Application of omics technologies for environmental risk assessment of genetically modified plants

Arabidopsis and modified defence mechanisms as a model study

Benyamin Houshyani

Thesis committee

Thesis supervisors
Prof. dr. ir. Harro J. Bouwmeester
Professor of Plant Physiology
Wageningen University

Prof. dr. Raoul J. Bino
Professor of Metabolomica of Plants
Wageningen University

Thesis co-supervisor
Dr. ir. Iris F. Kappers
Researcher at the laboratory of Plant Physiology
Wageningen University

Other members
Prof. dr. Nicole van Dam, Radboud University Nijmegen
Prof. dr. Robert Verpoorte, Leiden University
Prof. dr. Richard G.F. Visser, Wageningen University
Dr. ir. Esther J. Kok, RIKILT, Wageningen UR

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences

Application of omics technologies for environmental risk assessment of genetically modified plants

Arabidopsis and modified defence mechanisms as a model study

Benyamin Houshyani

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 16 March, 2012
at 1.30 in the Aula.

Benyamin Houshyani
Application of omics technologies for environmental risk assessment of genetically modified plants - *Arabidopsis* and modified defence mechanisms as a model study
230 pages.

Thesis, Wageningen University, Wageningen, NL (2012)
With references, with summaries in English

ISBN 978-94-6173-103-6

The research for this PhD dissertation has been funded by the Earth and Life Sciences Council of the Netherlands Organization for Scientific Research (NWO-ALW) under the ERGO program number 838.06.010.

Financial support of Wageningen University for printing this thesis is acknowledged.

With them the Seed of Wisdom did I sow,
And with my own hand labour'd it to grow:
And this was all the Harvest that I reap'd -
I came like Water, and like Wind I go.

Omar Khayyam (Persia 1048-1131)

	Contents	Page

To my beloved wife Parisa,

much-loved parents Mahin and Shaban and brother Bamshad

Chapter 1

General Introduction

Application of omics technologies for environmental risk assessment of genetically modified plants

Benyamin Houshyani

Introduction

Genetically modified (GM) crops recorded the firmest growth in application among biotechnology products with an 87-fold of hectarage increase in 2010 after 15 years of commercial practice (James 2010). Simultaneously, new developments in plant biotechnology continuously result in novel biotech crops with novel features (e.g. drought tolerant maize, Golden rice and Amflora starch potato) that progressively provide solutions for experienced problems. New developments also respond to the concerns of the society and critics about the potential associated risks *i.e.* by temporal and spatial control of transgene expression or by cisgenesis. However, in response to the concerns of society and criticizers, a detailed environmental risk assessment (ERA) is demanded by the European Union (EU) before permitting the release of a GM crop into the fields. Since GM plants are seen as new entities in the EU, assessment of their potential risks is independent of standard risk assessment procedures of plant protection products (EC 2001). In line with that, the Dutch Government sponsored the Ecology Regarding Gene-modified Organism (ERGO) research program. This research program started in 2007 to develop ecological guidelines for risk analysis of GM crops based on sound science.

During the implementation of the ERGO-programme, the European Food Safety Authority (EFSA) published the 'guidance of the environmental risk assessment of genetically modified plants' in 2010. This document addressed seven specific areas of concern, including the invasiveness of the GM crop, plant-to-microbe gene transfer, effects of the GM crop on human and animal health, and the *interaction of the GM crop with non-target organisms* (EFSA 2010). To improve studies on the effect of a GM crop on non-target organisms (Groot and Dicke 2002) EFSA provided guidelines on the inclusion of relevant groups of non-target species, endpoint (phenotypes or characters) selection for measurements and performance of experiments under laboratory conditions (EFSA 2010). The EFSA guidelines do not advise on the application of *omics* platforms in GM crops assessment not on methodological tools for their implementation. The use of *omics* therefore was considered an important aspect in the "multi-trophic interaction of genetically modified crops" theme of the ERGO-program. The focus of this thesis is mainly on the application of *omics* platforms and particularly metabolomics, in GM crop assessment. Two other projects of this theme focus mainly on the potential impact of GM crops on multi-trophic interactions in ecological contexts (above- and below ground). Collectively, these three projects should result in a multi-disciplinary and integrated approach for an ecology-based and sound-science evaluation guideline for regulators that can be used also later for assessment of non-target effects of GM crops under field conditions.

Application of omics platforms in GM crops assessment

Potential alterations in the cellular metabolism are one of the very first and still existing concerns when a recombinant DNA is introduced into the genome of a plant. Critics argue that some of these alterations are not traceable during the conventional agronomic safety evaluations and could potentially be a risk for living organisms. Omics platforms and particularly transcriptomics became of interest for these evaluations because they currently are the best tools for global molecular characterization (Kok and Kuiper 2003). Transcriptomics technologies became therefore the first applied tools for untargeted assessment of GM plants.

In a global gene expression profiling approach Baudo *et al.* (2006) compared the transcriptome of wheat (*Triticum aestivum*) lines overexpressing additional genes for glutenin biosynthesis with their non-GM counterparts and found that transgenesis did not significantly alter the transcriptome, and concluded that GM lines were substantially equivalent to their parental lines. In another study testing the transcriptome of a GM rice (*Oryza sativa*), overexpressing a gene encoding the anthranilate synthase alpha subunit had little effect on the overall transcriptome as expression of only 0.05% of the analysed genes was changed compared with the non GM counterpart (Dubouzet *et al.* 2007). In soybean (*Glycine max*) gene expression differed more between two conventional cultivars than between transgenics with glyphosate tolerance and their closest non-transgenic counterpart (Cheng *et al.* 2008). It has also been reported that genetic and environmental variations explain most of the transcriptome changes between MON810 transgenic and non-GM maize (*Zea mays*) varieties (Coll *et al.* 2010).

Metabolomics is a relatively new technique but a few records reported untargeted assessment of changes in metabolism in GM crops. NMR based comparison of wheat lines overexpressing a gene for glutenin biosynthesis with their non-GM counterparts demonstrated that differences between control and GM lines are generally within the same range as between controls of different genotypes grown in different environments (Baker *et al.* 2006). GM rice overexpressing a gene encoding the anthranilate synthase alpha subunit (Dubouzet, Ishihara, Matsuda, Miyagawa, Iwata and Wakasa 2007) showed minimal difference in its metabolome with the conventional parent as demonstrated by HPLC-PDA. In a recent study Kusano *et al.* (2011) used multiple analytical platforms and concluded that the difference between the transgenic lines of tomato (*Solanum lycopersicum*) transformed with a taste modifying enzyme and the 6 traditional cultivars was small compared to the difference between traditional cultivars.

Metabolomics as a more comprehensive tool

Omics technologies can be used to make a snapshot of the transcript, protein and metabolite composition of a plant. Each of these could contribute to GM crop assessment, but each has specific advantages and shortcomings. With the current developments in sequencing, the position of transcriptomics as the most global tool (Gregersen *et al.* 2005) is clear. Transcriptomics provides information on – sometimes genome wide - transcriptional changes and their biological relevance, especially in model plants can be inferred easier as information about gene function is increasingly available and far exceeds that of metabolites and proteins. However, transcriptomics does not provide information about the concentration of primary and secondary metabolites, the biological endpoints (Harrigan *et al.* 2007), an attribute that applies to a lesser extent to proteins and the least to transcripts. Untargeted analysis of the natural metabolome variation within a species has not only led to the discovery of correlations between individual metabolites and the elucidation of metabolic pathways (Hirai *et al.* 2007, Sønderby *et al.* 2007) but also to the discovery of genes, quantitative trait loci and metabolites that control or correlate with plant resistance and susceptibility (Ferrari *et al.* 2003, Figueiredo *et al.* 2008, Gatehouse 2002, Goodman *et al.* 2004, Kashif *et al.* 2009, Rohr *et al.* 2011, Rulhmann *et al.* 2002, Yang *et al.* 2011). In other words, there is increasing evidence that metabolites link genetic variation to ecological traits. Metabolomics techniques may be used in the near future as a relatively quick tool to select genotypes and environmental conditions that may represent as much variation as possible for ecological traits.

However, there are also inevitable shortcomings to metabolomics studies. An important question is whether a single metabolomics approach will provide a sufficiently diverse and representative profile of the metabolite variation in a plant. A solution for this could be the application of multiple analytical platforms to cover as much as possible metabolites from as many as possible classes. However, interpretation of these complete datasets will be difficult. No single data base claims to contain all existing metabolites in a single plant species. This problem can be overcome for GM assessment if consensus exists about the definition of risk-bearing metabolites. In this thesis, I consider a compositional change in metabolites risky only if a certain change exceeds the *consented range of variation* which is called the *baseline*.

Another underestimated issue in the application of metabolomics for GM plant assessment is the multidimensionality of the data. In fact, not just a few metabolites but the whole metabolome composition of a plant is under investigation. Therefore, comparison of two groups of plants and the judgment about the difference between them needs a more comprehensive approach than that what is usually being performed through univariate analyses such as ANOVA.

Non-target ecological effects and their assessment

Supporters of GM crop application often claim their "substantial equivalence" to their parental varieties. Challengers of GM crops refer to the substantial equivalence as being without any statistical basis and emphasize on the possible non-target effects to the surrounding environment due to genetic transformation. The current developments and trends after 16 years of first GM introduction shows that society's and among them the critic's opinion is of prime importance to create and maintain a sustainable ground for the growth of biotechnological crops. Therefore, the assessment of GM crop with respect to (potential) non-target effects, *i.e.* potential effects that were not targeted by transformation, is suggested (Craig *et al.* 2008, Hilbeck *et al.* 2011, Poppy 2000). Non-target ecological effects of biotech crops might include an effect on above- and below-ground non-target organisms, particularly when transformation enhances plant defence. As depicted by Kos *et al.* (2009) several ways exists in which an organism such as a carnivore or a pollinator can be exposed to a transgenic product (Fig. 1).

Even though we can postulate all the possible ways of how a GM crop can affect non-target organisms, there are no standardized guidelines how to test the non-target effects. A number of studies (for a review see Kos *et al.* 2009) propose a stepwise (tierd) approach for this purpose in GM crops engineered to produce insecticidal substances. The assessment in this approach starts in the laboratory and increases in complexity and realism towards semi-field and field set-ups. The knowledge that is gained in each step is used in subsequent steps and feedback between the tiers is necessary during the assessment. Conclusions from one level should be re-examined at another level and when the effects of GM crops on non-target organisms have been tested at all levels, extrapolation of the non-target effects of the transgenic crop can be made.

Methods for evaluation of non-target ecological effects under laboratory conditions are available, but validating the laboratory results needs methods for assessment under (semi-) field situations, which to date are lacking. A comprehensive evaluation of non-target ecological effects of biotech crops especially on the biotic environment is very complex due to its elaborate and time- and resource-consuming nature. Therefore, a universally consented and applicable alternative method for an estimation of non-target ecological effects of GM crops would be alternative and could be a used for a *first-pass* global assessment followed by a more limited non-target ecological effects assessment.

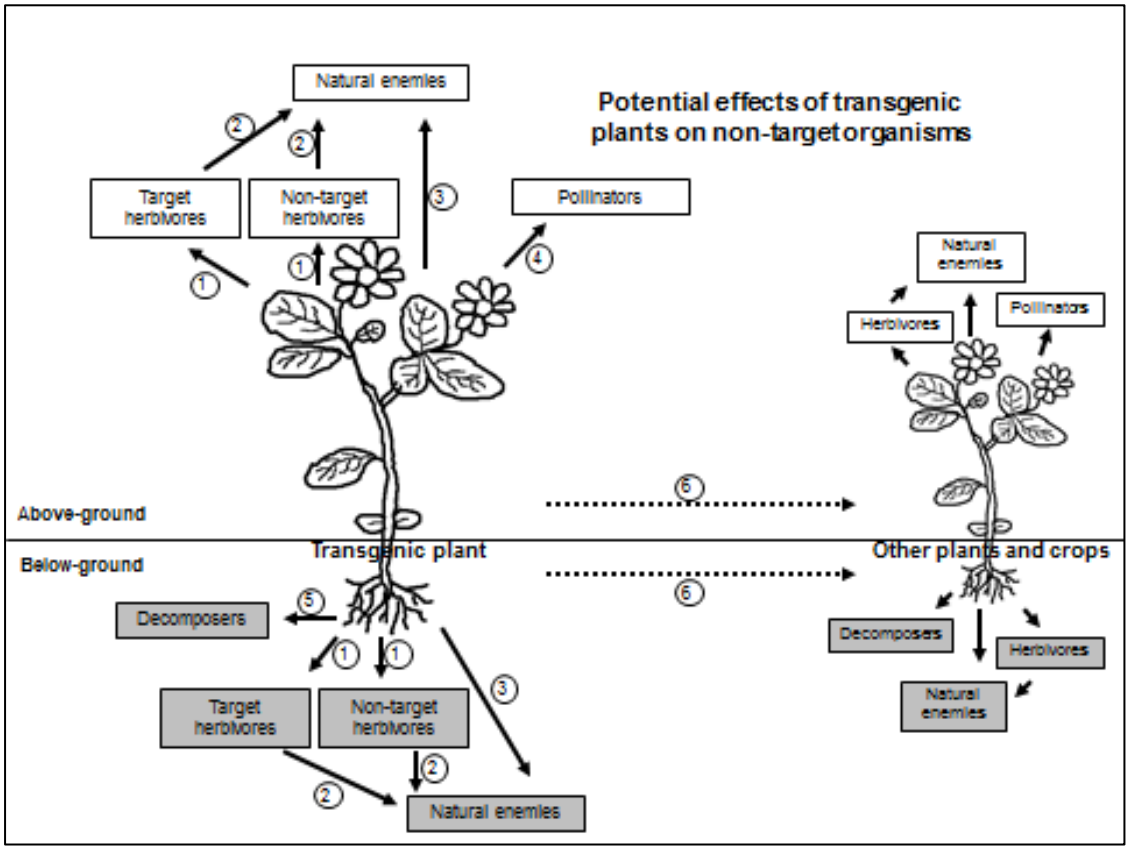


Figure 1 Potential effects of GM crops on non-target organisms (from Kos *et al*, 2009). The GM plant shown here represents plants either expressing insecticidal proteins or with altered VOC emission. The different possible effects of GM crops on target and non-target organisms are shown. (1) Direct effects on target and non-target herbivores arise (a) through plant feeding, (b) in response to VOCs, which can act as oviposition and feeding stimulants or as repellents, (c) by contact with soil exposed to transgenic products through plant decomposition or root exudation. (2) Host- or prey-mediated effects on natural enemies are indirect effects that arise through feeding on a host or prey that is feeding on a GM plant. (3) Direct effects on natural enemies of herbivores arise (a) from feeding on plant material, such as nectar and pollen, (b) in response to VOCs, (c) by contact with soil exposed to the transgenic product. (4) Direct effects on pollinators arise from feeding on nectar and pollen or in responding to VOCs. (5) Direct effects on decomposers arise from feeding on plant material or coming into contact with soil exposed to the transgenic product. (6) Effects on other plants and crops arise, for example, by transfer of pollen or by contact with soil exposed to the transgenic product. Exposure to VOCs from neighbouring plants can prime other plants for increased resistance to herbivores, which in turn could influence organisms associated with the plant or crop.

Baseline establishment

The assessment of non-target effects of a GM crop is commonly performed by comparing the GM line with the non-GM counterpart. However, any parameter or trait that could be analysed, such as the plant-insect interaction, the concentration of certain metabolite(s) or the whole metabolome is prone to variation due to genetic variation within the species. This variation, if explored, can be used to check if the observed changes in the GM crop fit within the boundaries of the (multidimensional) *baseline*. Therefore, we propose to first explore and establish the "*baseline* variation" for a trait within the crop species and its wild ancestors, jointly called "*baseline* genotypes".

Furthermore, growing conditions also have a profound effect on plants. Soil characteristics have a significant impact on plant quality and consequently on aboveground and belowground biota (Bruinsma *et al.* 2003, Kabouw *et al.* 2011, Pineda *et al.* 2010, van der Putten *et al.* 2001). In this document the term '*baseline* variation' is defined as 'the variation in a characteristic or trait observed among a selection of non-GM genotypes (*baseline* genotypes)/ across environmental conditions'. The *baseline* variation refers to the natural variation in any characteristic or trait within a crop species, such as plant-insect interaction, a specific transcript or metabolite level or even the whole transcriptome or metabolome.

As stated before, metabolites are the closest detectable entities to a plant's phenotype and its interaction with biota. The metabolome of a plant is known to be plastic and this reflects the adaptive potential of plants to a changing environment. The metabolome variation among accessions with a different history of adaptation to their environment represents the variation, which we assume does not pose any risk to the environment. Therefore, the difference or the *distance* of a metabolome to the baseline metabolome variation could be used as a first estimate of risk to the environment as a *first-pass* criterion, reducing the volume of non-target ecological assessments in later stages.

Thesis outline

In line with the above, the objective of this PhD research project was to apply new developments in science to address the mentioned shortcomings in the assessment of GM crops using omics platforms. This work should deliver tools that will help to pave the road towards a clear, efficient and sufficient procedure for GM crops risk assessment. As a model for a GM crop, *Arabidopsis thaliana* was employed throughout the whole project because of the straight-forward and fast genetic engineering possibilities. The traits studied are direct and

indirect insect defence traits. Direct insect defence of *Arabidopsis* was improved by introduction of a gene responsible for production of the cry1 protein (Bt toxin against leaf-chewing caterpillars) or overexpression of *MYB28* (encoding a transcription factor regulating aliphatic glucosinolate biosynthesis) into 4 *Arabidopsis* accessions. Indirect defence was improved by the introduction of a construct containing two genes responsible for sequential steps in *(E)*-nerolidol biosynthesis (a plant volatile attractive for natural enemies of insects). This thesis can be used as a methodological model for exploiting omics platforms in the assessment of the potential changes (risks) of genetic modification in plants.

To establish the metabolome baseline, **Chapter 2** examines the natural variation within the metabolome of nine *Arabidopsis thaliana* accessions grown under various environmental conditions and introduces a measure for whole metabolome differences (the *metabolic distance*) between accessions and across environments. This *metabolic distance* is used to characterize *Arabidopsis* accessions and is shown to be biologically relevant for plant-biota interaction. Based on the metabolic distance, three accessions with the most diverged metabolome and an average accession could be selected. Together, they represent the whole metabolome variation in *Arabidopsis* and can hence be used as the *baseline* for assessment purposes.

In **Chapter 3** the effect of indirect defence engineering on organic volatile compounds of the head space and semi-polar secondary metabolites of the aerial parts in *Arabidopsis* GM lines was studied. The GM lines harboured different combinations of transgenes involved in the sesquiterpenoid biosynthesis pathway leading to *de novo* emission of *(E)*-nerolidol. The effect of the emission of (E)-nerolidol on behaviour of a parasitoid wasp (*Diadegma semiclausum*) that was of interest in the sister projects, was investigated. The most efficient combination of pathway genes for indirect defence engineering was selected to be introduced into the genome of the four selected *Arabidopsis* accessions in chapter 2. Furthermore, two of the GM lines were selected for the assessment of non-target effects of the new trait at the transcriptome level (chapter 6).

Chapter 4 describes the development and characterization of *Arabidopsis* lines with an improved indirect defence mechanism. Hereto, mitochondrial *FPS2* (farnesyl diphosphate synthase2) and mitochondrial *FaNES1* (*Fragaria* X *ananassa* nerolidol synthase1) genes responsible for the last two steps of *(E)*-nerolidol biosynthesis were introduced. We transformed the four accessions selected in the study described in chapter 2. Subsequently, the

effect of genetic background and transformation on two tri-trophic interactions as a combination of target and non-target ecological effects was studied. Hereto, the herbivore aphid *Brevicoryne brassicae*, the parasitoid *Diaeretiella rapae* and the predator *Episyrphus balteatus* were exposed to the GM plants and their wild type background.

Chapter 5 describes the development and characterization of *Arabidopsis* GM lines with improved direct defence by overexpression of the gene encoding the MYB28, a positive regulator of the aliphatic glucosinolate pathway. We transformed all four accessions selected in chapter 2 and studied the effects of the genetic background and *MYB28* overexpression on transcript levels of aliphatic glucosinolate pathway genes, the concentration of glucosinolates and the overall impact of the engineering on *Brassicaceae* generalist and specialist caterpillars. As a result of this study we propose a certain class of glucosinolates as being more effective in the specific plant-insect interactions.

Chapter 6 employs the selected GM lines of chapter 3 that had enhanced emission of the sesquiterpenoid (*E*)-nerolidol to explore the potential of transcriptomics for assessment of the substantial equivalence of genetically modified crops. We applied the concept of metabolic distance to the concept of "*hyper-plane* distance" and used it to measure the distance between the transcriptome of the GM lines and the wild type plants described in chapter 3. Here, we propose a *meta data* analysis approach using the publicly available expression arrays data of *Arabidopsis* as natural transcriptome variation and as the *baseline* to which genetically modified plants can be compared. Using this approach we show that the transcriptome changes in the GM *Arabidopsis* are substantially equivalent to the naturally occurring variation and plasticity in gene expression among *Arabidopsis* accessions.

In **Chapter 7** we study the potential to exploit metabolomics in the substantial equivalence assessment of genetically modified plants using the characterized GM lines of *Arabidopsis* from chapter 4 and 5 along with GM line containing the gene responsible for Bt toxin. Based on the results of chapter 2, the GM lines and their wild type background were subjected to GC-MS and LC-MS analyses to gain information about a wide range of metabolites but ensuring redundancy in the data was minimal. Our data analysis shows that some of the transformation events result in lines that are – but also some that are not - substantially equivalent to the baseline with respect to their metabolome. We propose a *baseline* as a boundary for natural metabolome variation and the *metabolic distance* for assessment and statistical judgement of the metabolome differences

In **Chapter 8** the research conducted in three sister projects under the same theme of the ERGO program is jointly discussed. It provides specific guidelines for ecological assessment of non-target effects of GM-plants. In these guidelines, the plant family *Brassicaceae* was used as a model system. The guidelines address four main criteria that are important when evaluating non-target effects of newly developed GM crops: 1) correct selection of the *baseline* variation, 2) experimenting under representative testing conditions, 3) testing the right traits and organisms for ecological assessment and 4) correct interpretation of the data.

References

Baker, J.M., Hawkins, N.D., Ward, J.L., Lovegrove, A., Napier, J.A., Shewry, P.R. and Beale, M.H. (2006) A metabolomic study of substantial equivalence of field-grown genetically modified wheat. *Plant Biotechnology Journal*, **4**, 381-392.

Bruinsma, M., Kowalchuk, G.A. and van Veen, J.A. (2003) Effects of genetically modified plants on microbial communities and processes in soil. *Biology and Fertility of Soils*, **37**, 329-337.

Cheng, K.C., Beaulieu, J., Iquira, E., Belzile, F.J., Fortin, M.G. and StrÃ¶mvik, M.V. (2008) Effect of transgenes on global gene expression in soybean is within the natural range of variation of conventional cultivars. *Journal of Agricultural and Food Chemistry*, **56**, 3057-3067.

Coll, A., Nadal, A., Collado, R., Capellades, G., Kubista, M., Messeguer, J. and Pla, M. (2010) Natural variation explains most transcriptomic changes among maize plants of MON810 and comparable non-GM varieties subjected to two N-fertilization farming practices. *Plant Molecular Biology*, **73**, 349-362.

Craig, W., Tepfer, M., Degrassi, G. and Ripandelli, D. (2008) An overview of general features of risk assessments of genetically modified crops. *Euphytica*, **164**, 853-880.

Dubouzet, J.G., Ishihara, A., Matsuda, F., Miyagawa, H., Iwata, H. and Wakasa, K. (2007) Integrated metabolomic and transcriptomic analyses of high-tryptophan rice expressing a mutant anthranilate synthase alpha subunit. *Journal of Experimental Botany*, **58**, 3309-3321.

EC (2001) Directive 2001/18/EC of the European Parliament and of the Council of 12 March 2001 on the deliberate release into the environment of genetically modified organisms and repealing Council Directive 90/220/EEC: Official Journal of the European Communities, pp. 1-39.

EFSA (2010) Guidance on the environmental risk assessment of genetically modified plants. In *EFSA Journal* European Food Safety Authority (EFSA), Panel on Genetically Modified Organisms, Parma, Italy, pp. 1879.

Ferrari, S., Plotnikova, J.M., De Lorenzo, G. and Ausubel, F.M. (2003) Arabidopsis local resistance to Botrytis cinerea involves salicylic acid and camalexin and requires EDS4 and PAD2, but not SID2, EDS5 or PAD4. *Plant Journal*, **35**, 193-205.

Figueiredo, A., Fortes, A.M., Ferreira, S., Sebastiana, M., Choi, Y.H., Sousa, L., Acioli-Santos, B., Pessoa, F., Verpoorte, R. and Pais, M.S. (2008) Transcriptional and metabolic profiling of grape (Vitis vinifera L.) leaves unravel possible innate resistance against pathogenic fungi. *Journal of Experimental Botany*, **59**, 3371-3381.

Gatehouse, J.A. (2002) Plant resistance towards insect herbivores: A dynamic interaction. *New Phytologist*, **156**, 145-169.

Goodman, C.D., Casati, P. and Walbot, V. (2004) A multidrug resistance-associated protein involved in anthocyanin transport in Zea mays. *Plant Cell*, **16**, 1812-1826.

Gregersen, P.L., Brinch-Pedersen, H. and Holm, P.B. (2005) A microarray-based comparative analysis of gene expression profiles during grain development in transgenic and wild type wheat. *Transgenic Research*, **14**, 887-905.

Groot, A.T. and Dicke, M. (2002) Insect-resistant transgenic plants in a multi-trophic context. *Plant Journal*, **31**, 387-406.

Harrigan, G.G., Martino-Catt, S. and Glenn, K.C. (2007) Metabolomics, metabolic diversity and genetic variation in crops. *Metabolomics*, **3**, 259-272.

Hilbeck, A., Meier, M., Römbke, J., Jänsch, S., Teichmann, H. and Tappeser, B. (2011) Environmental risk assessment of genetically modified plants - Concepts and controversies. *Environmental Sciences Europe*, **23**.

Hirai, M.Y., Sugiyama, K., Sawada, Y., Tohge, T., Obayashi, T., Suzuki, A., Araki, R., Sakurai, N., Suzuki, H., Aoki, K., Goda, H., Nishizawa, O.I., Shibata, D. and Saito, K. (2007) Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 6478-6483.

James, C. (2010) Highlights of "Global Status of Commercialized Biotech/GM Crops: 2010. In *ISAAA Brief* ISAAA.

Kabouw, P., Kos, M., Kleine, S., Vockenhuber, E.A., van Loon, J.J.A., van der Putten, W.H., van Dam, N.M. and Biere, A. (2011) Effects of soil organisms on aboveground multitrophic interactions are consistent between plant genotypes mediating the interaction. *Entomologia Experimentalis et Applicata*, **139**, 197-206.

Kashif, A., Federica, M., Eva, Z., Martina, R., Young, H.C. and Robert, V. (2009) NMR metabolic fingerprinting based identification of grapevine metabolites associated with downy mildew resistance. *Journal of Agricultural and Food Chemistry*, **57**, 9599-9606.

Kok, E.J. and Kuiper, H.A. (2003) Comparative safety assessment for biotech crops. *Trends in Biotechnology*, **21**, 439-444.

Kos, M., van Loon, J.J.A., Dicke, M. and Vet, L.E.M. (2009) Transgenic plants as vital components of integrated pest management. *Trends in Biotechnology*, **27**, 621-627.

Pineda, A., Zheng, S.J., van Loon, J.J.A., Pieterse, C.M.J. and Dicke, M. (2010) Helping plants to deal with insects: the role of beneficial soil-borne microbes. *Trends in Plant Science*, **15**, 507-514.

Poppy, G. (2000) GM crops: Environmental risks and non-target effects. *Trends in Plant Science*, **5**, 4-6.

Rohr, F., Ulrichs, C., Schreiner, M., Nguyen, C.N. and Mewis, I. (2011) Impact of hydroxylated and non-hydroxylated aliphatic glucosinolates in Arabidopsis thaliana crosses on plant resistance against a generalist and a specialist herbivore. *Chemoecology*, **21**, 171-180.

Rulhmann, S., Leser, C., Bannert, M. and Treutter, D. (2002) Relationship between growth, secondary metabolism, and resistance of apple. *Plant Biology*, **4**, 137-143.

Sønderby, I.E., Hansen, B.G., Bjarnholt, N., Ticconi, C., Halkier, B.A. and Kliebenstein, D.J. (2007) A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. *PLoS ONE*, **2**.

van der Putten, W.H., Vet, L.E.M., Harvey, J.A. and Wackers, F.L. (2001) Linking above- and belowground multitrophic interactions of plants, herbivores, pathogens, and their antagonists. *Trends in Ecology & Evolution*, **16**, 547-554.

Yang, T., Stoopen, G., Yalpani, N., Vervoort, J., de Vos, R., Voster, A., Verstappen, F.W.A., Bouwmeester, H.J. and Jongsma, M.A. (2011) Metabolic engineering of geranic acid in maize to achieve fungal resistance is compromised by novel glycosylation patterns. *Metabolic Engineering*, **13**, 414-425.

Chapter 2

Characterization of the natural variation in Arabidopsis thaliana metabolome by the analysis of metabolic distance

Benyamin Houshyani · Patrick Kabouw · Dorota Muth · Ric C.H. de Vos · Raoul J. Bino · Harro J. Bouwmeester

In press, accepted by Metabolomics (2011).

Abstract

Metabolite fingerprinting is widely used to unravel the chemical characteristics of biological samples. Multivariate data analysis and other statistical tools are subsequently used to analyze and visualize the plasticity of the metabolome and/or the relationship between those samples. However, there are limitations to these approaches for example because of the multi-dimensionality of the data that makes interpretation of the data obtained from untargeted analysis almost impossible for an average human being. These limitations make the biological information that is of prime importance in untargeted studies be partially exploited. Even in the case of full exploitation, current methods for relationship elucidation focus mainly on between groups variation and differences. Therefore, a measure that is capable of exploiting both between- and within-group biological variation would be of great value.

Here, we examined the natural variation in the metabolome of nine *Arabidopsis thaliana* accessions grown under various environmental conditions and established a measure for the metabolic distance between accessions and across environments. This data analysis approach shows that there is just a minor correlation between genetic and metabolic diversity of the nine accessions. On the other hand, it delivers so far in Arabidopsis unexplored chemical information and is shown to be biologically relevant for resistance studies.

Introduction

Metabolites in living organisms together constitute the metabolome, which gives specific eco-physiological properties to the organism enabling it to interact with its kin and other species in the ecosystem. The signaling between plants (Belz 2007, Bouwmeester *et al.* 2003, Bouwmeester *et al.* 2007), and other living organisms such as predators and pollinators (Arimura *et al.* 2005, Cheng *et al.* 2007, Hilker and Meiners 2006, Raguso and Pichersky 1999) and defense against biotic agents (Chong *et al.* 2009, Cona *et al.* 2006, Ehrlich and Raven 1964, Kliebenstein *et al.* 2005, Rulhmann *et al.* 2002, Treutter 2005) are among the interactions in which metabolites play a pivotal role. The metabolome of an organism, however, is not a stable entity. Many different sources of variation including genetic, and the biotic and abiotic environment shape the metabolome resulting in a phenomenon that is referred to as metabolome plasticity. A number of approaches have been introduced to study this plasticity (Schripsema 2010). Metabolite fingerprinting or profiling, which is the unbiased global scanning of the metabolome, is being widely used in metabolomics research to unravel the metabolite composition (or metabolite "fingerprint") of biological samples. Multivariate analysis of these metabolite fingerprints is subsequently used to reduce the dimensionality of the fingerprint data to a number of components that explain the maximum variation, i.e. Principal Components (PCs). This is followed by visualization of the data and drawing conclusions based on the clustering of samples (Aliferis and Jabaji 2009, Garcia-Perez *et al.* 2008, Kashif *et al.* 2009).

Due to limitations in human imagination and visualization power, the conclusions about clustering are usually based on just two or rarely three PCs. Hence, although some of the additional PCs may contain relevant biological variation and are thus important for understanding the metabolic relationships between samples, they are usually excluded from the analysis. Even in the case of clear visual separation between clusters of samples along all PCs, there are no methods for the quantification of the distance between these clusters making quantitative comparisons between groups of samples impossible.

The lack of a measure to describe the distance between clusters can be partially circumvented by application of PCA and Hierarchical Cluster Analysis (HCA) simultaneously. However, the experimental design (such as a complete block design), the dimensionality of the data (in case of multivariate data), the relationship between and the ratio of inter- and intra-group variation, different levels of resolution to define the clusters and the method used for the clustering (Almeida *et al.* 2007) make HCA dendrograms not always compatible with the PCA clustering or the treatment structure. This is because HCA not only

takes into account the biological sources of variation (such as genotype and a treatment effect) but also non-biological or undesired sources of variation (such as technical variation and block effects). The non-biological sources of variation may hinder the dendrogram calculation if they have a comparable or greater influence on the variation than the biological sources. Some researchers filter their data statistically and select those data points with significant difference among predefined groups of samples (Boccard *et al.* 2007). Others use an arithmetic mean analysis and make dendrograms with a representative of each predefined group (Kim *et al.* 2009). As a consequence of these filtering techniques, biological information and/or the insight in the possible causes of variation may be lost. Therefore, a measure for the distance between clusters of samples that incorporates both between- and within-cluster biological variation would be of great value. For example, such a measure can be applied to determine the metabolic distance between genotypes, including genetically engineered organisms and their wild type relatives. The metabolic distance can also be used for correlation analysis between genetic and metabolic diversity.

In the present study, we examined the natural variation and plasticity in the metabolome of nine *A. thaliana* accessions in response to four different growing conditions. The objectives were: 1- to show the potential of metabolite fingerprinting and multivariate data analysis to characterize the effect of more than one source of variation on the diversity and plasticity of the metabolome, 2- to show the potential of metabolite fingerprinting and multivariate data analysis to establish the metabolic distance between accessions and different environmental conditions, 3- to estimate the correlation between the genetic and metabolic diversity of the nine accessions. Untargeted metabolite fingerprinting using three types of analytical platforms was employed to produce fingerprints of a wide range of metabolites in the nine accessions. A number of statistical methods were applied subsequently to the fingerprint data. Metabolites that contribute to the differences between the most diverged accessions were tentatively identified and the biological relevance of the observed differences in metabolic profiles of accessions assessed using a number of bioassays with biotic agents.

Materials and methods

Plant Material

Nine accessions of *A. thaliana* (supplementary information Table 1) were selected, based on habitat geographical distribution and variation in volatile headspace profile (Snoeren *et al.* 2010). Accessions were sown in 4 environments: on soil in a climate chamber (CC), a controlled-conditions greenhouse (GH), an uncontrolled-conditions greenhouse (UC) and on hydroponics in the climate chamber (HY). Supplementary information Table 2 lists the environmental conditions.

Seeds were sown in pot soil (heated to 60°C overnight before use; Lentse potgrond BV, Lent, the Netherlands) and placed in a climate chamber. Seedlings at stage 1.02 (Boyes *et al.* 2001), with 2 rosette leaves >1 mm, were transplanted to plastic containers (12 cm diameter, 4 seedlings per container) filled with the same soil. Containers were distributed randomly on shelves in CC, GH and UC and watered twice a week. For HY, seeds were sown on rock wool units fixed on a floating structure on the hydroponic solution (Tocquin *et al.* 2003). The rock wool absorbed the required water and nutrition for germination and growth. The hydroponics solution was refreshed weekly and aerated continuously.

Stage 3.70 to 3.90 plants (Boyes, Zayed, Ascenzi, McCaskill, Hoffman, Davis and GÃ¶rlach 2001) with 70-100% rosette formation were cut from the surface of the soil or rock wool. Roots were cut from the hypocotyl at the rock wool subsurface. Six biological replicates of shoot material were used for each accession in each environment. Each shoot biological replicate consisted of a pool of 4 plants that had been growing in the same pot in CC, GH and UC or were selected randomly in HY. For hydroponically grown roots, 4 biological replicates were used that consisted of the roots of 6 randomly pooled plants.

All biological replicates were flash frozen in liquid nitrogen, lyophilized for 72 hrs, homogenized in a steel jar containing two steel balls shaken at 20 s^{-1} by a MM300 mixer mill (Retsch) for 45 s at 21°C and ambient humidity. Samples were stored dry at 4°C until extraction for chemical analysis.

Extract preparation and LC-TOF-MS analysis

The protocol of Keurentjes *et al.* (Keurentjes *et al.*) was followed for extraction of semi-polar metabolites with some modifications. Fifty mg of shoot or 12.5 mg of root material, both lyophilized and homogenized, were mixed with 2 ml (for shoots) or 0.5 ml (for

roots) of ice-cold 75% methanol acidified with 0.1% (v/v) formic acid. After vortexing for 5 s, sonication for 15 min and centrifugation (2500 rpm) for 10 min, the extracts were filtered through syringe filters (Minisart SRP 4, 0.45 µm, Sartorius Stedim Biotech) and collected in glass vials. The filtered extract (150 µl) was transferred to a glass insert (300 µl) in a screw neck glass vial (1.5 ml) and then analyzed. In both positive and negative mode analyses, shoot samples were grouped in 3 sample blocks each containing 2 randomly selected biological replicates of each accession in each environment. A mixed sample of the 9 accessions was passed through the same extraction procedures and used as technical replicates. They were analyzed at the beginning, the end and as every 15th sample in the injection sequence.

Liquid chromatography was performed on a Waters Acquity Ultra Performance Liquid Chromatography system (Waters, Milford, MA, USA). Five µl of extract was injected automatically on an Acquity UPLC BEH C_{18} column (150 x 2.1 mm i.d., 1.7 µm particle size) (Waters), held at 50°C with a mobile phase flow of 0.4 $ml.min^{-1}$. The mobile phase consisted of water and acetonitrile containing 20 mM formic acid. The gradient applied started at 100% water for 0.5 min and subsequently changed to 10% acidified acetonitrile in 1 min, then rose linearly to 25% in 4 min, 65% in 3.5 min and 95% in 5 min, which was held for 6 min. Before the next run the column was equilibrated with starting conditions for 3 minutes.

Compounds eluting from the column were detected by a Waters LCT Premier TOF MS (Waters, Milford, MA, USA) equipped with a Z-spray interface and an electrospray ionization (ESI) source. The analysis was performed in both negative and positive ion modes in the range of *m/z* 80 to 1000 in separate runs, using a scan time of 200 ms. The parameters of the source were: desolvation gas temperature of 400 °C, nitrogen gas flow of 500 $l.h^{-1}$, capillary spray voltage of 2.5 keV, source temperature of 120°C, cone voltage of 50 eV, nitrogen gas flow of 50 $l.h^{-1}$, and aperture 1 voltage of 8 eV. The mass spectrometer was calibrated with 5 mM sodium formate in iso-propanol/water (9:1). A 1 $\mu g.ml^{-1}$ leucine enkephalin solution in acetonitrile/water (1:1) containing 0.1% formic acid, infused at a flow rate of 0.02 $ml.min^{-1}$, was used as a lock mass to continuously recalibrate the mass accuracy in both electrospray modes. The sampling rate of the lock mass solution was 0.4 s every 2 s. MassLynx software version 4.1 (Waters) was used to control the instruments and for data analysis.

Extract preparation and GC-TOF-MS analysis

Ten mg of shoot or 5 mg of root material, both lyophilized and homogenized, were weighed for extraction of polar metabolites. The instrument and protocol described in Fu *et al.* (2009) were used for extraction, derivatization and data acquisition by GC-TOF-MS with

minor changes. Shoot samples were grouped in three sample blocks, each containing two randomly selected biological replicates of each accession in each environment. Derivatized extracts (25 µl) were injected (2 µl) with an Optic3 injector (ATAS) at 70^{o}C with a gradient of 6^{o}C.s^{-1} to 240^{o}C. A split flow of 10 (1 ml:11 ml) was used for shoot or 5 (1 ml:6 ml) for root material with a column flow of 2 ml.min^{-1} in a GC6890N gas chromatograph (Agilent Technologies) on a ZB-50 capillary column (30 m x 0.32 mm i.d., 0.25 µm DF; Phenomenex). The column temperature was 70^{o}C for 2 min with a gradient of 10^{o}C.min^{-1} to 310^{o}C and a final time of 3 min. The GC was coupled to a Pegasus III time-of-flight mass spectrometer (LECO) and compounds were detected at a scanning rate of 20 spectra per second (*m/z* 50-600).

Data processing and analysis

The data of all analytical platforms were processed using MetAlign software (Lommen 2009) for peak detection and alignment of the data points. An in-house script called MetAlign Output Transformer (METOT; Plant Research International, Wageningen) was used for data filtration, missing value replacement, and data quality and analytical technique reproducibility verification. The post-METOT data matrix was subjected to multivariate mass spectra reconstruction (MMSR) for data size reduction and putative compound mass spectrum reconstitution (Tikunov *et al.* 2005). MMSR relates thousands of ion fragments in a chromatogram to their parental metabolites by clustering them based on retention time and peak intensity pattern across samples into reconstructed metabolites. These mass clusters were used for further analyses and putative identification of metabolites.

For multivariate data analysis, the intensity values of reconstructed metabolites were normalized by the dry weight of the sample. Subsequently, metabolite intensities of each sample were ^{10}log-transformed and scaled by dividing by the standard deviation of the metabolite intensities of the corresponding sample. An integrated dataset was constructed by combining the shoot data of three analytical platforms. Values in the integrated dataset were derived in the same manner with an additional scaling by the standard deviation of samples for a reconstructed metabolite subsequent to dry weight normalization. The ordination diagrams in CANOCO (ter Braak 1988) (Biometris, Wageningen, NL) were used to visualize the variation in the sample profiles. Detrended Correspondence Analysis (DCA) was used to check the gradient length (L) of the explanatory variables (accession and environment) and accordingly choose between the linear (L<4) or unimodal (L>4) ordination techniques (Smilauer 2003). In addition to the first two ordinates (or PCs), the third and fourth ordinates were incorporated into the analysis if they individually explained more than 10% of the total variation and the technical replicates were grouped along those ordinates on the scores plot. Partitioning of

explanatory variable effects such as accession, environment and their interaction on the observed variation was performed and tested statistically (*P*-value < 0.05) by Monte-Carlo Permutation (MCP) test using the Partial Redundancy Analysis (RDA) function of CANOCO (Smilauer 2003).

CANOCO scores and loading plots of shoot datasets were superimposed separately for three analytical methods which resulted in three biplots. Subsequently, 10 reconstructed metabolites that fitted more than 55% into the ordination space and showed to be accession-specific on the biplot were selected for further analysis.

Correlation between genetic and metabolic distance

A matrix of 149 genome-wide distributed SNP (single nucleotide polymorphism) markers from the Borevits lab (http://borevitzlab.uchicago.edu) was used to calculate the genetic distance between the nine accessions. The "Jukes & Kantor" distance and complete linkage clustering were determined using TREECON v.1.3b (Van De Peer and De Wachter 1997).

To compute the inter-accession metabolic distances, the inter-sample Euclidean distances in an ordination diagram were examined by taking the sample scores on the selected ordinates of the PCA scores plots (Kabouw *et al.* 2009). The inter-sample Euclidean distance matrices were computed for all platforms. The resulting matrices were used in an ANOSIM (analysis of similarity) by the program PAST (Hammer *et al.* 2001) to calculate the R-values as a measure for the metabolic distance between accessions. The Pearson correlation coefficient between two matrices (r) was determined by a Mantel test (10000 permutations).

In silico identification of reconstructed metabolites

Putative identification of selected metabolites from the shoot LC-MS datasets was done through the following steps: 1- Elimination of adduct ions; From all *m/z* ratios reconstructed in the centrotypes only mono-isotopic signals were selected and the respective neutral mass of the molecule was calculated using the Mass Spectrometry Adduct Calculator (http:// fiehnlab.ucdavis.edu/staff/kind/Metabolomics/MS-Adduct-Calculator/). 2- Molecular formula assignment; Putative molecular formulas for an accurate mass were predicted by the elemental composition tool of MassLynx (Waters) with a 5 ppm tolerance. 3- Molecular formula screening; The Seven Golden Rules software (http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/Software/) was used for heuristic filtering of the obtained molecular formulas (Kind and Fiehn 2007). Remaining possible

molecular formulas were scored by the software according to their isotopic abundance error. 4- Molecular formula ranking; The 5 highest ranking molecular formula were prioritized based on prior identification in *A. thaliana*, Brassicaceae species or other plant species, presence in the Dictionary of Natural Products (http://dnp.chemnetbase.com) and their score.

The Golm Metabolome Database (GMD@CSB.DB MSRI, http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/msri/gmd_msri.html), NIST library and an in-house mass spectral database for GC-TOF-MS were used to putatively identify reconstructed metabolites from GC-TOF-MS analysis, using NIST MS Search v.2.0. Both mass spectra and retention indices of the reconstructed metabolites were used to search for putative candidate metabolites already reported in *A. thaliana*. Matching factor (MF) and reverse matching factor (RMF) (Davies 1998) were exploited to select the best matching metabolite identity.

Metabolite identification with HPLC-PDA-QTOF-MS/MS system

Metabolites detected by LC-TOF-MS were further annotated using an HPLC- PDA-QTOF-MS/MS system from Waters (De Vos et al, 2006). Ten µl of the extracts were automatically injected on an HPLC Luna C18 analytical column (150 x 2.0, 3 mm) (Phenomenex). The chromatographic phases composition were water/formic acid (0.1% v/v) (A) and acetonitrile/formic acid (0.1% v/v) (B). The separation was performed at 40°C with a flow of 0.19 $ml.min^{-1}$ in a gradient starting with 5% of B which linearly increased to 75% in 45 min, than up to 90% in 2 min and continued isocratic for 5 min. The column was then equilibrated for 16 minutes under starting conditions.

The HPLC was linked to a PDA detector (Waters 2996) and a QTOF Ultima mass spectrometer (Waters Corporation). The ionization source parameters were: capillary voltage 2.75 keV, cone voltage 35 eV, source temperature 120°C and desolvation temperature 250°C. Cone gas and desolvation gas flows were 50 and 600 $L.h^{-1}$, respectively. MS/MS measurements were made with 0.40 s of scan duration and 0.10 s of interscan delay with increasing collision energies according to the following program: 5, 10, 15, 30, 50 V (ESI positive), or 10, 15, 25, 35, and 50 V (ESI negative). Leucine enkephalin was used as a lock mass and was continuously sprayed into a second ESI source using an LKB Bromma 2150 HPLC pump, and sampled every 10 s .

Bioassays

Accessions An-1, Cvi, Eri and Col-0 were selected to conduct bioassays. Inoculation by the powdery mildew pathogen *Oidium neolycopersici* and assessment of infection was

performed according to Bai *et al.* (2008). Inoculation by the downy mildew pathogen *Hyaloperonospora arabidopsidis* isolates Emoy2, Waco9 and Cala2 was done and infection assessed according to Van Damm *et al.* (2009). For *Botrytis cinerea* inoculation the protocol of Ferrari *et al.* (2003) was used with minor modifications. Plants were placed in darkness for 24 hrs after inoculation and subsequently kept at 9 hrs photoperiod. Scoring was done 3 d after inoculation by visual determination of the area of the lesions on the inoculated leaves.

Western flower thrips, *Frankliniella occidentalis* (Pergande), was reared according to De Vos *et al.* (2005). Sixteen thrips were transferred to each of twelve Petri dishes (9 cm) used as replicates. Each Petri dish contained two detached leaves of each accession on a 1.5% agar medium (Technical No.3). Every two leaves of the same accession were randomly distributed on four sides of the Petri dish keeping the same distance from each other. The number of thrips present on each accession was counted at 0.5, 1, 2, 3, 4, 5, 6, 21, 22, 23 and 24 hrs. The average number of observed thrips on each accession in the last four hours was used to compare the thrips attraction of accessions by generalized linear model analysis using PASW statistics 17.

Results and discussion

The phylogenetic relationship between accessions

Complete linkage clustering using 149 genome-wide SNPs resolved the phylogenetic relationship between the 9 *A. thaliana* accessions of this study (supplementary information Fig. 1). The level of confidence was estimated using a bootstrap of 500 replications. A separation between An-1, C-24, Cvi and Kyo-1 versus WS, Ler, Kond, Col-0 and Eri occurred at a low stringency level and formed two major clades with relatively close genotypes (accessions). Accessions An-1, Eri, Col-0 and C-24 diverged earliest from the rest of the genotypes in the clades indicating a larger genetic distance between them and the rest of the clade members.

Metabolome analyses

Shoot samples from all environments and root samples from hydroponics were subjected to three different profiling methods: LC-TOF-MS in positive and negative mode, and GC-TOF-MS. The number of entities (masses or reconstructed metabolites) after each step of the data processing workflow with MetAlign, METOT and MMSR is given in supplementary information Table 3. Reproducibility of the analyses and data processing were verified by two approaches: The approach of Vorst *et al.* (2005) and a PCA approach. Supplementary information Fig. 2 illustrates the methods and graphs of the first approach using root data of LC-MS negative mode - after MetAlign and METOT preprocessing - as an example. The amplitude scatter plots of all mass peaks of two technical replicates were made and compared with the same plot of two biological replicates (supplementary information Fig 2A and B as an example). The scatter plots showed that there is a close linear relationship between all signal intensities of any two technical replicates of an analytical technique (supplementary information Fig. 2A). For biological replicates there is also a linear relationship albeit with lower correlation (supplementary information Fig. 2B). The scatter plots also showed absence of hypo-alignment, as is concluded from the absence of satellite clouds around the diagonal axes of the aligned masses in the scatter plot. These clouds can be due to peaks of one technical replicate being aligned with noise peaks in the other technical replicate, instead of with its corresponding mass peak (Vorst, de Vos, Lommen, Staps, Visser, Bino and Hall 2005). Their absence indicates that masses were aligned efficiently and/or that misaligned masses were eliminated by METOT. Supplementary information Fig. 2C shows that more than 80% of the ion masses are present in at least five of the six technical replicates.

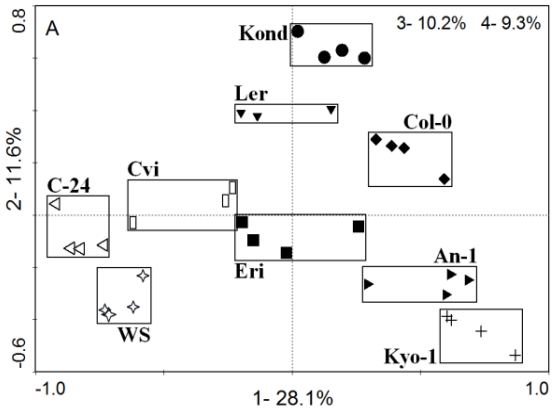
Around 80% of the ions that are detected in all six technical replicates show a variability in their measured mass of less than 5 ppm (supplementary information Fig. 2D) and the amplitude variation for more than 90% of the masses present in all six technical replicates was less than 20% (supplementary information Fig. 2E).

For the second approach, we visualized the variation in metabolite fingerprints of the technical and biological replicates by PCA (supplemental information Fig 2F, as an example). The technical replicates clustered more closely together than the biological replicates in the examined PCs showing that there was relatively minor variation due to extraction and instrument artifacts compared with the biological variation.

Visualization and quantification of the variation due to source effects

The gradient of the explanatory variables (accession and environment) was computed by DCA for all datasets. All of the gradients were short (L<4) suggesting a linear metabolite fingerprint response to the gradient of the underlying variation source(s). Accordingly, PCA was used for all data visualizations and analyses in which one or more independent gradients (PCs) represent predictors for fitting the regression model.

With the exception of the GC-TOF-MS data, all PCAs of shoot datasets showed separate clustering of HY samples from samples grown in the other environments along PC1 and PC2 (see supplementary information Fig. 2F as an example). Clustering of all hydroponically grown genotypes on one side of PC1 implies that there is a pronounced effect of this growth condition compared to the other sources of variation within the shoot LC-MS data. In other words, this environment has a stronger effect on metabolite composition than the other environments and its effect is stronger than the genotypic effect. Within the HY environment, however PCA showed clear clustering of the accessions along PC1 and PC2 in both LC-MS (Fig. 1 as an example) and GC-TOF-MS analyses (data not shown). In both root LC-MS modes, about 40% of the variation between accessions was explained by the first two PCs. In positive mode, all accessions were separated along PC1 and PC2 with C-24, Kyo-1 and Kond positioned at the extremes of the two dimensional PCA plot (Fig. 1A). In negative mode some accessions clustered together suggesting a close metabolic relationship (An-1 and Kyo-1, Eri and Ler, and WS and C-24) (Fig. 1B). PC3 (10.2 and 11.9% for positive and negative mode, respectively) and PC4 (9.3 and 10.2%) explained a similar amount of variation as PC2 (11.6 and 12.8%). Therefore, PC3 and PC4 were also used for ANOSIM and metabolic distance calculation, but only if the technical replicates did not separate along these PCs as much as the biological replicates.



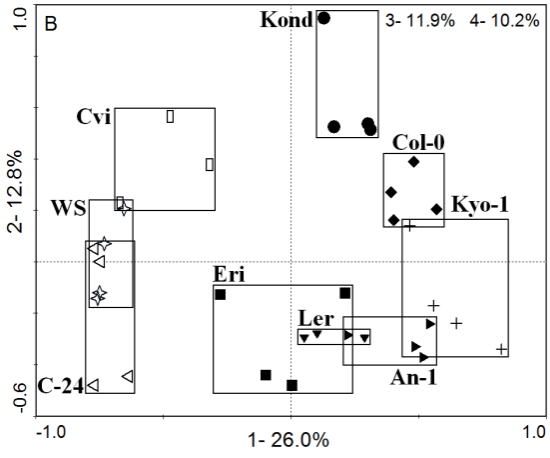


Figure 1 PCA scores plots of the root metabolite profile of nine accessions grown in hydroponics, analyzed by LC-MS in positive mode (**A**) and negative mode (**B**). Numbers along the axes indicate the PC number and percentage of explained variation. Boxes approximate the boundaries of within accession variation and illustrate clustering of samples belonging to the accession

Multivariate statistical representation of metabolite fingerprints by using PCA as shown in Figure 1 allows the visualization of the natural variation and plasticity in the metabolome. However, information about the magnitude of the variation caused by a known factor is difficult, if not impossible, to obtain in this way. Therefore, RDA was used to estimate the effect size of a single or multiple source(s) of variation such as accession or accession and environment on the metabolome in a multi-factorial experiment. Furthermore, in partial RDA, one or more of the explanatory variables can be defined as cofactor to remove the associated effect from the solution of the ordination model. For the root dataset only a single source of variation (accession) was defined and the explained variation in each dataset was calculated accordingly (supplementary information Table 4). Accession explained a bigger portion of the variation in both LC-MS datasets (79.0% and 80.6% in positive and negative mode, respectively) compared with the GC-TOF-MS dataset (53.9%). For shoot

datasets, first the effect of sample block was removed by partial RDA. This resulted in environment and accession explaining together 38.8%, 35.2% and 63.0% of the total variation in the GC-TOF-MS, LC-MS positive mode and LC-MS negative mode datasets, respectively. Using sample block and environment or sample block and accession as cofactor group in RDA, the explained variation by accession and environment was calculated, respectively (supplementary information Table 4). In conclusion, environment had a larger influence than the accession on the LC-MS determined metabolic variation in the shoot. For the GC-TOF-MS platform, which mainly detects polar primary metabolites, the environment effect was smaller than the accession effect.

Likewise the environment-accession interaction was defined as an explanatory variable and sample block, environment and accession main effects were added to the cofactor group. Environment-accession interaction explained a minor part of the total variation in GC-TOF-MS and LC-MS positive mode datasets, respectively, while it had no significant effect on the variation in the LC-MS negative mode dataset (supplementary information Table 4). In all datasets, the environment caused a change in the position of samples (Fig. 2C as an example with LC-MS negative mode dataset). This change is an indication of the capacity of each accession to respond to perturbations in the environment (metabolome plasticity). However, plant species might differ in the degree of phenotypic plasticity when exposed to the same environmental change (De Jong 2005). The amount of variation explained by the accession-environment interaction can be a measure for the difference between accessions in their ability to respond to a perturbation. The fact that the accession-environment interaction explained only little metabolite variation suggests that the metabolic responses of the different *A. thaliana* accessions towards changes in the environment is quite similar.

Metabolic distance between accessions

The PCA scores plots show that biological replicates of the same accession cluster together (Fig. 1). In many cases, however, the large within-accession variation (variation between biological replicates; indicated by boxes in Fig. 1) or small between-accession metabolic variation caused merging of samples belonging to different accessions along one or more PCs. The merging of clusters happened along different PCs (also along PC3 and PC4, which are not shown in Fig. 1). To take the relative position of samples to each other along all PCs, the scores of the samples on these PCs were exploited as sample properties. A matrix containing all these scores was used to calculate the Euclidian distance between samples (inter-sample distance). Using the inter-sample distance matrix, an R-value was assigned by the ANOSIM permutation test. This R-value is a geometric function that describes the

Table 1 R-values (metabolic distance) obtained by the ANOSIM permutation test on inter-sample Euclidian distances of root metabolite profiles obtained by LC-MS analysis in positive mode (values above the diagonal) and negative mode (values below the diagonal), based on the first four principle components of PCA, together explaining 59.3% and 60.9% of the variation, respectively. The numbers in parentheses indicate the ranking of the corresponding accession with regard to the average of its distances with the other accessions

LC-MS positive mode

LC-MS negative mode	An-1	Col-0	C-24	Cvi	Eri	Kond	Kyo-1	Ler	WS	Average -
An-1		0.34	0.20	1.00	0.44	0.69	0.07	0.48	0.55	0.65 (2)
Col-0	0.73		0.39	1.00	0.79	0.99	0.51	0.98	0.76	0.46 (6)
C-24	0.35	0.08		1.00	0.31	0.60	0.25	0.46	0.42	0.20 (9)
Cvi	1.00	1.00	0.61		0.83	1.00	1.00	1.00	1.00	0.85 (1)
Eri	0.94	0.35	0.13	1.00		1.00	0.70	0.37	1.00	0.59 (4)
Kond	0.50	0.26	0.08	0.33	0.42		0.57	1.00	0.65	0.31 (8)
Kyo-1	0.03	0.34	0.04	0.87	0.32	0.27		0.48	0.30	0.32 (7)
Ler	1.00	0.57	-0.02	1.00	0.87	0.33	0.30		1.00	0.63 (3)
WS	0.68	0.36	0.30	1.00	0.72	0.30	0.36	1.00		0.59 (4)
Average +	0.47 (8)	0.72 (4)	0.45 (9)	0.98 (1)	0.68 (6)	0.81 (2)	0.49 (7)	0.72 (3)	0.71 (5)	

distance between pairs of a *priori* groups of samples belonging to two accessions in the hyperplane. R-values can range from +1 to -1, with +1 indicating the maximal divergence of two accessions and -1 indicating that one group is the child of the other. R=0 occurs if the position of both accessions in the hyperplane completely overlaps. Table 1 shows the pair wise R-values for the root metabolite fingerprints obtained by LC-MS negative and positive mode analyses. Based on these R-values, for example accessions Kyo and An-1 had a very similar metabolome in both analytical methods ($R_{Positive\ mode}$=0.07 & $R_{Negative\ mode}$=0.03) (Table 1). Ler and Kond, on the other hand, were maximally different for the positive mode metabolic profile, whereas for the negative mode metabolic profile they were less different ($R_{Positive\ mode}$=1.00; $R_{Negative\ mode}$=0.33) (Table 1). On the contrary, Ler and An-1 were maximally different for the negative mode metabolic profile and less different for the positive mode metabolic profile ($R_{Positive\ mode}$=0.48; $R_{Negative\ mode}$=1.00) (Table 1).

The use of more than two PCs in the R-value calculation resulted in quite a different interpretation of the metabolic distance between some accessions compared with the two dimensional PCA plot. As an extreme example, the root samples of accessions C-24 and Col-0 clearly separated along the first PC in both LC-MS positive and negative mode (Fig. 1). However, the R-value of the negative mode analysis showed that these accessions are metabolically more similar in the hyperplane than what can be observed on a plain PCA plot

($R_{Negative\ mode}$=0.08) (Table 1). The distance is larger in positive mode (0.39) but still much smaller than for example the distance between C-24 and Cvi which in Fig. 1 are closer together than C-24 and Col-0.

Ranking of the average distance of one accession to all other accessions was used as a measure for its metabolic divergence from the rest of the accessions. Accessions Cvi and C-24, for example, were the most and the least diverging accessions, respectively, with regard to their root metabolite profiles obtained by both LC-MS modes ($R_{Positive\ mode\ average}$ =0.98 & $R_{Negative\ mode\ average}$ =0.85 for Cvi and $R_{Positive\ mode\ average}$ =0.45 & $R_{Negative\ mode\ average}$ =0.20 for C-24) (Table 1). Average metabolic distance depended on the analytical method used as was exemplified by accessions An-1 and Kond that changed from second to eighth and eighth to second as a result of the analytical method. This indicates the importance of the use of several different analytical platforms for elucidation of metabolic relationships between accessions .

Metabolic distance in complex datasets

So far, we analyzed the metabolic distance for genetic differences only. It would be of interest to look also at the metabolic distance between genotypes under different treatments. However, this implies the involvement of more than one source of variation and that makes such analyses complicated. Partial analysis, however, allows for reduction of the complexity by removing the effect of an undesired source of variation from the dataset. In Figure 2, three PCA plots of the shoot LC-MS negative mode data are shown (216 samples). If none of the sources of variation (sample block, accession, environment) was excluded from the model, only separate clustering of hydroponically grown samples could be observed (Fig. 2A). Hence, the first PC explains only the considerable change in metabolite fingerprint in the HY samples compared with the other environments.

Partial PCA of the LC-MS negative mode dataset with sample block and environment or sample block and accession as cofactor removed the cofactors' effect from the dataset and resulted in cluster formation based on the remaining sources of variation (Figs 2B & C). Hence, partial analysis reduced the complexity of the datasets in which more than one known variation source (accession and environment in this case) was responsible for the metabolic differences. Moreover, partial analysis enabled having more biological replicates for each accession or environment by combining the data from different environments or accessions, respectively, hence giving better insight in the within-accession or within-environment biological variation.

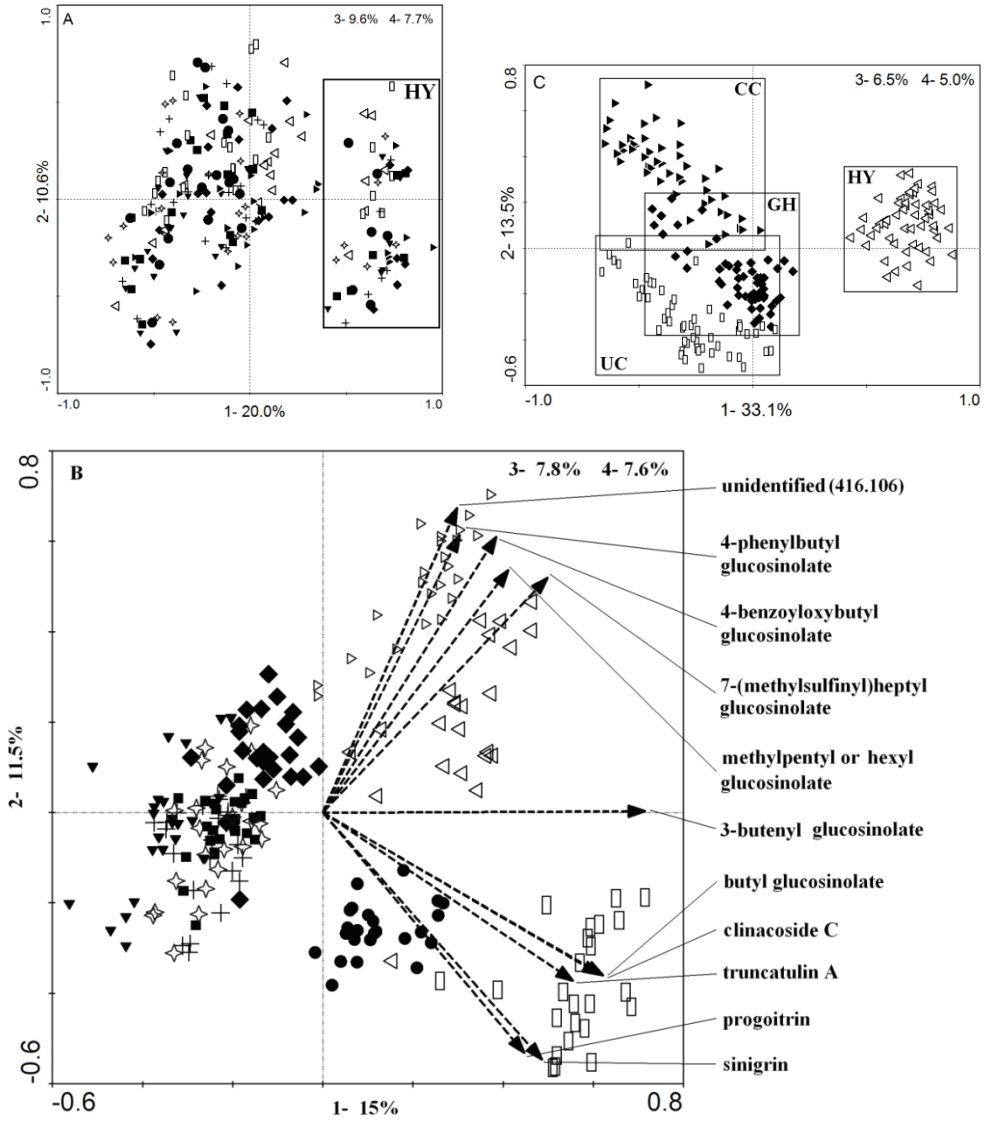


Figure 2 PCA plots of the metabolite profiles of the shoot of nine accessions grown in four environments, analyzed by LC-MS in negative-mode. **A** scores plot. **B** partial PCA biplot (superimposed scores and loadings plots) with environment and sample block as cofactor. Dashed arrows represent the 11 metabolites that more than 55% of their influence was represented by the first two PCs. Numbers along the axes indicate the PC number and percentage of variation explained. The accurate mass is given in parentheses for unidentified masses. **C** partial PCA scores plot with accession and sample block as cofactor. Boxes approximate the boundaries of within environment variation and illustrate clustering of samples belonging to the environment. Accessions in A & B: ▹An-1, ◆ Col-0, ◃ C-24, ▯ Cvi, ■ Eri, ● Kond, + Kyo-1, ▼ Ler, ✧ WS; Environments in C: CC climate chamber , GH controlled-conditions greenhouse, UC uncontrolled-conditions greenhouse, HY hydroponics in the climate chamber

Subsequently, the scores of the samples on PC 1 to 4 in these partial PCA analyses were used to calculate the distance between *a priori* groups (accessions or environments) of samples in the hyperplane and successively the average distance as a measure for metabolic divergence (Table 2 and supplementary information Table 5). The total of the average R-values provides a measure for the overall metabolic distance of an accession or environment from the others across all analytical methods. Among all accessions and across all analytical methods, Cvi (R_{total} = 4.93; ranking 1st) was the metabolically most diverged accession (Table 2). Accessions Kyo-1 (R_{total} = 2.41; ranking 7th) and Eri (R_{total} = 2.63; ranking 8th) had the lowest total R-value which makes them the least metabolically diverged accessions or in other words metabolically average accessions that in most of the analyses located close to the center of the hyperplane or in between most of the other accessions (Table 2 and Fig.2B).

Table 2 Average R-values (metabolic distance) of accessions in different analytical methods obtained by the ANOSIM permutation test on the inter sample distances. Numbers in parentheses show the ranking of the accession within the row

	An-1	Col-0	C-24	Cvi	Eri	Kond	Kyo-1	Ler	WS	PC[1]	Variance explained[2]
Root LC-MS+	0.47 (8)	0.72 (3)	0.45 (9)	0.98 (1)	0.68 (6)	0.81 (2)	0.49 (7)	0.72 (3)	0.71 (5)	4	0.593
Root LC-MS -	0.65 (2)	0.46 (6)	0.20 (9)	0.85 (1)	0.59 (4)	0.31 (8)	0.32 (7)	0.63 (3)	0.59 (4)	4	0.609
Root GC-MS	0.57 (5)	0.60 (4)	0.65 (3)	1.00 (1)	0.34 (8)	0.37 (7)	0.38 (6)	0.38 (6)	0.73 (2)	3	0.535
Shoot LC-MS+	0.38 (3)	0.29 (5)	0.53 (2)	0.58 (1)	0.25 (9)	0.27 (6)	0.32 (4)	0.25 (9)	0.25 (9)	3	0.239
Shoot LC-MS-	0.89 (2)	0.65 (5)	0.78 (4)	0.97 (1)	0.54 (8)	0.85 (3)	0.57 (7)	0.59 (6)	0.51 (9)	3	0.343
Shoot GC-MS	0.29 (6)	0.23 (9)	0.31 (4)	0.55 (1)	0.23 (9)	0.46 (2)	0.33 (3)	0.30 (5)	0.28 (7)	4	0.441
Total	3.25 (2)	2.95 (4)	2.92 (5)	4.93 (1)	2.63 (7)	3.07 (3)	2.41 (8)	2.87 (6)	3.07 (3)		

[1] the number of principal components (PC) used to calculate the Euclidean distances for ANOSIM

[2] % variation explained by the PCs used

Analogously among environments across all analytical methods, Hydroponics (R_{total} = 2.43) was the most distant environment from the others with the largest total R-value (Supplementary Information Table 5) as was also concluded from Fig. 2C. The total R-value of greenhouse (R_{total} = 1.63) samples was the smallest, indicating that this environment was the most "average" growth condition. (Supplementary Information Table 5 and Fig. 2C).

