Deliverable D5.1 Release through a server of a user-friendly and expert software product and its manual to calculate EBV and GEBV

Planned submission date: 31 August 2019 (M24)
Actual submission date: 10 September 2020 (M37)
Deliverable leader: Progeno BVBA
Versions: 1.0

Communication level	
PU Public	
PP Restricted to other programme participants (including the Commission Services)	
RE Restricted to a group specified by the consortium (including the Commission Services)	
CO Confidential, only for members if the consortium (including the Commission Services)	**CO**

Horizon 2020 of European Union
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

EUCLEG.eu

Table of content

Table of content 2
Introduction 3
I. EBV and GEBV analysis routines 4
1. EBVs 4
2. GEBVs 4
II. User interface 6
III. Manual 6
IV. Training 6

Introduction

Throughout the work packages 1 to 4 of the EUCLEG project, comprehensive phenotypic and genotypic data is being collected for five plant species, two forage legumes (red clover and alfalfa) and three grain legumes (soybean, pea and faba bean). These data points are uploaded, curated and managed by means of a dedicated Progeno server that has been established as part of WP2. This Progeno system represents the central data repository of the EUCLEG project and is on-line accessible for each project partner. Besides data management, Progeno offers various data analysis and visualization tools that allow to capitalize on the collected data.

This deliverable describes the development of specialized data analysis routines for the calculation of Estimated Breeding Values (EBVs) and Genomic Estimated Breeding Values (GEBVs). The EUCLEG project members have access to these routines by means of a user-friendly, graphical interface and a dedicated R (https://www.r-project.org/) library that allows expert users to integrate these functions into their own scripts and programs. A comprehensive tutorial provides instructions and examples for users of the graphical interface while the functions of the R library have been documented according to the R standards. Two on-line training sessions have been organized demonstrating the use of the analysis routines by means of examples and hands-on exercises.

I. EBV and GEBV analysis routines

The breeding value of an individual plant represents its expected contribution to the genetic performance of its offspring for a particular trait. Plant breeders combine parental plants with favourable breeding values to create the next generation which results in a gradual improvement of the breeding pool over each cycle. Unfortunately, obtaining the breeding value of a plant is not straightforward. Field trials allow to obtain phenotypic values (e.g. yield, disease resistance, …) for each individual accession but these are indirect measures of the genetic capacity of the plants. The environment in which each plant is raised generally impacts the phenotypic response and as such blurs the view on the breeding values. The combination of dedicated field trial designs and advanced analysis routines allow to separate genotypic effects from environmental influences and as such produce Estimated Breeding Values or EBVs. Genomic prediction is a relatively new approach that allows to estimate breeding values from the DNA code of the plants, producing GEBVs or Genomic Estimated Breeding Values.

1. EBVs

EBVs are generally obtained by fitting a linear mixed model to the phenotypic observations. The model specifications should match the design of the field trial and assume that the effects of (incomplete) blocks, rows and columns (i.e. the position of the plant in the trial) might affect the measured response value. The task of extracting EBVs from phenotypes therefore reduces to finding a model specification that fits the observations.

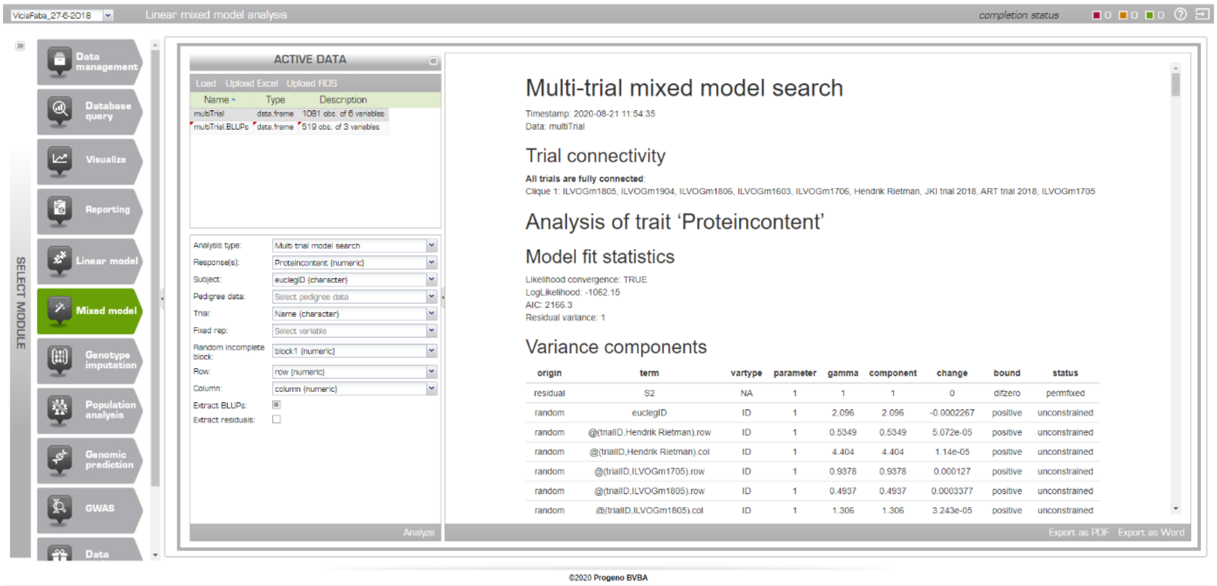
As part of Task 5.1 of the Eucleg project, the analysis routines of the Progeno Analytics module have been extended with a Linear Mixed Model toolbox. This toolbox allows to fit mixed model definitions to phenotypic observations collected in single and multi-environment trials. Users can specify the model they wish to use by including or excluding assumed effects for replications, incomplete blocks, rows and columns. The analysis of the specified model produces model fit statistics, allowing users to compare the various candidate models. The mixed model routines produce various diagnostic plots to visually assess modelling assumptions. The EBVs produced by the model are made available in a separate dataset.

