Zbrowse

Greg Ziegler

January 8, 2015

Contents

1 Description

Zbrowse is an interactive GWAS results viewer and web application that is an extension of the classic GWAS Manhattan plot. It allows for the rapid comparison of GWAS results from multiple phenotyping experiments and the rapid viewing and analysis of genes under peak SNPs. The tool currently supports the viewing of GWAS results in *Arabidopsis thaliana*, maize, and soybean, and we provide instructions to extend it to other organisms.

2 Introduction

The recent development of high-throughput phenotyping techniques coupled with the ability to genotype large populations of diverse individuals has revolutionized the way that forward genetics research is performed. Tools have rapidly become available to perform genome-wide association studies (GWAS) in a variety of species that can map traits to the genome with high enough resolution to quickly provide a tractable list of potential causal genes.

The result of a GWAS experiment is a calculation of the probability (P-value) of an association between each genotyped SNP and the observed phenotype.

The Manhattan plot is the standard way of viewing these datasets. The plot allows for the visualization of the entire dataset on a simple XY scatter plot. In a Manhattan plot, the genomic coordinate for each SNP is displayed along the X-axis and the negative logarithm of the P-value is displayed on the Y-axis. Plotted in this way, each SNP is visualized in its genomic context and highly associated loci form peaks that tower over the rest of the SNPs.

One of the first steps to analyzing GWAS results is determining what gene or genes fall under the SNP peaks that can be seen on the Manhattan plot. Analyzing GWAS data quickly becomes tedious when analyzing multiple phenotypes or relatively complex traits. One of the main goals of the browser described here is to find genotype-environment (GxE) interactions (e.g., those instances where an environmental condition causes a phenotypic effect, but only for individuals with a given genotype). Additionally, unexpected pleiotropic effects of a locus may become apparent when multiple phenotype GWAS results can be viewed easily on a single plot.

3   Installation

Requirements

- R (tested on R versions > 3.1)

- R Studio is recommended but not required to run Zbrowse.

Standard Installation

1. Navigate to http://www.baxterlab.org/

2. Click "Download Zbrowse"

   - This will download a zipped file of the most recent Zbrowse version to your computer

3. Unzip the file and open `interactiveGWASuploadable.Rproj`

   - If you haven't installed Rstudio, you may need to right-click the file and open with 'R'

4. At the R command prompt enter: `source("startup.R")`

   - This will check for and install any necessary libraries and open Zbrowse in your default browser.

Advanced Installation

Zbrowse can also be installed by cloning the Zbrowse git repository

Update Current Installation

The easiest way to update your current installation is to simply download the newest version of Zbrowse by following the instructions under Standard Installation.

To copy any saved datasets and column settings to the new installation, replace the files and folders in the `config` folder from the new installation with those from the old installation.

4 Quick Start Guide

Prepare GWAS Dataset

Your GWAS dataset is likely already in a comma-separated format. If not, open your dataset in spreadsheet software such as Excel and save as csv. Zbrowse is designed for plotting the most significant hits from a GWAS experiment and is therefore limited to plotting less than 5000 markers for a single trait. Most GWAS results files contain results from every marker, so it may be necessary to filter the non-significant markers before uploading to the browser.

Zbrowse also supports plotting of GWAS data which has interval ranges, such as Joint Linkage support intervals.

Load dataset into browser

In the browser, select the `.csv` radio button. Most likely you have a header row in the csv file for variable names, if not, deselect the `Header` check box. If the data are not comma separated you can also choose a semicolon or tab separated format. Select the organism from the dropdown list that your GWAS experiment was performed on. Then click `Choose file` and locate the csv file on your computer.

Select Appropriate Columns

The manage tab gives a look at the first ten rows of your data and provides selection boxes to tell the plotting software which columns to use for graphing. Note that traits may be a composition of multiple columns (e.g. a trait and a location), so multiple columns can be selected in the trait box. Select appropriate columns, then click the `Submit` button to switch to the `Whole Genome View` tab to look at your data.

