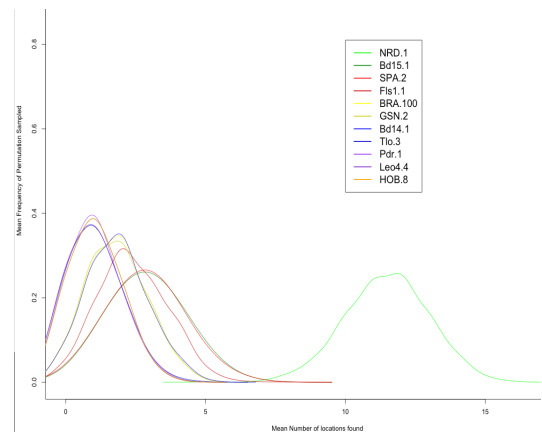
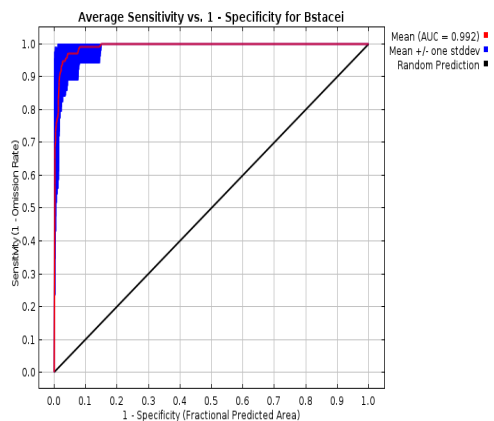
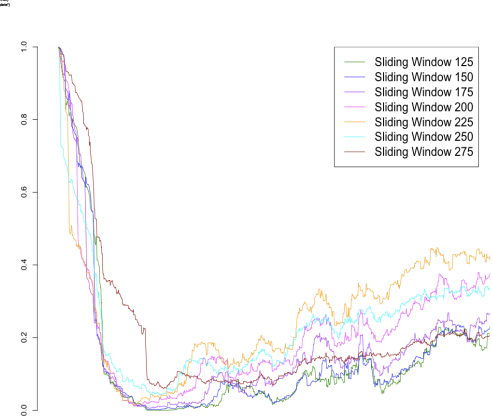
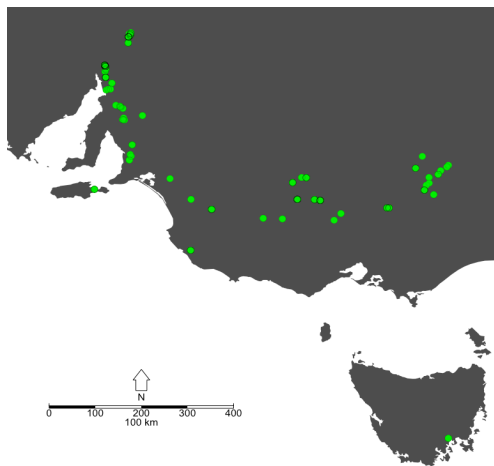
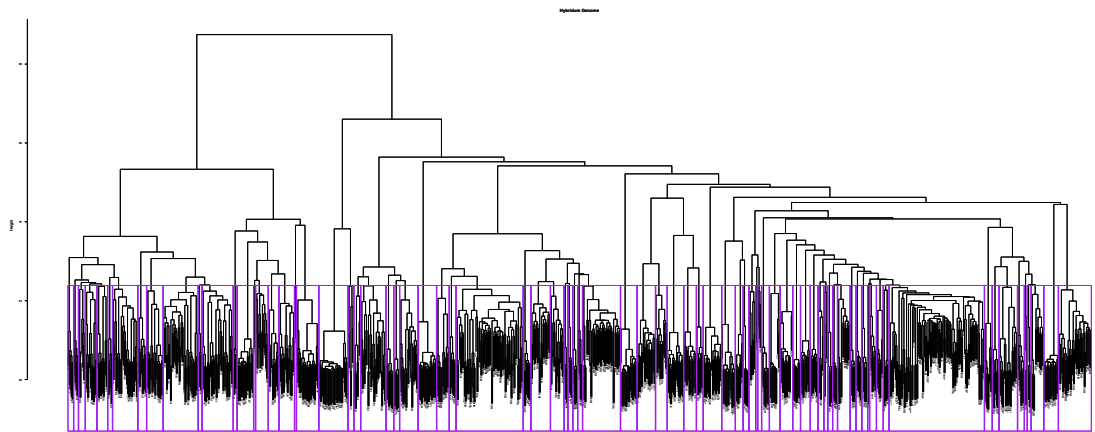


# Genetic, Geographic, and Climate Diversity of a Weedy Species: *The *Brachypodium distachyon* Species Complex*

Author: Jared Cameron Streich  
Borevitz Lab

Australian National University



A Thesis Submitted For the Degree of Doctor of Philosophy of  
Plant Biology of the Australian National University  
Submitted 21/03/2018  
© Jared Cameron Streich 2018  
All Rights Reserved

## Declaration by Author: Jared Cameron Streich


The content of this thesis is original work performed, analysed, documented, and described by the Author, me - Jared Cameron Streich, less information gathered from previously published material, and some not-yet published materials and is cited therein. Details of unpublished materials are discussed further in this body of work. The contribution of others to this work and its methods are noted accordingly and on page 3.

The aid and contribution of others is documented within the body of this work, which includes, study design, statistical analysis, laboratory procedures, computational processes designs, software, and all other services provided by others. The bulk majority of this text and my thesis is the result of my efforts to better my education and academic candidacy and merit the degree of Doctor of Philosophy of Biological Sciences at Australian National University in the Research School of Biology in Plant Science. No parts of this thesis are shared with another institution for other academic degrees.

I acknowledge that a digital version of my thesis must be lodged and submitted with the University Library and, abided by Australian National University's General Award Rules, and must be made available to the research community with accordance to the Copyright Act 1968.

I understand that a legal copyright of all material contained in this thesis belongs to the copyright holder(s). Within this thesis I have made agreements to hold copyright and reproduce material in this thesis.

The use of the reference genome for *Brachypodium stacei* ABR114 contained in this thesis was obtained with the agreement that it not be publically published unless consent was given by our collaborators Dr. John Vogel of the Vogel lab at The Joint Genome Institute at Walnut Creek, California, United States. That said the reviewers of material in this dissertation must not show this information as it is currently sensitive and should not be shared publically unless written permission is given by Dr. John Vogel at The Joint Genome Institute and/or said work is published. This thesis should not be publically available until after formal publication by Dr. Vogel and his team. Please be careful not to share this information until such time.



---

Jared Cameron Streich

## Contributions by Other Individuals to This Thesis

Contributor	Contribution/Service
Justin Borevitz	Lab leader and acting PhD mentor. Aided in editing and writing some R scripts.
Pip Wilson	Helped assimilate a functional germplasm for research use across our lab personnel, myself included, helped collect tissue from plants for sequencing.
Daniel Rosauer	Introduction to species modeling and github repositories to “kickstart” the learning of GIS in R.
Justin Borevitz, Adrienne Nicotra, Jason Bragg, and Owen Atken	Committee members, counseling and aiding project design and methods.
Kevin Murray and Megan Supple	Set up a quality control pipeline for the Borevitz lab sequence data and demultiplexing software. Helped in R and bash scripting.
Aaron Chuah, Steve Eichten, Kevin Murray, Jason Bragg, Justin Borevitz	Helped teach and write computer scripts for genotyping sequenced samples.
Cameron Jack, Aaron Chuah, Steve Eichten, Kevin Murray	Helped set up cluster nodes for my personal use of Computer space and hours on the Genomic Discovery Unit’s computer space.
Niccy Atkins and Norman Warthmann	Taught laboratory techniques associated with DNA sequencing. Some DNA extractions, and many but not all library preparations.
Pilar Catalan, Smadar Ezrati, David Garvin, Shuangshuang Liu and Kent Bradford, John Vogel, Hikmet Budak, and Samuel Hazen, Anna Caicedo.	Provided germplasm material for DNA sequencing and genomic analysis.
Josh Penalba, David Kainer, Tim Brown, Steve Eichten, Mikaela Atkinson, Kevin Murray, Joel Granados, Kieren Patchell, Caela Welsch, and Shuangshuang Liu.	Field Collection of germplasm material and whole plants.

## Acknowledgements

---

I am forever thankful to Dr. Justin Borevitz and his exceptional tolerance to being my mentor, his contribution to my education, well-being, positive effect on my mental growth under his leadership and example, and aiding my development into a member of academic society.

I am likewise thankful for Dr. Phillipa Wilson's influence and objective thinking which helped me better understand scientific methods and processes in modern science, and helping me organise our substantially sized germplasm.

I am very grateful for Kevin Murray's influence and aid in teaching me command line interface commands, R, python and computational shortcuts, and eating chocolate biscuits when I asked him to.

I am very thankful for Dr. Steve Eichten's help with learning R, linux command line, and computer operations, "How a computer looks at information."

My previous mentors, Dr. Aaron Liston for teaching me the foundations of genomics and speciation, landscape genomic ideas, as well as shape how I see plants in their native sphere.

Dr. Todd Mockler for being a very influential mentor, employing me in his lab and allowing me to conduct modern research while being an undergrad. Thank you for showing and tutoring me how to interrogate and look at problems from every angle with an unbiased perspective, and of course setting me on a path to academic success. Dr. Richard Halse who often spent extra time with me discussing plant anatomy, structure, evolution, and providing me teaching opportunities as an undergrad.

A huge thank you to my many friends in Canberra and Australia who we mutually help get through the thick and thin of PhD/Graduate student life. I absolutely could not have done it without you. As well as thank you to Angela Maruame who was a large part of my emotional support in the early years of my PhD.

A substantial thank you to Australian National University for taking a chance on me and allowing me to use your research facility to answer my scientific questions with your mentoring body, resources, and aiding disposition. The faculty, support staff, and resources are world class and deserve recognition, especially IT, NCRIS, and Plant Services.

## Keywords

---

Genetics, Genomics, Brachypodium, Brachypodium distachyon, Brachypodium stacei, Brachypodium hybridum, climate, ecotype, landscape genomics, hybridisation, niche breadth, invasion, invasion biology, Genotyping By Sequencing, species distribution modelling, polyploidy.

## Publications During PhD Candidature

\* Denotes first Author

---

Wilson, P. \*, **Streich, J.**, & Borevitz, J. (2015). Genomic Diversity and Climate Adaptation in *Brachypodium*. Genetics and Genomics of *Brachypodium*. Springer Books.

Shuangshuang Liu\*, **Jared Streich**, Justin Borevitz, Kevin Rice, Tingting Li, and Kent Bradford. (2018). "Linking molecular, phenotypic, and environmental variability to understand plant invasion trends: a population perspective" *Oikos*.

## Publications in Review

---

Wilson, P.B.\*, **Streich, J.C.\***, Murray, K.D.\*, Eichten, S.R., Cheng, R., Aitken, N.C., Spokas, K., Warthmann, N. and Borevitz, J.O., (2018). Population structure of the *Brachypodium* species complex and genome wide association of agronomic traits in response to climate. *bioRxiv*, p.246074. Submitted to *PLOS Genetics*

## Publications in Writing Phase

---

**Streich, J.\***, Liu, S., Wilson, P., Lopez, D., Eichten, S., Murray, K, Gordon, S., Ezrati, S., Budak, H., Catalan, P., Vogel, J., Bradford, K., Mur, L., Hazen, S., Garvin, D., Borevitz, J., (2017). Global Genetic and Climate Diversity of a Weedy Species: The *Brachypodium distachyon* Species Complex.

**Jared Streich\***, Kevin Murray, Justin Borevitz. (2018). GenoCLIM: "An R Package Integrating Climate and Genotype Data to the Landscape. Submitting to Journal of Open Source Software, JOSS. *In writing*  
website: <https://sites.google.com/site/genoclim/>

**Jared Streich\***, Nathan Nolte, Justin Borevitz. (2018). filterVCF: "An R Package for filtering and analysing genomic data. Submitting to Journal of Open Source Software, JOSS. *In writing*  
website: <https://sites.google.com/site/filtervcf/>.

**Jared Streich\***, Nathan Nolte, Justin Borevitz. (2018). Allele Richness: "An R Package for calculating allelic richness of a data set and plotting topographic maps of allelic diversity. Submitting to Journal of Open Source Software, JOSS. *In writing*  
website: <https://sites.google.com/site/allelerichness/>.

## Publications As Part of Thesis Body

---

None

## Abstract

---

The Introduction of novel species into non-native environments can have biodiversity and agricultural effects on landscapes costing billions of dollars in damage each year. Approximately 1.2 million hectares of land are currently deemed unusable globally because of invasive plants. The likelihood of introduced species becoming invasive isn't always understood, nor the effect of introductions immediately apparent. The environment is the primary selection force for screening habitability and is the primary selector for adaptation, but measuring all its components is complex. Therefore climate factors, precipitation and temperature, are the primary variables for determining a species distribution. The three model grasses in the *Brachypodium distachyon* complex species were used in this study because of their small sequenced genomes, classified as weedy and invasive in some regions, and were once native to the circum-Mediterranean, now global distributed. Genotyping by sequencing was used on 1,573 individuals to determine species identification and genetic diversity of each complex member. A total of 125 unique genotypes of *B. distachyon* were found from 479 individuals, eight unique genotypes of *B. stacei* from 50 individuals, and 80 unique genotypes of *B. hybridum* from 1,015 individuals. MaxEnt distribution modelling was used to find potential area using a *training specificity equals sensitivity* threshold both natively and globally. *B. stacei* was the most rare having the smallest potential area in its native range at 2,458,837 square kilometers and 3,207,524 globally. *B. distachyon* had the largest native potential area at 5,098,573 square kilometers, but rare outside its native range, Australia only. *B. hybridum* was modelled to have 3,935,266 square kilometers natively, but 6,705,946 square kilometers globally leaving 2,770,680 of potential habitat non-natively. Common genotypes of the polyploid complex member *B. hybridum* were permutation tested for global abundance across groups of regions, with the genotype NRD-1 being significantly more abundant geographically than random. NRD-1 was also used for global distribution modelling to determine global suitable regions that would be sensitive to NRD-1 introduction. The three complex species were compared for climate breadth where *B. hybridum* had the widest climate breadth of the three group members. The genotype NRD-1 was also compared to *B. hybridum* as a whole to see if the NRD-1 genotype had a similar climate breadth as the whole species, possibly defining the species climate breadth. The climate diversity within each species was used to designate climate type identities for sample locations to measure climate range a genotype occupies and the climate diversity of geographic space. The *B. hybridum* genotype NRD-1 was found in the most climate types through permutation testing and found to have a significantly larger climate breadth than average p-value <0.01. Geographic regions with high climate diversity were also found to have the most genotypes. As *B. hybridum* was found to be the most widely distributed of the three study species, many specific genotypes occurred in numerous climate types and

were sampled on multiple continents, particularly genotype NRD-1, thus were concluded as the most widely adapted *B. hybridum* and all other *B. distachyon* complex species genotypes.

### Australian and New Zealand Standard Research Classifications (ANZSRC)

---

060408 Genomics 40%

060411 Population, Ecological and Evolutionary Genetics 40%

050103 Invasive Species Ecology 20%

### Fields of Research (FoR) Classifications

---

FoR Codes: Description, Percent of Thesis

FoR code: 0602, ECOLOGY, 20%

FoR code: 0607, Plant Biology, 20%

FoR code: 0604, Genetics, 60%

## Table of Contents

---

I. Declaration by Author.....	2
II. Contributions by Others.....	3
III. Acknowledgements.....	4
IV. Keywords.....	4
V. Publications During this Thesis.....	5
VI. Abstract.....	6
VII. Australian and New Zealand Standard Research Classifications.....	7
IIX. Fields of Research Classifications.....	7
IX. Table of Contents.....	8-9

### Chapter I:

#### **Invasion Biology, Genomics, Distribution Modelling, Climate Analysis, and the *Brachypodium***

<b>distachyon species Complex.....</b>	<b>11</b>
1.1 Introduction.....	12
1.2 Species Introductions and Invasion Biology.....	13
1.3 The <i>Brachypodium</i> Species Complex: A Model for Invasion Biology.....	14
1.4 Landscape Genomics: Concepts, Practices and Current Uses.....	18
1.5 Case Studies of Landscape Genomics in Non- <i>Brachypodium</i> Species.....	21
1.6 Species Distribution Modelling and Climate Analysis.....	26
1.7 Climate Tolerance, Breadth, and Analysis of Species and Genotypes.....	31
1.8 Discussion and Questions, Hypothesis, and Aims for Each Chapter.....	34
1.9 Citation.....	39

### Chapter II:

#### **Germplasm Development, Species Identification and Regional Assessment of *Brachypodium***

<b>Species.....</b>	<b>50</b>
Abstract.....	50
2.1 Introduction.....	51
2.2 Methods.....	53
2.3 Results.....	58
Streich-Borevitz Germplasm.....	58
Species Identification by Sequencing.....	59
Distribution of Species Across Geography.....	61
Assigning Regional Identities To Collection Sites.....	62
2.4 Discussion.....	64
2.5 Data Sets and Script Links.....	67
2.6 Citation.....	67

### Chapter III:

#### **Genetic Diversity within *Brachypodium* Species.....**

<b>Abstract.....</b>	<b>70</b>
3.1 Introduction.....	71
3.2 Methods.....	76
3.3 Results.....	78
Genetic Diversity of each Species.....	79
Genetic Diversity across Geography of Common Genotypes.....	83
Dispersal Test of Common Genotypes.....	93
3.4 Discussion.....	97
3.5 Data Sets and Script Links.....	103
3.6 Citation.....	104



<b>Chapter IV:</b>	
<b>Genomic Biogeography.....</b>	<b>108</b>
Abstract.....	108
4.1 Introduction.....	109
4.2 Methods.....	115
4.3 Results.....	116
Potential Area of Species.....	117
Overlap of Suitable Ranges of Each Complex Member.....	121
Genotype Distribution Models.....	124
Genotypes per Regions.....	125
4.4 Discussion.....	129
4.5 Data Sets and Script Links.....	133
4.6 Citation.....	134
<b>Chapter V:</b>	
<b>Species Climate Classification and Diversity.....</b>	<b>137</b>
Abstract.....	137
5.1 Introduction.....	138
5.2 Methods.....	141
5.3 Results.....	143
Climate Comparison Between Species of Significant Environmental Variables.....	144
Climate Diversity of Collection Sites.....	146
Climate Permutation Tests of Climate Windows of Common Genotypes.....	147
5.4 Discussion.....	151
5.5 Data Sets and Script Links.....	154
5.6 Citation.....	154
<b>Chapter VI:</b>	
<b>Conclusion and Discussion.....</b>	<b>158</b>
6.1 Introduction .....	159
6.2 Discussion of Genetic Analysis .....	162
6.3 Discussion Biogeography .....	167
6.4 Climate to Genetic and Geographic Data.....	171
6.5 Final Discussion.....	173
6.6 Citation.....	177
<b>Appendix:.....</b>	<b>181</b>
Glossary.....	181
Abbreviations.....	183
Chapter II Supplemental Material.....	184
Chapter III Supplemental Material.....	192
Chapter IV Supplemental Material.....	205
Chapter V Supplemental Material.....	213



# Chapter I: Introduction: Invasion Biology, Genomics, Distribution Modelling, Climate Analysis, and the *Brachypodium distachyon* Species Complex

---

## 1.1 Thesis Introduction

## 1.2 Species Introductions and Invasion Biology

Global Impact

How Occurs

Control methods

Landscape Genomics for Invasion Biology Improvements

## 1.3 About the *Brachypodium* Species Complex And Their use as Invasion Model Species

*Brachypodium distachyon*

*Brachypodium stacei* & *Brachypodium hybridum*

*Brachypodium* Species as Models for Invasion

*Brachypodium* Collection Locations, Populations and Study Germplasm

## 1.4 Landscape Genomics: Concepts, Practices and Current Uses

Collection Sites and Landscape Coverage

Neutral Forces Affecting Genetic Signal of Climate Adaptation

Testing Adaptation to Climatic Range

Demography and Genetics

## 1.5 Case Studies of Landscape Genomics in *Non-Brachypodium* Species

Examples of Landscape Genomics to illustrate concepts

Collecting Sites and Landscape Coverage

Genotyping by Sequencing

Neutral forces affecting genetic signal of climate adaptation

Testing adaptation to climatic range

Demography and Genetics

Case Studies Of Landscape Genomics

*Model Species*

*Non-model species*

*Agricultural species and close relatives*

## 1.6 Species Distribution Modelling and Climate Analysis

Modelling Programs

MaxEnt

Modelling Native Range

Modelling Non-native Range

Modelling Invasive Species and Suitable non-native regions

## 1.7 Climate Analysis of Species and Genotypes

Climate Breadth

Determining the geographic range and climatic niche breadth of divergent genetic lineages

Environmental Data Layers: BioClim, WorldClim, and ALA

## 1.8 Thesis Questions, Hypothesis, and Aims by Chapter

Chapter 2: Collections and Germplasm

Chapter 3: Species Identification and Genetics Analysis

Chapter 4: Genomic Biogeography

Chapter 5: Climate Analysis

## 1.9 Discussion

## 1.10 Citation

## 1.1 Thesis Introduction

---

The emerging discipline of landscape genomics seeks to understand the multiple effects of environment on the geographic distribution of populations and alleles within species due to differential fitness underlying local adaptation. The use of landscape genomics has been used successfully to identify neutral processes of population migration and to identify the alleles under environmental selection (Fournier-Level, 2011; Shen, 2014; Platt, 2015).

Species introduction events can have positive, neutral, or negative impact on natural or agricultural landscape productivity. Many introduced species can become invasive and disrupt existing systems making them less biologically productive or less diverse by displacing native individuals. It is estimated that an approximate 1.2 million hectares are overrun each year to invasive plant species in the United States (USDA Forest Service, 2016). Invasive species cause ecological damage measured in the billions of dollars and Australia spends approximately \$3.4 billion a year to combat just invasive plants, while the United States spends up to \$34.7 billion (Pimentel, 2000; Schmidt, 2012; Australian Bureau of Statistics 2012). The study of introduction events of novel organisms to non-native habitats is commonly called *invasion biology*. In many ways invasion biology demonstrates many of the core principles of biology, such as adaptation and rapid evolution. Natural invasions are uncommon in nature and rarely measured. Prehistoric migration events can be disentangled through genomics, the variation and species divergence in phylogenetics, and the fluctuations in geologic features across large timescales. However, many organisms were introduced alongside the rapid geographic expansion of human society. Landscape genomics concepts and tools can inform invasion biology by revealing the number of introduced genetic lineages and their subsequent migration patterns (Bakker, 2009; Takahara, 2013).

The geographic locations of a species can be used to model the distribution of viable habitat (Phillips, 2004). This is often used to determine the existing and suitable range of a species. Locations that are predicted to be suitable, but not yet colonized, may be vulnerable to invasion. Theoretically, the predicted suitable range may differ for closely related species or even different genetic lineages within a species. This can be due to alternative preferences in climate or soil variables. Further, genetic lineages may have narrow or wider tolerances for climate variables. Thus, climate breadth can reveal widely adapted genetic lineages that tolerate a large range in climate variables. Widely adapted lineages may be more prone to become invasive. Alternatively, many different specialist lineages can also occupy a wide environmental range. Genomic approaches can distinguish between these possibilities and inform management of invasive species.

Understanding the interaction between landscape and organism is paramount in landscape genomics. The start of any landscape genomic study is a geographic and genetic diverse germplasm. Creating a meaningful landscape genomic data set is difficult without prior knowledge of sample locations or genetically diverse hotspots. Germplasm assembly can be optimised by pacing collection efforts with genetic screens and distribution modelling techniques that use previous sample locations to predict new regions of genetic, climatic, and geographic diversity. Many distribution-modeling programs can determine what climate variables have significance to the breadth of a species precipitation and temperature tolerance. However, wild collected individuals could express similar phenotypes to closely related species, but are misidentified at a species level. A well-designed genetic screen can identify species, ancestral lineages, family groups, and genotypes, thus greatly increases information about a germplasm.

The key tasks of this thesis are four fold. This project required a carefully assembled and globally diverse germplasm to represent the genetic diversity of both native and non-native regions. Once the germplasm collection was compiled, samples were grown and DNA sequenced to determine their species and genotype identity. Subsequently, collection locations were sorted by species and genotypes to interpolate the size of globally suitable geographic distributions revealing the differences in their fundamental range limits. The final task was to distinguish climate generalists from specialists by determining the number and type of climate classes that widespread genotypes were found in relation to a random or neutral expectation.

## **1.2 Species Introductions and Invasion Biology**

---

Society has become more cosmopolitan each year and countries around the world face introduction events at many ports globally. Introduction events can happen many times, and many have occurred before the concept of an invasive species was first termed in Charles Elton's book *The Ecology of Invasions by Animals and Plants* (Elton, 1958). Settlers have colonised the new world for centuries and the import and export of species to and from habitable spaces with little to no regard for the local impact an introduction event might inflict on the landscape. While the *Brachypodium distachyon* species complex is not classified as invasive in Australia, it is an introduced species to the continent and is a set of closely related model cereal C3 grass species.

Species can be introduced a variety of ways, but usually by contamination of a product or on purpose for aesthetic, agriculture, or personal means, as is the likely case with Burmese pythons (*Python molurus*) in southeastern United States, and Paterson's Curse (*Eichium plantagineum* and *Eichium vulgare*) in Australia (Wilson, 2011; Konarzewski, 2012). Most invasive plants

have been and are introduced by the horticulture trade (Burt, 2007; Hulme, 2009). However, the assessment of magnitude an introduced species impact has on novel locations is not immediately tangible. It could take many years or even decades to assess if a species is invasive and at that point it is likely too late to eradicate without significant efforts (Myers, 2000). It might also be that some species are naturalised in some areas but invasive in nearby habitats, so a high-resolution map of sensitive habitats of suspect invaders would be ideal resources for land managers.

Herbarium records are a great place to start describing the geographic and climate range of introduced species. Records often have metadata about microclimate, some phenotype data including whether the plant was flowering, in addition to when and where the plant was collected. However, herbariums rarely if ever have any genetic data that can identify cytological differences or cryptic species. To address this, records in herbariums can aid a researcher's decision about when and where to travel for future collections, what trait(s) to use for identification or analysis, and what locations they historically have occupied. Using previous collection points, researchers can travel back to the sample's original location, or similar locations based on computer based simulation models that predict species specific suitable geography like the program MaxEnt, discussed more below (Phillips, 2005; Banta, 2012). Once samples are obtained through collection trips a species can be more thoroughly analysed using formal analysis through genomics.

A landscape genomics approach comparing the genetic diversity of both the native and introduced range can identify the location of origin(s), the climate patterns and breadth of that species locally and natively, and the surrounding sensitive regions. The screening of genetic diversity across non-native landscapes can reveal regions that require more attention, such as locations with high genetic variation. Regions with more genetic diversity will have more opportunity to create unique genotypes. By using geographic coordinates of presence locations, global species distribution models can calculate vulnerable non-native geography. This is true for widespread genotypes that are found in a wide breadth of climate locations and they can be modelled at a genotype level for distribution modelling. The genomic analysis by DNA sequencing can help researchers; land managers and agricultural sectors better understand the introduction process and how a species colonises new habitats.

### **1.3 The *Brachypodium distachyon* Species Complex: A Model for Invasion Biology**

---

*Brachypodium distachyon* is an ideal model species for C3 grass monocots because of its phylogenetic placement in the Poaceae plant family, particularly agricultural grass crops like barley and wheat (Draper, 2001). Like *Arabidopsis thaliana*, *B. distachyon* has a small stature,

grows well in laboratory conditions, and a compact genome (~266mb) with a haploid chromosome number  $x=5$ . *B. distachyon* is also easily transformable, and a moderate array of accessions from diverse geographic regions (Draper 2001, Opanowicz, 2008; Vogel, 2009; Vogel 2010, Mur, 2011). Since its initial proposal as a model organism, numerous preceding papers have published using *B. distachyon* in a range of topics. A transformation protocol was developed in 2005 and improved in 2008 where multiple accessions were tested for transformation amenability and efficiency (Christiansen 2005; Vogel, 2006; Vogel, 2008). Several studies also show that several inbred maternal lines have been developed and tested for susceptibility to cereal grain pathogens with variation in phenotypes (Draper, 2001; Parker 2008; Peraldi, 2011; Alderman, 2013). The full genome was published in 2010 and now a nearly complete version 3.1 is available. Furthermore, there has been an uptick in collection efforts from many research groups generating large germplasms for research purposes (Vogel, 2009; Catalan, 2012; Tyler, 2016; Shiposha, 2016).

The *B. distachyon* complex species (*B. distachyon*, *B. stacei*, and *B. hybridum*) are also great models for the study of introduced and invasive species as they exhibit wide variation in many “weedy” traits, such as: life strategies, high seed yield, and high self-fertilisation rates (Draper, 2001; Vogel 2009; Vogel, 2010). Currently there is a substantial germplasm available for research from the USDA and many research groups created their own collections from native and non-native locations. Beyond studies of *B. distachyon* germplasm and a smaller study on *B. stacei*, there is no definitive paper that describes the genetic relationship of most or all known populations of published accessions of the *Brachypodium distachyon* species complex (Tyler, 2014; Shiposha; Vogel, 2009).

#### Genomes of *Brachypodium distachyon*, *Brachypodium stacei*, and *Brachypodium hybridum*

Initially *Brachypodium distachyon* was considered one species with variation in cytotypes, one diploid, one tetraploid, and one hexaploid, but later discovered that *B. distachyon* had various cytotypes (Hasterok 2004; Opanowicz, 2008; Filiz, 2009; Idziak, 2011). When the planning of this project started in late 2011 *Brachypodium distachyon* was still considered one species of various diploid and polyploid cytotypes and substantial phenotypic variation. Thus the species was expected to have significant genetic diversity, and based on collection locations where it was found a large geographic range. Later, it was discovered that there were two diploids and one allotetraploid (Idziak, 2011). The three species were assembled into a species complex composing of the most commonly researched *B. distachyon*. The other two species were described in 2012 and are *B. stacei* the other diploid with a haploid chromosome number of  $x=10$  with a genome size of approximately 240 mb, and the allotetraploid, which is a hybrid composing of both diploid complex members (Idziak, 2011; Catalan, 2012; Shiposha, 2016).

The polyploid *B. hybridum* is an allotetraploid of a *B. stacei*-like and a *B. distachyon*-like set of ancestors detected by FISH probes  $2n=4x=30$ , with a haploid chromosome number of  $x=15$  ( $x=10+5$ ) and about 510mb (Hastorak, 2008; Idziak, 2011; Catalan, 2012). When discussing these three species of *Brachypodium* in this thesis, *B. distachyon* and the *B. distachyon*-like subgenome in *B. hybridum* are referred to as the D genome collectively. Likewise, *B. stacei* and the *B. stacei*-like subgenome in *B. hybridum* is referred to as the S subgenome collectively. The naming regime employed for *Brachypodium* genomes involved in this thesis is subject to change once said genomes are fully described. The naming scheme employed here describes the current understanding of ploidy and genomes of these species.

*B. hybridum* has a manageable genome size of about 510mb, about the size of *Setaria* models (520mb) and Eucalyptus (600mb) thus making it an ideal polyploid model (Padovan, 2013; Huang, 2014). Many phenomena are not well understood in allopolyploid genomics and *B. hybridum* could be a suitable model for many polyploid effects: trans-chromosomal signaling, RNA and protein dosage effects, mixed protein complexes from multiple subgenomes, fixed heterosis, multiple subgenomes donating to protein complexes, chromosome dominance, and other traits found in other grass polyploid species, particularly in wheat- *Triticum* species. *B. hybridum* is also an ideal model for landscape genomics to answer questions about species introduction events, how polyploids can have more allelic variation that can expand niche breadth, have larger climate envelopes than close diploid relatives, and will be discussed in chapters three, four, and five.

#### Collections and Accessions of *Brachypodium*

Within *B. distachyon* there are many natural accessions that fall within two major clades, dubbed the A and B groups and each of these have east and west Mediterranean sub-groups (Wilson, 2015). Beyond this major split, another publication reported a C subgenome from central Europe (Tyler, 2016). *B. stacei* has been considered a rare species, but little was known about where it grows until more recorded locations were reported and its distribution mapped (Lopez, 2015). With more knowledge of where *B. stacei* grows more collections should be pursued to advance its use as a research organism and knowledge of its genetic variation. A recent paper Shiposha 2016 shows the genetic diversity in 19 western locations of the Mediterranean and Atlantic islands (Shiposha, 2016). The other previous study to demonstrate genetic diversity in *B. stacei*, only had three individuals from Sicily (Tyler, 2016). With these two published works and the analysis from this thesis, *B. stacei* could soon become a popular research model. *B. stacei* is considered to be the closest living relative to the first *Brachypodium* species as it maps well to *Oryza* species (personal comment from John Vogel, 2015). Learning more about *B. stacei* can help us better understand the evolution of the Triticeae (wheat, barley, oat species) from the Ehrhartoideae (*Oryza* species) since the Brachypoidieae (*Brachypodium*



species) lay between the two groups and each harbour many domesticated grain species that form most of the caloric intake of human nutrition (Catalan, 1997; Hands, 2012).

#### Using *Brachypodium* species as Models for Invasion Biology

*Brachypodium* species native range spans the mediterranean region, Europe, North Africa, the Middle East, and much of West Asia. However, *B. hybridum* has been introduced to novel regions of the world including the Southern African continent, North America, South America, and Australia. The first analysis of non-native *B. distachyon* polyploid cytotypes described the population structure and characterised the weediness traits in non-native regions of the United States North America and constitutes the only paper currently published of introduced *B. distachyon* polyploids (Bakker, 2009). Introduced *Brachypodium* species are reported in Southern Africa, South and North America, and Australia. Records in Australia date back to 1880 near Adelaide (ALA, 2013). As a suspected grain contaminant, *Brachypodium* possibly arrived by  $\approx$ 1780-1800 during early grain trials in Australia (Australian Bureau of Statistics, 2006). One report states that wheat was grown in Parramatta by 1789. Commercial cereal agriculture started by 1816 near Sydney, and Tasmania by 1826 (Morgan, 2003). Another source indicates wheat flour was imported to South Australia up until 1839 from Tasmania until wheat agriculture was developed in the Adelaide Plains, and that *Brachypodium* herbarium records indicate its presence as early as 1898 (ALA, 2016; PIR, 2013).

In Australia, the *B. distachyon* complex species are considered naturalised, except in Western Australia where it is considered “weedy” and possibly changes fire regimes (Hussey, 2007). In most wild populations these three species are obligate selfing plants with cleistogamous florets, small stature and quick life cycle, though it was reported in a study in modern day Tunisia that *B. hybridum* has higher than expected outcrossing rates at  $N_m=2.31$  (Draper, 2001; Garvin, 2008; Neji, 2015). It was also reported in two different publications that lineages of both *B. distachyon* and *B. hybridum* disperse widely with the same lineages being found across large geographic ranges (Dell’Acqua, 2014; Neji, 2015). *B. hybridum* appears to be the most widely adapted of the three species because of its climate diversity and expanded geographic distribution beyond its native range, probably from having fixed heterosis effects from its two sub-genomes (Garvin 2008; Catalan 2012). For *B. hybridum* to survive in many climates and vector to novel environments on non-native continents, it is expected to be a widely adapted species. *Brachypodium distachyon* species complex exhibits other interesting traits common with other weedy invasive species: high selfing capacities with rare occurrences of outcrossing, can make adventitious roots, shows dramatic phenotypic plasticity, and can vegetatively overwinter in some environments (Bakker 2009; Vogel 2009; Garvin 2008; Catalan 2012; personal observation). The progressive research, genomic tools, private and public germplasms from many continents make the *Brachypodium distachyon* complex species ideal models to study

landscape genomics and use as model plants for studying wide adaptability and invasive behavior in grasses (Garvin, 2008; Bakker 2009).

By understanding the genetic basis of local adaptation, we may find allelic variants that allow some groups to be adapted to specific environments (specialists), and other alleles may be important for generalists (Storz, 2005). *Brachypodium* species, as a whole, have a large geographic range, mostly spanning Mediterranean-like climates in North Africa and Eurasia. The species as a whole could be widely adapted, but groups living in isolation and thus genotypes are associated with geographic distance like in *A. thaliana*, which has genetic diversity associated with geographic and climatic space (Hancock, 2011). Some genotypes could be more frequent than others because of two possible reasons: (A) preferred climate type is common across geography, or (B) they have wide climate breadth (Platt, 2010; Horton, 2012; Banta, 2012). In either case, some variation in dispersal mechanisms could be affecting how widespread a genotype is. Since all three species are annual plants, there could also be variation in seed dispersal ability of common genotypes versus rare genotypes.

#### **1.4 Landscape Genomics, Concepts, Practices, and Current Uses**

---

The goal of landscape genomics is to identify the genetic basis of environmental adaptation using natural collections distributed across a variable landscape. Genomic variation is identified and then associated with environmental variables at the site of origin. This approach requires numerous independent samples across environmental gradients for statistical power (Bragg, 2015). Samples may be grouped by historical demographic processes that aren't independent, in which case wider sampling may be necessary. Environmental filtering at an adaptive locus can positively increase the allele frequency of a beneficial allele or negatively select against a deleterious allele. Most variants are neutral in most locations, however a few are positively or negatively selected in particular locations as seen in *A. thaliana* (Shen, 2014).

##### Collection Sites and Landscape Coverage

Study locations are environmentally complex, but can be simply described by temperature and precipitation- abiotic stress. Specifically, 19 derived BioClimatic variables summarizing temperature and precipitation are commonly used for species distribution modeling (BioClim, 2016; Busby, 1991; Beaumont, 2005). Often many researchers are involved in collections further complicating landscape genomic studies. This thesis is the product of eight research groups collaborating to meet mutual and/or overlapping goals. Each collaborating research group employed different sampling regimes; some groups sampled more locations with fewer individuals or vice versa. Some research groups sampled across smaller and larger geographic space. The inconsistency of sampling regimes across collaborating partners required grouping

geographic regions per each species and is further discussed in Chapter II to normalise geographic and climate analysis.

#### DNA Sequencing: A Quick Review of Genotyping by Sequencing

Sequence technologies have become more efficient and more financially attainable in the last decade (Shendure, 2008; Elshire, 2011). Large-scale genomic scans of many individuals can vary in technique depending on genome size and allocated research budget. Whole genome sequencing of smaller genomes like bacteria and yeast easily assemble and are amenable to many research budgets. Diploid genomes with moderate outcrossing rates in the  $\leq 500$  mb range could easily be scanned using low coverage whole genome sequencing of about 100 samples per lane using Illumina platforms. Whole genome sequencing starts to get more expensive beyond one gb genome size at 50-100 samples per lane. At this point genotyping by sequencing (GBS) becomes more practical, but only provides relatedness information and less specific quantitative trait loci mapping resolution. Once optimal resolution of trait mapping is reached with GBS, whole genome sequencing of trait carrying individuals can be performed. Likewise, datasets with many hundreds to thousands of samples of smaller genomes (100-500mb) also benefit by using genotyping by sequencing and multiplexing, having upwards of 196-384 samples per lane to describe relatedness, then switching to whole genome sequencing for fine scale mapping of quantitative traits of individuals of interest.

The use of sequence technology in ecological and environmental studies has provided strong statistical evidence to describe lineages of organisms across landscapes and environmental gradients. One example is the genetic analysis of 5,707 plants screened at 149 SNPs in the paper The Scale of Population Structure in *Arabidopsis thaliana* (Platt, 2010). Within this paper they conclude that *Arabidopsis thaliana* has 1,799 unique genotypes, thus for high volume phenotyping screens, these would be the core mapping set of phenotypic traits of this species (Platt, 2010). This study also sequenced plants from both the native and non-native range and compared genetic diversity between the two. Genetic diversity was found to be higher in the native range, but some locations in the non-native range harboured many different genotypes and possible admixed locations. Another study looked at *Brachypodium sylvaticum* to see if invaded locations had genetic bottlenecks and trace some lineages to central west Europe (Rosenthal, 2008). That study found three major genetic lineages in North America ( $K=3$ ) and five major genetic lineages globally ( $K=5$ ). They only traced one of the three lineages in the native range to the invaded range, leaving the origin of two prominent genetic lineages unresolved.

### Neutral forces affecting genetic signal of climate adaptation

The distribution of a genetic lineage is the result of both adaptive and neutral forces. With limited migration, founder effects and isolation by distance, prevent our ability to separate the chance historical demographic effects on the entire genome from the adaptive genes. We must consider the genetic divergence and population structure when selecting locations for intensive sampling (Bragg, 2015). Recently founded populations may not have enough diversity to separate specific chromosome segments, or haplotype blocks, associated with climate variables from the background of a structured population. In this case, adaptive loci cannot be partitioned from background variation. Geographic distance can often explain genomic isolation within a species, which is a neutral process. This has traditionally been detected using Mantel Tests or Partial Mantel Tests (Mantel, 1967), however recently it was shown that mantel tests do not remove the sampling bias and spatial structure (Guillot, 2013).

### Testing adaptation to climatic range

Typically common gardens and reciprocal transplant studies are used to test for local adaptation where the home genotype performs superior to the genotype from farther away, example is Clausen *et al.* 1940 (Clausen, 1940). Provenance trials do this on a larger scale, evaluating phenotypes from a broader range of genotypes across many locations. These studies are massive and are certainly hampered by starting conditions and weather variation. One solution is to use growth chambers that can mimic natural climatic conditions, without weather noise. Technology being developed in the Borevitz lab can synthesize environments in growth chambers. These specialised growth chambers are called The *SpectralPhenoClimatron* that allows fitness traits to be measured via advanced high-throughput phenotyping digital imaging techniques in multiple target environments (Brown, 2014). These advanced smart chamber experiments can be run multiple times at most yearly time points, thus bringing the environment to the researcher.

### Demography and Genetics

Isolation by distance analysis on a species with samples across its native range should show some signal of genetic diversity and aid population structure description. With *Brachypodium* having a large native range, spanning three continents, one can expect to find samples that are significantly diverged from one to another (Opanowicz *et al.*, 2008; Garvin 2008). The trends within a mantel test should be relatively linear and organised. It would be expected the introduced range of *Brachypodium* to be more randomised showing little genetic structure between collection sites, being more pervasive/aggressive genotypes have been vectored to new locations and possibly introduced multiple times. Mantel Tests between native and introduced ranges should be very dissimilar in shape and description, because one would assume

introduced locations have been randomly invaded. Partial mantel test can describe if geological features are significant factors in isolating populations/genotypes.

### 1.5 Case Studies of Landscape Genomics in *Non-Brachypodium* Species

---

The use of model species has greatly improved the consistency and reproducibility of science both past and present. The increased output, affordable costs, and quality of genomic sequencing has further improved model organism genomics to where specific loci responsible for phenotypes of interest can be interrogated for function and interaction among the organism as a whole. The use of model organisms extends to landscape genomics as they are derived from natural populations. Many successful studies have been performed on model species to interrogate the links and associations between genetics and environment. Beyond model organisms, techniques to call useable markers have improved in non-model species as well. While the use of second and next generation sequencing is rapidly being applied to non-models or even whole genome sequencing, SSR markers and exon capture have accurately called the basic population structure and relatedness of individuals in diversity studies. Though SSR markers are limited in their ability to find adaptive regions of the genome, other techniques like RNA sequencing can identify the transcriptome and align reads against *in-silico* RNA-DNA synthesized loci set. Other techniques also work using kmer comparisons of reads like in KWIP (Murray, 2017).

#### Model organisms

##### *Arabidopsis thaliana* and other species

Composing of five haploid chromosomes and a small genome approximately 119mb the model plant *Arabidopsis thaliana* has provided scientists with enormous amounts of information about plant metabolism, function, growth, physiology, evolution, and photosynthesis. The genome was first published in 2000 and the mapping and functions of genes have been explored even earlier (Arabidopsis Genome Initiative, 2000). *A. thaliana* is easily transformable and readily grows in laboratory conditions, requiring little light to grow and no significant soil associates needed to accommodate growth. The long-standing interest in the scientific community in *A. thaliana* pushed collection efforts from many locations, which lead to natural variation studies and landscape genomics (Mitchell-Olds, 2001; Tonsor, 2005; Ågren and Schemske 2012; Ågren, 2013).

Many landscape genomic studies have used *Arabidopsis* species to study genotype-by-environment effects. *A. thaliana* has been particularly successful in landscape genomic studies. The environmental variation across sample locations could predict patterns of polymorphisms across the whole genome as described in one study of *A. thaliana* (Lee & Mitchell-Olds, 2012)

Further in that same study, some polymorphisms were also predicted based on genomic structure and composition; and that environmentally relevant factors contribute to population divergence across populations, and locally adapted genotypes. A similar study showed a pattern across geographic space where suits of inherited genetic markers were present across specific landscape gradients (Hancock, 2011). Some locations overlapped geographically and levels of polymorphisms present per location would be predictive of fitness at one location. Thirty different biological processes were found ecologically relevant across numerous environmental factors with significant p-values. It was found that non-synonymous variants in climate associations were more common than synonymous variants, which proves that polymorphisms are more likely to change protein coding regions within genes to adaptive alleles (Lasky, 2012). A study in *A. thaliana* discovered an early stop codon in a methylation transferase CMT2 allele conveying a larger climate tolerance in those *A. thaliana* individuals carrying that allele (Shen, 2014). Variation in the gene DOG1 in *A. thaliana* was found to be associated with different soil types, where seasonal germination time was strongly correlated with delaying germination by temperature sensitivity and altering abscisic acid metabolism (Chiang, 2011).

#### *Setaria viridis* and *Setaria italica*

The new cosmopolitan model species *Setaria viridis* and *Setaria italica* are diploid C4 grass species composing of nine haploid chromosomes and about 512mb genomes (Bennetzen, 2012; Lata, 2013). *Setaria* species are in the Panicoideae subfamily of the Poaceae and are phylogenetically placed near many agriculturally relevant species such as *Sorghum bicolor*, switchgrass *Panicum* species, *Zea mays*, proso millet and perl millet (Bennetzen, 2012). In some parts of the world *Setaria* species are grown as important food sources as well (Jia, 2013). *Setaria* species have evolved C4 photosynthesis independently from other closely related members, but function as ideal models for more complicated genomes using C4 type metabolisms, however *Setaria* species are also ideal models for cell wall biosynthesis, response to drought, and particularly *S. viridis* in invasion biology (Bennetzen, 2012; Doust, 2017). The genus *Setaria* are transformable in multiple methods, small to medium stature plants, and have both diploid and polyploid relatives (Brutnell, 2010; Martins 2015; Saha, 2016).

Currently there are few landscape genomic studies of *Setaria* species due to the genome sequence becoming available in recent years and the building of a public germplasm. A substantial genomic diversity and population genetic study using 273 individuals showed the structure of *S. viridis* and *S. italica* across Europe, Asia, and North America, with outlier locations in South America (Huang, 2014). While no specific genetic correlations were associated to climate, the two model species grow in vastly different ecological environments. A comparative genomics study found flowering time variation in *S. italica* and *S. viridis* F7 RILs that showed variation in flowering time and morphology occurs in four different growing

environments, which is evidence that plasticity in flowering time and phenotypes should vary across ecological gradients and future studies could investigate adaptation to local environments in *Setaria* species (Mauro-Herrera, 2013). *S. italica* is most commonly grown in arid environments for human consumption, due to its sizeable range it should also be investigated for environmentally induced phenotypic plasticity or local adaptation (Jia, 2013).

#### *Pinus taeda*

Softwood trees are often considered agricultural species, but there are few to no domesticated lines. Due to the large genomes of most gymnosperm plants the genomic analysis is a significant challenge compared to ordinary model organisms as well as the lifespan and height of most tree species (Morse, 2009). The reference species *Pinus taeda* (Loblolly Pine) is a very large 22 Gb genome with a haploid chromosome number of  $x = 12$ , but many other *Pinus* species have much smaller genomes. In *P. taeda* linkage disequilibrium decays quickly, it has a large geographic and ecological breadth in its native range, and transformation protocols have been developed (Gould, 2002; Brown, 2004; Krutovsky and Neale 2005; Heuertz, 2006). All *Pinus* species are diploid with the same chromosome number. While induced polyploid individuals have poor fitness, and interspecific hybridization is successful (Williams, 2002).

Ecological variation occurs across the native range of *P. taeda* and several studies have published on climate to genotype interactions within the species. A subset of 1,730 genomic markers from 682 individuals across 54 locations was collected to investigate the ecological genetics of *P. taeda*, which revealed strong correlations between geography and climate (Eckart, 2010<sup>1</sup>). In this study, numerous variants were correlated with elevation or climate data and annotation reveals possible pathways that are associated with local adaptation most via abiotic stress, which would indicate some sort of environment based selection pressure. A separate study found five variants associated with aridity with significance to both biotic and abiotic stress response, 24 other variants were associated with high  $F_{st}$  and physiological processes (Eckert, 2010<sup>2</sup>).

Serotiny, the effect of a trigger response to induce seed dispersal from the maternal plant is a trait common in gymnosperms (Johnson, 1993; Bond, 2005). The measure of the serotinous phenotypes was conducted in *P. taeda* in three different ecologically and genetically distinct populations to investigate serotiny as an adaptive phenotype resulting in 11 loci that explain  $\approx 50\%$  of the phenotypic variation (Parchman, 2012).

#### *Eucalyptus species*

Tree species usually are long-lived and the duration to reach maturity is much longer than model species. Trees, nevertheless, are anthropocentrically important in both native and plantation forests for their timber and ecosystem services. Population studies in forest tree

species often have weak population structure, large sample numbers, and high neutral genetic diversity, which aides resolving locus-specific diversifying selection from a weak background of neutral noise (Potts, 1997; Krimi, 2006, Savolainen, 2007; Eckart<sup>1</sup>, 2010; Bradbury, 2013). As seen in *P. taeda*, landscape studies in *Eucalyptus* have provided significant insight to tree landscape genomics. *Eucalyptus* has superior pulp qualities for paper and a potential for biofuel resource (Myburg, 2011; Kainer 2015). *Eucalyptus* species are also ideal models with *E. grandis* (subtropical) and *E. globulus* (temperate) each having reference genomes (Myburg, 2011; Myburg, 2014). *Eucalyptus* species have a variety of uses for society mostly in oil synthesis, energy, and fiber (Kainer, 2015). Many biologically unique genomic qualities exist in *E. grandis* that make it an interesting model species. For example, more than 1/3 of all genes in most *Eucalyptus* species are tandem duplicates; a manageable diploid genome size of ~640 mb and a haploid chromosome number of  $x = 11$ ; close relative to many Myrtaceae; and synteny to other rosid genomes like *Vitis* species (Grattapaglia, 1994). *Eucalyptus* species inhabit temperate to tropical regions, spatially found across >200 million hectares in Australia, and are present on six continents globally (Pires, 2009). Most *Eucalyptus* species exhibit high levels of outcrossing and LD is relatively low, making environmental association genetics ideal across landscapes with high gene flow between individuals discussed below. There are many landscape genomic studies of *Eucalyptus* species, but few with genome coverage sufficient to comprehensively identify causal associations to landscape or climate beyond outliers from genomic background.

Using climate data sourced from ANUCLIM, a study of 274 individuals from nine geographic locations of *E. tricarpa* found 94 sequence tagged markers across the genome that were found adaptive along an aridity gradient across the southeastern region of the Australia continent (Steane, 2014). The detection of correlated adaptive loci was performed using BAYSCAN V2.1 and canonical analysis of principal coordinates. Also, *E. tricarpa* was found strongly correlated with geography,  $R^2=0.72$ ;  $p$ -value = 0.001, specifically along an east-west gradient. Another *Eucalyptus* study of the foundation species *E. globulus* used 16 microsatellite markers across 444 trees from 39 regions to resolve population structure and gene flow between regions of southeastern Australia including the island of Tasmania (Yeoh, 2012). Within this study, five distinct population groups were found by using Evanno's  $\Delta K$  from data created by STRUCTURE v. 2.3.1 when testing  $K=1-39$ . The five groups were followed up by a neighbor-joining tree which matched consistently with both geography and STRUCTURE  $K=5$ . Across all geographic locations no isolation by distance was detected, but this could be due to low quantity of markers. However, isolation by distance via mantel test did provide meaningful results between sub-groups of ancestral populations.



In a landscape study consisting of 596 individuals from 21 locations along the southwest coast of Australia, the species *E. gomphocephala* was examined for genetic diversity, population structure, and genetic association to the environment. A homolog of a *CONSTANS*-like gene was found as an  $F_{st}$  outlier to in four different climate variables: winter solar radiation, summer precipitation, aridity, and potential evaporation (Bradbury, 2013). The genetic diversity was calculated by using 18 SSR markers, also used in isolation by distance. Genetic diversity correlated to geographic distance with an  $R^2=0.362$  with p-value < 0.001. Population structure was calculated using the popular program STRUCTURE v.2.3.2.1. The optimal population number was calculated as  $\Delta K=2$  via the program Structure Harvester and the Evanno, *et al.* 2005 method. Genetic diversity and population structure largely confirm each other's results.

### Non-Model Species

#### *Helianthus species*

One of the most invasive and species rich plant families is the Asteraceae. Within this family is the *Helianthus* genus which teeters in many categories: agriculturally as a model organism for Asteraceae species is *Helianthus annuus*; as an agricultural crop commonly known as sunflower; and many species as an invasive (Blackman, 2011; Whitney, 2010). Species in the *Helianthus* genus also readily hybridises, creating admixed lineages with high genetic diversity (Baack, 2005; Prentis, 2008; Kane, 2009; Scascitelli, 2010). It has been found in *Helianthus* species that gene transfer can occur between two groups of the same species via another species hybridization (Scascitelli, 2010). Widely hybridizing species can be formidable pests for land managers as they likely can reinvent themselves once faced with a selection pressure, this could be especially true for *Helianthus* species. The hybridisation between domesticated *H. annuus* introduced natural strains creates an agricultural weed that can disrupt harvest yield by altering seed dormancy and seed shattering profiles in Spain and France (Muller 2009; Presotto 2014).

### Crop Species

#### *Oryza species*

The *Oryza* genus of plants, commonly called rice, is one of the most well studied and researched plants to date. It is also one of society's most consumed crops where approximately 50% of the world's population is fed by rice species (Zhang, 2005). The domesticated rice species *O. sativa* includes two subspecies *O. sativa ssp sativa* and *O. sativa ssp indica*. Due to the obvious anthropocentric necessity and small genome at  $\approx 380\text{mb}$ , *O. sativa* was one of the first plant genome assemblies and research on rice is active. There are more than 120,000 accessions in seed stock centers and research groups. Many studies focus on finding functions and biochemical pathways of adaptive genes. In such a useful genus many studies have explored agronomic phenotypes that are associated with specific genomic regions that are ideal for breeding.

*Oryza sativa* as a cultivated species has many developmental and morphological traits to improve yield and harvest. One study examined 413 highly unique *O. sativa* accessions from 82 countries using 44,100 SNPs to examine phenotypic variation in 34 developmentally important traits as well as resolve genetic diversity and ancestral history (Zhao, 2011). Dozens of variants were found to be influencing many complex traits by using GWAS and variation in traits was associated with population structure calculated by EIGENSOFT PCA. Interestingly admixture occurred within each subspecies and not between *O. sativa* and *O. indica*. Phenotypes were on many occasions associated with specific quantitative loci, and several of those were homologues to previously characterised genes in *A. thaliana*. Linkage disequilibrium varied greatly between populations and genomic loci, varying from 100kbp to over 2Mbp. The long LD is likely caused by the homozygosity of inbred agricultural lines. Using individuals from all ancestral groups LD decayed quickly compared to within groups, as expected given deeper shared ancestry.

### *Panicum species*

Switchgrass is both an agricultural and natural species. The investigations about its demography have been well studied despite its complicated genome and ploidy levels. Local adaptation to specific climates could be one of the driving forces of ploidy in *Panicum* species (Lu, 2013; Costich, 2010). *Panicum* species are widespread across North America and have been well sampled geographically (Morris 2011; Lu, 2013; Morris 2011; Grabowski, 2014). The use of the UNEAK pipeline was able to efficiently call markers without a reference genome in *P. virgatum* providing 29,221 genomic markers that were used to distinguish population structure and ploidy in 540 individuals from 66 collection locations. Genome analysis reveals four distinct sets with no admixture between groups and shows they are reproductively isolated and that hybridisation is not occurring, which was also found in previous studies using SSR and nuclear genomes (Young, 2011; Zalapa, 2011; Zhang 2011). Genome size between 4x and 8x groups is easily distinguished by flow cytometry and some aneuploidy samples could exist in rare instances as well as one hexaploid at one location (Costich, 2010).

## **1.6 Species and Genotype Distribution Modeling**

---

Evaluating the environmental suitability for a given species or genotype is a complex process that requires replicated trials of many lines across environmental gradients. For much of history, predicting local suitability was predominantly based on the physiology of plants (Sanderson, 1999; Köppen, 1936). Plants from various environments typically have convergent physical traits. Agronomic characteristics, such as last frost, or first seasonal rains were also suitable techniques of early farmers to optimise crop yield as plants are sensitive to their own set of environmental cues that greatly affect germination, vegetative growth, flowering, and

senescence, and is practiced today in simulated crop modelling (Mathews, 2013). Calculating the environmental suitability of a species across broad geographic range requires multiple environmental measurements. Biologically relevant abiotic climate variables (temperature and precipitation) tend to be the most reliable data for predicting the suitability of a local climate for a given species.

Computer based models can predict local suitability using two input data sets: species observations in coordinates, and climate data. Using a variety of statistical methods, a computer model will rank each locations probability of being suitable as similar to the training locations where a species is found. The science of predicting species locations based on climate patterns is often called species distribution modelling (SDM) (Hijmans, 2005; Phillips, 2004; Phillips, 2008; Elith, 2011, Phillips 2005). SDMs use comparisons of environmental variables to build predictive models. Since species are subject to many different types of abiotic stress, that could even be critical at certain annual time points, the input variables are often based on monthly, seasonal, or annual time scales of both temperature and precipitation. There are many SDM choices to calculate species distribution and suitability across geography, but MaxEnt is the most common due to its lack of bias in creating predicted climate windows and projecting suitable regions onto geographic space compared to other SDM software.

#### Prediction of Invasive Species Ranges

Predicting novel ranges of introduced species is challenging because it is a false assumption that a species climate breadth is fixed to its native habitat (Peterson, 2003). This likely is not the case as a species (or its direct lineages) could be previously adapted to paleo-climates, and/or their range has shifted with the previous fluctuations in changing global temperatures and even hybridising with relic groups as seen in *A. thaliana* (Sharbel, 2000; Lee, 2017). Therefore, input points from positive observations in non-native locations will greatly improve the predictive power of a species distribution model in both native and non-native ranges, and will also more accurately predict the breadth of a species climate tolerance. One of the interesting aspects of invasion biology is that having more locations beyond a species native range, that also have neutral to positive fitness, helps determine the true climate tolerances of that species. There are many other model types for predicting invasiveness, but usually requires previous calculations of plant density in various climate gradients and other knowledge of the species, like biotic interactions/sensitivities to other species with known ranges, phenotypic plasticity measurements across gradients, abundance/density, etc. In this thesis no prior knowledge was known about species phenotypes across gradients in non-native habitats. However, one study did show that *Brachypodium hybridum* has more phenotypic plasticity across climate gradients than diploid *B. distachyon* (Manzaneda, 2015). Also, Chapter V does investigate similarity in climate between native and non-native collection sites.

Rapid adaptation in introduced species has been characterised in publication and that range models of invasiveness often underpredict the fundamental niche (all possible habitable geographic space regardless of presence) of a species. As previously discussed, an anthropocentric assumption about the native range representing the fundamental niche is likely false. However, a publication reviewed nine different plant species for adaptation in non-native ranges, of the studies highlighted, introduced species often had phenotypic changes (Clements, 2011). Those traits analysed include: leaf shape, number and size increased; seeds often became larger, changes in perennial or annual life strategy; possible hybridisation with other species; and some had increased climate tolerances because the realised niche was now more descriptive of range and climate limits. The actual genetic causes in these studies are not carefully examined and most predate modern genomic analysis, but their phenotype changes are still relevant. Some of the phenotype variation in these studies has been observed in *B. distachyon* where some lines have different flowering time, and variation in leaf traits (Vogel, 2009). It is possible that some of the non-native adaptation mentioned is from admixture of individuals from geographically isolated native regions, and that the introduced genotypes have outcrossed and created novel genotypes in the non-native ranges and should be investigated. After all, *A. thaliana* has multiple genotypes in non-native ranges (Platt, 2010). A separate study that did analyse both genetic association and phenotypic variation found that *Lithrum salicaria*, a common North American invasive, had adapted to flower sooner in shorter northern seasons than locations as far as 1,000km south (Coulatti, 2013). In the case of a self-fertile outbreeding invasive species like *L. salicaria*, it should be noted that if non-native adaptive phenotypes do arise, it could quickly spread to other individuals and increase the fundamental range of said species.

#### MaxEnt Statistics and Basic Function

MaxEnt uses Maximum Entropy Modeling concepts via machine learning to calculate the climate suitability of geographic space based on species observation data in digital geospatial coordinate format (Phillips, 2004). For environmental inputs the program uses geospatial matrix maps called raster layers, the most common are precipitation and temperature climate data. The second input format is geographic coordinates of species observations. The basic premise is to calculate the upper and lower bounds of climate variables based on observation points, then weight each variable's contribution within the model based on entropy. To do this, MaxEnt first calculates the covariates conditional density at observation sites, and the unconditional density (marginal density) of the study area. Once both the covariate conditional density and marginal density are calculated, prevalence data (site observations) are used to calculate a conditional probability of environmental suitability. A ratio is first estimated for density of covariates across the study area  $f_1(z)$  and the marginal density of covariates  $f(z)$ . A comparison of one place versus another  $f_1(z)/f(z)$  through permutation, and optionally averaging sets of permutations, is

the primary operation that creates the initial data, or “raw data” of MaxEnt (Elith, 2011). Once the initial data and values associated with training filtering algorithms are complete, MaxEnt runs filtering programs to generate the final model(s). For more information about the machine learning of MaxEnt see Phillips, 2005, and for more about MaxEnt statistics read Elith *et al.* 2011. For more on how MaxEnt uses covariate analysis see Ward, 2007.

#### *Equal Sensitivity and Specificity Thresholding of MaxEnt to call Binary Suitability/Unsuitability*

Setting a binary limit on suitability could be used for a variety of different reasons. A binary classifier can also be used to filter out regions with low probability while retaining the probability values for each pixel point on the map outputs. By in large the reason for calling a binary threshold is based on the question being asked. Nearly every model will predict more geographic area with lower suitability scores and high suitability regions will be less common. If a researcher wants to describe the within species climate breadth they may use a binary classifier to limit the effect of low predicted climates to describe climate windows or geographic space. In the event of planning a collection trip or survey, researchers may want to optimise their chance at collecting species and set a higher threshold. Therefore, setting the minimum suitability threshold lower will result in a wider assumed climate window and more diverse climate types will be accepted as suitable as well as larger geographic area. In this thesis, the objective was to find suitable area and climate breadth across Chapters IV and V so a binary classifier was used based on a standard method used in specificity and sensitivity studies.

Setting binary suitability/unsuitability of regions with MaxEnt requires making a calculated assumption about a model’s accuracy to find suitable locations. Commonly used in machine learning (decision trees and neural nets), as well as medical diagnostic accuracy studies, MaxEnt uses a sensitivity and specificity algorithm to calculate model performance. The MaxEnt prediction algorithm will classify data into a 2x2 table of two classes, positive and negative results, and false negative and false positive. (Phillips, 2004; Hajian-Tilaki, 2013). Normal distributions of both the positive and negative classifications are plotted by their respective predicted values; the overlap of each positive and negative class represents the rate of false classifications. However, MaxEnt presents the predicted negative and positive probability values from the model output as a curved line on an XY plot scaled zero to one on each axis, often called a receiver operator curve (ROC). An ROC is a two-dimensional XY ordinal plot, where the predicted p-values of the false positives are on the y-axis, and the predicted false negative p-values are on the x-axis. The further the predicted p-values are from each other in both classes, the farther the area under the ROC curve is from a random 0.5, indicating a well performed model. Since p-values range from zero to one, the maximum area under the curve (AUC) is one, a perfect classifier. In most Diagnostic Accuracy Studies it is common to choose a threshold above 0.5, however the MaxEnt classifier is often more accurate than a standard threshold, and it can output the ideal threshold for overlapping tails of the positive and negative

rates (Phillips, 2006; Elith, 2011). Thus, a much lower rate can be used and is directly related to the AUC probability. MaxEnt does not default to show the two normal distributions since the ROC and AUC describe the model performance and the ideal threshold for calling a binary classifier with false positive and false negative information.

#### Local and global Modelling

MaxEnt functions to find spatial trends based on observation data. When modeling suitable habitat to create species distributions across a species native range, it is important to frame the study area boundaries proximal to the observation points (Ficetola, 2007; Medley, 2010; Elith, 2011). The further the observation points are from the boundaries of the study area, the more likely a model will sample regions with diverse non-suitable climates in the model. Oversampling more climate variation in non-predicted regions overfills the predicted negative class and will augment the model and create biased environmental variable contributions (Elith, 2011; Warren, 2011). One way to overcoming a bias towards one set of variables over another is using a tool like Environmental Niche Modelling Tools, (ENMTools) (Warren, 2010). ENMTools can trim a model distribution based on the maximal and average dispersal distance from observation points, if known. Doing so will remove locations that are actually beyond the physical limits of the study species dispersal ability. In this thesis ENMTools was not used because the focus was finding potential suitable habitat per species and genotypes requiring global climate layers, the assumption being that if a species or genotype were to travel beyond its normal range, what locations have suitable climates. In the case of finding new suitable regions in the native Mediterranean ranges, study boundaries were drawn slightly larger than a previous study that used ENMTools, the goal was to find new native regions that could harbour *Brachypodium* species of interest (Lopez-Alvarez, 2015).

#### Data Suitability Output

When performing global models it is important to assess what climate variables are describing suitability scores and that predicted suitable habitats compare in some way to climate data at the species observation locations. MaxEnt creates a diagnostic html sheet that can be read locally by a typical Internet browser and summarises the model performance and predicted geographic regions. The most scientifically relevant output from MaxEnt are the geographic maps that describe the predicted suitability, calculated as probability, represented as an image (.png) and an ASCII raster layer that can be further analysed by other programs and software depending on the specific question being addressed. Equally important are the diagnostic outputs describing the model performance: the percent contribution of each variable; jackknifing statistical methods; and the sensitivity-specificity ROC plots about model performance (Phillips, 2004; Phillips, 2006; Veloz, 2009; Medley 2010; Elith, 2011).

### Brachypodium distachyon Complex Range Models

Only one definitive study exists that calculated the likely native ranges and overlap of each complex species across various geologically recent timescales (Lopez-Alvarez, 2015). In this same publication, the predicted ranges of *B. distachyon* and *B. stacei* (diploids) rarely overlap through most of calculable history. The predicted ranges of *B. hybridum* often overlaps with both *B. stacei* and *B. distachyon*. Interestingly there are few locations that were predicted for only diploids but not suitable for *B. hybridum*, and could be that the allotetraploid *B. hybridum* inherited most of the diploid ranges, but what climate variable combinations are amenable to the diploids and not the polyploid should be investigated. Finally, that study also found that *B. hybridum* not only inhabits many of the same regions as either diploid, but that it expanded its geographic range post polyploidisation to new regions from an expanded climate breadth. A previous study described the likely suitable locations for the whole species complex and their likely realised niche based on presence locations of all cytotypes: spanning much of Europe, Central Asia, Sub-Continental India, North Africa, and non-native locations of North America, southern Africa, parts of South America near Uruguay, and much of southern Australia (Garvin, 2008). The breadth of climate tolerance of each species was also examined in Chapter V.

### **1.7 Climate Tolerance, Breadth, and Analysis of Species and Genotypes**

---

The environment is one of the primary stresses on a species geographic distribution and likely even extends to ancestral groups and genotypes (Phillips, 2005; Wisz, 2008; Elith 2011; Warren 2011; Brown, 2016). Uncovering which environmental variables underpin selection on a species requires complex field experiments from many locations and is fraught with many challenges. Field studies can be expensive, exhaustive, and require strenuous on-site measurements, experimentation, and consistency to make analysis meaningful. Extreme weather events can also confound field trials requiring postponing experiments until the next season. On a local scale, field studies can be difficult, but are possible and frequent in landscape genomics and ecology, but global field studies are especially difficult. Multiple continent sampling efforts are required, often via collaboration with other research groups. While the many challenges of onsite field studies is a surmountable task, they require careful experimental design and consistency in methods practiced between all individuals. Like in species distribution modelling, the use of WorldClim abiotic variables simplifies studies spanning broad geography to calculate the limits of climate tolerance. The current best global climate datasets for both climate analysis and SDMs are available at worldclim.org and at resolutions as small as 1km (Hijmans, 2005).

### Climate association studies in Brachypodium species

Several studies have been published on the climate variation of the *Brachypodium distachyon* species complex. Variation between different *B. distachyon* cytotypes, now three species, were

associated with different climate patterns, polyploids had a larger geographic breadth across the mediterranean and commonly present in warmer regions of the Iberian peninsula than diploid (*B. distachyon*) (Manzaneda, 2012). That same study found grain size and phenotypic variation was greater in polyploid samples and associated with climate patterns across gradients of precipitation and soil moisture. One study calculated the environmental niche of *B. stacei* and *B. distachyon*, finding diploids were statistically distinct with little to some overlap: *B. distachyon* is more common in cooler regions, while *B. stacei* is present in warmer regions with less precipitation (Catalan, 2012; Lopez-Alvarez, 2015; Catalan, 2015). The range of the polyploid *B. hybridum* overlapped significantly with both diploid species ranges. Despite the overlap between the polyploid and the diploids, *B. hybridum* also had an extended range from the diploids and is consistent with other studies (Manzaneda, 2011; Lopez-Alvarez, 2015; Catalan, 2015). A study in Turkey found 15 possible climate associated loci by scanning 82 wild collected individuals across nine climate unique locations calculated by the Ecocrop function in the program DIVA-GIS on an east-west longitudinal gradient to capture both climate and geographic isolation in a sampling transect using Bd21 as a control as well as four inbred lines (Dell'Acqua, 2014). While this study was conceptually ideal, having diverse sampling regions with multiple samples per location, with a species like *B. distachyon*, more sample locations would have improved statistical power to detect climate-associated loci. As mentioned in a book chapter regarding *Brachypodium* species research, very little about life history strategies and variation in ecological variation is currently published (Des Marais, 2015).

#### About the BioClim Climate Variables

Nineteen biologically relevant global climate variables are commonly used in SDMs and climate association studies. They are composed of precipitation and temperature values at annual, seasonal, and monthly intervals, and are readily available for near all global locations at 1+ km square resolution. Data can be mined for each sample collection location in a variety of ways. Data can be extracted from geographic coordinates via software like QGIS, R, and Atlas of Living Australia website (ala.org.au) as well as others. Specific descriptions of each BioClim layer can be found at the website worldclim.org and in the appendix section of this thesis.

#### Associating Climate to Species and Genotype

One of the central assumptions of Landscape Genomics is that natural processes have already conducted the experiment; local environmental stress has already filtered individuals by natural selection and the association of adaptive phenotypes and their causative genetic loci are patterned across selective environmental gradients. The association of environmental variables to adaptive traits is clearly possible, but little is known about how weedy species interact with climate and environment beyond SDMs. Though it could be litigated about how much *A. thaliana* qualifies as a weedy species, its genotypes do have variation in geographic and climate



breadth. The North American *A. thaliana* “Heartland” haplogroup was collected 1,041 times having broad geographic range (Platt, 2010; Anastasio, 2011). And as discussed above, native *A. thaliana* studies have shown variation in genotype presence across geography with some lineages associated with specific climates and regions, and some alleles convey larger climate tolerance (Shen, 2014). One of the crucial parts of this thesis is discovering if a whole species is widely adapted, or if there is genotype variation in widespread dispersal and climate breadth. What makes *Brachypodium* species more ideal than *A. thaliana* as an invasion model is that native *A. thaliana* genotypes are rarely found more than 1km from another individual in the native range, where previous studies of native *B. hybridum* and *B. distachyon* can be detected across thousands of kilometers and likely is true for their introduced ranges (Platt, 2010; Anastasio, 2011; Dell’Acqua, 2014; Neji, 2015). The causes of *Brachypodium* complex species high selfing rates are their self-compatibility and a mostly cleistogamous flower that stifles outcrossing (Garvin, 2008). Although *A. thaliana* is self-fertile, it has a non-cleistogamous flower that attracts pollinators, which theoretically increases the outcrossing rates of native range individuals, where grasses like *Brachypodium* are typically wind pollinated.

#### The Origins of Climate Classification

Climate classes are a basic metric to categorise the climate of any geographic location and are commonly used in species climate tolerances or even quantify effects of climate change in local regions (Diaz, 2007; Rubel, 2010; Brugger, 2013). The classification of climates dates back to the ancient Greeks as five climate types, more recently described by De Candolle a French plant scientist in 1906 (De Candolle, 1906; Sanderson, 1999). Around the same time as De Candolle, the plant physiologist Wladimir Köppen was compiling the first scientific climate classification system starting around 1884 and later published in 1936 (Sanderson, 1999; Köppen, 1936). Köppen built off of both De Candolle’s work and Greek philosophers to create at that time the most accurate description of global climate variation. Köppen used five climate classes that were centered on the general physiological properties of the flora in any given habitat. Köppen’s first groups were: A- torrid zones, B- dry zones, C- temperate zones, D and E were varying levels of arctic or frigid zones, the snow zone, and the polar zone (Köppen, 1936; Kottek, 2010). Rudolf Geiger expanded Köppen’s classification method in collaboration with Köppen (Geiger, 1954). Later in 1966 and updated in 1980 the Trewartha Climate Classification system was developed to better describe the variation in equatorial climates as the previous Köppen-Geiger system was considered too broad in these zones (Peel, 2007). Climate classes are helpful basic descriptors to classify the climate type of a local area and the physiology of the local flora. However, there are better than 550,000 known plant species and 31 climate classes are not likely to accurately describe the windows of climate tolerance of all each, but serve as a basic classifier of a given region. Therefore, to better understand the interaction between climate and

species variation irrespective of geographic distance, it is advantageous to calculate climate classes specific to each species and its genotypes.

### Per Species Climate classification

The process of creating species-specific climate classes is largely unexplored. Climate classes are a broad approach to determine a 'climate type' based on groups of plant species in a local habitat. A classification system can be very descriptive of climate breadth when comparing multiple species or other branches of science like Climate Change. However, a single plant species can theoretically have a broad or narrow breadth in a climate classification system like Köppen-Geiger, occupying many or few classes. Thus the climate limits of a particular species may not easily be described by climate classification. The breadth of a species theoretically could be smaller than the window defining a single class it inhabits, or its own tolerance limits could fall in the middle of two classes and not occupy both classes completely. For a single species study, the climate limits must be measured to more accurately describe the climate variation across individuals and genotypes. The use of BioClim variables of each collection location can be clustered into groups and used to design species-specific climate classes that more accurately reflect the climate diversity within that species. Also, regions that are predicted as suitable may have the same climate type despite geographic distance even beyond equatorial boundaries where seasons occur at opposite times of the year. A specialised classification system can more accurately describe the variation of climate within a species or species group with overlap in climate preferences. Chapter V expounds on this concept of species-specific climate classification similarity across geography to see if climate diversity is associated with genetic hotspots in the local and non-local ranges.

## **1.8 Discussion and Questions, Hypothesis, and Aims for Each Chapter**

---

To summarise many of the discussed topics above, the use of *Brachypodium* model species offers a powerful system to study invasion biology being now globally distributed across six continents. Invasive species often have unique life strategies from typical plants, reproducing from vegetative propagation as well as self-pollination and outcrossing. Not only do invasives usually have high phenotypic plasticity, they are often polyploid species with poorly characterised genomes making their genomic analysis complex and without a reference genome. An advantage of using *Brachypodium* species is that they are model organisms with many reference genomes with variation in cytotype and ploidy. Therefore, GBS data should provide ample amounts of mappable reads and simplify genomic analysis by comparison to non-model species with little documentation about their genetics. *Brachypodium* species often disperse long-distances, and exhibit phenotypic plasticity across ecological gradients. Like invasive species, *Brachypodium* complex members often self-pollinate from a closed flower (grass

floret) and based on other studies can self propagate vegetatively. Lastly, *Brachypodium* species are models for the grasses, there is significant interest in obtaining diverse sample sets, and this work will greatly improve the number of genotyped accessions that can be made available to the greater research community.

The three members of the *Brachypodium distachyon* species complex are challenging to identify from overlapping phenotypes, but sequenced individuals should be identifiable by their proportion of mappable reads to specific genomes and is explored in chapter II. Thus, post species classification, individuals belonging to one species can be aligned against their reference genome to call variants and describe diversity. As described above in *A. thaliana*, the same lineages can be collected across broad geography, and since *Brachypodium* species often have closed flowers causing self-pollination that create near clonal groups and may be found many times in non-native regions across broad geographic distance.

A significant portion of invasion biology is predicting vulnerable geography. Using species distribution models (SDMs), the regions that are sensitive to invasion can be predicted. The use of SDMs to predict sensitive non-native regions can be challenging. However, using observation locations in non-native regions should improve a model's ability to predict sensitive geography. This is because non-native regions may provide climate tolerance information beyond the known climate windows of the native range, and as mentioned above, the native range is not always a reliable data set to describe climate windows. Post model output, the total surface area can be calculated for each species to determine what group has the most predicted suitable habitat in both native and non-native ranges. Also, the predicted suitable surface area can be compared between models of common genotypes to that of a whole species. The goal being to test if a common genotype is, or could have a larger amount of habitable geography than random, or even defining the geographic and climate limits of genotypes within a species.

The correlation of SNPs to variables such as geographic distance or climate is the emerging discipline of landscape genomics. However, *Brachypodium* species have unconventional characteristics that are common in invasive plants like non-obligate out-crosser from a cleistogamous flowers, long LD, long-distance seed dispersal, and high plasticity across climate gradients (Tyler, 2016; Wilson, Streich, & Murray, 2018). Also, the use of GBS sequencing techniques may limit detecting regions of the genome carrying adaptive loci due to the distance between markers being beyond linkage disequilibrium. Recombinant genotypes will also limit the ability to separate causal SNPs from the genomic background (Brachi, 2011). Since *Brachypodium* species often are selfing and will have high inbreeding rates, other types of issues could arise that break from traditional methods: genomic variation between populations

could be confounded due to high selfing rates; little to no isolation by distance from long-distance dispersal, and could cause a whole genotype's genome to test as significant.

The analysis of climate diversity between species and common genotypes should also reveal what groups have larger or smaller windows of climate tolerance. Köppen-Geiger is still a valid descriptor of Earth climate classes and aid in multi-species comparisons. However, a novel method will be used in this thesis Chapter V to quantify climate variation within the scope of a single species by creating a species-specific climate classification system rather than a broad system like Köppen-Geiger. Ideally, an invasive species will have a wide breadth of climate classes in a traditional classification system. Even then, the windows defining each climate class are arbitrary for just one species and some classes cover more surface area than others. Theoretically, a species could be classified as invasive in a single geographically broad climate class. Therefore in a single species analysis, the upper and lower bounds of a climate classification system should be set to the species limits, and the variation within tolerable climates be set to the detectable variation within suitable climate windows. Based on the presence of genotypes across climate classes should reveal what lineages have smaller or larger climate windows with more meaningful ranges.

This thesis uses samples from many hundreds of global locations from three closely related *Brachypodium* species of varying genomic structure and ploidy. The species themselves will be screened for genetic, geographic, and climate diversity. The common genotypes within species can then be examined to see if some lineages are more widespread due to an abundance of geography that falls within its climate windows, or a lineage has large climate windows and increases the amount of habitable geography it can occupy. Even if a species or common lineage has large breadths of habitable geography, they must disperse well to occupy this space. The definition of generalists would be: "Genotypes or genotype families (groups of closely related genotypes) that have broad range in geography and/or climate." A specialist would be defined as, "A genotype or genotype family that tests as having low diversity in geography or climate." Only then can the more and less invasive-like groups be identified: by the abundance of habitable geography, their dispersal ability, and their breadth of climate tolerance. And so the questions, hypothesis, and aims of this thesis can be framed and are listed below.

