PHENOTYPIC AND GENETIC STUDIES OF GRAPEVINE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Konstantin Divilov

December 2017

PHENOTYPIC AND GENETIC STUDIES OF GRAPEVINE

Konstantin Divilov, Ph.D.

Cornell University 2017

Plant breeding is the science of altering a plant's genetics to attain a desired phenotype. In this dissertation, I explore what phenotypes to measure when breeding for downy mildew resistance and improved floral scent and how to measure these phenotypes accurately and efficiently. Traditionally, downy mildew resistance has been measured by visually rating sporulation and hypersensitive response on leaves or leaf discs. However, such manual ratings become intractable when dealing with thousands of samples. Therefore, to measure sporulation on leaf discs, I developed a computer vision system that reduced phenotyping time by more than 90% when compared to manual ratings, and also was found to work well for phenotyping leaf trichomes. If phenotypes are collected in the vineyard, spatial variation from inoculum, soil, and microclimate might have an effect on these phenotypes. Testing this assumption, spatial processes explained some variance in vineyard phenotypes, but accounting for the spatial variance might not lead to significantly more accurate phenotypes. Quantitative phenotyping of floral scent for large numbers of grapevines using headspace analysis is not economically feasible, so I evaluated the robustness of a hexane extraction followed by gas chromatography-mass spectrometry to identify floral volatiles and found that it was robust regardless of extraction time when flowers were sampled from the same inflorescence. After obtaining phenotypes and genotypes of vines, quantitative trait loci are found, traditionally using one phenotype at a time. In our case, understanding how sporulation, HR, and leaf trichomes affected each other was of interest, in addition to how genetic markers affected the phenotypes, so I used Bayesian networks to explore these interactions. In one of two F_1 families studied, HR had a positive effect on sporulation, and leaf trichomes had a negative effect on both HR and sporulation, suggesting that leaf trichome density can be selected for in breeding for downy mildew disease resistance. A breeding project was started with the intention of creating a dwarf grapevine with an attractive floral scent. With a complementary interest to understand what volatile compounds were responsible for the various floral scents in grapevine, a diverse set of genotypes from various *Vitis* spp. were phenotyped for floral scent and volatiles, and it was found that similar scents were generated from different sesquiterpene profiles. Overall, this dissertation spans key concepts in the science of plant breeding, from parental selection and hybridization, to phenotyping by computer vision and chemical analysis, to statistical analyses of interacting phenotypes, genotypes, and spatial variability, with the findings possibly enhancing grapevine breeding strategies and execution.

BIOGRAPHICAL SKETCH

Konstantin Divilov was born in Baku, Azerbaijan in 1991. He moved to Brooklyn, New York when he was four and spent many summers living near the Catskill Mountains. It was this time in Upstate New York that gave him a love for nature. He obtained a B.A. in Biology from Hunter College in 2012, and a M.S. in Crop Sciences from the University of Illinois at Urbana-Champaign in 2014, working in the soybean breeding program of David R. Walker. He returned to Upstate New York for his Ph.D. in Plant Breeding and Genetics at Cornell University, working on grapevine breeding and genetics under Bruce I. Reisch.

ACKNOWLEDGMENTS

I thank Bruce Reisch for giving me the independence to explore my research interests, for asking questions that have led to improvements in all chapters of this thesis, and for giving me sage advice. I thank Lance Cadle-Davidson for getting me started with my downy mildew work and for providing very good edits on our manuscripts. I thank Michael Gore for reviewing my manuscripts. I would like to thank Bruce, Lance, and Michael for taking the time to advise me.

I thank the Plant Breeding and Genetics graduate students for wonderful fraternity and all my friends and residents of Ithaca and Geneva, who have kept me company and taught me so much over the years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

COMPUTER VISION FOR HIGH-THROUGHPUT QUANTITATIVE PHENOTYPING: A CASE STUDY OF GRAPEVINE DOWNY MILDEW SPORULATION AND LEAF TRICHOMES

1.1 Abstract

Quantitative phenotyping of downy mildew sporulation is frequently used in plant breeding and genetic studies, as well as in studies focused on pathogen biology, such as chemical efficacy trials. In these scenarios, phenotyping a large number of genotypes or treatments can be advantageous, but is often limited by time and cost. We present a novel computational pipeline dedicated to estimating the percent area of downy mildew sporulation from images of inoculated grapevine leaf discs in a manner that is time and cost efficient. The pipeline was tested on images from leaf disc assay experiments involving two F_1 grapevine families, one that had glabrous leaves (*V. rupestris* B38 × 'Horizon' [RH]) and another that had leaf trichomes ('Horizon' × *V. cinerea* B9 [HC]). Correlations between computer vision and manual visual ratings reached 0.89 in the RH family and 0.43 in the HC family. Additionally, we were able to use the computer vision system prior to sporulation to measure the percent leaf trichome area. We estimate that an experienced rater scoring sporulation would spend at least 90% less time using the computer vision system compared to the manual visual method. This will allow more treatments to be phenotyped in order to better understand the genetic architecture of downy mildew resistance and of leaf trichome density. We anticipate this computer vision system will find applications in other pathosystems or traits

where responses can be imaged with sufficient contrast from the background.

1.2 Introduction

Downy mildew of grapevine (*Vitis* spp.), caused by the obligate biotrophic oomycete *Plasmopara viticola* (Berk. & M.A. Curtis) Berl. & de Toni, is a major disease in humid grape-growing regions around the world. The disease is particularly devastating to *Vitis vinifera* cultivars, which lack genetic resistance to the disease. On the other hand, most native North American species, e.g., *V. riparia*, and certain *V. vinifera* hybrids with North American species have genetic resistance to downy mildew (Cadle-Davidson 2008) and are commonly used as parents in resistance breeding efforts and in genetic studies. The pathogen can only infect the abaxial side of leaves and developing berries because it enters via stomata. With leaves, the stomata are located on the abaxial surface, while on developing berries, the stomata are only present before developing into lenticels, after which infection is no longer possible (Kennelly et al. 2005). After leaf infection, three possible symptoms and/or signs may appear. The first is the hypersensitive response, which presents as black necrotic specks on the underside of the leaf (Bellin et al. 2009). The second is sporulation of the pathogen, which appears as white fuzz on the abaxial side of the leaf (Kennelly et al. 2005). The third is yellow spots that have an oily appearance on the adaxial surface, and this symptom is spatially connected to the abaxial sporulation. This paper will focus on the occurrence of sporulation. The most common way to screen for sporulation in genetic studies has been to use a leaf disc assay where a leaf disc is extracted from a leaf using a cork borer and plated on a medium before inoculation with a single *P. viticola* isolate or multiple *P. viticola* isolates, after which a visual rating using a scale of sporulation area is taken (Bellin et al. 2009).

Grapevine species are diverse with respect to their leaf trichome morphology (Ma et al. 2016). Their hydrophobic leaf trichomes have been hypothesized to play a role in resistance to downy mildew by preventing germination of zoospores and preventing germinated sporangia from reaching stomata (Kortekamp and Zyprian 1999). However, determining causality between leaf trichome density and downy mildew resistance requires an experiment without the confounding effect between ancestry and leaf trichome density. Leaf trichomes have traditionally been measured using International Organisation of Vine and Wine (OIV) descriptors (Loughner et al. 2008; Paolocci et al. 2015). Due to the possible causal link between leaf trichome density and downy mildew resistance, accurately phenotyping leaf trichome density is important for downstream analyses in distinguishing effects of genomic variants due to morphological and non-morphological traits.

In the breeding of crops, one desires to decrease the time and money spent evaluating cross progeny while increasing the number of cross progeny being screened without sacrificing phenotype quality. A common downy mildew screening method in grapevine breeding programs is to visually screen grapevines in the field following natural infection periods (Eibach 1998). While inexpensive, the method can give inaccurate ratings for genotypes, especially in years with mild disease pressure. In quantitative trait loci (QTL) studies, increasing the number of individuals that are phenotyped increases the statistical power of analyses designed to locate significant QTL. High-throughput plant phenotyping strategies would help to accomplish these goals in a cost-effective manner (Pauli et al. 2016). Numerous methods to use computer vision to detect and quantify plant diseases and their symptoms have been developed (Barbedo 2013). The benefits of using computer vision rather than manual ratings to assess disease include greater precision and accuracy; decreased time needed for ratings; and reduced labor costs (Bock et al. 2010).

Two computer vision methods using the leaf disc assay have been developed for rating downy mildew sporulation, but they are limited in their use. In one method (Peressotti et al. 2011), a single image is required for every leaf disc that is being phenotyped, which is a time-consuming step if there are thousands of leaf discs to phenotype. In the other method (Khiook et al. 2013), the leaf discs need to be removed from their original medium and the images need to be taken on a photographic reproduction bench under artificial lighting, increasing the time and cost associated with the phenotyping. Additionally, proprietary software is used to do the analysis, which further increases the cost of the method. We present a method of using images taken from a smartphone camera that capture inoculated leaf discs in bulk, and open source scripts are then used to quantify percent area of downy mildew sporulation. We show that the same method can be used to quantify the percent leaf trichome area.

1.3 Materials and Methods

1.3.1 Plant Material

Two F_1 grapevine families, *V. rupestris* B38 × 'Horizon' (*Vitis* sp. interspecific hybrid) (RH) and 'Horizon' × *V. cinerea* B9 (HC), were grown unreplicated in a vineyard in Geneva, New York (Hyma et al. 2015). *V. cinerea* B9 and the HC family had white trichomes on the abaxial side of their leaves, while leaves of

‘Horizon’, *V. rupestris* B38, and the RH family were glabrous. In 2015, 163 and 152 F_1 genotypes from the RH and HC families, respectively, were phenotyped for downy mildew sporulation with three experiments conducted with the RH family and two with the HC family. In 2016, only 157 and 145 genotypes from the respective families were phenotyped with two experiments conducted with both families; some genotypes were not available due to cold injury. The starting dates of the experiments for the RH family were 22 June 2015, 29 June 2015, 13 July 2015, 5 July 2016, and 27 July 2016 and those for the HC family were 15 June 2015, 6 July 2015, 27 June 2016, and 12 July 2016.

1.3.2 *Plasmopara viticola* Isolation and Maintenance

In October 2014, a clonal isolate of *P. viticola* was obtained by single sporangiophore isolation as previously described (Cadle-Davidson 2008) from a leaf of an organically grown ‘Frontenac Gris’ (Luby and Hemstad 2006) vine at Cornell Orchards in Ithaca, NY. The isolate was maintained by weekly transfers to surface-sterilized, susceptible leaves plated abaxial side up, on a 1% agar Petri dish. The new leaf was misted with sterile water, and the sporulating leaf was inverted briefly onto the new leaf to touch-transfer sporangia. The susceptible genotypes of leaves used for maintenance were the *V. vinifera* cultivars Chardonnay, Cabernet Sauvignon, and Riesling and the interspecific hybrid cultivar Delaware.

1.3.3 Leaf Disc Assay

For each experiment, the fifth leaf from the growing tip of four shoots of each F_1 genotype of a family and susceptible and resistant controls was harvested and put in a flexible plastic compact disc holder with holes punched out (Cadle-Davidson 2008). Leaves were maintained at 4 to 8℃ until the following day, when they were surface sterilized in 0.5% NaOCl for 2 min and rinsed in sterile double distilled water thrice for 2 min per rinse. Two leaf discs per leaf were punched out using a 1 cm cork borer and plated abaxial side up on 30 × 20 cm Pyrex dishes filled with sterile 1% agar. Each leaf disc replicate (eight in total for each experiment) was placed in a separate dish such that every genotype was represented on every dish. Genotypes were randomized in blocks of 20 in order to control for possible dish location effects while still allowing for rapid plating of leaf discs. Susceptible and resistant controls were included to ensure inoculum quality. For the RH family, 'Cabernet Sauvignon' was the susceptible control for the first two experiments in 2015, while 'Chardonnay' was the susceptible control for all other experiments, and *V. rupestris* B38 was the resistant control. For the HC family, 'Chardonnay' was the susceptible control and *V. cinerea* B9 was the resistant control. On the day after plating, each leaf disc was inoculated with 50 µL containing 5×10^4 sporangia/mL of the *P. viticola* single-sporangial isolate using a HandyStep S repeating pipette (BrandTech Scientific, Essex, CT), and the Pyrex dishes were sealed with plastic wrap and placed in a temperature-controlled room at 23℃ ± 1℃. The inoculation droplets were absorbed the following day with tissue paper (Kimwipes). Sometimes the inoculation droplet failed to adhere to the leaf disc, and phenotypes from these leaf discs were not included in the analyses. Sporulation was evaluated between 3 days post-inoculation (dpi) and 6 dpi for the RH family in all experiments except the second one in 2016 where it was evaluated between 4 dpi and 7 dpi due to a delay in sporulation. Sporulation was evaluated between 4 dpi and 7 dpi for the HC family in all experiments. Evaluation was done using both manual and computer vision methods. Leaf trichomes were quantified with the computer vision method at 2 dpi when no sporulation was visible.

1.3.4 Manual and Computer Vision Rating Methods

The manual rating method followed an established protocol whereby sporulation area on leaf discs was rated on a 1 to 5 ordinal scale, with 1 being no sporulation area, 2 being small sporulation area, 3 being moderate sporulation area, 4 being moderately high sporulation area, and 5 being high sporulation area (Kono et al. 2015). Because all the leaf discs were about the same size, the manual rating is implicitly a percent area of sporulation phenotype. For the computer vision rating method, digital images of leaf discs were taken using the native Camera application (default settings; without flash) of a handheld Apple iPhone 5s with an image resolution of 3264 × 2448 pixels. Each image contained at most 20 leaf discs in four rows and five columns. Images were taken on a black laboratory bench within 0.75 m of a window with the fluorescent room lights on. No other supplemental lighting was used. All images for a particular time point of an experiment were analyzed simultaneously via a pipeline of four Python (Python Software Foundation) scripts developed using OpenCV, an open source computer vision library (Bradski 2000). Of the four scripts, crop.py is used to crop the initial images while values.py is used to find the percent area of sporulation per leaf disc. The other two scripts, circles.py and lines.py, are used to find the correct function parameters that are then used in the values.py script. circles.py is used to find the threshold and Hough circle transform algorithm parameters while lines.py is used to find the Hough line transform algorithm parameters. All the scripts are parallelized such that when run, they automatically use all available CPU cores for faster image processing. The scripts and a guide for the scripts can be found at https://github.com/kdivilov/downymildew-CV.

For the computer vision system, the images were initially cropped such that only leaf discs that were fully contained in an image were kept (Figure 1.1). The cropped images were then converted to Lab color space, which, unlike RGB color space, includes all colors visible to the human eye, with all the layers thresholded using user-specified values and masked over the original image to filter the background, which in our case was agar. The blue layer of the original images, which highlights sporulation better than the other layers, was then selected and filtered using a Wallis filter (Pratt 2007; Wallis 1976) to account for unequal lighting conditions during image capture. A threshold was then applied to the grayscale Wallis-filtered images to only keep the brightest pixels, i.e., only pixels with a value of 255 were kept with the rest set to 0. Leaf veins that were present in the Wallis-filtered images were removed using the Hough line transform algorithm. Leaf discs were detected as circles using the Hough circle transform algorithm. The phenotype obtained at the end was the number of white pixels within a leaf disc divided by the area of the disc in pixels. This corresponded to the percent area of sporulation for a leaf disc in the RH family, while in the HC family it corresponded to the percent area of sporulation and leaf trichomes for a leaf disc. We did not use the white pixel quantity as the phenotype because images were taken at slightly different distances from the leaf discs. The area of a leaf disc thus acted as a scaling factor to allow phenotype comparisons between images.

1.3.5 Statistical Analysis

The manual and computer vision phenotype distributions for both families were obtained using phenotypes averaged across leaf discs and compared using violin

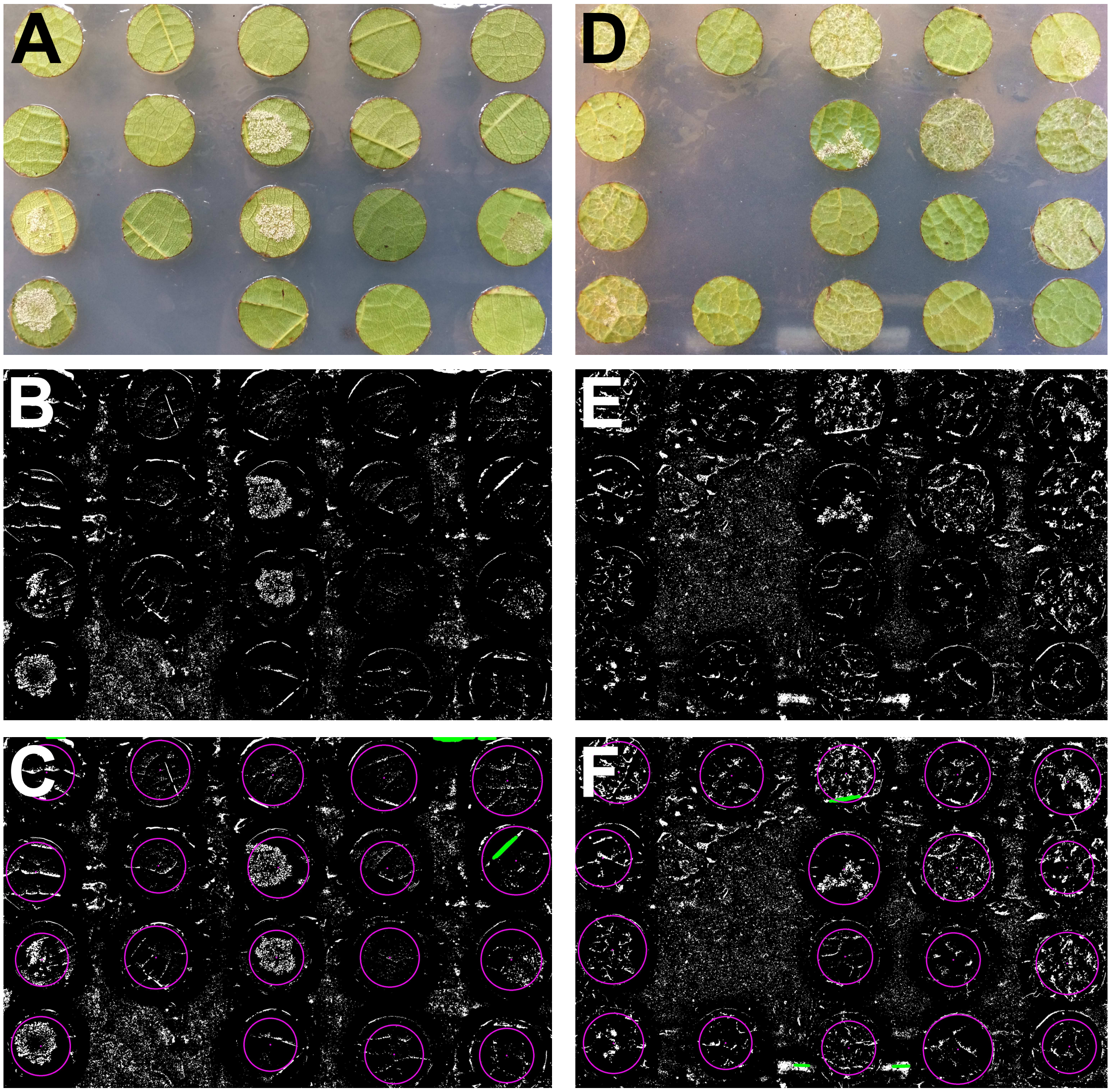

Figure 1.1: Computer vision phenotyping system for grapevine downy mildew sporulation caused by *Plasmopara viticola*. **A**, **B**, and **C**, *Vitis rupestris* B38 × Horizon family and **D**, **E**, and **F**, Horizon × *V. cinerea* B9 family. **A** and **D**, After the initial images have been cropped; **B** and **E**, after they have been passed through the Wallis filter; and **C** and **F**, the same as **B** and **E**, with the exception that the detected leaf discs and their centers and the detected leaf veins have been highlighted.

plots made using the ggplot2 (Wickham 2009) and RColorBrewer (Neuwirth 2014) packages in R (R Core Team). Normality of the distributions was evaluated using the ShapiroWilk test in R where the null hypothesis is that phenotypes are normally distributed. Significance of the test statistic was determined by a z-test with a Bonferroni-corrected alpha value of 0.0025 and 0.0031 for the RH and HC families, respectively, obtained by dividing 0.05 by the number of experimental time points across years and experiments. Because the computer vision phenotype for the HC family measured the percent area of sporulation and leaf trichomes, the effect of leaf trichomes was removed from that phenotype. To do this, a simple linear model was fit with the explanatory variable being the averaged computer vision phenotype for an experiment at 2 dpi, representing the percent area of leaf trichomes, and the response variable being the averaged computer vision phenotype for an experiment at a particular dpi. The residuals plus the intercept from the model added on were the new phenotypes that were used for the comparison. The addition of the intercept did not alter the distribution as it is a constant and was done so that the new phenotypes can have a more intuitive interpretation, viz., the percent area of sporulation when the correlation between the averaged computer vision phenotype and the averaged leaf trichome phenotype is zero. Because of this modification, the lower bound on the computer vision phenotype was no longer zero.

Comparisons of manual and computer vision phenotypes for each dpi within an experiment was conducted in R (Kim 2015) using Spearman's rank correlation coefficients for the RH family and semi-partial Spearman's rank correlation coefficients for the HC family with the significance of the correlations tested using a t-test with a Bonferroni-corrected alpha value of 0.0025 and 0.0031 for the RH and HC families, respectively. The covariate for the computer vision phenotype for the semi-partial Spearman's rank correlation coefficient was the leaf trichome phenotype. Spearman's rank correlation coefficient rather than Pearson's correlation coefficient was used because the association between the manual and computer vision phenotypes was non-linear. Comparisons were made using averaged and unaveraged phenotypes.

To test whether the computer vision phenotype at 2 dpi in each experiment was capturing the leaf trichome phenotype, QTL analyses were performed as in Hyma et al. (2015) using a previously created de novo curated genetic map from genotyping-by-sequencing SNP linkage data of the HC family (Hyma et al. 2015), except that the genotyping error rate used to calculate conditional genotype probabilities was set to 0.001. An approximate Bayes credible interval was used to find the region a QTL resides with probability 0.95. Physical locations of the SNPs were obtained using the 12X.2 version (URGI 2014) of the grapevine reference genome PN40024 (Jaillon et al. 2007).

1.4 Results

A majority of progeny and susceptible control samples had visual signs of sporulation within 7 dpi. The average inoculation failure rate, i.e., the rate at which the inoculation droplet failed to adhere to the leaf disc, was 8.9% across all experiments. The average manual ratings for the susceptible control for the experiments with the RH family at the final time point measured were 4.2, 5.0, and 3.9 in 2015 and 2.8 and 5.0 in 2016. The average manual ratings for the resistant control for the experiments with the RH family at the final time point measured were 1.4, 1.4, and 1.3 in 2015 and 2.3 and 1.5 in 2016. The average manual ratings for the susceptible control for the experiments with the HC family at the final time point measured were 5.0 and 4.9 in 2015 and 5.0 and 5.0 in 2016. The average manual ratings for the resistant control for the experiments with the HC family at the final time point measured were 1.0 and 1.0 in 2015 and 1.3 and 1.0 in 2016.

Using the manual visual rating method required 180 min for an experienced rater to phenotype 1336 leaf discs on eight Pyrex dishes and 30 min to input the phenotypes into a spreadsheet on a computer. Using the computer vision system from raw images required 3 min for an experienced rater to capture 72 images spanning these same samples and 15 min to phenotype the leaf discs by executing four scripts. Using all four logical cores on a 3.10 GHz Intel Core i3-2100 Processor, crop.py, circles.py, lines.py, and values.py took an average 1 min 38 s, 1 min 28 s, 2 min 45 s, and 2 min 16 s, respectively, when run on the 72 images. Across four consecutive days of data collection and analysis, manual rating time totaled 840 min while the computer vision pipeline required 72 min, an estimated time savings of 12.8 h (91%) per experiment and 115.2 h across all 9 experiments.

