

METHODOLOGY ARTICLE

Open Access



Advantages of continuous genotype values over genotype classes for GWAS in higher polyploids: a comparative study in hexaploid chrysanthemum

Fabian Grandke^{1,2*} , Priyanka Singh^{1,3}, Henri C. M. Heuven^{1,4}, Jorn R. de Haan¹ and Dirk Metzler²

Abstract

Background: Association studies are an essential part of modern plant breeding, but are limited for polyploid crops. The increased number of possible genotype classes complicates the differentiation between them. Available methods are limited with respect to the ploidy level or data producing technologies. While genotype classification is an established noise reduction step in diploids, it gains complexity with increasing ploidy levels. Eventually, the errors produced by misclassifications exceed the benefits of genotype classes. Alternatively, continuous genotype values can be used for association analysis in higher polyploids. We associated continuous genotypes to three different traits and compared the results to the output of the genotype caller SuperMASSA. Linear, Bayesian and partial least squares regression were applied, to determine if the use of continuous genotypes is limited to a specific method. A disease, a flowering and a growth trait with h^2 of 0.51, 0.78 and 0.91 were associated with a hexaploid chrysanthemum genotypes. The data set consisted of 55,825 probes and 228 samples.

Results: We were able to detect associating probes using continuous genotypes for multiple traits, using different regression methods. The identified probe sets were overlapping, but not identical between the methods. Bayesian regression was the most restrictive method, resulting in ten probes for one trait and none for the others. Linear and partial least squares regression led to numerous associating probes. Association based on genotype classes resulted in similar values, but missed several significant probes. A simulation study was used to successfully validate the number of associating markers.

Conclusions: Association of various phenotypic traits with continuous genotypes is successful with both uni- and multivariate regression methods. Genotype calling does not improve the association and shows no advantages in this study. Instead, use of continuous genotypes simplifies the analysis, saves computational time and results more potential markers.

Keywords: Association study, Polyploids, Linear regression, Bayz, Partial least squares, Continuous genotypes

Background

Many agriculturally and horticulturally important crops are polyploid [1, 2]. Polyploids have multiple sets of chromosomes and have arisen by extensive genomic alteration and genome duplication [3, 4]. Diploidization, the differentiation of duplicated loci, converts most polyploids

back to diploids on the long term [5, 6]. The phenomenon of polyploidy results in complex genomic architecture in many flowering plants thus complicates genomics-based breeding [7]. Research in polyploids is also limited by the available methods and technologies [8]. Most bioinformatic tools have been developed for diploids and cannot be applied to higher ploidy levels. Recently, several methods have been developed to overcome this limitation, but most of them are restricted to tetraploids [9, 10]. Association studies aim to determine a genetic origin for a

*Correspondence: grandke@biologie.uni-muenchen.de

¹Genetwister Technologies B.V., Wageningen, The Netherlands

²Fakultät für Biologie, University of Munich (LMU), Munich, Germany

Full list of author information is available at the end of the article

phenotypic trait [11]. While phenotyping is independent of the ploidy level, genotyping has been identified as a bottleneck in breeding of polyploid crops [12]. Genotyping describes the process of determining an organism's genotype [13] and is known to be erroneous [14]. It involves the extraction of genetic material, molecular biological processes and the assignment of genotypic classes.

The latter one is also referred to as genotype calling and is a challenging task for polyploids. While there are many methods available for diploids [15], only three open access tools have been developed for polyploids, namely fitTetra, beadarrayMSV and SuperMASSA [12, 16, 17]. The former two are restricted to tetraploids and optimized for data sets originating from Illumina GoldenGate™ and Illumina BeadArray™, respectively. Subsequently, they underperform for data originating from other technologies [16]. SuperMASSA is a web tool that requires upload of individual data files for each SNP resulting in a poor performance. Further, the source code is not available and the algorithm cannot be validated. Density-based spatial clustering algorithms like OPTICS and DBSCAN [18, 19], were successfully applied in hexaploid wheat [20]. Preliminary analysis showed that they did not succeed for our genotypes, because the data points do not segregate into clusters, which can be distinguished based on density (Fig. 1).

Generally, determination of genotype classes (genotype calling) reduces noise and leads to better associations. In polyploids, this task is not as straightforward and the advantage of noise reduction is reversed by the high risk of misclassification, i.e. assignment of wrong genotypes to samples. Therefore, we skipped genotype calling and used the continuous genotypic values (compare Eq. 6) directly. The aim of this study was to use these continuous values to detect probes associating to three traits in hexaploid chrysanthemum and compare the results to genotype classes to evaluate the advantage of our

approach. The traits have been selected to represent distinctive types (disease resistance, flowering and growth) and heritabilities (0.51, 0.78 and 0.92). We applied linear regression (LR), bayz (Bayesian regression) [21, 22] partial least squares regression (PLSR) [23, 24] and compared the results to avoid methodological bias. We showed that the assignment of genotype classes would not improve our findings, but lead to misclassification. In this article we demonstrate that we are able to use continuous genotype values to detect associations with three different traits in hexaploid chrysanthemum.

