Systems View Of The Soybean Genetic Mechanisms Involved In The Response To Plant Pathogen Infection

Konstantinos Krampis

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics and Computational Biology

M.A. Saghai Maroof
Brett M. Tyler
Ina Hoeschele
Naren Ramakrishnan
T.M. Murali

April 17, 2009
Blacksburg, VA

Keywords: soybean, gene expression, Single Feature Polymorphism, genetic markers, genetic mapping, pathogen resistance, pathway perturbation

Systems View Of The Soybean Genetic Mechanisms Involved In The Response To Plant Pathogen Infection

Konstantinos Krampis

ABSTRACT

This thesis involves the important crop plant soybean (*Glycine max*), and provides a rich information resource for breeders and geneticists working towards improving traits for pathogen resistance. Results reported here provide a systemic view at both the genetic and biochemical level, and were generated by data-mining gene expression data from soybean cultivars inoculated with plant pathogens and also recombinant inbred line (RIL) populations. The genome variability based on Single Feature Polymorphisms (SFPs) was measured for the first time in soybean, using a genetically diverse set of cultivated *G. max* lines and also a *G. soja* line. Additionally, a genetic map spanning all 20 soybean chromosomes groups were assembled in a large RIL population. The well studied metabolic pathways from the model plant *Arabidopsis thaliana*, were reconstructed in *G. max* based on sequence similarity comparison between the genomes of the two species. We performed algorithmic analysis of pathways in our set of soybean lines and RILs using the gene expression data, and acquired a systemic view of the metabolic response to pathogen infection in different genetic backgrounds. Significant differences in the patterns of pathway perturbation was observed in the different lines, and also between four different chromosomal regions that have been known to contain genetic elements contributing to pathogen resistance.

Acknowledgments

The author would like to thank all members of the advisory committee, for their time and effort to provide feedback and guidance, in order for the present work to come into completion. People whose names are not printed on the first page of this document, but without them critical pieces of this work would be missing, include: Dominic Tucker, Lecong Zhou, Sucheta Tripathy, Anne Dorance and Steve St. Martin of Ohio State and their students. Many thanks go to all members of the Plant Genome Project team, individuals as Felipe Arredondo, Ruslan Biyashev, Regina Hanlon, Lachelle Waller, and all the people of the Tyler and Maroof labs in VBI and CSES departments, who hosted me for the course of many years. Acknowledgements also to the National Science Foundation and Plant Genome Initiative which financially supported this work.

Much gratitude to my inner circle of fellow graduate students, easily identified around the GBCB program as ones that perform heavy computer coding for many hours. With whom we engaged in lengthy discussions for defining “Bioinformatics”, and for how as next generation professionals we can strive to make better software and push the field away from disconnected data silos.

Last but deserving the greatest thanks of all to Claudia, my life partner, for having to endure seeing me go through this, and also for listening to my monologues on whether the work we got ourselves into,will turn into a good deed for this world.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 2

Chapter 3

1. Introduction and background of the research

1.1. Soybean as an important crop and goals of this study

Glycine max (soybean) is one of the most important agricultural products in the United States, and the world's foremost provider of protein and edible oil. Plant breeding projects for crop yield improvement have been ongoing (Kumar 1999), and the sequencing of the soybean genome has been recently completed (http://www.phytozome.org). In addition, genomic databases such as SoyBase and the Legume Information System (http://www.soybase.org, Gonzales et al. 2005) have been developed, as data repositories and also providers of bioinformatics software tools for *G. max*. Despite these resources, more data are needed by soybean genetics and plant breeding programs, geared towards gene isolation for disease resistance and increased yield (Kumar 1999).

The goal of this thesis was to first add to the soybean genetic resources via analysis of high-throughput genomic data, based on a bioinformatic approach. Part of this goal was accomplished through the discovery of new genetic markers using microarray data, in order to add to the list of known landmarks (Song et al. 2004) on the soybean genome. In addition, dense genetic maps of the chromosomes of *G. max* were created with the identified genetic markers. These maps are a very valuable resource for Quantitative Trait Loci mapping (QTL, Doerge 2002), and constitute an important tool for soybean breeders working on discovery of novel disease resistance and increased yield genes. A second approach towards the goal of this study, was to quantify perturbation of the soybean metabolic pathways in response to pathogen infection, using microarray data and a novel algorithm. The plant defense against pathogens, involves a series of structural and physiological alterations in the cells as result of changes in the metabolic activity (Feys 2000). The current study identified metabolic pathways and genes playing a key role for defense response in *G. max*, adding therefore to the list of genes that can be used for engineering crops with improved pathogen resistance traits.

1.2. Genetic marker discovery using microarray technology

1.2.1. Microarray data for the detection of SFP genetic polymorphisms

Microarrays allow simultaneous probing for a large number of genomic regions, through hybridization of genomic DNA or cDNA generated from transcripts extracted from cells (Lashkari et al. 1997, Pollack et al. 1999). The microarray probe sequence segments are immobilized upon a non-reactive surface, and detect signal from the target DNA or cDNA which is in most cases fluorescently labeled. In this manner, the presence of certain gene transcripts within the cell can be identified and if the signal intensity is quantified appropriately, the number of transcripts or genomic copies can also be inferred. Moreover, the signal intensity can be used to deduce the affinity of the target to the probe sequences. Based on this fact, genetic polymorphisms were found in yeast (Winzeler et al. 1998), by identifying variable hybridization signals due to difference in gene sequences from different strains. In this application of microarray technology for detecting genetic polymorphisms, the new type of genetic markers was called Single Feature Polymorphism (SFP). The microarray probes that are identified as SFPs represent genomic regions, which might differ between the strains in single nucleotides or insertions/deletions of genomic sequence fragments. Therefore, in the study by Winzeler et al. (1998) it was shown for the first time that genetic polymorphisms can be identified without the traditional use of restriction enzymes or polymerase chain reaction, and without knowing the specific nature of the polymorphism.

Following the initial application of SFPs in yeast, Borevitz et al. (2003) showed that using microarrays for identifying genetic markers can also be applied successfully to organisms with more complex genomes, such as the plant *Arabidopsis thaliana*. Working again with yeast, Ronald et al. (2005) were the first to use cDNA generated directly from mRNA gene expression transcripts for the discovery of SFPs, instead of genomic DNA. In the study by Ronald et al., it was therefore demonstrated that genotyping with microarrays can be performed simultaneously with gene expression analysis. Subsequently, by using gene expression data SFPs were also discovered in more complex genomes than yeast, including those of barley, *Arabidopsis*, rice and tomato (Cui et al. 2005 and Rostoks et al. 2005, West et al. 2006, Kumar et al. 2007 and Oeveren et al. 2007, respectively). The first genetic map

based on SFPs was successfully assembled in *Arabidopsis* (West et al. 2003), by using expression data from a segregating Recombinant Inbred Line (RIL) population.

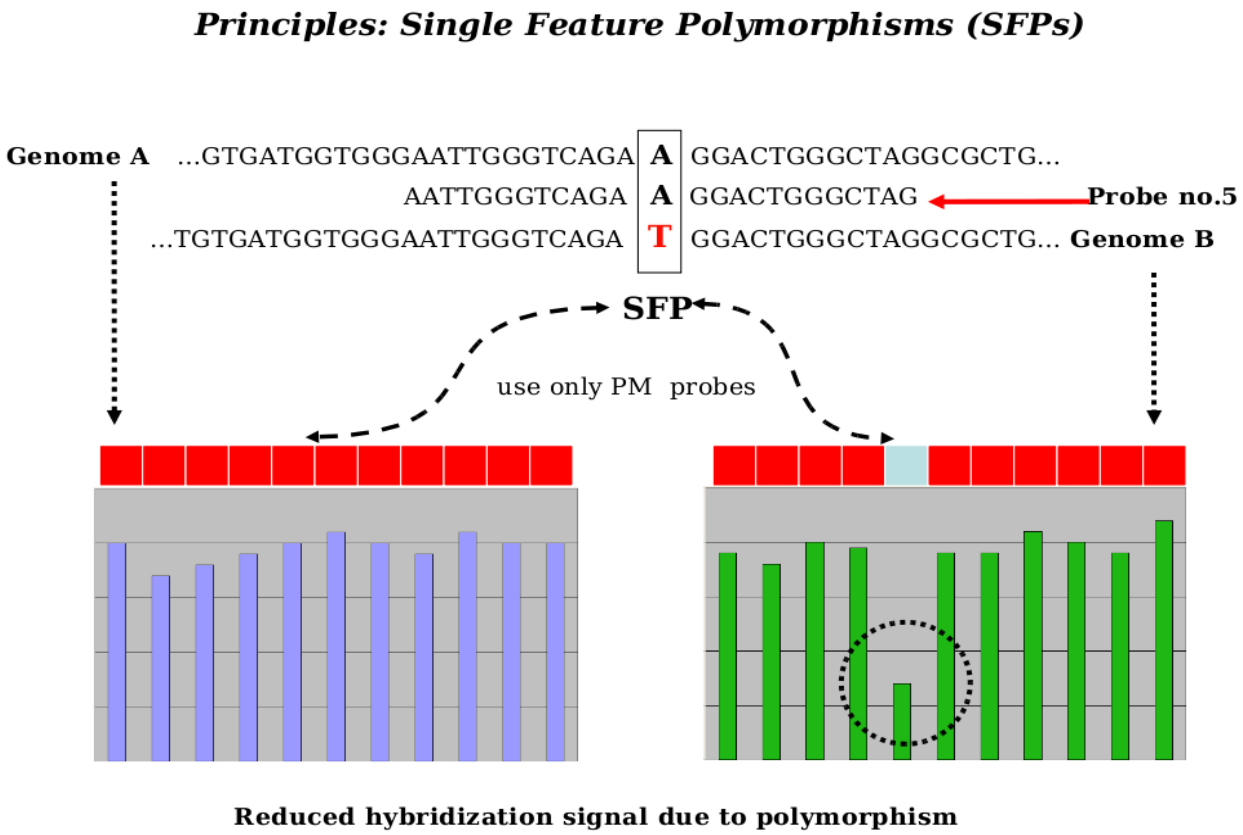
1.2.2. Algorithms available for polymorphism detection from gene expression data

A variety of approaches and algorithms have been implemented for quantifying differences in microarray hybridization signals, in order to identify potential SFPs. One of the methods used often in the various SFP studies involves the Significance Analysis of Microarrays statistic (SAM, Tusher et al. 2001). This method (used for SFP detection by Borevitz et al. 2003, Rostoks et al 2005, Kumar et al. 2007) assigns a score to each probe based on whether the signal difference observed between the experimental lines being compared, is significant relative to the standard deviation of repeated measurements within a given line. For probes with a score greater than a user-selected threshold, the difference between the lines is declared statistically significant. This method also uses permutations of the repeated measurements, in order to estimate the false discovery rate (FDR). While following a rigorous statistical approach, the SAM method was originally designed to compare signal levels using single probe per gene microarrays (otherwise called cDNA-spotted microarrays). Because the method operates on a probe-by-probe basis, gene expression variation between different genetic lines, and not real polymorphisms, can be mistakenly identified as SFPs (Luo et al). This can be avoided when microarrays are used with more than one probe per gene, so that we can discriminate whether hybridization is altered for a single probe and therefore is due to a polymorphism, or whether all probes in a probeset have altered levels designating a difference in gene expression.

In this respect, a more sensitive method was implemented for the detection of SFPs in *A.thaliana* (West et al. 2006), and can distinguish reduced hybridization signal because of a sequence polymorphism, or variation of gene expression between different genetic lines. In this method, the SFPdev statistic was introduced, which is based on the difference of signal intensity of each probe from the average of a probe-set for the gene (**Fig. 1**). What this statistic essentially captures is single probe signal deviations from the baseline of the expression levels of a gene, originating due to sequence mismatches in the gene region of each probe target. Probes with significant deviations in only one of the two lines being

compared, are identified as SFPs. Therefore, SFPdev allows separation of variability due to sequence polymorphism and gene expression variation. In a study by Ronald et al. (2005), another type of approach for SFP discovery was proposed by using an energy model for the hybridization efficiency of the oligonucleotides on the array. This method estimates the expression of a gene by modeling the energy of probe-target sequence duplex formation, and the expected intensity (Î) for each probe is also determined. A polymorphism is declared when there is a significant difference between the two strains, for the ratio of the observed to the expected probe signal. The drawback of this method is that a reference strain is needed based on which the microarray is designed. Therefore, application of this method is limited to experiments for which data can be collected from the reference strain. Furthermore, since this method by Ronald et al. (2005) operates on a probe by probe basis, gene expression polymorphisms might be mistakenly identified as SFPs (Luo et al. 2007).

Fig. 1 Principles of Single Feature Polymorphism discovery



1.2.3. Considerations for polymorphism detection using expression data

Some of the pitfalls for SFP detection include transcript abundance in the tissue type used for the microarray assay, cross-hybridization between gene families and the dominance of the SFP markers. The problem with gene families can be especially apparent in complex plant genomes where extended genome duplications, lead to multiple copies of each gene. While these duplicated regions are altered significantly by evolution and also plant breeding, still functional domains specific to a gene family with enough sequence similarity remain. SFPs are considered dominant markers, a definition used for genetic markers where a polymorphism can be detected only when a locus is homozygous (review in Hawley and Walker 2003). In the case of conserved gene sequences across gene families, a homozygous locus can be masked by the homologous non-polymorphic sequence found in a member of the gene family across the genome. In other words, a polymorphism that would be otherwise identified due to the reduced hybridization to a probe, can be masked by transcripts from genes within the same gene family that do not carry the polymorphism and hybridize efficiently to the probe. Finally, in the case where mRNA is used for the identification of SFPs, since only a subset of all genes are expressed in a specific tissue, polymorphism detection level varies with the type of cells used for the microarray assays (Luo et al 2007).

1.3. Soybean populations, traditional markers versus SFP and genetic mapping

1.3.1. Soybean populations and genetic mapping

A genetic map is an assembly of genetic markers, positioned relative to each other based on their recombination frequencies (review in Hawley and Walker 2003). Genetic recombination is a process where sections are exchanged through a cross-over between the two copies of a chromosome in a diploid organism. The further apart are the loci containing the genetic markers, the higher is the chance for recombination to occur between them, and therefore a higher frequency of recombination is observed. In chromosomes containing heterozygous loci and therefore polymorphic alleles, crossovers result in “swapping” alleles between the two copies of the chromosome. By counting the number of recombinants in a population, and since there is higher chance for recombination between loci further

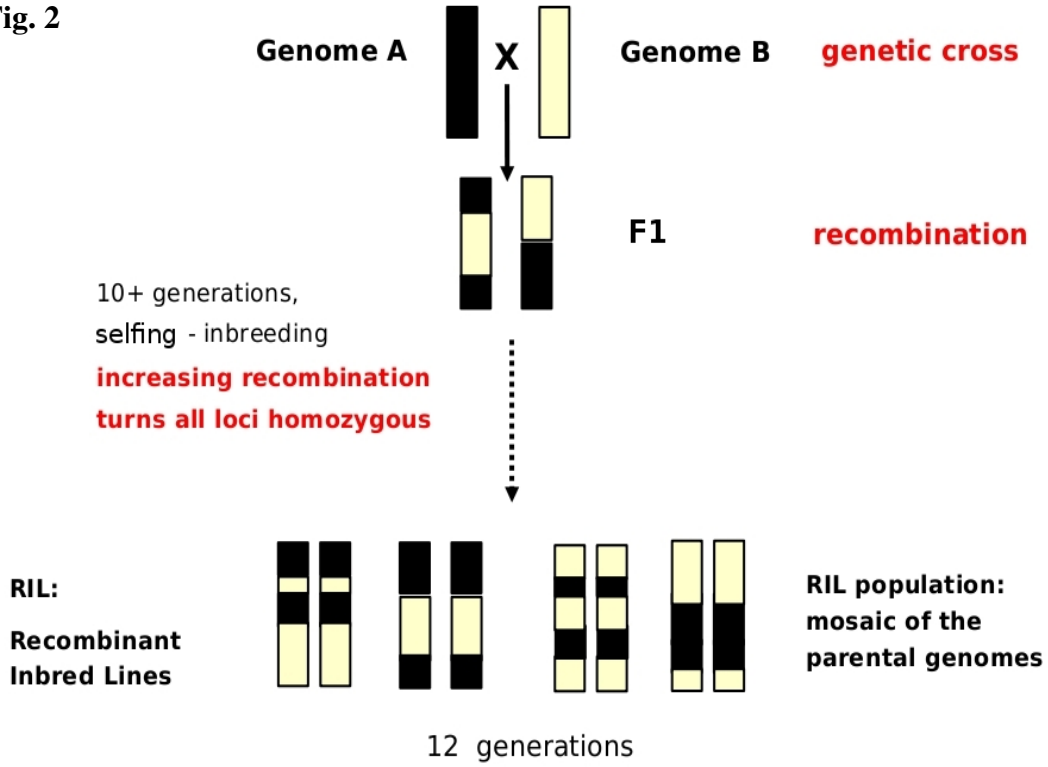
apart, it is possible to obtain a measure for the distance between the loci. This distance is measured using a genetic map unit called centiMorgan (cM), defined as the distance between genetic loci for which recombinant alleles are observed in 1 out of 100 individuals of the progeny. In other words, for loci that are 1cM apart we expect a frequency of 1% of the progeny to carry the recombinant alleles. In the case of multiple recombinations which cancel each other out, the observed frequency does not correspond to the true genetic distance. Finally, for genetic loci that are significantly far apart from each other on the same or different chromosomes, events of recombination between them are so often that are essentially random, and there are equal frequencies of recombinant versus non-recombinant progeny.

A genetic map is created by calculating the linear order between loci, based on recombination frequencies between them (review in Hawley and Walker 2003). The number of possible linear ordering of loci along a chromosome increases rapidly with the number of loci, since for N loci we have N! possible linear combinations. This makes genetic map construction difficult for large numbers of loci, when for example with N=10 genetic markers, we have 10!=3628800 possible linear combinations. Fortunately for constructing genetic maps in linear time, efficient heuristic algorithms such as Seriation, Branch & Bound and Local Reshuffling (Stam 1993) have been developed and implemented in modern genetic mapping software such as (JoinMap4.0, http://www.kyazma.nl/). Finally, across different organisms and even along the chromosomes of a single organism, the relationship of genetic distance (calculated based on the recombination frequencies) to the physical distance measured in DNA base pairs varies significantly. This is due to particular genomic regions where recombination takes place at high or low rate, such as for examples the centromere regions of the chromosomes.

The most commonly used populations for genetic mapping in soybean and plants in general are derived from the cross of two highly inbred, highly homozygous, and genetically distant parents (Dominique le Vienne, 2003). The resulting progeny from such a cross constitute the F1 generation, and have a hybrid diploid genome with one copy of a chromosome from the first parent, and the second copy from the other parent. Due to the genetic distance of the parents, the F1 individuals are highly heterozygous with

two distinct alleles in most loci, and similarly each allele coming from a different parent. Recombinat Inbred Line (RIL, Burr and Burr 1991) populations are obtained from randomly choosing a single seed from each individual progenies of the F1 generation. Each plant grown from the chosen seeds is self-pollinated (**Fig 2**.), and this process is repeated several more generations to reach homozygosity. Starting with two distinct alleles per locus in the F1 hybrids, during each round of selfing each genetic locus due to the meiosis taking place during the creation of gametes, has 50% chance of being advanced to the next generation and be fixed for a recombinant line. Therefore, heterozygocity is reduced by half in each generation, resulting in $1/2^{10} = 99.80\%$ of the loci being homozygous after 10 generations. Homozygous loci have the advantage that dominant alleles do not mask the recessive ones, whether in the case of trying to identify genetic polymorphisms or measuring phenotypic traits. Additionally, in parallel to the increased fixation of genetic loci in RIL populations during the inbreeding over the generations, recombination takes place between the different loci, resulting in multiple combinations of alleles across each chromosome (**Fig 2.**). The result is a population of genetically homozygous individuals for each locus, but with genomes that are a mosaic of the parental ones. Using the novel combinations of genetic loci within the RIL individuals, we can study their unique effects on the phenotype (Quantitative Trait Loci, QTL review in Doerge 2002). Finally, another advantage of an RIL population is that by continuous self pollination of the homozygous individuals, the population's genome can be preserved indefinitely, and therefore used for new mapping or other genetic experiments (Dominique de Vienne 2003).

Fig. 2



1.3.2. Traditional genetic markers in soybean genetic mapping

The first genetic map of soybean consisted of less than 63 isozyme, morphological and pigmentation genetic markers spanning several chromosomes (Palmer and Hedges 1993). Constructing chromosomal maps in early studies proved difficult as classic type of genetic markers are sparse, and even fewer reveal polymorphic loci, containing different alleles. Allele polymorphism is a requirement for genetic mapping, since genetic distance is calculated based on the frequencies of combinations of different alleles along the chromosomes (Hawley and Walker 2003). Subsequent genetic maps were constructed with molecular markers such as random amplified polymorphic DNA (RAPD) markers and restriction fragment length polymorphisms (RFLP, Shoemaker and Olson 1993). However, despite using the more precise molecular markers, the problem of low polymorphism persists when domesticated soybean lines of low genetic diversity are used. In addition, genetic mapping with these types of molecular markers can be cumbersome due to the multiple copies (multi-ploidy) of the soybean chromosomes.

This is due to single-base mutations occurring across the different copies of each chromosome, altering the restriction enzyme sites recognized by these markers and leading to erroneous output during the polymorphism screening (Cregan et al. 1999).

In order to eliminate the complexity of genetic map construction, simple sequence repeat (SSR or microsatellite) markers were developed (Cregan et al. 1999). These markers are based on the PCR amplification of tandem short-sequence repeats of 1-6 base pair length, usually within non-coding regions of the genome. SSRs are highly polymorphic molecular markers, and generally produce only one amplified product per chromosomal copy. Cregan et al. (1999) created a consensus map consisting of 606 SSRs, developed from three separate mapping populations, covering all 20 soybean chromosomes. An extension of this mapping study, incorporated additional Express Sequence Tag (EST) based SSRs to the consensus map (Song et al. 2004). The EST-derived SSR represent functional genes compared to conventional simple sequence repeat (SSR) markers, which are primarily based on amplifying non-coding regions of a genome.

In a recent study, PCR-based single nucleotide polymorphism (SNP) markers identifying single base changes and insertion/deletions within gene sequences, were shown to be more polymorphic than SSRs in soybean (Choi et al. 2007). The first SNP map of the soybean genome included SNPs from 1131 genes. Over a quarter of these genic SNP markers were found in gaps of the previous SSR genetic map of soybean, showing therefore great potential for complementing missing sections of the previous map. However, the SNP markers discovered comprised only a 21.5% fraction of all the examined sequences. This was partly due to the inability of amplifying the gene sequences, but also because of the low level of sequence variation in cultivated soybean. Moreover, most of the SNP markers were found to cluster in gene-rich regions along the chromosomes, therefore reducing their potential for use as genome landmarks since they are not widely dispersed on the genome (Choi et al. 2007).

1.4. Metabolic pathways and resistance to pathogen infection

Metabolism is the culmination of the genetic instructions carried by an organism, and is the means for interacting and persevering in its environment. The metabolic pathway diagram is also a blueprint that shows how the organism is adapted to environmental conditions such as nutrient availability, climate and potential pathogens (Lange and Ghassemian, 2005). Currently, the complete genome sequences of *A. thaliana* and *G. max* are available (http://www.tair.org and http://www.phytozome.org). The metabolic network of the model plant *A. thaliana* (Mueller et al. 2003), is one of the best studied in comparison to other plant species. On the other hand, while *G. max* is one of the most agronomically important plants, most research in this species has been geared toward genetic studies for crop yield improvement (Kumar 1999), resulting therefore in sparse data for its biological pathways.

Even when the metabolic pathway structure of a species is well known, still the real value comes from studying the pathways for how they operate temporally or against certain genetic backgrounds, leading to the manifestation of certain phenotypes (Lange and Ghassemian, 2005). One example of a phenotype of great importance for plant species is resistance against potential pathogens, where a series of structural and metabolic changes take place in the plant cells (Feys 2000). The cellular changes in this case, inhibit the multiplication and spread of the pathogen in the plant tissues. This is achieved mainly through the activation of the secondary metabolic pathways, where compounds are produced that reinforce the plant cell, but also chemical molecules with antibiotic activity or for defense signaling, such as salicylic acid or ethylene (Van Loon 1997).

A method for studying the patterns of the biological pathway perturbation under a certain biological condition, is by integrating pathway and gene expression data. Using computational methods the activation or repression of parts of metabolic network can be inferred from the changes in the transcriptional profile of the genes encoding the pathway enzymes. A newly developed algorithm towards this goal, is the Pathway Perturbation Algorithm (Corban Rivera 2009, in preparation for submission). This algorithm calculates perturbation of a biological pathways, based on contrasts of gene expression levels for the genes in adjacent steps of the pathway. In addition, this method does not

require every gene in a pathway to be differentially expressed. Conseqeuently, pathway perturbation can be detected when not all genes in the pathway alter their expression levels in response to a treatment, such as when biological modifications take place at the post-transcriptional or post-translational level.

1.5. References

Borevitz JO, Liang D, Plouffe D, Chang H-S, Zhu T, Weigel D, Berry CC, Winzeler E and Chory J (2003). "Large-scale identification of single-feature polymorphisms in complex genomes." Genome Res. **13**: 513-523.

Cregan PB, Jarvik T, Bush AL, Shoemaker RC, Lark KG, Kahler AL, Kaya N, VanToai TT, Lohnes DG and Chung J (1999). "An integrated genetic linkage map of the soybean genome." Crop Sci. **39**: 1464-1490.

Cui X, Xu J, Asghar R, Condamine P, Svensson JT, Wanamaker S, Stein N, Roose M and Close TJ (2005). "Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit." Bioinformatics **21**: 3852-3858.

Dominique le Vienne (2003) Molecular Markers in Plant Genetics and Biotechnology. Science Publisher. NY.

Doerge R. (2002) “Mapping and Analysis of Quantitative Trait Loci In Experimental Populations” Nature Rev Gen. **3**: 43-52.

Feys BJ and Parker JE. (2000) “Interplay of signaling pathways in plant disease resistance” Trends in Gen. **16**: 449-455.

Gonzales MD, Archuleta E, Farmer A, Gajendran K, Grant D, Shoemaker RC, Beavis WD and Waugh ME (2005) “The Legume Information System (LIS): an integrated information resource for comparative legume biology” Nucleic Acids Res. **33**: 660-665.

Hawley RS and Walker M.Y. (2003) Advanced genetic analysis, Blackwell Publishing, New York

Kumar LS (1999) “DNA markers in plant improvement: An overview” Biotechnol Adv. **17**: 143-182.

Lange BM, Ghassemian M. “Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps” (2005) Phytochemistry 66: 413-451.

Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO and Davis RW (1997). "Yeast microarrays for genome wide parallel genetic and gene expression analysis" Proc Natl Acad Sci USA **94**: 13057–13062.

Luo ZW, Potokina E, Druka A, Wise R, Waugh R and Kearsey MJ. (2007) "SFP genotyping from affymetrix arrays is robust but largely detects cis-acting expression regulators" Genetics **176** :789-800

Mueller LA, Zhang P, Rhee SY. (2003) "AraCyc: A Biochemical Pathway Database for Arabidopsis" Plant Physiology **132** :453-460

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D and Brown PO (1999) "Genome-wide analysis of DNA copy-number changes using cDNA microarrays" Nat Genet **23**: 41–46.

Ronald J, Akey J, Whittle EN and Smith G (2005). "Simultaneous genotyping gene-expression measurement and detection of allele-specific expression with oligonucleotide arrays." Genome Res. (15): 7.

Rostoks N, Borevitz J, Hedley P, Russell J, Mudie S, Morris J, Cardle L, Marshall D and Waugh R (2005). "Single-feature polymorphism discovery in the barley transcriptome." Genome Biology **6**: R54.

Somssich I, Hahlbrock K. (1998) Pathogen defense in plants - a paradigm of biological complexity. Trends Plant Sci. **3**:86–90.

Song QJ, Marek LF, Shoemaker RC, Lark KG, Concibido VC, Delannay X, Specht JE and Cregan PB (2004) "A new integrated genetic linkage map of the soybean" Theor. Appl. Genet. **109**: 122-128

Stam P (1993) "Construction of integrated genetic linkage maps by means of a new computer package JoinMap" Plant J. **3**: 739--744

Tusher VG, Tibshirani R and Chu G (2001). "Significance analysis of microarrays applied to the ionizing radiation response." PNAS **98**: 5-11

Van Loon LC. (1997) "Induced resistance in plants and role of pathogenesis-related proteins." Eur. J. Plant Pathol. **103**: 753–765.

Van Oeveren J (2007). Co-dominant SFP genotyping in tomato. Intelligent Systems for Molecular Biology. Vienna, Austria.

Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ and Davis RW (1998). "Direct allelic variation scanning of the yeast genome." Science **281**(5380): 1194-1197.

2. Mining gene expression data for Single Feature Polymorphism discovery and genetic map construction in soybean

2.1. Abstract

Single feature polymorphisms (SFPs) which are based on gene expression profiling data have recently provided an abundant new genic marker class in several plant species. As soybean (*Glycine max* L. Merr) has undergone several genetic bottlenecks through its domestication, identifying polymorphic, gene-based markers often proves difficult. In the current study, we assessed the levels of Single Feature Polymorphisms (SFPs) in a genetically diverse set of seven cultivated *G. max* lines, one *G. max* plant introduction and a *G. soja* line. Two transcript maps spanning all 20 soybean molecular linkage groups (MLGs) containing 941 and 1897 SFPs respectively were also created in an inter-specific recombinant inbred line (RIL) population. Polymorphic levels of SFPs were similar to these of conventional markers, as comparisons between well adapted lines found only low levels of polymorphisms, while wide crosses involving plant introductions or *G. soja* had higher levels. The newly created SFP map also included 113 simple sequence repeat (SSR) markers, and using these as anchoring points it was compared to the available public soybean maps for marker coverage and gap sizes. Linear order of the gene-based SFP markers were validated with the *G. max* whole genome sequence of Williams82. The algorithms developed in the current study allow large-scale data mining, and can be used for SFP discovery in other plant species with segregating RIL populations. This is the first report of the SFP methodology applied in soybean offering a new marker type based on annotated genic sequences.

