Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci

Keurentjes, Joost J.B.; Fu, Jingyuan; Terpstra, Inez R.; Garcia, Juan M.; Ackerveken, Guido van den; Snoek, L. Basten; Peeters, Anton J.M.; Vreugdenhil, Dick; Koornneef, Maarten; Jansen, Ritsert C.

Published in:
Proceedings of the National Academy of Sciences of the United States of America

DOI:
10.1073/pnas.0610429104

Genomic DNA Hybridizations. We determined variation in gene expression between two distinct accessions of *Arabidopsis* as well as within a RIL population derived from these accessions. However, the microarray probe set was designed on the sequenced accession Col and signal intensity ratios might therefore be affected by hybridization differences due to DNA polymorphisms between Col, L*er*, and Cvi. We therefore assessed this effect by hybridizing genomic DNA of the parental accessions.

Genomic DNA was isolated with CTAB-buffer (100 mM Tris·HCl/20 mM EDTA/1.4 M NaCl/2% wt/vol CTAB/1% wt/vol PVP-40/2% vol/vol 2-mercaptoethanol) followed by phenol/chloroform extraction and subsequently sheared in a nebulizer (Invitrogen, Valencia, CA) according to the manufacturer's protocol. Ten micrograms of genomic DNA was nebulized in 750 μl of shearing buffer for 90 seconds at 10 psi. Four μg of each DNA sample was amplified and labeled with the BioPrime Plus Array CGH Indirect Genomic Labeling System (Invitrogen). DNA probe immobilization was performed according to instructions from the Galbraith laboratory (http://ag.arizona.edu/microarray/). DNA microarrays were three times rehydrated over a 50°C water bath for 10 sec and snap dried on a 70°C heat block, followed by UV cross-linking (140 mJ). Four μg of each labeled DNA sample was combined with 2× hybridization buffer (50% formamide, 10× SSC and 0.2% SDS) and denatured for five minutes at 70°C. Samples were hybridized to the slides for 12-18 h in hybridization chambers at 42°C. Washes were performed at room temperature in 1× SSC/0.1% SDS, 0.2× SSC/0.1% SDS, and 0.2× SSC respectively. Arrays were scanned using a ScanArray Express HT (PerkinElmer, Wellesley, MA) and quantified with Imagene 6.0 (BioDiscovery, El Segundo, CA).

A comparison on 4 microarrays, including a dye swap, yielded no significant differential hybridization between the 2 genomes (q <0.05).

Alternatively we used CGH-Plotter, which has been designed to identify the deletion or amplification of groups of genes by applying k-means clustering and dynamic programming (1). Nineteen differentially hybridizing regions were identified containing 148 genes (SI Table 3). The larger regions are located on chromosome 3 (genes *At3g42240* to *At3g42570*) and chromosome 4 (genes *At4g16830* to *At4g17450*) and comprise numerous retrotransposon and transposase family members, as well as disease resistance genes and their homologs. Only 10 of the 148 genes are also differentially expressed between the parental accessions (1.2% of 855 differentially expressed genes) and 26 (0.65% of 4,066) showed an expression QTL with no enrichment for locally regulated genes. We therefore conclude that hybridization effects due to genomic polymorphisms only have a minor effect on gene expression analysis, as Kliebenstein *et al* also concluded. (2).

Automatic Cofactor Selection. In the multiple QTL model mapping, the cofactors were selected by using backward elimination process. A total of 55 evenly distributed markers with an average distance of 9.2 cM were preselected as initial cofactors. We then fitted the observed expression data to a model **<graphic1>**, where y is the log-ratio of signal intensities, x_i is the genotype comparison at the ith cofactor (i = 1.55), taking values 1 for Cvi/L*er*, -1 for L*er*/Cvi and 0 for L*er*/L*er* and Cvi/Cvi; and b_i is the substitution effect for the ith cofactor. This model was called a full model, which is an unbiased estimate of the noise, not suffering from overfitting. We then conducted an ANOVA test to compute the F and P values for explained variation by each factor, the estimate of residual variance (V_{full}) and the residual degree of freedom (df_{full}). Of all F statistics not significant at 99.9% confidence ($P < 0.001$) with 1 df_{full}, the factor with the lowest F value was eliminated from the model. *ANOVA*

analysis on this reduced model recomputed the F values for explained variance by using remaining factors and the residual variance (V_{red}). The F values in reduced models were adjusted by a factor V_{red}/V_{full}, and then these corrected F values could be analyzed by using the same cutoff as in the full model. This backward elimination process was repeated until all remaining markers were significant ($P < 0.001$) and the number of remaining cofactors was <10 because we would not expect the number of QTL for a single transcript to exceed 10. In the QTL mapping procedure, these remaining markers were used as cofactors for QTL detection.

