Open access • Book Chapter • DOI:10.1007/978-1-4939-6658-5_5

Mining Plant Genomic and Genetic Data Using the GnpIS Information System.
— Source link ↗

Anne-Françoise Adam-Blondon, Michael Alaux, Sophie Durand, Thomas Letellier ...+17 more authors

Institutions: Université Paris-Saclay, Agro ParisTech

Related papers:

- New Technologies InterStoreDB: A Generic Integration Resource for Genetic and Genomic Data

- SNPTrackTM : an integrated bioinformatics system for genetic association studies

- BASC: an integrated bioinformatics system for Brassica research

- AraPheno and the AraGWAS Catalog 2020: a major database update including RNA-Seq and knockout mutation data for Arabidopsis thaliana.

- Analysis and visualization of Arabidopsis thaliana GWAS using web 2.0 technologies.

Mining plant genomic and genetic data using the GnpIS information system

Anne-Francoise Adam-Blondon, Michael Alaux, Sophie Durand, Thomas Letellier, Guillaume Merceron, Nacer Mohellibi, Cyril Pommier, Delphine Steinbach, Françoise Alfama-Depauw, Joëlle Amselem, et al.

HAL Id: hal-03206731
https://hal.archives-ouvertes.fr/hal-03206731

Submitted on 23 Apr 2021

Mining plant genomic and genetic data using the GnpIS information system

Running Head : Mining GnpIS

Authors

A-F Adam-Blondon[1], M. Alaux[1*], S. Durand[1*], T. Letellier[1*], G. Merceron[1*], N. Mohellibi[1*], C. Pommier[1*], D. Steinbach[1,2*], F. Alfama[1], J. Amselem[1], D. Charruaud[1,3], N. Choisne[1], R. Flores[1], C. Guerche[1], V. Jamilloux[1], E. Kimmel[1], N. Lapalu[1,4], M. Loaec[1], C. Michotey[1], H. Quesneville[1]

* The authors have contributed equally to the work

[1] Research Unit in Genomics-Info UR1164, INRA, Université Paris-Saclay, Versailles, Route de Saint-Cyr, 78026 Versailles Cedex, France.

[2] Research Unit GQE-Le Moulon UMR 320, INRA, Université Paris-Sud, Université Paris-Saclay, CNRS, AgroParisTech, Ferme du Moulon, 91190 Gif-sur-Yvette, France

[3] ADRINORD Espace Recherche Innovation, 2 rue des Canonniers, 59800 Lille, France

[4] UMR BIOGER, UMR1290, INRA, AgroParisTech, 1 avenue Lucien Bretignières, 78850 Thiverval-Grignon, France

Corresponding author : A-F Adam-Blondon

Research Unit in Genomics-Info UR1164, INRA, Université Paris-Saclay, Versailles, Route de Saint-Cyr, 78026 Versailles Cedex, France.

Email : afadam@versailles.inra.fr

Summary

GnpIS is an information system designed to help scientists working on plants and fungi to decipher the molecular and genetic architecture of trait variations by facilitating the navigation through genetic, genomic and phenotypic information. The purpose of the present chapter is to illustrate how users can (i) explore data sets from phenotyping experiments in order to build new datasets for studying genotype x environment interactions in traits, (ii) browse into the results of other genetic analysis data such as GWAS to generate or check working hypothesis about candidate genes, or to identify important alleles and germplasms for breeding programs and (iii) explore the polymorphism in specific area of the genome using InterMine, JBrowse tools embedded in the GnpIS information system.

Key Words

GnpIS, GWAS, polymorphism, phenotypes, markers, genetic resources, GMOD genome browser, InterMine

1. Introduction

In the last decades, many research programs in biology have been profoundly affected by a major technological evolution: the very fast development of high-throughput technologies that produce massive amounts of data (e.g. new generation sequencing methods, high-density genotyping arrays, improvements in proteomics and metabolomics data acquisition, image treatments for phenotype observations). One of the consequences of the emergence of these technologies is the critical need for managing high volumes of data in efficient computer systems. In addition, strong initiatives from international policy makers stressed the commitment for academics research to provide open and interoperable data resources, which should enhance the reuse of datasets, hence reducing the cost of research projects (see for instance [1,2]).

