PROJECT DESCRIPTION

Improving Plant Breeding in Polyploid Crops Using Models for Demography and Selection

1 BACKGROUND

The process of domestication leaves distinctive signatures in the genomes of cultivated plants. The initial selection on traits of interest and subsequent interbreeding of suitable progeny simultaneously imprint a loss of genetic diversity from the original bottleneck as well as selection on the genes underlying the desired phenotype [Doebley et al. 2006; Meyer and Purugganan 2013]. This "domestication syndrome" is a conundrum for identifying domestication related genes owing to the confounding effects of historical demography and selection [Hammer 1984; Nielsen et al. 2009; Wang et al. 2017]. Previous studies aiming to identify genes under selection have employed genome-wide scans for statistical outliers, but other investigations have shown that these approaches are prone to false positives when demography, mutation, and recombination are not considered [Reich et al. 2002; Drake et al. 2005; Haasl and Payseur 2016]. Understanding the process of domestication is imperative for combatting these issues, and can help plant breeders by more accurately detect candidate genes for crop improvement.

Identifying genes and traits tied to domestication can benefit greatly from knowledge about the underlying demographic history. Inferring the demographic history of a population typically uses genomic data from multiple individuals to estimate changes in effective population size over time as well as the timing of population divergence, gene flow, and other historical events [Gutenkunst et al. 2009]. The diffusion approximation can be used to model the effects of this history on the site frequency spectrum (SFS), which summarizes the distribution of allele frequencies for genome-wide biallelic SNP data. The approximation relies on the mean or expected allele frequency trajectory and the variance around that expectation. Directional evolutionary forces such as mutation, selection, and migration are modeled by the expected allele frequency trajectory and the stochastic force of genetic drift is modeled by the variance [Kimura 1964]. This framework is especially appealing given the ease of collecting genomic SNP data using modern sequencing technologies such as restriction-site associated DNA sequencing (RADseq) [Miller et al. 2007; Baird et al. 2008].

A confounding factor for understanding the demographic history of domesticated plant species is that they are frequently polyploids [Salman-Minkov et al. 2016]. Methods for demographic inference and the detection of selection are often only implemented with diploids in mind, precluding the appropriate analysis of many of the world's most important crop species (e.g., maize, bread wheat, sugarcane, apple, strawberry, canola oil, etc.) [Hilu 1993]. Extending these methods to polyploids and implementing them in an open source software package would break down many barriers to crop improvement, and would facilitate inferences of demography and selection in polyploid organisms across the Tree of Life.

We propose to develop an integrative framework that uses demographic information for the inference of candidate domestication genes in polyploid crop species and to make it available to plant breeders through an easy-to-use software package. We will then apply these methods to the allotetraploid oilseed crop *Brassica napus* L. (canola oil). *Brassica napus* (2n=38, AACC) is a member of the *Brassica* "Triangle of U," formed by at least one hybridization event between its diploid parents: *B. rapa* (2n=20, AA) and *B. oleracea* (2n=18, CC) [Nagaharu 1935]. *Brassica napus* is a young species formed within the last few thousand years and has only been domesticated for about 400 years [Chalhoub et al. 2014]. It is second only to soybeans in terms

of agricultural production for consumption by humans and other animals, and is cultivated for not only vegetable and synthetic oils, but also for its roots (rutabaga) and leaves (Siberian kale) [Gazave et al. 2016]. The primary agronomic traits for *B. napus* focus on seed contents, and include oil, protein, acid (erucic, linolenic, stearic), and glucosinolate levels [Zou et al. 2016]. Other traits such as seed fibre [Liu et al. 2013], the apetalous character (increases seed production) [Wang et al. 2015], and fatty acid profile are targeted as well [Raboanatahiry et al. 2017]. Given the continued interest in cultivating diverse varieties of *B. napus*, our work will provide important insights on the demographic history of the species and will identify candidate domestication genes that can be used to guide future breeding efforts.

2 RESEARCH AIMS AND METHODS

The primary objective of this NSF-PRFB proposal is to develop a modeling framework for understanding the demographic history of polyploid crop species and to use this information to robustly identify genes under selection. Our methods and software are potentially transformative as they will be generally applicable to breeding efforts in any polyploid crop species. Our main system for applying these methods will be the allotetraploid oilseed crop *Brassica napus* (source of canola oil).

Our major research aims are to:

1. Develop methods and software to model the demography of polyploid crop species.
2. Model the demographic history of *Brassica napus* and its diploid parents.
3. Identify genes under selection in *B. napus* using whole-genome simulations based on the inferred demographic history of domestication within the species.