Results and discussion

We applied LR, bayz and PLSR to identify significant probes associated with the disease, flower and growth traits (Additional file 1). Later, we compared sets of significant probes identified by the above mentioned methods. Figure 2 gives an overview about total numbers of associated probes per method and overlap between them. We repeated the LR analysis with genotype calls by SuperMASSA and compared the results. Further, we simulated datasets with the same properties of our real dataset to determine the expected number of significant markers.

Disease trait

LR detected ten significant probes ($q\text{-value} \leq 0.01$) when we used the continuous values. We called the genotype classes with SuperMASSA and repeated the LR analysis leading to 2 significant probes. We compared the results to see which approach worked better. An example is shown in Fig. 3(a-c). The axes in A represent the raw values of the two alleles. Three genotypes (red squares, blue circles and green triangles) were identified and assigned to the samples. The three lines represent the expected angles for each cluster center. The clusters identified by SuperMASSA (colors) do not match the groups that are indicated by the shape of the scatter plot. We consider the

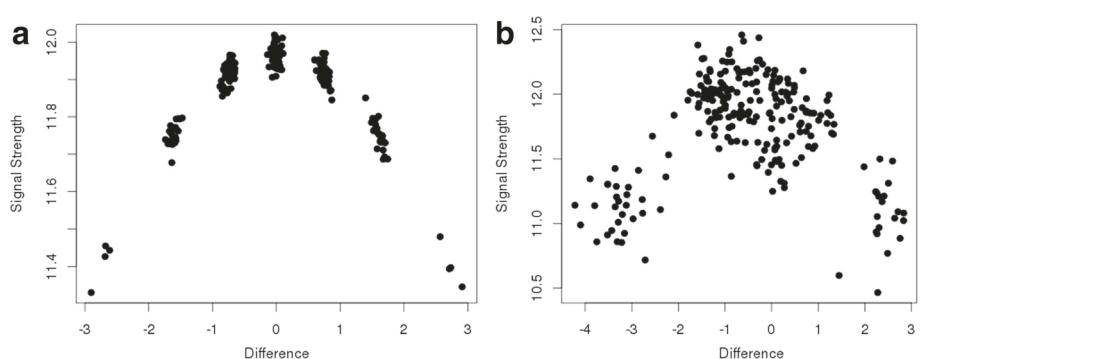


Fig. 1 Example probe. Example of genotype values for a hexaploid probe and 228 samples. The x-axis shows the difference between the signals of the two alleles. The y-axis shows the average signal strength per sample. The left and right sides show simulated and real data, respectively. **a** The simulation demonstrates how the seven genotype classes cluster into groups. **b** The real data shows the full segregation over the whole spectrum, but no clustering into seven genotype classes

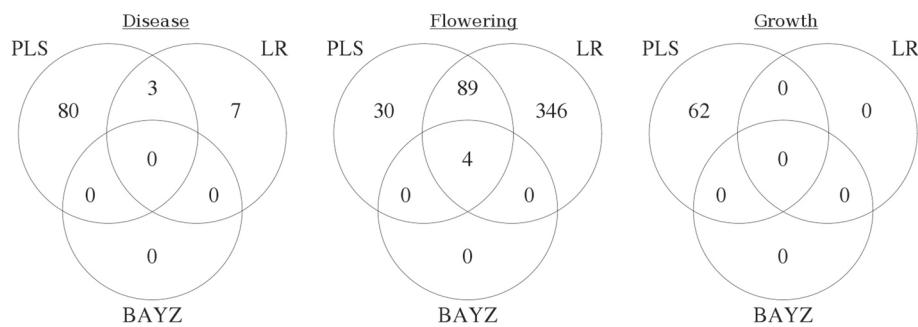


Fig. 2 Result comparison. Venn diagrams of significant probes for the disease (a), flowering (b) and growth (c) trait. The significance thresholds for LR(q -value), bayz (BF) and PLS(VIM score) were ≤ 0.01 , ≥ 10 and ≥ 2 , respectively

genotype calling as failed in that case. The blue cluster ranges over two groups, while the green cluster consists of outliers of the blue cluster. LR of the genotype class values with the phenotypes results in a p-value of 0.222 (Fig. 3c). Hence, the probe would be classified as non-significant, although it has not been corrected for multiple testing. The p- and q -values of the LR of the continuous genotypes for the same probe were 9.97×10^{-7} and 0.0078, respectively (Fig. 3b).

Additional file 2 shows a comparison of the significance values by SuperMASSA and the continuous values. Taken together, genotype calling distorted genotypic INFORMATION for some markers and the prevented their correct association. The desired noise reduction, which improves associations, could not be achieved. PLSR detected 83 probes with variance importance (VIM) scores ≥ 2 . Three of them overlap with the LR results. bayz did not find any significant probes with a Bayes factor (BF) threshold over 10. Even with a less-strict threshold of 5 there were no findings (compare Additional file 3).

