Theor Appl Genet (2010) 120:249–259
DOI 10.1007/s00122-009-1214-z

ORIGINAL PAPER

Enriched partial correlations in genome-wide gene expression profiles of hybrids (*A. thaliana*): a systems biological approach towards the molecular basis of heterosis

Sandra Andorf · Joachim Selbig · Thomas Altmann · Kathrin Poos · Hanna Witucka-Wall · Dirk Repsilber

Received: 1 April 2009 / Accepted: 30 October 2009 / Published online: 17 November 2009

Abstract Heterosis is a well-known phenomenon but the underlying molecular mechanisms are not yet established. To contribute to the understanding of heterosis at the molecular level, we analyzed genome-wide gene expression profile data of *Arabidopsis thaliana* in a systems biological approach. We used partial correlations to estimate the global interaction structure of regulatory networks. Our hypothesis states that heterosis comes with an increased number of partial correlations which we interpret as increased numbers of regulatory interactions leading to enlarged adaptability of the hybrids. This hypothesis is true for mid-parent heterosis for our dataset of gene expression in two homozygous parental lines and their reciprocal crosses. For the case of best-parent heterosis just one hybrid is significant regarding our hypothesis based on a resampling analysis. Summarizing, both metabolome and gene expression level of our illustrative dataset support our proposal of a systems biological approach towards a molecular basis of heterosis.

Communicated by F. van Eeuwijk.

Contribution to the special issue "Heterosis in Plants".

S. Andorf · D. Repsilber (✉)
Research Institute for the Biology of Farm Animals (FBN), Wilhelm-Stahl Allee 2, 18196 Dummerstorf, Germany
e-mail: repsilber@fbn-dummerstorf.de

J. Selbig · H. Witucka-Wall
University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany

T. Altmann
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr. 3, 06466 Gatersleben, Germany

K. Poos
University of Applied Sciences Gelsenkirchen, Site Recklinghausen, August-Schmidt-Ring 10, 45665 Recklinghausen, Germany

Introduction

The phenomenon of heterosis has already been known since the last century (Shull 1908). It was defined as "increased vigor, size, fruitfulness, speed of development, resistance to disease and to insect pests, or to climatic rigors of any kind, manifested by crossbred organisms compared with corresponding inbreds, as the specific results of unlikeness in the constitutions of the uniting parental gametes" by Shull (1952). This definition is restricted to describing the phenotypes that result when two different inbred lines are crossed. Therefore, it is often interpreted as not implying a genetic basis for heterosis (Lamkey and Edwards 1999). This was accomplished by Schnell and Cockerham (1992) defining heterosis as the difference in performance between hybrid and the mean of the two parents. Figure 1 displays such a quantitative genetics definition of heterosis. Mid-parent heterosis is the difference in phenotype value between the heterozygous offspring and the mean of the homozygous parents, while best-parent heterosis describes the situation where the hybrid exceeds the best parent.

Three different genetic models to explain heterosis have been suggested: dominance (Bruce 1910; Xiao et al. 1995), overdominance (Shull 1908; East 1936; Crow 1952) and epistasis (Schnell and Cockerham 1992; Li et al. 2001; Luo et al. 2001). These hypotheses can be divided into approaches based on dominance or overdominance and global approaches (epistasis) (for review see Lamkey and Edwards 1999 and Birchler et al. 2003). Towards a

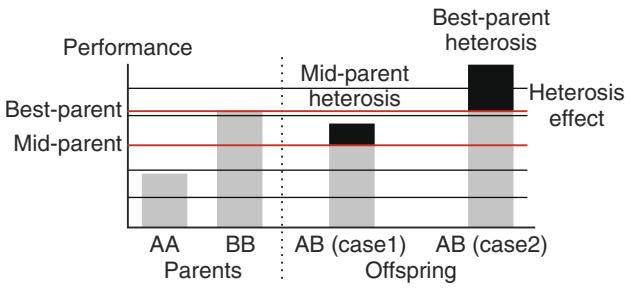

Fig. 1 Quantitative genetics definition of heterosis. The *black parts* of the performance of the heterozygous offspring denote the heterosis effect

molecular basis of heterosis, it has been analyzed which genes show the above genetic non-additivity in their expression levels (Vuylsteke et al. 2005), and if such genes are enriched in yield-related QTL (Wei et al. 2009). However, a molecular mechanistic model, which would be able to explain how the observed phenomena on the molecular level are integrated to result in heterosis on the phenotype level, is still lacking.

In our contribution, we use a systems biology approach to analyze heterosis in *Arabidopsis thaliana* plants based on patterns in genome-wide gene expression profiles. Already Robertson and Reeve (1952) suggested that heterozygotes are likely to posses greater biochemical versatility by carrying a greater diversity of alleles. Additional alleles at heterozygous loci may lead to additional regulatory interactions in the molecular network. Equipped with an enlarged repertoire of regulatory possibilities, hybrids may possibly be able to correctly respond to a higher number of environmental challenges leading to higher adaptability (individual acclimation ability) and, thus, the heterosis phenomenon.

Nowadays, high-throughput techniques, such as microarrays, allow measuring genome-wide feature profiles simultaneously. In global approaches, these datasets can be used to discover the interactions of molecules, how they are organized in networks and how the different networks are linked to each other (Barabási and Oltvai 2004). Partial correlations have been recommended to estimate regulatory interactions from observational data (Werhli et al. 2006).

Simple network models have been proposed to model the regulatory apparatus in a parsimonious way (Shubik 1996; Somogyi and Sniegoski 1996; Genoud and Métraux 1999). On this background we developed our "network hypothesis of heterosis" (Andorf et al. 2009). Our conceptual modeling results proposed that higher adaptability comes with an increased number of molecular interactions. To characterize the global interaction structure of regulatory networks, we use partial correlations (association networks). Based on the hypothesis of Robertson and Reeve (1952) and our conceptual model, we expect that the heterozygous genotypes show enriched partial correlations compared to the homozygous parents. These larger partial correlations represent the additional regulatory interactions in the molecular networks of the hybrids. Also, a gene set enrichment analysis is included to check for pathway-specific enlarged partial correlations.

The hypothesis was already tested on a metabolite dataset of samples of *A. thaliana* plants (Andorf et al. 2009). In this paper we will check if the hypothesis also holds true for gene expression data of the same genotypes. We use a certainly limited dataset, but aim to propose and illustrate a systems biological view which allows for an integrated hypothesis about the molecular basis of heterosis, complementing single gene and quantitative genetics approaches.