Correlation between metabolic distance matrices

The metabolic distance matrices (matrix of R-values) can be used also to analyze pattern similarity between analytical methods. A Mantel test showed a significant correlation between the metabolic distance matrices of shoot LC-MS positive and negative mode (r = 0.70 , *P*-value < 0.01) and between root LC-MS positive and negative mode (r = 0.47 , *P*-value < 0.05) (Table 3). Interestingly, there was also a significant correlation between the metabolic distance matrices of shoot LC-MS negative mode and GC-TOF-MS (r = 0.61 , *P-value* < 0.05) (Table 3).

Table 3 Mantel statistics "r", for the correlation between different datasets. *P*-value calculated by 10000 permutations

	r	***P*-value**
Root LC-MS+ Vs Root LC-MS-	0.47	0.04*
Root LC-MS+ Vs Root GC-MS	0.24	0.19
Root LC-MS- Vs Root GC-MS	0.39	0.08
Shoots LC-MS+ Vs Shoot LC-MS-	0.70	0.004*
Shoots LC-MS+ Vs Shoot GC-MS	0.38	0.08
Shoots LC-MS- Vs Shoot GC-MS	0.61	0.01*
Hyd[1]. shoots LC-MS- Vs Hyd roots LC-MS-	0.18	0.25
Hyd. shoots LC-MS+ Vs Hyd roots LC-MS+	0.29	0.08
Hyd. shoots GC-MS Vs Hyd roots GC-MS	0.39	0.10

A significant correlation between inter-accession metabolic distance matrices for the analytical methods indicates that the metabolite profiles obtained with these methods give a similar picture of the relationship between the accessions. This correlation does not imply that there is similarity in the measured metabolites by those methods. The lack of a significant correlation between some analytical methods, on the other hand, indicates that metabolic distances can be also method dependent and that the analytical methods are complementary, rather than redundant, in covering the metabolome. This is confirmed by the fact that the ranking of accessions across uncorrelated metabolic distance matrices varies more than across correlated matrices (Table 2). This also implies the importance of using different analytical methods for metabolome characterization. Table 3 provides some guidelines for the choice of analytical methods that avoid generation of redundant analytical data for metabolic relationship dissection. A combination of LC-MS positive mode and GC-TOF-MS analysis of derivatized polar extracts would, at least for *A. thaliana* accessions of this study, give reasonable coverage of the metabolome while avoiding too much redundancy.

No significant correlation was observed between metabolic distance matrices of shoot and root, in all three analytical methods (Table 3). The lack of a similarity between metabolite profiles of shoot and root has been shown in a number of studies (Kabouw, Biere, Van Der Putten and Van Dam 2009, Van Dam *et al.* 2009). However, here we show that there is also a lack of correlation between the matrices of metabolic distances of accessions based on root and shoot. This complements prior studies (Kabouw, Biere, Van Der Putten and Van Dam 2009, Van Dam, Tytgat and Kirkegaard 2009) and indicates that the metabolic relationship between accessions is dependent on the tissue under study.

Chemical relevance of metabolic distance

Key in estimating the metabolic distance between *a priori* groups of samples (accessions or environments) is the quantity of the analytical information and relevance of the calculated distance. The metabolite profiling using three analytical platforms was aimed at increasing the quantity of the information from different classes of metabolites. A metabolite identification strategy was followed to evaluate the chemical relevance of the calculated metabolic distance. Shoot metabolites responsible for separation of accessions were identified *in silico* as the first strategy. Hereto, we used partial PCA plots eliminating the effect of the environment from the dataset to find the metabolites responsible for grouping of the accessions. Then, the scores and loadings plots with the first two PCs were superimposed for shoot GC-MS and LC-MS positive and negative mode datasets separately. The biplots of shoot GC-MS data (supplementary information Fig. 3) and LC-MS negative mode data (Fig.

2B) are shown as examples. As a rule, only those metabolites were included into the ordination diagram that more than 55% of their influence was represented by the first two PCs. Consequently, ten metabolites were pinpointed in the GC-MS data that contributed most to the clustering of accessions Cvi and Kond on one sides of the scores plot (supplementary information Fig. 3 and supplementary information Table 6). They were identified and annotated as monosaccharides (fructose, 1-methyl-alpha-D-glucopyranoside and glucopyranose), a disaccharide (sucrose), an amino acid (L-glutamic acid) and its derivative (pyroglutamic acid), and four other, yet unknown, compounds that we could not unambiguously identify . All of them have previously been detected in *A. thaliana* (Fiehn *et al.* 2000). ANOVA showed that there is a statistical significant difference between the accessions ($\alpha = 0.05$) for these ten metabolites. The fact that these primary metabolites were among the most discriminating compounds suggests a fundamental difference in central carbon metabolism of Cvi and Kond compared with the other accessions, including Col-0. This difference is reflected in the divergence ranking of Cvi and Kond in the GC-MS shoot analysis (Table 2).

In the shoot LC-MS negative mode dataset, 11 metabolites were most responsible for the separate grouping of the accessions in the PCA biplot (Fig. 2B). The most likely elemental compositions of the corresponding parent ion mass of all these could be calculated by Seven Golden Rules (Kind and Fiehn 2007). Library search in the KnapSack database (http://kanaya.naist.jp/knapsack_jsp/top.html) and the Dictionary of Natural Products (http://dnp.chemnetbase.com) resulted in the annotation of 10 compounds of which 8 were denoted as glucosinolates (supplementary information Table 7). Glucosinolate profiles of *A. thaliana* have indeed been reported to vary across accessions (Kliebenstein *et al.* 2001). All putatively identified glucosinolates have been reported before in Col-0 (Matsuda *et al.* 2010) or other Brassicaceae genera (knapSack). The remaining three compounds with parent ion masses of 789.222 (tentatively identified as truncatulin A from *medicago truncatu*la) 338.093 (tentatively identified as clinacoside C) and 416.106 have so far not been reported in Brassicaceae or any plant species.

Analogously, in the shoot LC-MS positive mode dataset, ten metabolites responsible for separate grouping of An-1, Cvi and C-24 on the PCA biplot (not shown) were pin-pointed, of which nine could be tentatively identified. Only two of them, diptocarpilidine and 3-carboxytyrosine, have been reported before in at least one Brassicaceae species (supplementary information Table 7). The other *m/z* signals corresponded mainly to flavonoid derivatives and nitrogen containing compounds that have not been reported before in *A. thaliana*. All of the accession-specific metabolites in the LC-MS analyses were confirmed to be statistically different across accessions by ANOVA ($\alpha = 0.05$).

MS/MS fragmentation of the most relevant metabolites

MS/MS experiments were performed to further identify the differential metabolites. Mass and retention time-directed spectra were registered at different collision energies (from 5 up to 50 eV) for each compound and combined into one MS/MS spectrum. Not all of the selected ion masses could be fragmented perhaps due to structure stability although their presence was confirmed in the samples of independent experiments. The fragmentation pattern of 14 of the 21 ion masses was obtained successfully. The fragmentation pattern in combination with the corresponding retention time, isotopic pattern and UV-Vis spectrum allowed for further confirmation of the structure.

In negative mode, MS/MS confirmed the identity of six glucosinolates with their characteristic fragment signal at $[M-H]^-$ 96.9595 Da (~ 97, supplementary information Table 7). The isotopic pattern of the other putative glucosinolates, which failed to fragment in MS/MS analysis, confirmed the presence of sulfur making it likely they are indeed glucosinolates.

A phenolic compound with the elemental formula of $C_{37}H_{42}O_{19}$ and accurate mass of $[M-H]^-$ 789.224 Da was also detected in negative mode which was putatively annotated as an isomer of truncatulin A and not reported before in Brassicaceae. MS/MS analysis of the deprotonated molecule confirmed the presence of ferulic acid, guaiacylglycerol and genticic acid linked via two pentose sugars. (supplementary information Fig. 4A). This is the first report of this mass and elemental composition in *A. thaliana*. Hydroxybenzoic acids (such as genticic acid) combined with a guaiacylglycerol group have been found to occur as breakdown products of lignin (Katayama *et al.* 1981). Ferulic acid is a constituent of lignocellulose that crosslinks the lignin and polysaccharides conferring rigidity to the cell walls. These products may have some relation with defense (Fayos *et al.* 2006) but it's presence in some *A. thaliana* accessions remains to be further confirmed and explained.

In positive mode, three of the nitrogen containing compounds were annotated as alkaloids that were not reported before in *A. thaliana* (Supplementary Information Table 7). After analysis of MS/MS spectra it was possible to correlate the fragments and characteristic neutral losses with the structures proposed for one alkaloid, diptocarpilidine and also for one amino acid derivative, 3-carboxytyrosine, which have been reported before in Brassicaceae. To date, only the indolic alkaloid, camalexin (a phytoalexin), has been identified in A. thaliana (Glazebrook and Ausubel 1994, Hansen and Halkier 2005). However, the presence of a multitude of alkaloid biosynthetic gene homologues in the A. thaliana genome may suggest that more alkaloids can potentially be synthesized (Facchini *et al.* 2004). Annotated alkaloids

were particularly present in accessions Cvi, C-24 and An-1 and the fact that they have not been reported before may be simply due to their low abundance in Col-0.

Two putative flavonoid derivatives were identified in LC-TOF-MS positive mode (Supplementary Information Table 7). One of them was registered with accurate mass $[M+H]^+$ 947.2816 Da and was annotated *in silico* as an anthocyanin derivative not reported before in the Brassicaceae. Analysis of the MS/MS spectra (supplementary information Fig. 5A) shows this is a different flavonoid derivative, namely kaempferol 7-O-rhamnoside 3-O-rhamnosyl-(synapoyl)glucoside (Supplementary information Fig. 5B). Flavonoids have been studied to a large extent in *A. thaliana* and variants of this class of secondary metabolites have been identified including derivatives of kaempferol 3-glycosides (Matsuda *et al.* 2010, Sever *et al.* 2010). Here we report two new putative flavonoids in *A. thaliana* detected in LC-MS positive mode, with one of them also annotated as kaempferol glycoside by MS/MS fragmentation. The fact that these flavonoids have so far escaped identification in *A. thaliana* may be again due to their absence in the mostly studied accession Col-0.

Biological relevance of metabolic distance

Metabolites mediate the interaction of plants with their environment including biotic agents (Forlani 2010, Kashif, Federica, Eva, Martina, Young and Robert 2009, Macel and Klinkhamer 2009, Mahatma *et al.* 2009). Therefore metabolites are highly adaptive and hence strongly vary between genotypes and habitats/environments of the accessions (Menezes and Jared 2002). Hence, our strategy to evaluate the biological context of the observed metabolic divergence between accessions was to relate their differential metabolic profiles to the interaction with biotic agents. We characterized accessions in a non-induced metabolic state. The obtained metabolite fingerprints thus represent the innate immune or general defense system without a requirement for induction. Based on their metabolic distances, four accessions were selected that best represent the observed metabolome variation across the nine accessions. Cvi and An-1 were selected as the most diverged, Col-0 as the moderately diverged and most studied and Eri as an average accession (Table 2). The four accessions showed segregation with regard to resistance against or attractiveness to different biotic agents, including host and non-host and biotrophic and necrotrophic pathogens as well as an insect (supplementary information Table 8). None of the selected accessions was resistant or susceptible to all of the biotic agents. Subsequently, the resistance or attractiveness levels to the biotic agents were used in RDA as explanatory variables for the abundance of the reconstructed metabolites in the CC-environment samples as bioassays were done under the same conditions. Several metabolites were shown to significantly (MCP test, 10000 permutations, P-value < 0.01) correlate with the resistance level of the accessions to different

biotic agents (Fig. 3), some of which were not identified as accession-specific in our first strategy. Among the correlating metabolites based on the first two PCs two were detected in LC-MS positive mode whereas the rest were detected in negative mode with glucosinolates dominating the list (Fig. 3). Abundance of positively correlated compounds with *Botrytis* and thrips resistance was shown to be negatively correlated with the resistance against downy mildew isolates, while abundance of positively correlated compounds with *Oidium* resistance was shown to have no correlation with the resistance against other biotic agents (Fig. 3).

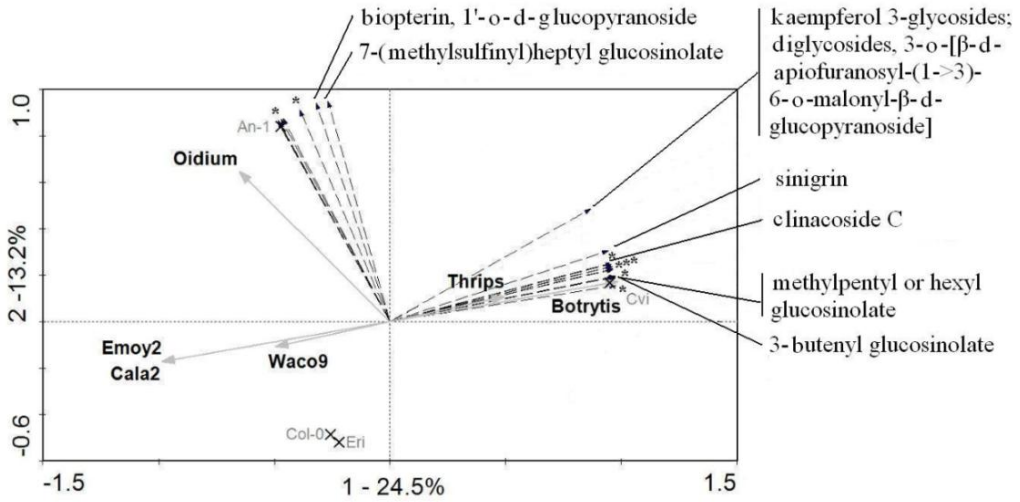


Figure 3. Ordination plot of accessions tested with a range of biotic agents: *Peronospora parasitica* isolates (Emoy2, Cala2, Waco9); *Oidium neolycopesici*; *Botrytis cinerea* and *Frankliniella occidentalis* (thrips). The resistance or repellence level of accessions to biotic agents (grey vectors) were set as explanatory variables and abundance of metabolites as response variables in the RDA plot. Numbers along the axes indicate the ordinate number and percentage of variation explained. X indicates the position of accessions on the scores plot with respect to their resistance level. Listed metabolites are accession-specific compounds identified in the present study and correlating with resistance; "*" indicates metabolites that were not identified in the present study but correlated with resistance. Vectors pointing in the same direction are positively correlated and those pointing in opposite directions are negatively correlated

Using the metabolic distance based on untargeted metabolite fingerprints, we selected accessions of *A. thaliana* that chemically diverged in the PCA plane. As a consequence of this chemical divergence, these accessions also differed in their interaction with biotic agents. This variation in chemical composition and biotic interaction could be subsequently used to identify compounds correlating with the interaction, some of them had

been found to be accession specific in the two dimensional PCA plane (Fig. 3 and supplementary Information Table 7), thus linking the metabolic distance of accessions with their differential response towards biotic stresses.

Correlation between genetic and metabolic distance

An integrated dataset that consisted of shoot metabolite fingerprints from all analytical methods was used in a partial PCA with environment as cofactor, in order to estimate the pair-wise metabolic distance (R-value) among the nine accessions across all analytical techniques (supplementary information Table 9). The 149 genome-wide distributed SNP markers were used to determine the genetic distance among the nine accessions (supplementary Information Table 9). A Mantel test showed that there is a small (correlation coefficient r=0.04) but significant (*P*-value < 0.01) pattern similarity between the shoot metabolic and genetic distance matrices. There was no significant correlation (r = 0.26 and *P*-value = 0.11) between the root metabolic distance (determined using all analytical methods) and the genetic distance matrices.

The weak correlation between the genetic and metabolic distance matrices shows that genetic diversity is not one to one translated into metabolic diversity. The phylogenetic tree of all used accessions (supplementary information Fig. 1) demonstrates that at the resolution level of 0.8, An-1 is the most genetically diverged accession and Eri is the second at a lower resolution (0.7). Moreover, there is a large genetic distance between these two accessions as shown in supplementary information table 9 and as they belong to the two different major clades of the tree (supplementary information Fig. 1). However, based on the combined metabolite datasets of the shoot, Cvi has the largest metabolic divergence and Eri is considered as an average accession with the least metabolic divergence (shown by the ranking of the average distance, Supplementary Information Table 9). An-1 is the second metabolically diverged accession, which would fit with the fact that it is genetically the most diverged. However, An-1 also has a large metabolic distance from Cvi (R=0.95) while genetically they belong to the same major clade (supplementary information Fig. 1). Such discrepancies imply that the genetic distance of two genotypes does not completely define their relationship and distance at the metabolic level. Our result is in accordance with a number of previous studies on closely related genotypes of two plant species (*Sesamum indicum* and *Oryza sativa*) and on several *Rhizobium* species (Laurentin *et al.* 2008, Mochida *et al.* 2009, Wolde-Meskel *et al.* 2004).We also observed convergence of "metabotypes", i.e. accessions with no metabolic distance although diverged genetically. As an example in the root dataset, negligible metabolic distance is observed between An-1 and Kyo-1 in both LC-MS modes (Table 1), while the two accessions of each of these pairs belong to the two

diverged subclades of the phylogenetic tree (supplementary information Fig. 1). A close metabolic relationship for genetically diverged accessions is in accordance with the hypothesis of phenotypic buffering (Fu, Keurentjes, Bouwmeester, America, Verstappen, Ward, Beale, De Vos, Dijkstra, Scheltema, Johannes, Koornneef, Vreugdenhil, Breitling and Jansen 2009). Although a genetic basis underlies the metabolome variation between *A. thaliana* accessions (our data) and mapping populations (Keurentjes, Fu, De Vos, Lommen, Hall, Bino, Van Der Plas, Jansen, Vreugdenhil and Koornneef 2006), the hypothesis of phenotypic buffering suggests the existence of breakpoints in a system that buffers them against too large effects of genetic variation on the phenotype.

Conclusion

We characterized the metabolic variation within 9 *A. thaliana* accessions grown under various growing conditions and established a statistical method for estimating a metabolic distance between genotypes or treatments. This method may help to evaluate and compare the effects of genetic (natural variation, breeding and genetic modification) or environmental perturbations on the metabolome. Metabolic distance can be used to quantify the metabolic diversity and plasticity among plant genotypes and environments and could be a useful tool in breeding programs and genetical genomics studies.

Supplementary Information Table 1 List of the *Arabidopsis thaliana* accessions used in this study and their geographical origin

Native Name	Short Name	Country of Origin	Region
Antwerpen	An-1	Belgium	West Europe
Columbia 0	Col-0	USA	Midwest USA
Cape Verdi Islands	Cvi	Cape Verdi Islands	Macaronesia ecoregion of the North Atlantic Ocean
Eriengsboda	Eri	Sweden	North Europe
Kondara	Kond	Tajikistan	Central Asia
Kyoto	Kyo-1	Japan	East Asia
Landsberger	Ler	Germany	Midwest Europe
Wassilewskija	WS	Russia	East Europe
C24	C-24	Portugal	South Europoe

Supplementary Information Table 2 Environmental conditions used to grow the *Arabidopsis thaliana* accessions of this study

	Light	Temperature	Humidity	Medium
Climate chamber (CC)	100 µM m^{-2} s^{-1}	22	Uncontrolled	Potting compost
Hydroponics (HY)	100 µM m^{-2} s^{-1}	Day: 22 Night: 18	70%	Hydroponics solution[1]
Greenhouse (GH)	> 100 µM m^{-2} s^{-1}	Day: 22 Night: 19	70%	Potting compost
Uncontrolled (OP)	Uncontrolled	Uncontrolled	Uncontrolled	Potting compost

1 Tocquin, P., Corbesier, L., Havelange, A., Pieltain, A., Kurtem, E., Bernier, G., (2003). A novel high efficiency, low maintenance, hydroponic system for synchronous growth and flowering of Arabidopsis thaliana. BMC Plant Biology, 3.

Supplementary Information Table 3 The number of masses or reconstituted metabolites after each step of the data processing workflow with MetAlign, METOT[1] and MMSR[2]

	Metalign	METOT[1]	MMSR[2]
Shoot LC-MS+	38972	15306	486
Shoot LC-MS-	44503	16357	456
Shoot GC-MS	12001	7576	222
Total Shoot	*95476*	*39239*	*1164*
Root LC-MS+	20227	7682	352
Root LC-MS-	20551	2294	483
Root GC-MS	8205	4960	225
Total Root	*48983*	*14936*	*1060*

1 Metabolic MetAlign Output Transformer

2 Multivariate Mass Spectra Reconstruction

Supplementary Information Table 4 Explained variation (in %) by the variation sources in different datasets

	Accession	Environment	Accession-Environment interaction
Shoot LC-MS+	17.0	18.3	0.5
Shoot LC-MS-	26.8	36.4	N.S.[1]
Shoot GC-MS	22.5	16.8	0.6
Root LC-MS+	79.0	N.A.[2]	N.A.
Root LC-MS-	80.6	N.A.	N.A.
Root GC-MS	53.9	N.A.	N.A.

[1]N.S. not significant (p-value > 0.05)
[2]N.A. not applicable

Supplementary Information Table 5 Average R-value (metabolic distance) of the environments in different analytical methods obtained by the ANOSIM permutation test. Numbers in parentheses show ranking of the environment within the row

	Climate chamber	Hydroponics	Greenhouse	Uncontrolled	PC[1]	Variance explained[2]
Shoot LC-MS+	0.72 (2)	0.88 (1)	0.65 (4)	0.67 (3)	3	0.375
Shoot LC-MS-	0.76 (2)	0.98 (1)	0.64 (4)	0.71 (3)	3	0.531
Shoot GC-MS	0.50 (2)	0.57 (1)	0.34 (4)	0.42 (3)	4	0.485
Total	1.98 (2)	2.43 (1)	1.63 (4)	1.8 (3)		

[1] the number of principle components (PC) used to calculate the Euclidean distances for ANOSIM

[2] % variation explained by the PCs used

Supplementary Information Table 6 List of *in silico* tentatively identified metabolites from the analyses by GC-TOF-MS that are associated with grouping of *A. thaliana* accessions Cvi and Kond

Putative ID	Retention index	Matching factor	Reverse Matching Factor	Library
fructose	1453.125	917	924	GMD
1-methyl-alpha-D-glucopyranoside	1548.525	904	916	GMD
glucopyranose	1618.017	901	975	GMD
sucrose	2160.211	842	931	GMD
L-glutamic acid	1390.882	970	970	GMD
pyroglutamic acid	1405.547	991	993	GMD
unknown (1)[1]	1983.757	818	818	GMD
unknown (2)[1]	1478.680	745	748	GMD
unknown (3)[1]	1617.983	916	971	GMD
unknown (4)[1]	1449.313	724	724	GMD

[1] Detected in *A. thaliana* but unidentified

Supplementary Information Table 7 List of differentially produced metabolites in *A. thaliana* accessions detected by LC-TOF-MS with putative *in silico* identified molecular formula and assigned names, some with MS\MS information. Putative IDs are *in silico* annotations unless stated otherwise

	Ret (min)	Detected Accurate Mass	Putative Elemental compositions	Putative ID	Δmass (ppm)	Library	Organism	specific accession	MS/MS fragments
LC-MS negative mode									
	1.47	388.039	C11H19NO10S2 C12H15N5O6S2 C20H11N3O4S C12H23NO5S4	progoitrin confirmed by MS/MS	1	KnapSack	*A. thaliana*	Cvi & Kond	275, 259, 195, 136, 97
	1.72	358.029	C10H17NO9S2 C11H13N5O5S2 C10H9N5O10	sinigrin confirmed by MS/MS	1.3	KnapSack	*A. thaliana*	Cvi & Kond	275, 259, 241, 195, 162, 116, 97
	2.63	338.093	C12H21NO8S C19H19N2S2 C13H17N5O4S C14H13N9S C16H13N5O4	clinacoside C (a sulfur containing amide)	1.5	KnapSack	*Clinacanthus nutans*, Unidentified in Brassicaceae	Cvi & Kond	NF[1]

Table 7 continued	Ret (min)	Detected Accurate Mass	Putative Elemental compositions	Putative ID	Δmass (ppm)	Library	Organism	specific accession	MS/MS fragments
	2.70	372.045	C11H19NO9S2 C16H15N5S3	3-butenyl glucosinolate confirmed by MS/MS	2	KnapSack	*A. thaliana*	Kond, Cvi, C-24 & An-1	275, 259, 241, 195, 179, 130, 119, 97
	2.92	374.059	C11H21NO9S2 C12H17N5O5S2 C11H13N5O10 C12H9N9O6 C20H13N3O3S	isobutyl glucosinolate, n-butyl glucosinolate, sec-butyl glucosinolate	0.1	KnapSack	*Brassica oleracea* (Brassicaceae)	Cvi & Kond	NF[1]
	2.99	478.088	C15H29NO10S3 C20H17N9O2S2 C30H23S3 C27H17N3O4S C14H25NO15S	7-(methylsulfinyl)heptyl glucosinolate confirmed by MS/MS	0	KnapSack	*A. thaliana*	An-1 & C-24	463, 414, 275, 252, 220, 97
	3.27	494.080	C18H25NO11S2 C19H21N5O7S2 C23H21N5O2S3 C22H17N5O7S C19H13N9O8	4-benzoyloxybutyl glucosinolate	0.4	KnapSack	*A. thaliana*	An-1 & C-24	NF[1]
	4.71	402.088	C13H25NO9S2 C16H21NO9S C17H17N5O5S C17H25NO4S3 C18H21N5S3	methylpentyl glucosinolate or hexyl glucosinolate confirmed by MS/MS	1	KnapSack	*Raphanus sativus* & *Wasabi japonica* (Brassicaceae)	An-1 & C-24	97

Table 7 continued	Detected Accurate Mass	Putative Elemental compositions	Putative Elemental compositions	Δmass (ppm)	Library	Organism	specific accession	MS/MS fragments	Detected Accurate Mass
	6.01	450.091	C17H25NO9S2 C18H21N5O5S2 C18H13N9O6 C26H17N3O3S C16H21NO14	4-phenylbutyl glucosinolate confirmed by MS/MS	1	KnapSack	*Armoracia rusticana* (Brassicaceae)	An-1 & C-24	97
	6.02	416.106	C15H23N5O5S2 C17H23NO9S C18H19N5O5S C15H15N9O6 C23H19N3O3S	unidentified	2.6	DNP	Unidentified	An-1, C-24 & Col-0	152, 97
	6.44	789.224	C37H42O19	truncatulin A annotated as truncatulin A isomer by MS/MS	3		*Medicago truncatula*	Cvi & Kond	593, 417, 399, 285, 193, 152,
LC-MS positive mode									
	2.66	667.150	C29H30O18 C28H32Cl2N6O7S	kaempferol 3-glycosides; diglycosides, 3-*O*-[β-D-apiofuranosyl-(1->3)-6-*O*-malonyl-β-D-glucopyranoside]	0.95	DNP	*Cicer arietinum, Unidentified in Brassicaceae*	An-1 & C-24	NF[1]

Table 7 continued	Ret (min)	Detected Accurate Mass	Putative Elemental compositions	Putative ID	Δmass (ppm)	Library	Organism	specific accession	MS/MS fragments
	2.68	226.070	C10H11NO5 C11H7N5O C6H7N7O3 C11H15NS2 C14H11NS	3-carboxytyrosine; (*S*)-form, MS/MS fragments fit with proposed structure	0		*Lunaria annua* (Brassicaceae)	An-1 & C-24	209, 191, 180, 162, 149, 134
	2.93	400.146	C15H21N5O8 C20H21N3O6 C19H21N5O3S C16H25N5O3S2 C16H17N9O4	biopterin, 1'-O-D-glucopyranoside (antibiotic activity)	0.1	DNP	*Spirulina platensis*, Unidentified in Brassicaceae	An-1, Cvi & C-24	238
	3.65	206.119	C12H15NO2 C9H19NO2S	bellendine (alkaloid)	2	DNP	*Bellendena montana*, Unidentified in Brassicaceae	Cvi & C-24	189
	4.18	174.094	C8H15NOS	diptocarpilidine (alkaloid), MS/MS fragments fit with proposed structure	1.6	DNP	*Diptychocarpus strictus* (Brassicaceae)	An-1, Cvi & C-24	130, 110
	4.50	346.090	C17H15NO7 C17H22Cl3N C13H11N7O5 C12H15N3O9 C10H15N7O5S	usnic acid amide (simple aromatic compound),3α-(4,4,4-trichloro-2 – butyryloxy)tropane	1.9, 1.0	DNP	*Cercosporidium henningsii*, *Scopolia tanguticus* Unidentified in Brassicaceae	Cvi & C-24	NF[1]

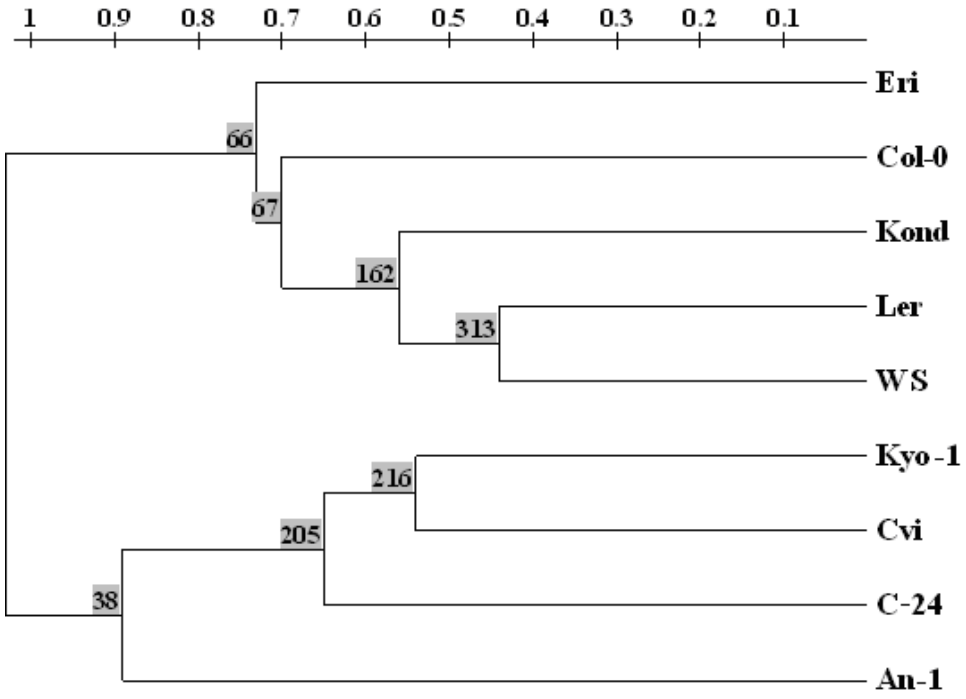
Table 7 continued	Ret (min)	Detected Accurate Mass	Putative Elemental compositions	Putative ID	Δmass (ppm)	Library	Organism	specific accession	MS/MS fragments
	5.23	561.211	C32H32O9 C27H32N2O11 C28H28N6O7 C22H28N10O8 C26H28N10O3S	lehmbachol C.	1.1	DNP	*Salacia lehmbachii*, Unidentified in Brassicaceae	Cvi & C-24	NF[1]
	5.41	188.109	C12H13NO C9H17NOS	4-hydroxy-2-propylquinoline (alkaloid)	1.3	DNP	*Boronia ternata*, Unidentified in Brassicaceae	Cvi & C-24	147, 124, 107
	6.17	730.268	C35H43N3O12S C36H39N7O8S	unidentified			Unidentified	Cvi & C-24	NF[1]
	6.05	947.282	C44H51O23(+)	3,4',5,7-tetrahydroxy-3',5'-dimethoxyflavylium(1+); 3-*O*-[4-hydroxy-*Z*-cinnamoyl-(->4)-α-L-rhamnopyranosyl-(1->6)-β-D-glucopyranoside], 5-*O*-β-D-glucopyranoside annotated as kaempferol 7-O-rhamnoside 3-O-rhamnosyl-(synapoyl)glucoside by MS/MS	1	DNP	*Petunia hybrida*, Unidentified in Brassicaceae	Cvi & C-24	801, 433, 369, 287, 207

Supplementary Information Table 8 Interaction of *Arabidopsis thaliana* accessions with biotic agents. Columns 1-3: isolates of *Peronospora parasitica* (1- Emoy2, 2- Cala2, 3- Waco9, downey mildew); 4: *Oidium neolycopesici* (powdery mildew); 5: *Botrytis cinerea* (botrytis); 6: *Frankliniella occidentalis* (thrips). R, resistant or unattractive; S, susceptible or attractive; M, mild sporulation or semi-attractive

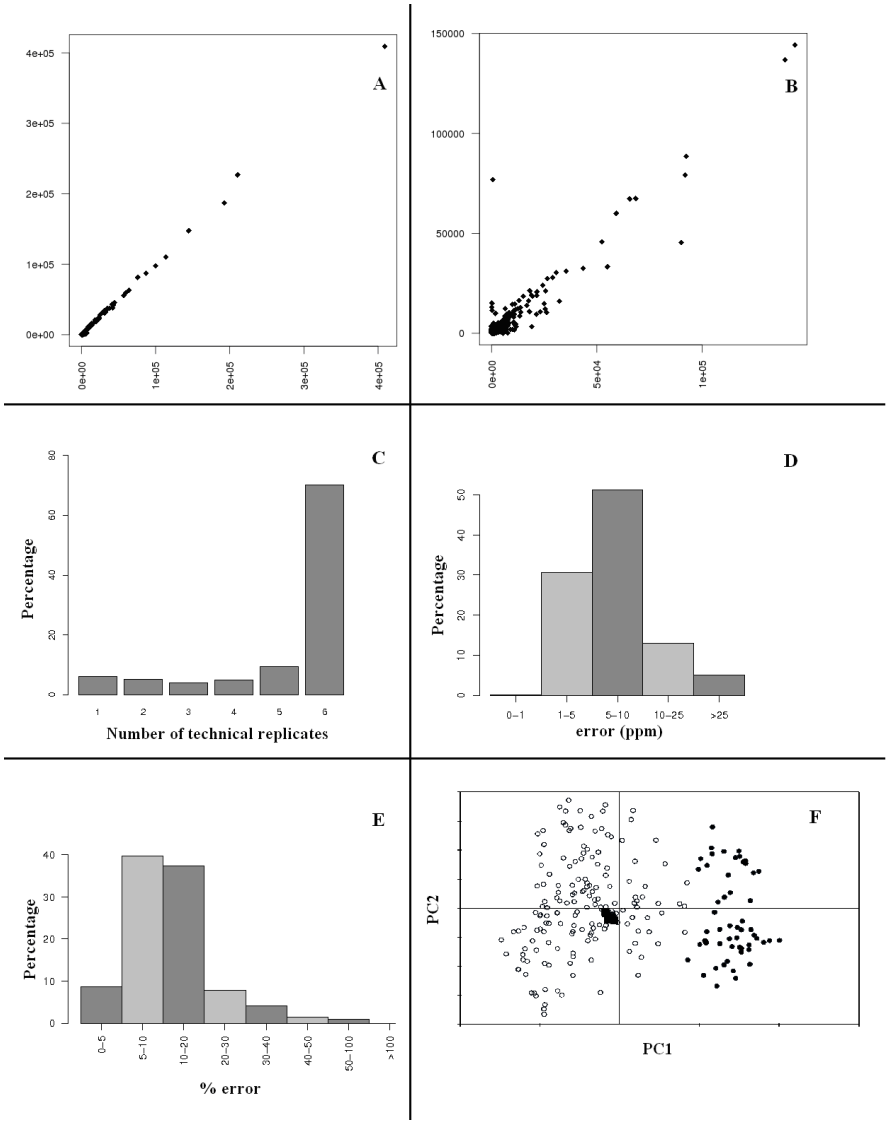
	1	**2**	**3**	**4**	**5**	**6**
An-1	R	R	M	R	S	S
Col-0	R	R	S	M	S	R
Cvi	S	M	S	S	M	R
Eri	R	R	R	S	R	M

Supplementary Information Table 9 Above the diagonal: Matrix of R-values (metabolic distance) obtained by ANOSIM using an integrated dataset of shoot samples analyzed with all analytical methods. Below the diagonal: Matrix of genetic distance determined using the "Juke & Kantor" method. Numbers in parentheses indicate the ranking of the corresponding accession among the other accessions with regard to its metabolic distance

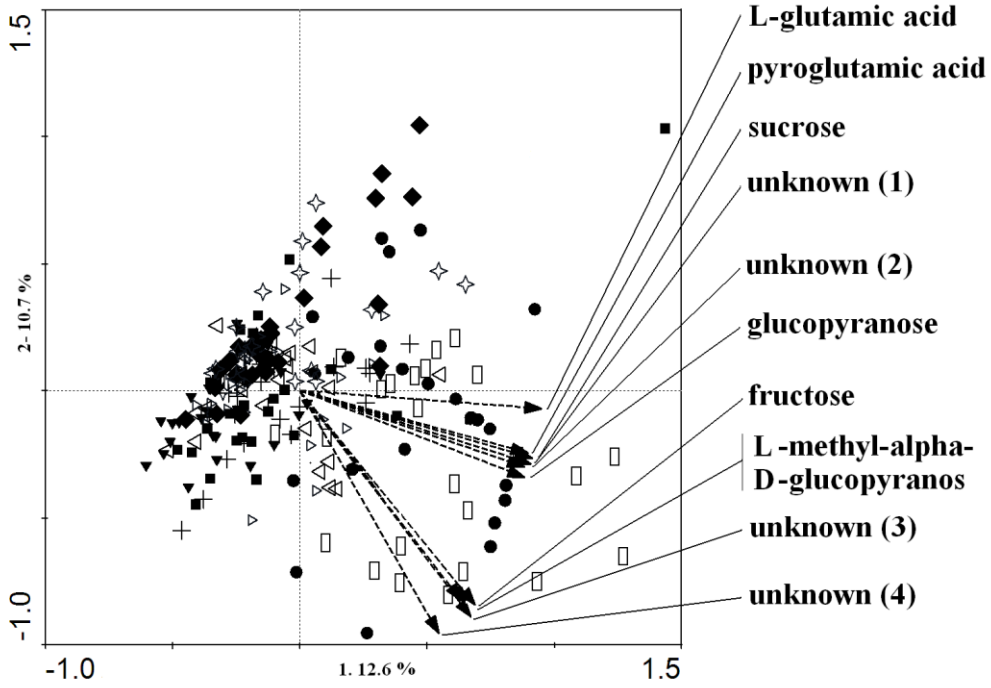
		Metabolic distance								
		An-1	**Col-0**	**C-24**	**Cvi**	**Eri**	**Kond**	**Kyo-1**	**Ler**	**WS**
Genetic distance	**An-1**		0.68	0.19	0.95	0.67	0.93	0.79	0.65	0.76
	Col-0	77.18		0.71	0.95	0.02	0.35	0.17	0.14	0.02
	C-24	83.49	83.49		0.77	0.64	0.79	0.78	0.72	0.71
	Cvi	84.6	74.16	64.78		0.85	0.83	0.95	0.96	0.89
	Eri	92.79	73.26	75.09	74.16		0.34	0.20	0.19	0.03
	Kond	89.21	65.01	61.373	54.39	65.011		0.51	0.65	0.25
	Kyo-1	85.75	69.36	104.15	86.94	62.34	102.51		0.12	0.12
	Ler	66.54	70.2	73.93	77.03	64.78	67.34	56.2		0.19
	WS	70.2	50.53	70.03	77.03	72.1	63.05	54.64	43.63	
	Average metabolic distance	0.70 (2)	0.38 (7)	0.66 (3)	0.89 (1)	0.37 (8)	0.58 (4)	0.46 (5)	0.45 (6)	0.37 (8)



Supplementary Information Figure 1 Phylogenetic tree of the accessions used in this study determined by using 149 genome-wide SNPs. The distance scale is shown above. The bootstrap values at the branching points denote the number of bootstrap trees after 500 bootstraps yielding a cluster of the same composition

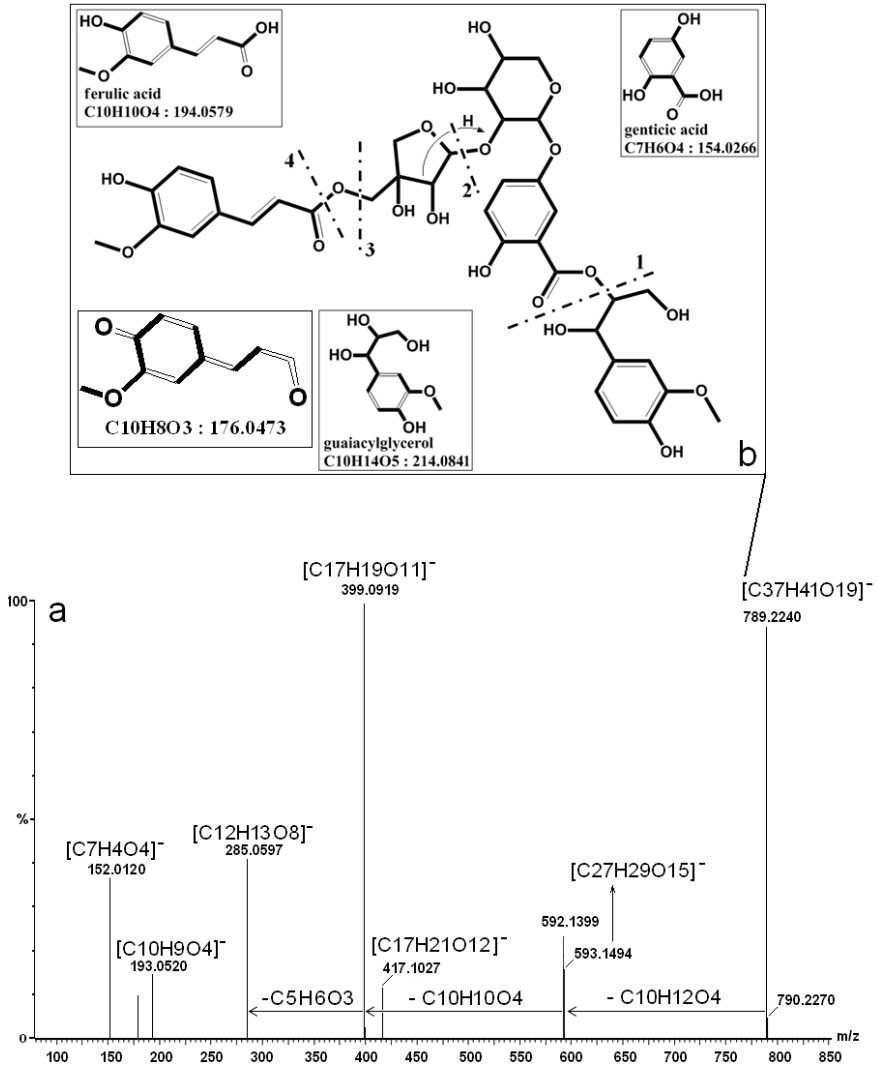


Supplementary Information Figure 2 Quality control of the data using root (A,B,C,D & E) and shoot (F) LC-MS negative mode datasets as examples. A; Peak intensity scatter plot of two technical replicates with an interval of 40 samples, B; Peak intensity scatter plot of two biological replicates of An-1 and Cvi, C; Percentage of masses present in 1 to 6 technical replicates, D; frequency distribution of mass-error (ppm) for masses present in all technical replicates, E; frequency distribution of percentage amplitude-error for masses present in all technical replicates, F; PCA scores plot projecting shoot LC-MS negative mode samples using PC1 and PC2, ■; technical replicates, ○; biological replicates of samples in CC (climate chamber), GH (controlled greenhouse) and UC (uncontrolled greenhouse), ●; biological replicates of HY (hydroponics)

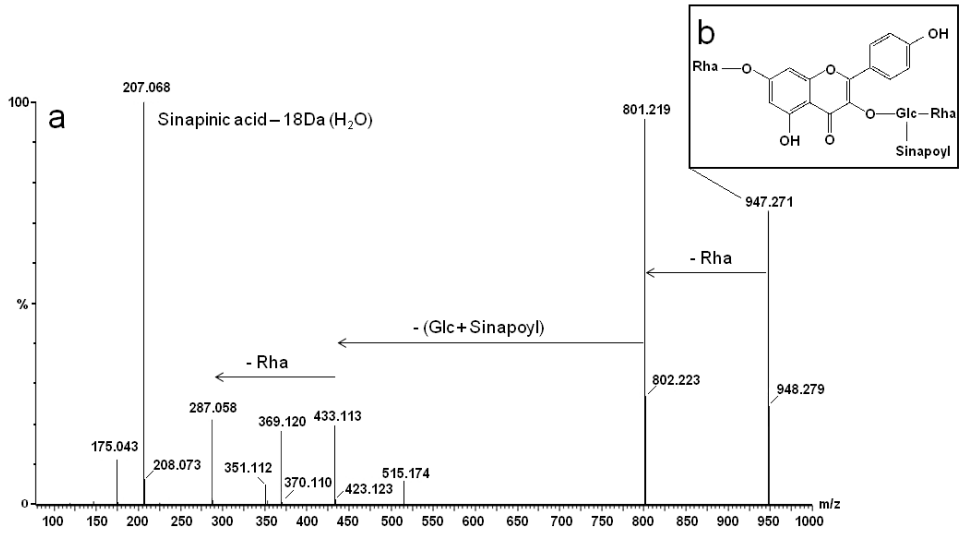


Supplementary Information Figure 3 Partial PCA biplot (superimposed scores and loadings plots) of the shoot of nine accessions grown in four environments, analyzed by GC-TOF-MS. Dashed arrows represent the 10 metabolites that more than 55% of their influence was represented by the first two PCs. Numbers along the axes indicate the PC number and percentage of variation explained. Metabolites indicated were tentatively identified as being responsible for separate clustering of accessions

▹ An-1, ◆ Col-0, ◃ C-24, ▯ Cvi, ■ Eri, ● Kond, + Kyo-1, ▼ Ler,



Supplementary Information Figure 4 MS/MS collective ion mass spectrum of the putative truncatulin A isomer with the parent ion $[M\text{-}H]^-$: 789.2240 (*a*). Proposed structure for truncatulin A isomer (*b*). Fragment $[C_{27}H_{29}O_{15}]^-$: 593.1494 was the result of the neutral loss of $[C_{10}H_{12}O_4]$: 196.0735 (derived from the dehydroguaiacylglycerol unit) (1) with the negative charge on $[COO]^-$. Subsequently, the fracture between two cyclic pentose units (beta elimination) and transfer of an H atom to the xylose unit results in the $[C_{12}H_{13}O_8]^-$: 285.0597 fragment (2). The fragment with m/z 399.0919 corresponds to a loss of ferulic acid $[C_{10}H_{10}O_4]$: 194.0579 from the $[C_{27}H_{29}O_{15}]^-$: 593.1494 fragment (3) resulting in the fragment with m/z 193.0520 i.e. $[C_{10}H_9O_4]^-$. Mass $[C_{27}H_{29}O_{15}]^-$: 593.1494 can also lose a ketene substituted ferulic acid $[C_{10}H_8O_3]$:176.0473 and leave the $[C_{17}H_{21}O_{12}]^-$: 417.1033 (4). Mass $[C_7H_4O_4]^-$: 152.0109 is the anion radical derived from genticic acid



Supplementary Information Figure 5 MS/MS collective ion mass spectrum of kaempferol 7-O-rhamnoside 3-O rhamnosyl-glucoside(sinapoyl) with parent ion $[M+H]^+$ 947.2816 Da (*a*). Proposed structure of kaempferol 7-O-rhamnoside 3-O rhamnosyl-glucoside(sinapoyl) (b). Glc: glucose residue (162 Da); Rha: rhamnose residue (146 Da). The assignment of the sugar linkage is based on intensities of ions resulting from sequential losses of rhamnose $[M+H-146]^+$, synapoyl glucoside $[M+H-146-368]^+$ and a second rhamnose residue $[M+H-146-368-146]^+$.Although it is not possible to unambiguously determine the aglycon substitution, the positions of the sugar linkage to the aglycon were based on the literature available in this field (Stobiecki *et al.* 2006 ;Vallejo *et al.* 2004)

Stobiecki, M., Skirycz, A., Kerhoas, L., Kachlicki, P., Muth, D., Einhorn, J., (2006). Profiling of phenolic glycosidic conjugates in leaves of Arabidopsis thaliana using LC/MS. *Metabolomics, 2*(4), 197-219

Vallejo, F., Tomas-Barberan, F. A., & Ferreres, F. (2004). Characterisation of flavonols in broccoli (Brassica oleracea L. var. italica) by liquid chromatography-UV diode-array detection-electrospray ionisation mass spectrometry. *Journal of Chromatography A, 1054*(1-2), 181-193

References

Aliferis, K.A. and Jabaji, S. (2009) 1H NMR and GC-MS metabolic fingerprinting of developmental stages of Rhizoctonia solani sclerotia. Metabolomics, 1-13.

Almeida, J.A.S., Barbosa, L.M.S., Pais, A.A.C.C. and Formosinho, S.J. (2007) Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. Chemometrics and Intelligent Laboratory Systems, **87**, 208-217.

Arimura, G.I., Kost, C. and Boland, W. (2005) Herbivore-induced, indirect plant defences. Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids, **1734**, 91-111.

Bai, Y., Pavan, S., Zheng, Z., Zappel, N.F., Reinstadler, A., Lotti, C., De Giovanni, C., Ricciardi, L., Lindhout, P., Visser, R., Theres, K. and Panstruga, R. (2008) Naturally occurring broad-spectrum powdery mildew resistance in a Central American tomato accession is caused by loss of Mlo function. Molecular Plant-Microbe Interactions, **21**, 30-39.

Belz, R.G. (2007) Allelopathy in crop/weed interactions - An update. Pest Management Science, **63**, 308-326.

Boccard, J., Grata, E., Thiocone, A., Gauvrit, J.Y., Lanteri, P., Carrupt, P.A., Wolfender, J.L. and Rudaz, S. (2007) Multivariate data analysis of rapid LC-TOF/MS experiments from Arabidopsis thaliana stressed by wounding. Chemometrics and Intelligent Laboratory Systems, **86**, 189-197.

Bouwmeester, H.J., Matusova, R., Zhongkui, S. and Beale, M.H. (2003) Secondary metabolite signalling in host-parasitic plant interactions. Current Opinion in Plant Biology, **6**, 358-364.

Bouwmeester, H.J., Roux, C., Lopez-Raez, J.A. and Becard, G. (2007) Rhizosphere communication of plants, parasitic plants and AM fungi. Trends in Plant Science, **12**, 224-230.

Boyes, D.C., Zayed, A.M., Ascenzi, R., McCaskill, A.J., Hoffman, N.E., Davis, K.R. and GÃ¶rlach, J. (2001) Growth stage-based phenotypic analysis of Arabidopsis: A model for high throughput functional genomics in plants. Plant Cell, **13**, 1499-1510.

Cheng, A.X., Lou, Y.G., Mao, Y.B., Lu, S., Wang, L.J. and Chen, X.Y. (2007) Plant terpenoids: Biosynthesis and ecological functions. Journal of Integrative Plant Biology, **49**, 179-186.

Chong, J., Poutaraud, A. and Hugueney, P. (2009) Metabolism and roles of stilbenes in plants. Plant Science, **177**, 143-155.

Cona, A., Rea, G., Angelini, R., Federico, R. and Tavladoraki, P. (2006) Functions of amine oxidases in plant development and defence. Trends in Plant Science, **11**, 80-88.

Davies, A.N. (1998) The new Automated Mass Spectrometry Deconvolution and Identification System (AMDIS). Spectroscopy Europe, **10**, 22-26.

De Jong, G. (2005) Evolution of phenotypic plasticity: Patterns of plasticity and the emergence of ecotypes. New Phytologist, **166**, 101-118.

De Vos, M., Van Oosten, V.R., Van Poecke, R.M.P., Van Pelt, J.A., Pozo, M.J., Mueller, M.J., Buchala, A.J., MÃ©traux, J.P., Van Loon, L.C., Dicke, M. and Pieterse, C.M.J. (2005) Signal signature and transcriptome changes of Arabidopsis during pathogen and insect attack. Molecular Plant-Microbe Interactions, **18**, 923-937.

Ehrlich, P.R. and Raven, P.H. (1964) Butterflies and plants: A study in coevolution. Evolution, **18**, 586-608.

Facchini, P.J., Bird, D.A. and St-Pierre, B. (2004) Can Arabidopsis make complex alkaloids? Trends in Plant Science, **9**, 116-122.

Fayos, J., Bellés, J.M., López-Gresa, M.P., Primo, J. and Conejero, V. (2006) Induction of gentisic acid 5-O-β-D-xylopyranoside in tomato and cucumber plants infected by different pathogens. Phytochemistry, **67**, 142-148.

Ferrari, S., Plotnikova, J.M., De Lorenzo, G. and Ausubel, F.M. (2003) Arabidopsis local resistance to Botrytis cinerea involves salicylic acid and camalexin and requires EDS4 and PAD2, but not SID2, EDS5 or PAD4. Plant Journal, **35**, 193-205.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. (2000) Metabolite profiling for plant functional genomics. Nature Biotechnology, **18**, 1157-1161.

Forlani, G. (2010) Differential in vitro responses of rice cultivars to Italian lineages of the blast pathogen Pyricularia grisea. 2. Aromatic biosynthesis. Journal of Plant Physiology.

Fu, J., Keurentjes, J.J.B., Bouwmeester, H., America, T., Verstappen, F.W.A., Ward, J.L., Beale, M.H., De Vos, R.C.H., Dijkstra, M., Scheltema, R.A., Johannes, F., Koornneef, M., Vreugdenhil, D., Breitling, R. and Jansen, R.C. (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nature Genetics, **41**, 166-167.

Garcia-Perez, I., Vallejo, M., Garciaa, A., Legido-Quigley, C. and Barbas, C. (2008) Metabolic fingerprinting with capillary electrophoresis. Journal of Chromatography A, **1204**, 130-139.

Glazebrook, J. and Ausubel, F.M. (1994) Isolation of phytoalexin-deficient mutants of Arabidopsis thaliana and characterization of their interactions with bacterial pathogens. Proceedings of the National Academy of Sciences of the United States of America, **91**, 8955-8959.

Hammer, O., Harper, D.A.T. and Ryan, P.D. (2001) Past: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica, **4**, XIX-XX.

Hansen, B.G. and Halkier, B.A. (2005) New insight into the biosynthesis and regulation of indole compounds in Arabidopsis thaliana. Planta, **221**, 603-606.

Hilker, M. and Meiners, T. (2006) Early herbivore alert: Insect eggs induce plant defense. Journal of Chemical Ecology, **32**, 1379-1397.

Kabouw, P., Biere, A., Van Der Putten, W.H. and Van Dam, N.M. (2009) Intra-specific differences in root and shoot glucosinolate profiles among white cabbage (Brassica oieracea var. capitata) Cultivars. Journal of Agricultural and Food Chemistry, **58**, 411-417.

Kashif, A., Federica, M., Eva, Z., Martina, R., Young, H.C. and Robert, V. (2009) NMR metabolic fingerprinting based identification of grapevine metabolites associated with downy mildew resistance. Journal of Agricultural and Food Chemistry, **57**, 9599-9606.

Katayama, T., Nakatsubo, F. and Higuchi, T. (1981) Degradation of arylglycerol-β-aryl ethers, lignin substructure models, by Fusarium solani. Archives of Microbiology, **130**, 198-203.

Keurentjes, J.J.B., Fu, J., De Vos, C.H.R., Lommen, A., Hall, R.D., Bino, R.J., Van Der Plas, L.H.W., Jansen, R.C., Vreugdenhil, D. and Koornneef, M. (2006) The genetics of plant metabolism. Nature Genetics, **38**, 842-849.

Kim, S.W., Min, S.R., Kim, J., Park, S.K., Kim, T.I. and Liu, J.R. (2009) Rapid discrimination of commercial strawberry cultivars using Fourier transform infrared spectroscopy data combined by multivariate analysis. Plant Biotechnology Reports, **3**, 87-93.

Kind, T. and Fiehn, O. (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics, **8**.

Kliebenstein, D.J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J. and Mitchell-Olds, T. (2001) Genetic control of natural variation in arabidopsis glucosinolate accumulation. Plant Physiology, **126**, 811-825.

Kliebenstein, D.J., Kroymann, J. and Mitchell-Olds, T. (2005) The glucosinolate-myrosinase system in an ecological and evolutionary context. Current Opinion in Plant Biology, **8**, 264-271.

Laurentin, H., Ratzinger, A. and Karlovsky, P. (2008) Relationship between metabolic and genomic diversity in sesame (Sesamum indicum L.). BMC Genomics, **9**.

Lommen, A. (2009) Metalign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Analytical Chemistry, **81**, 3079-3086.

Macel, M. and Klinkhamer, P.G.L. (2009) Chemotype of Senecio jacobaea affects damage by pathogens and insect herbivores in the field. Evolutionary Ecology, **24**, 237-250.

Mahatma, M.K., Bhatnagar, R., Dhandhukia, P. and Thakkar, V.R. (2009) Variation in metabolites constituent in leaves of downy mildew resistant and susceptible genotypes of pearl millet. Physiology and Molecular Biology of Plants, **15**, 249-255.

Matsuda, F., Hirai, M.Y., Sasaki, E., Akiyama, K., Yonekura-Sakakibara, K., Provart, N.J., Sakurai, T., Shimada, Y. and Saito, K. (2010) AtMetExpress development: A phytochemical atlas of Arabidopsis development. Plant Physiology, **152**, 566-578.

Menezes, H. and Jared, C. (2002) Immunity in plants and animals: Common ends through different means using similar tools. Comparative Biochemistry and Physiology - C Toxicology and Pharmacology, **132**, 1-7.

Mochida, K., Furuta, T., Ebana, K., Shinozaki, K. and Kikuchi, J. (2009) Correlation exploration of metabolic and genomic diversity in rice. BMC Genomics, **10**.

Raguso, R.A. and Pichersky, E. (1999) A day in the life of a linalool molecule: Chemical communication in a plant-pollinator system. Part 1: Linalool biosynthesis in flowering plants. Plant Species Biology, **14**, 95-120.

Rulhmann, S., Leser, C., Bannert, M. and Treutter, D. (2002) Relationship between growth, secondary metabolism, and resistance of apple. Plant Biology, **4**, 137-143.

Schripsema, J. (2010) Application of NMR in plant metabolomics: Techniques, problems and prospects. Phytochemical Analysis, **21**, 14-21.

Sever, A., Einhorn, J., Brunelle, A. and Laprevote, O. (2010) Localization of flavonoids in seeds by cluster time-of-flight secondary ion mass spectrometry imaging. Analytical Chemistry, **82**, 2326-2333.

Smilauer, J.L.a.P. (2003) Multivariate Analysis of Ecological Data using CANOCO
Cambridge University Press, Cambridge.

Snoeren, T.A.L., Kappers, I.F., Broekgaarden, C., Mumm, R., Dicke, M. and Bouwmeester, H.J. (2010) Natural variation in herbivore-induced volatiles in Arabidopsis thaliana. Journal of Experimental Botany, **61**, 3041-3056.

ter Braak, C.J.F. (1988) CANOCO-an extension of DECORANA to analyze species-environment relationships. Vegetatio, **75**, 159-160.

Tikunov, Y., Lommen, A., De Vos, C.H.R., Verhoeven, H.A., Bino, R.J., Hall, R.D. and Bovy, A.G. (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiology, **139**, 1125-1137.

Tocquin, P., Corbesier, L., Havelange, A., Pieltain, A., Kurtem, E., Bernier, G. and PeÌ • rilleux, C. (2003) A novel high efficiency, low maintenance, hydroponic system for synchronous growth and flowering of Arabidopsis thaliana. BMC Plant Biology, **3**.

Treutter, D. (2005) Significance of flavonoids in plant resistance and enhancement of their biosynthesis. Plant Biology, **7**, 581-591.

Van Dam, N.M., Tytgat, T.O.G. and Kirkegaard, J.A. (2009) Root and shoot glucosinolates: A comparison of their diversity, function and interactions in natural and managed ecosystems. Phytochemistry Reviews, **8**, 171-186.

Van Damme, M., Zeilmaker, T., Elberse, J., Andel, A., De Sain-van Der Velden, M. and Van Den Ackerveken, G. (2009) Downy mildew resistance in arabidopsis by mutation of Homoserine Kinase. Plant Cell, **21**, 2179-2189.

Van De Peer, Y. and De Wachter, R. (1997) Construction of evolutionary distance trees with TREECON for Windows: Accounting for variation in nucleotide substitution rate among sites. Computer Applications in the Biosciences, **13**, 227-230.

Vorst, O., de Vos, C.H.R., Lommen, A., Staps, R.V., Visser, R.G.F., Bino, R.J. and Hall, R.D. (2005) A non-directed approach to the differential analysis of multiple LC-MS-derived metabolic profiles. Metabolomics, **1**, 169-180.

Wolde-Meskel, E., Terefework, Z., Lindstrom, K. and Frostegard, A. (2004) Metabolic and genomic diversity of rhizobia isolated from field standing native and exotic woody legumes in southern Ethiopia. Systematic and Applied Microbiology, **27**, 603-611.

Chapter 3

Three-step pathway engineering results in more incidence rate and higher emission of nerolidol and improved attraction of *Diadegma semiclausum*

Benyamin Houshyani · Maryam Assareh · Antoni Busquets · Albert Ferrer · Harro J. Bouwmeester · Iris Kappers

Submitted to Metabolic Engineering

Abstract

The concentration and ratio of terpenoids in the headspace volatile blend of plants have a fundamental role in communication with insects. The sesquiterpene (*E*)-nerolidol is one of the important volatiles with effect on beneficial carnivores for biologic pest management in the field. To optimize *de novo* biosynthesis and reliable and uniform emission of (*E*)-nerolidol, we engineered 1-3 steps of the (*E*)-nerolidol biosynthesis pathway in *Arabidopsis thaliana*.

Introduction of a mitochondrial nerolidol synthase gene mediates *de novo* emission of (*E*)-nerolidol and linalool. Sequential co-expression of the mitochondrial *FPS1* and cytosolic *HMGR1* increased the number of transgenic plants emitting (incidence rate) and the emission rate of both volatiles. No association between the emission rate of transgenic volatiles and the toxicity of the transgenic volatiles could be established. We hypothesize that toxicity of (*E*)-nerolidol was diminished by metabolization to less or non-toxic derivatives, as there was a strong correlation between emission of (*E*)-nerolidol and biosynthesis of putative non-volatile conjugates.

Introduction

A main issue in ecology is to understand how interactions among individual organisms influence food webs and community dynamics. Plants are challenged by a wide range of herbivorous insects and have developed a multitude of defence strategies, including pre-existing physical and chemical barriers, tolerance mechanisms and induced defences that are activated upon attack, such as the biosynthesis of secondary metabolites. In induced direct defence mechanisms, plant-herbivore interaction is affected in various ways such as through the production of repellents, toxins and by hindering digestion (Vasconcelos and Oliveira, 2004; Wittstock and Gershenzon, 2002). Plants can also promote the efficiency of the second trophic level including predators and parasitoids of their enemies by using a so called induced indirect defence mechanism (Gatehouse, 2002).

This indirect induced defence involves biosynthesis of a blend of plant volatiles that attract antagonists of plant enemies upon herbivory; a phenomenon called "cry for help" (Dicke and Baldwin, 2010). This blend of herbivore induced plant volatiles (HIPV) can be complex and herbivore-specific, sometimes consisting of hundreds of compounds (Mumm and Dicke, 2010). The majority of compounds within the HIPV blend are annotated as terpenoids such as (3S)-(E)-nerolidol, (E)-4,8-dimethyl-1,3,7-nonatriene ((E)-DMNT), (E) - β-ocimene and (E,E)-α-farnesene. Other biochemical groups present in HIPV blends are green leaf volatiles, benzenoids and sulphur-or nitrogen-containing compounds. In the past decade, molecular genetic tools have been used to investigate about certain steps of the signalling pathways leading to HIPVs, which are then studied for their attraction of carnivorous arthropods (Halitschke and Baldwin, 2003; Kessler and Baldwin, 2004; Snoeren *et al.* 2009; Van Poecke and Dicke, 2004). In addition, molecular insight into the biosynthesis of HIPVs has allowed the genetic modification of the emission of these volatile compounds. These modified plants were successfully used to study the behaviour of carnivorous arthropods (Aharoni *et al.* 2005; Bouwmeester *et al.* 2003; Degenhardt *et al.* 2003; Kappers *et al.* 2005; Loivamaki *et al.* 2008; Schnee *et al.* 2006; Tholl *et al.* 2005). These studies demonstrate the effectiveness of HIPV modification through biotechnology in enhancement of biological control of above ground as well as below ground (Degenhardt *et al.* 2009) pests.

The C15-sesquiterpene (*E*)-nerolidol and the C11-homoterpene (*E*)- DMNT are examples of HIVPs detected in the odour blend of many plant species after herbivory including cucumber, lima bean, maize, tomato and some *Arabidopsis thaliana* accessions (Ament *et al.* 2004; Bouwmeester *et al.* 1999; Degenhardt and Gershenzon, 2000; Snoeren *et*

al. 2010). *De-novo* emission of these volatiles in the headspace of transgenic *A. thaliana* through genetic engineering alters the behaviour of predatory mites (Kappers *et al.* 2005).

FaNES1 (*Fragaria X ananassa* nerolidol synthase1) is a sesquiterpene synthase that *in vitro* catalyses the biosynthesis of (*E*)-nerolidol from FPP (farnesyl diphosphate) and the monoterpene linalool from GPP (geranyl diphosphate) (Aharoni *et al.* 2003). The failure to produce a reasonable amount of (*E*)-nerolidol in the cytosol or plastids was speculated to be due to the lack of available FPP (the precursor of sesquiterpenoids) in the cytosol or plastids (Aharoni *et al.* 2003). Therefore, Kappers *et al.* (2005) targeted FaNES1 to the mitochondria as *A. thaliana* possesses a long isoform of FPP synthase with a mitochondrial targeting signal (FPS1L) (Cunillera *et al.* 1997). Hereto, the CoxIV signal peptide from yeast (Kohler *et al.* 1997) was used. Transgenic *A. thaliana* plants harbouring the *CoxIV-FaNES1* construct produced substantial amounts of (*E*)-nerolidol and (*E*)-DMNT in their odour blend (Kappers *et al.* 2005). However, the first and next generations of transgenic plants displayed some growth retardation of the basal rosette, a phenomenon that was also observed in cytosolic and plastidic expressers of *FaNES1* (Aharoni *et al.* 2003).

Considering the potential of metabolic engineering to enhance a plant's indirect defence mechanisms, a consistent emission of engineered HIPVs and uniform growth of plants are important traits and growth retardation is hence an unwanted side-effect. One could speculate that the introduction of *CoxIV-FaNES1* diverts the mitochondrial FPP that is also required for ubiquinone and heme A biosynthesis and consequently causes growth inhibition. In order to reduce the harmful effects on plant growth and development, the levels of specific intermediates of the pathway, particularly those at branch-point positions should be controlled (Manzano *et al.* 2004). Knowing that an increase in precursor supply can elevate the level of terpenoids in plants in the presence of the corresponding terpene synthase gene (Aharoni *et al.* 2005; Chen *et al.* 2000; van Herpen *et al.* 2010), the undesirable phenotypes can possibly be diminished by over-expressing *FPS1L* to provide a large pool of FPP in the mitochondria. As isopentenyl diphosphate (IPP, the substrate for FPS1L) might be also a rate limiting factor, over-expression of *A. thaliana HMGR1S* (3-hydroxy-3-methylglutaryl coenzyme A reductase, short isoform) that encodes an earlier step in the mevalonate pathway and has been suggested to be rate limiting (Chappell *et al.* 1991) could further reduce the growth retardation. Over-expression of *HMGR1S* has been already shown to prevent necrosis in leaves of FPS1S (short isoform of FPS which is cytosolic) overexpression lines (Manzano *et al.* 2004) and reverse premature senescence on detached leaves of *FPS1L* (long isoform of FPS which is mitochondrial) overexpression lines under continuous illumination (Manzano *et al.* 2006; Manzano *et al.* 2004). Furthermore, transient co-expression of *tHMGR* (truncated form of

HMGR) and FPS with amorphadiene synthase has been shown to improve amorphadiene biosynthesis *Nicotiana benthamiana* (van Herpen *et al.* 2010).

