Plant Physiology Preview. Published on February 21, 2012, as DOI:10.1104/pp.112.194027

Matapax: An online high-throughput genome-wide association study pipeline

Liam H. Childs, Jan Lisec, and Dirk Walther

Corresponding Author:

Liam H. Childs
AG Bioinformatics
Am Mühlenberg 1
Golm 14476
Germany

Tel: +49 (0) 331 567 8624
Email: childs@mpimp-golm.mpg.de

Matapax: High-throughput, genome-wide association

Liam H. Childs*, Jan Lisec, and Dirk Walther

Max-Planck Institute for Molecular Plant Physiology, Am Mühlenberg 1, Golm, Germany.

*Corresponding author: childs@mpimp-golm.mpg.de

Abstract

High throughput sequencing and genotyping methods are dramatically increasing the number of observable genetic intraspecies differences that can be exploited as genetic markers. In addition, automated phenotyping platforms and OMICS profiling technologies further enlarge the set of quantifiable macroscopic and molecular traits at an ever increasing pace. Combined, both lines of technological advances create unparalleled opportunities to identify candidate gene regions and, ideally, even single genes responsible for observed variations in a particular trait via association studies. However, as of yet this new potential is not sufficiently matched by enabling software solutions to easily exploit this wealth of genotype-phenotype information. We have developed Matapax, a web-based platform to address this need. Initially, we built the infrastructure to support association studies in *Arabidopsis thaliana* based on several genotyping efforts covering up to 1,375 Arabidopsis accessions. Based on the user-supplied trait-information, associated SNP-markers and SNP-harbouring or neighbouring genes are identified using both the GAPIT and EMMA libraries developed for R. Additional interrogation is facilitated by displaying candidate regions and genes in a genome browser and by providing relevant annotation information. In the future, we plan to broaden the scope of organisms to other plant species as more genotype/phenotype information becomes available.

Availability: Matapax is freely available at http://matapax.mpimp-golm.mpg.de and can be accessed using any internet browser.

Keywords: Genome-wide associations studies (GWAS), SNP, genotyping, phenotyping, Arabidopsis thaliana, candidate gene identification, genomic marker, marker-trait associations

Introduction

Genome wide association studies (GWAS) are a powerful way to harness natural variation to detect the genomic causes for phenotypic variance by testing the relationship between sequence and phenotypic variation. Although, GWAS are criticised for being a data driven approach that potentially inflate Type I error through the sheer number of tests performed, they have enjoyed several successes in many organisms through the identification of experimentally determined associations (Aranzana et al., 2005; Zhao et al., 2007) and the identification of associations that were subsequently experimentally confirmed (Klein et al., 2005; 2007; Sladek et al., 2007). They also highlight many plausible novel associations (Atwell et al., 2010; Todesco et al., 2010).

The basic unit of GWAS is a statistical test of the association between the alleles of a genetic marker and the corresponding trait measurements. The number of these comparisons performed is set to rapidly rise due to two factors. 1) The advent of high-throughput macroscopic and molecular phenotyping platforms enabled the analysis of many more traits and accessions; i.e. organisms with a unique genetic makeup. For example, a single metabolomic study has the potential to produce measurements for hundreds to thousands of metabolites in as many accessions (Lisec et al., 2006; Sozzani and Benfey, 2011). 2) Novel de novo or re-sequencing technologies have yielded massively increased numbers of genetic markers at ever falling costs. Thus, the density of genetic markers may now be high enough to potentially allow single gene resolution of GWAS studies. These factors combined mean that a single GWAS may require over a million trait-marker comparisons. Although this allows the genetic causes for the trait variation to be much more tightly defined, a single study now demands significantly more computational resources to complete in a timely and statistically meaningful manner.

There are many tools available that perform GWAS (Aulchenko et al., 2007; Bradbury et al., 2007; Browning and Browning, 2007; Purcell et al., 2007; Muniz-Fernandez et al., 2011), two of which deserve particular mention as they are popular choices in plant and human GWAS. TASSEL (Bradbury et al., 2007) is a Java-based tool, which can be either downloaded or run from the home page, that provides a plethora of different analyses such as the pre-filtering of trait and marker data, phylogenetic analyses, general linear model and mixed linear model analyses. TASSEL was developed for maize, but is applicable to any organism. PLINK (Purcell et al., 2007) provides a set of tools to perform GWAS and supporting analyses and is heavily used in the HapMap project (2003). PLINK was developed for humans and is also applicable to any organism.

However, out-of-the-box, none of the aforementioned tools appear capable of performing the number of comparisons needed in a timely manner without extensive modification or knowledge of concurrent/asynchronous programming. They also require extensive data formatting by the users to

obtain the necessary input file formats. This poses a significant barrier to research groups who wish to perform GWAS but lack the necessary technical expertise or computing power.

Recently, the publication of three definitive *Arabidopsis thaliana* accession resources (Atwell et al., 2010; Cao et al., 2011; Horton et al., 2012) has made it possible to develop a platform that provides GWAS for practically any Arabidopsis trait data. Combined, these resources provide high density SNP-marker maps for over 1000 Arabidopsis accessions, which should be sufficient for many, if not most, Arabidopsis GWAS. However, these data require a prohibitive amount of pre-processing to assemble the data in a format that can be used by the aforementioned GWAS programs.