The computer vision phenotypes for the RH family were more often normally distributed than the manual phenotypes, which were often skewed (Figure 1.2A and B). For the HC family, the computer vision phenotypes were close to being normally distributed, but at times were skewed, while the manual phenotypes were never normally distributed (Figure 1.2C and D). Spearman's rank correlation coefficients for the RH family were between 0.10 and 0.81 for timepoints from 4 to 7 dpi for the unaveraged phenotypes (Figure 1.3A). Correlations at 3 dpi, and at 4 dpi in the third experiment in 2015, were low, and these dpi corresponded to time points when sporulation was sparse (Figure 1.2B). Semi-partial Spearman's rank correlation coefficients for the HC family were between -0.04 and 0.33 for the unaveraged scores (Figure 1.3C). For the unaveraged phenotypes, all correlations were significant with the exception of the correlations at 3 dpi for the second experiment in 2015 for the RH family and at 4 dpi for the second experiment in 2015 for the HC family, as well as the correlations at all time points for the second experiment in 2016 for the HC family. Averaging the eight scores for each individual gave higher Spearman's rank correlation coefficients for all time points in the RH family except at 3 dpi in the second experiment in 2015 and at 3 dpi and 4 dpi in the third experiment in 2015 (Figure 1.3B). Averaging also gave higher semi-partial Spearman's rank correlation for all time points in the HC family except at 4 dpi and 5 dpi in the second experiment in 2016 (Figure 1.3D). For the averaged phenotypes in the RH family, all correlations were significant with the exception of the correlations at 3 dpi for the first and second experiments in 2015 and at 3 and 4 dpi for the third experiment in 2016. For the averaged phenotypes in the HC family, all correlations were significant with the exception of the correlations at 4 and 5 dpi for the first experiment in 2015 and at 4, 5, and 7 dpi for the second experiment in 2015, as well as the correlations at all time points for the second experiment in 2016.

QTL analyses using the 2 dpi phenotype in each experiment of the HC family found two QTL in each experiment, one each on chromosomes 5 and 8 from 'Horizon'. The 95% approximate Bayes credible interval for each of the two QTL overlapped across experiments. The percent phenotypic variance explained by the QTL on chromosomes 5 and 8 ranged from 10.3% to 21.6% and 13.0% to 18.1%, respectively.

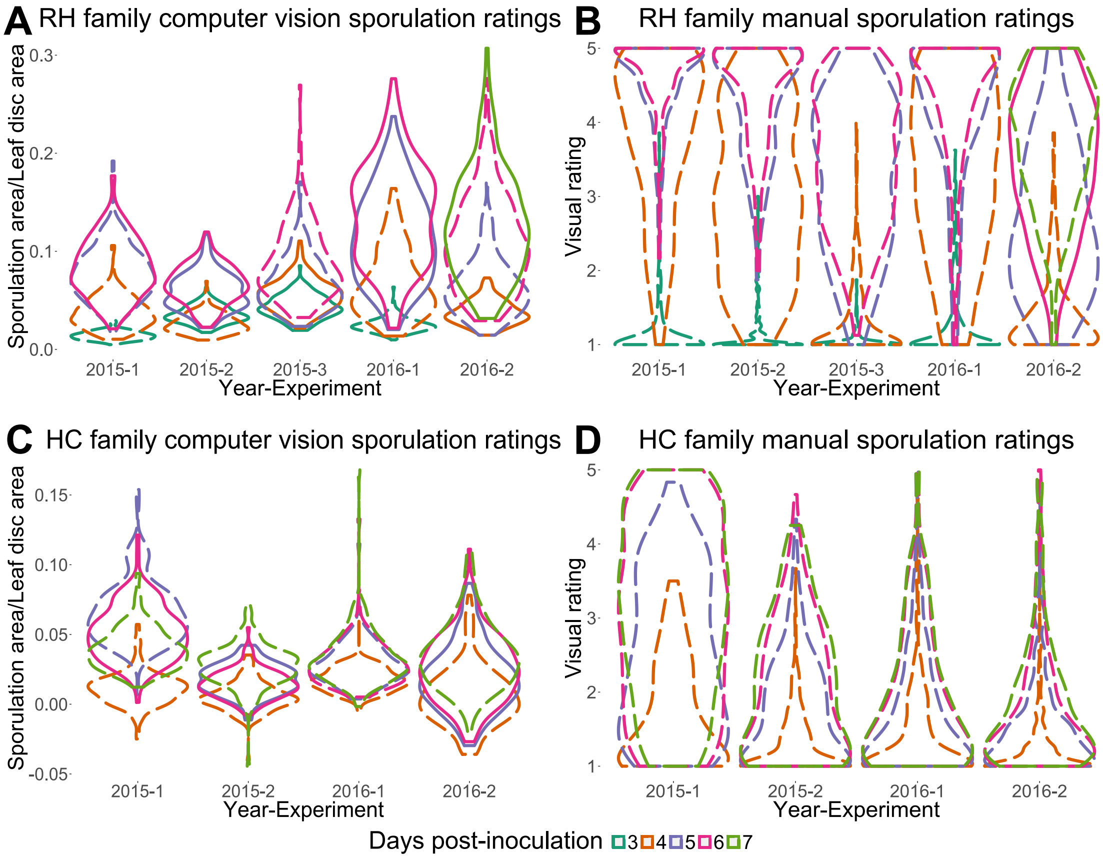

Figure 1.2: Violin plots showing the **A**, computer vision and **B**, manual phenotype distributions of the *Vitis rupestris* B38 × Horizon (RH) family for all experiments and the **C**, computer vision and **D**, manual phenotype distributions of the Horizon × *V. cinerea* B9 (HC) family for all experiments. The computer vision phenotypes for the HC family were processed to account for the variation in leaf trichome density. The computer vision phenotype measures the percent area of sporulation by *Plasmopara viticola* whereas the manual phenotype measures the area of sporulation using a 1-to-5 ordinal scale. Distributions with dashed lines indicate those significantly different than normal tested using a Shapiro-Wilk test of normality with a Bonferroni-corrected alpha value.

1.5 Discussion

Key challenges in phenotyping are selecting the response that reflects the treatment effect and accurately measuring that response at the right time (Cadle-Davidson et al. 2016). In our case, multiple factors from penetration to colonization to sporulation may affect grapevine downy mildew resistance (Boso and Kassemeyer

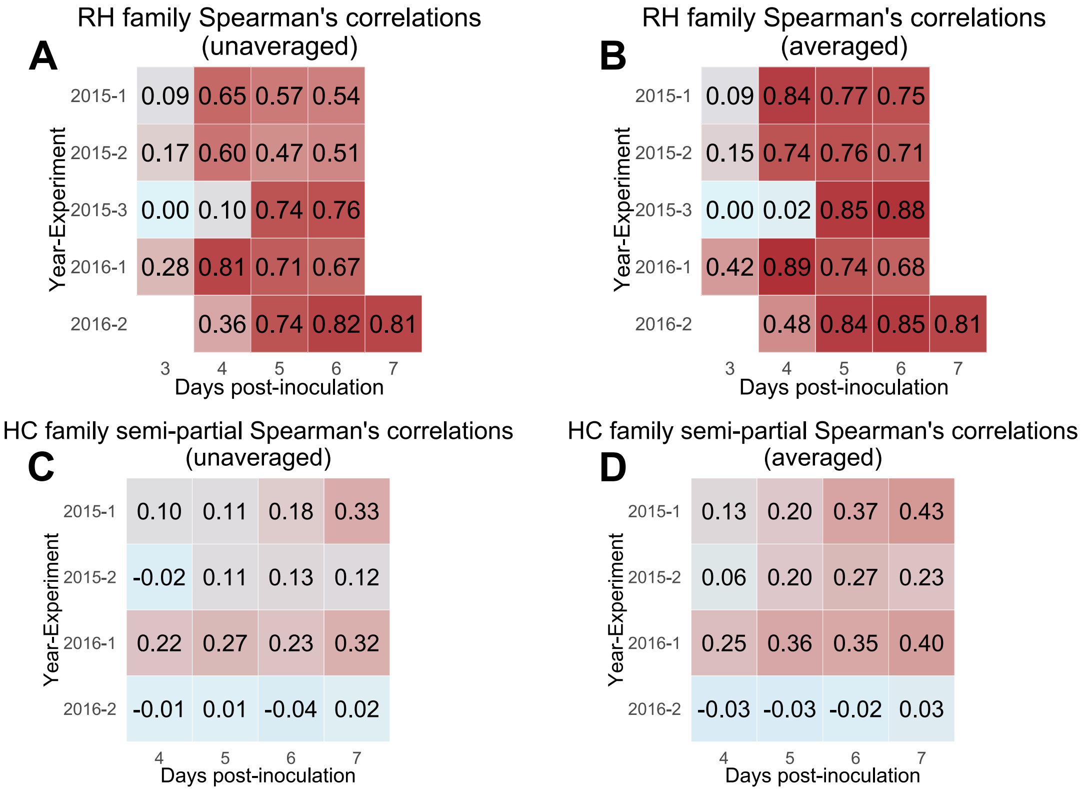

Figure 1.3: Correlations between computer vision and manual visual ratings at all experimental time points, shown as **A**, Spearman's rank correlation coefficients for the *Vitis rupestris* B38 × Horizon (RH) family unaveraged phenotypes; **B**, Spearman's rank correlation coefficients for the RH family averaged phenotypes; **C**, semipartial Spearman's rank correlation coefficients for the Horizon × *V. cinerea* B9 (HC) family unaveraged phenotypes; and **D**, semipartial Spearman's rank correlation coefficients for the HC family averaged phenotypes. The computer vision phenotype measures the percent area of sporulation by *Plasmopara viticola* whereas the manual phenotype measures the area of sporulation using a 1-to-5 ordinal scale.

2008). One approach is to use destructive assays, such as quantitative PCR analysis of *P. viticola* DNA concentration (Valsesia et al. 2005) or microscopy of stained samples to estimate the total amount of hyphae and spores (Boso and Kassemeyer 2008), or spore counts to quantify sporulation (Kono et al. 2015). Furthermore, one could concentrate on the hypersensitive response elicited by the leaf disc in response to *P. viticola* (Bellin et al. 2009). Alternatively, if one is interested in the area of sporulation, one can use either the computer vision or manual visual rating methods. The biochemical basis for all of these phenotypes is likely not the same, but correlations among them likely exist, as previously shown for manual visual ratings and spore counts (Kono et al. 2015). Thus, deciding what phenotype to use is at least partially subjective, but not trivial.

Here, we compared non-destructive methods to quantify sporulation using both a computer vision and a manual visual rating method, and showed computer vision can reduce the time to obtain the phenotypes by >90%. Additionally, the computer vision system allowed one the option to capture all the images first and then analyze them in bulk for all the experiments in a year, as was done for the experiments in this paper, if phenotypes are not needed immediately. The images also act as a form of documentation of an experiment that one could reexamine if necessary. Though not tested here, a camera other than the one on the iPhone 5s could be used provided it has enough resolution to capture sporulation, and phenotypes from two different cameras could be compared as the phenotypes are scaled by the leaf disc area. Another benefit of the computer vision rating system over the manual visual scale is that the phenotypes are quantitative. As seen in Figure 1.2, the computer vision ratings are less frequently skewed toward the most susceptible or resistant ratings compared to the manual visual ratings. While we cannot assume that the expected phenotype distribution should be normal, the extreme skew suggests that the manual rating scale does not work well in differentiating sporulation area when sporulation area is high or low.

While we do not have precise measurements of percent area of sporulation for comparison with our computer vision measurements, because we had multiple replications of the leaf discs, the law of large numbers (Bertsekas and Tsitsiklis 2002) tells us that we will approximately obtain the expected value of the manual or computer vision phenotype of a particular genotype. Because both the manual and computer vision ratings measured the same phenotype, although on different scales, given enough replication the average of the manual and computer vision ratings for a particular genotype should be highly correlated. In actuality, because we only had eight replications per genotype, we may not obtain a high correlation. However, the high correlation between the averaged manual and computer vision ratings in the RH family show that a computer vision system is a suitable replacement for the manual visual rating method. When trichomes are present, the computer vision ratings will approximate the expected value of the combined sporulation area and trichome phenotype, which is not the phenotype of interest. Additional testing of the computer vision system with genotypes with trichomes is needed to determine whether standardization of illumination can remove the influence of light on the calculated phenotype in order to increase the correlation between averaged manual and computer vision ratings when the effect of leaf trichomes is removed from each computer vision rating. Standardization of illumination may also produce more precise ratings for glabrous genotypes.

The current implementation of our computer vision system suffers from minor precision issues as seen in the violin plots for the second experiment for the RH family in 2015 at 3 and 4 dpi (Figure 1.2A). The plot for 4 dpi is shifted down toward lower values compared to the one at 3 dpi, though percent area of sporulation cannot decrease over time. Thus, the computer vision phenotypes between dpi within an experiment cannot be directly compared due to the presence of noise, which is normally distributed (data not shown), and different lighting conditions. However, manual visual ratings are not without precision issues. For example, in the case of phenotyping northern leaf blight on maize leaves, intra-rater precision, measured by Pearson's correlation coefficient, was on average 0.76 and 0.60 when using a quantitative or ordinal scale, respectively, while inter-rater correlations ranged from 0.65 to 0.93 and 0.58 to 0.82 when using a quantitative or ordinal scale, respectively (Poland and Nelson 2011). Similar precision issues were found in other pathosystems (Bock et al. 2008; Nita et al. 2003).

The leaf trichomes in the HC family presented a specific challenge, both for manual ratings and computer vision, but we still expected moderate correlation coefficients between the computer vision and manual visual ratings, which was not the case for the second experiment in 2016 even when phenotypes were averaged (Figure 1.3D). We found that if we used the 4 dpi computer vision phenotype as a covariate instead of the 2 dpi computer vision phenotype for the second experiment in 2016, the semi-partial Spearman correlations were 0.04, 0.08, and 0.18 between the manual and computer vision ratings at 5, 6, and 7 dpi for unaveraged phenotypes and 0.19, 0.26, and 0.29, respectively, for averaged phenotypes. The low correlation may have been due to the inability to obtain a sufficiently unbiased estimate of the leaf trichome phenotype for each day's unique lighting conditions. Yet, this is surprising given that we detected two QTL contributing to percent area of trichomes in the 2 dpi computer vision phenotype for the second experiment in 2016. Our results are consistent with a previous report on trichome density QTL in *Vitis* since each QTL's 95% approximate Bayes credible interval physically overlapped with previously reported QTL (Barba 2015). However, only 34.6% of the phenotypic variance was captured by the two QTL in the second experiment in 2016, leaving the possibility that variation other than that due to the leaf trichomes affected the 2 dpi computer vision phenotype.

In our experiments, we found that about 9% of the inoculation droplets did not stay on the leaf discs, and thus ratings could not be obtained. It is possible that spraying inoculum onto leaf discs, as done by Bellin et al. (2009), might be a better inoculation method. While our method had the advantage of placing roughly an equal amount of inoculum on each leaf disc, a comparison needs to be conducted to see if such precision has an effect on reducing the variance of ratings within genotypes, which is important if one has only a few leaf disc replicates per genotype. Additionally, our method was slow due to the need to individually pipette inoculum on each leaf disc, which does not scale well when a single person inoculates large numbers of leaf discs. The current state of the art in image recognition is a convolutional neural network (Krizhevsky et al. 2012). This technique requires large amounts of labeled training images and those images, as well as the test images, would have to be single leaf discs, which take longer to obtain than images of leaf discs in bulk. However, as shown by our work, a Hough circle transform algorithm would be an efficient way to isolate single leaf discs from an image. Our computer vision system, on the other hand, phenotypes the leaf discs in an unsupervised manner, so that previously phenotyped images are not needed in order to phenotype test images. Our computer vision system is also more suited to researchers having less technical knowledge in computer vision and machine learning, two topics that are seldom formally taught to plant pathologists and geneticists. We anticipate that with minor tweaking of parameters, most biologists could use this tool for pathosystems or traits where responses can be imaged with sufficient contrast from background.

1.6 Acknowledgments

We thank Tomàs Pallejà Cabrè for showing us the efficacy of using the blue layer for detection of disease symptoms and Konrad Wenzel for providing the source code for the Wallis filter. We also thank Mike Colizzi and Steve Luce for vineyard maintenance; Mike Colizzi, Corrigan Herbert, Aaron Green, Hema Kasinathan, Caitlyn Kumkey, Hanna Martens, Anne Repka, Michelle Schaub, and Mary Jean Welser for help in collecting leaves and plating leaf discs; and David Gadoury, Michelle Schaub, and Wayne Wilcox for supplying leaves for propagation of the *P. viticola* isolate. We thank Michael Gore for constructive feedback on the manuscript. This research was supported by the USDA-NIFA Specialty Crop Research Initiative (Award No. 2011-51181-30635), The New York Wine & Grape Foundation, Federal Capacity Funds and the Lake Erie Regional Grape Processor's Fund. Graduate Assistantship support for Konstantin Divilov was provided by the Charles R. Bullis Plant Hybridization Endowment and the Michael Nolan Endowment Fund.

1.7 References

Barba, P. (2015). "Genetic dissection of disease resistance and pest related traits in hybrid grapevine families". Ph.D. thesis. Cornell University.

Barbedo, J. G. A. (2013). "Digital image processing techniques for detecting, quantifying and classifying plant diseases". *SpringerPlus* 2 (660).

Bellin, D., E. Peressotti, D. Merdinoglu, S. Wiedemann-Merdinoglu, A.-F. Adam-Blondon, G. Cipriani, M. Morgante, R. Testolin, and G. Di Gaspero (2009). "Resistance to *Plasmopara viticola* in grapevine 'Bianca' is controlled by a major dominant gene causing localised necrosis at the infection site". *Theoretical and Applied Genetics* 120 (1), 163–176.

Bertsekas, D. P. and J. N. Tsitsiklis (2002). *Introduction to Probability.* 2nd ed. Belmont, MA: Athena Scientific.

Bock, C., P. Parker, A. Cook, and T. Gottwald (2008). “Visual rating and the use of image analysis for assessing different symptoms of citrus canker on grapefruit leaves”. *Plant Disease* 92 (4), 530–541.

Bock, C., G. Poole, P. Parker, and T. Gottwald (2010). “Plant disease severity estimated visually, by digital photography and image analysis, and by hyperspectral imaging”. *Critical Reviews in Plant Sciences* 29 (2), 59–107.

Boso, S. and H. Kassemeyer (2008). “Different suceptibility of European grapevine cultivars for downy mildew”. *Vitis* 47 (1), 39–49.

Bradski, G. (2000). “The OpenCV Library”. *Dr. Dobb's Journal of Software Tools* 25, 120–126.

Cadle-Davidson, L. (2008). “Variation within and between *Vitis* spp. for foliar resistance to the downy mildew pathogen *Plasmopara viticola*”. *Plant Disease* 92 (11), 1577–1584.

Cadle-Davidson, L., D. Gadoury, J. Fresnedo-Ramírez, S. Yang, P. Barba, Q. Sun, E. M. Demmings, R. Seem, M. Schaub, A. Nowogrodzki, et al. (2016). “Lessons from a phenotyping center revealed by the genome-guided mapping of powdery mildew resistance loci”. *Phytopathology* 106 (10), 1159–1169.

Eibach, R. (1998). “Investigations on the inheritance of resistance features to mildew diseases”. *VII International Symposium on Grapevine Genetics and Breeding*, 461–466.

Hyma, K. E., P. Barba, M. Wang, J. P. Londo, C. B. Acharya, S. E. Mitchell, Q. Sun, B. Reisch, and L. Cadle-Davidson (2015). “Heterozygous mapping strategy (HetMappS) for high resolution genotyping-by-sequencing markers: a case study in grapevine”. *PloS ONE* 10 (8), e0134880.

Jaillon, O., J.-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, et al. (2007). “The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla". *Nature* 449, 463–467.

Kennelly, M. M., D. M. Gadoury, W. F. Wilcox, P. A. Magarey, and R. C. Seem (2005). "Seasonal development of ontogenic resistance to downy mildew in grape berries and rachises". *Phytopathology* 95 (12), 1445–1452.

Khiook, I. L. K., C. Schneider, M.-C. Heloir, B. Bois, X. Daire, M. Adrian, and S. Trouvelot (2013). "Image analysis methods for assessment of H_2O_2 production and *Plasmopara viticola* development in grapevine leaves: application to the evaluation of resistance to downy mildew". *Journal of Microbiological Methods* 95 (2), 235–244.

Kim, S. (2015). "ppcor: an R package for a fast calculation to semi-partial correlation coefficients". *Communications for Statistical Applications and Methods* 22 (6), 665–674.

Kono, A., A. Sato, B. Reisch, and L. Cadle-Davidson (2015). "Effect of detergent on the quantification of grapevine downy mildew sporangia from leaf discs". *HortScience* 50 (5), 656–660.

Kortekamp, A. and E. Zyprian (1999). "Leaf hairs as a basic protective barrier against downy mildew of grape". *Journal of Phytopathology* 147 (7-8), 453–459.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks". *Advances in Neural Information Processing Systems*, 1097–1105.

Loughner, R., K. Goldman, G. Loeb, and J. Nyrop (2008). "Influence of leaf trichomes on predatory mite (*Typhlodromus pyri*) abundance in grape varieties". *Experimental and Applied Acarology* 45 (3), 111–122.

Luby, J. and P. Hemstad (2006). *Grape plant named 'Frontenac gris'.* US Patent PP16,478.

Ma, Z.-Y., J. Wen, S. M. Ickert-Bond, L.-Q. Chen, and X.-Q. Liu (2016). "Morphology, structure, and ontogeny of trichomes of the grape genus (*Vitis*, Vitaceae)". *Frontiers in Plant Science* 7.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes.* R package version 1.1-2. URL: `https://CRAN.R-project.org/package=RColorBrewer`.

Nita, M., M. Ellis, and L. Madden (2003). "Reliability and accuracy of visual estimation of Phomopsis leaf blight of strawberry". *Phytopathology* 93 (8), 995–1005.

Paolocci, M., M. Mugano, V. Alonso-Villaverde, and K. Gindro (2015). "Leaf morphological characteristics and stilbene production differently affect downy mildew resistance of *Vitis vinifera* varieties grown in Italy". *Vitis* 53 (3), 155–161.

Pauli, D., S. C. Chapman, R. Bart, C. N. Topp, C. Lawrence-Dill, J. Poland, and M. A. Gore (2016). "The quest for understanding phenotypic variation via integrated approaches in the field environment". *Plant Physiology* 172 (2), 622–634.

Peressotti, E., E. Duchêne, D. Merdinoglu, and P. Mestre (2011). "A semi-automatic non-destructive method to quantify grapevine downy mildew sporulation". *Journal of Microbiological Methods* 84 (2), 265–271.

Poland, J. A. and R. J. Nelson (2011). "In the eye of the beholder: the effect of rater variability and different rating scales on QTL mapping". *Phytopathology* 101 (2), 290–298.

Pratt, W. K. (2007). *Digital Image Processing.* 4th ed. Hoboken, NJ: John Wiley & Sons, Inc.

URGI (2014). *12X.2 version of the grapevine reference genome sequence from The French-Italian Public Consortium (PN40024).* URL: https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences.

Valsesia, G., D. Gobbin, A. Patocchi, A. Vecchione, I. Pertot, and C. Gessler (2005). “Development of a high-throughput method for quantification of *Plasmopara viticola* DNA in grapevine leaves by means of quantitative real-time polymerase chain reaction”. *Phytopathology* 95 (6), 672–678.

Wallis, R. (1976). “An approach to the space variant restoration and enhancement of images”. *Proceeding of the Symposium on Current Mathematical Problems in Image Science.* Naval Postgraduate School. Monterey, CA: Western Periodicals Co., 107–111.

