Human Immunology 77 (2016) 283–287

Contents lists available at ScienceDirect

journal homepage: www.elsevier.com/locate/humimm

Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An integrated case-control analysis pipeline

Derek J. Pappas [a,*], Wesley Marin [b,c], Jill A. Hollenbach [b,1], Steven J. Mack [a,1]

[a] Center for Genetics, Children's Hospital & Research Center Oakland, Oakland, CA, United States
[b] Department of Neurology, University of California, San Francisco, CA, United States
[c] Molecular and Cell Biology, University of California, Berkeley, CA, United States

ARTICLE INFO

Article history:
Received 3 June 2015
Revised 13 October 2015
Accepted 17 December 2015
Available online 18 December 2015

Keywords:
BIGDAWG
HLA KIR data analysis
R package
Web app
Hardy–Weinberg testing
Case-control analysis
Amino-acid analysis
Haplotype analysis

ABSTRACT

Bridging ImmunoGenomic Data-Analysis Workflow Gaps (BIGDAWG) is an integrated data-analysis pipeline designed for the standardized analysis of highly-polymorphic genetic data, specifically for the HLA and KIR genetic systems. Most modern genetic analysis programs are designed for the analysis of single nucleotide polymorphisms, but the highly polymorphic nature of HLA and KIR data require specialized methods of data analysis. BIGDAWG performs case-control data analyses of highly polymorphic genotype data characteristic of the HLA and KIR loci. BIGDAWG performs tests for Hardy–Weinberg equilibrium, calculates allele frequencies and bins low-frequency alleles for $k \times 2$ and 2×2 chi-squared tests, and calculates odds ratios, confidence intervals and *p*-values for each allele. When multi-locus genotype data are available, BIGDAWG estimates user-specified haplotypes and performs the same binning and statistical calculations for each haplotype. For the HLA loci, BIGDAWG performs the same analyses at the individual amino-acid level. Finally, BIGDAWG generates figures and tables for each of these comparisons. BIGDAWG obviates the error-prone reformatting needed to traffic data between multiple programs, and streamlines and standardizes the data-analysis process for case-control studies of highly polymorphic data. BIGDAWG has been implemented as the *bigdawg* R package and as a free web application at bigdawg.immunogenomics.org.

1. Introduction

The extensive polymorphism, linkage disequilibrium and genotyping ambiguity commonly associated with the HLA and KIR loci (described here collectively as *immunogenomic* loci) pose challenges for the consistent analyses of these data [1]. Modern genetic analysis programs are designed for use with bi-allelic single nucleotide polymorphisms (SNPs) or SNP haplotypes generated in genome-wide association studies (GWAS), but cannot be applied to highly polymorphic immunogenomic data. New tools are needed to leverage modern computational resources for the analysis of immunogenomic data, and to integrate the analysis of immunogenomic loci with genomic SNP/GWAS data. The few Ad-hoc tools designed to handle immunogenomic data, such as PyPop [2] and Arlequin [3] are limited by operating systems, outdated with spurious maintenance cycles, and often times require cumbersome data formatting.

A typical immunogenomic data analysis workflow involves the trafficking of data between several programs; this usually involves reformatting of these data for each program, a process that is time intensive, error prone and limits reproducibility. Quite often, this data-trafficking involves the use of Microsoft Excel, which is particularly poor choice for immunogenomic data-management [1]. In addition, the management of data in a typical workflow is often idiosyncratic to the analyst, which further limits reproducibility across studies. The automated manipulation of immunogenomic data in a single analysis workflow will reduce errors and allow true analytical reproducibility.

Abbreviations: ALD, asymmetric linkage disequilibrium; BIGDAWG, Bridging ImmunoGenomic Data-Analysis Workflow Gaps; BWA, BIGDAWG web application; DOF, degrees of freedom; GWAS, genome wide association study; HLA, human leukocyte antigen; HWEP, Hardy–Weinberg equilibrium proportions; IMGT, ImMunoGeneTics; IPD, immuno polymorphism database; KIR, killer-cell immunoglobulin-like receptor; LD, linkage disequilibrium; SNP, single nucleotide polymorphism.

* Corresponding author at: Children's Hospital Oakland Research Institute, 5700 Martin Luther King Jr Way, Oakland, CA 94609, United States.
E-mail address: djpappas75@gmail.com (D.J. Pappas).
[1] Contributed equally to manuscript.

http://dx.doi.org/10.1016/j.humimm.2015.12.006
0198-8859/

We have developed Bridging ImmunoGenomic Data-Analysis Workflow Gaps (BIGDAWG), an automated software pipeline that performs a suite of common case-control analyses of multi-locus highly polymorphic genetic data [4–6]. Unlike SNP/GWAS case-control analysis tools, BIGDAWG is tailored for use with immunogenomic data. In addition, BIGDAWG can be applied to any highly polymorphic genetic data, including SNPs and SNP haplotypes. BIGDAWG is implemented as an R package (named, *bigdawg*) and as a web application running at bigdawg.immunogenomics.org.

2. Methods

2.1. Implementation

BIGDAWG has been developed in the framework of the R statistical environment (http://www.r-project.org). The *bigdawg* R package provides documentation of all BIGDAWG functions, and includes a vignette detailing package use along with a sample dataset. The *bigdawg* vignette is included here as Supplementary Material. BIGDAWG's functionality depends on the *epicalc* [7] and *haplo.stats* [8] R packages, along with the R base package *parallel*. The R *XML* package [9] is required for updating the protein alignment object to adhere to the most current IMGT/HLA Database [10] (version 3.20.0 released 2015-04-17 as of this writing). The *bigdawg* R package (version 1.1) is covered under the GNU general public license version 3 or higher and has been made available through the Comprehensive R Archive Network (CRAN) repository.

