Euphytica (2008) 161:293–300
DOI 10.1007/s10681-007-9519-y

Genetic research in a public–private research consortium: prospects for indirect use of Elite breeding germplasm in academic research

Ralph van Berloo · Sjaak van Heusden · Arnaud Bovy · Fien Meijer-Dekens · Pim Lindhout · Fred van Eeuwijk

Received: 22 February 2007 / Accepted: 17 July 2007 / Published online: 9 August 2007

Abstract The creation of a public–private research partnership between plant breeding industry and academia can be beneficial for all parties involved. Academic partners benefit from the material contributions by industry and a practically relevant research focus, while industry benefits from increased insights and methodology tailored to a relevant set of data. However, plant breeding industry is highly competitive and there are obvious limits to the data and material partners are willing and able to share. This will usually include current and historic released cultivated materials, but will very often not include the elite germplasm used in-house to create new cultivars. Especially for crops where hybrid cultivars dominate the market, parental lines of hybrid cultivars are considered core assets that are never provided to outside parties. However, this limitation often does not apply to DNA or genetic fingerprints of these parental lines. We developed a procedure to take advantage of elite breeding materials for the creation of new promising research populations, through indirect selection of parents. The procedure starts with the identification of a number of traits for further study based on the presence of marker-trait associations and a priori knowledge within the participating companies about promising traits for quality improvement. Next, regression-based multi-QTL models are fitted to hybrid cultivar data to identify QTLs. Fingerprint data of parental lines of a limited number of specific hybrids are then used to predict parental phenotypes using the multi-QTL model fitted on hybrid data. The specific hybrids spanned the whole of the sensory space adequately. Finally, a choice of parental lines is made based on the QTL model predictions and new promising line combinations are identified. Breeding industry is then asked to create and provide progeny of these line combinations for further research. This approach will be illustrated with a case study in tomato.

Keywords Public–private partnership · Quantitative genetics · Research populations · Marker-assisted phenotype prediction · Product quality · Tomato · *Solanum lycopersicon*

R. van Berloo (✉) · S. van Heusden · F. Meijer-Dekens · P. Lindhout · F. van Eeuwijk
Plant Breeding, Wageningen University, Wageningen, The Netherlands
e-mail: ralph.van-berloo@keygene.com

R. van Berloo
Keygene NV, Wageningen, The Netherlands

A. Bovy
Plant Research International, Wageningen, The Netherlands

P. Lindhout
De Ruiter Vegetable seeds CV, Bergschenhoek, The Netherlands

F. van Eeuwijk
Biometris, Wageningen, The Netherlands

Introduction

In recent years, performing research in the framework of a public–private partnership has gained popularity. Obvious reasons are reductions in academic research budgets and a shift in focus of academia towards their stakeholders. In plant sciences, collaborations between industry and universities and research institutes have become more intense, but there are limits to the amount of information and resources consortium partners are willing to share. This may appear to hinder fruitful research progress, but solutions that circumvent sharing limitations can be thought of and applied.

In The Netherlands, the Centre of Biosystems Genomics (CBSG: see www.cbsg.nl for details on the goals and consortium partners of this centre) was established by the Dutch government as a 'centre of excellence'. Members of the CBSG consortium are Dutch universities and research institutes and over ten plants breeding companies active in the crop species potato and tomato. Both crops are important for Dutch agriculture and breeding, and were therefore selected as focal crops of the CBSG. One of the aims of the CBSG is to perform integrated research focusing on several aspects of environmental and consumer quality for these crop species. The involvement of research institutes allows efficient usage of specialized genomics techniques that are available at these institutes, while the involvement of industry facilitates usage of elite germplasm, together with the definition of clear research targets and a faster adoption of research results.