GnpIS is an original information system built to act as a central repository for plant science data. It implements an efficient and reliable information system over a computer infrastructure able to meet the challenge of huge data management. It was designed to integrate heterogeneous data types, allowing to bridge genomics, genetics and phenomics data [3]. As a reference data warehouse, it promotes data exchange and dissemination to the international scientific community. For this purpose, GnpIS uses internationally recognized formats, controlled vocabularies or ontologies, or other standards for data exchange [4].

The main focus of GnpIS is to help plant and fungi scientists to decipher the molecular and genetic architecture of trait variations by facilitating the navigation through genetic, genomic and phenotypic information. The purpose of the present chapter is to illustrate how users can (i) explore data sets from phenotyping experiments used for genetic analysis to build datasets for studying genotype x environment interactions in traits, (ii) browse into the data results of genetic analysis such as genome-wide association studies (GWAS) to generate new working hypothesis about candidate genes, or to identify important alleles and germplasms for breeding programs, and (iii) explore the polymorphism in specific area of the genome.

2. Materials

GnpIS relies on cutting edge data warehouse technologies using (i) a relational database to integrate and link data in a coherent and consistent framework, and (ii) data marts and (iii) NoSQL technologies to speed up the searches. It can be accessed through web graphical interfaces at the URL: https://urgi.versailles.inra.fr/gnpis. It guides the users to dedicated interfaces: https://urgi.versailles.inra.fr/Data/Genome/Genome-data-access for accessing genomes, https://urgi.versailles.inra.fr/GnpMap for genetic maps, https://urgi.versailles.inra.fr/GnpSNP for polymorphisms, https://urgi.versailles.inra.fr/ephesis/ for phenotyping experiments, https://urgi.versailles.inra.fr/association/ for phenotypes to genotypes associations, and https://urgi.versailles.inra.fr/siregal/ for the genetic resources.

It gives access to data about genomes via the GBrowse and JBrowse, two GMOD genome browsers [5]. Genetic resources data follows the MultiCrop Passport Data format (MCPD, [6]). Genetic maps, QTLs (Quantitative Trait Loci), phenotypes to genotypes associations, polymorphism data such as SSR (Simple Sequence Repeats) and SNP (Single-Nucleotide Polymorphism) use relevant ontologies or controlled vocabularies when they exist.

Moreover, the GnpIS portal allows intuitive google-like queries using a free text field, or complex queries on dedicated data subsets of the information system (grapevine data and wheat chromosome 3B data) using dedicated InterMines ([7]; http://urgi.versailles.inra.fr/GrapeMine/begin.do; http://urgi.versailles.inra.fr/Wheat3BMine/begin.do).

The datasets used to illustrate the present chapter were produced by different groups of researchers working on traits evaluation in wheat [8], on association genetics in tomato [9], on grapevine genomics [10, 11]. Many others are available in GnpIS, derived from projects in which the French Institute for Agronomical Research (INRA) is partner.

3. Methods

3.1. Building a dataset for the meta-analysis of phenotyping experiments

Studying genetic determinism of traits requires the statistically sound phenotypic evaluation of genotypes panels. It is also true for the meta-analysis of experimentations obtained in multiple environments, that may allow a better understanding of the influence of fluctuating environments (year/climate/location, management techniques, etc.). The building of such datasets for meta-analysis relies on a data integration process. This often implies to unambiguously identify in heterogeneous data sources the pivot scientific objects, such as plant materials or traits. Data integration is a long process that can be facilitated by integrative information systems such as GnpIS (Note 1). It also relies on the use of ontologies and controlled vocabularies for traits, and precise identifiers for germplasms, to ensure the consistency of the newly built datasets.

GnpIS gives access to multi-local and multi-annual phenotyping trials for several species and to the ontologies used. To select relevant data, three sets of filters are available: "Genotype" for selecting germplasm/plant material, "Variables" for trait and environment variables selection, and "Trial" for trials descriptors and metadata. The protocol below describes the way to build a dataset from a panel of evaluated genotypes, for a set of traits in a maximum number of experimental conditions, to better isolate the genetic component of the traits variation.