2.2. Introduction

Various molecular markers have been utilized to construct genetic linkage maps of soybean (*Glycine max* L. Merr), enhancing the ability of plant geneticists to improve agronomically important traits via marker-assisted selected (MAS) programs. The first genetic maps of soybean consisted of restriction fragment length polymorphism (RFLP), isozyme, and morphological markers spanning over 26 molecular linkage groups (MLGs) and consisting of fewer than 150 markers (Shoemaker and Olson 1993, Shoemaker and Specht 1995). Eliminating the complexity of the MLG construction, simple sequence repeat (SSR) or microsatellite markers provided highly polymorphic, genome specific and generally allowed for only one amplified product per soybean genotype. Cregan et al. , using data from three separate mapping populations, created a consensus map consisting of 606 SSRs covering all 20 soybean MLGs. Additional studies by Song et al. combined five mapping populations to create a genetic map spanning 2523.6 cM with 1,015 SSR, adding 12 to 29 additional SSR markers per MLG in attempts to fill gaps.

Expressed sequence tags (ESTs) represent functional genes, compared to conventional SSR markers which are primarily based on amplifying non-coding regions of a genome. A large database consisting of over 394,000 ESTs (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html) in soybean, is currently available providing a valuable recourse assisting in the development of new genic-based markers. Recently, single nucleotide polymorphism (SNP) markers based on single base changes and insertion/deletions in genic or EST sequences, have been shown to be more polymorphic than SSR markers in soybean. Choi et al. created this first transcript map of the soybean genome based on mapping SNPs from 1131 ESTs by utilizing 9459 primer pairs. A high proportion of these genic SNP markers mapped in gaps (>5 cM) in the SSR consensus map. Albeit, there was only 21.5% discovery rate for polymorphism overall, suggesting a low level sequence variation in cultivated soybean.

In contrast, highly polymorphic genetic markers called Single Feature Polymorphisms (SFPs), have recently been discovered using gene expression data in several plant species including *Arabidopsis* , barley, rice, tomato and cowpea (Borevitz et al. 2003 and West et al. 2006, Cui et al. 2005 and Rostoks

et al. 2005, Kumar et al. 2007, Oeveren et al. 2007, Das et al. 2008 respectively). The SFPs are identified based on reduced hybridization on gene-chip arrays (microarrays), as a result of sequence mismatches due to either SNPs or insertion/ deletions. The first application for SFP discovery in a plant genome was by Borevitz et al. (2003), who identified approximately 4000 SFPs in *Arabidopsis* in a comparison between strains Columbia and Landsberg *erecta* (L*er*), using Affymetrix GeneChip arrays. This type of microarray relies on the use of short oligonucleotide probes (25-mers) with multiple probes per gene, and is more suitable for identifying SFPs compared to the single probes found on spotted oligonucleotide cDNA arrays, which are less sensitive for detection of sequence mismatches. Similarly, West et al. identified and mapped 595 SFP markers in an *Arabidopsis* recombinant inbred line (RIL) population consisting of 148 individuals, with the resulting genetic map having a marker density on average 0.64 cM. This study also incorporated 38 SSR markers in addition to SFP markers, while the map was further validated using the *Arabidopsis* genome sequence. In regards to SFP discovery in plant species with larger and more complex genomes, Kumar et al. (2007) identified approximately 5,000 SFPs in rice, while Rostoks et al. (2005) identified double that number of SFPs between barley cultivars Morex and Golden Promise, both using Affymetrix microarrays. Recently over 1000 SFP markers were identified in the legume cowpea (*Vigna unquiculata*, Das et al. 2008), providing therefore a valuable genomic recourse in a species where few high density genetics maps are available.

A variety of algorithms have been implemented to date in the various SFP studies, in order to quantify the differences in hybridization intensities used for identifying potential SFP probes. In plant genomes specifically, one such method is the Significance Analysis of Microarrays (SAM, Tusher et al. 2001, applied by Borevitz et al. 2003, Rostoks et al 2005, Kumar et al. 2007 in *Arabidopsis*, barley and rice respectively). SAM asserts whether the difference observed between genetic lines is significant, relative to the standard deviation of repeated measurements within the same line. A more refined algorithm applied in *Arabidopsis* by West et al., improves over SAM by examining the ratio of the intensity for each probe to the overall gene-expression of the gene that the probe represents The statistic used by West et al. is called *SFPdev* and refines SFP detection by being able to distinguish between an altered intensity signal due to poor hybridization because of a sequence polymorphism, and a low-intensity signal due to low gene expression.

Currently, there are no studies involving discovery of SFP polymorphisms in soybean or any other *Glycine* species. In our study, we have assayed for SFPs a set of seven cultivated *G. max* lines, one *G. max* plant introduction and a *G. soja* line, in order to test the ability of the new marker to detect polymorphisms between various soybean lines. We also evaluated the segregation of SFPs in a RIL population generated by a cross of *G. max* by *G. soja*. Two genetic maps containing 941 and 1897 SFP markers were constructed in this population, spanning all 20 MLGs of soybean. Incorporating 113 publicly available SSR with SFP markers allowed integration and comparison of the new SFP genetic maps with publicly available maps. The SFP genetic maps were verified using the recently sequenced genome of Williams82 cultivar. Finally, we developed new software and implemented algorithms for the SFP discovery, which can be utilized for isolating polymorphisms in gene expression datasets from other segregating soybean and other plant species populations.

2.3. Results

2.3.1. SFP discovery in soybean lines

Seven *G. max* lines, one plant introduction (PI) of *G. max* and one *G. soja* accession, were used to identify SFPs and examine levels of polymorphism detected between these genetically similar and diverse lines. We used the SFPdev Min-Max Ratio algorithm (West et al 2006) which is based on the *SFPdev* statistic, an absolute difference of hybridization intensity values of each probe from the average of the probeset, divided by the probe value. *SFPdev* values are higher for polymorphic probes, as reduced hybridization intensity results in deviation from the average of the probeset. We declare a probe as SFP when the ratio of *SFPdev* in a pair of compared lines is greater than 2. This is an empirical threshold reported during a first implementation of the algorithm (West et al. 2006) and also was verified while applying this algorithm to our data. The numbers of SFPs discovered for the soybean lines is shown in **Table 1**. The 36 values below the diagonal of **Table 1**, correspond to genes for which transcripts from the lines in the table rows are polymorphic and have reduced hybridization intensity compared to those in the table columns (and vice-versa for the 36 values above the diagonal).

	Athow	Con-rad	Gen-eral	Ox20-8	PI291-237	Sloan	V71-370	William s	PI407-162
Athow	-	575	546	939	822	821	843	856	6298
Conrad	638	-	662	849	938	856	903	1100	7066
General	572	640	-	903	1064	935	956	1150	6140
Ox20-8	1521	1270	1491	-	898	626	1621	1280	6339
PI291237	1410	1346	1516	900	-	880	1490	1330	6523
Sloan	1374	1288	1424	614	791	-	1569	1254	6332
V71-370	1061	1089	1080	1232	1065	1178	-	1478	6181
Williams	893	1002	1014	763	744	681	1383	-	7181
PI407162	4421	5410	4513	6394	6403	6715	5244	5210	-

Table. 1. Number of SFPs discovered comparing a set of genetically seven diverse cultivated soybean lines, a plant introduction and a *G.soja* line.

The number of SFPs from **Table 1** were used as a genetic distance metric between the soybean lines, and hierarchical clustering was performed (details in Methods section) resulting in the dendrogram of **Fig.1a.**. The *G.sojae* PI407162 plant introduction was separated from the rest of the *G.max* lines (**Fig.1a.)**, and that also is apparent from the high number of polymorphisms this line has on **Table 1**. Two distinct groups are present on the dendrogram on **Fig.1a.**, one containing the Conrad, Athow and General lines, while the other is composed of the PI291237, Ox20-8 and Sloan. This grouping is also apparent from the number of polymorphisms between the seven *G. max* lines and PI291237 *G.max* shown on **Table 1**; Sloan, is highly diverse with Athow, Conrad and General (1374, 1288 and 1424 SFPs respectively) while it has fewer number of SFPs with Ox20-8 and PI291237 (614 and 791 respectively). The Ox20-8 line and PI291237 have high number of SFPs with the same lines above as Sloan. Within the group of Athow, Conrad and General, there are fewer polymorphisms and thus low genetic diversity when compared with each other (a maximum number of 662 SFPs). Similarly with above, the number of SFPs increases when Athow, Conrad and General are compared with Ox20-8, PI291237 or Sloan. Finally, the V71-370 and Williams *G.max* lines are placed in between these two groups on **Fig.1a.**. We also can see this from the number of polymorphisms they display on **Table 1** when they are compared for example with General from the first group (1080, 1014 SFPs with V71-370 and Williams respectively) or with Ox20-8 (1232, 763 SFPs) from the second group.

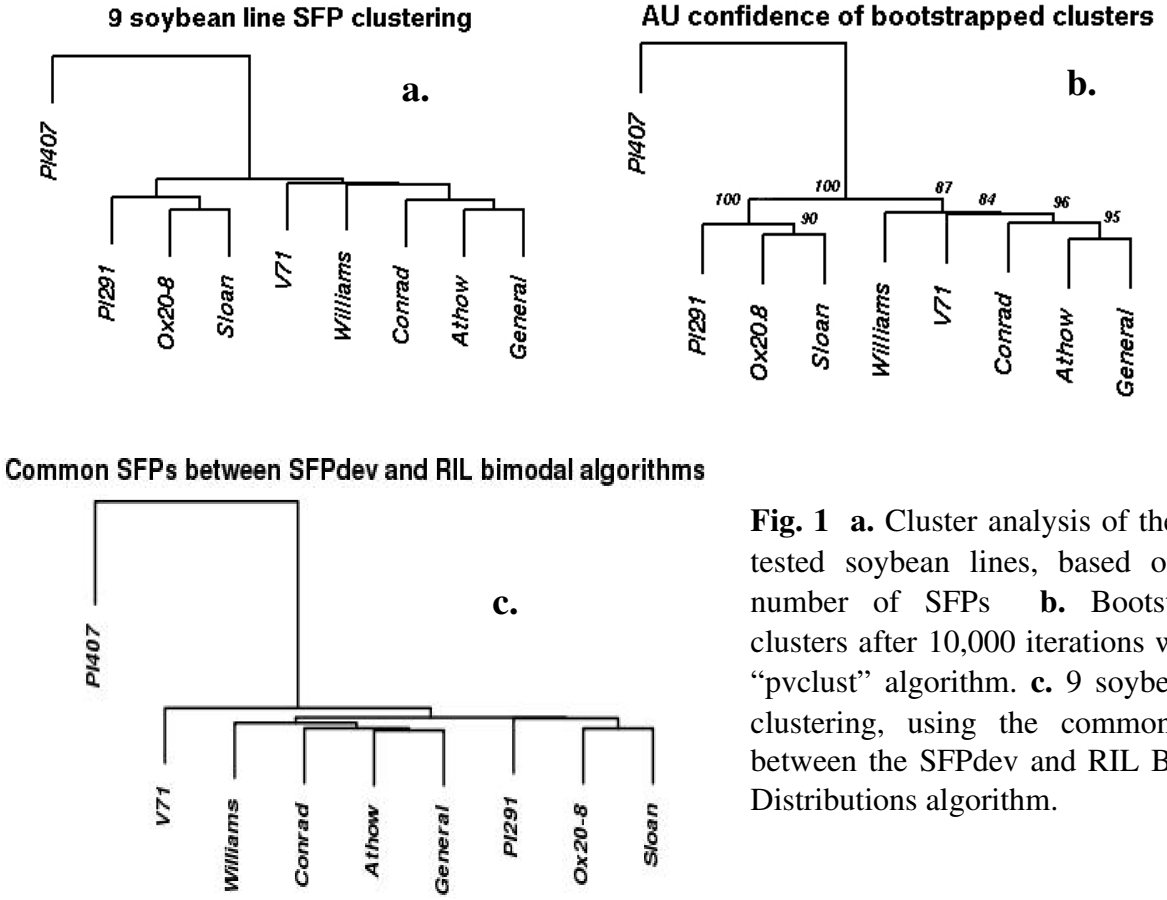


Fig. 1 **a.** Cluster analysis of the set of tested soybean lines, based on their number of SFPs **b.** Bootstrapped clusters after 10,000 iterations with the "pvclust" algorithm. **c.** 9 soybean line clustering, using the common SFPs between the SFPdev and RIL Bimodal Distributions algorithm.

In order to verify the clusters of the dendrogram on **Fig.1a.**, we used the "pvclust" method (Suzuki and Shimodaira 2006, details in the Methods section). This method in addition to verifying clusters by bootstrap resampling, calculates the Approximately Unbiased (AU) p-value, which indicates how strongly each cluster is supported by data. Following 10,000 iterations, we plotted the bootstrapped clusters along with their p-values on **Fig.1b.**, and examined the new dendrogram for significant clusters looking for branches with AU greater than 95% (0.05) confidence . The branch where the PI407162 plant introduction is separated from the rest of the soybean lines was found to have 100% AU on **Fig.1b.**, while the same is true at the branching point of PI291237 with Ox20-8 and Sloan. Another point on the right section of the tree at 96% AU, splits Williams82 and V71-370 from Conrad, Athow and General. These observations validate the grouping of soybean lines based on the number of SFPs

mentioned in the previous section, and additionally the bootstrapped dendrogram matches the one constructed with the original data **Fig.1a** .

For further validation of the clusters formed between the 9 soybean lines, we re-created the dedrogram using the intersection of genetic polymorphisms from **Table 1**, with the SFPs discovered in the Recombinant Inbred Line population (RIL, see subsequent section for details). While this population was generated by a cross of PI407162 and V71-370 and consequetively it contains only SFPs present for these lines, the sample size of the RIL population provides better statistical significance. Therefore, despite the SFPs found in the intersection of the results (see **SI Fig. S2** for detail) are biased towards these two soybean lines, we expect these data to contain minimum false positives. The intersection of SFPs from **Table 1** and the RIL population was used to construct the dendrogram on **Fig.1c**. Similarly with **Fig.1a**, the Conrad, Athow and General lines form a cluster, another cluster is composed by the PI291237, Ox20-8 and Sloan, while PI407162 is on a separate branch again due to its high number of polymorphisms. A minor difference between the two dendrograms concerns the V71-370 line, which clusters outside of these groups. This is because of the increased number of polymorphisms from V71-370 (see **SI Fig. S2** for details), which is an artifact since the intersection of the SFP datasets are biased towards the PI407162 and V71-370 lines.

Considering the two groups on **Fig.1a**., lines V71-370, Williams, Conrad, Athow and General display partial resistance to the *Phytophthora sojae* pathogen (data not shown), while PI407162, PI291237, Ox20-8 and Sloan are susceptible. It has been known from another study (Tucker et al. 2008, submitted) that a major disease-resistance QTL is located close to marker SATT529 on chromosome J (see also **Fig.2** for genetic map in subsequent section). For the SFPs located close to this marker , we found low polymorphism between the soybean V71-370, Williams, Conrad, Athow and General lines that have partial resistance to *P.sojae*. Additionally, the 83242.S1_2 SFP probe which maps 2.2cM above SATT529, is polymorphic between the PI407162 *G.soja* and all the *G.max* lines. This probe showed no polymorphism in any of the comparisons between the *G.max* cultivars, and therefore represents a unique allele on the *G.soja* germplasm. Furthermore, we found the transcript for the 83242.S1_2 SFP probe having homology to an *Arabidopsis* gene (AT3G21600) which is annotated

function related to senescence. Another probe 1650.1.S1_11 was found mapping 5.3cM above SATT529 on linkage group J (**Fig.2**), which was polymorphic in comparisons between Athow, Conrad, General, V71-370 and Williams (lines with partial resistance) with either of the Ox20-8 or Sloan (susceptible lines). In contrast, this probe was not found to be polymorphic when any of the Athow, Conrad, General, V71-370 and Williams are compared with each other, and similarly for comparison between Ox20-8 and Sloan. The homologous *Arabidopsis* gene (AT5G40370) to the 1650.1.S1_11 probe is annotated as a glutaredoxin.

2.3.2. SFP discovery in an interspecific cross of *Glycine max* by *Glycine soja*.

We then identified SFPs using the data for 293 F_{12} generation Recombinant Inbred Line (RIL) individuals, from an interspecific cross of V71-370 (*G .max*) and PI407162 (*G. soja*). We implemented an algorithm that finds bimodal distributions (West et al. 2006) in sets of *SFPdev* values, calculated from hybridization intensities for each probe across all individuals of the RIL population. The algorithm operates on the assumption than an *SFPdev* distribution has two modes when one parent has a polymorphic allele, which results in reduced hybridization affinities in the progeny inheriting this allele. The RIL bimodal distributions algorithm assigns genotypes to each RIL according to which parental value is positioned within the mode of the distribution the RIL is located (SI **Fig. S3** and **S4**). In order to assess the adequate separation between the two distribution modes the peak separation (*ps*) statistic is used, which is a metric incorporating the mean, variance and size of each distribution.

Numbers of detected SFPs at varying *ps* values using the above algorithm are summarized in **Table 2**. At lower thresholds of *ps* 30 and *ps* 50, more SFP polymorphisms are detected overall and on a per gene basis. We retained only the SFPs that have been successfully assigned parental genotypes, where SFPdev values calculated from the parental data were successfully positioned within the modes of the RIL distribution. At *ps* 30 cutoff, 12% of the SFPs could not be assigned a parental genotype, while the fraction with unassigned genotypes for *ps* 50 was 1%. The 1254 discovered SFPs at this cutoff were subsequently used for the map construction. In order to further ensure the quality of the

polymorphisms incorporated in the genetic map, we refined this set of SFPs and retained only 941 where only a single probe from each probeset was polymorphic.

*ps**	Total SFPs	SFP genes	SFPs / gene	SFP genes with assigned genotypes	% SFP genes without genotypes
> 80	208	194	1.1	190	3%
> 65	646	516	1.25	494	5%
> 50	1768	1254	1.4	1243	1%
> 30	7341	3874	1.9	3443	12%

Table 2. Number of SFPs detected at different peak separation thresholds using the expression data from the RIL population (ps = peak separation)

2.3.3. Map construction using SFPs in the RIL population

In order to identify each of the twenty soybean MLGs, the 941 SFPs from *ps* 50 were combined with 113 publicly available SSR markers of known order and distances on the MLGs. The SSR markers were also used to validate the newly constructed SFP maps according to their location in the publicly available soybean maps, and were selected at approximately 30 cM intervals and also to span the very distal ends of each linkage group. Data for all the SFP and SSR markers were loaded in JoinMap 3.0 software (Van Ooijen and Voorrips 2001), and genetic maps were generated following the marker grouping on the twenty soybean MLGs (**Fig. 2**, see also Methods section for details). A Chi square test (p=0.05) was run on all markers to examine any possible segregation distortion in the interspecific cross. Approximately 5% of all SSR and SFP markers mapping to MLG-I, -D2 and -D1B deviated from the expected 1:1 genotypic ratio, suggesting possible selection at these loci during earlier generations of the population. The markers on MLG-D1B are located near the *ms* locus conferring male sterility, suggesting this as a plausible reason for segregation distortions in the genotypic data.

SSR order and location in these newly constructed MLGs were consistent to publicly available maps (Song et al. 2004, Choi et al. 2007), with only slight arrangements of closely linked SSR markers. SFPs

did not appear to be extending or grouping several MLGs together, and the SSRs mapping at the ends of each MLG did not have SFPs further above or below them. Satt195 and Satt383 for example are two markers on MLG-A2 and MLG-O respectively, which are the most distal markers on all available public maps and also appear as such on the SFP MLGs. In cases where the upper or lower most SSR marker in a MLG was not polymorphic between the parental lines used in our study, or a SSR marker has not been mapped to these locations, mapped SFPs extending the MLG beyond the most distal SSR markers were compared to public maps. Satt163 on MLG-G was the uppermost polymorphic SSR marker, however seven SFP markers mapped above this position extending the MLG by 10 cM. A SNP marker (BARC-020027-04405) reported by Choi et al. also maps 6 cM above Satt163 on MLG-G, verifying the observation on our SFP map. The length from our MLGs spans a total 2572.4 cM, and compared to the 2550.3 cM size of the public map (**Table 3**) the difference is only 22.1 cM, showing good correlation between the two. As an example, in both the public and SFP maps MLG-A2 was the longest and MLG-J was relatively one of the shortest.

Since SFP markers correspond to SNP or insertion/deletions in gene sequences, to further verify our map we performed a comparison with the number of SNPs from the transcript map by Choi et al. in each linkage group (**Table 3**). This map was created by mapping SNPs derived from EST sequences of 1141 genes, and also combining map data from three separate RIL populations. Significant correlation was observed between the numbers of SNPs and SFPs in the various MLGs. More specifically, MLG-A2 had the highest number of markers in both studies with 78 SNPs and 69 SFPs. Choi et al. reported MLG-E, -J and -K had a significantly higher gene density than expected while a smaller number of genes mapped to MLG-M. Similarly, assuming SFP markers are randomly distributed throughout the genome based on its size at a density of one marker every 2.7 cM, the predicted number of SFPs for each linkage group is shown in **Table 3**. Both the actual and predicted number of SFPs on our maps was also high for MLG-E and -J , but MLG-K had lower than expected gene discovery level. Gene poor MLGs in the SFP maps were MLG-B1, -D1A and -D2.

Table 3. Existing Marker Linkage Groups (MLGs) lengths of Song et al. with preexisting molecular markers (SSR, RFLP, other) and new SNP markers present in the soybean consensus map, compared to newly constructed SFP genetic map.

Public Maps				New SFP map length including SSR markers				
MLG	cM	RFLP, SSR etc. (Song et al.)	SNPs (Choi et al.)	Length (cM)	SFPs (ps50 / ps30)	Ex-pected SFPs *	Gaps 5-10cM † (Song / ps50) (Choi / ps30)	Gaps >10cM † (Song / ps50) (Choi / ps30)
A1	102.3	87	53	113.9	50 / 67	42	(4/1) (0/0)	(0/1) (0/0)
A2	165.7	116	78	159.3	69 / 99	59	(0/3) (3/1)	(0/1) (0/0)
B1	131.8	71	47	146.7	30 / 59	54	(3/8) (2/4)	(2/3) (2/0)
B2	125.0	88	47	117.9	38 / 70	44	(2/7) (0/2)	(0/0) (0/0)
C1	136.1	70	51	123.2	55 / 107	46	(1/5) (0/2)	(0/0) (0/0)
C2	157.9	100	53	153.5	55 / 112	57	(4/4) (4/2)	(0/2) (0/0)
D1A	120.9	101	45	127.7	27 / 93	47	(2/4) (0/0)	(0/3) (0/0)
D1B	138.0	81	57	154.6	48 / 104	57	(6/5) (3/0)	(0/2) (0/0)
D2	140.9	87	77	153.8	40 / 109	57	(6/7) (2/0)	(0/4) (0/0)
E	71.3	103	70	128.0	55 / 117	47	(3/6) (2/0)	(0/0) (0/0)
F	151.4	113	72	121.0	54 / 92	45	(6/4) (6/0)	(0/1) (0/0)
G	126.9	129	68	119.4	55 / 118	44	(0/2) (1/0)	(0/1) (0/0)
H	124.0	84	48	112.8	33 / 89	42	(3/4) (1/1)	(0/3) (0/0)
I	120.9	76	54	109.0	35 / 89	40	(4/4) (3/0)	(1/3) (1/0)
J	91.2	98	74	110.1	53 / 86	41	(0/6) (0/3)	(0/0) (0/0)
K	120.0	92	69	128.7	39 / 79	48	(1/6) (0/1)	(0/2) (0/0)
L	117.0	99	41	111.4	45 / 77	41	(3/5) (1/1)	(0/0) (0/0)
M	146.3	78	39	131.7	53 / 120	49	(8/6) (3/1)	(0/1) (0/0)
N	117.2	82	48	119.0	59 / 95	44	(2/2) (3/2)	(2/1) (0/0)
O	146.4	93	51	130.7	48 / 115	48	(7/8) (6/0)	(0/1) (0/0)
Total	2550.3	1848	1141	2572.4	941/ 1897	952	(65/97) (40/20)	(5/29) (3/0)

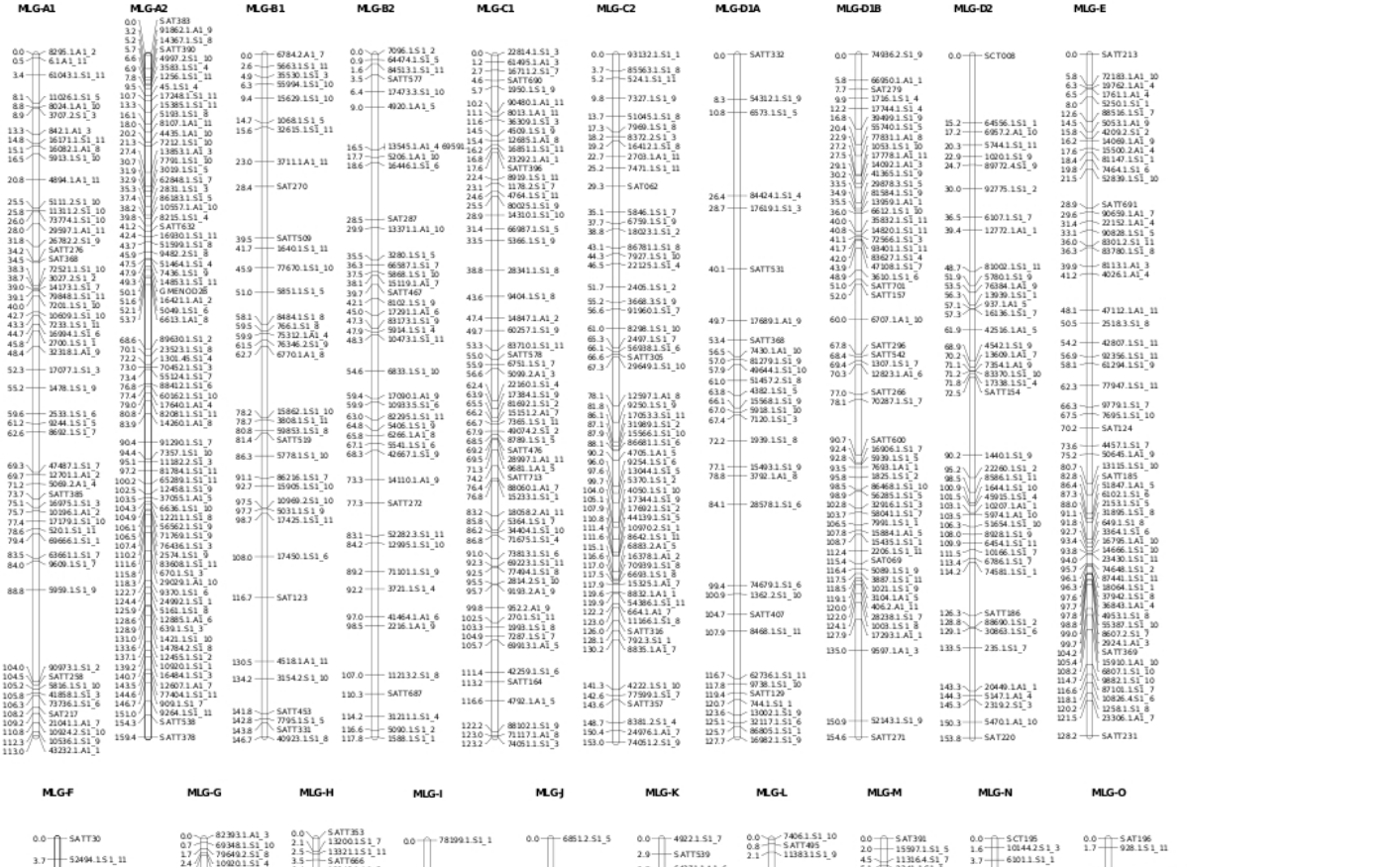
* Number of SFP markers per linkage group assuming random distribution and based on each MLG size.
† In the left paranthes are the number of gaps found on the soybean public map by Song et al. versus the SFP ps50 map, while on the right are the gaps on the SNP map by Choi et al. versus the SFP ps30 map.

2.3.4. Gap-filling in the SFP map and marker density compared to the public maps.

In order to identify more SFP markers in the soybean linkage groups, we saturated our map by adding markers discovered at the lower *ps* 30 threshold. Using the 3874 SFPs at this threshold (**Table 2.**) we followed an iterative procedure in JoinMap3.0: first we added markers in each MLG until the gaps were filled, and then removed those that distorted the marker order or length of the linkage groups. Including SSRs the updated *ps* 30 SFP map contained 1897 markers, and their distribution across linkage groups is also shown on **Table 3** and the maps are included within S5 .

We then compared the *ps* 50 and *ps* 30 SFP maps with the RLFP-SSR map by Song et al., and the transcript-SNP map by Choi et al. Although these maps have different marker types than our SFP map, using SSRs as anchor points allowed a direct comparison of distribution and sizes of gap regions across the different linkage groups (**Table 3**). For the *ps* 50 SFP map, smaller gaps of size 5-10cM are more frequent in our map compared to the one by Song et al., with only a few exceptions in MLGs A1, D1B, F and M**.** Similarly, our *ps* 50 SFP map has gaps greater than 10cM throughout the various MLGs, while on the Song et al. map they are present only on the B1, I and N linkage groups. We also compared our *ps* 30 SFP map to the SNP map by Choi et al. on **Table 3**. Both our *ps* 30 map and the map by Choi et al. use transcript-based markers, while also the Choi et al. map was constructed in order to fill the gaps in the map by Song et al. In this respect, most gaps on the *ps* 50 map were filled with SFP markers from the lower *ps* 30 threshold. More specifically, for smaller gaps of 5-10cM, they are less per MLG in the *ps* 30 SFP map, and overall they are $1/5^{th}$ of from the *ps* 50 one and also half in number from the Choi et al. map (20 versus 40). For larger gaps of greater than 10cM, none remained on the *ps* 30 SFP map, and even the ones present on the B1 and I MLGs on Choi et al. map were filled as shown on **Table 3**.

Fig. 2. Newly constructed SFP maps of the interspecific RIL of *Glycine max* (V71-370) by *Glycine soja* (PI407162).