Flowering trait

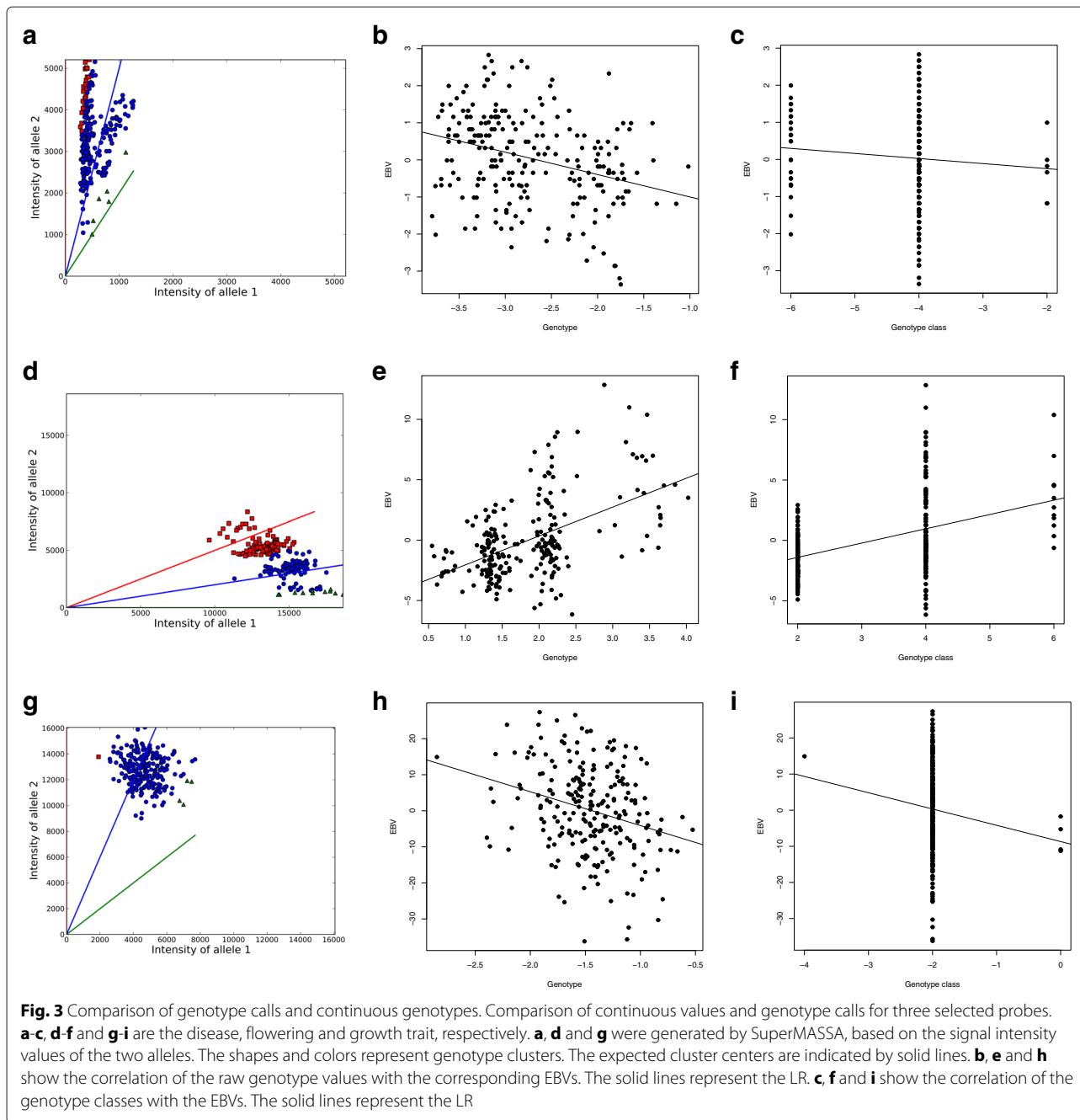
LR detected 439 significant probes for the flowering trait with the continuous values. SuperMASSA genotype calls led to 332 significant probes (Additional file 4). Again, the continuous genotypes generally lead to lower q -values than the genotype calls (compare Additional file 2). We simulated the experiment with varying numbers of significant probes (2-10) 100 times each. That way, we could observe how many significant probes we would detect if the true genotypes are known. The simulation results are shown in Additional file 5 and show that we expect around 100-2000 significant probes. Hence, our association results are in the correct magnitude. For our dataset continuous genotype values are advantageous over genotype classes because we obtain more significant probes and are less likely to miss trait related probes. However, our method is not too insensitive and does not result in thousands of false positive markers. In fact, the simulation

study detected even more false positives. It reduces the number of potential candidate probes from 55,825 to 439.

An example probe for the flowering trait is shown in Fig. 3d-f. SuperMASSA identified three different genotype classes. Continuous genotype values indicate four genotype classes, roughly centered at 0.75, 1.5, 2.2 and 3.5 (Fig. 3e). In contrast, the genotype classes by SuperMASSA combine the first two clusters (Fig. 3d). The blue one contains samples of the first and third cluster, which leads to its spread over the whole DEBV range (Fig. 3f). This leads to a lower p-value, but the probe is still highly significant. Nevertheless, in other cases this difference might determine whether the null hypothesis can be rejected or not.