In this study, fine-tuned production of (*E*)-nerolidol in *A. thaliana* was targeted. To achieve this, two rate-limiting genes of the mevalonate pathway, *HMGR1S* and mitochondrial *FPS1L,* are co-expressed with mitochondrial nerolidol synthase. The volatile compounds of the headspace and non-volatile compounds in the semi-polar extracts of wild type and the transgenic lines are analysed. The effect of the volatile blends from intact and herbivore-induced transgenic plants on *Diadegma semiclausum*, a beneficial parasitoid wasp against the Brassicaceae specialist caterpillar *Plutella xylostella,* is investigated. The results of this research are used to select the most efficient transformation strategy to produce a reliable and an efficient amount of transgenic volatiles for fortification of an indirect insect defence mechanism.

Materials and methods

Generation of transgenic plants

Transgenic plants from independent transformation events harbouring either *HMGR1S* (truncated version of hydroxylmethylglutaryl CoA reductase1 lacking its membrane domain) (Manzano *et al.* 2004), *FPS1L* (farnesyl-pyrophosphate synthase1 long isoform) (Manzano *et al.* 2006) or mitochondrial targeted *FaNES1* (*Fragaria* X *ananassa* nerolidol synthase) (Kappers *et al.* 2005) were crossed and selfed to produce independent transgenic lines of *A. thaliana* Col-3 with single, double or triple transgenes (Table 1). The transgenes were all under the control of the constitutive CaMV 35S promoter. Overexpression and presence of the transgenes were confirmed by RT-PCR (data not shown).

Table 1 Wild type (WT) and transgenic *Arabidopsis thaliana* lines in this study. All transgenes were under the control of the constitutive 35s Cauliflower Mosaic Virus promoter

Line	Group	Harbored transgenes
WT	Col-3 (WT)	-
A	FPS1L	FPS1L
B	HMGR1S::FPS1L	HMGR1S + FPS1L
C	HMGR1S::FPS1L	HMGR1S + FPS1L
D	COX	CoxIV[1]-FaNES1
E	COX	CoxIV[1]-FaNES1
F	COX	CoxIV[1]-FaNES1
G	COX	CoxIV[1]-FaNES1
H	COX+	FPS1L + CoxIV[1]-FaNES1
I	COX+	FPS1L + CoxIV[1]-FaNES1
J	COX+	FPS1L + CoxIV[1]-FaNES1
K	COX++	HMGR1S + FPS1L + CoxIV[1]-FaNES1
L	COX++	HMGR1S + FPS1L + CoxIV[1]-FaNES1
M	COX++	HMGR1S + FPS1L + CoxIV[1]-FaNES1

[1] CoxIV: mitochondrial signaling peptide from yeast

Seeds of wild type and transgenic lines were surface-sterilized, germinated on autoclaved 0.9% Daishin agar (Duchefa, The Netherlands) with 0.5 MS + vitamins (pH>6) (Duchefa, The Netherlands) and 10 µg.ml^{-1} BASTA (only transgenics) in 14 cm Petri dishes and incubated at 21 ± 2°C for ca. 10 days under a L8:D16 photoperiod with 80-110 µmol.m^{-2}.s^{-1} PPF. Four-leaf seedlings were transplanted into potting soil (Lentse potgrond BV, Lent, Netherlands, heated to 80°C overnight prior to use) and grown under the same conditions as above. Plants were watered twice a week. To prevent infestation by sciarid larvae, the soil was treated weekly with *Steinernema feltiae* entomopatogenic nematodes (Koppert B.V. Berkel en Rodenrijs, the Netherlands). Six to eight weeks after sowing, fully grown vegetative plants were used for experiments.

Herbivores and parasitoids

The herbivore *Pieris rapae*, the small cabbage white, was reared on Brussels sprouts plants (*Brassica oleracea* var. *gemmifera*, cv Cyrus) in a growth chamber (L16:D8; 20 ± 2°C and 70% RH). For herbivory induction, *Arabidopsis* plants were infested by equally distributing 20 second-instar *P. rapae* larvae per plant over the fully expanded leaves 24 hrs prior to the experiment. Herbivorous larvae were removed from plants before behavioural assays.

Parasitoid wasps, *Diadegma semiclausum*, were reared as described by (Bukovinszky *et al.* 2005). Emerging wasps were provided *ad libitum* with water and honey, and are referred to as 'naïve' wasps as they had neither been exposed to plant material, nor obtained oviposition experience (Bukovinszky *et al.* 2005). The parasitoid is known to be attracted to the volatiles emitted by *P. rapae*-infested *A. thaliana* Col-0 plant (Loivamaki *et al.* 2008; Snoeren *et al.* 2010).

Headspace trapping and analysis of volatiles by GC-MS

In the first headspace experiment, 12 plants of each of the 14 genotypes (wild type and 13 lines) (Table 1) in vegetative stage (6 wks) were sampled in groups (replicates) of 4 plants with similar rosette sizes, over a time period of four days. Headspace was sampled for 4 hrs between 11:00 am and 15:00 pm. In a second headspace experiment, single plants of each of the 14 genotypes, in vegetative stage with equal rosette sizes were sampled over a time period of three weeks. In this experiment, headspace was sampled for 1.5 hrs between 11:00 am and 13:00 pm. For headspace trapping, transgenic plants were carefully placed in a 2.5 L glass jar and the jar was closed with a Viton-lined glass lid having an inlet and outlet. Inlet air

was filtered by passing through stainless steel cartridges (Markes, Llantrisant, UK) containing 200 mg Tenax TA (20/35 mesh, Grace-Alltech, Deerfield, USA).

Headspace trapping was performed under the same conditions as the growing conditions. Collection of volatiles started 15 min after placing the plant(s) in the jar by sucking air out of the jar at a rate of 80 $ml.min^{-1}$ through a cartridge containing 200 mg Tenax TA. After the collection period, Tenax cartridges were dry-purged with nitrogen at 30 $ml.min^{-1}$ for 20 min at ambient temperature to remove water before thermodesorption of the trapped volatiles.

Headspace samples were analysed with a Thermo Trace GC Ultra (Thermo Fisher Scientific, Waltham, USA) connected to a Thermo Trace DSQ (Thermo Fisher Scientific, Waltham, USA) quadrupole mass spectrometer. Volatiles were desorbed from the cartridges using a thermal desorption system at 250°C for 3 min (Model Ultra Markes Llantrisant, UK) with a Helium flow of 30 $ml.min^{-1}$, focused at 0°C on an electronically-cooled sorbent trap (Unity, Markes, Llantrisant, UK). Volatiles were transferred in splitless mode to the analytical column (Rtx-5ms, 30m, 0.25 mm i.d., 1.0 µm film thickness, Restek, Bellefonte, USA) by rapid heating of the cold trap to 250°C. The GC was held at an initial temperature of 40°C for 3.5 min followed by a linear thermal gradient of $10°C.min^{-1}$ to 280°C and held for 2.5 min with a column flow of 1 $ml.min^{-1}$. The column effluent was ionized by electron impact ionization at 70 eV. Mass spectra were acquired by scanning from 45-400 m/z with a scan rate of 3 $scans.s^{-1}$.

Volatile compounds were identified using the deconvolution software AMDIS (NIST, USA). Identified volatiles were annotated using their retention index and mass spectrum. The retention indices and spectra were compared with those of authentic standards (Sigma Aldrich Chemie, Germany, for (*E*)-nerolidol and linalool), described in literature (Adams, 1995; Rohloff and Bones, 2005; Snoeren *et al.* 2010), NIST 2005 library, an in-house mass spectral library and The Pherobase (http://www.pherobase.com/database/kovats/kovats-index.php) on a column with (5%-Phenyl)-methylpolysiloxane stationary phase or equivalent. NIST MS Search v.2.0 was used for *in-silico* mass spectrum comparisons.

The amount of volatiles in the first headspace experiment was estimated semi-quantitatively by calculating the peak area of the characteristic ion mass (Supplementary Table 1) using the Quan Browse application of XCalibur and normalized by the fresh weight of the rosettes. Emitted amounts of (*E*)-nerolidol and linalool in the headspace of individual plants in the second experiment were quantified using characteristic ions for each compound (69 and 93 for (*E*)-nerolidol and linalool, respectively).

Metabolite analysis in leaf extracts using LC-MS

Hundred mg of flash frozen shoot material of each sample was powdered and mixed with 300 µl of ice-cold methanol acidified with 0.1% (v/v) formic acid. After vortexing for 5 s, sonication for 15 min and centrifugation (2500 rpm) for 10 min, the extracts were filtered through syringe filters (Minisart SRP 4, 0.45 µm, Sartorius Stedim Biotech) and collected in glass vials. Five µl of the filtered extract (150 µl) was injected and analysed by liquid chromatography coupled to quadrupole time-of-flight mass spectrometry (LC-QTOF-MS) operated in the negative ionization mode. This analytical system consists of a Waters Alliance 2795 HPLC connected to a Waters 2996 PDA detector and subsequently to a QTOF Ultima V4.00.00 mass spectrometer (Waters, MS technologies, UK). A C18 pre-column (2.0×4 mm, Phenomenex, USA) was coupled to an analytical column (2.0×150 mm, 100Å, particle size 3 µm, Phenomenex, USA) and degassed eluent A (1:1000 formic acid: ultrapure water) and eluent B (1:1000 formic acid: acetonitrile) were pumped at 0.19 $ml.min^{-1}$ into the HPLC system. The gradient was increased linearly from 5% to 35% eluent B in 45 min. Afterwards, the column was washed and equilibrated for 15 min before the next injection. A collision energy of 10 eV was used for full-scan LC-MS in the range of m/z 100 to 1,500. Leucine enkephalin ($[M - H]^-$ = 554.2620) was used as a lock mass for online mass calibration.

Data Analysis

GC-MS and LC-MS acquired data were processed using MetAlign (Lommen, 2009) for noise elimination and alignment of the data points. Data were normalized with the fresh weight of the analysed samples and subsequently Log10 transformed and scaled symmetrically by the standard deviation using CANOCO (Biometris, Wageningen, NL). Detrended Correspondence Analysis (DCA) was used to check the length of the gradient (L) and due to an $L < 4$, the linear ordination techniques, principal component analysis (PCA) and redundancy analysis (RDA) were selected to visualize variation across chromatograms and correlations between metabolites using CANOCO (Smilauer, 2003).

PASW statistics 17 from SPSS was used for ANOVA and other mean comparisons as well as correlation analysis.

Behavioural assays using parasitoid wasps

Parasitoid behaviour towards the volatile blend of transgenic lines was compared to that of wild type plants using a closed-system Y-tube olfactometer as described by (Bukovinszky *et al.* 2005). In short, filtered air was led through activated charcoal and split

into two air streams (4 $L.min^{-1}$) that were led through five-litre glass vessels containing the odour sources (4 plants of the same rosette size in each vessel). The olfactometer was illuminated with artificial light from above at an intensity of 60 ± 5 $\mu mol.m^{-2}$ sec^{-1} PPFD on the surface of the desk. All experiments were conducted in a ventilated room (20 ± 2 °C). To confirm the attractiveness of (*E*)-nerolidol in parasitoid behaviour, 96% (*E*)-nerolidol (Sigma-Aldrich Chemie, Germany) released from a glass capillary was compared with clean air in the same setup.

Naive, 3-7 days-old female *D. semiclausum* were individually introduced into the Y-tube olfactometer using a glass tube. Upon release in the olfactometer, parasitoid behaviour and parasitoid choice for one of the two odour sources was observed. Parasitoids that did not choose for any of the two arms of the olfactometer within five minutes after release, or choose an arm but did not make a final choice within ten minutes after release were considered as non-responding individuals, and were excluded from preference statistical analysis. Odour sources were interchanged after every five parasitoids to compensate for any unforeseen asymmetry in the set-up.

Parasitoid preference for odour sources was statistically analysed using a Chi-square test, with the null-hypothesis that parasitoids did not have a preference for any of the two odour sources.

Results

Phenotypic changes in transgenic lines

Average plant weights showed a mixture of retarded, accelerated and normal growing habits for the transgenic lines under short day (L8:D16) growing conditions. Fresh weight of some of the transgenic lines differed significantly ($\alpha = 0.05$) from that of the wild type Col-3 (Fig. 1). The observed differences were not associated with the introduced *FaNES*1 gene product as the significantly smallest lines were not among the highest (*E*)-nerolidol/linalool (derivatives) producers and vice versa (Fig. 1). Lines E (COX) and J (COX+) had on average a larger rosette compared with the wild type and line C (*HMGR1S::FPS1L*) and D (COX) that were not emitting (*E*)-nerolidol and line I (COX+) were significantly lighter than wild type plants; Lines F, H, K, L and M tend to be lighter on average (Fig. 1). These results were reproducible to some extent as in an independent experiment under the same growing conditions with 4 week old lines I (COX+, 170 mg) and L (COX++, 120 mg), they weighed significantly less than Col-3 (341 mg) (ANOVA, Tukey, n = 7, p-value < 0.01 for both comparisons), while there was no significant difference between the two transgenic lines (p-value = 0.137).

Effect of pathway engineering on the emission of linalool and (E)-nerolidol

Introduction of the mitochondrial targeted linalool/nerolidol synthase gene from strawberry (*CoxIV-FaNES1*) into *A. thaliana* is known to cause *de novo* emission of (*E*)-nerolidol and (*E*)-DMNT (Kappers *et al.* 2005).

In the first headspace experiment, 4 plants of each transgenic line were combined. The headspace of lines A (*FPS1L*), B and C (*HMGR1S::FPS1L*) was similar to the headspace of wild type plants as they produced neither linalool nor (*E*)-nerolidol (Fig. 1). Emission (*E*)-nerolidol was detected in transgenic lines E and F (COX, *CoxIV-FaNES1* expressing lines). Linalool was present in 3 of the 4 analysed COX transgenic lines (D, E & F), while it was absent in the wild type (Fig. 1).

Addition of *FPS1L* in *FPS1L::FaNES1* lines (COX+) increased the emission of both linalool and (*E*)-nerolidol (Fig. 1), which was statistically significant in line I compared with COX lines D, E, F and G (Fig. 1), while lines H and J did not differ significantly from COX lines in emission. The triple construct COX++, overexpressing *HMGR1S* in combination with *FPS1L* and *CoxIV-FaNES1* also increased the average emission of both linalool and (*E*)-

nerolidol (Fig. 1). Transgenic lines L and M showed significantly higher emission of linalool and (*E*)-nerolidol, respectively, compared with COX+ lines (Fig. 1).

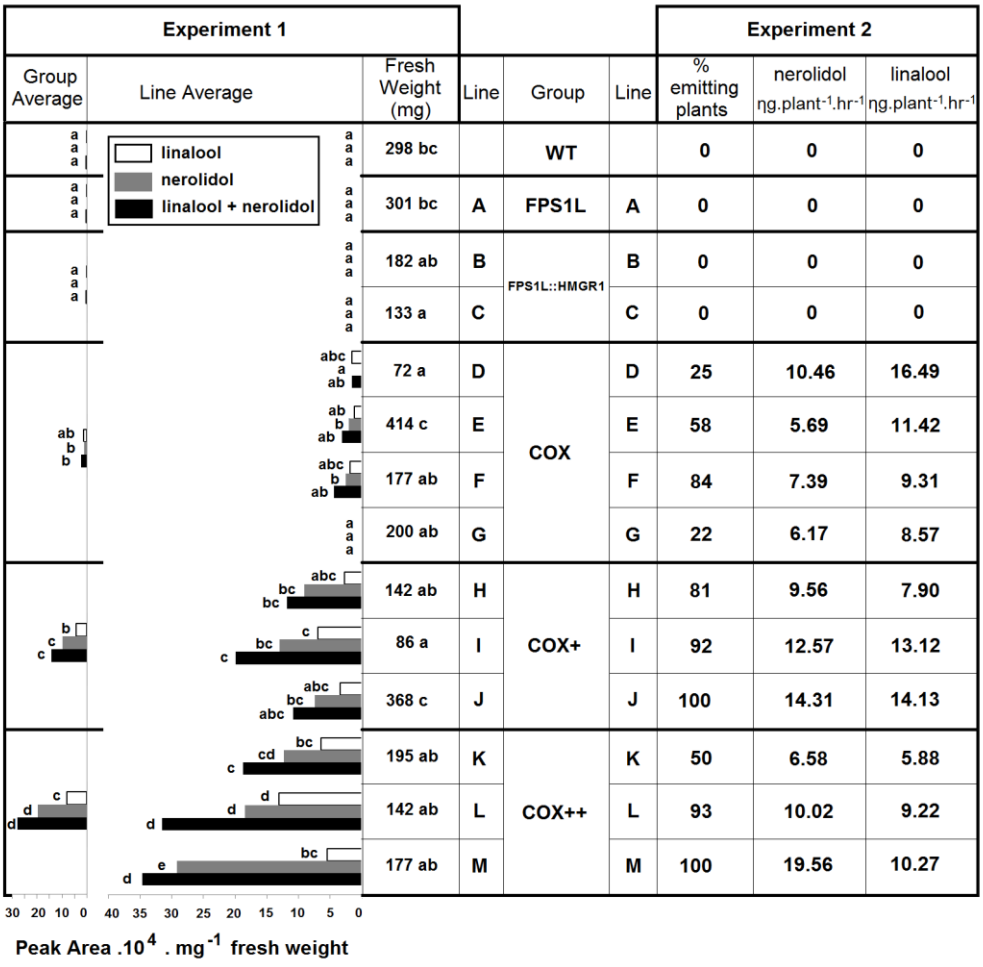


Figure 1 Linalool and (*E*)-nerolidol production bywild type (WT) and transgenic lines in two experiments. Experiment 1 compares transgenic lines based on the peak area per mg fresh weight of the examined plants and the average fresh weight of each plant 6 weeks after sowing. Different letters indicate significant differences for a volatile or the fresh weight between genotypes (ANOVA in conjunction with Tukey's test, $\alpha = 0.05$). Experiment 2 compares transgenic lines based on the percentage of examined plants that emitted the transgenic volatile at a detectable level during the experiment and the average quantity of linalool and (*E*)-nerolidol in emitting plants in ng per plant per hour. No significant difference was observed between transgenic volatile emitters in experiment 2.
COX, CoxIV- FaNES1 lines; COX+, FPS1L:: CoxIV-FaNES1 lines and COX++, HMGR1::FPS1L:: CoxIV-FaNES1 lines (CoxIV, the mitochondrial signal)

In a second headspace experiment the headspace of individual plants from each transgenic line was trapped. In two out of four analysed COX lines, less than half of the analysed plants showed (*E*)-nerolidol and linalool (Fig. 1) at a level above the detection limit during the period of headspace trapping (1.5 hrs). Emitting plants within the COX group did not differ in quantities of (*E*)-nerolidol and linalool emitted by COX+ and COX++ plants (Fig. 1). The percentage of emitting plants in line F (COX) was comparable with the percentages of emitting plants in the COX+ and COX++ groups (> 80%) (Fig. 1). Line F was also the highest emitter within the COX group in the first headspace experiment (Fig. 1).

More than 80% of the examined plants of both COX+ and COX++ lines emitted (*E*)-nerolidol and linalool, except for line K where only 50% of the plants showed emission (Fig. 1). The average amounts of emitted (*E*)-nerolidol and linalool did not differ significantly between lines and varied across plants from 0.3 to 43 ng.plant^{-1}.hr^{1} for (*E*)-nerolidol and 0.3 to 40 ng.plant^{-1}.hr^{-1}for linalool, respectively (Fig. 1).

The amount of (*E*)-nerolidol and linalool emitted by producing plants (Fig. 1) showed a significant correlation (R = 0.67, *p*-value < 0.01). As discussed above, the emission of (*E*)-nerolidol and linalool showed no significant differences between transgenic lines possibly due to the high variation. However, a chi-square test indicated that the assumption of transgenic volatile production by *all* COX plants was violated (*p*-value = 0.002) while this assumption did hold for the COX+ and COX++ lines (*p*-value = 0.92 and 0.28, respectively).

(*E*)-DMNT was clearly present in a number of COX, COX+ and COX++ plants (data not shown), but due to inconsistency in emission and high level of noise in the corresponding retention region we were unable to (semi) quantify the corresponding peak for further analysis.

Multivariate analysis of the headspace volatiles

To examine the overall effect of the introduced genes on the headspace profile, multivariate data analysis and ANOVA were performed on headspace data of the first experiment. Due to a low signal/noise ratio in the analysis, we initially browsed profiles and searched for known compounds based on previous studies on volatiles of *A. thaliana* (Rohloff and Bones, 2005; Snoeren *et al.* 2010). This resulted in a dataset consisting of semi-quantitative data of 53 putative volatiles representing various biochemical groups of compounds (Supplementary Table 1). The retention time and characteristic mass for each of these compounds were used for *in-silico* quantification of the corresponding peak area (semi-quantified data).

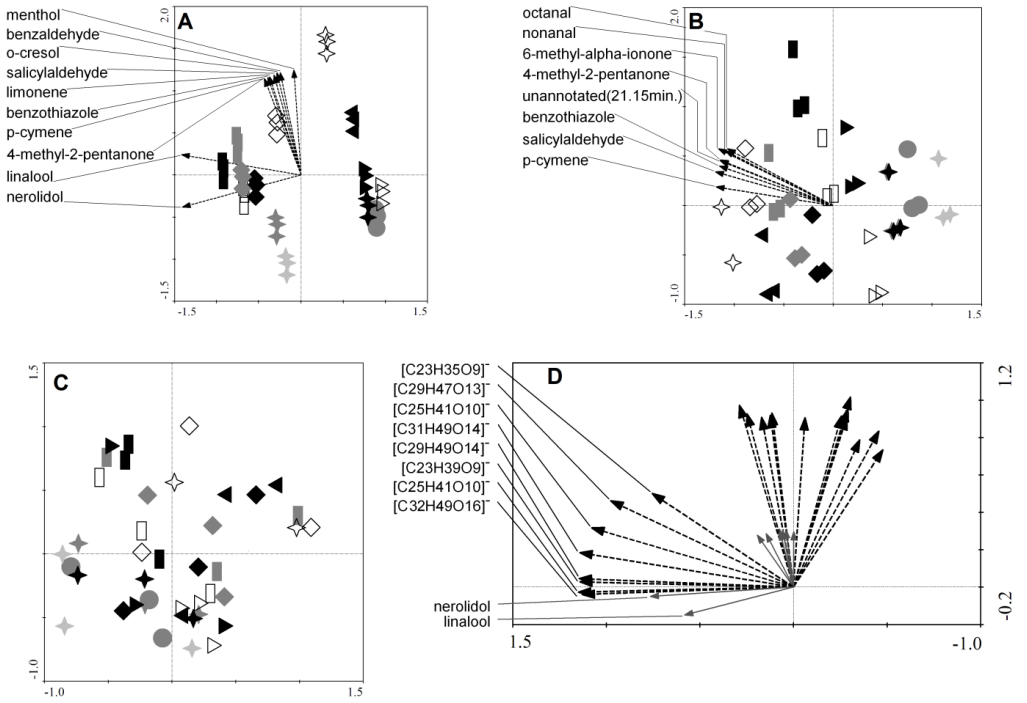


Figure 2 Ordination plots by 1) RDA using semi-quantified data including (A) and excluding (B) linalool and nerolidol with lines and biological replicates as explanatory variables for the observed variation in the peaks area. Metabolites indicated are putatively identified volatiles fitting more than 55% to the ordination model and important for the observed clustering of samples; 2) PCA using the semi-quantified data excluding linalool and nerolidol from the data set (C); 3) RDA showing the relationship between the non-volatile metabolites (dotted lines) that fitted more than 30% to the ordination diagram and head space volatiles (solid lines) that correlated between < -0.3 or > 0.3 with the ordination axes. Some of the proposed elemental compositions for non-volatile metabolites are indicated on the left. Arrows pointing the same direction indicate correlation. To focus on transgenic associated compounds only strongly correlated metabolites with linalool and nerolidol were putatively annotated.
● WT (wild type), ◄ A (FPS1L), ▷ B (HMGR1S::FPS1L), ▶ C (HMGR1S::FPS1L), ✧ D (COX), ✦ E (COX), ✦ F (COX) ✦ G (COX), ◇ H (COX+), ◆ I (COX+), ◆ J (COX+), ▯ K (COX++), ▮ L (COX++), ▮ M (COX++)

RDA was used to pinpoint the volatile compounds that separate transgenic lines from the wild type. Subsequently, any overlap with the list of significantly different volatiles was checked by ANOVA ($\alpha = 0.05$). RDA on semi-quantified data was performed with wild type and transgenic lines and biological replicates as explanatory variables for the observed variation in the semi-quantitative data. Lines and replicates together explained 53.5% of the variation in the data. RDA plot with the first (19.6%) and second (15.5%) ordinate showed a

clear separation between lines based on the abundance of (*E*)-nerolidol and linalool (Fig. 2A). *FPS1L* and *HMGR1S::FPS1L* and two COX lines (A, B, C, D and G) without (*E*)-nerolidol in their headspaces were clustered with wild type samples on one side of the ordination diagram. Volatiles that fit more than 60% (arbitrary threshold) to the ordination model are shown in figure 2A with linalool and (*E*)-nerolidol having the largest contribution to the observed separation between lines.

ANOVA on the semi-quantified data showed that only six of the annotated volatiles differ significantly across the lines ($\alpha = 0.05$) including 4-methyl-2-pentanone, 3-methyl-3-butenenitrile, 3-butenyl isothiocyanate, (*E*)-β-ocimene, linalool and (*E*)-nerolidol. From this set of volatiles only linalool and nerolidol differed significantly between (*E*)-nerolidol producing and non-producing lines and therefore the other four volatiles did not appear on the RDA plot (Fig. 2A). A similarity matrix with all significantly different volatiles showed no correlation between linalool or (*E*)-nerolidol abundance and other significantly different volatiles (data not shown). Excluding linalool and (*E*)-nerolidol from the dataset shortened the distance between (*E*)-nerolidol producing and non-producing lines as illustrated by the RDA diagram in figure 2B. Volatiles that fit more than 55% to the ordination model (arbitrary threshold, using a 60% threshold resulted in only one compound) and therefore important for the observed configuration on the ordination plot are pinpointed. These volatiles showed no significant difference across the lines (except for 4-methyl-2-pentanone). A PCA on the same dataset showed no clusters based on known structures (groups or lines) (Fig. 2C). All above observations suggest no major changes in the semi-quantified volatile profiles as a consequence of the introduction of the transgenes, except for the presence of (*E*)-nerolidol and higher quantities of linalool.

Analysis of conjugates of transgenic volatiles

To study metabolization of the transgenic volatiles *in planta*, methanol extracts of leaves of transgenic lines from the first experiment were analysed by LC-QTOF-MS. Biological replicates of extraction were the same as the replicates of the headspace trapping experiment. We followed a data integration and multivariate data analysis approach to search for any correlation between the analysed volatiles in the headspace and the analysed metabolites within the leaf. An RDA plot with headspace volatile and non-volatile metabolites. For eight of the non-volatile metabolites a high correlation with the emission of linalool and/or (*E*)-nerolidol was demonstrated in the RDA plot (Fig. 2D). ANOVA confirmed this correlation as the accumulated total ion count of all the RDA pinpointed non-volatile metabolites were significantly different across transgenic lines ($\alpha = 0.05$) (Fig. 3A).

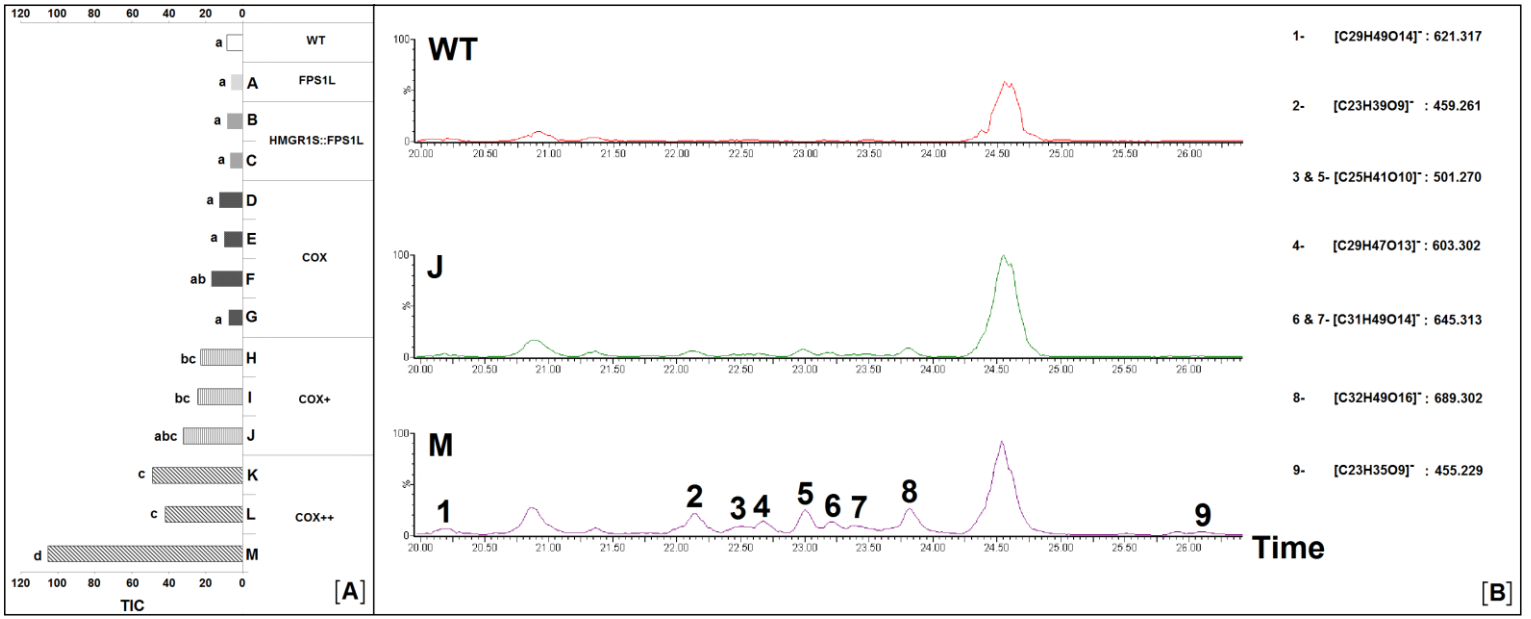


Figure 3 (A). Accumulated total ion count (TIC) of the LC-QTOF-MS metabolites that correlated with the abundance of linalool and nerolidol in the and transgenic lines. Different letters on the left side of the bars indicate a significant difference ($\alpha = 0.05$). (B). Representative chromatograms of a wild type (WT), COX+ (line J) and COX++ (line M) plant in the retention interval (~ 6 minutes) where all of the non-volatile metabolites correlating with linalool and nerolidol (indicated by numbers) were detected. Chromatograms were scaled to the highest TIC. The corresponding elemental compositions proposed for the parental negative ions of the metabolites are shown. COX, CoxIV- FaNES1 lines; COX+, FPS1L:: CoxIV-FaNES1 lines and COX++, HMGR1S::FPS1L:: CoxIV-FaNES1 lines (CoxIV, the mitochondrial signal)

Locating those reconstructed metabolites on the LC-QTOF-MS chromatograms pointed to a short interval within the chromatogram where all these metabolites were present in COX+ and COX++ lines while they were absent in all other transgenic lines and wild type plants (Fig. 3B).

The RDA plot (Fig. 2D) also shows other volatile and non-volatile metabolites that are correlating with each other. As the representing arrows are pointed to another direction (vertical), they do not correlate with the abundance of the transgenic volatiles. Hence, they were not selected for further investigations.

In-silico identification using the *Seven Golden Rules* software (http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/Software/) was used to propose the most likely elemental composition for the parental ion mass corresponding to the metabolites with high correlation to linalool and (*E*)-nerolidol on the RDA plot. In this way all selected metabolites could putatively be annotated as derivatives of (*E*)-nerolidol including glycosylated (with hexose or pentose), acetylated and malonylated derivatives of hydroxynerolidol (Table 2). A structure was also proposed for the annotated metabolites based on the most common structures of previously reported terpene conjugates (Aharoni *et al.* 2003; Yang *et al.* 2011) (Table 2). Following this approach we could not find any linalool derivative.

Behavioural assays using parasitoid wasps

The effect of the introduced volatile compounds on the behaviour of parasitoid wasps was investigated in a closed Y-tube olfactometer system using lines I and L, representative for COX+ and COX++ groups. Plants of these two lines consistently produced (*E*)-nerolidol and more linalool than the wild type plants and showed a significant difference towards each other in this respect in the second headspace experiment (Fig. 1). The attraction of naive female *D. semiclausum* parasitoid wasps to the volatile blend of non-infested transgenic lines was compared to that of the blend of wild type *A. thaliana* plants. To investigate the potential changes in plant-parasitoid interaction upon infestation, we conducted similar assays in parallel using *P. rapae*-infested transgenic and wild type plants.

First, the effect of (*E*)-nerolidol on the attraction of parasitoid wasps was examined in our set up by adding pure (*E*)-nerolidol, slowly released from a glass capillary into the air stream from an empty glass jar compared with empty jars; 69.0% of *D. semiclausum* wasps significantly preferred the (*E*)-nerolidol containing air over the air without (*E*)-nerolidol (chi-square test, $p < 0.01$, n=42) (Fig. 4).

Table 2 Putative annotation of the proposed elemental compositions for the selected parental ion masses in LC-QTOF-MS. Proposed structures are based on the most common structures of naturally occurring terpenes glycosides. A molecule of nerolidol is shown in the first row right.

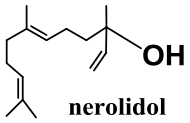
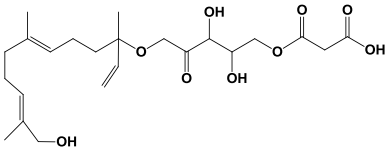
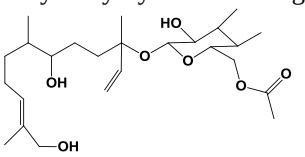
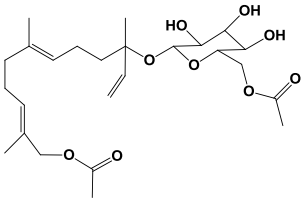
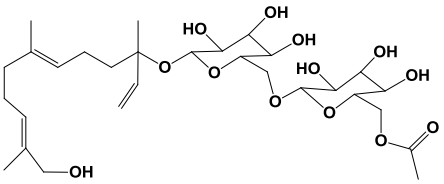
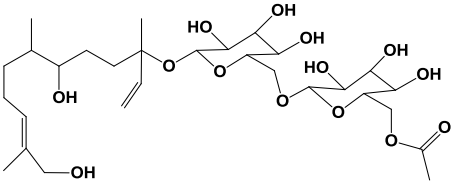
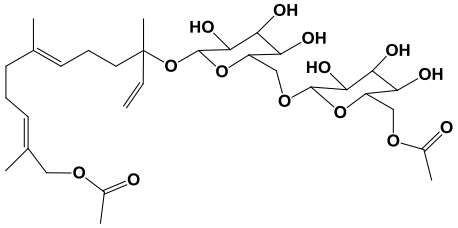
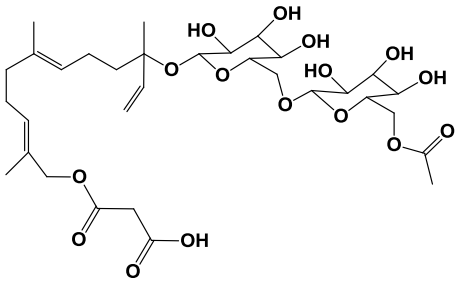
Negative ion mass	Proposed elemental composition	Putative annotation	 nerolidol
455.229	$[C_{23}H_{35}O_9]^-$	hydroxynerolidol-malonyl-ketopentoside	
459.261	$[C_{23}H_{39}O_9]^-$	dihydroxy-hydronerolidol-glucopyranoacetate	
501.270	$[C_{25}H_{41}O_{10}]^-$	hydroxynerolidol-acetyl-glucopyranoacetate	
603.302	$[C_{29}H_{47}O_{13}]^-$	hydroxynerolidol-glucopyranosylglucopyranoacetate	
621.317	$[C_{29}H_{49}O_{14}]^-$	dihydroxy-hydronerolidol-glucopyranosylglucopyranoacetate	

Table 2 Continued

Negative ion mass	Proposed elemental composition	Putative annotation
645.313	$[C_{31}H_{49}O_{14}]^-$	hydroxynerolidol-acetyl-glucopyranosylglucopyranoacetate 
689.302	$[C_{32}H_{49}O_{16}]^-$	hydroxynerolidol-malonyl-glucopyranosylglucopyranoacetate 

D. semiclausum parasitoid wasps significantly preferred the blend of volatiles emitted by non-infested COX+ transgenic plants to those of non-infested wild-type as 66.4% of responsive wasps chose the blend of non-infested COX+ transgenic plants compared to the blend of non-infested wild-type plants (chi-square test, $p < 0.01$, n=131, Fig. 4). Similarly, the blend of non-infested COX++ plants was significantly more attractive for parasitoid wasps than that of non-infested wild-type plants by attracting 64.2% of the responsive wasps (chi-square test, $p < 0.01$, n=134) (Fig. 4).

Herbivory by *P. rapae* caterpillars resulted in a higher attractiveness for parasitoid wasps as the blend of wild-type infested plants attracted 75% of the wasps in comparison to

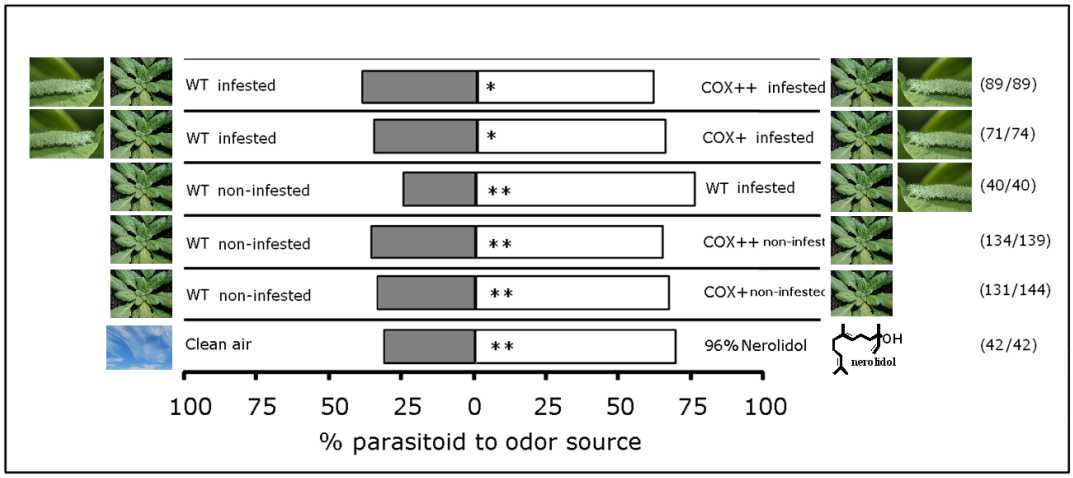


Figure 4 Responses of naive *D. semiclausum* females to the volatiles of two sets of odor sources, as assessed in a Y-tube olfactometer. Each bar represents the percentage of choices for each of the two odor sources as determined in 4-10 replicate experiments. The number of responding wasps of the total number of tested wasps is given between parentheses. Asterisks indicate significance within a choice test; * $p<0.05$ and ** $p<0.01$ (Chi-square test).
COX+, FPS1L:: CoxIV-FaNES1 lines and COX++, HMGR1::FPS1L:: CoxIV-FaNES1 lines (CoxIV, the mitochondrial signal)

the blend of non-infested wild-type plants (chi-square test, $p < 0.01$, n=40) (Fig. 4). When transgenic plants were infested by caterpillars, the behaviour of the parasitoids did not change in comparison with non-infested plants. Infested COX+ plants attracted 64.8% of the wasps (chi-square test, $p < 0.05$, n=71) and infested COX++ plants attracted 60.7% of wasps (chi-square test, $p < 0.05$, n=89) when tested against infested wild-type (Fig. 4).

Discussion

In a previous study, *A. thaliana* Columbia lines expressing mitochondrial targeted *FaNES1* emitted (*E*)-nerolidol and (*E*)-DMNT in their headspace (Kappers *et al.* 2005). This correlated with enhanced attraction of predatory mites and was proposed as a potential trait for biological control of pests in the field. However, transgenic lines emitted varying amounts of the introduced volatiles and showed retard growth compared with the wild type counterpart. These drawbacks were speculated to be due to the channelling of the mitochondrial FPP towards the biosynthesis of (*E*)-nerolidol by expression of *FaNES1*, decreasing the FPP pool for the other branches of the pathway that are vital for plant growth. The growth retardation would present a barrier against the potential application of volatile compounds pathway engineering for plant defence. In order to reduce the undesired effects, we over-expressed *FPS1L* and *HMGR1* to improve the flux through the mevalonate pathway.

Headspace analysis of individual plants showed that introduction of *FPS1L* in COX+ and COX++ lines made them more likely to emit (*E*)-nerolidol compared to COX lines. Having larger numbers of emitting plants made COX+ and COX++ lines produce significantly more (*E*)-nerolidol and also linalool in the pooled headspace analysis. These observations suggest that a minimum of precursor or its biosynthesis activity is required for (*E*)-nerolidol biosynthesis through constitutive expression of mitochondrial *FaNES1*.

Furthermore, over expression of *HMGR1* also significantly enhanced the production of (*E*)-nerolidol compared with over expression of *FaNES1* and *FPS1L* alone in COX+ (Fig. 1, experiment 1). However, the amount of the produced (*E*)-nerolidol by individual emitting plants in the COX++ lines does not significantly surpass the level produced by individual emitting plants in COX+ and COX group (Fig. 1, Experiment 2). Perhaps this is due to a negative feedback mechanism coupled to the accumulated products to halt the biosynthesis and/or the activation of conjugation of (*E*)-nerolidol to glycosides. Indeed, there is a continuous increase in hydroxynerolidol derivatives in COX, COX+ and COX++ plants. Also, free (*E*)-nerolidol levels could be balanced through conversion into DMNT and/or the uptake of cytosolic IPP by mitochondria could also be limiting.

Although nerolidol derivatives were present in leaves of transgenic plants, no correlation could be established between any putative linalool derivative and the emission of the transgenic volatiles in our LC-QTOF-MS analysis. Nevertheless, linalool conjugates were detected in plastidic *FaNES1* expressing *A. thaliana* plants by enzymatic hydrolysis (glycosidase) and subsequent GC-MS analysis of the released aglycones (Aharoni *et al.* 2003).

In this indirect method of detection, the nature of the original conjugations (glycones) and the hydrolysis effectiveness of all types of glycosides remained unknown. To know the metabolic fate of both linalool and (*E*)-nerolidol, we used an untargeted LC-MS approach, data integration and putative identification by Seven Golden Rules software to detect novel non-volatile compounds in the transgenic lines. As each untargeted method is also limited by extraction solvent affinities, behaviour of the chromatography column and the potential of molecular ionization, the presence of additional non-volatile compounds such as linalool conjugates cannot be excluded.

Physiologically, conjugation by glycosides diminishes the chemical reactivity of compounds (Von Rad *et al.* 2001) and together with malonylation facilitates accumulation, storage or transport of phytotoxic secondary metabolites (Von Rad *et al.* 2001) to reduce toxicity or store them temporarily for subsequent biotic-induced enzymatic release (Yang *et al.* 2011) such as attraction of beneficial insects (James, 2005) or repelling pests (Halbert *et al.* 2009; Wei *et al.* 2004). Here, we show formation of hydroxynerolidol conjugates in COX+ and COX++ lines, suggesting phytotoxic effects of this compound in high concentrations. The fact that conjugation of sesquiterpenoid volatiles have not commonly been under investigation could result in underestimating the functional efficiency of the introduced sesquiterpene biosynthesis genes (Kappers *et al.* 2005; Wu *et al.* 2006; Zhang *et al.* 2011).

Mitochondria are known sites for FPP biosynthesis (Cunillera *et al.* 1997) and no monoterpene synthase activity has been reported in these compartments (Yu and Utsumi, 2009). Moreover, the CoxIV signalling peptide is highly efficient in targeting proteins to mitochondria (Kohler *et al.* 1997) in such levels that no trace of linalool was reported in *CoxIV-FaNES1* lines of a previous study in which lines were selected without growth retardation for the sake of predatory mite behavioural bioassays (Kappers *et al.* 2005). Emission of linalool in wild type plants is limited to the inflorescence stage (Aharoni *et al.* 2003). Knowing this, the presence of linalool in vegetative stages of transgenic *FaNES1* lines was not expected. However, both (*E*)-nerolidol and linalool were detected in the headspace of COX, COX+ and COX++ plants in the vegetative stage. The dual monoterpene and sesquiterpene synthase activity of the FaNES1 recombinant protein allows the production of linalool in presence of GPP in *E. coli* (Aharoni *et al.* 2004) and transgenic plants with plastidial FaNES1 emitted high levels of linalool and (very) low levels of the (*E*)-nerolidol with a prominent retarded growth phenotype (Aharoni *et al.* 2003). As a result of the dual functionality of the FaNES1-encoded protein, we could simultaneously show the possibility of forming comparable and correlated amounts of both types of terpenoids strongly suggesting availability of GPP, the precursor for linalool in the mitochondria. *FPS1L* and its homolog *FPS2* can actually catalyse the two sequential condensation reactions: IPP with first DMAPP

to form GPP and GPP with the second IPP leading to FPP formation (Cunillera *et al.* 1996; Delourme *et al.* 1994). The results show that the amount of GPP is high enough to allow *de novo* linalool biosynthesis in substantial amounts, even if *FPS1L* is not over-expressed (in COX lines). The relatively high positive correlation (R = 0.67) between (*E*)-nerolidol and linalool emission indicates no competition between the branches of the biosynthetic pathways in mitochondria in the presence of *FaNES1*that can be another evidence for the presence of GPP in mitochondria

Nevertheless, this speculation needs experimental evidence and does not exclude yet the possibility of the exchange of prenyl precursors including GPP between the plastids, cytosol and the mitochondria, as demonstrated by several authors for plastids and cytosol (Adam and Zapp, 1998; Lichtenthaler, 2000). Wu. *et al.* (2006) targeted limonene synthase to cytosol and reported low but measurable levels of the monoterpene limonene, indicative of a small cytosolic pool of GPP. Therefore, leakage of the mitochondrial targeted *FaNES1* to cytosol can be also an explanation for the presence of linalool despite the high efficiency of CoxIV signalling peptide (Kohler *et al.* 1997) and substantial amount of linalool in the headspace.

Our data do not show any association between biosynthesis of *FaNES1* transgene products and growth retardation, as most of the transgenic lines did not significantly differ from the wild type counterpart or even appeared significantly larger in size. A smaller plant size was also observed within the lines in *HMGR1S::FPS1lL* group without any trace of (*E*)-nerolidol in the headspace. Our observations suggest that the retard growth phenotype can be due to the constitutive expression of introduced transgenes by using the CaMV 35s promoter and/or positional effect of the transgenes. However, these speculations do not exclude the additive effect of *FaNES1* transgene product toxicity yet such as inhibitory effects of linalool on the respiratory chain in mitochondria (Usta *et al.* 2009) and antioxidants against reactive oxygen species (Pompella *et al.* 2003) that already has been reported. Whether (*E*)-nerolidol has similar effects on the physiological state of the cell needs to be investigated.

Inconsistent emission of (*E*)-DMNT was observed by the analysed transgenic lines with Col-3 background, despite its more consistent emission in a previous study using *COX-FaNES1* expressing Col-0 lines (Kappers *et al.* 2005). Besides the genetic background difference, we can relate this to the fact that in the latter study, SPME GC-MS was used to analyse volatile compounds of the detached leaves. (*E*)-DMNT is an inducible volatile (Lee *et al.* 2010), having its responsible enzyme induced in the presence of (*E*)-nerolidol (its precursor) in the detached leaves could explain its more consistent detection in Kappers *et al.* (2005) in contrast to our study.

D. semiclausum wasps significantly prefer the volatile blend of transgenic COX+ and COX++ plants over that of the wild type. They are also more attracted to the volatile source when it either contains pure (*E*)-nerolidol (versus clean air) or the headspace blend of induced plants (versus non-induced). Wild type Col-3 vegetative plants are known to lack linalool (Aharoni *et al.* 2003). Although linalool has been reported to be attractive to predatory mites (Dicke *et al.* 1990), transgenic linalool producing *Arabidopsis* plants repel aphids (Aharoni *et al.* 2003) and *D. semiclausum* wasps are not attracted to linalool as well (unpublished data). As multivariate data analysis do not show any other major difference between the headspace of non-infested wild type and transgenic plants other than (*E*)-nerolidol and linalool, it seems that the presence of linalool has no effect on the attraction of wasps in the presence of (*E*)-nerolidol. Interestingly, this attraction for *D. semiclausum* is not disrupted when a non-host caterpillar, *P. rapae*, damages the transgenic lines and induces specific herbivory induced volatiles. Whether these lines are still attractive for host parasitoids of *P. rapae* (such as *Cotesia rubecula*) needs to be investigated.

Supplementary Table 1 Volatile compounds in the headspace of wild type and transgenic lines of this study

	Compound	CAS Nr.	Retention Time (min)	Experimental Retention Index	Literature Retention Index	Representative Ion	Ion Identification	
1	3-butene nitrile	109-75-1	5.51	657	658	67	MS	LRI
2	1-methoxy-2-propanol	107-98-2	5.89	673	672	47	MS	LRI
3	1-penten-3-ol	616-25-1	6.09	682	682	57	MS	LRI
4	2-pentanone	107-87-9	6.19	686	687	86	MS	LRI
5	acetoin	513-86-0	6.7	708	705	45	MS	LRI
6	methyl cyclohexane	108-87-2	7.15	728	726	83	MS	LRI
7	dimethyl disulfide	624-92-0	7.65	749	752	94	MS	LRI
8	1-pentanol	71-41-0	8.05	766	766	42	MS	LRI
9	2,2-dimethyl propanoic acid	75-98-9	8.73	796	790	102	MS	LRI
10	(*E*)-2-hexenal	6728-26-3	9.98	855	854	98	MS	LRI
11	2-Methylallyl acetate	820-71-3	10.01	856	x	72	MS	x
12	(*Z*)-Hex-3-en-1-ol	928-96-1	10.04	857	859	67	MS	LRI
13	1-methoxy-2-propyl acetate	108-65-6	10.24	867	857	72	MS	LRI
14	allyl isothiocyanate	57-06-7	10.67	887	887	99	MS	LRI
15	1-nonene	124-11-8	10.76	892	892	56	MS	LRI
16	anisole	100-66-3	11.4	924	918	108	MS	LRI
17	α-pinene	80-56-8	11.81	946	940	93	MS	LRI
18	5-ethyl-2(5H)-furanone	80-56-9	12.03	957	984	112	MS	LRI
19	propylbenzene	103-65-1	12.15	964	964	91	MS	LRI
20	3-butenyl isothiocyanate	3386-97-8	12.64	990	978	113	MS	LRI
21	mesitylene	108-67-8	12.66	991	995	105	MS	LRI
22	hemimellitene	526-73-8	12.91	1004	1007	105	MS	LRI
23	(*Z*)-4-hexen-1-ol, acetate	42125-17-7	12.93	1005	1005	67	MS	LRI
24	2,2'-dihydroxydipropyl ether	110-98-5	13.2	1021	1003	89	MS	LRI
25	pseudocumene	95-63-6	13.49	1034	1026	105	MS	LRI
26	2,2'-oxybis-1-propanol	108-61-2	13.53	1040	1051	59	MS	LRI
27	limonene	138-86-3	13.55	1041	1039	68	MS	LRI
28	trans-β-ocimene	3779-61-1	13.74	1052	1044	93	MS	LRI
29	indane	496-11-7	13.78	1054	1048	117	MS	LRI
30	phenylacetaldehyde	122-78-1	13.8	1055	1055	120	MS	LRI

Supplementary Table 1. continued

	Compound	CAS Nr.	Retention Time (min)	Experimental Retention Index	Literature Retention Index	Representative Ion	Ion Identification	
31	linalool	78-70-6	14.63	1103	1103	121	MS	LRI
32	(*E*)-DMNT	19945-61-0	14.9	1120	1113	150	MS	LRI
33	3-acetyl-2,5-dimethylfuran	10599-70-9	14.92	1121	1103	138	MS	LRI
34	4-methyl-2-undecene	91695-32-8	15.11	1133	1158	69	MS	LRI
35	(*S*)-(+)-6-methyl-1-octanol	110453-78-6	15.3	1145	x	97	MS	x
36	1-methylene-1H-Indene	2471-84-3	16.41	1215	x	128	MS	x
37	methyl cyclohexane	119-36-9	16.41	1215	1208	92	MS	LRI
38	2,4-dimethyl benzaldehyde	15764-16-6	16.75	1237	1180	134	MS	LRI
39	6-ethyltetralin	22531-20-0	18.61	1366	1340	131	MS	LRI
40	clovene	469-92-1	19.13	1403	1396	161	MS	LRI
41	longicyclene	1137-12-8	19.31	1417	1382	204	MS	LRI
42	isolongifolene	1135-66-6	19.56	1435	1402	148	MS	LRI
43	*β*-caryophyllene	87-44-5	19.7	1446	1440	133	MS	LRI
44	allo-isolongifolene	87064-18-4	19.74	1449	1412	95	MS	LRI
45	longifolene	475-20-7	19.81	1454	1448	161	MS	LRI
46	*α*-thujaplicinol	16643-33-7	20.21	1484	1509	165	MS	LRI
47	*α*-neoclovene	4545-68-0	20.44	1502	1497	189	MS	LRI
48	(*E*,*E*)-α-farnesene	502-61-4	20.62	1516	1520	93	MS	LRI
49	(*E*)-nerolidol	40716-66-3	21.37	1577	1571	93	MS	LRI
50	(*E*,*E*)-TMTT	62235-06-7	21.52	1589	1565	69	MS	LRI
51	hexyl salicylate	6259-76-3	22.89	1704	1682	222	MS	LRI
52	2-ethylhexyl salicylate	118-60-5	24.34	1834	1807	120	MS	LRI
53	farnesylacetaldehyde	66408-55-7	24.57	1856	1861	69	MS	LRI

References

Adam, K.P. and Zapp, J. (1998) Biosynthesis of the isoprene units of chamomile sesquiterpenes. Phytochemistry, **48**, 953-959.

Adams, R.P. (1995) Identification of Essential Oil Components by Gas Chromatography/Mass Spectroscopy.

Aharoni, A., Giri, A.P., Deuerlein, S., Griepink, F., De Kogel, W.J., Verstappen, F.W.A., Verhoeven, H.A., Jongsma, M.A., Schwab, W. and Bouwmeester, H.J. (2003) Terpenoid Metabolism in Wild-Type and Transgenic Arabidopsis Plants. Plant Cell, **15**, 2866-2884.

Aharoni, A., Giri, A.P., Verstappen, F.W.A., Bertea, C.M., Sevenier, R., Sun, Z., Jongsma, M.A., Schwab, W. and Bouwmeester, H.J. (2004) Gain and loss of fruit flavor compounds produced by wild and cultivated strawberry species. Plant Cell, **16**, 3110-3131.

Aharoni, A., Jongsma, M.A. and Bouwmeester, H.J. (2005) Volatile science? Metabolic engineering of terpenoids in plants. Trends Plant Sci., **10**, 594-602.

Ament, K., Kant, M.R., Sabelis, M.W., Haring, M.A. and Schuurink, R.C. (2004) Jasmonic acid is a key regulator of spider mite-induced volatile terpenoid and methyl salicylate emission in tomato. Plant Physiology, **135**, 2025-2037.

Bouwmeester, H.J., Matusova, R., Zhongkui, S. and Beale, M.H. (2003) Secondary metabolite signalling in host-parasitic plant interactions. Current Opinion in Plant Biology, **6**, 358-364.

Bouwmeester, H.J., Verstappen, F.W.A., Posthumus, M.A. and Dicke, M. (1999) Spider mite-induced (3S)-(E)-nerolidol synthase activity in cucumber and lima bean. The first dedicated step in acyclic C11-homoterpene biosynthesis. Plant Physiology, **121**, 173-180.

Bukovinszky, T., Gols, R., Posthumus, M.A., Vet, L.E.M. and Van Lenteren, J.C. (2005) Variation in plant volatiles and attraction of the parasitoid Diadegma semiclausum (HelleÌ • n). Journal of Chemical Ecology, **31**, 461-480.

Chappell, J., VonLanken, C. and Vögeli, U. (1991) Elicitor-inducible 3-hydroxy-3-methylglutaryl coenzyme a reductase activity is required for sesquiterpene accumulation in tobacco cell suspension cultures. Plant Physiology, **97**, 693-698.

Chen, D.H., Ye, H.C. and Li, G.F. (2000) Expression of a chimeric farnesyl diphosphate synthase gene in Artemisia annua L. transgenic plants via Agrobacterium tumefaciens-mediated transformation. Plant Science, **155**, 179-185.

Cunillera, N., ArrÃ³, M., Delourme, D., Karst, F., Boronat, A. and Ferrer, A. (1996) Arabidopsis thaliana Contains two differentially expressed farnesyl-diphosphate synthase genes. Journal of Biological Chemistry, **271**, 7774-7780.

Cunillera, N., Boronat, A. and Ferrer, A. (1997) The Arabidopsis thaliana FPS1 gene generates a novel mRNA that encodes a mitochondrial farnesyl-diphosphate synthase isoform. Journal of Biological Chemistry, **272**, 15381-15388.

Degenhardt, J. and Gershenzon, J. (2000) Demonstration and characterization of (E)-nerolidol synthase from maize: A herbivore-inducible terpene synthase participating in (3E)-4,8-dimethyl-1,3,7-nonatriene biosynthesis. Planta, **210**, 815-822.

Degenhardt, J., Gershenzon, J., Baldwin, I.T. and Kessler, A. (2003) Attracting friends to feast on foes: Engineering terpene emission to make crop plants more attractive to herbivore enemies. Current Opinion in Biotechnology, **14**, 169-176.

Degenhardt, J., Hiltpold, I., Köllner, T.G., Frey, M., Gierl, A., Gershenzon, J., Hibbard, B.E., Ellersieck, M.R. and Turlings, T.C.J. (2009) Restoring a maize root signal that attracts insect-killing nematodes to control a major pest. Proceedings of the National Academy of Sciences of the United States of America, **106**, 13213-13218.

Delourme, D., Lacroute, F. and Karst, F. (1994) Cloning of an Arabidopsis thaliana cDNA coding for farnesyl diphosphate synthase by functional complementation in yeast. Plant Molecular Biology, **26**, 1867-1873.

Dicke, M. and Baldwin, I.T. (2010) The evolutionary context for herbivore-induced plant volatiles: beyond the 'cry for help'. Trends Plant Sci., **15**, 167-175.

Dicke, M., Van Beek, T.A., Posthumus, M.A., Ben Dom, N., Van Bokhoven, H. and De Groot, A. (1990) Isolation and identification of volatile kairomone that affects acarine predatorprey interactions Involvement of host plant in its production. Journal of Chemical Ecology, **16**, 381-396.

Gatehouse, J.A. (2002) Plant resistance towards insect herbivores: A dynamic interaction. New Phytologist, **156**, 145-169.

Halbert, S.E., Corsini, D., Wiebe, M. and Vaughn, S.F. (2009) Plant-derived compounds and extracts with potential as aphid repellents. Annals of Applied Biology, **154**, 303-307.

Halitschke, R. and Baldwin, I.T. (2003) Antisense LOX expression increases herbivore performance by decreasing defense responses and inhibiting growth-related transcriptional reorganization in Nicotiana attenuata. Plant Journal, **36**, 794-807.

James, D.G. (2005) Further field evaluation of synthetic herbivore-induced plan volatiles as attractants for beneficial insects. Journal of Chemical Ecology, **31**, 481-495.

Kappers, I.F., Aharoni, A., Van Herpen, T.W.J.M., Luckerhoff, L.L.P., Dicke, M. and Bouwmeester, H.J. (2005) Plant science: Genetic engineering of terpenoid metabolism attracts bodyguards to Arabidopsis. Science, **309**, 2070-2072.

Kessler, A. and Baldwin, I.T. (2004) Herbivore-induced plant vaccination. Part I. The orchestration of plant defenses in nature and their fitness consequences in the wild tobacco Nicotiana attenuata. Plant Journal, **38**, 639-649.

Kohler, R.H., Zipfel, W.R., Webb, W.W. and Hanson, M.R. (1997) The green fluorescent protein as a marker to visualize plant mitochondria in vivo. Plant Journal, **11**, 613-621.

Lichtenthaler, H.K. (2000) Non-mevalonate isoprenoid biosynthesis: Enzymes, genes and inhibitors. Biochemical Society Transactions, **28**, 785-789.

Loivamaki, M., Mumm, R., Dicke, M. and Schnitzler, J.P. (2008) Isoprene interferes with the attraction of bodyguards by herbaceous plants. Proceedings of the National Academy of Sciences of the United States of America, **105**, 17430-17435.

Lommen, A. (2009) Metalign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Analytical Chemistry, **81**, 3079-3086.

Manzano, D., Busquets, A., Closa, M., Hoyerová, K., Schaller, H., Kamínek, M., Arró, M. and Ferrer, A. (2006) Overexpression of farnesyl diphosphate synthase in Arabidopsis mitochondria triggers light-dependent lesion formation and alters cytokinin homeostasis. Plant Molecular Biology, **61**, 195-213.

Manzano, D., Fernández-Busquets, X., Schaller, H., González, V., Boronat, A., Arró, M. and Ferrer, A. (2004) The metabolic imbalance underlying lesion formation in Arabidopsis thaliana overexpressing farnesyl diphosphate synthase [isoform 1S] leads to oxidative stress and is triggered by the developmental decline of endogenous HMGR activity. Planta, **219**, 982-992.

Mumm, R. and Dicke, M. (2010) Variation in natural plant products and the attraction of bodyguards involved in indirect plant defense. Canadian Journal of Zoology, **88**, 628-667.

Pompella, A., Visvikis, A., Paolicchi, A., De Tata, V. and Casini, A.F. (2003) The changing faces of glutathione, a cellular protagonist. Biochemical Pharmacology, **66**, 1499-1503.

Rohloff, J. and Bones, A.M. (2005) Volatile profiling of Arabidopsis thaliana - Putative olfactory compounds in plant communication. Phytochemistry, **66**, 1941-1955.

Schnee, C., Köllner, T.G., Held, M., Turlings, T.C.J., Gershenzon, J. and Degenhardt, J. (2006) The products of a single maize sesquiterpene synthase form a volatile defense signal that attracts natural enemies of maize

herbivores. Proceedings of the National Academy of Sciences of the United States of America, **103**, 1129-1134.

Smilauer, J.L.a.P. (2003) Multivariate Analysis of Ecological Data using CANOCO Cambridge University Press, Cambridge.

Snoeren, T.A.L., Kappers, I.F., Broekgaarden, C., Mumm, R., Dicke, M. and Bouwmeester, H.J. (2010) Natural variation in herbivore-induced volatiles in Arabidopsis thaliana. Journal of Experimental Botany, **61**, 3041-3056.

Snoeren, T.A.L., van Poecke, R.M.P. and Dicke, M. (2009) Multidisciplinary approach to unravelling the relative contribution of different oxylipins in indirect defense of Arabidopsis thaliana. Journal of Chemical Ecology, **35**, 1021-1031.

Tholl, D., Chen, F., Petri, J., Gershenzon, J. and Pichersky, E. (2005) Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from Arabidopsis flowers. Plant Journal, **42**, 757-771.

Usta, J., Kreydiyyeh, S., Knio, K., Barnabe, P., Bou-Moughlabay, Y. and Dagher, S. (2009) Linalool decreases HepG2 viability by inhibiting mitochondrial complexes I and II, increasing reactive oxygen species and decreasing ATP and GSH levels. Chemico-Biological Interactions, **180**, 39-46.

van Herpen, T.W.J.M., Cankar, K., Nogueira, M., Bosch, D., Bouwmeester, H.J. and Beekwilder, J. (2010) Nicotiana benthamiana as a production platform for artemisinin precursors. PLoS ONE, **5**.

Van Poecke, R.M.P. and Dicke, M. (2004) Indirect defence of plants against herbivores: Using Arabidopsis thaliana as a model plant. Plant Biology, **6**, 387-401.

Vasconcelos, I.M. and Oliveira, J.T.A. (2004) Antinutritional properties of plant lectins. Toxicon, **44**, 385-403.

Von Rad, U., Hüttl, R., Lottspeich, F., Gierl, A. and Frey, M. (2001) Two glucosyltransferases are involved in detoxification of benzoxazinoids in maize. Plant Journal, **28**, 633-642.

Wei, S., Reuveny, H., Bravdo, B.A. and Shoseyov, O. (2004) Hydrolysis of glycosidically bound volatiles from apple leaves (cv. Anna) by Aspergillus niger β-glucosidase affects the behavior of codling moth (Cydia pomonella L.). Journal of Agricultural and Food Chemistry, **52**, 6212-6216.

Wittstock, U. and Gershenzon, J. (2002) Constitutive plant toxins and their role in defense against herbivores and pathogens. Current Opinion in Plant Biology, **5**, 300-307.

Wu, S., Schalk, M., Clark, A., Miles, R.B., Coates, R. and Chappell, J. (2006) Redirection of cytosolic or plastidic isoprenoid precursors elevates terpene production in plants. Nature Biotechnology, **24**, 1441-1447.

Yang, T., Stoopen, G., Yalpani, N., Vervoort, J., de Vos, R., Voster, A., Verstappen, F.W.A., Bouwmeester, H.J. and Jongsma, M.A. (2011) Metabolic engineering of geranic acid in maize to achieve fungal resistance is compromised by novel glycosylation patterns. Metabolic Engineering, **13**, 414-425.

Yu, F. and Utsumi, R. (2009) Diversity, regulation, and genetic manipulation of plant mono- and sesquiterpenoid biosynthesis. Cellular and Molecular Life Sciences, **66**, 3043-3052.

Zhang, Y., Nowak, G., Reed, D.W. and Covello, P.S. (2011) The production of artemisinin precursors in tobacco. Plant Biotechnology Journal, **9**, 445-454.

Chapter 4

Genetic engineering of plant volatile terpenoids: effects on a herbivore, a predator and a parasitoid

Martine Kos · Benyamin Houshyani · Aart-Jan Overeem · Harro J. Bouwmeester · Berhane T. Weldegergis · Joop J. A. van Loon · Marcel Dicke · Louise E. M. Vet

Benyamin Houshyani generated the transformation constructs, produced and characterized the transgenic plants and contributed to the headspace GC-MS analysis, identification of volatile metabolites and writing the manuscript.

Submitted to Pest Management Science

Abstract

Transgenic plants that enhance the effectiveness of natural enemies can be developed by genetic engineering of the biosynthesis of volatile organic compounds (VOCs). Before the commercialisation of such transgenic plants can be pursued, detailed fundamental studies on their effects on herbivores and their natural enemies are necessary.

We constitutively expressed the linalool/nerolidol synthase gene *FaNES1* from strawberry in three *A. thaliana* ecotypes. We tested the behaviour of the aphid *Brevicoryne brassicae*, the parasitoid *Diaeretiella rapae* and the predator *Episyrphus balteatus* exposed to the transgenic plants and their isogenic ecotypes.

Transgenic *FaNES1*-expressing plants emitted (*E*)-nerolidol and larger amounts of (*E*)-DMNT and linalool. *Brevicoryne brassicae* was repelled by the transgenic lines of two of the ecotypes, whereas its performance was not affected. *Diaeretiella rapae* preferred aphid-infested transgenic plants over aphid-infested wild-type plants for two of the ecotypes. In contrast, female *E. balteatus* predators did not differentiate between aphid-infested transgenic or wild-type plants.

Our study suggests that genetically engineering plants to modify their emission of VOCs holds considerable promise for improving control of herbivores. However, our results have to be validated for a crop species before transgenic plants with enhanced attraction of natural enemies can provide interesting new components for Integrated Pest Management.