2.1 Aim 1: Develop Methods to Model Demographic History in Polyploids

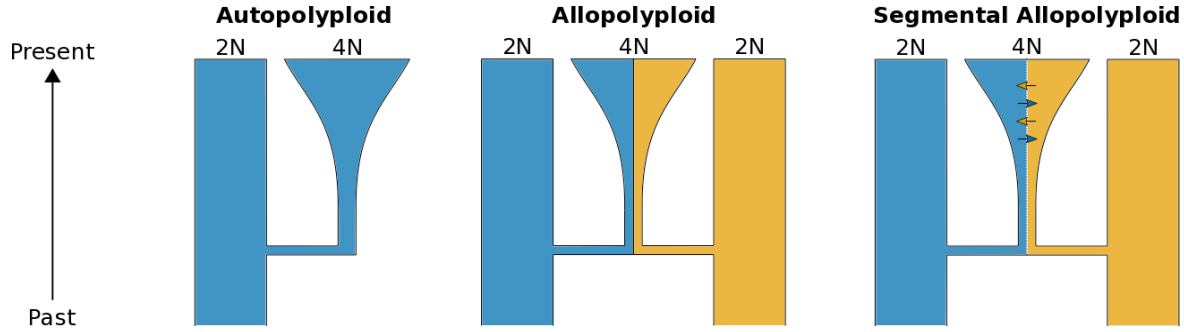

Figure 1. Demographic models for polyploid plant domestication. From left to right they represent autopolyploidy, allopolyploidy, and segmental allopolyploidy. Details for each model are in the text.

A powerful approach for inferring demographic history is to maximize a composite likelihood function that compares the SFS from a demographic model to the SFS obtained from observed SNP data [Sawyer and Hartl 1992]. A popular diffusion-based implementation of this approach is the Python package ∂a∂i, which uses a multidimensional SFS to model the demographic history of up to three populations simultaneously [Gutenkunst et al., 2009]. To extend this framework to polyploids, we propose several modifications (**§2.1.1–§2.1.4**; ordered by potential difficulty) to the standard diffusion model for diploids (depicted in **Figure 1**).

2.1.1 Autopolyploids: Autopolyploids are characterized by whole genome duplication (WGD) within a lineage such that all of the duplicated chromosomes have the same genetic ancestry. Assuming that homologous chromosomes are equally likely to recombine with one another (tetrasomic inheritance), the result will be a doubling of the effective population size. This change in effective population size affects the rate at which variation is lost in the population due to genetic drift. The autopolyploid model we propose includes a divergence event from its diploid ancestor with an initial bottleneck followed by expansion. Modifying the diffusion equation in this case only requires a simple rescaling of the effective population size.

2.1.2 Allopolyploids: Allopolyploids differ from autopolyploids in that they are formed through hybridization such that their duplicated chromosomes have different genetic ancestries (homoeologs). If homoeologous chromosomes do not recombine, we can model the SNP frequencies within the two subgenomes independently, but constrained to have the same demographic history. The allopolyploid model we propose includes the time of the hybridization event between the two diploid parents and a shared bottleneck with subsequent expansion for the two subgenomes. This can be modeled without directly modifying the diffusion approximation by coupling separate two-population models with a shared history for the allopolyploid.

2.1.3 Segmental Allopolyploids: It is often the case that the two parental species that form an allopolyploid are closely related, allowing for recombination to occur between homoeologous chromosomes. This type of polyploid, called a segmental allopolyploid [Stebbins 1951], has chromosomal inheritance and recombination patterns that are intermediate between autopolyploidy and allopolyploidy. To model this scenario, we propose to parameterize homoeologous recombination as migration between the subgenomes of the allopolyploid. In this case, we have a four population model that will require different numerical techniques to solve for the SFS (see note on *Moments* in **§2.1.6**).

2.1.4 Domestication Bottlenecks: The theoretical basis of the diffusion approximation is that the effective population size is large enough to treat changes in allele frequencies from generation to generation as infinitesimally small [Kimura 1964]. Because of this, population bottlenecks are difficult to model since they involve decreases in population size that can generate rapid drift. To address this issue, we will develop novel numerical and mathematical solutions to better estimate bottlenecks. In the simplest case, we will use forward simulations of the Wright-Fisher process during the bottleneck to accurately model changes in allele frequencies that the diffusion approximation cannot capture. We will also derive new mathematical formulas for the diffusion approximation under a scenario of small population size.

2.1.5 Adding Complexity: The scenarios that we have described above do not completely capture the complexity of WGD. However, we consider these demographic models to be suitable starting points upon which we will build greater complexity in the future. Future modeling efforts will include: (1) modeling higher ploidy levels, (2) incorporating gene loss and fractionation, and (3) using variable chromosome inheritance patterns during meiosis.

