THE WILD TOMATO CLADE OFFERS INSIGHTS INTO FLESHY FRUIT TRAIT EVOLUTION AT THE PHENOTYPIC AND MOLECULAR LEVELS

Jacob R. Barnett
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2

Part of the Evolution Commons

Recommended Citation

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
THE WILD TOMATO CLADE OFFERS INSIGHTS INTO FLESHY FRUIT TRAIT EVOLUTION AT THE PHENOTYPIC AND MOLECULAR LEVELS

A Dissertation Presented

by

JACOB R. BARNETT

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements of the degree of

DOCTOR OF PHILOSOPHY

February 2024

Organismic and Evolutionary Biology
THE WILD TOMATO CLADE OFFERS INSIGHTS INTO FLESHY FRUIT TRAIT EVOLUTION AT THE PHENOTYPIC AND MOLECULAR LEVELS

A Dissertation Presented

by

JACOB R. BARNETT

Approved as to style and content by:

__________________________________________
Ana L. Caicedo, Chair

__________________________________________
Lynn S. Adler, Member

__________________________________________
John G. Gibbons, Member

__________________________________________
Jason M. Kamilar, Member

Duncan Irschick, Graduate Program Director
Organismic and Evolutionary Biology
DEDICATION

To my family—especially David, Eileen, and Marie Barnett—for their steadfast support and encouragement, including some clutch assistance with pruning tomato plants that truly grew like wild. Even when I had my doubts, they always believed I could do hard things and accomplish this goal.
ACKNOWLEDGEMENTS

I am grateful to the many people who have helped and supported me along the way during this dissertation project. I couldn't possibly provide an exhaustive list of them here, but I will endeavor to mention some who are top of mind—apologies to anyone who didn't make this particular list!

First off I would like to thank the wonderful community of students, staff, and faculty at UMass Amherst for making this school such an excellent place to study and work. I will always treasure the time I spent here and feel privileged to have met many great friends and colleagues in the IDGP (OEB, PB, NSB and MCB Graduate Programs—especially my OEB cohort), Biology Department, Environmental Conservation Department, College of Natural Sciences, Kinesiology Department, Office of Professional Development, and the broader UMass community.

I would like to thank my advisor, Ana Caicedo, for welcoming me into her thriving lab community and providing constant support throughout my time at UMass. I truly appreciate all the opportunities she made possible for me, including growing tomatoes in greenhouses and on a farm, collaborating with top-notch scientists in Florida, working in a well-resourced molecular biology lab, and attending conferences in Rhode Island and Greece. She has helped me grow as a scientist and learn more than I ever thought I could.

I would also like to thank the other members of my committee, Lynn Adler, John Gibbons, and Jason Kamilar. When we first met as a group it felt like I had assembled an all-star team, with each person bringing a different area of expertise. It has been a pleasure to work with all of them, and I appreciate how they are always willing to help.
They each provided key support at different points in my project, whether it was help with phylogenetic comparative methods, statistics, next-generation sequence assembly and analysis, or manuscript revisions.

Denise Tieman at the University of Florida has been a fantastic collaborator, and I am truly thankful for her generosity, expertise on tomato flavor volatiles, and encouragement. Denise and her lab made the second chapter of my dissertation possible, running my wild tomato fruit samples on their GC-MS equipment and teaching me how to analyze the data, all during a pandemic! She also provided invaluable advice on sugar/acid assay protocols.

For advice and assistance growing tomato plants I am grateful to the UMass Vegetable Extension's Katie Campbell-Nelson, Sue Scheufele, and Genevieve Higgins; UMass Greenhouse staff Chris Joyner, Dave O'Neil, Chris Phillips, and Dan Jones; and UMass South Deerfield Research Farm staff Bob Skalbite, Neal Woodard, and Zach Zenke.

A huge thank you also goes to all the members of the Caicedo Lab, past and present, for sharing their expertise on everything from plant care to DNA extraction to bioinformatics, and for making the lab such an enjoyable place to work. Undergraduates Gina Buonauro and Rovin Sharma, as well as graduate student Maryam Rashidzade, did a tremendous amount of lab work and image analysis that were central to my project. I am also thankful for help from graduate student Xiang Li, postdoc Hamid Razifard, lab manager Sherin Perera, and undergraduates Jacob Scott, Carter Hissam, Dan Lowey, Jessica Lessard, Fabbiha Hossain, and Amanda Semple.
Others who helped at various stages of my project include: Ravi Ranjan in the UMass Genomics Resource Laboratory for performing the BluePippin size selection during the GBS sequencing library prep; the UMass Quantitative Sciences Group in the ECO Department and OEB student João Menezes for statistics advice; Norman Johnson at UMass for evolutionary biology expertise and fascinating conversations; Roger Chetelat at UC Davis for the suggestion to search the Tomato Genetics Resource Center collection notes database for mentions of animal dispersers; Charles Goulet at the University of Laval and OEB student Seanne Clemente for helpful suggestions with volatiles data analysis; and Jeremy Andersen in the ECO Department for helpful advice and encouragement with molecular selection analyses.

This research was supported by funding from the UMass Natural History Collections Grant to Jacob Barnett in 2020, UMass Graduate School Dissertation Research Grant to Jacob Barnett in 2020, Lotta Crabtree Fellowship in Production Agriculture to Jacob Barnett in 2019-2022, and National Science Foundation Grant #1564366 to Ana Caicedo.

Finally, I would not have completed this project without the support, encouragement, and love from friends and family, especially my parents, sister, and grandmother—they have provided much joy over these past few years and I appreciate them more than words can express.
Bite into a beautiful farm fresh bright red tomato on a sunny summer day, and that sublime flavor emanates from a complex blend of sugars, acids, and volatile compounds. Biologists have long been fascinated by nature's rich diversity of enticing fleshy fruits such as tomatoes, yet questions remain as to how this variety has evolved. Interestingly, certain suites of traits tend to covary together (such as bright color, small size, and high sugar content). According to the dispersal syndrome hypothesis, flowering plants have improved their reproductive success by producing fleshy fruits with appealing combinations of traits that attract animal dispersers (including humans) to spread their seeds. However, animal preferences may not be the only selective pressure driving fruit trait diversity—conflicting forces include damage-inflicting seed predators and pathogens, abiotic habitat conditions, or constraints stemming from non-adaptive mechanical, chemical, developmental, or phylogenetic limitations.
While a number of recent studies have advanced our understanding of the drivers of fruit trait evolution, few studies have examined the early stages of fleshy fruit evolution across an entire clade of recently diverged plant species. The tomato clade, a group of 14 congeneric species (Solanum sect. Lycopersicon) that includes the cultivated tomato (Solanum lycopersicum var. lycopersicum) and 13 species of wild relatives, presents an exciting opportunity to fill this gap in our understanding of fruit trait divergence. Furthermore, the findings are applicable to crop improvement efforts in the economically important tomato because the wild species can be interbred with cultivated varieties. The system has been well studied as a model of fruit development and genetics, providing a stronger knowledge of the molecular underpinnings of traits than is possible with most other fruit systems. For example, the vacuolar invertase gene TIV1 encodes an enzyme that breaks down sucrose into its components glucose and fructose and has been shown to underlie interspecies variation in ripe fruit sugar composition. Wild tomato species have adapted to a range of environments and are known to display multiple fruit colors, sugar profiles, and aromas that are quite different from the cultivated tomato. However, the extent of fruit trait and genetic diversity across all 13 wild species and the phylogenetic patterns in this variation have not previously been assessed.

This dissertation project presents a novel approach and scale to fruit evolution studies using the tomato clade. Specific aims are to: 1) examine the extent and pattern of variation in disperser-relevant ripe fruit phenotypes across wild tomato species and assess whether these traits vary together in syndromes, 2) quantify variation in ripe fruit volatile aroma compounds and their biochemical pathways across the clade as well as whether scent could communicate an honest signal of nutrient content to animal dispersers, and 3)
explore patterns of molecular evolution in a set of five genes known to affect different aspects of ripe fruit sugar content and composition in wild tomatoes. In addition to further developing the wild tomato system as a model for fleshy-fruited plant evolution, this research is applicable to plant breeders working on crop improvement because publicly available germplasm was used to quantify variations in flavor-relevant fruit traits and genes. These findings also have value beyond tomatoes since many other crop plants such as melons and blueberries share similar biochemical and developmental pathways.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>v</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xiv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xvi</td>
</tr>
</tbody>
</table>

## CHAPTER

1. EVIDENCE OF FRUIT SYNDROMES IN THE RECENTLY DIVERGED WILD TOMATO CLADE OPENS NEW POSSIBILITIES FOR THE STUDY OF FLESHY FRUIT EVOLUTION .............................................. 1

1.1 Abstract ........................................................................................................... 1
1.2 Introduction ..................................................................................................... 2
1.3 Methods ........................................................................................................... 6
    1.3.1 Plant material .................................................................................... 6
    1.3.2 Fruit phenotyping ............................................................................. 7
    1.3.3 DNA extraction, Genotyping-by-sequencing, and phylogenetic tree construction ................................................................................... 9
    1.3.4 Trait cluster analysis ......................................................................... 9
    1.3.5 Correlation analyses for pairs of traits ............................................ 10
    1.3.6 Abiotic variables analyses .................................................................. 10

1.4 Results ........................................................................................................... 11
    1.4.1 Extent of variation in individual fruit traits ................................... 11
    1.4.2 Trait variation relative to our GBS phylogeny ................................ 12
    1.4.3 Trait clustering ................................................................................... 14
    1.4.4 Correlations among pairs of traits .................................................. 16
    1.4.5 Abiotic factors associated with fruit traits ....................................... 18

1.5 Discussion ..................................................................................................... 19
1.6 Figures ........................................................................................................... 26

2. VARIATION IN RIPE FRUIT VOLATILES ACROSS THE WILD TOMATO CLADE: AN EVOLUTIONARY FRAMEWORK FOR STUDYING FRUIT SCENT DIVERSITY IN A CROP WILD RELATIVE ........................................... 33
2.1 Abstract ......................................................................................................... 33
2.2 Introduction ................................................................................................... 34
2.3 Methods ......................................................................................................... 39
  2.3.1 Study species and plant material ............................................................. 39
  2.3.2 Fruit volatile collection ........................................................................ 40
  2.3.3 GC-MS quantification of VOCs .............................................................. 40
  2.3.4 Data analyses ......................................................................................... 42
2.4 Results ........................................................................................................... 43
  2.4.1 Extent of variation in VOCs across the tomato clade ......................... 43
  2.4.2 Compounds and biosynthetic pathways differentiating species and the extent of phylogenetic conservatism ........................................... 45
  2.4.3 Associations between VOCs and putative dispersal syndromes ......... 47
  2.4.4 Associations between VOCs and fruit sugar content ........................ 51
2.5 Discussion ..................................................................................................... 53
2.6 Conclusions ................................................................................................... 59
2.7 Figures ........................................................................................................... 59

3: MOLECULAR EVOLUTION OF GENES UNDERLYING TOMATO RIPE FRUIT SUGAR TRAITS ACROSS 13 SPECIES OF WILD TOMATO RELATIVES. 67

3.1 Abstract ......................................................................................................... 67
3.2 Introduction .................................................................................................. 68
3.3 Results .......................................................................................................... 73
  3.3.1 Literature survey of genes known to affect tomato fruit sugar traits and the choice of five target genes .......................................................... 73
  3.3.2 Cladewide DNA sequence variation in the target genes: phylogenetic patterns and signs of selection .............................................................. 79
  3.3.2.1 Overall phylogenetic patterns in the target gene sequences ......... 81
  3.3.2.2 AGPL1: amino acid substitutions, signs of selection, and potential associations with the TSS fruit trait ......................................................... 83
  3.3.2.3 SUSI: amino acid substitutions, signs of selection, and potential associations with the TSS fruit trait ............................................................ 86
  3.3.2.4 SWEET1a: amino acid substitutions, signs of selection, and potential associations with the fructose-to-glucose ratio trait ....... 87
  3.3.2.5 TIV1: amino acid substitutions, signs of selection, and potential associations with the sucrose-to-hexose ratio fruit trait .......... 90
3.4 Discussion ..................................................................................................... 93
3.5 Methods ........................................................................................................ 97
LIST OF TABLES

Table 3.1 The five tomato fruit sugar-related genes chosen in this study for amplicon sequencing and molecular evolution analyses ................................................................. 110

Table 3.2 Nucleotide diversity in the new sequences obtained for the target genes ..... 111

Table 3.3 Protein sequence comparisons among the four genes for which full coding sequences were obtained in this study................................................................. 112

Table 3.4 Cladewide tests of selection on coding region (exons only) of each gene ... 113

Table 3.5 Comparison of substitution rates between coding region (exons only) and noncoding region (5' UTR and promoter) for pairs of accessions with differences in fruit sugar traits .......................................................................................... 114

Table 3.6 HKA test for signs of selection on 5' promoter region of target genes for colored-fruited species relative to green-fruited species ...................................... 116

*Note that the Appendix Tables listed below are found in separate Excel spreadsheets

Table A1 Data table of identifying information, collection locations and historical climate data for the 38 wild tomato accessions used in this study

Table A2 Data table of individual fruit values for the 21 morphology, color, and nutrition phenotypic traits measured in this study

Table A3 Qualitative notes on how ripening was determined for each accession used in this study

Table A4 Data table of accession mean values for the 21 phenotypic traits

Table A5 Historical climate data correlation coefficients

Table A6 Summary statistics for individual fruit trait values from Table A2, including phylogenetic signal

Table A7 Factor analysis for non-phylogenetic principal component analysis (PCA) of 21 morphology, color, and nutrition traits for individual fruit trait values

Table A8 Correlation coefficients for fruit trait phylogenetically independent contrasts (PICs) for the accession mean fruit trait values from Table A4

Table A9 Summary of fruit trait correlations highlighted in the Results section
Table A10 Correlation coefficients for selected climate and fruit trait variables

Table A11 Details of phylogenetic generalized least squares (PGLS) models with climate variables as predictors and one fruit trait as response variable

Table A12 Wild tomato fruit animal disperser evidence from Tomato Genetics Resource Center (TGRC) collection notes

Table B1 Accession IDs, germplasm collection locations, and VOC averages

Table B2 Notes on how fruit ripeness was determined

Table B3 VOC data for each of the 66 fruit samples

Table B4 Summary statistics, biochemical categories, and odor descriptors for VOCs

Table B5 Compounds considered key components of cultivated tomato aroma

Table B6 Random forest mean decrease accuracy for VOCs differentiating species

Table B7 Tasting notes on wild tomato fruits

Table C1 Gene IDs and genomic position for the five target genes

Table C2 Previously documented variation in the five target genes

Table C3 Wild tomato accessions sequenced in this study

Table C4 Primers used for amplicon sequencing

Table C5 PCR conditions used for amplicon sequencing

Table C6 AGPL1 protein sequence alignment showing amino acid substitutions

Table C7 SUSI protein sequence alignment showing amino acid substitutions

Table C8 SWEET1a protein sequence alignment showing amino acid substitutions

Table C9 TIV1 protein sequence alignment showing amino acid substitutions

Table C10 Results of SIFT analysis of tolerable/intolerable amino acid substitutions
<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1 Overview of the tomato clade (<em>Solanum</em> sect. <em>Lycopersicon</em>) and the 38 accessions used in this study.</td>
<td>27</td>
</tr>
<tr>
<td>Figure 1.2 Phenotypic diversity in ripe wild tomato fruits.</td>
<td>29</td>
</tr>
<tr>
<td>Figure 1.3 Wild tomato fruit traits cluster into two putative syndromes mainly defined by color, sugar type, and malic acid concentration, as shown by a non-phylogenetic principal components analysis (PCA) of 21 phenotypic traits for 208 fruits representing 38 accessions.</td>
<td>30</td>
</tr>
<tr>
<td>Figure 1.4 To assess the strength of the two fruit syndromes suggested by principal components analysis (PCA) (Fig. 3), these boxplots test differences between colored-fruited and green-fruited species groups for the nutrition and morphology traits that were the strongest differentiators in the PCA.</td>
<td>31</td>
</tr>
<tr>
<td>Figure 1.5 Temperature annual average and fruit color showed the strongest association between an historical climate variable and a syndrome-related fruit trait, leading us to hypothesize that the evolution of colored fruits may be related to warmer environments.</td>
<td>32</td>
</tr>
<tr>
<td>Figure 2.1 Phylogenetic relationships among species in the wild tomato clade (<em>Solanum</em> sect. <em>Lycopersicon</em>), with <em>Solanum lycopersicoides</em> included as the outgroup.</td>
<td>60</td>
</tr>
<tr>
<td>Figure 2.2 Wild tomato species are largely differentiated by their volatile organic compound (VOC) profiles, as shown by a principle components analysis (PCA) of centered and scaled raw values (in ng/gfw/hr) for all 66 compounds.</td>
<td>61</td>
</tr>
<tr>
<td>Figure 2.3 Heatmap of the 17 compounds with the greatest contribution to differentiating species.</td>
<td>62</td>
</tr>
<tr>
<td>Figure 2.4 Biosynthetic pathways drive differences in volatile organic compounds (VOCs) across the wild tomato clade, as shown by stacked bar plots of compound categories.</td>
<td>63</td>
</tr>
<tr>
<td>Figure 2.5 Boxplots showing differences between colored-fruited and green-fruited species groups in the amount of ester volatile organic compounds (VOCs).</td>
<td>64</td>
</tr>
<tr>
<td>Figure 2.6 Heatmap of the 15 compounds we quantified that are on the list of the 21 most important to cultivated tomato flavor according to Martina et al. (2021).</td>
<td>65</td>
</tr>
<tr>
<td>Figure 2.7 Association between total sugar concentration (glucose + fructose + sucrose, in mg/mL) and ester levels.</td>
<td>66</td>
</tr>
</tbody>
</table>
Figure 3.1 Differences in three ecologically relevant ripe fruit sugar traits among the 13 species of the tomato clade. ................................................................. 105

Figure 3.2 Maximum likelihood consensus gene trees for three regions of the target genes. ........................................................................................................... 109

Figure A1 Trait variation with respect to phylogeny for all 21 fruit traits. .......... 119

Figure A2 Separate principal component analysis (PCA) plots (non-phylogenetic) of morphology, color, and nutrition phenotypic traits for 208 fruits representing 38 accessions. ........................................................................................................ 131

Figure A3 Plots of the three strongest fruit-climate associations according to both phylogenetic generalized least squares (PGLS) models and non-phylogenetic Pearson correlation tests. ........................................................................... 133

Figure A4 Differences between colored-fruited and green-fruited species groups in total sugar concentration. ...................................................................................... 134

Figure B1 Map of collection locations for 38 wild accessions initially planted for this study. ........................................................................................................... 140

Figure B2 PCA of centered and scaled raw values (in ng/gfw/hr) for 66 compounds .141

Figure B3 Heatmaps of compounds grouped by biosynthetic pathway. ............... 142

Figure B4 Differences among 14 species in the level of cis-3-hexenal. ................. 146

Figure B5 Association between total sugar concentration (glucose + fructose + sucrose, in mg/mL) and ester levels. ................................................................. 147

Figure B6 Association between total sugar concentration (glucose + fructose + sucrose, in mg/mL) and terpene levels. ................................................................. 148

Figure B7 Association between total sugar concentration (glucose + fructose + sucrose, in mg/mL) and apocarotenoid levels. ................................................................. 149

Figure C1 Boxplots by species of three ecologically relevant ripe fruit sugar traits ..... 150

Figure C2 Boxplots of eight ripe fruit sugar traits for 38 accessions ....................... 151

Figure C3 Gene expression of five target genes at different fruit stages ............... 156
CHAPTER 1

EVIDENCE OF FRUIT SYNDROMES IN THE RECENTLY DIVERGED WILD TOMATO CLADE OPENS NEW POSSIBILITIES FOR THE STUDY OF FLESHY FRUIT EVOLUTION

1.1 Abstract

Fleshy fruits provide humans with many flavorful and nutritious crops. Understanding the diversity of these plants is fundamental to managing agriculture and food security in a changing world. This study surveyed fruit trait variation across species of tomato wild relatives and explored associations among color, size, shape, sugars, and acids. These wild tomato species native to South America can be interbred with the economically important cultivated tomato. Beyond its application to tomatoes, deepening our knowledge of how fruit traits evolve together is valuable to crop improvement efforts aimed at breeding more nutritious and appealing varieties of fruits.

Fleshy fruits display a striking diversity of traits, many of which are important for agriculture. The evolutionary drivers of this variation are not well understood, and most studies have relied on variation found in the wild. Few studies have explored this question on a fine-grained scale with a group of recently diverged species while controlling for environmental effects.

We developed the tomato clade as a novel system for fruit trait evolution research by presenting the first common garden-based systematic survey of variation and phylogenetic signal in color, nutrition, and morphology traits across all 13 species of tomato wild relatives (Solanum sect. Lycopersicon). We laid the groundwork for further
testing of potential evolutionary drivers by assessing patterns of clustering and
correlation among disperser-relevant fruit traits as well as historical climate variables.

We found evidence of two distinct clusters of associated fruit traits defined by
color, sugar type, and malic acid concentration. We also observed correlations between a
fruit's external appearance and internal nutrient content that could function as honest
signals to dispersers. Analyses of historical climate and soil variables revealed an
association between red/orange/yellow fruits and high annual average temperature.

Our results establish the tomato clade as a promising system for testing
hypotheses on the drivers of divergence behind early-stage fleshy fruit evolution,
particularly selective pressure from frugivores.

1.2 Introduction

Flowering plants produce an astounding variety of fruits that have long fascinated
biologists, yet questions remain as to how this diversity has arisen. Of particular interest
are fleshy fruits, which enable seed dispersal through animal consumers and also provide
great value to human society as agricultural food crops. Research across many species
has established that fruit traits do not vary independently. Instead, certain suites of traits
tend to occur together (for example small size, bright colors, and high sugar content) and
often involve correlations between external appearance and internal chemical
composition (Valido et al. 2011; Schaefer et al. 2014; Nevo et al. 2019; Sinnott-
Armstrong et al. 2020; Sinnott-Armstrong et al. 2022). Repeatedly observed associations
of certain traits are often referred to as fruit or dispersal syndromes (van der Pijl 1969;
Valenta and Nevo 2020; Rojas et al. 2022).
Decades of research have explored factors that may underlie these patterns in fleshy fruit trait covariation. Frugivorous animal preferences have long been considered an important evolutionary force, since plants better at attracting seed-dispersing fruit consumers improve their reproductive success (Ridley 1930; van der Pijl 1969; Janson 1983; Brodie 2017; Valenta et al. 2018; Nevo et al. 2018b). A conflicting force may be damage-inflicting seed predators and pathogens, in which case fruits that can deter or defend against these pests have a reproductive advantage (Mack 2000; Whitehead et al. 2016; Nevo et al. 2017). Adaptation to abiotic factors such as openness of habitat (Bolmgren and Eriksson 2005) or climatic variables (Zhao et al. 2018) may also underlie some fruit trait diversity. Additionally, fruit phenotypes may be constrained by mechanical (Valenta et al. 2022), chemical (Whitehead et al. 2016), developmental (Nevo et al. 2020), or other non-adaptive limitations due to shared genetic underpinnings from phylogenetic relatedness (Jordano 1995).

In recent years, our understanding of fruit trait evolution has been advanced by a number of studies (Nevo and Ayasse 2019; Valenta and Nevo 2020), including analyses of large global datasets (Sinnott-Armstrong et al. 2018; Nascimento et al. 2020; Onstein et al. 2020; Sinnott-Armstrong et al. 2021), but certain gaps have yet to be addressed. Most previous work has examined large sets of distantly related plant species. Some results from these broad datasets (e.g., Jordano 1995) suggested that congeneric species may have limited potential for fruit trait evolutionary response due to shared ancestry, but the fine-grained studies on closely related species needed to test this hypothesis are lacking. Some studies focus on only one habitat, such as a cloud forest (e.g., Rojas et al. 2022), where all species have adapted to similar abiotic conditions. These gaps could be
addressed by examining closely related species that have recently diverged to occupy a range of environments. This type of system would also provide clearer insight into the early stages of fruit evolution than has been possible with previous approaches.

Furthermore, few studies have used a controlled common garden setup in which all plants are grown under the same conditions to ensure that observed variations in fruit traits are mainly due to genetics rather than environment.

One system well-suited to this fine-grained approach to fruit diversity studies is the tomato clade, a group of 14 congeneric species (*Solanum* sect. *Lycopersicon*, Peralta et al. 2008) that includes the cultivated tomato (*Solanum lycopersicum* var. *lycopersicum*) and 13 species of wild relatives (Fig. 1.1). Throughout this paper, we use the word ‘tomato’ to refer to any of the 14 species in the group. The clade is a monophyletic group estimated to have diverged from a common ancestor ~2 to 2.5 million years ago based on fossil-calibrated molecular clock evidence (Särkinen et al. 2013; Pease et al. 2016), which enables consideration of phylogeny-related constraints at a finer time scale than many previous studies of fruit trait diversity. The 13 wild tomato species inhabit a range of environments across the coasts, deserts, and mountains of western South America, allowing us to account for abiotic factors as potential selective forces. The group displays interesting inter- and intra-species fruit differences such as colors ranging from red to orange to yellow to green—enough variation to make testing hypotheses of fruit trait evolution feasible. The tomato system has been well studied as a model of fruit development and genetics (Kimura and Sinha 2008) as well as ecological and evolutionary genomics (Moyle 2008), providing a stronger understanding of the molecular underpinnings of traits than is possible with most other fruit systems. Despite
these advantages of the tomato clade as a study system, the extent of fruit trait diversity across all 13 wild species and the potential evolutionary drivers of this variation have yet to be assessed. By including all species in the clade, we establish a complete evolutionary framework so that assumptions do not have to be made about how often traits evolved independently.

In this study we present the groundwork for a novel approach and scale to fruit trait evolution studies using the tomato system and a common garden. Given that wild tomatoes display multiple fruit colors and inhabit a range of environments, the system holds potential for testing the relative importance of animal preferences and abiotic conditions, two factors currently considered major drivers of fruit color syndromes (Sinnott-Armstrong et al. 2021). Our specific objectives were to: 1) Establish whether the tomato clade holds enough genetically-driven variation in disperser-relevant traits to make future hypothesis testing tractable. We did this by assessing the extent of variation in color, nutrition, and morphology traits in a common garden and whether the traits vary independently of phylogeny. 2) Explore whether patterns of covariation exist that could be tested for underlying evolutionary drivers. We did this by a) assessing whether any groups of disperser-relevant fruit traits cluster together in distinct syndromes, b) whether pairwise correlations exist between external appearance and internal composition, and c) whether any fruit traits are correlated with historical climate variables. The presence of fruit syndromes in the tomato clade, particularly if they involve associations among external color and internal nutrients after correcting for phylogeny and accounting for climate, could be used to test the hypothesis that animal disperser preferences were an evolutionary driver of fruit trait divergence. This so-called dispersal syndrome hypothesis
has been tested in a number of systems (Valenta and Nevo 2020), but not yet across an entire clade of closely related species that have adapted to different environments on a relatively recent timescale.

