Exploring adaptive genetic variation in exotic barley germplasm with landscape genomics

Dissertation to obtain the doctoral degree of Agricultural Sciences

(Dr. sc. agr.)

Faculty of Agricultural Sciences

University of Hohenheim

Institute of Plant Breeding, Seed Science and Population Genetics

Submitted by

Che-Wei Chang

from Taipei, Taiwan

2025

This thesis was accepted as a doctoral thesis (Dissertation) in fulfilment of the regulations to acquire the doctoral degree "Doktor der Agrarwissenschaften" by the Faculty of Agricultural Sciences at the University of Hohenheim, on 07^{th} of February 2025.

Date of the oral examination: 21^{th} of July 2025

Dean: Prof. Dr. Ralf Vögele

Examination Committee:

Chairperson of the oral examination:	Prof. Dr. Jörn Bennewitz
Supervisor and Reviewer:	Prof. Dr. Karl Schmid
Co-Reviewer:	Prof. Dr. Aurélien Tellier
Additional examiner:	Prof. Dr. Martin Hasselmann

Contents

Abbreviations **7**

Summary **8**

Zusammenfassung **10**

1 General Introduction **13**

1.1 Genetic bottleneck for barley breeding 14

1.2 Genetic resources for improving barley resilience 16

1.3 Evolutionary forces shaping patterns of genetic variation 17

1.4 Local adaptation of wild and cultivated barley 19

1.5 Tools for harnessing adaptive alleles 21

1.6 Objectives 23

2 Physical geography, isolation by distance and environmental variables shape genomic variation of wild barley (*Hordeum vulgare* L. ssp. *spontaneum*) in the Southern Levant **25**

2.1 Abstract 26

2.2 Introduction 27

2.3 Materials and Methods 29

2.3.1 Plant material and genotyping by sequencing 29

2.3.2 Environmental data 30

2.3.3 Inference of population structure 31

2.3.4 Analysis of gene flow 31

2.3.5 Partitioning genomic variation 33

2.3.6 Linkage disequilibrium 35

2.3.7 Identification of selection signatures 35

2.3.8 Gene ontology enrichment 36

2.4 Results 37

2.4.1 Summary of genotyping data 37

2.4.2 Population structure and spatial genetic pattern 37

2.4.3 Geographical pattern of gene flow 41

2.4.4 Genetic variation explained by environment and space 43

2.4.5 Adaptive candidates and GO enrichment 46

2.5 Discussion 47

2.5.1 Strong population structure of B1K+ and IPK genebank collections 49

2.5.2 Evidence for geographical pattern of gene flow 50

2.5.3 Effects of environment and geographical distance on SNP variation 51

2.5.4 Lack of strong evidence to pinpoint adaptive loci 53

2.5.5 Conclusion and outlook 55

2.6 Acknowledgments 56

2.7 Author contributions 56

2.8 Competing interests 57

2.9 Data archiving 57

3 *GGoutlieR*: an R package to identify and visualize unusual geo-genetic patterns of biological samples 58

3.1 Abstract 59

3.2 Statement of need 59

3.3 Algorithm of *GGoutlieR* 60

3.4 Example 61

3.4.1 Outlier identification 61

3.4.2 Visualization of unusual geo-genetic patterns 63

3.5 Availability 66

3.6 Acknowledgements 66

4 Predicting the geographical origins of barley genebank accessions using deep learning: Can large sample sizes improve genome-environment association studies? 67

4.1 Abstract 68

4.2 Introduction 69

4.3 Materials and Methods 71

4.3.1 Overview of *GEAplus* framework 71

4.3.2 Analysis of a global barley landrace collection 72

4.3.3 SLiM simulation 75

4.3.4 Accuracy of geographical origin inference 78

4.3.5 Performance of *GEAplus* 79

4.4 Results 82

4.4.1 Prediction accuracy of geographical origins of barley landraces . 82

4.4.2 GEA of barley landraces 83

4.4.3 Prediction accuracy of geographical origins with simulated data . 84

4.4.4 Limited benefit of imputing environmental data for GEA 87

4.5 Discussion 91

4.5.1 Performance of geographical origin inference 91

4.6 Acknowledgments 96

5 General Discussion **97**

5.1 Landscape genomics for barley germplasm 97

5.2 Strategies for future crop adaptation research 100

5.3 Future of barley pre-breeding 103

5.3.1 Utility of environmental data of geographical origins 103

5.3.2 Accelerating pre-breeding with technological advances 105

Acknowledgements **108**

Bibliography **109**

A Physical geography, isolation by distance and environmental variables shape genomic variation of wild barley (*Hordeum vulgare* L. ssp. *spontaneum*) in the Southern Levant - Supplementary information **138**

A.1 Genotypic data filtration and processing 139

A.2 Construction of synthetic environmental variables and variable selection 140

A.3 Classification of barrier and non-barrier pixel 142

A.4 Supplementary Tables 143

A.5 Supplementary Figures 153

A.6 Supplementary Files 166

B *GGoutlieR*: an R package to identify and visualize unusual geo-genetic patterns of biological samples - Supplementary information **167**

B.1 Methods 168

B.1.1 Overview 168

B.1.2 Assumptions 168

B.1.3 Geo-genetic outlier detection with K nearest neighbors 169

C Predicting the geographical origins of barley genebank accessions using deep learning: Can large sample sizes improve genome-environment association studies? - Supplementary information 174

C.1 SLiM simulation 175

C.1.1 Mutational effect of QTLs 175

C.1.2 Plasticity of selection 175

C.2 Supplementary Figures 175

D Curriculum Vitae 179

Abbreviations

CBI	Coalescent-based inference	LFMM	Latent factor mixed model
CEN	CENTRORADIALIS	LM	Linear model
dbMEM	distance-based Moran's eigenvector map	MAF	Minor allele frequency
EEMS	Estimated effective migration surfaces	MCMC	Markov chain Monte Carlo
FDR	False discovery rates	PCA	Principal component analysis
GEA	Genome-environment association	PC	Principal component
GEBV	Genomic estimated breeding value	PPD-H1	PHOTOPERIOD-H1
GBS	Genotyping-by-sequencing	QTL	Quantitative trait locus
GBLUP	Genomic best linear unbiased prediction	RDA	redundancy analysis
GO	Gene ontology	SNP	Single-nucleotide polymorphism
GPS	Global positioning system	SV	Structural variant
GWAS	Genome-wide association study	TDR	True positive rate
KNN	K-nearest neighbor	VIF	Variance inflation factor
LD	Linkage disequilibrium	VRN-H1	VERNALIZATION-H1

Summary

Understanding genetic variation underlying local adaptation is essential for improving crop resilience to address challenges posed by climate change. Barley (*Hordeum vulgare* L. ssp. *vulgare*), one of the most important crops, is suitable for studying local adaptation due to its remarkable adaptability. This PhD dissertation investigated adaptive genetic variation in exotic barley germplasm, including wild barley (*Hordeum vulgare* ssp. *spontaneum*) and barley landraces, from diverse environments and explored strategies to improve the use of genebank accessions for harnessing valuable genetic variants.

In the first study, local adaptation in wild barley populations from the Southern Levant was explored using landscape genomics approaches, combining genomic data with the climatic and soil properties of geographical origins. Through redundancy analysis (RDA), we found spatial autocorrelation explained 45% of genomic variation, and environmental factors accounted for 15%. Adaptive signatures were identified in the pericentromeric regions by the population-genetics-based scans and genome-environment association (GEA) scans, but they mostly disappeared when the population structure was considered. Our findings overall highlighted the role of nonselective forces in shaping the genetic variation of wild barley even in divergent environments.

The second study addressed challenges in passport data quality control for large-scale samples, such as germplasm collections in genebanks. The R package *GGoutlieR* was developed in this work to tackle the shortcomings of traditional manual data

cleaning. It efficiently detects and visualizes samples with unusual geo-genetic patterns by characterizing geography-genotype associations with distance-based statistics via K-nearest neighbors and calculating empirical p-values accordingly. By streamlining data cleaning and quality control, *GGoutlieR* supports more reliable landscape genomics studies, which is crucial for studying loci involved in local adaptation.

The third study explored the use of neural networks to predict geographical origins for genebank accessions lacking passport data, enabling their integration into genome-environment association (GEA) analyses. Neural network models demonstrated high prediction accuracy in cross-validation. Incorporating imputed environmental data (N = 11,032) into GEA, using barley flowering time genes as benchmarks, revealed complementary detection of genomic regions near flowering time genes compared to regular GEA (N = 1,626). Furthermore, simulations of polygenic local adaptation in selfing species showed that GEA power is insensitive to large sample sizes. These findings suggest that GEA with imputed environmental data can be a complementary approach for uncovering novel adaptive loci that might remain undetected in conventional GEA, rather than improving the statistical power of GEA.

Overall, this dissertation contributes to understanding the adaptive genetic variation in barley and expanding methodologies in landscape genomics, providing a direction for the future development of GEA approaches to better support allele mining for pre-breeding to enhance crop resilience.

Zusammenfassung

Ein besseres Verständnis der genetischen Variation, die der lokalen Anpassung zugrunde liegt, ist von entscheidender Bedeutung für die Verbesserung der Widerstandsfähigkeit von Kulturpflanzen, um den Herausforderungen des Klimawandels zu begegnen. Gerste (*Hordeum vulgare* L. ssp. *vulgare*), eine der wichtigsten Nutzpflanzen, eignet sich aufgrund ihrer bemerkenswerten Anpassungsfähigkeit für die Untersuchung der lokalen Anpassung. Diese Dissertation untersuchte die adaptive genetische Variation in exotischem Gerstenkeimplasma, einschließlich Wildgerste (*Hordeum vulgare* ssp. *spontaneum*) und Gerstenlandrassen, aus verschiedenen natürlichen Lebensräumen und erforschte Strategien zur Verbesserung der Nutzung von Genbankkeimplasma zur Nutzbarmachung wertvoller genetischer Varianten.

In der ersten Studie wurde die lokale Anpassung von Wildgerstenpopulationen aus der südlichen Levante mit Hilfe von landschaftsgenomischen Ansätzen untersucht, wobei genomische Daten mit den Klima- und Bodeneigenschaften des geografischen Herkunftsgebietes kombiniert wurden. Mit Hilfe einer Redundanzanalyse (RDA) konnten wir feststellen, dass die räumliche Autokorrelation 45 % der genomischen Variation erklärte, während Umweltfaktoren für 15 % verantwortlich waren. Adaptive Signaturen wurden in den perizentromerischen Regionen durch populationsgenetische Scans und Genom-Umwelt-Assoziations-Scans (GEA) identifiziert; diese verschwanden aber größtenteils, wenn die Populationsstruktur mitberücksichtigt wurde. Unsere Ergebnisse unterstrichen insgesamt die Rolle nicht-selektiver Kräfte bei der Gestaltung der genetischen Variation von Wildgerste selbst in sehr unterschiedlichen natür-

lichen Lebensräumen.

Die zweite Studie befasste sich mit den Herausforderungen im Umgang mit der Genauigkeit von Passdaten bei Proben von großem Umfang, wie z. B. Keimplasmasammlungen in Genbanken. Das R-Paket *GGoutlieR* wurde in der vorliegenden Arbeit entwickelt, um die Schwächen der traditionellen manuellen Datenbereinigung zu beheben. Es erkennt und visualisiert effizient Proben mit ungewöhnlichen geogenetischen Mustern, indem es geografisch-genotypische Assoziationen mit distanzbasierten Statistiken über K-nearest neighbors charakterisiert und entsprechend empirische p-Werte berechnet. Durch die Vereinfachung der Datenbereinigung und Qualitätskontrolle unterstützt *GGoutlieR* zuverlässigere landschaftsgenomische Studien, was für die Untersuchung von Loci, die an der lokalen Anpassung beteiligt sind, von entscheidender Bedeutung ist.

Die dritte Studie untersuchte die Verwendung neuronaler Netze zur Vorhersage der geografischen Herkunft von Genbank-Zugängen, für die keine Passdaten vorliegen, und ermöglichte so deren Integration in Genom-Umwelt-Assoziationsanalysen (GEA). Neuronale Netzwerkmodelle zeigten bei der Kreuzvalidierung eine hohe Vorhersagegenauigkeit. Die Einbeziehung von imputierten Umweltdaten (N = 11,032) in die GEA unter Verwendung von Gerstenblütezeitgenen als Benchmarks ergab eine ergänzende Erkennung von genomischen Regionen in der Nähe von Blütezeitgenen im Vergleich zur regulären GEA (N = 1,626). Darüber hinaus zeigten Simulationen der polygenen lokalen Anpassung bei sich selbst vermehrenden Arten, dass die GEA-Leistung unempfindlich gegenüber großen Stichproben ist. Diese Ergebnisse deuten darauf hin, dass GEA mit imputierten Umweltdaten ein ergänzender Ansatz zur Aufdeckung neuartiger adaptiver Loci sein kann, die bei herkömmlicher GEA unentdeckt bleiben könnten, und weniger dazu dient, die statistische Leistung von GEA zu verbessern.

Insgesamt trägt diese Dissertation zum Verständnis der adaptiven genetischen Variation in Gerste sowie zur Erweiterung der Methoden der Landschaftsgenomik bei. Sie gibt eine Richtung für die künftige Entwicklung von GEA-Ansätzen vor, um das

Allel-Mining vor der Züchtung besser zu unterstützen, mit dem Ziel, die Widerstandsfähigkeit von Pflanzen zu verbessern.

Chapter 1

General Introduction

Barley (*Hordeum vulgare* L. ssp. *vulgare*) is one of the world's most important cereal crops in terms of production. It has a total global production of 155 million tonnes and a harvested area of approximately 47 million hectares in 2022, making it the fourth-ranked cereal after maize, wheat, and rice (FAO 2022). Nowadays, barley production is predominantly concentrated in high and upper-middle-income economies, with leading producers including the Russian Federation, Australia, France, Germany, Canada, and Turkey (FAO 2022). The primary uses of barley are as animal feed (55-60%) and as a main ingredient in malts (30-40%) for the brewing industry (Verma et al. 2022). Nevertheless, barley is also a staple food in Central Asia, the Middle East, and North Africa because of its adaptability to low-input agriculture in marginal areas (Verma et al. 2022). In addition to low-input systems, barley can also adapt to diverse climates. Its cultivation spans various environments, from Scandinavia to semi-arid regions and even the Himalayas, due to its resilience to diverse climatic conditions (Verma et al. 2022). The strong adaptability makes barley an outstanding model for studying the potential of exotic genetic resources in addressing climate change (Dawson et al. 2015; Marok et al. 2021). In the past decades, traditional landraces and accessions of wild relatives, many of which are interfertile with cultivated barley, have been extensively collected and preserved as genetic resources in ex-situ genebanks (Dawson et al.

2015; Milner et al. 2019). These extensive germplasm collections provide researchers and breeders with ready access to diverse genetic material, facilitating advances in barley breeding.

1.1 Genetic bottleneck for barley breeding

Crop domestication is a co-evolutionary process that humans establish an environment to control the growth and reproduction of wild plant species to harvest resources (Purugganan 2022). In the past thousands of years, humans have selectively bred wild plant progenitors for desirable traits, such as erect plant structure, reduced shattering, and increased yield, leading to cultivated plants that are morphologically and genetically different from their wild ancestors (Doebley et al. 2006; Huang et al. 2022). This domestication process laid the foundation for modern breeding, which further refined crop varieties through controlled crosses, selection, and innovations in biotechnology, such as genetic transformation and gene editing.

Modern cultivated barley descends from wild barley (*Hordeum vulgare* L. ssp. *spontaneum*), dating back to the domestication approximately 10,000 years ago (Tanno and Willcox 2012; Zohary et al. 2012). Unlike other crops, domestication did not dramatically alter the morphology of barley. Researchers have traditionally distinguished domesticated barley by its non-brittle spike, a trait controlled by two linked loci, *Btr1* and *Btr2* (Takahashi and Hayashi 1964), which result from the deletion of two genes on chromosome 3H (Pourkheirandish et al. 2015). Archaeological findings of the earliest non-brittle spikes at the early Neolithic sites in the Southern Levant indicated Fertile Crescent as a primary origin of domestic barley (Pankin and von Korff 2017; Tanno and Willcox 2012). Genetic studies further supported Fertile Crescent, especially the Levant region, as the primary center of barley domestication, marked by the occurrence of *btr1* and *btr2* deletions in this area (Pourkheirandish et al. 2015) and the highest genetic similarity between the wild population from the region and modern cultivated

barley (Badr et al. 2000; Morrell and Clegg 2007). Additionally, other genetic studies hypothesized a secondary domestication center at the east of the Zagros Mountains in Iran (Morrell and Clegg 2007; Pankin et al. 2018; Pankin and von Korff 2017), with other possible centers in Morocco (Molina-Cano et al. 1999), the Horn of Africa (Orabi et al. 2007), and Tibet plateau (Dai et al. 2012), suggesting multiple instances of barley domestication. Genome-wide analyses further revealed the mosaic ancestry patterns across genome, implying the continuous introgressions from nearby wild barley populations into landraces across broad geographical ranges (Pankin et al. 2018; Poets et al. 2015). Overall, these findings highlight a complex demographic history for cultivated barley, likely involving recurrent admixture from multiple wild progenitor lineages that contributed adaptive alleles to landraces (Pankin and von Korff 2017; Poets et al. 2015; Russell et al. 2011).

While crop domestication and modern breeding have improved traits favorable for human beings, genetic variation has also been significantly eroded because of allele fixation. Linkage between selected loci and surrounding genomic regions further reduces diversity across the genome, which could make cultivated barley vulnerable to biotic and abiotic stresses (Bohra et al. 2022; Yahiaoui et al. 2014). Molecular evidence has demonstrated a substantial reduction in genetic diversity within cultivated barley compared to its wild progenitors, indicating a domestication bottleneck driven by human selection (Caldwell et al. 2006; Kilian et al. 2006). This genetic bottleneck limits potential crop improvement, particularly for enhancing resilience to environmental stresses, as breeding success depends on the genetic diversity accessible in the gene pool. To improve food security under the growing threat of climate change, enhancing crop resilience by broadening the genetic base of breeding populations is essential. In barley breeding, leveraging the genetic diversity present in wild relatives and landraces, presumably adapting to diverse environments through long-term natural selection, can enhance the resilience of cultivated barley (Dawson et al. 2015; McCouch et al. 2013; Yahiaoui et al. 2014).

1.2 Genetic resources for improving barley resilience

Wild relatives and landraces have not been extensively used in breeding programs despite being acknowledged as sources of valuable alleles for tolerance to abiotic stress, disease resistance, and adaptation to low-input agricultural environments (Bohra et al. 2022; Dwivedi et al. 2016; Kumar et al. 2020). A major hurdle for the exploitation of the adaptive genetic variation in exotic germplasm is the fear of undesired alleles (Kumar et al. 2020). During crossing, undesirable alleles, known as linkage drag, can be inadvertently transferred alongside the favorable genetic variation due to their proximity on the chromosome, consequently undermining the breeding progress in elite materials. To harness genetic diversity while managing the risk of linkage drag, it is imperative to identify and introgress beneficial alleles into elite genetic backgrounds, thereby creating genotypes that are immediately applicable in breeding programs (Schmidt et al. 2023). This attempt requires a comprehensive understanding of the loci of interest and a rigorous pre-breeding process.

Traditionally, researchers consider three *Hordeum* species, *H. vulgare* L. ssp. *vulgare*, *H. vulgare* L. ssp. *spontaneum*, and *H. bulbosum*, as the main genetic resources for barley breeding (Pourkheirandish and Komatsuda 2007; Wendler et al. 2014). The primary gene pool of barley consists of both cultivated barley (*H. vulgare* L. ssp. *vulgare*) and its wild relative (*H. vulgare* L. ssp. *spontaneum*), while *H. bulbosum* forms the secondary gene pool exhibiting reproductive barriers when crossed with cultivated barley (Ruge-Wehling and Wehling 2014; Wendler et al. 2014). Wild barley, as the progenitor of cultivated barley, is capable of producing fully fertile hybrids with cultivated barley (Asfaw and Bothmer 1990). Its cross-compatibility facilitates the transfer of genetic variants to domesticated barley background through backcrossing, enabling the characterization of exotic alleles and allele mining for various traits, including agronomic traits (Nice et al. 2017, 2016; Pham et al. 2024; Von Korff et al. 2004), stress tolerance (Kalladan et al. 2013; Lakew et al. 2013; Zhu et al. 2023), and disease resistance (Jost et al. 2020; Pan et al. 2021; Yuan et al. 2024). Landraces, on the other

hand, have also been used as donors of valuable genetic variation to elite lines (Bouhlal et al. 2022; Monteagudo et al. 2019; Verma et al. 2021), but are more favorable than wild barley because of less detrimental effects on yield-related traits (Monteagudo et al. 2019).

1.3 Evolutionary forces shaping patterns of genetic variation

Wild barley and barley landrace accessions are valuable breeding materials for improving the resilience of elite barley lines to biotic and abiotic stress, because of their broader genetic variation presumably adapting to diverse habitats (Kumar et al. 2020; Monteagudo et al. 2019; Wendler et al. 2014; Yuan et al. 2024). This genetic diversity in wild barley and landrace populations is shaped by the interplay of various evolutionary forces. Natural selection is a fundamental driving force in the evolutionary process of plant populations, particularly given the plant's sessile nature. The theory of natural selection, first articulated in Charles Darwin's work, *On the Origin of Species* (Darwin 1859), laid the foundation for evolutionary genetics. The natural selection theory predicts that individuals better adapted to their environment are more likely to survive and reproduce, while those less adapted have a lower chance to propagate. In the traditional selectionist view, the novel genetic variation affecting biological functions is usually deleterious and only a small fraction of mutations is beneficial to fitness. Through successive generations, natural selection purges deleterious alleles while increasing the frequency of beneficial ones, leading to local adaptation, a process that populations result in higher fitness in specific environments (Rellstab et al. 2015). For example, plants in dry lands may evolve deeper root systems to increase the chance of accessing water from deep soil layers (Lynch et al. 2014; Siddiqui et al. 2021; Uga et al. 2013).

Although natural selection is a driving force, local adaptation is a complex process also influenced by neutral processes, including mutation, gene flow, and genetic drift, rather than merely natural selection. Mutations are random changes in genome sequences that introduce new genetic variation to populations, serving as the fundamental source of all genetic diversity. However, not all mutations contribute to phenotypic variation. Contrary to the selectionist viewpoint, the neutral theory of Motoo Kimura contends most mutations on genome sequences are neutral, having no impact on the fitness of individuals. Moreover, it suggests genetic variation within a species is mainly the consequence of stochastic fluctuation of allele frequency within populations, namely genetic drift, instead of natural selection (Kimura 1979; Kimura et al. 1968). Tomoko Ohta extended the concept with her nearly neutral theory, asserting that a substantial fraction of mutations have very small effects on fitness rather than completely neutral, and most of those nearly neutral mutations are slightly deleterious (Ohta 1973).

Apart from selection and genetic drift, gene flow between populations also serves as a key factor driving evolution and sometimes determining a population to flourish or perish. Gene flow refers to the movement of alleles between populations in separate habitats, occurring in plants via pollen and seed dispersal, which introduces genetic variation and subsequently influences local adaptation. The early studies suggested gene flow as an inhibitor impeding local adaptation by swamping alleles that adapt to different environments (Kawecki and Ebert 2004; North et al. 2011). Limited gene flow, on the other hand, is predicted to facilitate the formation of locally adapted genotypes by promoting genetic differentiation. Meanwhile, local adaptation can result in migrants being less competitive in non-native environments, leading to isolation by environment (Kawecki and Ebert 2004). However, in contrast to the theoretical prediction, studies indicated that the strength of genetic drift determines whether gene flow facilitates local adaptation by replenishing genetic variation or hinders local adaptation through reducing genetic divergence (Alleaume-Benharira et al. 2006; Hereford 2009). Moreover, repeated gene flow could help non-native populations overcome environmental

constraints, aiding adaptation in foreign habitats (Smith et al. 2020). Together, natural selection along with neutral evolutionary mechanisms drive genetic diversity in plant populations.

1.4 Local adaptation of wild and cultivated barley

As mentioned, natural selection and non-selective evolutionary forces collectively result in genetic variation patterns in both wild barley and landraces. Understanding the genetic basis of local adaptation is essential to exploit the genetic diversity of exotic germplasm for barley breeding. Wild barley, a predominantly self-fertilizing species, is widely distributed across the Middle East (Harlan and Zohary 1966) and Tibet (Dai et al. 2012). Particularly, the wild barley in the Levant region serves as an excellent model for studying adaptive evolution because of its considerably high genetic diversity and its habitats having highly heterogeneous environments in a small geographical area, ranging from the Mediterranean climate with mild temperature and well watering to the desert climate with drought accompanying extreme temperature (Hübner et al. 2009; Volis et al. 2001). Recent genomic studies have shown that wild barley populations in the Levant exhibit significantly higher genetic diversity than those in other regions of the Fertile Crescent (Jakob et al. 2014; Pankin et al. 2018; Russell et al. 2016). Additionally, wild barley populations in the Southern Levant show population genetic structure and genetic differentiation associated with the environments of their habitats (Hübner et al. 2012, 2009). Using various genetic markers, previous studies have classified wild barley populations in the Southern Levant into two distinct genetic clusters associated with the latitude and precipitation gradient (Hübner et al. 2012, 2009; Jakob et al. 2014; Russell et al. 2016). Besides genetic markers, phenotypic analyses have supported the hypothesis of local adaptation in wild barley. In common garden experiments, wild barley accessions from the Southern Levant clustered into three ecotypes based on morphological traits, aligning with the environments in their

geographical origins and genetic structure (Hübner et al. 2013). These reproductive and vegetative traits were found significantly correlated with the environmental gradients, suggesting that local adaptation shapes morphological and fitness-related traits in response to distinct habitats(Hübner et al. 2013). Transplantation experiments further revealed that fitness varied among wild barley samples depending on their origin and the environment in which they were grown, indicating adaptive divergence driven by local selection pressures (Volis 2011; Volis et al. 2002a). Beyond broad-scale environmental gradients, fine-scale environmental differences may also contribute to genetic diversification in wild barley populations (Bedada et al. 2014; Nevo et al. 2005). Overall, these findings demonstrate the strong influence of local environmental selection on the genetic differentiation of wild barley in the Levant.

Alongside wild barley, local adaptation has also shaped the genetic variation of cultivated barley since natural selection acts as a pervasive evolutionary force in both wild and agricultural environments. Local adaptation of barley landraces is evident in the phenotype and genotype of accessions originating from diverse geography (Reviewed by Kumar et al. 2020). For example, Russell et al. (2016) disclosed the significant correlation of flowering and plant height with seasonal temperature and dryness variables of geographical origins in the common garden experiments. Their genomic analysis further indicated the variation of flowering gene haplotypes having a pivotal contribution to the adaptation to a wide spectrum of eco-geography (Russell et al. 2016). Moreover, a landmark study of a century-scale common garden experiment, initiated in 1929 in Davis, California, demonstrated the role of local adaptation in cultivated barley evolution (Landis et al. 2024). This experiment involved a barley landrace population founded by 28 varieties selected from diverse ecological and geographical origins, which were propagated across generations without human selection. By sequencing samples across multiple generations, Landis et al. (2024) indicated that natural selection, accompanied by linked selection, dominated the evolutionary trajectory, leading to the homogenization of the entire population, with more than half of the individuals descending from a single lineage after fifty generations. In sum, natural

selection is a pervasive evolutionary force affecting both wild barley and barley landraces, resulting in substantial genetic variation that can be useful in breeding (Bohra et al. 2022; Dwivedi et al. 2016; Kumar et al. 2020). However, the specific genomic regions involved in local adaptation of wild barley and landrace, as well as the relative contributions of environmental selection and non-selective forces to genetic variation, remain largely unclear.

1.5 Tools for harnessing adaptive alleles

To exploit the adaptive alleles in wild barley and landraces, comprehensive studies on loci controlling adaptive traits and the introgression of these alleles into elite genetic backgrounds are essential (Bohra et al. 2022; Schmidt et al. 2023). One of the strategies to identify adaptive loci is genome-wide association study (GWAS), which tests the associations between traits and genetic markers across the entire genome. This method enables researchers to directly pinpoint loci associated with traits of interest and has exhibited remarkable power in identifying valuable loci in barley genebank collection (Milner et al. 2019). However, although GWAS is a powerful tool to uncover associations between phenotype and genetic variants, it may not directly connect the identified loci to natural selection and specific environmental factors imposing selection.

An alternative to reveal adaptive loci is searching the genomic regions with the signatures of natural selection (Rellstab et al. 2015). Selective sweeps is a process of reducing genetic diversity at neutral loci linked to the selected loci when positive selection enhances the allele frequency of beneficial mutations (Stephan 2019). Methods for detecting selective sweeps include identifying genomic patterns like non-random allele association (so-called linkage disequilibrium; LD), reduced nucleotide diversity, and shifts in allele frequency distribution (Reviewed by Panigrahi et al. 2023; Pavlidis and Alachiotis 2017). These approaches often detect genomic regions deviating from

the expected neutrality inferred based on the neutral theory. However, the sequencing data treatment, such as alignment and variant calling, tends to skew allele frequency distribution and bias the population genetic inferences considerably for low-coverage sequencing data (Han et al. 2014).

Population-genetics-based genome scans can also detect adaptive loci by evaluating genetic divergence between sub-populations at individual genomic markers. Lewontin and Krakauer (1973) proposed that outliers of F_{ST} statistics (Wright 1949) may indicate loci under natural selection since neutral evolutionary forces should uniformly affect all loci, resulting in similar levels of allele frequency differentiation across a genome. In this regard, unusually high F_{ST} values may signal divergent selection, while low values could indicate balancing selection in certain genomic regions (Lotterhos and Whitlock 2014). The F_{ST} outlier approach has been further improved, like incorporating the covariance between populations to enhance the power of the detection (Günther and Coop 2013) and accounting for the sampling errors and non-independence among populations (Whitlock and Lotterhos 2015) to control the false positives. F_{ST}-based methods have been applied in barley to identify adaptive loci, such as in a study by Comadran et al. (2012), which used the genome scan of genetic divergence to reveal strong signals at loci involved in flowering time and verified the detected gene, CENTRORADIALIS (*HvCEN*), via the phenotypes of mutant lines. However, despite the effectiveness, F_{ST}-based approaches have some limitations, such as individuals have to be grouped into sub-populations beforehand (Luu et al. 2017).

Most importantly, F_{ST}-based genome scans cannot directly link the genetic signals to the environmental factors involved in the local adaptation. This limitation can be addressed by landscape genomics, an emerging discipline studying adaptive evolution through discovering the relationship between environmental factors and genomic variants (Rellstab et al. 2015). Adaptive loci are revealed by examining the statistical significance of genotype-environment associations (GEAs), associations between allele frequency and the environmental gradients at the sample's origin habitats (Lasky

et al. 2023). GEA approaches either test the effects of environmental gradients on allele frequencies (Caye et al. 2019; Coop et al. 2010; De Villemereuil and Gaggiotti 2015; Frichot et al. 2013; Günther and Coop 2013; Joost et al. 2007) or apply linear models traditionally used in phenotypic GWAS to examine genotype-environment associations (Agha et al. 2024; Lasky et al. 2015; Li et al. 2019; Sharma et al. 2021).

Conventionally, GEA approaches test the association of each genomic marker with individual environmental variables. However, since natural selection likely involves multiple environmental factors simultaneously, univariate approaches may be insufficient (Forester et al. 2018; Lasky et al. 2012). To address this shortcoming of univariate GEA, redundancy analysis (RDA) has been employed as a multivariate GEA approach. RDA is a multivariate technique used to study the linear relationship between two multivariate data sets (Legendre and Legendre 2012 Section 11.1). It has been adapted to landscape genomics, enabling statistical tests for the association between allele frequency and multiple environmental factors (Capblancq et al. 2018; Forester et al. 2018; Lasky et al. 2012). Moreover, the RDA framework allows researchers to quantify the genetic variation attributed to specific environmental factors and also to spatial auto-correlation caused by isolation-by-distance (Capblancq et al. 2018; Lasky et al. 2012, 2015).

1.6 Objectives

Barley is one of the major cereal crops and a valuable model for studying crop adaptation to climate change (Dawson et al. 2015; Marok et al. 2021). To effectively harness genetic diversity in wild barley and landrace, a comprehensive understanding of the genetic basis of adaptation and its interaction with environmental factors is essential (Bohra et al. 2022). Advances in sequencing technologies have made high-throughput genotyping increasingly affordable in recent years, enabling more extensive genotyping of germplasm in genebanks. The increase of genotyping efforts for barley ex-situ

collection provides a great opportunity to unveil the adaptive genetic variation that could support the breeding of resilient cultivated barley (Bernád et al. 2024; Milner et al. 2019; Russell et al. 2016). By integrating genomic data with eco-geographical information, it is possible to identify genes participating in the adaptive evolution and environmental factors driving selection in plant population (Bohra et al. 2022). The primary goal of this thesis was to bridge the knowledge of using landscape genomics as a tool for the discovery of adaptive loci to facilitate barley pre-breeding. The particular objectives of this thesis were:

1. Characterize the contributions of environmental gradients and spatial autocorrelation to genetic variation, and further identify potential adaptive loci using the GEA approaches in the wild barley population of the Southern Levant.

2. Extend the GEA approaches to barley accessions collected from a broader eco-geographical system, including Eurasia continent and North Africa, in order to reveal potential adaptive loci selected across diverse environmental conditions.

3. Explore and assess the strategy of integrating GEA and missing environmental information imputed through modeling the association patterns of accessible geographical origins and genetic variants.

Chapter 2

Physical geography, isolation by distance and environmental variables shape genomic variation of wild barley (*Hordeum vulgare* L. ssp. *spontaneum*) in the Southern Levant

This chapter is published as:
Chang, C.W., Fridman, E., Mascher, M. Himmelbach, A., and Schimd, K., Physical geography, isolation by distance and environmental variables shape genomic variation of wild barley (Hordeum vulgare L. ssp. spontaneum) in the Southern Levant. Heredity 128, 107–119 (2022). https://doi.org/10.1038/s41437-021-00494-x

2.1 Abstract

Determining the extent of genetic variation that reflects local adaptation in crop wild relatives is of interest for the purpose of identifying useful genetic diversity for plant breeding. We investigated the association of genomic variation with geographical and environmental factors in wild barley (*Hordeum vulgare L. ssp. spontaneum*) populations of the Southern Levant using genotyping-by-sequencing (GBS) of 244 accessions in the Barley1K+ collection. The inference of population structure resulted in four genetic clusters that corresponded to eco-geographical habitats and a significant association between lower gene flow rates and geographical barriers, e.g. the Judaean Mountains and the Sea of Galilee. Redundancy analysis (RDA) revealed that spatial autocorrelation explained 45% and environmental variables explained 15% of total genomic variation. Only 4.5% of genomic variation was solely attributed to environmental variation if the component confounded with spatial autocorrelation was excluded. A synthetic environmental variable combining latitude, solar radiation, and accumulated precipitation explained the highest proportion of genomic variation (3.9%). When conditioned on population structure, soil water capacity was the most important environmental variable explaining 1.18% of genomic variation. Genome scans with outlier analysis and genome-environment association studies were conducted to identify adaptation signatures. RDA and outlier methods jointly detected selection signatures in the pericentromeric regions, which have reduced recombination, of the chromosomes 3H, 4H, and 5H. However, selection signatures mostly disappeared after correction for population structure. In conclusion, adaptation to the highly diverse environments of the Southern Levant over short geographical ranges had a limited effect on the genomic diversity of wild barley. This highlighted the importance of non-selective forces in genetic differentiation.

2.2 Introduction

Local adaptation is an essential survival strategy for plants in stressful environments because they are sessile. Natural selection in heterogeneous environments leads to increased fitness of local genotypes. Gene flow can, however, offset genetic differentiation resulting from local adaptation and reduce fitness (Kawecki and Ebert 2004). In addition, genetic drift and demographic history contribute to genetic differentiation and confound adaptive variation with neutral variation (Günther and Coop 2013; Kawecki and Ebert 2004; López-Goldar and Agrawal 2021). Consequently, the combination of selective and nonselective forces simultaneously shapes genetic variation and leads to geographical patterns of population divergence and allele frequency distribution. Determining how different population genetic processes affect the geographical distribution of genetic variation is a key component in the study of plant adaptation. Investigating the role of adaptive and non-adaptive processes in genomic variation is of particular interest for wild relatives of crop plants, as this may allow the discovery of useful genetic variation for plant breeding (Turner-Hissong et al. 2020).

Wild barley (*Hordeum vulgare L. ssp. spontaneum*) is a highly suitable model species for studying the local adaptation of crop wild relatives, as it occurs over a wide geographical range in the Fertile Crescent and Central Asia (Harlan and Zohary 1966). Within this range, genotypes originating in Central Asia are genetically clustered with those from the eastern Fertile Crescent (Jakob et al. 2014; Pankin et al. 2018; Russell et al. 2016), There is a tendency of increasing genetic diversity from the east toward the west (Jakob et al. 2014). Wild barley in the western Fertile Crescent, i.e., the Levant, has the highest genetic diversity of the Fertile Crescent (Jakob et al. 2014; Pankin et al. 2018; Russell et al. 2016). It occupies heterogeneous environments, including Mediterranean and desert climates, within a short geographical distance (Hübner et al. 2009; Nevo et al. 1979; Volis et al. 2001). Wild barley populations from the Southern Levant show a strong correlation between genetic and environmental distances (Hübner et al. 2009). Population structure reflects eco-geographical habitats (Hübner et al.

2009, 2012) and distinguishes between northern and southern genetic clusters correlated with latitude and precipitation gradients (Jakob et al. 2014; Russell et al. 2016). Common garden experiments in previous studies revealed that eco-geography was correlated with morphological traits (Hübner et al. 2013), phenotypic plasticity (Galkin et al. 2018), and rhizosphere microbiota (Terrazas et al. 2020). Moreover, transplantation experiments showed a correlation between the geographical origin of wild barley ecotypes and fitness in different environments, suggesting local adaptation (Volis et al. 2002a,b, 2011). In addition to a broad geographical scale, environmental differences on a fine geographical scale also contribute to genetic diversification in wild barley (Bedada et al. 2014; Nevo et al. 2005; Wang et al. 2018). Overall, these results suggest a strong relationship between environmental differences, genetic divergence and phenotypic diversity of wild barley populations. This supports the hypothesis of local adaptation of wild barley in the Southern Levant. However, the relative contributions of environmental and non-selective forces to genetic variation and the genetic architecture of adaptive traits are still mostly unclear due to the lack of appropriate statistical approaches, fine-scale environmental data and sufficient genome-wide markers.

Wild barley is a valuable genetic resource for barley breeding because domestication and modern breeding have greatly reduced the genetic diversity of cultivated barley (*Hordeum vulgare L. ssp. vulgare*; Caldwell et al. 2006; Kilian et al. 2006). Since wild barley has no reproductive barrier to cultivated barley (Nevo et al. 1979), the genetic diversity of cultivated barley can be enhanced by introducing alleles from wild populations (Dawson et al. 2015). Numerous studies have shown evidence for local adaptation of wild barley to different environments (Bedada et al. 2014; Galkin et al. 2018; Hübner et al. 2013; Nevo et al. 1979; Volis et al. 2002a,b, 2004, 2011;Wang et al. 2018). Wild barley is, therefore, expected to possess considerable genetic variation that contributes to adaptation to various abiotic stresses (Dawson et al. 2015). Correspondingly, wild barley has been used as a source of novel alleles to improve stress tolerance in barley breeding (Baum et al. 2003; Pham et al. 2019). However, widespread use of wild barley has been limited due to its large genome size ($\sim$5.3

Gb) and undesirable traits (Schmid et al. 2018). To facilitate the utilization of favorable alleles in wild barley, it is important to take advantage of novel genomic technologies and eco-geographical information. Insight into the association between genetic variation and environments provides information that can help to guide the identification of valuable germplasms and the selection of core accessions for generating introgression lines and carrying out further genome-trait association studies (Bohra et al. 2022).

Genotyping-by-sequencing (GBS; Elshire et al. 2011; Poland et al. 2012) and a high-quality barley genome assembly (Jayakodi et al. 2020; Mascher et al. 2017) permit the exploration of genomic variation under environmental selection and the search for useful genetic variation in wild barley. In this study, we investigated genetic variation with high-density genome-wide markers that had not been used in earlier studies of wild barley in the Levant. We aimed to (1) describe the population structure of wild barley from the Southern Levant and place it in the context of a worldwide sample (Milner et al. 2019), (2) examine geographical patterns of gene flow in the Southern Levant, (3) characterize the relative contributions of environmental gradients and space to genomic variation and population structure, and (4) identify putative adaptive loci. Overall, our results indicated that geography and spatial autocorrelation were more important than selection for local adaptation in shaping genomic variation in wild barley in the Southern Levant. However, diverse environments, particularly water availability, show significant associations with genetic differentiation.

2.3 Materials and Methods

2.3.1 Plant material and genotyping by sequencing

We genotyped 244 wild barley accessions collected in the Southern Levant region (Fig. 2.1 A). These accessions, hereafter referred to as B1K+ accessions, included 191 accessions from Barley 1K (B1K) collection (Hübner et al. 2009) and 53 acces-

sions from an unpublished collection referred to as HOH, collected in 2005, 2009, and 2011 by K.S. (Fig. A.1; Appendix A File S1). The GBS library was constructed using genomic DNA digested with the restriction enzyme *PstI* and *MspI* as in Milner et al. (2019). In addition, published GBS data of 1,121 wild barley accessions from the IPK genebank (Milner et al. 2019) were included (Fig. 2.1 A). Because the IPK genebank contains accessions from Israel, we specified the source between IPK and B1K+ accessions to avoid confusion. Identification of single-nucleotide polymorphism (SNP) was performed as Milner et al. (2019). The detailed workflow of genotypic data filtration is described in Appendix A and summarized in Fig. A.2.

