Project Narrative

I. Training and Career Development Plan

After completing my PhD, I will pursue a post-doctoral position which will help me to achieve my goal of tenure-track professorship, working at the intersection of population and crop genomics. To prepare for this, I will continue to participate in i) research training activities to further develop my quantitative skills; ii) teaching and mentorship training to grow as an advisor; and iii) professional service activities. These activities will help hone my skills as a scientist and mentor, and provide models for creating a welcoming, diverse, and scientifically rigorous lab environment.

i) Research Training—My strength as a researcher is in applying cutting-edge genomic and computational approaches to long-standing classical genetic and evolutionary questions. In addition to my PhD major, at Indiana University (IU) I have completed a minor in bioinformatics, including applied classes on genomic data analysis, a theoretical class on statistical modeling, and two machine learning courses. I also attended the Summer Institute in Statistical Genetics (Univ. of Washington) on a competitive scholarship where I completed training modules on population and quantitative genetics. In the future, I plan to attend similar workshops, with a particular focus on machine learning and its application to genomics (e.g., the Harvard NSF Machine Learning and Biology Workshop).

ii) Teaching and mentorship—At IU I have served as an assistant instructor for six semesters, including in upper-level evolution (98 students), introductory biology lab (25 students), and diversity, evolution, and ecology (200 students). In addition, I am particularly passionate about teaching practical computational methods, and have assisted in the development and teaching of two bioinformatics workshops: The *R Programming Bootcamp* at the IU School of Public Health and the *Command Line Bioinformatics Workshop* hosted by the American Society of Naturalists. I have also served as a STEM career mentor to underrepresented students through IU's *Groups Scholars STEM Program*, and in the laboratory where I have mentored five IU undergraduates and two high school students. I plan to continue incorporating undergraduate and high school students in my research throughout my career. These activities have helped me grow as a teacher and mentor and will serve as a strong foundation as I begin my own independent group.

iii) Professional service—At IU, I have taken on several departmental leadership roles, including as the graduate student representative on two tenure-track hiring committees and, in the coming year, coordinating the IU Biology *Evolution Discussion Group*—a graduate student-led seminar series. Professionally, I have been a manuscript peer reviewer, and will continue this service to the academic community in the future, in addition to, for example, organizing conference symposia and early-career networking events.

II. Mentoring Plan

I will be mentored by Dr. Leonie Moyle, an expert in evolutionary genetics and the wild tomato clade, with a 15-year record of mentoring graduate research. We have an established, successful mentoring plan in which we meet weekly--to discuss ongoing and developing research, including conceptual and experimental design, manuscript preparation, teaching, and undergraduate mentoring--and biannually to complete and discuss a structured Individual Development Plan, which addresses progress and future plans for my academic and professional development. I will also continue to receive feedback from my lab group during our weekly meetings, and at the IU *Evolution Discussion Group*. In addition, I will hold monthly meetings with our Ecuadorian collaborator Dr. Maria de Lourdes Torres, and both Drs. Moyle and Torres will continue to connect me with their network of collaborators in the US and Ecuador, to further my academic development. ***Dr. Leonie Moyle's previous mentees:*** Meng Wu *(Research Scientist, Editas Medicine)*; Jamie Kostyun *(Post-doc, U. Mass)*; Jenna Hamlin *(Research scientist, CDC)*; Dean

Castillo *(Asst Prof, U. Utah)*; CJ Jewell *(Data scientist, AbbVie)*; Chris Muir *(Asst Prof, U. Hawaii Manoa)*; David Haak *(Asst Prof, Virginia Tech)*; Yaniv Brandvain *(Assoc. Prof; U. Minnesota)*.

III. PROJECT PLAN

Introduction—Plants adapt to their environments via evolutionary response mediated by genetic variation. Knowledge of the genes allowing for persistence in one environment over another allows us to test powerful hypotheses regarding the origins of biodiversity, as well as assist in the breeding of resilient crop species. Because they are sessilem wild plants must cope with harsh environmental conditions in place and thus are often subject to strong natural selection favoring traits suited to their local environment—a process termed *local adaptation*. Nearly a century of experiments has revealed that local adaptation is the rule rather than the exception (Turesson, 1922; Clausen et al., 1948). Nonetheless, our understanding of this process at the genetic level remains limited, despite the potential gains for modern crop breeding.

