­­Pan-genome of *Solanum habrochaites*, a wild tomato species

Christopher Neely - Boyce Thompson Institute

Advisers - Lei Gao, Zhangjun Fei

# Introduction and background

Tomatoes are one of the most important cultivated crops, with an annual production of 177 million tons in 2016, covering 4.7 million ha ([http://www.fao.org/faostat/en/#data/QC](http://www.fao.org/faostat/en/%23data/QC)). Due to its economic importance, the development of methods for increasing tomato production yields is at an all-time priority. Identifying novel alleles associated with important agronomical traits from wild tomato relatives and introducing these novel alleles to cultivated tomatoes provides a valuable approach for tomato improvement.

*Solanum habrochaites* is a diploid, wild tomato species. It grows on the slopes of the Andes Mountains, running from central Ecuador to central Peru, and has small, yellow flowers and bears clusters of green fruits. *S. habrochaites’* unique phenotype includes glandular trichomes on the fruit, and these trichomes have been shown to be related to sesquiterpenes and other chemicals that repel insects (Alba, Montserrat, & Fernández-Muñoz, 2009; Bleeker et al., 2011). When compared with the wild tomato, cultivated species have fruit that is a lot larger and that can vary in color. Because of these and other specifics, we are interested in better understanding the genomic differences between *S. habrochaites* and the cultivated tomato.

Using data sets from the recent sequencing of seven *S. habrochaites* accessions as a resource to mine for alleles and sources of genetic variation, we will construct a pan-genome for *S. habrochaites* (Aflitos et al., 2014). A pan-genome consists of core genes, which are shared by all accessions of the species, and variable genes, which are present in some, but not all, accessions. We will assemble and examine a pan-genome for *S. habrochaites* to better realize its full genomic potential.

# Explanation of proposed hypothesis

Our goal is to examine genetic differences between *S. habrochaite*s and *S. lycopersicum* Heinz 1706, the currently published reference genome for a domesticated tomato, by constructing a pan-genome for *S. habrochaites* using the Heinz 1706 genome as a reference. We will construct this pan-genome using the sequencing data of seven available accessions. We plan to catalog presence-absence variants (PAVs) between each of the accessions. We hypothesize that we will see genes absent in the Heinz 1706 reference that are present in *S. habrochaites*. Given the whitefly-repelling and potato psyllid-repelling abilities of *S. habrochaites*, as well as considering that *S. habrochaites*’ phenotype includes glandular trichomes, we also hypothesize that we will see defense-related genes and trichome-related or sesquiterpene synthesis-related genes in these seven accessions as well (Golicz et al., 2016; Gordon et al., 2017; Hurgobin et al., 2018; Levy & Tamborindeguy, 2014; McDowell et al., 2011).

# Experimental plan

Sequence data for the seven available accessions of *S. habrochaites* (LYC4, LA1777, LA0407, CGN157592, LA1718, TR00014, TR00015) will be downloaded from NCBI’s Sequence Read Archives (SRA) (<https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?>). These Illumina HiSeq 2000 raw reads will be quality-filtered with Trimmomatic v0.38 (Bolger, Lohse, & Usadel, 2014) to remove bases with Q < 25 over a sliding window of 10 bp, and then will be screened to remove reads less than 50 bp long. Sequence data per accession will be separately *de novo* assembled using SOAPdenovo2 (Luo et al., 2012). We will analyze quality metrics of the assemblies using Quast (Gurevich, Saveliev, Vyahhi, & Tesler, 2013).

We will compare results of each *de novo* assembly to the *S. lycopersicum* Heinz 1706 v3.0 reference genome using whole-genome alignment with NUCmer (Kurtz et al., 2004), and extract unaligned contigs from each individual assembly. We will then gather the unaligned contigs from all seven assemblies, removing redundant contigs in the process, to form our non-reference genome. These non-redundant contigs will be repeat masked using RepeatMasker (<http://www.repeatmasker.org>) and annotated using MAKER v2.31.10 (Holt & Yandell, 2011). MAKER involves using both *ab initio* and evidence-based gene prediction. For the former, we will use AUGUSTUS (<http://bioinf.uni-greifswald.de/augustus/>) and SNAP to make predictions (Leskovec & Sosic, 2016). For the latter, protein sequences from the *Solanales* order will be downloaded and mapped against the non-reference genome to make predictions. We will also use assemble RNA-seq evidence from *S. habrochaites* accessions using the Trinity assembler (Grabherr et al., 2011). These data sets will be used in the MAKER pipeline to generate high quality gene models. These models will be assessed with InterProScan (Finn et al., 2017), and high quality gene models will be functionally annotated with Blast2GO (<https://www.blast2go.com/>).