Figure 4 shows the detailed distribution of both probes and contigs for the 439 probes. 206 probes were duplicates, i.e. one codes for the forward and one for the reverse strand of 103 SNPs. The significance of both probes adds to the probability of the SNPs association. Accordingly, the 233 remaining probes code for unique SNPs. There are multiple scenarios when only one probe is selected. First, there is an additional SNP within the primer sequence of the failing probe and the hybridization is disturbed. Secondly, the other probe had a similar signal, but was above the significance threshold and filtered out. Thirdly, the SNP is not associated to the trait and the probe was selected erroneously. Lastly, the other probe has been filtered out in a preprocessing step, based on the segregation range of the θ values.

The probes were selected so that not more than three SNPs (six probes) lay on one contig. Thus, it is unlikely to detect two or more significant SNPs from the same contig by chance. The fact that we found up to three SNPs from the same contig is a good indicator for real association. We detected 5 probes from one contig, coding for three different SNPs. The white area in the second bar in Fig. 4 shows 28 single probes, which code for different markers, but are located on the same contig. Accordingly, the missing second probe is not required to proof association.



The PLSR and bayz associations detected 123 and 4 significant probes, respectively. The four probes, detected by bayz overlap with the LR and PLSR results. In addition there is an overlap of 89 probes between PLSR and LR. However, the associating probes have low R^2 values and explain only parts of the phenotypic variance. Table 1 shows the four probes, which have been identified by all three methods. The scores do not correlate because the methods base on different approaches. The R^2 values are obtained from the LR association.

Growth trait

LR did not result in any significant probes for the growth trait. Growth is known to be a polygenic trait [25]. Hence, we did not expect single probes to show strong association. The association with the SuperMASSA genotype calls did not output any significant probes, either (Additional file 6). An example probe is shown in Fig. 3g. All but five samples were assigned the same genotype class. We expect monomorphic probes, but in Fig. 3h we see that the genotype values span a large portion of the negative θ range. Thus, we expect multiple genotype

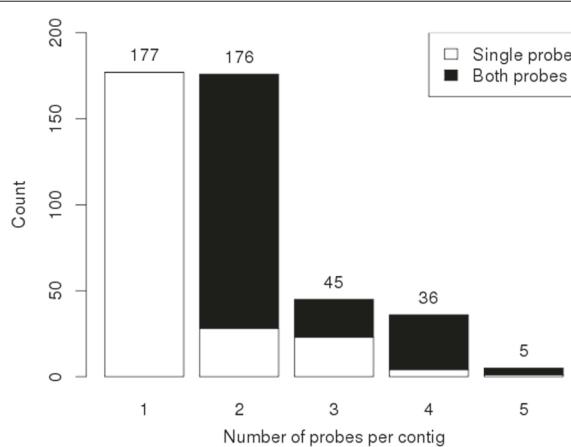


Fig. 4 Overview of contigs and probes for the flowering trait. Probe and contig distribution of the significant markers of the flowering trait. The five bars represent the number of markers that lay in the same contig. The colors distinguish between markers where both or only one of the probes were significant and are more reliable. Multiple markers from the same contig indicate its association to the trait

classes within that cloud of data points. However, they could neither be determined manually nor computationally. bayz did not detect any significant probes, either. In contrast, the PLSR association detected 62 significant probes. It includes low effect markers as well and is not limited to single loci. For polygenic traits PLSR is therefore advantageous.

General

We could have applied sparse partial least squares (SPLS) regression to deal with very high-dimensional data. The expected number of associating probes would be even higher than. We decided against it because the number of false positives and low impact probes (based on explained variance from the LR analysis) would have increased significantly. PLSR is less restrictive in the growing and disease trait but not in the flowering trait.

EBVs are established metrics in breeding, but have been criticized for genome-wide association studies (GWAS) because their naïve usage reduces power, increases the false positive rate and misestimates effect sized of quantitative trait loci (QTL) [26]. Hence, we used deregressed EBVs to account for fixed effects and repeated measurements, as described by Garrick et al. [27].

Table 1 Overlapping significant markers between all three methods for the flowering trait

Marker	R ²	q-value	BF	VIM
AX-89300609	0.31	1.72 × 10 ⁻¹⁵	11.03	3.23
AX-89215144	0.26	8.90 × 10 ⁻¹³	15.49	2.90
AX-89213862	0.18	1.74 × 10 ⁻⁸	24.82	4.12
AX-89256548	0.09	5.16 × 10 ⁻⁶	12.27	3.45

SuperMASSA misclassifies genotypes in some cases because it assumes clusters of equal distance at fixed positions. Instead, the data does not necessarily segregate into clusters (Fig. 3g). If clusters can be identified, they are not always at fixed positions, as assumed by SuperMASSA (Fig. 3d). Further, it does not account for outliers and assigns genotype classes to every sample. This results in clusters of only one sample in some cases and does not represent the genotype. Taken together, SuperMASSA does not cluster the data points properly for all probes. It seems to be optimized for data produced with two technologies and therefore performs less well on other data sets. The same situation was described by Voorrips et al. [16], when they compared fitTetra and beadarrayMSV. Thus, genotype calling has no advantage for polyploid data generated with the Affymetrix Axiom™ technology, as long as no method works properly. This underpins our preliminary analysis, which showed that the resolution of the signal intensities is not large enough to distinguish between the increased number of genotype classes expected in hexaploids (Fig. 1). To this end, association of the continuous genotypes is currently the best method.