Introduction

During the last decades, the demand for environmentally friendly pest control in agriculture has increased. To meet this demand, Integrated Pest Management (IPM) has been developed as an approach that integrates *e.g.* breeding for host plant resistance, biological and cultural control; including a chemical control component only as a last resort (van Lenteren, 1993; Dent, 1995; Koul *et al.* 2004). Recently, it was suggested that insect-resistant transgenic crops may become vital components of IPM (Romeis *et al.* 2008; Kos *et al.* 2009). At present, all commercially available insect-resistant transgenic crops employ direct resistance, and the majority of these transgenic crops express genes coding for proteins naturally occurring in *Bacillus thuringiensis* (*Bt*) that are lethal for target herbivores (Aronson & Shai 2001; Chen *et al.* 2008). In contrast, indirect resistance, which influences the effectiveness of natural enemies of herbivores (Karban & Baldwin, 1997; Dicke & Baldwin, 2010), has been largely neglected as a trait amenable to transgenesis.

Herbivore damage induces the emission of volatile organic compounds (VOCs) that have been shown to attract natural enemies of herbivores (Vet & Dicke, 1992; Heil, 2008; Dicke & Baldwin, 2010; Mumm & Dicke, 2010). Transgenic plants that enhance the effectiveness of natural enemies can be developed by using genetic engineering of VOC biosynthesis (Bouwmeester *et al.* 2003; Degenhardt *et al.* 2003; Poppy & Sutherland, 2004; Turlings & Ton, 2006). There are several benefits of incorporating transgenic crops that enhance biological control into IPM. For example, pest resistance to a biological control agent is not likely to evolve (Bale *et al.* 2008) and the VOCs that transgenic plants emit can repel herbivores (Dicke, 1986; Heil, 2004; Sanchez-Hernandez *et al.* 2006; Yang, 2008). However, there may also be ecological risks involved in using these transgenic plants, because the modified emission of VOCs might not only repel certain herbivores, it might also attract others (Carroll *et al.* 2006; Halitschke *et al.* 2008). Therefore, before transgenic crops with enhanced indirect resistance are implemented, information on the effects of the modification on (potential) pest herbivores is necessary.

Transgenic crops exhibiting enhanced attraction of natural enemies by the emission of novel VOCs or enhanced emission of native VOCs are not yet commercially available. Before commercialisation of these transgenic plants can be pursued, detailed fundamental studies on their effects on different groups of herbivores and their natural enemies are necessary. In the last decade, a few laboratory and field studies have been performed, mostly with the model plant *Arabidopsis thaliana*. These studies show that the novel or enhanced emission of VOCs by genetic engineering can increase the attraction of predatory mites

(Kappers *et al.* 2005), parasitoid wasps (Beale *et al.* 2006; Schnee *et al.* 2006) and entomopathogenic nematodes (Degenhardt *et al.* 2003; Degenhardt *et al.* 2009), and can repel aphids (Aharoni *et al.* 2003) and moths (Yang, 2008). However, comprehensive studies of the effects of transgenic plants with modified VOC emission on different functional groups of natural enemies simultaneously, as well as on their host or prey species, are needed.

The objectives of this study were to test the effects of transgenic *A. thaliana* plants with modified VOC emission on the behaviour of the specialist cabbage aphid *Brevicoryne brassicae*, the aphid parasitoid *Diaeretiella rapae* and the aphid predator *Episyrphus balteatus*. The model plant *A. thaliana* is ideally suited for testing effects of genetic engineering because it is easily transformed and has a short lifecycle (Aharoni *et al.* 2003). We constitutively expressed the linalool/nerolidol synthase gene *FaNES1* from strawberry in *A. thaliana* and targeted the enzyme to the mitochondria, which has been shown to result in the emission of two common herbivore-induced plant volatiles: the sesquiterpene (*E*)-nerolidol and its derivative, the homoterpene (*E*)-DMNT (4,8-dimethylnona-l,3,7-triene) (Kappers *et al.* 2005).

Furthermore, because FaNES1 is a dual-function enzyme that catalyses the formation of both nerolidol from FDP (farnesyl diphosphate) and linalool from GDP (geranyl diphosphate) with equal efficiency (Aharoni *et al.* 2005), the introduction of this enzyme in *A. thaliana* can also lead to emission of the monoterpene linalool if GDP is present (Aharoni *et al.* 2003). The modified VOC emission by transgenic *FaNES1*-expressing plants resulted in the attraction of predatory mites to these transgenic plants (Kappers *et al.* 2005), but other natural enemies have so far not been studied. In the literature, the effects of the manipulation of terpenoid biosynthesis in *A. thaliana* on insects are studied almost exclusively in the ecotype Col-0 (Aharoni *et al.* 2003; Kappers *et al.* 2005; Yang, 2008). Our study employed a novel approach of using different *A. thaliana* ecotypes as background for the transformation and testing the effects of the transformation on the emission of transgene-related products and subsequently on the behavioural response of aphids, parasitoids and predators.

Materials and methods

Plant material and growth conditions

Three *Arabidopsis thaliana* (L.) Heynh. ecotypes were selected, based on their maximal divergence in metabolite profiles (Houshyani *et al.* 2011): Antwerpen (An)-1 (obtained from the European Arabidopsis Stock Centre (http://nasc.nott.ac.uk/; An-1 = N944), Columbia (Col)-0 (provided by Dr. P. Reymond, Lausanne, Switzerland), and Eringsboda (Eri; collected in Sweden by members of the Laboratory of Genetics, Wageningen University; Eri-1 = CS22548).

Arabidopsis thaliana seeds were surface-sterilized overnight by a vapour-phase sterilization method. Hereto, the seeds were placed in a desiccator containing a mixture of 3 ml hydrochloric acid (37%, Merck KGaA, Darmstadt, Germany) and 100 ml bleach. Subsequentially, the seeds were inoculated on MS medium (purified agar 0.8% + 2.2 g L^{-1} 0.5 MS + vitamins; pH 6; 35 µg ml^{-1} kanamycin was added to the medium used for the T2 transgenic lines). After four days of stratification at 4 °C, plates were transferred to a growth chamber at 21 ± 2 °C, 50-70% relative humidity (RH) and a 8:16 light:dark (L:D) photoregime (with a light intensity of 200 µmol $m^{-2}s^{-1}$ photosynthetic photon flux density (PPFD)).

Two-week-old seedlings with two true leaves were transplanted to pots (5 cm diameter) containing autoclaved soil (80 °C for 4 h; Lentse potgrond, Lent, The Netherlands). Plants were watered three times a week and the soil was treated weekly with entomopathogenic nematodes (*Steinernema feltiae*; Koppert Biological Systems, Berkel en Rodenrijs, The Netherlands) to control infestation by larvae of sciarid flies. Plants used in the experiments were six to seven weeks old and remained in the vegetative state during the experiments.

Transgenic A. thaliana plants

Constructs for transformation and transgenic plants were generated as described in Houshyani *et al.* (2012). The strawberry linalool/nerolidol synthase gene *FaNES1* and the endogenous gene farnesyl diphosphate synthase (*FPS2*) were constitutively expressed using the 35S CaMV promotor and the enzymes were targeted to the mitochondria of *A. thaliana* (Houshyani *et al.* in prep.). *FPS2* was overexpressed to ensure that plants produced enough precursor for (*E*)-nerolidol biosynthesis. Methods S1 in the Supporting Information provides a detailed description of the generation of transgenic *FaNES1*-expressing plants. T1 generation

seedlings were selected on kanamycin plates (purified agar 0.8% + 2.2 g L^{-1} 0.5 MS + vitamins; pH 6; 50 µg ml^{-1} kanamycin) and the effectiveness of transformation was confirmed by gene-specific PCR. T2 generation seedlings of one positive line per ecotype (hereafter named An-FaNES1, Col-FaNES1 and Eri-FaNES1) were used in the experiments.

Insect rearing

Brevicoryne brassicae L. (Hemiptera: Aphididae) was reared on Brussels sprouts (*Brassica oleracea* L. var. *gemmifera* cv. Cyrus) in a greenhouse compartment.

Diaeretiella rapae McIntosh (Hymenoptera: Braconidae) was reared in gauze cages (30 x 40 x 60 cm) containing *B. brassicae*-infested *B. oleracea* plants in a climate chamber. Wasps were provided with water and honey. The *B. brassicae* and *D. rapae* cultures originated from individuals collected on *B. oleracea* in the vicinity of Wageningen (The Netherlands) in 2009.

Episyrphus balteatus de Geer (Diptera: Syrphidae) pupae were provided by Koppert Biological Systems and kept in gauze cages (67 x 50 x 67 cm) in a greenhouse compartment. Adults emerging from the pupae were provided with water, a *B. brassicae*-infested plant, organic sugar grains and bee-collected pollen provided by Koppert Biological Systems. All insect species were reared at 22 ± 2 °C, 60-70% RH and a 16:8 h L:D photoregime.

Chemical and morphological plant traits

Plant morphology

Of ten plants per *A. thaliana* line, we measured foliar biomass (fresh weight), plant diameter and the number of leaves and quantified trichome density by counting the number of trichomes in a 25 mm^2 area in the central part of the abaxial side of one mature leaf using a stereomicroscope (Leitz Dialux 20 EB, Wetzlar, Germany; magnification 40x).

Dynamic headspace collection of aphid-infested plants

While testing the preference of *D. rapae* in a Y-tube olfactometer (see below), we simultaneously collected the headspace of the plants. Volatiles were collected by sucking air out of each cuvette at a rate of 90 ml min^{-1} for 3 h through a stainless steel cartridge (Markes, Llantrisant, UK) filled with 200 mg Tenax TA (20/35 mesh; Grace-Alltech, Deerfield, USA). The fresh weight of the plant foliage in each cuvette was measured after volatile collection.

For each *A. thaliana* line, four to six replicates, each consisting of the headspace of four plants, were collected.

The headspace samples were analysed using a Thermo Trace GC Ultra (Thermo Fisher Scientific, Waltham, USA) coupled to a Thermo Trace DSQ (Thermo Fisher Scientific, Waltham, USA) quadrupole mass spectrometer (MS) (see Methods S2 for a detailed description of the analytical method and the identification of the volatile compounds). The peak areas of all compounds were expressed per unit plant fresh weight. In the GC-MS and statistical analysis, we particularly focussed on the emission of the transgene-related products (*E*)-nerolidol, (*E*)-DMNT, and linalool.

Aphid performance

We tested whether aphid performance differed between the transgenic and wild-type line of each ecotype, because this might (partly) explain the behaviour of the aphid's natural enemies. Aphid performance was assessed in a climate chamber at 21 ± 2 °C, 50-70% RH and a 8:16 L:D photoregime. Individual plants with insects were confined to cylindrical plastic containers (height 13 cm; diameter 11 cm) with a gauze lid. The light intensity at plant level was 200 µmol $m^{-2}s^{-1}$ PPFD. Plants were watered once a week. Several six-week-old plants of each ecotype/line were inoculated with 10 adult aphids. After 24 h, the adult aphids were removed and the produced offspring was allowed to develop for three days until they reached the second instar (L2). Three L2 nymphs were transferred to each of 25 plants of the same line as on which these nymphs had been feeding before. Daily, survival of the nymphs was recorded per plant until the first aphid on that plant reached the adult stage. The fastest developing adult was kept on the plant, while the other adults were removed. Any alate (winged) adults (ca. 5% of all adults) were excluded from the experiment. The development time until first reproduction (=T_d) of the remaining adult was recorded and the adult fresh weight was measured on a microbalance (Sartorius CP2P, Göttingen, Germany). The adult was allowed to feed on the plant and produce offspring, and after a certain number of days (equivalent to T_d), the number of offspring (=N) produced by the adult was counted. The estimated intrinsic rate of increase (r_m) was calculated for each aphid using the formula: $r_m = 0.738 \times (\ln N)/T_d$ (Karley *et al.* 2002).

Aphid preference

The preference of *B. brassicae* for the transgenic or the wild-type line of each ecotype was tested in two-choice bioassays in a greenhouse compartment (22 ± 2 °C, 60-70%

RH and a 16:8 h L:D photoregime). One transgenic and one wild-type plant were connected by a paper bridge (2 x 3 cm). Ten *B. brassicae* adults from the stock rearing were released in the centre of the bridge and were allowed to walk towards and feed on the plants. After 24 h the number of aphids on both plants was counted. For each ecotype at least 22 replicates were performed (An-1: 38, Col-0: 22, Eri: 30).

Parasitoid and predator preference

The preference of parasitoids and predators for volatiles from the transgenic or the wild-type line of each ecotype was tested in two-choice bioassays. Six-to-seven week-old *A. thaliana* plants were infested with 100 *B. brassicae* nymphs of mixed instars three days prior to the bioassays.

Parasitoid preference for aphid-induced plant volatiles

Parasitoid behaviour was assessed in a Y-tube olfactometer in a climatized room at 22 ± 2 °C as described by Bukovinszky *et al.* (2005). Compressed air was filtered over charcoal and split into two air streams each at a flow of 2 L min^{-1}. Each air stream was led through a 5 L glass cuvette that contained four aphid-infested plants of either the transgenic or the wild-type line of an ecotype. Each air stream was then led into one of the two arms of the Y-tube. The olfactometer was illuminated from above using artificial light at an intensity of 60 µmol $m^{-2}s^{-1}$ PPFD.

Naïve, mated two-day-old *D. rapae* females were allowed to oviposit for one h in aphids feeding on either the transgenic or the wild-type line of an ecotype (equally divided among the tested wasps) to increase their motivation to search for a host. Immediately after the training period, experienced wasps were individually released at the base of the Y-tube using a glass vial and their preference for one of both odour sources was recorded. A choice was recorded when a wasp crossed a finish line drawn one cm before the end of each arm, and did not return to the junction within 15 s. Wasps that did not move into one of the arms within 10 min or did not make a final choice within 15 min were considered as non-responsive and were omitted from the statistical analysis. Seven to nine new sets of transgenic and wild-type plants were used for each ecotype (An-1: 9; Col-0: 7; Eri: 7). For every new set of plants, 20 wasps were tested. After every ten wasps, the position of the odour sources was exchanged to compensate for any asymmetry in the set-up.

Predator oviposition preference

Mated female hoverflies from the stock rearing were used in the behavioural assays when they were two to three weeks old. Females were transferred to a plastic cage (30 x 30 x 30 cm) containing one transgenic plant and wild-type plant of the same ecotype, and 10% sugar solution in the same greenhouse compartment as described under *Aphid preference*. Females were allowed to oviposit on the plants. After 24 h, the number of eggs deposited on each plant was counted. Females that did not lay any eggs (about 30% of the females) were eliminated from the analysis. For each ecotype, at least 23 replicates with ovipositing females were obtained (An-1: 29, Col-0: 33, Eri: 23).

Statistical analysis

For most analyses, the transgenic line was compared with the wild-type line for each ecotype. Analyses were performed in SPSS for Windows (18th edition, Chicago, IL, USA), unless indicated otherwise. If variables were log-transformed to obtain normality and equal variance, this is indicated in the relevant table of the *Results* section. To test the effect of the line (transgenic or wild-type) on the measured variables, we used Student's t-tests. If assumptions on normality and equal variance were violated, Mann-Whitney U tests were used. Aphid survival was averaged per plant and differences in aphid survival between the transgenic and wild-type line of each ecotype were analysed by logistic regression in GenStat (13^{h} edition, VSN International, UK; dispersion estimated). T-probabilities were calculated to test pair-wise differences between means. Differences in the emission of transgene-related VOCs a) among the three transgenic lines and b) among the three wild-type lines were tested with ANOVA and post-hoc Tukey tests on the log-transformed data.

Aphid and parasitoid preference, measured by the number of aphids or wasps choosing for either the transgenic or the wild-type line of each ecotype, was analysed using a Chi-square test, with the null-hypothesis that the aphids or wasps did not have a preference for any of the two lines. Effects of parasitoid experience on the preference of the wasps was tested by logistic regression (dispersion estimated) in GenStat. Predator preference, measured by the number of predator eggs on either the transgenic or the wild-type line of each ecotype, was analysed with Wilcoxon's matched-pairs signed-rank tests.

To test if there were differences in the entire volatile profile between the transgenic line and the wild-type line of each ecotype, we used multivariate discriminant analysis Projection to Latent Structures-Discriminant Analysis (PLS-DA) in SIMCA-P (12th edition, Umetrics, Umeå, Sweden) (Eriksson *et al.* 2006). To pre-process data, volatile amounts were log-transformed, mean-centred and scaled to unit variance.

Results

Plant morphology

Transgenic An-FaNES1 plants had a higher biomass but a similar plant diameter compared to the An-1 wild-type plants, whereas Col-FaNES1 and Eri-FaNES1 plants had a lower biomass and smaller diameter than their respective wild-type plants (Student's t-test, n = 10, biomass: An-1: t = -3.04, P = 0.007; Col-0: t = 6.38, P < 0.001; Eri: t = 3.49, P = 0.003; diameter: An-1: t = 1.40, P = 0.179; Col-0: t = 9.24, P < 0.001; Eri: t = 5.69, P < 0.001; Table 1). Plants of all transgenic lines had a larger number of leaves than plants of the wild-type lines, although this was only significant for ecotypes An-1 and Col-0 (Mann-Whitney U-test, n = 10, An-1: U = 7.50, P < 0.001; Col-0: U = 21.00, P = 0.029; Eri: U = 44.50, P = 0.684; Table 1). For all ecotypes, leaves of transgenic *FaNES1*-expressing plants had a higher trichome density than leaves of wild-type plants (Student's t-test, n = 10, An-1: t = -2.74, P = 0.013; Col-0: t = -3.49, P = 0.003; Eri: t = -2.28, P = 0.035; Table 1).

Table 1 Mean (± SE) morphological plant characteristics (n = 10) of wild-type and transgenic FaNES1 lines of three Arabidopsis thaliana ecotypes

	***A. thaliana* line**					
Plant characteristic	**An-1**	**An-FaNES1**	**Col-0**	**Col-FaNES1**	**Eri**	**Eri-FaNES1**
Biomass (fresh weight; g)[a]	0.51 ± 0.02	0.63 ± 0.04 *****	0.49 ± 0.02	0.36 ± 0.01 *****	0.62 ± 0.05	0.43 ± 0.03 *****
Diameter (cm)[a]	8.0 ± 0.1	7.6 ± 0.2 **ns**	8.2 ± 0.1	6.1 ± 0.2 *****	8.7 ± 0.2	7.2 ± 0.2 *****
Number of leaves[b]	21.5 ± 0.3	25.8 ± 1.1 *****	22.1 ± 0.4	26.3 ± 1.6 *****	26.7 ± 0.9	28.8 ± 2.1 **ns**
Trichome density[a]	8.1 ± 0.6	10.7 ± 0.8 *****	13.7 ± 0.8	18.5 ± 1.2 *****	6.4 ± 0.3	7.8 ± 0.6 *****

[a] Performance parameter was averaged per plant before statistical analysis and analysed by logistic regression and post-hoc T-probability tests.
[b] Analysed by Mann-Whitney U-tests.
[c] Performance parameter was log-transformed in statistical analysis to obtain normality and analysed by Student's t-tests.
ns denotes no significant difference (P>0.05) in performance between aphids feeding on the wild-type and transgenic line for each ecotype

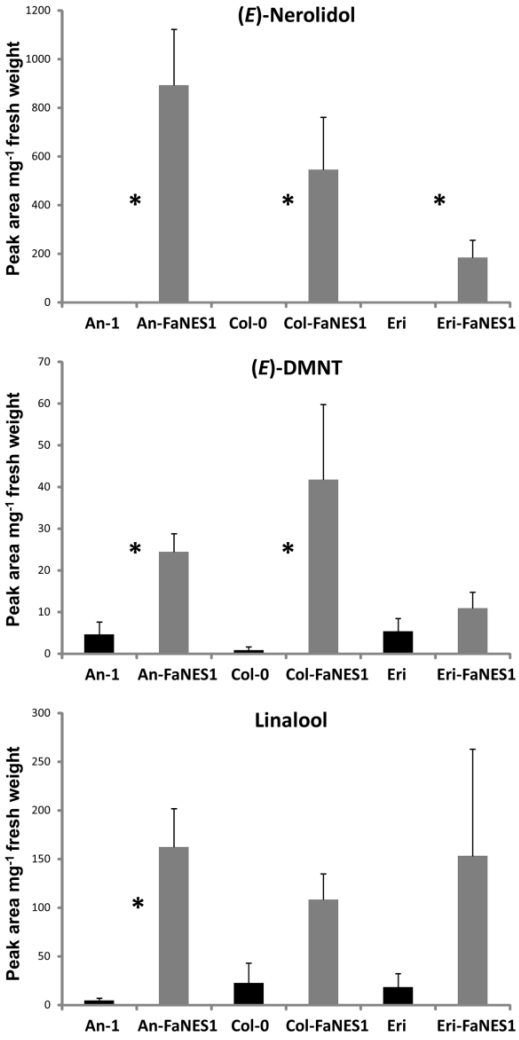


Figure 1 Emission of (*E*)-nerolidol, (*E*)-DMNT and linalool (peak area mg^{-1} fresh weight + SE) by aphid-infested wild-type or transgenic *Arabidopsis thaliana* lines of three ecotypes (An-1, Col-0 and Eri). Of each ecotype, a transgenic line (FaNES1) was created. *n* = 4-6 for each bar. An asterisk indicates a significant difference ($P<0.05$) in the emission of the compound between the wild-type and the transgenic *FaNES1*-expressing line for each ecotype, as analysed by Student's t-tests.

Dynamic headspace collection of aphid-infested plants

Transgenic *FaNES1*-expressing lines of the three ecotypes emitted (*E*)-nerolidol, whereas wild-type plants of the three ecotypes did not emit this compound (Fig. 1). An-FaNES1 plants emitted larger amounts of (*E*)-nerolidol than Eri-FaNES1 plants (ANOVA, $F_{2,11} = 6.01$; $P = 0.017$).

The transgenic FaNES1-expressing lines emitted larger amounts of (E)-DMNT than the corresponding wild-type plants, but the difference was only significant for ecotypes An-1 and Col-0 (Student's t-test, An-1: t = -3.38; P = 0.019; Col-0: t = -4.06; P = 0.007; Eri: t = -

1.65; P = 0.194; Fig. 1). While (E)-DMNT was detected in all transgenic samples, it was detected in only two of the six wild-type An-1 samples, one of the four wild-type Col-0 samples and two of the four wild-type Eri samples. Col-FaNES1 plants emitted larger amounts of (E)-DMNT than Eri-FaNES1 plants (ANOVA, $F_{2,11} = 4.26$; P = 0.043). The wild-type lines did not differ in emission of (E)-DMNT (ANOVA, $F_{2,11} = 0.295$; P = 0.750).

For each ecotype, the transgenic FaNES1-expressing line emitted larger amounts of linalool than the wild-type plants, but the difference was only significant for ecotype An-1, due to the high variation in emission of this compound among replicates (Student's t-test, An-1: t = -3.55; P = 0.015; Col-0 and Eri: P > 0.05; Fig. 1). In ca. 50% of the wild-type samples linalool was detected (three of the six An-1 samples, two of the four Col-0 samples and three of the four Eri samples), while this compound was detected in all the transgenic samples. The three transgenic lines did not differ in the emission of linalool, neither did the three wild-type lines (ANOVA, P > 0.05 for both analyses).

Apart from the reported differences in the emission of (E)-nerolidol, (E)-DMNT and linalool, there were no differences in the overall volatile profile between aphid-infested plants of the transgenic and wild-type line of each ecotype, as analysed by a multivariate discriminant analysis using all other volatile compounds (no significant PLS-DA components could be extracted; see Table S1 for an overview of the volatile compounds detected in the headspace of the aphid-infested plants).

Aphid performance and preference

There were no differences in any of the aphid-performance parameters between transgenic and wild-type plants for each of the ecotypes (P > 0.05 for every comparison; Table 2).

Aphids preferred wild-type Col-0 and Eri plants over plants of the corresponding transgenic line, but did not differentiate between An-1 and An-FaNES1 plants (Chi-square test, An-1: $\chi 2 < 0.01$, P = 1.000; Col-0: $\chi 2 = 10.47$, P = 0.001; Eri: $\chi 2 = 4.27$, P = 0.039; Fig. 2).

Parasitoid preference for aphid-induced plant volatiles

Parasitoids preferred volatiles from aphid-infested transgenic An-FaNES1 and Col-FaNES1 lines over volatiles from the corresponding aphid-infested wild-type lines, but did not differentiate between volatiles from aphid-infested Eri and aphid-infested Eri-FaNES1 plants

Table 2 Mean (± SE) performance characteristics of *Brevicoryne brassicae* on wild-type and transgenic FaNES1 lines of three *Arabidopsis thaliana* ecotypes

Performance parameter	*A. thaliana* line					
	An-1	**An-FaNES1**	**Col-0**	**Col-FaNES1**	**Eri**	**Eri-FaNES1**
Survival until adult stage (%)[a]	87	88 **ns**	93	89 **ns**	83	88 **ns**
Development time until first reproduction in days (*Td*)b	8.8 ± 0.1	8.8 ± 0.1 **ns**	9.7 ± 0.1	9.8 ± 0.1 **ns**	9.5 ± 0.1	9.6 ± 0.1 **ns**
Adult fresh weight in mg[c]	0.410 ± 0.017	0.417 ± 0.021 **ns**	0.329 ± 0.015	0.303 ± 0.011 **ns**	0.339 ± 0.011	0.320 ± 0.011 **ns**
Number of offspring (*N*) in time period equivalent to *Td*[c]	34.4 ± 3.4	34.0 ± 2.8 **ns**	24.8 ± 2.5	21.8 ± 2.2 **ns**	25.4 ± 2.0	23.0 ± 2.9 **ns**
Estimated intrinsic rate of population increase (*Rm*)[b]	0.288 ± 0.007	0.288 ± 0.008 **ns**	0.233 ± 0.009 **ns**	0.217 ± 0.010 **ns**	0.246 ± 0.006 **ns**	0.230 ± 0.008 **ns**

[a] Performance parameter was averaged per plant before statistical analysis and analysed by logistic regression and post-hoc T-probability tests.
[b] Analysed by Mann-Whitney U-tests.
[c] Performance parameter was log-transformed in statistical analysis to obtain normality and analysed by Student's t-tests.
ns denotes no significant difference (P>0.05) in performance between aphids feeding on the wild-type and transgenic line for each ecotype

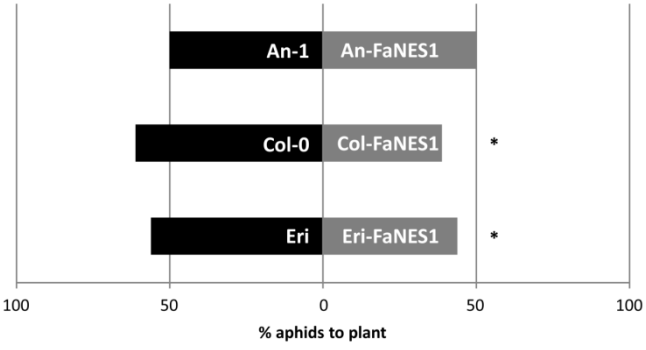


Figure 2 Preference of *Brevicoryne brassicae* adults in two-choice tests for the wild-type or transgenic line of three *Arabidopsis thaliana* ecotypes (An-1, Col-0 or Eri). Of each ecotype, a transgenic line with enhanced emission of volatiles (FaNES1) was created. Each bar represents the percentage of aphids that made a choice for the indicated plant line. The total number of tested aphids is indicated on the right. An asterisk indicates a significant preference (P<0.05) for one of the two lines in a combination, as analysed by Chi-square tests.

(Chi-square test, An-1: $\chi 2 = 4.12$, P = 0.042; Col-0: $\chi 2 = 5.28$, P = 0.022; Eri: $\chi 2 = 2.84$, P = 0.092; Fig. 3). There was no effect of previous oviposition experience on the preference of the wasps (logistic regression, P > 0.05 for every combination).

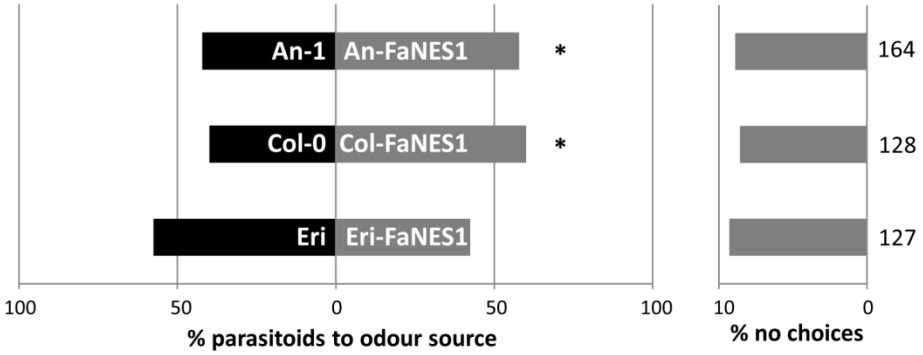


Figure 3 Responses of *Diaeretiella rapae* parasitoid females to volatile blends emitted by aphid-infested wild-type or transgenic *Arabidopsis thaliana* lines of three ecotypes (An-1, Col-0 and Eri) in a Y-tube olfactometer. Of each ecotype, a transgenic line with enhanced emission of volatiles (FaNES1) was created. Each bar represents the percentage of females that made a choice for the indicated odour sources. The percentage of parasitoids that did not make a choice ('% no choices') in each experiment and the total number of tested females is indicated on the right. An asterisk indicates a significant preference ($P<0.05$) for one of the two lines in a combination, as analysed by Chi-square tests. Photograph by Nina Fatouros; www.bugsinthepicture.com

Predator oviposition preference

Female *E. balteatus* did not differentiate between aphid-infested plants of the transgenic or wild-type line of any of the ecotypes (Wilcoxon's matched pair signed-rank test, P > 0.05 for every comparison; Fig. 4).

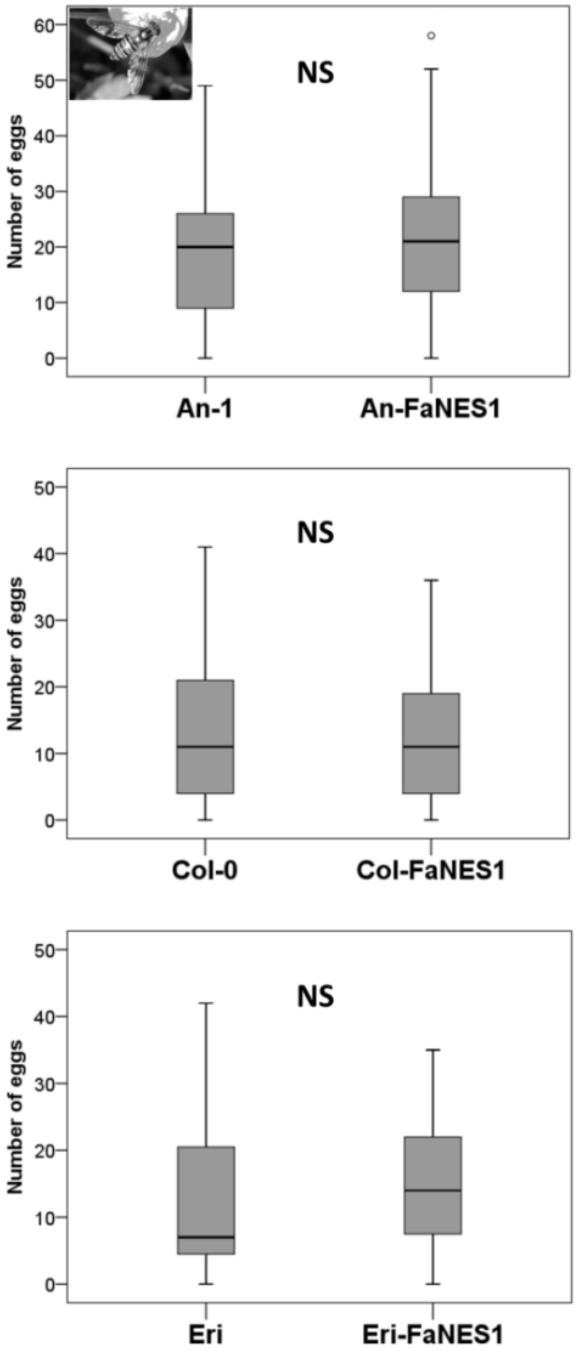


Figure 4. Oviposition preference (number of eggs) of the aphid predator *Episyrphus balteatus* in a two-choice assay with aphid-infested plants of wild-type and transgenic *Arabidopsis thaliana* lines of three ecotypes (An-1, Col-0 and Eri). Of each ecotype, a transgenic line with enhanced emission of volatiles (FaNES1) was created. The boxes span the first to third quartile range with the line across the box indicating the median. The whiskers represent the range. Open circles represent outliers. NS indicates a non-significant difference (P>0.05) between the number of eggs deposited on the wild-type and the transgenic line of each ecotype as analysed by the Wilcoxon's matched-pairs signed-rank test. Photograph by Han Endt; www.veluwe-insecten.nl/zweefvliegen/episyrphus/episyrphus.html

Discussion

Volatile emission by transgenic plants

Aphid-infested transgenic *FaNES1*-expressing plants emitted the sesquiterpene (*E*)-nerolidol, whereas this compound was not emitted by aphid-infested wild-type plants. Aphid-infested transgenic plants also emitted the homoterpene (*E*)-DMNT. (*E*)-DMNT is synthesised from (*E*)-nerolidol by endogenous *Arabidopsis* enzymes (Kappers *et al.* 2005). The emission of (*E*)-nerolidol and (*E*)-DMNT from transgenic *FaNES1*-expressing plants is in agreement with the study by Kappers *et al.* (2005) on ecotype Col-0, in which the same *FaNES1*-construct was used, but without the *FPS2*-construct.

Transgenic *FaNES1*-expressing plants also emitted larger amounts of the monoterpene linalool compared to wild-type plants, although this was only significant for ecotype An-1. FaNES1 is a dual-function enzyme that catalyses the formation of both linalool from GDP and nerolidol from FDP with equal efficiency (Aharoni *et al.* 2005). Our finding of higher linalool emission by transgenic plants suggests that there is a small pool of GDP present in the mitochondria of the *A. thaliana* ecotypes used. However, our finding could also mean that there was leaking of the linalool/nerolidol synthase from the mitochondria of the transformants in our study, as Kappers *et al.* (2005) did not report the emission of linalool in transgenic *FaNES1*-expressing plants.

The detection of (*E*)-DMNT and linalool in some of the wild-type samples is remarkable, as most previous studies report that *A. thaliana* plants do not produce these compounds in the vegetative state (Van Poecke *et al.* 2001; Aharoni *et al.* 2003; Kappers *et al.* 2005; Yang, 2008). However, these latter studies were done using an older version mass spectrometer (Van Poecke *et al.* 2001), or used solid-phase microextraction (SPME) GC-MS (Aharoni *et al.* 2003; Kappers *et al.* 2005; Yang, 2008). Perhaps their methods were less sensitive for (*E*)-DMNT and linalool than the method we used. In our study, (*E*)-DMNT and linalool were not detected in all wild-type samples, suggesting that the amounts were below the detection limit of the GC-MS, and the amounts in the samples in which these compounds were actually detected were small. A more recent study using the same GC-MS set-up as we did also recorded (*E*)-DMNT and linalool (and even (*E*)-nerolidol) in vegetative plants of the three *A. thaliana* ecotypes that we studied (Snoeren *et al.* 2010), suggesting that *A. thaliana* is indeed able to produce these compounds in the rosette stage.

The level of emission of (*E*)-nerolidol and (*E*)-DMNT by the transgenic plants was dependent on the genetic background of the transgenic line. It can be expected that plants with

a higher biomass emit larger amounts of volatiles. In agreement with this, An-FaNES1 plants emitted larger amounts of (*E*)-nerolidol than Eri-FaNES1 plants. However, the difference in biomass does not explain why Col-FaNES1 plants emitted larger amounts of (*E*)-DMNT than Eri-FaNES1 plants, as Col-FaNES1 plants had a lower biomass than Eri-FaNES1 plants. This suggests that the expression of the linalool/nerolidol synthase gene, the efficiency of the produced enzymes, or the availability of substrates differs among different ecotypes.

The biosynthesis of the novel terpenoids by transgenic *FaNES1*-expressing plants has been shown to impose costs on plant growth, and was speculated to be due to a reduction in the availability of substrates for other metabolites that play an important role in plant growth, or to direct toxic effects of the transgenic products to the plant (Aharoni *et al.* 2003; Kappers *et al.* 2005). In agreement with this, we observed growth retardation of transgenic plants for two of the three ecotypes (Col-0 and Eri). However, An-FaNES1 plants, which had the highest emission of (*E*)-nerolidol and linalool in their headspace, were actually larger than the wild-type An-1 plants. This suggests that the effects on plant growth are dependent on the genetic background of the transgenic line, or on the exact insertion position of the transgene in the genome.

Transgenic *FaNES1*-expressing plants of each ecotype exhibited a higher trichome density than the corresponding wild-type plants. This is probably due to the transgenic plants having a smaller diameter than the wild-type plants, suggesting that the leaves of the transgenic plants were smaller. It is known that trichomes are produced in the beginning of the development of a leaf and the growth of new trichomes is limited when the leaf starts to expand (Hülskamp & Schnittger, 1998).

Aphid performance and preference

Performance of the aphid *B. brassicae* did not differ between transgenic and wild-type plants for any of the three ecotypes. This suggests that the transgenic and wild-type plants did not differ in quality for aphid development, even though the transgenic plants had higher trichome densities, and were smaller (for two of the three ecotypes). Transgenic *FaNES1*-expressing plants that emitted large amounts of linalool (because the enzyme was targeted to the plastids, instead of the mitochondria) also did not affect the performance of caterpillars of the herbivore *Plutella xylostella* (Yang, 2008).

Brevicoryne brassicae aphids preferred wild-type Col-0 and Eri plants over plants of the corresponding transgenic line, suggesting that the volatiles (*E*)-nerolidol, (*E*)-DMNT and/or linalool emitted by Col-FaNES1 and Eri-FaNES1 plants acted as repellents to *B. brassicae*. This is in agreement with the study by Aharoni *et al.* (2003), in which transgenic

FaNES1-expressing plants with a high emission of linalool and linalool derivatives and a low emission of nerolidol repelled the aphid *Myzus persicae*. Linalool and/or nerolidol were also found to be repellent to spider mites, thrips, and moths (Dabrowski & Rodriguez, 1971; Doskotch *et al.* 1980; Kessler & Baldwin, 2001; Koschier *et al.* 2002; Yang, 2008). However, repellence of linalool and/or nerolidol to *B. brassicae* does not explain why aphids did not differentiate between An-1 and An-FaNES1 plants, the latter producing the largest amounts of (*E*)-nerolidol and linalool of the three transgenic lines. We cannot rule out effects of plant biomass or trichome density or of an interaction between these variables and VOC emission on aphid preference.

Parasitoid preference

The aphid parasitoid *D. rapae* preferred plants of An-FaNES1 and Col-FaNES1 over the wild-type plants. *Diaeretiella rapae* parasitizes many aphid species on at least 16 plant species (Pike *et al.* 1999), which could explain why this parasitoid is attracted towards plants emitting (*E*)-nerolidol, (*E*)-DMNT and linalool, all three common herbivore-induced volatiles. (*E*)-nerolidol, the most dominant transgene-related product in the headspace of the transgenic plants, is often reported as a component of the volatile blend of herbivore-induced plants. However, to our knowledge, attraction of natural enemies of herbivores specifically towards a source emitting this compound in pure form has only been reported for predatory mites (Kappers *et al.* 2005). The attractiveness of (*E*)-DMNT and linalool to several species of natural enemies has often been reported (Dicke *et al.* 1990a; Dicke *et al.* 1990b; Du *et al.* 1998; Hoballah *et al.* 2002; De Boer *et al.* 2004; Gouinguene *et al.* 2005; Kappers *et al.* 2005). Parasitoids did not differentiate between Eri and Eri-FaNES1 plants, probably because Eri-FaNES1 plants emitted only small amounts of (*E*)-nerolidol and (*E*)-DMNT.

Predator preference

In contrast to the parasitoid wasp, females of the aphid predator *E. balteatus* did not differentiate between aphid-infested plants of the transgenic and wild-type line of each ecotype. This suggests that (*E*)-nerolidol, (*E*)-DMNT and linalool do not affect the behaviour of *E. balteatus* at the emission rates tested. *Episyrphus balteatus* mainly uses aphid-derived chemicals to locate its prey (Almohamad *et al.* 2009). There are unfortunately only few published studies on the role of volatiles in hoverfly attraction. Verheggen *et al.* (2008) recorded that the sesquiterpenes α-humulene and β-caryophyllene did not evoke electroantennographic (EAG) responses in *E. balteatus*, whereas the aphid alarm pheromone (*E*)-β-farnesene did evoke responses. Monoterpenes elicited only a low response, compared to

the response to green leaf volatiles, and adding the monoterpene limonene to uninfested plants did not affect the oviposition behaviour of females (Verheggen *et al.* 2008). The low response to plant terpenoids might explain why *E. balteatus* did not differentiate between transgenic plants and wild-type plants in our study, which differed only in emission of three terpenoids.

In contrast to *D. rapae*, during the behavioural bioassays *E. balteatus* had access to the aphid-infested plants and could, therefore, also have used plant cues other than volatiles, as well as aphid cues, to select a plant for oviposition. We did not observe that hoverflies preferred the wild-type plants, which had lower trichome densities, over the transgenic plants, although it has been shown before that female hoverflies have difficulties in landing on plants with high trichome densities (Verheggen *et al.* 2009). Aphid body weight and population size did not differ between transgenic and wild-type plants, which may explain why hoverflies did not differentiate between the plants.

Potential of application of the transgenic approach in IPM

Our results indicate that genetically engineered plants with modified emission of VOCs can enhance the attraction of natural enemies of herbivores. Furthermore, a potential pest herbivore was repelled by the transgenic plant, while its performance was not affected, which increases the potential of these transgenic plants for pest control. Although our study suggests that genetically engineering plants to modify their emission of VOCs holds considerable promise for improving control of herbivores, it was performed with the model plant *A. thaliana*. Our results have to be validated with actual crop species before such a transgenic approach can be applied in IPM. The recent field study by Degenhardt *et al.* (2009) is an excellent example of how this technique could be successfully applied in a crop plant.

Before implementation of transgenic plants with modified emission of VOCs in IPM can be pursued, the technique for creating these transgenic plants should be optimised. Similar to previous studies (Aharoni *et al.* 2003; Kappers *et al.* 2005; Degenhardt *et al.* 2009), we used a constitutive promotor, which results in the constitutive production of the transgene-related products. Replacement of the constitutive promotor by an inducible promotor, for example one that is induced by herbivore feeding, will be essential. Natural enemies are capable of learning, and if the VOCs are emitted constitutively, the natural enemies likely learn to ignore the signal if it is not associated with a host or prey (Papaj *et al.* 1994; Kos *et al.* 2009). Furthermore, by using an inducible promotor, negative effects of the constitutive production of the transgenic products on plant growth may be prevented (Yang, 2008).

Conclusion

Our study shows that genetically transforming plants to modify the emission of terpenoids can repel a herbivore and enhance the attraction of one of its natural enemies, holding promise for improving control of herbivores in agriculture. However, our results have to be validated for a crop species before transgenic plants with enhanced attraction of natural enemies can provide interesting new components for IPM.

Supporting Information

Method S1: Generation of transgenic FaNES1-expressing plants.

Phusion enzyme (Finnzymes, Finland) was used for PCR when necessary. Standard cloning methods (restriction and ligation) were used for construction of all the plasmids and the vectors of final and intermediate stages and were checked by restriction analysis and sequencing.

The cDNA fragments encoding amino acid residues of the AtFPS2 (At4g17190) and FaNES1 proteins were cloned into two separate pGEM-T plasmid (Promega) (Cunillera *et al.* 1996; Kappers *et al.* 2005). The BamHI/NotI restricted fragment of resulted clones containing the FPS2 and FaNES1 coding sequences were gel purified and ligated in the pIV2B_2.5 and pIV2A_2.5 entry vectors, respectively, containing a CaMV 35S promoter, CoxIV mitochondrial targeting sequence and a RbcS1 terminator (www.pri.wur.nl/UK/products/ImpactVector/). Resulted entry clones were transferred to *E. coli* strain X1-Blue by heat shock for propagation. LR reaction (Invitrogen Life Technologies) was performed to recombine the sequences of interest in the entry clones between the right and left borders of the T-DNA in the pBINPLUS binary vector (Van Engelen *et al.* 1995). This resulted in the 2way plasmid contained the AtFPS2 and FaNES1 coding sequence each delimited by CaMV 35S promoter, CoxIV mitochondrial targeting sequence and a RbcS1 terminator. This binary vector was introduced to X1-Blue by heat shock and to *Agrobacterium tumefaciens* strain AglI by electroporation.

Plant transformations were performed using the *Arabidopsis* flower dipping method (Zhang *et al.* 2006). T1 seeds were harvested and T1 transgenic plants were selected on kanamycin plates (purified agar 0.8% + 2.2 g L^{-1} 0.5 MS + vitamins; pH 6; 50 µg ml^{-1} kanamycin) and confirmed by kanamycin resistance gene (NptII) specific primers. T2 generation seedlings of positive lines were used in the experiments.

References

Cunillera, N., Arro, M., Delourme, D., Karst, F., Boronat, A., & Ferrer, A. (1996) *Arabidopsis thaliana* contains two differentially expressed farnesyl-diphosphate synthase genes. *Journal of Biological Chemistry*, 271, 7774-7780.

Kappers, I.F., Aharoni, A., van Herpen, T., Luckerhoff, L.L.P., Dicke, M., & Bouwmeester, H.J. (2005) Genetic engineering of terpenoid metabolism attracts bodyguards to *Arabidopsis*. *Science*, 309, 2070-2072.

Van Engelen, F.A., Molthoff, J.W., Conner, A.J., Nap, J.P., Pereira, A., & Stiekema, W.J. (1995) pBINPLUS: An improved plant transformation vector based on pBIN19. *Transgenic Research*, 4, 288-290.

Zhang, X.R., Henriques, R., Lin, S.S., Niu, Q.W., & Chua, N.H. (2006) *Agrobacterium*-mediated transformation of *Arabidopsis thaliana* using the floral dip method. *Nature Protocols*, 1, 641-646.

Method S2: Description of the GC-MS method and the identification of volatile compounds.

GC-MS method

Prior to desorption of the volatiles, the samples were dry-purged under a flow of nitrogen at 20 ml min^{-1} for 20 min to remove moisture. The headspace samples were analysed using a Thermo Trace GC Ultra (Thermo Fisher Scientific, Waltham, USA) coupled to a Thermo Trace DSQ (Thermo Fisher Scientific, Waltham, USA) quadrupole mass spectrometer (MS). The collected volatiles were released from the Tenax TA thermally on an Ultra 50:50 thermal desorption unit (Markes, Llantrisant, UK) at 240ºC for 5 min under a helium flow of 30 ml min^{-1} while re-collecting the volatiles in an electronically-cooled sorbent trap (Unity, Markes, Llantrisant, UK) at 5ºC. Afterwards, the cold trap was rapidly heated at 40°C s^{-1} to 260°C and held for 7 min while the volatiles were transferred to a ZB-5MSi analytical column (30m x 0.25 mm I.D. x 1.0 µm film thickness; Phenomenex, Torrance, CA, USA), in a splitless mode for further separation. The analytical column was set at an initial temperature of

40°C, held for 3.5 min and raised at 10°C min^{-1} to 280°C and held for 2.5 min under a column flow of 1 ml min^{-1} in a constant flow mode. The DSQ MS was operated in scan mode with a mass range of 35 – 400 amu at 3.33 scans s^{-1} and ionization was performed in EI mode at 70 eV. The MS transfer line and ion source were set at 275 and 250°C, respectively.

Identification of compounds

Identification of compounds was based on comparison of mass spectra with those in the NIST 2005, Wiley and Wageningen Mass Spectral Database of Natural Products MS libraries. Experimentally calculated linear retention indices (LRI) were also used as additional criterion for confirming the identity of the compounds. Relative quantification (peak areas of individual compounds) was performed using a single (target) ion, in selected ion monitoring (SIM) mode.

Table S1: Volatile compounds detected in this study.

See chapter 3, supplementary Table 1.

References

Aharoni A, Giri AP, Deuerlein S, Griepink F, de Kogel WJ, Verstappen FWA, Verhoeven HA, Jongsma MA, Schwab W, Bouwmeester HJ. 2003. Terpenoid metabolism in wild-type and transgenic *Arabidopsis* plants. *Plant Cell* **15**: 2866-2884.

Aharoni A, Jongsma MA, Bouwmeester HJ. 2005. Volatile science? Metabolic engineering of terpenoids in plants. *Trends in Plant Science* **10**: 594-602.

Almohamad R, Verheggen FJ, Haubruge E. 2009. Searching and oviposition behavior of aphidophagous hoverflies (Diptera: Syrphidae): a review. *Biotechnologie Agronomie Societe et Environnement* **13**: 467-481.

Aronson AI, Shai Y. 2001. Why *Bacillus thuringiensis* insecticidal toxins are so effective: unique features of their mode of action. *FEMS Microbiology Letters* **195**: 1-8.

Bale JS, van Lenteren JC, Bigler F. 2008. Biological control and sustainable food production. *Philosophical Transactions of the Royal Society B-Biological Sciences* **363**: 761-776.

Beale MH, Birkett MA, Bruce TJA, Chamberlain K, Field LM, Huttly AK, Martin JL, Parker R, Phillips AL, Pickett JA, Prosser IM, Shewry PR, Smart LE, Wadhams LJ, Woodcock CM, Zhang YH. 2006. Aphid alarm pheromone produced by transgenic plants affects aphid and parasitoid behavior. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 10509-10513.

Bouwmeester HJ, Kappers IF, Verstappen FW, Aharoni A, Luckerhoff LLP, Lücker J, Jongsma MA, Dicke M 2003. Exploring multi-trophic plant-herbivore interactions for new crop protection methods. *Proceedings of the International Congress Crop Science and Technology, Vol. 2, 10-12 November 2003, Glasgow, British Crop Protection Council, Alton, UK*, 1123-1134.

Bukovinszky T, Gols R, Posthumus MA, Vet LEM, van Lenteren JC. 2005. Variation in plant volatiles and attraction of the parasitoid *Diadegma semiclausum* (Hellen). *Journal of Chemical Ecology* **31**: 461-480.

Carroll MJ, Schmelz EA, Meagher RL, Teal PEA. 2006. Attraction of *Spodoptera frugiperda* larvae to volatiles from herbivore-damaged maize seedlings. *Journal of Chemical Ecology* **32**: 1911-1924.

Chen M, Zhao JZ, Shelton AM, Cao J, Earle ED. 2008. Impact of single-gene and dual-gene *Bt* broccoli on the herbivore *Pieris rapae* (Lepidoptera: Pieridae) and its pupal endoparasitoid *Pteromalus puparum* (Hymenoptera: Pteromalidae). *Transgenic Research* **17**: 545-555.

Dabrowski ZT, Rodriguez JG. 1971. Studies on resistance of strawberries to mites. 3. Preference and nonpreference responses of *Tetranychus urticae* and *T. turkestani* to essential oils of foliage. *Journal of Economic Entomology* **64**: 387-&.

De Boer JG, Posthumus MA, Dicke M. 2004. Identification of volatiles that are used in discrimination between plants infested with prey or nonprey herbivores by a predatory mite. *Journal of Chemical Ecology* **30**: 2215-2230.

Degenhardt J, Gershenzon J, Baldwin IT, Kessler A. 2003. Attracting friends to feast on foes: engineering terpene emission to make crop plants more attractive to herbivore enemies. *Current Opinion in Biotechnology* **14**: 169-176.

Degenhardt J, Hiltpold I, Köllner TG, Frey M, Gierl A, Gershenzon J, Hibbard BE, Ellersieck MR, Turlings TCJ. 2009. Restoring a maize root signal that attracts insect-killing nematodes to control a major pest. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 13213-13218.

Dent DR. 1995. *Integrated pest management*. London, UK: Chapman & Hall.

Dicke M. 1986. Volatile spider-mite pheromone and host-plant kairomone involved in spaced-out gregariousness in the spider mite *Tetranychus urticae*. *Physiological Entomology* **11**: 251-262.

Dicke M, Baldwin IT. 2010. The evolutionary context for herbivore-induced plant volatiles: beyond the 'cry for help'. *Trends in Plant Science* **15**: 167-175.

Dicke M, Sabelis MW, Takabayashi J, Bruin J, Posthumus MA. 1990a. Plant strategies of manipulating predator-prey interactions through allelochemicals - Prospects for application in pest-control. *Journal of Chemical Ecology* **16**: 3091-3118.

Dicke M, van Beek TA, Posthumus MA, Bendom N, van Bokhoven H, de Groot AE. 1990b. Isolation and identification of volatile kairomone that affects acarine predator-prey interactions - Involvement of host plant in its production. *Journal of Chemical Ecology* **16**: 381-396.

Doskotch RW, Cheng HY, Odell TM, Girard L. 1980. Nerolidol: an antifeeding sesquiterpene alcohol for gypsy moth larvae from *Melaleuca leucadendron*. *Journal of Chemical Ecology* **6**: 845-851.

Du YJ, Poppy GM, Powell W, Pickett JA, Wadhams LJ, Woodcock CM. 1998. Identification of semiochemicals released during aphid feeding that attract parasitoid *Aphidius ervi*. *Journal of Chemical Ecology* **24**: 1355-1368.

Eriksson L, Johansson E, Kettaneh-Wold N, Trygg J, Wikström C, Wold S. 2006. *Multi- and megavariate data analysis. Part I: Basic principles and applications*. Umeå, Sweden: Umetrics Academy.

Gouinguene S, Pickett JA, Wadhams LJ, Birkett MA, Turlings TCJ. 2005. Antennal electrophysiological responses of three parasitic wasps to caterpillar-induced volatiles from maize (*Zea mays mays*), cotton (*Gossypium herbaceum*), and cowpea (*Vigna unguiculata*). *Journal of Chemical Ecology* **31**: 1023-1038.

Halitschke R, Stenberg JA, Kessler D, Kessler A, Baldwin IT. 2008. Shared signals - 'alarm calls' from plants increase apparency to herbivores and their enemies in nature. *Ecology Letters* **11**: 24-34.

Heil M. 2004. Direct defense or ecological costs: Responses of herbivorous beetles to volatiles released by wild lima bean (*Phaseolus lunatus*). *Journal of Chemical Ecology* **30**: 1289-1295.

Heil M. 2008. Indirect defence via tritrophic interactions. *New Phytologist* **178**: 41-61.

Hoballah MEF, Tamo C, Turlings TCJ. 2002. Differential attractiveness of induced odors emitted by eight maize varieties for the parasitoid *Cotesia marginiventris*: Is quality or quantity important? *Journal of Chemical Ecology* **28**: 951-968.

Houshyani B, Kabouw P, Muth D, de Vos RCH, Bino RJ, Bouwmeester HJ. 2011. Characterization of the natural variation in *Arabidopsis thaliana* metabolome by the analysis of metabolic distance. *Metabolomics* ***In press***.

Houshyani B, Assareh M, Busquets A, Ferrer A, Bouwmeester HJ, Kappers I. 2012. Three-step pathway engineering results in more incidence rate and higher emission of nerolidol and improved attraction of Diadegma semiclausum. Metabolic Engineering ***Submitted***

Hülskamp M, Schnittger A. 1998. Spatial regulation of trichome formation in *Arabidopsis thaliana*. *Seminars in Cell & Developmental Biology* **9**: 213-220.

Kappers IF, Aharoni A, van Herpen T, Luckerhoff LLP, Dicke M, Bouwmeester HJ. 2005. Genetic engineering of terpenoid metabolism attracts bodyguards to *Arabidopsis*. *Science* **309**: 2070-2072.

Karban R, Baldwin IT. 1997. *Induced Responses to Herbivory*. Chicago, USA: The University of Chicago Press.

Karley AJ, Douglas AE, Parker WE. 2002. Amino acid composition and nutritional quality of potato leaf phloem sap for aphids. *Journal of Experimental Biology* **205**: 3009-3018.

Kessler A, Baldwin IT. 2001. Defensive function of herbivore-induced plant volatile emissions in nature. *Science* **291**: 2141-2144.

Kos M, van Loon JJA, Dicke M, Vet LEM. 2009. Transgenic plants as vital components of integrated pest management. *Trends in Biotechnology* **27**: 621-627.

Koschier EH, Sedy KA, Novak J. 2002. Influence of plant volatiles on feeding damage caused by the onion thrips *Thrips tabaci*. *Crop Protection* **21**: 419-425.

Koul O, Dhaliwal GS, W. CG. 2004. *Integrated pest management: potential, constraints and challenges*. Wallingford, UK: CABI.

Mumm R, Dicke M. 2010. Variation in natural plant products and the attraction of bodyguards involved in indirect plant defense. *Canadian Journal of Zoology-Revue Canadienne De Zoologie* **88**: 628-667.

Papaj DR, Snellen H, Swaans K, Vet LEM. 1994. Unrewarding experiences and their effect on foraging in the parasitic wasp *Leptopilina heterotoma* (Hymenoptera, Eucoilidae). *Journal of Insect Behavior* **7**: 465-481.

Pike KS, Stary P, Miller T, Allison D, Graf G, Boydston L, Miller R, Gillespie R. 1999. Host range and habitats of the aphid parasitoid *Diaeretiella rapae* (Hymenoptera : Aphidiidae) in Washington state. *Environmental Entomology* **28**: 61-71.

Poppy GM, Sutherland JP. 2004. Can biological control benefit from genetically-modified crops? Tritrophic interactions on insect-resistant transgenic plants. *Physiological Entomology* **29**: 257-268.

Romeis J, Shelton AM, Kennedy GG. 2008. *Integration of insect-resistant genetically modified crops within IPM programs*. Dordrecht, The Netherlands: Springer.

Sanchez-Hernandez C, Lopez MG, Delano-Frier JP. 2006. Reduced levels of volatile emissions in jasmonate-deficient *spr2* tomato mutants favour oviposition by insect herbivores. *Plant Cell and Environment* **29**: 546-557.

Schnee C, Kollner TG, Held M, Turlings TCJ, Gershenzon J, Degenhardt J. 2006. The products of a single maize sesquiterpene synthase form a volatile defense signal that attracts natural enemies of maize herbivores. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 1129-1134.

Snoeren TAL, Kappers IF, Broekgaarden C, Mumm R, Dicke M, Bouwmeester HJ. 2010. Natural variation in herbivore-induced volatiles in *Arabidopsis thaliana*. *Journal of Experimental Botany* **61**: 3041-3056.

Turlings TCJ, Ton J. 2006. Exploiting scents of distress: the prospect of manipulating herbivore-induced plant odours to enhance the control of agricultural pests. *Current Opinion in Plant Biology* **9**: 421-427.

van Lenteren JC 1993. Integrated Pest Management: the inescapable trend. In: Zadoks JC ed. *Modern crop protection: developments and perspectives*. Wageningen, The Netherlands: Wageningen Press, 217-225.

Van Poecke RMP, Posthumus MA, Dicke M. 2001. Herbivore-induced volatile production by *Arabidopsis thaliana* leads to attraction of the parasitoid *Cotesia rubecula*: Chemical, behavioral, and gene-expression analysis. *Journal of Chemical Ecology* **27**: 1911-1928.

Verheggen FJ, Arnaud L, Bartram S, Gohy M, Haubruge E. 2008. Aphid and plant volatiles induce oviposition in an aphidophagous hoverfly. *Journal of Chemical Ecology* **34**: 301-307.

Verheggen FJ, Capella Q, Schwartzberg EG, Voigt D, Haubruge E. 2009. Tomato-aphid-hoverfly: a tritrophic interaction incompatible for pest management. *Arthropod-Plant Interactions* **3**: 141-149.

Vet LEM, Dicke M. 1992. Ecology of infochemical use by natural enemies in an tritrophic context. *Annual Review of Entomology* **37**: 141-172.

Yang LM. 2008. *Integration of host plant resistance and biological control: Using Arabidopsis-insect interactions as a model system.* PhD-thesis thesis Wageningen University.

Chapter 5

Overexpression of *HAG1/MYB28* in metabolically diverged *Arabidopsis* accessions: Effect on glucosinolates, gene expression profiles and performance of specialist and generalist herbivores

Benyamin Houshyani · Martine Kos · Belainesh Ararsa · Joop J. A. van Loon · Marcel Dicke · Jules Beekwilder · Harro J. Bouwmeester

In prep. for submission to Plant, Cell & Environment (2011)

Abstract

The R2R3-MYB transcription factor, HAG1/MYB28, regulates the expression of genes in the aliphatic glucosinolate biosynthetic pathway. The objectives of this study were to investigate the regulation of aliphatic glucosinolate biosynthesis in different genetic backgrounds of *Arabidopsis thaliana* by *HAG1/MYB28* overexpression and to investigate the effect of overexpression on a generalist and specialist herbivore. *HAG1/MYB28* was overexpressed in metabolically diverse accessions of *A. thaliana*, An-1, Col-0, Cvi and Eri. The expression of glucosinolate pathway genes as well as glucosinolate profiles of wild type and transgenic lines were analysed and the performance of a generalist lepidopteran, *Mamestra brassicae* and a Brassicaceae specialist lepidopteran, *Plutella xylostella*, was assessed.

The effect of *HAG1/MYB28* overexpression was to a large extent background-dependent both with regard to gene expression as well as metabolite levels. Overexpression generally increased aliphatic glucosinolate levels, but in a number of cases significantly reduced the glucosinolate level. Higher aliphatic glucosinolate concentrations significantly reduced the performance of the generalist but improved that of the specialist in some accessions. However, a higher glucosinolate concentration likely reduced the damage caused by both. We confirmed that chemical attributes of the side chain of aliphatic glucosinolates are more important for their effect on insects than the length of the side chain.

Introduction

Glucosinolates have been under investigation by plant breeders and plant and human biologists for a long time. These sulphur- and nitrogen-containing secondary metabolites mediate plant protection against various insects, pathogens and weeds and act as potent anticarcinogens (Hopkins *et al.* 2009; Vig *et al.* 2009; Redovniković *et al.* 2008). Brassicaceae including the model plant *Arabidopsis thaliana* contain a large variety of these compounds. Glucosinolates are stored in the vacuole, and upon cell damage enzymatic hydrolysis leads to the formation of volatile isothiocyanates, nitriles, epithiocyanates, thiocyanates and glucosides. Hydrolysis is catalysed by myrosinase, a glycosidase that is spatially separated from the glucosinolates in the plant , e.g. as a consequence of feeding by leaf-chewing herbivores (Kelly *et al.* 1998). Among other functions, glucosinolates and/or their hydrolysis products stimulate oviposition or feeding of specialist herbivores of Brassicaceae and are toxic, deterrent or repellent to generalist herbivores. The heterogeneous selection pressure originating from the mixed positive and negative impacts of these molecules on different herbivores in hostile and dynamic habitats (Daxenbichler *et al.* 1991 ;Jones *et al.* 1991) is possibly the reason for the large variation observed in glucosinolates in Brassicaceae (Newton *et al.* 2009; Poelman *et al.* 2009; Kliebenstein *et al.* 2001).

Aliphatic glucosinolates, the largest group within this class of phytochemicals in *A. thaliana*, are biosynthesised from deaminated methionine through three committed steps. These are: a) the side chain elongation in chloroplasts by serial condensation with acetyl-CoA, isomerization and decarboxylation involving enzymes such as methylthioalkylmalate synthases (MAM), an aconitase and an isopropylmalate dehydrogenase, b) core structure formation by cytochrome P450 mono-oxygenases such as cytochrome P450 79F1 (CYP79F1) and a C-S lyase and a glucosyltransferase and c) side chain modification by enzymes such as flavin-monooxygenase glucosinolate S-oxygenase 2 (FMO GS-OX2), alkenyl hydroxalkyl producing 2 (AOP2) and alkenyl hydroxalkyl producing 3 (AOP3) (Sønderby *et al.* 2010). Glucosinolates in different *A. thaliana* accessions vary in the length of the alkyl group (C3 to C8) of methylsulphinylalkyl glucosinolates and the ratio of methylsulphinylalkyl glucosinolates to the alkenyl glucosinolates (their derivatives) (Fig. 1), which is thought to be due to the polymorphism in the pathway genes and differential expression of the *AOP2* and *AOP3* genes (Kliebenstein *et al.* 2001; Neal *et al.* 2010; Chan *et al.* 2010).

The R2R3-MYB transcription factor, HAG1/MYB28, regulates the expression level of many genes in the aliphatic glucosinolate biosynthetic pathway in a complex interplay with

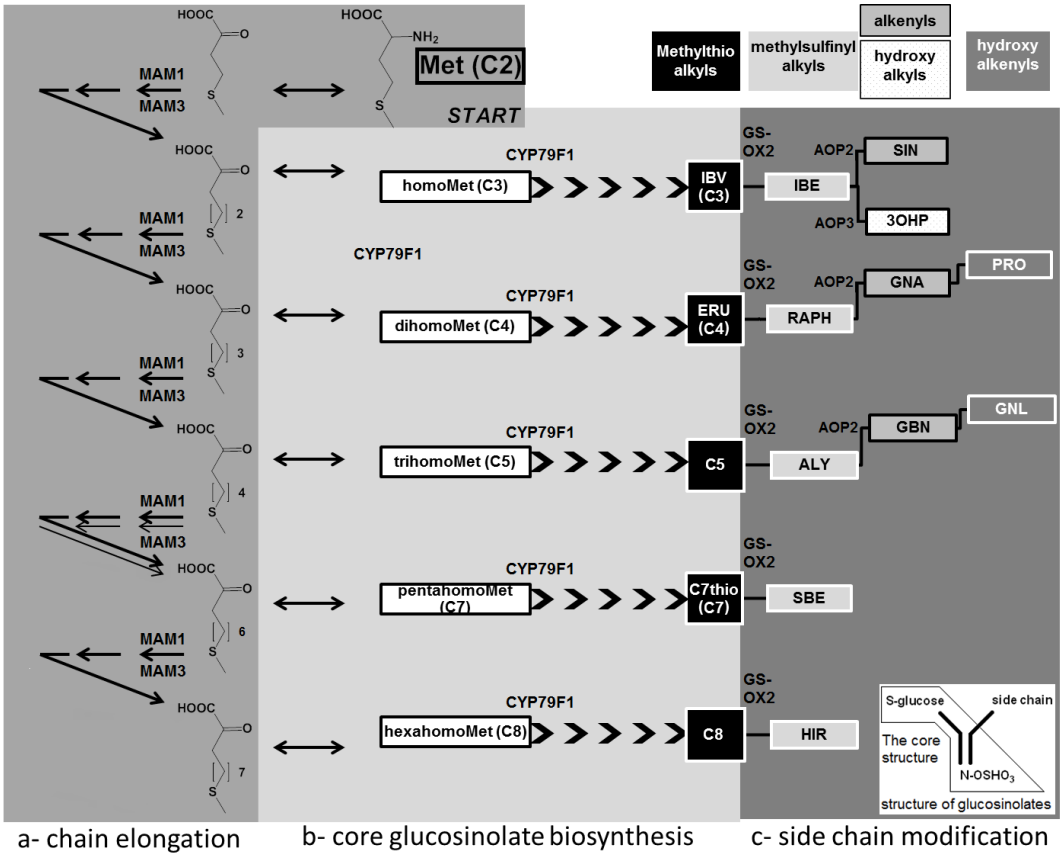


Figure 1 Schematic view of the three major steps in the aliphatic glucosinolate biosynthetic pathway (a, b and c) starting by deamination of methionine, Met (C2). Biosynthetic enzymes (MAM1, MAM3, CYP79F1, GS-OX2, AOP2 and AOP3) and aliphatic glucosinolates (see Table 2 for explanation of abbreviations) are indicated. The chain length of the aliphatic glucosinolates (C3-C8) is shown besides the primary methylthioalkyl substrate (black boxes, the final products of step b) and the structural characteristic is described above the side chain modification box (c) above the corresponding glucosinolate. The last step of the chain elongation is shown by double arrows indicating two consecutive elongation steps.

other R2R3-MYB members (Sønderby *et al.* 2010). The function of *MYB28* has been demonstrated using knockout and overexpression lines of *A. thaliana* accession Col-0. The impact of aliphatic glucosinolates on the performance of generalist herbivores was verified using these lines (Beekwilder *et al.* 2008; Gigolashvili *et al.* 2007; Hirai *et al.* 2007; Sønderby *et al.* 2010). Down-regulation of *MYB28* expression reduced the expression of aliphatic glucosinolate biosynthetic genes and consequently the levels of long-chain aliphatic glucosinolates were significantly reduced (Hirai *et al.* 2007). Larvae of the generalist

lepidopteran herbivore *Mamestra brassicae* feeding on knock-out mutants gained more weight than those feeding on wild type plants (Beekwilder ***et al.*** 2008). Over-expression of *MYB28* in *A. thaliana* resulted in higher amounts of aliphatic glucosinolates and a lower weight of larvae of the generalist *Spodoptera exigua* (Gigolashvili ***et al.*** 2007). The effect of these changes in glucosinolate profiles on the performance and response of herbivores specialized on Brassicaceae have not been tested. Specialist herbivores, such as *Plutella xylostella*, have adapted to glucosinolates by using detoxification (Ratzka *et al.* 2002; Wittstock *et al.* 2004), sequestration (Müller 2009) or by exploiting them and/or their hydrolysis products as oviposition or feeding stimulants (Barth *et al.* 2006; Miles *et al.* 2005; Zheng *et al.* 2011).

