Part of the transPLANT European Plant Genomics Infrastructure

Part 6

3rd transPLANT Training Workshop - October 2014
Exploiting and understanding Solanaceous genomes

Advanced breeding of solanaceous crops using BreeDB

Richard Finkers
Plant Breeding,
Wageningen UR

WAGENINGEN UR
For quality of life

BreeDB

A database supporting analysis of quantitative data

14 October 2014, Richard Finkers (@rfinkers)

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Outline

- Introduction BreeDB (15:30-15:50)
- BreeDB Tutorial (15:50-16:35)
- Wrap up (16:35-16:45)

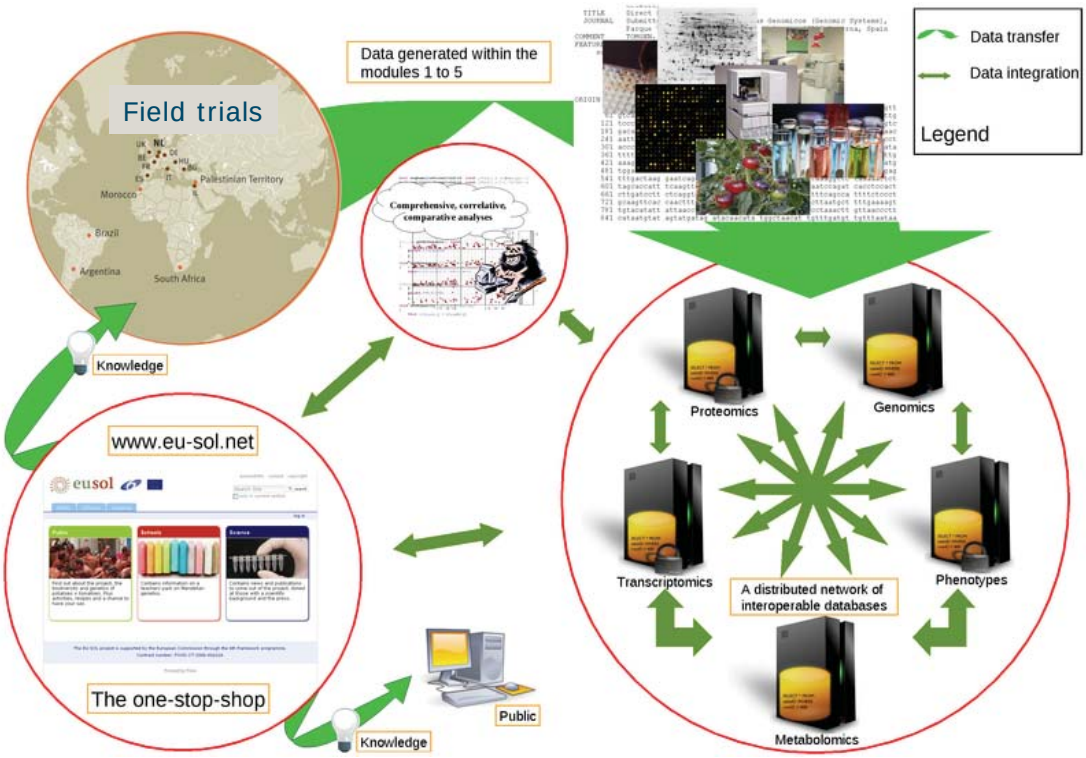
WAGENINGEN UNIVERSITY
WAGENINGEN UR

BreeDB

- BreeDB (https://www.eu-sol.wur.nl)
  - Structured storage of raw and analysed data
  - Exploration & Visualization of data
  - Integration of different data types (For example QTL analysis)
    - Real-time analysis using R
  - Adding meaning via integration with external resources
    - For example with SGN (tomato genome db)

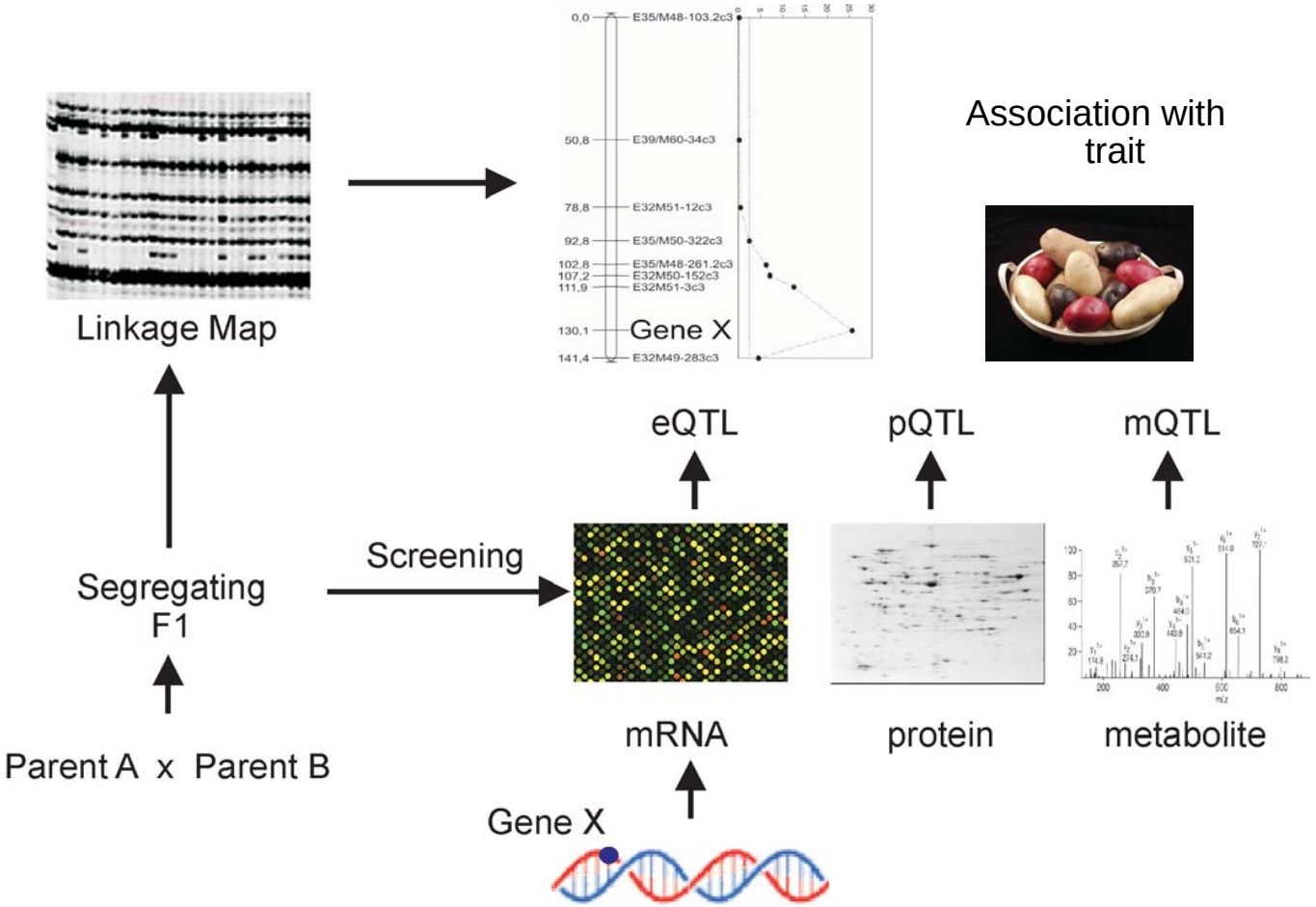
WAGENINGEN UNIVERSITY
WAGENINGEN UR

Data domains



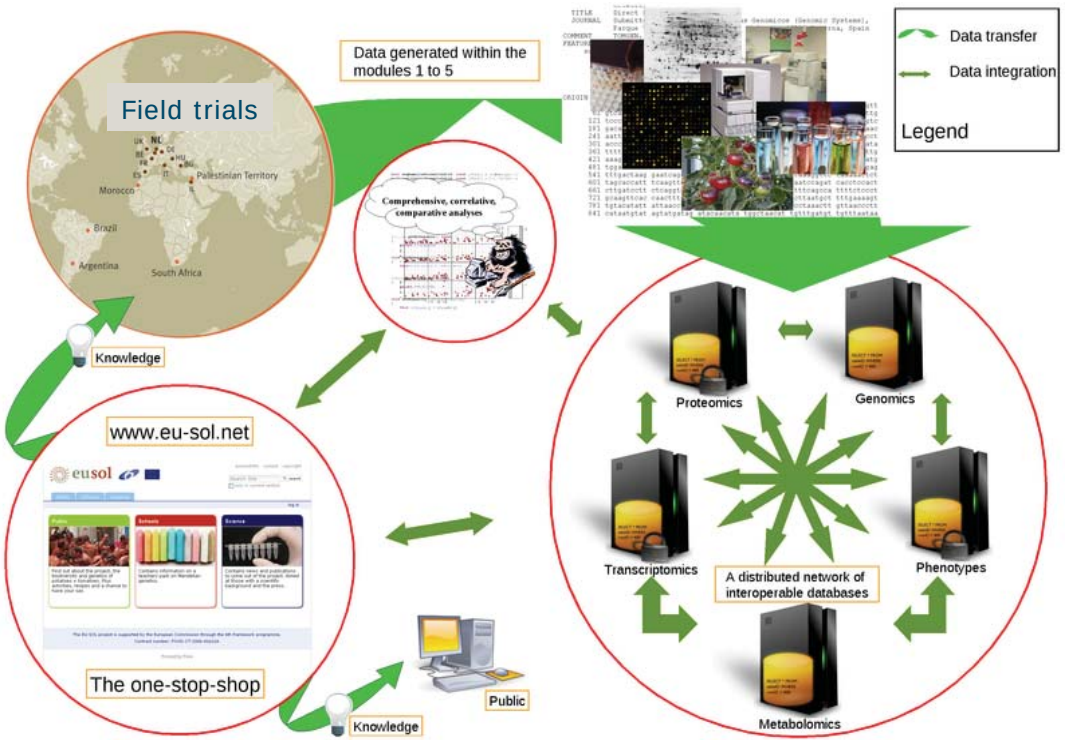
WAGENINGEN UNIVERSITY
WAGENINGEN UR

Acknowledgement: Jens Warfsman (EU-SOL)