2.3.5. Correlation of SFP genetic map with Williams82 physical sequence.

We verified the newly constructed SFP map, by calculating the linear correlation of the SFP order on the MLGs, with their location on the Williams82 genome sequence. This was achieved through a BLAST comparison of the ESTs represented by SFPs, to the twenty genomic sequence scaffolds assembled to the soybean linkage groups (http://www.phytozome.org). Due to the high degree of duplication in the soybean genome, we used a strict threshold of E-value of 1e-50 for the BLAST search. We then calculated the correlation of the positions of the BLAST hits on the Williams82 sequence, to the distances of the SFPs on our map. The results are summarized on **Table 4.**, and overall there was good correlation in both the *ps* 50 and *ps* 30 maps, ranging from to 0.73 to 0.96 and 0.71 to 0.93 respectively. Only MLG-B1 had a low correlation on the *ps* 30 map, but after close inspection we found the problem being due to the SFP from the unigene Gma.7780.1.A1. This marker was introduced from the *ps* 30 set in order to fill a gap on this linkage group, and was positioned at approximately 17cM. Based on the BLAST result this unigene contains many dispersed hits on scaffold Gm11 for MLG-B1. More specifically, the annotation for Gma.7780.1.A1 is for heat-shock protein, and possibly the multiple hits are due to protein domain similarities with members of this family dispersed along the scaffold. When this problematic marker was excluded from the calculations, the correlation was 0.8 for MLG-B1. Finally, on **Table 4** there is negative correlation for MLG-E, which means that the genetic order of markers on this linkage group is reversed, relative to the position on the physical sequence of their corresponding ESTs.

		SFP maps ps > 50		SFP maps ps > 30	
MLG	Scaffold	R	R^2	R	R^2
A1	Gm05	0.89	0.79	0.91	0.83
A2	Gm08	0.94	0.89	0.94	0.88
B1	Gm11	0.87	0.75	(**0.8) 0.4	0.16
B2	Gm14	0.92	0.85	0.93	0.86
C1	Gm04	0.88	0.77	0.89	0.79
C2	Gm06	0.88	0.78	0.91	0.82
D1A	Gm01	0.86	0.74	0.9	0.82
D1B	Gm02	0.97	0.93	0.93	0.87
D2	Gm17	0.87	0.75	0.91	0.83
E	Gm15	-0.88	0.78	-0.89	0.79
F	Gm13	0.85	0.72	0.84	0.71
G	Gm18	0.89	0.79	0.88	0.78
H	Gm12	0.95	0.9	0.93	0.87
I	Gm20	0.73	0.53	0.76	0.58
J	Gm16	0.93	0.86	0.93	0.87
K	Gm09	0.92	0.85	0.91	0.84
L	Gm19	0.89	0.79	0.71	0.51
M	Gm07	0.94	0.88	0.92	0.84
N	Gm03	0.89	0.79	0.91	0.83
O	Gm10	0.86	0.74	0.84	0.7

Table 4. Correlation of the SFP genetic maps with the Williams82 sequence.

2.4. Discussion

2.4.1. SFP genetic diversity between soybean lines

In the current study, genetic diversity between seven *G. max* lines, one plant introduction (PI) of *G. max* and one *G. soja* accession was assessed using SFPs, reporting for the first time application of this new class of molecular markers in soybean. Based on the levels of polymorphic SFPs detected between genetically similar or diverse lines, we plotted a dendrogram and two distinct groups were formed. More specifically, the first group contained the Conrad, Athow and General *G. max* lines, and the second was composed of the Ox20-8, Sloan and PI291237 plant introduction while the PI407162 (*G.*

soja) accession was separated from the rest. Similarly with the other marker types available in soybean such as RFLP , AFLP and SSR, we found the level of polymorphic SFPs low within the groups of well adapted lines from the dendrogram on **Fig.1a.** Concerning the separation between the cultivar groups, a study by Hyten et al. showed that while modern cultivated lines have lost 81% of rare alleles present in the progenitor Asian lines, they still retain up to 72% of their polymorphisms. The coefficient of parentage for Athow by Conrad and Athow by General is approximately 0.40, and 0.36 for Conrad by General (data not shown) suggesting a close common ancestor and less genetic variability between these lines. The Ox20-8 and Sloan are distantly related to these as they are Midwestern lines, while the V71-370 is a southern line justifying its position in the middle of the two groups on **Fig.1a**. Concerning the separation of the PI407162 plant introduction on our dendrogram, it is justified based on the fact that domestication of soybean resulted in change of allele frequencies in 60% of the *G. soja* genes (Hyten et al. 2006). Similarly, increased polymorphism was found between wild and 23 cultivated Japanese, US and Chinese cultivars through the use of 668 EST-derived microsatellite markers (Hishano et al. 2008).

We also observed increased number of polymorphisms between the two groups of *G. max* lines from the dendrogram on **Fig.1a**, for a specific region on MLG-J. This region contains a major resistance QTL for the pathogen *P. sojae* (Tucker et al., submitted), for which the G. max lines in the first group (V71-370, Williams, Athow, Conrad, General) have partial resistance, while the lines in the second group (PI407162, PI291237, Ox20-8 and Sloan) are susceptible. A polymorphism for the 83424.1.S1 SFP marker, was only found present in the PI407162 line. The EST corresponding to this marker is annotated as senescence protein and maps under the peak of the QTL, suggesting a unique allele possibly related to the susceptibility of *G. soja*. This protein also contains an InterPro domain (IPR:009686) studied in members of the DSA gene family of dailily (Panavas et al. 1999). In the same study, it was shown that DSA is one the major proteins responsible for cell death during the senescence of the flower petals. We also found a glutaredoxin gene (1650.1.S1) to be polymorphic only between the groups of resistant and susceptible cultivars on the **Fig.1a** dendrogram. As it was shown in previous studies in Arabidopsis (Ndamukong et al. 2006), glutaredoxin proteins play a significant role in the cross-talk of the salicylic and jasmonic acid disease resistance signal pathways, through interaction with TGA transcription factors. Additionally, TGAs are proteins that interact with NPR1, and it has

been reported that NPR1 is a central regulator of plant systemic acquired resistance (SAR) activated from redox changes (Mou et al. 2003).

2.4.2. SFP discovery and mapping in a segregating RIL population

Using expression data from an RIL population created from a cross of *G. max* by *G. soja*, we identified segregating SFPs markers and assembled a genetic map including all 20 soybean MLGs. The use of this interspecific and highly homozygous population allowed us to identify a large number of SFPs, which otherwise would be masked from transcripts of non-polymorphic alleles in the heterozygous parental lines. We implemented an algorithm that finds bimodal distributions, for polymorphic SFP probes with two distinct distriubtion modes within the RIL data. The large population size of 293 RIL individuals provides high numbers of replication for each microarray probe, and hence increased statistical significance. Using a strict peak separation (*ps*) 50 peak separation threshold for separating the two modes of the bimodal distribution, we identified and successfully genotyped a total of 1243 SFPs. The *ps* statistic guarantees identifying two truly distinct modes within the set of probe values, eliminating false positives from overlapping modes in the case of residual heterozygous genes the RIL population. From the 1243 SFPs at this threshold, in order to increase the quality of the genetic markers we retained 941, excluding those genes where multiple SFP probes were present. These markers were successfully assembled in the 20 soybean linkage groups, along with an additional 113 SSRs, used for validation and as anchoring points to the public maps (Song et al., Choi et al.). The average number of markers per gene at *ps* 50 was 1.3, which is similar to what have been reported previously (West et al. 2006) for SFPs discovered also in a RIL population in *Arabidopsis*. In order to increase the number of markers in the soybean linkage groups, we saturated our map by adding SFPs discovered at the lower *ps* 30 threshold. Using the 3874 genes with SFPs at this threshold and the available SSRs, we created a map containing 1897 markers. Overall, the polymorphism rate (number of genes with SFPs compared to the total 37,593 genes on the array), was increased from 3.3% at *ps* 50, to 9.1% at the lower threshold of *ps* 30.

False positives in SFP detection were also considered. In the smaller genomes of yeast and *Arabidopsis*, false discovery rates have been reported at 1 and 3%, respectively (Ronald et al. and West

et al., respectively). However in large, complex genomes this rate increases significantly as reported in the case of rice with 25% and barley with 20% false positives (Kumar et al. and Rostoks et al. respectively). A recent study in cowpea using the soybean Affymetrix arrays by Das et al. estimated that 68% of the SFPs can be validated and are true positives, while they inferred the remaining 32% as false positives. It was also reported by Das et al., that at least half of the non-validated cowpea SFPs were genes in multi-gene family members. In our study, a very conservative level peak separation threshold of *ps* 50 was chosen, and while a large number of SFPs were discovered, only 1% of those were rejected due to genotyping errors. Additionally a segregation distortion test was applied to all markers, and only 5% of all SSRs and SFPs deviated from the expected 1:1 genotypic ratio. Some of these specifically mapped to MLG-D1B near the *ms* (male sterility) locus, which suggests selection at earlier generations of the population at this locus, and a plausible reason for segregation distortion.

The new SFP genetic map was compared with the soybean public maps (Song et al., Choi et al.), based on the SSR marker positions. Overall, SFPs did not appear to be distorting MLGs, as the SSRs mapping near or at the very distal ends in the public maps did not have SFPs greatly above or below them. In only one case for MLG-G an SFP marker mapped 10cM above the most distal SSR, but it was verified based on the latest results by Choi et al. that mapped SNPs at this position extend the linkage group. There was also good agreement in marker density between the SFP and public maps, with their cumulative difference for the length of all linkage groups only at 22.1 cM. We then compared the distribution and sizes of gap regions for each linkage group, in the *ps* 50 SFP map versus the RLFP-SSR map by Song et al. Most MLGs in the SFP map had more gaps than the Song et al. map, and overall there were respectively 97 versus 65 gaps of size 5-10cM, and 29 versus 5 of size greater than 10cM. On the other hand, the saturated with SFPs *ps* 30 map filled most of the gaps when compared to the transcript-SNP map by Choi et al. This comparison was based on the rationale that the Choi et al. map uses transcript-based SNPs, in order to fill the gaps in the map by Song et al. We found that smaller gaps of 5-10cM are less per MLG in the *ps* 30 SFP map versus the Choi et al. map, and overall they are half in number (20 versus 40). For gaps greater than 10cM, none remained on the *ps* 30 SFP map, and even the ones present on MLG-B1 and -I on Choi et al. map were successfully filled. Other efforts for dense map creation in soybean have been recently reported in the literature (Hishano et al. 2008, Xia et al. 2008), using F_2 crosses between the Japanese cultivar 'Misuzudaizu' and the Chinese

line 'Moshidou Gong 503'. Despite the addition of many newly developed SSR and RFLP markers, gaps greater than 10cM remained in all linkage groups with especially large gaps in MLG-C1, -E and -M. In this respect we have demonstrated, the efficiency of SFPs for increasing marker coverage and density of the soybean genetic maps.

We further validated our SFP maps by comparing the order of SFP markers on the genetic map, with their locations on the genome sequence of the Williams82 line. By perfoming a BLAST similarity search with the EST sequences corresponding to the SFPs, we found good linear correlation of genetic versus physical distance. Since for this comparison all SFPs from both the *ps* 50 and *ps* 30 maps were used, the good correlation also validated the SFPs from the lower threshold.

2.4.3. Efficiency in the SFP detection, algorithm choices and implementation

In studies using mRNA for the detection of SFPs is important to identify whether the difference observed in the signal intensities between two genetic lines is due to a polymorphism or a gene expression variation. In this respect, the SFP algorithm must discriminate whether hybridization intensities differ between two lines on a single microarray probe, or in all probes of a probeset for a gene. The latter case would be an expression variation, while the former indicates a sequence mismatch on the specific gene region the probe corresponds to, and therefore a true polymorphism. The Significance Analysis of Microarrays algorithm (SAM, Tusher et al. 2001) has been used in a number of SFP studies in plant genomes (Borevitz et al. 2003, Rostoks et al 2005, Kumar et al. 2007). In these studies SAM was applied to each probe separately, without taking into account the average of each gene probeset. Therefore, an increased number of false positive SFPs have been identified from differentially expressed genes between two lines. For our analysis we compared soybean lines by using the method by West et al. first applied in *Arabidopsis*. This method implements the SFPdev statistic in Affymetrix microarrays to quantify single probe deviations from the overall probeset level, and therefore takes into account the expression level of each gene. With this statistic, the deviation values are normalized for the expression of the gene in each line, allowing discrimination of low probe intensity due to sequence mismatch or expression variability.

Another consideration is the availability of transcripts for SFP prediction, in the tissue type used for mRNA extraction. Since our plant material was young seedlings, there is increased gene expression in this early developmental stage. We expected therefore a great portion of the transcriptome to be active, and this is also apparent from the large number of SFPs detected by our algorithms. SFPs are dominant markers and polymorphic loci can be identified only if they are homozygous, since transcripts from mutant or wild-type non-polymorphic alleles can hybridize and mask the SFP probes. The RIL population used in our study was in advanced F_{12} generation, and single-seed selection has been applied from F_4 generation. Therefore, only a minimum amount of residual heterozygocity remains, and the majority of loci are fixed for alleles inherited from one or the other parent. For any heterozygous loci our algorithm does not identify a probe as SFP, due to the strict peak separation threshold which rejects any probes with insufficient separation in the bimodal distribution. Finally, in the case of duplicated genes and gene families, hybridizing transcripts from non-polymorphic homologous genes can similarly mask the presence of an SFP probe. As with heterozygous loci, our algorithm would not detect SFPs due to poor mode separation.

All the algorithms were implemented within the Postgresql (http://www.postgresql.org) database system. There are advantages in such an implementation, since database systems are designed to allow for creating custom views of certain portions of the data, intersecting sets of results, and integrating with additional datasets. Our algorithms were implemented using the standard SQL database system language (review in Chapple 2007), can be transferred without any modification and be fully functional to any other commercial or open-source database system. Therefore, the software developed with the two algorithms can be used for the discovery of SFPs in gene expression datasets from additional segregating populations of soybean, and also other plant species.

2.5. Materials and Methods

Plant Materials. The seven cultivated soybean lines Athow, Conrad, General, V71-370, Ox20-8 and Sloan, plant introduction PI291237, *G.soja* line PI407162 and RIL population were portions of a larger study involving transcriptional profiling during soybean infection with *Phytophthora sojae* the causal agent of root and stem rot disease (Zhou et al. 2008, submitted). The 293 RILs were developed from the interspecific cross between the *Glycine max* line (V71-370, Group V) and the *Glycine soja* plant introduction (PI407162, Group IV) using a modified single-seed descent method as described in Maroof et al. All plants were grown in a growth chamber with day and night temperatures settings of 27°C and 21°C, relative humidity averaging 75 to 90%, and a 14 h light:10 h dark cycle. Seven day-old seedlings for RNA extraction were mock-inoculated and wounded at 2 cm below the beginning of the root zone by scraping the root epidermis with a sterile scalpel for all nine lines and RILs in two separate experiments. Samples of root tissues from 10 plants from each individual were collected at 5 days post mock-inoculation, immediately frozen in liquid nitrogen and stored at -80ºC prior to RNA extraction.

RNA and DNA extractions The QIAGEN RNeasy® Plant Mini Kit (Qiagen Corporation, Valencia, CA) was used throughout for total RNA isolation from the root tissue sections. The manufacturer-provided protocol was followed with minor modifications in order to obtain a sufficient amount of high quality RNA. The quality of total RNA was checked in an Agilent 2100 Bioanalyser. RNA samples from the two inoculation replications were pooled in equal amounts. DNA was collected from a bulk sample of the 10 plants used for RNA extraction for all RILs from the first replication. DNA was extracted by as described in Maroof et al. (1994) nwith minor modifications.

Microarray Data Generation and Analysis. Microarray hybridization procedures were performed at the Core Laboratory Facility of Virginia Bioinformatics Institute (Blacksburg, VA) following the standard eukaryotic gene expression assay protocols described in the Affymetrix GeneChip®

Expression Analysis Technical Manual. In brief, the One-Cycle Target Labeling and Control Reagents (Affymetrix®) and 1 ug of total RNA were used to generate biotin-labeled cRNA. Twenty micrograms of labeled cRNA was fragmented in Fragmentation Buffer and then hybridized to the Affymetric Soybean Genome Array®, which assays simultaneously 37,593 soybean transcripts. Hybridization was performed at 45°C for 16 h in an Affymetrix® hybridization oven (model 640), and then microarrays were washed and stained with streptavidin-phycoerythrin using the fluidics protocol EukGE-WS2v5-450 in the Affymetrix® GeneChip® Fluidics Station 450. Stained chips were scanned with an Affymetrix GCS3000 7G Scanner. The Affymetrix GeneChip® Operating Software (GCOS, v1.4.0.036) was used to provide instrument control, first-level data analysis, and data management for the entire microrarray assay.

Pre-Processing and Low Level Microarray Data Analysis. Quality control of acquired gene expression data was performed using box-plots for each chip and ratio-versus-intensity plots for pairs of chips and by computing 3'/5' ratios of β -Actin to check for RNA degradation. Low-level analysis of raw expression data was performed by the following steps. For the SFP discovery in the nine soybean lines only, the Affymetrix Microarray Suite version 5 (MAS5) algorithm was used for retaining genes that had a Present Call (Affymetrix GeneChip® Expression Analysis Technical Manual). Implementation of MAS5 was from the Bioconductor package *affy* (online ref: http://bioconductor.org/packages/2.0/bioc/html/affy.html). A second step for both the RIL and soybean line expression data was background correction with the model-based procedure followed by quantile-normalization for probe-level data.

SFP marker discovery. The normalized gene expression data were uploaded into tables created within the POSTgres database system. The database tables also contained information on the soybean line or RIL, and the replicate experiment for each set of expression values. For the discovery of SFPs we implemented two algorithms (West et al. 2006), which are described in more detail in the subsequent sections. The first called SFPdev Min-Max Ratio while second RIL Bimodal Distributions algorithm. Both data mining algorithms were implemented in structured query language (SQL) routines, and their output were stored in the POSTgres database system. In addition, appropriate

indexing of relational tables and query optimization within the database was performed in order to make computations time-efficient. The SQL code routines are available upon request from the authors.

SFPdev Min-Max Ratio algorithm. In summary, this algorithm (SI **Fig. S1**) first calculates the ***SFPdev*** statistic, which is the absolute difference of hybridization intensity value of each probe from the average of the probeset, divided by the value for that probe. Each probeset corresponds to a gene and is composed of 11 Perfect Match probes (PM, details at http://www.affymetrix.com), while the MisMatch (MM) probes are not included in the calculation. The *SFPdev* values are calculated for each of the four replicate microarrays, and their distribution across the replicates is computed. The *SFPdev* value is higher in the case of a polymorphic probe, since the reduced hybridization results in greater deviation from the average intensity of the probeset. The calculation of this statistic is repeated for each of the lines separately. Then by comparing pairs of lines ***a***,***b***, we accept a probe as having a SFP polymorphism, if the ratio of the smallest value $SFPdev_a$ in the distribution of values from line ***a*** carrying the polymorphism, divided by the largest $SFPdev_b$ from line ***b*** (or vice versa) is greater than two-fold. This is an empirical threshold reported during the first implementation of the algorithm, and also verified while applying this algorithm to our data.

RIL Bimodal Distributions algorithm. This algorithm is similar to K-means clustering with K = 2. In summary, the *RIL Bimodal Distributions* (SI **Fig. S4**) algorithm first calculates the absolute values of probe intensity differences $\mathbf{d_{ijk}}$ (probe ***i***, probeset ***j***, RIL individual ***k***), from the average of each probeset. Similarly with the SFPdev, only PM probes are included in the calculation. In the next step, the algorithm computes the distribution of each $\mathbf{d_{ijk}}$ value across the 293 individuals of the RIL population. The median $\mathbf{M_{ij}}$ is initially used to split the distribution into an upper (***u***) and lower (***l***) subsets (SI **Fig. S3**). The averages $\boldsymbol{l}_{avg}$, $\boldsymbol{u}_{avg}$ of the ***l*** and ***u*** subsets respectively, are the seeding centers for the K-means clustering. Then the algorithm iterates in the same manner for eight times, but instead of the $\mathbf{M_{ij}}$, it uses the average $(\boldsymbol{l}_{avg} + \boldsymbol{u}_{avg})/2$ for splitting again into ***u***,***l*** subsets. After all iterations, the $\mathbf{d_{ijk}}$ values settle into a bimodal distribution, with each mode corresponding to a K-means cluster. These steps are repeated for probes in all the probesets, measured in the expression profile of each individual

of the population. In order to assess significant separation between the two distribution modes (or otherwise the two clusters), we use as metric the peak separation $ps = (A_l - A_u)/\sqrt{(S_l^2/n_l + S_u^2/n_u)}$ (A_l and A_u are distribution averages for the ***u***,***l*** modes respectively, standard deviations S_l and S_u, sample sizes n_l and n_u). We accept as SFPs the probes having $ps > 1.82$, which corresponds to the upper 5th percentile cutoff of a t-distribution.

The algorithm also computes the $\mathbf{d_{ij}}$ values (averaged across replicate microarrays) for the PI407162 and V71-370 parental data. For polymorphic probes, $\mathbf{d_{ijk}}$ values for the individuals of the RIL population are expected to cluster around the parental $\mathbf{d_{ij}}$, under the two modes of the distribution (SI **Fig. S3**). Since the RIL population was created by the cross of the genetically distant PI407162 and V71-370 soybean lines, the two modes originate due to the different parental alleles inherited to the RIL progeny. For each SFP probe, RIL individuals are assigned a genotype based on their clustering around one of the parental values (SI **Fig. S3**).

Clustering of 9 soybean lines based on SFPs. The R statistical analysis software (http://www.r-project.org) was used, and a matrix was created for the numbers of SFPs between soybean lines . The matrix was used as input to the "dist" function of the standard R library, and the euclidean distance between lines was calculated based on the numbers of SFPs. Next, a dendrogram was computed using the hierarchical clustering package "hclust" in R, based on the single linkage (nearest-neighbor) algorithm and using euclidean distance metric. The dendrogram was plotted using the "plot(dendro)" command.

We then validated the clusters in the dendrogram by performing bootstrap resampling, using the "pvclust" (Suzuki and Shimodaira 2006) package in R. The "pvclust" package calculates p-values for each cluster in hierarchical clustering, which indicate how strong the cluster is supported by data. Two types of p-values are provided : AU (Approximately Unbiased) and BP (Bootstrap Probability) p-value. The AU p-value is computed by multiscale bootstrap resampling, and is a better approximation to an unbiased p-value than BP, which is computed by normal bootstrapping. In our analysis we set the number of bootstrap iterations to 10, 000, and we plotted the bootstrapped clusters along with their p-

values . We then examined the plot for significant clusters using the AU p-value at 95% confidence (0.05).

Genetic map construction. Publicly available SSR markers, detecting polymorphism between the parental lines of the RIL population, were incorporated with SFP markers in map construction. The SSRs were selected approximately 30 cM apart from their map positions on the public consensus genetic map, and additional SSRs were chosen to cover the very distal ends of all twenty MLGs. PCR amplification, poly-acrylamide gel electrophoresis and SSR assays were performed as described by Yu et al. and Maroof et al. .

For map construction and MLG identification, 941 SFPs with a single polymorphic probe per gene at *ps* threshold greater than 50 were used. Also 113 SSR markers were added to this set, resulting in a total of 1054 genetic markers. The SFP and SSR markers were grouped to the 20 soybean MLGs, using LOD threshold of 5.0 in JoinMap3.0. However, MLG-I and -M grouped end to end and thus a higher LOD of 10 was chosen to separate the two. Recombination values were converted to genetic distance using Kosambi's mapping function in JoinMap3.0. Maps were similarly created following addition of markers from the lower *ps* 30 threshold in each linkage group, in order to fill the existing gaps. In case any of the additional SFP markers distorted the order or length of the linkage groups, they were removed and the maps were re-created. This was performed for a few iterations until a balance between gap-filling and validity of each linkage group was achieved.

Comparison with the Williams82 whole genome sequence. The 20 sequence scaffolds Williams82 sequence which correspond to the soybean linkage groups, were obtained from http://www.phytozome.org. They were used in a BLAST search with the Unigene-EST sequence corresponding to each SFP as query. Due to the high rate of duplication in the soybean genome we used a strict search threshold of 1e-50. We additionally filtered the BLAST hits, and kept only the ones from the scaffold corresponding to the linkage group (Table 4) where each SFP is located in our genetic maps. Finally, correlation was calculated for the SFP chromosomal positions (measured in cM of genetic distance), to their positions on the sequence scaffolds from the BLAST results.

2.6. References

Borevitz JO, Liang D, Plouffe D, Chang H-S, Zhu T, Weigel D, Berry CC, Winzeler E and Chory J (2003). "Large-scale identification of single-feature polymorphisms in complex genomes." Genome Res. **13**: 513-523.

Chapple M (2007). "SQL Fundamentals."

Choi I-Y, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon M-S, Hwang E-Y, Yi S-I, Young ND, Shoemaker RC, van Tassell CP, Specht JE and Cregan PB (2007). "A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis." Genetics **176**: 685-696.

Cregan PB, Jarvik T, Bush AL, Shoemaker RC, Lark KG, Kahler AL, Kaya N, VanToai TT, Lohnes DG and Chung J (1999). "An integrated genetic linkage map of the soybean genome." Crop Sci. **39**: 1464-1490.

Cui X, Xu J, Asghar R, Condamine P, Svensson JT, Wanamaker S, Stein N, Roose M and Close TJ (2005). "Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit." Bioinformatics **21**: 3852-3858.

Das S, Bhat PR, Sudhakar C, Ehlers JD, Wanamaker S, Roberts PA, Cui X and Close TJ (2008). "Detection and validation of single feature polymorphisms in cowpea (*Vigna unguiculata* L. *Walp*) using a soybean genome array." BMC Genomics **9**: 107.

Hisano H., Sato S., Isobe S., Sasamoto S., Wada T., Matsuno A., Fujishiro T., Yamada M., Nakayama S., Nakamura Y., Watanabe S., Harada K and Tabata S. (2007) **"**Characterization of the soybean genome using EST-derived microsatellite markers" DNA Res. **14**: 271-281.

Hyten DL, Song Q, Zhu Y, Choi I-Y, Nelson RL, Costa JM, Specht JE, Shoemaker RC and Cregan PB (2006). "Impacts of genetic bottlenecks on soybean genome diversity." PNAS (USA) **103**: 16666-16671.

Jin W, Palmer RG, Horner HT and Shoemaker RC (1998). "Molecular mapping of a male-sterile gene in soybean." Crop Sci. **38**: 1681-1685.

Keim P, Diers BW, Olson TC and Shoemaker RC (1990). "RFLP mapping in soybean: association between marker loci and variation in quantitative traits." Genetics **126**: 735-742.

Keim P, Shoemaker RC and Palmer RG (1989). "Restriction fragment length polymorphism diversity in soybean." Theor. Appl. Genet. **77**: 786-792.

Kumar R, Qiu J, Joshi T, Valliyodan B, Xu D and Nguyen HT (2007). "Single feature polymorphism discovery in rice." PLos ONE **14**: e284.

Lark KG, Weisemann JM, Matthews BF, Palmer R, Chase K and Macalma T (1993). "A genetic map of soybean (*Glycine max* L.) using an intraspecific cross of two cultivars: 'Minosy' and 'Noir 1'." Theor. Appl. Genet. **86**: 901-906.

Maroof MAS, Biyashev RM, Yang GP, Zhang Q and Allard RW (1994). "Extraordinarily polymorphic microsatellite DNA in barley: species diversity, chromosomal locations, and population dynamics." Proc. Natl. Acad. Sci. (USA) **91**: 5466-5470.

Maughan PJ, Saghai Maroof MA, Buss GR and Huestis GM (1996). "Amplified fragment length polymorphism (AFLP) in soybean: species diversity, inheritance, and near-isogenic line analysis." Theor. Appl. Genet. **93**: 392-401.

Mou Z, Fan W and Dong X (2003). "Inducers of plant systemic acquired resistance regulate NPR1 function through redox changes." Cell **7**: 9-15.

Ndamukong I, Abdallat AA, Thurow C, Fode B, Zander B, Weigel R, Gatz C and Von Haller A (2007). "SA-inducible Arabidopsis glutaredoxin interacts with TGA factors and suppresses JA-responsive PDF1.2 transcription." Plant J. **1**: 11-18.

Panavas T, Pikula A, Reid PD, Rubinstein B and Walker EL (1999). "Identification of senescense-associated genes from dailylily petals." Plant Mol. Biol. **40**: 11-15.

Ronald J, Akey J, Whittle EN and Smith G (2005). "Simultaneous genotyping gene-expression measurement and detection of allele-specific expression with oligonucleotide arrays." Genome Res. (15): 7.

Rostoks N, Borevitz J, Hedley P, Russell J, Mudie S, Morris J, Cardle L, Marshall D and Waugh R (2005). "Single-feature polymorphism discovery in the barley transcriptome." Genome Biology **6**: R54.

Shoemaker RC and Specht JE (1995). "Integration of the soybean molecular and classic genetic linkage groups." Crop Sci. **35**: 10.

Shoemaker RC and T.C O (1993). Molecular linkage map of soybean. Genetic maps: locus maps of complex genomes. Cold Spring Harbor Press, NY.

Song Q, Marek L, Shoemaker R, Lark K, Concibido V, Delannay X, Specht J and Cregan P (2004). "A new integrated genetic linkage map of the soybean." Theor. Appl. Genet. **109**(1): 122-128.

St. Martin SK (1982). "Effective poplation size for the soybean improvement program in maturity groups 00 to IV." Crop Sci. **22**: 151-152.

Suzuki R. SH (2006). "Pvclust: an R package for assessing the uncertainty in hierarchical clustering." Bioinformatics **12**: 2-9.

Tusher VG, Tibshirani R and Chu G (2001). "Significance analysis of microarrays applied to the ionizing radiation response." PNAS **98**: 5-11.

Van Oeveren J (2007). Co-dominant SFP genotyping in tomato. Intelligent Systems for Molecular Biology. Vienna, Austria.

Van Ooijen JW and Voorrips RE (2001). "JoinMap 3.0 software for the calculation of genetic linkage maps." Plant Research International, Wageningen, The Netherlands.

West MAL, van Leeuwen H, Kozik A, Kliebenstein DJ, Doerge RW, St. Clair DA and Michelmore RW (2006). "High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis." Genome Res. **16**: 787-795.

Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ and Davis RW (1998). "Direct allelic variation scanning of the yeast genome." Science **281**(5380): 1194-1197.

Yu Y, Saghai Maroof MA, Buss GR, Maughn PJ and Tolin SA (1994). "RFLP and microsatellite mapping of a gene for soybean mosaic virus resistance." Phytopathology **84**: 60-64.

Xia Z., Tsubokura Y, Hoshi M, Hanawa M, Yano C, Okamura K, Ahmed TA, Anai T, Watanabe S, Hayashi M, Kawai T, Hossain KG, Masaki H, Asai K, Yamanaka N, Kubo N, Kadowaki K, Nagamura N, Yano M, Sasaki T, and Harada K. (2007) “An integrated high-density linkage map of soybean with RFLP/ SSR/ STS/ and AFLP markers using a single F2 population” DNA Res. **14**: 257-269.

2.7. Supplementary Information

S1. Steps performed by the *SFPdev Min-Max Ratio* algorithm

probe p_{ijk}

i index of probe within the probeset (11 PM probes/probeset)
j probeset-gene index (37593 genes on the chip)
k_a, k_b replication index, parental lines ***a, b*** compared in pairs

- STEP.1

i. for each replication k_a

for each probeset **j**

calculate average of probe values $\mathbf{A}_{jka}$ within that probeset

for each probe **i**

calculate $SFPdev(ijk)a$ statistic as $SFPdev(ijk)a = abs\left(\frac{pijka - Ajka}{pijka}\right)$

repeat for k_b

ii. calculate distributions for the sets of the { $SFPdev(ijk)_a$ }, { $SFPdev(ijk)_b$ } values

- STEP.2, compare k_a vs k_b

for each probeset **j**

for each probe **i**

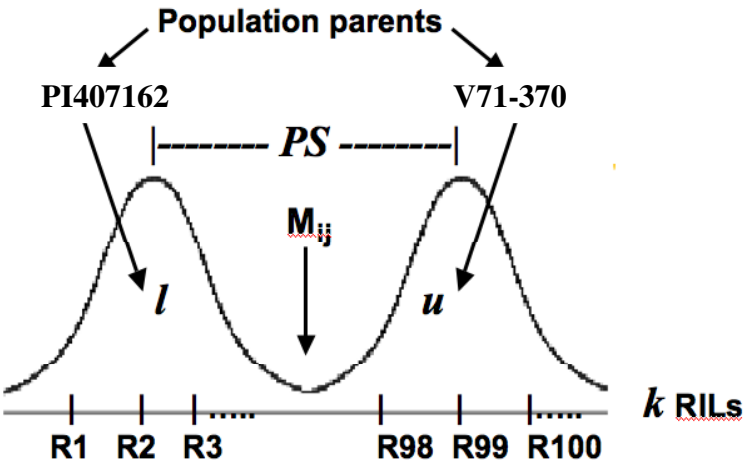
accept probe $\mathbf{p}_{ij}$ as SFP

$$\text{if} \quad \frac{\min\{SFPdev(ijk)_a\}}{\max\{SFPdev(ijk)_b\}} > 2 \quad \text{or} \quad \frac{\min\{SFPdev(ijk)_b\}}{\max\{SFPdev(ijk)_a\}} > 2$$

S2. SFPs found with both the SFPdev and RIL Bimodal Distributions (threshold of *ps* > 30) algorithms.

	Athow	Conrad	General	Ox20-8	PI291	Sloan	V71	Williams	PI407
Athow	0	117	109	148	150	127	180	159	3814
Conrad	130	0	145	150	187	137	203	195	3800
General	108	101	0	138	186	124	192	180	3806
Ox20-8	241	198	235	0	185	126	233	244	3796
PI291	255	255	274	184	0	178	250	276	3809
Sloan	234	196	207	117	160	0	242	228	3811
V71	278	252	253	239	245	225	0	313	3814
Williams	141	147	151	131	166	108	238	0	3814

S3. Schematic representing the RIL Bimodal Distributions algorithm. Gene expression values for individual RILs (R1, R2, R3,,, R98, R99, R100,,,) of the population are clustered around the parental values (PI407162 and V71-370) by the *RIL Bimodal Distributions* algorithm.



S4. Steps performed by the RIL Bimodal Distributions algorithm.

probe $\boldsymbol{p_{ijk}}$

i index of probe within the probeset (11 PM probes / probeset)
j probeset-gene index (37593 genes on the chip)
k number of RIL individual (293) RILs assayed in two replicate experiments)

- STEP.1

average probe values $\boldsymbol{p_{ijk}}$ from replicate experiments

- STEP.2

i. for each RIL **k**

for each probeset **j**

calculate average of probe values $\mathbf{A_{jk}}$ within that probeset

for each probe **i**

subtract the average of the probeset $\mathbf{d_{ijk} = p_{ijk} - A_{jk}}$

ii. for each probeset **j**

for each probe **i**

calculate distributions for the set of 293 $\mathbf{d_{ijk}}$ values

- STEP.3

for each probeset **j**

for each probe **i**

calculate median $\mathbf{M_{ij}}$ of set of $\mathbf{d_{ijk}}$ values

split in lower (***l***) and upper (***u***) sub-set, get averages $\boldsymbol{l_{avg}}$ and $\boldsymbol{u_{avg}}$

repeat for 8 iterations:

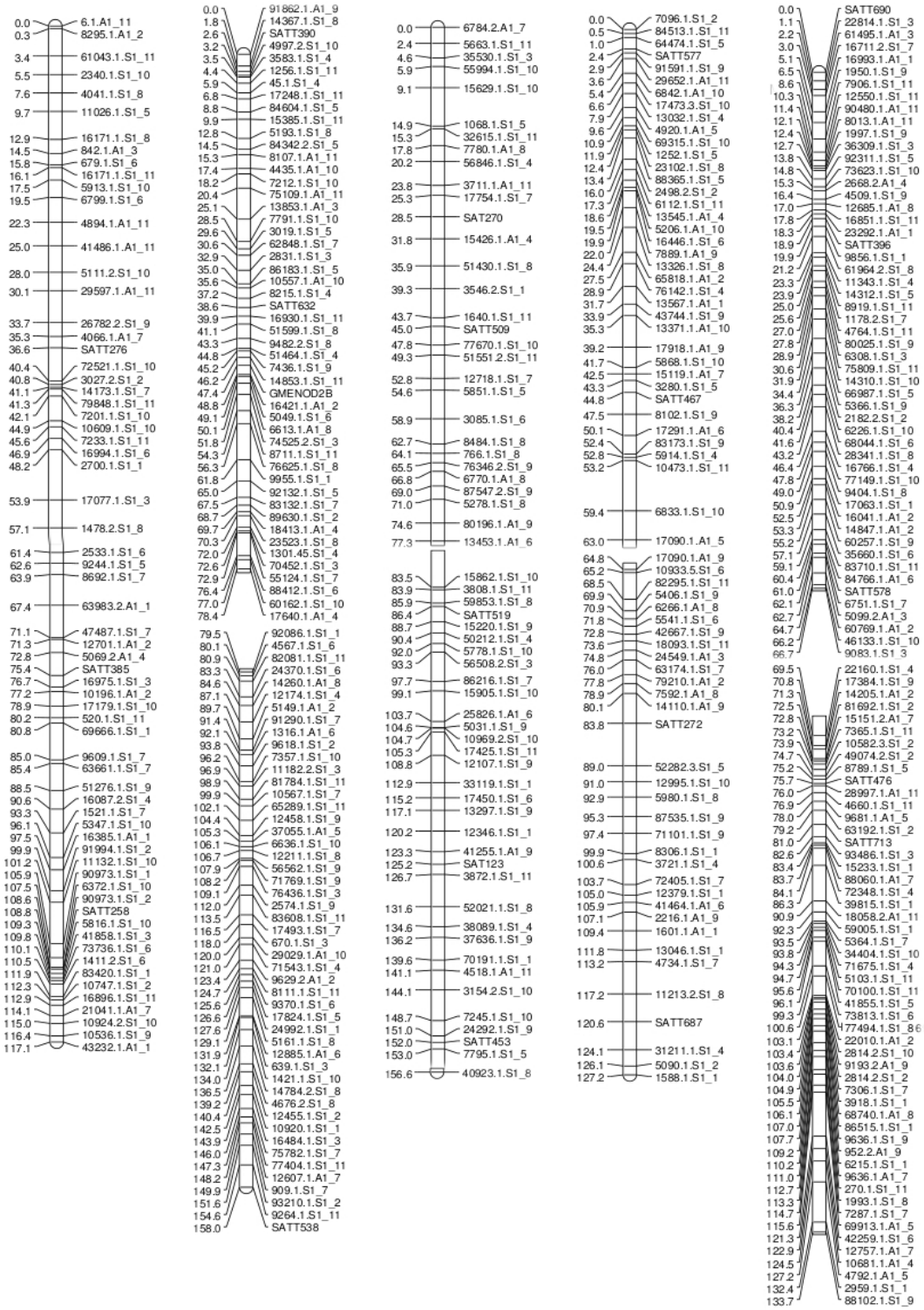
split again in new ***l*** and ***u*** based on $\mathbf{(\, l_{avg} + u_{avg}\,) / 2}$

compute peak separation $\boldsymbol{ps_{ij}} = (\boldsymbol{u_{avg}} - \boldsymbol{l_{avg}})/\sqrt{(\boldsymbol{S_l^2/n_l} + \boldsymbol{S_u^2/n_u})}$

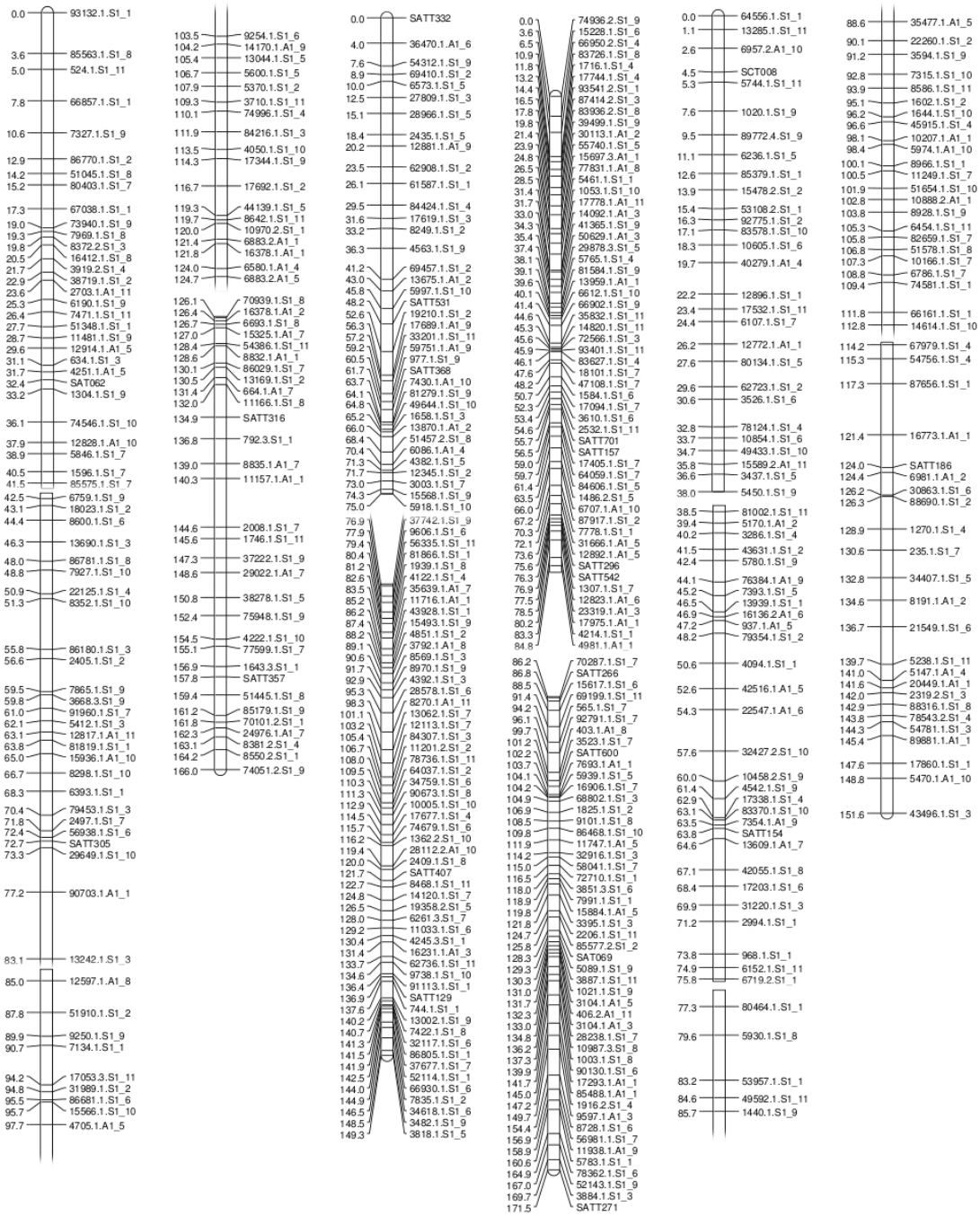
(where $\boldsymbol{S_l}$, $\boldsymbol{S_u}$ and $\boldsymbol{n_l}$, $\boldsymbol{n_u}$ the standard deviation and size of the lower and upper resulting modes of the bimodal distribution, respectively)

S5. Maps with gaps filled using SFPs from the *ps* 30 threshold.

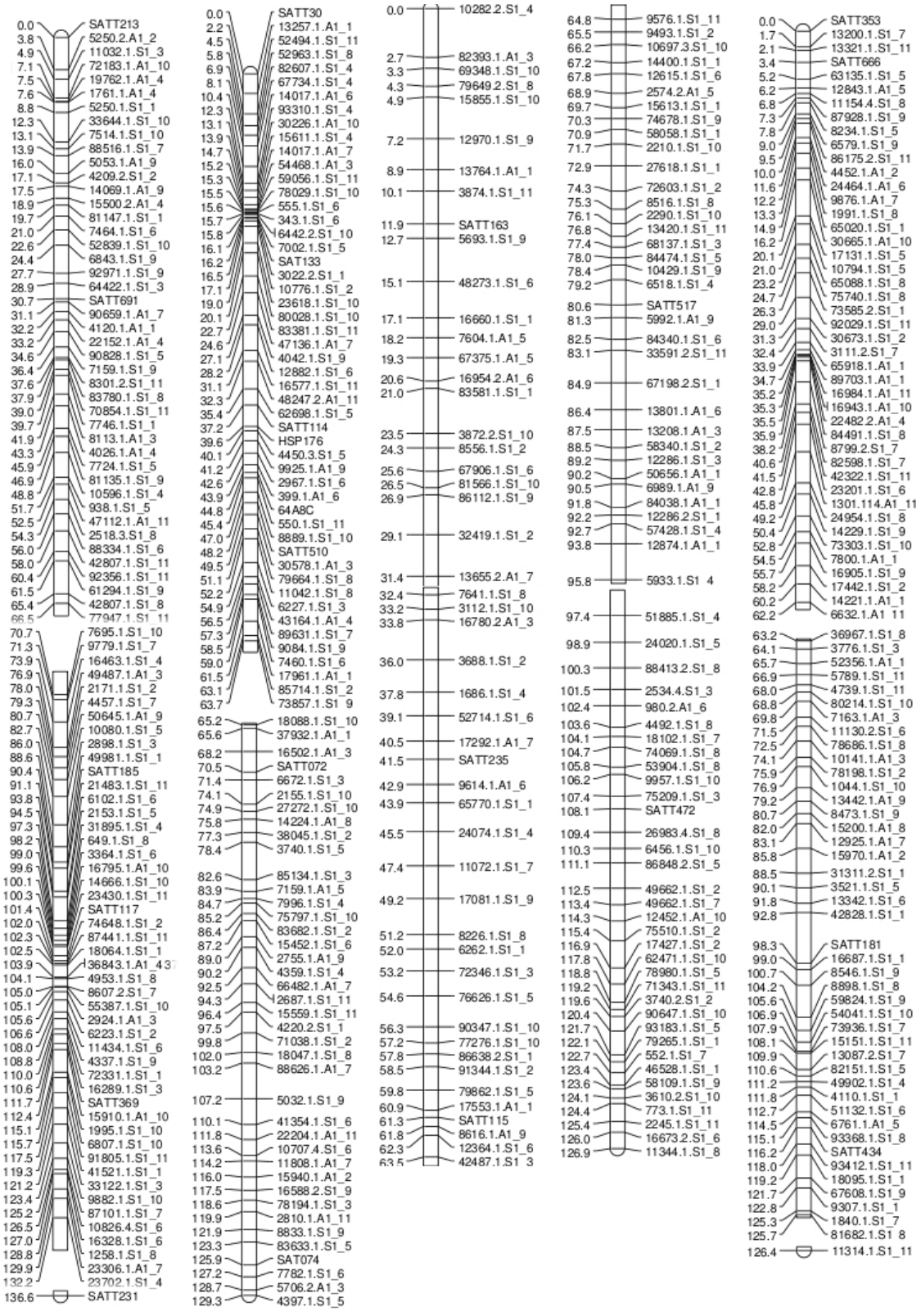
MLG-A1 **MLG-A2** **MLG-B1** **MLG-B2** **MLG-C1**