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

CHAPTER 2

SINGLE AND MULTIPLE PHENOTYPE QTL ANALYSES OF DOWNY MILDEW RESISTANCE IN INTERSPECIFIC GRAPEVINES

2.1 Abstract

Breeding grapevines for downy mildew disease resistance has traditionally relied on qualitative gene resistance, which can be overcome by pathogen evolution. Analyzing two interspecific F_1 families, both having ancestry derived from *Vitis vinifera* and wild North American *Vitis* species, across two years and multiple experiments, we found multiple loci associated with downy mildew sporulation and hypersensitive response in both families using a single phenotype model, and no locus explained more than 17% of the variance for either phenotype. For two loci, we used RNA-Seq to detect differentially transcribed genes and found that the candidate genes at these loci were likely not NBS-LRR genes. Additionally, using a multiple phenotype Bayesian network analysis, we found effects between the leaf trichome density, hypersensitive response, and sporulation phenotypes. Moderate to high heritabilities were found for all three phenotypes, suggesting that selection for downy mildew resistance is an achievable goal by breeding for either physical or non-physical-based resistance mechanisms, with the combination of the two possibly providing durable resistance.

2.2 Introduction

Downy mildew resistance in grapevine has been mapped to over a dozen loci from over half a dozen *Vitis* spp. (Buonassisi et al. 2017). Most loci found explain the majority of the variation in the disease phenotype in a particular experiment, but some explain only a small portion of the variation (Bellin et al. 2009; Blasi et al. 2011; Moreira et al. 2011; Venuti et al. 2013). Resistance dependent on a single dominant locus is not seen as being durable, especially to an airborne outcrossing pathogen like the one that causes grapevine downy mildew (Buonassisi et al. 2017; McDonald and Linde 2002). By contrast, quantitative resistance is controlled by many genes, each of which contributes a small portion to the resistance phenotype (Poland et al. 2009). Additionally, quantitative resistance can be controlled by different metabolic processes in the plant, including nucleotide-binding site leucine-rich repeat (NBS-LRR) resistance genes. A pathogen evolving to overcome quantitative resistance would face a more difficult path to affect the same phenotype, e.g., necrosis or sporulation, as it would on a susceptible plant because more effector genes in the pathogen would need to mutate or experience recombination in order to overcome the multifaceted resistance in the plant. It is especially important to prevent pathogen evolution from overcoming disease resistance in grapevine because it is not economically feasible to replant a vineyard due to loss of disease resistance if it occurs a few years after planting, unlike maize or wheat where one can change the cultivar that is grown yearly.

Grapevine downy mildew, caused by the obligate biotrophic oomycete *Plasmopara viticola* (Berk. & M.A. Curtis) Berl. & de Toni, is a common cause of yield loss for *Vitis vinifera* cultivars, which generally lack genetic resistance to the disease (Buonassisi et al. 2017). Wild *Vitis* species, e.g., *V. amurensis*, *V.*

cinerea, *V. riparia*, *V. rupestris*, on the other hand, have genetic resistance to downy mildew (Cadle-Davidson 2008) and are commonly used as parents in resistance breeding efforts and in genetic studies. Infection on grapevine leaves can be detected by observing sporulation or host necrosis, which on resistant vines is often associated with a hypersensitive response (HR) (Bellin et al. 2009; Buonassisi et al. 2017). Sporulation can be quantified either manually using human vision or by using computer vision algorithms (Divilov et al. 2017). Research in identifying downy mildew resistance quantitative trait loci (QTL) in grapevine has focused on disease phenotypes, e.g., sporulation and HR, but physical barriers produced by a plant can also play a role in the prevention of disease. Kortekamp and Zyprian (1999) used four wild *Vitis* accessions to demonstrate that trichomes on the abaxial side of grapevine leaves can play a role in disease resistance to downy mildew by forming a hydrophobic surface above the leaf that blocks *P. viticola* sporangia from reaching stomata. As with sporulation, leaf trichome density can also be quantified using computer vision algorithms (Divilov et al. 2017).

In order to breed grapevines that have either qualitative or quantitative resistance to downy mildew using marker-assisted selection, one needs to find significant associations between downy mildew resistance ratings and genetic markers. However, it is rare to find causal genes with QTL mapping due to linkage disequilibrium present in the region a QTL resides. Therefore, one often knows the physical location of the causal locus only within a range of one to four megabases. RNA-Seq (Wang et al. 2009) is a high-throughput RNA sequencing method that can be used to identify candidate genes for QTL by finding differentially transcribed genes in the region where a QTL resides. Here, we describe newly-identified QTL associated with downy mildew resistance phenotypes, and the use of RNA-Seq to find candidate genes for two QTL on chromosome 14.

2.3 Materials and Methods

2.3.1 Plant Material and Phenotyping Methods

The plant material used consisted of two F_1 grapevine families, *V. rupestris* B38 × 'Horizon' (RH) and 'Horizon' × *V. cinerea* B9 (HC), grown unreplicated in a vineyard in Geneva, New York. 'Horizon' ancestry is derived from *V. vinifera* and North American *Vitis* spp. (Reisch et al. 1983). The progeny of the HC family segregate for trichomes on the abaxial side of their leaves, while leaves of the RH family are glabrous. Both families were phenotyped for sporulation area and percent sporulation area using a leaf disc assay with manual and computer vision methods, respectively, where the manual rating was on a 1 to 5 ordinal scale and the computer vision ratings ranged between 0 and 1 (Divilov et al. 2017). The HC family was also phenotyped for percent leaf trichome area using a leaf disc assay with the same computer vision method at 2 days post-inoculation (dpi) (Divilov et al. 2017). Divilov et al. (2017) provided a description of the computer vision system used as well as a detailed description of the experimental design. Hypersensitive response (HR) was assessed in both families at 2 dpi using a visual manual rating method where leaf discs were scored on a 1 to 5 ordinal scale (Figure 2.1). The RH family was phenotyped in 2015 with 163 F_1 genotypes and with three experiments, and in 2016 with 157 F_1 genotypes and with two experiments. The HC family was phenotyped in 2015 and 2016 with 152 and 145 F_1 genotypes, respectively, and with two experiments in each year. Susceptible and resistant controls were included to ensure inoculum quality. For the RH family, 'Cabernet Sauvignon' was the susceptible control for the first two experiments in 2015, while 'Chardonnay' was the susceptible control for all other experiments, and *V. rupestris* B38 was the resistant control. For the HC family, 'Chardonnay' was the susceptible control and *V. cinerea* B9 was the resistant control. Each experiment took place at a different date and consisted of four phenotypes taken on successive days starting at either 3 or 4 dpi for the sporulation phenotype. A phenotype within each dpi consisted of the average rating of eight leaf discs, two from each of four leaves obtained from different shoots on a vine. The RH family was phenotyped for HR in all experiments while the HC family was phenotyped for HR in all experiments except the first one in 2015.

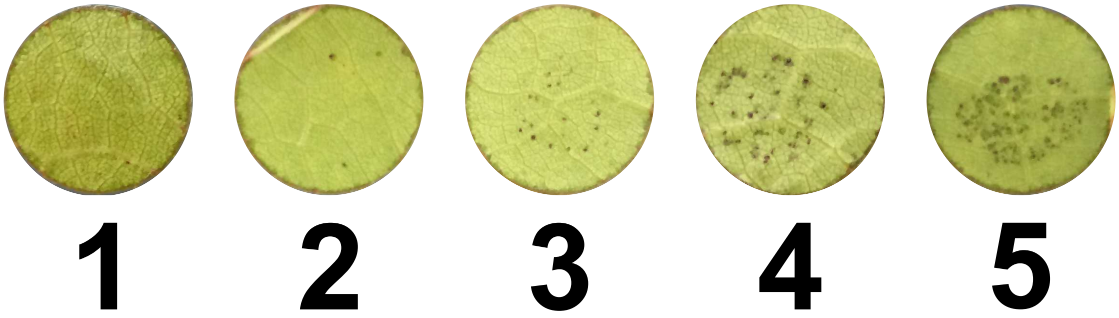

Figure 2.1: The ordinal visual scale for rating hypersensitive response on grapevine leaf discs.

2.3.2 Phenotype Modeling and Heritability Analysis

In order to analyze the 2015 and 2016 multi-experiment data for a phenotype, we fit a linear mixed model (McCulloch and Searle 2001) $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{Z}_{gy}\boldsymbol{u}_{gy} + \boldsymbol{\varepsilon}$ for each family individually. $\boldsymbol{X}$ was a design matrix and $\boldsymbol{b}$ was a vector of fixed effects. For the HR phenotype, $\boldsymbol{X}$ contained indicator variables for the intercept and experiments within years. For the manual and computer vision sporulation phenotypes, additional covariates of dpi within experiments within years were included. For the computer vision sporulation phenotype for the HC family, additional covariates of leaf trichomes within experiments within years were included. $\boldsymbol{Z}_g$ and $\boldsymbol{Z}_{gy}$ were incidence matrices relating genotypes to phenotypes and genotypes within years to phenotypes, respectively. $\boldsymbol{u}_g \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}\sigma^2_g)$ was a vector of normally distributed genetic effects, also known as estimated breeding values, corresponding to each genotype, where $\boldsymbol{I}$ is an identity matrix. We used an identity matrix as our genotype covariance matrix because the F_1 genotypes within each family had no population structure. $\boldsymbol{u}_{gy} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{E}\otimes\boldsymbol{I}\sigma^2_{gy})$ was a vector of normally distributed genotype-by-year interaction effects corresponding to each genotype in a year, where $\otimes$ is the Kronecker product and $\boldsymbol{E}$ is an identity matrix with the number of rows and columns equal to the number of years. The genotype-by-year interaction covariance matrix assumes that genotypic performance in 2015 was independent of genotypic performance in 2016. $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}\sigma^2_{\varepsilon})$ was a vector holding the normally distributed independent noise, or error, of the phenotypes. From this model, the broad-sense heritability was estimated as $\frac{\sigma^2_g}{\sigma^2_g + \frac{\sigma^2_{gy}}{y} + \frac{\sigma^2_{\varepsilon}}{n}}$ where σ^2_g, σ^2_{gy}, σ^2_{ε} were the genetic, genotype-by-year interaction, and error variances, respectively, and y and n were the number of years and number of experiments over the years, respectively. In addition to the sporulation and HR phenotypes, we also calculated the heritability of the leaf trichome phenotype. In that case, the only fixed effects in the linear mixed model were the same as that for HR. The models were fit using the EMMREML R package (Akdemir and Godfrey 2015) with the average information algorithm used to obtain estimates of variance components.

2.3.3 Single Phenotype QTL Analysis

In order to find QTL for the breeding values of a trait, a Haley-Knott linear regression model was built using forward and backward stepwise selection with R/qtl

(Broman et al. 2003). Interaction effects were not considered for inclusion in the model. The logarithm of odds (LOD) penalty for each trait was determined by 1,000 permutation tests with an alpha value of 0.05. Approximate Bayes credible intervals were calculated for QTL and represented the region in which a QTL resides with probability $\geq$0.95. The genotyping error rate used to calculate conditional genotype probabilities was set to 0.001. RH and HC family genetic maps used for the analyses were made with HetMappS and were previously published (Hyma et al. 2015). Imputation of missing single nucleotide polymorphism (SNP) data was performed using the expectation-maximization algorithm in rrBLUP (Endelman 2011). Physical locations of the SNPs in these maps were obtained using the 12X.2 version (URGI 2014) of the grapevine reference genome PN40024 (Jaillon et al. 2007). For comparison purposes, we performed the same stepwise regression analysis on manual and computer vision sporulation and HR phenotypes within years within experiments within dpi to observe what QTL were found using these data.

2.3.4 Multiple Phenotype Bayesian Network Analysis

In order to detect effects between multiple phenotypes, as well as the effect of SNPs on phenotypes, we constructed an averaged Bayesian network for each family using the bnlearn R package (Scutari 2010). A similar analysis has been previously done in wheat (Scutari et al. 2014). For the RH family, we constructed a Bayesian network using the manual sporulation and HR breeding values, as well as the SNPs in that family. For the HC family, the leaf trichome breeding values were included as well. We restricted the manual sporulation trait from affecting the HR and leaf trichome traits, and we restricted the HR trait from affecting the leaf trichome trait. This was done because necrosis is present on the leaf prior to the appearance of sporulation, and leaf trichomes are present on the leaf prior to inoculation. All traits were restricted from affecting the SNPs. We used the SI-HITON-PC algorithm (Aliferis et al. 2010) to find the Markov blanket of each trait individually. The Markov blanket (Pearl 1988) represents the set of traits and SNPs such that a given trait, conditional on its Markov blanket, is independent of all other traits and SNPs. Independence was determined by non-significance of the Pearson's correlation coefficient between a variable conditional on its Markov blanket and a variable outside the Markov blanket tested using a Student's t-test with an alpha value of 0.01. The hill-climbing algorithm (Scutari 2010) was then used to find the structure of the Bayesian network containing the traits and their Markov blankets. In the network, the distribution of a variable, or node, of interest conditional on its parents, i.e., nodes with arrows pointing to the node of interest, is parameterized as a linear regression model. The hill-climbing algorithm greedily adds arrows between nodes such that the total Bayesian information criterion (BIC), which is a function of the log likelihood of a linear regression model and a parameter penalization term, among the linear regression models achieves its highest value (the BIC is rescaled by -2 in bnlearn). The total BIC is the sum of the BICs of the individual linear regression models. We allowed the SNPs within all Markov blankets to affect any trait to account for possible pleiotropy. We constructed 1000 networks using a random set of 90% of the individuals in a family for each network and then created an averaged network where we kept the structural components, i.e., the arrows, present in at least half of the networks. QTL bootstrapped confidence intervals of the averaged network SNPs were obtained by calculating the range of physical locations of SNPs on the same chromosome as the averaged network SNPs found to affect the same trait in at least 5% of the networks. The networks were drawn using BayesNet (Luttinen 2013).

2.3.5 RNA-Seq Experimental Design and Analysis

On 11 July 2016 and 19 July 2016, two leaf disc assay experiments were conducted following the same phenotyping methodology (Divilov et al. 2017) used in the experiments described above. For RNA-Seq analyses, 28 RH genotypes were chosen that segregated for two QTL on chromosome 14, one on each of the RH parental maps found using the stepwise regression approach explained above with the computer vision sporulation trait. Specifically, 15 of the 28 genotypes were heterozygous and homozygous for the most significant marker within the QTL from *V. rupestris* B38 and 'Horizon', respectively, while the other 13 genotypes were homozygous and heterozygous for the most significant marker within the QTL from *V. rupestris* B38 and 'Horizon', respectively. Because markers exist in very close linkage to the most significant markers in both QTL credible intervals that are of opposite phase, the phase information of the most significant marker in both QTL credible intervals is not informative. Positive and negative control genotypes were included to ensure inoculum quality. At 7 hours post-inoculation, single leaf discs from each of the 28 genotypes were frozen in liquid nitrogen and stored at -80℃ prior to RNA extraction using the Spectrum^{TM} Plant Total RNA Kit (Sigma-Aldrich). Leaf discs of each genotype from the two experiments were combined prior to RNA extraction. Libraries were made using the protocol of Zhong et al. (2011) and sequenced using the Illumina NextSeq500 to obtain single-end 75 bp reads. Reads were aligned to the 12X.2 version of the grapevine reference genome PN40024 using HISAT2 (Kim et al. 2015) and the transcriptome was assembled using StringTie (Pertea et al. 2015) with the CRIBI functional annotation (Vitulo et al. 2014). Transcribed genes were analyzed by fitting a linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon}$ using Ballgown (Frazee et al. 2015) where $\boldsymbol{y}$ was a vector holding the $\log_2$(FPKM+1) values of genes with FPKM (Fragments Per Kilobase of transcript per Million reads sequenced) variances greater than one for each genotype; $\boldsymbol{X}$ was a design matrix that contained the indicator variables for the intercept and the genotype-phase grouping; $\boldsymbol{b}$ was a vector holding the mean and the effect of the group value; and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma_{\varepsilon}^2)$ was a vector holding the error. For each gene, the full model was compared to a model without the grouping covariate to derive an F statistic. The significance threshold for the F statistic was set to a q value of 0.05. We called genes that passed the significance threshold differentially transcribed genes. Only those physically located within the 95% approximate Bayes credible intervals of the two QTL on chromosome 14 from *V. rupestris* B38 and ‘Horizon’ were considered as possible candidate genes. The UniProt database (The UniProt Consortium 2017) was used to determine the names and GO terms of candidate genes.

2.3.6 Data Availability

The phenotypic and genetic data, as well as the code used to run the linear mixed models and single time point, single phenotype, and multiple phenotype Bayesian network analyses, are available at https://github.com/kdivilov. The RNA-Seq analysis pipeline is included in the repository as well.

2.4 Results

The manual and computer vision sporulation phenotype distributions for both families were previously published (Divilov et al. 2017). The RH family HR phenotypes and the HC family leaf trichome phenotypes were approximately normally distributed (Figure 2.2). The HC family HR phenotype distributions were approximately truncated normal distributions. The heritabilities for the RH family manual and computer vision sporulation and HR phenotypes were 0.40, 0.43, and 0.58, respectively. The heritabilities for the HC family manual and computer vision sporulation and HR phenotypes were 0.67, 0.21, and 0.73, respectively. The heritability of leaf trichomes in the HC family was 0.83. Single phenotype and multiple phenotype Bayesian network analyses in total identified ten significant QTL for these traits on chromosomes 5, 6, 7, 8, 11, 14 (two QTL), 15, 16, and 18, described below.

Single phenotype QTL from the RH family explained between 7% and 17% of the variation in the traits examined. Three QTL, one on chromosome 11 from 'Horizon' and two on chromosome 14 from *V. rupestris* B38 and 'Horizon' were found using the computer vision sporulation breeding values (Table 2.1). The same QTL with overlapping physical locations were found using the manual sporulation breeding values in addition to one on chromosome 18 from 'Horizon'. Two QTL, one on chromosome 8 from 'Horizon' and one on chromosome 11 from *V. rupestris* B38 and 'Horizon' were found using the HR breeding values. The HR QTL on chromosome 11 co-located with the sporulation QTL on chromosome 11. Single phenotype QTL from the HC family explained between 8% and 15% of the variation in the traits examined. Three QTL on chromosomes 5, 7, and 8 from 'Horizon' were found using the HC family manual sporulation breeding values, but no QTL

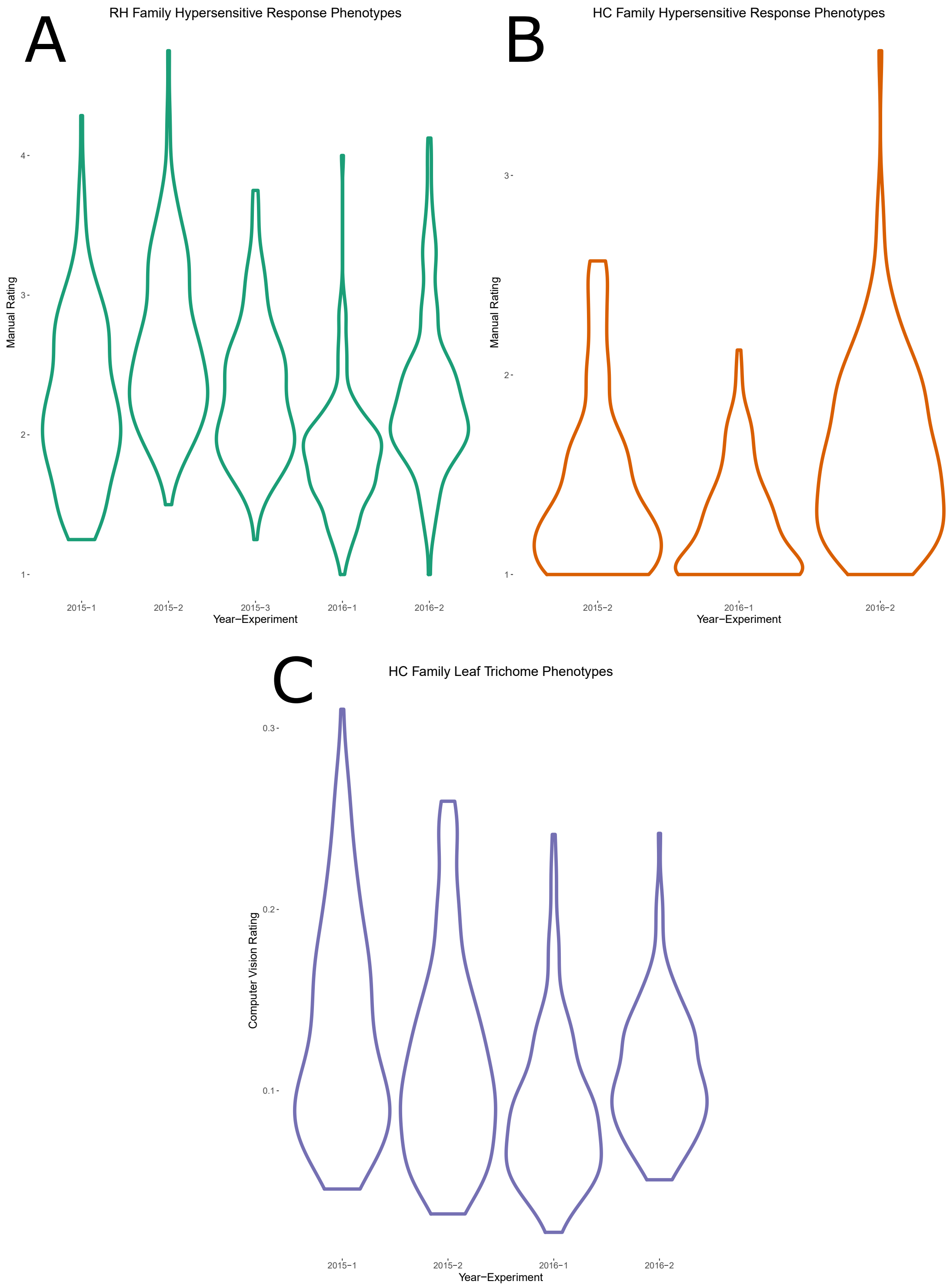

Figure 2.2: Violin plots showing the **A** *Vitis rupestris* B38 × Horizon (RH) and **B** Horizon × *V. cinerea* B9 (HC) hypersensitive response (HR) phenotype distributions and the **C** HC leaf trichome phenotype distribution across experiments within years. Each phenotype is represented as the average of eight leaf disc ratings.

were found using the computer vision sporulation breeding values. Three HR QTL from 'Horizon' were identified, one each on chromosomes 5 and 8 that co-located with those found with the HC manual sporulation breeding values, and one on chromosome 6. The QTL found on chromosome 8 from the HC family manual sporulation and HR breeding values did not co-locate with the one related to HR in the RH family.

Comparing QTL found using the breeding values derived from the linear mixed model that utilized all the data available to those found using phenotypes from a single dpi within an experiment within a year, fourteen and eight QTL were found with the latter data for RH and HC families, respectively (Table 2.2). Among those, seven and four of the QTL in their respective families were found only once. All five and four QTL found in the RH and HC families, respectively, using the linear mixed model were also found using the individual time point analysis. No QTL that was found only once using the individual time point analysis was found using the linear mixed model approach. For the RH family, the mean and median approximate Bayes credible intervals of QTL obtained using the breeding values were 5.7 Mbp and 2.7 Mbp wide, while those obtained using the individual phenotypes were 6.3 Mbp and 4.5 Mbp wide, respectively. For the HC family, the mean and median approximate Bayes credible intervals of QTL obtained using the breeding values were 5.6 Mbp and 4.7 Mbp, while those obtained using the individual phenotypes were 7.9 Mbp and 6.3 Mbp, respectively.