2.1.6 Software Development: We will implement all proposed extensions by modifying the existing code base in the ∂a∂i Python package. ∂a∂i is freely available and version-controlled on Bitbucket. All additions to the software will continue to be open source. We will also explore the possibility of implementing our models using the numerical machinery in the Python package *Moments*, an offshoot of ∂a∂i that can model the demographic history of up to five populations [Jouganous et al. 2017].

2.2 Aim 2: Model the Demographic History of *Brassica napus*

2.2.1 Data: We will use previously sequenced genome-wide SNP data sets collected for *B*. *napus*, *B*. *rapa*, and *B*. *oleracea* that are available through a collaboration between the Barker lab and Dr. Chris Pires at the University of Missouri. These data will be used to construct a multidimensional SFS to use as input for the polyploid models that we will implement in the new version of ∂a∂i that are described above.

2.2.2 Demographic Modeling: Previous work has shown that the homoeologous chromosomes in *Brassica napus* undergo recombination, making the segmental allopolyploid model the most appropriate demographic history to use [Xiong et al. 2011; Cai et al. 2014; Chalhoub et al. 2014]. To confirm this, we will also fit the strict allopolyploid model to the data and will use the Godambe Information Matrix (GIM) to perform model selection [Coffman et al. 2015]. For the segmental allopolyploid history, we will test models that include symmetric and asymmetric rates of homoeologous recombination (subgenome migration) between the subgenomes of *B*. *napus* to try to quantify differences in recombination between the A and C genomes. The fit of all models used to infer the demographic history of *B*. *napus* will be assessed using the GIM, as well as parametric bootstrapping [Gutenkunst et al. 2009; Coffman et al. 2015].

2.3 Aim 3: Identify Candidate Genes for Breeding in *Brassica napus*

2.3.1 Genome Scans for Selection: Genome scans for selection use sliding window-based approaches to identify regions of the genome that are statistical outliers in the genome-wide distribution of a given test statistic. However, these approaches are prone to false positives because they do not control for demographic history or heterogeneity in mutation and recombination rates [Reich et al. 2002; Drake et al. 2005; Schaffner et al. 2005; Sainudiin et al. 2007]. Below we describe a procedure to more robustly identify candidate genes under selection within each window, but we will still conduct genome scans for comparison to quantify the proportion of windows that are incorrectly identified as being under selection (i.e., regions that are outliers in the genome scans but do not pass our test described in **§2.3.2**) [Hsieh et al. 2016]. We will use sliding windows of 500 SNPs to calculate the G2D and integrated haplotype score (iHS) statistics and will consider regions in the 0.5% tails of the genome-wide distribution as outliers [Voight et al. 2006; Nielsen et al. 2009].

2.3.2 Simulation-Based Assessment of Statistical Outliers: Following Hsieh et al. (2016), we will perform whole-genome simulations that control for demographic history as well as variation in mutation and recombination rates using the program `msprime` [Kelleher et al. 2016]. Using 1000 samples of parameters from the confidence intervals of the estimated demographic history of *B*. *napus*, we will include previously developed recombination maps and estimates of local mutation rates to simulate whole chromosomes matching those from the A and C genomes in *B*. *napus* [Liu et al. 2014; Wang et al. 2015]. We will then take the same windows that were used in the genome scans and will extract the corresponding region from the simulated chromosomes. This will allow us to obtain a P-value for each window using the distribution of 1000 simulation replicates by calculating the proportion of the simulated distributions of the test statistics (G2D and iHS) that are greater than or equal to the observed values from the data. Among the list of significant regions, we hypothesize that we will find genes involved in the pathways for oil and fatty acid production [Liu et al. 2012], as well as genes controlling flowering time [Wang et al. 2011; Schiessl et al. 2014], seed protein content [Chao et al. 2017], and seed glucosinolate levels [Qu et al. 2015].

2.4 Project Significance

The methods and software we propose to develop are potentially transformative as they will be generally applicable for identifying candidate domestication genes while controlling for demography in any polyploid crop species. Past research on population genomics and demography has focused on diploids, leaving a substantial gap for the development of a modeling framework for polyploid species, which includes many economically important crops. We focus our application of these methods on *Brassica napus*, a significant but understudied agricultural system that provides global food resources, industrial grade oils, and biofuels [Gazave et al. 2016]. Gaining an understanding of domestication in *B. napus* will facilitate breeding efforts by allowing for the targeted introduction of genetic diversity into regions controlling traits of interest using crosses to other A and C genome carrying species in the Triangle of U [Qian et al., 2006; Rahman, 2013; Zou et al. 2017]. Furthermore, the integrative framework for modeling demography and selection that we present here can easily be employed by breeders using our software in both polyploids and diploids (or a combination), making it an especially appealing approach for applications in agricultural improvement.