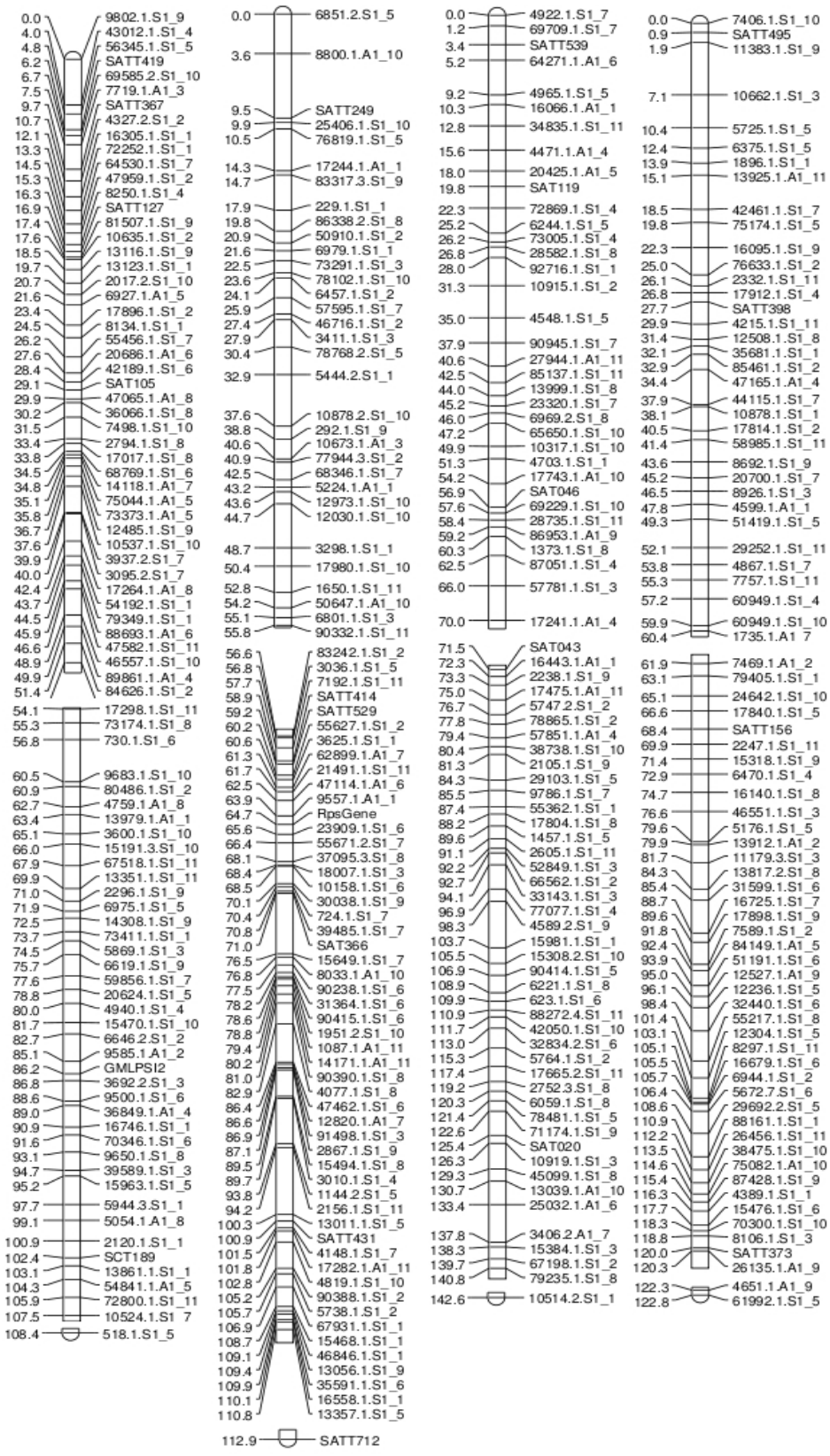


MLG-C2 **MLG-D1a** **MLG-D1b** **MLG-D2**

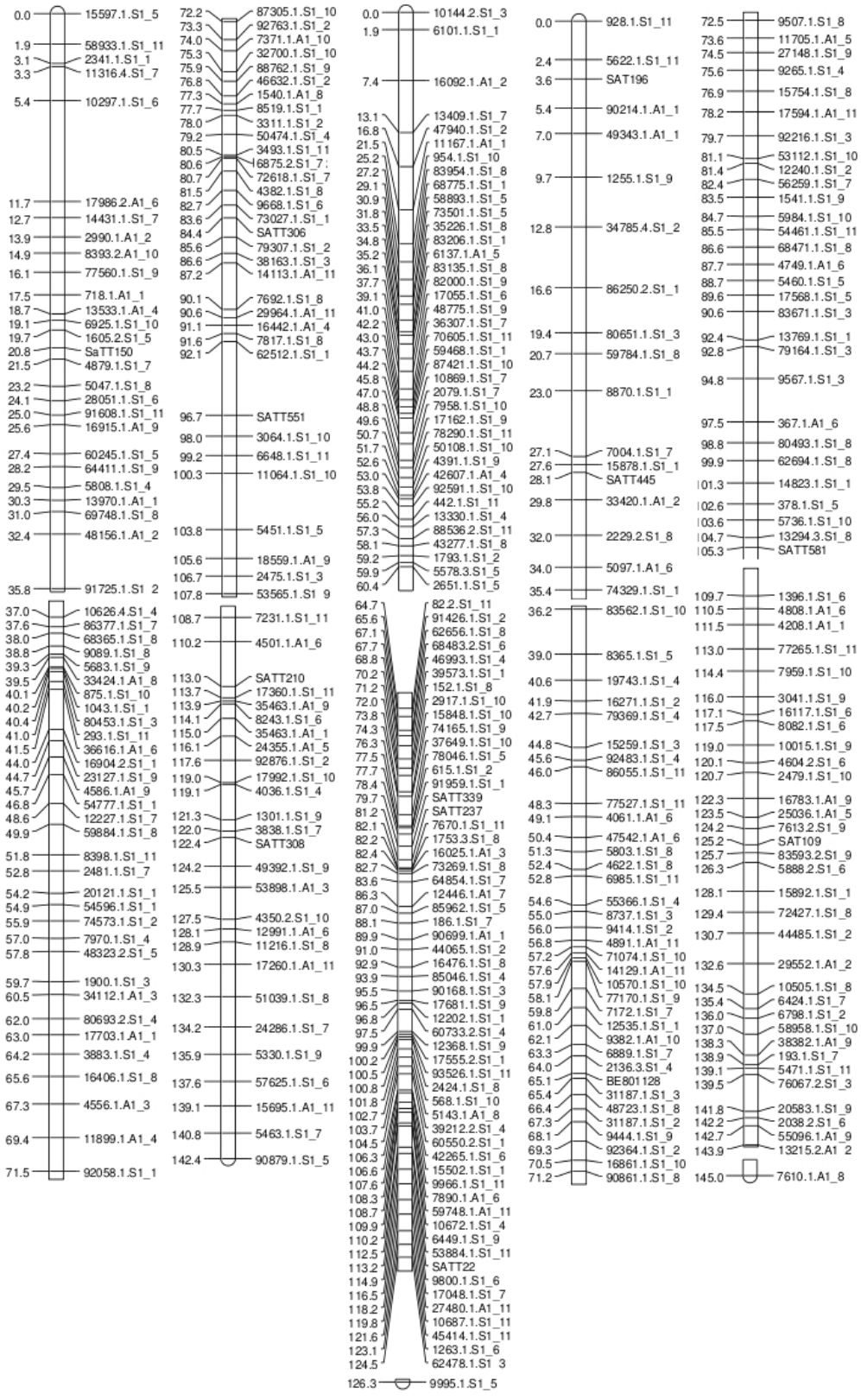


MLG-E **MLG-F** **MLG-G** **MLG-H**




MLG-I
MLG-J
MLG-K
MLG-L
0.0 9802.1.S1_9
4.0 43012.1.S1_4
4.8 56345.1.S1_5
6.2 SATT419
6.7 69585.2.S1_10
7.5 7719.1.A1_3
9.7 SATT367
10.7 4327.2.S1_2
12.1 16305.1.S1_1
13.3 72252.1.S1_1
14.5 64530.1.S1_7
15.3 47959.1.S1_2
16.3 8250.1.S1_4
16.9 SATT127
17.4 81507.1.S1_9
17.6 10635.1.S1_2
18.5 13116.1.S1_9
19.7 13123.1.S1_1
20.7 2017.2.S1_10
21.6 6927.1.A1_5
23.4 17896.1.S1_2
24.5 8134.1.S1_1
26.2 55456.1.S1_7
27.6 20686.1.A1_6
28.4 42189.1.S1_6
29.1 SAT105
29.9 47065.1.A1_8
30.2 36066.1.S1_8
31.5 7498.1.S1_10
33.4 2794.1.S1_8
33.8 17017.1.S1_8
34.5 68769.1.S1_6
34.8 14118.1.A1_7
35.1 75044.1.A1_5
35.8 73373.1.A1_5
36.7 12485.1.S1_9
37.6 10537.1.S1_10
39.9 3937.2.S1_7
40.0 3095.2.S1_7
42.4 17264.1.A1_8
43.7 54192.1.S1_1
44.5 79349.1.S1_1
45.9 88693.1.A1_6
46.6 47582.1.S1_11
48.9 46557.1.S1_10
49.9 89861.1.A1_4
51.4 84626.1.S1_2
54.1 17298.1.S1_11
55.3 73174.1.S1_8
56.8 730.1.S1_6
60.5 9683.1.S1_10
60.9 80486.1.S1_2
62.7 4759.1.A1_8
63.4 13979.1.A1_1
65.1 3600.1.S1_10
66.0 15191.3.S1_10
67.9 67518.1.S1_11
69.9 13351.1.S1_11
71.0 2296.1.S1_9
71.9 6975.1.S1_5
72.5 14308.1.S1_9
73.7 73411.1.S1_1
74.5 5869.1.S1_3
75.7 6619.1.S1_9
77.6 59856.1.S1_7
78.8 20624.1.S1_5
80.0 4940.1.S1_4
81.7 15470.1.S1_10
82.7 6646.2.S1_2
85.1 9585.1.A1_2
86.2 GMLPSI2
86.8 3692.2.S1_3
88.6 9500.1.S1_6
89.0 36849.1.A1_4
90.9 16746.1.S1_1
91.6 70346.1.S1_6
93.1 9650.1.S1_8
94.7 39589.1.S1_3
95.2 15963.1.S1_5
97.7 5944.3.S1_1
99.1 5054.1.A1_8
100.9 2120.1.S1_1
102.4 SCT189
103.1 13861.1.S1_1
104.3 54841.1.A1_5
105.9 72800.1.S1_11
107.5 10524.1.S1_7
108.4 518.1.S1_5
0.0 6851.2.S1_5
3.6 8800.1.A1_10
9.5 SATT249
9.9 25406.1.S1_10
10.5 76819.1.S1_5
14.3 17244.1.A1_1
14.7 83317.3.S1_9
17.9 229.1.S1_1
19.8 86338.2.S1_8
20.9 50910.1.S1_2
21.6 6979.1.S1_1
22.5 73291.1.S1_3
23.6 78102.1.S1_10
24.1 6457.1.S1_2
25.9 57595.1.S1_7
27.4 46716.1.S1_2
27.9 3411.1.S1_3
30.4 78768.2.S1_5
32.9 5444.2.S1_1
37.6 10878.2.S1_10
38.8 292.1.S1_9
40.6 10673.1.A1_3
40.9 77944.3.S1_2
42.5 68346.1.S1_7
43.2 5224.1.A1_1
43.6 12973.1.S1_10
44.7 12030.1.S1_10
48.7 3298.1.S1_1
50.4 17980.1.S1_10
52.8 1650.1.S1_11
54.2 50647.1.A1_10
55.1 6801.1.S1_3
55.8 90332.1.S1_11
56.6 83242.1.S1_2
56.8 3036.1.S1_5
57.7 7192.1.S1_11
58.9 SATT414
59.2 SATT529
60.2 55627.1.S1_2
60.6 3625.1.S1_1
61.3 62899.1.A1_7
61.7 21491.1.S1_11
62.5 47114.1.A1_6
63.9 9557.1.A1_1
64.7 RpsGene
65.6 23909.1.S1_6
66.4 55671.2.S1_7
68.1 37095.3.S1_8
68.4 18007.1.S1_3
68.5 10158.1.S1_6
70.1 30038.1.S1_9
70.4 724.1.S1_7
70.8 39485.1.S1_7
71.0 SAT366
76.5 15649.1.S1_7
76.8 8033.1.A1_10
77.5 90238.1.S1_6
78.2 31364.1.S1_6
78.6 90415.1.S1_6
78.8 1951.2.S1_10
79.4 1087.1.A1_11
80.2 14171.1.A1_11
81.0 90390.1.S1_8
82.9 4077.1.S1_8
86.4 47462.1.S1_6
86.6 12820.1.A1_7
86.9 91498.1.S1_3
87.1 2867.1.S1_9
89.5 15494.1.S1_8
89.7 3010.1.S1_4
93.8 1144.2.S1_5
94.2 2156.1.S1_11
100.3 13011.1.S1_5
100.9 SATT431
101.5 4148.1.S1_7
101.8 17282.1.A1_11
102.8 4819.1.S1_10
105.2 90388.1.S1_2
105.7 5738.1.S1_2
106.9 67931.1.S1_1
108.7 15468.1.S1_1
109.1 46846.1.S1_1
109.4 13056.1.S1_9
109.9 35591.1.S1_6
110.1 16558.1.S1_1
110.8 13357.1.S1_5
112.9 SATT712
0.0 4922.1.S1_7
1.2 69709.1.S1_7
3.4 SATT539
5.2 64271.1.A1_6
9.2 4965.1.S1_5
10.3 16066.1.A1_1
12.8 34835.1.S1_11
15.6 4471.1.A1_4
18.0 20425.1.A1_5
19.8 SAT119
22.3 72869.1.S1_4
25.2 6244.1.S1_5
26.2 73005.1.S1_4
26.8 28582.1.S1_8
28.0 92716.1.S1_1
31.3 10915.1.S1_2
35.0 4548.1.S1_5
37.9 90945.1.S1_7
40.6 27944.1.A1_11
42.5 85137.1.S1_11
44.0 13999.1.S1_8
45.2 23320.1.S1_7
46.0 6969.2.S1_8
47.2 65650.1.S1_10
49.9 10317.1.S1_10
51.3 4703.1.S1_1
54.2 17743.1.A1_10
56.9 SAT046
57.6 69229.1.S1_10
58.4 28735.1.S1_11
59.2 86953.1.A1_9
60.3 1373.1.S1_8
62.5 87051.1.S1_4
66.0 57781.1.S1_3
70.0 17241.1.A1_4
71.5 SAT043
72.3 16443.1.A1_1
73.3 2238.1.S1_9
75.0 17475.1.A1_11
76.7 5747.2.S1_2
77.8 78865.1.S1_2
79.4 57851.1.A1_4
80.4 38738.1.S1_10
81.3 2105.1.S1_9
84.3 29103.1.S1_5
85.5 9786.1.S1_7
87.4 55362.1.S1_1
88.2 17804.1.S1_8
89.6 1457.1.S1_5
91.1 2605.1.S1_11
92.2 52849.1.S1_3
92.7 66562.1.S1_2
94.1 33143.1.S1_3
96.9 77077.1.S1_4
98.3 4589.2.S1_9
103.7 15981.1.S1_1
105.5 15308.2.S1_10
106.9 90414.1.S1_5
108.9 6221.1.S1_8
109.9 623.1.S1_6
110.9 88272.4.S1_11
111.7 42050.1.S1_10
113.0 32834.2.S1_6
115.3 5764.1.S1_2
117.4 17665.2.S1_11
119.2 2752.3.S1_8
120.3 6059.1.S1_8
121.4 78481.1.S1_5
122.6 71174.1.S1_9
125.4 SAT020
126.3 10919.1.S1_3
129.3 45099.1.S1_8
130.7 13039.1.A1_10
133.4 25032.1.A1_6
137.8 3406.2.A1_7
138.3 15384.1.S1_3
139.7 67198.1.S1_2
140.8 79235.1.S1_8
142.6 10514.2.S1_1
0.0 7406.1.S1_10
0.9 SATT495
1.9 11383.1.S1_9
7.1 10662.1.S1_3
10.4 5725.1.S1_5
12.4 6375.1.S1_5
13.9 1896.1.S1_1
15.1 13925.1.A1_11
18.5 42461.1.S1_7
19.8 75174.1.S1_5
22.3 16095.1.S1_9
25.0 76633.1.S1_2
26.1 2332.1.S1_11
26.8 17912.1.S1_4
27.7 SATT398
29.9 4215.1.S1_11
31.4 12508.1.S1_8
32.1 35681.1.S1_1
32.9 85461.1.S1_2
34.4 47165.1.A1_4
37.9 44115.1.S1_7
38.1 10878.1.S1_1
40.5 17814.1.S1_2
41.4 58985.1.S1_11
43.6 8692.1.S1_9
45.2 20700.1.S1_7
46.5 8926.1.S1_3
47.8 4599.1.A1_1
49.3 51419.1.S1_5
52.1 29252.1.S1_11
53.8 4867.1.S1_7
55.3 7757.1.S1_11
57.2 60949.1.S1_4
59.9 60949.1.S1_10
60.4 1735.1.A1_7
61.9 7469.1.A1_2
63.1 79405.1.S1_1
65.1 24642.1.S1_10
66.6 17840.1.S1_5
68.4 SATT156
69.9 2247.1.S1_11
71.4 15318.1.S1_9
72.9 6470.1.S1_4
74.7 16140.1.S1_8
76.6 46551.1.S1_3
79.6 5176.1.S1_5
79.9 13912.1.A1_2
81.7 11179.3.S1_3
84.3 13817.2.S1_8
85.4 31599.1.S1_6
88.7 16725.1.S1_7
89.6 17898.1.S1_9
91.8 7589.1.S1_2
92.4 84149.1.A1_5
93.9 51191.1.S1_6
95.0 12527.1.A1_9
96.1 12236.1.S1_5
98.4 32440.1.S1_6
101.4 55217.1.S1_8
103.1 12304.1.S1_5
105.1 8297.1.S1_11
105.5 16679.1.S1_6
105.7 6944.1.S1_2
106.4 5672.7.S1_6
108.6 29692.2.S1_5
110.9 88161.1.S1_1
112.2 26456.1.S1_11
113.5 38475.1.S1_10
114.6 75082.1.A1_10
115.4 87428.1.S1_9
116.3 4389.1.S1_1
117.7 15476.1.S1_6
118.3 70300.1.S1_10
118.8 8106.1.S1_3
120.0 SATT373
120.3 26135.1.A1_9
122.3 4651.1.A1_9
122.8 61992.1.S1_5

MLG-M **MLG-N** **MLG-O**



3. Identifying *Arabidopsis* metabolic pathways present in soybean and their perturbation during *P.sojae* infection.

3.1. Abstract

Using sequence similarity comparison between the genomes of soybean (*Glycine max*) and *Arabidopsis thaliana*, 197 out of 221 pathways with assigned genes in the latter species were reconstructed in soybean. Using a novel algorithm and the gene expression data from eight cultivated lines and one soybean plant introduction (*G. sojae*) inoculated with *Phytophthora sojae*, we assessed perturbation of the transferred pathways during pathogen infection in different genetic backgrounds. We also used expression data from a soybean recombinant inbred line (RIL) population, and compared the pathway activity between groups of RILs with different parental genotypes under major disease-resistance QTL regions in chromosomes J, G, D1a and I. We found significant differences for secondary metabolism pathways in the cultivars versus the RILs, while also variation was observed between the different QTL chromosomal loci. These pathways include anthocyanin, chorismate biosynthesis and giberellin metabolism and where found perturbed in all the cultivar data, while only chromosome D1a showed perturbation of these pathways. In contrast, the suberin and phenylpropanoid biosynthesis pathways were perturbed in all four chromosomal regions except D1a. Uniquely perturbed for the QTL region in chromosome J are the carotenoid and IAA biosynthesis pathways. For the eight cultivar data, we observed perturbation of a large number of sugar metabolism pathways in the *P.sojae* susceptible cultivars Ox20-8, PI291237 and Sloan. This was not the case for the cultivars Athow, Conrad and General, which display partial resistance to *P.sojae*. Our results therefore suggest a unique genetic element under the resistance QTL in the D1a chromosome, and also the perturbation of the sugar metabolism pathways in the susceptible cultivars, as a result of the energy deficiency caused by the pathogenic infection.

3.2. Introduction

Metabolism, is the main means for plant species to interact and persevere within their environment. Through flexibility of their metabolic pathway network, plant organisms can adapt to changing environmental conditions such as nutrient and water availability, harsh climate and potential pathogens (Lange and Ghassemian, 2005). Currently, only a minimal dataset exists for the metabolic pathways of soybean (*G. max,* http://www.soybease.org), despite its being one of the most important agricultural products in United States. It has been shown that during the defense response against plant pathogens, structural and physiological changes take place in the cells (Feys 2000), as a result of activation or shutdown of metabolic pathways.

The contribution of the metabolic network function for a plant species' specific phenotype, can be identified by studying how the pathways operate in different genetic backgrounds. One example of such phenotype of great importance for plant species is the resistance against pathogens. Studying perturbation patterns of pathways during pathogen infection, can identify physiological mechanisms of defense used by the plants. A standard method for studying pathway activity (Ideker et al. 2002), is by integrating gene expression data measured under a certain biological perturbations, with data for the metabolic network. Using a computational methodology the perturbation or repression of parts of a metabolic network can be deduced based on changes in the transcriptional profile of the genes encoding the pathway enzymes, in response to the particular perturbation. A newly developed algorithm for this purpose, is the Pathway Perturbation Algorithm (Rivera et al. 2009, in preparation). This algorithm calculates the perturbation of the totality of a biological pathway or parts of it, based on the change of gene expression levels corresponding to the enzymes in adjacent steps of the pathway. Furthermore, it can identify perturbed pathways under a variety of biological treatments, including cancer and inoculation with pathogens.

Our first goal with the current study was to extend the available dataset for the soybean metabolic pathways. In contrast to soybean, the metabolic network of the model plant *A. thaliana* (Mueller et al. 2003), is the best studied of all plant species. The complete genome sequences of *A.thaliana* and *G.max* are currently available (http://www.tair.org, http://www.phytozome.org). Based on the sequence

homology between these two species, we identified conserved proteins between their genomes. Following that, we used the data for the *Arabidopsis* metabolic pathways and their homologous proteins in soybean, in order to infer pathways potentially present in soybean. Our second goal was to study perturbation of the transferred pathways in soybean, in a variety of genetic backgrounds. We used the Pathway Perturbation Algorithm with the expression data from 8 soybean cultivars, and from a population created from an inter-specific cross of *G.max* with *G.sojae*. These soybean lines are genetically diverse, and have various levels of resistance to the pathogen *Phytophthora sojae*. In this manner, we identified pathways that are perturbed in response to infection by the pathogen, and therefore contribute to pathogen resistance by provoking cellular modifications involved in plant defense. Overall, our study points to the underlying metabolic mechanisms involved in pathogen resistance for *G. max*, and can help pinpoint genetic factors contributing to resistance of soybean to the *Phytophthora* pathogen.

3.3. Results

3.3.1. Pathway transfer from *A.thaliana* to soybean using sequence homology search.

We initially performed a BLASTP protein sequence comparison between the genomes of the two plant species, in order to identify *A. thaliana* proteins homologous to the ones predicted from the soybean genome assembly. A BLASTN nucleotide sequence similarity search was also performed between predicted soybean genes, and the soybean Unigenes that correspond to the probes of the Affymetrix Soybean Genome Array. For the BLASTN search we used an E-value threshold of less than 10e-200, while in the BLASTP comparison of soybean versus *Arabidopsis* proteins, the threshold was 10e-20. Additionally, we required at least 90% of each unigene sequence matching to a soybean gene, while 80% of each *Arabidopsis* protein was required to match its homologous protein from soybean. The results from both homology searches were stored in a relational database (http://www.postgres.org), for reconstructing the metabolic network of *Arabidopsis* in soybean.

Also added to the database were the 313 *A. thaliana* pathways from the Fall 2008 release of AraCyc (http://www.tair.org). We first identified pathways having at least a pair of adjacent metabolic reactions

with assigned *A. thaliana* genes, and 195 out of the 313 pathways met this criterion. This is a requirement by the Pathway Perturbation Algorithm, where genes need to form a network which is provided as input to the algorithm. From the remaining 118 pathways in AraCyc we where able to identify 26 which had assigned *A.thaliana* genes in at least a pair of reactions (see supplementary **SI.1**), but not in adjacent steps of the pathway. These reactions were treated as being adjacent during the pathway transfer in soybean, in order to increase the number of pathways available for the algorithmic analysis. We performed SQL (Navathe and Elmasri 2002) queries, to integrate the database tables containing the data for gene homologies, with the data for the AraCyc metabolic network. In total, 197 pathways out of the 221 (195 plus 26) were transfered in soybean, based on the gene homology between the two species. Besides the soybean gene identifiers throughout the metabolic network, we also added identifiers for the corresponding microarray probes, based on the data from the unigene sequence similarity search. This allowed us to superimpose the gene expression data on the metabolic pathways and perform the algorithmic analysis for the pathway perturbation.

3.3.2. Visualization of *A.thaliana* pathways found in soybean

We visualized the pathways transfered to soybean using Cytoscape v2.6.1 (http://www.cytoscape.org) as shown on **Fig 1**. The majority of the transferred pathways are found within a big cluster on sections A-H of **Fig 1**, which contains the part of the metabolic network responsible for sugar metabolism and secondary metabolite generation. More specifically, on A-B of **Fig 1**, are the pathways for sugars synthesis such as TCA cycle, glyconeogenesis, starch biosynthesis, oxidative and non-oxidative branches of pentose phosphate pathway, glyoxylate cycle, glutathione redox reactions and photosynthesis. On the other hand, section C contains the pathways responsible for sugar catabolism, such as sucrose degradation to pyruvate, inositol oxidation, homogalacturonan and galactose degradation, plus UDP-sugars interconversion. The right section F-H of the large cluster on **Fig 1**, contains pathways related to secondary metabolite production, including for example anthocyanin, cutokinin, cellulose, carotenoid, flavonoid, flavonol, phenyl-propanoid, giberelline, ethylene and IAA biosynthesis. Interesting is the fact that the sub-cluster containing the sugar catabolism pathways, is connected with the pathways for the secondary metabolism via the mannose/manitol degradation

pathway at point D. Three smaller clusters are located on sections I-K of **Fig 1**. In more detail, on section I of the metabolic network we identified pathways related to amino acids, including amine metabolism (urea cycle, superpathway of polyamine biosynthesis) and for amino acid biosynthesis (arginine, alanine, proline, serine, tryptophan, histidine). Cluster J contains pathways for the *de novo* biosynthesis of both purine/pyrimidine nucleotides and pyrimidine ribonucleotides. On cluster K, are pathways related to the metabolism of lipids as choline, phospholipids, but also glycolipids and triacyglycerol. The complete list of pathways for each section of the metabolic network from **Fig 1**. can also be found on **SI.1**.

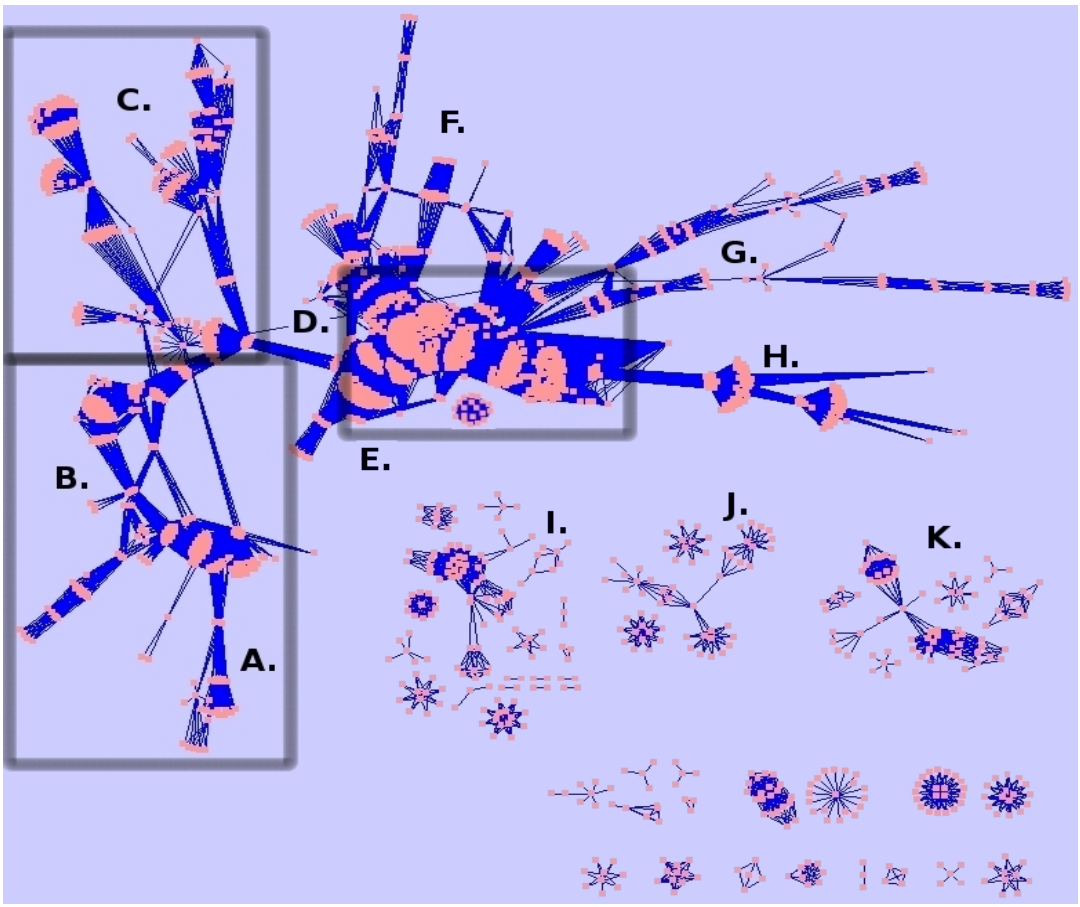


Fig.1 The 195 *Arabidopsis thaliana* pathways transferred in soybean based on gene homologies **(A-B).** Energy biosynthesis **(C).** Sugar catabolism **(D-H).** Secondary metabolism pathways. One the center of this cluster at **E.**, pathways related to disease resistance are found. **(I-K).** Amino acid, lipid, nucleotide metabolism

3.3.3. Algorithmic analysis of pathway perturbation using gene expression data

In order to study perturbation of the transfered pathways in soybean under different biological treatments, we used the Pathway Perturbation Algorithm (Rivera et al., in preparation). Given a set of pathways and gene expression measurements for a treatment and a control phenotype, this method identifies pathways that are most perturbed between the two phenotypes. In summary, the algorithm first computes the differential expression of each gene in treatment versus control, by performing a t-test and converting the t-score to a p-value. The Liptak-Stouffer Z-score statistic is used to aggregate p-values for genes that are connected in adjacent reactions along the metabolic network. In this manner, both the structure of the pathways and the differential expression of each gene are taken into account. The algorithm then maximizes the Z-score by identifying the most perturbed sub-pathway, based on a simulated annealing method. Finally, a permutation test is used, where the genes are replaced with randomly selected genes from the expression dataset, and the Z-scores are re-calculated for the randomized pathways. A p-value is computed as the fraction of random trials where the z-score of a permuted pathway, is larger that the z-score of the original pathway. The false discovery rate is controlled using the method of Benjamini and Hochberg (1995), and the adjusted p-value is used for reporting the pathway perturbation in all of the subsequent analysis.

As input to this algorithm we used gene expression data from eight cultivated soybean lines: Athow, Conrad, General, V71-370, Ox20-8, Sloan, Williams and plant introduction PI291237. Additionally, data from the expression profiling of a population of 297 Recombinant Inbred Lines (RIL) were used. Both datasets were generated as part of study for the partial resistance of soybean cultivars to the *Phytophthora sojae* pathogen (Zhou et al. 2009). More specifically, the Conrad, Athow, General and V71-370 lines show partial resistance to this pathogen, while the Ox20-8, Sloan, Williams and PI291237 are susceptible. The RIL population consisted of 297 individuals and was developed from an inter-specific cross between V71-370 (*Glycine max* line) and PI407162 (*Glycine soja* plant introduction), which are respectively resistant and susceptible. We then invoked the Pathway Pertubation algorithm using as input the 195 *Arabidopsis* pathways transfered to soybean, along with gene expression data for each of the soybean cultivars. The expression data consisted of measurements

on plant seedlings from each cultivar, 3 days after inoculation or mock-inoculation with the *P.sojae* pathogen.

For the RIL individuals, we used expression data only for the mock-inoculated plants, and we artificially created a treatment versus control input dataset. This was achieved by separating the population based on the marker genotype of each individual, under the major pathogen resistance Quantitative Trait Loci (QTL) region. These QTLs have been identified in the soybean chromosomes or Marker Linkage Groups, MLG-J, -G, -D1a and -I respectively (Tucker et al., in preparation). We used genotypic data from a previous study, where the same RIL individuals were assayed for Single Feature Polymorphism markers (SFPs, Krampis et al., in preparation) and created a genetic map. Since our population was based on a cross between V71-370 and PI407162 soybean lines, one or the other of these parental genotypes was identified for each individual under each QTL, and two groups were formed. Expression data for the two groups were provided as input to the algorithm, in lieu of the treatment versus control data of the eight cultivars. Finally, in order to establish a negative control, we also grouped the RIL individuals as described above, but used random loci from the rest of the twenty soybean chromosomes.

3.3.4. Results of pathway perturbation in different genetic backgrounds

Using expression data from an RIL population of 297 individuals, we calculated pathway perturbation from differential expression between two groups of RILs. Each group was formed based on the parental genotype (V71-370 or PI407162) found in the RIL individuals, under the major disease-resistance QTLs on the MLG-J, -G, -D1a and -I. The output of the Pathway Perturbation Algorithm was visualized using Cytoscape (http://www.cytoscape.org), and pathways were colored on **Fig.2** based on their perturbation p-values. Red color was used for p-values less than 0.01, while pathways were painted light brown for p-values between 0.01 and 0.05. Additional figures for randomly chosen loci in the rest of soybean MLGs, are provided on **SI.2**. Pathways with perturbation p-values less than 0.01 in at least one of the eight cultivars (see details below) or RIL datasets, are also summarized on **Table.1**. Overall, for all MLGs the majority of perturbed pathways were located in the region of the metabolic network for plant secondary metabolism (sub-cluster E. of **Fig.1**). In the case of MLG-J, pathways with

p-value less than 0.01 are depicted with red color on **Fig.2a**, and include suberin, free phenylpropanoid acid, carotenoid, flavonol and leuco-pelargonidin / cyanidin biosynthesis, leucine and valine degradation. For the same linkage group pathways with p-values between 0.01 and 0.05 are shown in light brown color on **Fig.2a**, and include IAA biosynthesis I, phenylpropanoid biosynthesis and isoleucine degradation. The exact p-values for the pathways on MLG-J are shown on **Table 1**. In the case of MLG-G and in similar colors to MLG-J, perturbed pathways with p-value less than 0.01 include the suberin and free phenylpropanoid acid biosynthesis, and are shown with arrow in center of **Fig.2b**. Additionally, the flavonol biosynthesis with p-value just above 0.01, is also common between these two MLGs, while the glucosinolate breakdown pathway is unique to MLG-G. Next, for MLG-D1a the leuco-pelargonidin / cyanidin biosynthesis pathway is perturbed similarly with MLG-J, albeit at a lower p-value between 0.01 and 0.05. As shown on **Table 1** and also colored light brown in the center of **Fig.2c**, the gibberellin biosynthesis (I, II, III), gibberellin inactivation and anthocyanin biosythesis pathways are perturbed only in MLG-D1a, but in none of the other MLGs. In addition, MLG-D1a displays strong perturbation with p-value less than 0.01 of the chorismate and phenyalanine, tyrosine, tryptofan biosynthesis pathways (red on **Fig.2c**), which are also unique in this linkage group. Furthermore, the trehalose biosynthesis pathway is perturbed only on MLG-D1a and is shown with an arrow on **Fig.2c**. Concerning MLG-I on **Fig.2d** no pathways were found significantly perturbed, except the suberin and free phenylpropanoid acid biosynthesis slightly above p-value 0.01, similar to MLG-J and -G (**Table 1**). Finally, as negative control we randomly chose loci from MLG-A1 and MLG-N. In this case, only few pathways showed weak perturbation such as valine biosynthesis, SAM cycle, purine and tri-hydrofolate biosynthesis, which were painted light brown and are indicated with arrows on **Fig.2e** and **Fig.2f**. Similar results with pathways perturbed at very low levels and spread throughout the metabolic network, were found for random loci chosen from each of the soybean chromosomes (Supplementary Information **SI2**). Finally, the phytyl-PP biosynthesis pathway was found strongly perturbed in all linkage groups as shown on **Table 1**. This is not surprising since this pathway is essential in the developmental processes of the young seedlings used in our study, such as membrane biosynthesis (Keller et al. 1998).

Table 1. Pathways with perturbation p-values < 0.01 for the 8 cultivar and RIL data. Underlined lettering indicates perturbation values with one order of magnitude lower than this threshold, which are less than 0.001.

	Athow	Conrad	General	Ox20-8	PI291	Sloan	V71	Williams	MLG-J	MLG-G	MLG-D1a	MLG-I
phytyl-PP biosynthesis	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.13E-05***	***5.13E-05***	***5.13E-05***	***5.13E-05***
suberin biosynthesis	***5.16E-05***	***5.17E-05***	1.49E-03	***5.13E-05***	***5.13E-05***	***5.14E-05***	***2.56E-04***	***5.13E-05***	***4.62E-04***	2.67E-03	---	1.89E-02
free phenylpropanoid acid biosynthesis	***5.16E-05***	***5.17E-05***	1.96E-03	***5.13E-05***	***5.13E-05***	***5.14E-05***	***3.59E-04***	***5.13E-05***	***5.64E-04***	3.03E-03	---	1.99E-02
carotenoid biosynthesis	7.80E-03	---	---	4.76E-02	1.09E-02	---	---	2.21E-03	1.59E-03	---	---	---
leucine degradation	---	---	---	---	---	---	---	---	2.15E-03	---	---	---
flavonol biosynthesis	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	7.33E-03	2.31E-02	---	---
valine degradation	---	1.24E-03	---	---	---	---	---	6.75E-02	8.00E-03	---	---	---
leucopelargonidin and leucocyanidin biosynthesis	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	9.74E-03	---	2.99E-02	---
photosynthesis, light reaction	---	5.37E-03	---	2.21E-02	---	1.04E-02	6.00E-03	4.64E-02	1.09E-02	---	---	7.44E-03
IAA biosynthesis I	---	4.81E-03	9.00E-03	3.49E-03	***9.23E-04***	3.83E-02	---	1.62E-02	1.27E-02	---	---	---
phenylpropanoid biosynthesis	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	1.46E-02	---	---	---
isoleucine degradation	---	***3.10E-04***	---	1.84E-02	---	4.37E-02	---	1.16E-02	3.42E-02	---	---	---
gibberellin biosynthesis I (non C-3, non C-13 hydroxylation)	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	---	---	2.67E-02	---
gibberellin biosynthesis II (early C-3 hydroxylation)	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	---	---	3.03E-02	---
GA12 biosynthesis	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	---	---	---	---
superpathway of GA12 biosynthesis	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	---	---	---	---
gibberellin biosynthesis III (early C-13 hydroxylation)	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	---	---	3.58E-02	---
gibberellin inactivation	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	---	---	3.69E-02	---
superpathway of gibberellin biosynthesis	***5.16E-05***	***5.17E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	---	---	3.57E-02	---
anthocyanin biosynthesis	***5.16E-05***	***5.17E-05***	***5.14E-05***	***2.56E-04***	1.08E-03	***5.14E-05***	***5.13E-05***	***5.13E-05***	---	---	4.11E-02	---
acetyl-CoA biosynthesis (from pyruvate)	***5.16E-05***	***5.17E-05***	***1.54E-04***	2.72E-03	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	---	---	---	---
superpathway of acetyl-CoA biosynthesis	***5.16E-05***	***5.17E-05***	***1.54E-04***	2.72E-03	***5.13E-05***	***5.14E-05***	***5.13E-05***	***5.13E-05***	---	---	---	---
ethylene biosynthesis from methionine	***5.16E-05***	1.50E-03	***5.14E-05***	9.44E-03	2.10E-03	3.45E-03	***6.15E-04***	5.64E-03	---	---	---	---
glycolysis, pyruvate dehydrogenase, TCA, glyoxylate bypass	***1.55E-04***	***5.17E-05***	***5.14E-05***	***9.23E-04***	***5.13E-05***	***5.14E-05***	***5.13E-05***	***3.59E-04***	---	---	---	---
flavonoid biosynthesis	***1.55E-04***	***5.17E-05***	***5.14E-05***	1.54E-03	***6.67E-04***	***2.06E-04***	***5.13E-05***	***1.03E-04***	---	---	---	---
salvage pathways of purine nucleosides	***2.07E-04***	1.36E-02	***5.14E-04***	---	3.59E-03	1.70E-03	***5.13E-05***	4.36E-03	---	---	---	---
chorismate biosynthesis	***4.13E-04***	***9.30E-04***	***8.75E-04***	***3.59E-04***	***1.03E-04***	***4.11E-04***	***7.18E-04***	1.90E-03	---	---	1.08E-03	---
phenylalanine, tyrosine and tryptophan biosynthesis	***5.16E-04***	***4.65E-04***	***8.75E-04***	***4.62E-04***	***5.13E-05***	***2.06E-04***	2.77E-03	2.31E-03	---	---	2.21E-03	---
TCA cycle	1.03E-03	***8.78E-04***	2.21E-03	2.29E-02	***7.18E-04***	1.54E-03	***5.13E-04***	3.74E-03	---	---	---	---
trans,trans-farnesyl diphosphate biosynthesis	1.65E-03	1.40E-03	***6.69E-04***	9.33E-03	9.49E-03	6.64E-03	1.90E-03	2.31E-03	---	---	---	---
oxidative ethanol degradation	3.41E-03	2.11E-02	***7.20E-04***	---	---	---	***5.13E-04***	---	---	---	---	---
glucosinolate breakdown	3.98E-03	---	---	---	3.92E-02	---	---	7.72E-02	---	3.91E-02	---	---
leucine biosynthesis	5.99E-03	---	1.08E-02	---	1.32E-02	---	---	1.26E-02	---	---	---	---
purine degradation	6.92E-03	---	1.13E-03	---	---	---	***4.10E-04***	---	---	---	---	---
cytokinins-O-glucoside biosynthesis	8.62E-03	---	1.08E-02	---	---	3.86E-03	2.77E-03	1.49E-03	---	---	---	---
geranylgeranyldiphosphate biosynthesis II (plastidic)	1.08E-02	4.50E-03	4.63E-03	2.08E-02	1.13E-02	1.25E-02	1.19E-02	6.36E-03	---	---	---	---
glyoxylate cycle	1.10E-02	4.55E-03	1.59E-02	---	2.92E-03	4.58E-03	1.14E-02	1.14E-02	---	---	---	---
non- / oxidative branches of pentose phosphate pathway	1.24E-02	1.53E-02	3.24E-03	7.95E-03	2.62E-03	1.49E-03	***1.03E-04***	---	---	---	---	---
phenylalanine biosynthesis	1.28E-02	---	---	3.33E-03	1.16E-02	---	1.17E-02	---	---	---	---	---
cellulose biosynthesis	1.41E-02	3.80E-02	***3.60E-04***	3.50E-02	***5.13E-05***	1.76E-02	***1.03E-04***	***5.13E-05***	---	---	---	---
superpathway of gluconate degradation	1.62E-02	3.61E-02	1.96E-03	3.80E-02	3.33E-02	2.47E-03	5.54E-03	8.46E-03	---	---	---	---
branched-chain alpha-keto acid dehydrogenase complex	1.85E-02	5.48E-03	2.63E-02	1.25E-02	8.62E-03	7.61E-03	---	4.41E-03	---	---	---	---
non-oxidative branch of the pentose phosphate pathway	2.01E-02	---	---	2.57E-02	1.39E-02	---	3.08E-03	5.13E-02	---	---	---	---
abscisic acid biosynthesis	2.28E-02	2.94E-02	3.01E-02	1.21E-02	2.67E-03	1.21E-02	1.79E-03	1.14E-02	---	---	---	---
trehalose biosynthesis	3.59E-02	3.20E-03	1.58E-02	4.51E-03	1.90E-02	5.55E-03	7.95E-03	4.15E-03	---	---	3.39E-02	---
sucrose degradation to ethanol and lactate (anaerobic)	---	1.40E-03	---	7.59E-03	***5.13E-05***	***1.03E-04***	***1.03E-04***	---	---	---	---	---
superpathway of sucrose degradation to pyruvate	---	1.03E-02	---	9.38E-03	***5.13E-05***	***5.14E-05***	1.85E-03	7.73E-02	---	---	---	---
formylTHF biosynthesis	---	1.74E-02	---	1.39E-02	4.92E-03	1.50E-02	---	---	---	---	---	---
sucrose degradation	---	1.81E-02	---	1.95E-02	6.72E-03	***3.60E-04***	---	2.42E-02	---	---	---	---
fatty acid beta-oxidation II (unsaturated, even number)	---	***5.17E-05***	---	2.71E-02	---	---	---	3.85E-02	---	---	---	---
glutamate degradation II	---	1.12E-02	4.94E-03	2.74E-02	1.62E-02	2.84E-02	1.49E-02	9.74E-03	---	---	---	---
glycolysis II (plant plastids)	---	6.56E-03	---	2.79E-02	***2.56E-04***	***3.09E-04***	***5.13E-05***	---	---	---	---	---
fatty acid beta-oxidation III (unsaturated, odd number)	---	***5.17E-05***	---	2.80E-02	---	---	---	3.84E-02	---	---	---	---
polyisoprenoid biosynthesis	---	1.08E-02	8.85E-03	3.18E-02	---	4.34E-02	---	---	---	---	---	---
photosynthesis	---	1.06E-02	1.00E-02	3.51E-02	4.31E-03	1.02E-02	***2.05E-04***	---	---	---	---	---
lysine degradation II	---	***1.03E-04***	---	3.53E-02	---	---	---	5.07E-02	---	---	---	---
glycolysis I (plant cytosol)	---	3.78E-02	---	4.38E-02	1.90E-03	***4.63E-04***	3.44E-03	---	---	---	---	---
quercetin glucoside biosynthesis	---	---	---	---	---	4.37E-03	4.05E-02	5.49E-03	---	---	---	---
Calvin cycle	---	3.67E-02	2.17E-02	---	6.10E-03	2.12E-02	1.08E-03	---	---	---	---	---
kaempferol glucoside biosynthesis	---	---	---	---	---	4.37E-03	4.05E-02	5.49E-03	---	---	---	---
gluconeogenesis	---	---	---	---	6.05E-03	6.38E-03	2.26E-03	---	---	---	---	---
fatty acid beta-oxidation I (saturated)	---	1.65E-03	---	---	---	---	---	6.05E-02	---	---	---	---
superpathway of starch degradation to pyruvate	---	---	---	---	3.95E-03	3.81E-03	1.64E-03	---	---	---	---	---
dolichyl-diphosphooligosaccharide biosynthesis	---	---	---	---	---	4.11E-03	---	7.13E-03	---	---	---	---
mannose degradation	---	---	---	---	---	9.05E-03	2.27E-02	3.51E-02	---	---	---	---
oxidative branch of the pentose phosphate pathway	---	---	---	---	4.92E-02	9.31E-03	1.12E-02	3.38E-02	---	---	4.39E-02	---
ureide degradation	---	3.16E-02	2.72E-02	---	6.05E-03	4.17E-03	1.65E-02	5.13E-02	---	---	---	---
glycine biosynthesis	---	---	---	---	---	6.74E-03	---	4.74E-02	---	---	---	---
chlorophyll a biosynthesis II	---	4.52E-02	---	---	---	***8.74E-04***	---	---	---	---	---	---