1.3 Methods

1.3.1 Plant material

Our sampling included 38 accessions chosen to span the known phylogenetic and geographic diversity across the tomato clade, consisting of 13 taxa: 12 species of wild tomato and the semi-wild *S. lycopersicum* var. *cerasiforme* (Fig. 1.1). Three accessions of each taxa were selected, representing different parts of their geographic ranges when possible (Fig. 1.1c). Seeds were obtained from the C. M. Rick Tomato Genetics Resource Center at the University of California, Davis, USA (TGRC, http://tgrc.ucdavis.edu) and the Universitat Politècnica de València, València, Spain (accession IDs and collection locations listed in Appendix A, Table A1). Each accession is representative of an independently sampled population in the wild. For autogamous (self-fertilizing) self-compatible species (*S. cheesmaniae, S. galapagense, some S. pimpinellifolium, S. lycopersicum* var. *cerasiforme, S. neorickii*) each plant of an accession is likely genetically identical. For others, plants are not identical, as the facultative self-compatible or allogamous self-incompatible accessions are maintained through "mass sibling" pollination in germplasm centers.

Seeds were soaked in a 2.7% sodium hypochlorite solution (per TGRC protocol) and germinated in a greenhouse at the University of Massachusetts Amherst. After eight weeks, seedlings were transplanted into a high tunnel greenhouse at the University of
Massachusetts Crop and Animal Research and Education Farm in South Deerfield, MA, where they were irrigated twice a week and fertilized once a week via drip lines. For each accession, three separate plants were grown in randomized locations. The study plants grew throughout the summer farm season (May–September 2020). Self-incompatible accessions were hand-pollinated several times during the first few weeks of flowering to facilitate fruit set, otherwise flowers were accessible to natural pollinators.

1.3.2 Fruit phenotyping

We chose to quantify 21 fruit traits representing morphology, color, and nutrition (Appendix A, Table A2), the three disperser-relevant categories commonly used in fruit studies (e.g., Valido et al. 2011; Rojas et al. 2022). We measured size and shape because they affect an animal's ability to handle and consume the fruit. Another trait of potential interest to dispersers is the proportion of a fruit's cross-section taken up by fleshy pericarp rather than the seed-containing central locules, known as pericarp area ratio, as this could affect the overall nutritional reward found in one fruit. We quantified color into three variables with the CIE L*C*h system (lightness, chroma, and hue) as they provide unambiguous color measures commonly used in comparative studies (Valido et al. 2011). Because the main caloric reward of tomatoes is sugar (fruits consist of about 90% water and are low in protein and lipids (García-Alonso et al. 2020), we quantified glucose and fructose (i.e., hexoses) as well as sucrose, the three most abundant sugars in tomatoes (Beckles et al. 2012). We also measured the organic acids citric acid and malic acid because they are another major component of cultivated tomato flavor (Tieman et al. 2017) and thus may be relevant to attracting animal dispersers. Ratios between the types
of sugars or acids, as well as the sugar to acid ratio, were considered distinct traits because these proportions have important effects on human taste preferences (Malundo et al. 1995; Anthon et al. 2011), and thus possibly on other animals' tastes.

We collected fruits as they turned fully ripe, which we determined qualitatively for each species (see Appendix A, Table A3). A total of five fruits per plant (and thus 15 fruits per accession) were phenotyped when possible. After measuring diameter (latitudinally across widest part) and length (longitudinally from stem end to blossom end) with digital calipers, and fresh weight with an analytical balance, fruits were cut in half latitudinally and scanned at 400 dpi with a color document scanner. Then seeds and pulp were removed, and the remaining fruit pericarps were frozen and stored at −80 °C until we later thawed and quantified their liquid extracts for concentrations of glucose, fructose, sucrose, citric acid and malic acid with absorbance-based assay kits, and Brix (total soluble solids) with a refractometer (Appendix A, Methods A1). Color, lobedness degree, and pericarp area ratio were calculated from scanned images of fruit cross-sections using Tomato Analyzer version 4.0 software (Darrigues et al. 2008) after manually adjusting the pericarp boundary. Color was calibrated with a ColorChecker Classic (X-Rite, Grand Rapids, MI, USA).

All data analyses were conducted in R v3.6.3 (R Core Team 2022). Summary statistics for each of the 21 variables were computed with built-in R functions and plots produced with the R package GGPPLOT2 v3.3.3 (Wickham 2016). Accession mean trait values were calculated (Appendix A, Table A4) for use in later analyses that required one value per accession. To account for uneven sampling, first the means for all fruits from
each plant (generally 5 fruits per plant) were computed, then per-plant means (generally 3 plants per accession) were averaged to arrive at a per-accession mean value for each trait.

1.3.3 DNA extraction, Genotyping-by-sequencing, and phylogenetic tree construction

We constructed an original Genotyping-by-Sequencing (GBS) phylogeny to provide a fine-grained, accession-level resolution of evolutionary relatedness. DNA was extracted from ground frozen tissue of young unexpanded leaves using the CTAB method (Porebski and Bailey 1997). A reduced representation GBS library was prepared according to a protocol modified from (Elshire et al. 2011) using the ApeKI restriction enzyme (Appendix A, Methods A2), and sequenced on one lane of an Illumina HiSeq PE 150 (Illumina Inc., San Diego, CA, USA) by Novogene, Inc. (Sacramento, CA, USA). Sequences were processed and aligned to the SL4.0 cultivated tomato reference genome (Hosmani et al. 2019) to produce a variant call file (VCF), from which a set of filtered SNPs was used to construct a maximum likelihood consensus tree (Appendix A, Methods A3). To quantify phylogenetic signal, we used the function phyloSignal from the R package PHYLOSIGNAL v1.3 (Keck et al. 2016) to calculate Blomberg’s K (Blomberg et al. 2003) and Pagel’s λ (Pagel 1999) from accession mean values of each trait and the original phylogenetic tree described above. Trees with trait values were produced with the function contMap from the R package PHYTOOLS v0.7-70 (Revell 2012).

1.3.4 Trait cluster analysis

We performed a principal components analysis (PCA) on the three color, 11 nutrition, and seven morphology traits. We used centered and scaled raw trait values for
208 individual fruit samples in the function `prcomp` from the R package `STATS`. A correlation matrix was used to normalize the covariance to a constant scale since variables were measured in different units.

To test fruit trait differences between the two color groups, we produced phylogenetic generalized least squares (PGLS) models that estimated lambda. We used the functions `comparative.data` and `pgls` from the R package `CAPER v1.0.1` (Orme et al. 2013) with lambda = "ML" to simultaneously optimize phylogenetic signal and regression parameters as recommended by (Revell 2009).

1.3.5 Correlation analyses for pairs of traits

To assess correlations between pairs of variables while taking phylogeny into account, as in Valido et al. (2011) and Rojas et al. (2022), we calculated phylogenetically independent contrasts (PICs) for each trait with the function `pic` from the R package `APE v5.6-2` (Paradis and Schliep 2019), using centered and scaled accession mean values as input. We then calculated Pearson's r for these PIC trait values using the function `corr.test` in the R package `PSYCH v2.2.9` (Revelle 2022) with p-values adjusted for multiple testing via the false discovery rate method (Benjamini and Hochberg 1995).

1.3.6 Abiotic variables analyses

Historical climate and soil data (36 variables total) for each accession's collection coordinates were obtained from the WORLDCLIM and SoilGrids databases (Appendix A, Methods A4; Appendix A, Table A1). We examined Pearson's correlation coefficients for pairs of variables (Appendix A, Table A5; calculated from centered and scaled values
using the `corr.test` function in R with false discovery rate p-value adjustment) and in cases where $r > |0.65|$ we kept only one variable, resulting in these eight uncorrelated variables: temperature annual average, precipitation annual total, temperature seasonality, precipitation seasonality, solar radiation annual average, wind speed coefficient of variation, soil bulk density, and soil coarse fragment volume.

We produced PGLS models with one fruit trait as the response variable and the eight uncorrelated climate variables as predictors using the `ppls` function with lambda = "ML". All variables were centered and scaled prior to analysis. Top models based on AIC scores were compared with the `dredge` function, and each predictor's sum of Akaike weights over all models was calculated with the `importance` function, both from the R package `MuMIn` v1.43.1 (Barton 2022).

1.4 Results

1.4.1 Extent of variation in individual fruit traits

Because the extent of fruit trait diversity across all wild tomato species had not previously been surveyed, we first quantified the ranges of variation across all fruits for each trait separately. We used coefficient of variation (CV) to compare variability across traits, and found that, in general, nutrition traits showed greater variability than color and morphology (Figs. 1.1a, 1.2; Appendix A, Table A6). Total sugars ranged in concentration from 4.1 ($S. neorickii$) to 168.0 ($S. arcanum$) mg/ml, and total acids from 0.24 ($S. huaylasense$) to 30.4 ($S. habrochaites$) mg/ml (Appendix A, Table A2). However, the most variable nutrition trait in terms of CV was sucrose/hexose ratio (Appendix A, Table A6). Color variations included light green, dark green, yellow,
orange, red, and some purple streaking on a few otherwise green fruits (Fig. 1.2). Chroma, which measures how vivid colors are, had the greatest CV, followed closely by hue (color of the visual spectrum; Appendix A, Table A6). In general, greenish fruits tended to be less pure/vivid and more white (lightness) than red/orange/yellow fruits.

Tomato fruit sizes mainly fell between 10–20 millimeters in diameter and 1–3 grams in fresh weight, ranging from minimums of 7.64 mm/0.24 g in S. galapagense to maximums of 28.29 mm/11.43 g in S. lycopersicum var. cerasiforme. Shapes were quite spherical for most species (diameter/length ratio 1.0–1.2), except for S. pennellii which tended to be more oval with a ratio as high as 1.5. The greatest CV for morphology traits was observed for fresh weight and seed count (Appendix A, Table A6).

1.4.2 Trait variation relative to our GBS phylogeny

We established a fine-grained phylogenetic framework by creating an original phylogeny of the accessions phenotyped in this study. This enabled us to account for the statistical nonindependence of data points from related taxa due to shared ancestry (Felsenstein 1985; Huey et al. 2019), as well as explore how patterns in fruit variation aligned with evolutionary relatedness. A total of 64,745 filtered SNPs were used to produce the maximum likelihood consensus tree, with TVM+F+ASC+R4 selected as the best-fit nucleotide substitution model.

Our tree recovered the current paradigm of relationships among wild tomato species (Fig. 1.1b). The four colored-fruited species (S. cheesmaniae, S. galapagense, S. lycopersicum var. cerasiforme, and S. pimpinellifolium) fell within a well-resolved monophyletic clade, with S. cheesmaniae and S. galapagense, two species endemic to the
Galápagos, as sister species. Sister to that clade was a monophyletic group consisting of *S. neorickii*, *S. chmielewskii*, and *S. arcanum*. Of note was the placement of *S. arcanum* LA2153 within the *S. peruvianum* clade, suggesting that the taxonomic designation of this accession may need revision. Our tree was consistent with other evidence that genetic identity can conflict with taxonomic designation for some accessions of *S. arcanum*, *S. corneliomulleri*, *S. huaylasense*, and *S. peruvianum* (Labate et al. 2014; Pease et al. 2016). Our *S. chilense* accessions formed a distinct group within the larger "peruvianum complex" clade, consistent with evidence that *S. chilense* evolved from *S. peruvianum* ancestors (Stam et al. 2019). Finally, *S. habrochaites* and *S. pennelli* appeared as well-differentiated sister species that grouped together in a clade sister to all the other tomato species.

The trait phylogenetic signal results based on this tree confirmed the need to incorporate phylogeny into our analyses when possible, as 13 of 21 traits had significant values for both Blomberg’s K and Pagel’s lambda (p < 0.01; 10 of those 13 traits were also significant at a Bonferroni-corrected cutoff of p < 0.002) (Appendix A, Table A6). These patterns also provided insight into which traits were labile (able to change over short evolutionary timescales) versus which were potentially more constrained. Traits with low phylogenetic signal values (K < 0.8, lambda < 0.65) indicative of lability were the nutrition variables of total sugars, total acids, citric acid, Brix, and sugars/acids ratio, as well as the size variables of diameter, length, and fresh weight. For those nutrition traits, high and low values were spread randomly when plotted across the tree (Appendix A, Fig. A1n,o,q,r,s). For the size traits, the low phylogenetic signal was mainly due to the smallest- and largest-fruited species (*S. galapagense* and *S. lycopersicum* var.)
*cerasiforme* being closely related (Appendix A, Fig. A1a-b,d). The remaining traits displayed high phylogenetic signal values (K > 0.8, lambda > 0.65), although this was due to several different distribution patterns across the tree depending on the trait (Appendix A, Fig. A1). For the three color variables, the high signal was largely due to all of the color-fruited accessions belonging to the same monophyletic group. A similar distribution was seen for the three sugars and malic acid (as well as their associated ratios), with most or all accessions within the colored-fruited clade showing high glucose (Appendix A, Fig. A1k) and fructose values but low sucrose and malic acid values relative to the green-fruited accessions. Interestingly, the green-fruited *S. pennellii* was an exception to this pattern as the LA1809 and LA2580 accessions were relatively high in glucose and fructose but low in sucrose (although they were high in malic acid like the rest of the green-fruited accessions). In contrast, the high signal in the case of diameter/length ratio (Appendix A, Fig. A1c), lobedness degree, and seed count was likely due to only the closely related *S. pennellii* accessions displaying the highest trait values, as low trait values were more scattered throughout the tree.

1.4.3 Trait clustering

We looked for evidence of fruit syndromes through PCA, to determine how fruits clustered together based on phenotypic similarity and which traits contributed most to this positioning. A PCA with all 208 individual fruits measured for all 21 traits as points (Fig. 1.3) showed two distinct clusters differentiated primarily along PC1, and factor loadings (Appendix A, Table A7) provided insight into which traits covaried together. One group consisted of fruits primarily green in hue, white in lightness, and high in malic
acid/sucrose/seed count, while the other contained fruits reddish in hue, more vivid in chroma, high in glucose/fructose, and lower in seed count. There was no obvious distinction along PC2, suggesting that the size and total sugar concentration traits most influential to that axis were not part of the two observed clusters. With the caveat that analyzing subsets of data may introduce bias, clustering patterns in PCAs conducted separately for each trait category (Appendix A, Fig. A2) also supported the existence of two fruit trait clusters differentiated primarily by color, sugar type, malic acid, and possibly seed count and shape.

Because these two clusters appeared to correspond to the informal colored-fruited and green-fruited species sub-groupings within the clade, we assessed the strength of these two putative syndromes by testing differences between the two color groups in sugar types, malic acid, seed count, and shape with PGLS models, which account for phylogenetic relatedness (Fig. 1.4). There were statistically significant (p < 0.008 after Bonferroni correction) differences in glucose (lambda = 0, p = 1e-7), sucrose (lambda = 0.09, p = 8.1e-7), and malic acid (lambda = 0.584, p = 0.0032), but not fructose (lambda = 1, p = 0.0496), seed count (lambda = 0.946, p = 0.346), and diameter/length ratio (lambda = 1, p = 0.475). In summary, sugar type and malic acid concentration were the key traits along with color that defined the two broad fruit clusters within the clade, with high glucose and low malic acid associated with colored fruits, while high sucrose and high malic acid were associated with green fruits.

Looking at how species were distributed within the two broad PCA clusters, the colored-fruited group showed more species-specific differentiation than the green-fruited group when all traits were considered together (Fig. 1.3). Different fruit trait types
appeared to vary in their effects on species differentiation when PCAs were conducted separately (see caveat above) for each trait category (Appendix A, Fig. A2). Noteworthy species with unique traits were *S. lycopersicum* var. *cerasiforme* (large size), *S. galapagense* (small size and high glucose/fructose), *S. pimpinellifolium* (high glucose/fructose), and *S. pennellii* (more oval shape and high seed count). Interestingly, the various shades of green displayed by the green-fruited species are not distinct enough to differentiate the species.

### 1.4.4 Correlations among pairs of traits

We further examined associations among fruit traits in a phylogenetic context by conducting Pearson correlation tests on each trait's phylogenetically independent contrasts (PICs). To assess whether correlation patterns differed within color groups, calculations were made separately for: 1) all 38 accessions, 2) only the 12 colored-fruited accessions, and 3) only the 26 green-fruited accessions (Appendix A, Table A8). Stars following correlation coefficients denote significance levels (**p<0.01, *p<0.05) after false discovery rate adjustment for multiple testing.

Variables that were very highly correlated (Pearson r > 0.8***) in all three groupings revealed some measurement redundancy among the 21 traits. The size variables of diameter, length, and fresh weight all had r values > 0.9***. The total sugars-Brix correlation (r > 0.87***) suggested that glucose, fructose, and sucrose make up the majority of soluble solids in wild tomatoes, while total acids-citric acid (r > 0.86***) showed malic acid to be the substantially less abundant of the two acids quantified.
We next looked at pairs of color or size and sugar/acid traits to explore whether a fruit's external appearance may provide an honest signal of its nutrient content. The only pairs of color-nutrition traits that were consistently correlated ($r > 0.4$) across all three groupings (Appendix A, Table A9) were chroma-total sugar ($r = 0.48^*, 0.70, 0.54$ for all, colored-fruited, and green-fruited groups, respectively) and lightness-sugars/acid ratio ($r = 0.48^*, 0.43, 0.49$). These associations suggested that throughout the clade, more vivid fruits tend to have higher sugar concentrations and more whitish fruits tend to have more sugars relative to acids. The only morphology-nutrition pairing consistently correlated ($r > |0.38|$) across all three groupings was diameter-glucose ($r = -0.49^*, -0.66, -0.39$), showing that smaller fruits tend to have higher glucose concentrations, particularly within the colored-fruited group.

Other pairs of traits showed correlation patterns that differed between the two color groups (Appendix A, Table A9). Associations that were present only in the colored-fruited group were related to color and sugar, most notably chroma-glucose, chroma-fructose, hue-glucose, and hue-fructose ($r > |0.45|^*$). Additionally, chroma-hue were correlated in the colored-fruited group ($r = -0.70^*$) but not the green-fruited group ($r = 0.19$). In contrast, an association only present in the green-fruited group was lobedness degree-malic/citric ratio ($r = 0.57$). Interestingly, size and sugar were correlated in both groups but in opposite directions: diameter-sucrose ($r = -0.68$) and diameter-total sugars ($r = -0.71$) were negative within colored-fruited and positive within green-fruited ($r = 0.48$ and $r = 0.41$, respectively), suggesting that larger fruits tend to have lower sugar concentrations in the colored-fruited group but higher sugar concentrations in the green-
fruited group. These associations between a fruit's external appearance and internal nutrient content could potentially function as honest signals to animal dispersers.

1.4.5 Abiotic factors associated with fruit traits

To explore whether the fruit trait covariations seen in our clustering and correlation analyses may have been related to adaptation to environmental niches, we assessed which historical climate variables were most associated with particular fruit traits via two methods.

We first conducted Pearson correlation tests for eight uncorrelated climate variables (described in Methods) and 12 fruit variables (chosen because they were important in clustering and correlation results). A non-phylogenetic correlation chart of these 20 variables (centered and scaled) with all 38 accessions included (Appendix A, Table A10) showed that seven climate-fruit trait pairs had $r \geq |0.4|$, of which two had $r \geq |0.5|$. Stars following correlation coefficients denote significance levels (**$p<0.01$, *$p<0.05$) after false discovery rate adjustment for multiple testing. Temperature annual average was correlated ($r > |0.40|**$) with lightness, chroma, glucose, fructose, and sucrose. The only other climate-fruit trait pairs with $r \geq |0.4|$ were temperature seasonality-fructose ($r = 0.40*$) and precipitation annual total-diameter ($r = 0.42**$).

Before concluding whether a pair of traits was associated, we also examined PGLS models in which one of the 12 syndrome-relevant fruit traits was the response variable and the eight climate variables were predictors. For each model, predictor variables with sum of weights $> 0.6$ were considered strongly associated with the fruit
trait response variable. Six of the 12 response variables had one or more predictor climate variables above the cutoff (Appendix A, Table A11).

After comparing results from the two methods, only three fruit-climate associations (Appendix A, Fig. A3) were above our cutoffs in both the non-phylogenetic Pearson correlation tests and the PGLS models: 1) diameter-precipitation annual total (Pearson $r = 0.42^{**}$, PGLS sum of weights 0.87), 2) chroma-temperature annual average (Pearson $r = 0.50^{**}$, PGLS sum of weights 0.67), and 3) fructose-temperature seasonality (Pearson $r = 0.40^*$, PGLS sum of weights 0.63). Two of those three associations involved color or sugar variables that were part of the fruit trait clusters identified above. Chroma-temperature annual average was the stronger of the two, leading us to hypothesize that the evolution of the colored-fruited cluster may be related to warmer climates. To visualize whether closely related accessions inhabit similar environments in terms of temperature annual average, we plotted these temperature values on our phylogeny (Fig. 1.5a). The majority of accessions in the monophyletic colored-fruited clade were collected from warmer locations than the green-fruited accessions, with a significant difference (PGLS p-value < 0.001) between the color groups (Fig. 1.5b).

1.5 Discussion

Fleshy fruits display a striking diversity of traits whose evolutionary drivers are not well understood despite their relevance to agriculture. Using a common garden and a fine-grained phylogenetic framework, we developed a novel approach and scale for fruit evolution research by conducting the first clade-wide systematic survey of disperser-relevant fruit traits across wild tomato species. Contrary to expectations that fruit traits
have low evolutionary potential due to constraints of shared ancestry, as suggested by taxonomically broad studies (e.g., Jordano 1995), we found substantial variation in color and nutrition traits across the recently diverged tomato clade. Morphology traits did not vary as widely across wild species, which is notable given the many shapes and sizes seen in cultivated tomato varieties. Total sugars, total acids, and size varied independently of phylogeny, while differences in color, sugar type, and malic acid concentration were aligned with the evolutionary split between the four colored-fruited species and the nine green-fruited species. After correcting for non-independence due to shared ancestry, the differences in glucose, sucrose, and malic acid between those two groups of species were statistically significant, suggesting that the differences have been maintained by selective pressure. Furthermore, these trait differences clustered into syndromes involving covariations among external color and internal nutrients that could function as honest signals to dispersers, a pattern in line with the hypothesis that animal preferences are an underlying evolutionary driver. Historical climate and soil variables were not strongly correlated with fruit traits, so selective pressure from frugivores may have been a more important evolutionary force than adaptation to abiotic conditions. We did find one notable association between colored-fruited species and high annual average temperature that should be considered when trying to tease apart the biotic and abiotic factors underlying the two wild tomato fruit syndromes.

To our knowledge there is currently no systematic data on which animals eat wild tomato fruits, although there are some anecdotal reports. The TGRC database contains collection notes from over 1700 total accessions of species we studied, out of which we found only 35 records that mention the words “eaten”, “disperse”, “animal”, “bird”, or
“rodent” (Appendix A, Table A12), providing evidence of inferred or reported fruit consumption by humans, grackles, other birds, tortoises, rodents, and “other animals”.

These TGRC notes do not reveal any clear patterns in disperser preferences: birds, mammals, and reptiles appear to consume both colored-fruited and green-fruited species but the prevalence of any given animal cannot be determined from these limited anecdotal reports. Beyond collection notes, studies of captive animals showed that passage through the guts of Galápagos tortoises (Rick and Bowman 1961) and mockingbirds (Rick 1964) can improve germination of *S. cheesmaniae* seeds, although dispersal in the wild is still unstudied (Heleno et al. 2011). There has been one recent video captured of a Galápagos mockingbird eating a yellow wild tomato fruit (Matthew Gibson 2019, personal communication). Recent field observations of *S. chilense* plants in Chile noted that ripe fruits always hang low to the ground and may be dispersed by *Microlophus* lizards (Remco Stam and Edeline Gagnon 2023, personal communication). One field study in northern Chile (Chetelat et al. 2009) found partially eaten fruit remains from *Solanum sitiens*, a green-fruited relative of wild tomatoes, in a rodent burrow in northern Chile. Beyond that we are not aware of any published data of potential tomato seed dispersers. Future studies that employ cameras to monitor animal visitation at plants with ripe fruits (e.g., Levey et al. 2006) would be a valuable contribution to the limited data currently available.

The two distinct trait clusters we observed involve associations among color and nutrient content in line with those found in other fruit systems where selective pressure from animal dispersers (i.e., the dispersal syndrome hypothesis) has been inferred (Valido et al. 2011; Sinnott-Armstrong et al. 2020; Valenta and Nevo 2020). The colored and
green tomato fruit groupings correspond to distinct balances of sugar and acid types, which are generally consistent with conventional syndromes (Janson 1983) based on bird vs. mammal preferences. One observed cluster consists of fruits that are red/orange/yellow, high in glucose and fructose, and low in sucrose and malic acid—traits thought to be preferred by birds, which are expected to be the main dispersers of the colored-fruited tomatoes because most birds have excellent color vision that they use to locate food (Lomáscolo and Schaefer 2010), and some birds are unable to digest sucrose (Martinez del Rio and Stevens 1989). The other cluster includes fruits greenish and lighter in color, high in sucrose and malic acid but low in glucose and fructose - traits generally thought to be more in line with fruits primarily dispersed by mammals. Despite the difference in sugar types, both groups of fruits produce similar total sugar concentrations (Appendix A, Fig. A4), suggesting both syndromes offer comparable overall caloric rewards. The fact that green-fruited species devote energy into sugary fruits could mean that they benefit from attracting animal dispersers, a hypothesis contrary to speculations that green-ripe wild tomato fruits are not consumed at all (e.g., Kamiyoshihara et al. 2020).

Our correlation analyses show that a wild tomato fruit's external appearance could also provide reliable information about its internal nutrient content on a finer, more quantitative scale, with a few different nuances for the colored vs. green species groups. Regardless of species grouping, high chroma fruits tend to have greater total sugar concentrations and high lightness fruits tend to have more sugars relative to acids. Within each color group, however, hue and diameter appear to signal different nutritional traits: for the colored-fruited species only, less reddish, smaller fruits tend to be higher in
glucose, fructose and total sugar concentration; while for the green-fruited species only, less greenish, larger fruits tend to be higher in total sugar and lower in citric acid concentration. Interestingly, the colored-fruited species showing higher sugar concentration in smaller fruits is consistent with cultivated tomatoes (Levin and Schaffer 2013), but the green-fruited species displayed the opposite trend with sugar concentration tending to be higher in larger fruits. These different nuances in signaling between species groups merit further testing as to whether they may have been influenced by distinct pressures from dispersers. Additionally, it would be valuable to assess whether signal-reward correlations exist for fruits at the intraspecies level because that scale of variation may be more ecologically relevant to frugivores (Nevo et al. 2022).