WAGENINGEN UR

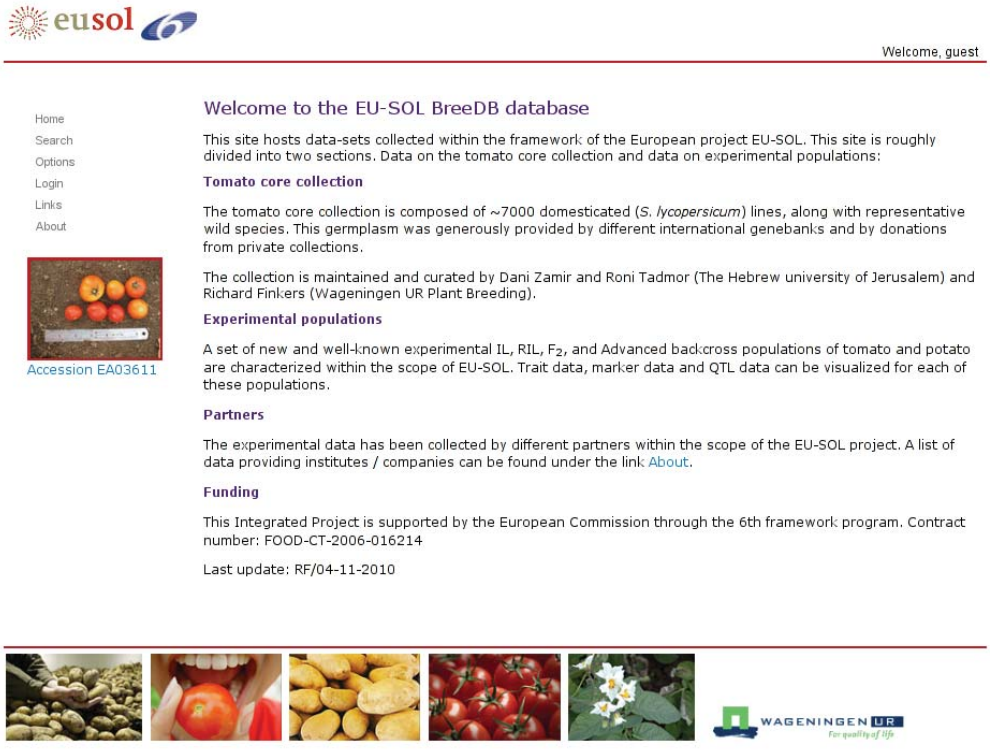
Data domains



WAGENINGEN UNIVERSITY
WAGENINGEN UR

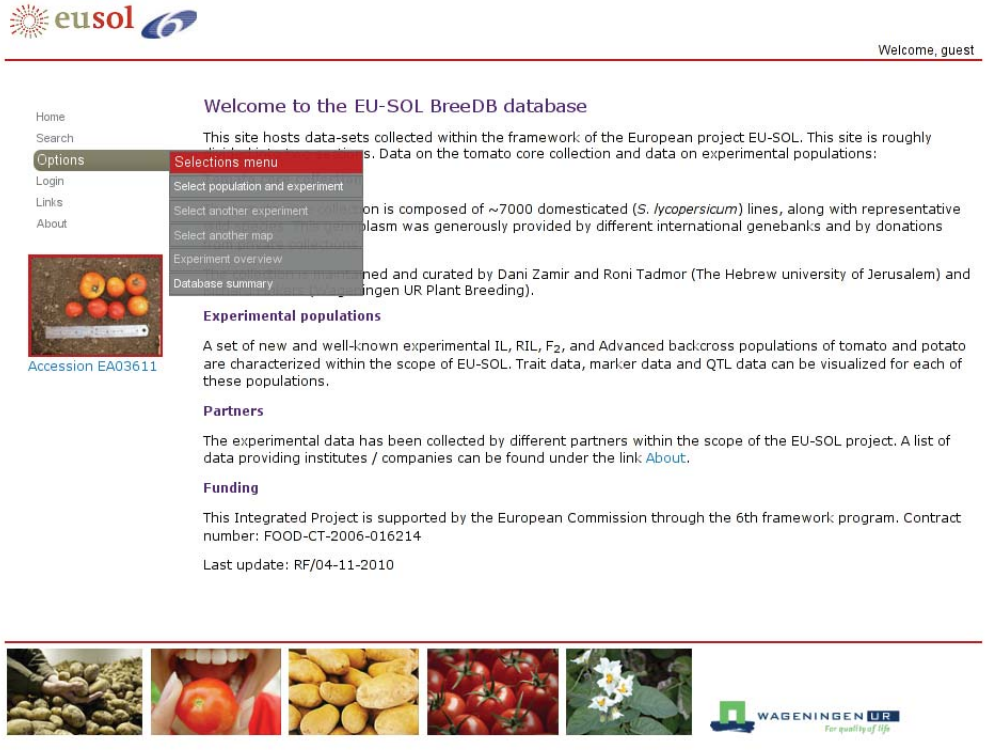
Acknowledgement: Jens Warfsman (EU-SOL)

BreeDB Interface



WAGENINGEN UNIVERSITY
WAGENINGEN UR

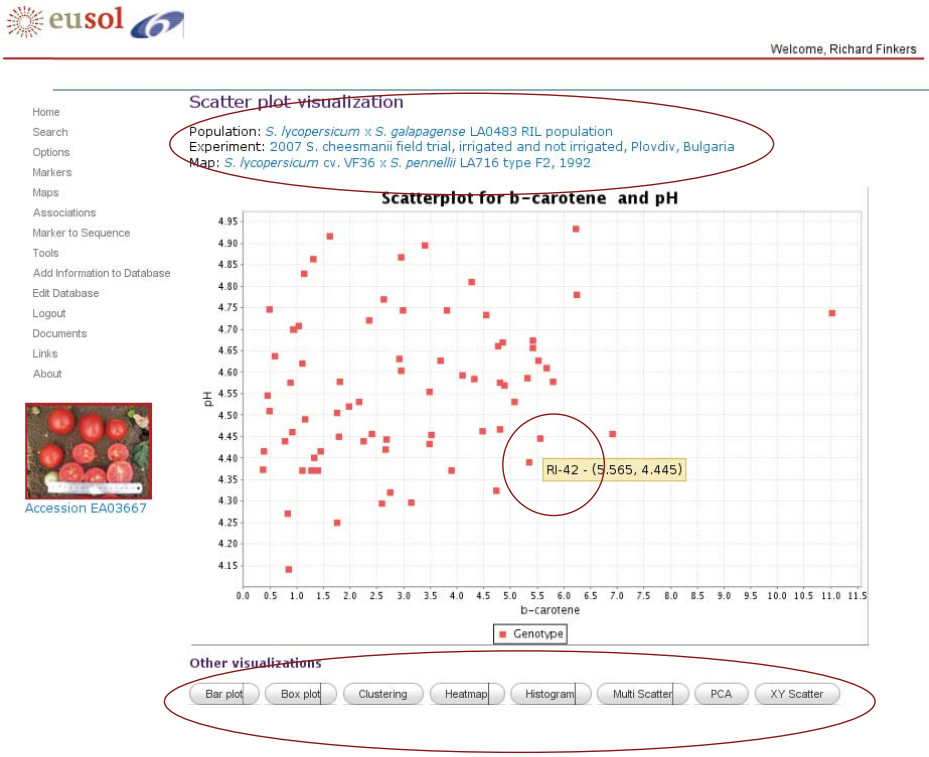
BreeDB Navigation (1)



