Breeding View

A visual tool for running analytical pipelines

User Guide

Darren Murray, Roger Payne & Zhengzheng Zhang

VSN International Ltd

January 2015

1. Introduction

The Breeding View is a visual tool for running analytical pipelines. It provides a user-friendly interface to access breeding data analysis tools, such as field trial analysis and QTL mapping. The Breeding View contains a visual pipeline representation of the steps involved for an analysis ranging from data quality checks to producing final reports. Each visual pipeline includes a set of *nodes* that allow interaction to control the flow of the analysis. The results from an analysis are summarized within a combination of html reports, data export files (e.g. Microsoft Excel files) and png image files. The Breeding View runs each analytical pipeline using an underlying statistical analysis engine. The main application used for the statistical analysis is GenStat. However, the Breeding View also includes a facility to use R.

The Breeding View can be run as part of the Breeding Management System (BMS) within the Integrated Breeding Platform (https://www.integratedbreeding.net/) or as a standalone application. This guide provides an introduction to using the Breeding View as a standalone application but the concepts also apply to when the Breeding View is launched from the BMS. In the guide examples are used to illustrate the different analysis pipelines that are available:

- Single trait field trial analysis - single trait single environment field trial analysis using mixed models. This pipeline can also be used to run a sequence of analyses for different environments.
- GxE anaylsis – genotype-by-environment (GxE) analysis.
- Single trait QTL mapping (single environment) – single trait single environment linkage analysis for inbred populations. This pipeline can also be used to run a sequence of analyses for different environments.

2. Getting Started

2.1 A Tour of the Breeding View Interface

In this section we will give a tour of the Breeding View interface. The figure below shows the interface after starting the Breeding View.

The interface includes a menu bar and a toolbar to provide a quick shortcut to some frequently used functions. A status bar is located at the bottom of the interface and displays information about pipelines when there are run. The panel on the left-hand side contains a tab called **Project** which displays information about a project and names of data structures available to run within a pipeline. The main part of the interface contains two tabs. The **Analysis Pipeline** tab displays a visual interactive layout of the analytical pipeline, and the **Output** tab displays output from the most recent analysis.

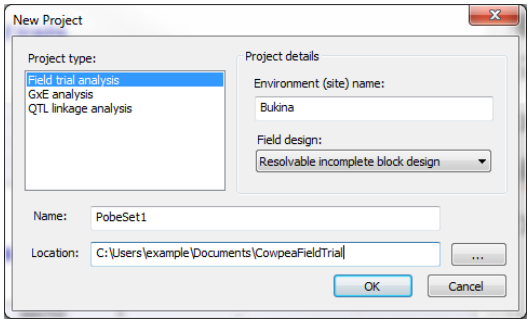
After an analysis has been run additional tabs will appear. If an analysis produces any graphs, then these will appear within a **Graph** tab. Each analysis produces an html report of results, which will appear within a **Report** tab.

2.2 Breeding View Project

To run an analysis for a particular pipeline, a Breeding View project needs to be created. A Breeding View project contains the information, settings and data required to run an analysis.

2.2.1 Creating a Breeding View Project

A new project can be created by selecting **File | New Project** from the menu bar, or by clicking on the **New Project** tool button. This will open a dialog, where there are three different types of project that can be created.



Each project requires a name and a location to store the working files.

2.2.2 Adding data to a Breeding View Project

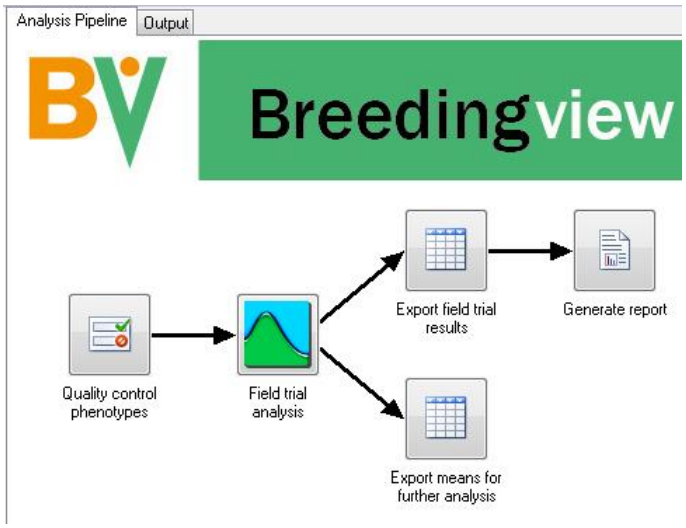
Before an analysis can be run, the data must be imported into the Breeding View. Data can be loaded by selecting **Project | Add Data** from the menu, or by clicking on the **Add to Project** tool button. This will open a window that allows you to search for, and select the file to open. Clicking **Open** on this window will open a dialog which can be used to select the appropriate data structures for the analysis.

2.2.3 Saving/opening a Breeding View project

A Breeding View project and the associated data can be saved, so that an analysis pipeline can be run on another occasion. A project is saved within two files: an xml file, containing details of the project attributes, and an associated data file. To save a project, select **File | Save Project** from the menu, and enter a filename. To open a project, select **File | Open Project** from the menu, or click on the **Open Project** tool button. Select the xml project file, and click **Open**. The Breeding View will automatically look for the associated data file and, if found, it will import the associated data stored within that file.

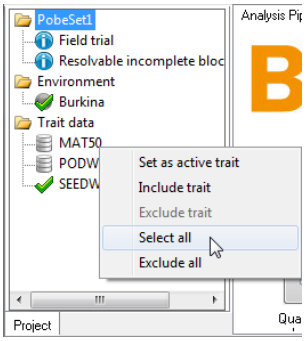
2.3 Analysis Pipelines

When a project has been created or opened, a visual representation of the analytical pipeline will be displayed on the **Analysis Pipeline** tab. The analysis pipeline includes a set of connected *nodes* which can be used to run and configure pipelines. The figure below shows the analysis pipeline for a single trait field trial analysis. Here there are 5 nodes representing the different components of the analytical pipeline.



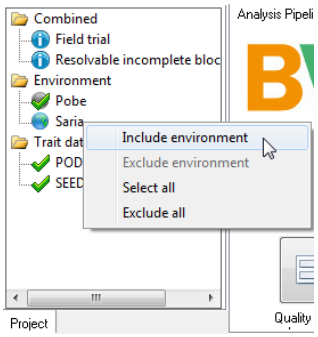
2.3.1 Selecting data to analyse

To run an analysis the data structures to analyse need to be selected. The data are displayed in the **Project** tab on the left-hand side of the interface. The traits available to run within a pipeline are listed within the **Trait data** folder. When a trait has been selected to be analysed it will have a green tick displayed next to its name in the list. You can select one or more traits to be run in an analysis.



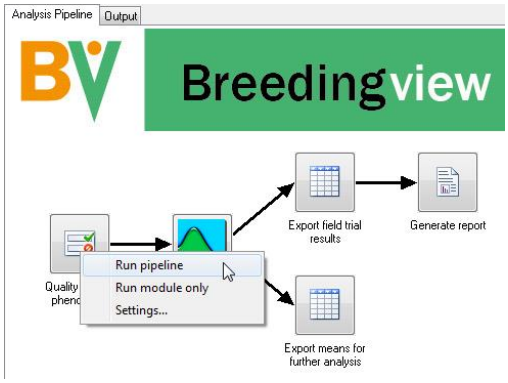
To change the selection of traits to be analysed, right-click on any of the traits and choose the action from the shortcut menu.

If the project contains data from more than one environment then a list of the environments will be displayed within the **Environment** folder in the **Project** tab. Similar to the list of traits, when an environment has been selected to run, it will have a green tick displayed next to its name. You can choose a selection of environments to be analysed by right-clicking on any of the environment names, and choosing the appropriate action from the shortcut menu.

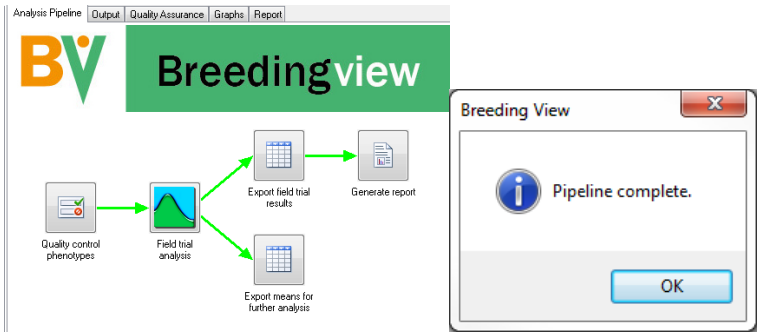


2.3.2 Running an analysis

To run the analysis pipeline, click on the right-mouse button on the first node, and select **Run pipeline** from the shortcut menu. This will run the whole analysis pipeline, by performing the task at each node in turn.



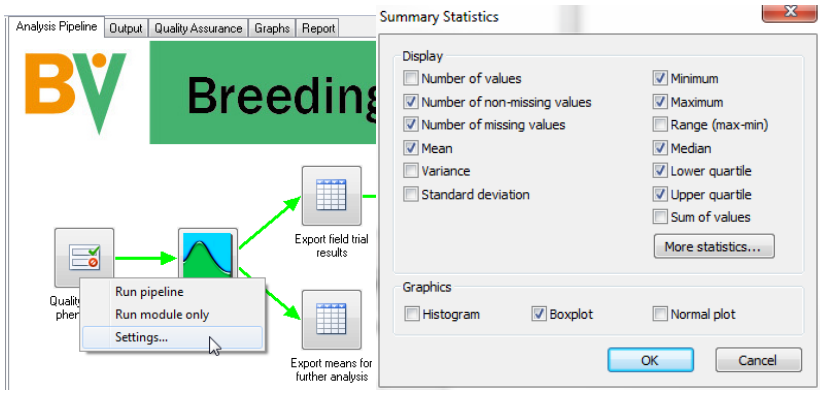
As the task at each node is completed, the connecting arrow will change colour to indicate that the pipeline is moving onto the next task. When the pipeline completes, a prompt will appear to indicate it has finished.



Individual traits can be run at a node at a time, by selecting **Run module only** from the shortcut menu.

2.3.3 Changing settings

Some of the nodes have options to control they way an analysis is performed and output that is displayed. To access the options, you can right-click on a node and select the **Settings** item from the shortcut menu.



The changes to the options are retained during the current session and are saved to the project file.

2.4 Output and Results

When an analysis is run, the output and results are displayed within the **Output**, **Graph** and **Report** tabs. In addition to this, the results are written to a set of files that are stored within a time stamped folder within the working folder. The files that are stored within this folder include the report in html format, data export files in csv or Microsoft Excel format and graphs in png format.

3. Single Trait Field Trial Analysis Pipeline

3.1 Introduction

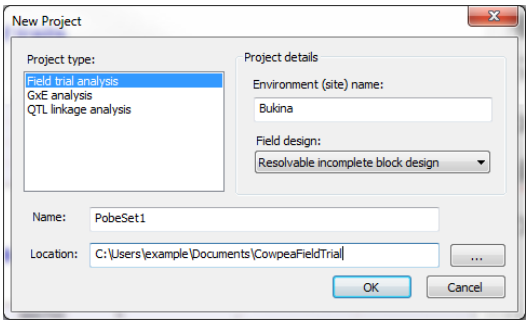
The objective here is to illustrate the analysis of a single trial, taking into account its design. We introduce the use of mixed models to account for extraneous source of variation in the trial, including replicates and incomplete blocks. The ultimate aim of this exercise would be to produce adjusted means per genotypes, that can be used within a GxE or QTL analysis pipeline.

3.2 Data

The data for this example are from a field trial of cowpea progeny conducted within a resolvable incomplete-block (alpha) design. The file *BF2011MARS_PobeSet1.csv* is a comma separated-values file, containing three traits along with the genotype ids and the design information. The traits are days to 50% maturity (MAT50), total weight of unthreshed pods (PODWT) and seed weight per plot (SEEDWT). The column GENOTYPES contains the genotype ids, and the columns REP and BLOCK contain the design information for the resolvable incomplete-block design.

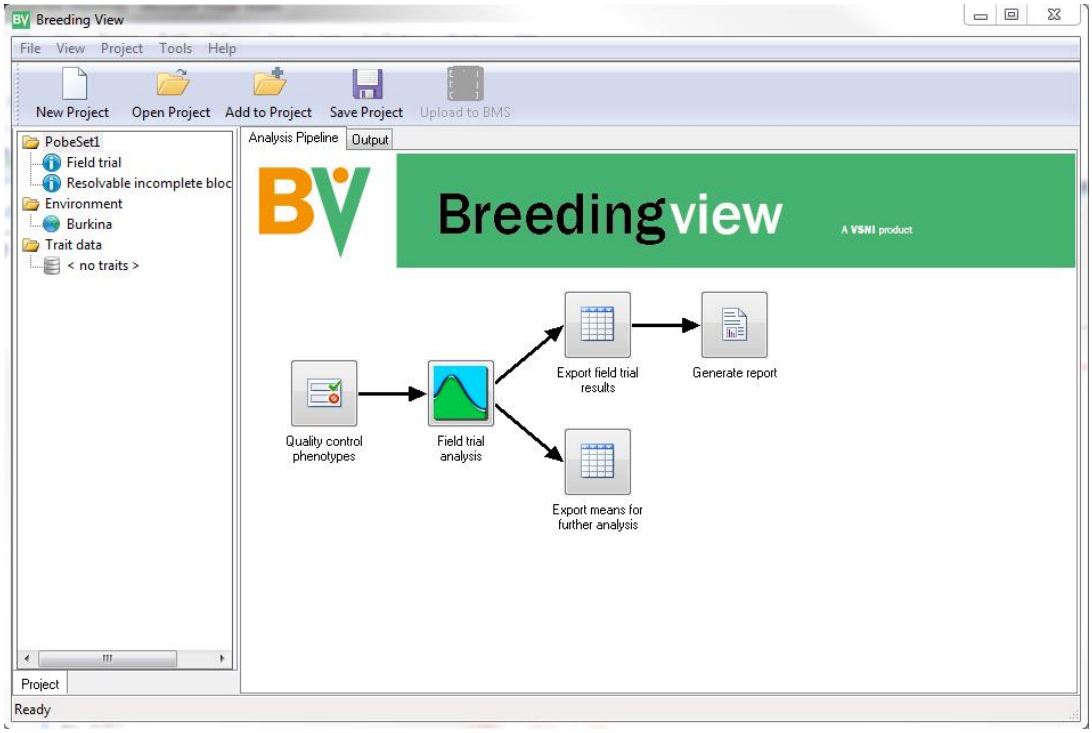
3.3 Breeding View Project

To analyse the field trial data, a Breeding View project needs to be set up. To create a new project, select **File | New Project** from the menu bar, or click on the **New Project** tool button.



In the **New Project** dialog, select the **Field trial analysis** item within the **Project type** list. Enter *Burkina* as the **Environment (site) name,** and select the **Resolvable incomplete block design** item from the **Field design** list. Enter *PobeSet1* as the **Name** of the project, and then browse for a folder to specify the **Location** where the working files will be stored. In the dialog above, we have specified a folder called *CowpeaFieldTrial* within *My Documents (C:\Users\example\Documents)* for the Location. Click on **OK** to create the project.

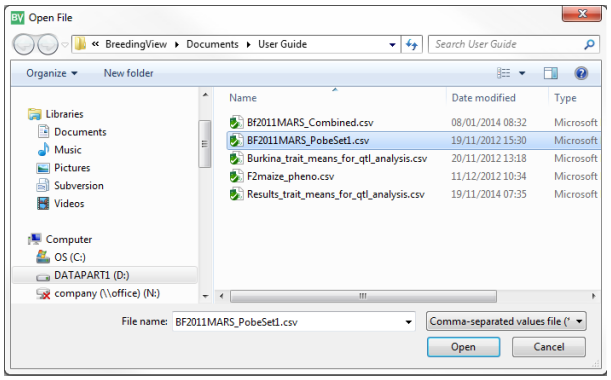
When the project has been created, the details are displayed in the **Project** tab. The *PobeSet1* folder is the project name, and contains attributes about the project. The *Environment* folder contains the environment name for the field trial. The **Trait data** folder lists all the available traits for analysis. On the right-hand side, the **Analysis Pipeline** tab contains a visual representation of the analysis pipeline for the field trial data.