Chapter II: Germplasm Development, Species Identification, and Regional Assessment of Brachypodium Species

**Question:** *How can the cryptic species of the Brachypodium distachyon species complex best be classified using genomics?*

**Hypothesis:** *I hypothesize that each species of the Brachypodium distachyon species complex will be distinguishable by their proportion of uniquely mapping reads to either one diploid genome, or both B. stacei and B. distachyon reference genomes as a polyploid.*

**Aim:** *By using highly conservative read mapping thresholds to filter away non-unique reads, the proportion of uniquely mappable reads will indicate whether a sample is one of two possible diploids if reads land almost exclusively to one reference genome, or reads map to both genomes indicating the allotetraploid B. hybridum.*

Chapter III: Genetics Analysis of Brachypodium distachyon Complex Members: Species Identification and Diversity

**Questions:** *What is the genetic diversity in relation to geography of the Brachypodium distachyon species complex and do genotypes of any species trend more as high or low dispersers?*

**Hypothesis:** *I hypothesize that since polyploid species often have larger stature and wider distributions geographically, the polyploid complex member B. hybridum should be more globally distributed than diploid species. Furthermore, some lineages will be more dispersed than others.*

**Aim:** *Obtain DNA sequence from individuals of each species to determine the genetic diversity across geography and test to see if some genotypes are more abundant than others by being better at dispersal. Use pairwise genetic distance among accessions to cluster whole genome genotype groups.*

#### Chapter IV: Genomic Biogeography: Predicting Invasion Sensitive Geography to

##### Species and Genotype

**Question:** *What are the regions with suitable climate across native and non-native geography for each *Brachypodium distachyon* complex species and do certain species or whole genome genotypes have larger ranges than would occur by chance?*

**Hypothesis:** *I hypothesize that since *B. hybridum* is a polyploid, and that it is known to demonstrate more phenotypic plasticity across native climate gradients, it will have a larger range and predicted surface area than diploid complex members. Further, some common genotypes of each species will have larger ranges than random.*

**Aim:** *First, calculate the surface area of predicted suitable habitat of all samples of each species and common genotypes. Then compare the total surface area of native and non-native habitat to see what species and genotypes are more prevalent and have larger fundamentally suitable surface area. Permute sampling locations to generate null distributions for common whole genome genotypes and compare to actual distribution.*

#### Chapter V: Testing Climate Classes and Windows in the *Brachypodium distachyon* Complex

**Question:** *What are the climate tolerance limits and variation of *Brachypodium distachyon* species using comparable bioclimatic variables and can certain whole genome genotypes be classified as specialists or generalists?*

**Hypothesis:** *I hypothesize that since *B. hybridum* is a polyploid with larger predicted suitable surface area, it will have larger climate tolerance limits. In addition, some genotypes of *B. hybridum* will occur in more climate classes than chance and thus are climate generalists, while others will be specialists with restricted climate breath.*

**Aim:** *Calculate the climate limits of each species using comparable climate variables and the occurrence of genotypes of each species across geography and species-specific climate classes. Then test the presence of genotypes across climate classes to see if some have wider climate windows than others.*

## 1.9 Citation

---

ala.org.au. (2012). Atlas of Living Australia, False Brome, *Brachypodium distachyon* (L). P.Beauv. Occurrence Records Map. <https://biocache.ala.org.au/occurrences/cc8eee49-f150-4f19-97eb-28be69e93471>

Al-Rabab'ah, M. A., & Williams, C. G. (2002). Population dynamics of *Pinus taeda* L. based on nuclear microsatellites. *Forest Ecology and Management*, 163(1-3), 263-271.

Ågren, J., & Schemske, D. W. (2012). Reciprocal transplants demonstrate strong adaptive differentiation of the model organism *Arabidopsis thaliana* in its native range. *New Phytologist*, 194(4), 1112-1122.

Ågren, J., Oakley, C. G., McKay, J. K., Lovell, J. T., & Schemske, D. W. (2013). Genetic mapping of adaptation reveals fitness tradeoffs in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, 110(52), 21077-21082.

Alderman, S., Garvin, D. F., Pfender, W. F., & Figueroa, M. (2013). Infection of *Brachypodium distachyon* by Formae Speciales of *Puccinia graminis*: Early Infection Events and Host-Pathogen Incompatibility.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *nature*, 408(6814), 796.

Australian Bureau of Statistics 2012 Invasive Plant Statistics. (2012). Land And Biodiversity, Environment. 1301.0 - Year Book Australia, 2012. Release Date : 24/05/2012. <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1301.0~2012~Main%20Features~Land%20and%20biodiversity~278>

Baack, E. J., Whitney, K. D., & Rieseberg, L. H. (2005). Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in *Helianthus* homoploid hybrid species. *New Phytologist*, 167(2), 623-630.

Bakker, Erica G., Brooke Montgomery, Tracy Nguyen, Kathleen Eide, Jeff Chang, Todd C. Mockler, Aaron Liston, Eric W. Seabloom, and Elizabeth T. Borer. (2009). "Strong population structure characterizes weediness gene evolution in the invasive grass species *Brachypodium distachyon*." *Molecular Ecology* 18, no. 12: 2588-2601.

Banta, Joshua A., Ian M. Ehrenreich, Silvia Gerard, Lucy Chou, Amity Wilczek, Johanna Schmitt, Paula X. Kover, and Michael D. Purugganan. "Climate envelope modelling reveals intraspecific relationships among flowering phenology, niche breadth and potential range size in *Arabidopsis thaliana*." *Ecology Letters* 15, no. 8 (2012): 769-777.

Beaumont, L. J., Hughes, L., & Poulsen, M. (2005). Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. *Ecological modelling*, 186(2), 251-270.

Bennetzen, J.L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A.C., Estep, M., Feng, L., Vaughn, J.N., Grimwood, J. and Jenkins, J., (2012). Reference genome sequence of the model plant *Setaria*. *Nature biotechnology*, 30(6), p.555.

Blackman, B.K., Rasmussen, D.A., Strasburg, J.L., Raduski, A.R., Burke, J.M., Knapp, S.J., Michaels, S.D. and Rieseberg, L.H., 2011. Contributions of flowering time genes to sunflower domestication and improvement. *Genetics*, 187(1), pp.271-287.

- Bomblies, K., Yant, L., Laitinen, R. A., Kim, S. T., Hollister, J. D., Warthmann, N., ... & Weigel, D. (2010). Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genet*, *6*(3), e1000890.
- Bond, W. J., & Keeley, J. E. (2005). Fire as a global 'herbivore': the ecology and evolution of flammable ecosystems. *Trends in ecology & evolution*, *20*(7), 387-394.
- Brachi, B., Morris, G. P., & Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome biology*, *12*(10), 232.
- Bradbury, D., Smithson, A., & Krauss, S. L. (2013). Signatures of diversifying selection at EST-SSR loci and association with climate in natural *Eucalyptus* populations. *Molecular ecology*, *22*(20), 5112-5129.
- Brown, G. R., Gill, G. P., Kuntz, R. J., Langley, C. H., & Neale, D. B. (2004). Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(42), 15255-15260.
- Brown, T.B., Cheng, R., Sirault, X.R., Rungrat, T., Murray, K.D., Trtilek, M., Furbank, R.T., Badger, M., Pogson, B.J. and Borevitz, J.O., (2014). TraitCapture: genomic and environment modelling of plant phenomic data. *Current opinion in plant biology*, *18*, pp.73-79.
- Brown, C.J., O'connor, M.I., Poloczanska, E.S., Schoeman, D.S., Buckley, L.B., Burrows, M.T., Duarte, C.M., Halpern, B.S., Pandolfi, J.M., Parmesan, C. and Richardson, A.J., (2016). Ecological and methodological drivers of species' distribution and phenology responses to climate change. *Global change biology*, *22*(4), pp.1548-1560.
- Brutnell, T.P., Wang, L., Swartwood, K., Goldschmidt, A., Jackson, D., Zhu, X.G., Kellogg, E. and Van Eck, J., (2010). *Setaria viridis*: a model for C4 photosynthesis. *The Plant Cell*, *22*(8), pp.2537-2544.
- Brugger, K., & Rubel, F. (2013). Characterizing the species composition of European Culicoides vectors by means of the Köppen-Geiger climate classification. *Parasites & vectors*, *6*(1), 333.
- Burt, J. W., Muir, A. A., Piovia-Scott, J., Veblen, K. E., Chang, A. L., Grossman, J. D., & Weiskel, H. W. (2007). Preventing horticultural introductions of invasive plants: potential efficacy of voluntary initiatives. *Biological Invasions*, *9*(8), 909-923.
- Busby, J. (1991). BIOCLIM—a bioclimate analysis and prediction system. *Plant Protection Quarterly (Australia)*.
- Catalán, P., Kellogg, E. A., & Olmstead, R. G. (1997). Phylogeny of Poaceae Subfamily Pooideae Based on ChloroplastndhF Gene Sequences. *Molecular phylogenetics and evolution*, *8*(2), 150-166.
- Catalán, P., Müller, J., Hasterok, R., Jenkins, G., Mur, L. A., Langdon, T., López-Alvarez, D. (2012). Evolution and taxonomic split of the model grass *Brachypodium distachyon*. *Annals of Botany*, *109*(2), 385-405.
- Catalan, P., López-Álvarez, D., Díaz-Pérez, A., Sancho, R., & López-Herránz, M. L. (2015). Phylogeny and evolution of the genus *Brachypodium*. In *Genetics and genomics of Brachypodium* (pp. 9-38). Springer, Cham.
- Clausen, J. C., Keck, D. D., & Hiesey, W. M. (1940). *Effect of varied environments on western North American plants*. Carnegie Institution of Washington.
- Clements, D. R., & Ditommaso, A. (2011). Climate change and weed adaptation: can evolution of invasive plants lead to greater range expansion than forecasted?. *Weed Research*, *51*(3), 227-240.
- Chiang, George CK, Melanie Bartsch, Deepak Barua, Kazumi Nakabayashi, Marilyne Debieu, Ilkka Kronholm, Maarten Koornneef, Wim JJ Soppe, Kathleen Donohue, and Juliette de Meaux. (2011). "DOG1



expression is predicted by the seed-maturation environment and contributes to geographical variation in germination in *Arabidopsis thaliana*." *Molecular Ecology* 20, no. 16: 3336-3349.

Convertino, M., Muñoz-Carpena, R., Chu-Agor, M. L., Kiker, G. A., & Linkov, I. (2014). Untangling drivers of species distributions: Global sensitivity and uncertainty analyses of MaxEnt. *Environmental Modelling & Software*, 51, 296-309.

Costich, D. E., Friebe, B., Sheehan, M. J., Casler, M. D., & Buckler, E. S. (2010). Genome-size variation in switchgrass (*Panicum virgatum*): flow cytometry and cytology reveal rampant aneuploidy. *The Plant Genome*, 3(3), 130-141.

Colautti, R. I., & Barrett, S. C. (2013). Rapid adaptation to climate facilitates range expansion of an invasive plant. *Science*, 342(6156), 364-366.

Draper, J., Mur, L. A., Jenkins, G., Ghosh-Biswas, G. C., Bablak, P., Hasterok, R., & Routledge, A. P. (2001). *Brachypodium distachyon*. A new model system for functional genomics in grasses. *Plant physiology*, 127(4), 1539-1555.

DeC, R. (1906). The Classification of Climates: II. *Bulletin of the American Geographical Society*, 465-477.

Dell'Acqua, M., Zuccolo, A., Tuna, M., Gianfranceschi, L., & Pè, M. E. (2014). Targeting environmental adaptation in the monocot model *Brachypodium distachyon*: a multi-faceted approach. *BMC genomics*, 15(1), 1.

Des Marais, D. L., & Juenger, T. E. (2015). *Brachypodium* and the abiotic environment. In *Genetics and Genomics of Brachypodium* (pp. 291-311). Springer, Cham.

Diaz, J. R., Weatherhead, E. K., Knox, J. W., & Camacho, E. (2007). Climate change impacts on irrigation water requirements in the Guadalquivir river basin in Spain. *Regional Environmental Change*, 7(3), 149-159.

Doust, A. N., Mauro-Herrera, M., Hodge, J. G., & Stromski, J. (2017). The C4 model grass *Setaria* is a short day plant with secondary long day genetic regulation. *Frontiers in plant science*, 8, 1062.

Eckert<sup>1</sup>, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González-Martínez, S. C., & Neale, D. B. (2010). Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*, 185(3), 969-982.

Eckert<sup>2</sup>, A.J., Wegrzyn, J.L., Cumbie, W.P., Goldfarb, B., Huber, D.A., Tolstikov, V., Fiehn, O. and Neale, D.B., (2012). Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. *New Phytologist*, 193(4), pp.890-902.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, 17(1), 43-57.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS one*, 6(5), e19379.

Elton, C. S. (1958). The Ecology of Invasions by Plants and Animals. *Methuen, London*, 18.

Flowers, T. J. (2004). Improving crop salt tolerance. *Journal of Experimental botany*, 55(396), 307-319.

Ficetola, G. F., Thuiller, W., & Miaud, C. (2007). Prediction and validation of the potential global distribution of a problematic alien invasive species—the American bullfrog. *Diversity and Distributions*, 13(4), 476-485.

Fick, S.E. and R.J. Hijmans, 2017. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*.

- Filiz, E., Ozdemir, B. S., Budak, F., Vogel, J. P., Tuna, M., & Budak, H. (2009). Molecular, morphological, and cytological analysis of diverse *Brachypodium distachyon* inbred lines. *Genome*, *52*(10), 876-890.
- Fitter, A. H., & Hay, R. K. (2012). *Environmental physiology of plants*. Academic press.
- Hussey, B.M.J., Keighery, G.J., Dodd, J., Lloyd, S.G. & Cousens, R.D. (2007). *Western Weeds. A guide to the weeds of Western Australia*. 2nd Edition. The Plant Protection Society of Western Australia, Victoria Park.
- Fournier-Level, A., Korte, A., Cooper, M. D., Nordborg, M., Schmitt, J., & Wilczek, A. M. (2011). A map of local adaptation in *Arabidopsis thaliana*. *Science*, *334*(6052), 86-89.
- Garvin, D. F., Gu, Y. Q., Hasterok, R., Hazen, S. P., Jenkins, G., Mockler, T. C., ... & Vogel, J. P. (2008). Development of genetic and genomic research resources for, a new model system for grass crop research. *Crop Science*, *48*(Supplement\_1), S-69.
- Gould, J. H., Zhou, Y., Padmanabhan, V., Magallanes-Cedeno, M. E., & Newton, R. J. (2002). Transformation and regeneration of loblolly pine: shoot apex inoculation with *Agrobacterium*. *Molecular Breeding*, *10*(3), 131-141.
- Geiger, R., & Pohl, W. (1954). Eine neue Wandkarte der Klimagebiete der Erde nach W. Köppens Klassifikation (A New Wall Map of the Climatic Regions of the World According to W. Köppen's Classification). *Erdkunde*, 58-61.
- Grabowski, P. P., Morris, G. P., Casler, M. D., & Borevitz, J. O. (2014). Population genomic variation reveals roles of history, adaptation and ploidy in switchgrass. *Molecular ecology*, *23*(16), 4059-4073.
- Grattapaglia, D., & Bradshaw Jr, H. D. (1994). Nuclear DNA content of commercially important *Eucalyptus* species and hybrids. *Canadian Journal of Forest Research*, *24*(5), 1074-1078.
- Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, *4*(4), 336-344.
- Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, *4*(2), 627.
- Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C., Roux, F. and Bergelson, J., (2011). Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, *334*(6052), pp.83-86.
- Hands, P., & Drea, S. (2012). A comparative view of grain development in *Brachypodium distachyon*. *Journal of Cereal Science*, *56*(1), 2-8.
- Hasterok, R., Draper, J., & Jenkins, G. (2004). Laying the cytotoxic foundations of a new model grass, *Brachypodium distachyon* (L.) Beauv. *Chromosome Research*, *12*(4), 397-403.
- Hasterok, R., Marasek, A., Donnison, I.S., Armstead, I., Thomas, A., King, I.P., Wolny, E., Idziak, D., Draper, J. and Jenkins, G., (2006). Alignment of the genomes of *Brachypodium distachyon* and temperate cereals and grasses using bacterial artificial chromosome landing with fluorescence in situ hybridization. *Genetics*, *173*(1), pp.349-362.
- Heuertz, M., De Paoli, E., Källman, T., Larsson, H., Jurman, I., Morgante, M., Lascoux, M. and Gyllenstrand, N., 2006. Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics*, *174*(4), pp.2095-2105.
- Horton, M.W., Hancock, A.M., Huang, Y.S., Toomajian, C., Atwell, S., Auton, A., Mulyati, N.W., Platt, A., Sperone, F.G., Vilhjálmsson, B.J. and Nordborg, M., (2012). Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature genetics*, *44*(2), p.212.

- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, 25(15), 1965-1978.
- Huang, P., Feldman, M., Schroder, S., Bahri, B. A., Diao, X., Zhi, H., ... & Kellogg, E. A. (2014). Population genetics of *Setaria viridis*, a new model system. *Molecular ecology*, 23(20), 4912-4925.
- Hulme, P. E. (2009). Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, 46(1), 10-18.
- Hussey, B.M.J., Keighery, G.J., Dodd, J., Lloyd, S.G. & Cousens, R.D. (2007) *Western Weeds. A guide to the weeds of Western Australia*. 2nd Edition. The Plant Protection Society of Western Australia, Victoria Park.
- Idziak, Dominika, Alexander Betekhtin, Elzbieta Wolny, Karolina Lesniewska, Jonathan Wright, Melanie Febrer, Michael W. Bevan, Glyn Jenkins, and Robert Hasterok. (2011). "Painting the chromosomes of *Brachypodium*—current status and future prospects." *Chromosoma* 120, no. 5: 469-479.
- Jia, G., Huang, X., Zhi, H., Zhao, Y., Zhao, Q., Li, W., ... & Zhu, C. (2013). A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nature genetics*, 45(8), 957.
- Johnson, S. R., & Young, D. R. (1993). Factors contributing to the decline of *Pinus taeda* on a Virginia barrier island. *Bulletin of the Torrey Botanical Club*, 431-438.
- Kainer, D., Bush, D., Foley, W. J., & Külheim, C. (2017). Assessment of a non-destructive method to predict oil yield in *Eucalyptus polybractea* (blue mallee). *Industrial crops and products*, 102, 32-44.
- Kaul, S., Koo, H.L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L.J., Feldblyum, T., Nierman, W., Benito, M.I., Lin, X. and Town, C.D., (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *nature*, 408(6814), pp.796-815.
- Konarzewski, T. K., Murray, B. R., & Godfree, R. C. (2012). Rapid development of adaptive, climate-driven clinal variation in seed mass in the invasive annual Forb *Echium plantagineum* L. *PloS one*, 7(12), e49000.
- Krutovsky, K. V., & Neale, D. B. (2005). Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. *Genetics*, 171(4), 2029-2041.
- Lasky, J. R., Des Marais, D. L., McKAY, J. O. H. N., Richards, J. H., Juenger, T. E., & Keitt, T. H. (2012). Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate. *Molecular Ecology*, 21(22), 5512-5529.
- Lata, C., & Prasad, M. (2013). *Setaria* genome sequencing: an overview. *Journal of plant biochemistry and biotechnology*, 22(3), 257-260.
- Lee, C. R., & Mitchell-Olds, T. (2012). Environmental adaptation contributes to gene polymorphism across the *Arabidopsis thaliana* genome. *Molecular biology and evolution*, 29(12), 3721-3728.
- Lee, C.R., Svardal, H., Farlow, A., Exposito-Alonso, M., Ding, W., Novikova, P., Alonso-Blanco, C., Weigel, D. and Nordborg, M., (2017). On the post-glacial spread of human commensal *Arabidopsis thaliana*. *Nature communications*, 8, p.14458.
- López-Alvarez, D., Manzaneda, A. J., Rey, P. J., Giraldo, P., Benavente, E., Allainguillaume, J., ... & Ezrati, S. (2015). Environmental niche variation and evolutionary diversification of the *Brachypodium distachyon* grass complex species in their native circum-Mediterranean range. *American journal of botany*, 102(7), 1073-1088.

- Lovell, J. T., Grogan, K., Sharbel, T. F., & McKay, J. K. (2014). Mating system and environmental variation drive patterns of adaptation in *Boechera spatifolia* (Brassicaceae). *Molecular ecology*, *23*(18), 4486-4497.
- Landguth, E. L., & Balkenhol, N. (2012). Relative sensitivity of neutral versus adaptive genetic data for assessing population differentiation. *Conservation Genetics*, *13*(5), 1421-1426.
- Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S. and Costich, D.E., (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS genetics*, *9*(1), p.e1003215.
- Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in ecology & evolution*, *18*(4), 189-197.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, *27*(2 Part 1), 209-220.
- Martins, P. K., Ribeiro, A. P., da Cunha, B. A. D. B., Kobayashi, A. K., & Molinari, H. B. C. (2015). A simple and highly efficient Agrobacterium-mediated transformation protocol for *Setaria viridis*. *Biotechnology Reports*, *6*, 41-44.
- Matthews, R. B., Rivington, M., Muhammed, S., Newton, A. C., & Hallett, P. D. (2013). Adapting crops and cropping systems to future climates to ensure food security: The role of crop modelling. *Global Food Security*, *2*(1), 24-28.
- Mauro-Herrera, M., Wang, X., Barbier, H., Brutnell, T. P., Devos, K. M., & Doust, A. N. (2013). Genetic control and comparative genomic analysis of flowering time in *Setaria* (Poaceae). *G3: Genes, Genomes, Genetics*, *3*(2), 283-295.
- Manzaneda, A. J., Rey, P. J., Bastida, J. M., Weiss-Lehman, C., Raskin, E., & Mitchell-Olds, T. (2012). Environmental aridity is associated with cytotype segregation and polyploidy occurrence in *Brachypodium distachyon* (Poaceae). *New Phytologist*, *193*(3), 797-805.
- Manzaneda, A. J., Rey, P. J., Anderson, J. T., Raskin, E., Weiss-Lehman, C., & Mitchell-Olds, T. (2015). Natural variation, differentiation, and genetic trade-offs of ecophysiological traits in response to water limitation in *Brachypodium distachyon* and its descendent allotetraploid B. hybridum (Poaceae). *Evolution*, *69*(10), 2689-2704.
- Medley, K. A. (2010). Niche shifts during the global invasion of the Asian tiger mosquito, *Aedes albopictus* Skuse (Culicidae), revealed by reciprocal distribution models. *Global ecology and biogeography*, *19*(1), 122-133.
- Méndez-Vigo, B., Picó, F. X., Ramiro, M., Martínez-Zapater, J. M., & Alonso-Blanco, C. (2011). Altitudinal and climatic adaptation is mediated by flowering traits and FRI, FLC, and PHYC genes in *Arabidopsis*. *Plant physiology*, *157*(4), 1942-1955.
- Mitchell-Olds, T. (2001). *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. *Trends in Ecology & Evolution*, *16*(12), 693-700.
- Montesinos-Navarro, A., Wig, J., Xavier Pico, F., & Tonsor, S. J. (2011). *Arabidopsis thaliana* populations show clinal variation in a climatic gradient associated with altitude. *New Phytologist*, *189*(1), 282-294.
- Morgan, S. (2003). *Land Settlement in Early Tasmania. Creating an Antipodean England*. Selwun College, Cambridge, Cambridge University Press, Sydney.
- Morris, G. P., Grabowski, P. P., & Borevitz, J. O. (2011). Genomic diversity in switchgrass (*Panicum virgatum*): from the continental scale to a dune landscape. *Molecular Ecology*, *20*(23), 4938-4952.

- Muller, M.H., Delieux, F., Fernandez-Martinez, J.M., Garric, B., Lecomte, V., Anglade, G., Leflon, M., Motard, C. and Segura, R., (2009). Occurrence, distribution and distinctive morphological traits of weedy *Helianthus annuus* L. populations in Spain and France. *Genetic Resources and Crop Evolution*, 56(6), pp.869-877.
- Munns, R. (2002). Salinity, growth and phytohormones. In *Salinity: environment-plants-molecules* (pp. 271-290). Springer Netherlands.
- Munns, R., & Tester, M. (2008). Mechanisms of salinity tolerance. *Annu. Rev. Plant Biol.*, 59, 651-681.
- Murray, K. D., Webers, C., Ong, C. S., Borevitz, J., & Warthmann, N. (2017). kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLoS computational biology*, 13(9), e1005727.
- Mur, Luis AJ, Joel Allainguillaume, Pilar Catalán, Robert Hasterok, Glyn Jenkins, Karolina Lesniewska, Ianto Thomas, and John Vogel. (2011). "Exploiting the *Brachypodium* Tool Box in cereal and grass research." *New Phytologist* 191, no. 2: 334-347.
- Myburg, A., Grattapaglia, D., Tuskan, G., Jenkins, J., Schmutz, J., Mizrahi, E., Hefer, C., Pappas, G., Sterck, L., Van De Peer, Y. and Hayes, R., (2011). December. The Eucalyptus grandis Genome Project: Genome and transcriptome resources for comparative analysis of woody plant biology. In *BMC proceedings* (Vol. 5, No. 7, p. I20). BioMed Central.
- Myburg, A.A., Grattapaglia, D., Tuskan, G.A., Hellsten, U., Hayes, R.D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D. and Goodstein, D.M., (2014). The genome of Eucalyptus grandis. *Nature*, 510(7505), p.356.
- Myers, J.H., Simberloff, D., Kuris, A.M. & Carey, J.R. (2000). Eradication revisited: dealing with exotic species. *Trends in Ecology & Evolution*, 15, 316 – 320.
- Neji, Mohamed, Filippo Geuna, Wael Taamalli, Yosra Ibrahim, Remo Chiozzotto, Chedly Abdelly, and Mhemmed Gandour. (2015). "Assessment of genetic diversity and population structure of Tunisian populations of *Brachypodium hybridum* by SSR markers." *Flora-Morphology, Distribution, Functional Ecology of Plants* 216: 42-49.
- Nordborg, M., Borevitz, J. O., Bergelson, J., Berry, C. C., Chory, J., Hagenblad, J., ... & Stahl, E. A. (2002). The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature genetics*, 30(2), 190-193.
- Novembre, J., & Ramachandran, S. (2011). Perspectives on human population structure at the cusp of the sequencing era. *Annual review of genomics and human genetics*, 12, 245-274.
- Opanowicz, M., Vain, P., Draper, J., Parker, D., & Doonan, J. H. (2008). *Brachypodium distachyon*: making hay with a wild grass. *Trends in plant science*, 13(4), 172-177.
- Padovan, A., Keszei, A., Foley, W. J., & Külheim, C. (2013). Differences in gene expression within a striking phenotypic mosaic Eucalyptus tree that varies in susceptibility to herbivory. *BMC plant biology*, 13(1), 29.
- Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular ecology*, 21(12), 2991-3005.
- Parker, David, Manfred Beckmann, David P. Enot, David P. Overy, Zaira Caracuel Rios, Martin Gilbert, Nicholas Talbot, and John Draper. (2008). "Rice blast infection of *Brachypodium distachyon* as a model system to study dynamic host/pathogen interactions." *Nature Protocols* 3, no. 3: 435.
- Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Ågren, J., Bossdorf, O., Byers, D., Donohue, K. and Dunning, M., (2010). The scale of population structure in *Arabidopsis thaliana*. *PLoS genetics*, 6(2), p.e1000843.

- Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and earth system sciences discussions*, 4(2), 439-473.
- Peraldi, A., Beccari, G., Steed, A., & Nicholson, P. (2011). *Brachypodium distachyon*: a new pathosystem to study Fusarium head blight and other Fusarium diseases of wheat. *BMC Plant Biology*, 11(1), 100.
- Peterson, A. T., Papes, M., & Kluza, D. A. (2003). Predicting the potential invasive distributions of four alien plant species in North America. *Weed Science*, 51(6), 863-868.
- Pires, J., Estevinho, M. L., Feás, X., Cantalapiedra, J., & Iglesias, A. (2009). Pollen spectrum and physico-chemical attributes of heather (*Erica* sp.) honeys of north Portugal. *Journal of the Science of Food and Agriculture*, 89(11), 1862-1870.
- Prentis, P. J., Wilson, J. R., Dormontt, E. E., Richardson, D. M., & Lowe, A. J. (2008). Adaptive evolution in invasive species. *Trends in plant science*, 13(6), 288-294.
- Presotto, A., Poverene, M., & Cantamutto, M. (2014). Seed dormancy and hybridization effect of the invasive species, *Helianthus annuus*. *Annals of applied biology*, 164(3), 373-383.
- Kane, N.C., King, M.G., Barker, M.S., Raduski, A., Karrenberg, S., Yatabe, Y., Knapp, S.J. and Rieseberg, L.H., (2009). Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution*, 63(8), pp.2061-2075.
- Köppen, W. (1936). The geographical system of climate. *Berlin, Germany*.
- Krimi, Z., Raio, A., Petit, A., Nesme, X., & Dessaux, Y. (2006). *Eucalyptus occidentalis* plantlets are naturally infected by pathogenic *Agrobacterium tumefaciens*. *European journal of plant pathology*, 116(3), 237-246.
- PIR, (2013). History of Agriculture in South Australia. Cereals and Grains. Government of South Australia. Primary Industries and Regions. [http://pir.sa.gov.au/aghistorical/cereals\\_\\_and\\_\\_grains/wheat](http://pir.sa.gov.au/aghistorical/cereals__and__grains/wheat)
- Phillips, S. J., Dudík, M., & Schapire, R. E. (2004, July). A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning* (p. 83). ACM.
- Phillips, S. J. (2005). A brief tutorial on MaxEnt. *AT&T Research*.
- Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. *Ecography*, 31(2), 161-175.
- Pimentel, D., Lach, L., Zuniga, R., & Morrison, D. (2000). Environmental and economic costs of nonindigenous species in the United States. *BioScience*, 50(1), 53-65.
- Potts, B. M., & Wiltshire, R. J. (1997). Eucalypt genetics and genecology. *Eucalypt ecology: individuals to ecosystems*, 56-91.
- Ream, T. S., Woods, D. P., Schwartz, C. J., Sanabria, C. P., Mahoy, J. A., Walters, E. M., ... & Amasino, R. M. (2014). Interaction of photoperiod and vernalization determines flowering time of *Brachypodium distachyon*. *Plant physiology*, 164(2), 694-709.
- Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular ecology*, 24(17), 4348-4370.
- Richards, C. L., Rosas, U., Banta, J., Bhambhra, N., & Purugganan, M. D. (2012). Genome-wide patterns of *Arabidopsis* gene expression in nature. *PLoS Genet*, 8(4), e1002662.

- Rosenthal, D. M., Ramakrishnan, A. P., & Cruzan, M. B. (2008). Evidence for multiple sources of invasion and intraspecific hybridization in *Brachypodium sylvaticum* (Hudson) Beauv. in North America. *Molecular ecology*, *17*(21), 4657-4669.
- Rubel, F., & Kottek, M. (2010). Observed and projected climate shifts 1901–2100 depicted by world maps of the Köppen-Geiger climate classification. *Meteorologische Zeitschrift*, *19*(2), 135-141.
- Saha, D., Gowda, M. C., Arya, L., Verma, M., & Bansal, K. C. (2016). Genetic and genomic resources of small millets. *Critical Reviews in Plant Sciences*, *35*(1), 56-79.
- Sanderson, M. (1999). The classification of climates from Pythagoras to Köppen. *Bulletin of the American Meteorological Society*, *80*(4), 669-673.
- Savolainen, O., & Pyhäjärvi, T. (2007). Genomic diversity in forest trees. *Current opinion in plant biology*, *10*(2), 162-167.
- Scascitelli, M., Whitney, K. D., Randell, R. A., King, M., Buerkle, C. A., & Rieseberg, L. H. (2010). Genome scan of hybridizing sunflowers from Texas (*Helianthus annuus* and *H. debilis*) reveals asymmetric patterns of introgression and small islands of genomic differentiation. *Molecular Ecology*, *19*(3), 521-541.
- Schmidt, J. P., Springborn, M., & Drake, J. M. (2012). Bioeconomic forecasting of invasive species by ecological syndrome. *Ecosphere*, *3*(5), 1-19.
- Sharbel, T. F., Haubold, B., & Mitchell-Olds, T. (2000). Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Molecular Ecology*, *9*(12), 2109-2118.
- Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, *26*(10), 1135-1145.
- Shen, X., De Jonge, J., Forsberg, S. K., Pettersson, M. E., Sheng, Z., Hennig, L., & Carlborg, Ö. (2014). Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. *PLoS genetics*, *10*(12), e1004842.
- Shiposha, V., Catalán, P., Olonova, M., & Marques, I. (2016). Genetic structure and diversity of the selfing model grass *Brachypodium stacei* (Poaceae) in Western Mediterranean: out of the Iberian Peninsula and into the islands. *PeerJ*, *4*, e2407.
- Steane, D. A., Potts, B. M., McLean, E., Prober, S. M., Stock, W. D., Vaillancourt, R. E., & Byrne, M. (2014). Genome-wide scans detect adaptation to aridity in a widespread forest tree species. *Molecular Ecology*, *23*(10), 2500-2513.
- Storz, J. F. (2005). INVITED REVIEW: Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, *14*(3), 671-688.
- Takahara, T., Minamoto, T., & Doi, H. (2013). Using environmental DNA to estimate the distribution of an invasive fish species in ponds. *PLoS one*, *8*(2), e56584.
- Tester, M., & Davenport, R. (2003). Na<sup>+</sup> tolerance and Na<sup>+</sup> transport in higher plants. *Annals of botany*, *91*(5), 503-527.
- Tonsor, S. J., Alonso-Blanco, C., & Koornneef, M. (2005). Gene function beyond the single trait: natural variation, gene effects, and evolutionary ecology in *Arabidopsis thaliana*. *Plant, Cell & Environment*, *28*(1), 2-20.
- Tyler, L., Fangel, J. U., Fagerström, A. D., Steinwand, M. A., Raab, T. K., Willats, W. G., & Vogel, J. P. (2014). Selection and phenotypic characterization of a core collection of *Brachypodium distachyon* inbred lines. *BMC plant biology*, *14*(1), 25.

- Tyler, L., Lee, S. J., Young, N. D., Delulio, G. A., Benavente, E., Reagon, M., ... & Caicedo, A. L. (2016). Population structure in the model grass *Brachypodium distachyon* is highly correlated with flowering differences across broad geographic areas.
- USDA Forest Service. (2016). Region 8. Invasive Species, US Forest Service. United States Department of Agriculture. <http://www.fs.usda.gov/detail/r8/forest-grasslandhealth/invasivespecies/?cid=stelprdb5326137>
- Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, *36*(12), 2290-2299.
- Vogel, J. P., Tuna, M., Budak, H., Huo, N., Gu, Y. Q., & Steinwand, M. A. (2009). Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. *BMC plant biology*, *9*(1), 1.
- Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., Bevan, M. W., ... & Tice, H. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, *463*(7282), 763-768.
- Warren, D. L., Glor, R. E., & Turelli, M. (2010). ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography*, *33*(3), 607-611.
- Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in MaxEnt: the importance of model complexity and the performance of model selection criteria. *Ecological applications*, *21*(2), 335-342.
- Ward, D. F. (2007). Modelling the potential geographic distribution of invasive ant species in New Zealand. *Biological Invasions*, *9*(6), 723-735.
- Whitney, K. D., Randell, R. A., & Rieseberg, L. H. (2010). Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. *New Phytologist*, *187*(1), 230-239.
- Willson, J. D., Dorcas, M. E., & Snow, R. W. (2011). Identifying plausible scenarios for the establishment of invasive Burmese pythons (*Python molurus*) in southern Florida. *Biological Invasions*, *13*(7), 1493-1504.
- Wilson, P., Streich, J., & Borevitz, J. (2015). Genomic Diversity and Climate Adaptation in *Brachypodium*. In *Genetics and Genomics of Brachypodium* (pp. 107-127). Springer International Publishing.
- Wilson, P.B., Streich, J.C., Murray, K.D., Eichten, S.R., Cheng, R., Aitken, N.C., Spokas, K., Warthmann, N. and Borevitz, J.O., (2018). Population structure of the *Brachypodium* species complex and genome wide association of agronomic traits in response to climate. *bioRxiv*, p.246074.
- Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., & Guisan, A. (2008). Effects of sample size on the performance of species distribution models. *Diversity and distributions*, *14*(5), 763-773.
- Woods, D. P., Ream, T. S., & Amasino, R. M. (2014). Memory of the vernalized state in plants including the model grass *Brachypodium distachyon*. *Frontiers in plant science*, *5*.
- Woods, D. P., Ream, T. S., Minevich, G., Hobert, O., & Amasino, R. M. (2014). PHYTOCHROME C is an essential light receptor for photoperiodic flowering in the temperate grass, *Brachypodium distachyon*. *Genetics*, *198*(1), 397-408.
- Yeoh, S. H., Bell, J. C., Foley, W. J., Wallis, I. R., & Moran, G. F. (2012). Estimating population boundaries using regional and local-scale spatial genetic structure: an example in *Eucalyptus globulus*. *Tree genetics & genomes*, *8*(4), 695-708.
- Young, H. A., Lanzatella, C. L., Sarath, G., & Tobias, C. M. (2011). Chloroplast genome variation in upland and lowland switchgrass. *PLoS one*, *6*(8), e23980.



Zalapa, J. E., Price, D. L., Kaepler, S. M., Tobias, C. M., Okada, M., & Casler, M. D. (2011). Hierarchical classification of switchgrass genotypes using SSR and chloroplast sequences: ecotypes, ploidies, gene pools, and cultivars. *Theoretical and Applied Genetics*, 122(4), 805-817.

Zhang, N., Xu, Y., Akash, M., McCouch, S., & Oard, J. H. (2005). Identification of candidate markers associated with agronomic traits in rice using discriminant analysis. *Theoretical and applied genetics*, 110(4), 721-729.

Zhang, Y., Zalapa, J.E., Jakubowski, A.R., Price, D.L., Acharya, A., Wei, Y., Brummer, E.C., Kaepler, S.M. and Casler, M.D., (2011). Post-glacial evolution of *Panicum virgatum*: centers of diversity and gene pools revealed by SSR markers and cpDNA sequences. *Genetica*, 139(7), p.933.

Zhao, K., Tung, C.W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., Norton, G.J., Islam, M.R., Reynolds, A., Mezey, J. and McClung, A.M., (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature communications*, 2, p.467.

## Chapter II: Germplasm Development, Species Identification, And Regional Assessment of *Brachypodium* Species

---

### Abstract

A germplasm represents seed collections sourced from many geographic locations that provide valuable genetic resources for breeding. I collected my own samples and formed collaborations with eight different research groups from six continents to build a *Brachypodium* germplasm collection. The collection is composed of 2,772 single seed descent lines as individual accessions combining pre-existing published accessions, private collections, and my personal collections. In addition, bulk accessions were also collected that could be further used to derive maternal lines. The true species identity of these accessions was largely unknown, and required molecular identification. A two-reference genome identification pipeline was created to rapidly classify individuals as one of the three species: *B. distachyon*, *B. stacei*, or *B. hybridum*. I concatenated the *B. stacei* and *B. distachyon* reference genomes into one *in-silico* polyploid reference. Then, genotyping by sequencing data for each individual barcoded accession was aligned to the polyploid reference. Species were identified by the proportion of reads mapping to either one or both diploid genomes. Of the 1,970 accessions sequenced, 1,719 samples received enough sequence coverage to be classified. A total of 528 *B. distachyon*; 50 *B. stacei*; and 1,147 *B. hybridum* samples were classified. Since each research group independently developed their own germplasm, integrating collections introduced discontinuity in: the number of locations, the defined location radius, samples per site, proximity to a road, and distance between collection locations. To group sites into species specific regions geographic coordinates of sample locations were clustered into major sub-continental groups. A total of 115 collection locations of *B. distachyon* were categorised into 20 regions across the Mediterranean and one outlier was found in Australia. Most of *B. stacei* was found in a single region of the eastern Mediterranean comprised of 23 close geographic coordinates. *B. hybridum* was widespread around the globe, found in 303 locations across 35 regions from six continents.

### Chapter Outline

---

#### 2.1 Introduction

#### 2.2 Methods

#### 2.3 Results

- Streich-Borevitz Germplasm
- Publically Available Germplasms
- Privately Available Germplasms
- Species Identification By Sequencing
- The Distribution of Species across Geography

## 2.4 Discussion

## 2.5 Data Sets and Script Links

## 2.6 Citation

## 2.1 Introduction

---

Collecting and maintaining very large germplasm collections is the role of dedicated stock centers, which employ numerous staff members, to focus on commercially important and/or rare biodiversity. Evaluating the phenotypic variation of entire collections is even more laborious and costly. Further, this is often confounded by differences within species complexes and overlapping geographic ranges. Random subsampling techniques can be used to analyse phenotypic variation in diversity studies (El Bouhssini, 2010). Alternatively, the Focused Identification of Germplasm Strategy (FIGS) screens collection locations by local climate and soil attributes to select suitable accessions for phenotypic screens (Khazaei, 2013). In an example study FIGS was used to analyse the climate diversity of *Vicia faba* to find locations with collected individuals of two different climate regimes across a large geographic space (Khazaei, 2013). The lines were then tested for physiological differences in water stress conditions and many lines were found to have drought resistant traits ideal for breeding water stress tolerant crops. However, to interbreed adaptive traits from the wild, germplasm species must be compatible and accessions within a species complex should be categorised.

A powerful way to screen germplasm collections is to first determine distinct genetic lineages and then select a balanced genetic diversity core set of lines for phenotyping (Brachi, 2011). Different accessions may not be different genotypes when they have nearly identical genome sequences, as is typical in self reproducing species. By DNA sequencing first, largely genetically redundant accessions are removed from the experimental design leaving a reduced set of roughly equidistant and genetically distinct samples that retain the natural diversity. An example being 5,707 accessions of *Arabidopsis thaliana* that were reduced to 1,799 diverse genotypes based on 149 markers (Platt, 2010). Genome -Wide Association Studies (GWAS) were optimised by reducing germplasm to a core set of diverse lineages. More obviously, DNA sequencing can first be used to find genetic anomalies that may indicate a collected specimen is actually a cryptic species within a species complex.

Initial studies of *Brachypodium distachyon* used few accessions, mainly Bd1-1, Bd2-3, Bd3-1, Bd18-1, Bd21, Bd21-3, and Bd29 (Mur, 2011). This lack of diverse study material led to the development of more collections, which coincided with the increased popularity of natural diversity studies. From this, the number of common accessions of *Brachypodium* species has increased dramatically over the last decade but species identity was largely unknown. Notable

contributions include a study in 2009 where 165 diverse lines from 45 locations were examined for genetic diversity (Vogel, 2009). A second study with some overlapping sample material further increased available accessions to 195 lines (Feliz, 2009; Mur, 2011). Seven ABR lines were made available via the Catalan and Stace groups, and subsequently 44 new locations were bulk collected, subsampled, and scanned for genetic diversity (Mur, 2011). By the start of the 2010's several hundred accessions of *B. distachyon* were available through public and private germplasm collections. *B. distachyon* became solidified as the grass model organism having numerous accessions coupled with the 2010 publication featuring a complete reference genome (Vogel, 2010). An initial core diversity set of 46 genotypes was created from 166 accessions (Tyler, 2014). Selecting diverse genetic material can improve the chances of mapping causative loci for a desired phenotype. It can also reduce the number of samples needed for successful mapping studies as seen in *Arabidopsis thaliana*, *Oryza sativa*, and *Zea mize* (Garris, 2005; Jin, 2010; Atwell, 2010; Brachi, 2011; Gross, 2014; Zhang, 2016).

Aside from GWAS in wild collections, genetic analysis of cultivated material can also help distinguish what groups were selected for breeding in agriculture as in *Setaria* and rice (*Oryza*) species (Zhao, 2011; Huang, 2014). In *Setaria* the species *S. italica* has been domesticated in comparison to its wild ancestor *S. viridis*, but is sexually compatible and many individual lineages of *S. italica* are admixed between the two species (Huang, 2014). Likewise, many rice lineages are treated as different subspecies, but when bred and introgressed with other closely related species they make viable offspring (Takano-Kai, 2009; Kovach, 2009; Zhao, 2011). In these cases the different founder genomes can be identified within segregating populations using the popular program STRUCTURE that calculates the number of ancestral groups within a study set of individuals.

The central aim of this chapter is to dissect the *Brachypodium* species complex into molecularly identifiable lineages including two diploids and a hybrid allotetraploid. Bulk collected individuals were separated into maternal lines and assigned to species groups based on the proportions of reads mapping to either one or both diploid genomes.

**Question:** *How can the cryptic species of the Brachypodium distachyon species complex best be classified using genomics?*

**Hypothesis:** *I hypothesize that each species of the Brachypodium distachyon species complex will be distinguishable by their proportion of uniquely mapping reads to either one diploid genome, or both B. stacei and B. distachyon reference genomes as a polyploid.*

**Aim:** *By using highly conservative read mapping thresholds to filter away non-unique reads, the proportion of uniquely mappable reads will indicate whether a sample is one of two possible diploids if reads land almost exclusively to one reference genome, or reads map to both genomes indicating the allotetraploid *B. hybridum*.*

## 2.2 Methods

---

The organization of germplasm from multiple continents, researchers, and across species is no small task. Most of the collections shared with the Borevitz lab were already developed into inbred lines, many of which were already past many cycles of selfing to insure homozygosity. Recent or bulk collections require investigation, as species identification of *Brachypodium* can be difficult in non-controlled conditions (Catalan; 2012). Samples were grown and DNA sequenced to identify species. The germplasm was assembled from eight different research groups each with different collection methods, small or large distances between sites. Thus, sample locations were first separated by species and then assigned regional identities to aid analysis in later chapters.

### *Borevitz lab Collections in Australia, Europe, and North America*

All accessions that I collected for the Borevitz Lab were collected in bulk from field locations, then later sorted at Australian National University. Approximately eight single seed descent maternal lines were subsequently developed from each sample location. Each collection site is approximately 30 meters in radius. In Australia, whole plants were harvested from a minimum of three different randomly chosen sub-locations within each site. If the location had unique microclimate features, then that gradient was sampled across. Examples of within site variation include: gradients toward a riverbank or creek; variation in overhead tree cover; or a sloping hillside. If the landscape features were qualitatively distinct within 30 meters such that they could disrupt gene flow, a new collection point would be allocated even if within the site radius. Examples of observed qualitative landscape features that might stop gene flow would include: a cliff-side, opposite sides of a river, or samples found in disturbed habitats near natural habitats. Sampling across microclimate gradients, but also as diverse as possible within locations was done to maximise the possibility of collecting genetic diversity. It should be noted that certain genotypes may have higher abundance in one microhabitat over another within a diverse collection location though this was not formally tested.

Collection efforts in Australia span 83 different locations across the states of New South Wales, Victoria, Tasmania, and South Australia. Collection planning started by looking at herbarium records on the Atlas of Living Australia website (ala.org.au) to plan route and timing. During onsite sampling, whole plants were harvested on location in bulk. Later, when developing maternal lines at ANU glass houses, bulk collections of each location were sorted, taking seed from the most diverse phenotypes. The most commonly selected phenotype was height, where

the tallest and shortest individuals were found first, then six random plants were drawn and organised between by height. This typically left eight plants per location. In some locations, if more phenotypic variety was found up to nine or ten plants were used to create maternal lines. Variation in phenotypes includes: unusual branching patterns, number of leaves, or other unique physical traits in the maternal plant. All plant material selected for creating a maternal line required the presence of root, stem, leaf, and floral tissue, and mature seed. The only exception being the CFW location near Flinders Ranges in South Australia where samples were scarce and any viable mature seed found was collected. For selecting maternal lines all phenotypes were done by eye, acknowledging a sample could be the same genotype, but having different phenotypes from microclimate or developmental effects.

I collected plants in France, Spain, Portugal, and Italy. In 2013, I planned a trip in Europe using records from numerous herbariums across Western Europe totaling 35 different locations. Only seeds are allowed import into Australia for research purposes, vegetative material was not allowed. Before returning to Australia, I harvested all seed from whole plant tissue with many hundreds of seeds per location. When developing bulk collections, I chose seeds with morphological variation in seed traits by sight to develop maternal lines. The most distinguishing traits were seed size, colour, and awn length. At each collection location I chose eight of the most physically diverse seeds to start maternal lines. All European collected lines were created and grown at Australian National University, Canberra ACT Australia.

I collected at six different locations in North America in the State of California. Like Europe, importation of whole plant material to Australia was not permitted. Also like Europe, I chose seed to make maternal lines based on diverse seed morphology, usually variation in seed size, awn length, and lemma colour ranging from green to a red-like hue. Like European and Australian lines, I chose eight diverse seeds per collection location to create maternal lines.

#### Publicly Available Accessions

*USDA-ARS; Vogel lab, Joint Genome Institute, University of California at Berkeley:*

Many accessions were accessed through the United States Department of Agriculture and most have been cytotyped for ploidy, and species. The previously identified accessions were used as controls for identifying other accessions in our collective germplasm. I received 442 lines from 92 locations from the Vogel lab and USDA-ARS.

#### Privately Available Accessions

The Garvin Lab shared non-maternal descent bulk accessions of 18 previously developed locations typically called the Bd lines. The Catalan Lab shared 292 accessions from 11 locations mostly from the northeast regions of the Iberian Peninsula. All of the samples shared by the Catalan lab had been identified as *B. distachyon* except two samples of *B. stacei* and one *B.*

*hybridum*. The Mur Lab shared samples exclusively from the Iberian peninsula totaling 153 accessions from 32 locations. Introduced *Brachypodium distachyon* complex species samples were collected by the Bradford Lab of University of California Davis at 20 locations totaling 182 accessions, from the state of California, United States. The Hazen and Caicido labs provided six samples from five locations of *B. distachyon* from: Armenia, Greece, Italy, Russia, and Spain. The Greece and Italy samples are helpful since much of the central European regions are not well collected from. Also, having true *B. distachyon* from Russia and Armenia are helpful as geographic outliers since most samples come from Turkey and Iberia. Unfortunately, samples from Italy and Greece did not sequence well and were not included in this thesis. I received 375 accessions from 183 locations from the Ezrati Lab that cover most of the landscape of Israel and parts of Armenia, Lebanon, and Greece. The Budak lab shared a private collection of samples from four locations and 25 maternal lines. See table 2.1 for a breakdown of locations, accession quantity, and numbers of known and unknown species before starting this thesis.

Lab	Number of Locations	Number of Accessions	Number of Unidentified	<i>B. distachyon</i>	<i>B. stacei</i>	<i>B. hybridum</i>
Borevitz	124	640	640	---	---	---
Bradford	20	182	---	---	---	182
Budak	4	25	25	---	---	---
Garvin	18	bulk	bulk	---	---	---
Caicido	4	4	---	4	---	---
Ezrati	183	375	188	---	---	187
Catalan	11	292	---	292	1	1
Vogel	92	442	---	---	---	---
Mur	32	153	---	---	---	---
<b>Totals</b>	<b>488</b>	<b>2,113*</b>	<b>852</b>	<b>296</b>	<b>2</b>	<b>370</b>

**Table 2.1:** Known metadata about each research group germplasm prior to this thesis analysis. Bulk seed was provided by the Garvin Lab from the collection locations of the common *Bdn* lines available from the USDA-ARS and accounts for most of the discrepancy between 2,113 and 2,772. A total of 1,897 samples were sequenced with enough depth for some comparison.

#### *Library preparation for DNA Sequencing*

All samples were grown at Australian National University from seed. DNA was extracted using DNEasy 96-well kits. Post DNA extraction, concentration was quantified using QuBit 2.0 fluorometer on 10 randomly chosen wells and averaged. A Genotyping By Sequencing (GBS) analysis method was used in the library preparation using a PstI six-cutter restriction enzyme to digest DNA into fragments. Cleaved DNA was then ligated with identification oligo barcodes,

Illumina Y adapters, and PCR primers for PCR amplification. Post PCR amplification DNA fragment concentrations were assessed per well per plate via a Shimadzu MCE-202 Multina 96 well plate reader with the DNA-12000 chemistry, and a Perkin-Elmer GXII Assay Chip. After concentration assessment sample wells were pooled for gel-based fragment size selection and sequencing. DNA fragments were gel filtered to 100-300 base pairs. Prepared DNA was then sequenced on Illumina HiSeq2000, HiSeq2500, and NextSeq500 platforms with paired end format.

#### Post Sequencing Sample Demultiplexing

Raw sequence data was demultiplexed using the software AXE (<https://github.com/kdmurray91/axe>). Reads are cross-referenced with an index of forward and reverse barcodes allowing one mismatch that is not shared with another barcode, then binned to a fastq file for each individual with the individual's name. Reads with barcodes with two or more mismatches are excluded from analysis. Once reads are binned into fastq files per each individual their restriction sites are removed and are then run through a customised genotyping pipeline using the TASSEL 3.0 software.

#### An In-silico Polyploid Reference Genome for Species Identification

The *Brachypodium distachyon* species complex is composed of two diploids with unique genomes that hybridised to form the allotetraploid *B. hybridum*. Due to the volume of bulk collected and unidentified accessions, a rapid species identification method was developed. Their proportion of sub-genome specific reads that uniquely map to a concatenated in-silico reference genome classified unknown samples. The reference genome was made from combining the two diploid reference genomes. The *B. stacei* sub-genome reference genome accession ABR114 v1.0, and the *B. distachyon* subgenome is represented by the accession Bd1-1 v1.0 as chromosomes 11-15. A highly conservative threshold was set, and a read's BWA mapping quality was used as a second filter. Ultimately, each individual was classified by their proportion of reads mapping to either one of, or both reference genomes to assign species identity.

#### First Attempts to Classify Species

Typically to distinguish two species with the same or similar cytotype the program STRUCTURE could be used to distinguish species by having different ancestral histories and fixed markers between two groups as seen in Rice and *Setaria* species (Huang, 2014). Since this species complex is composed of two diploids and one allotetraploid a STRUCTURE based method would not likely work because each species will have varying ancestral histories and a previous study found multiple hybridizations in *B. hybridum* (Catalan, 2012). In addition, allotetraploids will have orthologous haplotypes and could confound variant calls against a single reference genome having unusual levels of pseudo-heterozygosity against when markers



are created by mapping reads against one genome. A clustering method was first attempted to distinguish individuals to each species using principal component analysis using only one genome, Bd21 v2.0. A principal component analysis (PCoA) successfully identified most samples to species classes, but on many occasions failed to separate known vernalization-requiring *B. distachyon* groups like BdTR8 lines, Tek lines, and Bd1-1 from *B. hybridum* polyploids (further investigated in Chapter III). Also, *B. stacei* clustered close to *B. hybridum* in PCoA and was not easily distinguishable. Ultimately PCoA was not used for species identification. However, these results instigated the use of the Bd1-1 reference genome instead of the Bd21 reference as the *B. distachyon*-like subgenome in the aforementioned *in-silico* polyploid reference genome. Theoretically if the *B. distachyon*-like subgenome in *B. hybridum* is a closer extant relative to Bd1-1 or other vernalization-requiring lines, then BWA is more likely to align subgenome specific reads in the allopolyploid with this genome with higher mapping quality scores. This should give *B. hybridum* samples a more even proportion of reads to the expected values of 47% *B. stacei* and 53% *B. distachyon* because of different size genomes between the two species ( $\approx 240\text{mb}$  and  $\approx 270\text{mb}$ ).

#### Alignment Pipeline Thresholds for Species Classification

The alignment process used the Tassel reference genome pipeline coupled with Burrows Wheeler Aligner (BWA) to filter out non-unique reads mapping to multiple loci (Bradburry, 2007; Shendure, 2008; Elshire, 2011). There were many positive control samples used in our study of each species *B. distachyon*, *B. stacei*, and *B. hybridum*. Eighteen individuals were found called incorrectly for species categorisation; they were flagged and removed from the analysis, but many were biological replicates (seed planted and DNA sequenced from the same maternal line seed packet). The only Tassel pipeline parameter that was changed is the minor allele frequency set to (0.001). Within the BWA parameters, a conservative miss-match rate was used, allowing only 1%, or 1 of 100 base mismatch rate. Detected variants that are present in  $\geq 50\%$  of samples are used, thus  $< 50\%$  shared variants are removed. By removing non-unique loci, remaining loci are sub-genome specific and used for species identification. The proportion of reads in either or both subgenomes distinguished each individual for species.

The species identification pipeline outputs sequence data about variants of each individual across all chromosomes. Each sample's percentage of reads mapping to either genome was used to classify species. The summation of both genomes total markers post-filtered was set to a minimum of 10,000 to insure that remaining samples have enough resolution between subgenomes to decipher species and ploidy. To keep sample coverage from affecting species identification each individual's total number of reads from each subgenome is divided by read total to normalise the data set. The R package "kmeans" in the core stats package was used to call clusters based on the ratio of each genome per sample as shown in k-means clustering

algorithm (Hartig, 1979). The samples that are diploid will cluster closer to their respective species *K* having a majority of variants represented in one of the sub-genomes, while polyploids will have equal numbers of variants from both sub-genomes.

Thresholds were drawn to filter out questionable sample calls when calculating species. A few samples had intermediate or unusual sub-genome specific marker ratios and were flagged and removed. Intermediate samples were scanned individually for variation in markers per each chromosome in case some chromosomes were possibly deleted or missing, but no such trend was found. The final categories for species identification were: low-coverage samples, *B. hybridum*, *B. stacei*, *B. distachyon*, and intermediate. Once individuals are assigned a species identity, each species is organised geographically into regions for analysis.

#### Categorising Collection Locations into Regions

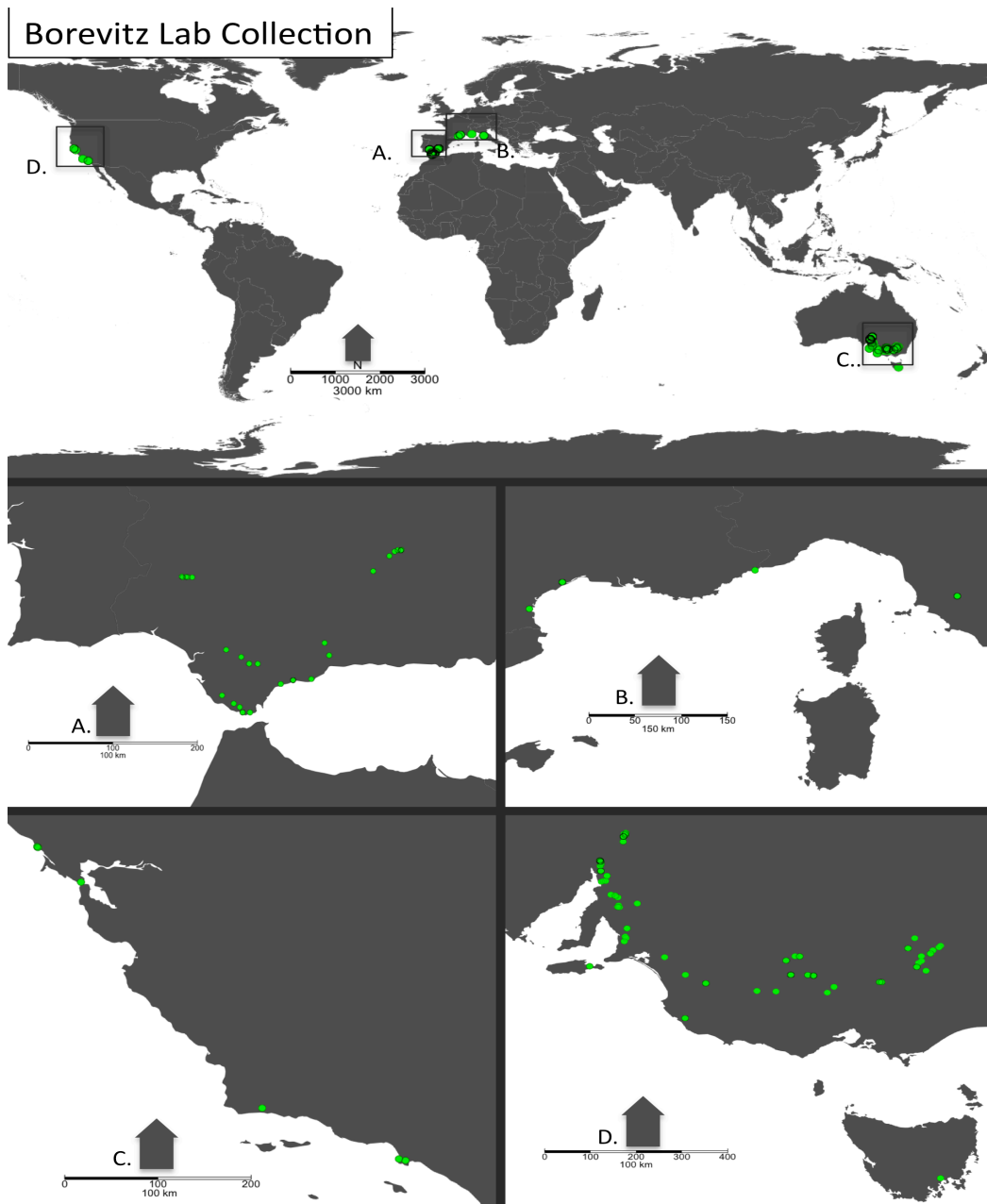
There is enormous variation in collection material techniques and practices from each participating collaborator. To normalize coverage variation across geography, collection locations were assigned region identities independently for each species. For each species analysis, locations were first parsed by continent, then locations were clustered by a typical clustering algorithm in the R package 'Mclust' using their latitude and longitude coordinates as inputs (Fraley, 2005). Clustering coordinates by their proximity to each other created groups of collection locations despite collector origin. These regions can be used to show how much genetic and climate diversity there is per region and will be necessary for future chapters.

### **Results 2.3**

---

#### Borevitz Lab, Development method of maternal line accessions

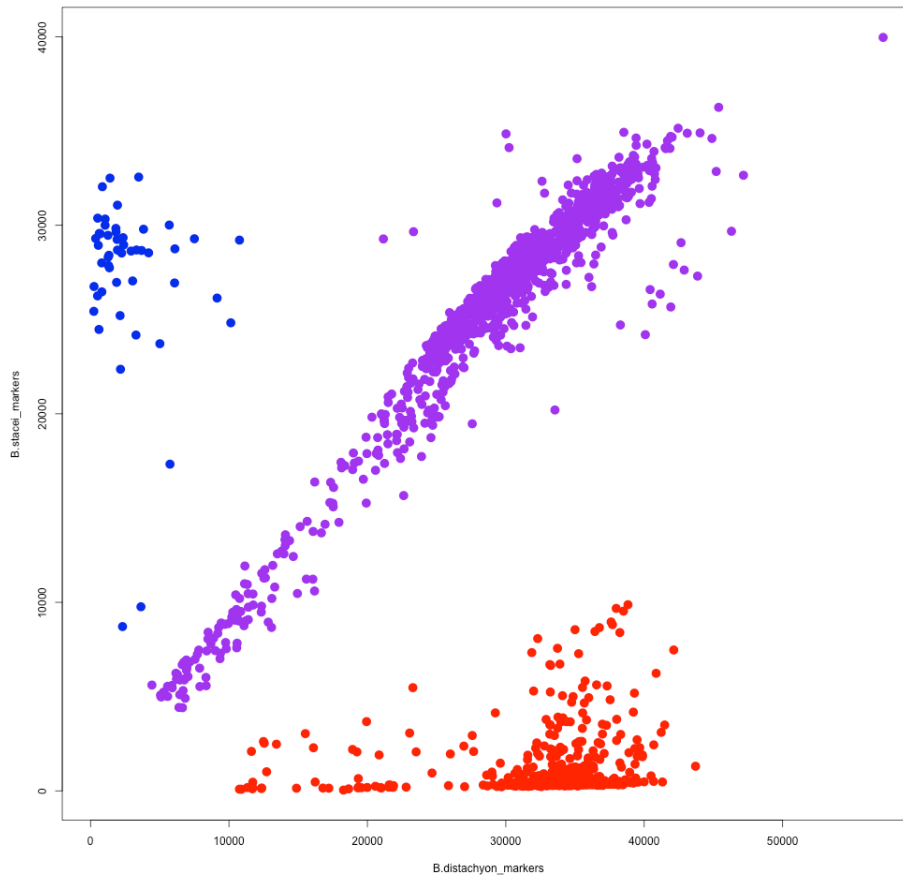
Within this project three different continents were sampled from. The Borevitz Lab, led by myself, collected extensively across SE Australia to capture as much diversity as possible. Since a large part of the collection effort was based on locations from the Atlas of Living Australia website (ala.org.au), most of the locations visited for sampling had species present. Also using species distribution models increased the sampling efforts by the Borevitz Lab. Collection efforts in Europe were not as successful for total yield of sample locations. Approximately 40% of locations examined had *Brachypodium* species present.



**Figure 2.2:** The Borevitz global collection of *Brachypodium*: (A) 30 sites were collected from in Southern Iberia; (B) five in modern day France and Italy; (C) six on North American in the state of California; (D) and most collection locations are in Australia totaling 83 sites.

### Species identification by Sequencing

The alignment of 1,897 samples against the *in-silico* *B. hybridum* reference genome was able to classify previously unknown samples as one of three species: *B. stacei*, *B. distachyon*, and *B. hybridum*, or other types of flagged samples. A total of 65,711 variants were found in the *B. distachyon* sub-genome, and 56,918 variants in the *B. stacei* sub-genome, which is proportional to genome size per each species, 266mb and 240mb respectively. Three distinct groups emerge by calculating the proportion of subgenome specific variants, plotted in X and Y coordinates (Figure 2.3). Fifty *B. stacei* samples aligned along the Y-axis in blue. *B. distachyon* trails along the X-axis plotted in red. Allotetraploid *B. hybridum* has both *B. stacei*-like and *B. distachyon*-like subgenomes and had near equal proportions of subgenome variants.



**Figure 2.3:** Samples plotted by most likely species and normalised by total calls. Samples plotted by total markers in each sub-genome and coloured by their candidate species identity. *B. stacei* is plotted as blue found near the Y-axis and near the X-axis in red is *B. distachyon*. *B. hybridum* has proportionally equal parts of reads mapping to each genome and is coloured purple. The diploid reference genomes are not equal in total base pair, *B. stacei* is  $\approx 240\text{Mb}$  and *B. distachyon* is 266 mb and each sub-genome was normalized accordingly.

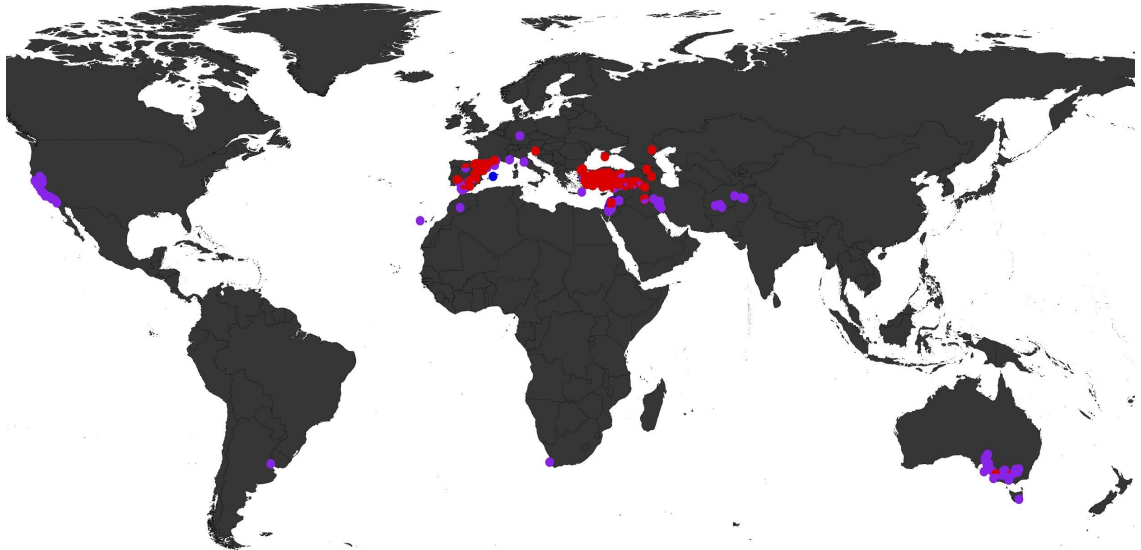
**Detected Species Identification Errors in Samples; Flagged and Removed**

Fourteen known *B. hybridum* were classified as *B. distachyon* and were flagged and removed from analysis based on previous publications (Vogel, 2009; Feliz, 2009; Mur, 2011; Tyler, 2016). One sample of *B. stacei* was classified as *B. hybridum* and was removed from analysis also based on previous publication (Hasterok, 2008; Vogel, 2009). Ninety-seven samples were between species classifications and 71 were low coverage samples with less than 10k variants from both subgenomes. A cumulative of 202 samples were flagged and removed from any one of these above-mentioned causes.

In study	<i>B. dis</i>	<i>B. sta</i>	<i>B. hyb</i>	Low Cov.	Between Species	Wrong Species	Flagged and Removed
1,970	528	60	1,180	71	116	15	202

**Table. 2.4:** Table of final counts of each category from species analysis. A total of 528 *B. distachyon* were identified, 50 *B. stacei*, and 1,180 *B. hybridum*. 165 samples received low sequence coverage to be deciphered. A total of 116 individuals fell between species and 15 classified as the wrong species. 202 samples were flagged and removed. *B. dis* = *B. distachyon*, *B. sta* = *B. stacei*, and *B. hyb* = *B. hybridum*.

### Plotting Identified species on maps



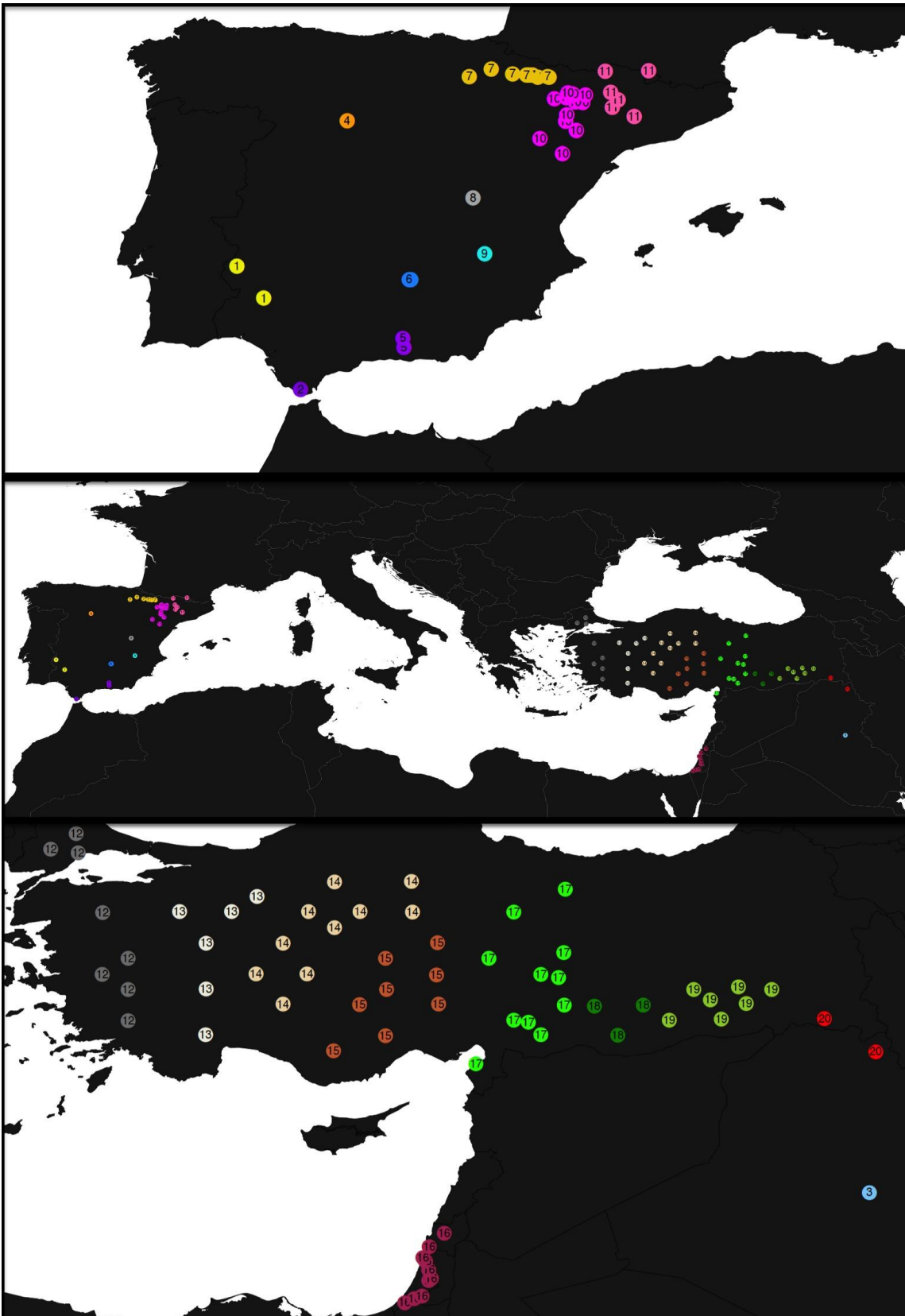
**Figure 2.5:** Locations by species pie charts. Individuals plotted are coloured by their candidate species identity. *B. stacei* in blue, *B. distachyon* in red, and *B. hybridum* in purple. See supplemental figures S2.8-S2.10 in the appendix for more pictures of pies on maps from species identification.

### Species Classification across Geography

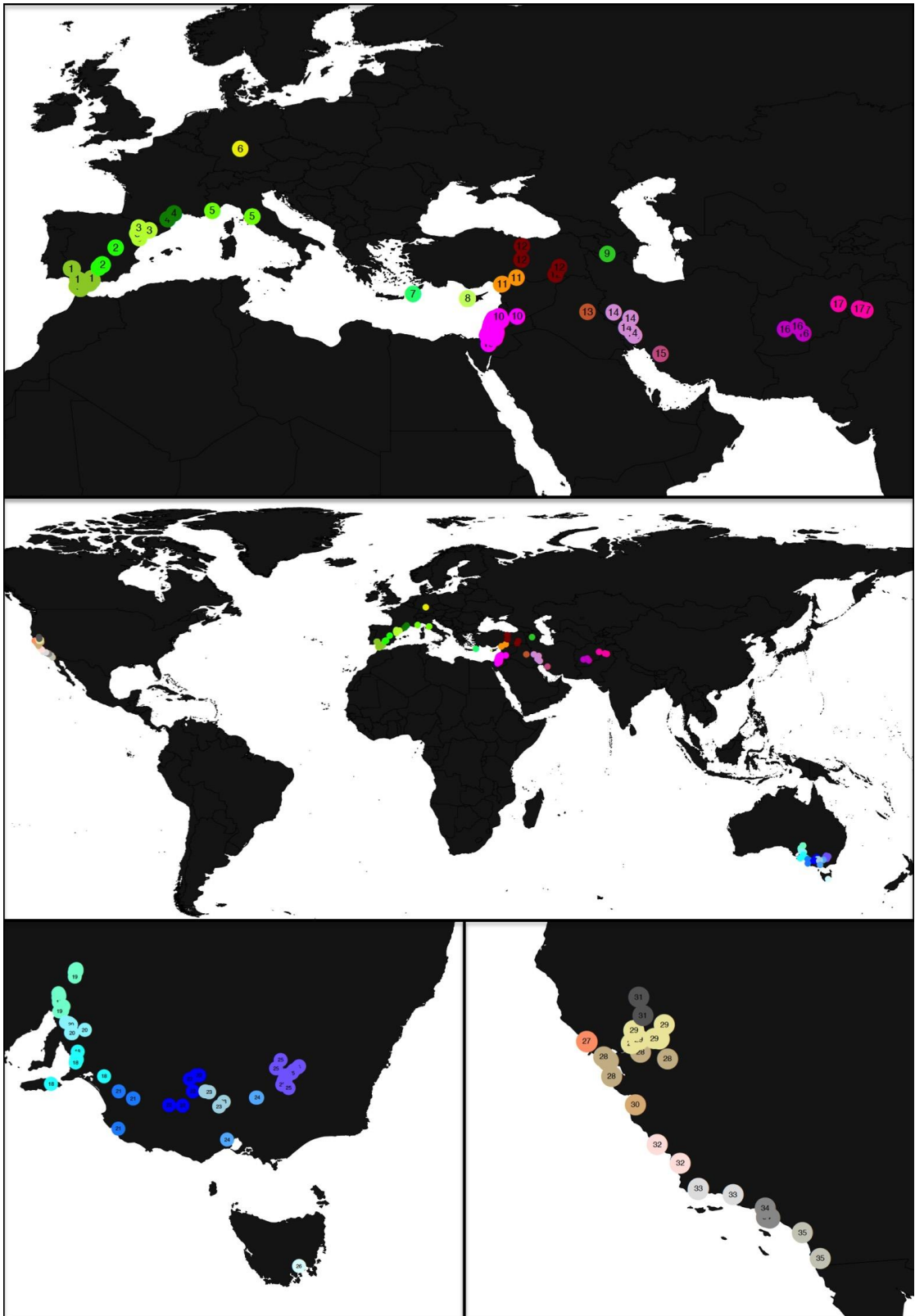
Species were plotted as pie charts on maps to show overlap in species presence in collection locations. The 115 *B. distachyon* sites were common in northern Mediterranean regions, and one location in Australia. The 23 *B. stacei* locations are mostly in the Southeast Mediterranean with two locations in western sites. *B. hybridum* was found on all mediterranean climate bearing continents being North and South America, Australia, Asia, Europe, and Africa. There were many locations in Iberian yet few in modern day Turkey that had both *B. hybridum* and *B. distachyon*. Many locations containing both *B. distachyon* and *B. hybridum* were found in the southeast Mediterranean, while most *B. stacei* locations were shared with *B. hybridum*. One southeast Mediterranean location had all three species.

### Assigning Regional Identities to Pooled Germplasms

*B. hybridum* and *B. distachyon* were found in multiple areas and were broken up into geographic regions. *B. distachyon* was found in 115 locations and classified into 20 regions (figure 2.6). This study uses source material from many research groups each with different sampling techniques, distances between geographic locations, overlapping ranges of individual seed collections, and number of samples collected per site. To normalise sampling across geography, collection sites were clustered into groups to overcome sampling bias and testing of presence of lineages in further chapters. A total of 313 *B. hybridum* global collection locations were partitioned into 35 regions (figure 2.8). Since *B. hybridum* was found on many continents with many locations, continental regions were assigned themed colour hues: Green-Yellow as Western Europe, Magenta-Red as Asia, Cyan-Blue as Australia, Grey/Brown as North America. As *B. stacei* appears rarely in the four geographic areas surveyed, it will largely be omitted from further analyses.



**Figure 2.6:** Assigned regions for *B. distachyon* samples. A total of 20 regions were designated for *B. distachyon*. Outlier samples like ABR2 (France), ABR9 (Croatia), and PYR2.6 and WLE2-2 (Australia) were removed and not assigned regional identities.



**Figure 2.7:** The regional identities of collection locations of *B. hybridum*. Locations have colour themes by the continent they were found in. A total of 35 regions were created with 9 locations in Europe (green), 8 region in Asia (red-magenta), 9 regions in Australia (blues), and North America with 9 regions (grey-brown).

## 2.4 Discussion

---

The rapid species identification by sequencing and a substantial addition of accessions increases the breadth of the *Brachypodium distachyon* species complex as model organisms. Typical studies in *Brachypodium* have less than 100-200 accessions, while this study encompasses 1,897 accessions that are species categorised with relatively high confidence. This is important, as most diversity studies require many hundreds of equidistant diverse individuals. Species are also visualised across geography and clear patterns are present in where each species grows and will be interrogated in further chapters, hence the break down of collections from collaborators into regions which is crucial for chapters three and five.