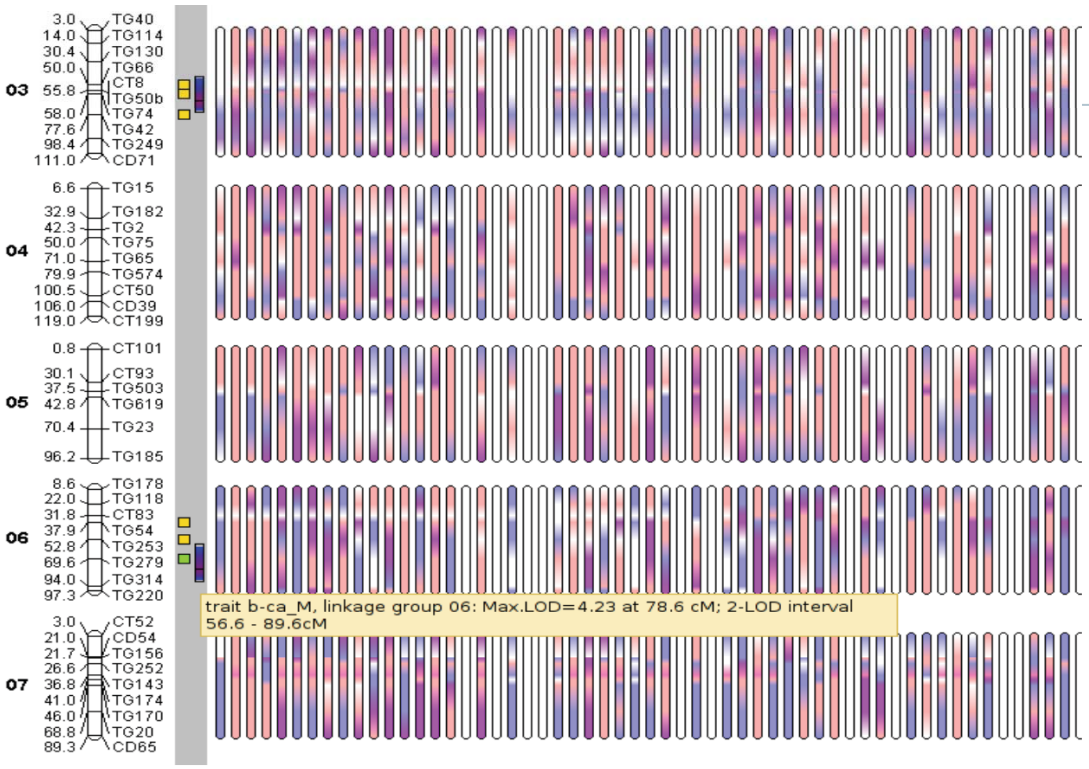
WAGENINGEN UNIVERSITY
WAGENINGEN UR

BreeDB Navigation (2)

Data exploration

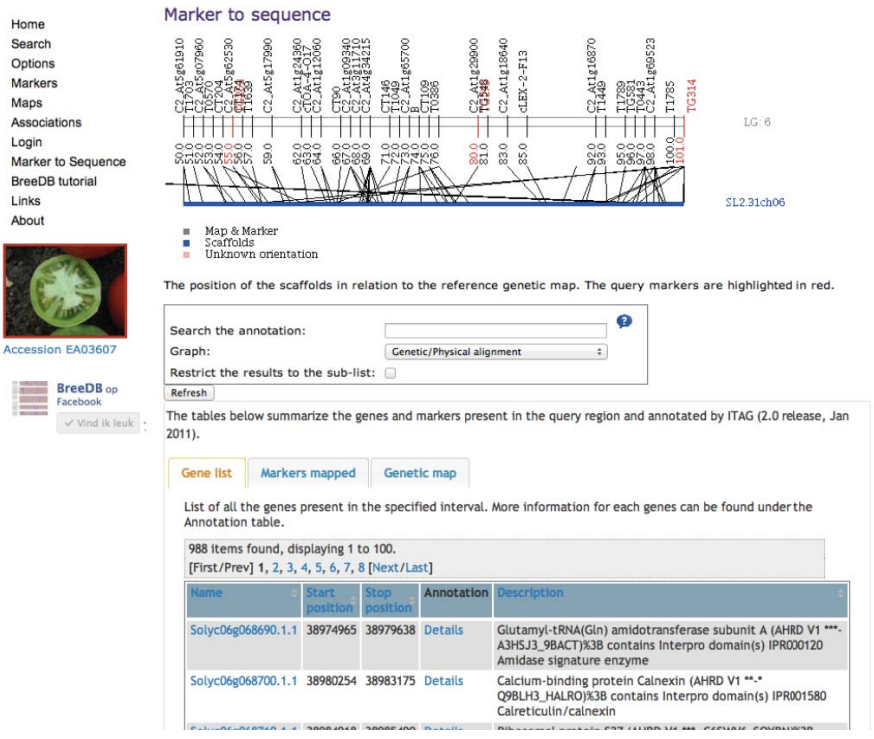


Kruskal-Wallis & Interval mapping



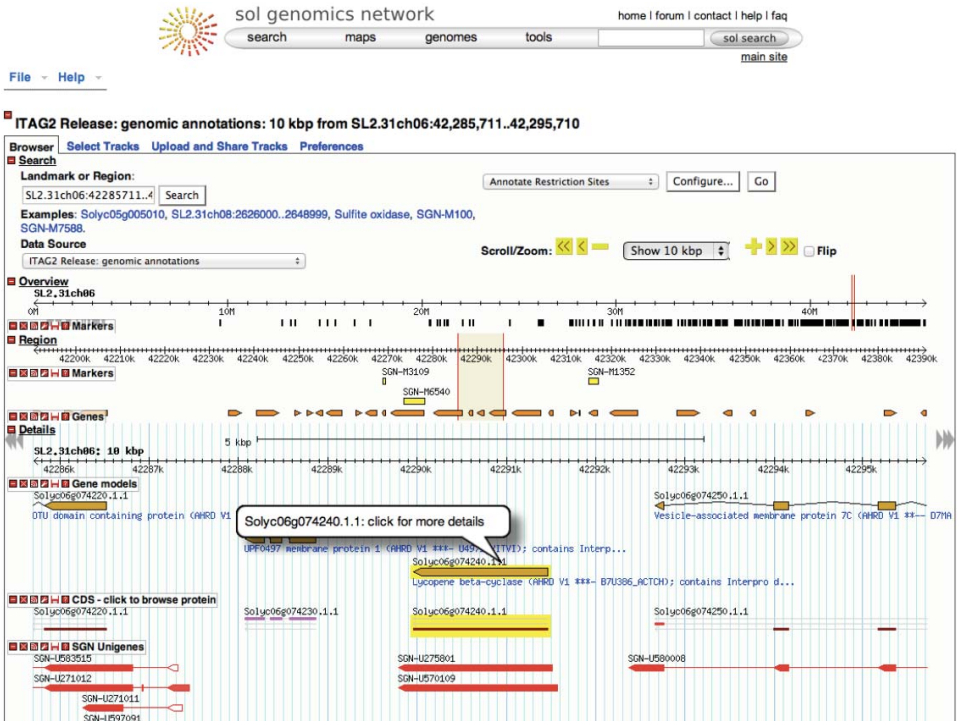
WAGENINGEN UNIVERSITY
WAGENINGEN UR

QTL region → Genome annotation



WAGENINGEN UNIVERSITY
WAGENINGEN UR

Tomato genome browser (GBrowse)



BreeDB Exercise Topics

- Germplasm collection
- Data exploration
- QTL analysis
- QTL LOD2 interval → physical interval

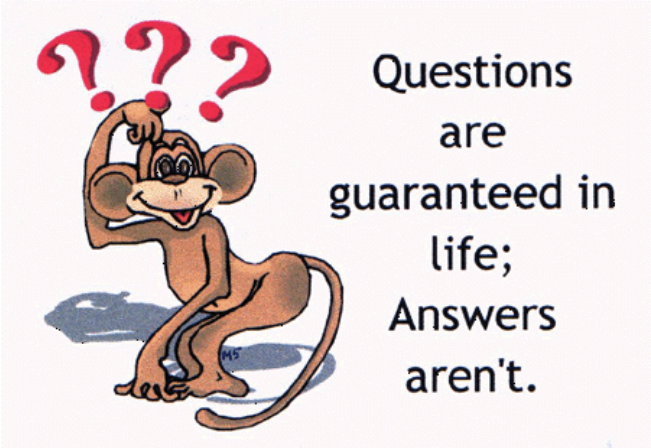
WAGENINGEN UNIVERSITY
WAGENINGEN UR

Hands on Time

WAGENINGEN UNIVERSITY
WAGENINGEN UR

BreeDB Exercises → Questions?

- Germplasm collection
- Data exploration
- QTL analysis
- QTL → physical interval



WAGENINGEN UNIVERSITY
WAGENINGEN UR

Introduction

Plant breeding research nowadays deals with a lot of data, resulting from phenotyping in field and greenhouse trials but also from high-throughput analysis of molecular markers, RNA transcripts (microarrays), proteins and metabolites. Databases are becoming indispensable to manage these data and to use them for the selection of plant material and for identification of markers, metabolites and mechanisms associated with important agronomic traits.

Within Wageningen UR Plant Breeding, we are developing a relational database which aims to support breeding for quantitative traits. The name of this database system is BreeDB.

What is BreeDB

BreeDB is a relational database which aims to support breeding for quantitative agronomical traits. The database can be explored through a web-based interface, which offers tools to present basic statistical overviews such as box plots and histograms, but also multivariate tools. Graphical genotyping tools are available to show molecular marker data and QTL data in relation to genetic linkage maps. In addition, photos of each accession can be shown together with a detailed report of observations made on this accession.

BreeDB is designed to store data from both inbreeding and out-breeding crop species and the analysis and visualization methods adapt automatically to the type of population on hand. For some features of BreeDB, integration with third party database is required. Therefore, a short introduction on database interoperability is included.