Studies on the function of *MYB28* have so far been carried out in *A. thaliana* accession Col-0. However, there is significant variation in glucosinolate profiles among *A. thaliana* accessions. Kliebenstein *et al.* (2001a) analysed 39 *A. thaliana* accessions and identified 34 different glucosinolates forming 14 qualitatively different glucosinolate profiles. Analysing 9 accessions in an untargeted metabolomics approach, we also found that glucosinolates are one of the important classifiers of the metabolome of the genotypes (Houshyani *et al.* 2011). To evaluate the effect of the genetic background on the resulting phenotype after *MYB28* introduction, in the present study we have overexpressed *MYB28* in *A. thaliana* accessions An-1, Col-0, Cvi and Eri that have quite different glucosinolate profiles (Houshyani *et al.* 2011). The effect of *MYB28* overexpression in some of these different accessions is studied both at the transcript level of a number of glucosinolate biosynthetic pathway genes and at the metabolite level. The possible impact of naturally different and genetically altered glucosinolate profiles in the 4 accessionss on the performance of the generalist lepidopteran *M. brassicae* and the Brassicaceae specialist lepidopteran *P. xylostella* is studied.

Materials and Methods

Plant material

Four *A. thaliana* (L.) Heynh. accessions, An-1 (Antwerpen-1), Col-0 (Columbia 0), Cvi (Cape Verdi Island) and Eri (Eringsboda) were used that represent most of the metabolome variation in a set of nine previously analysed accessions by an untargeted metabolomics approach (Houshyani *et al.* 2011). They were obtained from the European Arabidopsis Stock Centre (http://nasc.nott.ac.uk/) (An-1=N944, Cvi=N8580), collected in Sweden by members of the Wageningen Genetics Laboratory (Eri-1=CS22548) or provided by P. Reymond (Lausanne, Switzerland) (Col-0). Seeds were surface sterilized by vapour phase sterilization and germinated on autoclaved 0.8% purified agar (OXOID) with 0.5 MS + vitamins (pH > 6) (Duchefa, The Netherlands) and 50 $\mu g.ml^{-1}$ kanamycin (the latter was only for selection plates) in 14 cm Petri dishes. Petri dishes were placed in a ventilated growth chamber at 21 ± 2 °C, and L16:D8 photoperiod with ± 120 $\mu mol.m^{-2}.s^{-1}$ photosynthetic photon flux (PPF). Seedlings with two true leaves were transferred from Petri dishes to plastic cups (5 cm diameter) filled with potting soil (Lentse potgrond BV, Lent, Netherlands) and grew in the same chamber. The soil was heated to 80°C overnight prior to transplanting. Plants were watered twice a week. Three weeks after transplanting, fully grown vegetative plants were used for expression profiling, glucosinolate analysis or insect bioassays.

Generation of transgenic HAG1/MYB28 over-expression plants

The binary plasmid *Pro_{35S}:MYB28* was kindly provided by Dr. Tamara Gigolashvili (group of Prof. Ulf-Ingo Flügge) (University of Cologne, Germany) (Gigolashvili *et al.* 2007).

Pro_{35S}:MYB28 was transformed into *Agrobacterium tumefaciens* strain (Agl0) by electroporation and, after digestion and PCR confirmation, the *Pro_{35S}:MYB28* clone was transferred into the *A. thaliana* (L.) Heynh accessions by flower dipping (Zhang *et al.* 2006). T1 seeds were harvested and transgenic lines were selected on medium with kanamycin (50 $\mu g\ ml^{-1}$) and confirmed by kanamycin resistance gene (NptII) specific primers (Table 1). T2 generation seeds of positive lines were harvested and T2 generation transgenic plants were used in the experiments.

Glucosinolate analysis

Glucosinolate content of wild type and transgenic plants was analysed using shoot material of five 4-5 weeks old plants as biological replicates. Plants were cut just above the soil surface, immediately frozen in liquid nitrogen and subsequently transferred to a freeze-drier for 3 days. Lyophilized samples were ground using a micro-dismemberator U (B. Braun Biotech International). Glucosinolates were extracted from 100 mg of ground sample as described before (Kabouw *et al.* 2010) and glucosinolate content was assessed using high-performance liquid chromatography (HPLC). Glucosinolates were detected by a photodiode array detector (integration wavelength 229 nm). A concentration gradient of sinigrin was used as external standard for quantification (Acros, New Jersey, USA). To calculate the concentration of individual glucosinolates, we used the correction factors at 229 nm from Buchner (1987) and the European Community (1990). Desulfo-glucosinolates were identified by comparison of HPLC retention times and ultraviolet spectra with standards provided by M. Reichelt (Max Planck Institute for Chemical Ecology, Jena, Germany) and a certified rapeseed standard (Community Bureau of Reference, Brussels, Belgium, code BCR-367 R).

Gene expression analysis

Three biological replicates (consisting of a pool of 4 leaves from three 4-5 weeks old plants) of each wild type accession and transformation event (hereafter, collectively referred to as genotypes) were used. Following flash-freezing in liquid nitrogen, replicates were finely ground (2 min, 20 s^{-1} frequency) using a micro-dismemberator U (B. Braun Biotech International). RNA was isolated using 1 ml Tripure reagent (Roche) according to the manufacturer's instructions. The isolated RNA was treated with DNaseI (Invitrogen) and purified again using RNeasy (Qiagen). The concentration of purified RNA was measured (Nanodrop), quality was checked (gel analysis) and 1 µg RNA was used for cDNA synthesis using the iscript cDNA synthesis kit (BioRad).

Gene-specific primers for quantitative RT-PCR were designed using Beacon Designer 7.0 (Premier Biosoft Int), checked for gene specificity by blasting against the *A. thaliana* genome and RefSeq RNA database and by performing melt curve analysis on a MyIQ Single-Color Real-Time PCR Detection System (BioRad) (Table 1). The presence of SNPs in the annealing sites of the primers in An-1 and Cvi was checked using the genome browser of Salk *Arabidopsis* 1001 genomes (http://signal.salk.edu/atg1001/3.0/gebrowser.php). SNP data of Eri were not available. Due to the high level of polymorphism in the *At1g62540* locus (*GS-OX2* gene) the best primer combination had one SNP in the annealing site in Cvi. The amplification efficiency of

quantitative RT-PCR primers was determined by performing quantitative RT-PCR on dilution series of a Col-0 template. Quantitative RT-PCR analysis was carried out in optical 96-well plates on the MyIQ Single-Color Real-Time PCR Detection System, using SYBR Green Supermix Reagent (BioRad). Data analysis and threshold cycle (Ct) values were calculated using the IQ5 Optical System software (version 2.0, BioRad). Subsequently, Ct values were normalized for differences in cDNA synthesis by subtracting the Ct value of the constitutively expressed gene *β-Tubulin* (Table 1) from the Ct value of the gene of interest (δCt). The δδCt value of the gene of interest relative to Col-0 was calculated based on Livak and Schmittgen (Livak *et al.* 2001), multiplied by 10 and then Log2 transformed.

Table 1 Primers for quantitative RT-PCR

Gene Name	Sequence (5' to 3')
At5g61420, MYB28	F: TTGATGGAACAGGGTATTG
At5g61420, MYB28	R: GTATTGCTTGTCGGAACTA
At5g23010, MAM1	F: GGAATTGGTGAAAGAAGTG
At5g23010, MAM1	R: TTGTGTAAACCCCATTGAT
At5g23020, MAM3	F: AGAATCTCTGATGGATGGT
At5g23020, MAM3	R: GGCTTATGTGGTTGAACA
AT1G16410, CYP79F1	F: GCAAGAATCAGACATACCA
AT1G16410, CYP79F1	R: GGGACATAATGAGCACTT
At1g62540, GS.-OX2	F: GGAAATAGACACGGCATAT
At1g62540, GS.-OX2	R: CAAGGAATGTGAAGCGATA
At4g03060, AOP2	F: TGTGCTCTTATGAATGGTAG
At4g03060, AOP2	R: GCTGTGTATCTTGTCCTC
At4g03050, AOP3	F: CTGGATATGATGGTGAGAAG
At4g03050, AOP3	R: AATCGCATACGGTAATACG
AT5G62690, β-Tubulin	F: CTCAAGAGGTTCTCAGCAGTA
AT5G62690, β-Tubulin	R: TCACCTTCTTCATCCGCAGTT

Insect performance and plant damage analysis

Performance analyses of the generalist *M. brassicae* L. (Lepidoptera: Noctuidae) and Brassicaceae specialist *P. xylostella* L. (Lepidoptera: Yponomeutidae) feeding on different genotypes were performed using 4-5 weeks old plants and second instar larvae of both species

reared on Brussels sprouts (*Brassica oleracea* L. var. *gemmifera* cv. Cyrus) in climate rooms at 22 ± 2 °C, 40-50% RH and a 16:8 h L:D photoregime (Laboratory of Entomology, Wageningen University). As experiments were performed in two chambers at 2-week intervals, Col-0 was included in all experiments to check for reproducibility. The experiments for both species of insects took place on trays containing *A. thaliana* growing in pots, allowing the larvae to move freely between plants. Escaped larvae were prevented to go to the next tray by placing the tray in a layer of water and by applying sticky traps. Each tray contained 50 plants of the same genotype on which 1 *M. brassicae* larva or 2 *P. xylostella* larvae (both 2nd instar) were released per plant. The *M. brassicae* larvae were collected 7 days and the *P. xylostella* larvae 6 days post-release. Collected larvae were freeze-dried and weighed using a microbalance (Sartorius CP2P, Göttingen, Germany). All performance analyses were done under the same conditions as the plant growing conditions.

Damaged leaf area was calculated by taking a photo of ten randomly selected plants in the performance experiments 4 days post-release of the larvae. Photos were analysed and the damaged leaf area was calculated for each plant by Image 1.40g (NIH, Bethesda, MD, USA). To correct for impact of possible unwanted effects on the parameters due to performing experiments in different chambers and dates, all measured parameters were expressed as deviation from Col-0 average as the control genotype.

Statistical analysis

The expression data, glucosinolate concentrations and leaf damage area data were analysed by ANOVA in conjunction with post-hoc tests (Tukey's HSD). The larval weights were log-transformed and analysed by ANOVA in conjunction with Tamhane's T2 test and all analyses were done using PASW Statistics 17 SPSS package. All the multivariate data analyses were performed by CANOCO 4.5 for Windows (ter Braak 1988). The dataset was symmetrically scaled by the standard deviation (SD) of samples and individual glucosinolates or insect parameters without log transformation. An indirect gradient analysis, principal component analysis (PCA) using only the aliphatic glucosinolate concentrations, was performed to visualize the difference in the glucosinolate profiles of the samples. A direct gradient analysis, redundancy analysis (RDA), was used to examine and visualize the power of individual aliphatic glucosinolates in explaining the variation in insect parameters. We used only those explanatory variables (aliphatic glucosinolates) for RDA plot generation of which correlation with predicted ordinates was outside the arbitrarily selected range of 0.5 and -0.5.

Results

Overexpression of Pro$_{35S}$::MYB28

An overexpression construct of *MYB28* was introduced into *A. thaliana* accessions An-1, Col-0, Cvi and Eri. T2 transgenic lines were established for each accession. Transformation efficiency differed between accessions as 2 (An-1), 40 (Col-0), 1 (Cvi-1) and 10 (Eri) T1 transgenic plants (successful transformation events) were obtained by transformation.

In the transgenic plants, *MYB28* expression was up-regulated up to 19-fold compared with the wild type counterpart. Figure 2a compares expression levels of *MYB28* in a selection of genotypes. Total glucosinolate levels of maximum 10 transformation events per accession with clear overexpression of *MYB28* were determined. Surprisingly, alongside with transformation events showing enhanced glucosinolate levels (Col-0-enhanced, Eri-enhanced), also events with glucosinolate levels lower than in the corresponding wild type with a significant difference with enhanced event were observed (Col-0-inhibited, Eri-inhibited), despite the fact that they all overexpressed *MYB28* (Figure 2b). Among the 40 transformation events with Col-0 background, 3 events exhibited a significantly, 1.75, 2- and 2.5-fold, higher concentration of aliphatic glucosinolates at T2 generation, 1 had a significant, 4-fold, lower level and the rest did not show any difference ($\alpha = 0.05$, n=5). Among the 10 Eri T2 lines, two events showed 2-fold and 3-fold enhanced levels of aliphatic glucosinolates (the latter significantly higher but abnormal seed production), two showed 1.5-fold and 2-fold inhibited level (the latter significantly less but abnormal seed production) and the rest showed no significant change ($\alpha = 0.05$, n=5). In An-1 background, 1 transformation event of the two successful events showed a significant 3-fold inhibition and the other line showed no difference compared with the wild type ($\alpha = 0.05$, n=5). The only Cvi successful transformation event showed a slight reduction in total aliphatic glucosinolates concentration (data not shown). This level was still higher than the most enhanced overexpression event (Col-0-enhanced) on average. Wild type Cvi contained the highest total aliphatic glucosinolates concentration of the four accessions investigated, and its level was even higher than that of most of the transgenic lines (Fig. 2b). There was a significant difference between wild type accessions in indole glucosinolate levels, but overexpression of *MYB28* did not change these levels except for Eri-enhanced. Eri had the highest concentration of indole glucosinolates and Cvi the lowest (Table 2).

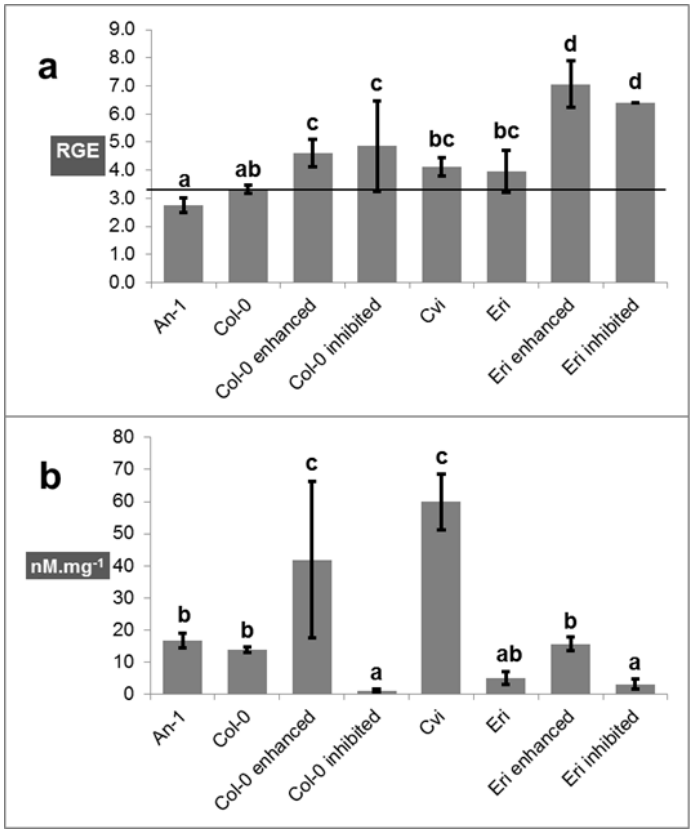


Figure 2 Relative gene expression (RGE) of *MYB28* (a) and total aliphatic glucosinolate concentration in nM per mg shoot dwt (b) in *Arabidopsis thaliana* accessions and selected transformation events. The baseline expression value (Col-0) is 3.33 (LOG_2, x10) and is indicated by the horizontal line in a. Enhanced and inhibited transformation events of Col-0 and Eri refer to the lines that displayed statistically significant different aliphatic glucosinolate concentration (ANOVA in conjunction with Tukey's test). Bars show the mean ± standard error and different letters indicate significant differences ($\alpha = 0.05$, n=5)

Expression of pathway genes

To study if the effect of *MYB28* overexpression on the expression level of aliphatic glucosinolates pathway is background-specific, quantitative RT-PCR was performed to assess the expression of a number of pathway genes. Primers were designed for genes with a major role in the aliphatic glucosinolate pathway (Kliebenstein *et al.* 2001). Primers for quantitative RT-PCR should be gene-specific and cover mRNA segments without polymorphism in different accessions. No polymorphisms were observed in Cvi, Col-0 and An-1 accessions, except that Cvi had a SNP in the annealing site of the reverse primer for the GS-OX2 gene. No SNP data was available for Eri, therefore poor detection of *MAM1* and *GS-OX2* genes expression in Eri could be due to mismatches in the primer annealing site (Fig. 3).

Overexpression of *MYB28* significantly up- and down-regulated *MAM1* expression in enhanced and inhibited lines of Col-0, respectively ($\alpha = 0.05$, n=3). *MYB28* overexpression

did not elevate *MAM1* expression in Eri, maybe due to poor detection of *MAM1* transcripts (Fig. 3). *MYB28* overexpression did not change *MAM3* expression in Col-0- and Eri-enhanced lines but down-regulated *MAM3* in Col-0-inhibited ($\alpha = 0.05$, n=3) (Fig. 3). Overexpression of *MYB28* also affected *CYP79F1* expression in Col-0 background by up- and down-regulating it in the enhanced and inhibited line, respectively, but had no effect in Eri ($\alpha = 0.05$, n=3) (Fig. 3). *MYB28* overexpression down-regulated *GS-OX2* expression in the Col-0-enhanced line only, while it had no effect in Eri ($\alpha = 0.05$, n=3) (Fig. 3).

Overexpression of *MYB28* increased and decreased *AOP2* expression in both Col-0 and Eri-enhanced and -inhibited transformation events, respectively. *AOP2* has a functional allele in Cvi and had the highest expression in Cvi and An-1, while its transcripts were also detected in Col-0 wild type ($\alpha = 0.05$, n=3). Overexpression of *MYB28* influenced the transcript level of the other deoxygenase gene, *AOP3* in Eri-enhanced and inhibited lines and it was detected at a lower level in the other genotypes ($\alpha = 0.05$, n=3) (Fig. 3).

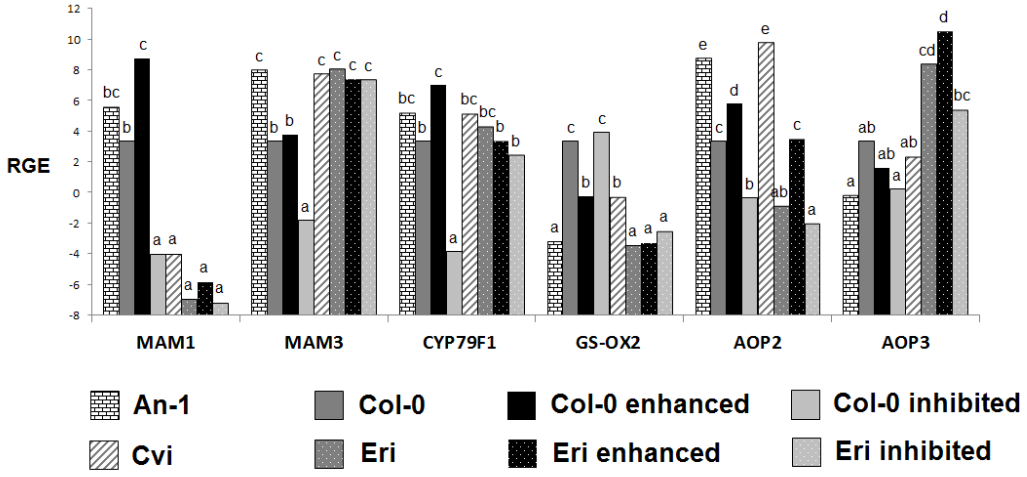


Figure 3 Relative gene expression (RGE) of *Arabidopsis thaliana* accessions and *MYB28* overexpressing transgenic lines for 6 genes involved in aliphatic glucosinolate biosynthesis. Expression of each gene is shown relative to Col-0. The baseline expression value (Col-0) for all genes is 3.33 (LOG_2, x10). Different letters indicate significant differences (ANOVA in conjunction with Tukey's test) between genotypes for the same transcript ($\alpha = 0.05$, n = 3) with "a" meaning a Ct value higher than 30, thus very low expression

Glucosinolate composition

A set of genotypes including wild type accessions and an enhanced and an inhibited aliphatic glucosinolate transgenic line in Col-0 and Eri background, were profiled for

glucosinolates that were previously detected in Col-0 and Cvi (Table 2 and Supplementary Table 1) (Kliebenstein *et al.* 2001). Among the 20 analysed glucosinolates no IBV (glucoiberverin, C3), EPRO (epiprogoitrin, C4) and HIR (glucohirsutin, C8) could be detected. Among the methylthioalkyl glucosinolates (Fig. 1), only ERU (glucoerucin, C4) and C7thio (C7) were detected with Col-0 and An-1 having the highest levels, respectively (Table 2). For the methylsulphinylalkyl glucosinolates that are derived from the methylthioalkyls including IBE (glucoiberin, C3), RAPH (glucoraphanin, C4), ALY (glucoalyssin, C5) and SBE (glucosiberin, C7) (Fig. 1), Eri, Col-0, Col-0 and An-1 were the richest, respectively (Table 2). For alkenyl glucosinolates, SIN (sinigrin, C3), GNA (gluconapin, C4) and GBN (glucobrassicanapin, C5), which are derived from the methylsulfinylalkyls (Fig. 1), Cvi and then An-1 had the highest concentrations for SIN, while GNA and GBN were detected only in An-1 (Table 2). Among the hydroxyalkyl glucosinolates that are also derived from methylsulfinylalkyls (Fig. 1), only 3OHP (C3) was detected in all analysed Eri wild type plants and overexpression lines.

Among the wild type accessions, Cvi had the highest concentration of C3 and C4 aliphatic glucosinolates with significantly higher alkenyl glucosinolates SIN (C3) and GNA (C4) (Fig. 1 and Table 2). Eri was the lowest C4 aliphatic glucosinolate producing accession and An1 was the richest in C5 and C7 aliphatic glucosinolates compared with the other accessions (Fig. 1 and Table 2).

PCA biplots using the aliphatic glucosinolate dataset and the first 3 principal components (explaining 90% of the variation) showed clustering of samples based on genetic background (Fig. 4). Cvi and An-1 clustered separately from Col-0 and Eri along the first PC. Col-0 and Eri were more similar than An-1 and Cvi in aliphatic glucosinolate profile. The most important glucosinolates for clustering of accessions were the methylsulphinylalkyl glucosinolates IBE (C3) and RAPH (C4) with significantly higher abundance in Col-0 followed by Eri ($\alpha = 0.05$, n=5), while their alkenylated derivatives SIN (C3) and GNA (C4) were significantly more abundant in Cvi followed by An-1 ($\alpha = 0.05$, n=5) (Fig. 1 and 4 and Table 2).

The biplots (Fig. 4) also show that the short-chain aliphatic glucosinolates IBE (C3), 3OHP (C3) RAPH (C4) and ALY (C5) are the endogenous glucosinolates mostly affected by *MYB28* overexpression. No significant change occurred in the concentration of the long-chain aliphatic glucosinolates (C7) (Table 2). In contrast to Col-0-enhanced, biosynthesis of IBE (C3) was decreased in Eri-enhanced (Table 2). There were no new glucosinolates synthesized in *MYB28* overexpression lines except for trace amounts of alkenylated glucosinolates GNA (C4) and GBN (C5) in Col-0-enhanced that were not detected in any of the wild type plants of this accession (Table 2).

Table 2 Concentration of the measured glucosinolates (nmol/ mg shoot dwt) in the wild types accessions and transgenic lines. Letters after each value indicate significant difference between accessions/lines for the corresponding glucosinolate (Tukey's HSD, $\alpha = 0.05$, $n = 5$).

	Aliphatic													Indolic					Aromatic
Accessions/Lines	C3			C4				C5			C7		Total Aliphatic					Total Indole	
	IBE	3OHP	SIN	ERU	RAPH	GNA	PRO	ALY	GBN	GNL	C7thio	SBE		GBC	4OH	4MeOH	NEO		NAS
An-1	0.01a	0.00a	3.61b	0.03a	0.03a	7.37b	3.46b	0.06a	0.34b	0.25b	0.65b	0.77b	16.58b	3.47bc	0.02a	0.33 ab	0.30abc	4.12bc	0.02a
Col-0	1.42a	0.00a	0.00a	2.57b	9.21b	0.00a	0.00a	0.28a	0.00a	0.00a	0.07a	0.17a	13.72b	2.13ab	0.03b	0.77c	0.31abc	3.24ab	0.00a
Col-0 enhanced	2.53b	0.00a	0.00a	0.16a	43.92c	0.21a	0.00a	2.26b	0.02a	0.00a	0.04a	0.74b	49.88c	1.19a	0.00a	0.65c	0.31abc	2.15a	0.05a
Col-0 inhibited	0.02a	0.00a	0.00a	0.21a	0.68a	0.02a	0.00a	0.04a	0.00a	0.00a	0.00a	0.00a	0.97a	2.37ab	0.00a	0.61bc	0.68c	3.66bc	0.00a
Cvi	0.09a	0.00a	20.46c	0.00a	0.15a	38.60c	0.00a	0.02a	0.00a	0.00a	0.14a	0.41ab	59.87d	2.09ab	0.00a	0.37ab	0.25ab	2.71a	0.00a
Eri	2.57b	1.94b	0.00a	0.15a	0.22a	0.00a	0.00a	0.00a	0.00a	0.00a	0.01a	0.05a	4.94ab	4.14c	0.00a	0.31a	0.37abc	4.82c	0.00a
Eri enhanced	1.96b	12.63c	0.00a	0.09a	0.70a	0.00a	0.00a	0.00a	0.00a	0.00a	0.04a	0.19a	15.61b	2.70ab	0.00a	0.25a	0.18a	3.13ab	0.00a
Eri inhibited	1.58ab	1.32b	0.00a	0.02a	0.11a	0.00a	0.00a	0.00a	0.00a	0.00a	0.00a	0.01a	3.04a	3.19b	0.00a	0.24a	0.17a	3.6bc	0.00a

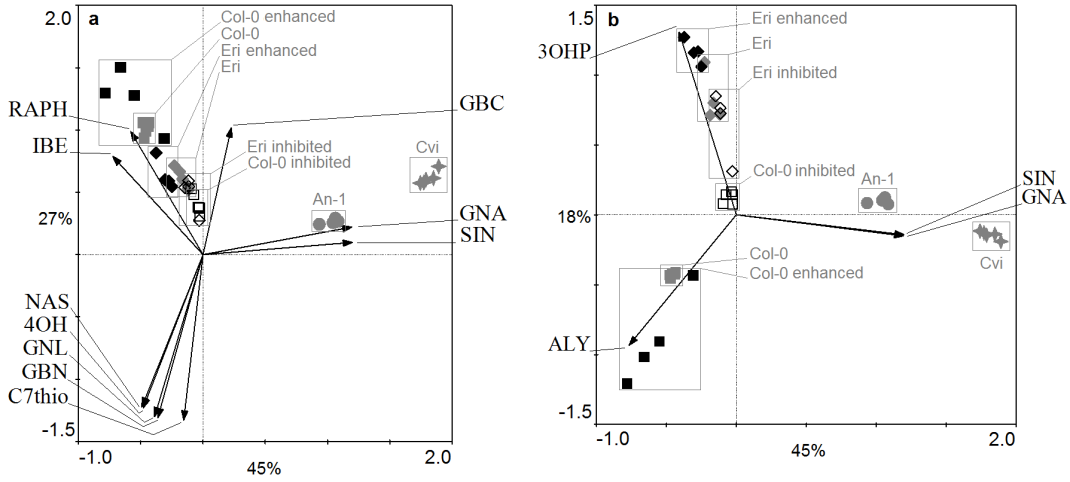


Figure 4 PCA biplots with the 1st and 2nd (a) and 1st and 3rd (b) principal components (PCs) showing clustering of *Arabidopsis thaliana* genotypes based on their glucosinolate profile. Percentages indicate the share of each PC in explaining the observed variation in aliphatic glucosinolates. Arrows indicate the glucosinolates that fit more than 55% to the ordination model. Boxes inside the plots delimit the variation of the corresponding genotype across the PCs. ● An-1, ■ Col-0, ■ Col-0 enhanced, □ Col-0 inhibited, ✦ Cvi, ◆ Eri, ◆ Eri enhanced, ◇ Eri inhibited

Effect of genetic background and Pro$_{35S}$:MYB28 overexpression on insect performance

Selected genotypes (with either higher or lower glucosinolate levels) were used to investigate the effect of glucosinolates on larval weight and damage by the generalist *M. brassicae* and the specialist *P. xylostella*. Feeding behaviour of *M. brassicae* and *P. xylostella* was different with *M. brassicae* starting feeding from the older leaves of both wild type and transgenic lines and never feeding on young leaves throughout the whole experiment, while *P. xylostella* started feeding from the young leaves and then continued with the older leaves (personal observation). *Mamestra brassicae* larvae feeding intensity (leaf damage area) varied slightly across wild type accessions (Fig. 5a). In Col-0, the leaf area damaged by *M. brassicae* is lower in the enhanced line with a higher concentration of aliphatic glucosinolates. The damage was even significantly lower than for Cvi ($\alpha = 0.05$, n=5) even though the latter had a significantly higher aliphatic glucosinolate concentration ($\alpha = 0.05$, n=5) (Fig. 5a and Fig. 2b). *Mamestra brassicae* larvae feeding on Eri-enhanced plants tended to escape making judgement on their damaged leaf dubious and larval weight comparison impossible.

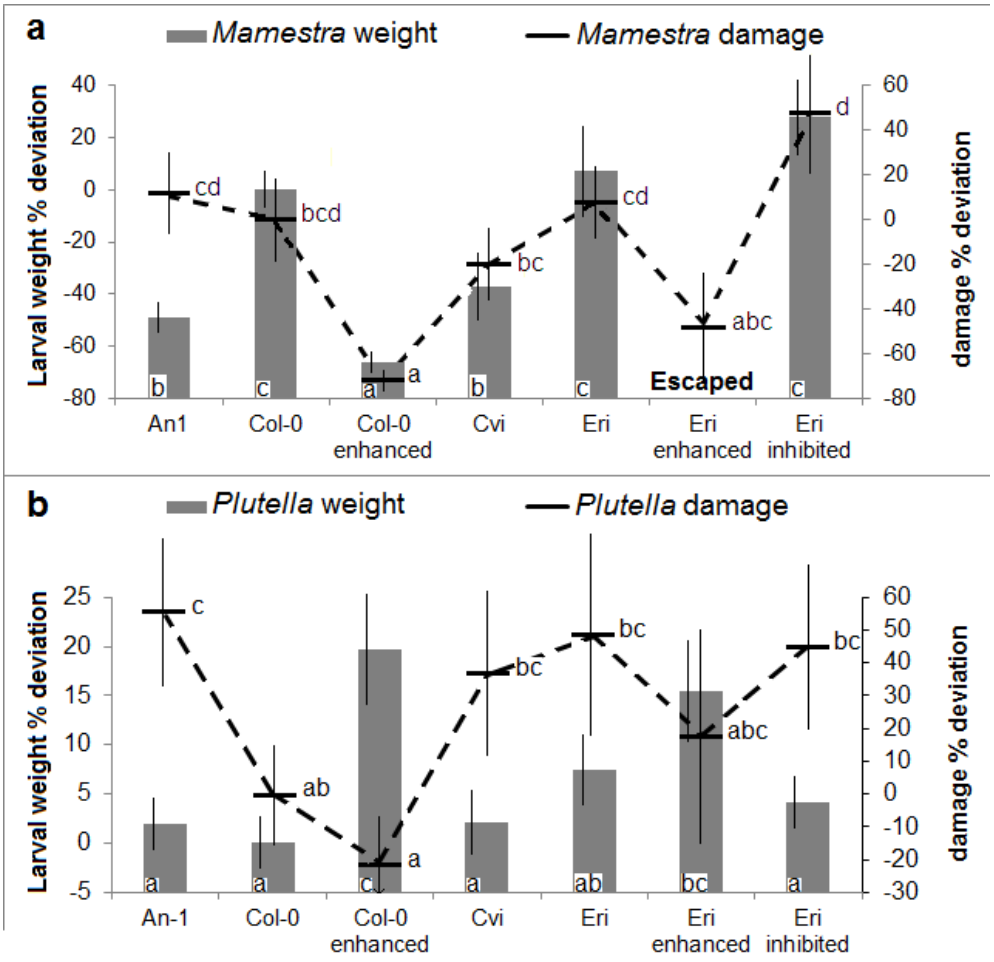


Figure 5 Larval weight (bars) and leaf damage (horizontal lines) on *Arabidopsis thaliana* accessions and *MYB28* overexpressing transgenic lines of *Mamestra brassicae* (a) and *Plutella xylostella* (b). Both parameters are expressed as percentage of deviation from the wild type Col-0 average (control). Different letters indicate a significant difference between genotypes for the larval weight (inside the bar) and leaf damage (beside the line) of the corresponding insect (ANOVA in conjunction with Tamhane's or Tukey's, respectively, for multiple comparison, $\alpha = 0.05$). Standard errors of larval weights are centered on the top of the corresponding bar and those of leaf damage on the corresponding line. Data for leaf damage were connected by dotted lines to indicate the similarity in the pattern of leaf damage changes from genotype to genotype between *M. brassicae* and *P. xylostella*

Mamestra brassicae larval weight also varied significantly across wild type accessions; larval weight on Col-0 and Eri was significantly higher than on An-1 and Cvi ($\alpha = 0.05$, n=5) (Fig. 5a). *Mamestra brassicae* larvae on Col-0-enhanced showed the lowest weight

while those feeding on Eri-inhibited were on average the heaviest, but did not differ significantly with wild type Col-0 and Eri.

Feeding intensity by *P. xylostella* followed the same pattern as *M. brassicae* among genotypes (Fig. 5a & b). *Plutella xylostella* larval weight on wild type accessions did not differ significantly. However, the weight of *P. xylostella* larvae was higher on Col-0-enhanced lines (α = 0.05, n=5) that had increased levels of aliphatic glucosinolates even though feeding intensity on these lines did not differ with the wild type background (Fig. 5b). *P. xylostella* larvae feeding on the Eri-inhibited line, damaged leaves and gained weight to the same level as on the wild type background, but their weight was significantly less than those feeding on Eri enhanced lines with increased levels of aliphatic glucosinolates (Fig. 5b).

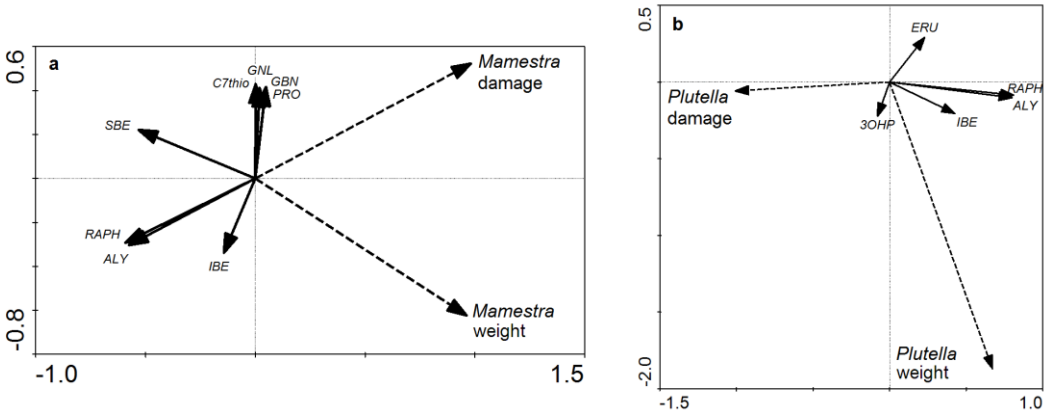


Figure 6 RDA biplot showing the correlation between *M. brassicae* (a) or *P. xylostella* (b) parameters (dashed arrows) and aliphatic glucosinolates (solid arrows). Arrows pointing in the same direction indicate a positive correlation, whereas an opposite direction indicates a negative correlation. Short projection length of arrows on ordination axes indicate a low contribution to the explained variance by the ordination axes.

Correlating glucosinolate concentration with insect parameters

The data on wild type and transgenic glucosinolate profiles in combination with the insect behaviour data allow for canonical correlation analysis to find the glucosinolates that explain most of the observed variation in insect parameters. The resulting RDA plot and the direction of arrows (that represent insect parameters) show that among the aliphatic glucosinolates in *A. thaliana*, the 4 methylsulfinyl glucosinolates RAPH (C4), ALY (C5) and to a lesser extent IBE (C3) and SBE (C7) have the highest negative relationship with the

damage by *M. brassicae* (Fig. 6a). SBE had also the strongest negative correlation with *M. brassicae* growth (Fig. 6a). RAPH, ALY and IBE also had the strongest negative correlation with the damage by *P. xylostella* (Fig. 6b). However, only one of the analysed glucosinolates (ERU) correlated with *P. xylostella* growth (negatively). Otherwise, *P. xylostella* performance correlated positively with the glucosinolates (arrows pointing in the same direction) (Fig. 6b). Our data also showed a positive relationship between feeding by *M. brassicae* and its performance while this was not the case for *P. xylostella* (Fig. 6a and b).

Discussion

We introduced the *Pro$_{35S}$::MYB28* construct into the genome of four *A. thaliana* accessions, Col-0, Cvi, Eri and An-1. These accessions differ strongly at the whole metabolome level and among other metabolites, glucosinolates play a substantial role in the differentiation between them (Houshyani *et al.* 2011). This allows for comparison of the effect of *MYB28* overexpression in different genetic/metabolic backgrounds. *MYB28* expression varied across the wild type accessions, but there was no association between *MYB28* transcript level and the total concentration of aliphatic glucosinolates. Accession Eri with the lowest aliphatic glucosinolate concentration is one of the highest *MYB28* expressers.

Accessions Cvi and Col-0 have already been characterized with respect to the major genetic loci controlling the biosynthesis of aliphatic glucosinolates (Gigolashvili *et al.* 2007) and could thus be used as references in our approach for genetic characterization of An-1 and Eri. The amount of the C4 aliphatic glucosinolates in the profile of An-1(high) and Eri (low) (Table 2) suggests that An-1 has the same allele as Col-0 and Cvi at the *GS-ELONG* locus, which is responsible for elongation of the carbon chain from C3 to C4 (Fig. 1), whereas Eri seems genetically diverged from the other three genotypes at this locus. The accessions also show variation in alkenyl aliphatic glucosinolates SIN (C3), GNA (C4), PRO (progoitrin, C4), GBN (C5) and GNL (napoleiferin, C5) (Table 2) that are products of AOP2 (Fig. 1) and are present in An-1 and Cvi but not in Col-0 and Eri. This suggests genetic variation in the *AOP2* locus which is known to harbour a functional *AOP2* gene in Cvi. 3OHP, the product of AOP3 that is responsible for hydroxylation of methylsulfinylalkyl glucosinolates (Fig. 1), is present in Eri genotypes only (Table 2). It indicates that the *AOP3* locus is functional in Eri only, in agreement with the hypothesis that *A. thaliana* displays differential AOP leaf expression whereby a particular accession expresses either *AOP2* or *AOP3* or neither, but not both (Kliebenstein *et al.* 2001). The aliphatic glucosinolate profile of wild type Col-0 suggests that this accession has neither the *AOP2* nor the *AOP3* functional allele as no trace of alkenyl or hydroxylated aliphatic glucosinolates could be detected (Tables 2 and 3). However, the presence of GNA (3-butenyl) which is an *AOP2* product (Fig 1) in Col-0-enhanced indicates that a functional allele is present in this accession but is just less effective than in An-1 and Cvi. The genetic diversity across alkenyl producing accessions, An-1 and Cvi, is also extendable to the *GS-OH* locus that mediates conversion of alkenyl aliphatic glucosinolates to the hydroxylated products such as PRO (4C) and GNL (5C) (Fig. 1). It appears that only An-1 has a functional allele of the *GS-OH* locus as Cvi lacks these two modified aliphatic glucosinolates (Table 2).

Surprisingly some lines that overexpress *MYB28* exhibit aliphatic glucosinolate pathway inhibition compared with the corresponding wild type even though they have a comparable level of *MYB28* expression as lines with enhanced glucosinolates (Fig. 2). This effect could be a symptom of translational inhibition of the *MYB28* mRNA. Many investigations have demonstrated that miRNAs play an important role in plant post-transcriptional gene regulation, particularly of transcription factors, by targeting mRNAs for cleavage or repressing translation (Sun *et al.* 2011; Zhang *et al.* 2006). When miRNAs perfectly or near-perfectly bind to mRNAs they inhibit translation and/or mediate the cleavage of mRNAs and therefore repress protein formation, whereas a miRNA that imperfectly binds to a mRNA only inhibits translation without mRNA cleavage (Brodersen *et al.* 2008). The observed inhibition can be due to a similar mechanism that is triggered by the high levels of *MYB28* mRNA. This would then suppress translation of *MYB28* mRNA and hence down-regulate the aliphatic glucosinolate pathway as a consequence. The suppression seems to occur to such an extent that the total concentration of glucosinolates is reduced to below or a comparable amount as the previously reported concentrations in *myb28*, *myb29* and *myb28*/*myb29* knockouts (Beekwilder *et al.* 2008; Müller *et al.* 2010). In line with this, Gigolashvilli *et al.* (2007) reported an increase in *MYB28* transcript level 1 minute after mechanical induction while no effect was observed on the level of *CYP79F1* expression within 120 min. As *CYP79F1* is strongly regulated by MYB28, also this suggests post-transcriptional control such as translation inhibition of MYB28.

Effect of MYB28 overexpression is dependent on the genetic background

In accordance with the expectations, constitutive overexpression of *MYB28* altered the concentration of aliphatic glucosinolates and the expression of some pathway genes in Col-0 and Eri background to a considerable extent while its impact on indole glucosinolates was small (Table 2). The alterations in aliphatic glucosinolate concentration were dependent on the genetic background. That is, mostly only endogenous aliphatic glucosinolates were changed quantitatively (Figure 2b and Table 2). As an example, expression of *AOP3* was affected by *MYB28* overexpression in Eri (Fig. 3), whereas no effect was observed in Col-0 transgenic lines due to the silence state of *AOP3* in Col-0 (Grubb *et al.*, 2006). As a consequence, we observed enhanced biosynthesis of the corresponding glucosinolate (3OHP) - that is also present in wild type Eri – in *MYB28*-overexpressing Eri but not in Col-0 (Table 2). One exception to this is the production of *de novo* alkenyl aliphatic glucosinolates (GNA and GBN) in *MYB28*-overexpressing Col-0, probably due to up-regulation of the *GS-ALK* gene *AOP2* (Fig. 3). However, despite the fact that *AOP2* expression was also induced in the Eri-enhanced line (Fig. 3), no traces of alkenyl aliphatic glucosinolate were observed in this

accession (Table 2). The fact that we detected *AOP2* expression is not completely in accordance with the suggestion by Kliebenstein and co-workers (2001b) that the two *AOP* genes cannot both be functional in a single *A. thaliana* accession. However, despite the expression of *AOP2* indeed we do not see any product of AOP2 in Eri, perhaps due to malfunction of AOP2 or the higher efficiency of AOP3 (of which expression is also induced by *MYB28* over expression) which competes with AOP2 for the substrate. Taken together this suggests that the impact of overexpression of *MYB28* on aliphatic glucosinolates is the same as after elicitation in wild type *Arabidopsis*. Induction by insect attack also leads to quantitative differences in the glucosinolate profile of the induced plant without *de novo* glucosinolate biosynthesis (Gols *et al.*, 2008; Mewis *et al.*, 2006; Textor *et al.*, 2009).

Other pathway genes that are involved in side-chain elongation *(MAM3*) and core glucosinolate biosynthesis (*CYP79F1*) were also differentially affected by *MYB28* overexpression in Col-0 and Eri background (Fig. 3). *MAM3* is mainly involved in the biosynthesis of long-chain aliphatic glucosinolates (Textor *et al.*, 2007). Sønderby *et al.* (2007) reported a decrease in *MAM3* expression in *MYB28* overexpression lines whereas there is no change in *MAM3* expression in *myb28/*myb29 double knockouts in Col-0 background. This is in contrast to our findings on Col-0 transgenic lines but is in agreement with our Eri-inhibited transformation event (Fig. 3). Beekwilder *et al.* (2008), however, reported comparable results for *MYB28*-affected *MAM3* expression in *myb28/*myb29 double knockouts as in our Col-0-inhibited. We cannot translate any of our observations on *MAM3* transcript level to the glucosinolate profile of the wild type and corresponding transgenic lines, but we can assume involvement of other regulatory mechanisms for *MAM3* expression that could be affected by the genetic background as well. *CYP79F1*, one of the targets of *MYB28* (Fig. 1) (Sønderby *et al.*, 2007), was also only affected in Col-0 background and not in Eri (Fig. 3). In conclusion, the effect of *MYB28* overexpression on the transcript level of genes in the chain elongation and core glucosinolate biosynthetic pathway and the biosynthesis of the corresponding aliphatic glucosinolates is background dependent.

Effect of MYB28 overexpression on generalist and specialist lepidopteran herbivores

Glucosinolates and/or their hydrolysis products were deterrent to the generalist *M. brassicae* and have a stimulating effect on the specialist *P. xylostella*. The latter even fed on younger *A. thaliana* leaves with higher concentration of glucosinolates (Lambdon *et al.*, 2005). In line with this observation, the generalist *M. brassicae* larvae gained relatively the least weight on Col-0-enhanced that contained a higher concentration of endogenous aliphatic

glucosinolates while the specialist *P. xylostella* larvae became relatively heaviest when feeding on this Col-0-enhanced. The generalist was relatively heaviest when feeding on wild type Col-0, Eri and Eri-inhibited, while the specialist grew less when feeding on these genotypes. These observations not only confirm the toxic effect of aliphatic glucosinolates on the generalist but could also indicate that high aliphatic glucosinolate containing genotypes are a better nutritious source for the specialist *P. xylostella*. *Pletella xylostella* larvae relatively gained most weight on the high glucosinolate producing lines even though they caused less damage on these genotypes. Whether the positive effect is caused by glucosinolates or is an indirect effect of other physiological changes in the plants needs to be investigated. The specialist *P. xylostella* possesses a glucosinolate detoxification mechanism, consisting of a sulfatase that largely prevents the formation of toxic hydrolysis products (Ratzka *et al.*, 2002). Despite this detoxification mechanism in the specialist *P. xylostella*, similar to the generalist *M. brassicae*, a negative correlation between aliphatic glucosinolate concentrations and the damage caused by the specialists can be observed. This is an important observation as it indicates potential for practical application of high aliphatic glucosinolate producing plants to control the damage caused by both generalists and specialists. However, before any practical recommendation, the negative effect of high aliphatic glucosinolate on the damage by Brassicaceae specialists should be repeated and the effect on the reproduction and the dynamics of offspring of the specialist must be studied as well. Poelman *et al.* (2009) also reported similar pattern of change in oviposition preference in response to the cultivar-specific variation in foliar glucosinolate profiles of *Brassica oleracea* by the specialist herbivore *Pieris rapae* and the generalist *M. brassicae*.

Overexpression of *MYB28* and the consequently higher glucosinolate concentration in Col-0 is mainly associated with lower growth rate of the generalist feeding on them while *MYB28* overexpression and the consequently higher aliphatic glucosinolate concentration in Eri is mainly associated with deterrence and escape of the generalist. On plants from all genetic backgrounds the specialist gained relatively more weight on plants having higher concentrations of aliphatic glucosinolates in *MYB28* overexpressing lines but still caused less damage. Exploiting the observed quantitative and qualitative differences between the genotypes used, we can shed light on the effect of specific groups of glucosinolates on the larval weight or damage caused by the generalist and the specialist. Cvi had the highest aliphatic glucosinolate concentration of all genotypes of this study but larval weight and damage caused by *M. brassicae* on Cvi were still higher than on Col-0-enhanced. Also the specialist caused the lowest damage and highest weight when feeding on Col-0-enhanced. This comparison also applies to a lesser extent to Cvi and Eri-enhanced. Correlation analysis showed a high association between insect parameters and sulfinyl glucosinolates, IBE (C3), RAPH (C4), ALY (C5) and SBE (C7) (Fig. 6). We therefore assume that chemical attributes

of the side chain of aliphatic glucosinolates are more important for their effect on insects than the length of the side chain. Simmonds et al (2001) also reported a relationship between the indolyl-, benzyl- glucosinolates of the leaf surface and oviposition preference by cabbage root fly females. We observed that methylsulfinyl glucosinolates are more effective in conferring resistance to caterpillars than hydroxyalkyl and alkenyl glucosinolates. This assumption is supported by a study on a *A. thaliana* Mr-0/Sap-0 population that showed that methylsulfinyl glucosinolate containing lines are more resistant to the generalist caterpillar *Spodoptera exigua* (Hübner) and to the specialist caterpillar *Pieris brassicae* (L.) than lines containing hydroxypropyl glucosinolates as main compounds (Rohr *et al.*, 2009). In another study on a different population, Rohr *et al.* (2011) show that leaf consumption of the same generalist was higher on 3-methylsulfinylpropyl-producing lines with low glucosinolate levels than on 3-hydroxypropyl-producing lines that contained a more than 2-fold higher concentration of glucosinolates (Rohr *et al.*, 2011). The latter also supports our observation on Eri-enhanced with the highest concentration of 3-hydroxypropyl glucosinolate, which inhibited feeding by *M. brassicae*. Our observation agrees with a study on *Brassica oleracea* cultivars with high concentration of IBE (C3 methylsulfinyl glucosinolate) that show low herbivore diversity (Poelman *et al.* 2009) compared to those genotypes containing less IBE.

Conclusion

This study shows that the effect of *HAG1/MYB28* overexpression is to a large extent background-dependent both with regard to gene expression as well as metabolite levels. Higher aliphatic glucosinolate concentration reduced the larval weight and damage by the generalist as expected, while it led to higher larval weight of the specialist only. Our data show that variation in the damage by the generalist and specialist particularly depends on the concentration of certain classes of glucosinolates.

Supplementary Table 1 List of glucosinolates analysed based on their classification

Class	Abbreviation	Trivial Name	Scientific Name	Precursor
Aliphatic	IBV (3MTP)	Glucoiberverin	3-(Methylthio)propyl	
	IBE (3MSOP)	Glucoiberin	3-(Methylsulphinyl)propyl	3-(Methylthio)propyl
	3OHP		3-hydroxypropyl	3-(Methylsulphinyl)propyl
	SIN	Sinigrin	2-Propenyl	3-(Methylsulphinyl)propyl
	ERU (4MTB)	Glucoerucin	4-(Methylthio)butyl	2-Oxo-6-methylthio-thiohexanoic acid
	RAPH (4MSOB)	Glucoraphanin	4-(Methylsulphinyl)butyl	4-(Methylthio)butyl
	GNA	Gluconapin	3-Butenyl	4-(Methylsulphinyl)butyl
	PRO	Progoitrin	(2R)-2-Hydroxy 3-butenyl	3-Butenyl
	EPRO	Epiprogoitrin	(2S)-2-Hydroxy 3-butenyl	3-Butenyl
	ALY (5MSOP)	Glucoalyssin	5-(Methylsulphinyl)pentyl	5-(Methylthio)pentyl (5MTP)
	GBN	Glucobrassicanapin	4-Pentenyl	5-(Methylsulphinyl)pentyl
	GNL	Napoleiferin	2-Hydroxy-4-pentenyl	4-Pentenyl
	C7thio (7MTH)		7-(Methylthio)heptyl	2-Oxo- 9- methylthio- nonanoic acid
	SBE (7MSOH)	Glucosiberin	7-(Methylsulfinyl)heptyl	7-(Methylthio)heptyl
	HIR (8MSOO)	Glucohirsutin	8-(Methylsulfinyl)octyl	8-(Methylthio)octyl
Indolic	GBC	Glucobrassicin	3-Indolylmethyl	

Supplementary Table 1 continued

Class	Abbreviation	Trivial Name	Scientific Name	Precursor
	4OH	4-hydroxyglucobrassicin	4-Hydroxy-3-indolylmethyl	
	4MeOH	4-methoxyglucobrassicin	4-Methoxy-3-indolylmethyl	
	NEO	Neo-glucobrassicin	1-Methoxy-3-indolylmethyl	
Aromatic	NAS	Gluconasturtiin	2-Phenylethyl	

References

Beekwilder, J., van Leeuwen, W., van Dam, N.M., Bertossi, M., Grandi, V., Mizzi, L., Soloviev, M., Szabados, L., Molthoff, J.W., Schipper, B., Verbocht, H., de Vos, R.C.H., Morandini, P., Aarts, M.G.M. and Bovy, A. (2008) The Impact of the Absence of Aliphatic Glucosinolates on Insect Herbivory in Arabidopsis. *Plos One*, **3**.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L. and Voinnet, O. (2008) Widespread translational inhibition by plant miRNAs and siRNAs. *Science*, **320**, 1185-1190.

Buchner, R. (1987) Approach to determination of HPLC response factors for glucosinolates. In *Glucosinolates in rapeseeds* (Wathelet, J.P. ed. Dordrecht, The Netherlands: Martinus Nijhoff Publishers, pp. 50-58.

Chan, E.K.F., Rowe, H.C. and Kliebenstein, D.J. (2010) Understanding the evolution of defense metabolites in Arabidopsis thaliana using genome-wide association mapping. *Genetics*, **185**, 991-1007.

Daxenbichler, M.E., Spencer, G.F., Carlson, D.G., Rose, G.B., Brinker, A.M. and Powell, R.G. (1991) GLUCOSINOLATE COMPOSITION OF SEEDS FROM 297 SPECIES OF WILD PLANTS. *Phytochemistry*, **30**, 2623-2638.

European Community (1990) Oilseeds-determination of glucosinolates High performance liquid chromatography. *Official Journal of the European Community L* **170/28**, Annex VIII: 03.07.27-34.

Gigolashvili, T., Yatusevich, R., Berger, B., Müller, C. and Flugge, U.I. (2007a) The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis in *Arabidopsis thaliana*. *Plant Journal*, **51**, 247-261.

Gols, R., Bukovinszky, T., Van Dam, N.M., Dicke, M., Bullock, J.M. and Harvey, J.A. (2008) Performance of generalist and specialist herbivores and their endoparasitoids differs on cultivated and wild Brassica populations. *Journal of Chemical Ecology*, **34**, 132-143.

Grubb, C.D. and Abel, S. (2006) Glucosinolate metabolism and its control. *Trends Plant Sci.*, **11**, 89-100.

Hirai, M.Y., Sugiyama, K., Sawada, Y., Tohge, T., Obayashi, T., Suzuki, A., Araki, R., Sakurai, N., Suzuki, H., Aoki, K., Goda, H., Nishizawa, O.I., Shibata, D. and Saito, K. (2007) Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 6478-6483.

Hopkins, R.J., van Dam, N.M. and van Loon, J.J.A. (2009) Role of Glucosinolates in Insect-Plant Relationships and Multitrophic Interactions. In *Annual Review of Entomology*, pp. 57-83.

Houshyani B, Kabouw P, Muth D, de Vos RCH, Bino RJ, Bouwmeester HJ. 2011. Characterization of the natural variation in *Arabidopsis thaliana* metabolome by the analysis of metabolic distance. *Metabolomics **In press***.

Jones, C.G. and Firn, R.D. (1991) On the evolution of plant secondary chemical diversity. *Philosophical Transactions - Royal Society of London, B*, **333**, 273-280.

Kabouw, P., Biere, A., Van Der Putten, W.H. and Van Dam, N.M. (2009) Intra-specific differences in root and shoot glucosinolate profiles among white cabbage (Brassica oieracea var. capitata) Cultivars. *Journal of Agricultural and Food Chemistry*, **58**, 411-417.

Kelly, P.J., Bones, A. and Rossiter, J.T. (1998) Sub-cellular immunolocalization of the glucosinolate sinigrin in seedlings of Brassica juncea. *Planta*, **206**, 370-377.

Kliebenstein, D.J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J. and Mitchell-Olds, T. (2001a) Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant Physiology*, **126**, 811-825.

Kliebenstein, D.J., Lambrix, V.M., Reichelt, M., Gershenzon, J. and Mitchell-Olds, T. (2001b) Gene duplication in the diversification of secondary metabolism: Tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in arabidopsis. *Plant Cell*, **13**, 681-693.

Lambdon, P.W. and Hassall, M. (2005) How should toxic secondary metabolites be distributed between the leaves of a fast-growing plant to minimize the impact of herbivory? *Functional Ecology*, **19**, 299-305.

Livak, K.J. and Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method. *Methods*, **25**, 402-408.

Mewis, I., Tokuhisa, J.G., Schultz, J.C., Appel, H.M., Ulrichs, C. and Gershenzon, J. (2006) Gene expression and glucosinolate accumulation in Arabidopsis thaliana in response to generalist and specialist herbivores of different feeding guilds and the role of defense signaling pathways. *Phytochemistry*, **67**, 2450-2462.

Miles, C.I., Del Campo, M.L. and Renwick, J.A.A. (2005) Behavioral and chemosensory responses to a host recognition cue by larvae of Pieris rapae. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, **191**, 147-155.

Müller, C. (2009) Interactions between glucosinolate- and myrosinase-containing plants and the sawfly Athalia rosae. *Phytochemistry Reviews*, **8**, 121-134.

Müller, R., de Vos, M., Sun, J.Y., Sønderby, I.E., Halkier, B.A., Wittstock, U. and Jander, G. (2010) Differential effects of indole and aliphatic glucosinolates on lepidopteran herbivores. *Journal of Chemical Ecology*, **36**, 905-913.

Neal, C.S., Fredericks, D.P., Griffiths, C.A. and Neale, A.D. (2010) The characterisation of AOP2: a gene associated with the biosynthesis of aliphatic alkenyl glucosinolates in Arabidopsis thaliana. *Bmc Plant Biology*, **10**.

Ratzka, A., Vogel, H., Kliebenstein, D.J., Mitchell-Olds, T. and Kroymann, J. (2002) Disarming the mustard oil bomb. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11223-11228.

Redovniković, I.R., Glivetić, T., Delonga, K. and Vorkapić-Furač, J. (2008) Glucosinolates and their potential role in plant. *Periodicum Biologorum*, **110**, 297-309.

Rohr, F., Ulrichs, C. and Mewis, I. (2009) Variability of aliphatic glucosinolates in Arabidopsis thaliana (L.)-Impact on glucosinolate profile and insect resistance. *Journal of Applied Botany and Food Quality*, **82**, 131-135.

Rohr, F., Ulrichs, C., Schreiner, M., Nguyen, C.N. and Mewis, I. (2011) Impact of hydroxylated and non-hydroxylated aliphatic glucosinolates in Arabidopsis thaliana crosses on plant resistance against a generalist and a specialist herbivore. *Chemoecology*, **21**, 171-180.

Sønderby, I.E., Geu-Flores, F. and Halkier, B.A. (2010) Biosynthesis of glucosinolates - gene discovery and beyond. *Trends Plant Sci.*, **15**, 283-290.

Sønderby, I.E., Hansen, B.G., Bjarnholt, N., Ticconi, C., Halkier, B.A. and Kliebenstein, D.J. (2007) A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. *Plos One*, **2**.

Sun, G. (2011) MicroRNAs and their diverse functions in plants. *Plant Molecular Biology*, 1-20.

ter Braak, C.J.F. (1988) CANOCO-an extension of DECORANA to analyze species-environment relationships. *Vegetatio*, **75**, 159-160.

Textor, S., De Kraker, J.W., Hause, B., Gershenzon, J. and Tokuhisa, J.G. (2007) MAM3 catalyzes the formation of all aliphatic glucosinolate chain lengths in Arabidopsis. *Plant Physiology*, **144**, 60-71.

Textor, S. and Gershenzon, J. (2009) Herbivore induction of the glucosinolate-myrosinase defense system: Major trends, biochemical bases and ecological significance. *Phytochemistry Reviews*, **8**, 149-170.

Vig, A.P., Rampal, G., Thind, T.S. and Arora, S. (2009) Bio-protective effects of glucosinolates - A review. *LWT - Food Science and Technology*, **42**, 1561-1572.

Wittstock, U., Agerbirk, N., Stauber, E.J., Olsen, C.E., Hippler, M., Mitchell-Olds, T., Gershenzon, J. and Vogel, H. (2004) Successful herbivore attack due to metabolic diversion of a plant chemical defense. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4859-4864.

Zhang, B., Pan, X., Cobb, G.P. and Anderson, T.A. (2006a) Plant microRNA: A small regulatory molecule with big impact. *Developmental Biology*, **289**, 3-16.

Zhang, X., Henriques, R., Lin, S.S., Niu, Q.W. and Chua, N.H. (2006b) Agrobacterium-mediated transformation of Arabidopsis thaliana using the floral dip method. *Nature Protocols*, **1**, 641-646.

Chapter 6

Assessment of transcriptome perturbations in *Arabidopsis* lines with genetically engineered indirect insect defence

Benyamin Houshyani · Sander A.R. van der Krol · Raoul J. Bino · Harro J. Bouwmeester

Submitted to Plant Journal (2011)

Abstract

Molecular characterization is an essential step of risk/safety assessment of genetically modified (GM) crops. Holistic approaches for molecular characterization using *omics platforms* can be used to confirm the intended impact of the genetic engineering, but also can reveal the unintended changes at the *omics* level as a first assessment of potential risks. The potential of *omics* platforms in risk assessment of GM crops has rarely been used for this purpose because of lack of a consensus reference and statistical methods to judge the significance or importance of the pleiotropic changes in GM plants. Here we propose a *meta data* analysis approach of GM plants, measuring the transcriptome distance to untransformed parental plants. In the statistical analysis of the distances, values are compared to natural occurring transcriptome distances in non-GM counterparts. Using this approach we show that the pleiotropic effect of genes involved in indirect insect defence traits are substantially equivalent to the variation and plasticity in gene expression occurring naturally in *Arabidopsis*.

Introduction

According to the consensus document on the assessment of plants derived from modern biotechnology, a molecular characterization must be included to provide assessors with the possibility to predict phenotypes and risk/safety concerns (OECD 2010). To be able to do so, reliable methods and tools for characterization and risk/safety assessment are essential. To assess the risk of *novel foods*[2] (such as a *GM food*[3]), the term "substantial equivalence" was introduced by the OECD in 1991. This concept implies that a novel food should be considered the same as and as safe as a conventional food (the safe counterpart) if it has the same characteristics and composition. Development of reliable methods and tools to analyse the equivalence thus is important from a regulatory point of view because it may be used as an argument for GM foods to be covered by the same regulatory framework as its conventional counterpart **(**CRS Report for Congress: 2005; http://ncseonline.org/nle/crsreports/05jun/97-905.pdf).

The untargeted measurement techniques collectively referred to as ''*omics*" (proteomics, metabolomics and transcriptomics) can be used for characterizing and evaluating the effects of transgene insertion and compositional equivalence of GM crops relative to their conventional (safe) counterparts (Baker *et al.* 2006, Catchpole *et al.* 2005, Kuiper *et al.* 2001). These techniques should confirm the intended impact of the novel trait[4] (has the intended change occurred) but can also reveal the unintended changes. To judge whether an unintended change is significant, however, the magnitude of the changes should be evaluated and judged against a *baseline* representing the natural variation in the trait under evaluation (proteome, metabolome and/or transcriptome) in the natural parental lines (Baudo *et al.* 2006), wild relatives (Cheng *et al.* 2008, Harrigan *et al.* 2007), populations derived from the safe counterparts and populations exposed to naturally occurring biotic and/or abiotic stress factors (Coll *et al.* 2010). Such a comparison is not trivial, as for each sample thousands of data points are generated with each of these *omics* technologies.

Here, we studied the transcriptome changes in two genetically modified (GM) *Arabidopsis* lines. Both lines expressed a mitochondrial targeted nerolidol synthase (a sesquiterpene synthase) gene (*COX-FaNES1*) from strawberry (*Fragaria x ananassa*) (Kappers *et al.* 2005). On top of that, two strategies were followed to boost expression of the trait by improving substrate availability. In one line (COX+) the precursor availability was boosted by overexpression of a mitochondrial targeted farnesyl diphosphate (FPP) synthase 1 long isoform (*FPS1L, At5g47770*) (Fig. 1a) (Manzano *et al.* 2006). In the other (COX++), precursor availability was boosted by overexpression of both *FPS1L* and a cytosolic

hydroxymethylglutaryl CoA reductase 1 short isoform (*HMGR1S, At1g76490*) (Fig. 1a). *HMGR1S* encodes an enzyme catalyzing a step of the mevalonate pathway and *FPS1L* encodes an enzyme catalyzing the last step before sesquiterpene biosynthesis. Both are pleiotropic and rate limiting enzymes (Closa *et al.* 2010, Leivar *et al.* 2005). Both lines emit the volatile compound, (*E*)-nerolidol in the head space and are similarly efficient in attraction of the endolarval parasitoid of *Plutella xylostella* (diamondback moth), *Diadegma semiclausum* (this thesis). The use of different strategies to generate the same trait, as for the COX+ and COX++, can help to classify potential changes in the transcriptome into changes that are specifically associated with the novel trait(s) or changes that are non-specific.

The following workflow was used to analyse the transcriptome changes in *Arabidopsis*: 1- changes in the transcriptome of the transformed lines were first compared to the Col-3 (wild type) background, 2- intended and unintended changes that form the overall perturbation were analysed and 3- the substantial transcriptome equivalence of the overall perturbation was statistically evaluated and assessed by comparing the transcriptome changes with natural *Arabidopsis* transcriptome variation, the *baseline*. Hereto, we used microarray data that were obtained specifically for this study as well as data from publically available databases on 16 *Arabidopsis* accessions and four groups of lines derived from an *Arabidopsis* RIL population (Ler/Cvi). The transcriptomics data for the *Arabidopsis* accessions represent the genetic transcriptome variability caused by diversification of a common ancestor's genome. This variability is achieved by natural mutations combined with local evolutionary selection pressure, resulting in diverse but supposedly balanced genome compositions of the different accessions and consequently different transcriptional profiles (Fig. 2a). The RIL population represents the genetic diversity caused by mixing the Cvi and Ler genomes and hence resembles domestication of plant species through conventional and modern breeding activities (Fig. 2b). The recently developed statistical method to determine the metabolic '*hyper-plane distance*' (Houshyani *et al,* 2011) was used to calculate a '*transcriptome distance*' between groups of samples which allows comparison of the substantial transcriptome equivalence of GM lines with the *baseline*.

[1] OECD, The Organization for Economic Co-operation and Development (OECD) is an international economic organization to stimulate economic progress and world trade. It is a forum of countries committed to democracy and the market economy, providing a platform to compare policy experiences, seek answers to common problems, identify good practices, and co-ordinate domestic and international policies of its members.

[2] Novel food is a type of food that does not have a significant history of consumption or is produced by a method that has not previously been used.

[3] Genetically modified (GM) foods are foods derived from genetically modified organisms. Genetically modified organisms have had specific changes introduced into their DNA by genetic engineering techniques.

[4] Trait in this paper is the characteristic that is the aim of genetic modification.