Materials and methods

Experimental data and preprocessing

Gene expression data were measured using Agilent's *Arabidopsis thaliana* Microarray Kit 4x44k, P/N G2519F (Agilent Microarray Designs ID 021169, arrays contain four subarrays where each represents a different hybridization). To isolate the RNA the innuPREP Plant RNA Kit (845-KS-2060250, Analytik Jena) was used. The RNA was obtained from seedlings of *A. thaliana* of two homozygous lines C24 and Columbia (Col-0; depicted as Col in the following) and the reciprocal crosses C24 × Col and Col × C24. Gene expression profiles were measured during early development at seven time points [4, 6, 10, 15, 20, 25 and 30 days after sowing (DAS)]. For each measurement, a group of seedlings (Petri dish, pot) was grown and fully harvested after every specific time of growing.

Figure 2 shows the experimental design, a multiple nested loop design. Each arrow represents one subarray, where the arrowhead symbolizes that the sample was labeled with one color and the root of the arrow symbolizes the other color. For each genotype–time point combination 2 or 4 biological replicates were measured. For the time points of 4, 10, 20 and 30 DAS, we had four replicates each. Part of the subarray that contains the samples of C24 at the time points 15 and 20 DAS (dashed arrow in Fig. 2) was covered by an air bubble and therefore, this subarray was excluded from the analysis.

Figure 3 summarizes the workflow of our analysis, beginning with the raw data from these microarray hybridizations.

During reading the raw data with the function *read.maimages* of the Bioconductor (Gentleman et al. 2004) R package *limma* (Smyth 2005), low quality spots were detected using eight quality features (see Table 1)

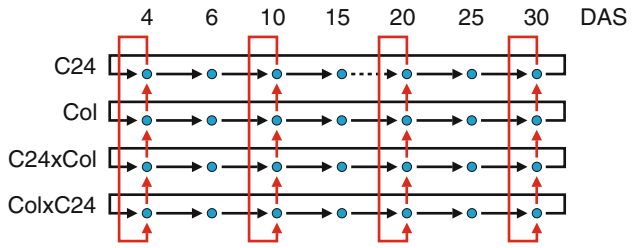

Fig. 2 Experimental design with multiple loops. Each *arrow* symbolizes one subarray (*dashed arrow* subarray was excluded from analysis)

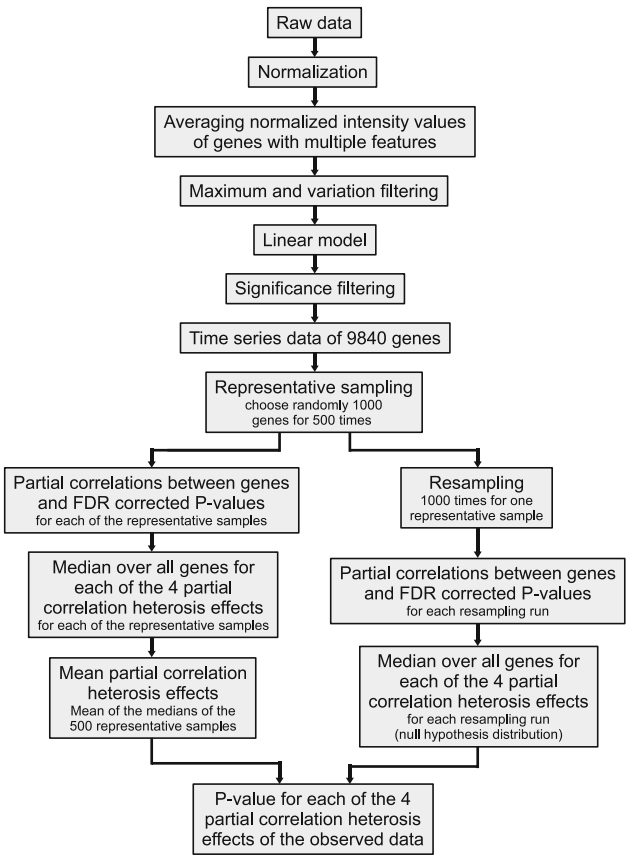

Fig. 3 Workflow of our analysis

described in the reference guide of the Agilent Feature Extraction Software (Agilent Technologies Inc. 2008). Afterwards, the raw intensities of the spots that were not flagged out, were background corrected using the method *normexp* (Ritchie et al. 2007) of the R package *limma*. Background corrected values were lowess normalized to get as unbiased red/green-ratios as possible. For global comparability, the data of all arrays were quantile normalized (Smyth and Speed 2003). For 3651 genes, more than 20% of the measured values were flagged out and therefore these genes were excluded from further analysis.

As proposed by Yang et al. (2002), normalized gene intensities were obtained as in Eqs. 1 and 2 from re-parameterizing the normalized log-ratios (M) and mean log-intensities (A) from the *limma* analysis.

Table 1 Eight Boolean variables related to outliers (Agilent Technologies Inc. 2008)

Green channel	Red channel
gIsFeatNonUnifOL	rIsFeatNonUnifOL
gIsBGNonUnifOL	rIsBGNonUnifOL
gIsFeatPopnOL	rIsFeatPopnOL
gIsBGPopnOL	rIsBGPopnOL

$$M = \log I_{Cy5} - \log I_{Cy3} \qquad A = \frac{1}{2}(\log I_{Cy5} + \log I_{Cy3}) \quad (1)$$

$$\log I_{Cy5} = A + \frac{1}{2}M \qquad \log I_{Cy3} = A - \frac{1}{2}M. \quad (2)$$

Regarding the locus IDs, 6,647 genes are represented by two or more spots on each subarray. The normalized intensity values for all measurements of these genes are replaced by the average of the values of the multiple spots. This leaves 33,445 genes for the further analysis.

Subsequently, profiles of non-expressed genes as well as approximately constant profiles were cleaned out applying two further filtering steps. In the first step, genes with very small maximum intensity values were screened out. Based on the distribution of the maximum intensity values of all genes (data not shown), we required a minimum intensity of $\log I \geq 7$ for at least one measurement of the gene. In the second step, genes were excluded from the subsequent analysis which showed low variation in their normalized intensities. In this filtering step we applied a cutoff of 2.8.