2.3.2 Environmental data

To investigate the relationship between genetic variation and environmental gradients, we used environmental data including (1) climate data from the *WorldClim2* database (Fick and Hijmans 2017) with a resolution of 30 arcseconds [$\sim$1 km], (2) soil data from the *SoilGrids* database (Hengl et al. 2017) with a resolution of 250m, (3) topographic variables based on elevation data from the *SRTM* database (`https://srtm.csi.cgiar.org/`) with a resolution of 90m, and (4) geographical coordinates of collection points (Appendix A File S1). To mitigate the problem of collinearity for redundancy analysis (RDA; Legendre and Legendre 2012), highly correlated environmental variables were grouped by hierarchical clustering using a customized clustering index (Appendix A; Fig. A.3). We then selected the eigenvector of the first principal axis for each collinear group as a synthetic variable to represent highly correlated environmental variables. Next, all environmental variables, including the synthetic variables, were selected based on variance inflation factors (VIFs) until all VIFs were less than 5. Details of this procedure are given in Appendix A. Finally, 12 environmental variables, including 7 synthetic and 5 nonsynthetic variables (Table A.1 and A.2), were selected for environmental association analyses.

2.3.3 Inference of population structure

The number of ancestors and ancestry coefficients were estimated using the model-based method *ALStructure* (Cabreros and Storey 2019). *ALStructure* uses a likelihood-free algorithm to derive estimates from minimal model assumptions. It is generally superior to existing likelihood-based methods in terms of accuracy and computational speed. The method does not assume Hardy-Weinberg equilibrium within populations, but defines the number of ancestral populations (*K*) as the rank of a matrix consisting of individual-specific allele frequencies (Leek 2011). Optimal *K* was calculated using the *estimate_d* function of the R package *alstructure* (Cabreros and Storey 2019), and ancestry coefficients were estimated using the *alstructure* function. A range of *K* values, from 2 to 8, was also used to examine the stratification of population structure. In addition to *ALStructure*, principal component analysis (PCA) and neighbor-joining (NJ) were also performed. Missing genotypic values ($\sim$3% of the dataset) were replaced by the average number of alternative alleles at each SNP locus before performing PCA.

To analyze genetic differentiation, we calculated F_{ST} and Nei genetic distance between genetic clusters defined by *ALStructure*. Accessions were assigned to genetic clusters according to the highest ancestry coefficient calculated by *ALStructure* with the optimal *K* value. F_{ST} values were calculated as ratio of average values (Bhatia et al. 2013) and Nei's genetic distances were calculated using the function *stamppNeisD* of the R package *StAMPP* (Pembleton et al. 2013).

2.3.4 Analysis of gene flow

To identify gene flow barriers that may explain observed population structure, an analysis of estimated effective migration surfaces (*EEMS*; Petkova et al. 2016) was done. First, B1K+ accessions were clustered into 58 demes that correspond to the location of collection sites. *EEMS* was then conducted in three independent runs of Markov chain

Monte Carlo (MCMC), and the results of the three runs were averaged. Each MCMC chain encompassed 10 million burn-in iterations and 10 million post-burn-in iterations thinned by an interval of 5000 iterations. Outputs of *EEMS* were processed using the R package *rEEMSplots*. To examine whether geographical barriers contribute to genetic isolation, we separated map pixels into barrier and non-barrier pixels according to geographical elevation (details in Appendix A; Fig. A.4). We then conducted a Wilcoxon test to examine the hypothesis that geographical barriers are significantly associated with lower gene flow rates.

To account for non-independence between observations, we used *ResistanceGA* (Peterman 2018) to assess the support for isolation by geographical barriers. In short, *ResistanceGA* optimizes resistance surfaces based on genetic distances and transformed landscape features with a genetic algorithm. Pairwise genetic distances between 58 demes were calculated as $D_{ps} = 1 - p$, where p is the proportion of shared alleles. In the analysis of *ResistanceGA*, we converted elevation and slope to continuous resistance surfaces, respectively, with inverse ricker and inverse-reverse monomolecular transformation. In addition, surface water data obtained from Global Surface Water (`https://global-surface-water.appspot.com/download`; Pekel et al. 2016) were used as the categorical resistance surface for *ResistanceGA*. To assess the model fit of different landscape feature combinations, we carried out two bootstrap analyses that separately used R^2 and Akaike information criterion (AIC) as the model ranking standard with 1,000 iterations. Default parameters were used for the *ResistanceGA* framework.

As a complementary method to *EEMS*, unbundled principal components (unPC; House and Hahn 2018) were employed to reveal potential long-distance migration. *unPC* scores, a ratio of PCA-based genetic distance on population level to geographical distance between demes, were computed with the R package *unPC*. Original *unPC* scores were transformed by Box-Cox transformation into an approximate Gaussian distribution. Subsequently, an outlier test based on Student's t-distribution with a two-

tailed significance level of 0.05 was performed to identify extreme population pairs. In this test, we assumed dependence between genetic and geographical distances. A null hypothesis is that samples from a pair of collection sites display an isolation-by-distance pattern. In other words, outliers identified with this test are considered to result from the violation of isolation-by-distance, which could constitute either long-distance migration or isolation due to unknown factors.

To infer asymmetric gene flows, we utilized the coalescent-based inference (CBI; Lundgren and Ralph 2019). We manually grouped accessions into ten geographical regions (Fig. A.5) such that each region covered roughly equal geographical areas as suggested by Lundgren and Ralph (2019). We likewise considered the gene flow pattern inferred by *EEMS*. The sample sizes in each region ranged from 10 to 41 with an average of 24.4. Next, we created an adjacency matrix to allow gene flow between adjacent regions (Fig. A.5). For CBI input, pairwise genetic distances were computed as the average number of different alleles across SNPs. CBI was performed using the R package *gene.flow.inference* (Lundgren and Ralph 2019) with 2 million pre-burn-in iterations, 60 million burn-in iterations, and 100 million post-burn-in iterations followed by a thinning process for every 5,000 iterations to rule out serial correlations. Medians of gene flow rates and coalescence rates were computed from posterior distributions and 95% credible intervals were calculated with the highest density interval method by using the R package *bayestestR* (Makowski et al. 2019).

2.3.5 Partitioning genomic variation

To partition genomic variation into components explained by different factors, we conducted redundancy analysis (RDA), a multivariate method for studying a linear relationship between two or more matrices (Legendre and Legendre 2012). Specifically, we used simple RDA and RDA conditioned on covariates, i.e., partial RDA, to estimate the proportion of SNP variation explained by environmental variables, spatial autocor-

relation, and population structure. RDA was performed with the *rda* function of the R package *vegan* (Oksanen et al. 2019). For all RDA models in this study, we carried out 5,000 permutations to test the significance of explanatory variables with the R function *anova.cca*.

To model the effect of spatial autocorrelation on SNP variation, distance-based Moran's eigenvector maps (dbMEMs) were used in RDA (Dray et al. 2006; Legendre and Legendre 2012). First, a network of 58 collection sites was built with the Gabriel graph, and a spatial weighting matrix of inverse geographical distances (km^{-1}) was constructed in line with the method of Forester et al. (2018). Next, the spatial weighting matrix was decomposed to generate dbMEMs. Subsequently, forward selection was performed to identify dbMEMs that associate significantly with spatial genetic structure by using *forward.sel* function (Dray et al. 2019). The selected dbMEMs with positive and negative eigenvalues, corresponding to broad-scale and fine-scale spatial structures, were both used in RDA to capture comprehensive spatial autocorrelation. In addition, to partition observed population structure, ancestry coefficients estimated by *ALStructure* with the optimal *K* values were used in the RDA on SNPs as covariates.

Since genetic clusters were largely congruent with eco-geographical habitats, we were interested in the degree of population structure that could be attributed to environmental and spatial factors. By fitting RDA models on ancestry coefficients instead of SNPs, we excluded recent genetic variation within populations to better quantify the relative contributions of environments and spatial autocorrelation to population structure. To carry out this analysis, SNPs were replaced by ancestry coefficients inferred by *ALStructure* with the optimal *K* as the new response variables in RDA models.

To evaluate the effects of individual environmental variables on SNP variation, we sequentially fitted one environmental variable at a time as the explanatory variable and treated ancestry coefficients as covariates in RDA models. Considering the correlation between environmental variables (Fig. A.3 C and D), we conducted additional permutation tests for marginal effects of environmental variables in a model includ-

ing all environmental variables by setting the parameter *by = 'margin'* for *anova.cca*. This method tested the significance of each environmental variable while removing the confounding effect with the other environmental variables.

2.3.6 Linkage disequilibrium

Linkage disequilibrium (LD) was evaluated as the pairwise r^2 of SNPs by using the *snpgdsLDMat* of the R package *SNPRelate* (Zheng et al. 2012) with a window size of 250 markers. To evaluate the genome coverage of markers, genome-wide LD decline against physical distance was fitted by using local polynomial regression and the formula of Hill and Weir (1988). Local polynomial regression was carried out by using the R function *loess* with a smoothing parameter of 0.005.

2.3.7 Identification of selection signatures

As a genome-environment association (GEA) method, RDA has high detection power and a low false-positive rate in identifying adaptation signatures (Capblancq et al. 2018; Forester et al. 2016, 2018). We therefore performed genome scans with simple and partial RDA. Simple RDA was done by treating 27,147 SNPs as response variables and twelve environmental variables as explanatory variables. To control for false positives due to population structure, partial RDA was performed by using ancestry coefficients estimated with the optimal *K* as covariates. A statistical framework proposed by Capblancq et al. (2018) was used for statistical tests and controlling for false discovery rates (FDR). Briefly, the loadings of SNPs in the first four RDA axes, selected according to the proportion of explained variation (Fig. A.6), were converted into Mahalanobis distances that approximated to a chi-squared distribution with four degrees of freedom. Next, p-values and q-values were computed accordingly, and SNPs with FDR $<$ 0.05 were considered as candidate adaptive SNPs. The statistical test was

conducted using the R function *rdadapt* (Capblancq et al. 2018).

Besides RDA, the latent factor mixed model (LFMM; Caye et al. 2019), which is an univariate GEA method, was performed by using the R package *lfmm* (Caye et al. 2019) with parameter *K = 4* to correct population structure and q-values were subsequently computed. SNPs with FDR < 0.05 were considered to be candidate adaptive SNPs.

As a complement to GEA methods, outlier SNPs with an extreme divergence between genetic clusters were detected by the X^TX statistics (Günther and Coop 2013). We assigned accessions to genetic clusters according to the highest ancestry coefficient estimated by *ALStructure* with the optimal *K* and calculated X^TX by using *BAYPASS* ver2.1 (Gautier 2015). *BAYPASS* was run by setting 25 short pilot runs, 100,000 burn-in iterations and 100,000 post-burn-in iterations with a thinning interval of 40 iterations. A significance threshold of X^TX was determined by the 99.5% quantile of pseudo-observed X^TX (Gautier 2015) calculated from neutral markers simulated by *simulate.baypass* (Gautier 2015).

2.3.8 Gene ontology enrichment

To investigate biological functions related to putatively adaptive loci, we conducted gene ontology (GO) enrichment analysis with gene annotations of the barley 'Morex v2' genome (Mascher 2019). Over-representation of GO terms for genes within 500 bp adjacent intervals of candidate SNPs was tested by Fisher's exact test with 10,000 runs using *SNP2GO* (Szkiba et al. 2014). GO terms with an FDR $<$ 0.05 were regarded as significantly enriched. Annotations of genes within 500 bp upstream and downstream of the candidate SNPs were also reported.

2.4 Results

2.4.1 Summary of genotyping data

SNP calling and preliminary filtration resulted in 101,711 SNPs for 1,365 accessions, including 1,121 IPK accessions and 244 B1K+ accessions. Depending on the analytical requirements, we selected different subsets from 101,711 SNPs as follows. For the joint population structure analysis of IPK and B1K+ accessions, we selected 4,793 SNPs with minor allele frequency (MAF) $\geq$ 0.05 among 72 IPK accessions originating from 13 countries (Russell et al. 2016). This joint dataset had a missing proportion of 0.043 and was LD-pruned with PLINK using a r^2 threshold of 0.1. For analyses of B1K+ accessions, we selected 58,616 SNPs with an overall missing proportion of 0.029 and maximal individual missing proportion of 0.059. Further filtration resulted in 19,601 SNPs (LD-pruned; MAF $\geq$ 0.01) and 27,147 SNPs (unpruned; MAF $\geq$ 0.05; Details in Appendix A; Fig. A.2)

2.4.2 Population structure and spatial genetic pattern

The inference of population structure among B1K+ accessions with *ALStructure* (Cabreros and Storey 2019) identified four clusters (Fig. 2.1 B and C) corresponding to the Mediterranean northern region, semi-arid coastal region, Judaean Desert, and Negev Desert (Fig. 2.1D). Hereafter, we named the four B1K+ clusters as North, Coast, Eastern Desert, and Southern Desert. With *K = 4*, 174 of 244 (71.3 %) accessions had a highest ancestry coefficient of less than 0.9. The first three principal components (PCs) represented the clusters corresponding to the *ALStructure* results (Fig. 2.1 B and C). On the first PC axis, the northern cluster was separated from two desert clusters, and on the second PC axis, the coastal cluster was separated from the others. On the third PC axis, the southern desert cluster was separated from the eastern desert cluster. The three PC axes explained 4.73%, 3%, and 2.83% of the variation,

Table 2.1 F_{ST} and Nei's genetic distances between four genetic clusters.

		F_{ST}			
		North	Coast	Eastern Desert	Southern Desert
Nei's Distance	North	-	0.1124	0.1149	0.2593
	Coast	0.0209	-	0.1321	0.2533
	Eastern Desert	0.0216	0.0248	-	0.2125
	Southern Desert	0.0508	0.0482	0.0389	-

respectively. A hierarchical population structure was evident in the NJ tree (Fig. 2.2A) and in the *ALStructure* analysis with K = 2-8 (Fig. 2.2B). To evaluate the importance of marker density and additional samples for population structure analysis, we performed PCA and *ALStructure* by either including or removing the HOH accessions with random selection of 100 and 5,000 SNPs, in addition to the original dataset. The dataset with 100 SNPs did not allow to identify genetic clusters while datasets with 5,000 SNPs separated into four genetic clusters by using the first four PCs even without the HOH accessions. However, *ALStructure* could only identify three ancestral populations ($K = 3$) if the HOH accessions were excluded (Appendix A File S2).

To quantify the extent of genetic differentiation among the four clusters, we computed pairwise F_{ST} and Nei's genetic distances. The Southern Desert cluster was most strongly isolated from the other three clusters (Table 2.1). With respect to the genomic pattern of differentiation, F_{ST} values were highest in the pericentromeric regions of the chromosomes 2H, 3H, 4H, 5H, and 6H. The Southern Desert cluster differentiated from the other three clusters in most of the genome except the pericentromeric regions of the chromosomes 3H and 4H (Fig. A.7).

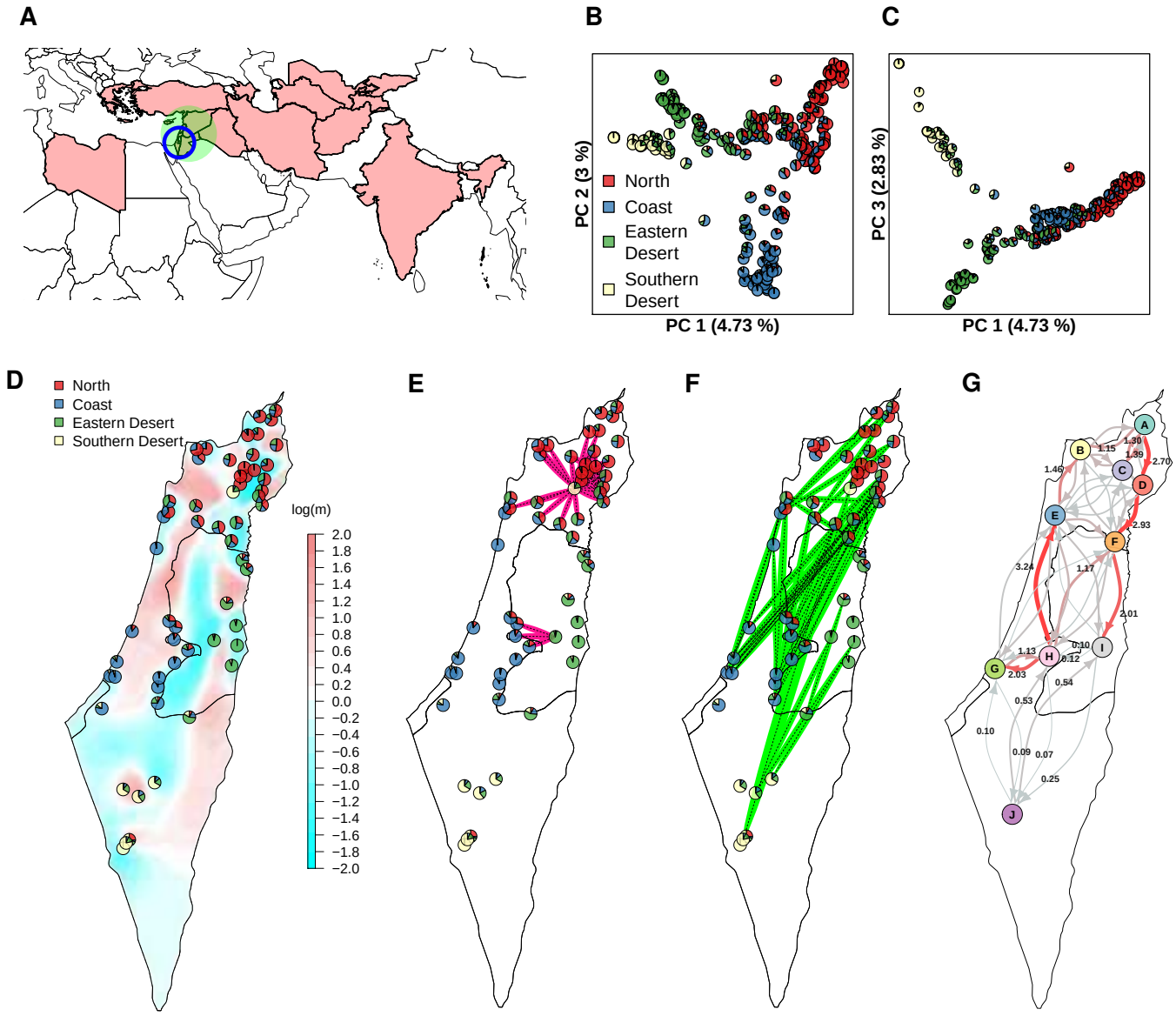

Figure 2.1 Spatial genetic structure of 244 B1K+ accessions and results of gene flow analysis. **A** A geographical map of accession origins. Countries of origin of IPK accessions are colored in light red. The green close circle represents the Levant region. The blue open circle indicates the origin of B1K+ collection. **B** PCA plot of the first and second PC axes. **C** PCA plot of the first and third PC axes. Pie charts in PCA plots represent ancestry coefficients of individuals estimated by *ALStructure* with $K = 4$. **D** Distribution of genetic clusters and effective migration surface. Pie charts give the average ancestry coefficients of individuals in collection sites. Color gradient represents gene flow rates estimated by *EEMS*. **E** Population pairs with *unPC* scores higher than the top 2.5% threshold that indicates a significantly low genetic similarity over a short geographical distance. **F** Population pairs with *unPC* scores lower than the bottom 2.5% threshold that indicates significantly high genetic similarity over a long geographical distance. **G** Gene flow rates inferred by the coalescent-based inference method, representing the probabilities per unit of time that individuals in a region i are descended from a region j (Lundgren and Ralph 2019). The thickness of arrows and the depth of red color is proportional to gene flow rates. The full results are given in Table A.4.

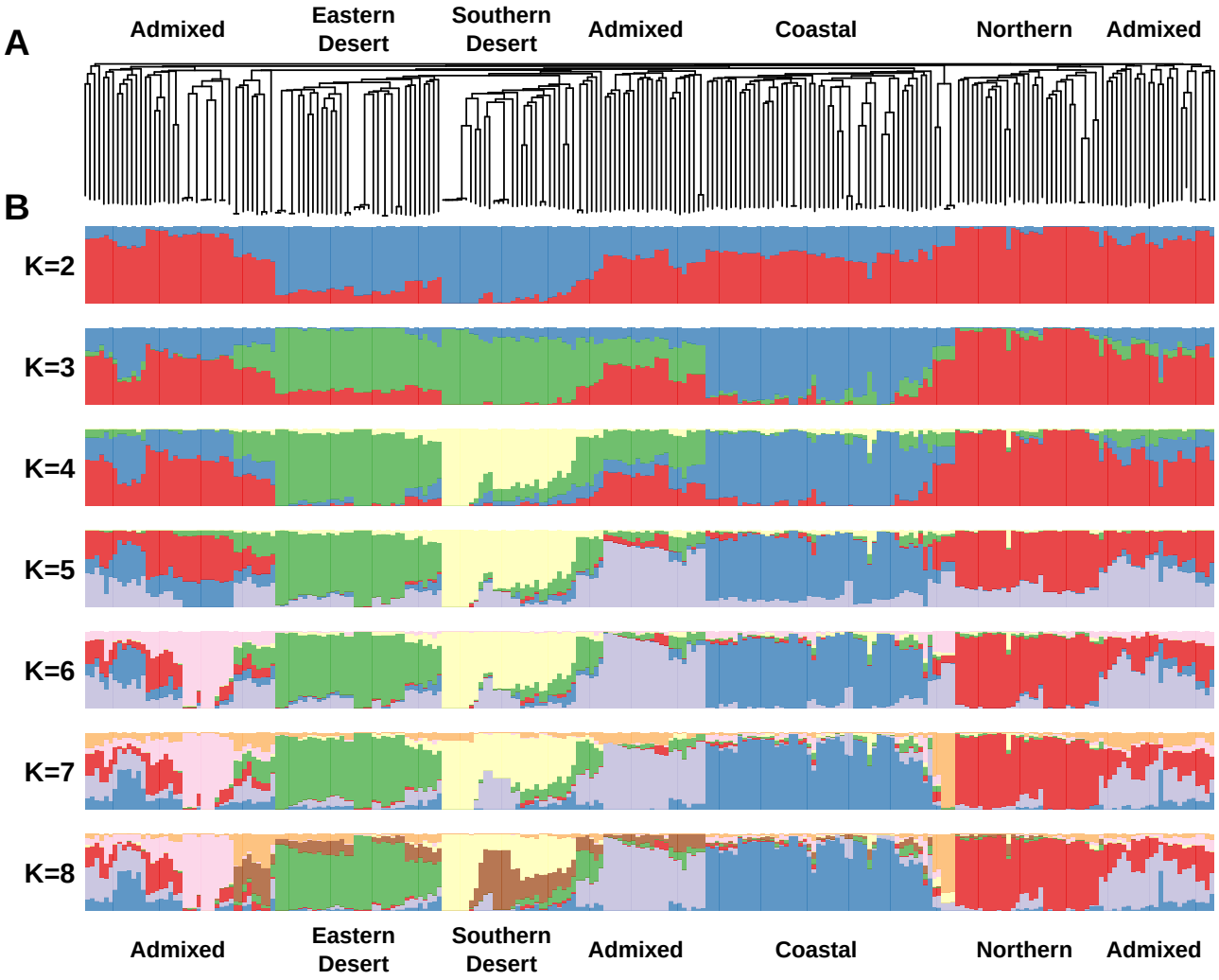

Figure 2.2 Inference of population structure. **A** Unrooted neighbor-joining (NJ) tree and **B** ancestry coefficients of 244 B1K+ accessions estimated by using *ALStructure* with *K = 2-8*. Accessions are sorted according to the NJ tree. With *K = 4*, red, blue, green, and yellow bars correspond to the Northern, Coastal, Eastern Desert and Southern Desert genetic clusters, respectively, as in Fig. 2.1.

A joint PCA of B1K+ and IPK accessions was consistent with major clusters identified in B1K+ and showed that B1K+ accessions overlapped with a large proportion of the IPK collection (Fig. 2.3A). To visualize the genetic relationship between B1K+ accessions and IPK accessions of different origins, we selected 72 geographically distinct accessions used in a previous study (Russell et al. 2016). On the first PC axis, most of the 72 geographically diverse accessions collected from western and central Asian countries were separated from B1K+ accessions but clustered more closely to the two desert clusters than to the northern and coastal clusters (Fig. 2.3A). Because

an unbalanced sample size of accessions from Israel (616 out of 1,365 accessions) might bias the PC axes, we performed another joint PCA by projecting 1,293 accessions onto PC spaces of 72 geographically distinct accessions. The PC projection was done by calculating inner products between genotypic values of 1,293 accessions and eigenvectors obtained from the PCA of 72 geographically distinct accessions. This approach could avoid the misinterpretation of sample origin and migration based on PC (McVean 2009). The PC projection showed that accessions typically clustered by geographical origin, as reported in previous studies (Milner et al. 2019; Russell et al. 2016), and B1K+ accessions were concentrated in a small area of PC space (Fig. 2.3B).

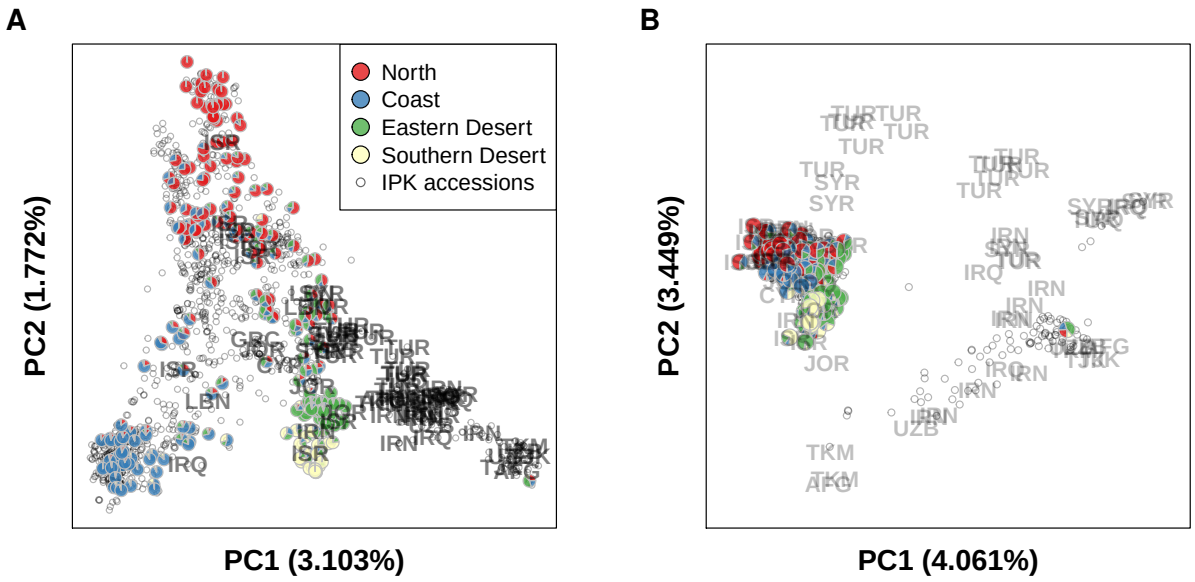

Figure 2.3 PCA plots of 244 B1K+ accessions and 1,121 accessions from the IPK genebank. **A** PCA performed with all of the available accessions. **B** PCA performed with 72 geographically diverse accessions and projection of the remaining accessions to PC spaces. Pie charts represent ancestry coefficients of 244 B1K+ accessions estimated by using *ALStructure* with $K = 4$. Gray open dots represent IPK accessions.

2.4.3 Geographical pattern of gene flow

To identify barriers limiting gene flow within the Levant region, we performed an *EEMS* (Petkova et al. 2016) analysis that revealed uneven gene flow across the landscape. The area of low gene flow rates corresponded closely to geographical barriers, includ-

ing the Sea of Galilee, the Jordan Valley, and the Judea and Samaria mountain ridges (Fig. 2.1D). A Wilcoxon test supported the association between geographical barriers and lower gene flow rates ($p < 2.2 \times 10^{-22}$; Fig. A.8 A). In addition, *EEMS* analysis showed that genetic dissimilarity between demes did not have a simple linear relationship with geographical distances (Fig. A.8 B), indicating that isolation-by-distance was not sufficient to explain genetic differentiation. This result was supported by a *ResistanceGA* analysis indicating that landscape features explained genetic distances better than geographical distances. A composite resistance surface consisting of elevation and slope was suggested as the best predictor according to R^2 in all bootstrap iterations ($\bar{R^2} = 0.51$; Table A.3.1), whereas the model of geographical distances had $\bar{R^2} = 0.21$. Model selection with AIC suggested a model with surface water as resistance surface as best model in 96.8% of bootstraps with $\bar{R^2} = 0.31$ (Table A.3.2).

The *EEMS* analysis also showed that effective genetic diversity, which is the expected genetic dissimilarity of two individuals sampled from a site (Petkova et al. 2016), decreased from north to south (Fig. A.8 C), suggesting higher genetic diversity in the north than the south. Furthermore, we performed *unPC* (House and Hahn 2018), the ratio of PC-based genetic distances to geographical distances, which is more sensitive to long-distance migration than *EEMS*. The population pairs with high *unPC* score supported regions of low gene flow identified by *EEMS* (Fig. 2.1E). This was particularly true for the majority of significant population pairs with located in the region around the Sea of Galilee in northern Israel (Fig. 2.1E). In addition, the population pairs with low *unPC* scores suggested potentially long-distance migration events in the north-south direction (Fig. 2.1F).

To evaluate asymmetric gene flows, we used CBI (Lundgren and Ralph 2019), which suggested unequal gene flows in a North-South direction (Fig. 2.1G; Table A.4). There was a trend for gene flow from South (region H) to North (region B) in the western region (region $H \rightarrow E \rightarrow B$; Fig. 2.1G) and an opposite trend from North (region A) to South (region I) in the eastern region (region $A \rightarrow D \rightarrow F \rightarrow I$; Fig. 2.1G). The

strongest gene flow (3.24 with the 95% credible interval of 0.51-6.32; Table A.4) was observed from populations close to Jerusalem (region H) to the surrounding areas of Mount Carmel (region E). However, the gene flow in the opposite direction (region $E \rightarrow H$) was much weaker (0.77 with the 95% credible interval of 0-2.32; Table A.4). Low gene flow rates of connections across geographical barriers, such as $C \rightleftharpoons D$ and $H \rightleftharpoons I$, correlated with the results of *EEMS* (Fig. 2.1D) and *unPC* (Fig. 2.1 E and F). Furthermore, low gene flow rates between the Negev desert (region J) and its adjacent regions indicated the isolation of Southern Desert accessions, consistent with the high genetic differentiation suggested by the F_{ST} values (Table 2.1).

2.4.4 Genetic variation explained by environment and space

SNP variation partitioning with redundancy analysis (RDA)

To quantify the relative contributions of the environment and space to genomic variation, we performed RDA on SNPs by taking all environmental variables as a whole and incorporating spatial autocorrelation. RDA showed that environmental variables explained 15.12% ($R^2_{adj} = 0.107$; $p = 0.0002$) of SNP variation while spatial autocorrelation captured by dbMEMs, which are eigenfunctions of a spatial network (Dray et al. 2006; Legendre and Legendre 2012), explained 44.95% ($R^2_{adj} = 0.285$; $p = 0.0002$; Fig. 2.4A). We found 10.63% of SNP variation is jointly explained by environmental variables and spatial autocorrelation, and 4.49% ($R^2_{adj} = 0.013$; $p = 0.0002$) was solely explained by environmental variables (Fig. 2.4A). Considering the confounding effect between environment and population structure, we treated ancestry coefficients (*K = 4*) as covariates in partial RDA when examining the effect of environmental variables on SNP variation. The partial RDA indicated that population structure explained 15.43% ($R^2_{adj} = 0.148$; $p = 0.0002$) of SNP variation, and environmental variables solely explained 8.71% ($R^2_{adj} = 0.048$; $p = 0.0002$) of SNP variation when conditioned on population structure (Fig. 2.4A).

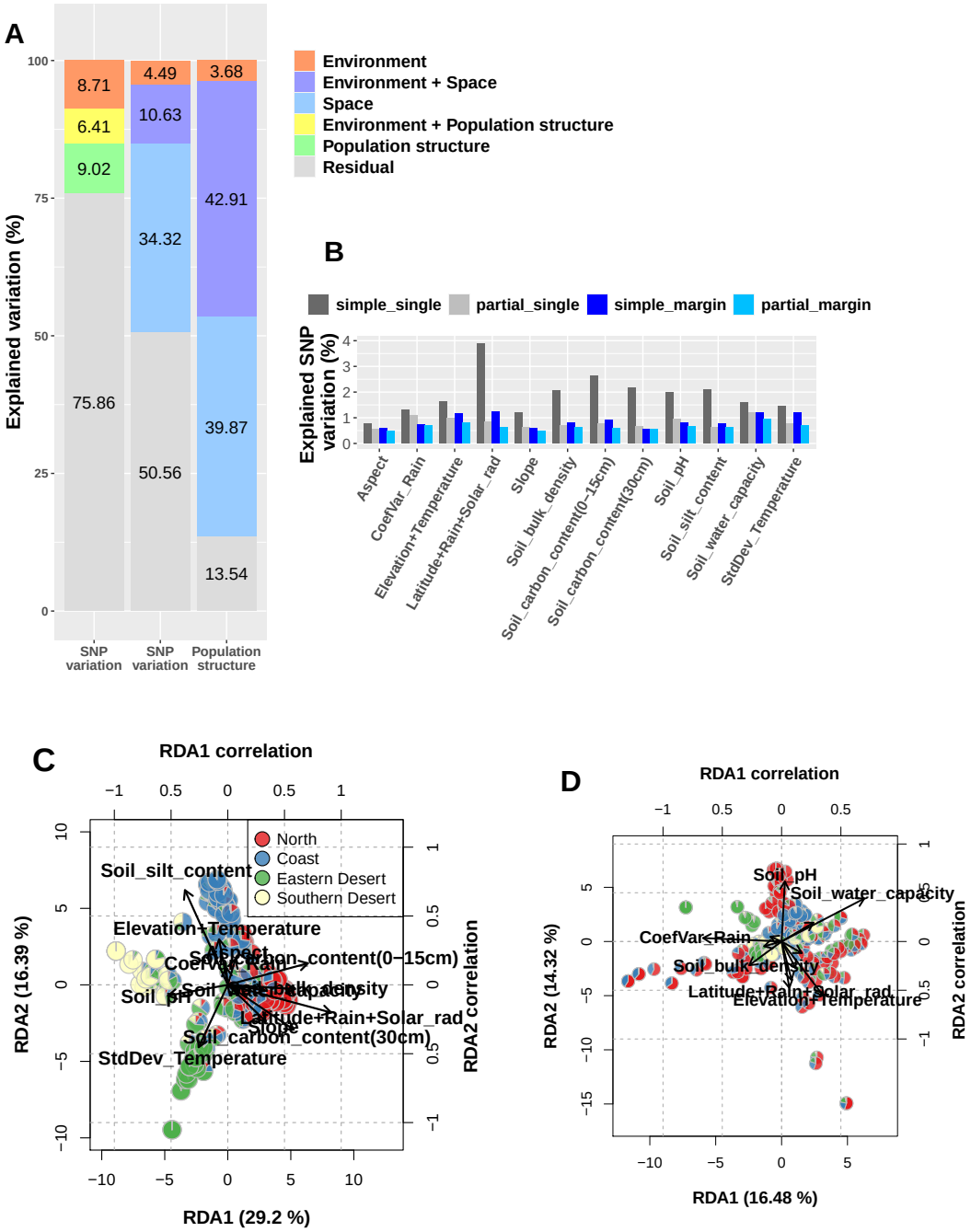

Figure 2.4 Results of variation partitioning and RDA biplots. **A** Variation partitioning of SNP variation and population structure. Left and middle columns: explained SNP variation estimated by RDA models using population structure and spatial autocorrelation as covariates, respectively. Right column: population structure explained by environment and space. Environment, space and population structure are represented by twelve environmental variables, dbMEMs and ancestry coefficients ($K = 4$) in RDA models. **B** Percentage of SNP variation explained by environmental variables. The *simple_single* and *partial_single* show individual effects estimated based on RDA models fitting one environmental variable at a time. The *simple_margin* and *partial_margin* show marginal effects estimated based on RDA models fitting all environmental variables. The *partial_single* and *partial_margin* are estimated based on partial RDA conditioned on population structure. **C** Biplot of simple RDA. **D** Biplot of partial RDA conditioned on population structure. The arrows represent correlations of the environmental variables with RDA axes that are shown in greater detail in Table A.6. Abbreviation in the biplots: Asp Aspect, CVR CoefVar_Rain, ET Elevation+Temperature, LRS Latitude+Rain+Solar_rad, Slp Slope, SBD Soil_bulk_density, SCC1 Soil_carbon_content(0-15cm), SCC2 Soil_carbon_content(30cm), SpH Soil_pH, SSC Soil_silt_content, SWC Soil_water_capacity, SDT StdDev_Temperature

Relative importance of individual environmental variables for SNP variation

After confirming an association between genomic variation and the environment, we further investigated the effects of individual environmental variables. In simple RDA models with separate fitting of each environmental variable, permutation tests showed that all of the 12 environmental variables were significantly associated with SNP variation ($p < 0.005$; Table A.5.1). Without constraining on population structure, the synthetic variable '*Latitude+Rain+Solar_rad*' (Table A.1 and A.2) explained the highest proportion of SNP variation (3.89%; Fig. 2.4 B; Table A.5.1). In contrast, in partial RDA models conditioned on population structure, '*Soil_water_capacity*' explained the highest proportion of SNP variation (1.18%; Fig. 2.4 B; Table A.5.2) whereas the proportion of SNP variation explained by '*Latitude+Rain+Solar_rad*' reduced to 0.86%. The variable '*Aspect*' presented the lowest but significant association with SNP variation in both simple and partial RDA conditioned on population structure (Table A.5.1 and A.5.2). The explained variation of '*Soil_water_capacity*', '*CoefVar_Rain*', which refers to coefficients of variation of precipitation in the growing season, and '*Aspect*' decreased less than other environmental variables after conditioned on population structure. This indicates that they correlated less with population structure. We also investigated marginal effects in models that incorporated all environmental variables by considering correlations between environmental variables. The variables '*Latitude+Rain+Solar_rad*' and '*Soil_water_capacity*' once again showed the highest marginal effect in the simple RDA and partial RDA conditioned on population structure, respectively (Fig. 2.4 B; Table A.5.3 and A.5.4).

RDA biplots provided further information on the relative importance of environmental gradients. The biplot of the simple RDA (Fig. 2.4C) showed a population structure consistent with the four genetic clusters identified by *ALStructure*. The first and second RDA axes corresponded to genetic differentiation in the north-south and west-east directions, respectively (Fig. 2.4C). The first RDA axis was strongly (r = 0.911; Table A.6) correlated with the variable "*Latitude+Solar_rad*" (LSR; Fig. 2.4C). When conditioned

on population structure, two water-related variables, "*Soil_water_capacity*" (SWC; r = 0.697) and "*CoefVar_Rain*" (CVR; r = -0.662), were the most influential predictors on the first RDA axis (Fig. 2.4D and Table A.6). However, if conditioned on spatial autocorrelation rather than on population structure, the effects of all environmental variables decreased significantly (Fig. A.9 and Table A.6). This indicated a strong correlation of environmental gradients with spatial autocorrelation.

Association of population structure with environment and space

With reference to our hypothesis that the diverse environments in the Southern Levant were an important factor in shaping populations, we quantified the relative contributions of environment and space to population structure ($K = 4$) with RDA on ancestry coefficients. As expected, a high proportion of population structure that was explained by environmental variables (42.91 of 46.59%; the right column of Fig. 2.4A) could not be separated from the component explained by spatial autocorrelation. Only 3.68% ($R^2_{adj} = 0.0358$; $p = 0.0002$) of population structure could be solely explained by environments whereas spatial autocorrelation accounted solely for 39.87% ($R^2_{adj} = 0.374$; $p = 0.0002$) of population structure (Fig. 2.4A). This result suggested that spatial autocorrelation had a larger effect on population differentiation of wild barley in the Southern Levant than environmental diversity.

2.4.5 Adaptive candidates and GO enrichment

The association between genomic variation and environment prompted us to perform genome scans to identify putative adaptive loci. Given the large genome size ($\sim$5.3 Gb), we first estimated LD decay to assess whether the marker density of the reduced representation data was sufficient to accurately identify adaptive genes in these scans. We fitted the *loess* model and Hill-Weir formula with 27,147 genome-wide SNPs. We then observed a rapid decay in LD because r^2 values dropped to half of the highest

values of 0.377 and 0.454 after pairwise SNP distances of 213 bp and 125 bp, respectively (Fig. A.10). Given the large size of the barley genome, this result indicates a possible difficulty in detecting the precise locations of adaptive loci, except for closely linked loci with the current marker density.

Three GEA methods, simple RDA, partial RDA, and LFMM, identified 352, 364, and 307 candidate SNPs (FDR < 0.05), respectively, and the outlier method, *BAYPASS*, identified 279 candidate SNPs (X^TX > 11.05). However, candidate SNPs detected by the four methods hardly overlapped, except simple RDA and *BAYPASS* with 125 common SNPs, 91 of which were located in pericentromeric regions of chromosomes 3H, 4H, and 5H (Fig. 2.5; Appendix A File S3). By searching 500 bp adjacent intervals of candidate SNPs, the four methods jointly identified two genes on the chromosome 4H. The first gene *HORVU.MOREX.r2.4HG0308420* locates closely to SNPs associated with the variable '*Latitude+Rain+Solar_rad*' in the LFMM analysis (Appendix A File S4) which encodes an ATP-dependent RNA helicase. The second gene *HORVU.MOREX-.r2.4HG0314300* is linked to SNPs associated with '*Elevation+Temperature*'. It encodes a nucleolar GTP-binding protein (Fig. A.11; Table A.7; Appendix A File S4). GO term enrichment analysis identified 2 and 10 enriched GO terms based on candidate SNPs detected by simple RDA and *BAYPASS*, respectively (Table A.8). No GO term was enriched based on the results of partial RDA and LFMM.