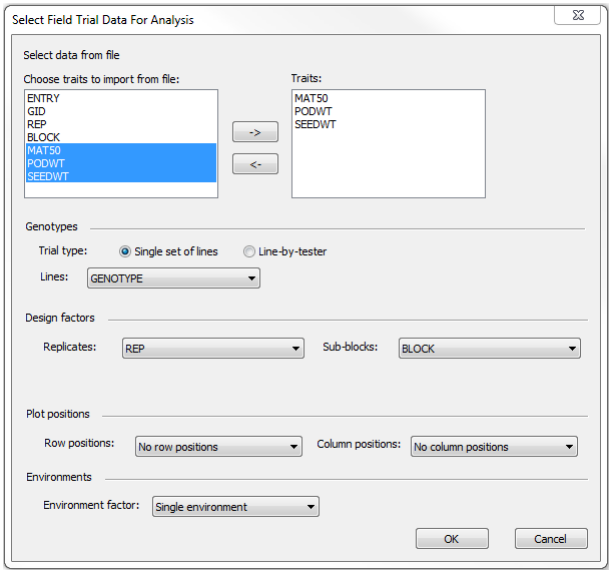


The nodes in the **Field trial analysis** pipeline are as follows:

Nodes	Description
Quality control phenotypes	Summary statistics for the trait(s)
Field trial analysis	Mixed model analysis of field trial for the trait(s)
Export field trial results	Stores results in external files
Generate report	HTML report of results including means and summaries for the trait(s)
Export means for further analysis	Stores means in an external file using a format ready to use in a GxE or QTL analysis pipeline

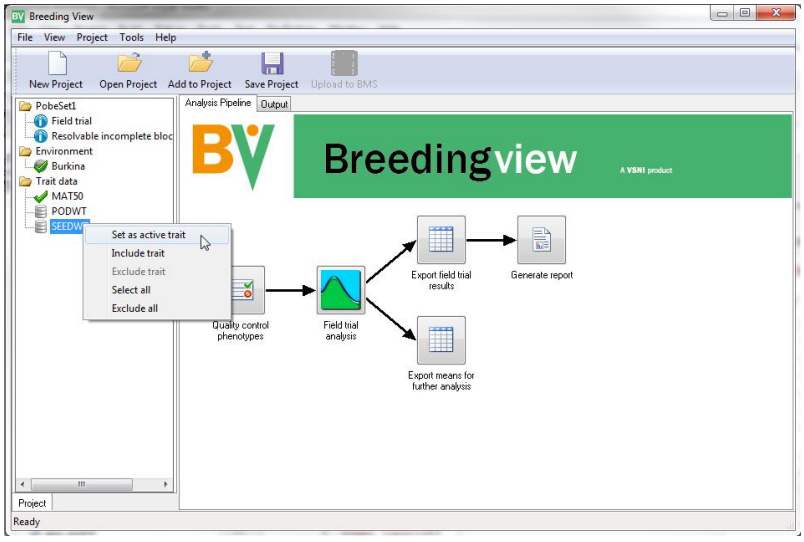
Before an analysis can be run, the field trial data must be added to the project. To do this, select **Project | Add Data** from the menu, or click on the **Add to Project** tool button. Locate the file *BF2011MARS_PobeSet1.csv*, and click **Open**. This will open a dialog called **Field trial data,** which is used to select the traits, genotypes and design information for the analysis.





When a file is opened, it is scanned to determine whether it contains default names for the genotype and design terms for the analysis. If the data within a file uses default names (for a resolvable incomplete block design these are Genotypes, Reps and Blocks), then they are automatically selected within the dialog. The file *BF2011MARS_PobeSet1.csv* contains the column names GENOTYPE, REP and BLOCK, which are recognised as standard names. So these are automatically selected in the dialog. To choose the traits, select on the names *MAT50*, *PODWT* and *SEEDWT* in the **Columns in file** list. Click on the **->** button to transfer the names to the **Traits** list. Click **OK** to import the data.

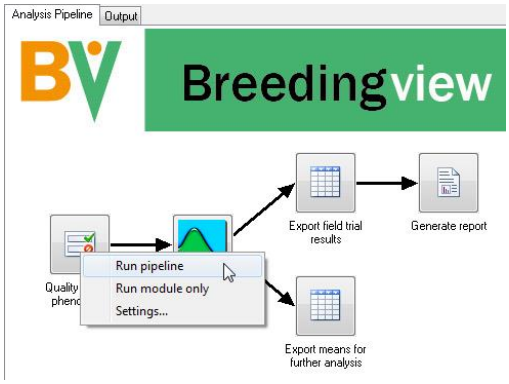
After the data have been imported, the traits will now appear within the **Trait data** folder in the **Project** tab on the left-hand side. By default the trait (*MAT50*), that has a green tick next to its name, has been selected to be used within the analysis pipeline. To change this to the *SEEDWT* trait, click on the name *SEEDWT* with the right mouse button, and select the **Set as active trait** item from the shortcut menu.



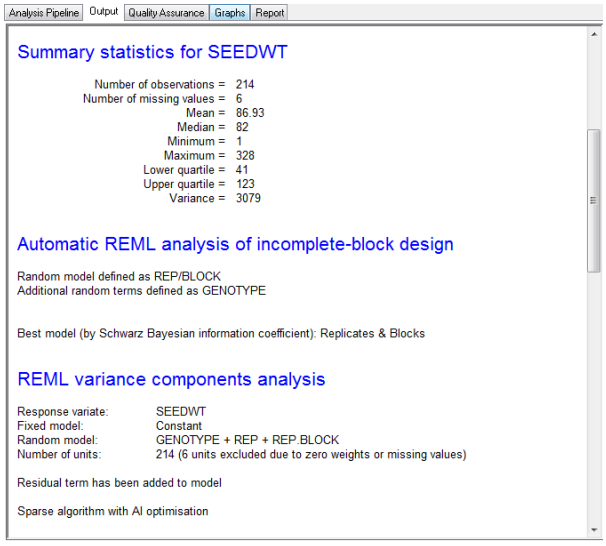
3.4 Running the Field Trial Analysis pipeline

3.4.1 Single trait and environment

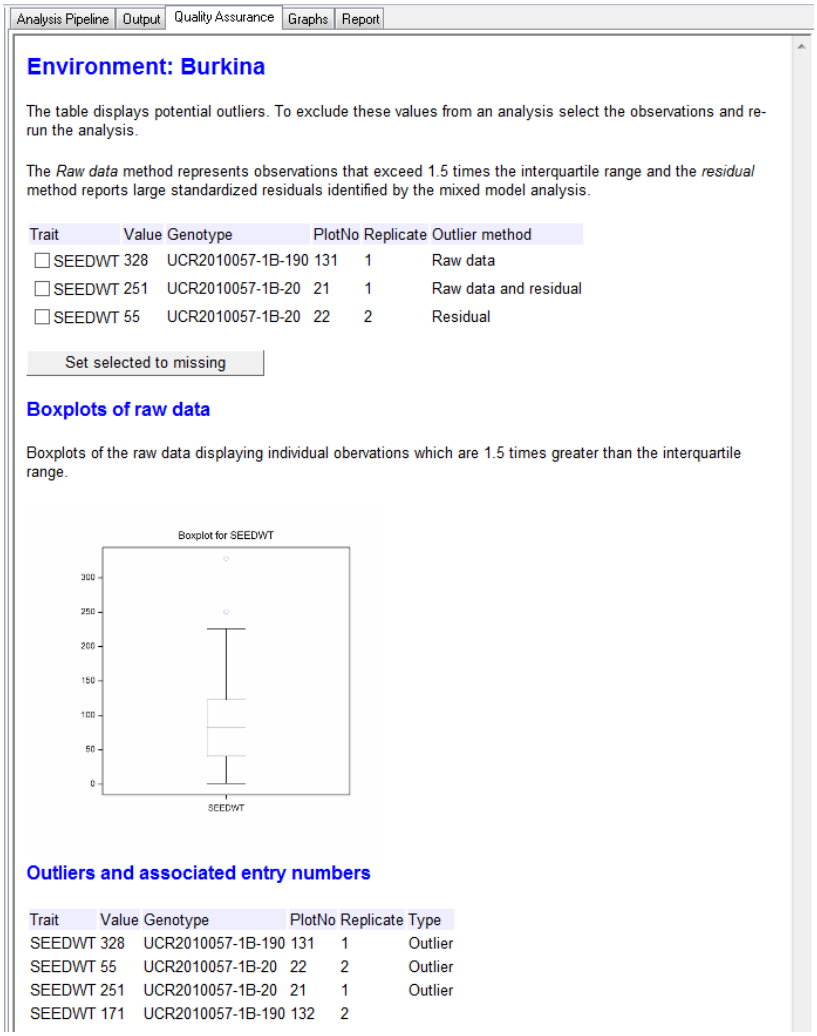
To run the analysis pipeline, click on the right-mouse button on the first node (*Quality control phenotypes*), and select **Run pipeline** from the shortcut menu. This will run the whole analysis pipeline, by performing the task at each node in turn.



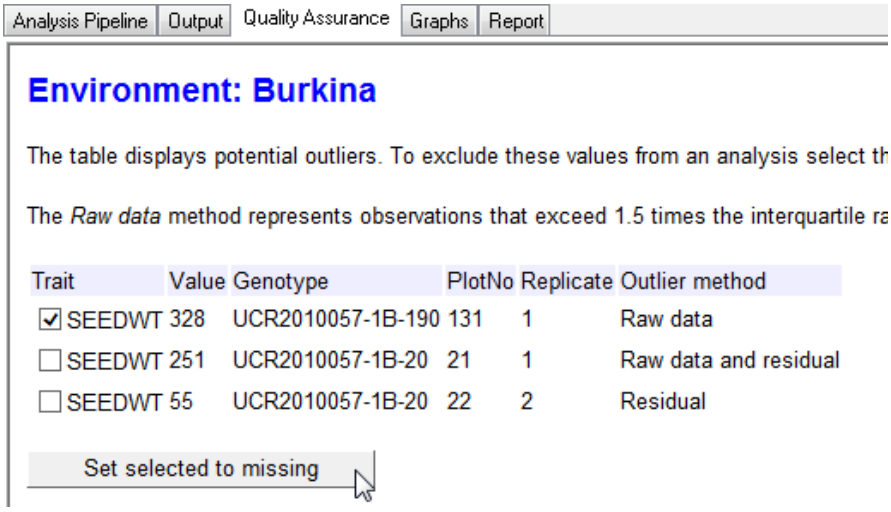
When the analysis pipeline has completed, all the intermediate output from the analysis at each node appears in the **Output** tab. For the *SEEDWT* trait, this shows the summary statistics and the mixed model analysis (see section 3.5) of the resolvable incomplete block design.



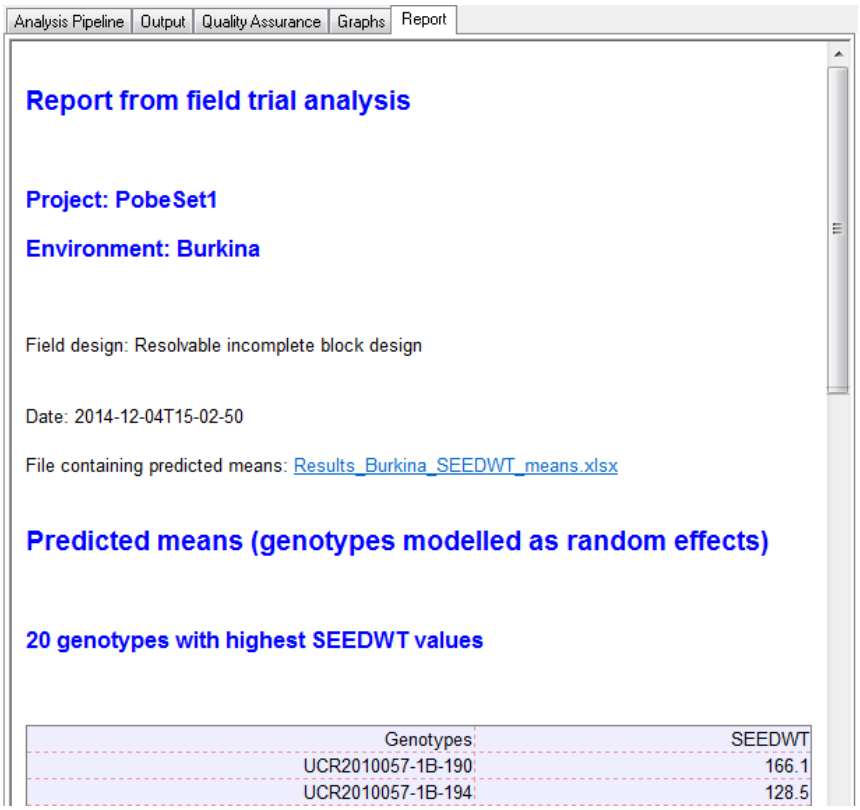
A **Quality Assurance** tab will appear, which displays a report of potentially outling values. This provides details about two types of outliers. The *raw data* outliers are observations that exceed 1.5 times the interquartile range, and can be seen on the accompanying boxplot. The *residual* outliers are observations that have been reported as a large standardized residual by the mixed model analysis. Diagnostic plots for the mixed model analysis can be viewed in the **Graph** or **Report** tabs. At the bottom of the report, a table displays the entries/genotypes that have outlier observations along with their associated values in other plots within the field design.



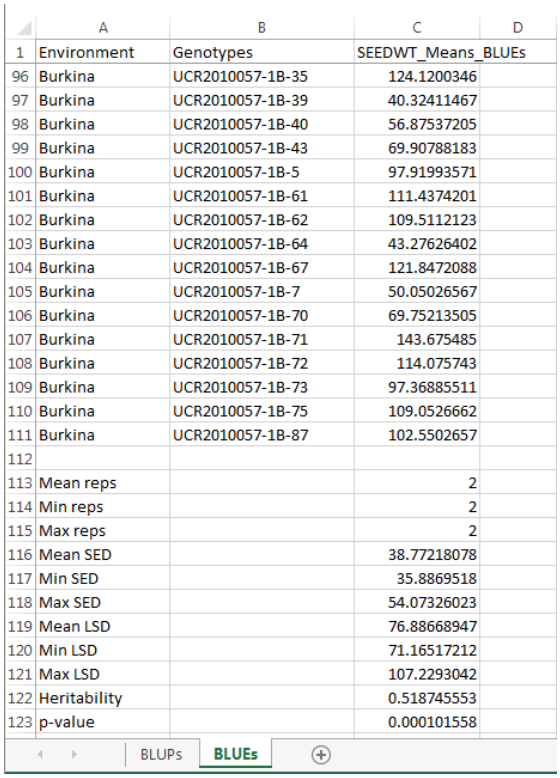
The QA report can be used to change observations to become missing for the analysis. To do this, select the observations to set as missing from the list of traits, and click the **Set selected as missing** button. The next time the analysis is run those observations will be excluded from the analysis.



The final report of the means and standard errors for the genotypes appears in a table within the **Report** tab.