Genetic Regulatory Network Construction. Combined expression-trait correlations and expression-quantitative trait locus mapping has been used to increase the power of identifying the candidate regulator gene or novel target genes. Bing and Hoeschele (3) computed the Spearman rank correlation coefficient between the expression profiles of genes in an eQTL region and the profile of the gene mapped to that region. At least one candidate was retained with a significant and highest correlation coefficient. However, the gene affected by an eQTL region and all locally regulated genes in this eQTL region are probably coexpressed due to a linked genetic effect. The true regulator may be among the top regulators, but will not always be the superlative. To improve the reliability in predicting regulators, especially for those master regulators with a pleiotropic effect, we considered the function-related genes in a group. We made an assumption that the function-related genes mapping to the same eQTL region are likely to have one and the same regulator in that region. The most likely candidate is the one that best correlates to the whole group. Another reason for starting with subsets of function-related genes is that genome-wide studies always have to face the conflict between the power of detection (in favor of a less stringent threshold) and the control of FDR (in favor of a stringent threshold). The function-related gene can be selected by different ways based on, e.g., gene family, keywords, or Gene Ontology terms. To illustrate our method, we first consider the gene regulatory network for flower genes as an example. Flower genes here are defined as the genes annotated with keywords "circadian rhythm," "flower development," or "photoreceptor" in The *Arabidopsis* Information Resource database. This initial subset was complemented with literature mining (4-24). The final set contained 192 genes, 175 of which were measured in our experiments. A total of 83 genes showed significant linkage at the genome-wide threshold of P value 2.23×10^{-3}. The eQTL support intervals of each mapped gene were determined by setting left and right border positions associated with $\max\{-\log_{10}P\}-1.5$. The regulator candidates are the genes physically located in the eQTL intervals. The candidates were sorted by using iGA, which was initially proposed to identify the functional classes of genes that are significantly changed in a microarray experiment (25). We postulated that, among all possible regulators, the best candidates are those that correlate particularly well to a large number of their potential target genes. We calculated all pairwise Spearman rank correlations on expression profiles (80 log ratios of cohybridized RILs) between each of the 83 mapped flower genes and all potential regulators in their eQTL intervals. The number of potential regulators is y_i (i = 1,...,83) for the ith flower gene, and the total number of correlation coefficients (**<graphic2>**) is 105,899, with 23,306 potential candidate genes. These values were then rank-ordered so that the strongly correlated gene-candidate pairs were at the top of the list. We moved along the rank list of all correlation coefficients from top to bottom, counting the genes mapped to one given candidate regulator, and each time we encountered a new member we asked how likely it was to observe this many members of this given candidate that high up in the list by chance. This probability (P value) is exactly

<graphic3>;**<graphic4>**,

where n is the total number of correlation coefficients (n took value 105,899 in the case of flower genes), x is the total number of flower genes mapping to a given candidate gene, and t is the rank of the zth member. The notation **<graphic5>** indicates the binomial coefficient (i.e., the number of ways of picking y unordered items from a list of x items). By this method, we got a vector of P values associated with the comapped flower genes for each candidate regulator. For each given candidate, we determined the position, z, in the vector that yielded the smallest P value and assigned this value as the iGA possibility of change value (PC value) of this given candidate gene. Then, the top z comapping flower genes made a contribution to this PC value. The PC-value threshold was 2.15×10^{-6}, Bonferroni adjusted as $0.05/m$, where m is the total number of candidate genes. Any candidate with a significant PC value can be a putative regulator, and the flower genes contributing to this PC value are its potential target genes. We retained the regulator with the lowest PC value and defined the regulatory relation in terms of the sign of the correlation coefficient. If the correlation coefficient is negative, regulation is repressive; otherwise, it is activating.