Genotyping by sequencing (GBS) results in similar datasets, where the continuous genotypes are replaced by read counts. In general, our approach can be applied to this kind of data, but it is limited to bi-allelic SNPs and requires an extension to account for multi-allelic SNPs. Assuming an additive genetic effect [28] and a multi-allelic SNP, θ needs to be upgraded to θ_i :

$$\theta_i = \log_2(G_i) - \sum_{j=1, j \neq i}^4 \log_2(G_j) \quad (1)$$

where G is the set of alleles and G_i is the read count of allele i . Alternatives to the log-transformations might be more effective and need to be investigated [17].

The higher the ploidy level, the smaller is the advantage of genotype classes over the continuous values. With increasing numbers of genotype classes, the effect of noise reduction declines. For instance, in a diploid we expect the clusters AA, AB and BB at around 2, 0 and -2, respectively. A value of 1.2 would be assigned to cluster AA, so we correct for a large proportion of the signal. For any ploidy level n we expect up to $2n + 1$ clusters on a similar range, because the overall signal strength is limited by the used technology (e.g. amount of genetic material, GBS read depth). Consequently the distances between clusters decrease and the correction accounts for smaller proportions of the signal. In addition, the risk of misclassifications increases, because there is less tolerance for variation in the signal intensity or clusters overlap. Further, the distribution of genotype values approximates a continuous distribution with increasing ploidy levels. Figure 1 shows an example of a simulated hexaploid marker and one from the real data set. We were not able to identify the genotype classes for all samples in that case, because the clusters are indistinguishable. Nevertheless, the data points spread over the whole range of θ and provide genotypic information. From a biological perspective, continuous genotypes are difficult to interpret, because the number of alleles is discrete and should fall in one of the genotype classes. One explanation are tri- and tetra-allelic SNPs, where more than two nucleotides are present at the same position [29, 30]. If they are measured with bi-allelic technology (e.g. genotyping arrays), the sum of the two allele counts does not necessarily add up to the expected number (ploidy level). Alternatively, we might observe fractionation, the deletions in sub-genomes of allopolyploids [31]. Both result in data points outside of the expected clusters. For the association we mean-centered the genotype values.

Skipping genotype calling leads to further challenges with current linkage mapping and haplotype phasing methods, because they require genotype classes. Nevertheless, the choice of tools that work for polyploids is very limited anyways and new solutions need to be developed. Further, low-coverage sequencing and imputations of genotypes add more difficulties [32].

Conclusions

We showed that continuous genotype data can be used successfully in an association study of a polyploid crop and validated our findings in a simulation study. Application of different regression tools show that our approach is not limited to a specific method, but the results vary to a large extend.

Genotype calling leads to misclassification and false association results in some cases, where significant markers could not be detected. However, the majority of markers lead to similar results with genotype classes and continuous values, indicating that genotype calling is not adversely in general. In this study genotype calling has no advantage and can be skipped unless better methods are developed. Instead, use of continuous genotypes simplifies the analysis, saves computational time and results more potential markers. Nevertheless, the overlapping clusters of the given data set remain a challenge and the use of continuous genotypes is a successful solution to that problem.

Methods

A hexaploid chrysanthemum population consisting of 228 F1 offspring was used for our study. The cultivated plant material was provided by Dümmen Orange and all experiments have been performed according to legal guidelines.

Phenotypes

Three different traits have been used, as shown in Table 2. They represent the three areas that are relevant in a horticultural crop association study: disease, growth and flowering. Further, they span a wide range of heritability values. Details about the traits are not provided, because they are confidential and not important for the methodology itself. All traits' distributions are bell shaped and can be approximated by a normal distribution. The replicated measurements have been transformed into deregressed estimated breeding values (DEBV).

The estimation of the breeding values (EBVs) was performed using ASReml-R [33]. The EBVs for the individuals were derived by fitting a mixed linear model using the REML (residual maximum likelihood) procedure (Additional file 7). The asreml model was fitted to optimally use the information available for each individual, while simultaneously adjusting for environmental effect i.e. block and plate numbers. The mixed model for calculation of EBVs can be presented as

$$y_i = \alpha + \beta_i^1 + \beta_i^2 + g_i + e_i \quad (2)$$

Where y_i is the observed trait value, α is the population mean, β_i^1 is the fixed block effect, β_i^2 is the fixed plate number effect, g_i is the random accession effect, where

Table 2 Overview of traits (DEBV)

Trait	h^2	Mean	SD	Range	Samples
Disease	0.51	0.05	1.2	6.19	228
Flowering	0.78	0.00	3.3	19.01	228
Growth	0.92	0.21	12.2	63.57	228

$g \sim N(0, \sigma_g^2)$ and e_i is the random error of the observed trait value, where $e \sim N(0, \sigma_e^2)$. In order to calculate DEBV s as described by Garrick et al. [27], the predictive error variance (PEV) was calculated from the model parameters. Here, we have used variances of EBVs as a measure for PEV. The DEBV s were calculated using

$$debv_i = \frac{ebv_i}{r_i^2} \quad (3)$$

With

$$r_i^2 = 1 - \frac{PEV_i}{\sigma_g^2} \quad (4)$$

and

$$PEV_i = \text{var}(ebv_i) \quad (5)$$

where r_i^2 is the reliability of the EBV of plant i and σ_g^2 is the additive genetic variance.

