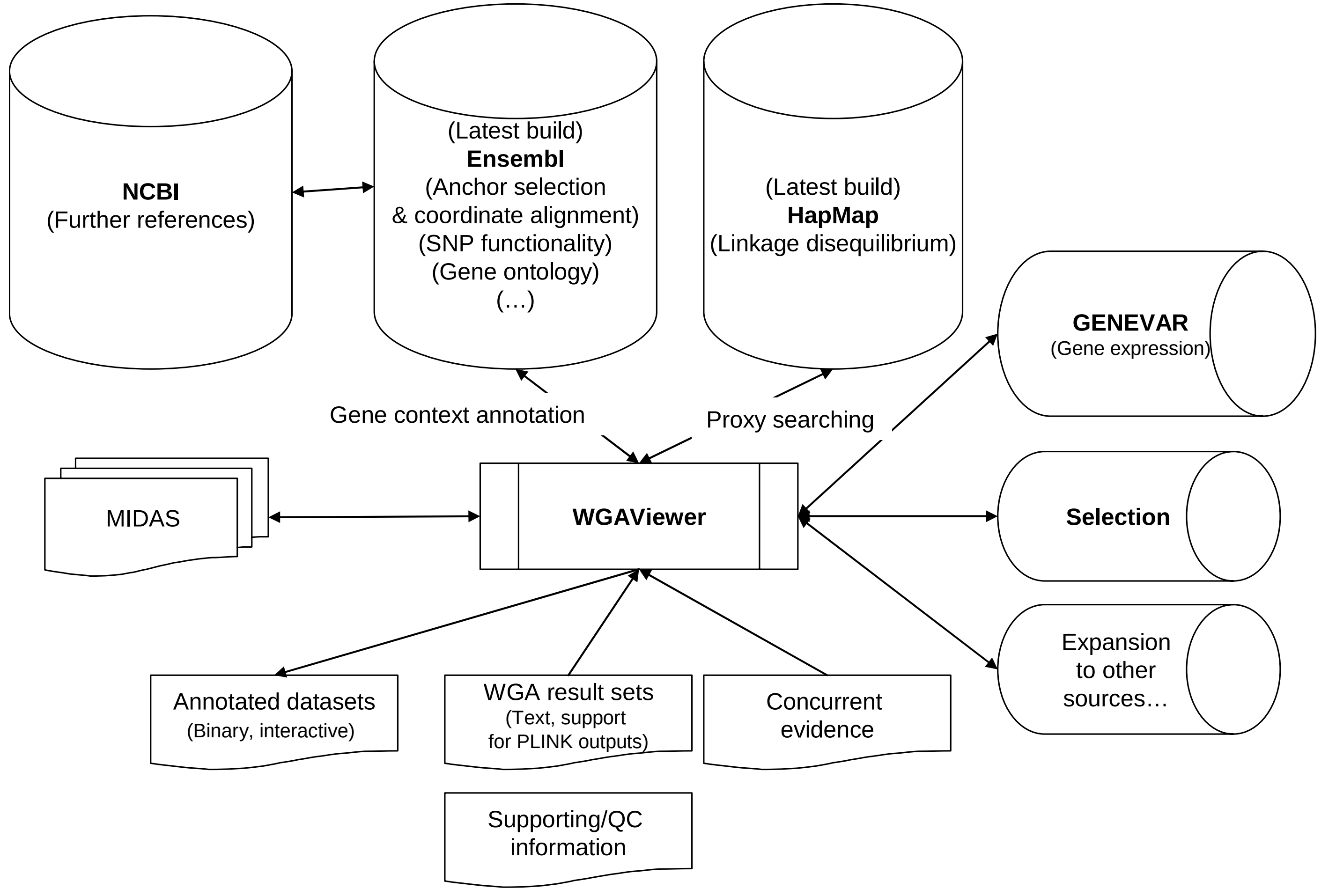
WGAViewer: A Software for Genomic Annotation of Whole Genome Association Studies

Supplementary Figure 1. Scheme for WGAViewer annotation.

The anchor selection and coordinate alignment procedures are accomplished through the MySQL connection and queries (http://www.mysql.com/) to the remote Ensembl databases (http://www.ensembl.org). We access the HapMap individual genotype data and perform the LD proxy searching through a standard Hypertext Transfer Protocol connection with HapMap database (http://www.hapmap.org). Gene expression and selection data are accessed through a local database connection, with the ability to expand to further sources when available. An association test between HapMap genotype and GENEVAR gene expression data was implemented in WGAViewer. In addition to the local annotations, WGAViewer also provides the access to further reference databases through hyperlink, including PubMed and dbSNP from the National Center for Biotechnology Information (NCBI). All the annotation results and interactive features are stored to a serialized JAVA object and can be reopened at a later date.

Cylinders denote various sources of bioinformatic databases. Arrows represents the data flow direction. NCBI: The National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/); GENEVAR: Gene Expression Variation (http://www.sanger.ac.uk/humgen/genevar/); MIDAS: The Mart for IGSP Data from Association Studies (http://midas.genome.duke.edu/biomart/martview);
QC: quality control. PLINK is a whole genome association analysis toolset developed by Shaun Purcell in Harvard University (Purcell et al. 2007).



Supplementary Figure 1.

Supplementary Table 1. Questions leading to summary features implemented in the WGAViewer software

What are the top hits and their p values?
Are these top hits located in or near any gene?
If they are located in a gene, what type of SNPs are they? Are they of known function?
If they are not non-synonymous coding SNPs, nor located in a known splice site, how far are they from the closest exon?
If they are not in a known gene, how far are they from the closest known gene?
What exactly is the genic context for each hit? What are the surrounding genes?
Is there any evidence for evolutionary conservation/selection of the surrounding region?
What are the p values of the surrounding SNPs?
What is the LD context among these SNPs?
How far does the LD extend for each hit? Does this LD extension cover other genes?
Are there (perhaps ungenotyped) proxies for the associated SNP that are in a more interesting genomic context?
Do these hits or their proxies show any association with available functional data, for example, gene expression levels?
After all, is there a way to conveniently annotate these hits in an automatic and batch manner? Is there a way to automatic filter their proxies by their function?
There are many candidate gene studies published on the same or related phenotypes. What are the p values for SNPs in and around these associated genes in our WGA project? Can we replicate previous findings?
Can we replicate previous associations of particular SNPs?
If the previously -associated SNPs have not been included in our WGA project, are there any correlated proxies or tags for these candidate SNPs? What are their p values?
Is there evidence of population stratification effects?
We have association data for replication cohorts. There are also cohorts with related but not identical phenotypes. Is there a way to compare them easily?
We have genome-wide HWE test results. We have effect size, effect direction, etc. Is there a way to list them alongside our association findings?
We don’t have a WGA set. But I want to annotate a SNP in such a way too. I also want to test LD among a list of SNPs. I want to test SNP-gene expression associations. Are there any convenient bioinformatic tools that WGAViewer can offer?

REFERENCES

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, and P.C. Sham. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet.* 81.