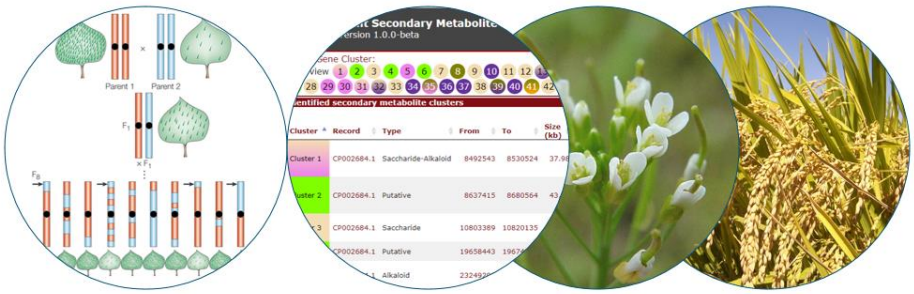
Integrating metabolite and expression quantitative trait loci with biosynthetic gene clusters

MSc thesis report



Written by:

Lotte Witjes[1], BSc.

Supervised by:

dr. Harm Nijveen[1]

dr. Marnix Medema[1]

[1] Department of Bioinformatics, Wageningen University & Research, Radix West, Droevendaalsesteeg 1, 6708 PB Wageningen,

Contents

Abstract

Plant phytochemicals and their synthesis are of increasing interest since these secondary metabolites can be used in a wide variety of applications, they can be applied as medicines, dyes, flavorings, cosmetics and much more. Therefore, knowledge of the biosynthetic pathways of these compounds and their regulation is important for improving yield of certain antibiotics in crops or for improving crop resistance and quality, for example. Recent studies found the tendency of genes coding for enzymes involved in secondary metabolite biosynthesis pathways to physically cluster on the genome. These so-called biosynthetic gene clusters (BGCs) consist of at least three different enzymes executing some consecutive enzymatic reactions in a biosynthetic pathway. Since metabolite levels are considered complex quantitative traits, metabolite quantitative trait analysis (mQTL) can be used to elucidate genomic regions influencing these traits. In a same manner, expression QTL (eQTL) analysis can be used to identify genomic regions that affect expression levels. Combining mQTL with eQTL (metabolite and expression quantitative trait loci) and biosynthetic gene clusters (BGCs) predicted with PlantiSMASH might give more insights in biochemical synthesis and regulation than using any of these sources alone. Annotation of putative BGCs can be done by taking into account the mass spectrometry (MS) data of the (known) metabolites and expression levels and domains of (known) genes. This leads to the possibility to generate hypotheses of genes encoding enzymes that might play a role in the biosynthesis of the secondary metabolite. The latter is important if exploitation of a plant for its secondary metabolites, for earlier mentioned applications, is desired since many of the biosynthetic genes for secondary metabolites are still unknown. Here, strategies based on genomic, location overlap were developed to efficiently integrate mQTL and eQTL with BGCs using a recombinant inbred line population of *Oryza sativa* and a genome-wide association study of *Arabidopsis thaliana*. Genomic overlap was detected between BGCs and xQTLs (either e- or mQTLs), leading to interesting links between BGCs and secondary metabolites as their putative products. Several validation methods were used to strengthen the evidence for the links. Colocation networking was applied, by looking for genomic overlap between mQTL-mQTL and mQTL-eQTL pairs in *A. thaliana*, leading to hypotheses proposing new possible substrates of OMT1 and BGLU6. Therefore, these methods can generate hypotheses for linking BGCs to their products *in silico*, which can be used to design targeted lab experiments for validation.

Introduction

Plants produce a large variety of biochemical compounds, compared to other organisms. These biochemical compounds are involved in primary and secondary metabolism. Primary metabolites ensure proper growth, reproduction and development, whereas secondary metabolites, or phytochemicals, are important for plant defense and also act in attracting other organisms and signaling (Nützmann et al., 2016).

Metabolite (and expression) levels are considered complex quantitative traits and genomic regions that affect these levels can be elucidated by quantitative trait locus (mQTL; metabolite QTL, eQTL; expression QTL) analysis. They are called complex quantitative traits in the sense that the trait is affected by multiple genes and can be affected by environmental factors. It can also be, in the case of epistasis, that the effect of one gene depends on the expression of other genes, which adds another layer of complexity. xQTL (mQTL or eQTL) analysis involves annotating and measuring metabolite/mRNA levels combined with genotyping of inbred lines. Recombinant inbred lines (RILs) by single-seed descent (SSD) are often used to achieve transgressive segregation of traits by crossing two parents with distinct phenotypes. The RILs are generated by selfing individual members of the second generation (F_2) until nearly complete homozygous individuals are obtained. In a combined analysis, because of the differences between genome composition and metabolite/mRNA levels in the RILs, genomic regions can be identified that affect certain metabolite/mRNA levels (Carreno-Quintero et al., 2013). Metabolic profiling of plants during mQTL

analyses is often done by mass spectrometry (MS) following gas chromatography (GC) or liquid chromatography (LC). However, it is important to note that the combination of GC-MS mostly detects primary metabolites, whereas LC-MS detects a wide array of secondary metabolites. This has to do with the fact that most secondary metabolites are not volatile and therefore cannot be detected with GC (Rowe et al., 2008).

Recent studies found physical clustering of genes in plants involved in sequential enzymatic reactions, as is also true for Bacteria in the case of operons, of biosynthetic pathways for secondary metabolite synthesis. These clustered genes are called biosynthetic gene clusters (BGCs) whenever they contain genes coding for at least three different types of enzymes (Nützmann et al., 2016). Plant BGCs often contain the gene encoding the enzyme responsible for the first step in the pathway and two or more other genes for enzymes downstream in the pathway, often interspersed with unrelated genes. There are some hypotheses about the reason behind clustering of these pathway genes, like: less risk of disruption (and thus loss of the pathway) by recombination, less chance of toxic intermediates, higher chance of co-regulation because of co-localization, and benefits for formation of multi-enzyme complexes (Nützmann et al., 2016). Recently tools, like PlantiSMASH (Kautsar et al., 2017) and PlantClusterFinder (Schläpfer et al., 2017), have been developed to both discover BGCs related to known clusters in databases and predict BGCs *de novo*.

Integrating mQTL, eQTL and predicted BGCs might lead to the discovery of novel biosynthetic pathways or extend the knowledge about known ones. This knowledge can in turn be used to exploit a plant for its secondary metabolites, for e.g. application as drugs or cosmetics. mQTLs make it possible to link BGCs to metabolites. If the mQTL is for the content of a secondary metabolite and mapped to a specific (clustered) biosynthetic gene of known function or containing known domains, this gene might be causal for the variation in secondary metabolite content. This information can then be used to up- or downregulate the production of the secondary metabolite in the plant by genomic engineering. eQTLs might illuminate the regulation of both BGCs and mQTLs. If the eQTL is for the expression of a biosynthetic gene and mapped to the same biosynthetic gene or a location elsewhere in the genome, the mapping locations might be causal for the variation in the biosynthetic gene. These mapping locations might be promoter regions, transcription factors, enhancers or other regulatory elements influence the biosynthetic gene's expression level. This knowledge can in turn be used to elucidate regulation of biosynthetic pathways. Hence, *in silico* integration of these datatypes seems promising for generating hypotheses of biosynthesis of secondary metabolites and its regulation.

A combination of mQTL and eQTL analyses, and linking these, has already been done by Wentzell et al. in 2007 in *Arabidopsis thaliana* RILs, in which they associated polymorphisms influencing expression and metabolite levels in two glucosinolate biosynthetic networks (aliphatic and indolic). They found that all eQTLs for genes in these specific pathways also affected the accumulation of the corresponding metabolites. Furthermore, epistasis was more often detected for metabolites and their broad sense heritability (H^2, ratio between genetic variance and total variance) was lower in comparison to gene expression levels. These results show that metabolic traits are more affected by environmental factors than gene expression levels. However, combining mQTLs with eQTLs allowed them to understand the regulation of these two biosynthetic pathways better, despite the complexity of epistasis and low H^2 for metabolites (Wentzell et al., 2007).

Here, methods were developed to integrate xQTL datasets with BGCs predicted with PlantiSMASH based on genomic locational overlap. Datasets of both a RIL population and a Genome-Wide Association Study (GWAS) of *Oryza sativa* and *Arabidopsis thaliana* respectively were used. Hereafter, the detected overlap was validated and further analyzed to explain the overlap using literature, genome annotations, and biochemistry knowledge.

Methodology

Realization of the proposed goals requires integrative analyses. The methods consist of preprocessing the (xQTL) data, running PlantiSMASH, finding genomic overlap between BGCs and xQTLs, and validation of the discovered overlap. Whenever settings are not specified for the used tools, default settings were used. The workflow and specific commands can be found below in appendices **1.** and **2. Commands**. Most graphs were made with ggplot2 (Wickham, 2009) in the R programming language (R Core Team, 2017), whereas the rest of the methods have been implemented in the Python programming language (Python Software Foundation, https://www.python.org/).

Data xQTL description

The *Oryza sativa* eQTL dataset was taken from the study of Wang et al. in 2014. The dataset contains 13,648 significant eQTLs (5,079 *cis*-eQTLs and 8,568 *trans*-eQTLs) with an average size of 1.50 Mb. Expression profiling was done with an Affymetrix Rice Genome Array (GPL2025) on flag leaf tissue at heading date of rice plants grown under normal agricultural conditions. The results of the expression profiling was later used for the coexpression analysis within PlantiSMASH, the data is stored in the NCBI GEO database with accession number GSE49020. The *O. sativa* mQTL dataset was derived from the study of Gong et al. in 2013. The dataset contains 2,822 significant mQTLs with an average size of 2.20 Mb. Metabolite profiling was done with LC-MS/MS on flag leaf tissue at heading date of rice plants grown under normal agricultural conditions. Both studies used the same RIL population consisting of 210 lines from a cross between Zhenshan 97 and Minghui 63. Furthermore, the same strategy for QTL mapping was used, namely, composite interval mapping with R/qtl (Broman et al., 2003) based on 1,619 recombinant bins. The locations of the xQTLs correspond to the MSUv6.1 (Kawahara et al., 2013) version of the rice genome (size: 373.2 Mb, gene density: 6.66 Kb/gene, number of genes: 55,986). The eQTL and mQTL densities were 0.027 and 0.132 Mb/QTL, respectively.

Besides the *O. sativa* RIL population, data from a GWAS of *Arabidopsis thaliana* were used as well. The eQTL dataset was derived from the study of Kawakatsu et al. in 2016, in which 1,227 different *A. thaliana* accessions were geno- and phenotyped (1,673,530 markers). Expression profiling was done with RNA-seq using the Illumina HiSeq 2500 sequencer (Illumina, Inc., San Diego, CA) on leaf tissue from rosettes just before bolting under normal conditions. A linear mixed model in the LIMIX Python package (Lippert et al., 2014) was applied to the genotype and gene expression matrix, resulting in 2,185 significant eQTLs. The *A. thaliana* mQTL dataset was taken from an unpublished GWAS (Kooke et al, unpublished data), in which 349 different *A. thaliana* accessions were geno- and phenotyped (214,051 markers). Metabolite profiling was done with GC-MS and LC-MS on full rosette leaf tissue under normal conditions. Raw MS spectral data were processed with MSClust (Tikunov et al., 2012) and these data were used later for validation/annotation. MSClust uses unsupervised fuzzy clustering to extract putative metabolite mass spectra (Tikunov et al., 2012). Linear mixed models in EMMAX (Kang et al., 2008) and the GAPIT R package (Lipka et al., 2012) were applied to the genotype and metabolite profiling matrix, resulting in 1,897 significant mQTLs. The locations of both studies' xQTLs correspond to the TAIR10 version (Swarbreck et al., 2008) of the thale cress genome (size: 119.7 Mb, gene density: 4.35 Kb/gene, number of genes: 33,602). A total number of 175 *A. thaliana* accessions overlap between the two GWASs. The eQTL and mQTL densities were 0.055 and 0.063 Mb/QTL, respectively.

Data xQTL formatting

Both the eQTL and mQTL data were stored in tab-separated text files containing the following columns. The first column contains locus tag names (OsXXgXXXXX, AtXgXXXXX) in case of eQTLs and metabolite names for mQTLs. Whenever the metabolite name is unknown, an artificial ID was made with the mass

and retention time. The second column contains the chromosome numbers of the QTLs. The third, fourth and fifth columns contain locations of the peak, start (inferior boundary) and end (superior boundary) of the QTL (in Mb), respectively. The LOD-scores are stored in the last column. For the *A. thaliana* GWASs, the method of linear mixed models only records the best association between a SNP and expression/metabolite levels. The actual causal SNP of the variation in expression/metabolite level can be anywhere between the previous and next marker with respect to the associated marker. This was taken into account when posing hypotheses of gene/metabolite – BGC associations. The average marker density of the entire *A. thaliana* genome was 1 marker every 565.47 basepairs (for the mQTL study, the marker density was higher for the eQTL study, namely 1 marker every 71.19 basepairs). The QTL regions were artificially extended, to match the format, by adding and subtracting the average marker density specific to the QTL analysis. For the mQTL study, the marker list that was used with exact positions was available. An example of the data format can be seen in appendix **3. Input data format xQTLs**.