We used a linear model (adjusted to Kerr et al. 2000; Kerr and Churchill 2001) to analyze the experimental loop design and estimate the gene expression profiles. It contains the factors g denoting the four genotypes, factor $t \in \{1,..., 7\}$ denoting the seven time points of the developmental time series, their interaction $g \times t$, factor $AR \in \{1,..., 11\}$ denoting the array (containing the four subarrays) and factor $DcRNA \in \{1,..., 7\}$ denoting the date of cRNA synthesis. The fitting of the linear regression was done on a gene-wise basis for the following model where $y_{i,j,k,l,m}$ depicts the normalized gene intensities.

$$y_{i,j,k,l,m} = \mu + g_i + t_j + (g \times t)_{i,j} + AR_k + DcRNA_l + \varepsilon_{i,j,k,l,m}. \quad (3)$$

In this model, μ gives the overall gene-wise mean, the four genotypes are denoted with index i, the seven time points with index j, the array with index k, the date of cRNA synthesis with index l and the replicates with index m (between 1 and 4 biological replicates; see Fig. 2 for details). A factor *dye* was not included in this model

because it was not significant. Estimated gene expression values, $y^*_{i,j}$ were obtained from model 3 as in Eq. 4

$$y^*_{i,j} = g_i + t_j + (g \times t)_{i,j}. \quad (4)$$

Afterwards, we applied an additional filtering step on the estimated effects of the linear model. In this significance filter we filtered out genes that do not show a significant time and/or genotype–time interaction effect. We corrected the *P* values for these effects using the FDR correction described by Benjamini and Hochberg (1995). We choose a liberal cutoff of 0.2 as significance level to only exclude genes which show nearly no time dependency or $g \times t$ interaction. After this filtering step, 9,840 genes remained for all further analyses, a number inline with our expectations from earlier expression studies in *A. thaliana* (Ma and Sun 2005).

The analyses were performed using R (R Development Core Team 2008) (version 2.8.1) on an openSUSE Linux 11.0 (x86_64) server with 32GB RAM. Raw gene expression data, estimated profiles as well as scripts are available upon request.

Network statistics

Werhli et al. (2006) suggested that partial correlations of features of time series profiles can be used to study causal regulatory interactions. Simulation results for metabolite time series data confirmed this (Andorf et al. 2009). So, we based our investigation of additional regulatory interactions in hybrids on the estimation of regulatory interactions through partial correlations. To calculate partial correlations we employed the approach as proposed by Opgen-Rhein and Strimmer (2007). Their algorithm is implemented in the R package *GeneNet* (Opgen-Rhein et al. 2007). We used this package to obtain partial correlations from the normalized gene intensities of the seven time points. In *GeneNet*, partial correlations are calculated as in Eq. 5

$$\tilde{\rho}_{a,b} = \frac{-\omega_{a,b}}{\sqrt{\omega_{a,a}\omega_{b,b}}} \quad (5)$$

$\tilde{\rho}_{a,b}$ is the partial correlation between the genes a and b. $\omega_{a,b}$ is the element of the inverse covariance matrix. It is estimated using a shrinkage approach (Schäfer and Strimmer 2005b) within the package *GeneNet*. For the shrinkage estimator of the partial correlations we used the default option "static" in the method *ggm.estimate.pcor* of the package *GeneNet*. Because we do not include any a priori information about the partial correlations in the shrinkage process, the covariance matrix is shrunk towards the identity matrix. To demonstrate the validity of this estimation procedure for the dimension of our data we conducted a methodology simulation study.

1. Construction of covariance matrices with constant covariance values of 0.25 and 0.4 for 1,000 nodes (the diagonal values were set to unity).
2. Cholesky decomposition approach (Parrish et al. 2009) to simulate gene expression data out of these matrices for seven time points.
3. Calculation of the partial correlations using the R package *GeneNet*.
4. Calculation of the difference between the mean of the partial correlations from the 0.4 covariance matrix and the one with 0.25 values.
5. Repeat of (1)–(4) for 100 times. The difference between the mean of the partial correlations of both simulated gene expression data had the same order of magnitude as the differences between the mean of partial correlations of the homozygous and heterozygous genotypes in our experimental data.

The simulation study described above is capable of showing that we are able to use the shrinkage estimator of the partial correlations as implemented in the package *GeneNet* in our study in a valid way. The resulting histogram of the differences between the mean partial correlations calculated from the 0.4 and the 0.25 covariance matrix for each of the 100 repeats is shown in Fig. 4. In 77 cases from the 100 repeats, the mean difference of the simulated data was positive. In these cases, the stronger correlated data (0.4) lead to a detection of larger partial correlations in our simulation study. The means of all partial correlation values of the 100 repeats for the 0.25 and 0.4 covariance matrices, respectively, were 6.5×10^{-4} and 8.1×10^{-4}. For 1,000 randomly chosen genes of our experimental data, we calculated a mean of the means of the partial correlations for all four genotypes of 8.5×10^{-5}. Thus, we have shown that the shrinkage approach for the estimation of partial correlations by Schäfer and Strimmer (2005b) can be used for the dimension of 1,000 nodes and 7 time points as in our data. The power to identify enriched partial correlations in our simulation was 77%.

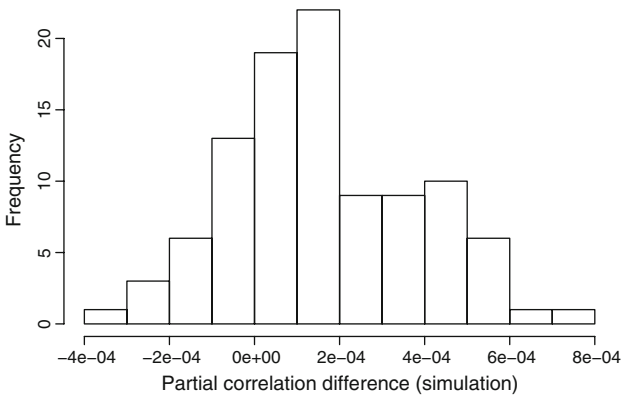

Fig. 4 Differences between the mean of the partial correlations calculated from the simulated 0.4 covariance matrix and the one with 0.25 values for each of the 100 repeats. A positive difference was detected for 77% of the repeats

Within *GeneNet*, two-sided *P* values for the test of non-zero correlation (null hypothesis: zero partial correlations) are calculated (Schäfer and Strimmer 2005a; Strimmer 2008). The *P* values were corrected using the FDR correction described by Benjamini and Hochberg (1995). Like Werhli et al. (2006), we are interested in roughly estimating which regulatory interactions exist. Therefore, regarding our hypothesis that heterozygous genotypes contain more regulatory interactions, our focus is on the number of existing regulatory interactions, estimated as significant partial correlations, and not on the value of each partial correlation itself. Hence, the further analysis of mid-parent and best-parent heterosis effects is based on *s*-values that are calculated like in Eq. 6

$$s_{d,f,u} = 1 - p^{FDR}_{d,f,u} \tag{6}$$

$p^{FDR}_{d,f,u}$ donates the FDR estimates according to Benjamini and Hochberg (1995) for the partial correlation between two genes ($f, u \in \{1,..., N\}$, N genes in the analysis) of genotype $d \in \{C24 \times C24,\ Col \times Col,\ C24 \times Col,\ Col \times C24\}$. Using *s*-values, we get a high value for regulatory interactions that are most probably present (low corrected *P* value) and low values for regulatory interactions that are probably not present in the regulatory network (high corrected *P* value).