Manhattan Plots

Navigation through the graphs and tables in Zbrowse is done using the tabs at the top of the page, adjusting options on the sidebar panel, or clicking points on the plots. The user interface is designed to be intuitive and allow the user to quickly zoom into a point of interest anywhere on the genome. Please try navigating through the various tabs to gain a feel for the browser functionality or read on for a more in depth explanation of all of the browser functions.

5 Troubleshooting and Frequently Asked Questions

- The browser isn't loading
  - Check your R version and package versions
    * In the console type: `sessionInfo()` to see your R version and the versions of loaded packages.
    * We have extensively tested on R version 3.1.2. Some features may not work with older versions of R.
    * Required packages have been tested using these package versions: rCharts (0.4.5) and shiny (0.10.2.2). These packages are under active development and older or newer versions often are not compatible. Please let us know if you have a newer version of a package that is not working with the browser.
  - Try using a different browser. Most of our testing was on the most recent version of Firefox.
  - If the above steps don't work please email us at baxterlabZbrowse@danforthcenter.org and include your `sessionInfo()` output as well as the browser or browsers you have tried using.
- I don't see any points in the plot
  - Check that you have the appropriate columns selected
  - Make sure you have points that fall within the Y-axis range
  - Make sure you aren't trying to plot more than 5000 points for a single trait

6 Documentation

6.1 Browser UI and Navigation

Zbrowse utilizes a tab-based navigation format to make accessing different aspects of the browser fast, efficient, and intuitive. There is a sidebar panel on the left of the page that updates with a set of options specific to the tab being displayed.

Manage Tab

The first tab in the list, and the landing page when the application is first loaded, is the `Manage` tab. This tab allows a new GWAS dataset to be uploaded into the application or a pre-loaded dataset from a dropdown menu can be selected. Data to be uploaded can be in a flat file delimited with either commas, semicolons or tabs or an RData object. Before uploading, also select the organism from which the GWAS data being uploaded comes.

Once uploaded, a preview of the first ten rows of the dataset will appear in the main panel. Below this table is a series of selection boxes that allow the user to specify which columns in the file to use for plotting. Select a chromosome and base pair for determining the location of each SNP in the genome. Currently,

only numeric chromsome values are allowed. If the uploaded dataset is data from only one GWAS trait, there is a checkbox to include all data as one trait. Otherwise, the user can select one or more trait columns to group the data by when plotting. For example, you might be interested in comparing NAM GWAS results from multiple NAM growout locations, or in comparing results from multiple GWAS experiments performed on different traits, or both multiple locations and multiple traits; simply select the column or columns with the label for the trait that the SNP corresponds to. Finally, select the y-axis column with the significance value against which to plot each SNP. Usually, this is the negative logarithm of the P-value, but can also be the number of bootstrap models that include this SNP or any other measure of trait significance, such as effect size. The final parameter allows for user selectable values for the Y-axis, if left unset, the software will automatically scale the y-axis based on the range of the selected data.

Once all of the correct columns have been selected, click submit to be brought to the `Whole Genome View` tab. Once submitted, the software will remember the selected settings for this dataset on future visits and automatically populate the fields with the previously selected values.

Finally, the `Save Current Dataset` button on the sidebar will save the dataset to the `config` folder of Zbrowse so it will be available from the dropdown list everytime the browser is launched.

Data Table Tab

The `Data Table` tab provides a fast way to browse the dataset you have uploaded in table form. The table is sortable and searchable and can also be exported using the buttons at the top of the page. However, note that when using the `copy` and `save` buttons at the top of the table, the exported table will only contain what is displayed on the current page. To export a large table, change the number of rows viewed to a larger number (maximum 500).