We started with a subset of known function-related genes, i.e., 192 flower genes. This is a knowledge-driven selection. The next natural step is to find whether there are any novel genes comapping with this functional class or coregulated, if linkage was not significant. A method proposed by Lan *et al.* (26) was used. Instead of finding seed transcripts by clustering QTL profiles, we used the regulators and target genes obtained from iGA study as seed transcripts. The log ratio gene expression profile matrix (axb) was then split into two parts: one is the a_1xb matrix for seed transcripts; the other is the a_2xb matrix for other genes, where a is the total number of gene transcripts (a = 24,065 in our case); a_1 is the number of seed transcripts; a_2 is the number of other genes ($a_1+a_2 = a$); and b is the number of arrays (b = 80 in our case). We computed the Spearman correlation coefficient and its corresponding P value for each a_1 seed gene with each a_2 transcript. A permutation test was used to compute an empirical threshold and estimate its corresponding FDR (27) for which we randomly permutated the b columns in the seed transcripts matrix. Therefore, the correlations between seed transcripts and other genes were interrupted but the correlation structures within these two sets were intact. The correlation coefficients were computed as described above, and the highest absolute coefficient was recorded. The process was repeated 1,000 times, and the 95th percentile of the rank-ordered coefficients generated an empirical threshold. The transcripts passing this threshold were potential novel target genes.

1. Autio R, Hautaniemi S, Kauraniemi P, Yli-Harja O, Astola J, Wolf M, Kallioniemi A (2003) *Bioinformatics* 19:1714-1715.

2. Kliebenstein DJ, West MA, van Leeuwen H, Kim K, Doerge RW, Michelmore RW, St Clair DA (2006) *Genetics* 172:1179-1189.

3. Bing N, Hoeschele I (2005) *Genetics* 170:533-542.

4. Ausin I, Alonso-Blanco C, Jarillo JA, Ruiz-Garcia L, Martinez-Zapater JM (2004) *Nat Genet* 36:162-166.

5. Ausin I, Alonso-Blanco C, Martinez-Zapater JM (2005) *Int J Dev Biol* 49:689-705.

6. Baurle I, Dean C (2006) *Cell* 125:655-664.

7. Boss PK, Bastow RM, Mylne JS, Dean C(2004) *Plant Cell* 16:S18-S31.

8. Edwards KD, Lynn JR, Gyula P, Nagy F, Millar AJ (2005) *Genetics* 170:387-400.

9. Hayama R, Coupland G(2003) *Curr Opin Plant Biol* 6:13-19.

10. Imaizumi T, Schultz TF, Harmon FG, Ho LA, Kay SA (2005) *Science* 309:293-297.

11. Kevei E, Gyula P, Hall A, Kozma-Bognar L, Kim WY, Eriksson ME, Toth R, Hanano S, Feher B, Southern MM, *et al.* (2006) *Plant Physiol* 140:933-945.

12. Kobayashi Y, Kaya H, Goto K, Iwabuchi M, Araki T (1999) *Science* 286:1960-1962.

13. Komeda Y (2004) *Annu Rev Plant Biol* 55:521-535.

14. Levy YY, Dean C (1998) *Plant Cell* 10:1973-1990.

15. Levy YY, Dean C (1998) *Curr Opin Plant Biol* 1:49-54.

16. McClung CR (2006) *Plant Cell* 18:792-803.

17. Michael TP, Salome PA, Yu HJ, Spencer TR, Sharp EL, McPeek MA, Alonso JM, Ecker JR, McClung CR (2003) *Science* 302:1049-1053.

18. Mizoguchi T, Wright L, Fujiwara S, Cremer F, Lee K, Onouchi H, Mouradov A, Fowler S, Kamada H, Putterill J, Coupland G (2005) *Plant Cell* 17:2255-2270.

19. Mouradov A, Cremer F, Coupland G (2002) *Plant Cell* 14:S111-S130.

20. Parcy F (2005) *Int J Dev Biol* 49:585-593.

21. Simpson GG, Dean C (2002) *Science* 296:285-289.

22. Simpson GG, Gendall AR, Dean C (1999) *Annu Rev Cell Dev Biol* 15:519-550.

23. Swarup K, Alonso-Blanco C, Lynn JR, Michaels SD, Amasino RM, Koornneef M, Millar AJ (1999) *Plant J* 20:67-77.

24. Yanovsky MJ, Kay SA (2003) *Nat Rev Mol Cell Biol* 4:265-275.

25. Breitling R, Amtmann A, Herzyk P (2004) *BMC Bioinformatics* 5:34.

26. Lan H, Chen M, Flowers JB, Yandell BS, Stapleton DS, Mata CM, Mui ET, Flowers MT, Schueler KL, Manly KF, *et al.* (2006) *PLoS Genet* 2:e6.

27. Storey JD, Tibshirani R (2003) *Proc Natl Acad Sci USA* 100:9440-9445.