BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 14 2010, pages 1786–1787
doi:10.1093/bioinformatics/btq280

Genetics and population analysis Advance Access publication May 27, 2010

Evoker: a visualization tool for genotype intensity data

James A. Morris[1], Joshua C. Randall[2], Julian B. Maller[2] and Jeffrey C. Barrett[1,*]

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1HH and [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, UK

Associate Editor: Martin Bishop

ABSTRACT

Summary: Genome-wide association studies (GWAS), which produce huge volumes of data, are now being carried out by many groups around the world, creating a need for user-friendly tools for data quality control (QC) and analysis. One critical aspect of GWAS QC is evaluating genotype cluster plots to verify sensible genotype calling in putatively associated single nucleotide polymorphisms (SNPs). Evoker is a tool for visualizing genotype cluster plots, and provides a solution to the computational and storage problems related to working with such large datasets.

Availability: http://www.sanger.ac.uk/resources/software/evoker/

Contact: barrett@sanger.ac.uk

Received on March 31, 2010; revised on May 21, 2010; accepted on May 24, 2010

1 INTRODUCTION

Genome-wide association studies (GWAS) have recently transformed the landscape of complex human disease genetics by identifying hundreds of genes affecting risk for common diseases and traits (Hirschhorn, 2009). While these studies have become commonplace, the large volumes of data they produce still pose analytical and computational challenges. Among the most important of these is the importance of rigorous quality control (QC) procedures (de Bakker *et al.*, 2008) which help to reduce false positives arising from poor quality DNA, population structure, hidden confounders and genotyping artifacts. This last category, genotype calling artifacts, persist despite the considerable effort which has been invested in genotype calling algorithms (Korn *et al.*, 2008; Teo *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007) and automated QC filters. While filters like genotype call rate and Hardy–Weinberg equilibrium can flag a proportion of poorly clustered SNPs, the most reliable means of verifying the quality of genotypes is to visualize the genotype cluster plots for each SNP individually (Wellcome Trust Case Control Consortium, 2007).

While genotype cluster visualization is critical to GWAS QC, it remains challenging for many GWAS research groups to do so efficiently. The normalized intensity files from which genotype cluster plots are generated are extremely large and unwieldy in the default formats from SNP chip providers (uncompressed text-format intensities from a GWAS of 10 000 individuals would be hundreds of gigabytes). Extracting subsets of data and plotting hundreds of SNPs of interest is typically a tedious procedure requiring some computational sophistication. Moreover, the rapid proliferation of GWAS has exposed a general need for freely available, easy-to-use tools for routine tasks, as illustrated by the widespread adoption of programs like PLINK (Purcell *et al.*, 2007) for QC and association analyses. Therefore, we have developed Evoker, a Java program which supports two simple and compact binary data formats and is designed to make genotype cluster plot inspection an easy and efficient process.

*To whom correspondence should be addressed.

2 FEATURES

2.1 Overview

While many groups have implemented *ad hoc* genotype cluster plotting solutions (e.g. producing static images via the statistical software package R), they suffer from a number of drawbacks. First, because the source data files are enormous and cannot be stored locally on users' machines, the plots cannot easily be generated in an interactive fashion. Second, no convenient interface exists for standard GWAS tasks, such as rapidly scoring a large number of SNPs. Finally, static images provide little flexibility for exploring the details of genotype calling, especially in borderline SNPs. Evoker addresses these problems with remote data access via SSH, a streamlined workflow, and additional interactive features.

2.2 File formats and data access

Evoker requires four data file types, which provide information about samples, SNPs, genotype calls and X/Y allelic intensities. These can be formatted either as PLINK (Purcell *et al.*, 2007) 'binary pedigree' files or in a similar format used by the WTCCC2 project (Consortium and WTCCC2, 2009). Both of these formats utilize a binary format (rather than text) to reduce the size of the large genotype and intensity files, while preserving rapid indexing to any particular SNP. While these formats are substantially smaller than source text files, they are still too large to be practically stored on local workstations. Evoker can therefore connect to a remote server via SSH with appropriately formatted data. When accessing data remotely, Evoker only copies the relatively small sample and SNP information files while very large genotype and intensity files remain on the remote host. A helper perl script extracts genotype and intensity data for specific SNPs when requested by the client, and only this small slice of data is passed back via the SSH connection, keeping the client responsive.

