nature genetics

System-wide molecular evidence for phenotypic buffering in *Arabidopsis*

Jingyuan Fu[1,2,9], Joost J B Keurentjes[3–5,9], Harro Bouwmeester[4–6], Twan America[5,6], Francel W A Verstappen[4–6], Jane L Ward[7], Michael H Beale[7], Ric C H de Vos[5,6], Martijn Dijkstra[1], Richard A Scheltema[1], Frank Johannes[1], Maarten Koornneef[3,8], Dick Vreugdenhil[4], Rainer Breitling[1] & Ritsert C Jansen[1,2]

We profiled 162 lines of *Arabidopsis* for variation in transcript, protein and metabolite abundance using mRNA microarrays, two-dimensional polyacrylamide gel electrophoresis, gas chromatography time-of-flight mass spectrometry, liquid chromatography quadrupole time-of-flight mass spectrometry, and proton nuclear magnetic resonance. We added all publicly available phenotypic data from the same lines and mapped quantitative trait loci (QTL) for 40,580 molecular and 139 phenotypic traits. We found six QTL hot spots with major, system-wide effects, suggesting there are six breakpoints in a system otherwise buffered against many of the 500,000 SNPs.

Complex traits can be the consequence of heritable variations in the genome and probably include many human diseases. The Human Genome Project and its successor, the Human Variome Project[1], have started to catalog the millions of possible variations in the human genome, notably SNPs, indels, copy number variations (CNVs) and alternative splicing variants. Other projects have started to associate these variations with disease traits in order to detect disease-associated alleles, and with transcript abundance traits in order to elucidate the molecular networks that cause diseases[2,3]. Complementary information is obtainable from similar studies on model organisms such as mouse, yeast and *Arabidopsis thaliana*, in which it is often easier to create and collect appropriate samples.

Here we report the results from the first system-wide genetical genomics[4] study of molecular variation in a model organism, in which we integrate transcript, protein and metabolite data generated by our groups with publicly available phenotypic data from a population of 162 L*er* × Cvi recombinant inbred lines (RILs) of *Arabidopsis thaliana* (**Supplementary Methods** online). All the molecular observations were based on the same biological samples of seedlings (40,580 molecular traits in total), and all the phenotypic data came from a wide range of growth stages, derived from many different studies on the same RIL population (139 phenotypic traits belonging to 35 phenotypic trait categories; **Supplementary Table 1** online). We observed profound and widespread genetic control at all levels, from expressed gene to phenotype (**Fig. 1**). QTLs of over 2,000 transcript traits mapped close to the gene itself (local eQTL), whereas QTLs for almost 3,000 transcript traits mapped to a location different from the gene position (distant eQTL). Notably, these 5,000 transcripts on average show one QTL, as do the proteins (pQTL). Metabolites, however, on average show two QTLs (mQTLs) and phenotypes even show three QTLs (phQTLs) per trait. There are several significant QTL hot spots on the genome underlying variation in many molecular and phenotypic traits, such as the location of the *ER* marker gene on chromosome 2 ($P < 10^{-6}$). Another hot spot is located near the GH.473C marker on chromosome 5 ($P < 10^{-6}$), the most prominent hot spot for protein variation. These hot spots propagate to the metabolite and phenotype level. There are additional hot spots for metabolites and phenotypes at the DF.77C marker on chromosome 3 ($P < 2.7 \times 10^{-6}$), at the EC.66C marker on chromosome 1 ($P = 4.3 \times 10^{-4}$) and at the GH.121L marker on chromosome 5 ($P = 1.5 \times 10^{-4}$), and there is a hot spot for the phenotype level at the *CRY2* marker gene on chromosome 1 ($P = 7.0 \times 10^{-3}$). Together, these six hot spots influence 16%, 25%, 55% and 77% of 4,832 transcript, 253 protein, 7,158 metabolite and 116 phenotypic traits with QTLs, respectively, when a window of 5 cM around the hot spot is used to account for imperfect mapping resolution in the QTL analysis (**Fig. 1** and **Supplementary Fig. 1** online). The hot spots also seem to act in concert and to apply across a wide range of phenotypic trait categories (**Supplementary Fig. 2** online). The hot spots are not caused by redundant reporting of results in certain phenotypic trait categories; indeed, they influence 94% of the 33 phenotypic trait categories with QTLs (**Supplementary Fig. 3** online).