1. Go to https://urgi.versailles.inra.fr/gnpis and click on "Phenotypes" and on the left menu on "experimental data" or go directly to https://urgi.versailles.inra.fr/ephesis/ephesis/viewer.do

2. Plant material selection: the "Genotype" tab from the form is selected by default. In the example shown in fig. 1, the focus is on bread wheat by selecting the genus "*Triticum* L." and the species "*Triticum aestivum aestivum*".

3. Variables selection: clicking on the "Variable" tab shows all the variables consistent with the current search parameters, i.e. "*Triticum aestivum aestivum*". To focus on yield as a trait of bread wheat (for instance), it is possible to find in the Small grain cereal

Network Ontology the three following traits: "yield", "thousand grain weight" and "test weight". Clicking on these three variables and then on the "Results" button allows retrieving results from 635 trials (at the date this chapter was written, as in the rest of this document).

4. <u>Refine the selection</u>: go back to the selection form using the "Back to form" button and select the "Trial" tab. It is possible, for instance, to refine by data set: selecting "INRA Wheat Network BRC accession (A series)" in the "Add trial by data set" box allows retrieving 230 trials after clicking on the "Results" button (fig. 2). It is also possible to filter by years and/or locations, which is a way to focus on some macro environment variables. Leaving empty a filter box means not filtering on the corresponding criteria.

5. <u>Data export:</u> the result page of GnpIS gives a summary of the filters applied and a preview of the results of the query. Two tables are provided: the "Trial table" with the average results from several repetitions and the "Rep table" that gives access to the raw data for each repetition. It is possible to download all the results either in a simple tabular format or in an ISA Tab exchange data format (fig. 2). The ISA Tab exchange data format ensures the traceability of the metadata associated to the data [4]. The downloaded data includes the year, the genotype (LotNumber), information on the position of the lot in the experiment (X and Y) and information on the treatments used in the experiment (here in the "itk" column). It allows seeking statistical correlation between those different factors (fig. 2). It is possible for users having a private access (Note 2) to the database to get climatic data associated to the experiments selected (Note 3).

3.2. GWAS data exploration

With the increased number of plant genome sequencing projects and the development of high-throughput genotyping methods, it is now possible for a growing number of crops to genotype enough SNP to conduct association mapping approaches, either at the whole-genome scale

(GWAS) or in regions in which QTL have been detected by genetic linkage approaches. Thanks to the combination of approaches based on association genetics and QTL mapping, it is now possible i) to fine map QTL ii) to identify new markers useful for selection, iii) to find interesting alleles in genetic resources collections and to use them in new material dedicated to selection. Nowadays, GWAS is becoming a mandatory approach in the strategies developed by scientists and breeders to understand the genetic architecture of traits, and the relationships between genotype and phenotype. GnpIS was recently extended to store GWAS experiments results including data and metadata.

GnpIS provides as metadata (i) the plant material used in the GWAS experiment with unambiguous identification of accessions, (ii) the name of the panel in which are gathered these genotypes and some information on its genetic structure: kinship between individuals and linkage disequilibrium (LD) between polymorphisms, and (iii) the statistical model used to detect associations between markers and phenotypes. The data provided consist in (i) the genotypic and the phenotypic values used as entry in the model and (ii) the *p*-values of the association between each marker and the studied traits. The interface allows exploring GWAS data to discover the best markers and alleles associated to traits of interest with Q-Q Plots (quantile-quantile plots) and Manhattan Plots (−log10(P) genome-wide association plots). In turn, such marker identification allows finding candidate genes underlying the phenotypic variation through their integration with other data types present in the GnpIS data warehouse, such as genomic annotation data, expression data, genetic mapping data or other QTL data.

1. Go to https://urgi.versailles.inra.fr/gnpis and click on "Association" or go directly to https://urgi.versailles.inra.fr/association.

2. In the left menu, in "Queries", select "Associations" and then select the genus of the species of interest (e.g. *Solanum* L.)

3. It is then possible to filter the data by traits, markers or panel. For instance, selecting Ascorbate content and Fructose and filtering on the "solcap" panel with the Panel tab, retrieved at the date the chapter was written 5524 unique markers with a result of association to one or both traits. All the results are displayed in two tables and

downloadable into a csv format (fig. 3). It is possible to visualize a boxplot of the phenotypic value of each allele at the marker (fig. 3). Note that at the bottom of the boxplot, a table sorts the accessions of the panel according to their genotypes at the marker and displays of each of them, their phenotypic value (fig 3). The name of the accessions is clickable to display more information.