Genotypes

The genotypes were measured with a customized Affymetrix Axiom™ microarray. It provides $\sim 100k$ probes for hexaploid chrysanthemum. We filtered out probes with a θ range below 2, because association requires segregation. The final data set consists of 55,825 probes. Each SNP is represented by two probes, upstream and downstream, respectively. We genotyped 228 samples and preprocessed them with Affymetrix Power Tools [34]. This includes quantile normalization and transformation of the microarray measurements. The genotype calling step from Affymetrix Power Tools was not performed, because it is limited to diploids and cannot detect more than three clusters. The microarray provides one value for each of the two alleles for every probe. The two measurements A and B are transformed into difference values θ , where

$$\theta = \log_2(B) - \log_2(A) \quad (6)$$

and a signal strength s where

$$s = \frac{\log_2(A) + \log_2(B)}{2} \quad (7)$$

An example of a bi-allelic probe from a hexaploid chrysanthemum data set is shown in Fig. 1. The x-axis represents θ , the difference between the two alleles A and B. The values span the whole range of potential genotypes and represent seven different genotype classes. The leftmost samples are homozygous A, while the rightmost ones are homozygous B. The intermediates are heterozygous in varying proportions. The y-axis shows the mean signal strength s . The homozygous s values are lower, because logarithmic values are used.

The genotype calling was done with the web application of SuperMASSA without population-level information (<http://statgen.esalq.usp.br/SuperMASSA/>) [12]. The ploidy range was set from 2 to 6; the other parameters were used with the default parameters. We used the raw values of the two alleles as input. The resulting genotypes represented the numbers of the two alleles. For the association we used the difference between the counts of the first and second allele.

Association methods

Three different methods to associate the continuous genotypic values with the phenotypes were used: LR, bayz and partial least squares regression (PLSR). The model to calculate the LR for all three traits was

$$Y_i = \alpha + \beta x_i + \epsilon_i \quad (8)$$

where Y_i are the DEBV s, α is the population mean, β is the regression coefficient, x_i the mean-centered, continuous genotype value and ϵ_i the residual error. The function lm from the R package stats (Version 3.1.3) with the default parameters was used for the regression [35]. The resulting p-values were transformed into q -values with the function qvalue of the R-package qvalue (Version 1.43.0) with default parameters [36, 37]. We applied a threshold of 0.01 to select the significantly associating probes.

The effect of each SNP was estimated using Bayesian Variable Selection method as implemented in the bayz software [21, 22] and described by Schurink et al. [38]. The applied method is similar to the BayesC π method [39], except the prior of π was changed from a uniform(0,1) distribution to a slightly informative prior distribution $\sim \text{Beta}(10, 1)$. In bayz, shrinkage of allele effects was done by applying a mixture distribution. Many SNP effects were shrunk to nearly zero to obtain high sparsity in SNP effects and only a small part of the SNP effects were less severely shrunken, thereby identifying SNPs with important associations. The prior mixture distribution was

$$a_k \begin{cases} N(0, \sigma_{g0}^2), \text{with probability } \pi_0 \\ N(0, \sigma_{g1}^2), \text{with probability } \pi_1 = (1 - \pi_0) \end{cases} \quad (9)$$

Where the ‘null’ distribution modeled the majority of SNP with (virtually) no effect using prior settings $\pi_0 = 0.99$ and $\sigma_{g0}^2 = 0.001$. The second distribution modeled SNPs with large effects where prior settings were $\pi_1 = 0.01$ and $\sigma_{g1}^2 = 0.1$. Variances of the mixture distribution and other model effects were estimated using a uniform prior and sampled with a Monte Carlo Markov Chain (MCMC) using Gibbs sampling. The MCMC was

run for 50,000 iterations with a burn-in of 10,000 iterations and a thin-interval of 200. A Bernoulli distribution specified probabilities for a SNP belonging to the 'null' or second distribution and proportions for the mixture were set to have a slightly informative prior distribution $\sim \text{Beta}(10, 1)$.

For the PLSR analysis we used the most significant probes of the LR analysis, based on F-statistic values ≥ 4 . The numbers of probes were 4517, 4546 and 4957 for the disease, flowering and growth trait, respectively. We used the *pls* functions of the R package caret (version 6.0-47) [40]. The association was accomplished in three steps. First, the data was split into a calibration (80 %) and a test set (20 %). Second, the calibration set was used to select the optimal latent variables (LV). We repeated a 10-fold cross validation 20 times and assessed 1-10 LVs based on the lowest root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\gamma_i - \hat{\gamma}_i)^2} \quad (10)$$

Where n is the number of samples, γ_i is the observed and $\hat{\gamma}_i$ is the predicted phenotypic value. In the third step the model was build and the significant probes were predicted based on their variable importance measurement (VIM) scores with a lower threshold of 2 [41, 42]. The scores were determined with the *varImp* function from the R-package caret [40]. The calculation is based on the weighted sums of the absolute regression coefficients.