Running PlantiSMASH and formatting its output

BGCs were predicted with PlantiSMASH (Kautsar et al., 2017) for both the *O. sativa* and *A. thaliana* genomes. A separate GFF3 and FASTA file with genome version MSUv6.1 (Kawahara et al., 2013) was used as the input for rice BGC prediction. For thale cress, the NCBI GenBank genome file (GBFF) of TAIR10 (Swarbreck et al., 2008) was used as input, as PlantiSMASH accepts both. The gene expression profiling by array of all *O. sativa* RILs including parents was given as input for PlantiSMASH as it is capable of performing a coexpression analysis. For both species, clusterBLAST and knownclusterBLAST were turned on, making PlantiSMASH search for similar BGCs in other plants and known BGCs by referring to MiBiG (Medema et al., 2015), respectively. The settings for BGC prediction were made less strict by lowering the minimum number of unique domains per BGC from two to one, and by increasing the CD-HIT cutoff from 0.5 to 0.6. The output files `XX_BGC.txt` (where `XX` refers to the chromosomes) in the folder named `txt` produced by PlantiSMASH was used to parse the BGC data for finding genomic overlap. The following elements were parsed from these files: clusterID as given by PlantiSMASH, clustertype (saccharide, terpene, polyketide, alkaloid, putative etc.), chromosome, start and end of the BGC in bp, and a list of genes belonging to the BGC. The coexpression analysis output was used for validation of the genomic overlap, as well as the output of the knownclusterBLAST analysis (see the sections **Coexpression analysis of genes in BGCs with overlapping xQTLs** and **Confirming known BGC-product pairs** below).

Genomic overlap between BGCs and xQTLs

Three possibilities for genomic overlap between BGCs and xQTLs were taken into account. The first possibility is when the peak of an xQTL is within the boundaries of the BGC. The second possibility is when the inferior boundary of the xQTL is within the boundaries of the BGC, with a minimum overlap of 30% (arbitrary, but led to half the detected overlap for some BGCs in rice without losing too many other BGCs with overlap) of the BGC's size. The last possibility is when the superior boundary of the xQTL is within the boundaries of the BGC, with a minimum overlap of 30% of the BGC's size. Visualization of these rules can be seen in appendix **4. Overlap rules**. The rule for minimum overlap of 30% of BGC's size was only applied to the rice xQTL data as these xQTLs are much larger than the thale cress ones. The output is a tab-separated text file containing BGC information (clusterID, clustertype, chromosome, and start and end in bp), followed by information on the xQTLs (locus tag/metabolite name, p-value, adjusted p-value, LOD-score, locus annotation, locus start bp, locus end bp, and locus status) found to be overlapping. The p-values are derived from permutation tests, which will be explained in the validation section. The locus annotation is parsed from the GFF3 file, as well as the locus start and end position. Locus status indicates whether the gene for which the eQTL was found belongs to the BGC, if 'true' the status is local, when

'false' the status is distant. An example of the output format can be seen in appendix **5. Output data format xQTLs and BGCs overlap**.

Colocation network: genomic overlap between xQTLs independent of BGCs

Another option is to look for genomic overlap between xQTLs independent of BGCs. This might potentially lead to the discovery of BGCs that were not predicted by PlantiSMASH, but can as well give hints to regulatory regions. One can look at three possibilities: overlap between mQTLs, eQTLs, and mQTLs and eQTLs. Here, overlap was again defined as overlapping genomic regions with the same rules applied as for the overlap finding between BGCs and xQTLs. Again, it made sense to apply the minimum overlap rule solely to the *O. sativa* xQTLs. Due to time constraints, we only looked at overlap between mQTLs and between mQTLs and eQTLs in *A. thaliana*. The detected overlapping xQTL pairs were given a(n) (adjusted) p-value by using the permutation test procedure described below, only using the Benjamini-Hochberg correction. A colocation network was constructed of the outcome with Cytoscape v3.4.3 (Shannon et al., 2003) where nodes represent xQTLs and the edges represent the overlap between the xQTLs with edge weight as the $-\log_{10}$ of the adjusted p-values. An example of the output format can be seen in appendix **6. Output data format overlapping mQTLs**.

Validation

Several approaches were used to validate the genomic overlap that was found and assess the feasibility of the described method: permutation test, confirming known BGC-product pairs, analysis of MSClust data, variant calling, and a coexpression analysis of genes in BGCs. Hereafter, these methods are described in more detail in the abovementioned order.

Permutation test

A permutation or randomization test was applied to test for likelihood of random occurrence of genomic overlap between xQTLs and BGCs. The randomization test involved generating independent random xQTL and BGC data by shuffling chromosome number and chromosomal location independently, keeping the size of the regions the same. The actual chromosomal number and length were used as the input for shuffling. Hereafter the random data were used as an input for the function to detect genomic overlap between xQTLs and BGCs. This procedure is equal to one permutation. Thousand permutations were performed to calculate the likelihood of a detected overlap. This likelihood was calculated by dividing the number of times the specific overlap was found by the number of eQTLs or mQTLs (depending on the type of overlap). Both Benjamini-Hochberg (Benjamini et al., 1995) and Bonferroni (Bonferroni, 1936) multiple testing corrections were applied to the resulting p-values. The adjusted p-value threshold was set to 0.05 for both correction methods, however, because of the exploratory nature of this study no overlaps were discarded but the adjusted p-values of both correction methods were taken into account for validation. Furthermore, the Bonferroni correction seemed too conservative as seen by the number of insignificant overlaps, even for the *A. thaliana* QTL data.

Confirming known BGC-product pairs

For some BGCs in *O. sativa* and *A. thaliana* the product is known. These known BGC-product pairs can be used as a way of validation of the genomic overlap method that was described here. If the known products are detected in the MS analyses and associated with a genomic region, then there will be an mQTL for this product in the datasets that were used. If this mQTL overlaps with the BGC that it was already associated with, this strengthens the confidence in the method developed here. The following products are already associated with BGCs in *O. sativa*: phytocassane/oryzalide and momilactone. The following products are already associated with BGCs in *A. thaliana*: arabidiol/baruol, tirucalla, marneral and thalianol. A more

detailed description of the BGCs with known products is presented in appendix **7. Detailed overview of BGCs with known products in *O. sativa* and *A. thaliana*.**

Analyzing MS(Clust) data

Raw spectral LC-MS data processed with MSClust (Tikunov et al., 2012) were available for the *A. thaliana* GWAS. As not all metabolites were annotated in the mQTL dataset and appeared as cluster IDs referring to the output of MSClust, the latter dataset was used to find the cluster IDs and the masses of the measured compounds belonging to that cluster. As a cluster in MSClust can contain multiple fragments, first, the parent ion had to be identified. The parent ion mass was then used to search the following plant-specific MS databases for hypothetical annotation: ReSpect (Sawada et al., 2012) and KNapSacK (Nakamura et al., 2013). The exact masses of the metabolites having an overlapping mQTL in the *O. sativa* dataset were also searched again in these databases. The structures of the potential metabolite candidates were analyzed as well, looking for specific groups matching genes in the BGC (for example, finding hydroxy groups and having dioxygenases in the BGC). PubChem (Kim et al., 2015) was used to search for structural information of compounds.

Variant calling

As Illumina HiSeq 2000 (Illumina, Inc., San Diego, CA) WGS paired reads with multiple insert sizes were available for both parents of the recombinant inbred line population of *O. sativa* and since their xQTLs are large, a SNP calling procedure was applied to verify if the genes in the biosynthetic gene clusters are actually variable between the parents of the population. Variation in these genes might then be responsible for the variation in expression and/or metabolite levels that were found with the xQTLs. The following NCBI SRA (Leinonen et al., 2011) datasets were used for the Minghui 63 parent: SRR3234369, SRR3234370 and SRR3234371. Whereas the following were used for the Zhenshan 97 parent: SRR3234372, SRR3234373 and SRR32374. For both parents paired-end libraries were sequenced with three different insert sizes: 300 bp, 5 kb and 10 kb. The total coverage for Minghui 63 before trimming and filtering is approximately 185x, whereas for Zhenshan 97 it is 261x. Hereafter, the methodology will be described briefly, specific command and settings can be found in appendix **2. Commands.** Reads were trimmed and filtered with Trimmomatic v0.36 (Bolger et al., 2014), possible adapters were removed as well. Read quality was checked before and after trimming with FastQC v0.11.7 (Andrews, 2010). Trimmed reads of both parents were mapped (`--sensitive`) with Bowtie2 v2.2.6 (Langmead et al., 2012) against the indexed MSUv6.1 rice genome reference (Kawahara et al., 2013). The resulting SAM files were converted to sorted BAM files and indexed with SAMtools v1.7 (Li et al., 2009). SNPs and indels were called with SAMtools `mpileup` and BCFtools `call` v1.6 (Narasimhan et al., 2016). IGV v2.4.8 (Thorvaldsdóttir et al., 2013) was used to visualize and search through the called variants, BCFtools `stats` was used to create some statistics of the variant calling. The goal was to look for non-synonymous variants in the domains of genes coding for enzymes important for the BGC. Low quality variants (< 50) were not considered.

Coexpression analysis of genes in BGCs with overlapping xQTLs

The last method of validation that was applied, was a coexpression analysis on the transcription profiling array data from the RIL population of *O. sativa* (GSE49020), consisting of 216 samples (one replicate from each RIL, and three replicates from each parent). Predicting BGCs is one thing, however when the genes in the BGCs are actually coexpressed, this indicates that the BGC is active and its genes might be co-regulated where the latter is an important characteristic and/or benefit of BGCs. However, gene expression still depends on the conditions that were used in the experiments, some genes are solely expressed under very specific (environmental) conditions. The coexpression analysis method based on Pearson Correlation (PC) implemented in PlantiSMASH was used with default settings. However, first, the

Affymetrix probe names in the expression matrix were parsed in to locus tag as previously used (LOC_OsXXgXXXXX). Not all probes had a matching locus tag in Ensembl BioMart (Kinsella et al., 2011). The output graphs of the PlantiSMASH coexpression analysis were used to interpret results: expression heatmaps, coexpression networks (edges are drawn when the PC coefficient reached a certain threshold) and hive plots (showing inter-cluster coexpression).

Source code

All written code can be found on https://github.com/lottewitjes/MSc_thesis.

Results

Since the method for finding overlap described above was executed for both *O. sativa* and *A. thaliana*, but not all validation methods, the results section is divided in two subsections, one for each organism/study.

Oryza sativa, recombinant inbred line population

The overlap between xQTLs and BGCs analysis for *O. sativa* resulted in finding overlap with an xQTL for 31 (13 with mQTLs) out of 49 BGCs predicted by PlantiSMASH. **Figure 1** shows the number of overlapping xQTLs on the x-axis versus the 49 BGCs represented by their cluster type on the y-axis. The color represents the different QTL types, whereas the alternating background color represents the different chromosomes, starting from one at the origin to twelve at the top. Local eQTLs (red) are eQTLs of genes that belong to the BGC. Distant eQTLs (blue) are eQTLs of genes outside the BGC. Metabolite QTLs (mQTLs) are represented in green. The black arrow shows the known BGC-products pair of phytocassane, the BGC for momilactone was not found in this genome. It can be seen that the number of BGCs and overlapping xQTLs are not equally distributed throughout the genome. PlantiSMASH predicted only one BGC on chromosome twelve, whereas there were six on chromosome six. The number of overlapping xQTLs that was found for the different BGCs varied as well: the terpene cluster on chromosome two had over 50 xQTLs, and for some clusters no or little overlap was found. There were more overlapping distant eQTLs than local eQTLs and mQTLs. These observations matched with the number of xQTLs in the dataset, graphs of the xQTL distributions can be seen in **Supplemental figure 6 A and B** in appendix
8. Distribution of xQTLs throughout the genomes of *O.* sativa and *A. thaliana*.

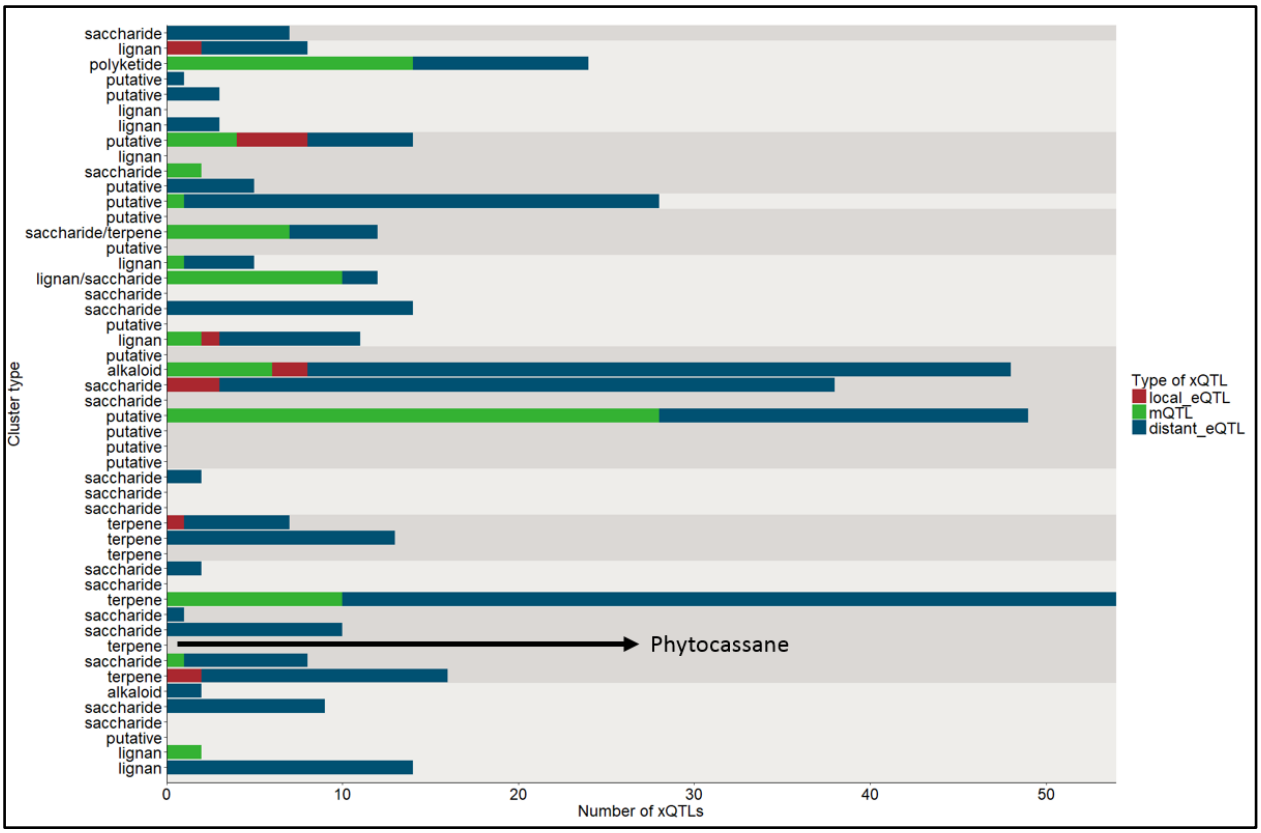


Figure 1 The number of overlapping xQTLs on the x-axis versus the BGCs represented by their cluster type on the y-axis for *O. sativa*. The alternating background color represents the chromosomes, starting with chromosome one at the origin to chromosome twelve at the top of the figure. Local eQTLs (red) are eQTLs of genes that belong to the BGC. Distant eQTLs (blue) are eQTLs of genes outside the BGC. Metabolite QTLs (mQTL, green) are mQTLs of metabolites. The BGC with a known product (phytocassane) is indicated with the black arrow.