4. The detail of the phenotyping and genotyping experiments used as inputs can be displayed by clicking on the panel name in the table. It displays the panel card and all the experiments linked to it. The phenotyping and genotyping data are displayed (Note 4) and downloadable in GnpIS (see paragraph 3.1).

5. Several tools allows going deeper into the results:

    a. Graphical and dynamic display of the associations with a Q-QPlot or a Manhattan Plot (fig. 3).

    b. New filters (by value, by chromosome, by trait…) facilitate selection and display.

    c. Links to markers positions on the reference genome of the selected genus (fig. 4). A direct link to the browser of the reference genome (when available) is clickable in the "Association results" table, column "Genome Browser" (fig. 4). In the Manhattan plot, a pop-up window giving basic information on a marker (name, chromosome, position, *Pval, -logPval*), a link to the marker card and again to the browser can be obtained by pointing the marker (zoom until it works; fig. 4). Depending on genomes, the browser gives access to gene annotations, expression data, epigenetic marks, polymorphisms, etc. The marker card gives access to other types of information linked to the marker: genetic and physical maps, associated QTLs, sequences.

3.3. Exploring sequence diversity in a genomic region

GnpIS allows bridging regions identified in QTL analysis, or association mapping, with genomic annotations and other associated genetic information. For instance, the GSVIVG01013466001

grapevine gene (other ID: Vv18s0122g00190) is located in a genome region under study in relation with the Fleshless berry locus [9].

1. One way to retrieve information on the region of interest is to use the GrapeMine (http://urgi.versailles.inra.fr/GrapeMine/) by clicking on InterMine, then on Grape and typing Vv18s0122g00190 in the search box. The tool gives access to a page gathering all the information on the gene (structure, sequence, annotation etc., fig. 5) and allows discovering other data stored in GnpIS via the links available in the right panel (i.e. the Vitis Gbrowse, datamart or else using the quick search, see below). A specific tutorial on GrapeMine (and on Wheat3BMmine) is available via a "documentation" link at the top right side of the window. One of the main interests of the InterMine tool is that it allows retrieving information linked to a genome interval or to lists of features and not only to a single feature as shown in figure 5.

2. The second way to find useful data is to go directly to https://urgi.versailles.inra.fr/gnpis and enter the gene ID (GSVIVG01013466001) in the text-based quick search box. Clicking on an exact match sends on a zoom of the GSVIVG01013466001 gene in the Vitis 12X Genome browser (GBrowse) (fig. 6).

3. Go to the menu "Select Tracks" at the top of the search GBrowse box (Note 5). A list of tracks is provided that can be added or hidden from the GBrowse view. Add the "Genoscope annotation" and the "V1_NR" annotation tracks. Click "All on" in the polymorphism check box to show all the tracks of this category or select the three tracks from the "berry size project" (example in fig. 6). Click on the "Back to the browser" link to get the genomic view updated with the selected tracks.

4. The gene transcripts predicted by two different automatic annotations corresponding to the same gene (the gene ID is GSVIVG01013466001 in the Genoscope annotation and Vv18s0122g00190 in the CRIBI V1 annotation) are shown together with all the SNPs and insertions/deletions that were detected in the different parts of the gene.

5. It is possible to select any of the SNP markers in the area and to click on it to open a pop-up that gives a minimum information on the SNP (SNP ID, the type of

polymorphism, the position of the genome) and additional links like for instance towards Ensembl Plants. In the example shown in fig. 6, the SNP chosen is Vv_1274223. Clicking on "Link to SNP card in GnpIS" in the pop-up allows to open the "Polymorphic locus card" of the GnpIS database in which additional information can be found (https://urgi.versailles.inra.fr/GnpSNP/snp/card/snp.do?name=Vv_1274223; fig. 7)

   a. The flanking sequences of this SNP that can be downloaded in a FASTA format. These sequences are based on the reference sequence that was used for SNP calling, which is also downloadable.

