > biallelic SNPs. To use the software the user simply 1) Uploads a SNPchip file (DMET raw output or .PED compatible), 2) Sets parameters for missingness and Hardy-Weinberg Equilibrium, 3) Selects options for multiple testing correction, 4) Selects 'Execute Workflow' and then once the analysis is complete selects 'Download Files'.

navigation menu on the website. A searchable preview of uploaded data is available by selecting the 'Raw Upload Data Table' tab under the navigation menu. For DMET chip data (or other genetic data on a similar scale, with about 2000 variants), the whole QC process can be completed in only a few minutes. The software is designed for candidate gene data, and would not run efficiently for genome-wide scale data. In addition, the software does not test for cryptic relatedness, gender checks, or perform genotyping concordance with reference samples (e.g. HapMap). However, these tools could be implemented into future iterations of PGxClean.

PGxClean has default parameter choices implemented, but can be easily tailored based on user preference by adjusting the parameters in the sidebar widget. Additional explanations of the user parameters are available by selecting the 'Documentation' tab under the navigation menu.

(1) *Genotyping efficiency/Missing Data.* The first QC step is screening both variants and individuals for high levels of missing data. By default, if variants have more than 5% data missing across all individuals, they are removed from the dataset. Then, individuals are checked for missing data, and if more than 5% of variants are missing, these individuals are removed. The percentage can be user-specified to be specific to the needs of a particular study.