Finding a good model can be a tedious process, especially for multi-environment trials where a different model is needed for each individual trial location. Therefore, the mixed model toolbox offers single- and multi-trial model search routines. These routines will iteratively fit all members of a predefined set of candidate models and return the fit statistics and EBVs for the model specification that demonstrates the best fit. These automated search routines allow even non-experienced users to extract reliable EBVs from their phenotypic observations. Figure 1 shows a screenshot of the graphical interface of the Progeno Analytics module in which a multi-trial mixed model search has been performed on protein content observations collected in 9 EUCLEG field trials.

2. GEBVs

To obtain GEBVs from molecular marker scores (i.e. DNA codes at specific locations of the genome) one requires a trained genomic prediction model. The training of such a model requires the availability of a set of accessions for which both marker scores and phenotypes (preferably EBVs) are available. The required marker scores are deliverables of WP1 while phenotypic observations are produced in WP3, WP4 and WP5. The collected phenotypic and genotypic data sources are centralized in the Progeno database system and readily available for training and validating genomic prediction models.

There are numerous genomic prediction methods that differ extensively in assumptions about the genetic architecture of the trait, computational performance, prediction accuracy and reliability. In terms of prediction accuracy, there is no single prediction method that outperforms the others as results tend to differ widely between datasets. As part of task 5.1, the Progeno Analytics module has been extended with a genomic prediction module that provides 4 different prediction methods:



Name	Type	Description
multiTrial	data.frame	1081 obs. of 6 variables
multiTrial.BLUPs	data.frame	519 obs. of 3 variables

Analysis type: Multi trial model search
Response(s): Proteincontent {numeric}
Subject: euclegID {character}
Pedigree data: Select pedigree data
Trial: Name {character}
Fixed rep: Select variable
Random incomplete block: block1 {numeric}
Row: row {numeric}
Column: column {numeric}
Extract BLUPs: ☒
Extract residuals: □

Multi-trial mixed model search

Timestamp: 2020-08-21 11:54:35
Data: multiTrial

Trial connectivity

All trials are fully connected:
Clique 1: ILVOGm1805, ILVOGm1904, ILVOGm1806, ILVOGm1603, ILVOGm1706, Hendrik Rietman, JKI trial 2018, ART trial 2018, ILVOGm1705

Analysis of trait ‘Proteincontent’

Model fit statistics

Likelihood convergence: TRUE
LogLikelihood: -1062.15
AIC: 2166.3
Residual variance: 1

Variance components

origin	term	vartype	parameter	gamma	component	change	bound	status
residual	S2	NA	1	1	1	0	difzero	permfixed
random	euclegID	ID	1	2.096	2.096	-0.0002267	positive	unconstrained
random	@(trialID,Hendrik Rietman).row	ID	1	0.5349	0.5349	5.072e-05	positive	unconstrained
random	@(trialID,Hendrik Rietman).col	ID	1	4.404	4.404	1.14e-05	positive	unconstrained
random	@(trialID,ILVOGm1705).row	ID	1	0.9378	0.9378	0.000127	positive	unconstrained
random	@(trialID,ILVOGm1805).row	ID	1	0.4937	0.4937	0.0003377	positive	unconstrained
random	@(trialID,ILVOGm1805).col	ID	1	1.306	1.306	3.243e-05	positive	unconstrained

©2020 Progeno BVBA

Figure 1: Screenshot of a multi-trial mixed model search performed on EUCLEG protein content observations.