We seek to introduce a platform for the genome wide analysis of trait-marker interaction that simultaneously addresses the issues of computation time, renders GWAS analysis more easily accessible to biologists, provides informative post-hoc analyses of the results and to make such studies widely accessible to the broader scientific community regardless of in-house technical capability or analytical expertise.

Results

We developed the Matapax web-based platform to initially support GWAS in *Arabidopsis thaliana*. It was implemented as a web-based solution utilising the R library EMMA (Kang et al., 2008) and GAPIT (Lipka et al., 2011). EMMA is the current tool of choice for GWAS on Arabidopsis and provides efficient computation of associations using mixed models and kinship matrices for population structure control. GAPIT enhances EMMA with algorithms designed to boost statistical power and further decrease computation time. Processing of user data is performed in parallel on the server using the Torque resource management system. For use, the user requires an internet browser and need only upload a matrix of trait measurements including trait and accession names. The developed platform was named MArker-Trait Association Platform And eXplorer (Matapax).

Initial data check

Matapax incorporates three marker sets from the *Arabidopsis thaliana* POLYmorphism DataBase (AtPolyDB, referred to in this text as SNP-Set1 and SNP-Set2) (Atwell et al., 2010; Horton et al., 2012), the 1001-Genomes Project marker database (SNP-Set3) (Cao et al., 2011) into the pipeline with the option to choose either one or the other for association mapping. Both of these databases provide high quality genotyping data for an extensive number of accessions and markers (single nucleotide polymorphisms, SNPs) that have been produced using either SNP chips (SNP-Set1 and SNP-Set2) or resequencing techniques (SNP-Set3). Matapax currently utilizes 199, 1,307 and 91 accessions from SNP-Set1 SNP-Set2 and SNP-Set3, respectively. The number of genotyped accessions in SNP-Set3 is set to increase to 1001 accessions once initial, pre-publication use

limitations are lifted by the data providers. Matapax uses the public marker information as provided by the respective data providers thereby relying on their proper SNP calling quality control. However, we have included an option in the results for users to filter out markers that do not meet a specified Minimum Allele Frequency. In future, this option will also be available for the Hardy-Weinberg equilibrium.

There is currently no naming standard for Arabidopsis accessions, thus the three different marker databases that Matapax uses occasionally have slightly different names for the same accessions. Matapax attempts to automatically match user accessions to the database accessions, however, when there is no direct match, the option to interactively select the correct accession is offered. In the marker databases, some accessions have been genotyped more than once. In this case, the user is presented with an option to choose from the different unique ecotype IDs that the marker databases provide. If the user's accession is not present in the marker database, the option to ignore the accession is available. The user can also specify that an accession is a cross between two other homozygous accessions by writing the accession name in the form "<accession 1> x <accession 2>" (i.e. separating the two accessions with an "x").

In cases where the distribution of trait measurements is skewed, we provide the option to perform a Box-Cox transformation, which lessens the influence of outliers and produces a more symmetric distribution. The fit of the trait data to a normal distribution is estimated using the Shapiro-Wilk test (Shapiro and Wilk, 1965) and a plot of the measurement distribution is provided.

Results analysis

Post-hoc analysis is assisted through the use of a graphical genome browser (Figure 1) and a detailed results table (Figure 2) listing the p-values for the association of each marker with each trait. The user is able to load and manipulate this table entirely within their own internet browser and can recall the results of previous studies by supplying a job identification number that is provided after job submission. The results can also be downloaded as a compressed table for local analysis.

The genome browser is based on AnnoJ (http://www.annoj.com), a system that utilises REpresentational State Transfer (REST) principles. This allows AnnoJ to be customised towards our purposes and to draw data from many sources. The genome browser displays gene models provided by the Salk Institute (http://pbio.salk.edu/pbioe) along with the obtained $-\log_{10}$ transformed association p-values and allows the user to browse, scale, zoom and search the results.

The results table provides several features including filter and multiple sort capabilities on the traits, chromosomes, positions and p-values. Searching of TAIR10 Arabidopsis annotation is enabled along with links to TAIR annotation (TAIR; http://www.arabidopsis.org). Additionally, polymorphism information is provided as base changes in intergenic and intronic regions, and amino acid changes in

coding regions. Further information is provided as box-plots of marker-trait segregation and quantile-quantile (Q-Q) plots of the association values.

As the very nature of GWAS involves hundreds of thousands of accession-phenotype association tests, Multiple Testing Correction needs to be addressed. To this end, the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) is employed to correct the provided p-values. Another approach is to visually observe how the most significant p-values deviate from the expected distribution. Ideally, a higher than expected fraction of marker-associated p-values will be skewed towards significant values. To assist in this form of analysis, Q-Q plots of the association values per trait are available. Such plots are a scatter-plot of the expected and observed distributions that allow deviations to be clearly identified. Points that are noticeably different from the diagonal or deviate from the trend are potential candidates for further examination.

