A peer-reviewed version of this preprint was published in PeerJ on 27 May 2015.

View the peer-reviewed version (peerj.com/articles/cs-3), which is the preferred citable publication unless you specifically need to cite this preprint.

Ziegler GR, Hartsock RH, Baxter I. 2015. Zbrowse: an interactive GWAS results browser. PeerJ Computer Science 1:e3
https://doi.org/10.7717/peerj-cs.3

Zbrowse: an interactive GWAS results browser

Greg R Ziegler, Ryan H Hartsock, Ivan Baxter

The growing number of genotyped populations, the advent of high-throughput phenotyping techniques and the development of GWAS analysis software has rapidly accelerated the number of GWAS experimental results. Candidate gene discovery from these results files is often tedious, involving many manual steps searching for genes in windows around a significant SNP. This problem rapidly becomes more complex when an analyst wishes to compare multiple GWAS studies for pleiotropic or environment specific effects. To this end, we have developed a fast and intuitive interactive browser for the viewing of GWAS results with a focus on an ability to compare results across multiple traits or experiments. The software can easily be run on a desktop computer with software that bioinformaticians are likely already familiar with. Additionally, the software can be hosted or embedded on a server for easy access by anyone with a modern web browser.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.902v1 | CC-BY 4.0 Open Access | rec: 17 Mar 2015, publ: 17 Mar 2015

Greg R Ziegler
greg.ziegler@ars.usda.gov
United States Department of Agriculture Agricultural Research Service
St. Louis,
Missouri,
United States

Ryan H Hartsock
rhartsock423@gmail.com
Donald Danforth Plant Science Center,
St. Louis,
Missouri,
United States

Ivan Baxter
ivan.baxter@ars.usda.gov
United States Department of Agriculture Agricultural Research Service
St. Louis,
Missouri,
United States

CORRESPONDING AUTHOR
Ivan Baxter
975 N Warson Rd, St. Louis, MO 63132
(314) 587-1438
ivan.baxter@ars.usda.gov

Introduction

The recent development of high-throughput plant 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 (Kang et al., 2010; Segura et al., 2012; Lipka et al., 2012) that can map traits to the genome with high enough resolution to quickly provide a tractable list of potential causal genes.

One of the first steps an analyst will take is determining what gene or genes fall under the SNP peaks that can be seen on the Manhattan plot. Unfortunately, these plots are not interactive and identifying the peaks of interest usually involves sifting through the results table for the coordinates of the peak of interest and then filtering a large gene annotation file for a range around the coordinates of the peak. The extra steps involved in exploring the data in this way makes it more likely that interesting associations may be missed either due to 1) mistakes made in attempting to mine the large results files or 2) the dataset not being mined deeply enough due to the difficulty of looking for genes under less significant peaks. Additionally, this method quickly becomes tedious when analyzing multiple phenotypes or relatively complex traits. Some web applications provide tools for viewing Manhattan plots interactively, but they are all either specific to a single species (Childs, Lisec & Walther, 2012; Li, Sham & Wang, 2012; Seren et al., 2012) or are part of larger software suites that make fast and straightforward viewing of GWAS results difficult (Stein et al., 2002). These resources also do not allow for easy viewing and comparison of GWAS results across phenotypes and studies, a situation that frequently arises with structured populations.

Goals

We approached the construction of a new GWAS browser with the goal of giving the users the following tools, all of which were focused on versatility and adaptability:

1. *Ability to plot multiple traits in the same panel.* We wanted to enable users 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 allele) and loci with pleiotropic effects (the same loci affecting multiple phenotypes)..

2. *Ability to rapidly move between scales (thousands of bps to billions).*
3. *Ability to to find overlaps or commonalities among sets.*
4. *Ability to interact directly with the plots*. We wanted the ability to look at the annotations of genes inline easily and link to additional information.
5. *Ability to download plots and gene lists.*
6. Finally, we wanted all of this information and functionality to be available in one browser window using tools that are common and freely available to the community.

Here, we present an interactive GWAS results viewer 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. *Arabidopsis thaliana*, maize, soybean, and sorghum are bundled with the software but we provide instructions and tools to easily add support for other organisms.