For the eight cultivars, extended pathway perturbation was found in comparison to the four linkage linkage groups above, as shown on **Fig.3a-h** and **Table 1**. This was not surprising, since the input dataset for the Pathway Perturbation Algorithm consisted of gene expression from mock versus inoculated plants of the eight cultivars, instead of only mock data used in the case of the RILs. This resulted in greater contrasts and therefore more significant p-values for the pathway perturbation (see also Methods section for algorithm details). For all the soybean cultivars and similarly with the linkage groups, we observed increased perturbation (p-values less than 0.01, colored as red) in the central cluster containing the secondary metabolism pathways. More specifically, these pathways include suberin, free phenylpropanoid acid, flavonol, leuco-pelargonidin / cyanidin, phenylpropanoid, gibberellin biosynthesis I, II, III and inactivation, anthocyanin, acetyl-CoA, ethylene, flavonoid, chorismate, cellulose, abscisic acid and trehalose biosynthesis. The carotenoid biosynthesis pathway from the same group of secondary metabolism pathways, was found only perturbed on Athow, Ox20-8, PI291237 and Williams as shown on **Table 1**. The IAA biosynthesis was perturbed in all cultivars except Athow and V71-370. Concerning the Ox20-8, PI291237, Sloan, V71-370 and Williams cultivars, they are susceptible to the *P.sojae* pathogen,while Athow, Conrad and General display some resistance (Dorrance A., unpublished results). Differences between these two groups of cultivars were observed mainly for the sugar metabolism pathways, displayed on the lower half of **Table 1**, and also indicated by arrows on the left of **Fig.3a-h**. More specifically, the super-pathways of aerobic sucrose degradation to ethanol and lactate, and of sucrose degradation to pyruvate, were found perturbed only in General, Ox20-8, PI291237, Sloan, V71-370 and Williams cultivars. Similarly for the gluconeogenesis, starch degradation to pyruvate and oxidative branch of the pentose phosphate pathway, are perturbed in PI291237, Sloan and V71-370 as shown on the lower half of **Table 1**. Finally, the quercetin and kaempferol glucoside biosynthesis, where perturbed only in Sloan, V71-370 and Williams. As with the data in the four linkage groups, the phytyl-PP biosynthesis was found perturbed in each of the eight soybean cultivars.

Fig.2 Pathway perturbation of the 195 *Arabidopsis* pathways transferred in soybean, using RIL population data, for the major resistance QTLs to *P.sojae*. Red colors are for perturbation p-value < 0.01 and light brown for p < 0.05. See text for more details on the pathways depicted here.

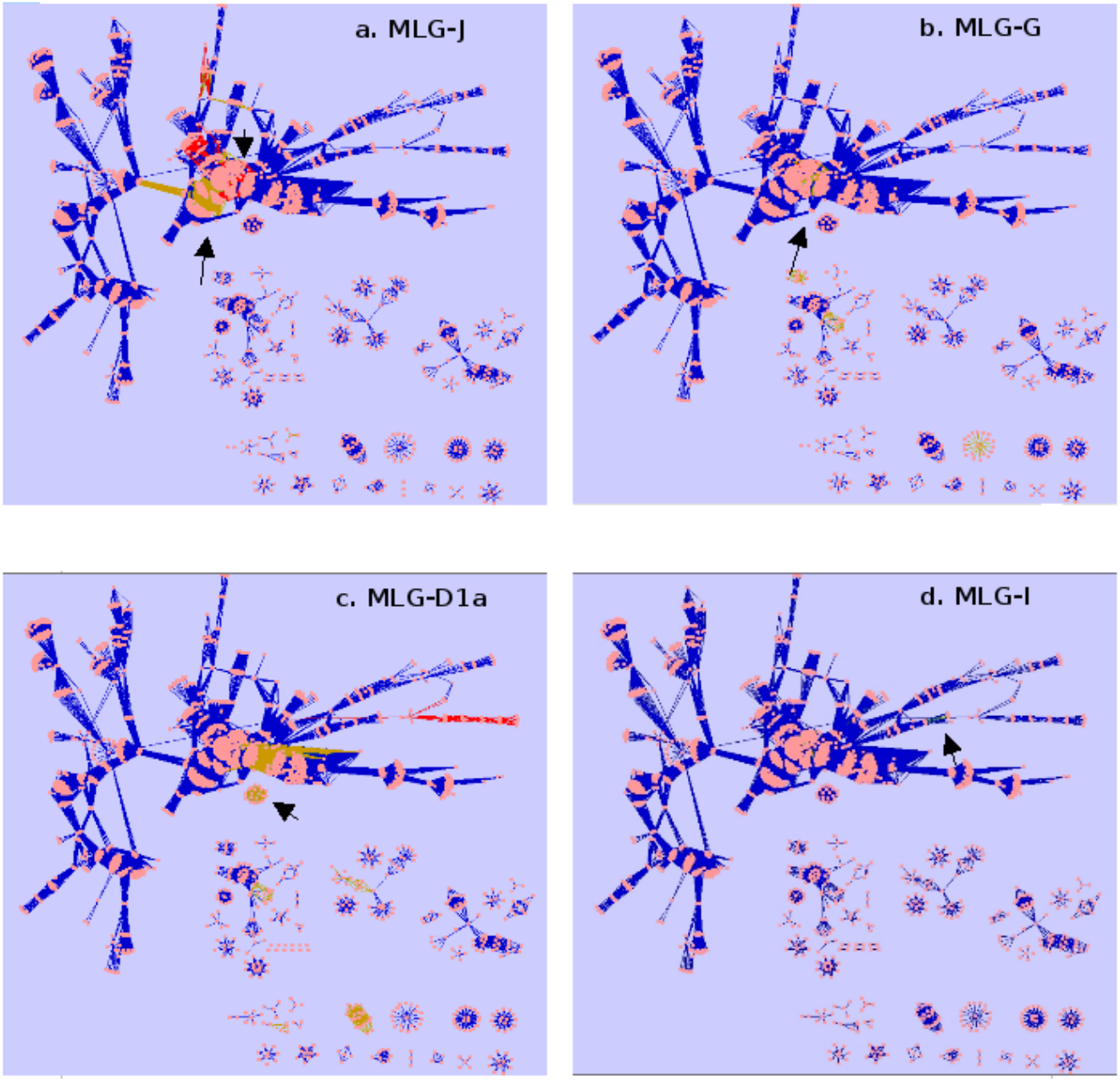


Fig.2 (continued)

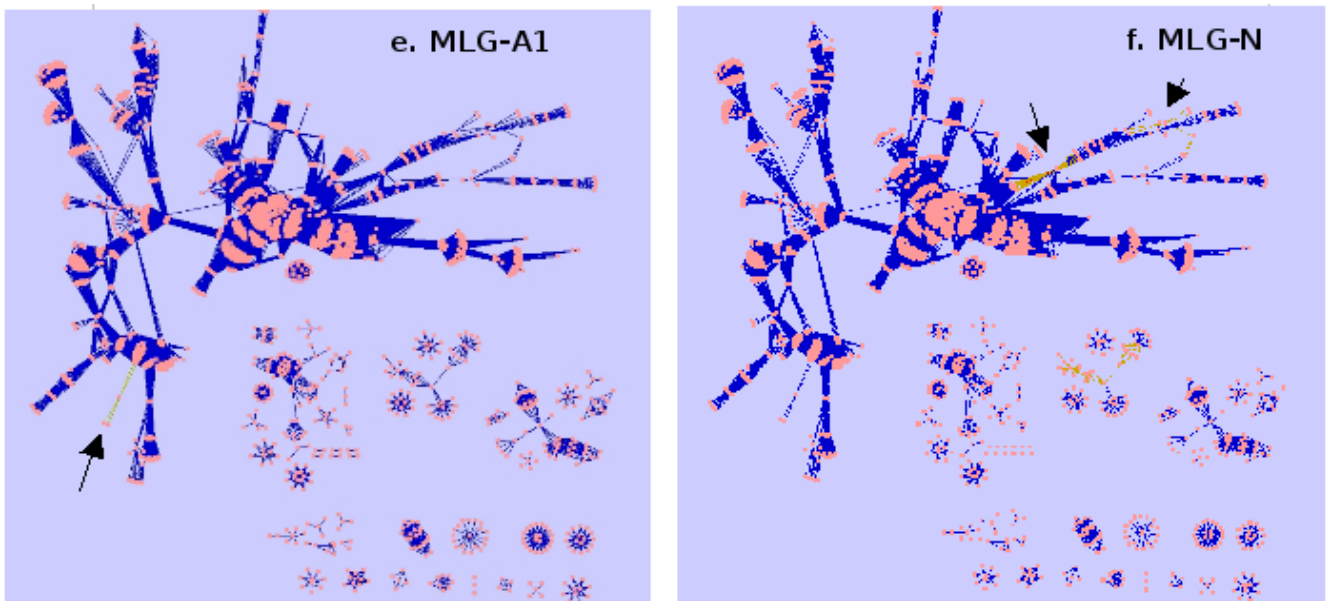


Fig.3 Pathway perturbation of the 195 *Arabidopsis* pathways transferred in soybean, using data from the 8 soybean cultivated lines. Red colors are for perturbation p-value < 0.01 and light brown for p < 0.05. See text for more details on the pathways and cultivars depicted here.

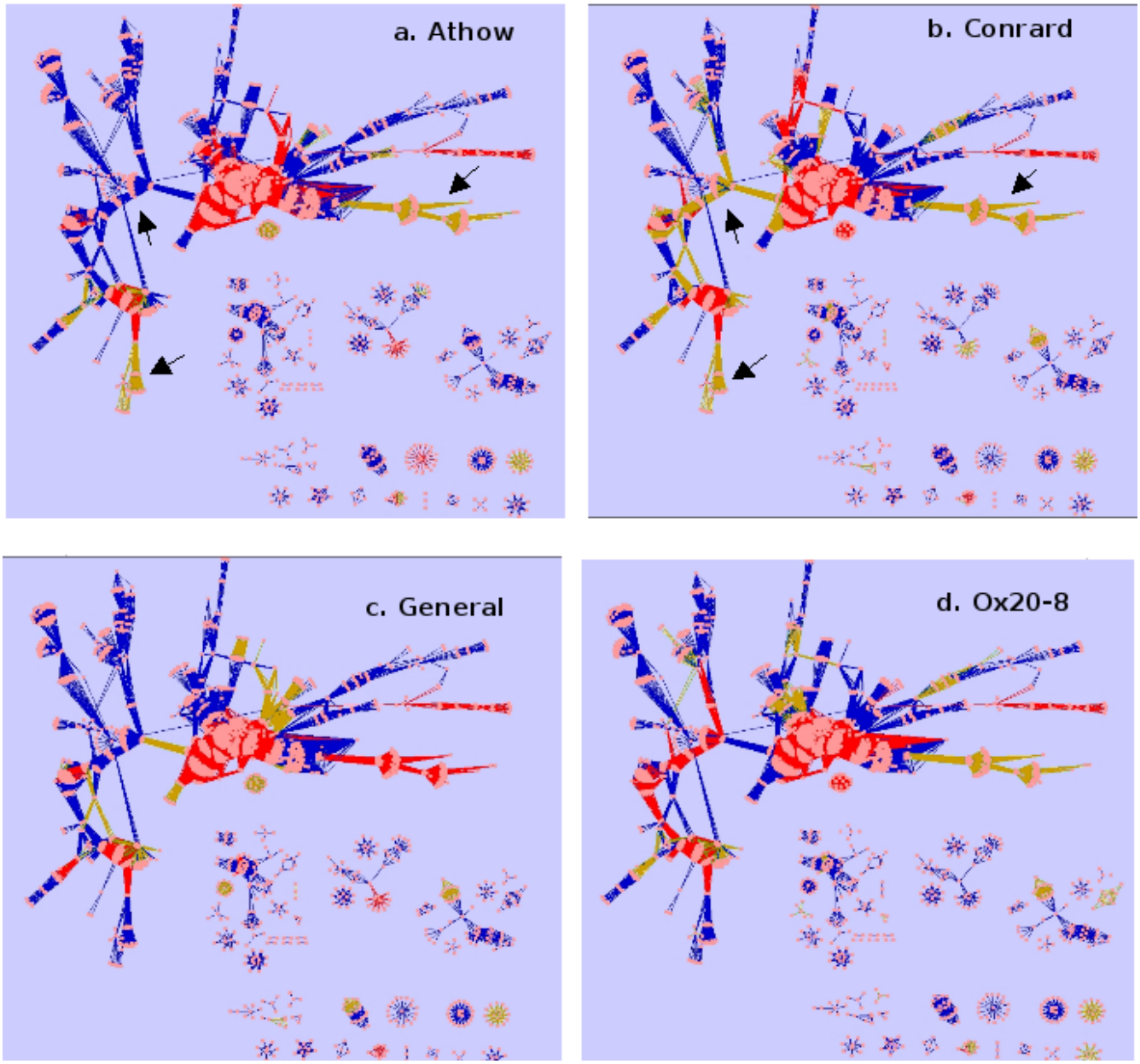
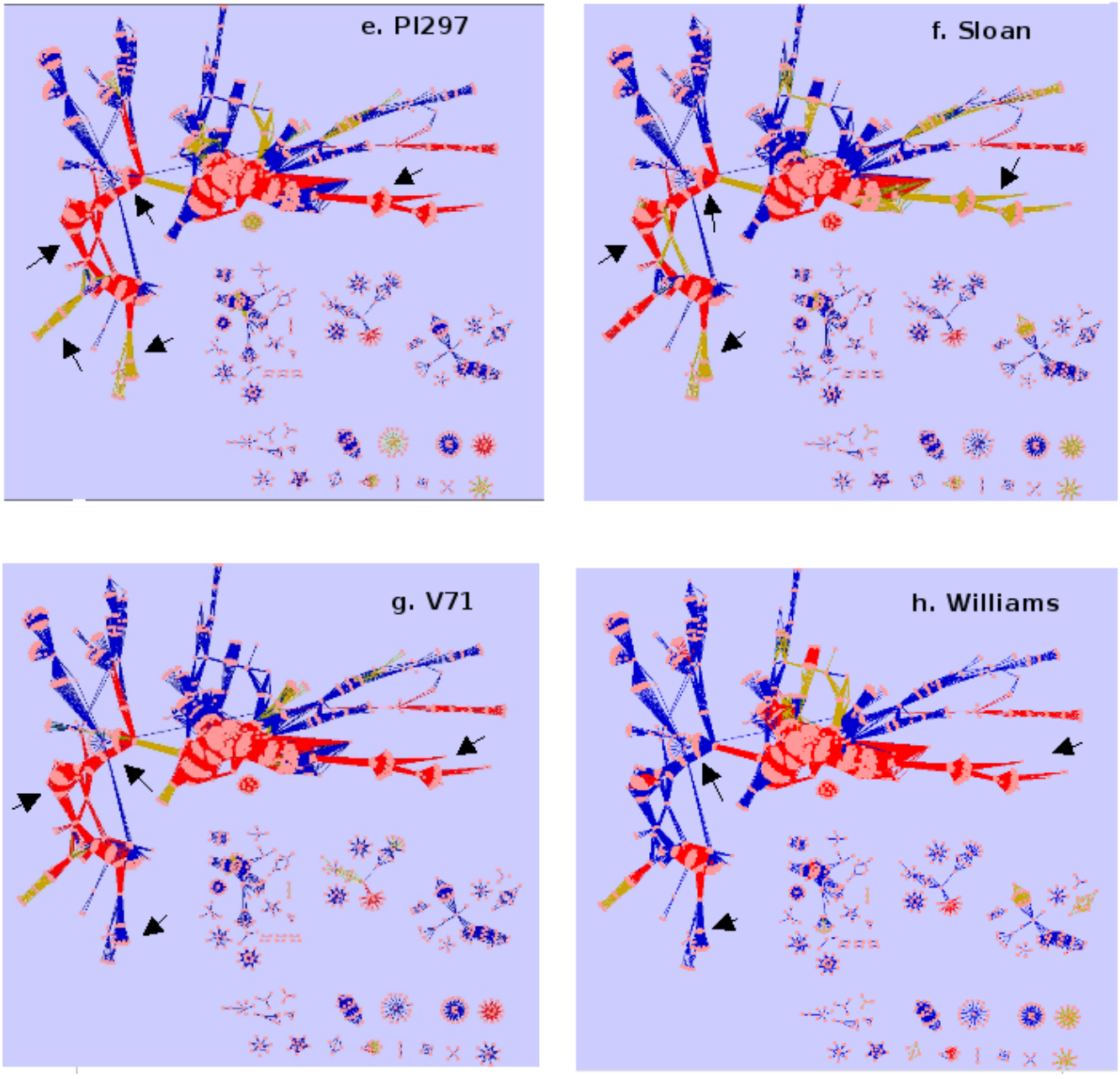


Fig.3 (continued)



3.3.5. Details of pathway perturbation on the gene level

In order to identify differences in pathway perturbation at the gene level, we visualized the up- or down-regulation of gene expression in the datasets used by the Pathway Perturbation Algorithm. We chose the chorimate biosynthesis pathway, since it showed consistent perturbation across all soybean cultivars and at least for one of the linkage groups. For this purpose, we calculated the t-scores for the differential expression of the genes in each pathway between pathogen versus mock inoculated plants in the eight cultivars, or between RIL groups with PI407162 versus V71-370 genotype. Based on the direction of our comparisons, positive t-score denotes more expression in inoculated cultivars or the PI407162 group of RILs, while negative corresponds to more expression in the mock or V71-370 group of RILs We then converted the t-scores to p-values and using the Kyoto Encyclopedia of Genes and Genomes (KEGG), genes with differential expression at p-value < 0.01 were visualized. Each enzyme of the KEGG pathways corresponding to our genes, was painted red or green depending on whether the t-score was respectively positive or negative. All the visualizations are shown on **Fig.4a-g**, while also on **Fig.4a** additional pathways are depicted in the vicinity of chorismate biosynthesis on the metabolic network.

The Athow, Conrad and General cultivars, display the same pattern of gene up- and down- regulation. More specifically, the genes coding for the enzymes 3-deoxy-7-phosphoheptulonate synthase (Enzyme Commission number, EC: 2.5.1.54), 3-dehydroquinate synthase (4.2.3.4) and 3-phosphoshikimate 1-carboxyvinyltransferase (2.5.1.19), were found to be down-regulated during infection in these cultivars, shown with green color on **Fig.4a**. On the same figure, the 3-dehydroquinate dehydratase I (4.2.1.10) and shikimate 5-dehydrogenase (1.1.1.25) enzymes where up-regulated during infection with *P. sojae*, therefore where colored red. Similar are the results for PI292371, with the addition of shikimate kinase (2.7.1.71) that was found down-regulated, as shown on **Fig.4b**. The Ox20-8, Sloan and V71-370 cultivars also display down-regulation of the shikimate kinase gene as shown on **Fig.4c**. Unlike Athow, Conrad, General and PI291237 there was not a change in the expression of 3-deoxy-7-phosphoheptulonate synthase (2.5.1.54) in either Ox20-8, Sloan or V71-370. It is interesting to note the result of **Fig.4d** and **e**, where cultivar Williams and MLG-J, display opposite regulation with each other, of the exact same enzymes (2.5.1.54, 4.2.3.4, 2.5.1.19).

Fig.4 Chorismate biosynthesis pathway, with red color for more expression of the genes ($p<0.01$) in inoculated cultivars or PI407162 group of RILs, while green is for more expression in the mock or V71-370 group of RILs. The chorismate pathway is included within the dashed rectangle, connecting pathways are outside. **a.** Athow, Conrad, General **b.** PI291237 **c.** Sloan, V71-370 **d.** Williams **e.** MLG-J **f.** MLG-D1a **g.** MLG-I.

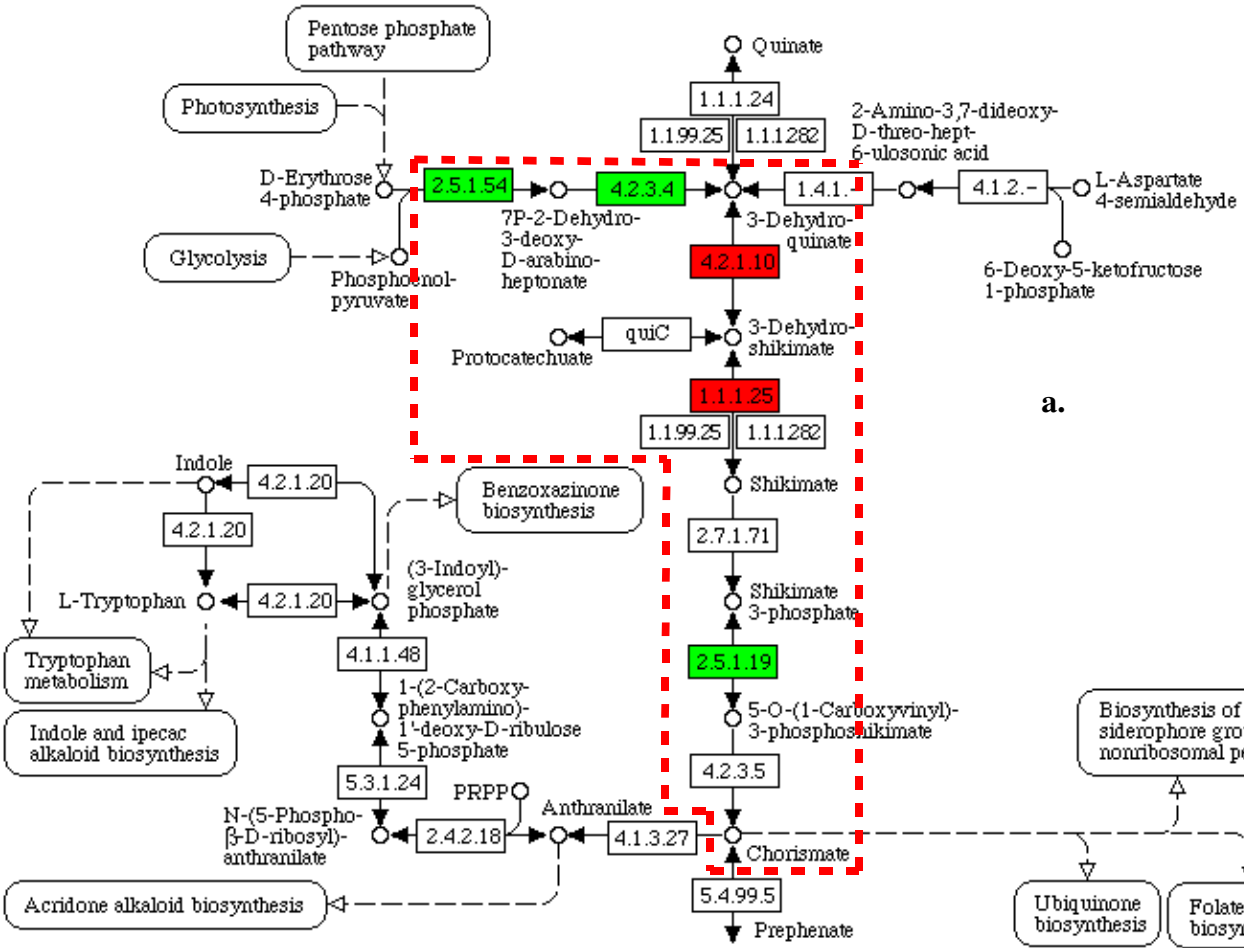
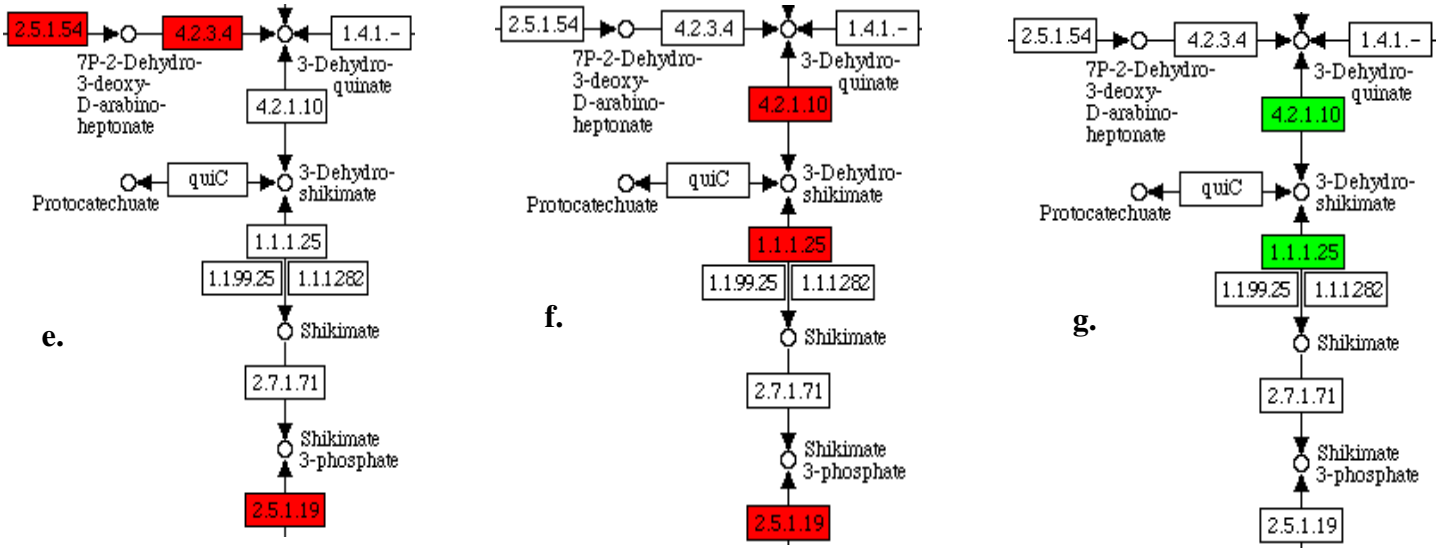


Fig.4 (continued)



More specifically on **Fig.4d** and **e** are depicted with green or red colors respectively, the genes down-regulated during inoculation of seedlings from the Williams cultivar, or up-regulated in RIL individuals with PI407162 genotype respectively under the QTL in MLG-J. The chorismate biosynthesis pathway, is not perturbed in MLG-J, and for Williams it shows one order of magnitude less perturbation than the rest of the cultivars (**Table 1**.). On **Fig.4f** for MLG-D1a, the 3-phosphoshikimate 1-carboxyvinyltransferase (2.5.1.19) gene was up-regulated in RIL individuals with PI407162 genotype similar to MLG-J. On the other hand, MLG-D1a showed up-regulation of the 3-dehydroquinate dehydratase I (4.2.1.10) and shikimate 5-dehydrogenase (1.1.1.25), similarly to the eight soybean cultivars. As shown on **Table 1**., MLG-D1a was the only linkage group to have chorismate biosynthesis pathway found perturbed by the Pathway Perturbation Algorithm. In MLG-I while this pathway is not perturbed, the pattern was different from all cultivars and MLGs, with only two genes up-regulated in the group of RILs with V71-370 genotype in this locus (4.2.1.10, 1.1.1.25). Finally, MLG-G did not show any changes in the expression of the genes in the chorismate pathway, while also in this case the pathway was not found perturbed by the algorithm.