Matapax run-time

To assess Matapax processing times, we ran the pipeline while varying the numbers of traits and accessions independently. Increasing the number of traits submitted in a single job, resulted in a linear increase in run-time with jumps every 12 traits corresponding to the number of nodes on the server (Figure 3a). There was also a gradual overall increase due to the overhead of querying the marker data and inserting the results data into a relational database management system.

Increasing the number of accessions resulted in an exponential increase in run-time for both GAPIT and EMMA, although GAPIT performed much faster than EMMA (Figure 3b). Due to the exponential computational complexity of EMMA, we were unable to test EMMA with more than 200 accessions. In future, the next iteration of EMMA (EMMAx) will be used once it leaves beta stage providing a significant increase in computational speed. These results suggest that using GAPIT for any number of accessions over 200 is advisable. With GAPIT, it is possible to perform GWAS for 1000 accessions in a little over 4 hours per trait, implying use of all accessions in SNP-Set1, SNP-Set2 and SNP-Set3 is computationally feasible.

Matapax-desktop comparison

A simple test between the results of Matapax and a desktop computer was conducted using simulated phenotypes for all 199 accessions in SNP-Set1 resulting in a perfect correlation. The generated trait values, the association R script for a desktop computer and the results for both Matapax and the desktop are available in the supplementary (Supplementary File 2). The marker data can be obtained from AtPolyDB (https://cynin.gmi.oeaw.ac.at/home/resources/atpolydb) and the EMMA R script can be obtained from the EMMA home page (http://mouse.cs.ucla.edu/emma).

Case study

To test Matapax, we replicated an earlier association study (Atwell et al., 2010) that associated 107 Arabidopsis traits covering resistance, ionomics, flowering and developmental phenotypes with the high density SNP marker set SNP-Set1 using a kinship matrix for population structure correction. By comparing our results to those in this study, we were not only able to test our pipeline over several traits, but also assess how closely our method agrees with established results. Through Matapax, we associated the trait data with SNP-Set1 using EMMA to perform the associations and a K-matrix to correct for population structure.

The vast majority of Matapax results had an almost 1:1 correlation with the Atwell 2010 data (Figure 4, Supplementary Table 1). For these traits, observations could be drawn that are highly similar or identical to the original results concerning the location and significance of the association. Further information about the nature of the polymorphism was also provided including whether it is intergenic, synonymous or non-synonymous.

There were a few traits, however, that remained difficult to reproduce, though the results clearly correlated. Due to the complicated nature of GWAS, there are several places where the discrepancies could have arisen: 1. Differences in the way the genomic data was calculated could alter how the measurements were compared to each other in the mixed linear model. Such differences could stem from the use of different marker or sequence data versions or differing missing value imputation methods. 2. Differences in the kinship matrix calculation would change the way the fixed effects are calculated in the model. 3. Other potential sources of differences are the algorithm version and exact parameters used.

Discussion

By developing Matapax we sought to create a platform that is easy to use, that makes the available rich genotypic information readily available, where the user requires little-to-no technical knowledge about GWAS or programming and that assists the user in answering and further developing hypotheses from the results.

Although we are able to reproduce the results of the Atwell 2010 study with high accuracy (87 out of 107 traits with a correlation greater than 0.80), there were 7 traits that correlated rather poorly (correlation less than 0.60) (Figure 3). Due to the complex nature of a GWAS, it is difficult to test all the possible combinations of different parameters and data sets to try and tease out the workflow that the original study followed. This lack of reproducibility highlights the need for standards that could be easily provided by a platform such as Matapax.

As discussed in (Zhao et al., 2007), the effectiveness of correction for population structure is widely debated and top results require a significant degree of scepticism. This is hopefully mitigated with an informed and annotated review of the results, and the application of the right population structure correction(s). Matapax provides two forms of correction (kinship matrix – K, population structure matrix – P), which can be newly calculated or uploaded by the user and appear to be sufficient to reduce inflated p-values in many traits (Zhao et al., 2007).

Aside from population structure correction, a great deal of manual inspection of the results is necessary to produce meaningful and biologically relevant results. To assist in manual inspection, we sought to include all relevant information about the results of a GWAS in a highly manipulable manner. With input from users, we hope to refine this aspect of Matapax further.

Matapax presents three high coverage Arabidopsis marker database choices for association that provide whole genome coverage of the Arabidopsis genome; AtPolyDB (SNP-Set1 and SNP-Set2) (Atwell et al., 2010; Huang et al., 2011; Horton et al., 2012) and the 1001-Genomes Project (SNP-Set3) (Ossowski et al., 2008; Weigel and Mott, 2009) using different methods for genotyping. The choice of marker database for association can be informed by the differences between the three databases. SNP-Set1 uses a 250k SNP resequencing chip that interrogates 250k positions in the Arabidopsis genome that were chosen based upon earlier genotyping work (Kim et al., 2007). This allows the rapid genotyping of Arabidopsis accessions but suffers from incomplete coverage, as is evidenced by the presented case study. There are only 199 accessions available in this dataset. SNP-Set2 uses the same experimental procedure for detecting SNPs as SNP-Set1 but uses a later calling method and has 1,307 accessions. The SNP-Set3 project markers are called using next-generation sequencing technology potentially identifying all markers in the sequenced accessions. This high-density genomic coverage comes at the cost of accession coverage. Currently, there are 447 accessions available in this dataset although only 91 are permitted for use.