Wild germplasm and crop wild relatives have been a rich resource of beneficial crop traits and have historically served as a powerful counterpart to artificial selection. However, many traits relevant to adaptation are considered to be highly complex, making the identification of causal loci in germplasm resources exceptionally difficult. To overcome this major barrier to modern crop breeding, and to maintain year-to-year increases in crop yields, more than a single approach is needed. Methods derived from population and ecological genomics represent an important yet underdeveloped complement to traditional methods of exploiting wild germplasm. These methods leverage existing adaptive variation in wild populations to identify genes subject to selection. To this end, my overarching goal for this project is to **use population genomics to identify and verify adaptive alleles for the breeding of complex traits in an important crop species.**

The wild tomato clade, native to the Andean region of South America, is an ideal system for studying the genetic architecture of local adaptation. Abundant climatic variation in the region—ranging from cool, wet highland environments to dry, salty coastal environments—has been shown to be an important driver of local adaptation (Blanca et al., 2012, 2015; Gibson & Moyle, 2020; Nakazato et al., 2008; Zuriaga et al., 2009), including in traits valuable for breeding. In particular, the wild species *S. pimpinellifolium (S. pim.)*—the closest relative of domesticated tomato—has been the source for many agricultural traits (e.g., fruit quality and disease resistance; Capel et al., 2015; Peirce, 1971). However, complex phenotypes such as drought and salinity tolerance—which naturally vary among populations of *S. pim* (Nakazato et al., 2008)—have not been well characterized, and are largely undeveloped in domesticated lines. In the context of rapidly increasing global temperatures, sustained droughts are becoming more frequent across the United States. With this in mind, my project has two major objectives aimed at identifying and functionally assessing genes underlying abiotic stress tolerance in tomato, with the end-goal of improving US tomato crop resilience:

> **Objective 1:** *Identify candidate loci under natural selection in S. pim along three replicate coastal-inland environmental gradients using whole-genome sequencing*
>
> **Objective 2:** *Evaluate the functional effect of candidate loci by phenotyping stress tolerance traits in a targeted germplasm panel*

Rationale and Significance—This project will complement more traditional approaches to identifying abiotic stress tolerance genes by leveraging population genomic tools designed for studying the association between genome-wide variants and environmental gradients in natural populations. At their core, these methods work by identifying associations between allele frequency (at each gene in the entire genome) and environmental variables. I propose to use a powerful replicated sampling scheme along three parallel inland-coastal gradients to identify

sequence variants with strong, consistent environmental associations across all three clines. Signals shared across multiple gradients indicate variants more likely to be beneficial across multiple genetic backgrounds and environmental contexts— i.e., globally-beneficial gene variants--a necessary quality if used for plant breeding. This replicated design also circumvents the sensitivity of these methods to confounding factors, such as historical population genetic structure, which can otherwise lead to false-positive associations (Price et al., 2019). Once I identify a strong set of candidate genes, I will then test them for functional relevance using genomic prediction and directed phenotyping experiments, allowing for identification of potential genic targets for future transgenic modification.

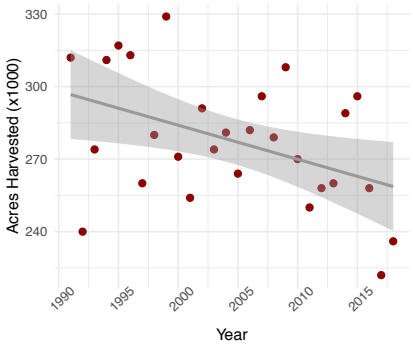

Figure 1: California tomato harvest (in thousands of acres) from 1990-2018. Data from USDA (2018).

