Integrating Bioinformatic Approaches to Promote Crop Resilience

Chenming Cui

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 the
School of Plant and Environmental Sciences

David C. Haak, Committee Chair

Aureliano Bombarely Gomez
Song Li
Boris A. Vinatzer

May 31th 2019
Blacksburg, Virginia

Keywords: transcriptome, machine learning, stress tolerance, Nanopore, genome

Integrating Bioinformatic Approaches to Promote Crop Resilience

Chenming Cui

ABSTRACT

Even under the best management strategies contemporary crops face yield losses from diverse threats such as, pathogens, pests, and environmental stress. Adding to this management challenge is that under current global climate projections these impacts are predicted to become even greater. Natural genetic variation, long used by traditional plant breeders, holds great promise for adapting high performing agronomic lines to these stressors. Yet, efforts to bolster crop plant resilience using wild relatives have been hindered by time consuming efforts to develop genomic tools and/or identify the genetic basis for agronomic traits. Thus, increasing crop plant resilience requires developing and deploying approaches that leverage current high-throughput sequencing technologies to more rapidly and robustly develop genomic tools in these systems. Here we report the integration of bioinformatic and statistical tools to leverage high-throughput sequencing to 1) develop a machine learning approach to determine factors impacting transcriptome assembly and quantitatively evaluate transcriptome completeness, 2) dissect complex physiological pathway interactions in *Solanum pimpinellifolium* under combined stresses—using comparative transcriptomics, and 3) develop a genome assembly pipeline that can be deployed to rapidly assemble a more contiguous genome, unraveling previously hidden complexity, using *Phytopthora capsici* as a model. As a result, we have generated strategic guidelines for transcriptome assembly and developed an orthologue and reference free, machine learning based tool “WWMT” to quantitatively score

transcriptome completeness from short read data. Secondly, we identified “hub genes” and describe genes involved with “cross-talk” between drought and herbivore stress response pathways. Finally, we demonstrate a protocol for combining long-read sequencing from the Oxford Nanopore Technologies MinION, and short-read data, to rapidly assembly a cost-effective, contiguous and relatively complete genome. Here we uncovered hidden variation in a well-known plant pathogen finding that the genome was 92% bigger than previous estimates with more than 39% of duplicated regions, supporting a hypothesized recent whole genome duplication in this clade. This community resource will support new functional and evolutionary studies in this economically important pathogen.

Integrating Bioinformatic Approaches to Promote Crop Resilience

Chenming Cui

GENERAL AUDIENCE ABSTRACT

Meeting the food production demands of a burgeoning population in a changing environment, means adapting crop plants to become more resilient to environmental stress. One of the greatest barriers to understanding and predicting crop responses to future environmental change is our poor understanding of the functional and genomic basis of stress resistance traits for contemporary crops. This impediment presents a barrier for rapid crop improvement technologies, such as, gene editing or genomic selection, that is only partially overcome by generating large amounts of sequencing data. Here we need tools that allow us to process and evaluate huge amounts of data generated from next generation sequencing studies to help identify genomic regions associated with agronomic traits. We also need technical approaches that allow us to disentangle the complex genetic interactions that drive plant stress responses. Here we present work that used statistical analysis and recent advances of artificial intelligence to develop a bioinformatic approach to evaluate genomic sequencing data prior to downstream analyses. Secondly, we used a reductionist approach to filter thousands of genes to key genes associated with combined stress responses (herbivory and drought), in the most widely used vegetable in the world, tomato. Finally, we developed a method for generating whole genome sequences that is low-cost and time sensitive and tested it using a well-known plant pathogen genome, wherein we unraveled significant hidden complexity. Overall this work provides community-wide genomic tools and information to promote crop resilience.

DEDICATION

I dedicate this dissertation to my beloved parents and family.

To myself, for my boldness and perseverance in the past ten years, since my sophomore year to now.

ACKNOWLEDGEMENTS

Throughout my journey of pursuing a doctorate degree at Virginia Tech, I have received a large amount of support from mentors, family and friends. First of all, I would like to express my gratitude to my advisor, Dr. David Haak for his incredible advocacy and training during my Ph.D.. Aside from his professional training, his generosity, respect, and decency substantially impacted me and inspired me to be a better person in my life.

I appreciate my Ph.D. advisory committee members, Drs. Aureliano Bombarely, Song Li and Boris Vinatzer, who collectively trained me with computational techniques, knowledge in genomics, bioinformatics and biology. They deeply broadened my horizons and fundamentally developed my critical thinking.

I would like to thank the Translational Plant Sciences program and the department of Plant Pathology, Physiology, and Weed Science which have provided me with great opportunities to learn, explore and communicate science.

Last, I must give special acknowledgment to my father and mother during my years of being abroad pursuing degrees. Without their love this would not have been possible.

TABLE OF CONTENTS

Chapter 1. Executive Summary

Equipping contemporary crops with enhanced resilience demands prompt solutions especially under the scenario of global climate change. Yet, gleaning insight on plant stress responses requires identifying the genetic underpinnings of important functional traits. By elucidating the genes and gene networks associated with stress resistance in crops and wild crop relatives, we can begin to build resilient sustainable crops. The promise, as yet incompletely filled, of sequencing technology is to enable the rapid identification of genomic regions associated with important traits. Developing tools and approaches to leverage technological and computational advances will enable the full promise of next generation sequencing to be realized. This research employs comparative, experimental transcriptomics and genomics, leveraging advanced sequencing technology and integrated bioinformatic analysis, with an aim to provide genomics and informatics resources that can be used to develop resilient agricultural plants. We integrated three projects into our work:

1) generate a guideline on transcriptome assembly. Namely, address how assembly approach, genomic divergence and sequencing depth impact the completeness of transcriptome assembly from lllumina short read data. *De novo* and reference guided transcriptome assembly approaches utilize different algorithms and result in assemblies with different utilities. Understanding how these approaches impact assemblies and downstream uses depends on several factors. First, we evaluate these two assembly approaches from our data; secondly, choosing a close related species as reference or assembly reference free (*de novo*) remains challenging especially for people who work

with non-model organisms. In other words, it is uncertain “how far is too far?” (phylogenic distance) for using a closely related species in reference guided assembly. Thirdly, of wide concern for many users is, how many reads or how much sequencing depth/coverage is enough for transcriptome assembly? Some research has shown that high coverage levels can lead to short-read redundancy in transcriptome assembly. However, it is uncertain if increasing input read numbers proportionally increases completeness of transcriptome assemblies. We interrogate these questions with various statistical analyses and an accurate transcriptome evaluation tool, BUSCO, which quantitatively evaluates transcriptome completeness according to a universal orthologs database OrthoDB. Though BUSCO is a popular tool, the database-wide orthologus alignments are computationally challenging and time consuming for large datasets and, importantly, there are no informative metrics allowing users to understand why their assemblies reach a particular quality score. Therefore, we developed a machine learning based tool to predict BUSCO-like completeness score (96% accuracy on test data) with simple evaluation matrices. It is computationally efficient removing laborious ortholog mapping steps and providing an evaluation from sequencing reads and assemblies alone. Such an implementation greatly aids people who work with large scale transcriptome datasets, where computational resource and time efficiency matters, such that these quality control measures will be skipped. Overall, we present guidelines for transcriptome assembly approaches that are based in real data and provide a new tool for high-throughput evaluation of transcriptome assembly completeness.

2) Transcriptome-wide identification of 'cross-talk' genes in Solanum *pimpinellifolum* under combined biotic and abiotic stress responses. Global change is an irreversible process that is broadly impacting the natural environment and threatening food security. In particular, the periodicity of drought (abiotic stress) and disease (biotic stress) pressure is increasing, which is predicted to have dramatic impacts on agricultural production. Plants have a unique challenge in this regard as they must survive biotic and abiotic stressors simultaneously. Yet, we find that, at the interface of stress responses ('cross-talk'), plants are constrained in their response to elicitors. For instance, under drought stress many plants demonstrate an increased susceptibility to disease and herbivory. Dissecting the molecular basis for crosstalk is challenging as specific physiological, molecular, and metabolic responses elicited from the combined biotic and abiotic stress are not-predictable from single stressor responses. Further, these traits are controlled by a number of genes. Therefore, interrogating loci independently will not comprehensively capture the mechanisms involved. Members of the species, *Solanum pimpinellifolium*, have evolved a broad range of stress tolerance mechanisms, resulting in tremendous variation with which, when combined with next generation sequencing, we can begin to dissect these complex relationships. Additionally, these may provide a novel source of stress resistance for agronomic improvement. In this context, we characterized the transcriptional profiles and expression networks using RNA-Seq data, from experimental treatments of S. *pimpinellifolium* accessions LA1269 and LA1589, which are methyl-jasmonate (JA) sensitive and insensitive respectively. Replicate plants were exposed to drought-stress (DS), mimicked herbivory (JA), and the combined stressors (JADS). To this end, we compared the differential gene expression profiles and

gene co-expression networks across experimental treatments and accessions. The “hub-genes” of protein-protein interactions and biological pathways were identified at the biotic and abiotic “cross-talk” interface. Eventually this research will provide informatics resources for plant defense and tolerance to global climate change.

3) Finally, we developed a draft assembly of *Phytopthora capsici* from long-read sequencing and in doing so uncovered previously unresolved genomic complexity. *P. capsici* is an oomycete plant pathogen that causes blight and fruit rot targeting a large spectrum of vegetable crops. Sequencing and assembly of a completed *P. capsici* genome are very important steps towards understanding its genetic associations with hosts and fungicide resistance. In order to characterize the pathogenicity of *P. capsici*, Lamour et al (2012a). generated a reference genome using Sanger Sequencing data in 2012. This assembly strategically avoided highly repetitive regions as these are notoriously difficult to resolve with short read data. Thus, some of the complexity associated with adaptation in this system was not sufficiently resolved. Oxford Nanopore Technologies (ONT) is a pore-based long read sequencing platform that generates high throughput long reads (average read length 3-15kbp), cost-effectively. With these long reads, we will be able to span repetitive regions and assemble a more complete genome. In addition, we can leverage the wealth of publicly available data that is being generated through various sequencing efforts to improve the assembly and annotation. We sequenced *P. capsici* using a single flow cell on the ONT MinION generating ~10Gb long read data. In contrast to current read correction approaches, which are computationally expensive, we generated an overlapped reads assembly and then successively polished

using high coverage long reads and publicly available short read data. This approach reduced assembly time by over 500%. In addition to saving time, the assembly was 92% bigger than the previous reference genome—confirmed by flow cytometry—was contained in fewer contigs and had an overall higher completeness score than the current reference genome. This work exemplified a protocol of rapidly improving pathogen draft genomes by combining cost-effective long read sequencing and publicly available short read data with an improved bioinformatic pipeline, to bring us closer to real-time sequencing of important crop pathogens.

In summary, integrating multiple bioinformatic approaches, this dissertation centers on improving crop resilience. This was done by 1. providing strategical approaches for crop transcriptome assembly and building tools in transcriptome completeness prediction; 2. identifying resistance genes and gene networks biotic and abiotic stress response “cross-talk”; 3. developing a cost-effective protocol to rapidly assemble a complex pathogen genome and unravel its hidden complexity.

Chapter 2. A Machine Learning Based Method for Transcriptome Assembly Assessment

Chenming Cui[1], Aureliano Bombarely[1,2], David C. Haak[1*]

1. School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061 USA

2. Present address: Department of Bioscience/Dipartimento di Bioscienze University of Milan/Universita degli Studi di Milano (UNIMI), III piano / torre B Via Celoria, 26 Milano, Italy, 20133

Chenming Cui: chenmc1@vt.edu

Aureliano Bombarely: aureliano.bombarely@unimi.it

David Haak: dhaak@vt.edu

*To whom correspondence should be addressed: dhaak@vt.edu

ABSTRACT

The advent of whole transcriptome sequencing (RNA-Seq) is providing genomic scale information for non-model systems, where complexity and/or cost precluded whole genome sequencing efforts. While a reduced representation of the whole genome, the captured profile is a collection of all the expressed transcripts for any given tissue or collection of tissues. However, RNA-Seq requires researchers to make decisions at multiple points. An optimal practice guideline is thus needed, especially for non-model organisms (reference guided and *de* novo using the popular tool Trinity). This study also proposes a machine learning based approach in transcriptome completeness assessment. It is a reference free, BUSCO score alike tool that quantitatively assess completeness of transcriptome from short read data without using an ortholog database. *Results:* 68 accessions of Solanum and Capsicum were assembled in both *de novo* and reference guided assembly approaches (using a common tool, Trinity). We assessed assembly quality via the BUSCO completeness score. Assembly approaches were comparable in terms of BUSCO completeness score across the entire dataset, but each of them offered significant advantages for particular metrics. Multi-model inference and correlation analysis suggested 1) the importance of input read number, where more reads tend to assemble more complete transcriptomes; 2) Phylogenetic distance matters in reference guided assemblies, with ~5% of phylogenetic divergence as the limit before transcriptome completeness begins to fall; 3) we trained a random forest model in machine learning to predict assembly BUSCO completeness scores, resulting in a model with 95% accuracy.

Conclusions: This study provides a practice guideline for transcriptome assembly from RNAseq short-read data and a new tool for reference free evaluation of transcriptome assembly completeness. Here we found that input read number is positively correlated with assembly completeness, never reaching a plateau in our study even when using up to ~200M reads. Also, *de novo* and reference guided assembly approaches offer distinct advantages with different aspects and the optimal strategy should align with downstream objectives. Phylogenetic divergence is important in reference guided assembly wherein increasing distance over ~5% offered no advantage over *de* novo assembly, suggesting that researchers in this situation should opt for *de novo* assembly approaches. Finally, we provided a reference-free machine learning tool for estimating transcriptome completeness that provides a BUSCO-like score as well as identifying features impacting this score. This tool will greatly benefit transcriptome assembly quality evaluation for large-scale datasets and non-model organisms.

Keywords: Transcriptome assembly, De novo assembly, Reference guided assembly, Non-model organisms, Transcriptome assessment, Machine learning

INTRODUCTION

Next generation sequencing has resulted in a drastic increase in sequencing throughput and subsequently has continued to drive down costs of generating sequencing data. Illumina RNA-Seq (Wang, Gerstein et al. 2009), massively parallel sequencing of cDNA and Iso-seq (Eid, Fehr et al. 2009), single-molecule long-read isoform sequencing (Iso-seq) are important applications of high-throughput sequencing. They capture the global gene expression profile from many tissue types and organisms of interest (Trapnell, Williams et al. 2010). Hence, the captured profile is a collection of all the expressed transcripts, providing a reduced representation of the genome. Accordingly, the collection is defined as a transcriptome(Huang, Chen et al. 2016). This reduced representation of the genome has provided an exceptional cost-effective opportunity for bringing genomic tools to non-model systems (Collins, Biggs et al. 2008, Cahais, Gayral et al. 2012). RNA-Seq technology has been applied to many fields of biological science and proved its success in gene regulatory networks (Bhardwaj, Josse et al. 2018), single nucleotide polymorphism discovery (Fan, Lee et al. 2018), alternative splicing capture (Codina-Fauteux, Beaudoin et al. 2018), and differential gene expression analysis (Jones, Usman et al. 2018).

Assembling short RNA-Seq reads into transcriptomes is computationally challenging (Grabherr, Haas et al. 2011). Currently, there are two computational strategies for reconstructing transcriptomes from RNA-Seq short reads, alignment-based methods (reference guided) or graph-based assembly (*de novo*) (Song, Catlin et al. 2018). The choice between the two strategies often depends on the availability of a reference and the intended downstream use of the data. The reference guided approach, such as

Scripture (Guttman, Garber et al. 2010) or Cufflinks (Trapnell, Williams et al. 2010), aligns reads to an existing reference genome, taking possible splicing events into consideration, then merges overlapping alignments. The boundaries between exons (splice junctions) are bridged with junction spanning reads (Grabherr, Haas et al. 2011). A number of tools exist for reference guided transcriptome assembly and several recent reviews compare these tools (Conesa, Madrigal et al. 2016) (Castel, Levy-Moonshine et al. 2015). The second approach, *de novo* assembly, assembles transcripts directly from the RNA-Seq reads (Grabherr, Haas et al. 2011). Many of these tools use de Brujin graphs, an advanced computational algorithm that parses the sequence fragments into individual graphs, which are comprised of nodes and edges. Nodes are nucleotide sequences with length of k. If k-1 nucleotides are the overlaps between two nodes, they are defined as an edge that connects the nodes (Grabherr, Haas et al. 2011). And then graphs are routed independently to stitch together full-length isoforms. That is, these tools attempt to assemble reads into transcripts directly.

Most downstream applications of transcriptome data, e.g., phylogenomic, differential gene expression, and candidate gene exploration, etc., all depend on the integrity and completeness of the assembled transcriptome. Yet, most decisions rely on the availability of a reference genome with the implicit assumption that reference guided assembly approaches are generally better at reconstructing full-length transcriptomes from short-read data. Thus, a better understanding of the comparative strengths and biases in assembly strategy is needed to understand how assembly approach impacts downstream analyses. For example, a *de novo* assembly program takes advantage of de Brujin graphs to assemble transcripts directly, providing the opportunity to discover novel transcripts

that are not contained in the reference assembly, yet these may come at the cost of gene model completeness. In other words, *de novo* approaches may drive the formation of assemblies with more fragmented transcripts. Also, it may incur chimeras in de Bruijn graph connection by self-connection at the end of a transcript or connection with hairpin's reverse-complement sequence, which lead to misassemblies (Bombarely, Edwards et al. 2012, Bushmanova, Antipov et al. 2018). Secondly, RNA-Seq enables us to capture gene expression profiles with great depth, however, previous reports indicate that too much read depth can negatively impact transcriptome assemblies by reducing both median contig length and contig integrity (Huang, Chen et al. 2016). Thirdly, genomic resources from closely related species can serve as a reference to guide transcriptome assembly for non-model organisms, but, "How far is too far?" and what are the impacts of phylogenetic distance on transcriptome completeness.

Addressing these questions requires a way to comparatively evaluate transcriptome assembly quality. While multiple tools exist, each has a different assessment strategy including, comparative assessment using rnaQUAST (Bushmanova, Antipov et al. 2016) and BUSCO (Simão, Waterhouse et al. 2015), reference-independent metrics using TransRate (Smith-Unna, Boursnell et al. 2016) and a likelihood based assessment, RSEM-eval (Li, Fillmore et al. 2014) have been applied when the ground truth is unknown. Of these tools, BUSCO and TransRate are widely used because of their flexibility and relative ease of use. BUSCO is often considered the gold standard because it aligns universal single-copy orthologs selected from a comprehensive catalog of orthologs database, OrthoDB v9 (Zdobnov, Tegenfeldt et al. 2016), to assesses the completeness of genome and transcriptome assemblies. Conversely, TransRate evaluates assembly

integrity by mapping sequencing reads to the assembly and providing a global quality score which is a weighted mean of various comprehensive metrics such as contig length, number of reads mapped, contig N50, etc. All of these quality evaluation tools rely on metrics that are time consuming to produce, computationally intensive, or both, Only TransRate provides insights on the metrics that impact the completeness score.

Machine learning enables the development of supervised (using a training set) or unsupervised computational algorithms and applies them to deconvolute highly complex interactions among feature sets (Signal, Gloss et al. 2017). This approach has been successfully applied to a wide range of areas within genetics and genomics (Sonnenburg, Rätsch et al. 2002, Rätsch, Sonnenburg et al. 2007, Reinbolt, Sonis et al. 2018). In particular, the random forest (RF) algorithm (Breiman 2001) is an ensemble method that is comprised of many individual decision trees, wherein each tree is built from a sample drawn with replacement to average out variance leading to an overall better model. These approaches have already been applied to genomic and transcriptomic data to classify coding and noncoding RNA (Pan, Xiong et al. 2018), identify cell type (Herman and Grün 2018), characterize disease progression (van Galen, Hovestadt et al. 2019), among other insights.

Here we have generated data from wild tomato and pepper species to address the role of assembly approach on the completeness of transcriptome assemblies and devise a machine learning approach to assess assembly quality. Specifically, we ask--How does sequencing depth impact quality of the assembly for both reference guided and *de novo* approaches? What is the impact of phylogenetic distance on assembly quality within reference guided assemblies? Does the difference between assembly algorithms drive a

difference in assembly quality? We then use these outcomes to inform a random forest machine learning algorithm that returns a BUSCO compatible quality score and the aspects of the assembly with the greatest impact on the completeness score.

RESULTS

Transcriptome completeness assessment using BUSCO and TransRate

We used BUSCO to quantitatively assess the completeness of the transcriptome assemblies. All of our accessions were assembled using both *de novo* and reference guide approaches as implemented within Trinity v2.0.6 (Grabherr, Haas et al. 2011). Reference guided assemblies were first aligned to the reference genome using STARv2.5a (Dobin, Davis et al. 2013) before being transferred to the Trinity GG (Genome-guided assembly) pipeline. BUSCO completeness scores for reference guided assemblies ranged from 52.90 to 88.50 with a mean of 76.28 (Figure 1). *De novo* assemblies were reported BUSCO completeness scores with range of 52.60 to 88.90, mean 76.30 (Figure 1). To test, if Trinity *de novo* assembly and reference guided assembly approaches lead to significantly differences in transcriptome's completeness, we examined our assemblies with Wilcoxon signed-rank test. Reference guided assemblies were significantly better than *de novo* assembly in transcriptome completeness (Wilcoxon signed-rank test, V = 676, p-value = 0.023). To investigate specific differences among the assemblies we examined the performance of some typical assembly quality metrics using TransRate 1.0.3 (Smith-Unna, Boursnell et al. 2016)(Figure 2). Here, we found that contig N50 size was not significantly different between *de novo* assemblies and reference guided assemblies (Wilcoxon signed-rank

test, V = 1489, p-value = 0.054). For Number of bases assembled, proportion of mapping to reference and number of contigs, *de novo* assemblies were significantly larger than reference guided assemblies with V = 67 (p-value = 1.4e-11), V = 2300 (p-value = 5.8e-12) and V = 171.5 (p-value = 9.5e-10) respectively in Wilcoxon signed-rank tests. However, reference guided assemblies were significantly better than *de novo* assembly in proportion of contigs hitting reference and mean contig length with V = 2103 (p-value = 1.3e-08) and V = 1637 (p-value = 0.0046) in Wilcoxon signed-rank test respectively. Taken together, these results suggest that *de novo* and reference guided assembly approaches have specific advantages over each other with different metrics.

To investigate the impact of each metric on the BUSCO completeness score, we correlated BUSCO scores with the TransRate assembly optimal score, which is also intended to reflect assembly completeness. Surprisingly, there was no correlation between these assembly completeness metrics across the entire dataset (Figure 3A; reference guided; lm, slope = 7.7, adj.R^2 = -0.01, p-value = 0.64 & Figure 3B; *de novo*; lm, slope = -4.1, adj.R^2 = 0.01, p-value = 0.72). To deconvolute the relationship between the weighted metrics associated with the TransRate assembly score we performed principle component analysis (PCA) among all the metrics (Figure 4). Metrics like “contigs with lowcovered”, “contigs_segmented” and “gc_skew” were clustered together with ‘negative’ loading in the plot, while, in contrast, the proportion of low covered contigs (“p_contigs_lowcovered”) showed an opposite loading pattern. Secondly, we found that the TransRate optimal score showed a week positive correlation with proportion of low covered contigs in Figure S1A (lm, slope = 0.55, adj.R^2 = 0.08, P-value = 4.7e-04). However, this same metric showed a negative correlation between BUSCO completeness

score and proportion of low covered contigs in Figure S1B (lm, slope = -97.94, adj.R^2 = 0.23, p-value = 1.54e-09). Overall, we found that the TransRate assembly score did not show any correlation with the BUSCO score. Fruther, the correlation of "p_contigs_lowcovered" with the TransRate score was orthogonal to the relationship between "p_contigs_lowcovered" and BUSCO score.

Dissecting the effects of phylogenetic distance and input read numbers in transcriptome assembly

To assess the impact of phylogenetic distance and input read number on assembly completeness we used a mixed modeling framework with BUSCO completeness scores as the dependent variable. Assembly approach, *de novo* or reference guided was treated as random effect in the models. Phylogenetic distance and read numbers, which denoted by species variables and read numbers were tested as fixed effects in the models. The most complex model was established with "Species" and "Reads" variables getting a superior fitness with AICc score of 786.75. With the reduction of fixed effect in the following "reads" model and "Null" model, the model fitness (AICc values) decreased notably. The alteration of model fitness from "Species + Reads" model to "Reads" model and "Null" model, which were reflected in "Delta AIC" from 0.00 to 49.73 and 125.43 indicated the weight of phylogenetic distance and reads numbers in impacting transcriptome completeness.

To further test the effects of input read number on transcriptome assembly we evaluated the relationship between input read number and several important quality measures. First, we found a positive correlation between input read number and total bases assembled

(Figure 5A; lm, slope = 2.36, adj.R^2 = 0.55, p-value = 3.52e-25). Secondly, the number of input reads was positively correlated with the BUSCO completeness score up to 200 million reads with no plateau (Figure 5B; slope = 1.11e-06, adj.R^2 = 0.33, p-value = 1.45e-13). Similarly, a positive correlation was found between input read number and the number of contigs with a CRB-BLAST hit (Figure 5C; slope = 6.27e-04, adj.R^2 = 0.59, p-value = 7.28e-28). The CRB-BLAST hit denotes the number of reciprocal best hits against the S. *lycopersicum* reference genome, therein reflecting the importance of input read number in assembled contig completeness. Taken together, both phylogenetic distance between the reference genome and the sequencing organism and number of reads in the transcriptome assembly are all important for generating complete transcriptomes when using reference guided assembly. Importantly, the number of reads continuous to contribute to transcriptome completeness up to using 200 million read pairs, which was our maximal detection power in the study.

A machine learning approach in transcriptome completeness prediction

While BUSCO only produces a single vector to determine assembly completeness, TransRate produces a large matrix (N= 54) of quality measures including “Contig metrics”, “Read mapping metrics” and “Comparative metrics”. “Contigs metrics” provides statistics for contigs in an assembly, accounting for 18 features, for example, the number of contigs in the assembly (“n_seqs”), the mean length of the contigs (“mean_len”), and N50 value. “Read mapping metrics” provided alignment statistics between reads and the assembled contigs, accounting for 19 features, for instance, the number and proportion of reads mapping that suggested a bad assembly (“bad mappings”), the number of contigs with >

50% of chance of being segmented ("contigs segmented"). "Comparative metrics" incorporated reference-based comparisons, to compare the assembly and accounted for 17 features. "CRBB hit" is an example of this category, which is an estimate of the number of reciprocal best hits between assembled contigs and a reference set. Using all of these features in our RF regressor to predict BUSCO completeness score, our model reached an accuracy of 0.97, 0.94 and 0.94 on the training, validation, and test datasets respectively. However, the feature importance plot (Figure 6A) suggested that a large portion of our initial features were not informative. In our second attempt, we reduced the number of features, excluding uninformative features from "Comparative metrics". The model with the best performance retained just 15 of the original 54 features, returning an accuracy of 0.99, 0.95 and 0.95 on training dataset, validation, and test datasets respectively. The most informative features were "contigs segmented", "bad mapping", and "n50" contig size (Figure 6B). Note that our second trained model was a reference-free model, reporting a BUSCO-like completeness score without mapping assembly against orthologues database. We lunched this machine learning tool, "WWMT" ("What's Wrong with My Transcriptome") in Jupyter notebook with Python.

DISCUSSION

In this study, we used 136 non-independent transcriptome assemblies from 68 Solanaceae accessions encompassing *Solanum* and *Capsicum* to show that assembly approach can have important impacts on transcriptome integrity and completeness, therein impacting downstream uses. Using two common quality assessment tools, BUSCO and TransRate we were able to comparatively assess the impacts of assembly

approach (*de novo* vs. reference guided), phylogenetic distance to reference species (for reference guided) and input read number in transcriptome assembly quality. When TransRate optimized scores showed a certain bias, we took advantage of TransRate output matrices to generate a machine learning model for transcriptome completeness assessment.

De novo and reference guided transcriptome assembly

Comparison of BUSCO completeness scores suggest that reference guided assemblies are more complete than *de novo* assemblies (p-value = 0.023). The distribution of scores suggest that these two methods are equitable across the quantile distribution with very close ranges between the two (Figure 1). *De novo* assembly provides a reasonable alternative when the reference species is not available or incomplete, but assembly should be evaluated for quality prior to implementation. Further, careful consideration should be given to the intended downstream use of the assemblies. For example, our study found that for Trinity *de novo* assemblies, contig N50 size was not significantly better than *de novo* assembly (p-value = 0.054). Whereas in other important metrics, for instance, “number of bases assembled”, “proportion mapping to reference” an “number of contigs”, *de novo* assemblies were superior to reference guided assembly (p-value < 0.05, Figure 2). However, reference guided assemblies tended to generate more complete transcriptome, likely because a guided reference approach makes it feasible to assemble low coverage regions (Martin, J.A., Wang, Z., 2011).

Because reference guided approaches may do a better job capturing gene models (Figure 2), in systems lacking genomic resources, researchers will often choose to use a

reference from a closely related organism. While simulation data suggests that gene capture declines with increasing divergence about ca. 8% (Vijay, Poelstra et al. 2013), limited empirical data exist on the impact of nucleotide divergence on assembly quality. Our data generally support this finding, including phylogenetic distance as a fixed effect in mixed models improved the fit as evidenced by a decrease in deltaAIC (Table 2). This suggests that nucleotide distance is an important factor affecting assembly quality, when using the reference guided approach.

Effect of read number on assembly quality

Minimum sequencing depth is important for transcriptome assembly, as both k-mer based (Trinity *de novo* assembly) and mapping based approaches (STAR, Trinity reference guided assembly) rely on read consensus. However, a recent study suggested that increasing read number should reach an asymptote where further increases could result in redundancy of the assembly, therein impairing median contig length and identity between contigs and reference protein sequences (Huang, Chen et al. 2016). However, we did not observe similar trends in our study. Instead, we found that the number of bases assembled directly correlated to input read number (up to ~ 200M; Figure 5A). Quality metrics support the integrity of these assembled bases, for example CRB-BLAST hits, a direct reflection of the identity between contigs and the reference protein sequences, increased with increasing read input number (Figure 5B). Finally, the BUSCO completeness score was positively correlated with the number of input reads (Figure 5C). Thus, while input read numbers above 200M may negatively impact assemblies, we found no loss in completeness or integrity when using up to 200M reads.