The BIGDAWG web application (BWA) is a *shiny* [11] implementation of the *bigdawg* R package. As such, BWA requires only a modern web browser and Internet access to function, and does not require the R environment to be installed on a user's system. BWA input data and analytical parameters (described in Section 2.5) are specified in the user's web-browser, and the results files can be downloaded from the browser.

2.2. Functions

BIGDAWG accepts unambiguous genotype data for case-control groups as input, and calculates allele frequencies for chi-square (χ^2) testing, along with odds ratios, confidence intervals and *p*-values for each allele (for a processing flowchart see Supplementary Material Fig. 1). BIGDAWG combines rare alleles into a common class ("binning"; see Section 2.3) which are included for testing, performs overall locus-level ($k \times 2$) tests of significance, followed by a series of allele-level (2×2) tests of significance for each locus. In addition, the control group is tested for deviations from expected Hardy–Weinberg equilibrium proportions (HWEP) at the allele level. When multi-locus genotype data are available, BIGDAWG estimates user-specified haplotypes and performs the same binning and statistical calculations for each haplotype [$k \times 2$ tests at the multi-locus level (e.g. HLA-A–HLA-B or HLA-DR B1–HLA-DQA1–HLA-DQB1) followed by 2×2 tests at the haplotype level]. For HLA data, BIGDAWG integrates protein sequence alignments from the IMGT/HLA database to run case-control association tests on individual amino-acid positions within exon 2 and exon 3 (class I) or exon 2 (class II) ($k \times 2$ tests for each polymorphic amino-acid position, followed by 2×2 tests for each amino-acid residue). For these amino acid analyses, HLA allele names must conform to the colon-delimited HLA allele name nomenclature as defined by the WHO Nomenclature Committee for Factors of the HLA System in April 2010 [12].

2.3. Statistics

All HWEP and phenotype association (haplotype, locus and amino acid) analyses are currently based on a traditional χ^2 test. For HWEP deviation testing, BIGDAWG combines rare genotypes into a single common class (binning) for analysis and performs a goodness-of-fit test. The degrees of freedom (dof) are calculated as $\text{dof} = g - (a - 1)$, where g is the number of unique non-binned genotypes and a is the number of unique non-binned alleles.

For testing phenotype associations, BIGDAWG runs a test-of-independence, automatically tabulating the $k \times 2$ contingency tables, where k is the number of unique haplotypes, alleles or amino acids. For either testing scenario, rare cells (with expected counts less than five) are combined into a common class (binned) prior to computing the χ^2 statistic, except in cases of the test-of-independence where all cells of a given $k \times 2$ contingency table are $\geqslant 1$ and fewer than 20% of the cells have expected counts less than five. BIGDWG's haplotype estimation function requires the R *haplo.stats* package, whereas calculation of the individual haplotype/allele/residue confidence intervals, odds ratios, and *p*-values requires the R *epicalc* package.

2.4. Input and output data structures

BIGDAWG input files are tab delimited text files with columns for subject IDs, phenotype association analysis (labeled 1 or 0), and column pairs of unambiguous, unphased alleles for each locus. Allele names can be of any format (e.g., 1, 2, 3, a, A, b, B, s, S, t, T, p, P, q, Q, etc. can be supplied as allele names). For HLA data, allele names (with or without a locus prefix) can include from a single field up to the full length name for a given allele (e.g., "01", "01:01". "01:01:01" and "01:01:01:01" are all recognized as valid alleles). BIGDAWG treats the absence of a locus (e.g., resulting from structural variation) as an allele of that locus, and recognizes "00:00" as a convention for identifying absent loci. This can be especially relevant for HLA loci such as *HLA-DRB3*, *HLA-DBR4*, *HLA-DRB5* as well as members of the *KIR* gene family, where locus absence may be informative and associated with the pertinent phenotype.

After input data have been read, BIGDAWG provides a short summary of the relevant architecture of the supplied data (e.g., the number of unique alleles and the number of instances of missing data at each locus), and runs a set of data consistency checks to ensure the most compatible data set for analysis (e.g., identifying large-scale discrepancies between the number of HLA allele-name fields in case and control groups). An example of this summary is shown in Fig. 1. The *bigdawg* vignette, included in the Supplementary Material, provides more detailed description of input file requirements.

Summaries of each analysis are displayed on the R console/terminal window (Fig. 2), or web-browser pane. However, all analytical results are recorded as tab delimited text files, which include more detailed descriptions of each analysis. In addition, each BIGDAWG analysis generates a "run parameters" file identifying the options used in that run, allowing each analysis to be reproduced. Descriptions of each BIGDAWG result file are included in the Supplementary Material as part of the *bigdawg* vignette.

2.5. Parameters

BIGDAWG offers considerable flexibility in the selection of parameters for running an analysis. Users can specify individual levels of analysis (for Hardy–Weinberg ("HWE") or for case-control at the haplotype ("H"), locus ("L") or amino-acid ("A") levels) or combinations of these tests (data permitting) using the Run.Tests parameter in the *bigdawg* R package, or using checkboxes for