The averaged Bayesian network for the RH family showed no effect on sporulation by HR (Figure 2.3). A QTL was found to affect sporulation on chromosome 14 from *V. rupestris* B38 that co-located with the one from the single phenotype analysis (Table 2.3). Three QTL were found to affect HR on chromosomes 8 and

Table 2.1: Statistics for the QTL found from the single phenotype stepwise regression analysis performed on breeding values of the *V. rupestris* B38 × 'Horizon' (RH) and 'Horizon' × *V. cinerea* B9 (HC) F_1 families incorporating data across two years and multiple experiments.

Name	Family	Chr	Heterozygous Parent[a]	Phenotype	LOD Threshold[b]	LOD Score[c]	% Var. Explained[d]	Effect Size[e]	95% Credible Interval (Mbp)[f]
Rpv17	RH	8	Horizon	Hypersensitive response (HR)	3.60	6.70	12.94	0.149	11.369-11.721-12.184
Rpv18	RH	11	Horizon	Sporulation (manual)	3.59	4.69	8.77	0.159	7.038-8.138-19.921
Rpv18	RH	11	Horizon	Sporulation (computer vision)	3.65	4.21	8.51	0.006	7.038-16.995-19.921
	RH	11	*V. rupestris* B38	Hypersensitive response (HR)	3.60	4.11	7.66	0.117	7.609-17.753-19.857
Rpv18	RH	11	Horizon	Hypersensitive response (HR)	3.60	8.71	17.33	0.169	15.397-15.397-16.994
	RH	14	Horizon	Sporulation (manual)	3.59	7.75	15.14	0.208	23.275-24.823-25.002
	RH	14	Horizon	Sporulation (computer vision)	3.65	5.76	11.91	0.007	24.366-25.002-25.778
Rpv19	RH	14	*V. rupestris* B38	Sporulation (manual)	3.59	6.19	11.83	0.184	27.119-29.790-29.790
Rpv19	RH	14	*V. rupestris* B38	Sporulation (computer vision)	3.65	7.34	15.51	0.008	27.085-29.543-29.790
	RH	18	Horizon	Sporulation (manual)	3.59	3.79	6.99	0.146	6.598-9.684-14.528
	HC	5	Horizon	Sporulation (manual)	3.43	5.58	11.27	0.314	0.844-3.112-5.511
	HC	5	Horizon	Hypersensitive response (HR)	3.44	6.91	15.10	0.177	0.114-2.388-3.642
Rpv20	HC	6	Horizon	Hypersensitive response (HR)	3.44	4.00	8.37	0.142	0.770-0.907-13.625
Rpv21	HC	7	Horizon	Sporulation (manual)	3.43	5.42	10.90	0.312	0.994-2.129-3.150
	HC	8	Horizon	Sporulation (manual)	3.43	5.81	11.77	0.323	16.814-19.217-22.458
	HC	8	Horizon	Hypersensitive response (HR)	3.44	4.86	10.29	0.147	17.766-19.609-22.458

[a] The heterozygous parent is the one that has the heterozygous allele. SNPs for the genotypes in the families either were homozygous for one allele or heterozygous because only pseudo-testcross markers were used to build the genetic maps (Hyma et al. 2015).

[b] Calculated using 1,000 permutation tests with an alpha value of 0.05.

[c] Given for the most significant marker in a QTL credible interval.

[d] Calculated as $\frac{\text{Type III SS}}{\text{Total SS}} \times 100$ for the most significant marker in a QTL credible interval.

[e] The absolute value of the effect of the most significant marker in a QTL credible interval when the other QTL listed for a particular phenotype within a family are included in the model. The absolute value is given because phase information of the most significant marker is not informative due to linkage with markers of opposite phase.

[f] Location intervals are based on the 12X.2 version (URGI 2014) of the grapevine reference genome. The middle value represents the location of the most significant marker.

Table 2.2: Statistics for the QTL found using stepwise regression performed on the single time point phenotypes of the *V. rupestris* B38 × 'Horizon' (RH) F_1 family incorporating data across two years and multiple experiments.

Year	Exp	Dpi	Chr	Heterozygous Parent[a]	Phenotype	LOD Threshold[b]	LOD Score[c]	% Var. Explained[d]	Effect Size[e]	95% Credible Interval (Mbp)[f]
2016	1	3	2	Horizon	Sporulation (computer vision)	3.48	4.13	11.41	0.0052	0.833-1.327-2.711
2015	2	2	5	Horizon	Hypersensitive response (HR)	3.53	7.61	13.59	0.4445	2.437-2.852-5.708
2015	1	5	7	*V. rupestris* B38	Sporulation (computer vision)	3.56	5.23	7.01	0.0155	22.112-22.750-23.228
2016	2	6	7	Horizon	Sporulation (computer vision)	3.51	3.91	10.91	0.0319	2.809-5.357-15.270
2015	1	2	8	*V. rupestris* B38	Hypersensitive response (HR)	3.53	4.45	9.30	0.3881	9.153-10.905-14.130
2015	1	2	8	Horizon	Hypersensitive response (HR)	3.53	4.95	10.40	0.3905	9.060-11.721-13.507
2015	1	3	8	Horizon	Sporulation (manual)	2.90	4.70	9.68	0.1989	16.823-19.001-21.130
2015	1	4	8	Horizon	Sporulation (computer vision)	3.56	4.41	7.32	0.0094	4.298-17.589-19.192
2015	1	4	8	Horizon	Sporulation (manual)	3.50	5.94	10.83	0.7422	15.927-17.589-20.865
2015	1	5	8	Horizon	Sporulation (computer vision)	3.56	6.76	9.28	0.0173	4.298-16.390-18.320
2015	2	2	8	Horizon	Hypersensitive response (HR)	3.53	5.71	9.93	0.3667	7.219-8.077-12.184
2015	3	2	8	Horizon	Hypersensitive response (HR)	3.56	6.94	16.18	0.3958	9.335-11.656-18.216
2016	2	6	9	Horizon	Sporulation (manual)	3.68	3.77	10.54	0.6191	16.428-21.899-22.356
2015	1	5	10	Horizon	Sporulation (computer vision)	3.56	5.17	6.92	0.0147	1.990-4.249-14.698
2015	1	6	10	Horizon	Sporulation (computer vision)	3.67	4.48	8.02	0.0176	3.606-11.379-18.923
2015	1	2	11	Horizon	Hypersensitive response (HR)	3.53	6.57	14.14	0.4468	16.995-16.995-17.753
2015	2	2	11	*V. rupestris* B38	Hypersensitive response (HR)	3.53	5.08	8.75	0.3549	8.203-17.389-18.918
2015	2	2	11	Horizon	Hypersensitive response (HR)	3.53	8.00	14.37	0.4350	8.288-15.397-16.994
2015	2	5	11	Horizon	Sporulation (computer vision)	3.55	4.00	9.66	0.0117	5.898-8.155-17.753
2015	3	2	11	Horizon	Hypersensitive response (HR)	3.56	4.38	9.84	0.3027	11.179-15.397-17.753
2015	1	3	12	*V. rupestris* B38	Sporulation (manual)	2.90	3.59	7.28	0.1731	0.735-0.832-2.469
2015	1	4	12	*V. rupestris* B38	Sporulation (computer vision)	3.56	7.60	13.23	0.0121	0.832-2.469-5.321
2015	1	4	12	*V. rupestris* B38	Sporulation (manual)	3.50	4.93	8.85	0.6534	0.832-3.110-3.372
2015	1	5	12	*V. rupestris* B38	Sporulation (computer vision)	3.56	6.08	8.26	0.0161	0.735-3.050-5.321
2015	1	6	12	*V. rupestris* B38	Sporulation (computer vision)	3.67	5.29	9.59	0.0186	1.504-3.953-6.824
2015	3	6	13	Horizon	Sporulation (manual)	3.65	3.98	6.85	0.4962	12.412-22.830-28.836
2015	1	4	14	*V. rupestris* B38	Sporulation (computer vision)	3.56	4.16	6.88	0.0088	25.272-29.543-29.790
2015	1	4	14	Horizon	Sporulation (computer vision)	3.56	4.39	7.28	0.0098	22.173-24.981-27.565
2015	1	4	14	*V. rupestris* B38	Sporulation (manual)	3.50	4.26	7.58	0.6085	7.966-29.543-29.790
2015	1	4	14	Horizon	Sporulation (manual)	3.50	7.54	14.07	0.8078	21.233-21.439-23.867
2015	1	5	14	*V. rupestris* B38	Sporulation (computer vision)	3.56	4.04	5.33	0.0131	25.272-27.478-29.790
2015	1	5	14	Horizon	Sporulation (computer vision)	3.56	7.68	10.68	0.0190	27.369-27.369-27.698
2015	1	6	14	Horizon	Sporulation (computer vision)	3.67	4.74	8.54	0.0180	22.488-27.474-28.791
2015	2	4	14	Horizon	Sporulation (manual)	3.60	4.79	12.74	0.8148	4.090-17.855-26.072
2015	2	5	14	*V. rupestris* B38	Sporulation (computer vision)	3.55	4.01	9.70	0.0117	5.774-25.535-29.790
2015	2	6	14	*V. rupestris* B38	Sporulation (computer vision)	3.70	6.50	13.90	0.0147	3.395-3.395-28.215
2015	2	6	14	Horizon	Sporulation (computer vision)	3.70	4.37	9.06	0.0117	23.275-25.486-29.407
2015	3	5	14	*V. rupestris* B38	Sporulation (computer vision)	3.61	4.67	9.78	0.0163	26.362-27.478-29.790
2015	3	5	14	Horizon	Sporulation (computer vision)	3.61	4.30	8.96	0.0164	24.934-25.002-26.764
2015	3	5	14	*V. rupestris* B38	Sporulation (manual)	3.72	4.80	10.48	0.7044	25.272-27.085-29.759
2015	3	6	14	*V. rupestris* B38	Sporulation (computer vision)	3.53	5.21	10.32	0.0242	26.807-29.543-29.790
2015	3	6	14	Horizon	Sporulation (computer vision)	3.53	5.82	11.64	0.0254	24.934-25.069-26.072
2015	3	6	14	*V. rupestris* B38	Sporulation (manual)	3.65	6.08	10.79	0.6236	26.807-27.118-29.759
2015	3	6	14	Horizon	Sporulation (manual)	3.65	4.52	7.85	0.5873	23.275-24.981-26.235
2016	1	4	14	*V. rupestris* B38	Sporulation (computer vision)	3.68	6.23	16.69	0.0281	27.085-28.604-29.759
2016	1	4	14	*V. rupestris* B38	Sporulation (manual)	3.72	4.23	11.67	0.8124	25.628-28.604-29.790
2016	1	5	14	*V. rupestris* B38	Sporulation (computer vision)	3.59	6.23	16.69	0.0281	26.362-28.604-29.790
2016	1	6	14	*V. rupestris* B38	Sporulation (computer vision)	3.54	5.59	14.12	0.0397	26.362-28.604-29.543

Table 2.2 (continued): Statistics for the QTL found using stepwise regression performed on the single time point phenotypes of the *V. rupestris* B38 × 'Horizon' (RH) F_1 family incorporating data across two years and multiple experiments.

Year	Exp	Dpi	Chr	Heterozygous Parent[a]	Phenotype	LOD Threshold[b]	LOD Score[c]	% Var. Explained[d]	Effect Size[e]	95% Credible Interval (Mbp)[f]
2015	1	4	16	*V. rupestris* B38	Sporulation (computer vision)	3.56	4.36	7.24	0.0091	16.131-19.348-22.753
2015	1	5	16	*V. rupestris* B38	Sporulation (computer vision)	3.56	5.91	8.00	0.0161	19.080-19.348-19.531
2015	1	6	16	*V. rupestris* B38	Sporulation (computer vision)	3.67	4.92	8.88	0.0183	19.094-19.923-20.009
2015	2	3	16	*V. rupestris* B38	Sporulation (manual)	3.27	3.31	8.97	0.2169	13.270-20.649-22.369
2015	2	5	16	*V. rupestris* B38	Sporulation (manual)	3.62	4.36	11.66	0.5242	19.080-20.649-23.306
2015	2	6	16	*V. rupestris* B38	Sporulation (computer vision)	3.70	3.97	8.18	0.0114	17.699-20.649-23.165
2015	2	6	16	*V. rupestris* B38	Sporulation (manual)	3.53	5.13	13.57	0.5274	17.497-20.649-22.145
2016	1	6	16	*V. rupestris* B38	Sporulation (computer vision)	3.54	4.06	10.02	0.0346	19.746-21.309-22.369
2015	2	2	17	Horizon	Hypersensitive response (HR)	3.53	3.79	6.40	0.3288	17.294-17.948-18.085
2015	1	3	18	*V. rupestris* B38	Sporulation (manual)	2.90	7.05	15.03	0.9274	6.314-7.537-7.952
2015	1	3	18	*V. rupestris* B38	Sporulation (manual)	2.90	5.96	12.51	0.8374	7.952-7.985-8.676
2015	3	5	18	Horizon	Sporulation (computer vision)	3.61	5.99	12.80	0.0185	3.249-6.598-10.064
2015	3	5	18	Horizon	Sporulation (manual)	3.72	7.22	16.33	0.8370	3.917-6.598-8.633
2015	3	6	18	Horizon	Sporulation (computer vision)	3.53	7.15	14.58	0.0284	4.382-6.598-8.577
2015	3	6	18	Horizon	Sporulation (manual)	3.65	9.51	17.74	0.7913	4.382-6.598-8.633
2016	1	3	18	Horizon	Sporulation (manual)	3.47	4.05	11.19	0.3412	4.894-8.736-10.296

[a] The heterozygous parent is the one that has the heterozygous allele. SNPs for the genotypes in the families either were homozygous for one allele or heterozygous because only pseudo-testcross markers were used to build the genetic maps (Hyma et al. 2015).

[b] Calculated using 1,000 permutation tests with an alpha value of 0.05.

[c] Given for the most significant marker in a QTL credible interval.

[d] Calculated as $\frac{\text{Type III SS}}{\text{Total SS}} \times 100$ for the most significant marker in a QTL credible interval.

[e] The absolute value of the effect of the most significant marker in a QTL credible interval when the other QTL listed for a particular phenotype within a family are included in the model. The absolute value is given because phase information of the most significant marker is not informative due to linkage with markers of opposite phase.

[f] Location intervals are based on the 12X.2 version (URGI 2014) of the grapevine reference genome. The middle value represents the location of the most significant marker.

Table 2.3: Statistics for the QTL found using stepwise regression performed on single time point phenotypes of the 'Horizon' × *V. cinerea* B9 (HC) F_1 family incorporating data across two years and multiple experiments.

Year	Exp	Dpi	Chr	Heterozygous Parent[a]	Phenotype	LOD Threshold[b]	LOD Score[c]	% Var. Explained[d]	Effect Size[e]	95% Credible Interval (Mbp)[f]
2015	1	4	5	Horizon	Sporulation (manual)	3.60	3.99	9.92	0.4527	0.316-1.061-6.609
2015	1	5	5	Horizon	Sporulation (manual)	3.62	4.85	11.28	0.7258	0.316-1.061-6.609
2015	1	7	5	Horizon	Sporulation (manual)	3.53	3.86	9.66	0.7545	0.316-1.061-6.661
2015	2	2	5	Horizon	Hypersensitive response (HR)	3.56	3.59	9.18	0.2541	0.114-2.910-9.112
2016	1	2	5	Horizon	Hypersensitive response (HR)	3.46	4.75	12.45	0.1869	0.114-3.155-6.661
2016	1	4	5	Horizon	Sporulation (manual)	3.24	4.23	10.04	0.3990	0.316-0.601-10.450
2016	1	5	5	Horizon	Sporulation (manual)	3.44	3.60	10.81	0.5121	0.114-0.924-16.347
2016	1	6	5	Horizon	Sporulation (manual)	3.41	4.16	12.38	0.6232	0.114-2.388-16.347
2016	1	7	5	Horizon	Sporulation (manual)	3.55	4.05	12.08	0.6216	0.114-2.388-16.347
2016	2	2	5	Horizon	Hypersensitive response (HR)	3.47	5.18	13.49	0.3237	0.114-2.388-6.912
2016	2	4	5	Horizon	Sporulation (manual)	3.24	3.66	10.97	0.3237	0.316-0.924-20.469
2016	2	5	5	Horizon	Sporulation (manual)	3.39	5.16	15.12	0.5211	0.844-0.924-16.347
2016	2	6	5	Horizon	Sporulation (manual)	3.39	5.51	14.17	0.6132	0.844-0.924-15.445
2016	2	7	5	Horizon	Sporulation (manual)	3.43	7.05	14.85	0.6497	0.844-0.924-5.446
2015	2	2	6	Horizon	Hypersensitive response (HR)	3.56	4.34	11.23	0.2841	0.770-6.848-14.939
2016	1	2	6	Horizon	Hypersensitive response (HR)	3.46	4.71	12.36	0.1978	0.802-0.907-9.327
2015	1	6	7	Horizon	Sporulation (manual)	3.54	4.64	10.06	0.7830	0.994-2.129-3.150
2015	1	7	7	Horizon	Sporulation (computer vision)	3.45	5.05	9.12	0.0113	23.341-27.311-27.340
2015	2	5	7	Horizon	Sporulation (manual)	3.50	5.28	14.79	0.5529	0.994-1.537-5.334
2015	2	6	7	Horizon	Sporulation (manual)	3.56	5.81	16.13	0.6957	2.610-3.494-3.509
2015	2	7	7	Horizon	Sporulation (manual)	3.56	6.06	16.78	0.7171	2.486-3.494-3.509
2016	2	6	7	Horizon	Sporulation (manual)	3.39	3.73	9.34	0.4955	2.566-2.610-16.988
2016	2	7	7	Horizon	Sporulation (manual)	3.43	4.65	9.41	0.5493	0.994-2.486-4.237
2015	1	4	8	Horizon	Sporulation (manual)	3.60	5.46	13.88	0.5272	16.814-19.609-22.458
2015	1	5	8	Horizon	Sporulation (manual)	3.62	7.52	18.23	0.9091	17.142-17.766-21.588
2015	1	6	8	Horizon	Sporulation (manual)	3.54	5.95	13.16	0.9019	16.814-18.737-20.067
2015	1	7	8	Horizon	Sporulation (manual)	3.53	5.14	13.12	0.8663	16.814-17.766-20.865
2016	1	4	8	Horizon	Sporulation (manual)	3.24	4.05	9.58	0.7880	15.579-15.804-16.155
2016	1	4	8	Horizon	Sporulation (manual)	3.24	6.79	16.82	1.0166	16.542-17.766-19.486
2016	2	2	8	Horizon	Hypersensitive response (HR)	3.47	4.23	10.86	0.2901	17.766-19.751-21.954
2016	2	7	8	Horizon	Sporulation (manual)	3.43	3.92	7.84	0.4723	16.542-20.865-22.458
2015	1	6	15	*V. cinerea* B9	Sporulation (manual)	3.54	5.37	11.77	0.9112	15.905-16.962-19.560
2016	2	7	16	*V. cinerea* B9	Sporulation (manual)	3.43	4.23	8.52	0.4986	0.507-2.800-8.056
2015	2	7	18	Horizon	Sporulation (computer vision)	3.33	3.50	1.93	0.0095	3.950-9.684-23.781

[a] The heterozygous parent is the one that has the heterozygous allele. SNPs for the genotypes in the families either were homozygous for one allele or heterozygous because only pseudo-testcross markers were used to build the genetic maps (Hyma et al. 2015).

[b] Calculated using 1,000 permutation tests with an alpha value of 0.05.

[c] Given for the most significant marker in a QTL credible interval.

[d] Calculated as $\frac{\text{Type III SS}}{\text{Total SS}} \times 100$ for the most significant marker in a QTL credible interval.

[e] The absolute value of the effect of the most significant marker in a QTL credible interval when the other QTL listed for a particular phenotype within a family are included in the model. The absolute value is given because phase information of the most significant marker is not informative due to linkage with markers of opposite phase.

[f] Location intervals are based on the 12X.2 version (URGI 2014) of the grapevine reference genome. The middle value represents the location of the most significant marker.

11 from ‘Horizon’ that co-located with those from the single phenotype analysis and on chromosome 16 from *V. rupestris* B38 that was not found using the single phenotype analysis. The averaged Bayesian network for the HC family showed leaf trichomes having a negative effect on both sporulation and HR and HR having a positive effect on sporulation (Figure 2.4). A QTL on chromosome 6 from ‘Horizon’ was found to have an effect on HR and a QTL on chromosome 7 from ‘Horizon’ was found to have an effect on both sporulation and leaf trichomes. Two additional QTL on chromosomes 8 and 15 from ‘Horizon’ had an effect on leaf trichomes. The QTL on chromosomes 6, 7, and 8 co-located with those from the single phenotype analysis.

Of the RNA-Seq reads from the 28 genotypes, on average 86.5% aligned to the 12X.2 reference genome, ranging from 84.3% to 88.3%. In total, 30,402 genetic loci were detected in the transcriptome, 45% of which had a FPKM variance across the genotypes of less than one. There were 95 differentially transcribed genes, of which 75 were located on chromosome 14. Twenty-seven were within the 95% credible intervals for the two QTL on chromosome 14 from *V. rupestris* B38 and ‘Horizon’. Among these, two did not correspond to those existing in the CRIBI annotation. The 25 remaining genes are listed in Table 2.5 and their protein GO terms are given in Table 2.6. Nine of the 25 genes were in the credible interval of the QTL from *V. rupestris* B38, while the other 16 were in the credible interval of the QTL from ‘Horizon’.

Table 2.4: The physical locations of QTL found using the multiple phenotype Bayesian network analysis performed on breeding values of the *V. rupestris* B38 × 'Horizon' (RH) and 'Horizon' × *V. cinerea* B9 (HC) F_1 families incorporating data across two years and multiple experiments.

Name	Family	Chr	Heterozygous Parent[a]	Phenotype(s)	95% Confidence Interval (Mbp)[b]
Rpv17	RH	8	Horizon	Hypersensitive response (HR)	11.656-11.656-11.961
Rpv18	RH	11	Horizon	Hypersensitive response (HR)	15.397
Rpv19	RH	14	*V. rupestris* B38	Manual sporulation (Sp)	29.543
	RH	16	*V. rupestris* B38	Hypersensitive response (HR)	22.124
Rpv20	HC	6	Horizon	Hypersensitive response (HR)	6.64
Rpv21	HC	7	Horizon	Manual sporulation (Sp) and leaf trichomes (Lt)	1.455-2.610-4.080
	HC	8	Horizon	Leaf trichomes (Lt)	17.545-17.766-21.504
	HC	15	Horizon	Leaf trichomes (Lt)	17.663

[a] The heterozygous parent is the one that has the heterozygous allele. SNPs for the genotypes in the families either were homozygous for one allele or heterozygous because only pseudo-testcross markers were used to build the genetic maps (Hyma et al. 2015).

[b] Location intervals are based on the 12X.2 version (URGI 2014) of the grapevine reference genome. The middle value represents the location of the markers in the networks in Figs. 2 and 3. The single location for certain QTL is due to the absence of other SNPs affecting the phenotype in at least 5% of the networks.