The variant calling analysis between the parents of the RIL population resulted in 3,738,137 SNPs and 763,542 indels, and the ratio between transitions and transversions was 2.44. The latter ratio is as expected, since transitions occur approximately twice as often as transversions (Collins et al., 1994). The PlantiSMASH coexpression analysis resulted in having six BGCs with coexpression networks (Pearson correlation coefficient, PCC, threshold 0.5). The fifth BGC of type saccharide had a coexpression network containing five genes (with edges ranging from PCC 0.50-0.63), all other networks had only two nodes (with edges ranging from PCC 0.52-0.64). A coexpression network was visible whenever a BGC is expressed in the gene expression profiling analysis and edges were visible whenever genes within the BGC were coexpressed. The Bonferroni correction on the p-values of the permutation test seemed too conservative for the exploratory nature of this study, and therefore only the Benjamini-Hochberg corrected p-values (BH) will be named together with the LOD-score of the QTL analysis. There were mQTLs for both phytocassane A and C, however no overlap was found between these mQTLs and the phytocassane BGC.

When looking closer at the found overlap between BGCs and xQTLs, there are some examples that give some confidence to the methods described here. Still, all that will be described hereafter is hypothetical. The first example involves the second BGC of type lignan on the first chromosome. This BGC showed overlap with two mQTLs whereof one was lehmbachol A (BH: 0.01343, LOD-score: 3.3). Lehmbachol A (PubChem ID: 102066461) is a stilbenolignan, which matched the BGC type predicted by PlantiSMASH: lignan. **Figure 2** shows the 2D structure of lehmbachol A. The BGC contained two dirigent enzymes and two dioxygenases. Dirigent enzymes are important for plant secondary metabolism and lack catalytic

activity but are capable of directing the outcome of bimolecular coupling reactions (Pickel et al., 2013). One of the dioxygenases (LOC_Os01g25010) had three non-synonymous SNPs (methionine to threonine, alanine to glycine, and valine to phenylalanine) in the dioxygenase domain (2OG_FeII_Oxy) of the resulting protein and no variants in the other dioxygenase domain (DIOX_N). Since lehmbachol has hydroxy groups, these SNPs might be causal for the variation in the lehmbachol A content and therefore this BGC might be involved in its biosynthesis. **Supplemental figure 7** presents an example of the variant browsing in IGV, the variants in LOC_Os01g25010 are shown. The PlantiSMASH coexpression analysis showed no coexpression between genes in this BGC.

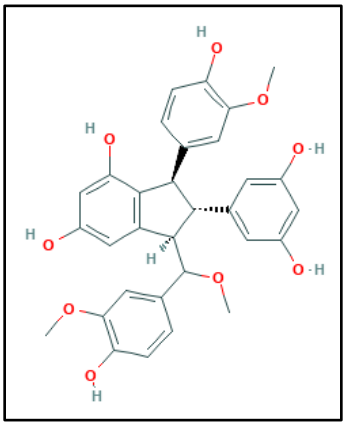


Figure 2 The 2D structure of lehmbachol A. Taken and adjusted from PubChem (Kim et al., 2015).

The second example involves BGC 42 of type 'putative' on the tenth chromosome. This BGC showed overlap with ten eQTLs whereof three eQTLs were of genes with dioxygenase domains within the BGC: LOC_Os10g40960 (BH: 0.03, 0.03, LOD-score: 20.47, 24.07) and LOC_Os10g40990 (BH: 0, LOD-score: 63.65). There were four overlapping mQTLs as well, where three masses (611.1607, 655.2133, 801.2237 Da) were searched in the ReSpect and KNapSacK databases. The 611 Da mass gave a hit with cyanidin 3,5-diglucoside, the 655 Da mass with malvidin 3,5-diglucoside, and the 801 Da mass with malvidin 3-(coumaroyl)-5-diglucoside. All three mQTLs corresponding to the named masses had significant BH adjusted p-values and LOD-scores ranging from 4.1-5.6. Since the difference between cyanidin (PubChem ID: 128861) and malvidin (PubChem ID: 159287) is one methoxy and methyl group, it might be that the six dioxygenases present in the BGC are involved in adding an oxygen atom to cyanidin, where after methyltransferases (situated elsewhere) finish the conversion. **Figure 3A and 3B** show the 2D structures of cyanidin and malvidin, respectively. This BGC might be involved in the conversion of cyanidin sugars to malvidin sugars. LOC_Os10g40990 (UniProt ID: Q336S9_ORYSJ) is a putative flavonol synthase and had one non-synonymous SNP (glutamine to histidine) in the dioxygenase domain (DIOX_N), that might be responsible for the cis-eQTL and the mQTLs. There are no known active site residues for this dioxygenase, therefore it is unknown if this non-synonymous SNP is at an active site residue. Homology modelling with SWISS-MODEL (Biasini et al., 2014), based on a template with 39% protein sequence similarity (PDB ID: 5O7Y, protein name: thebaine 6-O-demethylase, and active site prediction with COFACTOR (Zhang et al., 2017) lead to no predicted active residues. The model is presented in **figure 4**. The PlantiSMASH coexpression analysis showed no coexpression between genes in this BGC.

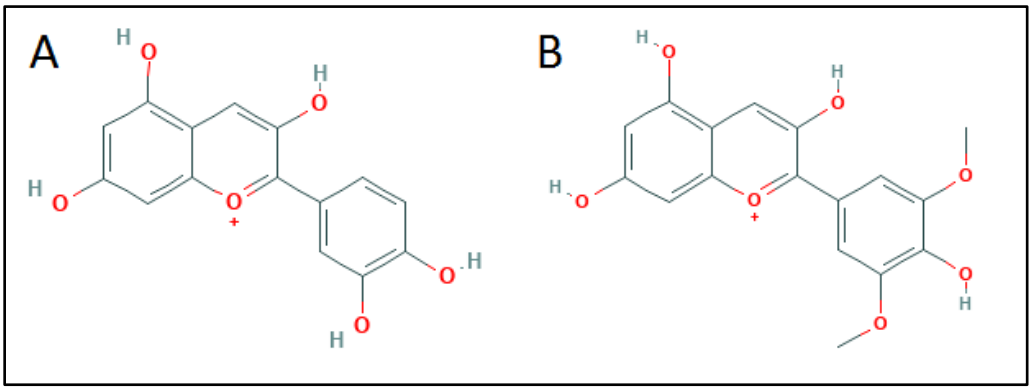


Figure 3 The 2D structures of **A** cyanidin and **B** malvidin. Taken and adjusted from PubChem (Kim et al., 2015).

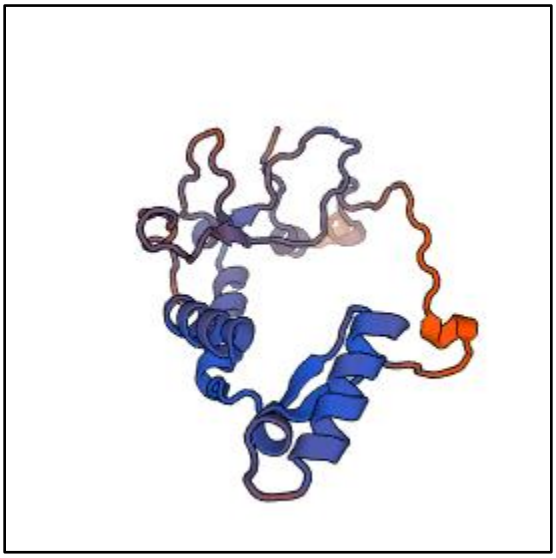


Figure 4 A 3D structural model of the protein encoded by LOC_Os10g40990 modelled with SWISS-MODEL (Biasini et al., 2014).

The last example involves BGC 47 of type polyketide on the eleventh chromosome. This BGC showed overlap with 10 eQTLs and 14 mQTLs. One of the mQTLs is for isogemichalcone B (BH: 0.00870, LOD-score: 4.5). Isogemichalcone B (PubChem ID: 42607532) is a compound of type polyketide, which matched the BGC type predicted by PlantiSMASH. **Figure 5** shows the 2D structure of isogemichalcone B. The BGC contained three ketosynthases whereof one had one non-synonymous SNP (glycine to serine) in the ketosynthase domain (Chal_sti_synth_N), which might be the causality of the variation in the isogemichalcone B content.

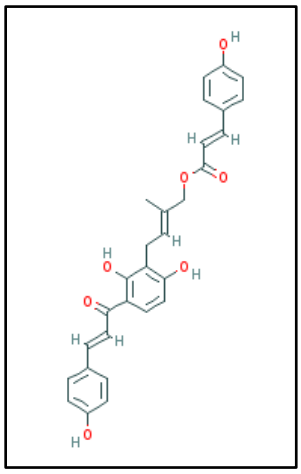


Figure 5 The 2D structure of isogemichalcone B. Taken and adjusted from PubChem (Kim et al., 2015).

Arabidopsis thaliana, genome-wide association study

The overlap between xQTLs and BGCs analysis for *A. thaliana* resulted in finding overlapping xQTLs for 36 (33 with mQTLs) out of 48 BGCs predicted by PlantiSMASH. **Figure 6** shows the number of overlapping xQTLs on the x-axis versus the 48 BGCs represented by their cluster type on the y-axis. The color represents the different QTLs, whereas the alternating background color represents the different chromosomes, starting from one at the origin to five at the top. Local eQTLs (red) are eQTLs of genes that belong to the BGC. Distant eQTLs (blue) are eQTLs of genes outside the BGC. Metabolite QTLs (mQTLs) are represented in green. The black arrows point towards known BGC-products pairs, all that are known for *A. thaliana* were detected by PlantiSMASH. Here as well it can be seen that the number of BGCs and overlapping xQTLs were not evenly distributed throughout the genome. PlantiSMASH predicted fewer BGCs on chromosome four in comparison with the others. The number of overlapping xQTLs that was found for the different BGCs varied as well, the second saccharide cluster on chromosome two had over 20 xQTLs, and for some clusters no or little overlap was found. There were more overlapping mQTLs than local and distant eQTLs. Graphs of the xQTL distributions can be seen in **Supplemental figure 6 C and D** in **8. Distribution of xQTLs throughout the genomes *of O*. sativa *and A.* *thaliana*.** There were no overlapping local eQTLs on chromosome three, four and five, however **Supplemental figure 6 C** shows that there were numerous eQTLs associated with these chromosomes.