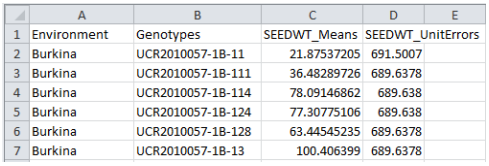
The report includes a link to the file Results_Burkina_SEEDWT_means.xlsx, which is a Microsoft Excel file that contains the predicted means from the mixed model analysis. The file can be opened by clicking on the link if your browser supports it. Otherwise the file can be found in time stamped folder within your working folder. The file contains two tabs: BLUPS (best linear unbiased predictors) and BLUEs (the best linear unbiased estimates). The sheets of BLUEs includes a summary of the results.



	A	B	C	D
1	Environment	Genotypes	SEEDWT_Means_BLUEs	
96	Burkina	UCR2010057-1B-35	124.1200346	
97	Burkina	UCR2010057-1B-39	40.32411467	
98	Burkina	UCR2010057-1B-40	56.87537205	
99	Burkina	UCR2010057-1B-43	69.90788183	
100	Burkina	UCR2010057-1B-5	97.91993571	
101	Burkina	UCR2010057-1B-61	111.4374201	
102	Burkina	UCR2010057-1B-62	109.5112123	
103	Burkina	UCR2010057-1B-64	43.27626402	
104	Burkina	UCR2010057-1B-67	121.8472088	
105	Burkina	UCR2010057-1B-7	50.05026567	
106	Burkina	UCR2010057-1B-70	69.75213505	
107	Burkina	UCR2010057-1B-71	143.675485	
108	Burkina	UCR2010057-1B-72	114.075743	
109	Burkina	UCR2010057-1B-73	97.36885511	
110	Burkina	UCR2010057-1B-75	109.0526662	
111	Burkina	UCR2010057-1B-87	102.5502657	
112				
113	Mean reps		2	
114	Min reps		2	
115	Max reps		2	
116	Mean SED		38.77218078	
117	Min SED		35.8869518	
118	Max SED		54.07326023	
119	Mean LSD		76.88668947	
120	Min LSD		71.16517212	
121	Max LSD		107.2293042	
122	Heritability		0.518745553	
123	p-value		0.000101558	

BLUPs BLUEs

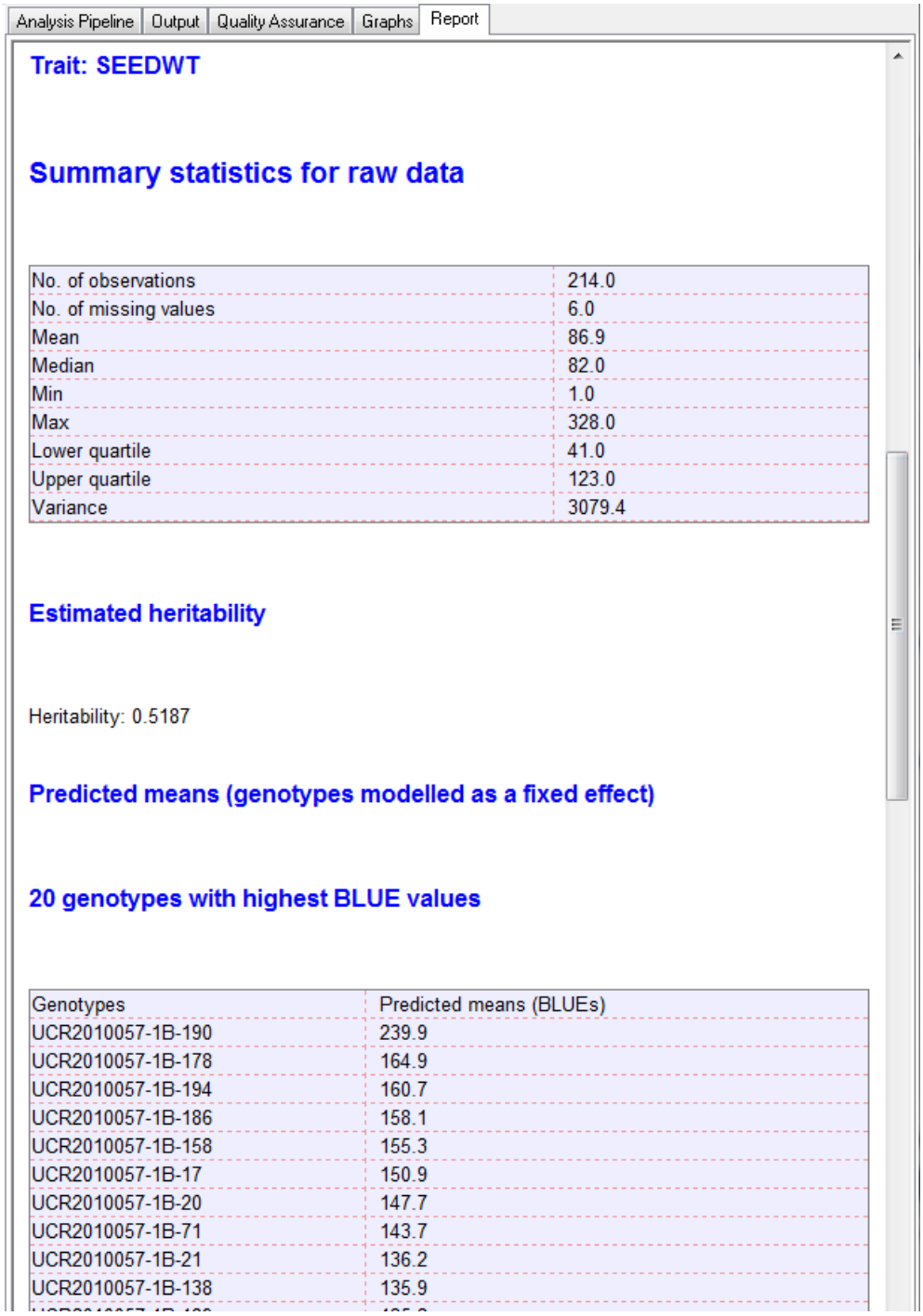
The adjusted means, required for further analysis in the QTL analysis pipeline, are saved within a comma separated values (csv) file in a time stamped folder in your working folder. The name for the file is automatically generated using the current time and date the environment and trait name. So, for this example, it will be called *Burkina_SEEDWT_means_for_qtl_analysis.csv*. To view the contents of the file, open it within Notepad or Microsoft Excel.



	A	B	C	D	E
1	Environment	Genotypes	SEEDWT_Means	SEEDWT_UnitErrors	
2	Burkina	UCR2010057-1B-11	21.87537205	691.5007	
3	Burkina	UCR2010057-1B-111	36.48289726	689.6378	
4	Burkina	UCR2010057-1B-114	78.09146862	689.638	
5	Burkina	UCR2010057-1B-124	77.30775106	689.638	
6	Burkina	UCR2010057-1B-128	63.44545235	689.6378	
7	Burkina	UCR2010057-1B-13	100.406399	689.6378	

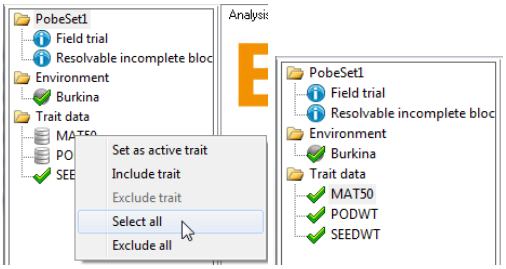
The file *Burkina_SEEDWT_means_for_qtl_analysis.csv* contains four columns: the environment name (Environment), genotypes (Genotypes), the means (SEEDWT_Means) and associated unit errors (SEEDWT_UnitErrors).

Other output displayed in the report includes summary statistics of the raw data, the heritability (see Section 3.5) for the trial, the means for the best genotypes and diagnostic plots from the mixed model analysis.

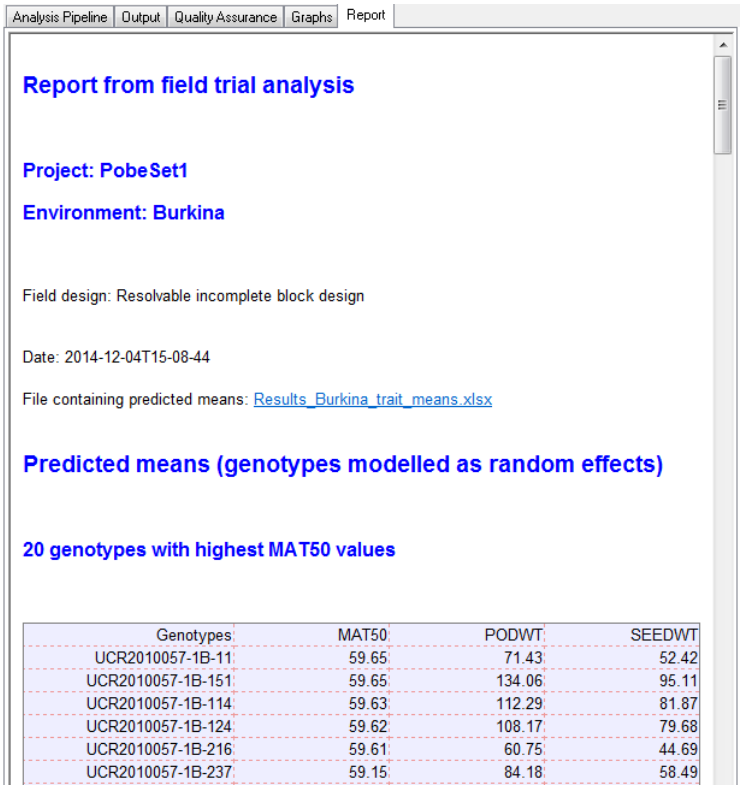


3.4.1 Multiple traits in a single environment

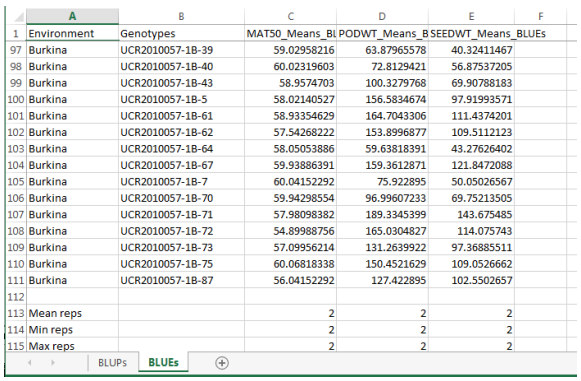
Multiple traits can be run simultaneously, with the results combined into a single report. To run the analysis pipeline on multiple traits, click on the mouse button on any of the traits, and choose the **Select all** item on the shortcut menu. Each trait will now have a green tick indicating that it will be included in the run. To run the analysis pipeline, click on the right-mouse button on the first node (*Quality control phenotypes*), and select **Run pipeline** from the shortcut menu. Note that, when multiple traits have been selected, the whole pipeline must be run.



Each trait is analysed in turn, and the results are collated into a single report.



The report contains a summary for all the traits analysed in the run, and then produces a more detailed set of results for each trait in turn. An Excel (.xlsx) file is generated, containing the combined predicted means and associated summary statistics. The file can be accessed by clicking on the link in the report (if supported by your browser).



	A	B	C	D	E	F
1	Environment	Genotypes	MAT50_Means_BL	PODWT_Means_B	SEEDWT_Means_BLUEs	
97	Burkina	UCR2010057-1B-39	59.02958216	63.87965578	40.32411467	
98	Burkina	UCR2010057-1B-40	60.02319603	72.8129421	56.87537205	
99	Burkina	UCR2010057-1B-43	58.9574703	100.3279768	69.90788183	
100	Burkina	UCR2010057-1B-5	58.02140527	156.5834674	97.91993571	
101	Burkina	UCR2010057-1B-61	58.93354629	164.7043306	111.4374201	
102	Burkina	UCR2010057-1B-62	57.54268222	153.8996877	109.5112123	
103	Burkina	UCR2010057-1B-64	58.05053886	59.63818391	43.27626402	
104	Burkina	UCR2010057-1B-67	59.93886391	159.3612871	121.8472088	
105	Burkina	UCR2010057-1B-7	60.04152292	75.922895	50.05026567	
106	Burkina	UCR2010057-1B-70	59.94298554	96.99607233	69.75213505	
107	Burkina	UCR2010057-1B-71	57.98098382	189.3345399	143.675485	
108	Burkina	UCR2010057-1B-72	54.89988756	165.0304827	114.075743	
109	Burkina	UCR2010057-1B-73	57.09956214	131.2639922	97.36885511	
110	Burkina	UCR2010057-1B-75	60.06818338	150.4521629	109.0526662	
111	Burkina	UCR2010057-1B-87	56.04152292	127.422895	102.5502657	
112						
113	Mean reps		2	2	2	
114	Min reps		2	2	2	
115	Max reps		2	2	2	

BLUPs BLUEs

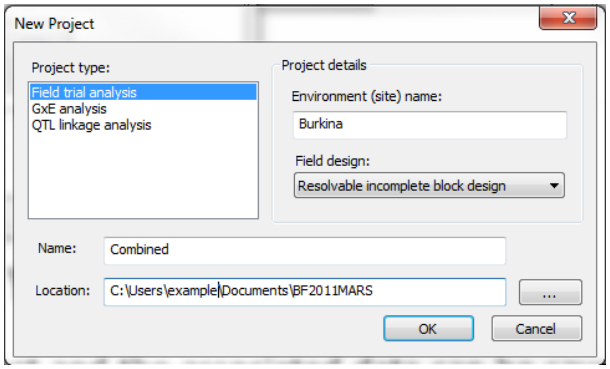
In addition, the adjusted means for each trait are saved within a single csv file called *Burkina_trait_means_for_qtl_analysis.csv*, which is in a format that can be used for further analysis in other pipelines.

3.4.1 Multiple traits in multiple environments

Sometimes field trials with the same design properties may have been conducted at different environments. Rather than creating a separate project to run the analysis for each environment/site, it may be desirable to run these all in sequence. For example, a field trial was conducted at two sites for the cowpea progeny data: Pobe and Saria. In each field trial an alpha design was used, with the same genotypes grown at each site and the same traits measured. Rather than creating two projects to analyse these data, both sites can be combined into a single project and run in sequence.

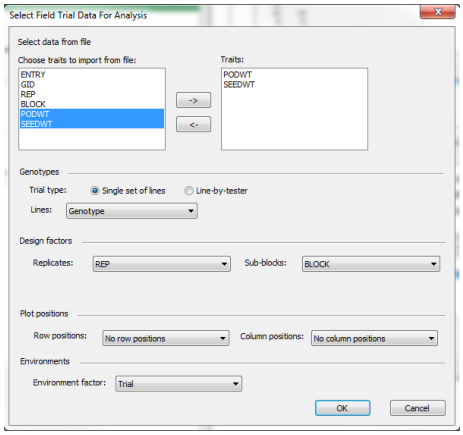
To analyse a series of trials, the data need to be supplied within a stacked format. The data should include a column to reference the different environments/sites, and the remaining columns should contain the genotypes, design factors and traits stacked by environment. The example below shows the file layout where a column called *Trial* has been used to identify the different environments.

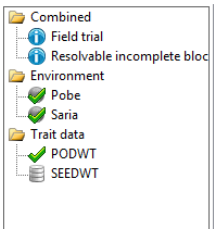
	A	B	C	D	E	F
1	Trial	Genotype	REP	BLOCK	PODWT	SEEDWT
2	Pobe	UCR2010057-1B-2	1	3	71	41
3	Pobe	UCR2010057-1B-2	2	1	8	2
4	Pobe	UCR2010057-1B-3	1	3	181	131
5	Pobe	UCR2010057-1B-3	2	5	17	8
6	Pobe	UCR2010057-1B-5	1	3	176	135
7	Pobe	UCR2010057-1B-5	2	2	147	*
8	Pobe	UCR2010057-1B-7	1	4	185	128
9	Pobe	UCR2010057-1B-7	2	3	29	19
10	Pobe	UCR2010057-1B-11	1	6	47	33



To create a new project for a sequential analysis of field trials, select **File | New Project** from the menu bar, or click on the **New Project** tool button. In the **New Project** dialog (see below), select the **Field trial analysis** item within the **Project type** list. Enter *Burkina* as the **Environment (site) name,** and select the **Resolvable incomplete block design** item from the **Field design** list. Enter *Combined* as the **Name** of the project, and then browse for a folder to specify the **Location** where the working files will be stored. In the dialog below a folder has been specified called *BF2011MARS* within *My Documents (C:\Users\example\Documents)* for the Location. Click on **OK** to create the project.