Simulation

We simulated 228 F1 offspring genotypes (55,825 probes on 18 chromosomes) based on the parental genotypes with PedigreeSim [43]. We selected 2 to 10 associating probes randomly and calculated phenotypic values Y_i for each offspring i with an adapted formular by Günter et al. [44]

$$Y_i = \sqrt{1 - \sum_j \pi_j} * N(0, 1) + \sum_j a_{ij} \sqrt{\frac{\pi_j}{6 * f_j(1 - f_j)}}, \quad (11)$$

where π_j is the explained variance ($\sum_j \pi_j$ is the heritability), f_j is the allele frequency and a_{ij} is the genotype of sample i for probe j . The heritability was set to 0.78 as for the flowering trait. Afterwards, we associated the simulated phenotypes with the genotypes using LR. This simulation procedure was repeated 100 times for each parameter combination.

Additional files

Additional file 1: Association scores. Tab-separated data file with q -values (LR), VIM-scores (PLS) and BFs (bayz) for all probes for all three traits. (TSV 4970 kb)

Additional file 2: SuperMASSA results. PDF including the plots of the significant calls of the LR analysis. (PDF 17 kb)

Additional file 3: Genotype calling Disease. PDF including 10 images of genotype calls produced with SuperMASSA. (PDF 735 kb)

Additional file 4: Genotype calling Flowering. PDF including 11 images of genotype calls produced with SuperMASSA. (PDF 773 kb)

Additional file 5: Simulation results. PDF including the simulation results. (PDF 33 kb)

Additional file 6: Genotype calling Growth. PDF including 12 images of genotype calls produced with SuperMASSA. (PDF 960 kb)

Additional file 7: Estimated Breeding Values. Tab-separated data file of the deregressed estimated breeding values for the Disease, Flowering and Growth trait for all 228 samples. (TSV 15 kb)

Abbreviations

BF, Bayes factor; PLSR, partial least squares regression; DEBV, deregressed estimated breeding value; GBS, genotyping by sequencing; LR, linear regression; MCMC, Monte Carlo Markov Chain; PEV, predictive error variance; REML, residual maximum likelihood; SNP, single nucleotide polymorphism; SPLS, sparse partial least squares

Acknowledgements

This project (INTERCROSSING) has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 289974. The APC fee for this article has been funded by the EC FP7 Post-Grant Open Access Pilot. We thank Dümmen Orange for providing the data set and Pascal Duenk for assistance with the DEBV's and bayz. We acknowledge Luc Janssen for the suggestion of an appropriate model for handling continuous genotypes and the critical comments by two anonymous reviewers on an earlier version of the manuscript.

Availability of data and materials

The genotype data set supporting the results of this article is available in the Zenodo repository, <http://zenodo.org/record/46285>.

Authors' contributions

FG did the data preprocessing, the LR analysis, simulation study and drafted the manuscript. HH advised on the study design and helped to draft the manuscript. PS calculated the DEBV's and did the association with PLSR and bayz. JRH participated in the study design and coordination. DM reviewed the manuscript and designed the simulation study. All authors read and approved the final manuscript.

Authors' information

Some authors are employed at Genetwister Technologies B.V. and the data is owned by Dümmen Orange, a shareholder of the company.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

¹Genetwister Technologies B.V., Wageningen, The Netherlands. ²Fakultät für Biologie, University of Munich (LMU), Munich, Germany. ³Institute for Molecules and Materials (IMM), Radboud University, Nijmegen, The Netherlands. ⁴Animal Breeding and Genetics, Wageningen UR, Wageningen, The Netherlands.