Within CBSG, the tomato quality research focuses on improvement of tomato as it is grown in the Dutch system of greenhouse cultivation with regard to consumer quality. Consumer quality in this respect deals with consumer decision strategies in the supermarket, but also includes sensory sensations related to the tasting of tomato fruits as well as metabolic measurements that were performed to capture soluble and volatile tomato fruit content. These aspects have received more attention from plant breeding industry lately, partly induced by negative publicity about poorly tasting tomatoes during the late 1990s. The CBSG tomato quality project is executed by a consortium of private and academic partners. This structure required that compromises would be found for issues related to use of input material and distribution and publication of results. Part of these compromises yielded an experimental setup that consisted of a screening of a diverse set of cultivated (mostly hybrid) tomato cultivars, as a phenotypic screening of the elite parental germplasm of commercial cultivars was not possible due to legal and practical considerations. In contrast to other crops like maize, where the use of elite germplasm has been described before, for instance in diversity studies (Melchinger et al. 1992; Lübberstedt et al. 2000), to our knowledge this is the first paper that reports on the use of industry derived elite inbred germplasm for academic research in greenhouse tomatoes. In this paper, we describe the approach we developed to use privately owned elite germplasm for the creation of new genetic resources that can be applied in academic plant breeding research.

This approach can be divided into three consecutive steps:

1. Identification of traits that show associations and identification of associated markers (QTLs) and their relative contributions to the phenotype.
2. Prediction of phenotypic values for (elite) inbred lines using the models identified under Step 1.
3. Choice of suitable (elite) parents and creation of new populations.

These steps are explained and discussed in detail in the methodology section.

Methodology and application within the tomato quality project

Data

Plant material and observations

During the first phase of the tomato quality project a set of 94 tomato cultivars was provided by the plant breeding companies from their collections of current and historic germplasm. The large majority of these cultivars were hybrids, but the set also contained a few non-hybrids, mainly old and no longer commercially grown cultivars. The set was selected, based on prior info and on expert judgment of the breeders, to contain a high amount of diversity with regard to many tomato fruit quality aspects. As all mayor

European vegetable breeding companies participated in the project we expect that the set will be representative for the diversity present among (European) greenhouse cultivated tomatoes. So the set contained both very 'good' as well as very 'bad' tomatoes with regard to taste, smell, appearance, etc. The set of tomato genotypes was grown in three replicated experiments, and DNA samples were extracted from each genotype. Fruit samples of ripe fruits (fruits that had just passed 'turning' stage) were gathered in each of the three experiments and used for a variety of assessments. These assessments consisted of organoleptic (sensory) evaluations by a trained panel of judges, quantitative scoring of metabolic fruit contents using liquid and gas chromatography combined with mass spectrometry (GCMS & LCMS; see Tikunov et al. 2005 for more details), consumer enquiries in which consumers were asked to judge tomato fruits by answering a questionnaire (Van den Heuvel et al. 2006), and several other morphological and commonly scored traits like fruit size, amount of soluble content (Brix) and fruit firmness.

Molecular markers

DNA samples of the cultivars were used to obtain genetic fingerprints. AFLP marker scoring with 50 primer-combinations was performed and yielded a total of 1,200 polymorphic markers. A large proportion of the AFLP markers could be scored co-dominantly, meaning that homozygous presence of a band could be distinguished from heterozygous presence and complete absence. Roughly 25% of the AFLP markers had a known position on a genetic linkage map. More details on the molecular markers scored in the set of hybrid tomato varieties are presented in R. van Berloo et al. (submitted) and Zhu et al. (in preparation) but, in summary, these authors concluded from the marker analyses that sufficient prospects for association mapping within the selected germplasm were present.

Results of earlier association studies

After 2 years of experimentation a complete dataset for 94 tomato cultivars was obtained and aggregated mean values for each characteristic, corrected for all non-genetic factors, were derived. These aggregated datasets were used for association mapping, which is reported in Van Berloo et al. (in preparation). The approach that was taken was similar to the one earlier applied in a set of barley cultivars for yield stability traits by Kraakman et al. (2004). Main conclusion from the association mapping was that for a number of important quality traits clear associations with genomic markers were observed. Type of tomato fruit (fruit size) strongly determined many phenotypic traits as well as marker band frequencies and created thereby a strong subpopulation effect in the association analyses. Overall marker-trait associations were high for many traits but many of these associations were due to genomic regions that differed between cherry tomatoes on the one hand, and beef/round tomatoes on the other hand. Therefore Van Berloo et al. also reported on associations studies using adjusted models that accounted for subpopulation structure, in which the tomato type was included as a covariable. These studies yielded more reliable genomic associations with quality traits, especially for a number of metabolic compounds. A study that distinguished additive and dominance effects revealed that dominance effects were negligible for nearly all of the observed associations.