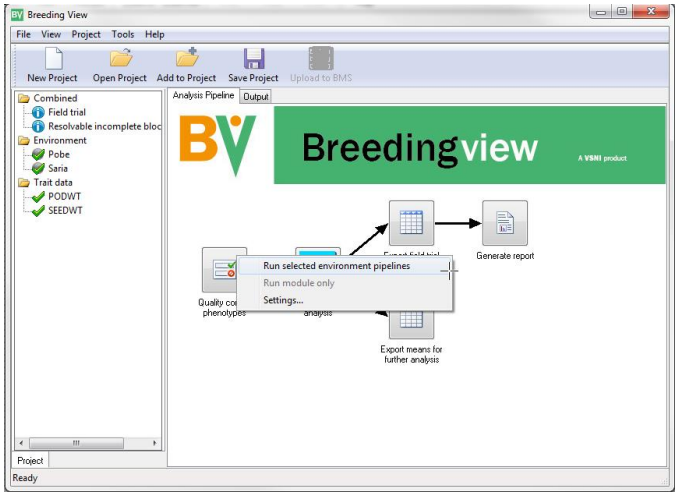
To import the data, select **Project | Add Data** from the menu, or click on the **Add to Project** tool button. Locate the file *BF2011MARS_Combined.csv*, and click **Open**. In the **Field trial data** dialog, select the names *PODWT* and *SEEDWT* in the **Columns in file** list, and click on the **->** button to transfer the names to the **Traits** list. Next select *Genotype* from the drop list for the **Genotypes factor**, *REP* for the **Replicates**, *BLOCK* for the **Sub-blocks** and *Trial* from the drop list for the **Environments factor** Click **OK** to import the data.





After the data have been imported, the traits *PODWT* and *SEEWT* will appear within the **Trait data** folder in the **Project** tab on the left-hand side. In addition, the environments/sites *Pobe* and *Saria* appear within the **Environment** folder in the **Project** tab.

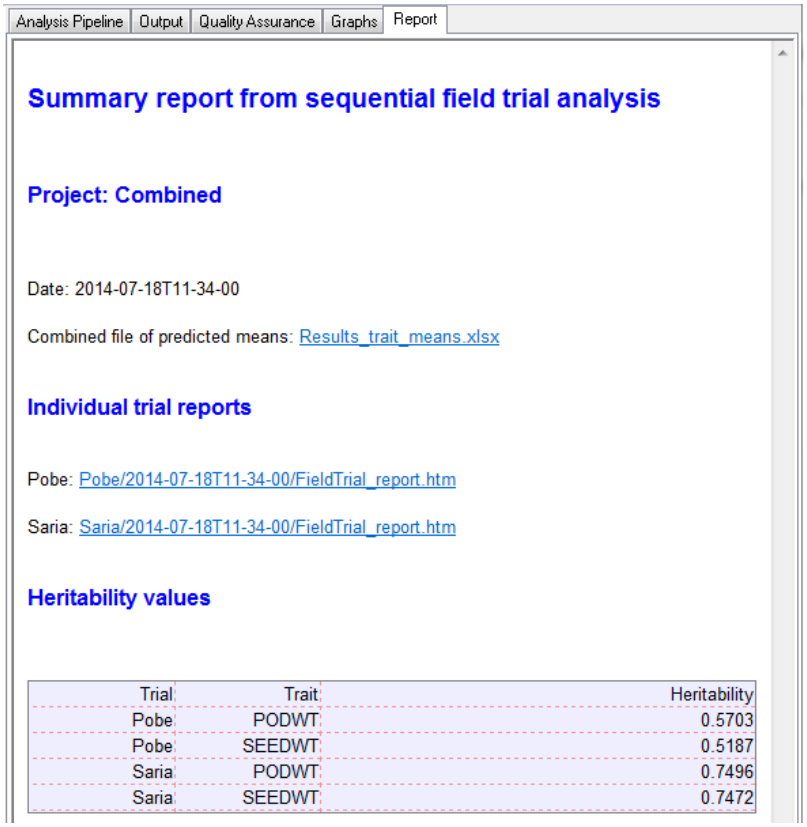
To run the sequential analysis for all traits and environments, first click on the right-mouse button on any trait and choose **Select all** from the shortcut menu. Next click on the right-mouse on the first node (*Quality control phenotypes*), and select **Run selected environment pipelines** from the shortcut menu.



When a sequence of trials is analysed, each environment is analysed in turn, and the results are stored within a folder using the name of the environment. After all environments have been run, a folder called *Combined* is created, which stores the combined results for the trials. On completion of the analysis, summary reports for all the environments are produced.

The **Quality Assurance** tab provides links to the individual QA reports within each environment. The individual reports provide details of potential influential values for each trait within the selected environment.

The **Report** tab provides links to each individual environment report and a summary of the heritability values for each trait within each environment. The individual reports contain the combined results for all traits analysed within an environment.



An Excel (.xlsx) file is generated containing the predicted means, stacked by environment. The file can be accessed by clicking on the link in the report (if supported by your browser).

3.5 Methods

3.5.1 Mixed model analysis

The Field trial analysis node does two mixed model analyses: in the first (Step 1) the Genotypes are fitted as a random term, while in the second (Step 2) the Genotypes are fitted as a fixed term. The Step 1 model is used to obtain estimates of variance parameters. By default, these variance parameters are then used in the Step 2 model, when the Genotypes are fitted as a fixed term. The rationale for this process is that we would prefer to fit Genotypes as a random term, as this avoids selection bias and results in better estimates of the variance parameters, particularly for unreplicated designs with check plots. However, the shrinkage associated with predictions of random effects is undesirable when predictions are to be carried forward to a second stage analysis such as a QTL analysis. For this reason, we use a compromise: we set Genotypes as a fixed term in order to obtain unbiased estimates but use variance parameters estimated from the model with Genotypes fitted as random.

3.5.2 Heritability

Heritability is commonly used to quantify the degree of genetic determination of the trait of interest, and in the simplest case can be interpreted as the proportion of observed variation that can be attributed to genetic differences. Heritability (denoted h^2) ranges from 0 (no genetic determination of the trait) to 1 (total genetic determination with no measurement error). Heritability depends on both the trait measured and the precision of the trial. In addition, there are many different definitions of heritability, depending on whether it relates to total genetic variation (broad-sense), or additive genetic effects (narrow-sense), or to genotype predictions (mean line), or individual observations. Because heritability relates to genetic variation, it can only be obtained when Genotypes are fitted as random (e.g. from the first analysis performed by the Field trial analysis node). For simple variance component models, heritability can be obtained directly from the estimated variance components. For more complex models a more general definition is require. We therefore use the generalized heritability measure described by Cullis, Smith & Coombes (2006). This quantity can be interpreted as a broad-sense mean line heritability, derived from an estimate of the correlation between the genotype BLUPs and their unknown true value.

4. Genotype-by-environment analysis

The objective of this section is to illustrate genotype-by-environment (GxE) analysis.

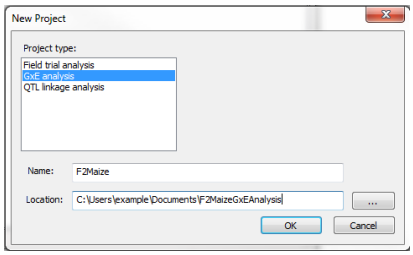
4.1 Data

The data for this example come from a maize drought stress breeding programme of CIMMYT. The population is an F_2 generated by crossing a drought tolerant parent (P1) with a drought susceptible one (P2). Seeds harvested from each of 211 F_2 lines were used to test F_3 families in 8 different environments; well watered, intermediate and severe water stress trials in 1992 (WW92a, IS92a, and SS92a respectively), intermediate and severe water stress trials in 1994 (IS94a, SS94a), and low and high nitrogen in 1996 (LN96a, LN96b, and HN96a). The suffix 'a' indicates a winter trial, and 'b' a summer trial. The measured traits were: yield in kg/plot (yld), anthesis silking interval in days (asi), number of ears per plant (eno), days to male flowering (mflw) and plant height in m (ph). Trait means for each genotype from each trial are held in file *F2Maize_pheno.csv.*

When importing data, the trait means should be stacked to provide a separate column for each trait together with a column identifying the environments and another identifying the genotypes. An example of the layout for a file can be seen in *F2Maize_pheno.csv.*

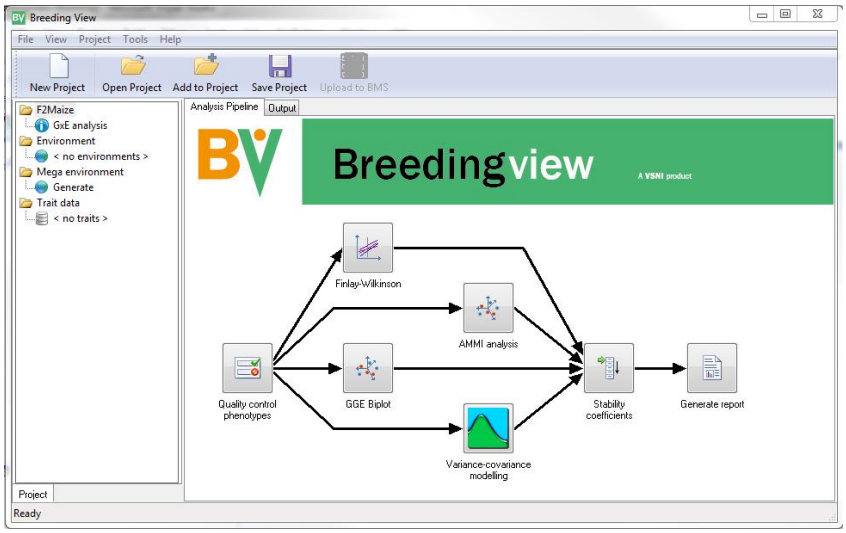
4.2 Breeding View Project

To run the GxE analysis, a new Breeding View project needs to be set up. To create a new project, select **File | New Project** from the menu bar, or click on the **New Project** tool button.



In the **New Project** dialog, select the **GxE analysis** item within the **Project type** list. Enter *F2Maize* as the **Name** of the project, and then browse for a folder to specify the **Location** where the working files will be stored. In the dialog above, we have specified a folder called *F2MaizeGxEAnalysis* within *My Documents (C:\Users\example\Documents)* for the Location. Click on **OK** to create the project.

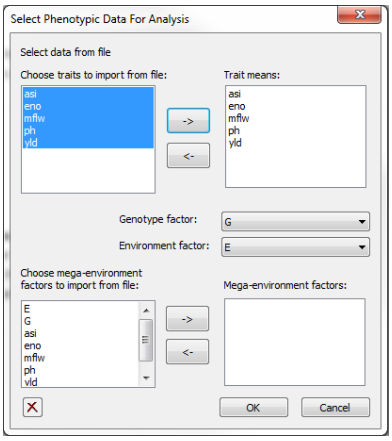
When the project has been created, the Project tab will contain 4 folders. The folder **F2Maize** is the project name, and contains attributes about the project. The **Trait data** and **Environment** folders will display the available traits and environments when the data are loaded. The **Mega environment** folder contains a single item called *Generate*, which can be used to form mega environments.



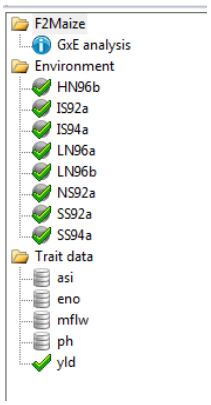
The nodes in the **GxE analysis** pipeline are as follows:

Nodes	Description
Quality control phenotypes	Summary statistics within and between environments for the trait(s)
Finlay-Wilkinson	Performs a Finlay-Wilkinson joint regression
AMMI analysis	To perform an AMMI analysis and produce biplots
GGE biplot	Produces a GGE biplot
Variance-covariance modelling	Models and selects the best covariance structure for genetic correlations between environments
Stability coefficients	Estimates different stability coefficients to assess genotype performance
Generate report	HTML report of the results

The phenotypic data need to be imported into the project before the analysis pipeline can be run. To import the data, select **Project | Add Data** from the menu, or click on the **Add to Project** tool button. In the **Open File** dialog, navigate to the folder containing the file *F2Maize_pheno.csv*, and click **Open**. This opens the **Open Phenotypic Means (Traits)** dialog, where the data to be imported can be selected. Select the names *asi, eno, mflw, ph and yld* in the **Columns in file** list, and click on the **->** button to transfer the names to the **Trait means** list. Next select *G* from the drop list for the **Genotype factor** and *E* from the drop list for the **Environment factor**. Click **OK** to import the data.



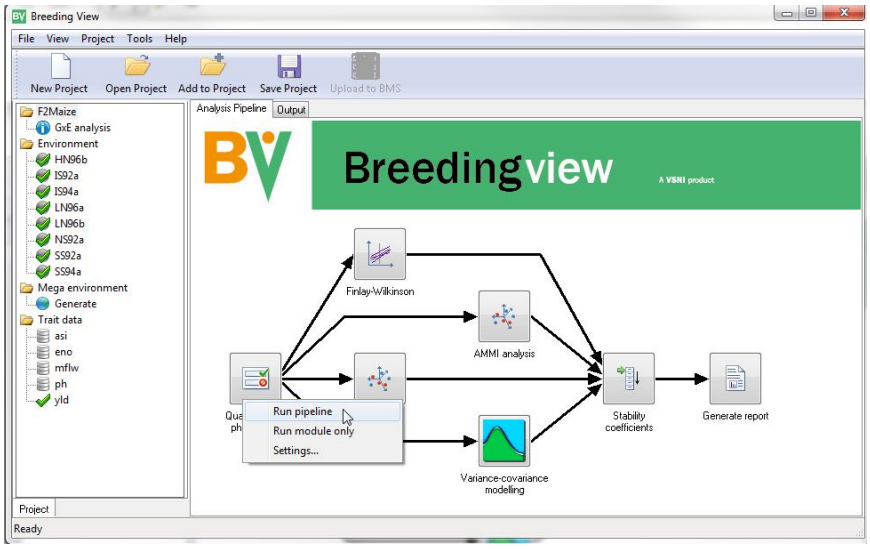
After the data have been imported, the traits *asi, eno, mflw, ph and yld* will now appear within the **Trait data** folder in the **Project** tab on the left-hand side. In addition, the environments *HN96b, IS92a, IS94a, LN96a, LN96b, NS92a, SS92a* and *SS94a* now appear within the **Environment** folder in the **Project** tab on the left-hand side. Right-click on the trait *yld*, and select the **Set as active trait** menu item on the short-cut menu.