To determine the partial correlation mid-parent heterosis effect (see Fig. 1) of each gene pair, we first calculated for each genotype separately for every single gene ($f \in \{1,...,N\}$) the mean value of the *s*-values of its pairwise partial correlations to all other genes:

$$h_{d,f} = \frac{1}{N-1} \sum_{u \in \{1,...,N\}, f \neq u} s_{d,f,u}. \tag{7}$$

Second, the mid-parent value for each gene was built out of the mean values calculated before for the homozygous genotypes:

$$h^{\text{mid-parent}}_{f} = \frac{1}{2} \sum_{v \in \{C24 \times C24, Col \times Col\}} h_{v,f}. \tag{8}$$

Finally, the partial correlation mid-parent heterosis effects were calculated as the difference between the mean values from Eq. 7 of either hybrid and the mid-parent values:

$$h^{\text{mid-heterosis}}_{w,f} = h_{w,f} - h^{\text{mid-parent}}_{f} \tag{9}$$

w denotes the respective heterozygous line ($w \in \{C24 \times Col, Col \times C24\}$).

Simultaneously, we calculated the partial correlation best-parent heterosis effect values (see Fig. 1). Here, instead of the mid-parent value, we determined the best-parent value (the maximum values of the mean values of the two homozygous genotypes; from Eq. 7):

$$h^{\text{best-parent}}_{f} = \max_{v \in \{C24 \times C24, Col \times Col\}} h_{v,f}. \tag{10}$$

Afterwards, the partial correlation best-parent heterosis effect values were calculated as the difference between the mean values from Eq. 7 of either heterozygous genotype and the best-parent values. w denotes again the heterozygous line ($w \in \{C24 \times Col, Col \times C24\}$):

$$h^{\text{best-heterosis}}_{w,f} = h_{w,f} - h^{\text{best-parent}}_{f}. \tag{11}$$

As calculating partial correlations involves large matrices and, hence, a lot of working memory, we were not able to analyze partial correlations for all 9,840 genes that remain after filtering. Instead, we selected representative samples of 1,000 randomly chosen genes. This is displayed in the left chain of the workflow in Fig. 3. To show that randomly selecting 1,000 genes indeed results in a representative sample, we selected 500 times randomly 1,000 genes and analyzed the variation of the features of interest. For each of the 500 repeats we calculated the partial correlation mid-parent and best-parent heterosis effects for either heterozygous genotype. For each of these four cases we determined the median of the calculated heterosis effect values. Thus, we got four median values for each of the 500 repeats; one median for each hybrid for the mid-parent as well as the best-parent heterosis effect. For each of the four cases, we then calculated the mean of the before determined 500 median values, $\bar{h}_r$ (r indexing the four cases), as well as the 95% confidence interval (2.5 and 97.5% quantiles). If these confidence intervals exclude the value of zero partial correlation heterosis effects, we would be confident both to be able to show a robust effect and that our sampling approach yields representative samples in our sense.

To determine the significance of the observed partial correlation heterosis effects, we resampled the data of one randomly drawn representative sample of 1,000 genes in such a way that the genotype origins of the data are randomly re-assigned (right chain in Fig. 3). For each gene in the set of 1,000, the estimated time profiles of the four genotypes were randomly re-assigned to the four genotypes (with replacement). This resampling was done 1,000 times. We calculated median values over the chosen 1,000 genes for each of the 1,000 resampling runs. This distribution of median partial correlation heterosis effects constitutes the null hypothesis distribution to establish a one-sided *P* value for the originally observed partial correlation heterosis effects:

$$p^{\#} = \#(h_{\text{resampled}} \geq \bar{h}_r)/1,000. \tag{12}$$

A gene set enrichment analysis was performed to investigate if genes that show large partial correlation heterosis effect values (Eqs. 9, 11) are particularly

enriched in single pathways. We used gene sets (based on locus IDs) for 79 pathways. 30 of them were based on a MapMan annotation file (Usadel et al. 2009; Thimm et al. 2004), which, in turn, is based on the TAIR database version 8 (Swarbreck et al. 2008). 49 gene sets were built upon Plant Ontology (PO) terms (The Plant Ontology Consortium 2002). Pathways that contained less than 10 or more than 4,000 of the genes we analyzed were excluded from this analysis, because too few genes in one pathway would make this pathway easily significant even if it just contains one or two genes with high partial correlation heterosis effect values. Too large pathways are not specific enough. The partial correlation heterosis effect values for each gene were determined using the first 100 representative samples of 1,000 randomly chosen genes each. Each time the mid-parent as well as best-parent heterosis effect values for either hybrid were saved and averaged. We got one partial correlation mid-parent and best-parent heterosis effect value per heterozygous genotype for each of our 9,840 genes. However, our gene set enrichment analysis was based on just 8,500 genes because for the other genes we did not have a locus ID and, thus, we could not assign them to the pathways. The median values of the mid-parent and best-parent heterosis effect values for either hybrid for all 8,500 genes are very close to the mean values shown in Fig. 5.

We performed our gene set enrichment analysis using the hypergeometric distribution according to Draghici et al. (2003) and Backes et al. (2007). This over-representation analysis measures enrichment by cross-classifying genes according to the membership in a functional category (gene set) and the membership in a selected list. We chose as selected list the 850 genes (10% of all genes in this analysis) that show the largest partial correlation mid-parent as well as best-parent heterosis effect for either heterozygous genotype. The resulting *P* values were corrected using the FDR correction described by Benjamini and Hochberg (1995).

Results

As proposed by Werhli et al. (2006), an increase in molecular interactions can be measured as increase in partial correlations (also shown in a simulation study in Andorf et al. 2009). Therefore, we investigated partial correlations according to Opgen-Rhein and Strimmer (2007) of our experimental data, to test our hypothesis that regulatory networks of hybrids show enriched molecular interactions compared to their parental homozygous genotypes. The investigation was based on *s*-values (Eq. 6) to determine how many molecular interactions are probably present in the different genotypes.