In its first release, Matapax focuses on the model plant *Arabidopsis thaliana*. Evidently, the infrastructure can easily be expanded to include other species as well pending increased availability of broad information on genetic variability in the respective species. Furthermore, upon positive reception of Matapax by the scientific community, creating a central repository for GWAS around the Matapax nucleus may also be worthwhile especially in the light of the increased need to report testable and reproducible GWAS study results.

Conclusion

We have developed a genome-wide association pipeline that performs all essential steps for basic GWAS, is capable of handling genotypic crosses as well as a large number of requests and presents

the results in an easy-to-use manner that assists in the development of hypotheses. The entry level requirements to GWAS on Arabidopsis has been significantly lowered as investigators no longer need to format the myriad datasets to fit the specifications of the particular tool they are using, nor are they required to maintain computational resources beyond a computer with internet browser capability. The results analysis is being assisted by the provision of necessary annotation and contextual information and the possibility to search the annotation using keywords.

We present Matapax with the core capabilities required for GWAS and with all initial development goals met. Future development of Matapax will require user input on functionality and, with enough interest, Matapax will be placed under active development where we expect to work closely with users to plan many improvements and features that will be developed for future release. Matapax is freely available at http://matapax.mpimp-golm.mpg.de.

Materials and Methods

Matapax design

Matapax is a pipeline that ties together several technologies and publicly available tools and resources to achieve the desired goals and features.

We incorporate three genome-wide marker databases into the pipeline, both generated from Arabidopsis. SNP-Set1 and SNP-Set2 provide marker data for 199 and 1,307 accessions, respectively. The SNPs are determined using a 250k Affymetrix genotyping chip thus providing a high resolution map of genomic variation across a large number of accessions. Currently, SNP-Set3 provides marker data for 91 accessions. The SNPs are determined through re-sequencing thus providing an extremely high resolution map of genomic variation, although not as many accessions are available. However, in future, this number will increase. The marker databases are formatted as SQLite (http://www.sqlite.org) relational databases for efficient storage, quick retrieval times and scalability. At the time of submission, Matapax utilizes 91 accessions from SNP-Set3 because of pre-publication use restrictions imposed by the original data providers. After publication by the original data providers, the entire sets will readily be made available via Matapax.

Association tests are performed in the R statistical computing environment (http://www.r-project.org) using the EMMA library (Kang et al., 2008). Through GAPIT and EMMA we are able to provide efficient mixed-model association with optional corrections for kinship (K) and population structure (P). The K-matrix is a distance matrix where the difference is defined as the mean of the similarities plus the differences between two accessions. In the current version of Matapax, the difference is calculated as a simple genetic distance:

Eq. 1

$$d_{ij} = \frac{1}{n}\sum_{k}^{n} G_{ik} \times G_{jk} + (1 - G_{ik}) \times (1 - G_{jk})$$

where G_i and G_j are the accessions being compared and n is the number of markers. The values of G range between 0 and 1, typically falling on the values 0, 0.5 and 1 indicating a genotype with two major alleles, a heterozygous genotype and a genotype with two minor alleles. Heterozygous genotypes are standardised before calculating the K and P matrices. If another form of kinship matrix is desired, then it is possible to upload custom matrices. In future, Matapax will include other kinship matrices such as the Loiselle kinship matrix. Matapax derives the P-matrix by taking the first three principal components calculated from the genotypic data. To improve computational time, the NIPALS algorithm is used. However, power users should consider determining the number of PCAs separately and uploading their own P-matrix.

In addition to using efficient association algorithms, we implemented parallelisation with the Torque Resource Manager (http://www.clusterresources.com) ensuring that all jobs are processed without conflict and that the pipeline remains scalable in the event that more nodes are added in future. Currently, Matapax processes the traits of each submission in parallel, but in future, each trait-marker comparison will be parallelised.

Result analysis is assisted with the implementation of the genome browser AnnoJ (http://www.annoj.org). Each trait is displayed as a separate track along with graphical and textual genome annotation. An association track can be scaled, zoomed and browsed allowing the user to peruse the results and see how well genomic regions associate with the given traits.

Results are also presented as a table where each row displays the association between a trait and a marker. The columns cover the tested *Trait*, the *Chromosome* of the marker, the *Position* of the marker on the chromosome, the *p-value* of the association, the *Annotation* of the marker as The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org) AGI codes and the type of *Polymorphisms* that occur at that position. All columns except *Annotation* and *Polymorphisms* support multiple sorting and filtering to allow users to display the information they wish. The *Annotation* column supports searching allowing the user to search for genes or biological processes/molecular functions of interest. Filtering on the *Polymorphism* column will be implemented in a future release. Entries in the *Trait* column are hyperlinked to the Q(uantile)-Q(uantile) plot for that trait, entries in the *Position* column are hyperlinked to the genome browser, entries in the *p-value* column are hyperlinked to a box-plot of the trait-marker segregation and entries in the *Annotation* are hyperlinked to the TAIR website.