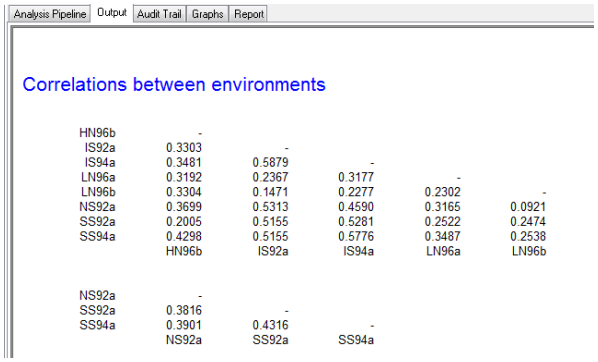
4.3 Running the GxE Analysis pipeline

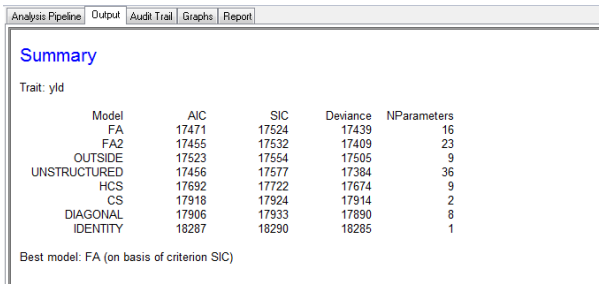
4.3.1 Single trait

To run the analysis pipeline, click on the right-mouse button on the first node (*Quality control phenotypes*), and select **Run pipeline** from the shortcut menu. This will run the whole analysis pipeline, by performing the task at each node in turn. As the task at each node is completed, the connecting arrow will change colour to indicate that the pipeline is moving onto the next task.

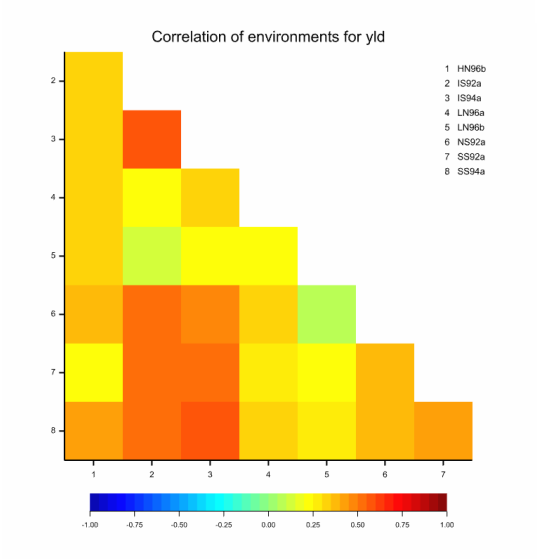


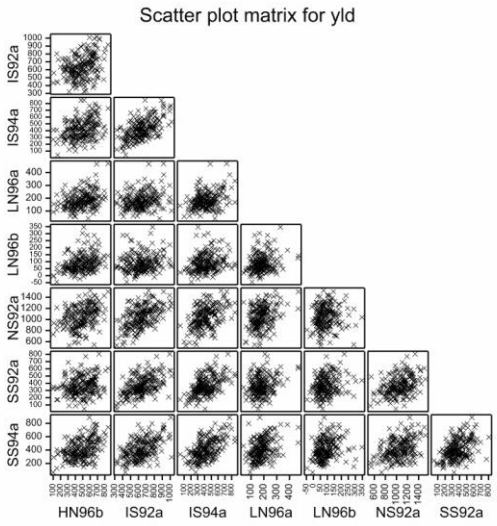
When the analysis pipeline has completed, all the intermediate output from the analysis at each node appears in the **Output** tab. For the GxE analysis pipeline, it will contain summary statistics for the *yld* trait both within and between the environments, a Finlay-Wilkinson joint regression, an AMMI analysis, stability statistics and a summary of the different variance-covariance models.





By default, the summary statistics node provides a graph of correlations and a scatter matrix between environments for the raw data, allowing exploration of the relationship between environments. The graphs are displayed within the **Graph** tab. Environments with similar characteristics will induce similar results, which will result in higher genetic correlations. The images below show high correlation between some environments, such as IS92a and IS94a as compared to low correlation between some environments, such as IS92a and LN96b.





The Finlay-Wilkinson node produces a joint regression analysis (see Finlay & Wilkinson, 1963) to investigate the interaction between genotypes and environments. The analysis aims to characterize the *sensitivity* of each genotype to environmental effects, by fitting a regression of the environment means for each genotypes on the average environmental means. Sensitivity provides a way of assessing the stability of the genotypes. The responses of genotypes with low sensitivity values are more stable with respect to changes of environment. The output includes an analysis of variance table for the regression fit, a list of the sensitivity values for each genotype and some associated plots. In the estimates a value of 1 represents the average sensitivity and genotypes with a value greater than one exhibit higher than average sexitivity and genotypes with a value less than 1 are less senstive than average. In the analysis of variance table, the Sensitivities are highly significant according to the F test ($F_{210,1260}$=2.37; $p < 0.001$). The table below shows a selection of the sensitivity values from the analysis:

Finlay and Wilkinson modified joint regression analysis

Response variate:	yld
Number of genotypes:	211
Number of environments:	8
Convergence criterion:	0.001000
Number of iterations:	3

Analysis of variance

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Genotypes	210	13821018.2745	65814.3727	6.32	<0.001
Environments	7	127771686.9700	18253098.1386	1753.13	<0.001
Sensitivities	210	5178199.4885	24658.0928	2.37	<0.001
Residual	1260	13118797.2995	10411.7439		
Total	1687	159889702.0325	94777.5353		

Estimates

Genotype	Mean	s.e.	Sensitivity	s.e.	Mean square deviation	Rank
G001	510.4	36.08	1.2367	0.1311	7344	184
G002	485.3	36.08	1.1141	0.1311	10753	146
G003	474.9	36.08	1.1136	0.1311	3438	144

Genotype	Response in average environment	Sensitivity
G207	405.4	0.61
G121	604.1	0.98
G012	377.3	1.01
G018	446.8	1.02
G025	573.1	1.27

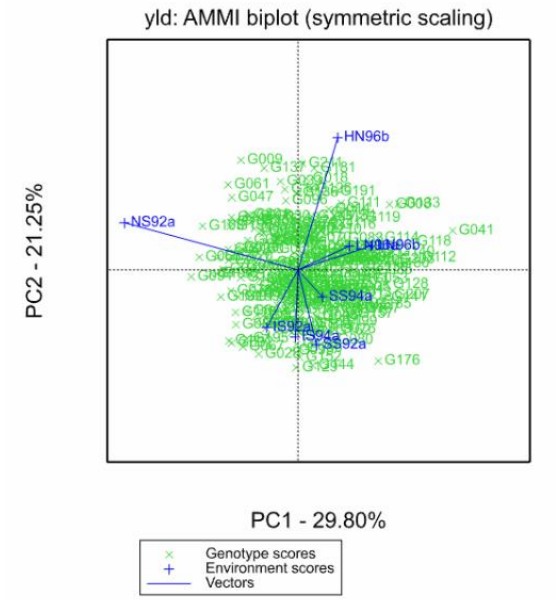
In the average environment G025 and G121 have the highest means, and perform better than G041, G012 and G018. Although G121 performs slightly better than G025 in the average environment, G025 has a higher sensitivity value and will exploit improved environmental conditions better than G121. Similarly, G207 performs better than G012 in the average environment, but has low sensitivity and will not benefit from the better environmental conditions and hence will not perform as well as G012 in these conditions.

The AMMI analysis node fits a model which involves the Additive Main effects of ANOVA with the Multiplicative Interaction effects of principal components analysis (PCA). The AMMI model is more flexible than the Finlay-Wilkinson model in that more than one environmental quality variable is allowed where they are explained using multiplicative terms. A desriable property of the AMMI model is that genotype and environmental scores can be used to construct biplots to help interpret genotype-by-environment interaction. In the biplot, genotypes that are similar to each other are closer in the plot than genotypes that are different. Similarly, environments that are similar will group together as well. When environment scores are connected to the origin of the plot, an acute angle between lines indicate a positive correlation between environments. A right angle between lines indicates low or no correlation between environments, and an obtuse angle indicates negative correlation. The projection of a genotype onto the environmental axis reflects performance in that particular environment. The AMMI node produces an Analysis of Variance and associated biplot.

AMMI Analysis

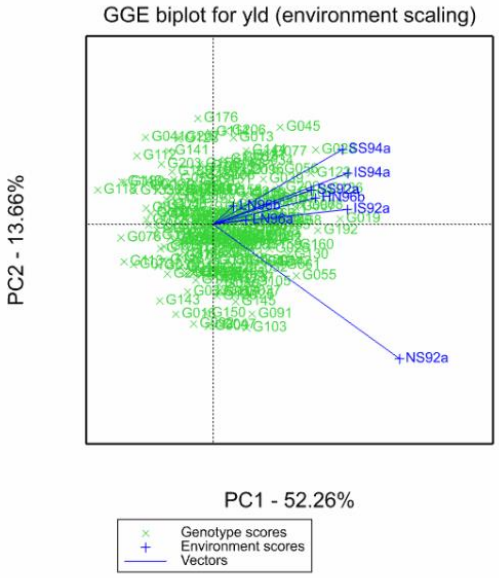
ANOVA table for AMMI model

Source	d.f.	s.s.	m.s.	v.r.	F pr
Genotypes	210	13821018	65814	5.29	<0.001
Environments	7	127771687	18253098	1466.47	<0.001
Interactions	1470	18296997	12447		
IPCA 1	216	5451796	25240	2.93	<0.001
IPCA 2	214	3888148	18169	2.11	<0.001
Residuals	1040	8957053	8613		



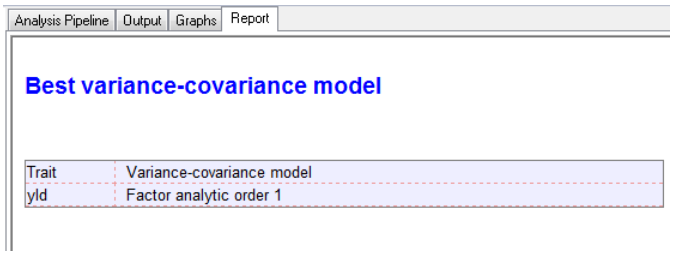
The analysis-of-variance table for the AMMI shows that the genotype-by-environment interaction is explained by two terms (IPCA 1 and IPCA 2) that are both highly significant ($F_{216,1040}$=2.93; $p < 0.001$ and $F_{214,1040}$=2.11; $p < 0.001$). In the AMMI biplot the angle between the environmental axes for IS94a and IS96a) is acute, indicating high correlation between these environments. There is a right angle between NS92a and HN96b, indicating no or little correlation between environments. The obtuse angle between HN96b and IS92a indicates negative correlation. The projection of genotype G041 to the NS92a enviroment axis shows a negative interaction with the environment, as compared to G061 which has a positive adaption.

The GGE biplot node provides a modification of the model fitted in the AMMI node, which joins the effects of the genotypic main effects and the genotype-by-environment interaction (Yan & Kang, 2003). As this describes both the genotypic main effects and genotype-by-environment interaction together, it is known as a GGE model, and the biplots are called GGE biplots. The interpretion of the GGE biplot is similar to the AMMI biplot, but now the genotypes are distributed according to overall performance in each environment, rather than just genotype-by-environment interaction.

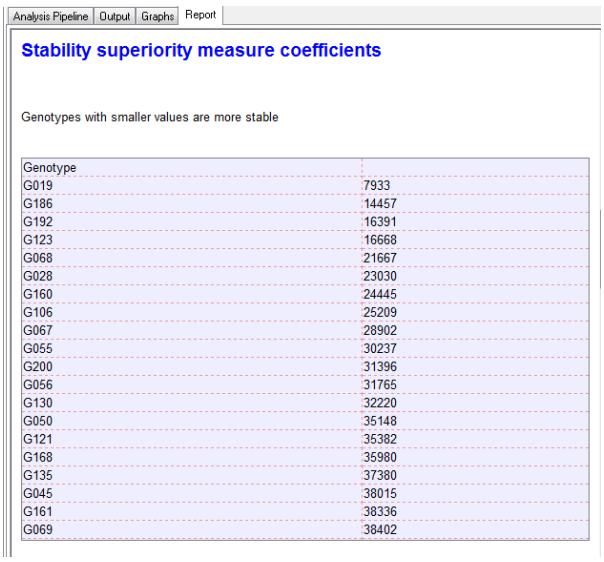


In the GGE biplot, the best performing genotypes are on the right-hand side of the plot. It can be seen that poor performing genotypes, such as G_041, are on the left of the plot and higher yield genotypes, such as G_055, are on the right-hand side.

The variance-covariance modelling node produces an alternative method for modelling genotype-by-environment interaction. The AMMI, GGE and Finlay-Wilkinson nodes model the mean response, but this node uses mixed models to model the genotype-by-environment interaction in terms of heterogeneity of variances and covariances. The node evaluates a range of possible variance-covariance models and their goodness of fit is used to select the best based on an information criterion. This is an optional node and can be disabled from the analysis pipeline by using the settings for the node.



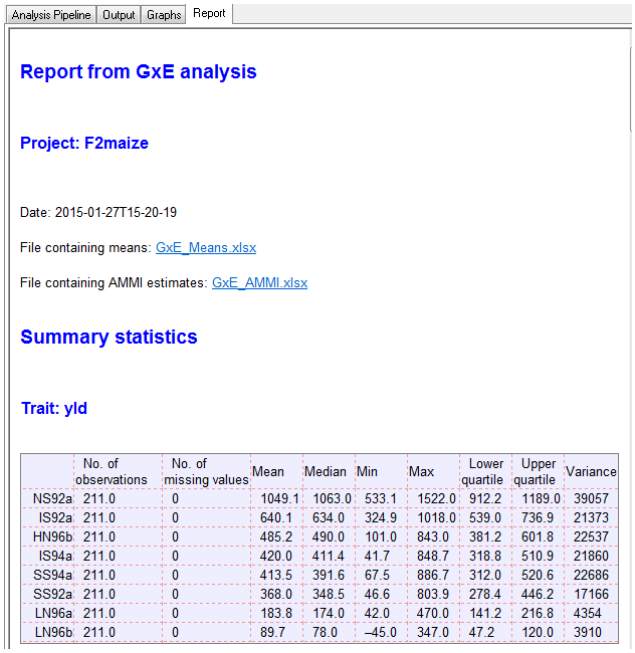
The stability coefficients node produces tables showing three different stability coefficient measures for the genotypes, to assess their overall reliability or stability with a trait. The superiority performance calculates the *cultivar-superiority measure* of Lin & Binns (1988). For each genotype, this is the sum of the squares of the difference between its mean in each environment and the mean of the best genotype there, divided by twice the number of environments. Genotypes with the smallest values of the superiority tend to be more stable, and closer to the best genotype in each environment. The *static stability coefficient* is defined as the variance between its mean in the



Genotype	
G019	7933
G186	14457
G192	16391
G123	16668
G068	21667
G028	23030
G160	24445
G106	25209
G067	28902
G055	30237
G200	31396
G056	31765
G130	32220
G050	35148
G121	35382
G168	35980
G135	37380
G045	38015
G161	38336
G069	38402

various environments. This provides a measure of the consistency of the genotype, but without taking account of how good its performance is. Wricke's ecovalence produces Wricke's (1962) *ecovalence stability coefficient*. This is the contribution of each genotype, to the genotype-by-environment sum of squares, in an unweighted analysis of the genotype-by-environment means. A low value indicates that the genotype responds in a consistent manner to changes in environment. The output shows the top 20 genotypes using the superiority performance. The genotype G_019 has the smallest value and is the most stable.

All the results are collated and displayed in the **Report** tab within an HTML document. The HTML report is saved in a file within the working project folder, in a subfolder using the date and time. The file is automatically named *GxE_report.htm*, and any associated graphs are saved in the same folder.