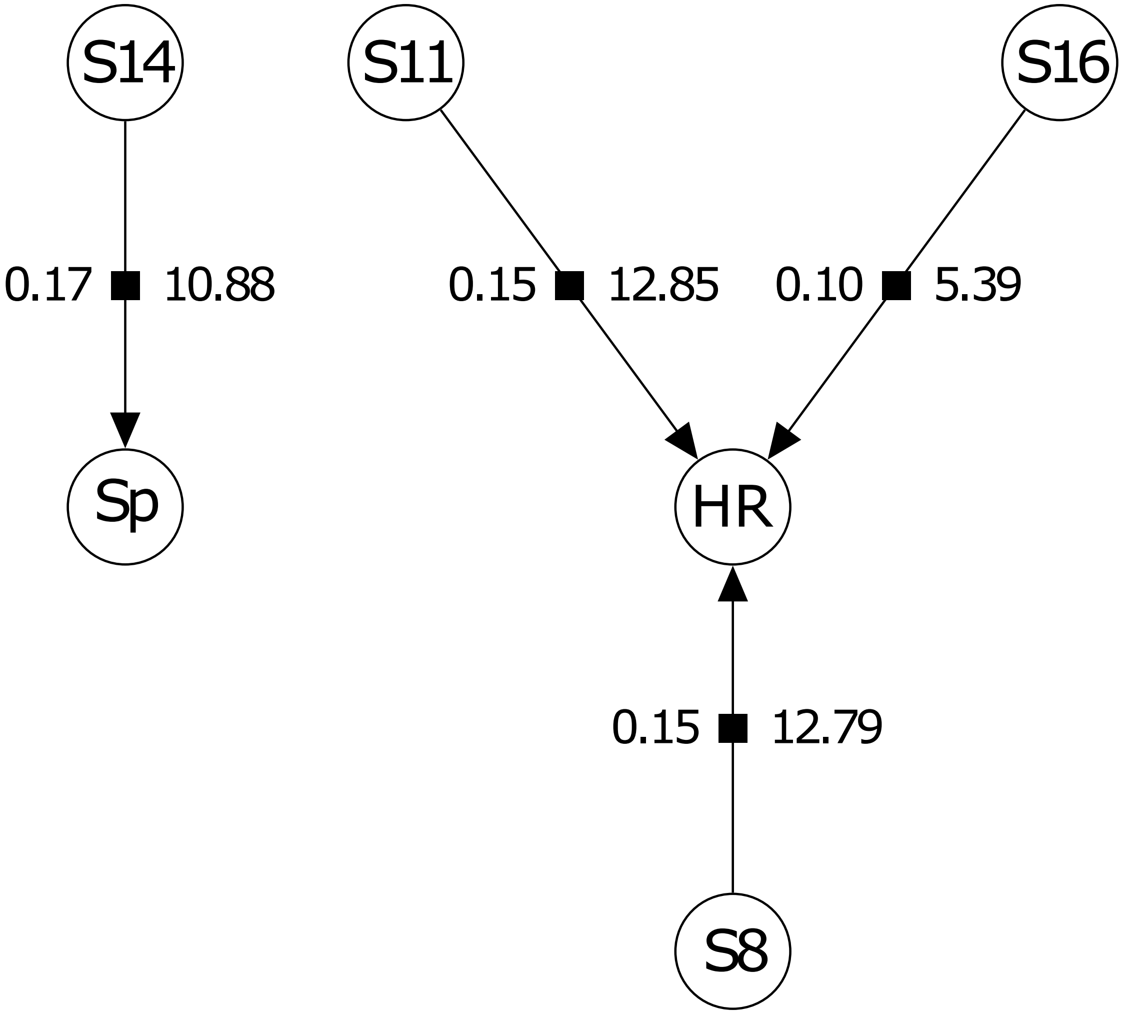

Figure 2.3: The averaged Bayesian network for the RH family manual sporulation (Sp) and hypersensitive response (HR) traits. S8, S11, S14, and S16 correspond to SNPs on chromosomes 8, 11, 14, and 16, respectively. The number to the left of an edge pointing from a SNP to a trait represents the absolute effect size of the SNP on the trait while the number to the right represents the percent variance of the trait explained by the SNP calculated as $\frac{\text{Type III SS}}{\text{Total SS}} \times 100$. The SNP confidence intervals are given in Table 2.4.

Table 2.5: A list of differentially transcribed candidate genes within two QTL credible intervals on chromosome 14 from the *V. rupestris* B38 × ‘Horizon’ F_1 family ordered by significance. GO terms for these genes are given in Table 2.6.

Gene ID	UniProt ID	Protein Name	Heterozygous Parent	Fold Change[a]	q Value	Transcript Range (bp)[b]
VIT_14s0108g01130	F6H5V6	Putative uncharacterized protein	*V. rupestris* B38	0.33—2.97	1.11E-11	29767850-29771739
VIT_14s0068g00980	D7SVG4	Putative uncharacterized protein	Horizon	1.50—0.67	0.000004	24740077-24753861
VIT_14s0068g01970	F6H3X5	Putative uncharacterized protein	Horizon	0.57—1.76	0.000005	25633234-25634255
VIT_14s0108g00040	D7SX63	FACT complex subunit SSRP1	*V. rupestris* B38	0.85—1.17	0.000044	28871463-28878208
VIT_14s0066g02550	D7TX08	Plasma membrane ATPase	*V. rupestris* B38	1.50—0.66	0.000229	28745606-28754974
VIT_14s0068g01730	D7SVM8	Putative uncharacterized protein	Horizon	1.51—0.66	0.000245	25424955-25439887
VIT_14s0068g00800	D7SVE8	Putative uncharacterized protein	Horizon	0.81—1.23	0.000316	24574253-24580946
VIT_14s0108g00150	F6H5P3	Putative uncharacterized protein	*V. rupestris* B38	1.51—0.66	0.000628	28944177-28951278
VIT_14s0068g01720	F6H465	Putative uncharacterized protein	Horizon	0.65—1.53	0.002021	25424822-25425652
VIT_14s0068g01100	F6H429	Putative uncharacterized protein	Horizon	1.19—0.84	0.002182	24882721-24885537
VIT_14s0068g02110	A5BE40	Putative uncharacterized protein	Horizon	1.43—0.70	0.002228	25707772-25712673
VIT_14s0068g01940	A5BCW2	Putative uncharacterized protein	Horizon	0.81—1.24	0.002607	25605002-25609466
VIT_14s0066g02560	A5BD80	Thioredoxin h4	*V. rupestris* B38	0.83—1.20	0.008721	28755934-28756776
VIT_14s0068g00660	Q19N38	WD repeat 2	Horizon	1.30—0.77	0.010981	24473597-24477950
VIT_14s0108g01100	F6H5V3	Putative uncharacterized protein	*V. rupestris* B38	1.37—0.73	0.012521	29753062-29756587
VIT_14s0108g00070	D7SX65	Putative uncharacterized protein	*V. rupestris* B38	0.71—1.40	0.013738	28889440-28894772
VIT_14s0066g00830	D7TWK7	Putative uncharacterized protein	*V. rupestris* B38	0.75—1.33	0.017500	27308643-27313582
VIT_14s0068g01310	F6H442	Putative uncharacterized protein	Horizon	0.82—1.22	0.021602	25043826-25044557
VIT_14s0068g01710	D7SVM6	Putative uncharacterized protein	Horizon	0.75—1.33	0.021602	25408934-25417805
VIT_14s0066g02380	A5AEQ0	Putative uncharacterized protein	*V. rupestris* B38	1.16—0.86	0.022765	28580980-28583873
VIT_14s0068g01840	A5BFT4	Putative uncharacterized protein	Horizon	0.72—1.39	0.024415	25554099-25554809
VIT_14s0068g01260	F6H440	Putative uncharacterized protein	Horizon	0.58—1.71	0.027264	25012954-25022155
VIT_14s0068g01990	D7SVP7	Putative uncharacterized protein	Horizon	0.72—1.39	0.028077	25641499-25645266
VIT_14s0068g01270	F6H441	Putative uncharacterized protein	Horizon	1.84—0.54	0.029598	25022156-25022479
VIT_14s0068g01160	D7SVI1	Putative uncharacterized protein	Horizon	2.20—0.45	0.048404	24943888-24944747

[a] Calculated as two raised to a power equal to that of the effect size of the grouping covariate for a given gene. Two fold change values are given due to the uncertainty of which phase the effect is coming from for a given QTL. Each column left or right of the vertical bar represents one phase.

[b] The range was calculated by finding the start and end locations of all transcripts for a gene and finding the most proximal start location and the most distal end location.

Table 2.6: A list of differentially transcribed candidate genes and their corresponding GO terms within two QTL credible intervals on chromosome 14 from the *V. rupestris* B38 × ‘Horizon’ F_1 family ordered by significance.

Gene ID	UniProt ID	GO - Molecular Function
VIT_14s0108g01130	F6H5V6	NA
VIT_14s0068g00980	D7SVG4	NA
VIT_14s0068g01970	F6H3X5	NA
VIT_14s0108g00040	D7SX63	DNA binding
VIT_14s0066g02550	D7TX08	ATP binding; hydrogen-exporting ATPase activity, phosphorylative mechanism; metal ion binding
VIT_14s0068g01730	D7SVM8	ATPase activity; ATP binding
VIT_14s0068g00800	D7SVE8	SNAP receptor activity; SNARE binding
VIT_14s0108g00150	F6H5P3	NA
VIT_14s0068g01720	F6H465	NA
VIT_14s0068g01100	F6H429	NA
VIT_14s0068g02110	A5BE40	molybdenum ion binding; sulfite oxidase activity
VIT_14s0068g01940	A5BCW2	P-P-bond-hydrolysis-driven protein transmembrane transporter activity
VIT_14s0066g02560	A5BD80	oxidoreductase activity, acting on a sulfur group of donors, disulfide as acceptor; protein disulfide oxidoreductase activity; protein-disulfide reductase activity; thioredoxin-disulfide reductase activity
VIT_14s0068g00660	Q19N38	NA
VIT_14s0108g01100	F6H5V3	NA
VIT_14s0108g00070	D7SX65	NA
VIT_14s0066g00830	D7TWK7	sirohydrochlorin cobaltochelatase activity
VIT_14s0068g01310	F6H442	NA
VIT_14s0068g01710	D7SVM6	transmembrane transporter activity
VIT_14s0066g02380	A5AEQ0	NA
VIT_14s0068g01840	A5BFT4	NA
VIT_14s0068g01260	F6H440	phosphate ion transmembrane transporter activity
VIT_14s0068g01990	D7SVP7	NA
VIT_14s0068g01270	F6H441	NA
VIT_14s0068g01160	D7SVI1	NA

Table 2.6 (continued): A list of differentially transcribed candidate genes and their corresponding GO terms within two QTL credible intervals on chromosome 14 from the *V. rupestris* B38 × ‘Horizon’ F_1 family ordered by significance.

Gene ID	UniProt ID	GO - Biological Process
VIT_14s0108g01130	F6H5V6	NA
VIT_14s0068g00980	D7SVG4	NA
VIT_14s0068g01970	F6H3X5	xylan biosynthetic process
VIT_14s0108g00040	D7SX63	DNA repair; DNA replication; regulation of transcription, DNA-templated; transcription, DNA-templated
VIT_14s0066g02550	D7TX08	ATP biosynthetic process
VIT_14s0068g01730	D7SVM8	NA
VIT_14s0068g00800	D7SVE8	intracellular protein transport; vesicle docking; vesicle fusion
VIT_14s0108g00150	F6H5P3	NA
VIT_14s0068g01720	F6H465	NA
VIT_14s0068g01100	F6H429	NA
VIT_14s0068g02110	A5BE40	chlorophyll metabolic process; nitrate assimilation; response to sulfur dioxide; sulfur compound metabolic process
VIT_14s0068g01940	A5BCW2	protein targeting
VIT_14s0066g02560	A5BD80	cell redox homeostasis; cellular response to oxidative stress; glycerol ether metabolic process
VIT_14s0068g00660	Q19N38	NA
VIT_14s0108g01100	F6H5V3	NA
VIT_14s0108g00070	D7SX65	NA
VIT_14s0066g00830	D7TWK7	cobalamin biosynthetic process
VIT_14s0068g01310	F6H442	NA
VIT_14s0068g01710	D7SVM6	NA
VIT_14s0066g02380	A5AEQ0	NA
VIT_14s0068g01840	A5BFT4	NA
VIT_14s0068g01260	F6H440	mitochondrial transport; response to salt stress
VIT_14s0068g01990	D7SVP7	NA
VIT_14s0068g01270	F6H441	NA
VIT_14s0068g01160	D7SVI1	NA

Table 2.6 (continued): A list of differentially transcribed candidate genes and their corresponding GO terms within two QTL credible intervals on chromosome 14 from the *V. rupestris* B38 × 'Horizon' F_1 family ordered by significance.

Gene ID	UniProt ID	GO - Cellular Component
VIT_14s0108g01130	F6H5V6	integral component of membrane; plasma membrane; trans-Golgi network
VIT_14s0068g00980	D7SVG4	NA
VIT_14s0068g01970	F6H3X5	integral component of membrane
VIT_14s0108g00040	D7SX63	chromosome; nucleus
VIT_14s0066g02550	D7TX08	integral component of membrane
VIT_14s0068g01730	D7SVM8	NA
VIT_14s0068g00800	D7SVE8	endomembrane system; integral component of membrane; SNARE complex
VIT_14s0108g00150	F6H5P3	NA
VIT_14s0068g01720	F6H465	NA
VIT_14s0068g01100	F6H429	NA
VIT_14s0068g02110	A5BE40	mitochondrion; peroxisome
VIT_14s0068g01940	A5BCW2	integral component of membrane; intracellular
VIT_14s0066g02560	A5BD80	cytoplasm
VIT_14s0068g00660	Q19N38	nucleus
VIT_14s0108g01100	F6H5V3	NA
VIT_14s0108g00070	D7SX65	NA
VIT_14s0066g00830	D7TWK7	NA
VIT_14s0068g01310	F6H442	NA
VIT_14s0068g01710	D7SVM6	integral component of membrane; vacuolar membrane
VIT_14s0066g02380	A5AEQ0	NA
VIT_14s0068g01840	A5BFT4	integral component of membrane
VIT_14s0068g01260	F6H440	cell wall; chloroplast; integral component of membrane; mitochondrial inner membrane; vacuolar membrane
VIT_14s0068g01990	D7SVP7	NA
VIT_14s0068g01270	F6H441	NA
VIT_14s0068g01160	D7SVI1	NA

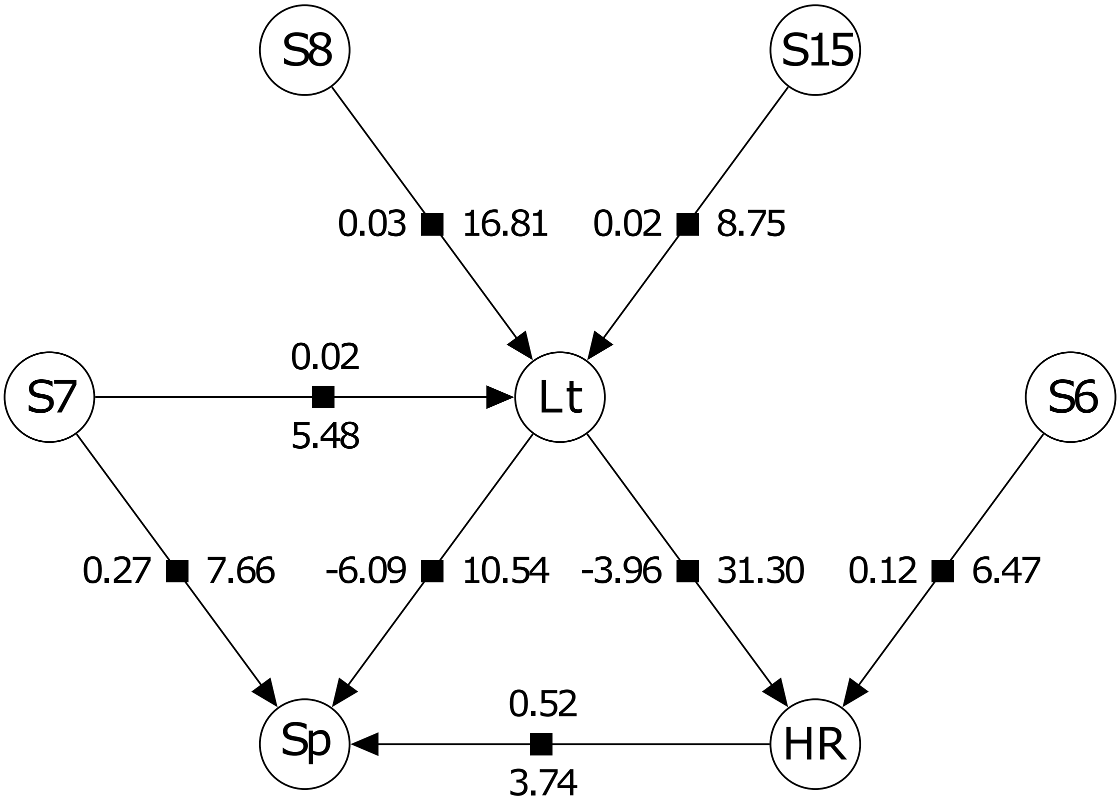

Figure 2.4: The averaged Bayesian network for the HC family manual sporulation (Sp), hypersensitive response (HR), and leaf trichome (Lt) traits. S6, S7, S8, and S15 correspond to SNPs on chromosomes 6, 7, 8, and 15, respectively. The numbers above and to the left of an edge pointing from a trait/SNP to a trait represents the effect size of the trait/SNP on the trait while the numbers below and to the right represents the percent variance of the trait explained by the trait/SNP calculated as $\frac{\text{Type III SS}}{\text{Total SS}} \times 100$. Effect sizes from SNPs are absolute values while those from traits are not. The SNP confidence intervals are given in Table 2.4.

2.5 Discussion

When searching for QTL to introgress into grapevine germplasm, we are particularly interested in those QTL that show a consistent effect across years. In our linear mixed model approach, we were able to take into account the effects of year and genotype-by-year interaction because we had two years of phenotypic data. However, the model was naïve with respect to the assumption of independence between years. While we replicated within individual time points since each phenotype was the average of eight leaf disc ratings, this replication does not take into account the effect of the environmental conditions prior to leaf harvest. Thus, when analyzing experiments individually for QTL, as we did using the individual time point analysis, there is a greater level of uncertainty surrounding the phenotypes, which play the role of breeding values in the individual time point analysis, than if the genotypes were replicated across multiple experiments. As we replicate genotypes, we are more certain of their estimated breeding values, and, consequently, of the estimated marker effects. This may explain why none of the QTL found only once using the individual time point analysis were found using the single phenotype analysis and why the QTL found using the individual time point phenotypes have much larger credible intervals.

The QTL found in the RH family using the single phenotype analysis were the same for the manual or computer vision sporulation breeding values with the exception of one additional QTL for the manual sporulation breeding values (Table 1). This is consistent with the finding that the two phenotypes are highly correlated (Divilov et al. 2017). The failure to find any QTL using the computer vision sporulation breeding values in the HC family and the low heritability found for that phenotype reflects the poor accounting of lighting conditions in the computer vision system discussed previously (Divilov et al. 2017). The QTL on chromosomes 5 and 8 from 'Horizon' found previously for leaf trichome density using the 2 dpi computer vision rating (Divilov et al. 2017) were found to be QTL for disease resistance when using the manual sporulation and HR breeding values in the single phenotype analysis. This is consistent with our finding that leaf trichomes have an effect on manual sporulation and HR using the multiple phenotype Bayesian network approach. The averaged Bayesian network suggests that the QTL on chromosome 8 operates through the modulation of leaf trichomes (Figure 3). In the same network, we found that the presence of HR is associated with an increase in sporulation. While HR is a component of a resistance reaction, it does not imply that the plant is able to completely stop pathogen colonization. However, we did not find this association in the RH family. Further experimentation would be required to determine the cause of this discrepancy. Likewise, the suggested pleiotropic effect of the QTL on chromosome 7 on sporulation and leaf trichomes would be another interesting mechanism to dissect.

While most QTL found using the single phenotype analysis were found using the Bayesian network analysis, some of the confidence intervals in the latter analysis were not informative (Table 2). For the RH family, the SNP found in the averaged network to affect a trait was on a different parental map than the other SNPs on the same chromosome found to affect the trait in at least 5% of the networks. For the HC family, only one SNP was found for the QTL on chromosomes 6 and 15 in at least 5% of the networks. This does not mean that the QTL location is known with certainty, but rather is a result of the robustness of the Markov blanket found for the traits. To obtain a more informative level of uncertainty of the QTL physical positions in such situations, one can obtain an approximate Bayesian credible interval as was done in the single phenotype approach.

There are currently 16 known QTL reported to be associated with downy mildew disease resistance (VIVC 2017). No QTL have been found on chromosomes 6, 8, 11, or 16, where we found QTL associated with sporulation and HR. While QTL on chromosome 14 have previously been found, the QTL found by Blasi et al. (2011) and Venuti et al. (2013) do not physically co-locate to the QTL we found on chromosome 14. The QTL on chromosome 7 found using the HC family single phenotype and Bayesian network analyses did not co-locate to one previously found on chromosome 7 by Moreira et al. (2011). We have designated the QTL on chromosome 8 from 'Horizon' *Rpv17*; the QTL on chromosome 11 from 'Horizon' *Rpv18*; the QTL on chromosome 14 from *V. rupestris* B38 *Rpv19*; the QTL on chromosome 6 from 'Horizon' *Rpv20*; and the QTL on chromosome 7 from 'Horizon' *Rpv21.* Because the stepwise regression and Bayesian network analyses disagreed on whether there are one or two QTL on chromosome 14, we only assigned a name to one QTL on that chromosome. Additionally, we have not proposed naming any QTL found only using stepwise regression or only using the Bayesian network analysis.

Since the QTL on chromosome 14 did not explain a large portion of the phenotypic variance, we did not expect to find differentially transcribed NBS-encoding genes, which tend to play a qualitative role in disease resistance, although this role is not absolute (Poland et al. 2009). While a gene encoding FACT (FAcilitates Chromatin Transcription) complex subunit SSRP1 (Structure Specific Recognition Protein 1), which has a GO term associated with DNA binding, was differentially transcribed, we could not find any literature showing its role in disease resistance. Thioredoxin h4, which is part of a family of proteins that interacts with peroxidases (Arnér and Holmgren 2000), is a promising candidate protein because there is evidence that peroxidase activity is associated with downy mildew disease resistance (Kortekamp et al. 1998). The protein A5BE40 likely plays a related role as it is localizes to the peroxisome. Another promising candidate gene is VIT_14s0068g00800, whose protein is predicted to interact in a SNARE (SNAP [Soluble NSF [N-ethylmaleimide-Sensitive Factor] Attachment Protein] REceptor) complex. Proteins in *Arabidopsis thaliana* interacting in SNARE complexes were shown to be responsible for non-host resistance to powdery mildew of barley (Collins et al. 2003). VIT_14s0068g01970, which is involved in xylan biosynthesis, is another promising candidate gene because the presence of xylan, a type of hemicellulose, was shown to be associated with increased infection of *Fusarium herbarum* on wheat (Wingard 1941). Unfortunately, most differentially transcribed genes were associated with proteins of unknown functionality.

In this study using a multi-year, multi-experiment analysis, we have shown that disease resistance in the genotypes studied is a trait controlled by many QTL with small effect sizes, and is influenced by leaf trichomes as well. Five QTL not previously described were identified. For one F_1 family studied, we have some evidence that the underlying candidate genes are likely not NBS-encoding genes. We used a susceptible *V. vinifera* genome sequence for our analysis as no resistant wild *Vitis* genome has been assembled, so our gene search was partially biased. We believe breeding with these quantitative genes will help to generate durably resistant grapevine cultivars compared to breeding solely with R genes (Poland et al. 2009). Because we found that leaf trichomes have an effect on disease resistance in one F_1 family, breeding for leaf trichomes presents another opportunity to select for downy mildew disease resistance. We expect that stacking QTL for leaf trichomes and sporulation and HR disease resistance would produce more durable resistance than breeding for one mechanism of resistance alone.

2.6 Acknowledgments

We thank Nicholas Santantonio for discussions about statistical methodology and Alisson Kovaleski for help with RNA extraction. We thank Mike Colizzi and Steve Luce for vineyard maintenance; Mike Colizzi, Corrigan Herbert, Aaron Green, Hema Kasinathan, Caitlyn Kumkey, Hanna Martens, Anne Repka, Michelle

Schaub, and Mary Jean Welser for help collecting leaves and plating leaf discs; and David Gadoury, Michelle Schaub, and Wayne Wilcox for supplying leaves for propagation of the *P. viticola* isolate used for inoculation. The authors are grateful for the prior work on downy mildew resistance performed with a subset of the same families by Atsushi Kono and Ming-Te Lu. We also thank Michael Gore for feedback on the manuscript. This research was supported by the USDA-NIFA Specialty Crop Research Initiative (Award No. 2011-51181-30635), The New York Wine & Grape Foundation, Federal Capacity Funds and the Lake Erie Regional Grape Processor's Fund. Graduate Assistantship support for Konstantin Divilov was provided by the Charles R. Bullis Plant Hybridization Endowment and the Michael Nolan Endowment Fund.