The assembled germplasm from each lab is the foundation that the rest of this thesis is built on. The questions in the following chapters of this thesis can only be answered with an organised set of accessions that are single seed maternal descent lines. The level of resolution is also subject to the methods and quality control of each research group, their own methods for selecting accessions and maternal lines, and the quality of work each individual practices. The sampling level is as accurate as each collector's ability to sample strategically for genetic diversity within each location. It is my opinion that each research group has taken diligent care in developing their research material and their shared germplasms are of high quality.

The sampling of material across landscapes is no easy task and requires skilled training in identifying not only species and their habitats, but also the possible variation within location habitats that a specific species is found. This is only achieved by working with others knowledgeable about a study species, its habits, life strategy, locations with suitable climate, but even other indicator species that are commonly found in the same climates. Finding *Brachypodium* species in the desert regions of South Australia where they are rare, often near river and streambeds and under trees, was often challenging. Then in the higher elevations near and north of Adelaide, samples are extremely abundant as well as in southwest New South Wales. Finding *Brachypodium* in wetter and more temperate Victoria is slightly less easy where samples are more sparse or growing near taller and abundant plants. Finding remnant individuals in Tasmania based on herbarium observations was extremely difficult. The microclimates within a collection site could also have a significant contribution to where and what species are found. To improve landscape genomic techniques and analysis, the sampling area should be well documented, photographed, or scanned depending on what equipment is available.

The identification of species by genomic markers was a relatively straightforward task and assigned species with high accuracy. The use of two genomes made species classification of



individuals by the proportion of reads to each subgenome simple and quick. These results also align with previous reports that *Brachypodium distachyon* is in fact composed of three different species (Catalan, 2012). Species identification by sequencing via two reference genomes is ideal for allotetraploids and their diploid counterparts and worked efficiently in this case. In practice, some errors were encountered, but many factors could cause species identification mishaps, the most likely root cause is human error. A sample could have had their tissue placed in the wrong plate well at sampling, a well could have had the wrong barcode added during library preparation, an early mutation in PCR amplification of DNA reads could cause false over-coverage of another sample with the same or similar barcode. Errors could also of happened in the growing/bulking stage: labeling errors on growing pots, miss-planted seeds in pots, or an alternate accession contaminating another accession seed packet.

Cryptic species identification by sequencing with common chromosome sets requires the detection and genotyping of many species-specific markers. In this case, a hard species boundary is not clear as divergent loci accumulate over time and in small populations. This situation is typical in other recently diverged species within Poaceae, Salicaceae, Asteraceae and many others that are known to share and spread haplotypes between species (Yatabe, 2007). Using programs like STRUCTURE to analyze genomic data can identify recently diverged species (Pritchard, 2000). For lesser known genre or species, STRUCTURE analysis where the ancestral population was set to the number of expected species theoretically could identify or categorise individuals to species and has been attempted before in *Brachypodium* (Huang, 2014; Tyler, 2016). However, one species could be an older lineage, under or over sampled, and thus biases in allele frequency could alter results. An ideal way to use STRUCTURE for species categorisation is to reduce the study set to a genetically distinct subset of individuals by removing genetically redundant accessions, set the ancestral population parameter to the expected number of species, then impute the ancestral composition onto closely related, near clonal or genotype level, samples. This should alleviate allele frequency bias and call species more accurately. It is likely however that a two-reference genome technique of mapping loci to either diploid genome would work for any combination of two species in this species complex.

Theoretically, different species have diverged over time and occupy different climates and habitats, but if these two species are similar in phenotypes they could be difficult to decipher, often called cryptic species (Bickford, 2007). Therefore, other plants classically defined as a single species, especially those introduced to new habitats, should be more thoroughly investigated as multiple cryptic species with unique and possibly overlapping ranges. In addition, one of the two cryptic species will be more common than other species. It is possible that some introduced species may indeed be different species, but genetically similar enough to occasionally hybridise and create a locally diverse gene pool. This would generate many unique

genotypes as seen in Asteraceae *Helianthus* species, *Eucalyptus* species in the Myrtaceae, and *Setaria* species as well (Potts, 2004; Field, 2011; Yatabe, 2007; 2008; Li, 2011). Any introduced species belonging to a plant family that readily hybridises and contains many invasive species should also be tested for species diversity, examples being Salicaceae, Asteraceae, and Poaceae plant families (Ainouche, 2004; Yatabe, 2007; Soltis, 2009; Wang, 2009; Hardig, 2010; Wang 2014). In this study case, what is now considered *B. distachyon* is extremely rare outside its native range, where *B. hybridum* (polyploid) was most common outside the native range, and was only recently described in 2012. Many more noxious pests could be multiple species not yet analysed for genetic or cytological diversity.

Since true *B. distachyon* was found in Australia, it is possible more locations contain diploids and should be examined. It could be possible that some regions of South Africa, South America, or North America could also harbour introduced *B. stacei* or *B. distachyon* that simply have not been found yet. A large emphasis was made in the Borevitz lab to sample widely in Australia, thus the resolution of this study may have made it possible to detect diploids outside the native range. Therefore, it would be ideal to look in more regions of non-native suitable space for other species.

Many different collection habits were used to assemble individual germplasms between collaborators, varying in sample coverage and distance between collection points. Collection locations were classified into regions despite their research group for each species to simplify and more normalise geographic analysis for later chapters (Chapters IV and V). When a study species becomes increasingly popular, like a model species, it would be ideal if more standard collection practices were instituted to facilitate thorough analysis. Standard practice methods are commonplace in many fields of biology, such as the medical field with tissue sampling, bacteria sampling, or blood sampling from patients (Bruneau, 2001; Gašová, 2005; Mager, 2007). Particularly to diversity studies, standard practice should include more than one sample be harvested per each collection location (Eckart, 2009; Eckart, 2010; Bragg, 2015).

While approximately two-thirds of the germplasm was identified by sequencing, many lines still require profiling. Regardless of plans to identify currently unknown samples, numerous accessions were sequenced to identify species and could be mapped geographically. The actual distribution of our study species became increasingly known during the course of this project and sampling effects, biases, trends, and recommended regions to collect more samples from will be acknowledged and discussed in the coming chapters. Finally, the hypothesis that species easily split into groups by their proportion of reads to one or both genomes is accepted by classification.

## 2.5 Data Download and Scripts

### GitHub Repository

A source code repository was created for the scripts and data sets in this chapter. The script for assessing species category is listed in the online repository GBSFilter at the Borevitz GitHub webpage. Each script will require their directory edited to read files. Raw figures are generated from these scripts and edited in various software for enhanced visualisation, power point preview, etc. Raw data can also be found in this same repository and specifically in links provided below.

### Repository

- <https://github.com/borevitzlab/GBSFilter>.

### Species ID Script

- [https://github.com/borevitzlab/GBSFilter/blob/master/Brachypodium\\_Species\\_ID.R](https://github.com/borevitzlab/GBSFilter/blob/master/Brachypodium_Species_ID.R)

### Genetic Marker Data

- B. distachyon subgenome:

[https://github.com/borevitzlab/GBSFilter/blob/master/myGBSGenos\\_D.txt.gz](https://github.com/borevitzlab/GBSFilter/blob/master/myGBSGenos_D.txt.gz)

- B. stacei subgenome: [https://github.com/borevitzlab/GBSFilter/blob/master/myGBSGenos\\_S.txt.gz](https://github.com/borevitzlab/GBSFilter/blob/master/myGBSGenos_S.txt.gz)

## 2.6 Citation

---

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T. and Jiang, R. (2010). Genome-wide association study of 107 phenotypes in a common set of Arabidopsis thaliana inbred lines. *Nature*, 465(7298), 627.

Bruneau, C., Perez, P., Chassaing, M., Allouch, P., Audurier, A., Gulian, C., Janus, G., Boulard, G., De Micco, P., Salmi, L.R. and Noel, L., (2001). Efficacy of a new collection procedure for preventing bacterial contamination of whole-blood donations. *Transfusion*, 41(1), pp.74-81.

Bickford, D., Lohman, D.J., Sodhi, N.S., Ng, P.K., Meier, R., Winker, K., Ingram, K.K. and Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in ecology & evolution*, 22(3), 148-155.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633-2635.

Bragg, J. G., Supple, M. A., Andrew, R. L., & Borevitz, J. O. (2015). Genomic variation across landscapes: insights and applications. *New Phytologist*, 207(4), 953-967.

Brachi, B., Morris, G. P., & Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome biology*, 12(10), 232.

Brkljacic, J., Grotewold, E., Scholl, R., Mockler, T., Garvin, D.F., Vain, P., Brutnell, T., Sibout, R., Bevan, M., Budak, H. and Caicedo, A.L., (2011). Brachypodium as a model for the grasses: today and the future. *Plant Physiology*, pp.pp-111.

Catalán, Pilar, Jochen Müller, Robert Hasterok, Glyn Jenkins, Luis AJ Mur, Tim Langdon, Alexander Betekhtin, Dorota Siwinska, Manuel Pimentel, and Diana López-Alvarez. (2012). Evolution and taxonomic split of the model grass Brachypodium distachyon. *Annals of Botany*, 109(2), 385-405.

Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, St Clair JB, Neale DB. (2009). Association genetics of coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. *Genetics* **182**: 1289–1302.

- Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González-Martínez SC, Neale DB. (2010). Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics* **185**: 969–982.
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS one*, 6(5), e19379.
- Filiz, E., Ozdemir, B. S., Budak, F., Vogel, J. P., Tuna, M., & Budak, H. (2009). Molecular, morphological, and cytological analysis of diverse *Brachypodium distachyon* inbred lines. *Genome*, 52(10), 876-890.
- Field, D. L., Ayre, D. J., Whelan, R. J., & Young, A. G. (2011). Patterns of hybridization and asymmetrical gene flow in hybrid zones of the rare *Eucalyptus aggregata* and common *E. rubida*. *Heredity*, 106(5), 841.
- C. Fraley, A. E. Raftery and R. Wehrens (2005). Incremental model-based clustering for large datasets with small clusters. *Journal of Computational and Graphical Statistics* 14:1:18
- Garris, A. J., Tai, T. H., Coburn, J., Kresovich, S., & McCouch, S. (2005). Genetic structure and diversity in *Oryza sativa* L. *Genetics*, 169(3), 1631-1638.
- Gašová, Z., Marinov, I., Vodvářková, Š., Böhmová, M., & Bhuyian-Ludvíková, Z. (2005). PBPC collection techniques: standard versus large volume leukapheresis (LVL) in donors and in patients. *Transfusion and apheresis science*, 32(2), 167-176.
- Gordon, S.P., Priest, H., Des Marais, D.L., Schackwitz, W., Figueroa, M., Martin, J., Bragg, J.N., Tyler, L., Lee, C.R., Bryant, D. and Wang, W. (2014). Genome diversity in *Brachypodium distachyon*: deep sequencing of highly diverse inbred lines. *The Plant Journal*, 79(3), 361-374.
- Gross, B. L., & Zhao, Z. (2014). Archaeological and genetic insights into the origins of domesticated rice. *Proceedings of the National Academy of Sciences*, 111(17), 6190-6197.
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), 100-108.
- Huang, Pu, Maximilian Feldman, Stephan Schroder, Bochra A. Bahri, Xianmin Diao, Hui Zhi, Matt Estep, Ivan Baxter, Katrien M. Devos, and Elizabeth A. Kellogg. (2014). Population genetics of *Setaria viridis*, a new model system. *Molecular ecology*, 23(20), 4912-4925.
- JGI, (2017). *Brachypodium stacei* ABR114, Surveying natural diversity of the model grass *Brachypodium distachyon* (Proposal ID: 277), <https://genome.jgi.doe.gov/BrastaABR114/BrastaABR114.info.html>
- Jin, L., Lu, Y., Xiao, P., Sun, M., Corke, H., & Bao, J. (2010). Genetic diversity and population structure of a diverse set of rice germplasm for association mapping. *Theoretical and Applied Genetics*, 121(3), 475-487.
- Kovach, M. J., Calingacion, M. N., Fitzgerald, M. A., & McCouch, S. R. (2009). The origin and evolution of fragrance in rice (*Oryza sativa* L.). *Proceedings of the National Academy of Sciences*, 106(34), 14444-14449.
- Li, P., & Brutnell, T. P. (2011). *Setaria viridis* and *Setaria italica*, model genetic systems for the Panicoid grasses. *Journal of experimental botany*, 62(9), 3031-3037.
- Liang, S. U. N., Jie, C. H. E. N., Kai, X. I. A. O., & Wencai, Y. A. N. G. (2017). Origin of the Domesticated Horticultural Species and Molecular Bases of Fruit Shape and Size Changes during the Domestication, Taking Tomato as an Example. *Horticultural Plant Journal*, 3(3), 125-132.
- Mager, S. R., Monique HA Oomen, Manuel M. Morente, Cathy Ratcliffe, Kyle Knox, David J. Kerr, Francesco Pezzella, and Peter HJ Riegman. (2007). Standard operating procedure for the collection of fresh frozen tissue samples. *European Journal of Cancer*, 43(5), 828-834.

- Mur, Luis AJ, Joel Allainguillaume, Pilar Catalán, Robert Hasterok, Glyn Jenkins, Karolina Lesniewska, Ianto Thomas, and John Vogel. (2011). "Exploiting the Brachypodium Tool Box in cereal and grass research." *New Phytologist* 191, no. 2: 334-347.
- Pankin, A., & von Korff, M. (2017). Co-evolution of methods and thoughts in cereal domestication studies: a tale of barley (*Hordeum vulgare*). *Current opinion in plant biology*, 36, 15-21.
- Potts, B. M., & Dungey, H. S. (2004). Interspecific hybridization of *Eucalyptus*: key issues for breeders and geneticists. *New Forests*, 27(2), 115-138.
- Sauvage, Christopher, Andrea Rau, Charlotte Aichholz, Joël Chadoeuf, Gautier Sarah, Manuel Ruiz, Sylvain Santoni, Mathilde Causse, Jacques David, and Sylvain Glémin. (2017). Domestication rewired gene expression and nucleotide diversity patterns in tomato. *The Plant Journal*.
- Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10), 1135.
- Takano-Kai N, Jiang H, Kubo T, Sweeney M, Matsumoto T, Kanamori H, Padhukasahasram B, Bustamante C, Yoshimura A, Doi K, McCouch S. (2009). Evolutionary history of GS3, a gene conferring grain length in rice. *Genetics*, 182(4), 1323-1334.
- Tyler, Ludmila, Jonatan U. Fangel, Alexandra Dotson Fagerström, Michael A. Steinwand, Theodore K. Raab, William GT Willats, and John P. Vogel. (2014). "Selection and phenotypic characterization of a core collection of *Brachypodium distachyon* inbred lines." *BMC plant biology* 14, no. 1: 25.
- Vogel, J. P., Tuna, M., Budak, H., Huo, N., Gu, Y. Q., & Steinwand, M. A. (2009). Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. *BMC plant biology*, 9(1), 88.
- Yatabe, Y., Kane, N. C., Scotti-Saintagne, C., & Rieseberg, L. H. (2007). Rampant gene exchange across a strong reproductive barrier between the annual sunflowers, *Helianthus annuus* and *H. petiolaris*. *Genetics*, 175(4), 1883-1893.
- Zhang, Fantao, Tao Xu, Linyong Mao, Shuangyong Yan, Xiwen Chen, Zhenfeng Wu, Rui Chen, Xiangdong Luo, Jiankun Xie, and Shan Gao. (2016). Genome-wide analysis of Dongxiang wild rice (*Oryza rufipogon* Griff.) to investigate lost/acquired genes during rice domestication. *BMC plant biology*, 16(1), 103.
- Zhao, K., Wright, M., Kimball, J., Eizenga, G., McClung, A., Kovach, M., Tyagi, W., Ali, M.L., Tung, C.W., Reynolds, A. and Bustamante, C.D. (2010). Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS one*, 5(5), e10780.
- Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., ... & McClung, A. M. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature Communications*, 2, 467.

## Chapter III: Genetic Diversity within *Brachypodium* Species

### Abstract

Genetic diversity screens within species can unveil genomic patterns associated with historical migration across geographic regions, such as isolation by distance. *Brachypodium distachyon*, *B. stacei*, and *B. hybridum* were genotyped by sequencing and variants were identified by alignment to individual species reference genomes. For *B. hybridum* an *in-silico* reference genome was used by concatenating the *B. distachyon* (Bd1-1) and *B. stacei* (ABR114) reference genomes to call uniquely mapping subgenome specific variants. Within each species, each accession was compared against each other to determine pairwise genetic distance. Distinct genotypes were identified using a distance threshold set as the most divergent of technically replicated samples. A total of 125 unique genotypes of *B. distachyon* were calculated using 14,436 SNPs from 479 samples, eight genotypes of 50 *B. stacei* samples were calculated using 4,744 SNPs, and for the allotetraploid 80 genotypes were calculated out of 1,015 samples using 18,525 variants. The most common genotypes of *B. hybridum* and *B. distachyon* were mapped to show their geographic breadth including areas outside of the native range for *B. hybridum*. Non-native ranges were exclusively invaded by *B. hybridum* except one location in Australia having two *B. distachyon*. A permutation test of observed sites showed that the most common *B. hybridum* genotype NRD-1 was more widespread across regions than would be expected by chance (p-value < 0.01) suggesting it has wide dispersal ability.

### Chapter Outline

---

#### 3.1 Introduction

- Description of *Brachypodium distachyon* species complex

#### 3.2 Methods

- Laboratory Methods for DNA Extraction and Sequencing

- Computational Methods for Genotyping

  - Genotyping Pipeline Overview

  - Settings and Filtering Parameters

  - Tassel Reference Alignment Analysis

#### 3.3 Results

- Per Species Overall Genetic Diversity

- Nested Analysis amongst groups

- Re-analysing groups based on genetic distance

- Core Diversity Set and Genotypes

- In-Silico* Polyploid Donor Genome Detection

#### 3.4 Discussion

### 3.1 Introduction

---

The global human population is expected to breach nine billion before 2050 requiring an increase in global food production and agronomic efficiency (Tillman, 2011) to deliver both food and environmental security (Rivers, 2015). Grass species already provide the majority of human caloric intake via wheat, rice, and indirectly maize. Grasses also account for many of the noxious agricultural and natural invasive species pests (Daehler, 1998; Pyšek, 1998; United Nations, 2013; Sawar, 2013; ACP-FAO, 2016). The monocot genomes of wheat, oat, and barley can be challenging and laborious to work with due to larger stature, high ploidy levels, and complicated very large genomes. However, the closely related *Brachypodium distachyon* species complex has easily manageable genomes and a small stature making them ideal agricultural models. The *Brachypodium distachyon* species complex has been introduced globally and reported as weedy and invasive, thus are ideal models for introduced species (Bakker, 2009). The *Brachypodium distachyon* complex is a close relative to many agriculturally relevant grass species and as a noxious weed or invasive (Opanowicz, 2008; Bakker, 2009)

#### Invasion Biology Concepts

The term for the study of introduced species, their migration, and their environmental effects is commonly called Invasion Biology. The spread of invasive species is threatens global food security. Roughly 70% of invasive plants are from the global horticulture trade, leaving 30% to other means, often agricultural contaminants and tourism (Carlton, 2003). No matter their path to novel locations, introduced species can disrupt a variety of landscapes, as is the case in Australia where Patterson's Curse (*Eichium plantagineum* and *Eichium vulgare*) were introduced as horticultural species, and now disrupts many agricultural landscapes (Konarzewski, 2012). Weedy and invasive plants account for approximately four billion dollars in damage annually to agricultural and natural landscapes across Australia, and similar patterns are seen across the globe (Australian Bureau of Statistics, 2012). Another example of their effects, in the United States it is estimated that  $\approx 3$  million acres, 1.2 million hectares, of land are lost each year to invasive plants (USDA Forest Service, 2016). The control and spread of invasive species is thus of great concern, but how they migrate, how much genetic diversity they bring, and how often they are introduced requires resolution (Barrett, 1991; Allendorf, 2003; Simberloff, 2013).

Across native ranges it is assumed random mating occurs between individuals and that patterns of genetic diversity coincide with geographic distance and/or environmental gradients (Wright, 1934; Kimura and Maruyama, 1971; Rousset, 1997; Bragg, 2015). Theoretically, populations evolve neutrally, by both genetic isolation by distance and isolation by environment within the

natural native range. This will result in a level of population structure. Widespread species may thus contain either 1) many locally adapted genotypes that are partially isolated, or 2) have less population structure share adaptive alleles across native ranges, or 3) contain plastic and resilient genotypes with wide environmental tolerance. Introduced species in non-native habitats can have population structure when multiple introduction events have occurred and they remain genetically isolated. Genetic studies within *Brachypodium* species would elucidate mechanisms underlying invasion biology concepts of a widespread species by analysing its geographic and genetic patterns across native and non-native ranges.

#### About *Brachypodium* Model Species

Model plant species have greatly increased our understanding in all aspects of plant science research. Model organisms must have a manageable stature, be amenable to genetic transformation, and easily grow in laboratory or experimental settings. Ideally, a model has extensive collections such as the global germplasm collection of *Arabidopsis thaliana* composed of thousands of accessions. The plant *A. thaliana* has been an immeasurably valuable model, but is a dicot and is a poor model for monocot systems. *B. distachyon* has many of the same qualities as *A. thaliana*, but is a temperate C3 monocot grass, is amenable to crossing, highly inbreeding due to a cleistogamous flower, is readily transformable, and is closely related to many agriculturally important cereals with complex genomes: *Triticum species* (wheat), *Hordeum species* (barley), and *Avena species* (oat) (Catalan, 1997; Draper, 2001; Opanowicz, 2008; Vogel, 2008; Alves, 2009; Bragg, 2012; Mochida, 2013; Fitzgerald, 2015). *Brachypodium* species are used as biofuel research having ideal cell wall composition, physiology, and plant pathology as *Brachypodium* and have a variety of pathogen responses to common cereal fungal pests (Gomez, 2008; Lee, 2012; Marriott, 2014). *Brachypodium* is an ideal model species, but when compared to *A. thaliana* germplasm resources, it is limited to a few hundred accessions from predominantly modern day Turkey, Italy, and the Iberian Peninsula (Mur, 2011; Tyler, 2016). A large and widespread germplasm collection is required to use *Brachypodium species* as an invasion model. This thesis significantly increases the number of accessions for all *Brachypodium distachyon* complex species and expands its use as new model for invasion biology.

#### About The *Brachypodium distachyon* Species Complex

*Brachypodium distachyon* has had increased interest as a model species for grass genomics since its first proposal in 2001. *B. distachyon* has a small stature, grows well in laboratory conditions, amenable to transformation, and has a compact genome 266mb with a haploid chromosome number of  $x=5$  (Draper 2001, Opanowicz, 2008; Vogel, 2009; Vogel 2010). The full genome was published in 2010 and recently a pangenome from multiple genomes has been



published. Publications of the species complex are dominated by *B. distachyon*, but it has two other member species (Gordon, 2017).

*B. stacei* was originally regarded as an auto- or allo- polyploid relative of *B. distachyon*, but cytological studies revealed *B. stacei* as a  $2n=2x=20$  diploid with a haploid chromosome number  $x=10$ , twice that of *B. distachyon*, but a smaller genome size of  $\approx 240\text{mb}$  (Hasterok 2004; Hasterok 2006; Vogel, 2009; Catalan, 2012). Phenotypically *B. stacei* has a larger stature than *B. distachyon*, more similar to the allotetraploid *B. hybridum* (Catalan, 2012; Catalan 2015). Since *B. stacei* was regarded as an unusual polyploid cytotype of *B. distachyon* little effort was put forth to collect it and is rare in collections. Now identified as a different species, *B. stacei* only recently had its native range described having little geographic overlap with *B. distachyon* (Lopez-Alvarez, 2015).

Detected by probes, *B. hybridum* is a composite allotetraploid species with *B. stacei*-like and *B. distachyon*-like subgenomes (Hasterok, 2004, Idziak, 2011). It has similar traits and growth patterns to that of *B. stacei* and *B. distachyon* and an estimated genome size of  $\approx 510\text{mb}$ , about the size of *Setaria viridis* (Huang, 2014; Hasterok, 2004, Idziak, 2012; Catalan, 2012). Unlike *B. distachyon*, there are no known vernalization requiring lines of either *B. hybridum* or *B. stacei* (Vogel, 2009; Woods, 2013). Like *B. distachyon*, a significant increase in *B. hybridum* and *B. stacei* accessions are present in this study and greatly contributes to the future germplasm of the *Brachypodium distachyon* species complex.

#### Genetic Diversity of the *Brachypodium distachyon* Species Complex

Currently there are few scientific studies describing the relatedness of each complex species, which is needed for genome wide association studies (GWAS) and quantitative trait loci mapping (QTL). *B. distachyon* is the primary model of the three complex species having the most publications, yet most diversity studies cover sub regions of the circum-Mediterranean usually modern day Turkey or the Iberian Peninsula (Mur, 2011; Gordon, 2014; Tyler, 2016). One study calculated a core diversity set of 46 equidistant *B. distachyon* lines from 166 accessions and assessed phenotypic variation (Tyler, 2014). A total of 213 lines were published in 2016 from the same research group with previously un-sampled locations in central and northeast Mediterranean regions of Europe and Asia, finding three distinct ancestral lineages via the software STRUCTURE (Tyler, 2016).

Little is known about the genetic diversity of *B. hybridum*. One study highlights the genetic diversity in Iberia comparing cytotypes of *B. distachyon* before *B. hybridum* was described as a separate species (Mur, 2011). A *B. hybridum* study in Northern Mediterranean Africa in modern day Tunisia analysed 145 lines by 15 (SSR) markers across nine natural locations (Neji, 2015).

Interestingly, gene flow between locations was high,  $N_m=2.31$ , but geographic distance did not explain any significant genetic variation and was attributed to long-distance seed dispersal. Likewise, the genetic diversity of *B. stacei* is not well known, but a study in 2016 described much of the diversity in the western Mediterranean (Shiposha, 2016). Collections have been rare of *B. stacei* most likely because its ecological niche was not well described or collected from, but eventually described in 2015 (Lopez-Alvarez, 2015). While collections of *B. hybridum* have been common, being a polyploid cytotype has hindered its genetic exploration. Now with the *B. stacei* reference genome concatenated with a *B. distachyon* reference genome an *in-silico* *B. hybridum* reference genome can create a functional polyploid reference genome for read mapping (further described in methods).

### Landscape Genomic Concepts

Landscape genomics investigates patterns of genetic variation and gene flow between populations that result in local adaptation across geographic and environmental gradients (Manel, 2003; Manel, 2013; Rellstab, 2015; Bragg, 2015). While GWAS benefits from diverse subsets of unique individuals, landscape genomics requires the actual collected samples from all locations (Bragg, 2015). Individuals should differ genetically at adaptive loci across environmental gradients. After time, this can create population structure across environmental gradients, masking the adaptive alleles on a diverged genomic background (Wright, 1934; Kimura and Maruyama, 1971; Bragg, 2015). These ancestral groups might re-encounter each other in overlapping ranges and sexually recombine, creating admixed individuals. Here, traits from each group are shared by shuffling of genetic material between lineages. Environmental distance can also cause isolation leaving mutations to accumulate within confined individuals as seen in a study on *Alnus glutinosa* where phenotypes were associated with environmental variation (De Kort, 2014). Environmental or geographic distance can alter allele frequencies between populations due to genetic drift. Most non-lethal mutations are neutral and cause no phenotypic change, however it has been shown that some mutations are neutral in some environments and adaptive in others. In *A. thaliana*, an early stop codon in chromomethyltransferase two (CMT2) changed the stress response to cold and increased tolerance (Shen, 2014). The presence of the CMT2 mutation was more abundant in regions where *A. thaliana* would encounter extreme cold, but also present in non-cold extreme locations.

### Genomic Diversity and Patterns Across Geography

Natural Diversity studies like GWAS require knowledge of existing genetic diversity as seen in other model species (Kim, 2007; Brachi, 2011; Jia, 2013; Huang, 2014). Information pertaining to a study system's diversity, linkage disequilibrium, ancestral composition can inform researchers about the creation of recombinant inbred lines (RILs), diversity sets, and reverse genetics (Nordborg, 2005; Platt, 2010; Bomblies, 2010; Long, 2013; Huang, 2014). Once these concepts are understood the genome can be further disentangled to examine processes that once

affected its composition and structure. Example studies used *Arabidopsis* species to analyse genotype-by-environment effects and has been particularly successful in landscape genomic studies. One study showed a pattern across geographic space where specific genomic markers varied across specific landscape gradients (Hancock, 2011). In the same study, polymorphisms within a location were predictive of fitness at that location.

#### Self pollination: multiple inbred accessions descending from a few whole genome genotypes

A variety of phenotypes are common in invasive plants and one of the most common is self-pollination (Rogers, 2011). This is mostly due to the Allele Effect where the necessity of outcrossing can be detrimental in isolated populations with few individuals, and especially if there is little genetic diversity (Sutherland, 1999). This can even be the case if a native pollinator is abundant and carrying pollen from local individuals (Viet, 1996; Leung, 2004; Taylor, 2004). Since invasive plants often are introduced in small quantities at first with little to no others to mate with, there is little change they will become invasive if the species is an obligate out-crosser, worse if a pollinator is necessary. Depending on the species and the system, in theory many individuals could be introduced at once, as is hypothesized with *Brachypodium* species being introduced as a grain contaminant. It is known that *Brachypodium* complex species have cleistogamous flowers that often remain closed whereby reducing outcrossing and increasing self-pollination (Garvin, 2008; Vogel, 2009). Variation does exist in weedy and invasive species with regards to pollination. *A. thaliana* has been introduced to North America and as a facultative self-compatible species with low outcrossing rates, many of the same genotype were found across large geographic range over 1000 times, but in native populations of *A. thaliana* the same genotype is rarely found farther than a few kilometers (Platt, 2010; Anastasia, 2011). In the case of another invasive self-pollinating species *Lithrum salicaria*, it is capable of rapidly outcrossing with neighboring individuals (Coulatti, 2013). Also, *L. salicaria* was found to have locally adaptive mutations of non-native origin that conveyed a larger range along a north-south gradient and those mutations were found to be vectored by pollen, indicating that gene flow by pollen or seed can cause an invasive to flourish.

#### Question, Hypothesis and Aim of Chapter

The nature of this dissertation is to investigate the genetic, geographic, and climate diversity of three *Brachypodium* species in the context of landscape genomics and invasion biology. This required genomic analysis and genetic variation across geography on a global scale. The first goal is to calculate within species genetic diversity. The *Brachypodium distachyon* complex species has travelled extensively out of their native range and a focused question can be addressed using this species group as a model.

**Questions:** *What is the genetic diversity in relation to geography of the *Brachypodium distachyon* species complex and do genotypes of any species trend more as high or low dispersers?*

**Hypothesis:** *I hypothesize that since polyploid species often have larger stature and wider distributions geographically, the polyploid complex member *B. hybridum* should be more globally distributed than diploid species. Furthermore, some lineages will be more dispersed than others.*

**Aim:** *Obtain DNA sequence from individuals of each species to determine the genetic diversity across geography and test to see if some genotypes are more abundant than others by being better at dispersal. Use pairwise genetic distance among accessions to cluster whole genome genotype groups.*

### 3.2 Methods

---

#### DNA Preparation, Library Preparation, DNA Sequencing, and Demultiplexing

The DNA sequencing preparation is exactly the same as described in Chapter II. To recap: DNA was extracted via DNEasy 96-well kits, quantified by QuBit 2.0 fluorometer and restriction enzyme digested into fragments via genotyping by sequencing techniques (GBS). DNA was ligated to barcodes with Illumina Y adapters and PCR primers for PCR amplification. Post PCR amplification, DNA fragment concentrations were assessed per well per plate on a Shimadzu MCE-202 Multina 96 well plate reader with the DNA-12000 chemistry, and a Perkin-Elmer GXII Assay Chip. Sample wells were pooled for gel-based fragment size selection and sequenced on Illumina HiSeq2000, HiSeq2500, and NextSeq500 platforms with paired end format. Raw fastq sequence files were demultiplexed using the software AXE (<https://github.com/kdmurray91/axe>). Reads are cross referenced to an index of forward and reverse barcodes allowing one mismatch not shared with another barcode, then binned to each sample's fastq files.

#### Reference Genomes and Their Properties

Three reference genomes were used in this study. For genotyping *B. distachyon* samples, the Bd21 *de-novo* assembly genome and its SNP-corrected Bd1-1 version 1.0 (Vogel, 2010; Gordon, 2014). The *B. stacei* ABR114 genome is an early release first edition of the *B. stacei* reference genome created by the Vogel Lab at The Joint Genome Institute University of California, Berkeley (JGI, 2016). An *in-silico* reference genome was created for the polyploid *B. hybridum*, where the *B. stacei* reference genome was combined with the Bd1-1 *B. distachyon* reference genome to make an *in-silico* polyploid reference genome: *B. stacei* ( $x = 1-10$ ) and *B. distachyon* ( $x = 11-15$ ) to make the *B. hybridum* haploid  $x=15$ .

#### The TASSEL Genotyping Pipeline and Variant Caller Settings

The TASSEL pipeline is an effective and quick method for marker detection for genotyping samples (Shendure, 2008; Elshire, 2011). A minimum of five identical read counts across all samples were required to validate a marker. Before alignment, reads are trimmed to 64 bases

then aligning software used was the Burrows-Wheeler Aligner platform (Heng, 2009). In BWA reads are assigned a mapping quality score from 0-37 based on a variety of factors, mostly by how uniquely they map to one plus loci. Thus reads with quality scores below 10 were removed. For diploid species, the miss-match rate was set to the standard 4% or four mismatches out of 100 bases. For the polyploid *B. hybridum* the miss-match rate was cut in half to 2% to reduce the quantity of reads mapping to both sub-genomes. Once reads pass the minimum presence rate and miss-match rate, they must pass an allele frequency rate. The minor allele frequency rate was set to 0.001 of presence across all samples, however, more stringent filters were applied post TASSEL in R. The final output from the TASSEL pipeline is a matrix of SNP alleles in rows and samples in columns.

#### Whole Genome Genotype Clustering and Population Analysis in R

The SNP matrix of samples was analysed using custom filtering scripts in R. To insure that samples are comparable two metrics are applied. First, samples need to have enough sequence depth to meet a minimum threshold for quantity of genomic variants, and secondly markers must be present in a minimum number of samples to be valid. In *B. distachyon* and *B. stacei* the minimum number of markers was set at 10,000 and 4,000 respectively. For *B. hybridum* the minimum marker threshold was set to 21,803 variants because the size difference between the *B. stacei* subgenome (S subgenome @  $\approx 240\text{mb}$ ) and the *B. distachyon* subgenome (D subgenome @  $\approx 266\text{mb}$ ):  $10,000\text{ variants} * 266\text{mb} / 240\text{mb} = 11,803\text{ variants}$ . Valid markers must be present in  $\geq 50\%$  of samples to become usable markers, all markers with  $< 50\%$  of presence across all samples are then removed. Rare paralogous variants are also removed. For the polyploid *B. hybridum*, each subgenome is filtered independently and checked for equal proportion of variants per each subgenome. Few samples had unequal ratios of subgenome markers that still passed filtering and were flagged and removed due to uneven coverage or alignment to subgenomes. Once each subgenome was filtered, they were merged into a genomic matrix of variants.

#### Genetic Diversity Calculations in R

In each species analysis, individuals are run through a principal coordinate analysis (PCoA) using the R package 'cmdscale'. The principle coordinate vectors are also used to calculate pairwise distance using the R base package cor() and as.dist() using the cor() function "pairwise.complete.obs". Pairwise distance matrices were converted to hierarchical clustered objects using the R package 'hclust' and genotype boundaries are created using rect.hclust() on dendrograms. Technical replicates, two seeds from the same maternal plant, were grown and independently DNA sequenced to detect dendrogram resolution and accuracy. In each species, the technical replicate with the highest branch height was used to call genotype across all samples.

### Geographic Regions of Genotypes

The partitioning of geographic coordinates into regions in Chapter II is used to show the genotype diversity across geography. To recap: the region partitioning method in the previous chapter, each collaborator collected samples with different methods, to normalise the collection locations, they were first partitioned into sub-continental sets, then clustered via a common clustering algorithm into regions by their latitude and longitude coordinates in the R package Mclust (Fraley, 2002; Fraley, 2012). This created 20 locations for *B. distachyon* in the European and Asian Mediterranean regions, and four continents for *B. hybridum* and 35 regions. *B. stacei* was rare and most in close proximity and is excluded for most geographic analysis of this thesis.

### Permutation Tests to calculate Genotype Abundance across Geography

Genotype abundance is a central topic to this thesis and sheds light on the variation in dispersal ability. A permutation test was used to calculate what genotypes had high or low abundance across geography thus would be considered good or bad at dispersal. Genotypes had various levels of detection ranging from n=55 unique geographic sites or less, thus the number of randomly drawn geographic locations per iteration was set to the total number of observations in the true data set for each genotype- a genotype found 55 times, had a random sample draw of 55 points per iteration, and a genotype with 33 sites would have 33 randomly drawn geographic sites per iteration in the permutation test. Samples present in less than nine locations were removed from testing due to low statistical power and considered under sampled. Each genotype was tested individually. In each iteration a randomised subset of geographic sites were drawn and the frequency of all common genotypes counted and used in a normal distribution to calculate probability of drawing a specific genotype by chance. This probability was averaged across iterations to define the final p-value. The mean number of sites and standard deviation a genotype could be found in was also reported.

## **3.3 Results**

---

### Genetic Diversity of Each Species, high quality accessions

Once samples were assigned a species identify, they were run through our species genotyping pipeline. Species were run through a standard Tassel 3.0 reference genome genotyping pipeline except the minor allele frequency was changed from 0.01 to 0.001. The diploid species *B. distachyon* and *B. stacei* used the standard setting in BWA. The *B. hybridum* BWA settings were standard settings except the miss-match setting was set to two mismatches instead of four. This allows BWA to map reads with less ambiguity between sub-genomes of a tetraploid and still provide enough easily mapped subgenome specific markers. Samples are moved from Tassel to R for further filtering of genotyping data. Of the 528 samples that qualified for *B. distachyon* candidacy 479 passed genotype filtering having a minimum of 10,000 variants. Of the 60 candidates for *B. stacei* species assignment, 50 passed filtering having a minimum of

4,000 variants. Of the 1,201 samples that received candidacy for *B. hybridum* 1,015 samples passed filtering by having equal or greater than 21,803 variants (See table 3.1).

Genotyping By Sequencing Marker Data: high quality SNPs

Each species final genotype data was filtered to descriptive markers for calling diversity and genotype. *B. stacei* had the least total markers called at 4,744 differentially called loci. *B. distachyon* had nearly a power of ten increase of sample number, thus had higher markers for genotyping samples. *B. hybridum* had substantial coverage between subgenomes having 18,525 differentially called markers. The average distance between markers in *B. stacei* was the greatest at 50,590 bases. *B. hybridum* averaged 28,070 bases between markers across both subgenomes. *B. distachyon* had the least distance at 18,426 bases (see Table 3.1).

Genotype identity was called conservatively based on the resolution of relatedness between replicate individual accessions of the same germline. The dendrogram cut height was set at the highest branch of two individuals of the same maternal accession replicate. Replicates with poor sequencing quality may not be identified as closely related. Thus, a conservative threshold for our whole genome genotype categorization. While most sample replicates are of close proximity to each other, the conservative threshold splits genetic relatedness at a level higher than typical for inbred line genotyping studies (Platt, 2010). This threshold can group distinct but related lines within a family, such as recombinant inbred lines which share common parents. For simplicity the term *genotype* will still be used interchangeably with *genotype family*. A total of 479 individuals of *B. distachyon* were condensed to 125 genotypes. From 60 individuals of *B. stacei*, eight genotypes were found, and 80 genotypes of *B. hybridum* were found from 1,147 individuals. The total number of individuals per genotypes was calculated per each species. Though under sampled in this dataset, *B. stacei* had a total of 0.16 genotypes/individuals. *B. distachyon* had the highest diversity of genotypes/accessions at 0.26. *B. hybridum* had the lowest number of genotypes/accessions at 0.07 (See Table 3.4).

Sequencing and SNP/Variant Data	<i>B. distachyon</i>	<i>B. stacei</i>	<i>B. hybridum</i>
Input Samples	528	60	1,201
Raw Marker Count	54,131	10,530	97,949
Minimum Variants in Quality Samples	10,000	4,000	21,803
Samples with Marker Count Greater than Min. Variants	479	50	1,180
Minimum Samples in quality Variants	240	26	508
Valid Variants with Sample Counts Greater than Minimum Samples	14,436	4,744	18,525
Unique Genotype Families	125	8	80
Genotypes per Accessions	0.26	0.16	0.07

**Table 3.1.** Table of Genotyping Properties per Species: The data used to genotype samples shows many thousands of markers to profile relatedness. *B. distachyon* accrued the most genotypes of the three species despite having a smaller sample size. *B. stacei* was not common in our data sets and *B. hybridum* was abundant.

Figure 3.2: *Brachypodium distachyon*

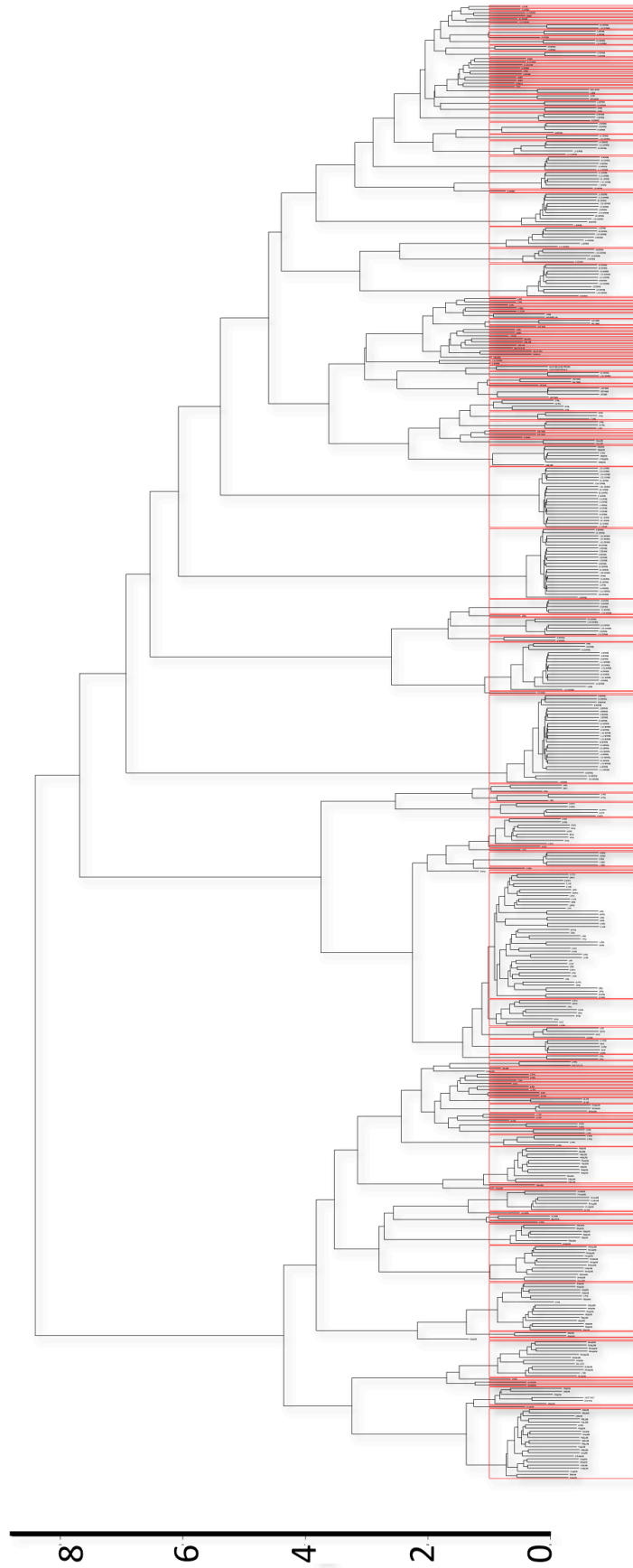


Figure 3.2 *Brachypodium distachyon*: A dendrogram of the 129 individuals comprising 125 genotypes: The sequencing and genomic analysis of *B. distachyon* yielded 129 genetically distinct individuals, of 4 were crosses were included for comparison of heterozygosity. A total of 125 genotypes were calculated from 479 individuals using 14,436 SNPs. The dendrogram is cut at the highest branch of a technical replicate, two individuals from the same maternal line.



Figure 3.3: *Brachypodium stacei*

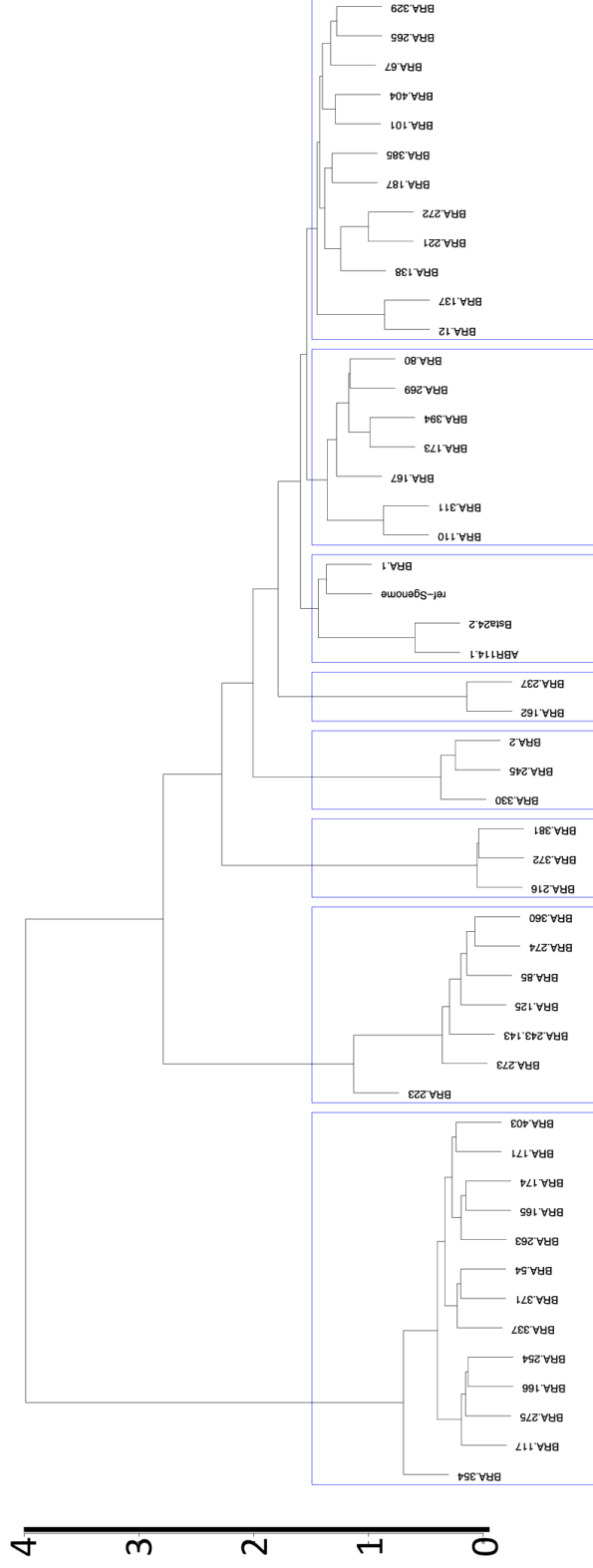


Figure 3.3 *Brachypodium stacei*: A dendrogram of the 50 individuals comprising eight genotypes: The sequencing and genomic analysis of *B. stacei* yielded 8 genotypes that were calculated from 60 individuals using 4,744 SNPs. The only known technical replicate of the same accession was the reference genome and a sample of ABR114 and is the height the the dendrogram was cut at.

Figure 3.4: *Brachypodium hybridum*

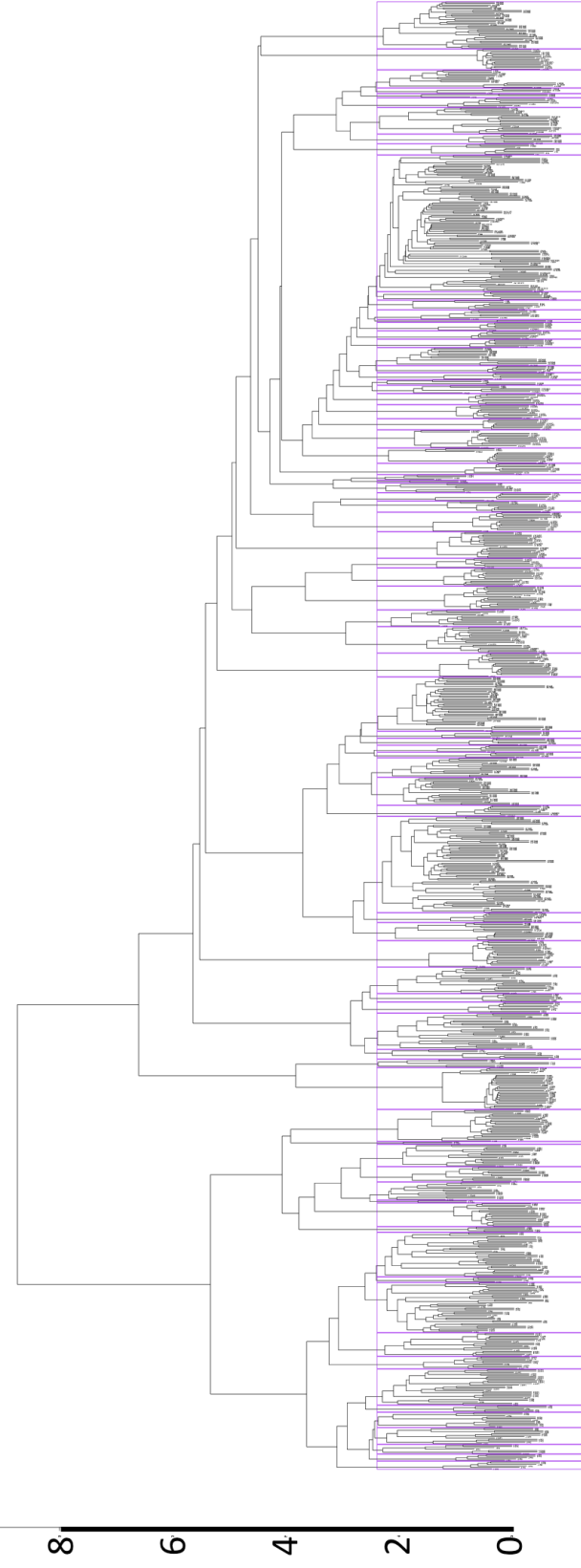


Figure 3.4 *Brachypodium hybridum*: A dendrogram of 1,052 individuals comprising 80 genotypes: The sequencing and genomic analysis of *B. hybridum* yielded 80 genotypes that were calculated from 1,052 individuals using 18,525 variants. The dendrogram is cut at the highest branch of a technical replicate, two individuals from the same maternal line.

### Genotypes by Continent

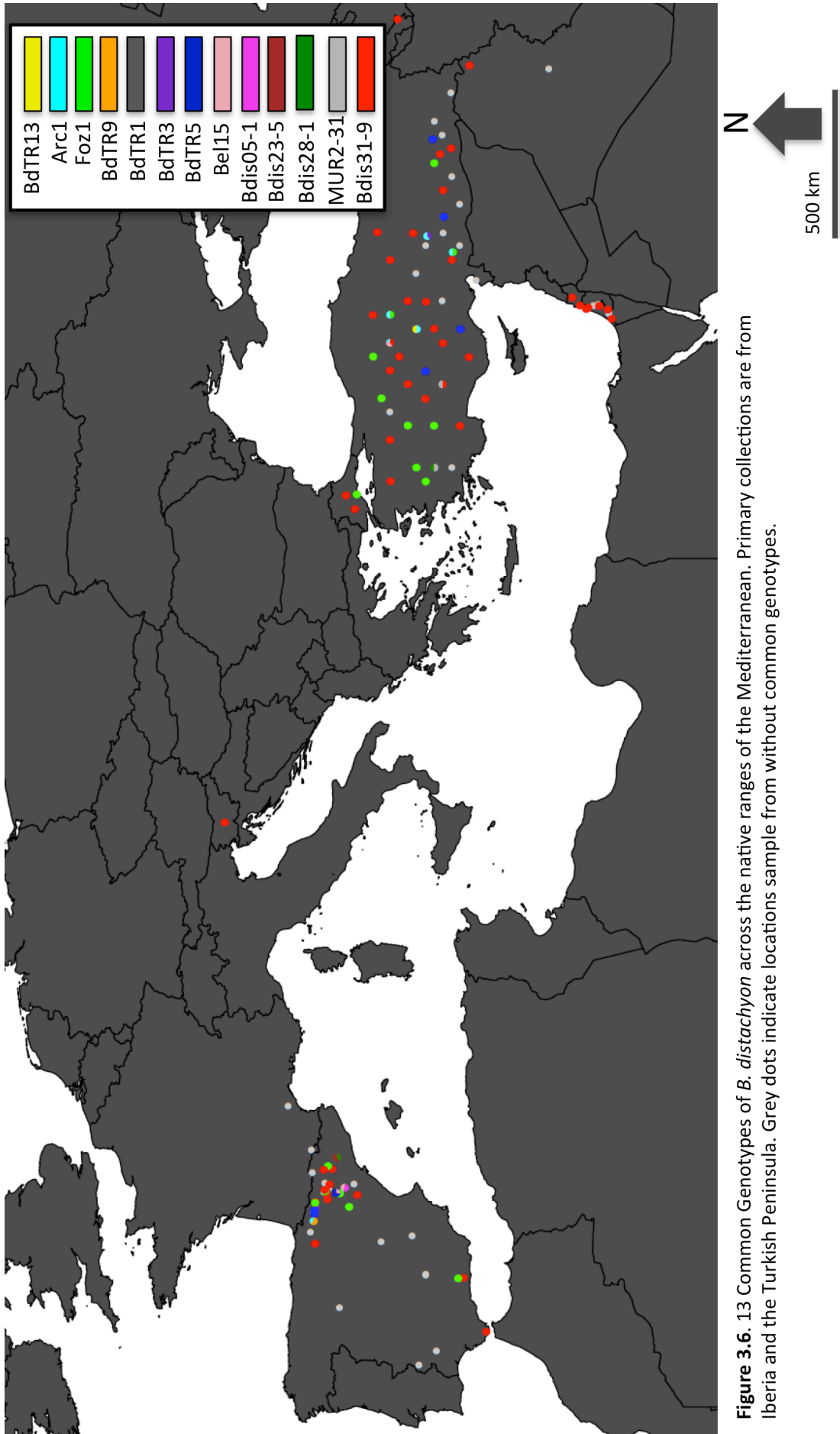
Diploid species were mostly found in their native circum Mediterranean ranges. *B. distachyon* had the highest diversity in Europe with many locations having multiple genotypes per location, while less diversity was found in West Asia. Two genotypes of *B. distachyon* were found in Australia and constitute the only non-native diploids in this study. *B. stacei* was predominantly sourced from the eastern Mediterranean regions of modern day Israel and Palestine. Genotype diversity is highest in both Europe (native) and Australia (non-native), both having more genotypes than captured in Asia. Of the three species in the complex, *B. hybridum* is the dominant species in all non-native ranges. Australia had the most number of genotypes in a non-native range at 38, but had many more sample sites to capture diverse genotypes. North America had the highest number of genotypes per sites, but could be due to diverse admixed individuals creating many genotypes that shuffled the same haplotypes from few or more founders. South America and South Africa had large amounts of genotypes per sample location, but both locations are represented by one single collection location each.

---	<i>B. distachyon</i>	<i>B. stacei</i>	<i>B. hybridum</i>
Location	Genotypes/ Locations	Genotypes/ Locations	Genotypes/ Locations
Europe	100/47	1/1	38/195
Asia	38/66	7/2	28/325
North Africa	---	1/1	4/1
Australia	2/2	---	38/83
North America	---	---	25/26
South America	---	---	6/1
Southern Africa	---	---	4/1

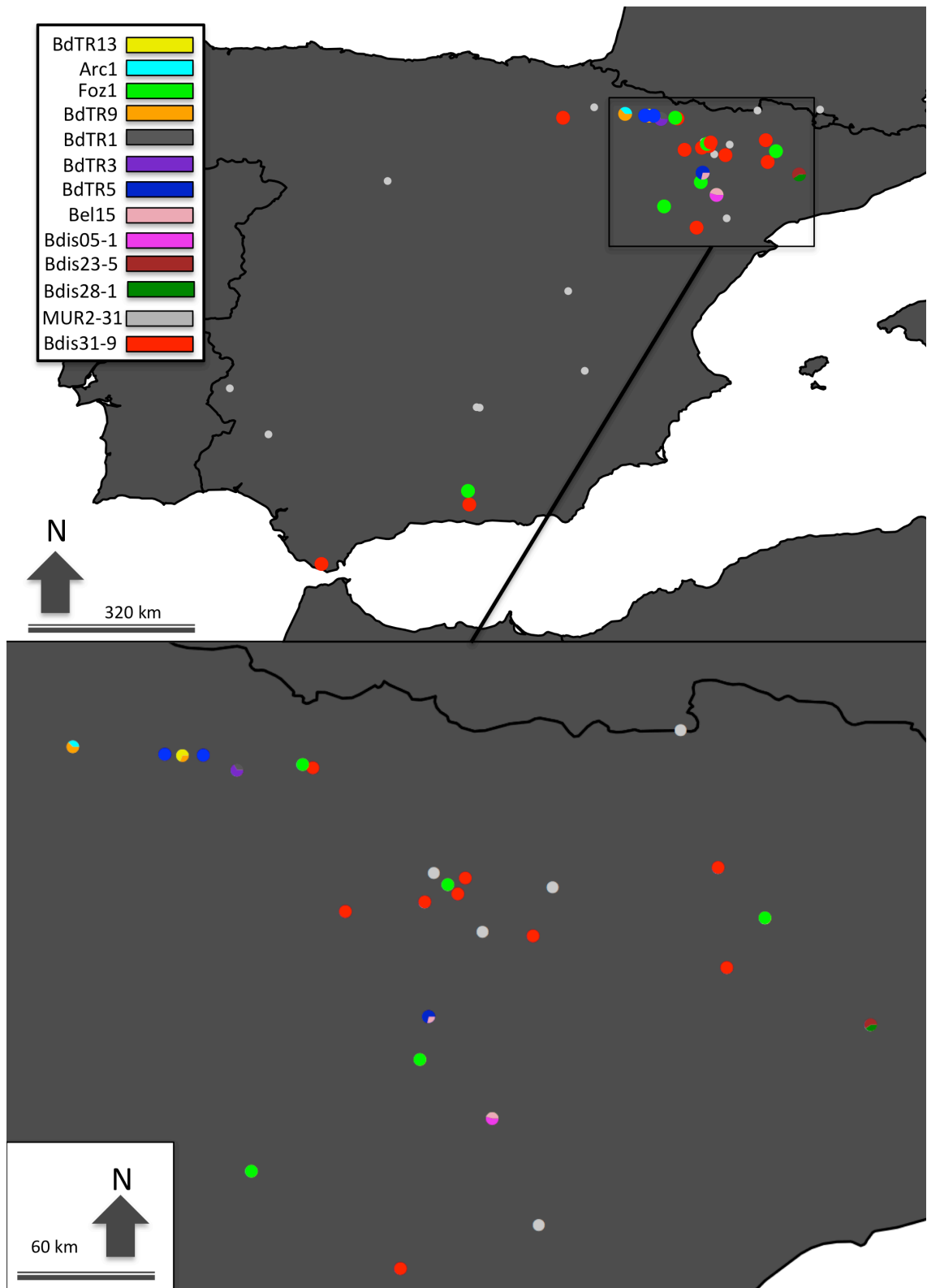
**Table 3.5.** The total number of Genotypes per species and their location per region: Diversity per each species is listed above with location number per continent. The most genetic diversity for both *B. hybridum* and *B. distachyon* was in Europe. North America and Australia both had substantial quantities of genotypes rivaling the native range in some cases.

### Pie Chart Maps Show the 13 Most Widely Distributed Genotypes Groups of *B. distachyon*

The 13 most common genotypes or genotype groups in this data set were mapped via pie charts of where they were found in their native range. The areas of high diversity for *B. distachyon* per this study's sampling resolution are in northeast Spain and central Turkey. Northeast Spain in particular had heavy sampling per location and some of the sites used were sampled from more than one collaborator allowing sizable diversity to be sampled. Much of Turkey was only sampled one time across three different collection efforts. The genotype group Mon3 was the most abundant across much of the range analysed having 41 samples. More rare groups like BdTR13 and Arc1, were found the least at 10 times each (See figure 3.6) For pie charts on the native range of *B. distachyon*, pie colours represent genotypes, grey dots indicate locations sampled from, but did not contain a common genotype.



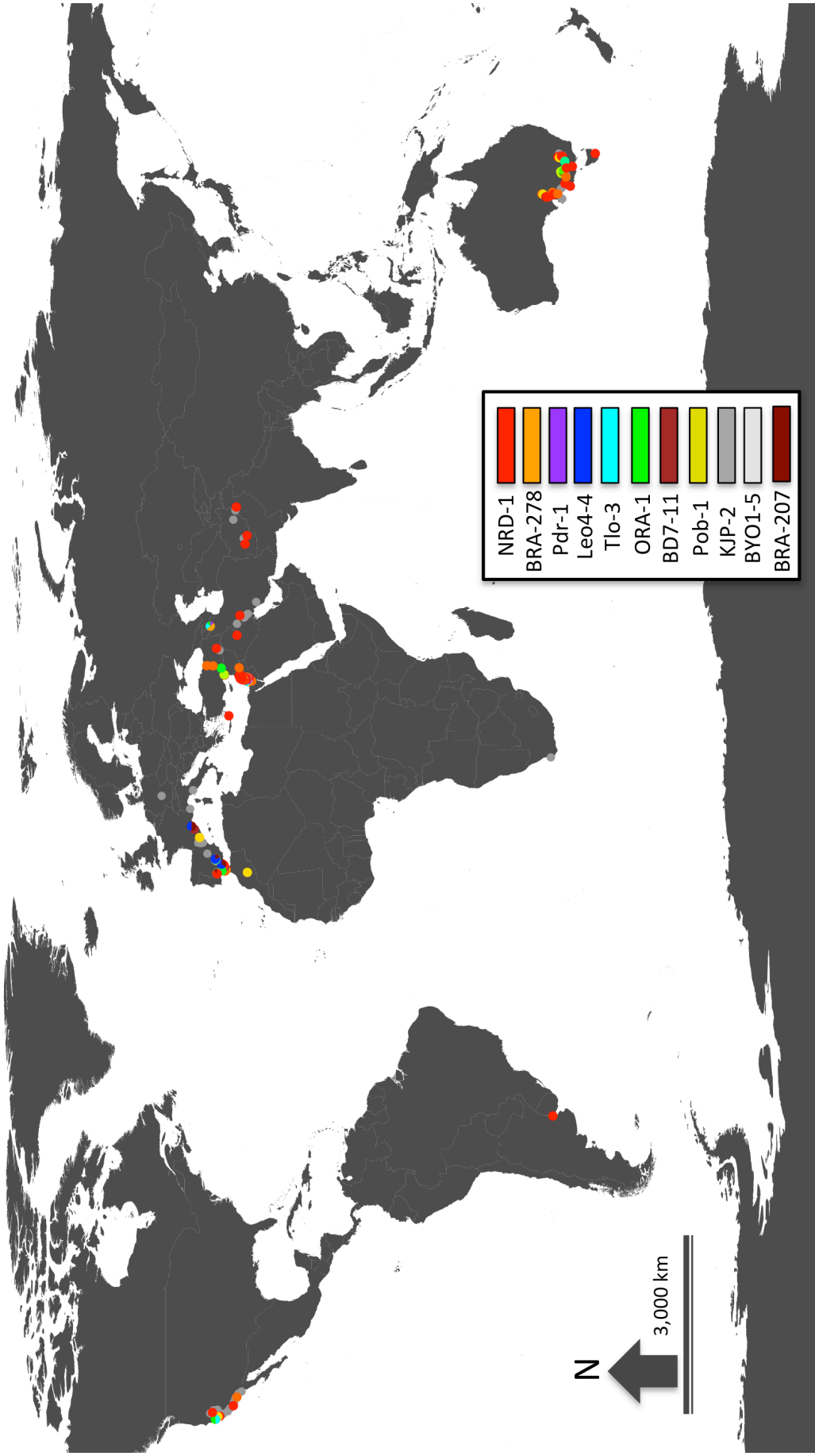
**Figure 3.6.** 13 Common Genotypes of *B. distachyon* across the native ranges of the Mediterranean. Primary collections are from Iberia and the Turkish Peninsula. Grey dots indicate locations sample from without common genotypes.



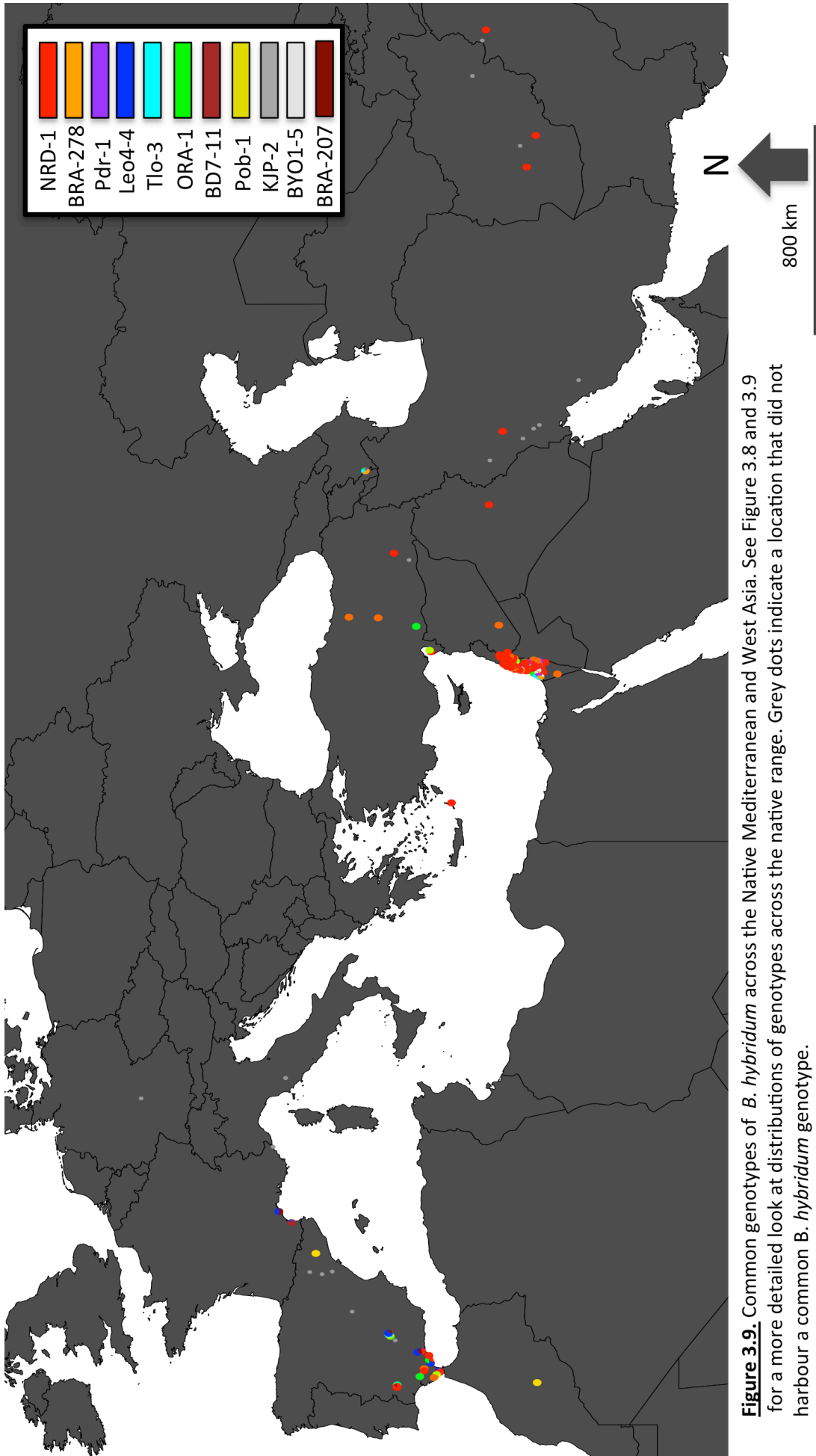
**Figure 3.7.** 13 Common *B. distachyon* genotypes across the Iberian Peninsula drawn as pie charts. In both figures, sample locations without a common genotype are indicated by a grey dot. Above: A significant amount of sampling efforts were put into sampling the NE section of the peninsula. Below: A close view of genotypes in the North Eastern Iberian peninsula.

*Pie Chart Maps of 11 Most Common Distributed Genotypes Groups of B. hybridum*

Eleven of the most common genotypes globally were plotted in pie charts on maps for *B. hybridum*. In the native range of *B. hybridum* there is an obvious difference in most of the common genotype composition with some types being more common in some areas than others. Common genotypes found across many locations will have multiple accession names while still being the same genotype. If an accession name was previously published, it was given priority in naming a genotype. Genotype NRD-1 was found across both east and west Mediterranean locations, but was more common in the east. NRD-1 was also the most common globally and was found on four continents indicated by a red colour. Figure 3.8 shows the distribution globally of the 11 most common *B. hybridum*. While much of the diversity found in the state of California in the US is traceable to Eastern Europe, many of the common genotypes in California are found on both sides of the Mediterranean.

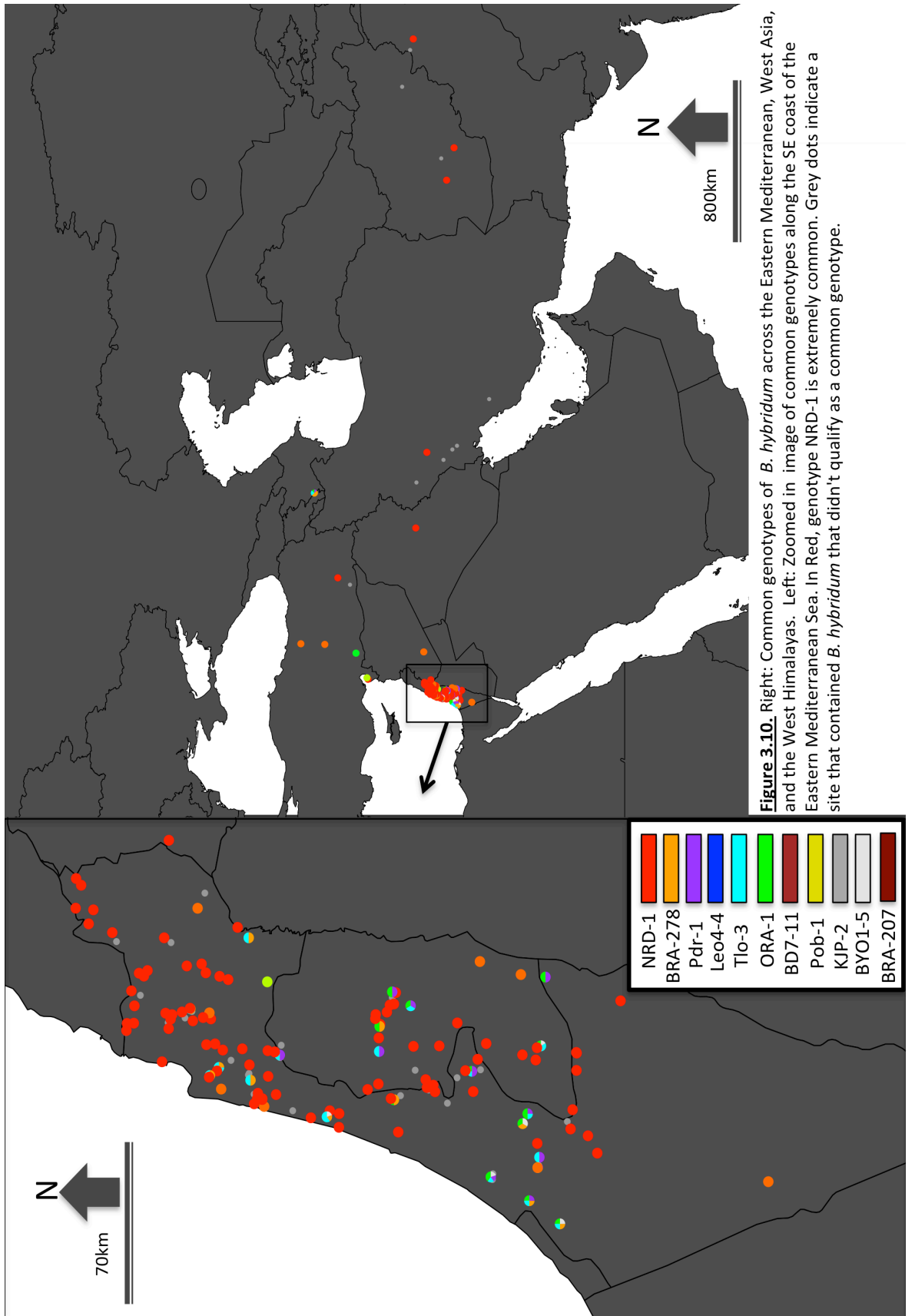


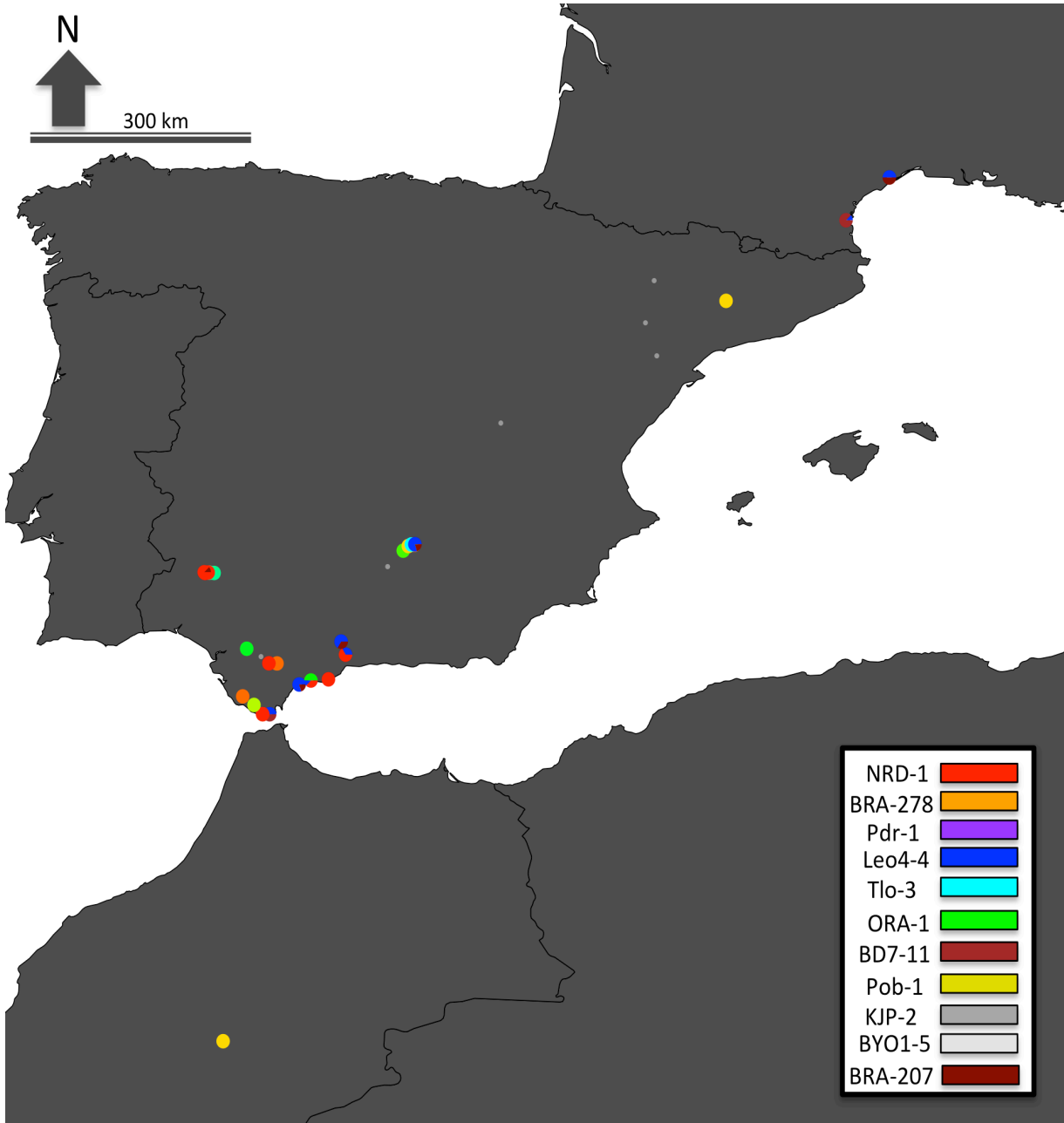
**Figure 3.8.** Eleven Common genotypes of *B. hybridum* and their global distribution plotted as pie charts on collection sites. Grey dots are locations that did have *B. hybridum*, but not a common genotype.



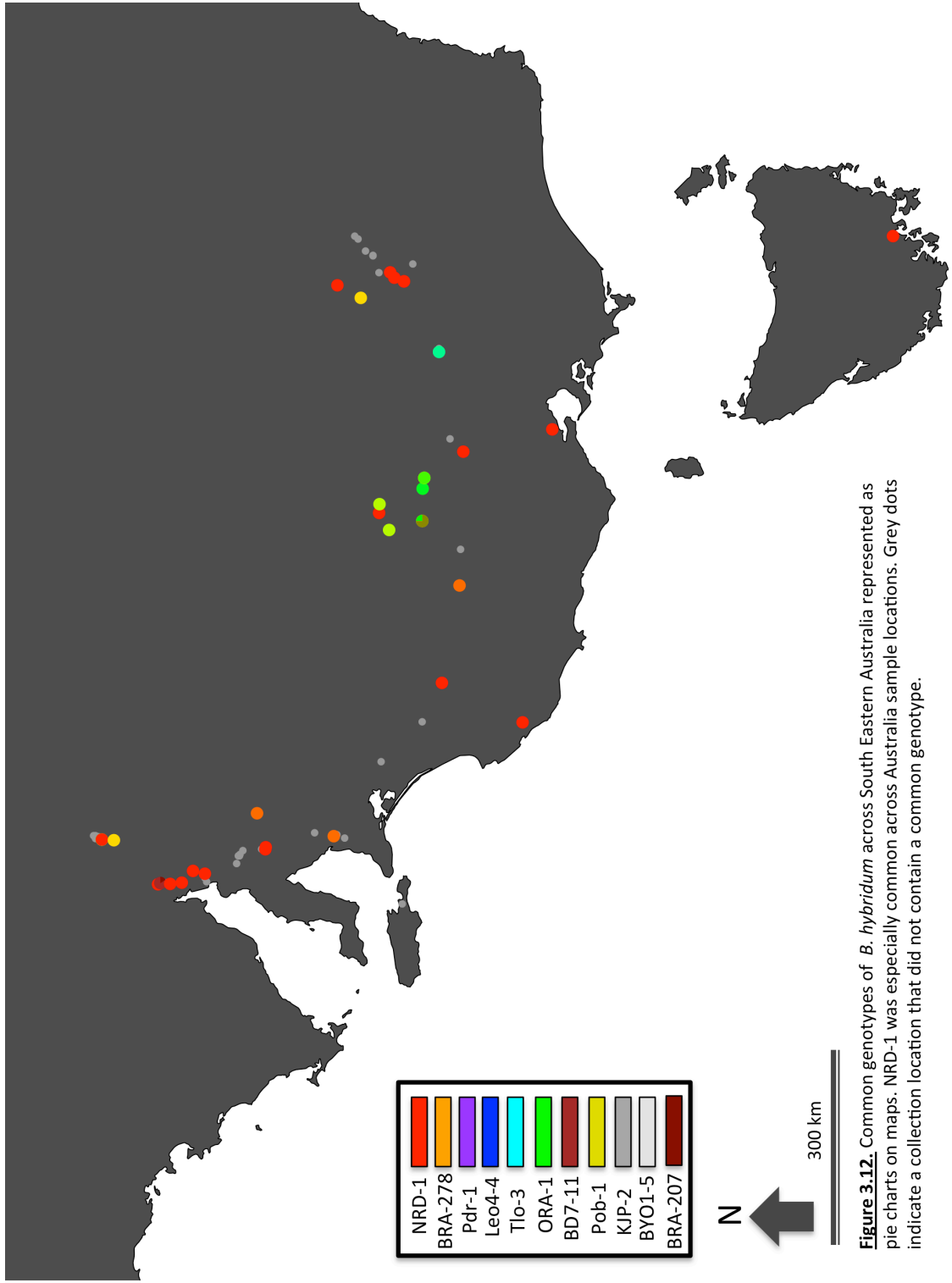
**Figure 3.9.** Common genotypes of *B. hybridum* across the Native Mediterranean and West Asia. See Figure 3.8 and 3.9 for a more detailed look at distributions of genotypes across the native range. Grey dots indicate a location that did not harbour a common *B. hybridum* genotype.