The proposed project directly addresses core goals in the "Plant Health and Production and Plant Products" AFRI Farm Bill Priority Area: to understand how plants grow, how to improve productivity, and how to use them in new ways. As the second most consumed vegetable in the United States (USDA), the tomato crop is central to national food production and food security; in 2017, US processed tomato consumption was 73.3 pounds per capita. While tomato yields per hectare have steadily increased over the last three decades due to modernized farming and breeding practices (USDA), this positive yield trajectory is threatened by accelerating climate change, including rising temperatures and increased drought severity. California, the largest global producer of processing tomatoes (USDA), has experienced recent and sharp harvest declines, mostly attributable to unpredictable drought and temperature conditions (**Figure 1**), and sensitivity to abiotic stress remains a major limiting factor for tomato production in the US, as well as globally. New approaches to crop breeding which leverage natural diversity to improve productivity are critical for maintaining yield under these changing climates.

Approach

Objective 1—Identifying candidate loci for adaptation to abiotic climates. To identify loci underlying adaptation to divergent abiotic climates, I will perform whole genome sequencing of 15 populations of *S. pimpinellifolium* arrayed along 3 replicate coastal-inland environmental gradients in Ecuador (**Figure 2**). Eight populations have been previously collected in 2019 (but not yet sequenced) and I will collect a minimum of six additional populations. Field expeditions will be conducted in conjunction with our collaborator Dr. María de Lourdes Torres (Universidad San Francisco de Quito [USFQ], Ecuador) with whom I have completed three previous field trips (*see attached letter from Dr. Torres*). DNA will be extracted at USFQ and genomic libraries for 10 individuals/population (150 total) will be created by the Indiana University Center for Genomics and Bioinformatics and sequenced to an average depth of 6X (based on 950 Mb genome size) using Illumina NextSeq technology. Reads will be quality checked using standard procedures previously implemented by PD Matthew Gibson (Gibson & Moyle, 2020), and mapped to the domesticated tomato reference genome SL4.0 (Tomato Genome Consortium).

Given that moderately low sequencing depths are expected, errors in genotype calling could inflate rates of false-positive associations. To account for error inflation, I will apply two methods to detect SNP-environment associations: (i) a standard model-based approach using latent factor mixed models (*LFMM*; Frichot et al., 2013) which will rely on genotype calls and (ii) a likelihood-based approach using the software *ANGSD* (analysis of next generation sequencing

data; Korneliussen et al., 2014) which accounts for the elevated uncertainty in genotype assignment at low depth. I will control for historical population structure in both analyses, using sparse nonnegative matrix factorization (*sNMF*; Frichot et al., 2014) to define latent factors in *LFMM* and using the first three multi-locus principal component axes as covariates in *ANGSD*.

I will run genome-wide scans for each cline against a set of six abiotic environmental variables which I have previously found to contribute to genomic divergence across Ecuador (Gibson & Moyle, 2020) and will retain only signals detected by both *LFMM* and *ANGSD* after false-discovery rate correction. Focal environmental variables include vapor pressure, evapotranspiration, soil texture, and precipitation seasonality, among others. SNPs passing correction will represent our top candidates for adaptive loci in each group. These lists will then be compared between groups to identify shared signals of adaptation using the following procedure. For each outlier SNP present in two or more clines, I will fit linear models of the relationship between allele frequency and longitude. Because I am interested in alleles which define coastal vs. inland habitats, I will remove alleles with slope 0 (as determined by a t-test) and those with slopes of differing signs in multiple clines (as determined by an analysis of covariance). The final set of SNPs will be strong candidates for parallel natural selection along the coastal-inland gradient.

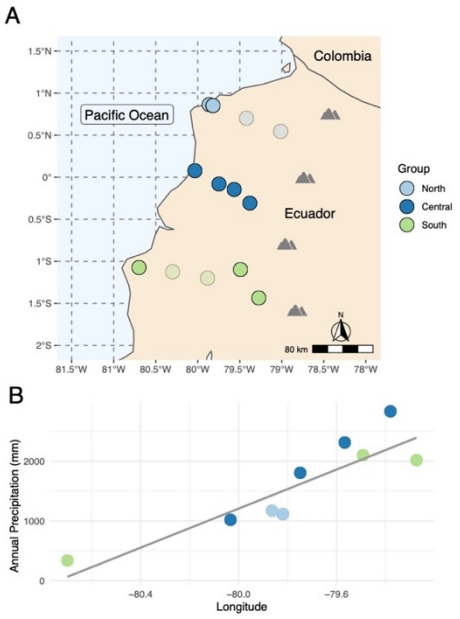

Figure 2: (A) Sampling localities in Ecuador. Shaded points indicate populations previously sampled, transparent points indicate proposed future collection sites. (B) Annual precipitation averages for all collections sites vs. longitude.