Identification of parents for construction of promising offspring population

Step 1: Identification of traits and initial set of associated markers.

Single marker–single trait associations, corrected for subpopulation structure, helped to define the set of phenotypic traits to be used for further study. Traits were selected when they showed strong phenotypic variation between cultivars and when a number of significantly associated markers could be identified… The list of traits was supplemented with traits that were of special interest to breeding companies, although not yet showing clear associations with markers. A brief overview of selected traits is given in Table 1, while Fig. 1 shows the distribution of phenotypic observations that was observed over the 94 cultivars.

For subsequent analyses, we preferred to keep sensory and metabolite data apart. The hybrid by trait

Table 1 List of selected traits with genomic associations and the proportion of explained variation by a multi-QTL association model fitted to the whole of the set of 94 tomato cultivars and to each of three subgroups of tomato cultivars

Trait name	R^2	R^2_{beef}	R^2_{cherry}	R^2_{round}
1 Penten 3 one	0.73	0.30	0.43	0.57
3 Methylbutanol	0.70	0.44	0.68	0.74
Aftertaste sweet	0.80	0.16	0.65	0.23
Avg fruitweight	0.93	0.07	0.83	0.61
Brix	0.89	0.02	0.82	0.42
Citric acid	0.62	0.10	0.68	0.23
Fruit firmness	0.69	0.61	0.06	0.54
Glucose	0.67	0.26	0.55	0.20
Mouthfeel mealy	0.62	0.05	0.35	0.49
Phenylethanol	0.78	0.34	0.53	0.57
Scent spicy	0.57	0.19	0.57	0.35
Scent sweet	0.57	0.16	0.42	0.09
Taste earthy	0.45	0.08	0.15	0.39
Taste pungent	0.55	0.19	0.65	0.41
Taste spicy	0.88	0.31	0.73	0.61
Taste sweet	0.84	0.40	0.74	0.39
Taste unripe	0.59	0.42	0.12	0.48
Taste watery	0.86	0.57	0.76	0.67

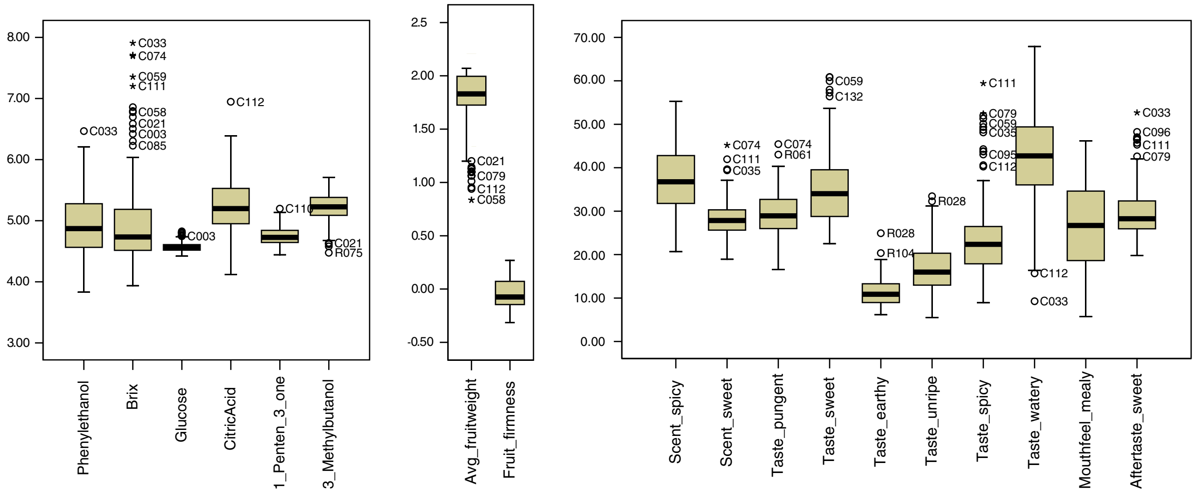

Fig. 1 *Box plots* describing phenotypic observations over 94 hybrid cultivars for selected traits. The graph is divided into three sections to allow clear representation of the different scales for different traits. First section contains log transformed metabolic observations, representing observed abundance of specific metabolic compounds. Second section contains fruit related observations. Third section contains sensory observations which were scored by a panel of judges on a scale between 0 and 100. Individual labels for outliers indicate cultivar code