2.5 Discussion

Our study indicated that geography and spatial autocorrelation were better predictors of genomic variation than environmental gradients even though the diverse environments of the Southern Levant are expected to impose strong natural selection (Hübner et al. 2009; Nevo et al. 1979). These findings imply that genomic variation of wild barley in the Southern Levant was mainly driven by neutral processes consistent with a neutralist perspective (e.g., Volis et al. 2001, 2003, 2005). However, environmental

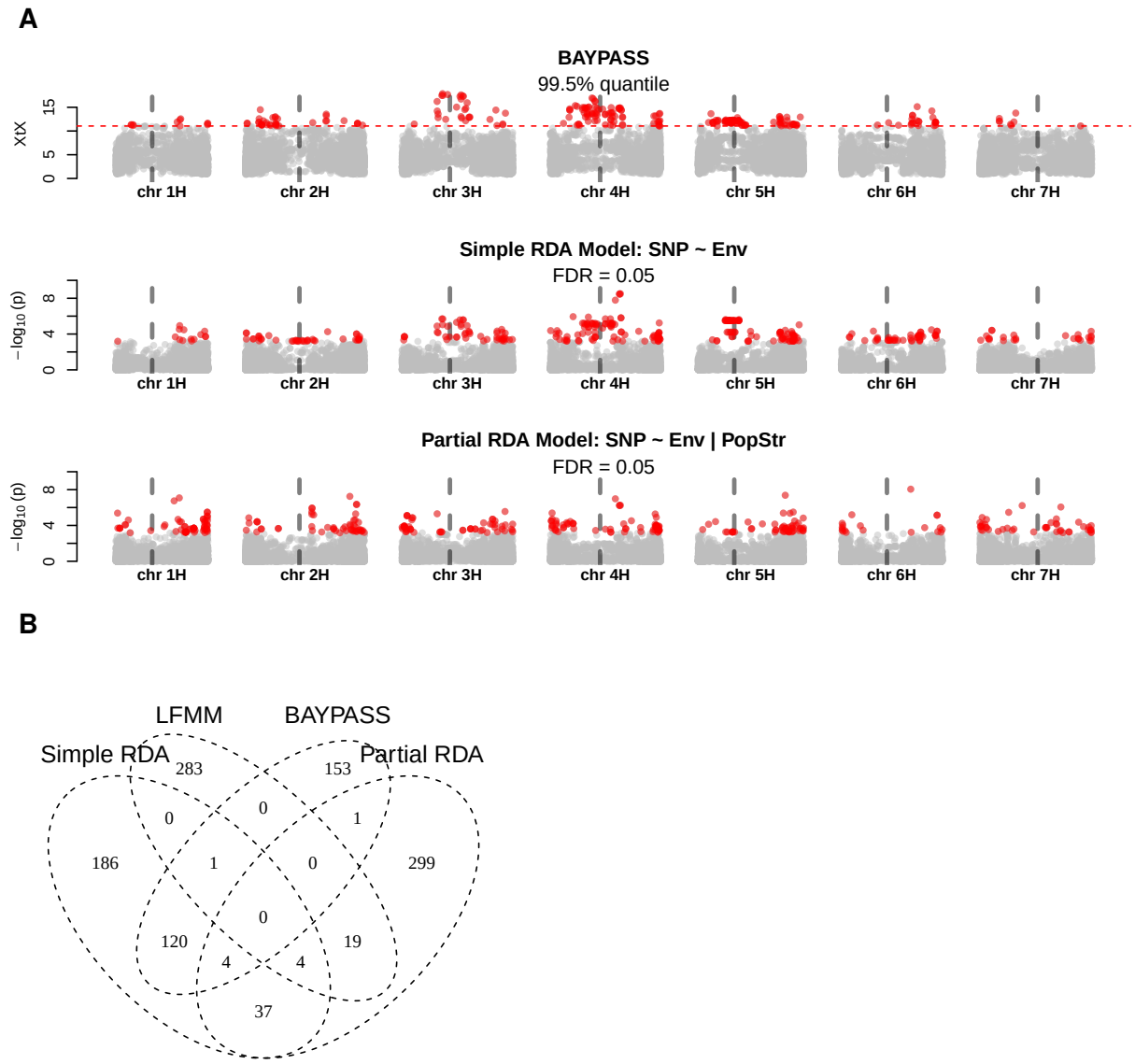

Figure 2.5 Inference of selection. **A** Genome scans for adaptation signatures. Three Manhattan plots correspond to the *BAYPASS*, simple RDA, and partial RDA conditioned on population structure. Significant SNPs are highlighted as red dots. The positions of centromeres are indicated with vertical gray dash lines. **B** Numbers of significant SNPs detected by four different methods for genome scans.

variables were still associated with a relatively small but considerable proportion of genomic variation (15.12%; Fig. 2.4A), suggesting that natural selection and hitch-hiking may have a detectable effect on the structure of genetic diversity.

2.5.1 Strong population structure of B1K+ and IPK genebank collections

Three clusters of wild barley from the Levant region that correspond to eco-geographical habitats were previously characterized using SSR markers and an SNP array developed for cultivated barley (Hübner et al. 2012) and morphological traits (Hübner et al. 2013). Our results are consistent with previous findings, except that the previously reported desert cluster (Hübner et al. 2012) was split into two clusters (Fig. 2.1 B-D), which was evident in the *ALStructure* analysis with *K = 3* and *K = 4* (Fig. 2.2). Difference to previous study resulted from an increased marker number but also to the inclusion of additional accessions collected in the Negev Desert in 2011 (Appendix A File S2). Although genetic clusters were consistent with eco-geographical habitats, caution should be exercised when interpreting the results of model-based methods. First, the number of ancestral populations might be overestimated due to isolation by distance (Bradburd et al. 2018). Second, the high proportion of admixed accessions (174 of 244 B1K+ accessions; 71.3%) might not result from admixture. Both spatial autocorrelation (Bradburd et al. 2018) and demographic history (Lawson et al. 2018) such as bottlenecks that likely occur in self-pollinating species (Hartfield et al. 2017), may lead to high admixture proportions in model-based methods.

The joint PCA incorporating the IPK wild barley collection indicated a strong effect of an unbalanced sample size of accessions from Israel (616 of 1,365 accessions) on a PCA (Fig. 2.3 A and B). This comparison highlighted the importance of balanced sampling when analyzing population structure because unequal sample sizes among groups could lead to the distortion of PCs (McVean 2009). The PCA, based on all accessions, compressed the accessions with a broad geographical origin across the whole distribution range of wild barley into a cluster (Fig. 2.3A) that did not appropriately reflect their wide geographical origin. In contrast, a PCA with a more balanced sample of accessions from the whole species range revealed that the wild barley from the Southern Levant regions contained only a small proportion of the total diversity of

wild barley (Fig. 2.3B). However, the PCA of the complete sample revealed that accessions from Greece and Cyprus clustered with accessions from the Southern Levant (Fig. 2.3 A and A.12 A). This suggested they originated from the Southern Levant or adjacent areas without a sufficiently long history of differentiation from the ancestral populations. Likewise, 579 IPK accessions of unknown origins might be closely related to the Levant region as they strongly overlapped with our B1K+ population (Fig. A.12).

2.5.2 Evidence for geographical pattern of gene flow

We expected that gene flow among wild barley populations was limited because of a low rate of outcrossing (<2%; Abdel-Ghani et al. 2004), and seed dispersal occurred mainly within 1.2 m (Volis et al. 2010). However, self-fertilizing plants can establish a population with a single seed after long-distance dispersal (Baker 1967), and the long spiky awns attached to seeds of wild barley facilitates dispersal by zoochory. Over a sufficiently long period, gene flow across landscapes might accumulate via occasional dispersal and outcrossing. *EEMS* (Petkova et al. 2016) was previously used to identify gene flow barriers in plant populations across large geographical ranges, e.g., rice (Gutaker et al. 2020) and spruce (Tsuda et al. 2016). In our data, *EEMS* revealed fine-scale patterns of gene flow attributable to geography, particularly of the Sea of Galilee and the Jordan Valley, which had not been previously identified by inferring gene flow between genetic clusters (Hübner et al. 2012). These geographical separations appear to promote genetic differentiation within a short geographical distance that interfered with isolation-by-distance patterns (Fig. A.8 B). The analysis of *ResistanceGA*, which accounted for the non-independence of samples, also suggested a stronger effect on genetic differentiation by geographical barriers than by isolation-by-distance (Table A.3).

Coalescent-based inference (CBI; Lundgren and Ralph 2019) detected trends of gene flow in opposite directions in eastern and western regions (Fig. 2.1G). This con-

tradicted the net gene flow from north to south identified by Hübner et al. (2012). The different conclusions regarding gene flow directions in western Israel were likely due to the manner in which geographical information was incorporated into the analyses. While Hübner et al. (2012) assigned accessions according to genetic clustering, our assignment emphasized geographical origin. CBI gene flow rates expressed the probability that a population descended from another population per unit time (Lundgren and Ralph 2019). Consequently, a history of recent colonization might explain gene flow trends in our data. Incorporating historical genome recombination to infer gene flow at different time periods might provide a clearer picture (Al-Asadi et al. 2019). Errors in gene flow inference could result from sampling biases, missing and erroneous genotypic values caused by low sequencing depth, and also from uneven distributions of markers due to the nature of GBS (Elshire et al. 2011; Poland et al. 2012). However, imbalanced sampling should not bias our results because *EEMS* and CBI are insensitive to unequal sample numbers (Lundgren and Ralph 2019; Petkova et al. 2016).

2.5.3 Effects of environment and geographical distance on SNP variation

RDA analysis indicated that environmental gradients explained a substantial portion of SNP variation and population structure (Fig. 2.4A). This analysis did not include all possible environmental effects because comprehensive environmental data were not available. For example, the adaptive trait *drought stress recovery* is associated with the rainfall predictability in wild barley (Galkin et al. 2018), but such data were only available for some collection sites. In addition, the control for collinearity and nonlinear environmental effects that RDA did not account for, might lead to unexplained genetic variation in our analysis.

Phenotypic studies suggested the importance of rainfall in the evolution of wild barley in the Southern Levant (Galkin et al. 2018; Hübner et al. 2013; Volis 2011;

Volis et al. 2002a,b). Our RDA analysis indicated that variables related to water availability ('*Latitude+Rain+Solar_rad*' and '*Soil_water_capacity*') were the most important drivers of genomic variation (Fig. 2.4 B-D; Table A.5). It was not possible to specify the effects of individual environmental gradients because they were highly correlated. For example, we could not separate the effect of precipitation from latitude, which is highly relevant for the timing of flowering in barley (Russell et al. 2016). Unlike other environmental variables, '*Aspect*' had few confounding effects with other gradients and population structure (Fig. 2.4B). '*Aspect*' was the strongest predictor when conditioned on spatial autocorrelation (Fig. A.9; Table A.5). In the Southern Levant, south-facing slopes might be more exposed to drought and heat than north-facing slopes due to higher solar radiation, resulting in significantly stronger selection within only a few hundred meters, referred to as the Evolution Canyon model (Bedada et al. 2014; Nevo et al. 2005). Our results suggest that '*Aspect*' might reflect a minor but pervasive effect of microclimate in the Southern Levant that could not be represented by climate data at the current resolution. In *Mimulus guttatus*, an important locus of microgeographical adaptation was successfully identified by integrating quantitative trait loci mapping and population genomic analyses (Hendrick et al. 2016). A similar approach might be used to investigate the genetic architecture of adaptation to microclimatic conditions in wild barley.

By using dbMEMs, which model the effects of spatial autocorrelation on SNP variation, our RDA revealed that high proportions of SNP variation (45%) and population structure (83%) were explained by spatial autocorrelation (Fig. 2.4A). The lower proportion of SNP variation attributed to environments (Fig. 2.4A) indicated that environmental selection might be an influential but not a dominant driver of genetic differentiation. In contrast to our findings, environment had a significantly stronger effect than geographical distance on diversity in *Boechera stricta* (Lee and Mitchell-Olds 2011). However, in *Arabidopsis thaliana* (Lasky et al. 2012), sorghum (Lasky et al. 2015), rice (Gutaker et al. 2020) and wild tomato (Gibson and Moyle 2020), the contribution of the environment was comparable and highly overlapped with geographical distance. This

suggested that isolation-by-distance was a robust and widespread pattern in a small geographical range like in our case and over a large geographical scale (Gibson and Moyle 2020; Gutaker et al. 2020; Lasky et al. 2012, 2015). Complex spatial structures confounding with environmental gradients are a pervasive challenge in the study of local adaptation (de Villemereuil et al. 2014; Excoffier et al. 2009). In particular, population genetic analyses tend to be biased by spatial structure (Battey et al. 2020b). For this reason, phenotypic studies using crosses between accessions and common garden experiments are also required to distinguish between genetic variation attributed to local adaptation and spatial autocorrelation. Additionally, we noted that a high percentage of SNP variation (51%; Fig. 2.4A) remained unexplained even after incorporating dbMEMs. This could be due to either unknown evolutionary forces that are independent of spatial autocorrelation or to the limitations of our current dataset and methodologies.

2.5.4 Lack of strong evidence to pinpoint adaptive loci

The rapid decay of LD within a few hundred base pairs (Fig. A.10) was consistent with similar studies of wild barley populations from the Middle East and Central Asia (Morrell et al. 2005). Reduced-representation sequencing approaches tend to have limited power in identifying adaptive loci, especially for genomes with high levels of recombination (Tiffin and Ross-Ibarra 2014). Rapid LD decay and a large genome size of $\sim$5.3 Gb indicate that the marker density of this study might not allow precise genome scans. To account for this caveat and to control for false-positive rates, we combined the results from multiple methods of genome scans (Forester et al. 2018; Lotterhos and Whitlock 2015; Rellstab et al. 2015). Although different methods detected different signals, we considered only overlapping signals between scans to be promising adaptive signatures because our goal was to identify stress-tolerance loci that could be useful in barley breeding. In particular, the correlation between populations raised a concern of false positives in genome scans (de Villemereuil et al. 2014) as we studied

populations from a small geographical range.

Significant polymorphisms hardly overlapped between methods (Fig. 2.5B). This observation might be explained by (1) lack of adaptive loci with large effects, (2) strong confounding effect of population structure, and (3) limitations of the dataset. Although there was no robust evidence of the identification of adaptation genes, the genome scans based on X^TX and simple RDA identified significant correlations with environmental variables and strong genetic differentiation in the pericentromeric regions of the chromosome 3H, 4H, and 5H (Fig. 2.5A). However, these associations were not observed in the partial RDA and LFMM analyses. Although the X^TX statistics accounted for the covariance of allele frequencies (i.e., population structure) among populations (Günther and Coop 2013), spurious signals of selection might arise if self-fertilization inflated false-positive values via strong genetic drift (Hodgins and Yeaman 2019). For this reason and because of a strong association between population structure and environments (Fig. 2.4A), false positives were expected for the outlier and GEA methods even with a correction for population structure. In spite of the concern about false positives, the high degree of putative selection-driven differentiation was still remarkable. Similar patterns of genetic differentiation in pericentromeric regions were reported in previous studies of barley (Contreras-Moreira et al. 2019; Fang et al. 2014; Wang et al. 2018), teosinte (Pyhäjärvi et al. 2013) and maize (Navarro et al. 2017). Theoretical studies suggested that adaptation with gene flow could result in divergent linkage groups of locally beneficial alleles in low recombination regions (Akerman and Bürger 2014; Bürger and Akerman 2011; Yeaman and Whitlock 2011). These conclusions were supported by simulation and empirical studies, e.g., in stickleback, sunflower, and *Arabidopsis lyrata* (Berner and Roesti 2017; Hämälä and Savolainen 2019; Renaut et al. 2013; Samuk et al. 2017). Low-recombination pericentromeric regions of wild barley were reported to have significantly higher ratios of non-synonymous to synonymous substitution (π_a/π_s) than other genomic regions (Baker et al. 2014). This suggested a tendency to accumulate genetic load in pericentromeric regions. Moreover, in terms of conditional neutrality, the accumulation of conditionally deleterious

mutations in habitats where they are neutral could lead to genotype-environment interactions of fitness if migration is weak relative to genetic drift (Mee and Yeaman 2019). Taken together, given weak gene flow, high rates of self-fertilization and variable recombination rates over the genome, a long-term accumulation of conditionally deleterious mutations might result in locally neutral linkage of alleles in low-recombination genomic regions. This could create a pattern of polymorphism that might resemble local adaptation and explain our observations in the pericentromeric regions.

2.5.5 Conclusion and outlook

We observed a stronger effect of non-selective factors such as geography and isolation-by-distance on total genetic diversity in the wild barley populations of the diverse and stressful environments of the Southern Levant. Nevertheless, natural selection has a small but significant influence on genomic variation. This might be potentially valuable for barley breeding because water availability, i.e., precipitation and soil water capacity, was the most strongly correlated environmental variable. Outlier test and simple RDA identified genomic regions that might contribute to local adaptation, but these regions were not robustly identified by the different tests applied. One limitation of our study was therefore that only a small proportion of the wild barley genome was sequenced by the GBS approach. This was suitable for analyzing genome-wide patterns of variation and mapping of causal genes (Milner et al. 2019), but was not powerful enough for pinpointing genomic targets of local adaptation. In the near future, whole genome sequencing of wild barley accessions (Sato et al. 2021) and the development of a barley pangenome (Jayakodi et al. 2020) will greatly increase the ability of population genomic approaches to understand wild barley adaptation and facilitate the mining of useful alleles for plant breeding. Such approaches can be combined with common garden and transplantation experiments of wild barley genotypes to measure fitness effects in different environments (Hübner et al. 2013; Volis 2011), gene expression studies of differentially adapted genotypes (Hübner et al. 2015) and

mapping populations. Given the major impact of isolation-by-distance on genomic variation, adaptive genetic variation was likely confounded with population structure. Mapping populations with sufficient genome recombination evaluated in different environments permitted the disentangling of adaptive and neutral variation, as shown in such populations developed from wild and cultivated barley (Herzig et al. 2018; Wiegmann et al. 2019). Whole genome resequencing followed by computational analysis can be rationalized to analyze a large number of genotypes such as the complete B1K population. Consequently, we believe that population genomic analysis of differentially adapted crop-wild relatives will complement other approaches to understanding plant adaptation and enable the use of this information for breeding (Bohra et al. 2022).

2.6 Acknowledgments

We are grateful to Amit Shtern for plant growth and leaf sampling, Elisabeth Kokai-Kota for DNA extraction, and Anne Fiebig for data submission. The project at Schmid lab was supported by a Gips Schüle Foundation award "Freiräume für die Forschung" and funds from the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the Federal Programme for Ecological Farming and Other Forms of Sustainable Agriculture (Project number 2818202615). The project at Fridman lab was supported by the Israeli Science Foundation (ISF) (project number 1270/17).

2.7 Author contributions

CWC and KS designed the study. KS and EF arranged and coordinated leaf sample collection. KS supervised DNA extractions and GBS library establishment. AH orga-

nized DNA sequencing. MM performed SNP calling. CWC carried out data analyses. CWC and KS wrote and revised the manuscript. KS and EF obtained funding.

2.8 Competing interests

The authors declare no competing interests.

2.9 Data archiving

The GBS data of B1K+ accessions collected in this study have been archived at the European Nucleotide Archive (ENA) with project ID PRJEB47405. The ENA sample IDs are available in Appendix A File S5. The geographical coordinates and environmental data are available in Appendix A File S1. The R code used for analysis is archived at `https://kjschmidlab.gitlab.io/b1k-gbs/`.

Chapter 3

GGoutlieR: an R package to identify and visualize unusual geo-genetic patterns of biological samples

This chapter is published as:
Chang, C.W. and Schmid, K., GGoutlieR: an R package to identify and visualize unusual geo-genetic patterns of biological samples. Journal of Open Source Software, 8(91), 5687 (2023). https://doi.org/10.21105/joss.05687

3.1 Abstract

Landscape genomics is an emerging field of research that integrates genomic and environmental information to explore the drivers of evolution. Reliable data on the geographical origin of biological samples is a prerequisite for accurate landscape genomics studies. Traditionally, researchers discover potentially questionable samples using visualization-based tools. However, such approaches cannot handle large sample sizes due to overlapping data points on a graph and can hinder reproducible research. To address this shortcoming, we developed **G**eo-**G**enetic **outlier** (*GGoutlieR*), an R package of a heuristic framework for detecting and visualizing samples with unusual geo-genetic patterns. Outliers can be identified using either geography-based K-nearest neighbors (KNNs) or genetics-based KNNs. The framework calculates empirical p-values for each sample, allowing users to easily identify outliers in data sets with thousands of samples. The package also provides a plotting function to display the geo-genetic patterns of outliers on a geographical map. GGoutlieR has the potential to significantly minimize the data cleaning required by researchers prior to conducting landscape genomics analyses.

3.2 Statement of need

Landscape genomics is a thriving field in ecological conservation and evolutionary genetics (Aguirre-Liguori et al. 2021; Lasky et al. 2023), providing insights into the links between genetic variation and environmental factors. This methodology requires reliable geographical and genomic information on biological samples. To determine whether data are reliable, researchers can examine associations between genetic similarities and the geographic origin of biological samples before proceeding with further studies. Under the assumption of isolation-by-distance, pairwise genetic similarities of samples are expected to decrease with increasing geographical distance between

the sample origins. This assumption may be violated by long-distance migration or artificial factors such as human activity or data/sample management errors.

Visualization-based tools such as *SPA* (Yang et al. 2012), *SpaceMix* (Bradburd et al. 2016), *unPC* (House and Hahn 2018) allow to identify samples with geo-genetic patterns that violate the isolation-by-distance assumption, but these tools do not provide statistics to robustly label outliers. Advances in genome sequencing technologies lead to much larger sample sizes, such as in geo-genetic analyses of genebank collections of rice (Gutaker et al. 2020; Wang et al. 2018), barley (Milner et al. 2019), wheat (Schulthess et al. 2022), soybean (Liu et al. 2020) and maize (Li et al. 2019). Visualization-based approaches may not be suitable to display unusual geo-genetic patterns in big datasets due to the large number of overlapping data points on a graph. To overcome this problem, we developed a heuristic statistical framework for detecting **G**eo-**G**enetic **outlier**, named *GGoutlieR*. Our *GGoutlieR* package computes empirical p-values for violation of the isolation-by-distance assumption for individual samples according to prior information on their geographic origin and genotyping data. This feature allows researchers to easily select outliers from thousands of samples for further investigation. In addition, *GGoutlieR* visualizes the geo-genetic patterns of outliers as a network on a geographical map, providing insights into the relationships between geography and genetic clusters.

3.3 Algorithm of *GGoutlieR*

Assuming isolation by distance, the geographical origins of samples can be predicted from their patterns of genetic variation, and vice versa (Battey et al. 2020a; Guillot et al. 2016). In this context, prediction models should result in large prediction errors for samples that violate the isolation-by-distance assumption. Based on this concept, we developed the GGoutlieR framework to model anomalous geo-genetic patterns.

Briefly, *GGoutlieR* uses *K*-nearest neighbor (KNN) regression to predict genetic components with the *K* nearest geographical neighbors, and also predicts in the opposite direction. Next, the prediction errors are transformed into distance-based (*D*) statistics and the optimal *K* is identified by minimizing the sum of the *D* statistics. The *D* statistic is assumed to follow a gamma distribution with unknown parameters. An empirical gamma distribution is obtained as the null distribution by finding optimal parameters using maximum likelihood estimation. With the null gamma distribution, *GGoutlieR* tests the null hypothesis that the geo-genetic pattern of a given sample is consistent with the isolation-by-distance assumption. Finally, p-values are calculated for each sample using the empirical null distribution and prediction error statistics. The details of the *GGoutlieR* framework are described step by step in Appendix B.

3.4 Example

3.4.1 Outlier identification

For demonstration, we used the genotypic and passport data of the global barley landrace collection of 1,661 accessions from the IPK genebank (König et al. 2020; Milner et al. 2019). The full analysis of the barley dataset with *GGoutlieR* is available in the vignette (`https://github.com/kjschmidlab/GGoutlieR/blob/master/vignettes/outlier_detection.pdf`). Outliers were identified using the *ggoutlier* function. The function *summary_ggoutlier* was then used to obtain a summary table of outliers by taking the output of *ggoutlier*.

```
library(GGoutlieR)
data("ipk_anc_coef") # get ancestry coefficients
data("ipk_geo_coord") # get geographical coordinates

pthres = 0.025 # set a p-value threshold

## run GGoutlieR
ggoutlier_result <- ggoutlier(geo_coord = ipk_geo_coord,
                              gen_coord = ipk_anc_coef,
                              plot_dir = "./fig",
                              p_thres = pthres,
                              cpu = 4,
                              klim = c(3,50),
                              method = "composite",
                              verbose = F,
                              min_nn_dist = 1000)

## print out outliers
head(summary_ggoutlier(ggoutlier_result))

#>                 ID     method      p.value
#> 1  BRIDGE_HOR_2827      geoKNN 0.0002534661
#> 2 BRIDGE_HOR_12795      geoKNN 0.0002875591
#> 3    BRIDGE_BCC_37      geoKNN 0.0003014085
#> 4 BRIDGE_HOR_10557      geoKNN 0.0003502037
#> 5 BRIDGE_HOR_10555      geoKNN 0.0003697646
#> 6        BTR_FT519 geneticKNN 0.0003828147
```

3.4.2 Visualization of unusual geo-genetic patterns

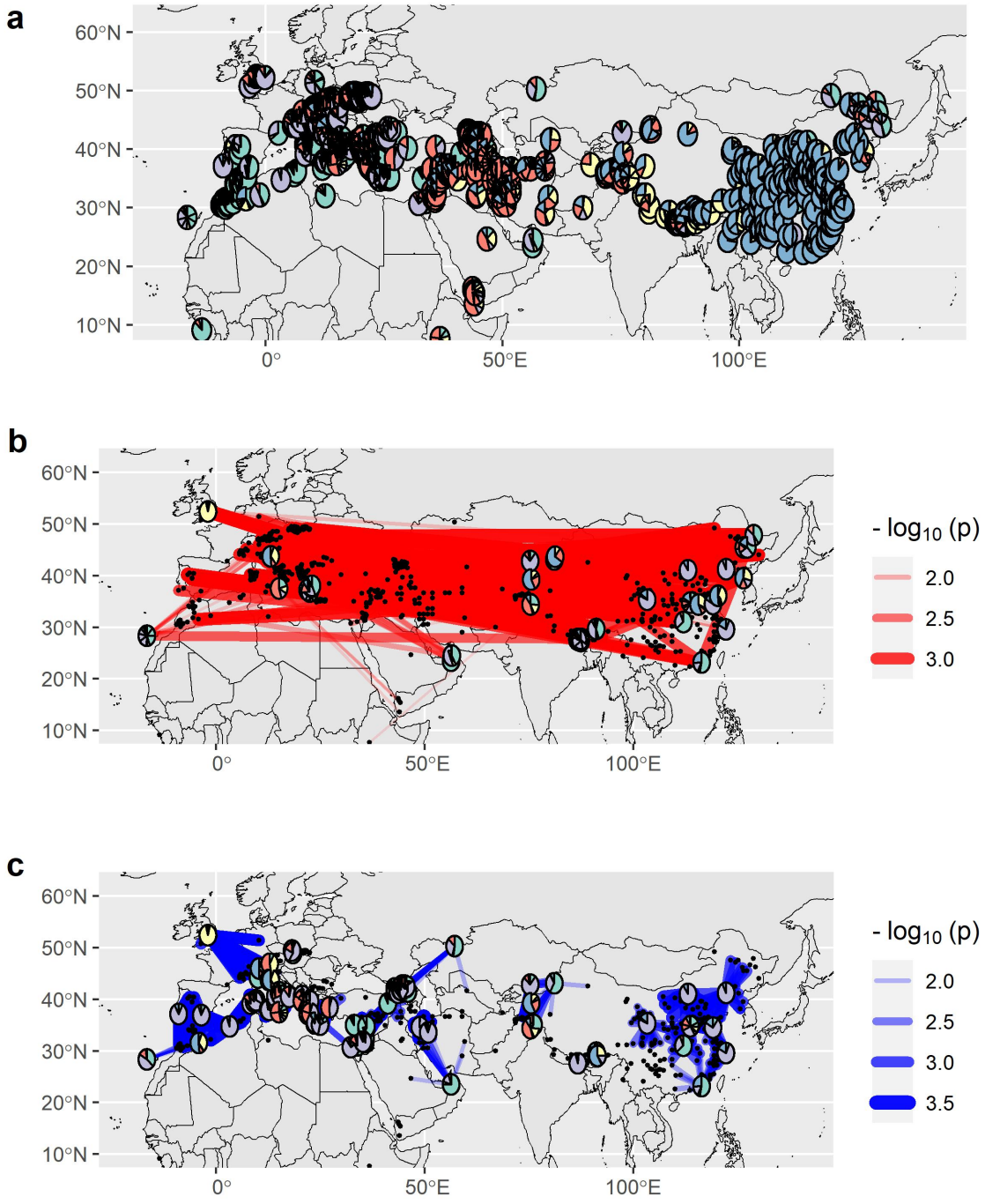

Figure 3.1 Visualization example of *GGoutlieR* with IPK barley landrace data. (a) Geographical map with ancestry coefficients of landraces presented by pie charts. (b) and (c) Unusual geo-genetic associations identified by GGoutlieR. The red lines show the individual pairs with unusual genetic similarities across long geographical distances. The blue lines indicate the unusual genetic differences between geographical neighbors. Pie charts present the ancestry coefficients of outliers.

The unusual geo-genetic patterns detected by *GGoutlieR* can be presented on a geographical map with the function *plot_ggoutlier* (Fig. 3.1).

Moreover, the function *plot_ggoutlier* allows users to gain insight into outliers from a selected geographical region (Fig. 3.2).

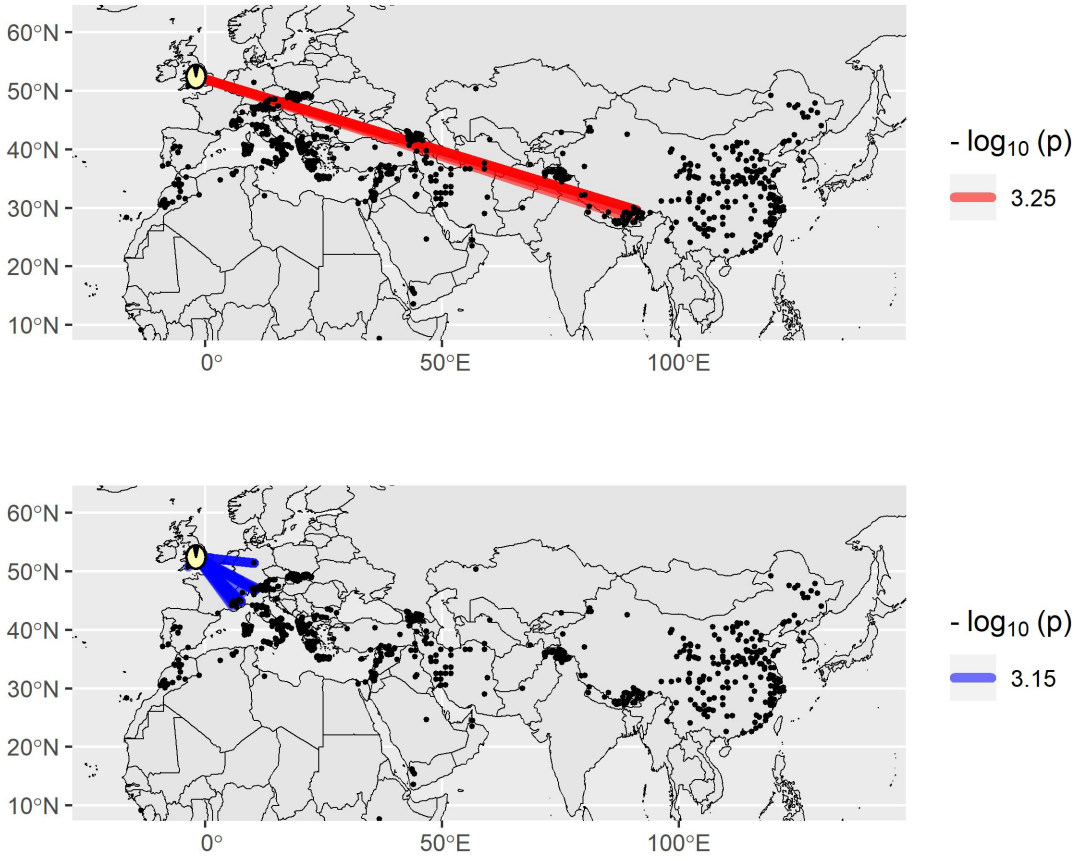

Figure 3.2 Visualization example of IPK barley landrace data with a highlight of samples from UK. The red lines show that the outliers in UK are genetically similar to accessions from Southern Tibet.

```
## Visualize GGoutlieR results
## Figure 1: visualize all outliers
plot_ggoutlier(ggoutlier_res = ggoutlier_result,
               gen_coord = ipk_anc_coef,
               geo_coord = ipk_geo_coord,
               p_thres = pthres,
               map_type = "both",
               select_xlim = c(-20,140),
               select_ylim = c(10,62),
               plot_xlim = c(-20,140),
               plot_ylim = c(10,62),
               pie_r_scale = 2,
               map_resolution = "medium")

## Figure 2: highlight outliers in UK with `select_xlim` and `select_ylim`
plot_ggoutlier(ggoutlier_res = ggoutlier_result,
               gen_coord = ipk_anc_coef,
               geo_coord = ipk_geo_coord,
               p_thres = pthres,
               map_type = "both",
               select_xlim = c(-12,4),
               select_ylim = c(47,61),
               plot_xlim = c(-20,140),
               plot_ylim = c(10,62),
               pie_r_scale = 2,
               map_resolution = "medium",
               add_benchmark_graph = F,
               plot_labels = NA)
```

3.5 Availability

The *GGoutlieR* package and vignette are available in our GitHub repository (`https://github.com/kjschmidlab/GGoutlieR`) and CRAN (`https://cran.r-project.org/web/packages/GGoutlieR/index.html`).

3.6 Acknowledgements

We are grateful to Dr. Martin Mascher and Max Haupt of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) for providing raw VCF data of barley landraces used in the example. This work was supported by the funds from the Federal Ministry of Food and Agriculture (BMEL) according to a decision of the parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the Federal Programme for Ecological Farming and Other Forms of Sustainable Agriculture (Project number 2818202615). C.W.C was supported by the Study Abroad Fellowship from the Education Ministry of Taiwan (R.O.C.) (Project number 1100123625).

Chapter 4

Predicting the geographical origins of barley genebank accessions using deep learning: Can large sample sizes improve genome-environment association studies?

This chapter will be submitted to Molecular Ecology Resources as:
Chang C.W. and Schmid, K., Evaluating the effect of predicting the geographic origin of barley genebank accessions using deep learning on genome-environment association studies.

4.1 Abstract

Genome-environment association (GEA) is an approach for identifying adaptive loci by combining genetic variation and environmental parameters. Application of GEA on crop genetic resources has demonstrated its potential for improving crop resilience in a changing climate. However, missing data on the geographic origin of genetic resources, e.g., in ex situ genebanks limits the effectiveness of GEA. We explored the use of neural network models and genomic data to overcome this limitation by predicting the genetic origin of genebank accessions. Neural networks demonstrated high prediction accuracy of geographical origins in cross-validation. However, its predictions occasionally lack ecological plausibility, as the current model solely focuses on geographical proximity between predicted and true origins without considering whether the predicted locations are viable for barley growth. For example, some predicted origins were located in the Mediterranean Sea. Using barley flowering time genes as benchmarks, GEA integrating imputed environmental data (N=11,032) displayed partially concordant yet complementary detection of genomic regions near flowering time genes compared to regular GEA (N=1,626), highlighting the potential of GEA with imputed data to complement regular GEA in uncovering novel adaptive loci. Also, contrary to our initial hypothesis that anticipates a significant improvement in GEA performance can be achieved by increasing sample size with a recovery of missing geographical origins, our simulations yield unexpected insights. Our study suggests potential limitations in the sensitivity of GEA approaches to the considerable expansion in sample size achieved through predicting missing geographical data.

Overall, our study provides insights into leveraging incomplete geographical origin data by integrating deep learning with GEA. Our findings indicate the need for further development of GEA approaches to optimize the use of imputed environmental data, such as incorporating regional GEA patterns instead of solely focusing on global associations between allele frequencies and environmental gradients across large-scale landscapes.

4.2 Introduction

Crop domestication and improvement have led to genetic bottlenecks, resulting in reduced genetic diversity in modern crop cultivars compared to their wild progenitors (Khoury et al. 2022; Tanksley and McCouch 1997). The erosion of genetic diversity poses a significant challenge for contemporary crop breeding, particularly in view of climate change that threatens agricultural productivity and resilience. To address the loss of adaptive genetic variation during domestication, plant breeders seek novel genetic variation in exotic genetic resources, such as traditional landraces and wild relatives (Bohra et al. 2022; Kumar et al. 2020). However, the introduction of useful alleles from these resources is hindered by linkage drag and potential reproductive barriers (Bohra et al. 2022; Dempewolf et al. 2017; Saad et al. 2022). In the genomics era, the integration of advanced sequencing technologies and computational approaches provides a promising avenue for introgressing favorable traits by mapping and selection of beneficial alleles using molecular breeding methods.

Genome-environment association (GEA), or environmental genome-wide association (EnvGWAS) studies, identify candidate adaptive loci associated with local adaptation by correlating allele frequencies with environmental variables (Coop et al. 2010; Lasky et al. 2012, 2015). GEA studies are based on the assumption that plants adapt to their local environments through natural selection, which drives changes in allele frequencies at causal genes influencing environmental adaptation across different regions in the distribution range of a species (Lasky et al. 2023). The application of GEA to crops and their wild relatives, such as sorghum (Lasky et al. 2015), rice (Gutaker et al. 2020), wild tomato (Gibson and Moyle 2020), sunflower (Todesco et al. 2022), and barley (Russell et al. 2016), indicates its potential to facilitate the identification and introgression of useful genetic variation into advanced breeding populations.

Genebanks preserve genetic diversity through extensive ex-situ collections of traditional crop cultivars, and they are an important source of novel and useful genetic variation for inclusion into plant breeding programs (Mascher et al. 2019). However, a significant challenge lies in efficiently identifying valuable genetic variants among thousands of genebank accessions, particularly given the limited resources available for their evaluation (Longin and Reif 2014; Schulthess et al. 2022). Recent advancements in affordable high-throughput sequencing have

allowed to unlock novel genetic diversity (Mascher et al. 2019) and have enabled comprehensive genomic analyses of genebank collections for various crops, including barley (Milner et al. 2019), wheat (Sansaloni et al. 2020; Schulthess et al. 2022), rice (Gutaker et al. 2020; Wang et al. 2018), pepper (Tripodi et al. 2021), and chickpea (Varshney et al. 2021).

Despite the rapid increase in genome-wide sequencing data, a relatively small proportion of genebank collections has a detailed record of their geographical origin, as many were collected before the establishment of modern collection data standards. For instance, among the 20,000 barley accessions from the German genebank at IPK that were genotyped by sequencing (Milner et al. 2019), only about 13% are geo-referenced (König et al. 2020). This limitation hinders the application of computational genome-environment association (GEA) approaches to identify adaptive genetic variants in plant genetic resources. However, the advent of deep learning methods, such as neural networks, along with advancements in computer hardware like graphics processing units (GPUs), has enabled computationally efficient inference of geographical origins for large datasets. For example, Battey et al. (2020a) developed *Locator*, a neural network-based tool that predicts geographical origins using complex allele distribution patterns. It outperforms conventional model-based approaches, such as SPASIBA (Guillot et al. 2016), in both computational efficiency and prediction accuracy. These advancements create the opportunity to recover missing geographical origin information for genebank accessions directly from genome sequencing data. The inferred geographic coordinates for additional accessions can subsequently be used in GEA, potentially enhancing the statistical power of environmental association studies by leveraging a much larger sample size (Cockram and Mackay 2018).

In this study, we explore the potential of applying geographical origin inference using the Locator method by Battey et al. (2020a) to improve the detection of adaptive loci reflecting selection by environmental factors using GEA approaches in genebank collections. We refer to this approach as *GEAplus* framework and apply it using the data of the global barley collection (Milner et al. 2019). Barley is one of the most important crops worldwide and its data set is suitable for this approach because of its large size with more than 10,000 genotyped accessions. We first assess the prediction accuracy of geographical origin inference by deep learning in barley landraces and compare the results of conventional GEA with those of *GEAplus* to test

the hypothesis that a larger sample leads to the discovery of additional genomic regions involved in local adaptation. To validate the observed results *GEAplus*, we perform conventional GEA and *GEAplus* on populations generated through forward-in-time simulations of genetic variation under different models of domestication, and then compare the power of conventional GEA and *GEAplus* to identify causal single nucleotide polymorphisms (SNPs) affecting fitness in different environments.

4.3 Materials and Methods

4.3.1 Overview of *GEAplus* framework

We developed the *GEAplus* framework (Figure 4.1) to investigate genomic adaptation by using genebank samples that lack geographical origin data. The framework begins by constructing a fully connected neural network model trained on geo-referenced samples with both geographical origin and genotypic data. This predictive model infers the geographical origins of a prediction set—samples without passport data—using genome-wide genetic variants, without relying on an explicit isolation-by-distance model (Battey et al. 2020a). This step involves imputing the geographical origin of accessions based on the spatial distribution of SNPs. Once the geographical origins are inferred, environmental data for the prediction set samples are retrieved from public databases. Finally, genome-wide association studies using a mixed linear model are conducted across all samples, including both the training set and the prediction set, to identify potential adaptive loci by testing associations between genetic variants and environmental variables. For clarity, we referred to the traditional GEA approach using only geo-referenced accessions as *regular GEA* in this work to distinguish it from *GEAplus*, which incorporates predicted geographical coordinates.