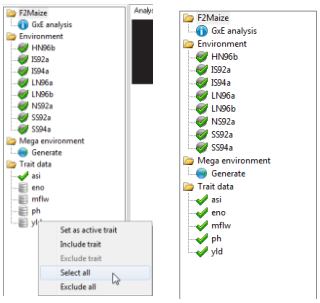
	No. of observations	No. of missing values	Mean	Median	Min	Max	Lower quartile	Upper quartile	Variance
NS92a	211.0	0	1049.1	1063.0	533.1	1522.0	912.2	1189.0	39057
IS92a	211.0	0	640.1	634.0	324.9	1018.0	539.0	736.9	21373
HN96b	211.0	0	485.2	490.0	101.0	843.0	381.2	601.8	22537
IS94a	211.0	0	420.0	411.4	41.7	848.7	318.8	510.9	21860
SS94a	211.0	0	413.5	391.6	67.5	886.7	312.0	520.6	22686
SS92a	211.0	0	368.0	348.5	46.6	803.9	278.4	446.2	17166
LN96a	211.0	0	183.8	174.0	42.0	470.0	141.2	216.8	4354
LN96b	211.0	0	89.7	78.0	−45.0	347.0	47.2	120.0	3910

The report contains links to two Excel files. The first contains the means used for the analysis within two tabs. The tab labelled **Averaged** lists the genotype values averaged over environments, and the tab **yld** displays the means for each environment along with their ranking. The second Excel file contains estimates from AMMI analysis. Similarly to the file of means the tab labelled **Averaged** lists the genotype values averaged over environments, and the tab **yld** displays the means for each environment along with their ranking.

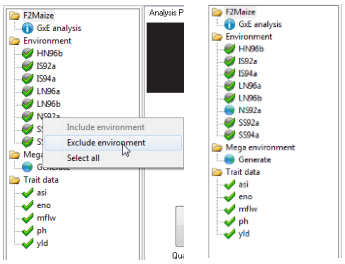
The remainder of the report contains summary statistics, Finlay-Wilkinson joint regression analysis, an AMMI analysis, a GGE biplot, tables of stability coefficients and a summary of the best variance-covariance model.

4.3.2 Multiple traits

Multiple traits can be run simultaneously, with the results combined into a single report. To run the analysis pipeline on multiple traits, click on the mouse button on any of the traits, and choose the **Select all** item on the shortcut menu. Each trait will now have a green tick indicating that it will be included in the run. To run the analysis pipeline, click on the right-mouse button on the first node (*Quality control phenotypes*), and select **Run pipeline** from the shortcut menu. Note that when multiple traits have been selected, the whole pipeline must be run.



By default, all environments are selected for a GxE analysis, however a subset can be chosen. To exclude environments, click on the right mouse button on the environment to exclude, and select **Exclude environment** from the shortcut menu. The example below shows the environment *NS92a* being excluded.



Afer running the pipeline, a combined report is produced. The output for the stability coefficients, that is displayed in the report, can be controlled using the settings. The coefficients can be sorted and subset to display only the best genotypes.

Analysis Pipeline | Output | Audit Trail | Graphs | Report

Best variance-covariance model

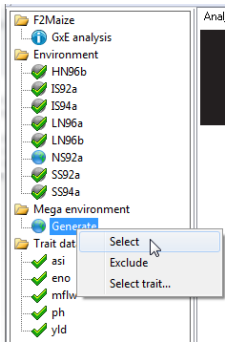
Trait	Variance-covariance model
asi	Factor analytic order 1
eno	Uniform correlations with unequal variances
mflw	Factor analytic order 2
ph	Factor analytic order 2
yld	Factor analytic order 1

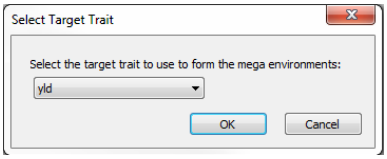
Stability superiority measure coefficients

fTrait Genotype	asi	eno	mflw	ph	yld
G162	35	13	9	539	45969
G139	33	14	13	1246	112205
G134	32	18	27	1740	109739
G156	30	14	38	559	65123
G171	30	11	30	2050	111451
G132	29	16	29	1518	74609
G178	28	10	22	688	90768
G022	28	15	25	1201	83776
G129	28	8	31	466	50171
G116	27	6	4	395	43818
G200	27	5	8	1133	28567
G136	27	7	21	1263	86560
G198	26	15	46	338	51206

4.3.3 Mega environments

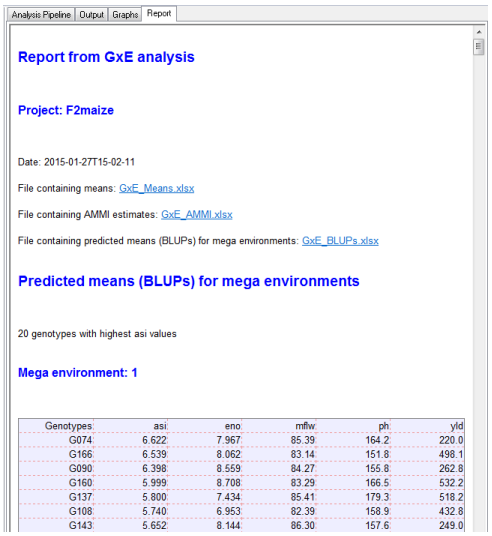
An analysis can peformed to determine the presence of mega environments, and then produce results based on those mega environments. To perform this analysis, right-click on the **Generate** item within the **Mega environment** folder, and choose the **Select** item from the shortcut menu. When the **Generate** item is selected, mega environments will be formed based on the winning genotype from each environment using an AMMI-2 model. If there are two or more traits being analysed, then a *target* trait needs to be selected from which the mega environments will be formed. To select a *target* trait, right-click on the **Generate** item, and choose the **Select trait** item from the shortcut menu. This will open a dialog where the *target* trait to use to form the mega environments can be selected. In this dialog select *yld* for the *target* trait.





To run the analysis right-click on the **Quality control phenotypes** node, and select **Run pipeline** from the shortcut menu.

The analysis produces additional output to include the predicted means (BLUPs) within each mega environment based on a mixed model. The report is extended to include a link to an Excel file (.xlsx) containing the predicted means (BLUPs) for the mega environments from fitting a mixed model. The file contains a sheet for the predicted means for all traits within each mega enviroment for, and a sheet labelled **All BLUPs** which combines all the data from the other sheets into a single table. Additional output is produced where tables of the predicted means (BLUPs) are displayed for each mega environment.



5. Single Trait QTL Mapping (Single Environment)

The objective of this example is to illustrate the simplest situation for QTL mapping by linkage analysis, which is the detection of QTLs in a single environment.

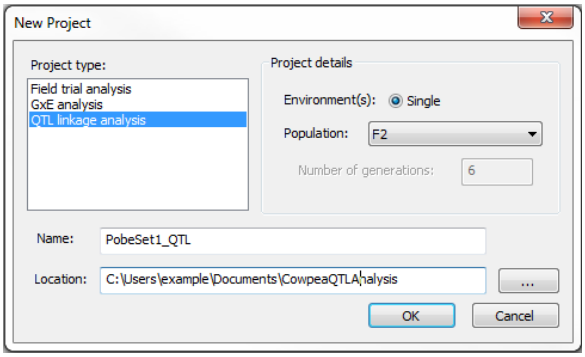
5.1 Data

The three basic data files that are necessary to perform a QTL analysis are the map file, the genotype file and the phenotype file. The map file contains information of marker locations (linkage group and position within linkage group), and the genotype file contains the marker scores of each individual in the population. The map and genotype files are plain text files in Flapjack format; details of the layout can be found at: http://bioinf.scri.ac.uk/flapjack/help/projects.shtml. The phenotype data file contains the observations of one or more traits of each individual in the population. The genotypic data for this example are for an F_2 population of 288 individuals. The population has been genotyped with 164 markers, stored in the file *Cowpea_genotypes.txt*. The map file *Cowpea_map.txt* contains 159 of the 164 markers mapped onto 11 linkage groups. The traits for the analysis are stored in *Burkina_trait_means_for_qtl_analysis.csv* (generated in Section 1).

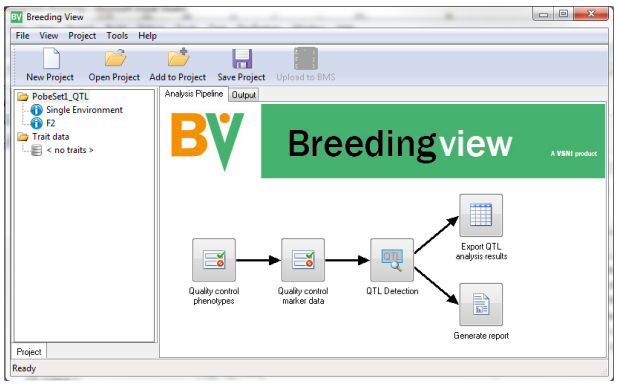
5.2 Breeding View Project

To run a QTL linkage analysis, a new Breeding View project needs to be set up. To create a new project, select **File | New Project** from the menu bar, or click on the **New Project** tool button.

In the **New Project** dialog, select the **QTL linkage analysis** item within the **Project type** list. Each type of project requires details to be supplied. Select **F2** from the **Population** list. The project needs to be given a name and a location to store the working files. Enter *PobeSet1_QTL* as the **Name** of the project, and then browse for a folder to specify the **Location** where the working files will be stored. In the dialog above, a folder has been specified called *CowpeaQTLAnalysis* within *My Documents (C:\Users\example\Documents)* for the Location. Click on **OK** to create the project.

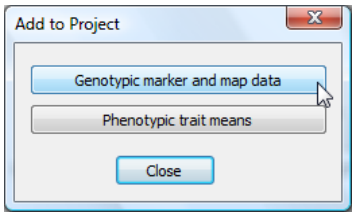


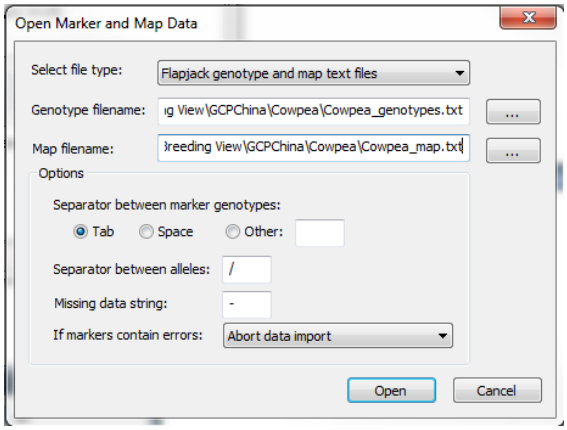
When a project has been created, the details are displayed in a tab called **Project** on the left-hand side. The *PobeSet1_QTL* folder is the project name, and contains attributes about the project. The **Trait data** folder lists all the available traits for analysis. The nodes in the **QTL linkage analysis** pipeline are as follows:



Nodes	Description
Quality control phenotypes	Summary statistics for the active trait
Quality control marker data	Summary statistics for marker data
QTL Detection	Performs genome-wide scan using SIM and CIM
Export QTL analysis results	Allows data to be exported into a Flapjack project file and automatically displayed in Flapjack
Generate report	HTML report of QTL results

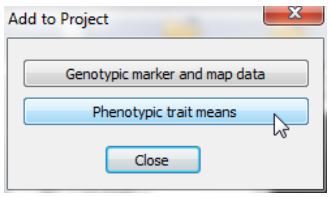
The genotypic and phenotypic data need to be imported into the project before the analysis pipeline can be run. To import the genotypic data, select **Project | Add Data** from the menu, or click on the **Add to Project** tool button. On the **Add to Project** dialog, click on the **Genotypic and map data** button.

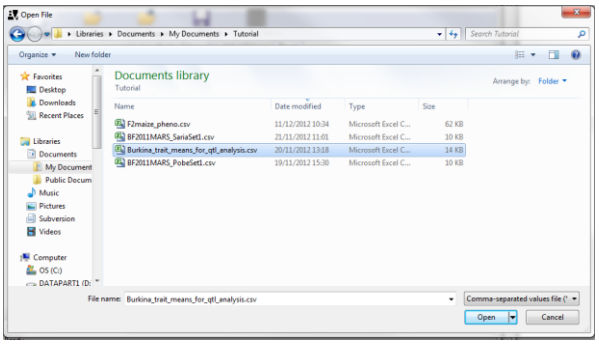




In the **Open Marker and Map Data** dialog, select the **Flapjack genotype and map text files** item from the **Select file type** list. For the **Genotype filename** browse for the *Cowpea_genotypes.txt* file, and for the **Map filename** browse for the *Cowpea_map.txt file*. You can browse for a filename by clicking on the **[...]** button. Click **Open** to import the genotypic data.

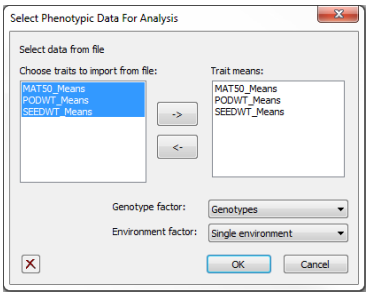
To import the phenotypic data, select **Project | Add Data** from the menu, or click on the **Add to Project** tool button. On the **Add to Project** dialog, click on the **Phenotypic trait means** button.



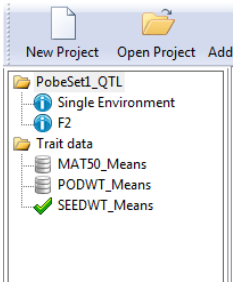


In the **Open File** dialog, go to the folder containing the *Burkina_trait_means_for_qtl_analysis.csv* file and click **Open**.

This opens the **Open Phenotypic Means (Traits)** dialog, where the data to be imported can be selected. Select the names *MAT50_Means, PODWT_Means and SEEDWT_Means* in the **Columns in file** list, and click on the **->** button to transfer the names to the **Trait means** list. Next select *Genotypes* from the drop list for the **Genotype factor**. Click **OK** to import the data.



After the data have been imported, the traits *MAT50_Means, PODWT_Means and SEEDWT_Means* will appear within the **Trait data** folder in the **Project** tab on the left-hand side. Right-click on the trait *SEEDWT_Means*, and select the **Set as active trait** menu item on the short-cut menu. This trait will then have a green tick next to it, to indicate that it is the trait that will be used within the analysis pipeline.



5.3 Running the QTL Linkage Analysis pipeline

5.3.1 Automatic Subset of Data

When there is a difference between the genotype ids for the phenotypic and genotypic data, the analysis pipeline will use a subset of the data, that includes only the genotype ids in common between the genotypic and phenotypic data.

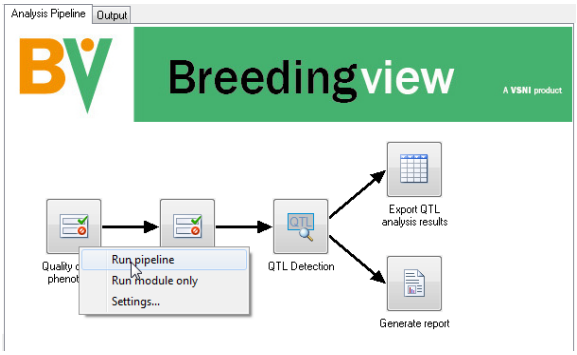
The node for the **Quality control of markers** checks the markers scores for missing observations and, if either the individuals or markers contain 50% or more missing observations, they are removed from the analysis. The threshold, and whether a subset is applied, can be controlled by right-clicking on the **Quality control of markers** node, and selecting the **Settings** item on the shortcut menu.