Outcomes: My previous studies of adaptation in this system indicated that a substantial proportion of genotypic variation among *S. pim* populations can be explained by variation in abiotic climate (Gibson & Moyle, 2020). While non-selective processes (i.e., drift in isolated subpopulations) were also shown to impact the partitioning of genetic diversity across space, their impact was small relative to climate. Because of this, I expect to detect substantial genetic signals of adaptation between inland and coastal sites which I know to be highly diverged for several abiotic variables. Given that adaptations to such environmental factors are often considered exceptionally complex, I expect to detect hundreds of associated SNPs per cline, each of relatively small effect, rather than a handful of highly explanatory loci. I also expect that most signals of adaptation will be idiosyncratic and unique to individual clines. This is because, in addition to environmental factors which vary between the coast and inland, other factors vary by latitude which may modulate the response to selection. Similarly, other patterns of local adaptation (e.g., to unmeasured biotic factors) are likely also at play. Although I expect a minority of signals to be shared between clines, shared signals will be highly likely to have functional effects on phenotypic variation relevant to coastal-inland adaptation.

Caveats and Considerations: The continued spread of COVID-19 represents a potential barrier to making further field collections in Ecuador. In the event that travel is not safe by the proposed travel date of Spring 2021, one of two possible scenarios will be enacted: 1) if travel within Ecuador is safe but international travel prohibited, field collections will be undertaken solely by our Ecuadorian collaborator Dr. María de Lourdes Torres and her group who have substantial

experience identifying and collecting wild tomato or 2) if travel within Ecuador remains unsafe, all subsequent sequencing and analysis will proceed with the existing 9 populations (90 samples) sampled in January 2020. While this would lead to reductions in the robustness of my study design, the spatial extent of population sampling would still be greater than most published population genomic studies. My advisor—Dr. Leonie Moyle—and I have deemed my existing collections sufficient for all downstream objectives if existing university travel bans remain in place.

Objective 2—Evaluating the functional significance of key candidate loci using a genomics-informed germplasm assessment. In order to test whether candidate loci from objective 1 have direct effects on survival under abiotic stress (drought and heat), and to identify phenotypic variation associated with plant fitness, I will construct predictive models and measure trait variation and survival in a set of additional Ecuadorian accessions. (*Due to exporting restrictions imposed by the Ecuadorian Department of Environment, I am unable to directly phenotype the offspring of sampled plants*). For each cline, 8 Ecuadorian *S. pim* accessions from the TGRC (4 coastal and 4 inland)—originating from areas within 5km of my objective 1 sampling sites—will be sequenced concurrently with samples collected in objective 1 (150 novel population samples + 24 accessions = 174 total). Environment-associated SNPs and their effect sizes from objective 1 will be used as training data to fit polygenic models for each TGRC accession. For each focal environmental variable, I will calculate a stress tolerance score (STS) representing my projection of the tolerance of each accession under experimental stress based on its genotypes at outlier loci. For example, accessions with alleles associated with low levels of precipitation (i.e., alleles at high frequency in coastal sites) will be predicted to be more tolerant to drought.

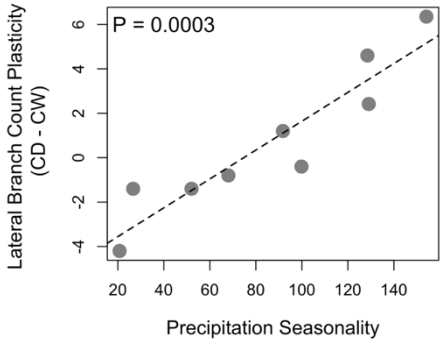

Figure 3: Pilot root experiments reveal significant variation in root branching responses to drought among accessions that is strongly associated with precip. seasonality.