Inconsistencies between BUSCO and TransRate

Even though BUSCO and TransRate scores are generated from different metrics, we expected them to have a positive relationship. Instead, we found no relationship between TransRate optimal score and BUSCO score (Figure 3A, 3B). This was surprising because, even though TransRate scores are a composite of several different metrics, there are components that should result in a higher BUSCO score. To dissect this discrepancy, we performed Principal Components Analysis (PCA) on the TransRate scores to identify which metrics could be driving these differences. Intriguingly, we found that many metrics clustered in the anticipated direction, for example CRB-BLAST hits (Figure 4). Yet, other metrics, such as the "proportion of contigs with low coverage" did not cluster with negative indicators like "contigs with low coverage" or "fragmented contigs" (Figure 4). We suspected that the TransRate algorithm was over-weighting some low coverage metrics and this was artificially increasing the optimal score. We confirmed this by generating skimmed datasets representing different read depths for the same sample. Here we found that, counterintuitively, TransRate reported higher optimal scores for assemblies with fewer input reads (Figure S2A, S2B), while BUSCO scores were negatively correlated, as predicted (Figure 5C). Thus, TransRate may be overestimating assembly scores based on read depth and this likely explains why TransRate optimal scores are inconsistent with BUSCO scores.

Metrics in transcriptome completeness assessment via machine learning

Many metrics such as N_{50} size, total contig number and average contig size are commonly used in whole-genome assembly evaluation, people adapt these metrics to *de novo*

transcriptome assemblies directly (T O'Neil and Emrich 2013). However, the assumptions used for evaluating genome assembly differs from transcriptome assembly, as such, the utility of these metrics is questionable. For example, contig coverage is expected to be exponentially distributed for transcriptome assembly (Smith-Unna et al., 2016), since gene expression differs by nature, so interrogating coverage statistics for these assemblies is less useful than for genome assemblies. Additionally, assessing these metrics individually may also lead to bias. For instance, high rates of chimeras from mis-assembly could inflate N_{50} and length statistics. Similarly, a high rate of failure to assemble low abundance transcripts could also bias these statistics. Coupled with these problems in using metrics to evaluate transcriptome assembly is the time-consuming nature of evaluating against a reference set (e.g., BUSCO) or mapping raw reads to the assembly. Therefore, we decided to take advantage of the metrics generated from TransRate to train a reference free machine learning model in transcriptome completeness evaluation. Here we investigated the potential of machine learning with pooled features from TransRate contig metrics (e.g. "number of contigs", "the mean length of the contigs"), read mapping metrics (e.g. "the proportion of the provided read pairs that mapped successfully", "the number and proportion of reads pairs mapping in a way indicative of bad assembly") and comparative metrics (e.g. "number of reference proteins with at least X% of their bases covered by a CRB-BLAST hit"). With these metrics the accuracy of a random forest regressor on training, validation, and test data sets reached 0.99, 0.95, and 0.95, respectively. From this model we estimated the impact on assembly quality and found that the top significant features were "bad_mapping", "contigs_segmented" and "n50" (Figure 6A). "bad_mapping" reflected accuracy of the

contigs that assembled from its input reads. “contigs_segmented” indicated completeness of the contigs, whereas “n50” was the reflection of contig length. Interestingly, read mapping related features (“comparative metrics”) were not found in the top 10 significant features. These are features such as, “p_cov95” which referred to proportion of reference proteins with at least 95% of their bases covered by a CRB-BLAST hit (Figure 5B). We refined our model to exclude the read mapping components and identified final feature importance, which again retained the features identified above, but reordered. (Figure 6B).

BUSCO is a widely accepted method for uncovering the ground truth of assemblies’ completeness, but practically when people handle large scale of transcriptome datasets, the computational power and time (~1h/per transcriptome assembly(~75M)) is still a headache for researchers. Here we trained our model with BUSCO score, building a reference-free, BUSCO-free BUSCO score-alike machine learning tool in transcriptome completeness assessment. We expect this tool to greatly improve large scale dataset quality checks and non-model organism transcriptome assembly.

CONCLUSION

This study provided a practice guideline for transcriptome assembly from RNAseq short-read data. The greater number of input reads tends to lead a better transcriptome completeness with maximal detection of ~200M reads. *De novo* and reference guided assembly approaches advantages over each other with different aspects, the optimal strategy should align with downstream objectives. Cautions should be taken choosing a close related species in reference guided assembly since the advance of *de novo*

assembly tends to match up 5% of sequence divergency cost of a reference species. TransRate generated examinations on contigs but we didn't see any correlation between TransRate score and BUSCO completeness score. We provided a reference-free, BUSCO-free BUSCO score-alike machine learning tool in transcriptome completeness assessment, it will greatly benefit large scale datasets quality check and non-model organism transcriptome quality assessment.

MATERIALS & METHODS

Sample collection, plant RNA extraction and sequencing

We sampled single or multiple accessions from 15 *Solanum* species and 10 *Capsicum* species (Table S1) in South America, along the west slope of the Andes to represent a broad spectrum of diversity in wild tomato and its wild relatives. RNA was extracted from pooled tissues including leaf, root, flower, buds and fruits of three stages. RNA extraction was performed using RNeasy plant QIAGEN kit according to manufacturer's protocol. RNA quality was accessed using NanoDrop 2000. Bar-coded 150bp cDNA libraries were prepared using TruSeq RNA sample preparation kit ver.2 according to manufacturer's protocol and recommendations. All libraries were sequenced on the Illumina HiSeq4000 platform. The generated raw reads were submitted to the sequence read archive (SRA).

Assembly of transcriptomes

Sequencing generated FASTQ files were quality checked using FASTQC v0.11.3 (Andrews 2016). Then the raw reads were trimmed to clean the reads with the following

parameters of Trimmomatic v0.33 (Bolger, Lohse et al. 2014)"--ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:150".

Paired ends sequences of each accession were assembled into transcriptome using Trinity v2.0.6 (Grabherr, Haas et al. 2011) with the default *de novo* assembly parameters. Meanwhile, reference guided assembly was performed on all the accessions as well. We used STAR v2.4 to align reads of each accession to their coordinate reference genomes S. *lycopersicum* v2.50 or C. *annuum* v2.0 (Fernandez-Pozo, Menda et al. 2014) with default setting but retain the unmapped reads by "--outReadsUnmapped" option and output coordinate-sorted bam files by "--BAM_SortedByCoordinate" option. With the sorted bam file, we used "--genome_guided_bam" option in Trinity to assemble into transcriptome at each locus.

Assessment of assemblies' quality

TransRate v1.0.3 (Smith-Unna, Boursnell et al. 2016) was used for transcriptome assembly quality evaluation. Reference protein sequence file of *S.lycopersicum* and *C.annuum* were set with --reference flag for Solanum and Capsicum species in the data set respectively; Pair ends reads were used to enhance the evaluation with –left and –right flags. We evaluated the *de novo* and reference guided assemblies of each accession together using --assembly flag. To measure the completeness of these transcriptome assemblies, we used BUSCO v3 with 'embryophyta_odb9' as reference database.

Statistical analysis

To compare the BUSCO scores of reference guided assemblies and de novo assemblies, we conducted hypothesis tests using Wilcoxon signed-rank test in R 3.2.2 (R Core Team, 2013). Prior to this, the normality of the BUSCO scores was examined with Shapiro-Wilk test in R. The linear mix-effects models (LMM) approach was chosen to examine the influence of assembly methods (*de novo* and reference guided), phylogeny distance, which represented by species' category, and numbers of input reads that used on assembly completeness performance (BUSCO). Accession was coded as a mixed effect in the LMMs. Model selection was executed according to lower AICc value, which suggested a higher explanatory power of the model. The significance level of the fixed effects was tested by likelihood ratio tests. We started with a global LMM that included all fixed effects and simplified the starting model in sequential steps. In the following steps, we fit all possible sub-models. The associated change of AICc value (delta AICc) with each effect reduction reflected the significance level of the effect. All analyses were conducted by using lme4 v1.1-21 (Bates, Sarkar et al. 2007) and AICcmodavg v2.2-1 (Mazerolle and Mazerolle 2019) packages in R software.

Machine learning

All of the assemblies described above were used in machine learning model. Assembly methods (*de novo* or reference guided), input read numbers and TransRate's outputs matrix including "Contigs metrics", "Read mapping metrics" and "Comparative metrics" were collected as features. We started our naïve attempt with all these features and began to reduce features that were in "Comparative metrics". Then, to predict transcriptome completeness score (BUSCO), an ensemble regressor Random Forest

was employed. The entire dataset was split in training dataset, validation dataset and test dataset with 60%, 20% and 20% of the dataset respectively. The performance of our regressor was measured by coefficient of the model. All the machine learning steps were performed using scikit-learn v0.20.0 (Pedregosa, Varoquaux et al. 2011) in Python3.6.

TABLES & FIGURES

Model	K	Log-likelihood	AICc	∆AICc	AICc.Wt
Species + Reads	29	-356.16	786.75	0.00	1
Reads	4	-414.08	836.48	49.73	0
Null	3	-453.00	912.18	125.43	0

Table 1. Multi-model inference comparing models that incorporated input read number and species as a proxy for phylogenetic distance. In “Species + Reads” model, assembly method (*de novo* or reference guided) testing fixed effects of “Species” (phylogenetic distance) and “Reads” (input read number). Fixed effects were reduced in the following “Reads” model and “Null” model. AICc was the measured for model fit. ΔAICc indicated the value change of AICc comparing to the “Species + Reads” model. AICc.Wt (abbreviation for weight) stranded for the support for the model.

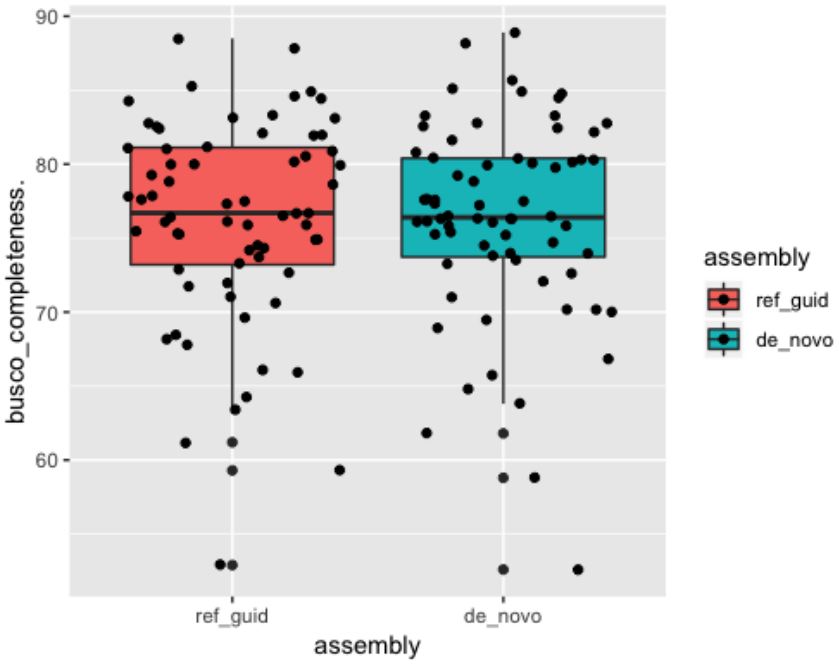


Figure 1. Boxplots depicting BUSCO completeness scores for reference guided versus *de novo* assemblies (N = 68). Reference guided assemblies were significantly better than *de novo* assemblies in BUSCO completeness scores (Wilcoxon signed-rank test, V = 676, p-value = 0.023).

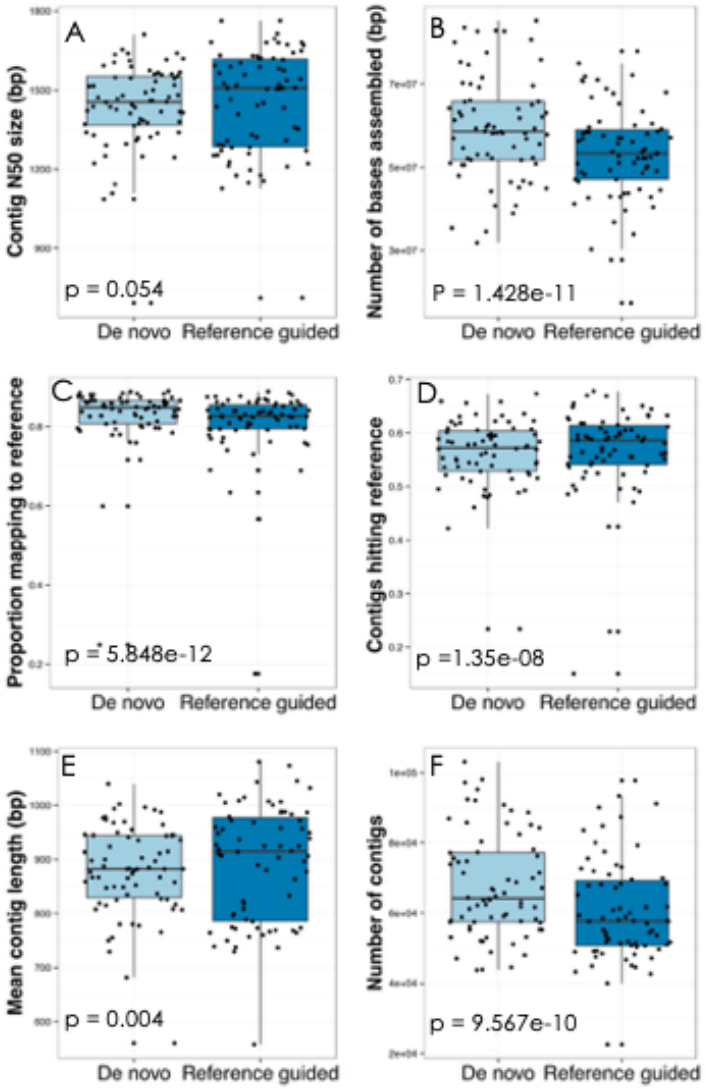


Figure 2. Boxplots showing the relationship between various quality metrics and assembly approach. A). compared contig N50 size, no significant difference between *de novo* and reference guided assembly (Wilcoxon signed-rank test, V = 1489, p-value = 0.054); B). compared number of bases assembled, *de novo* assembled larger number of bases (Wilcoxon signed-rank test, V = 67, p-value = 1.4e-11); C). compared proportion of mapping to reference genome*, S.lycopersicum* and *C.annuum accordingly*, *de novo* had larger mapping proportion than reference guided (Wilcoxon signed-rank test, V = 2300,

p-value = 5.8e-12); D). compared proportion of contigs hitting reference, reference guided had lager hitting proportion (Wilcoxon signed-rank test, V=2103, p-value = 1.3e-08); E). compared mean contig length, reference guided generated longer contigs (Wilcoxon signed-rank test, V= 1637, p-value = 0.0046); F). compared number of contigs, *de novo* assembled transcriptome with more contigs than reference guided (Wilcoxon signed-rank test, V= 171.5, p-value =9.5e-10)

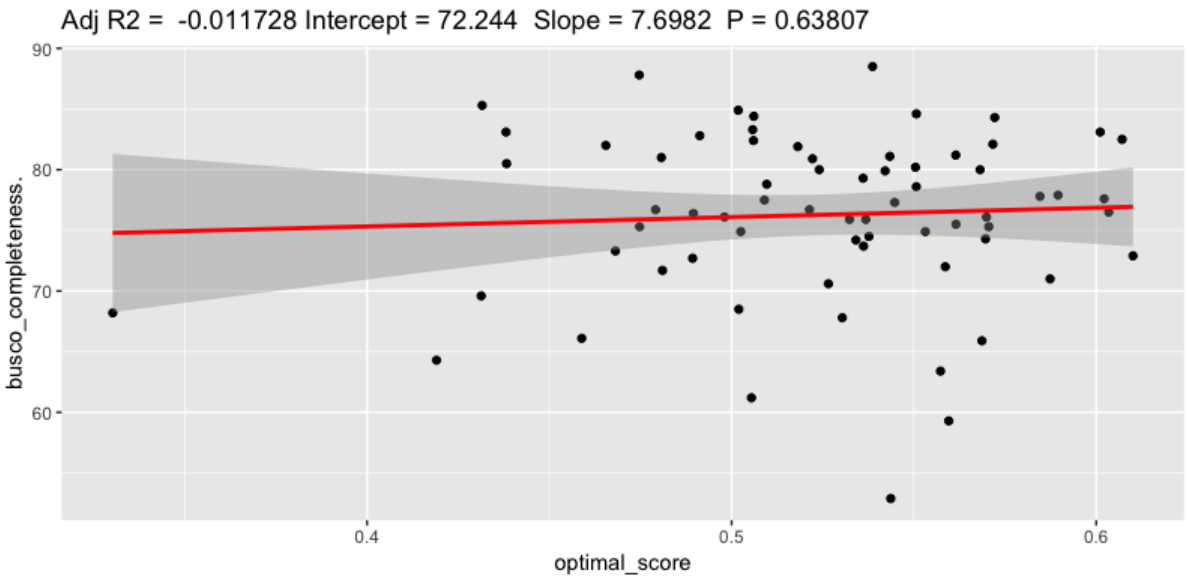


Figure 3A.

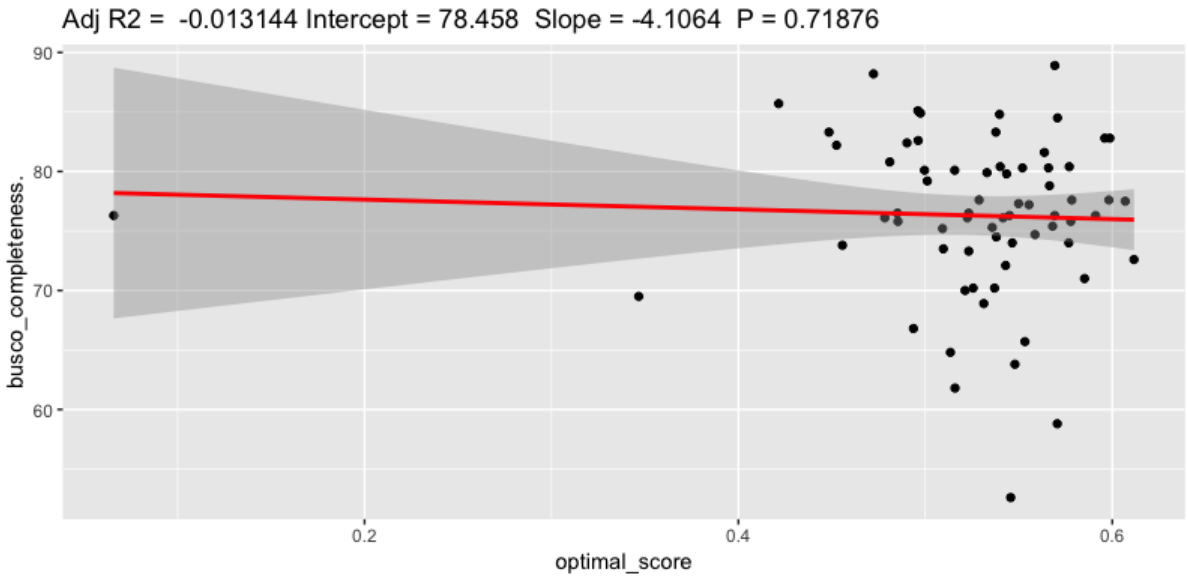


Figure 3B.

Figure 3. BUSCO completeness score versus TransRate assembly optimal score. No relationship is seen between the two quality scores. Figure 3A. Reference guided on (lm, slope = 7.70, adj. R^2 = -0.01, p-value = 0.64) Figure 3B. *de novo* on f (lm, slope = -0.41, adj. R^2 = -0.01, p-value = 0.72)

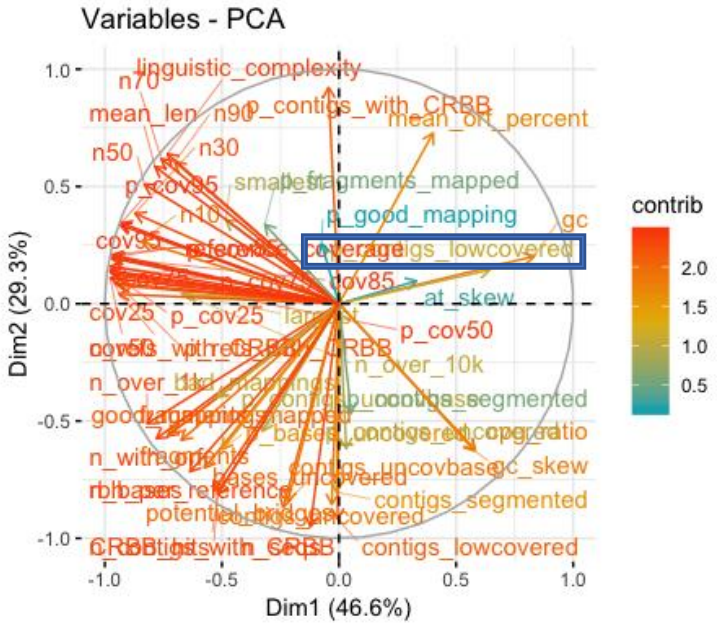


Figure 4. PCA analysis among metrics variables. Positive correlated variables pointed to same side of the plot, and negative correlated variables pointed to the opposite side. “contrib” denoted for the variable contributions to the principal components. Variable “p_contigs_lowcovered” (proportion of low covered contigs) was marked.

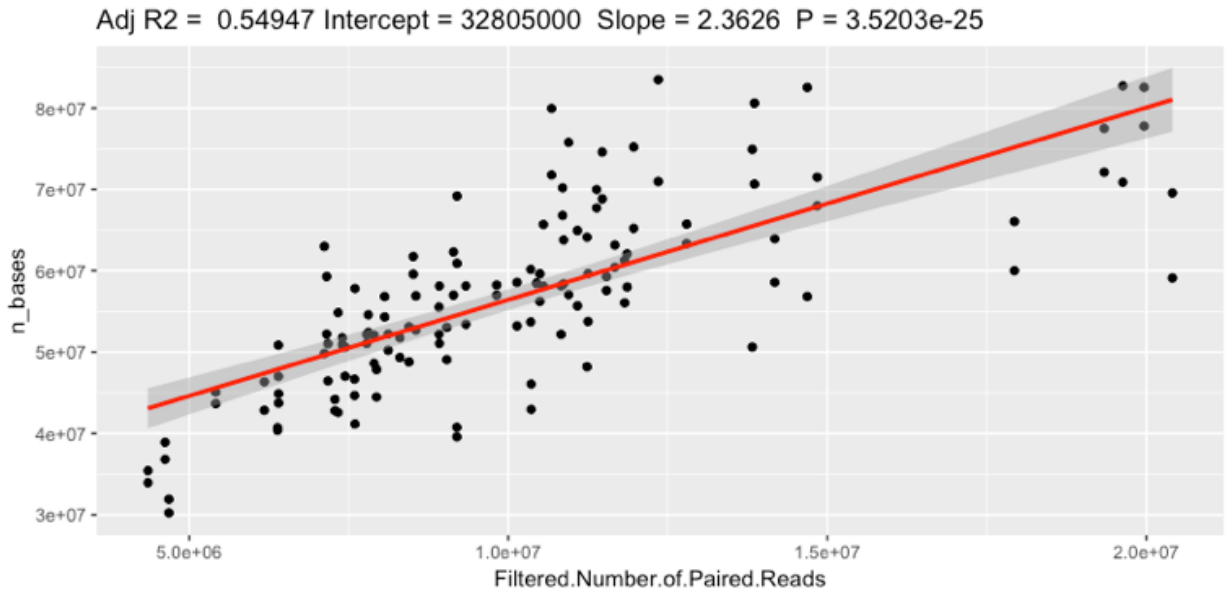


Figure 5A.

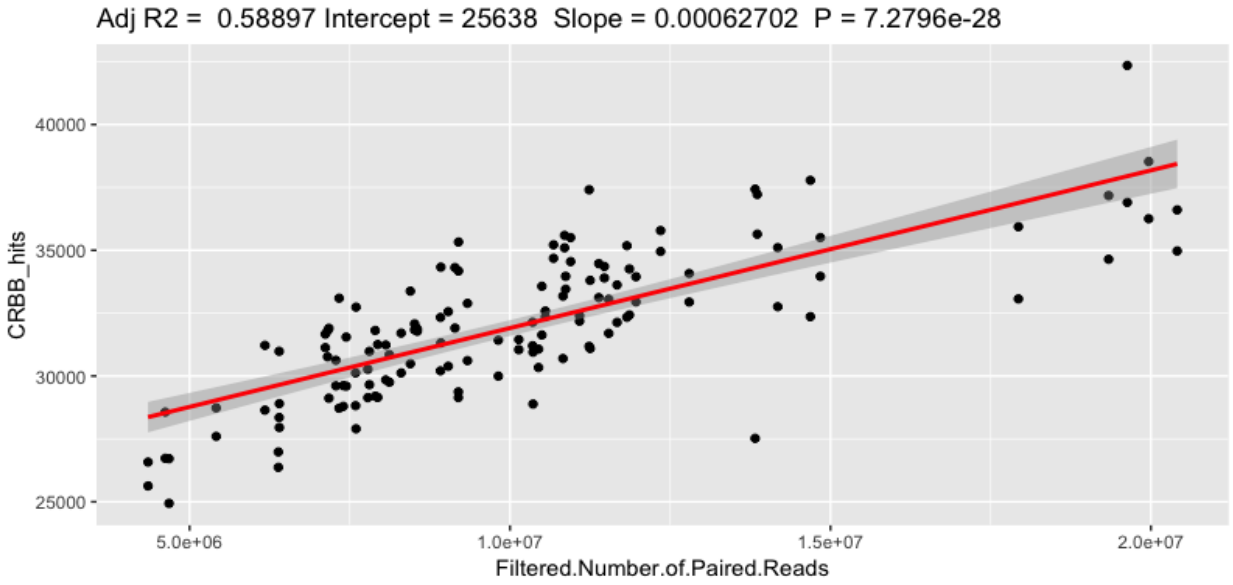


Figure 5B.

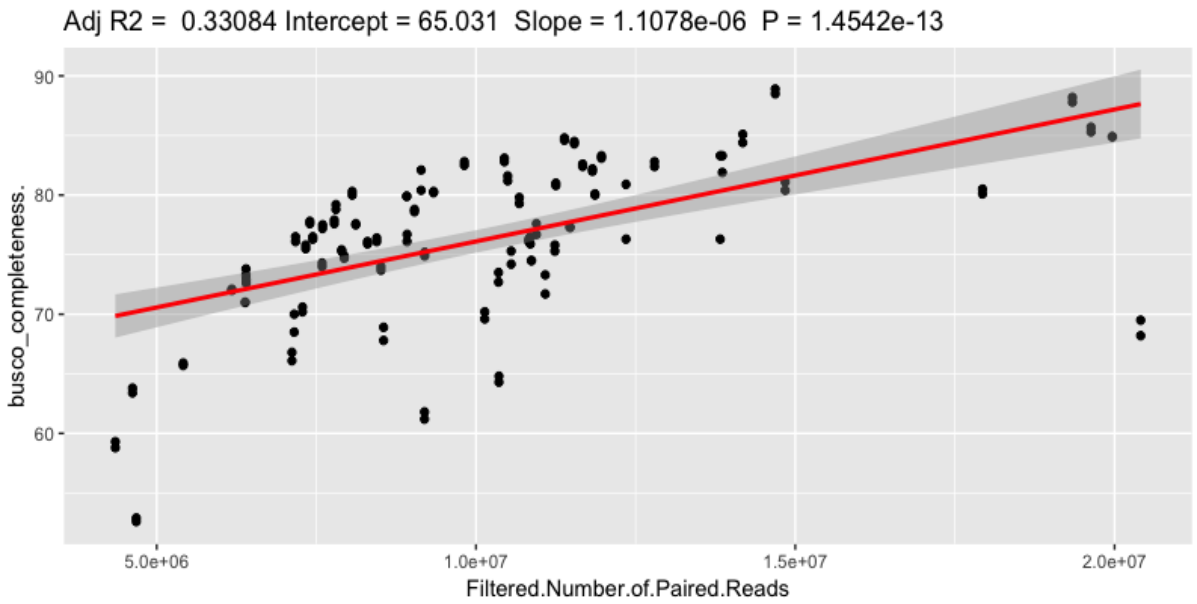


Figure 5C.

Figure 5. The relationship between input read number and A). number of bases assembled (lm, slope = 2.36, adj.R^2 = 0.55, p-value = 3.52e-25), B). BUSCO completeness score (lm, slope = 1.11e-06, adj.R^2 = 0.33, p-value = 1.45e-13) , and C). conditional reciprocal best blast (CRBB) hits (lm, slope = 6.27e-04, adj.R^2 = 0.59, p-value = 7.28e-28)

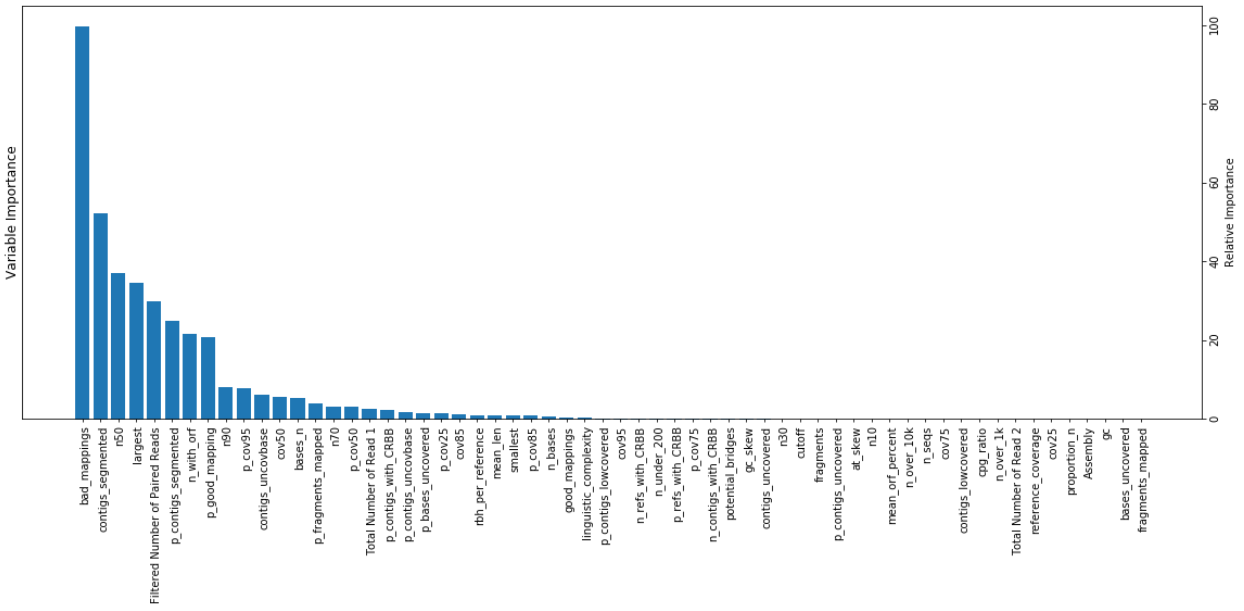


Figure 6A.