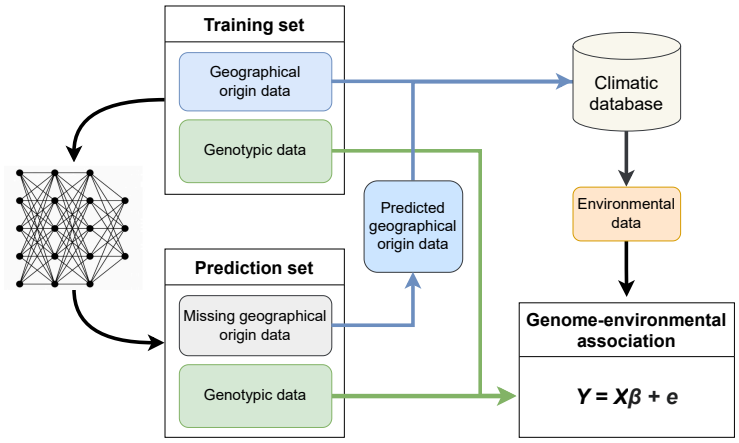

Figure 4.1 Flowchart of the *GEAplus* framework. Accessions with both geographical origin and genotypic data are used as a training set to build a prediction model using a fully connected neural network (Battey et al. 2020a). Once the model is trained, it predicts the geographical origins of accessions lacking passport data based on their genotypic information. Subsequently, environmental data corresponding to the predicted geographical origins are retrieved from public databases. Finally, genome-environment association (GEA) analyses are performed, incorporating genotypic data and various environmental parameters, including those derived through the imputed geographical origins.

4.3.2 Analysis of a global barley landrace collection

Available data of the barley landrace accessions

To evaluate the effectiveness of *GEAplus*, we applied both the regular GEA and *GEAplus* to a global collection of barley landraces. We identified 12,129 landrace accessions from the BRIDGE database (König et al. 2020) by selecting those classified as *Traditional cultivar/landrace*. For the GEA analysis, we retrieved geographical coordinates for 1,661 geo-referenced accessions from the same database (König et al. 2020). As a result, only about 14% of the landrace accessions included in this study had precise geographical origin data. It is important to note that caution should be exercised with the collection site data due to incomplete documentation of seed exchanges or potential errors in passport data, as highlighted in a previous study (Milner et al. 2019).

All selected landrace accessions were genotyped using genotyping-by-sequencing (GBS) by Milner et al. (2019). Single-nucleotide polymorphisms (SNPs) were identified as Milner et al. (2019) with aligning sequences against the 'Morex' V3 genome assembly (Mascher et al. 2021). Subsequently, SNPs were filtered using VCFtools v0.1.17 (Danecek et al. 2011) with a

minor allele count $\geq$ 10 and a maximum missing rate < 95%. Missing values were then imputed for all landrace accessions using BEAGLE v5.2 (Browning et al. 2018) with default settings.

In this study, we used the deep learning model *Locator* (Battey et al. 2020a) to infer missing geographical origin data. This model predicts geographical origins from genotypic data using a fully connected neural network with a loss function designed to minimize the Euclidean distance between the predicted and true locations of samples in the training data. Simulations in Battey et al. (2020a) indicate that *Locator* requires a high density of SNPs to achieve accurate predictions of geographical origin. To meet this requirement, we applied a relaxed filtering criterion, resulting in 557,991 SNPs with a minor allele frequency (MAF) > 0.01 among geo-referenced accessions across the entire landrace collection. For genome-environment association (GEA) analyses, we further filtered these to select 87,036 SNPs with a minor homozygous genotype frequency > 0.01 and a heterozygosity < 0.1 among geo-referenced accessions. Two datasets were prepared with the selected 87,036 SNPs: one for the geo-referenced accessions and another for the entire landrace collection. These datasets were used to perform regular GEA and *GEAplus*, respectively.

We then retrieved bioclimatic data for individual landrace accessions from the WorldClim 2.1 database (Fick and Hijmans 2017). We extracted 19 environmental variables based on the geographical coordinates of origin using the *raster* package (Hijmans 2018). Subsequently, we performed a principal component analysis (PCA) on the 19 environmental variables and used the first three principal components in the downstream GEA analysis to address the multicollinearity of the environmental variables (Hoban et al. 2016).

Data cleaning

Due to seed exchange and other historical human activities, the genetic similarity of barley germplasm may conflict with their documented geographical origins (Milner et al. 2019). This disruption of the geo-genetic association weakens the isolation-by-distance pattern, which is expected to reduce the accuracy of geographical origin inference. To enhance prediction accuracy, we removed outliers exhibiting unusual geo-genetic patterns using the *GGoutlieR* R package (version 1.0.2) (Chang and Schmid 2023). As input for *GGoutlieR*, we first calcu-

lated ancestry coefficients for the 1,661 geo-referenced accessions using 29,219 representative SNPs. These SNPs were pruned with *PLINK* 1.9 (Purcell et al. 2007) using an $r^2 < 0.1$ threshold. The optimal number of ancestral populations (K) was determined with the *estimate_d* function, and ancestry coefficients were estimated using the *alstructure* function from the R package *ALStructure* (Cabreros and Storey 2019). Outlier samples that violated the isolation-by-distance assumption were identified using *GGoutlieR* with the parameter *p_thres = 0.025*, while all other settings were kept at their default values. These outliers were subsequently excluded from the training of the *Locator* model.

Inference of geographical origin

We used all accessions that passed the *GGoutlieR* filter to train the neural network model, *Locator* v1.2 (Battey et al. 2020a). In assessing the usefulness of geographical origin inference with *Locator*, our main focus was on the general accuracy and robustness, rather than obtaining the most accurate model by tuning the hyperparameters of the neural network model, so the default setting was used.

We evaluated the prediction accuracy using ten-fold cross-validation with ten replicates. The prediction accuracy was measured as the R^2 between a true coordinate and a predicted coordinate. To extract environmental data for GEA analysis (Fig. 4.1), we averaged the predicted coordinates from all prediction models for each accession. To train the model and perform the prediction of geographical origin inference, we used four Nvidia A100 GPUs and 256 Gb working memory of the high-performance computing clusters of the State of Baden-Württemberg (bwHPC; `https://www.bwhpc.de/`), Germany.

GEA analysis with barley landraces

Due to the large sample size, we employed the computationally efficient genome-wide association approach *REGENIE* (Mbatchou et al. 2021) for GEA analysis. To fit the kinship effect, *REGENIE* uses representative SNPs to conduct genome-wide ridge regression in the first stage of its algorithm. We used a set of SNPs with $r^2 < 0.2$, extracted using *PLINK*, for

the genome-wide ridge regression of *REGENIE*. We also used the first three principal components of genotypic data, computed with *PLINK*, for controlling the population structure in the first and second stages of *REGENIE*. To control false positive discoveries, we computed adjusted p-values based on chi-square statistics from *REGENIE* using the genomic inflation factor method (François et al. 2016). Chi-square statistics were divided by an inflation factor λ, calculated as $\lambda = Median(z^2)/\chi^2_{0.5,\ df=1}$, where z^2 is the estimated chi-square statistics and $\chi^2_{0.5,\ df=1}$ is the middle quantile of a chi-square distribution with one degree of freedom. Adjusted p-values were calculated using the R code of the genomic inflation factor method from Capblancq et al. (2018). We conducted regular GEA and *GEAplus* analysis separately using the geo-referenced accessions and the entire landrace collection. The environmental variables for the entire landrace collection included data from both known geographical origins of geo-referenced samples and predicted origins of non-geo-referenced samples. Since flowering time genes are crucial for barley adaptation (Russell et al. 2016), we used them as benchmarks for successful identification of adaptive loci. To obtain the positions of flowering time genes on the 'Morex' V3 genome assembly, we queried the NCBI database (www.ncbi.nlm.nih.gov/) with the following keywords: *heading[All Fields] OR CEN-like[All Fields] OR flowering[All Fields] OR vernalization[All Fields] OR PPD[All Fields] OR VRN-H1[All Fields] AND "Hordeum vulgare"[porgn] AND alive[prop]*. For our analysis, we parsed the physical positions of eleven flowering time genes on the 'Morex' V3 genome assembly, including *FT3*, *ELF3*,*PPD-H1*, *FT4*, *CEN*, *FT2*, *FT5*,*TFL1*, *ABAP1*, *VRN-H1*, and *FT1* (reviewed by Fernández-Calleja et al. 2021).

4.3.3 SLiM simulation

To assess the performance of *GEAplus* framework, we simulated populations under environmental selection with gene flow using *SLiM* v4.0 (Haller and Messer 2023a). SLiM is a tool for forward genetic simulations of evolutionary processes with defined parameters, such as mutation rates and migration rates. We implemented a spatial stepping-stone model to approximate gene flow on a continuous landscape, with sub-populations located at real collection sites of barley landraces in the Fertile Crescent and surrounding regions. The simulation focused on a relatively small geographical range instead of a global level to ensure it was computationally feasible. We obtained the geographical coordinates of collection sites from the *Genesys*

database (`https://www.genesys-pgr.org/`) and selected 312 sites with spatial sampling to ensure that each site was at least 40 km away from the others. Since we did not simulate long-distance migration between sites and mislabeling of samples, additional data cleaning with *GGoutlier* was not needed.

We performed simulations under two demographic scenarios. The first scenario involved range expansion from a single refugium (1R) near an archaeological site of domesticated barley in the Israel-Jordan area (Mascher et al. 2016; Sallam et al. 2024) (Fig. C.1a). The second scenario involved expansion from two refugia (2R), with the second refugium located around the Zagros Mountains based on the hypothesis of a second domestication center (Morrell and Clegg 2007; Fig. C.1b). The migration rates between adjacent sites per generation were defined as a function of inverse geographical distances. We conducted three replicates of simulations for each demographic scenario by setting different numeric seeds for each replicate. Our simulations were performed with the *nonWF* model of *SLiM*, which supports dynamic population sizes, allowing simulations of sub-populations to propagate from a single individual and enabling the possibility of population extinction.

We ran simulations for 20,000 generations, with each run initiated with a burn-in phase to establish standing variation followed by 10,000 post-burn-in generations as barley was domesticated $\sim$10,000 years ago (Badr et al. 2000; Sallam et al. 2024). During the burn-in phase, no migration events were allowed, and the carrying capacity of the initial site was set to 5,000 individuals. Population expansion was initiated in the post-burn-in phase, with migration allowed for 10,000 generations. We controlled the total individual number by setting a carrying capacity of 250 individuals for each site in the post-burn-in phase, resulting in approximately 78,000 individuals in our simulation.

We simulated 10^6 loci for each individual, including 999,900 neutral loci and 100 biallelic selected SNPs, or quantitative trait loci (QTLs), evenly distributed across 10 linkage groups. To compensate for the small individual number, we used the method of Matz et al. (2020) by setting the mutation rate to $6.5 \times 10^{-}7$, which is 100-fold higher than the estimation in wild barley (Li et al. 2020). The outcrossing rate was set to 0.01, as the estimated outcrossing rate ranged from 0-1.8% (Abdel-Ghani et al. 2004). The mutation effects of QTLs were drawn from

a normal distribution $N(0, 0.45)$ (see Appendix C for details). We calculated the phenotype of an individual as the sum of QTL effects and a random value drawn from the normal distribution $N(0, 0.5)$ to simulate environmental noise.

To simulate local adaptation, we transformed phenotypes into fitness values for each individual based on local environmental conditions, following the approach described by Haller and Messer (2016). We treated the mean temperature of the warmest quarter from the WorldClim2 (Fick and Hijmans 2017) as an ideal local optimum phenotype for each site. We calculated the relative fitness of an individual as the ratio of its phenotype to the local optimum phenotype. Fitness was defined as the probability density of the normal distribution $N(y_{opt,j}, \sigma_{plasticity})$, where $y_{opt,j}$ is the local optimum phenotype for the jth site and $\sigma_{plasticity}$ is a parameter that controls the strength of selection, which was set to 2.85 (see Appendix C for details). With this setting, the relative fitness of an individual approached to 1 at a given site if its phenotype was close to the locally optimal phenotype.

To reduce the computational demands of our simulations, we adopted a hybrid strategy that combined forward and backward simulations, as described by Haller et al. (2019). In brief, we conducted forward simulations using *SLiM* with the tree-sequence recording method (Haller et al. 2019; Kelleher et al. 2018) to record pedigrees. Neutral mutations were suppressed in the forward simulation and overlaid in the backward simulation according to the tree sequences. This approach ignored neutral mutations in the forward simulation, as neutral mutations were assumed not to affect genealogies (Haller et al. 2019). We then used *pyslim* v1.0.1 and *tskit* v0.5.3 (Kelleher et al. 2018) within a *Python* 3.8 environment to read and manipulate the tree sequences from *SLiM*. We performed the backward simulation of neutral mutations using *msprime* v1.2.0 (Baumdicker et al. 2022) with a mutation rate of $6.5 \times 10^{-}7$. At the end of our simulations, we sampled 50 individuals from each site and recorded their genotypes, resulting in a VCF file with 15,600 individuals. We removed loci with MAF less than 0.005 using *vcftools* v0.1.17 (Danecek et al. 2011) as we were not interested in rare alleles. The scripts of simulations are available at `https://gitlab.com/CheWeiChang1992/geaplus`.

4.3.4 Accuracy of geographical origin inference

To simulate the absence of geographical origin data for genebank accessions, we masked the geographical origin information for 95% of individuals in the simulated dataset. First, we randomly selected 50% of the simulated sub-populations (156 out of 312) for further sampling, reflecting the fact that genebank accessions may originate from populations without recorded geographical coordinates, such as those collected before the advent of the Global Positioning System (GPS). Next, we assumed that geographical origin records were largely missing for accessions from the selected sub-populations. Based on this assumption, we randomly selected five individuals from each chosen site to form a training set and masked the geographical origins of the remaining individuals to create a prediction set. By repeating this sampling procedure, we generated three replicates of training sets ($N = 780$) and prediction sets ($N = 14,820$) for each simulation run.

To infer the missing geographical origins, we employed the *Locator* deep neural network model (Battey et al. 2020a). To prevent bias in the predictions, QTLs were excluded from the *Locator* model during training, ensuring they were not used to predict geographical coordinates. For model training, the individuals in the training set were randomly divided into ten groups, with one group sequentially designated as the validation set in each iteration. This procedure was repeated eight times, resulting in a total of 80 trained models for each training set. Following prediction, the geographical origins inferred by the 80 models for each masked individual were averaged to produce the final estimates. The accuracy of geographical origin inference was assessed using the R^2 value, comparing the true locations with the predicted locations.

Since 50% of the simulated sub-populations were excluded from the model training, individuals originating from these removed sub-populations were labeled as Type 1 samples (Fig. 4.2) for clarity. Conversely, individuals sampled from sub-populations included in the model training were categorized as Type 2 samples (Fig. 4.2). Predicting origins for Type 2 samples reflects a scenario where samples with unknown origins might actually belong to a population present in the training set, introducing the potential for data leakage in prediction models (Bernett et al. 2024). To evaluate performance, we computed the R^2 separately for Type 1 and Type

2 samples, alongside measuring prediction errors as the geographical distance between true and predicted locations.

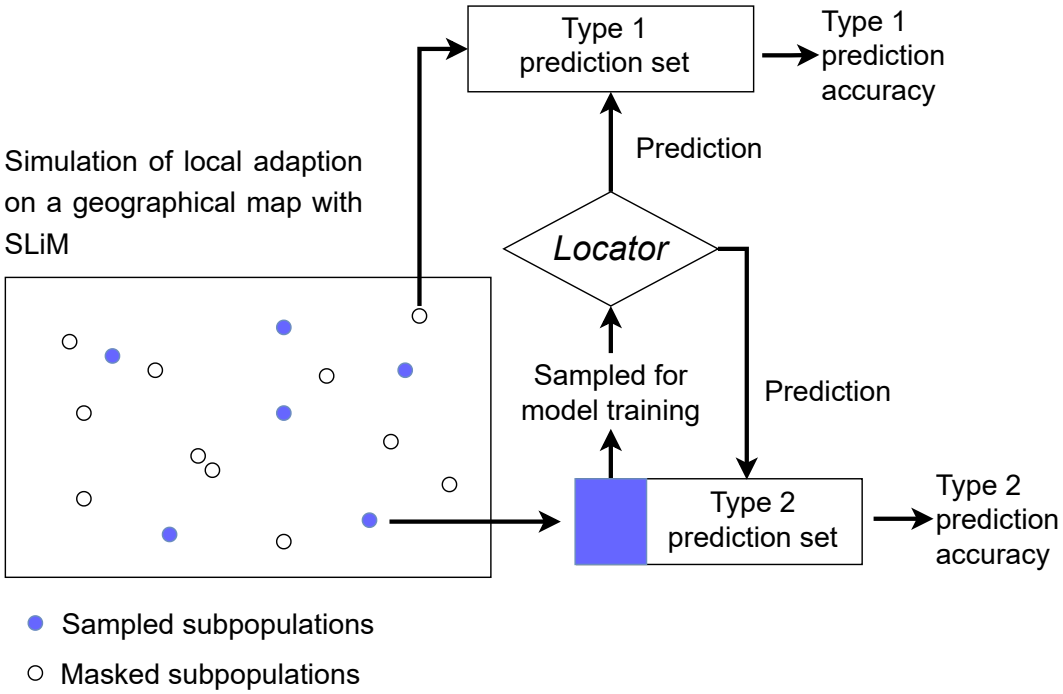

Figure 4.2 Design of Type 1 and Type 2 prediction for geographical origin inference with SLiM simulated data. The Type 1 prediction set consists of individuals from sub-populations that are not involved in model training, whereas the Type 2 prediction set contains individuals from sub-populations partially used as a training set. Blue dots and a blue block represent individuals sampled as a training set for *Locator*. Open dots and blocks represent individuals used as prediction sets.

4.3.5 Performance of *GEAplus*

To assess the performance of GEA, we conducted genome scans using simulated data filtered with a minor allele frequency (MAF) threshold of >0.01. To manage computational demands effectively, we implemented individual-based GEA with *REGENIE* and two population-based GEA approaches, as outlined below.

For individual-based GEA, we applied the same *REGENIE* settings used for the IPK landrace collection. In particular, SNPs with $r^2 < 0.2$ were used for genome-wide regression to account for kinship effects, while the first three principal components of genotypic values were included as covariates to control for population structure.

For population-based GEA, sub-populations were defined based on geographical origins, followed by the calculation of allele frequencies. The sub-populations were determined by grouping individuals into geographical clusters using complete-linkage hierarchical clustering with the R function *hclust* based on geographical distances. In cases where individuals were part of the prediction set, their predicted geographical origins were used. The clustering process began by grouping individuals with pairwise geographical distances of less than 10 km. To minimize environmental heterogeneity, further subdivision of geographical groups was performed through hierarchical clustering whenever the standard deviation of the environmental variable (SD_{env}) within a group exceeded 1% of the total SD_{env} across all individuals. After defining sub-populations, allele frequencies and mean environmental variables were calculated to perform the population-based GEA.

We employed two population-based GEA approaches: redundancy analysis (RDA) and a likelihood ratio test based on linear models (LM). RDA has been demonstrated to be a robust method for identifying adaptive loci with high statistical power (Forester et al. 2018). Two types of RDA were performed: simple RDA and partial RDA, the latter being conditioned on covariates. In both approaches, environmental data were used as explanatory variables, while allele frequencies served as the response variables. For partial RDA, the first three principal components of the genotypic data matrix were included as covariates to account for the confounding effects of population structure. To assess the statistical significance of SNPs, we applied chi-squared tests with 1 degree of freedom, following the framework proposed by Capblancq et al. (2018), to compute p-values.

We used four linear models for likelihood ratio tests, referred to as *LM_naive*, *LM_RR*, *LM_P*, and *LM_PRR*. The tests were conducted using our R function *geascan_rrlm*, which is available at `https://gitlab.com/CheWeiChang1992/geaplus/-/blob/main/SLiM_gea/R_script/GEAscan_RRLMv2.R`. The *LM_naive* model considered only the direct association between allele frequencies of geographical groups and environmental variation. To account for kinship and population structure, we adopted an approach similar to the individual-based GEA performed with *REGENIE*. In the *LM_P* and *LM_PRR* models, we accounted for population structure by excluding the components of environmental variation explained by the first three principal components of the genotypic data matrix. For the *LM_RR* and *LM_PRR* models,

we further corrected for kinship effects by removing components of environmental variation explained by genome-wide ridge regression (RR) before performing the likelihood ratio tests.

The four models were formulated as:

LM_naive: $Env \sim f_m$

LM_P: $Env \sim f_m + Env_{PC}$

LM_RR: $Env \sim f_m + Env_{Kinship}$

LM_PRR: $Env \sim f_m + Env_{PC} + Env_{Kinship}$

Here, Env represents a vector of environmental variables, and $f_m = \{f_{m,1}, f_{m,2}, f_{m,3}, ..., f_{m,n}\}$ denotes the allele frequencies of n geographical groups at the m-th SNP locus. Env_{PC} captures the environmental variation explained by the first three principal components of the allele frequency matrix $F_{genotype} = [f_1, f_2, f_3, ..., f_m]$. $Env_{Kinship}$ represents the environmental variation explained by genome-wide ridge regression.

Likelihood ratio tests were performed by calculating the test statistic:

$$\lambda_{LR} = -2 \ln \frac{L(M_{null})}{L(M_{alt})}$$

To control false positive discoveries, we adjusted λ_{LR} using the genomic inflation factor method (François et al. 2016) and calculated p-values based on a chi-squared distribution with one degree of freedom. To compare regular GEA with *GEAplus*, we evaluated the power and true positive rate of the seven GEA approaches described above: *REGENIE*, simple RDA, partial RDA, *LM_naive*, *LM_P*, *LM_RR*, and *LM_PRR*. Regular GEA was conducted using only the training sets ($N = 780$) with true environmental variables. In contrast, *GEAplus* analyses combined the training sets ($N = 780$) with true environmental variables and the prediction sets ($N = 14,820$), where environmental variables were derived based on predicted geographical origins.

The power of GEA was calculated as:

$$power = \frac{N_{true}}{N_{segQTL}}$$

where N_{true} is the number of detected QTLs, and N_{segQTL} is the number of segregating QTLs. Similarly, the true positive rate (TDR) was computed as:

$$TDR = \frac{N_{true}}{N_{sigSNPs}}$$

where $N_{sigSNPs}$ represents the total number of significant SNPs, including both neutral SNPs and QTLs.

4.4 Results

4.4.1 Prediction accuracy of geographical origins of barley landraces

We assessed the prediction accuracy (R^2) of *Locator* (Battey et al. 2020a) for inferring the geographical origins of IPK barley landraces using ten-fold cross-validation with ten replicates, resulting in one hundred R^2 values. For data quality control, we identified 117 accessions as outliers with geo-genetic patterns that deviated from the assumption of isolation-by-distance by using *GGoutlieR*. The prediction accuracy of *Locator* was evaluated separately for the original dataset of 1,661 geo-referenced accessions and a cleaned dataset of 1,544 geo-referenced accessions using cross-validation. Through the data cleaning procedure, we improved the mean R^2 values for longitude prediction from 0.932 (SD=0.039) to 0.987 (SD=0.009), and for latitude prediction from 0.839 (SD=0.062) to 0.921 (SD=0.033) (Fig. 4.3). The range of R^2 values after data cleaning also showed improvement, increasing from 0.819-0.988 to 0.954-0.996 for longitude, and from 0.637-0.947 to 0.803-0.975 for latitude (Fig. 4.3). Furthermore, we estimated prediction errors as the geographical distances between the predicted locations and the true geographical origins. The average and median prediction errors of *Locator* models after data cleaning were 203.47 km and 84.24 km, respectively (Fig. C.2).

Table 4.1 Correlations between the loading of the first three principal components of bioclimatic variables.

Geo-referenced landraces	All landraces		
	PC1	PC2	PC3
PC1	0.00	0.87	-0.37
PC2	0.82	0.42	0.31
PC3	-0.39	0.18	0.88

4.4.2 GEA of barley landraces

We used the first three principal components (PCs) of 19 bioclimatic variables for subsequent GEA analyses, referring to them as environmental PC1, PC2, and PC3. Among the geo-referenced landraces, 35 out of 1,661 accessions lacked bioclimatic data because their geographical coordinates fell in regions without available data. This resulted in 1,626 geo-referenced accessions being included in the environmental PCA. Similarly, after inferring missing geographical origins, we obtained bioclimatic data for 11,032 accessions from the entire landrace collection ($N = 12,129$) for environmental PCA. Accessions with predicted locations in regions lacking climatic data were excluded. These environmental PCs explained 79.08% and 81.69% of the environmental variance in the geo-referenced landraces ($N = 1,626$) and the entire landrace collection ($N = 11,032$), respectively (Fig. 4.4 A). This indicates that the first three PCs captured the majority of the environmental variation. The PCA loadings revealed that PC1 of the geo-referenced landraces was more strongly associated with PC2 (r = 0.87; Table 4.1) compared to PC1 of the entire landrace set (r = 0.00; Table 4.1; Fig. 4.4 B and C). Using environmental PCs as response variables in *REGENIE*, we performed both regular GEA and *GEAplus*, using eleven flowering time genes as benchmarks. A flowering time gene was considered successfully identified if a significant SNP (FDR < 0.05) was located within a 500 kb region adjacent to the gene, based on the decay of linkage disequilibrium in domesticated barley reported in Milner et al. (2019).

The regular GEA analysis of IPK accessions ($N = 1,626$) identified *PHOTOPERIOD-H1* (*PPD-H1*) on chromosome 2H and *VERNALIZATION-H1* (*VRN-H1*) on chromosome 5H as being associated with environmental PC1 (Fig. 4.5). The most significant SNPs were located

55.6 kb downstream of *PPD-H1* and 244.3 kb upstream of *VRN-H1*, respectively. No flowering time genes were identified in the regular GEA analysis using environmental PC2 or PC3 (Fig. C.3). In contrast, *GEAplus* ($N = 11,032$) identified only *PPD-H1*, and this association was with environmental PC2 (Fig. 4.6). No flowering time genes were detected by *GEAplus* with environmental PC1 or PC3 (Fig. C.4). The nearest significant SNP to *PPD-H1* identified by *GEAplus* was located 225.7 kb downstream of the gene. Comparing the two methods, regular GEA identified more significant SNPs that were closer to *PPD-H1* than those detected by *GEAplus*. Additionally, regular GEA successfully identified *VRN-H1*, which *GEAplus* did not detect. However, *GEAplus* identified four significant SNPs located 1.17 Mb upstream of *CENTRORADIALIS* (*CEN*; Fig. 4.6), whereas regular GEA did not detect any significant SNPs near *CEN* (Fig. 4.5, Fig. C.3).

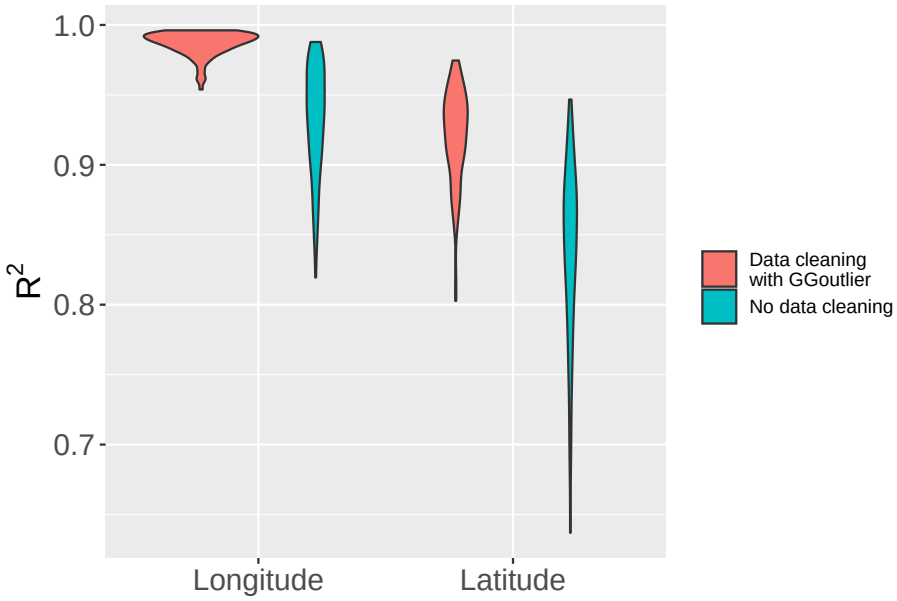

Figure 4.3 Prediction accuracy of geographical origin inference for IPK accessions. The accuracy was estimated by ten-fold cross-validation with ten replicates. The size of the population is 1,661 and 1,544, respectively, before and after data cleaning.

4.4.3 Prediction accuracy of geographical origins with simulated data

Accessions with unknown geographical origins may belong to the same sub-populations as those used to train prediction models, potentially inflating prediction accuracy due to data leak-

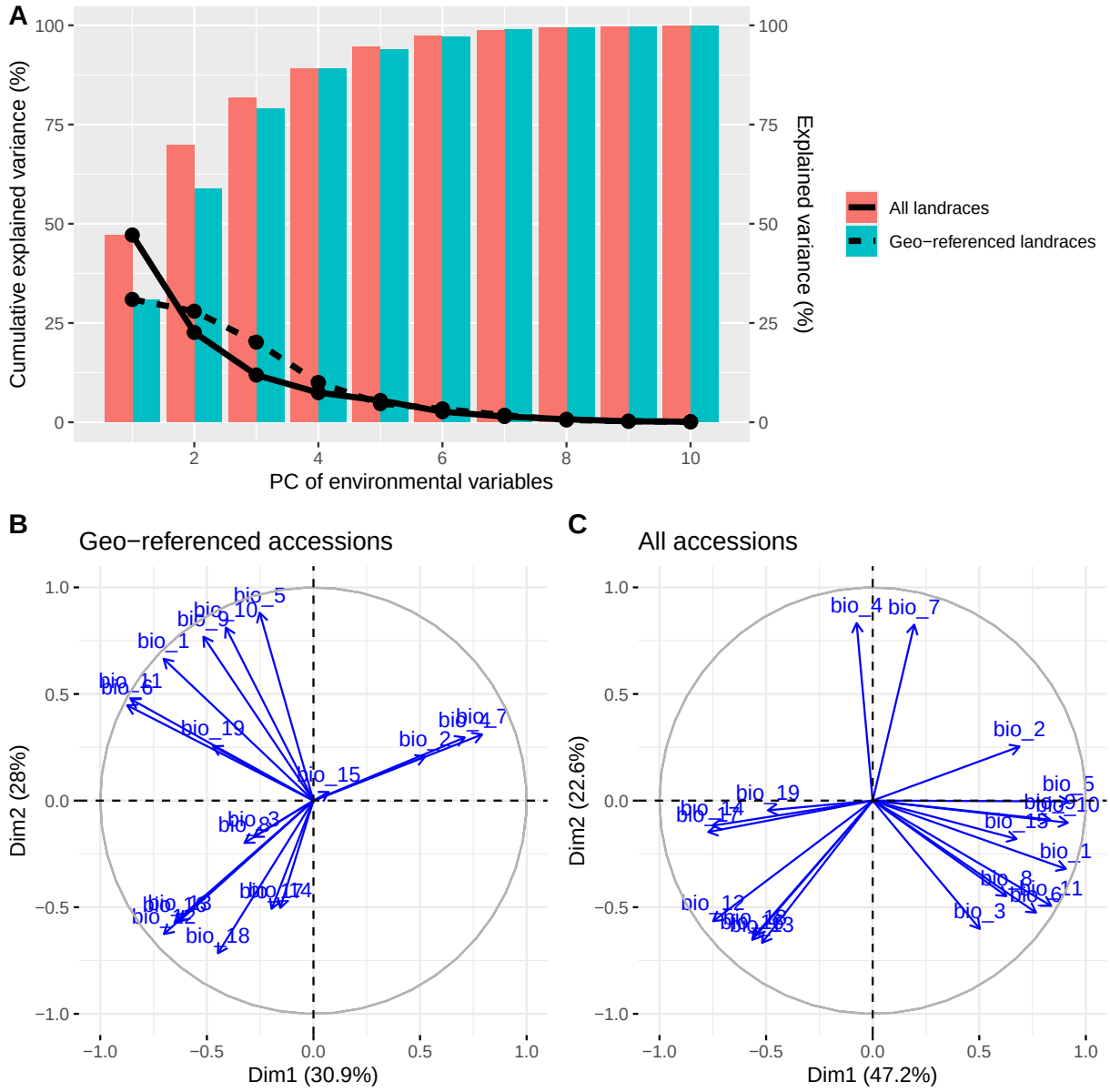

Figure 4.4 Summary of principal components (PCs) of 19 bioclimatic variables. (A) Variance explained by the first ten principal components. Colored bars represent cumulative explained variance, while lines indicate the variance explained by individual PCs. (B) Loading plot of the geo-referenced accessions (N=1,626). (C) Loading plot of the entire landrace collection with imputed environmental data (N=11,032).

age (Bernett et al. 2024). To address this issue, we evaluated prediction accuracy using two validation approaches. Type 1 predictions introduced minimal data leakage, as the prediction set was derived from sub-populations distinct from those in the training set. In contrast, Type 2 predictions accounted for significant data leakage, with the training and prediction sets overlapping in sub-populations (Fig. 4.2). This design reflects the realistic scenario where accessions with missing origins may not always come from sub-populations present in the training set. As

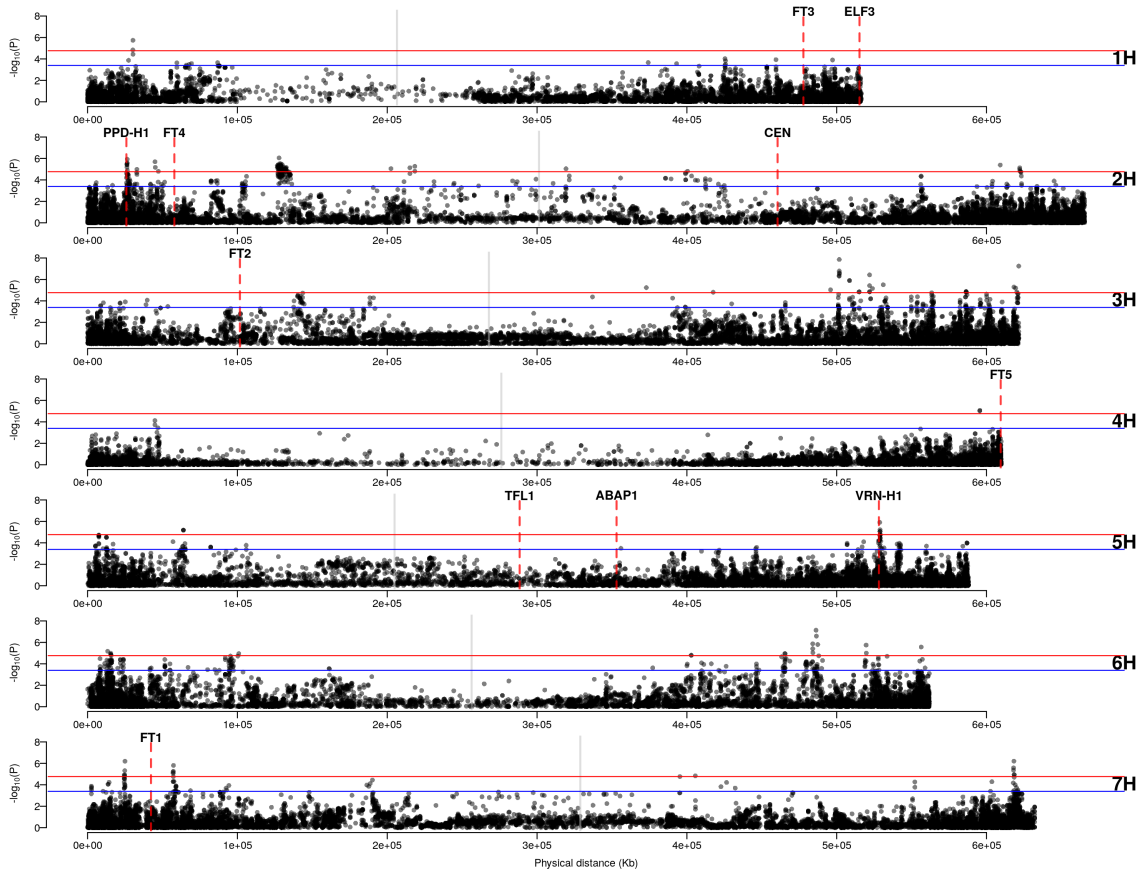

Figure 4.5 Regular GEA of IPK accessions (N=1,626) with the first environmental principal component (PC1). Blue and red horizontal lines are the significant levels of FDR = 0.05 and FDR = 0.01. Grey vertical lines indicate the positions of centromeres. Red dashed lines indicate the positions of flowering time genes.

expected, Type 1 predictions resulted in lower accuracy for both latitude and longitude coordinates compared to Type 2 predictions (Table 4.2 and Fig. 4.7). Furthermore, Type 1 predictions exhibited standard deviations over ten times larger than those observed for Type 2 predictions (Table 4.2 and Fig. 4.7).

Since we used imputed environmental data in GEA, we evaluated the accuracy of predicting environmental variables based on inferred geographical origins. The imputed environmental variables exhibited strong correlations with the true environmental variables, with average R^2 values ranging from 0.88 to 0.975 across four scenarios (Type 1 and Type 2 predictions paired with 1R and 2R demographic models; Fig. 4.8). Environmental data imputation was more accurate in Type 2 predictions than in Type 1 predictions, regardless of the demographic scenario (Fig. 4.8). Additionally, simulations of different demographic scenarios revealed that predictions were more accurate when population expansion originated from two initial sites (2R scenario)

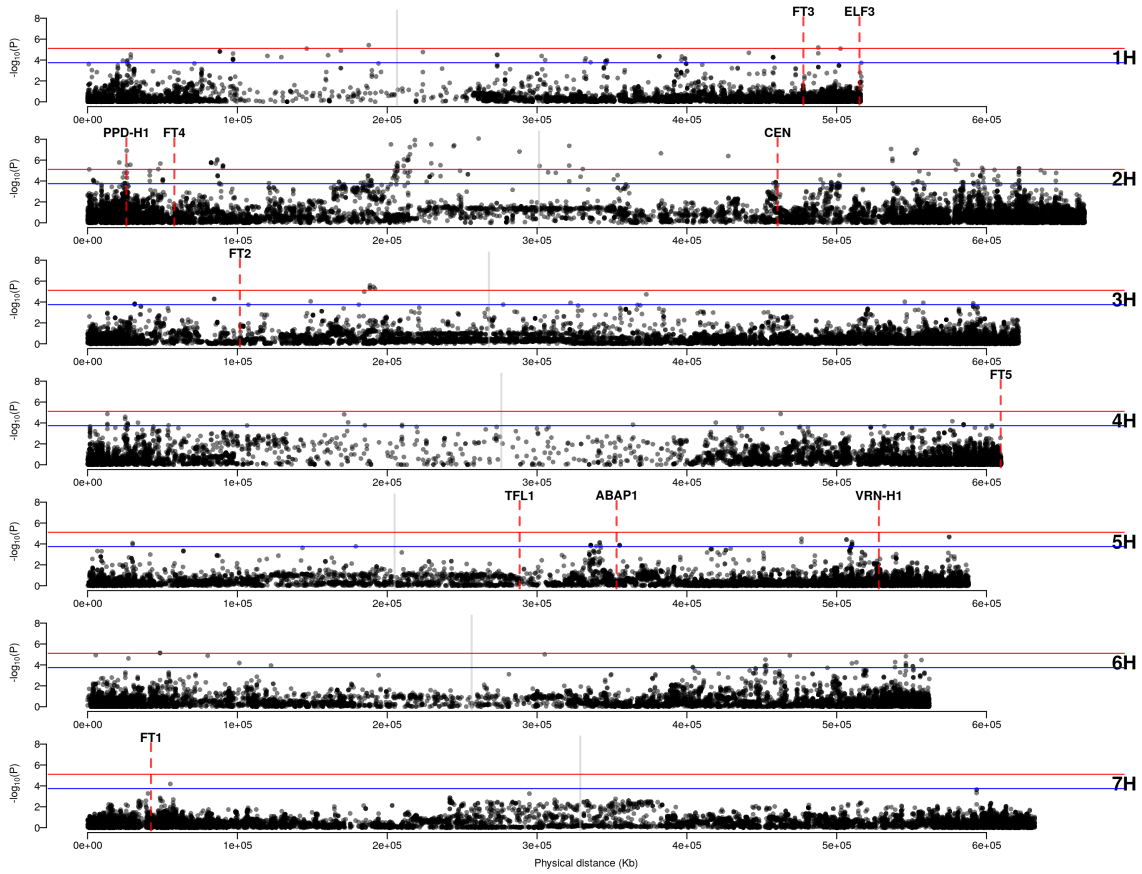

Figure 4.6 *GEAplus* of IPK accessions (N=11,032) with the second environmental principal component (PC2). Blue and red horizontal lines are the significant levels of FDR = 0.05 and FDR = 0.01. Grey vertical lines indicate the positions of centromeres. Red dashed lines indicate the positions of flowering time genes.

compared to a single initial site (1R scenario) (Fig. 4.7 and Fig. 4.8).

4.4.4 Limited benefit of imputing environmental data for GEA

We tested the hypothesis that imputing environmental variables could enhance the performance of GEA by evaluating the power and true discovery rate (TDR) of regular GEA and *GEAplus*. In the simulations, QTL detection was defined as the correct identification of selected SNPs, while the detection of neutral SNPs was considered a false positive. Our results indicated that imputing environmental variables through geographical origin inference did not improve the power of GEA, regardless of the demographic scenarios or GEA approaches used (Fig. 4.9). Both regular GEA and *GEAplus* exhibited similar performance across all tested models, except for *REGENIE*. Using population-based approaches, *GEAplus* demonstrated

Table 4.2 Prediction accuracy of geographical origin inference in SLiM simulated data.