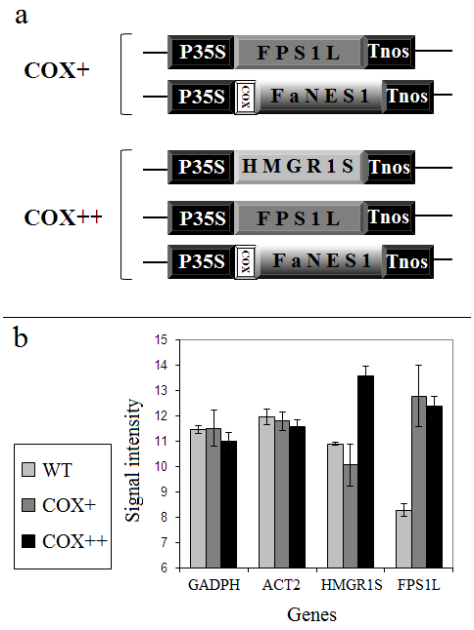


Figure 1 Constructs used for transgenes expression in *Arabidopsis*. (a) Genes stacked in COX+ and COX++ lines by crossing: *HMGR1S*, short (cytosolic) isoform of 3-hydroxy-3-methylglutaryl coenzyme A reductase 1; *FPS1L*, farnesyl diphosphate synthase 1 long isoform; *FaNES1*, nerolidol synthase 1 from *Fragaria* X *ananassa*; COX, mitochondrial signal peptide; P35S, CaMV ***35S promoter; Tnos, terminator*** of the *Agrobacterium tumefaciens* nopaline synthase. (b) Average signal intensities of the overexpressed (*HMGR1S* and *FPS1L*) and housekeeping (*GADPH* and *ACT2*) genes in Col-3 (WT), COX+ and COX++ lines. Bars indicate the 95% confidence interval

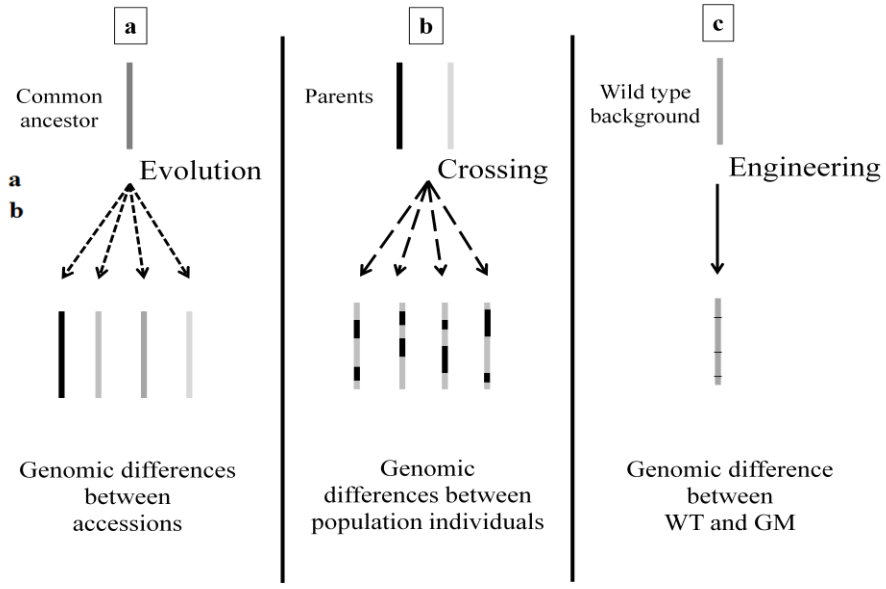


Figure 2 Potential genetic sources for transcriptome variation. Each bar represents the whole genome of an individual line. WT, wild type. GM, genetically modified

Experimental Procedure

Plant material

The generation of COX+ and COX++ transgenic lines in *Arabidopsis thaliana* (accession Col-3) background is described elsewhere (Fig. 1) (Houshyani *et al,* 2012). Transgenic lines and wild type plants were grown on LB medium (purified agar 0.8% + 2.2 gr L^{-1} 0.5 MS), supplemented with the herbicide BASTA (10 µg mL^{-1}) for transgenic line selection. Plates were placed in a growth chamber (21 ± 2 °C with L8:D16 photoperiod with 80-110 µmol.m^{-2}.s^{-1} PPF). When seedlings had 2 true leaves they were transferred to soil (Lentse potgrond, Lent, The Netherlands) and grown for 6 weeks under the same conditions. Growth of COX+ and COX++ lines was slightly retarded during the first 4 weeks after transplanting, but this difference largely disappeared in the last 2 weeks before sampling.

Total RNA extraction, reverse transcription and qPCR analysis

Sampling of four young fully expanded leaves was done if insect (*Diadegma semiclausum*) preference for its volatile blend was observed, indicating that the plant was producing the expected volatiles (Houshyani *et al,* 2012). Therefore sampling was done during a period of five days in the same time of the day and under the same environmental conditions as described above. Samples were immediately flash frozen in liquid nitrogen, stored at -80°C, ground in liquid nitrogen and total RNA was isolated using the protocol of TriPure Isolation Reagent (Roche Applied Science, www.roche-applied-science.com). Total RNA was treated with DNaseI (Invitrogen, www.invitrogen.com) according to the manufacturer's instructions and purified using the RNeasy Mini Kit (Qiagen, www.qiagen.com) and the RNA Cleanup protocol. RNA samples were quantified (UV absorption at 260 nm) using an ND1000 Spectrometer (Nanodrop technologies, www.nanodrop.com). OD 260/280 nm absorption ratio (>2) and agarose gel electrophoresis analysis confirmed the purity and integrity of RNA samples.

One µg total RNA was subsequently converted to cDNA using the iScript cDNA synthesis kit (Bio-Rad, www.bio-rad.com). Gene-specific primers were designed using Beacon Designer 7.0 (Premier Biosoft, www.premierbiosoft.com) (Supplementary Table 1). Primers were checked for gene specificity by blasting against the *Arabidopsis* genome and RefSeq RNA database and by performing melt curve analysis on a MyIQ Single-Color Real-Time PCR Detection System (BioRad). The amplification efficiency of PCR primers was determined by performing RT-PCR on dilution series of a template.

Quantitative RT-PCR analysis was carried out in optical 96-well plates with a MyIQ Single-Color Real-Time PCR Detection System (BioRad), using SYBR Green. Each reaction contained 10 µl 2x iQTM SYBR Green Supermix Reagent (BioRad), 20 ng cDNA and 150 nM of each gene-specific primer in a final volume of 20 µl. All qRT-PCR experiments were performed in duplicate. The following PCR program was used for all PCR analyses: 95°C for 3 min; 40 cycles of 95°C for 10 s and 60°C for 30 s. Threshold cycle (Ct) values were calculated using the MyIQ Optical System software (version 2.0, BioRad). Subsequently, Ct values were normalized for differences in cDNA synthesis by subtracting the Ct value of the reference gene β-tubulin (Chen *et al.* 2003) from the Ct value of the gene of interest. Normalized Ct values (δCt) were used for statistical comparison of wild type and transgenic lines using SPSS (t-test, α=0.05) (IBM, www.ibm.com).

RNA labelling, microarray hybridization and data processing

The total RNA of samples in which expression of *FaNES1*, over-expression of *FPS1L* and *HMGR1S* and emission of nerolidol were confirmed were provided to ServiceXS (www.servicexs.com), who performed labelling, hybridization, quality control and data acquisition. Briefly, the RNA concentration and 260/280 nm absorbance ratio measured by the NanoDrop ND-1000 Spectrophotometer (NanoDrop technologies) and electropherograms (plot of results from an electrophoresis analysis) and RNA integrity number produced by the Agilent 2100 Bioanalyzer (Lab-on-a-chip Technology, Agilent, www.chem.agilent.com) were used to re-check the RNA quality before labelling. An RNA sample was considered suitable for array hybridization if it had a concentration of 100-500 ng/µl, 260/280 ratio of around 2.0, intact bands on the gel corresponding to 18S and 28S ribosomal RNA subunits and no chromosomal peaks or RNA degradation products (RIN>5.0) (http://www.chem.agilent.com/RIN/). Hundred ng of RNA was used to synthesize cDNA and Biotin-labelled cRNA using the Affymetrix 3' IVT-Express labelling Kit (www.affymetrix.com). The labelling controls were added to the RNA before labelling. The possibility of very short cRNA formation that can cause a 3' – 5' bias and influences the data analysis was ruled out by lab-on-a-chip analysis (Agilent). Fifteen µg cRNA was used for further fragmentation to prevent secondary structure and probe proximity interference and finally 10 µg for the hybridization to the Affymetrix *Arabidopsis* Genome ATH1 Arrays (Redman *et al.* 2004). The GeneChip Hybridization, Wash and Stain Kit (Affymetrix) was used for the hybridization, washing, staining and scanning of the chips. Thirty µl of labelled material was added to 270 µl hybridization cocktail having hybridization controls added. The Affymetrix protocols were strictly followed. Labelling and hybridization controls showed that the processes were reliable.

The Affymetrix Command Console (v1.1) and Expression Console software (v1.1) provided signal estimation and Quality Control (QC) functionality for the GeneChip Expression Arrays. To check quality and applicability of the generated microarray data, the distribution of the log2-transformed intensities was viewed by boxplots and smoothed histograms. They showed no shift in the distribution of the RMA normalized data (Supplementary Fig. 2). To identify outliers relative to the bulk of samples in the data set, each sample was compared to a reference. Since no experimental reference sample was included, it was generated *in-silico* by calculating the median for each gene across all samples. Subsequently, RLE (Relative Log Expression) and NUSE (Normalized Unscaled Standard Errors) plots were generated for RMA normalized data. Samples centered at zero in an RLE plot showing no outlyers. The histogram of the normalized data using the quantile normalization employed in the RMA algorithms also showed no bias for the amplified samples.

Meta data preparation

The ArrayExpress database for gene expression experiments in EMBL-EBI (http://www.ebi.ac.uk/arrayexpress) was used to query data obtained with Affymetrix ATH1 chips using the following keywords: "*Arabidopsis"* AND "accession OR ecotype". Five experiments from about 1000 hits were selected that have used the same tissue in a similar developmental stage and of plants grown in almost similar experimental conditions as in our study. The 5 experiments (E-GEOD-5728 and 12676, E-MEXP-1799 and 2144 and ETABM-18) consisted of CEL files of 64 chips (samples).

The expression data of a Cvi/Ler RIL population (160 lines) were kindly provided by Dr. J. Keurentjes (Wageningen University). For the sake of statistical analysis and *hyper-plane* distance calculation, 4 groups of RILs were selected from the RIL population each comprising of 5 lines. Two of these groups (GPs) were separated on the first three PCs of a PCA plot performed on just the expression data of the RIL population (expression GPs) and the other two were separated on the first two PCs of a PCA plot which was produced by the molecular marker data of the population (genetic GPs) (https://cbsgdbase.wur.nl/*Arabidopsis*/demo/marker/markers-index.php).

Expression data of the RIL population, of the wild type plants of this study and of the ecotypes from studies in the public databases were used to generate the *baseline;* addition of the expression data of transgenic lines to *baseline* formed the *meta data* of this study. Systematic biases resulting from different sources of RNA and batches of microarrays in the *meta data* were removed by the "remove effect" function of GeneMaths XT.

Data analysis

GeneMaths XT (www.applied-maths.com) was used for pre-processing of the image data in the CEL files, visualization and PCA. The normalization of all arrays was performed using the Robust Multichip Average (RMA) algorithm with standard settings. All statistical comparisons were performed in PASW statistics 17 (SPSS) using ANOVA in conjunction with Tukey's test (α=0.05). Gene set enrichment analysis (Mootha *et al.* 2003, Subramanian, *et al.* 2005) was done using the GSEA desktop application (http://www.broadinstitute.org) to determine whether an *a priori* defined set of genes shows statistically significant difference between wild type plants and COX+ or COX++ transgenic lines. A database of 555 gene sets was used for GSEA (Groot and van der Graaff 2009). Over- and Under-Representation Analysis (ORA) of functional categories (Keller *et al.* 2008) in the test set (a set of differentially expressed genes) was done using GeneTrail (http://genetrail.bioinf.uni-sb.de) and a reference set comprising all gene IDs of *Arabidopsis* ATH1 GeneChip. Venn Diagram Generator (http://www.pangloss.com/seidel/Protocols/venn.cgi) was used for producing Venn diagrams to identify overlapping genes in the set of differentially expressed genes. The method of Houshyani *et al.* (2011) for *hyper-plane distance* was used to calculate the *transcriptome distance* between all used groups and genotypes using scores of samples on the first 9 PCs of a PCA plot using the *meta data*. The 9 principle components were selected using Horn's Parallel Analysis (http://doyenne.com/Software/files/PA for PCA *Vs* FA.pdf). The *transcriptome distance* calculation was done using non-weighted PC scores, as well as weighted. For the latter, the scores of samples on each PC were multiplied by the variation that was explained by that PC.

The *transcriptome distance* between two groups varies between 1.00 and -1.00 with 1.00 indicating the maximum distance between two groups or no overlap in the location of the group in the *hyper-plane*, 0 indicating two completely overlapping groups and -1.00 indicating two groups where one group is a child of the other group.

Results

Transformation of Col-3 plants used in this study was done by the insertion of 2 (in COX+) or 3 (in COX++) constructs, each containing 2 transgenes including the gene of interest (*COX-FaNes1*, *FPS1L* or *HMGR1S*) (Fig. 1a) and a selection-marker gene (Houshyani *et al*, 2012). Quantitative RT-PCR showed *de novo FaNES1* expression in the transgenic lines. *FPS1L* expression was at least 64-fold higher in COX+ and COX++ and expression of *HMGR1S* was at least16-fold higher in COX++ lines than in Col-3 (the wild type) (Houshyani *et al*, 2012).

For the analysis of the transcriptional activity in the GM and wild type plants, RNA of Col-3, COX+ and COX++ transgenic lines was isolated for *Arabidopsis* ATH1 GeneChip hybridization. Sampling was in a stage at which the GM plants produced a volatile blend that attracted parasitoid wasps (*Diadegma semiclausum*) (Houshyani *et al*, 2012). *FPS1L* and *HMGR1S* are *Arabidopsis* genes and are present on the *Arabidopsis* micro-array. The average signal intensity for *FPS1L* was 22.6 fold (4.5 units) higher in COX+ or COX++ than in the Col-3 and for *HMGR1S* the increase in signal intensity was 5.7 fold (2.5 units), which is similar to the quantitative RT-PCR results (Fig. 1b).

Transcriptome changes

Quantitative changes in the transcriptome

Comparing COX+ lines with Col-3 plants, 545 probe sets (genes) were differentially expressed (2.38% of the total) of which 139 (0.60%) and 35 (0.15%) were more than two fold up- or down-regulated in COX+ lines, respectively. In the COX++ versus Col-3 comparison, the number of differentially expressed genes was 485 (2.12% of total) of which 13 (0.05%) were more than two fold up- and 85 (0.37%) were more than two fold down-regulated. Only 131 genes (0.57% of total) were differentially expressed in both modifications of which 47 (0.20%) were more than two fold up- or down-regulated in at least one of the modifications. A closer look at the list of 47 genes and their pattern of change revealed that only four genes have the same pattern of change, one up-regulated (*At5g47770*, *FPS1*) and three down-regulated (*At4g29020*, *At3g30720*, *At3g50360*) in both COX+ and COX++ lines and the rest showed a reverse pattern. Combined, these results indicate a negligible overlap or a remarkable difference between the two transgenic strategies with respect to their impact on the transcriptome of Col-3.

XY scatter plots allow visualization and comparison of the global gene expression variability across Col-3 and the transgenic lines. First the variability between Col-3 replicates was used to establish the *baseline* variation in these XY scatter plots (Fig. 3a, 1st row). XY scatter plot of an individual Col-3 replicate and a transgenic line of COX+ and COX++ group with the smallest R^2 (therefore the highest variability) are also shown in Figure 3a 2nd and 3rd

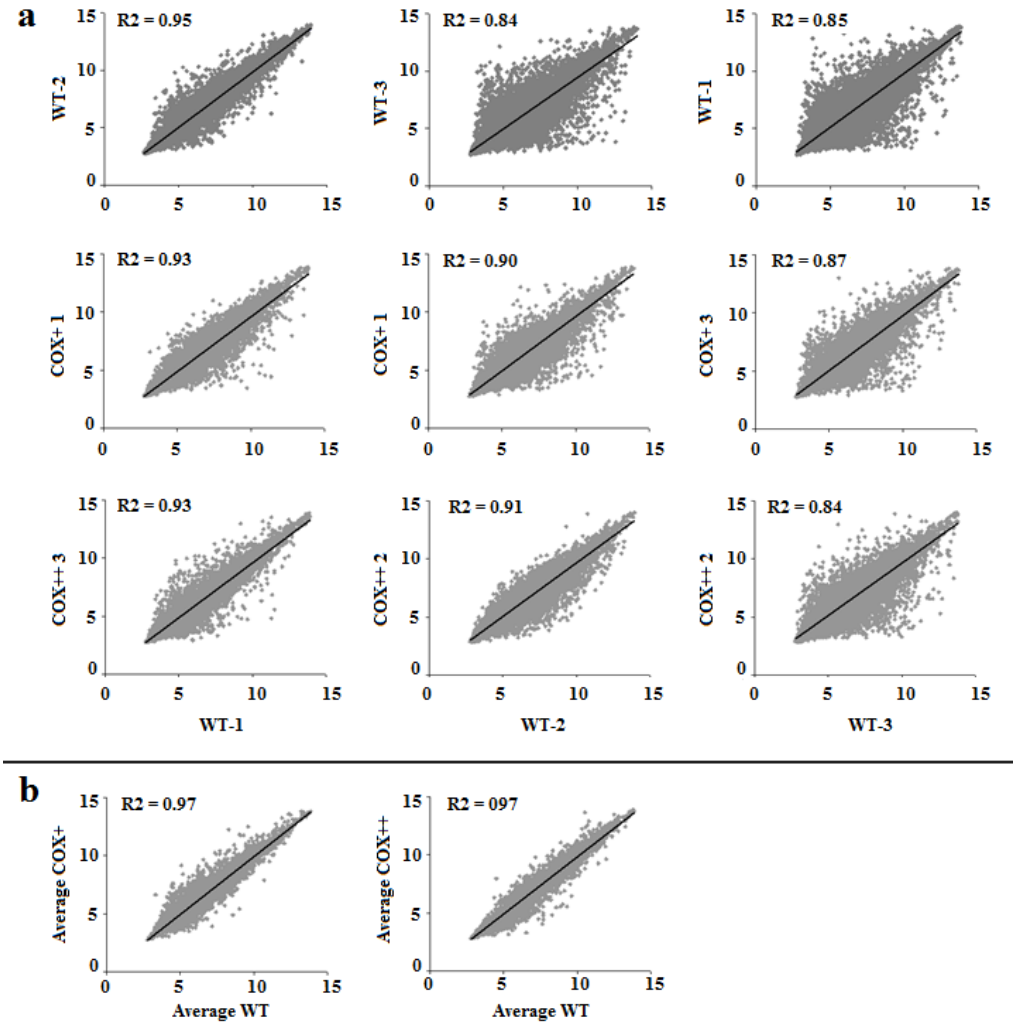


Figure 3 XY scatter plots with transcriptome data. The correlation constant (R^2) represents the variability of the global gene expression profile of two samples or groups. A small R^2 indicates large variation, a large R^2 indicated small variation. (a). XY scatter plots of the Col-3 replicates (WT) with each other (1st row) and with a replicate of COX+ and COX++ with the smallest R^2 value (2nd and 3rd row, respectively). (b). XY scatter plot for the averages of the groups.

row, respectively. Comparison of the Col-3 scatter plots (*baseline*) with that of transgenic lines showed no obvious influence of the genetic modification on the global gene expression pattern as Col-3 correlations have an equal or smaller R^2 (thus larger variation) than the transgenic lines (Fig 3a). XY scatter plots for the group averages show that the averaged Col-3 gene expression values are globally a reliable predictor for the expression of most of the genes in both transgenic groups with an R^2 value of 0.97 for both the Col-3 *Vs.* COX+ and COX++ scatter plots (Fig. 3b).

Gene Set Enrichment analysis of transcriptional changes

Gene Set Enrichment analysis (GSEA) (Mootha *et al.* 2003, Subramanian *et al.* 2005) was used to determine whether any *a priori* defined sets of genes (555 sets) (Groot and van der Graaff 2009) are differentially expressed between Col-3 and COX+ and COX++ transgenic lines. Instead of analysing the correlation of a single gene with the new biological state (GM), GSEA derives its power from looking at the effect of genetic modification in sets of genes that share a common biological function, cellular localization, chromosomal location or regulation.

None of the gene sets showed a significant change in the transgenic lines compared with Col-3 in GSEA. The absence of a significant difference can be due to either negligible variation caused by the introduction of the transgenes compared to other sources of variation. GSEA also resulted in a list of genes ranked by their correlation with *FPS1L* expression. The heat map (Supplementary Fig. 1) shows that 17 genes have more than 90% correlation with the expression level of *FPS1L* (six positively and 11 negatively). This list also includes the three genes that were more than two fold down-regulated in both transgenic lines, *At4g29020* (encoding an endomembrane glycine-rich protein), *At3g30720* (encoding qua-quine starch, a cytosolic protein) *and At3g50360* (encoding centrin2, the plasma membrane calcium binding protein) none of which we can currently give a clear biological link to either *FaNES1* expression or *FPS1L* overexpression.

Analysis of functional categories in transcriptional changes

Using over-representation analysis (ORA), the enrichment of functional categories in the set of significantly up- or down-regulated genes was analysed. A set comprising all probe set IDs of *Arabidopsis* ATH1 GeneChip was used as a reference for statistical evaluation.

Ignoring the fold change criteria, 318, 154 and 29 genes were significantly up-regulated in COX+, COX++ and both lines, respectively, compared with the wild type Col-3

(ANOVA in conjunction with Tukey, α= 0.05). The first two sets significantly represented 198 and 16 categories in COX+ and COX++ lines compared with the reference set, respectively (ORA with FDR, 0.05). No functional over-represented category was present in both transgenic lines confirming the lack of similarity between the transcriptomes of the COX+ and COX++ lines. Ignoring fold change in down-regulation, 227, 330 and 34 genes were significantly different in COX+, COX++ and both lines, respectively, compared with the wild type Col-3 (ANOVA in conjunction with Tukey, α= 0.05). The first two sets significantly represented 59 and 242 categories in COX+ and COX++ compared with the reference set, respectively (ORA with FDR, 0.05). Only the "protein metabolic process" category was significantly represented in both lines but under-represented in COX+ and over-represented in COX++.

In order to narrow down our search for functional categories that could be specifically related to the molecular mechanism behind the novel trait, ORA was done on the commonly up- or down-regulated genes in both COX+ and COX++ lines. Among the 29 up-regulated and the 34 down-regulated genes in both COX+ and COX++ lines, no specific functional category was over- or under-represented (ORA with FDR, 0.05). This indicates that common transcriptome changes in COX+ and COX++ lines are rare and that the novel trait does not represent defined changes in certain biological categories.

Multivariate analysis of GM transcriptome data

A PCA plot of the RMA normalized data showed distinct clustering of COX+ and COX++ lines along PC1, PC2 and PC3 (Fig. 4a and b). Col-3 replicates separated from neither of the transgenic lines along these PCs. Clustering of COX+ and COX++ lines on different sides of Col-3 suggests a difference in the impact of the two strategies on the transcriptome. The PCA plots also show that the variation in the transcriptome of the transgenic lines along PC1, PC2 and PC3 is within the Col-3 variation, except for variation of COX+ lines along PC2 (Fig. 4a) and COX++ lines along PC3 (Fig. 4b). Col-3 samples showed larger variation along the first PC compared with the transgenic lines (Fig. 4a). This variation coincided with the larger number of highly variable genes in Col-3 replicates as 917 (4.0%), 237 (1.0%) and 242 (1.0%) of the genes had a CV of 20% up to 70% (maximum) in Col-3, to 61% in COX+ and to 44% in COX++ , respectively. A PCA after excluding the highly variable genes (with a CV > 20%) shortened the visual distance between the Col-3 replicates (Fig. 4c). However, it did not change the overall conformation of the groups relative to each other. To check whether the observed variation within the Col-3 plants of this study had a true biological origin, we performed ORA on the set of genes with CV > 20% in Col-3. The most represented

subcategories in the test set belonged to the catalytic activity, primary metabolic process and response to abiotic stimulus categories which may be due to the effect of sampling day.

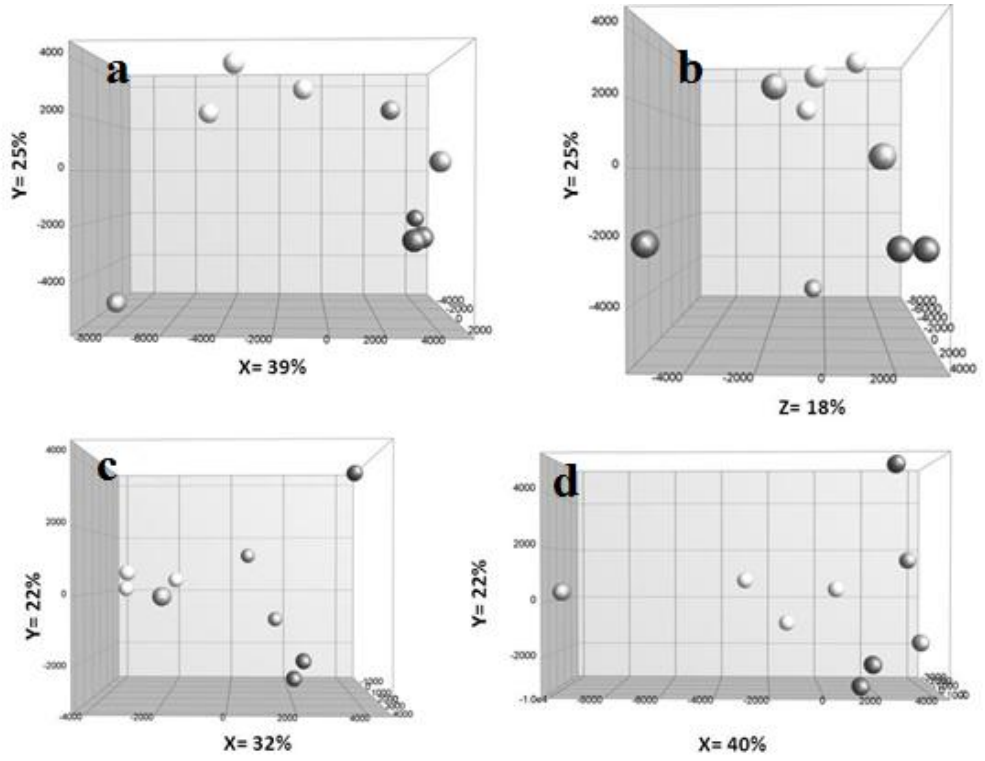


Figure 4 PCA plots using RMA normalized genes data of Col-3 plants and COX+ and COX++ lines. a, PCA plot with PC1 (X) and PC2 (Y). b, PCA plot with PC3 (X) and PC2 (Y). c, PCA plot using genes with CV < 20% across wild type samples with PC1 (X) and PC2 (Y). d, PCA plot using ANOVA insignificant genes (α =5%) with PC1 (X) and PC2 (Y). Percentages are the variation explained by the corresponding PC. Dark grey circles: wild type samples, light grey: COX+ and black: COX++ samples

To confirm that the observed variation within Col-3 replicates is independent from the variation in differential genes between Col-3, COX+ and COX++, all significantly different genes were filtered out by ANOVA (α = 0.05). PCA on the remaining genes showed that groups became closer, but a large variation was still observed within Col-3 replicates (Fig. 4d). This observation confirms independence of the large variation within Col-3 replicates from between group variation.

Table 1. *Transcriptome distances* between *Arabidopsis* genotypes. Values above the diagonal are the distances based on intact (un-weighted) PC scores and values below the diagonal are distances based on the weighted scores on the first 9 PCs of a PCA on gene expression data of the wild type and transgenic lines of this study (a), wild type accessions of public databases (b) and groups of the Cvi/Ler RIL population (c). Significant values are shown in bold (permutation test, P-value=<0.05). a, b and c correspond to the source of the data which is delimited by blocks. Accessions shaded by the same color belong to the same experiment.

		Col-3	COX+	COX++	Bur0-23	Pf0-23	Pf0-27	Cal-0	Col-4	GM	Bay	C-24	Col-0	Cvi	Estland	Kindaville	Ler	Niederzenz	Shahdara	Vancouver	PCA RIL1	PCA RIL2	genetic RIL1	genetic RIL2
Study number						1799			5728	2144							18							
Col-3	**a**		0.19	-0.07	-0.33	-0.33	-0.33	0.06	**0.41**	0.15	0.19	0.22	0.22	0.11	0.15	0.22	0.07	0.15	0.19	0.11	**0.59**	**0.70**	**0.59**	**0.59**
COX+		0.26		1.00	0.50	0.58	0.58	0.30	**0.89**	0.63	0.70	0.56	0.78	0.56	0.70	0.85	0.78	0.67	0.59	0.56	**0.92**	**0.81**	**0.89**	**0.86**
COX++		0.15	1.00		0.33	0.25	0.17	**0.48**	**0.54**	0.48	0.37	0.81	0.48	0.63	0.44	0.56	0.48	0.48	0.67	0.70	**0.70**	**0.89**	**0.75**	**0.78**
Bur0-23	**b**	-0.33	0.50	0.83		0.50	0.50	-0.14	0.14	0.00	0.00	1.00	1.00	1.00	0.00	0.33	0.00	0.00	0.17	1.00	**1.00**	**0.99**	**0.55**	0.00
Pf0-23		-0.42	0.50	0.75	0.25		-0.75	-0.39	-0.21	-0.25	0.00	0.58	0.75	0.75	-0.08	0.50	-0.25	-0.25	0.17	0.67	**0.82**	**0.89**	0.49	-0.11
Pf0-27		-0.42	0.50	0.50	0.00	-0.50		-0.32	-0.29	-0.25	0.00	0.42	0.75	0.58	0.00	0.50	-0.25	-0.25	0.08	0.67	0.67	**0.90**	0.60	0.02
Cal-0		0.04	0.26	**0.65**	-0.14	-0.32	-0.32		0.10	-0.13	0.13	0.15	**0.48**	0.06	0.09	0.43	-0.07	-0.11	-0.06	0.15	**0.56**	**0.62**	**0.36**	0.15
Col-4		0.24	**0.80**	0.59	-0.14	-0.25	-0.25	0.07		-0.09	0.00	0.39	0.44	0.44	0.02	0.52	0.24	0.15	0.07	0.37	**0.46**	**0.80**	**0.37**	0.21
GM		-0.07	0.52	0.70	0.00	-0.25	-0.25	-0.09	-0.04		0.22	0.41	0.78	0.56	0.19	0.70	0.15	-0.11	-0.04	0.56	**0.69**	**0.81**	0.33	-0.07
Bay		-0.04	0.70	0.52	0.00	0.00	0.00	0.15	-0.02	0.04		0.81	0.56	0.78	0.04	0.48	0.41	0.19	0.33	0.78	**0.69**	**0.88**	**0.66**	0.16
C-24		0.22	0.44	1.00	1.00	0.58	0.42	0.13	0.39	0.41	0.74		1.00	0.70	0.78	1.00	0.70	0.63	0.37	0.78	**1.00**	**1.00**	**1.00**	**0.73**
Col-0		0.22	0.78	0.78	1.00	0.75	0.75	**0.48**	0.41	0.78	0.56	1.00		1.00	0.22	1.00	0.78	0.78	0.78	1.00	**1.00**	**1.00**	**1.00**	**0.84**
Cvi		0.22	0.52	1.00	1.00	0.67	0.25	0.06	0.44	0.56	0.78	0.63	1.00		0.78	1.00	0.56	0.56	0.56	1.00	**1.00**	**1.00**	**0.99**	**0.50**
Estland		-0.04	0.67	0.63	0.00	-0.08	0.00	0.09	-0.06	0.00	-0.11	0.74	0.07	0.78		0.89	0.37	0.15	0.15	0.78	**0.69**	**0.90**	**0.59**	0.26
Kindaville		0.15	0.78	0.67	0.33	0.42	0.42	0.33	0.20	0.33	0.11	1.00	0.93	1.00	0.26		0.93	0.78	0.96	1.00	**0.91**	**0.99**	**0.84**	**0.66**
Ler		-0.07	0.70	0.78	0.00	-0.25	-0.25	-0.06	0.04	0.15	0.30	0.56	0.78	0.56	0.22	0.33		-0.04	0.26	0.56	**0.69**	**0.81**	**67**	0.12
Niederzenz		0.04	0.59	0.81	0.00	-0.25	-0.25	-0.09	0.00	-0.19	0.11	0.56	0.78	0.56	0.11	0.63	0.04		0.11	0.63	**0.82**	**0.82**	0.38	-0.16
Shahdara		0.22	0.56	0.96	0.00	0.08	0.17	-0.15	0.07	-0.04	0.26	0.33	0.78	0.56	0.22	0.70	0.07	0.04		0.63	**0.86**	**0.80**	**0.64**	0.15
Vancouver		0.22	0.56	1.00	1.00	0.58	0.58	0.04	0.35	0.48	0.78	0.74	1.00	1.00	0.78	1.00	0.56	0.56	0.52		**1.00**	**1.00**	**0.94**	**0.51**
PCA RIL1	**c**	**0.59**	**0.90**	**0.78**	**1.00**	**0.84**	0.62	**0.56**	**0.50**	**0.69**	**0.69**	**1.00**	**1.00**	**1.00**	**0.69**	**0.89**	**0.69**	**0.82**	**0.86**	**1.00**		**1.00**	**0.76**	**0.60**
PCA RIL2		**0.70**	**0.77**	**1.00**	**1.00**	**0.89**	**0.91**	**0.64**	**0.81**	**0.83**	**0.90**	**1.00**	**1.00**	**1.00**	**0.90**	**1.00**	**0.83**	**0.83**	**0.80**	**1.00**	**1.00**		**0.76**	**0.27**
genetic RIL1		**0.59**	**0.84**	**0.91**	**0.58**	0.55	0.56	**0.38**	**0.37**	0.31	0.59	**1.00**	**1.00**	**0.98**	0.56	**0.76**	**0.55**	0.29	**0.59**	**0.91**	**0.82**	**0.90**		0.11
genetic RIL2		**0.53**	**0.70**	**0.92**	-0.22	-0.02	0.16	0.14	0.26	-0.10	0.20	**0.79**	**0.86**	**0.56**	0.26	**0.51**	0.05	-0.20	0.15	**0.52**	**0.63**	**0.34**	-0.01	

Transcriptome distance between GM lines and the wild type background

PCA can visualize the relationship between samples. However, it is only possible to judge similarity or differences between samples based on three PCs (for our study explaining ~20% of the variation). We used the approach of Houshyani *et al.* (2011), which was used to calculate metabolic *hyper-plane distance*, to calculate transcriptome *hyper-plane distance* using the sample scores on the first 9 PCs. For *hyper-plane distance* calculation, the microarray signals representing *FPS1L* and *HMGR1S* were removed from the dataset. Col-3 and transgenic lines' sample scores on the first 9 PCs of a PCA with *meta data* was used for calculation of the *transcriptome distance* between Col-3 plants and COX+ and COX++ lines (Table 1a). Using either intact or weighted scores, the Col-3 *transcriptome distance* to COX+ (0.19 and 0.26, respectively) and COX++ (-0.07 and 0.15) were not significant (permutation test, P-value=<0.05). The set of independently transformed COX+ and COX++ lines showed the maximum difference to each other (1.0, 1.0) although still not significant (Table 1a). This indicates the absence of any similarity between the COX+ and COX++ GM plants in the response to transformation.

Transcriptome difference assessment

To assess the distances between transgenic lines and Col-3, we evaluated them in the context of natural variation (different accessions) or conventional breeding practices (RILs). For this purpose we used published transcriptome data of different accessions and RILs (the *meta data*) and analysed the differences and distances between accessions and groups of RILs. For transcriptome analysis the RIL population was divided into two groups (GPs) based on molecular marker differences (Genetic GPs) and based on transcriptional expression differences (Expression GPs). ANOVA on the accessions, Genetic GPs and Expression GPs and Venn diagram showed that 137 of 174 and 85 of 98 differentially expressed genes (> 2 fold) in COX+ and COX++ vs. Col-3, respectively, were also significantly > 2 fold different in at least one of 3 selected experiments from the public database (where statistical analysis was possible) and/or between Genetic GPs and/or Expression GPs. This indicates that a considerable number of differences between the COX+ or COX++ lines and the Col-3 also occur naturally and are not specific to the genetic modification. For interpretation of the specific changes, the reader is kindly referred to the ORA results section.

Subsequently, the *transcriptome distance* between all groups of the *meta data* was calculated and tested for statistical significance (Table 1b and c). Within the accessions of the public database, only two accessions (Col-0 in study 18 and Cal-0 in study 5728) had a significant *transcriptome distance* to each other (permutation test, P-value=<0.05) with both

intact (0.48) and weighted (0.48) scores (Table 1b). Although the rest of the pairwise distances were not statistically significant, strong variation was observed, particularly in study number 18. *Transcriptome distances* based on weighted scores in this study were varying from -0.11 between Bay and Estland to 1.00 (the maximum) between several pairs such as Col-0 and Cvi. Using intact scores, Bay and Estland showed a distance of 0.04 which was one of the smallest distances after Niederzenz and Ler (-0.04). Ler is an accession of *Arabidopsis* widely used for both molecular and genetic studies (Torii *et al.* 1996). It was isolated from a mutagenized seed population and harbours the *erecta (er)* mutation that causes strong phenotypes such as altered organ shape, compact inflorescence with flowers clustering at the top and round leaves with short petioles and short and blunt siliques (Torii, Mitsukawa, Oosumi, Matsuura, Yokoyama, Whittier and Komeda 1996). A look at the Ler *transcriptome distance* by weighted scores and comparing with the rest of the accessions in study 18 revealed a distance range between 0.04 with Niederzenz and 0.78 with Col-0. There are several other pairs of accessions with the maximum distance (1.0) in the same study (such as Cvi and Col-0) (Table 1b).

The Genetic GPs and Expression GPs comprised RILs that were grouped based on genetic or expression profile similarity, respectively. As every individual of a RIL population represents randomly 50% of each parent's genome, it was impossible to choose two groups that are genetically the most distant groups. Therefore, AFLP molecular marker data of the RIL population were used in a PCA analysis to select the two genetically most distant groups (Genetic GPs). Two other groups (Expression GPs) were selected on the first three PCs of a PCA plot performed on expression data of the RIL population.

Transcriptome distances between Genetic GPs were 0.11 and -0.01 using intact and weighted scores, respectively (Table 1c). The *transcriptome distance* between Expression GPs were significant and 1.0 (the maximum) for both intact and weighted scores (permutation test, P-value=<0.05) (Table 1C). There are also significant distances between Genetic GPs and Expression GPs ranging from 0.27 to 0.76 and 0.34 to 0.90 for intact and weighted PCA scores, respectively (permutation test, P-value=<0.05) (Table 1c).

Discussion

Pleiotropic transcriptional effects in the GM plants are smaller than pleiotropic variation in nature

Here we have used different methods (uni-variate statistics, GSEA, ORA, multivariate data analysis and *hyper-plane distance*) to detect, interpret and assess the unintended changes in the transcriptome of transgenic lines compared with Col-3. To assess the global differences, the *hyper-plane distance* method as previously described (Houshyani *et al*, 2011) was put into context by comparing the Col-3 *vs.* transgenic lines *transcriptome distance* with that between different accessions and groups of a RIL population. Results show that the largest *transcriptome distances* and statistically significant differences are found between the two groups of the RIL population (Table 1). This demonstrates that the cross of two different parental lines, resulting in a population of individual offspring plants of mixed genome composition, has a larger pleiotropic effect on gene transcriptional activity than introduction of two or three genes. Also the global transcriptome differences between the individual accessions of *Arabidopsis* that can be found in the nature are larger than the transcriptional differences between the two GM lines and the wild type Arabidopsis Col-3.

The difference in transcriptome activity of the different *Arabidopsis* accessions represents the genetic variation derived from differences in evolutionary genetic drift between accessions (Fig. 2a). This variation still results in very similar phenotypes of the different accessions at a macroscopic level. Presumably, this is due to the multiple levels of feedback regulation that occurs at the transcript, protein and metabolite levels, a phenomenon referred to as phenotypic buffering (Fu *et al.* 2009). In contrast, the macroscopic phenotypic differences between members of the RIL population (Fig. 2b) are much larger and can exceed those of the original parental lines. For instance, individual members of two RIL populations (Ler/Cvi and Ler/Col-0) showed extensive variation in clock period and phase, while the parental lines were similar (Swarup *et al.* 1999). With the same Cvi/Ler population it was also shown that the balanced gene expression pattern in the parents became transgressive among the segregants of the RIL population as 15% of the genes for which the parents did not show a significant difference in expression levels had an eQTL in the population. Also, much lower heritabilities were calculated from the parental data than with those from the RIL population (Keurentjes *et al.* 2006). Our results on the transcriptional distances are in agreement with results obtained from metabolomics: analysis of different masses in the same RIL population showed that 40% of the detected masses were specific to the RIL individuals and were not detected in the parental lines (Keurentjes *et al.* 2006).

As the variation between accessions and within RILs are considered to be of 'natural' origin (in contrast to the genetic modifications in GM plants), the transcriptional distances in that germplasm can be considered as a *baseline*. When the pleiotropic effects on gene expression in GM plants are evaluated in the context of this *baseline*, the results show that pleiotropic global transcriptome changes in the two GM lines fall well within the range of transcriptome distances that occur in nature. As such, the GM lines may therefore be called substantially equivalent to one or some of the naturally occurring Arabidopsis plants.

Natural variations covers large part of the transcriptome variation

Although the quantity of intended and unintended changes in gene expression are minor compared with the number of analysed genes, there are still many unintended changes in gene activity in the two GM lines. Without any *p-value* adjustment for multiple tests or filtering based on fold change in the expression, 2.38% and 2.12% of the genes were differentially expressed in COX+ and COX++ lines, respectively, compared with the wild type. The use of two fold as filter, decreased these to 0.75% and 0.42% for COX+ and COX++ lines, respectively. Similar levels of transcriptome changes were observed by introduction of a marker gene only (El Ouakfaoui and Miki 2005).

The specificity of the transcriptional changes was evaluated by GSEA and ORA, which did not result in any consistent change. Moreover, for a considerable number of the genes that showed differences between the GM and WT plants, similar differences were identified in natural plant populations or a plant population constructed through classical breeding. The fact that a similar difference in gene activity can be found in natural populations in itself does not say anything about the specific impact of such difference in gene activity, which could still be a matter of concern in relation to evaluation of GM plants. However, it does show that any concern about pleiotropic effects on gene activity is more relevant for the output of a classical breeding strategy than in GM plants.

Different cause of pleiotropic changes in transcriptome of GM plants

Quantitative transcriptome comparison revealed different patterns of up-regulated and down-regulated genes in COX+ and COX++, with only three down-regulated genes (> 2 fold) in common in both lines. The more global analysis by GSEA or ORA did not yield any commonly changed gene set in COX+ and COX++. These findings show that the two strategies that were used to create the same (insect resistance) trait have a different impact on the transcriptome. The pleiotropic changes in the COX+ and COX++ transgenic lines seem to

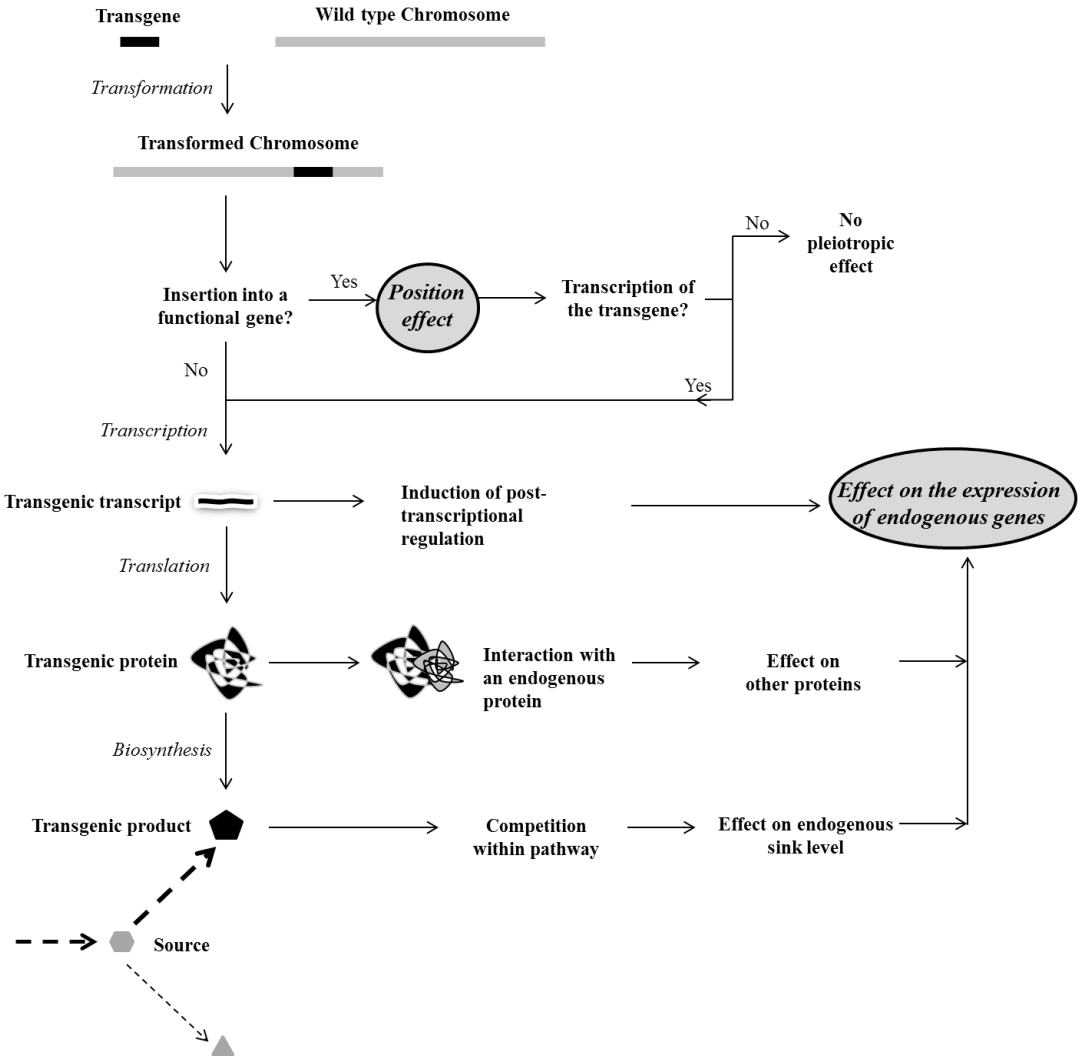


Figure 5 A schematic overview of the possible position and pleiotropic effects of inserted construct and/or gene(s)

be non-specific to the novel trait, as these lines cluster separately from each other (Fig. 4). This also suggests a consistent change in the transcriptome of the biological replicates (within each set), which therefore seems to be related to the effect of the introduced gene(s). The deviation from a tight clustering of sets could be the result of additional pleiotropic effect that are independent of the introduced gene(s), and may be the result of pleiotropic effects of

different genomic insertion sites (position effect) (Fig. 5). Alternatively, deviation from tight clustering of biological replicates can be the result of micro-environmental differences within the plant population, in which case they are not specific to our modification and not reproducible.

Environmental effects increase baseline variation and further reduce significance of transcriptional distances

Transcription is a polymorphic trait that is under the control of many genetic, epigenetic and environmental factors (Rockman and Kruglyak 2006, Zhang and Borevitz 2009) and natural polymorphism among *Arabidopsis* accessions has been already reported to be considerable at the genome (Platt *et al.* 2010) as well as transcriptome levels (Zhang *et al.* 2008). We note that our *baseline* was constructed using public data of samples of plants that were grown under strict environmental control and samples were harvested at similar developmental stages. In real agricultural practice with a higher variability in environmental conditions (Barros *et al*, 2010, Houshyani *et al*, 2011), the *baseline* variation would increase and hence the *transcriptome distances* between two sets of genetically different plants would show reduced size as well as significance. Therefore, when environmental effects on transcription would be included, the GM lines most likely would show even higher substantial equivalence to natural occurring genotypes than what they shoe here.

Conclusions

The current study explored and assessed the changes in the transcriptome of genetically engineered lines of *Arabidopsis* using a holistic *omics* analytical and statistical approach and comparison to a natural transcriptome *baseline*. Results show that the pleiotropic (and sometime unintended) changes in the transcriptome of GM plants are non-significant when compared to natural occurring genotypes and that transcriptional distances between GM and untransformed WT are much smaller than the transcriptional distances occurring within the *baseline* group (accessions, RILS). Combined, the results show that critical arguments against GM plants, based on fear of unknown pleiotropic effects on gene activity are more relevant for classical breeding than for GM approaches.

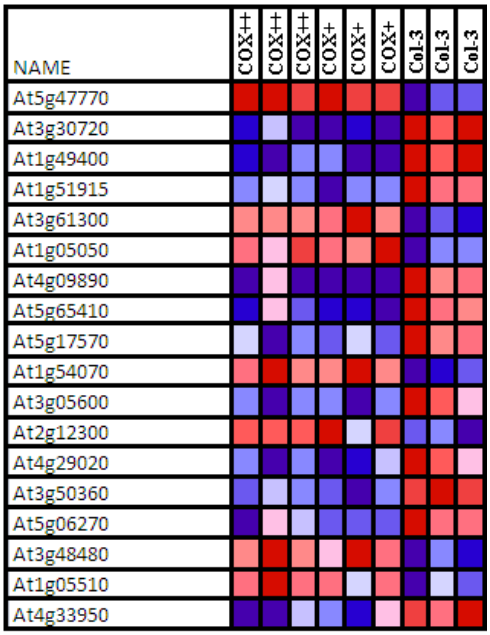
This case study contributes to the ERGO program (http://www.nwo.nl/nwohome.nsf/pages/NWOA_6N4LKX_Eng) that in line with the guidance on the environmental risk assessment of genetically modified plants (EFSA 2010)

aims to increase public knowledge about novel crops and develop a multidisciplinary approach for assessment of crops generated using modern biotechnology. Results of this program will be also transferred to authorities in The Netherland who are responsible for regulation considering genetically modified crops.

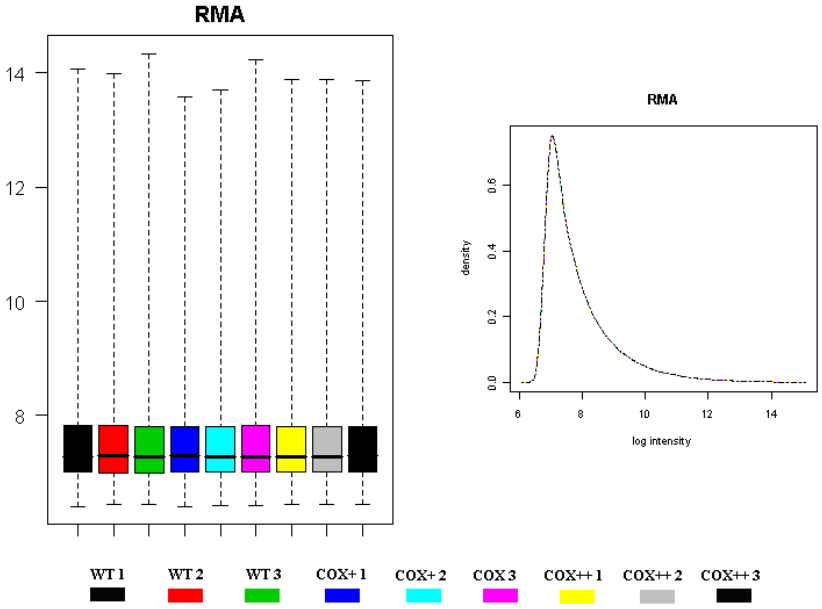
Supporting Information

Supplementary Table 1 Primers for quantitative RT-PCR analysis

Gene Name	Sequence (5' to 3')
FaNES1	
Fragaria ananassa	F: CGACTACTGAGACAACATGGTTAC
Nerolidol Synthase 1	R: TCCTCACCCAGAACTTGCTTG
FPS1L	
AT5G47770	F: ATGAGTGTGAGTTGTTGTTGTAGG
long isoform	R: GCTTTGGATACGACGACGATAG
HMGR	
AT1G76490	F: AAGGGTTACCGTTGGATGGATTTG
	R: CAAGCAACAATGGACCAGCAATC
β-Tubulin	
AtxG	F: CTCAAGAGGTTCTCAGCAGTA
	R: TCACCTTCTTCATCCGCAGTT



Supplementary Figure 1 Heat map of the gene identities with more than 90% positive or negative correlation with the expression of *FPS1* over-expressed in COX+ and COX++ transgenic lines.



Supplementary Figure 2 Boxplots and smoothed histograms for quality check of the microarray data shows the distribution of the log2-transformed intensities after RMA normalization viewed by boxplots and smoothed histograms.

References

Baker, J.M., Hawkins, N.D., Ward, J.L., Lovegrove, A., Napier, J.A., Shewry, P.R. and Beale, M.H. (2006) A metabolomic study of substantial equivalence of field-grown genetically modified wheat. *Plant Biotechnology Journal*, **4**, 381-392.

Baudo, M.M., Lyons, R., Powers, S., Pastori, G.M., Edwards, K.J., Holdsworth, M.J. and Shewry, P.R. (2006) Transgenesis has less impact on the transcriptome of wheat grain than conventional breeding. *Plant Biotechnology Journal*, **4**, 369-380.

Catchpole, G.S., Beckmann, M., Enot, D.P., Mondhe, M., Zywicki, B., Taylor, J., Hardy, N., Smith, A., King, R.D., Kell, D.B., Fiehn, O. and Draper, J. (2005) Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 14458-14462.

Chen, F., Tholl, D., D'Auria, J.C., Farooq, A., Pichersky, E. and Gershenzon, J. (2003) Biosynthesis and emission of terpenoid volatiles from Arabidopsis flowers. *Plant Cell*, **15**, 481-494.

Cheng, K.C., Beaulieu, J., Iquira, E., Belzile, F.J., Fortin, M.G. and StrÃ¶mvik, M.V. (2008) Effect of transgenes on global gene expression in soybean is within the natural range of variation of conventional cultivars. *Journal of Agricultural and Food Chemistry*, **56**, 3057-3067.

Closa, M., VranovÃ¡, E., Bortolotti, C., Bigler, L., ArrÃ³, M., Ferrer, A. and Gruissem, W. (2010) The Arabidopsis thaliana FPP synthase isozymes have overlapping and specific functions in isoprenoid biosynthesis, and complete loss of FPP synthase activity causes early developmental arrest. *Plant Journal*, **63**, 512-525.

Coll, A., Nadal, A., Collado, R., Capellades, G., Kubista, M., Messeguer, J. and Pla, M. (2010) Natural variation explains most transcriptomic changes among maize plants of MON810 and comparable non-GM varieties subjected to two N-fertilization farming practices. *Plant Molecular Biology*, **73**, 349-362.

EFSA (2010) Guidance on the environmental risk assessment of genetically modified plants plants. *EFSA Journal*, **8**.

El Ouakfaoui, S. and Miki, B. (2005) The stability of the Arabidopsis transcriptome in transgenic plants expressing the marker genes nptII and uidA. *Plant Journal*, **41**, 791-800.

Fu, J., Keurentjes, J.J.B., Bouwmeester, H., America, T., Verstappen, F.W.A., Ward, J.L., Beale, M.H., De Vos, R.C.H., Dijkstra, M., Scheltema, R.A., Johannes, F., Koornneef, M., Vreugdenhil, D., Breitling, R. and Jansen, R.C. (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nature Genetics*, **41**, 166-167.

Groot, E.P. and van der Graaff, E. (2009) *Arabidopsis thaliana* gene set database version 1.1 *http://omnibus.uni-freiburg.de/~eg470/ath_gsdb.html*.

Harrigan, G.G., Martino-Catt, S. and Glenn, K.C. (2007) Metabolomics, metabolic diversity and genetic variation in crops. *Metabolomics*, **3**, 259-272.

Houshyani B, Kabouw P, Muth D, de Vos RCH, Bino RJ, Bouwmeester HJ. 2011**.** Characterization of the natural variation in *Arabidopsis thaliana* metabolome by the analysis of metabolic distance. *Metabolomics **In press***. **ISSN:** 1573-3882

Houshyani B, Assareh M, Busquets A, Ferrer A, Bouwmeester HJ, Kappers I. 2012. Three-step pathway engineering results in more incidence rate and higher emission of nerolidol and improved attraction of Diadegma semiclausum. Metabolic Engineering ***Submitted***

Kappers, I.F., Aharoni, A., Van Herpen, T.W.J.M., Luckerhoff, L.L.P., Dicke, M. and Bouwmeester, H.J. (2005) Plant science: Genetic engineering of terpenoid metabolism attracts bodyguards to Arabidopsis. *Science*, **309**, 2070-2072.

Keller, A., Backes, C., Al-Awadhi, M., Gerasch, A., KÃ¼ntzer, J., Kohlbacher, O., Kaufmann, M. and Lenhof, H.P. (2008) GeneTrailExpress: A web-based pipeline for the statistical evaluation of microarray experiments. *BMC Bioinformatics*, **9**.

Keurentjes, J.J.B., Fu, J., De Vos, C.H.R., Lommen, A., Hall, R.D., Bino, R.J., Van Der Plas, L.H.W., Jansen, R.C., Vreugdenhil, D. and Koornneef, M. (2006) The genetics of plant metabolism. *Nature Genetics*, **38**, 842-849.

Keurentjes, J.J.B., Fu, J., Terpstra, I.R., Garcia, J.M., Van Den Ackerveken, G., Snoek, L.B., Peeters, A.J.M., Vreugdenhil, D., Koornneef, M. and Jansen, R.C. (2007) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 1708-1713.

Kuiper, H.A., Kleter, G.A., Noteborn, H.P.J.M. and Kok, E.J. (2001) Assessment of the food safety issues related to genetically modified foods. *Plant Journal*, **27**, 503-528.

Leivar, P., GonzÃ¡lez, V.M., Castel, S., Trelease, R.N., LÃ³pez-Iglesias, C., ArrÃ³, M., Boronat, A., Campos, N., Ferrer, A. and FernÃ ndez-Busquets, X. (2005) Subcellular localization of arabidopsis 3-hydroxy-3-methylglutaryl-coenzyme A reductase. *Plant Physiology*, **137**, 57-69.

Manzano, D., Busquets, A., Closa, M., Hoyerová, K., Schaller, H., Kamínek, M., Arró, M. and Ferrer, A. (2006) Overexpression of farnesyl diphosphate synthase in Arabidopsis mitochondria triggers light-dependent lesion formation and alters cytokinin homeostasis. *Plant Molecular Biology*, **61**, 195-213.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., RidderstrÃ¥le, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D. and Groop, L.C. (2003a) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34**, 267-273.

OECD (2010) Consensus Document on Molecular Characterisation of Plants Derived from Modern Biotechnology. *Series on Harmonisation of Regulatory Oversight in Biotechnology*, **51**.

Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Ã…gren, J., Bossdorf, O., Byers, D., Donohue, K., Dunning, M., Holub, E.B., Hudson, A., Le Corre, V., Loudet, O., Roux, F., Warthmann, N., Weigel, D., Rivero, L., Scholl, R., Nordborg, M., Bergelson, J. and Borevitz, J.O. (2010) The scale of population structure in Arabidopsis thaliana. *PLoS Genetics*, **6**.

Redman, J.C., Haas, B.J., Tanimoto, G. and Town, C.D. (2004) Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. *Plant Journal*, **38**, 545-561.

Rockman, M.V. and Kruglyak, L. (2006) Genetics of global gene expression. *Nat. Rev. Gen.*, **7**, 862-872.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550.

Swarup, K., Alonso-Blanco, C., Lynn, J.R., Michaels, S.D., Amasino, R.M., Koornneef, M. and Millar, A.J. (1999) Natural allelic variation identifies new genes in the Arabidopsis circadian system. *Plant Journal*, **20**, 67-77.

Torii, K.U., Mitsukawa, N., Oosumi, T., Matsuura, Y., Yokoyama, R., Whittier, R.F. and Komeda, Y. (1996) The arabidopsis ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. *Plant Cell*, **8**, 735-746.

Zhang, X. and Borevitz, J.O. (2009) Global analysis of allele-specific expression in Arabidopsis thaliana. *Genetics*, **182**, 943-954.

Zhang, X., Byrnes, J.K., Gal, T.S., Li, W.H. and Borevitz, J.O. (2008) Whole genome transcriptome polymorphisms in Arabidopsis thaliana. *Genome Biology*, **9**.

Chapter 7

Whole-metabolome difference assessment of GM *Arabidopsis* lines with three novel insect defence traits using natural metabolome variation as a reference

Benyamin Houshyani · Patrick Kabouw · Harro J. Bouwmeester

In preparation for submission

Abstract

The metabolome of plants is highly dynamic and is influenced by the genotype as well as the biotic and abiotic environment. This natural metabolome variation can be considered as the baseline variation to which changes in the metabolome upon genetic modification (GM) could be compared for substantial equivalence.

Here we used a holistic approach, metabolomics, to explore the root and shoot metabolome of *Arabidopsis thaliana* genotypes including four wild type accessions and 18 lines with 3 GM traits in 3 backgrounds. The wild type accessions were metabolically diverged and their metabolome collectively formed the baseline metabolome as a reference for comparison of the GM lines. The metabolome data were used to calculate the "metabolic distance" of the GM lines to the baseline as a statistical approach to quantify the change in the metabolome of a GM line compared with the natural metabolome variation.

Our analysis shows that some of the transformation events result in lines that are substantially equivalent to the baseline. In other words, these lines did not show any significant metabolic distance to the baseline in all of the metabolomics analyses applied. The other transformation events resulted in lines that showed a significant metabolic distance to the baseline in at least one of the analyses on shoot or root. We discuss that this significant difference likely disappears when the effect of the environment on the baseline is also included in the analysis. We propose that metabolic distance could be used as a first-pass criterion in ecological risk assessment of GM crops.

Introduction

Genetically modified crops are the fastest growing biotechnology products in the history of modern agriculture recording an 87-fold increase in acreage in 2010 in their 15th anniversary year after commercialization (James 2010). New developments in plant biotechnology will result in novel biotech crops with novel features (such as drought tolerance, pathogen or insect resistance, and nutritionally enriched and industrially more efficient crops) that may no longer just address producer problems but also the interests and concerns of the society. The latter is achieved in a number of ways such as exploiting less controversial sources for transgenes in the cisgenic approach (Schouten *et al.* 2006), the omission of anti-biotics selection markers and temporal or spatial control of transgene expression (Conner *et al.* 2007; Hensel *et al.* 2011). This progress slowly results in the view that biotech crops should be considered as vital components of integrated pest management for environmental-friendly control of pests (Kos *et al.* 2009; Lundgren *et al.* 2009). However, the concerns of society about application of biotech products require the full attention of all involved to prevent future complications around this issue and support a sustainable development of GM applications.

Assessment of non-target effects of biotech crops on the environment, *i.e.* effects that were not intended by the transformation, is one of the requirements that has been suggested for the evaluation of the safety of GM crops (Craig *et al.* 2008; Hilbeck *et al.* 2011; Poppy 2000). Non-target environmental or ecological effects of biotech crops might include an effect on above- and below-ground non-target organisms, particularly when transformation enhances plant defence through direct or indirect mechanisms. However, a complete evaluation of all possible non-target effects of biotech crops on the biotic environment is likely not feasible due to its elaborate and time- and resource-consuming nature. Therefore, there is a need for methods that could be used for an efficient initial screen of GM lines before elaborate ecological studies are done. The proteome and metabolome of plants are known to mediate the interaction of plants with their living environment (Figueiredo *et al.* 2008; Kuśnierczyk *et al.* 2008; Leiss *et al.* 2009; Leiss *et al.* 2011; Macel *et al.* 2010; Kos *et al.* 2011a). Metabolites reliably reflect biological endpoints (observable or measurable phenotypes) (Harrigan *et al.* 2007) and the metabolite fingerprint can be used to select closely related genotypes that nevertheless segregate for interaction with multiple organism (Houshyani *et al.* 2011). Metabolite fingerprints have also been used for discovery of genes, quantitative trait loci and metabolites that control or correlate with the plant-biota interaction such as plant resistance and susceptibility (Ferrari *et al.* 2003; Figueiredo *et al.* 2008; Gatehouse 2002; Goodman *et al.*

2004; Kashif *et al.* 2009; Rohr *et al.* 2011; Rulhmann *et al.* 2002; Yang *et al.* 2011), which further corroborates the importance of the metabolome as a predictor of plant-biotic environment interaction. Hence, metabolite fingerprinting, depending on the percentage of metabolome coverage, offers opportunities for a 'first-pass' evaluation of compositional similarity between transgenic plants and their wild type counterpart (Enot and Draper 2007).

The assessment of non-target effects of genetic modification on the metabolome of a GM crop is commonly performed by comparing the GM line with the non-GM counterpart (Barros *et al.* 2010; Dubouzet *et al.* 2007). However, the metabolome is highly variable, even within a species, for example between ecotypes and varieties (Barros *et al.* 2010; Kusano *et al.* 2011; Houshyani *et al.* 2011). Furthermore, the growing conditions of a plant have a profound effect on the metabolome (Houshyani *et al.* 2011) and also affect its interactions with other organisms. Soil characteristics, for example, have a significant impact on plant quality and consequently on aboveground and belowground biota (Kabouw *et al.* 2010; van der Putten *et al.* 2001; Bruinsma *et al.* 2003; Pineda *et al.* 2010). This natural variation in the metabolome should be used to assess if changes in the metabolome of a GM crop remain within the boundaries of this natural variation, which we define as the (multi-dimensional) baseline. Metabolite fingerprinting of germplasm related to the transgenic crop will help in finding the most metabolically diverged genotypes that based on the above reasoning can then be used to establish the baseline not only for assessment of metabolome changes but also to be used as a reference for non-target ecological parameters such as plant-insect interaction in ecological risk assessment.

To be able to compare the metabolome of a GM crop with this baseline, a statistical approach to quantify the similarity between the metabolome of a genetically modified (GM) lines and an established baseline metabolome in the setting of high dimensional data is required. In an earlier study, we developed and applied a statistical method to assess the distance between genotypes or treatments, using complex metabolomics and transcriptomics data (Houshyani *et al.* 2011; Houshyani 2012b). This method is employed in the present study to assess the effects of genetic engineering on the metabolome of *Arabidopsis thaliana* in relation to a baseline.

As a model for genetic engineering, we used genetic modification of *A. thaliana* with three different traits related to direct and indirect insect defence. Direct defence traits in plants are chemical or physical means such as the production of toxins, repellents and digestibility reducers or the production of trichomes and epicuticular waxes to directly affect the performance or behaviour of herbivores (Vasconcelos and Oliveira 2004; Wittstock and Gershenzon 2002). In this study, direct defence of *A. thaliana* was enhanced by two approaches. The first approach was overexpression of the transcription factor MYB28

(Houshyani 2012b) to boost the biosynthesis of aliphatic glucosinolates. Glucosinolates and their hydrolysis products (isothiocyanates and nitriles) negatively affect the performance of a variety of generalist herbivores (Kelly *et al.* 1998). The effect on insect resistance and the presumed positive effect of glucosinolates on human health have triggered many attempts to engineer glucosinolates in cruciferous plants or even introduce it into non-cruciferous plants (Geu-Flores *et al.* 2009; Shukla and Mattoo 2009; Wielanek *et al.* 2009). The second strategy that was used to enhance the direct defence in *A. thaliana* was the introduction of the SN19 gene encoding cry1 protein from Bacillus thuringiensis (Bt toxin) (Naimov *et al.* 2001) which is commercially used for the control of herbivores in a number of crops in a number of countries.

If a herbivore is not stopped by any direct defence mechanism, the resulting damage to the plant induces the so-called indirect defence (Gatehouse 2002). This involves biosynthesis and emission of a blend of plant volatile organic compounds (VOCs) that attracts the attention of antagonists of the plant enemies (Dicke and Baldwin 2010). Indirect defence of *A. thaliana* was enhanced by introduction of a construct containing two genes responsible for sequential steps in sesquiterpene biosynthesis, one from strawberry (FaNES1, Fragaria X ananassa nerolidol synthase 1) and one from *A. thaliana* (AtFPS2, farnesyl diphosphate synthase 2). The introduced construct mediates the biosynthesis of (E)-nerolidol and DMNT which are attractants of natural enemies of herbivores (Aharoni *et al.* 2003; Kappers *et al.* 2005; Houshyani 2012a; Degenhardt *et al.* 2003; Degenhardt *et al.* 2009).

The modified pathways of this study (terpenoid and glucosinolate) are also model pathways for quantitative genetics; their molecular and biochemical aspects are elucidated to a large extent and their ecological role has been firmly established (Chan *et al.* 2010; Nagegowda 2010; Ober 2010; Tholl 2006; Chan *et al.* 2011) providing practical advantages of exploiting them in the development of new generations of GM crops.

To assess the effect of the genetic background, we introduced the direct and indirect defence traits into four accessions of *A. thaliana*, An-1, Col-0, Cvi and Eri that based on a previous study were demonstrated to be significantly metabolically divergent (Houshyani *et al.* 2011). The consequences of the genetic modification for biosynthesis of the transgenic products and/or the effect on target organisms were described in independent studies (Houshyani *et al.* 2012a; Houshyani 2012b; Kos *et al.* 2011a; Kos *et al.* 2011b). A whole metabolome fingerprinting approach was employed for the analysis of wild types and independent transgenic lines by two metabolomics platforms GC-TOF-MS and LC-TOF-MS in positive mode. Subsequently, the metabolome data were used for multivariate data analysis to produce the output required for calculating the metabolic distance (Houshyani *et al.* 2011) between all studied groups such as a GM accession or a transformation event of a GM

accession and the baseline group, composed of the wild type backgrounds. This approach provides a statistical tool to evaluate the substantial equivalence of the metabolome of GM plants given the metabolome baseline.