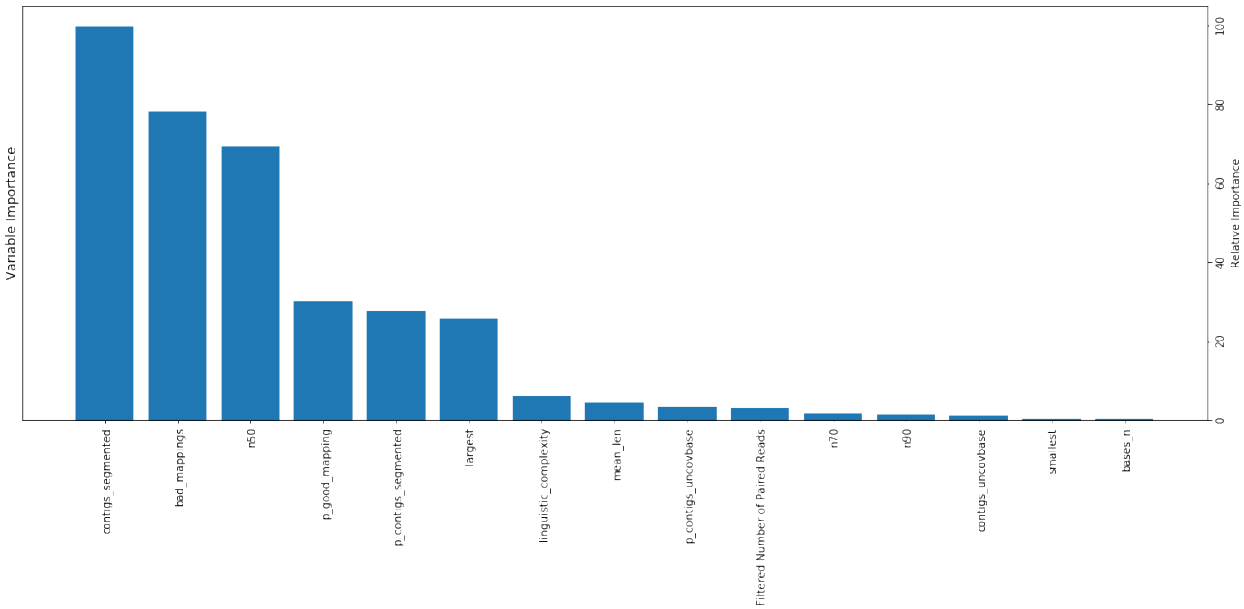


Figure 6B.

Figure 6. Variable importance of our machine learning predictive model- A). naïve model, features selected from “Contig metrics”, “Read mapping metrics” and “Comparative metrics”, the top three important features were “bad_mapping”, “contigs_segmented” and “n50” sequentially, B). refined model, features selected from “Contig metrics”, “Read mapping metrics”, the top three important features were “contigs_segmented”, “bad_mapping”, and “n50” sequentially.

SUPPLEMENTAL MATERIALS

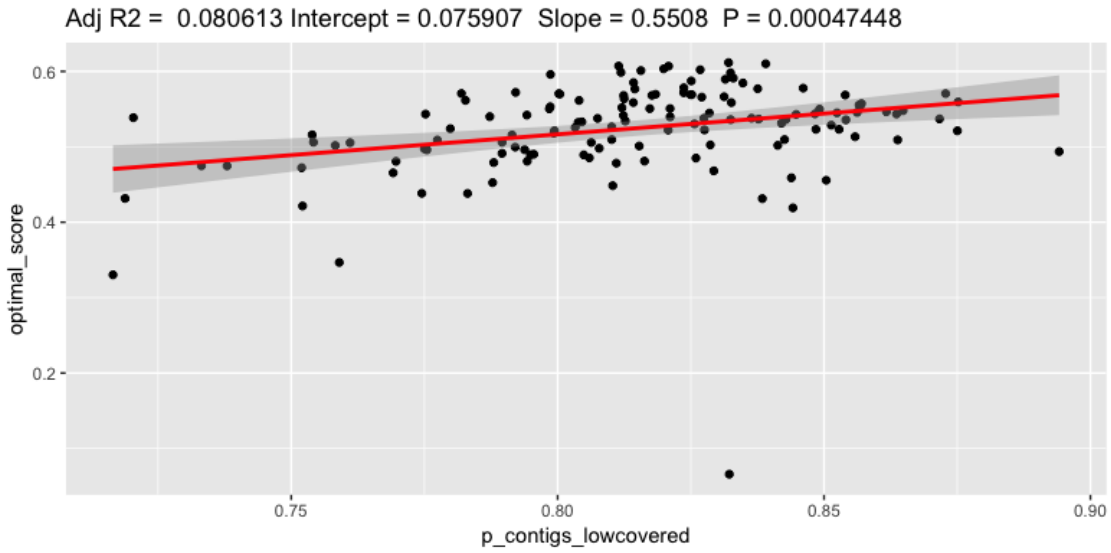


Figure S1A. Correlation between TransRate optimal score and proportion of low covered contigs (lm, slope = 0.55, adj.R^2 = 0.08, P-value = 4.7e-04).

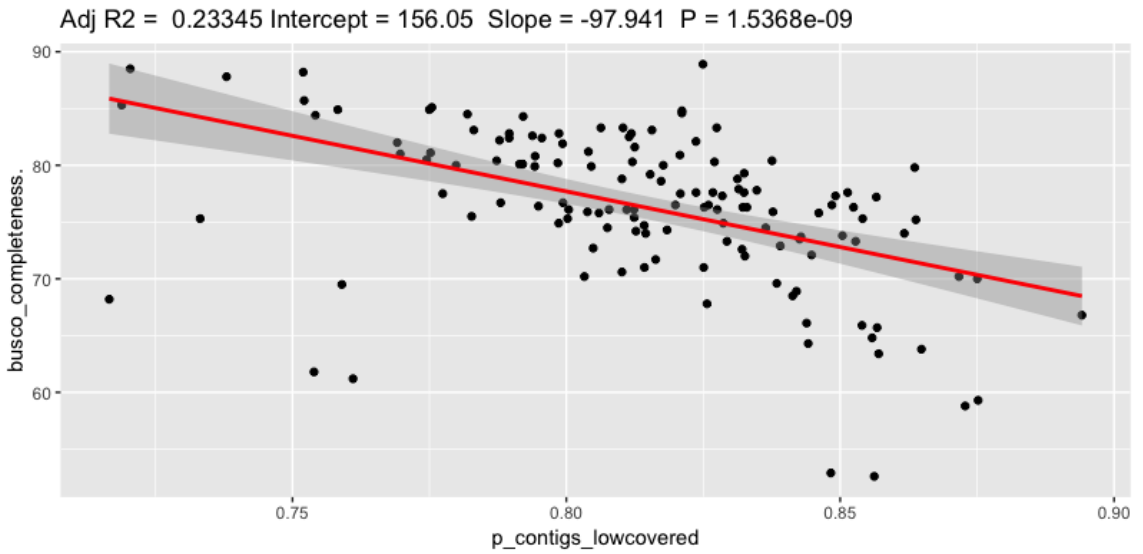


Figure S1B. Correlation between BUSCO completeness score and proportion of low covered contigs (lm, slope = -97.94, adj.R^2 = 0.23, P-value = 1.54e-09).

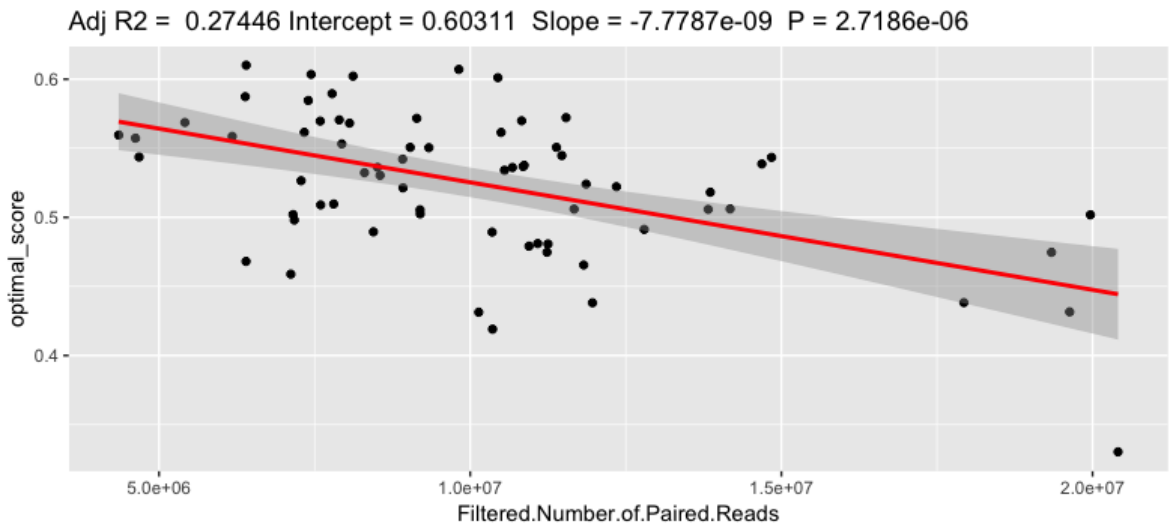


Figure S2A.

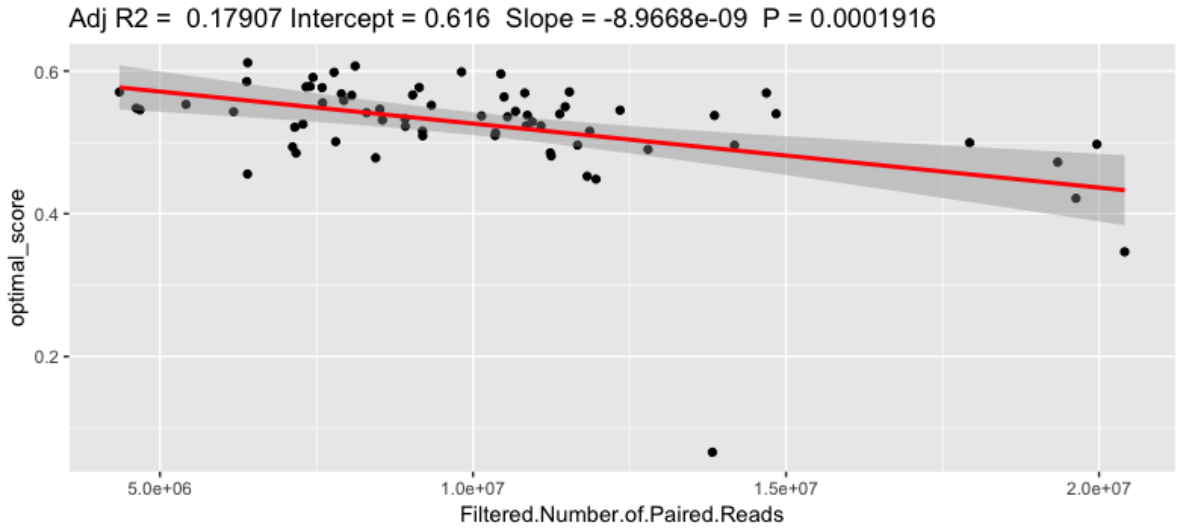


Figure S2B.

Figure S2. The relationship between input read number and TransRate optimal score A). reference guided assembly (lm, slope = -7.78e-09, adj.R^2 = 0.27, p-value = 2.72e-06), B). *de novo* assembly (lm, slope = -8.97e-09, adj.R^2 = 0.18, p-value = 1.9e-04)

Solanum	*Capsicum*
pimpinellifolium	*pubescens*
peruvianum	*pratermissium*
pennellii	*peruvianum*
ochranthum	*minutiflorum*
neorickii	*galapagense*
lycopersicumcer	*eximium*
lycopersicum	*chinense*
lycopersicoides	*chacoense*
huaylasense	*baccatum*
habrochaites	*annuum*
galapagense	
corneliomuelleri	
chmielewskii	
chilense	
cheesmanieae	
arcanum	

Table S1. Species that were sampled for sequencing and transcriptome assembly, 15 and 10 species were selected from *Solanum* genus and *Capsicum* genus respectively.

REFERENCES

Andrews, S. (2016). FastQC: a quality control tool for high throughput sequence data. 2010.

Bates, D., et al. (2007). "The lme4 package." R package version **2**(1): 74.

Bhardwaj, A., et al. (2018). RNA-seq based mapping strategies to uncover heterogeneity in survival among Pancreatic Ductal Adenocarcinoma (PDAC) patients. Proceedings of the 9th International Conference on Computational Systems-Biology and Bioinformatics, ACM.

Bolger, A. M., et al. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120.

Bombarely, A., et al. (2012). "Deciphering the complex leaf transcriptome of the allotetraploid species Nicotiana tabacum: a phylogenomic perspective." BMC genomics **13**(1): 406.

Breiman, L. (2001). "Random forests." Machine learning **45**(1): 5-32.

Bushmanova, E., et al. (2016). "rnaQUAST: a quality assessment tool for de novo transcriptome assemblies." Bioinformatics **32**(14): 2210-2212.

Bushmanova, E., et al. (2018). "rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data." bioRxiv: 420208.

Cahais, V., et al. (2012). "Reference-free transcriptome assembly in non-model animals from next-generation sequencing data." Molecular ecology resources **12**(5): 834-845.

Castel, S. E., et al. (2015). "Tools and best practices for data processing in allelic expression analysis." Genome biology **16**(1): 195.

Codina-Fauteux, V.-A., et al. (2018). "PHACTR1 splicing isoforms and eQTLs in atherosclerosis-relevant human cells." BMC medical genetics **19**(1): 97.

Collins, L. J., et al. (2008). An approach to transcriptome analysis of non-model organisms using short-read sequences. Genome Informatics 2008: Genome Informatics Series Vol. 21, World Scientific**:** 3-14.

Conesa, A., et al. (2016). "A survey of best practices for RNA-seq data analysis." Genome biology **17**(1): 13.

Dobin, A., et al. (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.

Eid, J., et al. (2009). "Real-time DNA sequencing from single polymerase molecules." Science **323**(5910): 133-138.

Fan, J., et al. (2018). "Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data." Genome research **28**(8): 1217-1227.

Fernandez-Pozo, N., et al. (2014). "The Sol Genomics Network (SGN)—from genotype to phenotype to breeding." Nucleic acids research **43**(D1): D1036-D1041.

Grabherr, M. G., et al. (2011). "Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data." Nature biotechnology **29**(7): 644.

Guttman, M., et al. (2010). "Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs." Nature biotechnology **28**(5): 503.

Herman, J. S. and D. Grün (2018). "FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data." Nature methods **15**(5): 379.

Huang, X., et al. (2016). "Comparative performance of transcriptome assembly methods for non-model organisms." BMC genomics **17**(1): 523.

Jones, J. W., et al. (2018). "Differential Gene Expression and Pathway Analysis in Juvenile Nasopharyngeal Angiofibroma Using RNA Sequencing." Otolaryngology–Head and Neck Surgery **159**(3): 572-575.

Li, B., et al. (2014). "Evaluation of de novo transcriptome assemblies from RNA-Seq data." Genome biology **15**(12): 553.

Mazerolle, M. J. and M. M. J. Mazerolle (2019). "Package 'AICcmodavg'."

Pan, X., et al. (2018). "WebCircRNA: Classifying the circular RNA potential of coding and noncoding RNA." Genes **9**(11): 536.

Pedregosa, F., et al. (2011). "Scikit-learn: Machine learning in Python." Journal of machine learning research **12**(Oct): 2825-2830.

Rätsch, G., et al. (2007). "Improving the Caenorhabditis elegans genome annotation using machine learning." PLoS Computational Biology **3**(2): e20.

Reinbolt, R. E., et al. (2018). "Genomic risk prediction of aromatase inhibitor‐related arthralgia in patients with breast cancer using a novel machine‐learning algorithm." Cancer medicine **7**(1): 240-253.

Song, Q. A., et al. (2018). "Computational analysis of alternative splicing in plant genomes." Gene.

Signal, B., et al. (2017). "Machine learning annotation of human branchpoints." Bioinformatics **34**(6): 920-927.

Simão, F. A., et al. (2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." Bioinformatics **31**(19): 3210-3212.

Smith-Unna, R., et al. (2016). "TransRate: reference-free quality assessment of de novo transcriptome assemblies." Genome research **26**(8): 1134-1144.

Sonnenburg, S., et al. (2002). New methods for splice site recognition. International Conference on Artificial Neural Networks, Springer.

T O'Neil, S. and S. J. Emrich (2013). "Assessing De Novo transcriptome assembly metrics for consistency and utility." BMC genomics **14**(1): 465.

Team, R. C. (2013). "R: A language and environment for statistical computing."

Trapnell, C., et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature biotechnology **28**(5): 511.

van Galen, P., et al. (2019). "Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity." Cell **176**(6): 1265-1281. e1224.

Vijay, N., et al. (2013). "Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments." Molecular ecology **22**(3): 620-634.

Wang, Z., et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature reviews genetics **10**(1): 57.

Zdobnov, E. M., et al. (2016). "OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs." Nucleic acids research **45**(D1): D744-D749.

Chapter 3. Transcriptome-wide Identification of 'Cross-talk' Genes on Solanum *pimpinellifolium* in Biotic and Abiotic Stress Response

Chenming Cui[1], David C. Haak[1*]

1. School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061 USA

Chenming Cui: chenmc1@vt.edu

David Haak: dhaak@vt.edu

*To whom correspondence should be addressed: dhaak@vt.edu

ABSTRACT

Understanding how plants integrate responses to environmental stressors can provide critical insight on plant resilience. Antagonism between abiotic and biotic stress signaling pathways is the result of a complex interaction between abscisic acid and jasmonic acid cascades. Here we leverage intraspecific natural variation in induced defense responses between two wild tomato accessions, LA1589 and LA1269 to examine changes in gene expression under methyljasmonate (JA), drought stress (DS), and the combined (JADS) stresses. We find that gene expression profiles for our JA treatment match the phenotypic responses. In accession LA1589, which exhibited a week induced defense response, just 137 genes were differentially expressed, in contrast 1,204 were identified in LA1269, an accession with strong inducibility. Supporting our phenotypic data, GO annotation revealed that the differentially expressed genes in accession LA1269 were significantly enriched for terms related to defense while LA1589 was not. Unsurprisingly, drought stress dominated differential gene expression in both accessions. Network analysis of gene expression from the JADS treatment in LA1269 revealed that drought stress invokes transcription factors associated with gene expression changes that regulate jasmonic acid signaling. Together, these suggest that abiotic stress exerts multilevel regulation over biotic stress responses.

INTRODUCTION

The Solanaceae contains a number of major crops, including chili peppers, tomato, potato, tobacco, and petunia. Tomato, *Solanum lycopersicum*, (formerly *Lycopersicon esculentum*), is one of the most important vegetables in the world (Haliński and

Stepnowski 2016) and is well known for its nutritional value, taste, and texture. However, they are susceptible to various diseases that results in annual yield losses (Wang, Vinocur et al. 2003). Adding in the additional effects of drought stress, annual crop losses are estimated at hundreds of billions of dollars. Mitigating these effects via adapted crops relies on altered management practices and suitable genetic variation for novel resistance or tolerance. Yet, the domestication of many crop plants has narrowed available genetic diversity (Tanksley 2004). Wild crop relatives are a source of natural variation for crop improvement (Gur and Zamir 2004). The dissection of genomic variation between domesticated crops and their wild progenitors can provide the identification of alternative alleles for agronomics traits, insights into the history of domestication, breeding potential, local adaptation processes (Gepts 2002, Weigel and Nordborg 2005), and the molecular basis of plant immunity (Seo, Kim et al. 2016). Genomics-assisted breeding, which routinely turned to wild relatives as an additional source of genetic variation, facilitates the rapid development of improved cultivars through the identification of novel agronomic genomic regions associated with agronomic traits.

Both biotic and abiotic factors contribute to tomato differentiation and adaptation to the large spectrum of environments they inhabit (Peralta, Spooner et al. 2008, Haak, Ballenger et al. 2014). In order to tolerate these stressors, tomatoes have acquired adaptive traits over the course of evolution and domestication (Rejeb, Pastor et al. 2014). Biotic and abiotic stresses alter plant metabolism, by eliciting the activation of stress response pathways, therein causing decreases in yield and fitness (Heil and Bostock 2002, Shao, Chu et al. 2008, Bolton 2009, Massad, Dyer et al. 2012). Abiotic stresses such as extreme temperatures, drought, and salinity, are a major cause of yield loss

globally, accounting for an estimated 50% of annual losses of major crops (Wang, Vinocur et al. 2003). Abiotic stress invokes abscisic acid (ABA) as a signaling molecule which stimulates wide ranging interconnections in genetic, physiological, and biochemical responses (Wang, Vinocur et al. 2000). For instance, water deficit results in osmotic stress in plant cells and consequentially disrupts homeostasis resulting in cellular damage (Zhu, 2001). Similarly, biotic stress, such as herbivory or pathogen infection, invokes jasmonic acid (JA) as a signaling molecule which results in a cascade that includes genetic and physiological shifts leading to the production of defensive compounds (Maron and Kauffman, 2006; Strauss and Zangerl, 2002; Brown and Hovmoller, 2002). In real-world environments, for both natural and agricultural ecosystems, plants must integrate their responses to simultaneous stressors (Mauck, Smyers et al. 2015). Thus, an intricate 'crosstalk' scenario occurs in plants experiencing a combination of biotic and abiotic stress. Though some mechanisms of how individual stressors impact plants have been uncovered (Qin, Shinozaki et al. 2011, Thakur and Sohal 2013), it is still not clear when interactions among pathways are antagonistic, synergistic, or additive and under what conditions it may move through this continuum (Anderson, Badruzsaufari et al. 2004). Interestingly, responses under combined biotic and abiotic stress cannot be predicted from plant responses under a single stress (Atkinson and Urwin 2012), leading to the hypothesis that 'crosstalk' itself is the product of adaptive evolution (Thaler, Humphrey et al. 2012). Here we leverage the tomato system to identify the genomic components of drought stress and jasmonic acid signaling pathways interaction.

Cultivated tomato (*S*. *lycopersicum*) and its 13 wild relatives are native to the west slope of the Andes spanning 27 degrees latitude in south America, from northern Ecuador to

Chile, and the Galapagos Islands (Haak, Ballenger et al. 2014). These wild relatives have long been a source of genes for improving domesticated tomato. This is particularly true for Solanum *pimpinellifolium*, one of the closest wild relatives of the cultivated tomato (Blanca, Cañizares et al. 2012, Consortium, Aflitos et al. 2014) that is adapted to very diverse local environments, such as tropical rainforests in northern coastal Ecuadorian and desert in coastal Peru (Zuriaga, Blanca et al. 2009). Various traits of S. *pimpinellifolium* appear to have been lost for cultivated tomato during the process of domestication (Bai and Lindhout 2007, Razali, Bougouffa et al. 2018). Therefore, *S. pimpinellifolium* has important genetic variation that can potentially improve domesticated tomato. Indeed, *S. pimpinellifolium* is well known as the source of the *PTO* pathogen resistance gene that was introgressed into tomato nearly 40 years ago (Pedley and Martin 2003). Additional traits improved by using resources from *S. pimpinellifolium*, resistance genes for spider mite (Salinas, Capel et al. 2013), antioxidant traits (Top, Bar et al. 2014), and heat resistance genes (Lin, Yeh et al. 2010).

To unravel the genomic basis of “cross-talk” and identify genes involved under combined biotic and abiotic stress, we selected two populations of *S. pimpinellifolium* that were about 500 meters apart and used experimental transcriptomics under four conditions, control, simulated herbivory, drought stress, and the combined stresses. We mimic herbivory responses using methyljasmonate, the active form of jasmonic acid (JA), an important signaling molecule in plants (Baldwin 2001, Wasternack 2007), that can be applied to elicit a replicable induced herbivory defense response in Solanaceae (Farmer and Ryan 1992, Haak, Ballenger et al. 2014). Drought stress (DS) was imposed by holding soil moisture to 25% field capacity. In a combined treatment both stresses were

imposed with JA treatment preceding the imposition of DS. The tobacco hornworm *Manduca sexta,* a natural specialist for Solanaceae, which elicits constitutive herbivore signaling (Kawahara et al. 2009; David Haak, 2014) can be utilized together with JA induction to study plant constitutive and plastic herbivore stress response.

RESULTS

Differential constitutive resistance to herbivory

To estimate defense responses we use a bioassay, the tobacco hornworm *Manduca sexta,* a natural specialist for Solanaceae, that can be used to estimate levels of constitutive and induced defenses (Kawahara et al. 2009; David Haak, 2014). Plants of accession LA1589 and LA1269 were grown under stress treatment of the Jasmonic acid (JA), drought stress (DS) by holding water to 25% field capacity, and the combined conditions of JA and DS (JADS) to the 8-10 leaf stage. Well-watered plants were set as control group in this study. For each of five replicate plants from each condition, leaves were removed weighed and placed in GladWare containers with perforated lids. Ten 2nd instar larvae were then allowed to feed on the leaves for 48 hours. Caterpillar initial and post-feeding mass (final mass) were collected. The mean values from the larvae were used in the calculation of relative growth rate (RGR) using the formula: log (final mass / initial mass). As a result, we quantitatively measured plants resistance strength by 1-RGR (supplemental material). Both accessions exhibited similar constitutive resistance levels (light grey bars on LA1269 and LA1589: 0.77 and 0.75, respectively), however, accession LA1269 showed a much stronger level of induced defense, while LA1589 appeared to be insensitive to our JA treatment (Figure 1, LMM $p < 0.05$). Plants from accession LA1269

treated with JA, showed an increase in resistance from 0.78 to 0.95 (light grey and brown bars on the left); whereas in accession LA1589 the non-significant mean increase was from 0.76 to 0.78 on accession LA1589 (light grey and brown bars on the right). 2). Under drought stress alone, defense levels were not significantly different from constitutive resistance levels for either accession (Figure 1, dark grey bars). The combined stress (JADS) differentially impacted the two accessions (Figure 1, light red bars). Accession LA1269 showed a 20% decline in resistance when compared to JA induced resistance only. Conversely, LA1589 had no significant change in resistance under the co-stress treatment (JADS) 3). Overall there was a significant interaction of JADS between accessions that was driven by the attenuation of the induced defense response in accession LA1269.

Drought stress dominates transcriptome wide stress responses

RNA-Seq experiments were conducted to characterize transcriptome wide gene expression patterns in response to individual and combined stresses. Expression analysis largely reflected our phenotypic data wherein accession LA1589 was less sensitive to exogenous JA treatment compared to accession LA1269 showing much greater changes in overall gene expression. Filtering differentially expressed genes (DEG) at 2-fold log change and q-value of 0.05, we identified 137 DEGs (25 were up-regulated, 112 of them were down-regulated in JA treatment) from 3 biological replicates consisting of 4-5 plants each of accession LA1589 (Figure 2A) and 1204 (447 were up-regulated, 757 were down-regulated in JA treatment) DEGs in accession LA1269 (Figure 2B). However, both accessions extensively responded to the DS and JADS treatments. In LA1589, we detected 7366 (3689 were up-regulated, 3677 were down-regulated) and 8032 (4029

were up-regulated, 4003 were down-regulated) DEG respectively (Figure 2A). In LA1269 we identified 4947 (2149 were up-regulated, 2798 were down-regulated) and 6874 (3001 were up-regulated, 3873 were down-regulated) DEGs in DS and JADS treatments respectively. (Figure 2B).

To explore gene expression patterns across treatments, we plotted the 50 most highly expressed genes (Figure 3). Similar expression patterns were captured for plants with (DS and JADS) or without DS (WW and JA) in the top 50 expressed genes of accession LA1589 (Figure 3A). Similar patterns were observed for accession LA1269 as well (Figure 3B). Together these show that the DS treatment dominated the expression profiles for both accessions. This was supported by principle component analysis (PCA) wherein the separation of DS and JADS along PC1 explained the majority of the variance for both accessions with LA1589 at 83.1% and LA1269 at 72.3% (Figure 4A and B). Interestingly, we did not observe the same pattern for plants under the JA treatment. Principle component analysis (PCA) of accession LA1589 revealed little discrimination between JA and control (WW) conditions, with one replicate, JA_B, clustering within WW replicates (Figure 4A). In contrast, despite only having two biological replicates, treatments were more clearly separated for accession LA1269 (Figure 4B). These differences are reflected in the corresponding PC2 axes where just 4.5% of the variation in expression patterns was captured for LA1589 in contrast to the 17.7% captured for LA1269.

Expression patterns for genes in response to treatments and the interfaces are shown in Figure 5. Of the total 9109 DEGs identified in LA1589 ca. 81% (7366) were identified in response to DS. Of these 6327 (ca. 86% of DS DEGs) were shared between the JA, DS, and JADS treatments (Figure 5A). DEG shared by both DS and JADS treatments

dominated this set with 6245. Only 82 were shared between JA, DS, and JADS. Interestingly, just ca. 28% (38) of the JA induced DEGs were unique to the JA treatment for this accession, while nearly 60% (82) were shared across all three conditions. Similarly, of the 8032 DEGs associated with the combined stress (JADS) ca. 21% (1705) were uniquely expressed in the JADS condition, with ca. 79% of DEGs shared with DS (6327).

Similarly, DS dominated the DEG patterns in accession LA1269 but not at the same magnitude. The interface between DS and JADS (4532 DEGs) for LA1269 (Figure 5B) contributed to ~92% and ~66% of single DS and combined JADS induced DEGs. While only ca. 1.5% of DEG were associated with JA in LA1589, nearly 16% (1204) of all DEG were associated with the JA treatment in LA1269. Of these a similar proportion of DEG ca. 19% (229) were unique to the JA condition. The unique 1185 DEGs of LA1269 (aside from the shared DEGs between LA1589 and LA1269) under JA stress explained its versatility of herbivore resistance.