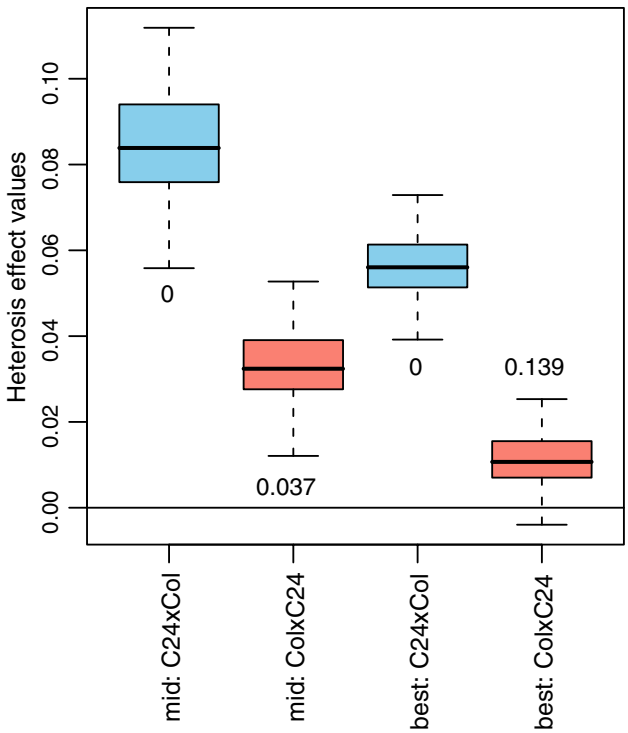

Fig. 5 Distribution of the the median values of the partial correlation heterosis effects of 500 repeated analysis of 1,000 randomly chosen genes. Mean values and 95% confidence intervals as well as the *P* values are given

For 500 different sets of 1,000 randomly chosen genes we calculated the partial correlation mid-parent heterosis effect values (Eq. 9) as well as the partial correlation best-parent heterosis effect values (Eq. 11). Figure 5 displays the distribution of the 500 median values for the partial correlation heterosis effect values from the 500 repeated measurements of 1,000 genes (representative samples). Furthermore, the mean value and the 95% confidence interval are shown for each case. For the heterozygous genotype C24 × Col, the 95% confidence intervals exclude the zero for the mid-parent as well as the best-parent partial correlation heterosis effect. The 95% confidence intervals for the genotype Col × C24 exclude the zero just for the mid-parent partial correlation heterosis effect values and not for the best-parent partial correlation heterosis effect values. These three cases for which the 95% confidence intervals exclude the zero show the effect of enrichment of partial correlations in the transcriptome of the heterozygous lines and, furthermore, we are confident that choosing 1,000 genes randomly out of 9,840 genes leads to representative samples in this sense. For the last case we cannot decide on this basis if selecting 1,000 genes randomly is not representative or if this genotype does not show a best-parent partial correlation heterosis effect. We also determined the significance of the partial correlation

heterosis effects of the observed data. This analysis was based on the resampling of one set of 1,000 randomly chosen genes. For the genotype C24 × Col we calculated a *P* value of zero for the mid-parent as well as the best-parent partial correlation heterosis effect. Hence, both effects are significant for this heterozygous genotype. For the other heterozygous genotype (Col × C24), only the mid-parent partial correlation heterosis effect is significant with a *P* value of 0.037. For the best-parent partial correlation heterosis effect of this genotype, we determined a *P* value of 0.139. Thus, the best-parent heterosis effect is not significant for the genotype Col × C24. The *P* values are also given in Fig. 5.

Figure 6 shows the partial correlation mid-parent as well as best-parent heterosis effect values for either heterozygous genotype for one set of 1,000 representative genes in detail. The histograms show that most of the partial correlation mid-parent heterosis effects for both heterozygous genotypes (C24 × Col: Fig. 6a; Col × C24: Fig. 6b) are positive. The shift to the right is not as big for the partial correlation best-parent heterosis effects (C24 × Col: Fig. 6c; Col × C24: Fig. 6d) as for the mid-parent heterosis effects but still noticeable.

In our gene set enrichment analysis we investigated if the partial correlation heterosis effects are enriched in some particular pathways. Table 2 shows the pathways that are enriched in either case of the partial correlation mid-parent and the best-parent heterosis effect for both hybrids.

Discussion

Our study aims at contributing to the understanding of heterosis at the molecular level by proposing a systems biological approach to analyze molecular profile data in hybrids. We estimated regulatory interactions between genes as partial correlations of their transcript profiles in a genome-wide approach for the early development of two homozygous *A. thaliana* lines and their reciprocal crosses. Results show a genome-wide global increase in the significance of partial correlations between transcript profiles in the hybrid lines as compared to the mid-parent as well as best-parent expectations. Moreover, in some functional groups of TAIR as well as PO terms, both hybrid lines show a particularly high partial correlation heterosis effect. These results confirm earlier findings on the metabolite level (Andorf et al. 2009) and provide further support to a *molecular network hypothesis of heterosis* which we developed as systems biological approach contributing to a better understanding of the molecular basis of heterosis.