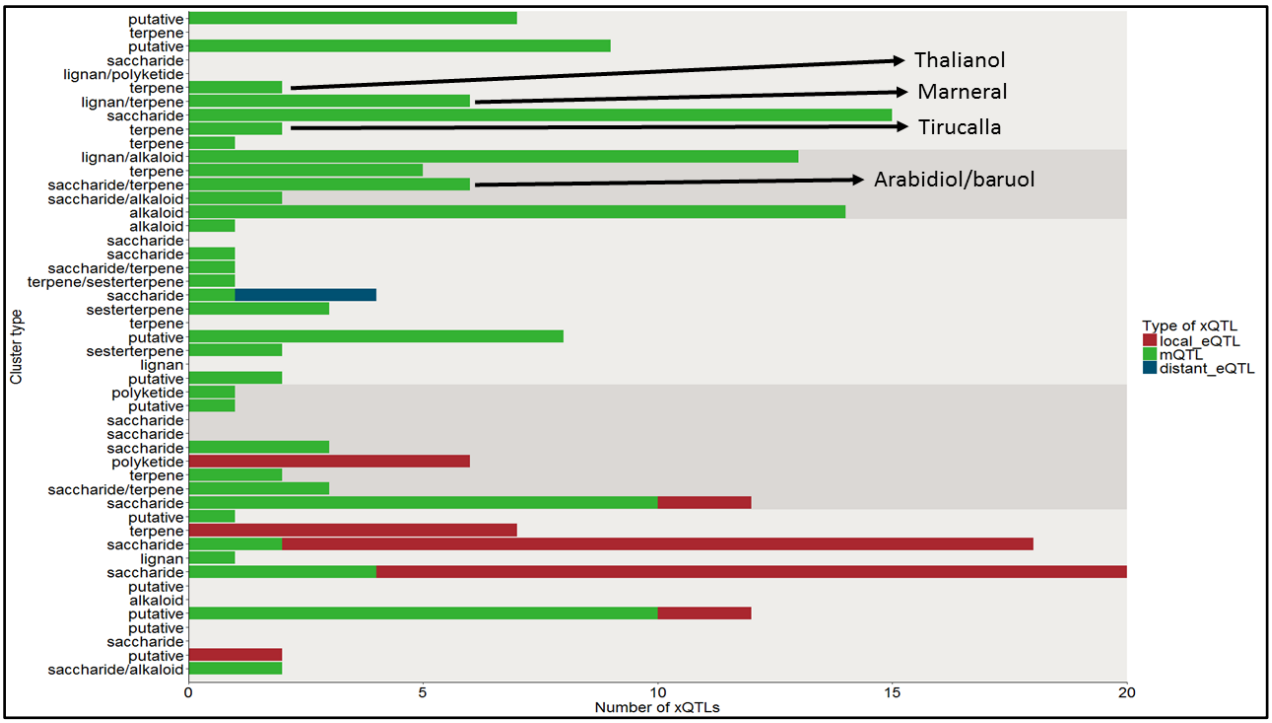


Figure 6 The number of overlapping xQTLs on the x-axis versus the BGCs represented by their cluster type on the y-axis for *A. thaliana*. The alternating background color represents the chromosomes, starting with chromosome one at the origin to chromosome five at the top of the figure. Local eQTLs (red) are eQTLs of genes that belong to the BGC. Distant eQTLs (blue) are eQTLs of genes outside the BGC. Metabolite QTLs (mQTL, green) are mQTLs of metabolites. The BGCs with known products are indicated with black arrows.

When taking a closer look at the found overlap between BGCs and xQTLs, there are again examples that give some confidence to the methods described here, both for RIL-based and GWA studies. Still, all that will be described hereafter is hypothetical, however *A. thaliana*'s genome is better annotated than the rice genome and the GWASs had higher resolution. No mQTLs were annotated in the dataset for arabidiol/baruol, tirucalla, marneral and thalianol, however they still might be amongst the mQTL with unidentified masses.

The first example involves the thirteenth BGC of type saccharide on the second chromosome. It showed overlap with two cis-eQTLs for SCPL12 (serine carboxypeptidase-like 12, however it is a Scl acyltransferase) and 10 mQTLs. One of the mQTLs was for kaempferitrin, which showed an association with the marker for SCPL11 (serine carboxypeptidase-like 11, however it is a Scl acyltransferase). Kaempferitrin (PubChem ID: 21159160) is a kaempferol with two sugar groups on the third and seventh carbon atom, as **Figure 7** shows. The BGC contained a UDP-glycosyltransferase (UniProtKB ID: O81010, ORF name: T20K9.14, gene name: UGT79B8) which has the following GO molecular function terms (Ashburner et al., 2000): quercetin 3-O-glucosyltransferase (GO:0080043) and quercetin 7-O-glucosyltransferase (GO:0080044) activity, both inferred from biological aspect of ancestor (IBA). It might be that this glycosyltransferase also acts on kaempferol to add sugar groups since kaempferol and quercetin only differ in one hydroxy group. Therefore, this BGC might be involved in the production of kaempferitrin.

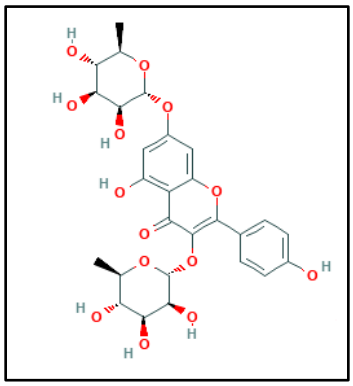


Figure 7 The 2D structure of kaempferitrin. Taken and adjusted from PubChem (Kim et al., 2015).

The second example involves BGC 36 of type saccharide/terpene on the fourth chromosome. The knownclusterBLAST analysis of PlantiSMASH indicated that this BGC is the known BGC for arabidiol/baruol (83% similarity). The BGC showed overlap with six mQTLs. All mQTLs had unidentified masses, however one of the mQTLs associated with a particular MSClust cluster (1_246) contained masses around 420-500 Da, it might be that this cluster contains arabidiol (PubChem ID: 25245907) and/or baruol (PubChem ID: 25203718), since the masses of arabidiol and baruol are 444.744 and 426.729 Da, respectively. One of the masses in this cluster was 503.1238 Da, this could be arabidiol with acetic acid as an adduct (M + acetic acid - H, mass: 503.7579 Da), since the LC-MS analysis was in negative mode. **Figure 8A and 8B** show the 2D structures of arabidiol and baruol. Cluster 1_246 was associated with the marker for the pentacyclic triterpene synthase 1 (PEN1), which is known to convert oxidosqualene to arabidiol (Xiang et al., 2006).

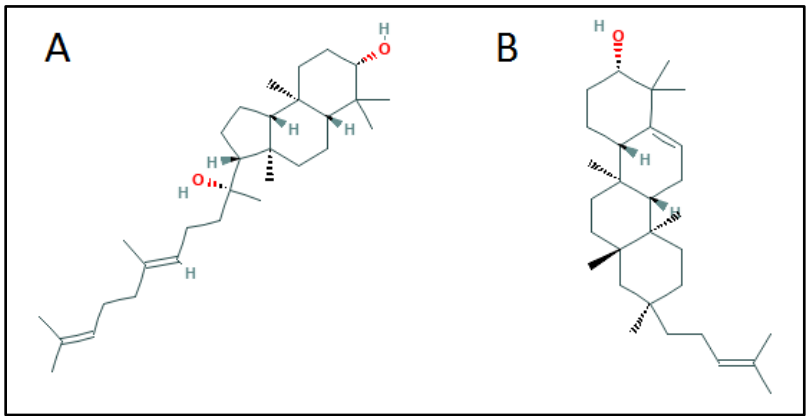


Figure 8 The 2D structures of **A** arabidiol and **B** baruol. Taken and adjusted from PubChem (Kim et al., 2015).

The last example involves BGC 46 of type putative on the fifth chromosome. It showed overlap with 9 mQTLs whereof one mQTL (BH: 0, LOD-score: 7.463) was for methoxyglucobrassicin (PubChem ID glucobrassicin: 5484743), a glucosinolate. **Figure 9** shows the 2D structure of glucobrassicin. Methoxyglucobrassicin was found to be associated with the marker for CYP81F2. CYP81F2 codes for a indol-3-yl-methylglucosinolate hydroxylase and was previously found to be involved in the biosynthesis of glucobrassicin in *Brassica oleracae* (Sotelo et al., 2016). CYP81F2 was also previously proven to be capable of hydroxylating the glucosinolate indole ring in *A. thaliana* (Pfalz et al., 2011). The BGC also contained the gene MJB24.4 (UniProtKB ID: Q9LVD5) coding for a putative thioredoxin superfamily protein, that might act on the sulfur atoms in (methoxy)glucobrassicin as well. At last, the BGC also contained several genes with methyltransferase domains, however none of them were previously identified as capable of transferring methyl groups to glucobrassicins (or other glucosinolates).

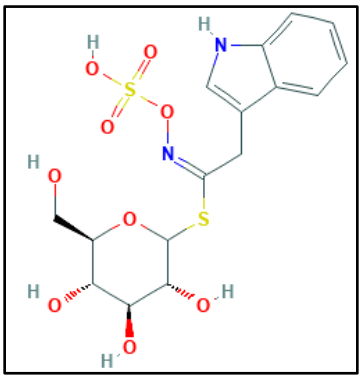


Figure 9 The 2D structure of glucobrassicin. Taken and adjusted from PubChem (Kim et al., 2015).

Finally, a colocation network analysis was done to look for overlap between mQTLs and between mQTLs and eQTLs independent of the predicted BGCs in *A. thaliana*. This approach might lead to novel BGC discovery and links between metabolites and genes. The resulting network is shown in **Supplemental figure 8**, and had 671 nodes and 2734 edges. Some hubs can be seen in the network and two of those, involving secondary metabolites, are presented here. Unfortunately, those hubs did not contain links between mQTLs and eQTLs.

Figure 10 presents the first network hub of overlapping mQTLs. This network hub contained some putative anthocyanidin sugars: malvidin, cyanidin and delphinidin sugars. It also contained ferulic acid and sinapoyl esters, and some flavonoid sugars (kaempferol, luteolin). All mQTLs were associated with the marker for OMT1 in the QTL analysis. OMT1 codes for O-methyltransferase 1 that is known to act on caffeic acid, hydroxyferulic acid, sinapoyl esters, and lignins (Zhang et al., 1997) (Goujon et al., 2003), however it also acts as a flavonol 3'-O-methyltransferase (Muzac et al., 2000). The associations between the metabolites and genes in the *A. thaliana* dataset thus correspond with literature, and it might be that this OMT1 also acts on anthocyanidin sugars. The latter is hypothesized here based on the network hub. In close physical proximity (within 20 kb) of OMT1 are some transcription factors, and the genes PTAC15 and PORA. PTAC15 encodes for a mTERF protein important in development processes (Kleine, T., 2012), whereas PORA encodes a light-dependent NADPH oxidoreductase that is involved in chlorophyll synthesis (Kim et al., 2012). Therefore OMT1 does not seem to be part of a BGC.

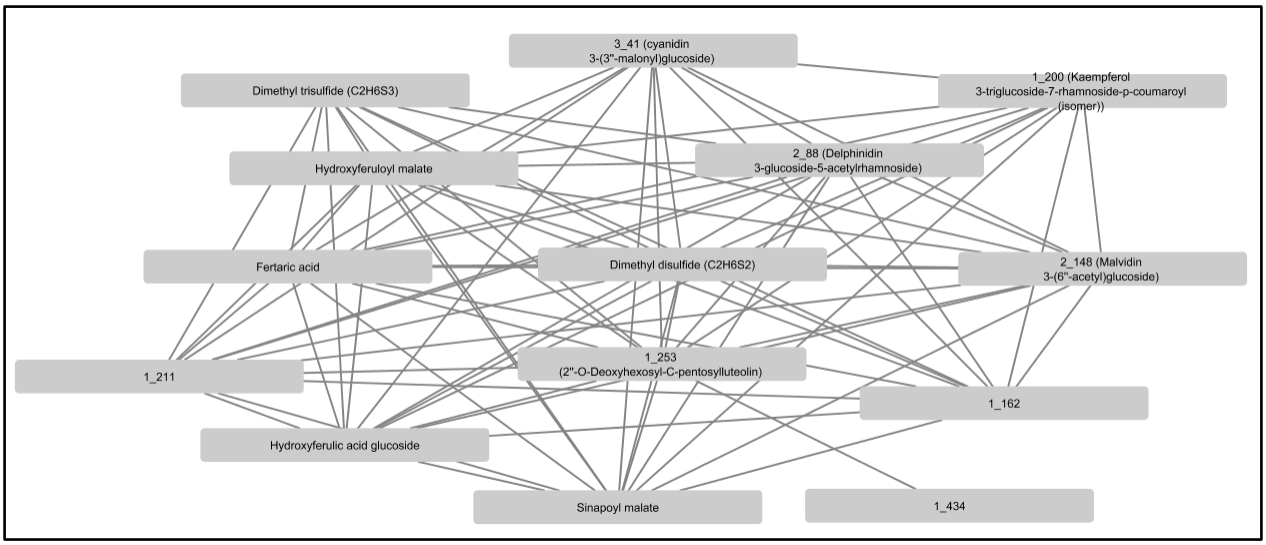


Figure 10 A network hub containing mQTLs (nodes) for anthocyanidin related compounds, flavonoid sugars, and ferulic acid and sinapoyl esters. The edges represent locational overlap between the mQTLs. All mQTLs in this hub are linked to the marker for OMT1.

Figure 11 presents the second network hub of overlapping mQTLs. This network hub contained some flavonoid sugars and one glucosinolate. All mQTLs were associated with the marker for BGLU6 in the QTL analysis. BGLU6 encodes a flavonol O-glucosyltransferase and was found to be important for the production of flavonol 3-O-gentiobioside 7-O-rhamnosides (both kaempferol and quercetin derived) in the study of Ishihara et al., from 2016. This network hub contains two kaempferol and one quercetin sugars, which is in line with the findings in the study. A hypothesis would be that this BGLU6 is also capable of O-glycosylation of (2-phenylethyl) glucosinolates. In close physical proximity (within 20 kb) to the BGLU6 gene are the genes coding for BGLU5 and NAC023. The latter is a transcription factor that plays a role in the determination of the position of shoot apical meristems (Ooka et al., 2003), whereas BGLU5 appears to be a pseudogene since it lacks certain motifs necessary for its glycosidase activity (Xu et al., 2004). Therefore, BGLU6 does not seem to be part of a BGC.