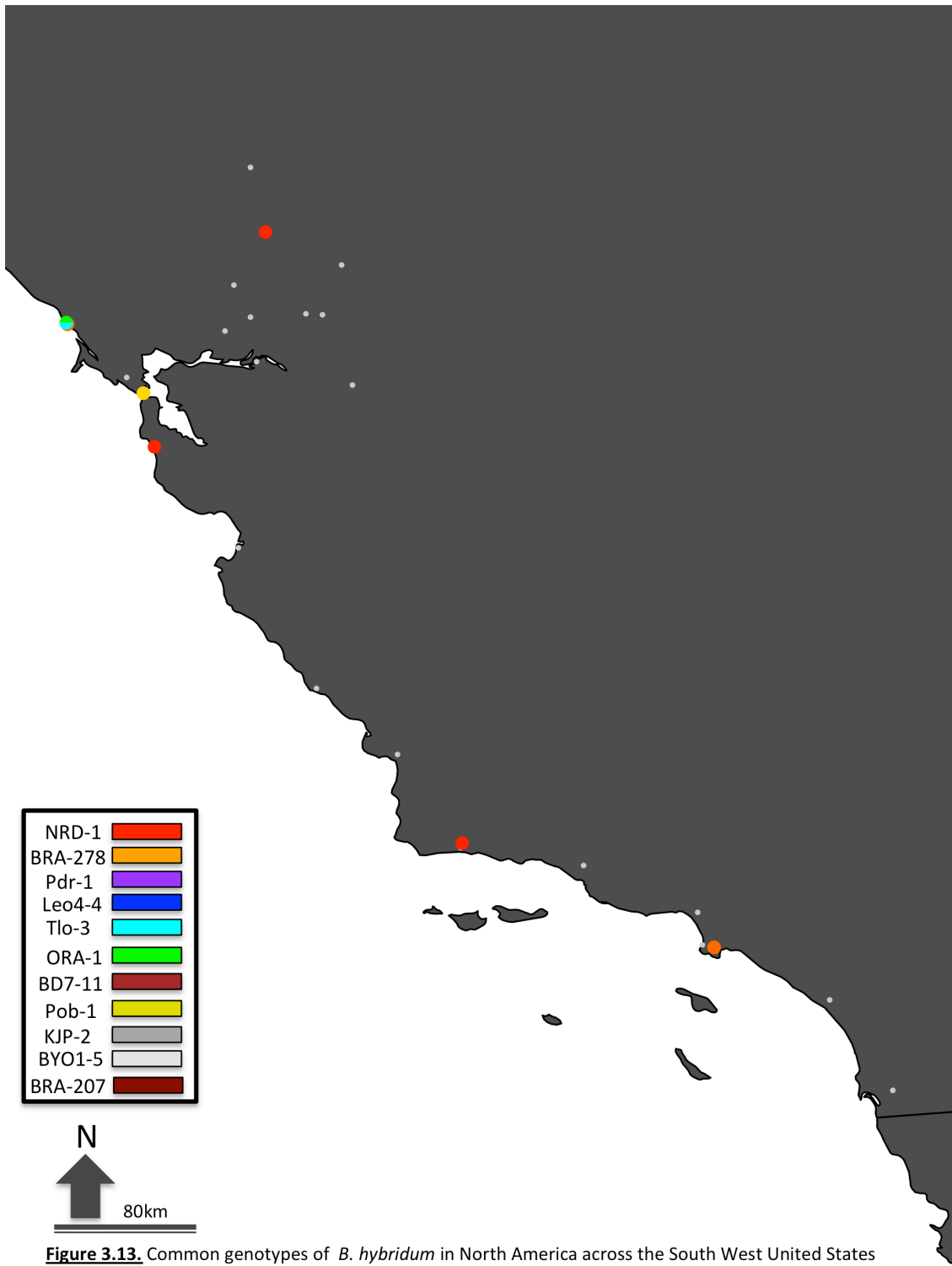




**Figure 3.11** Right: Common genotypes of *B. hybridum* across the Western Mediterranean and North Africa. Grey dots indicate a location that did harbour *B. hybridum*, but didn't not qualify as a common genotype.

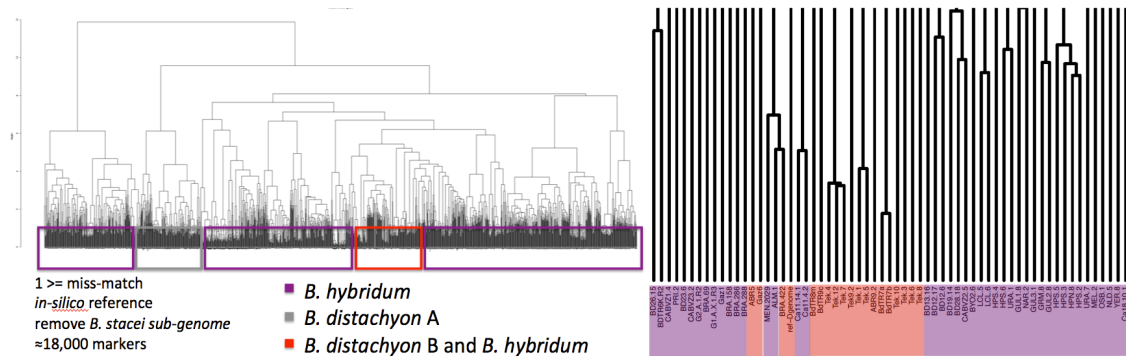


**Figure 3.12.** Common genotypes of *B. hybridum* across South Eastern Australia represented as pie charts on maps. NRD-1 was especially common across Australia sample locations. Grey dots indicate a collection location that did not contain a common genotype.



### Origins of the *B. hybridum* subgenome

In searching for the origin of the *B. hybridum* subgenomes, vernalization requiring *B. distachyon* lines aligned as close extant relatives to the D subgenome of many *B. hybridum*. Another group of *B. hybridum* had distinctly different D subgenomes on the left of figure 3.14. The origins of the *B. stacei*-like subgenome (S genome) were also investigated, but little is known about *B. stacei* and extant relatives due to the low sample coverage of *B. stacei* and few studies about its origins.



**Figure 3.14.** Origins of the *B. hybridum* D Genome. Left: The dendrogram of both *B. hybridum* D subgenome specific variants and *B. distachyon* individuals. Individuals in purple are true *B. hybridum*. In grey and red are known *B. distachyon*. Right: A close up of known *B. distachyon* B groups that are known to require vernalization are found among *B. hybridum* D subgenomes. Also in the left figure a separate group of *B. distachyon* A group are found between a large out-group of *B. hybridum*.

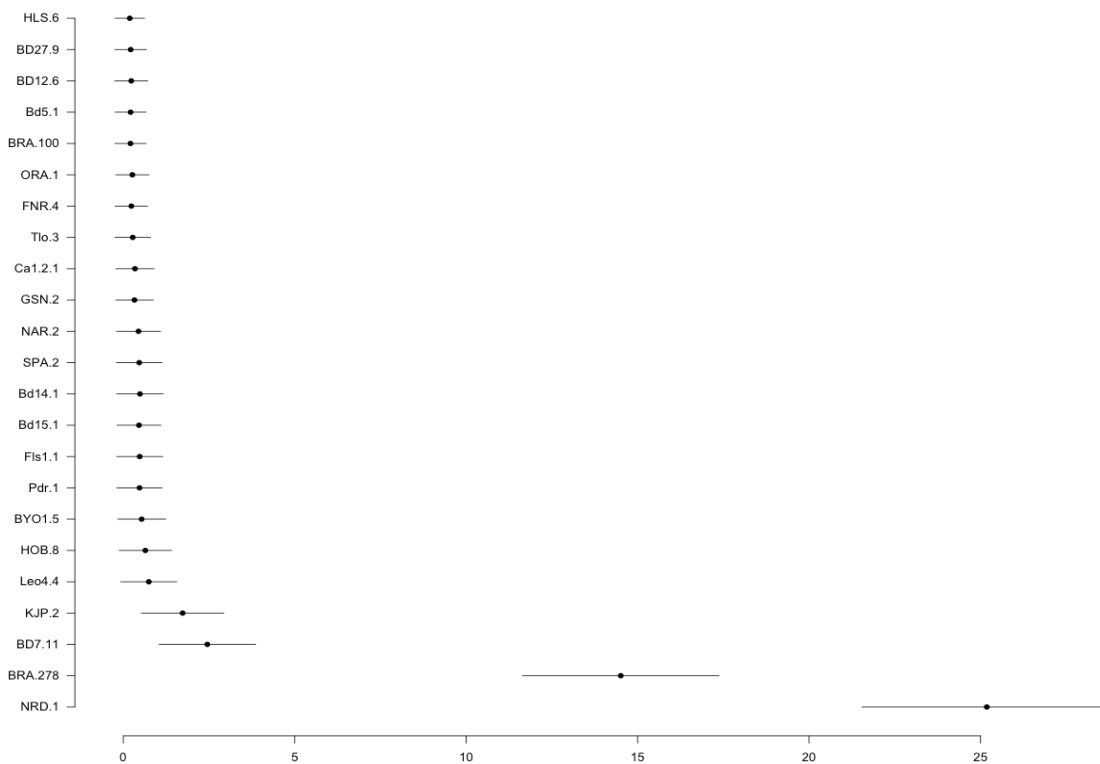
### Genetic association to Geography

A mantel test using the R package ‘mantel.rtest’ of genetic distance to geography revealed that geographic distance explained 4.7% of the overall genetic difference in *B. distachyon* with a p-value of 0.005 (Chessel, 2004; Mantel, 1967). There is little geographic distance explaining genetic distance indicating individuals are migrating randomly across sampling regions. *B. hybridum* was also tested for geographic variation explaining genetic distance and an association was found at 9.0% with a p-value = 0.01. It should be noted that other methods do exist for detecting isolation by distance or association between climate or phenotypes than mantel tests, and that mantel tests do have bias (Guillot, 2013). Sampling bias is the root cause of false association between two plus matrices of data meaning both mantel and partial mantel tests are both susceptible to bias.

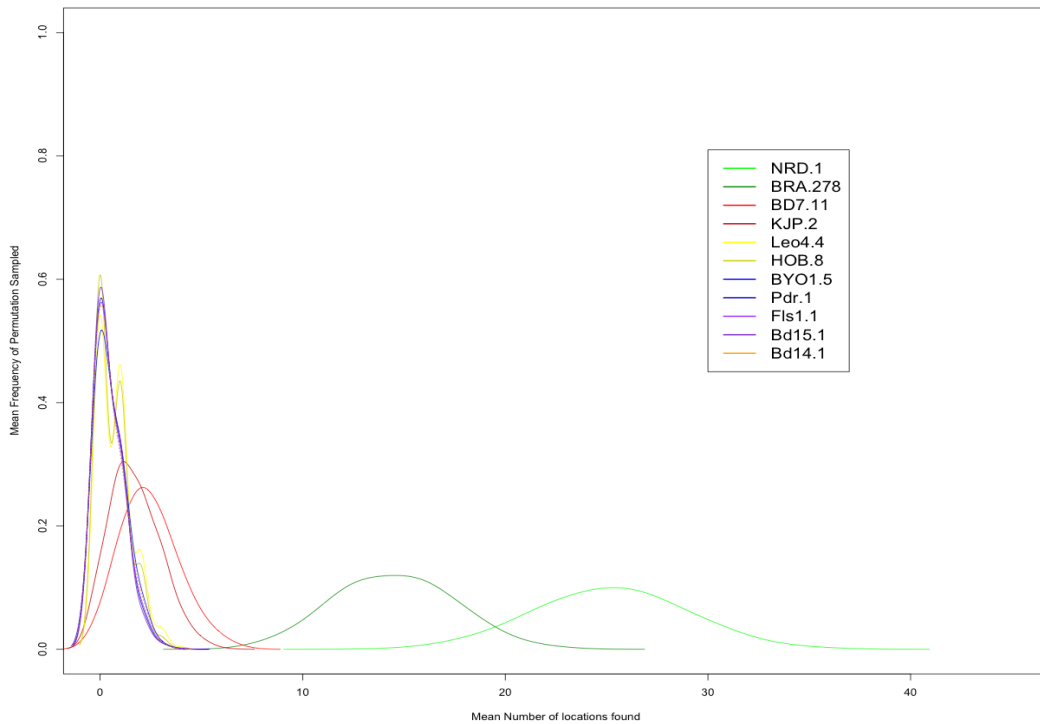
### Permutation tests of geography in *B. hybridum*, all locations:

Permutation tests were performed to test the geographic breadth of common genotypes, in both geographic abundance across site locations and regions as described in Chapter II. Specific qualifications were needed for genotypes to be tested. For location testing genotypes must be at seven or more locations. For regional testing a genotype needed to be present at a minimum of three regions. Random draws of sites from all locations were sampled at each genotypes presence number per iteration of 1,000 iterations. Then the actual presence of that genotype per

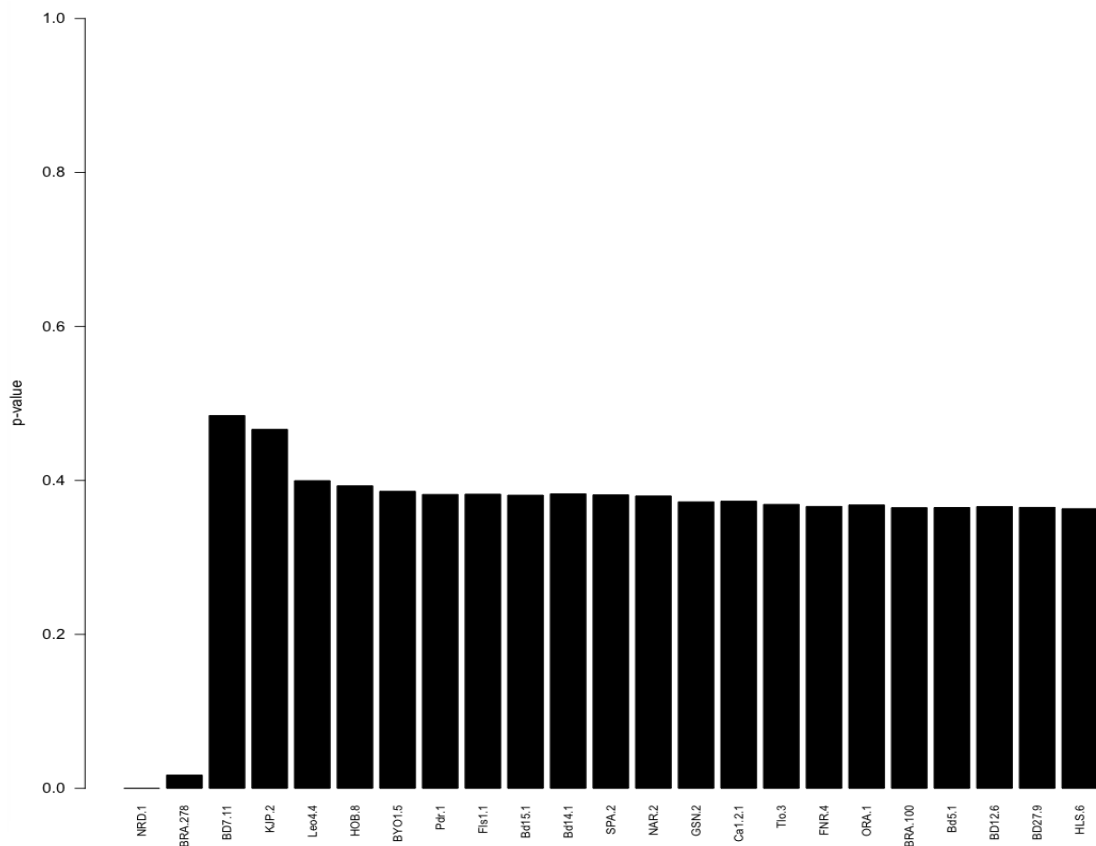
iteration was calculated and compared to the average number of genotypes present. The mean presence of a genotype was compared to mean average of all genotypes resulting in two genotypes with significant ( $0.05 > p\text{-value}$ ) more abundance across sites, NRD-1 and BRA-278, 0.000039 and 0.0172 respectively (See Figures 3.15 and 3.16). However, this test only shows what individuals are abundant across sites and does little to describe long-distance dispersal across regions so a regional test was also performed. Only NRD-1 was significantly present across regions with  $p\text{-value} = 3.550459e^{-09}$  as compared to the average genotypes found across regions. Note: In permutation figures common genotypes are not the same between regions and locations, because some genotypes are more abundant across local sites, while across regions some genotypes may be rare locally, but abundant across broad geography. Thus, the colours to denote genotype will not be consistent since some genotypes are not present in both categories.



**Figures 3.15.** Common genotypes permutation tested for presence across all geographic sites individually with mean and +/- 1 standard deviation. Genotypes NRD-1 and BRA-278 both had broad presence across individual geographic locations, NRD-1 being the larger of the two.

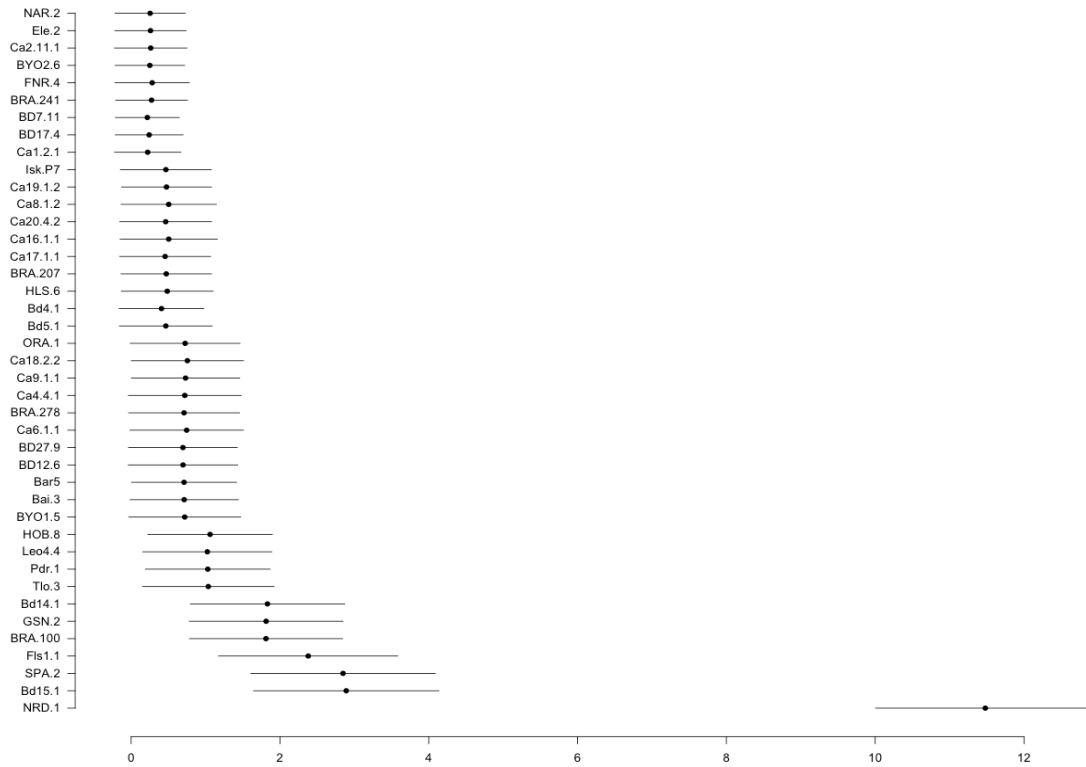


**Figures 3.16.** Common genotypes permutation test density plots for presence across all geographic sites individually. Genotypes NRD-1 and BRA-278 both had broad presence across individual geographic locations, NRD-1 being the larger of the two.

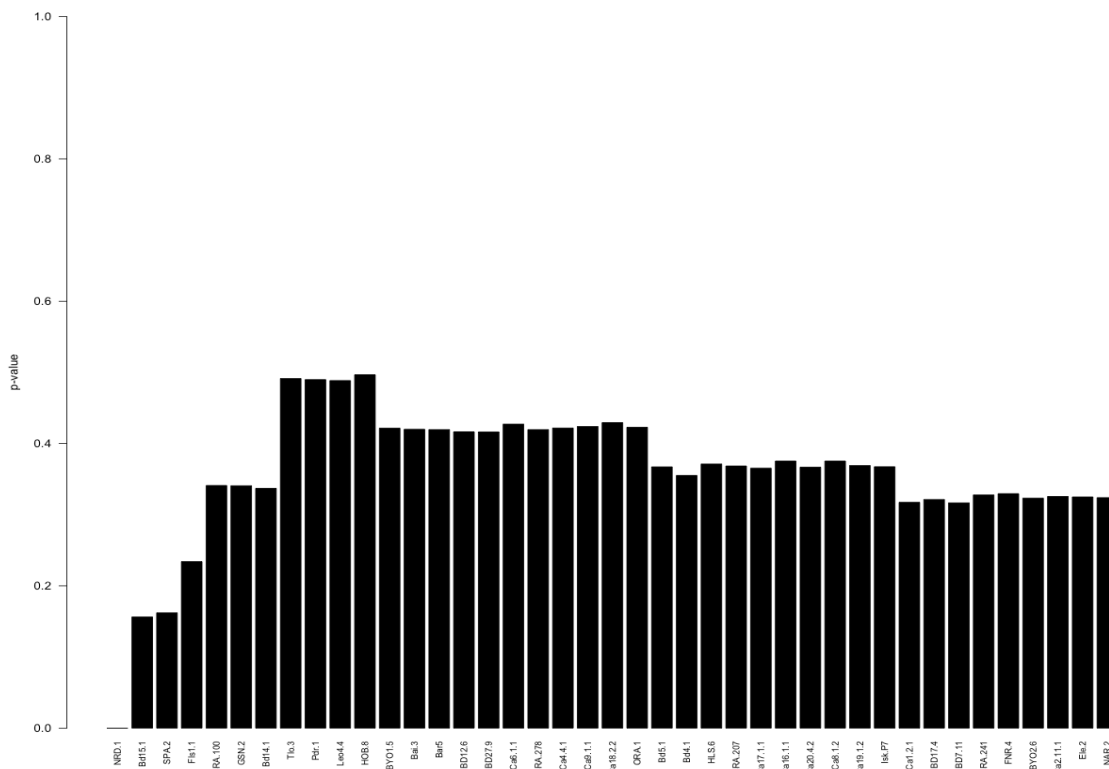


**Figure 3.17.** Permutation test p-values per each genotype across geographic regions. NRD-1 is the only accession with a p-value of significance,  $0.01 > p$  as compared to the average presence of genotypes across regions.

Genotype Permutation Test Results Across 35 Global Regions of *B. hybridum*

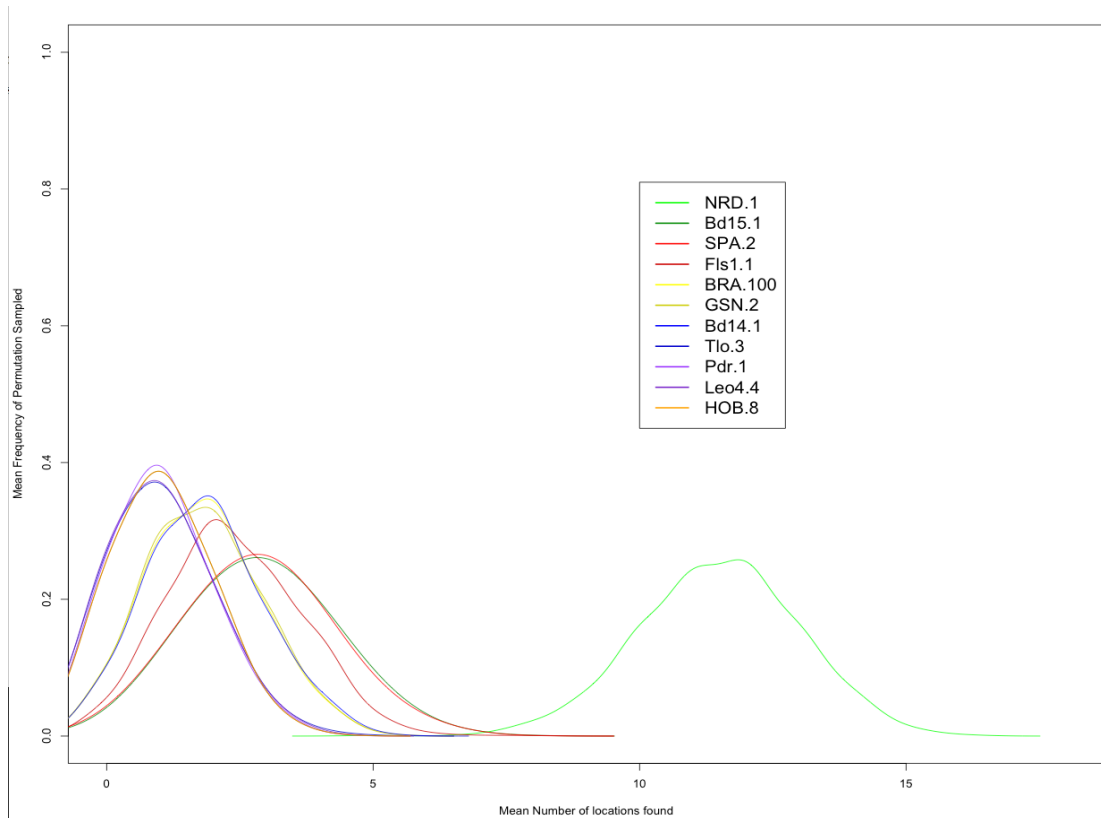


**Figure 3.18.** Permutation results for genotype presence in geographic regions. When calculating distribution across regions, NRD-1 was far more present than all other genotypes. To review regions and their description see Chapter II methods.



**Figure 3.19.** Permutation test p-values per each genotype across geographic regions. NRD-1 is the only accession with a p-value of significance,  $0.01 > p$  as compared to the average presence of genotypes across regions.





**Figure 3.20.** Permutation test density plots of the 11 most common genotypes for geographic abundance across 35 global regions. NRD-1 is the only genotype with a significant abundance compared to random sampling.

### 3.4 Discussion

Genetic analysis of each species reveals different patterns of diversity and dispersal across geography despite variation in sample size for each species. Having an individual reference genome for each species provides genotyping accuracy and ordered markers. *B. stacei* was rare in collection sites and little can be determined about its patterns of geographic diversity, while the other diploid, *B. distachyon* was much more common. *B. hybridum*, and specifically genotype NRD-1, was found globally, but multiple genotype introductions were identified on each continent. Since multiple genotypes were found traceable to the native range, a model that 'all introduced individuals are from a single dominant genotype' can be rejected.

#### Level and Resolution to call genotypes

With all large-scale genotyping studies a certain level of human error will occur, as highlighted in *A. thaliana* where sample locations could not be trusted (Anastasio, 2011). The genotyping accuracy required to call species, shown in Chapter II, is much lower than the same resolution needed for calling genotypes accurately. I used technical replicates, plants from seed of the same maternal plant, to determine how genotypes should cluster. In addition, accessions from the same site, were often biological replicates of the same genotype. This study's germplasm does have high overlap with previous publications of *B. distachyon* and trends in genetic placement in dendrograms is largely the same between publications in Turkey, Iberia, and other

larger studies of the circum Mediterranean ranges (Vogel, 2009; Feliz, 2009; Mur, 2011; Tyler, 2014; Tyler, 2016).

The ability to call genotypes in this study is based off of the highest branch length in a dendrogram of technical replicated samples. Samples also varied in sequence depth and total markers detected. The introduction of new more diverse samples will have further effects on marker calling, allele frequency, and overall diversity. The new tool SNPrelate is an ideal software package to show the branching quality across a dendrogram via z-scores. The user can set the z-score threshold to call genotypes and the higher quality the data is the lower the z-score can call genetically unique individuals (Zheng, 2012). Various ways have been developed to more accurately mine and weight useful markers and alleles. Though still in development, I created a genotype calling function in R that finds the two most diverged samples and orders all remaining samples between them. Then the software will call genotype based on a user set threshold of percent difference using only the shared markers between two samples. The script will keep scanning for percent difference across samples until the specified threshold is broken. The samples within the threshold will be assigned a unique genotype ID. Once a sample breaks the percent difference threshold it becomes the next most divergent individual and all others are ordered between them. This process continues until all samples are called to genotype. This package will provide statistics and plots that characterised the average distance, the number of markers typically used and so forth. It is my intention to release this tool with many others in an R package that reads standard VCF files and is currently being developed under the name “GenoCLIM”. GenoCLIM will also have many of the R features and functions used in this thesis.

#### Conclusions about *Brachypodium distachyon*

*B. distachyon* had the most genotypes per accession. It should be noted that *B. distachyon* has also been the primary target for collection efforts by research scientists working with *Brachypodium* species and many collection sites had few replicates, often only having a single sample. In future collections of *B. distachyon* it would be ideal to have more samples collected per location to identify rare genotypes including admixed samples to determine outcrossing rates. Resampling at sites now known to contain multiple genotypes and/or species would be particularly useful to identify inter and intraspecific hybridization. The deep population structure seen in *B. distachyon* makes identifying adaptive loci apart from neutral background loci difficult, as alleles are generally not segregating across geographic boundaries. With little outcrossing, genetic information is often migrating at a whole genome level. This is further confounded with many of the samples being long-standing inbred lines. Since outcrossing rates are low, native locations to study adaptation across climate and geographic boundaries will have to be in targeted regions where populations have already interbred. Or by synthetic field

experiments that introduce recombinant genotypes. Within local areas of both Spain and Turkey, previous studies show similar lack of recombinant genetic diversity seen in the *B. distachyon* here (Mur, 2011; Vogel, 2009).

#### Conclusions about *Brachypodium stacei*

Due to the rareness and under representation of *B. stacei* this study has limited ability to describe the diversity of the species. Since *B. stacei* is a small genome diploid species, now has a reference genome, and has been found to hybridise with *B. distachyon*, more collections of *B. stacei* would be ideal to better understand its genetic history and diversity. Species distribution modelling and predicting of *B. stacei* will be covered in the next chapter and will highlight new possible collection locations. Currently the only genetic diversity papers to cover *B. stacei* is Shiposha, 2016 and Tyler, 2016. The first study had samples from the western Mediterranean and Canary Islands in the Atlantic (Shiposha, 2016). Tyler *et al.* had a handful of samples from the central European areas on the island of Sicily. Combining the samples and known diversity in these two studies with this work would be a landmark achievement to further bring *B. stacei* to the model species level as there are very few known *B. stacei* in public repository germplasms. As discussed in the next two chapters we are finding out more and more about where *B. stacei* grows and its climate preferences.

#### Conclusions about *Brachypodium hybridum*

*Brachypodium hybridum* was found to have the fewest number of genotypes per accession of the three species. However, being an allotetraploid of two closely related species to *B. stacei* and *B. distachyon*, this could be expected as polyploidy itself is an isolating mechanism. A certain amount of diversity already existed in both *Brachypodium* diploid species before hybridisation. Since mutations rates are often higher in polyploid organisms one possibility is that the rate of genetic mutations increased once the two donor species formed *B. hybridum* (Dubcovsky, J., & Dvorak, J., 2007; Otto, S. P., 2007). Alternatively multiple polyploidisation events could have occurred and subsequently the different *B. hybridum* lineages could have recombined to generate novel recombinant genotypes. These alternative histories have different effects on the site frequency spectrum. The hybridisation event creating the polyploid is a hot topic in *Brachypodium* studies. Here I report that the hybridization event likely has two different events with close living relatives in the public germplasms common in *Brachypodium* research. The biogeography of these lines in relation to *B. stacei* lines could help untangle the creation of the complex's polyploid. In the Appendix section is a dendrogram of mitochondrial markers generated using the GBS data in this study to see how known diploids align with polyploids using mitochondrial markers against the wheat mitochondrial genome, figure S6.2. This thesis was not designed to conclusively identify genetic components of this hybridisation event, but there is strong evidence to say it has happened more than once and similar evidence

was reported previously on two occasions (Catalan, 2012; Tyler *et. al*, 2016). Based on this evidence the subgenome origins should be investigated more thoroughly. The hypothesis of *B. hybridum* being more abundant than other species is accepted by presence across continents.

#### Widespread genotypes

Permutation tests of genotype presence across geographic regions show genotype NRD-1 as the most widespread lineage of *B. hybridum* and all complex members with significant p-values compared to average presence of genotypes ( $0.01 > p\text{-value}$ ). NRD-1 is present in 21 regions by frequency and in permutation averaged 13 regions. Compared to all accessions NRD-1 was many times the standard deviation compared to the average presence across regions of all other genotypes. The exact mechanism phenotypically that makes NRD-1 such a wide disperser is not known, however should be investigated. The hypothesis that specific lineages of *B. hybridum* have more abundance than random is accepted by statistical significance.

#### Areas of High Diversity

To our current understanding, the region presently known as Israel is a genetic hotspot for both *B. stacei* and *B. hybridum*. However, that does not mean these regions harbour the most diversity and other locations should be investigated. Southern Spain in Andalucía is also a genetic hotspot for *B. hybridum*. *B. distachyon* had two known regions of high biodiversity being in present day Iberia and Turkey. Topics about distribution and future collections will also be discussed in the next chapter. Regions in northeast Spain showed the most likely locations of higher outcrossing rates based on genotypes being found in multiple locations and shared alleles between individuals. The Adi-*n* (Adiyaman, Turkey) location had significant amounts of genetic diversity, as well as the Bdis25-*n* and the Bdis23-*n* locations in northeast Spain. Non-native regions near Adelaide in Australia showed lots of genetic diversity within collection locations for *B. hybridum* and many locations towards the Flinders Ranges. In California 70% (14 of the 20) locations are mixed genotype location with only one spot not having a common genotype. Thus multiple introductions have occurred in Australia and California.

#### Geographic distance to Genetic Distance

Geography didn't explain much of the genetic diversity within *B. distachyon* or *B. hybridum*. This could be because some genotypes of either species are traversing a large amount of space. For *B. hybridum*, the geographic distance explained slightly more genetic variation, but both species scored very low compared to other studies. Even *A. thaliana* has significant isolation by distance in the native range, but is also a rapid cycling species with many weed like traits (Platt, 2010). Landscape studies with tree species with significant numbers of individuals often have significant isolation by distance as discussed in Chapter I with *Eucalyptus glaberrima*, *Eucalyptus tricarpa*, and to a lesser extent *Eucalyptus gomphocephala* (Yeoh, 2011; Bradbury,

2013; Steane, 2014). However, other studies in grass species found little genetic diversity was explained by geographic distance. Rice species, while could be culturally isolated, had little geographic isolation by distance and while crossable *Oryza sativa* and *Oryza indica* don't appear to have any admixture between lineages (Zhao, 2011). A similar story was found in *Setaria viridis* (wild close relative) and *Setaria italica* (domesticated) (Huang, 2014). In the two *Setaria* species many admixed samples were found between species, but ancestral lineages were widely dispersed similarly to both *Brachypodium* species and *Oryza* species. Meaning that the grasses likely have a more unique life strategy than conventional landscape genomics stories compared to other weedy short-lived species or even long-lived species: grasses are wind-pollinated species that often are stable selfers; they make many seeds per generation; they can have both annual and perennial life strategies; and high plasticity in growth. This may not always be the case, as isolation by distance would scale with the age of the species and size of the fundamental niche. In a species like switchgrass (*Panicum virgatum*) with a broad geographic range in North America, lineages dispersed large distances, but isolation by distance was still found do to an expansive range (Grabowski, 2011). However, within isolated regions occupied by specific ancestral groups, *P. virgatum* had little correlation of genetic diversity to geographic distance.

#### Compare Results to Other Species

*Brachypodium* species are a unique look at invasion biology systems; having overlapping and different traits and life strategies to other commonly researched species. *Arabidopsis thaliana* is a common weedy species in temperate disturbed habitats, especially in North America (Platt, 2010, Anastasia, 2011) The long standing interest in the scientific community in *A. thaliana* pushed collection efforts from many locations, which lead to natural variation studies and landscape genomics (Mitchell-Olds, 2001; Tonsor. 2005; Ågren and Schemske 2012; Ågren, 2013). Within the native ranges of *A. thaliana*, the environmental variation across sample locations could predict patterns of polymorphisms across the whole genome as described in one study of *A. thaliana* (Lee & Mitchell-Olds, 2012). Polymorphisms were also predicted based on genomic structure and composition; and that environmentally relevant factors contribute to population divergence across populations, and locally adapted genotypes. Polymorphisms are slightly correlated with geography in *B. distachyon* and *B. hybridum*, 4.7% and 8.97% respectively. However, *Brachypodium* species have been shown to disperse long distances and rarely outcross (Del Aqua, 2014; Neji, 2015). The measures of outcrossing rates in *B. hybridum* are likely biased in publication since only one paper measures outcrossing and found it to be low. Other geographic locations could have higher outcrossing rates. The use of study material in this thesis is often inbred lines that were developed for more general purposes of phenotypes than measures of genetic diversity nearly a decade ago. Forcing natural populations to inbreed for many generations will not resemble natural phenomenon. The exact reason why some

genotypes of *B. hybridum* are dispersing farther than others is not known and is likely related to a phenotype. Growth traits should be measured to see if seed quantity is a mappable trait, and ideally achieved through whole genome sequencing. While it is possible this dispersal trait is actually based on environment, it is likely physical. In the case of *A. thaliana* an early stop codon in a methylation transferase CMT2 allele conveying a larger climate tolerance in those *A. thaliana* individuals carrying that allele (Shen, 2014). If the isolation is by environment in *B. hybridum* then that could be a cause of some lineages dispersing farther.

A substantial genomic diversity and population genetic study using 273 individuals showed the structure of *S. viridis* and *S. italica* across Europe, Asia, and North America, with outlier locations in South America (Huang, 2014). While no specific genetic correlations were associated to climate, the two model species grow in vastly different ecological environments. Like *B. distachyon* and *B. hybridum*, *S. viridis* showed little isolation by distance despite having multiple ancestral groups. Many *Setaria* species are millet crops, and thus could have been dispersed by nomadic people, and since *Brachypodium* species are often found near agriculture, they could also have migrated as seed contaminants in nomadic and modern crops.

#### Other Genomic Features of Brachypodium in the Appendix Section: Population Structure and LD

The Appendix section of this thesis shows the linkage disequilibrium of *B. hybridum* and *B. distachyon* (See figures S3.13 to S3.20). The population structure analysis was not a prime focus of this thesis and the scope of genetic analysis is at the current standing genetic variation rather than the ancestral history. However, these details will briefly be discussed. The popular program STRUCTURE v.2.3.4 was used to determine the ancestral groups of *B. hybridum* subgenomes. The S genome was calculated to have optimal K at  $K=2$  using the Evanno, 2005 method and the D subgenome was  $K=3$ . This difference gives more evidence that there are indeed multiple origins of *B. hybridum* as seen in previously published hapmap studies using chloroplast markers and also matches an alignment of all Complex members against the *Triticum aestivum* mtGenome in the Appendix figure S3.7 (Catalan, 2012). Linkage disequilibrium was also attempted on *B. hybridum* and was calculated to  $\approx 50\text{kb}$  across all samples for both subgenomes. The overall results for LD in *B. hybridum* subgenomes were difficult to calculate and 50kb is a very loose approximation. The SNP density of *B. hybridum* averages at 27,638 bases and LD dropped within one window motion to  $R^2 = 0.10$  using seven averaged windows of 200kb (125kb, 150kb, 175kb, 200kb, 225kb, 250kb, and 275). The quick decay, but low genetic diversity of *B. hybridum* hits an interesting spot between few alleles per LD block, yet high enough outcrossing to show quicker LD decay than *B. distachyon* by an order of magnitude. The quicker decay in *B. hybridum* could also be from more recently wild collected material while much of the *B. distachyon* germplasm has undergone many rounds of selfing (Vogel, 2009). Whole genome sequencing of a genetically diverse subset of individuals would be ideal to resolve LD decay and the true bases per variant across the *B. hybridum*

subgenomes. This study's marker density for *B. distachyon* averages about 18,426 bases per loci as seen in the Appendix section (Figures S3.19 and S3.20).

Linkage disequilibrium for *B. distachyon* was calculated to decay to 0.10 via  $R^2$  at  $\approx 320$ kb using seven different sized sliding windows averaged together with a mean length of 200kb (125kb-275kb, like *B. hybridum*). This leaves few markers within long LD blocks to calculate associations to climate or phenotype via GBS. Also, many of the lines used are previously developed homozygous lines that are eight-plus generations selfed and will have little reflection of true heterozygosity or true representatives of local diversity. Whole genome data would also help to resolve the true bases per SNP for *B. distachyon*. The shorter LD in *B. distachyon* compared to other studies is slightly alarming as LD was found to be longer in previous publications (800kb to over 5mb to reach  $R^2$  of 0.1 using similar sliding window sizes) and within genetically distinct subgroups (Tyler, 2016; Wilson, Streich, and Murray, 2017). Population structure of *B. distachyon* was also investigated and is summarised in the Appendix section figure S3.22. Population structure was also calculated in a recent *Brachypodium distachyon* pan genome publication with an Evanno's method  $\Delta K$  of  $K=3$  (Gordon, 2017). In the Appendix section of this thesis the  $\Delta K = 4$ , however this thesis uses nearly five times as many samples. It can further be noted that in another study the  $\Delta K$  of *B. distachyon* was also  $K=3$  and was from a diverse subset where the near clonal individuals of each lineage were removed such that specific groups were not oversampled (Wilson, Streich, and Murray, 2017). The true  $\Delta K$  for *B. distachyon* is likely between  $K=3$  and  $K=4$ , however the actual population structure of *Brachypodium* species is not within the scope of this study. Lastly, the SNP density across chromosomes of *B. distachyon* was also calculated and featured in the Appendix section (See figures S3.8-S3.12).

### 3.5 Data Sets and Script Links

---

#### GitHub Repository

A source code repository was created for the scripts and data sets in this chapter. The script for assessing species category is listed in the online repository GBSFilter at the Borevitz GitHub webpage. Each script will require their directory edited to read files. Raw figures were generated from these scripts and edited in a variety of software: power point, preview, and etcetera. Raw data can also be found in this same repository and specifically in links provided below.

#### **Repository**

- <https://github.com/borevitzlab/GBSFilter>.

#### **Per Species**

##### *B. distachyon* script and data

[https://github.com/borevitzlab/GBSFilter/blob/master/Streichj\\_PhD\\_ANU\\_BorevitzLab\\_Genotyping\\_Bdistachyon.R](https://github.com/borevitzlab/GBSFilter/blob/master/Streichj_PhD_ANU_BorevitzLab_Genotyping_Bdistachyon.R)  
<https://github.com/borevitzlab/GBSFilter/blob/master/kmeansDistachyon.txt.zip>

### B. stacei script and data

[https://github.com/borevitzlab/GBSFilteR/blob/master/Streichj\\_PhD\\_ANU\\_BorevitzLab\\_Genotyping\\_Bstacei.R](https://github.com/borevitzlab/GBSFilteR/blob/master/Streichj_PhD_ANU_BorevitzLab_Genotyping_Bstacei.R)  
[https://github.com/borevitzlab/GBSFilteR/blob/master/streichj\\_Stacei\\_hapmap.txt](https://github.com/borevitzlab/GBSFilteR/blob/master/streichj_Stacei_hapmap.txt)

### B. hybridum script and data

[https://github.com/borevitzlab/GBSFilteR/blob/master/Streichj\\_PhD\\_ANU\\_BorevitzLab\\_Genotyping\\_Bhybridum.R](https://github.com/borevitzlab/GBSFilteR/blob/master/Streichj_PhD_ANU_BorevitzLab_Genotyping_Bhybridum.R)  
<https://github.com/borevitzlab/GBSFilteR/blob/master/kmeansHybridumH.txt.zip>

## 3.6. Citation

---

Agriculture and Consumer Protection - ACP. (2016). Dimensions of need - Staple foods: What do people eat?. Food and Agriculture Organization of the United Nations, Retrieved July 27, 2016, from <http://www.fao.org/docrep/u8480e/u8480e07.htm>

Allendorf, F. W., & Lundquist, L. L. (2003). Introduction: population biology, evolution, and control of invasive species. *Conservation Biology*, 17(1), 24-30.

Alves, S. C., Worland, B., Thole, V., Snape, J. W., Bevan, M. W., & Vain, P. (2009). A protocol for Agrobacterium-mediated transformation of *Brachypodium distachyon* community standard line Bd21. *Nature protocols*, 4(5), 638.

Australian Bureau of Statistics 2012 Invasive Plant Statistics. (2012). Land And Biodiversity, Environment. 1301.0 - Year Book Australia, 2012. Release Date : 24/05/2012.  
<http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1301.0~2012~Main%20Features~Land%20and%20biodiversity~278>

Bakker, Erica G., Brooke Montgomery, Tracy Nguyen, Kathleen Eide, Jeff Chang, Todd C. Mockler, Aaron Liston, Eric W. Seabloom, and Elizabeth T. Borer. (2009). "Strong population structure characterizes weediness gene evolution in the invasive grass species *Brachypodium distachyon*." *Molecular ecology* 18, no. 12: 2588-2601.

Barrett, P. R. F., Murphy, K. J., & Wade, P. M. (1990). The management of aquatic weeds. *The management of aquatic weeds.*, 473-490.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

Bragg, J.N., Wu, J., Gordon, S.P., Guttman, M.E., Thilmony, R., Lazo, G.R., Gu, Y.Q. and Vogel, J.P., (2012). Generation and characterization of the Western Regional Research Center *Brachypodium* T-DNA insertional mutant collection. *PLoS One*, 7(9), p.e41916.

Bragg, J. G., Supple, M. A., Andrew, R. L., & Borevitz, J. O. (2015). Genomic variation across landscapes: insights and applications. *New Phytologist*, 207(4), 953-967.

Carlton, J. (2003). *Invasive species: vectors and management strategies*. Island Press.

Catalán, P., Kellogg, E. A., & Olmstead, R. G. (1997). Phylogeny of Poaceae Subfamily Pooideae Based on ChloroplastndhF Gene Sequences. *Molecular phylogenetics and evolution*, 8(2), 150-166.

Catalán, P., Müller, J., Hasterok, R., Jenkins, G., Mur, L. A., Langdon, T., López-Alvarez, D. (2012). Evolution and taxonomic split of the model grass *Brachypodium distachyon*. *Annals of Botany*, 109(2), 385-405.

Draper, J., Mur, L. A., Jenkins, G., Ghosh-Biswas, G. C., Bablak, P., Hasterok, R., & Routledge, A. P. (2001). *Brachypodium distachyon*. A new model system for functional genomics in grasses. *Plant physiology*, 127(4), 1539-1555.

Daehler, C. C. (1998). The taxonomic distribution of invasive angiosperm plants: ecological insights and comparison to agricultural weeds. *Biological Conservation*, 84(2), 167-180.



- De Kort, H., Vandepitte, K., Bruun, H. H., Closset-Kopp, D., Honnay, O., & Mergeay, J. (2014). Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the tree species *Alnus glutinosa*. *Molecular Ecology*, *23*(19), 4709-4721.
- Dubcovsky, J., & Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, *316*(5833), 1862-1866.
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS one*, *6*(5), e19379.
- Fitzgerald, T. L., Powell, J. J., Schneebeli, K., Hsia, M. M., Gardiner, D. M., Bragg, J. N., ... & Vogel, J. P. (2015). Brachypodium as an emerging model for cereal–pathogen interactions. *Annals of botany*, *115*(5), 717-731.
- Fraley, Chris, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation Technical Report No. 597, Department of Statistics, University of Washington.
- Gomez, L. D., Bristow, J. K., Statham, E. R., & McQueen-Mason, S. J. (2008). Analysis of saccharification in *Brachypodium distachyon* stems under mild conditions of hydrolysis. *Biotechnology for biofuels*, *1*(1), 15.
- Gordon, S. P., Priest, H., Des Marais, D. L., Schackwitz, W., Figueroa, M., Martin, J., ... & Wang, W. (2014). Genome diversity in *Brachypodium distachyon*: deep sequencing of highly diverse inbred lines. *The Plant Journal*, *79*(3), 361-374.
- Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., ... & Martin, J. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature communications*, *8*(1), 2184.
- Hammami, R., Jouve, N., Soler, C., Frieiro, E., & González, J. M. (2014). Genetic diversity of SSR and ISSR markers in wild populations of *Brachypodium distachyon* and its close relatives *B. stacei* and *B. hybridum* (Poaceae). *Plant Systematics and Evolution*, *300*(9), 2029-2040. JGI, 2016
- Hasterok, R., Draper, J., & Jenkins, G. (2004). Laying the cytotoxic foundations of a new model grass, *Brachypodium distachyon* (L.) Beauv. *Chromosome Research*, *12*(4), 397-403.
- Hasterok, R., Marasek, A., Donnison, I.S., Armstead, I., Thomas, A., King, I.P., Wolny, E., Idziak, D., Draper, J. and Jenkins, G., (2006). Alignment of the genomes of *Brachypodium distachyon* and temperate cereals and grasses using bacterial artificial chromosome landing with fluorescence in situ hybridization. *Genetics*, *173*(1), pp.349-362.
- Huang, P., Feldman, M., Schroder, S., Bahri, B. A., Diao, X., Zhi, H., ... & Kellogg, E. A. (2014). Population genetics of *Setaria viridis*, a new model system. *Molecular ecology*, *23*(20), 4912-4925.
- Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, *4*(4), 336-344.
- Idziak, Dominika, Alexander Betekhtin, Elzbieta Wolny, Karolina Lesniewska, Jonathan Wright, Melanie Febrer, Michael W. Bevan, Glyn Jenkins, and Robert Hasterok. (2011). "Painting the chromosomes of *Brachypodium*—current status and future prospects." *Chromosoma* *120*, no. 5: 469-479.
- Lee, C. R., & Mitchell-Olds, T. (2012). Environmental adaptation contributes to gene polymorphism across the *Arabidopsis thaliana* genome. *Molecular biology and evolution*, *29*(12), 3721-3728.
- Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." *Bioinformatics* *25*.14 (2009): 1754-1760.

- Lichstein, J. W. (2007). Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology*, 188(2), 117-131.
- López-Alvarez, D., Manzaneda, A. J., Rey, P. J., Giraldo, P., Benavente, E., Allainguillaume, J., ... & Ezrati, S. (2015). Environmental niche variation and evolutionary diversification of the *Brachypodium distachyon* grass complex species in their native circum-Mediterranean range. *American journal of botany*, 102(7), 1073-1088.
- Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in ecology & evolution*, 18(4), 189-197.
- Manel, S., & Holderegger, R. (2013). Ten years of landscape genetics. *Trends in ecology & evolution*, 28(10), 614-621.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1), 209-220.
- Marriott, P.E., Sibout, R., Lapierre, C., Fangel, J.U., Willats, W.G., Hofte, H., Gómez, L.D. and McQueen-Mason, S.J., (2014). Range of cell-wall alterations enhance saccharification in *Brachypodium distachyon* mutants. *Proceedings of the National Academy of Sciences*, 111(40), pp.14601-14606.
- Mochida, K., & Shinozaki, K. (2013). Unlocking Triticeae genomics to sustainably feed the future. *Plant and Cell Physiology*, 54(12), 1931-1950.
- Mur, L. A., Allainguillaume, J., Catalán, P., Hasterok, R., Jenkins, G., Lesniewska, K., ... & Vogel, J. (2011). Exploiting the *Brachypodium* Tool Box in cereal and grass research. *New Phytologist*, 191(2), 334-347.
- Neji, M., Geuna, F., Gordon, S. P., Taamalli, W., Vogel, J. P., Ibrahim, Y., ... & Gandour, M. (2016). Insertion/Deletion markers for assessing the genetic variation and the spatial genetic structure of Tunisian *Brachypodium hybridum* populations. *Recent Research in Science and Technology*, 8, 14-23.
- Otto, S. P., 2007
- Opanowicz, M., Vain, P., Draper, J., Parker, D., & Doonan, J. H. (2008). *Brachypodium distachyon*: making hay with a wild grass. *Trends in plant science*, 13(4), 172-177.
- Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell*, 131(3), 452-462.
- Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Ågren, J., Bossdorf, O., Byers, D., Donohue, K. and Dunning, M., (2010). The scale of population structure in *Arabidopsis thaliana*. *PLoS genetics*, 6(2), p.e1000843.
- Pyšek, P. (1998). Is there a taxonomic pattern to plant invasions?. *Oikos*, 282-294.
- Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular ecology*, 24(17), 4348-4370.
- Sarwar, M. H., Sarwar, M. F., Sarwar, M., Qadri, N. A., & Moghal, S. (2013). The importance of cereals (Poaceae: Gramineae) nutrition in human health: A review. *Journal of cereals and oilseeds*, 4(3), 32-35.
- Shen, X., De Jonge, J., Forsberg, S. K., Pettersson, M. E., Sheng, Z., Hennig, L., & Carlborg, Ö. (2014). Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. *PLoS genetics*, 10(12), e1004842.
- Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10), 1135-1145.
- Simberloff, D. (2013). *Invasive species: what everyone needs to know*. Oxford University Press.

- Shiposha, V., Catalán, P., Olonova, M., & Marques, I. (2016). Genetic structure and diversity of the selfing model grass *Brachypodium stacei* (Poaceae) in Western Mediterranean: out of the Iberian Peninsula and into the islands. *PeerJ*, 4, e2407.
- Tilman, D., Balzer, C., Hill, J., & Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences*, 108(50), 20260-20264.
- Tyler, L., Lee, S. J., Young, N. D., Delulio, G. A., Benavente, E., Reagon, M., ... & Caicedo, A. L. (2016). Population structure in the model grass *Brachypodium distachyon* is highly correlated with flowering differences across broad geographic areas.
- Vogel, J. P., Tuna, M., Budak, H., Huo, N., Gu, Y. Q., & Steinwand, M. A. (2009). Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. *BMC plant biology*, 9(1), 1.
- Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., Bevan, M. W., ... & Tice, H. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, 463(7282), 763-768.
- Wilson, P. B., Streich, J. C., Murray, K. D., Eichten, S. R., Cheng, R., Aitken, N. C., ... & Borevitz, J. O. (2018). Population structure of the *Brachypodium* species complex and genome wide association of agronomic traits in response to climate. *bioRxiv*, 246074.
- United Nation. (2013). United Nations Report: World population projected to reach 9.6 billion by 2050, Department of Economic and Social Affairs, <http://www.un.org/en/development/desa/news/population/un-report-world-population-projected-to-reach-9-6-billion-by-2050.html>
- USDA Forest Service. (2016). Region 8. Invasive Species, US Forest Service. United States Department of Agriculture. <http://www.fs.usda.gov/detail/r8/forestgrasslandhealth/invasivespecies/?cid=stelprdb5326137>
- Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics* (2012); doi: 10.1093/bioinformatics/bts610

## Chapter IV: Genomic Biogeography

---

### Abstract

A large part of invasion biology is calculating vulnerable locations and the total surface area at risk from a given species. Computer modeling and analysis can predict and quantify both of these factors. Maximum Entropy modelling with the program MaxEnt was the choice method for finding suitable habitat for each member of the *Brachypodium distachyon* complex with my own R scripts used to calculate sensitive suitable surface area. Models were run for two reasons: identify potentially new collection locations in their native Mediterranean region, and secondly to identify globally sensitive areas in non-native regions. MaxEnt outputs suitability predictions across geography in probability format (zero to one), thus to calculate potential surface area a minimum threshold is needed to set binary presence/absence. This study used the popular method setting probability where sensitivity and specificity training thresholds are equal. In the native Mediterranean range *B. distachyon* had the largest potential area at 5,098,573 square kilometers in potential area, *B. stacei* had 2,458,837 square kilometers in potential area (similar to the global area planted in wheat), and *B. hybridum* had 3,935,266 square kilometers. In global models *B. distachyon* had 6,517,340 potential square kilometers globally with most of the potential area still in the native range, while most of the non-native suitable habitat in the northwest United States, south central and northeast China. The species *B. stacei* had 3,207,524 square kilometers globally, with 748,687 square kilometers in non-native habitat, mostly in coastal Angola in Africa. *B. hybridum* had 6,705,946 square kilometers globally with many regions in Australia, South America, North America and Southern Africa being suitable locations. *B. hybridum* had the largest non-native suitable habitat at 2,770,680 kilometers. Two genotypes of *B. hybridum* were observed on multiple continents and in high abundance to make global models to search for potential non-native habitats per each genotype. Eight *B. distachyon* genotypes and genotype families were also observed enough in multiple locations to make distribution models in study region centered on the country Turkey. Species level models were also combined to search for overlap between species with areas of the Southeastern Mediterranean and the Iberian Peninsula showing overlap of all three species in current conditions. Modelling of specific genetic lineages and the combining of models aids germplasm development for bio-diverse locations or unique habitats for future collections. Overall, *Brachypodium hybridum*, was had the largest suitable geographic area globally of the three species. Non-native areas in North America and Australia had nearly the same amount of unique genotypes as the native range. The genotype NRD-1 was modelled geographically for potential suitable habitats with many regions occupied, but many more without current records of introduction indicating that some of its extended niche is not yet realised.

## Chapter IV Outline

---

### 4.1 Introduction

### 4.2 Methods

- Niche Breadth prediction: Climate only Maximum Entropy Predictions

  - MaxEnt Settings

  - MaxEnt Outputs

  - Post MaxEnt R Analysis

- Region Clustering of Collection Sites

### 4.3 Results

- Native and Global Potential Area

- Overlap of potential area of Species Native Range

- Genotype Distributions

- Genetic diversity by geographic regions

### 4.4 Discussion

## 4.1 Introduction

---

The assessment and analysis of local environmental suitability for a given species was once a complex process that required replicated trials and reciprocal transplants of individuals from across gradients. In fact, for much of history, assessing habitat suitability was predominantly in agricultural systems relying on environmental cues. Natural phenomenon like last frost or first seasonal rain are examples of agronomic signifiers of critical decision-making time points for growers: when to plant or harvest, when to fertilise or apply specific treatments. Such agronomic techniques were suitable to the needs of farmers to optimise crop yield as plants are sensitive to their own set of environmental cues that greatly effect/affect germination, vegetative growth, flowering, and senescence, and is practiced today in simulated crop modelling (Mathews, 2013). The environment is both biotic and abiotic stresses, and calculating the environmental suitability of a species across broad geographic range requires many complex measurements. Knowing what affects environmental variables have can be cryptic even on a local scale. Thus, biologically relevant abiotic climate variables are the most reliable data for predicting the suitability of a local climate for a given species and can be performed by computer simulation.

Modern techniques to assess suitability of local environments use mathematical models and two input data sets: species observations in coordinates, and climate data from satellites and weather stations. Based on various statistical methods, the model will calculate species suitability across geography. This process is often called species distribution modelling (SDM) (Hijmans, 2005; Phillips, 2004; Phillips, 2008; Elith, 2011, Phillips 2005). The ratios and comparisons of abiotic climate variables across various monthly, seasonal and annual time scales are what most SDMs use to calculate potential suitability of a given geographic space. There are a multitude of programs to calculate species distribution and suitability across geography, but most commonly used is a program called MaxEnt due to its lack of bias compared to other SDM software.

### Prediction of Invasive Species Ranges

One of the difficulties of predicting novel ranges of introduced species is assuming its climate breadth is fixed to native habitat (Peterson, 2003). This likely is not the case as a species could be pre-adapted to paleo-climates, and/or their range has shifted with the previous fluctuations in changing global temperatures and even hybridising with relic groups as seen in *A. thaliana* (Sharbel, 2000; Lee, 2017). The interesting aspect of invasion biology is how an introduced species responds to novel climates beyond their climate breadth. Having variation to neutral fitness helps determine the true climate tolerances of that species, which previously would have been based on native ranges. There are many other model types for predicting invasiveness, but usually requires *a priori* knowledge of the species, like biotic interactions/sensitivities, phenotypic plasticity measurements across gradients, abundance, etc. In this thesis no prior knowledge was known about species phenotypes across gradients in non-native habitats. However, one study did show that *Brachypodium hybridum* has more phenotypic plasticity across climate gradients than diploid *B. distachyon* (Manzaneda, 2015). Also, Chapter V does investigate similarity in climate between native and non-native collection sites.

Predicting novel ranges of introduced species is challenging because it is a false assumption that a species climate breadth is fixed to its native habitat (Peterson, 2003). This likely is not the case as a species (or its direct lineages) could be previously adapted to paleo-climates, and/or their range has shifted with the previous fluctuations in changing global temperatures and even hybridising with relic groups as seen in *A. thaliana* (Sharbel, 2000; Lee, 2017). Therefore, input points from positive observations in non-native locations will greatly improve the predictive power of a species distribution model in both native and non-native ranges, and will also more accurately predict the breadth of a species climate tolerance. One of the interesting aspects of invasion biology is that having more locations beyond a species native range, that also have neutral to positive fitness, helps determine the true climate tolerances of that species. There are many other model types for predicting invasiveness, but usually requires previous calculations of plant density in various climate gradients and other knowledge of the species, like biotic interactions/sensitivities to other species with known ranges, phenotypic plasticity measurements across gradients, abundance/density, etc. In this thesis no prior knowledge was known about species phenotypes across gradients in non-native habitats. However, one study did show that *Brachypodium hybridum* has more phenotypic plasticity across climate gradients than diploid *B. distachyon* (Manzaneda, 2015). Also, Chapter V does investigate similarity in climate between native and non-native collection sites.

Rapid adaptation in introduced species has been characterised in publication and that range models of invasiveness often under predict the fundamental niche (all possible habitable geographic space regardless of presence) of a species. As previously discussed, an

anthropocentric assumption about the native range representing the fundamental niche is likely false. However, a publication reviewed nine different plant species for adaptation in non-native ranges, of the studies highlighted, introduced species often had phenotypic changes (Clements, 2011). Those traits analysed include: leaf shape, number and size increased; seeds often became larger, changes in perennial or annual life strategy; possible hybridisation with other species; and some had increased climate tolerances because the realised niche was now more descriptive of range and climate limits. The actual genetic causes in these studies are not carefully examined and most predate modern genomic analysis, but their phenotypic changes are still relevant. Some of the phenotype variation in these studies has been observed in *B. distachyon* where some lines have different flowering time, and variation in leaf traits (Vogel, 2009). It is possible that some of the non-native adaptation mentioned is from admixture of individuals from geographically isolated native regions, and that the introduced genotypes have outcrossed and created novel genotypes in the non-native ranges and should be investigated. After all, *A. thaliana* has multiple genotypes in non-native ranges (Platt, 2010). A separate study that did analyse both genetic association and phenotypic variation found that *Lithrum salicaria*, a common North American invasive, had adapted to flower sooner in shorter northern seasons than locations as far as 1,000km south (Coulatti, 2013). In the case of a self-fertile outbreeding invasive species like *L. salicaria*, it should be noted that if non-native adaptive phenotypes do arise, it could quickly spread to other individuals and increase the fundamental range of said species. It should be noted again that training a computer model to predict fundamental niche is always subject to the data set and outputs in this chapter are probably similar to the true fundamental niche, but would likely require a larger data set with phenotype and plant density information to properly estimate true fundamental niche of a plant species.

#### *Brachypodium distachyon* Complex Range Models

There are few studies describing the geographic ranges of *Brachypodium distachyon* complex species. A study described the likely suitable locations for the whole species complex and their likely realised niche, spanning much of Europe, Central Asia, Sub-Continental India, North Africa, and non-native locations of North America, southern Africa, parts of South America near Uruguay, and much of southern Australia (Garvin, 2008). The only definitive study calculated the likely native ranges and overlap of each complex species across various geologically recent timescales (Lopez-Alvarez, 2015). In this same publication the predicted ranges of *B. distachyon* and *B. stacei* (diploids) rarely overlap through most of calculable history. The predicted ranges of *B. hybridum* often overlaps with both *B. stacei* and *B. distachyon*. Interestingly there are few locations that were predicted for only diploids but not suitable for *B. hybridum*, and could be that the allotetraploid *B. hybridum* inherited most of the diploids ranges. Finally, that study also found that *B. hybridum* not only inhabits many of the same regions as either diploid, but that it expanded its geographic range post polyploidisation to

new regions from expanding its climate breadth. The expansion of climate breadth was also examined in Chapter V.

#### MaxEnt Statistics and Basic Function

MaxEnt uses Maximum Entropy Modeling concepts via machine learning to calculate the climate suitability of geographic space based on species observation data in digital geospatial coordinate format (Phillips, 2004). For environmental inputs the program uses geospatial matrix maps called raster layers (further described below and in the appendix glossary), the most common are precipitation and temperature climate data. The second input format is geographic coordinates of species observations. The basic premise is to calculate the upper and lower bounds of climate breadth variables based on observation points, then weight each variable's contribution within the model based on entropy.

#### Equal Sensitivity and Specificity Thresholding of MaxEnt to call Binary Suitability/Unsuitability

Setting binary suitability/unsuitability of regions with MaxEnt requires making a calculated assumption about the model's ability to find suitable locations. Commonly used in machine learning (decision trees and neural nets), as well as medical diagnostic accuracy studies, MaxEnt uses a sensitivity and specificity algorithm to calculate model performance. The MaxEnt prediction algorithm will classify data into a 2x2 table of two classes, positive and negative results, and false negative and false positive. (Phillips, 2004; Hajian-Tilaki, 2013). Normal distributions of both the positive and negative classifications are plotted by their respective predicted values, the overlap of each positive and negative class represents the rate of false classifications. However, MaxEnt presents the predicted negative and positive probability values from the model output as a curved line on an XY plot scaled zero to one on each axis, often called a receiver operation characteristic (ROC). An ROC is a two-dimensional XY ordinal plot, where the predicted p-values of the false positives are on the y-axis, and the predicted false negative p-values are on the x-axis (See Figure 4.3). The further the predicted p-values are from each other in both classes, the farther the ROC curve is from a slope of 0.5, indicating a well performed model. Since p-values range from zero to one, the maximum area reachable in an ROC plot is one, or the area under the curve (AUC), see figure 4.2. The AUC score is the probability that a classifier will rank a random positive observation higher than a random low observation. In most Diagnostic Accuracy Studies it is common to choose a threshold of 0.5, however the MaxEnt classifier is often more accurate than a standard threshold, and it can output the ideal threshold for overlapping tails of the positive and negative rates (Phillips, 2006; Elith, 2011). Thus a much lower rate can be used and is directly related to the AUC probability. MaxEnt does not default to show the two normal distributions since the ROC and AUC describe the model performance and the ideal threshold for calling a binary classifier with false positive and false negative information.

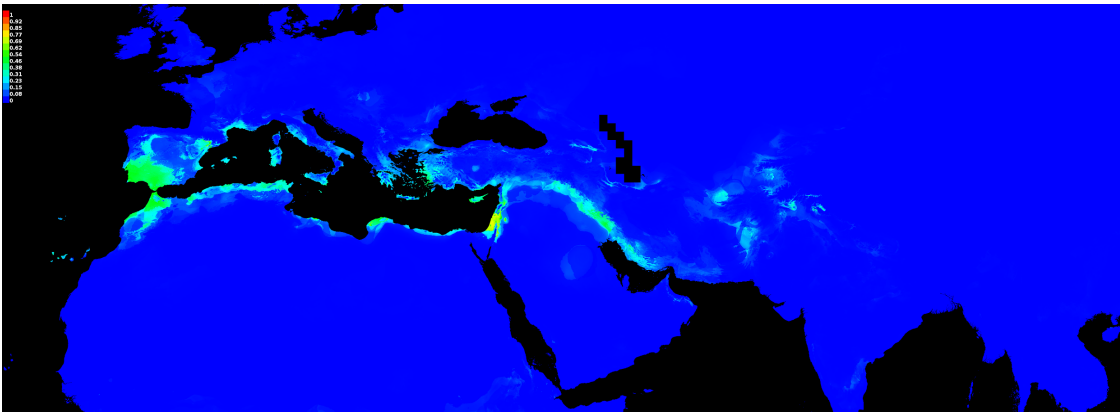


### Local and global Modelling

MaxEnt functions to find spatial trends based on observation data. When modeling suitable habitat to create species distributions across a species native range, it is important to frame the study area boundaries proximal to the observation points (Ficetola, 2007; Medley, 2010; Elith, 2011). The further the observation points are from the boundaries of the study area, the more likely a model will sample regions with diverse non-suitable climates in the model. Oversampling more climate variation in non-predicted regions overfills the predicted negative class and will augment the model and create biased environmental variable contributions (Elith, 2011; Warren, 2011). One way to overcoming a bias towards one set of variables over another is using a tool like Environmental Niche Modelling Tools, (ENMTools) (Warren, 2010). ENMTools can trim a model distribution based on the maximal and average dispersal distance from observation points, if known. Doing so will remove locations that are actually beyond the physical limits of the study species. In this thesis ENMTools was not used because the focus was finding potential suitable habitat per species and genotypes requiring global climate layers, the assumption being that if a species or genotype were to travel beyond its normal range, what locations have suitable climates. In the case of finding new suitable regions in the native Mediterranean ranges, study boundaries were drawn slightly larger than a previous study that used ENMTools, the goal was to find new native regions that could harbour *Brachypodium* species of interest (Lopez-Alvarez, 2015).

### Data Suitability Output

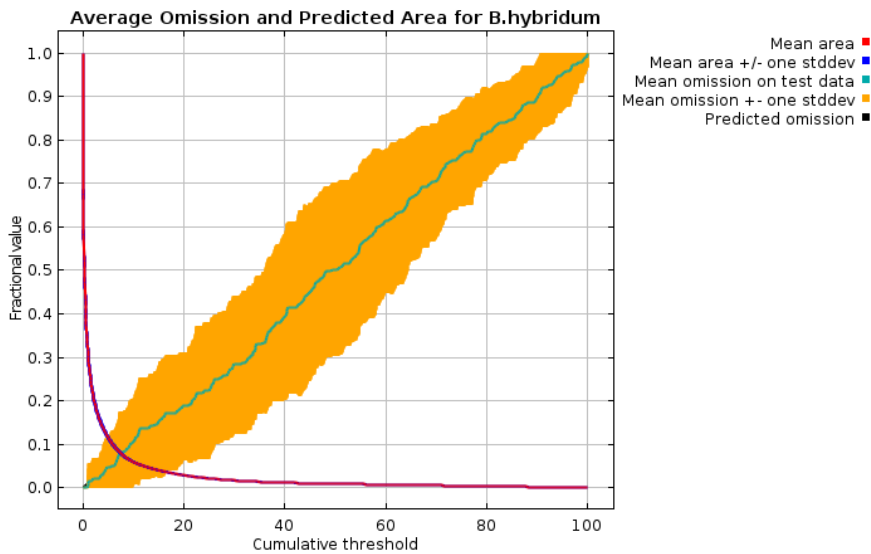
When performing global models it is important to assess what climate variables are describing suitability scores and that predicted suitable habitats compare in some way to climate data at the species observation locations, that is why MaxEnt can calculate the percent contribution of each variable to the model through entropy and jackknifing statistical methods (Veloz, 2009; Medley 2010). One of the most important outputs are the descriptive digital maps showing the predicted suitability in the context of probability space as .png format and ASCII raster layer (.asc) format (the required input format for MaxEnt raster layers). Bellow is an example of the typical geographic output of suitable space in the study area of the species *B. hybridum* native range (See Figure 4.1). The legend indicates the suitability score associated with specific colours that are in the model output.



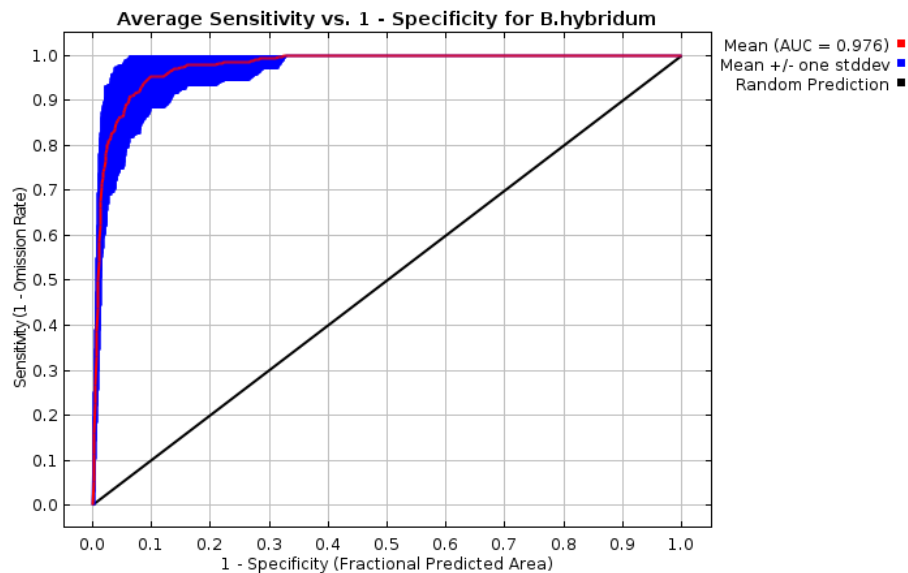
**Figure 4.1.** Example output from MaxEnt, Potential Suitable Area of *B. hybridum* Native Range: The above map indicates the geographic locations in the study area most likely to be suitable habitats for *B. hybridum* in units of probability. The suitability/probability and colour legend is in the upper left corner.

Average Omission and Predicted Area: ROC and AUC Plots

MaxEnt also outputs information about model accuracy and validity to indicate the quality and “comparable-ness” of the model to other models. MaxEnt outputs a graph plotting the comparison of species observation locations to predicted suitable area titled *Average Omission and predicted Area* (See figure 4.2 for an example for *B. hybridum* native range). Below is a graph describes the ratio of similarity between observation points and predicted suitable regions, and ideally there is a 1:1 ratio following a slope of 0.5 slope. The graph also shows the mean omission of +/- one unit of standard deviation. The Sensitivity and specificity of *B. hybridum* native model is also plotted below (Figure 4.3). Like standard deviation in the AUC plot, the ROC plot also averages each multiple models' mean AUC score across thresholds. In the ROC, the further the slope is from 0.5 the better the model performed indicating that the model's predicted area was far from similar omitted area.



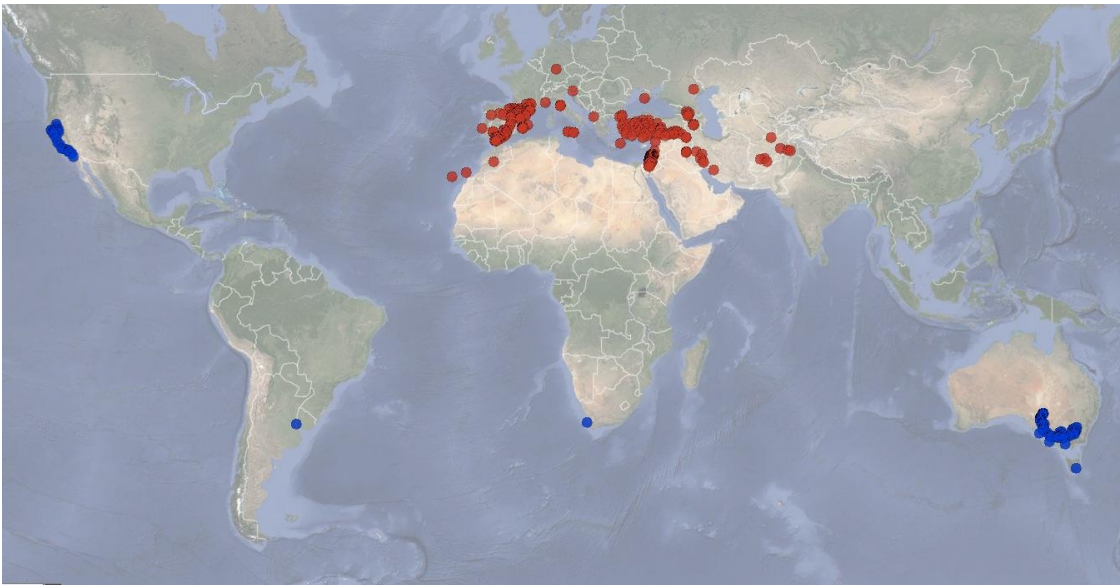
**Figure 4.2.** Example output from MaxEnt, The Average Omission and Predicted Area for *B. hybridum* Native Range: The comparison of input points to predicted suitability areas is forming a high quality 1:1 trend line across all cumulative thresholds. The standard deviation forms an ideal shape for an informative model.



**Figure 4.3.** Example output from MaxEnt, Average Sensitivity vs. 1- Specificity for *B. hybridum* Native Range: The comparison of background area of low suitability compared to predicted high suitability is found to be dissimilar across Specificity of Predicted area and not background points of low suitability are qualifying as high suitability. If the sensitivity vs. 1-specificity graph trends at a 0.5 slope then the predicted suitable habitat in the study area is random. The further from a 0.5 slope the more accurate the model is at correctly classifying false negatives. When both of these graphs show ideal trends we can start to understand how informative the model is, and can compare them to other models with different parameter settings.

#### WorldClim, BioClim: Environment and Climate Layers

WorldClim and BioClim are spatially interpolated digitised climate grids, sometimes referred to as ‘climate surfaces’ and are often the main sources of climate analysis of agricultural and environmental analysis (Hijmans, 2005). The data grids are freely downloadable for research use at worldclim.org. The datasets can be used as raster layers and input into analysis programs like QGIS, R, MaxEnt, and others with little to moderate effort for file format conversion. Each climate surface is composed from analytics across 50 years (1950-2000) of temperature, precipitation, and solar radiation within specified annual time ranges on a global scale from all continents except Antarctica. Climate surfaces are currently accurate to 30 arc seconds, or 1km resolution globally though some regions have higher resolution than others, United States and Australia as examples. Nineteen of the forty WorldClim climate surfaces are deemed biologically relevant, known as BioClim, and were used in this study for climate analysis. Definitions for all BioClim layers are in the glossary section of the Appendix. For further reading about BioClim and WorldClim layers visit: ([www.bioclim.org](http://www.bioclim.org)).



**Figure 4.4.** The 817 Collection locations of *Brachypodium distachyon* complex species, 488 were used in this study. Blue dots indicate a non-native collection location, red dots indicate a native collection location.

#### Questions, Aims, and Hypothesis of Chapter

The goal of this chapter is to search for suitable locations across geography via the SDM software MaxEnt in both native and non-native ranges for all three complex species. For species that have common genotypes, models were also created to predict suitable locations.