Received: 19 February 2016 Accepted: 11 July 2016

Published online: 24 August 2016

References

- Soltis DE, Soltis PS, Tate JA. Advances in the study of polyploidy since plant speciation. *New Phytologist*. 2003;161(1):173–91.
- Leitch AR, Leitch IJ. Genomic plasticity and the diversity of polyploid plants. *Science*. 2008;320(5875):481–3.
- Acquaah G. Principles of plant genetics and breeding. Malden: Wiley-Blackwell; 2012.
- Comai L. The advantages and disadvantages of being polyploid. *Nat Rev Genet*. 2005;6(11):836–46.
- Wang X, Shi X, Hao B, Ge S, Luo J. Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytologist*. 2005;165(3):937–46.
- Paterson AH, Bowers JE, Chapman BA. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A*. 2004;101(26):9903–8.
- Ramsey J, Schemske DW. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu Rev Ecol Syst*. 1998;29(1):467–501.
- Soltis DE, Buggs RJA, Doyle JJ, Soltis PS. What we still don't know about polyploidy. *Taxon*. 2010;59(5):1387–403.
- Dufresne F, Stift M, Vergilino R, Mable BK. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Mol Ecol*. 2014;23(1):40–69.
- Grandke F, Ranganathan S, Czech A, de Haan JR, Metzler D. Bioinformatic tools for polyploid crops. *J Agric Sci Technol B*. 2014;4:593–601.
- Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*. 2005;6(2):95–108.
- Serang O, Mollinari M, Garcia AAF. Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. *PLoS ONE*. 2012;7(2):30906.
- Syvänen AC. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet*. 2001;2(12):930–42.
- Pompanon F, Bonin A, Bellemain E, Taberlet P. Genotyping errors: causes, consequences and solutions. *Nat Rev Genet*. 2005;6(11):847–6.
- Lamy P, Grove J, Wiuf C. A review of software for microarray genotyping. *Human Genomics*. 2011;5(4):304–9. 21712191.
- Voorrips RE, Gort G, Vosman B. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics*. 2011;12(1):172.
- Gidskehaug L, Kent M, Hayes BJ, Lien S. Genotype calling and mapping of multisite variants using an atlantic salmon iSelect SNP array. *Bioinformatics*. 2011;27(3):303–10.
- Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: Ordering points to identify the clustering structure. In: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. Philadelphia: ACM Press; 1999. p. 49–60.
- Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland: AAAI Press; 1996. p. 226–31.
- Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L, Mastrangelo AM, Whan A, Stephen S, Barker G, Wieske R, Plieske J, International Wheat Genome Sequencing Consortium, Lillemo M, Mather D, Appels R, Dolferus R, Brown-Guedira G, Korol A, Akhunova AR, Feuillet C, Salse J, Morgante M, Pozniak C, Luo MC, Dvorak J, Morell M, Dubcovsky J, Ganap M, Tuberous R, Lawley C, Mikoulitch I, Cavanagh C, Edwards KJ, Hayden M, Akhunov E. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol J*. 2014;12(6):787–96.
- Heuven HCM, Janss LLG. Bayesian multi-QTL mapping for growth curve parameters. *BMC Proc*. 2010;4:12. 20380755.
- Kapell DN, Sorensen D, Su G, Janss LL, Ashworth CJ, Roehle R. Efficiency of genomic selection using Bayesian multi-marker models for traits selected to reflect a wide range of heritabilities and frequencies of detected quantitative traits loci in mice. *BMC Genet*. 2012;13(1):42.
- Jöreskog KG, Wold HOA. Systems under indirect observation: causality, structure, prediction. Amsterdam: North-Holland; 1982.
- Kuhn M, Johnson K. Linear regression and its cousins. In: Applied Predictive Modeling. New York: Springer; 2013. p. 112–21.
- Salas Fernandez MG, Becroft PW, Yin Y, Lübbertsdorf T. From dwarves to giants? Plant height manipulation for biomass yield. *Trends Plant Sci*. 2009;14(8):454–61. doi:10.1016/j.tplants.2009.06.005.
- Ekine CC, Rowe SJ, Bishop SC, de Koning D-J. Why breeding values estimated using familial data should not be used for genome-wide association studies. *G3: Genes|Genomes|Genetics*. 2013;4(2):341–7. 24362310.
- Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol*. 2009;41(1):55. 20043827.
- Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet*. 2008;4(2):1000008.
- Casci T. Population genetics: SNPs that come in threes. *Nat Rev Genet*. 2010;11(1):8–8.
- Phillips C, Amigo J, Carracedo A, Lareu MV. Tetra-allelic SNPs: informative forensic markers compiled from public whole-genome sequence data. *Forensic Sci Int Genet*. 2015;19:100–6. 26209763.
- Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M. Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics*. 2004;166(2):935–45. 15020478.
- Pasanici B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, Li H, Gupta N, Neale BM, Daly MJ, Sklar P, Sullivan PF, Bergen S, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Purcell SM, Haas DW, Liang L, Sunyaev S, Patterson N, de Bakker PIW, Reich D, Price AL. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet*. 2012;44(6):631–5.
- Gilmour AR, Thompson R, Cullis BR. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*. 1995;51(4):1440–50. doi:10.2307/2533274.
- Affymetrix Power Tools. 2015. http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx. Accessed 25 Jul 2015.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. <http://www.R-project.org/>. Accessed 13 Aug 2015.
- Storey JD. Qvalue: Q-value Estimation for False Discovery Rate Control. 2015. R package version 2.0.0. <http://qvalue.princeton.edu/>. Accessed 13 Aug 2015.
- Storey JD. The positive false discovery rate: a bayesian interpretation and the q-value. *Ann Stat*. 2003;31(6):2013–35. 3448445.
- Schurink A, Janss LL, Heuven HC. Bayesian Variable Selection to identify QTL affecting a simulated quantitative trait. *BMC Proc*. 2012;6(Suppl 2):8. doi:10.1186/1753-6561-6-S2-S8.
- Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*. 2011;12(1):186.
- Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A. Caret: Classification and Regression Training. 2012. R package version 5.15-044. <http://CRAN.R-project.org/package=caret>.
- Mehmood T, Liland KH, Snipen L, Sæbø S. A review of variable selection methods in partial least squares regression. *Chemometr Intell Lab Syst*. 2012;118:62–9.
- Stephen Milborrow: Notes on the earth package. 2015. <http://www.milbo.org/doc/earth-notes.pdf>. Accessed 13 Aug 2015.
- Voorrips RE, Maliepaard CA. The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics*. 2012;13(1):248. Accessed 30 Apr 2014.
- Günther T, Gawenda I, Schmid KJ. phenosim - A software to simulate phenotypes for testing in genome-wide association studies. *BMC Bioinformatics*. 2011;12(1):265. doi:10.1186/1471-2105-12-265. Accessed 2015-12-18.