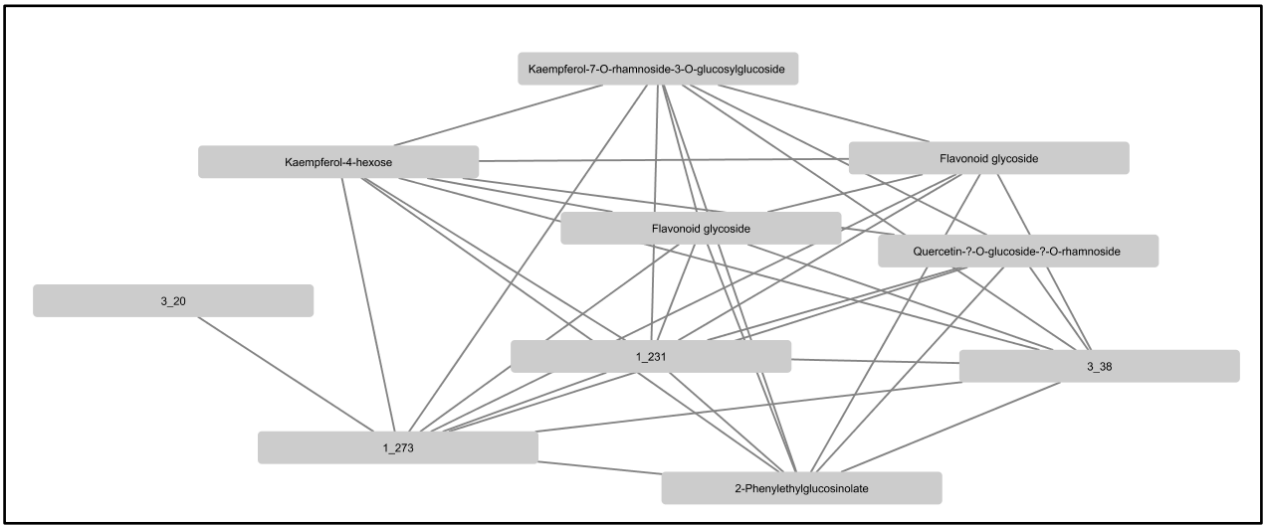


Figure 11 A network hub containing mQTLs (nodes) for flavonoid sugars and one glucosinolate. The edges represent locational overlap between the mQTLs. All mQTLs in this hub are linked to the marker for BGLU6.

All examples described in this section indicate that the method of integrating BGCs with eQTLs and mQTLs might work to link BGCs to putative products and regulating genes for both RIL populations and GWASs. Furthermore, the method to find overlapping xQTLs independent of predicted BGCs might work as well. Still all methods show some limitations and culprits as discussed in the next section.

Discussion

Since plants produce an ample array of mostly unknown secondary metabolites, natural product discovery is of high interest so that metabolites with beneficial characteristics for human can be exploited for applications, e.g. drug discovery (Nützmann et al., 2016). The goal of this study was to explore integration methods for predicted BGCs by PlantiSMASH and metabolite and expression QTL data for RIL populations and GWASs. This integration can lead to hypotheses generation of genes involved in the biosynthesis of secondary metabolites. The idea was that, by looking at genomic overlap of the BGC and xQTL regions, BGCs could be linked to metabolites and genes that regulate the expression of the BGCs. The **Results** section showed some examples that might indicate the methods developed here work for posing new hypotheses, albeit rather speculative. This section will discuss some remarkable discoveries, limitations, and future work.

Differences can be noticed between the overlap BGCs-xQTLs of the *O. sativa* and *A. thaliana* datasets, **Figure 1 and 5**. Some of these differences are explained by the number and distribution of the xQTLs in the genome (**Supplemental figure 6**). However, some are more difficult to explain, like the observation that there are no overlapping eQTLs on chromosome 4 and 5 of *A. thaliana*. This observation cannot be explained by the lack of eQTLs on these chromosomes, but it might be that the BGCs's genes on these chromosomes are more specialized and only expressed in certain conditions (abiotic or biotic stress) or tissues. The *A. thaliana* QTL analyses were executed under normal conditions, providing enough nutrients and without any added stresses. The same is true for the conditions during the *O. sativa* QTL analyses. This is a limitation as well for novel BGC product discovery, since secondary metabolites are mostly produced under stress and thus their biosynthesis genes too are mostly expressed under stress. Both studies used leaf tissue for gene expression and metabolite profiling.

In general, more overlap between BGCs and xQTLs was found for *O. sativa* despite the applied overlap cutoff, the larger genome and almost the same number of predicted BGCs in comparison to *A. thaliana*. However, this can be explained by the large number of xQTLs in the dataset and their much larger sizes (in the order of Mb) in comparison with the sizes of the *A. thaliana* xQTLs (in de order of kb). The difference in the number of local and distant overlapping eQTLs can as well be explained by the difference in eQTL sizes between the two datasets. The eQTL density is two times higher for *O. sativa* (0.027 versus 0.055 Mb/QTL), however A. thaliana's mQTL density is two times higher (0.063 vs 0.132 Mb/QTL). This, and difference in the total number of genes in the genomes (55,986 versus 33,602 genes), explain why more overlap is found with eQTLs in *O. sativa* compared to *A. thaliana*, and more overlap with mQTLs in *A. thaliana* compared to *O .sativa*.

The low resolution of the *O. sativa* xQTL datasets and the lack of an extensive genome annotation (in comparison to *A. thaliana*) caused limitations for the interpretation of the discovered overlap. Causal genes for metabolite content and gene expression variation could not be designated with any certainty. The variant calling procedure and the coexpression analyses for *O. sativa* added another source of information, making it easier to point out causal genes but still all links that are presented in the **Results** section are hypothetical.

The variant calling procedure between the parents of the *O. sativa* RIL population was used to assess if important genes in BGCs with overlapping xQTLs were actually variable. Non-synonymous SNPs were found in domains of some of these genes, however more work is needed to verify if these SNPs are actually causing problems in the (active sites of the) resulting enzymes and therefore can alter metabolite levels

and be the causality of the mQTLs. In case of the eQTLs, more research is needed to verify if there are SNPs in regulatory regions of the BGC's genes that might cause the variation in the gene expression levels and thus eQTLs. However, there is still the problem of the low resolution of the *O. sativa* datasets with xQTLs spanning several genes, which adds a lot of uncertainty to pointing out causal genes.
The coexpression analysis of the 216 samples for the *O. sativa* study did not show a lot of coexpression between the genes in the BGCs. This can be explained again by the conditions that were used in the experiments. Some genes are only expressed under certain stresses or in certain tissues or a combination of both. Furthermore, not all Affymetrix probe names were linked with a locus tag name of format OsXXgXXXXX in Ensembl BioMart (Kinsella et al., 2011), this caused a significant loss of gene expression data since only expression data of 11,465 genes out of 57,381 probes were kept and there are 55,986 genes (loci) in version MSUv6.1 of the rice genome (Kawahara et al., 2013).
The problem of uncertainty was smaller for the *A. thaliana* datasets since the marker density was much higher. There the xQTLs were about 1 kb in size, which is actually the region between marker with best association in GWAS ± chromosomal marker density. In theory, it is possibly that the regions are much larger since previous and next markers can be significantly associated with the trait as well, again adding uncertainty to the selection of causal genes. Due to the GWAS mapping procedure based on the linear mixed models LIMIX (Lippert et al., 2014) and EMMAX (Kang et al., 2008) that were used, only the best association is reported and therefore the region was extended with the chromosomal marker density in this study. Nevertheless, the links between BGCs and xQTLs for *A. thaliana* can be hypothesized with more certainty due to these smaller xQTL regions, but more importantly due to the extensive research that is already done in thale cress. However, these smaller xQTL regions caused problems in the colocation network analysis. The size of the regions might span one gene, but they won't span an intergenic region between two genes. This causes limitations in the colocation network analysis, in this way it is difficult to discover novel BGCs since the xQTL regions are too small. A solution would be to extend the xQTL regions by the actual distance to the previous and next marker, maybe even the 2 markers before and after. Another problem in the colocation network analysis was that there was little overlap found between mQTLs and eQTLs, which also makes it difficult to discover novel BGCs. An explanation could be that the QTL analyses for *A. thaliana* were performed on two different sets of accessions. It might be that the metabolites are present in one set, and the genes are expressed in the other set, causing missing links between mQTLs and eQTLs in the network.
A last limitation, shared by both datasets, is that the (*in silico*) LC-MS and structural databases for secondary metabolites are sparse and therefore the annotations of the detected masses in the *O. sativa* and *A. thaliana* MS analyses are uncertain or even unknown. A solution might be to use the substructure exploration tool MS2LDA (Van der Hooft et al., 2017) for metabolomics data to identify substructures in the unknown mass spectra, this might then aid in linking BGCs to their products, if enzymes in the BGCs are known to act on certain molecular subgroups and the latter are present in the compound's mQTL that overlapped with the BGC.
Despite all uncertainties, methods were developed in this study that might successfully integrate BGCs with xQTLs in both RIL population-based studies and GWASs in plants. These methods can be used *in silico* to generate hypotheses to design targeted lab experiments for validation. The examples in the **Results** section can be validated by targeted knock-outs of genes or with near isogenic lines to assess the causality of these genes in the variation of gene expression or metabolite levels. Fine-mapping of the xQTL datasets could lead to higher resolutions, whereas improvements in mass spectrometry (data analysis) can lead to more efficient natural product discovery. Furthermore, the QTL studies can be optimized for integration with BGC prediction. Growth conditions can include abiotic (e.g. rhizosphere soil composition, temperature, light conditions, humidity) and/or biotic stress (e.g. rhizobiome composition, insect or herbivore exposure) depending on the use of the secondary metabolites for which a BGC wants to be found. The tissue that is used for the expression and metabolite profiling analyses in these QTL studies is

important too, not all secondary metabolites and their biosynthesis genes are produced/expressed in every tissue. These improvements together would make the integration methods more efficient. Besides the two described methods, the integration method can be extended by looking for genomic overlap of trans-eQTLs of genes within BGCs. This can potentially lead to discovery of regulators for the BGCs. Another option is to use BGCs predicted with PlantClusterFinder (Schläpfer et al., 2017) as one more data source. Furthermore, the integration methods need to be optimized for speed, efficiency and applicability in other plants and maybe even Bacteria and fungi by using AntiSMASH (Weber et al., 2015) instead of PlantiSMASH together with GWASs.

Acknowledgements

I want to thank the Bioinformatics group and the Laboratory of Plant Physiology from Wageningen University & Research for insights, supervision and data. More specifically: dr. Marnix Medema, dr. Harm Nijveen, dr. Justin van der Hooft, dr. Aalt-Jan van Dijk, Satria Kautsar MSc., Hernando Suarez Duran MSc., dr. ing. Joost Keurentjes, dr. Rik Kooke, dr. Ric de Vos, dr. ir. Wilco Ligterink, and dr. ir. Guusje Bonnema.

References

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., … Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics, 25(1), 25–29. http://doi.org/10.1038/75556

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289-300. Retrieved from http://www.jstor.org/stable/2346101

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., … Schwede, T. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Research, 42(Web Server issue), W252–W258. http://doi.org/10.1093/nar/gku340

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. http://doi.org/10.1093/bioinformatics/btu170

Bonferroni, C. E. (1936). 1936 Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commerciali Di Firenze, 8, 3–62.

Broman, K. W., Wu, H., Sen, Ś., & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics, 19(7), 889–890. https://doi.org/10.1093/bioinformatics/btg112

Carreno-Quintero, N., Bouwmeester, H. J., & Keurentjes, J. J. B. (2013, January). Genetic analysis of metabolome-phenotype interactions: From model to crop species. Trends in Genetics. https://doi.org/10.1016/j.tig.2012.09.006

Collins, D. W., & Jukes, T. H. (1994). Rates of transition and transversion in coding sequences since the human- Rodent divergence. Genomics, 20(3), 386–396. https://doi.org/10.1006/geno.1994.1192

Goujon, T., Sibout, R., Pollet, B., Maba, B., Nussaume, L., Bechtold, N., … Jouanin, L. (2003). A new Arabidopsis thaliana mutant deficient in the expression of O-methyltransferase impacts lignins and sinapoyl esters. Plant Molecular Biology, 51(6), 973–989. https://doi.org/10.1023/A:1023022825098

Gong, L., Chen, W., Gao, Y., Liu, X., Zhang, H., Xu, C., … Luo, J. (2013). Genetic analysis of the metabolome exemplified using a rice population. Proceedings of the National Academy of Sciences, 110(50), 20320–20325. https://doi.org/10.1073/pnas.1319681110

Ishihara, H., Tohge, T., Viehöver, P., Fernie, A. R., Weisshaar, B., & Stracke, R. (2016). Natural variation in flavonol accumulation in Arabidopsis is determined by the flavonol glucosyltransferase BGLU6. Journal of

Experimental Botany, 67(5), 1505–1517. http://doi.org/10.1093/jxb/erv546

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. Genetics, 178(3), 1709–1723. https://doi.org/10.1534/genetics.107.080101

Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). PlantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Research, 45(W1), W55–W63. http://doi.org/10.1093/nar/gkx305

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., … Matsumoto, T. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice, 6(1), 1–10. https://doi.org/10.1186/1939-8433-6-1

Kawakatsu, T., Huang, S. shan Carol, Jupe, F., Sasaki, E., Schmitz, R. J. J., Urich, M. A. A., … Ecker, J. R. (2016). Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell, 166(2), 492–506. https://doi.org/10.1016/j.cell.2016.06.044

Kim, C., & Apel, K. (2012). Arabidopsis light-dependent NADPH: Protochlorophyllide oxidoreductase A (PORA) is essential for normal plant growth and development: An addendum. Plant Molecular Biology, 80(2), 237–240. https://doi.org/10.1007/s11103-012-9944-8

Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. Nucleic Acids Res. 2016 Jan 4; 44(D1):D1202-13. Epub 2015 Sep 22 [PubMed PMID: 26400175] http://doi.org/10.1093/nar/gkv951

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., … Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database: The Journal of Biological Databases and Curation, 2011, bar030. http://doi.org/10.1093/database/bar030

Kleine, T. (2012). Arabidopsis thaliana mTERF proteins: evolution and functional classification. Frontiers in Plant Science, 3, 233. http://doi.org/10.3389/fpls.2012.00233

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. http://doi.org/10.1038/nmeth.1923

Leinonen, R., Sugawara, H., & on behalf of the International Nucleotide Sequence Database Collaboration, M. (2011). The Sequence Read Archive. Nucleic Acids Research, 39(Database issue), D19–D21. http://doi.org/10.1093/nar/gkq1019

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. http://doi.org/10.1093/bioinformatics/btp352

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. Bioinformatics 28(18), 2397–2399, https://doi.org/10.1093/bioinformatics/bts444

Lippert, C., Casale, F. P., Rakitsch, B., & Stegle, O. (2014). LIMIX: genetic analysis of multiple traits. BioRxiv, 003905. https://doi.org/10.1101/003905

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., … Glöckner, F. O. (2015). Minimum Information about a Biosynthetic Gene cluster. Nature Chemical Biology, 11(9), 625–631. http://doi.org/10.1038/nchembio.1890

Nakamura, K., Shimura, N., Otabe, Y., Hirai-Morita, A., Nakamura, Y., Ono, N., … Kanaya, S. (2013). KNApSAcK-3D: A three-dimensional structure database of plant metabolites. Plant and Cell Physiology, 54(2). https://doi.org/10.1093/pcp/pcs186

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics, 32(11), 1749–1751. http://doi.org/10.1093/bioinformatics/btw044

Nützmann, H. W., Huang, A., & Osbourn, A. (2016). Plant metabolic clusters – from genetics to genomics. New Phytologist, 211(3), 771–789. https://doi.org/10.1111/nph.13981

Ooka, H., Satoh, K., Doi, K., Nagata, T., Otomo, Y., Murakami, K., … Kikuchi, S. (2003). Comprehensive Analysis of NAC Family Genes in Oryza sativa and Arabidopsis thaliana. DNA Research, 10(6), 239–247. https://doi.org/10.1093/dnares/10.6.239

Pfalz, M., Mikkelsen, M. D., Bednarek, P., Olsen, C. E., Halkier, B. A., & Kroymann, J. (2011). Metabolic Engineering in Nicotiana benthamiana Reveals Key Enzyme Functions in Arabidopsis Indole Glucosinolate Modification. The Plant Cell, 23(2), 716–729. http://doi.org/10.1105/tpc.110.081711

Pickel, B., & Schaller, A. (2013). Dirigent proteins: Molecular characteristics and potential biotechnological applications. Applied Microbiology and Biotechnology. https://doi.org/10.1007/s00253-013-5167-4

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rowe, H. C., Hansen, B. G., Halkier, B. A., & Kliebenstein, D. J. (2008). Biochemical Networks and Epistasis Shape the Arabidopsis thaliana Metabolome. THE PLANT CELL ONLINE, 20(5), 1199–1216. https://doi.org/10.1105/tpc.108.058131

Sawada, Y., Nakabayashi, R., Yamada, Y., Suzuki, M., Sato, M., Sakata, A., Akiyama, K., Sakurai, T., Matsuda, F., Aoki, T., Hirai, M., Saito, K. (2012). RIKEN tandem mass spectral database (ReSpect) for phytochemicals: A plant-specific MS/MS-based data resource and database. Phytochemistry, 82, 38-45. https://doi.org/10.1016/j.phytochem.2012.07.007.

Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., … Rhee, S. Y. (2017). Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. Plant Physiology, 173(4), 2041–2059. http://doi.org/10.1104/pp.16.01942

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research, 13(11), 2498–2504. http://doi.org/10.1101/gr.1239303

Sotelo, T., Velasco, P., Soengas, P., Rodríguez, V. M., & Cartea, M. E. (2016). Modification of Leaf Glucosinolate Contents in Brassica oleracea by Divergent Selection and Effect on Expression of Genes Controlling Glucosinolate Pathway. Frontiers in Plant Science, 7, 1012. http://doi.org/10.3389/fpls.2016.01012

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., … Huala, E. (2008). The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. Nucleic Acids Research, 36(SUPPL. 1). https://doi.org/10.1093/nar/gkm965

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, 14(2), 178–192. http://doi.org/10.1093/bib/bbs017

Tikunov, Y. M., Laptenok, S., Hall, R. D., Bovy, A., & de Vos, R. C. H. (2012). MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. Metabolomics, 8(4), 714–718. http://doi.org/10.1007/s11306-011-0368-2

Van der Hooft, J. J. J., Wandy, J., Young, F., Padmanabhan, S., Gerasimidis, K., Burgess, K. E. V., … Rogers, S. (2017). Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics. Analytical Chemistry, 89(14), 7569–7577. http://doi.org/10.1021/acs.analchem.7b01391

Wang, J., Yu, H., Weng, X., Xie, W., Xu, C., Li, X., … Zhang, Q. (2014). An expression quantitative trait loci-guided co-expression analysis for constructing regulatory network using a rice recombinant inbred line population. Journal of Experimental Botany, 65(4), 1069–1079. https://doi.org/10.1093/jxb/ert464

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., … Medema, M. H. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research, 43(Web Server issue), W237–W243. http://doi.org/10.1093/nar/gkv437

Wentzell, A. M., Rowe, H. C., Hansen, B. G., Ticconi, C., Halkier, B. A., & Kliebenstein, D. J. (2007). Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genetics,

3(9), 1687–1701. https://doi.org/10.1371/journal.pgen.0030162

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

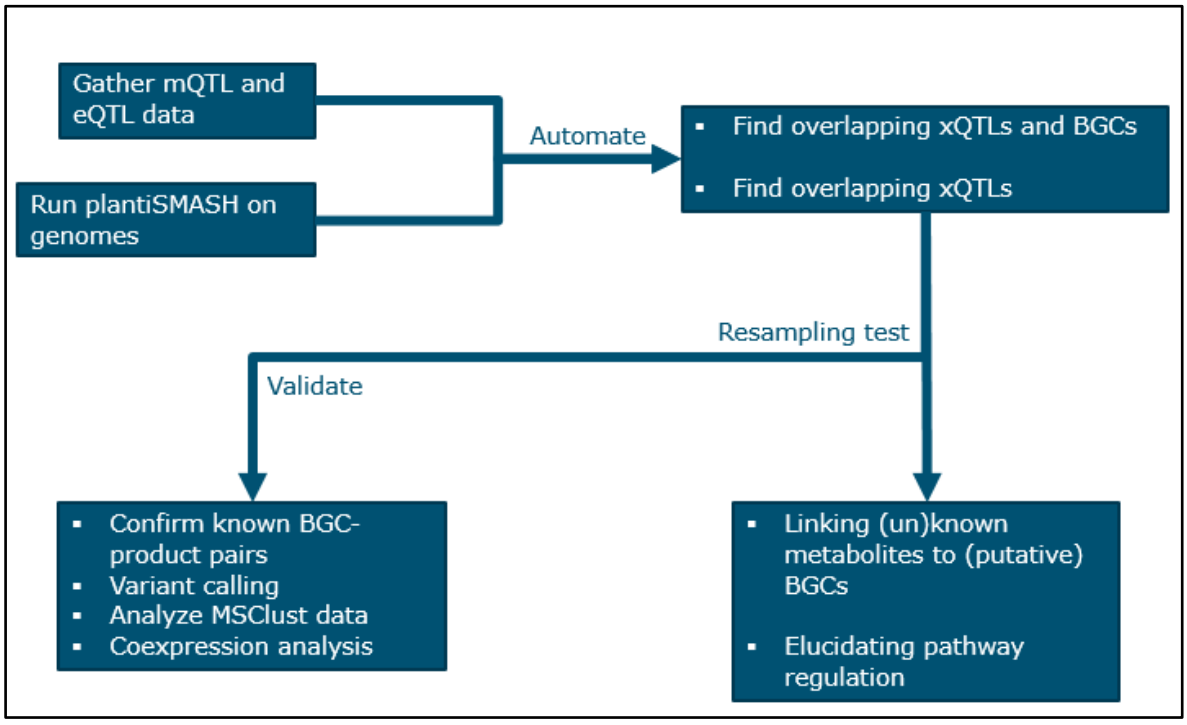
Xiang, T., Shibuya, M., Katsube, Y., Tsutsumi, T., Otsuka, M., Zhang, H., … Ebizuka, Y. (2006). A new triterpene synthase from Arabidopsis thaliana produces a tricyclic triterpene with two hydroxyl groups. Organic Letters, 8(13), 2835–2838. https://doi.org/10.1021/ol060973p

Zhang, C., Freddolino, P. L., & Zhang, Y. (2017). COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic Acids Research, 45(Web Server issue), W291–W299. http://doi.org/10.1093/nar/gkx366

Zhang, H., Wang, J., & Goodman, H. M. (1997). An Arabidopsis gene encoding a putative 14-3-3-interacting protein, caffeic acid/5-hydroxyferulic acid o-methyltransferase. Biochimica et Biophysica Acta - Gene Structure and Expression, 1353(3), 199–202. https://doi.org/10.1016/S0167-4781(97)00096-1

Appendices

1. Workflow



Supplemental figure 1 General workflow of this study.

2. Commands

Running PlantiSMASH on *O. sativa*

```
python run_antismash.py --taxon plants --gff3 oryza_sativa_indica_MSUv6.1.gff3
--coexpress --coexpress-csv_file GSE49020_matrix.csv --clusterblast --
knownclusterblast --min-domain-number 1 --cdh-cutoff 0.6 --outputfolder
<some_name> oryza_sativa_indica_MSUv6.1.fa
```

Running PlantiSMASH on *A. thaliana*

```
python run_antismash.py --taxon plants --clusterblast --knownclusterblast --
min-domain-number 1 --cdh-cutoff 0.6 --outputfolder <some_name>
arabidopsis_thaliana_TAIR10.gbff
```

SNP calling

Retrieve MH63 data, Illumina HiSeq 2000 paired-end reads with 3 insert sizes

```
fastq-dump –I –split-files SRR3234371        (10kb insert-size)
fastq-dump –I –split-files SRR3234370        (5kb insert size)
fastq-dump –I –split-files SRR3234369        (300bp insert size)
```

Retrieve ZS97 data, Illumina HiSeq 2000 paired-end reads with 3 insert sizes

```
fastq-dump –I –split-files SRR3234374        (10kb insert-size)
fastq-dump –I –split-files SRR3234373        (5kb insert size)
fastq-dump –I –split-files SRR3234372        (300bp insert size)
```

Build an index for the Oryza sativa reference genome (MSUv6.1) with Bowtie2

```
bowtie2-build -f oryza_sativa_indica_MSUv6.1.fa MSUv6.1
```

Trim reads of MH63 and ZS97

```
java -jar trimmomatic-0.36.jar PE -phred33 SRR3234369_1.fastq
SRR3234369_2.fastq SRR3234369_1_trimmed_paired.fastq
SRR3234369_1_trimmed_unpaired.fastq SRR3234369_2_trimmed_paired.fastq
SRR3234369_2_trimmed_unpaired.fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

java -jar trimmomatic-0.36.jar PE -phred33 SRR3234370_1.fastq
SRR3234370_2.fastq SRR3234370_1_trimmed_paired.fastq
SRR3234370_1_trimmed_unpaired.fastq SRR3234370_2_trimmed_paired.fastq
SRR3234370_2_trimmed_unpaired.fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

java -jar trimmomatic-0.36.jar PE -phred33 SRR3234371_1.fastq
SRR3234371_2.fastq SRR3234371_1_trimmed_paired.fastq
SRR3234371_1_trimmed_unpaired.fastq SRR3234371_2_trimmed_paired.fastq
SRR3234371_2_trimmed_unpaired.fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

java -jar trimmomatic-0.36.jar PE -phred33 SRR3234372_1.fastq
SRR3234372_2.fastq SRR3234372_1_trimmed_paired.fastq
SRR3234372_1_trimmed_unpaired.fastq SRR3234372_2_trimmed_paired.fastq
SRR3234372_2_trimmed_unpaired.fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

java -jar trimmomatic-0.36.jar PE -phred33 SRR3234373_1.fastq
SRR3234373_2.fastq SRR3234373_1_trimmed_paired.fastq
SRR3234373_1_trimmed_unpaired.fastq SRR3234373_2_trimmed_paired.fastq
SRR3234373_2_trimmed_unpaired.fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

java -jar trimmomatic-0.36.jar PE -phred33 SRR3234374_1.fastq
SRR3234374_2.fastq SRR3234374_1_trimmed_paired.fastq
SRR3234374_1_trimmed_unpaired.fastq SRR3234374_2_trimmed_paired.fastq
SRR3234374_2_trimmed_unpaired.fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Map reads of MH63 and ZS97 to the MSUv6.1 reference genome with Bowtie2