Database interoperability

BreeDB is primarily designed to store phenotypes and genotypes. However, breeding research involves integration of many other types of data (Figure 1). Ideally, each type of ~omics will have its own specialized database system for storage of observations and meta-information. BreeDB aims to integrate with these databases via web-service technology (Figure 1). For this, we implemented a MOBY (htttp://www.biomoby.org) compliant client and within BreeDB. BreeDB will also be accessible via a set of MOBY web-services.

The MOBY client is used to bridge the gap between for example, a QTL and the underlying genome sequence. Visualization of candidate genes, for QTLs identified in tomato or potato, will be integrated within BreeDB. For this, MOBY web-services will be implemented for sequence data stored in the *Solgenomics* network database (SGN, http://solgenomics.wur.nl). In addition, the moby client is also used to obtained metabolite data which is stored within the Golm EU-SOL database (EU-SOL only). Integrative efforts, with other databases, are ongoing.

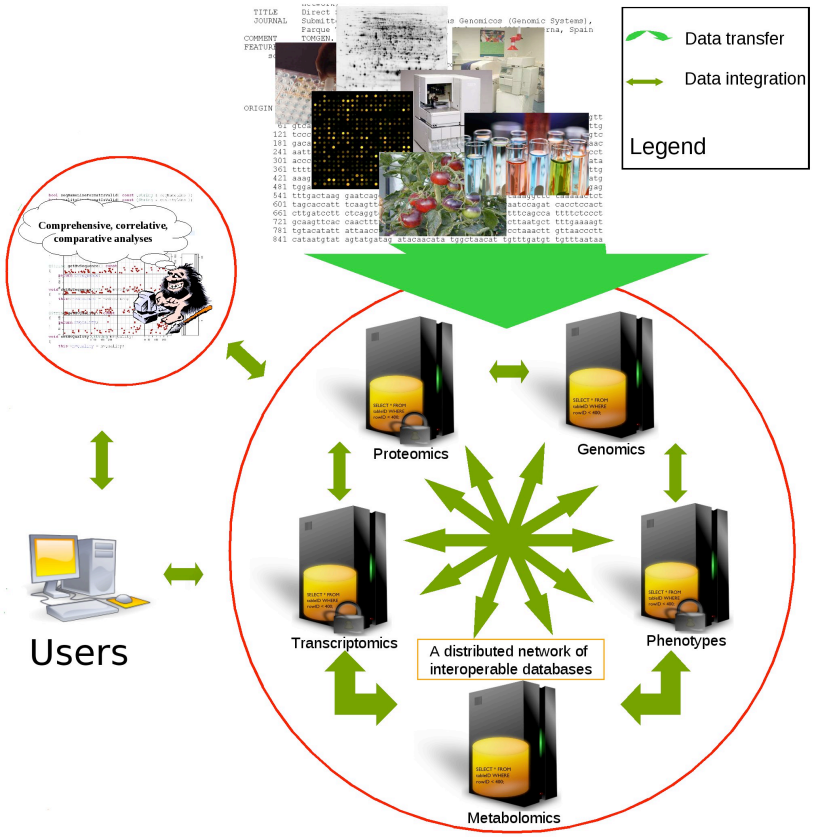


Figure 1: Interoperability among biological databases. Data of ~omics experiments, and their specific meta-information, should be stored within domain specific databases. Integration of these different resources can be achieved using web-service technology.

Data exploration & Visualization basics

It is often the case that some effort must be made to focus your attention on pertinent aspects of your data before true analysis can begin. This is almost universally true for large data sets, especially that data which was not gathered in a controlled or focused manner. But it is often also true for small data sets gathered with very rigid and specific techniques. Even very small data sets have myriad subsets, each of which might be especially pertinent to a given study.

Narrowing your focus so that you can thoroughly analyze data is problematic because you may lose important perspectives in doing so. But it remains an important task. The question is how to go about it. Data exploration is a methodology in which manual techniques are utilized to find one's way through a data set and bring important aspects of that data into focus for further analysis. Though such a methodology can be applied to data sets of any size or type, its manual nature makes it more reasonable for smaller data sets, especially those in which the data has been carefully gathered and constructed.

BreeDB contains a number of methods to aid end-users in exploration of the available data. These tools are in general tailor made for the type of data than can be explored. Each of them will be introduced in relation to their respective type of data / database module. However, before starting to discuss the different data exploration and analysis methods, we will start with giving a quick overview of the BreeDB user interface.

The BreeDB user interface

The BreeDB user interface is an interface that can be accessed using a web-browser. The interface consists mainly out of four regions: The header, the menu, the content and the footer.

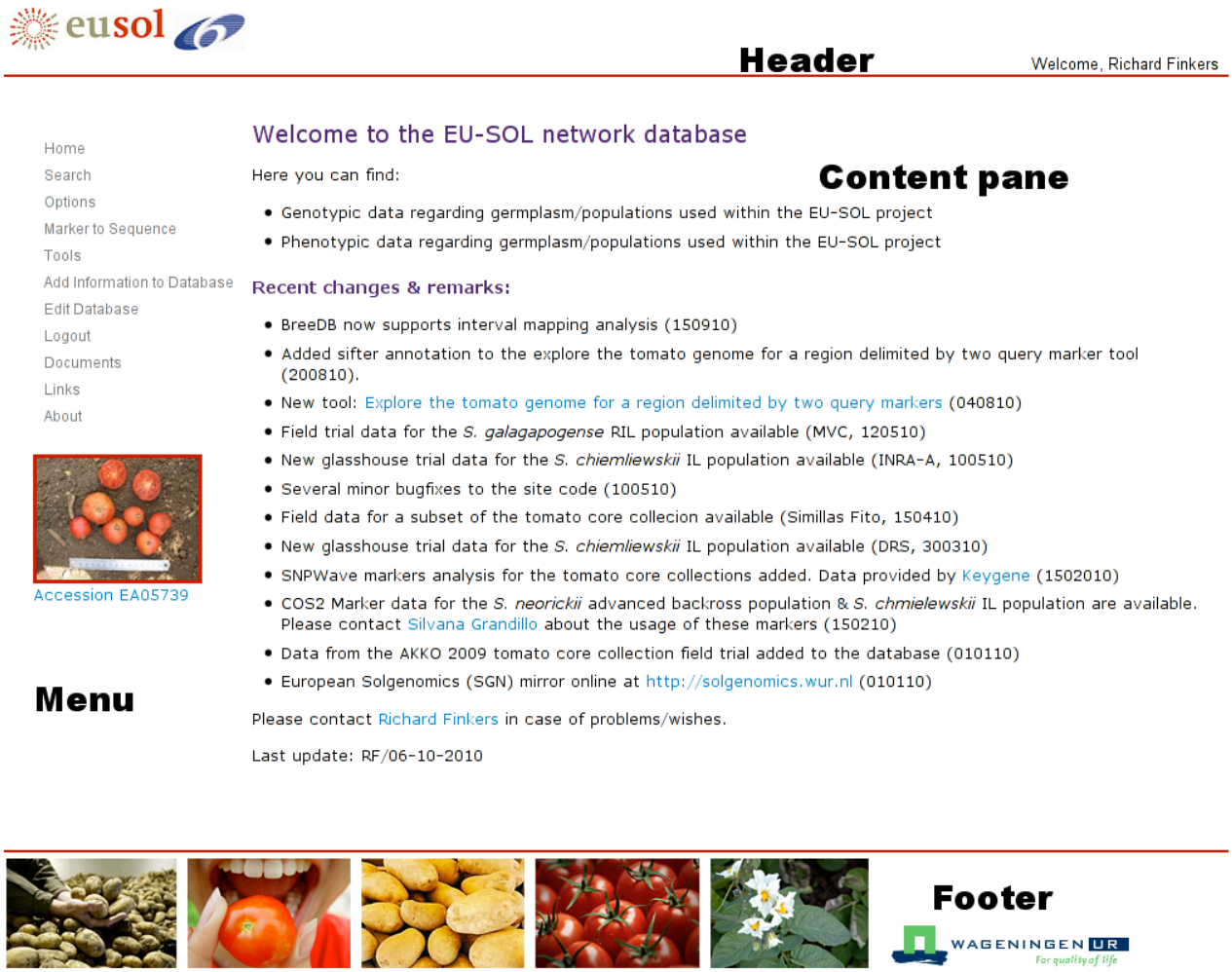


Figure 2: Outline of the BreeDB web-based user interface. The header shows the user that is logged-in into the system, the menu pane is context sensitive and differs on the basis of user authorization and user selection. The content pane shows all relevant output.

Convention: Options, selectable from the menu pane, are underlined in this document

Note: BreeDB is mainly developed using Firefox. Although we try to prevent this as much as possible, some features does not always work the way that they intended to work on other web browsers.