   b. All the genotypes observed at the same genome position for a re-sequencing experiment are listed ("Linked with variations" block) and can be viewed in more details (click on "view list") in a clickable table. The genotype table (https://urgi.versailles.inra.fr/GnpSNP/snp/genotypeAction.do?action=BATCH &batchId=1608) can be obtained by clicking on "view list of variations", then in the table appearing, on the name of the experiment (in the example shown in fig. 7, "VVC2972A-batch1") and finally on "view the genotype table" (fig.8). This matrix of the genotypes for all the markers of the region and all the plant material involved in this resequencing analysis can be exported in a csv format (export button at the bottom of the page).

   c. All the corresponding sequenced accessions ("Linked with lines" block in fig. 7) and their taxon name are listed in a table (click on "view list"). Clicking on the accession name gives access to all the experiments of SNP detection for the given accession.

6. To know more about the accessions, go back to the text-based quick search box in GnpIS portal (https://urgi.versailles.inra.fr/gnpis) and enter the accession ID (e.g. 324Mtp43, or "Charger"). The system returns a table with several options "phenotypes, genetic resources, polymorphisms…", depending on the data linked to the term found in the database. Genetic resources links to all the information associated to the accession: accession name, taxon, holding institute, origin, evaluation data, etc. (see

fig. 9). Clicking on the basket at the bottom right allows ordering the accession to a Biological Resource Center if the accession is maintained at INRA.

7. Another way is to click on the "Genetic Resources" module of GnpIS and on "Taxons" in the left menu. Typing the cultivar name "Cabernet franc" in the "Scientific name" box returns all the accessions available for this cultivar together with the linked data (fig 10).

4. Notes

[1]GnpIS is under continuous improvement: problems can be indicated by an Email to urgi-contact@versailles.inra.fr

[2]The management and access to private data is made possible in the frame of project in collaboration with INRA.

[3]Access to climatic data must be asked only for a short list of experiments as the query retrieves a large number of data.

[4]Some link may display an error message. This is often due to the fact that the linked data are private.

[5]Some unwanted old configurations may be stored in your cache for some applications or pages of GnpIS interface. You may have to clean the cache of your web explorer to solve the problem or type Ctrl+F5.

5. Acknowledgements

Charcosset, Patrice This, Gilles Charmet, François-Xavier Oury, Arnaud Gauffreteau, Etienne Paux and Frédéric Choulet. GnpIS has been developed in the last 4 years with the support of INRA, the ANR projects of the "Investment for the Future" call Phenome (https://www.phenome-fppn.fr/), Aker (http://www.aker-betterave.fr/en/), Amaizing (ANR-10-BTBR-03), Breedwheat (ANR-10-BTBR-02) and Peamust (ANR-11-BTBR-02), the ANR project GnpAsso (ANR-10-GENM-0006), the TransPLANT FP7 European infrastructure project (project n°283496).

6. References

1- http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

2- Kaiser J (2015) U.S. agencies fall in line on public access. Science 349 :167

3- Steinbach, D; Alaux, M; Amselem, J; Choisne, N; Durand, S; Flores, R; Keliet, AO; Kimmel, E; Lapalu, N; Luyten, I; Michotey, C; Mohellibi, N; Pommier, C; Reboux, S; Valdenaire, D; Verdelet, D; Quesneville, H (2013) GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database (Oxford), Bat058. doi: 10.1093/database/bat058

4- Krajewski P, Chen D, Ćwiek H, van Dijk ADJ, Fiorani F, Kersey P, Klukas C, Lange M, Markiewicz A, Nap JP, van Oeveren J, Pommier C, Scholz U, van Schriek M, Usadel B, Weise S (2015) Towards recommendations for metadata and data handling in plant phenotypic experiments. J Exp Bot, 66(18):5417-27. doi:10.1093/jxb/erv271

5- Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH (2009) JBrowse: a next-generation genome browser. Genome Res, 19(9):1630-8. doi: 10.1101/gr.094607.109.

6- http://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v2-mcpd-v2/

7- Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X, Micklem G (2012) InterMine:

a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics, 28(23):3163-5. doi: 10.1093/bioinformatics/bts577.