Finally, we will align quality-filtered reads for each accession to these annotated, non-redundant, unaligned contigs, and we will identify PAVs for these annotated genes using the SGSGeneLoss package (<http://appliedbioinformatics.com.au/index.php/SGSGeneLoss>). We will catalog genes that are accession-specific and genes that are shared among more than one accession.

# Justification

A pan-genome gives researchers an increased awareness of the possible variation inherent to a species. The creation of a pan-genome for *Solanum habrochaites* will provide a valuable resource for researchers looking to study diversity within tomatoes. It will also provide additional insight into diversity as a means of improving cultivated tomato crops. Understanding the nature of this variation is a benefit to crop breeders who wish to introgressively hybridize their crops for insect resistance, added durability, or increased longevity. As the need for higher crop yields increases, so does the importance of a more complete understanding of the available genomic potential afforded to cultivars from their wild species relatives.

# Anticipated Results

A pan-genome will be constructed for the wild tomato *Solanum habrochaites*, and we will characterize PAVs within the accessions. We expect to see unique genes related to defense response and insect resistance, likely related to sesquiterpene synthesis and other chemical pathways. This pan-genome will allow us to see which genes are shared among the wild accessions and which genes are unique to individual accessions. Our expectation is that we will have an improved data set that will more completely describe the variation within *S. habrochaites*.

# References

Aflitos, S., et al. (2014). Exploring genetic variation in the tomato ( Solanum section Lycopersicon ) clade by whole-genome sequencing. *The Plant Journal*, *80*(1), 136–148. https://doi.org/10.1111/tpj.12616

Alba, J. M., Montserrat, M., & Fernández-Muñoz, R. (2009). Resistance to the two-spotted spider mite (*Tetranychus urticae*) by acylsucroses of wild tomato (*Solanum pimpinellifolium*) trichomes studied in a recombinant inbred line population. *Experimental and Applied Acarology*, *47*(1), 35–47. https://doi.org/10.1007/s10493-008-9192-4

Bleeker, P. M., et al. (2011). RNA-seq discovery, functional characterization, and comparison of sesquiterpene synthases from *Solanum lycopersicum* and *Solanum habrochaites* trichomes. *Plant Molecular Biology*, *77*(4–5), 323–36. https://doi.org/10.1007/s11103-011-9813-x

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

Finn, R. D., et al. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research*, *45*(D1), D190–D199. https://doi.org/10.1093/nar/gkw1107

Golicz, A. A., et al. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, *7*, 13390. https://doi.org/10.1038/ncomms13390

Gordon, S. P., et al. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-02292-8

Grabherr, M. G., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 491. https://doi.org/10.1186/1471-2105-12-491

Hurgobin, B., et al. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. *Plant Biotechnology Journal*. https://doi.org/10.1111/pbi.12867

Kurtz, S., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2), R12. https://doi.org/10.1186/gb-2004-5-2-r12

Leskovec, J., & Sosic, R. (2016). SNAP: A general purpose network analysis and graph mining library. Retrieved from http://arxiv.org/abs/1606.07550

Levy, J., & Tamborindeguy, C. (2014). *Solanum habrochaites*, a Potential Source of Resistance Against *Bactericera cockerelli* (Hemiptera: Triozidae) and "*Candidatus* Liberibacter solanacearum”. *Journal of Economic Entomology*, *107*(3), 1187–1193. https://doi.org/10.1603/EC13295

Luo, R., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(1), 18. https://doi.org/10.1186/2047-217X-1-18

McDowell, et al. (2011). Comparative functional genomic analysis of Solanum glandular trichome types. *Plant Physiology*, *155*(1), 524–39. https://doi.org/10.1104/pp.110.167114