While some correlations between color and nutrition could be due to shared biochemistry rather than natural selection from animals (Cazetta et al. 2012), this seems unlikely to be the case for tomato fruits because color, sugar, and acid are controlled by distinct biochemical pathways. Red/orange/yellow colors are produced via the carotenoid pathway from isoprenoid precursor molecules (Sun and Li 2020), sugar levels are altered by sucrose and starch metabolism (Beckles et al. 2012), and citric and malic acid are intermediates in the tricarboxylic acid cycle (Zhang and Fernie 2018). Molecular mapping has revealed only a few genomic regions affecting both sugar and acid content (Grandillo et al. 2013). Humans have taken advantage of simply inherited variations in these disparate pathways to produce different color, sugar, and acid combinations in domesticated cultivars (Levin and Schaffer 2013), demonstrating that changes in only a few loci can result in large fruit trait differences. For example, altered expression levels of the \textit{TIV1} invertase gene cause tomato fruits to shift between sucrose or hexose
accumulation (Moy et al. 2007). Given that the biochemical basis of color, sugar, and acid traits allows for evolutionary lability, the persistence of two clusters of associated traits with high phylogenetic signal within the clade suggests a stabilizing selective force.

The correlations among color, sugar type, and malic acid traits may be linked to closely related species adapting to environments with similar selective pressures. Our results suggest that colored-fruited accessions evolved in locations with warmer annual average temperature (which correlates with lower elevation and higher annual average vapor pressure) than green-fruited accessions, consistent with previous work showing the mean annual temperature across species distributions was higher for *S. cheesmaniae*, *S. galapagense*, *S. lycopersicum* var. *cerasiforme*, and *S. pimpinellifolium* relative to the green-fruited species (Nakazato et al. 2010; Ramírez-Ojeda et al. 2021). Thus something about warmer locales may have been necessary for the evolution of colored fruits.

Perhaps warmer environments enable fruits to complete a more energetically expensive developmental color sequence due to a longer growing season (Sinnott-Armstrong et al. 2018), which could offset the lost photosynthetic capability in fruits that become brightly colored upon ripening (Cipollini and Levey 1991). Warmer climates could also have the indirect effect of unique biotic pressures, such as a greater proportion of animal dispersers that prefer colored fruits (Willson and Whelan 1990).

It is also worth considering what evolutionary forces may be behind some of the more unique species within the clade. *S. pennellii* fruits stand out for their oval shape and high seed count. The species inhabits environments with relatively high annual average temperature (compared to other green-fruited species) and very low annual precipitation (second lowest in the clade behind *S. chilense*) (Nakazato et al. 2010; Ramírez-Ojeda et
al. 2021); perhaps this unique combination plays host to a different assemblage of animal dispersers. *S. lycopersicum* var. *cerasiforme* produces the largest fruits in the clade and is found in locales with the highest annual precipitation; this correlation could be due to greater water availability being necessary for the production of larger fruits (tomato fresh weight is ~90% water), or perhaps locations with more rainfall have different resident dispersers that are able to consume larger fruits. The two Galápagos species are notable for stark differences in fruit traits between very closely related species - *S. cheesmaniae* fruits have low chroma, light orange/yellow/greenish hue, and low total sugar concentrations, while *S. galapagense* fruits have high chroma, darker orange hue, small size, and high total sugar concentration. While genetic drift may be a powerful force in these small island populations, whether selection from dispersers has influenced the differences is an open question.

Our evidence of an association among color, sugar type, and malic acid traits in wild tomato fruits lays the groundwork for future tests of the dispersal syndrome hypothesis. More data on wild tomato frugivores and their preferences would enable the hypothesis to be tested on a more fine-grained phylogenetic scale than has yet been used. Furthermore, data on disperser preferences and a more thorough quantification of intraspecific fruit trait variation could test whether the correlations among color, nutrition, and size we observed do in fact serve as honest signals linking a fruit's external appearance to its internal content. Some abiotic factors such as temperature annual average may be related to the trait associations, although the influence of climate is difficult to disentangle from the effects of unique assemblages of frugivores inhabiting different environments. Another pressure we did not explore but would be worth
investigating is defense, as pathogens or seed predators may have influenced variation in
traits such as trichomes, acylsugars, and glycoalkaloids.

Cultivated tomatoes have long been considered the predominant model organism
for fleshy-fruited plants, leading to many ongoing discoveries of the genetic mechanisms
and developmental processes controlling fruit fleshy fruit traits. Our survey shows that
the wild tomato clade harbors a rich diversity of fruit traits, offering an exciting
opportunity to explore the evolutionary drivers that made the eventual domestication of
this beloved fruit crop possible. Further expanding the study of model systems to
incorporate wild relatives, particularly through ecological field studies of biotic
interactions, as is now possible in tomatoes, can provide unparalleled opportunities to
understand the genetic, developmental, and ecological factors that have shaped fleshy
fruit evolution.

1.6 Figures
Figure 1.1 Overview of the tomato clade (*Solanum* sect. *Lycopersicon*) and the 38 accessions used in this study. (a) Representative ripe fruit from each of the three accessions phenotyped per species (except for *S. chilense* (CHI), for which only one of the two phenotyped accessions is shown). Red box denotes colored-fruited group, blue box green-fruited group, white box cultivated tomato. (b) Maximum likelihood consensus
tree of accessions (color-coded by species, see below) based on 64,745 SNP markers obtained through genotyping-by-sequencing, constructed in IQ-TREE with 1000 ultrafast bootstraps using the TVM+F+ASC+R4 nucleotide substitution model selected by ModelFinder with ascertainment bias correction and SH-aLRT likelihood ratio test. Numbers at nodes represent bootstrap percentages; x-axis shows tree scale. The outgroup species *S. lycopersicoides* and the cultivated tomato *S. lycopersicum* "Ailsa Craig" are included in the tree for reference but were not phenotyped for this study. Accession LA0441, which we originally included in the study as an *S. peruvianum*, has now been reclassified as *S. arcanum* in the Tomato Genetics Resource Center (TGRC) database, consistent with its placement on our tree. Only two accessions of *S. chilense* are shown because the third accession we planted (LA1932) failed to produce any fruit. IDs correspond with those provided in Appendix A, Table A1. (c) Map of collection locations for the 38 accessions used in this study, color-coded by species. Countries shown are the southern part of Colombia, Ecuador, Peru, and the northern part of Chile. Colors used in (b) and (c) represent species, abbreviated as follows: ARC = *S. arcanum*, CER = *S. lycopersicum* var. cerasiforme, CHE = *S. cheesmaniae*, CHI = *S. chilense*, CHM = *S. chmielewskii*, COR = *S. corneliomulleri*, GAL = *S. galapagense*, HAB = *S. habrochaites*, HUA = *S. huaylasense*, NEO = *S. neorickii*, PEN = *S. pennellii*, PER = *S. peruvianum*, PIM = *S. pimpinellifolium*. ‘LA’ or ‘BGV’ numbers represent TGRC accession IDs as listed in Appendix A, Table A1.
Figure 1.2 Phenotypic diversity in ripe wild tomato fruits. (a)-(h) Fruit cross-section scans showing fruits with the minimum and maximum values for chroma, hue, lightness, and pericarp area ratio: (a) *S. arcanum* LA2153, (b) *S. galapagense* LA0528, (c) *S. peruvianum* LA0111, (d) *S. arcanum* LA2157, (e) *S. lycopersicum* var *cerasiforme* BGV008189, (f) *S. peruvianum* LA1474, (g) *S. arcanum* LA2157, (h) *S. corneliomulleri* LA1945. (i)-(h) Select photos highlighting the range of variation across the clade: (i) *S. pimpinellifolium* LA0373, (j) *S. galapagense* LA0528, (k) *S. arcanum* LA2157, (l) *S. habrochaites* LA2329, (m) *S. cheesmaniae* LA0428, (n) *S. peruvianum* LA1474, (o) *S. pennellii* LA2963. Scale bars represent 10 millimeters.
Figure 1.3 Wild tomato fruit traits cluster into two putative syndromes mainly defined by color, sugar type, and malic acid concentration, as shown by a non-phylogenetic principal components analysis (PCA) of 21 phenotypic traits for 208 fruits representing 38 accessions. All variables were centered and scaled. In left-hand plot: points = individual fruits; 3-letter abbreviations denote species, abbreviated as follows: ARC = S. arcanum, CER = S. lycopersicum var. cerasiforme, CHE = S. cheesmaniae, CHI = S. chilense, CHM = S. chmielewskii, COR = S. corneliomulleri, GAL = S. galapagense, HAB = S. habrochaites, HUA = S. huaylasense, NEO = S. neorickii, PEN = S. pennellii, PER = S. peruvianum, PIM = S. pimpinellifolium; colored polygons = species. Red circles surround the colored-fruited species, blue circles the green-fruited species. Right-hand plot shows PCA factor loadings of the variables (detailed in Appendix A, Table A7).
Figure 1.4 To assess the strength of the two fruit syndromes suggested by principal components analysis (PCA) (Fig. 1.3), these boxplots test differences between colored-fruited and green-fruited species groups for the nutrition and morphology traits that were the strongest differentiators in the PCA. Colored points represent accessions and are color-coded by species, abbreviated as follows: ARC = *S. arcanum*, CER = *S. lycopersicum* var. *cerasiforme*, CHE = *S. cheesmaniae*, CHI = *S. chilense*, CHM = *S. chmielewskii*, COR = *S. corneliomulleri*, GAL = *S. galapagense*, HAB = *S. habrochaites*, HUA = *S. huaylasense*, NEO = *S. neorickii*, PEN = *S. pennellii*, PER = *S. peruvianum*, PIM = *S. pimpinellifolium*. Red boxplots denote colored-fruited group, blue boxplots green-fruited group. Diameter/length ratio was chosen to represent shape because it was highly correlated ($r = 0.81^{***}$; Appendix A, Table A8) with lobedness degree, the one other shape-related variable we measured. Raw (untransformed) accession mean values were used; sample sizes were 12 for the colored group and 26 for the green group. Lambda and p-values are from phylogenetically controlled generalized least-squares (PGLS) models with lambda="ML".
Figure 1.5 Temperature annual average and fruit color showed the strongest association between an historical climate variable and a syndrome-related fruit trait, leading us to hypothesize that the evolution of colored fruits may be related to warmer environments. (a) Plot showing a heatmap of temperature annual average raw values at accession collection sites (in degrees Celsius) distributed across our phylogenetic tree. Pagel's lambda phylogenetic signal test results (with p-value) for temperature annual average are shown below the tree. (b) Boxplot showing the significant difference in temperature annual average between colored-fruit (12 accessions, denoted by red box) and green-fruited (26 accessions, denoted by blue box) species groups. Colored points represent accessions and are color-coded by species, abbreviated as follows: ARC = S. arcanum, CER = S. lycopersicum var. cerasiforme, CHE = S. cheesmaniae, CHI = S. chilense, CHM = S. chmielewskii, COR = S. corneliomulleri, GAL = S. galapagense, HAB = S. habrochaites, HUA = S. huaylasense, NEO = S. neorickii, PEN = S. pennellii, PER = S. peruvianum, PIM = S. pimpinellifolium. Pagel's lambda and p-values are from a phylogenetically controlled generalized least-squares (PGLS) model with lambda = "ML". The majority of accessions in the monophyletic colored-fruited clade were collected from warmer locations than the green-fruited accessions, although there were three accessions that deviated from this trend—one S. lycopersicum var. cerasiforme (BGV008189) collected at an unusually cold site for the species, as well as one S. habrochaites (LA2098) and one S. pennellii (LA1809) collected at warm sites near the equator at the northern edge of these species' ranges.
CHAPTER 2

VARIATION IN RIPE FRUIT VOLATILES ACROSS THE TOMATO CLADE: AN EVOLUTIONARY FRAMEWORK FOR STUDYING FRUIT SCENT DIVERSITY IN A CROP WILD RELATIVE

2.1 Abstract

**Premise:** The scents of volatile organic compounds (VOCs) are an important component of ripe fleshy fruit attractiveness, yet their variation across closely related wild species is poorly understood. Phylogenetic patterns in these compounds and their biosynthetic pathways offer insight into the evolutionary drivers of fruit diversity, including whether scent can communicate an honest signal of nutrient content to animal dispersers. We assessed ripe fruit VOC content across the tomato clade (*Solanum* sect. *Lycopersicon*), with implications for crop improvement since these compounds are key components of tomato flavor.

**Methods:** We analyzed ripe fruit volatiles from 13 species of wild tomato grown under common garden conditions. Interspecific variation in 66 compounds and their biosynthetic pathways was assessed in 32 accessions, with an accession-level phylogeny accounting for relatedness.

**Results:** Wild tomato species can be differentiated by their VOCs, with *Solanum pennellii* notably distinct. Phylogenetic conservatism exists to a limited extent. Major clade-wide patterns corresponded to divergence of the five brightly colored-fruited species from the nine green-fruited species, particularly for nitrogen-containing compounds (higher in colored-fruited) and esters (higher in green-fruited), the latter appearing to signal a sugar reward.
**Conclusions:** We established a framework for fruit scent evolution studies in a crop wild relative system, showing that each species in the tomato clade has a unique VOC profile. Differences between color groups align with fruit syndromes that could be driven by selection from frugivores. The evolution of colored fruits was accompanied by changes in biosynthetic pathways underlying esters and nitrogen-containing compounds, volatiles important to tomato flavor.

2.2 Introduction

Plants produce a vast array of secondary metabolites that perform many important functions. For fleshy-fruited plants, the scents of volatile organic compounds (VOCs) in ripe fruits are generally thought to help attract seed dispersers by making mature fruits noticeable and appealing to animal consumers, while also aiding in defense against pathogens and seed predators (Rodriguez et al. 2013; Nevo and Ayasse 2019). Hundreds of different VOCs can be present in fleshy fruits, although this array of compounds is produced by relatively few main biosynthetic pathways (Jiang and Song 2010). From the human perspective, VOCs are key components of fruit flavor in agricultural crops (Klee and Tieman 2018). Assessing variation in these compounds across species is valuable to evolutionary biologists studying the drivers of fruit diversity as well as plant breeders working to improve fruit flavor.

Researchers have been exploring the evolutionary forces behind fruit trait diversity for decades, but have only recently begun identifying and quantifying the VOCs comprising wild fruit scent (Nevo and Ayasse 2019). Fruit-eating animal preferences have long been considered an important driver, because plants better at attracting fruit
consumers improve their seed dispersal (Ridley 1930; van der Pijl 1969; Janson 1983; Valenta and Nevo 2020). In line with this hypothesis, a number of studies have shown that VOCs are important for attracting non-human seed dispersers and signaling fruit ripeness (Nevo and Ayasse 2019). Evidence that fruit scent chemicals evolved in tandem with animal disperser consumption has come from recent work on frugivorous bats (Hodgkison et al. 2013; Santana et al. 2021) and lemurs (Nevo et al. 2018b). A mechanism underlying this fruit-frugivore coevolution could be that fruit scent communicates an honest signal of internal nutrient content to potential animal consumers (Nevo et al. 2019; Nevo et al. 2020a). At the same time, it is important to consider the possibility that some fruit VOCs serve a defensive rather than attractive function (Nevo et al. 2017). Additionally, fruit scent may be constrained by non-adaptive chemical, developmental, or phylogenetic limitations on VOC production (Nevo et al. 2020b). The degree of phylogenetic constraint, or conservatism (i.e., similarity due to inheritance from common ancestors) appears to differ depending on the system, with some multi-genera community-based studies not finding much conservatism (Nevo et al. 2018; Nevo et al. 2020b) and other narrower lineage-focused studies either finding it in Ficus (Hodgkison et al. 2013) or not finding it in Piper (Santana et al. 2021). Despite recent interest in wild fruit VOCs, the field is still young and data only exists for a small number of non-agricultural systems.

The VOCs of agricultural fruits have received considerable attention due to the economic importance of fruit flavor (Klee and Tieman 2018), but the wild relatives of these crop fruits have been far less studied and usually only in the context of domestication or crop improvement. A crop wild relative system has yet to be surveyed
as a framework for studying wild fruit scent evolution, presenting an exciting opportunity to understand the evolutionary drivers of VOC composition in a system with direct relevance for agriculture. Furthermore, few studies of interspecies VOC variation have employed a common garden (Schwinning et al. 2022) in which all plants are grown under the same conditions to ensure that differences are due mainly to genetics rather than environment. This controlled approach is particularly powerful in a crop relative system because the genetic, biosynthetic, developmental underpinnings of fruit VOCs are better studied than in most wild systems.

One system well-positioned to fill this gap in wild fruit scent evolution studies is the tomato clade, a group of 14 species (Solanum sect. Lycopersicon, family Solanaceae, Peralta et al. 2008) composed of the cultivated tomato (Solanum lycopersicum var. lycopersicum) and 13 species of wild relatives (Fig. 2.1). The clade is a monophyletic group estimated to have diverged from a common ancestor 2–2.5 million years ago based on fossil-calibrated molecular clock evidence (Särkinen et al. 2013; Pease et al. 2016). The 13 wild tomato species inhabit a range of environments in western South America (Appendix B, Fig. B1), from deserts to forests and coastlines to high mountains. Two informal sub-groups of species within the clade are generally recognized based on ripe fruit color: 1) a monophyletic colored-fruited group consisting of the red Solanum pimpinellifolium, red Solanum lycopersicum var. cerasiforme, orange Solanum galapagense, and yellowish/orange Solanum cheesmaniae, and 2) a paraphyletic green-fruited group encompassing the remaining nine species. Wild tomatoes offer a valuable source of genetic diversity for crop improvement, as most species can be interbred with each other and with cultivated tomatoes (Baek et al. 2015). The tomato system has
received much attention as a model of fruit development and genetics (Kimura and Sinha 2008), enabling a more in-depth consideration of the molecular basis of fruit VOCs than is possible with most other fruit systems. In the cultivated tomato, more than 400 VOCs have been detected (Petro-Turza 1986; Baldwin et al. 2000), although only about 20 of these compounds have been identified as major contributors to human flavor preferences (Martina et al. 2021). The four main categories of VOCs in tomatoes are based on the precursor molecules lipids, carotenoids/terpenes, non-aromatic amino acids (leucine and isoleucine), and phenylalanine (Colantonio et al. 2022); these categories are consistent with those commonly used for fleshy fruit VOCs in general (Rodríguez et al. 2013; Nevo and Ayasse 2019). The ~20 compounds considered to have a strong impact on cultivated tomato flavor come from across the four precursor categories and impart aromas including green/viney/earthy, nutty/stale, and fruity/floral/sweet/tropical (Baldwin et al. 2008). In terms of chemical classification, VOCs can also be grouped into esters, alcohols, aldehydes, ketones, lactones, and terpenoids (Jiang and Song 2010). Among these chemical classes, esters are particularly important for attracting animal dispersers to wild fruits due to their fruity/floral aromas (Rodríguez et al. 2013), which are especially enticing to scent-oriented mammals (Peris et al. 2017) including lemurs (Nevo et al. 2018) and elephants (Nevo et al. 2020). Interestingly, while humans respond positively to ester aromas in certain crops such as apple, banana, citrus, and strawberry (Jiang and Song 2010), esters have a negative impact on human flavor preference in cultivated tomatoes (Klee and Tieman 2018). Previous work on three of the nine green-fruited wild tomato species showed they produced higher percentages of acetate esters than the two
red-fruited species, with intermediate levels produced in the two yellow/orange species (Goulet et al. 2012). However, a survey of ripe fruit VOCs across all 13 wild species has not yet been carried out. Given this preliminary evidence of interspecies differences in disperser-relevant scent compounds, establishing a complete evolutionary framework of VOC variation for all species in the clade would enable hypothesis testing without making assumptions about how often evolutionary changes occurred.

In this study, we present a novel framework for wild fruit scent evolution research by surveying variation in ripe fruit volatiles across all 13 wild tomato species in a common garden. The combination of a recently diverged, phylogenetically tight clade and a common growing environment distinguishes this study from many others that used a phylogenetically broader community approach and a wild setting. By quantifying variations in individual VOCs and their categories across the tomato clade, our objectives were to: (1) examine which compounds and biosynthetic pathways are most important for differentiating species and determine the extent of phylogenetic conservatism in the clade, (2) assess whether fruit scent variation aligns with two previously hypothesized wild tomato fruit syndromes (colored-fruited vs. green-fruited) that could have been driven by disperser preferences (Barnett et al. 2023a), and (3) explore whether scent could communicate an honest signal to animal dispersers by testing for an association between VOCs and fruit sugar content. We predicted each species would display a distinct VOC profile and that the colored-fruited group would differ from the green-fruited group, particularly for disperser-relevant ester compounds and VOCs important to human-perceived flavor, as could be expected if animal preferences were an evolutionary driver of fruit trait diversity.
2.3 Methods

2.3.1 Study species and plant material

Our sampling included 32 accessions chosen to span the known phylogenetic and geographic diversity across the tomato clade (Solanum sect. Lycopersicon), consisting of 14 taxa: 12 species of fully wild relatives, the semi-wild S. lycopersicum var. cerasiforme, and the cultivated tomato variety 'Ailsa Craig' (Fig. 2.1; Appendix B, Table B1). Three accessions of each taxa representing different parts of their geographic ranges were initially selected and planted (Appendix B, Fig. B1), but not all accessions produced enough fruit for VOC quantification; thus for some taxa only one or two accessions were measured. Seeds were obtained from the C. M. Rick Tomato Genetics Resource Center at the University of California, Davis, USA (TGRC, http://tgrc.ucdavis.edu) and the Universitat Politècnica de València, València, Spain (accession IDs and collection locations listed in Appendix B, Table B1). Each accession is representative of an independently sampled population in the wild. For autogamous (self-pollinating) self-compatible species (S. cheesmaniae, S. galapagense, some S. pimpinellifolium, S. lycopersicum var. cerasiforme, Solanum neorickii) each plant of an accession is likely genetically identical. For others, plants are not identical, as the facultative self-compatible or allogamous self-incompatible accessions are maintained through "mass sibling" pollination in germplasm centers.

Seeds were soaked in a 2.7% sodium hypochlorite solution (as per the TGRC protocol) and germinated in a greenhouse at the University of Massachusetts Amherst. After eight weeks, seedlings were transplanted into a high tunnel greenhouse at the
University of Massachusetts Crop and Animal Research and Education Farm in South Deerfield, MA, where they were irrigated twice a week and fertilized once a week via drip lines. For each accession, three separate individual plants were grown in different randomized locations. Our study plants grew throughout the summer farm season (May–September 2020) and were pruned as needed. Self-incompatible accessions were hand-pollinated several times during the first few weeks of flowering to facilitate fruit set, and all flowers were accessible to natural pollinators.

2.3.2 Fruit volatile collection

Ripe fruits were collected on three separate days (August 3rd, August 31st, and September 14th, 2020), packed in bubble wrap in ventilated plastic clamshell containers, and shipped overnight from Massachusetts to the University of Florida for VOC analysis. While still in Massachusetts, fruit ripeness was determined qualitatively for each species based on a combination of fruit or calyx color change, softening, and/or falling off at the touch (Appendix B, Table B2). For most accessions, fruits from several plants were pooled together to get as close as possible to 100 grams of fruit per sample, although for some small-fruited samples such as *S. galapagense* the total weight was as low as 8 grams. A total of 66 samples representing 14 taxa and 32 accessions were collected (Appendix B, Table B3).

2.3.3 GC-MS quantification of VOCs

Samples were processed at the University of Florida and analyzed with gas chromatography-mass spectrometry (GC-MS) for compound identification as detailed in
(Tieman et al. 2006). They were also analyzed on a gas chromatograph equipped with a flame ionization detector for quantification. Briefly, air filtered through a hydrocarbon trap flowed over chopped fruit enclosed in glass tubes for one hour with the aid of a vacuum pump for volatiles collection on a Super Q column. The trapped volatiles were then eluted with methylene chloride after adding nonyl acetate as an internal standard, and analyzed on an Agilent (Palo Alto, California, USA) 6890N gas chromatograph equipped with an Agilent 30m DB-5 column. Peaks were quantified using Agilent GC ChemStation Rev. A.10.02 [1757] software based on retention times compared to known standards. Compound identity was confirmed for each accession via mass spectrometry using Agilent MSD Productivity ChemStation v.D.03.xx software and the NIST05 MS Library database / MS Search Program v.2.0d.

A total of 66 compounds with clear and consistent peaks in most samples were chosen to be quantified (Appendix B, Table B4). Most of these compounds were known to be present in cultivated tomatoes, although we also included additional compounds not usually found in cultivated tomatoes that were identified in multiple wild species samples. Peak areas were adjusted relative to the nonyl acetate standard and converted to units of nanograms/grams of fruit weight/hour (ng/gfw/hr).

Compounds were categorized based on their precursor molecules as in (Colantonio et al. 2022). For each category of compounds, ng/gfw/hr values for all the compounds within a category were added together to get a category sum. Category percentages were computed by dividing the category sum by the sum of all 66 compounds.
2.3.4 Data analysis

Analyses were carried out in R v4.2.2 (R Core Team 2022) and RStudio Desktop v2022.12.0.353 (Posit team 2022), with plots produced using the R packages GGPLOT2 v3.4.0 (Wickham 2016) and PHEATMAP v1.0.12 (Kolde 2019).

Differentiating species and assessing phylogenetic patterns

We performed a principal components analysis (PCA) on centered and scaled accession mean values of the 66 VOC compounds using the function `prcomp` from the R package STATS, plotting the first two PC axes with GGPLOT2.

To determine which VOCs were most important for differentiating species, we quantified the Random Forest mean decrease accuracy for all 66 compounds using MetaboAnalyst 5.0 (Pang et al. 2021). Data were log (base 10) transformed, but data scaling and sample normalization were not used. Only 61 of 66 samples (representing 11 of 14 species) were used for this analysis because *Solanum chilense*, *Solanum huaylasense*, and *Solanum peruvianum* had fewer than three samples. In addition to conducting the analysis with all 11 species included, the same quantifications were made for the five colored-fruited species only and the six green-fruited species only to determine which compounds were top species differentiators within each of the two color groups.

We used an accession-level phylogenetic tree from Barnett et al. (2023a) to place VOC variation in a phylogenetic context. To quantify the degree of phylogenetic conservatism in VOC compounds and categories, we used the function `phyloSignal` from the R package PHYLOSIGNAL v1.3 (Keck et al. 2016) to calculate Blomberg's K.
(Blomberg et al. 2003) and Pagel's λ (Pagel 1999) from accession mean values of each compound or category.

To test the difference in ester content between colored-fruited and green-fruited accessions, phylogenetic generalized least squares (PGLS) models with color group as the predictor variable and ester sum (or percent) as the response variable were produced with the functions `comparative.data` and `pghs` from the R package `CAPER` v1.0.1 (Orme et al. 2013) with lambda="ML" to simultaneously optimize phylogenetic signal and regression parameters as recommended by (Revell 2009).