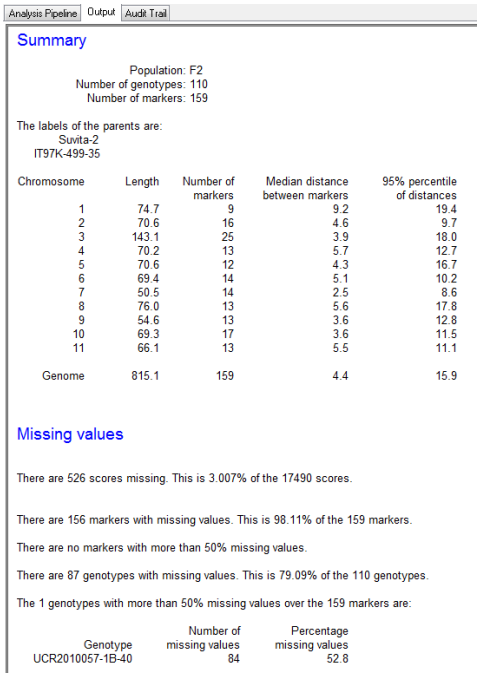
5.3.2 QTL Detection

The search of QTLs is done by testing the presence of a QTL at each of the different evaluation positions over the chromosomes. When the test is performed only at marker positions, this is called **marker-based QTL detection**; when in addition to the marker positions tests are done in-between markers, this is called **Interval Mapping.** To increase the power of the genome-wide QTL search, **composite interval mapping** (CIM) can be used. The idea of CIM is to include a number of cofactors in the QTL scan model that control for the variation caused by the genetic background (i.e. variation caused by QTLs outside the region where the QTL is tested). For example, after an initial scan of QTLs by SIM, one can perform CIM using the candidate QTLs detected in the SIM scan as cofactors. This can be repeated one or more times, until the list of detected QTLs does not change. By default, the QTL analysis uses two rounds of CIM. To avoid co-linearity between cofactors and tested positions, cofactors have to be removed temporarily from the model when testing for QTLs close to cofactor positions. The window within which cofactors are removed is set by default to 50 cM.

5.3.3. Running the Pipeline

To run the QTL analysis pipeline, click on the right-mouse button on the first node (**Quality control phenotypes**), and select **Run pipeline** from the shortcut menu.




Analysis Pipeline | Output | Audit Trail

Summary

Population: F2
Number of genotypes: 110
Number of markers: 159

The labels of the parents are:
Suvita-2
IT97K-499-35

Chromosome	Length	Number of markers	Median distance between markers	95% percentile of distances
1	74.7	9	9.2	19.4
2	70.6	16	4.6	9.7
3	143.1	25	3.9	18.0
4	70.2	13	5.7	12.7
5	70.6	12	4.3	16.7
6	69.4	14	5.1	10.2
7	50.5	14	2.5	8.6
8	76.0	13	5.6	17.8
9	54.6	13	3.6	12.8
10	69.3	17	3.6	11.5
11	66.1	13	5.5	11.1
Genome	815.1	159	4.4	15.9

Missing values

There are 526 scores missing. This is 3.007% of the 17490 scores.

There are 156 markers with missing values. This is 98.11% of the 159 markers.

There are no markers with more than 50% missing values.

There are 87 genotypes with missing values. This is 79.09% of the 110 genotypes.

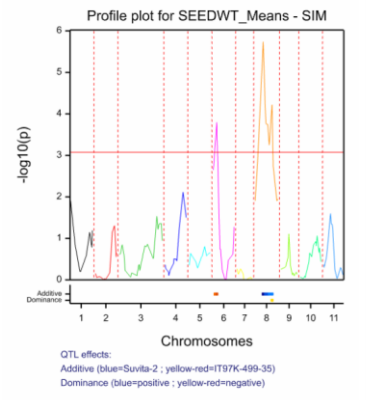
The 1 genotypes with more than 50% missing values over the 159 markers are:

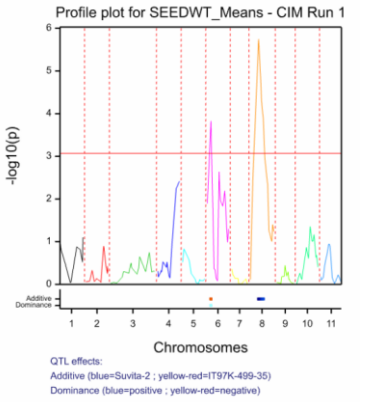
Genotype	Number of missing values	Percentage missing values
UCR2010057-1B-40	84	52.8

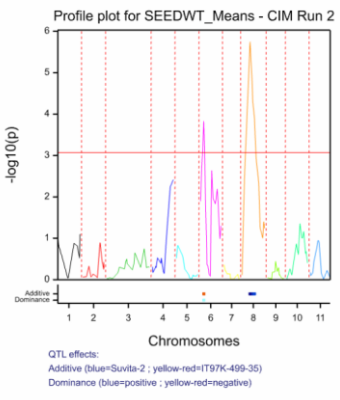
When the analysis pipeline has completed, all the intermediate output from the analysis at each node appears in the **Output** tab. For the QTL analysis pipeline, it will contain summary statistics for the *SEEDWT_Means* trait, summary statistics for the marker data and output from the QTL detection.

The summary for the marker data, displayed in the Output tab, provides details of the numbers of markers within each linkage groups and the number of missing observations. In this analysis, there are no markers with 50% or more missing observations. However, there is one genotype reported (UCR2010057-1B-40) that contains more than 50% missing observations, so this has been removed from the QTL detection.

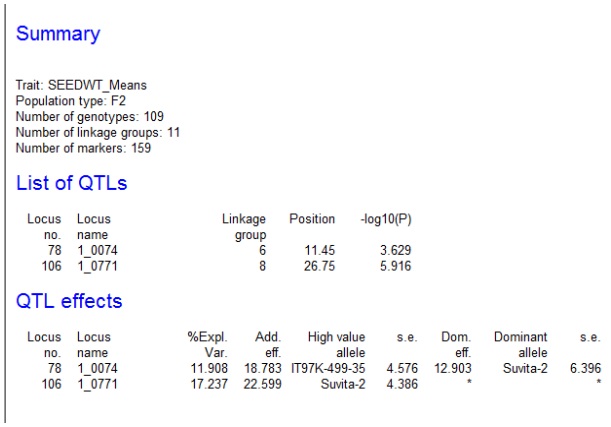
The QTL detection produces graphs of the scan profiles for the simple interval mapping and each round of composite interval mapping. The graphs are displayed within the **Graph** tab. The profiles for the SIM and two rounds of CIM for the additive effects are shown below. Each graph displays the profile within each linkage group, the threshold of detection and parent contributor of the high value allele.







Output is produced for each round of SIM and CIM, with a summary of the of the loci greater than the detection threshold and candidate QTLs. After the last round of CIM, the next step is to estimate their effects. This is done by including all the QTLs in a model simultaneously, and then using backward selection to determine whether all the QTLs are significant. The final step is to estimate the effects of significant set of QTLs. This analysis shows two significant QTLs: for 1_0074 in linkage group 6, and 1_0771 in linkage group 8.



Summary

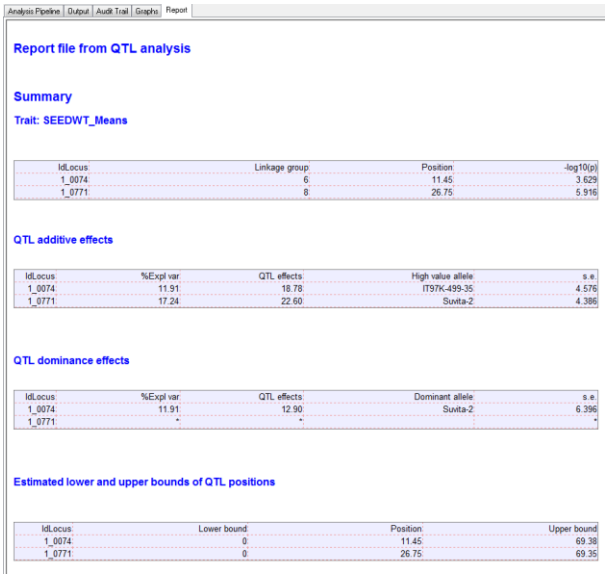
Trait: SEEDWT_Means
Population type: F2
Number of genotypes: 109
Number of linkage groups: 11
Number of markers: 159

List of QTLs

Locus no.	Locus name	Linkage group	Position	-log10(P)
78	1_0074	6	11.45	3.629
106	1_0771	8	26.75	5.916

QTL effects

Locus no.	Locus name	%Expl. Var.	Add. eff.	High value allele	s.e.	Dom. eff.	Dominant allele	s.e.
78	1_0074	11.908	18.783	IT97K-499-35	4.576	12.903	Suvita-2	6.396
106	1_0771	17.237	22.599	Suvita-2	4.386	*		*



Report file from QTL analysis

Summary

Trait: SEEDWT_Means

IdLocus	Linkage group	Position	-log10(p)
1_0074	6	11.45	3.629
1_0771	8	26.75	5.916

QTL additive effects

IdLocus	%Expl var	QTL effects	High value allele	s.e.
1_0074	11.91	18.78	IT97K-499-35	4.576
1_0771	17.24	22.60	Suvita-2	4.386

QTL dominance effects

IdLocus	%Expl var	QTL effects	Dominant allele	s.e.
1_0074	11.91	12.90	Suvita-2	6.396
1_0771	*	*		*

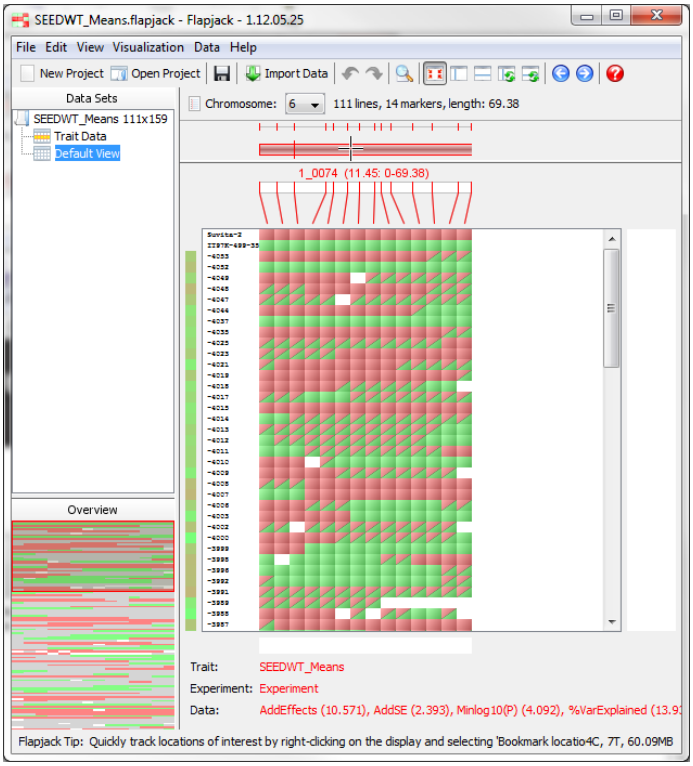
Estimated lower and upper bounds of QTL positions

IdLocus	Lower bound	Position	Upper bound
1_0074	0	11.45	69.38
1_0771	0	26.75	69.35

The final results are collated and displayed in the **Report** tab within an HTML document. The HTML report is saved in a file within the working project folder, in a subfolder using the date and time. The file is automatically named *SEEDWT_Means_Report.htm*, and the associated graph for the genetic map is called *SEEDWT_Means_Report_GeneticMap001.png*.

5.3.4 Exporting to Flapjack

If you have Flapjack installed on your PC, the marker and map data can be exported to a Flapjack project file along with the traits and QTL results. Flapjack (http://bioinf.scri.ac.uk/flapjack/) is a tool for graphical genotyping and haplotype visualization, that can routinely handle the large data volumes generated by high throughput SNP and comparable genotyping technologies. Its visualizations are rendered in real-time, allowing for rapid navigation and comparisons between lines, markers and chromosomes. The Flapjack project file is saved in a file within the working project folder in a

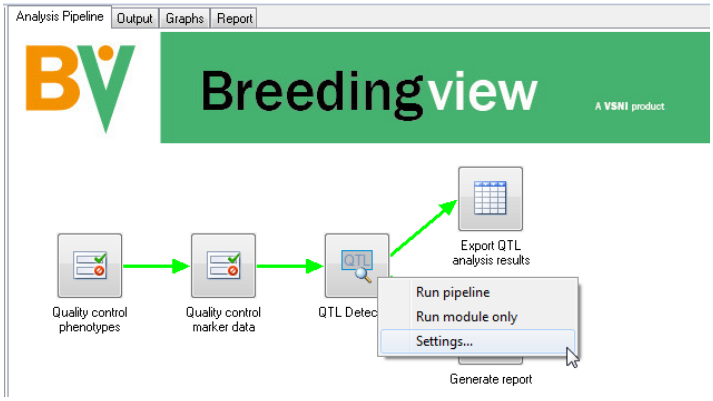


subfolder using the date and time of the run. The file is automatically named *SEEDWT_Means.flapjack.* By default, once the Flapjack file has been created, it is automatically opened within Flapjack.

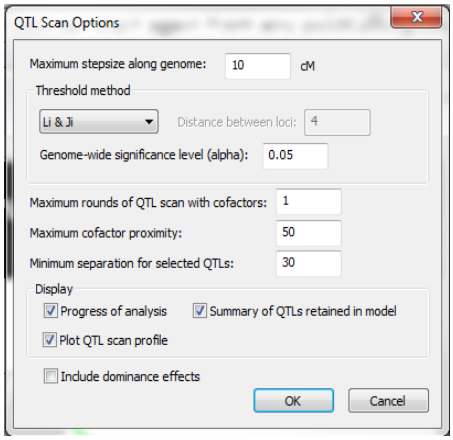
By default, the Breeding View will search for Flapjack in the standard installation folders. If Flapjack has been installed within a non-standard installation folder, then the Breeding View will need to be configured to use the non-standard installation folder. This can be done by using the *Component Manager*, which can be accessed by selecting the **Configuration** item on the **Tools** menu.

5.3.5. Changing Options for the QTL Linkage Analysis pipeline

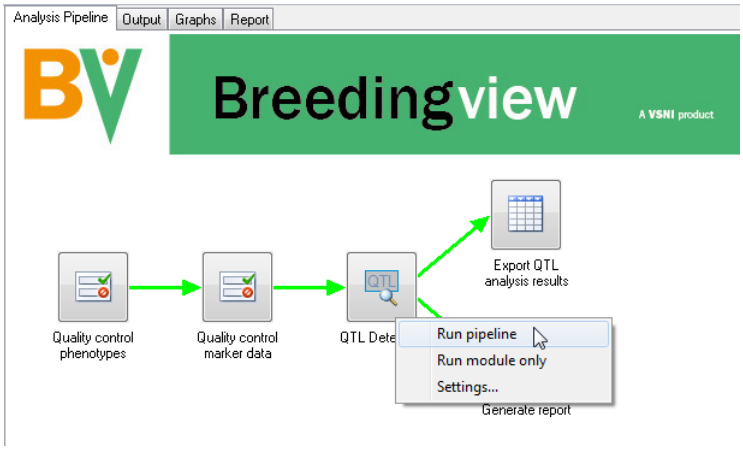
By default, the QTL analysis pipeline is set to include both additive and dominance effects for an F_2 population. To change to only include additive effects, we can change the option within the settings for the **QTL Detection** node. To do this, right-click on the **QTL Detection** node, and select the **Settings** item on the shortcut menu.



This opens a dialog called **QTL Scan Options**. To just use the additive effects, remove the selection of the **Include dominance effects** checkbox. In the first run of the pipeline, the QTL detection was testing only at marker positions. To change to interval mapping, i.e. also test between marker positions, enter a stepsize of 10 in the **Maximum stepsize along genome** box. With this value an evaluation position will be automatically created, whenever the gap between two consecutive markers is larger than 10 cM. This dialog can also be used to specify the number of rounds of CIM. To change to a single round of CIM, enter the value 1 in the **Maximum rounds of QTL scan with cofactors** box.