3 TRAINING OBJECTIVES AND PLAN

As a researcher with a quantitative background, the training objectives for the PI are (1) to formally move his research focus into agricultural systems, (2) to gain an understanding of plant breeding practices and how methods in evolutionary genetics are applied in agriculture, and (3) to enhance his computational and mathematical skills through the development of new tools for research in plant sciences. The PI will work closely with both the Barker and Gutenkunst labs to develop, implement, and apply the new methods described in this Fellowship to *B. napus*. Mentoring from the co-sponsors will be complementary, with Dr. Gutenkunst advising the PI on working with the diffusion approximation, demographic models, and software development, and Dr. Barker advising him on the biology, data analysis, and plant breeding applications in *B. napus*. An essential part of the PI's training will also be to network and establish connections with researchers that are part of the National Plant Genome Initiative. These collaborations will be invaluable in the latter stages of the project as we disseminate and test our methods in other polyploid crop systems.

4 CAREER DEVELOPMENT, FUTURE RESEARCH, AND PROJECT DISTINCTNESS

Since earning my Bachelor's degree in mathematics, I have always been fascinated by the application of computational and statistical methods in biology. While working on my PhD, the focus of my dissertation has been on developing bioinformatic tools to analyze high throughput sequencing data for SNP genotyping and haplotype inference for applications in phylogenetic systematics in non-model plant groups. I am currently using these tools to model hybridization, and to infer phylogenetic networks, in the genus *Penstemon* (Plantaginaceae). **The research proposed here will give me the opportunity to move my research into agricultural systems, allowing me to develop computational tools that can be used to enable global crop improvement.** The mentoring, training, and collaboration provided through this Fellowship will also help me to become part of the agricultural research community. In the future, I will continue to foster these goals and collaborations as a faculty member with a research program in computational evolutionary genetics focused on understanding the process of domestication through the development and integration of new methods for analyzing genomic and phenotypic data.

5 PROJECT SPONSORS AND HOST INSTITUTION

Dr. Michael Barker and Dr. Ryan Gutenkunst are renowned experts in plant and human genetics, respectively, with applied and computational research foci on the analysis of genetic data in many areas of evolutionary biology. Dr. Barker has played a vital role in studying whole genome duplication at multiple timescales in *Brassica*, as well as in developing bioinformatic tools to study plant evolution. Likewise, Dr. Gutenkunst has been instrumental in the use of the diffusion approximation through his development of ∂a∂i and its application to study demography and selection in humans. They have collaborated on demographic modeling in *B*. *rapa* [Qi et al., 2017] and are an ideal team to accomplish the research objectives described here. I have visited them at The University of Arizona (UA) and they have been actively advising me on developing the ideas in this Fellowship. Furthermore, UA has a history of excellence in postdoctoral training, with access to computational resources, diverse faculty, and interdisciplinary research infrastructure (BIO5 Institute), making it the perfect institution for this NSF-PRFB.

6 BROADER IMPACTS

Interdisciplinary training in biology, statistics, and informatics is paramount for gaining a deeper understanding of the factors affecting the success of modern agriculture. By developing an integrative framework for the analysis of demography and selection in polyploids and implementing it in freely available, open source software, we are providing an essential tool to the agricultural community. The training provided by this Fellowship will also equip me with the expertise necessary to build tools that will help solve global issues in agriculture and will prepare me to train future students to do the same.

Throughout the Fellowship, I will engage in undergraduate and graduate student mentoring and will focus in particular on teaching skills in statistics and bioinformatics. In addition, I will develop modular teaching materials aimed at instructing students studying plant breeding and quantitative genetics. Specifically, I will develop lesson plans on (1) historical demography and domestication and (2) identifying candidate genes under selection with genome-wide SNP data. The effectiveness of the lesson plans will be assessed through guest lectures in classes taught by Dr. Barker and Dr. Gutenkunst with pre- and post-surveys given to students to determine comprehension. All teaching materials and code will be open source and version controlled on GitHub, so that other instructors can easily access, update, and reuse the content as they see fit.

7 TIMELINE

Project Objective	**Year 1**	**Year 2**	**Year 3**
Develop diffusion approximation for polyploid models.	X		
Improve bottleneck estimation and add new features to ∂a∂i.	X		
Infer demographic history for *B*. *napus*, *B*. *rapa*, and *B*. *oleracea*.	X	X	
Use whole-genome simulations to ID candidate genes in *B*. *napus*.		X	X
Translate research into plant breeding solutions for *B*. *napus*.		X	X
Develop and distribute teaching materials.		X	X
Communicate and distribute work through publications and talks.		X	X
Develop research program and apply for faculty positions.		X	X