**Associations between VOCs and fruit sugar content**

We examined potential associations between VOCs and nutrient content using sugar data from Barnett et al. (2023a). Briefly, concentrations (in mg/mL) of glucose, fructose, and sucrose were quantified from liquid extracts of ripe fruit pericarps using absorbance-based assay kits (Megazyme, Bray, Ireland); "total sugars" represents the sum of those three concentrations. Raw/untransformed accession mean values for VOC categories and sugars were used for analyses in R. Plots were produced with `GGPLOT2`, and Pearson's $r$ was calculated with the built-in R function. PGLS models, with VOC category (ester, terpene, or apocarotenoid) sum (or percent) as the predictor variable and total sugars as the response variable, were produced with `CAPER` with lambda = "ML".

2.4 Results

2.4.1 Extent of variation in VOCs across the tomato clade

To establish a framework for studying fruit scent evolution across the tomato clade, we quantified 66 of the VOCs found in fruit from *S. lycopersicum* var.
lycopersicum and 13 related wild tomato species. In order to assess how different biosynthetic pathways affected patterns of variation in these compounds, we categorized VOCs by their biosynthetic precursors: 31 lipid-derived, 7 carotenoid/terpene-derived, 12 non-aromatic amino acid (leucine and isoleucine)-derived, 10 phenylalanine-derived, and 6 other (Appendix B, Table B4). Separately we labeled 17 of the 66 compounds as esters since multiple precursor types (lipid, non-aromatic amino acid, phenylalanine, and other) can be esterified. Esters are a group of interest because they tend to have fruity or floral scents and are known to be attractive to vertebrate dispersers (Peris et al. 2017), so may be subject to selective pressure from animals. Also of note from a human perspective, fifteen of the compounds we quantified (Appendix B, Table B5) are considered key components of cultivated tomato aroma (Martina et al. 2021); we did not quantify all 21 on that list because not all of those compounds had clear, consistently detectable peaks in our GC-MS samples. Compounds not usually found in cultivated tomatoes that showed clear peaks in multiple wild species samples were the non-aromatic amino acid-derived compounds acetoin and 2-propylthiazole; the terpene-derived compounds alpha-pinene, alpha-curcumene, and zingiberene; and the other-derived compound 2-pentylfuran.

Across all 66 samples and all 66 compounds, amounts (in ng/gfw/hr) varied from not detected (a total of 53 compounds had a 0 value in at least one sample) to 688.54 (zingiberene) (Appendix B, Table B3). Note that for zingiberene, Solanum habrochaites LA2329 was the only accession with such high levels; the 2nd highest accession was S. cheesmaniae LA0428 with 24.1 ng/gfw/hr. Besides zingiberene, the compounds with the next three highest maximum values were 3-methyl-1-butanol (614.73 in S. galapagense LA0528), 2-methyl-1-butanol (532.72 S. neorickii LA1626), and hexyl alcohol (410.74 in
S. neorickii LA2200); all three of these compounds have odors described as fusel, alcoholic, and sweet (Appendix B, Table B4). On the other end of the spectrum, the compounds with the three lowest maximum values (in ng/gfw/hr) were 2-ethyl-1-hexanol (0.669 in Solanum arcanum LA0441), phenethyl acetate (1.489 in S. galapagense LA0528), and 3-methyl-2-butenal (1.773 in S. peruvianum LA1474).

Only 13 of the 66 compounds were detected in all 66 samples; 12 of them were either lipid or non-aromatic amino acid derived, and none of them were esters. Included in these 12 were the four least variable compounds in terms of coefficient of variation (CV): 3-pentanone (12.565 mean ng/gfw/hr, 0.622 CV), 2-methylbutyraldehyde (21.941 mean ng/gfw/hr, 0.669 CV), 1-penten-3-ol (9.362 mean ng/gfw/hr, 0.734 CV), and hexanal (66.965 mean ng/gfw/hr, 0.921 CV). Compound summary statistics are listed in Appendix B, Table B4.

2.4.2 Compounds and biosynthetic pathways differentiating species and the extent of phylogenetic conservatism

To visualize how well species could be differentiated by levels of the 66 compounds, we performed a PCA with sample averages for the 32 accessions as points, drawing color-coded polygons around accessions from the same species (Fig. 2.2; Appendix B, Fig. B2). Most species occupied a distinct area of the PCA space, with some overlap among the green-fruited Solanum chmielewskii, S. neorickii, S. arcanum, Solanum corneliomulleri, and S. chilense. Noticeably separate from all other species was Solanum pennellii.
To determine which individual VOCs were most important for differentiating species, we quantified the Random Forest mean decrease accuracy for all 66 compounds. A total of 17 compounds had scores above 0.01 (Appendix B, Table B6; Fig. 2.3) and were thus considered major differentiators among species; these compounds included representatives from all five precursor molecule categories, seven of them were esters, and one was nitrogen-containing.

When the sum of all 66 compound levels was calculated for each species, *S. neorickii* and *S. pennellii* had the two highest totals (Fig. 2.4); interestingly, these two species are not closely related. On the other end of the spectrum, the more closely related *S. huaylasense* and *S. peruvianum* produced the lowest sum totals. All four of these species are green-fruited; the five monophyletic colored-fruited species all had intermediate sums.

To explore whether particular biosynthetic pathways were important drivers of VOC differences across the clade, we compared each compound category's sum (total amount in ng/gfw/hr) and percentage (category sum divided by sum of all compounds) among species. Most species showed a similar partitioning of VOC categories in terms of sums and percentages (Fig. 2.4), with lipid-derived and non-aromatic amino acid-derived being the two major categories in all species. *Solanum habrochaites* was unique for its high level of carotenoid/terpene compounds, mainly driven by very high amounts of the sesquiterpenes alpha-curcumene and zingiberene in one accession (LA2329).

In terms of esters (a category that includes compounds from multiple precursor categories), *S. huaylasense* stood out for having the highest percentage of esters, despite producing the lowest total VOC sum of any species (Fig. 2.4). *S. pennellii* was notable
for high values of both sum and percentage of esters. Interestingly, *S. neorickii* had the lowest percentage of esters for a green-fruited species while its sister species *S. chmielewskii* had one of the highest percentages. Among colored-fruited species, the two Galápagos Islands endemics *S. galapagense* and *S. cheesmaniae* produced more esters than the red-fruited mainland species.

The similarities in distantly-related species and differences in closely-related species suggest that ripe fruit scent does not show much phylogenetic conservatism. Phylogenetic signal results also support the conclusion that conservatism is limited, because only five of 66 compounds had strong (> 0.7) and 12 of 66 compounds had moderate (>0.4 and <0.7) values for both Blomberg’s K and Pagel’s lambda (Appendix B, Table B4). Interestingly, four of the five high-signal and five of the 12 moderate-signal compounds were esters. Of the seven compound categories tested, only “ester” and “other” had at least moderate (>0.4) values for both K and lambda.

2.4.3 Associations between VOCs and putative dispersal syndromes

There were stark differences between the colored-fruited and green-fruited species groups for a subset of the species-differentiating compounds described above, in line with the putative fruit syndromes known to distinguish these two groups for other fruit traits (Barnett et al. 2023a). In particular, the nitrogen-containing 2-isobutyrlthiazole was high in colored-fruited but low in green-fruited species, while the esters methyl caproate and ethyl caproate as well as the lipid-derived *trans*-2-nonenal were low in colored-fruited but high in green-fruited species (Fig. 2.3).
We quantified Random Forest mean decrease accuracy separately for just the colored-fruited and just the green-fruited species to determine if certain compounds were more important for differentiating species within each group. These additional analyses revealed that valeraldehyde (lipid-derived), cis-3-hexenal (lipid-derived) and methyl salicylate (phenylalanine-derived) were the top three differentiators among colored-fruited species only, while 1-pentanol (lipid-derived), trans-2-nonenal (lipid-derived), and 2-methyl-1-butanol (non-aromatic amino acid-derived) were the top three for green-fruited species only (Appendix B, Table B6). The only compound that was a major differentiator within both color groups was 3-pentanone (lipid-derived). Among the colored-fruited species, S. galapagense stood out with high levels of alpha-pinene, ethyl propionate, and propyl acetate. Notable green-fruited species were S. chmielewskii with high levels of trans-2-pentenal, methyl caproate, ethyl caproate, methyl butyrate, and valeraldehyde; S. neorickii with high levels of methyl caproate, ethyl caproate, trans-2-nonenal, methyl disulfide, methyl butyrate, 2-methyl-1-butanol, and alpha-pinene; and S. pennellii with high levels of ethyl caproate, trans-2-nonenal, isobutyl acetate, and propyl acetate.

For a finer-grained look at interspecies patterns among biochemical precursor categories, we also examined how individual compounds within each category or biosynthetic pathway varied across accessions (Appendix B, Fig. B3). Apocarotenoids were generally higher in colored-fruited accessions, while terpenes were higher in green-fruited accessions (Appendix B, Fig. B3a). Among non-aromatic amino acid-derived compounds, the nitrogen-containing 2-propylthiazole and 2-isobutylthiazole were mostly absent from green-fruited accessions (Appendix B, Fig. B3b); these two compounds are
also largely absent from fruits other than tomatoes (Liscombe et al. 2022). The most abundant lipid-derived compounds across the clade contained five or six carbons, while longer chain compounds such as octanal, nonanal, and trans-2-nonenal were present in higher levels in green-fruited species (Appendix B, Fig. B3d). Most phenylalanine-derived compounds were not very abundant across the clade, with guaiacol and salicylaldehyde completely absent from the green-fruited species (Appendix B, Fig. B3e).

Esters were generally highest in green-fruited accessions, with some nuance for compounds such as cis-3-hexenyl acetate, phenethyl acetate, ethyl caprylate, and hexyl hexanoate—these were absent from most green-fruited species but high in two or three green-fruited species (Appendix B, Fig. B3f).

Overall, green-fruited species produced higher total sums and percentages of esters compared to the colored-fruited species (Fig. 2.5). Phylogenetic generalized least squares (PGLS) models that tested the difference between colored-fruited and green-fruited groups did not show significant p-values when all species were included (P = 0.29, lambda = 0.822 for esters sum; P = 0.17, lambda = 0.767 for esters percent; 30 degrees of freedom). However, when the two accessions of S. pennellii were removed, the p-values became highly significant (P = 0.009, lambda = 0 for esters sum; P = 0.002, lambda = 0 for esters percent; 28 degrees of freedom); this difference in p-values was due to S. pennellii's high ester measurements relative to its sister species S. habrochaites.

Five of the compounds we quantified were nitrogen-containing (Appendix B, Fig. B3g) and were mainly present only in colored-fruited (particularly red) accessions, with a few interesting exceptions. 3-methylbutanenitrile (isovaleronitrile) was present at low levels in almost all green-fruited accessions, but the other four compounds were mostly...
absent from the green-fruited accessions. Interestingly, the green-fruited *S. corneliomulleri* LA1945 produced appreciable levels of four of the five nitrogen-containing compounds and was the only green-fruited accession in which we detected 1-nitro-2-phenylethane. *S. chilense* LA2750 was also notable for being the only green-fruited accession to produce fairly high levels of both 2-isobutylthiazole and 2-propylthiazole. Also noteworthy is the lack of 2-propylthiazole, 1-nitro-2-phenylethane, and benzyl cyanide in the orange/yellow-fruited *S. galapagense* and *S. cheesmaniae* relative to their red-fruited sister species, suggesting that the loss of red fruit color on the Galápagos Islands may be correlated with decreased production of those three nitrogen-containing compounds.

Inter-accession patterns of variation in the 15 compounds important to cultivated tomato human flavor preference (Appendix B, Table B5) are shown in Fig. 2.6. The carotenoid-derived compounds geranylacetone and 6-methyl-5-hepten-2-one are more abundant in the colored-fruited species compared to the green-fruited species, which is not surprising since red/orange/yellow fruit color is due to carotenoid accumulation. Other compounds more abundant in colored-fruited relative to green-fruited species were 2-isobutylthiazole, 1-nitro-2-phenylethane, and guaiacol. The compound *cis*-3-hexenal (the most abundant VOC in cultivated tomato) showed an interesting pattern of being highest in the three red-fruited species as well as *S. habrochaites* and *S. pennellii*, but low in the other green-fruited species (Appendix B, Fig. B4).

Taken together, these phylogenetic patterns revealed that many of the major differences in VOCs and their biosynthetic pathways across the tomato clade are associated with the divergence of the five monophyletic colored-fruited species from the
remaining nine green-fruited species, with *S. pennellii* particularly distinct from all other species. Interestingly, the divergence of the two color groups is also associated with differences in fruit sugar type and malic acid levels that may represent dispersal syndromes (Barnett et al. 2023a). In general, the color-fruited group has evolved lower levels of esters and higher levels of nitrogen-containing volatiles relative to the rest of the clade, although there are a few interesting exceptions. Decreased ester production appears to have evolved independently at least once within the green-fruited species in *S. neorickii*. Increased nitrogen-containing VOC production may have arisen independently in the green-fruited *S. corneliomulleri* and *S. chilense*, while the production of some nitrogen-containing compounds appears to have been lost in the yellow/orange-fruited *S. galapagense* and *S. cheesesmaniae*.

2.4.4 Associations between VOCs and fruit sugar content

To explore whether fruit scent may communicate an honest signal about fruit nutrient content to animal dispersers, we examined the association between VOCs and sugars. For scent we focused on the ester, terpene, and apocarotenoid VOC categories because all three can be important for attracting wild animal dispersers (Rodriguez et al. 2013; Nevo and Ayasse 2019). We quantified nutrient content as "total sugars": the summed concentrations of glucose, fructose, and sucrose, the three main sugar types in tomato (Beckles et al. 2012). These sugars comprise the fruit's main caloric reward since tomatoes are low in protein and lipids (García-Alonso et al. 2020). We hypothesized that scent-based signaling would be more evident in green-fruited tomato species compared to colored-fruited ones, because the former may depend on smell to attract animal dispersers.
to their visually inconspicuous fruits. Other plants with dull-colored fruits known to attract animals via scent include some species of *Piper* (e.g., Santana et al. 2021) and *Ficus* (e.g., Lomáscolo et al. 2010).

We found that accessions very high in esters were moderately high in total sugars, while accessions low in esters can range from very low total sugars to very high total sugars (Fig. 2.7; Appendix B, Fig. B5). Thus a fruit with a strong ester scent does appear to signal the presence of a sugar reward. In contrast, a fruit without a strong ester scent does not seem to provide any information about sugar content, as low ester accessions can either be high or low in sugar. Consistent with our hypothesis, all of the accessions high in esters (>15%) were from green-fruited species, particularly *S. chmielewskii* and *S. pennellii*. Interestingly, there was notable within-species variation in ester levels for most green-fruited species other than *S. pennellii*; for example *S. habrochaites* had one high, one medium, and one low ester accession.

We did not find any clear association between terpenes and total sugars (Appendix B, Fig. B6). The accession *S. habrochaites* LA2329 was a lone outlier with very high levels of terpenes. With this accession excluded, there was still no clear trend, as the remaining accessions that were high in terpenes displayed both high and low total sugar levels.

We did not find evidence of an association between apocarotenoids and sugars (Appendix B, Fig. B7). We were only able to quantify two carotenoid-derived compounds for this study, and these were almost exclusively detected in colored-fruited species, so only those 12 accessions were included in this analysis. The red-fruited *S. pimpinellifolium* accessions were generally high in both apocarotenoids and total sugars.
Both *S. galapagense* and *S. cheesmaniae* were low in apocarotenoids, but the former was high in sugars while the latter was low in sugars.

2.5 Discussion

Our results represent the first clade-wide survey of ripe fruit volatiles across cultivated tomato and its 13 species of wild relatives, establishing an evolutionary context for the patterns of variation in these aroma-producing compounds and their biosynthetic pathways. By growing all plants in a common garden, we ensured that interspecies differences were mainly due to genetics rather than environment. We found that species can be differentiated by their ripe fruit VOC profiles and that phylogenetic conservatism only exists to a limited degree, with many of the major clade-wide patterns corresponding to the divergence of the five monophyletic colored-fruited species from the remaining nine green-fruited species, and *S. pennellii* quite distinct from all other species.

Phylogenetic conservatism in fruit scent only exists to a limited extent within the wild tomato clade, with a number of similarities in distantly-related species and differences in closely-related species. The VOCs that do show some evidence of conservatism are mainly associated with the split of the monophyletic colored-fruited group from the paraphyletic green-fruited group. The low levels of esters and high levels of nitrogen-containing compounds in the color-fruited group could be explained by shared genetic changes inherited from a common ancestor; however it is difficult to disentangle this phylogenetic inertia from stabilizing selection due to similar ecological conditions such as similar disperser preferences. There does not seem to be as much phylogenetic conservatism within the green-fruited group; for example, *S. chmielewskii*
has high ester levels consistent with many other green-fruited species while its sister species *S. neorickii* has quite low ester levels (Fig. 2.4).

The differences between colored-fruited and green-fruited species support the existence of two wild tomato fruit syndromes that could be linked to disperser preferences (Barnett et al. 2023a). Esters and nitrogen-containing compounds showed particularly notable contrasts between colored-fruited and green-fruited species, suggesting that the evolution of colored fruits was accompanied by changes in those two biosynthetic pathways. For example, the nitrogen-containing 2-isobutylthiazole was mainly produced only in colored-fruited species, while the ester methyl caproate showed high levels only across green-fruited species. There were a few interesting exceptions to these trends, however, and compounds from all categories of molecules varied in ways that contributed to each species' unique aroma profile. In addition, high ester levels in green-fruited species appeared to signal the presence of a sugar reward, although low ester levels were found in fruits with both high and low sugar content.

Our findings that green-fruited species can be differentiated by VOC profiles and that many of these species produce high levels of esters are noteworthy because it suggests these species can use scent to attract animal dispersers to their visually inconspicuous fruits. According to the dispersal syndrome hypothesis (Valenta and Nevo, 2020), green fruits are thought to be mainly dispersed by mammals rather than birds because many mammals have an acute sense of smell and poor color vision. Ester compounds have been shown to attract scent-oriented mammals (Peris et al. 2017), which could be the selective force behind the high levels of esters in some wild tomato species. We confirmed that colored-fruited tomato species produced significantly lower
percentages of esters than green-fruited species, consistent with the hypothesis that more brightly-colored fruits evolved to primarily attract birds that have excellent color vision but a poor sense of smell. A possible genetic basis of this difference in ester production between colored- and green-fruited species has been identified; a retrotransposon insertion in the promoter of the SlCXE1 carboxylesterase gene is found in all four colored-fruited species but none of the green-fruited species that have thus far been sequenced (S. chmielewskii, S. neorickii, S. habrochaites, and S. pennellii) (Goulet et al. 2012). That change in the promoter was shown to increase gene expression, leading to higher levels of esterase enzyme and thus more breakdown of esters. Apart from limited anecdotal reports, however, little is known about the animal dispersers of wild tomatoes in their native ranges in South America (Barnett et al. 2023a), so ecological field studies of these species would be an exciting opportunity to test the hypothesis that animal preferences may have driven fleshy fruit evolution.

Disperser attraction might explain the particularly unique VOC profile of S. pennellii compared to the rest of the clade. This species appears to produce the strongest scent of any wild tomato species, with the highest total sum of compounds. Many of these compounds are fruity-smelling esters; S. pennellii displayed the highest sum and second highest percentage for the ester category, with particularly high levels of isobutyl acetate and propyl acetate compared to other species. Interestingly, it produced more cis-3-hexenal (the most abundant VOC in cultivated tomato, with a green/grassy odor) than any other wild species. The environments inhabited by S. pennellii are distinct from the rest of the clade, with relatively high annual average temperature (the highest among green-fruited species) and very low annual precipitation (second lowest in the clade.
behind *S. chilense* (Nakazato et al. 2010; Ramírez-Ojeda et al. 2021). Perhaps dispersers are more spread out in these desert habitats and need to be attracted from great distances, while relatively warm conditions enable the production of strong scents in part because biosynthetic enzyme kinetics are faster at higher temperatures, as has been shown for floral VOCs (Farré-Armengol et al. 2014). VOCs would also be released more quickly at high temperatures since they would be more volatile.

It is worth noting that our VOC measurements were made on chopped fruits, which likely have different VOC emissions than intact fruits, as has been shown in cultivated tomatoes (Baldassarree et al. 2015). While future studies measuring VOCs from intact wild tomato fruits (as well as the leaves) would provide a more complete picture of the scent profiles experienced by frugivores in nature, we think chopped fruits are still relevant ecologically for two reasons. First, animals that do not swallow the fruit whole would break the fruit when chewing/tasting it and presumably then decide whether to eat more and disperse the seeds. Second, ripe or overripe fruits may drop off the plant and break open, releasing volatiles that attract animals to the plant. Anecdotally, we observed fallen fruits under the plants of many wild species (particularly *S. arcanum*, *S. chilense*, *S. chmielewskii*, *S. galapagense*, *S. habrochaites*, *S. neorickii*, and *S. pennellii*), and noticed distinctive scents from most of the wild species while collecting fruits.

In regards to whether scent provides an honest signal of nutrient content, our finding that accessions with high ester levels are also high in sugars is consistent with previous studies of animal-dispersed fleshy fruits in Madagascar (Nevo et al. 2019; Nevo et al. 2022) that found a positive correlation between ester and sugar levels. This may be due to a yet undetermined biochemical link between fruit sugars and ester production.
Others have proposed that alcohols, which are necessary for ester synthesis, may accumulate in fruits with high sugar—for example, ethanol can be produced by microbial fermentation of sugars (Nevo et al. 2019). Fermentation seems unlikely to explain the high ester emissions of wild tomatoes, however, because the fruits we measured were not overripe and several of the most abundant esters (e.g., isopentyl acetate, propyl acetate, isobutyl acetate) are not derived from ethanol. While wild tomato fruits with a strong ester scent do appear to be an honest signal of high sugar reward, it is worth noting that fruits with less ester-driven scents can have a wide range of sugar contents. Whether these low-ester fruits with high sugar contents have a different way of signaling to dispersers is an intriguing question.

Our results that the majority of green-fruited species produce almost no nitrogenous volatiles were consistent with expectations (Liscombe et al. 2022), although we were surprised to discover that some activity in the pathway may have independently evolved in S. corneliomulleri LA1945 and S. chilense LA2750. These accessions may be worth investigating for variations in the SITNH1 gene recently shown to underlie the biosynthesis of many of the nitrogen-containing VOCs in cultivated tomato fruits (Liscombe et al. 2022). The potential ecological function of nitrogen-containing VOCs in wild tomato fruits also merits further study; while herbivore damage has been shown to elicit production of nitrogenous volatiles in leaves of some plants (van den Boom et al. 2004), these compounds are not commonly emitted by fruits (Nevo and Ayasse 2019).

Interestingly, humans have selected against esters and in favor of nitrogenous volatiles over the course of domestication. The fruity/floral aromas of esters are negatively associated with cultivated tomato flavor (Klee and Tieman 2018), while
several nitrogen-containing volatiles contribute positively to flavor (Liscombe et al. 2022). This trend toward lower ester and higher nitrogenous VOC levels in red-fruited species appears to have begun before humans started eating tomatoes, however, as our findings show this pattern present in the wild red-fruited *S. pimpinellifolium*. Intriguingly, the two yellow/orange-fruited Galápagos species may be reverting back towards the green-fruited norm; *S. galapagense* and *S. cheesmaniae* produce more esters and fewer nitrogenous volatiles than their red-fruited sister taxa. These results are consistent with previous work showing low ester levels in *S. pimpinellifolium* and intermediate ester levels in the two Galápagos species (Goulet et al. 2012). Whether the decrease in esters is connected to the opposing trend in nitrogenous volatiles is an interesting question. It is possible these changes in colored-fruited VOCs were a by-product of early red-fruited plants evolving to attract bird dispersers, which likely rely on vision (Lomáscolo and Schaefer 2010) rather than scent when locating food. Whatever the original cause, human selection seems to have accelerated the red-fruited trend towards lower ester and higher nitrogenous volatile production.

From a modern human perspective, our VOC measurements reveal that several wild species could be worth exploring as potential germplasm sources for crop improvement due to their high levels of compounds positively associated with flavor: notable examples were *S. arcanum* and *S. corneliomulleri* for 2-phenyl ethanol; *S. cheesmaniae*, *S. habrochaites*, and *S. pennellii* LA2580 for phenylacetaldehyde; and *S. chmielewskii* for 1-penten-3-one. In addition, we tasted fruits from many of the accessions used in this study and compiled anecdotal notes on their flavors and aromas (Appendix B, Table B7).
2.6 Conclusions

The clade-wide patterns reported here provide a framework for future studies of wild tomato fruit VOCs in a crop relative system. Our survey identifies compounds underlying the rich variety of scents produced by these species, and provides clues as to the evolutionary context in which this variation arose. The wild tomato system is particularly powerful for investigating fruit evolution because cultivated tomatoes have been well studied as a model system for fleshy fruit development and genetics. Complementing this knowledge with field studies in these species' native ranges would be particularly informative, as the ecological context in which these fruits produce their scents is largely unknown. In addition, furthering our understanding of wild relatives would enhance plant breeders' efforts to improve fruit flavor.

2.7 Figures
Figure 2.1 Phylogenetic relationships among species in the wild tomato clade (*Solanum* sect. *Lycopersicon*), with *Solanum lycopersicoides* included as the outgroup. Maximum likelihood consensus tree based on 64,745 SNP markers obtained through genotyping-by-sequencing; constructed in IQ-TREE with 1000 ultrafast bootstraps using the TVM+F+ASC+R4 nucleotide substitution model selected by ModelFinder with ascertainment bias correction and SH-aLRT likelihood ratio test (Barnett et al. 2023a). Numbers at nodes represent bootstrap percentages. Three-letter species abbreviations and species color-coding are used throughout this paper. Photographs show ripe fruits of each species; three photographs of *Solanum cheesmaniae* are included to show the different colors of the three accessions measured in this study. Scale bar represents 10 millimeters.
Figure 2.2 Wild tomato species are largely differentiated by their volatile organic compound (VOC) profiles, as shown by a principle components analysis (PCA) of centered and scaled raw values (in ng/gfw/hr) for all 66 compounds. Produced with prcomp function in R. Points represent accession averages (32 total), color-coded by species (14 total) and labeled with 3-letter species abbreviations: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum NEO = Solanum neorickii, PEN = Solanum pennelli, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium. Points from the same species are connected to facilitate viewing of species differentiation. See Appendix B, Fig. B2 for a version of this figure with accession numbers labeled for each point.
Figure 2.3 Heatmap of the 17 compounds with the greatest contribution to differentiating species, based on Random Forest Mean Decrease Accuracy values greater than 0.01 (Appendix B, Table B6). See Appendix B, Table B4 for details on the compounds. Categories are based on the compound's precursor molecule; whether the compound is an ester is also indicated. Data values (in ng/gfw/hr) were log (base 10) transformed but not scaled. Columns represent accession averages, with color-coding along the top denoting species (see Fig. 2.1). Accessions are arranged phylogenetically based on a tree from Barnett et al. (2023a), with colored-fruited species on the left and green-fruited species on the right. Three-letter abbreviations denote species as follows: ARC = *Solanum arcanum*, CER = *Solanum lycopersicum* var. *cerasiforme*, CHE = *Solanum cheesmaniae*, CHI = *Solanum chilense*, CHM = *Solanum chmielewskii*, COR = *Solanum corneliomulleri*, GAL = *Solanum galapagense*, HAB = *Solanum habrochaites*, HUA = *Solanum huaylasense*, LYC = *Solanum lycopersicum* var. *lycopersicum*, NEO = *Solanum neorickii*, PEN = *Solanum pennellii*, PER = *Solanum peruvianum*, PIM = *Solanum pimpinellifolium*. Numbers after 3-letter species abbreviations are Tomato Genetics Resource Center (TGRC) 'LA' accession numbers, except for CER_B8189 (full accession ID BGV008189) which does not have an 'LA' number.
Figure 2.4 Biosynthetic pathways drive differences in volatile organic compounds (VOCs) across the wild tomato clade, as shown by stacked bar plots of compound categories based on either precursor molecules (top row), or ester/non-ester (bottom row), for 14 species abbreviated as follows: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum, NEO = Solanum neorickii, PEN = Solanum pennellii, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium. Species are arranged phylogenetically along the y-axis with cultivated tomato at the top (see Fig. 2.1). Species values were computed by first calculating accession averages from 66
samples and then averaging those accession averages (32 total). Left-hand plots show category sums of compound values in ng/gfw/hr. Right-hand plots show category percentages of the sum of all 66 compounds.