sensory and volatile data matrices were then each subjected to a multivariate analysis in the form of a principal components analysis (PCA; see for instance Digby and Kempton 1987; Graffelman and van Eeuwijk 2005) to investigate the relationships between the hybrids, between the traits, and between hybrids and traits. Data were not standardized as within the set of sensory and metabolite data the traits were already expressed at the same scale. Results of the PCA were represented graphically in biplots, in which hybrids and traits were positioned in relation to their scores on the first two PCA axes. We used the biplots to identify sets of four hybrids belonging to a single company that spanned the phenotypic ranges for sensory and metabolite data adequately, i.e. whose convex hulls included the total set of hybrids to an acceptable degree for both types of data. The coverage of the sensory range was deemed more important than coverage of the metabolite range. Figure 2 gives the biplot for the sensory data (63% of the variation is represented in the biplot), and Fig. 3 shows the biplot for metabolic observations (80% of the variation is represented). The squares in these plots indicate the cultivars, and the dark squares indicate a set of four cultivars, stemming from a single company, which gave a good coverage of the multivariate sensory range. The coverage of the metabolite space was of less quality than the coverage of the sensory space. Our graphical procedure for choosing four hybrids contains some trial and error aspects. However, the method is based on a well defined multivariate dimension reduction technique whose results are visually presented. There may appear to exist some arbitrariness in the interpretation of the biplots, but in general different observers will arrive at comparable selections of hybrids.

All hybrid cultivars that were selected in this way were highly heterozygous, so any segregating offspring population derived from these cultivars themselves would show a complex pattern of