8- Oury FX, Godin C, Mailliard A, Chassin A, Gardet O, Giraud A, Heumez E, Morlais JY, Rolland B, Rousset M, Trottet M, Charmet, G (2012) A study of genetic progress due to selection reveals a negative effect of climate change on bread wheat yield in France. Europ J Agronomy, 40 , 28-38 . doi: 10.1016/j.eja.2012.02.007

9- Sauvage C, Segura V, Bauchet G, Stevens R, Do PT, Nikoloski Z, Fernie AR, Causse M (2014) Genome-Wide Association in Tomato Reveals 44 Candidate Loci for Fruit Metabolic Traits. Plant Physiol, 165, 1120-1132. DOI: 10.1104/pp.114.241521

10- Houel C, Bounon R, Chaïb J, Guichard C, Péros JP, Bacilieri R, Dereeper A, Canaguier A, Lacombe T, N'Diaye A, Le Paslier MC, Vernerey MS, Coriton O, Brunel D, This P, Torregrosa L, Adam-Blondon AF (2010) Patterns of sequence polymorphism in the fleshless berry locus in cultivated and wild *Vitis vinifera* accessions. BMC Plant Biol, 10:284. doi:10.1186/1471-2229-10-284

11- Adam-Blondon A-F, Jaillon O, Vezzulli S, Zharkikh A, Troggio M, Velasco R (2011) Genome Sequence Initiatives. in: A-F Adam-Blondon, JM Martinez-Zapater, Chittaranjan Kole (eds) Genetics, Genomics and Breeding of Grapes. Science Publishers and CRC Press. pp 211-234. ISBN 9781578087174
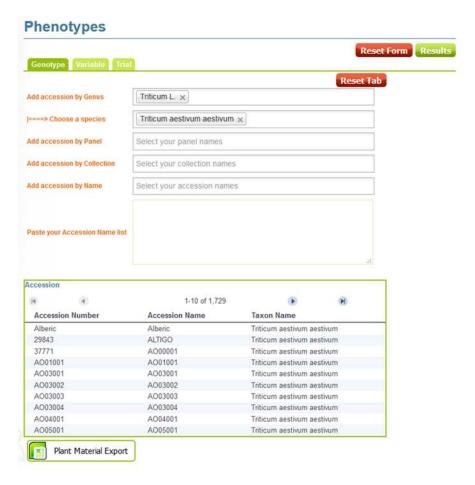
Figures Captions

figure 1. Query page of the GnpIS module for phenotypes (https://urgi.versailles.inra.fr/ephesis/ephesis/viewer.do). A filter on experiments involving bread wheat genotypes has been applied. It is possible to export the list of the selected accessions (plant material) and their identifiers in a tabular format.
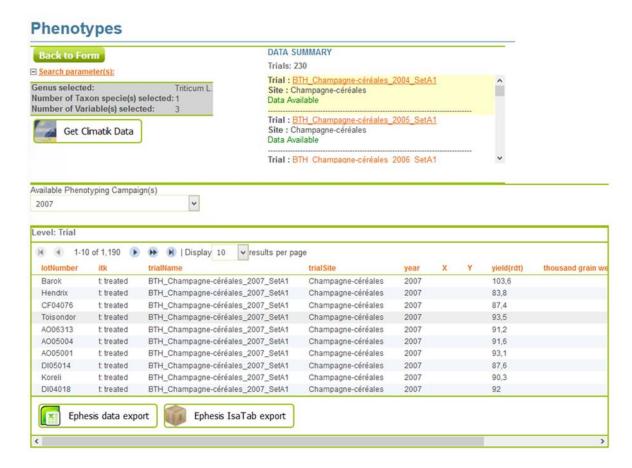
figure 2. Result table of the GnpIS module for phenotypes showing a summary of the Search parameters and a preview of the selected trials.