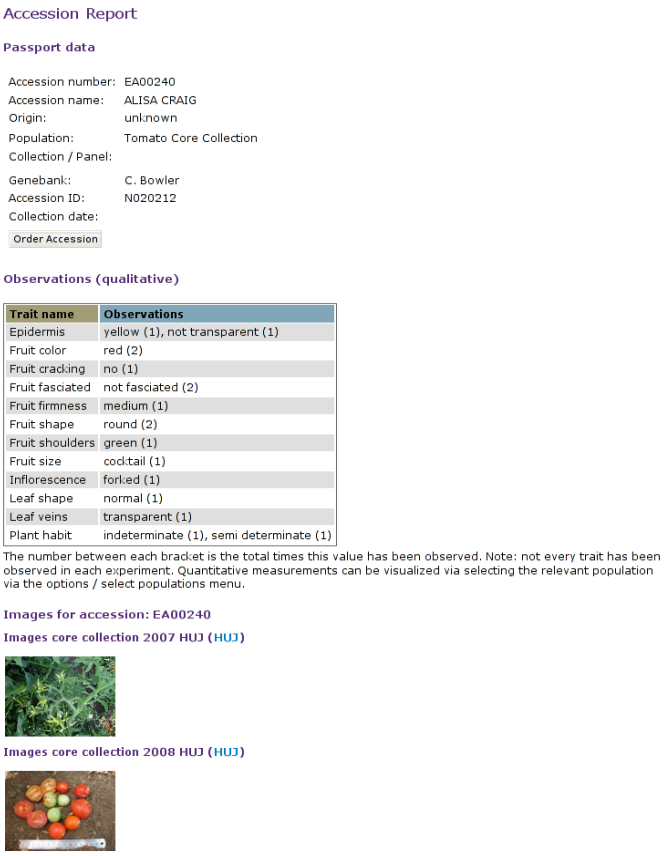
Note: Although there are more than one way to access the different data exploration and data analysis methods, this manual will only refer to selection of methods using the menu pane which is located on the left site of the screen. This menu consist out of a number of sub-menu's which can be opened by hovering over the different menu options. The content of this menu depends on the current authorization level and selected population / experiment / map. The contents of some of the results pages are also depending on the current authorization level. The manual describes only options that are available to all users.

The germplasm module

The germplasm module contains the data for a collection of genetic resources. These are typically accessions obtained from gene banks, private collections, but also include immortal experimental populations. Besides accessions, BreeDB also includes information about genotypes. If an accession is heterozygous, this leads to multiple genotypes. Transient populations are usually not defined as an accession, but only defined as a genotype.

The germplasm module provides a number of basic search functions, which are available from the menu: search. The two main options are: search by accession name and search by accession number. These methods can be used to find accessions using its original name or using a known BreeDB specific accession identifier. Alternatively, an accession can also be search using an accession number from the gene bank from which the accession originates. This option can be found under the Search; Search accessions by passport data. In addition, a collection can be searched by country of origin or original gene bank from this page as well.

All search options might return you one or a list containing multiple accessions. If more than one result has been found, a downloadable list of all matching accessions will be shown which contains the BreeDB accession number, the accession name and the country of origin. The BreeDB accession ID is hyper-linked and can be clicked to obtain detailed information about this accession (Figure 3). The detailed accession page will be shown directly if just one hit is found within BreeDB.



Accession Report

Passport data

Accession number:	EA00240
Accession name:	ALISA CRAIG
Origin:	unknown
Population:	Tomato Core Collection
Collection / Panel:	
Genebank:	C. Bowler
Accession ID:	N020212
Collection date:	

Order Accession

Observations (qualitative)

Trait name	Observations
Epidermis	yellow (1), not transparent (1)
Fruit color	red (2)
Fruit cracking	no (1)
Fruit fasciated	not fasciated (2)
Fruit firmness	medium (1)
Fruit shape	round (2)
Fruit shoulders	green (1)
Fruit size	cocktail (1)
Inflorescence	forked (1)
Leaf shape	normal (1)
Leaf veins	transparent (1)
Plant habit	indeterminate (1), semi determinate (1)

The number between each bracket is the total times this value has been observed. Note: not every trait has been observed in each experiment. Quantitative measurements can be visualized via selecting the relevant population via the options / select populations menu.

Images for accession: EA00240

Images core collection 2007 HUJ (HUJ)

Images core collection 2008 HUJ (HUJ)

Figure 3: Accession report for accession EA00240. The upper part of the report summarizes passport data for this accession. The middle table describes the different qualitative characteristics which have been observed on this

accession. The lower area shows all available photographs for this accession, in this case from two different experiments.

The detailed accession report consists of three major parts: passport information, qualitative observations and available images. The passport information contains data about the origin of the accession as well as information of the collection / gene bank that this accession is part of. Gene bank derived accessions are usually hyperlink to their original gene bank as these resources usually also publish their own observations. The qualitative observations describe all descriptive results for this accession. The frequency of an observation is given between brackets. In addition to these text based descriptors, photographs, from multiple experiments, are shown in this report as well. Together, these descriptions provide an good impression of each accession.

The last germplasm exploration option shows the origin of the accessions shown on a world map. This method can be found under search; show accessions on Google map. The color coded world-map shows an impression of how many accessions were originating from a given country. The color scale ranges from purple (many), via red to blue (a few) and gray (none). Exact numbers of accessions / country can be obtained from the downloadable table.

Exploration of qualitative characters

It has already been shown that in the accession report, qualitative characteristics are summarized. However, it is also possible to select accessions on the bases of these descriptors. For this, the Search; search by phenotype menu option can be selected. One or more phenotypes might be selected to obtain a subset of accessions according to that criteria. Depending on the number of results, an table containing a list of accessions or an report for one accession will be shown.

Selecting population, experiment & map

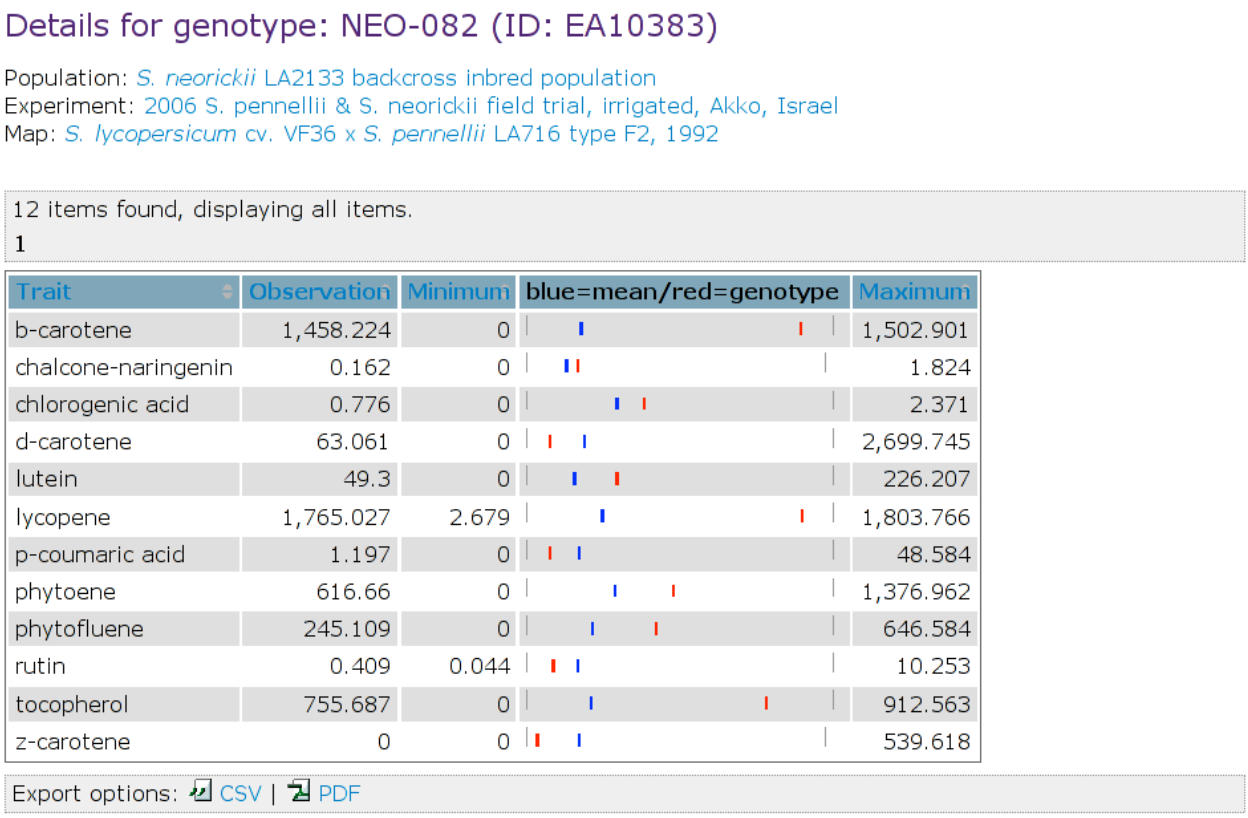
Thus far, we have been mainly focusing on exploration of passport data and qualitative data. However, BreeDB is mainly designed for analysis of quantitative data. As BreeDB contains data for many experiments, it usually does not make sense to visualize everything at once. Therefore, BreeDB provides the possibility to select specific subsets from the stored data. The selection wizard can be started by clicking the Select population and experiment sub-menu in the Options menu. The first step of the selection wizard allows you to select any of the available natural or experimental populations. Based on this selection, experiments performed using this population can be selected. Based on the availability of marker data, for the selected population, the option to select a genetic linkage map or physical map is show. All subsequent analysis will use this selection. A genetic map or experiment can be changed using the options; select another map or options, select another experiment respectively. Alternatively, the wizard can be restarted using options; select population and experiment.