**Question:** *What are the regions with suitable climate across native and non-native geography for each *Brachypodium distachyon* complex species and do certain species or whole genome genotypes have larger ranges than would occur by chance?*

**Hypothesis:** *I hypothesize that since *B. hybridum* is a polyploid, and that it is known to demonstrate more phenotypic plasticity across native climate gradients, it will have a larger range and predicted surface area than diploid complex members. Further, some common genotypes of each species will have larger ranges than random.*

**Aim:** *Calculate the surface area of predicted suitable habitat of all samples of each species and common genotypes. Then compare the total surface area of native and non-native habitat to see what species and genotypes are more prevalent and have larger fundamentally suitable surface area.*

## 4.2 Methods

### Potential Area

The potential suitable area of each species was modeled using MaxEnt in both the native range and globally. Native range models are created to search for new locations to expand collection efforts and obtain new genetic material of each species. Global models were created to search for suitable habitat beyond the native range of each species and genotype with enough sampling. Earlier research calculated the current distribution of all three study species based on presence of samples and trimmed using ENMTools (Lopez, 2015). The background size to

model the native range, what was considered native space is derived from the Lopez, 2015 study, but the study area was expanded based on herbarium records to search for new collection locations (ALA, 2016; GBIF, 2016). Custom scripts in R that account for earth curvature performed the surface area calculations used in this study.

#### Potential Area Per Genotype

While other studies have examined distributions of specific genotypes of suits of genes, as of to date, we are unaware of any lab or research group modelling genotype distribution at this level of power to describe genotype, and relatedness amongst genotypes. It's possible that modelling a specific genotype distribution could be tested for environment specificity or preference by alleles contained within genotypes occupying those regions by resampling the predicted suitable habitats in the model. The ability to test genotype specific distribution models is still under development and discussed further in the discussion of this chapter and Chapter VI.

#### Settings for *B. distachyon* Genotypes

The study area for *B. distachyon* genotypes was trimmed to a square area surrounding the country currently known as Turkey and based on the combined observation points of all *B. distachyon* genotypes. MaxEnt model parameters were set to standard settings except: the number of replicates to average was set to 25, the number of permutations was set to 1,000, the calling threshold was set to *Equal Training Specificity and Sensitivity* to call binary suitability to non-suitability. All BioClim variables were included for *B. distachyon* genotype models.

#### Settings for *B. hybridum* Genotypes

Global models to predict potential suitable area for *B. hybridum* were un-cropped and framed at maximum potential space. MaxEnt model parameters were set to standard settings except: the number of replicates to average was set to 10, the number of permutations was set to 1,000, the calling threshold was set to *Specificity equals Sensitivity* to call binary suitability to non-suitability (Elith, 2011). All BioClim variables were included for *B. hybridum* genotype models and compared to NRD-1 models using random points to see if the distribution of widespread *B. hybridum* genotypes were larger than by chance. Methods to properly test this concept are still in development and were not included in this dissertation, but were discussed in Chapter VI.

### **4.3 Results**

---

#### Maximum Entropy Predictions for each Species, Native Range and Non-Native Range

The distribution for each species was modelled to compute the potential area in their native range to search for areas that could harbour individuals of each species, and to calculate potential surface area. The minimum suitability threshold was calculated by MaxEnt for each species and binary presence/absence was calculated by the equal specificity-sensitivity method (Elith, 2011). While any region close to the minimum threshold is unlikely to be suitable

geography it is deemed valid for potential habitable geography. Thus, any region above the minimum threshold was counted as suitable and used in area calculations. In theory, the fundamental niche of a species is larger than the realised niche. However, to truly gauge what of the three complex species has the most potential suitable geography, each species was investigated for surface area calculations to determine range size in both native and non-native locations. *B. distachyon* had the largest potential area by climate in its native range compared to all other species. Nearly all of the area of *B. distachyon* ( $\approx 78\%$ ) was found on its native range indicating that the relative climate variables that *B. distachyon* is sensitive to are isolated in western Eurasia. The area of *B. hybridum* probable fundamental niche is not as large as *B. distachyon* at  $\approx 3.9$  million km<sup>2</sup>, however geography with climate amenable to *B. hybridum* is common globally. *B. hybridum* *B. stacei* probable fundamental niche was the smallest of the three species on both global and native scales (See table 4.5).

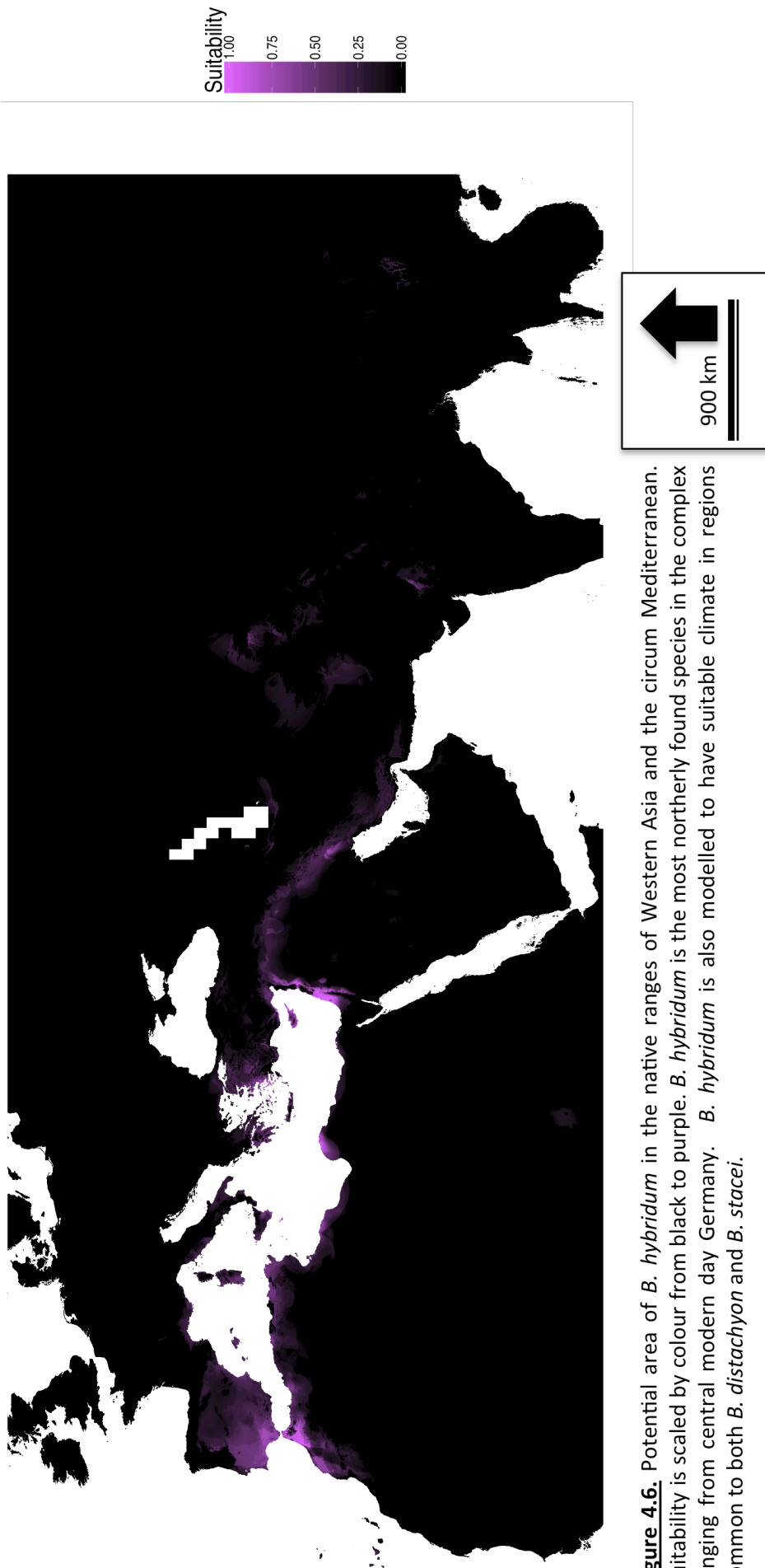
Species	Predicted Native Area km	Non-native Predicted Area	Global Predicted Area
<i>B. distachyon</i>	5,098,573	1,418,767	6,517,340
<i>B. stacei</i>	2,458,837	748,687	3,207,524
<i>B. hybridum</i>	3,935,266	2,770,680	6,705,946

**Table 4.5.** The potential suitable habitat of each species measured in km<sup>2</sup>: Each species was modelled in MaxEnt to show potential native and non-native habitats that could harbour *Brachypodium* species. *B. hybridum* had the largest surface area globally and in non-native ranges.

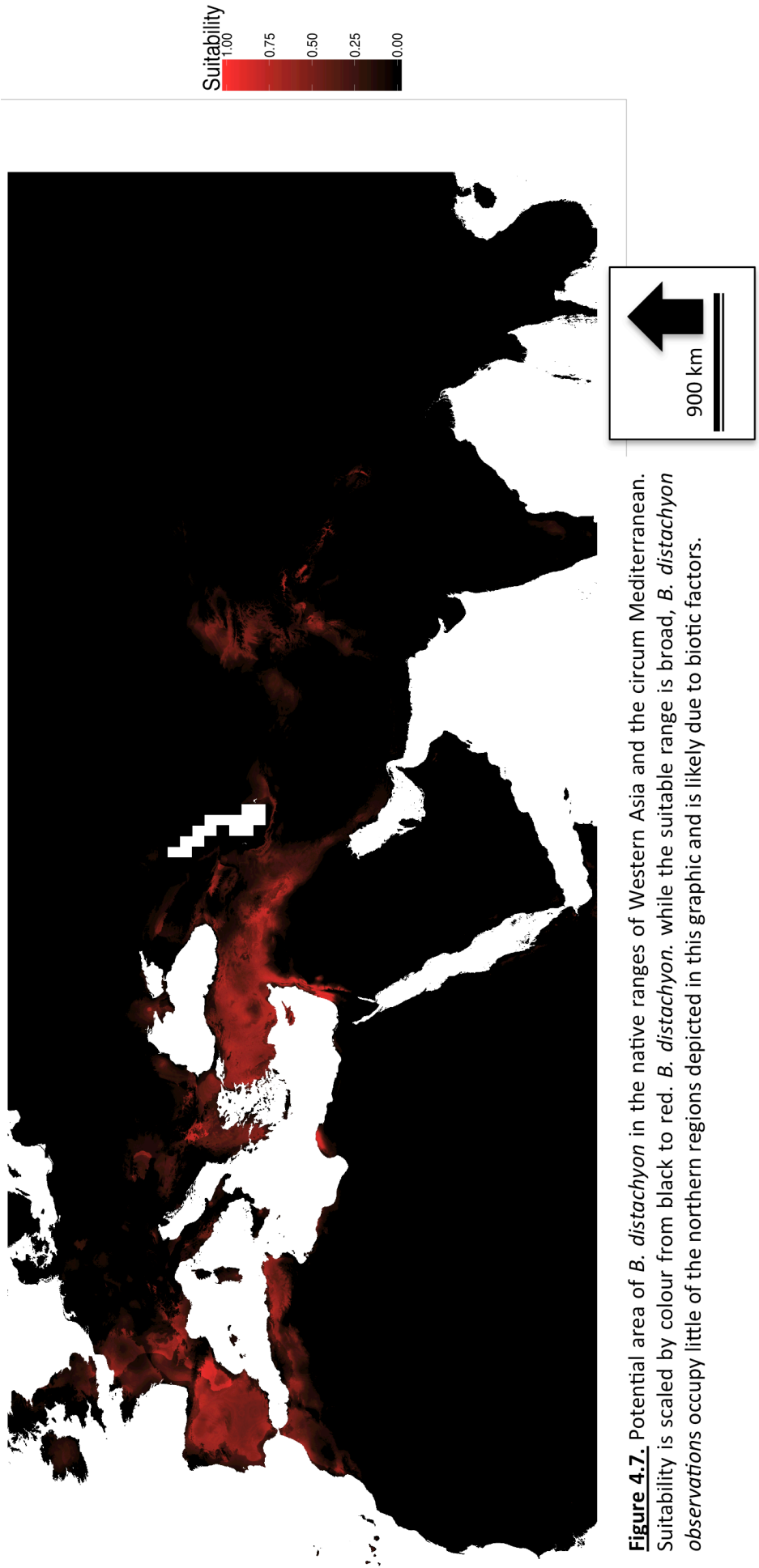
#### Maps of Potential Area of *Brachypodium distachyon* Complex Species and Overlapping Ranges

MaxEnt modelling outputs .ascii raster layers with suitability values in each pixel. *B. hybridum* native models show most of the suitable geography near the coastal areas of the Mediterranean sea, much of the Iberian Peninsula, parts of North Africa including modern day Liberia, Algeria, Morocco, and Tunisia. *B. hybridum* also has amenable climate in the Middle East and West Asia (See Figure 4.6). Native models show that suitable climate for *B. distachyon* does occur quite far north of the northern most *B. distachyon* in modern day France (See Figure 4.7). *B. stacei* is almost exclusively in coastal areas overlaps little with *B. distachyon*, but overlaps frequently with *B. hybridum* (See Figures 4.8, 4.9, and 4.10). Figures of geographic maps were calculated using base R and the 'raster' package (Hijmans, 2014).

The overlap of each species ranges, where they potentially co-occur was calculated in R using the 'raster' package (Hijmans, 2014). Independently calculated geographic ranges of each species can be overlaid and colour coded in regions of overlap. To simplify visualization of overlap, a different colour code was used to represent each species. RGB colours have ideal colour mixing to represent multiple combinations of three independent data sets. The colour scheme used for each species and their combination is: Blue = *B. distachyon*, Red = *B. stacei*, and Green = *B. hybridum*. Overlap of species is represented by between colours: Cyan = *B. hybridum* + *B. distachyon*, Yellow = *B. hybridum* + *B. stacei*, Magenta (Rare) = *B. stacei* + *B. distachyon*, Purple All Species. Both binary presence/absence from equal sensitivity thresholding and linear gradients were plotted (See Figures 4.9-4.11).

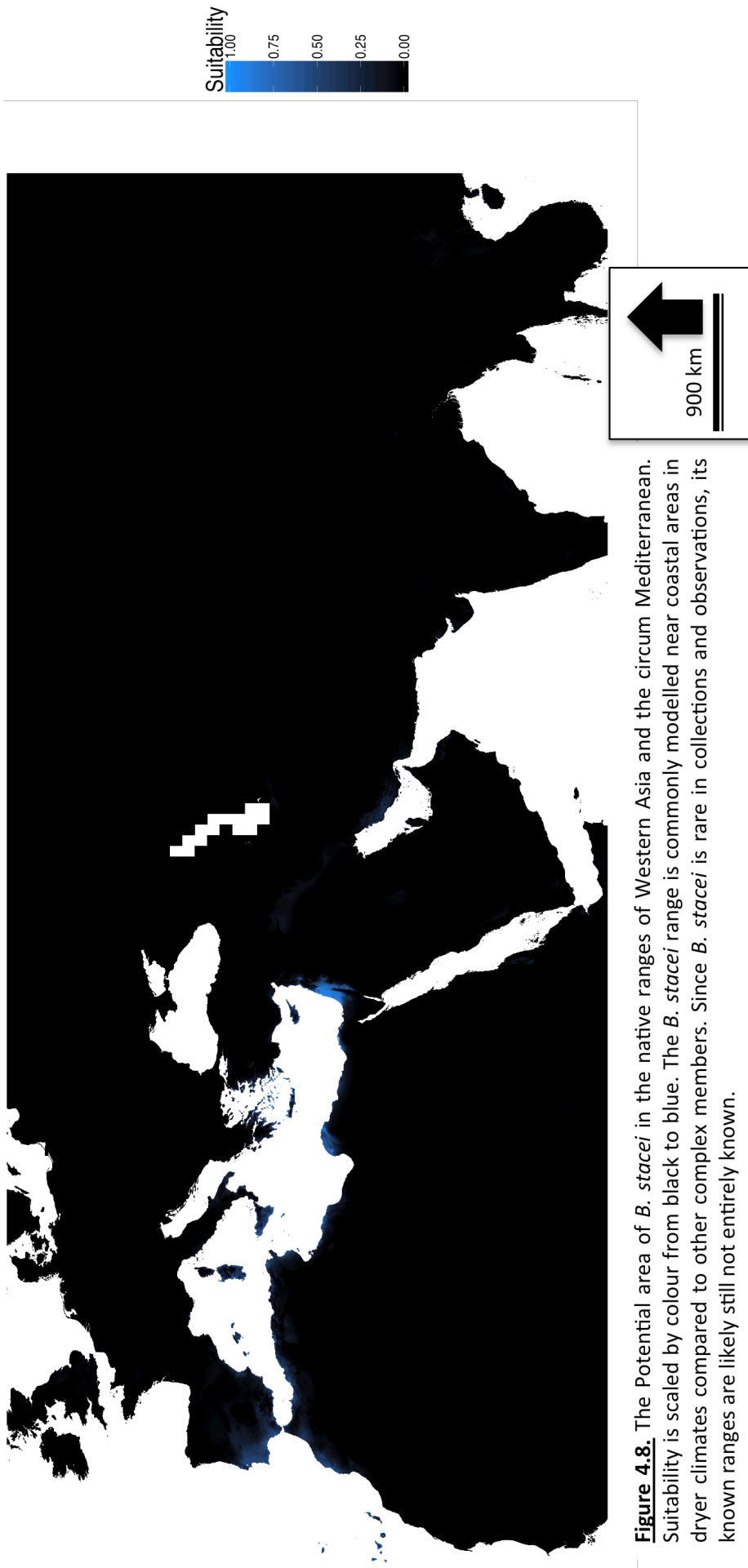


**Figure 4.6.** Potential area of *B. hybridum* in the native ranges of Western Asia and the circum Mediterranean. Suitability is scaled by colour from black to purple. *B. hybridum* is the most northerly found species in the complex ranging from central modern day Germany. *B. hybridum* is also modelled to have suitable climate in regions common to both *B. distachyon* and *B. stacei*.

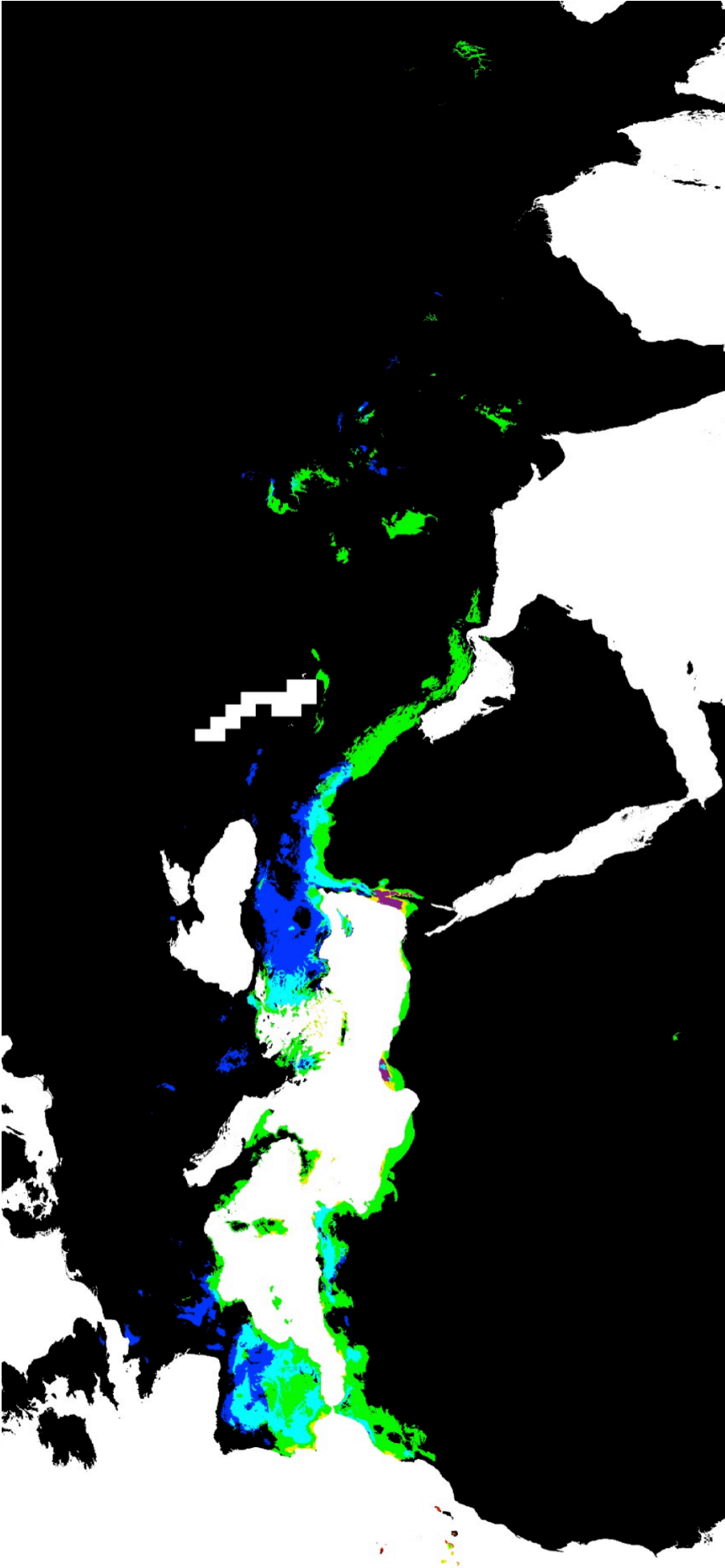


**Figure 4.7.** Potential area of *B. distachyon* in the native ranges of Western Asia and the circum Mediterranean. Suitability is scaled by colour from black to red. *B. distachyon*. while the suitable range is broad, *B. distachyon* observations occupy little of the northern regions depicted in this graphic and is likely due to biotic factors.

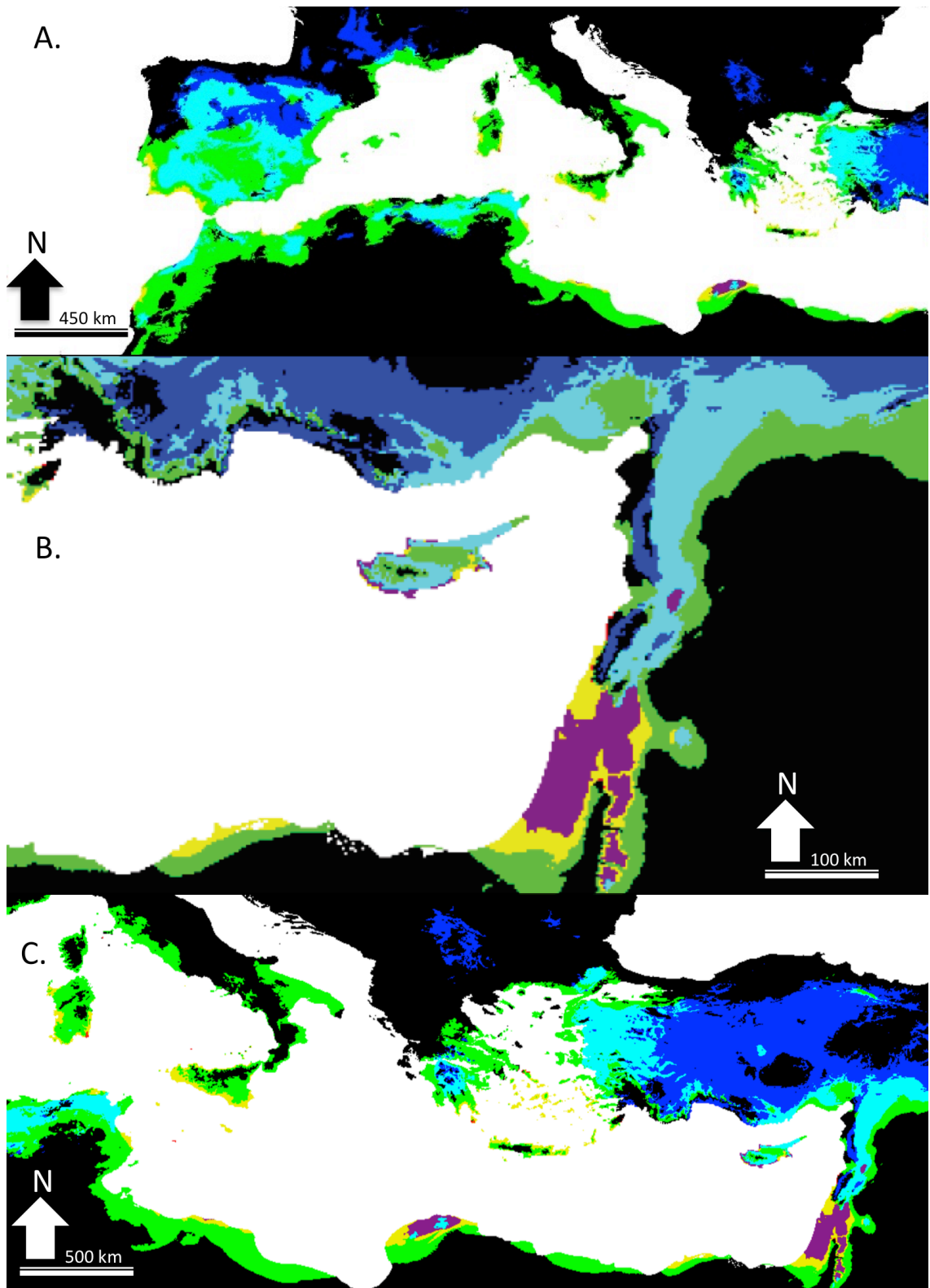




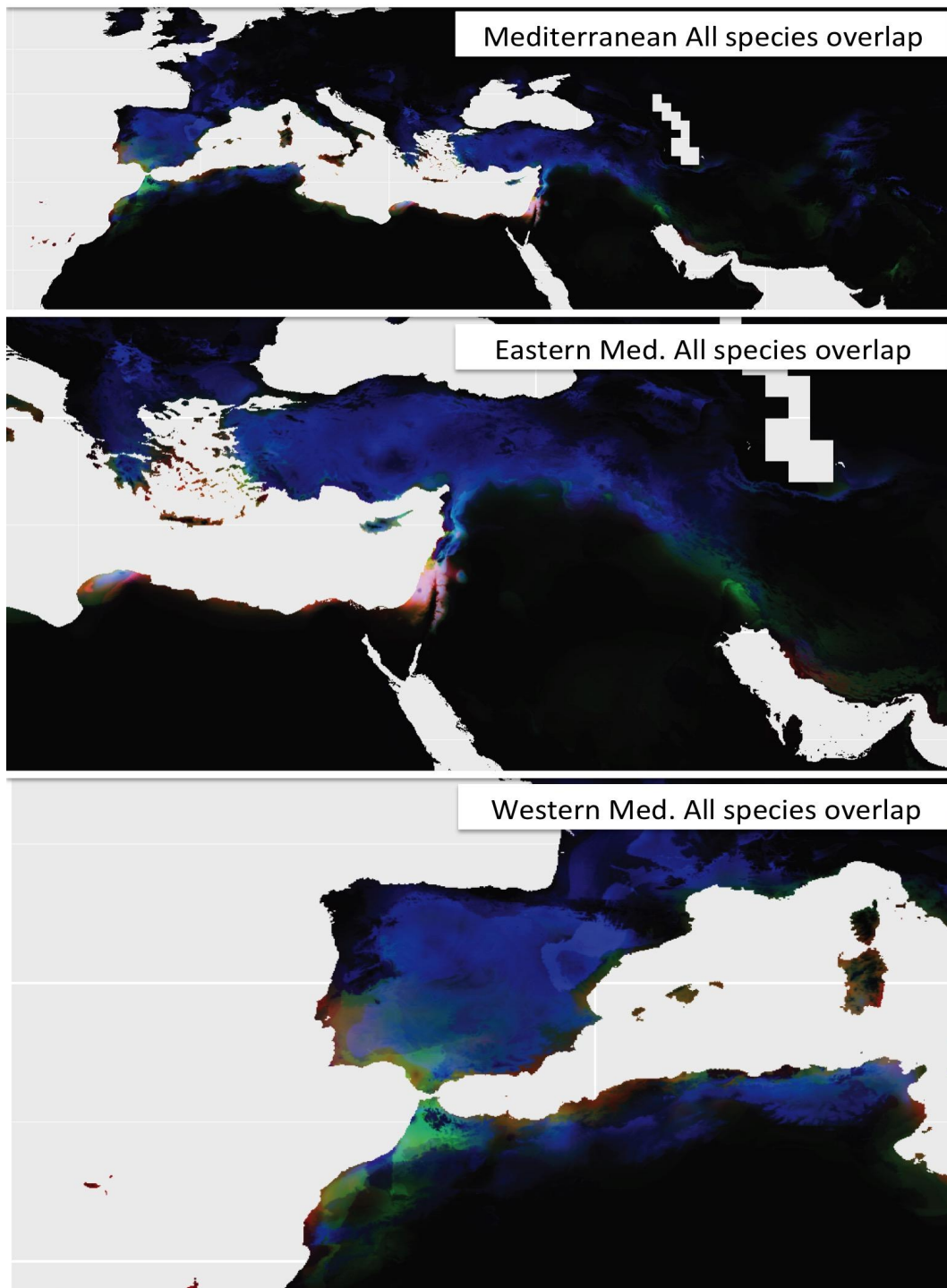
**Figure 4.8.** The Potential area of *B. stacei* in the native ranges of Western Asia and the circum Mediterranean. Suitability is scaled by colour from black to blue. The *B. stacei* range is commonly modelled near coastal areas in dryer climates compared to other complex members. Since *B. stacei* is rare in collections and observations, its known ranges are likely still not entirely known.



**Figure 4.9.** The Overlap of all three *B. distachyon* complex species across the native range. Each species potential area was set to binary suitability:non-suitability based on their individual sensitivity = specificity metric from MaxEnt. To make colour profiling of individual and overlapping ranges more distinct, a red-green-blue (RGB) colour scale was used instead of the previously allocated colours. Each species is represented by individual colours, Blue = *B. distachyon*, Red = *B. stacei*, and Green = *B. stacei*. Overlap of species is represented by between colours: Cyan = *B. hybridum* + *B. distachyon*, Yellow = *B. hybridum* + *B. stacei*, Magenta (Rare) = *B. stacei* + *B. distachyon*, Purple All Species.



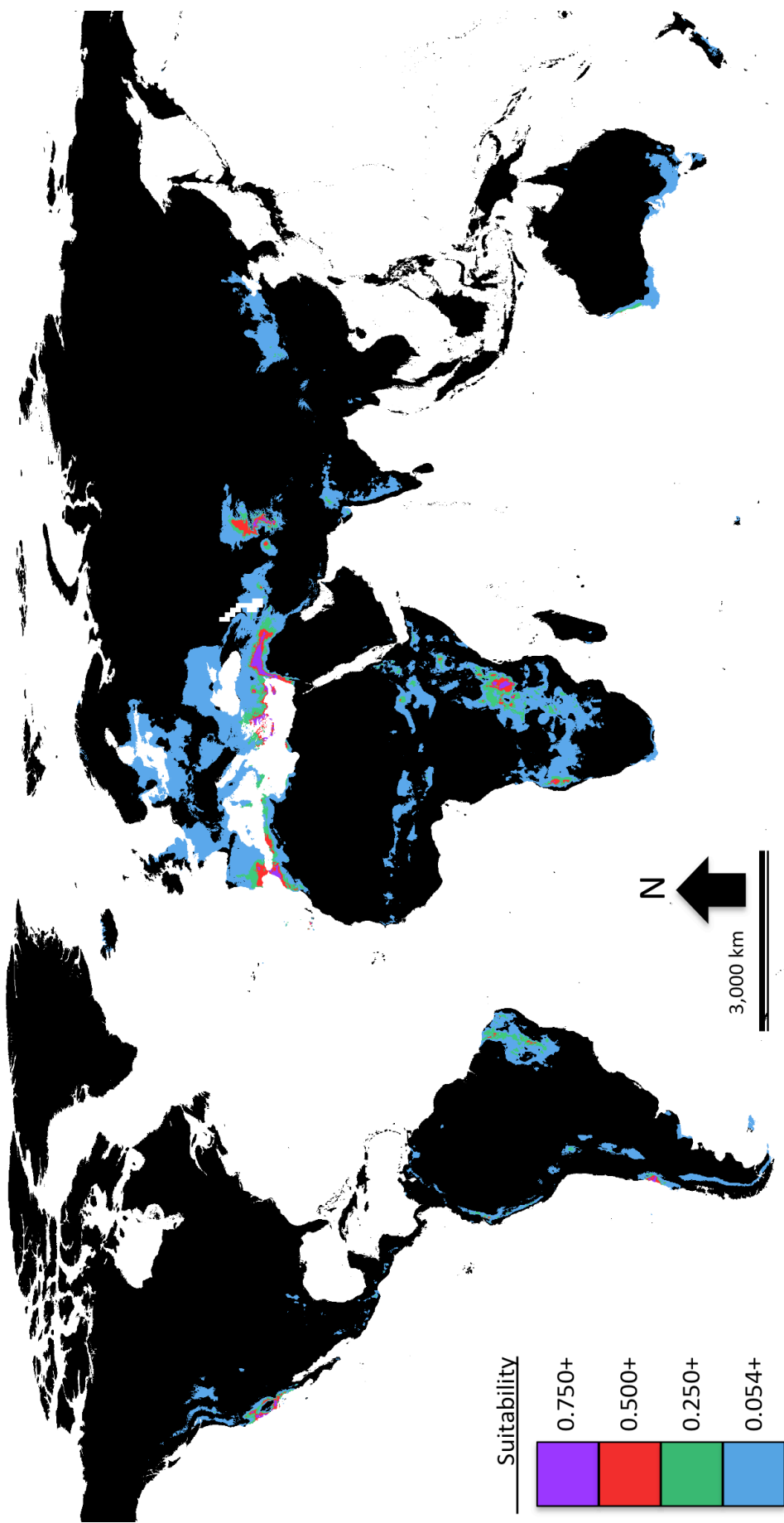
**Figure 4.10.** The Overlap of all three *B. distachyon* complex species across the native range. Each species potential area was set to binary suitability:non-suitability based on their individual sensitivity = specificity metric from MaxEnt. To make colour profiling of individual and overlapping ranges more distinct, a red-green-blue (RGB) colour scale was used instead of the previously allocated colours. Each species is represented by individual colours, Blue = *B. distachyon*, Red = *B. stacei*, and Green = *B. hybridum*. Overlap of species is represented by between colours: Cyan = *B. hybridum* + *B. distachyon*, Yellow = *B. hybridum* + *B. stacei*, Magenta (Rare) = *B. stacei* + *B. distachyon*, Purple All Species. A: Zoomed in view of the Western Mediterranean of species overlap. B: The overlap of species across the Eastern Mediterranean. C: View of the Central Mediterranean and Turkish Peninsula.



**Figure 4.11.** All *Brachypodium* study species overlap with gradient RGB colouring: The above maps indicate the overlap of all three *Brachypodium distachyon* complex members in their native circum-Mediterranean range in RGB colour gradients. The colour regime is the same as Figure 4.10 except colours are blended based on potential suitability for each species. The input data and colour regime are the same as in Figure 4.10.

#### Maximum Entropy Predictions for each Species, Globally

The MaxEnt parameters to predict potential global area for each species were set the same as Lopez, 2015 for native range distribution for each species except the study area was left uncropped to maximize potential area prediction. The only other difference is that the observation data is different and the predictions might not match exactly.



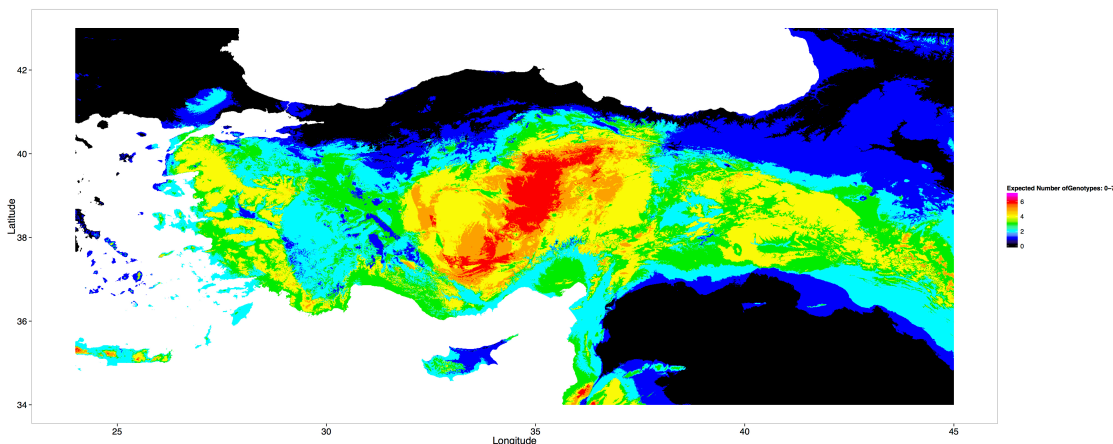
**Figure 4.12.** Global Distribution Model for the genotype NRD-1 of *B. hybridum*. A maximum entropy distribution model for the most common genotype to predict the sensitive geographic range to invasion. The minimum suitability score as calculated by Sensitivity = Specificity through MaxEnt was 0.054 and is coloured in Blue. Three other binary threshold were set to show variation in predicted geographic locations at different suitability considerations:  $p > 0.25$  in green,  $p > 0.50$  in red, and  $p > 0.75$  in purple.

Suitability scores	NRD-1 (km)	All <i>B. hybridum</i> (km)
Minimum Specificity = Sensitivity: 0.054 and 0.058	18,247,672	6,705,946
0.25	2,998,690	3,184,070
0.50	1,114,262	677,697
0.75	317,966	3,413

**Table 4.13.** Table of suitability thresholds for three random models and NRD-1 for potential area in km<sup>2</sup>. Suitability thresholds from 25 distribution models (1,000 iterations each) were averaged together to calculate the potential area of *B. hybridum* (303 locations) and NRD-1 (51 locations). Across four thresholds of suitability scores NRD-1 had larger potential area with regard to climate data. The minimum scores for equal sensitivity = specificity are nearly identical, however the rest of the model statistics indicate that averages between model sensitivity across thresholds varied widely.

#### Expected Genotypes Density of *B. distachyon* in Turkey

Eight different genotype families of *B. distachyon* were independently modelled via MaxEnt and then combined to show the locations across Turkey that could harbour the most genetic diversity and most likely locations to have outcrossing individuals and gene flow. To see each individual genotype model see appendix S4.16 through S4.23. Regions with a high-expected number of genotypes would be ideal for high-resolution landscape genomics studies and optimising collection trips for genetic diversity. Being modelled for genetic diversity could actually have more than the expected number of genotypes because some locations could have more measurable outcrossing rates.

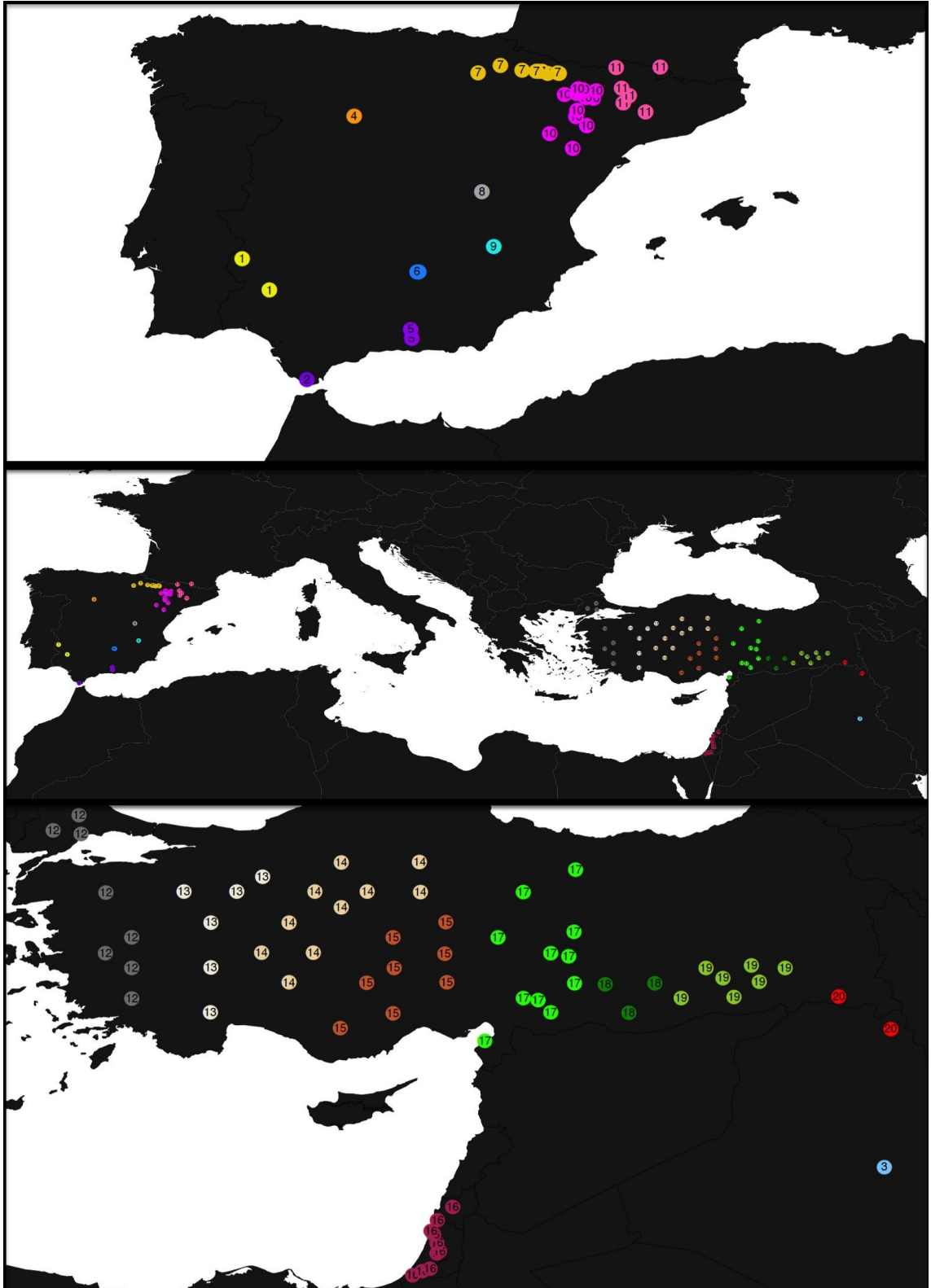


**Figure 4.14.** Expected Genotype Density of *B. distachyon*: The suspected most genetically diverse locations of Turkey for *B. distachyon* created from eight different genotype MaxEnt models. See Appendix for individual models per genotype.

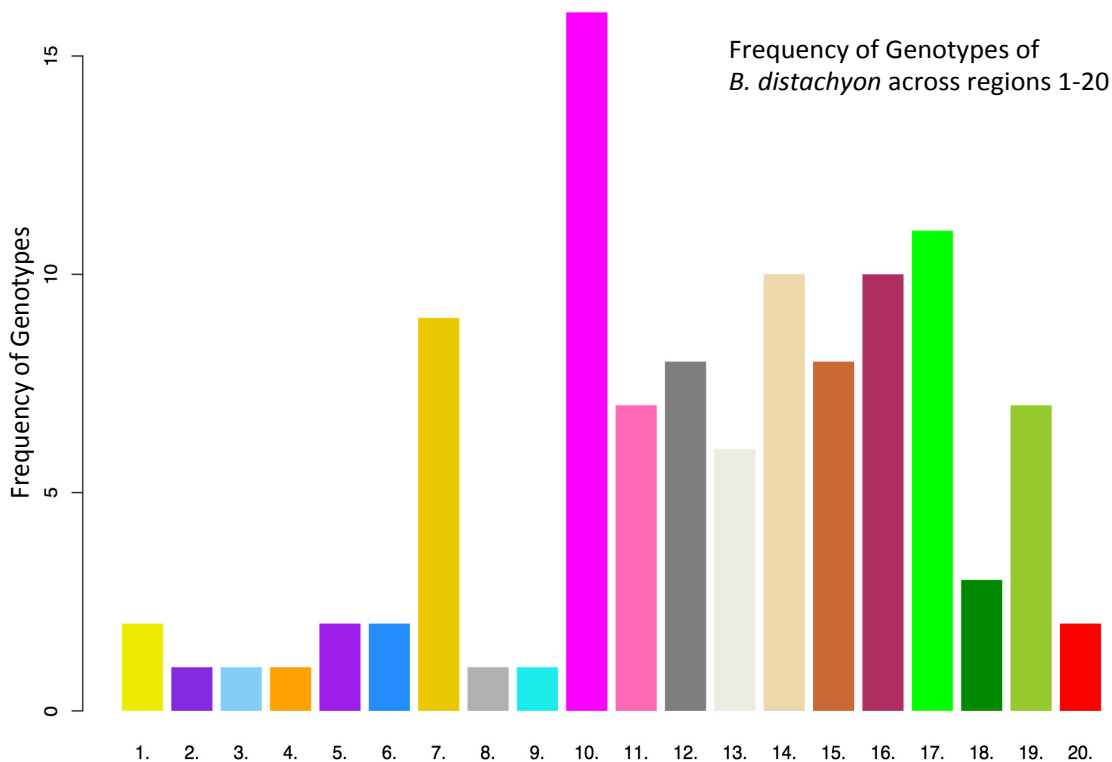
#### Genotype diversity to geographic region

The germplasm in this study is a composition of nine different research groups combining their seed material to study genetic diversity across geographic and climate space (Chapter V). Since each research group had one or more collectors and different collection practices were employed per each group, collection locations were clustered by their proximity to each other. By clustering locations by their geographic distance from one another, regions can be assigned

in specific areas with varying sampling coverage, and the local diversity in these regions can be better understood. *B. distachyon* was found to be in 20 different regions in the native range with outlier locations removed. *B. hybridum* was assigned 35 distinct regions globally and outlier location in South Africa and South America were removed since they both composed of only one sample location each.



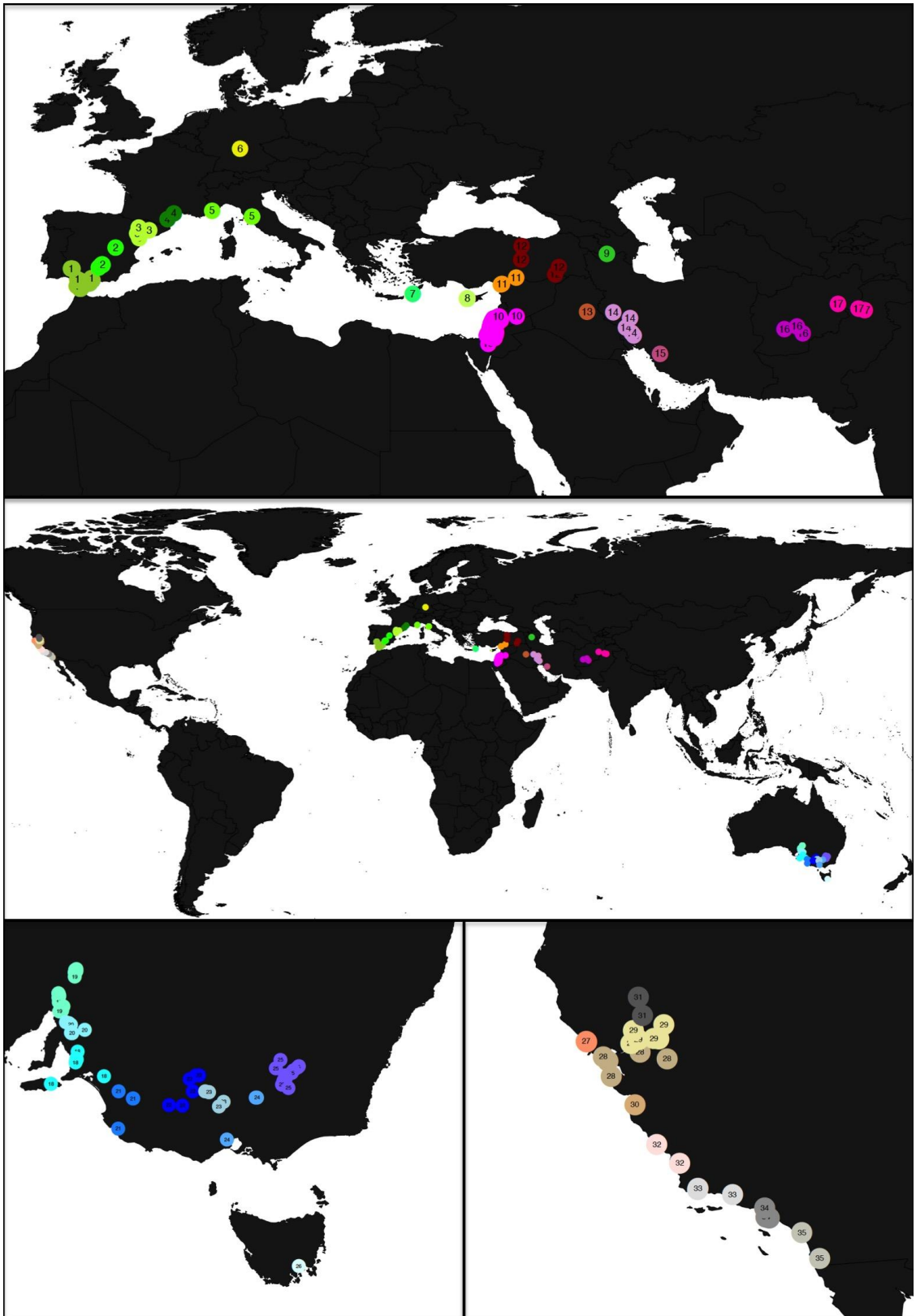
**Figure 4.15.** Reprise of regions as described in Chapter II. Figure 4.16 barplots colour corresponds to regions to show what regions had higher numbers of unique genotypes.



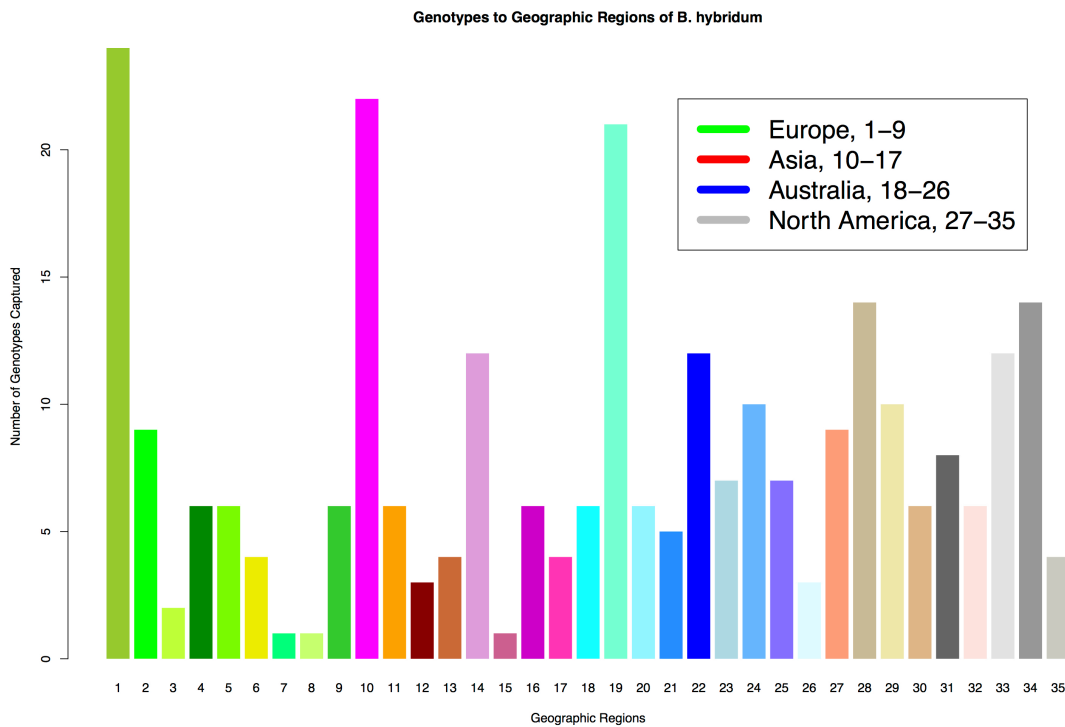
**Figure 4.16.** Bar Chart of Genotypes found per *B. distachyon* regions: The total number of genotypes are counted per each region and placed in barplots of their region's identity. Areas near the NE of Spain received intensive sampling that provided many unique genotypes. Genotypes can be found in more than one region and counted multiple times across locations. y-axis = frequency of genotypes, x-axis = regions.

Region 7, 10, and 11 were the most genotype abundant in the western Mediterranean. The eastern Mediterranean showed significant diversity in regions 12 through 17. Region 10 had the most genotypes showing a genetic hotspot for *B. distachyon*. Much of the Iberian Peninsula is absent of sampling, or existing samples did not make it to this study. The southern Pyrenees showed 38 genotypes across three regions. The country Turkey showed substantial diversity across most of the country with much of the unique genotypes in the central part of the country. Regions with lower diversity should still be further investigated as some could just be under-sampled or subject to biased sampling regimes. Regions of the Pyrenees showed dynamic variation in number of genotypes as well indicating pockets of diversity are near areas with substantial differences in lineages and could be the result of recent or adaptive migration if few groups are found in specific areas.





**Figure 4.17.** Reprise of regions as calculated in Chapter II. The regional identities of collection locations of *B. hybridum*. Locations have colour themes by the continent they were found in. A total of 35 regions were created with 9 locations in Europe (green), 8 region in Asia (red-magenta), 9 regions in Australia (blues), and North America with 9 regions (grey-brown).



**Figure 4.18.** Barplots of Sample Counts per Each Region: The sampling regime of each collaborator is different and some locations received significantly more coverage than others, and some regions had more sample locations than others. Region 10 in Eastern Mediterranean had many more samples and locations than others. The x-axis = regions, and the y-axis = frequency of unique genotypes.

#### Genetic Diversity Across Regions of *B. hybridum*

Much of the genetic diversity of *B. hybridum* across geography was low in diversity compared to a few locations natively: 1 and 10, while in the non-native sites region 19 near Adelaide Australia showed significant diversity. A large proportion of genotypes were found in non-native regions, but could be from admixture outcrosses of two or more genetically wide individuals resulting in highly diverse offspring.

## 4.4 Discussion

### Current distribution and future collection locations

The regions currently collected from in public and private germplasms constitute much of the expected native range of the three *Brachypodium* species with the exception of North Africa, and regions near the central Mediterranean (Greece and Italy). With much of the distribution of *B. distachyon* at 40.48 latitude in our study, the central Mediterranean regions at this latitude are mostly water and land is scarce. The northern regions of the central mediterranean coastline are towards the upper latitude limit for much of *B. distachyon* and sightings of the species will be less probable. Even more so for *B. stacei* whose northernmost predicted locations are most limited of the three study species. *B. stacei* should be common across much of the eastern Mediterranean region and Northern Africa, near countries: Israel, Jordan, Syria, Lebanon, Egypt, Libya, Tunisia, Algeria, and Morocco. Regions of the western Himalaya were also predicted by models for both *B. distachyon* and *B. hybridum*, but so far no *B. distachyon* have

been found that far east. The output maps created in these models could optimise collection efforts in these regions to capture more *B. distachyon* in geographically unique regions like central Europe and the western Himalayan range. There were a few locations further north that could be investigated as well. Northern France near Paris and Nancy showed trace regions suitable for *B. distachyon* as well as parts of southern England and these regions were predicted to have suitable paleo-climates (Lopez-Alvarez, 2015). Areas near Macedonia, northern Serbia, southern Hungary, and Bulgaria also showed measurable suitability for *B. distachyon*.

The native range of *B. hybridum* overlapped much of the suitable habitat of both *B. stacei* and *B. distachyon*. Simply adding the two distributions of the diploids does not perfectly predict the suitable area of *B. hybridum*. Meaning after the hybridisation event that created *B. hybridum*, some regions were no longer suitable to the tetraploid species, and new habitat was predicted as suitable beyond the distribution of the diploid species, and sample collections confirm the presence of *B. hybridum* beyond the areas that the tetraploid overlaps with diploids.

#### Sensitive areas to invasion

By modelling each species the potential overlap was shown in the native range, but also the potential area outside the native range for genotypes and species. Numerous geographic locations were calculated as suitable; *B. hybridum* having the most global area found potentially habitable. *Brachypodium distachyon* had sporadic non-native locations qualify as potentially habitable, and *B. stacei* had few locations almost exclusively on the African continent.

#### Global sensitive areas to invasion from *B. distachyon*

Regions of the United States showed possible suitability mainly in the Northwestern US in the state of Washington, near the south central part of the state and parts of the State of Oregon. Though this region showed suitability in modelling, these regions were inspected and no *Brachypodium* species of any kind were found. Other locations in the NW also modelled as suitable near the US Canada border near Vancouver British Columbia, which also showed mild suitability. In the southwest part of the U.S. near Los Angeles also modelled suitable climate and habitat in the San Gabriel Mountains. Also near San Diego Mountains near the Cleveland National Forest. Areas near the Rio Grande river delta in the US state of Texas and the Mexican province of Tamaulipas and city of Matamoros. A few patchy locations near Delaware and Maryland also showed mild suitability scores. China had moderate suitability scores in the North to Northeast provinces of East Shandong near Jining into the Shanxi and Shaanxi provinces near the Houhe and Huanglian Rivers. Further towards the Ghansu province also showed mild suitability towards the cities of Longnan and Tianshui. South central China also showed mild to moderate suitability in most regions of the province of Yunnan near the Yuangjiang river and into North Vietnam. In the far north of Vietnam showed mild suitability near the China-Vietnam border close to the Phang Xi Pang National Park. Argentina showed

little to no suitability but only in the southern regions of the Buenos Aires Provinces near Mar Del Plata to Santa Teresita coastal areas and parts of the coastal areas near the borders of provinces Rio Negro and Buenos Aires Provinces. Inland Argentina also showed mild suitability near where the Chubut and Rio Negro provinces both border with Chile. Very few areas of Australia showed suitability for *B. distachyon* only being in the State of Victoria near the Grampians to the coast west of Melbourne and a few mild patches near the Margret River in Western Australia. The only other location is the mountains near Hobart in the island state of Tasmania. The South Island of New Zealand showed very little suitability of *B. distachyon* consisting of a narrow strip of climate space from the town of Lumsden in the Southland Province to the Otago Province town of Cromwell.

#### Global sensitive areas to invasion from *B. hybridum*

Nearly all of southern continent of Australia is sensitive to invasion of *B. hybridum*, with the only exception is the far southeast coastal areas of Victoria. Starting in Western Australia the Zuytdorp Nature Reserve and diagonal southeast to Jilbadji Nature Reserve are sensitive, over to Flinders and gammon Ranges in South Australia, and across the southern Hay Plain to Wagga Wagga in New South Wales and south to the city of Melbourne in the state of Victoria. Also the island state of Tasmania is sensitive near the city Hobart. Approximately one sixth to one eighth of the continent is sensitive to invasion. Few areas in the United States are sensitive to invasion of *B. hybridum*. The primary suitable space is in the state of California composing the majority of the San Fernando Valley from Sacramento to Bakersfield. The coastal areas of California are also sensitive from the cities of San Francisco to San Diego and into the Country Mexico. Mexico shows little suitability, primarily near the United States Mexico Border near Tijuana, patchy areas near Caborca in the Province of Sonora, and two isolated patches near Tuxpan in Nayarit province and Cabo San Lucas in the Baja California Sur Province. *Brachypodium hybridum* showed significant suitability in South Africa near Cape Town in the Western Province near the coastal regions. Some areas along the southern coastline in the Western and Eastern Provinces showed patchy suitability as well as small sporadic portions in the Province Limpopo from Pretoria to Zimbabwe. Very few areas in Mozambique and Zimbabwe showed any suitability only a few mild climate spaces showed habitat potentially suitable, but were in the Matabeleland South and Masvingo provinces and Mozambique province of Gaza. Regions near Santiago and Villarrica showed high suitability for *B. hybridum* in the provinces of Araucania and Los Rios provinces from coastal areas to the foothills of the Andes. Very similar to the predicted suitable habitat of *B. distachyon* in Argentina, *B. hybridum* showed mild to moderate suitability in the southern regions of the Buenos Aires Provinces near Mar Del Plata to Santa Teresita coastal areas and parts of the coastal areas near the borders of provinces Rio Negro and Buenos Aires Provinces. Inland Argentina scored mild suitability south of the intersection of the Chubut and Rio Negro provinces and the Chilean Border.

#### Global sensitive areas to invasion from *B. stacei*

*Brachypodium stacei* habitat is the least common of the three species globally based on the input points available to model its potential planet wide distribution. Within Angola, only the coastal areas of the central west showed high suitability especially in the Quicama National Park in the Bengo Province and moderate suitability near the city Lobito in the Namibe province. The Cape Verde Islands are the only other location to show high suitability of *B. stacei*. The highest suitability came from the north western islands of Santo Antao and Ilha de Sao Vicente. However, this region could potentially be considered native habitat considering the proximal distance the Cape Verde islands have to native populations residing on the northern African continent.

#### Genotype Specific Distributions of *B. distachyon* in Turkey

The genotype diversity modelling in figure 4.14 moderately follows the trends seen in figure 4.7, but slightly over predicts the diversity in some areas and leaving others under predicted. This is because the genotype diversity modelling was composed of eight genotype models. Thus, the maximum output value would be eight genotypes. If more genotypes were available in enough abundance, then the diversity modelling would likely be more accurate. Furthermore, the amount of physical diversity that is sampled physically and made it to this study is also a limiting factor to consider. Had more samples been taken at each location then more diversity is more likely to be captured. As genotype models continue via the author's anticipated future work and maybe others, it could be possible to compute probable genetic diversity per regions of a study areas based on mixing distribution models and find areas of high genetic diversity. Coupling this form of predicting genetic diversity analysis with mapping climate gradients could also improve sampling regimes and optimise landscape analysis.

\*Note: During the course of this program, I have been working on mapping climate gradients within species distribution models to reveal climate boundaries and diversity across geographic space. These models are not a required part of this dissertation, but I anticipate furthering this concept to create a much-needed tool in climate analysis of landscapes. The argument being that one species may have different sensitivities to suits of climate variables to another species, these models would show the climate gradients in general, but also per a specified species sensitivities to particular climate types. These models can also work at a genotype level. Lastly, these climate gradient models would also be helpful for designing transects and optimise sampling efforts by capturing samples from multiple locations of climate types. Ideally this will eventually become a research tool written as an R package and likely python as I code well with these two languages. Here is a link to the software I'm developing:

<https://sites.google.com/site/climtools/>

### *B. hybridum* genotypes In Global Areas

It is yet to be seen how well a genotype level model can actually predict the true habitat of a genotype versus a species. However, designing a system to create several hundred to a thousand potential area predictions at random compared to several of a specific genotype should prove that there is a difference in potential area that at random. It should also maybe considered that a genotype that is believed to be wide-spread might be more plastic in growth and development and lower suitability scores are part of what make it more successful. Simulation data might be the key to investigating widespread species trends by testing multiple hypothetical scenarios as is done in genomic modelling and landscapes. It should also be noted that actual experiments such simulated climates and phenotype data as well as true plant density across gradients can more accurately predict the fundamental niche of a genotype.

### *B. hybridum* Genotypes Models of NRD-1

The fact that there were so many genotypes in non-native regions does help determine the possible fundamental niche of the common genotype NRD-1 as found in Chapter III. However, the level to call genotype in this study may not be accurate enough to truly determine the climate breadth of a genotype, because the cut height of the dendrogram is not based on percent difference of markers, but on the highest branch of two individuals of a set of technical replicate individuals. Thus, the genotype could be more like a lineage or subfamily. It could be reasoned that a study could be performed that checks if the fundamental niche expands or contracts as the number of genotypes called is lowered causing more individuals to be members of sub-lineages, families, family groups, and eventually major branches. The more inputs MaxEnt has the more accurately it can predict the true range. As discussed above, the more surface area that is not predicted further causes the true negative to bias the false positive and false negative rate (Warren, 2010; Elith, 2011). This would cause an over-prediction of the total surface area of that would be suitable for NRD-1. Chapter V will investigate this further in the context of climate and the breadth of climate variables of common genotypes. For now the hypothesis of *B. hybridum* being having a larger global fundamental niche is accepted by having the most modelled vulnerable habitat. The hypothesis that NRD-1 does have a larger climate breadth than *B. hybridum* as a whole is accepted, but it's actual climate breadth as opposed to its modelled breadth is compared and discussed in Chapter V.

## **4.5 Data Sets and Script Links**

---

[https://github.com/jstreich/GIS\\_And\\_LandScapes](https://github.com/jstreich/GIS_And_LandScapes)

## 4.6 Citation

---

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, 17(1), 43-57.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874.

Ficetola, G. F., Thuiller, W., & Miaud, C. (2007). Prediction and validation of the potential global distribution of a problematic alien invasive species—the American bullfrog. *Diversity and Distributions*, 13(4), 476-485.

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews*, 29(Suppl 1), S83.

Fraley, C., & Raftery, A. E. (2006). *MCLUST version 3: an R package for normal mixture modeling and model-based clustering*. WASHINGTON UNIV SEATTLE DEPT OF STATISTICS.

Garvin, D. F., Gu, Y. Q., Hasterok, R., Hazen, S. P., Jenkins, G., Mockler, T. C., ... & Vogel, J. P. (2008). Development of genetic and genomic research resources for, a new model system for grass crop research. *Crop Science*, 48(Supplement\_1), S-69.

GBIF.org (2016). Global Biodiversity Information Facility, Brachypodium distachyon search results.[http://www.gbif.org/occurrence/search?taxon\\_key=5290143&HAS\\_COORDINATE=true&HAS\\_GEOSPATIAL\\_ISSUE=false&display=map&COUNTRY.offset=10](http://www.gbif.org/occurrence/search?taxon_key=5290143&HAS_COORDINATE=true&HAS_GEOSPATIAL_ISSUE=false&display=map&COUNTRY.offset=10)

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2), 627.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, 25(15), 1965-1978.

Hijmans, R. J., & van Etten, J. (2014). raster: Geographic data analysis and modeling. *R package version*, 2(8).

Lee, Cheng-Ruei, Hannes Svardal, Ashley Farlow, Moises Exposito-Alonso, Wei Ding, Polina Novikova, Carlos Alonso-Blanco, Detlef Weigel, and Magnus Nordborg. (2017). "On the post-glacial spread of human commensal *Arabidopsis thaliana*." *Nature Communications* 8.

López-Alvarez, D., Manzaneda, A. J., Rey, P. J., Giraldo, P., Benavente, E., Allainguillaume, J., ... & Ezrati, S. (2015). Environmental niche variation and evolutionary diversification of the *Brachypodium distachyon* grass complex species in their native circum-Mediterranean range. *American journal of botany*, 102(7), 1073-1088.

Manzaneda, A. J., Rey, P. J., Anderson, J. T., Raskin, E., Weiss-Lehman, C., & Mitchell-Olds, T. (2015). Natural variation, differentiation, and genetic trade-offs of ecophysiological traits in response to water limitation in *Brachypodium distachyon* and its descendent allotetraploid *B. hybridum* (Poaceae). *Evolution*, 69(10), 2689-2704.

- Matthews, R. B., Rivington, M., Muhammed, S., Newton, A. C., & Hallett, P. D. (2013). Adapting crops and cropping systems to future climates to ensure food security: The role of crop modelling. *Global Food Security*, 2(1), 24-28.
- Lichstein, J. W. (2007). Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology*, 188(2), 117-131.
- Medley, K. A. (2010). Niche shifts during the global invasion of the Asian tiger mosquito, *Aedes albopictus* Skuse (Culicidae), revealed by reciprocal distribution models. *Global ecology and biogeography*, 19(1), 122-133.
- Peterson, A. T., Papes, M., & Kluza, D. A. (2003). Predicting the potential invasive distributions of four alien plant species in North America. *Weed Science*, 51(6), 863-868.
- Phillips, S. J., Dudík, M., & Schapire, R. E. (2004, July). A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning* (p. 83). ACM.
- Phillips, S. J. (2005). A brief tutorial on MaxEnt. *AT&T Research*.
- Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. *Ecography*, 31(2), 161-175.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945-959.
- Sharbel, T. F., Haubold, B., & Mitchell-Olds, T. (2000). Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Molecular Ecology*, 9(12), 2109-2118.
- Šimundić, A. M. (2009). Measures of diagnostic accuracy: basic definitions. *EJIFCC*, 19(4), 203.
- Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, 36(12), 2290-2299.
- Ward, D. F. (2007). Modelling the potential geographic distribution of invasive ant species in New Zealand. *Biological Invasions*, 9(6), 723-735.
- Warren, D. L., Glor, R. E., & Turelli, M. (2010). ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography*, 33(3), 607-611



## Chapter V: Climate Analysis and Genotype Climate Windows

### **Abstract**

Climate factors of temperature and precipitation are the primary abiotic stress acting on organisms. Understanding the variation in climate tolerance at a species level and even a genotype level can aid our understanding of adaptation. The breadth of annual temperature, annual precipitation, and temperature seasonality are the three shared climate variables that contributed highly to each species MaxEnt models in their native range and were used to contrast climate envelopes between each *Brachypodium distachyon* complex member. *B. stacei* was found in warmer drier environments with the smallest amount of seasonal change in temperature. *B. distachyon* was found in the wettest environments, but had similar annual and seasonal temperature values to *B. hybridum*. *B. hybridum* had the largest annual mean temperature range to all species in this study. Multi-linear regression models and Partial multi-linear regression models tests were performed on *B. hybridum* and *B. distachyon* for association to BioClim variables. BioClim4, seasonal temperature, explained 3.96% of the genetic variation in *B. distachyon* in a partial multi-linear regression models test when geography was included in the association model. BioClim4 explained the most genetic variation in *B. hybridum* as well, but  $R^2$  values using all genetic markers yielded 0.33% of genetic variation is explained by BioClim4. A method was developed to create species-specific climate classes. Within collection locations, variation in climate was calculated where each location's climate variables were clustered into groups to reveal what locations had similar climate despite geographic distance. The breadth of each common genotype was permutation tested across geographic locations where the breadth of climate types was measured across sites. Genotype NRD-1, which was previously shown to be geographically wide, also was present in the most climate classes and had the largest climate breadth,  $p < 0.01$ .

### **Chapter Outline**

---

#### **5.1 Introduction**

#### **5.2 Methods**

- Definitions of WorldClim Variables
- Niche-Breadth per species and genotype
- Climate Data Collection
- Filtering Climate Data Methods
- Mathematical Processes

#### **5.3 Results**

- The Global Climate Diversity of Collection locations per species
- Climate Diversity of *B. distachyon* and *B. hybridum* Genotype

#### **5.4 Discussion**

#### **5.5 Data Sets and Script Links**

#### **5.6 Citation**

## 5.1 Introduction

---

Climate is one of the primary selector pressures of a species or even a genotype's geographic distribution (Phillips, 2005; Wisz, 2008; Elith 2011; Warren 2011; Brown, 2016). Studying the environmental requirements and tolerances of a species requires extensive local measurements; examples include soil type, mycorrhizal associates, herbivory, and many others. Such exhaustive studies can be expensive, complex, and require lots of on-site measurements, experimentation, and analysis. This is especially true for global studies where multiple continental sample efforts are required. Climate, precipitation and temperature, is the current best data type to understand species distribution and data sets are available on a global scale at resolutions as small as 1km (Phillips, 2004; Hijmans, 2005; Phillips, 2005; Elith, 2011).

### Climate association studies in *Brachypodium* species

Previous studies have examined the effects of climate on the *Brachypodium distachyon* species complex. Different cytotypes of *B. distachyon*, now different species, were associated with different climate patterns where polyploids had a larger breadth across the mediterranean and were found in warmer regions of the Iberian Peninsula (Manzaneda, 2011). Within that study grain size and phenotypic variation was greater in polyploid samples and associated with climate variation in precipitation and environmental effects in soil moisture. Post species reclassification in 2012, the environmental niche modelling of *B. stacei* and *B. distachyon* diploids were statistically distinct with some overlap, *B. distachyon* is common in cooler regions, while *B. stacei* is found in drier and warmer regions (Catalan, 2012; Lopez-Alvarez, 2015; Catalan, 2015). The polyploid *B. hybridum* overlapped significantly with both diploid species, but also had an expanded ecological range and is consistent with other studies (Manzaneda, 2011; Lopez-Alvarez, 2015; Catalan, 2015). A study in Turkey found 15 possible climate associated loci by scanning 82 wild collected individuals across nine climate unique locations calculated by the Ecocrop function in the program DIVA-GIS on an east-west longitudinal gradient to capture both climate and geographic isolation in a sampling transect using Bd21 as a control as well as four inbred lines (Dell'Acqua, 2014). As mentioned in a book chapter regarding *Brachypodium* species research, very little about life history strategies and variation in ecological variation is currently published (Des Marais, 2015).

### About the BioClim Climate Variables

Nineteen different global climate data sets composed of precipitation and temperature at annual, seasonal, and monthly intervals are readily available for most all global land surface locations at 1+ km square resolution and are frequently used in species to climate studies. Data can be mined for each sample collection location from the 19 biologically relevant global raster layers from BioClim (<http://www.worldclim.org/bioclim>) layers via software like QGIS, R, Atlas of

Living Australia website (ala.org.au) and others. Specific descriptions of each BioClim layer can be found at the website and in the appendix section of this thesis.

### Associating Climate to Species and Genotype

Landscape genomics assumes that natural processes have already conducted the experiment of natural selection by environment and seeks to find causative genetic variation of adaptive phenotypes. Association studies relating coding regions to climate have found loci responsible for adaptive changes in model and non-model organisms. Notably in *Arabidopsis thaliana* several studies revealed adaptive traits caused by variation in coding regions. An early stop codon in CMT2, a methylation transferase, enabled a larger climate tolerance in *A. thaliana* (Shen, 2014). In another study, the environmental variation across sample locations could predict patterns of polymorphisms across the whole genome, as well as variation in GO terms per environment; some polymorphisms were also predicted based on genomic structure and composition; that environmentally relevant factors contribute to population divergence across populations and locally adapted genotypes (Mitchell-Olds, 2012). A similar study showed a pattern across geographic space where suites of inherited genomic markers were present across specific landscape gradients (Hancock, 2011). Some locations overlapped geographically and levels of polymorphisms present per location would be predictive of fitness at one location. Thirty different biological processes were found ecologically relevant across numerous environmental factors with significant p-values. Another study found that non-synonymous variants in climate associations were more common than synonymous variants, which proves that polymorphisms are more likely to change protein coding regions within genes of adaptive alleles (Lasky, 2012).

Significant ecological variation occurs across the native range of *P. taeda* and several studies have published on climate to genotype interactions within the species. A total of 1,730 genomic markers were derived from 682 individuals sampled across 54 locations to investigate the ecological genetics of *P. taeda*, which revealed strong correlations between geography and climate (Eckart, 2010<sup>1</sup>). In this study, numerous variants were correlated with elevation or climate data and annotation reveals possible pathways that are associated with local adaptation most via abiotic stress, which would indicate some sort of environment based selection pressure. A separate study found five variants associated with aridity with significance to both biotic and abiotic stress response, 24 other variants were associated with strong  $F_{st}$  and physiological processes (Eckert, 2010<sup>2</sup>).

Serotiny, the effect of a trigger response to induce seed dispersal from the maternal plant is a trait common in gymnosperms (Johnson, 1993; Bond, 2005). The measure of the serotinous phenotypes was conducted in *P. taeda* in three different ecologically and genetically distinct

populations to investigate serotiny as an adaptive phenotype resulting in 11 loci that explain  $\approx 50\%$  of the phenotypic variation (Parchman, 2012).

### *The Origins of Climate classification*

The classification of climates dates back to the ancient Greeks as five climate types, more recently described by De Candolle and a French plant scientist in 1906 (De Candolle, 1906; Sanderson, 1999). Around the same time as De Candolle, Wladimir Koeppen, a plant physiologist, was compiling the first climate classification system by 1884 (Sanderson, 1999; Koeppen, 1936). Koeppen later built off of both De Candolle's work and Greek philosophers to create at that time the most accurate description of global climate variation by 1936. The five climate classes originally were centered around the general physiological properties of the flora in any given habitat. Koeppen's first groups were: A- torrid zones, B- dry zones, C- temperate zones, D and E were varying levels of arctic or frigid zones, the snow zone, and the polar zone (Koeppen, 1936; Kottek, 2010). Koeppen's classification method was eventually expanded by Rudolf Geiger in collaboration with Koeppen and was published in 1954 (Geiger, 1954). Later in 1966 and updated in 1980 the Trewartha Climate Classification system was developed to better describe the variation in equatorial climates as the previous Koeppen-Geiger system was considered too broad in these zones (Peel, 2007). The current and most popular version of climate classification is a Koeppen Geiger classification system using 31 different climate classes (Halenka, 2013).

### *Per Species Climate classification*

Climate classes are a broad approach to determine climate type based on groups of plant species in a local habitat. A classification system can be very descriptive of climate breadth when comparing multiple species or other branches of science like Climate Change. However, a single plant species can theoretically have broad or narrow breadth in a climate classification system like Koeppen-Geiger, occupying many or few classes. Thus the climate limits of a particular species may not easily be described by climate classification. The breadth of a species theoretically could be smaller than the window defining the classification it inhabits, or between the upper limits of one class and the lower limits of another and would not occupy both classes completely. For a single species study, the climate limits must be measured to more accurately describe the climate variation across individuals and genotypes. The use of BioClim variables of each collection location can be clustered into groups and used to design species-specific climate classes that more accurately reflect the climate diversity within that species.

The use of region identities in Chapter III showed the amount of genotypes per region- groups of local collection locations. The same was performed here, but instead of genotypes, climate classes were used to show what regions had the most climate diversity. Theoretically, if a region

is climate diverse, having many diverse climate-types in close proximity, local individuals would have to disperse smaller distances to encounter a new set of abiotic climate stresses. Thus have more chances to test locally derived mutations against previously uninhabitable environments. One aspect of this chapter is to see if climate diversity is associated with genetic hotspots in the local and non-local ranges.

Calculations using BioClim variables can reveal climate preference and tolerance of a study group. Climate being a significant part of environment, can aid our description of the climate niche breadth of a species. Further, the use of genetic analysis to find not only what species is more widespread, but also what genotypes. For species that rarely outcross, spread widely, or have low genetic diversity even when crossing, this chapter aims to find what groups of each species have wide and narrow climate breadth, and also compare regions of genetic diversity to regions of climate diversity.

#### Chapter Question, Hypothesis, and Aim

**Question:** *What are the climate tolerance limits and variation of *Brachypodium distachyon* species using comparable bioclimatic variables and can certain whole genome genotypes be classified as specialists or generalists?*

**Hypothesis:** *I hypothesize that since *B. hybridum* is a polyploid with larger predicted suitable surface area, it will have larger climate tolerance limits. In addition, some genotypes of *B. hybridum* will occur in more climate classes than chance and thus are climate generalists, while others will be specialists with restricted climate breath.*

**Aim:** *Calculate the climate limits of each species using comparable climate variables and the occurrence of genotypes of each species across geography and species-specific climate classes. Then test the presence of genotypes across climate classes to see if some have wider climate windows than others.*

## 5.2 Methods

---

### Climate-Envelopes Per Species

Climate envelopes per species were contrasted using BioClim variables found significant by MaxEnt models in Chapter IV and previous publication (Lopez-Alvarez, 2015). All study species had three overlapping climate variables that contributed significant information for distribution modelling: BioClim1 Annual Mean Temperature, BioClim4 Temperature Seasonality (standard deviation of temperature range compared to the annual mean), and BioClim12 Annual Precipitation. The three variables each species shares are the ideal variables for quantifying precipitation and temperature annually, plus temperature seasonality. Each species has a different number of observation locations, *B. distachyon*  $n=115$ , *B. stacei*  $n=90$ , *B. hybridum*  $n=303$ .