Co-expression analysis and GO term enrichment

To capture treatment-specific co-expression profiles, we performed a cluster analysis using WGCNA1.67 (Zhang and Horvath 2005) in R. Using correlations among co-expressed genes and dynamic tree cut analysis, we identified various modules under each treatment as shown in Figure 6. This approach allowed us to identify modules whose expression patterns were very similar, and then we choose a height cut of 0.25 on the dendrogram to merge modules whose correlation was above 0.75. We retained modules that were highly correlated with a significance level of $p < 0.05$ across treatment

conditions. We detected two modules ($p < 0.05$) from LA1589 DS (Figure 6A), blue2 module ($r = -0.96$, $p = 0.003$) and thistle module ($r = 0.95$, $p = 0.004$); one module ($p < 0.05$) from LA1589 JA (Figure 6B), turquoise ($r = -0.95$, $p = 0.004$) and; two modules ($p < 0.05$) from LA1589 JADS (Figure 6C), firebrick4 module ($r = 0.99$, $p = 1e-04$) and maroon module ($r = -0.99$, $p = 1e-04$). Similarly, we detected two modules ($p < 0.05$) from LA1269 DS (Figure 6D), greenyellow module ($r = 0.98$, $p = 0.02$) and darkmagenta module ($r = -0.97$, $p = 0.03$); one module ($p < 0.05$) from LA1269 JA (Figure 6E), red module ($r = 1$, $p = 0.001$); two modules ($p < 0.05$) from LA1269 JADS(Figure 6F), sienna3 module ($r = -0.98$, $p = 0.02$) and magenta module ($r = 1$, $p = 0.004$).

To determine gene ontology (GO) enrichment in modules, and compare GO terms across modules, conditions and accessions, we performed GO enrichment analysis using Metascape (Zhou, Zhou et al. 2019). 1). Comparing the GO terms that were enriched under all the conditions of LA1589, we discovered that RNA transcription and translation related GO terms were frequently enriched in the blue module (Figure 7). For example, ‘ribosomal large subunit biogenesis’ (GO: 0042273), ‘cleavage involved in rRNA processing’ (GO: 0000469), and ‘mRNA surveillance pathway’ (GO: 0071028), were enriched in this module. Not surprisingly, such RNA related GO terms were found in the JADS maroon module only (we did not find these in the JADS firebrick4 module), such as ‘regulation of mRNA metabolic process’ (GO: 1903311), ‘production of siRNA involved in RNA interference’ (GO: 0030422). Defense related GO terms greatly enriched in firebrick4 module, for example we observed ‘defense response to other organism’ (GO: 0098542) and ‘immune effector process’ (GO:0002252). Meanwhile, water efficiency related GO terms (GO:0009414, GO:0010118) were enriched in the firebrick4 module as

well. Interestingly, both firebrick4 and maroon modules were clustered under JADS condition, but maroon module was very similar with blue module of DS, representing large basic biochemistry metabolism (RNA activities' related GO enrichment). Whereas, the firebrick4 module was enriched with more defense related GO terms. We did not identify any functional enrichment of GO terms for the JA condition module.

Comparing the GO terms that were enriched under all the conditions of LA1269 we found similar enrichment patterns (Figure 7). We captured water use efficiency and abiotic related GO enrichment in greenyellow and darkmagenta modules, for example 'response to abscisic acid' (GO: 0009737), 'response to osmotic stress'(GO:0006970), 'cellular response to abiotic stimulus' (GO: 0071214). Interestingly, abiotic defense related GO terms were captured in the red module of JA, for example 'response to heat' (GO:0009408), 'heat acclimation' (GO:0010286), though biotic defense related GO terms were enriched as well (GO:0010337, 1900424GO). In the JADS condition, we observed immune (GO:0002376) and biotic stress (GO: 0050829, GO:0002237) related GO terms enriched in the sienna3 and magenta modules.

Comparing the GO terms that were enriched across accessions we found that the DS and JADS conditions exhibited diverged GO enrichment. For example, the top three significantly ($-\log_{10}$(p-value)>6) enriched GO terms in DS treatment were "ribosome, eukaryotes"(KEGG module: M00177), "ribosomal large subunit biogenesis" (GO:0042273) and "anchored component of plasma membrane (GO:0046658) for LA1589 (Figure 7A) and "photosynthetic electron transport chain" (GO:0009767), "regulation of response to stimulus" (GO:0048583) and "response to abscisic acid" for LA1269 (Figure 7D). Under JADS combined stress, "plastid organization" (GO:0009657), "organelle inner membrane"

(GO:0019866) and “chloroplast stroma” (GO:0009570) were the top three significantly ($-\log_{10}$(p-value)>10) enriched for LA1589 (Figure 7C), whereas, “cell surface receptor signaling pathway” (GO:0007166), “immune system process” (GO:0002376) and “regulation of response to stimulus” (GO:0048583) were the top three for LA1269 (($-\log_{10}$(p-value)>6, Figure 7F&G). Notably, jasmonic acid signaling related GO term (GO:0009867) was captured specific for LA1269 JADS treatment (((-$\log_{10}$(p-value)>3, Figure 7G). Interestingly, we did not detect any functional enrichment in the JA module for LA1589, but the JA module from LA1269 was enriched with various defense related functions as described above. We queried the term ‘defense’ across several ontology lists (GO biological process, GO cellular components, GO Molecular Functions and so on) in Metascape. The JA module from LA1269 was significantly enriched for defense related functions (Figure 8B, p = 0.0078). Predictably, there was no significant enrichment of defense related functions in the JA module of LA1589 (Figure 8A, p = 0.29). The same analysis was performed across all modules of the two accessions. All of them, with the exception of the DS and JA modules from LA1589, were significantly enriched with defense related GO terms.

Hub genes and KEGG pathways identified in response to ‘cross-talk’

To identify hubs associated with overlap between the defense response pathways we evaluated connections between clusters of inferred function for the JADS condition from each accession. We captured highly connected clusters in the maroon module using Metascape (Zhou, Zhou et al. 2019) (Figure 9A), a web-based portal that provides gene list annotation and classification. Clusters of “organelle inner membrane”, “intracellular protein transmembrane transport”, “chloroplast envelope” and “mitochondrion

organization" were identified. The only cluster connecting the three others was "mRNA metabolic process". Clusters from the firebrick4 module (Figure 9B) exhibited few connections, although 'transport related' clusters strongly interacted. Defense related clusters were connected via "regulation of response to stimulus", into an interacting network, specifically, "response to alcohol", "regulation of developmental process", "immune effector process" and "defense response to other organism" were directly or indirectly linked by the red node.

Similar clustering analysis was conducted on LA1269 as well. We found that clusters of "cellular response acid chemical", "regulation of response to stimulus", "response to osmotic stress" and "jasmonic acid mediated signaling pathway" were highly connected in magenta module (Figure 9C). Notably, such jasmonic acid signaling related cluster was not identified in LA1589's. The rest of the clusters in magenta module were relatively independent such as "regulation of developmental process", "leaf development" (Figure 9C). Fewer clusters were identified in the sienna3 module (Figure 9D), the only dense connections centered at "inorganic molecular entity transmembrane transporter", linking with "inorganic ion homeostasis" and "regulation of hormone metabolic process". Some other defense related clusters "immune system process" and "defense response to Gram-negative bacterium" were also identified.

Additionally, we identified predicted protein interaction patterns using Metascape (Zhou, Zhou et al. 2019) from modules of JADS for both of the two accessions, using the identified gene list and the database BioGrid, Metascape implemented the MCODE algorithm to identify densely connected network components (Figure 10). Independent interaction networks were assigned unique MCODE ids and colors for the component in

the network. To understand the biological process that these protein interactions might involve, we carried out GO enrichment analysis for each MCODE. Three MCODEs were detected with enrichment in maroon module's protein interaction networks. We can see these protein interactions integrated with plant development (MCODE 3,6) and RNA metabolism (MCODE 5) (Figure 10A). Whereas plant hormone signal transduction associated with ABA and water use efficiency terms were identified in Friebrick4 module protein interactions (Figure 10B). Interestingly, MCODE that were identified from LA1269 (magenta module) were exclusively associated with the jasmonic acid response. The relevant genes "JAZ1", "JAZ3" and "JAZ8" were recovered in MCODE3 (Figure 10C). The only MCODE we identified from sienna3 module was functionally associated with methylation and demethylation process (Figure 10D).

To identify the gene pathways involved under the JADS treatment we mapped gene profiles (DEGs from JADS of both accessions) onto KEGG pathways to illustrate the biological functions involved. This analysis identified genes associated with carbohydrate metabolism, energy metabolism, lipid metabolism, nucleotide metabolism, amino acid metabolism, metabolism of cofactors and vitamins and biosynthesis of secondary metabolites in both accessions (Table 2). These represent a host of activities associated with drought responses such as ABA metabolism, for example, we have DEGs from JADS of LA1589 and LA1269 mapped onto the Fatty Acid biosynthesis pathway (Figure 11A&B; Table 3). Interestingly, cutin, suberin and wax biosynthesis pathways were incorporated as well (Figure 11C&D; Table 3). Here we see differential expression differences between LA1589 and LA1269, as evidenced by the difference in color of the same key enzymatic steps (Figure 11B).

DISCUSSION

The advent of high-throughput gene expression profiling is enabling the development of experimental designs aimed at elucidating the genes involved in complex responses such as plant stress responses. Identifying the genes involved in these pathways, and in particular the genes involved in pathway interactions, is paramount to developing resilient crop plants. Here we have identified two accessions of *S. pimpinellifolium*, LA1589 and LA1269, with differential responses to the exogenous application of methyljasmonate. Leveraging this intraspecific variation, we have used RNASeq in a factorial experiment to identify genes, networks and pathways associated with induced defense, drought stress, and the combination of these stressors. Using these two different accessions has allowed us to identify key genetic targets within these stress response pathways that generate, in this case, antagonist interactions between defense and drought response pathways.

Natural populations of *S. pimpinellifolium*, are found across diverse natural environments, including gradients in abiotic and biotic conditions. In this study, we found that accession LA1269 mounted induced defense responses when treated with methyljasmonate (meJA) and that this defense response was attenuated when co-stressed with drought. In contrast, accession LA1589 did not mount a meJA associated defense response and was relative treated (Figure 1). Thus, we predicted that these two accessions would exhibit differential gene expression profiles and provide an opportunity to dissect the genes associated with the 'cross-talk' between JA and ABA signaling pathways.

Differential meJA induced expression profiles

Consistent with the phenotypic data (Figure 1), the number of DEG associated with meJA

treatment was 10-fold greater in LA1269 than in LA1589, suggesting that there is an expression network difference between these accessions. Indeed, co-expression network analysis clustered the 137 DEG from LA1589 into a module, however enrichment analysis did not reveal any significantly enriched defense genes. In contrast, the JA associated co-expression module for LA1269 was significantly enriched for defense related GO terms. This suggests that LA1589 lacks some component of induced defense responses while maintaining some key targets in JA signaling. While JA signaling is often associated with secondary metabolism for plant defense, it also has well documented signaling roles in other plant processes such as stomatal opening and closing (Junttila, Li et al. 2008) Intriguingly, our analysis identified a network of JAZ genes in LA1269 (Figure 10C) that was not found in the analysis of LA1589. The JAZ proteins play a central role in the JA signaling cascade and these candidate genes may provide important insights into the moderated defense response seen in LA1589. An alternative explanation is that this could result from insufficient treatment, however this scenario is unlikely for two reasons, first we do detect DEG with roughly the same proportionate overlap across conditions as found in LA1269, and secondly plants were treated with JA prior to randomization to drought stress and we have identified a similar number of JADS DEG between accessions. Finally, we know from phenotypic data that other species within this clade also appear to lack an induced defense response (Haak et al. 2014).

Drought stress dominates expression profiles

As with previous studies (Shaik and Ramakrishna 2013, Lu, Zhou et al. 2017, Xiao, Hu et al. 2019), the imposition of drought stress resulted in the greatest change in expression

profiles, for both accessions. This was also true for the JADS treatment as evidenced by PCA clustering along the first principle component, where the JADS treatments tended to cluster with the DS treatments. Network analysis in this cluster revealed modules with significantly enriched GO terms associated with desiccation tolerance, membrane integrity, and cellular reprogramming. Connecting these DEGs to metabolic pathways gives us some insight on the complicated signaling under this condition. For example, we find that in both accessions predicted protein-protein interaction networks capture relevant ABA signaling processes and growth/developmental shifts. In LA 1589, there is a cluster representing ABA binding including ABA associated binding factors, ABF2 and ABF4. ABFs are widely recognized as central transcription factors in ABA signaling that regulate gene expression under drought responses (Haak, Fukao et al. 2017). Similarly, in LA1269 a cluster nuclear factor Y (NF-Y) transcription factors was identified. In Arabidopsis, the NF-Y gene family is responsive to ABA signaling and are (generally) associated with promoting drought tolerance, growth and development (Zhao, Wu et al. 2017). In fact, NF-YC which represent 2 of the 3 NF-Y genes identified here have been manipulated in *Paspalum* to enhance drought and salt tolerance (Wu, Shi et al. 2018). Thus, in both accessions we see broad abiotic stress responses that are commonly associated with drought. While many of the major gene families involved here are evolutionarily conserved (Haak et al. 2017), the diversity of responses between accessions could be leveraged for detailed studies of drought response mechanisms.

Attenuation of the JA response in co-stressed LA1269

Identifying the gene networks associated with signaling pathway interactions is an

important step in developing resilient crops. Here we used our co-stressed (JADS) treatment in accession LA1269 to identify gene expression changes associated with attenuation of the JA mediated defense response. The putative protein-protein interaction network analysis revealed a key hub of JAZ genes, namely JAZ 1, 3, and 8. JAZ genes are a well-known family of repressors that regulate JA signaling, where in the depressed state JAZ expression declines and JA signaling proceeds (Grunewald, Vanholme et al. 2009). Interestingly, within this network the JAZ genes were interacting with AT4G00870 (BHLH14), a basic helix loop helix transcription factor that is also a negative regulator of JA signaling (Song, Qi et al. 2013). This evidence suggests that ABA mediated JA signaling pathway antagonism is happening very early in the pathway. Consistent with this view, we found additional transcription factors associated with negative regulation of JA signaling, namely the Teosinte branched1 / Cycloidea / Proliferating cell factor 1 (TCP) family of transcription factors (Danisman, Van der Wal et al. 2012). Here we find a network that includes 3 TCP transcription factors, TCP 4, 8, 20. Thus, we see potential multilevel antagonistic regulation of JA defense signaling when accession LA1269 is co-stressed.

CONCLUSION

Plant responses to multiple environmental stressors are integrated through complex signaling pathways. We leveraged RNAseq to identify the gene co-expression networks associated with differential pathway induction. Using modular co-expression analysis to retain information but reduce dimensionality we were able to identify key expression networks associated with environmental stress responses. Importantly, we confirmed that observational differences in phenotype are underpinned by concordant shifts in gene

expression. Here, important antagonism between signaling pathways that limits plant defense responses under drought stress, was supported by multiple early pathway transcription factors suggesting multilevel regulation of stress pathway interaction.

MATERIALS & METHODS

Plant growth conditions and stress treatments

S. *pimpinellifolium* plants of accessions LA1589 and LA1269 were germinated in a germination chamber (Percival) and then transplanted to a potting soil mix at the seedling stage. Plants were grown under well-watered conditions until the 7-leaf stage when stress treatments were imposed. Plants were randomized to receive: 1. well-watered conditions, 2. Jasmonic acid as 1mM methyljasmonate, 3. drought stress by dry down and holding water at 25% filed capacity (gravimetric determination) or the combination of JA and drought stress. Plants were grown at room temperature (22℃ – 28℃) with 14-h daylight cycle in a growth chamber at Virginia Tech (Blacksburg, VA, USA). Replicate sets of plants were grown for the bioassay and gene expression studies.

JA preparation and spray

Methyljasmonate (1 mmol/L) was prepared from JA stock solution (predisolved in 100% ethanol; Santa Cruz Biotechnology, Dallas, Texas, USA) and double-distilled water (following, Haak et al. 2014). We sprayed JA on leaves of plants until they were damp. Plants not receiving the JA treatment were sprayed with a control spray that contained double-distilled water and ethanol but no methyljasmonate. The spray was conducted at the 7-10 leaf stages, at 24 hours and 48 hours. Afterwards, Air-dried plants were returned

to the growth chamber.

Bioassay of herbivore stress defense

To assay defense levels we used the herbivore *Manduca sexta*. We obtained eggs from Carolina Biological Supply (Burlington, North Carolina, USA) and hatched in rearing chambers. Larvae were fed tobacco hornworm bulk media for seven to eight days. Larvae were kept without food for 3 hours before an initial mass was taken. Leaves from each of ten replicate plants were collected in 5 cups and 2nd instar larvae were added and allowed to feed for 48 hours. Final larvae mass was measured after the 48h feeding period. Relative growth rate (RGR) (Crawley, 2012:570; Haak et al., 2014), determined as the log(final mass/initial mass) was used to estimate caterpillar growth. Hence, plant defense level/strength was estimated as 1-RGR (Morris et al. 2006; David Haak, 2014).

RNA isolation and sequencing

Leaf sample of each individual plant was collected and frozen in liquid nitrogen before RNA extraction. DNeasy plant QIAGEN kit was used to extract RNA according to the manufacturer's protocol. RNA integrity and quality were assessed using NanoDrop 1000 (Thermo Fisher Scientific) and Agilent 2100 BioAnalyzer (Agilent Technologies Inc., Germany). Paired end (150 bp) bar-coded cDNA libraries were prepared for multiplex sequencing on the Illumina HiSeq4000 sequencer (Duke Center for Genomic and Computational Biology, Durham, NC, USA) using the TruSeq Stranded Total Library Prep Kit with ribodepletion (Illumina, San Diego, USA) according to manufacturer's standard protocols.

Sequencing raw read data process

Paired end RNA-Seq data was generated from Illumina HiSeq4000 sequencing and stored in Fastq files. FastQC 0.11.3 (Andrews 2010) was used to perform quality check on all Fastq files in the command line user interface. Raw read trimming was executed using Trimmomatic 0.33 (Bolger, Lohse et al. 2014) with the following arguments in command line tool: "ILLUMINACLIP: TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:150" to remove Illumina adapters and low quality bases. The trimmed, paired end reads were aligned to a reference species genome, S.*lycopersicum* (SL3.0, https://solgenomics.net/) using the STAR 2.4.2 aligner (Dobin, Davis et al. 2013) with default argument setting, retaining only the uniquely mapped reads. The aligned SAM files were converted to BAM format and sorted using SAMtools 1.3 (Li, Handsaker et al. 2009) with arguments of "-Sb" and "sort" respectively. Read counts were generated form the sorted deduplicated BAM files using featureCounts 1.5.0-p2 (Liao, Smyth et al. 2013) with the following arguments: "-p -t exon -g gene_id" and reference SL3.0's annotation gtf (ITAG3.20, https://solgenomics.net/).

Differential Transcriptional Response analysis

Read count tables with 2 and 3 replicates for each condition (WW, DS, JA, JADS) were generated for accessions LA1269 and LA1589, respectively. Transcriptome-wide differential gene expression detection was performed using DESeq2 1.22.2 package (Love, Huber et al. 2014) in R. Counts were filtered for low counts (threshold = 1) and zeros and then normalized using the 'rlogTransform' function in DESeq2. We used a minimal fold change of 2 and adjusted p-value cutoff (FDR) of 0.05 were applied to identify

DEGs among all treatment constrasts (WW and DS, WW and JA, WW and JADS). PCA was employed to explore the transcriptomes across replicates and conditions on the top 1000 genes. To visualize differentially expressed genes (DEGs) at the interfaces among conditions we developed venn diagrams using VennDiagram 1.6.20 package (Chen and Boutros 2011) in R.

Co-expression analysis, GO enrichment analysis and KEGG pathway annotation

To perform condition specific gene co-expression analysis, we employed Weighted Gene Co-expression Network Analysis WGCNA1.67 (Zhang and Horvath 2005) in R. DEG lists from each treatment of the two accessions were subjected to such network analysis. We set 'minModuleSize = 30' to cluster co-expressed genes from the DGE lists (DESeq2 output) of each condition. Only modules with a significant ($p < 0.05$) correlation with our experimental conditions were retained.

These modules were subjected to gene ontology analysis using Metascape (Zhou, Zhou et al. 2019). We carried out enrichment analysis on the background of ontology sources: GO biological process, GO cellular components, Go molecular Functions and KEGG functional sets. Similarly, we used the search term "defense" to query ontology sources as well to determine if "defense" related terms were enriched. Terms were clustered based on their membership similarities, where terms with a similarity > 0.3 were connected by edges. Protein and protein interactions analysis was carried out with a database "BioGrid 3.5" (Stark, Breitkreutz et al. 2006) and the Molecular Complex Detection (MCODE) algorithm (Bader and Hogue 2003) through Metascape. Pathway analysis was conducted to map our module gene profiles to KEGG database pathways

(Kanehisa and Goto 2000) using Pathview 3.9 (Luo and Brouwer 2013). We converted the *Solanum* gene ids to Uniprot accession using NCBI blast (Camacho, Coulouris et al. 2009) and Entrez ids using DAVID 6.8 (Huang, Sherman et al. 2009) in the GO enrichment and pathway analysis.

TABLES & FIGURES

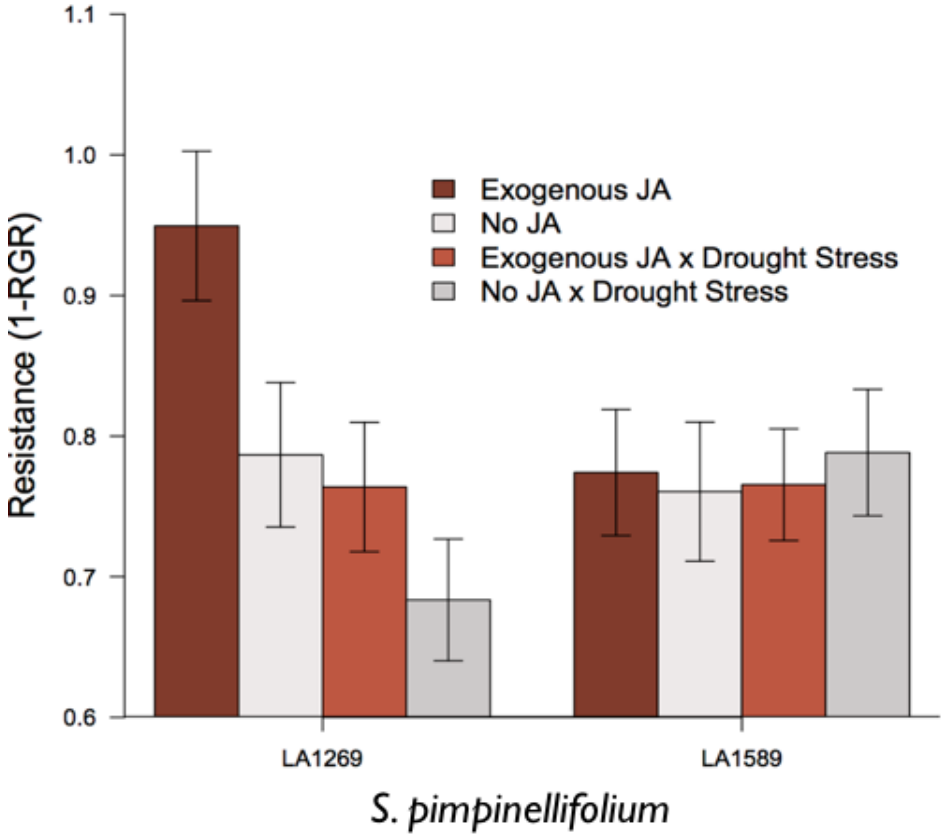


Figure 1. Plants resistance levels (1-RGR) were measured on all replicates of each accessions. Resistance of accession LA1269 was highly inducible with JA, whereas accession LA1589 was not JA sensitive (Brown and light grey bars). Accession LA1269 lost resistance largely due to co-stress JA with DS, whereas accession LA1589 didn't (Brown and red bars).

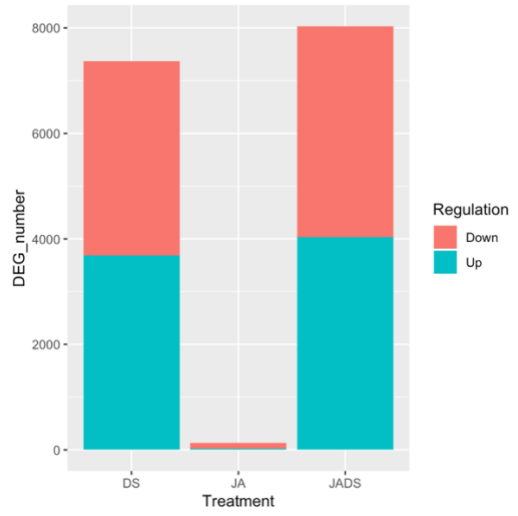


Figure 2A.

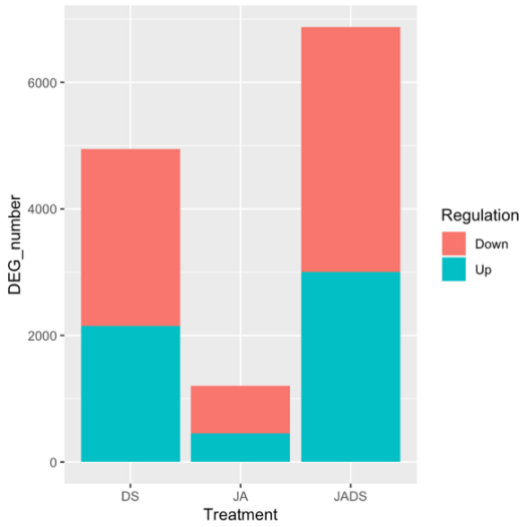


Figure 2B.

Figure 2. Total gene numbers of transcriptome wide DGE in response to DS, JA and JADS for A). accession LA1589 and B). accession LA1269. Red and green proportions of bars denoted for the down and up regulated DEGs in each of the treatment.

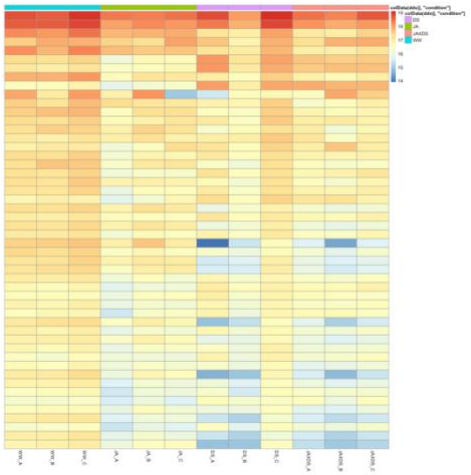


Figure 3A.

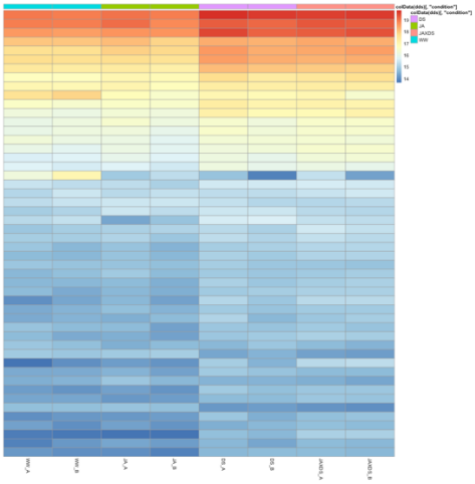


Figure 3B.

Figure 3. Top 50 highly expressed genes on heatmap across replicates and conditions for A.) accession LA1589 and B). accession LA1269. Color key denoted for $\log_2$ normalized count and treatment groups.

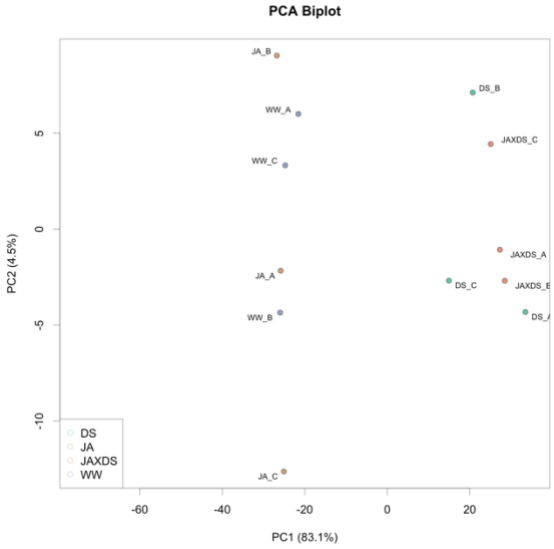


Figure 4A.

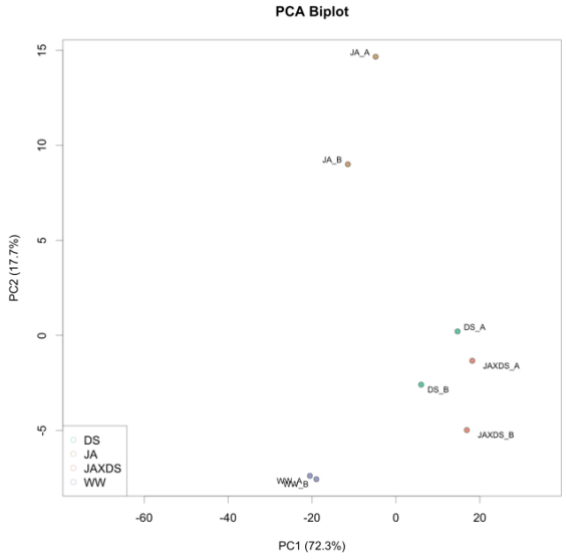


Figure 4B.

Figure 4. First and second principle components (PC) analysis of regularized log transformed gene expression data. Each treatment had 3 biological replicates, denoted

in same color for A). accession LA1589 Each treatment had 2 biological replicates, denoted in same color for B). accession LA1269.

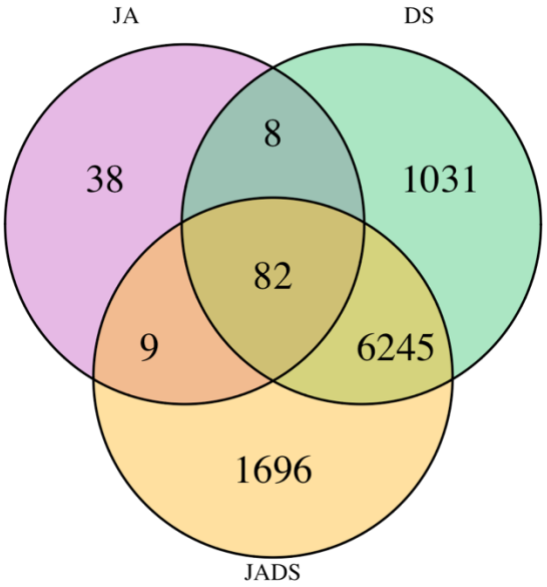


Figure 5A.