On the majority of the results pages, the current selections are shown. The population, experiment and map details are usually hyper-linked. If you follow this hyperlink, you will go to an page on which detailed information for each selection is provided.

Note: Relevant genetic linkage maps or physical maps are shown if at least 60 markers are overlapping between a map and marker data set. However, we currently do not order the map names by the number of overlapping markers and we also don't show the number of overlapping markers between the two types of information. This functionality is on the to do list.

The data exploration module

The data exploration module contains a set of basic visualization tools. These tools include bar plot, box plot, hierarchical clustering, heat map, principal component bi-plots and (multiple) scatter plots. These data exploration tools will become available under the Associations menu after selection of a population, an experiment and a genetic or physical map. The first set of visualization tools show the variation per trait and are accessible from the menu Associations, Trait visualization. Methodology to explore variation between at least two traits is available under the menu Associations, Trait-Trait associations. A tool to show the variation between different experiments conducted on the same population is available from the menu: Associations, Experiment-trait associations. The bar plot, histogram and single scatter plot graphs provide additional information by hover-over of each data point. In addition, the bar plot and single scatter plot are hyper-linked to an detailed report for a selected genotype (Figure 4).



Details for genotype: NEO-082 (ID: EA10383)

Population: *S. neorickii* LA2133 backcross inbred population
Experiment: 2006 S. pennellii & S. neorickii field trial, irrigated, Akko, Israel
Map: *S. lycopersicum* cv. VF36 x *S. pennellii* LA716 type F2, 1992

12 items found, displaying all items.
1

Trait	Observation	Minimum	blue=mean/red=genotype	Maximum
b-carotene	1,458.224	0		1,502.901
chalcone-naringenin	0.162	0		1.824
chlorogenic acid	0.776	0		2.371
d-carotene	63.061	0		2,699.745
lutein	49.3	0		226.207
lycopene	1,765.027	2.679		1,803.766
p-coumaric acid	1.197	0		48.584
phytoene	616.66	0		1,376.962
phytofluene	245.109	0		646.584
rutin	0.409	0.044		10.253
tocopherol	755.687	0		912.563
z-carotene	0	0		539.618

Export options: CSV | PDF

Figure 4: detailed quantitative phenotyping report for accession EA10383 (individual NEO0082 of the *S. neorickii* backcross inbred population). Columns will be discussed from left-to-right. The trait column contains the name of the trait. The observation column shows the observed value for this specific individual, within the selected experiment. The minimum column contains the minimum observation within the whole population while the maximum column contains the maximum observation within the whole population. The bar with the red and green lines is a quantitative representation of the score of the selected genotype (red) and the population average (blue).

The graphical genotyping module

BreeDB contains a visualization module that shows **g**raphical **g**eno**t**ypes (GGT). A graphical representation of molecular marker data can be an important tool in the process of selection and evaluation of plant material. GGT enables the representation of molecular marker data by simple

chromosome drawings (Figure 5). This GGT image consists out of two parts. The left pane shows the genetic or physical map while the right area shows the marker scores for each individual. Each vertical line symbolizes the set of chromosomes for that individual, while the color coding indicates if an segment is homozygous like the mother allele (light pink), heterozygous (shocking pink) or homozygous like the father allele (blue).

The QTL analysis module

Quantitative trait loci (QTL) analysis in an important tool to study the association between a set of genotypes and a trait of interest. Within BreeDB, two types of QTL analysis methodology have been implemented, namely the Kruskall-Wallis test and interval mapping. Since the Kruskall-Wallis is a non-parametric method, it does not assume a normal population. This methodology is therefore available for all different types of populations (for example IL, AB, RIL, F_2, etc). However, interval mapping makes use of the assumption that the quantitative phenotype follows a normal distribution with equal variance in both parental strains and is therefore only available for F_2 and RIL populations. The QTL analysis menu option comes available under the Associations; Trait-Marker associations menu once a population, experiment and a **linkage map** has been selected. One or more traits can be selected for an QTL analysis, and results for each of these traits will be visualized in one overview (Figure 5). By default, we perform both Kruskall-Wallis and interval mapping analysis simultaneously on-the-fly. The results of the Kruskall-Wallis test is visualized for each marker, while the LOD2 interval is shown for each QTL detected using interval mapping methodology.

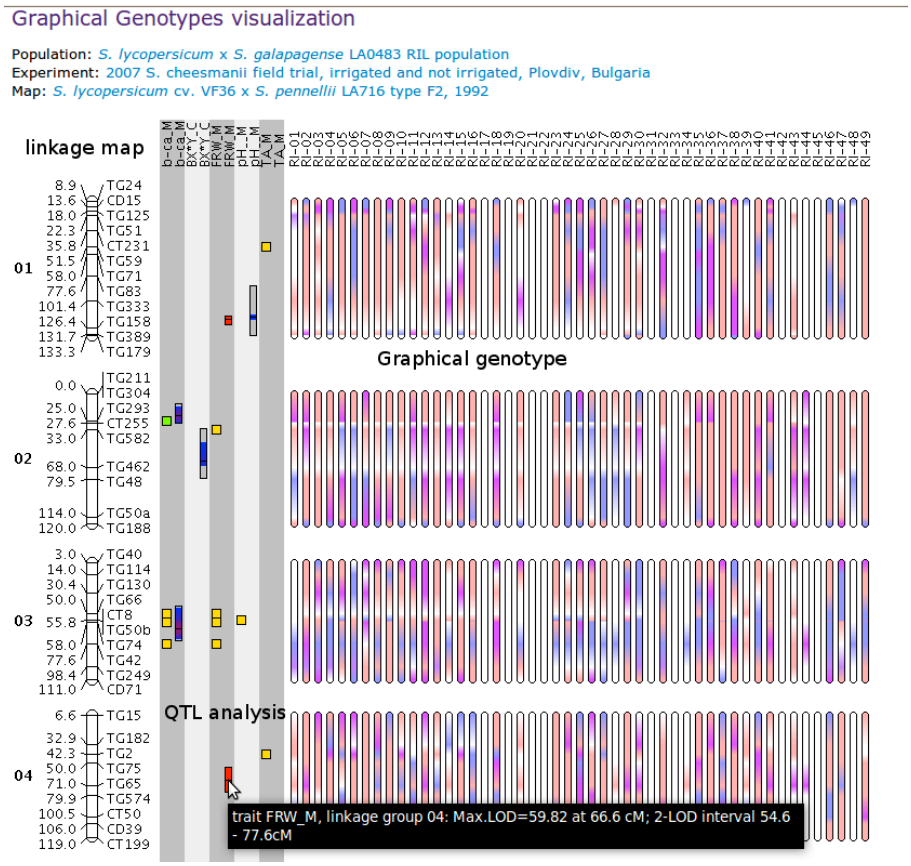


Figure 5: Graphical genotypes of Chromosome 1 to 4 from demo data of an RIL population. Each vertical bar within the

graphical genotype pane represents one individual of this population. Introgressions are shown using color codes while putative QTL regions are shown between the genetic linkage map pane and the graphical genotype pane. Chart elements are hyper-linked to drill through to individual QTLs, chromosomes, individuals or traits.

We provide several ways to further inspect results of the QTL analysis. Almost all regions, within the results visualization, are hyperlinked to report pages. In general, hovering the mouse over a region of the visualization will show more information. For example, hovering over the a genotype bar will show you which individual and linkage group is under the mouse. Hovering over the result of a QTL analysis result will provide more details about the analysis. In addition, many regions of this visualization are also hyperlinked. The genotype bar is hyperlinked to the genotype report (Figure 4). Each marker is hyperlinked to a marker detail pages (*Solanaceae* network, http://solgenomics.wur.nl). While the results of a Kruskall-Wallis test is hyperlinked to an visualization that shows the distribution for this trait / allele (Figure 6). The LOD2 interval is hyperlinked to the genetic to physical region tool and will be discussed in more detail in the next paragraph. Interval mapping is performed using the R package R/QTL (Broman et al. 2003).