2.7 References

Akdemir, D. and O. U. Godfrey (2015). *EMMREML: Fitting Mixed Models with Known Covariance Structures*. R package version 3.1. URL: `https://CRAN.R-project.org/package=EMMREML`.

Aliferis, C. F., A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos (2010). "Local causal and markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation". *Journal of Machine Learning Research* 11, 171–234.

Arnér, E. S. and A. Holmgren (2000). "Physiological functions of thioredoxin and thioredoxin reductase". *The FEBS Journal* 267 (20), 6102–6109.

Bellin, D., E. Peressotti, D. Merdinoglu, S. Wiedemann-Merdinoglu, A.-F. Adam-Blondon, G. Cipriani, M. Morgante, R. Testolin, and G. Di Gaspero (2009). "Resistance to *Plasmopara viticola* in grapevine 'Bianca' is controlled by a major dominant gene causing localised necrosis at the infection site". *Theoretical and Applied Genetics* 120 (1), 163–176.

Blasi, P., S. Blanc, S. Wiedemann-Merdinoglu, E. Prado, E. H. Rühl, P. Mestre, and D. Merdinoglu (2011). "Construction of a reference linkage map of *Vitis amurensis* and genetic mapping of *Rpv8*, a locus conferring resistance to grapevine downy mildew". *Theoretical and Applied Genetics* 123 (1), 43–53.

Broman, K. W., H. Wu, Ś. Sen, and G. A. Churchill (2003). "R/qtl: QTL mapping in experimental crosses". *Bioinformatics* 19 (7), 889–890.

Buonassisi, D., M. Colombo, D. Migliaro, C. Dolzani, E. Peressotti, C. Mizzotti, R. Velasco, S. Masiero, M. Perazzolli, and S. Vezzulli (2017). "Breeding for grapevine downy mildew resistance: a review of "omics" approaches". *Euphytica* 213 (5), 103.

Cadle-Davidson, L. (2008). "Variation within and between *Vitis* spp. for foliar resistance to the downy mildew pathogen *Plasmopara viticola*". *Plant Disease* 92 (11), 1577–1584.

Collins, N. C., H. Thordal-Christensen, V. Lipka, S. Bau, et al. (2003). "SNARE-protein-mediated disease resistance at the plant cell wall". *Nature* 425, 973–977.

Consortium, T. U. (2017). "UniProt: the universal protein knowledgebase". *Nucleic Acids Research* 45 (D1), D158–D169.

Divilov, K., T. Wiesner-Hanks, P. Barba, L. Cadle-Davidson, and B. I. Reisch (2017). "Computer vision for high-throughput quantitative phenotyping: a case study of grapevine downy mildew sporulation and leaf trichomes". *Phytopathology* 107 (12), 1549–1555.

Endelman, J. B. (2011). "Ridge regression and other kernels for genomic selection with R package rrBLUP". *The Plant Genome* 4 (3), 250–255.

Frazee, A. C., G. Pertea, A. E. Jaffe, B. Langmead, S. L. Salzberg, and J. T. Leek (2015). “Ballgown bridges the gap between transcriptome assembly and expression analysis”. *Nature Biotechnology* 33 (3), 243–246.

Hyma, K. E., P. Barba, M. Wang, J. P. Londo, C. B. Acharya, S. E. Mitchell, Q. Sun, B. Reisch, and L. Cadle-Davidson (2015). “Heterozygous mapping strategy (HetMappS) for high resolution genotyping-by-sequencing markers: a case study in grapevine”. *PloS ONE* 10 (8), e0134880.

Jaillon, O., J.-M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, et al. (2007). “The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla”. *Nature* 449, 463–467.

Kim, D., B. Langmead, and S. L. Salzberg (2015). “HISAT: a fast spliced aligner with low memory requirements”. *Nature Methods* 12 (4), 357–360.

Kortekamp, A., R. Wind, and E. Zyprian (1998). “Investigation of the interaction of *Plasmopara viticola* with susceptible and resistant grapevine cultivars”. *Journal of Plant Diseases and Protection* 105 (5), 475–488.

Kortekamp, A. and E. Zyprian (1999). “Leaf hairs as a basic protective barrier against downy mildew of grape”. *Journal of Phytopathology* 147 (7-8), 453–459.

Luttinen, J. (2013). *BayesNet.* URL: `https://github.com/jluttine/tikz-bayesnet`.

McCulloch, C. E. and S. R. Searle (2001). *Generalized, Linear, and Mixed Models.* New York, NY: John Wiley & Sons, Inc.

McDonald, B. A. and C. Linde (2002). “Pathogen population genetics, evolutionary potential, and durable resistance”. *Annual Review of Phytopathology* 40 (1), 349–379.

Moreira, F. M., A. Madini, R. Marino, L. Zulini, M. Stefanini, R. Velasco, P. Kozma, and M. S. Grando (2011). “Genetic linkage maps of two interspecific grape crosses (*Vitis* spp.) used to localize quantitative trait loci for downy mildew resistance”. *Tree Genetics & Genomes* 7 (1), 153–167.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Pertea, M., G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg (2015). “StringTie enables improved reconstruction of a transcriptome from RNA-seq reads”. *Nature Biotechnology* 33 (3), 290–295.

Poland, J. A., P. J. Balint-Kurti, R. J. Wisser, R. C. Pratt, and R. J. Nelson (2009). “Shades of gray: the world of quantitative disease resistance”. *Trends in Plant Science* 14 (1), 21–29.

Reisch, B., W. Robinson, K. Kimball, R. Pool, and J. Watson (1983). “‘Horizon’ Grape”. *HortScience* 18, 108–109.

Scutari, M. (2010). “Learning Bayesian networks with the bnlearn R Package”. *Journal of Statistical Software* 35 (3), 1–22.

Scutari, M., P. Howell, D. J. Balding, and I. Mackay (2014). “Multiple quantitative trait analysis using bayesian networks”. *Genetics* 198 (1), 129–137.

URGI (2014). *12X.2 version of the grapevine reference genome sequence from The French-Italian Public Consortium (PN40024).* URL: `https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences`.

Venuti, S., D. Copetti, S. Foria, L. Falginella, S. Hoffmann, D. Bellin, P. Cindrić, P. Kozma, S. Scalabrin, M. Morgante, et al. (2013). “Historical introgression of the downy mildew resistance gene *Rpv12* from the Asian species *Vitis amurensis* into grapevine varieties”. *PLoS ONE* 8 (4), e61228.

Vitulo, N., C. Forcato, E. C. Carpinelli, A. Telatin, D. Campagna, M. D'Angelo, R. Zimbello, M. Corso, A. Vannozzi, C. Bonghi, et al. (2014). "A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype". *BMC Plant Biology* 14 (1), 99.

VIVC (2017). *Table of loci for traits in grapevine relevant for breeding and genetics.* URL: `http://www.vivc.de/index.php?r=dbsearch%2Fdataonbreeding`.

Wang, Z., M. Gerstein, and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". *Nature Reviews Genetics* 10 (1), 57–63.

Wingard, S. (1941). "The nature of disease resistance in plants. I". *The Botanical Review* 7 (2), 59–109.

Zhong, S., J.-G. Joung, Y. Zheng, Y.-r. Chen, B. Liu, Y. Shao, J. Z. Xiang, Z. Fei, and J. J. Giovannoni (2011). "High-throughput Illumina strand-specific RNA sequencing library preparation". *Cold Spring Harbor Protocols* 2011 (8), 940–949.

CHAPTER 3

VINEYARD SPATIAL ANALYSIS USING ARMA PROCESSES

3.1 Abstract

In vineyards, spatial effects are likely involved in most measurable phenotypes. Separation of spatial effects from phenotypes has been shown to improve selection accuracy in field trials of agronomic crops. For one out of three vineyard data sets analyzed, which had genetic data, the use of autoregressive-moving average (ARMA) processes in a Gaussian process regression model improved genomic prediction accuracy. Quantitative trait loci (QTL) identified were the same using either spatially unadjusted or adjusted phenotypes. For the two data sets with only phenotypic data, ARMA processes explained some of the variance in the phenotypes. Differential evolution was used to optimize the log marginal likelihood of the models, presenting an alternative to derivative-based optimization.

3.2 Introduction

Spatial variation due to soil and environment has long been known to affect phenotypes measured in the field (Fisher 1935). Two general strategies have been developed to account for this variation. The first strategy consists of dividing a field into blocks and replicating genotypes within the blocks, which are then used as fixed effects. As grapevines are clonally propagated, such a strategy would slow down a grapevine breeding program. The second strategy is to assume that the spatial variation in the field is generated by a stochastic process (Gleeson and Cullis 1987). Such a strategy does not require replication of vines. One then needs to decide which stochastic process best fits the data. Two options exist for possible objective functions. The first is the marginal likelihood of an estimated model. A common model used in spatial analysis is a Gaussian process regression model (Rasmussen and Williams 2006), also known as a kriging model in geostatistics or a best linear unbiased predictor (BLUP) model in animal and plant breeding (Morota and Gianola 2014). The second possible objective function is the cross-validation accuracy of genomic prediction (Lado et al. 2013). In genomic prediction, genetic markers are used to build a covariance matrix that is then used as a prior in a Gaussian process regression model (Morota and Gianola 2014), although other models have also been used (Meuwissen et al. 2001). Genomic prediction accuracy is of particular importance to plant breeders because improving genomic prediction accuracy improves the genetic gain of genomic selection. Here, three data sets were used to quantify how well a set of stochastic processes captured spatial variation in vineyards.

3.3 Vineyard Data Sets

The first data set consisted of phenotypic and genetic data from a vineyard planted in 1990 in Geneva, NY, with a 'Horizon' × Illinois 547-1 F_1 family as well as 'Chancellor', 'Concord', 'Steuben', and PI 200569 checks. The F_1 family genotypes were unreplicated while 'Chancellor', 'Concord', and 'Steuben' were replicated seven times and PI 200569 was replicated twice in various locations. The vineyard was planted in six rows and vine spacing was 1.2 m within rows and 2.7 m between rows, except the spacing between the first and second row, which was 5.5 m. In 2002, 475 vines, including the controls, were phenotyped for powdery mildew resistance in the field on a 1 to 5 scale representing percent area of sporulation on the vine with 1 = 0-3%, 2 = 3-12%, 3 = 12-25%, 4 = 25-50%, and 5 = >50%. Of the 475 vines, 358 were genotyped using genotyping-by-sequencing to obtain 9,876 high quality single nucleotide polymorphisms (SNPs) (Hyma et al. 2015). Missing SNP data were imputed using the expectation-maximization algorithm in the rrBLUP R package (Endelman 2011). Phenotypic data were available for 350 of 358 genotyped vines.

The second data set originated with a vineyard planted in 2016 in Geneva, NY, with a NY84.0101.03 × *Vitis rupestris* 'Pillans' F_1 family as well as NY84.0101.03, *V. rupestris* 'Pillans', 'Chancellor', and 'Concord' checks. NY84.0101.03 is a hybrid selection from the Cornell grapevine breeding program. The F_1 family genotypes were unreplicated, while NY84.0101.03 and *V. rupestris* 'Pillans' were replicated four times and 'Chancellor' and 'Concord' were replicated twice. The vineyard was planted in four rows with 76 vines per row. Spacing was 1.8 m within rows and 2.7 m between rows. In 2017, 297 surviving vines, controls inclusive, of the 304 initially planted were phenotyped for powdery mildew resistance using the same 1 to 5 scale used in the first data set.

The third data set was collected from a vineyard planted in 2016 in Geneva, NY with a NY84.0100.03 × 'Himrod' F_1 family as well as NY84.0100.03 and 'Himrod' checks. The F_1 family genotypes were unreplicated while the checks were replicated four times. The vineyard was planted in three rows with 61 vines per row. Each vine was 1.8 m apart within a row, and between row spacing was 2.7 m. In 2017, 177 vines, controls inclusive, which had never been hedged, were phenotyped for plant height by measuring the distance from the base of each vine to the final node on the longest shoot.

3.4 Spatial Analysis

In a randomized vineyard, the objective of spatial analysis is to find the function $f(\boldsymbol{X})$ that explains the phenotypes $\boldsymbol{y}$ of the vines in the vineyard, where $\boldsymbol{X}$ is a matrix specifying the rows and columns of the vines. $f(\boldsymbol{X})$ can be either a fixed (deterministic) or random (stochastic) function/process or a combination of the two. The spatial variation is likely the additive and/or multiplicative effect of many functions of soil, weather, and genetics, so by the Central Limit Theorem (Bertsekas and Tsitsiklis 2002), the spatial effect between any subset of locations can be assumed to be produced from a Gaussian distribution, i.e., $f(\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$, where $\boldsymbol{m}$ are possible fixed effects and $\boldsymbol{\Sigma}$ is the covariance matrix. This is equivalent to modeling the spatial effects as a Gaussian process, $f(\boldsymbol{X}) \sim \mathcal{GP}(\boldsymbol{m}, \boldsymbol{\Sigma})$. A Gaussian process is a collection of variables any subset of which has a Gaussian distribution (Rasmussen and Williams 2006).

If there are no fixed effects to consider, the phenotypic mean can be subtracted from the phenotypes and the spatial function is now $f(\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Here, the spatial function is only controlled by the covariance matrix $\boldsymbol{\Sigma}$ of the vine locations, and this matrix will encode the prior assumptions about how the locations affect each other. Assuming that the covariance between two locations, i and j, within a row or column, $|i-j|$ locations apart is the same for any two locations, within a row or column, $|i-j|$ locations apart, the spatial stochastic process can be said to be stationary (Box et al. 2015). Stationary processes have a constant mean and thus are specified by their covariance matrix. Two processes, one of which can be made stationary and the other always stationary, are autoregressive processes of order p, or AR(p) processes, and moving average processes of order q, or MA(q) processes. These two processes can be generalized into ARMA(p,q) processes where AR(p)

and MA(q) are ARMA(p,0) and ARMA(0,q), respectively. ARMA processes will be the focus of the present analyses. These processes are common in time series analysis (Box et al. 2015; Murray-Smith and Girard 2001), but have been applied in spatial analysis as well (Bernal-Vasquez et al. 2014; Gleeson and Cullis 1987).

In an AR(p) process, the spatial effect is an additive function of the spatial effect of p adjacent locations, in either direction within a row or column, and white noise, or independent error (Box et al. 2015). The influence of the adjacent locations is modulated by the parameters $\boldsymbol{\phi}$, which are estimated as explained in the next section, with the number of parameters in $\boldsymbol{\phi}$ equal to p. The covariance matrix for an AR(1) process is a symmetric Toeplitz matrix with elements

$$\Gamma_{i,j} = \frac{\phi_1^{|i-j|}}{1-\phi_1^2}\sigma_n \tag{3.1}$$

where σ_n is the noise variance. In a Toeplitz matrix, the elements along each diagonal are equal. The process is stationary when $|\phi_1| < 1$. The covariance matrix for an AR(2) process is symmetric Toeplitz with elements (Muendler 2000)

$$\Gamma_{i,j} = \begin{cases} \frac{1-\phi_2}{(1+\phi_2)((1-\phi_2)^2-\phi_1^2)}\sigma_n^2, & \text{if } |i-j| = 0 \\ \frac{\phi_1}{(1+\phi_2)((1-\phi_2)^2-\phi_1^2)}\sigma_n^2, & \text{if } |i-j| = 1 \\ (\phi_1\Gamma_{i,|i-j|-1} + \phi_2\Gamma_{i,|i-j|-2})\sigma_n^2, & \text{if } |i-j| > 1 \end{cases} \tag{3.2}$$

with recursion used to generate the covariances when $|i-j| > 1$. The process is stationary when $\phi_2 + \phi_1 < 1$, $\phi_2 - \phi_1 < 1$, and $-1 < \phi_2 < 1$.

In a MA(q) process, the spatial effect is an additive function of the white noise of q adjacent locations, in either direction within a row or column, and white noise. The influence of the adjacent locations is modulated by the parameters $\boldsymbol{\theta}$, with the number of parameters in $\boldsymbol{\theta}$ equal to q (Box et al. 2015). All MA processes are stationary regardless of the parameter values. The covariance matrix for a MA(1) process is symmetric Toeplitz with elements

$$\Gamma_{i,j} = \begin{cases} (1+\theta_1^2)\sigma_n^2, & \text{if } |i-j| = 0 \\ -\theta_1\sigma_n^2, & \text{if } |i-j| = 1 \\ 0 & \text{if } |i-j| > 1 \end{cases} \tag{3.3}$$

The covariance matrix for a MA(2) process is symmetric Toeplitz with elements

$$\Gamma_{i,j} = \begin{cases} (1+\theta_1^2+\theta_2^2)\sigma_n^2, & \text{if } |i-j| = 0 \\ (-\theta_1+\theta_1\theta_2)\sigma_n^2, & \text{if } |i-j| = 1 \\ -\theta_2\sigma_n^2 & \text{if } |i-j| = 2 \\ 0 & \text{if } |i-j| > 2 \end{cases} \tag{3.4}$$

A property of AR and MA processes is that any AR(p) process can be written as a MA(∞) process and any MA(q) process can be written as an AR(∞) process (Box et al. 2015). Under certain values of $\boldsymbol{\theta}$, AR(∞) processes associated with MA(q) processes do not converge to finite values and are said to not be invertible. More importantly, when a MA(q) process is not invertible, the $\boldsymbol{\phi}$ parameters in the corresponding AR(∞) process increase as $|i-j|$ increases, which is not realistic in most applications of spatial analysis. The constraints for $\boldsymbol{\theta}$ in order for MA(1) and MA(2) processes to be invertible are the constraints given for $\boldsymbol{\phi}$ above for AR(1) and AR(2) processes to be stationary.

In an ARMA(p,q) process the spatial effect is an additive function of p adjacent locations, the white noise of q adjacent locations, and white noise (Box et al. 2015). The covariance matrix for a ARMA(1,1) process is symmetric Toeplitz with elements

$$\Gamma_{i,j} = \begin{cases} \frac{1+\theta_1^2-2\phi_1\theta_1}{1-\phi_1^2}\sigma_n^2, & \text{if } |i-j| = 0 \\ \frac{(1-\phi_1\theta_1)(\phi_1-\theta_1)}{1-\phi_1^2}\sigma_n^2, & \text{if } |i-j| = 1 \\ \phi_1\Gamma_{i,|i-j|-1})\sigma_n^2, & \text{if } |i-j| > 1 \end{cases} \tag{3.5}$$

The process is stationary when $|\phi_1| < 1$ and invertible when $|\theta_1| < 1$.

As there are multiple rows and columns in the vineyard data sets, and because $\boldsymbol{\Gamma}$ is the same for any row and column, an entire vineyard's spatial covariance matrix can be specified by a Kronecker product of the row-specific and column-specific covariance matrices. For example, for an AR(p) process, the entire vineyard's spatial covariance matrix can be written as $\boldsymbol{\Sigma}_{AR(p)} = \boldsymbol{\Gamma}_r \otimes \boldsymbol{\Gamma}_c$, where $\boldsymbol{\Gamma}_r$ and $\boldsymbol{\Gamma}_c$ are the row-specific and column-specific AR(p) covariance matrices. The assumption for the present analyses is that there is no spatial covariance between adjacent rows within a column. Thus, the complete covariance matrices are $\boldsymbol{\Sigma} = \boldsymbol{I} \otimes \boldsymbol{\Gamma}_c$. This assumption is made due to the small number of rows in all the vineyard data sets, which would make estimates of row-specific $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ inaccurate.

3.5 Gaussian Process Regression

To find the spatial process that best fits the phenotypes of each data set, a Gaussian process regression model will be used (Rasmussen and Williams 2006). In the model, one of the spatial processes explained in the previous section will be the prior for the spatial function. Following Bayes' Theorem, posterior = $\frac{\text{likelihood}\times\text{prior}}{\text{marginal likelihood}}$, or $p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{f},\boldsymbol{X})p(\boldsymbol{f}|\boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})}$, where $\boldsymbol{f}$ are the spatial effects. The prior is $p(\boldsymbol{f}|\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$, where $\boldsymbol{K}$ is a spatial covariance matrix, and the likelihood is $p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{f}, \sigma_n^2\boldsymbol{I})$, where σ_n^2 is the estimated noise variance that is different than the one used in the spatial processes. Maximization of the marginal likelihood, rather than the likelihood, is performed as the marginal likelihood does not necessarily increase with increasing degrees of freedom used by the model, so the model will not overfit the data (Bishop 2006). Thus, the marginal likelihood will select the true model, assuming that the true model is within the space of models being considered. As the phenotypes are the additive combination of the spatial effects and independent noise, i.e., $p(\boldsymbol{y}|\boldsymbol{X}) = \mathcal{N}(\mathbf{0}, \boldsymbol{K}) + \mathcal{N}(\mathbf{0}, \sigma_n^2\boldsymbol{I}) = \mathcal{N}(\mathbf{0}, \boldsymbol{K} + \sigma_n^2\boldsymbol{I})$, the marginal likelihood is

$$p(\boldsymbol{y}|\boldsymbol{X}) = e^{-\frac{1}{2}\boldsymbol{y}^T(\boldsymbol{K}+\sigma_n^2\boldsymbol{I})^{-1}\boldsymbol{y}} \det(\boldsymbol{K} + \sigma_n^2\boldsymbol{I})^{-\frac{1}{2}}(2\pi)^{-\frac{n}{2}} \tag{3.6}$$

Taking the logarithm,

$$\log p(\boldsymbol{y}|\boldsymbol{X}) = -\frac{1}{2}\boldsymbol{y}^T(\boldsymbol{K} + \sigma_n^2\boldsymbol{I})^{-1}\boldsymbol{y} - \frac{1}{2}\log\det(\boldsymbol{K} + \sigma_n^2\boldsymbol{I}) - \frac{n}{2}\log 2\pi \tag{3.7}$$

To avoid performing a matrix inversion, which is a computationally intensive matrix operation, the log marginal likelihood can be computed by (Rasmussen and Williams 2006)

$$\log p(\boldsymbol{y}|\boldsymbol{X}) = -\frac{1}{2}\boldsymbol{y}^T \operatorname{solve}(\boldsymbol{L}^T, \operatorname{solve}(\boldsymbol{L}, \boldsymbol{y})) - \sum_{i=1}^{n} \log L_{ii}) - \frac{n}{2}\log 2\pi \tag{3.8}$$

where $\boldsymbol{L}\boldsymbol{L}^T$ is the Cholesky decomposition of $(\boldsymbol{K} + \sigma_n^2\boldsymbol{I})$ and $\operatorname{solve}(\boldsymbol{A}, \boldsymbol{b})$ solves for $\boldsymbol{c}$ in the equation $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{b}$.

In order to maximize the log marginal likelihood function, one can use either derivative or derivative-free methods. For the present analyses, differential evolution (Storn and Price 1997), a derivative-free optimization algorithm, from the DEoptim R package (Mullen et al. 2011) was used that works by iteratively transforming candidate sets of parameters, in this case $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and $\boldsymbol{\sigma}^2$. Differential evolution has previously been used to optimize the log marginal likelihood function in a Gaussian process regression model (Petelin et al. 2011). During optimization, the log marginal likelihood was penalized when the spatial processes were not stationary or invertible.