Figure 5B.

Figure 5. Three-way Venn diagrams of DEGs on A). accession LA1589 and B). accession LA1269 on the right).

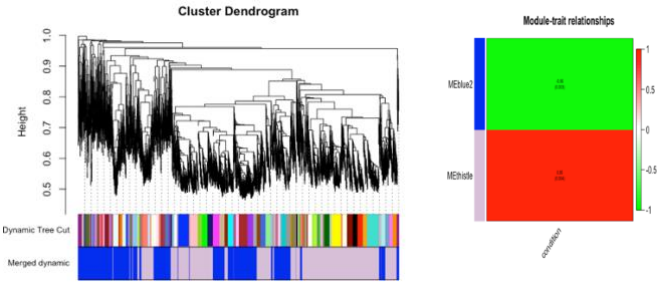


Figure 6A. (LA1589-DS)

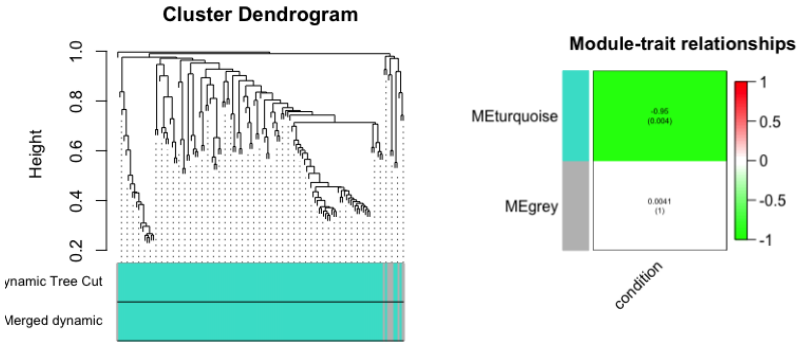


Figure 6B. (LA1589-JA)

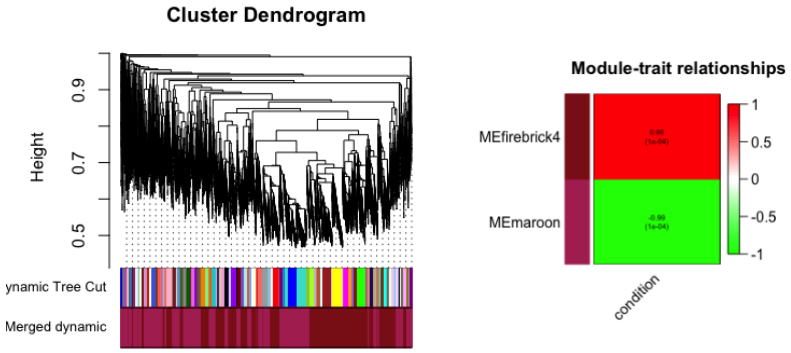


Figure 6C. (LA1589-JADS)

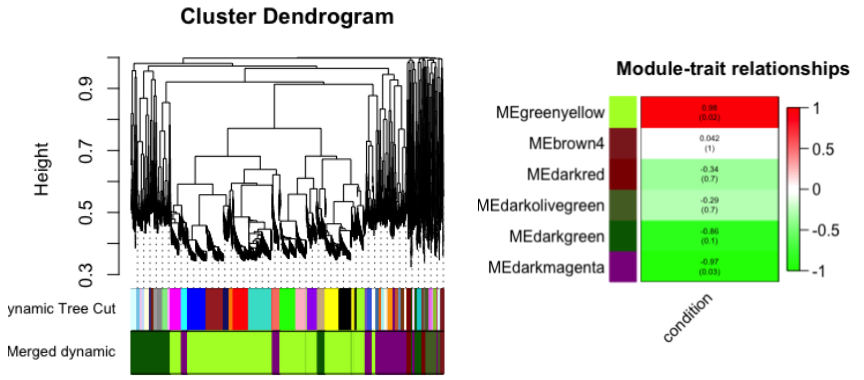


Figure 6D (LA1269-DS)

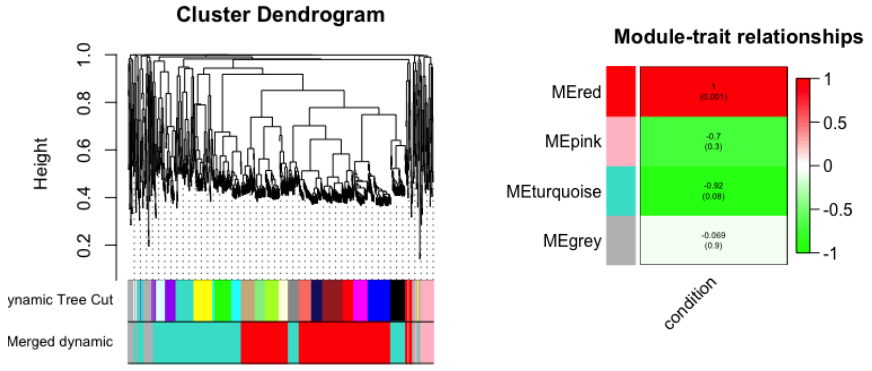


Figure 6E(LA1269-JA)

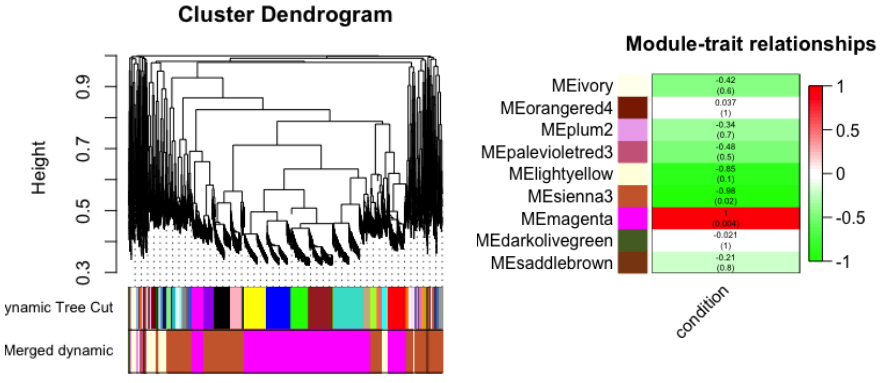


Figure 6F(LA1269-JADS)

Figure 6. Clustering dendrograms from DEGs of each treatment for LA1589 and LA1269. Modules of LA1589 and LA1269 clustered under DS, JA, JADS conditions in A, B, C and D, E, F respectively.

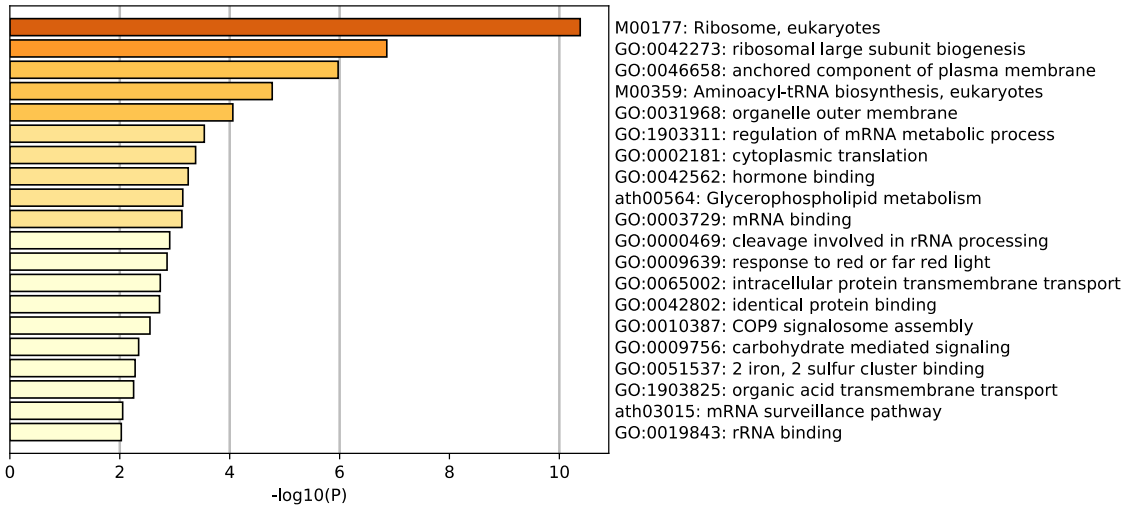


Figure 7A. (Blue module in LA1589 DS)

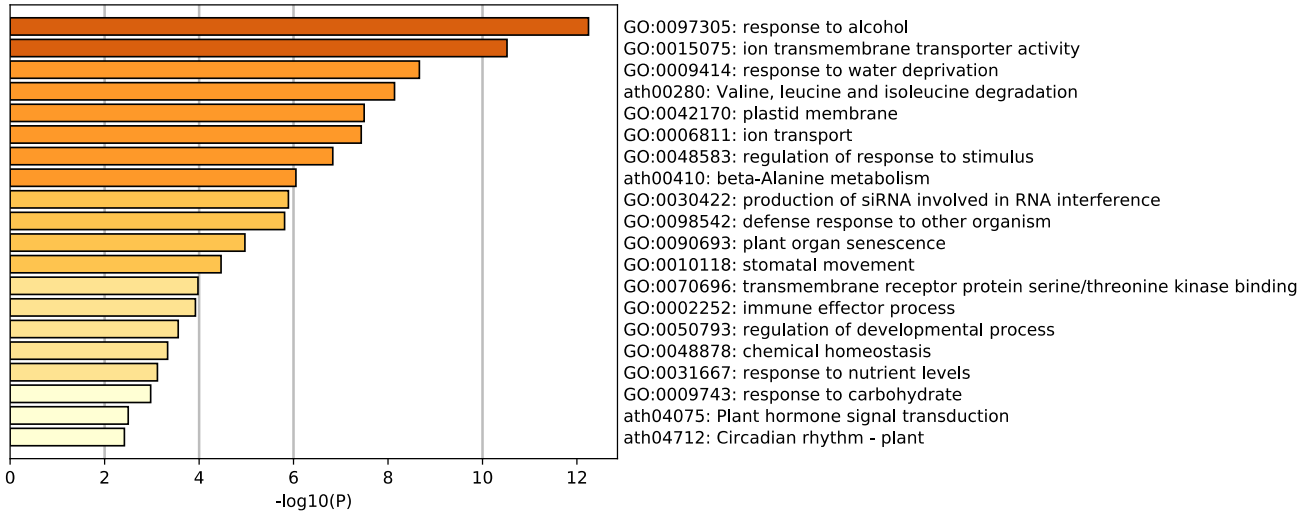


Figure 7B. (firebrick4 module in LA1589 JADS)

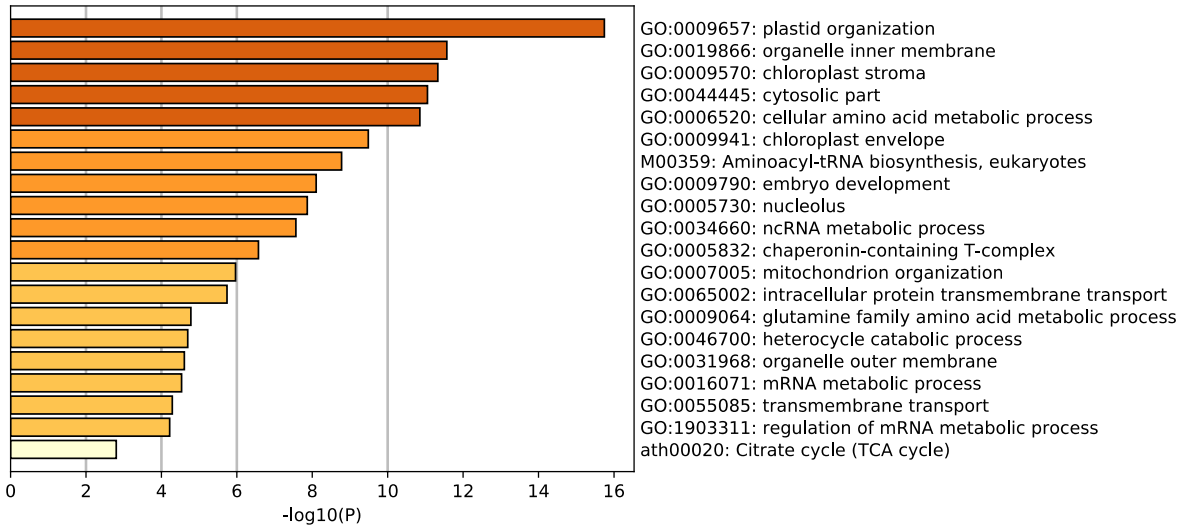


Figure 7C. (maroon module in LA1589 JADS)

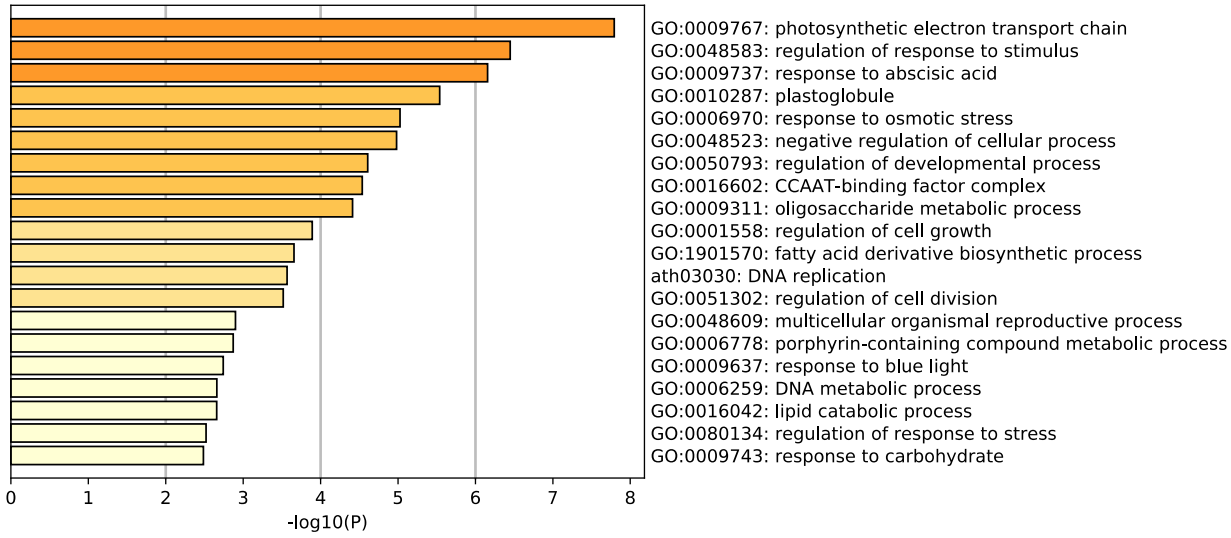


Figure 7D. (greenyellow module in LA1269 DS)

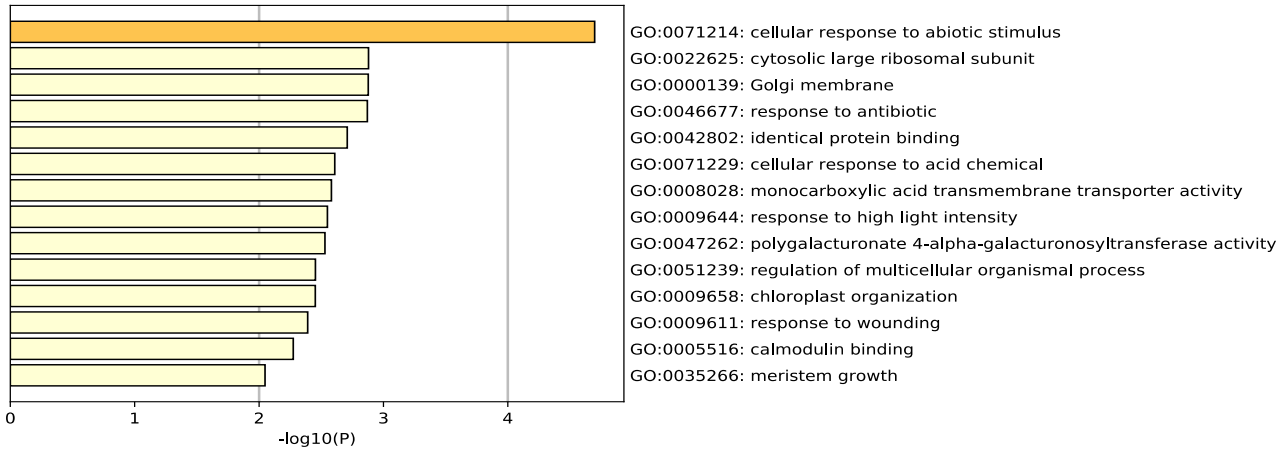


Figure 7E. (darkmagenta module in LA1269 DS)

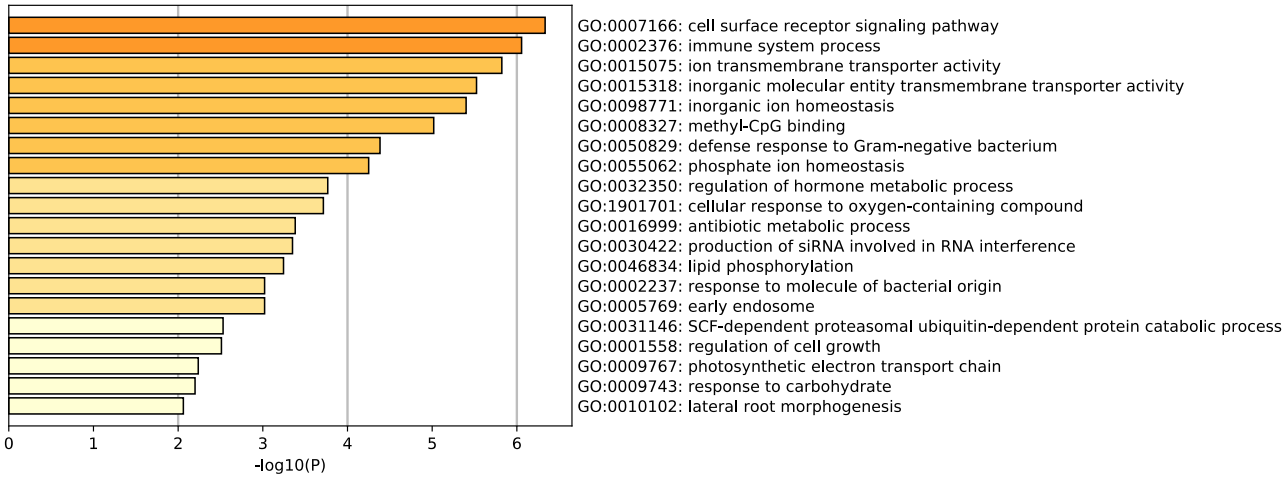


Figure 7F. (sienna3 module in LA1269 JADS)

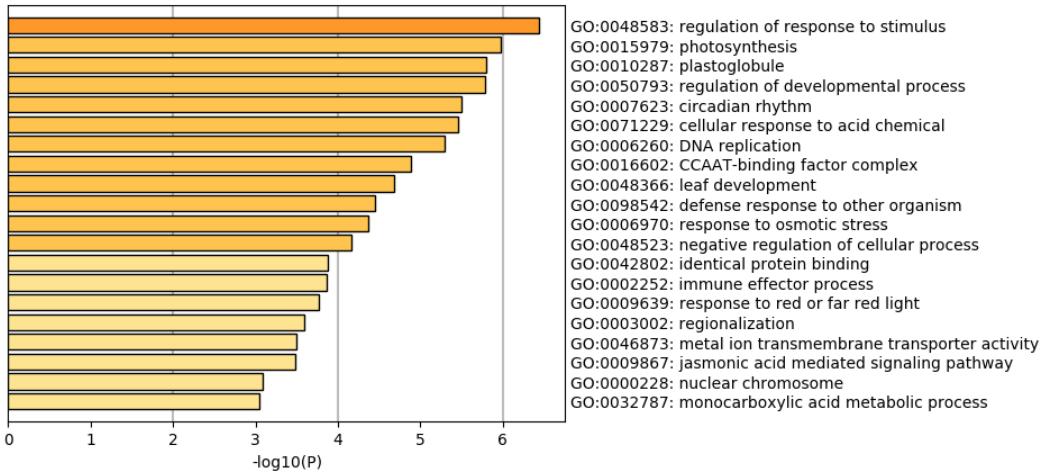


Figure 7G. (magenta module in LA1269 JADS)

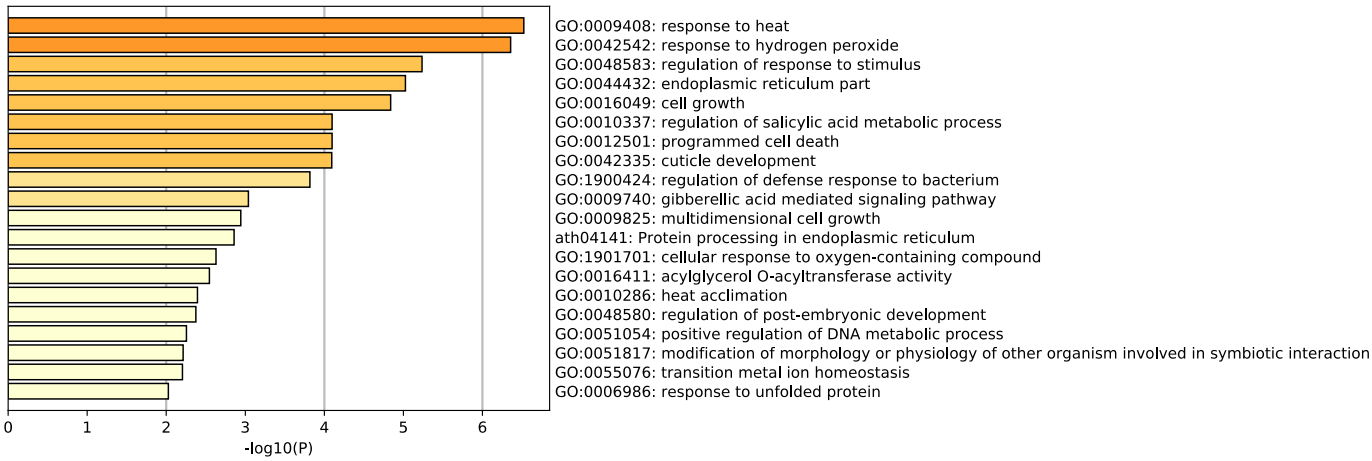


Figure 7H. (red module in LA1269 JA)

Figure 7. Heatmap of enriched terms across LA1589 modules (A, B, C) and LA1269 modules (D, E, F, G, H), colored by p-values.

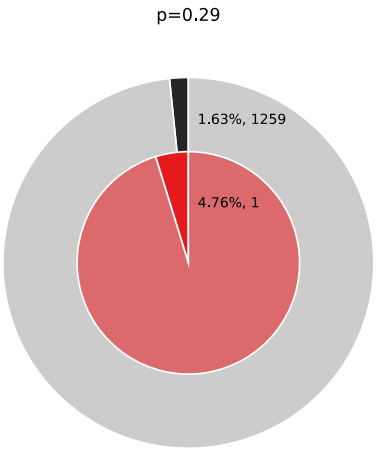


Figure 8A.

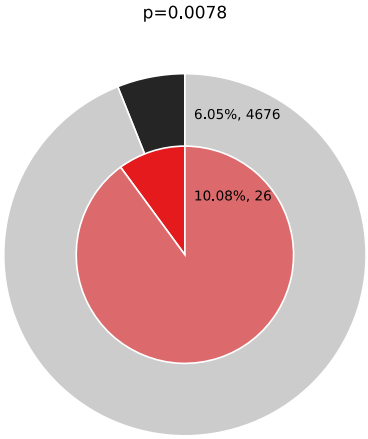


Figure 8B.

Figure 8. Enrichment of genes matching membership term: defense. (A: turquoise module of LA1589 JA, B: red module of LA1269 JA). The outer pie shows the number and the percentage of genes in the background that are associated with the membership (in black); the inner pie shows the number and the percentage of genes in the individual input gene list that are associated with the membership. The p-value indicates whether the membership is statistically significantly enriched in the list.

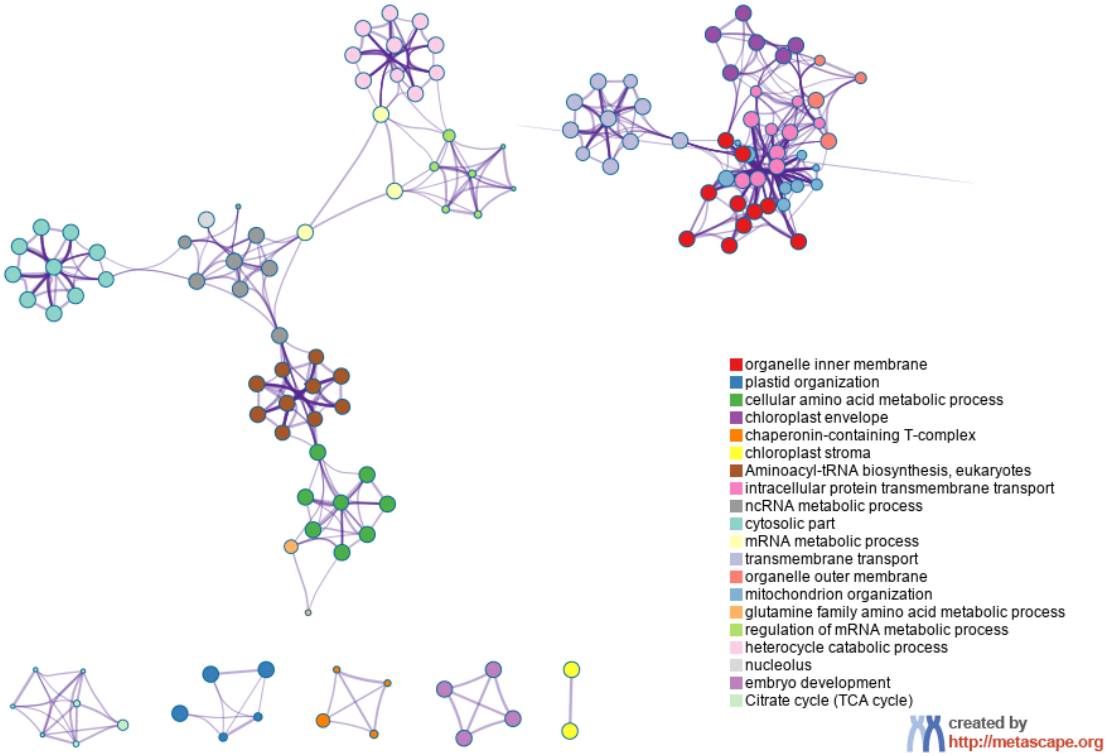


Figure 9A. (maroon module)

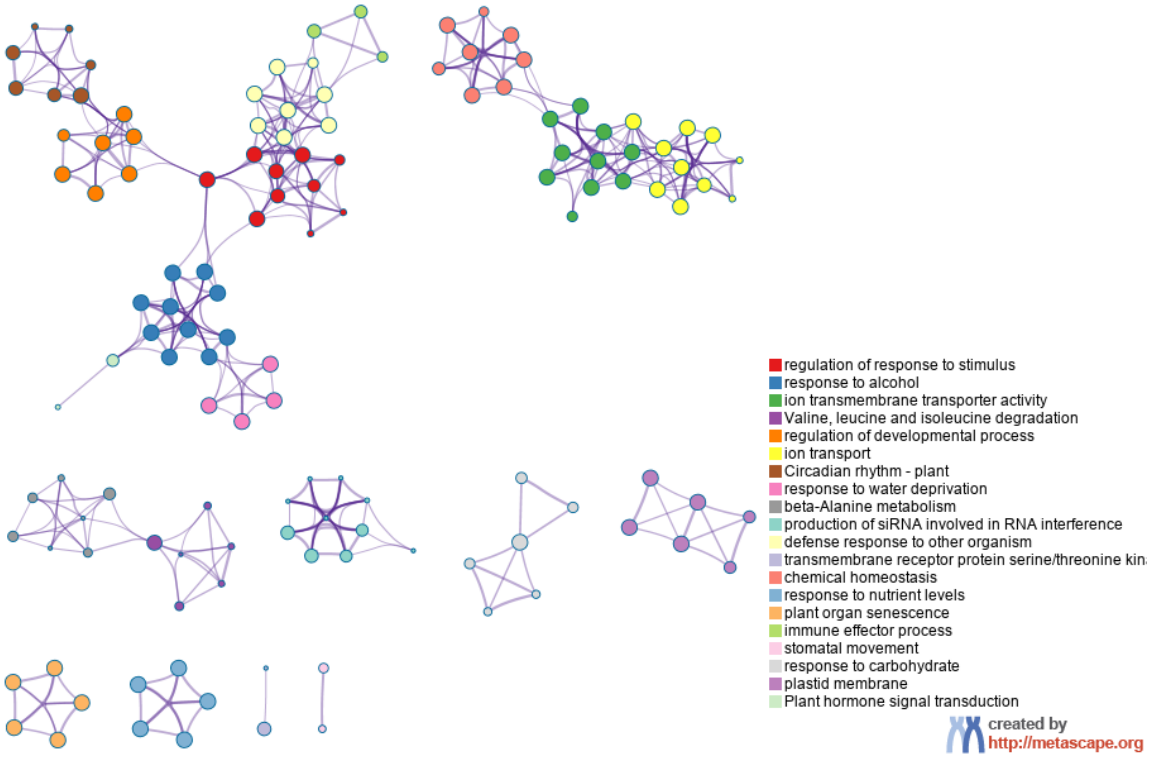


Figure 9B. (firebrick4 module)