2.3 Streamlined workflow

Upon loading data, Evoker will automatically recognize and load multiple sample sets present in the data directory (e.g. cases and controls, samples from different collections or samples that have undergone different experimental processes). This allows the user

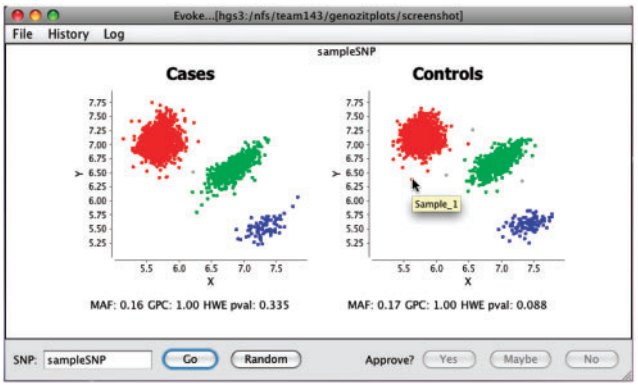

Fig. 1. Screenshot of Evoker running on Mac OS X displaying the intensity plots for a SNP in two collections.

to easily view multiple collections side by side to compare genotype calls across sample sets. Plots for can be quickly saved in the PNG graphic format for either individual collections or all displayed collections.

Plots for particular markers can be quickly called up by searching on the marker name, along with summary statistics on minor allele frequency, genotyping call rate and Hardy–Weinberg equilibrium *P*-value. A common application of the program is to view dozens or hundreds of SNPs showing evidence for association to rule out false positives caused by clustering artifacts. Although it is possible for the user to view all the SNPs they wish to using the simple search interface, the most efficient way to accomplish this task is using a SNP list. Users can load a simple text list of SNP IDs of interest to be plotted one at a time. The user can quickly view, assess and make a decision of yes, no or maybe for each SNP via buttons or keyboard shortcuts. These decisions are recorded in a separate file and the next SNP in the list is automatically displayed. Using SNP lists while connected to a remote data source is fast and responsive, because Evoker will download the data for all SNPs in the list in the background.

2.4 Additional features

Unlike most available cluster plot visualizations, Evoker provides a number of useful interactive advantages. Hovering over a particular point pops up a 'tooltip' with the corresponding sample ID (Fig. 1). Dragging a rectangle down and to the right zooms in on the highlighted area of the plot, and dragging up and to the left zooms back out. Finally, a history of the most recently viewed SNPs allows the user to quickly switch between different plots.

Evoker can also be used to visualize the effect that excluding certain samples (such as those on the borderline of particular QC thresholds) has on the genotype clusters. Sets of samples can be removed from cluster plots by loading simple text lists of sample IDs. Once an exclusion list is loaded, the samples in the list can be shown or hidden by toggling a checkbox. This feature can be a useful diagnostic of whether particular individuals have a consistent effect on the shape of genotype clusters, in which case the entire dataset may need to be recalled after their exclusion.

3 IMPLEMENTATION

Evoker is written in Java and will work on any platform with Java 1.5 or later installed. The Evoker bundle also includes a number of script written in Perl which require an installation of Perl 5.005 or later. The scripts included in the Evoker bundle enable users to easily create the files Evoker requires, such as binary intensity files, from widely used file formats such as CHIAMO (Wellcome Trust Case Control Consortium, 2007), Illuminus (Teo *et al.*, 2007) and Birdsuite (Korn *et al.*, 2008). Evoker is open source software under the MIT license, hosted at http://sourceforge.net/projects/evoker/ (details for downloading the source code are available on the web site).

4 CONCLUSION

Evoker is an open source program for visualizing genotype cluster plots designed to be integrated into QC workflows for GWAS. It provides a fast, user-friendly and interactive interface to these large and cumbersome datasets, enabling researchers without computational backgrounds to undertake this key QC step consistently while reducing the amount of time spent on the task.

Funding: Wellcome Trust (grant WT08912/Z/09/Z to J.A.M. and J.C.B.).

Conflict of Interest: none declared.

REFERENCES

UK IBD Genetics Consortium and WTCCC2 (2009) Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.*, **41**, 1330–1334.

de Bakker,P.I. *et al*. (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.*, **17**, R122–R128.

Hirschhorn,J.N. (2009) Genomewide association studies–illuminating biologic pathways. *N. Engl. J. Med.*, **360**, 1699–1701.

Korn,J.M. *et al*. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, **40**, 1253–1260.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Teo,Y.Y. *et al*. (2007) A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, **23**, 2741–2746.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.