Figure. 2.5 Boxplots showing differences between colored-fruited and green-fruited species groups in the amount of ester volatile organic compounds (VOCs), displayed as total sum (in ng/gfw/hr) of the 17 ester compounds quantified (left plot), and percentage of the total sum of all 66 compounds (right plot), based on raw/untransformed accession mean values. Colored points represent accessions and are color-coded by species, abbreviated as follows: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum NEO = Solanum neorickii, PEN = Solanum pennellii, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium. All 32 accessions are included; sample sizes are 13 for the colored group and 19 for the green group. Phylogenetic generalized least squares (PGLS) models were produced from raw/untransformed accession mean values, with lambda = "ML", p-values were on 30 degrees of freedom. †Note that S. pennellii (PEN) had a strong effect on the PGLS models due to its high values; with the two accessions of that species (olive-green-colored points) removed, the PGLS p-values became 0.009 for esters sum and 0.002 for esters percent; both had lambda = 0 and 28 degrees of freedom.
Figure 2.6 Heatmap of the 15 compounds we quantified that are on the list of the 21 most important to cultivated tomato flavor according to Martina et al. (2021). Category is based on precursor molecules. Flavor effect comes from Klee and Tieman (2018), NSS = not statistically significantly correlated with consumer liking. See Appendix B, Table B5 for further details on the compounds. Data values (in ng/gfw/hr) were log (base 10) transformed but not scaled. Columns show accession averages, with color-coding along the top denoting species, arranged phylogenetically based on a tree from Barnett et al. (2023a) with colored-fruited species (red box) on the left and green-fruited species (blue box) on the right. Three-letter abbreviations denote species as follows: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum NEO = Solanum neorickii, PEN = Solanum pennellii, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium. Numbers after 3-letter species abbreviations are Tomato Genetics Resource Center (TGRC) 'LA' accession numbers, except for CER_B8189 (full accession ID BGV008189) which does not have an 'LA' number.
Figure 2.7 Association between total sugar concentration (glucose + fructose + sucrose, in mg/mL) and ester levels, displayed as A) total sum (in ng/gfw/hr) of the 17 ester compounds quantified, and B) percentage of the total sum of all 66 compounds, based on raw/untransformed accession mean values. Colored points represent accessions and are color-coded by species, abbreviated as follows: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum NEO = Solanum neorickii, PEN = Solanum pennelli, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium; see Appendix S1, Fig. B5 for a version of this plot with accession numbers labeled. 31 accessions are included; the cultivated variety Ailsa Craig was removed for this analysis because it was influenced by human selection. Pearson's r and phylogenetic generalized least squares (PGLS) models were produced from raw/untransformed accession mean values, with 29 degrees of freedom and lambda = "ML" for the PGLS models.
3.1 Abstract

Ripe fruit sugar traits help make fleshy fruits attractive to animals and can be part of syndromes of correlated traits that may have evolved together due to selective pressure from animal disperser preferences. The genes underlying fruit sugar traits have been studied extensively in cultivated tomato (Solanum lycopersicum) due to their relevance in breeding more flavorful crop varieties. However, few studies have assessed the evolutionary dynamics of these fruit sugar genes, which are also ecologically relevant, across all 13 wild species in the tomato clade (Solanum sect. Lycopersicon). This group shows substantial interspecies variation in ripe fruit sugar content and composition and therefore provides a unique opportunity to explore fleshy fruit trait evolution at the molecular level. In this study we present a cladewide evolutionary analysis of a set of sugar-related genes likely to affect fruit phenotypes in wild tomato species. We tested for molecular signs of selection on four genes and found evidence of purifying selection on their protein-coding sequences, although the vacuolar acid invertase TIV1 and the glucose transporter SWEET1a appear to be under less stringent purifying selection than the sucrose synthase SUSI and the ADP-glucose pyrophosphorylase subunit AGPL1. In contrast, there were signs of positive selection on the 5' promoters of SWEET1a, AGPL1, and especially TIV1. For the invertase TIV1, selection appears to have acted on a retrotransposon insertion in the 5' promoter that is present in the four hexose-
accumulating colored-fruiting species but absent from the nine sucrose-accumulating
green-fruiting species—suggesting that the derived trait of ripe fruit hexose-accumulation
conferred an evolutionary advantage. Taken together, our analyses indicate that selection
has acted in differing ways on the protein-coding and regulatory regions of fruit sugar-
related genes over the course of wild tomato evolution.

3.2 Introduction

The cultivated tomato (Solanum lycopersicum var. lycopersicum) is one of the
world's most valuable horticultural crops (Bergougnoux 2014). It is also a workhorse of
scientific research—the cultivated tomato and its 13 wild relative species that together
comprise the tomato clade (Solanum sect. Lycopersicon, family Solanaceae; Peralta et al.
2008; Fig. 3.1) are an important model system for fleshy fruit development and genetics
(Carrari and Fernie 2006; Kimura and Sinha 2008; Quinet et al. 2019; Zhu et al. 2022) as
well as ecological and evolutionary genomics (Moyle 2008; Pease et al. 2016). Of
particular interest are ripe fruit sugar traits, which are key components of cultivated
tomato flavor and consumer preference (Klee and Tieman 2018). These fruit sugar traits
are also relevant to evolutionary biologists because they can be part of syndromes of
correlated traits that may have evolved together due to selective pressure from animal
disperser preferences (Valido et al. 2011; Cazetta et al. 2012; Valenta and Nevo 2020;
Barnett et al. 2023a). Because sugar is a caloric reward that can increase a fruit's chances
of being chosen by animal dispersers (Schaefer et al. 2003) and therefore influence the
plant's seed dispersal and reproductive success (Valenta and Nevo 2020), the genes
underlying ripe fruit sugar content and composition may bear signatures of natural selection.

Variation in ripe fruit sugar traits upon which selection can act is known to exist across wild species in the tomato clade as well as among cultivated tomato varieties (Schauer et al. 2005; Beckles et al. 2012; hereafter we use the term ‘tomato’ to refer to both cultivated and wild species as a group). The three main sugar compounds found in tomato fruits are sucrose and its two components, the hexoses glucose and fructose. Differences in fruit sugar composition are a major component of putative dispersal syndromes in the tomato clade (Barnett et al. 2023a), most notably the contrast between the four brightly colored-fruited species (Solanum cheesmaniae, Solanum galapagense, Solanum lycopersicum, and Solanum pimpinellifolium) that accumulate primarily hexoses and the nine green-fruited species that accumulate primarily sucrose (Fig. 1; Appendix C, Fig. C1a, Fig. C2a). Fructose-to-glucose ratio also varies across the clade, with S. cheesmaniae notably higher than the other colored-fruited species and several green-fruited species (Solanum arcanum, Solanum habrochaites, Solanum huaylasense, and Solanum peruvianum) higher than the rest of the clade (Fig. 1; Appendix C, Fig. C1b, Fig. C2b). Ratios among the three sugar compounds are important tomato fruit phenotypes relevant to flavor and animal preferences because 1) each sugar can have a different perceived sweetness (Biester et al. 1925), and 2) some animals have difficulty digesting the more complex disaccharide sucrose because of either rapid gut passage (Afik and Karasov 1995) or lack of a sucrose-digesting enzyme (Martinez del Rio and Stevens 1989).
Fruit sugar content, or the combined amount of all sugar compounds, is another important trait because it affects the overall caloric content of a fruit and thus the nutritional reward an animal disperser obtains by eating a fruit. Sugar content is commonly measured as total soluble solids (TSS), or degrees Brix. While TSS also includes some non-sugar compounds such as organic acids, it is a convenient proxy for sugar content because sugars make up 65% of TSS (Beckles 2012). Tomatoes show substantial differences in TSS at both the inter- and intra-species level; the trait varies independently of phylogeny and does not show any clear cladewide patterns (Fig. 3.1; Appendix C, Fig. C1c, Fig. C2c; Barnett et al. 2023a). For example, the lack of phylogenetic inertia in TSS is clear in the two Galápagos species, with *S. cheesmaniae* showing some of the lowest TSS values in the clade (~4-5° Brix) while its sister species *S. galapagense* has some of the highest values (~11-12° Brix).

Tomato fruit sugar traits are complex phenotypes influenced by many genes, but decades of work have identified a number of key genes likely to underlie the phenotypic variation seen across the clade. While the focus of much of this research has been on cultivated tomato due to its importance for breeding more flavorful crop varieties (reviewed in Carrari and Fernie 2006; Kanayama 2017; Zhu et al. 2022), a number of studies leveraged natural interspecies variation to identify genes through crosses between a cultivated tomato and a wild species with a different fruit sugar phenotype. The genes found through these interspecies crosses are particularly relevant to the question of how fruit sugar traits evolved because variation was present in the wild rather than induced through artificial mutagenesis. Three well-studied crosses (and their associated introgression lines) include those between the cultivated tomato and 1) *Solanum*
chmielewskii, a green-fruited species notable for high sucrose accumulation (Yelle et al. 1991; Chetelat et al. 1993; Chetelat et al. 1995), 2) Solanum habrochaites, a green-fruited species notable for high fructose-to-glucose ratio (Hadas et al. 1995, Levin et al. 2000), and 3) Solanum pennellii, a green-fruited species notable for high TSS (Eshed and Zamir 1995, Fridman et al. 2004). These studies used map-based cloning to pinpoint several genes that can affect ripe fruit sugar traits and performed transgenic experiments to confirm the genes' effect on phenotype in cultivated tomato.

While identifying these sugar-related genes and their variation in wild species has proved valuable to modern cultivated tomato breeding efforts, studies have yet to synthesize this knowledge from an evolutionary perspective. Given that ripe fruit sugar traits are key components of putative dispersal syndromes in wild tomatoes (Barnett et al. 2023a), such a synthesis would provide an exciting opportunity to test for molecular signs of selection on genes known to affect ecologically relevant fruit sugar phenotypes. Although a few studies have sequenced a sugar-related gene across most of the 13 wild tomato species (e.g. TIV1 in Slugina et al. 2017; LIN7 in Slugina et al. 2018; SUS1 in Slugina et al. 2019), a cladewide synthesis of inter- and intra-species variation in multiple fruit sugar genes is lacking.

Tests of selection on fruit sugar-related genes have yet to be applied in the context of speciation across the entire tomato clade. Some studies have found signs of selection on a sugar-related gene (e.g. the invertase LIN5 which is associated with ripe fruit TSS levels) during the domestication of cultivated tomatoes from the red-fruited wild S. pimpinellifolium and semi-wild S. lycopersicum var. cerasiforme (Tieman et al. 2017; Razifard et al. 2020; Zhao et al. 2022), but these tests did not include any green-fruited
wild species so could not provide insights into evolutionary dynamics prior to domestication. In the case of LIN5, the allele found mostly in cultivated tomato (which is associated with lower fruit TSS) likely increased in frequency as an indirect effect of selection for larger fruit size during domestication and improvement (Tieman et al. 2017; Zhu et al. 2018).

In this study, we synthesize knowledge of the genetic basis of ripe fruit sugar traits in wild tomatoes and present a cladewide analysis of evolutionary dynamics among a set of fruit sugar-related genes. Because ripe fruit sugar traits are key parts of syndromes of correlated traits that may have evolved due to selective pressure from animal disperser preferences (Barnett et al. 2023a), we sought to test for molecular signs of selection on genes likely to influence these ecologically relevant fruit phenotypes. First, we reviewed the literature to determine which tomato fruit sugar-related genes have been demonstrated to affect variation in fruit sugar traits. We chose five of the most well-studied of these genes, including four that had been identified through interspecies crosses and thus are known to possess phenotypically relevant variation in at least one wild species. We then filled gaps in our knowledge of cladewide variation in these five target genes and their promoters through next generation amplicon sequencing of several accessions of each of the 13 species of wild tomatoes, which enables detection of structural variations. With these data for the target genes, we assess: (1) patterns of phylogenetic variation, (2) nucleotide variants that may affect protein structure, (3) whether any of these protein-coding sites show evidence of selection, and (4) whether there is evidence of selection on the 5' promoter regions of genes hypothesized to be associated with fruit sugar trait differences among species. Given that sugar-related genes
are important to primary metabolism, we predicted that all five genes would show
evidence of purifying selection and that few variants would affect protein structure.
However, because not all genes are expressed in all tissues and some may have paralogs,
we expected that some genes would be under less stringent purifying selection than
others. We also predicted that 5' regulatory regions would show evidence of selection in
comparisons of species with known differences in fruit sugar traits because interspecies
gene expression differences have been associated with variations in those phenotypes.

3.3 Results

3.3.1 Literature survey of genes known to affect tomato fruit sugar traits
and the choice of five target genes

Plants contain a number of genes that could potentially affect sugar traits if
expressed in fruits. These sugar-related genes encode proteins that fall within two broad
categories: (1) metabolic enzymes, which chemically modify sugar molecules by
breaking them apart, linking them together, or attaching other molecules such as
phosphate groups; and (2) transporters, which facilitate movement of sugar molecules
across membranes. Decades of work has examined a number of these genes and their
proteins throughout the tomato fruit development process, mostly in cultivated varieties
(reviewed in Carrari and Fernie 2006; Kanayama 2017; Zhu et al. 2022). To our
knowledge, ten sugar-related genes in tomato have been functionally validated (through
transgenics in a cultivated variety and/or protein assays in interspecies introgression
lines) to affect ripe fruit sugar content and/or composition (Zhu et al. 2022). A number of
other sugar-related genes show evidence that they may affect fruit sugar traits, either
because of significant GWAS associations with a fruit sugar trait (such as phosphoenolpyruvate carboxylase (PEP); Zhao et al. 2022) or validated effects on other fruit traits (such as SlSWEET15 knockouts with decreased fruit size; Ko et al. 2021); however, those genes have not yet been directly demonstrated to affect a ripe fruit sugar trait. Ripe fruit is particularly important in the context of fruit trait evolution because it is the stage at which animal dispersers are most likely to consume a fruit and potentially spread seeds to new areas.

We were particularly interested in genes identified through interspecies crosses, because this indicates the gene shows phenotypically relevant variation in at least one wild species. Four of the ten genes validated to control tomato fruit phenotypes (TIV1, LIN5, AGPL1, SWEET1a; reviewed in Zhu et al. 2022) were previously identified using map-based cloning approaches based on crosses between a cultivated tomato variety and one of its wild relatives. Two of the genes encode different invertases—metabolic enzymes which irreversibly cleave sucrose into its component hexoses (glucose and fructose)—the vacuolar acid invertase TIV1 and the cell wall acid invertase LIN5. The gene AGPL1 encodes the large subunit of the ADP-glucose pyrophosphorylase (AGPase) metabolic enzyme, which plays an important role in converting glucose into starch. Finally, SWEET1a encodes a glucose transporter from the SWEET protein family that facilitates movement of glucose across membranes.

The earliest of these genes to be identified was the vacuolar acid invertase gene TIV1 (Solyc03g083910), which affects ripe fruit sucrose-to-hexose ratio. Crosses between the hexose-accumulating S. lycopersicum and the sucrose-accumulating Solanum chmielewskii accession LA1028 (Yelle et al. 1991; Chetelat et al. 1993; Chetelat
et al. 1995) were used to pinpoint genetic control of the ripe fruit trait of sucrose-to-hexose ratio to a single major recessive gene. Follow-up studies showed that in sucrose-accumulating fruits with the *S. chmielewskii* allele, the vacuolar acid invertase enzyme was inactive and the *TIV1* gene was transcriptionally silent (Klann et al. 1993); this same lack of *TIV1* gene expression and enzyme activity was also found in ripe fruits of the sucrose-accumulating *S. habrochaites* (Miron et al. 2002). The difference in gene expression between the cultivated and green-fruited species alleles appears to be due to a retrotransposon insertion in the *TIV1* 5' promoter in cultivated tomato, which was shown through a PCR test to also be present in the colored-fruited wild species *S. pimpinellifolium* and *S. cheesmaniae*, but absent from the green-fruited species *S. chmielewskii, S. habrochaites, S. pennellii*, and *S. peruvianum* (Moy et al. 2007). Recent studies of *TIV1* gene expression in ripe fruits of wild species found high expression in the colored-fruited hexose-accumulating *S. lycopersicum, S. pimpinellifolium*, and *S. cheesmaniae* but low expression in the green-fruited sucrose-accumulating *S. arcanum, S. chmielewskii, S. habrochaites, S. pennellii*, and *S. peruvianum* (Appendix C, Fig. C3; Slugina et al. 2017; Doron-Faigenboim et al. 2023). These results suggest that the presence of the retrotransposon in colored-fruited species is correlated with increased *TIV1* gene expression and hexose accumulation in ripe colored fruits. The gene's effect on the sucrose-to-hexose ratio trait was validated in transgenic cultivated tomato lines with an antisense *TIV1* gene that produced fruits with increased sucrose and decreased hexoses (Klann et al. 1996). The functional mechanism driving this phenotypic effect is *TIV1* invertase enzyme activity in the vacuole, which breaks down sucrose into its component hexoses.
Another well-studied invertase gene, the cell wall acid invertase *LIN5* (Solyc09g010080), can affect ripe fruit TSS. The gene was identified through introgression lines from a cross between *S. lycopersicum* cultivar M82 and the high-TSS *Solanum pennellii* accession LA0716 (Eshed and Zamir 1995). The *LIN5* allele in *S. pennellii* has an amino acid substitution from Glu to Asp at position 348 that resulted in enhanced ability to degrade sucrose (thus increasing fruit import capacity from phloem) when the altered invertase enzyme was tested in genetically engineered yeast (Fridman et al. 2004). *LIN5* gene expression levels in fruits were statistically similar between introgression lines with the two different alleles, suggesting that the difference in ripe fruit TSS was not due to regulatory changes (Fridman et al. 2004). A role for *LIN5* in TSS has also been inferred from a GWAS including *S. lycopersicum*, the semi-wild *S. lycopersicum* var. *cerasiforme*, and wild *S. pimpinellifolium*, where a missense SNP (resulting in an amino acid change from Asn to Asp) at position 366 has different variants among cultivated and wild accessions (Tieman et al. 2017; Razifard et al. 2020; Pereira et al. 2021). Overexpression of the *LIN5* variant less common in cultivated plants results in higher fruit sugar content (Tieman et al. 2017), and silencing of *LIN5* in cultivated tomatoes lowers TSS, glucose, and fructose levels (Zanor et al. 2009). Interestingly, at both position 348 and 366 most cultivated tomatoes have the allele with less active invertase and correspondingly lower ripe fruit TSS. Biochemical analysis of fruits from the introgression line with the *S. pennellii* allele of *LIN5* concluded that the more active invertase enzyme cleaved more sucrose and thus increased the fruit's capacity to import additional sucrose from the phloem (Baxter et al. 2005).
Variation in the ADP-glucose pyrophosphorylase large subunit gene AGPL1 (Solyc01g109790) has also been shown to affect ripe fruit TSS, although through a different mechanism than the LIN5 invertase—increased starch accumulation rather than increased sucrose cleavage. The AGPL1 gene was identified via introgression lines derived from a cross between S. lycopersicum and the high-TSS Solanum habrochaites accession LA1777 (Hadas et al. 1995), in conjunction with the aforementioned S. lycopersicum x S. pennellii lines (Schaffer et al. 2000). Subsequent functional work revealed that lines with the S. habrochaites allele of AGPL1 showed higher gene expression and higher AGPase enzyme activity in developing fruit relative to the cultivated tomato allele, which correlated with increased starch accumulation in immature fruit and higher TSS in ripe fruit (Petreikov et al. 2006). A later study of gene expression in wild tomato fruits showed that AGPL1 is mostly expressed at the immature green stage, confirming that S. habrochaites has higher expression levels than cultivated tomato—and also revealing interspecies expression differences among S. pimpinellifolium, S. cheesmaniae, S. chmielewskii, S. peruvianum, and S. pennellii (Appendix C, Fig. C3; Doron-Faigenboim et al. 2023). While protein structure differed by four amino acid changes between the S. habrochaites and cultivated alleles in the introgression lines mentioned above, these differences in amino acid sequence did not change the kinetic properties of the purified enzymes (Petreikov et al. 2006), although the changes may serve a modulatory or stabilizing function. Regardless of the exact underlying mechanism, more AGPase enzyme activity in developing tomato fruits increases starch accumulation which appears to enable higher TSS levels in ripe fruits.
Finally, the SWEET glucose transporter gene *SWEET1a* (Solyc04g064610) is known to affect another fruit trait, fructose-to-glucose ratio. That trait was mapped to a major locus on chromosome four with crosses between *S. lycopersicum* and the high fructose-to-glucose *S. habrochaites* accession LA1777 (Levin et al. 2000), followed by subsequent fine mapping that identified *SWEET1a* and verified its effect on fructose-to-glucose ratio via transgenic overexpression (Shammai et al. 2018). Introgression lines with the *S. habrochaites* allele of *SWEET1a* had higher gene expression and a higher fruit fructose-to-glucose ratio than lines with the cultivated tomato allele (Shammai et al. 2018). *SWEET1a* gene expression also varied considerably among five other wild species, with expression levels in ripe fruit positively correlated with ripe fruit fructose-to-glucose ratio (Shammai et al. 2018). As for differences in protein structure, the same study found that six amino acid sites and two indels varied in at least one wild species, but did not discuss whether these changes might affect protein function. Shammai et al. (2018) also localized the transporter protein encoded by the gene to the plasma membrane, concluding that it functions to move glucose out of cells while keeping fructose in, resulting in a high ratio of fructose relative to glucose.

Other sugar-related genes that have not been mapped to a fruit trait-related QTL through interspecies crosses, but were nevertheless hypothesized to affect tomato fruits because of orthology with genes in other plant species, have also been functionally validated in cultivated tomato. A well-known example is the sucrose synthase gene *SUS1* (Solyc12g009300); transgenics with an antisense *SUS1* gene show decreased sucrose import into young fruit (D'Aoust et al. 1999) but intriguingly do not exhibit changes in ripe fruit sugar levels (Chengappa et al. 1999). However, another study correlated
increased SUS1 expression with higher ripe fruit TSS and hexose levels in a near-isogenic line from a cross between cultivated tomato and S. pennellii, suggesting that the region introgressed from the wild species played a key role in altering SUS1 gene expression and ripe fruit TSS (Ikeda et al. 2016). The other five genes with a demonstrated effect on a fruit sugar trait include a vacuolar invertase inhibitor (VIF; Qin et al. 2016), a cell wall invertase inhibitor (INVINHI, Jin et al. 2009), tomato ripening-associated membrane protein (TRAMP, Chen et al. 2001), a sucrose transporter (SUT2; Hackel et al. 2006), and malate dehydrogenase (MDH; Centeno et al. 2011).

From the ten functionally validated fruit sugar-related genes described above, we chose five genes to explore more deeply. We included the four genes identified from wild species introgression lines (AGPL1, LIN5, SWEET1a, TIV1) because they are well-studied and are known to possess meaningful variation in at least one wild species. We also included SUS1 because 1) it is another one of the most extensively studied sugar-related genes in tomato, 2) a wild species introgression line showed differences in SUS1 expression and ripe fruit TSS relative to cultivated tomato, and 3) it encodes a sucrose synthase, a type of sugar-related enzyme not represented among the other four selected genes. All five of the genes are known to be expressed in the fruits of at least some tomato species, with considerable variation in gene expression among species and across fruit developmental stages (Appendix C, Fig. C3; Doron-Faigenboim et al. 2023). Details on the five target genes, each of which encodes a different type of enzyme or transporter, are summarized in Table 3.1 and Appendix C, Tables C1 and C2.
3.3.2 Cladewide DNA sequence variation in the target genes: phylogenetic patterns and signs of selection

To gain a more complete cladewide picture of variation in the five target genes and enable molecular evolution analyses, we obtained new sequences for each gene and ~2 kb of its 5' promoter from 40 accessions—three of each of the 13 wild tomato species plus one cultivated tomato variety, Ailsa Craig (Appendix C, Table C3). We used amplicon sequencing with custom-designed primers (Appendix C, Tables C4, C5) and de novo assembled the NGS reads from pooled amplicons. This cost-efficient approach enabled us to pick up potential structural variants that would not be detected by aligning reads to a single reference genome, improving upon the approach of others (e.g., Dinh et al. 2018) that used whole genome resequencing to obtain gene sequences.

We were not able to amplify and assemble sequences for all 40 accessions for all five target genes, but for four of the genes we did get the full coding region for at least 32 accessions and the 5' promoter region for at least 27 accessions (Table 3.2). We were unable to obtain sequences for the full coding region or the 5' promoter of the LIN5 gene, so did not include it in our molecular evolution analyses because these require a complete protein-coding sequence. The difficulty in amplifying and assembling sequences of LIN5 may be related to the similarity with other LIN paralogs in tomato and the presence of a CACTA transposon-like insertion in the first intron. There are at least nine cell wall acid invertase genes in tomato (Wan et al. 2018), with one family consisting of two tandem pairs of segmentally duplicated CWIN genes on chromosomes 9 (LIN5 and LIN7) and 10 (LIN6 and LIN8) that all appear to be involved in sucrose sink metabolism (Godt and Roitsch 1997; Fridman and Zamir 2003). In addition, a CACTA-transposon-like insertion
is present in the first intron of LIN5 in cultivated tomato (Proels and Roitsch 2006), although whether this insertion is also present in any wild tomato species was not tested in that study. Given that tomato cell wall invertase genes are likely targets for transposon insertions due to their high AT nucleotide content, as well as the fact that CACTA transposon-like insertions appear to be ubiquitous in Solanaceae genomes (Proels and Roitsch 2006), the presence of the transposon in LIN5 in cultivated tomato could explain some of the difficulty in sequencing and assembling amplicons of this gene across wild tomato species.