User Interface

The ZBrowse GWAS results viewer is an interactive application that runs on a local machine using R and is rendered in any modern web browser. Because the browser runs on the users local machine, the data will remain private. The focus of the first version is a local installation, the browser display allows for easy sharing of the application. The browser is designed to be a streamlined environment that provides fast access to visualization tools to allow the quick analysis of GWAS results. ZBrowse utilizes a tab-based navigation format to make accessing different aspects of the browser fast, efficient, and intuitive. There is also a sidebar panel on the left of the page that updates with a set of options specific to the tab being displayed.

The first tab in the list, and the landing page when the application is first loaded, is the Manage tab (Figure 1). 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. Before uploading, the user selects the appropriate organism from a dropdown menu. Data can be uploaded in a flat file delimited with either commas or tabs or an RData object.

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. The user needs to select a chromosome and base pair for determining the location of each SNP in the genome. 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, a researcher might be interested in comparing GWAS results from multiple experiments, or in comparing results from multiple traits measured in the same experiment, or both. The user simply needs to select the column or columns with the label for the trait that the SNP corresponds to. Finally, the user needs to 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 (RMIP, Valdar et al., 2009) or any other measure of trait significance, such as effect size. The final parameter allows for user selectable values for the Y-axis scale. By default, the software will automatically scale the y-axis based on the range of the selected data.

After the user has selected the appropriate parameters, clicking the submit button will trigger a tab change to the Whole Genome View visualization tab (Figure 2). Conveniently, once submitted, the software will remember the selected settings for this dataset on future visits and automatically populate the fields with the previously selected parameters. 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 highlight chromosome breaks. 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. There is also an option for showing only overlapping SNPs with the ability to adjust both the overlap size around each point and the minimum number of overlaps.

When the user scrolls 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 (Figure 3). 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 will change tabs to the Chromosome View tab (Figure 4). 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 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. The displayed chromosome can be changed without returning to the Whole Genome View tab using the drop-down menu in the sidebar panel. 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 SNP of interest. The tooltip for genes in this track displays information about the gene location, strand, and function, if known. For maize, arabidopsis and soybean, 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) (Lamesch et al., 2011), soybean links to Soybase (Grant et al., 2010), and maize links to the Maize Genetics and Genomics Database (MaizeGDB) (Lawrence et al., 2004). In addition, clicking genes in organisms added from Phytozome (Goodstein et al., 2012) via the add organism application described below opens the Phytozome description page for that gene. ZBrowse can be easily modified to link out to other species-specific databases that can accept a query string in the URL.

In addition to the visual browser, annotation data can be explored in tabular form in the Annotation Table tab (Figure 5). This table provides an interactive table of the genes found in the window around the selected point. The table is sortable and searchable and can also be exported as a comma-separated file. A similar table viewer is available in the Data Table tab for analysis of the selected GWAS dataset.

Adding Organisms

Currently, maize, soybean, arabidopsis and sorghum are downloaded with the browser source package. We have developed an application to quickly add organisms to the browser from annotations downloaded from the Plant Genomics Portal (Phytozome) to the local installation of ZBrowse. Additionally, we will be formatting requested and popular organisms and releasing the files on GitHub. These will be easy to download and incorporate into your existing browser installation.

Technical Foundation

The GWAS browser is written in the R programming language using packages that provide wrappers around popular javascript web applications including shiny (RStudio Inc., 2013) and rCharts (Vaidyanathan, 2013). Because of this, the browser can be run locally with only R and any modern web browser. Internal data processing makes use of the plyr package (Wickham, 2011). The javascript plots are drawn using Highcharts (highcharts.com) and are available for use under the Creative Commons Attribution-NonCommercial 3.0 License. Tables are generated using the javascript library Datatables (datatables.net) and xtable (Dahl, 2013). All of the tools and software used are either free or open source. The use of R to build the web application makes it more easily accessible to bioinformaticians to extend than if it was written in pure javascript. Many GWAS programs are written in R (Kang et al., 2008; Segura et al., 2012; Lipka et al., 2012). So, many scientists performing GWAS will already have some familiarity with R constructs, even if they are not computational biologists. This familiarity will hopefully make it easier for the community who is using the browser to extend it and modify it for their purposes.