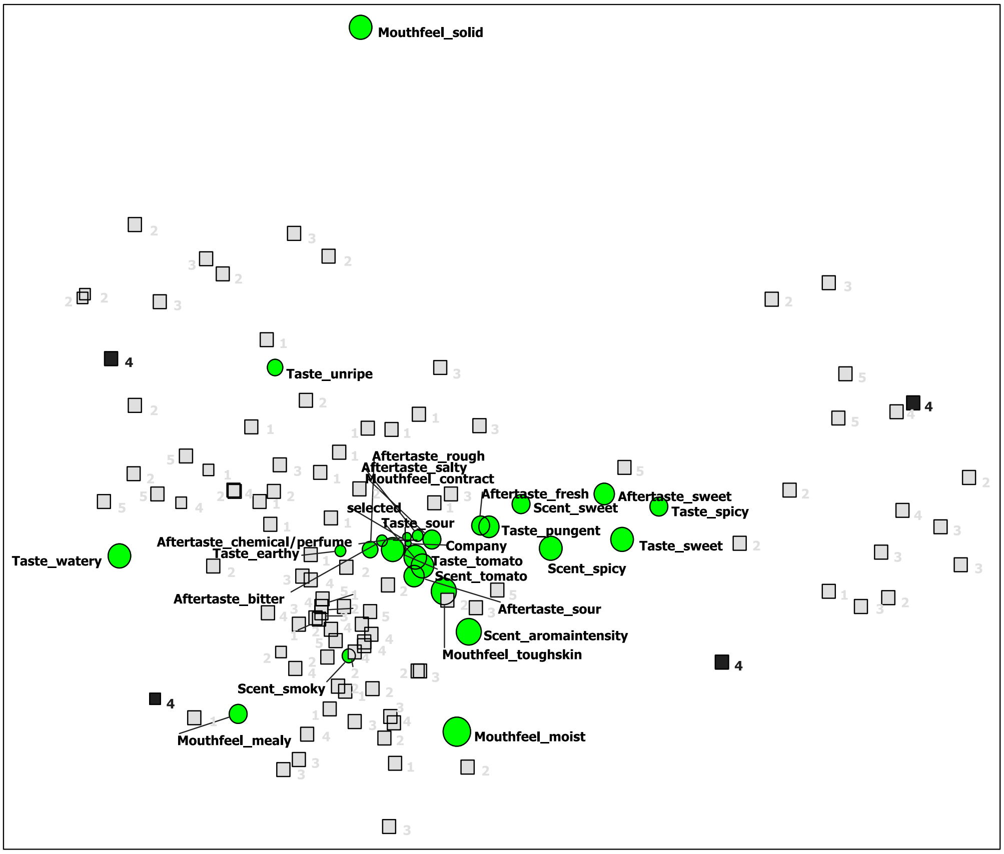

Fig. 2 *Biplot* showing the result of a PCA analysis for sensory traits. *Circles* indicate the positions of the traits. *Square* indicators represent the hybrid genotypes. *Genotype labels* indicate the company that provided the cultivar. The most contrasting hybrids with regard to sensory traits within the subset of hybrids originating from the same company were selected from this kind of plots. The *dark* coloured *squares* indicate the cultivars that were finally selected (see text). The amount of variation represented by this PCA plot is 63%

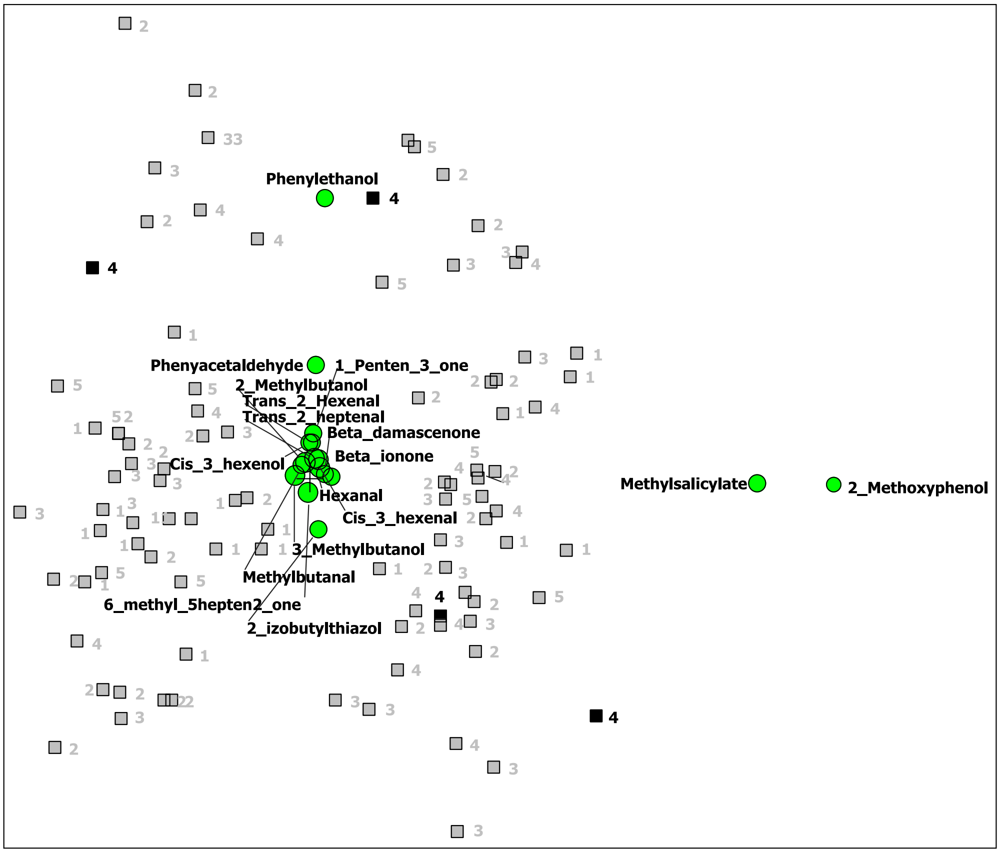

Fig. 3 *Biplot* showing the result of a PCA analysis for metabolic traits. *Circles* indicate the positions of the traits. *Square* indicators represent the hybrid genotypes. *Genotype labels* indicate the company that provided the genotype. The most contrasting hybrids with regard to sensory traits within the subset of hybrids originating from the same company were selected from this kind of plots. The *dark* coloured *squares* indicate the cultivars that were finally selected in this project. The amount of variation represented by this plot is 80%