3.3.2.1 Overall phylogenetic patterns in the target gene sequences

To assess the phylogenetic structure of variation in the target genes and determine how well this structure aligns with species distinctions, we produced maximum likelihood gene trees for exons (Fig. 3.2a), exons and introns (Fig. 3.2b), and 5' promoters (Fig. 3.2c); LIN5 is shown separately in Fig. 3.2d because only partial CDS sequences and no 5' promoter sequences were obtained. For most genes the trees largely followed species phylogeny, and showed the brightly colored-fruitied accessions as a monophyletic group separate from the paraphyletic green-fruitied accessions. The exception was the sucrose synthase gene SUS1, for which the exon tree and exon plus intron tree showed the Galápagos species surprisingly closer to the more basal green-fruitied species than the other colored-fruitied species (Figs. 3.2a, 3.2b). This unusual pattern was due to a number of synonymous substitutions in SUS1 that were shared among those species but not present in the red-fruitied species. Interestingly, SUS1 appears to be evolving differently than the other three genes analyzed and accumulating
more synonymous substitutions, with its exons showing the highest nucleotide diversity (Table 3.2) but lowest percentage of amino acid substitutions relative to protein length (Table 3.3). The unusual phylogenetic pattern does not extend to the 5′ promoter sequences of *SUS1*, which closely follow species expectations (Fig. 3.2c).

Not surprisingly, for all five genes the nine green-fruited species showed more sequence variation than the five colored-fruited species. Nucleotide diversity among the green-fruited species ranged from a minimum of 0.00799 (*AGPL1* exons) to a maximum of 0.05939 (*AGPL1* 5′ promoter), while among the colored-fruited species the minimum was 0.00069 (*AGPL1* exons) and the maximum was 0.01418 (*SUS1* 5′ promoters; Table 3.2).

When amino acid substitutions were detected across the clade these tended to have low frequency alleles, with about half of the substitutions present in only one accession. The four genes for which we obtained full coding sequences displayed between 17 and 63 amino acid substitutions (summary in Table 3.3; alignments in Appendix C, Tables C6-C9). To identify substitutions that could have a radical effect on protein structure, we carried out a SIFT analysis (Ng and Henikoff 2001) with cultivated tomato as the query and predicted the following number of intolerable/deleterious substitutions: four in *AGPL1*, four in *SUS1*, six in *SWEET1a*, and 14 in *TIV1* (Appendix C, Table C10). Nearly all those 28 deleterious substitutions were present in only 1–3 accessions. These results suggest that across the clade there is selection to preserve protein function in these four genes, and potentially deleterious amino acid variants are kept at low frequencies.
We next tested for evidence of natural selection in the sequences of each of the target genes, reasoning that variations in these genes could alter fruit sugar phenotypes likely to affect the fruit's chances of being eaten and dispersed by animals (and thus ultimately alter a plant's reproductive success). In the coding regions we expected to see mostly evidence of purifying selection because properly functioning sugar-related proteins are important to primary metabolism in the plant, and thus variants that affected the protein's functionality would likely be quickly selected against. Positive selection could still occur if an amino acid change resulted in enhanced protein function, but this seemed less likely than a deleterious amino acid change. For the four target genes for which we obtained full coding sequences, we tested for signs of selection at individual sites in the protein-coding region using Selecton (Doron-Faigenboim et al. 2005) and HyPhy (Kosakovsky Pond et al. 2020), both of which are based on Ka/Ks (the ratio of nonsynonymous to synonymous substitutions). We also tested for signs of selection in the 5' promoter region of genes where the associated fruit sugar phenotype showed clear interspecies differences, reasoning that nucleotide substitutions in this regulatory region could cause heritable changes in gene expression that alter the fruit sugar trait. These included TIVI and the higher sucrose-to-hexose ratio in green-fruited vs. colored-fruited species; AGPLI, SUS1, and the high TSS in S. galapagense relative to its sister species S. cheesmaniae; and SWEET1a and the high fructose-to-glucose ratio in S. cheesmaniae, S. arcanum, S. habrochaites, S. huaylasense, and S. peruvianum.
3.3.2.2 *AGPL1*: amino acid substitutions, signs of selection, and potential associations with the TSS fruit trait

For the coding sequence of the ADP-glucose pyrophosphorylase large subunit gene *AGPL1*, which has been shown to affect ripe fruit TSS (Petreikov et al. 2006), there were a total of 14 amino acid sites with substitutions in at least two accessions in the clade according to our new sequence data, two of which were predicted to be intolerable by SIFT (Appendix C, Table C6; note that there were 16 amino acid sites with substitutions in only one accession). One of the substitutions that was present in multiple species was a three amino acid insertion (Glu-Lys-Lys at positions 54-56) that was absent from the four colored-fruited species, *S. chmielewskii*, *S. neorickii*, and all but one of the *S. arcanum* accessions, but was present in the remaining six green-fruited species. Earlier work has shown that this Glu-Lys-Lys insertion in an allele of *AGPL1* in *S. habrochaites* (along with three single amino acid changes, Ile27Met, Thr412Ser, and Ile506Thr) did not change the enzyme's kinetics relative to the cultivated tomato version of the protein, at least according to *in vitro* chemical assays of purified protein (Petreikov et al. 2006). However, the change in protein structure may serve a modulatory or stabilizing function that results in increased starch accumulation in developing tomato fruit.

Tests of selection on the coding region of *AGPL1* suggested fairly strong purifying selection, with a cladewise Ka/Ks ratio of 0.1399 (Table 3.4) that was well below the 1.0 that would be expected in the absence of selection. There was almost no evidence of positive selection—only one of the sites with an amino acid substitution (codon 513) showed evidence of pervasive (i.e. assuming selection pressure is constant across phylogeny) positive selection according to the FUBAR method in HyPhy, zero
sites showed evidence of episodic (i.e. evolving under selection in a proportion of branches) positive selection according to the MEME method in HyPhy, and a test of positive selection on the protein in Selecton was non-significant (Table 3.4). The codon 513 site had Ile in cultivated tomato and most of the other species in the clade, with the exception of *S. chilense* and one *S. huaylasense* accession (change to Val), and *S. habrochaites* and one *S. pennellii* accession (change to Thr; Appendix C, Table C6). The accessions with the change at codon 513, which was predicted to be tolerable according to SIFT (Appendix C, Table C10), do not appear to have any obvious differences in fruit sugar traits relative to the rest of the clade (Fig. 3.1).

Because *AGPL1* has been associated with the TSS trait and *S. galapagense* displays high fruit TSS relative to its sister species *S. cheesmaniae*, we were particularly interested in whether any amino acid changes distinguished those two species. The three *S. cheesmaniae* accessions had identical *AGPL1* protein sequences, while two of the three *S. galapagense* accessions had a two amino acid deletion at position 21-22 and one of the *S. galapagense* accessions had a Ser67Thr substitution not found in any other accessions in the clade. These protein changes unique to *S. galapagense* may be worth further investigation into whether they could be related to the high ripe fruit TSS values observed in the species.

For the 5' promoter sequence of *AGPL1*, a comparison between the red-fruiting *S. pimpinellifolium* and green-fruiting *S. pennellii* showed that substitution rates were much higher in the 5' promoter relative to the coding region, with a Kn/Ks ratio of 6.960 (Table 3.5). The Kn/Ks ratio was also fairly high (3.862) for a comparison between *S. pennellii* and another green-fruiting species, *S. corneliomulleri*, and slightly lower (1.966) but still
above the neutral expectation of 1.0 for a comparison between red-fruited *S. pimpinellifolium* and green-fruited *S. arcanum*. These results of a consistently elevated substitution rate suggest that selection may be acting on the 5' promoter of *AGPL1*, especially in the case of *S. pennellii*.

3.3.2.3 *SUS1*: amino acid substitutions, signs of selection, and potential associations with the TSS fruit trait

For the coding sequence of the sucrose synthase gene *SUS1*, which may affect ripe fruit TSS (D'Aoust et al. 1999; Ikeda et al. 2016), there were a total of 8 amino acid sites with substitutions in at least two accessions in the clade according to our sequence data, one of which was predicted to be intolerable by SIFT (Appendix C, Table C7; note that there were 17 amino acid sites with substitutions in only one accession, three of which were predicted intolerable by SIFT). One of the substitutions present in multiple species was Arg11Ser; at that site Ser appears to be the ancestral amino acid because only cultivated tomato and one of the *S. l. v. cerasiforme* accessions had the Arg. A previous study measured the enzyme kinetics of the Arg11Ser amino acid change and found no difference between the two versions of the protein (Dinh et al. 2018).

Similar to *AGPL1*, tests of selection on the coding region of *SUS1* suggested fairly strong purifying selection, with a cladewide Ka/Ks ratio of 0.1447 (Table 3.4). There was almost no evidence of positive selection—only one of the sites with an amino acid substitution (codon 126) showed evidence of pervasive positive selection according to FUBAR in HyPhy, zero sites showed evidence of episodic positive selection according to MEME in HyPhy, and Selecton found no positively selected sites in the protein (Table
3.4. The codon 126 site had Ala in cultivated tomato and most of the other species in the clade, with the exception of S. neorickii LA1626, S. huaylasense LA1365, and S. corneliomulleri LA1305, which all had a change to Val. The accessions with the change to Val at codon 126, which SIFT predicted to be tolerable (Appendix C, Table C10), all have fairly high TSS but do not stand out relative to the rest of the clade (Appendix C, Fig. C2).

Because SUS1 has been associated with the TSS trait and S. galapagense displays high fruit TSS relative to its sister species S. cheesmaniae, we were particularly interested in whether any amino acid changes distinguished those two species, but found that their protein sequences were identical.

For the 5' promoter sequence of SUS1, a comparison between the red-fruited S. pimpinellifolium and green-fruited S. pennellii showed that substitution rates were lower in the 5' promoter relative to the coding region, with a Kn/Ks ratio of 0.517 (Table 3.5). The Kn/Ks ratio was also low (0.551) for a comparison between S. pennellii and another green-fruited species, S. corneliomulleri. However, Kn/Ks was 1.957 for a comparison between red-fruited S. pimpinellifolium and green-fruited S. arcanum, above the neutral expectation of 1.0.

3.3.2.4 SWEET1a: amino acid substitutions, signs of selection, and potential associations with the fructose-to-glucose ratio fruit trait

For the coding sequence of the glucose transporter gene SWEET1a, which is associated with fructose-to-glucose ratio (Shammai et al. 2018), there were a total of 11 amino acid sites with substitutions in at least two accessions in the clade according to our
sequence data, four of which were predicted to be intolerable by SIFT (Appendix C, Table C8; note that there were 8 amino acid sites with substitutions in only one accession, two of which were predicted intolerable by SIFT). While nearly all of the amino acid substitutions predicted deleterious by SIFT for the four genes analyzed were present in only 1–3 accessions (Appendix C, Table C10), the two exceptions were in SWEET1a, with Thr219Ala in five accessions (only the Galápagos species had Ala) and Arg246Lys in 31 accessions (only five red-fruited accessions had the Arg). Neither of those two substitutions correlate with differences in the fructose-to-glucose ratio phenotype that is associated with SWEET1a.

Tests of selection on the coding region of SWEET1a suggested less stringent purifying selection than the other genes analyzed, with a cladewide Ka/Ks ratio of 0.4357 (Table 3.4). There was limited evidence of positive selection—none of the sites with an amino acid substitution showed evidence of pervasive positive selection according to FUBAR in HyPhy, and one site (codon 18) showed evidence of episodic positive selection according to MEME in HyPhy. Selecton found 17 sites with a Ka/Ks ratio > 1, computing a significance level of 0.05 for the likelihood ratio test between the model allowing positive selection and the null model (Table 3.4). The codon 18 site has Thr in cultivated tomato and all but one of the other species in the clade, S. pennellii; the change to Phe was predicted tolerable by SIFT (Appendix C, Table C10). The fructose-to-glucose ratio in S. pennellii is fairly low compared to its closest relative species in the green-fruited group, but is comparable to the colored-fruited species and the green-fruited species more closely related to the colored group (Appendix C, Fig. C2).
Our data did not show any clear correlations between the high fructose-to-glucose ratio in fruits of *S. cheesmaniae* and four of the green-fruited species and any protein sequence differences, but revealed a few notable substitutions. For the four green-fruited species with high fructose-to-glucose ratio, there were no sites with changes that clearly distinguished them from the rest of the clade, although amino acid substitutions at positions 226, 237, and 238 were present in several accessions and may be worth further investigation. For *S. cheesmaniae*, which is expected to have little genetic variation because it is a highly selfing species, two of the accessions we sequenced had identical *SWEET1a* proteins to the other colored-fruited species, but accession LA0428 had a premature stop codon at position 226 and a notably higher fructose-to-glucose ratio relative to the other *S. cheesmaniae* accessions. This result is unexpected because a higher fructose-to-glucose ratio has been shown to occur when the *SWEET1a* glucose transporter protein is more active (Shammai et al. 2018), and it seems unlikely that a truncated protein would have improved functionality. Also intriguing is that our phenotype data show *S. cheesmaniae* LA1412 had a fairly low fructose-to-glucose ratio (Appendix C, Fig. C2) while Shammai et al. (2018) reported that it had a high fructose-to-glucose ratio.

For the 5' promoter sequence of *SWEET1a*, substitution rates were consistently higher in the 5' promoter relative to the coding region, with Kn/Ks ratios of 9.345, 3.530, and 2.345 for the comparisons between *S. pimpinellifolium* vs. *S. arcanum*, *S. pimpinellifolium* vs. *S. pennellii*, and *S. pennellii* vs. *S. corneliomulleri*, respectively (Table 3.5). Because those ratios are all higher than the neutral expectation of 1.0, they
suggest that selection may be acting on the promoter region of this gene in several species.

3.3.2.5 TIV1: amino acid substitutions, signs of selection, and potential associations with the sucrose-to-hexose ratio fruit trait

For TIV1, the gene first associated with the high sucrose-to-hexose ratio in green-fruited species thanks to an interspecies cross between cultivated tomato and S. chmielewskii (Chetelat et al. 1995, Klann et al. 1996), it was notable that the colored-fruited species had nearly identical protein sequences (only four amino acid sites showed any variation with the colored-fruited species) while the green-fruited species were quite variable (~30 sites varied in at least two accessions within the green-fruited species). Cladewise, TIV1 had the most variable protein sequence of the four genes analyzed (Table 3.3)—there were a total of 31 amino acid sites with substitutions in at least two accessions in the clade according to our new sequence data, two of which were predicted to be intolerable by SIFT (Appendix C, Table C9; note that there were 32 amino acid sites with substitutions in only one accession, 12 of which were predicted intolerable by SIFT).

Selection tests on the coding region of TIV1 showed a cladewise Ka/Ks ratio of 0.2592, suggesting purifying selection that was less stringent than AGPL1 and SUS1 but more stringent than SWEET1a (Table 3.4). Of the four genes analyzed, TIV1 showed the most evidence of positive selection in its coding region. There were five amino acid sites in the TIV1 protein that had changes with evidence for pervasive positive selection according to FUBAR in HyPhy (Table 3.4); interestingly, all of the colored-fruited
accessions were identical at these five sites. Two of these sites also showed evidence of positive selection according to MEME in HyPhy (Table 3.4): one is codon 342 within the N-terminal domain of the protein, which is fixed as Gly in all colored-fruited species and is either Ala or a Thr in most (but not all) of the green-fruited species; the other is codon 433 within the C-terminal domain, which is fixed as Val in all colored-fruited species and is Ala in most (but not all) accessions with the peruvianum species complex. The changes at both amino acid sites are considered tolerable according to SIFT (Appendix C, Table C10). Both changes were also found by Slugina et al. (2017), and neither of the changes were predicted to be deleterious by PROVEAN according to those authors. Neither change alters a putative active site within the TIV1 invertase enzyme, but whether the changes affect enzymatic activity has not been tested (Slugina et al. 2017). Given that TIV1 is expressed at very low levels in the fruits of the five green-fruited species that have thus far been tested (Appendix C, Fig. C3; Slugina et al. 2017; Doron-Faigenboim et al. 2023), it may be that the gene is under less stringent purifying selection in the green-fruited species relative to the colored-fruited species.

Interestingly, there was an amino acid substitution in TIV1 unique to S. galapagense, a species notable for its high fruit TSS relative to its sister species S. cheesmaniae. At position 440, S. galapagense had Leu where S. cheesmaniae had Gln (like the rest of the colored-fruited species and three of the green-fruited species). This protein change may be worth further investigation into whether it could be related to the high ripe fruit TSS observed in S. galapagense.

We were particularly interested in testing for signs of selection on the TIV1 5' promoter because previous work had used PCR-based presence/absence tests to suggest
that a retrotransposon insertion associated with altered gene expression was present in the 5' promoter of colored-fruited species but not green-fruited species (Moy et al. 2007). Our sequence data confirmed the presence of this insertion in all of the colored-fruited species but none of the green-fruited species. To test whether the ~700 bp of 5' promoter sequence between the retrotransposon insertion and the TIV1 transcription start site showed any signs of selection, we compared nucleotide substitution rates between that promoter region and the coding region of TIV1. We found a Kn/Ks ratio of 4.442 in a comparison between the red-fruited, hexose-accumulating S. pimpinellifolium and the green-fruited, sucrose-accumulating S. arcanum (Table 3.6); this Kn/Ks ratio above 1.0 is suggestive of positive selection because the rate of nucleotide substitution in the promoter region is several times higher than the background rate of synonymous substitution in the adjacent coding region. In a comparison between S. pimpinellifolium and a more distantly related green-fruited species, S. pennellii, the Kn/Ks ratio for the TIV1 5' promoter is 5.019, again well above 1.0.

As another way to quantify whether selection may have acted on the TIV1 5' promoter, we compared polymorphism and divergence as calculated for the HKA test (Table 3.6). The colored-fruited accessions were considered one "species" and the green-fruited accessions were considered the second "species". To estimate a genome-wide background ratio of polymorphism to divergence, we used a set of 54900 SNPs from a GBS dataset (Barnett et al. 2023a). The polymorphism/divergence ratio for the TIV1 5' promoter was 0.207, significantly lower (chi-square test p-value = 0.0237) than the ratio of 0.896 for the genome-wide background ratio. This lower ratio in TIV1 indicates that the TIV1 5' promoter in the colored-fruited group has experienced a reduction in genetic
diversity, which is suggestive of selection because a smaller number of haplotypes have become fixed in the colored-fruited group.

3.4 Discussion

Our results present a cladewise synthesis of ecologically relevant fruit-sugar related genes and their evolutionary dynamics across 13 species of tomato wild relatives. We tested for molecular evidence of selection on a set of genes likely to influence ripe fruit sugar traits, reasoning that these genes may be under selection because they affect phenotypes important to syndromes of correlated fruit traits that could have arisen through evolutionary pressure from animal disperser preferences. As expected due to the genes' importance in primary metabolism, we found evidence of purifying selection on the protein-coding sequences of four sugar-related genes, although the vacuolar acid invertase TIV1 and the glucose transporter SWEET1a appear to be under less stringent purifying selection than the sucrose synthase SUS1 and the ADP-glucose pyrophosphorylase subunit AGPL1. In contrast, there were signs of positive selection on the 5' promoters of TIV1, SWEET1a, and AGPL1, indicating that selection has acted in differing ways on the regulatory and protein-coding regions of fruit sugar-related genes over the course of wild tomato evolution. In addition, we found a number of amino acid substitutions throughout the clade that did not show signs of selection, some of which may be correlated with phenotypic differences.

We were particularly interested in whether species with notable differences in fruit sugar traits shared any amino acid substitutions because such substitutions could represent ecologically meaningful genotype to phenotype associations. The most obvious
cladewide fruit trait difference is the low sucrose-to-hexose ratio in colored-fruited species relative to the higher sucrose-to-hexose ratio in green-fruited species. For AGPL1 and SUS1, there were no amino acid substitutions that distinguished the color groups. In SWEET1a all but one of the colored-fruited accessions had a 4-amino-acid deletion at position 228-231, and in TIV1 there were 4 separate amino acid substitutions that distinguished the color groups (R14H, F220Y, T437I, A463V). These sequence changes shared by the colored-fruited group may be worth further investigation into whether they affect protein function in the SWEET1a glucose transporter or the TIV1 invertase enzyme. Another notable phenotypic difference is the high TSS in S. galapagense relative to its sister species S. cheesmaniae. The two species had identical protein sequences for SUS1 and SWEET1a, while in TIV1 there was one change unique to S. galapagense (Q440L) and in AGPL1 there were two changes found only in S. galapagense; these protein differences may be worth further investigation into whether they could help explain the high ripe fruit TSS values in S. galapagense. The final fruit trait difference we focused on is the high fructose-to-glucose ratio in S. cheesmaniae and four of the green-fruited species, but we did not find any SWEET1a glucose transporter protein variants that clearly distinguished these species.

Tests of selection on the protein-coding regions of the four analyzed genes revealed very little evidence of positive selection overall, and varied slightly depending on the test (Table 3.4), with only TIV1 showing more than one codon site with evidence of positive selection. Interestingly, for all four genes only green-fruited species showed variation at the sites with evidence of positive selection, suggesting that protein sequences are strongly conserved among the colored-fruited species.
On a cladewide scale, coding regions appear to be evolving more quickly in
*SWEET1a* and *TIV1* relative to *AGPL1* and *SUS1*. There were relatively more cladewide
nonsynonymous substitutions in *SWEET1a* and *TIV1*; the mean Ka/Ks for all branches
was higher in *SWEET1a* (0.44) and *TIV1* (0.26) than *AGPL1* (0.14) and *SUS1* (0.14; Table
3.4). These Ka/Ks numbers are in line with the differences among genes in terms of
percentages of amino acid substitutions relative to protein length, with *TIV1* (9.9%) and
*SWEET1a* (6.8%) higher than *AGPL1* (5.5%) and *SUS1* (3.1%; Table 3.3). These results,
together with Ka/Ks values consistently lower in pairwise comparisons for *AGPL1* and
*SUS1* (0.02–0.15) relative to *SWEET1a* and *TIV1* (0.16–0.92; Table 3.5), suggest that
*AGPL1* and *SUS1* are evolving under more stringent purifying selection because
nonsynonymous substitutions are selected against more strongly.

Intriguingly, however, *SUS1* appears to be evolving differently by accumulating
more synonymous substitutions than the other three genes; its exons have the highest
nucleotide diversity (Table 3.2) but lowest percentage of amino acid substitutions relative
to protein length (Table 3.3). *SUS1* also has lower nucleotide diversity in the 5' promoters
(~0.03) than the other three genes (~0.05). One possible explanation for this unusual
pattern could be that *SUS1* is located in a region of the genome with a high mutation rate
but strong purifying selection on the promoter and coding regions.

Also notable is the observation that among the four genes analyzed, *AGPL1* has
the highest nucleotide diversity in the 5' promoters (0.056) but the lowest nucleotide
diversity in the exons (0.0065; Table 3.2). It appears that the *AGPL1* protein's function
has been strongly conserved through purifying selection while in contrast the regulatory
region may have experienced positive selection. This pattern makes sense given the
protein's importance in the starch accumulation pathway; effects on protein function could have negative repercussions throughout the plant.

For the cladewise patterns in the 5' promoters of the four genes analyzed, Kn/Ks ratios greater than 1.0 are suggestive of positive selection in the regulatory regions of TIV1, SWEET1a, and AGPL1 (Table 3.5). Interestingly, interspecies expression differences in these genes (but not protein structure changes) are correlated with fruit sugar trait differences between introgression line genotypes (Klann et al. 1993 for TIV1; Petreikov et al. 2006 for AGPL1; Shammai et al. 2018 for SWEET1a), suggesting that regulatory changes may be more likely targets for positive selection than coding sequence changes. More recent work on gene expression differences among wild species shows considerable interspecies variation for these three genes (Appendix C, Fig. C3; Doron-Faigenboim et al. 2023), consistent with the hypothesis that regulatory changes may have been under selection during wild tomato speciation.

Perhaps the most noteworthy result of our analyses is the evidence of selection on the 5' promoter region of the vacuolar acid invertase TIV1, which we confirmed to have a retrotransposon insertion in all of the hexose-accumulating colored-fruited species but none of the sucrose-accumulating green-fruited species. An HKA test comparing the TIV1 5' promoter region (before the start of the retrotransposon insertion) between colored-fruited and green-fruited groups (Table 3.6) showed a significant reduction in genetic diversity in the colored-fruited group that was suggestive of selection. Interestingly, for the other three genes the HKA test was not significant between the colored-fruited and green-fruited groups (Table 3.6), suggesting that selection has not acted on the 5' promoters of those genes in the colored-fruited group as a whole. The
Kn/Ks ratio for the TIV1 5' promoter region (Table 3.5) was greater than the neutral expectation of 1.0, which is also consistent with selection. Given that the difference in sucrose-to-hexose ratio between the two color groups is a key part of putative fruit dispersal syndromes in the tomato clade (Barnett et al. 2023a), the evidence of selection on a regulatory region of a gene known to be associated with the trait suggests that differences in TIV1 gene expression were an important evolutionary driver of fruit sugar trait divergence in the tomato clade.

Few fleshy fruit systems have been the subject of as much genetic research as the tomato clade, but the wild species in the group have been understudied from a molecular evolutionary perspective. Our analyses synthesize current knowledge on the genes underlying ripe fruit sugar traits in wild tomatoes, finding some signs of selection associated with fruit syndrome traits—particularly in regulatory regions. Future studies can build on this work by investigating genes known to affect other ecologically relevant wild tomato fruit traits such as color and scent. Integrating these molecular studies with field data on wild tomato animal dispersers could provide an unparalleled picture of the evolutionary forces driving fleshy fruit diversity.

3.5 Methods
3.5.1 Plant material and DNA extraction

Our sampling included 42 accessions chosen to span the known phylogenetic and geographic diversity across the tomato clade, consisting of 14 taxa: 12 species of wild tomato, the semi-wild S. lycopersicum var. cerasiforme, and the cultivated tomato S. lycopersicum var. lycopersicum (Fig. 3.1; see Barnett et al. 2023a for a phylogeny of all
42 accessions). Three accessions of each taxa were selected, representing different parts of their geographic ranges when possible. Seeds were obtained from the C. M. Rick Tomato Genetics Resource Center at the University of California, Davis, USA (TGRC, http://tgrc.ucdavis.edu) and the Universitat Politècnica de València, València, Spain (accession IDs and collection locations listed in Appendix C, Table C3). Each accession is representative of an independently sampled population in the wild. For autogamous self-compatible species (S. cheesmaniae, S. galapagense, some S. pimpinellifolium, S. lycopersicum var. cerasiforme, S. lycopersicum var. lycopersicum, S. neorickii) each plant of an accession is likely genetically identical. For others, plants are not identical, as the facultative self-compatible or allogamous self-incompatible accessions are maintained through "mass sibling" pollination in germplasm centers.