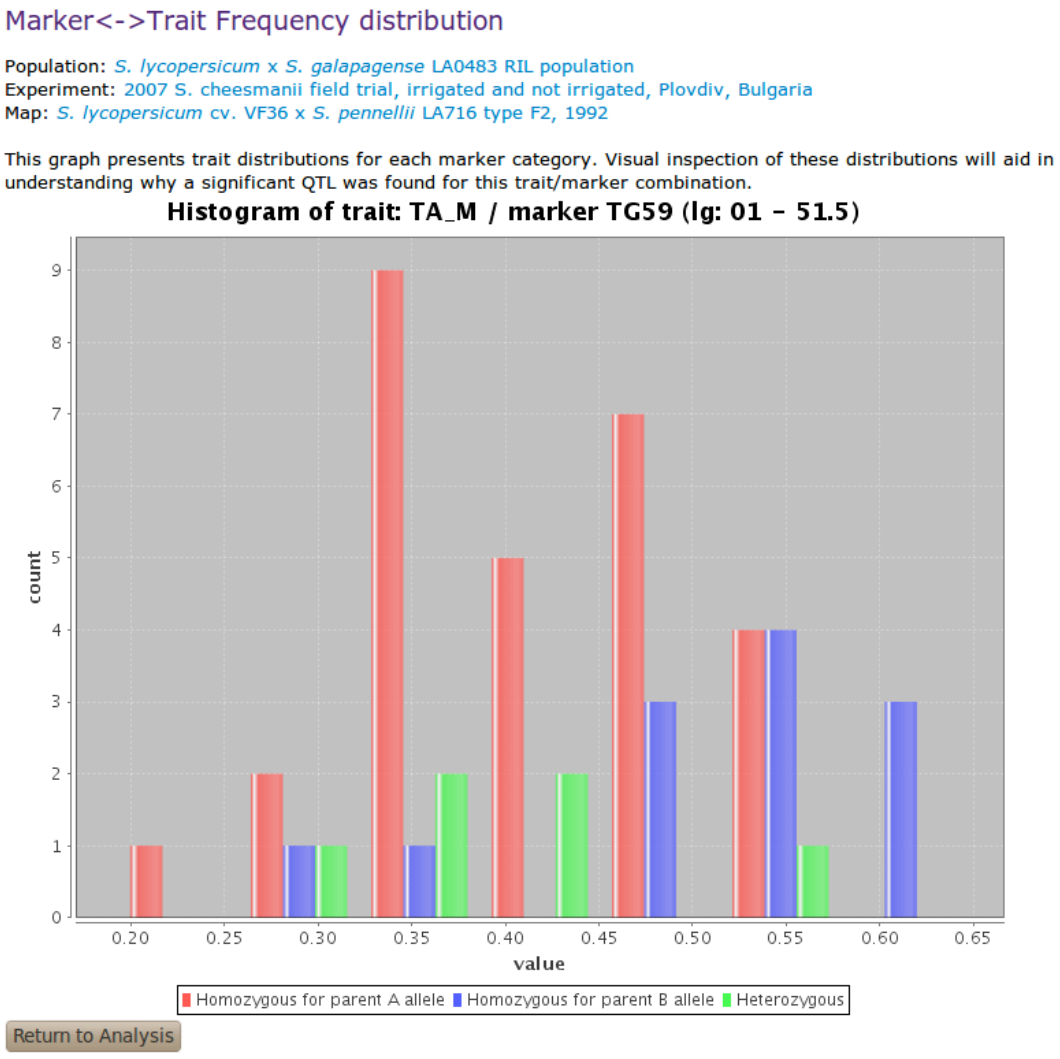


Figure 6: Frequency distribution of a marker trait combination which is significant according to the Kruskal-Wallis analysis. A separation of the distribution for the group homozygous like the mother allele and the distribution of the group h omozygous like the father allele can be seen. The heterozygous group is intermediate. This suggests that the alleles f or this QTL might work in an additive manner.

The genetic to physical region module

The International Tomato Genome Sequencing Project was started in 2004 by an international consortium including participants from Korea, China, the United Kingdom, India, the Netherlands, France, Japan, Spain, Italy and the United States. In 2009, a whole-genome shotgun approach was initiated, which in conjunction with other data yielded high quality assemblies. The ITAG consortium has annotated the different builds. Currently, this data is released under the data release agreement which can be found at: http://solgenomics.wur.nl/genomes/Solanum_lycopersicum/index.pl. The genetic to physical region module relies heavily on the data provided by this consortium. Please note that by using this tool, you agree to this data access agreement.

The aim of the genetic to physical region module is to bridge the gap between a genetic and a physical map. Traditionally, a lot of QTL studies has been taken place without the potential to easily mine the physical region for candidate genes. This module should aid is such efforts. The tool can be accessed using two different approaches. The first approach allows users to enter two markers, flanking the genetic region of interest. This approach is available under the Marker to sequence menu. The second approach requires an user to perform a QTL analysis on their population/trait of interest. The LOD2 interval, of each QTL, can be clicked directly. The genetic and physical interval for the QTL will be shown (Figure 7).

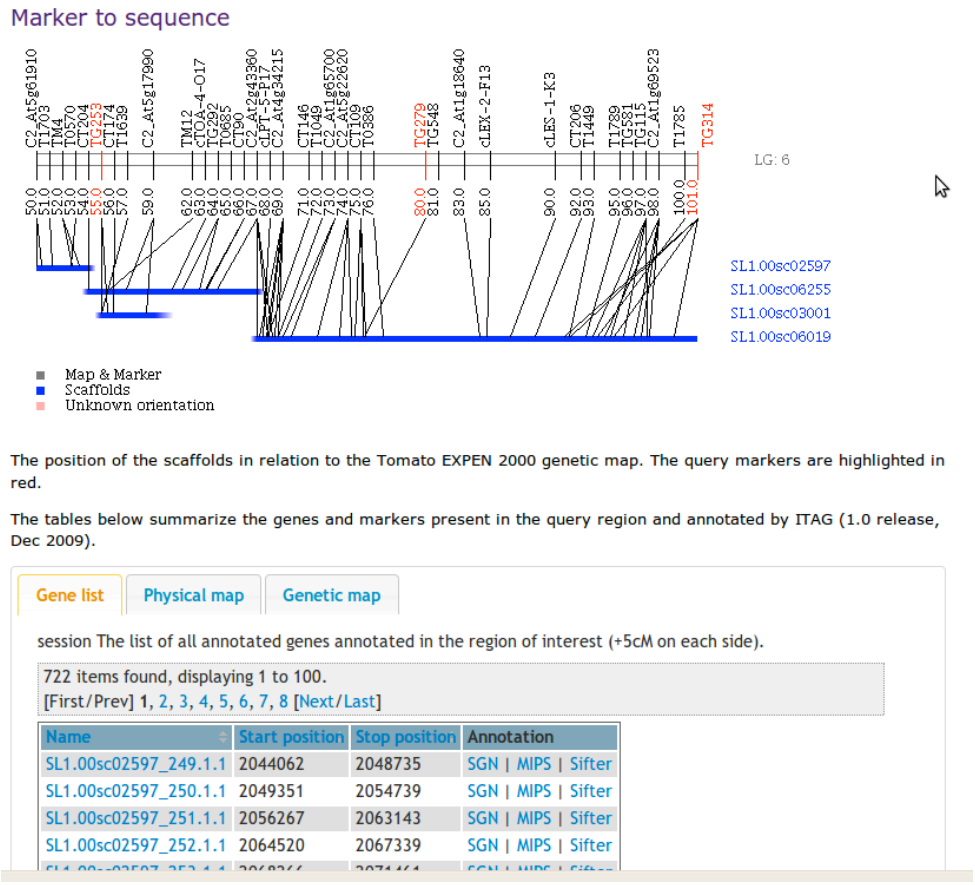


Figure 7: Output of the genetic to physical region module. The upper graph shows the genetic linkage map, aligned to the scaffolds of the *Solanum lycopersicum* cv. Heinz assembly version 1.00. Marker in red are markers that were in common with the map that was used for QTL analysis and the EXPEN 2000 reference map used for this visualization. The markers TG253 and TG134 are the markers flanking the LOD2 interval. The lower pane of the visualization

contains a table with all genes, using their ITAG identifier and containing links to functional, found on scaffolds matching the LOD2 interval In addition, genetic and physical information about markers spanning the LOD2 interval are also available in table form using the selection tabs.

The physical map is drawn on the basis of a number of assumptions. The most important assumption is that markers should be available on the tomato EXPEN 2000 genetic linkage map (*Solanaceae* genomics network, http://solgenomics.wur.nl). The genetic interval is extended 5cM on either site of the QTL interval / flanking markers to aid in drawing the alignment between genetic and physical map. Annotations, for all scaffolds within the area of interest, are obtained from the ITAG annotation database which is available from the *Solanaceae* genomics network. These annotations are available in table form and can be downloaded. Currently, we are adding technology that will allow an user to perform queries on this list so that it will be possible to zoom in for candidates of interest. Associations between lists of candidate gene and knowledge about the trait under investigation should allow the formulation of new hypothesis for gene(s) controlling the trait of interest.

Association mapping module

Tools to perform association mapping, also know as linkage disequilibrium mapping, are currently under development.

The SNP Discovery Module

The role of the SNP discovery module is to visualize SNPs in sequences that were obtained from two or more accessions. Note: this module is not used within all BreeDB instances. Basic usage of this module includes visualization of all potential SNPs between two lines for which sequence data is available. The number of SNPs between both lines and physical location of the SNP are summarized for each sequence region. The sequence, including the location of each SNP is shown at the level of the sequence. In addition, interesting sequences can be downloaded in a (multiple) fasta format.

Marker Diversity Module

Diversity in the accessions can not only be shown at the phenotype level, but also on the genotype level using molecular marker datasets. For this, we make use of the R package adegenet (Jombart T. 2008). Adegenet is a package dedicate to multivariate analysis of genetic markers and is used to obtain a dissimilarity matrix for an set of markers. This dissimilarity matrix can subsequently be used as the input for an hierarchical clustering analysis (Figure 8) or PCO analysis. As individual users might want to zoom-in on only subsets from the available set of accessions, this tool allows end-users to analyze their own sub-set on-the-fly. Analysis starts by selection either Markers; Marker diversity; Hierarchical clustering or Markers; Marker diversity; PCO visualization from the menu. In addition, frequencies for each marker allele might also be queried using the Markers; Marker statistics option an results are shown in table format (Figure 9).