inheritance, making analyses of these populations difficult and little representative. It was therefore decided to resort to the parental lines of the selected hybrids. These (elite) parental inbred lines had, however, not been phenotyped, as none of the breeding lines had been made available to the CBSG consortium. We therefore chose to produce marker-based phenotype predictions for the selected traits. Out of the total germplasm of the five participating companies, we selected three sets of hybrid cultivars that each by itself covered sufficiently well the phenotypic range with regard to the traits of interest. This procedure led to three equivalent sets of four cultivars. We retained three sets of cultivars, in order not to limit ourselves at this stage but to retain options for additional choices at a later stage, when comparing the three sets on subsequent criteria like viability of offspring, seed availability, etc. As each hybrid cultivar is the progeny of two parents, we needed to extend our parental evaluations and predictions to a maximum of 3 × 8 parental lines. Most promising parental inbred lines for further crossing and production of segregating populations were then chosen using predicted phenotypes from multi-QTL models, which is explained in detail in the next section.

Step 2: Prediction of phenotypic values for (elite) inbred lines.

To identify the most promising parental lines, predicted phenotypic values for the parental lines were needed. We used the hybrid set to construct a predictive model for the parental lines. Linear regression models for individual traits were identified, using as the predictor set of variables selections from the set of mapped markers. The model for an individual trait, using only the hybrid information was:

$$Y_{\text{hybrids}} = \mu + \sum_{i=1\ldots n} \beta_i m_i + \varepsilon. \tag{1}$$

In this formula Y_{hybrids} is a trait response vector, μ represents the intercept term, the β_i's represent QTL effects of associated markers m_i, while ε is an error term. In Eq. 1, n QTLs contribute to the trait variation. Of course, n will vary between traits. No covariable was included for the cherry versus round/beef contrast, as the purpose of the model was just prediction of the response, and this contrast is most easily included by a marker representing exactly this contrast. In addition, for most traits substructure seemed to play no important role.

We had, for each phenotypic trait, a response vector of 94 values (hybrids) that could be regressed on potentially 304 markers. To identify a predictive model, we used a stepwise selection procedure. The criterion for inclusion of markers in the predictive model was $P = 0.05$, while for removal $P = 0.1$ was used. The model was run on a standard PC using SPSS 12 (SPSS 2003).

The selected 3 × 8 = 24 parental lines were genotyped using 16 AFLP primer combinations. From the identified predictive multi-QTL models, predictions for the phenotypic values of the parental lines were created by inserting the appropriate marker scores for the parental lines:

$$\hat{Y}_{\text{Lines}} = \hat{\mu} + \sum_{i=1\ldots n} \hat{\beta}_i m_i. \tag{2}$$

In formula 2, $\hat{\mu}$ represents the intercept while the $\hat{\beta}_i$'s represent the estimated QTL effects of the QTLs associated with markers $m_{i.}$

The quality of the fitted regression models was good when judged by the proportion of explained variation in the hybrid data (Table 1). To give some idea about the influence of population substructure on the trait variation and the quality of hybrid prediction from markers, we also show the proportions explained variation within the three groups of tomatoes. It is remarkable that for whichever trait, there is always at least one group of hybrids that showed the same quality of fit as the total of the three groups of hybrids together. This means that the problem of substructure is relatively small, the multi-QTL models for marker-trait association were never due to exclusively the contrast between cherry and beef/round.

We are aware of the fact that the followed approach may lead to a too optimistic estimate of parental performance and that the figures in Table 1 are in fact upwardly biased. However this is then true for the overall quality of the model fit and predictions. We do not think that the most essential part of those predictions, the ranking of parent lines, was compromised by a possible over-optimism following from high values for correlations between predictions and observations for the hybrids.

Step 3: Choice of suitable parents from the set of (elite) inbred lines.