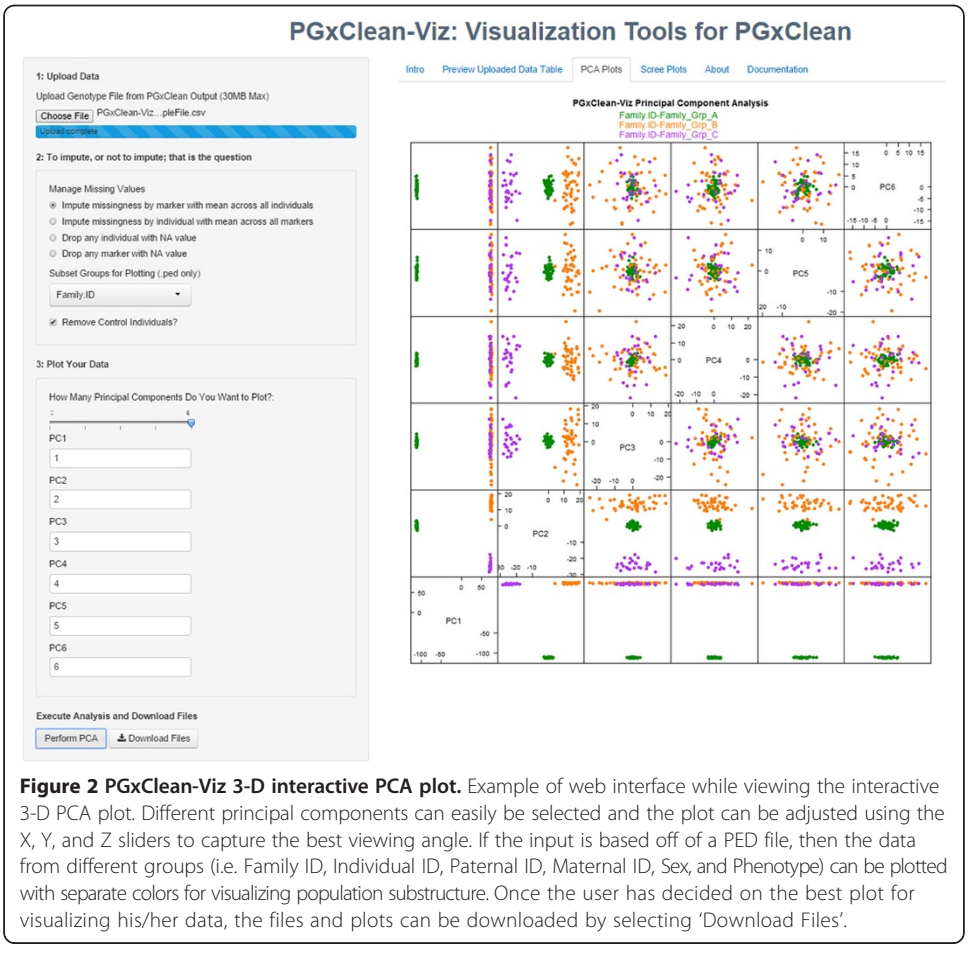

Figure 2 PGxClean-Viz 3-D interactive PCA plot. Example of web interface while viewing the interactive 3-D PCA plot. Different principal components can easily be selected and the plot can be adjusted using the X, Y, and Z sliders to capture the best viewing angle. If the input is based off of a PED file, then the data from different groups (i.e. Family ID, Individual ID, Paternal ID, Maternal ID, Sex, and Phenotype) can be plotted with separate colors for visualizing population substructure. Once the user has decided on the best plot for visualizing his/her data, the files and plots can be downloaded by selecting 'Download Files'.

(2) *Test for Deviation from Hardy-Weinberg Proportions.* Testing markers for deviation from proportions expected under Hardy-Weinberg equilibrium [8] has become an important check for overall genotyping quality, and PGxClean will test for deviations using Fisher's Exact test (so they are valid even in very small samples or for low allele frequencies) and will report results with or without correction for multiple comparisons. Typical implementations of tests for Hardy-Weinberg disequilibrium assume that variants are biallelic. For PGxClean, we used expanded versions of the Hardy-Weinberg equation for multiple alleles to calculate expected values. Additionally, this analysis can be performed on stratified portions of the dataset, specified in a "PED" format. For example, if a case–control study was performed, this filter should be performed on only the control samples, or you might want to perform this analysis separately for different ethnic groups in a heterogeneous sample. This can be accomplished by selecting the group you would like to stratify using the drop down menu that appears once the PED file is uploaded. Additionally, if a DMET formatted file is uploaded, the "case–control study" box can be selected (Figure 1). This will stratify the analysis based on IDs with "control" in the sample name. Additional details regarding this step are available under the "documentation" tab at www.pgxclean.com.

After processing data using PGxClean, the user can download a zipped file containing several outputs including, the newly 'cleaned' data file, allele frequencies, the results of the Hardy-Weinberg equilibrium test data, and a genotype file. The genotype output

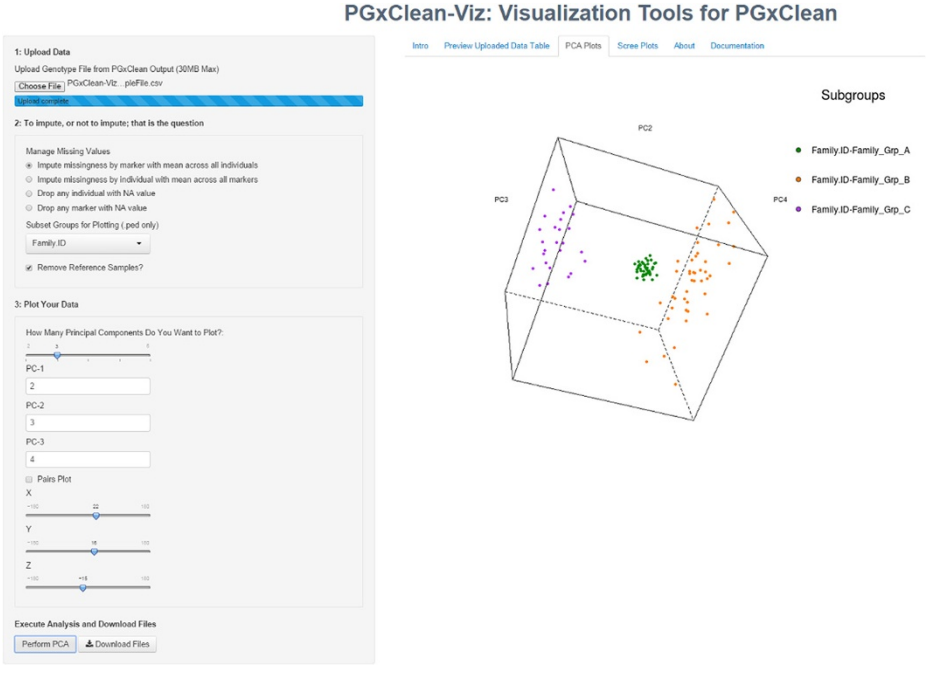

Figure 3 PGxClean-Viz PCA pairs plot. Example of web interface while viewing contrasts of up to 6 principal components. If the input is based off of a PED file, then the data from different groups (i.e. Family ID, Individual ID, Paternal ID, Maternal ID, Sex, and Phenotype) can be plotted with separate colors for visualizing population substructure. Once the user has decided on the best plot for visualizing his/her data, the files and plots can be downloaded by selecting 'Download Files'.

file can be used in PGxClean-Viz, an extension of PGxClean that provides tools for performing principal component analysis (PCA) and other visualizations.