### Multi-linear regression models and Partial Multi-linear regression models Tests

Association between data matrices of geography and climate to genomic data was performed via the multi-linear regression models test function MRM in the R package 'ecodist' to report p-values,  $R^2$  values, and F-test values (Lichstein, 2007). The attempt was to search for climate variables and geographic distance measures that explain genetic variation. Association scans were performed genetically per all markers cumulative. Scans were also performed in multi-linear regression models and partial multi-linear regression models tests against geographic distance and all BioClim variables. *B. distachyon* samples were found to have some genetic variation explained by BioClim4. At 50% percent shared markers 3.96% of the genome was explained by BioClim4. Allopolyploid *B. hybridum* genomic data was filtered down to just chromosome specific markers. When using a pairwise genetic distance matrix of all markers cumulatively, no BioClim variables explained more than 0.33% (BioClim4) of the genetic variation in partial multi-linear regression models tests in *B. hybridum*. As noted earlier, there are instances where over fitting occurs in mantel tests (Guillot, 2013).

### Climate Data Collection

Climate data was sampled per collection location a few different ways. The quickest method was to use the website Atlas of Living Australia. Using their spatial portal page .csv files were uploaded with geospatial collection location points in latitude longitude digital format. Once geographic coordinates are uploaded the user can sample and download BioClim values specific for each location as well as elevation and other environmental data. Atlas of Living Australia has global BioClim data for nearly the entire Earth. The data can be downloaded to a spreadsheet and modified to be an R readable .txt file. A metadata table for all 817 sample locations was created to perform analysis of BioClim and spatial data to genomic data. On some occasions climate data was extracted from BioClim layers using R.

### Classifying Climate Data Methods

Climate classification was performed using all 19 BioClim variables and clustered using the R package 'Mclust' (Fraley, 2006). Mclust also assigns probability values of data points belonging to other cluster centers as well as center locations. This analysis was done per each species, and using all species. The regions of each species described in Chapter IV also had their climate variables clustered to search for climate variation per region.

### Overview of Climate Analysis

The similarity of precipitation and temperature values associated with each collection location was assessed with pairwise distance measures to show climate relatedness. Each collection location was analysed in two ways, one using all the 19 BioClim variables with equal weighting, and secondly with weighted values for each BioClim variable. The equal weighting of all BioClim variables analysis will decipher the similarity of climate per each location

independent of what the species growing requirements are. The weighted analysis shows the similarity of climate per location by factors deemed selective on the species as a whole by MaxEnt's output.

#### Multidimensional Scaling, Hierarchical Clustering, and Climate

Climate classification was performed by using clustering algorithms via the R package 'Mclust' using all 19 BioClim variables sampled at each location, because Mclust provides cluster center measures (Fraley, 2006). Each location was assigned a numerical value based on its cluster identity to show what climate types were present at locations and regions as seen in Chapter IV. Climate groups were often found to overlap or have close cluster centers to each other despite the number of cluster centers called or designated by computer or user settings. *B. hybridum* locations clustered at 51 unique climate types, but the centers of each were often close together, complex, and difficult to distinguish visually. To simplify climate classification the climate data for each location was forced to smaller clusters, 14 for *B. hybridum* and 11 for *B. distachyon* and specific clusters are proprietary to each species. The goal of climate classification is to see which genotypes or lineages are found in more climate types than others. By forcing the clustering to distinct centers, the range of climate types becomes more visually describable by colour and value designation, this clustering of climate type for each sample location could be biased by possible oversampling of some genotypes and sampling locations. However, despite the resolution of this data set and the genotypes contained some *B. hybridum* and *B. distachyon* were found in more climate clusters than others.

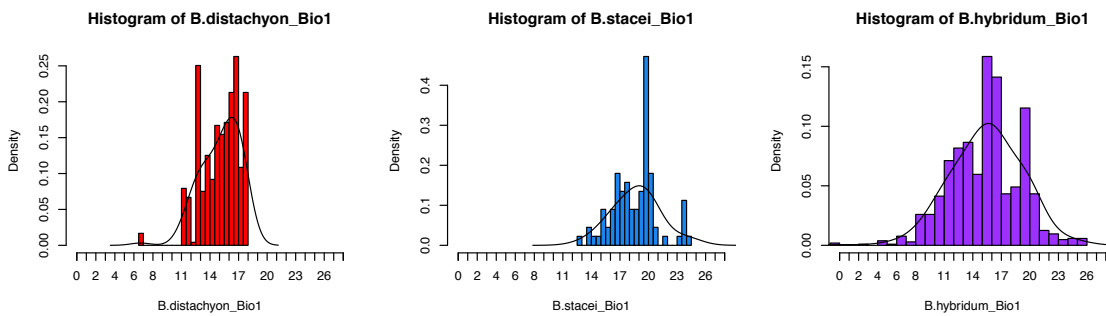
### **5.3 Results**

---

#### Climate Breadth by Significant BioClim Variables

Each species was analysed by their shared BioClim climate variables that were found significant by MaxEnt in *B. hybridum*, *B. distachyon*, and *B. stacei*. Each species had more than three BioClim variables deemed significant in describing the overall species distribution, but three were shared, BioClim1 mean annual temperature, BioClim4 the percent standard deviation to the annual mean temperature, and BioClim 12 the mean annual precipitation. Their climate distributions were plotted in histograms and t-tests were performed to find significant differences in climate breadth per species (See figures 5.1-6). BioClim1 was statistically different between *B. distachyon* and *B. stacei* with a p-value <0.01. Likewise, the difference of mean in *B. stacei* and *B. hybridum* was also statistically significant with a p-value < 0.01. *B. hybridum* and *B. distachyon* did have a statistical significance in differences in mean annual temperature, but the likelihood that these are physiologically significant is not likely and should be experimentally tested as they differ by less than half a degree (See Figure 5.1 and Table 5.2). Temperature seasonality (BioClim4) was found significant between *B. hybridum* and both diploids, but should also be experimentally tested (See Figure 5.3 and Table 5.4). *B. distachyon*

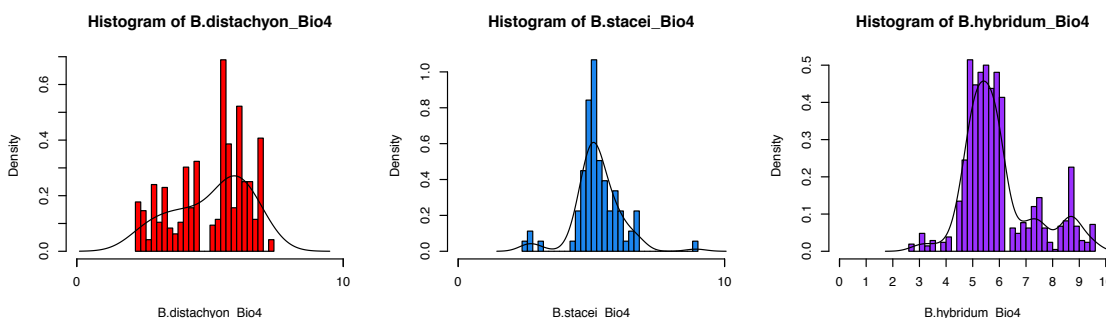
appears to be bi-modal in BioClim4 and there are two main flowering types in *B. distachyon*, however the bi-modal nature does not seem to correlate with flowering time based on published literature, but could be experimentally tested (Woods, 2013; Ream, 2014). There was a non-significant difference in temperature seasonality between *B. distachyon* and *B. stacei*. Lastly, there was a statistical significant difference in BioClim12 between each combination of species all having a t-test p-value < 0.05, and two having a p-value < 0.01 (See Figure 5.5 and Table 5.6). The number of collection sites of *B. hybridum* could be influencing the significance of these results having 303 locations compared to 90 *B. stacei* sites and 115 in *B. distachyon*.



**Figure 5.1.** *B. distachyon* and *B. hybridum* both had similar yearly average temperature, but *B. hybridum* was slightly warmer and one standard deviation was found to be higher. *B. stacei* was the warmest and its mean temperature was more than two standard deviations from *B. distachyon*. Both *B. stacei* and *B. hybridum* had similar maximum annual temperature extremes. *B. hybridum* had the largest difference between maximum and minimum annual mean temperature. See Table 5.2 for a description of these histograms in a t-test output and coefficient of variation.

<i>B. distachyon</i> mean	<i>B. stacei</i> mean	<i>B. hybridum</i> mean	<i>t</i>	<i>df</i>	p-value	95% conf. interval	95% conf. interval	CV%
15.09	18.69	-	-13.34	112.49	<2.2e <sup>-16</sup>	-4.137	-3.067	9.33
15.09	-	15.44	-2.38	1471.75	0.02	-0.636	-0.061	9.20
-	18.69	15.44	-11.70	126.50	2.2e <sup>-16</sup>	-3.804	-0.061	38.02

**Table 5.2.** BioClim1 t-test and percent coefficient of variation: BioClim1 Average mean temperature is significantly different per each species in different attributes. *B. distachyon* and *B. hybridum* have similar average mean temperature, but *B. hybridum* has a larger standard deviation and coefficient of variation to the mean indicating a larger breadth in yearly mean temperature than *B. distachyon* and *B. stacei*. The highest mean annual temperature was *B. stacei* at 18.81 degrees C°.

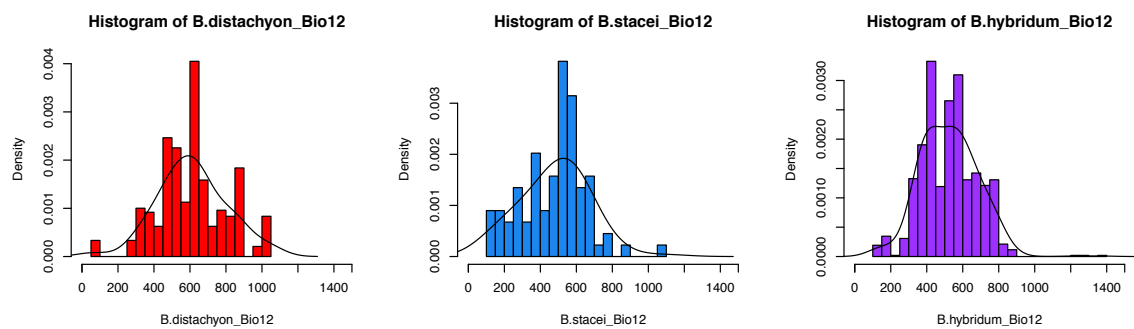


**Figure 5.3.** *B. stacei* had the smallest variation in Temperature seasonality with most samples in +/- 1 standard deviation and few outliers. *B. hybridum* and *B. distachyon* nearly have the same temperature seasonality value and standard deviation. *B. distachyon* was found to have the most temperature extremes, but both *B. distachyon* and *B. hybridum* can be found in very neutral to very extreme temperature shifts across a year. *B. hybridum* had three peaks indicating a level of tri-modality but one peak was significantly larger.



B. distachyon mean	B. stacei mean	B. hybridum mean	t	df	p-value	95% conf. interval	95% conf. interval	CV%
5.05	5.22	-	-1.60	183.31	0.112	-0.398	0.042	189.61
5.05	-	5.94	-1.04	891.04	<2.2e <sup>-16</sup>	-1.040	-0.744	114.15
-	5.22	5.94	7.09	125.94	8.34e <sup>-11</sup>	0.514	0.912	160.31

**Table 5.4.** BioClim4 t-test and percent coefficient of variation: BioClim4 temperature seasonality is relatively narrow for *B. stacei* and *B. distachyon* with few exceptions, while *B. hybridum* has more broad breadth in variation in annual mean temperature, tolerating large differences in annual mean temperature indicating it can survive in more extreme temperature regimes than the other two species.



**Figure 5.5.** *B. distachyon* was found to be in the wettest environments, while *B. stacei* was found in the driest then *B. hybridum*. One positive standard deviation of *B. stacei* overlaps with one negative standard deviation of *B. distachyon*, so approximately 50% of samples from each species could overlap in total annual precipitation. While the breadths of each species between maximum and minimum values are nearly identical, *B. distachyon* appears to tolerate the widest breadth of annual precipitation in total. *B. stacei* has largest coefficient of variation to the mean indicating that per the amount precipitation it has the most variability, but could be due to outlier sample locations.

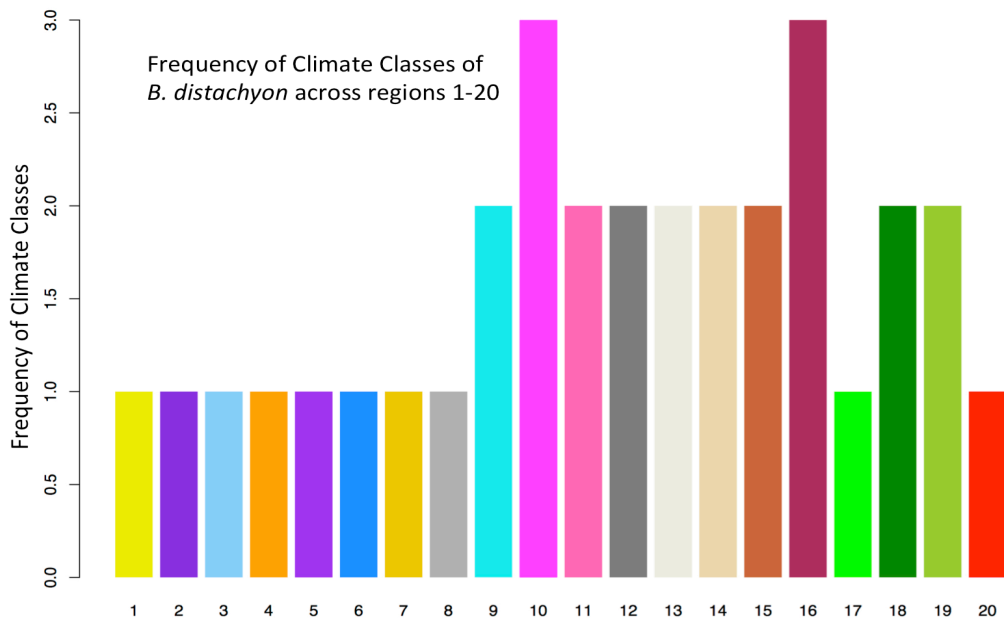
<i>B. distachyon</i> mean	<i>B. stacei</i> mean	<i>B. hybridum</i> mean	t	df	p-value	95% conf. interval	95% conf. interval	%CV
612.22	477.40	-	6.40	126.77	2.76e <sup>-9</sup>	93.136	176.48	20.18
612.22	-	523.14	8.99	778.52	<2.2e <sup>-16</sup>	69.622	108.547	26.70
-	477.40	523.14	2.31	99.26	0.02	6.500	84.961	19.99

**Table 5.6.** Bio12 t-test and percent coefficient of variation: *B. distachyon* occurs on average in locations with higher rainfall per year. *B. hybridum* has the largest breadth of rainfall experienced per year, but with few rare outliers. *B. stacei* has the smallest amount of average annual rainfall per year across its populations and the largest coefficient of variation at 26.70%. However, the distribution of *B. stacei* with *B. hybridum* overlaps heavily with *B. hybridum* averaging approximately 46.54mm more precipitation.

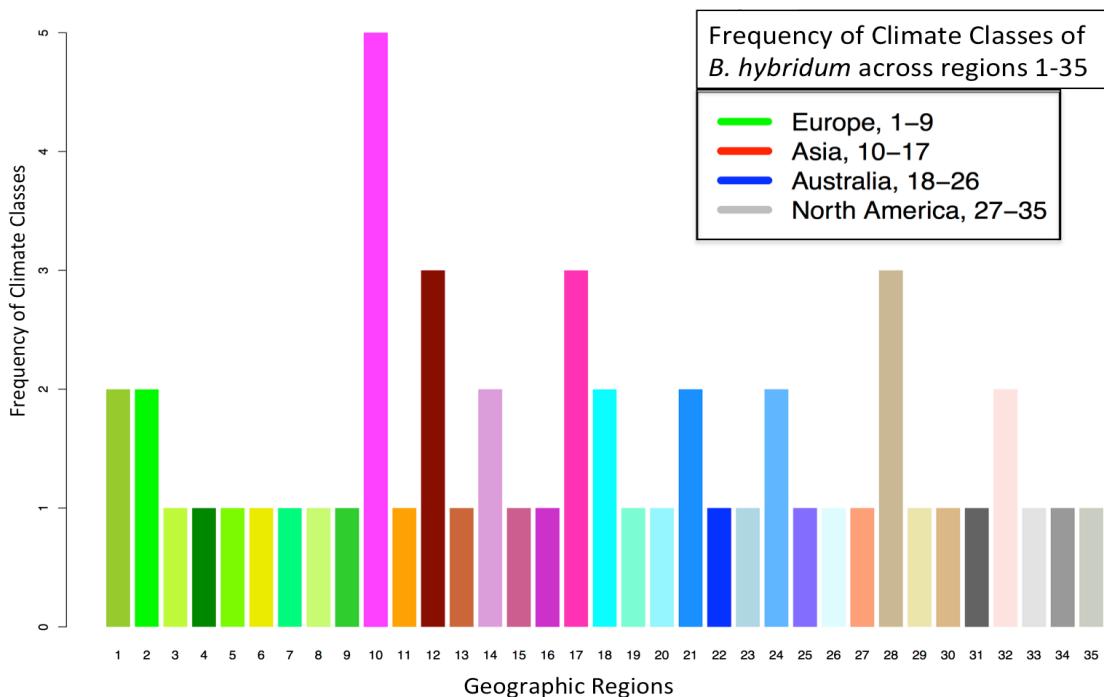
### Species-Specific Climate Classification

Classification of climate classes was performed for *B. distachyon* and *B. hybridum*. *B. stacei* was rare in collections with few sites and was not used in tests. Since *B. hybridum* is much more geographically broad than *B. distachyon* having hundreds of samples and locations globally, climate classes were mostly utilised in *B. hybridum*. However, the number of genotypes in *B. distachyon* were still tabled to regions and compared to climate diversity also within those regions as seen in Chapter IV. *B. distachyon* climate data was split into 11 climate classes

regions near the Pyrenees had the highest climate diversity, which also harboured the most genetically diverse regions (See Appendix Supplemental figures S5.8 for *B. hybridum* and S5.9 for *B. distachyon* climate class clustering in dendrogram format).



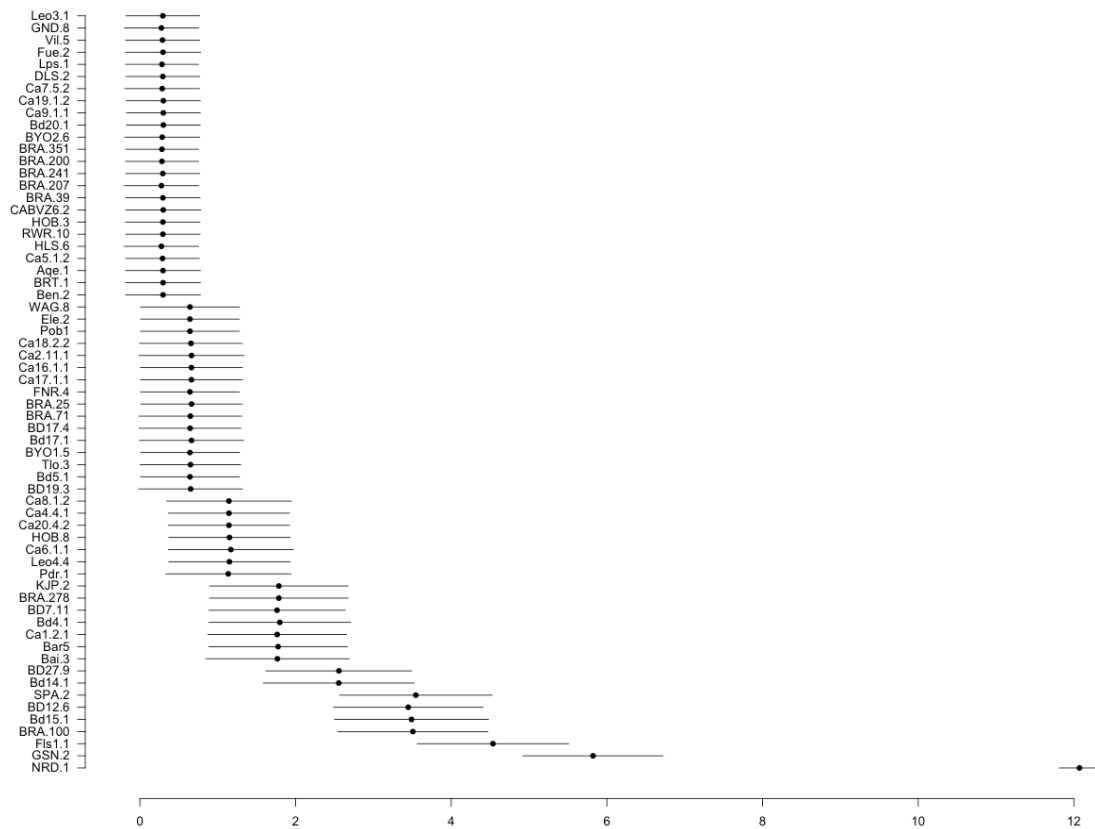
**Figure 5.7.** Climate diversity per region of *B. distachyon*. Regions near they Pyrenees in Iberia and Southeastern Mediterranean had the highest climate diversity. See corresponding figures 2.6 or 4.15



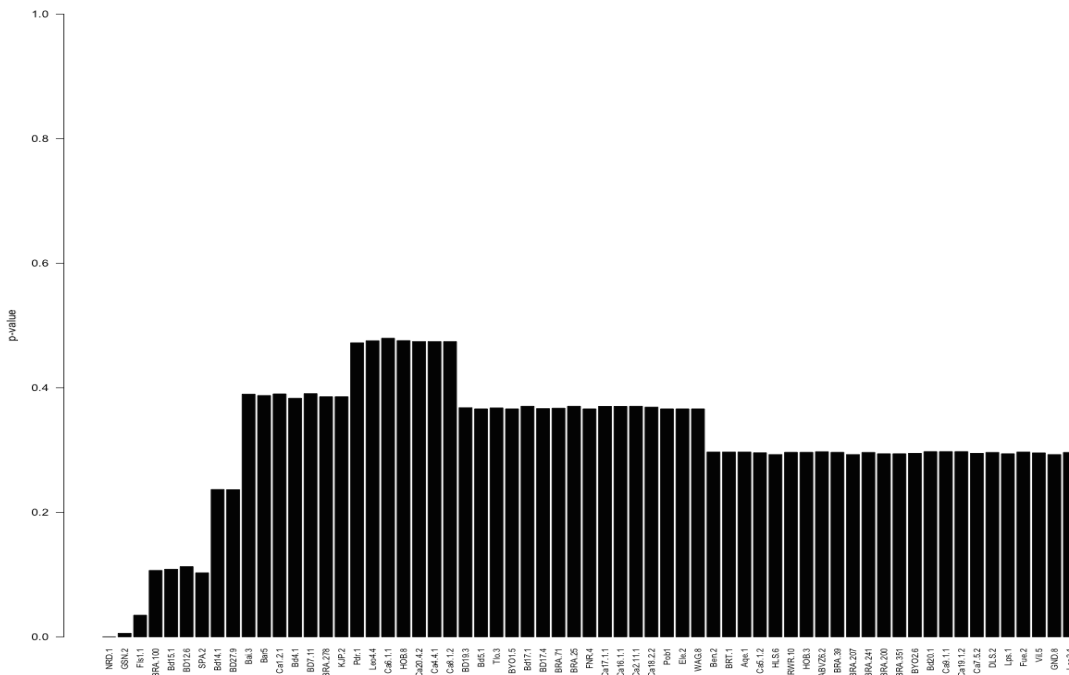
**Figure 5.8.** Climate diversity per region of *B. distachyon*. Regions near they Pyrenees in Iberia and Southeastern Mediterranean had the highest climate diversity. See corresponding figures 2.7 or 4.17.

### Genotypes with wide and narrow climate breadth

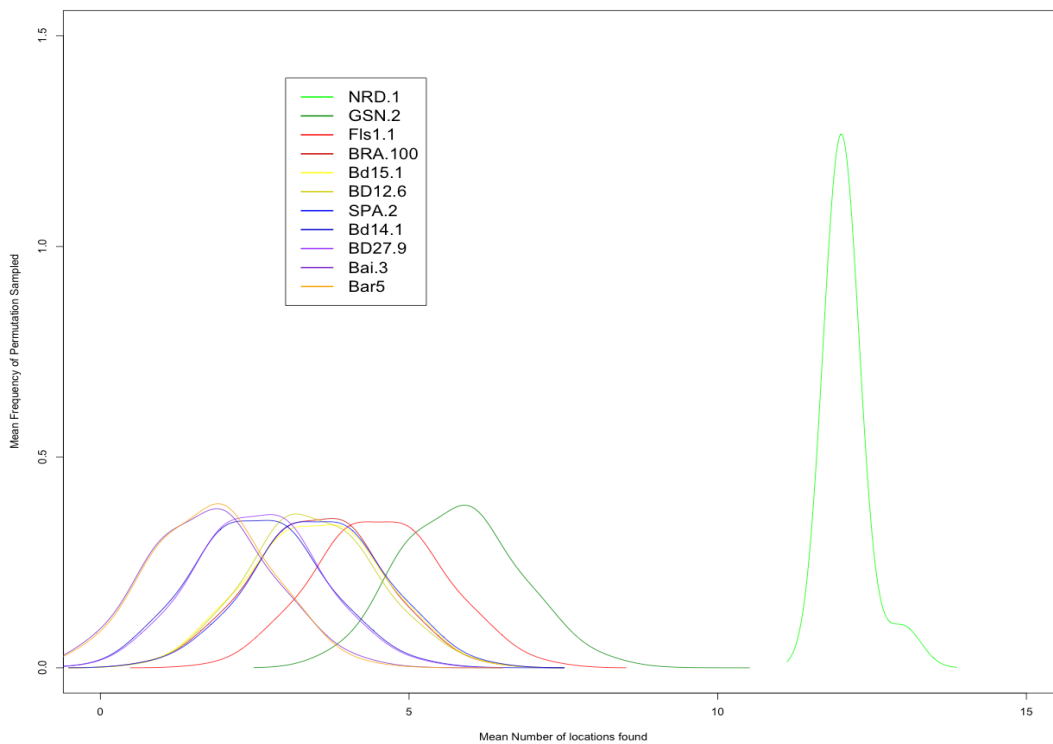
Permutation tests were used to determine breadth of climate from climate classes. Each genotype was tested independently across 1,000 iterations. In each iteration, a random number generator seed was set to a new number to insure the same sampling combinations were not reused. To qualify as testable groups, genotypes that were present in only one climate class were removed from the data set. The same 303 locations that *B. hybridum* was found in were used for *B. hybridum* tests, sites that didn't contain *B. hybridum* were not used. For each iteration, a random set of locations was drawn at the same rate each genotype was found, NRD-1 was found in 51 locations, thus 51 random locations were drawn per iteration. Each time a genotype is present in a location, the climate class it was found in is counted once and totaled. The total number of unique climate classes a genotype was found in was compared to the average number of classes from all genotypes. The genotype NRD-1 averaged 12.1 climate classes per 1,000 iterations as compared to the average of 2.3.



**Figure 5.9.** Permutation test of common genotypes and their individual climate breadth compared to the species as a whole and +/-1 standard deviation. NRD-1 was most present across climate breadth of the species.



**Figure 5.10.** P-values of permutation tests calculated by presence of a genotype compared to random sampling. NRD-1 was the only genotype that had a p-value lower than 0.05 significance and geographic abundance.

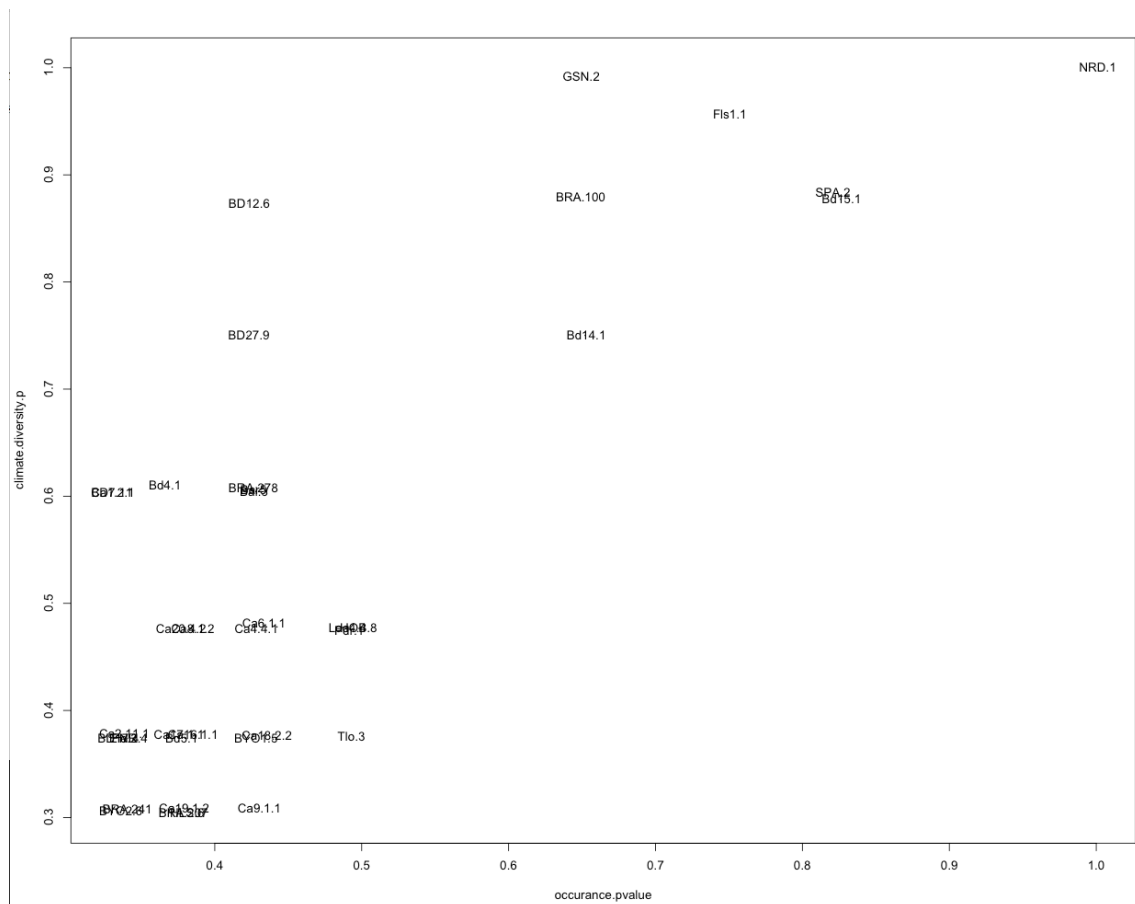


**Figure 5.11.** Density plots of the 11 most common genotypes and their permutation test average. Genotype NRD-1 had the largest number of climate presence across groups averaging >12 climate classes.

Genotypes of Geographic Diversity and Climate Diversity

Permutation tests are used to show the variance of a sample set and the variation of samples within a data set. As calculated in Chapter III, BRA-278 and NRD-1 had larger than normal geographic presence, and NRD-1 was the only genotype common across broad regions. Genotype NRD-1 was the only genotype to have significant geographic distribution by total

regions rather than just collection sites. To properly gauge the geographic abundance and the climate abundance of common genotypes, they can be plotted by their probability values of having presence across regions on a scale of 0-1 and also for their climate diversity on a scale of 0-1. These two scales can be used orthogonally to show what genotypes are more likely to be widespread and are plotted below (See figure 5.12). NRD-1 is in the upper right corner having both high geographic and climate diversity. Abundant individuals across geography always had higher climate diversity indicating that having a larger climate breadth is a sign of a group that could become widespread. However, there are genotypes that are climate abundant but were not found across broad geography indicating that dispersal ability is also a factor in becoming a dominant or "invasive"-like genotype.



**Figure 5.12.** Regional Occurrence by Climate presence from permutation tests. The x-axis is probability value of presence in collection locations. The y-axis is the probability of presence across climate classes. Each axis is scaled from zero to one by their probability scores compared to average within each test. Genotype NRD-1 was both the most abundant across regions and had the largest climate breadth.

Genotypes with Significant p-values for Climate Class Diversity

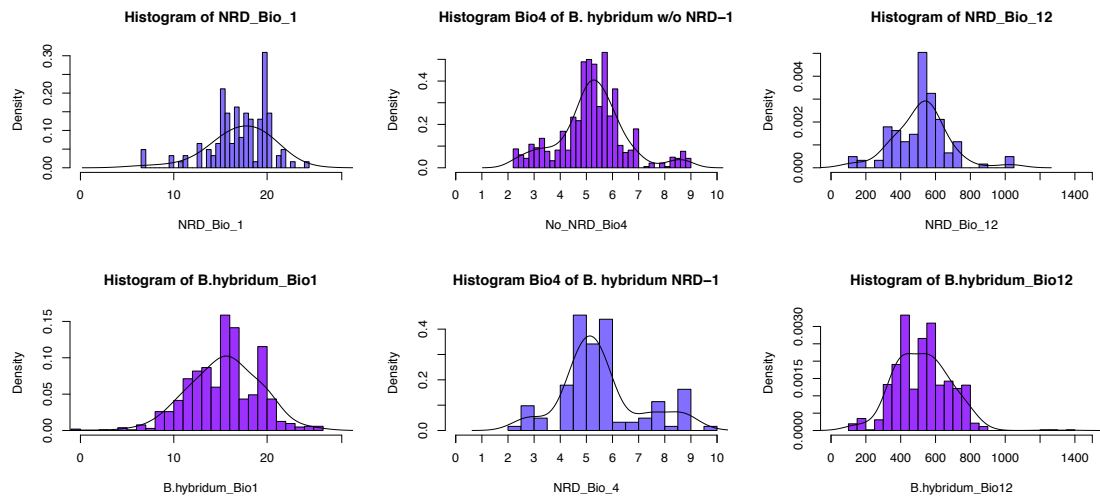
Three genotypes of *B.hybridum* tested as having significantly larger climate breadth than average. Genotypes NRD-1, GSN-2 and Fls1-1 all had a p-value < 0.01. Though GSN-2 and Fls1-1 have significant climate windows, neither had particularly broad geographic test results in regional diversity as calculated in Chapter IV. Leaving only NRD-1 with significant abundance across sites, regions, and climate classes.

Comparison of Mean and Standard Deviation in NRD-1 and all other *B. hybridum*

Tests of identicalness of mean, and standard deviation between NRD-1 and all other *B. hybridum* climate variables known to be significant to *B. hybridum*. If the data is assumed in a normal distribution, a t-test shows that all means of the climate variables are within significant limits to consider them identical. Assuming the data to be non-normally distributed, using a Wilcox test calculates that Bio1 is not significantly different between all other *B. hybridum* and NRD-1. The standard deviation between NRD-1 and all other *B. hybridum* in an F-test shows no significant difference between NRD-1 and all other *B. hybridum*. (See table 5.14). These tests show that NRD-1 mean and standard deviation are not that different than the species as a whole. However, it does show that NRD-1 is present across the species whole climate breadth of significantly important climate variables.

	p-value	Test	NRD-1	<i>B. hybridum</i>	other	Significant
<b>T-test</b>	<b>p =</b>	<b>T =</b>	<b>Mean/95% conf</b>	<b>Mean/95% conf</b>	<b>df</b>	<b>Significant</b>
Bio1	0.039	-2.087	16.52 / -1.20	17.14 / -0.03	142.13	F
Bio4	0.045	-2.020	5.22 / -0.568	5.51 / -0.01	146.41	F
Bio12	0.322	0.993	533.45 / -14.92	518.36 / 45.10	163.89	F
<b>Wilcox</b>	<b>p =</b>	<b>W =</b>	---	---	---	<b>Significant</b>
Bio1	< 0.001	45468	---	---	---	F
Bio4	0.460	54322	---	---	---	F
Bio12	0.503	58744	---	---	---	F
<b>F-test</b>	<b>p =</b>	<b>F =</b>	<b>Variance</b>	<b>Variance</b>	<b>Ratio Variance</b>	<b>Significant</b>
Bio1	0.001	0.598	0.451	0.771	0.598	F
Bio4	0.011	0.543	0.543	0.928	0.720	F
Bio12	0.192	1.206	0.910	1.555	1.206	F

**Table 5.13.** Tests of identicalness of mean, and standard deviation between NRD-1 and all other *B. hybridum* climate variables known to be significant to *B. hybridum*. T-test results show if a significant difference exists in normally distributed data. The significance of identicalness between two data sets of non-normally distributed data can be tested through a Wilcox test. The identicalness of standard deviation can be calculated by the F-test.



**Figure 5.14.** Comparison of Bio1, Bio4, and Bio12 between NRD-1 and all remaining *B. hybridum* as histograms with density curves. There are subtle differences in means and standard deviations, but cumulatively speaking there are insignificant differences in mean and standard deviation of all *B. hybridum* and NRD-1.

#### Does Regional Climate Variation Explain Genetic Variation

For *B. hybridum*, regions with high climate diversity often had more genotypes. To test if regions with high genetic diversity are associated with climate diversity an  $R^2$  test was performed on two distance matrices of regions. The 303 *B. hybridum* collection sites are grouped into 35 regions and individuals were reduced to a diversity set of 80 genotypes, and its collection locations by climate were grouped into 14 climate classes. Thus, distance matrix A used genotypes per regions, and distance matrix B used climate types per regions. An  $R^2$  test of these distance matrices with a p-value=0.06 showed a 4.07% explanation of climate diversity per region to genotype. Thus, climate diversity doesn't explain genetic diversity in *B. hybridum*.

## 5.4 Discussion

### Climate variable contrast of Species

Each species climate was compared based on BioClim variables that showed significant importance in species modelling individually in MaxEnt. By chance these climate layer variables are reasonably useful for describing the general climate trends for contrasting their breadth in precipitation and temperature annually. Three-way t-tests on the climate breadth of each species show where they overlap and differ. *B. stacei* was found in the warmest locations and averaged nearly 3°C warmer than *B. distachyon* and *B. hybridum*. However, *B. hybridum* had the largest breadth in annual temperature mean, spanning from 4°C to over 22°C average. This breadth in temperature exceeds the upper and lower limits of both diploids combined.

### Larger climate breadth of B. hybridum

The larger global surface area calculated for *B. hybridum* is reflected in its presence globally. Nearly seven million square km are deemed suitable climate, meaning that the larger breadth of possible climate opens more land space for colonisation. *B. distachyon* also had a high surface

area as suitable climate approximately 6.5 million km<sup>2</sup>, but the breadth of climate that *B. distachyon* is likely what limits the global models. The fact that *B. distachyon* habitat is so high in the native range, means that the climate types it prefers happen to be common in the Mediterranean region given that over five million km<sup>2</sup> are deemed suitable. It should be acknowledged that phenotypic traits are potential selection factors for “invasiveness” and that *B. hybridum* species phenotypes are what aids its breadth advantage of *B. distachyon* to survive in more diverse climates. If field trials in non-native regions were performed, the true climate breadth for *B. distachyon* could be further investigated. The climate breadth seen in *B. stacei* reflects the habitat difference between *B. stacei* and *B. distachyon*. *B. stacei* is common in warmer locations with lower precipitation, which are usually lower latitude and closer to the 30° +/- mark globally as seen in Koppen-Geiger Climate Classifications (Peel, 2007). The hypothesis that *B. hybridum* has larger climate breadth is accepted by statistical analysis of relevant climate variables.

#### Multi-linear regression models on Climate Variables

Both *B. distachyon* and *B. hybridum* were tested for association to all climate variables BioClim 1-19. The highest value that *B. hybridum* scored in partial multi-linear regression models is 0.33% to BioClim4 which does not conclude much genetic diversity is explained by this climate variable. *B. stacei* was found four times with intensive sampling in some places so association tests were not performed. *B. stacei* will need more sampling from across it's range to find adaptive alleles associating with a particular climate variable. *B. distachyon* was tested for association to climate with BioClim4 being the strongest partial multi-linear regression models test signal, the next closest was BioClim7 at 2.63%. A rotation of geographic coordinates against climate data was tested to see if climate variation was influenced by isolation by distance, however more tests are needed to truly claim BioClim4 explains genetic diversity in *B. distachyon* (See supplemental figure S5.9). Associated loci to specific climate variables will likely require whole genome sequencing to properly derive the causative variants, as GBS data is relatively sparse. Mapping of specific loci in association to BioClim variables was trialed and no significant regions had high association. However other studies have shown association to specific climate variables and listed loci and candidate genes (Dell'Acqua, 2014; Wilson, Streich, and Murray, 2017).

#### Association of Bio4 and *B. distachyon*

BioClim 4, or Temperature Seasonality, is the degree of temperature variation annually quantified by standard deviation of monthly average temperature. Essentially, the coefficient of variation by taking the mean temperature and dividing it by the standard deviation of monthly mean temperatures across a year in degrees Kelvin. In the natural world this would be analogous to how temperate areas have more mean temperature variation from month to month across the year than areas closer to the equator where temperature mean changes little from



month to month. Temperate regions can sometimes get nearly as warm as more equatorial climates but often for shorter durations and mean temperature will follow seasonal changes. It should be noted that deserts often have large daily temperature range/seasonality, being very warm in the day and cold in the evenings so temperature fluctuation monthly could require the same physiology for daily life year round in arid areas. However, in *B. distachyon* an association was found where 3.96% of genetic variation was explained by the percentage the standard deviation of monthly temperatures annually is as to the mean temperature. The exact cause of this association is still unknown at this time and is being further investigated beyond this body of work. This was also tested using a mantel and partial mantel test to account for geography and the mantel test has been proven to provide false positive results and other methods should be used to properly test climate association like Bayesian statistical methods (Guillot, 2013).

#### Geography and Climate Diversity

The overall diversity of both *B. distachyon* and *B. hybridum* helps elucidate why *B. hybridum* is more widespread than *B. distachyon* even though in Chapter IV *B. distachyon* was found to have a larger native potential area. Being that *B. hybridum* has a larger climate breadth than the other two species it makes sense that it would be more widespread and as in Catalan 2012 reports *B. hybridum* has a larger physical stature (Catalan, 2012). One potential reason for the larger climate diversity is the fact that *B. hybridum* is a polyploid of two species with different climate preferences with some overlap to encounter each other. When they hybridized much of the native range of either diploid was also habitable by *B. hybridum*. The fact that *B. hybridum* also showed significant genetic diversity in areas with climate diversity also seems logical. The more climate types a species can encounter within a small space, the more likely it would be to adapt to rapidly changing climate gradients in a short distance. This logic holds mostly true with *B. distachyon* as well in regions 10 and 11 where substantial genetic diversity was found in regions with climate diversity. The above-mentioned areas of high genetic diversity could make excellent field study regions for their respective species.

#### Climate Classes and Genetic Diversity

Permutation tests were able to reveal statistical significance that some genotypes had broad and narrow geographic breadth in Chapter IV. *B. hybridum* was the only species to have large presence across continents with hundreds of locations to use as test data, thus *B. distachyon*, with only two foreign locations, was not included in permutation test. The genotype NRD-1 is a geographically widespread lineage both in number of sites and in regions. A similar test was applied to see if NRD-1 or other genotypes were also present across broad climate breadth of the species *B. hybridum*. Climate data was obtained for each geographic location of *B. hybridum* and clustered into 14 groups, thus spanning the climate breadth of all collection locations in this study. Through 1,000 iterations, the random sampling of locations for the presence of a

genotype and the number of climate types revealed that climate classes specific to a species does determine the climate breadth of a sub-lineage. It should also be noted that regions with high genetic diversity were strongly associated with regions with high climate diversity. This could be a new recognised phenomenon in landscape genomics to see how correlated a species is with local climate variation within distribution limits. The *B. hybridum* genotype NRD-1 was found in more climate classes than random, therefore the hypothesis that some genotypes are more climate diverse than change is accepted.

Defining the genetic subgroup is an area of some subjectivity. The use of population structure for ancestral lineages could be one method of analysing climate diversity, however ancestral groups tend to be within  $K2$ - $K13$  groups and lower numbers may not be that intuitive in studies with large numbers of individuals (Pritchard, 2000; Janes, 2017). Also, some ancestral groups could be significantly more abundant and bias the results of testing abundance and climate breadth simply because they are older lineages, which could also harbour more allelic diversity. However, the use of population structure should be used to scan for adaptive variants associated with wide-dispersal or broad climate tolerance. In this study near clonal groups were used as common genotypes; the genetic variation between near clonal groups could distort the statistical ability to detect causative variants. For instance, the level of resolution to call genotype in this study was considerably high from using highest branch length of technical replicates. It is possible that the cut height in this study was too high. The height used to cut a dendrogram would classify near-clonal lineages as the same genotype, but is still too broad to declare a group as having wide climate-breadth, or wide geographic breadth. Using whole genome data would remedy the resolution to compartmentalise individuals into genetic groups and possibly more coverage across the genome.

## 5.5 Data Sets and Script Links

---

### Repository

- <https://github.com/borevitzlab/GBSFilteR>.

### Per Species

#### *B. distachyon* script and data

[https://github.com/borevitzlab/GBSFilteR/blob/master/Streichj\\_PhD\\_ANU\\_BorevitzLab\\_Genotyping\\_Bdistachyon.R](https://github.com/borevitzlab/GBSFilteR/blob/master/Streichj_PhD_ANU_BorevitzLab_Genotyping_Bdistachyon.R)

<https://github.com/borevitzlab/GBSFilteR/blob/master/kmeansDistachyon.txt.zip>

#### *B. stacei* script and data

[https://github.com/borevitzlab/GBSFilteR/blob/master/Streichj\\_PhD\\_ANU\\_BorevitzLab\\_Genotyping\\_Bstacei.R](https://github.com/borevitzlab/GBSFilteR/blob/master/Streichj_PhD_ANU_BorevitzLab_Genotyping_Bstacei.R)

[https://github.com/borevitzlab/GBSFilteR/blob/master/streichj\\_Stacei\\_hapmap.txt](https://github.com/borevitzlab/GBSFilteR/blob/master/streichj_Stacei_hapmap.txt)

#### *B. hybridum* script and data

[https://github.com/borevitzlab/GBSFilteR/blob/master/Streichj\\_PhD\\_ANU\\_BorevitzLab\\_Genotyping\\_Bhybridum.R](https://github.com/borevitzlab/GBSFilteR/blob/master/Streichj_PhD_ANU_BorevitzLab_Genotyping_Bhybridum.R)

<https://github.com/borevitzlab/GBSFilteR/blob/master/kmeansHybridumH.txt.zip>

## 5.6 Citation

---

Akin, WE, (1991). *Global Patterns: Climate, Vegetation, and Soils*. University of Oklahoma Press. p. 52. ISBN 0-8061-2309-5.

Bond, W. J., & Keeley, J. E. (2005). Fire as a global 'herbivore': the ecology and evolution of flammable ecosystems. *Trends in ecology & evolution*, 20(7), 387-394.

Brown, T.B., Cheng, R., Sirault, X.R., Rungrat, T., Murray, K.D., Trtilek, M., Furbank, R.T., Badger, M., Pogson, B.J. and Borevitz, J.O., (2014). TraitCapture: genomic and environment modelling of plant phenomic data. *Current opinion in plant biology*, 18, pp.73-79.

Catalán, P., Müller, J., Hasterok, R., Jenkins, G., Mur, L.A., Langdon, T., Betekhtin, A., Siwinska, D., Pimentel, M. and López-Alvarez, D. (2012). Evolution and taxonomic split of the model grass *Brachypodium distachyon*. *Annals of Botany*, 109(2), 385-405.

Catalan, P., López-Álvarez, D., Díaz-Pérez, A., Sancho, R., & López-Herránz, M. L. (2015). Phylogeny and evolution of the genus *Brachypodium*. In *Genetics and genomics of Brachypodium* (pp. 9-38). Springer, Cham.

Catalán, P., López-Álvarez, D., Bellosta, C., & Villar, L. (2016). Updated taxonomic descriptions, iconography, and habitat preferences of *Brachypodium distachyon*, *B. stacei*, and *B. hybridum* (Poaceae). In *Anales del Jardín Botánico de Madrid* (Vol. 73, No. 1). Consejo Superior de Investigaciones Científicas.

Dell'Acqua, M., Zuccolo, A., Tuna, M., Gianfranceschi, L., & Pè, M. E. (2014). Targeting environmental adaptation in the monocot model *Brachypodium distachyon*: a multi-faceted approach. *BMC genomics*, 15(1), 801.

DeC, R. (1906). The Classification of Climates: II. *Bulletin of the American Geographical Society*, 465-477.

Des Marais, D. L., & Juenger, T. E. (2015). *Brachypodium* and the abiotic environment. In *Genetics and Genomics of Brachypodium* (pp. 291-311). Springer International Publishing.

Eckert<sup>1</sup>, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González-Martínez, S. C., & Neale, D. B. (2010). Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*, 185(3), 969-982.

Eckert<sup>2</sup>, A.J., Wegrzyn, J.L., Cumbie, W.P., Goldfarb, B., Huber, D.A., Tolstikov, V., Fiehn, O. and Neale, D.B., (2012). Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. *New Phytologist*, 193(4), pp.890-902.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, 17(1), 43-57.

Geiger, R. (1954). "Klassifikation der Klimate nach W. Köppen" [Classification of climates after W. Köppen]. Landolt-Börnstein – Zahlenwerte und Funktionen aus Physik, Chemie, Astronomie, Geophysik und Technik, alte Serie. Berlin: Springer. 3. pp. 603–607.

Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, 4(4), 336-344.

Halenka, T., Belda, M., Kalvova, J., & Holtanova, E. (2013, April). Climate classification revisited: From Köppen to Trewartha for models evaluation. In *EGU General Assembly Conference Abstracts* (Vol. 15, p. 13374).

Hancock, A.M., Brachi, B., Faure, N., Horton, M.W., Jarymowycz, L.B., Sperone, F.G., Toomajian, C., Roux, F. and Bergelson, J., (2011). Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, 334(6052), pp.83-86.

- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, 25(15), 1965-1978.
- Janes, J. K., Miller, J. M., Dupuis, J. R., Malenfant, R. M., Gorrell, J. C., Cullingham, C. I., & Andrew, R. L. (2017). The K= 2 conundrum. *Molecular ecology*.
- Johnson, S. R., & Young, D. R. (1993). Factors contributing to the decline of *Pinus taeda* on a Virginia barrier island. *Bulletin of the Torrey Botanical Club*, 431-438.
- Köppen, W. (1936). The geographical system of climate. *Berlin, Germany*.
- Rubel, F., & Kottek, M. (2010). Observed and projected climate shifts 1901–2100 depicted by world maps of the Köppen-Geiger climate classification. *Meteorologische Zeitschrift*, 19(2), 135-141.
- Lasky, J. R., Des Marais, D. L., McKAY, J. O. H. N., Richards, J. H., Juenger, T. E., & Keitt, T. H. (2012). Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate. *Molecular Ecology*, 21(22), 5512-5529.
- López-Alvarez, D., López-Herranz, M. L., Betekhtin, A., & Catalán, P. (2012). A DNA barcoding method to discriminate between the model plant *Brachypodium distachyon* and its close relatives *B. stacei* and *B. hybridum* (Poaceae). *PLoS one*, 7(12), e51058.
- López-Álvarez, D., Zubair, H., Beckmann, M., Draper, J., & Catalán, P. (2016). Diversity and association of phenotypic and metabolomic traits in the close model grasses *Brachypodium distachyon*, *B. stacei* and *B. hybridum*. *Annals of botany*, 119(4), 545-561.
- Manzaneda, A. J., Rey, P. J., Anderson, J. T., Raskin, E., Weiss-Lehman, C., & Mitchell-Olds, T. (2015). Natural variation, differentiation, and genetic trade-offs of ecophysiological traits in response to water limitation in *Brachypodium distachyon* and its descendent allotetraploid *B. hybridum* (Poaceae). *Evolution*, 69(10), 2689-2704.
- Lee, C. R., & Mitchell-Olds, T. (2012). Environmental adaptation contributes to gene polymorphism across the *Arabidopsis thaliana* genome. *Molecular biology and evolution*, 29(12), 3721-3728.
- Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular ecology*, 21(12), 2991-3005.
- Peel MC, Finlayson BL, McMahon TA. (2007) Updated world map of the Köppen-Geiger climate classification. *Hydrol Earth Syst Sci* 11: 1633–1644
- Phillips, S. J., Dudík, M., & Schapire, R. E. (2004, July). A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning* (p. 83). ACM.
- Phillips, S. J. (2005). A brief tutorial on MaxEnt. *AT&T Research*.
- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945-959.
- Ream, T.S., Woods, D.P., Schwartz, C.J., Sanabria, C.P., Mahoy, J.A., Walters, E.M., Kaeppler, H.F. and Amasino, R.M., 2014. Interaction of photoperiod and vernalization determines flowering time of *Brachypodium distachyon*. *Plant physiology*, 164(2), pp.694-709.
- Sanderson, M. (1999). The classification of climates from Pythagoras to Koeppen. *Bulletin of the American Meteorological Society*, 80(4), 669-673.

Shen, X., De Jonge, J., Forsberg, S. K., Pettersson, M. E., Sheng, Z., Hennig, L., & Carlborg, Ö. (2014). Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. *PLoS genetics*, *10*(12), e1004842.

Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in MaxEnt: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, *21*(2), 335-342.

Wilson, P. B., Streich, J. C., Murray, K. D., Eichten, S. R., Cheng, R., Aitken, N. C., ... & Borevitz, J. O. (2018). Population structure of the *Brachypodium* species complex and genome wide association of agronomic traits in response to climate. *bioRxiv*, 246074.

Woods, D. P., Ream, T. S., & Amasino, R. M. (2014). Memory of the vernalized state in plants including the model grass *Brachypodium distachyon*. *Frontiers in plant science*, *5*, 99.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., & Guisan, A. (2008). Effects of sample size on the performance of species distribution models. *Diversity and distributions*, *14*(5), 763-773.

## Chapter VI: Final Discussion and Conclusions

---

### **6.1 Introduction section**

- Review concepts of Literature in Relation to this Thesis
- Introduction Events
- Isolation by Distance and Long-Distance Dispersal
- Genetic Diversity Studies in Relation to this Thesis
- Species Distribution Modelling
- Climate Biology Science

### **6.2 Discussion of Genetic Analysis**

- Species Identification by Sequencing
- Genetic Diversity for each species
- Genotype Resolution achieved
- Evidence of Non-Native Genotype Origins

### **6.3 Genomic Biogeography**

- Species Potential Areas Native and Globally
- Geographically Diverse Genotypes
- Species and Genotype Distribution
- Testing Distribution of Genotypes

### **6.4 Climate to Genetic and Geographic Data**

- Associating Climate to Genotype Data
- Climate Diversity and Breadth
- Climate Diverse Genotypes
- Climate Envelopes as a Species and Genotype
- Species to Climate Via Partial Mantel
- Scanning for Adaptive Genetic Loci

### **6.5 Final Discussion**

- Identifying species using genomic information
- Future Work in Genotyping
- Future Collections of Each Species
- Improving distribution modelling to a genotype level
- Association Studies Using Climate Data
- How this Dissertation Contributes to Invasion Biology
- Final notes

## 6.1 Introduction

---

### *Review of concepts of Literature in Relation to This Thesis*

Scans of genetic diversity can provide insight into the biological diversity of a species. Using genetic information from multiple locations can reveal patterns of genetic diversity across climate and geographic space, even across oceans and continents. This is especially true with studies of long-distance dispersal events of introduced species. Each scenario is unique with some studies finding clonal or near clonal diversity in non-native habitats and others unveiling more genetic diversity than expected. Saint Patterson's curse (*Eichium plantagineum* and *Eichium vulgare*) is an example of a low diversity introduction event where just a few introduced plants turned into an invasive pest (Konarzewski, 2012). A study of *Actotheca populifolia* and *Petrorhagia nanteuilii* also showed that introduced plants don't have to carry significant genetic diversity to be potential invasive species (Rollins, 2013). However, having low or high genetic diversity can have different consequences to a species ability to adapt. The degree of heterozygosity of introduced individuals, the genetic groups they hail from, and the number of introduction events, will ultimately define the introduced species level of plasticity. Even epigenetic factors could play a role in plasticity of adaptive traits, but many epigenetic marks have been found to be associated with genetic polymorphisms in the genome (Rollins, 2013; Eichten 2013; Eichten, 2016). *Brachypodium distachyon* showed significantly more potential area in the native range and genetic diversity than *Brachypodium hybridum*, yet *B. hybridum* had more climate breadth than *B. distachyon*. What phenotypes cause *B. hybridum* to be a more successful coloniser than the two other species was not investigated in this study, but *B. hybridum* is clearly more wide-spread than the other species based on herbarium records in GBIF and ALA (GBIF, 2016; ALA, 2016). The likely reason is that *B. hybridum* is a polyploid of two species that have relatively different climate preferences with little overlap in the native range, but are similar enough in genomes that their function as an allotetraploid genome is highly efficient. For *B. hybridum*, its success could both be a fixed heterosis state that allopolyploids have, on top of two genomes with an optimal amount of heterozygous-like effect of orthologous alleles between subgenomes. Studies have shown that *B. hybridum* has a larger stature and seed yield than *B. distachyon* and usually more stature and seed yield than *B. stacei* (Catalan, 2012; Catalan, 2016). Seed yield and physical size could be the functional difference between *B. distachyon* and *B. hybridum* given that *B. hybridum* has a larger climate breadth as seen in this study and others (Lopez *et. al.*, 2015). The climates that *B. stacei* prefers has been found to be rare in the Mediterranean area, but the areas that were modelled in Chapter IV of this thesis and in Lopez *et.al.* 2014 are not well sampled from. In fact, much of the sampling efforts in previous studies have largely missed the *B. stacei* range. The under representation of *B. stacei* makes its assessment of climate preferences and geographic diversity unresolved.

### Genetic Scans for Species Identity and Genetic Diversity

By using two reference genomes to identify species, the maximum amount of diversity was captured in each group, including the polyploid *B. hybridum*. Having a reference genome and samples from all three species genetic components make the *Brachypodium distachyon* species complex ideal for studying polyploid genetics, especially for grasses where the *Brachypoideae* is closely related to many agricultural species and the diverse climate and geographic space they occupy (Mur 2011; Draper, 2001). The two diploids also have annotation and small well-mapped genomes at 240mb and 266mb, leaving the tetraploid with a genome size  $\approx$ 512mb (Vogel, 2010). Once species were identified genetically, their genetic data was re-mapped to the species-specific genome and genetic relatedness was called. In *B. distachyon* 125 genotypes were found from 479 individuals, eight genotypes were found in 50 individuals of *B. stacei*, and 80 genotypes were found in 1,015 individuals of *B. hybridum*.

### Australia and North America Introduction Events

There are always new angles and perspectives to take to interrogate information collected in a study. What is considered to be a single species might be three species, possibly more and genetic testing with multiple genomes can give the dimensions needed to interrogate species diversity and identifying cryptic species as seen in Chapter II. If *Brachypodium distachyon* had not become a model species, its genome sequenced and cytology experiments to count chromosomes, those trying to curb its destruction in the United States would have less information about what it is and where it can and can't spread, what pathogens it might be susceptible to, and so forth. If an introduced species requires management and risk assessment, then a genetic analysis could be useful to understand the level of diversity a species carries in local areas. As a proposed example, if a species carries phenotypes that are localised in a specific geographic region that make its management more challenging, like herbicide resistance, then knowing where that genetic group is distributed would be helpful for land managers. Knowing where that groups geographic range overlaps with other groups of the same species could make for a more advanced and efficient management system to combat herbicide resistant weeds by preventing the gene flow of herbicide resistant alleles through different genetic groups with different climate preferences. To combat introduced species we must understand more of their genomic diversity, how and when they migrate, where they come from specifically, and their climate and geographic breadth.

As seen in a previous study, multiple introduced *B. hybridum* groups were found in the state of California in the United States (Bakker, 2008). The total number of genotypes in this study was not reported, but the structure value was at  $K=4$ , indicating a significant amount of genetic diversity (Bakker, 2008). The Bradford lab and the Borevitz lab further investigated the genetic diversity of California *B. hybridum*. In total from 26 locations and 187 individuals, we found 25



genotypes in North America. In Australia, the Borevitz lab surveyed 83 locations and found 38 genotypes in the SE region of the continent. Genotype NRD-1 was collected in both Australia and North America and is traceable to the Greater Mediterranean area, mostly near the eastern coastlines of the Mediterranean. Population structure was calculated for *B. hybridum* per subgenome and is shown in the Appendix section, as not being a core part of this thesis. The ancestral history of each species was not a focus of investigation, because the origins of *B. hybridum* are not as important as the current standing genetic diversity. An ancestral *K* could have little significance about a current invasive lineage. The combination of ancestral groups could show the admixture of specific lineages in native ranges, however the predominant ancestral groups didn't show much significance in admixture as a causative effect of widespread lineages. The much discussed genotype NRD-1's subgenomes are predominantly composed of a single *K* each with little admixture between groups (See figures S3.13-S3.17 in the Appendix).

#### Species and Genotype Biogeography

The search for new potential habitat in the native and non-native range yielded many locations that could harbour the study species. While these locations are speculative and subject to how informative species distribution models are, they do show what areas ideal for searching for more diversity. Distribution modelling at the level of genotype is especially speculative because there are little resources available to test a genotype's preferred climate without reciprocal transplant methods or use of modified growth chamber to simulate climate. The comparison of different genotype models to the expected distribution of a species as a whole using all observation locations could better inform researchers about how much of a species total distribution is represented by a few genotypes that are more widespread. It should be noted that the actual climate breadth of *B. hybridum* as calculated in Chapter V was just as wide as genotype NRD-1. This is especially important to note, because the outer climate limits of *B. hybridum* were not defined by NRD-1, and NRD-1 was not wider than average. So in the context of genotype modelling in Chapter IV, the larger predicted suitable geography, fundamental niche, of NRD-1 was only from having a smaller set of sample points in the model (NRD-1 had 51 observation locations, and *B. hybridum* as a whole had 303). What can be concluded is that NRD-1 has as wide of a climate breadth as the species as a whole, and that other genotypes through permutation tests had smaller climate breadth and were not as widely distributed.

#### Climate analysis of the *Brachypodium distachyon* Species Complex

The overall diversity in comparable BioClim variables showed that *B. hybridum* tolerated a wider range of annual precipitation, annual mean temperature, and the amount of variation in annual temperature compared to the mean. When searching for genetic variation associated with climatic variables only BioClim4 showed effect on *B. distachyon*. *B. stacei* is under represented

in this study to test alleles associating with BioClim variables and *B. hybridum* showed almost no association between BioClim data and genetic variation. However, after sorting sample locations into geographic regions, some of those regions harboured more climate diversity than others. Regions with climate diversity tended to have more genetic diversity as seen in Chapter IV and V. Specific genotypes had broader climate breadth as tested through permutation shown in Chapter V, but only NRD-1 had both broad geographic and climatic diversity with p-values < 0.01 in both tests.

## 6.2 Discussion of Genetic Analysis

---

### Species Identification

The identification of a species by genomic markers was complicated using only one reference genome and previous attempts to species identify all samples correctly were mildly accurate. The confounding of species identification by genetic analysis was not initially straightforward due to high amounts of structured diversity in each species and some samples aligning like polyploids that were usually *B. distachyon* B groups as seen in Chapter II polyploid subgenotype detection. With a small quantity of individuals representing each species, their identification by genetic analysis was relatively simple via principal components analysis and clustering methods. However, with many hundreds of samples and the addition of many genotypes, species identity was confounded by shared alleles in each species. Initially using principal component clustering system yields a distinction between *B. distachyon* and other species, but even some *B. distachyon* samples would cluster with other species. The methods I used were not written about in this thesis but was trialed many times. Once the *B. stacei* reference genome became available separation of the three species was simplified and measurable. Identification of species by using multiple reference genomes is becoming normal for metagenomics, such as the program FR-HIT that can align reads against multiple reference genomes (Nu, 2011).

### Genetic Diversity *Brachypodium distachyon* Species Complex

#### *B. distachyon*

Due to the strong structure between lineages, it's likely that *B. distachyon* is gathering genetic diversity by mutation accumulation with rare occasions of outcrossing. A total of 125 diverse genotypes are suitable to start for association studies, however more would be helpful and targeted collection efforts in diverse areas predicted by modelling should provide insight into the total genetic diversity of the species. Also, the use of whole genome sequencing rather than GBS would provide better resolution to find causative loci, being that this study's marker density averages about 18,426 bases per loci. As shown in the Appendix (Figures S3.19 and S3.20) section LD was calculated to decay to 0.10 via  $R^2$  at  $\approx 320\text{kb}$  using seven different sized

sliding windows averaged together with a mean length of 200kb (125kb-275kb). This leaves few markers within long LD blocks to calculate associations to climate or phenotype via GBS. Also, many of the lines used are previously developed homozygous lines that are eight-plus generations selfed and will have little reflection of true heterozygosity or true representatives of local diversity.

The regions near the Eastern Pyrenees Mountains in Spain and Southern France could be a great location for studying landscape genomics and sampling more diverse locations. Likewise, central Turkey north of Merçin also showed lots of diversity and should be re-sampled. The fact that two different accessions of *B. distachyon* were found in Australia is also significant with one having close relatives in Spain, and the other having close relatives in Turkey. There were 56 genotypes only found once and could be due to low coverage in these regions. In addition, re-sequencing these genotypes could reduce the number of actual genotypes called in this study, but would improve the resolution of known diversity of this species.

I will personally say that I believe one accession, PYR6-2, is likely from an internal error during harvesting or DNA/library prep. When I collected samples in this region and examined the maternal plant, none had any of the features common in *B. distachyon*, such short stature, or pubescent flowers, which are not definitively *B. distachyon* characteristics, but are common in the species. PYR6-2 did have small stature compared to any *B. hybridum*. It was even smaller than most diploids when grown in a glass house at ANU, and in my opinion, had similar growth features to the UKR lines that are known *B. distachyon* lines from the USDA and grow extremely slowly with very small leaves. However, an accession in that same library prep known to be a Pyrenees accession, ABR7, came up polyploid in species identification. Thus, I suspect these two lines were likely switched somehow. The accessions WLE1-1 and WLE1-2 both had the correct proportion of reads and physical traits (pubescent flowers) to qualify as *B. distachyon* physically and genetically. Due to the ambiguousness of PYR2-6, I would normally recommend it be re-sequenced. However, the S1 generation that was sequenced had an *Ergot*-like fungus on the seed and the plant was destroyed. The maternal collected plant from the location site is still at the Borevitz lab at ANU and can have more seed drawn to be planted and re-sequenced. Also, the location is documented and can be resampled from as well.

### *B. stacei*

The small amount of samples of *B. stacei* makes it hard to assess its true diversity. Having only 50 identifiable samples that pass filtering thresholds by sequencing from three areas also decreases the likelihood that this study composes a snapshot of the genetic diversity of the species. Ten more samples could be sequenced at higher coverage to improve our understanding of the genetic diversity from our current germplasm, as these samples failed genotype filtering,

but had enough sequence depth to obtain *B. stacei* candidature. At the start of this dissertation little was known about the  $2n=2x=20$  cytotype now known as *Brachypodium stacei*. After the species distribution modelling in Chapter IV, and the modelling in Lopez, 2015, the ranges of this species is better described and more collection efforts can be focused on where it resides. The collection from the Ezrati lab proved to be fruitful as it contained many *B. stacei* samples and reasonable diversity. More collection efforts can greatly improve the scientific community's understanding of *B. stacei*, plus the hybridisation event that created the allotetraploid *B. hybridum*.

### *B. hybridum*

Multiple genotypes were found of *B. hybridum* and collecting this species was the most simple in the native and non-native range. The fact that there are only 80 genotypes captured by 1,105 samples indicates it probably doesn't have much diversity compared to other *Brachypodium* species. This is probably because the species is expected to be about one million years old (personal comment, John Vogel). The true age of the species and if multiple hybridizations have occurred has yet to be shown. However, aligning to the *Triticum aestivum* mitochondrial genome revealed 26 GBS markers that showed strong evidence that the hybridization between a *B. distachyon* and *B. stacei*-like plants did occur more than once. As seen in other publications, many known *B. hybridum* samples aligned near diploid (Mur, 2011) In that study some *B. distachyon* and *B. hybridum* lines mapped similarly. Population Structure was also calculated in *B. hybridum*, however the focus of this thesis is about the standing genetic diversity rather than the ancestral origins. As mentioned above the population structure analysis is featured in the Appendix section of this thesis (See figures S3.14-S3.17). Linkage disequilibrium was also attempted on *B. hybridum* and was calculated to  $\approx 50$ kb across all samples for both subgenomes. The overall results for LD in *B. hybridum* subgenomes were difficult to calculate and 50kb is a very loose approximation. The SNP density of *B. hybridum* averages at 27,638 bases and LD dropped within one window motion using seven averaged windows of 200kb (125kb, 150kb, 175kb, 200kb, 225kb, 250kb, and 275). The quick decay, but low genetic diversity of *B. hybridum* hits an interesting sweet spot between few alleles per LD block, yet high enough outcrossing to show quicker LD decay than *B. distachyon* by an order of magnitude. Whole genome sequencing of a genetically diverse subset of individuals would be ideal to resolve LD decay and the true bases per variant across the *B. hybridum* subgenomes.

### Genotype Resolution Achieved

The use of genotyping by sequencing to capture sporadic genetic markers across the genome provided the necessary resolution to describe the genetic diversity of our study groups. Various levels of multiplexing and different sequencers were used as well across species, which required strict filtering regimes to keep samples comparable. Starting with many tens of thousands of

markers and reducing to 20-40% post filtering of the original marker count yielded 4,744 markers in *B. stacei*, 14,436 markers in *B. distachyon*, and 18,525 markers in *B. hybridum*. For genotyping purposes and relatedness calculations, this was more than enough and the Tassel pipeline proved to be adequate. As with any diversity study, and this study is no exception, the tracking of samples from start to finish can be a challenge for any researcher. Mistakes were likely made in small number per each step of the genotyping process. Since this study spans the efforts of collecting and sharing samples from nine labs, it's possible human errors were made. Up to 2,722 samples were collected in total, of which 1,818 samples were used and 1,573 accrued enough sequence depth and genome marker proportions to qualify for species candidature. Across all these lines and movement of material, it's no doubt that errors occurred, either by samples in the wrong packet, samples planted in the wrong pot, samples were harvested in the wrong well, samples possibly mixed up at sensitive moments of library dilution. Many individual samples of previously sequenced or cytological analysed accessions ended up in the wrong classification were flagged and removed. These samples should be sequenced again to properly determine their true relation to the collective growing germplasm of the *Brachypodium distachyon* species complex.

I currently am working on an R package that would greatly improve the ability to call genotype. The package would consist of a series of functions that would filter SNPs by their presence across samples, and also filter out low coverage samples. There will be many other features in this R package and some have already been developed in script format to test output and processing time. The key function to genotype samples will first order all individuals between the two most diametrically opposite individuals (least related two individuals). Then it will iteratively scan through each individual from top to bottom. During each scan two individuals will be compared for relatedness only by the markers shared between those two samples, any marker missing in an individual will be removed. This is a particularly ideal examination of relatedness because it accurately examines two samples with vastly different sequence depths. If all samples have a minimum of 10k variants then reducing already closely related samples to their shared markers will only reduce the amount of used markers to a very small extent. The percent difference in alleles across shared and covered loci will easily be examined and can be set by a user, ideally 1.5-2%. If two samples shared marker/alleles are under this threshold they would be classified as the same genotype. This process would continue as less and less related sample pairs are compared until the threshold is exceeded and then the samples that didn't break the threshold are: removed from analysis, listed in a matrix, and labeled a specific genotype ID. This method has been trialed on this data set and calls 399 genotypes of *B. distachyon* and 125 genotypes of *B. hybridum*. Since this is for discussion purposes the data output from these scripts is not listed here or in the Appendix section, but I have a website for the R package 'FilterVCF' I am developing: <https://sites.google.com/site/filtervcf/>

### Multiple introductions

When a species is introduced, they may or may not be known to survive locally and is dependent a priori knowledge about their native range climate dynamics. One study found that multiple introduction are quite rare and many cases have been found where introductions single events and still cause prolific damage to ecosystems and agriculture, example St. Patterson's Curse which was a single event and was traced to a garden in Victoria Australia (Konarzewski, 2012; Rollins, 2013). This could also be partially true since most introduced species are from the horticultural trade (Reichard, 2001; Drew, 2010). When a species is introduced, and possibly more than once, it's more likely to have genetic diversity per each introduction event and example is *Ambrosia artemisifolia* in Genton 2005, and *Brachypodium sylvaticum* in the U.S. states or California, Oregon, and Washington (Genton, 2005; Rosenthal, 2008). Both of these studies found significant genetic diversity of introduced populations of their respective species. These outcomes might not immediately be apparent nor any negative impacts manifested. When a species like *B. distachyon* (*B. hybridum*) is introduced to the new world as early as the 1780s in Australia and initially classified as native, then later reclassified as introduced near ≈1950, then multiple introductions likely have occurred and should be investigated genetic diversity (ALA, 2016).

### Isolation by Distance and Long-Distance Dispersal

*Brachypodium* species, particularly *B. hybridum* and *B. distachyon* appear to have little isolation by distance across their native range as seen in Chapter III. This could in part be due to anthropocentric and paleo-anthropocentric distribution of seeds (Opanowicz, 2008). There is some account of seeds of *Brachypodium* species being used as a food source in ancient human habitation sites and could explain part of why these species are so widely dispersed in the circum Mediterranean area (Draper, 2001). It should also be noted that *B. hybridum* has successfully colonised many parts of the globe in regions where wheat is commonly grown, likely being a contaminant of seed stock in the last few centuries during the colonisation of the new worlds (GBIF, 2016). *B. distachyon* does have some lineages that are isolated to specific areas, but they are still close relatives of trans-mediterranean groups. *B. hybridum* genotypes are even more widespread in their native range indicating larger dispersal.

### Origins of Long Distance Distribution

Australian *B. hybridum* are easily traceable to both east and west Mediterranean, mostly eastern, and to some extent Spain. If PYR6-2 is truly in Australia, there must be at least two introduction events, given that we found *B. distachyon* two times. The accessions PYR6-2 and WLE2-2 have easily traceable origins: one from Turkey WLE2-2, the other from Spain PYR6-2, and neither of these genotypes are found on both sides of the greater Mediterranean area. They likely came to Australia independently via two different ports and probably from two different events. To insure that these two *B. distachyon* lines are truly in Australia and not a chance mistake in the

first round of analysis, it would be ideal to sample these locations a second time to confirm their existence. Given the amount of genetic diversity in Australia in *Brachypodium* species, there likely have been many introduction events. The same goes for the state of California in the United States. We found several genotypes in California, however most of the California genotypes are traceable to the eastern Mediterranean regions with only a few lineages tracing to the whole Mediterranean, which were some of the most widespread genotypes globally. We also observed that the most widespread genotypes in the native range were common in introduced locations, example NRD-1 was found in 51 locations on four continents and 13 climate types.

### 6.3 Genomic Biogeography

---

#### Species Native and Global Potential Areas

The modeling of species over the last two decades is becoming more and more of an exact science, but improvements can be made. Reducing the size of a species model to the local study area greatly increases the likelihood of predicting the true range of the species (Elith, 2011). Global models can be very informative, but increasing the amount of surface area the model covers introduces more climate variation in non-predicted areas resulting in a biased model (Elith, 2011; Phillips, 2005; 2006; Phillips 2008). The larger the surface area the more points are needed, however too many points can make the model over-fit the designated surface area. This study was modelled after the parameters and study size of the Lopez, 2015 study specifically about these species. The study area was chosen to reflect the accepted limits of all three species distributions with the exception of extending the range further east. Based on GBIF records, there are more records of *B. distachyon* complex members further east than currently in the public germplasm and publications (GBIF, 2016). The GBIF repository doesn't distinguish the three species and states the species only as *B. distachyon*. Given that these locations listed in the repository for all locations only use the name *B. distachyon*, even those known to only have *B. hybridum*, it should be considered that any one of these species should be tested for potential suitability in these regions. The GBIF repository shows over 17k records of *B. distachyon* globally and many of these locations were observed in global models for *B. hybridum*. Thus, most of the GBIF locations are likely *B. hybridum*.

To properly expand and correct species distribution models, other methods can be proposed. The distribution modelling of each ancestral group individually and add them together, like in Chapter IV with each species, can better show the climate breadth and limits of each sub-group. Tools like ENMTools will try to un-bias a model based on diverse geographic locations. If the geographic locations are a minimum distance apart, the output from MaxEnt should have more climate diversity in the model. However, as demonstrated in this study, not all genotypes have the same climate breadth. While the exact climate breadth for each tested genotype is likely

larger than what was detected in this study, some individuals had little climate diversity. Thus some sort of genetic subset would better reflect the sub-groups and the species distribution as a whole. Another option is to examine the whole species, but only use the most climate diverse sites with equal weighting. Two locations could be less than a kilometer apart, but have radically different climates, examples being rapid elevation change, or proximity to oceans. Thus locations should be used based on their climate variables. In this study climate classes were assigned to each collection site, thus a minimum of 10 sites from each class could be used to model a species distribution and would bypass the need for genetic sequencing, though that may also be of interest.

### *B. distachyon*

The native potential area for *B. distachyon* has mostly been sampled in the east and west Mediterranean. Since much of the central Mediterranean regions are sea, there are few locations where *B. distachyon* will be found. North Africa in Algeria, Morocco, and Tunisia all model as suitable habitats for *B. distachyon* in this study and in Lopez, 2015. Sampling from these regions could reveal more genetic diversity and unique lineages that are not yet in the public germplasm. It could also be possible that the Northern regions of the African continent were refuge for *B. distachyon* as glacial maximums would have pushed the species further south out of Europe and Asia (Mitchell-Olds, 2001; Beck, 2008). These regions could be harbouring individuals or rare lineages that didn't migrate back to northern suitable space. Southern France and Northern Italy also showed potential habitat and would be interesting sample locations given their distance to the Iberian Peninsula. Accession ABR2 was found near Montpellier in southern France so the species is confirmed in these regions as well as ABR8 found in Tuscany Italy near Siena. Parts of Macedonia, Bulgaria and mainland Greece showed high suitability for *B. distachyon* and could help close the east-central gaps of the collected range of *B. distachyon*. Since ABR9 was collected in Slovenia and is further from Turkey than this region, it's likely that these countries could harbour *B. distachyon* as well. Many islands also showed suitability scores high enough to potentially contain *B. distachyon*, particularly islands of Greece, Cyprus, and islands of Italy and France: Corsica, Sicily, and Sardinia. These locations often modelled better for other species so their investigation could be merited for multi-species collections.

The potential global distribution of *B. distachyon* could be higher than expected. In my own experience many people made personal comments that only *B. hybridum* would be found outside of Europe and Asia based on it being found already numerous times in the past outside of its native environment. Finding true *B. distachyon* on two independent occasions in Australia shows potential global areas that are modelled as suitable should be investigated. Areas of the United States in the state of Washington were investigated extensively by the Borevitz Lab, but not to a full extent and could be searched further. Areas of China and the



Himalayas modelled as suitable space and being native to parts of Eurasia, these easterly locations could harbour distant lineages both geographically and genetically.

#### *B. stacei*

The small amount of sample locations (four locations with 50 samples) of *B. stacei* available for species modelling make its predicted potential native area have less confidence. The habitat of *B. stacei* compared to *B. distachyon* show they overlap very little in their native range. Only in select parts of the eastern and western Mediterranean do these species ranges overlap: Israel, Libya, Morocco, Spain, Greece, and Cyprus. Many parts of southern Spain and much of Morocco show high suitability for *B. stacei* as well as the Canary Islands in the Atlantic. The presence of *B. stacei* in North Africa in Lopez, 2015, and climate analysis shows that *B. stacei* prefers drier warmer habitat at lower elevations than *B. distachyon* and is also reflected in Chapters IV and V in this thesis. Much of Greece and southwest Turkey show high suitability as well for *B. stacei*.