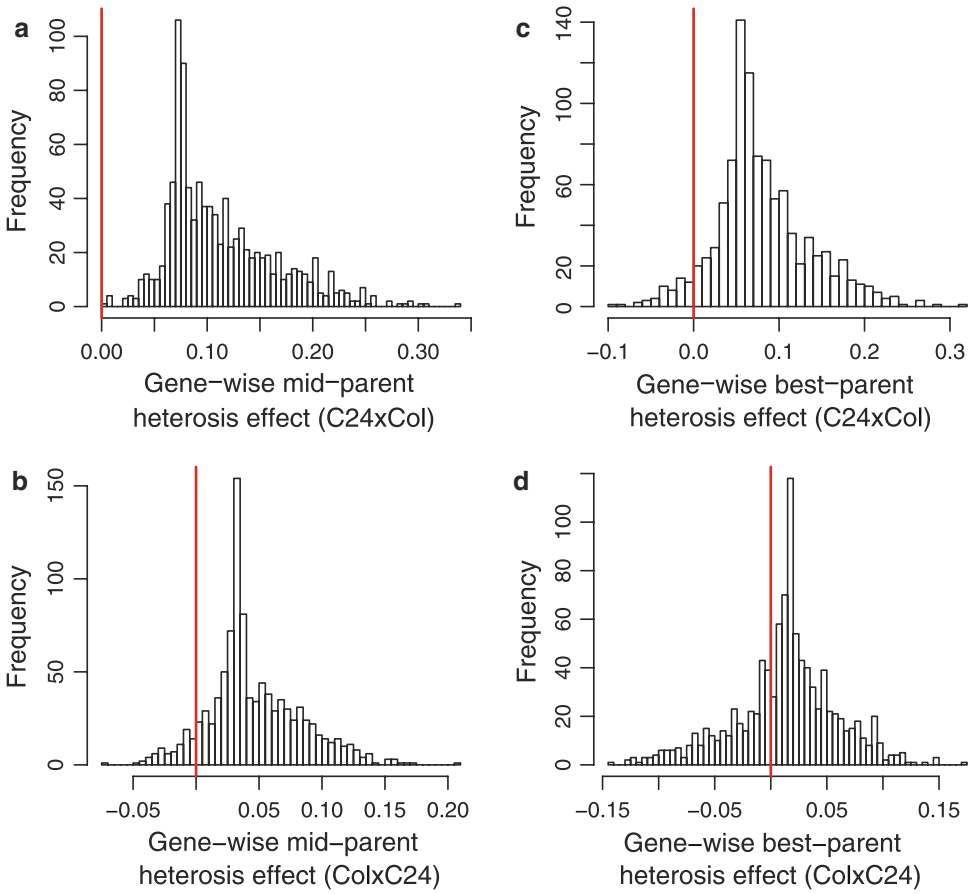

Fig. 6 Display of partial correlation mid-parent heterosis effects (see Eq. 9) as well as partial correlation best-parent heterosis effects (see Eq. 11) for one representative set of 1,000 genes. For both hybrids most of the genes show a larger significance of the partial correlations than the mid-parent values or best-parent values, respectively

Table 2 Results of gene set enrichment analysis for partial correlation heterosis effects

C24 × Col mid-parent	Col × C24 mid-parent	C24 × Col best-parent	Col × C24 best-parent
Male gametophyte	Male gametophyte	Male gametophyte	Male gametophyte
Stress	Lateral root primordium	Sperm cell	Lateral root primordium
Sperm cell	Transport	Stress	Sperm cell
Redox regulation	Sperm cell		
Photosynthesis	Primary root apical meristem		
	Lipid metabolism		
	Ovule		

Enriched pathways of TAIR and PO

Within the existing diversity of explanatory hypotheses towards a molecular basis for heterosis, our approach aims to investigate changes in regulatory interaction on a global level, rather than searching for single responsible loci. Existence of regulatory interactions on a global scale is estimated through significance of partial correlations. We therewith follow a line of argumentation taken as early as in the 1950s when Robertson and Reeve (1952) or Maynard Smith (1956) suggested that genetic heterozygosity might result in greater biochemical versatility in development and for reacting to environmental challenges. A larger repertoire in regulatory possibilities on the molecular level could result in the observed superior hybrid vigor. Also, recent discussions about possible molecular causes of heterosis include the notion of altered regulatory effects in hybrids and the positive effects of an enlarged repertoire of regulatory responses (Birchler et al. 2003; Song and Messing 2003). The emphasis of our study is on substantiating this hypothesis as to enable to experimentally measure the enlarged regulatory versatility in hybrids as global structures on the molecular level.

In an earlier contribution, we proposed a systems biological approach contributing to an understanding of heterosis at the molecular level which we termed "network hypothesis of heterosis" (Andorf et al. 2009). Taking a very simplistic parsimonious view, we considered the Boolean network approach, following Genoud and Métraux (1999), to demonstrate how the enhanced possibility to correctly respond to environmental challenges is linked to an enlarged number of regulatory interactions. These were estimated as significant partial correlations of metabolite profiles in the same design as for the current study at some earlier time points of development. Summarizing the earlier results with those of the current study, we were now able to show a global enrichment of the number/significance of the partial correlations in the hybrid lines on both metabolome and transcriptome level for our illustrative datasets.

Regulatory interactions can only be estimated from correlation structures of a regulatory network in such parts where ongoing regulatory processes lead to measurable changes in the respective molecular profiles. As both datasets concern the early development of *A. thaliana*, where Meyer et al. (2004) showed that the foundations of biomass heterosis are laid, it might be speculated that it is the nature of this biomass phenotype that it concerns a global adaptation process of the seedling. Later developmental stages and adaptation processes, such as flowering, fruit ripening or other more specific phenotypes may require more local, limited molecular responses, e.g., restricted to special pathways or gene regulatory modules. The current study as well as the results of Andorf et al. (2009) mostly show *global* changes in partial correlation structures, i.e., increase in estimated regulatory interactions.

However, as result of our gene set enrichment analysis several gene sets appeared to be specifically enriched. We hypothesize that these genes are among the subset of highly regulated genes during the specific developmental interval of our study. With the small-powered study design in mind, we do not want to speculate about biological interpretations of specific enriched gene sets.

In other species, such as *Drosophila* or mice, it became evident early in heterosis research that stress conditions were prone to cause pronounced heterosis effects (Harrison 1962; Maynard Smith 1956). A possible reason is that under such conditions the regulatory system is challenged to its limits. Hence, it is then necessary to make full use of the spectrum of regulatory possibilities. This may lead to inferior performance of the homozygous parental lines based on their limited regulatory possibilities when compared to their heterozygous offspring. In the setting of the current study, establishing a viable seedling under laboratory conditions, such as a climatic chamber opposed to the natural environment, may represent such an environmental challenge capable to show the enhanced potency of the hybrids' molecular regulatory repertoire.

When confronting hybrid genotypes with the environment, e.g., when recording performance in interesting environments for breeding and exploitation, *functional* data

such as gene expression or metabolite profiles allow an additional, deeper characterization of potentially advantageous crosses. For example, Thiemann et al. (2010) search for gene expression signals of single genes correlated with hybrid performance in maize and functionally study their candidates using GO terms. Frisch et al. (2010) follow an alternative strategy. Parental gene expression values are used to build a distance measure which is used to predict hybrid performance with a linear model. Further approaches exist to combine functional and genetic data for hybrid performance prediction (Steinfath et al. 2010). Hence, these studies might complement respective results from QTL studies. As Melchinger et al. (2007) found in their quantitative genetics study of *Arabidopsis* heterosis, QTL heterosis effects are to a large extent dependent on the whole genetic background. If a given genetic background is advantageous or not, certainly is dependent on the environment. This dependency is only accessible via *functional* tests. Several groups (Vuylsteke et al. 2005; Swanson-Wagner et al. 2006; Guo et al. 2006; Wei et al. 2009) investigated hybrids in comparison to their homozygous parents on the functional level, measuring genome-wide gene expression levels. In addition to their findings about proportions of realized modes of gene action in hybrids, our own contribution can be seen as proposing an idea for a systems biological heterosis analysis of the functional domain or gene expression level. We propose a hypothesis how molecular correlation structures specific for heterozygotes could be understood as mechanistic link between molecular and phenotypic manifestation of heterosis.