Limitations

The browser takes a fundamentally different approach from current state of the art browsers. It is focused on the ability to quickly plot a variety of GWAS experiments on a single Manhattan plot. A caveat to this ability, however, is that it cannot plot every SNP in a genotype dataset. Due to memory, time, and plotting constraints the current browser is limited to approximately 5000 data points per trait, which is significantly less than most genotype datasets. Of course, only the most strongly associated SNPs are typically of interest, so this problem can be easily mitigated by trimming the input file to contain only significant associations (e.g., $p<0.05$). Future improvements to the browser could support the plotting of more information by binning points when zoomed out to a point where over plotting is an issue and

only loading individual data points asynchronously when the zoom level is sufficient to see individual points.

The generality of the browser allows for it to be used with any SNP dataset. Only chromosome number and base pair information needs to be provided for each SNP. However, this means that specific information about the genotype dataset being used, such as minor allele frequency or linkage disequilibrium information, cannot be displayed on the plot. Of course, the flexibility of the browser would make it easy to build personalized solutions that could display additional information for specific SNP datasets and additional tracks could be added to display linkage disequilibrium decay around the selected SNP.

One obvious extension of the browser that would address many of the limitations listed above would be to connect it to a database designed to quickly and efficiently handle all of the data that goes into a GWAS experiment. Database support would allow custom subsetting of entire GWAS datasets and if the GWAS genotype files are available, then summary data about each particular SNP could also be displayed. This would allow the browser to be incorporated into a much larger ecosystem that could take a GWAS experiment from phenotypic dataset, through running a GWAS experiment, to final analysis and visualization.

While the limitations identified above may constrain the use of the browser for certain applications, there are a number of use cases that are enabled by its current functionality. Using open source tools and GitHub for the code distribution, the browser functionalities can be enhanced by the authors or by other members of the user community.

Available resources:

Download at:

http://www.baxterlab.org/#!/cqi0

Code is also available on github at

https://github.com/baxterlabZbrowse/ZBrowse

manual can be found here:

http://media.wix.com/ugd/52737a_2a65d0deb3bd4da2b5c0190c0de343ca.pdf

For support email:

baxterlabZbrowse@danforthcenter.org

References

Childs LH, Lisec J, Walther D. 2012. Matapax: An Online High-Throughput Genome-Wide Association Study Pipeline. *Plant Physiology* 158:1534–1541.

Dahl DB. 2013. *xtable: Export tables to LaTeX or HTML*.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* 40:D1178–D1186.

Grant D, Nelson RT, Cannon SB, Shoemaker RC. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Research* 38:D843–D846.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* 178:1709–1723.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* 42:348–354.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E. 2011. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* 40:D1202–D1210.

Lawrence CJ, Dong Q, Polacco ML, Seigfried TE, Brendel V. 2004. MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Research* 32:D393–D397.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28:2397–2399.

Li MJ, Sham PC, Wang J. 2012. Genetic variant representation, annotation and prioritization in the post-GWAS era. *Cell Research* 22:1505–1508.

RStudio Inc. 2013. *shiny: Web Application Framework for R*.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* 44:825–830.

Seren Ü, Vilhjálmsson BJ, Horton MW, Meng D, Forai P, Huang YS, Long Q, Segura V, Nordborg M. 2012. GWAPP: A Web Application for Genome-Wide Association Mapping in Arabidopsis. *The Plant Cell Online*:tpc.112.108068.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. 2002. The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Research* 12:1599–1610.

Vaidyanathan R. 2013. *rCharts: Interactive Charts using Polycharts.js*.

Valdar W, Holmes CC, Mott R, Flint J. 2009. Mapping in Structured Populations by Resample Model Averaging. *Genetics* 182:1263–1277.

Wickham H. 2011. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software* 40:1–29.