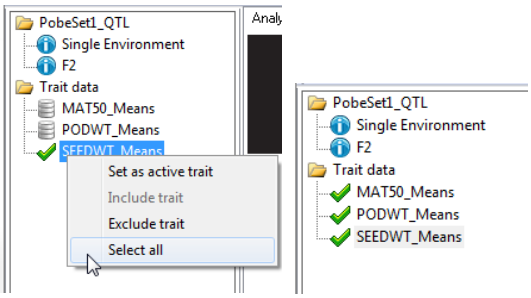


After a pipeline has been run, you can start again from any point along the pipeline. So, to run just the QTL detection and reporting, you should start the analysis again from the **QTL Detection** node. To do this, click on the right-mouse button on the **QTL Detection** node, and select **Run pipeline** from the shortcut menu.

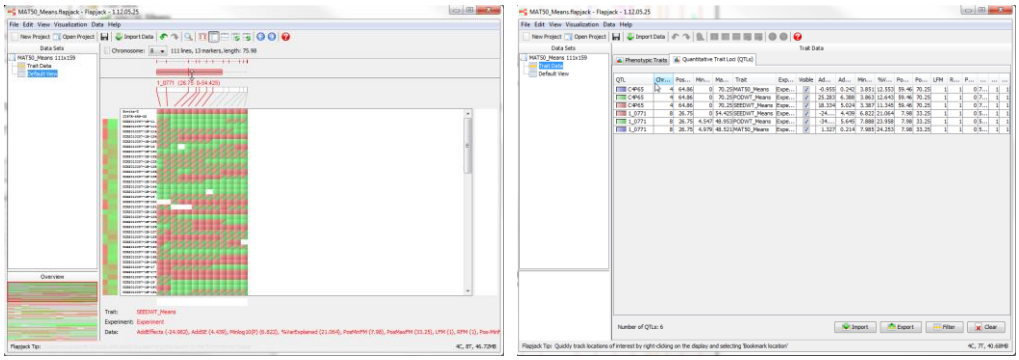


5.3.6 Running multiple traits

Multiple traits can be run simultaneously with the results combined into a single report. To run the analysis pipeline on multiple traits, click on the mouse button on any of the traits, and choose the **Select all** item on the shortcut menu. Each trait will now have a green tick indicating that it will be included in the run. To run the analysis pipeline, click on the right-mouse button on the first node (*Quality control phenotypes*), and select **Run pipeline** from the shortcut menu.



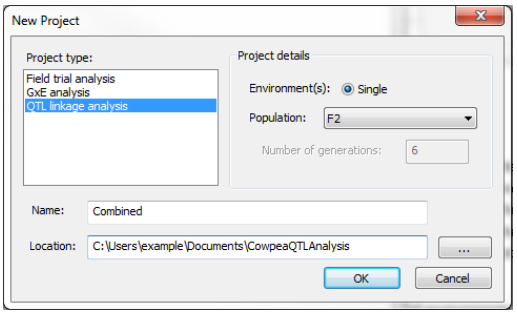
Each trait is analysed in turn and the results are collated into a single report. In addition the resulting qtls for each trait are combined into a single Flapjack file.



5.3.7 Running multiple traits in multiple environments

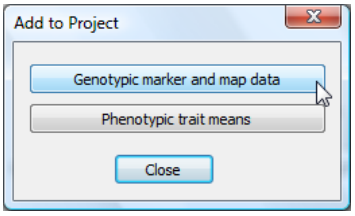
When adjusted means for traits are available from trials conducted at different environments rather than creating a separate project to run the analysis for each environment/site, it may be desirable to run these all in sequence. To analyse a series of environments, the data need to be supplied within a stacked format. The data should include a column to reference the different environments/sites, and the remaining columns should contain the genotypes, and adjusted means for the traits stacked by environment.

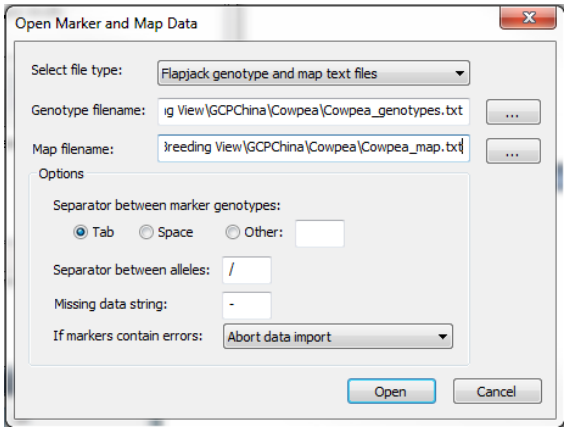
To create a new project for a sequential analysis of environments, select **File | New Project** from the menu bar, or click on the **New Project** tool button. In the **New Project** dialog (see below), select the **QTL linkage analysis** item within the **Project type** list. Select **F2** from the **Population** list. The project needs to be given a name and a location to store the



working files. Enter *Combined* as the **Name** of the project, and then browse for a folder to specify the **Location** where the working files will be stored. In the dialog above, a folder has been specified called *CowpeaQTLAnalysis* within *My Documents (C:\Users\example\Documents)* for the Location. Click on **OK** to create the project.

To import the genotypic data, select **Project | Add Data** from the menu, or click on the **Add to Project** tool button. On the **Add to Project** dialog, click on the **Genotypic and map data** button.

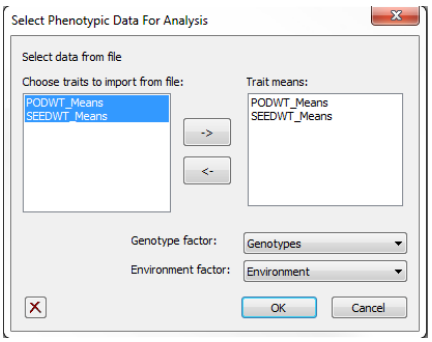


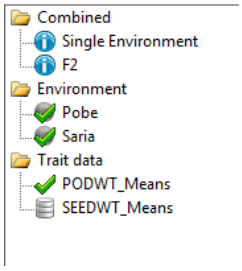


In the **Open Marker and Map Data** dialog, select the **Flapjack genotype and map text files** item from the **Select file type** list. For the **Genotype filename** browse for the *Cowpea_genotypes.txt* file, and for the **Map filename** browse for the *Cowpea_map.txt file*. You can browse for a filename by clicking on the **[...]** button. Click **Open** to import the genotypic data.

To import the phenotypic data, select **Project | Add Data** from the menu, or click on the **Add to Project** tool button. On the **Add to Project** dialog, click on the **Phenotypic trait means** button. In the **Open File** dialog, go to the folder containing the *Results_trait_means_for_qtl_analysis.csv* file and click **Open**.

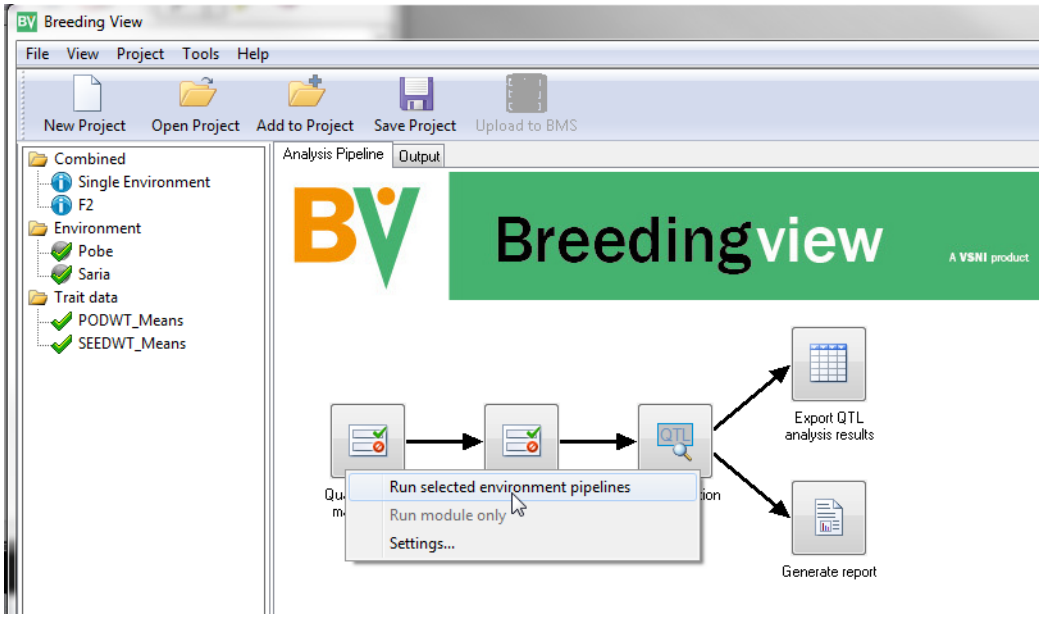
In the **Open Phenotypic Means (Traits)** dialog, select the names *PODWT_Means and SEEDWT_Means* in the **Columns in file** list, and click on the **->** button to transfer the names to the **Trait means** list. Next select *Genotypes* from the drop list for the **Genotype factor** and select *Environment* from the drop list for the **Environment factor**. Click **OK** to import the data.





After the data have been imported, the traits *PODWT_Means* and *SEEWT_Means* will appear within the **Trait data** folder in the **Project** tab on the left-hand side. In addition, the environments/sites *Pobe* and *Saria* appear within the **Environment** folder in the **Project** tab.

To run the sequential analysis for all traits and environments, first click on the right-mouse button on any trait and choose **Select all** from the shortcut menu. Next click on the right-mouse on the first node (*Quality control phenotypes*), and select **Run selected environment pipelines** from the shortcut menu.



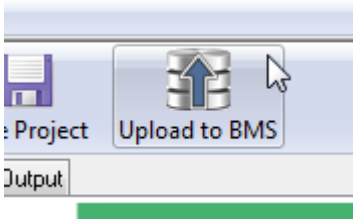
When a sequence of environments are analysed, each environment is analysed in turn, and the results are stored within a folder using the name of the environment. After all environments have been run, a folder called *Combined* is created, which stores the combined results. On completion of the analysis, summary reports for all the environments are produced.

The **Report** tab provides links to each individual environment report and the individual reports contain the combined results for all traits analysed within an environment.

6. Working with the BMS

The Breeding View can be run as part of the Breeding Management System (BMS) within the Integrated Breeding Platform (https://www.integratedbreeding.net/). The BMS is a Workbench, comprising of software tools linked to a database for access to pedigree, phenotypic and genotypic data, developed by GCP's IBP. The BMS incorporates both statistical analysis tools and decision-support tools. These tools are assembled in a way to allow data to flow from one application to the next for the different stages of the crop-breeding process. The Breeding View forms part of the analytical pipeline for statistical analysis of field trials, GxE interaction and QTL mapping.

The Breeding View can be launched from the BMS with a pre-loaded project containing data from the BMS database. After running the analysis the results can be uploaded to the BMS database using the upload button on the toolbar.



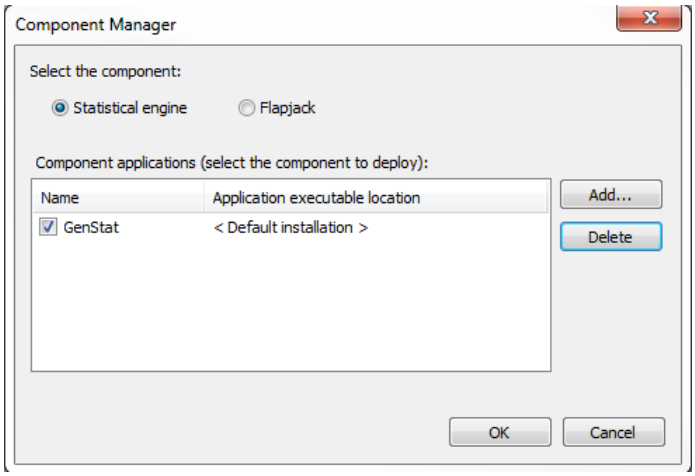
For more details about the BMS please visit:
https://www.integratedbreeding.net/.

7. Statistical Engine

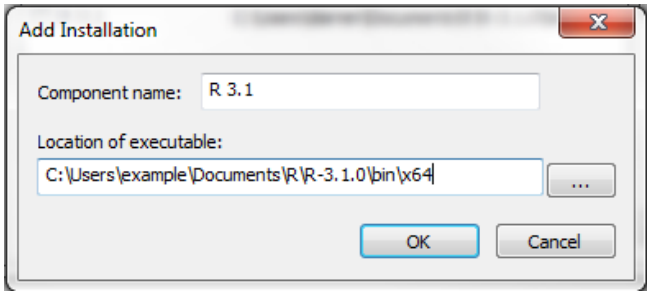
The Breeding View uses GenStat to perform the statistical analysis. The single environment field trial analysis pipeline has been designed to also run using R.

7.1 Running the R statistical engine

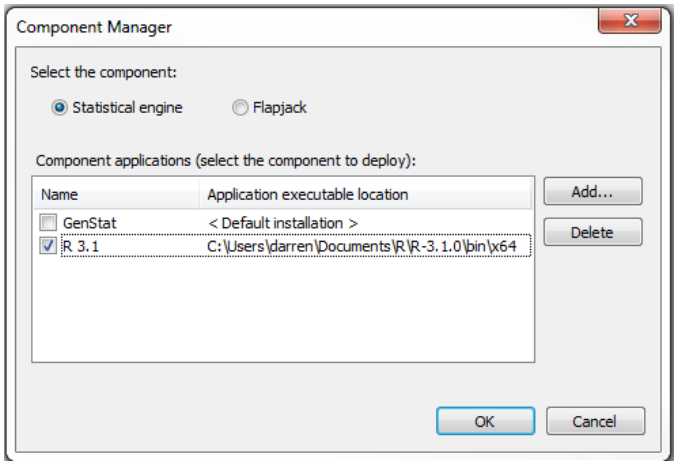
To use R as the statistical the BVRAP package must be installed in the version of R that is to be linked to the Breeding View. To connect R to the Breeding View the location of R needs to be provided using the component manager. To open the component manager select **Tools | Configuration** from the menu.



The link to R needs to be added as a new component by clicking on the **Add** button. This opens the dialog shown below where you can enter the name for the component and the location of binaries for R (usually this will be in bin\i386 or bin\x64 within the R installation).



In the above example a link has been specified for the 64-bit version of R 3.1.0. Clicking Ok will add the new component to list on the main dialog. To change to use R select the R 3.1 item within the component applications list.



7.2 Viewing commands

When the Breeding View is run using either GenStat or R, an audit trail of the commands can be viewed. To view the commands select **View | Audit Trail** from the menu. This will add a tab called **Audit Trail,** and the next time an analysis is run the commands will then be copied into that tab window.

8. Further reading

Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. *Journal of Agricultural Biological and Environmental Statistics* **11**(4), 381-393

Gauch, H.G. (1992). Statistical Analysis of Regional Yield Trials – AMMI analysis of factorial designs. Elsevier, Amsterdam.

Finlay, K.W. & Wilkinson, G.N. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, **14**, 742-754.

Lin, C.S. & Binns. M.R. (1988). A superiority performance measure of cultivar performance for cultivar x location data. *Canadian Journal of Plant Science*, **68**, 193-198.

Wricke, G. (1962). Uber eine method zur erfassung der okogischen streubreite in feldversuchen. Zeitschriff Fur Pflanzenzuchtung, **47**, 92-96.

Yan, W. & Kang, M.S. (2003). *GGE Biplot Analysis: a Graphical Tool for Breeders, Geneticists and Agronomists*. CRC Press, Boca Raton.

Acknowledgements

We are very grateful to those providing us with example data sets. For the cowpea data set we thank Jeff Ehlers, Tim Close, Philip Roberts, Bao Lam Huyuh (University of Riverside team) and Issa Drabo (INERA CRREA, Burkina Faso). The statistical algorithms in the Breeding View were developed in collaboration with the Biometris group at University of Wageningen, whose assistance is gratefully acknowledged.

Feedback

We will greatly appreciate any feedback you may have on the current version of this tool. For questions or feedback, please contact us at breedingview@vsni.co.uk.