Pipeline flow

Only one input file is needed to start the pipeline and there is a choice of marker database available for the user. Traits are submitted as a simple tab delimited flat file where each row is a separate accession and each column the trait measurements. Accessions can be named by their native name or by the IDs used by the groups who produced the marker databases. A basic normality check is performed on each trait and a p-value indicating how well the trait measurements fit a normal distribution is given for each trait along with the option to perform a Box-Cox transformation that may improve the fit. A selection of population structure correction matrices is also presented to the user. Once all checks are complete, the markers for the matching accessions are extracted from pre-formatted SQL databases. Each trait is submitted individually to a cluster that is managed by the Torque Resource Manager (http://www.clusterresources.com) ensuring that all jobs are processed properly without resource conflicts.

Timing the pipeline

To test the run-time of Matapax, we independently increased the number of traits and accessions over several submissions. The run-time response to increasing traits was tested for random values generated for 50 accessions using all SNPs from SNP-Set1, GAPIT and a K matrix over increasing numbers of traits from 2 to 25. Each set of traits was submitted one at a time to ensure that all the nodes on the server were dedicated to a single submission. The run-time response to increasing accessions was tested for random values generated for a single trait using all SNPs from SNP-Seq1 and a K matrix over increasing numbers of accessions and for both GAPIT and EMMA. For reasons described previously, each set of accessions was submitted one at a time.

Result comparisons

As an initial comparison, we compared the association p-values of Matapax and a desktop computer. The trait data was created by generating random values for all 199 accessions in SNP-Set1. Matapax was run using a K-matrix and an R script was written to do the same on a desktop computer. The p-values resulting from both were then correlated.

To test the pipeline on published data, we obtained phenotypic and genomic data from (Atwell et al., 2010). These are publicly available from the AtPolyDB website. In addition, we obtained the association p-values via direct communication. The phenotype data covers disease resistance, ionomics, flowering and developmental phenotypes in Arabidopsis and the accession data was SNP-Set1. The trait and marker data were fed into the Matapax pipeline and the resulting association p-values were correlated with the published association p-values.

Supplemental Material

Supplementary Table 1: Trait names and correlation of Matapax with published data

Supplementary File 2: Desktop R script, trait values, and desktop and Matapax results for Matapax-desktop comparison

Acknowledgements

We would like to acknowledge Thomas Degenkolbe for his role in helping us test Matapax. We would also like to acknowledge Uemit Seren of the Nordborg lab for providing the association values of the Atwell 2010 article. We wish to thank Detlef Weigel and colleagues for graciously allowing us use of the generated genotyping information prior to publication. Finally, we would like to acknowledge Reviewer 1 of our manuscript whose detailed suggestions have helped shape and improve many aspects of Matapax.

References

Aranzana MJ, Kim S, Zhao KY, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang CL, Toomajian C, Traw B, Zheng HG, Bergelson J, Dean C, Marjoram P, Nordborg M (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. Plos Genetics **1:** 531-539

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature **465:** 627-631

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics **23:** 1294-1296

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) **57:** 289-300

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics **23:** 2633-2635

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet **81:** 1084-1097

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Muller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet **advance online publication**

The International HapMap Consortium (2003) The International HapMap Project. Nature **426:** 789-796

Wellcome Trust Case Consortium, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Barrett JC, Davison D, Easton D, Evans D, Leung H-T, Marchini JL, Morris AP, Spencer CCA, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, St Clair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskvina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop DT, Iles MM, Maqbool A, Yudasheve N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, Satsangi J, Fisher SA, Forbes A, Lewis CM, Onnie CM, Prescott NJ, Sanderson J, Mathew CG, Barbour J, Mohiuddin MK, Todhunter CE, Mansfield JC, Ahmad T, Cummings FR, Jewell DP, Webster J, Brown MJ, Lathrop GM, Connell J, Dominiczak A, Marcano CAB, Burke B, Dobson R, Gungadoo J, Lee KL, Munroe PB, Newhouse SJ, Onipinla A, Wallace C, Xue M, Caulfield M, Farrall M, Barton A, Bruce IN, Donovan H, Eyre S, Gilbert PD, Hider SL, Hinks AM, John SL, Potter C, Silman AJ, Symmons DPM, Thomson W, Worthington J, Dunger DB, Widmer B, Frayling TM, Freathy RM, Lango H, Perry JRB, Shields BM, Weedon MN, Hattersley AT, Elliott KS, Groves CJ, Lindgren CM, Rayner NW, Timpson NJ, Zeggini E, Newport M, Sirugo G, Lynons E, Vannberg F, Brown MA, Franklyn JA, Heward JM, Simmonds MJ, Hill AVS, Bradbury LA, Farrar C, Pointon JJ, Wordsmith P,

Gough SCL, Seal S, Stratton MR, Rahman N, Ban M, Goris A, Sawcer SJ, Compston A, Conway D, Jallow M, Bumpstead SJ, Chaney A, Downes K, Ghori MJR, Gwilliam R, Inouye M, Keniry A, King E, McGinnis R, Potter S, Ravindrarajah R, Whittaker P, Withers D, Easton D, Pereira-Gale J, Hallgrimsdottir IB, Howie BN, Su Z, Teo YY, Vukcevic D, Bentley D, Caulfield M, Mathew CG, Worthington J (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447:** 661-678