3.4. Discussion

3.4.1. *Arabidopsis* pathways transferred in soybean based on gene homology

Currently, only a sparse dataset exists for the metabolic and regulatory pathways of soybean (Soybase, http://www.soybase.org). Furthermore, this dataset has a small number of pathways related to disease resistance against pathogens, while none are available in a standard data format. Our goal with the current study was first to fill gaps in the metabolic network of soybean, using the abundant pathway data of the model plant *Arabidopsis thaliana*, from the AraCyc database (Mueller et al., 2003). The AraCyc database was originally created by computationally predicting plant metabolic pathways, using the Pathway Tools software (Karp et al. 2002). In summary, this software predicts pathways for an organism by matching the annotation names of its genes to enzyme names in the Metacyc reference database (Caspi et al. 2008). Since the soybean genome has only been recently sequenced and its annotation is still at an early stage, in our study we chose to perform metabolic reconstruction in soybean based on gene homologies with the *Arabidopsis* genome. Using a protein sequence similarity search between the two species, we were able to identify 197 *Arabidopsis* metabolic pathways that are potentially present in soybean. While the Fall 2008 version of AraCyc used in our study contains 313 pathways (http://www.tair.org), only 221 contain at least a pair of assigned genes to be practically useful for any analysis. Therefore, approximately 90% of the *Arabidopsis* pathways containing identified genes, were transferred to soybean. Additionally, we integrated the transferred network in soybean with the unigene identifiers from the Affymetrix Soybean Genome microarray, to the predicted genes in the soybean genome assembly. This allowed us to superimpose gene expression data on the metabolic network, and calculate perturbation of each pathway using the Pathway Perturbation Algorithm.

Overall, we found *Arabidopsis* pathways from all categories of metabolism having homologous genes in soybean. Examples include sugar anabolism and catabolism, secondary metabolites as well as amino acid and nucleotide biosynthesis. On the website hosting the AraCyc database, the complete metabolism of *Arabidopsis* is hierarchically divided based on the overall metabolic functions in the cell. For example, one category termed Biosynthesis and Generation of Precursor Metabolites includes

both the Amines/Polyamines and Carbohydrates Biosynthesis. On the other hand, in our visualization of the pathways of *A.thaliana* transferred in soybean, we clustered the pathways based on the nature of biochemical conversions taking place in each pathway. In this manner, we formed a section in our graph for the part of the metabolic network responsible for sugar metabolism, another section with the secondary metabolite pathways, another with amino acid biosynthesis and so forth. This provided us with the advantage of being able to visualize directly which cellular functions are altered in each biological condition, based on the results from the Pathway Pertubation algorithm. Overall, the transferred network in soybean had good connectivity, with the sugars and secondary metabolite pathways forming a large cluster connected by the mannitol degradation pathway, while amino acid, nucleotide and lipid metabolism pathways aggregated in smaller clusters .

3.4.2. Algorithm for identifying pathway perturbation in soybean

For our study of pathway perturbation in soybean, we chose the Pathway Perturbation Algorithm (Corban et al., in preparation) since this method does not require every gene in a pathway to be differentially expressed. If only some part of a pathway is significantly perturbed under a certain biological treatment, it still can be identified since this algorithm calculates the combined effect of multiple expression changes in the genes of a pathway. The advantage of this approach is that it covers cases where not all genes in the pathway alter their expression levels in response to a treatment, such as when biological modifications take place at the post-transcriptional or port-translational level. Furthermore, the algorithm by Rivera et al. computes pathway perturbation by summarizing changes in gene expression along each pathway based on the topology of the metabolic network, by taking into account the connectivity of each gene with its adjacent genes in the pathway. One such example are genes that act as “hubs” in the metabolic network, meaning that they iteract with many genes within the pathway and their perturbation can have pleiotropic effects (Jeong et al. 2000).

Other methods also exist that identify perturbed pathways, using differential gene expression between biological treatments. One of the approaches commonly used is Active Modules algorithm, originally developed by Ideker et al. (2002). In the study by Rivera et al. (in preparation) it was shown that the statistical significance achieved using the Pathway Perturbation Algorithm, exceeds the significance of

the most perturbed pathways identified using the Active Modules algorithm. Another common method used is the Gene Set Enrichment Analysis (GSEA, Subramanian et al. 2005). This approach compares two phenotypes by first sorting all the genes, from most to least differentially expressed based on a t-test between the two biological treatments. Then given a gene set of interest such as those participating in a biological pathway, GSEA tests whether these genes are ranked toward the top or the bottom of the sorted list. Rivera et al. showed that significant sets of genes might be missed by GSEA, since in the perturbed pathways genes can be up- or down regulated with high positive or negative t-scores respectively, ranking separately on top or bottom of the list. In addition, GSEA uses a strict null hypothesis that distribution of the t-scores in a particular gene set is the same as the distribution of the rest of the genes in the set. On the other hand, the Pathway Perturbation Algorithm uses a less strict null hypothesis based on the distribution of randomly-selected genes from the complete set of genes measured with the microarray, and which are not included in the pathways.

3.4.3. Soybean pathway perturbation in different genetic backgrounds

In order to study pathway perturbation, we used the Pathway Perturbation Algorithm and analyzed expression data collected before and after infection with the *P. sojae* pathogen in the eight soybean cultivars. Additionally, we used the same algorithm in order to identify pathway perturbation between two groups of individuals in a soybean RIL population. This population was created by an inter-specific cross of *G. max* (V71-370 line) to *G. sojae* (PI407162 line), and it was grouped based on the parental genotypes under the major disease resistance QTLs on MLG-J, G, D1a, I (Tucker et al. in preparation, Krampis et al. in preparation). Overall, we observed more pathway perturbation on the soybean cultivars versus the RIL groups. This was expected, since the input dataset for the eight cultivars consists of gene expression from mock versus inoculated plants, resulting in greater contrasts between expression levels in the two treatments. Conseqeuently, the increased contrast provides better signal to noise ratio for the algorithm to identify perturbed pathways, than in the case of the RILs where only mock data were used.

We found different perturbed pathways in the cultivars versus the MLGs, while also variation was observed among the four different MLGs. More specifically the giberellin metabolism, anthocyanin and

chorismate biosynthesis pathways were perturbed in all the cultivar data, but only MLG-D1a showed perturbation with the RIL data. On the other hand, the suberin and phenylpropanoid biosynthesis pathways were identified perturbed in all MLGs except D1a. These results suggest a possibly unique genetic element in the resistance QTL located in MLG-D1a, controlling the giberellin metabolism, anthocyanin and chorismate biosynthesis pathways. In a recent study (Hu et al. 2009), it was shown that the chorismate pathway provides resistance to powdery mildew in barley, while also this compound is a precursor of salicylic acid (Wildermuth et al. 2001), an important hormone signaling molecule for plant defense. Athocyanins are important antioxidants, and it has been shown that they are related to controlling the rate of cell turnover or otherwise apoptosis, effectively making malignant cell cells die faster (Hou 2003). This suggests that elements in the chromosomal region under MLG-D1a, play a role in disease resistance through controlling pathways responsible for signaling affecting cell fate. Cellular apoptosis in plants is a major mechanism employed in the defense against pathogens, through the hypersensitive response and formation of necrotic lesions (Heath 2000). On the other hand, suberin and phenylpropanoid biosynthesis were found perturbed based on data from all linkage groups except D1a. These are essential metabolites for the structure of cell walls (Graca and Pereira 2000), which are the major barrier to pathogen entry into the plant cells. Therefore, our results show potential control of pathways by the QTLs in MLG-J, -G and -I that resist further infection of healthy plant tissues by the pathogen by fortification of cells. On the contrary, the QTL locus on MLG-D1a is possibly related to the metabolic mechanism resulting in apoptosis of already infected cells. Finally, unique to MLG-J are the carotenoid and IAA biosynthesis pathways, while the glucosinolate breakdown pathway is unique for MLG-G. The IAA is an important auxin molecule for tissue growth such as the cell mentioned above, while glucosinolate is a natural pesticide and plant defense chemical (Bones and Rossiter 1996).

For the 8 cultivar data, we observed perturbation of a large number of sugar metabolism pathways in the Ox20-8, PI291237 and Sloan cultivars. These cultivars have been shown to be more susceptible to the *P.sojae* pathogen, when compared to Athow, Conrad and General (Dorrance A., unpublished). Results from different studies (reviewed in Berger et al. 2007), show increase in the sugar metabolism during infection of a variety of plant species, compensating for the uptake of plant nutrients from the pathogen. Therefore, the perturbation of the sugar metabolic pathways in our susceptible cultivars, might similarly make up for the nutrient loss caused by *P. sojae* as infection progresses. Additionally,

sugars are not only nutrients but also signals for the pathogen (Berger et al. 2007). Therefore, reduction in sugar levels may propagate the defense response in the the resistant cultivars, and therefore play an important role in specific cultivar–pathogen interactions.

We then tried to isolate details of the pathway perturbation at the gene level, by visualizing the gene expression changes during infection with the *P. sojae* pathogen in the different cultivars, or with different RIL genotypes in the MLGs. We used as example the chorismate biosynthesis pathway, which has been shown to modulate the carbon flux from primary to secondary metabolism (Weaver and Herrmann 1997). In addition, this pathway has been shown to be the first step in the production of a number of specialized metabolites, such as quinones, phenylpropanoids and indoles. We found the genes for 3-dehydroquinate dehydratase I (Enzyme Commission number, EC: 4.2.1.10) and shikimate 5-dehydrogenase (EC: 1.1.1.25), up-regulated during infection in all cultivars except in Williams, and in RILs with PI407162 genotype in MLG-D1a. In contrast to that, for MLG-I these genes were found to be up-regulated in individuals with V71-370 genotype, and based on the Pathway Perturbation Algorithm output in this case the chorismate pathway was found not perturbed. No changes in the expression of the 3-dehydroquinate dehydratase I or shikimate 5-dehydrogenase genes where observed in MLG-J or cultivar Williams, and similarly with MLG-I the chorismate pathway was also not perturbed. Therefore, up-regulation of these two genes during infection, might be a requirement for pathway perturbation during infection. Furthermore, since they were up-regulated in RIL individuals with PI407162 genotype under the QTL in MLG-D1a, a genetic element specific to the PI407162 germplasm, might reside in this specific locus and controlling expression of these genes. The up- regulation of the 3-dehydroquinate dehydratase I and shikimate 5-dehydrogenase in individuals with V71-370 genotype in MLG-I, suggests a counter-balancing control element in this QTL locus, originating from the germplasm of the second parent of the population. According to the data from AraCyc, the enzymic activities of 3-dehydroquinate dehydratase (EC: 4.2.1.10) and shikimate 5-dehydrogenase (EC: 1.1.1.25) are performed by the DHQase/SORase bi-functional enzyme (AraCyc gene id: AT3G06350).

As it has been reported in the literature, in plants the regulation of the chorismate biosynthesis pathway seems to take place exclusively on the genetic level, whereas microorganisms regulate the pathway via chemical feedback inhibition (Herrmann and Weaver 1999). Some of the enzymes for this pathway

have been characterized in plants, and they have been shown to have greater amino acid identity with prokaryote homologues than with yeast and fungal homologues (Herrmann and Weaver 1999). The 3-deoxy-7-phosphoheptulonate synthase (EC:2.5.1.54) which is the first enzyme in the chorismate biosynthesis pathway, based on AraCyc has two paralogues in Arabidopsis name DHS1 and DHS2. For DHS1 it has been reported that it is chloroplastic and requires the ferredoxin/thioredoxin (Fd/TRX) system of the chloroplast to be functional (Entus et al. 2002), while also it is activated during *Pseudomonas* infection (Keith et al 1991). On the other hand, DHS2 is cytocolic and was found downregulated during *Pseudomonas* infection (Keith et al 1991). It is therefore possible since root tissues were used in our study, that during the microarray assay transcripts from the soybean gene corresponding to DHS2 from *Arabidopsis* were measured. If this is the case indeed, based on our results (**Fig.4**), it is similarly down-regulated during infection as in *Arabidopsis*. For the 3-dehydroquinate synthase (EC 4.2.3.4) which is the second enzyme in the pathway (**Fig.4**), it has been shown to be sensitive to plant pathogen elicitors in tomato (Biscoff et al. 2001). The third pathway enzyme, the DHQase/SORase bi-functional enzyme for which details have been presented in the previous paragraph, has been experimentally characterized in tobacco (Bonner et al. 1994). Furthermore, it has been shown in pepper, that this enzyme is induced after wounding (Diaz et al. 1998). The DHQase/SORase enzyme has also enough genetic variability, that it has been used to determine the validity or extent of outcrossing in *Vicia faba*, to evaluate genetic variation within a population of rapeseed, and to derive evolutionary relationships between cultivars, ecotypes, and species in *Phaseolus cossineus* and cotton (review in Herrmann and Weaver 1999). Finally, since the chorimate biosynthesis pathway modulates the transition from primary to secondary metabolism (Weaver and Herrmann 1997), and most of its enzymes respond to pathogen elicitors and wounding, we can speculated that it has a key role for sensing and regulating the defense response.

3.5. Materials and Methods

Gene homologies between *A. thaliana* **and** *G. max.* An initial BLAST nucleotide search was performed using the *G. max* unigenes corresponding to probes of the Affymetrix Soybean Genome Array (Affymetrix, Santa Clara CA), compared against the predicted gene sequences (cDNA) from the soybean genome sequence (http://www.phytozome.org). A second BLAST search was used to compare the amino acid sequences of the predicted soybean genes to those of *A. thaliana* (TAIR-Aracyc, http://www.tair.org). The results from both searches were uploaded to the tables of a POSTgres relational database (http://www.postgres.org), and the BLAST hits were filtered at different E-value and similarity thresholds during the pathway integration.

Pathway transfer from *A.thaliana* **to** *G.max.* The A.thaliana pathway database hosted at TAIR-Aracyc (http://www.tair.org) is a highly curated resource, which was generated based on the annotations of the sequenced genome and the Pathway Tools software (Karp et al. 2002). Initially, the pathways where downloaded from Aracyc (Mueler et al. 2003) in the BioPAX-RDF/OWL format (http://www.biopax.org). Using the Sesame RDF database (http://www.openrdf.org) queries where formulated, in order to extract the pathways in a pairwise interaction format (see SI Fig. for more information). Each pair represents the interaction of two genes catalyzing adjacent biochemical reactions in a pathway, and for reactions catalyzed by more than one genes all possible pairs where identified. The pairwise interactions where also stored in the same POSTgres database with the BLAST results. Next, a query was used that joined the database tables with the BLAST result from the soybean unigenes versus the predicted genes, to the result of soybean peptides from the predicted genes versus *Aradidopsis* proteins, and finally to the table with the pairwise pathway interactions. Entries from the database table containing the unigenes were filtered out if their E-value was greater than 10e-200 and no less than 90% of the unigene sequence was included in the each BLAST High Scoring segment Pair (HSP). Similarly, results from the table containing the BLAST output of the soybean peptides against the Arabidopsis proteins, were kept only if they had E-value less than 10e-20 and at 80% of the soybean peptides length was included in the HSP. In this manner we identified the A.thaliana genes which the soybean unigenes corresponded to, and we extracted a new table that contained part of the

A.thaliana pathway network transferred in soybean based on gene homologies. Additionally, this table stores the pathways using unigene identifiers from the Affymetrix Soybean Genome Array, therefore allowing us to directly superimpose the gene expression data on the transferred metabolic pathways.

Plant Materials and RNA extractions. Seven cultivated soybean lines Athow, Conrad, General, V71-370, Ox20-8 and Sloan, plant introduction PI291237, *G.soja* line PI407162, and a Recombinant Inbred Lines (RILs) soybean population were used in this study. The population consisted of 297 RILs and was developed from and interspecific cross between the *Glycine max* line (V71-370, Group V) and the *Glycine soja* plant introduction (PI407162, Group IV) using a modified single-seed decent method as described in (Maroof et al.). All plants were grown in a growth chamber with day and night temperatures settings of 27°C and 21°C, relative humidity averaging 75 to 90%, and a 14 h light : 10 h dark cycle (Zhou et al. 2008). Seven day-old seedlings for RNA extraction were inoculated or mock-inoculated with *P. sojae*, and wounded at 2 cm below the beginning of the root zone by scraping the root epidermis with a sterile scalpel for all eight lines and RILs in two separate experiments. Samples of root tissues from 10 plants from each individual were collected at 3 and 5 days post inoculation or mock-inoculation for the 8 soybean lines, and 5 days post mock-inoculation for the RILs. Then they were immediately frozen in liquid nitrogen and stored at -80ºC prior to RNA extraction. The QIAGEN RNeasy® Plant Mini Kit (Qiagen Corporation, Valencia, CA) was used throughout for total RNA isolation from the root tissue sections. The manufacturer-provided protocol was followed with minor modifications in order to obtain a sufficient amount of high quality RNA. The quality of total RNA was checked in an Agilent 2100 Bioanalyser. RNA samples from the two inoculation replications were pooled in equal amounts.

Microarray Data Generation and Analysis. Microarray hybridization procedures were performed at the Core Laboratory Facility of Virginia Bioinformatics Institute (Blacksburg, VA) following the standard eukaryotic gene expression assay protocols described in the Affymetrix GeneChip® Expression Analysis Technical Manual. In brief, the One-Cycle Target Labeling and Control Reagents (Affymetrix®) and 1 ug of total RNA were used to generate biotin-labeled cRNA. Twenty micrograms of labeled cRNA was fragmented in Fragmentation Buffer and then hybridized to the Affymetric Soybean Genome Array®, which assays simultaneously 37,593 soybean transcripts. Hybridization was

performed at 45°C for 16 h in an Affymetrix® hybridization oven (model 640), and then microarrays were washed and stained with streptavidin-phycoerythrin using the fluidics protocol EukGE-WS2v5-450 in the Affymetrix® GeneChip® Fluidics Station 450. Stained chips were scanned with an Affymetrix GCS3000 7G Scanner. The Affymetrix GeneChip® Operating Software (GCOS, v1.4.0.036) was used to provide instrument control, first-level data analysis, and data management for the entire microrarray assay. Quality control of acquired gene expression data was performed using box-plots for each chip and ratio-versus-intensity plots for pairs of chips and by computing 3'/5' ratios of ▪-Actin to check for RNA degradation. Low-level analysis of raw expression data was performed by the following steps. For the SFP discovery in the eight soybean lines only, the Affymetrix Microarray Suite version 5 (MAS5) algorithm was used for retaining genes that had a Present Call (Affymetrix GeneChip® Expression Analysis Technical Manual). Implementation of MAS5 was from the Bioconductor package *affy* (online ref: http://bioconductor.org/packages/2.0/bioc/html/affy.html). A second step for both the RIL and soybean line expression data was background correction with the model-based procedure followed by quantile-normalization for probe-level data.

Pathway perturbation analysis in *G. max.* In order to study pathway perturbation of our pathways under different biological treatments, we used the Pathway Perturbation Algorithm by Rivera et al. Given a pathway and gene expression measurements for a treatment and a control phenotype, this method computes the sub-pathway that is most perturbed between the two phenotypes. In the first step for each gene g, the p-value (p_g) of its differential expression in the two phenotypes is computed. For obtaining the p-value, a two-sided t-test is performed under the null hypothesis that the distributions of expression values in the samples of the two phenotypes have identical means (different variances are allowed). Then the p-value is converted into a z-score using the formula $z_g = N^{-1}(1 - p_g)$, where N^{-1} is the inverse of the normal cumulative distribution function. No cutoff is imposed on this z-score, and all genes are included in a subsequent analysis for finding the most perturbed sub-pathways. In order to compute an aggregate Z(P) value for the perturbation of each pathway, by aggregating the Z-scores of the individual genes, the algorithm uses the weighted Liptak-Stouffer Z-score statistic. This statistic is the ratio of the sum of weighted z-scores for each gene, using as weight (W_g) the number of interactions (edges) each gene has with others in the pathway:

$$Z(P) = \frac{\sum W_g Z_g}{\sqrt{\sum W_g^2}}$$

With denominator set as square root of the sum of squared weights, the statistic follows a normal distribution with mean 0 and standard deviation 1. Additionally, the use of weights incorporates in the final statistic besides the perturbation of each gene, the connectivity and therefore its contribution to the metabolic network. This algorithm does not require a significant z-score across the whole pathway, and therefore not every gene needs to be differentially expressed in order to declare the pathway perturbed. This is achieved by calculating the most perturbed sub-pathway, by following a heuristic approach based on simulated annealing in order to identify the sub-pathway that maximizes the Liptak-Stouffer Z-score.

We initially used as input to the algorithm expression data for all 8 soybean lines from 3 or 5 days post inoculation and mock-inoculation. Following this, using as criterion the genotype of each RIL individual in the genetic markers found under the 4 major disease-resistance Quantitative Trait Loci (QTL) on chromosomes J, G, I and D1A of soybean, and separated the RIL population in two groups for each QTL. The different genotypes originate from either the V71-370 or PI407162 parent of the population. Expression data (5 days post mock-inoculation) from the two groups of RILs formed for each QTL, were provided as input to the algorithm in lieu of the mock versus inoculated groups used for the 8 soybean lines. We then selected the z-scores for the contrasts of mock versus inoculated in 3 or 5 days expression data for the 8 soybean lines, and also the z-scores for the two groups of the RILs for each chromosome, from the algorithm output. The z-score were used to colorize the transferred metabolic network of *A.thaliana* in soybean using Cytoscape (http://www.cytoscape.org). We used a transition of spectrum colors (violet to red) to represent z-score continuous increments from 0 to 22, and white to represent all z-scores greater than 22 (see SI, Fig. for colorization of the z-score scale).

3.6. Rerefences

Benjamini , Hochberg Y. (1995) "Controlling the false discovery rate: a practical and powerful approach to multiple testing" Journal of the Royal Statistical Society **57**: 289–300.

Berger S, Sinha AK, Roitsch T. (2007) "Plant physiology meets phytopathology: plant primary metabolism and plant pathogen interactions" J of Exp Botany **58** :4019-4026.

Bischoff M, Schaller A, Bieri B, Kessler F, Amrhein N, Schmid J. (2001) "Molecular Characterization of Tomato 3-Dehydroquinate Dehydratase Shikimate: NADP Oxidoreductase" Plant Phys. **125** :1891–1900.

Bones AM, Rossiter JT. (1996) "The myrosinase-glucosinolate system - an innate defense system in plants" Physiologia plantarum **97**: 194-208.

Bonner CA, Jensen RA. (1994) "Cloning of cDNA encoding the bifunctional dehydroquinase shikimate dehydrogenase of aromatic amino acid biosynthesis in *Nicotiana tabacum*" Biochem J. **302**:11-4.

Brilli M, Fani R, Liò P. (2008) "Current trends in the bioinformatic sequence analysis of metabolic pathways in prokaryotes" Brief Bioinform. **9** :34-45.

Caspi R, Foerster H, Fulcher CA, Kaipa ., Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer A, Tissier C, Walk TC, Zhang P, Karp PD. (2008) "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases" Nucleic Acids Research **36** :623-631.

Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. (2008) "Identifying functional modules in protein-protein interaction networks: an integrated exact approach" Bioinformatics **24** :223–231.

Diaz J, Merino F. (1998) Wound-induced shikimate dehydrogenase and peroxidase related to lignification in pepper (*Capsicum annum* L.) J. Plant Physiol **152**:51–57

Entus R, Poling M, Herrmann KM. (2002) "Redox regulation of Arabidopsis 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase" Plant Physiol. **129**:1866-1871.

Graca J, Pereira H. (2000). "Suberin structure in potato periderm: glycerol, long-chain monomers, and glyceryl and feruloyl dimers." J Agric Food Chem **48**:5476-5483.

Heath, MC. (2000). "Hypersensitive response-related death." Plant Molecular Biology **44**: 321–34.

Herrmann KM, Weaver LM. (1999). "The Shikimate Pathway." Annu Rev Plant Physiol Plant Mol Biol **50**: 473-503

Hou DX (2003). "Potential mechanisms of cancer chemoprevention by anthocyanins" Curr. Mol. Med. 3: 149–159.

Hu P, Meng Y, Wise RP. (2009) "Functional contribution of chorismate synthase, anthranilate synthase, and chorismate Mutase to penetration resistance in barley–powdery mildew interactions" Molec Plant Microbe Interact **22** :311-320.

Ideker T, Ozier O, Schwikowski B, Siegel AF. (2002) "Discovering regulatory and signalling circuits in molecular interaction networks" Bioinformatics **18**: 233-240.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. (2000) "The large-scale organization of metabolic networks" Nature. **407** : 651-654.

Keith B, Dong XN, Ausubel FM, Fin GR (1991) "Differential induction of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase genes in Arabidopsis thaliana by wounding and pathogenic attack" Proc Natl Acad Sci USA **88**: 8821–8825.

Klee HJ, Muskopf YM, Gasser CS. (1987) "Cloning of an *Arabidopsis thaliana* gene encoding 5-enolpyruvylshikimate-3-phosphate synthase: sequence analysis and manipulation to obtain glyphosate-tolerant plants" Mol Gen Genet. **210** :437-442.

Kasai K, Kanno T, Akita M, Ikejiri-Kanno Y, Wakasa K, Tozawa Y. (2005) "Identification of three shikimate kinase genes in rice: characterization of their differential expression during panicle development and of the enzymatic activities of the encoded proteins" Planta. **222** : 438-447.

Karp PD, Paley S, Romero P. (2002) "The Pathway Tools software" Bioinformatics **18** : 225-232

Keith B, Dong XN, Ausubel FM, Fink GR. (1991) "Differential induction of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase genes in Arabidopsis thaliana by wounding and pathogenic attack" Proc Natl Acad Sci U S A. **88** :8821-8825.

Keller Y, Bouvier F, d'Harlingue A, Camara B. (2008) "Metabolic compartmentation of plastid prenyllipid biosynthesis--evidence for the involvement of a multifunctional geranylgeranyl reductase" Eur J Biochem **251**: 413-417.

Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. (2004) "PathBLAST: a tool for alignment of protein interaction networks" Nucleic Acids Res. **32**: 83-88.

Krampis K., Tucker D., Maroof S., Tyler B. M. and Tripathy S. (2009) "High-density haplotyping with inSilico discovered polymorphisms in Soybean" (in preparation for submission)

Macheroux P, Schmid J, Amrhein N, Schaller A. (1999) "A unique reaction in a common pathway: mechanism and function of chorismate synthase in the shikimate pathway" Planta. **207** :325-334.

Maraziotis IA, Dimitrakopoulou K, Bezerianos A. (2007) "Growing functional modules from a seed protein via integration of protein interaction and gene expression data" BMC Bioinformatics **8**: 408.

Mueller LA, Zhang P, Rhee SY. (2003) "AraCyc: a biochemical pathway database for *Arabidopsis*" Plant Physiology **132** :453-460.

Navathe SB Elmasri R (2002) "Fundamentals of Database Systems" Addison-Wesley, Longman, NY

Oehm S, Gilbert D, Tauch A, Stoye J, Goesmann A. (2008) "Comparative Pathway Analyzer--a web server for comparative analysis, clustering and visualization of metabolic networks in multiple organisms" Nucleic Acids Res. **36**: 433-7.

Rivera CG, Tyler BM, Murali TM (2009) "Sensitive detection of pathway perturbations in cancers" (in preparation for submission)

Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, Paulovich A, Pomeroy S, Golub T, Lander E, Mesirov J. (2005) "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles" Proc Natl Acad Sci USA **102**: 15545-15550.

Tucker DM, Maroof MAS, Mideros S, Skoneczka JA, Nabati DA, Buss GR, Hoeschele I, Tyler BM, St. Martin SK, Dorrance AE. (2009) "Mapping Quantitative Trait Loci for Partial Resistance to *Phytophthora sojae* in a Soybean Inter-specific Cross" (in preparation for submission)

Ulitsky I, Shamir R. (2007) "Identification of functional modules using network topology and high-throughput data" BMC Systems Biology **1**: 8.

Weaver LM, Herrmann KM. (1997) "Dynamics of the shikimate pathway in plants" Trends in Plant Sci. 2: 346-351.

Wildermuth MC, Dewdney J, Wu G, Ausubel FM (2001) "Isochorismate synthase is required to synthesize salicylic acid for plant defence" Nature **414**: 562-565.

Zhou L, Mideros SX, Bao L, Hanlon R, Arredondo FD, Tripathy S, Krampis K, Jerauld A, Evans C, St. Martin SK, Maroof MAS, Hoeschele I, Dorrance AE, Tyler BM. (2009) "Infection and genotype remodel of the entire soybean genome" BMC Genomics 10:49-55.