```
bowtie2 -x MSUv6.1 -1 SRR3234369_1_trimmed_paired.fastq,
SRR3234370_1_trimmed_paired.fastq, SRR3234371_1_trimmed_paired.fastq -2
SRR3234369_2_trimmed_paired.fastq, SRR3234370_2_trimmed_paired.fastq,
SRR3234371_2_trimmed_paired.fastq -S MH63.sam -I 100 -X 10000 --phred33 --
sensitive --threads 8 --met-file MH63_metrics.txt

bowtie2 -x MSUv6.1 -1 SRR3234372_1_trimmed_paired.fastq,
SRR3234373_1_trimmed_paired.fastq , SRR3234374_1_trimmed_paired.fastq -2
SRR3234372_2_trimmed_paired.fastq, SRR3234373_2_trimmed_paired.fastq,
SRR3234374_2_trimmed_paired.fastq -S ZS97.sam -I 100 -X 10000 --phred33 --
sensitive --threads 8 --met-file ZS97_metrics.txt
```

`--sensitive` is the same as the following individuals settings: `-D 15 -R 2 -L 22 -i S,1,1.15`

Conversion of SAM files to BAM files

```
samtools view -bS -@ 8 MH63.sam > MH63.bam
samtools view -bS -@ 8 ZS97.sam > ZS97.bam
```

Sorting BAM files

```
samtools sort -@ 8 MH63.bam -o MH63_sorted.bam
samtools sort -@ 8 ZS97.bam -o ZS97_sorted.bam
```

Indexing BAM files

```
samtools index -@ 10 MH63_sorted.bam
samtools index -@ 10 ZS97_sorted.bam
```

Variant calling with SAMtools mpileup and BCFtools call

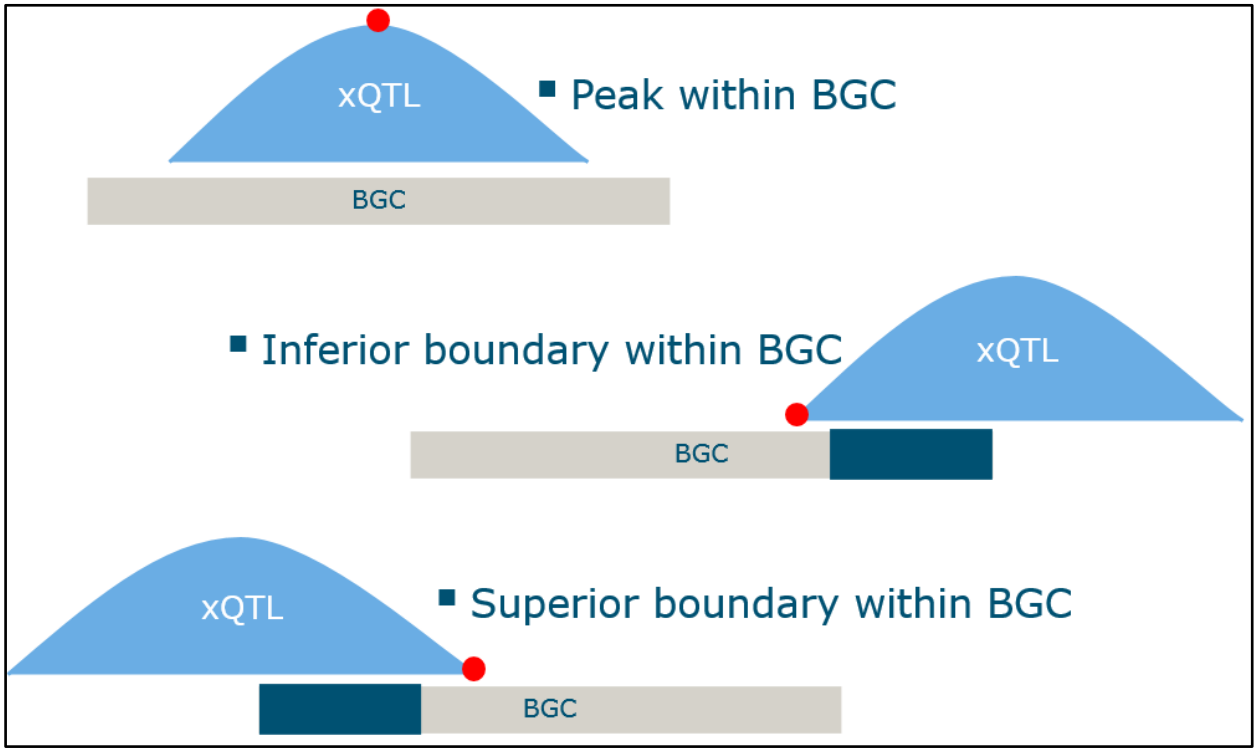
```
samtools mpileup -f oryza_sativa_indica_MSUv6.1.fa -g MH63_sorted.bam
ZS97_sorted.bam -o MH63_ZS97.bcf

bcftools call -m -Ov MH63_ZS97.bcf -o MH63_ZS97.vcf
bcftools stats MH63_ZS97.vcf
```

3. Input data format xQTLs

BGIOSGA029658	1	0.2825	0	0.565	52.26727907	Choline	7	17.0	16.2	17.0	11.7
BGIOSGA002596	1	0.461	0	0.922	12.70386823	Choline	7	28.6	28.4	29.7	5.0
BGIOSGA002596	1	0.2995	0	0.599	23.83486275	Choline	4	15.8	13.3	15.9	4.5
BGIOSGA002586	1	0.2995	0	0.599	16.02618613	Choline	4	21.8	21.3	21.8	3.5
LOC_Os01g01689	1	0.2995	0	0.599	27.62156402	L-Serine	6	7.6	7.1	9.7	5.5
LOC_Os01g01800	1	0.2825	0	0.565	105.1423	L-Serine	1	3.7	3.0	4.3	4.7
BGIOSGA002525	1	0.6105	0	1.221	8.661246234	L-Serine	4	35.3	33.5	35.3	3.1
LOC_Os01g01700	1	0.6105	0	1.221	5.405967275	L-Proline	1	6.3	5.3	8.4	4.8
LOC_Os01g01010	1	0.2825	0	0.565	46.96352091	N-Hydroxysuccinimide	1	6.3	5.3	6.6	4.1
LOC_Os01g01312	1	0.2995	0	0.599	42.92679436	L-Valine	10	3.5	2.6	5.3	6.2
BGIOSGA002572	1	0.2995	0	0.599	11.259097	L-Valine	6	9.4	7.3	9.7	3.7
BGIOSGA002573	1	0.461	0	0.922	15.7081864	L-Threonine	1	6.3	5.3	8.4	3.5
BGIOSGA002618	1	0.461	0	0.922	44.60994433	L-Threonine	4	27.9	24.4	29.5	3.2
LOC_Os01g01720	1	0.2995	0	0.599	27.78506751	Benzamidine	2	12.0	10.8	15.9	5.0
BGIOSGA002583	1	0.5375	0	1.075	13.03836186	Benzamidine	4	27.9	24.4	29.5	3.0
BGIOSGA002571	1	0.2825	0	0.565	58.38060568	m0012-L_122.0269_6.94	12	14.6	10.8	16.3	36.1
BGIOSGA002589	1	0.461	0	0.922	23.89472723	m0012-L_122.0269_6.94	6	2.3	1.8	2.3	7.6
LOC_Os01g01780	1	0.461	0	0.922	11.41736864	m0012-L_122.0269_6.94	7	29.7	28.1	29.7	3.6
LOC_Os01g01780	1	0.2825	0	0.565	32.839435	L-Leucine	10	5.3	2.6	5.5	5.1

Supplemental figure 2 Examples of the input data format of xQTLs. The left side of the figure shows eQTLs of O. sativa. The right side of the figure shows mQTLs of O. sativa. For each xQTL the following elements are reported: gene or metabolite name, chromosome number, peak location in Mb, inferior boundary location in Mb, superior boundary location in Mb, and the LOD-score.

4. Overlap rules



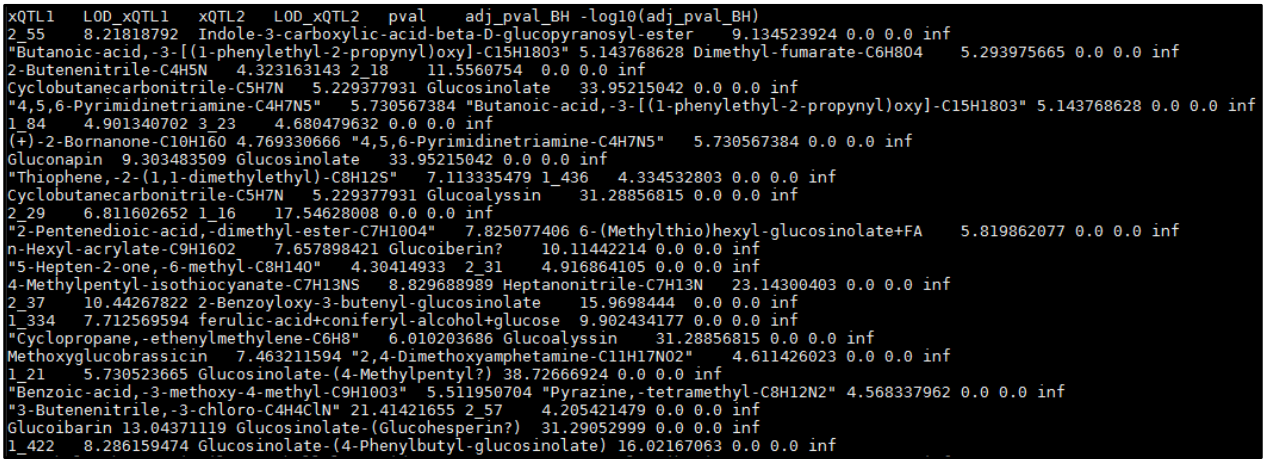
Supplemental figure 3 Overlap rules for detecting of genomic overlap between xQTLs and BGCs, however these rules also apply for finding overlap between xQTLs independent of BGCs in which the BGC in this figure is another xQTL. There are three possible ways of overlap as shown. The blue areas in the BGCs represent the minimum overlap size that is required. This minimum overlap is calculated as a percentage of the total BGC size.

5. Output data format xQTLs and BGCs overlap

```
#clusterID  cluster type    chromosome  BGC start bp    BGC end bp
33  lignan;saccharide   7   25423186    25502862

#xQTL   p-value adjusted p-value    LOD-score   locus annotation    locus start bp  locus end bp    locus status
LOC_Os05g13520  0.001   0.03    9.025   white-Dbrown complex homolog protein 12, putative, expressed    7492627 7498325 distant
LOC_Os01g57430  0.001   0.03    5.379   hypothetical protein    33193040    33193358    distant
m0536-L_NaN_9.29    0.002   0.0158  5.1
m0744-L_NaN_10.6    0.001   0.009369    3.4
Tricin-O-rutinoside 0.01    0.03    9.4
m0812-L_NaN_10.15   0.003   0.02232 5.6
m0868-L_NaN_23.3    0.0 0.0 3.5
Tricin-O-rhamnosyl-O-malonylhexoside    0.001   0.009338    4.2
m0942-L_NaN_9.37    0.0 0.0 5.2
Sucrose 0.004   0.02856 4.9
LPC(1-acyl-18:2)    0.003   0.0222  3.5
LPC(1-acyl-18:2)    0.003   0.0222  3.7
```

Supplemental figure 4 Example of the output data format of overlap between xQTLs and BGCs. A tab-separated text file as presented in this figure is generated for every BGC for which overlap is found. The top of the file contains information about the BGC: clusterID, cluster type, chromosome number, BGC start position in bp and BGC end position in bp. The bottom of the file contains information about the xQTLs that are overlapping with the BGCs: gene/metabolite name, p-value, adjusted p-value, LOD-score, locus annotation (eQTL), locus start position in bp (eQTL), locus end position in bp (eQTL) and locus status (eQTL).

6. Output data format overlapping mQTLs in *A. thaliana*