Grant MR, Godiard L, Straube E, Ashfield T, Lewald J, Sattler A, Innes RW, Dangl JL (1995) Structure of the Arabidopsis RPM1 gene enabling dual specificity disease resistance. Science **269**: 843-846

Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, Muliyati NW, Platt A, Sperone FG, Vilhjalmsson BJ, Nordborg M, Borevitz JO, Bergelson J (2012) Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. Nat Genet **44:** 212-216

Huang YS, Horton M, Vilhjálmsson BJ, Seren Ã, Meng D, Meyer C, Ali Amer M, Borevitz JO, Bergelson J, Nordborg M (2011) Analysis and visualization of Arabidopsis thaliana GWAS using web 2.0 technologies. Database **2011**

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

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M (2007) Recombination and linkage disequilibrium in Arabidopsis thaliana. Nat Genet **39:** 1151-1155

Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science **308:** 385-389

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore M, Buckler ES, Zhang Z (2011) User Manual of GAPIT: Genome Association and Prediction Integrated Tool.

Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie A (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. Nature Protocols **1:** 387-396

Muniz-Fernandez F, Carreno-Torres A, Morcillo-Suarez C, Navarro A (2011) Genome-wide association studies pipeline (GWASpi): a desktop application for genome-wide SNP analysis and management. Bioinformatics **27:** 1871-1872

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res **18:** 2024-2033

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81:** 559-575

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika **52:** 591-611

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature **445:** 881-885

Sozzani R, Benfey P (2011) High-throughput phenotyping of multicellular organisms: finding the link between genotype and phenotype. Genome Biology **12:** 219

Todesco M, Balasubramanian S, Hu TT, Traw MB, Horton M, Epple P, Kuhns C, Sureshkumar S, Schwartz C, Lanz C, Laitinen RAE, Huang Y, Chory J, Lipka V, Borevitz JO, Dangl JL, Bergelson J, Nordborg M, Weigel D (2010) Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. Nature **465:** 632-636

Weigel D, Mott R (2009) The 1001 genomes project for Arabidopsis thaliana. Genome Biol **10:** 107

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An Arabidopsis example of association mapping in structured samples. PLoS Genet **3:** e4