3.7. Supplementary Information

S1 Names of pathways found in soybean and are drawn the various sections of **Fig.1**

Pathway	**Fig. section**
sucrose degradation to ethanol and lactate (anaerobic)	**a-b**
photosynthesis	**a-b**
ascorbate glutathione cycle	**a-b**
superdrogenase, TCA, and glyoxylate bypass	**a-b**
(deoxy)ribose phosphate degradation	**a-b**
branched-chain alpha-keto acid dehydrogenase complex	**a-b**
superpathway of glyperpathway of sucrose degradation to pyruvate	**a-b**
superpathway of oxidative and non-oxidative branches of pentose phosphate pathway	**a-b**
gluconeogenesis	**a-b**
glutathione redox reactions	**a-b**
superpathway of ribose and deoxyribose phosphate degradation	**a-b**
superpathway of starch degradation to pyruvate	**a-b**
superpathway of radation to pyruvate	**a-b**
TCA cycle	**a-b**
valine biosynthesis	**a-b**
superpathway of sucrose degradation to pyruvate	**a-b**
superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass	**a-b**
superpathway of glyoxylate cycle	**a-b**
alpha-ketoglutarate dehydrogenase complex	**a-b**
superpathway of ascorbate biosynthesis	**c**
mannose degradation	**c**
ascorbate glutathione cycle	**c**
GDP-L-fucose biosynthesis I (fuvate	**c**
galactose degradation I	**c**
UDP-sugars interconversion	**c**
superpathway of sucrose degradation to pyruvate	**c**
homogalacturonan degradation	**c**
GDP-L-fucosesucrose degradation to pyruvate	**c**
inositol oxidation pathway	**c**
trehalose degradation	**c**
starch biosynthesis	**c**
sucrose degradation to ethanol and lactate (anaerobic)	**c**
dTDP-L-rhamnose biosynthesis	**c**
GDP-L-fucose biosynthesis I (from GDP-D-mannose)	**c**
UDP-glucose biosynthesis (from glucose -phosphate)	**c**
sucrose biosynthesis	**c**
quercetin glucoside biosynthesis	**e**
flavonoid biosynthesis	**e**
suberin biosucocyanidin biosynthesis	**e**

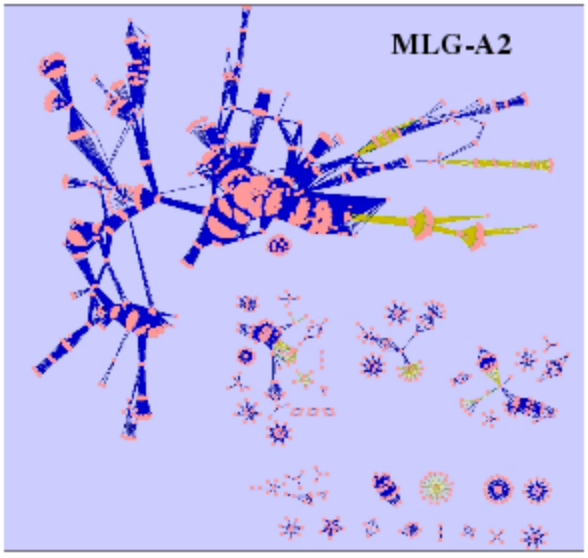
Pathway	Fig. section
GA۱۲ biosynthesis	**e**
flavonol biosynynthesis	**e**
sinapate ester biosynthesis	**e**
S-methylmethionine cycle	**e**
suberin biosynthesis	**e**
cellulose biosynthesis	**e**
superpathway of lysine, threonine, and methionine biosynthesis	**e**
quercetin glucoside biosynthesis	**e**
anthocyanin biosynthesis (pelargondin and leucocyanidin biosynthesis	**e**
superpathway of aspartate and asparagine biosynthesis	**e**
IAA biosynthesis I	**e**
superpathway of GA۱۲ biosynthesis	**e**
superpathway of GA۱۲ biosynthesthesis	**e**
flavonol biosynthesis	**e**
free phenylpropanoid acid biosynthesis	**e**
trehalose biosynthesis	**e**
flavonohenylpropanoid biosynthesis	**e**
fatty acid beta-oxidation I (saturated)	**e**
gibberellin inactivation	**e**
glucosinolate biosynthesis from tryptophan	**e**
mannitol degradation	**e**
superpathwaperpathway of GA۱۲ biosynthesis	**e**
flavonol biosynthhesis	**e**
carotenoid biosynthesis	**e**
brassinosteroid biosynthesis II	**e**
ehylene biosynthesis from methionine	**e**
removal of superoxide radicals	**e**
oxidative ethanol degradation	**e**
leucine biosynthesis	**e**
phenylpropanoid biosyntesis	**e**
methionine salvage pathway	**e**
leucopelargonidin and leucocyanidin biosynthesis	**e**
phenylpropanoid biosynthesis	**e**
superpathway of threonine biosynthesis	**e**
anthocyanin biosynthesis (pelargonidin ۳-O-glucoside, cyanidin ۳-O-glucoside)	**e**
superpathway of gibberellin biosynthsynthesis	**e**
valine degradation	**f**
very long chain fatty acid biosynthesis	**f**
brassinosteroid biosynthesis II	**f**

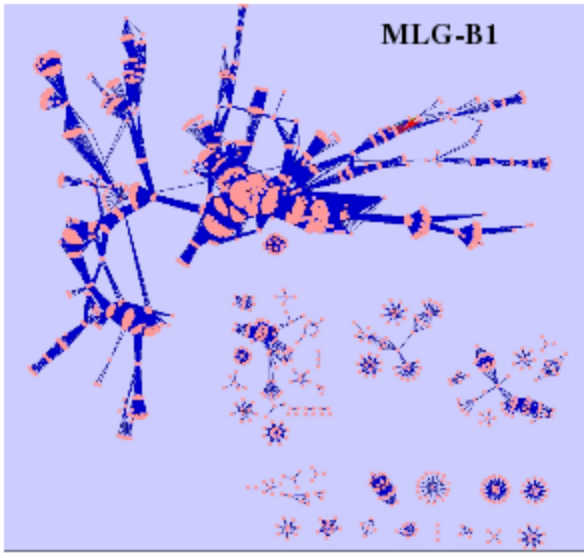
Pathway	Fig. 1 section
mevalonate pathway	**f**
lysine degradation II	**f**
superpathway of lysine, threonine, and methionine biosynthesis	**f**
superpathway of threonine biosynthesis	**f**
carotenoid biosynthesis	**f**
tyrosine biosynthesis II	**f**
asparagine biosynthesis I	**f**
folate transformations	**f**
isoleucine degradation	**f**
homomethionine biosynthesis	**f**
superpathway of aspartate and asparagine biosynthesis	**f**
leucine degradation	**f**
leucine biosynthesis	**f**
aldoxime degradation	**f**
trans-zeatin biosynthesis	**f**
sterol biosynthesis	**f**
SAM cycle	**f**
side chain elongation cycle aliphatic glucosinolates (Arabidopsis)	**f**
cysteine biosynthesis	**f**
phenylalanine biosynthesis	**f**
fatty acid beta-oxidation I (saturated)	**f**
folate transformations	**g**
threonine biosynthesis from homoserine	**g**
superpathway of lysine, threonine, and methionine biosynthesis	**g**
SAM cycle	**g**
S-methylmethionine cycle	**g**
glycine biosynthesis	**g**
superpathway of threonine biosynthesis	**g**
photorespiration	**g**
tetrahydrofolate biosynthesis	**g**
superpathway of phenylalanine, tyrosine and tryptophan biosynthesis	**g**
folate polyglutamylation II	**g**
formylTHF biosynthesis	**g**
folate polyglutamylation I	**g**
phenylalanine biosynthesis	**g**
cellulose biosynthesis	**h**
arginine biosynthesis II (acetyl cycle)	**i**
superpathway of lysine, threonine, and methionine biosynthesis	**i**
citrulline biosynthesis	**i**

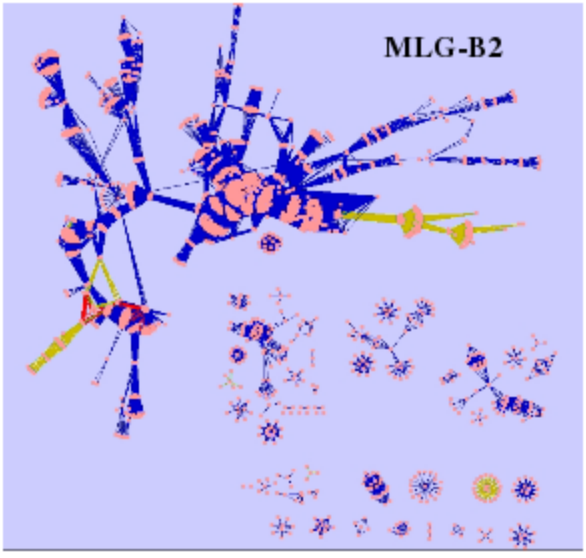
Pathway	Fig. 1 section
histidine biosynthesis	**i**
glutamate degradation II	**i**
tryptophan biosynthesis	**i**
ascorbate biosynthesis	**i**
gamma-glutamyl cycle	**i**
ureide degradation	**i**
urea cycle	**i**
ammonia assimilation cycle	**i**
proline biosynthesis II (from arginine)	**i**
glutathione degradation	**i**
serine biosynthesis	**i**
superpathway of polyamine biosynthesis	**i**
biotin biosynthesis	**i**
beta-alanine biosynthesis I	**i**
lysine degradation II	**i**
UDP-N-acetyl-D-glucosamine biosynthesis	**i**
arginine degradation II	**i**
thiamine biosynthesis	**i**
pyrimidine salvage pathway	**i**
ureide biosynthesis	**i**
pyridine nucleotide cycling (plants)	**j**
superpathway of histidine, purine and pyrimidine biosynthesis	**j**
purine nucleotide metabolism (phosphotransfer and nucleotide modification)	**j**
de novo biosynthesis of pyrimidine deoxyribonucleotides	**j**
pyrimidine nucleotide metabolism (phosphotransfer and nucleotide modification)	**j**
de novo biosynthesis of pyrimidine ribonucleotides	**j**
de novo biosynthesis of purine nucleotides	**j**
salvage pathways of purine nucleosides	**j**
cholesterol biosynthesis	**k**
sterol biosynthesis	**k**
triacylglycerol biosynthesis	**k**
phospholipid biosynthesis	**k**
phospholipases	**k**
superpathway of fatty acid biosynthesis	**k**
dolichyl-diphosphooligosaccharide biosynthesis	**k**
glycolipid biosynthesis	**k**
superpathway of choline biosynthesis	**k**
superpathway of lipid-dependent phytate biosynthesis	**k**

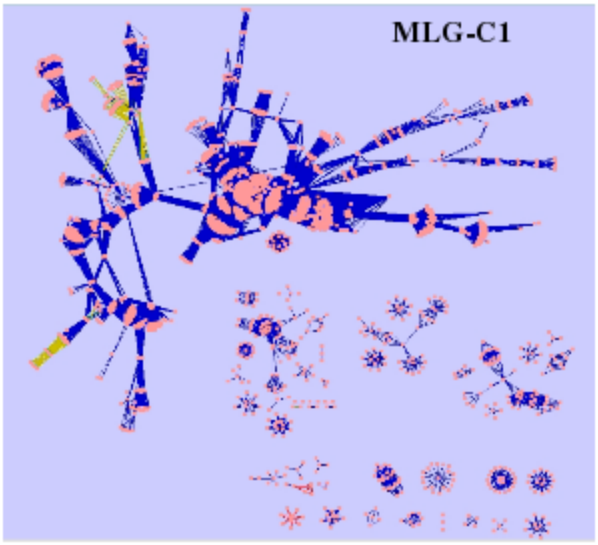
Pathway	Fig. 1 section
superpathway of pantothenate and coenzymeA biosynthesis	**remaining**
stachyose biosynthesis	**remaining**
methylglyoxal degradation	**remaining**
heme biosynthesis	**remaining**
aerobic respiration	**remaining**
abscisic acid biosynthesis	**remaining**
aerobic respiration -- al	**remaining**
monoterpene biosynthesis	**remaining**
glucosinolate breakdown	**remaining**
aerobic respiration -- alternative oxidase pathway	**remaining**
phytyl-PP biosynthesis	**remaining**
carbon tetrachloride degradation	**remaining**
polyisoprenoid biosynthesis	**remaining**
photosynthesis, light reaction	**remaining**
chlorophyllide a biosynthesis	**remaining**
glycine degradation	**remaining**
trans,trans-farnesyl diphosphate biosynthesis	**remaining**
molybdenum cofactor biosynthesis	**remaining**
methylerythritol phosphate pathway	**remaining**
chlorophyll a degradation	**remaining**
chlorophyll cycle	**remaining**
geranylgeranyldiphosphate biosynthesis II (plastidic)	**remaining**

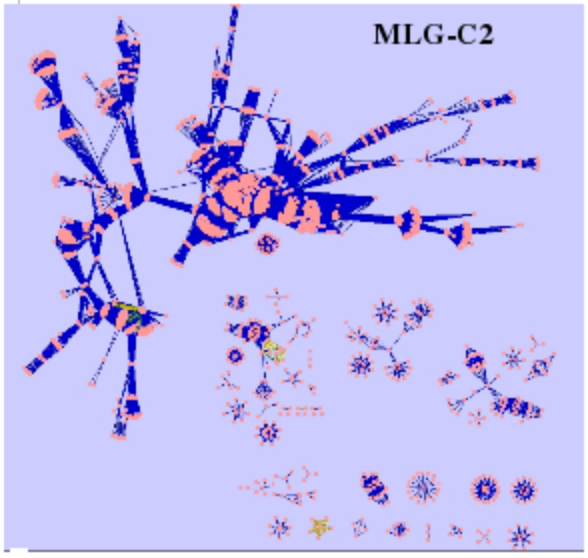
S2 Visualization of pathway perturbation in the14 soybean chromosomes except MLG-J, -G, -I, D1a and MLG-A1, -N. Loci in each of these chromosomes were chose in random as negative control.

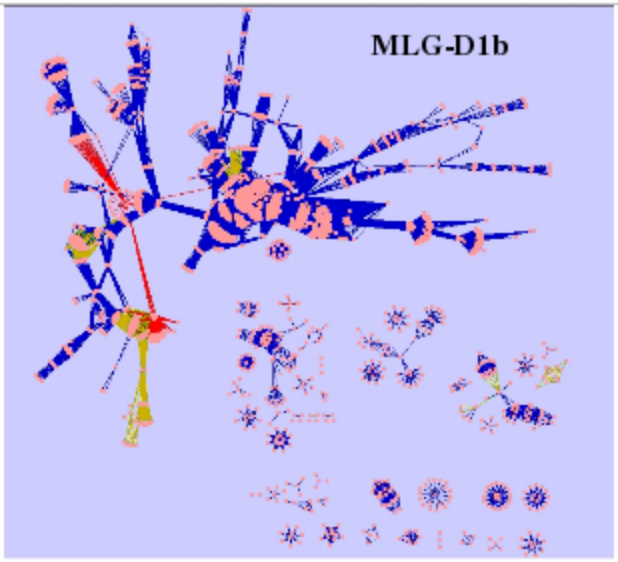


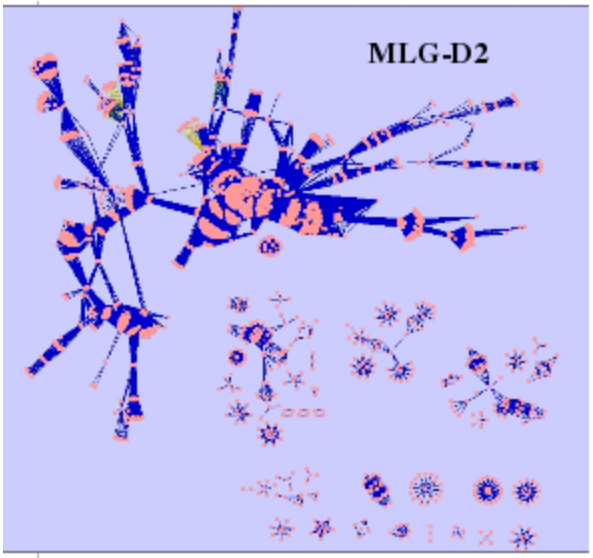


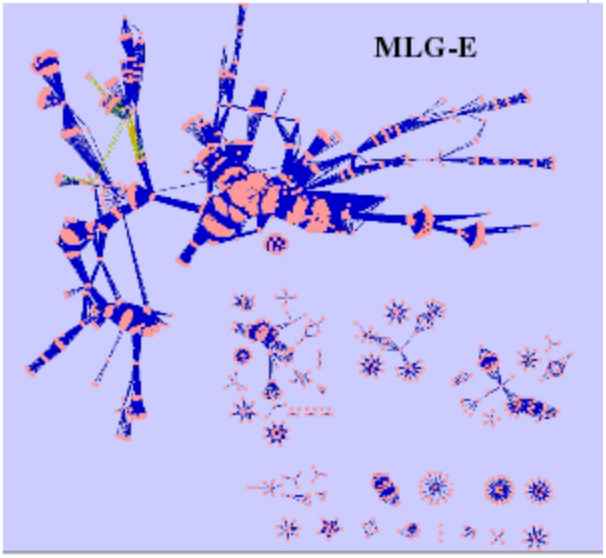


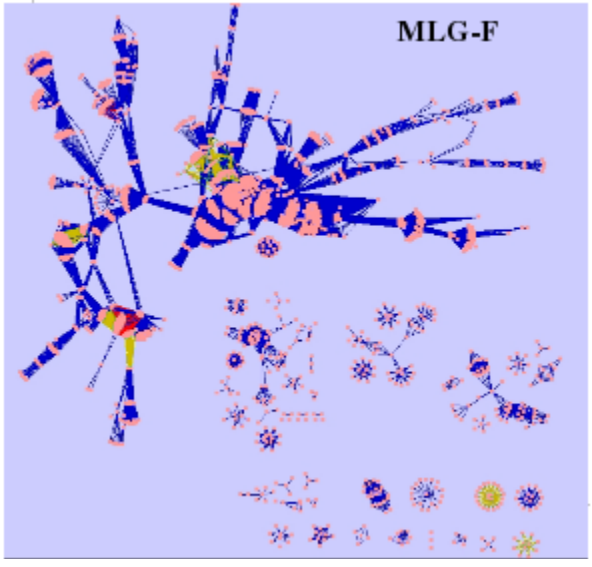


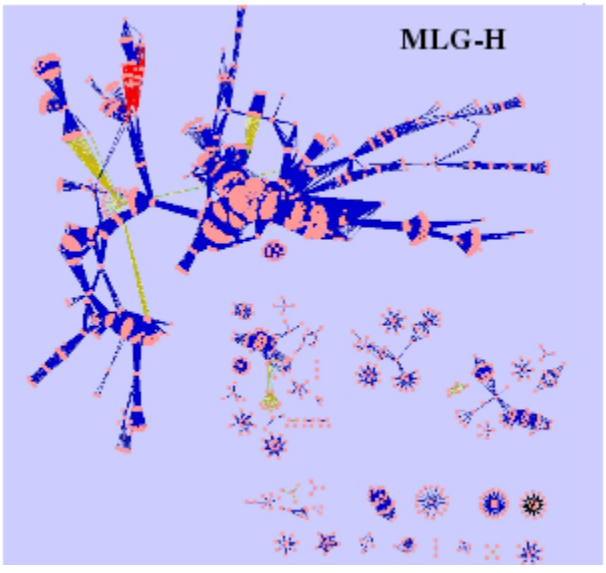

MLG-C2

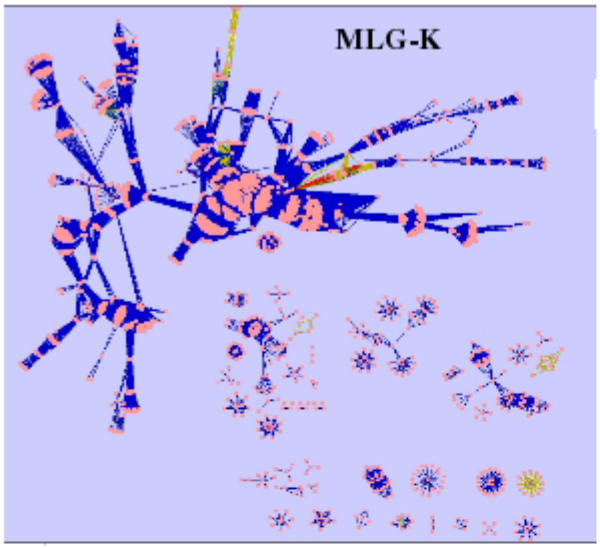

MLG-D1b

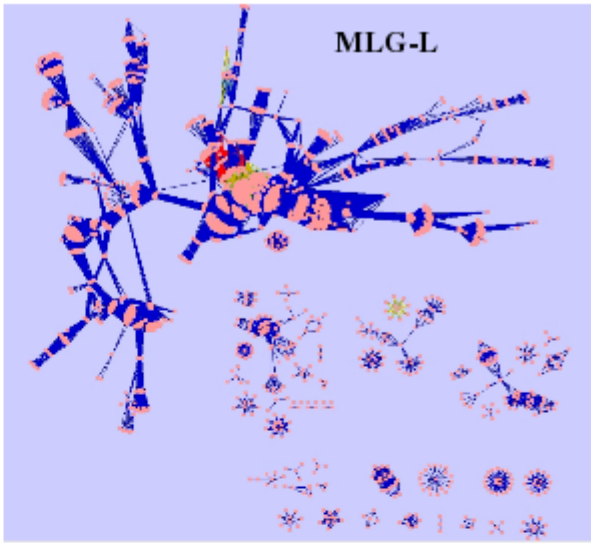

MLG-D2

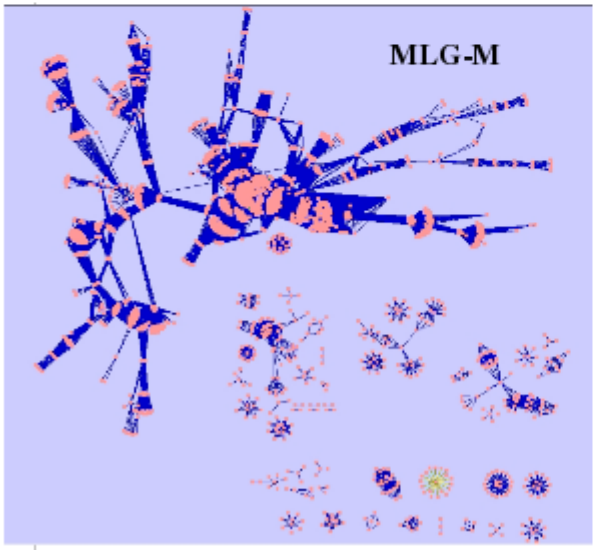

MLG-E

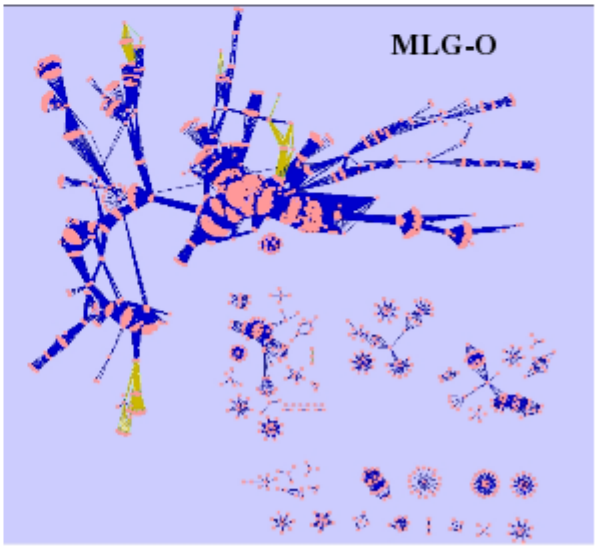

MLG-F


MLG-H


MLG-K


MLG-L


MLG-M


MLG-O

4. Conclusion and summary

4.1. Advantages of the new SFP genetic markers for soybean

In the current study, a large number of SFPs were successfully mapped to the twenty soybean chromosomes, providing a large set of new genetic markers in addition to the collection on the current soybean maps. These new class of marker identified for the first time in soybean here, offers a great genetic resource for this crop species where until recently only small portions of the genome were sequenced. In the SFP genetic map, areas containing large gaps in the soybean public map and considered marker poor, were successfully filled. This suggests the robustness of this new type of marker to fill gene poor or areas of low recombination, such as those present around centromeres. In addition, as it has been demonstrated by various studies using expression data for identifying polymorphisms, the SFP approach is an efficient way to screen for a large number of potential polymorphisms in plant species with complex, multiploid genomes. The use of microarray probes for genetic marker screening has the advantage of being able to detect polymorphisms even in the case of single base or sequence repeat changes. This creates problems with traditional marker screening when detecting RFLP restriction sites or performing SSR amplification. Furthermore, microarrays can screen thousands of genetic loci simultaneously, providing great potential for polymorphism discovery even in domesticated cultivars with low genetic diversity. This was verified through our study, where eight domesticated soybean cultivars were successfully assayed with SFP markers and successfully separated in a phylogenetic tree.

For the SFP markers no additional cost exists other than the implementation of the appropriate algorithms, in contrast to the labor and cost-intensive laboratory screening for traditional genetic markers. Our new soybean markers were identified by using available gene expression data, therefore leveraging the value of existing data. The newly constructed SFP genetic map provides additional information on genome organization in soybean, as it represents actual gene coding sequences rather than non-coding regions as in the case of SSRs. Conseqeuently, SFPs have double value since they are both genetic markers and genes with known sequence and annotations in most cases. Thus, the mapped SFPs can provide breeders with additional markers for mapping qualitative and quantitative traits, and

can point to candidates for the traits under consideration. Our new gene-based markers can be used for initial screening of a large number of polymorphisms, and once SFPs are identified that are associated with traits or genomic regions of interest, they can then be followed in more detail. Based on the sequence of the microarray probes, PCR primers can be designed that will isolate the sequence of a gene containing an SFP polymorphism. Finally, positioning QTLs over an SFP genetic map can present the opportunity for isolating genes that are directly (cis- effect) controlling the trait phenotype. In this manner, genetic maps created using SFPs are an invaluable addition to the toolbox of soybean geneticists and breeders, working towards gene isolation for disease resistance and other economically important crop traits.

4.2. Pathway perturbation and mechanisms of plant pathogen resistance

We have performed pathway perturbation analysis in soybean in a variety of genetic backgrounds, in order to uncover the underlying metabolic mechanisms involved in pathogen resistance. By using the abudant data for the *Arabidopsis* metabolic pathways we first identified pathways potentially present in soybean, based on the sequence homology between the two species. We then used expression data from genetic lines of soybean with different levels of resistance to the pathogen *P. sojae*, and also a recombinant inbred line (RIL) population created from a cross of *G. max* with *G. sojae*. Using the Pathway Perturbation Algorithm, we identified pathways that significantly alter their activity in response to infection by the pathogen, and therefore contribute to pathogen resistance by provoking cellular modifications involved in plant defense. Since the soybean lines and individuals of population, have different genetic backgrounds, our results can help pinpoint genetic and biochemical factors contributing to the resistant soybean phenotype.

In all soybean lines used in our study, we found perturbation of secondary metabolism pathways, with the majority of these related to production of molecules for cell wall fortification. This is expected, since the cell wall is a major barrier and the first line of defense against pathogens trying to evade the plant cells. Interesting is the result for the cultivars that are susceptible to *P. sojae*, which were found to have up-regulated sugar metabolism. This is a side effect of the increased presence of the pathogen within the plant tissues in comparison to the resistant cultivars. As the pathogen consumes nutrients

from the plant, sugars need to be metabolised in order to compensate for the loss of nutrients and energy.

Specific differences where also found in regards to perturbed pathways, between the four chromosomal loci used with the expression data from the RIL population created by the cross of PI407162 and V71-370. These loci are within QTL regions of the genome known to contain genes for resistance to *P. sojae.* With our experimental design we tried to uncover the distinct contribution to resistance from the perspective of metabolic changes, of each parental genotype under in the QTL loci. For MLG-D1a in particular, between the two groups of RILs with either PI407162 or V71-370 genetoype in this locus, pathways for plant hormones such as giberellin and anthocyanin where found perturbed. Unique for MLG-G was the glucosinolate pathway, a molecule that can act as pesticide. While pathways metabolizing molecules for the creation of cell walls such as phenylpropanoid biosyntesis were found perturbed in all MLGs, these remain un-altered in MLG-D1a. The carotenoid biosyntesis pathway was perturbed only in MLG-J. These results hint that there is a possibility that each QTL in the four loci used in our study, controls a separate aspect of the metabolic response to pathogen infection, and therefore a different cellular modification that provides resistance. We also studied in more detail the contrast of expression levels between the cultivars or MLGs for genes participating in the chorismate biosynthesis pathway, a key pathway for carbon flux from primary to secondary metabolism. For the 3-dehydroquinate dehydratase I and shikimate 5-dehydrogenase, we observed up-regulation in the resistant cultivars during infection, where also this pathway was perturbed. Concerning the QTL loci only MLG-D1a showed significant perturbation of the chorismate biosythesis pathway, and in this case the two genes where up-regulated in RIL individuals with PI407162 genotype. When no changes in the expression of the genes where observed, or they where up-regulated in indiciduals with V71-370 genotypes (for example in MLG-I), this pathway was not perturbed.

Follow up research can build on the data produced by the present study, where contrasts in the expression of the individual genes participating in the perturbed pathways we identified between the MLGs, can be analyzed in more detail. Such genes that show distinct patterns of expression between soybean cultivars are for example the 3-dehydroquinate dehydratase I and shikimate 5-dehydrogenase. For these detailed studies at the genomic level can be performed. Since the soybean genomic sequence

is available exact possition of the pathway genes on the genome can be identified, and also whether they reside within the QTL region controlling the pathway. This would be a case of cis- or otherwise localized regulation of the pathway. Even if this is not true, by following the annotation of the genome sequence on the QTL locus, information can be gathered about genes that based on their annotated function might perform remote (trans-) regulation of the pathway activity. This would provide a good set of hypotheses for experimental testing, and therefore candidate genes for soybean breeders for developing new soybean lines with increased *P. sojae* resistance.