1) RRBLUP: Ridge-Regression BLUP is a standard method in the field of genomic prediction. This technique is stable, computationally efficient and generally results in a prediction accuracy that equals or is slightly under that of the best performing method.
2) GBLUP: Also a standard method that has been shown to be mathematically equivalent to RRBLUP. This approach is, however, considerably more demanding in terms of computations but produces a reliability for each produced GEBV.
3) Q-GBLUP: A variant of GBLUP that allows to include the effects of one or more molecular markers that are associated to QTLs of the trait of interest. Standard GBLUP assumes that each QTL only has a small effect on the trait of interest. In the presence of QTLs with large effects, this assumption does not hold resulting in an underestimation of the QTL effects and reduced prediction accuracy.
4) Random Forest: A prediction technique from the domain of machine learning. Contrary to the other methods, Random forest does not rely on a linear model but uses a set (i.e. a forest) of decision trees to make predictions. The method is very powerful in terms of prediction accuracy but has considerable computational requirements. Several parameters (e.g. number of trees, tree size, …) require manual optimisation which can be a time-consuming activity.

The Genomic prediction module allows to train one or more genomic prediction models and validate their prediction accuracy by means of cross-validation routines. The resulting models can be stored, reused and shared with other users of the system. The trained models can be fed with the molecular marker profiles of new selection candidates to produce their GEBVs and rank them accordingly.

The Progeno Analytics module has been extended with other toolboxes to further capitalize on the collected phenotypic and genotypic data in the EUCLEG project. The Imputation module provides various routines for replacing missing values in molecular marker matrices which is crucial for training genomic prediction models. The GWAS (Genome-Wide Association Study) module, developed as part of WP4, allows to detect associations between markers and phenotypes which can then be used to improve genomic prediction accuracy by means of the Q-GBLUP approach. The Population Analysis module allows to detect and visualize possible population structures in the genotyped accessions which is in turn crucial knowledge for obtaining reliable GWAS results.

II. User interface

Statistical data analysis is a multidisciplinary field combining mathematics, statistics, computer science and knowledge of the domain, quantitative genetics in this case. The R scripting language has become a preferred tool for domain practitioners to analyse their datasets with up-to-date analysis techniques. Most R libraries and functions require little background knowledge concerning the mathematics and statistics that drive the analysis routines while at the same time the source code is readily available, allowing user to take a deeper dive into the technicalities when desired.

The analysis routines that have been implemented under Task 5.1 have been mainly developed in R. Some internal routines have been developed in C++ for computational efficiency but these can also be called from R. EUCLEG members that are knowledgeable in R can integrate the aforementioned EBV/GEBV estimation routines in their own data analysis scripts and workflows.

The Progeno Analytics module provides a graphical, web-based user interface that has been built on top of the R library. This point-and-click interface allows users that are not familiar to scripting in R to analyse their datasets with the exact same routines. As Progeno is a web application, users do not need to install any software on their own computer. All analysis jobs are executed on the powerful EUCLEG server that also centralises all collected data sources. Users can switch back and forth between the graphical user interface and the R interface within their analysis workflows, combining the best of both worlds.

III. Manual

Both the R routines and the graphical interface to the analysis routines of the Progeno Analytics module have been extensively documented. The documentation of the R functions are available from within R and provide detailed specification of the function arguments and output. The complete function manual is also available as PDF at https://www.progeno.net/EUCLEG/PGSP.pdf.

The graphical interface to the analysis routines have been documented by means of a dedicated tutorial. This 35 page document provides a detailed description of all the point-and-click actions that are needed to perform the various analysis types in the Progeno Analytics module. Most example datasets that are used in the tutorial have been extracted from the EUCLEG Progeno database. In fact, the querying of the data in the database is part of the tutorial itself. The document has been incrementally improved by incorporating the comments and suggestions of several consortium scientists who carried out all the tasks and exercises described in the tutorial. The final version was made available on the EUCLEG collaborative platform and https://www.progeno.net/EUCLEG/ProgenoAnalyticsTutorial.pdf.

IV. Training

Trainings for statisticians and EUCLEG partners involved in data analysis were initially scheduled to be organized in Ghent (Belgium) in April 2020. Because of COVID-19 outbreak, these trainings were cancelled. A remote training session was organized on May 20 and May 26 of 2020. The participants were asked to train themselves with the tutorial before the sessions. There were 30 participants in total, equally distributed over two training days. Each training day consisted of four 1,5 hour sessions in which the participants were demonstrated how to perform the various analyses that are relevant to the EUCLEG project. They made hands-on exercises and the obtained results were interpreted and discussed in group. The sessions were intentionally interactive allowing participants to ask questions or suggest improvements on the available analysis routines and user interface.