In central West Africa, Angola showed high suitability of *B. stacei*. There are other *Brachypodium* species known to inhabit the southern parts of the African continent like *B. flexim*. It could be that *B. flexim* is a close relative of *B. stacei* and inhabits this region because of similarity of preferred habitat. In the a previous study, *B. flexum* and *B. mexicanum* both aligned as close relatives of *B. stacei* (Catalan, 2015). It could be worth the effort to collect from the native range of each of these species to investigate how certain *Brachypodium* species tolerate high temperatures and little annual rainfall. This is especially true since they are all found in dry warm areas and are close relatives to *B. stacei*, but found on different continents: *B. stacei* in North Africa, Asia, and Europe, *B. flexum* in south and central Africa, and *B. mexicanum* in North and South America. Regardless, the coastal areas of Angola south of Luana are the only highly suitable habitats found for *B. stacei* outside its native range. A few other locations showed mild suitability on the South American continent in Chile near the city Santiago, and the Australian continent near the city Perth on the southwestern coastline. The only other location that modelled mild suitability is near the city Los Angeles in North America in the state California in the United States.

#### *B. hybridum*

As a species, *B. hybridum* was modelled to have less potential area in the native range than *B. distachyon*. However, global models showed significant more potential area than *B. distachyon*. With the above-mentioned discussion about GBIF records in their species repository, it's likely that nearly all the observed non-native individuals are not *B. distachyon*, but *B. hybridum*. Given that true *B. distachyon* were observed in Australia and some non-native regions were also modelled as suitable climates, it would be worthwhile investigating the species composition in these locations. In Australia, the locations that harboured *B. distachyon* also contained *B.*

*hybridum* and much of their native range models in Chapter IV and Lopez *et al* 2015 show these species overlap significantly, and Chapter V shows their overlap in climate breadth. Native models also showed a significant amount of the Eastern Himalayas being potential habitat in Nepal and India which matches the GBIF repository observations that were not included in the model for *B. hybridum* native and global predictions.

#### Geographically Diverse Genotypes

Eleven *B. hybridum* genotypes were found across five continents and found more than a hundred times cumulatively. One genotype, NRD-1 in particular, was represented by 121 samples and found in 51 locations on four continents. Other more rare genotypes were found in non-native locations and could be in areas currently unexamined. Based on the resolution of the data set in this study it is clear that some genotypes are much more common globally than others, but that doesn't mean that other locations that have not been examined don't harbour what are currently deemed less common genotypes. In fact, it could be the opposite. As with *B. stacei* we are only just starting to understand where it grows as seen in this study and again in Lopez, 2015. Once the climate breadth of a species is better understood the breadth of a species can be described. A few genotypes in *B. distachyon* transcend the east west divide of the Mediterranean. To truly say these samples represent one cumulative genotype could be stretching the description used to call genotype. Nevertheless, these samples are very closely related and this lineage is spanning large distances. The resolution to call genotype in this study is subject to the comparable-ness of loci GBS captured, and as discussed next, could be improved with whole genome sequencing.

#### Testing distribution of genotypes

Genotype level modelling of potential area is no doubt controversial, and methods will need serious testing to conclude what is truly the “genotype distribution” compared to the species distribution. One proposed method is to create a permutation system to see if surface area is predicted differently than by chance. MaxEnt models, especially global models, generate enormous amounts of data that can quickly fill a computer hard drive. By designing a custom pipeline through shell scripting it could be possible to generate a random list of input points from a list of coordinates of species observations. These random subsets of observation points will be a set length about the same as the number of observations of a genotype. By permuting MaxEnt sessions and purging excess MaxEnt data and calling a custom R script to calculate surface area statistics and save them in a data matrix, one could hypothetically test if a genotype distribution is different than the species with measureable differences and also plot these differences. This method was in development during the end of this study and will be developed further in the near future. This form of metric could also show the differences in biodiversity with more accuracy than the method employed in Figure 4.18 in Chapter IV. The idea of sub-setting models isn't necessarily unheard of. A similar concept was found in a paper from the

Purugganan lab in Banta et al. (Banta, 2012). However, the sampling methods employed by our collaborators across the Turkish landscape were thorough enough to show that genotypes were found multiple times in different geographic areas and in enough abundance to create distribution maps of each genotype. Update: I have started building this method, however calling MaxEnt from bash-shell command scripts appears to be more challenging on our local cluster computer and fails to launch MaxEnt. I've had to re-configure how to launch MaxEnt from an shell script. I have easily been able to get this functioning on a personal computer, but are too slow to proceed with 1,000 models, completing about 2-3 model iterations per day and would take too long. Therefore, efforts to make this technique functional have not been pursued, but a modified version can likely get nearly the same result, but changing the parameters of MaxEnt. This technique is still ongoing in development.

Like genotype modelling performed in this study, other studies have analysed the overlap of species. The Moritz lab is known for their aggressive progress in detecting hybridisation zones and species overlap modelling (Carnival, 2016; Rosaur, 2015). Without describing the genetic diversity of a study group, running a MaxEnt analysis or species distribution model could be biased based on the input of the samples by relatedness. One or two widespread genotypes can't represent the whole of a species distribution, but as was seen in *Brachypodium distachyon* it was actually found to be three species after investigation. On top of being three species, each of these species had different ranges, climate tolerances, and some of the genetic groups within true *B. distachyon* and *B. hybridum* were more widespread than others, which would bias the average species range. Another point to make is that one genetic group could be under sampled, thus certain genetic components of that species are also biased against in species distribution modelling. It could also be proposed to jackknife test a species model by removing observation points of genotype groups to measure their contribution to the final distribution model. This could better show the importance that climate variables have per genotype and the species cumulative model.

## **6.4 Climate to Genetic and Geographic Data**

---

### Associating Climate to Genotype Data

Different genetic lineages may occupy different geographic ranges and corresponding climate envelopes such as what was seen in *Setaria* species in Huang *et. al*, (Huang, 2014). These can highlight potential adaptive differences however alleles providing the advantage of large climate envelopes may be fixed within each lineage. Hybrid zones between groups or long range dispersal followed by admixture could provide the opportunity to uncouple adaptive alleles from their genomic background and using genotype level modelling could help detect those regions. Digital herbarium records are a great place to start describing the climate range of a

species utilizing existing collections. Records often have metadata about microclimate, some phenotype data including whether the plant was flowering, in addition to when and where the plant was collected. This can aid a researcher's decision about when and where to travel for collecting, what trait(s) to look for, and what locations they occupy. With collection points, researchers can also travel back to the same location or similar locations based on model predictions like those created in computer programs to predict niche breadth using climate envelopes and species distribution modelling software like MaxEnt (Phillips, 2006; Joost, 2007; Banta, 2012).

#### Species to Climate via Partial Mantel Tests

No significant amount of genetic variation in tetraploid *B. hybridum* was explained by climate with only BioClim4, temperature seasonality, at 0.33%. It might be better to test each sub-genome independently since each sub-genome was once a diploid species that faced its own selection regime before hybridisation. Suites of subgenome alleles found in a tetraploid like *B. hybridum* could correspond more with some diploid genotype distributions than using both subgenomes cumulatively. It has already been seen that *B. distachyon* and *B. stacei* have different and rarely overlapping distributions, it could be that specific genetic groups with their own proprietary climate preferences participated in the polyploidisation event and some of those markers remain present in the genome. Their effect could be diluted and non-significant in their current polyploid state, but at one time could have been adaptive. If enough surviving relic close relative diploids of a polyploid subgenome were captured, their current location's climate data could be used to investigate potential paleo origins.

#### Scanning for Adaptive Genetic loci controlling survival of fitness

Fitness is the result of selection at many different life stages and often differs across environments and involves many loci and possibly intensive surveys in native landscapes or modified growth chambers. An indirect way is to use the presence of a plant in a location as an observation that a particular combination of alleles can survive there. When a large collection of plants spanning a suitable range of climate variables is sequenced, one can use environment data as a fitness phenotype to test if certain alleles are associated. A genome wide association study scan can then be performed with climate data to identify adaptive loci or if genetic variation is explained by a climate variable. Though an association was found with *Brachypodium distachyon* with a climate variable. There was not enough time in the project to investigate that here. However example landscape studies and reviews do discuss in detail the ability to capture alleles associated with adaptive traits examples are Weigel, 2015 and Kesari, 2012.

### At What Point Do Climate Envelopes Apply

When thinking about climate, the concept of a species is a lot like the concept of a genotype. A species can be thought as having a certain potential and realised distribution across habitable space - there are the places it can and cannot grow. A genotype would have equal or smaller climate breadth than the limit of that species as a whole. A good example of this is the *B. hybridum* genotype NRD-1, which was found many times on multiple continents and in many climate types. NRD-1 was found in so many locations and in many climate types it had just a wide of climate breadth than the species *B. hybridum* as a whole. NRD-1 could have a significant role in defining the fundamental niche of *B. hybridum* as compared to other genotypes.

Some alleles or genes can be adaptive in some locations or they can be neutral or maladaptive in other locations or scenarios (Smith, 2008). This makes testing the effect a gene has difficult in a species like those of the *Brachypodium distachyon* complex with long LD and rare outcrossing events (Tyler, 2016). Adaptive alleles are likely inherited with neutral alleles and with rare outcrossing events many loci will appear adaptive, the whole genome may even statistically prove to be adaptive by not outcrossing enough for accurate detection of genetic loci causing a phenotype. The best way to test for adaptive alleles in *B. distachyon* would be a high resolution landscape study with multiple transects leading across a significant environmental gradient where two plus genotypes are found on both sides of the gradient. The country of Turkey would be ideal due to the significant topographical variation in geography and genetic variation. Figure 4.17 in Chapter IV shows a map of the region of Turkey and the likely locations to find high and low genetic diversity. By sampling across and between locations in that model could provide ideal sampling locations to test adaptive alleles. The same sort of study would be true for *B. stacei* and *B. hybridum* in Israel given the amount of genetic diversity and climate diversity found across the country.

## **6.5 Final Discussion**

---

### Future work

The science of landscape genomics has made tremendous strides in the last few decades with sequencing techniques becoming more tangible and amenable to more research interests. Beyond sequencing being affordable, the computational tools have increased as well. The statistics program R now has many packages and functions for population genetics, landscape analysis, GIS studies, and more. The same goes for python and other languages. The days of whole genome analysis are basically here and the existing tools for analysing genomic data are improving. Thus the previous tools like MaxEnt and other landscape/GIS analysis software should be improved to accommodate more accurate descriptions of species. Though the use of GBS data has performed well to analyse the three study species for relatedness and association

to climate variables in *B. distachyon* with BioClim4, the investigation of causative variants is at a standstill. Some attempts to map to specific loci were made, but were not included in requirements of this dissertation and removed. The investigation continues beyond this body of work as mentioned before.

### Future Collections

Extensive public collections of *Brachypodium* exist with over 1,060 accessions published and at least 181 accessions available from the USDA National Plant Germplasm System, including 141 *B. distachyon* (ars-grin.gov/npgs/) (Catalan, 2012; Dell'Acqua, 2014; Draper, 2001; Filiz, 2009; Hammami, 2014; Mur, 2011; Vogel, 2006a). At least a further 3,000 accessions are estimated to be present in private collections and available by request or through collaborations. However, many geographic regions remain unrepresented or under-sampled. Having those spaces filled would be beneficial for the whole *Brachypodium* community. As mentioned, above there is a strong division in the current public collection between Eastern and Western European accessions. Hence, it would be advantageous to have collections across the Middle East and North Africa as it is highly possible that *B. distachyon* complex species were pushed south during the last ice age and extant lineages could be sources of new maternal lines for research as seen in other species like *Arabidopsis thaliana* (Lee, 2017). Central Southern Europe might provide a source of admixture populations between the Eastern and Western genotypes or completely new genotype groups. There are records of *B. distachyon* complex species occurring in northern latitudes in the UK, Belgium, Germany, and France and also warrant investigation and collection. Non-native locations that could benefit from collection are South America in Argentina near the wine growing regions in the Pampas, the wheat and wine growing regions of Chile, and the lower elevations of the Andes from Peru to Colombia. Also, western Australia has many locations known to harbour *Brachypodium* species as well as coastal South Africa (ALA, 2016; GBIF, 2016).

With genotype data, we can focus further collections on hybrid zones and polymorphic sites to increase the number of recombinant genotypes, as seen in Chapter III and IV in this thesis, regions have been identified of high genetic and climate diversity and more sampling and smart transects could provide better insight in local adaptation in *Brachypodium distachyon* complex species. The species as model provides natural genetic mapping resources to dissect complex adaptive traits. Admixed populations also inform about the evolutionary history of *Brachypodium* and provide an opportunity to test natural selection on segregating variation. To study the natural variation of stress tolerance in *Brachypodium distachyon* complex species, it would be good to have further collections from more extreme environments. This would include higher altitude/latitude environments such as Northern Europe and more arid environments in the Middle East and North Africa. Areas with multiple, recent introductions are also of interest

as they may provide examples of strong selection on standing variation as well as insights to climate breadth of lineages as introduced sites are likely not the exact same climate as where the introduced groups are from.

#### NRD-1, The Most Geographically and Climate Diverse Genotype

The investigation of why genotype NRD-1 is so extensive and climate diverse should be dissected to find the causative reasons of its colonisation success. Being that this genotype is so widespread, locations where it was found could be remote surveyed on location to follow and quantify its expressed phenotypes in the field (Brown, 2016). Another more replicable analysis could be using modified climate chambers the Borevitz lab has been developing to recreate the natural environment in a controlled growth facility. Using a few chambers and a diverse set of genotype lines, a study can be implemented to measure fitness across different climate thresholds replicated from real world locations (Borevitz Lab, 2016). Other possibilities would be creating crosses and RILs to find causative loci associated with phenotypes or phenotypic plasticity.

#### The *Brachypodium* Future

Previously, one of the drawbacks of *Brachypodium* research was the lack of collections and low geographic diversity of samples in the public germplasm. Early in the development of *B. distachyon* as a model there were only a handful researchers developing it. Now through our efforts and the efforts of our collaborators, there is ample genetic material for the *Brachypodium distachyon* species complex, but there are many locations that require more attention highlighted above. With diversity of the species more understood, the development of a large set of naturally diverse lines could create an ideal GWAS set to interrogate phenotypes across landscapes or search for causative orthologs between species. The Vogel lab, Ezrati lab, Mur Lab, Hazen Lab, Mockler lab, Garvin lab, Borevitz lab, and Catalan lab are just a few of the many research groups bringing the *Brachypodium* genus into the model species mainstream.

#### Collaborations

This thesis is built upon the collaborative efforts of nine research groups sharing research material. Some of the collections are still private in most regards, but our collaborators are eager to push their *Brachypodium* germplasms into new collaborations. The cumulative efforts of many people make this thesis possible, from technicians, to field collectors, even people now deceased. Without their dedication and assembly of germplasms this project wouldn't be possible. The motivation to their efforts goes beyond this work, but their contributions and accuracy of information provided is substantial.

#### Final Discussion

The region currently known as Palestine/Israel was heavily sampled. Despite its heavy coverage in our data set, the region is very climate diverse and harbours lots of genetic diversity of *B.*

*hybridum*. This is logical because the area is a coastal maritime climate that quickly raises to high elevation in a very seasonal climate, then become very dry and greatly reduces to negative altitude approaching the Dead Sea. This is an ideal location to find climate diverse lineages because within a short distance many climate types can be found and a species living in one region has to travel very little distance to experience and test a new climate type. Though biodiversity is the most common in Mediterranean like climates, areas with rapid changes in elevation or proximity to coastlines will further extend the possibility of finding genetic diversity. The other locations in the native range that had climate diversity are the Western Himalayas and the NE areas of Turkey.

In the non-native range significant genetic diversity is also found in the San Francisco Bay area of the US of North America and the metropolitan area of Adelaide on the Australian continent. Both of these areas served as introduction events and harboured multiple genotypes. Like Israel, these locations likely have diverse climate gradients between coastal areas and quickly raising to mountainous regions and/or deserts. Once a species is introduced to a climate diverse region, there are many climate options in a small geographic space. If an introduced species lands in one of these climate diverse regions they may only need to survive one or two generation in a non-ideal climate until they disperse the short distance to find their ideal location across a climate gradient.

Furthermore, the genetic and climate diversity of *B. hybridum* found in the SE Mediterranean could be the result of having so much climate diversity in a small space. New alleles under selection are rapidly tested at a population's climate boundaries. Having a short dispersal distance to test new alleles against novel climates would likely accelerate the selection of new mutations. Therefore it could be suggested that native areas of a known invasive species that have significant climate diversity probably have the most invasive lines. Given the evidence presented here in this work, and the amount of geographic space that the *B. hybridum* genotype NRD-1 is found in and common across the SE Mediterranean, a landscape analysis of *B. hybridum* in this area could provide some insight to the evolution of invasive species in their native range.

The degree of genetic diversity found in non-native habitats is astonishing. The possible reasons for *B. hybridum* to have travelled so well could be it's anonymous appearance as a weedy grass species. Since it is a grass, it is difficult to identify without flowers, limiting the amount of time available annually to identify it. It is also an annual plant that rarely vegetatively overwinters, so it is physically present as a small anonymous seed for a large part of the year. Also, its overlap in climate breadth and regions it is introduced in overlaps with other agricultural species. Though any literature pertaining to the actual dispersal of *Brachypodium* species to the new



worlds is next to non-existent, the records of its appearance in Australia do overlap with the introduction of humans and agriculture, suggesting a strong correlation. Regardless, the amount of diversity of introduced locations was higher than expected.

The distribution of a species or series of species in new environments is no trivial matter. The movement of species to novel habitats is known to potentially cause damage, but it seems logical that the eradication of an unwanted species is more difficult if multiple introductions do occur and more genetic diversity is introduced to a non-native gene pool. The use of *Brachypodium distachyon* complex species has been fruitful because of the availability of resources made possible by previous projects and collection efforts. The within and between species diversity was easy to map genetic loci, and having samples of both extant diploid partners in the polyploid species was of benefit. The geographic and climate diversity was also helpful in searching for genotypes that span large geographic and climate space. The scanning for genetic, geographic and climate diversity of native and non-native habitats provided a glimpse into the potential diversity of other introduced species.

## 6.6 Citation

---

ALA (2016). *Brachypodium distachyon* search. ala.org. Spatial Portal. Atlas of Living Australia. <http://bie.ala.org.au/search?q=brachypodium+distachyon>

Bakker, E. G., Montgomery, B., Nguyen, T., Eide, K., Chang, J., Mockler, T. C., ... & Borer, E. T. (2009). Strong population structure characterizes weediness gene evolution in the invasive grass species *Brachypodium distachyon*. *Molecular ecology*, *18*(12), 2588-2601.

Banta, J. A., Ehrenreich, I. M., Gerard, S., Chou, L., Wilczek, A., Schmitt, J., ... & Purugganan, M. D. (2012). Climate envelope modelling reveals intraspecific relationships among flowering phenology, niche breadth and potential range size in *Arabidopsis thaliana*. *Ecology letters*, *15*(8), 769-777.

Beck, J. B., Schmutz, H., & Schaal, B. A. (2008). Native range genetic variation in *Arabidopsis thaliana* is strongly geographically structured and reflects Pleistocene glacial dynamics. *Molecular Ecology*, *17*(3), 902-915.

Borevitz Lab. (2016). TraitCapture. Borevitz Lab, Plant genomics for Climate Adaptation. <http://borevitzlab.anu.edu.au/spectralphenoclimatron/>

Brown, T. B., Hultine, K. R., Steltzer, H., Denny, E. G., Denslow, M. W., Granados, J., ... & Sánchez-Azofeifa, A. (2016). Using phenocams to monitor our changing Earth: toward a global phenocam network. *Frontiers in Ecology and the Environment*, *14*(2), 84-93.

Carnaval, A. C., Waltari, E., Rodrigues, M. T., Rosauer, D., VanDerWal, J., Damasceno, R., ... & Pie, M. R. (2014). Prediction of phylogeographic endemism in an environmentally complex biome. In *Proc. R. Soc. B* (Vol. 281, No. 1792, p. 20141461). The Royal Society.

- Catalán, P., Müller, J., Hasterok, R., Jenkins, G., Mur, L. A., Langdon, T., ... & López-Alvarez, D. (2012). Evolution and taxonomic split of the model grass *Brachypodium distachyon*. *Annals of Botany*, *109*(2), 385-405.
- Catalan, P., López-Álvarez, D., Díaz-Pérez, A., Sancho, R., & López-Herránz, M. L. (2015). Phylogeny and evolution of the genus *Brachypodium*. In *Genetics and genomics of Brachypodium* (pp. 9-38). Springer International Publishing.
- Catalán, P., López-Álvarez, D., Bellosta, C., & Villar, L. (2016, March). Updated taxonomic descriptions, iconography, and habitat preferences of *Brachypodium distachyon*, *B. stacei*, and *B. hybridum* (Poaceae). In *Anales del Jardín Botánico de Madrid* (Vol. 73, No. 1, p. e028).
- Dell'Acqua, M., Zuccolo, A., Tuna, M., Gianfranceschi, L., & Pè, M. E. (2014). Targeting environmental adaptation in the monocot model *Brachypodium distachyon*: a multi-faceted approach. *BMC genomics*, *15*(1), 1.
- Draper, J., Mur, L. A., Jenkins, G., Ghosh-Biswas, G. C., Bablak, P., Hasterok, R., & Routledge, A. P. (2001). *Brachypodium distachyon*. A new model system for functional genomics in grasses. *Plant physiology*, *127*(4), 1539-1555.
- Drew, J., Anderson, N., & Andow, D. (2010). Conundrums of a complex vector for invasive species control: a detailed examination of the horticultural industry. *Biological Invasions*, *12*(8), 2837-2851.
- Eichten, S. R., Stuart, T., Srivastava, A., Lister, R., & Borevitz, J. O. (2016). DNA Methylation profiles of diverse *Brachypodium distachyon* aligns with underlying genetic diversity. *bioRxiv*, 039602.
- Eichten, S. R., Briskine, R., Song, J., Li, Q., Swanson-Wagner, R., Hermanson, P. J., ... & Myers, C. L. (2013). Epigenetic and genetic influences on DNA methylation variation in maize populations. *The Plant Cell*, *25*(8), 2783-2797.
- Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, *17*(1), 43-57.
- Filiz, E., Ozdemir, B. S., Budak, F., Vogel, J. P., Tuna, M., & Budak, H. (2009). Molecular, morphological, and cytological analysis of diverse *Brachypodium distachyon* inbred lines. *Genome*, *52*(10), 876-890.
- Fraley, C., & Raftery, A. E. (2006). *MCLUST version 3: an R package for normal mixture modeling and model-based clustering*. WASHINGTON UNIV SEATTLE DEPT OF STATISTICS.
- GBIF.org (2016). Global Biodiversity Information Facility, *Brachypodium distachyon* search results. [http://www.gbif.org/occurrence/search?taxon\\_key=5290143&HAS\\_COORDINATE=true&HAS\\_GEOSPATIAL\\_ISSUE=false&display=map&COUNTRY.offset=10](http://www.gbif.org/occurrence/search?taxon_key=5290143&HAS_COORDINATE=true&HAS_GEOSPATIAL_ISSUE=false&display=map&COUNTRY.offset=10)
- Genton, B. J., Shykoff, J. A., & Giraud, T. (2005). High genetic diversity in French invasive populations of common ragweed, *Ambrosia artemisiifolia*, as a result of multiple sources of introduction. *Molecular ecology*, *14*(14), 4275-4285.
- Hammami, R., Jouve, N., Soler, C., Frieiro, E., & González, J. M. (2014). Genetic diversity of SSR and ISSR markers in wild populations of *Brachypodium distachyon* and its close relatives *B. stacei* and *B. hybridum* (Poaceae). *Plant Systematics and Evolution*, *300*(9), 2029-2040.

- Huang, P., Feldman, M., Schroder, S., Bahri, B. A., Diao, X., Zhi, H., ... & Kellogg, E. A. (2014). Population genetics of *Setaria viridis*, a new model system. *Molecular ecology*, *23*(20), 4912-4925.
- Kesari, R., Lasky, J. R., Villamor, J. G., Des Marais, D. L., Chen, Y. J. C., Liu, T. W., ... & Verslues, P. E. (2012). Intron-mediated alternative splicing of *Arabidopsis* P5CS1 and its association with natural variation in proline and climate adaptation. *Proceedings of the National Academy of Sciences*, *109*(23), 9197-9202.
- Konarzewski, T. K. (2012). *Clinal variation in life-history traits of the invasive plant species Echinochloa polystachyon L* (Doctoral dissertation).
- Lee, C. R., & Mitchell-Olds, T. (2012). Environmental adaptation contributes to gene polymorphism across the *Arabidopsis thaliana* genome. *Molecular biology and evolution*, *29*(12), 3721-3728.
- López-Alvarez, D., Manzaneda, A. J., Rey, P. J., Giraldo, P., Benavente, E., Allainguillaume, J., ... & Ezrati, S. (2015). Environmental niche variation and evolutionary diversification of the *Brachypodium distachyon* grass complex species in their native circum-Mediterranean range. *American journal of botany*, *102*(7), 1073-1088.
- Mitchell-Olds, T. (2001). *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. *Trends in Ecology & Evolution*, *16*(12), 693-700.
- Mur, L. A., Allainguillaume, J., Catalán, P., Hasterok, R., Jenkins, G., Lesniewska, K., ... & Vogel, J. (2011). Exploiting the *Brachypodium* Tool Box in cereal and grass research. *New Phytologist*, *191*(2), 334-347.
- Neji, M., Geuna, F., Taamalli, W., Ibrahim, Y., Chiozzotto, R., Abdelly, C., & Gandour, M. (2015). Assessment of genetic diversity and population structure of Tunisian populations of *Brachypodium hybridum* by SSR markers. *Flora-Morphology, Distribution, Functional Ecology of Plants*, *216*, 42-49.
- Niu, B., Zhu, Z., Fu, L., Wu, S., & Li, W. (2011). FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics*, *27*(12), 1704-1705.
- Opanowicz, M., Vain, P., Draper, J., Parker, D., & Doonan, J. H. (2008). *Brachypodium distachyon*: making hay with a wild grass. *Trends in plant science*, *13*(4), 172-177.
- Phillips, S. J. (2005). A brief tutorial on MaxEnt. *AT&T Research*.
- Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161-175.
- Reichard, S. H., & White, P. (2001). Horticulture as a Pathway of Invasive Plant Introductions in the United States Most invasive plants have been introduced for horticultural use by nurseries, botanical gardens, and individuals. *BioScience*, *51*(2), 103-113.
- Rollins, L. A., Moles, A. T., Lam, S., Buitenwerf, R., Buswell, J. M., Brandenburger, C. R., ... & Thomson, F. J. (2013). High genetic diversity is not essential for successful introduction. *Ecology and evolution*, *3*(13), 4501-4517.
- Rosenthal, D. M., Ramakrishnan, A. P., & Cruzan, M. B. (2008). Evidence for multiple sources of invasion and intraspecific hybridization in *Brachypodium sylvaticum* (Hudson) Beauv. in North America. *Molecular ecology*, *17*(21), 4657-4669.

Smith, C. R., Anderson, K. E., Tillberg, C. V., Gadau, J., & Suarez, A. V. (2008). Caste determination in a polymorphic social insect: nutritional, social, and genetic factors. *The American Naturalist*, 172(4), 497-507.

Tyler, L., Lee, S. J., Young, N. D., DeLulio, G. A., Benavente, E., Reagon, M., ... & Caicedo, A. L. (2016). Population structure in the model grass *Brachypodium distachyon* is highly correlated with flowering differences across broad geographic areas.

Vogel, J. P., Garvin, D. F., Leong, O. M., & Hayden, D. M. (2006). *Agrobacterium*-mediated transformation and inbred line development in the model grass *Brachypodium distachyon*. *Plant Cell, Tissue and Organ Culture*, 84(2), 199-211.

## Appendix

---

### Appendix Outline

Glossary  
List of Abbreviations  
Chapter II Supplementary Materials  
Chapter III Supplementary Materials  
Chapter IV Supplementary Materials  
Chapter V Supplementary Materials

### **Glossary**

---

**Climate:** The abiotic factors contributing to the overall environment type, usually precipitation, temperature, sometimes elevation and solar radiation.

**Climate type:** The climate cluster designation given by clustering locations by their climate variables using Mclust to find cluster centers.

**Environment:** The surroundings quantifiable conditions affecting an individual in an specific location.

**Genotype:** A unique set of genetic variants occurring within one or more individual's genome that indicate they share the same genetic material within the bounds of statistical error to call variants.

**Phenotype:** A quantifiable expressed trait that can be measured.

**BioClim 1:** Annual Mean Temperature: The combined average of all months of the year

**BioClim 2:** Mean Diurnal Range: The mean range between the average monthly maximum minimum average minimum temperature.

**BioClim 3:** Isothermality: Isothermality is the mean diurnal range divided by the temperature average range. Some call it the "evenness of temperature" across a year. Another way to think of it is the day:night temperature oscillate compared to the summer:winter oscillations.

**BioClim 4:** Temperature Seasonality: The degree of temperature variation annually quantified by standard deviation of monthly average temperature as compared to the mean annual temperature.

**BioClim 5:** Maximum Temperature of Warmest Month: The monthly max temperature over a yearly period.

**BioClim 6:** Minimum Temperature of Coldest Month: The minimum temperature of a month across one a year period.

**BioClim 7:** Temperature Annual Range: The amount of temperature range between maximum warmest monthly temperature and minimum coldest monthly temperature across a one-year period.

**BioClim 8:** Mean Temperature of Wettest Quarter: A measure of an annual period's wettest seasonal temperature.

**BioClim 9:** Mean Temperature of Driest Quarter: The average temperature across a yearly period's driest season.

**BioClim 10:** Mean Temperature Warmest Quarter: An approximate mean of temperatures within the warmest season of a year.

**BioClim 11:** Mean Temperature of Coldest Quarter: The mean temperature of the coldest season of a year.

**BioClim 12:** Annual Precipitation: The total of all monthly precipitation measurements across one year.

**BioClim 13:** Precipitation of Wettest Month: The total precipitation occurring in the wettest month of a year.

**BioClim 14:** Precipitation of Driest Month: The total precipitation of the driest month across a year period.

**BioClim 15:** Precipitation Seasonality: This is the variation in monthly precipitation totals over a year. A ratio of the standard deviation of monthly total precipitation and the mean monthly precipitation expressed as a percentage. Sometimes described as the coefficient of variation in precipitation.

**BioClim 16:** Precipitation of Wettest Quarter: The measure of total precipitation of the wettest season per a year interval.

**BioClim 17:** Precipitation of Driest Quarter: A seasonal index of approximate precipitation that prevails at the driest season of a year

**BioClim 18:** Precipitation of Warmest Quarter: The total precipitation of a season with the warmest mean temperature of a year.

**BioClim 19:** Precipitation of Coldest Quarter: The total precipitation of a season with the coldest mean temperature per a year period.

**SNP:** Single nucleotide polymorphism. A mutational change of base at a specific loci.

**Variant:** A genetic polymorphism at a specific loci, that might possibly be orthologous to a loci shared between one plus genomes of multiple species.

**Climate Type:** A distinct partition of a climate gradient that collection locations would be categorised under.

**Collection Location:** An area sampled for specimens. Distinct sizes for location area are specific to each collaborative research group.

**Genotype:** A genetically distinct individual comprised of identical to near identical single nucleotide polymorphisms that falls within the resolution of the employed genotyping platform.

**Potential Area:** The cumulative area in square kilometers derived from MaxEnt modelling. The minimum presence probability threshold used was set at where training specificity equals sensitivity. All area calculation in this work was corrected for earth curvature.

**Sampling Region:** A cluster of collection locations that fall within specified confidence intervals of pairwise geographic distance from each other compared to other locations.

### List of Abbreviations

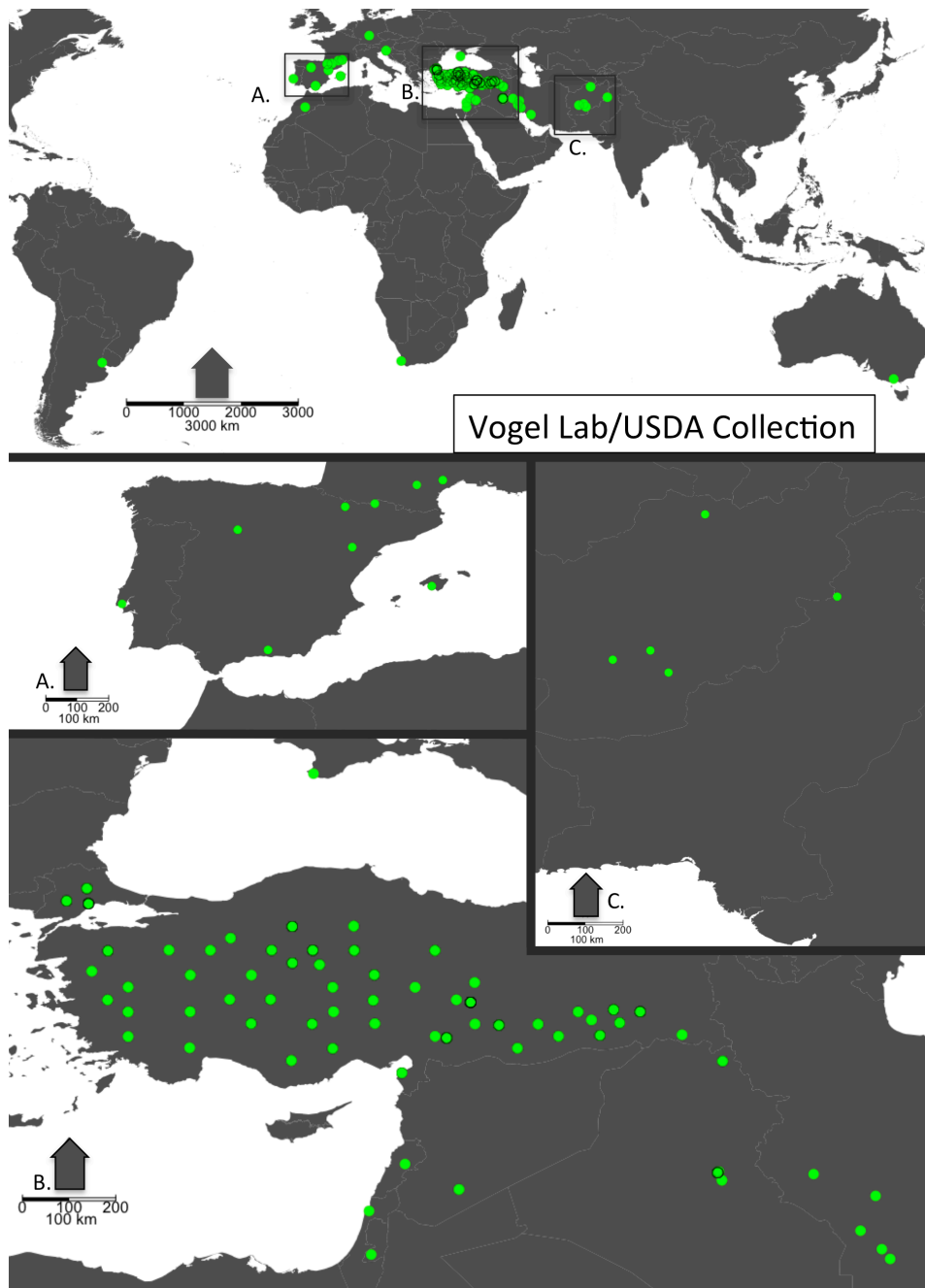
---

GBS: Genotyping By Sequencing

GIS: Global Information System

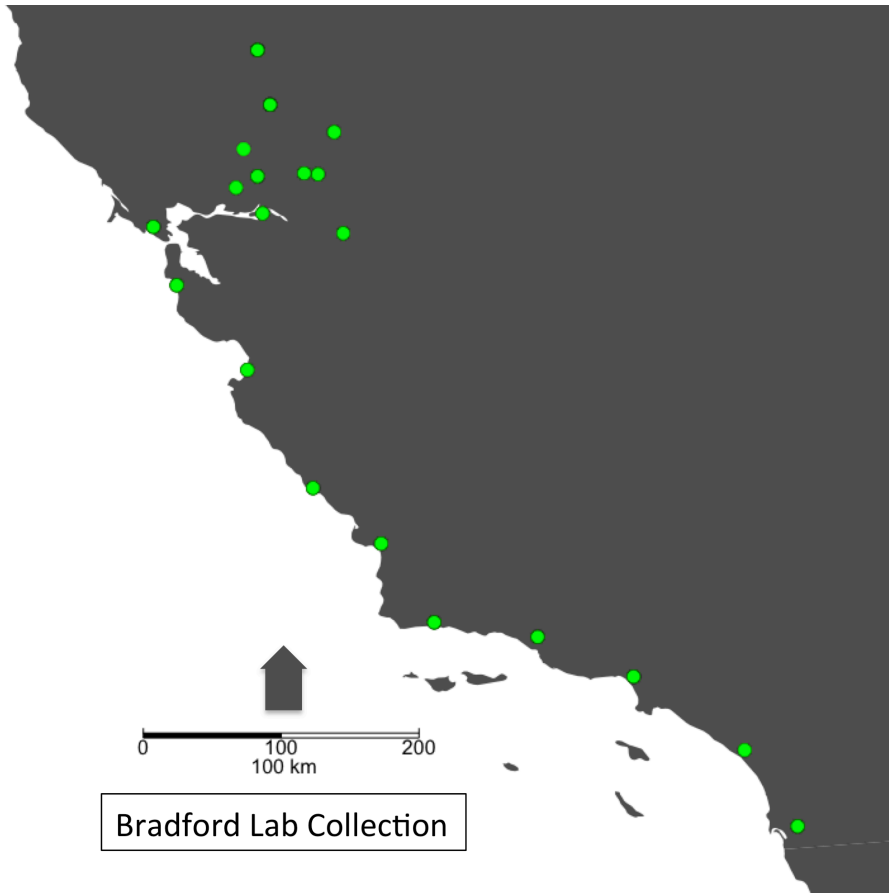
SDM: Species Distribution Model

SNP: Single Nucleotide Polymorphism

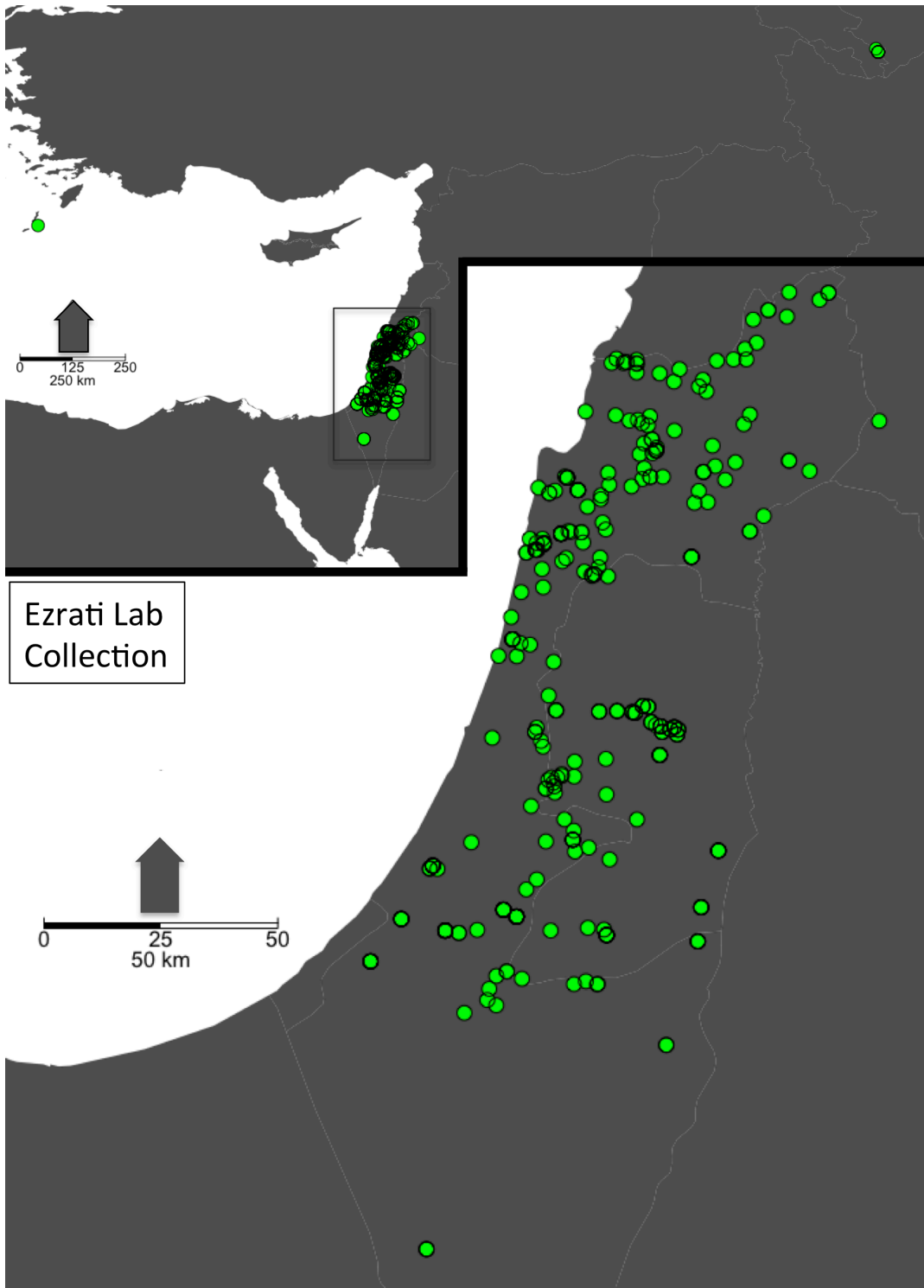


S2.1 Vogel lab collection sites. Global locations across the Mediterranean, Europe, West Asia, Middle East, Iberia, Africa, South America, and Australia

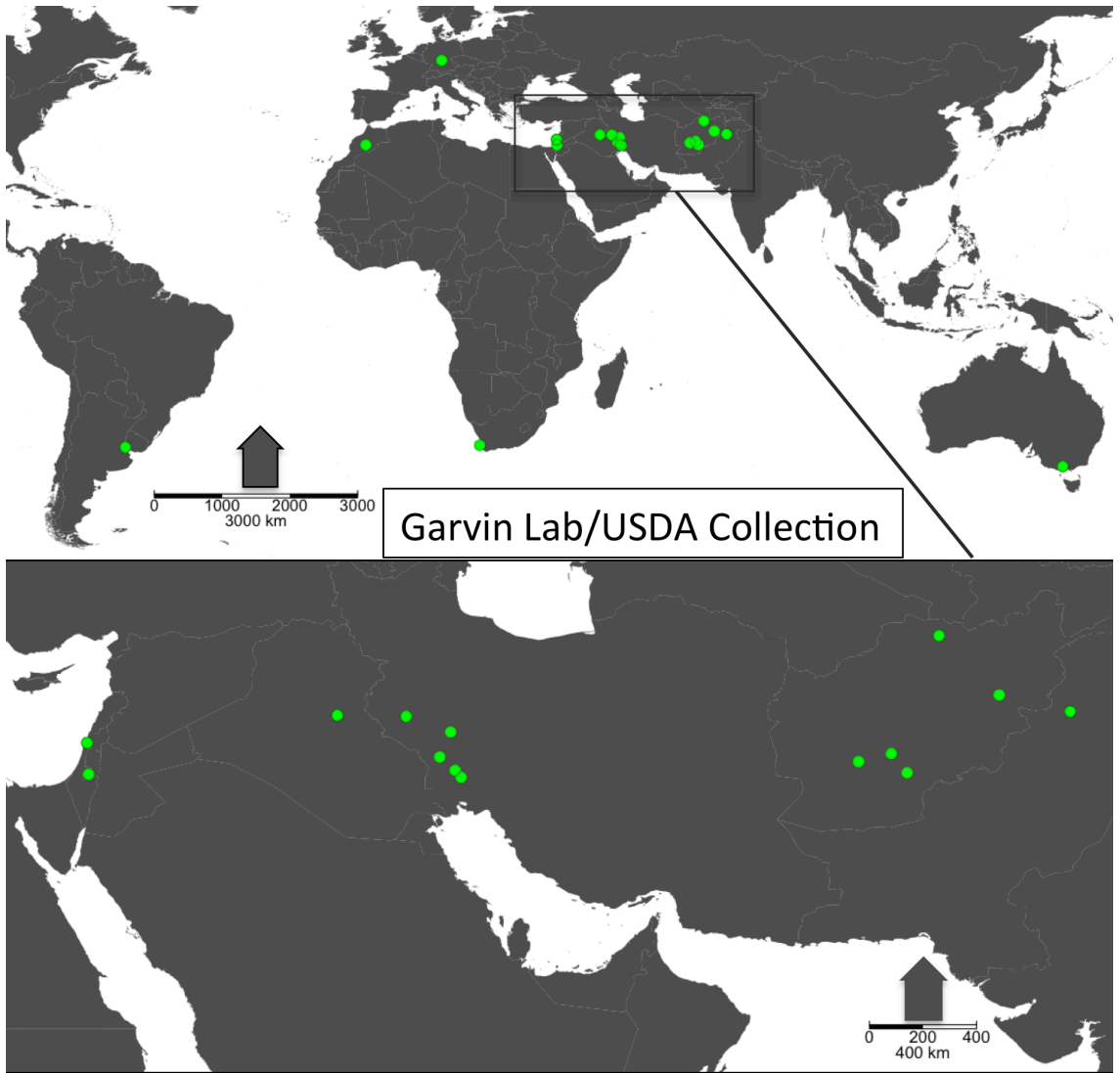




S2.2 Bradford lab collection, locations across North America in the United States, California

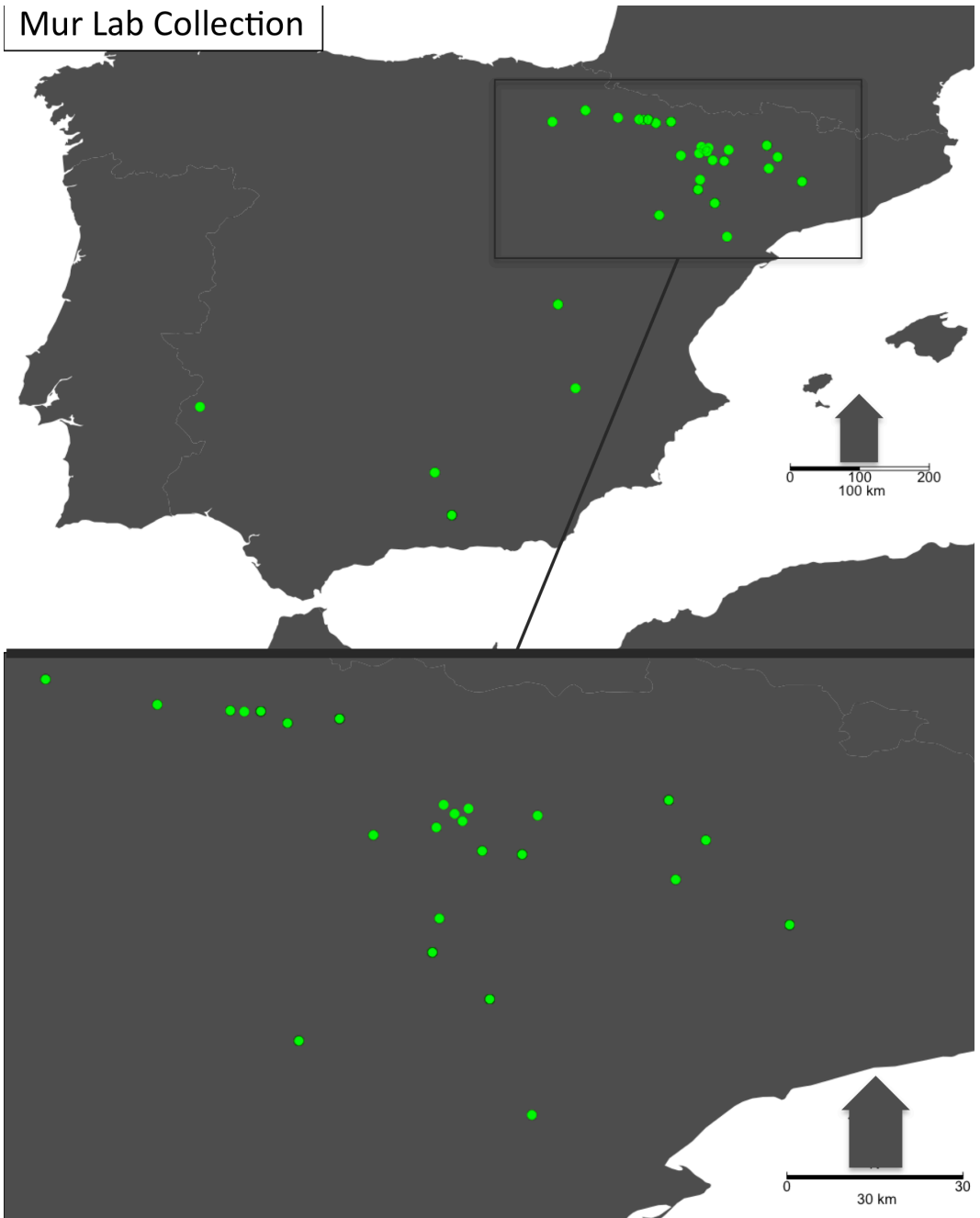


S2.3. Ezrati lab Collection. Collection sites across modern day Israel/Palestine, Greece, and Armenia

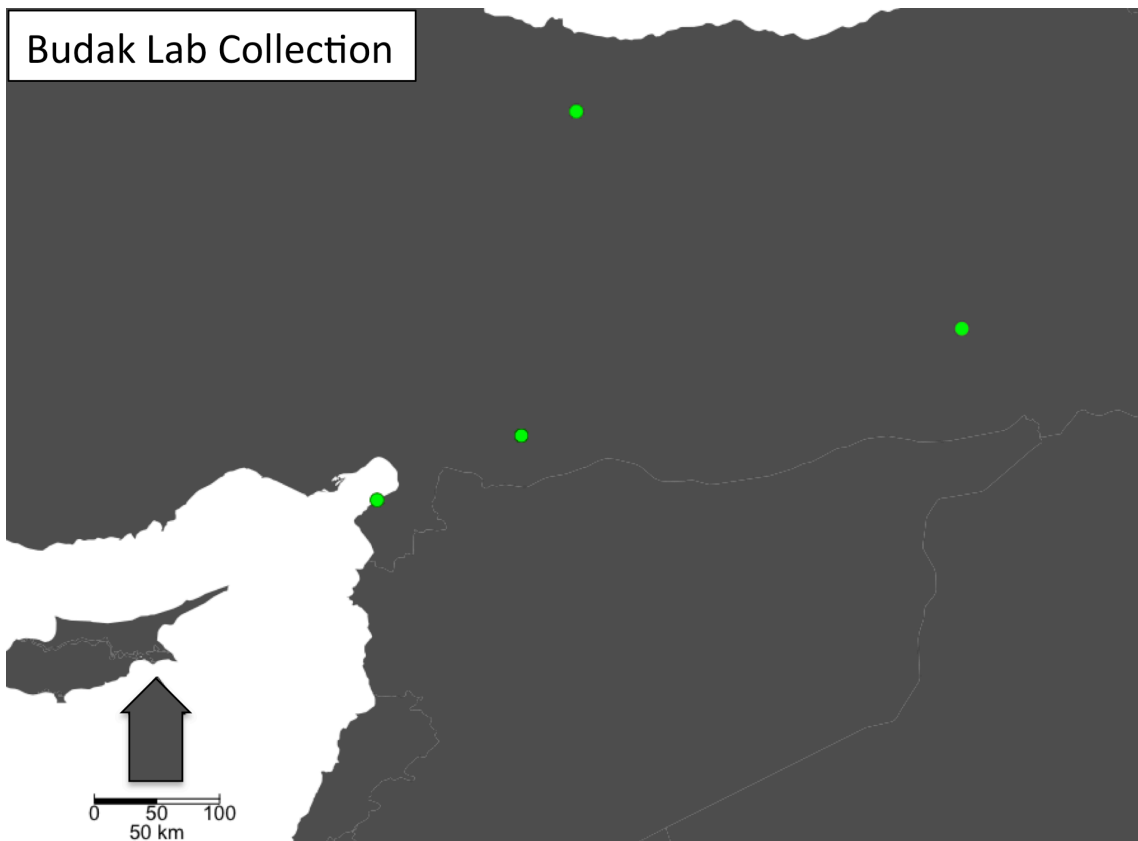


S2.4 Garvin Lab Collection/USDA Bulk Collections. Collection Sites across West Asia, Modern day Germany, Africa, South America, and Australia.

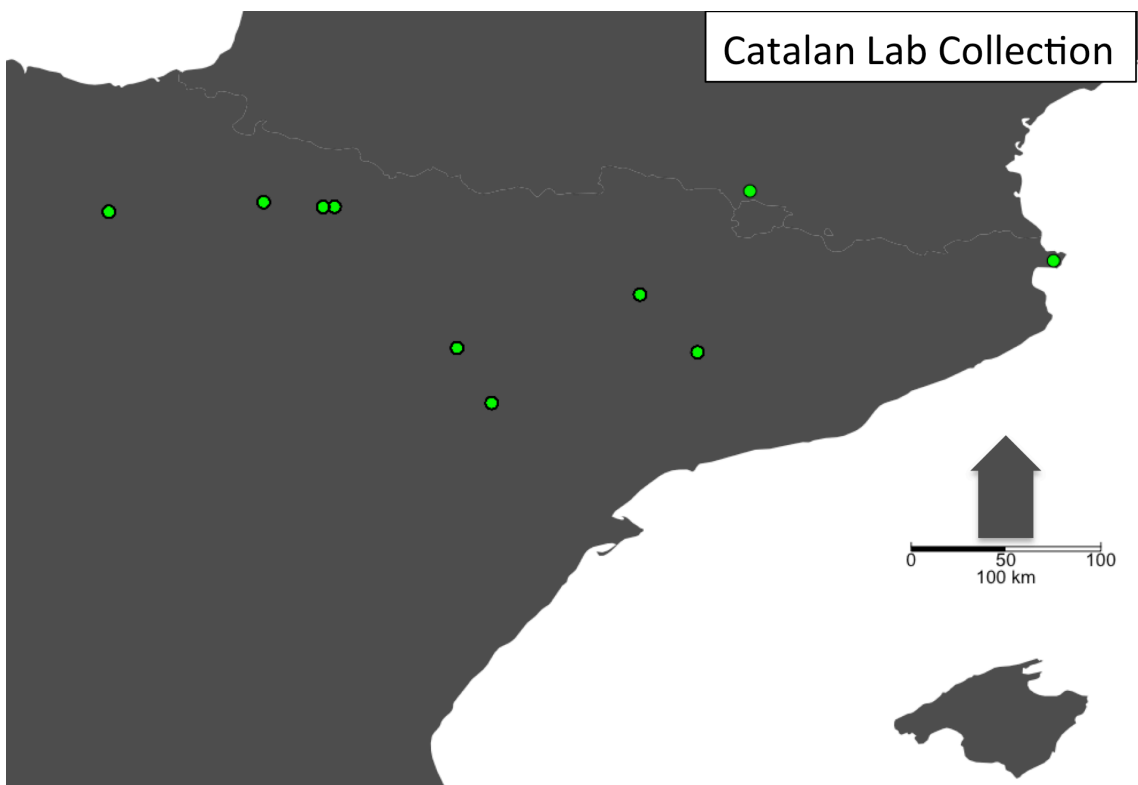
Mur Lab Collection



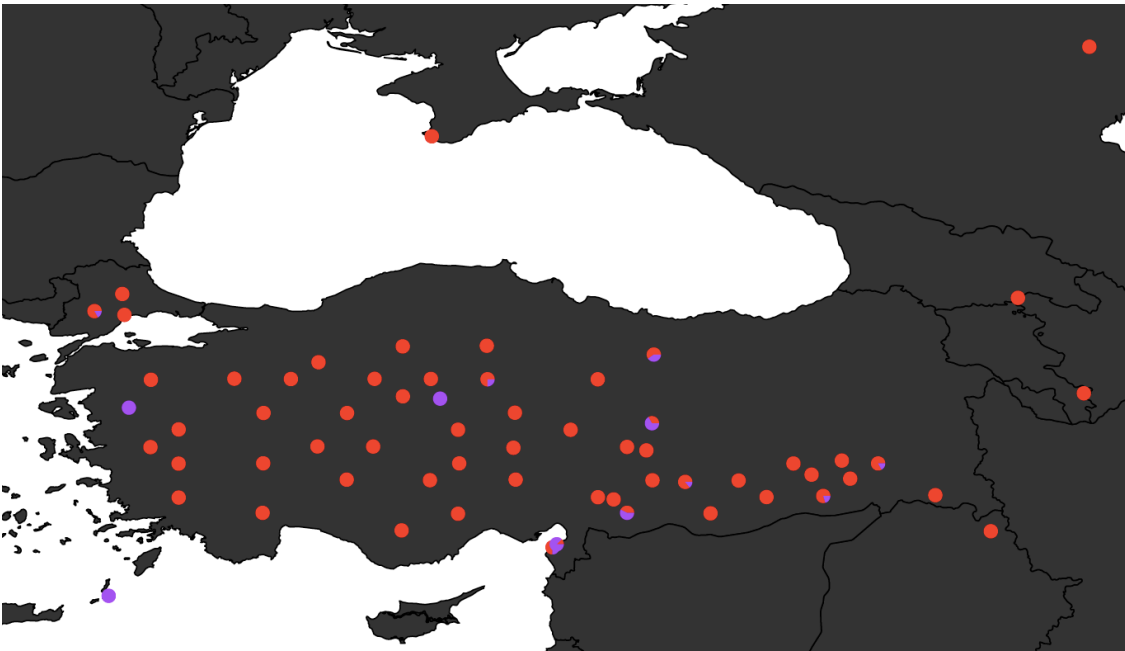
S2.5 Mur Lab Collection. Collection Sites from across Modern day Spain on the Iberian Peninsula.



S2.6 Budak lab. Collection Locations across Eastern modern day Turkey



S2.7 Catalan Lab Collection. Locations across North-eastern Iberian Peninsula.



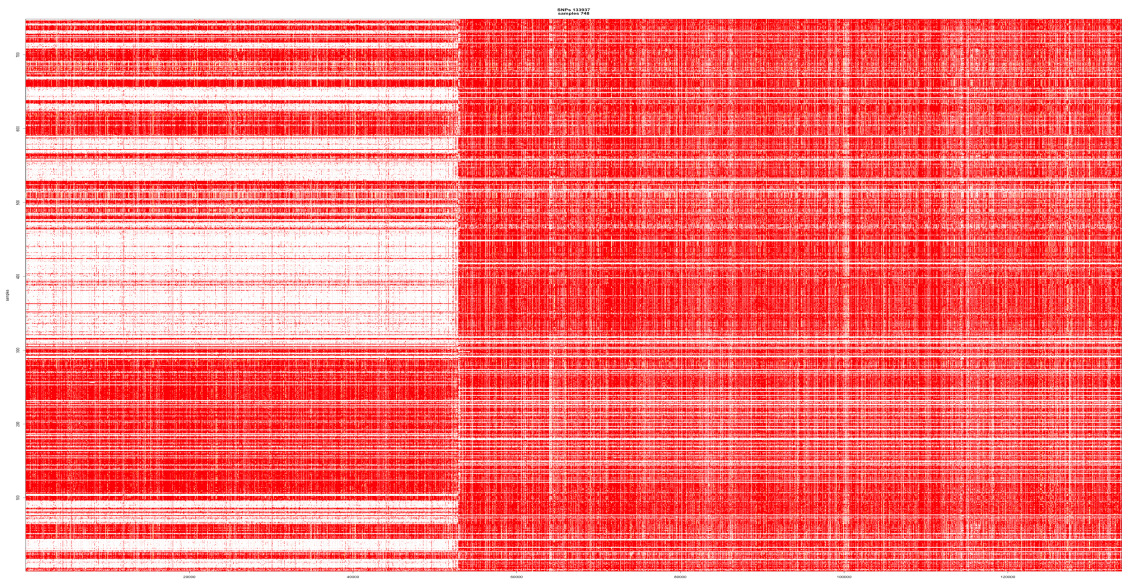
S2.8 Pie charts on maps of species identification across the NE Mediterranean.  
 Red = *B. distacyon*, Purple = *B. hybridum*



S2.9 Pie charts on maps of species identification across the SE Australian Continent.  
 Red = *B. distacyon*, Purple = *B. hybridum*



**Figure S2.10** Pie charts on maps of species identification across the southern Iberian Peninsula. Red = *B. distachyon*, Purple = *B. hybridum*



**Figure S2.11.** Raw Data Matrix of *B. distachyon* Sequence Data, Stacei genome, then Distachyon, X is variants, Y is samples. White is absence of data, red is presence of an allele.



**Figure S3.1.** Raw Data Matrix of *B. distachyon* Sequence Data, Stacei genome, then Distachyon, X is variants, Y is samples:

Clustering and Relatedness of *B. hybridum* and sub-genomes:

	D Subgenome	S Subgenome	Cumulative
<b>Samples</b>	1,201	1,201	1,201
<b>Raw Marker Count</b>	54,223	43,726	97,949
<b>Valid Samples with Marker Count &gt; 21,803</b>	1,015	1,015	1,015
<b>Valid Markers called in &gt; n Samples</b>	545	531	1,076
<b>Average Markers Per Sample</b>	30,338	26,258	56,596
<b>Minimum Variants</b>	11,803	10,000	21,803
<b>High Quality Variants</b>	9,273	9,252	18,525
<b>Genotypes</b>	---	---	80

**Table S3.2** *B. hybridum* Genome and Subgenomes Analysis Properties



Figure S3.3. Dendrogram of the D subgenome of *B. hybridum*

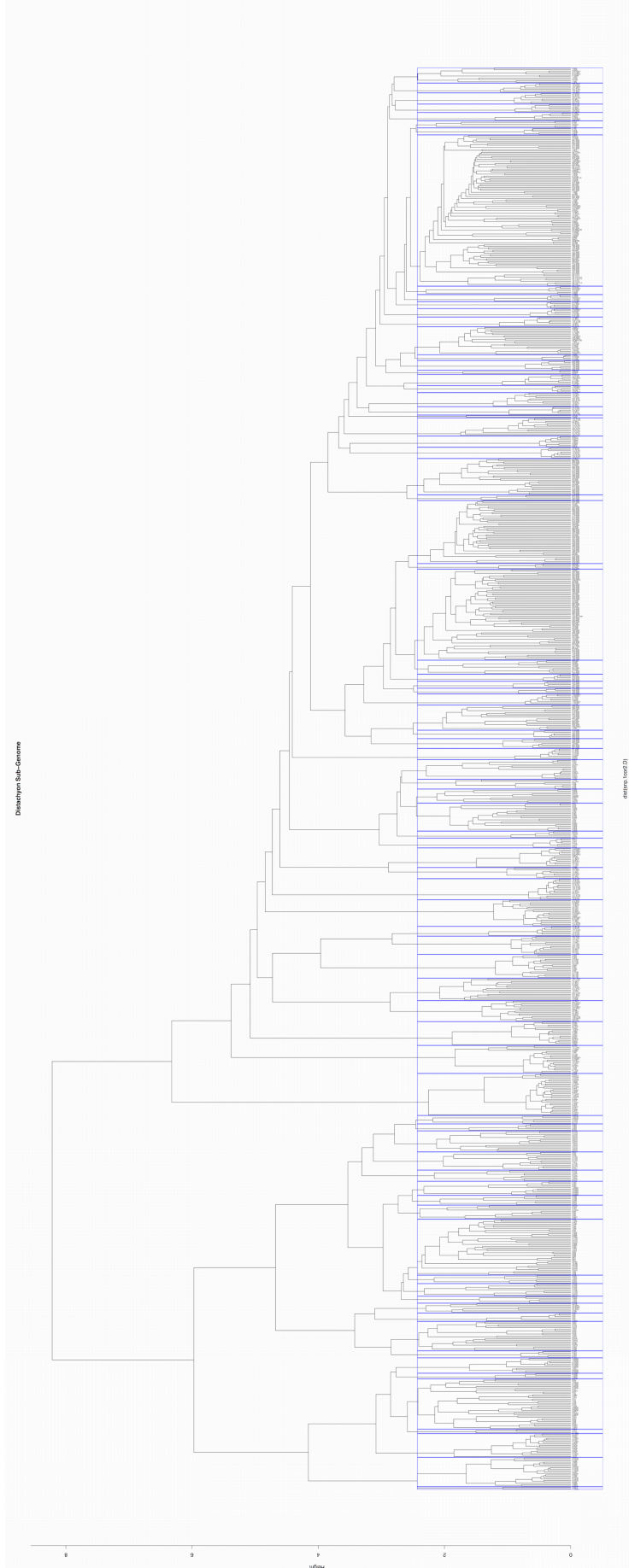


Figure S3.4. Dendrogram of the S subgenome of *B. hybridum*

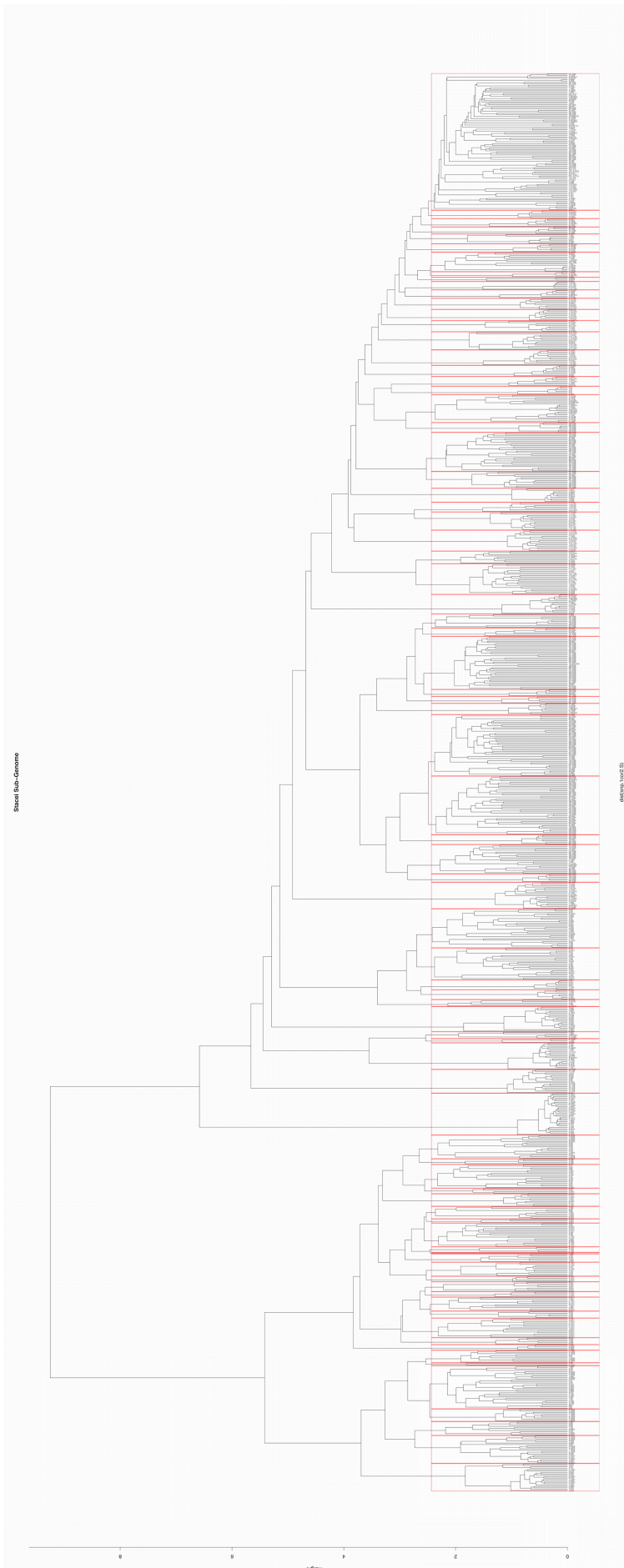


Figure S3.5, Final Genotype Names and Genotyping Cluster Identities for *B. distachyon*:

Cluster Number	Genotype	Cluster Number	Genotype	Cluster Number	Genotype
1	BD21	43	BRA.422	86	Bdis22.16
2	10Geo.G30i2	44	Ban2	88	Bdis22.18
4	BdTR13	45	Ban3	89	Bdis23.5
5	ABR2	46	Bd3.1	90	Bdis22.1
6	ABR3	47	BdTR10	91	Bdis22.3
7	ABR4	48	BdTR9	92	Bdis22.4
8	ABR5	49	BdTR10j	93	Bdis22.5
9	ABR7	50	BdTR11f	94	Bdis23.4
10	ABR9	51	BdTR12	95	Bdis23.2
11	ARN1018	52	BdTR13d	96	Bdis25.2
12	Arc1	53	BdTR13h	97	Bdis28.1
13	Foz1	54	BdTR13	98	Bdis25.6
14	Abz1021.2	55	BdTR1	99	Bdis25.4
15	Adi.10	56	BdTR1d	100	Bdis25.5
16	Adi.11	57	BdTR2e	101	Bdis31.2
17	BdTR11	58	BdTR2i	102	MUR2.31
18	Adi-14	59	BdTR3	103	Mon3
19	Adi.15	60	BdTR3i	104	Bdis31.9
20	Adi.16	61	BdTR3t	105	Sig7
21	Adi.17	62	BdTR3o	106	Foz15
22	Adi.1	63	BdTR5	107	Gaz.3
23	Adi.2	64	BdTR5o	108	G1DX-2
24	Adi.3	65	BdTR7	109	Luc3
25	Adi.6	66	BdTR8	110	Gaz.2
26	Adi.7	67	BdTR9c	111	Gaz.8
27	Adi.8	68	BdTR9e	112	Gaz1
28	Adi.9	69	BdTR9h	113	Kah.1
29	Ald2.2	70	BdTR9i	114	Kah.2
30	Ald2.5	71	Bel15	115	Kah.3
31	Arn2	72	Bdis03.5	117	Kah.6
32	Cel1	73	Bdis03.4	118	Koz.3
33	Ban1	74	Bdis03.6	119	Luc1
34	BD2.15	75	Bdis03.9	120	Mig1
35	Ban1	76	Bdis05.7	121	Mon17
36	BD21.3	77	Bdis05.2	122	Mon1
37	BD30.1	78	Bdis05.13	123	Mon4
38	BRA.178	79	Bdis05.1	124	PYR2.6
39	BRA.419	80	Bdis05.15	125	Per2
40	BRA.81	82	Bdis05.19	126	Tek.8
41	BRA.88	83	Bdis05.3	127	Tek.4
41	BRA.88	84	Bdis22.10	128	Tek.1
42	BRA.308	85	Bdis22.14	129	Uni8

Figure S3.5, Final Genotype Names and Genotyping Cluster Identities for *B. distachyon*:

A list of final genotype assignment of *B. distachyon* from 479 samples condensed to 129 unique samples, and 125 genotypes, some of which are crosses and were removed from this list. Note: This list is only as accurate as this study, which used genotyping by sequencing, which is a reduced representation approach to calculating relatedness.

Figure S3.6, Final Genotype Names and Genotyping Cluster Identities for *B. hybridum*:

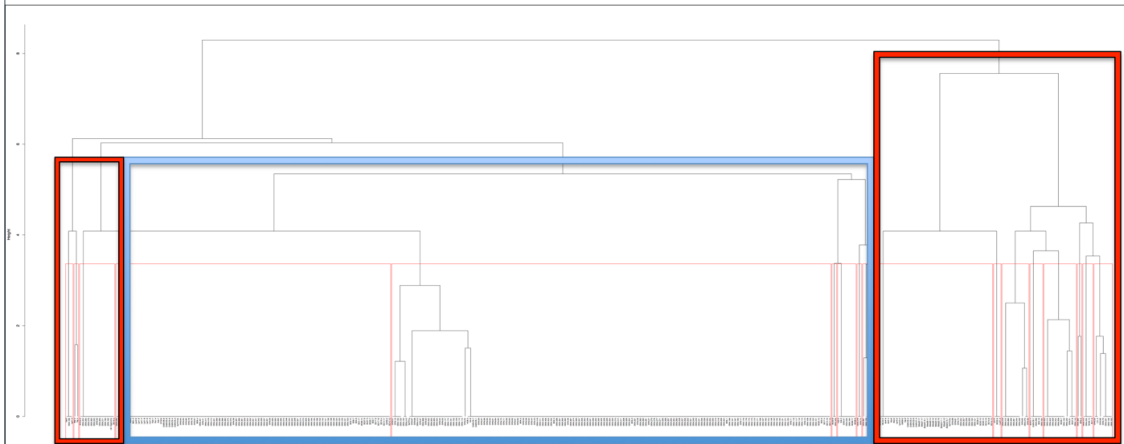
Cluster Number	Genotype	Cluster Number	Genotype
1	NRD.1	41	BRA.207
2	BD19.3	42	BRA.99
3	BRA.100	43	BRA.241
4	Bd5.1	44	BRA.200
5	Tlo.3	45	BRA.351
6	Pdr.1	46	FNR.4
7	BYO1.5	47	BYO2.6
8	Ben.2	48	Isk.P2
9	BRT.1	49	Bd20.1
10	Aqe.1	50	Ca17.1.1
11	Leo4.4	51	Ca16.1.1
12	Fac.2	52	Ca2.11.1
13	Fls1.1	53	Ca20.4.2
14	Bai.3	54	Ca4.4.1
15	GSN.2	55	Ca8.1.2
16	Bar5	56	Ca9.1.1
17	Bd15.1	57	Ca13.2.1
18	Bd14.1	58	Ca19.1.2
19	BD12.6	59	Ca18.2.2
20	BD27.9	60	Ca7.5.2
21	Ca1.2.1	61	Pob1
22	Ca6.1.1	62	Lpn2.3
23	Bd17.1	63	DLS.2
24	Ca5.1.2	64	Mde.4
25	Bd4.1	65	Ele.2
26	SPA.2	66	Lps.1
27	BD17.4	67	Fue.2
28	BD7.11	68	Vil.5
29	HLS.6	69	GND.8
30	HOB.8	70	NAR.2
31	BRA.278	71	WAG.8
32	BRA.71	72	Isk.P7
33	KJP.2	73	QUO.1
34	ELI.1	74	LCL.1
35	RWR.10	75	ORA.1
36	HOB.3	76	Leo3.1
37	CABVZ6.2	77	MAG.3
38	BRA.39	78	SWN.1
39	BRA.25	79	SHP.1
40	BRA.356	80	Roj.8

**Figure S3.6.** Final Genotype Names and Genotyping Cluster Identities for *B. hybridum*:

A list of final genotype assignment of *B. hybridum* from 1,015 samples condensed to 80 unique genotypes, Note: This list is only as accurate as this study which used genotyping by sequencing, which is a reduced representation approach to calculating relatedness.

Alignment of 1,573 *Brachypodium* samples against the *Triticum aestivum* mtGenome to detect multiple hybridisation events using 26 markers.

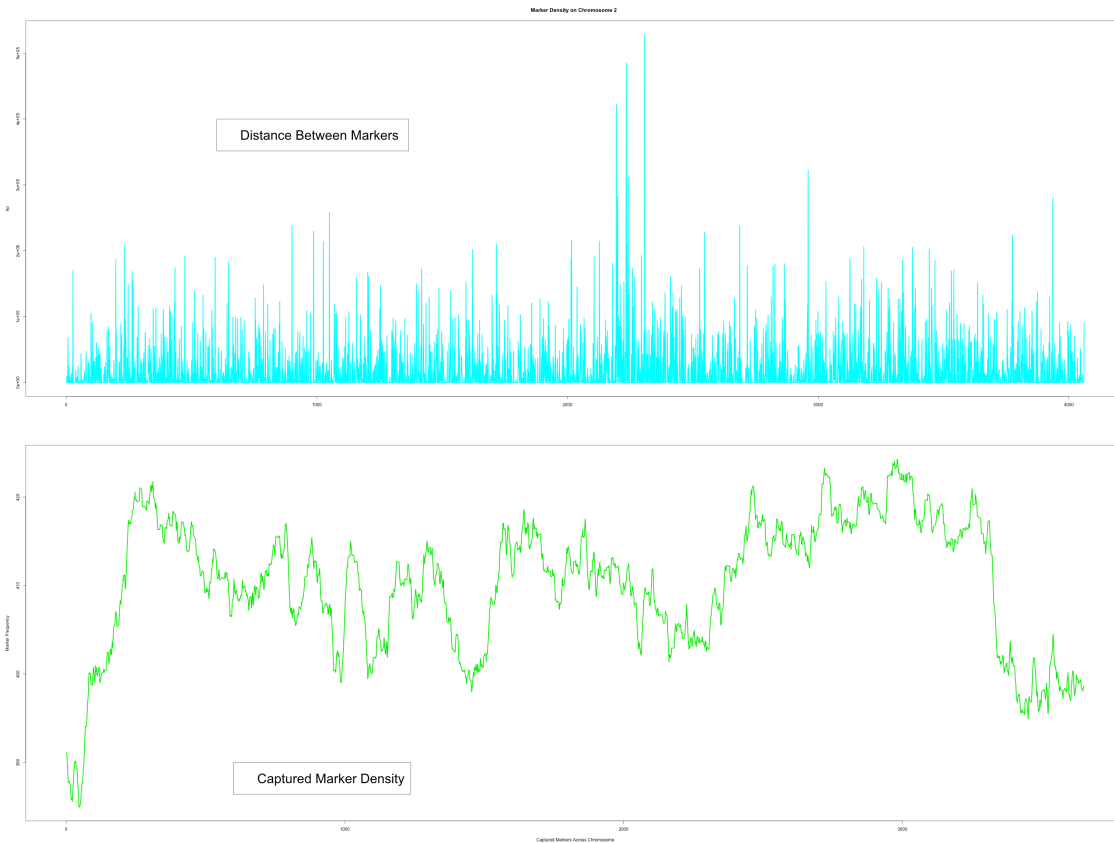
*B. hybridum* and *B. stacei*  
 *B. hybridum* and *B. distachyon*



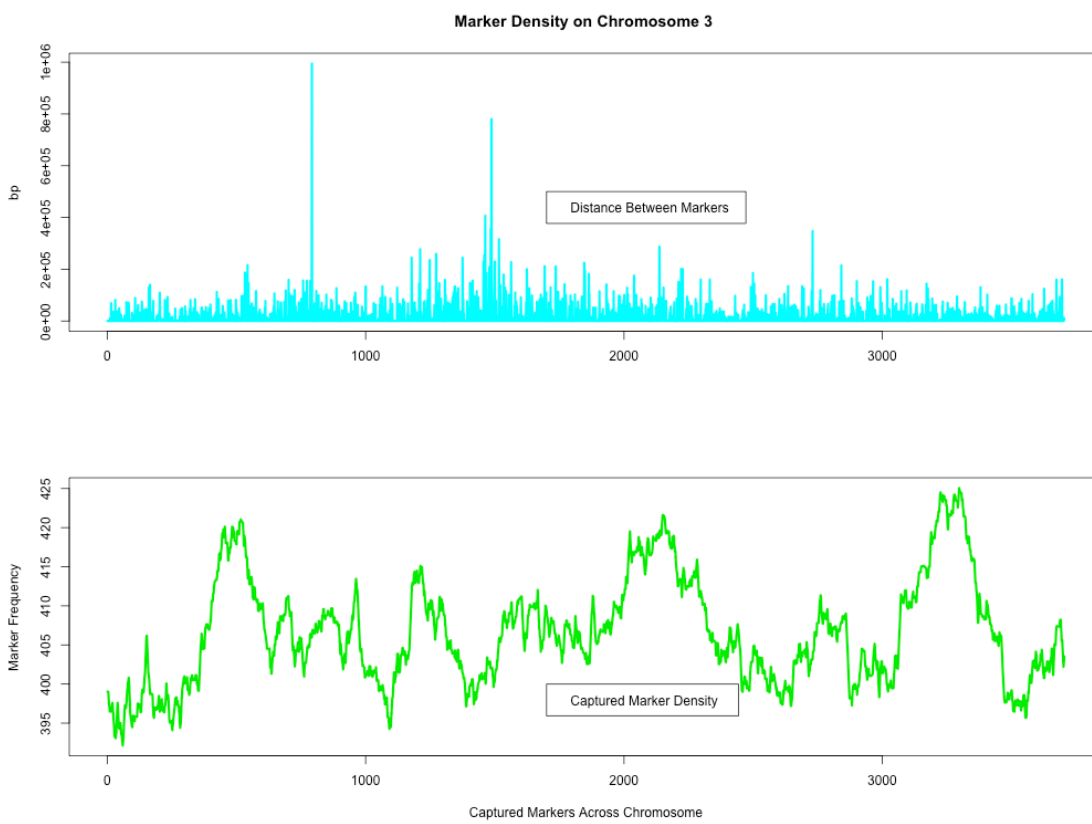
**Figure S3.7.** Dendrogram of 1,573 species tested samples with valid markers above each species thresholds for genotyping. 26 Variants called against the *Triticum aestivum* mitochondrial genome reveals likely patterns of maternal species and multiple hybridisations.