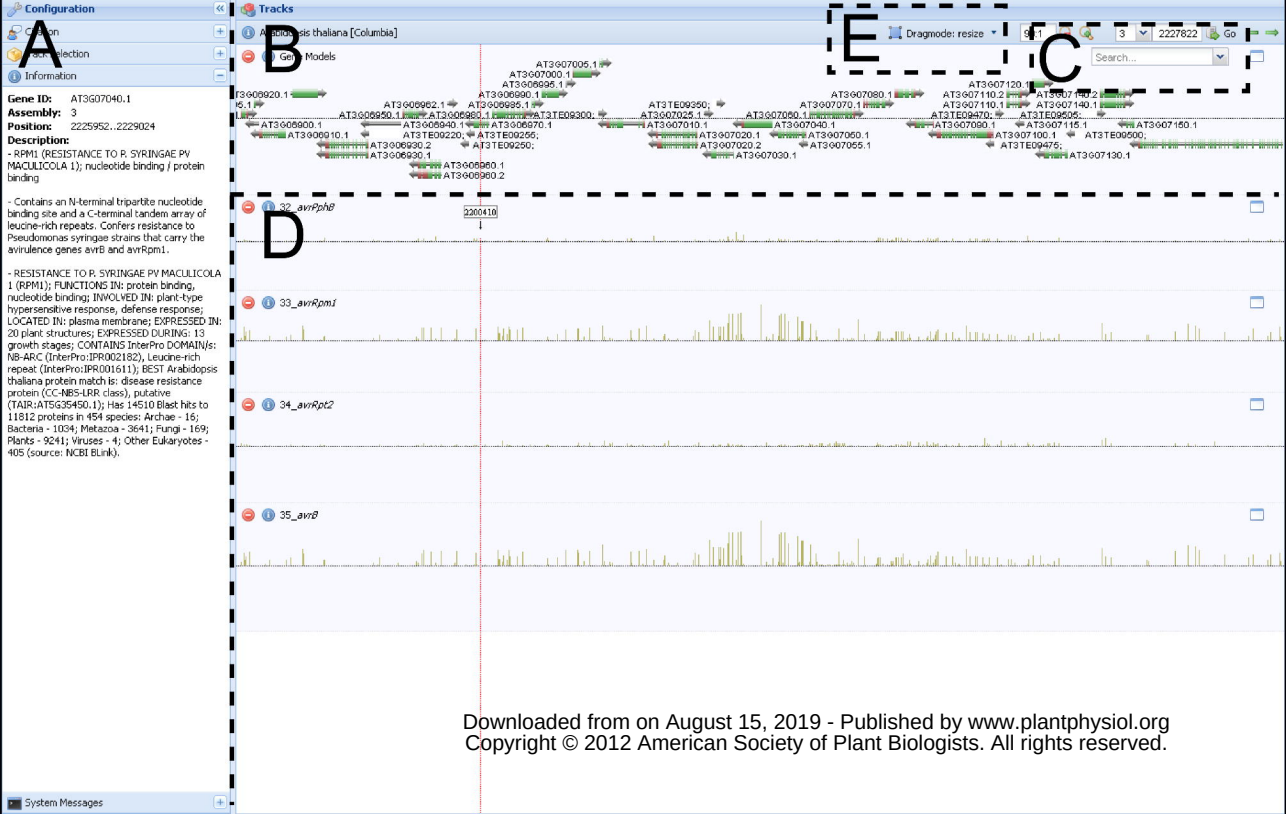
Figures

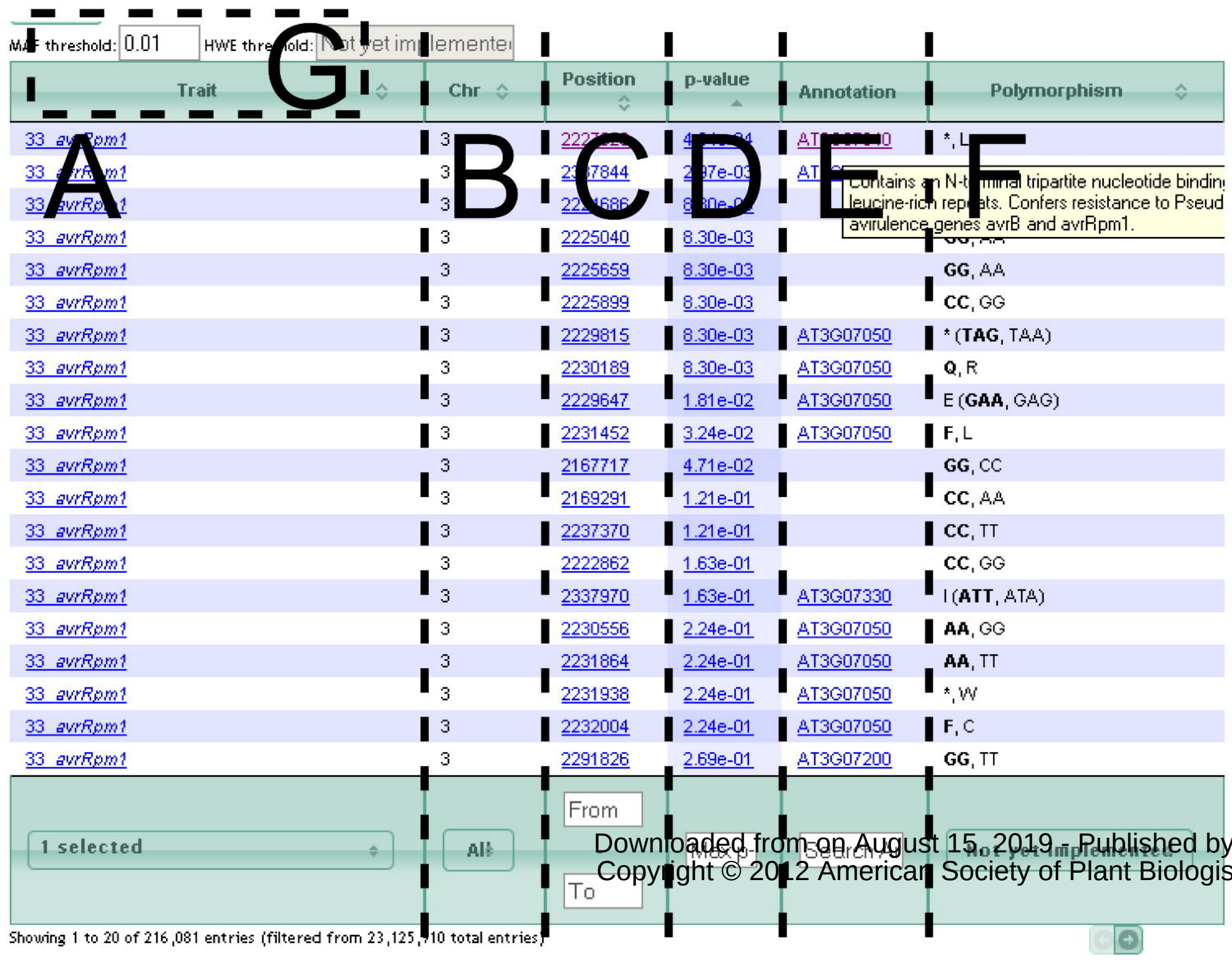
Figure 1: A screenshot of the genome browser. Matapax allows users to view association p-values in the context of the genome with the aid of the AnnoJ genome browser. This browser supports several features. **A:** The visible tracks can be selected in the side menu and any information about the track source or gene annotation is also displayed here. **B:** The default TAIR genome annotation is the default track. The TAIR version is dependent on the marker database selected for association. Selecting the gene model names will display the model annotation in the side menu. **C:** The annotation track can be searched using the search tool. Searches can be made on both AGI codes and model annotation terms. **D:** The remaining tracks display the marker association strengths of each analysed trait. Each analysed trait is displayed in a separate track. The association p-values are $-\log_{10}$ transformed meaning the higher the bar, the stronger the trait-marker association. **E:** The genome browser enables browsing by "dragging" the tracks horizontally, scaling the values of the individual tracks as well as resizing the track height and zooming on areas of interest. Users can also go directly to positions of interest. The displayed figure shows high association between the avrRpm resistance phenotype and the *rpm1* resistance gene, which has been shown to play a significant role in *Pseudomonas Syringae* resistance (Grant et al., 1995).