Predicted phenotypes for the parental lines were created in the previous step. Subsequently, PCA was performed on groups of selected traits (sensory, fruit content and metabolites), and the results were represented in biplots to allow a direct multivariate assessment which was then used to choose a subset of lines spanning maximum phenotypic range, i.e. a large convex hull. Selection of parental lines was performed in a similar way as described earlier for the hybrid phenotypes. From each set of parental lines, i.e. from each of the three companies, four lines were selected. Finally, based on predicted phenotypic range, availability of germplasm and marker diversity, one set of four lines was selected for continuation of the tomato quality experiments. The four finally selected parental lines were intercrossed in a half-diallel scheme, and F_1 genotypes were obtained. Each of the F_1 genotypes was then selfed to obtain six segregating F_2 populations. This approach, using a set of connected F_2 populations in which each parental line is present in three populations, was also used by Blanc et al. (2006, 2007) and is very similar to the approach employed by Paulo et al. (2007) in *Arabidopsis*. The chosen setup should result in a higher power for QTL analysis due to additional segregation of multiple alleles. More advanced statistical models will be necessary for the QTL analysis.

Our selected parental lines contained regular sized (round) tomatoes as well as small fruited (cherry) tomatoes. Therefore, several of the crosses between the parental lines are of the cherry-round type, which will yield segregation for fruit size as well as many other related traits of interest in the progeny. Genetic, phenotypic, metabolic and sensory evaluation of the resulting F_2 genotypes is currently being performed by two of the academic CBSG consortium partners; Wageningen University and Plant Research International.

Concluding remarks

Public private-partnerships in plant science can be advantageous for all parties involved: industry benefits from results that are more tailored to their research demands and the use of research materials that are closer to the materials they also use in-house, while academia benefits from increased funding and availability of the (vast) resources gathered by industry. We have described some of the compromises that are inherently required when entering such collaboration, and also presented an approach that allowed research to proceed in the desired direction. Although this is only an example of the use of molecular markers as 'information transfer tools' in plant breeding research, and we are also aware that our approach requires substantial commitment from all partners, we believe that in many more cases like the one described in this paper, molecular markers can be put to use to circumvent sharing limitations.

References

Blanc G, Charcosset A, Mangin B, Gallais A, Moreau L (2006) Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. Theor Appl Genet 113:206–224

Blanc G, Charcosset A, Gallais A, Moreau L (2007) QTL detection and marker assisted selection in a multiparental maize design. Euphytica (in press)

Digby PGN, Kempton RA (1987) Multivariate analysis of ecological communities. Chapman and Hall, London

Graffelman J, van Eeuwijk FA (2005) Calibration of multivariate scatter plots for exploratory analysis of relations within and between sets of variables in genomic research. Biom J 47:863–879

Kraakman ATW, Niks RE, Van den Berg PMMM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435–446

Lübberstedt T, Melchinger AE, Dußle C, Vuylsteke M, Kuiper M (2000) Relationships among early European maize inbreds: IV. Genetic diversity revealed with AFLP markers and comparison with RFLP, RAPD, and pedigree data. Crop Sci 40:783–791

Melchinger AE, Boppenmaier J, Dhillon DS, Pollmer WG, Herrmann RG (1992) Genetic diversity for RFLPs in European maize inbreds. Theor Appl Genet 84:672–681

Paulo MJ, Huang X, van Eeuwijk FA, Koornneef M (2007) Analysis of quantitative traits in an F4 population derived from crossing four accessions of *Arabidopsis thaliana*. Euphytica (in press)

SPSS (2003) SPSS for Windows, Rel. 12.0.1. SPSS Inc., Chicago

Tikunov Y, Lommen A, de Vos CHR, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiol 139:1125–1137

Van Berloo R, Zhu A, Gort G, Ursem R, Bovy A, van Heusden S, van Eeuwijk FA (in prep) Association mapping for fruit quality traits in a set of cultivated tomato hybrids

Van Berloo R, Zhu A, Verbakel H, Gort G, van Eeuwijk FA (under review) Diversity analysis in a large set of cultivated tomatoes. Submitted to Theor Appl Genet

Van den Heuvel T, van Trijp H, Gremmen B, Rennes RJ, van Woerkum C (2006) Why preferences change: beliefs become more salient through provided (genomics) information. Appetite 47:343–351

Zhu A, van Berloo R, Xiong H, van Eeuwijk FA (in prep) Haplo-block loci and alleles construction in a selected set of cultivated tomatoes with application in diversity studies