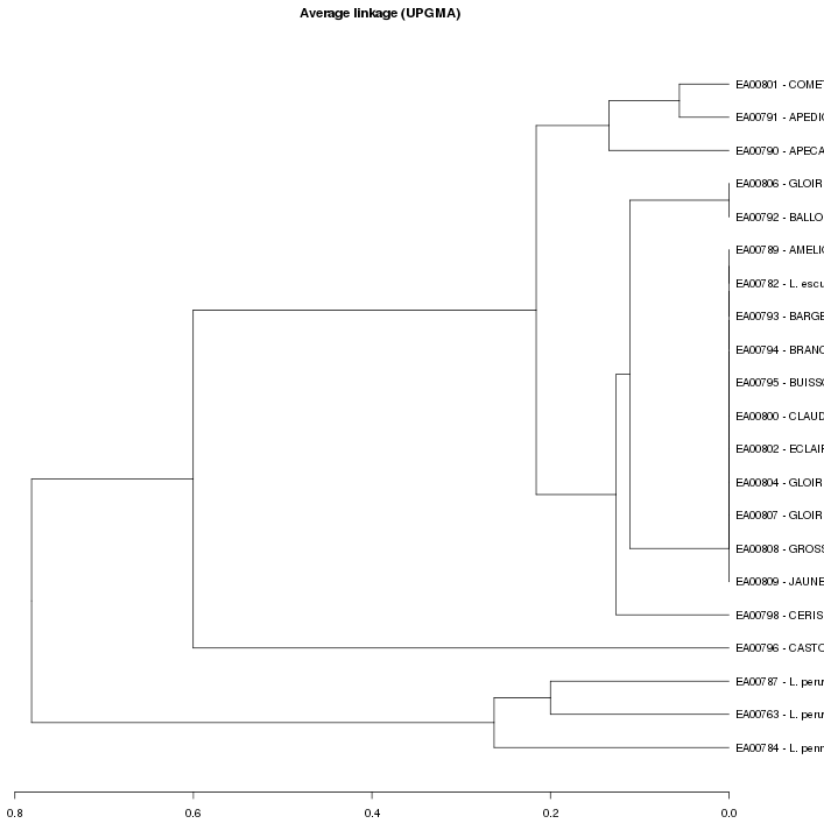


Figure 8: Results of an on-the-fly genetic diversity analysis on a subset of the tomato core collection using a set of SNPWave markers (Keygene N.V.®). A set of wild relatives of the cultivated tomato, *S. lycopersicum* and *S. lycopersicum* land-races have been selected for this analysis. The diversity is show using an hierarchical clustering method. The wild relatives are clearly distinct from the group of *S. lycopersicum land-races.*

Marker Overview

Population: Tomato Core Collection
Experiment: 2008 Core Collection Field trial - De Ruiter Seed
Map: No map for this population

17 items found, displaying all items.
1

MarkerName	Frequency Allel 1	Frequency Heterozygous	Frequency Allel 2	Frequency Unknown
KGe0149	2462	13	467	189
KGe0532	2236	10	713	172
KGe0717	2823	16	125	167
KGe0769	2937	2	18	174
KGe0787	2874	7	85	165
KGe0832	2954	1	7	169
KGe0906	2328	28	606	169
KGe0952	2737	7	217	170
KGe1393	2818	14	65	234
KGe2133	2584	20	347	180
KGe2217	2886	23	56	166
KGe2336	2846	5	116	164
KGe2563	2952	8	6	165
KGe2925	2884	2	79	166
KGe2937	2827	18	120	166
KGe2960	2880	10	70	171
KGe3140	2736	14	209	172

Export options: CSV | PDF

Figure 9: Allele frequencies for a set of SNPWave markers (Keygene N.V.®) on the EU-SOL tomato core collection of land-races and wild accessions. The occurrence of homozygous allele 1, heterozygous , homozygous allele2 or unknown is counted and displayed. This table is

downloadable in an excel compatible format.

Data upload module

Within BreeDB there is a beta version of data upload. This upload allows users to add data to the database. At this point, the upload has been programmed for adding experiments, observations, methods and traits, however these are still in a testing stage. The upload works using standardized template files, which will be available from BreeDB. The template consists of a file with one or more excel sheets. In these sheets the first row is filled with names, these should stay in the file. The data you want to upload should be added underneath this header and the information should be entered in the appropriate column. The file can then be uploaded to the database, here can be set if the dataset is publicly available and a few extra settings, which are easily understandable.

Tutorial

The aim of this tutorial is to give you an easy, yet comprehensive overview of the major features of BreeDB. The tutorial cannot show you everything, so some engagement and a will to experiment are required. Also, the introduction or the overview presentation might provide you a reference on how to use the functionality of BreeDB. During the tutorial, we usually refer to usage of the menu which is in the left pane of the screen, although, some of the options are repeated elsewhere as well.

Exercise 1 – Explore the germplasm collection

The main role of the germplasm module is to store passport data for each accession. In this exercise, we go through some of the basic search functionality of BreeDB. For this, we will use the BreeDB instance developed within the EU-SOL project. The URL is: https://www.eu-sol.wur.nl

Show all accessions on a Google map.

1) From which country are the majority of accessions ? How many are from Brazil?

Alternatively, accessions can be search using their passport data.

2) Which accessions originate from Norway?

The collection can also be searched for your own favorite cultivar

3) How many accessions with the name “Black cherry” are part of the collection?
4) What can you say about the origin of these accessions?

Exercise 2 – Explore trait variation in experimental or natural populations

A set of simple data exploration tools are available in BreeDB. These tools have the aim to make the stored information easily minable and are available for both natural populations as for populations derived from a bi-parental cross.

We will explore the data using the following settings (Options, Select populations and experiment):

- Population: *S. lycopersicum* x *S. galapagense* LA0483 RIL population
- Experiment: 2007 *S. cheesmanii* field trial, irrigated and not irrigated, Plovdiv, Bulgaria
- Map: *S. lycopersicum* cv. VF36 x *S. pennellii* LA716 type F2, 1992

The first task will be to generate a bar plot for the trait ß-carotene.

1) Which line has the highest ß-carotene content, and how much?

Now, click on the bar for the genotype with the highest ß-carotene content.

2) For which other traits does this genotype have a higher value? Does this make sense and how can this interpretation easily be made?

Perform an single scatter (XY scatter) analysis for ß-carotene content and Total pigments

3) What do you observe?

Note: each point of the scatter plot shows additional information when hovered-over. Each point can also be clicked to obtain more information about this specific individual.

Exercise 3 – Graphical genotyping & QTL visualization

In this exercise, we are going to perform a QTL study for the ß-carotene content in Tomato fruit. We will still use the same population and experiment as selected in exercise 2. Before starting the QTL analysis, try to answer the following questions:

1) The growing season of this experiment was not normal. What was extraordinary for this field trial?

The studied population is a RIL population. Use GGT (show map) to visualize the marker scores.

2) Does the graphical genotyping image follow your expectation?

Perform an QTL analysis for the trait ß-carotene.

3) How many QTLs do you detect?
4) Do the results from the Kruskal-Wallis analysis correspond to the results of the Interval mapping?

Show the Marker ↔ trait frequency distribution for the most significant marker, according to the Kruskal-Wallis analysis on Chromosome 6.

5) Which is the most significant marker, according to the Kruskal-Wallis analysis, on Chromosome 6?
6) Do you consider this QTL to be of interest?

Exercise 4 – From QTL to physical map (Tomato).

The purpose of this exercise is to show how to jump from an genetic interval of interest to the corresponding region on the genome. Our use case still includes the *S. galapagense* population, the Bulgarian 2007 field experiment and the trait ß-carotene content in Tomato fruit. In case that the QTL of interest is not on your screen anymore, perform a QTL analysis once more.

In the previous exercise, we have been mainly exploring the results of the Kruskal-Wallis analysis. Now, we will focus on the results from the interval mapping approach, specifically on the QTL interval for ß-carotene content on chromosome 6.

1) How significant is the QTL for ß-carotene content on chromosome 6 according to the interval mapping analysis. How large is the LOD2 support interval?

This QTL interval can be further explored by clicking on the LOD2 bar.

2) How many genes map to the selected genetic interval?

BreeDB offers functionality to assist in mining this list of potential candidate genes. For example, we can search within this list with the keyword carotene.

3) How many genes have the keyword carotene associated. Can you tell why these genes were included in this short-list?

You can search the list with alternate forms of the keyword. Search the list with the keword ”beta-carotene”.

4) Explain the result.

Acknowledgments

This project was co-financed by the 6th framework EU project “High Quality *Solanaceous* crops for consumers, processors and producers by exploration of Biodiversity”. Contract number: FOOD-CT-2006-016214, Wageningen UR Plant Breeding and the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative / Netherlands Organization for Scientific Research”. The functionality of the genetic to physical map tool relies on database integration with the *Solanacea* genomics network (SGN) and was possible because of the SGN mirror we host in Wageningen (http://solgenomics.wur.nl).

Contact

BreeDB is a system that is still actively under development. Updated information about BreeDB can be found at: http://www.plantbreeding.wur.nl/UK/software_breedb.html or by contacting Dr. Richard Finkers directly (richard.finkers@wur.nl). Although great care is taken to ensure optimal data exploration possibilities, we might have overlooked some details. We greatly appreciate if you can contact us to improve our contents and site quality.

References

Broman KW, Wu H, Sen Ś, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889-890

Jombart T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403-1405. doi: 10.1093/bioinformatics/btn129