Materials and Methods

Constructs and transformation

Phusion enzyme (Finnzymes, Finland) was used for PCR when necessary. Standard cloning methods (restriction and ligation) were used for construction of all the plasmids and the vectors of final and intermediate stages which were checked by restriction analysis and sequencing.

CoxIV-FPS2 and CoxIV-FaNES1 single and double (2way) constructs: The cDNAs of AtFPS2 (At4g17190) and FaNES1 were cloned into two separate pGEM-T plasmids (Promega) (Cunillera *et al.* 1996; Kappers *et al.* 2005). The BamHI/NotI restricted fragments of FPS2 and FaNES1 were gel purified and ligated into the pIV2B_2.5 and pIV2A_2.5 entry vectors, respectively, containing a CaMV 35S promoter, CoxIV mitochondrial targeting sequence and an RbcS1 terminator (www.pri.wur.nl/UK/products/ImpactVector/). These entry clones were transferred to E. coli strain X1-Blue by heat shock for propagation. An LR reaction (Invitrogen Life Technologies) was used to recombine the sequences of interest between the right and left borders of the T-DNA in the pBINPLUS binary vector (Van Engelen *et al.* 1995). This resulted in the 2way plasmid containing the AtFPS2 and FaNES1 coding sequences each driven by the CaMV 35S promoter and equipped with a CoxIV mitochondrial targeting sequence and an RbcS1 terminator. This binary vector was introduced into X1-Blue by heat shock and to *Agrobacterium tumefaciens* AglI by electroporation.

MYB28 construct: The expression clone was kindly provided by Prof. Ulf-Ingo Flügge (University of Cologne, Germany). Briefly, the coding sequence of HAG1/MYB28 (At5g61420.2) was amplified by RT-PCR and cloned into the pDONOR207 vector. LR reaction (Invitrogen Life Technologies) between pDONOR207 and pGWB2 recombined the insert from the entry clone into the destination vector. The binary plant transformation vector pGWB2 contained the CaMV 35S promoter and kanamycin and hygromycin resistance genes (Gigolashvili *et al.* 2007). Subsequently, the expression clone was confirmed by digestion analysis (EcoRI + HindIII) and sequencing. A confirmed CaMV 35S::HAG1 clone was used to transfect *A. tumefaciens* strain Agl0.

SN19 construct: SN19 is encoded by a hybrid gene and consists of domain I and III of cry1Ba and domain II of cry1Ia (Naimov *et al.* 2001). The binary vector pTC12 (with pBINPLUS backbone) containing the domain I and III modified SN19 for optimized expression in planta was kindly provided by Dr Ruud de Maagd (Plant Research International, Wageningen University). Modifications took place on domain I and III by directed

mutagenesis. The positive effect of modification on protein biosynthesis and insect resistance has already been justified in the model plant *A. thaliana* (data not published). pTC12 was used to transfect *A. tumefaciens* strain Agl0 after restriction analysis and confirmation. The introduced construct mediated mortality of Pieris rapae second instar larvae feeding on the generated transgenic lines of *A. thaliana* (data not shown).

All plant transformations were performed using the *A. thaliana* flower dip method (Zhang *et al.* 2006). T1 seeds were harvested and T1 transgenic plants were selected on medium with kanamycin (50 µg ml^{-1}) and confirmed by kanamycin resistance (NptII) specific primers. T2 generation seeds of positive lines were harvested and used in the experiments.

Choice of A. thaliana accessions and growing conditions

Four accessions of *A. thaliana*, An-1, Col-0, Eri and Cvi, that represented the natural genetic metabolome variation of a group of 9 accessions in a previous study (Houshyani *et al.* 2011) were selected for transformation. *A. thaliana* seeds (wild type and T2 generation of 2 transformation events of each construct in An-1, Col-0 and Eri background) were surface-sterilized overnight by chlorine gas (3 mL of 37% HCl in 100 ml bleach) in an airtight desiccator and spread on the growth medium in agar plates (purified agar 0.8% + 2.2 g L^{-1} 0.5 MS + vitamins + 50 µg ml^{-1} kanamycin only in selection medium; pH 6). After four days of stratification at 4 °C, plates were transferred to a growth chamber at 21 ± 2 °C, 50-70% relative humidity (RH) and a 16 h:8 h light:dark (L:D) photoperiod with a light intensity of ± 120 µmol m^{-2} s^{-1} photosynthetic photon flux density (PPFD). Two-week old seedlings with two true leaves were transplanted to rock wool in a hydroponics system with 0.5 Hoagland's solution, 10 h:14 h L:D to delay flowering and otherwise the same conditions as described above. Hydroponics solution was renewed once a week. Shoot and root materials of individual plants were harvested 4 weeks after transfer to hydroponics, flash frozen in liquid N2 and lyophilized for 48 h before extraction for chemical analysis.

Lyophilized shoot and root samples of Eri, Cvi, Col-0 and Eri that were grown in 2009 under the same conditions (hydroponics, climate room) and were analysed before (Houshyani *et al.* 2011) (and had been stored at 4°C since then) were included in the present study in all of the analyses.

Extraction for chemical analysis

The lyophilized plant material was ground and ± 50 mg of shoot and ± 10 mg of root material was weighed for extraction using a protocol derived from two protocols for GC-TOF-MS and LC-TOF-MS analysis (Houshyani *et al.* 2011). Briefly, for LC-MS analysis shoot and

root samples were mixed with 2 and 1 ml of the extraction solution, respectively (75% v/v methanol, 0.1% v/v FA, 2.65 x 10^{-5}% w/v leucine enkephalin, 9 x 10^{-4}% w/v ribitol), vortexed, sonicated (15 min) and centrifuged (15 min). Of the supernatant 150 µl was filtered (Minisart SRP 4, 0.45 µm, Sartorius Stedim Biotech) and collected in a plastic vial with insert (300 µl) and screw neck for the analysis.

For GC-TOF-MS, 350 µl of H_2O (to reach a 40% methanol/water ratio) and 525 µl of $CHCl_3$ were added to 400 µl of the LC-MS extract, which was vortexed and centrifuged (5 min). Of the supernatant 50 µl (shoot) or 200 µl (root) was transferred to a glass insert in a glass vial, dried in a speed vacuum O/N and capped in the presence of argon gas for the analysis.

A mixed sample of wild type and transgenic plants was passed through the same procedure for injection as every 20th sample in the loading sequence of both analyses as technical replicates.

GC-TOF-MS analysis

Samples were grouped in sample blocks, each containing one randomly selected biological replicate of every accession. The instrument and protocol described in Houshyani *et al.* (2011) were used for derivatization and data acquisition. On-line derivatized extracts (25 µl) were injected (2 µl) with an Optic3 injector (ATAS) at 70^{o}C with a gradient of 6^{o}C.s^{-1} to 240^{o}C. A split flow of 10 (1 ml:11 ml) was used with a column flow of 2 ml.min^{-1} in a GC6890N gas chromatograph (Agilent Technologies) on a FactorFour capillary column (VF-5ms, 30 m x 0.32 mm i.d., 0.25 µm DF + 10 m EZ-Guard; Varian). The column temperature was 70^{o}C for 2 min with a gradient of 10^{o}C.min^{-1} to 310^{o}C and a final time of 3 min. The GC was coupled to a Pegasus III time-of-flight mass spectrometer (LECO) and compounds were detected at a scanning rate of 20 spectra per second (m/z 50-600).

LC-TOF-MS analysis

Liquid chromatography was performed on a Waters Acquity Ultra Performance Liquid Chromatography system (Waters, Milford, MA, USA). Ten µl of extract was injected automatically on an Acquity UPLC BEH C18 column (150 x 2.1 mm i.d., 1.7 µm particle size) (Waters), held at 50°C with a mobile phase flow of 0.4 ml.min-1. The mobile phase consisted of water (A) and acetonitrile (B) containing 20 mM formic acid. The gradient applied started at 90% A / 10% B for 1.5 min and was subsequentely changed linearly to 25% B in 4 min, to

65% in 3.5 min and to 95% in 5 min, which was held for 6 min. Before the next run the column was equilibrated with starting conditions for 3 minutes.

Compounds eluting from the column were detected by a Bruker micrOToF-Q MS (Bruker Daltonik, Bremen, Germany) equipped with an electrospray ionization (ESI) source. The analysis was performed in positive ion mode in the range of 80 to 1000 m/z, using a scan time of 500 ms. The parameters of the source were: drying gas temperature of 200°C, drying gas flow of 8 L min-1, nebulizer gas pressure of 2.0 bar, capillary spray voltage of 3.5 kV, transfer time of 75 µs. The mass spectrometer was tuned and calibrated before analysis. Furthermore, at the beginning of each chromatographic run, an internal mass calibration solution containing 5 mM sodium formate/acetate mixture was added via an injection port. A set of 30 reference masses of selected formate and acetate cluster ions were used to for mass correction. Bruker MicrOTOF control software was used for data collection and instrument control and Compass Hystar Post Processing software for data analysis.

Data analysis

The 4 metabolome data sets (2 analysis techniques x 2 tissues) were processed and analysed and the metabolic distance between genotypes were calculated according to the method described in Houshyani *et al.* (2011) with slight modifications in data preparation.

Datasets were processed independently using MetAlign (Lommen and Kools 2011) for peak detection and alignment of the data points. An in-house script called MetAlign Output Transformer (METOT; Plant Research International, Wageningen) was used for data filtration, missing value replacement, and data quality and analytical technique reproducibility verification. The post-METOT data matrix was subjected to multivariate mass spectra reconstruction (MMSR) for data size reduction and putative compound mass spectrum reconstitution (Tikunov *et al.* 2005).

For multivariate data analysis, the intensity values of reconstituted metabolites were normalized by the dry weight of the sample. Subsequently, metabolite intensities of each sample were scaled by the standard deviation. An integrated dataset containing the analytical data of root and shoot by GC-TOF-MS and LC-TOF-MS positive analyses was made for a combined assessment. Before PCA or RDA, the data were symmetrically scaled by the standard deviation.

The ordination diagrams in CANOCO (ter Braak 1988) were used to measure and visualize the variation in the sample profiles. In addition to the first two ordinates (or PCs), the third and fourth ordinates were incorporated into the analysis if their P-value > 0.05 (Monte Carlo Permutation test). Partitioning of explanatory variable effects such as accession,

environment and their interaction on the observed variation were performed and tested statistically (P-value < 0.05) by Monte-Carlo Permutation test using the Partial Redundancy Analysis (RDA) function of CANOCO (Smilauer 2003).

To compute the metabolic distances, the inter-sample Euclidean distances in an ordination diagram were examined by taking the sample scores on the selected ordinates of the PCA scores plots (Kabouw *et al.* 2010). The inter-sample Euclidean distance matrices were computed for all tissues/platforms. The resulting matrices were used in an ANOSIM (analysis of similarity) by the program PAST (Hammer *et al.* 2001) to calculate the R-values as a measure for the metabolic distance between groups (baseline and GM groups or transformation events).

Results

Effect of the sources of variation

The analyses on GC-MS and LC-MS positive of shoot and root samples resulted in 4 datasets, in all of which the effect of 4 sources of variation could be estimated: 1) technical variation, 2) variation caused by the growing year, 3) variation across wild type accessions, the baseline and 4) variation caused by the introduction of the new traits and transformation *i.e.*, introduction of 2way, MYB28 or SN19 constructs. The effect of every source of variation was estimated and compared between and within datasets using dry weight normalized data of all samples and RDA (Table 1).

Technical variation

The GC-MS and LC-MS analyses were carried out during 5 and 3 uninterrupted days, respectively and for each plant material separately. The long duration of the analyses could possibly cause variation with non-biological origin *i.e.* technical variation. To check for this we followed the procedure of Houshyani *et al.* (2011) using scatter plot to check for hypo-alignment and PCA approach to control the clustering of technical replicates along PCs. There was no visible difference in the position of the technical replicates along the first 4 PCs of a PCA, suggesting absence of strong technical variation in the analyses. When using the intensity of the internal standard present in all the samples as explanatory variable and experimental year and accessions as cofactors (variables of which effects are counted and eliminated from the dataset) in an RDA, the internal standard explained less than 1% of the remaining variation in all datasets. Because of the low technical variation, technical replicates were excluded from further data analysis but peak intensity of the internal standard was used as a cofactor to correct for technical variation, despite its relatively small effect.

Growing year

To have an estimate about the potential effect of different growing years on the metabolome, we had included samples from a previous study that were collected under almost the same growing conditions. As the previous experiment did not include GM samples, we eliminated the GM samples of the present experiment from the dataset and performed RDA on the remaining data with experimental year as explanatory variable and internal standard intensity and accession as cofactors. The difference between wild type samples of the present

Table 1 Explained variance by individual sources of variation as a percentage of the total variance in the dataset, measured by RDA. Numbers in parenthesis are the average explainable variance by the groups that formed the corresponding source of the variation in a dataset. Trait + Transformation and Trait, samples of all transformation events of all backgrounds

Dataset	wild type	GM (2way) + wild type			GM (MYB28) + wild type			GM (SN19) + wild type		
Source	Year	Wild type	Trait + Transformation	Trait	Wild type	Trait + Transformation	Trait	Wild type	Trait + Transformation	Trait
GC-MS Shoot	29.1	7.1 (1.77)	3.3 (0.55)	2 (0.33)	5.4 (1.35)	7.2 (1.2)	2.2 (0.36)	5.2 (1.3)	4.3 (0.71)	3.5 (0.58)
GC-MS Root	22.5	2.8 (0.7)	19.2 (3.2)	7.5 (1.25)	3.1 (0.77)	7.6 (1.26)	4.2 (0.7)	1.6 (0.4)	6.3 (1.05)	4.6 (0.76)
LC-MS positive Shoot	28.6	9.1 (2.27)	14 (2.33)	8.9 (1.48)	6.9 (1.72)	4.7 (0.78)	1.7 (0.28)	10.2 (2.55)	7.1 (1.18)	4.2 (0.7)
LC-MS positive Root	33.6	26.3 (6.57)	22.5 (3.75)	13.4 (2.23)	8.6 (2.15)	12.9 (2.16)	1.5 (0.25)	5.4 (1.35)	4.4 (0.73)	1.8 (0.3)
Integrated data	N.A.	12.1 (3.02)	0.39 (0.06)	0.26 (0.04)	12.0 (3.00)	0.5 (0.8)	0.1 (0.02)	11.1 (2.77)	1.3 (0.22)	1.1 (0.18)

experiment and the previous experiment accounted for a large part of the metabolome variation (Table 1) varying between 22.5% (in the GC-MS root dataset) to 33.6% (the LC-MS shoot dataset) of the total variance in the corresponding dataset of wild type samples.

Variation across accessions

The metabolome variation between the wild type accessions of the present study represents the genetic metabolome variation, the baseline. This baseline was used as a reference for the assessment of the changes in the metabolome due to transformation. To mimic a situation in which the metabolome of a GM crop is compared with the baseline, we analysed the variation in the baseline metabolome for each genetic modification (construct) separately by removing the other GM samples from the dataset. RDA was performed with wild type accession as explanatory variable for the observed variation in the remaining dataset and internal standard intensity as cofactor, resulting in the percentage of baseline pertained metabolome variation for each transformation/tissue/analysis (Table 1).

Transformation and new trait

The same datasets that were used for calculation of the baseline variation were used for calculation of the variation due to transformation (with 2way, MYB28 or SN19 construct) and/or the newly introduced trait. In RDA, the introduced construct was used as the explanatory variable while internal standard intensity and wild type accession were used as cofactors. Some of the introduced constructs (transformation + new trait) explain more variation than the accessions in some datasets (Table 1). The 2way construct caused a larger variation than accessions in GC-MS root and LC-MS shoot metabolites, the MYB28 construct in GC-MS root and LC-MS root and the SN19 construct in GC-MS root metabolites. Having samples from independent transformation events for each construct in An-1, Col-0 and Eri backgrounds (except for MYB28 in An-1) we could eliminate the variation that is not specific to the new trait but is caused by the transformation, such as positional effect of the introduced construct. In RDA, the introduced construct was used as explanatory variable and internal standard intensity, accessions and transformation event were set as cofactors. Some of the constructs (such as 2way in GC-MS root dataset) still have a larger trait-specific variation than the accession variation (Table 1). Whether this implies that they are also significantly different from the baseline will be discussed below.

Metabolic distance to the baseline

The approach of Houshyani *et al.* (2011) was used to assess the difference between the metabolome of the GM lines and the baseline metabolome constituted by the wild type accessions. In this approach, the metabolite fingerprint of samples is subjected to multivariate data analysis (PCA). Subsequently, a metabolic distance is calculated and statistically tested between groups of samples using scores of samples on the selected PCs and ANOSIM (analysis of similarities) to assess metabolome similarity, R. ANOSIM provides a way to test statistically whether there is a significant difference between two or more groups of sampling units.

When R=0, the members of two groups completely overlap in the multi-dimensional space and thus are identical. A positive value occurs when the members of two groups do not completely overlap and are not completely separated in the multi-dimensional space with +1 the maximum for no overlap and a complete separation between two groups within the experiment. A negative value occurs when members of one of the groups fall within the members of the other group in all dimensions *i.e.* one group is a subset of the other group in the multi-dimensional space. Since the priority was to assess the effect of transformation on the metabolome, the accession samples of the previous experiment were excluded from the analysis. However, it is clear that growing year has a large influence on the metabolome (Table 1) and this should hence be used to establish a robust baseline for GM risk assessment.

GC-MS metabolic distance of GM groups to the baseline

Table 2 shows the GC-MS based metabolic distances of the GM samples to the baseline and to each other. Each GM group in this analysis consisted of all the samples with the same genetic back ground and the same construct, regardless of being from different transformation events. For the shoot, the metabolic distances to the baseline were low for most constructs (indicating complete overlap with the baseline) but they were significantly positive (P-value < 0.05) for “Eri MYB28” and “Eri SN19” (Table 2), indicating a significant separation in the multi-dimensional space. For the root GC-MS metabolites all of the GM groups had a positive GC-MS metabolic distance (R) to the baseline (although all below 0.5) indicating there was separation, but not complete, of the GM groups from the baseline in the multi-dimensional space. This separation was significant (P-value < 0.05) for all constructs except “An-1 MYB28” and “Col-0 MYB28” (Table 2).

Table 2 GC-MS metabolic distances between groups (*baseline* and GM lines) based on shoot (above diagonal) and root (below diagonal) GC-MS profiles. *A. thaliana* accessions (An-1, Col-0 and Eri) were transformed by introduction of different constructs (2way, MYB28 and SN19). The baseline consists of the wild type An-1, Col-0, Cvi and Eri replicates. Metabolic distances are obtained by the ANOSIM permutation test on inter-sample Euclidian distances based on shoot and root GC-MS PCA plots. Significant distances (permutation test, *P*-value < 0.5) are shown in bold.

		GC-MS Shoot									
		BASELINE	**An-1 2way**	**An-1 MYB28**	**An-1 SN19**	**Col-2way**	**Col-0 MYB28**	**Col-0 SN19**	**Eri 2way**	**Eri MYB28**	**Eri SN19**
GC-MS Root	**BASELINE**		-0.09	0.06	0.02	-0.07	-0.06	0.02	0.11	**0.14**	**0.36**
	An-1 2way	**0.44**		**0.25**	-0.05	-0.04	-0.02	-0.03	0.12	-0.03	**0.15**
	An-1 MYB28	0.10	-0.13		0.03	0.09	0.15	0.10	0.19	-0.07	0.00
	An-1 SN19	**0.33**	-0.01	-0.14		-0.05	0.00	-0.08	-0.01	-0.04	0.04
	Col-2way	**0.19**	0.06	0.09	-0.03		-0.04	-0.01	0.09	-0.05	**0.14**
	Col-0 MYB28	0.07	0.10	-0.06	0.04	-0.02		0.07	**0.17**	-0.07	**0.20**
	Col-0 SN19	**0.32**	-0.02	-0.04	-0.07	-0.05	0.04		0.05	0.02	0.09
	Eri 2way	**0.16**	0.05	0.03	0.01	-0.03	-0.03	-0.01		-0.04	0.00
	Eri MYB28	**0.16**	**0.16**	0.02	0.08	-0.08	-0.01	0.01	-0.07		0.02
	Eri SN19	**0.33**	0.00	-0.12	-0.03	-0.05	0.03	-0.07	-0.03	0.04	

Table 3 LC-MS positive metabolic distances between groups (*baseline* and GM lines) based on shoot (above diagonal) and root (below diagonal) LC-MS profiles. *A. thaliana* accessions (An-1, Col-0 and Eri) were transformed by introduction of different constructs (2way, MYB28 and SN19). The baseline consists of wild type An-1, Col-0, Cvi and Eri replicates. Metabolic distances are obtained by the ANOSIM permutation test on inter-sample Euclidian distances based on shoot and root LC-MS positive PCA plots. Significant distances (permutation test, *P*-value < 0.5) are shown in bold.

		LC-MS positive Shoot									
		BASELINE	**An-1 2way**	**An-1 MYB28**	**An-1 SN19**	**Col-2way**	**Col-0 MYB28**	**Col-0 SN19**	**Eri 2way**	**Eri MYB28**	**Eri SN19**
LC-MS positive Root	**BASELINE**		**0.15**	0.19	**0.15**	0.12	**0.25**	0.12	0.02	**0.12**	0.05
	An-1 2way	**0.50**		**0.28**	-0.05	**0.26**	**0.20**	**0.24**	**0.40**	**0.59**	**0.32**
	An-1 MYB28	-0.14	0.07		0.25	0.03	-0.11	0.16	0.08	0.20	0.03
	An-1 SN19	**0.39**	-0.04	0.05		**0.14**	**0.20**	**0.16**	**0.33**	**0.49**	**0.26**
	Col-2way	**0.33**	0.01	0.03	-0.03		0.04	-0.03	0.05	**0.16**	0.07
	Col-0 MYB28	**0.17**	0.13	-0.03	0.02	-0.02		0.05	**0.13**	**0.28**	0.10
	Col-0 SN19	**0.48**	0.05	**0.43**	-0.05	-0.03	0.06		**0.23**	**0.35**	**0.27**
	Eri 2way	**0.31**	0.08	**0.30**	-0.03	-0.02	-0.02	0.04		0.01	-0.08
	Eri MYB28	**0.25**	**0.32**	0.22	**0.12**	0.06	0.01	0.06	0.05		0.07
	Eri SN19	**0.41**	0.00	0.15	-0.06	-0.04	0.02	-0.08	-0.03	0.06	

LC-MS metabolic distance of GM groups to the baseline

Each GM group again consisted of all the samples with the same genetic back ground and the same construct, regardless of being from different transformation events. For the LC-MS data none of the GM groups had a metabolic distance over 0.5 to the baseline showing that there was no complete separation from the baseline of any of the GM groups in the multi-dimensional space (Table 2). Just as for GC-MS, also for LC-MS there were more significant metabolic distances for root than for shoot metabolites (Table 2 and 3). Three GM groups that had no significant shoot GC-MS based metabolic distance ($P > 0.05$) to the baseline did have a significant shoot LC-MS metabolic distance (An-1 2way, An-1 SN19 and Col-0 MYB28) while Eri SN19 that exhibited a significant shoot GC-MS distance (P-value < 0.05) had no significant LC-MS distance to the baseline (Table 3). Almost similar to the observations in root GC-MS based metabolic distances, all of the root LC-MS metabolic distances of GM groups to the baseline (except for "An-1 MYB28") were significant ($P < 0.05$).

Metabolic distance of the transformation events to the baseline

The metabolic distance of a GM group to the baseline as described above was derived using the metabolite fingerprint of baseline samples and GM samples with the same genetic background and construct regardless of being from different transformation events. RDA analysis showed that the transformation event had a considerable contribution to the explained variation (Table 1). Therefore, samples of different transformation events with a similar background could be separated in the hyper-plane of PCA plots, such as in An-1 2way transformation (Fig. 1). Therefore, we also calculated the metabolic distance of each independent transformation event to the baseline (Table 4).

A clear distinction between the terms "explained variation" and "metabolic distance" is necessary for fair assessment of genetic modification. The ability of a source to explain a large portion of variation (such as genetic modification compared with baseline) does not necessarily mean that samples that are affected by that source will have a significant metabolic distance to the samples that are not affected. The variation accounted for can be between- or within-group effects or a mixture of both. The more effect on between-group variation, the larger the metabolic distance can be expected between the groups.

In An-1 both transformation events with the 2way construct resulted in at least one significant metabolic distance to the baseline in the different tissues and/or analysis platforms (Table 4), which is in agreement with the previous observation for An-1 2way group distance to the baseline (Tables 2 and 3) and the visual evaluation in Fig. 1. The same applies for An-1

Table 4 Metabolic distances of individual transformation events to the *baseline.* Metabolic distances are based on PCA plots with shoot and root GC-MS and LC-MS positive mode metabolite profiles of transformed *A. thaliana* accessions (An-1, Col-0 and Eri) by introduction of different constructs (2way, MYB28 and SN19) through different transformation events (-1, -2 or -3). The *baseline* consists of the wild type samples of An-1, Col-0, Cvi and Eri accessions of *A. thaliana*. Metabolic distances are obtained by the ANOSIM permutation test on inter-sample Euclidian distances based on shoot and root PCA plots obtained by the mentioned analyses. Significant distances (permutation test, *P*-value < 0.5) are shown in bold.

	Background	**An-1**		**Col-0**		**Eri**		
		2way-1	2way-2	2way-1	2way-2	2way-1	2way-2	
2way	GC-MS Shoot	-0.07	-0.16	-0.02	-0.15	0.11	0.08	
	GC-MS Root	**0.65**	**0.49**	0.22	0.21	0.28	0.10	
	LC-MS positive Shoot	**0.23**	0.11	0.10	0.11	-0.11	0.10	
	LC-MS positive Root	**0.32**	**0.89**	0.32	**0.44**	**0.37**	0.27	
		MYB28 1		MYB28-1	MYB28-2	MYB28-1	MYB28-2	MYB28-3
MYB28	GC-MS Shoot	0.06		-0.15	-0.02	0.22	0.07	**0.30**
	GC-MS Root	0.10		**0.33**	0.05	0.31	0.20	**0.35**
	LC-MS positive Shoot	0.19		0.09	**0.42**	0.17	-0.04	-0.01
	LC-MS positive Root	-0.14		0.18	0.23	-0.12	**0.36**	**0.48**
		SN19-1	SN19-2	SN19-1	SN19-2	SN19-1	SN19-2	
SN19	GC-MS Shoot	0.15	-0.09	0.04	-0.01	0.21	**0.57**	
	GC-MS Root	0.22	**0.81**	**0.87**	0.03	**0.53**	**0.40**	
	LC-MS positive Shoot	0.01	**0.31**	0.14	0.02	0.00	0.08	
	LC-MS positive Root	**0.39**	**0.60**	**0.67**	0.32	**0.61**	**0.39**	

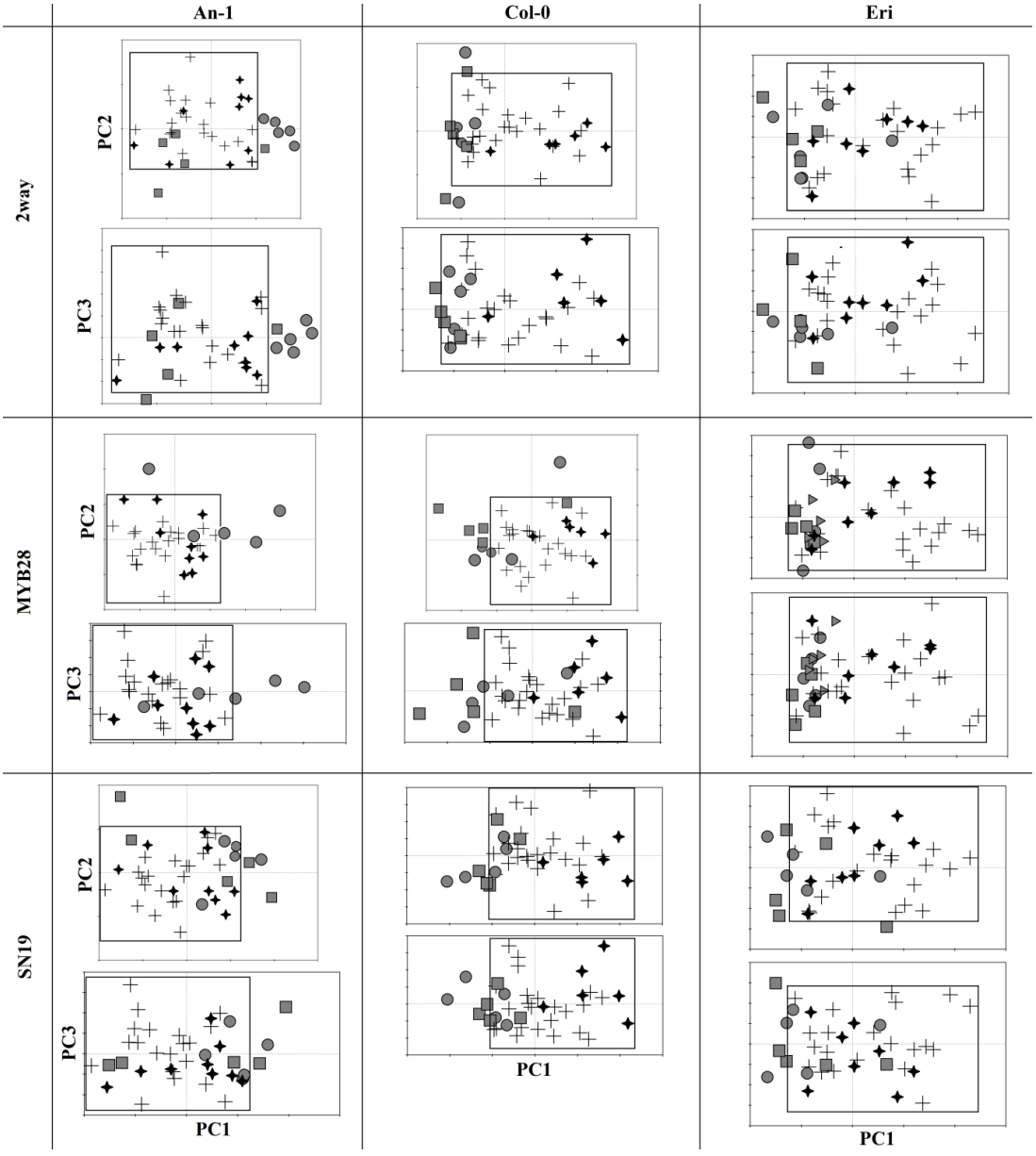


Figure 1 Score plots with the first three PCs of PCAs using different datasets of shoot LC-MS positive mode. Every dataset consists of all wild type samples forming the *baseline* (An-1, Col-0, Cvi and Eri) and GM samples in a specific background (An-1, Col-0 or Eri) and carrying one constructs (2way, MYB28 or SN19). The inner box delimits the variation of wild type samples (*baseline*) along the used PCs. ✦, wild type samples with the same background of GM samples. ✛, other wild type samples used to establish the *baseline*. ●, first transformation event samples, ■, second transformation event samples and ►, third transformation event samples

transformation with SN19. As sample scores on more than two PCs are used for calculation of the metabolic distance, a two-dimensional PCA does not necessarily visualise the calculated distance. However here, most of the distances could be confirmed by evaluating the corresponding PCAs just using the first three PCs (Fig. 1). For example, for the shoot LC-MS data, the distance between the baseline and the first event of An-1 2way transformation was larger (and significant) than for the second event (non-significant). This is reflected also in the position of the first event samples in the PCA as all of the 5 replicates are separated from the baseline samples along PC1 (but not along PC2 and PC3) (Fig. 1), whereas 3 of the 5 second event replicates are positioned within the variation of baseline samples along the first 3 PCs (Fig. 1). The single An-1 MYB28 transformation event had a non-significant metabolic distance to the baseline in both GM group and transformation event approaches, and distances were the same for both approaches as it was expected (Table 2, 3 and 4).

The Col-0 2way and Col-0 SN19 groups always had significant metabolic distance based on root data (Tables 2 and 3). However, the metabolic distances of the individual transformation events to the baseline showed that Col-0 2way-1 and Col-0 SN19-2 events have non-significant metabolic distance to the baseline in all analyses (Table 4) which is also reflected in the PCA plots of shoot LC-MS analysis shown as an example (Fig. 1). In another interesting observation, 4 of 5 replicates of Col-0 MYB28-2 transformation event were located outside the baseline variation along PC1 (Fig. 1). Indeed, the shoot LC-MS based metabolic distance of this transformation event to the baseline is relatively large and significant (Table 4).

The difference in metabolic distance between groups and transformation events can also be observed for Eri 2way and Eri MYB28 samples. These two groups had at least two significant metabolic distances to the baseline (Tables 2 and 3). Nevertheless, the metabolic distances of the individual transformation events to the baseline showed that Eri 2way-2 and Eri MYB28-1 events have non-significant metabolic distance to the baseline in all analyses (Table 4). These differences suggest a strong effect of transformation on the metabolome and consequently on the metabolic distance.

Combined analysis

When the data of root and shoot obtained on the two analytical platforms were analysed as an integrated dataset, the variation that could be explained by the new trait and transformation or the new trait alone was minimal compared to the variation in the data that could be explained by the baseline (Table 1). In this analysis, none of the transformation events had a significant metabolic distance to the baseline.

Discussion

We used the model plant *A. thaliana* for genetic modification and investigated the substantial equivalence of GM lines harbouring three different constructs to the baseline at the whole metabolome level. The objective was to develop tools to evaluate the impact of transformation on the metabolome. Evaluation of the effect of genetic modification on the metabolome and comparison of GM plants with traditionally bred lines is quite new. In a holistic approach performing PCA on transcriptomics, proteomics and metabolomics data of a Bt and a glyphosate-tolerant Roundup Ready maize and their wild type background, Barros *et al.* (2010) reported that growing season and agricultural practice individually explain larger amounts of variation than genetic modification and hardly any PC could indicate effect of genetic modification on mRNAs, proteins or metabolites. The large magnitude of environmental effect compared with genotypic effect has also been reported in other studies (Coll *et al.* 2010; Houshyani *et al.* 2011; Zeller *et al.* 2010) and was also confirmed in the present study as growing year explained much more variation than genetic modification and accession (Table 1). Our study also gives specific information about the effect of genetic modification on the root and shoot metabolome. As is evident form Table 1, the share of genetic modification in explaining the whole variation in a dataset was usually larger in root than in the shoot samples. Moreover, the number of significant distances was much higher when comparing the metabolome of the root to the baseline than that of the shoot (Table 2-4). The difference between the root and shoot metabolome has also been reported before in targeted (Kabouw *et al.* 2010; Van Dam *et al.* 2009) as well as untargeted studies (Houshyani *et al.* 2011), but the differential impact of genetic modification on subterranean and aerial parts of GM plants has not been reported before. An explanation for this phenomenon may be that the principle of phenotypic buffering proposed by Fu *et al.* (2009) is stronger in the shoot than in the roots.

Determination of the biological relevance of changes is difficult and a big gap exists between the statistical significance and the ecological impact of these changes making risk assessment impossible. In the case of metabolites it is known that the plasticity that evolved in the metabolome across *A. thaliana* accessions correlates with or determines the adaptive potential of the accessions to their dynamic environment including biotic factors. We consider the wide range of metabolome variation among accessions with a history of adaptation to their environment as the boundaries of a baseline for metabolome changes. Any change within these boundaries may cause a change in the interaction with the biotic environment, but most probably does not pose any risk to that environment. Hence, the panel of analysed *A. thaliana*

accessions in the present study (and our earlier work, Houshyani *et al.* 2011) in fact shapes a dynamic and natural boundary for metabolome variation and a baseline for substantial equivalence assessment of the GM metabolome. So far just a few groups started using natural variation in omics data as a tool for risk assessment (Baudo *et al.* 2006; Cheng *et al.* 2008; Coll *et al.* 2010; Kusano *et al.* 2011). However, there is a lack of a community consented, systematic and standard approach to measure and statistically test the similarity between groups of GM samples and a baseline for high dimensional omics data and therefore application for GM non-target effect assessment has not been realized so far. Enot and Draper (2007) suggested a complementary approach introducing margin measures and scatter matrices eigenvalues in conjunction with estimates of classification accuracy and model sensitivity. In the present paper we took a more comprehensive approach, the metabolic distance (Houshyani *et al.* 2011), to statistically measure the impact of genetic modification on the whole metabolome of GM lines.

One set of the GM lines in this study harboured the SN19 construct, which should lead to the production of Bt (Naimov *et al.* 2001), a protein and not a metabolite and not reported to exist in *A. thaliana*. The other two constructs, 2way and MYB28, were used to engineer two indigenous metabolic pathways (terpenoids and glucosinolates, respectively) and mediate biosynthesis of metabolites that can be found in at least one of the accessions of *A. thaliana* (Snoeren *et al.* 2010; Houshyani 2012b). Therefore, targeted and untargeted effects of transformation were not differentiated – that is the metabolites expected to be increased by the transformation (target effect) were not removed from the dataset - for four reasons: 1- there are always debates and lack of consensus about the definition of targeted and predictable untargeted effects (Hoekenga 2008), 2- the opponents of the substantial equivalence paradigm (that mainly focuses on targeted effects) stress the possibility that unpredictable untargeted effects can occur (Hoekenga 2008), 3- although a large part of the untargeted effects is unpredictable it always exists and 4- removing any of the GM affected metabolites from this analysis involves elimination of the natural variation from the dataset.

Our assessment shows that having GM lines with a "substantially equivalent to the baseline" metabolome is plausible (Table 4). The baseline used contained only the genetic metabolome variation while other factors can also cause large variation in the dataset, such as environment or growing year/season. Indeed, our previous study showed that environmental conditions have a dramatic effect on the metabolome and cause a large metabolic distance between groups that can easily exceed the distance between genotypes (Houshyani *et al.* 2011). It is clear that addition of the data obtained on wild type samples of the previous experiment as well as samples grown under different environments to the dataset would result

in a substantially wider baseline. This in turn would decrease the metabolic distances of the GM lines to the baseline and make them insignificant.

Our study also shows that for GM metabolome assessment it is recommended to include samples from independent transformation events. As an example, while Eri MYB28 could be rejected based on group-based metabolic distances, there was a transformation event of Eri MYB28 that did not produce any significant metabolic distance when assessed individually, suggesting therefore it is safe or not rejectable (Table 2-4). The Col-0 2way-1, Eri 2way-2, An-1 MYB28-1, Eri MYB28-1 and Col-0 SN19-2 transformation events discovered to be not significantly distant from the baseline metabolome based on all analyses (Table 4). Therefore, they could be inferred as substantially equivalent to the baseline metabolome, while the group that they belong to it had a significant metabolic distance in at least one of the analyses.

We also observed a good correlation between the amount of the transgenic product biosynthesis and its metabolic distance assessment. An1 2way-1 and Col-0 MYB28-2 transformation events showed a significant metabolic distance based on their shoot LC-MS profile to the baseline. They were also among the highest producers of (E)-nerolidol and aliphatic glucosinolates, respectively, amongst the other transformation events in the same and other backgrounds (Houshyani 2012a; Kos *et al.* 2011b). The traits were also more effective, *i.e.* Col-0 MYB28-2 had a stronger negative effect on the performance of leaf chewing caterpillars than other transgenic lines (Houshyani 2012b). This demonstrates the biological relevance of metabolic distance as it was also evident from the study of Houshyani *et al.* (2011) that shows metabolically diverged genotypes segregate in their interaction with pathogens and insects and confirms the claims about the close relationship between an organism's phenotype and relevant metabolites (Hoekenga 2008; Kok and Kuiper 2003). We would like to emphasize that the observed large influence of the transformation event on the metabolome (unpredictable unintended effect) might sometimes distort this correlation between metabolic distance and the actual GM product biosynthesis.

As metabolites are closely related to the organism's phenotype, the set of analysed metabolites to calculate the metabolic distance must be sufficiently diverse and representative of the metabolic diversity in the plant. To be able to claim that this is true for the dataset produced in the current study, complete information about the diversity of metabolites in *A. thaliana* is required. As we lack such a standard library of *A. thaliana*, we used GC-TOF-MS and LC-TOF-MS positive mode platforms to detect a range of primary and secondary polar metabolites. A former study based on three metabolomics technologies including these two, showed that the metabolic distances based on GC-MS and LC-MS negative mode correlate significantly for both root and shoot (Houshyani *et al.* 2011). Therefore, LC-MS negative

mode analysis was excluded in the present study. Because of the availability of more sensitive LC-MS positive mode equipment, the number of detected reconstructed (putative) metabolites strongly increased further reducing the need for two LC-MS modes.

Conclusion

The current study explored and assessed the changes in the metabolome of genetically engineered *A. thaliana* using a natural metabolome baseline for reference and a holistic analytical and statistical approach for evaluation. The metabolome is closely related to the biological end point and the organism's phenotype. The wide range of metabolome variation among accessions with a history of adaptation could therefore determine a boundary for metabolome changes that most probably do not pose any risk to the environment. We used this metabolome variation among *A. thaliana* accessions to establish a baseline, the limits in the multi-dimensional metabolome data. Metabolome compositions outside this baseline could possibly pose a risk to the environment. Therefore, our approach can be used as a first- pass criterion in the whole process of GM crop ecological risk assessment. Our analysis showed that the metabolome difference between some transformation events and the baseline was significant for some lines and not for others. However, our baseline only consisted of four accessions and did not include the environmental effect on the metabolome. In agricultural practice the variability in environmental conditions is much broader (Barros *et al.* 2010; Houshyani *et al.* 2011) and hence the baseline variation would be bigger. The establishment of a representative metabolome baseline is not easy. In our study on the transcriptomics substantial equivalence we created a baseline using publicly available data (Houshyani *et al.* 2012b). In the present study we only used the four wild type accessions in the baseline group that were also used for genetic modification. It would be extremely valuable if a similar database as for the transcriptomics data in *A. thaliana* could be created for metabolomics data. Initially this would probably be done for *A. thaliana* but we could foresee that such a database would later be expanded with crop species such that these data could be used by any evaluator to create a metabolome baseline for GM risk assessment.

References

Aharoni, A., Giri, A.P., Deuerlein, S., Griepink, F., De Kogel, W.J., Verstappen, F.W.A., Verhoeven, H.A., Jongsma, M.A., Schwab, W. and Bouwmeester, H.J. (2003) Terpenoid Metabolism in Wild-Type and Transgenic Arabidopsis Plants. *Plant Cell*, **15**, 2866-2884.

Barros, E., Lezar, S., Anttonen, M.J., van Dijk, J.P., Röhlig, R.M., Kok, E.J. and Engel, K.H. (2010) Comparison of two GM maize varieties with a near-isogenic non-GM variety using transcriptomics, proteomics and metabolomics. *Plant Biotechnology Journal*, **8**, 436-451.

Baudo, M.M., Lyons, R., Powers, S., Pastori, G.M., Edwards, K.J., Holdsworth, M.J. and Shewry, P.R. (2006) Transgenesis has less impact on the transcriptome of wheat grain than conventional breeding. *Plant Biotechnology Journal*, **4**, 369-380.

Bruinsma, M., Kowalchuk, G.A. and van Veen, J.A. (2003) Effects of genetically modified plants on microbial communities and processes in soil. *Biol. Fertil. Soils*, **37**, 329-337.

Chan, E.K.F., Rowe, H.C., Corwin, J.A., Joseph, B. and Kliebenstein, D.J. (2011) Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana. *PLoS Biology*, **9**.

Chan, E.K.F., Rowe, H.C. and Kliebenstein, D.J. (2010) Understanding the evolution of defense metabolites in Arabidopsis thaliana using genome-wide association mapping. *Genetics*, **185**, 991-1007.

Cheng, K.C., Beaulieu, J., Iquira, E., Belzile, F.J., Fortin, M.G. and StrÃ¶mvik, M.V. (2008) Effect of transgenes on global gene expression in soybean is within the natural range of variation of conventional cultivars. *Journal of Agricultural and Food Chemistry*, **56**, 3057-3067.

Coll, A., Nadal, A., Collado, R., Capellades, G., Kubista, M., Messeguer, J. and Pla, M. (2010) Natural variation explains most transcriptomic changes among maize plants of MON810 and comparable non-GM varieties subjected to two N-fertilization farming practices. *Plant Molecular Biology*, **73**, 349-362.

Conner, A.J., Barrell, P.J., Baldwin, S.J., Lokerse, A.S., Cooper, P.A., Erasmuson, A.K., Nap, J.P. and Jacobs, J.M.E. (2007) Intragenic vectors for gene transfer without foreign DNA. *Euphytica*, **154**, 341-353.

Craig, W., Tepfer, M., Degrassi, G. and Ripandelli, D. (2008) An overview of general features of risk assessments of genetically modified crops. *Euphytica*, **164**, 853-880.

Cunillera, N., Arró, M., Delourme, D., Karst, F., Boronat, A. and Ferrer, A. (1996) Arabidopsis thaliana Contains two differentially expressed farnesyl-diphosphate synthase genes. *Journal of Biological Chemistry*, **271**, 7774-7780.

Degenhardt, J., Gershenzon, J., Baldwin, I.T. and Kessler, A. (2003) Attracting friends to feast on foes: Engineering terpene emission to make crop plants more attractive to herbivore enemies. *Current Opinion in Biotechnology*, **14**, 169-176.

Degenhardt, J., Hiltpold, I., Köllner, T.G., Frey, M., Gierl, A., Gershenzon, J., Hibbard, B.E., Ellersieck, M.R. and Turlings, T.C.J. (2009) Restoring a maize root signal that attracts insect-killing nematodes to control a major pest. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 13213-13218.

Dicke, M. and Baldwin, I.T. (2010) The evolutionary context for herbivore-induced plant volatiles: beyond the 'cry for help'. *Trends in Plant Science*, **15**, 167-175.

Dubouzet, J.G., Ishihara, A., Matsuda, F., Miyagawa, H., Iwata, H. and Wakasa, K. (2007) Integrated metabolomic and transcriptomic analyses of high-tryptophan rice expressing a mutant anthranilate synthase alpha subunit. *Journal of Experimental Botany*, **58**, 3309-3321.

Enot, D.P. and Draper, J. (2007) Statistical measures for validating plant genotype similarity assessments following multivariate analysis of metabolome fingerprint data. *Metabolomics*, **3**, 349-355.

Ferrari, S., Plotnikova, J.M., De Lorenzo, G. and Ausubel, F.M. (2003) Arabidopsis local resistance to Botrytis cinerea involves salicylic acid and camalexin and requires EDS4 and PAD2, but not SID2, EDS5 or PAD4. *Plant Journal*, **35**, 193-205.

Figueiredo, A., Fortes, A.M., Ferreira, S., Sebastiana, M., Choi, Y.H., Sousa, L., Acioli-Santos, B., Pessoa, F., Verpoorte, R. and Pais, M.S. (2008) Transcriptional and metabolic profiling of grape (Vitis vinifera L.) leaves unravel possible innate resistance against pathogenic fungi. *Journal of Experimental Botany*, **59**, 3371-3381.

Fu, J., Keurentjes, J.J.B., Bouwmeester, H., America, T., Verstappen, F.W.A., Ward, J.L., Beale, M.H., De Vos, R.C.H., Dijkstra, M., Scheltema, R.A., Johannes, F., Koornneef, M., Vreugdenhil, D., Breitling, R. and Jansen, R.C. (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nature Genetics*, **41**, 166-167.

Gatehouse, J.A. (2002) Plant resistance towards insect herbivores: A dynamic interaction. *New Phytologist*, **156**, 145-169.

Geu-Flores, F., Olsen, C.E. and Halkier, B.A. (2009) Towards engineering glucosinolates into non-cruciferous plants. *Planta*, **229**, 261-270.

Gigolashvili, T., Yatusevich, R., Berger, B., Müller, C. and Flügge, U.I. (2007) The R2R3-MYB transcription factor HAG1/MYB28 is a regulator of methionine-derived glucosinolate biosynthesis in Arabidopsis thaliana. *Plant Journal*, **51**, 247-261.

Goodman, C.D., Casati, P. and Walbot, V. (2004) A multidrug resistance-associated protein involved in anthocyanin transport in Zea mays. *Plant Cell*, **16**, 1812-1826.

Hammer, O., Harper, D.A.T. and Ryan, P.D. (2001) Past: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*, **4**, XIX-XX.

Harrigan, G.G., Martino-Catt, S. and Glenn, K.C. (2007) Metabolomics, metabolic diversity and genetic variation in crops. *Metabolomics*, **3**, 259-272.

Hensel, G., Himmelbach, A., Chen, W., Douchkov, D.K. and Kumlehn, J. (2011) Transgene expression systems in the Triticeae cereals. *Journal of Plant Physiology*, **168**, 30-44.

Hilbeck, A., Meier, M., Römbke, J., Jänsch, S., Teichmann, H. and Tappeser, B. (2011) Environmental risk assessment of genetically modified plants - Concepts and controversies. *Environmental Sciences Europe*, **23**.

Hoekenga, O.A. (2008) Using metabolomics to estimate unintended effects in transgenic crop plants: problems, promises, and opportunities. *Journal of biomolecular techniques : JBT*, **19**, 159-166.

Houshyani B, Kabouw P, Muth D, de Vos RCH, Bino RJ, Bouwmeester HJ. 2011. Characterization of the natural variation in *Arabidopsis thaliana* metabolome by the analysis of metabolic distance. *Metabolomics **In press***.

Houshyani B, Assareh M, Busquets A, Ferrer A, Bouwmeester HJ, Kappers I. (2012a). Three-step pathway engineering results in more incidence rate and higher emission of nerolidol and improved attraction of Diadegma semiclausum. Metabolic Engineering ***Submitted***

Houshyani, B. (2012b) Application of omics technologies for environmental risk assessment of genetically modified plants, *Arabidopsis* and modified defence mechanisms as a model study. In *Plant Physiology*. Wageningen: Wageningen University.

James, C. (2010) Highlights of "Global Status of Commercialized Biotech/GM Crops: 2010. In *ISAAA Brief* ISAAA.

Kabouw, P., Biere, A., Van Der Putten, W.H. and Van Dam, N.M. (2010) Intra-specific differences in root and shoot glucosinolate profiles among white cabbage (Brassica oieracea var. capitata) Cultivars. *Journal of Agricultural and Food Chemistry*, **58**, 411-417.

Kappers, I.F., Aharoni, A., Van Herpen, T.W.J.M., Luckerhoff, L.L.P., Dicke, M. and Bouwmeester, H.J. (2005) Plant science: Genetic engineering of terpenoid metabolism attracts bodyguards to Arabidopsis. *Science*, **309**, 2070-2072.

Kashif, A., Federica, M., Eva, Z., Martina, R., Young, H.C. and Robert, V. (2009) NMR metabolic fingerprinting based identification of grapevine metabolites associated with downy mildew resistance. *Journal of Agricultural and Food Chemistry*, **57**, 9599-9606.

Kelly, P.J., Bones, A. and Rossiter, J.T. (1998) Sub-cellular immunolocalization of the glucosinolate sinigrin in seedlings of Brassica juncea. *Planta*, **206**, 370-377.

Kok, E.J. and Kuiper, H.A. (2003) Comparative safety assessment for biotech crops. *Trends in Biotechnology*, **21**, 439-444.

Kos, M., Broekgaarden, C., Kabouw, P., Oude Lenferink, K., Poelman, E.H., Vet, L.E.M., Dicke, M. and van Loon, J.J.A. (2011a) Relative importance of plant-mediated bottom-up and top-down forces on herbivore abundance on Brassica oleracea. *Functional Ecology*, **25**, 1113-1124.

Kos, M., Houshyani, B., Overeem, A.-J., Bouwmeester, H.J., Weldegergis, B.T., Loon, J.J.A.v., Dicke, M. and Vet, L.E.M. (2011b) Genetic engineering of plant volatile terpenoids: effects 1 on a herbivore, a predator and a parasitoid. *New Phytologist*, **Submitted**.

Kos, M., van Loon, J.J.A., Dicke, M. and Vet, L.E.M. (2009) Transgenic plants as vital components of integrated pest management. *Trends in Biotechnology*, **27**, 621-627.

Kusano, M., Redestig, H., Hirai, T., Oikawa, A., Matsuda, F., Fukushima, A., Arita, M., Watanabe, S., Yano, M., Hiwasa-Tanase, K., Ezura, H. and Saito, K. (2011) Covering chemical diversity of genetically-modified tomatoes using metabolomics for objective substantial equivalence assessment. *PLoS ONE*, **6**.

Kuśnierczyk, A., Winge, P., Jørstad, T.S., Troczyńska, J., Rossiter, J.T. and Bones, A.M. (2008) Towards global understanding of plant defence against aphids - Timing and dynamics of early Arabidopsis defence responses to cabbage aphid (Brevicoryne brassicae) attack. *Plant, Cell and Environment*, **31**, 1097-1115.

Leiss, K.A., Choi, Y.H., Abdel-Farid, I.B., Verpoorte, R. and Klinkhamer, P.G.L. (2009) NMR metabolomics of thrips (frankliniella occidentalis) resistance in senecio hybrids. *Journal of Chemical Ecology*, **35**, 219-229.

Leiss, K.A., Choi, Y.H., Verpoorte, R. and Klinkhamer, P.G.L. (2011) An overview of NMR-based metabolomics to identify secondary plant compounds involved in host plant resistance. *Phytochemistry Reviews*, **10**, 205-216.

Lommen, A. and Kools, H.J. (2011) MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics*, 1-8.

Lundgren, J.G., Gassmann, A.J., Bernal, J., Duan, J.J. and Ruberson, J. (2009) Ecological compatibility of GM crops and biological control. *Crop Protection*, **28**, 1017-1030.

Macel, M., van dam, N.M. and Keurentjes, J.J.B. (2010) Metabolomics: The chemistry between ecology and genetics. *Molecular Ecology Resources*, **10**, 583-593.

Nagegowda, D.A. (2010) Plant volatile terpenoid metabolism: Biosynthetic genes, transcriptional regulation and subcellular compartmentation. *FEBS Letters*, **584**, 2965-2973.

Naimov, S., Weemen-Hendriks, M., Dukiandjiev, S. and De Maagd, R.A. (2001) Bacillus thuringiemis Delta-Endotoxin Cry1 Hybrid Proteins with Increased Activity against the Colorado Potato Beetle. *Applied and Environmental Microbiology*, **67**, 5328-5330.

Ober, D. (2010) Gene duplications and the time thereafter - examples from plant secondary metabolism. *Plant Biology*, **12**, 570-577.

Pineda, A., Zheng, S.J., van Loon, J.J.A., Pieterse, C.M.J. and Dicke, M. (2010) Helping plants to deal with insects: the role of beneficial soil-borne microbes. *Trends Plant Sci.*, **15**, 507-514.

Poppy, G. (2000) GM crops: Environmental risks and non-target effects. *Trends in Plant Science*, **5**, 4-6.

Rohr, F., Ulrichs, C., Schreiner, M., Nguyen, C.N. and Mewis, I. (2011) Impact of hydroxylated and non-hydroxylated aliphatic glucosinolates in Arabidopsis thaliana crosses on plant resistance against a generalist and a specialist herbivore. *Chemoecology*, **21**, 171-180.

Rulhmann, S., Leser, C., Bannert, M. and Treutter, D. (2002) Relationship between growth, secondary metabolism, and resistance of apple. *Plant Biology*, **4**, 137-143.

Schouten, H.J., Krens, F.A. and Jacobsen, E. (2006) Cisgenic plants are similar to traditionally bred plants: International regulations for genetically modified organisms should be altered to exempt cisgenesis. *EMBO Reports*, **7**, 750-753.

Shukla, V. and Mattoo, A.K. (2009) Potential for engineering horticultural crops with high antioxidant capacity. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources*, **4**.

Smilauer, J.L.a.P. (2003) *Multivariate Analysis of Ecological Data using CANOCO*
Cambridge University Press, Cambridge.

Snoeren, T.A.L., Kappers, I.F., Broekgaarden, C., Mumm, R., Dicke, M. and Bouwmeester, H.J. (2010) Natural variation in herbivore-induced volatiles in Arabidopsis thaliana. *Journal of Experimental Botany*, **61**, 3041-3056.

ter Braak, C.J.F. (1988) CANOCO-an extension of DECORANA to analyze species-environment relationships. *Vegetatio*, **75**, 159-160.

Tholl, D. (2006) Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. *Current Opinion in Plant Biology*, **9**, 297-304.

Tikunov, Y., Lommen, A., De Vos, C.H.R., Verhoeven, H.A., Bino, R.J., Hall, R.D. and Bovy, A.G. (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiology*, **139**, 1125-1137.

Van Dam, N.M., Tytgat, T.O.G. and Kirkegaard, J.A. (2009) Root and shoot glucosinolates: A comparison of their diversity, function and interactions in natural and managed ecosystems. *Phytochemistry Reviews*, **8**, 171-186.

van der Putten, W.H., Vet, L.E.M., Harvey, J.A. and Wackers, F.L. (2001) Linking above- and belowground multitrophic interactions of plants, herbivores, pathogens, and their antagonists. *Trends in Ecology & Evolution*, **16**, 547-554.

Van Engelen, F.A., Molthoff, J.W., Conner, A.J., Nap, J.P., Pereira, A. and Stiekema, W.J. (1995) pBINPLUS: An improved plant transformation vector based on pBIN19. *Transgenic Research*, **4**, 288-290.

Vasconcelos, I.M. and Oliveira, J.T.A. (2004) Antinutritional properties of plant lectins. *Toxicon*, **44**, 385-403.

Wielanek, M., Królicka, A., Bergier, K., Gajewska, E. and Skłodowska, M. (2009) Transformation of Nasturtium officinale, Barbarea verna and Arabis caucasica for hairy roots and glucosinolate-myrosinase system production. *Biotechnology Letters*, **31**, 917-921.

Wittstock, U. and Gershenzon, J. (2002) Constitutive plant toxins and their role in defense against herbivores and pathogens. *Current Opinion in Plant Biology*, **5**, 300-307.

Yang, T., Stoopen, G., Yalpani, N., Vervoort, J., de Vos, R., Voster, A., Verstappen, F.W.A., Bouwmeester, H.J. and Jongsma, M.A. (2011) Metabolic engineering of geranic acid in maize to achieve fungal resistance is compromised by novel glycosylation patterns. *Metabolic Engineering*, **13**, 414-425.

Zeller, S.L., Kalinina, O., Brunner, S., Keller, B. and Schmid, B. (2010) Transgene x environment interactions in genetically modified wheat. *PLoS ONE*, **5**.

Zhang, X., Henriques, R., Lin, S.S., Niu, Q.W. and Chua, N.H. (2006) Agrobacterium-mediated transformation of Arabidopsis thaliana using the floral dip method. *Nature Protocols*, **1**, 641-646.

Chapter 8

General discussion: guidelines for the assessment of ecological non-target effects of genetically modified crops

Benyamin Houshyani[*] *· Martine Kos*[*] *· Patrick Kabouw*[*] *· Arjen Biere · Harro J. Bouwmeester · Nicole M. van Dam · Marcel Dicke · Joop J. A. van Loon · Wim H. van der Putten · Louise E. M. Vet*

* Authors contributed equally to the manuscript

In preparation for submission to COGEM

Introduction

Genetically modified (GM) plants are produced by inserting DNA from an organism belonging to a different species, or by overexpressing or knocking-out endogenous genes. GM plants can contain genes that confer resistance to insects, pathogens, herbicides, or environmental stresses, or genes that improve the nutritional value, storability or flavour (Nap *et al.* 2003). These beneficial traits can be attained much faster by GM approaches than by conventional breeding programs. This may partly explain the rapid increase in the cultivation of GM crops from 2 million hectares in 1996 to 148 million hectares in 2010 (James 2010).

The rapid increase in the cultivation of GM crops and the many new developments in biotechnology arose a discussion on the urgency of obtaining ecological knowledge on potential side effects of GM crop. This ecological knowledge was considered essential in the assessment of the environmental risks of GM crops, but the acquisition of the data lagged behind the biotechnological developments (Schuttelaar & Partners *et al.* 2004). Several ecological issues have to be considered before releasing GM crops into the environment (Conner *et al.* 2003, Dutton *et al.* 2003, Romeis *et al.* 2006, Scholte and Dicke 2005, Snow *et al.* 2005). These concerns relate to the invasion of GM crops into natural habitats, hybridization of the GM crops with wild relatives, horizontal gene transfer, development of resistance in target organisms and effects on non-target organisms (Conner, Glare and Nap 2003, Snow, Andow, Gepts, Hallerman, Power, Tiedje and Wolfenbarger 2005).

European Union (EU) regulatory bodies require detailed environmental risk assessment (ERA) before they permit the introduction of a GM crop into the environment. As EU sees GM plants as new entities, risk assessment for GM crops is different from the standard risk assessment that applies to conventional plant protectants in that they require the evaluation of potential adverse effects of the introduction of the GM crops on the environment (EC 2001). However, until recently, the EU did not provide clear guidelines with regard to how this risk assessment should be carried out.

The Dutch Government considered it problematic that there is a lack of ecological knowledge on transgenic crops and of clear guidelines for ERA. Therefore, in 2007, the Dutch Government fostered a research programme aimed at strengthening the ecological risk analysis for GM plants: the ERGO (Ecology Regarding Gene-modified Organisms) programme. The main objective of the ERGO-programme was to study the ecology of GM crops and to develop guidelines to assess the ecological effects of new GM crops.

In 2010, during the execution of the ERGO-programme, the EFSA (European Food Safety Authority) published the 'Guidance of the environmental risk assessment of genetically modified plants'. This document addressed seven specific areas of concern, such as invasiveness of the GM plant, plant-to-microbe gene transfer, effects of the GM plant on human and animal health, and the interaction of the GM plant with non-target organisms (EFSA 2010). To improve studies on the effect of the GM plant on non-target organisms, which is considered of paramount importance within the different ecological effects of GM crops (Groot and Dicke 2002), the EFSA provided guidelines that specify what functional groups of non-target species to include, what measurement endpoints to include, and how to perform experiments under laboratory conditions (EFSA 2010). However, the document did not provide much information on the development of field experiments, use of novel techniques such as *omics* analyses, the appropriate *baseline* against which the ecological significance of any non-target effects should be assessed and correct interpretation of the results of the non-target studies. These aspects were considered important in the ERGO-programme, and therefore, the guidelines presented in this document focus specifically on these aspects.

This document provides guidelines for assessing the non-target effects of a new GM crop and can be used when a request for permission for field trials or commercialisation of a new GM crop is submitted by a producer (the applicant) to the Dutch Advisory Commission on Genetic Modification (COGEM). The guidelines address four main questions that are important in evaluating the non-target effects of newly developed GM crops: 1) Did the applicant provide data representing the *baseline* variation? 2) Did the applicant measure the traits under representative testing conditions? 3) Did the applicant test relevant organisms and traits? and 4) Did the applicant interpret the results correctly? A summarizing step-wise question list is added to facilitate the use of these guidelines by the expert advisors at COGEM (Supplementary Material 1). This document is based on the results from three connected PhD-theses in which the plant family Brassicaceae was used as a model system (Houshyani 2012, Kabouw 2012, Kos 2012). Therefore, the guidelines have been validated for brassicaceous plant species, and most of the results that are discussed in this document were obtained from experiments with brassicaceous plants. Although the Brassicaceae family was used as a model system, this document has broader application possibilities than for brassicaceous species only.

Did the applicant provide data representing the baseline variation?

What is *baseline* variation?

Commonly, the assessment of non-target effects of a GM crop is done by comparing the GM line with the relevant isogenic line (*i.e.* the original genotype into which the transgene was introduced). However, a certain range of variation in the effects on non-target organisms also exists among conventional varieties of the crop species. This variation might be more extensive than between a GM line and its isogenic original line. Consequently, the observed effects of the GM plant may well fit within the variation that exists among conventional varieties. For this reason, it is necessary to first represent the '*baseline* variation' in order to properly evaluate the ecological effects of a GM crop within the range of variation within the crop species. In this document the term '*baseline* variation' is defined as **'the variation in effects observed among a selection of non-GM varieties, across a set of environmental conditions'**. The *baseline* variation refers to the natural variation in any plant trait, such as the metabolome of a plant or the interaction with an organism.

Information on *baseline* variation is necessary in order to assess whether the GM variety is disproportionally different from the varieties that were produced by traditional breeding. In *Brassica oleracea* L., for instance, centuries of breeding have resulted in a large number of varieties (cultivars). These varieties differ in the level of direct resistance, based *e.g.* on the biosynthesis of glucosinolates (Gols *et al.* 2008, Kabouw *et al.* 2010b, Kos *et al.* 2011a, Kushad *et al.* 1999, Poelman *et al.* 2009b) and indirect resistance, *i.e.* attraction of natural enemies (Chin and Lindsay 1993, Geervliet *et al.* 1997, Kalule and Wright 2004, Poelman *et al.* 2009a). The variation in effects on non-target organisms among these conventional cabbage varieties may potentially be larger than between a GM cabbage line and its isogenic line. Comparing the effects of the GM crop with the *baseline* variation in effects is, therefore, crucial in order to make realistic impact assessments of GM crops.

How to represent *baseline* variation?

The *baseline* variation can be represented by selecting a range of conventionally bred and/or wild-type genotypes/varieties that have a close (*i.e.* can be crossed with) taxonomic relationship at the species level with the GM plant. The environmental conditions under which a plant grows affect its interactions with other organisms. For example, soil characteristics do not only influence plant quality, but also both aboveground and belowground biota and the plant-mediated interactions between these biota (Bruinsma *et al.* 2003, Kabouw *et al.* 2011, Pineda *et al.* 2010, van der Putten *et al.* 2001). Therefore, the *baseline* variation should cover both genotypic variation and environmental variation. This environmental variation can refer to both the geographical zone, *e.g.* climate and soil type, and the management system, *e.g.* land use and cultivation practices (EFSA 2010; see also question 2). Representation of the *baseline* variation must be based on proper knowledge of the conventional agro-ecosystem.

An important aspect for *baseline* establishment is determination of the relevant traits. This is, amongst others, dependent on the modification used in the GM crop. If the modification concerns enhanced resistance to herbivores, the varieties to be used for representation of the *baseline* variation can, for example, be selected on the basis of known (preferably experimentally assessed) varying levels of resistance to herbivores. If GM plants are generated by introducing heterologous genes, it is preferable to include into the *baseline* varieties spanning the complete range of the effects on herbivores (from very susceptible to very resistant). This applies, for instance, to GM plants containing genes coding for *Bacillus thuringiensis* (*Bt*)-toxins.

In addition, untargeted metabolomics analysis could be useful for representation of the *baseline* variation. The metabolome of plants (the complete set of primary and secondary metabolites) displays genetic variation and - on top of that - varies in response to changes in the abiotic environment and also has a two-way interaction with the surrounding biota. Metabolomics can be used as a quick tool to select genotypes/varieties and environmental conditions that will encompass and enforce as much possible variation. This variation does not only reflect the plasticity in the species' metabolome but also the potential variation in its interaction with the (biotic) environment. Therefore, it can be used in target and non-target studies for *baseline* establishment (Houshyani *et al.* 2011). However, it is not possible to guarantee that the full range of variation has been covered, because the sample of genotypes/varieties included in the metabolomics analysis is, at present, necessarily limited.

Did the applicant measure the traits under representative testing conditions?

Tiered approach

Based on suggestions from the literature (Andow and Zwahlen 2006, Birch *et al.* 2007, Charleston and Dicke 2008, Dutton, Romeis and Bigler 2003, Poppy and Sutherland 2004), a stepwise (tiered) approach for testing the non-target effects of GM crops is proposed (Figure 1; based on Kos *et al.* (2009)). In this tiered approach, the assessment increases in complexity and realism based on the knowledge that is gained in previous steps. Between the tiers, sufficient feedback is necessary rather than considering the tiers as steps in a sequential, linear approach. If necessary, results from one level could be re-examined at another level to fill certain knowledge gaps (Birch, Griffiths, Caul, Thompson, Heckmann, Krogh and Cortet 2007).