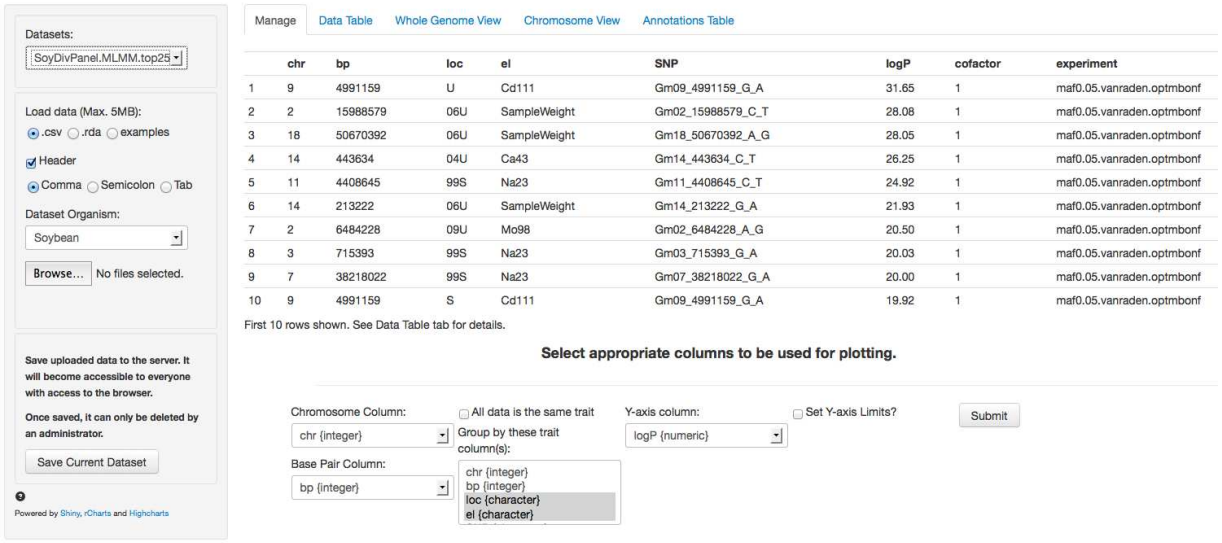

Fig. 1. Landing page for Interactive GWAS Results Browser.

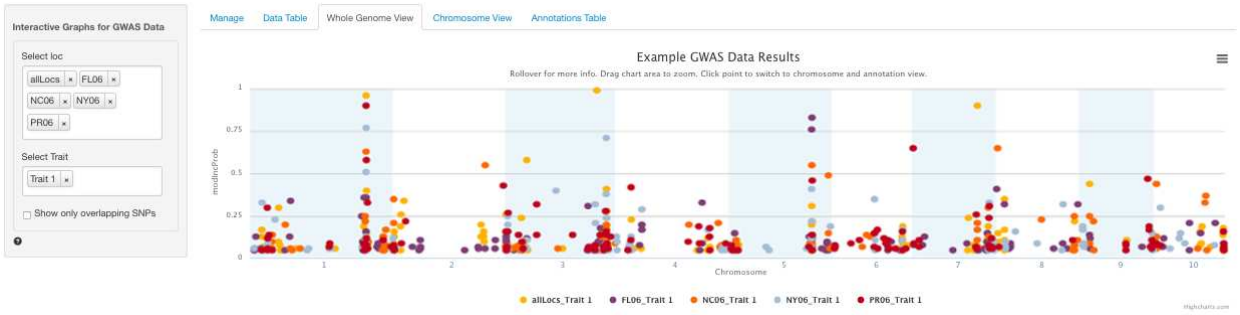

Fig. 2. Genome wide view of Interactive GWAS Results Browser. The legend at the bottom of the figure displays the color of points that correspond to the combination of traits and locations selected in the sidebar on the left hand side of the figure. Clicking the points in the legend allows a user to easily show or hide points from that trait. modIncProb=Random Model Inclusion Probability, RMIP, the fraction of times each SNP displayed was returned out of 100 GWAS analyses performed on a random subset of 80% of the data. The title of the plot is automatically generated from the filename of the dataset provided by the user. This makes it easy to determine which GWAS experiment is being plotted. Base Pairs = base position along the chromosome.