Genomic DNA was extracted from ground frozen tissue of young unexpanded leaves using the CTAB method (Porebski & Bailey, 1997), then diluted to ~40 ng/ul for use in PCR reactions.

3.5.2 PCR amplification and sequencing

To obtain sequences of the five target sugar genes (Appendix C, Table C1) for each tomato accession, we produced a set of amplicons that spanned the target gene regions, pooled them together by accession, sequenced each pool with short read Next-Gen sequencing, then recovered the original amplicon sequences computationally.

First we designed primers to amplify at least the transcribed region of each gene, plus ~2 kb upstream (5') of the transcription start site and ~1 kb downstream (3') of the transcription end site if possible, using Geneious Prime v2022.0.2 software (Biomatters, Inc., San Diego, CA, USA). Primer sequences (Appendix C, Table C4) were based on an
alignment of target gene sequence extractions from the cultivated tomato 'SL4.0' (Hosmani et al., 2019) and *S. pennellii 'Spenn'* (Bolger et al., 2014b) reference genomes, using loci with identical or nearly identical sequences in both references to increase the chances of primers working in multiple species. For most of the genes we needed several overlapping sets of primers to cover the region because we were not able to obtain large enough single amplicons. Furthermore, some primers did not work in all species so we designed additional primers to obtain amplicons from as many accessions as possible. We performed 20 ul PCR reactions (recipe and cycling conditions in Appendix C, Table C5) using Thermo Phire Green Hot Start II Polymerase (Thermo Fisher Scientific, Waltham, MA, USA) with amplicon-specific (and in some cases accession-specific) annealing and extension temperatures. PCR products were run on a 1% agarose gel, and concentrations of bands matching the expected target amplicon size were estimated by eye from the gel image relative to a lambda DNA-HindIII digest ladder (New England Biolabs, Ipswich, MA, USA).

We then pooled together the best amplicons into one tube per accession, aiming for equivalent amounts of each amplicon by adjusting the volume added to the pool based on the amplicon's length and concentration. Each pool contained between 7–16 amplicons. To prepare for sequencing, the pool was purified with a Zymo Clean & Concentrator 25 kit (Zymo Research, Irvine, CA, USA), eluted in 38 ul LTE buffer (10 mM Tris-HCl, pH 8.5, 0.1 mM EDTA), quantified with Qubit dsDNA BR assay (Thermo Fisher Scientific, Waltham, MA, USA), and if necessary diluted to a concentration of 30-60 ng/ul in LTE buffer. 35 ul of the diluted sample was sequenced on an Illumina MiSeq PE 150 (Illumina Inc., San Diego, CA, USA) by the Massachusetts General Hospital
DNA Core Complete Amplicon Sequencing service (Cambridge, MA, USA), which produced paired end raw reads of 142 base pairs each.

3.5.3 Assembly and alignment of target sugar gene sequences

To recover the sequences of the original amplicons from these short reads, we first checked for quality with FastQC (Andrews, 2010) and trimmed with Trimmomatic v0.39 (Bolger et al., 2014a) using the options: ILLUMINACLIP:TruSeq2-PE.fa:2:30:10:2:keepBothReads LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:35. We then de novo assembled the trimmed paired reads with Platanus-allee v2.0.2 (Kajitani et al., 2019) using the default assemble and phase options; we chose phased rather than consensus contigs to preserve heterozygotes. To identify which gene each contig corresponded to, the phased contigs were queried against a reference genome, either cultivated tomato SL4.0 for colored-fruited accessions or Spenn for green-fruited accessions, using BLAST+ v2.7.1 (Altschul et al., 1990) with settings -evalue 10 -word_size 9. A custom R script was used to find BLAST hits in the resulting .tsv files that started or ended within the primer-amplified regions of any of the five target sugar genes (coordinates based on the SL4.0 or Spenn reference genomes used to design the primers), then pull out from the allPhaseBlock.fa files any contigs with hits, and finally create a separate .fa file for each gene.

Because not all of these 'gene hit' contigs assembled by Platanus-allee spanned the entire target region in one or two (for heterozygotes) contigs due to the limitations of the assembler, we manually concatenated 'short' contigs when possible. First, the 'gene hit' contigs were imported into Geneious Prime v2023.0.1 (Biomatters Ltd.) and divided into
folders by gene, within which we produced separate alignments for colored-fruited and green-fruited accessions, using MAFFT v7.490 (Katoh et al., 2002) with settings Algorithm = AUTO, Scoring matrix = 200 PAM / k=2, Gap open penalty = 1.53, Offset value = 0.123, and 'Automatically determine sequences' direction' set to on. Reference sequences were extracted from either SL4.0 or Spenn and bounded by the positions of the two most distant primers used to produce the amplicons of that gene. These alignments were sorted by accession to visualize how the 'short' contigs were positioned relative to each other and the target gene region.

We separated each alignment into three regions: CDS plus introns (start codon to stop codon, which included both exons and introns), 5' promoter (anything upstream of start codon, including UTR), and 3' promoter (anything downstream of stop codon, including UTR). We then manually concatenated the 'short' contigs to the extent possible depending on the situation, mainly by masking gaps or ambiguous repeats with Ns.

3.5.4 Molecular evolution analyses

For each gene, separate alignments were produced for the exons only, exons plus introns, and ~2 kb of 5' promoter region using the MAFFT v7.490 (Katoh et al., 2002) algorithm implemented in Geneious Prime v2023.0.1 (Biomatters Ltd.). Prior to analyses, alignments were checked by eye to confirm that gaps (marked as '-') looked like real deletions rather than something produced by the alignment algorithm, and missing data was marked as 'N'.

Maximum likelihood gene trees were made for each region of each gene using the W-IQ-TREE web interface
with the following settings: sequence type/model = nucleotides/find best and apply; branch support = bootstrap, ultrafast, number of replicates = 1000; single branch test: SH-aLRT branch test with number of replicates = 1000; approximate bayes test = TRUE; tree search = perturbation strength 0.5, number of unsuccessful iterations to stop = 100; tree type = phylogram. *Solanum lycopersicoides* was specified as the outgroup; its sequences were obtained through BLAST searches of the cultivated tomato coding sequence for a given gene against the *S. lycopersicoides* chromosomes v1.0 database on SolGenomics (https://solgenomics.net/tools/blast/). The consensus tree files were used to create tree diagrams with the R package GGTREE v3.10.0 (Xu et al. 2022) in R v4.3.2 (R Core Team, 2023).

Nucleotide diversity (pi) was calculated with the *nuc.div* function in the R package PEGAS v1.2 (Paradis 2010) with pairwise.deletion = TRUE, using separate multiple sequence alignment fasta files for each gene region. The nucleotide diversity calculation was done separately for alignments containing (1) all species, (2) colored-fruited species only, and (3) green-fruited species only. The *S. lycopersicoides* outgroup sequences were not included in these analyses, but appropriate extractions from the cultivated tomato 'SL4.0' reference sequence and the *S. pennellii* 'Spenn' reference sequence were included.

Protein sequences were predicted from our nucleotide sequences using Geneious. After copying the exon annotations from the SL4.0 cultivated tomato reference to each of the wild species in a given alignment, exon sequences were extracted and used to produce a translation alignment. Amino acid substitutions were identified from this alignment
using Geneious. To predict whether amino acid changes would be tolerable or intolerable (i.e. deleterious) to the protein, we used the SIFT (Ng and Henikoff 2001) single protein browser tool (https://sift.bii.a-star.edu.sg/www/SIFT_seq_submit2.html) with the cultivated tomato reference protein sequence (from SolGenomics) as the query and default parameters (database = UniProt-SwissProt_2010_09, median conservation of sequences = 3.00, and remove sequences more than 90 percent identical to query).

For selection analyses on protein-coding sequences, the 'exons only' alignments were exported from Geneious in fasta format and were used in the HyPhy v2.5 software suite (www.hyphy.org; Kosakovsky Pond et al. 2020) via the Datamonkey v2.0 server (www.datamonkey.org; Weaver et al. 2018). For each gene, the length of the 'exons only' nucleotide alignment was checked if it was divisible by three and the stop codon was removed. The *S. lycopersicoides* outgroup sequences were not included in these analyses.

As a pre-processing step to screen for the presence of recombination, each alignment was run through GARD (Kosakovsky Pond et al. 2006) and partitioned data files were exported for use in subsequent selection analyses.

As an additional way to test for evidence of selection on coding sequences for the cladewide 'exons only' alignment of each gene, we used the Selecton server (http://selecton.bioinfo.tau.ac.il/; Doron-Faigenboim et al. 2005) to calculate: (1) the number of sites with $\text{Ka/Ks} > 1$, and (2) the likelihood ratio significance test comparing the M8 model (allowing positive selection) with the M8a model (null model, not allowing positive selection). The *S. lycopersicoides* outgroup sequences were not included in these analyses.
As one way of assessing evidence of selection on the 5' promoter for certain interspecies comparisons with known differences in fruit sugar traits, nucleotide substitution rates were compared between the coding region (exons only) and the adjacent 5' promoter region (including 5' UTRs) using the KnKs function with default settings in KaKs_Calculator v3.0 (Zhang 2022).

As an additional way of assessing potential evidence for selection in the 5' promoter region of the target genes, we performed the HKA test comparing polymorphism and divergence using DnaSP v 6.0 (Rozas et al. 2017). Thirty accessions that had clean sequence data for the 5' promoter region of TIV1 were used, with the "intraspecific" group defined as the 12 accessions from the colored-fruited group and the "interspecific" group defined as the 18 accessions from the green-fruited group. "Locus 1" was the 5' promoter sequence (696 bases for TIV1), and "Locus 2" was a set of 54980 genome-wide GBS SNPs (from Barnett et al. 2023a). For each accession, the 5' promoter sequence was concatenated to the GBS SNP sequence (as a string of nucleotides from a fasta file), resulting in a single fasta file with both "loci" on the same line for each accession. The concatenated alignment was imported into DnaSP and the HKA test was run, with the two sequence sets defined as the colored-fruited and green-fruited groups. For TIV1, Locus 1 was defined as site 1 to 696 (the 5' promoter) and Locus 2 as site 697 to 55689 (the GBS SNPs).
3.6 Figures

Figure 3.1 Differences in three ecologically relevant ripe fruit sugar traits among the 13 species of the tomato clade (*Solanum* sect. *Lycopersicon*). The maximum likelihood phylogenetic tree on the left is based on 64,745 SNPs obtained through genotyping by sequencing; numbers at nodes represent bootstrap percentages (Barnett et al. 2023a). Heatmaps show species means of three major ripe fruit sugar traits: sucrose-to-hexose ratio, fructose-to-glucose ratio, and total soluble solids (in degrees Brix). Species means were computed from a total of 206 ripe fruits (~18 per species) from 38 accessions (~3 per species) grown in a common garden in Barnett et al. (2023a); to account for uneven sampling, the means for all fruits from each plant (generally two per plant) were computed first, then per-plant means (generally three plants per accession) were averaged, and finally per-accession means (generally three accessions per species) were averaged to arrive at a per-species mean value for each trait. See Appendix C, Table C3 for list of accessions, and for a more fine-grained look at fruit trait data see boxplots for each species (Appendix C, Fig. C1) and accession (Appendix C, Fig. C2).
Figure 3.2 Maximum likelihood consensus gene trees for three regions of the target genes: (A) CDS (exons) only, (B) CDS (exons) plus Introns, (C) 5' promoters, and (D) LIN5 CDS only and CDS plus Introns, shown separately because only partial CDS sequences and no 5' promoter sequences were obtained. Trees were constructed in IQTree with the following parameters: 'Find best and apply' nucleotide substitution model selected by ModelFinder; 1000 ultrafast bootstraps; SH-aLRT single branch test with 1000 replicates. Trees were rooted on outgroup *Solanum lycopersicoides* (LYD). Tree diagrams were produced with the GGTree package in R. Red text denotes brightly colored-fruited accessions and blue text denotes green-fruited accessions; blue numbers at nodes show bootstrap percentages. Species are abbreviated as follows: ARC = *Solanum arcanum*, CER = *S. lycopersicum* var. cerasiforme, CHE = *S. cheesmaniae*, CHI = *S. chilense*, CHM = *S. chmielewskii*, COR = *S. corneliomulleri*, GAL = *S. galapagense*, HAB = *S. habrochaites*, HUA = *S. huaylasense*, LYC = *S. lycopersicum* var. *lycopersicum* (cultivated tomato), LYD = *S. lycopersicoides*, NEO = *S. neorickii*, PEN = *S. pennelli*, PER = *S. peruvianum*, PIM = *S. pimpinellifolium*. Numbers after species abbreviation represent Tomato Genetics Resource Center ‘LA’ accession IDs. Reference sequences publicly available from SolGenomics are denoted by 'ref'. A lowercase 'a' or 'b' after the numbers denotes the two alleles of a heterozygote.
### 3.7 Tables

Table 3.1 The five tomato fruit sugar-related genes chosen in this study for amplicon sequencing and molecular evolution analyses. nt. = nucleotides, a.a. = amino acids

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Gene ID</th>
<th>Gene length (nt.), transcript on start to end, SL4.0</th>
<th>Type of protein encoded by gene</th>
<th>Protein length (a.a.), ITAG4.0 on Phytozome</th>
<th>Metabolic enzyme or transporter</th>
<th>References for gene's identification and functional validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGPL1</td>
<td>Solyc01g109790</td>
<td>4788</td>
<td>ADP-glucose pyrophosphorylase large subunit</td>
<td>524</td>
<td>Metabolic enzyme</td>
<td>Schaffer et al. 2000; Petreikov et al. 2006</td>
</tr>
<tr>
<td>LIN5</td>
<td>Solyc09g010080</td>
<td>3988</td>
<td>Cell wall acid invertase</td>
<td>607</td>
<td>Metabolic enzyme</td>
<td>Fridman et al. 2004; Zanor et al. 2009</td>
</tr>
<tr>
<td>SUS1</td>
<td>Solyc12g009300</td>
<td>5812</td>
<td>Sucrose synthase</td>
<td>805</td>
<td>Metabolic enzyme</td>
<td>D'Aoust et al. 1999</td>
</tr>
<tr>
<td>SWEET 1a</td>
<td>Solyc04g064610</td>
<td>1836</td>
<td>SWEET glucose transporter</td>
<td>246</td>
<td>Transporter</td>
<td>Levin et al. 2000; Shamai et al. 2018</td>
</tr>
<tr>
<td>TIV1</td>
<td>Solyc03g083910</td>
<td>4518</td>
<td>Vacuolar acid invertase</td>
<td>692</td>
<td>Metabolic enzyme</td>
<td>Chetelat et al. 1995; Klann et al. 1996</td>
</tr>
</tbody>
</table>
Table 3.2 Nucleotide diversity in the new sequences obtained for the target genes, calculated with the nuc.div function in the pegas R package with pairwise.deletion = TRUE. Note: LIN5 is not included in the "exons only" and "5' promoter" regions due to difficulty obtaining sequences that covered those regions. Total number of accessions in alignment includes the cultivated tomato SL4.0 reference sequence and the Solanum pennellii Spenn reference sequence.

<table>
<thead>
<tr>
<th>Region</th>
<th>Gene name</th>
<th>Length of alignment (nucleotides)</th>
<th>Number of accessions in alignment</th>
<th>Nucleotide diversity (pi), all species included</th>
<th>Nucleotide diversity (pi), colored-fruited species only</th>
<th>Nucleotide diversity (pi), green-fruited species only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exons only</td>
<td><em>AGPL1</em></td>
<td>1582</td>
<td>40</td>
<td>0.00654</td>
<td>0.00069</td>
<td>0.00799</td>
</tr>
<tr>
<td>Exons only</td>
<td><em>SUS1</em></td>
<td>2416</td>
<td>32</td>
<td>0.01267</td>
<td>0.00704</td>
<td>0.01361</td>
</tr>
<tr>
<td>Exons only</td>
<td>SWEET1a</td>
<td>751</td>
<td>36</td>
<td>0.00719</td>
<td>0.00259</td>
<td>0.00813</td>
</tr>
<tr>
<td>Exons only</td>
<td>TIV1</td>
<td>1900</td>
<td>36</td>
<td>0.01135</td>
<td>0.00103</td>
<td>0.01272</td>
</tr>
<tr>
<td>Exons and introns</td>
<td><em>AGPL1</em></td>
<td>3838</td>
<td>40</td>
<td>0.01634</td>
<td>0.00162</td>
<td>0.01801</td>
</tr>
<tr>
<td>Exons and introns</td>
<td><em>SUS1</em></td>
<td>4055</td>
<td>33</td>
<td>0.02361</td>
<td>0.01359</td>
<td>0.02441</td>
</tr>
<tr>
<td>Exons and introns</td>
<td>SWEET1a</td>
<td>1600</td>
<td>37</td>
<td>0.01692</td>
<td>0.00311</td>
<td>0.01984</td>
</tr>
<tr>
<td>Exons and introns</td>
<td>TIV1</td>
<td>4150</td>
<td>37</td>
<td>0.02397</td>
<td>0.00275</td>
<td>0.02615</td>
</tr>
<tr>
<td>Exons and introns (partial)</td>
<td>LIN5</td>
<td>2069</td>
<td>30</td>
<td>0.02525</td>
<td>0.00506</td>
<td>0.02511</td>
</tr>
<tr>
<td>5' promoter</td>
<td><em>AGPL1</em></td>
<td>2720</td>
<td>40</td>
<td>0.05626</td>
<td>0.01233</td>
<td>0.05939</td>
</tr>
<tr>
<td>5' promoter</td>
<td><em>SUS1</em></td>
<td>5595</td>
<td>27</td>
<td>0.03065</td>
<td>0.01418</td>
<td>0.03593</td>
</tr>
<tr>
<td>5' promoter</td>
<td>SWEET1a</td>
<td>3871</td>
<td>38</td>
<td>0.04877</td>
<td>0.00398</td>
<td>0.05326</td>
</tr>
<tr>
<td>5' promoter</td>
<td>TIV1</td>
<td>8575</td>
<td>37</td>
<td>0.04481</td>
<td>0.00445</td>
<td>0.04953</td>
</tr>
</tbody>
</table>
Table 3.3 Protein sequence comparisons among the four genes for which full coding sequences were obtained in this study. a.a. = amino acids

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Total accessions sequenced in this study (Number of which were heterozygotes)</th>
<th>Length of protein in a.a.</th>
<th>Total number of indels</th>
<th>Total number of a.a. substitutions (Percent of protein length)</th>
<th>Number of a.a. substitutions present in only one accession (Percent of total a.a. substitutions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGPL1</td>
<td>39 (7)</td>
<td>528</td>
<td>2</td>
<td>29 (5.5%)</td>
<td>16 (55%)</td>
</tr>
<tr>
<td>SUS1</td>
<td>31 (2)</td>
<td>806</td>
<td>0</td>
<td>25 (3.1%)</td>
<td>17 (68%)</td>
</tr>
<tr>
<td>SWEET 1a</td>
<td>36 (5)</td>
<td>251</td>
<td>4</td>
<td>17 (6.8%)</td>
<td>8 (47%)</td>
</tr>
<tr>
<td>TIV1</td>
<td>36 (5)</td>
<td>634</td>
<td>0</td>
<td>63 (9.9%)</td>
<td>32 (51%)</td>
</tr>
</tbody>
</table>
Table 3.4 Cladewide tests of selection on coding region (exons only) of each gene with Selecton and HyPhy. Tests in HyPhy were done on a partitioned alignment after inferring recombination breakpoints with GARD. Note: *LIN5* is not included due to difficulty obtaining a complete coding sequence.

<table>
<thead>
<tr>
<th>Gene (number of sequences in alignment)</th>
<th>Total number of amino acid sites</th>
<th>Selecton: Number of sites with Ka/Ks &gt;1</th>
<th>Selecton: results of likelihood ratio significance test comparing M8 model (allowing positive selection) with M8a model (null model, not allowing positive selection).</th>
<th>HyPhy: Mean tested omega (Ka/Ks) of BUSTED unconstrained models (CoV in parentheses) Colored / Green = colored- or green- fruited species only</th>
<th>HyPhy: Evidence of episodic diversifying selection at at least one site on one branch, based on BUSTED likelihood ratio test (LRT), all branches selected?</th>
<th>HyPhy: Number of sites under pervasive positive selection according to FUBAR</th>
<th>HyPhy: Number of sites under episodic positive selection according to MEME</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGPL1 (47)</td>
<td>527</td>
<td>9</td>
<td>Positive selection in the protein is non-significant</td>
<td>All branches: 0.1399 (0.0) Colored: 0.3323 (1.5) Green: 0.1364 (0.0)</td>
<td>No, p=0.50; 0 sites with ER&gt;10 suggestive of positive selection</td>
<td>1 site (codon 513)</td>
<td>0 sites with significant (p&lt;0.1) positive selection</td>
</tr>
<tr>
<td>SUS1 (34)</td>
<td>805</td>
<td>0</td>
<td>No positively selected sites found in the protein</td>
<td>All branches: 0.1447 (9.8) Colored: 0.184 (8.5) Green: 0.126 (10.1)</td>
<td>Yes, p=0.00005; 5 sites with ER&gt;10</td>
<td>1 site (codon 126)</td>
<td>0 sites with significant (p&lt;0.1) positive selection</td>
</tr>
<tr>
<td>SWEET1a (41)</td>
<td>250</td>
<td>17</td>
<td>Likelihood ratio test shows significance level of 0.05</td>
<td>All branches: 0.4357 (10.8) Colored: 1.694 (6.1) Green: 0.3865 (14.1)</td>
<td>Yes, p=0.0305; 1 site with ER&gt;10 (codon 18)</td>
<td>0 sites</td>
<td>1 site with significant (p&lt;0.1) positive selection (codon 18)</td>
</tr>
<tr>
<td>TIV1 (41)</td>
<td>633</td>
<td>53</td>
<td>Likelihood ratio test shows significance level of 0.001</td>
<td>All branches: 0.2592 (0.0) Colored: 0.4521 (1.3) Green: 0.2548 (0.0)</td>
<td>No, p=0.50; 0 sites with ER&gt;10</td>
<td>5 sites (codons 342, 378, 381, 433, 478)</td>
<td>2 sites with significant (p&lt;0.1) positive selection (codons 342 and 433)</td>
</tr>
</tbody>
</table>
Table 3.5 Comparison of substitution rates between coding region (exons only) and noncoding region (5' UTR and promoter) for pairs of accessions with differences in fruit sugar traits that were hypothesized to show signs of selection in associated genes. Computed with the KnKs function in KaKs_Calculator (Zhang 2022). Among the four genes within each comparison, dark shading = highest value and light shading = lowest value. Note: LIN5 not included due to difficulty obtaining a complete coding sequence.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Kn/Ks</th>
<th>Ka/Ks</th>
<th>Kn</th>
<th>Ka</th>
<th>Ks</th>
<th>Non-coding length</th>
<th>Non-coding subs.</th>
<th>Coding length</th>
<th>Coding subs.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison: Red-fruited (PIM4717_SPMB6) vs. Green-fruited (ARC2157_SA18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGPL1</td>
<td>1.966</td>
<td>0.030</td>
<td>0.053</td>
<td>0.0008</td>
<td>0.027</td>
<td>2179</td>
<td>110</td>
<td>1572</td>
<td>10</td>
</tr>
<tr>
<td>SWEET 1a</td>
<td>9.345</td>
<td>0.339</td>
<td>0.052</td>
<td>0.0019</td>
<td>0.006</td>
<td>2710</td>
<td>136</td>
<td>717</td>
<td>2</td>
</tr>
<tr>
<td>SUS1</td>
<td>1.957</td>
<td>0.075</td>
<td>0.031</td>
<td>0.0012</td>
<td>0.016</td>
<td>3598</td>
<td>108</td>
<td>2415</td>
<td>13</td>
</tr>
<tr>
<td>TIV1</td>
<td>4.442</td>
<td>0.924</td>
<td>0.035</td>
<td>0.0074</td>
<td>0.008</td>
<td>2239</td>
<td>77</td>
<td>1899</td>
<td>14</td>
</tr>
<tr>
<td>Comparison: Red-fruited (PIM0373_SPM01 or PIM4717_SPMB6) vs. S. pennellii (PEN2963_SPE21 or PEN1809_SPE16)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGPL1</td>
<td>6.960</td>
<td>0.159</td>
<td>0.069</td>
<td>0.0016</td>
<td>0.0099</td>
<td>2172</td>
<td>142</td>
<td>1572</td>
<td>5</td>
</tr>
<tr>
<td>SWEET 1a</td>
<td>3.530</td>
<td>0.565</td>
<td>0.062</td>
<td>0.0099</td>
<td>0.0177</td>
<td>2703</td>
<td>160</td>
<td>687</td>
<td>8</td>
</tr>
<tr>
<td>SUS1</td>
<td>0.517</td>
<td>0.025</td>
<td>0.036</td>
<td>0.0017</td>
<td>0.0700</td>
<td>3755</td>
<td>132</td>
<td>2247</td>
<td>37</td>
</tr>
<tr>
<td>TIV1</td>
<td>5.019</td>
<td>0.692</td>
<td>0.091</td>
<td>0.0126</td>
<td>0.0181</td>
<td>2203</td>
<td>186</td>
<td>1899</td>
<td>26</td>
</tr>
<tr>
<td>Comparison: S. pennellii (PEN2963_SPE21 or PEN1809_SPE16) vs. another green-fruited (COR1305_SCR09)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGPL1</td>
<td>3.862</td>
<td>0.110</td>
<td>0.059</td>
<td>0.0017</td>
<td>0.015</td>
<td>2072</td>
<td>117</td>
<td>1581</td>
<td>8</td>
</tr>
<tr>
<td>SWEET 1a</td>
<td>2.345</td>
<td>0.165</td>
<td>0.062</td>
<td>0.0044</td>
<td>0.026</td>
<td>2747</td>
<td>161</td>
<td>624</td>
<td>6</td>
</tr>
<tr>
<td>SUS1</td>
<td>0.551</td>
<td>0.026</td>
<td>0.037</td>
<td>0.0017</td>
<td>0.068</td>
<td>3807</td>
<td>138</td>
<td>2247</td>
<td>37</td>
</tr>
<tr>
<td>TIV1</td>
<td>3.300</td>
<td>0.446</td>
<td>0.070</td>
<td>0.0094</td>
<td>0.021</td>
<td>1730</td>
<td>113</td>
<td>1899</td>
<td>23</td>
</tr>
<tr>
<td>Comparison: S. cheesmaniae (CHE1412_SC19) vs. S. galapagense (GAL1044_SG30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
*Note: the sequences of these two accessions are so similar, the Ks values are extremely low.*

| AGPL1 | 118.216 | 50 | 0.0018 | 0.0007 | 0.0000 | 2282 | 4 | 1563 | 1 |
| SWEET 1a | 667.550 | 0.094 | 0.0007 | 0.0000 | 0.0000 | 2681 | 2 | 735 | 0 |
| SUS1  | 0.253 | 0.001 | 0.0009 | 0.0000 | 0.0034 | 3492 | 3 | 2271 | 1 |
| TIV1 | 159.445 | 50 | 0.0021 | 0.0006 | 0.0000 | 5330 | 11 | 1899 | 1 |
Table 3.6 To assess potential evidence for selection in the 5' promoter region of the target genes in the colored-fruited species group relative to the green-fruited group, we performed the HKA test comparing polymorphism and divergence using DnaSP v 6.0 (Rozas et al. 2017). Accessions that had clean sequence data for the 5' promoter region of the target gene were used, with the "intraspecific" group defined as the accessions from the colored-fruited group (12-13 accessions) and the "interspecific" group defined as the accessions from the green-fruited group (12-24 accessions). "Locus 1" was the 5' promoter sequence and "Locus 2" was a set of 54980 genome-wide GBS SNPs (from Barnett et al. 2023a).