figure 3. Display of the results of association genetic analysis in GnpIS: summary of the experiment, a set of filters allowing refining the display, and the table of the *pValues* for each marker x trait combination. It is also possible to display the boxplot of the trait values observed for each allelic combination, a Q-Q Plot of the *pValues* or a Manhattan plot of the *–log(pValues)* of the markers x traits associations.
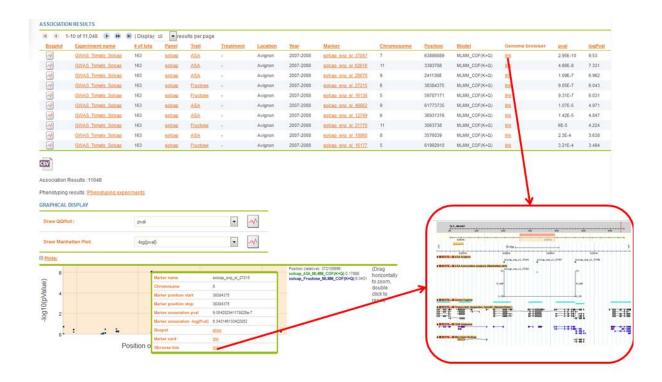
figure 4. GnpIS allows easy navigation between the association results presented in the table or plots and the annotation of the corresponding genomic regions displayed in genome browsers.
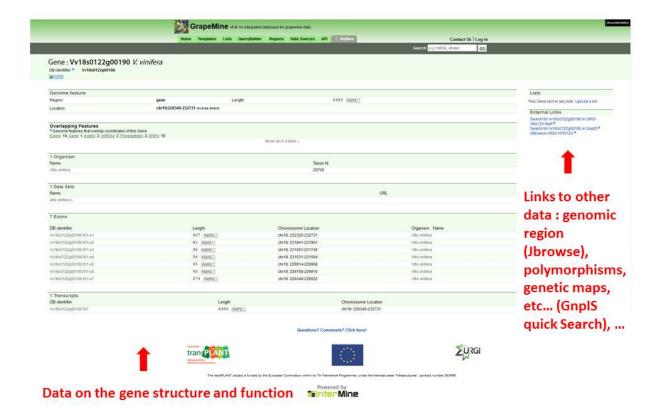
figure 5. Result page for one example of very simple search, information on one gene, using the GrapeMine.
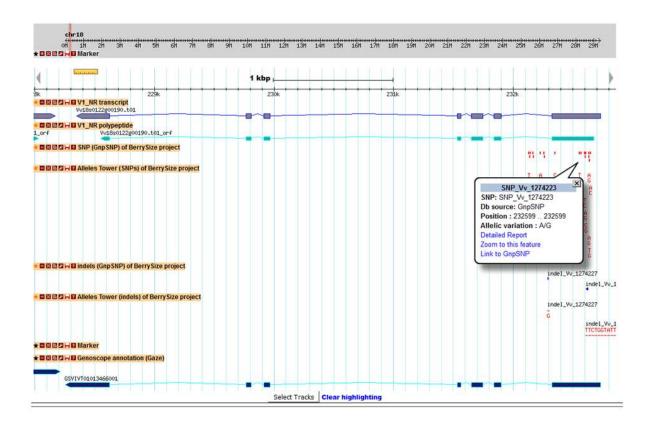
figure 6. Grapevine GBrowse showing a track with SNP markers and their linked information in a pop up.

figure 7. Polymorphic locus card of GnpIS: name, reference sequence, flanking sequences, variation observed, on which line it has been scored etc…
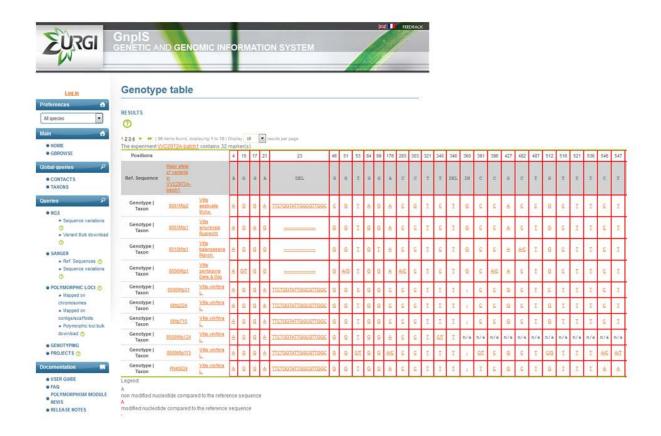
figure 8. Table describing all the genotypes of accessions stored in GnpIS for a given polymorphic marker.

figure 9. Information retrieved in the GnpIS database for the accession code "324Mtp43".

figure 10. Hits found in GnpIS for a taxon containing the string "Carbernet franc".