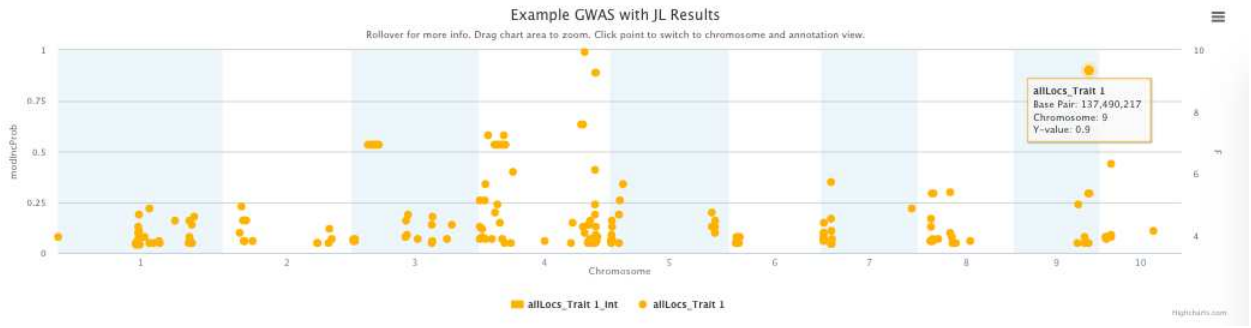

Fig. 3. Example of tooltip displaying top SNP for Trait 1 in a Maize NAM GWAS experiment. The legend at the bottom of the figure displays the color of points that correspond to the combination of traits and locations selected in the sidebar on the left hand side of the figure. Clicking the points in the legend allows easily show or hide points from that trait. modIncProb=Random Model Inclusion Probability, RMIP, the fraction of times each SNP displayed was returned out of 100 GWAS analyses performed on a random subset of 80% of the data. The title of the plot is automatically generated from the filename of the file provided by the user. This makes it easy to determine which GWAS experiment is being plotted.

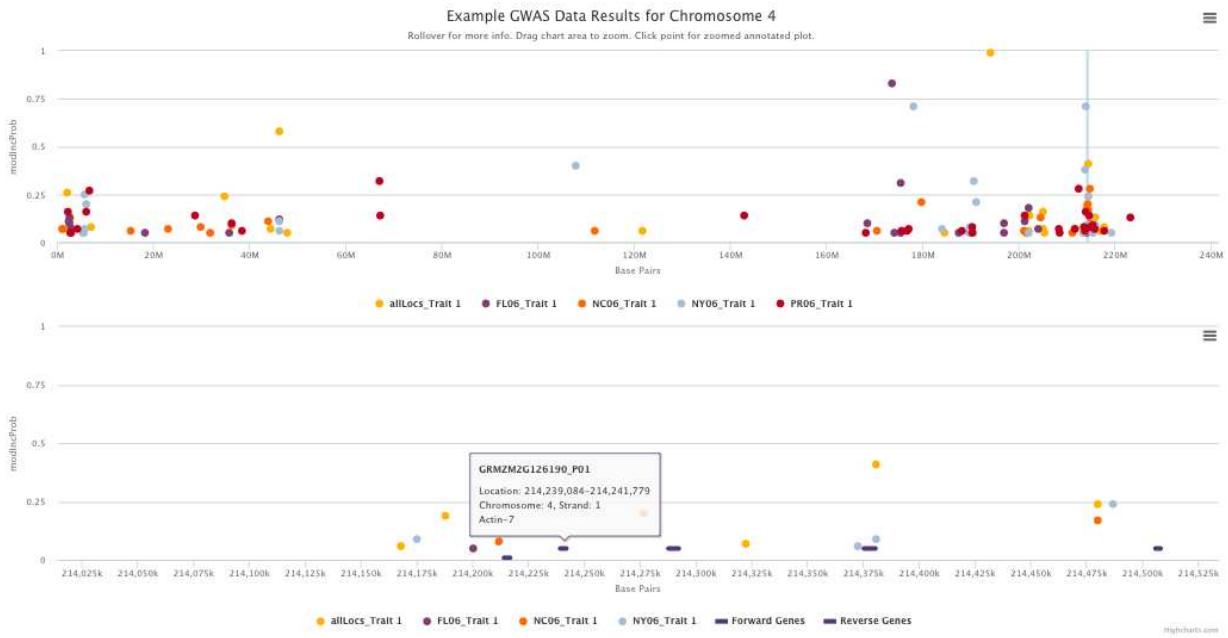

Fig. 4. Chromosome View tab of the Interactive GWAS Results Browser displaying a peak above the ion transporter. The legend at the bottom of the figure displays the color of points that correspond to the combination of traits and locations selected in the sidebar on the left hand side of the figure. Clicking the points in the legend allows a user to easily show or hide points from that trait. modIncProb=Random Model Inclusion Probability, RMIP, the fraction of times each SNP displayed was returned out of 100 GWAS analyses performed on a random subset of 80% of the data. The title of the plot is automatically generated from the filename of the dataset provided by the user. This makes it easy to determine which GWAS experiment is being plotted. Base Pairs = base position along the chromosome.