Considering regulatory interactions and possibilities to infer their global structure from molecular profile data, it is evident that a lot of existing regulatory interactions either involve molecular species which are not measured or act across different layers of the molecular regulatory apparatus profiled. Here, we adapt the view proposed by Somogyi and Sniegoski (1996), who emphasize the fact that the interactions deduced from molecular profiles of a specific level, e.g., metabolome or transcriptome, map regulatory processes of other molecular levels onto the one under consideration. Hence, the deduction of regulatory interactions for the specifically measured features may be wrong in detail, because the effects of molecules from other molecular levels are masked. This is especially so in the case of our study, as the number of time points sampled does not at all suffice to draw any strong conclusions on the level of a single estimated regulatory interaction. We think, however, that using our findings to build hypotheses about *global* structures of the molecular regulatory apparatus, such as an increased number of regulatory interactions in heterozygotes, is still allowed.

Partial correlations, also called *association networks*, are just one of several possibilities for estimating global regulatory interaction structures. The related so-called *relevance networks* (Butte et al. 2000), where Pearson correlations are measured to describe global correlation structures, are, however less eligible for our task. In contrast to partial correlations where indirect correlations are explicitly excluded, these remain an important factor when considering Pearson correlations. To emphasize this difference, it might be stated that when considering Pearson correlations it is save to talk about structures of *missing* correlations, whereas considering partial correlations reveals structures of *existing* correlations without being contaminated with indirect correlations. Werhli et al. (2006) recommended the use of partial correlations for the estimation of molecular interaction of regulatory networks from observational data, also contrasting it with a Bayesian network approach. In our study we follow this recommendation and use an algorithm proposed by Schäfer and Strimmer (2005b) which employs a shrinkage approach to estimate the partial correlations (R package *GeneNet*). Their approach is suitable for data with small sample size and large numbers of variables, as our genome-wide gene expression profiles. As we do not have any a priori information about the covariance structure of our transcriptome data, we chose the identity as canonical shrinkage target. Moreover, the time points of our time series data are not close enough in time to make additional use of the time series character—hence we chose the option "static" for application of the shrinkage estimator in the *GeneNet* package.

Time series data with only seven time points are a poor basis for investigating correlation structures of thousands of features. In our case we were concerned with nearly 10,000 gene expression profiles from which we chose a set of 1,000 genes as representative sample. However, we refrained from interpreting partial correlations for single pairs of features, as due to the shortness of our time series, we were not able to carry out a more accurate network reconstruction analysis. Instead, we were interested in the global structures down to the level of a set of coarse grained pathways. This way, we feel that this kind of investigation of overall structure is still valid. A medium scale number of false positives or negatives may not disturb this coarse grained analysis results.

Regarding the number of features analyzed, it is necessary to also discuss the feature selection, or filtering process, which was performed previously to partial correlation analysis. Our filtering procedure has been chosen such as to filter out gene expression profiles which were likely representing features not expressed or regulated during the time interval of early development in our experiment. We chose cutoffs for the filtering process such that around 10,000 genes remained for further analyses. This number matches what is expected from existing

studies regarding proportions of actively expressed genes in different tissues of *Arabidopsis* (Ma and Sun 2005).

Our dataset is also a compromise from another point of view. The plant tissue used for feature extraction (RNA as well as metabolite isolation procedures) was the complete young seedling. Hence, we assessed only the average of tissues constituting the seedling. Inferences about the regulatory structure are therefore possibly exclusively valid on the global scale we address, most likely not for many specific single features and their correlations.

Furthermore, we are aware of the fact that only a single cross is a poor basis to draw general conclusions.

Summarizing the methodological considerations, it remains to emphasize that the dataset of the current investigation could be analyzed with valid results only at the coarse grained global level. However, at this level, gene expression as well as metabolite profiles jointly pointed towards an increase in the significance of partial correlations. This, based on Werhli et al. (2006), we interpret as increase in number of interactions allowing for an increased adaptability to environmental challenges during early seedling development.

Future investigations should certainly involve multiple lines, multiple species, multiple time windows of development or different environmental challenges for homozygous parents and their hybrids to be proven on the *functional* level. Also, longer time series should be investigated. Moreover, when more detailed time series data become available, an analysis of regulatory structures on a smaller scale could become possible where more valid investigations could be taken on the levels of special pathways, regulatory modules or motifs (Hartwell et al. 1999; Milo et al. 2002; Lee et al. 2002). Also, integrative bioinformatic approaches involving the combination of gene expression with metabolite profiles and hQTL data could reveal promising results, especially for more local heterosis phenotypes affecting only a small part of the regulatory network (see for example Gärtner et al. 2009; Wei et al. 2009). The discovery of functional groups of genes with particularly enriched partial correlations could complement and refine quantitative genetics analysis about non-additive gene actions and help to approach an understanding of the molecular basis of heterosis.

Hence, the systems biological approach towards finding the molecular basis of heterosis introduced with the current investigation should be seen as a methodological proposal illustrated with a small dataset, complementary to the quantitative genetics approach, which is not taking into account global structures of the various OMICS levels, and the single-gene centered approaches, which involve data of much higher resolution for the price of neglecting the global view.

Acknowledgments This work was supported by the German Research Council (DFG) under Grants RE 1654/2-1 and SE 611/3-1. We want to thank Dirk Hincha (MPIMP-Golm) and his lab for supporting our gene expression experiments.