There are at least 500,000 SNPs between the two parental lines[5] that have given rise to one or more eQTLs for each of about 5,000 transcripts. However, only a handful of these genetic effects have propagated to the phenotype level and led to QTL hot spots. Initially, this is a surprising observation, because each of the eQTLs can be considered a molecular perturbation of the biological system and could have general consequences in the densely connected regulatory networks. On the other hand, it may be less surprising when considered from an engineering or system-biology perspective: our genome-wide and system-wide results fit the predictions of robustness theory[6–9] and experimental mutagenesis results[10] that indicate that much of the genetic variation in gene expression networks is hidden

[1]Groningen Bioinformatics Centre, University of Groningen, 9751NN Haren, The Netherlands. [2]Department of Genetics, University Medical Centre Groningen, University of Groningen, 9700RB Groningen, The Netherlands. [3]Laboratory of Genetics, Wageningen University, 6703BD Wageningen, The Netherlands. [4]Laboratory of Plant Physiology, Wageningen University, 6703BD Wageningen, The Netherlands. [5]Centre for Biosystems Genomics, Wageningen, 6708PB Wageningen, The Netherlands. [6]Plant Research International, Wageningen, 6708PB Wageningen, The Netherlands. [7]Rothamsted Research, National Centre for Plant and Microbial Metabolomics, AL5 2JQ Harpenden, Herts, UK. [8]Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. [9]These authors contributed equally to this work. Correspondence should be addressed to R.C.J. (r.c.jansen@rug.nl).

Received 17 June 2008; accepted 10 November 2008; published online 25 January 2009; doi:10.1038/ng.308

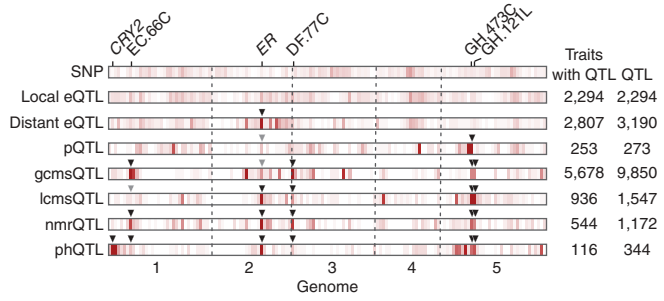

Figure 1 System-wide view on the effects of genome variation. A heat map of the genomic locations of 481,299 known SNPs in gene regions is shown in the top panel. Below it, seven heat maps of QTL distributions are shown: for 24,065 transcript abundance traits (local and distant eQTL), for 2,843 protein abundance traits (pQTL), for 13,672 metabolite abundance traits (gcmsQTL, lcmsQTL and nmrQTL) and for 139 phenotypic traits (phQTL). The number of traits with a QTL and the total number of QTLs for each heat map are shown on the right-hand side of the heat maps. Genomic positions of several QTL hot spots are indicated by the corresponding marker names and arrows above the heat maps. A black arrow indicates that the hot spot effect passes level-wise and system-wide significance thresholds; a gray arrow indicates that the hot spot effect contributes to the system-wide significance but does not pass the level-wise significance threshold (see **Supplementary Methods**). In total, there are 144 markers along the genome at a spacing of 3.5 centiMorgan (cM) on average and the heat maps show the number of SNPs and QTLs at each marker position; we have plotted the 144 markers at equal spacing, not at cM spacing, to ease viewing (see **Supplementary Fig. 4** online for the genetic map and **Supplementary Fig. 5** online for QTL profiles of individual traits). Dashed vertical lines crossing heat maps indicate chromosome boundaries.

by nonlinearity in response functions and is selectively neutral or nearly neutral[11]. Such system properties suppress the propagation of variation to the phenotypic level, even in the absence of specific selection or molecular feedback mechanisms[9]. Robustness is essential to keep processes and traits in any living organism within acceptable and quite narrow limits, even in the case of major genetic variation: without a robust system design, minor variations in thousands of genes would regularly lead to massive changes at the phenotype level, causing dysfunction in recombined individuals. Our system-wide data provide the first genome-wide and system-wide empirical evidence for this robustness, as most heritable variation in individual molecular traits is only associated with downstream variation in molecular and phenotypic traits to a minor extent.

