WGAViewer: Software for genomic annotation of whole genome association studies

Dongliang Ge,1,2,5 Kunlin Zhang,3 Anna C. Need,1 Olivier Martin,4 Jacques Fellay,1,2 Thomas J. Urban,1,2 Amalio Telenti,2,3 and David B. Goldstein1,2,5

1 Center for Population Genomics & Pharmacogenetics, Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina 27708, USA; 2 Center for HIV/AIDS Vaccine Immunology, Duke University, Durham, North Carolina 27708, USA; 3 Institute of Microbiology, University Hospital, University of Lausanne, 1011 Lausanne, Switzerland; 4 Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

To meet the immediate need for a framework of post-whole genome association (WGA) annotation, we have developed WGAViewer, a suite of JAVA software tools that provides a user-friendly interface to automatically annotate, visualize, and interpret the set of P-values emerging from a WGA study. Most valuably, it can be used to highlight possible functional mechanisms in an automatic manner, for example, by directly or indirectly implicating a polymorphism with an apparent link to gene expression, and help to generate hypotheses concerning the possible biological bases of observed associations. The easily interpretable diagrams can then be used to identify the associations that seem most likely to be biologically relevant, and to select genomic regions that may need to be resequenced in a search for candidate causal variants. In this report, we used our recently completed study on host control of HIV-1 viral load during the asymptomatic set point period as an illustration for the heuristic annotation of this software and its contributive role in a successful WGA project.

[Supplemental material is available online at www.genome.org. WGAViewer is available at http://www.genome.duke.edu/centers/pg2/downloads/wgaviewer.php.]
hyperlink. This annotation also displays concurrent evidence from multiple databases for convenient comparison. Third, the *individual SNP annotation* annotates the associated SNP with a graphical representation of its LD score with all HapMap SNPs in a specified surrounding region (The International HapMap Consortium 2005), so that non-genotyped associated SNPs in surrounding genes can be easily identified. This class of annotation also tests whether associated polymorphisms show any association with the expression (in immortalized B-lymphocytes) of the closest (or an alternatively specified) gene currently represented in the GENEVAR database (Stranger et al. 2007). This feature will be expanded as additional data become available, for example, to include data on both expression in other tissue types and alternative splicing. Displayed alongside this plot is additional SNP-related information, such as ancestral allele, concurrent evidence from multiple databases, SNP function (synonymous, non-synonymous, splice-site, etc.), validation status, and other identifiers. Fourth, the *gene/SNP finding annotation* provides a convenient way to locate and annotate candidate genes/SNPs of specific interest in a WGA project, and to align with the physical coordinates from the latest genome build. Moreover, this annotation enables a search for specific SNPs, and if they are not present in a WGA project, it then displays the results for the SNPs that are associated at a user-specified level (their “LD proxies”). These functions make a handy and reliable comparison with existing reports, for example, the reports from previous candidate gene studies. When specific SNPs of interest have not been genotyped in a WGA project, these annotations for searching LD proxies are especially convenient and valuable. Fifth, *concurrent evidence from multiple databases* allows the user to load multiple databases simultaneously, with one of them as the “core” database to be considered. The supporting databases could include replication studies, projects with related phenotypes, other publicly available projects, even studies with different marker sets or different phenotypes. These data sets can then be listed and plotted alongside the core database as concurrent evidence. WGAViewer provides a convenient way to select, sort, and annotate the SNPs with evidence at specified significance level across all the core and reference databases. Finally, *user-customized supporting/QC databases* can be loaded by WGAViewer to facilitate the interpretation of the WGA results, including important information such as HWE *P*-values, effect size, effect direction, QC scores, or other user-customized data.