Manage | Data Table | Whole Genome View | Chromosome View | Annotations Table

15 records per page

Search: Copy Print Save

translation_id	chromosome	gene_id	transcript_id	is_canonical	gene_class	transcript_class	transcript_start	transcript_end	transcript_strand	evidence_type	cds_status	V2
GRMZM2G024624_P01	1	GRMZM2G024624	GRMZM2G024624_T01	yes	NH	NH	244787884	244789920	1	cdna|est|omrna	good	Hypothetical gene of unknown function
GRMZM2G024680_P01	1	GRMZM2G024680	GRMZM2G024680_T01	yes	WH	WH	244770698	244772023	1	cdna|est|protein	good	Jasmonate ZIM-domain protein 3
GRMZM2G024705_P01	1	GRMZM2G024705	GRMZM2G024705_T01	yes	WH	WH	244758653	244759635	-1	abinitio	good	GRMZM2G024705_P01
GRMZM2G070825_P01	1	GRMZM2G070825	GRMZM2G070825_T01	yes	WH	WH	244738411	244740169	-1	cdna|est|omrna|protein	good	Esterase
GRMZM2G083091_P01	1	GRMZM2G083091	GRMZM2G083091_T01	yes	WH	WH	244982406	244987142	-1	est|mrna	good	Transmembrane protein 14
GRMZM2G083156_P01	1	GRMZM2G083156	GRMZM2G083156_T01	yes	WH	WH	244976754	244978209	1	abinitio|omrna|protein	good	similar to putative sulfate transporter
GRMZM2G083182_P01	1	GRMZM2G083182	GRMZM2G083182_T01	yes	WH	WH	244908744	244974544	1	abinitio	good	ABC transporter
GRMZM2G139744_P01	1	GRMZM2G139744	GRMZM2G139744_T01	yes	WH	WH	245159703	245165127	1	est|protein	good	Sec20 family protein
GRMZM2G145412_P01	1	GRMZM2G145412	GRMZM2G145412_T01	yes	WH	WH	244826496	244827697	-1	cdna|est|omrna|protein	good	ZIM motif family protein
GRMZM2G145458_P01	1	GRMZM2G145458	GRMZM2G145458_T01	yes	WH	WH	244834016	244834761	-1	abinitio|cdna|est|omrna	good	ZIM motif family protein
GRMZM2G146994_P01	1	GRMZM2G146994	GRMZM2G146994_T01	yes	WH	WH	244748663	244758524	1	est	good	GRMZM2G146994_P01
GRMZM2G174549_P01	1	GRMZM2G174549	GRMZM2G174549_T01	yes	WH	WH	245032159	245037328	-1	cdna|est|omrna|protein	good	Purple acid phosphatase
GRMZM2G174574_P02	1	GRMZM2G174574	GRMZM2G174574_T02	yes	WH	WH	245027693	245031977	1	cdna|est|omrna|protein	good	GRMZM2G174574_P02
GRMZM2G373779_P01	1	GRMZM2G373779	GRMZM2G373779_T01	yes	WH	WH	244734090	244737572	-1	abinitio	good	BSD domain containing protein
GRMZM2G472403_P02	1	GRMZM2G472403	GRMZM2G472403_T02	yes	TE	TE	245039264	245043621	-1	abinitio|est|omrna|protein	good	GRMZM2G472403_P02
translation_id	chromosome	gene_id	transcript_id	is_canonical	gene_class	transcript_class	transcript_start	transcript_end	transcript_strand	evidence_type	cds_status	V2

Fig. 5. Annotation tab of the Interactive GWAS Results Browser.