Demography	R^2			
	Type 1		Type 2	
	Longitude	Latitude	Longitude	Latitude
1R	0.946 (0.026)	0.864 (0.035)	0.997 (0.0005)	0.991 (0.002)
2R	0.977 (0.005)	0.901 (0.023)	0.998 (0.0001)	0.991 (0.002)

power comparable to regular GEA. However, with the individual-based GEA approach using *REGENIE*, regular GEA exhibited higher power than *GEAplus* in both the 1R and 2R demographic scenarios (Fig. 4.9 A and B). This finding contradicts our hypothesis that imputing environmental data would enhance the identification of adaptive loci with GEA.

We also evaluated the true positive rate (TDR) to assess the accuracy of correctly identifying adaptive loci. Simple RDA and partial RDA showed a tendency toward higher TDR in *GEAplus* compared to regular GEA, while other approaches did not exhibit this pattern (Fig. 4.9 C and D). However, *GEAplus* with partial RDA failed to identify any significant SNPs in two of the simulation runs for both 1R and 2R scenarios, resulting in missing TDR values, which are not shown in Fig. 4.9 C and D.

In addition to investigating *GEAplus*, we examined a hypothetical scenario of perfect imputation of environmental variables, referred to as *perfect GEAplus*. To simulate this condition, we used the true environmental variables to conduct GEA on the entire simulated population and evaluated the power and TDR of *perfect GEAplus*. We found that the power of *perfect GEAplus* was not significantly different from that of regular GEA in the 1R scenario across all GEA approaches (Fig. 4.9 E). However, in the 2R scenario, individual-based GEA using *REGENIE* demonstrated slightly higher power compared to regular GEA (Fig. 4.9 F). Among all approaches, individual-based GEA with *REGENIE* exhibited the highest power in both the 1R and 2R scenarios (Fig. 4.9 E and F). When comparing TDR, no significant differences were observed between *perfect GEAplus* and regular GEA (Fig. 4.9 G, and H). However, population-based partial RDA tended to achieve a higher TDR compared to other approaches in our simulations, particularly in the 2R scenario (Fig. 4.9 G and H). Overall, our findings suggest that increasing the sample size does not significantly enhance the performance of GEA under our simulated scenarios, consistent with our empirical analysis.

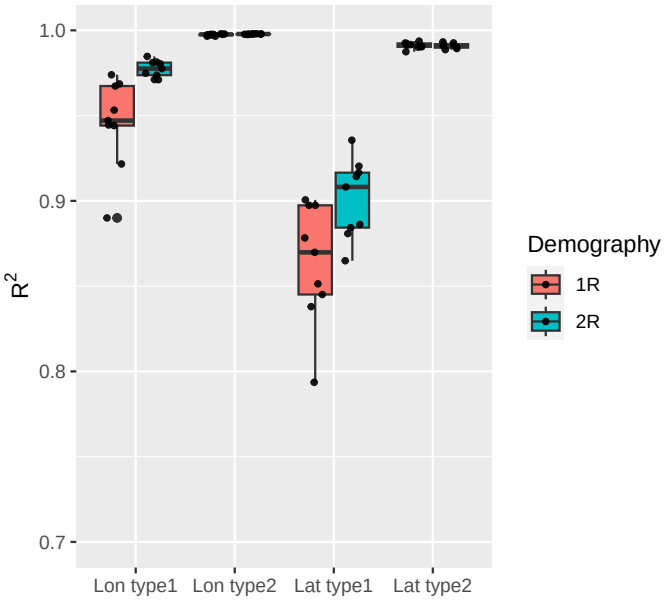

Figure 4.7 Prediction accuracy of *Locator* in SLiM simulated data. R^2 is calculated between true locations and predicted locations of prediction sets separately for longitude and latitude. Prediction accuracy is evaluated with Type 1 and Type 2 prediction sets as illustrated in Fig. 4.2.

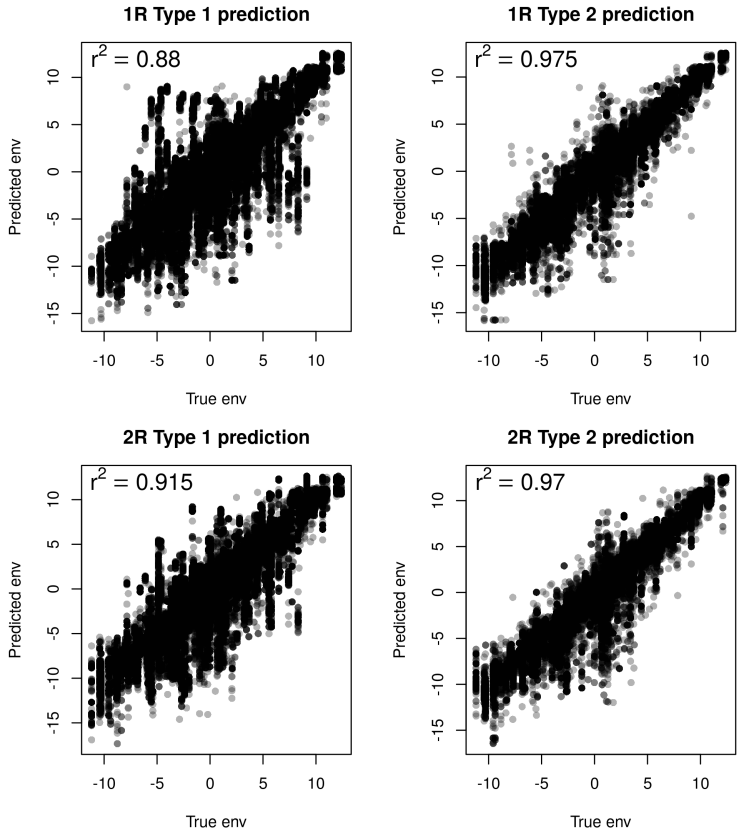

Figure 4.8 Imputation accuracy of environmental data based on predicted geographical origins in SLiM simulated data.

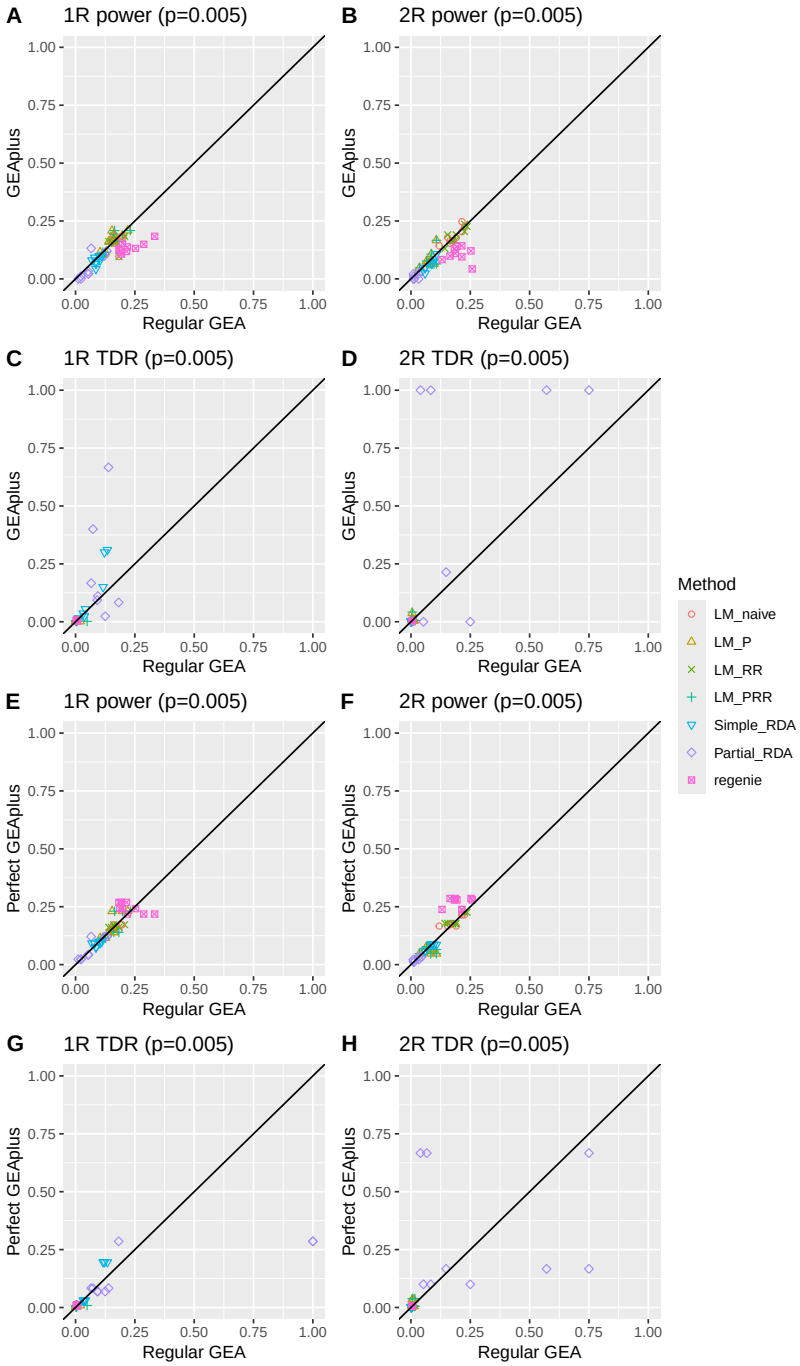

Figure 4.9 Comparison of power and true discovery rate (TDR) between regular GEA ($N = 780$) and *GEAplus*/*perfect GEAplus* ($N = 15,600$). (A-D) Pairwise comparisons between regular GEA and *GEAplus* under 1R and 2R demographic scenarios. Power and TDR are defined as the proportion of detected QTLs ($p = 0.005$) relative to the total number of segregating QTLs and the number of significant SNPs, respectively. TDR of *GEAplus* with RDA is missing in four simulations, two with 1R setting and two with 2R setting, because of no significant SNP. (E-H) Pairwise comparisons between regular GEA and *perfect GEAplus* under 1R and 2R demographic scenarios. *Perfect GEAplus* assumes environmental information is imputed with no error.

4.5 Discussion

Our approach of integrating imputation of environmental data and GEA rests on two key assumptions: first, that genomic sequences can accurately predict missing geographical origins based on the isolation by distance expectation, i.e. the majority of genetic variation are neutral and reflect demographic history instead of adaptation (Wright 1943), and second, that GEA can achieve higher power with increased sample sizes (Lotterhos and Whitlock 2015). While we observed good accuracy in geographical origin inference, the significant expansion in sample size did not translate into substantial benefits for GEA.

4.5.1 Performance of geographical origin inference

Using genome-wide sequencing data, we employed the deep-learning-based geographical origin inference approach *Locator* (Battey et al. 2020a) to recover missing location data for over ten thousand accessions. Our cross-validation results showed high accuracy ($R^2 > 0.9$) in predicting both longitude and latitude. We also evaluated the accuracy of geographical origin prediction using forward genetic simulations with a scaling approach to ensure computational feasibility. A recent study highlighted the need to increase the number of burn-in iterations in forward simulations, depending on the effective population size and sequence length, to obtain unbiased genetic diversity when using the scaling approach under the Wright-Fisher model (Ferrari et al. 2024). However, our study used the *nonWF* model of SLiM (Haller and Messer 2019) and focused on GEA rather than genetic-diversity-based metrics, making reduced genetic diversity due to scaling less of a concern. Our simulations revealed that *Locator* performed better under a population expansion scenario originating from two locations (2R scenario) compared to a single origin (1R scenario; Table 4.2). This suggests that the presence of population genetic structure may enhance the accuracy of geographical origin inference using neural networks. Additionally, we observed that prediction accuracy was higher for individuals derived from sub-populations overlapping with the training set (Type 2 prediction set) than for those sampled from entirely excluded sub-populations (Type 1 prediction set; Table 4.2). This finding indicates that prediction accuracy can vary depending on the genetic similarity between

tested samples and the individuals in the training set. To optimize the training set, the ideal sampling strategy should include as many sub-populations as possible, even if the sample size from each sub-population is small.

In our empirical analysis of genebank accessions, cross-validation revealed that the accuracy of *Locator* was strongly influenced by the extent to which the training set adhered to the isolation-by-distance assumption. We observed a significant improvement in prediction accuracy after removing outliers that violated this assumption. This finding highlights the critical role of data quality and population history in accurately inferring geographical origins. Populations with a history of recent long-distance migration events pose a particular challenge for geographical origin prediction. For instance, if migrants recently established a large colony in a distant environment, the deep-learning approach *Locator* may incorrectly infer geographical origins due to high genetic similarity between the original and newly established populations. Such scenarios are plausible for plant species that have dispersed through human activities or seed exchange in recent decades as postulated for Ethiopian barley traditional cultivars (Teklemariam et al. 2022). Therefore, recovering missing geographical origins from genotypic data requires a solid understanding of population history and careful data preprocessing to ensure reliable predictions.

Improving geographical origin inference with ecological knowledge

In our cross-validation of geographical origin inference for barley landraces, we observed some predictions that were ecologically implausible. The deep learning approach, *Locator*, minimizes prediction error by reducing the distance between predicted and true locations. However, it does not account for ecological constraints, such as whether the predicted locations are suitable for barley growth. For example, in our analysis of the barley landrace collection, 1,097 accessions were erroneously projected to the Mediterranean Sea, resulting in failed environmental data imputation. These errors stem from the limitations of the current deep learning model, which attempted to predict geographical origins for samples that might represent an admixture of landraces from Europe and North Africa. Additionally, the model overlooks the impact of microclimates, such as those caused by altitude in mountainous regions, where cli-

mate conditions can vary significantly over short distances and lead to distinct evolutionary outcomes (Hämälä and Savolainen 2019; Lampei et al. 2019). These limitations present a significant obstacle to using geographical origin inference in GEA studies. Inaccurate environmental data imputation, even when geographical origin inference achieves high accuracy, can introduce substantial noise into GEA analyses.

To improve the quality of environmental data imputation, incorporating knowledge of natural habitats is essential. A recent study proposed a statistical framework to estimate the credibility of individual geographical records for hundreds of species (Arlé et al. 2021). This framework assesses the likelihood of geographical origin records being accurate by leveraging species distribution data from biological databases. Integrating such probability-based plausibility measures into the training process of deep learning models could significantly enhance the reliability of geographical origin inference.

Effect of imputing environmental data on GEA

Using imputed environmental data, GEA has the potential to uncover distinct adaptive signatures, although this does not necessarily translate into increased power for genome-wide scans. In our implementation of *GEAplus* on barley landraces, we defined signals near flowering time genes as successful identifications of adaptive loci. Both regular GEA and *GEAplus* identified significant signals near *PPD-H1*, a gene associated with adaptation to photoperiod, temperature, and drought (Gol et al. 2021; Ochagavía et al. 2022; Turner et al. 2005; Wiegmann et al. 2019). Recent research involving the re-sequencing of *PPD-H1* in a global germplasm collection revealed associations between *PPD-H1* haplotypes and precipitation-related bioclimatic variables (Sharma et al. 2020). In contrast to the consistent signatures near *PPD-H1*, no significant SNPs close to *VRN-H1* were found with *GEAplus*, while new SNPs emerged upstream of *CEN* (Fig. 4.5 and 4.6). Although the possibility of false-positive discoveries cannot be ruled out, this observation may reflect the impact of geographical diffusion of mutations on GEA approaches. The efficacy of GEA depends on the differentiation of allele frequencies across distinct environments. Non-monotonic clinal patterns in allele frequencies along environmental gradients can result in the failure to detect adaptive loci using GEA (Lot-

terhos 2023). Moreover, the spread of mutations is often constrained by geographical distance, limiting adaptive alleles to specific areas. This can lead to non-clinal patterns where adaptive alleles are absent in geographically distant regions with comparable environmental conditions. Such scenarios are plausible in global germplasm collections, where accessions may originate from geographically distant yet ecologically similar environments.

We can conclude that environmental data imputation can influence GEA by either revealing new adaptive loci or eliminating previously observed adaptive signatures due to changes in allele frequencies. This issue is not unique to the *GEAplus* framework but also applies to regular GEA when samples are collected across a broad geographical range, as monotonic clinal patterns are a prerequisite for the success of GEA (Lotterhos 2023). To address this limitation, one potential solution is to perform multiple GEA tests within pre-defined smaller geographical areas using a sliding window approach, rather than conducting a single GEA test with all samples combined. This localized approach may improve the identification of adaptive loci by uncovering region-specific variations. However, this method introduces a large number of statistical tests, as it requires evaluating tens of thousands of SNPs for each environmental variable across hundreds of geographical windows. Such extensive testing could lead to a proliferation of spurious signals. Therefore, further development and rigorous validation of this methodology are crucial to ensure its effectiveness and reliability.

Validation of GEA approaches by simulations

In our comparison between *GEAplus* and regular GEA using simulations, we observed a slight improvement in power for individual-based GEA with *REGENIE* through environmental data imputation in the 2R scenario, but no such improvement for population-based GEA approaches, even with perfectly imputed environmental data (Fig. 4.9 F). This finding is consistent with observations reported by Lotterhos and Whitlock (2015), who noted that increasing the sample size improved the power of LFMM (Frichot et al. 2013), an individual-based method, but not Bayenv2.0 (Günther and Coop 2013), which is population-based. The difference in power was attributed to the individual-based versus population-based units of analysis in these approaches. Despite differences in simulation scenarios, our results align with Lotter-

hos and Whitlock (2015), suggesting that individual-based GEA tends to achieve higher statistical power than population-based GEA when sample sizes are large. Aside from the slight power improvement observed for the individual-based approach in the *perfect GEAplus* of the 2R scenario, our simulations indicated no general advantage of *GEAplus* over regular GEA, even with perfectly imputed environmental data (Fig. 4.9 E, F, G, and H). Previous simulation studies have suggested that GEA approaches are generally insensitive to sample size, except under island model scenarios (Forester et al. 2018; Lotterhos and Whitlock 2015). Since our simulations approximated an isolation-by-distance demographic structure in a confined geographical region, increasing the sample size did not significantly benefit GEA performance, as also observed in earlier studies (Forester et al. 2018; Lotterhos and Whitlock 2015). This was true even though the sample size in the *GEAplus* scheme was twenty times larger than that in the regular GEA scheme. However, if GEA were applied on a continent-wide scale with sufficient isolation between populations, larger sample sizes could potentially improve performance. In addition to demographic effects, various other factors, such as mating systems (Hodgins and Yeaman 2019), genome recombination rates (Lotterhos 2019), and sampling strategies (Lotterhos and Whitlock 2015), can influence the detection of adaptive loci. Future research incorporating diverse simulation settings that account for these factors will be essential to gain a more comprehensive understanding of the conditions under which GEA performs effectively.

Conclusion and outlook

In summary, our study demonstrates the potential of using extensive genomic data and deep learning models to recover missing geographical origins in global germplasm collections, providing ecological insights for identifying alleles associated with tolerance to environmental stresses. However, we also acknowledge the limitations of current geographical origin inference methods. Existing neural networks prioritize minimizing geographical distance prediction errors without considering the ecological plausibility of the predicted habitats. Our findings suggest that environmental data imputation primarily benefits individual-based GEA. Additionally, our simulations indicate that large-scale genotyping is unlikely to improve GEA performance

if samples are indiscriminately pooled for analysis. Instead, increasing the volume of data is more likely to enhance the detection of region-specific associations rather than global GEA patterns. The development of innovative approaches, such as an individual-based sliding window GEA method, will be crucial for effectively leveraging imputed environmental data while accounting for the geographical distribution of alleles.

4.6 Acknowledgments

We sincerely thank Dr. Martin Mascher and Max Haupt of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) for providing the raw VCF data of barley landraces used in this study. This research was supported by funding from the Federal Ministry of Food and Agriculture (BMEL) through the Federal Office for Agriculture and Food (BLE) under the Federal Programme for Ecological Farming and Other Forms of Sustainable Agriculture (Project number 2818202615). Additional support for C.W.C. was provided by the Study Abroad Fellowship from the Ministry of Education, Taiwan (R.O.C.) (Project number 1100123625).

Chapter 5

General Discussion

This work aimed to investigate adaptive genetic variants in exotic barley germplasm, focusing on wild barley population from the Southern Levant and a barley landrace collection of the IPK genebank. We used genotype-environment associations (GEAs), an approach combining genomic data with environmental variables based on the geographical origins of samples to identify loci involved in environmental adaptation (Lasky et al. 2023; Rellstab et al. 2015). Moreover, we developed a statistical framework to discover unusual geo-genetics associations for a systematic GEA data pre-treatment and proposed a novel GEA strategy to maximize the use of genotypic data for accessions lacking geographical data, which was assessed using forward genetic simulation. In general, this work extended the knowledge of barley adaptive evolution and offered insights into using GEA to investigate genetic variants in crop germplasm.

5.1 Landscape genomics for barley germplasm

In our wild barley study (Chapter 2, Chang et al. 2022), we conducted a comprehensive investigation of population genetic structure, gene flow, and natural selection signatures of the wild barley in the Southern Levant using genotyping-by-sequencing (GBS) data, the Global Positioning System (GPS) data, and environmental data of sample's geographical origins. The analysis revealed a stronger influence of neutral evolutionary processes than natural selection

on wild barley in this region. Redundancy analysis (RDA) showed that 15.12% of genomic variation was associated with environmental variables, of which approximately 4% of genomic variation was explained by a synthetic variable of latitude, rainfall, and solar radiation and 1.2% linked to soil water capacity.

Alongside the noticeable association with environmental variables, our finding also demonstrated the pervasive effect of isolation-by-distance, evident in the 45% genomic variation explained by the spatial autocorrelation. This observation is concordant with previous studies that emphasize neutrality in wild barley evolution in the Southern Levant (Volis et al. 2003, 2005). Moreover, about two-thirds of the environment-associated genomic variation was confounded with spatial autocorrelation, likely due to the collinearity between spatial and environmental gradients, as observed in other plant species like *Arabidopsis* (Lasky et al. 2012), sorghum (Lasky et al. 2015), and rice (Gutaker et al. 2020).

Besides, our study highlighted a limitation of GEA in cases of strong population structure that correlates with genotype-environment associations. We observed 83% of population genetic structure linked to spatial autocorrelation, likely reflecting the effect of genetic drift in selfing species. In self-fertilizing species, strong genetic drift due to reduced independent allele sampling could cause confounding between neutral genetic differentiation of populations and selection signatures (Hodgins and Yeaman 2019). Approximately 40% (6.41% out of 15.12%) of environment-associated genomic variation in our wild barley population was confounded with population structure. This confounding effect implies that correcting for population genetic structure may obscure some true selection signatures in GEA scans, as evidenced by the difference in detection patterns between RDA conditioned on population structure and other methods without controlling population structure effects. This limitation of GEA was reported in a previous study that showed the population structure correction in a GEA scan can strongly impact the statistical power to detect known adaptive genes associated with environmental gradients (Lasky et al. 2015). However, since strong genetic drift in selfing species can raise spurious selection signatures (Hodgins and Yeaman 2019), controlling for population structure remains essential to control for false positives.

In our study, genome-wide scans without controlling for population genetic structure identi-

fied selection signatures in the pericentromeric regions of chromosome 3H, 4H, and 5H. Given the weaker purifying selection in pericentromeric regions discovered in wild barley (Baker et al. 2014), the signatures detected by our genome scans may reflect selection rather than merely the consequence of genetic drift. In addition to local adaptation, heterogeneous natural selection by diverse environments can cause non-local maladaptation. It results from conditionally deleterious mutations, which have either neutral or deleterious effects depending on environmental conditions (Mee et al. 2024; Mee and Yeaman 2019). Such mutations may accumulate in specific environments where they are neutral but are purged in others, potentially mimicking phenotypic patterns and allelic differentiation patterns of local adaptation (Mee and Yeaman 2019). Low recombination in pericentromeric regions may also promote the accumulation of conditionally deleterious mutations (Mee et al. 2024), supporting our hypothesis that selective signatures in these regions might result from non-local maladaptation. However, our current data and computational methods do not allow further testing of this hypothesis. In the future, the development of computational methods, such as machine-learning-based methods, to decouple population structure and selective signatures and also to disentangle local adaptation and non-local maladaptation in selfing species will be needed.

To enhance the use of barley genebank germplasm, we developed the *GEAplus* framework which recovers the missing geographical origins of accessions for GEA analyses (Chapter 3, Chang and Schmid 2023; Chapter 4). Since GEA is based on the association between genetic variation and sample origin environments, the framework includes an automatic data quality control step to exclude samples with error-prone geographical origin data. Traditionally, such data cleaning tasks rely on data visualization and manual identification of unusual patterns, but this approach is infeasible to handle tens of thousands of accessions in genebanks. Our machine-learning-based tool, *GGoutlieR* (Chang and Schmid 2023), discovers samples with unusual geo-genetic patterns that may result from seed exchange and documentation errors. Also, this tool can assist researchers in visualizing long-distance migrations if migration events occurred in recent history. However, the method has limitations in cases of complex migration histories, as it assumes samples predominantly follow an isolation-by-distance model.

Following the data cleaning, the *GEAplus* framework imputes missing geographical origins using fully connected neural networks trained on samples with known origins and then con-

ducts GEA scans. In brief, accessions with genomic data and known geographical origins are used as training data for neural networks to predict the unknown collection sites of accessions without clear passport data based on their genotypic data. Next, GEA scans are performed by using the environmental data procured from climatic databases according to geographical origins for all genotyped accessions, including those with imputed geographical data. Our cross-validation analysis indicated the neural networks had high prediction accuracy of geographical origins (mean $R^2 > 0.9$) in the global barley landrace collection, consistent with the findings in other species and simulations (Battey et al. 2020a). We further compared the regular GEA approach with the *GEAplus* framework using flowering time genes as benchmarks. Both approaches detected signatures near *PHOTOPERIOD-H1* (*PPD-H1*), whereas *GEAplus* also identified new GEA signatures in the upstream of *CENTRORADIALIS* (*CEN*), which has been suggested involved in local adaptation in multiple studies (Bustos-Korts et al. 2019; Comadran et al. 2012; Landis et al. 2024; Russell et al. 2016). However, only the standard GEA detected signatures near *VERNALIZATION-H1* (*VRN-H1*), which *GEAplus* did not capture. These different patterns of GEA signatures from the two approaches may reflect geographical constraints on allele diffusion. This finding suggests that larger sample sizes do not necessarily increase statistical power but may reveal different adaptive associations by integrating new relationships of environments and allele frequencies into GEA tests. Our simulation for local adaption also indicated that the extremely large sample size does not benefit the power of GEA.

5.2 Strategies for future crop adaptation research

In our studies, we employed GEA to identify adaptive loci in wild barley and landrace accessions. However, our findings highlighted the limitations of this approach, particularly its reliance on linear relationships between allele frequencies and environmental gradients (Lotterhos 2023), so it may overlook non-linear association patterns. Moreover, our results in wild barley presented confounding effects between environments and population genetic structure, which can lead to difficulties for discovery of adaptive loci. Given the capability of capturing complicated non-linear patterns, applying machine learning and deep learning algorithms to

GEA may enhance the power to detect adaptive loci and better distinguish the spurious selective signatures resulting from genetic drift.

Traditionally, population genetics uses model-based statistical approaches, often with model assumptions simplifying the realistic complexity of evolutionary processes to make computation feasible (Huang et al. 2024). As a drawback, the implementation of model-based approaches may be limited by simplified model assumptions, such as fixed population size and random mating within populations, making them less ideal for very complicated genetic and evolutionary scenarios. Additionally, advances in high-throughput sequencing technologies lead to exponential growth in genomic data, presenting new opportunities to investigate complex evolutionary processes, but also posing computational challenges. Traditional model-based approaches often struggle to handle such high-dimensional datasets, highlighting the need for novel computational tools capable of managing the scale and complexity of modern genomic data (Huang et al. 2024; Korfmann et al. 2023; Schrider and Kern 2018). To address these challenges, machine learning techniques have been increasingly applied in population genetics, leading to transitioning the field from a predominantly theory-based discipline to a more data-driven paradigm (Korfmann et al. 2023; Schrider and Kern 2018). Machine learning approaches have shown promising results, demonstrating equal or superior performance compared to traditional approaches (Schrider and Kern 2018), such as identification of selective signatures (Kern and Schrider 2018; Lin et al. 2011; Ronen et al. 2013; Schrider and Kern 2016). Machine learning algorithms can identify genomic regions under natural selection by recognizing intricate patterns from a variety of summary statistics (Kern and Schrider 2018; Lin et al. 2011; Schrider and Kern 2016) or site frequency spectrum (Ronen et al. 2013). Such capability of revealing hidden features from a vast amount of high-dimensional data makes machine learning have great potential in population genetics (Schrider and Kern 2018).

Moreover, deep learning, a rapidly evolving branch of machine learning, has gained vast interest across various research domains. Deep learning algorithms use neural networks comprising multiple hidden layers of non-linear transformations to automatically extract complex features from raw data (LeCun et al. 2015; Sapoval et al. 2022). The potential of deep learning in population genetics has been widely explored in recent years, including the identification of selection signatures due to its strength to discover subtle signatures that may not be cap-

tured by traditional summary statistics (Huang et al. 2024; Korfmann et al. 2023). For instance, *DeepGenomeScan* employs deep learning to model the non-linear relationship between genomic information and environmental data and then extracts the weight of each genomic region from neural networks for statistical tests to discover selection signatures (Qin et al. 2022). Deep learning algorithms, usually convolutional neural networks, can also be used as a classifier to distinguish genomic regions under selection from neutrality by learning genomic variation patterns from training data, such as *ImaGene* (Torada et al. 2019), *disc-pg-gan* (Riley et al. 2024), and *Timesweeper* (Whitehouse and Schrider 2023). Furthermore, as deep neural networks are capable of revealing complex patterns, applications of deep learning may accurately disentangle genetic drift and real GEA signatures resulting from selection for low-coverage sequencing datasets, such as the results reported in our study (Chang et al. 2022). However, deep learning models typically require large training datasets with known evolutionary parameters, which is usually unpragmatic in real-world studies (Huang et al. 2024).

Training datasets for deep-learning-based approaches are often generated by large-scale realistic simulations (Huang et al. 2024). Evolutionary simulations can be performed in two ways: backward-time and forward-time simulations. Backward approaches, such as *ms* (Hudson 2002) and *msprime* (Baumdicker et al. 2022), simulate evolutionary processes with a backward manner generation-by-generation based on coalescent. On the contrary, forward simulations, like *SLiM* (Haller and Messer 2019, 2023b), begin with initial populations and evolve them forward in time. The forward simulations offer great flexibility since it is capable of simulating every individual and mutation explicitly (Haller and Messer 2023b), but they are computationally much less efficient than the backward approaches. A hybrid approach that combines forward and backward simulations addresses this issue by using forward simulations to model quantitative trait loci (QTLs) under selection and backward simulations to overlay neutral mutations (Haller et al. 2019; Haller and Messer 2019). This hybrid strategy greatly reduces the required computational resources and computing time by excluding the simulation of neutral mutations, making the simulations of large samples computationally feasible. The success of hybrid simulations relies on the recent development of tree-sequence recording, allowing efficient storage of entire population histories in forward simulations (Kelleher et al. 2018).

On this basis, a recent study further demonstrated that graph convolutional neural networks

can directly learn features from tree sequences for population genetics inference (Whitehouse et al. 2024) This advance paves the way for more scalable model training for the identification of selection signatures in the future. Additionally, emerging self-supervised plant DNA language models, like *PlantCaduceus* (Zhai et al. 2024), may advance crop adaptation research by offering new methods for functional annotation and mutation effect prediction, even with limited training data.

5.3 Future of barley pre-breeding

Genetic diversity of breeding populations is one of the pivotal factors determining the success of breeding programs. The breeder's equation illustrates the genetic response to selection as $\Delta G = ih\sigma_g$ (Falconer 1996), which is a function of selection intensity (i), the square root of heritability (h), and the additive genetic variation within the breeding population (σ_g). This equation predicts that genetic improvement by selection can be constrained by limited genetic variance within a breeding population even if a strong selection intensity is applied. Given concerns over genetic diversity loss due to domestication and modern breeding practices (Bohra et al. 2022; Flint-Garcia et al. 2023), expanding genetic variation in breeding populations is vital for achieving continuous breeding success in the future.

5.3.1 Utility of environmental data of geographical origins

Pre-breeding is a crucial process to expand genetic diversity of breeding populations by bringing favorable alleles from exotic germplasm into breeding materials that can be directly used in breeding programs. This process starts with the identification of plant genetic resources carrying traits or alleles of interest. The focused identification of germplasm strategy (FIGS) leverages knowledge of the relationship between environmental conditions at geographical origins and plant traits to enhance the probability of discovering accessions with specific characteristics (Bohra et al. 2022; Endresen 2010; Sunitha et al. 2024). Traditionally, this strategy relies on the accurate geographical information of collection sites, but such data may be miss-

ing or inaccurate for accessions collected before the invention of GPS technology or due to documentation errors. To optimize the utility of genebank germplasm, those accessions with missing or error-prone GPS data should be included in the consideration rather than leaving them out. This gap can be addressed by machine learning and deep learning approaches. For data cleaning, machine learning offers a scalable approach to identify accessions with error-prone data, such as *GGoutliR* introduced in our study (Chapter 3, Chang and Schmid 2023). Next, this refined data can then serve as training data to impute the geographical origins of accessions with unclear collection sites, allowing for the further extraction of environmental data. Deep learning methods have shown impressive accuracy in predicting geographical origins by capturing complex relationships between geography and genomic variants for both empirical and simulated populations in Battey et al. (2020a) and also in our study (Chapter 4). The enriched geographical origin information from this approach can facilitate FIGS for harnessing the genetic diversity of genebank germplasm. Additionally, GEA analysis could indicate accessions potentially carrying adaptive alleles as donors if selection signatures are successfully identified. However, given the large genomic regions identified in our study (Chapter 2; Chang et al. 2022), the direct use of our current results in barley pre-breeding is limited.

Alternatively, the environmental data can be obtained from the available GPS information and used as proxies for phenotypes in genomic prediction, a technique of predicting traits of individuals from genome-wide markers. As such, the predicted environmental variables may reveal the environmental conditions that accessions adapt to, which can be used to assist the parent selection (Halpin-McCormick et al. 2024). The recent studies have demonstrated moderate to high accuracy in genomic prediction of individual environmental variables for the maize core collection of CIMMYT International Germplasm Bank (Li et al. 2024) and the USDA barley core collection (Halpin-McCormick et al. 2024). However, a limitation of this approach is that environmental variables may not always directly correlate with the traits of interest. This is evident in the poor performance of predicting yield-related traits of maize using environmental variables (Li et al. 2024). Therefore, instead of taking environmental information of accessions' origins as predictors, environmental data may be used to supplement genomic prediction by fitting variance-covariance matrix of environments as a non-additive term in genomic best linear unbiased prediction (GBLUP) models (Kehel et al. 2020) or considering environmental variables as covariates in a multivariate GBLUP model.

5.3.2 Accelerating pre-breeding with technological advances

Once the donors are selected from wild relatives or landraces, subsequent recurrent selection has to be performed to minimize linkage drag and develop ready-to-use breeding materials. Traditional multi-year recurrent selection can be accelerated with marker-assisted selection, genomic selection, speed breeding, and enhanced recombination (Bohra et al. 2022). Genomic selection is a breeding strategy making selections based on the genomic estimated breeding values (GEBVs) of complex quantitative traits that are predicted using genome-wide markers (Meuwissen et al. 2001). Genomic selection has been widely accepted and used in modern breeding programs because it enables breeding decisions to be made based solely on genotype, leading to cost savings from less phenotyping and more genetic gain per unit of time by shortening breeding cycles (Alemu et al. 2024). For pragmatic use, breeders may initiate a genomic selection pre-breeding program from the direct crosses between exotic donor accessions and elite lines to rapidly purge deleterious donor alleles (Gorjanc et al. 2016). However, such a strategy is more feasible for traits controlled by loci with large effect and less suitable for polygenic loci as allele frequencies may quickly shift toward elite alleles, resulting in the loss of donor alleles (Gorjanc et al. 2016).

Speed breeding is an approach to fasten plant development by growing plants in a greenhouse or a growth chamber with controlled temperature and photoperiod, enabling more generations per year (Watson et al. 2018). Since genomic selection allows breeding decisions without field trials, the integration of genomic selection with speed breeding can further improve the rate of genetic gain (Jähne et al. 2020; Watson et al. 2019). Furthermore, inducing crossover events across chromosomes during meiosis could facilitate pre-breeding by breaking linkages between favorable and undesirable alleles. More generations with speed breeding can result in a higher number of chromosome recombinations per year, which increases the probability of removing linkage drag and expanding the genetic variance of populations. In this regard, increasing the crossover number per generation further accelerates the progress of pre-breeding, which can be achieved by altering the expression of crossover regulator genes (Epstein et al. 2023). By integrating such mutants with genomic selection and speed breeding, the efficiency of pre-breeding can be greatly improved.

In addition, pre-breeding of barley could benefit from the remarkable advance in barley pan-genome assembly with chromosome-scale sequences of 76 accessions consisting of domesticated and wild barley (Jayakodi et al. 2024). With the high-quality pan-genome assembly, structural variants (SVs), such as translocation, present-absence variation, and copy number variation, can be discovered, improving the power of genome-wide association mapping for quantitative trait loci (QTLs) that may be unable to be detected by single nucleotide polymorphism (SNP) markers (Della Coletta et al. 2021; Schreiber et al. 2024; Zhou et al. 2022). Likewise, since SVs may contribute to phenotypic variation not associated with SNPs, incorporating SVs into genomic prediction has the potential to improve prediction accuracy (Della Coletta et al. 2021; Schreiber et al. 2024; Zhou et al. 2022). Also, SVs could be involved in local adaptation by contributing to traits that increase fitness in specific habitats. As our study solely used SNPs in GEA analyses, incorporating SVs from barley pan-genome assemblies in the future study may uncover overlooked adaptive loci.

Moreover, super pan-genome assemblies, comprising diverse species within a genus or a family, provide valuable insights for breeding purposes from evolutionary perspectives. By uncovering SVs highly divergent between domesticated species and wild relatives, researchers can target domestication genes for genetic engineering to improve cultivated varieties (Li et al. 2023; Schreiber et al. 2024; Tang et al. 2022). For instance, Li et al. (2023) identified a gene predominately existing in wild tomato and generated transgenic lines under cultivated tomato background, resulting in a higher number of fruit-bearing branches. Domestication genes discovered through pan-genomics can also facilitate "de novo domestication" (Della Coletta et al. 2021), a progress of introducing domestication-related traits into wild relatives with desirable characteristics in order to generate new crops. By using gene editing, desired genetic diversity in wild relatives can be rapidly accessible for breeding through de novo domestication (Zsögön et al. 2018). This methodology could be applied to wild barley to facilitate the introgression of adaptive traits into barley cultivars.

Lastly, with comparative genomics, super pan-genome may improve the accuracy of genomic selection by disclosing deleterious mutations related to traits of interest. The potential of this approach is evident in a recent potato study, which increased the prediction accuracy of yield by 24.7% via integrating deleterious burden into the GBLUP model (Wu et al. 2023). This

framework could be adapted to barley pre-breeding to effectively eliminate deleterious mutations that might be harmful to productivity. Overall, as genomic technologies, computational methods, and genome assemblies continue to advance, these innovations will pave the way for more efficient barley breeding programs.

Acknowledgements

First and foremost, I wish to express my deepest gratitude to my PhD supervisor, Prof. Dr. Karl Schmid. His invaluable advice, encouragement to explore new ideas, and insightful feedback on my manuscripts have greatly shaped my academic journey. His inspiring guidance and constant support have been crucial for my growth as a researcher.

I am sincerely grateful to the 350b lab members – Katharina Bondel, Fabian Freund, Max Haupt, Ali Baturaygil, Flavio Lozano-Isla, Sonja Kersten, Niharika Rakasi, and Anurag Daware – for their stimulating scientific discussions and collaboration over the years. Special thanks go to Barbel Hessenauer and Anja Oehlmann for their kind assistance with administrative and bureaucratic matters. I'd also like to express my gratitude to Elisabeth Kokai-Kota and Vanessa Haseneder for their expertise in DNA extraction and library preparation, and to Mireia Vidal for her bioinformatic analysis of raw DNA sequences. Viola Abraham's support in growing barley seedlings was invaluable to my research, and I deeply appreciate her help.

I am profoundly thankful to Marco Roelcke, my dear "old" friend, for his encouragement and support. He not only helped me address the bureaucratic challenges of extending my residence permit but also shared his knowledge of German history and politics. Besides, his assistance with German translations is very much appreciated.

Many thanks to all my colleagues at the institute, Thea, Xintian, Muhammad, Cleo, Sandra, Felicien, Khaoula, Rodrigo, Belen, Anna, Jan, Paul, Felix, Johannes, Lea, and all unmentioned colleagues, for accompanying me throughout this journey. Your companionship has made my PhD experience truly enjoyable and unforgettable.

I am also grateful to my coauthors, Prof. Eyal Fridman, Dr. Martin Mascher, and Dr. Axel Himmelbach, for their invaluable contributions to the wild barley study. Collaborating with them has been an honor, and their expertise greatly enriched this work.

Finally, I want to express my profound gratitude to my family in Taiwan for their unconditional love, support, and encouragement. Their belief in me has been my greatest strength throughout this journey.

Bibliography

Abdel-Ghani, A. H., Parzies, H. K., Omary, A. and Geiger, H. H. (2004). Estimating the outcrossing rate of barley landraces and wild barley populations collected from ecologically different regions of Jordan, *Theoretical and Applied Genetics* **109**: 588–595.

Agha, H. I., Endelman, J. B., Chitwood-Brown, J., Clough, M., Coombs, J., De Jong, W. S., Douches, D. S., Higgins, C. R., Holm, D. G., Novy, R. et al. (2024). Genotype-by-environment interactions and local adaptation shape selection in the US national chip processing trial, *Theoretical and Applied Genetics* **137**(5): 99.