To test these predictions, I will perform a two-phase experiment of accession tolerance/survivability under simulated heat and water stress in both seedling (*phase I*) and juvenile (*phase II*) plants. Phase I will consist of a full-factorial manipulation of water and heat (four treatments: cool/wet, cool/dry, hot/wet, hot/dry) in seedlings, using custom rhizome boxes to phenotype root development. The total experiment (3 reps/accession × 4 treatments × 24 accessions = 288 plants) can be performed using two identical Percival growth chambers, in less than four weeks. Root traits (incl. rooting depth and branching architecture) previously identified as relevant to wild tomato abiotic tolerance (Nakazato et al., 2008; Rick, 1973)—including divergence between coastal and inland sites—will be measured alongside survival. Survival will be measured via a resurrection assay (returning plants to control setting post-treatment) and root traits will be assessed using a flatbed scanner. Experimental feasibility of this design was confirmed in a pilot experiment which revealed abundant genotypic variation among accessions in root branching responses to drought stress that was associated with key abiotic environmental variation (precip. seasonality), indicating that such traits are heritable, differ among accessions, and may underlie key adaptive strategies (**Figure 3**).

In phase II I will use a replicated, fully randomized experimental design to assess juvenile drought tolerance in an outdoor rainout shelter at Indiana University (4 reps × 2 treatments × 24 accessions = 192 plants). At 30 days post germination (DPG), plants will be transferred to the

shelter where they will be well-watered for 10 days prior to treatment. At 40 DPG, water will be restricted to drought treatment plants. Plants will be monitored twice daily and days to wilting will be determined. I will continue drought treatment until 55 DPG. Water will then be returned and plants will be scored for survival (dead, resurrected, or alive).

To assess the predictive power of associated SNPs on genotypic stress tolerance, I will use linear models to test for significant associations between each accession's STS and its measured tolerance to drought, temperature, and their combination. Stress tolerance for phase I will be defined as the propensity to maintain or exceed control treatment growth under stress. For phase II, tolerance will be defined as days to wilting and as proportion survived under drought relative to proportion survived under the control treatment, where the 'resurrected' case will be assigned a value of 1/2. Strong associations between predictions and common garden measurements will indicate that the environment-associated SNPs used to fit the model have physiologically relevant effects on tolerance to abiotic stress.

Outcomes: Based on my own preliminary data and previous studies, I expect to find abundant genotypic variation for survival under abiotic stress and in its associated phenotypes. Because these traits vary by accession, are heritable, and have previously been shown to vary between the coast and inland, I expect to detect correlations between predicted STSs and observed phenotypes indicative of adaptation.

Timeline

Activity	Fall 2020	Spring 2021	Summer 2021	Fall 2021	Spring 2022	Summer 2022
Collect samples *(Obj.1)*		X				
Library prep. *(Obj.1)*			X			
Sequencing *(Obj.1)*			X	X		
Scan for outlier SNPs *(Obj.1)*			X	X		
Build polygenic models *(Obj.2)*				X		
Seedling tolerance assay *(Obj.2)*	X	X	X			
Juvenile tolerance assay *(Obj.2)*		X	X			
Data analysis *(Obj.1&2)*				X	X	
Publish *(Obj.1&2)*					X	X
Attend SSE conferences			X			X

IV. Evaluation Plan

Evaluating progress—I will track my progress by committing to the above timeline for gathering samples, generating sequence data, and performing experiments, with the one potential exception being field work which may be delayed due to COVID-19. My progress will also be evaluated directly by my advisor during weekly one-on-one meetings via formal project update documents and during our weekly lab meetings where I will discuss my experiments, share my results, and solicit feedback from other lab members.

Dissemination plan—The results of my USDA-supported work will be published in high-impact, peer-reviewed journals. I expect that two original research manuscripts will come from this project, data from which will also serve as a basis for future work in the system. I will present my findings at international conferences such as the SSE Evolution or ASB Botany conferences in 2021 and 2022 and Gordon Research Conferences. Additionally, project results will be shared in departmental discussion groups, such as the IU biology Evolution Discussion Group.