As an illustration of the heuristic annotation of WGAViewer and its contribution to a real WGA project, we use a recently completed study on host control of human immunodeficiency virus-1 (HIV-1) viral load during the asymptomatic set point period (Fellay et al. 2007). A screenshot from WGAViewer (Fig. 1) shows the individual SNP annotation results for rs9264942 on chromosome 6 spanning from 31,182 kb to 31,582 kb. First, an overview of chromosome 6 shows two genome-wide significant SNPs (plotted in red points). The further annotation then shows that one of the two hits, rs9264942, is located in the 5′ region of the *HLA-C* (major histocompatibility complex, class I, C) gene and is ~35 kb away from the transcription initiation site. Most strikingly, WGAViewer explicitly shows that this SNP shows a high correlation with the expression levels of *HLA-C*, which is the closest gene to this SNP. These annotations immediately lead to a hypothesis that rs9264942 may function through controlling *HLA-C* expression level and thus point to a most promising direction for functional investigations. However, it should also be noted that in the case of a highly variable gene like *HLA-C*, apparent expression effects could result from either an association with expression or an association with variant sites affecting probe binding (T. Urban, W. Yoon, K.V. Shianna, J. Fellay, D. Ge, B.H. Haynes, A.J. McCMichael, M. Carrington, A. Tenenti, and D.B. Goldstein, unpubl.). Furthermore, these annotations also enable the convenient identification of SNPs in LD with rs9264942, which are associated with *HLA-C* expression level and HIV-1 viral load, for example, rs2249742 (*r*^2^ = 0.74, 1 kb upstream). Finally, in addition to the individual result as shown in Figure 1, this class of annotations can be sequentially performed for a number of ranked SNPs at a specified cutoff.

**Discussion**

In this report we introduce WGAViewer, a JAVA software tool that provides a user-friendly interface to annotate and help interpret the set of *P*-values emerging from a whole genome association (WGA) study. Using the six classes of annotation that are provided by this software, different pipelines for annotation can be employed to interpret the results from a WGA project (see Supplemental material, WGAViewer User’s Guide). Most importantly, this annotation environment provides a convenient and heuristically accessible summary of information that is relevant to the interpretation of association results. In our own experience, carrying out simple steps such as asking whether an associated polymorphism is in linkage disequilibrium with a functional polymorphism or whether it associates with the expression of nearby genes can take an unreasonably long time to reconcile necessary databases and make the comparisons. For example, when a specific polymorphism is implicated in an association study, an obvious and important immediate question is whether the polymorphism associates with expression levels of nearby genes. One data set available for performing this is the GENEVAR data set (Stranger et al. 2007), but extracting the necessary data from GENEVAR to perform the test is laborious. Specifically, it is necessary to first search for the closest gene to target SNPs using the latest genome build (for example, using the Ensembl database). Secondly, one has to search for the probes for this gene in the GENEVAR database and retrieve the individual normalized expression data in each HapMap cell line in the appropriate population. Thirdly, one has to search for the HapMap database to retrieve the SNP genotype for each individual. And finally, one has to merge these data from disparate data sources and perform an association test. In our experience, these steps can take 75 min on average. WGAViewer, however, dramatically expedites these classes of interpretations in an automated manner and presents results in heuristic graphs. Using this software, it is extremely convenient to search for candidate genes, SNPs, and their LD proxies, requiring no input for genome coordinates or other information. The work session, along with downloaded annotation and interactive features, can be saved and reopened at a later date.

As such, this software tool can be viewed primarily as expediting the work that would always follow the discovery of strong associations with a phenotype of interest. Most of the features to be found in WGAViewer emerge directly from real studies in the lab where we asked specific questions for specific discoveries and found that the process would substantially benefit from automation (see Supplemental Table 1).

In addition to these expediting annotations, there is, however, one class of annotation that goes beyond convenient summaries of work that could be carried out manually and individually
ally. In our own experience, the manual annotations are not feasible beyond the few top discoveries. What happens with a polymorphism that achieves a P-value of only $10^{-5}$, but which is itself in strong LD with an ungenotyped SNP that is thought to have definite functional consequence? This SNP would warrant special consideration, but could be missed in most manual settings. We have established a comprehensive annotation (1–5 min per SNP) meant to pluck out this kind of suggestive associations, and this is one of the few features of WGAViewer that aims toward real automated consideration of the full sets of results, as opposed to only expediting and summarizing analyses. This function features a serial annotation for a set of SNPs ranked by their P-values. In addition to the gene context, LD extension, and expression annotation routines for each SNP, this process will also check and filter the functions for both the original genotyped SNPs and their LD proxies. Once the annotation is done, the interactive filtering and other annotation features can be saved for convenient later use.

The rapidly increasing number of genome scans using identical SNP sets allows assessment of common genetic determinants across different diseases and phenotypes. In order to facilitate searches for polymorphisms that impact multiple pheno-
types, we have incorporated features for comparing the evidence across databases, including between original findings and replication cohorts, between studies in various populations, between original findings and related phenotypes, or between original findings and results from different studies. For these purposes, WGAViewer has been designed to read and annotate multiple databases and to identify all SNPs that show a user-specified degree of association in two or more studies. WGAViewer provides a free and convenient platform to access and annotate all the WGA databases hosted by and released from the Duke Institute for Genome Sciences & Policy, powered by the Mart for IGSP Data from Association Studies (an instance of BioMart; Kasprzyk et al. 2004). In addition to this, WGAViewer also provides a convenient interface that allows a user to load and list different kinds of supporting information that may facilitate the interpretation.