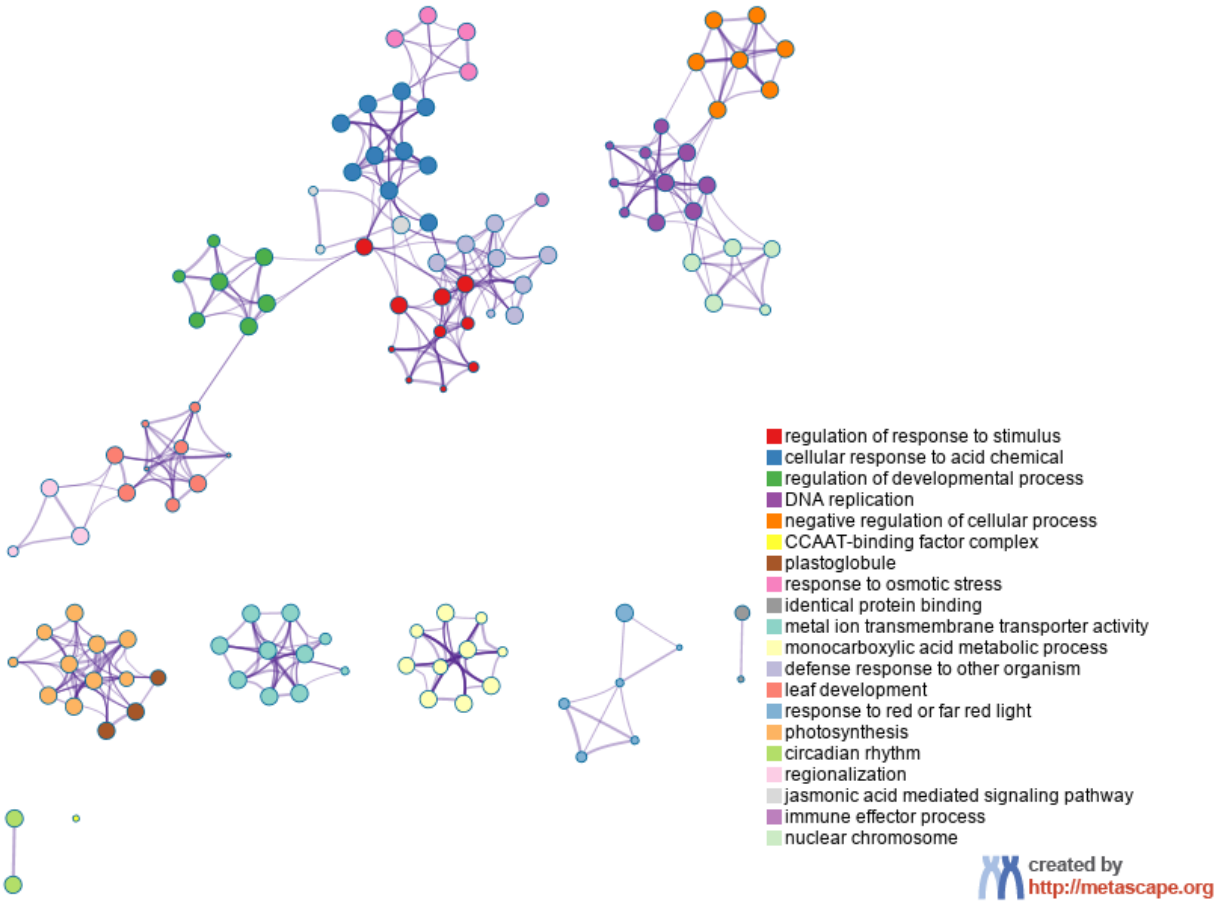


Figure 9C. (magenta module)

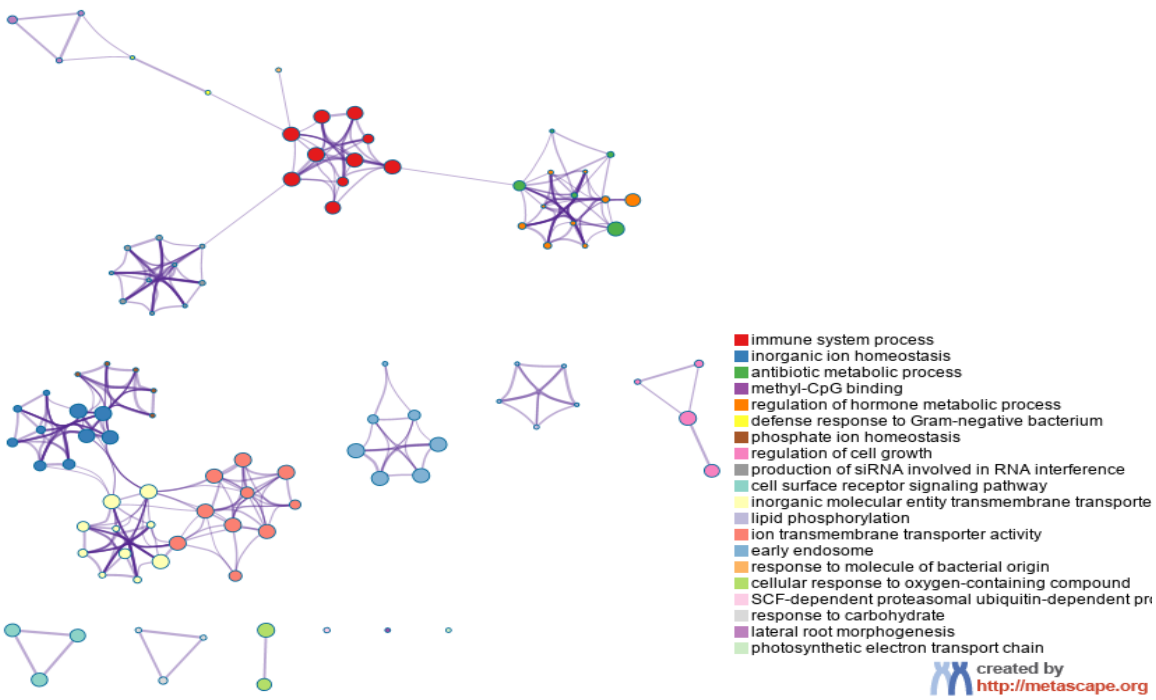
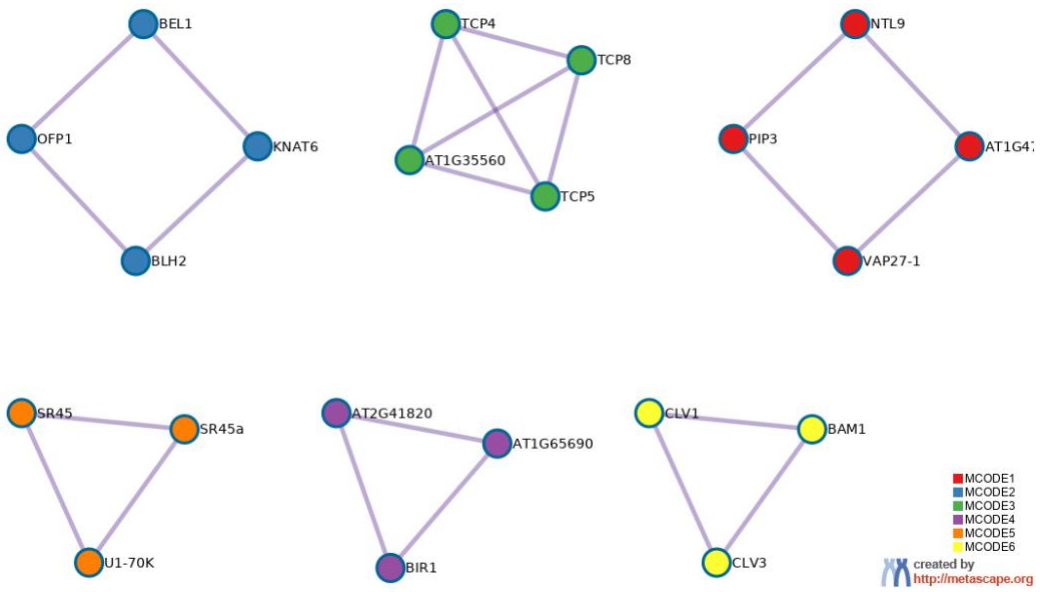


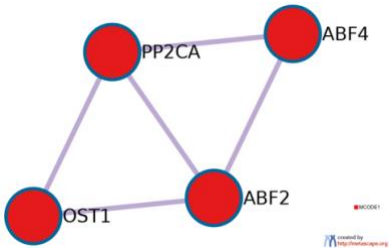
Figure 9D. (sienna3 module)

Figure 9. Sub-networks of enriched terms in maroon module (A: LA1589 JADS), firebrick4 module (B: LA1589 JADS), magenta module (C:LA1269 JADS) and sienna3 module (D: LA1269 JADS). Nodes connected closely were assigned same cluster ID and colored by cluster ID.



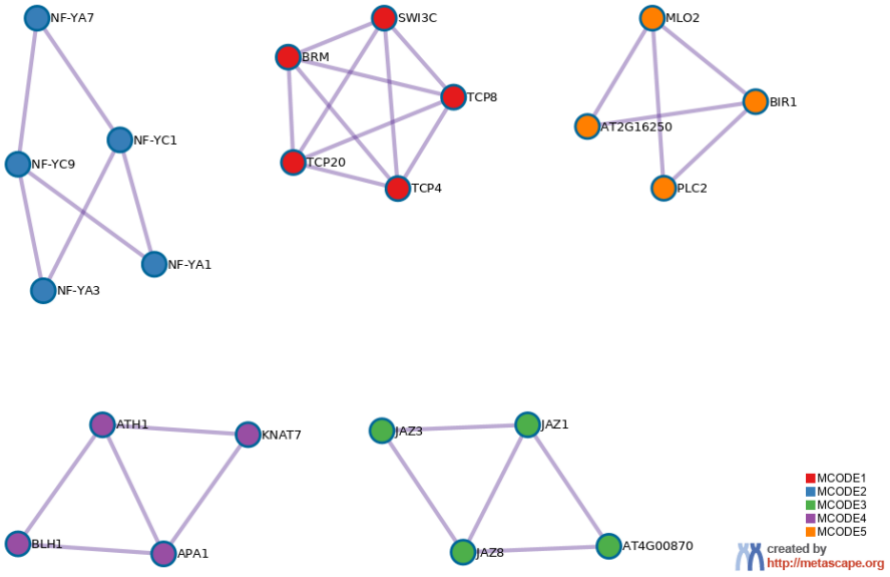
Network	Annotation
Input ID	GO:0044445|cytosolic part|-13.7;GO:0005832|chaperonin-containing T-complex|-11.8;GO:0101031|chaperone complex|-10.5
Input ID_MCODE_ALL	GO:0033612|receptor serine/threonine kinase binding|-6.8;GO:0005102|signaling receptor binding|-5.8;GO:0009934|regulation of meristem structural organization|-5.8
Input ID_SUB4_MCODE_3	GO:0040034|regulation of development, heterochronic|-7.1;GO:0050793|regulation of developmental process|-3.6
Input ID_SUB1_MCODE_5	GO:0016607|nuclear speck|-7.0;GO:0016604|nuclear body|-6.6;GO:0008380|RNA splicing|-5.6
Input ID_SUB12_MCODE_6	GO:0009934|regulation of meristem structural organization|-8.9;GO:0033612|receptor serine/threonine kinase binding|-7.9;GO:0009933|meristem structural organization|-7.2

Figure 10A. (maroon module)



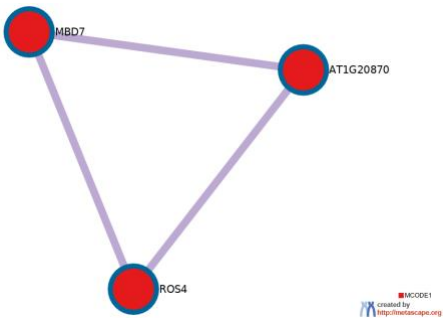
Network	Annotation
Input ID	GO:0097305|response to alcohol|-13.5;GO:0009737|response to abscisic acid|-13.0;GO:0071229|cellular response to acid chemical|-10.4
Input ID_MCODE_ALL	ath04075|Plant hormone signal transduction|-5.8;GO:0009733|response to auxin|-4.7;GO:0015293|symporter activity|-3.7
Input ID_SUB1_MCODE_1	GO:0015318|inorganic molecular entity transmembrane transporter activity|-4.4;GO:0015075|ion transmembrane transporter activity|-4.3;GO:0015291|secondary active transmembrane transporter activity|-4.1
Input ID_SUB1_MCODE_3	ath04075|Plant hormone signal transduction|-7.3;GO:0009737|response to abscisic acid|-5.9;GO:0097305|response to alcohol|-5.9
Input ID_SUB1_MCODE_5	GO:0009733|response to auxin|-6.6
_FINAL	GO:1901701|cellular response to oxygen-containing compound|-15.5;GO:0071229|cellular response to acid chemical|-14.7;GO:0009737|response to abscisic acid|-14.6
_FINAL_MCODE_ALL	ath04075|Plant hormone signal transduction|-7.3;GO:0009414|response to water deprivation|-6.7;GO:0009415|response to water|-6.7
_FINAL_SUB1_MCODE_1	ath04075|Plant hormone signal transduction|-7.3;GO:0009414|response to water deprivation|-6.7;GO:0009415|response to water|-6.7

Figure 10B. (Firebrick4 module)



Network	Annotation
Input ID	GO:0048583|regulation of response to stimulus|-11.7;ath04075|Plant hormone signal transduction|-11.0;GO:0042802|identical protein binding|-10.3
Input ID_MCODE_ALL	GO:0016602|CCAAT-binding factor complex|-8.9;GO:0048583|regulation of response to stimulus|-6.1;GO:0090575|RNA polymerase II transcription factor complex|-5.7
Input ID_SUB2_MCODE_2	GO:0016602|CCAAT-binding factor complex|-12.1;GO:0090575|RNA polymerase II transcription factor complex|-8.8;GO:0044798|nuclear transcription factor complex|-8.5
Input ID_SUB1_MCODE_3	GO:0009867|jasmonic acid mediated signaling pathway|-6.2;GO:0071395|cellular response to jasmonic acid stimulus|-6.1;GO:0009753|response to jasmonic acid|-5.0
Input ID_SUB1_MCODE_5	GO:0098542|defense response to other organism|-3.6

Figure 10C. (magenta module)



Network	Annotation
Input ID	GO:0071229|cellular response to acid chemical|-5.9;GO:0009867|jasmonic acid mediated signaling pathway|-5.9;GO:0071395|cellular response to jasmonic acid stimulus|-5.8
Input ID_MCODE_ALL	GO:0044728|DNA methylation or demethylation|-7.2;GO:0006304|DNA modification|-7.1;GO:0051052|regulation of DNA metabolic process|-6.5
Input ID_SUB5_MCODE_1	GO:0044728|DNA methylation or demethylation|-7.2;GO:0006304|DNA modification|-7.1;GO:0051052|regulation of DNA metabolic process|-6.5

Figure 10D. (sienna3 module)

Figure 10. Protein-protein interaction networks and MCODE components in maroon module (A: LA1589 JADS) firebrick4 module (B. LA1589 JADS), magenta module (C: LA1269 JADS) and sienna3 module (D: LA1269 JADS)

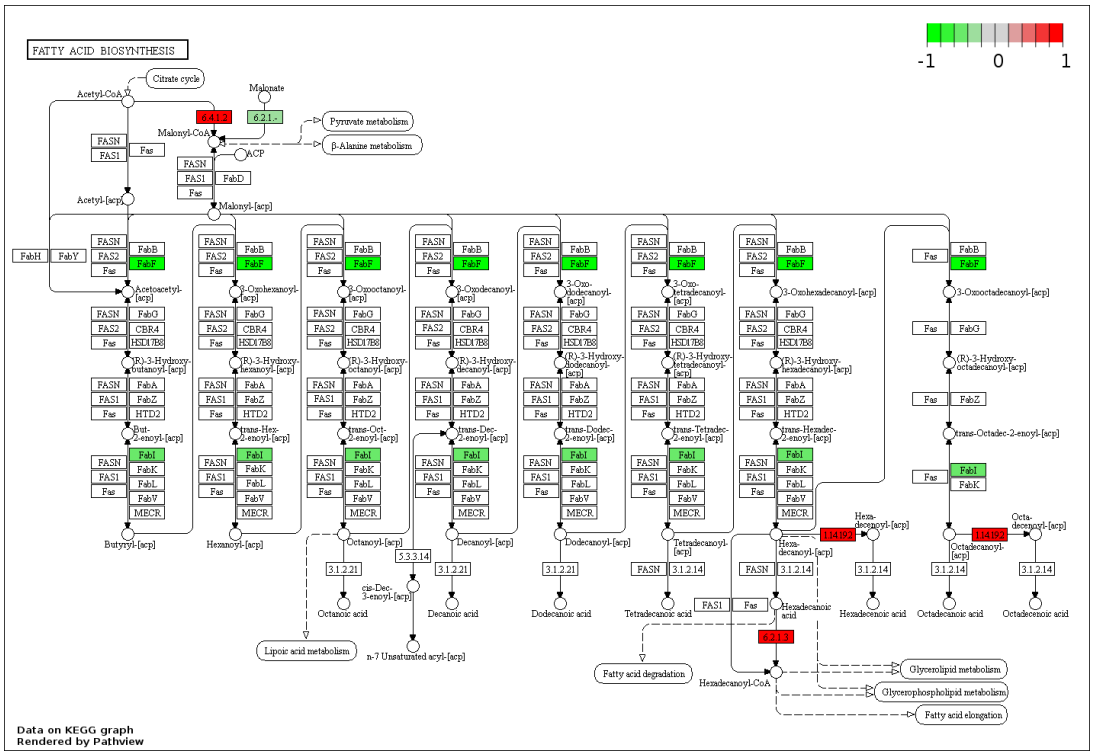

FATTY ACID BIOSYNTHESIS
-1
0
1
Citrate cycle
Acetyl-CoA
Malonate
6.4.1.2
6.2.1.-
Malonyl-CoA
Pyruvate metabolism
β-Alanine metabolism
ACP
FASN
FAS1
Fas
FabD
Acetyl-[acp]
Malonyl-[acp]
FabH
FabY
FabB
FabF
FAS2
FabG
CBR4
HSD17B8
FabA
FabZ
HTD2
FabI
FabK
FabL
FabV
MECR
Acetoacetyl-[acp]
(R)-3-Hydroxy-butanoyl-[acp]
But-2-enoyl-[acp]
Butyryl-[acp]
3-Oxohexanoyl-[acp]
(R)-3-Hydroxy-hexanoyl-[acp]
trans-Hex-2-enoyl-[acp]
Hexanoyl-[acp]
3-Oxooctanoyl-[acp]
(R)-3-Hydroxy-octanoyl-[acp]
trans-Oct-2-enoyl-[acp]
Octanoyl-[acp]
3-Oxodecanoyl-[acp]
(R)-3-Hydroxy-decanoyl-[acp]
trans-Dec-2-enoyl-[acp]
Decanoyl-[acp]
3-Oxo-dodecanoyl-[acp]
(R)-3-Hydroxy-dodecanoyl-[acp]
trans-Dodec-2-enoyl-[acp]
Dodecanoyl-[acp]
3-Oxo-tetradecanoyl-[acp]
(R)-3-Hydroxy-tetradecanoyl-[acp]
trans-Tetradec-2-enoyl-[acp]
Tetradecanoyl-[acp]
3-Oxohexadecanoyl-[acp]
(R)-3-Hydroxy-hexadecanoyl-[acp]
trans-Hexadec-2-enoyl-[acp]
Hexadecanoyl-[acp]
3-Oxooctadecanoyl-[acp]
(R)-3-Hydroxy-octadecanoyl-[acp]
trans-Octadec-2-enoyl-[acp]
Octadecanoyl-[acp]
1.14.19.2
Hexa-decenoyl-[acp]
Octa-decenoyl-[acp]
3.3.3.14
cis-Dec-3-enoyl-[acp]
3.1.2.21
3.1.2.14
Octanoic acid
Decanoic acid
Dodecanoic acid
Tetradecanoic acid
Hexadecanoic acid
Hexadecenoic acid
Octadecanoic acid
Octadecenoic acid
6.2.1.3
Hexadecanoyl-CoA
Lipoic acid metabolism
n-7 Unsaturated acyl-[acp]
Fatty acid degradation
Glycerolipid metabolism
Glycerophospholipid metabolism
Fatty acid elongation
Data on KEGG graph
Rendered by Pathview

Figure 11A.

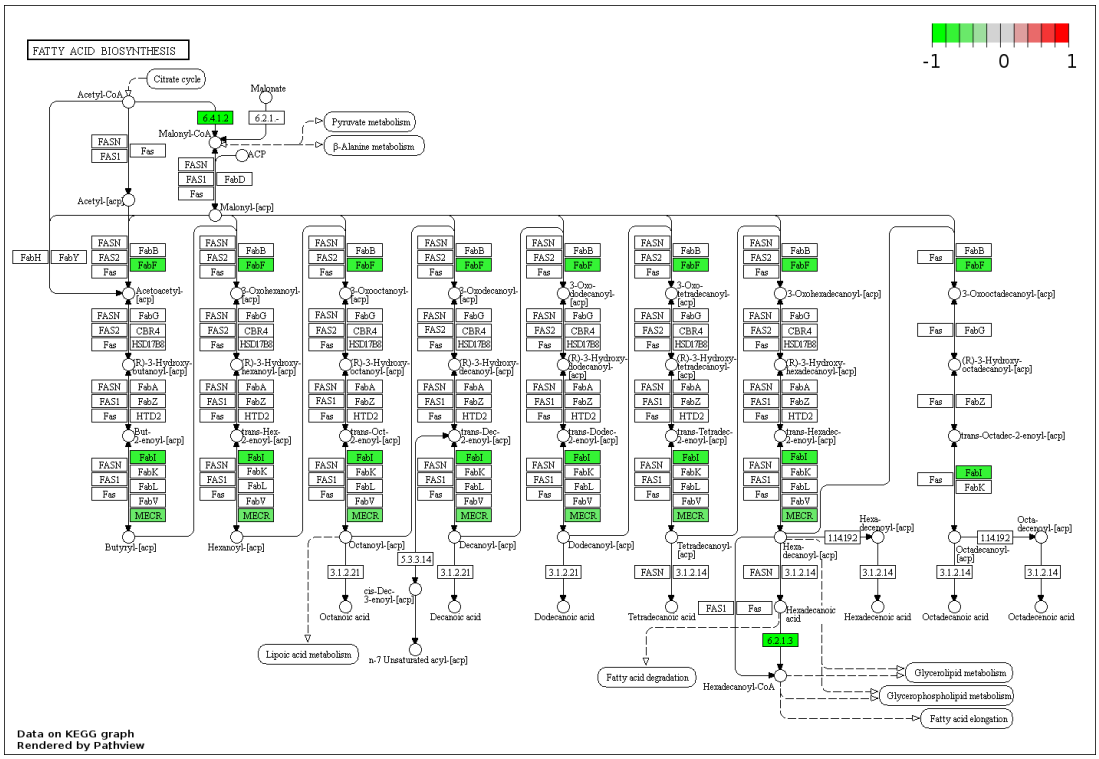

FATTY ACID BIOSYNTHESIS
-1
0
1
Citrate cycle
Acetyl-CoA
Malonate
6.4.1.2
6.2.1.-
Malonyl-CoA
Pyruvate metabolism
β-Alanine metabolism
ACP
FASN
FAS1
Fas
FabD
Acetyl-[acp]
Malonyl-[acp]
FabH
FabY
FabB
FabF
FabG
CBR4
HSD17B8
FabA
FabZ
HTD2
FabI
FabK
FabL
FabV
MECR
Acetoacetyl-[acp]
3-Oxohexanoyl-[acp]
3-Oxooctanoyl-[acp]
3-Oxodecanoyl-[acp]
3-Oxo-dodecanoyl-[acp]
3-Oxo-tetradecanoyl-[acp]
3-Oxohexadecanoyl-[acp]
3-Oxooctadecanoyl-[acp]
(R)-3-Hydroxy-butanoyl-[acp]
(R)-3-Hydroxy-hexanoyl-[acp]
(R)-3-Hydroxy-octanoyl-[acp]
(R)-3-Hydroxy-decanoyl-[acp]
(R)-3-Hydroxy-dodecanoyl-[acp]
(R)-3-Hydroxy-tetradecanoyl-[acp]
(R)-3-Hydroxy-hexadecanoyl-[acp]
(R)-3-Hydroxy-octadecanoyl-[acp]
But-2-enoyl-[acp]
trans-Hex-2-enoyl-[acp]
trans-Oct-2-enoyl-[acp]
trans-Dec-2-enoyl-[acp]
trans-Dodec-2-enoyl-[acp]
trans-Tetradec-2-enoyl-[acp]
trans-Hexadec-2-enoyl-[acp]
trans-Octadec-2-enoyl-[acp]
Butyryl-[acp]
Hexanoyl-[acp]
Octanoyl-[acp]
Decanoyl-[acp]
Dodecanoyl-[acp]
Tetradecanoyl-[acp]
Hexadecanoyl-[acp]
Octadecanoyl-[acp]
5.3.3.14
3.1.2.21
3.1.2.14
cis-Dec-2-enoyl-[acp]
Octanoic acid
Decanoic acid
Dodecanoic acid
Tetradecanoic acid
Hexadecanoic acid
Hexadecenoic acid
Octadecanoic acid
Octadecenoic acid
1.14.19.2
Hexa-decenoyl-[acp]
Octa-decenoyl-[acp]
Lipoic acid metabolism
n-7 Unsaturated acyl-[acp]
Fatty acid degradation
6.2.1.3
Hexadecanoyl-CoA
Glycerolipid metabolism
Glycerophospholipid metabolism
Fatty acid elongation
Data on KEGG graph
Rendered by Pathview

Figure 11B.

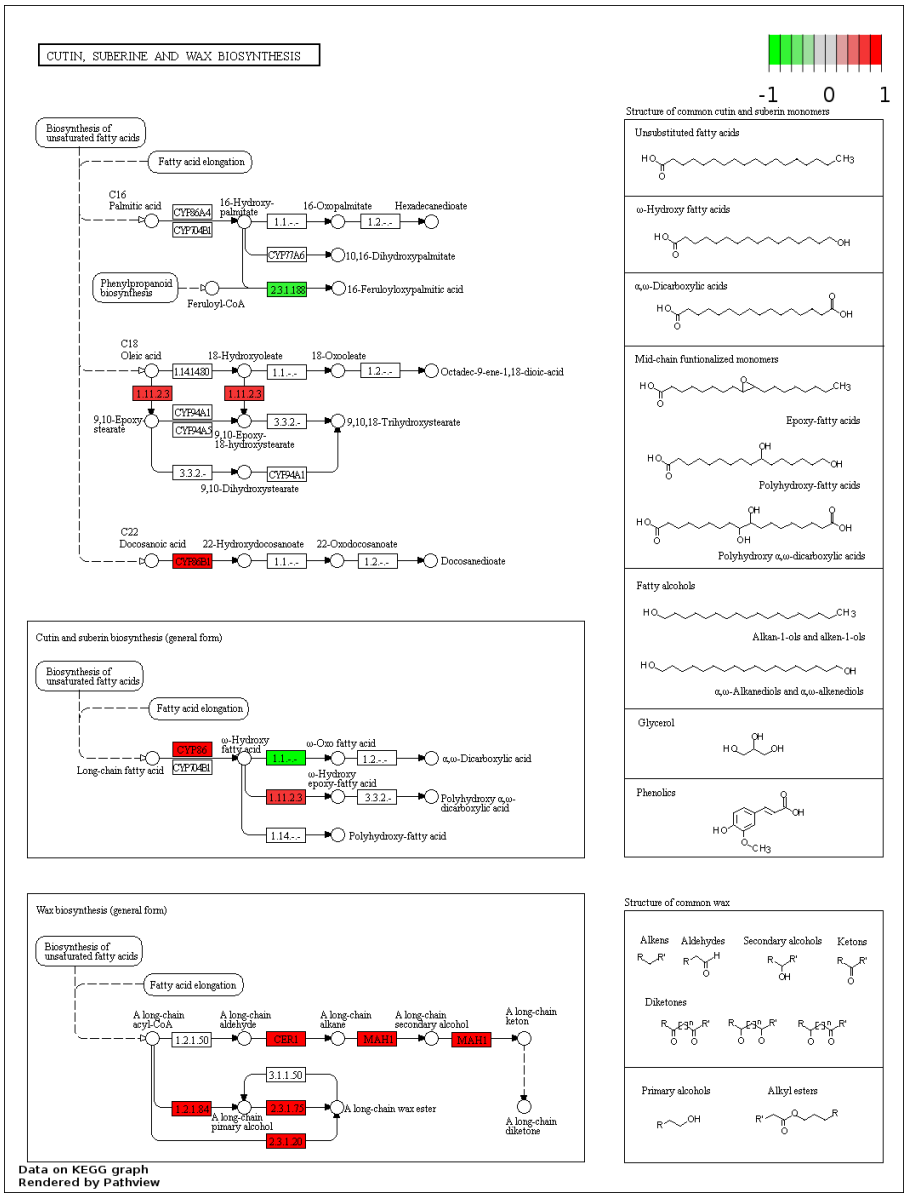

CUTIN, SUBERINE AND WAX BIOSYNTHESIS
-1
0
1
Structure of common cutin and suberin monomers
Unsubstituted fatty acids
ω-Hydroxy fatty acids
α,ω-Dicarboxylic acids
Mid-chain funtionalized monomers
Epoxy-fatty acids
Polyhydroxy-fatty acids
Polyhydroxy α,ω-dicarboxylic acids
Fatty alcohols
Alkan-1-ols and alken-1-ols
α,ω-Alkanediols and α,ω-alkenediols
Glycerol
Phenolics
Biosynthesis of unsaturated fatty acids
Fatty acid elongation
C16
Palmitic acid
CYP86A4
CYP704B1
16-Hydroxy-palmitate
16-Oxopalmitate
Hexadecanedioate
1.1.-.-
1.2.-.-
CYP77A6
10,16-Dihydroxypalmitate
Phenylpropanoid biosynthesis
Feruloyl-CoA
2.3.1.188
16-Feruloyloxypalmitic acid
C18
Oleic acid
18-Hydroxyoleate
18-Oxooleate
1.14.14.30
Octadec-9-ene-1,18-dioic-acid
1.11.2.3
CYP94A1
CYP94A5
9,10-Epoxy stearate
3.3.2.-
9,10,18-Trihydroxystearate
9,10-Epoxy-18-hydroxystearate
9,10-Dihydroxystearate
C22
Docosanoic acid
22-Hydroxydocosanoate
22-Oxodocosanoate
CYP86B1
Docosanedioate
Cutin and suberin biosynthesis (general form)
CYP86
ω-Hydroxy fatty acid
ω-Oxo fatty acid
α,ω-Dicarboxylic acid
Long-chain fatty acid
ω-Hydroxy epoxy-fatty acid
Polyhydroxy α,ω-dicarboxylic acid
1.14.-.-
Polyhydroxy-fatty acid
Wax biosynthesis (general form)
A long-chain acyl-CoA
A long-chain aldehyde
A long-chain alkane
A long-chain secondary alcohol
A long-chain keton
1.2.1.50
CER1
MAH1
3.1.1.50
1.2.1.84
2.3.1.75
A long-chain wax ester
A long-chain primary alcohol
2.3.1.20
A long-chain diketone
Structure of common wax
Alkens
Aldehydes
Secondary alcohols
Ketons
Diketones
Primary alcohols
Alkyl esters
Data on KEGG graph
Rendered by Pathview

Figure 11C.