Aguirre-Liguori, J. A., Ramírez-Barahona, S. and Gaut, B. S. (2021). The evolutionary genomics of species' responses to climate change, *Nature Ecology & Evolution* **5**(10): 1350–1360.

Akerman, A. and Bürger, R. (2014). The consequences of gene flow for local adaptation and differentiation: a two-locus two-deme model, *Journal of Mathematical Biology* **68**(5): 1135–1198.

Al-Asadi, H., Petkova, D., Stephens, M. and Novembre, J. (2019). Estimating recent migration and population-size surfaces, *PLoS Genetics* **15**(1): e1007908.

Alemu, A., Åstrand, J., Montesinos-Lopez, O. A., y Sanchez, J. I., Fernandez-Gonzalez, J., Tadesse, W., Vetukuri, R. R., Carlsson, A. S., Ceplitis, A., Crossa, J. et al. (2024). Genomic selection in plant breeding: Key factors shaping two decades of progress, *Molecular Plant* .

Alleaume-Benharira, M., Pen, I. and Ronce, O. (2006). Geographical patterns of adaptation within a species' range: interactions between drift and gene flow, *Journal of Evolutionary Biology* **19**(1): 203–215.

Arlé, E., Zizka, A., Keil, P., Winter, M., Essl, F., Knight, T., Weigelt, P., Jiménez-Muñoz, M. and Meyer, C. (2021). bRacatus: A method to estimate the accuracy and biogeographical status of georeferenced biological data, *Methods in Ecology and Evolution* **12**(9): 1609–1619.

Asfaw, Z. and Bothmer, R. v. (1990). Hybridization between landrace varieties of ethiopian barley (*hordeum vulgare ssp. vulgare*) and the progenitor of barley (*h. vulgare ssp. spontaneum*), *Hereditas* **112**(1): 57–64.

Badr, A., Sch, R., Rabey, H. E., Effgen, S., Ibrahim, H., Pozzi, C., Rohde, W. and Salamini, F. (2000). On the origin and domestication history of barley (*Hordeum vulgare*), *Molecular Biology and Evolution* **17**(4): 499–510.

Baker, H. G. (1967). Support for Baker's law—as a rule, *Evolution* **21**(4): 853–856.

Baker, K., Bayer, M., Cook, N., Dreißig, S., Dhillon, T., Russell, J., Hedley, P. E., Morris, J., Ramsay, L., Colas, I. et al. (2014). The low-recombining pericentromeric region of barley restricts gene diversity and evolution but not gene expression, *The Plant Journal* **79**(6): 981–992.

Battey, C. J., Ralph, P. L. and Kern, A. D. (2020a). Predicting geographic location from genetic variation with deep neural networks, *eLife* **9**: e54507.

Battey, C., Ralph, P. L. and Kern, A. D. (2020b). Space is the place: Effects of continuous spatial structure on analysis of population genetic data, *Genetics* **215**(1): 193–214.

Baum, M., Grando, S., Backes, G., Jahoor, A., Sabbagh, A. and Ceccarelli, S. (2003). QTLs for agronomic traits in the mediterranean environment identified in recombinant inbred lines of the cross' Arta'× *H. spontaneum* 41-1, *Theoretical and Applied Genetics* **107**(7): 1215–1225.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G. et al. (2022). Efficient ancestry and mutation simulation with msprime 1.0, *Genetics* **220**(3): iyab229.

Bedada, G., Westerbergh, A., Nevo, E., Korol, A. and Schmid, K. J. (2014). DNA sequence variation of wild barley *Hordeum spontaneum* (L.) across environmental gradients in Israel, *Heredity* **112**(6): 646–655.

Bernád, V., Al-Tamimi, N., Langan, P., Gillespie, G., Dempsey, T., Henchy, J., Harty, M., Ramsay, L., Houston, K., Macaulay, M. et al. (2024). Unlocking the genetic diversity and population structure of the newly introduced two-row spring European heritage barley collection (ExHIBiT), *Frontiers in Plant Science* **15**: 1268847.

Berner, D. and Roesti, M. (2017). Genomics of adaptive divergence with chromosome-scale heterogeneity in crossover rate, *Molecular Ecology* **26**(22): 6351–6369.

Bernett, J., Blumenthal, D. B., Grimm, D. G., Haselbeck, F., Joeres, R., Kalinina, O. V. and List, M. (2024). Guiding questions to avoid data leakage in biological machine learning applications, *Nature Methods* **21**(8): 1444–1453.

Bhatia, G., Patterson, N., Sankararaman, S. and Price, A. L. (2013). Estimating and interpreting FST: the impact of rare variants, *Genome Research* **23**(9): 1514–1521.

Bohra, A., Kilian, B., Sivasankar, S., Caccamo, M., Mba, C., McCouch, S. R. and Varshney, R. K. (2022). Reap the crop wild relatives for breeding future crops, *Trends in Biotechnology* **40**(4): 412–431.

Bouhlal, O., Visioni, A., Verma, R. P. S., Kandil, M., Gyawali, S., Capettini, F. and Sanchez-Garcia, M. (2022). CGIAR barley breeding toolbox: A diversity panel to facilitate breeding and genomic research in the developing world, *Frontiers in Plant Science* **13**: 1034322.

Bradburd, G. S., Coop, G. M. and Ralph, P. L. (2018). Inferring continuous and discrete population genetic structure across space, *Genetics* **210**(1): 33–52.

Bradburd, G. S., Ralph, P. L. and Coop, G. M. (2016). A spatial framework for understanding population structure and admixture, *PLoS Genetics* **12**(1): e1005703.

Brown, A., Zohary, D. and Nevo, E. (1978). Outcrossing rates and heterozygosity in natural populations of *hordeum spontaneum koch* in Israel, *Heredity* **41**(1): 49–62.

Browning, B. L., Zhou, Y. and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels, *The American Journal of Human Genetics* **103**(3): 338–348.

Bürger, R. and Akerman, A. (2011). The effects of linkage and gene flow on local adaptation: A two-locus continent–island model, *Theoretical Population Biology* **80**(4): 272–288.

Bustos-Korts, D., Dawson, I. K., Russell, J., Tondelli, A., Guerra, D., Ferrandi, C., Strozzi, F., Nicolazzi, E. L., Molnar-Lang, M., Ozkan, H. et al. (2019). Exome sequences and multi-environment field trials elucidate the genetic basis of adaptation in barley, *The Plant Journal* **99**(6): 1172–1191.

Cabreros, I. and Storey, J. D. (2019). A likelihood-free estimator of population structure bridging admixture models and principal components analysis, *Genetics* **212**(4): 1009–1029.

Caldwell, K. S., Russell, J., Langridge, P. and Powell, W. (2006). Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*, *Genetics* **172**(1): 557–567.

Capblancq, T., Luu, K., Blum, M. G. and Bazin, E. (2018). Evaluation of redundancy analysis to identify signatures of local adaptation, *Molecular Ecology Resources* **18**(6): 1223–1233.

Caye, K., Jumentier, B., Lepeule, J. and François, O. (2019). LFMM 2: fast and accurate inference of gene-environment associations in genome-wide studies, *Molecular Biology and Evolution* **36**(4): 852–860.

Chang, C.-W., Fridman, E., Mascher, M., Himmelbach, A. and Schmid, K. (2022). Physical geography, isolation by distance and environmental variables shape genomic variation of wild barley (*hordeum vulgare* l. ssp. *spontaneum*) in the Southern Levant, *Heredity* **128**(2): 107–119.

Chang, C.-W. and Schmid, K. (2023). GGoutlieR: an R package to identify and visualize unusual geo-genetic patterns of biological samples, *Journal of Open Source Software* **8**(91): 5687.

Cockram, J. and Mackay, I. (2018). Genetic Mapping Populations for Conducting High-Resolution Trait Mapping in Plants, *in* R. K. Varshney, M. K. Pandey and A. Chitikineni (eds), *Plant Genetics and Molecular Biology*, Vol. 164, Springer International Publishing, Cham, pp. 109–138.

Comadran, J., Kilian, B., Russell, J., Ramsay, L., Stein, N., Ganal, M., Shaw, P., Bayer, M., Thomas, W., Marshall, D. et al. (2012). Natural variation in a homolog of *antirrhinum centroradialis* contributed to spring growth habit and environmental adaptation in cultivated barley, *Nature Genetics* **44**(12): 1388–1392.

Contreras-Moreira, B., Serrano-Notivoli, R., Mohammed, N. E., Cantalapiedra, C. P., Beguería, S., Casas, A. M. and Igartua, E. (2019). Genetic association with high-resolution climate data reveals selection footprints in the genomes of barley landraces across the Iberian Peninsula, *Molecular Ecology* **28**(8): 1994–2012.

Coop, G., Witonsky, D., Di Rienzo, A. and Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation, *Genetics* **185**(4): 1411–1423.

Dai, F., Nevo, E., Wu, D., Comadran, J., Zhou, M., Qiu, L., Chen, Z., Beiles, A., Chen, G. and Zhang, G. (2012). Tibet is one of the centers of domestication of cultivated barley, *Proceedings of the National Academy of Sciences* **109**(42): 16969–16973.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T. et al. (2011). The variant call format and VCFtools, *Bioinformatics* **27**(15): 2156–2158.

Darwin, C. (1859). On the origin of species.

Dawson, I. K., Russell, J., Powell, W., Steffenson, B., Thomas, W. T. and Waugh, R. (2015). Barley: a translational model for adaptation to climate change, *New Phytologist* **206**(3): 913–931.

de Villemereuil, P., Frichot, É., Bazin, É., François, O. and Gaggiotti, O. E. (2014). Genome scan methods against more complex models: when and how much should we trust them?, *Molecular Ecology* **23**(8): 2006–2019.

De Villemereuil, P. and Gaggiotti, O. E. (2015). A new FST-based method to uncover local adaptation using environmental variables, *Methods in Ecology and Evolution* **6**(11): 1248–1258.

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M. B. and Hirsch, C. N. (2021). How the pan-genome is changing crop genomics and improvement, *Genome Biology* **22**: 1–19.

Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C. and Guarino, L. (2017). Past and future use of wild relatives in crop breeding, *Crop Science* **57**(3): 1070–1082.

Doebley, J. F., Gaut, B. S. and Smith, B. D. (2006). The molecular genetics of crop domestication, *Cell* **127**(7): 1309–1321.

Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N. and Wagner, H. H. (2019). *adespatial: Multivariate Multiscale Spatial Analysis*. R package version 0.3-7.
URL: *https://CRAN.R-project.org/package=adespatial*

Dray, S., Legendre, P. and Peres-Neto, P. R. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM), *Ecological Modelling* **196**(3-4): 483–493.

Dwivedi, S. L., Ceccarelli, S., Blair, M. W., Upadhyaya, H. D., Are, A. K. and Ortiz, R. (2016). Landrace germplasm for improving yield and abiotic stress adaptation, *Trends in Plant Science* **21**(1): 31–42.

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

Endresen, D. T. F. (2010). Predictive association between trait data and ecogeographic data for Nordic barley landraces, *Crop Science* **50**(6): 2418–2430.

Epstein, R., Sajai, N., Zelkowski, M., Zhou, A., Robbins, K. R. and Pawlowski, W. P. (2023). Exploring impact of recombination landscapes on breeding outcomes, *Proceedings of the National Academy of Sciences* **120**(14): e2205785119.

Excoffier, L., Hofer, T. and Foll, M. (2009). Detecting loci under selection in a hierarchically structured population, *Heredity* **103**(4): 285–298.

Falconer, D. S. (1996). *Introduction to quantitative genetics*, Pearson Education India.

Fang, Z., Gonzales, A. M., Clegg, M. T., Smith, K. P., Muehlbauer, G. J., Steffenson, B. J. and Morrell, P. L. (2014). Two genomic regions contribute disproportionately to geographic differentiation in wild barley, *G3: Genes, Genomes, Genetics* **4**(7): 1193–1203.

FAO (2022). Crops and livestock products.
URL: *https://www.fao.org/faostat/*

Fernández-Calleja, M., Casas, A. M. and Igartua, E. (2021). Major flowering time genes of barley: allelic diversity, effects, and comparison with wheat, *Theoretical and Applied Genetics* **134**: 1867–1897.

Ferrari, T., Feng, S., Zhang, X. and Mooney, J. A. (2024). Towards simulation optimization: An examination of the impact of scaling on coalescent and forward simulations, *bioRxiv* pp. 2024–04.

Fick, S. E. and Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, *International Journal of Climatology* **37**(12): 4302–4315.

Flint-Garcia, S., Feldmann, M. J., Dempewolf, H., Morrell, P. L. and Ross-Ibarra, J. (2023). Diamonds in the not-so-rough: Wild relative diversity hidden in crop genomes, *PLoS Biology* **21**(7): e3002235.

Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L. and Lasky, J. R. (2016). Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes, *Molecular Ecology* **25**(1): 104–120.

Forester, B. R., Lasky, J. R., Wagner, H. H. and Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations, *Molecular Ecology* **27**(9): 2215–2233.

François, O., Martins, H., Caye, K. and Schoville, S. D. (2016). Controlling false discoveries in genome scans for selection, *Molecular Ecology* **25**(2): 454–469.

Frichot, E., Schoville, S. D., Bouchard, G. and François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models, *Molecular Biology and Evolution* **30**(7): 1687–1699.

Galkin, E., Dalal, A., Evenko, A., Fridman, E., Kan, I., Wallach, R. and Moshelion, M. (2018). Risk-management strategies and transpiration rates of wild barley in uncertain environments, *Physiologia Plantarum* **164**(4): 412–428.

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates, *Genetics* **201**(4): 1555–1579.

Gibson, M. J. and Moyle, L. C. (2020). Regional differences in the abiotic environment contribute to genomic divergence within a wild tomato species, *Molecular Ecology* **29**(12): 2204–2217.

Gol, L., Haraldsson, E. B. and von Korff, M. (2021). Ppd-H1 integrates drought stress signals to control spike development and flowering time in barley, *Journal of Experimental Botany* **72**(1): 122–136.

Gorjanc, G., Jenko, J., Hearne, S. J. and Hickey, J. M. (2016). Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations, *BMC Genomics* **17**: 1–15.

Guillot, G., Jónsson, H., Hinge, A., Manchih, N. and Orlando, L. (2016). Accurate continuous geographic assignment from low-to high-density SNP data, *Bioinformatics* **32**(7): 1106–1108.

Günther, T. and Coop, G. (2013). Robust identification of local adaptation from allele frequencies, *Genetics* **195**(1): 205–220.

Gutaker, R. M., Groen, S. C., Bellis, E. S., Choi, J. Y., Pires, I. S., Bocinsky, R. K., Slayton, E. R., Wilkins, O., Castillo, C. C., Negrão, S. et al. (2020). Genomic history and ecology of the geographic spread of rice, *Nature Plants* **6**(5): 492–502.

Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W. and Ralph, P. L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes, *Molecular Ecology Resources* **19**(2): 552–566.

Haller, B. C. and Messer, P. W. (2019). SLiM 3: forward genetic simulations beyond the Wright–Fisher model, *Molecular Biology and Evolution* **36**(3): 632–637.

Haller, B. C. and Messer, P. W. (2023a). SLiM 4: Multispecies eco-evolutionary modeling, *The American Naturalist* **201**(5): E127–E139.

Haller, B. C. and Messer, P. W. (2023b). SLiM 4: multispecies eco-evolutionary modeling, *The American Naturalist* **201**(5): E127–E139.

Haller, B. and Messer, P. (2016). SLiM: an evolutionary simulation framework, *URL: http://benhaller. com/slim/SLiM_Manual. pdf* .

Halpin-McCormick, A., Campbell, Q., Negrao, S., Morrell, P. L., Hubner, S., Neyhart, J. and Kantar, M. B. (2024). Back to the future: Environmental genomic selection to take advantage of polygenic local adaptation, *bioRxiv* pp. 2024–10.

Hämälä, T. and Savolainen, O. (2019). Genomic patterns of local adaptation under gene flow in *Arabidopsis lyrata*, *Molecular Biology and Evolution* **36**(11): 2557–2571.

Han, E., Sinsheimer, J. S. and Novembre, J. (2014). Characterizing bias in population genetic inferences from low-coverage sequencing data, *Molecular Biology and Evolution* **31**(3): 723–735.

Harlan, J. R. and Zohary, D. (1966). Distribution of wild wheats and barley, *Science* **153**(3740): 1074–1080.

Hartfield, M., Bataillon, T. and Glémin, S. (2017). The evolutionary interplay between adaptation and self-fertilization, *Trends in Genetics* **33**(6): 420–431.

Hendrick, M. F., Finseth, F. R., Mathiasson, M. E., Palmer, K. A., Broder, E. M., Breigenzer, P. and Fishman, L. (2016). The genetics of extreme microgeographic adaptation: an integrated approach identifies a major gene underlying leaf trichome divergence in Yellowstone *Mimulus guttatus*, *Molecular Ecology* **25**(22): 5647–5662.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B. et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning, *PLoS One* **12**(2): e0169748.

Hereford, J. (2009). A quantitative survey of local adaptation and fitness trade-offs, *The American Naturalist* **173**(5): 579–588.

Herzig, P., Maurer, A., Draba, V., Sharma, R., Draicchio, F., Bull, H., Milne, L., Thomas, W. T. B., Flavell, A. J. and Pillen, K. (2018). Contrasting genetic regulation of plant development in wild barley grown in two European environments revealed by nested association mapping, *Journal of Experimental Botany* **69**(7): 1517–1531. Publisher: Oxford Academic.

Hijmans, R. J. (2018). *raster: Geographic Data Analysis and Modeling*. R package version 2.7-15.
URL: *https://CRAN.R-project.org/package=raster*

Hijmans, R. J. (2019). *geosphere: Spherical Trigonometry*. R package version 1.5-10.
URL: *https://CRAN.R-project.org/package=geosphere*

Hill, W. and Weir, B. (1988). Variances and covariances of squared linkage disequilibria in finite populations, *Theoretical Population Biology* **33**(1): 54–78.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A. and Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions, *The American Naturalist* **188**(4): 379–397.

Hodgins, K. A. and Yeaman, S. (2019). Mating system impacts the genetic architecture of adaptation to heterogeneous environments, *New Phytologist* **224**(3): 1201–1214.

House, G. L. and Hahn, M. W. (2018). Evaluating methods to visualize patterns of genetic differentiation on a landscape, *Molecular Ecology Resources* **18**(3): 448–460.

Huang, X., Huang, S., Han, B. and Li, J. (2022). The integrated genomics of crop domestication and breeding, *Cell* .

Huang, X., Rymbekova, A., Dolgova, O., Lao, O. and Kuhlwilm, M. (2024). Harnessing deep learning for population genetic inference, *Nature Reviews Genetics* **25**(1): 61–78.

Hübner, S., Bdolach, E., Ein-Gedy, S., Schmid, K., Korol, A. and Fridman, E. (2013). Phenotypic landscapes: phenological patterns in wild and cultivated barley, *Journal of Evolutionary Biology* **26**(1): 163–174.

Hübner, S., Günther, T., Flavell, A., Fridman, E., Graner, A., Korol, A. and Schmid, K. J. (2012). Islands and streams: clusters and gene flow in wild barley populations from the Levant, *Molecular Ecology* **21**(5): 1115–1129.

Hübner, S., Höffken, M., Oren, E., Haseneyer, G., Stein, N., Graner, A., Schmid, K. and Fridman, E. (2009). Strong correlation of wild barley (*Hordeum spontaneum*) population structure with temperature and precipitation variation, *Molecular Ecology* **18**(7): 1523–1536.

Hübner, S., Korol, A. B. and Schmid, K. J. (2015). RNA-seq analysis identifies genes associated with differential reproductive success under drought-stress in accessions of wild barley hordeum spontaneum, *BMC Plant Biology* **15**(1): 1–14.

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation, *Bioinformatics* **18**(2): 337–338.

Jähne, F., Hahn, V., Würschum, T. and Leiser, W. L. (2020). Speed breeding short-day crops by LED-controlled light schemes, *Theoretical and Applied Genetics* **133**(8): 2335–2342.

Jakob, S. S., Rödder, D., Engler, J. O., Shaaf, S., Özkan, H., Blattner, F. R. and Kilian, B. (2014). Evolutionary history of wild barley (*Hordeum vulgare* subsp. *spontaneum*) analyzed using multilocus sequence data and paleodistribution modeling, *Genome Biology and Evolution* **6**(3): 685–702.

Jayakodi, M., Lu, Q., Pidon, H., Rabanus-Wallace, M. T., Bayer, M., Lux, T., Guo, Y., Jaegle, B., Badea, A., Bekele, W. et al. (2024). Structural variation in the pangenome of wild and domesticated barley, *Nature* pp. 1–9.

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V. S., Gundlach, H., Monat, C., Lux, T., Kamal, N., Lang, D., Himmelbach, A., Ens, J., Zhang, X.-Q., Angessa, T. T., Zhou, G., Tan, C., Hill, C., Wang, P., Schreiber, M., Boston, L. B., Plott, C., Jenkins, J., Guo, Y., Fiebig, A., Budak, H., Xu, D., Zhang, J., Wang, C., Grimwood, J., Schmutz, J., Guo, G., Zhang, G., Mochida, K., Hirayama, T., Sato, K., Chalmers, K. J., Langridge, P., Waugh, R., Pozniak, C. J., Scholz, U., Mayer, K. F. X., Spannagl, M., Li, C., Mascher, M. and Stein, N. (2020). The barley pan-genome reveals the hidden legacy of mutation breeding, *Nature* pp. 1–6. Publisher: Nature Publishing Group.

Joost, S., Bonin, A., Bruford, M. W., Després, L., Conord, C., Erhardt, G. and Taberlet, P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation, *Molecular Ecology* **16**(18): 3955–3969.

Jost, M., Singh, D., Lagudah, E., Park, R. F. and Dracatos, P. (2020). Fine mapping of leaf rust resistance gene Rph13 from wild barley, *Theoretical and Applied Genetics* **133**: 1887–1895.

Kalladan, R., Worch, S., Rolletschek, H., Harshavardhan, V. T., Kuntze, L., Seiler, C., Sreenivasulu, N. and Röder, M. S. (2013). Identification of quantitative trait loci contributing to yield and seed quality parameters under terminal drought in barley advanced backcross lines, *Molecular Breeding* **32**: 71–90.

Kawecki, T. J. and Ebert, D. (2004). Conceptual issues in local adaptation, *Ecology Letters* **7**(12): 1225–1241.

Kehel, Z., Sanchez-Garcia, M., El Baouchi, A., Aberkane, H., Tsivelikas, A., Charles, C. and Amri, A. (2020). Predictive characterization for seed morphometric traits for genebank accessions using genomic selection, *Frontiers in Ecology and Evolution* **8**: 32.

Kelleher, J., Thornton, K. R., Ashander, J. and Ralph, P. L. (2018). Efficient pedigree recording for fast population genetics simulation, *PLoS Computational Biology* **14**(11): e1006581.

Kern, A. D. and Schrider, D. R. (2018). diploS/HIC: an updated approach to classifying selective sweeps, *G3: Genes, Genomes, Genetics* **8**(6): 1959–1970.

Khoury, C. K., Brush, S., Costich, D. E., Curry, H. A., de Haan, S., Engels, J. M., Guarino, L., Hoban, S., Mercer, K. L., Miller, A. J. et al. (2022). Crop genetic erosion: understanding and responding to loss of crop diversity, *New Phytologist* **233**(1): 84–118.

Kilian, B., Özkan, H., Kohl, J., von Haeseler, A., Barale, F., Deusch, O., Brandolini, A., Yucel, C., Martin, W. and Salamini, F. (2006). Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication, *Molecular Genetics and Genomics* **276**(3): 230–241.

Kimura, M. (1979). The neutral theory of molecular evolution, *Scientific American* **241**(5): 98–129.

Kimura, M. et al. (1968). Evolutionary rate at the molecular level, *Nature* **217**(5129): 624–626.

König, P., Beier, S., Basterrechea, M., Schüler, D., Arend, D., Mascher, M., Stein, N., Scholz, U. and Lange, M. (2020). BRIDGE–a visual analytics web tool for barley genebank genomics, *Frontiers in Plant Science* **11**: 701.

Korfmann, K., Gaggiotti, O. E. and Fumagalli, M. (2023). Deep learning in population genetics, *Genome Biology and Evolution* **15**(2): evad008.

Kumar, A., Verma, R. P. S., Singh, A., Sharma, H. K. and Devi, G. (2020). Barley landraces: Ecological heritage for edaphic stress adaptations and sustainable production, *Environmental and Sustainability Indicators* **6**: 100035.

Lakew, B., Henry, R. J., Ceccarelli, S., Grando, S., Eglinton, J. and Baum, M. (2013). Genetic analysis and phenotypic associations for drought tolerance in *hordeum spontaneum* introgression lines using SSR and SNP markers, *Euphytica* **189**: 9–29.

Lampei, C., Wunder, J., Wilhalm, T. and Schmid, K. J. (2019). Microclimate predicts frost hardiness of alpine *Arabidopsis thaliana* populations better than elevation, *Ecology and Evolution* **9**(23): 13017–13029.

Landis, J. B., Guercio, A. M., Brown, K. E., Fiscus, C. J., Morrell, P. L. and Koenig, D. (2024). Natural selection drives emergent genetic homogeneity in a century-scale experiment with barley, *Science* **385**(6705): eadl0038.

Lasky, J. R., Des Marais, D. L., McKAY, J. K., Richards, J. H., Juenger, T. E. and Keitt, T. H. (2012). Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate, *Molecular Ecology* **21**(22): 5512–5529.

Lasky, J. R., Josephs, E. B. and Morris, G. P. (2023). Genotype–environment associations to reveal the molecular basis of environmental adaptation, *The Plant Cell* **35**(1): 125–138.

Lasky, J. R., Upadhyaya, H. D., Ramu, P., Deshpande, S., Hash, C. T., Bonnette, J., Juenger, T. E., Hyma, K., Acharya, C., Mitchell, S. E. et al. (2015). Genome-environment associations in sorghum landraces predict adaptive traits, *Science Advances* **1**(6): e1400218.

Lawson, D. J., Van Dorp, L. and Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots, *Nature Communications* **9**(1): 1–11.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning, *Nature* **521**(7553): 436–444.

Lee, C.-R. and Mitchell-Olds, T. (2011). Quantifying effects of environmental and geographical factors on patterns of genetic differentiation, *Molecular Ecology* **20**(22): 4631–4642.

Leek, J. T. (2011). Asymptotic conditional singular value decomposition for high-dimensional genomic data, *Biometrics* **67**(2): 344–352.

Legendre, P. and Legendre, L. (2012). Canonical analysis, *Numerical ecology, 3rd English Edition*, The Netherlands: Elsevier Science BV, chapter 11, pp. 625–710.

Lewontin, R. C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms, *Genetics* **74**(1): 175–195.

Li, F., Gates, D. J., Buckler, E. S., Hufford, M. B., Janzen, G. M., Rellán-Álvarez, R., Rodríguez-Zapata, F., Navarro, J. A. R., Sawers, R. J., Snodgrass, S. J. et al. (2024). The utility of environmental data from traditional varieties for climate-adaptive maize breeding, *bioRxiv* pp. 2024–09.

Li, J., Chen, G.-B., Rasheed, A., Li, D., Sonder, K., Zavala Espinosa, C., Wang, J., Costich, D. E., Schnable, P. S., Hearne, S. J. et al. (2019). Identifying loci with breeding potential across temperate and tropical adaptation via EigenGWAS and EnvGWAS, *Molecular Ecology* **28**(15): 3544–3560.

Li, K., Ren, X., Song, X., Li, X., Zhou, Y., Harlev, E., Sun, D. and Nevo, E. (2020). Incipient sympatric speciation in wild barley caused by geological-edaphic divergence, *Life Science Alliance* **3**(12).

Li, N., He, Q., Wang, J., Wang, B., Zhao, J., Huang, S., Yang, T., Tang, Y., Yang, S., Aisimutuola, P. et al. (2023). Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species, *Nature Genetics* **55**(5): 852–860.

Lin, K., Li, H., Schlotterer, C. and Futschik, A. (2011). Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics, *Genetics* **187**(1): 229–244.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., Shi, M. et al. (2020). Pan-genome of wild and cultivated soybeans, *Cell* **182**(1): 162–176.

Longin, C. F. H. and Reif, J. C. (2014). Redesigning the exploitation of wheat genetic resources, *Trends in Plant Science* **19**(10): 631–636.

López-Goldar, X. and Agrawal, A. A. (2021). Ecological interactions, environmental gradients, and gene flow in local adaptation, *Trends in Plant Science* **0**(0). Publisher: Elsevier.

Lotterhos, K. E. (2019). The effect of neutral recombination variation on genome scans for selection, *G3: Genes, Genomes, Genetics* **9**(6): 1851–1867.

Lotterhos, K. E. (2023). The paradox of adaptive trait clines with nonclinal patterns in the underlying genes, *Proceedings of the National Academy of Sciences* **120**(12): e2220313120.

Lotterhos, K. E. and Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests, *Molecular Ecology* **23**(9): 2178–2192.

Lotterhos, K. E. and Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method, *Molecular Ecology* **24**(5): 1031–1046.

Lundgren, E. and Ralph, P. L. (2019). Are populations like a circuit? Comparing isolation by resistance to a new coalescent-based method, *Molecular Ecology Resources* **19**(6): 1388–1406.

Luu, K., Bazin, E. and Blum, M. G. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis, *Molecular Ecology Resources* **17**(1): 67–77.

Lynch, J. P., Chimungu, J. G. and Brown, K. M. (2014). Root anatomical phenes associated with water acquisition from drying soil: targets for crop improvement, *Journal of Experimental Botany* **65**(21): 6155–6166.

Makowski, D., Ben-Shachar, M. and Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework, *Journal of Open Source Software* **4**(40): 1541.

Marok, M. A., Marok-Alim, D. and Rey, P. (2021). Contribution of functional genomics to identify the genetic basis of water-deficit tolerance in barley and the related molecular mechanisms, *Journal of Agronomy and Crop Science* **207**(6): 913–935.

Mascher, M. (2019). Pseudomolecules and annotation of the second version of the reference genome sequence assembly of barley cv. morex [morex v2].
URL: *https://doi.ipk-gatersleben.de:443/DOI/83e8e186-dc4b-47f7-a820-28ad37cb176b/d1067eba-1d08-42e2-85ec-66bfd5112cd8/2*

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P. E., Russell, J. et al. (2017). A chromosome conformation capture ordered sequence of the barley genome, *Nature* **544**(7651): 427–433.

Mascher, M., Schreiber, M., Scholz, U., Graner, A., Reif, J. C. and Stein, N. (2019). Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding, *Nature Genetics* **51**(7): 1076–1081.

Mascher, M., Schuenemann, V. J., Davidovich, U., Marom, N., Himmelbach, A., Hübner, S., Korol, A., David, M., Reiter, E., Riehl, S. et al. (2016). Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley, *Nature Genetics* **48**(9): 1089–1093.

Mascher, M., Wicker, T., Jenkins, J., Plott, C., Lux, T., Koh, C. S., Ens, J., Gundlach, H., Boston, L. B., Tulpová, Z. et al. (2021). Long-read sequence assembly: a technical evaluation in barley, *The Plant Cell* **33**(6): 1888–1906.

Matz, M. V., Treml, E. A. and Haller, B. C. (2020). Estimating the potential for coral adaptation to global warming across the Indo-West Pacific, *Global Change Biology* **26**(6): 3473–3481.

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O Dushlaine, C., Barber, M., Boutkov, B. et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits, *Nature Genetics* **53**(7): 1097–1103.

McCouch, S., Baute, G. J., Bradeen, J., Bramel, P., Bretting, P. K., Buckler, E., Burke, J. M., Charest, D., Cloutier, S., Cole, G. et al. (2013). Agriculture: feeding the future, *Nature* **499**(7456): 23.

McVean, G. (2009). A genealogical interpretation of principal components analysis, *PLoS Genetics* **5**(10).

Mee, J. A., Carson, B. and Yeaman, S. (2024). Conditionally deleterious mutation load accumulates in genomic islands of local adaptation but can be purged with sufficient genotypic redundancy, *The American Naturalist* **204**(1): 000–000.

Mee, J. A. and Yeaman, S. (2019). Unpacking conditional neutrality: genomic signatures of selection on conditionally beneficial and conditionally deleterious mutations, *The American Naturalist* **194**(4): 529–540.

Meuwissen, T. H., Hayes, B. J. and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps, *Genetics* **157**(4): 1819–1829.

Milner, S. G., Jost, M., Taketa, S., Mazón, E. R., Himmelbach, A., Oppermann, M., Weise, S., Knüpffer, H., Basterrechea, M., König, P. et al. (2019). Genebank genomics highlights the diversity of a global barley collection, *Nature Genetics* **51**(2): 319–326.

Molina-Cano, J. L., Moralejo, M., Igartua, E. and Romagosa, I. (1999). Further evidence supporting Morocco as a centre of origin of barley, *Theoretical and Applied Genetics* **98**: 913–918.

Monteagudo, A., Casas, A. M., Cantalapiedra, C. P., Contreras-Moreira, B., Gracia, M. P. and Igartua, E. (2019). Harnessing novel diversity from landraces to improve an elite barley variety, *Frontiers in Plant Science* **10**: 434.

Morrell, P. L. and Clegg, M. T. (2007). Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the fertile crescent, *Proceedings of the National Academy of Sciences* **104**(9): 3289–3294.

Morrell, P. L., Toleno, D. M., Lundy, K. E. and Clegg, M. T. (2005). Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization, *Proceedings of the National Academy of Sciences* **102**(7): 2442–2447.

Navarro, J. A. R., Willcox, M., Burgueño, J., Romay, C., Swarts, K., Trachsel, S., Preciado, E., Terron, A., Delgado, H. V., Vidal, V. et al. (2017). A study of allelic diversity underlying flowering-time adaptation in maize landraces, *Nature Genetics* **49**(3): 476.

Nevo, E., Beharav, A., Meyer, R., Hackett, C., Forster, B., Russell, J. and Powell, W. (2005). Genomic microsatellite adaptive divergence of wild barley by microclimatic stress in 'Evolution Canyon', Israel, *Biological Journal of the Linnean Society* **84**(2): 205–224.

Nevo, E., Zohary, D., Brown, A. and Haber, M. (1979). Genetic diversity and environmental associations of wild barley, *Hordeum spontaneum*, in Israel, *Evolution* pp. 815–833.

Nice, L. M., Steffenson, B. J., Blake, T. K., Horsley, R. D., Smith, K. P. and Muehlbauer, G. J. (2017). Mapping agronomic traits in a wild barley advanced backcross–nested association mapping population, *Crop Science* **57**(3): 1199–1210.

Nice, L. M., Steffenson, B. J., Brown-Guedira, G. L., Akhunov, E. D., Liu, C., Kono, T. J., Morrell, P. L., Blake, T. K., Horsley, R. D., Smith, K. P. et al. (2016). Development and genetic characterization of an advanced backcross-nested association mapping (AB-NAM) population of wild× cultivated barley, *Genetics* **203**(3): 1453–1467.

North, A., Pennanen, J., Ovaskainen, O. and Laine, A.-L. (2011). Local adaptation in a changing world: The roles of gene-flow, mutation, and sexual reproduction, *Evolution* **65**(1): 79–89.

Ochagavía, H., Kiss, T., Karsai, I., Casas, A. M. and Igartua, E. (2022). Responses of barley to high ambient temperature are modulated by vernalization, *Frontiers in Plant Science* **12**: 776982.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution, *Nature* **246**(5428): 96–98.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. and Wagner, H. (2019). *vegan: Community Ecology Package*. R package version 2.5-6.
URL: *https://CRAN.R-project.org/package=vegan*

Orabi, J., Backes, G., Wolday, A., Yahyaoui, A. and Jahoor, A. (2007). The Horn of Africa as a centre of barley diversification and a potential domestication site, *Theoretical and Applied Genetics* **114**: 1117–1127.

Pan, Y., Zhu, J., Hong, Y., Zhang, M., Lv, C., Guo, B., Shen, H., Xu, X. and Xu, R. (2021). Identification of novel QTL contributing to barley yellow mosaic resistance in wild barley (*hordeum vulgare spp. spontaneum*), *BMC Plant Biology* **21**: 1–11.

Panigrahi, M., Rajawat, D., Nayak, S. S., Ghildiyal, K., Sharma, A., Jain, K., Lei, C., Bhushan, B., Mishra, B. P. and Dutt, T. (2023). Landmarks in the history of selective sweeps, *Animal Genetics* **54**(6): 667–688.

Pankin, A., Altmüller, J., Becker, C. and von Korff, M. (2018). Targeted resequencing reveals genomic signatures of barley domestication, *New Phytologist* **218**(3): 1247–1259.

Pankin, A. and 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.

Pavlidis, P. and Alachiotis, N. (2017). A survey of methods and tools to detect recent and strong positive selection, *Journal of Biological Research-Thessaloniki* **24**: 1–17.

Pekel, J.-F., Cottam, A., Gorelick, N. and Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes, *Nature* **540**(7633): 418–422.

Pembleton, L., Cogan, N. and Forster, J. (2013). StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations, *Molecular Ecology Resources* **13**: 946–952.

Peterman, W. E. (2018). ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms, *Methods in Ecology and Evolution* **9**(6): 1638–1647.

Petkova, D., Novembre, J. and Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces, *Nature Genetics* **48**(1): 94.

Pham, A.-T., Maurer, A., Pillen, K., Brien, C., Dowling, K., Berger, B., Eglinton, J. K. and March, T. J. (2019). Genome-wide association of barley plant growth under drought stress using a nested association mapping population, *BMC Plant Biology* **19**(1): 134.

Pham, A.-T., Maurer, A., Pillen, K., Nguyen, T. D., Taylor, J., Coventry, S., Eglinton, J. K. and March, T. J. (2024). A wild barley nested association mapping population shows a

wide variation for yield-associated traits to be used for breeding in Australian environment, *Euphytica* **220**(2): 24.

Poets, A. M., Fang, Z., Clegg, M. T. and Morrell, P. L. (2015). Barley landraces are characterized by geographically heterogeneous genomic origins, *Genome Biology* **16**: 1–11.

Poland, J. A., Brown, P. J., Sorrells, M. E. and Jannink, J.-L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach, *PloS One* **7**(2): e32253.

Pourkheirandish, M., Hensel, G., Kilian, B., Senthil, N., Chen, G., Sameri, M., Azhaguvel, P., Sakuma, S., Dhanagond, S., Sharma, R. et al. (2015). Evolution of the grain dispersal system in barley, *Cell* **162**(3): 527–539.

Pourkheirandish, M. and Komatsuda, T. (2007). The importance of barley genetics and domestication in a global perspective, *Annals of Botany* **100**(5): 999–1008.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data, *Genetics* **155**(2): 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses, *The American Journal of Human Genetics* **81**(3): 559–575.

Purugganan, M. D. (2022). What is domestication?, *Trends in Ecology & Evolution* **37**(8): 663–671.

Pyhäjärvi, T., Hufford, M. B., Mezmouk, S. and Ross-Ibarra, J. (2013). Complex patterns of local adaptation in teosinte, *Genome Biology and Evolution* **5**(9): 1594–1609.

Qin, X., Chiang, C. W. and Gaggiotti, O. E. (2022). Deciphering signatures of natural selection via deep learning, *Briefings in Bioinformatics* **23**(5): bbac354.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M. and Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics, *Molecular Ecology* **24**(17): 4348–4370.

Renaut, S., Grassa, C., Yeaman, S., Moyers, B., Lai, Z., Kane, N., Bowers, J., Burke, J. and Rieseberg, L. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers, *Nature Communications* **4**(1): 1–8.

Riley, R., Mathieson, I. and Mathieson, S. (2024). Interpreting generative adversarial networks to infer natural selection from genetic data, *Genetics* **226**(4): iyae024.

Ronen, R., Udpa, N., Halperin, E. and Bafna, V. (2013). Learning natural selection from the site frequency spectrum, *Genetics* **195**(1): 181–193.

Ruge-Wehling, B. and Wehling, P. (2014). The secondary gene pool of barley (*hordeum bulbosum*): Gene introgression and homoeologous recombination, *Biotechnological Approaches to Barley Improvement*, Springer, pp. 331–343.

Russell, J., Dawson, I. K., Flavell, A. J., Steffenson, B., Weltzien, E., Booth, A., Ceccarelli, S., Grando, S. and Waugh, R. (2011). Analysis of> 1000 single nucleotide polymorphisms in geographically matched samples of landrace and wild barley indicates secondary contact and chromosome-level differences in diversity around domestication genes, *New Phytologist* **191**(2): 564–578.

Russell, J., Mascher, M., Dawson, I. K., Kyriakidis, S., Calixto, C., Freund, F., Bayer, M., Milne, I., Marshall-Griffiths, T., Heinen, S. et al. (2016). Exome sequencing of geographically diverse barley landraces and wild relatives gives insights into environmental adaptation, *Nature Genetics* **48**(9): 1024–1030.

Saad, N. S. M., Neik, T. X., Thomas, W. J., Amas, J. C., Cantila, A. Y., Craig, R. J., Edwards, D. and Batley, J. (2022). Advancing designer crops for climate resilience through an integrated genomics approach, *Current Opinion in Plant Biology* **67**: 102220.