After maximization of the log marginal likelihood, the estimated parameters can be used to obtain the spatial effects by $\boldsymbol{K}(\boldsymbol{K} + \sigma_n^2\boldsymbol{I})^{-1}\boldsymbol{y}$ and the number of degrees of freedom used by the model (Hastie and Tibshirani 1990) by $\mathrm{tr}(\boldsymbol{K}(\boldsymbol{K} + \sigma_n^2\boldsymbol{I})^{-1})$, where tr is the trace of a matrix.

3.6 Genomic Prediction and QTL Analysis

In order to see the effects of spatial adjustment for the first vineyard data set, which had genetic data, the phenotypes, subtracted by the spatial effects, were used to cross-validate a genomic prediction model as well as to find QTL. For cross validation, 100 random sets of 90% of the genotypes with phenotypes were used as the training data in a Gaussian process regression model to predict the remaining 10%. The covariance matrix used was $\frac{\boldsymbol{Z}\boldsymbol{Z}^T}{2\sum_{i=1}^{s} m_i(1-m_i)}$ where $\boldsymbol{Z}$ is the centered matrix of allele values with the number of rows equal to the number of genotypes and the number of columns equal to the number of SNPs, m_i is the minor allele frequency for a particular SNP, and s is the number of SNPs (VanRaden 2008). Prediction accuracy was measured by the Pearson correlation coefficient. In this model, the log marginal likelihood function was optimized using the Efficient Mixed Model Association (EMMA) algorithm (Kang et al. 2008) in the sommer R package (Covarrubias-Pazaran 2016). The EMMA algorithm is a fast optimizer when one has a single, static covariance matrix, apart from the diagonal noise matrix, such as the case here. QTL analysis was performed using stepwise regression in R/qtl (Broman et al. 2003) using a previously constructed genetic map (Hyma et al. 2015). Significance of QTL for each phenotype was determined using 1,000 permutation tests with an alpha value of 0.05.

The phenotypic and genetic vineyard data and the code used to perform the analyses can be obtained at https://github.com/kdivilov.

3.7 Results and Discussion

For the three data sets, AR(1), AR(2), MA(1), MA(2), and ARMA(1,1) processes were used as priors in a Gaussian process regression model and the log marginal likelihoods and degrees of freedom of the models were obtained. For the first data set, we also obtained the mean and standard deviations of the cross-validation genomic prediction accuracies using both unadjusted and spatially adjusted phenotypes in a Gaussian process regression model and the number and location of QTL found using stepwise regression.

Across the data sets, AR(p) processes generally fit the data better than MA(q) processes as they achieved higher log marginal likelihood values, with AR(2) processes fitting better than the rest (Table 1). While previous researchers have used the Akaike information criterion (AIC) to penalize the number of parameters being optimized (Leiser et al. 2012), this is not necessary as the model is able to select the optimal number of parameters to use. The fitting of an AR(2) process in the first data set is an example of such selection, where ϕ_2 was set to zero after maximization of the log marginal likelihood. Similarly, the degrees of freedom used by the models were not related to how well they fit the data as expected since maximizing the log marginal likelihood does not lead to overfitting. The reason for this is that $\frac{1}{2} \log \det \left(\boldsymbol{K} + \sigma_n^2 \boldsymbol{I} \right)$ in the log marginal likelihood penalizes model complexity

(Rasmussen and Williams 2006). As mentioned previously, when selecting a model using the log marginal likelihood, one assumes that the true model is within the space of models being considered. This is false as the true spatial process underlying the complex interaction of soil, weather, and genetics cannot be as simple as an AR(1) process, for example. However, it is likely that ARMA processes are good approximations for the true spatial process in most circumstances, and so selection of the prior can be done solely based on the log marginal likelihood, in the absence of genetic data. Such is the case in the second and third data sets where the model estimated spatial variance parameters greater than zero. However, one does not know if removal of the estimated spatial effects from the measured phenotypes will improve selection accuracy because the marginal likelihood, whether penalized using the AIC or the Bayesian information criterion (BIC) or not, does not provide information about model misspecification. The value of cross-validation using genetic data as a model selection check on the prior can be seen in the first data set where the increase in the log marginal likelihood corresponded to an increase in prediction accuracy. For this data set, the AR(1), AR(2), and ARMA(1,1) spatial effects were the same, and thus the accuracies were equivalent. While using any spatial process increased the mean genomic prediction accuracy for powdery mildew resistance in the first data set, the increase was within one standard deviation from the mean genomic prediction accuracy for the model with unadjusted phenotypes, suggesting that additional data need to be analyzed to confirm the utility of the spatial correction for powdery mildew resistance.

The marginal log likelihoods achieved for the ARMA(1,1) processes for all data sets were equivalent to those achieved with an AR(1) process (Table 1). The likely reason for this is that an ARMA(1,1) process can be written in terms of a sum of an independent AR(1) process and white noise (Granger and Morris 1976). Therefore,

Table 3.1: Log marginal likelihood estimates and degrees of freedom for five spatial processes for three data sets and the mean genomic prediction accuracies for 100 cross-validation test sets, with the standard deviations in parentheses, for the first data set.

Phenotype	Spatial Process	Log ML	DF	Accuracy
2002 Powdery Mildew	None			0.449 (0.126)
2002 Powdery Mildew	AR(1)	-826.854	25.007	0.487 (0.117)
2002 Powdery Mildew	AR(2)	-826.854	25.007	0.487 (0.117)
2002 Powdery Mildew	MA(1)	-838.723	281.105	0.473 (0.119)
2002 Powdery Mildew	MA(2)	-836.719	267.647	0.474 (0.118)
2002 Powdery Mildew	ARMA(1,1)	-826.854	108.822	0.487 (0.117)
2017 Powdery Mildew	AR(1)	-451.605	297.000	
2017 Powdery Mildew	AR(2)	-450.908	75.459	
2017 Powdery Mildew	MA(1)	-451.673	146.457	
2017 Powdery Mildew	MA(2)	-451.467	114.851	
2017 Powdery Mildew	ARMA(1,1)	-451.605	111.633	
2017 Height	AR(1)	-1076.017	40.989	
2017 Height	AR(2)	-1075.012	148.771	
2017 Height	MA(1)	-1081.685	135.653	
2017 Height	MA(2)	-1076.651	82.172	
2017 Height	ARMA(1,1)	-1076.017	104.258	

there is a white noise component from the ARMA(1,1) prior and another one from the likelihood function. When the likelihood white noise captures all the white noise variance, the ARMA(1,1) process acts as an AR(1) process, explaining the equivalent log marginal likelihoods achieved by the two processes. Inclusion of white noise from the likelihood function is necessary as otherwise adjustment of the phenotypes is not possible.

In the Cornell grapevine breeding program, marker-assisted selection is performed to screen susceptible unphenotyped seedlings using QTL found in various F_1 families phenotyped for disease resistance. The number of QTL found in the first vineyard data set using either spatially unadjusted or adjusted phenotypes was three, regardless of the spatial process used to calculate the spatial effects. The QTL were in the same locations across all phenotypes, specifically on chromosomes 14 and 15 (2 QTL). This suggests that spatial variation is likely not a limitation to successful mapping of disease resistance QTL from field data.

In this study, only data from vineyards with complete F_1 families (having had no seedling selection) were analyzed. The reason for this is that Cornell grapevine breeding first stage selection vineyards are not randomized and the vines in those vineyards are discarded as it becomes apparent that a vine does not have desirable characteristics. Second stage selection vineyards are likewise not randomized. First stage selection vineyards contain unreplicated vines that have passed initial nursery screening. Second stage selection vineyards contain replicated vines that have passed first stage screening. The issue with analyzing a vineyard that is not randomized is that the spatial and genetic effects become confounded. The issue with discarding vines is that, other than the initial year of planting, large gaps in the vineyard will arise, and this will lead to worse estimates of the spatial function covariance parameters than if all, or almost all, vines were present. However, a vineyard with gradually fewer vines is less expensive to maintain, and thus there is a tradeoff between having phenotypic data for more vines, although with fewer phenotypes per year as vines get discarded, and having more precise phenotypic data for fewer vines.

From the limited set of vineyard data analyzed, we recommend an AR(1) process as a sufficient process to account for spatial variation in a vineyard. An AR(1) process was also found to capture most of the spatial variation in sorghum trials (Leiser et al. 2012). The benefit of accounting for spatial variation is that one obtains phenotypes less corrupted by spatial variation. Such adjusted phenotypes should improve genetic gain regardless of whether one uses whole-genome prediction for selection. Similarly, adjustment of phenotypes should aid in QTL mapping, although, as seen here, it might not translate to differences in the QTL used for marker-assisted selection. A limitation to the ARMA processes presented here is the assumption of stationarity. For example, the covariance function between two vines in one end of a vineyard is assumed to be the same as two in another end. However, development of more realistic priors is often precluded by the lack of information on the physiological, pedological, and climatic bases of spatial variation.

3.8 Acknowledgments

We thank Mike Colizzi and Steve Luce for vineyard maintenance and for obtaining the powdery mildew phenotypes. We thank Andrew Gordon Wilson for helpful discussions about Gaussian processes. We thank Lance Cadle-Davidson for suggesting use of the male parent in the second data set. We also thank Simon Divilov and Michael Gore for feedback on the manuscript. This research was supported by the USDA-NIFA Specialty Crop Research Initiative (Award No. 2011-51181-30635), The New York Wine & Grape Foundation, Federal Capacity Funds and the Lake Erie Regional Grape Processors Fund. Graduate Assistantship support for Konstantin Divilov was provided by the Charles R. Bullis Plant Hybridization Endowment and the Michael Nolan Endowment Fund.

3.9 References

Bernal-Vasquez, A.-M., J. Möhring, M. Schmidt, M. Schönleben, C.-C. Schön, and H.-P. Piepho (2014). “The importance of phenotypic data analysis for genomic prediction-a case study comparing different spatial models in rye". *BMC Genomics* 15 (1), 646.

Bertsekas, D. P. and J. N. Tsitsiklis (2002). *Introduction to Probability*. 2nd ed. Belmont, MA: Athena Scientific.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time Series Analysis: Forecasting and Control*. 5th ed. John Wiley & Sons.

Broman, K. W., H. Wu, Ś. Sen, and G. A. Churchill (2003). "R/qtl: QTL mapping in experimental crosses". *Bioinformatics* 19 (7), 889–890.

Covarrubias-Pazaran, G. (2016). "Genome-assisted prediction of quantitative traits using the R package *sommer*". *PLoS ONE* 11 (6), 1–15.

Endelman, J. B. (2011). "Ridge regression and other kernels for genomic selection with R package rrBLUP". *The Plant Genome* 4 (3), 250–255.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Gleeson, A. C. and B. R. Cullis (1987). "Residual maximum likelihood (REML) estimation of a neighbour model for field experiments". *Biometrics* 43 (2), 277–287.

Granger, C. W. and M. J. Morris (1976). "Time series modelling and interpretation". *Journal of the Royal Statistical Society. Series A (General)*, 246–257.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. New York, NY: Chapman and Hall.

Hyma, K. E., P. Barba, M. Wang, J. P. Londo, C. B. Acharya, S. E. Mitchell, Q. Sun, B. Reisch, and L. Cadle-Davidson (2015). "Heterozygous mapping strategy (HetMappS) for high resolution genotyping-by-sequencing markers: a case study in grapevine". *PloS ONE* 10 (8), e0134880.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin (2008). “Efficient control of population structure in model organism association mapping”. *Genetics* 178 (3), 1709–1723.

Lado, B., I. Matus, A. Rodríguez, L. Inostroza, J. Poland, F. Belzile, A. del Pozo, M. Quincke, M. Castro, and J. von Zitzewitz (2013). “Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data”. *G3: Genes, Genomes, Genetics* 3 (12), 2105–2114.

Leiser, W. L., H. F. Rattunde, H.-P. Piepho, and H. K. Parzies (2012). “Getting the most out of sorghum low-input field trials in West Africa using spatial adjustment”. *Journal of Agronomy and Crop Science* 198 (5), 349–359.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard (2001). “Prediction of total genetic value using genome-wide dense marker maps”. *Genetics* 157 (4), 1819–1829.

Morota, G. and D. Gianola (2014). “Kernel-based whole-genome prediction of complex traits: a review”. *Frontiers in Genetics* 5.

Muendler, M. (2000). *Econ 202a Problem Set 1: Suggested Solutions*. URL: `http://econweb.ucsd.edu/muendler/teach/00s/ps1-prt1.pdf`.

Mullen, K., D. Ardia, D. Gil, D. Windover, and J. Cline (2011). “DEoptim: an R Package for global optimization by differential evolution”. *Journal of Statistical Software* 40 (6), 1–26.

Murray-Smith, R. and A. Girard (2001). “Gaussian process priors with ARMA noise models”. *Irish Signals and Systems Conference*. Maynooth, Ireland, 147–152.

Petelin, D., B. Filipič, and J. Kocijan (2011). “Optimization of Gaussian process models with evolutionary algorithms”. *International Conference on Adaptive and Natural Computing Algorithms*. Ed. by A. Dobnikar, U. Lotrič, and B.

Šter. Lecture Notes in Computer Science vol. 6593. Springer Berlin Heidelberg, 420–429.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Storn, R. and K. Price (1997). “Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces”. *Journal of Global Optimization* 11 (4), 341–359.

VanRaden, P. M. (2008). “Efficient methods to compute genomic predictions”. *Journal of Dairy Science* 91 (11), 4414–4423.

CHAPTER 4

TOWARD AN ORNAMENTAL DWARF GRAPEVINE WITH AN ATTRACTIVE FLORAL SCENT

4.1 Introduction

The use of grapevine (*Vitis* spp.) is mostly limited to the production of consumable products, e.g., wine and table grapes. Vineyards near urban centers often host events, such as weddings, and the existence of continuously flowering grapevines with exceptional floral scent would increase the quality of the services these vineyards provide. Additionally, dwarf grapevine varieties for ornamental purposes can be a unique addition in a nursery catalog. Currently, 'Pixie', a dwarf grapevine derived from the L1 cell layer of the *Vitis vinifera* chimera 'Pinot Meunier', is available for viticultural and research purposes (Cousins 2012). The dwarfism is caused by mutation in the gene VvGAI1, which causes the vine to be insensitive to gibberellic acid (Boss and Thomas 2002). The mutation also causes meristematic uncommitted primordia in an actively growing vine to differentiate into mostly inflorescences, which is different than wild type vines that only produce inflorescences in latent buds.

The scent of *V. vinifera* flowers has been described as having a "beautiful floral and fruity-fresh note as well as a mignonette-like, acrid-dusty and green side-note" (Buchbauer et al. 1995). The attraction of beetles for pollination purposes to grapevine floral scent is a possible evolutionary reason for the existence of the scent (Branties 1978). Valencene, a sesquiterpene found in Valencia oranges, is the compound most frequently found to be of greatest abundance evolved from *V. vinifera* flowers and is localized within pollen grains (Barbagallo et al. 2014;

Buchbauer et al. 1995; Buchbauer et al. 1994b; Buchbauer et al. 1994a; Martin et al. 2009). However, other compounds are also present in the headspace of the flowers, and, to our knowledge, it is not yet known what combination and concentration of compounds are necessary to produce the floral scent, or bouquet, typical of *V. vinifera*. Additionally, there is a lack of research into the compounds evolved from flowers of grapevine species other than *V. vinifera* whose flowers are not hermaphroditic.

Extraction of floral compounds can be performed in a variety of ways. Two common methods are 1) extraction using a solvent in physical contact with flowers that are detached from the plant, and 2) extraction where the air, or headspace, surrounding flowers that are not detached from the plant is sampled (Buchbauer et al. 1994b; Martin et al. 2009). While the latter method produces more biologically relevant results, it is also more expensive and technical. Buchbauer et al. (1994b) used both methods to extract floral volatiles in *V. vinifera* and found that valencene was the compound with the greatest abundance regardless of the extraction method used. After extraction, gas chromatography-mass spectrometry (GC-MS) is performed to separate the extracted compounds by boiling point and to identify the compounds.

Our goal was to breed a dwarf grapevine with a floral scent similar to what is found in ‘Couderc 3309’ (*Vitis riparia* × *Vitis rupestris*). Here, we report on the progress of that effort. Additionally, we wanted to explore what floral compounds are released from some *Vitis* spp. native to North America.

4.2 Materials and Methods

In the spring of 2015, the grapevine germplasm in Geneva, NY was partially screened for floral scent intensity and aroma by sniffing of flowers by the author. Couderc 3309, a male rootstock cultivar, was selected as the male parent in crossing because of its pleasant, perfume-like aroma, which was even present around the flowers after anthesis. Pollen of the male parent was collected and stored at -20℃ in the same season. Cuttings of *V. rupestris* 187G × ‘Pixie’, a female hybrid, were rooted in a greenhouse on November 2015 and pollinated with ‘Couderc 3309’ in the winter of 2015 and spring of 2016. ‘187G’ is a female rootstock cultivar. In December 2016, 852 seed from the crosses were humidified for 24 h, soaked in 1.5% H_2O_2 for 24 h, and soaked in 5000 ppm gibberellic acid (90+% purity) for 24 h. The seed were then stratified for two months at 4℃ and then kept at 20℃ for a week prior to germination in a greenhouse. Out of the 243 seedlings that germinated, 60 were dwarf, as determined by internode length. The dwarf seedlings were planted in a nursery in Geneva, NY, along with six ‘187G’ × ‘Pixie’ vines.

In order to find the optimal length of time for a hexane extraction of floral volatiles, 10 open flowers from *V. cinerea* B9, which had a characteristic *V. vinifera* scent as determined by the author, growing in the field were placed in a glass tube with 0.5 mL of hexane and kept in the hexane at 4℃ for either 1, 3, 6, 12, or 24 h with three replications per time point (except the 3 h time point that only had two replications). The flower collection was done on 17 July 2015 at 1 pm. This method was similar to that of Martin et al. (2009). The flowers for all the time points came from the same flower cluster. After the allotted time passed, the solution was filtered through glass wool and stored at -20℃ before being sent to the Abby and Howard P. Milstein Synthetic Chemistry Core Facility at Cornell

University for GC-MS analysis with a temperature program similar to Martin et al. (2009).

Using the same flower processing methodology, a diversity panel of female and male grapevine accessions from the USDA National Germplasm Repository in Geneva, NY were sampled for their floral volatiles. The panel included *V. aestivalis* (‘Winnebago’ and Rem 30-77), *V. cinerea* (C-66-14), *V. labrusca* (Rem 26-75 and Grem-5), *V. riparia* (‘Tom’s Favorite’), and *V. rupestris* (‘Wichita Refuge’ and R-66-24). Additionally, ‘Couderc 3309’ was sampled from the repository as well as ‘Chardonnay’ and ‘Riesling’ (*V. vinifera*), both of which are hermaphroditic, from a nearby vineyard. Rem 30-77, C-66-14, R-66-24, and ‘Couderc 3309’ are male accessions while ‘Winnebago’, Rem 26-75, Grem-5, ‘Tom’s Favorite’, and ‘Wichita Refuge’ are female accessions. Recently opened flowers were sampled at 9 am at various dates in June and July of 2016 and kept in hexane for 1 hour. Due to genetic factors, flowering dates varied and so flowers could not be sampled on the same date for all accessions. The floral aroma of the sampled inflorescence for each accession was recorded by the author prior to sampling.

Analysis of the GC-MS chromatograms to find the most likely compounds in solution was performed using the Automated Mass Spectral Deconvolution and Identification System (AMDIS) (National Institute of Standards and Technology [NIST]) and the NIST 2008 Mass Spectral Library. Chromatogram plots were made using the ggplot2 (Wickham 2009) and RColorBrewer (Neuwirth 2014) packages in R. Prior to plotting, the baseline for each spectrum was removed using the MALDIquant R package (Gibb and Strimmer 2012).

4.3 Results and Discussion

Chromatograms were similar for all time points in the experiment with *V. cinerea* B9, suggesting that 1 hour is enough to extract floral volatiles from grapevine flowers (Figure 4.1). Germacrene D was by far the compound of greatest abundance in the samples, which was unexpected because, as the floral scent in *V. cinerea* B9 was similar to *V. vinifera* flowers, we expected valencene to be of greatest abundance. Running one of the samples in the GC-MS with pure valencene confirmed that the peak was not valencene. While the abundance of compounds in *V. cinerea* B9 was similar for all samples, that was not the case for the diversity panel (Figure 4.2). Specifically, 'Winnebago', C-66-14, R-66-24, and 'Couderc 3309' had much less abundance than the rest of the panel and had compounds, such as etamiphyllin and metacetamol, that were not found in the other accessions. We believe this low abundance was due to sampling flowers that had no or little pollen left, assuming that the floral volatiles arise from the pollen grains (Martin et al. 2009). However, prior to sampling, the flowers of all accessions had a scent.

Most of the compounds identified in the diversity panel and *V. cinerea* B9, e.g., germacrene D, valencene, δ-cadinene, β-selinene, α-bergamotene, levomenol, viridiflorene, are known sesquiterpenes. Sesquiterpenes share the common precursor farnesyl diphosphate (Caspi et al. 2008; Kanehisa and Goto 2000). Similar to *V. cinerea* B9, Rem 30-77 produced a high level of germacrene D without valencene. Rem 30-77 had a mild earthy, perfume-like scent unlike that of *V. cinerea* B9. Both *V. labrusca* accessions, which produced germacrene D in high quantities, also had a peak determined to be valencene. Like *V. cinerea* B9, these accessions had a scent similar to *V. vinifera.* 'Riesling' had a chromatogram that we expected, i.e., it had a high abundance of valencene. Interestingly, 'Chardonnay' had nei-

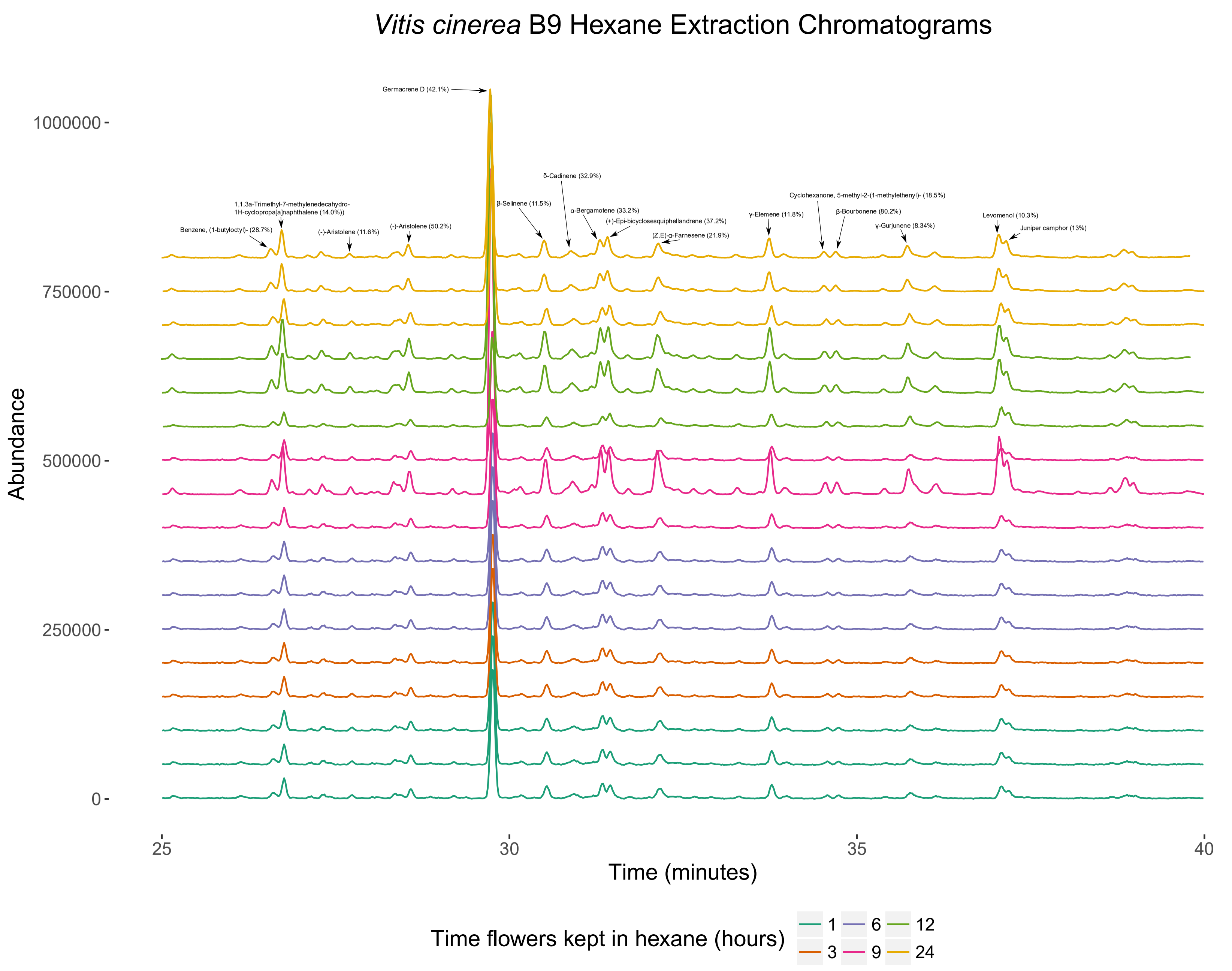

Figure 4.1: GC-MS chromatograms from *V. cinerea* B9 flowers in a hexane solvent. The most likely compounds of peaks are given along with the probability of the matching in parentheses. The y-axis is not specific to a particular spectrum and each spectrum is separated from adjacent ones by an abundance value of 50000.