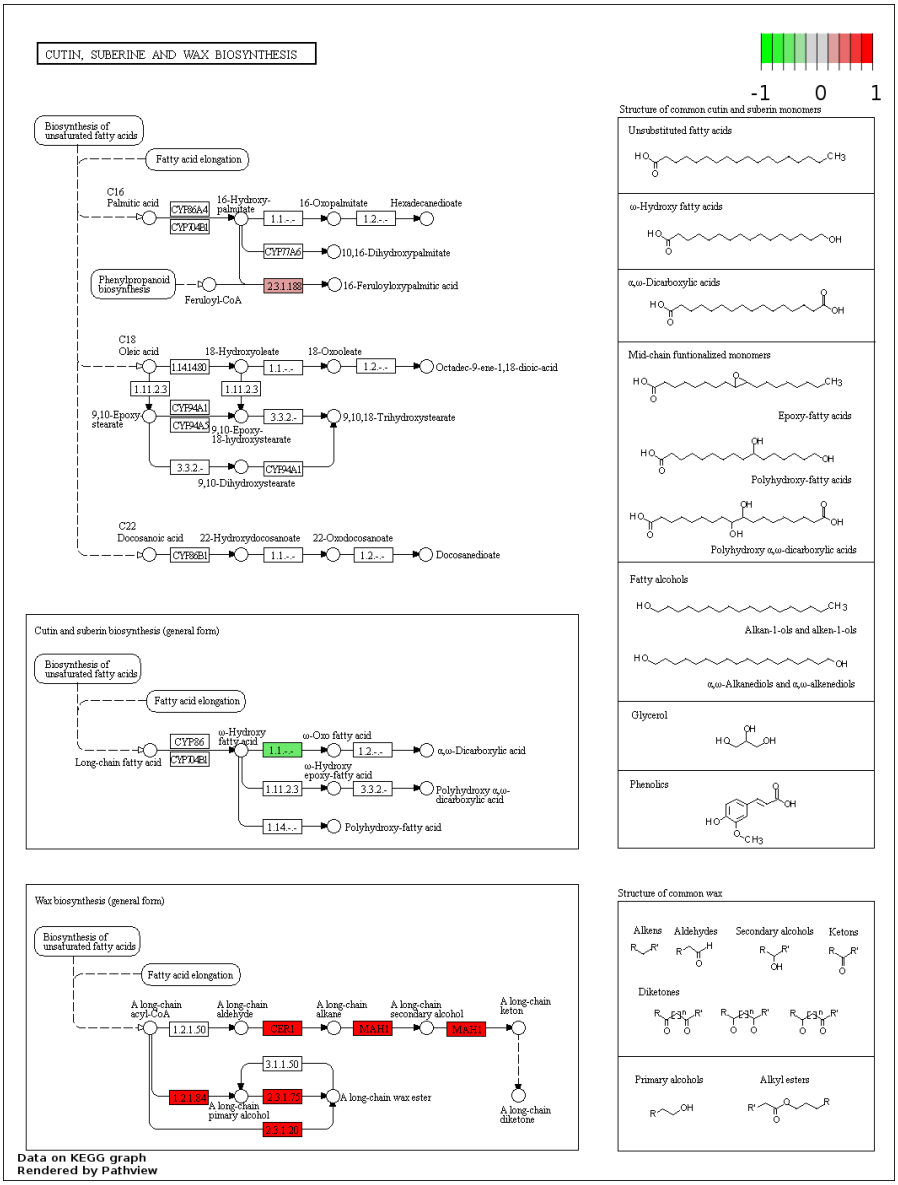


Figure 11D.

Figure 11. Differential expression genes from JADS treatment mapped to KEGG “fatty acid biosynthesis” pathway (A. LA1589; B. LA1269) and “Cutin, suberine and wax biosynthesis” pathway (C. LA1589; D. LA1269). Colored key denoted for $\log_2$ FoldChange of gene expression.

Carbohydrate Metabolism	Energy metabolism	Lipid metabolism	Metabolism of cofactors and vitamins	Amino acid metabolism	Nucleotide metabolism	Biosynthesis of other secondary metabolites
Glycolysis / Gluconeogenesis	Oxidative phosphorylation	Fatty acid biosynthesis	Ubiquinone and other terpenoid-quinone biosynthesis	Arginine biosynthesis	Purine metabolism	Caffeine metabolism
Citrate cycle (TCA cycle)	Photosynthesis	Fatty acid elongation				
Pentose phosphate pathway	Photosynthesis - antenna proteins	Fatty acid degradation				
Pentose and glucuronate interconversions		Synthesis and degradation of ketone bodies				
Fructose and mannose metabolism		Cutin, suberine and wax biosynthesis				
Galactose metabolism		Steroid biosynthesis				
Ascorbate and aldarate metabolism						

Table 2: KEGG metabolism pathways identified from differential expressed genes under JADS treatment of LA1589 and LA1269. The highlight cell indicated pathways that only presented in LA1589 otherwise presented on both accessions.

Pathway	**LA1589**	**LA1269**
Fatty acid biosynthesis	ACC1 AAE13 FAB1 MOD1 FTM1 LACS8	BCCP2 KASI MOD1 AT3G45770 LACS2
Cutin, suberine and wax biosynthesis	RWP1 AT1G70680 CYP86B1 HTH MS2 CER1 WSD1 CYP96A15	RWP1 HTH MS2 CER1 WSD1 CYP96A15

Table 3: Differential expressed genes under JADS treatment involved in pathways.

SUPPLEMENTAL MATERIALS

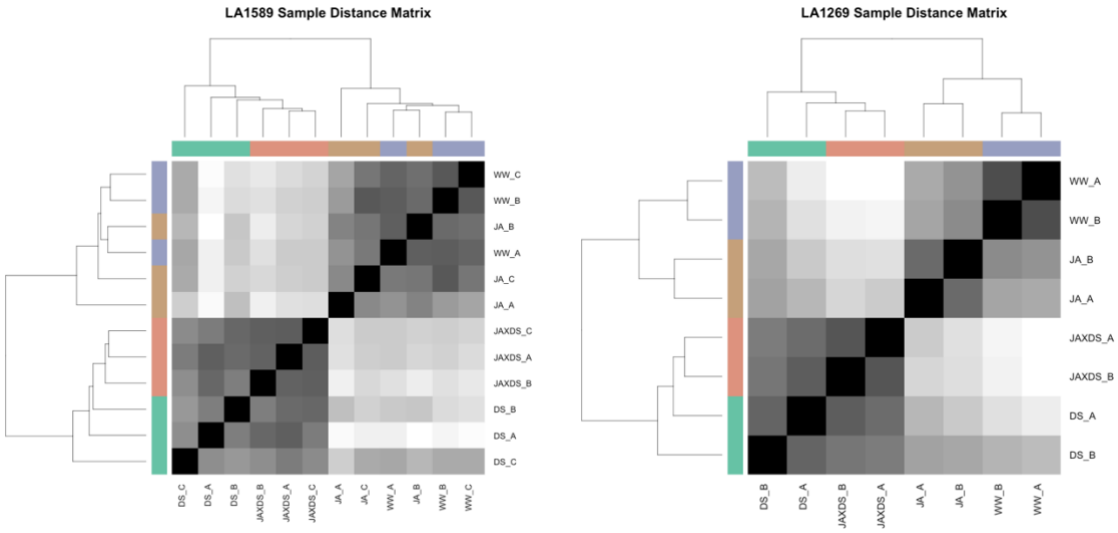


Figure S1**.** Sample to sample distance. Euclidean distance calculated from regularized log transformation of read count across replicates and treatments.

REFERENCES

Anderson, J. P., et al. (2004). "Antagonistic interaction between abscisic acid and jasmonate-ethylene signaling pathways modulates defense gene expression and disease resistance in Arabidopsis." The plant cell **16**(12): 3460-3479.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Andrews, S. (2016). FastQC: a quality control tool for high throughput sequence data. 2010.

Atkinson, N. J. and P. E. Urwin (2012). "The interaction of plant biotic and abiotic stresses: from genes to the field." Journal of experimental botany **63**(10): 3523-3543.

Bader, G. D. and C. W. Hogue (2003). "An automated method for finding molecular complexes in large protein interaction networks." BMC bioinformatics **4**(1): 2.

Bai, Y. and P. Lindhout (2007). "Domestication and breeding of tomatoes: what have we gained and what can we gain in the future?" Annals of botany **100**(5): 1085-1094.

Baldwin, I. T. (2001). "An ecologically motivated analysis of plant-herbivore interactions in native tobacco." Plant physiology **127**(4): 1449-1458.

Bates, D., et al. (2007). "The lme4 package." R package version **2**(1): 74.

Bhardwaj, A., et al. (2018). RNA-seq based mapping strategies to uncover heterogeneity in survival among Pancreatic Ductal Adenocarcinoma (PDAC) patients. Proceedings of the 9th International Conference on Computational Systems-Biology and Bioinformatics, ACM.

Blanca, J., et al. (2012). "Variation revealed by SNP genotyping and morphology provides insight into the origin of the tomato." PLoS One **7**(10): e48198.

Bolger, A. M., et al. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120.

Bolton, M. D. (2009). "Primary metabolism and plant defense—fuel for the fire." Molecular plant-microbe Interactions **22**(5): 487-497.

Breiman, L. (2001). "Random forests." Machine learning **45**(1): 5-32.

Cahais, V., et al. (2012). "Reference-free transcriptome assembly in non-model animals from next-generation sequencing data." Molecular ecology resources **12**(5): 834-845.

Camacho, C., et al. (2009). "BLAST+: architecture and applications." BMC bioinformatics **10**(1): 421.

Castel, S. E., et al. (2015). "Tools and best practices for data processing in allelic expression analysis." Genome biology **16**(1): 195.

Chen, H. and P. C. Boutros (2011). "VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R." BMC bioinformatics **12**(1): 35.

Codina-Fauteux, V.-A., et al. (2018). "PHACTR1 splicing isoforms and eQTLs in atherosclerosis-relevant human cells." BMC medical genetics **19**(1): 97.

Collins, L. J., et al. (2008). An approach to transcriptome analysis of non-model organisms using short-read sequences. Genome Informatics 2008: Genome Informatics Series Vol. 21, World Scientific**:** 3-14.

Conesa, A., et al. (2016). "A survey of best practices for RNA-seq data analysis." Genome biology **17**(1): 13.

Consortium, T. G. S., et al. (2014). "Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing." The Plant Journal **80**(1): 136-148.

Danisman, S., et al. (2012). "Arabidopsis class I and class II TCP transcription factors regulate jasmonic acid metabolism and leaf development antagonistically." Plant physiology **159**(4): 1511-1523.

Dobin, A., et al. (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.

Eid, J., et al. (2009). "Real-time DNA sequencing from single polymerase molecules." Science **323**(5910): 133-138.

Fan, J., et al. (2018). "Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data." Genome research **28**(8): 1217-1227.

Farmer, E. E. and C. A. Ryan (1992). "Octadecanoid precursors of jasmonic acid activate the synthesis of wound-inducible proteinase inhibitors." The plant cell **4**(2): 129-134.

Gepts, P. (2002). "A comparison between crop domestication, classical plant breeding, and genetic engineering." Crop Science **42**(6): 1780-1790.

Grabherr, M. G., et al. (2011). "Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data." Nature biotechnology **29**(7): 644.

Grunewald, W., et al. (2009). "Expression of the Arabidopsis jasmonate signalling repressor JAZ1/TIFY10A is stimulated by auxin." EMBO reports **10**(8): 923-928.

Gur, A. and D. Zamir (2004). "Unused natural variation can lift yield barriers in plant breeding." PLoS biology **2**(10): e245.

Guttman, M., et al. (2010). "Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs." Nature biotechnology **28**(5): 503.

Haak, D. C., et al. (2014). "No evidence for phylogenetic constraint on natural defense evolution among wild tomatoes." Ecology **95**(6): 1633-1641.

Haak, D. C., et al. (2017). "Multilevel regulation of abiotic stress responses in plants." Frontiers in plant science **8**: 1564.

Haliński, Ł. P. and P. Stepnowski (2016). "Cuticular hydrocarbons and sucrose esters as chemotaxonomic markers of wild and cultivated tomato species (Solanum section Lycopersicon)." Phytochemistry **132**: 57-67.

Heil, M. and R. M. Bostock (2002). "Induced systemic resistance (ISR) against pathogens in the context of induced plant defences." Annals of botany **89**(5): 503-512.

Herman, J. S. and D. Grün (2018). "FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data." Nature methods **15**(5): 379.

Huang, D. W., et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature protocols **4**(1): 44.

Huang, X., et al. (2016). "Comparative performance of transcriptome assembly methods for non-model organisms." BMC genomics **17**(1): 523.

Jones, J. W., et al. (2018). "Differential Gene Expression and Pathway Analysis in Juvenile Nasopharyngeal Angiofibroma Using RNA Sequencing." Otolaryngology–Head and Neck Surgery **159**(3): 572-575.

Junttila, M. R., et al. (2008). "Phosphatase-mediated crosstalk between MAPK signaling pathways in the regulation of cell survival." The FASEB Journal **22**(4): 954-965.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids research **28**(1): 27-30.

Li, B., et al. (2014). "Evaluation of de novo transcriptome assemblies from RNA-Seq data." Genome biology **15**(12): 553.

Li, H., et al. (2009). "The sequence alignment/map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Liao, Y., et al. (2013). "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." Bioinformatics **30**(7): 923-930.

Lin, K.-H., et al. (2010). "Quantitative trait loci influencing fruit-related characteristics of tomato grown in high-temperature conditions." Euphytica **174**(1): 119-135.

Love, M. I., et al. (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome biology **15**(12): 550.

Lu, X., et al. (2017). "RNA-seq Analysis of Cold and Drought Responsive Transcriptomes of Zea mays ssp. mexicana L." Frontiers in plant science **8**: 136.

Luo, W. and C. Brouwer (2013). "Pathview: an R/Bioconductor package for pathway-based data integration and visualization." Bioinformatics **29**(14): 1830-1831.

Massad, T. J., et al. (2012). "Costs of defense and a test of the carbon-nutrient balance and growth-differentiation balance hypotheses for two co-occurring classes of plant defense." PLoS One **7**(10): e47554.

Mauck, K. E., et al. (2015). "Virus infection influences host plant interactions with non-vector herbivores and predators." Functional Ecology **29**(5): 662-673.

Mazerolle, M. J. and M. M. J. Mazerolle (2019). "Package 'AICcmodavg'."

Pan, X., et al. (2018). "WebCircRNA: Classifying the circular RNA potential of coding and noncoding RNA." Genes **9**(11): 536.

Pedley, K. F. and G. B. Martin (2003). "Molecular basis of Pto-mediated resistance to bacterial speck disease in tomato." Annual review of phytopathology **41**(1): 215-243.

Pedregosa, F., et al. (2011). "Scikit-learn: Machine learning in Python." Journal of machine learning research **12**(Oct): 2825-2830.

Peralta, I. E., et al. (2008). "Taxonomy of wild tomatoes and their relatives (Solanum sect. Lycopersicoides, sect. Juglandifolia, sect. Lycopersicon; Solanaceae)." Systematic botany monographs **84**.

Qin, F., et al. (2011). "Achievements and challenges in understanding plant abiotic stress responses and tolerance." Plant and Cell Physiology **52**(9): 1569-1582.

Rätsch, G., et al. (2007). "Improving the Caenorhabditis elegans genome annotation using machine learning." PLoS Computational Biology **3**(2): e20.

Razali, R., et al. (2018). "The genome sequence of the wild tomato Solanum pimpinellifolium provides insights into salinity tolerance." Frontiers in plant science **9**.

Reinbolt, R. E., et al. (2018). "Genomic risk prediction of aromatase inhibitor‐related arthralgia in patients with breast cancer using a novel machine‐learning algorithm." Cancer medicine **7**(1): 240-253.

Rejeb, I., et al. (2014). "Plant responses to simultaneous biotic and abiotic stress: molecular mechanisms." Plants **3**(4): 458-475.

Salinas, M., et al. (2013). "Genetic mapping of two QTL from the wild tomato Solanum pimpinellifolium L. controlling resistance against two-spotted spider mite (Tetranychus urticae Koch)." Theoretical and applied genetics **126**(1): 83-92.

Seo, E., et al. (2016). "Genome-wide comparative analyses reveal the dynamic evolution of nucleotide-binding leucine-rich repeat gene family among Solanaceae plants." Frontiers in plant science **7**: 1205.

Shaik, R. and W. Ramakrishna (2013). "Genes and co-expression modules common to drought and bacterial stress responses in Arabidopsis and rice." PLoS One **8**(10): e77261.

Shao, H.-B., et al. (2008). "Water-deficit stress-induced anatomical changes in higher plants." Comptes rendus biologies **331**(3): 215-225.

Signal, B., et al. (2017). "Machine learning annotation of human branchpoints." Bioinformatics **34**(6): 920-927.

Simão, F. A., et al. (2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." Bioinformatics **31**(19): 3210-3212.

Smith-Unna, R., et al. (2016). "TransRate: reference-free quality assessment of de novo transcriptome assemblies." Genome research **26**(8): 1134-1144.

Song, Q. A., et al. (2018). "Computational analysis of alternative splicing in plant genomes." Gene.

Song, S., et al. (2013). "The bHLH subgroup IIId factors negatively regulate jasmonate-mediated plant defense and development." PLoS genetics **9**(7): e1003653.

Sonnenburg, S., et al. (2002). New methods for splice site recognition. International Conference on Artificial Neural Networks, Springer.

Stark, C., et al. (2006). "BioGRID: a general repository for interaction datasets." Nucleic acids research **34**(suppl_1): D535-D539.

T O'Neil, S. and S. J. Emrich (2013). "Assessing De Novo transcriptome assembly metrics for consistency and utility." BMC genomics **14**(1): 465.

Tanksley, S. D. (2004). "The genetic, developmental, and molecular bases of fruit size and shape variation in tomato." The plant cell **16**(suppl 1): S181-S189.

Thakur, M. and B. S. Sohal (2013). "Role of elicitors in inducing resistance in plants against pathogen infection: a review." ISRN biochemistry **2013**.

Thaler, J. S., et al. (2012). "Evolution of jasmonate and salicylate signal crosstalk." Trends in plant science **17**(5): 260-270.

Top, O., et al. (2014). "Exploration of three solanum species for improvement of antioxidant traits in tomato." HortScience **49**(8): 1003-1009.

Trapnell, C., et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature biotechnology **28**(5): 511.

van Galen, P., et al. (2019). "Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity." Cell **176**(6): 1265-1281. e1224.

Vijay, N., et al. (2013). "Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments." Molecular ecology **22**(3): 620-634.

Wang, W., et al. (2003). "Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance." Planta **218**(1): 1-14.

Wang, W., et al. (2000). Biotechnology of plant osmotic stress tolerance physiological and molecular considerations. IV International Symposium on In Vitro Culture and Horticultural Breeding 560.

Wang, Z., et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature reviews genetics **10**(1): 57.

Wasternack, C. (2007). "Jasmonates: an update on biosynthesis, signal transduction and action in plant stress response, growth and development." Annals of botany **100**(4): 681-697.

Weigel, D. and M. Nordborg (2005). "Natural variation in Arabidopsis. How do we find the causal genes?" Plant physiology **138**(2): 567-568.

Wu, X., et al. (2018). "Overexpression of a NF-YC Gene Results in Enhanced Drought and Salt Tolerance in Transgenic Seashore Paspalum." Frontiers in plant science **9**: 1355.

Xiao, J., et al. (2019). "Genome-wide identification and expression profiling of trihelix gene family under abiotic stresses in wheat." BMC genomics **20**(1): 287.

Zdobnov, E. M., et al. (2016). "OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs." Nucleic acids research **45**(D1): D744-D749.

Zhang, B. and S. Horvath (2005). "A general framework for weighted gene co-expression network analysis." Statistical applications in genetics and molecular biology **4**(1).

Zhao, H., et al. (2017). "The Arabidopsis thaliana nuclear factor Y transcription factors." Frontiers in plant science **7**: 2045.

Zhou, Y., et al. (2019). "Metascape provides a biologist-oriented resource for the analysis of systems-level datasets." Nature communications **10**(1): 1523.

Zuriaga, E., et al. (2009). "Genetic and bioclimatic variation in Solanum pimpinellifolium." Genetic Resources and Crop Evolution **56**(1): 39.

Chapter 4. Draft Assembly of *Phytopthora capsici* from Long-read Sequencing Uncovers Complexity.*

*This manuscript has been submitted to *Molecular Plant Microbe Interactions*.

Chenming Cui[1], John H. Herlihy[1], Aureliano Bombarely[1,2], John M. McDowell[1], David C. Haak[1**]

1. School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061 USA

2. Present address: Department of Bioscience/Dipartimento di Bioscienze University of Milan/Universita degli Studi di Milano (UNIMI), III piano / torre B Via Celoria, 26 Milano, Italy, 20133

**To whom correspondence should be addressed: dhaak@vt.edu

Keywords: Nanopore sequencing, Minion, miniasm, pathogen, oomycete

ABSTRACT

Resolving complex plant pathogen genomes is important for identifying the genomic shifts associated with rapid adaptation to selective agents such as host and fungicide, yet assembling these genomes remains challenging and expensive. *Phytophthora capsici* is an important, globally distributed plant pathogen that exhibits wide-spread fungicide resistance and a broad host range. As with other pathogenic oomycetes, *P. capsici* has a complex life history and a complex genome. Here we leverage Oxford Nanopore Technology (ONT) and existing short read resources, to rapidly generate a low-cost, improved assembly. We generated 10Gbp from a single MinIon flow cell resulting in > 1.25 million reads with an N50 of 13kb. The resulting assembly is 124Mbp in 906 scaffolds with an N50 length of 232kb. This assembly is 92% bigger than the current draft genome of 64Mbp. We confirmed this larger genome size using flow cytometry, with an estimated size of 110Mbp. BUSCO analysis identified 96.5% complete orthologs (39.7% duplicated). Evolutionary analysis supports a recent whole genome duplication in this group. Our work provides a blueprint for rapidly integrating benchtop long-read sequencing with existing short-read data, to dramatically improve assembly quality and integrity of complex genomes and offer novel insights into pathogen genome function and evolution.

INTRODUCTION

With one of the widest host ranges in the genus, *Phytophthora capsici* is an important destructive pathogen in diverse cropping systems including, pepper, tomato, and potato (Leonian 1922, Lamour et al. 2012a). To understand the mechanisms involved in host-adaptation a Sanger sequencing-based reference genome was generated in 2012, using

an inbred isolate LT1534 (Lamour et al. 2012b). Comparative genomics revealed that like other *Phytophthora* spp., *P. capsici* has a diverse array of effector proteins that are associated with pathogenicity (Lamour et al. 2012b,Stam et al. 2013). Tyler et al. (2006) and others have demonstrated that in *Phytophthora* spp. these are often associated in gene poor regions interspersed within repetitive regions (Raffaele and Kamoun 2012). Resolving genomes at this level is important for identifying key genomic regions associated with pathogenicity traits. Yet, these complex regions of repeats are computationally challenging for using high-throughput short-read (150-300bp) data as the reads often do not span the repetitive region therein prohibiting accurate assembly.

Recent advances in sequencing technologies are improving our ability to generate long-reads (3-15kbp) that allow us to overcome these complexities by spanning more of the repetitive region, allowing accurate assembly and uncovering previously hidden genomic information. For example, using the Pacific Biosciences Sequel platform, Yang et al. 2018) generated a full-length assembly for *Phytophthora cactorum* that spanned 121.5 Mb and consisted of ca. 46% repetitive sequence. Novel findings from this sequencing project include the identification of a recent whole genome duplication (WGD) and subsequent gene loss in this lineage, as well as expanded gene families associated with pathogenicity, relative to the more host-specific *P. sojae* (Yang et al. 2018). Similarly, Malar et al. (Malar et al. 2019) generated a haplotype-phased assembly of *Phytophthora ramorum* from PacBio long-reads identifying an overall repeat increase from 29% (prior assembly) to 48% and described newly identified RXLR effector genes. The obvious downside of long-read sequencing is both the high per-base cost to generate the reads

(depending on sequencing cell output) and the time and technical expertise needed for library preparation (Fletcher et al. 2019).

The Oxford Nanopore Techologies (ONT) MinIon platform is a so-called third generation long read platform that is small and aimed at lab-based users and requires minimal technical expertise. Importantly, sequencing flow cell output can be improved by the end user through adjustments to extraction and preparation protocols. Finally, the time from sample preparation to data acquisition is greatly diminished and eliminates the need for sample shipment to a sequencing facility. Here we report the use of ONT MinIon sequencing technology and a streamlined bioinformatics pipeline that includes minimap2 and miniasm (Li 2016) to develop an improved reference genome for *P. capsici*, in only nine days. This cost-effective, improved assembly revealed novel gene-space and genomic complexity and enabled a substantial revision of *P. capsici* genome size.

RESULTS

Genome sequencing and assembly

The isolate used in this study, LT263, was originally isolated from infected pumpkin in Tennessee in 2004. Sequencing was completed on the ONT MinIon platform. A single 1D R9.4 flowcell generated 1,258,480 reads (~10Gb) at 70X with an N50 read length of 11507bp, the longest read was 99,577bp, and the mean read length was 7,114bp.

Basecalled nanopore raw reads were assembled into 124.2 Mb using a custom bash pipeline that included overlapping via minimap2 (Li 2018), overlap consensus *de novo* assembly via miniasm (Li 2016), and successive sequence polishing via Racon (Vaser et al. 2017). Because miniasm concatenates overlaps produced by minimap2 (or other

overlappers), raw assemblies are contiguous but retain the error rate of the input reads. Thus, the *P. capsici* raw assembly was 123.4 Mb contained in 1132 contigs (Table 1a) with an N50 length of 158.9kb, however, it retained a ~14.4% error rate and therefore captured only 127 of 234 (53.9%, Table 1b) complete BUSCOs (alveolata_stramenophiles_ensembl database implemented in BUSCO v 3.0; Simão et al. 2015).

Polishing the assembly with the uncorrected nanopore long reads using Racon (Vaser et al. 2017) improved the BUSCO score to 156 out of 234 (66.6%, Table 1b). Completeness was improved substantially to 226 out of 234 (96.6%, Table 1b) complete BUSCOs after polishing with publicly available short read data (Lamour et al. 2012b). The increased duplication rate is in contrast with the published LT1534 assembly (Lamour et al. 2012b) were the single copy BUSCO score of 91.0% and duplicated COGs at 0.00% (Table 1b). The polished assembly was scaffolded using SSPACE-Long (Boetzer and Pirovano 2014) resulting in 906 scaffolds with an L50 of 199.6 kbp, and 99% of scaffolds > 10kb (Table 1a). Assembly, error correction, and scaffolding took just 72 hours using 20 threads and 100Gb RAM on a 48-core sever with 1 Tb of RAM available. In contrast, assembling corrected reads with Canu v1.7 (Koren et al. 2017), completed in 132 hours, and raw assembly with Canu v1.7 using internal read correction exceeded 30 days.

Confirmation of P. capsici LT263 genome size

The 124 Mb size of the assembly was surprising, because it is almost two-fold larger than the previous estimate of 64 Mb derived from *P. capsici* LT1534 (Lamour et al. 2012b). Flow cytometry was conducted to confirm the genome size of this isolate. Using *Sinningia*

speciosa (1C size ~392Mb) as an internal standard the haploid nuclear content of this isolate was estimated at 110Mb (Figure 1).

Genome sequence analysis and comparative genomics

A combination of homology and *ab initio* methods implemented in Maker (Yandell and Ence 2012), predicted 24,160 protein coding regions from 906 scaffolds (Table 2). This represents a gene content increase of 4,355 from the previous estimate of 19,805 (Lamour et al. 2012b) and results in a gene density of 195/Mb which is within the range of other members of the genus (Yang et al. 2018). Thus, genomic architecture was not substantially altered with strong synteny detected between the current assembly, reference genome, and other *Phytophthora spp.* genomes (Figure 2), similar to other comparative studies (Lamour et al. 2012b).

A P. capsici-P. infestans Ks plot (Figure 3) supports the long-standing hypothesis of an ancestral whole genome duplication leading to the clade containing *P. ramorum*, *P. infestans*, *P. sojae*, and *P. capsici* (Martens and Van de Peer 2010). The peak at Ks 0.1-0.5 is consistent with the speciation event between *P. capsici* and *P. infestans*. A second peak at Ks 1.7-2.0 is consistent with an ancestral WGD event. Though these values of Ks are close to saturation, they are consistent with other studies supporting a WGD in this clade (Martens and Van de Peer 2010; Yang et al. 2018).

DISCUSSION

A long-standing goal in genomics has been the development of sequencing and assembly approaches that allow the development of contiguous genome sequences in a reasonably short timeframe. We leveraged Oxford Nanopore MinIon generated long-read and

available short-read sequencing data to assemble the complex genome from *P. capsici*. In addition, using recently introduced algorithms, the assembly was completed in 1/6th of the time required for standard approaches. Importantly, we found that the genome was 1.9x the size of previous estimates and represented an increase in the number of genes and repeat content.

While prior estimates of the genome size were much smaller our assembly size was confirmed by a flow cytometry-estimated size of 110Mb. Previous estimates were based on assemblies that in part used short-read data where repetitive regions are often collapsed (Treangen and Salzberg 2011). Conversely, generating overlaps from high error long read data can lead to false expansion from partial overlaps at repeats (Chu et al. 2017), which may, in part, explain the discrepancy between our assembly size and the flow cytometry estimate (124 and 110 Mb, respectively). We anticipate that the repeat-rich regions in the new assembly will enable identification of additional genes with functions in plant host interactions, as shown recently for *P. ramorum* (Malar et al. 2019).