```
xQTL1   LOD_xQTL1   xQTL2   LOD_xQTL2   pval    adj_pval_BH -log10(adj_pval_BH)
2_55    8.21818792  Indole-3-carboxylic-acid-beta-D-glucopyranosyl-ester    9.134523924 0.0 0.0 inf
"Butanoic-acid,-3-[(1-phenylethyl-2-propynyl)oxy]-C15H18O3" 5.143768628 Dimethyl-fumarate-C6H8O4    5.293975665 0.0 0.0 inf
2-Butenenitrile-C4H5N   4.323163143 2_18    11.5560754  0.0 0.0 inf
Cyclobutanecarbonitrile-C5H7N   5.229377931 Glucosinolate   33.95215042 0.0 0.0 inf
"4,5,6-Pyrimidinetriamine-C4H7N5"   5.730567384 "Butanoic-acid,-3-[(1-phenylethyl-2-propynyl)oxy]-C15H18O3" 5.143768628 0.0 0.0 inf
1_84    4.901340702 3_23    4.680479632 0.0 0.0 inf
(+)-2-Bornanone-C10H16O 4.769330666 "4,5,6-Pyrimidinetriamine-C4H7N5"   5.730567384 0.0 0.0 inf
Gluconapin  9.303483509 Glucosinolate   33.95215042 0.0 0.0 inf
"Thiophene,-2-(1,1-dimethylethyl)-C8H12S"   7.113335479 1_436   4.334532803 0.0 0.0 inf
Cyclobutanecarbonitrile-C5H7N   5.229377931 Glucoalyssin    31.28856815 0.0 0.0 inf
2_29    6.811602652 1_16    17.54628008 0.0 0.0 inf
"2-Pentenedioic-acid,-dimethyl-ester-C7H10O4"   7.825077406 6-(Methylthio)hexyl-glucosinolate+FA    5.819862077 0.0 0.0 inf
n-Hexyl-acrylate-C9H16O2    7.657898421 Glucoiberin?    10.11442214 0.0 0.0 inf
"5-Hepten-2-one,-6-methyl-C8H14O"   4.30414933  2_31    4.916864105 0.0 0.0 inf
4-Methylpentyl-isothiocyanate-C7H13NS   8.829688989 Heptanonitrile-C7H13N   23.14300403 0.0 0.0 inf
2_37    10.44267822 2-Benzoyloxy-3-butenyl-glucosinolate    15.9698444  0.0 0.0 inf
1_334   7.712569594 ferulic-acid+coniferyl-alcohol+glucose  9.902434177 0.0 0.0 inf
"Cyclopropane,-ethenylmethylene-C6H8"   6.010203686 Glucoalyssin    31.28856815 0.0 0.0 inf
Methoxyglucobrassicin   7.463211594 "2,4-Dimethoxyamphetamine-C11H17NO2"    4.611426023 0.0 0.0 inf
1_21    5.730523665 Glucosinolate-(4-Methylpentyl?) 38.72666924 0.0 0.0 inf
"Benzoic-acid,-3-methoxy-4-methyl-C9H10O3"  5.511950704 "Pyrazine,-tetramethyl-C8H12N2" 4.568337962 0.0 0.0 inf
"3-Butenenitrile,-3-chloro-C4H4ClN" 21.41421655 2_57    4.205421479 0.0 0.0 inf
Glucoibarin 13.04371119 Glucosinolate-(Glucohesperin?)  31.29052999 0.0 0.0 inf
1_422   8.286159474 Glucosinolate-(4-Phenylbutyl-glucosinolate) 16.02167063 0.0 0.0 inf
```

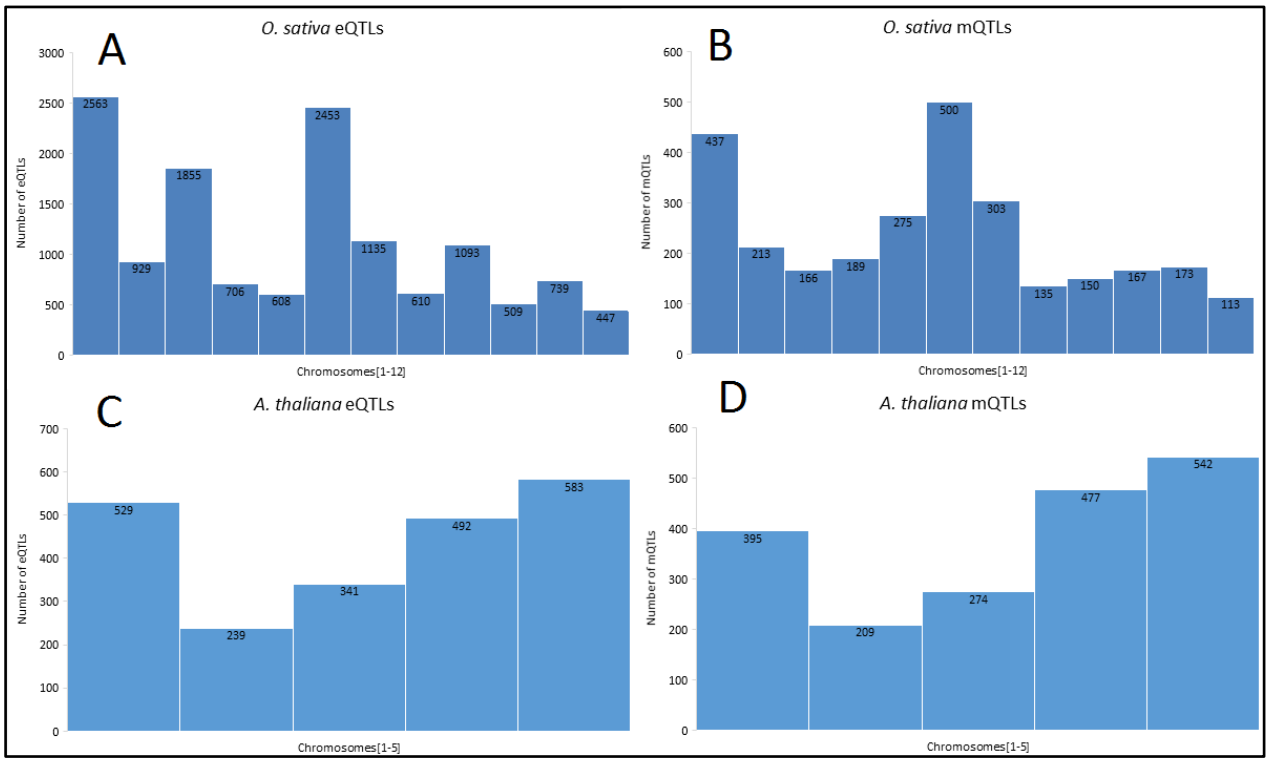
Supplemental figure 5 Example of the output data format of overlapping mQTLs. A tab-separated text file as presented in this figure is generated. The headers of the file are: xQTL1, LOD_xQTL1, xQTL2, LOD_xQTL2, pval, adj_pval_BH and -log10(adj_pval_BH). xQTL1 and xQTL2 are the mQTLs and their LOD-scores in the QTL analysis (LOD_xQTL1, LOD_xQTL2) that showed locational overlap. The significance of the overlap was tested with a randomization test, which results in p-values (pval). The p-values were correct with Benjamini-Hochberg (adj_pval_BH) and -log10 transformed (-log10(adj_pval_BH).

7. Detailed overview of BGCs with known products in *O. sativa* and *A. thaliana*

Supplemental table 1 A detailed overview of BGCs with known products in *O. sativa* and A. thaliana. The biosynthetic class, molecular formula, average molecular mass (Da) and PubChem ID are given for the BGC's products. The chromosomal location and the MiBiG accessions are given for the BGCs.

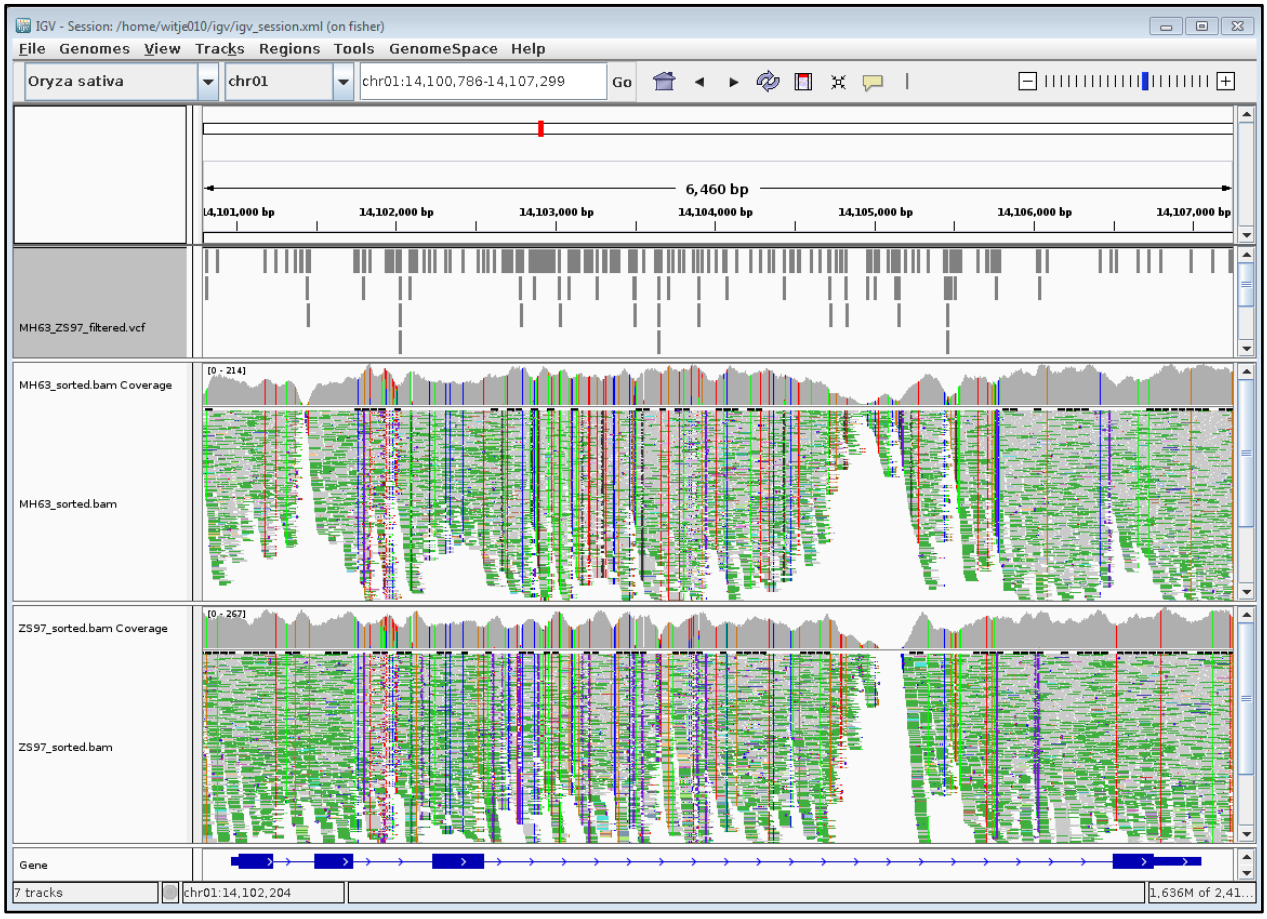
	Biosynthetic class	**Chromosomal location (Mb)**	**Molecular formula**	**Average molecular mass (Da)**	**MiBiG accession**	**PubChem ID**
O. sativa						
Phytocassane/oryzalide	terpene	2: 22.52-22.76	$C_{20}H_{28}O_3$	316.43452	BGC0000672	10313699
Momilactone	terpene	4: 5.31-5.58	$C_{20}H_{26}O_3$	314.41864	BGC0000671	162644
A. thaliana						
Arabidiol/baruol	terpene	4: 8.73-8.82	NA	NA	BGC0001313	NA
Tirucalla	terpene	5: 14.19-14.25	$C_{30}H_{50}O$	426.7174	BGC0001314	12302184
Marneral	terpene	5: 17.02-17.06	$C_{30}H_{50}O$	426.7174	BGC0000669	25001002
Thalianol	terpene	5: 19.43-19.46	$C_{30}H_{50}O$	426.7174	BGC0000670	25229600

8. Distribution of xQTLs throughout the genomes of *O. sativa* and *A. thaliana*



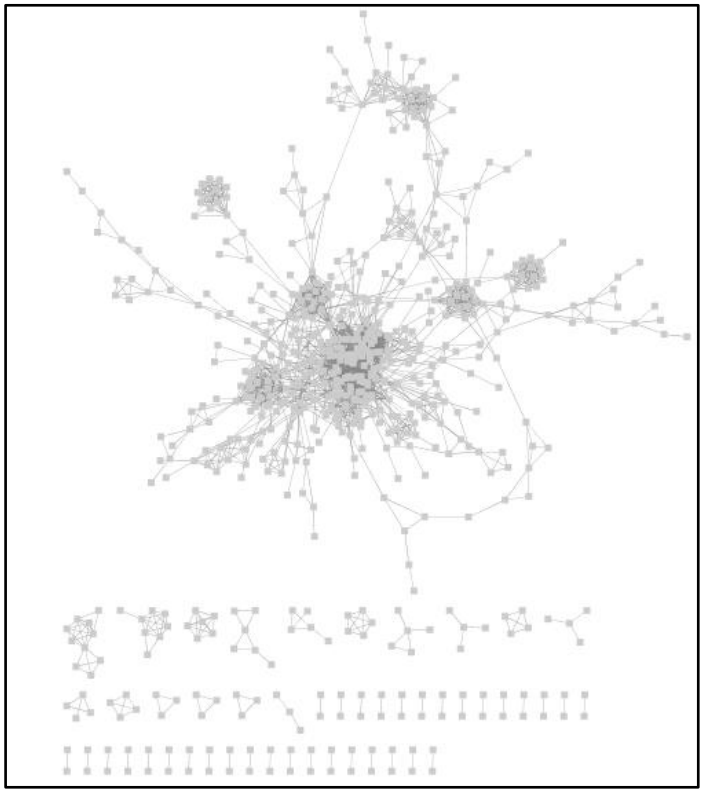
Supplemental figure 6 Distributions of xQTLs for the genomes of *O. sativa* and *A. thaliana*. **A.** Distribution of eQTLs in rice. **B.** Distribution of mQTLs in rice. **C.** Distribution of eQTLs in thale cress. **D.** Distribution of mQTLs in thale cress.

9. Example of variant browsing with IGV in *O. sativa*



Supplemental figure 7 An example of variant browsing with IGV in *O. sativa*. The gene shown is LOC_Os01g25010, which is a dioxygenase with three non-synonymous SNPs in the DIOX_N domain as was presented in the **Results** section. The sorted BAM files of both parents are shown as well. The top track presents the variants between the two variants in a VCF file, filtered means here that only variants were kept (ALT != ".").

10. Overview colocation network for overlapping mQTLs and eQTLs in *A. thaliana*



Supplemental figure 8 An overview of the entire colocation network of overlapping mQTLs and eQTLs in *A. thaliana*. Within the large network are some smaller hubs. Outside the large network (bottom of figure) are some smaller networks. Nodes represent mQTLs and edges represent genomic overlap. The edge width is determined by the -log10 of the BH adjusted p-values derived from the permutation test. The network was created with Cytoscape v3.4.0. (*Shannon et al., 2003*).