WGAViewer is an ongoing project under continuous development. Our future development plans currently include: transcription factor binding sites, miRNA sites, comparative genetics, and splicing regulators. We welcome suggestions and collaborations.

We view the publicly available WGAViewer tool as a first step toward facilitating the interpretation of genome-wide association studies in the context of the rapidly evolving knowledge of the human genome and the disparate databases that house that knowledge.

Methods

Annotation implementation

WGAViewer is a stand-alone software developed using the JAVA language by Sun Microsystems (http://java.sun.com/) and an integrated development environment (IDE) NetBeans (http://www.netbeans.org/). WGAViewer was compiled and should be used in a computer system with a JAVA environment with version 1.5.0_07-b03 or later. The current version has 186 JAVA classes and is released in one executable JAR file. Microsoft Windows users have an option to launch the program through a wrapper program: WGAViewer.exe, while other users can start the program through a shell program, WGAViewer.sh. We detail the scheme of the WGAViewer annotation methods in Supplementary Figure 1.

Alignments between genome builds

Instead of using a fixed genome build version as the genomic coordinate source, we chose to always apply the latest genome build from Ensembl at the time of each annotation, and map every coordinate from other sources to this dynamic core database. This dynamic procedure helps to interpret the WGA results using the most updated transcripts and SNP coordinates, and avoids discrepancy between different major builds used in different sources (for example, build 36 and build 35), or even between different subversions (for example, build 36: Ensembl version 46.36h and 43.36e). This procedure is implemented in all the annotation routines, most importantly, in the gene and SNP searching functions. We noticed that sometimes these discrepancies could be up to –200 kb, which could potentially lead to very different gene/SNP contexts and hence result in different interpretations and hypotheses. Therefore, instead of directly using genome coordinates from diverse sources, we have treated the genetic markers as anchors and always used a fast hashtable-matching to map these anchors into the uniform Ensembl build version. This uniform build version, together with the version from other sources, is stored in the annotated project file and can be referred to in the later interpretation.

Testing

We have used and tested WGAViewer on typical Windows- and Linux-based PCs (2-GHz processor, 2 GB of memory). A brief annotation for ranked hits without graphical features can take 0.5–1.0 sec per SNP depending on Web connection speed. A comprehensive annotation with a spanning region –200 kb can take 1–5 min per SNP. For a typical WGA result set with ~50,000 SNPs, WGAViewer requires at least 512 MB of memory to load and store the annotation results, with a tool for monitoring memory usage, and will alert the user when the system is about to run out of memory. WGAViewer does require that the user computer system has the ability to establish a standard MySQL database connection, which is through the most common MySQL port 3306 and 5306. In cases where such connections cannot be directly established, WGAViewer provides a substitute proxy connection through the servers hosted in the Duke Institute for Genome Sciences & Policy.

WGAViewer has been applied in a number of projects (Cavalleri et al. 2007; Fellay et al. 2007). A data set based on host control of HIV-1 viral load has been included in the released package as an example.

Acknowledgments

We thank Darin London, Robert J. Wagner, and Mark R. DeLong from the Duke Institute for Genome Sciences & Policy for their work on setting up the query engine of the Mart for IGSP Data from Association Studies (MIDAS), provided by the Biomart Data Integration System (http://www.biomart.org/). WGAViewer is supported by the Center for HIV-AIDS Vaccine Immunology (CHAVI, http://www.chavi.org/) and the Duke Institute for Genome Sciences & Policy (IGSP, http://www.genome.duke.edu/).

References


Received September 25, 2007; accepted in revised form January 25, 2008.
WGAViewer: Software for genomic annotation of whole genome association studies

Dongliang Ge, Kunlin Zhang, Anna C. Need, et al.

Genome Res. 2008 18: 640-643
Access the most recent version at doi:10.1101/gr.071571.107

Supplemental Material
http://genome.cshlp.org/content/suppl/2008/03/04/gr.071571.107.DC1

References
This article cites 10 articles, 3 of which can be accessed free at:
http://genome.cshlp.org/content/18/4/640.full.html#ref-list-1

License

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.