Whole Genome View Tab

The plot on this tab is formatted as a standard genome-wide Manhattan plot. The x-axis is ordered by chromosome and base pair within each chromosome. The background of the plot has alternating blue/white shading for the even and odd chromosomes to make discerning chromosome breaks simple. The panel on the left contains a box for each trait column selected in the Manage tab. Each of these boxes is populated with the values found in that column and any combination of one to many traits can be selected for viewing on the plot on the right.

Often the goal of analyzing multiple GWAS experiments on a single plot is to look for markers that overlap between locations or phenotypes. When many points are plotted and being viewed at the whole genome level it can be difficult to see which points actually overlap or are within a certain distance from one another. To filter for only SNPs that overlap between traits, check the `Show only overlapping SNPs` box. This will bring up two additional options, the `Overlap size around each point` specifies a window size around each point, if a point has no other points within its window, it will be removed from the

plot. The minimum number of traits that need to overlap can also be specified. If set to one, then overlaps of a trait with itself will be considered.

The plot contains a variety of interactive features. In addition to selecting traits to view in the sidebar panel, traits can be quickly hidden by clicking their entry in the legend. When many points are plotted on the same graph, overplotting can make it difficult to discern points clustered around the same peak. To alleviate this, the plot can be easily zoomed by clicking and dragging a zoom box anywhere in the plot. This makes it much easier to see the relationship between tightly grouped points. When you scroll over points in the plot, a tooltip will display that shows information about the trait that SNP is associated with, the Y-axis value, and the exact chromosome and base pair for the SNP. If the tooltip gets in the way of the viewing or selecting of points, clicking the plot will temporarily hide the tooltip box.

Clicking any point in the `Whole Genome View` tab will change tabs to the `Chromosome View` tab.

Chromosome View Tab

This tab contains two plots: one is a chromosome-wide view displaying the data from the chromosome clicked in the genome-wide view, the other plot is an annotation plot of the region around the clicked base pair. A blue band in the chromosome-wide plot highlights the region being displayed in the annotation plot. The displayed chromosome can be changed without returning to the `Whole Genome View` tab using the drop-down menu in the sidebar panel. Similar to the genome-wide plot, points can be clicked to redraw the annotation plot around new points of interest.

The annotation plot is a variable width plot that defaults to showing the region 250,000 base pairs on either side of the point of interest. The width of this region can be adjusted between 1,000 and 500,000 base pairs using the slider on the sidebar panel. The bottom of this plot has a track that shows the position of coding sequences around the gene of interest on both the forward and reverse strand. The tooltip for genes in this track displays information about the gene location, strand, and function, if known. Clicking on the gene will open a new browser tab that links to the gene description page specific to the organism being viewed. Arabidopsis links to The Arabidopsis Information Resource(TAIR), soybean links to Soybase, and maize links to the Maize Genetics and Genomics Database (MaizeGDB), other organisms link to the Phytozome page for that gene.

Annotations Table

The `Annotations Table` tab provides a fast way to browse and search the genes around the currently selected point. The table is sortable and searchable and can also be exported using the buttons at the top of the page. Note that when using the `copy` and `save` buttons at the top of the table, the exported table will only contain what is displayed on the current page. The entire table can be exported as a csv file by clicking the `Download` button in the sidebar.

6.2   Adding a new organism

We are currently implementing a simple method for adding organisms using phytozome annotations. For now, if there is an organism you would like to see added to the browser please email your request to `baxterlabZbrowse@danforthcenter.org`.

7   Credits

Zbrowse was conceived by Ivan Baxter and Greg Ziegler. Implemented by Greg Ziegler and Ryan Hartsock. The browser would not be possible without these software packages: R, R Studio, Shiny, Highcharts, and Rcharts.

8   License

Zbrowse is available under the creative commons non-commercial license CC BY-NC 3.0. The Highsoft software package, Highcharts, which is used to make the plots, is also licensed under CC BY-NC 3.0 and is not available for commercial use. Other R packages and JavaScript libraries Zbrowse uses are licensed under their own terms.