Sallam, A. H., Guo, Y., Jayakodi, M., Himmelbach, A., Fiebig, A., Simmons, J., Bethke, G., Lee, Y., Spanner, R., Badea, A., Baum, M., Belzile, F., Ben-David, R., Brueggeman, R., Case, A., Cattivelli, L., Davis, M., Dockter, C., Dolezel, J., Dreiseitl, A., Gavin, R., Glick, L., Greiner, S., Hamilton, R., Hayes, P. M., Heisel, S., Henson, C., Kilian, B., Komatsuda, T., Li, C., Liu, C., Mahalingam, R., Maruschewski, M., Matny, O., Maurer, A., Mayer, K. F. X., Mayrose, I., Morrell, P., Moscou, M., Muehlbauer, G. J., Oono, Y., Ordon, F., Ozkan, H., Pecinka, A.,

Perovic, D., Pillen, K., Pourkheirandish, M., Russell, J., Šafář, J., Salvi, S., Sanchez-Garcia, M., Sato, K., Schmutzer, T., Scholz, U., Scott, J., Brar, G. S., Smith, K. P., Sorrells, M. E., Spannagl, M., Stein, N., Tondelli, A., Tuberosa, R., Tucker, J., Turkington, T., Valkoun, J., Verma, R. P. S., Vinje, M. A., Schmising, M. v. K., Walling, J. G., Waugh, R., Wise, R. P., Wulff, B. B. H., Yang, S., Zhang, G., Mascher, M. and Steffenson, B. J. (2024). Whole-genome sequencing of the wild barley diversity collection: A resource for identifying and exploiting genetic variation for cultivated barley improvement.

Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J. and Schluter, D. (2017). Gene flow and selection interact to promote adaptive divergence in regions of low recombination, *Molecular Ecology* **26**(17): 4378–4390.

Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petroli, C., Campos, J., Dreher, K., Payne, T., Marshall, D. et al. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints, *Nature Communications* **11**(1): 4572.

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C., Dannenfelser, R., Dun, C., Edrisi, M. et al. (2022). Current progress and open challenges for applying deep learning across the biosciences, *Nature Communications* **13**(1): 1728.

Sato, K., Mascher, M., Himmelbach, A., Haberer, G., Spannagl, M. and Stein, N. (2021). Chromosome-scale assembly of wild barley accession 'OUH602', *G3 Genes|Genomes|Genetics* (jkab244).

Schmid, K., Kilian, B. and Russell, J. (2018). Barley domestication, adaptation and population genomics, *The Barley Genome*, Springer, pp. 317–336.

Schmidt, S. B., Brown, L. K., Booth, A., Wishart, J., Hedley, P. E., Martin, P., Husted, S., George, T. S. and Russell, J. (2023). Heritage genetics for adaptation to marginal soils in barley, *Trends in Plant Science* **28**(5): 544–551.

Schreiber, M., Jayakodi, M., Stein, N. and Mascher, M. (2024). Plant pangenomes for crop improvement, biodiversity and evolution, *Nature Reviews Genetics* pp. 1–15.

Schrider, D. R. and Kern, A. D. (2016). S/HIC: robust identification of soft and hard sweeps using machine learning, *PLoS Genetics* **12**(3): e1005928.

Schrider, D. R. and Kern, A. D. (2018). Supervised machine learning for population genetics: a new paradigm, *Trends in Genetics* **34**(4): 301–312.

Schulthess, A. W., Kale, S. M., Liu, F., Zhao, Y., Philipp, N., Rembe, M., Jiang, Y., Beukert, U., Serfling, A., Himmelbach, A. et al. (2022). Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement, *Nature Genetics* **54**(10): 1544–1552.

Sharma, R., Cockram, J., Gardner, K. A., Russell, J., Ramsay, L., Thomas, W. T., O'Sullivan, D. M., Powell, W. and Mackay, I. J. (2021). Trends of genetic changes uncovered by Env-and Eigen-GWAS in wheat and barley, *Theoretical and Applied Genetics* pp. 1–12.

Sharma, R., Shaaf, S., Neumann, K., Go, Y., Mascher, M., David, M., Al-Yassin, A., Özkan, H., Blake, T., Hübner, S. et al. (2020). On the origin of photoperiod non-responsiveness in barley, *bioRxiv* pp. 2020–07.

Siddiqui, M. N., Léon, J., Naz, A. A. and Ballvora, A. (2021). Genetics and genomics of root system variation in adaptation to drought stress in cereal crops, *Journal of Experimental Botany* **72**(4): 1007–1019.

Smith, A. L., Hodkinson, T. R., Villellas, J., Catford, J. A., Csergő, A. M., Blomberg, S. P., Crone, E. E., Ehrlén, J., Garcia, M. B., Laine, A.-L. et al. (2020). Global gene flow releases invasive plants from environmental constraints on genetic diversity, *Proceedings of the National Academy of Sciences* **117**(8): 4218–4227.

Stephan, W. (2019). Selective sweeps, *Genetics* **211**(1): 5–13.

Sunitha, N., Prathibha, M., Thribhuvan, R., Lokeshkumar, B., Basavaraj, P., Lohithaswa, H. and Anilkumar, C. (2024). Focused identification of germplasm strategy (FIGS): A strategic approach for trait-enhanced pre-breeding, *Genetic Resources and Crop Evolution* **71**(1): 1–16.

Szkiba, D., Kapun, M., von Haeseler, A. and Gallach, M. (2014). SNP2GO: functional analysis of genome-wide association studies, *Genetics* **197**(1): 285–289.

Takahashi, R. and Hayashi, J. (1964). Linkage study of two complementary genes for brittle rachis in barley, *Berichte des Ohara Instituts für Landwirtschaftliche Biologie, Okayama Universität* **12**(2): 99–105.

Tang, D., Jia, Y., Zhang, J., Li, H., Cheng, L., Wang, P., Bao, Z., Liu, Z., Feng, S., Zhu, X. et al. (2022). Genome evolution and diversity of wild and cultivated potatoes, *Nature* **606**(7914): 535–541.

Tanksley, S. D. and McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild, *Science* **277**(5329): 1063–1066.

Tanno, K.-i. and Willcox, G. (2012). Distinguishing wild and domestic wheat and barley spikelets from early Holocene sites in the Near East, *Vegetation history and archaeobotany* **21**: 107–115.

Teklemariam, S. S., Bayissa, K. N., Matros, A., Pillen, K., Ordon, F. and Wehner, G. (2022). The genetic diversity of Ethiopian barley genotypes in relation to their geographical origin, *PLoS One* **17**(5): e0260422.

Terrazas, R. A., Balbirnie-Cumming, K., Morris, J., Hedley, P. E., Russell, J., Paterson, E., Baggs, E. M., Fridman, E. and Bulgarelli, D. (2020). A footprint of plant eco-geographic adaptation on the composition of the barley rhizosphere bacterial microbiota, *Scientific Reports* **10**(1): 1–13.

Tiffin, P. and Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation, *Trends in Ecology & Evolution* **29**(12): 673–680.

Todesco, M., Bercovich, N., Kim, A., Imerovski, I., Owens, G. L., Ruiz, Ó. D., Holalu, S. V., Madilao, L. L., Jahani, M., Légaré, J.-S. et al. (2022). Genetic basis and dual adaptive role of floral pigmentation in sunflowers, *eLife* **11**: e72072.

Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S. and Fumagalli, M. (2019). ImaGene: a convolutional neural network to quantify natural selection from genomic data, *BMC Bioinformatics* **20**(Suppl 9): 337.

Tripodi, P., Rabanus-Wallace, M. T., Barchi, L., Kale, S., Esposito, S., Acquadro, A., Schafleitner, R., van Zonneveld, M., Prohens, J., Diez, M. J. et al. (2021). Global range expansion

history of pepper (*Capsicum* spp.) revealed by over 10,000 genebank accessions, *Proceedings of the National Academy of Sciences* **118**(34): e2104315118.

Tsuda, Y., Chen, J., Stocks, M., Källman, T., Sønstebø, J. H., Parducci, L., Semerikov, V., Sperisen, C., Politov, D., Ronkainen, T. et al. (2016). The extent and meaning of hybridization and introgression between Siberian spruce (*picea obovata*) and norway spruce (*picea abies*): cryptic refugia as stepping stones to the west?, *Molecular Ecology* **25**(12): 2773–2789.

Turner, A., Beales, J., Faure, S., Dunford, R. P. and Laurie, D. A. (2005). The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley, *Science* **310**(5750): 1031–1034.

Turner-Hissong, S. D., Mabry, M. E., Beissinger, T. M., Ross-Ibarra, J. and Pires, J. C. (2020). Evolutionary insights into plant breeding, *Current Opinion in Plant Biology* **54**: 93–100.

Uga, Y., Sugimoto, K., Ogawa, S., Rane, J., Ishitani, M., Hara, N., Kitomi, Y., Inukai, Y., Ono, K., Kanno, N. et al. (2013). Control of root system architecture by *deeper rooting 1* increases rice yield under drought conditions, *Nature Genetics* **45**(9): 1097–1102.

Varshney, R. K., Roorkiwal, M., Sun, S., Bajaj, P., Chitikineni, A., Thudi, M., Singh, N. P., Du, X., Upadhyaya, H. D., Khan, A. W. et al. (2021). A chickpea genetic variation map based on the sequencing of 3,366 genomes, *Nature* **599**(7886): 622–627.

Verma, R. P. S., Lal, C., Malik, R., Kharub, A. S., Kumar, L. and Kumar, D. (2022). Barley improvement: current status and future prospects in changing scenario, *New Horizons in Wheat and Barley Research: Global Trends, Breeding and Quality Enhancement* pp. 93–134.

Verma, S., Yashveer, S., Rehman, S., Gyawali, S., Kumar, Y., Chao, S., Sarker, A. and Verma, R. P. S. (2021). Genetic and agro-morphological diversity in global barley (*Hordeum vulgare L.*) collection at ICARDA, *Genetic Resources and Crop Evolution* **68**: 1315–1330.

Volis, S. (2011). Adaptive genetic differentiation in a predominantly self-pollinating species analyzed by transplanting into natural environment, crossbreeding and QST-FST test, *New Phytologist* **192**(1): 237–248.

Volis, S., Mendlinger, S. and Ward, D. (2002a). Adaptive traits of wild barley plants of Mediterranean and desert origin, *Oecologia* **133**(2): 131–138.

Volis, S., Mendlinger, S. and Ward, D. (2002b). Differentiation in populations of *Hordeum spontaneum* along a gradient of environmental productivity and predictability: life history and local adaptation, *Biological Journal of the Linnean Society* **77**(4): 479–490.

Volis, S., Shulgina, I., Ward, D. and Mendlinger, S. (2003). Regional subdivision in wild barley allozyme variation: adaptive or neutral?, *Journal of Heredity* **94**(4): 341–351.

Volis, S., Verhoeven, K., Mendlinger, S. and Ward, D. (2004). Phenotypic selection and regulation of reproduction in different environments in wild barley, *Journal of Evolutionary Biology* **17**(5): 1121–1131.

Volis, S., Yakubov, B., Shulgina, I., Ward, D. and Mendlinger, S. (2005). Distinguishing adaptive from nonadaptive genetic differentiation: comparison of Q ST and F ST at two spatial scales, *Heredity* **95**(6): 466–475.

Volis, S., Yakubov, B., Shulgina, I., Ward, D., Zur, V. and Mendlinger, S. (2001). Tests for adaptive RAPD variation in population genetic structure of wild barley, *Hordeum spontaneum* Koch, *Biological Journal of the Linnean Society* **74**(3): 289–303.

Volis, S., Zaretsky, M. and Shulgina, I. (2010). Fine-scale spatial genetic structure in a predominantly selfing plant: role of seed and pollen dispersal, *Heredity* **105**(4): 384–393.

Von Korff, M., Wang, H., Léon, J. and Pillen, K. (2004). Development of candidate introgression lines using an exotic barley accession (*hordeum vulgare ssp. spontaneum*) as donor, *Theoretical and Applied Genetics* **109**: 1736–1745.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R. R., Zhang, F. et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice, *Nature* **557**(7703): 43–49.

Watson, A., Ghosh, S., Williams, M. J., Cuddy, W. S., Simmonds, J., Rey, M.-D., Asyraf Md Hatta, M., Hinchliffe, A., Steed, A., Reynolds, D. et al. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding, *Nature Plants* **4**(1): 23–29.

Watson, A., Hickey, L. T., Christopher, J., Rutkoski, J., Poland, J. and Hayes, B. J. (2019). Multivariate genomic selection and potential of rapid indirect selection with speed breeding in spring wheat, *Crop Science* **59**(5): 1945–1959.

Wendler, N., Mascher, M., Nöh, C., Himmelbach, A., Scholz, U., Ruge-Wehling, B. and Stein, N. (2014). Unlocking the secondary gene-pool of barley with next-generation sequencing, *Plant Biotechnology Journal* **12**(8): 1122–1131.

Whitehouse, L. S., Ray, D. and Schrider, D. R. (2024). Tree sequences as a general-purpose tool for population genetic inference, *Molecular Biology and Evolution* p. msae223.

Whitehouse, L. S. and Schrider, D. R. (2023). Timesweeper: accurately identifying selective sweeps using population genomic time series, *Genetics* **224**(3): iyad084.

Whitlock, M. C. and Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of F ST, *The American Naturalist* **186**(S1): S24–S36.

Wiegmann, M., Maurer, A., Pham, A., March, T. J., Al-Abdallat, A., Thomas, W. T., Bull, H. J., Shahid, M., Eglinton, J., Baum, M. et al. (2019). Barley yield formation under abiotic stress depends on the interplay between flowering time genes and environmental cues, *Scientific Reports* **9**(1): 6397.

Wright, S. (1943). Isolation by distance, *Genetics* **28**(2): 114.

Wright, S. (1949). The genetical structure of populations, *Annals of Eugenics* **15**(1): 323–354.

Wu, Y., Li, D., Hu, Y., Li, H., Ramstein, G. P., Zhou, S., Zhang, X., Bao, Z., Zhang, Y., Song, B. et al. (2023). Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding, *Cell* **186**(11): 2313–2328.

Yahiaoui, S., Cuesta-Marcos, A., Gracia, M. P., Medina, B., Lasa, J. M., Casas, A. M., Ciudad, F. J., Montoya, J. L., Moralejo, M., Molina-Cano, J. L. et al. (2014). Spanish barley landraces outperform modern cultivars at low-productivity sites, *Plant Breeding* **133**(2): 218–226.

Yang, W.-Y., Novembre, J., Eskin, E. and Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data, *Nature Genetics* **44**(6): 725–731.

Yeaman, S. and Whitlock, M. C. (2011). The genetic architecture of adaptation under migration–selection balance, *Evolution: International Journal of Organic Evolution* **65**(7): 1897–1911.

Yuan, Z., Rembe, M., Mascher, M., Stein, N., Jayakodi, M., Börner, A., Oldach, K., Jahoor, A., Jensen, J. D., Rudloff, J. et al. (2024). Capitalizing on genebank core collections for rare and novel disease resistance loci to enhance barley resilience, *Journal of Experimental Botany* **75**(18): 5940–5954.

Zhai, J., Gokaslan, A., Schiff, Y., Berthel, A., Liu, Z.-Y., Miller, Z. R., Scheben, A., Stitzer, M. C., Romay, M. C., Buckler, E. S. et al. (2024). Cross-species modeling of plant genomes at single nucleotide resolution using a pre-trained DNA language model, *bioRxiv* pp. 2024–06.

Zheng, X., Levine, D., Shen, J., Gogarten, S., Laurie, C. and Weir, B. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data, *Bioinformatics* **28**(24): 3326–3328.

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K. et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding, *Nature* **606**(7914): 527–534.

Zhu, J., Zhou, H., Fan, Y., Guo, Y., Zhang, M., Shabala, S., Zhao, C., Lv, C., Guo, B., Wang, F. et al. (2023). *HvNCX*, a prime candidate gene for the novel qualitative locus *qs7.1* associated with salinity tolerance in barley, *Theoretical and Applied Genetics* **136**(1): 9.

Zohary, D., Hopf, M. and Weiss, E. (2012). *Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin*, Oxford University Press on Demand.

Zsögön, A., Čermák, T., Naves, E. R., Notini, M. M., Edel, K. H., Weinl, S., Freschi, L., Voytas, D. F., Kudla, J. and Peres, L. E. P. (2018). De novo domestication of wild tomato using genome editing, *Nature Biotechnology* **36**(12): 1211–1216.

Appendix A

Physical geography, isolation by distance and environmental variables shape genomic variation of wild barley (*Hordeum vulgare* L. ssp. *spontaneum*) in the Southern Levant - Supplementary information

A.1 Genotypic data filtration and processing

Six accessions are included in both the IPK (Milner et al. 2019) and B1K collections (Hübner et al. 2009). We considered the 6 overlapping accessions as part of B1K collection for this study, therefore no accessions overlap between the 244 B1K+ accessions and 1,121 IPK's accessions in this context.

Identification of single nucleotide polymorphism (SNP) was carried out with the same pipeline as in Milner et al. (2019) based on the barley 'Morex V2' assembly. The raw VCF file was first filtered by using an AWK script (bitbucket.org/ipk_dg_public/vcf_filtering) to remove loci with QUAL (mapping quality) less than 40 and genotypic calls with DP (read depth) less than 2. SNPs with a missing proportion higher than 0.2 and accessions with a missing proportion higher than 0.3 were excluded. Next, genotypic data of duplicates were merged according to the following rules: (1) retain a genotypic value if genotypes of duplicates are identical at a given SNP locus, (2) turn a given SNP locus to missing value if any genotype of duplicate conflicts with another, and (3) retain the genotypic value if only one duplicate is not missing at a given SNP locus. Considering that wild barley is predominantly self-fertilizing (Brown et al. 1978), accessions with more than 0.05 heterozygosity among all polymorphic sites were discarded, and then remaining heterozygotes in the dataset were treated as missing values. To investigate the extent of B1K+ accessions covering the genetic variation of the wild barley collection of the IPK genebank, a dataset consisting of 1,121 IPK accessions and 244 B1K+ accessions with 4,793 geographically diverse SNPs was extracted. Briefly, SNPs with a missing proportion $<$0.2 across the whole panel and minor allele frequency (MAF) $>$0.05 among 72 geographically diverse accessions, collected from 13 countries (Russell et al. 2016), were selected. Then, the selected SNPs were pruned by using *PLINK 1.9* (Purcell et al. 2007) to remove SNPs in linkage-disequilibrium (LD) with an r^2 threshold of 0.1, a window size of 50, and a step size of 5, and eventually resulted in 4,793 SNPs.

For analyses of wild barley from the southern Levant, we first extracted 58,616 SNPs with MAF higher than 0.01 and a missing proportion lower than 0.1 among 244 B1K+ accessions. For population structure analysis, LD pruning was performed to exclude redundant markers with an r^2 threshold of 0.1 using *PLINK 1.9* (Purcell et al. 2007) by considering the model

assumption of our downstream analysis (Cabreros and Storey 2019). LD-pruning reduced the unimputed 58,616 SNPs to 19,601 SNPs. For redundancy analysis (RDA), another dataset with no missing values was prepared because RDA requires a non-missing dataset. The whole wild barley panel, including the B1K+ and IPK collections, was imputed by using *BEAGLE 5.1* (Browning et al. 2018) with default parameters. Next, the 244 B1K+ accessions with the same 58,616 SNPs mentioned above, which have the original missing proportion lower than 0.1 among 244 accessions, were extracted. Lastly, SNPs with MAF lower than 0.05 among 244 B1K+ accessions were removed, resulting in 27,147 imputed SNPs. The imputed SNP data were coded as 0, 1, and 2 according to the counts of alternative alleles and treated as the response variable for RDA. In addition to the RDA, the imputed dataset with 27,147 SNPs was also applied to other genome scan analyses.

A.2 Construction of synthetic environmental variables and variable selection

The climatic data including temperature, precipitation and solar radiation was downloaded from *WorldClim2* (Fick and Hijmans 2017) with the resolution of 30 arc-seconds ($\sim$1 km). The climatic data for growing period of wild barley in Israel, typically between late October and April, was extracted to calculate the bioclimatic variables. Soil property data including a total of 11 variables of four soil layers (0 cm, 5 cm, 15 cm, and 30 cm) with a resolution of 250 m was obtained from SoilGrids (*soilgrids.org*;Hengl et al. 2017). Elevation data with the resolution of 90 m was downloaded from the SRTM database (`https://srtm.csi.cgiar.org/`). Aspects and slopes were calculated based on the elevation data using the *terrain* function in the R package *raster*. Radians of aspects were further converted to cosine values, denoting north-facing and south-facing slopes as 1 and -1, respectively. Environmental data for 244 B1K+ accessions was extracted with the function *extract* in the R package *raster* according to the geographic coordinates of their collecting locations.

To resolve the problem with collinearity, we first combined the highly correlated environmental variables to generate synthetic environmental variables. A synthetic environmental

variable was defined as the first principal component scores of a group of highly correlated environmental variables. To cluster the highly correlated environmental variables that can be combined into a synthetic environmental variable, we first grouped all environmental variables by using the complete-linkage clustering based on the dissimilarity between environmental variables. The dissimilarity of a pair of environmental variables was calculated as $1 - R^2$. The R^2 represents the coefficient of determination of two environmental variables. With the dendrogram of the complete-linkage clustering, we could have various clustering combinations by cutting the tree at different levels. Next, we searched the optimal clustering combinations along the dendrogram by using a customized index. The index was calculated as

$$\frac{\sum_{n=1}^{k} SS_{PC1,n}}{SS_{Total} - \sum_{n=1}^{k} SS_{PC1,n}}$$

where *k* is the number of clusters with more than one environmental variable when cutting the dendrogram at a given level. The $SS_{PC1,n}$ and SS_{Total} terms represent the sum of squares of the first principal component calculated from n-th cluster of standardized environmental variables and the total sum of squares of all standardized environmental variables respectively. The $SS_{PC1,n}$ and SS_{Total} terms were calculated by using the *svd* function in R. This index was designed by regarding the sum of squares as the amount of information. Thus, the index can be interpreted as the ratio of information captured by the first PCs of grouped variables to the total information of the original data. The decrease of this index can suggest the loss of information when using the first PCs to represent the grouped variables. In other words, by maximizing the index, we could supposedly find clustering combinations that produced the most informative synthetic environmental variables to represent highly correlated environmental variables. The optimal cuttoff $1 - R^2$ cutoff was searched between 0 and 1 (A.3A). Subsequently, the first PCs, or synthetic environmental variables, were computed according to the optimal clustering combination (A.3B). After that, the synthetic environmental variables and the non-synthetic environmental variables were further selected until variance inflation factors (VIFs) of all variables were less than 5. The selection was done by sequentially removing the environmental variable with the highest VIF, but we manually kept the variable related to the accumulated precipitation and the mean temperature ('Latitude+Rain+Solar_rad' and 'Elevation+Temperature') because the precipitation and temperature have been suggested as the most important gradi-

ents (Hübner et al. 2009). The VIFs were calculated as $\frac{1}{1-R_j^2}$, where the R_j^2 is the coefficient of determination obtained by regressing variable j on all the other variables (Legendre and Legendre 2012). The procedure resulted in 12 environmental variables, including 7 synthetic environmental variables and 5 non-synthetic environmental variables (Table A.1). The rotations used to generate synthetic variables are shown in Table A.2.

A.3 Classification of barrier and non-barrier pixel

To test the hypothesis that geographical barriers are significantly associated with the lower migration rate, we performed a Wilcoxon test by using migration rates estimated by *EEMS* and the geographical barriers defined according to geographical elevation. We assumed the relatively low migration areas on the landscape result from a drastic change in elevation, such as valleys and mountains, therefore we identified geographical barriers by selecting the map pixels with elevation deviating from the majority of pixels (A.4). Specifically, we selected the pixels with the top 25%, a cut-off where pixel counts drastically decreased (A.4 A), of the absolute values of the standardized elevation and treated selected pixels as the geographical barriers (A.4 B). Thereby, pixels of the geographical map were classified as barrier pixels and non-barrier pixels, and the corresponding migration rate of each pixel was extracted from the result of *EEMS*. Finally, a Wilcoxon test was carried out to test if the migration rates are lower at the barrier pixels than the non-barrier pixels.

A.4 Supplementary Tables

Table A.1 Description of environmental variables used in redundancy analysis (RDA) and genome-environment association (GEA) analysis.

Environmental variable	Description	Proportion of explained variance[a]
Aspect	Geographical aspect	-
CoefVar_Rain	Coefficient of variation of precipitation of growing season	-
Elevation+Temperature	First PC of elevation, mean temperature, maximal temperature, and minimal temperature of growing season	0.931
Latitude+Rain+Solar_rad	First PC of latitude, accumulated precipitation, and mean solar radiation of growing season	0.910
Slope	Geographical slope	-
Soil_bulk_density	First PC of soil bulk density of the depths of 0 cm and 5 cm	0.947
Soil_carbon_content(0-15cm)	First PC of soil organic carbon content of the depth of 0 cm, 5 cm, and 15 cm	0.883
Soil_carbon_content(30cm)	Soil organic carbon content of the depth of 30 cm	-
Soil_pH	First PC of soil pH of the depth of 0 cm, 5 cm, 15 cm, and 30 cm	0.969
Soil_silt_content	First PC of silt content (2–50 micro meter) mass fraction in percentage of the depth of 0 cm, 5 cm, 15 cm, and 30 cm	0.970
Soil_water_capacity	First PC of soil water capacity with with FC = pF 2.0, 2.3, and 2.5 until wilt point of the depth of 0 cm	0.962
StdDev_Temperature	Standard deviation of temperature of growing season	-

Table A.2 Rotations of first principal component (PC1) used to generate the synthetic environmental variables.

Environmental Variable	Rotation of PC1
Aspect	-
CoefVar_Rain	-
Elevation+Temperature	$0.508 \times Elevation - 0.513 \times avg\_temp - 0.491 \times max\_temp - 0.487 \times min\_temp$
Latitude+Rain+Solar_rad	$0.588 \times Latitude + 0.557 \times sum\_prec - 0.587 \times avg\_srad$
Slope	-
Soil_bulk_density	$-0.707 \times sbd(0cm) - 0.707 \times sbd(5cm)$
Soil_carbon_content(0-15cm)	$0.557 \times socc(0cm) + 0.590 \times socc(5cm) + 0.585 \times socc(15cm)$
Soil_carbon_content(30cm)	-
Soil_pH	$0.495 \times pH(0cm) + 0.502 \times pH(5cm) + 0.503 \times pH(15cm) + 0.500 \times pH(30cm)$
Soil_silt_content	$-0.499 \times sltc(0cm) - 0.502 \times sltc(5cm) - 0.503 \times sltc(15cm) - 0.495 \times sltc(30cm)$
Soil_water_capacity	$-0.575 \times swc(pF2.0; 0cm) - 0.582 \times \times swc(pF2.3; 0cm) - 0.575 \times \times swc(pF2.5; 0cm)$
StdDev_Temperature	-

sum_prec: Accumulated precipitation of growing season
avg_srad: Average solar radiation of growing season
avg_temp: Average temperature of growing season
max_temp: Maximum temperature of growing season
min_temp: Minimum temperature of growing season
bd: Soil bulk density (depth of soil layer is shown in the parenthesis)
swc: Soil water capacity (field capacity, referred to pF, and depth of soil layer are shown in the parenthesis)
$socc$: Soil organic carbon content (depth of soil layer is shown in the parenthesis)
pH: Soil pH value (depth of soil layer is shown in the parenthesis)
$sltc$: Silt content (depth of soil layer is shown in the parenthesis)

Table A.3 Bootstrap analysis of *ResistanceGA* with 1,000 iterations using random resampling of 75% of sampled populations. The leftmost column shows resistance surfaces used in linear mixed effect models. "Avg.AIC" is the average Akaike information criterion. "Avg.R^2_m" is the average marginal R^2. "Avg.LL" is the average log-likelihood. "n" and "Percent top" is the number and percentage of iterations that a model is selected as the best-supported model. "k" is the number of parameters fitted in a model.

Table A.3.1 Bootstrapping with model ranking based on R^2_m

Surface	Avg.AIC	Avg.R^2_m	Avg.LL	n	Percent top	k
Elev+Slope	-5725.286	0.5083	2869.643	1000	100	7
Elev+Slope+Water	-5734.975	0.4414	2876.487	0	0	9
Slope+Water	-5736.288	0.3311	2874.144	0	0	6
Elev	-5741.162	0.3159	2874.581	0	0	4
Water	-5783.527	0.3042	2894.764	0	0	3
Elev+Water	-5739.697	0.3010	2875.849	0	0	6
Distance	-5753.420	0.2113	2878.710	0	0	2
Slope	-5730.958	0.2074	2869.479	0	0	4

Table A.3.2 Bootstrapping with model ranking based on AIC

Surface	Avg.AIC	Avg.R^2_m	Avg.LL	n	Percent top	k
Water	-5783.397	0.3055	2894.698	968	96.8	3
Distance	-5752.230	0.2115	2878.115	9	0.9	2
Elev	-5739.792	0.3169	2873.896	7	0.7	4
Elev+Water	-5738.073	0.3014	2875.036	12	1.2	6
Elev+Slope+Water	-5732.737	0.4412	2875.369	0	0.0	9
Slope+Water	-5733.853	0.3317	2872.926	4	0.4	6
Slope	-5729.806	0.2077	2868.903	0	0.0	4
Elev+Slope	-5723.318	0.5080	2868.659	0	0.0	7

Table A.4 Gene flow rate and coalescence rate of 10 geographical clusters inferred by coalescence-based method. The rows and columns respectively indicate the source and sink of gene flow. The coalescence rates are shown in diagonal. The 95% credible intervals are shown in the parentheses.

	To									
From	A	B	C	D	E	F	G	H	I	J
A	0.9304 (0-2.04)	1.1481 (0-2.99)	1.3852 (0-4.46)	2.6973 (1.1-4.69)	-	-	-	-	-	-
B	0.66 (0-2.27)	1.2995 (0.42-2.81)	0.7353 (0-2.73)	-	0.7096 (0-2.15)	0.2949 (0-0.89)	-	-	-	-
C	1.2967 (0-3.99)	0.7587 (0-2.65)	0.9776 (0.28-1.96)	0.5604 (0-1.71)	0.3461 (0-1.29)	0.5628 (0-1.29)	-	-	-	-
D	0.4697 (0-1.84)	-	0.4442 (0-1.61)	1.3294 (0.42-2.64)	0.4024 (0-1.62)	2.9335 (1.49-4.88)	-	-	-	-
E	-	1.4597 (0-3.5)	0.3548 (0-1.42)	0.3587 (0-1.32)	0.4086 (0-1)	0.8163 (0-1.91)	0.5671 (0-1.81)	0.7746 (0-2.32)	0.5595 (0-1.71)	-
F	-	0.534 (0-1.6)	0.4686 (0-1.59)	0.6024 (0-2.35)	0.4509 (0-1.89)	2.3576 (0.88-4.33)	0.2834 (0-0.83)	0.5712 (0-1.8)	2.0077 (0.84-3.55)	-
G	-	-	-	-	0.4679 (0-1.39)	0.1673 (0-0.56)	1.2041 (0.4-2.2)	1.1336 (0-3.12)	-	0.0865 (0-0.36)
H	-	-	-	-	3.2381 (0.51-6.32)	1.1729 (0-2.36)	2.0253 (0-3.92)	1.7529 (0.61-3.27)	0.0983 (0-0.41)	0.0655 (0-0.3)
I	-	-	-	-	0.154 (0-0.54)	0.242 (0-0.8)	-	0.1216 (0-0.4)	0.5207 (0.18-1.14)	0.253 (0-1.16)
J	-	-	-	-	-	-	0.1044 (0-0.46)	0.5311 (0.04-0.91)	0.538 (0-0.99)	0.6713 (0.2-1.19)

Table A.5 Effect of environmental variables estimated by RDA models. Environmental variables are ordered according to the explained variation. P-values are computed by permutation tests with 5,000 iterations.

Table A.5.1	Environmental variable	F	Var	Var (%)	p-value
Individual effect of environmental variables	Latitude+Rain+Solar_rad	9.8046	553.6372	3.8937	0.0002
	Soil_carbon_content(0-15cm)	6.5739	376.0355	2.6447	0.0002
	Soil_carbon_content(30cm)	5.3230	306.0195	2.1522	0.0002
	Soil_silt_content	5.1947	298.7987	2.1015	0.0002
	Soil_bulk_density	5.1081	293.9198	2.0671	0.0002
	Soil_pH	4.8677	280.3609	1.9718	0.0002
	Elevation+Temperature	3.9666	229.2979	1.6127	0.0002
	Soil_water_capacity	3.8998	225.4961	1.5859	0.0002
	StdDev_Temperature	3.5253	204.1527	1.4358	0.0002

	CoefVar_Rain	3.2164	186.4980	1.3116	0.0002
	Slope	2.9324	170.2299	1.1972	0.0002
	Aspect	1.8359	107.0534	0.7529	0.0002

Table A.5.2 Individual effect of environmental variables conditioned on population structure

Environmental variable	F	Var	Var (%)	p-value
Soil_water_capacity	3.3934	168.3390	1.1839	0.0002
CoefVar_Rain	3.1455	156.2026	1.0986	0.0002
Elevation+Temperature	2.8432	141.3632	0.9942	0.0002
Soil_pH	2.7389	136.2396	0.9582	0.0002
Latitude+Rain+Solar_rad	2.4531	122.1676	0.8592	0.0002
StdDev_Temperature	2.2437	111.8356	0.7865	0.0002
Soil_carbon_content(0-15cm)	2.2237	110.8457	0.7796	0.0002
Soil_bulk_density	1.9666	98.1368	0.6902	0.0002
Soil_carbon_content(30cm)	1.8504	92.3804	0.6497	0.0002
Slope	1.8143	90.5932	0.6371	0.0002
Soil_silt_content	1.7954	89.6590	0.6306	0.0002
Aspect	1.6250	81.2028	0.5711	0.0002

Table A.5.3 Marginal effect of environmental variables

Environmental variable	F	Var	Var (%)	p-value
Latitude+Rain+Solar_rad	3.3906	177.1453	1.2459	0.0002
Soil_water_capacity	3.2845	171.6034	1.2069	0.0002
StdDev_Temperature	3.2423	169.3995	1.1914	0.0002
Elevation+Temperature	3.1535	164.7580	1.1587	0.0002
Soil_carbon_content(0-15cm)	2.4493	127.9675	0.9000	0.0002
Soil_pH	2.2410	117.0829	0.8234	0.0002
Soil_bulk_density	2.1565	112.6705	0.7924	0.0002
Soil_silt_content	2.1446	112.0493	0.7880	0.0004
CoefVar_Rain	1.9776	103.3233	0.7267	0.0002
Aspect	1.6364	85.4935	0.6013	0.0018

	Slope	1.5674	81.8895	0.5759	0.0024
	Soil_carbon_content(30cm)	1.5571	81.3511	0.5721	0.0054

Table A.5.4 Marginal effect of environmental variables conditioned on population structure

Environmental variable	F	Var	Var (%)	p-value
Soil_water_capacity	2.8405	134.3772	0.9451	0.0002
Elevation+Temperature	2.4234	114.6438	0.8063	0.0002
CoefVar_Rain	2.1510	101.7564	0.7157	0.0002
StdDev_Temperature	2.1193	100.2602	0.7051	0.0002
Soil_pH	2.0162	95.3822	0.6708	0.0006
Soil_bulk_density	1.8989	89.8333	0.6318	0.0006
Latitude+Rain+Solar_rad	1.8864	89.2404	0.6276	0.0002
Soil_silt_content	1.8743	88.6696	0.6236	0.0004
Soil_carbon_content(0-15cm)	1.7986	85.0871	0.5984	0.0006
Soil_carbon_content(30cm)	1.6314	77.1770	0.5428	0.0034
Slope	1.4805	70.0406	0.4926	0.0080
Aspect	1.4777	69.9055	0.4916	0.0090

Table A.6 Correlation of environmental variables with fitted site scores on first four RDA axes. The correlations were evaluated based on the RDA models (1) without covariates (Simple RDA), (2) conditioned on spatial autocorrelation with dbMEMs (Partial RDA dbMEM), and (3) conditioned on population structure with ancestry coefficients (Partial RDA PopStr).

	Simple RDA				Partial RDA dbMEM				Partial RDA PopStr			
	RDA1	RDA2	RDA3	RDA4	RDA1	RDA2	RDA3	RDA4	RDA1	RDA2	RDA3	RDA4
Aspect	0.067	0.185	0.251	0.3	0.274	-0.365	-0.121	0.237	0.269	0.193	0.072	0.038
CoefVar_Rain	-0.015	0.126	0.543	-0.604	0.013	0.018	0.017	0.008	-0.662	0.035	0.106	0.584
Elevation+Temperature	-0.077	0.33	-0.751	-0.193	0.001	-0.011	0.008	0.004	0.353	-0.557	0.265	-0.046
Latitude+Rain+Solar_rad	0.911	-0.201	0.023	-0.197	-0.004	-0.007	0.005	0.002	0.065	-0.475	-0.049	0.027
Slope	0.335	-0.264	-0.027	-0.178	0.038	-0.065	0.106	-0.069	-0.146	-0.049	0.371	-0.198
Soil_bulk_density	0.59	-0.03	0.037	-0.44	0.041	-0.036	0.021	0.044	-0.272	-0.243	0.176	-0.045
Soil_carbon_content(0-15cm)	0.706	0.157	0.153	-0.218	-0.019	0.022	-0.020	-0.001	0.084	-0.396	0.107	-0.21
Soil_carbon_content(30cm)	0.582	-0.325	0.209	0.221	0.086	0.044	-0.015	0.029	0.169	-0.118	-0.23	-0.254
Soil_pH	-0.526	-0.078	-0.224	0.326	0.021	0.077	0.025	0.079	0.028	0.615	0.354	-0.144
Soil_silt_content	-0.377	0.687	0.203	-0.238	0.149	-0.095	0.027	-0.078	-0.093	0.049	0.12	0.382
Soil_water_capacity	0.325	-0.039	-0.254	0.845	-0.016	-0.004	-0.042	-0.053	0.697	0.44	-0.18	-0.072
StdDev_Temperature	-0.257	-0.453	-0.113	0.216	-0.020	-0.038	-0.006	-0.003	0.118	-0.272	-0.544	-0.267

Table A.7 Genes within 500 bp upstream or downstream of the candidate SNPs jointly detected by four methods.

Gene	Chr.	Start (bp)	End (bp)	Position of candidate SNP (bp)	Method	Environmental variable	Gene annotation
HORVU.MOREX.r2.4HG0308420	chr4H	312,111,636	312,189,673	312,146,232	Simple RDA (p = 2.007e-06; q = 0.002)	-	ATP-dependent RNA helicase
				312,146,232	Partial RDA (p = 2.641e-04; q = 0.031)	-	
				312,146,215	LFMM (p = 1.940e-06; q = 0.028)	Latitude+Rain+Solar_rad	
				312,146,232	BAYPASS (X^TX = 11.303)	-	
HORVU.MOREX.r2.4HG0314300	chr4H	392,019,895	392,026,683	392,022,112	Simple RDA (p = 3.183e-09; q = 3.245e-05)	-	Nucleolar GTP-binding protein 2
				392,022,112	Partial RDA (p = 5.620e-07; q = 0.001)	-	
				392,021,665	LFMM (p = 1.372e-4; q = 0.0364)	Elevation+Temperature	
				392,021,634	BAYPASS (X^TX = 14.690)	-	

Table A.8 Result of gene ontology (GO) term enrichment.

	Enriched GO	P-value	FDR	GO definition
Simple RDA	GO:0004556	2.7640e-07	0.0014	Catalysis of the endohydrolysis of (1->4)-alpha-D-glucosidic linkages in polysaccharides containing three or more alpha-(1->4)-linked D-glucose units.
	GO:0103025	2.3737e-07	0.0014	Catalysis of the reaction: n H2O + a 1,4-alpha-D-glucan = alpha-maltohexaose + a 1,4-alpha-D-glucan
BAYPASS	GO:0006631	8.2636e-06	0.0077	The chemical reactions and pathways involving fatty acids, aliphatic monocarboxylic acids liberated from naturally occurring fats and oils by hydrolysis.
	GO:0009735	1.1465e-06	0.0029	Any process that results in a change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a cytokinin stimulus.
	GO:0009236	2.5754e-05	0.0203	The chemical reactions and pathways resulting in the formation of cobalamin (vitamin B12), a water-soluble vitamin characterized by possession of a corrin nucleus containing a cobalt atom.
	GO:0090351	1.18776e-07	0.0004	The process whose specific outcome is the progression of the seedling over time, beginning with seed germination and ending when the first adult leaves emerge.
	GO:0006635	2.4051e-06	0.0038	A fatty acid oxidation process that results in the complete oxidation of a long-chain fatty acid. Fatty acid beta-oxidation begins with the addition of coenzyme A to a fatty acid, and occurs by successive cycles of reactions during each of which the fatty acid is shortened by a two-carbon fragment removed as acetyl coenzyme A; the cycle continues until only two or three carbons remain (as acetyl-CoA or propionyl-CoA respectively).
	GO:0009062	5.8859e-06	0.0063	The chemical reactions and pathways resulting in the breakdown of a fatty acid, any of the aliphatic monocarboxylic acids that can be liberated by hydrolysis from naturally occurring fats and oils. Fatty acids are predominantly straight-chain acids of 4 to 24 carbon atoms, which may be saturated or unsaturated; branched fatty acids and hydroxy fatty acids also occur, and very long chain acids of over 30 carbons are found in waxes.
	GO:0009845	1.1054e-07	0.0004	The physiological and developmental changes that occur in a seed commencing with water uptake (imbibition) and terminating with the elongation of the embryonic axis.
	GO:0004325	2.4976e-06	0.0038	Catalysis of the reaction: protoheme = Fe(2+) + protoporphyrin IX.
	GO:0046487	1.1351e-09	1.1649e-05	The chemical reactions and pathways involving glyoxylate, the anion of glyoxylic acid, HOC-COOH.
	GO:0019395	1.7175e-05	0.0147	The removal of one or more electrons from a fatty acid, with or without the concomitant removal of a proton or protons, by reaction with an electron-accepting substance, by addition of oxygen or by removal of hydrogen.