References

Agilent Technologies Inc. (2008) Agilent feature extraction software: reference guide, 6th edn. USA, G4460-90020

Andorf S, Gärtner T, Steinfath M, Witucka-Wall H, Altmann T, Repsilber D (2009) Towards systems biology of heterosis: a hypothesis about molecular network structure applied for the arabidopsis metabolome. EURASIP J Bioinform Syst Biol 147157

Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Müller R, Meese E, Lenhof HP (2007) Genetrail—advanced gene set enrichment analysis. Nucleic Acids Res 35(Web Server issue):W186–W192

Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57(1):289–300

Birchler JA, Auger DL, Riddle NC (2003) In search of the molecular basis of heterosis. Plant Cell 15:2236–2239

Bruce AB (1910) The mendelian theory of heredity and the augmentation of vigor. Science 32(827):627–628

Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci USA 97(22):12182–12186

Crow JF (1952) Heterosis. In: Dominance and overdominance. Iowa State College Press, Ames, pp 282–297

Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA (2003) Global functional profiling of gene expression. Genomics 81(2):98–104

East EM (1936) Heterosis. Genetics 21(4):375–397

Frisch M, Thiemann A, Fu J, Schrag TA, Scholten S, Melchinger AE (2010) Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. Theor Appl Genet (accepted)

Gärtner T, Steinfath M, Andorf S, Lisec J, Meyer RC, Altmann T, Willmitzer L, Selbig J (2009) Improved heterosis prediction by combining information on DNA- and metabolic markers. PLoS One 4(4):e5220

Genoud T, Métraux JP (1999) Crosstalk in plant cell signaling: structure and function of the genetic network. Trends Plant Sci 4(12):503–507

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80

Guo M, Rupe MA, Yang X, Crasta O, Zinselmeier C, Smith OS, Bowen B (2006) Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. Theor Appl Genet 113(5):831–845

Harrison GA (1962) Heterosis and adaptability in the heat tolerance of mice. Genetics 47(4):427–434

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402(SUPP):C47–C52

Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. Biostatistics 2(2):183–201

Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comput Biol 7(6):819–837

Lamkey KR, Edwards JW (1999) The quantitative genetics of heterosis. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in crops. ASA, CSSA, and SSSA, Madison, pp 31–48

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298(5594):799–804

Li ZK, Luo LJ, Mei HW, Wang DL, Shu QY, Tabien R, Zhong DB, Ying CS, Stansel JW, Khush GS, Paterson AH (2001) Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. I. Biomass and grain yield. Genetics 158:1737–1753

Luo LJ, Li ZK, Mei HW, Shu QY, Tabien R, Zhong DB, Ying CS, Stansel JW, Khush GS, Paterson AH (2001) Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. II. Grain yield components. Genetics 158:1755–1771

Ma L, Sun N (2005) Organ-specific expression of Arabidopsis genome during development. Plant Physiol 138:80–91

Maynard Smith J (1956) Acclimatization to high temperatures in inbred and outbred *Drosophila subobscura*. J Genet 54(1):497–505

Melchinger AE, Utz HF, Piepho HP, Zeng ZB, Schön CC (2007) The role of epistasis in the manifestation of heterosis: a systems-orientated approach. Genetics 177:1815–1825

Meyer RC, Törjék O, Becher M, Altmann T (2004) Heterosis of biomass production in arabidopsis. Establsiment during early development. Plant Physiol 134:1813–1823

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827

Opgen-Rhein R, Strimmer K (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Syst Biol 1:37

Opgen-Rhein R, Schäfer J, Strimmer K (2007) GeneNet: Modeling and Inferring Gene Networks. http://strimmerlab.org/software/genenet/

Parrish RS, Spencer III HJ, Xu P (2009) Distribution modeling and simulation of gene expression data. Comput Stat Data Anal 53(5):1650–1660

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0. http://www.R-project.org

Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK (2007) A comparison of background correction methods for two-colour microarrays. Bioinformatics 23:2700–2707

Robertson FW, Reeve EC (1952) Heterozygosity, environmental variation and heterosis. Nature 170(4320):286

Schäfer J, Strimmer K (2005a) An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 21(6):754–764

Schäfer J, Strimmer K (2005b) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4:32

Schnell FW, Cockerham CC (1992) Multiplicative vs. arbitrary gene action in heterosis. Genetics 131(2):461–469

Shubik M (1996) Simulations, models and simplicity. Complexity 2(1):60

Shull GH (1908) The composition of a field of maize. Am Breeders Assoc Rep 4:296–301

Shull GH (1952) Beginnings of the heterosis concept. In: Gowen JW (ed) Heterosis: a record of researches directed toward explaining and utilizing the vigor of hybrids. Iowa State College Press, Ames, pp 14–48

Smyth GK (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W (eds) Bioinformatics and computational biology solutions using R and bioconductor. Springer, New York, pp 397–420

Smyth GK, Speed T (2003) Normalization of cDNA microarray data. Methods 31:265–273

Somogyi R, Sniegoski CA (1996) Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. Complexity 1:45–63

Song R, Messing J (2003) Gene expression of a gene family in maize based on noncollinear haplotypes. Proc Natl Acad Sci USA 100:9055–9060

Steinfath M, Gärtner T, Lisec J, Meyer RC, Altmann T, Willmitzer L, Selbig J (2010) Prediction of hybrid biomass in *Arabidopsis thaliana* by selected parental SNP and metabolic markers. Theor Appl Genet (accepted)

Strimmer K (2008) A unified approach to false discovery rate estimation. BMC Bioinformatics 9:303

Swanson-Wagner RA, Jia Y, DeCook R, Borsuk LA, Nettleton D, Schnable PS (2006) All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. Proc Natl Acad Sci USA 103(18):6805–6810

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E (2008) The Arabidopsis information resource (TAIR): gene structure and function annotation. Nucleic Acids Res 36(Database issue): D1009–D1014. http://www.arabidopsis.org

The Plant Ontology Consortium (2002) The plant ontology consortium and plant ontologies. Comp Funct Genomics 3:137–142. http://www.plantontology.org

Thiemann A, Fu J, Schrag TA, Melchinger AE, Frisch M, Scholten S (2010) Correlation between parental transcriptome and field data for the characterization of heterosis in *Zea mays* L. Theor Appl Genet (accepted)

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37(6):914–939

Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A, Stitt M (2009) A guide to using MapMan to visualize and compare omics data in plants: a case study in the crop species, Maize. Plant Cell Environ 32(9):1211–1229

Vuylsteke M, van Eeuwijk F, Van Hummelen P, Kuiper M, Zabeau M (2005) Genetic analysis of variation in gene expression in *Arabidopsis thaliana*. Genetics 171(3):1267–1275

Wei G, Tao Y, Liu G, Chen C, Luo R, Xia H, Gan Q, Zeng H, Lu Z, Han Y, Li X, Song G, Zhai H, Peng Y, Li D, Xu H, Wei X, Cao M, Deng H, Xin Y, Fu X, Yuan L, Yu J, Zhu Z, Zhu L (2009) A transcriptomic analysis of superhybrid rice LYP9 and its parents. Proc Natl Acad Sci USA 106(19):7695–7701

Werhli AV, Grzegorczyk M, Husmeier D (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics 22(20):2523–2531

Xiao J, Li J, Yuan L, Tanksley SD (1995) Dominance is the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers. Genetics 140:745–754

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30(4):e15