Principal Component Analysis (PCA)

PCA has become a standard procedure for looking at population substructure to identify confounders such as batch effects, and to identify cryptic relatedness in population based cohorts [9]. PCA is incorporated into PGxClean-Viz and can be accessed through PGxClean or www.pgxclean.com/viz. The user simply uploads the genotype file that is created as a PGxClean output file, selects how missing values should be managed, and selects the number of principal components to be plotted and presses 'Perform PCA'. Additionally, if the PGxClean file is uploaded as a "PED" formatted file, the plots can be stratified by family ID, individual ID, paternal ID, maternal ID, sex, or phenotype. These data are recorded in the first 6 columns of the PGxClean output file (consistent with the PLINK PED file) and offer the option for observing various forms of stratification (e.g. race, ethnicity, or sex). The corresponding PCA plot is rendered on the screen as either a two dimensional scatterplot, 3D scatter plot (with the ability to interactively rotate the plot), or pairs plot with up to 6 principal components plotted based on user specification. In addition, scree plots provide the resulting percentage of variance explained by each principal component can be seen by selecting the 'Scree Plots' tab in the navigation menu. Once the user is satisfied with their visualization of the PCA, a zipped file containing all of the figures can be downloaded by selecting 'Download Files'. The downloaded file also contains the corresponding rotation data and eigenvectors from the PCA if additional follow-up analysis is desired. A screenshot of the console for the software, along with example figures from the PCA tools are shown in Figures 1, 2 and 3.

While this software was designed for the DMET Plus chip, it also allows for PED format files, so it should be readily useable for a wide range of genotype data and is unique in its ability to process non-biallelic variants.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AMR, SC, and NC conceived of the original design for the software. DR led the software development, drafted the manuscript. JJ aided in the web tool application. All authors assisted with editing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Kevin Long and David Reif for testing the software.

Author details

[1]Bioinformatics Research Center, Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA. [2]Gentris LLC, a CGI Company, Morrisville, NC 27560, USA. [3]NCSU Department of Statistics, 2311 Stinson Drive, Campus Box 7566, Raleigh, NC 27695, USA.

Received: 8 June 2014 Accepted: 18 October 2014
Published: 6 November 2014

References

1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges.** *Nat Rev Genet* 2008, **9:**356–369.
2. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559–575.
3. Hodgkinson A, Eyre-Walker A: **Human triallelic sites: evidence for a new mutational mechanism?** *Genetics* 2010, **184:**233–241.

4. Peters EJ, McLeod HL: **Ability of whole-genome SNP arrays to capture 'must have' pharmacogenomic variants.** 2008, **9**(11):1573–1577.
5. Oetjens MT, Denny JC, Ritchie MD, Gillani NB, Richardson DM, Restrepo NA, Pulley JM, Dilks HH, Basford MA, Bowton E, Masys DR, Wilke RA, Roden DM, Crawford DC: **Assessment of a pharmacogenomic marker panel in a polypharmacy population identified from electronic medical records.** *Pharmacogenomics* 2013, **14:**735–744.
6. RStudio and Inc: *shiny: Web Application Framework for R.* 2013.
7. R Development Core Team: *A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* 2013. ISBN 3-900051-07-0, URL http://www.R-project.org/.
8. Hardy GH: **Mendelian proportions in a mixed population.** *Science* 1908, **28:**49–50.
9. Pearson K: **Principal components analysis.** *Lond Edinb Dublin Philos Mag J Sci* 1901, **6:**559.

doi:10.1186/1756-0381-7-24
Cite this article as: Rotroff *et al.*: PGxClean: a quality control GUI for the Affymetrix DMET chip and other candidate gene studies with non-biallelic alleles. *BioData Mining* 2014 **7**:24.