Figure 2: A screenshot of the results table. The results displayed in the results table of Matapax are highly configurable to assist users in the interpretation of their results. **A:** The Trait column can be both sorted and filtered. Trait names link to Q(uantile)-Q(uantile) plots of the trait. **B:** The *Chromosome* column can be both sorted and filtered. **C:** The *Position* column can be both sorted and filtered on a range. Each position is linked to the genome browser enabling the user to visualise the genomic context of the marker. **D:** The *p-value* column displays the association strength. Higher values have stronger association. This column can be both sorted and filtered with a minimum value. The association valued link to a box-plot of the marker-trait segregation. **E:** The *Annotation* column displays an AGI code if the marker can be found in a gene model. The AGI code is linked to the TAIR website and holding the mouse over the AGI code will display the TAIR annotation. The TAIR version displayed is dependent on the marker dataset version. The Annotation column can be filtered by both AGI codes and model annotation terms. **F:** The *Polymorphism* column displays the polymorphisms at the current position either as the nucleotide change from Col-0 if the marker is non-coding, or as the amino acid change if the marker is coding. Currently, this column can be neither sorted nor filtered. **G:** It is also possible to filter the markers based on their Minimum Allele Frequency (MAF). In future, this filter will be extended to the Hardy-Weinberg Equilibrium.

Figure 3: Computational run-time for Matapax. The run time of Matapax was tested for increasing numbers of traits and accessions. The results were plotted as a LOWESS fit of the runtime as the number of traits and accessions increased where the solid line is the fit and the dotted lines the 95% confidence interval. **A.** The run-time over a number of traits increasing from 2 to 25. As 12 nodes were available on the server when the jobs were run, there is a rough doubling of computation time every 12 traits. There is also a slight extra cost for each trait in the post-processing step of Matapax as the results are inserted into a relational database. **B:** The run-time for both GAPIT and EMMA over increasing numbers of accessions; 2, 5, 10, 20, 50, 100, 200 and 1000. EMMA was unable to complete running within reasonable time when tested with 500 and 1000 accessions.

Figure 4: Correlation of published with Matapax association p-values. The association p-values of the published data were correlated with the association p-values obtained by Matapax. For the vast majority of traits, Matapax is able to obtain identical results to the published data. However, Matapax is unable to reproduce the association results for a few traits. Possible reasons are discussed in the body of the text.





MAF threshold: 0.01 HWE threshold: Not yet implemented

Trait	Chr	Position	p-value	Annotation	Polymorphism
33 avrRpm1	3	2227[illegible]	4.[illegible]e-04	AT[illegible]0	*, L
33 avrRpm1	3	2337844	2.97e-03	AT[illegible]	
33 avrRpm1	3	2221686	8.30e-03		
33 avrRpm1	3	2225040	8.30e-03		
33 avrRpm1	3	2225659	8.30e-03		**GG**, AA
33 avrRpm1	3	2225899	8.30e-03		**CC**, GG
33 avrRpm1	3	2229815	8.30e-03	AT3G07050	* (**TAG**, TAA)
33 avrRpm1	3	2230189	8.30e-03	AT3G07050	**Q**, R
33 avrRpm1	3	2229647	1.81e-02	AT3G07050	E (**GAA**, GAG)
33 avrRpm1	3	2231452	3.24e-02	AT3G07050	**F**, L
33 avrRpm1	3	2167717	4.71e-02		**GG**, CC
33 avrRpm1	3	2169291	1.21e-01		**CC**, AA
33 avrRpm1	3	2237370	1.21e-01		**CC**, TT
33 avrRpm1	3	2222862	1.63e-01		**CC**, GG
33 avrRpm1	3	2337970	1.63e-01	AT3G07330	I (**ATT**, ATA)
33 avrRpm1	3	2230556	2.24e-01	AT3G07050	**AA**, GG
33 avrRpm1	3	2231864	2.24e-01	AT3G07050	**AA**, TT
33 avrRpm1	3	2231938	2.24e-01	AT3G07050	*, W
33 avrRpm1	3	2232004	2.24e-01	AT3G07050	**F**, C
33 avrRpm1	3	2291826	2.69e-01	AT3G07200	**GG**, TT

Contains an N-terminal tripartite nucleotide binding leucine-rich repeats. Confers resistance to Pseud avirulence genes avrB and avrRpm1.

1 selected | All | From | To | Max p- | Search | Not yet implemented

Showing 1 to 20 of 216,081 entries (filtered from 23,125,710 total entries)