<table>
<thead>
<tr>
<th>Region/Locus (number of sites used in the HKA test calculation)</th>
<th>Polymorphism: observed segregating sites within colored-fruited group</th>
<th>Divergence: observed average number of differences between colored-fruited and green-fruited groups</th>
<th>Ratio of Polymorphism / Divergence</th>
<th>p-value for chi-squared test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus 1: TIV1 5' promoter (525 sites)</td>
<td>6</td>
<td>28.92</td>
<td>0.207</td>
<td><strong>0.0237</strong></td>
</tr>
<tr>
<td>Locus 2: Genome-wide GBS SNPs (11552 sites)</td>
<td>1115</td>
<td>1244.14</td>
<td>0.896</td>
<td></td>
</tr>
<tr>
<td>Locus 1: SWEET1a 5' promoter (1752 sites)</td>
<td>22</td>
<td>47.05</td>
<td>0.468</td>
<td>0.1427</td>
</tr>
<tr>
<td>Locus 2: Genome-wide GBS SNPs (26146 sites)</td>
<td>2859</td>
<td>2880.89</td>
<td>0.992</td>
<td></td>
</tr>
<tr>
<td>Locus 1: SUS1 5' promoter (2261 sites)</td>
<td>52</td>
<td>70.40</td>
<td>0.739</td>
<td>0.6929</td>
</tr>
<tr>
<td>Locus 2: Genome-wide GBS SNPs (13237 sites)</td>
<td>1369</td>
<td>1546.27</td>
<td>0.885</td>
<td></td>
</tr>
<tr>
<td>Locus 1: AGPL1 5' promoter (1435 sites)</td>
<td>50</td>
<td>74.18</td>
<td>0.674</td>
<td>0.4711</td>
</tr>
<tr>
<td>Locus 2: Genome-wide GBS SNPs (10481 sites)</td>
<td>1116</td>
<td>1182.71</td>
<td>0.943</td>
<td></td>
</tr>
</tbody>
</table>
This dissertation further develops the tomato clade as a model system for fleshy fruit evolution, synthesizing knowledge of variation in animal disperser-relevant ripe fruit traits across all 13 wild tomato species at the phenotypic and molecular levels. Results show that two syndromes of covarying ripe fruit traits appear to exist: 1) red/orange/yellow fruits that are high in glucose and fructose, low in sucrose and malic acid, low in esters, and high in nitrogen-containing VOCs; and 2) green fruits that are low in glucose and fructose, high in sucrose and malic acid, high in esters, and low in nitrogen-containing VOCs. Furthermore, fruity-scented esters could attract animals and may be an honest signal of a sugar reward. The difference in sugar composition between the brightly colored-fruited and green-fruited groups may owe its genetic origins to a retrotransposon insertion in the 5' promoter region of the vacuolar invertase gene TIV1; the insertion is shared by the colored-fruited species, alters gene expression in a way that shifts ripe fruit sugar composition from sucrose accumulation to hexose accumulation, and bears molecular signs of selection.

These patterns establish the clade as a promising system for testing hypotheses about the evolutionary drivers behind early-stage fleshy fruit trait divergence, particularly the influence of animal disperser preferences. Ecological field studies in the native ranges of wild tomato species would be especially informative because little is known about which animals eat the fruits in the wild. Additional work on the cladewide molecular evolutionary dynamics of genes affecting ripe fruit color and scent would also enable a more complete picture of how disperser-relevant fruit traits are evolving.
APPENDIX A

SUPPORTING INFORMATION FOR CHAPTER 1: EVIDENCE OF FRUIT SYNDROMES IN THE RECENTLY DIVERGED WILD TOMATO CLADE OPENS NEW POSSIBILITIES FOR THE STUDY OF FLESHY FRUIT EVOLUTION

(a)
(d) FreshWeight_grams (Page's lambda = 0, p value = 1)

(e) SeedCount (Page's lambda = 0.953, p value = 0)
Fig. A1 Trait variation with respect to phylogeny for all 21 fruit traits: (a) Diameter in millimeters, (b) Length in millimeters, (c) Diameter to length ratio, (d) Fresh weight in grams, (e) Seed count, (f) Lobedness degree, (g) Pericarp area ratio, (h) Color lightness
average, (i) Color hue average, (j) Color chroma average, (k) Glucose concentration in milligrams/milliliter (l) Fructose concentration in milligrams/milliliter, (m) Sucrose concentration in milligrams/milliliter, (n) Total sugars (glucose+fructose+sucrose) concentration in milligrams/milliliter, (o) Citric acid concentration in milligrams/milliliter, (p) Malic acid concentration in milligrams/milliliter, (q) Total acids (citric+malic) concentration in milligrams/milliliter, (r) Brix in degrees, (s) Sugars to acids ratio, (t) Sucrose to Hexose (glucose+fructose) ratio, (u) Malic to citric acid ratio. Left-hand column: plots showing heatmaps of accession mean trait values distributed across our phylogeny, with Pagel's lambda phylogenetic signal results for that fruit trait variable. Right-hand column: boxplots of individual fruit trait raw values for each accession; colors represent species, abbreviated as follows: ARC = Solanum arcanum, CER = S. lycopersicum var. cerasiforme, CHE = S. cheesmaniae, CHI = S. chilense, CHM = S. chmielewskii, COR = S. corneliomulleri, GAL = S. galapagense, HAB = S. habrochaites, HUA = S. huaylasense, NEO = S. neorickii, PEN = S. pennellii, PER = S. peruvianum, PIM = S. pimpinellifolium. ‘LA’ or ‘BGV’ numbers represent Tomato Genetics Resource Center accession IDs.
Fig. A2 Separate principal component analysis (PCA) plots (non-phylogenetic) of morphology, color, and nutrition phenotypic traits for 208 fruits representing 38 accessions. All variables were centered and scaled. In left-hand plots: points = individual fruits; 3-letter abbreviations = species names, abbreviated as follows: ARC = Solanum arcanum, CER = S. lycopersicum var. cerasiforme, CHE = S. cheesmaniae, CHI = S. chilense, CHM = S. chmielewskii, COR = S. corneliomulleri, GAL = S. galapagense, HAB = S. habrochaites, HUA = S. huaylasense, NEO = S. neorickii, PEN = S. pennellii, PER = S. peruvianum, PIM = S. pimpinellifolium; colored polygons = species. Red circles surround the colored-fruited species, blue circles the green-fruited species. Right-hand plots show PCA factor loadings of the variables used to create the corresponding left-hand plot. When only seven morphology traits are considered, (a), most fruits cluster together in the middle of the plot, but fruits that are large and oval with a low pericarp area ratio and high seed count, principally from S. pennellii, cluster in the top right, while fruits that are large and spherical with a low pericarp area ratio, primarily from S. lycopersicum var. cerasiforme, cluster in the bottom right. With just three color variables (b), red/orange/yellow fruits cluster together fairly tightly and are mostly distinct from the greenish fruits, which are more spread out. The nutrition-only PCA (c) shows quite a bit of overlap in the middle of the plot, but along PC2 there is a somewhat distinct cluster of fruits high in glucose and fructose but low in malic acid.
Fig. A3 Plots of the three strongest fruit-climate associations according to both phylogenetic generalized least squares (PGLS) models and non-phylogenetic Pearson correlation tests. Regression line slope, standard error (SE), p-value, and Pagel's λ are from PGLS models with lambda = "ML" on 36 degrees of freedom. Points represent accession mean raw values, with colors denoting species, abbreviated as follows: ARC = Solanum arcanum, CER = S. lycopersicum var. cerasiforme, CHE = S. cheesmaniae, CHI = S. chilense, CHM = S. chmielewskii, COR = S. corneliomulleri, GAL = S. galapagense, HAB = S. habrochaites, HUA = S. huaylasense, NEO = S. neorickii, PEN = S. pennellii, PER = S. peruvianum, PIM = S. pimpinellifolium. ‘LA’ or ‘BGV’ numbers represent Tomato Genetics Resource Center accession IDs.
Fig. A4 Differences between colored-fruited and green-fruited species groups in total sugar concentration. Colored points represent accessions and are color-coded by species, abbreviated as follows: ARC = *Solanum arcanum*, CER = *S. lycopersicum* var. *cerasiforme*, CHE = *S. cheesmaniae*, CHI = *S. chilense*, CHM = *S. chmielewskii*, COR = *S. corneliomulleri*, GAL = *S. galapagense*, HAB = *S. habrochaites*, HUA = *S. huarayasense*, NEO = *S. neorickii*, PEN = *S. pennellii*, PER = *S. peruvianum*, PIM = *S. pimpinellifolium*. Raw (untransformed) accession mean values were used; sample sizes were 12 for the colored group and 26 for the green group. Lambda and p-value is from a phylogenetically controlled generalized least-squares (PGLS) model with lambda="ML".
Methods A1 Sugar and acid assay details.

We quantified concentrations of glucose, fructose and sucrose as well as citrate and malate in liquid extracts from thawed pericarps. Samples from individual fruits were measured in duplicate with absorbance-based assay kits (Megazyme, Bray, Ireland), UV transparent 96-well plates (Greiner Bio-One, Monroe, North Carolina, USA), and a SpectraMax M5 Microplate Reader (Molecular Devices, San Jose, CA, USA). Totals and ratios were calculated mathematically from the raw values of the individual sugars or acids. For samples with values of zero for one of the sugars or acids, infinite ratios (due to a zero in the denominator) were replaced with the maximum finite sample value for that ratio. For log transformations of the data, raw values of zero (resulting in a log value of -Inf) were replaced with log values equal to the minimum finite log value for that variable. Brix, or total soluble solids, was measured for drops of the liquid pericarp extracts using a MA871 Digital Brix Refractometer (Milwaukee Instruments, Rocky Mount, North Carolina, USA).

Methods A2 GBS protocol details.

A reduced representation Genotyping-by-Sequencing (GBS) library was prepared according to a protocol modified from Elshire et al., (2011). Genomic DNA was digested with the ApeKI restriction enzyme (New England Biolabs, Ipswich, MA, USA, product # R0643S) for 2 hours at 75 °C. Common and barcode adapters (listed in Table below) were ligated to the digested DNA using T4 DNA Ligase (New England Biolabs, product # M0202S) in 45 µl reactions for 60 minutes at 22 °C then 20 minutes at 65 °C. Ligated samples were quantified for dsDNA with a Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), then pooled into a single tube containing 70 ng DNA of each sample, and cleaned with a QIAquick column PCR purification kit (Qiagen, Germantown, MD, USA) and eluted in 35 µl EB Buffer. Size-selection for 200-500 bp was done with a BluePippin 1.5% agarose cassette (Sage Science, Beverly, MA, USA). The size-selected pool was then amplified in 20 µl volumes via PCR with Phusion polymerase (New England Biolabs, product # M0530S; primer sequences listed in Table below) and a thermocycler protocol of initial denaturation at 72 °C for 5 minutes; 40 cycles of 98 °C for 30 sec., 65 °C for 30 sec., and 72 °C for 45 sec.; and final extension at 72 °C for 5 minutes. The resulting amplicons were cleaned with a QIAquick column PCR purification kit and size selected for 300-600 bp with BluePippin. After confirmation of fragment size distribution with BioAnalyzer (Agilent Technologies, Palo Alto, CA, USA), the library was sequenced on one lane of an Illumina HiSeq PE 150 (Illumina Inc., San Diego, CA, USA) by Novogene, Inc. (Sacramento, CA, USA).

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Barcode</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common forward adapter</td>
<td>5’-CWGAGATCGGAAGAGCGGTTCAGCAGGAATGCCCGAG</td>
<td>BC01</td>
<td>AGGC</td>
</tr>
<tr>
<td>Common Reverse adapter</td>
<td>5’-CTCGGCATTCTCTGCTGAACCGCTCTTCCGATCT</td>
<td>BC02</td>
<td>GATT</td>
</tr>
<tr>
<td>GBS PCR primer #1</td>
<td>5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT</td>
<td>BC03</td>
<td>CGAA</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------------------------------------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>GBS PCR primer #2</td>
<td>5'-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATCT</td>
<td>BC04</td>
<td>GTCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC05</td>
<td>ACAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC06</td>
<td>GCGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC07</td>
<td>TGCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC08</td>
<td>CCTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC09</td>
<td>AACA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC10</td>
<td>CTGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC11</td>
<td>TAAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC12</td>
<td>CTCGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC13</td>
<td>ACCGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC14</td>
<td>CGTCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC15</td>
<td>TCGCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC16</td>
<td>CGCAT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC17</td>
<td>GCACT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC18</td>
<td>AGCTA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC19</td>
<td>CAGCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC20</td>
<td>TCAGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC21</td>
<td>CATGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC22</td>
<td>GCTAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC23</td>
<td>CTGAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC24</td>
<td>GACTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC25</td>
<td>ACGTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC26</td>
<td>CGATC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC27</td>
<td>ATGCGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC28</td>
<td>TCCAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC29</td>
<td>GCTTGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC30</td>
<td>TCCGTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC31</td>
<td>TCGGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC32</td>
<td>GTCTTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC33</td>
<td>CGGAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC34</td>
<td>CGTCTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC35</td>
<td>CACGTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC36</td>
<td>GTTAGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC37</td>
<td>AGCATT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC38</td>
<td>CTCCGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC39</td>
<td>TTGGCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC40</td>
<td>CCACGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC41</td>
<td>GATGTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC42</td>
<td>TGTTAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC43</td>
<td>CAGTGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC44</td>
<td>GCCTAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC45</td>
<td>AGTGGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC46</td>
<td>TGACCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC47</td>
<td>TTGCAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC48</td>
<td>CTAGCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC49</td>
<td>TAGTGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC50</td>
<td>GTTCCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC51</td>
<td>TGAATC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC52</td>
<td>TGCCTA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC53</td>
<td>ATCGTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC54</td>
<td>GTCACT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC55</td>
<td>CATCGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC56</td>
<td>AGGTCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC57</td>
<td>CCTGCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC58</td>
<td>GTACTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC59</td>
<td>ACCTGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC60</td>
<td>CTTGAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC61</td>
<td>ACCGCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC62</td>
<td>GCTCAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC63</td>
<td>CTATGC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC64</td>
<td>CAGCCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC65</td>
<td>GCCATC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC66</td>
<td>CGCTCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC67</td>
<td>ACGCTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC68</td>
<td>CTGATC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC69</td>
<td>CCAGTC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC70</td>
<td>TACGCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC71</td>
<td>CCGTAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC72</td>
<td>TCGACC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC73</td>
<td>GCATCC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC74</td>
<td>CGTACC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC75</td>
<td>ATTCAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC76</td>
<td>TATCTGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC77</td>
<td>TTCAAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC78</td>
<td>GAGCAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC79</td>
<td>TCTAGGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC80</td>
<td>ACACGGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC81</td>
<td>GACGTGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC82</td>
<td>TCATAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC83</td>
<td>TTACGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>Sequence processing, variant calling, and phylogenetic tree construction details.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequence raw read quality was verified with FastQC (Andrews, 2010). Reads were demultiplexed, trimmed, aligned to the most recent version of the cultivated tomato reference genome, SL4.0 (Hosmani et al., 2019), and variant called with the Fast-GBS.v2 pipeline (Torkamaneh et al., 2020) using default settings except for the imputation step, which was skipped. The resulting variant call file (VCF) was filtered with vcftools v0.1.14 (Danecek et al., 2011) to include only bi-allelic SNPs present in &gt;90% of individuals with a minor allele count of at least two and heterozygote counts of &lt;20 out of 86 individuals.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The set of filtered SNPs described above was used to construct a maximum likelihood consensus tree with IQ-TREE v1.6.3 (Nguyen et al., 2015) after removing sites considered invariant by the program. ModelFinder (Kalyaanamoorthy et al., 2017) selected the best nucleotide substitution model with ascertainment bias correction (+ASC), and a consensus tree was produced via ultrafast bootstrap (Hoang et al., 2018) and SH-aLRT likelihood ratio test (Guindon et al., 2010) with 1000 replicates. A phylogenetic tree diagram was created with the R package GGTREE v1.16.6 (Yu et al., 2017).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>Historical climate and soil data.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
|        | Data for 19 bioclimatic variables (provided as annual values), as well as solar radiation, wind speed, and water vapor pressure (provided as monthly values from which annual average, max, min, and coefficient of variation were calculated) were downloaded.
(in January 2022) from WORLDCLIM version 2.1 historical data for 1970-2000 at 30 seconds resolution. Soil variables were downloaded from SoilGrids 2.0 (Poggio et al., 2021) at 250 m cell size. Accession collection site latitude/longitude coordinates were used to assign values to each accession (Appendix A, Table A1).
Fig. B1 Map of collection locations for 38 wild accessions initially planted for this study, color-coded by species abbreviated as follows: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum var. lycopersicum NEO = Solanum neorickii, PEN = Solanum pennellii, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium; numbers after 3-letter species abbreviations are Tomato Genetics Resource Center (TGRC) 'LA' accession numbers, except for CER_B8189 (full accession ID BGV008189) which does not have an 'LA' number. Only 31 of these accessions (Appendix B, Table B1) produced enough fruit for volatiles quantification. Countries shown are the southern part of Colombia, Ecuador, Peru, and the northern part of Chile.
Fig. B2 PCA of centered and scaled raw values (in ng/gfw/hr) for all 66 compounds. Produced with prcomp function in R. Points represent accession averages (32 total), with colored polygons denoting species (14 total). Three-letter abbreviations denote species as follows: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum NEO = Solanum neorickii, PEN = Solanum pennellii, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium. Numbers after species abbreviations are Tomato Genetics Resource Center (TGRC) 'LA' accession numbers, except for CER_B8189 (full accession ID BGV008189) which does not have an 'LA' number.
A) Carotenoid / Terpene derived compounds

B) Non-aromatic amino acid derived compounds
C) Lipid derived compounds

D) Lipid derived non-ester compounds
E) Phenylalanine-derived compounds

F) Ester compounds
G) Nitrogen-containing compounds

Fig. B3 Heatmaps of compounds grouped by biosynthetic pathway. A) Carotenoid / Terpene derived compounds, B) Non-aromatic amino acid derived compounds, C) Lipid derived compounds, D) Lipid derived non-ester compounds, E) Phenylalanine derived compounds, F) Ester compounds, G) Nitrogen-containing compounds. See Appendix B, Table B4 for details on the compounds. Categories are based on the compound's precursor molecule. Data values (in ng/gfw/hr) were log (base 10) transformed but not scaled. Columns represent accession averages, with color-coding along the top denoting species. Accessions arranged phylogenetically based on a tree from Barnett et al. (2023a) with colored-fruited species on the left and green-fruited species on the right. Three-letter abbreviations denote species as follows: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum NEO = Solanum neorickii, PEN = Solanum pennellii, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium. Numbers after 3-letter species abbreviations are Tomato Genetics Resource Center (TGRC) 'LA' accession numbers, except for CER_B8189 (full accession ID BGV008189) which does not have an 'LA' number.
Fig. B4 Differences among 14 species in the level of cis-3-hexenal. Raw (untransformed) accession mean values are shown. A) Boxplots of each species; 32 accessions total, some species only had one accession. Species are abbreviated as follows: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum NEO = Solanum neorickii, PEN = Solanum pennellii, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium. ANOVA p-values for comparisons between LYC and CER (p = 0.015), and LYC and PIM (P = 0.099) are shown (18 degrees of freedom). Lambda and p-value is from a phylogenetically controlled generalized least-squares (PGLS) model with lambda="ML". B) Plot showing a heatmap of how cis-3-hexenal levels were distributed across an accession-level phylogenetic tree based on Barnett et al. (2023a), produced with
the function `contMap` from the R package `phytools` (Revell, 2012). Pagel's lambda phylogenetic signal test results (with p-value) for cis-3-hexenal accession mean values are shown above the tree; key to the heatmap is shown in the bottom left, with accession mean values ranging from 0 to 46.993 ng/gfw/hr.

Fig. B5 Association between total sugar concentration (glucose + fructose + sucrose, in mg/mL) and ester levels, displayed as A) total sum (in ng/gfw/hr) of the 17 ester compounds quantified, and B) percentage of the total sum of all 66 compounds, based on raw/untransformed accession mean values. Colored points represent accessions and are color-coded by species, abbreviated as follows: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum, NEO = Solanum neorickii, PEN = Solanum pennellii, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium; numbers after 3-letter species abbreviations are Tomato Genetics Resource Center (TGRC) 'LA' accession numbers, except for CER_B8189 (full accession ID BGV008189) which does not have an 'LA' number. 31 accessions are included; the cultivated variety Ailsa Craig was removed for this analysis because it was influenced by human selection. Pearson's r and phylogenetic generalized least squares (PGLS) models were produced from raw/untransformed accession mean values, with 29 degrees of freedom and lambda = "ML" for the PGLS models.
Fig. B6 Association between total sugar concentration (glucose + fructose + sucrose, in mg/mL) and terpene levels, displayed as total sum (in ng/gfw/hr) of the five terpene compounds quantified (left plot), and percentage of the total sum of all 66 compounds (right plot), based on raw/untransformed accession mean values. Colored points represent accessions and are color-coded by species, abbreviated as follows: ARC = *Solanum arcanum*, CER = *Solanum lycopersicum* var. *cerasiforme*, CHE = *Solanum cheesmaniae*, CHI = *Solanum chilense*, CHM = *Solanum chmielewskii*, COR = *Solanum corneliomulleri*, GAL = *Solanum galapagense*, HAB = *Solanum habrochaites*, HUA = *Solanum huaylasense*, LYC = *Solanum lycopersicum* var. *lycopersicum*, NEO = *Solanum neorickii*, PEN = *Solanum pennellii*, PER = *Solanum peruvianum*, PIM = *Solanum pimpinellifolium*. 30 out of 32 accessions are included in these plots; the cultivated variety Ailsa Craig was removed for this analysis because it was influenced by human selection, and *S. habrochaites* LA2329 was removed because of exceptionally high terpene values (602.4 ng/gfw/hr sum and 54.7 percent). Pearson’s $r$ and phylogenetic generalized least squares (PGLS) models were produced from raw/untransformed accession mean values, with lambda = “ML” for the PGLS models. The five mono- and sesqui-terpenes compounds included were alpha-caryophyllene, alpha-curcumene, alphahapinene, beta-caryophyllene, zingiberene.
Pearson's $r = 0.121$
PGLS $p$-value = 0.717, lambda = 0

Fig. B7 Association between total sugar concentration (glucose + fructose + sucrose, in mg/mL) and apocarotenoid levels, displayed as total sum (in ng/gfw/hr) of the 2 carotenoid-derived compounds quantified (left plot), and percentage of the total sum of all 66 compounds (right plot), based on raw/untransformed accession mean values. Colored points represent accessions and are color-coded by species, abbreviated as follows: ARC = Solanum arcanum, CER = Solanum lycopersicum var. cerasiforme, CHE = Solanum cheesmaniae, CHI = Solanum chilense, CHM = Solanum chmielewskii, COR = Solanum corneliomulleri, GAL = Solanum galapagense, HAB = Solanum habrochaites, HUA = Solanum huaylasense, LYC = Solanum lycopersicum var. lycopersicum NEO = Solanum neorickii, PEN = Solanum pennellii, PER = Solanum peruvianum, PIM = Solanum pimpinellifolium. Only 12 colored-fruited accessions are included because the apocarotenoids were not detected in most green-fruited accessions; the cultivated variety Ailsa Craig was removed for this analysis because it was influenced by human selection. Pearson's $r$ and phylogenetic generalized least squares (PGLS) models were produced from raw/untransformed accession mean values, with lambda = "ML" for the PGLS models. The two VOC compounds included were geranyacetone and 6-methyl-5-hepten-2-one.
APPENDIX C

SUPPORTING INFORMATION FOR CHAPTER 3: MOLECULAR EVOLUTION OF GENES UNDERLYING TOMATO RIPE FRUIT SUGAR TRAITS ACROSS 13 SPECIES OF WILD TOMATO RELATIVES

Fig. C1 Boxplots by species of three ecologically relevant ripe fruit sugar traits, based on 206 individual fruits (~18 per species; data from Barnett et al. 2023a). (A) sucrose-to-hexose ratio, (B) fructose-to-glucose ratio, (C) total soluble solids (in degrees Brix). ANOVA tests were performed with the `aov()` function in R. Significance codes: '****' = <0.001, '***' = <0.01, '**' = <0.05.
Fig. C2 Boxplots of eight ripe fruit sugar traits for each of the 38 accessions used in this study, based on 206 individual fruit raw trait values (~6 per accession). With the
exception of the fructose-to-glucose ratio plot, these plots were previously published in a slightly different form as Figure S1 in Barnett et al. 2023a). Colors represent species, arranged phylogenetically and abbreviated as follows: ARC = *Solanum arcanum*, CER = *S. lycopersicum* var. cerasiforme, CHE = *S. cheesmaniae*, CHI = *S. chilense*, CHM = *S. chmielewskii*, COR = *S. corneliomulleri*, GAL = *S. galapagense*, HAB = *S. habrochaites*, HUA = *S. huaylasense*, NEO = *S. neorickii*, PEN = *S. pennellii*, PER = *S. peruvianum*, PIM = *S. pimpinellifolium*. ‘LA’ or ‘BGV’ numbers represent Tomato Genetics Resource Center accession IDs. (A) sucrose-to-hexose(glucose+fructose) ratio, (B) fructose-to-glucose ratio, (C) total soluble solids (in degrees Brix), (D) glucose concentration in milligrams/milliliter, (E) fructose concentration in milligrams/milliliter, (F) sucrose concentration in milligrams/milliliter, (G) total sugars (glucose+fructose+sucrose) concentration in milligrams/milliliter, (H) sugars (glucose+fructose+sucrose) to acids (citric+malic) ratio.
Fig. C3 Gene expression at different fruit developmental stages (B = breaker, IG = immature green, MG = mature green, R = ripe) for the five target genes. Boxplots are color-coded by species. Data from Doron-Faigenboim et al. 2023.


Revelle, W. 2023. Procedures for Psychological, Psychometric, and Personality Research [R package psych version 2.3.9].


Ridley, H. 1930. The dispersal of plants throughout the world. Reeve and Ashford, Kent, UK.