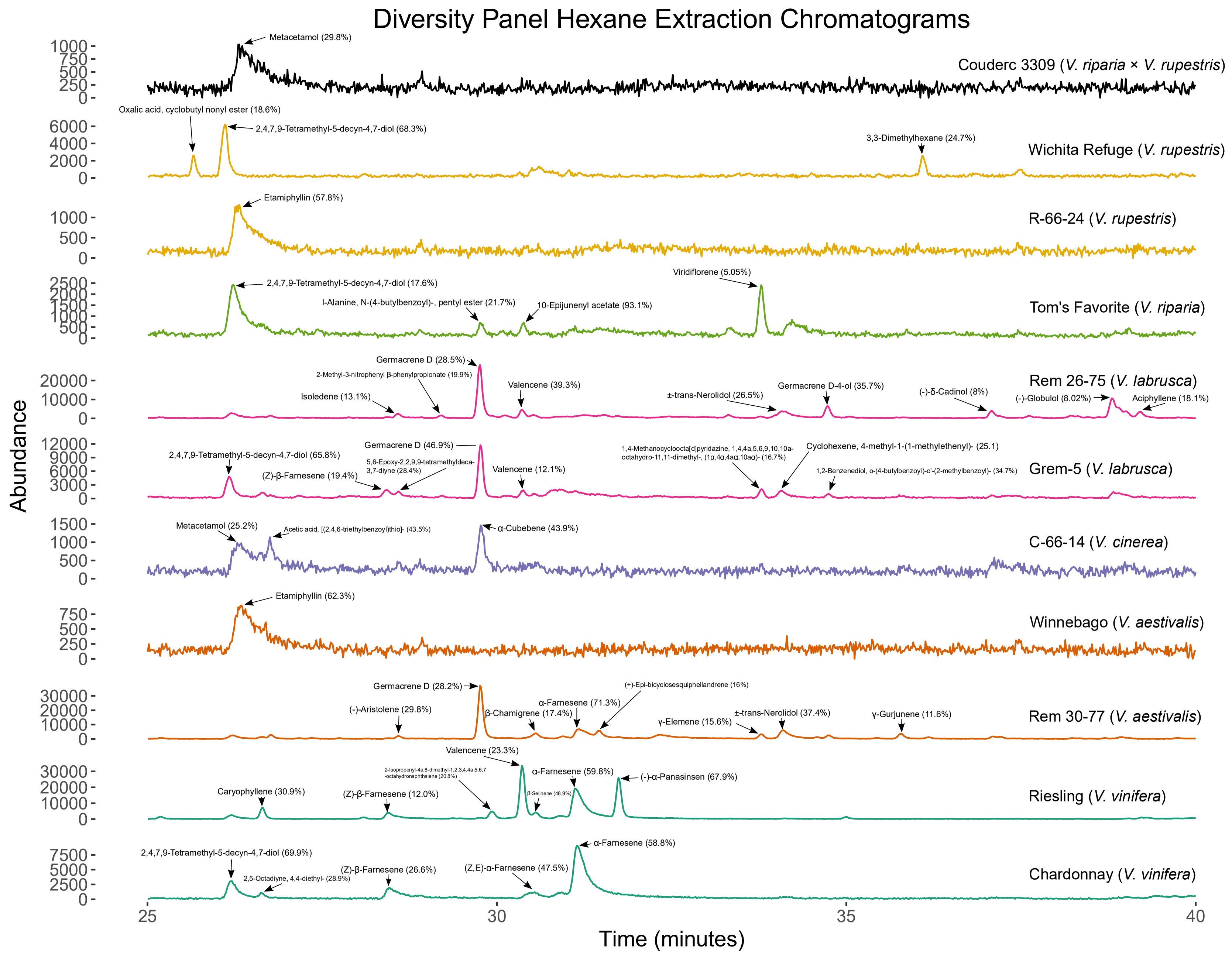

Figure 4.2: GC-MS chromatograms from the diversity panel flowers in a hexane solvent. The most likely compounds of peaks are given along with the probability of the matching in parentheses. The different colors of the spectra represent different *Vitis* spp.

ther a valencene nor germacrene D peak. For ‘Chardonnay’, α-farnesene was the compound of greatest abundance. These results suggest that the aroma common to *V. vinifera* is not only attributable to valencene.

‘Wichita Refuge’ and ‘Tom’s Favorite’, both of which did not have many sesquiterpene peaks, did not have similar floral scents. ‘Wichita Refuge’ had a mild perfume-like scent while ‘Tom’s Favorite’ had a scent similar to *V. vinifera*. This suggests that the search for the compounds that cause the flower bouquet in grapevine should be expanded to compounds outside the sesquiterpenes. Of the accessions with particularly low abundance values, ‘Winnebago’ and ‘Couderc 3309’ had a perfume-like scent while C-66-14 and R-66-24 had a *V. vinifera*-like scent.

The methodology used for the floral volatile study was very qualitative and should not be taken to be definitive. First, only one date was sampled for each accession and at only one time of the day. It is known that volatile emission in grapevine flucuates during a day (Martin et al. 2009), and the composition of the emission might vary as well. Second, the hexane extraction might not be a good representation of the true floral emission composition. Validating the results with headspace analysis would make them more biologically meaningful. Lastly, the floral aroma was only rated by the author. Rating of aroma by a diverse group of people would make the chromatogram to aroma phenotype conclusions stronger.

Out of the 60 (‘187G’ × ‘Pixie’) × ‘Couderc 3309’ seedlings planted in the nursery, 35 were vigorous in the field and were transplanted to the greenhouse in the fall of 2017 for floral scent evaluation. Early in the season, all the seedlings in the field initially produced inflorescences that looked like tendrils except with very few flowers at the ends (Figure 4.3). Later in the season, some seedlings produced inflorescences with a greater number of flowers (Figure 4.4). From two seedlings with enough flowers, we found that the floral scent was not *V. vinifera*-like, but rather perfume-like, suggesting that segregation might occur in the greenhouse screening.

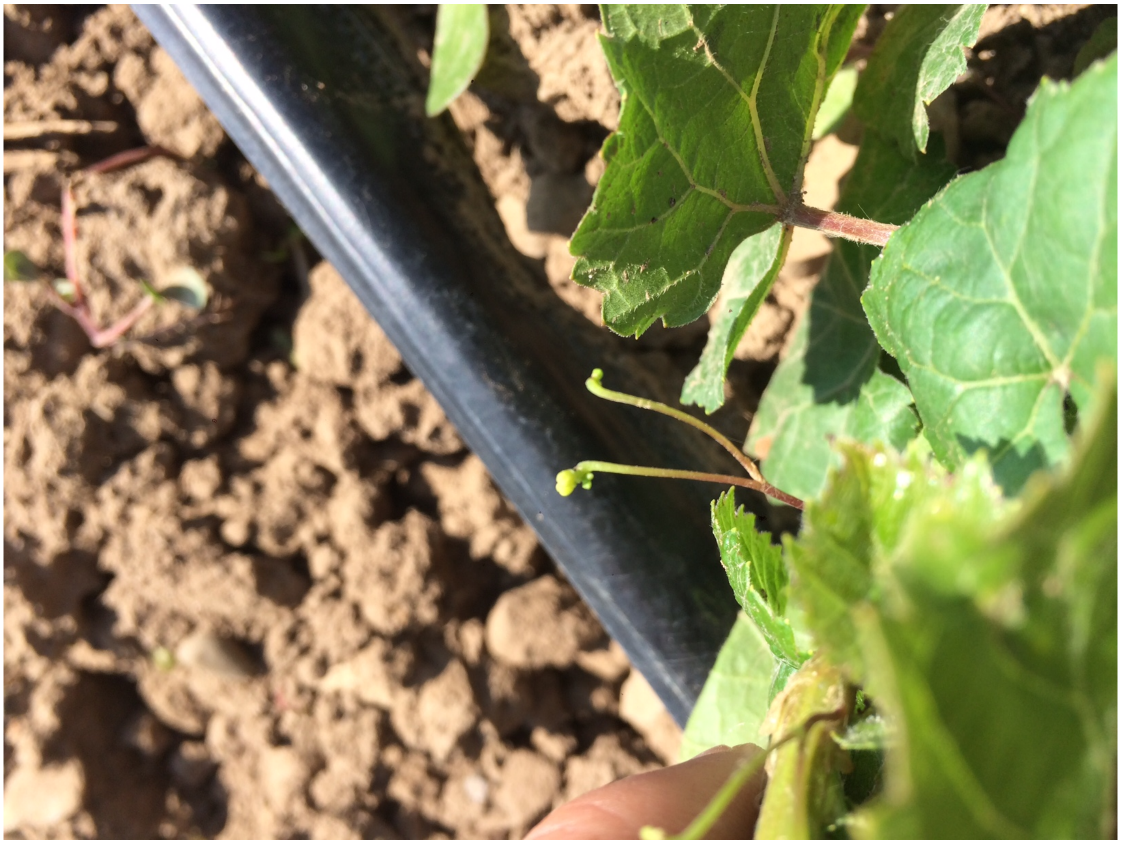

Figure 4.3: Flowers of a (‘187G’ × ‘Pixie’) × ‘Couderc 3309’ seedling on 14 August 2017.

4.4 Acknowledgments

We thank Robert Raguso for helpful discussions about floral scent phenotyping, Debra Johnston for providing the rooted ‘187G’ × ‘Pixie’ cuttings, and Peter Cousins for advice on seed germination. We thank Mike Colizzi for managing the nursery. We also thank J. David Warren for helping with the GC-MS analysis

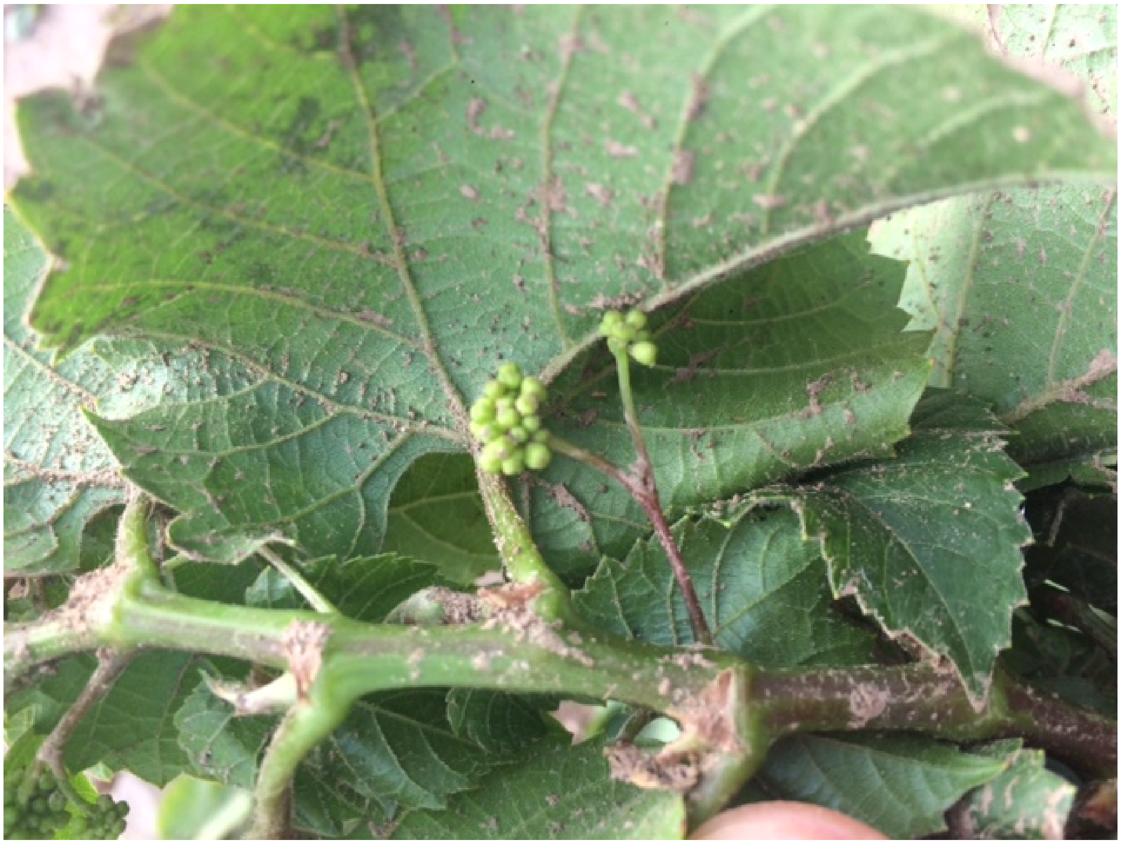

Figure 4.4: Flowers of a (‘187G’ × ‘Pixie’) × ‘Couderc 3309’ seedling on 13 September 2017.

and C. Thomas Chao for making the grapevine germplasm available for our use. This research was supported by the USDA-NIFA Specialty Crop Research Initiative (Award No. 2011-51181-30635), The New York Wine & Grape Foundation, Federal Capacity Funds and the Lake Erie Regional Grape Processor’s Fund. Graduate Assistantship support for Konstantin Divilov was provided by the Charles R. Bullis Plant Hybridization Endowment and the Michael Nolan Endowment Fund.

4.5 References

Barbagallo, M. G., A. Pisciotta, and F. Saiano (2014). “Identification of aroma compounds of *Vitis vinifera* L. flowers by SPME GC-MS analysis”. *Vitis* 53 (2), 111–113.

Boss, P. K. and M. R. Thomas (2002). “Association of dwarfism and floral induction with a grape ‘green revolution’ mutation”. *Nature* 416, 847–850.

Branties, N. B. M. (1978). “Pollinator attraction of *Vitis vinifera* subsp. silvestris”. *Vitis* 17, 229–233.

Buchbauer, G., L. Jirovetz, M. Wasicky, A. Herlitschka, and A. Nikiforov (1994a). “Aroma von Weissweinblüten: Korrelation sensorischer Daten mit Headspace-Inhaltsstoffen”. *Zeitschrift für Lebensmittel-Untersuchung und Forschung* 199 (1), 1–4.

Buchbauer, G., L. Jirovetz, M. Wasicky, and A. Nikiforov (1994b). “Headspace analysis of *Vitis vinifera* (Vitaceae) flowers”. *Journal of Essential Oil Research* 6 (3), 311–314.

Buchbauer, G., L. Jirovetz, M. Wasicky, and A. Nikiforov (1995). “Aroma von Rotweinblüten: Korrelation sensorischer Daten mit Headspace-Inhaltsstoffen”. *Zeitschrift für Lebensmittel-Untersuchung und Forschung* 200 (6), 443–446.

Caspi, R., H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang, and P. D. Karp (2008). “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases”. *Nucleic Acids Research* 36, D623–D631.

Cousins, P. (2012). “Small but mighty: ‘Pixie’ grapevine speeds up the pace of grape genetics research and breeding”. *Appellation Cornell* (10).

Gibb, S. and K. Strimmer (2012). “MALDIquant: a versatile R package for the analysis of mass spectrometry data”. *Bioinformatics* 28 (17), 2270–2271.

Kanehisa, M. and S. Goto (2000). “KEGG: Kyoto encyclopedia of genes and genomes”. *Nucleic Acids Research* 28 (1), 27–30.

Martin, D. M., O. Toub, A. Chiang, B. C. Lo, S. Ohse, S. T. Lund, and J. Bohlmann (2009). “The bouquet of grapevine (*Vitis vinifera* L. cv. Cabernet Sauvignon) flowers arises from the biosynthesis of sesquiterpene volatiles in pollen grains”. *Proceedings of the National Academy of Sciences* 106 (17), 7245–7250.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes.* R package version 1.1-2. URL: `https://CRAN.R-project.org/package=RColorBrewer`.

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

CHAPTER 5
CONCLUSION AND FUTURE DIRECTIONS

This dissertation explored grapevine breeding through concepts from computer vision, graphical models, time series analysis, and traditional breeding methodology. Adapting methodology from annual crops to grapevines can be difficult due to the perenniality and intensive management of grapevines. Here, we discuss some limitations of the current work and future directions of disease resistance and floral scent research in grapevines.

The computer vision system developed can only detect white pixels, which correspond to sporulation or leaf trichomes, but can be modified, with some effort, to detect other colors. One can generalize the detection of disease by modeling the pixels in a leaf disc as coming from a mixture of Gaussian distributions. This model might also be used to remove the parameter search for thresholding the background in the images. Most of the time spent using the computer vision system was on finding a set of parameters for the Hough circle transform algorithm that detected all the leaf discs as circles in a particular experiment's image set. Recently, an algorithm called Ellipse and Line Segment Detector, with Continuous validation (ELSDc) (Pătrăucean et al. 2017) was developed that may be able to remove the need for this parameter search.

A limitation of our execution of the leaf disc assay was the time and human resources spent setting up an experiment from harvest of the leaves (4 hours, 3 people) to punching out the leaf discs and plating them on agar (6 hours, 3 people). It would be very valuable to see how much of the work in the phenotyping procedure was unnecessary. For example, how would the ratings of each genotype compare to our experimental results if only one leaf disc per genotype was obtained in the field without any surface sterilization of leaves? Using a smaller amount of resources per experiment can be used to address other limitations in our experiments, such as the use of only one isolate of *Plasmopara viticola*. In addition to the time constraints during the growing season, there is a human constraint to keeping multiple isolates of the oomycete due to its obligate biotrophic nature, which means that all the isolates must be cultured on fresh leaves every week. Moreover, the oomycete is constantly evolving, so it is not straightforward when one should change the isolates used to evaluate genotypes, except for when resistance is seen as deteriorating, which can take years to determine since the deterioration might be a genotype-by-environment (G×E) effect.

We found evidence of G×E effects in our experiments, which was not expected because the leaf discs were inoculated in the laboratory and kept in controlled conditions. Thus, a more detailed G×E study is needed. Planting an F_1 family in multiple locations is usually not affordable because each vineyard requires a substantial investment to build and maintain. A feasible alternative would be to develop a diverse set of dwarf grapevines, as we have done in the floral scent project, and place these vines in pots with different soil types and in Personal Food Computers (Ferrer et al. 2017), which are environment-controlled devices, programmed to replicate growing conditions in various years and locations. This concept is similar to the Tree Computer (OpenAg 2017) and would possibly allow tractable evaluation of G×E effects on downy mildew resistance.

Another possibility to explore G×E effects would be to evaluate grapevine data from multiple F_1 families from various grapevine breeding programs that already are collaborating as part of VitisGen (www.vitisgen.org). One substantial hurdle to overcome would be the standardization of vineyard maintenance, e.g., training, pruning, pesticide application, as much as possible and good documentation of how each vineyard was treated. With the availability of phenotypes, which would also be standardized, and soil and weather data, hydrological and/or geochemical processes, in addition to genetic data, can be integrated in modeling the phenotypes in order to determine if such processes improve the likelihood of the model than using genetic data only. Such data collection is similar to what is being performed with the Genomes To Fields Initiative (www.genomes2fields.org). This integration would help in determining where to deploy QTL found in the future multi-state evaluations. Additionally, this data can be used to explore the importance of genotype-by-management and genotype-by-environment-by-management effects for disease resistance. These effects can be important if it is found, for example, that certain training systems in certain environments lead to greater disease susceptibility of certain genotypes.

Successful implementation of long-term marker-assisted selection of the downy mildew resistance QTL found in our work is complicated by the fact that the QTL have small effect sizes. For large-effect NBS-LRR loci, e.g., the *Rpv1* locus, empirical breeding data from the Cornell grapevine breeding program show that they produce an effect on downy mildew disease resistance regardless of other loci in a vine. For the small effect QTL, the simplest assumption is that all the QTL will have a small additive effect after introgression, regardless of other loci in a vine. To test this assumption, one can test all possible combinations of the QTL in various genetic backgrounds. To avoid this time-consuming and expensive experiment, if one knew the genes behind the QTL and their corresponding mechanistic pathways, one can determine the likelihood of a QTL having an effect in a vine given the sequences of the genes in the pathway as one can estimate the likelihood of the resulting protein being non-functional. However, as seen by our very small

RNA-Seq experiment, many of the genes in the *Vitis vinifera* reference genome that possibly affect disease resistance currently have unknown functionality, which makes pathway reconstruction difficult. The best approach available to us might simply be to perform truncation selection during marker-assisted selection, keeping, for example, 20% of the seedlings with the greatest amount of resistance QTL and evaluating the combinations each year in the nursery as well as in later stages of the breeding program.

We emphasize the importance of prior knowledge, for example, in selecting QTL or integrating soil and weather data into hydrological and/or geochemical processes, because of the small sample sizes common in grapevine breeding. If large population sizes were commonplace, we could, for example, conduct the combinatorial experiment stated above. If large numbers of tested environments were common, it might be possible to estimate the underlying G×E effects directly from soil and weather data without biophysical priors. Thus, we believe grapevine breeding will need to integrate more knowledge from diverse disciplines to generate superior varieties in a shorter amount of time.

An important limitation in our breeding of dwarf grapevines for floral scent was that the phenotyping was limited to a single person. Phenotyping germplasm in the field with a large group of people is difficult because of the dynamic nature of grapevine floral scent and human olfactory fatigue. Floral scent in a grapevine accession might change within a day or between days. Additionally, not all grapevine accessions flower at the same time. Olfactory fatigue limits the number of accessions that can be tested per person per day. All these factors, as well as the difficulty of obtaining a large group of people trained to evaluate scent in a standardized manner, make floral scent phenotyping difficult in the field. However, dwarf grapevines flower constantly in the greenhouse, making phenotyping the F_1 progeny from our cross manageable as there are fewer time constraints. Therefore, the cross progeny can be used to evaluate floral scent phenotyping methods as well as how floral scent changes with respect to inflorescence phenology. While evaluations in a greenhouse would eliminate the possibility of evaluating many G×E effects, partly breeding for home ornamental purposes makes this limitation less important as homes are kept in relatively constant environmental conditions, although breeding for environment-dependent floral scent would be an interesting future direction.

5.1 References

Ferrer, E. C., J. Rye, G. Brander, T. Savas, D. Chambers, H. England, and C. Harper (2017). "Personal Food Computer: A new device for controlled-environment agriculture". *arXiv preprint arXiv:1706.05104*.

OpenAg (2017). *Tree Computer*. URL: `https://www.media.mit.edu/projects/tree-computer/overview/`.

Pătrăucean, V., P. Gurdjos, and R. G. von Gioi (2017). "Joint A Contrario Ellipse and Line Detection". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (4), 788–802.