Our data also expose a number of exceptions to this robustness rule: the hot spots discussed above, where genetic variation is apparently 'unlocked' and becomes clearly visible at the phenotypic level. These hot spots seem to correspond to a few molecular 'breakpoints' of an otherwise robust regulatory system. Such fragilities have also been predicted by robustness theory[6–9], and our study is a first step toward their identification in a model organism. In general, hot spots seem to be quite rare. We found only six hot spots in our system-wide study involving two rather diverse parents, and none have been detected so far in one of the largest human genetical genomics studies on gene expression[2]; in other organisms, between zero and eight were found[12]. It is notable that most of the hot spots in our study can be linked to well-studied genes such as *CRY2*, *INV* (EC.66C), *ER*, *MAM1/2* (GH.473C) and *HUA2* or *FRL1* (GH.121L), which are known for their pleiotropic effects on plant metabolism (*MAM1/2*, *INV*), physiology (*CRY2*) and morphology and development (*ER*, *HUA2*, *FRL1*) due to their function at central cellular network hubs[13]. This indicates that classic experiments with mutants could be equally biased toward a few fragile elements of cellular circuitry. Owing to the imperfect resolution of QTL mapping, it cannot be excluded that the hot spot regions contain other causal genes with less apparent effects than the currently known hub genes. Now that our genetic analysis has identified these hot spots, future molecular studies and systems biology analyses should aim at elucidating their roles and interplay in cellular networks.

In conclusion, our study shows that the huge genome and transcriptome variation between two accessions of *Arabidopsis* is subject to pervasive genetic buffering. The largest fraction of molecular variants is silent at the phenotypic level, and only a few influential 'hot spot' regions cause major phenotypic variation across a range of environmental conditions. Whether such hot spots form the basis for evolutionary adaptation to changing environments remains to be determined. A generalization of genetical genomics, which studies the effect of genetic variation at the hot spots in multiple well-chosen environments, will be a useful approach toward answering this question[14]. Our results are also in agreement with recent findings that many human diseases share their genetic origin with other diseases to some extent[15]. Fragilities at crucial nodes in the molecular networks may underlie this phenomenon. After all, molecular networks may be complex, but complex traits could well be a lot simpler than previously thought: variation in a multitude of *Arabidopsis* complex traits can be explained to a considerable extent by only a few QTL hot spots.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This research was supported by the Netherlands Organization for Scientific Research (NWO), the Netherlands Genomics Initiative (NGI) and the UK Biotechnology and Biological Sciences Research Council (BBSRC).

AUTHOR CONTRIBUTIONS

J.F. analyzed the system-wide data, J.J.B.K. managed the system-wide data production, H.B. and F.W.A.V. produced the GC-MS data, T.A. produced the 2D-PAGE data, J.L.W. and M.H.B. produced the NMR data, R.C.H.d.V. produced the LC-MS data, M.D. and R.A.S. helped analyze the GC- and LC-MS data, F.J. helped analyze the NMR data, M.K. coordinated the QTL-Express project and provided the biological materials, D.V. contributed to the biological interpretation, R.B. contributed to the system-wide interpretation and writing of the manuscript, R.C.J. conceived the project and coordinated the analysis and writing of the manuscript.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Anonymous. *Nat. Genet.* **39**, 423 (2007).
2. Emilsson, V. *et al. Nature* **452**, 423–428 (2008).
3. Chen, Y. *et al. Nature* **452**, 429–435 (2008).
4. Jansen, R.C. & Nap, J.P. *Trends Genet.* **17**, 388–391 (2001).
5. Clark, R.M. *et al. Science* **317**, 338–342 (2007).
6. Kitano, H. *Nat. Rev. Genet.* **5**, 826–837 (2005).
7. Kitano, H. *Mol. Syst. Biol.* **3**, 137 (2007).
8. Gjuvsland, A.B., Plahte, E. & Omholt, S.W. *BMC Syst. Biol.* **1**, 57 (2007).
9. Bergman, A.M. & Siegal, M.L. *Nature* **424**, 549–552 (2003).
10. Queitsch, C., Sanger, T.A. & Rutherford, S.L. *Nature* **417**, 618–624 (2002).
11. Khaitovich, P. *et al. PLoS Biol.* **2**, e132 (2004).
12. de Koning, D.J. & Haley, C.S. *Trends Genet.* **21**, 377–381 (2005).
13. El-Assal, E.S. *et al. Nat. Genet.* **29**, 435–440 (2001).
14. Yang, L., Breitling, R. & Jansen, R.C. *Trends Genet.* **24**, 518–524 (2008).
15. Goh, K.I. *et al. Proc. Natl. Acad. Sci. USA* **21**, 8685–8690 (2007).