Annotation of our assembly using publicly available RNAseq data (Lamour et al. 2012b) identified an increased gene content of about 4,000 genes. In addition, we report a substantial increase in the number of duplicated conserved orthologous genes detected. These differences, however, were similar to long-read genome assemblies for other closely related members of the genus (Yang et al. 2018). The increase in repeats across the genome is most likely the result of a prior whole genome duplication (WGD) which is supported in data from other species *P. ramorum*, *P. cactorum*, and *P. sojae* (Yang et al. 2018; Malar et al. 2019; Martens and Van de Peer 2010). For our assembly this is supported by evidence from *P. sojae*-*P. capsici* Ks plots, wherein we find two Ks peaks,

one representing the split between *P. capsici* and *P. sojae* and a second peak consistent with the retention of genes from an older duplication event. Yang, et. al. (2018) found that this WGD lead to an expansion of gene families in the clade that contains *P. cactorum* and *P. sojae*. Further comparative genomic analysis within this clade will identify the timing of this WGD and its impact on gene family diversification leading to adaptation.

Resolving the complex genomes of plant pathogens is an important step toward understanding the mechanisms through which they adapt to host plants. We have coupled a lab-based sequencing approach and efficient assembly algorithms to generate a *de novo* assembly for *P. capsici* that captured previously undescribed complexity. An important part of this assembly was the availability of public data for genome polishing and annotation. We anticipate that costs associated with hybrid sequencing approaches, such as presented here, will continue to decrease as an increasing number of plant pathogen sequencing projects are completed and those data are added to repositories. In total, our sequencing effort was just nine days from DNA extraction to polished assembly pushing us closer to 'real-time' whole genome sequencing of plant pathogens in the field.

MATERIALS & METHODS

Strain selection and cultivation conditions

P. capsici isolate LT263 is used globally for its virulence on a wide range of hosts, and its sexual and asexual fecundity amendable for genetic studies. The isolate was obtained from Kurt Lamour, and maintained on 10% V8 agar plates at 26 ˚C in the dark. For DNA extraction, flasks with 10% V8 liquid media were inoculated and grown in a shaker

incubator at 26 ˚C in the dark for seven days. Hyphae were collected from the flasks in 50ml centrifuge tubes, immediately frozen in liquid nitrogen, then stored at -80 ˚C until use.

DNA extraction and nanopore sequencing

High quality *P. capsici* DNA was isolated using modified DNeasy Plant Mini Kit, according to the manufacturer's instructions (Qiagen). The genomic DNA was sequenced using the MinION platform. Sequencing was preceded by library preparation from 1.5ug gDNA using 1D Genomic DNA sequencing kit SQK-LSK108 from Oxford Nanopore Technologies. DNA fragmentation was not performed in order to retain longer fragments for sequencing. The extracted DNA was repaired using the FFPE Repair Mix (New England Biolabs), followed by end repair and dA-tailing using the NEBNext End Prep Module (New England Biolabs). Then, the adapter was ligated to the wash cleaned DNA using Blunt/TA Ligase Master Mix (New England Biolabs). All bead washing steps were completed using AMPure XP beads (Beckman Coulter). The prepared libraries were sequenced using an R9.4 Flow cell on the MinION device. Sequencing was performed using 48 run time protocol of MinKNOW2.2 software. In total, 1.258 million reads, which translated to 10.06 gigabases was generated from a single flow cell.

Genome assembly and assessment

Basecalling was performed using Albacore v2.3.1 (Oxford Nanopore Technologies) and subsequent raw FAST5 files were converted into a single combined FASTQ file. Read lengths that were below 1000 bp were filtered out prior to genome assembly. Raw reads

were assembled via a custom bash pipeline (supplemental material) that used minimap2 -x ava-ont to generate overlaps, miniasm to assemble overlaps, and two rounds of racon polishing. The first round of polishing used the raw reads and a second round used publicly available Illumina sequence data. The polished assembly was scaffolded using SSPACE-Long (Boetzer and Pirovano 2014) and the scaffolds were annotated using Maker (Yandell and Ence 2012). For comparison, raw reads were also assembled using Canu v.1.7 (Koren et al. 2017).

Genome scaffolds were assessed for integrity and length using Quast v. 4.6.3 (Gurevich et al. 2013) and the quality of gene space capture was assessed using BUSCO v.3.0 (Simão et al. 2015). Assembly, polishing, and scaffolding of version pcapsici_VT1.0 were completed on a Linux server running Ubuntu v18.04 with 96 cores and 1TB RAM available. Evolutionary analyses were completed using Mauve (Darling et al. 2004) to generate whole genome alignments and synteny plots and CoGe (https://genomevolution.org/coge/) to generate comparative Ks plots.

Genome size estimation

We modified the protocol of Galbraith et al. (1983) for use with *P. capsici* in culture (Zhang, Makris, Herlihy, and Haak, in prep). In short, P. capsici LT263 was maintained on 10% clarified V8 plates (1.5% agar) with 30mg/L beta-sitosterol and then transferred to the same liquid medium in a shaker incubator (28^{o}C) and kept in the dark for 7 days. Approximately 0.8-1.2g (wet weight) of hyphae were combined with 0.5g (wet weight) fresh leaves of *Sinningia speciosa* (1C = 392Mb; T. Hasing, unpublished) and co-chopped. After chopping, samples were filtered by columns and transferred to chilled mortar and

ground for 15s. After grinding, 1 mL of DeLaat's buffer (see de Laat and Blaas 1984) containing cell constituents was added and the resulting suspension was successively passed through a 50um and 10um filter. A 1:1 v/v amount of staining solution containing propidium iodide (50 ug/ul), RNase (50 ug/ml), and ß-mercaptoethanol (1.1 ul/ml) was added. Samples were gently mixed and incubated in the dark at room temperature for 20 minutes. Samples were then kept in the dark at 4°C until processed in the Flow Cytometry Resource Laboratory at Virginia Tech. Relative fluorescence was measured with the FL2 detector, and DNA content was quantified with FL2-area (integrated fluorescence) and displayed on histograms (Baldwin and Husband 2013).

All data and processing scripts associated with this project have been deposited with VTechData under doi: XXX-XXXXXX. The assembly, CDS, and annotation files are publicly available at https://genomevolution.org/coge/.

ACKNOWLEDGEMENTS

The authors thank Qian Zhang for her work developing a Flow Cytometry protocol for *P. capsici*, as well as Melissa Makris and the Flow Cytometry Resource Lab in the Virginia Tech College of Veterinary Medicine, for conducting the flow analyses. We thank Kurt Lamour for providing P. capsici isolate LT263. We are grateful to the *Phytophthora* community for providing public access to genomic datasets. This work was supported by a grant to DCH, ABG, and JMM from the Fralin Life Science Institute at Virginia Tech.

TABLES & FIGURES

a) Assembly	Total assembly (Mbp)	No. of contigs	Longest contig (Mbp)	N50 of contigs	L50 of contigs (kbp)
Scaffolds	124.2	906	1.26	174	199.6
Contigs	123.4	1132	1.26	232	158.9
b) Completeness	**Complete BUSCOs**	**Single copy**	**Duplicated**	**Fragmented**	**Missing**
Raw assembly	53.90%	43.60%	10.30%	4.30%	41.80%
Long read polish	66.60%	47.40%	19.20%	3.00%	30.40%
Short read polish	96.50%	56.80%	39.70%	0.90%	2.60%
LT1534 assembly	91.00%	91.00%	0.00%	2.10%	6.90%

Table 1. Genome assembly and completeness summary data for the *P. capsici* genome.

Annotation	
GC content of the genome (%)	50.95
No. predicted protein-coding genes	24,160
No. protein-coding mRNAs	40,159

Table 2. Genome assembly annotation summary data for the P. capsici genome.

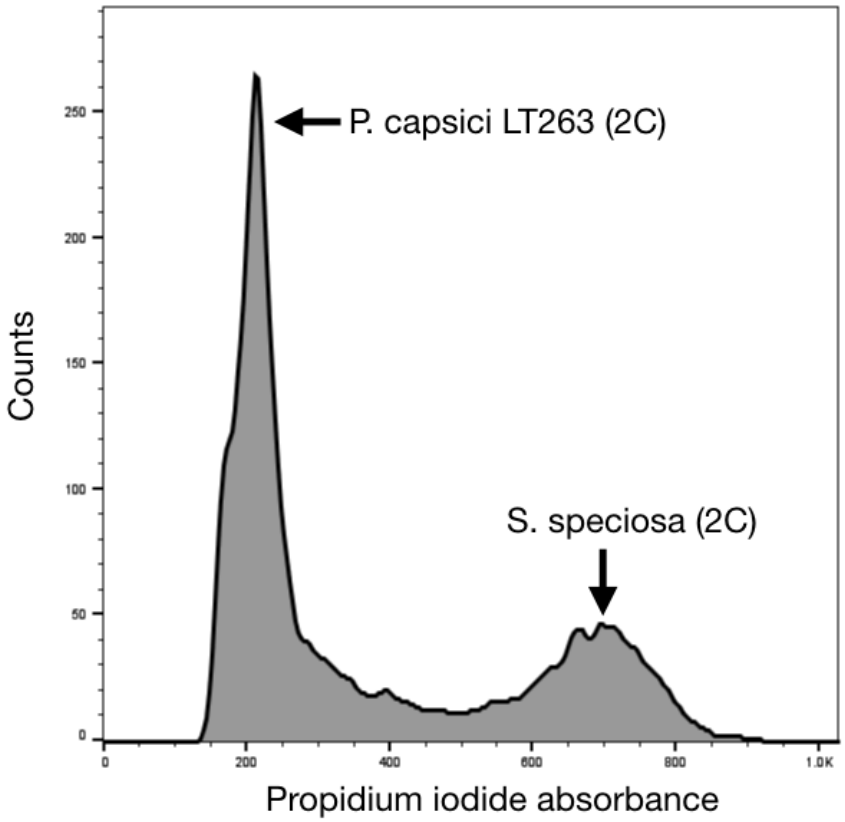


Figure 1. Flow cytometry plot confirming the larger assembly size for *P. capsici* LT263. The first peak is *P. capsici* (~110Mb) the second peak is an internal standard from *Sinningia speciosa* (~392Mb).

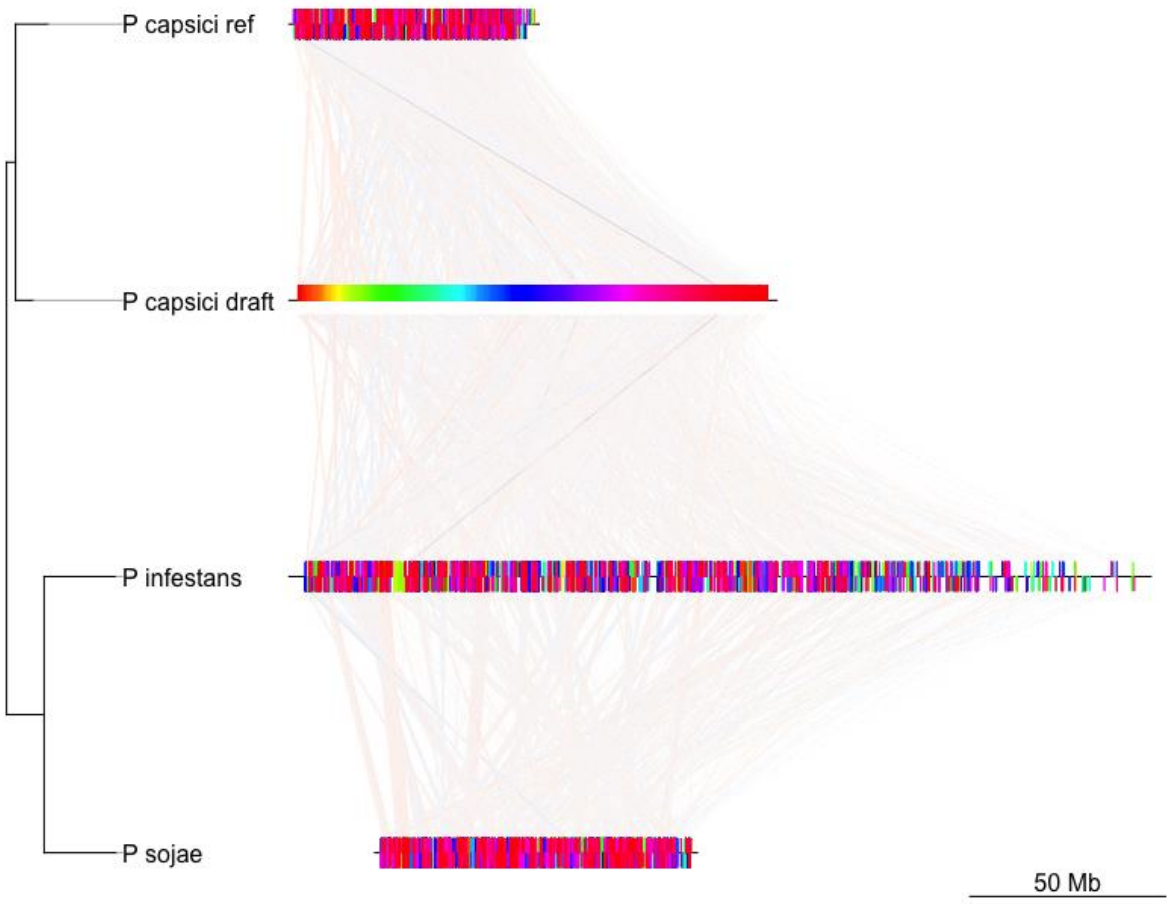


Figure 2. Syntenic plot between *P. sojae*, *P. capsici* (draft), and the published *P. capsici* genome. Regions connected by lines are considered shared blocks between the genomes. Bars above the line in the reference match the orientation in the draft while colored bars below the horizontal line represent inverted genes. The extended length of the horizontal line in the draft assembly represents the expanded genome.

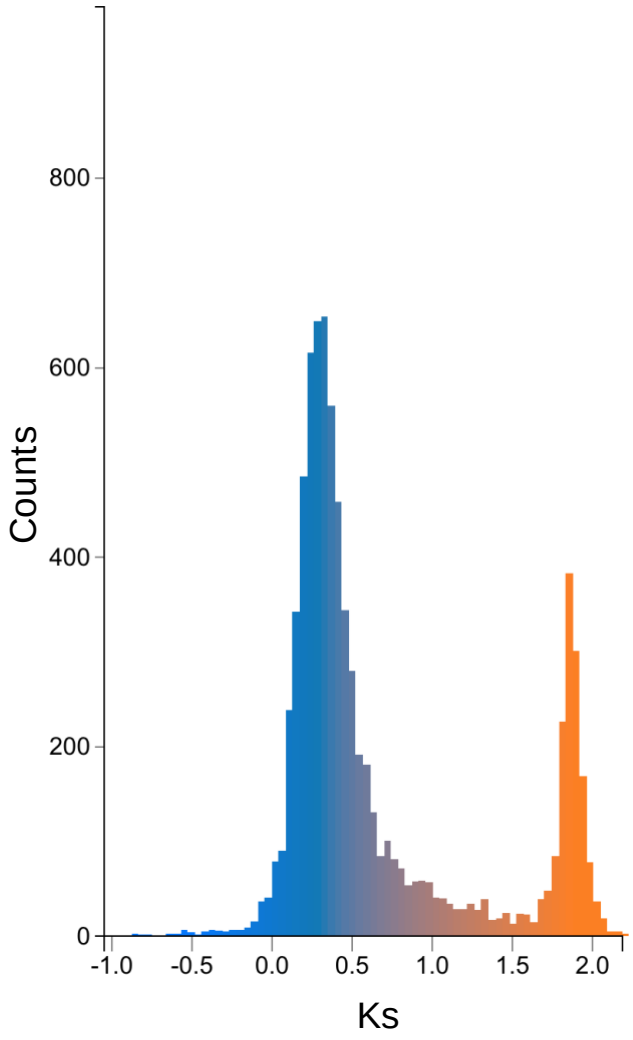


Figure 3. Synonymous substitution (Ks) plot indicating the presence of duplication in this lineage. Visualization was produced in CoGe and can be reproduced here: https://genomevolution.org/coge/SynMap.pl

SUPPLEMENTAL MATERIALS

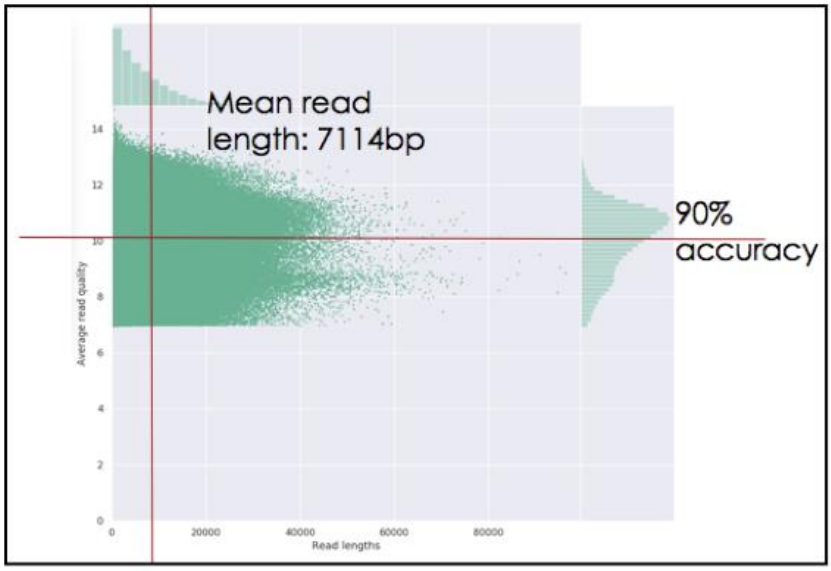


Figure S1. Raw read lengths generated from Oxford Nanopore sequencing versus raw read quality. The average read length was 7114bp with 90% read accuracy.

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

06 April 2019, Saturday, 07:00:51

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best ☑ Show heatmap

Statistics without reference	pcap_canu_scaffolds	pcap_canu_scaffolds_broken	racon_sr_scaffolds	racon_sr_scaffolds_broken
# contigs	1358	1977	906	1132
# contigs (>= 0 bp)	1358	1977	906	1132
# contigs (>= 1000 bp)	1358	1977	906	1132
# contigs (>= 5000 bp)	1315	1908	906	1130
# contigs (>= 10000 bp)	1269	1799	905	1123
# contigs (>= 25000 bp)	1076	1386	846	1024
# contigs (>= 50000 bp)	734	798	677	776
Largest contig	1 075 963	890 108	1 255 726	1 255 726
Total length	133 884 261	131 692 394	124 231 288	123 385 214
Total length (>= 0 bp)	133 884 261	131 692 394	124 231 288	123 385 214
Total length (>= 1000 bp)	133 884 261	131 692 394	124 231 288	123 385 214
Total length (>= 5000 bp)	133 791 603	131 535 699	124 231 288	123 378 017
Total length (>= 10000 bp)	133 445 618	130 687 451	124 225 275	123 326 505
Total length (>= 25000 bp)	130 040 483	123 336 713	123 003 374	121 318 098
Total length (>= 50000 bp)	117 597 163	101 978 239	116 880 513	112 365 994
N50	173 812	111 141	199 566	158 856
N75	86 473	53 363	113 112	91 983
L50	199	300	174	232
L75	473	736	382	490
GC (%)	51.06	51.06	50.95	50.95
Mismatches				
# N's	2 191 942	75	846 121	47
# N's per 100 kbp	1637.19	0.06	681.09	0.04

Figure S2. Quast quality comparison between the custom pipeline used to assemble the ONT generated *P. capsici* draft and the standard pipeline Canu v1.7. Total assembly metrics indicate that the custom pipeline generated a more contiguous draft genome.

	Complete BUSCOs	***Single copy***	***Duplicated***	***Fragmented***	***Missing***
ONT-Raw	53.9%	43.6%	10.3%	4.3%	41.8%
Long read polish	66.6%	47.4%	19.2%	3.0%	30.4%
Short read polish	96.5%	56.8%	39.7%	0.9%	2.6%
Reference	91.0%	91.0%	0%	2.1%	6.9%

Table S1. Comparison of BUSCO analyses between the draft assembly of LT263 and the reference assembly from LT1534.

REFERENCES

Baldwin, S. J., and Husband, B. C. 2013. The association between polyploidy and clonal reproduction in diploid and tetraploid Chamerion angustifolium. Mol Ecol. 22:1806–1819

Boetzer, M., and Pirovano, W. 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. 15:211

Chu, J., Mohamadi, H., Warren, R. L., Yang, C., and Birol, I. 2017. Innovations and challenges in detecting long read overlaps: an evaluation of the state-of-the-art. Bioinformatics. 33:1261–1270

Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Research, 14(7), pp.1394-1403.

de Laat, A. M., and Blaas, J. 1984. Flow-cytometric characterization and sorting of plant chromosomes. TAG Theoretical and Applied Genetics. 67:463–467

Fletcher, K., Gil, J., Bertier, L.D., Kenefick, A., Wood, K.J., Zhang, L., Reyes-Chin-Wo, S., Cavanaugh, K., Tsuchida, C., Wong, J. and Michelmore, R., 2019. Genomic signatures of somatic hybrid vigor due to heterokaryosis in the oomycete pathogen, Bremia lactucae. bioRxiv, p.516526.

Galbraith, D. W., Harkins, K. R., Maddox, J. M., Ayres, N. M., Sharma, D. P., and Firoozabady, E. 1983. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. Science. 220:1049–1051

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29:1072–1075

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27:722–736

Lamour, K.H., Stam, R., Jupe, J. and Huitema, E., 2012a. The oomycete broad-host-range pathogen Phytophthora capsici. *Molecular plant pathology*, *13*(4), pp.329-337.

Lamour, K. H., Mudge, J., Gobena, D., Hurtado-Gonzales, O. P., Schmutz, J., Kuo, A., Miller, N. A., Rice, B. J., Raffaele, S., Cano, L. M., Bharti, A. K., Donahoo, R. S., Finley, S., Huitema, E., Hulvey, J., Platt, D., Salamov, A., Savidor, A., Sharma, R., Stam, R., Storey, D., Thines, M., Win, J., Haas, B. J., Dinwiddie, D. L., Jenkins, J., Knight, J. R., Affourtit, J. P., Han, C. S., Chertkov, O., Lindquist, E. A., Detter, C., Grigoriev, I. V., Kamoun, S., and Kingsmore, S. F. 2012b. Genome Sequencing and Mapping Reveal Loss of Heterozygosity as a Mechanism for Rapid Adaptation in the Vegetable Pathogen Phytophthora capsici. Molecular Plant-Microbe Interactions. 25:1350–1360

Leonian, L.H., 1922. Stem and fruit blight of peppers caused by Phytophthora capsici sp. nov. *Phytopathology*, *12*(9).

Li, H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics. :1–8

Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.

Malar, C. M., Yuzon, J. D., Das, S., Das, A., Panda, A., Gosh, S., Tyler, B. M., Kasuga, T., and Tripathy, S. 2019. Haplotype-phased genome assembly of virulent

Phythophthora ramorum isolate ND886 facilitated by long-read sequencing reveals effector polymorphisms and copy …. Molecular Plant-Microbe Interactions.

Martens, C., and Van de Peer, Y. 2010. The hidden duplication past of the plant pathogen Phytophthora and its consequences for infection.

Raffaele, S. and Kamoun, S., 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nature Reviews Microbiology, 10(6), p.417.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212

Stam, R., Jupe, J., Howden, A. J. M., Morris, J. A., Boevink, P. C., Hedley, P. E., and Huitema, E. 2013. Identification and Characterisation CRN Effectors in Phytophthora capsici Shows Modularity and Functional Diversity D. Arnold, ed. PLoS ONE. 8:e59517–13

Treangen, T. J., and Salzberg, S. L. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 13:36–46

Tyler, B.M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R.H., Aerts, A., Arredondo, F.D., Baxter, L., Bensasson, D., Beynon, J.L. and Chapman, J., 2006. Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*, *313*(5791), pp.1261-1266.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27:737–746

Yandell, M., and Ence, D. 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13:1–14

Yang, M., Duan, S., Mei, X., Huang, H., Chen, W., Liu, Y., Guo, C., Yang, T., Wei, W., Liu, X., He, X., Dong, Y., and Zhu, S. 2018. The Phytophthora cactorum genome provides insights into the adaptation to host defense compounds and fungicides. Sci. Rep. :1–11

Chapter 5. Conclusions

Contemporary crops are exposing to severely unfavorable conditions under the projection of global climate change. Unlocking the natural genetic diversity holds the promise to equip our modern crop system with enhanced resilience. Yet, one of the greatest barriers is our poor understanding on genetic basis of agronomic functional traits. The promise, as yet incompletely filled, of recent advances in sequencing technology allow us gating access to the genetic basis of crops defensive traits and pathogenesis. Here we integrated bioinformatic and genomic approaches to improve plant resilience to leverage next generation sequencing to uncover natural genomic variation in non-model plants. Specifically, we developed tools that improve the ability to detect full-length gene models from non-model species, characterize complex stress pathway gene networks, and deploy a rapid sequencing and assembly method that goes from sampling to whole genome assembly in just 6 days. First, we develop a machine learning method for evaluating transcriptome assembly and quality, therein improving gene models for downstream uses of RNAseq data such as identifying differentially expressed genes. Next, we identify the complex network of genes that are associated with antagonistic stress response pathway interactions. We also identified natural genetic variation in this response that can be used to improve cultivated tomato. Finally, we leverage the third generation of sequencing, single molecule approaches, to identify genomic variation that was previously hidden by genomic complexity, in a crop plant pathogen.

Transcriptome assembly completeness is a function of the relative integrity of the gene models developed by the assembly algorithm and is essential to downstream applications

of RNAseq data. Coupled to completeness is assembly quality which includes completeness but additionally estimates the integrity of the assembly. These are metrics such as, number of input bases assembled into contigs, or the proportion of raw reads that map to the transcriptome assembly. Identifying the factors that drive assembly quality and integrity is paramount to improving the application of RNAseq tools to non-model systems, such as wild crop relatives. To identify the factors that impact transcriptome assemblies we used comparative transcriptomics to assess factors like assembly approach (*de novo* and reference guided), phylogenetic distance (nucleotide diversity), and number of input reads, and common quality scores as the response variable. Here we found that reference guided approaches are better than *de novo* approaches up to about 5% nucleotide divergence. Additionally, we identified the most important factors shaping assembly quality and completeness and devised some best practices for RNAseq experiments.

We also used the factors involved in shaping assembly quality to develop a reference-free machine learning approach for transcriptome quality evaluation. This approach provides assembly quality scores that are on par with the gold standard tool BUSCO but is much more efficient and operates without the need for reference sequences. Further, this machine learning tool “WWMT” archived an accuracy of 0.99, 0.95 and 0.95 in predicting BUSCO completeness score on, the training, validation, and test datasets. Thus, we have provided a set of best practices and an evaluation tool for use in non-model systems to improve the accuracy of identifying important genomic variation in these systems.

Identifying the genomic basis for complex traits like stress responses, a long-standing issue, is finally attainable using next generation sequencing approaches. This is made even more complicated through the interaction between stress response pathways. Thus, to understand the gene networks involved in 'real-world' stress conditions we have to unravel these interactions which, in tomato, are often antagonistic. We leverage our transcriptome assembly tools and experimental transcriptomics to generate transcriptomic responses between two wild tomato accessions under drought stress, mimicked herbivory stress, and the combined stressors. In doing so, we have identified the co-expression networks that exist for each stress alone and combined and the set of genes uniquely expressed in the combined condition, wherein we find the targets of stress pathway interaction. Here we pinpoint the clear downregulation of genes controlling the biotic stress response pathway under the combined stress of drought and mimicked herbivory. This downregulation is targeted at several major genes in the pathway and across several regulatory mechanisms, thus exhibiting a multilevel regulation.

Interestingly, we uncover natural variation in the mimicked herbivory treatment and again in combined stress treatment where one accession of wild tomato (LA1589) shows no phenotypic response. Yet, under the mimicked herbivory treatment biotic stress pathway genes are differentially expressed, but to a much lower extent than in the responsive accession. This coupled with physiological changes under the mimicked herbivory treatment suggests that the primary genetic components of the pathway are intact, but that secondary genetic components associate with plant defense are not. Further, there does not appear to be downregulation of the same key regulatory genes under the combined stress treatment. This suggests that these two accessions might be important

sources of genetic variation for secondary defense and breaking stress pathway antagonism in cultivated tomato.

Finally, rapid identification of genetic varation in plant pathogens is an important step in identifying durable resistance mechanisms. The Oxford Nanopore Technologies (ONT) MinIon platform has drastically dropped the cost of long read sequencing to an unprecedent level. We have generated 1,258,480 reads (~10Gb) from a single flowcell with an average read length of 7,114bp. Taking advantage of these long reads, we were able to span more repetitive regions in the genome. Therefore, we assembled this draft genome into 124 Mbp (scaffold N50 = 232kb), which is almost twice the size of the current draft genome, 64Mbp.

Though Oxford Nanopore long reads effectively resolved the complex repetitive regions in our draft genome, assemblies generated from such error prone reads need to be polished to generate a completer and more accurate genome. Notably, polishing with uncorrected nanopore long reads has limited potential to improve assembly completeness. However, because of sequencing projects aimed at pathogens we leveraged publicly available short read data, to perform a second round of polishing that tremendously improved assembly quality, uncovering 96.6% of complete BUSCOs. Our approach demonstrated a protocol in using high error rate sequencing reads assembly genome.

Another challenge in noisy long read assembly is that it is quite computationally intensive. Here we streamlined a computational pipeline with the existing tools to rapidly complete assembly, scaffold and polish our assembly. Our computational pipeline takes advantage

of these state of art algorithms that greatly accelerates computation efficiency. Together our approach demonstrated a cost-effective protocol to rapidly recover complex regions of a pathogen genome, revealing hidden variation.

Overall, we have developed a series of integrated approaches aimed at identifying important natural genomic variation that can be leveraged to improve agricultural plant resilience. In Chapters 2 and 3, our focus is on developing genomic tools that can be used to rapidly identify natural genetic variation among wild crop relatives. This variation has long stood as an important source of genetic improvement for plant breeders. The data generated from our work will be useful for genomic improvement of cultivars, as well as the re-domestication of additional wild species. Herein, uncovering genetic diversity that is needed to combat an ever-changing climate. In Chapter 4, we turn our attention toward methods that will assist in developing robust disease resistance in cropping systems. Namely, developing and deploying approaches that can be used to identify genomic variation among pathogens, in nearly real-time. Here, our approach led to a more complete and deconvoluted genome assembly than the previously established reference genome, that was rapid and cost-effective. Identifying important genomic variation among pathogen isolates will allow the directed discovery of control methods, including genetic based resistance in the crop host. The integrated approaches to uncovering genomic variation developed in this dissertation will be instrumental in the development of resilient crop plants in the future.