Predicting effects under field conditions from greenhouse data

The tiered approach starts with laboratory and greenhouse studies. The main reasons to conduct such studies are the containment of the GM plant and the relatively small costs of greenhouse studies. If greenhouse studies can be used to predict non-target effects in the field, this will greatly facilitate future ERAs for new GM crops. Several studies have compared results from the greenhouse with results from the field. For example, four white cabbage cultivars were tested in both greenhouse and field studies for their metabolite composition and their suitability for herbivores and their natural enemies. The profiles of root glucosinolates, the characteristic secondary metabolites of brassicaceous plants, were generally comparable

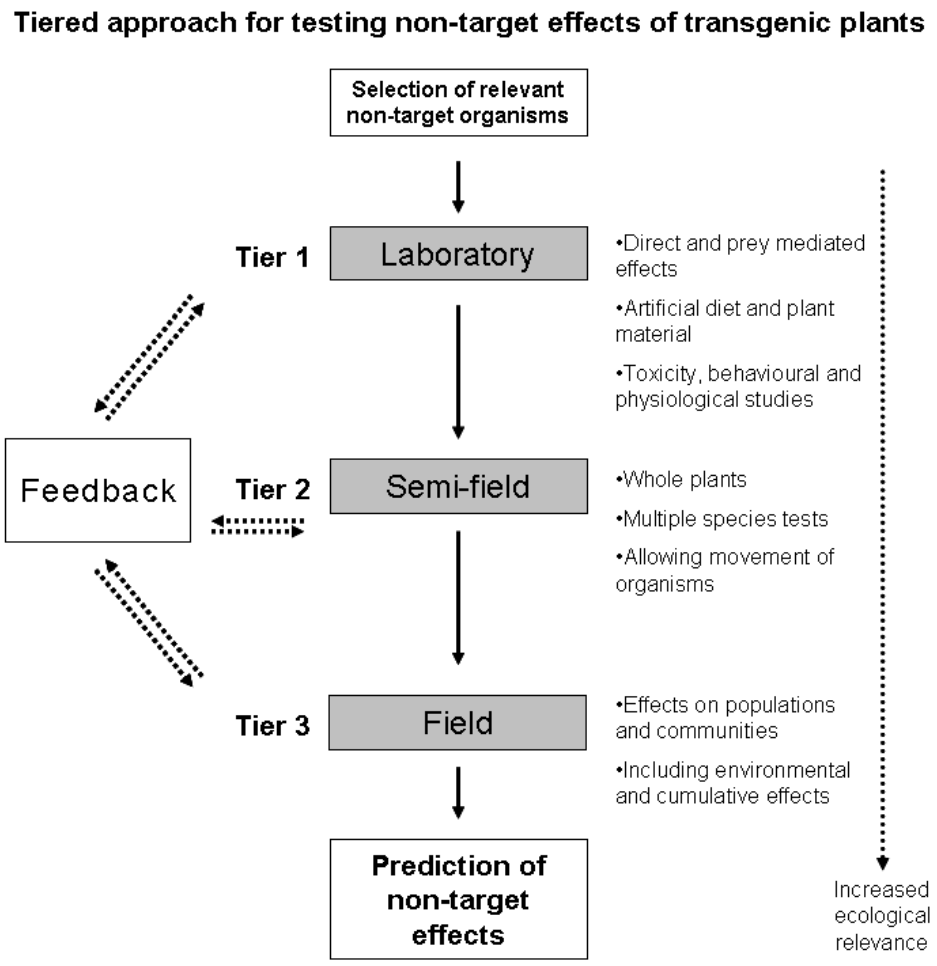


Figure 1 Stepwise (tiered) approach for testing effects of transgenic plants on non-target organisms from Kos *et al.* (2009)

between greenhouse studies and field trials, whereas shoot glucosinolate profiles were highly variable and not comparable (Kabouw *et al.* 2010a). For performance/abundance of aboveground herbivores and their natural enemies, a similar ranking of the cultivars for both greenhouse and field experiments was observed (Kabouw, Kos, Kleine, Vockenhuber, van Loon, van der Putten, van Dam and Biere 2011, Kos, Broekgaarden, Kabouw, Oude Lenferink, Poelman, Vet, Dicke and van Loon 2011a, Kos *et al.* 2011b). It has also been found that field parasitism rates of caterpillars on these four cultivars were reliably predicted by the behaviour of the parasitoids in the greenhouse (Poelman, Oduor, Broekgaarden, Hordijk, Jansen, van Loon, van Dam, Vet and Dicke 2009a).

In another study the metabolome profiles of *Arabidopsis thaliana* plants grown in four different environments (a climate chamber in soil or hydroponics, a controlled-conditions greenhouse, an uncontrolled-conditions greenhouse) were compared. The interaction between ecotype and environment explained little of the variation among metabolite profiles, suggesting that the metabolic responses of the different ecotypes to changes in the environment were quite similar (Houshyani *et al.* 2011).

The cited studies show that greenhouse studies can often, but not always, be a good indicator for the effects that can be expected in the field. However, in the greenhouse, mostly single plant-organism interactions are tested, which might not be representative for the complexity of interactions in the field. For example, early season herbivores differentially affect the responses of plant resistance to subsequently colonising herbivores, and thereby differentially affect the abundance of herbivores that occur later in the growing season (Poelman *et al.* 2008). Such effects cannot be tested in the greenhouse. Furthermore, in greenhouse studies often homogenised and sterilised potting soil is used that lacks the complex plant-soil interactions that can be present in the field. Thus, despite the fact that greenhouse studies are a good start and can save costs associated with field experiments in the beginning of the risk assessment procedure, field studies are always required later in the risk assessment procedure to validate the results from the greenhouse.

Environmental conditions

Similar to what was described for the representation of the *baseline* variation, it is important to include several environmental growing conditions (*e.g.* soil type and climatic region) that are representative of the cultivation area of the crop when testing non-target effects. EFSA advises applicants to identify the geographical regions within the EU where the GM plant is planned to be introduced, and to select several of these regions which reflect the appropriate climatic, ecological and agricultural conditions for testing effects of the GM crop

on non-target organisms (EFSA 2010). Growing conditions that are not commercially used for a crop (*e.g.* hydroponics for maize) have a limited predictive value and could be excluded, because these might lead to a biased view on the non-target effects that might not be observed under 'normal' cultivation conditions.

Did the applicant test relevant organisms and traits?

Selection of non-target organisms

There are several publications that suggest criteria for selecting aboveground and belowground non-target organisms to be included in risk assessments (Table 1). An example of such a criterion is that the selected non-target species are of key importance in ecosystem functioning, such as natural enemies of herbivores, pollinators and decomposers. Other criteria indicate that the selected species are not only exposed to, but are also susceptible to the transgenic product, which is dependent on the expression of the gene in the plant, the specificity of the transgenic product and the feeding mechanism of the organism.

Evaluation of below- and aboveground systems

Plants function as essential links between aboveground and belowground organisms (Bezemer *et al.* 2005, Bezemer and van Dam 2005, van der Putten, Vet, Harvey and Wackers 2001, Wardle *et al.* 2004). Therefore, even if the transgene is expressed in specific tissues only, it is advised that the effects of the GM crop are tested for both aboveground and belowground organisms (Bruinsma, Kowalchuk and van Veen 2003).

Selection of non-target traits

The fitness of an organism is the most important non-target trait to quantify. Several performance traits, such as mortality, development time and adult weight can be used as proxies for the fitness of an organism (Roitberg *et al.* 2001). Changes in the values of these traits can have large effects on the populations of non-target organisms. However, many studies on non-target effects of GM plants only assessed mortality of the organisms. Sub-lethal effects, such as a change in development time or the sex-ratio of the population are often ignored. This can lead to an underestimation of the effect of the GM crop on the non-target organism (Charleston and Dicke 2008, EFSA 2010). Therefore, it is important to study lethal as well as sub-lethal effects on non-target organisms. The intrinsic rate of population growth,

for example, is considered to be a good and realistic parameter for non-target studies, because it combines both lethal and sub-lethal effects (Charleston and Dicke 2008).

Effects on non-target organisms are preferably studied during multiple generations, for instance to unravel the longer term impact of sub-lethal effects (EFSA 2010). For non-target effects at the population level, long-lasting field experiments are required. For multivoltine species, which have two or more generations per year, a field experiment that is performed over an entire field season is required, and results are preferably validated during a second field season. For univoltine species, which have one generation per year, more than one field season is always required. Alternatively, modelling studies could play a role in predicting the effects of a change in the development time or fecundity on the population dynamics of non-target organisms.

Although most non-target studies focus on effects on the fitness of the non-target organisms in confined, no-choice situations, non-target testing should preferably also include effects on the selection behaviour of the organisms (Jongsma *et al.* 2010, Poppy and Sutherland 2004). For instance, the organisms might prefer to forage on the isogenic or a conventionally bred variety over the transgenic variety. Such behavioural factors might limit negative effects of GM plants under field conditions, and ignoring effects on behaviour of non-target organisms may lead to a biased view of the potential environmental effects of GM plants.

Use of untargeted metabolomics

Untargeted metabolomics can be used to study the (unintended) effects of the genetic modification on the plant's chemical properties. In this way, it can be assessed whether the metabolic profile of the GM crop falls outside the metabolic *baseline* variation, established as described above. Untargeted metabolomics can potentially also be predictive for the effect of plants on herbivores and pathogens. In a study on Arabidopsis ecotypes, their resistance against pathogens and insects could be correlated with concentrations of several metabolites known to be important in plant-insect interaction (Houshyani *et al.*, 2011). Although untargeted metabolomics can be a useful tool in non-target testing of GM crops, for the time being metabolomics cannot replace actual testing of non-target effects until more evidences about the predictive value of untargeted metabolomics are obtained during future GM crops assessments.

Did the applicant interpret the results correctly?

Hypothesis testing

After completion of the non-target studies, it is important that the results are interpreted correctly. Statistical analysis is the final decision-making tool to assess non-target effects of the GM crop. In statistical analyses, a null hypothesis (set prior to the experiment) is evaluated. The null hypothesis for testing the effects of a GM crop on non-target organisms can be: 'the effects of the GM crop are not significantly different from the full complement of effects found in the *baseline* variation'. It is important to understand that a statistical analysis of a data set can only reject a null hypothesis or fail to reject it. If the comparison of the *baseline* variation and the GM plant reveals no statistically significant difference between them, the null hypothesis should not be rejected. Nevertheless, it does not mean that there is no difference in reality. It only means that there is not enough evidence to reject the null hypothesis.

When the null hypothesis is rejected when it should not be, a type I error occurs. This comprises a risk for the producer of the GM crop, because the producer has to develop a new crop in which no negative effects on non-target organisms are observed before permission for introduction in the agro-ecosystem will be granted. When the null hypothesis is not rejected when it should be, a type II error occurs. This is a risk for the environment, because the GM crop causes non-target effects that were not detected. The statistical power of an experiment refers to the ability to prevent type I and II errors. By increasing the sample size of an experiment, the statistical power can be enhanced. It is important to use a proper experimental design to make proper statistical analyses feasible.

Use of multivariate statistics (MVS) to assess non-target effects

When only one non-target trait is selected, a univariate analysis such as ANOVA (or the non-parametrical Kruskal-Wallis test) with a post-hoc multiple comparison test can be used to test whether effects of the GM crop differ from the *baseline* variation. However, experiments that address the non-target effects of GM crops can result in large datasets involving several traits. These large datasets might be difficult to analyse or interpret. Multivariate statistics (MVS) based on ordinations is helpful to analyse and visualise differences in suites of traits and to assess whether the genetic modification of a plant results in effects that are within or outside the *baseline* variation. In the context of representation *baseline* variation, MVS can also be used to identify varieties that represent extremes or averages for a set of measured traits. In general, MVS is a useful tool to classify samples and to visualise differences and relationships between samples for which multiple variables have

been recorded. Principal Component Analysis (PCA), for example, converts a set of observations (variables) into a set of uncorrelated variables, called principal components. The first principal component explains the highest percentage of the total observed variation in variables, followed by the second principal component, etc. The resulting PCA plot shows how the observations are related and if there are any deviating observations in the data. In further steps of the MVS, the differences between the groups of observations (*e.g.* genotypes or environments) can be statistically analysed. In this way, it can be analysed whether the effects of the GM crop are different from the effects found in the *baseline* variation. MVS was for example used to analyse the large datasets of untargeted metabolomics and interpret the meaning of these multivariate data (Houshyani *et al.* 2011).

Although MVS is a useful tool for ecologists to evaluate non-target effects of GM crops, there are a number of challenges when using MVS. In particular, it is important to not ignore important variation and to correctly assess whether observed differences are statistically significant. Publications using MVS frequently show and interpret only the first two principal components. However, third and higher principal components of ordination diagrams might still contain relevant information. On the other hand, analysis of too many principal components will result in analysing meaningless information (*i.e.* the random error). To determine the correct number of principal components to include, it is necessary to have an objective criterion for deciding how many components to retain for correct interpretation. There are several statistical methods to evaluate how many components should be included in MVS models, namely scree plots, Horn's parallel analysis, broken stick, Kaiser Method, and cross validation. The use of these methods can increase the accuracy of MVS interpretation. Horn's parallel analysis was found to be one of the best to evaluate the significance of components (Hayton *et al.* 2004, Lautenschlager 1989, Watkins 2000, Zwick and Velicer 1986). Nevertheless, some researchers include control replicates to their multivariate analysis that in principle are completely similar to each other. As long as these samples cluster along a principle component, it can be included in the analysis.

The second challenge is to determine whether the observed differences, *e.g.* between a GM plant and the *baseline* variation, are statistically significant. Frequently, this is tested by using the scores of the samples (*i.e.* the position of the sample in the ordination space) in more "traditional" (univariate) statistics such as ANOVA after an ordination. An alternative is to use Monte Carlo permutation tests to determine if the overall differences between groups of samples (*e.g.* plant varieties) in an MVS are statistically significant. The advantage of the Monte Carlo permutation test is that it is performed on the raw data underlying the ordination diagram and does not transform the data. This can result in a more accurate analysis of the underlying trends (Anderson and Legendre 1999, Legendre and Anderson 1999).

If non-target effects differ from the effects found in the *baseline* variation

If the non-target effects of the GM crop differ from the effects found in the *baseline* variation, the null hypothesis is rejected (see stepwise question list in the Supplementary Material 1), this indicates that cultivating the GM crop could pose an ecological risk for non-target organisms.

Conclusions

We advise regulators to include several novel aspects in the assessment of non-target effects of GM crops. These are**:**

Compare non-target effects of GM crops with the baseline variation

Instead of only comparing the non-target effects of the GM crop with those of its genetic background, it is recommended to compare the non-target effects of a transgenic crop with the *baseline* variation. The *baseline* information is necessary in order to assess whether the GM plant is disproportionally affecting non-target organisms compared with varieties that were produced by traditional breeding and/or were grown under a range of common environmental conditions.

Select of several plant varieties and environments for representing the baseline variation

The *baseline* variation can be represented by selecting a range of genotypes/varieties that should be as much as possible different in the trait(s) under study from the GM crop (with a taxonomic relationship at the species level but still crossable) and by including several appropriate environmental conditions, *e.g.* different soil types and climates.

Use metabolomics data for baseline representation and/or as a part of the non-target testing.

Untargeted metabolomics, in which the entire metabolome of a plant is studied, is a useful tool to select the genotypes/varieties and environmental conditions to be included in the

set used for assessment of non-target *baseline* effects (as described under 2), as well as to study the effect of genetic modification on the plant's chemical properties.

Use multivariate statistical approaches for baseline representation and for assessment of non-target effects.

Multivariate statistics (MVS) based on ordinations can be used to analyse and visualise differences in a suite of traits, which can be valuable for *baseline* representation and for comparing the effects of a GM crop with the *baseline* variation.

Table 1 Criteria for selecting non-target organisms in risk assessments of GM crops

Criterion	**Explanation**	**References**
Ecological importance	Importance of organism in ecosystem functioning, *e.g.* pollinators, natural enemies of pest species and decomposers	(Andow and Hilbeck 2004, Andow and Zwahlen 2006, Birch, Griffiths, Caul, Thompson, Heckmann, Krogh and Cortet 2007, Bruinsma, Kowalchuk and van Veen 2003, Dutton, Romeis and Bigler 2003, EFSA 2010, Prasifka *et al.* 2008, Scholte and Dicke 2005)
Economic importance	Importance of organism in agriculture, *e.g.* honeybees and pollinators of fruit and seed crops	(Andow and Hilbeck 2004, Dutton, Romeis and Bigler 2003, EFSA 2010, Prasifka, Hellmich, Dively, Higgins, Dixon and Duan 2008, Scholte and Dicke 2005)
Potential exposure to product of the transgene	Depending on expression of the gene in the plant, on feeding behaviour of the organisms, and in the case of carnivores on abundance and feeding behaviour of their prey/host	(Andow and Hilbeck 2004, Bruinsma, Kowalchuk and van Veen 2003, Dutton, Romeis and Bigler 2003, EFSA 2010, Prasifka, Hellmich, Dively, Higgins, Dixon and Duan 2008, Scholte and Dicke 2005)
Susceptibility to transgenic product	Depending on specificity of the transgenic product	(Bruinsma, Kowalchuk and van Veen 2003, Dutton, Romeis and Bigler 2003, EFSA 2010, Scholte and Dicke 2005)
Functional group	Selected organisms should be representatives of different functional groups, *e.g.* herbivores, pollinators, predators, parasitoids and decomposers	(Andow and Hilbeck 2004, Andow and Zwahlen 2006, Bruinsma, Kowalchuk and van Veen 2003, Dutton, Romeis and Bigler 2003, EFSA 2010, Scholte and Dicke 2005)

Table 1 Continued

Criterion	Explanation	References
Cultural status	Based on protected status of organism (*e.g.* endangered species) and cultural or symbolic value for society	(Andow and Hilbeck 2004, Andow and Zwahlen 2006, Dutton, Romeis and Bigler 2003, EFSA 2010, Prasifka, Hellmich, Dively, Higgins, Dixon and Duan 2008, Scholte and Dicke 2005)
Abundance in ecosystem	Selected organisms should occur in the agro-ecosystem, and abundant species are considered more likely to be important in the system	(Andow and Hilbeck 2004, Charleston and Dicke 2008, Dutton, Romeis and Bigler 2003, EFSA 2010, Scholte and Dicke 2005)
Availability of knowledge	Information on the organism is necessary in the risk assessment study, *e.g.* for assessing ecological importance.	(Dutton, Romeis and Bigler 2003, Scholte and Dicke 2005)
Availability and amenability	Selected organism should be commercially available or easy to keep and rear, and amendable for use in laboratory, greenhouse and field tests.	(Birch, Griffiths, Caul, Thompson, Heckmann, Krogh and Cortet 2007, Bruinsma, Kowalchuk and van Veen 2003, Dutton, Romeis and Bigler 2003, EFSA 2010, Prasifka, Hellmich, Dively, Higgins, Dixon and Duan 2008, Scholte and Dicke 2005)

Supplementary Material 1 Step-wise question list to facilitate the use by regulators

If the answer to any of the following questions is 'no', this indicates a flaw in the set-up of the study that was performed to test the non-target effects of the GM crop. If the answer to all questions is 'yes', the study was performed properly.

Did the applicant select the correct baseline variation?

- Was a range of varieties selected for establishment of the baseline variation?
- Were the selected varieties of the same (or a closely related) species?
- Were several environmental conditions (e.g. soil types and climates) used for baseline establishment?
- Was the selection of varieties/environments based on a relevant trait (for example resistance to herbivores or metabolomics)?
- Was the baseline variation established based on correct use of statistics (e.g. multivariate statistics?)

*Did the applicant measure the traits under representative testing **conditions?***

- Was a tiered (stepwise) approach used (see Figure 1)?
- Were results from one level re-examined at another level to fill in certain knowledge gaps, if necessary?
- Were field studies performed?
- Were multiple environmental conditions (e.g. soil types) selected for field studies?
- Were the selected environmental conditions relevant for the crop species (e.g. commercially practised)?

Did the applicant test the right organisms and traits?

- Were relevant non-target organisms selected? (see Table 1)
- Were both aboveground and belowground organisms included?
- Was fitness of the non-target organisms quantified, i.e. by using several performance parameters as proxies?
- Do the selected performance parameters have the ability to demonstrate both lethal and sub-lethal effects?
- Were the non-target effects studied during multiple generations (e.g. during multiple field seasons)?
- Was behaviour of the non-target organisms studied?

Did the applicant interpret the results correctly?

- Was the sample size large enough (e.g. > 40 replicates for laboratory studies)?
- Was pseudo-replication prevented?
- Was the formulated null hypothesis comparable to: 'the effects of the GM crop are not significantly different from the full complement of effects found in the baseline variation'?
- Were relevant statistical methods used and were the data interpreted correctly to accept (i.e. a failure to reject) the null hypothesis?
- Were multivariate statistics used to analyse multiple traits simultaneously, instead of using univariate statistics for every trait separately?

Reference

Anderson, M.J. and Legendre, P. (1999) An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, **62**, 271-303.

Andow, D.A. and Hilbeck, A. (2004) Science-based risk assessment for nontarget effects of transgenic crops. *Bioscience*, **54**, 637-649.

Andow, D.A. and Zwahlen, C. (2006) Assessing environmental risks of transgenic plants. *Ecology Letters*, **9**, 196-214.

Bezemer, T.M., de Deyn, G.B., Bossinga, T.M., van Dam, N.M., Harvey, J.A. and van der Putten, W.H. (2005) Soil community composition drives aboveground plant-herbivore-parasitoid interactions. *Ecology Letters*, **8**, 652-661.

Bezemer, T.M. and van Dam, N.M. (2005) Linking aboveground and belowground interactions via induced plant defenses. *Trends in Ecology & Evolution*, **20**, 617-624.

Birch, A.N.E., Griffiths, B.S., Caul, S., Thompson, J., Heckmann, L.H., Krogh, P.H. and Cortet, J. (2007) The role of laboratory, glasshouse and field scale experiments in understanding the interactions between genetically modified crops and soil ecosystems: A review of the ECOGEN project. *Pedobiologia*, **51**, 251-260.

Bruinsma, M., Kowalchuk, G.A. and van Veen, J.A. (2003) Effects of genetically modified plants on microbial communities and processes in soil. *Biology and Fertility of Soils*, **37**, 329-337.

Charleston, D.S. and Dicke, M. (2008) Designing experimental protocols to investigate the impact of GM crops on non-target arthropods: Committee on Genetic Modification, pp. 61.

Chin, H.W. and Lindsay, R.C. (1993) Volatile sulfur compounds formed in disrupted tissue of different cabbage cultivars. *Journal of Food Science*, **58**, 835-&.

Conner, A.J., Glare, T.R. and Nap, J.P. (2003) The release of genetically modified crops into the environment - Part II. Overview of ecological risk assessment. *Plant Journal*, **33**, 19-46.

Dutton, A., Romeis, J. and Bigler, F. (2003) Assessing the risks of insect resistant transgenic plants on entomophagous arthropods: *Bt*-maize expressing Cry1Ab as a case study. *BioControl*, **48**, 611-636.

EC (2001) Directive 2001/18/EC of the European Parliament and of the Council of 12 March 2001 on the deliberate release into the environment of genetically modified organisms and repealing Council Directive 90/220/EEC: Official Journal of the European Communities, pp. 1-39.

EFSA (2010) Guidance on the environmental risk assessment of genetically modified plants. In *EFSA Journal* European Food Safety Authority (EFSA), Panel on Genetically Modified Organisms, Parma, Italy, pp. 1879.

Geervliet, J.B.F., Posthumus, M.A., Vet, L.E.M. and Dicke, M. (1997) Comparative analysis of headspace volatiles from different caterpillar-infested or uninfested food plants of *Pieris* species. *Journal of Chemical Ecology*, **23**, 2935-2954.

Gols, R., Wagenaar, R., Bukovinszky, T., van Dam, N.M., Dicke, M., Bullock, J.M. and Harvey, J.A. (2008) Genetic variation in defense chemistry in wild cabbages affects herbivores and their endoparasitoids. *Ecology*, **89**, 1616-1626.

Groot, A.T. and Dicke, M. (2002) Insect-resistant transgenic plants in a multi-trophic context. *Plant Journal*, **31**, 387-406.

Hayton, J.C., Allen, D.G. and Scarpello, V. (2004) Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, **7**, 191-205.

Houshyani, B., Kabouw, P., Muth, D., de Vos, R.C.H., Bino, R.J. and Bouwmeester, H.J. (2011) Characterization of the natural variation in *Arabidopsis thaliana* metabolome by the analysis of metabolic distance. *Metabolomics*, ***In press***.

Houshyani, B. (2012) Application of omics technologies for environmental risk assessment of genetically modified plants, *Arabidopsis* and modified defence mechanisms as a model study. In *Plant Physiology*. Wageningen: Wageningen University.

James, C. (2010) Global status of commercialized biotech/GM crops: 2010(Applications), I.I.S.f.t.A.o.A. ed: International Service for the Acquisition of Agribiotech Applications.

Jongsma, M.A., Gould, F., Legros, M., Yang, L.M., van Loon, J.J.A. and Dicke, M. (2010) Insect oviposition behavior affects the evolution of adaptation to *Bt* crops: consequences for refuge policies. *Evolutionary Ecology*, **24**, 1017–1030.

Kabouw, P. (2012).

Kabouw, P., Biere, A., van der Putten, W.H. and van Dam, N.M. (2010a) Intra-specific differences in root and shoot glucosinolate profiles among white cabbage (*Brassica oleracea* var. *capitata*) cultivars. *Journal of Agricultural and Food Chemistry*, **58**, 411-417.

Kabouw, P., Kos, M., Kleine, S., Vockenhuber, E.A., van Loon, J.J.A., van der Putten, W.H., van Dam, N.M. and Biere, A. (2011) Effects of soil organisms on aboveground multitrophic interactions are consistent between plant genotypes mediating the interaction. *Entomologia Experimentalis et Applicata*, **139**, 197-206.

Kabouw, P., van der Putten, W.H., van Dam, N.M. and Biere, A. (2010b) Effects of intraspecific variation in white cabbage (*Brassica oleracea* var. capitata) on soil organisms. *Plant and Soil*, **336**, 509-518.

Kalule, T. and Wright, D.J. (2004) The influence of cultivar and cultivar-aphid odours on the olfactory response of the parasitoid *Aphidius colemani*. *Journal of Applied Entomology*, **128**, 120-125.

Kos, M. (2012).

Kos, M., Broekgaarden, C., Kabouw, P., Oude Lenferink, K., Poelman, E.H., Vet, L.E.M., Dicke, M. and van Loon, J.J.A. (2011a) Relative importance of plant-mediated bottom-up and top-down forces on herbivore abundance on *Brassica oleracea*. *Functional Ecology*, **25**, 1113-1124.

Kos, M., Kabouw, P., Noordam, R., Hendriks, K., Vet, L.E.M., van Loon, J.J.A. and Dicke, M. (2011b) Prey-mediated effects of glucosinolates on aphid predators. *Ecological Entomology*, **36**, 377–388.

Kos, M., van Loon, J.J.A., Dicke, M. and Vet, L.E.M. (2009) Transgenic plants as vital components of integrated pest management. *Trends in Biotechnology*, **27**, 621-627.

Kushad, M.M., Brown, A.F., Kurilich, A.C., Juvik, J.A., Klein, B.P., Wallig, M.A. and Jeffery, E.H. (1999) Variation of glucosinolates in vegetable crops of *Brassica oleracea*. *Journal of Agricultural and Food Chemistry*, **47**, 1541-1548.

Lautenschlager, G.J. (1989) A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. *Multivariate Behavioral Research*, **24**, 365-395.

Legendre, P. and Anderson, M.J. (1999) Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments (vol 69, pg 1, 1999). *Ecological Monographs*, **69**, 512-512.

Nap, J.P., Metz, P.L.J., Escaler, M. and Conner, A.J. (2003) The release of genetically modified crops into the environment - Part I. Overview of current status and regulations. *Plant Journal*, **33**, 1-18.

Pineda, A., Zheng, S.J., van Loon, J.J.A., Pieterse, C.M.J. and Dicke, M. (2010) Helping plants to deal with insects: the role of beneficial soil-borne microbes. *Trends in Plant Science*, **15**, 507-514.

Poelman, E.H., Broekgaarden, C., van Loon, J.J.A. and Dicke, M. (2008) Early season herbivore differentially affects plant defence responses to subsequently colonizing herbivores and their abundance in the field. *Molecular Ecology*, **17**, 3352-3365.

Poelman, E.H., Oduor, A.M.O., Broekgaarden, C., Hordijk, C.A., Jansen, J.J., van Loon, J.J.A., van Dam, N.M., Vet, L.E.M. and Dicke, M. (2009a) Field parasitism rates of caterpillars on *Brassica oleracea* plants are reliably predicted by differential attraction of *Cotesia* parasitoids. *Functional Ecology*, **23**, 951-962.

Poelman, E.H., van Dam, N.M., van Loon, J.J.A., Vet, L.E.M. and Dicke, M. (2009b) Chemical diversity in *Brassica oleracea* affects biodiversity of insect herbivores. *Ecology*, **90**, 1863-1877.

Poppy, G.M. and Sutherland, J.P. (2004) Can biological control benefit from genetically-modified crops? Tritrophic interactions on insect-resistant transgenic plants. *Physiological Entomology*, **29**, 257-268.

Prasifka, J.R., Hellmich, R.L., Dively, G.P., Higgins, L.S., Dixon, P.M. and Duan, J.J. (2008) Selection of nontarget arthropod taxa for field research on transgenic insecticidal crops: Using empirical data and statistical power. *Environmental Entomology*, **37**, 1-10.

Roitberg, B.D., Boivin, G. and Vet, L.E.M. (2001) Fitness, parasitoids, and biological control: an opinion. *Canadian Entomologist*, **133**, 429-438.

Romeis, J., Meissle, M. and Bigler, F. (2006) Transgenic crops expressing *Bacillus thuringiensis* toxins and biological control. *Nature Biotechnology*, **24**, 63-71.

Scholte, E.J. and Dicke, M. (2005) Effects of insect-resistant transgenic crops on non-target arthropods: first step in premarket risk assessment studies: Committee on Genetic Modification, pp. 57.

Schuttelaar & Partners, Dicke, M., van Tienderen, P.H. and van Veen, J.A. (2004) Ecologisch onderzoek noodzakelijk voor de risicobeoordeling van toekomstige genetisch gemodificeerde planten.

Snow, A.A., Andow, D.A., Gepts, P., Hallerman, E.M., Power, A., Tiedje, J.M. and Wolfenbarger, L.L. (2005) Genetically engineered organisms and the environment: Current status and recommendations. *Ecological Applications*, **15**, 377-404.

van der Putten, W.H., Vet, L.E.M., Harvey, J.A. and Wackers, F.L. (2001) Linking above- and belowground multitrophic interactions of plants, herbivores, pathogens, and their antagonists. *Trends in Ecology & Evolution*, **16**, 547-554.

Wardle, D.A., Yeates, G.W., Williamson, W.M., Bonner, K.I. and Barker, G.M. (2004) Linking aboveground and belowground communities: the indirect influence of aphid species identity and diversity on a three trophic level soil food web. *Oikos*, **107**, 283-294.

Watkins, M.W. (2000) Monte Carlo PCA for Parallel Analysis: State College, PA: Ed & Psych Associates.

Zwick, W.R. and Velicer, W.F. (1986) Comparison of 5 rules for determining the number of components to retain. *Psychological Bulletin*, **99**, 432-442.

Summary

As a result of rapid biotechnological developments in the past century, genetically modified (GM) crops were developed and introduced for field application. Despite the advantages of these crops and the professional marketing policies, people also started questioning the safety of GM products for humans and the environment. In response to that, scientific advisory bodies (such as COGEM, The Netherlands Commission on Genetic Modification) suggested that, among other measurements, an environmental risk assessment (ERA) of a GM crop should be done before introduction into the field. Ecological knowledge about the possible effects was considered a vital component of that assessment. In 2007, the Dutch Government initiated the ERGO (Ecology Regarding Gene-modified Organisms) research programme to generate a scientific basis for a sound ecological risk analysis. The main objective of the ERGO-programme was to develop ecology-based guidelines for how to best assess the possible ecological side-effects of new GM crops. Also the European Food Safety Authority (EFSA) recognised the interaction of a GM crop with non-target organisms as a potential environmental risk and therefore they provided guidelines for selection of a range of non-target organisms and phenotypes to be studied under laboratory conditions as part of a GM crop risk assessment study. These guidelines formed the basis for the ERGO research themes.

Parallel to the new biotechnological developments leading to the introduction of GM plants into the environment, new analytical techniques were also introduced that revolutionized the field of analytical biology. High throughput analytical platforms, collectively called *omics* technologies, created opportunities for untargeted analysis of cellular components with biological and ecological functions including mRNAs (transcriptomics), proteins (proteomics) and metabolites (metabolomics). These analytical platforms were recommended by several researchers in the field of GM food/feed safety for the analysis and comparison of a GM product with its safe counterpart. However, EFSA failed to formulate concrete rules about the application of the *omics* platforms in GM risk assessment perhaps due to a lack of consensus about where and how to employ these technologies in the whole ERA of GM plants. In the ERGO programme, exploration of the potential to apply *omics* platforms for ERA of GM crops was therefore one of the objectives.

This PhD thesis originates from one of the ERGO themes, assessment of the effect of genetic modification on non-target organisms. Under this theme with three PhD students a multidisciplinary approach was pursued to provide guidelines for how to evaluate non-target effects of GM crops altered in insect resistance using ecological methods as well as *omics*

platforms. In this PhD thesis, I set out to find solutions for some of the limitations in the application of *omics* platforms such as the lack of a statistical method to evaluate the differences between GM *vs.* wild type plants at the *omics* level and the question what would be a fair reference for the judgement about the effect of genetic modification. As a model for the evaluation of the impact of genetic modification on the *omics* phenotype we used three insect defence traits that we introduced using genetic modification into several different *Arabidopsis thaliana* accessions. The first trait, indirect defence, was the production of the volatile (*E*)-nerolidol which has been shown to attract predatory mites that can control spider mites. The other two traits were direct defence traits and consisted of overexpression of the transcription factor (MYB28) to boost aliphatic glucosinolate biosynthesis and the introduction of *Cry1*from *Bacillus thuringiensis* encoding the Bt toxin that is effective against lepidopteran insects (caterpillars). As a reference for comparison of the effects of the genetic modification, we used a panel of wild type *A. thaliana* accessions that were selected in this study and publically available data of different accessions and individuals of a RIL population that together constitute the *baseline*, the variation present in the non-GM background germplasm.

To allow for comparison of large datasets with this *baseline*, in **Chapter 2** a statistical measure was developed, which we coined *hyper-plane distance* and which was used to assess the non-target effects of our genetic modification in transcriptomics as well as metabolomics analyses. In *omics* untargeted analyses, multivariate, hyper-dimensional data are generated, making global comparison of samples or groups of samples very difficult. In chapter 2 a method was developed to calculate a distance between the metabolome - analysed on three different metabolomics platforms - of genotypes or environments. Hereto, we employed principal component analysis (PCA) to reduce the number of analysed metabolites to a series of principal components (PCs) or dimensions of a PCA plot. The scores of the samples on a number of PCs, representing the relative position of samples to each other on those PCs, were subsequently used in an analysis of similarity (ANOSIM). In this manner, we used the variation in the samples' PC scores to derive a distance between groups of samples on a multi-dimensional plot, the *hyper-plane distance*, in the case of metabolites called the *metabolic distance*. This distance represents between-group differences as well as within-group differences and therefore is a measure of the overlap between groups in a multi-dimensional context. Furthermore, it was also possible to statistically test the calculated distance in ANOSIM by permuting the samples' scores to produce a *P*-value for the calculated distance. *Hyper-plane distance* gives a single measure for the difference between groups of samples in a PCA hyper-plane, something that is impossible visually with many samples of many groups in a multi-dimensional context. The *metabolic distance* was used to select metabolically diverged accessions of *A. thaliana* and to determine the impact of the

environment on the metabolome of *A. thaliana*. The accessions thus selected (An-1, Col-0, Cvi and Eri) are representative for the metabolome diversity across the set of analysed accessions, and hence represent the *baseline* metabolome.

Engineering *A. thaliana* to produce the volatile (*E*)-nerolidol was used to alter indirect defence in *A. thaliana*. In **Chapter 3** several genetic engineering strategies were used to generate transgenic lines that uniformly emit sufficient amount of the volatile. Combination of the gene responsible for (*E*)-nerolidol biosynthesis (*FaNES1*) with the gene responsible for biosynthesis of its precursor, farnesyl diphosphate synthase (*FPS1L*), both equipped with mitochondrial targeting signal, resulted in higher production of (*E*)-nerolidol than with *FaNES1* alone. The transgenic production of (*E*)-nerolidol in Arabidopsis also resulted in the formation of non-volatile conjugates. Adding also 3-hydroxy-3-methylglutaryl CoA reductase 1 (*HMGR1*), a rate limiting enzyme of the mevalonate pathway, resulted in a further increase in the production of (*E*)-nerolidol as well as its non-volatile conjugates. Transgenic *A. thaliana* plants emitting (*E*)-nerolidol were more attractive to the insect *Diadegma semiclausum*, which is an important endoparasitoid of the larvae of *Plutella xylostella* (cabbage moth).

In **Chapters 4 and 5** the chemical changes in and effects of transgenic *A. thaliana* accessions altered in indirect or direct defence on insect behaviour were characterised. In **Chapters 4** the mitochondrial-targeted nerolidol synthase (*COX-FaNES1*) and the gene encoding the enzyme for the substrate (FPP) biosynthesis in mitochondria (*COX-FPS2*) were introduced into three *A. thaliana* accessions. Transgenic plants also emitted (*E*)-DMNT and linalool in addition to (*E*)-nerolidol. The aphid, *Brevicoryne brassicae*, was repelled by the transgenic lines of two of the accessions, although its performance on the transgenic lines was not affected. The aphid parasitoid, *Diaeretiella rapae*, preferred aphid-infested transgenic plants over aphid-infested wild-type for two of the accessions. Although another aphid predator, *Episyrphus balteatus*, did not differentiate between aphid-infested transgenic or wild-type plants, the results suggest that genetically engineering plants to modify their emission of VOCs holds promise for improving control of herbivores.

In **chapter 5,** *MYB28* was overexpressed in three *A. thaliana* accessions. *MYB28* overexpression had different effects (positive as well as negative) on the total aliphatic glucosinolate level in different transformation events of the same genetic background, possibly as a result of tight post-transcriptional regulation of MYB28. Furthermore, enhancement of the aliphatic glucosinolate pathway seems to be genetic background specific. Leaf damage by Brassicaceae generalist *Mamestra brassicae* and specialist *Plutella xylostella* were negatively affected by *MYB28* overexpression, giving promises for improvement of chewing pest damage control. Higher glucosinolate levels as a result of *MYB28* overexpression affected insect performance positively in the specialist and negatively in the generalist. Statistical analysis

revealed the differential influence of certain structural groups of aliphatic glucosinolates on the two different insects.

Chapter 6 demonstrates the application of the *hyper-plane distance* for the assessment of GM-mediated effects on the transcriptome. In this case, publicly available *meta data* containing the natural transcriptome variation in *A. thaliana* were proposed as a reference. Using this approach we showed that GM *Arabidopsis* lines with a novel indirect defence trait display changes in the transcriptome due to introduction of pleiotropic transgenes. However, the observed changes were well within the range of variation and plasticity in gene expression occurring naturally in *A. thaliana*. We also showed that unintended changes in the transcriptome are the result of other factors than the novel trait itself. This is an important observation because it implies that untargeted effects could be avoided or changed by using other strategies for transformation.

In **Chapter 7** all the transgenic lines generated in my thesis work were included in a metabolomics approach to study the effect of genetic modification on the metabolome level. The primary selected accessions of *A. thaliana* (Chapter 2) formed the *baseline* metabolome and the *hyper-plane distance* measurement was employed for analysis of differences. Untargeted metabolomics analyses using GC-TOF-MS and LC-TOF-MS of shoot and root material showed that the metabolome of most of the transgenic lines was substantially equal to the baseline even though the baseline did not yet include environment-induced metabolome variation. We suggest that substantial equivalence of a GM line's metabolome with the baseline can be used to infer a low or even no risk of the particular genetic modification for non-target organisms and can be used as a first-pass criterion in the assessment of non-target ecological effects.

Chapter 8 was written in collaboration with the two other PhD students from the same ERGO project. It summarizes and discusses the most important conclusions of the research done by the three PhD students and integrates the results in the form of guidelines for assessing the non-target ecological effects of a new GM crop. These guidelines suggest rules that must be taken into consideration when a request for permission for field trials or commercialisation of a new GM crop is submitted to COGEM.

Samenvatting

Toepassing van omics technologieën voor de beoordeling van milieu risico's van genetisch gemodificeerde planten

Arabidopsis en gemodificeerde afweermechanismen als model studie

Als gevolg van de snelle biotechnologische ontwikkelingen in de afgelopen eeuw, werden genetisch gemodificeerde (GM) gewassen ontwikkeld voor toepassing in de landbouw. Ondanks de voordelen van deze gewassen en een professioneel marketingbeleid, hebben mensen hun bedenkingen bij de veiligheid van GM producten voor mens en milieu. In reactie hierop stelden verschillende wetenschappelijke adviesorganen (zoals de COGEM, de Nederlandse Commissie Genetische Modificatie) voor, dat naast andere bepalingen en metingen, een milieu-risico-analyse (MRA) moet worden uitgevoerd vóórdat een GM-gewas mag worden geïntroduceerd in het veld. Ecologische kennis over de mogelijke effecten werd beschouwd als een essentieel onderdeel van deze beoordeling. In 2007 heeft de Nederlandse regering het ERGO (Ecologie Rond Genetisch gemodificeerde Organismen) onderzoeksprogramma geïnitieerd om een wetenschappelijke basis voor een goede ecologische risico-analyse te genereren. De belangrijkste doelstelling van het ERGO-programma bestond uit het genereren van hoogwaardige fundamentele ecologische kennis, om mogelijke ecologische neveneffecten van toekomstige generaties GM gewassen zo goed mogelijk te beoordelen. Ook de Europese Autoriteit voor voedselveiligheid (EFSA) erkent de potentiele interactie van een GM gewas met niet-doelorganismen als een mogelijk milieurisico en heeft daarom richtlijnen opgesteld voor GM gewas risico-evaluatie studies waarin een selectie aan niet-doelorganismen is opgenomen die bestudeerd moeten worden in het laboratorium. Deze richtlijnen vormen de basis voor de ERGO onderzoeksthema's.

Parallel aan de nieuwe biotechnologische ontwikkelingen die geleid hebben tot de introductie van GM planten in het milieu, werden nieuwe analytische technieken geïntroduceerd die eveneens een revolutie op het gebied van de analytische biologie betekenden. 'High-throughput' analytische methodes, die gezamenlijk *omics* technologieën worden genoemd, hebben mogelijkheden gecreëerd voor het niet-gericht analyseren van cellulaire componenten met biologische en ecologische functies, waaronder mRNA's (transcriptomics), eiwitten (proteomics) en metabolieten (metabolomics). Deze analytische methodes zijn aanbevolen door verschillende onderzoekers van GM levensmiddelen en -diervoeders voor de analyse en vergelijking van GM producten met een veilige (niet-GM)

tegenhanger. Echter, de EFSA heeft geen concrete regels over de toepassing van *omics* platforms in de GM risicobeoordeling opgesteld, wat wellicht te wijten is aan een gebrek aan consensus over hoe deze technologieën toe te passen in de MRA van GM planten. Eén van de doelstellingen van het ERGO-programma was dan ook om de mogelijkheden te onderzoeken om *omics* platforms toe te passen voor de MRA van GM gewassen.

Het werk dat is beschreven in dit proefschrift valt binnen één van de ERGO thema's, de beoordeling van het effect van genetische modificatie op niet-doelorganismen. Binnen dit thema werd door drie promovendi een multidisciplinaire aanpak nagestreefd om richtlijnen op te stellen hoe niet-beoogde effecten van GM insect-resistente gewassen geëvalueerd kunnen worden door middel van ecologische methoden en *omics* platforms. In dit proefschrift beschrijf ik mogelijke oplossingen voor enkele van de beperkingen in de toepassing van *omics* platforms, zoals het ontbreken van een statistische methode om verschillen tussen GM en wild-type planten op *omics* niveau te evalueren en de vraag wat een eerlijke referentie zou zijn om effecten van genetische modificatie te beoordelen. Als model voor de evaluatie van de impact van genetische modificatie met behulp van de *omics* fenotypering zijn drie insect-verdedigingseigenschappen gebruikt, die door genetische modificatie in verschillende *Arabidopsis thaliana* accessies zijn ingebracht. De eerste eigenschap, indirecte verdediging, is de productie en emissie van het vluchtige (*E*)-nerolidol, waarvan is aangetoond dat het roofmijten aantrekt die spintmijt kunnen bestrijden. De andere twee eigenschappen zijn beide directe verdediging eigenschappen en bestaan uit de overexpressie van de transcriptiefactor (*MYB28*) om de biosynthese van alifatische glucosinolaten te stimuleren en de introductie van *Cry1* uit *Bacillus thuringiensis* dat codeert voor het Bt-toxine dat effectief is tegen *Lepidoptera* insecten (rupsen). Als referentie voor de vergelijking van de effecten van de genetische modificatie, hebben we in deze studie een selectie gemaakt van verschillende wild type *A. thaliana* accessies en daarnaast gebruik gemaakt van publiekelijk beschikbare data van verschillende accessies en individuen van een RIL populatie die tezamen de *baseline* vormen, de variatie die aanwezig is in de niet-GM genetische achtergrond van *A. thaliana*. Om grote datasets te kunnen vergelijken met deze *baseline* variatie, werd in **hoofdstuk 2** een statistische methode ontwikkeld waarin we de *hyper-plane distance* introduceren, een parameter waarmee de non-target gevolgen van onze genetische modificatie op transcriptoom en metaboloom niveau worden beoordeeld. In *omics* analyses, worden multivariate, hyper-dimensionale datasets gegenereerd, waardoor een globale vergelijking van monsters of groepen van monsters erg moeilijk is. In hoofdstuk 2 wordt een methode ontwikkeld om een afstand te berekenen tussen het metaboloom van verschillende genotypen of van een genotype in verschillende omgevingen gebaseerd op drie verschillende metabolomic platforms. Hiervoor werd principale-componentenanalyse (PCA) gebruikt, een multivariate analyse methode om het aantal geanalyseerde metabolieten te beschrijven met een kleiner aantal relevante

grootheden, de hoofdcomponenten of principale componenten (PCs). De scores van de monsters op een aantal PCs, die de relatieve positie van monsters ten opzichte van elkaar vertegenwoordigen, werden vervolgens statistisch getoetst voor groepsverschillen met een ANOSIM (Analysis of Similarity). Op deze wijze wordt de variatie in de PC scores van de monsters gebruikt om de afstand tussen groepen monsters te bepalen in een multi-dimensionale ruimte, wat we *hyper-plane distance* noemen. In het geval van een metabolietanalyse noemen we dit de metabole afstand (*metabolic distance*). Deze afstand beschrijft zowel verschillen tussen groepen als verschillen binnen groepen en is daarom een maat voor de overlap tussen groepen in een multi-dimensionale context. Daarnaast was het mogelijk om de berekende afstanden statistisch te testen met ANOSIM door permutatie van PC scores om zo een *P*-waarde te berekenen. *Hyper-plane distance* geeft een maat voor het verschil tussen groepen van monsters in een PCA hyper-vlak, iets wat onmogelijk te visualiseren is wanneer er veel groepen zijn in een multi-dimensionale context. De *metabolic distance* werd gebruikt om *A. thaliana* accessies te selecteren die in hun metabolietenprofiel zo veel mogelijk van elkaar verschillen en om de impact van het milieu op het metaboloom van *A. thaliana* te bepalen. De op deze wijze geselecteerde accessies (An-1, Col-0, Cvi en Eri) zijn representatief voor de metabole diversiteit binnen de groep van geanalyseerde accessies, en vertegenwoordigen het *baseline* metaboloom.

Transformatie van *A. thaliana* planten zodat deze het vluchtige (*E*)-nerolidol produceren werd gebruikt om indirecte verdediging te wijzigen in *A. thaliana*. In **hoofdstuk 3** werden verschillende genetische manipulatie strategieën toegepast om transgene lijnen te generen die een uniforme en voldoende emissie van de geurstof vertoonden. De combinatie van het gen dat codeert voor de biosynthese van (*E*)-nerolidol (*FaNES1*) met het gen dat codeert voor de biosynthese van het substraat, farnesyl difosfaat synthase (*FPS1L*), beide voorzien van een mitochondriaal targeting-signaal, resulteerde in een hogere productie van (*E*)-nerolidol dan met *FaNES1* alleen. De productie van transgene (*E*)-nerolidol in Arabidopsis resulteerde ook in de vorming van niet-vluchtige conjugaten. Het toevoegen van 3-hydroxy-3-methyl CoA reductase 1 (*HMGR1*), een cruciaal enzym in de mevalonate biosynthese route resulteerde in een verdere toename van zowel de productie van (*E*)-nerolidol als van de niet-vluchtige conjugaten. Transgene *A. thaliana* planten die (*E*)-nerolidol produceren waren aantrekkelijker voor *Diadegma semiclausum*, een insect dat (endo) parasiteert op de larven van *Plutella xylostella* (koolmot).

In de **hoofdstukken 4 en 5** werden de chemische veranderingen gekarakteriseerd van transgene *A. thaliana* accessies die veranderd zijn in hun indirecte of directe verdediging tegen insecten en daarnaast werden de effecten daarvan op het gedrag van insecten bestudeerd. In **hoofdstuk 4** werd het mitochondriale nerolidol synthase (*COX-FaNES1*) en het gen dat codeert voor het enzym voor de biosynthese van het substraat (FPP) in de mitochondria (*COX-*

FPS2) geïntroduceerd in drie *A. thaliana* accessies. Transgene planten produceerden behalve (*E*)-nerolidol ook (*E*)-DMNT en linalool. De bladluis, *Brevicoryne brassicae*, werd afgestoten door transgene lijnen van twee accessies, hoewel zijn groei op geen van de transgene lijnen werd beïnvloed. De bladluis sluipwesp, *Diaeretiella rapae*, had een voorkeur voor bladluis-geïnfesteerde transgene planten boven bladluizen-geïnfesteerde wild-type planten in twee van de drie accessies. Hoewel een andere bladluispredator, *Episyrphus balteatus*, geen onderscheid maakte tussen luis-geïnfesteerde transgene of wild-type planten, suggereren de resultaten dat genetisch gemodificeerde planten aangepast in hun geurstof emissie wellicht succesvol gebruikt kunnen worden in biologische controle van herbivoren.

In **hoofdstuk 5** werd in drie *A. thaliana* accessies *MYB28* tot overexpressie gebracht. De *MYB28* overexpressie had verschillende effecten (zowel positieve als negatieve) op het totale gehalte aan alifatische glucosinolaten in verschillende transformaties binnen dezelfde genetische achtergrond, mogelijk als gevolg van een sterke post-transcriptionele regulatie van *MYB28*. Bovendien lijkt optimalisatie van de alifatische glucosinolatenroute specifiek voor de genetische achtergrond. Bladschade door de Brassicaceae generalist *Mamestra brassicae* en de specialist *Plutella xylostella* werd in beide gevallen negatief beïnvloed door *MYB28* overexpressie, wat aangeeft dat schade-controle van kauwende insecten geoptimaliseerd kan worden. Hogere glucosinolaat niveaus, als gevolg van *MYB28* overexpressie, beïnvloedden de insectengroei positief in de specialist en negatief in de generalist. Statistische analyse toonde aan dat bepaalde structurele groepen van de alifatische glucosinolaten verschillende effecten hebben op de twee insectensoorten.

Hoofdstuk 6 beschrijft de toepassing van de *hyper-plane distance* voor de beoordeling van de GM-gemedieerde effecten op het transcriptoom. In dit geval werden openbaar beschikbare *meta data* van de natuurlijke transcriptoom variatie in *A. thaliana* gebruikt als referentie. Met deze aanpak laten we zien dat GM *Arabidopsis* lijnen met een nieuwe directe verdedigingseigenschap veranderingen vertonen in het transcriptoom als gevolg van de invoering van pleiotrope transgenen. De waargenomen veranderingen vallen echter ruim binnen de variatie en plasticiteit in genexpressie die van nature aanwezig is in *A. thaliana*. Ook bleek dat onbedoelde veranderingen in het transcriptoom het resultaat zijn van andere factoren dan van de nieuwe eigenschap zelf. Dit is een belangrijke constatering, omdat het impliceert dat niet-gerichte effecten kunnen worden voorkomen of gewijzigd met behulp van andere transformatiestrategieën.

In **hoofdstuk 7** werden alle transgene lijnen die gegenereerd zijn in het werk beschreven in mijn proefschrift, opgenomen in een metabolomics aanpak om het effect van genetische modificatie op het metaboloom te bestuderen. De in eerste instantie geselecteerde accessies van *A. thaliana* (hoofdstuk 2) vormden het *baseline* metaboloom en de *hyper-plane distance* meting werd gebruikt voor de analyse van verschillen. Ongerichte metabolomics

analyses van spruit- en wortelmateriaal met behulp van GC-TOF-MS en LC-QTOF-MS, toonde aan dat het metaboloom van de meeste transgene lijnen nagenoeg gelijk is aan de *baseline*, ondanks dat deze nog geen milieu-geïnduceerde metaboloom variatie omvatte. We suggereren dat wanneer een GM-lijn een wezenlijk gelijkwaardig metaboloom als dat van de *baseline* vertoont, dit kan worden gebruikt om het risico van een bepaalde genetische modificatie voor niet-doelorganismen als laag in te schatten.

Hoofdstuk 8 is geschreven in samenwerking met de twee andere promovendi uit hetzelfde ERGO-project. Het bespreekt de belangrijkste conclusies van het onderzoek dat door de drie promovendi is uitgevoerd en integreert hun resultaten in de vorm van richtlijnen voor de beoordeling van niet-bedoelde ecologische effecten van een nieuw GM gewas. Hierin wordt een procedure voorgesteld die moeten worden uitgevoerd wanneer een nieuw GM gewas voor verzoek om toestemming voor veldproeven of commercialisering wordt voorgelegd bij COGEM.

Biography

Benyamin Houshyani was born on August 26, 1974 in Esfahan-Iran. He attended high school in Iran and UAE (Dubai) during which time he reached the best rank.

During his bachelors program in Mohaghegh Ardabili University (UMA) in Ardebil-Iran, Benyamin gained experience in dry-land farming and seed production, and in February of 1996 received his Bachelor of Science in Agronomy and Plant Breeding. Afterwards, he attended the Iranian navy where he accomplishes his military service in 1998, during which he prepared himself for the nation-wide admission examination for a masters program in Agro-technology Management.

After study, he worked at Bazargan Co. in Tehran as a field technical expert for vegetable seeds and stayed in vegetable seeds industry until 2005. During this period, he travelled around the world and to all major cities in Iran. These provided valuable professional experience for him. Meanwhile, he married to Parisa Rezaie in 2002. His last position in the industry was vegetable seeds advisor for Nunhems in Iran.

His interest in new developments in plant sciences, particularly genetics, led him to Wageningen University in the Netherlands, where he pursued his masters degree in Plant Breeding and Genetic Resources. At the end of his graduate studies he applied for a PhD position in the department of Plant Physiology in Wageningen University under the supervision of Prof. Harro J. Bouwmeester. There, he realized implementation of omics technologies for the assessment of genetically modified plants. His work resulted in 8 papers which were or will be published.

During his graduate studies, Benyamin kept his interest in vegetable seeds industry and seed science. In 2011, Benyamin was offered a position in Monsanto B.V. and started his new career as a seed technology scientist and is ambitious to become a full scientist in vegetable seeds technology.

List of publications

Houshyani B, van der Krol SAR, Bino RJ, Bouwmeester HJ. *Assessment of transcriptome perturbations in Arabidopsis lines with genetically engineered indirect insect defence.* ***In prep.***

Houshyani B, Kos M, Ararsa B, van Loon JJA, Dicke M, Beekwilder J, Bouwmeester HJ. *Overexpression of HAG1/MYB28 in metabolically diverged Arabidopsis accessions: Effect on glucosinolates, gene expression profiles and performance of specialist and generalist herbivores.* ***In prep.***

Houshyani B, Kabouw P, Bouwmeester HJ. (2012) *Whole-metabolome difference assessment of GM Arabidopsis lines with three novel insect defence traits using natural metabolome variation as a reference.* ***In prep.***

Houshyani B, Assareh M, Busquets A, Ferrer A, Bouwmeester HJ, Kappers I. (2012) *Three-step pathway engineering results in more incidence rate and higher emission of nerolidol and improved attraction of Diadegma semiclausum.* Metabolic Engineering ***Submitted***

Kos M, Houshyani B, Overeem AJ, Bouwmeester HJ, Weldegergis BT, van Loon JJA, Dicke D, Vet LEM. (2012) *Genetic engineering of plant volatile terpenoids: effects on a herbivore, a predator and a parasitoid.* Pest Management Science. ***Submitted***

Kos M, Houshyani B, Wietsma R, Kabouw P, Vet LEM, van Loon JJA, Dicke M. (2012) *Effects of glucosinolates on a generalist and specialist leaf-chewing herbivore and an associated parasitoid.* Phytochemistry. ***In press***

Kos M, Houshyani B, Achhami BB, Wietsma R, Gols R, Weldegergis BT, Kabouw P, Bouwmeester HJ, Vet LEM and Dicke M. (2012) *Herbivore-Mediated Effects of Glucosinolates on Different Natural Enemies of a Specialist Aphid.* Journal of Chemical Ecology ***38(1), p100-115***

Houshyani B, Kabouw P, Muth D, de Vos RCH, Bino RJ, Bouwmeester HJ. (2011**).** Characterization of the natural variation in *Arabidopsis thaliana* metabolome by the analysis of metabolic distance. *Metabolomics* ***In press***.

Jamil M, Charnikhova T, Houshyani B, van Ast A Bouwmeester HJ. (2011) *Genetic variation in strigolactone production and tillering in rice and its effect on Striga hermonthica infection.* Planta. ***In press***

Faino L, Azizinia S, Houshyani B, Verzaux E and Ercolano MR, Visser RGF, Bai Y. (2011) *Fine mapping of two major QTLs conferring resistance to powdery mildew in tomato.* Euphytica. ***In press***

Houshyani B, Araqi MK, Minaei S. (2002) *Evaluation of the effects of rangeland seedling planter parameters on planting attributes of atriplex.* Iranian journal of natural resources. ***55(2):283-291.***

Acknowledgment

I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my supervisor, Prof. Harro Bouwmeester who gave me the opportunity to be at this stage. I have been amazingly fortunate to have an advisor who gave me the freedom to explore on my own and at the same time the guidance to recover when my steps wavered. His insightful comments and constructive criticisms were inspiring and helped me focus my ideas. Harro, you taught me when to trust others and how to make people feel confident on me.

My co-supervisors, Prof. Raoul Bino and Dr. Iris Kappers, have been always there to listen and give advice. I am thankful to them for the discussions and feedback to sort out the technical details of my work. I am also thankful to them for encouraging and for carefully reading and commenting on revisions of my manuscripts. Although not in my supervision team, I should acknowledge Prof. Sander van der Krol for his priceless advice and intellectual collaboration.

I extend my gratitude to Dr. Maarten Jongsma for his positive opinion during my job interview and later for his guidance and to Dr. Ric de Vos, Dr. Jules Beekwilder, Dr. Ruud de Maagd, Prof. Marcel Dicke and Prof. Joop van Loon for holding me to a high research standard.

Geert Stoopen, Francel Verstappen, Miriam Goedbloed and Bert Schipper, you helped me to settle down in the lab, enrich my ideas and perform my experiments as good as I could. In fact the whole experiments were not possible without your logistic and practical support. Rina (Anthonijsz), thank you for all you have done for me to make my stay in the department free of paper work and easy.

I am also indebted to the members of the Department of Genetics as well as *metabolic engineering group, metabolomics group, parasitic weeds group and Wageningen seed lab*, the members of Department of Plant Physiology, with whom I have interacted scientifically during the course of my PhD studies. Particularly, I would like to acknowledge Dr. Henk Hilhorst, Dr. Dick Vreugdenhil, Dr. Wilco Ligterink, Dr. Leonie Bentsink, Dr. Bas Dekkers, Dr. Wouter Kohlen, Dr. Joost Keurentjes, Anna Undas, Natalia Carreno Quintero, Desalegn Etalo and Dr. Thierry Delatte for the many valuable discussions that helped me better understand my research area and that of others. I appreciate the efforts of the many students who helped me in conducting the research practically and intellectually; my sincere

thanks to Maryam Assareh, Belainesh Ararsa, Panwasa Suriya (Sa) for their contributions to the various domains.

Of course, my PhD study could not be as successful as it is now without proactive collaboration and intellectual contribution of Martine Kos and Patrick Kabouw, with whom I discussed and decided a lot of things. My heartfelt thanks for everything.

Many friends in different groups shared their valuable scientific views with me that I appreciate a million, but also have helped me stay sane. Neli and Jeroen, Natalia and Padraic, Desalegn, Anna, Catarina and Bart, Jimmy, Qing and Ting, Eric, Lemeng, Wei, Julio, Wouter and Charles, you created a wonderful community around me through these difficult years. This supported and helped me overcome setbacks and stay focused on my study. I deeply appreciate your belief in me. Liping, Mohammad Majdi, Muhammad Jamil, Rashid, Noorullah, Ralph, Imran, Rik, Ronny, Manickam, Phuong, Myriam, Aldana, Hanzie and Yanxia, I greatly value your friendship and look forward to continue it.

Most importantly, none of this would have been possible without the love and patience of my wife. Parisa has been a constant source of love and support all these years. I would like to express my heartfelt gratitude to my parents, parents in law, brother and brother and sisters in law, who encouraged and supported me mentally throughout this endeavor. Maman Mahin, baba Shaban, maman Minoo and baba Mehdi, you were the source of strength for me. Bamshad, thanks for designing the cover. It was a great help. I liked it a lot. Kamran, Mahsa, Babak, Mahnoush and Vahid, thank you all for making these years pleasant for us. It was crucial to keep the balance between private life and study.

I also convey my thanks to those who enlightened my defence ceremony by their presence in the Aula and those I forgot to mention in this acknowledgment, my apologies. Finally, I thank again my supervisor for encouraging the use of correct grammar and consistent notation in my writings and for carefully reading and commenting on revisions of this thesis and my manuscripts.

Thank you all

Education Statement of the Graduate School

Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to: **Benyamin Houshyani Hassanzadeh**
Date: **16 March 2012**
Group: **Plant Physiology, Wageningen University & Research Centre**

1) Start-up phase	*date*
▸ **First presentation of your project**	
Characterization of transgenic Arabiidopsis lines altered in direct and indirect resistance traits.	Feb 11, 2008
▸ **Writing or rewriting a project proposal**	
PSG Business Challenge 2010: COAG, Center for Omics Assessment of GM crops	2010
▸ **Writing a review or book chapter**	
▸ **MSc courses**	
▸ **Laboratory use of isotopes**	
Subtotal Start-up Phase	*4.5 credits**

2) Scientific Exposure	*date*
► **EPS PhD Student Days**	
International PhD students retreat, Wageningen	Oct 02-03, 2008
EPS PhD student day, Leiden University	Feb 26, 2009
PhD retreat in Cologne, Cologne, Germany	Apr 15-16,2010
EPS PhD student day, Utrecht University	Jun 01, 2010
EPS PhD student day, Wageningen University	May 20, 2011
PhD retreat in Orsay, France	Jul 05-08, 2011
► **EPS Theme Symposia**	
EPS theme 3 day, 'Metabolism and Adaptation', Wageningen University	Nov 06, 2007
EPS theme 2 day, 'Interactions between Plants and Biotic Agents', Utrecht University	Jan 22, 2009
EPS theme 3 day, 'Metabolism and Adaptation', Univeristy of Amsterdam	Feb 18, 2009
EPS theme 2 day, Interactions between 'Plant and Biotic Agents', Utrecht University	Jan 15, 2010
EPS theme 3 day, 'Metabolism and Adaptation', Leiden University	Feb 19, 2010
EPS theme 4 day, 'Genome Biology', Wageningen University	Dec 10, 2010
EPS theme 3 day, 'Metabolism and Adaptation'	Feb 10, 2011
► **NWO Lunteren days and other National Platforms**	
ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 07-08, 2008
ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 07-08, 2009
ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 19-20, 2010
ALW meeting 'Experimental Plant Sciences', Lunteren	Apr 04-05, 2011
► **Seminars (series), workshops and symposia**	
Plant -Insect Interaction workshop, Amsterdam	Oct 26, 2007
Ecogenomics national day	Feb 29, 2008
Seminar in Entomology by Dr. Martin Heil	May 2008
Keygene symposium "green meets white"	Feb 09, 2009
Seminar by Justin Boveritz	Jan 12, 2010
Seminar "frontiers in plant-microbe interactions" by Christine Gebhardt	Feb 05, 2010
Plant sciences seminars (4 seminars)	Feb-Apr 2011
ERGO seminars	Jun 30, 2009
Ergo Annual Symposium	Feb 25-26, 2010
SIMCA-P workshop in PRI	Mar 01, 2010
Medica genomics seminar in Rotterdam	May 27, 2010
EPS seminar	May 31, 2010
Prof. Richard Michelmore presentation in Keygene	Jun 03, 2010
Systems biology day, Wageningen WICC	Jun 16, 2010

Seminars Strigolactones by Koichi Yoneyama, Takahito Nomura and Kaori Yoneyama	Jun 25, 2010
Seminar for rhizosphere plant parasitic interaction by John Yoder	Sep 01, 2010
Seminar "The value of biodiversity" by Bas Haring	Sep 16, 2010
Seminar "Studying the genetics of root growth in rice" by Adam Price	Sep 17, 2010
Seminar 'Genetic and functional analysis of disease resistance in Brassica' by Régine Delourme (INRA, Rennes, France)	Oct 05, 2010
Mini symposium: How to write a world class paper, Wageningen University	Oct 27, 2010
Plant-Insect Interaction Workshop, Wageningen University	Nov 11, 2010
EPS ExPectationS day, Wageningen University	Nov 19, 2010
Dr. Kirsten Bomblies seminar 'Genetic incompatibility and the plant immune system'	Mov 18, 2010
Seminar in Forum about relationship and phygenetic trees for languages	Nov 18, 2010
Semper florens seminar Seed and seed technology (Nunhems and Incotec)	Feb 22, 2011
Plant Soil Interactions GXE interaction seminar by Fred van Eeuwijk	Mar 15, 2011
Schilperoort Lectures: Entrepreneurship in the Agrofood sector organized by DAFNE, Wageningen	Mar 23, 2011
► **Seminar plus**	
► **International symposia and congresses**	
Conference Technology Transfer in the Plant Sciences: Enhancing National and International Benefits, Wageningen, The Netherlands	May 20-21, 2010
Metabolomics Congress, Amsterdam, The Netherlands	Jul 04-09, 2010
NERN meeting, Lunteren, The Netherlands	Feb 08, 2011
► **Presentations**	
Poster presentation in PRI	2007
Presentation in EPS PhD students day in Leiden	Feb 26, 2009
Presentation in theme 3 symposium "genotype dependent whole metabolome response to environmental perturbation"	Feb 19, 2010
Mutual presentation in Ergo symposium	Feb 26, 2010
Poster presentationin PhD retreat in Cologne	Apr 15-16,2010
Poster + oral presentation in metabolomics congress	Jul 04-09, 2010
Presentation in plant insect interaction workshop	Nov 11, 2010
Presentation in NERN meeting	Feb 08, 2011
► **IAB interview**	Dec 04, 2009
► **Excursions**	
Excursion to Monsanto	2010
Subtotal Scientific Exposure	*26.4 credits**

3) In-Depth Studies	*date*
► **EPS courses or other PhD courses**	
System Biology: Statistical Analysis of Omics Data	Dec 08-11, 2008
Metabolomics course in Leiden (5 days)	Apr 21-29, 2008
EPS PhD summer school 'Environmental Signaling', Utrecht	Aug 24-26, 2009
► **Journal club**	
participating in literature discussion group	2007-2011
► **Individual research training**	
Subtotal In-Depth Studies	*6.6 credits**

4) Personal development	*date*
► **Skill training courses**	
Scientific Writing	Jan 2010
Workshop "IP in a PPP context in Hotel & Congress Centre 'Hof van Wageningen', Wageningen	Mar 31, 2011
workshop "Business plan service providers" in Hotel & Congress Centre 'Hof van Wageningen',	Jun 17 & 22, 2011
Dutch Language A2 certificate from ROC	2008-2010
► **Organisation of PhD students day, course or conference**	
► **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*5.3 credits**

TOTAL NUMBER OF CREDIT POINTS*	**42.8**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements

** A credit represents a normative study load of 28 hours of study.*

Layout: Benyamin Houshyani
Design of the cover: Bamshad Houshyani, an outsider’s imagination of metabolic engineering for pest resistance in plants