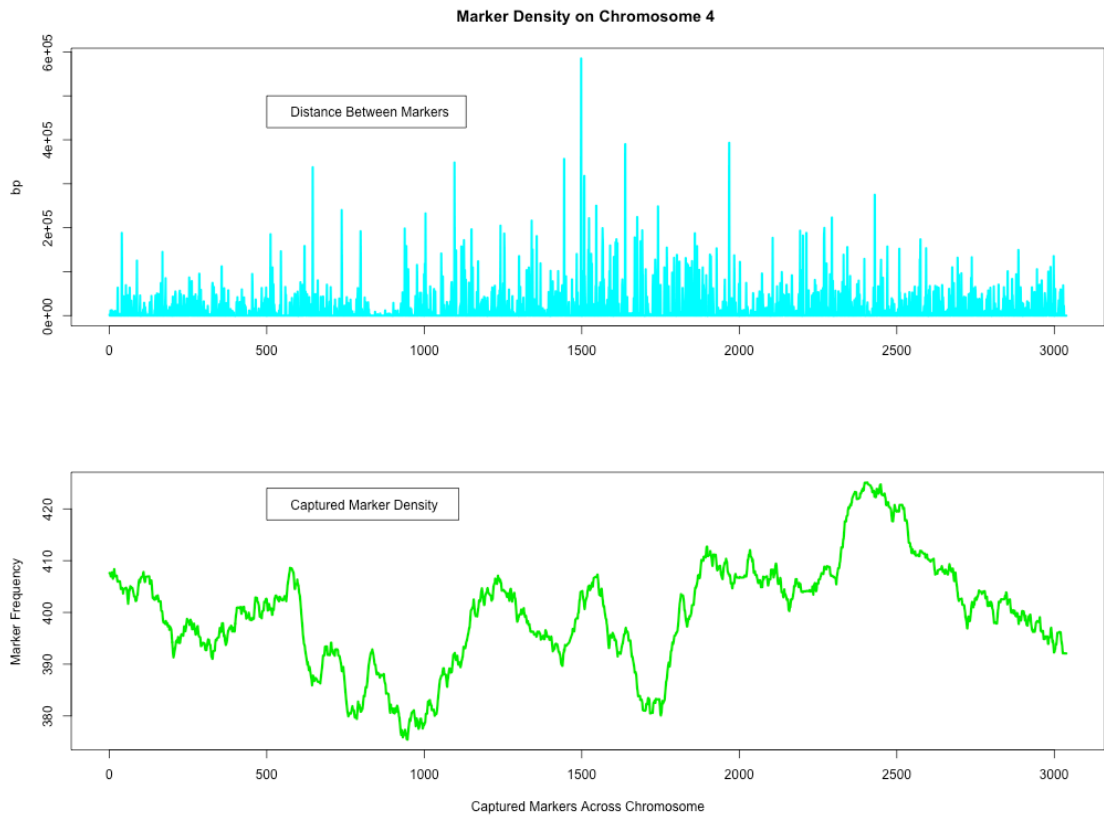
**Figure S3.8.** Marker Density of *B. distachyon* chromosome 1.



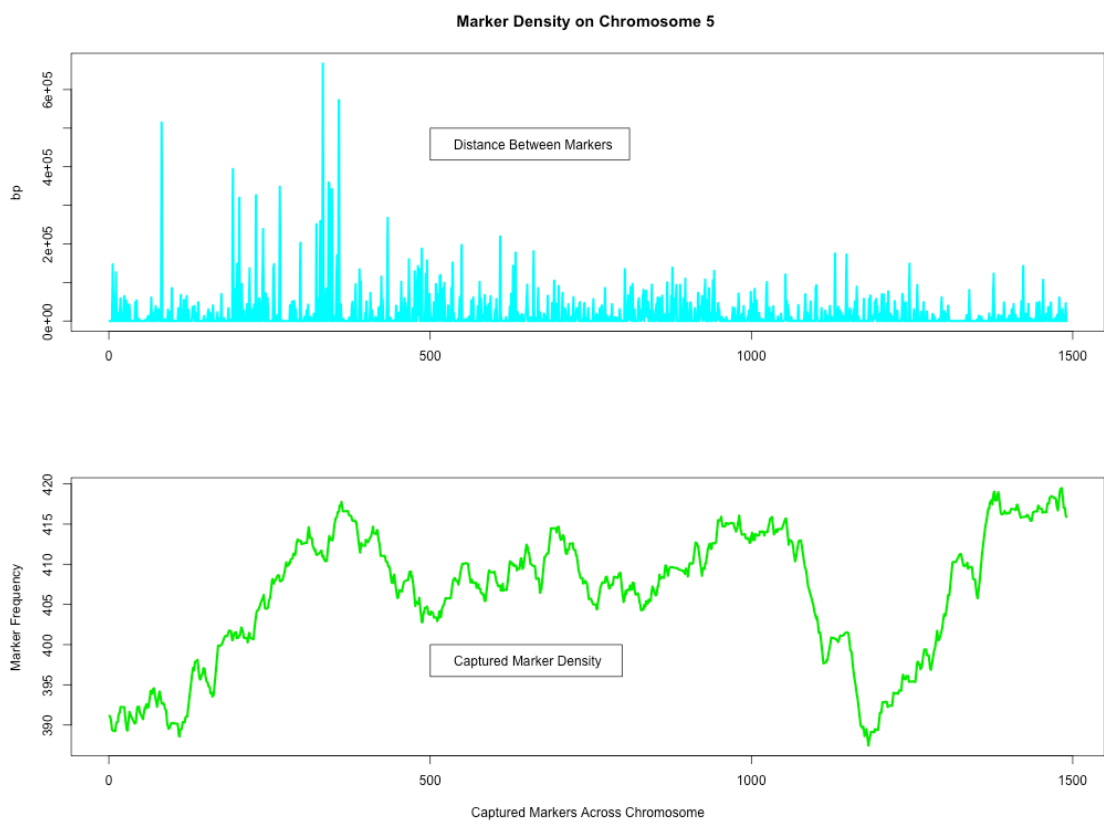
**Figure S3.9.** Marker Density of *B. distachyon* chromosome 2.



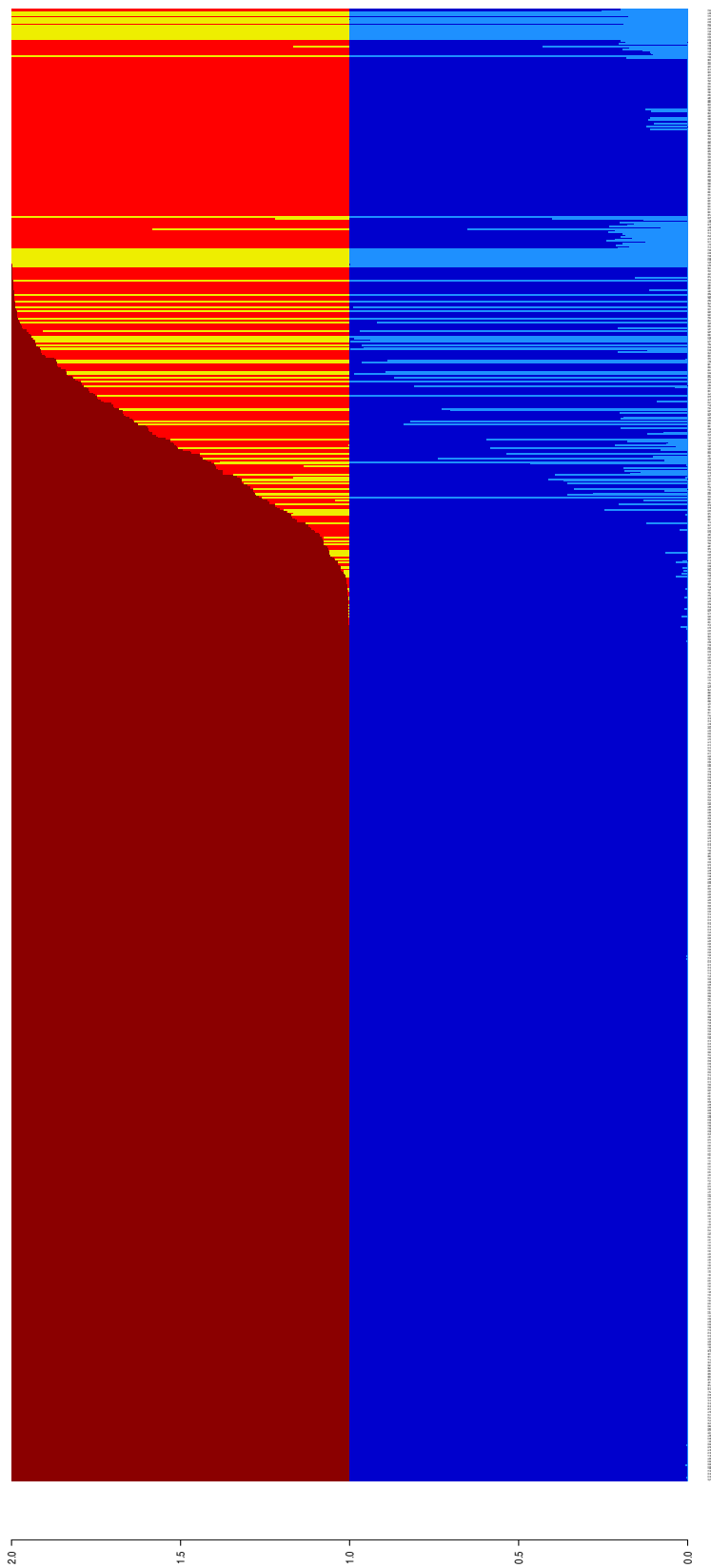
**Figure S3.10.** Marker Density of *B. distachyon* chromosome 3



**Figure S3.11.** Marker Density of *B. distachyon* chromosome 4



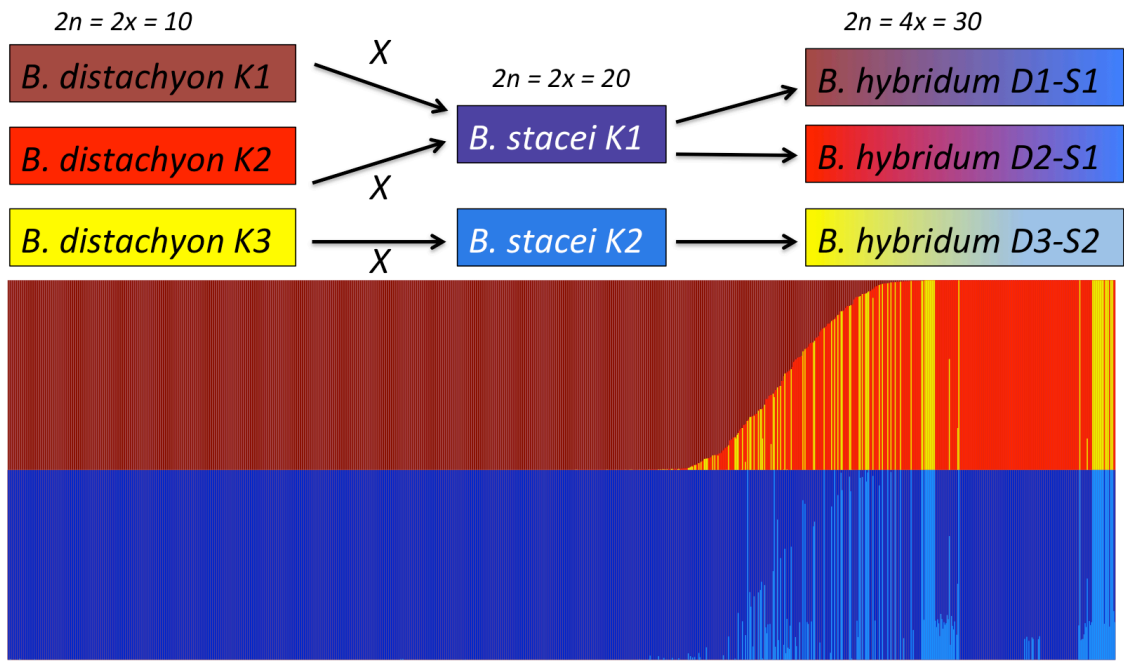
**Figure S3.12.** Marker Density of *B. distachyon* chromosome 5



**Figure S3.13.** Population structure of *B. hybridum* subgenomes using 1,008 individuals. The D genome calculated as  $K=3$  (brown, red, and yellow) and the S genome calculated as  $K=2$  (cyan and blue) using Evanno's delta  $K$  method. The difference in delta  $K$  indicates that there could be three or more hybridisation events. Structure was run in 16 replicates of 100k burnin and 100k iterations for K1-13 on each subgenome.

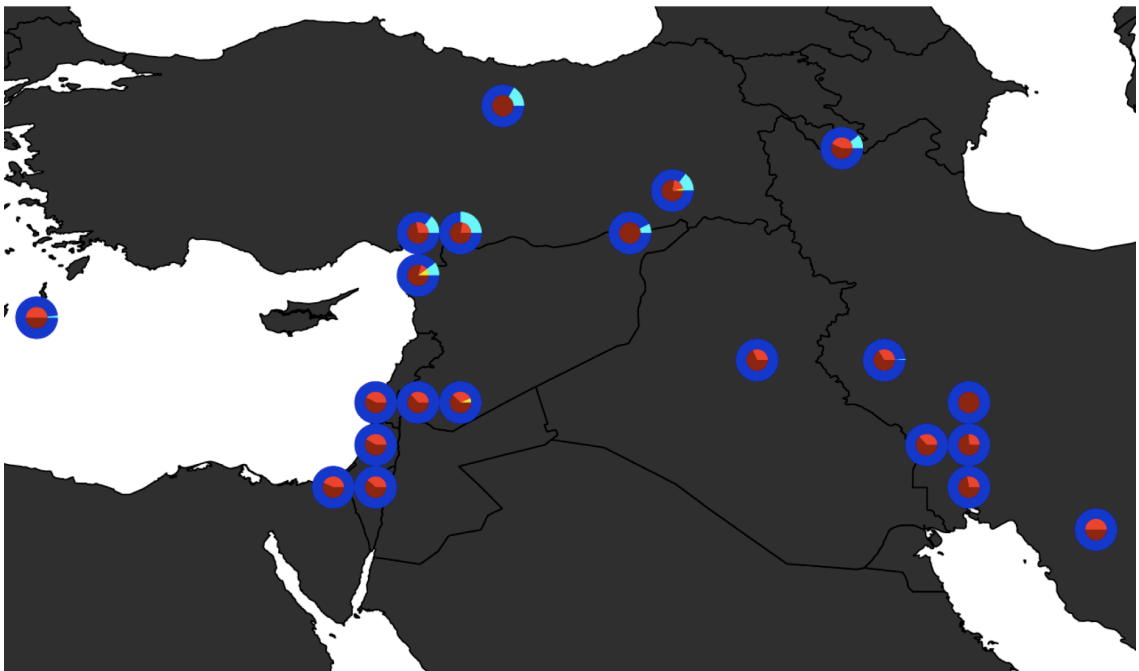


## Brachypodium species complex and multiple origins of *B. hybridum* using STRUCTURE



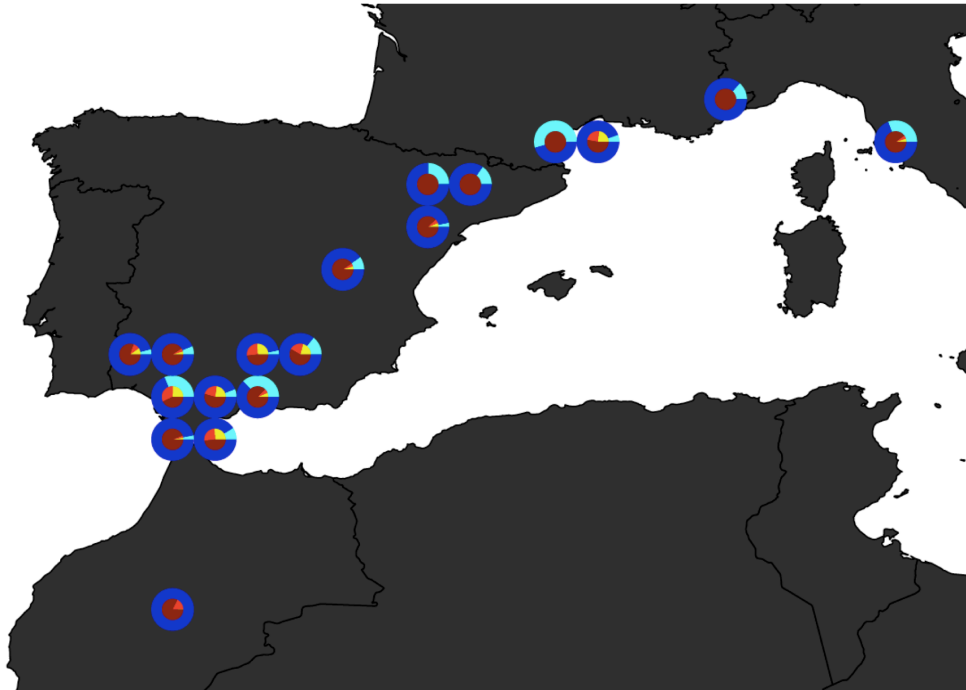
**Figure S3.14.** Population structure of *B. hybridum* subgenomes using 1,008 individuals to show likely hybridisation pattern.

## Eastern Mediterranean



**Figure S3.14.** Population structure of *B. hybridum* subgenomes using 1,008 individuals across the Eastern Mediterranean. Pie charts on maps indicate both subgenomes with one pie laid over another larger pie chart. Red-Yellow hues are the *B. distachyon*-like D subgenome, and the Blue hues are the *B. stacei*-like subgenome.

## Western Mediterranean



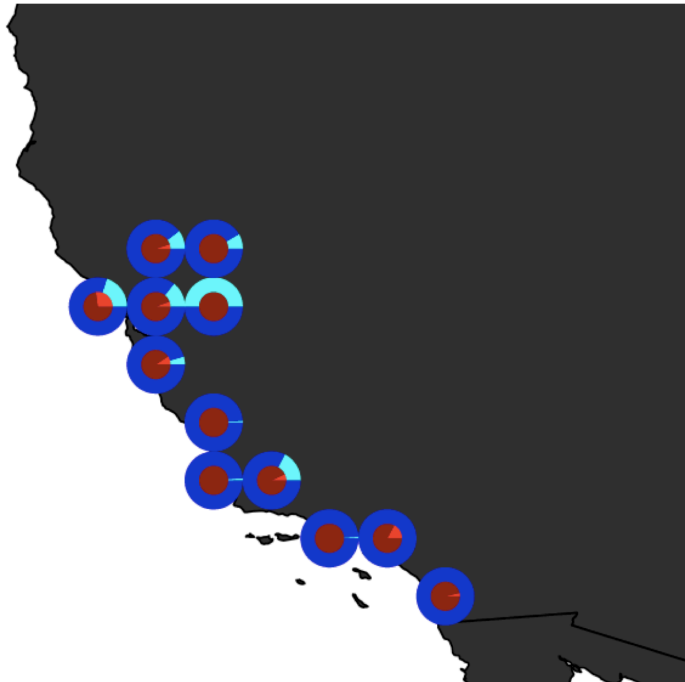
**Figure S3.15.** Population structure of *B. hybridum* subgenomes using 1,008 individuals across the Western Mediterranean. Pie charts on maps indicate both subgenomes with one pie laid over another larger pie chart. Red-Yellow hues are the *B. distachyon*-like D subgenome, and the Blue hues are the *B. stacei*-like subgenome.

## Introduced locations in Australia

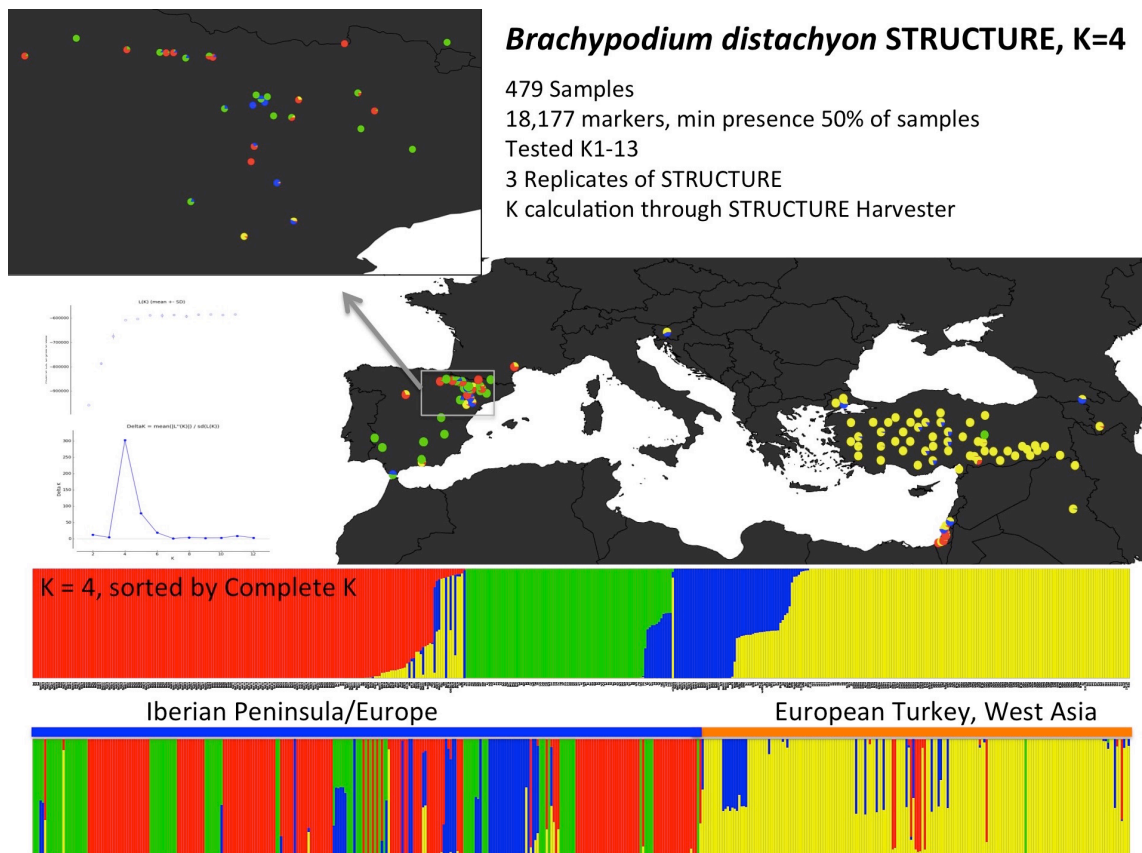


**Figure S3.16.** Population structure of *B. hybridum* subgenomes using 1,008 individuals across the South Eastern Australian Continent. Pie charts on maps indicate both subgenomes with one pie laid over another larger pie chart. Red-Yellow hues are the *B. distachyon*-like D subgenome, and the Blue hues are the *B. stacei*-like subgenome.

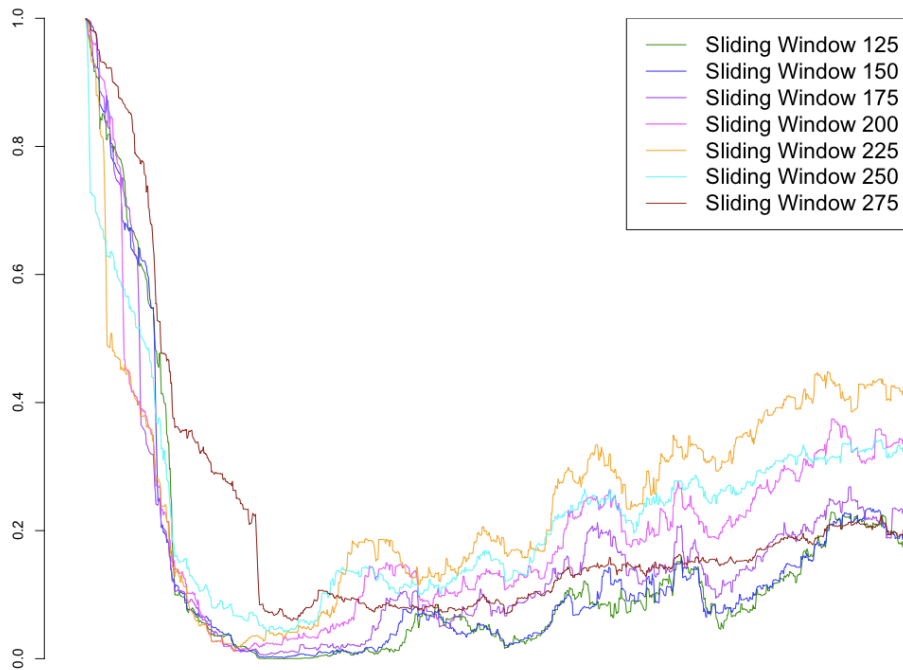
Introduced locations in North America, California State United States



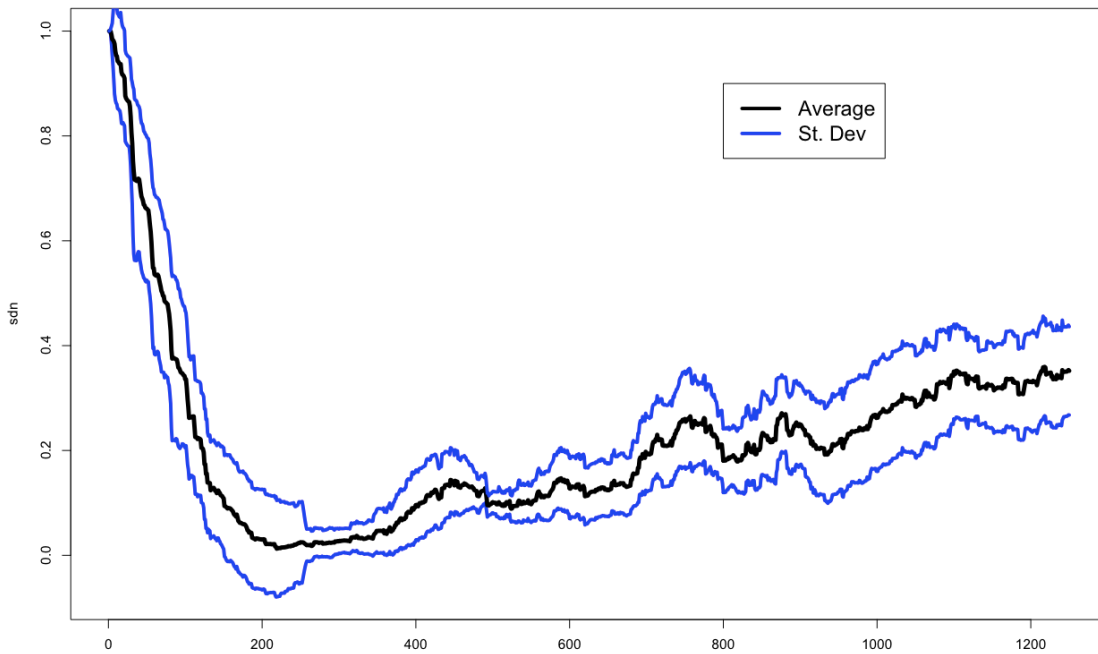
**Figure S3.17.** Population structure of *B. hybridum* subgenomes using 1,008 individuals across the United States, California in North America. Pie charts on maps indicate both subgenomes with one pie laid over another larger pie chart. Red-Yellow hues are the *B. distachyon*-like D subgenome, and the Blue hues are the *B. stacei*-like subgenome.



**Figure S3.18.** Population structure of *B. distachyon* using 479 individuals. Pie charts on maps indicate the present composition of that location's ancestral groups. Top barplot is sorted by present K, bottom barplot is sorted from west to east.



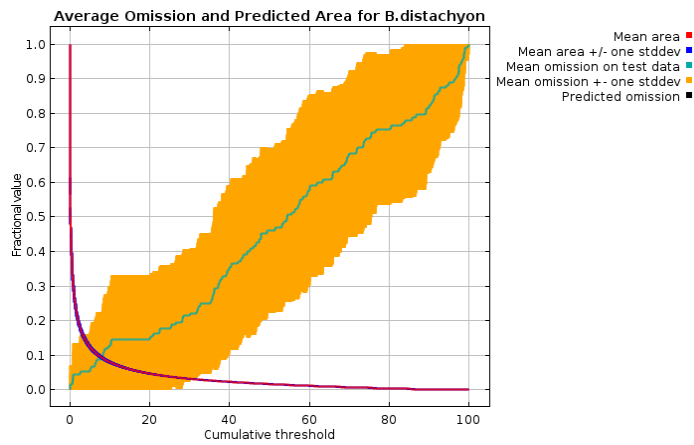
**Figure S3.19.** Linkage disequilibrium in all *B. distachyon* using seven different sliding window sizes measured in base pairs: 125k, 150k, 175k, 200k, 225k, 250k, and 275k. LD reached a  $R^2$  score of 0.1 at  $\approx 317$ k bases averaged across all windows.



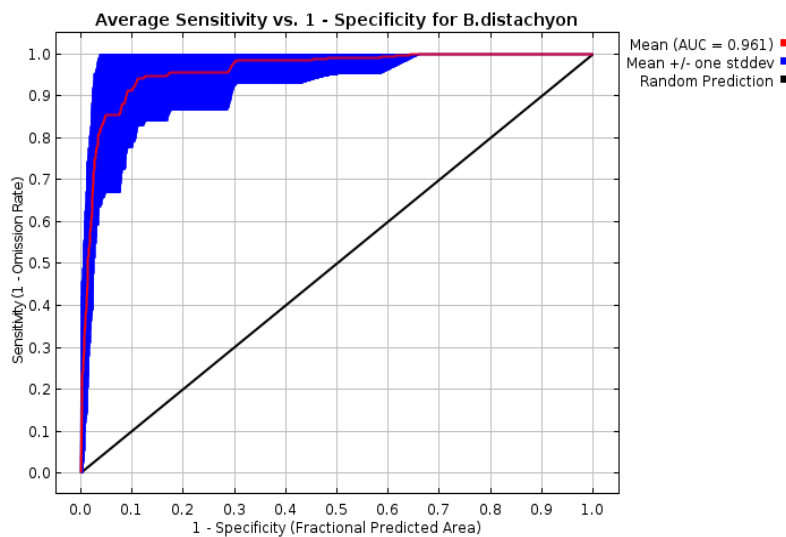
**Figure S3.20.** Linkage disequilibrium in all *B. distachyon* using seven different sliding window sizes averaged together measured in base pairs: 125k, 150k, 175k, 200k, 225k, 250k, and 275k. LD reached a  $R^2$  score of 0.1 at  $\approx 317$ k bases averaged across all windows. Average in black and  $\pm 1$  standard deviation in blue.

## Chapter IV Supplementary Material

### Chapter 4, Supplemental Figures and Tables



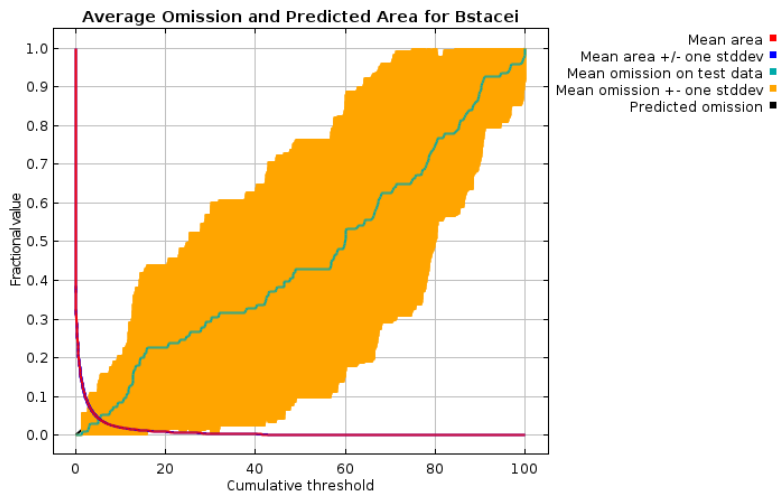
**Figure S4.1.** The Average Omission and predicted area for *B. distachyon* native range model.



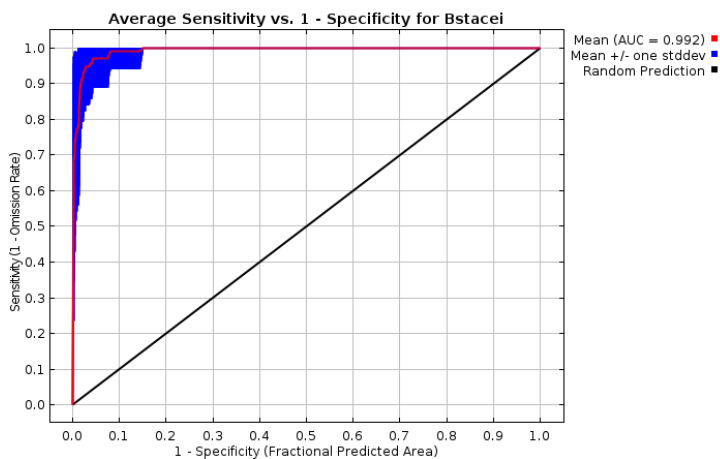
**Figure S4.2** The Average Sensitivity vs. 1 Minus the Specificity for *B. distachyon* Native.

Variable	Percent contribution	Permutation importance
bio_12_EU_WAsia_NAfrica	23.6	19.4
bio_1_EU_WAsia_NAfrica	19.6	13.9
bio_2_EU_WAsia_NAfrica	18	13.7
bio_8_EU_WAsia_NAfrica	18	5.3
bio_4_EU_WAsia_NAfrica	11.3	13.9
bio_3_EU_WAsia_NAfrica	7.5	29.8
bio_14_EU_WAsia_NAfrica	1.7	3.1
bio_9_EU_WAsia_NAfrica	0.3	0.8

**Figure S4.3.** Climate Variable Contribution and Permutation importance in the *B. distachyon* local potential area model.



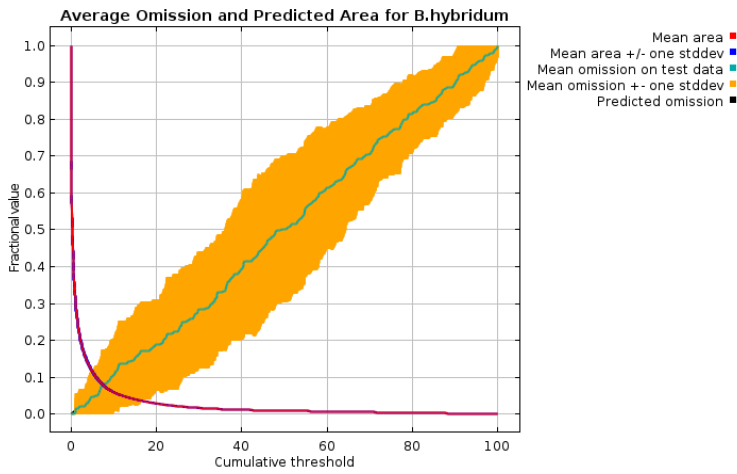
**Figure S4.4**, The Average Omission and Predicted Area for *B. stacei*.



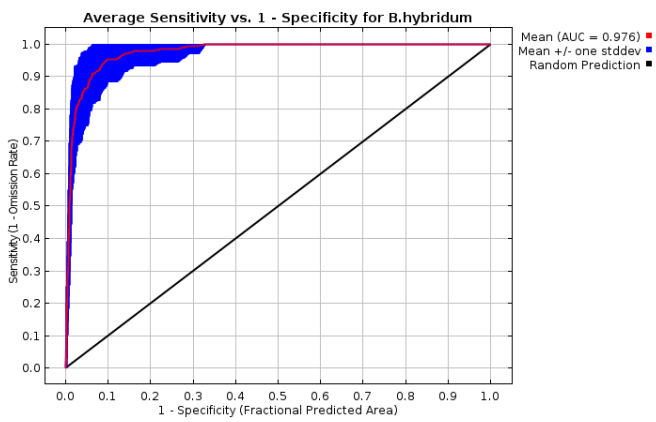
**Figure S4.5** Average Sensitivity vs. One Minus Specificity for *B. stacei* in the native range.

Variable	Percent contribution	Permutation importance
bio_1_EU_WAsia_NAfrica	24.5	20.7
bio_12_EU_WAsia_NAfrica	21.2	10.6
bio_4_EU_WAsia_NAfrica	14.1	2
bio_6_EU_WAsia_NAfrica	12.2	62
bio_17_EU_WAsia_NAfrica	12	3.5
bio_2_EU_WAsia_NAfrica	9.5	0.4
bio_3_EU_WAsia_NAfrica	6.5	0.8

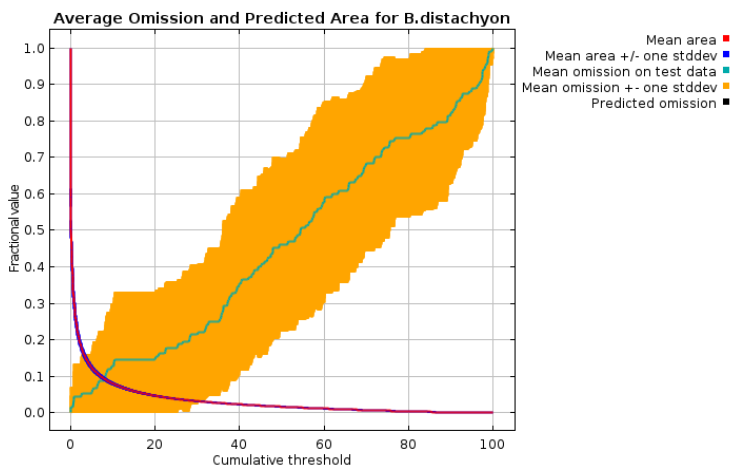
**Figure S4.6**. Climate Variable Contribution and Permutation importance in the *B. stacei* local potential area model.



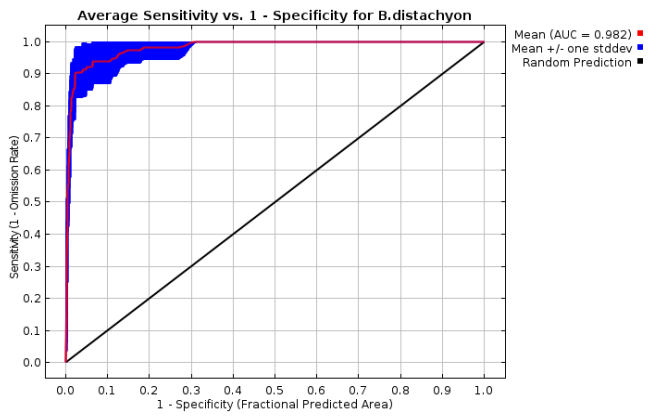
**Figure S4.7.** The Average Omission and predicted area for native potential area of *B. hybridum*.



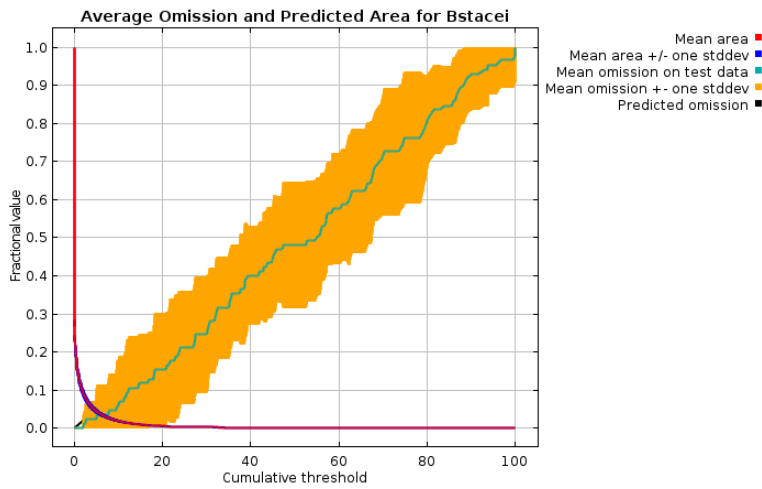
**Figure S4.8.** Average Sensitivity and One Minus Specificity of *B. hybridum*.



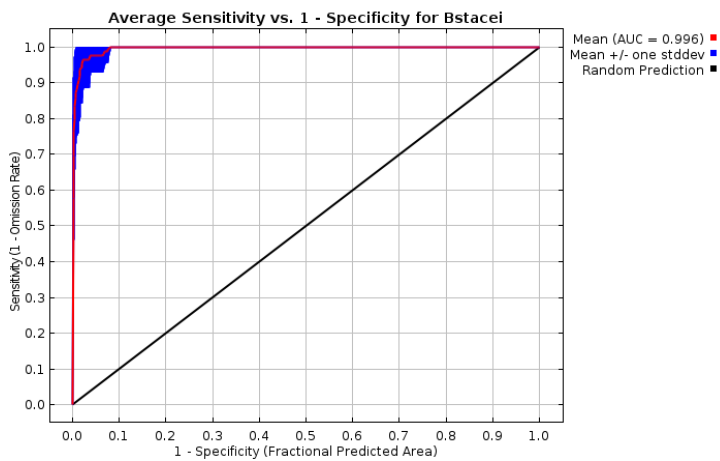
**Figure S4.9.** Average Omission and predicted area for *B. distachyon* Global Model.



**Figure 4.10.** Average Sensitivity and One Minus Specificity of *B. distachyon*.



**Figure S4.11.** *B. stacei* Global model, Average Omission and predicted area.

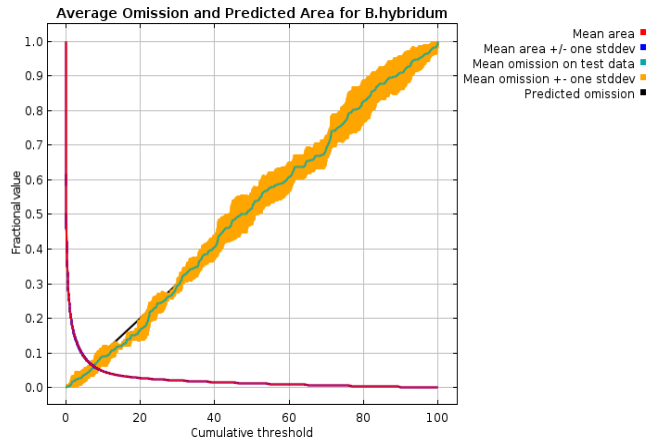


**Figure S4.12.** Global model, Average Sensitivity and One Minus Specificity of *B. stacei*.

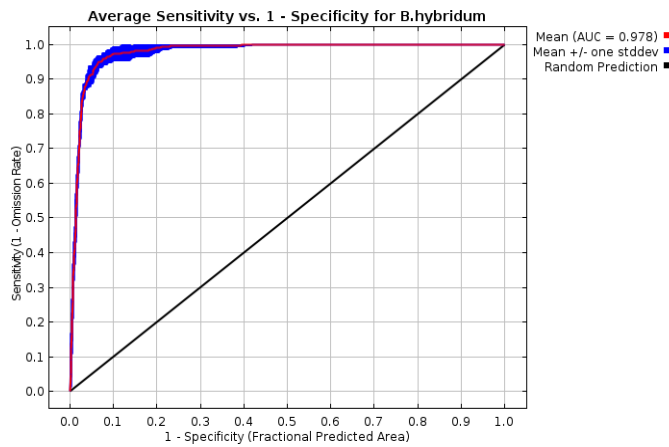


Variable	Percent contribution	Permutation importance
bio_17	25.7	2.7
bio_2	17.8	0.1
bio_6	15.6	71.4
bio_4	13.8	7.8
bio_12	13.3	2.7
bio_1	11.9	14.8
bio_3	1.9	0.5

**Figure S4.13.** *B. hybridum* Global model, Table of variable contribution to Global Modelling.



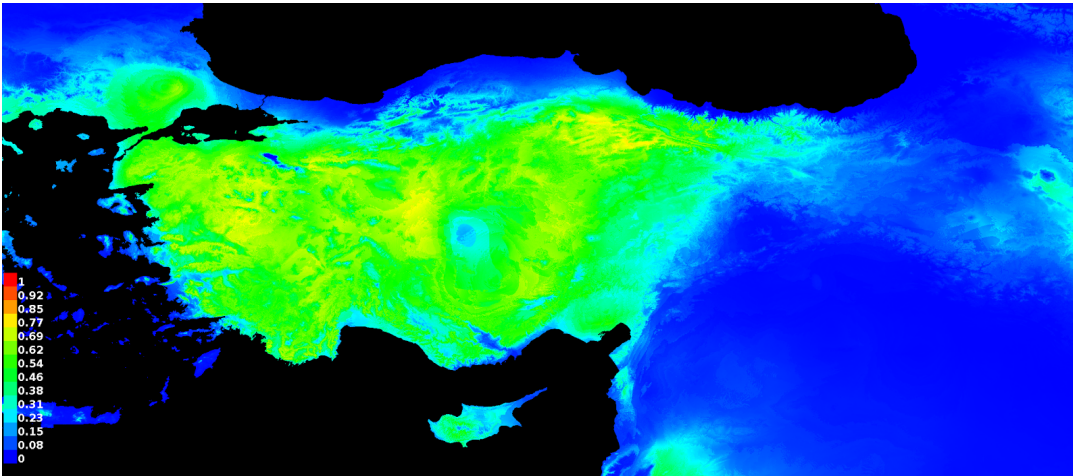
**Figure S4.14.** *B. hybridum B. hybridum* Global model, Average Omission and predicted area.



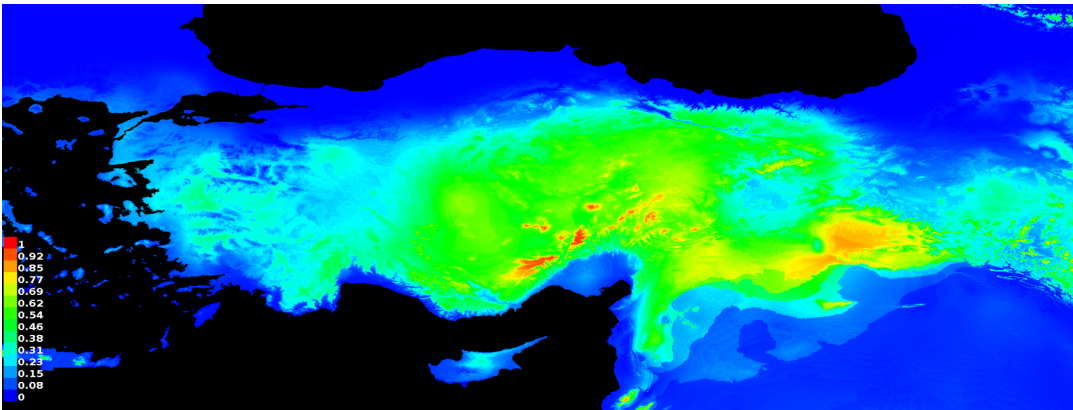
**Figure S4.15.** *B. hybridum* Global model, Average Sensitivity and One Minus Specificity.

## Maximum Entropy Predictions per Genotype

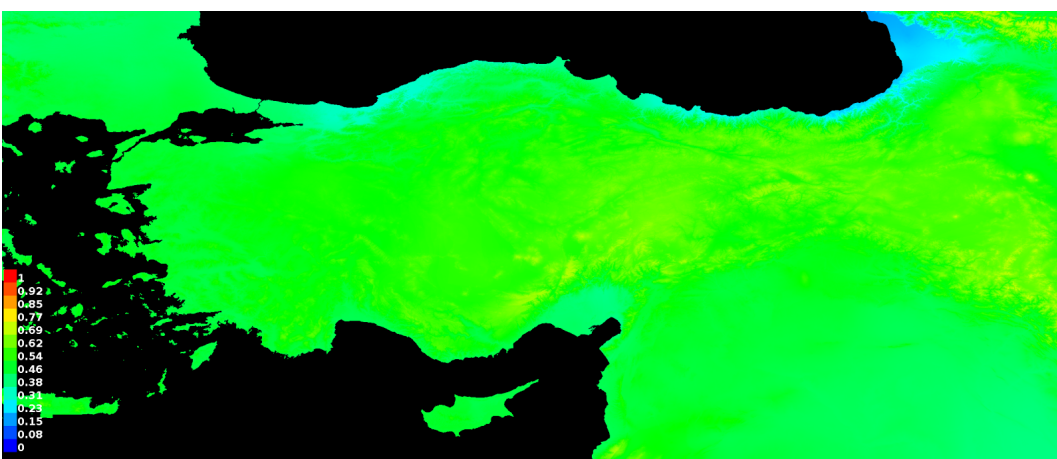
*B. distachyon* Genotypes, Turkey



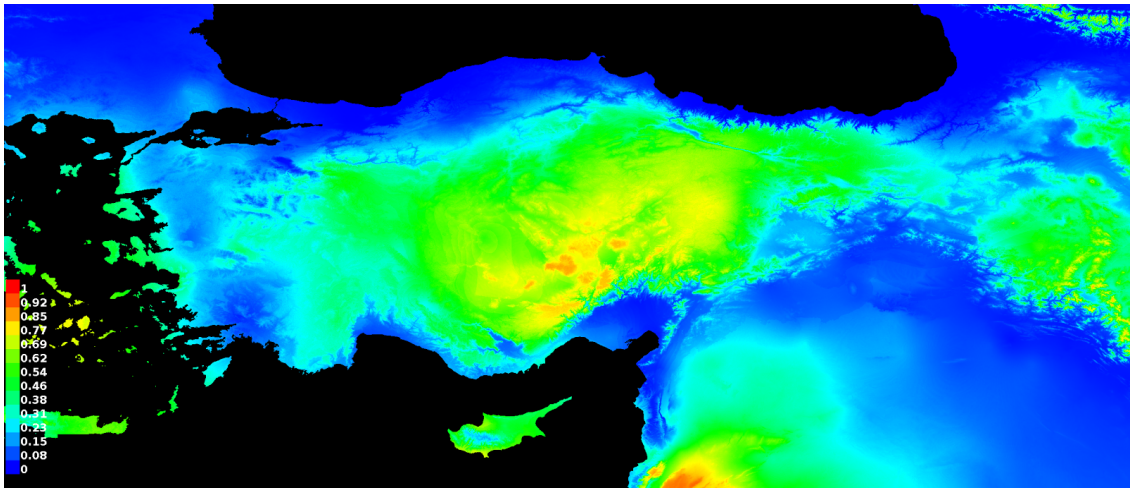
**Figure S4.16.** BdTR1-2 Potential Area.



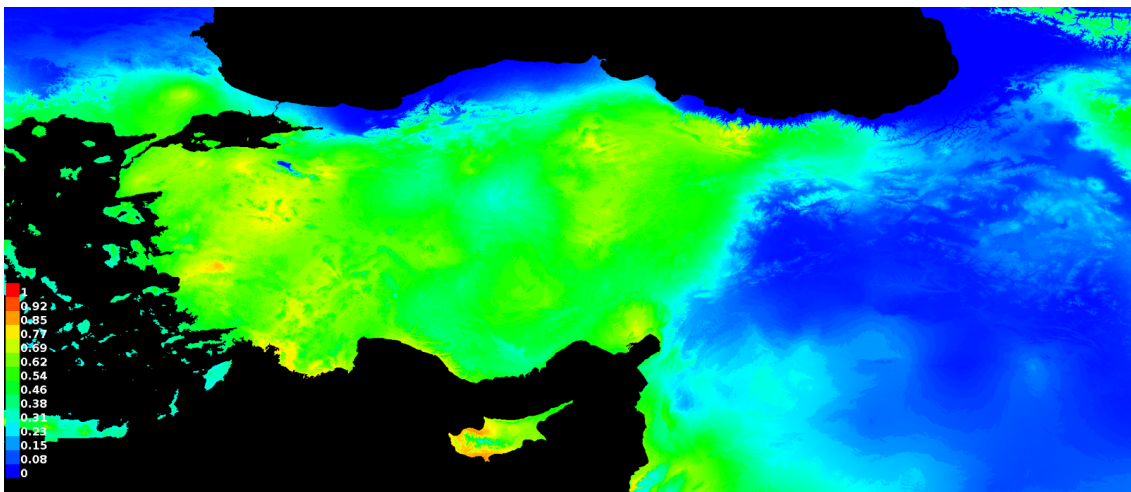
**Figure S4.17.** BdTR3 Potential Area.



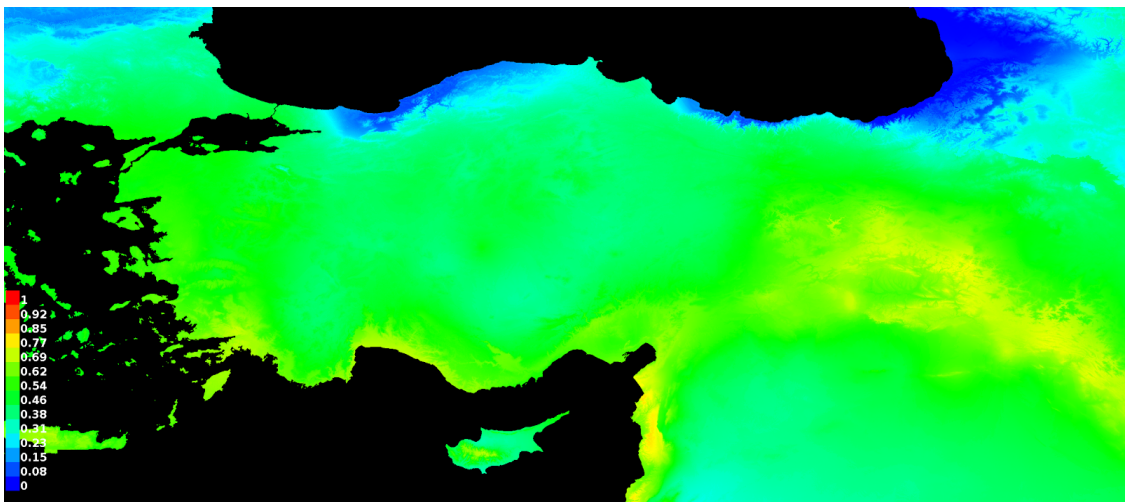
**Figure S4.18.** BdTR5 Potential Area.



**Figure S4.19.** BdTR8 Potential Area.



**Figure S4.20.** BdTR9 Potential Area.



**Figure S4.21.** BdTR10 Potential Area.

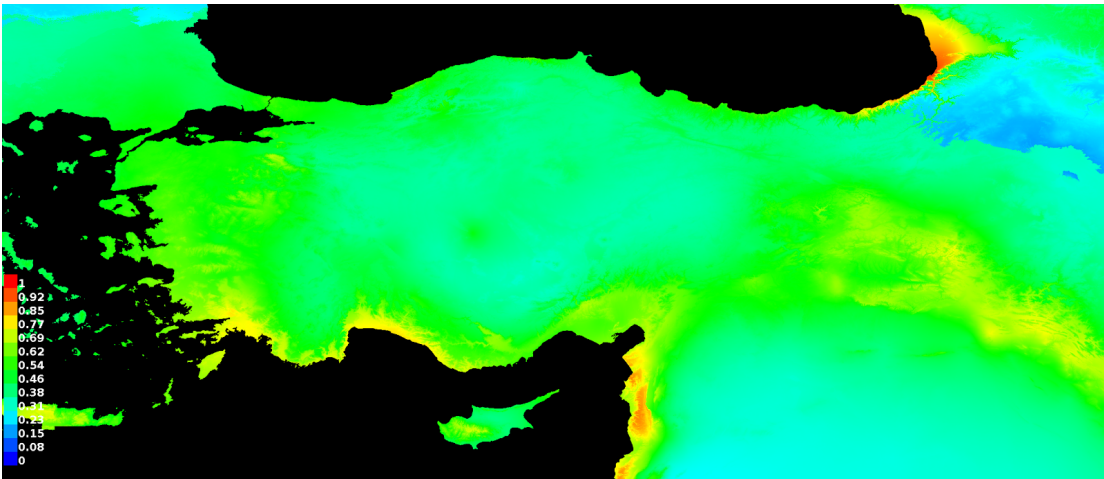


Figure S4.22. BdTR11 Potential Area.

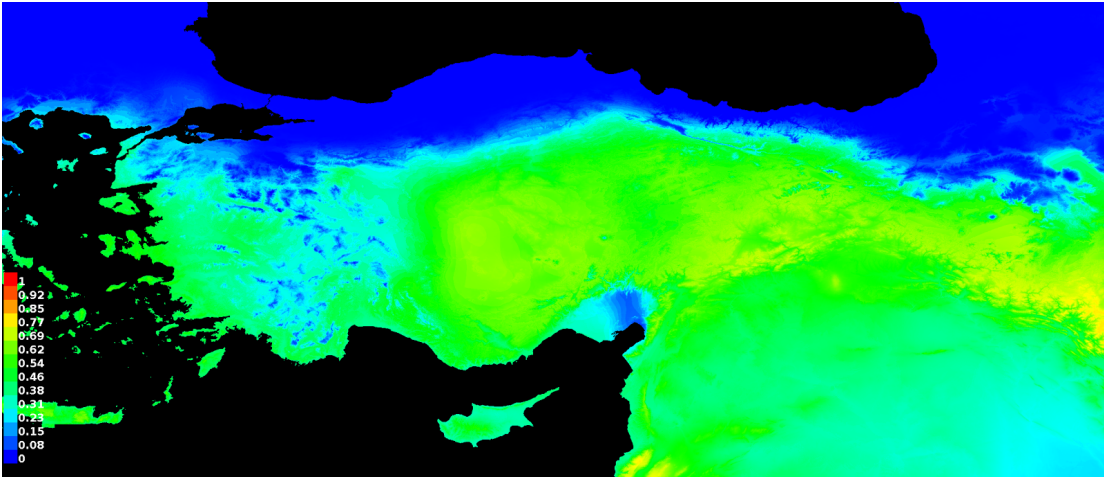
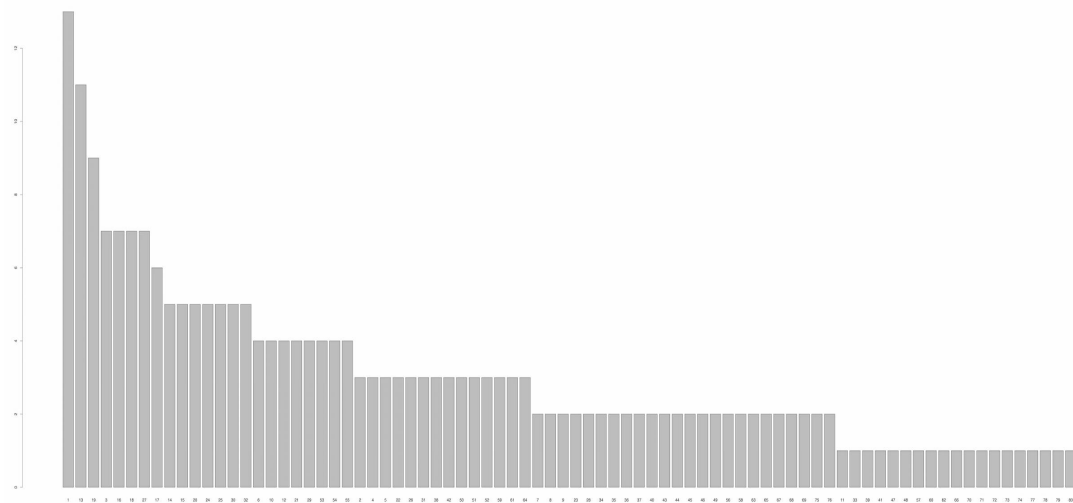


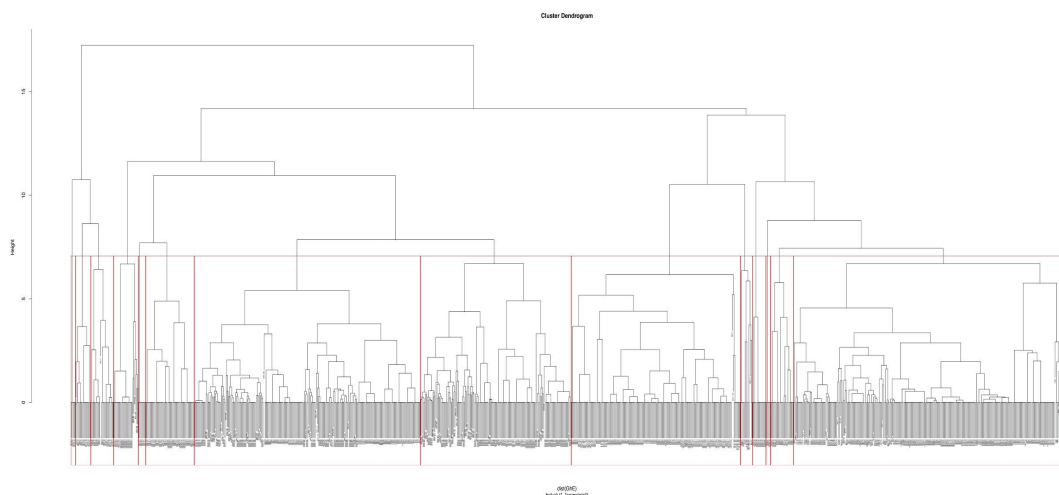
Figure S4.23 BdTR13 Potential Area.

## Chapter V Supplementary Materials

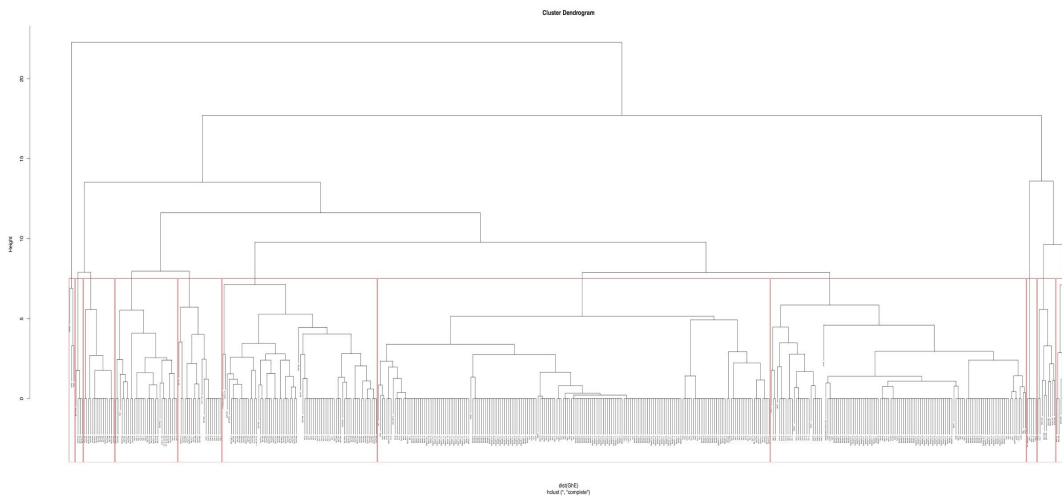
### *B. hybridum*



**Figure S5.1.** Barplots of genotypes to climate types ordered: A barplot of *B. hybridum* genotypes ordered by their number of occupied climates. Genotype one, or NRD-1 the far left, was found in 13 different climates and sampled 121 times across many locations.



**Figure S5.2.** *Brachypodium hybridum* dendrogram of Climate Variables: A climate diagram describing the relatedness of sample locations of *B. hybridum* where each leaf is an individual. *B. hybridum* locations were classified into 14 groups. By keeping each sample as a replicate of each climate group, the amount of representation each cluster has in the dendrogram is preserved and climate groups can be visualised by their number of representative locations.



**Figure S5.3** *Brachypodium distachyon* dendrogram of Climate Variables: A non-genetic diagram composed of climate data from each sample location. By clustering the 19 BioClim variables of each sample's location into eleven groups, the climate diversity of each genotype can be quantified, especially for widespread genotypes that occupy more than one cluster. Each of the 476 leaves in this dendrogram represents an individual.

---	Value	p-value
Integer	55.769925097	0.065
BioClim1 distance	-0.089455354	0.001
Geological distance R <sup>2</sup>	0.001252523	0.001
BioClim1 R <sup>2</sup>	0.01666654	0.001
F-test	949.9756	0.001

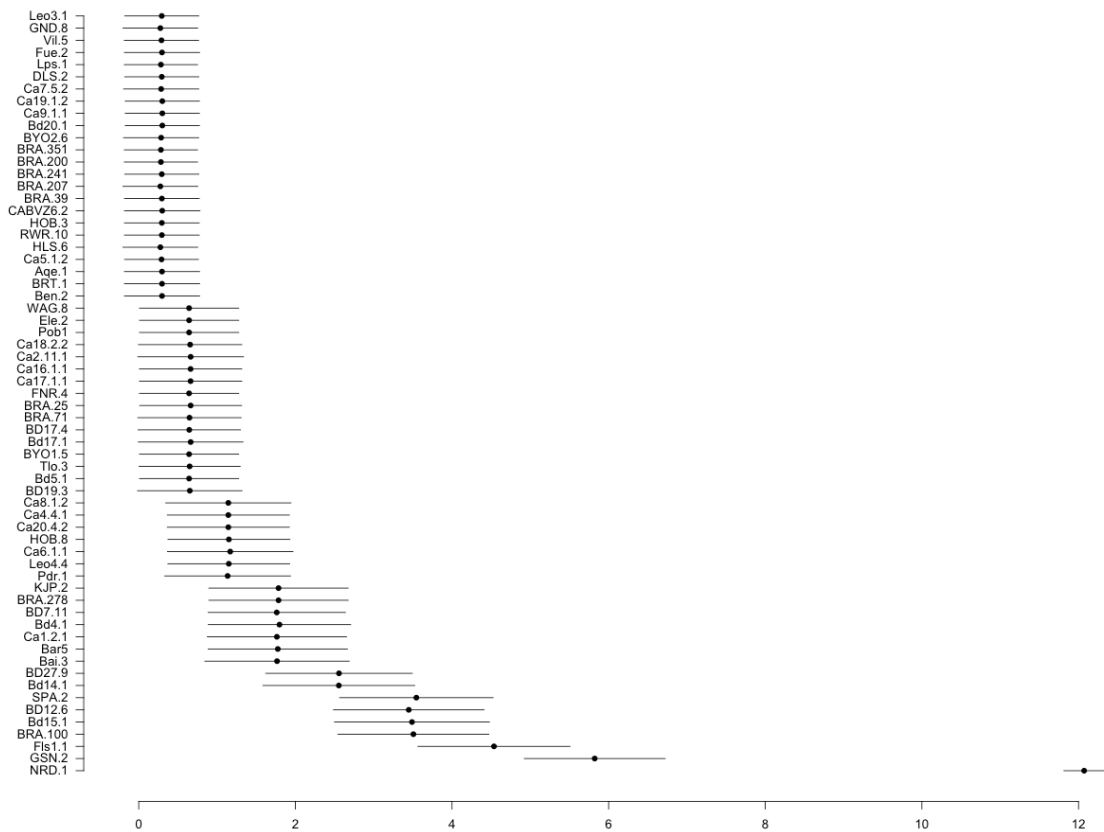
**Table S5.4.** *B. distachyon* BioClim1 and Geographic Distance Partial multi-linear regression models Test: BioClim1 annual mean temperature was tested as a variable for genetic association due to its strong correlation to explaining the distribution of *B. distachyon* in literature and MaxEnt modelling in chapter 4. BioClim1 was not found to explain significant genetic variation.

---	Value	p-value
Integer	55.725624276	0.003
BioClim4 distance	-0.005135141	0.001
Geological distance R <sup>2</sup>	0.003634472	0.001
BioClim4 R <sup>2</sup>	0.03963121	0.001
F-test	2312.955	0.001

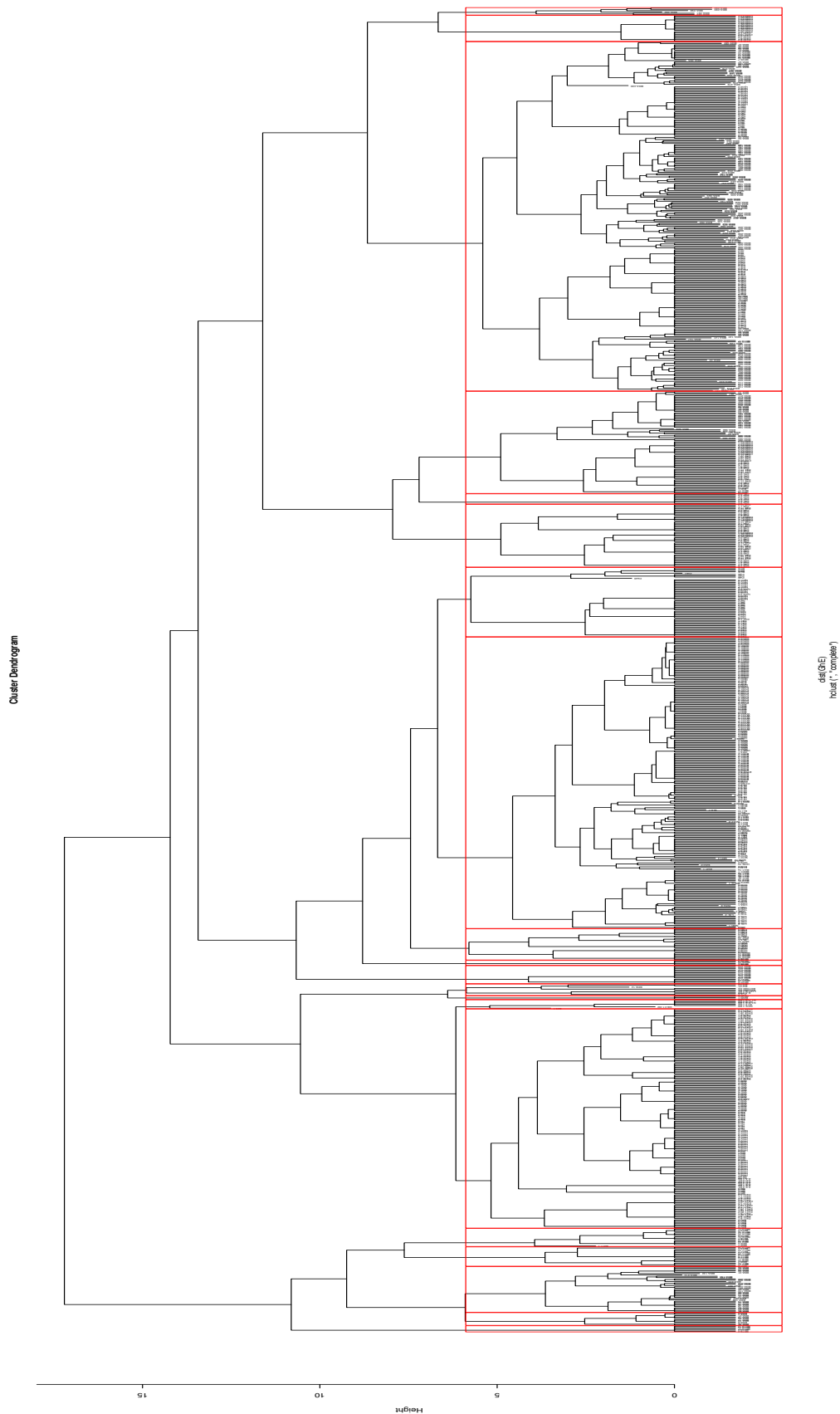
**Table S5.5.** BioClim4 and Geographic Distance Partial multi-linear regression models Test: BioClim4 the percent standard deviation of monthly annual temperatures compared to the annual mean was found to explain nearly 4% of genetic variation in *B. distachyon*. The explanation of this is not known yet in this study.

---	Value	p-value
Integer	5.138934e+01	1.00
BioClim12 distance	1.460350e-02	0.001
Geological distance R <sup>2</sup>	6.098444e-04	0.008
BioClim12 R <sup>2</sup>	0.01420049	0.001
F-test	807.3883	0.001

**Table 5.6.** BioClim12 and Geographic Distance Partial multi-linear regression models Test: BioClim12 annual mean precipitation was tested as a climate variable testing genetic association due to its strong correlation in species distribution modelling in literature and this study's MaxEnt modelling in Chapter IV. BioClim 12 wasn't found to explain significant genetic variation in *B. distachyon*.

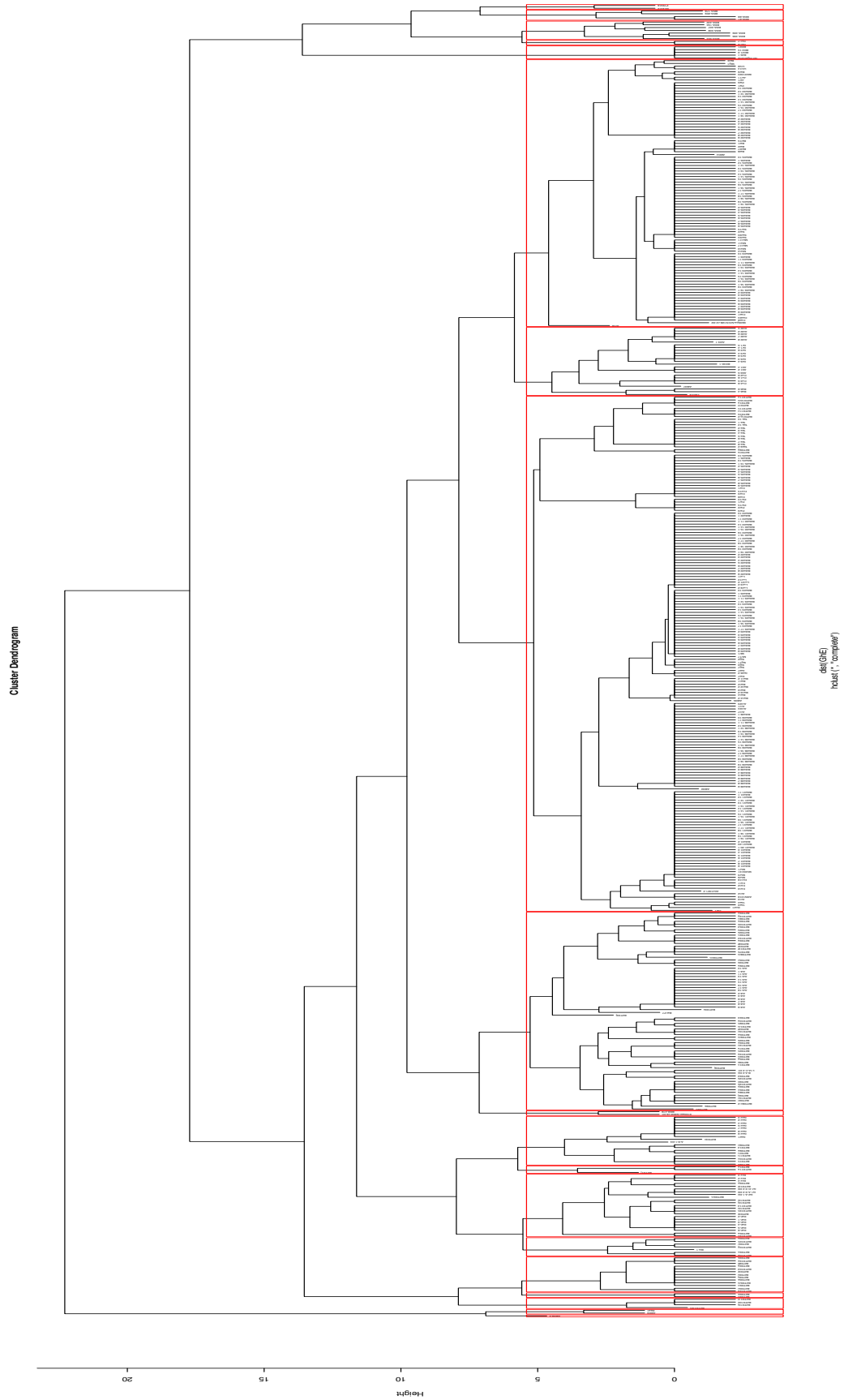


**Figure S5.7.** Larger image of permutation test for climate diversity.

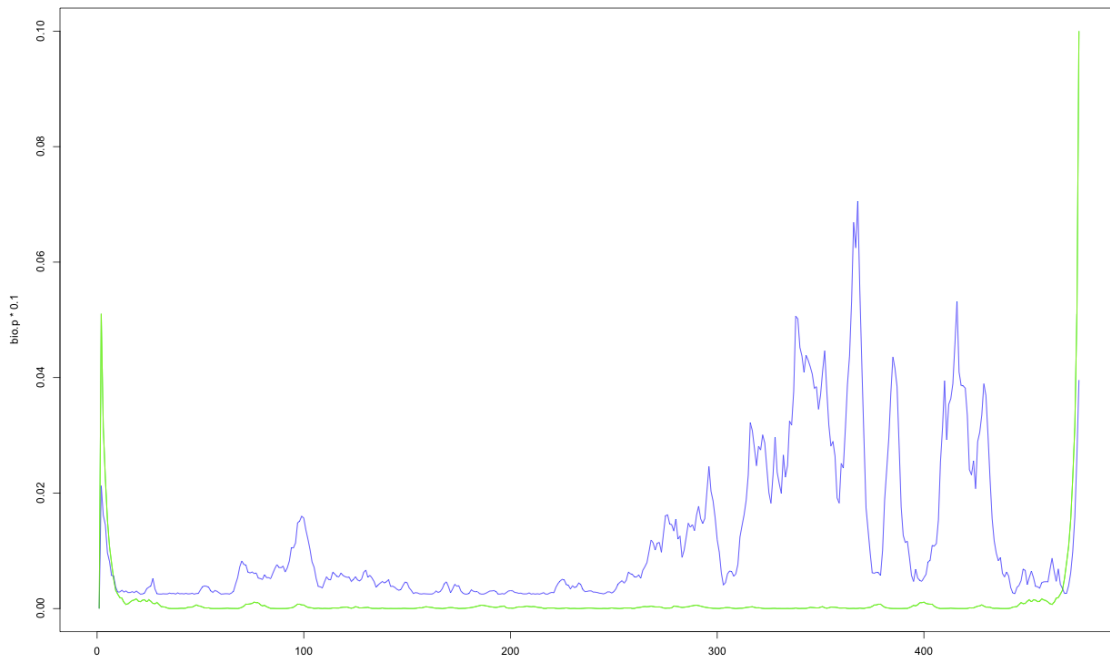


**Figure S5.8** Dendrogram of 1,015 individuals from 303 collection sites and their 19 BioClim variables as vectors that were used to cluster into 14 climate classes. Some classes are rare and others more common.





**Figure S5.8** Dendrogram of 479 individuals from 115 collection sites and their 19 BioClim variables as vectors that were used to cluster into 14 climate classes. Some classes are rare and others more common.



**Figure S5.9** Multi-linear regression test of geographic distance to climate in green to original data set, and multi-linear regression on climate data to genetic data. By rotating each sample at each position against the original data set, true associations will occur when both the original climate data set and the genetic data set are associated at the same time as the original indicated by both colours spiking at the same time as seen at the far right of the figure. This should indicate that some true climate variable association is not because of geographic distance.