A.5 Supplementary Figures

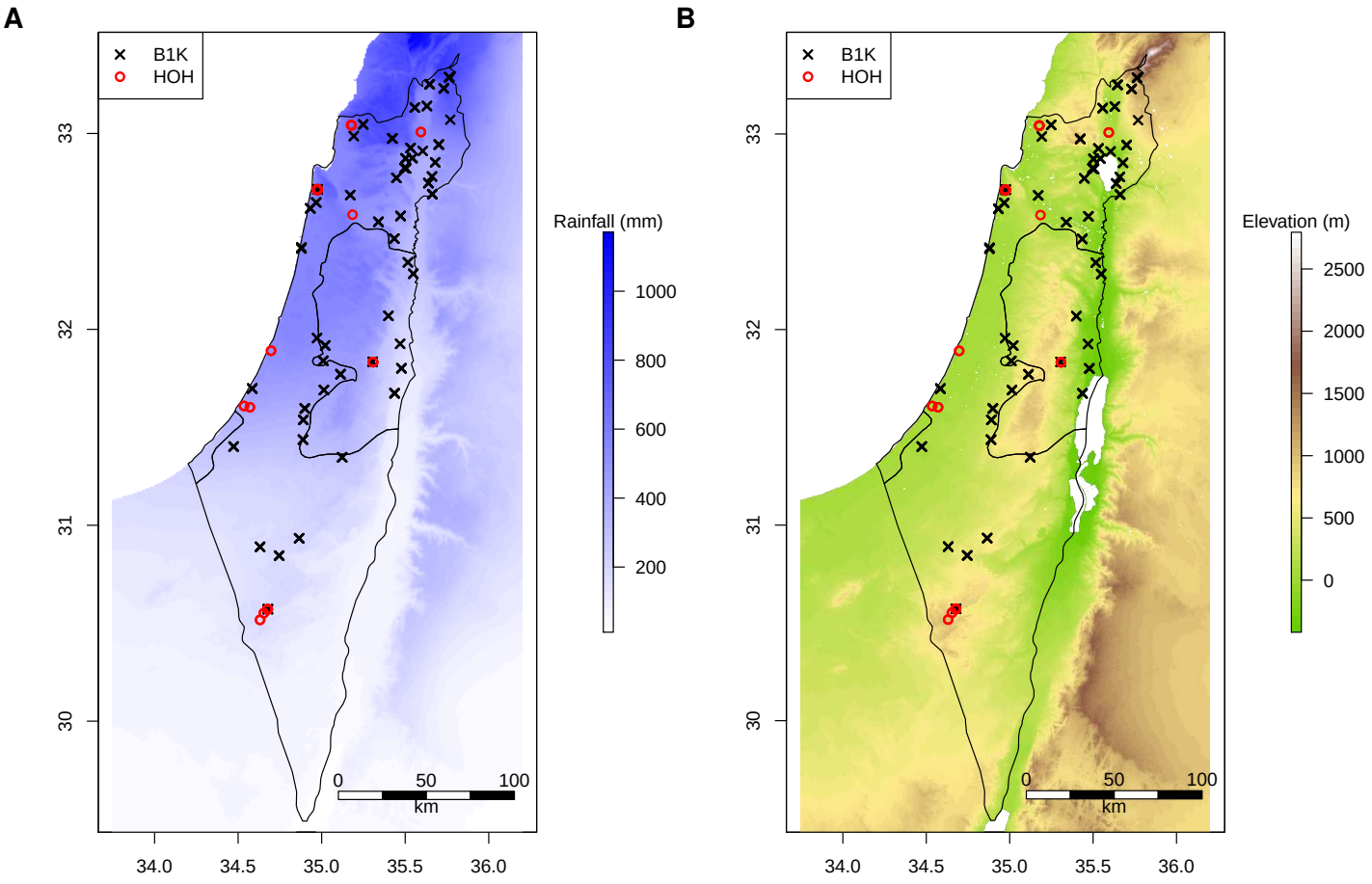

Figure A.1 Collection sites of 244 B1K+ accessions. (A) Map with the accumulated rainfall between October and April. (B) Map with the geographical elevation. The black crosses and red circles represent the collection sites of B1K accessions and unpublished accessions stored at the University of Hohenheim (HOH collection), respectively.

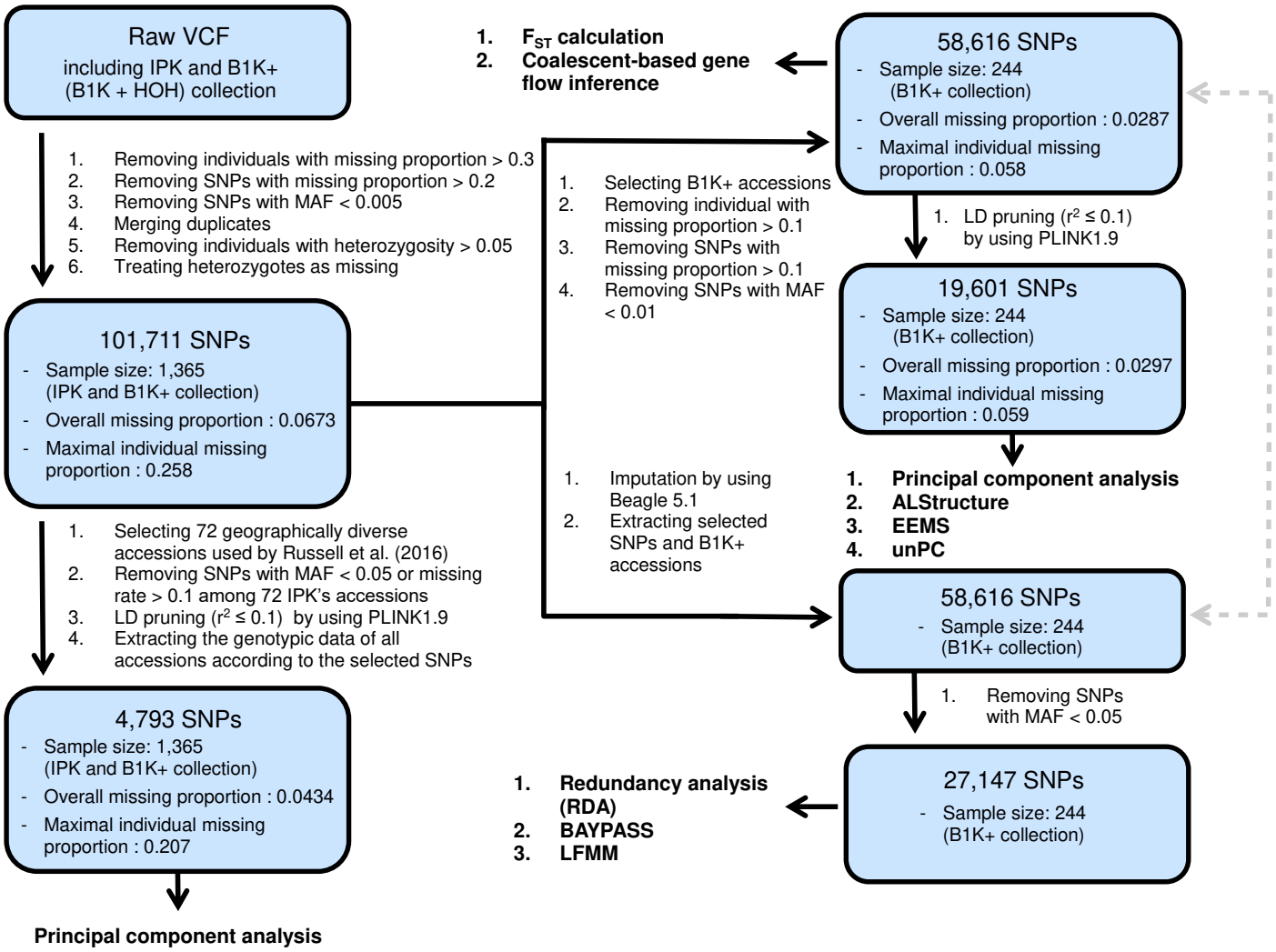

Figure A.2 Workflow of genotypic data filtration. The double arrow with the gray dash line represents that two datasets are consisting of the same 58,616 SNP loci.

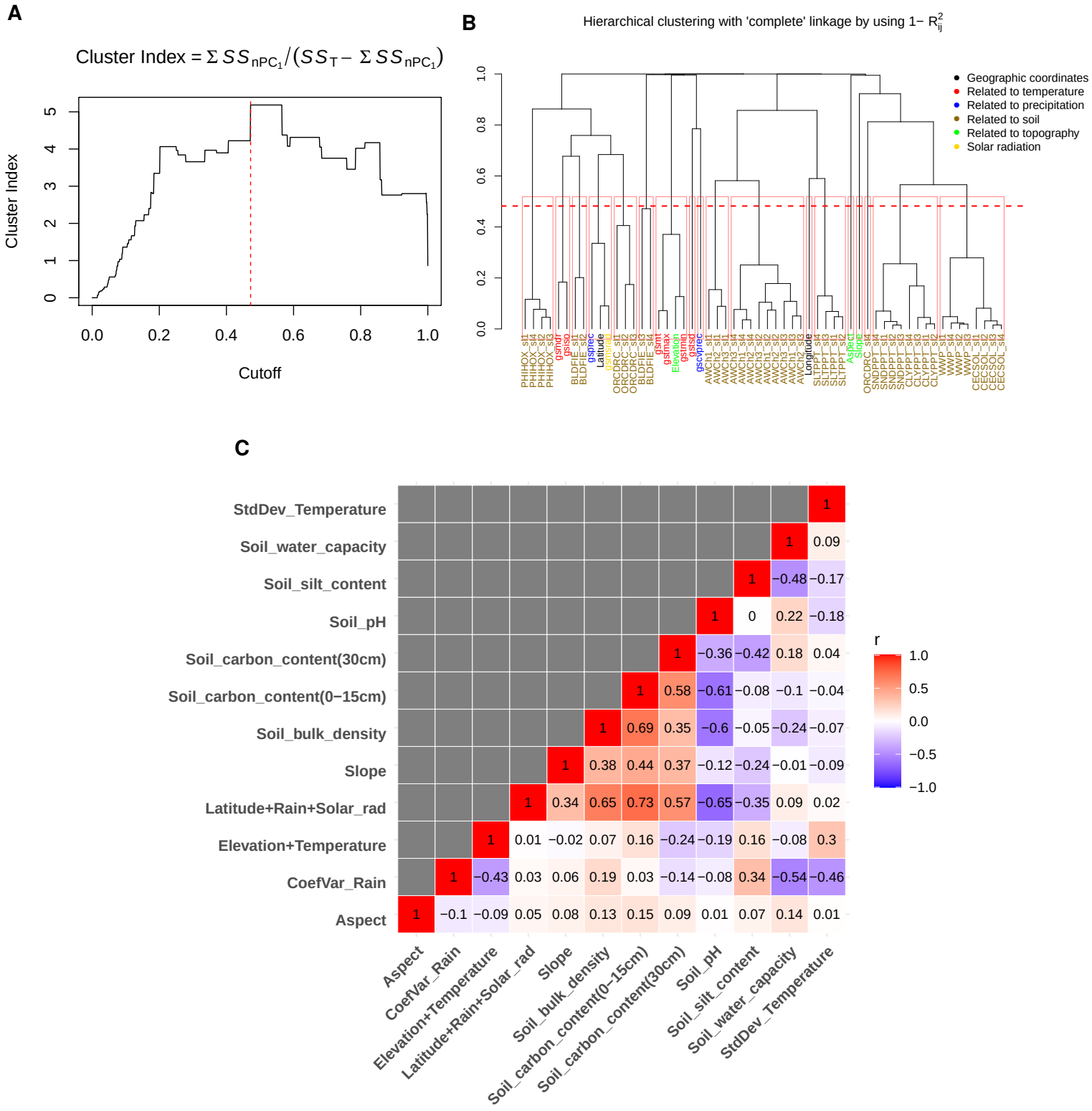
A
Cluster Index = Σ SS_nPC1 / (SS_T − Σ SS_nPC1)
Cluster Index
Cutoff
B
Hierarchical clustering with 'complete' linkage by using 1− R²ij
Geographic coordinates
Related to temperature
Related to precipitation
Related to soil
Related to topography
Solar radiation
C
StdDev_Temperature
Soil_water_capacity
Soil_silt_content
Soil_pH
Soil_carbon_content(30cm)
Soil_carbon_content(0−15cm)
Soil_bulk_density
Slope
Latitude+Rain+Solar_rad
Elevation+Temperature
CoefVar_Rain
Aspect
r

Figure A.3

D

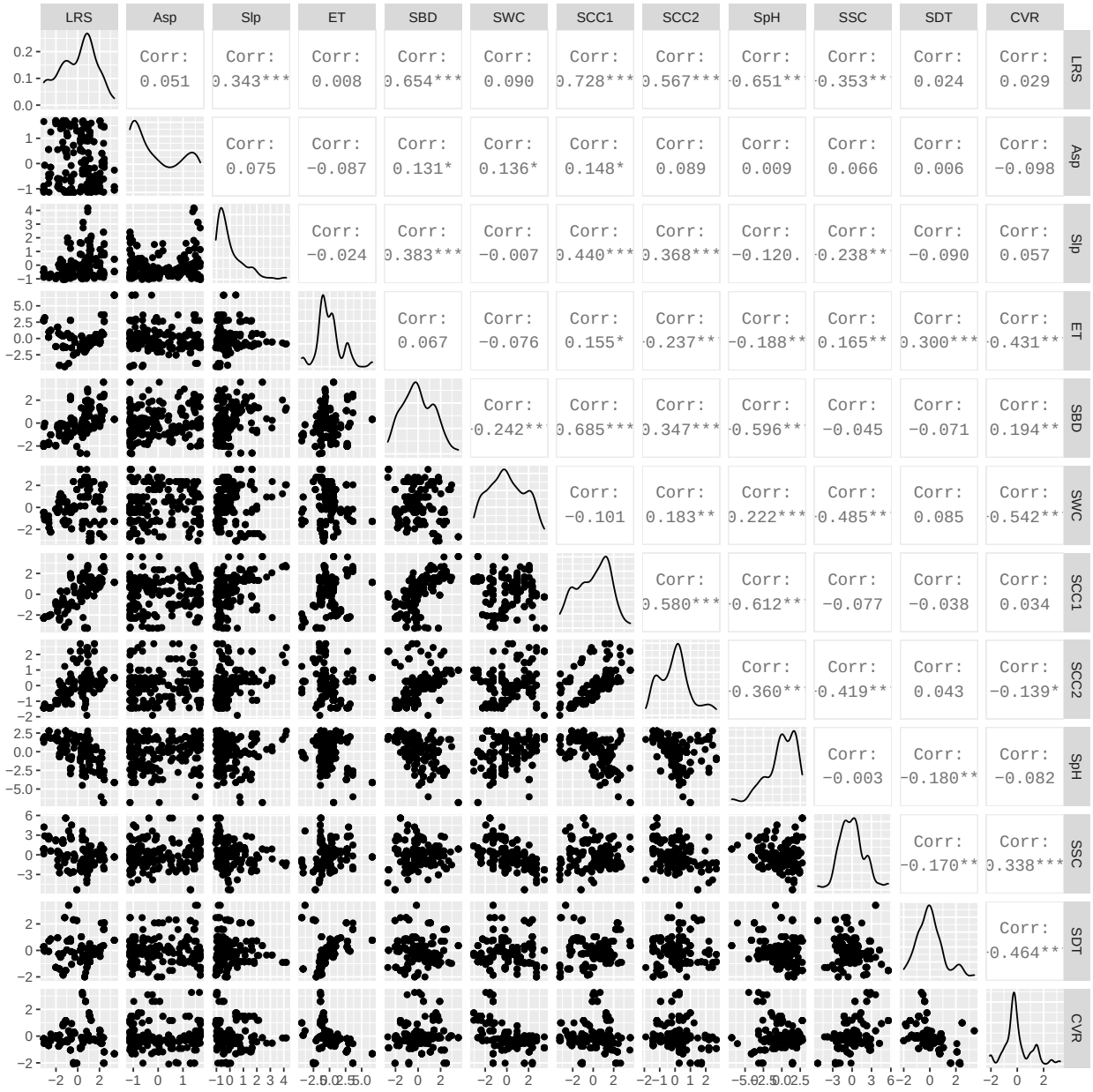

Figure A.3 Procedure of generating synthetic environmental variables. (A) Relationship of cutoff for clustering and corresponding cluster index. The y-axis represents the customized cluster index, and the x-axis represents $1 - R^2$ that is used as a cutoff for the dendrogram. The vertical red dash line shows the optimal cutoff, which is 0.472 (B) Dendrogram of hierarchical clustering based on the $1 - R^2$. The horizontal red dash line shows the optimal cutoff used to determine the clustering combinations for generating synthetic environmental variables. (C) Correlation between the retained 12 environmental variables. (D) Scatterplot panel of the retained 12 environmental variables. Abbreviations: Asp:Aspect, CVR:CoefVar_Rain, ET:Elevation+Temperature, LRS:Latitude+Rain+Solar_rad, Slp:Slope, SBD:Soil_bulk_density, SCC1:Soil_carbon_content(0-15cm), SCC2:Soil_carbon_content(30cm), SpH:Soil_pH, SSC:Soil_silt_content, SWC:Soil_water_capacity, SDT:StdDev_Temperature

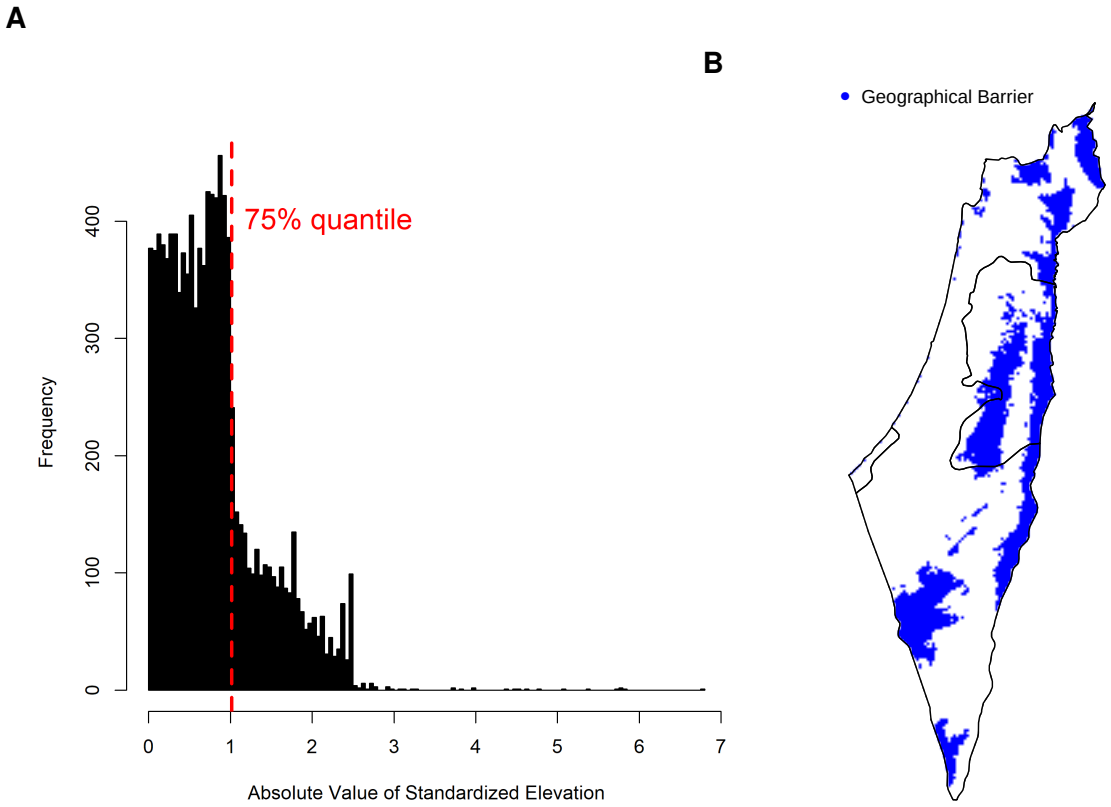

Figure A.4 Classification of barrier and non-barrier pixels. (A) Histogram of absolute value of standardized elevation. The map pixels with elevation deviating from the majority of pixels are defined as geographical barriers. The top 25% cutoff is determined by the drastic decrease of pixel counts. (B) Geographical distribution of barrier pixels.

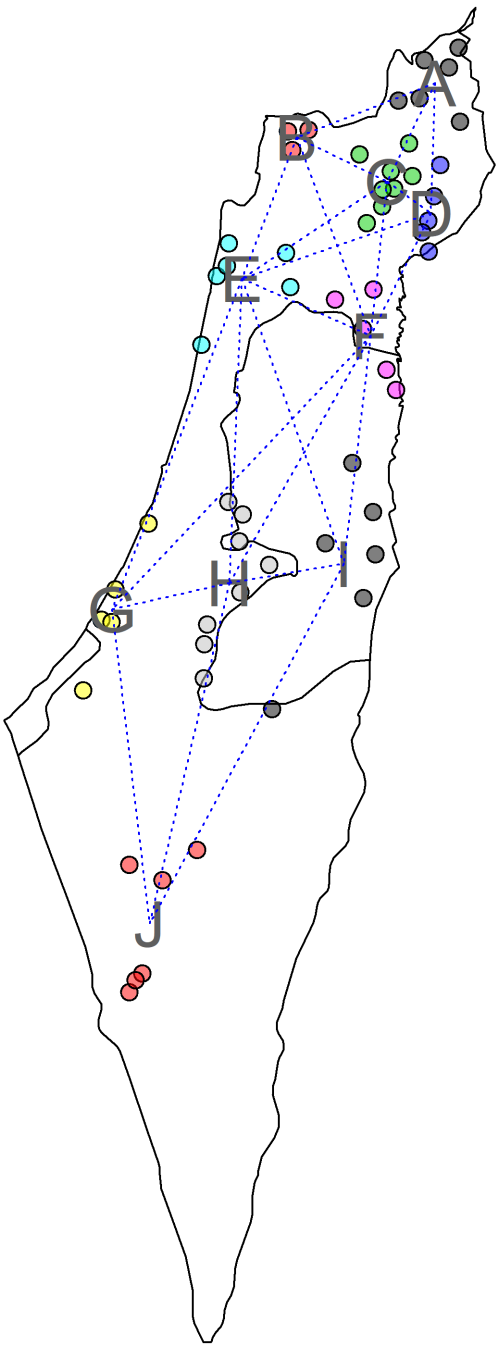

Figure A.5 Geographical clusters and network used for the coalescent-based gene flow inference. The blue dash lines represent the edges of network which allows the movement between the connected geographical regions in the gene flow model.

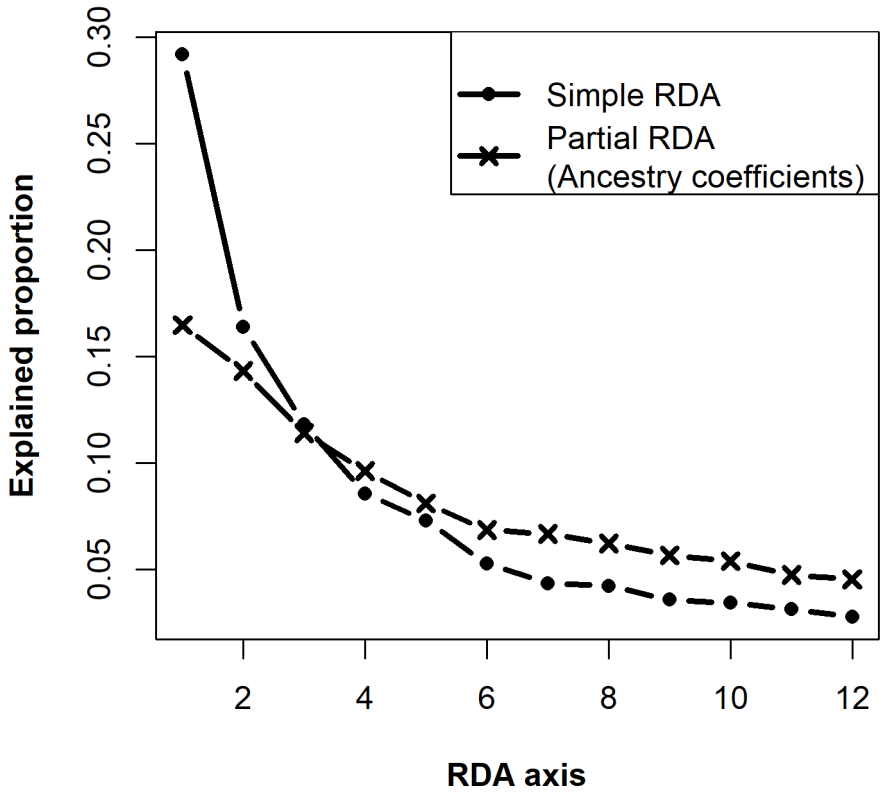

Figure A.6 Explained proportion of variation of RDA axes. The lines with dots and crosses show the result of simple RDA and partial RDA conditioned on population structure, respectively. The first four RDA axes are selected to compute Mahalanobis distances.

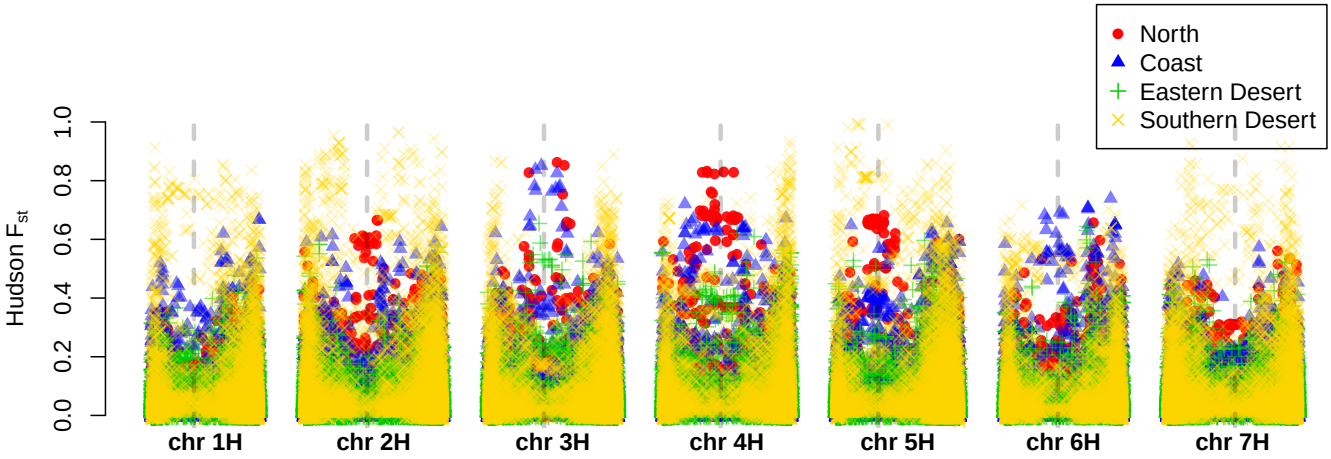

Figure A.7 F_{ST} values along genome. The position of centromeres is shown by the vertical gray dash lines. The dots represent F_{ST} calculated between a specific genetic cluster and three remaining genetic clusters as a whole.

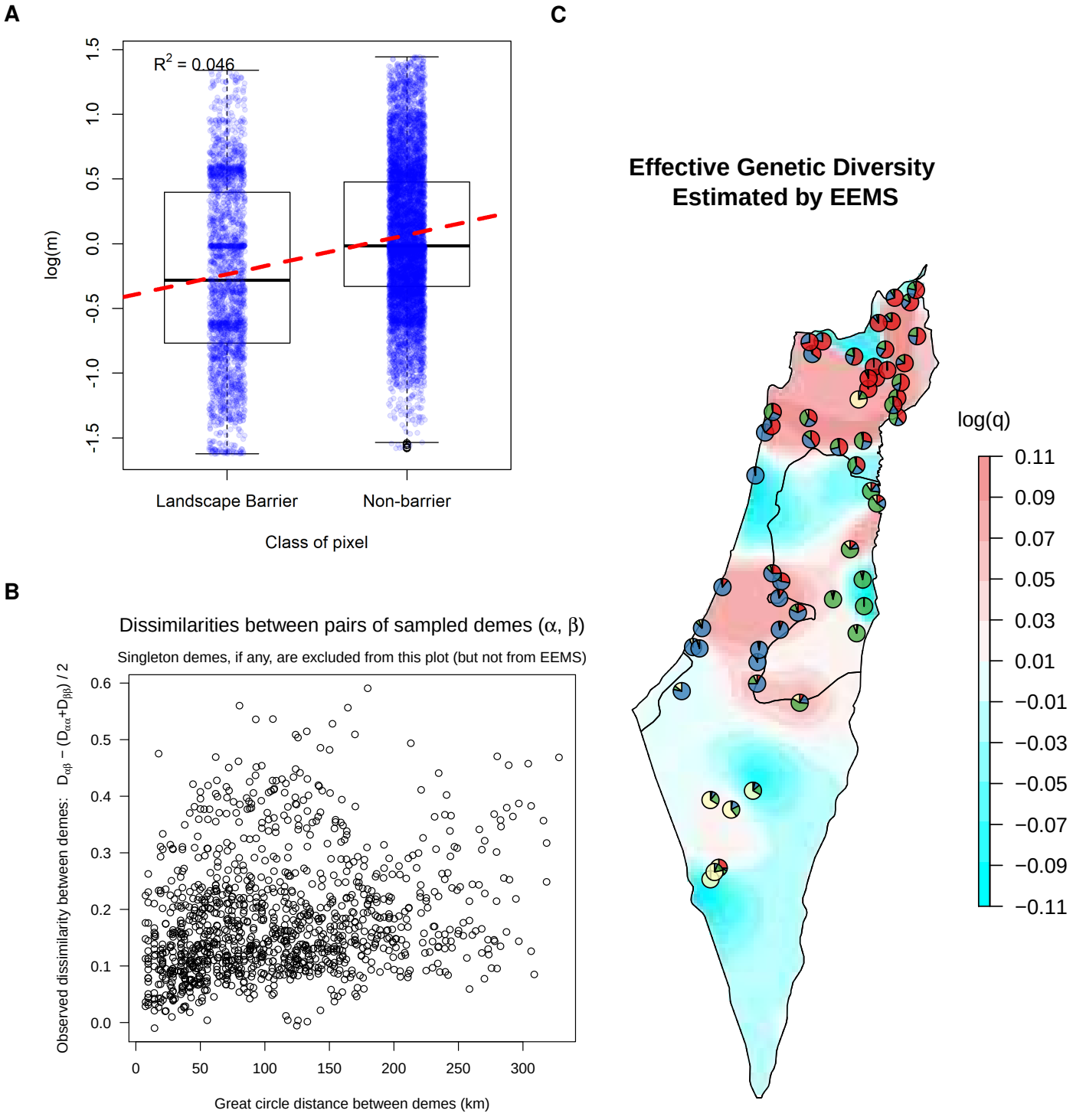

Figure A.8 Result of EEMS. (A) Logarithm of gene flow rate ($log[m]$) estimated by EEMS classified according to barrier and non-barrier pixels. (B) Relationship between observed genetic dissimilarity computed by EEMS and geographical distances. (C) Effective genetic diversity surface estimated by EEMS. Effective genetic diversity (q) is the expected genetic dissimilarity of two individuals sampled from a location. The pie charts represent the average ancestry coefficients of accessions in each deme.

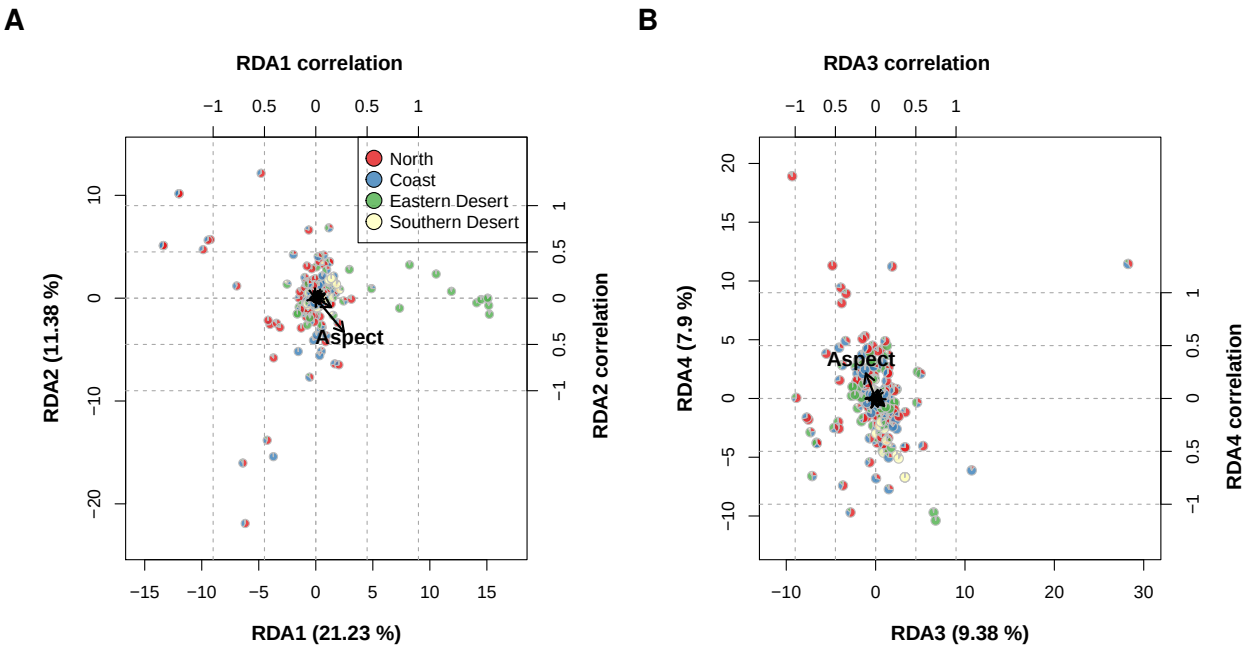

Figure A.9 Biplot of RDA conditioned on spatial autocorrelation. (A) Biplot of the first and second RDA axes. (B) Biplot of the third and fourth RDA axes. The arrows represent correlations of the environmental variables with RDA axes that are shown in Table A.6 in details. The pies represent ancestry coefficients of individuals. The coordinates of pies correspond to site scores of individuals on RDA axes.

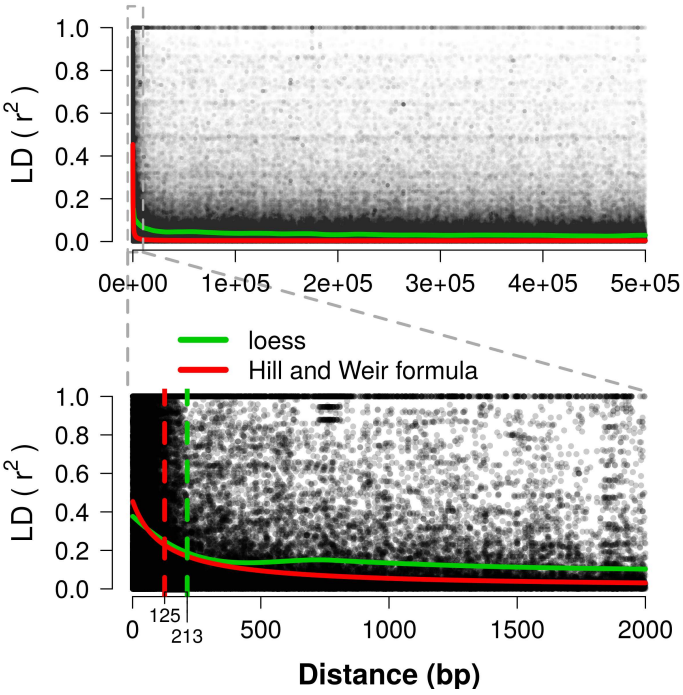

Figure A.10 Linkage disequilibrium decay of the B1K+ collection across genome zooming in the first 500 kb and 2 kb. The green and red lines represent the fitted r^2 values by using *loess* method and Hill and Weir formula. The green and red dash lines represent the decay to the half of the highest fitted r^2.

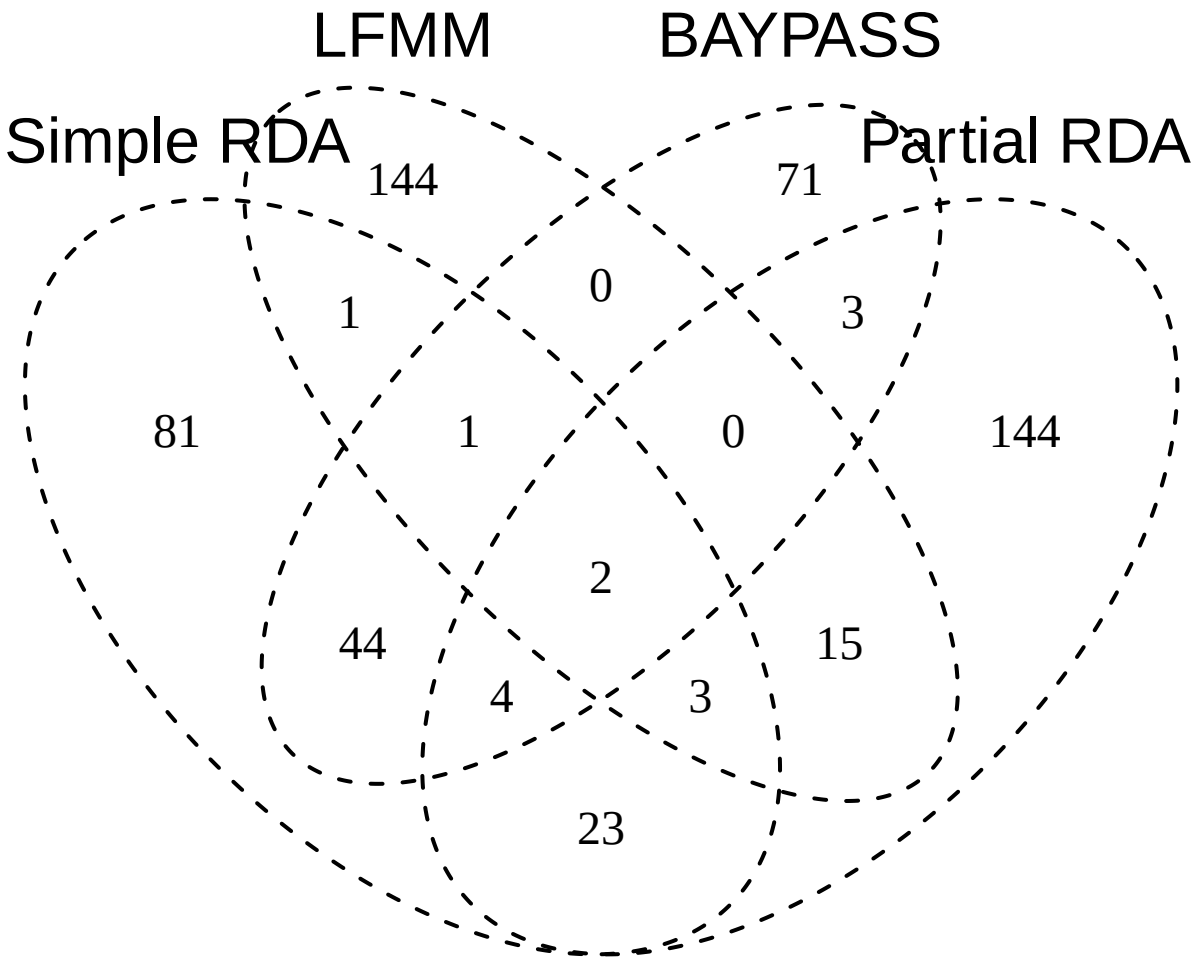

Figure A.11 Number of genes locating within 500 bp adjacent intervals of significant SNPs detected by the GEA and outlier methods.

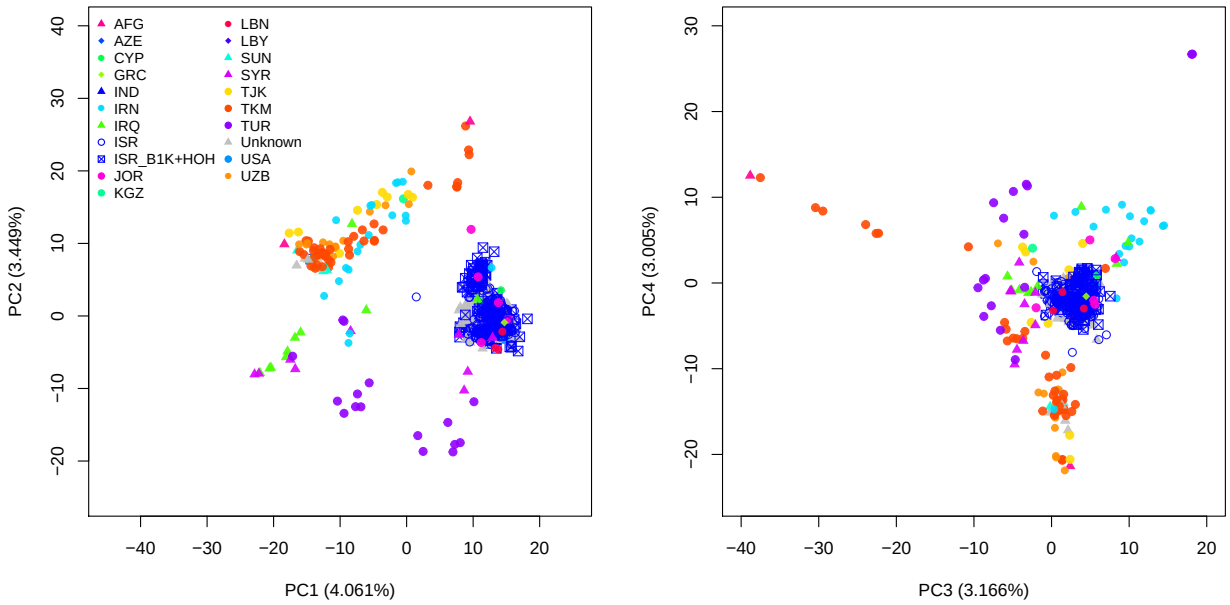

Figure A.12 PCA plots of 244 B1K+ accessions and 1,121 accessions from IPK's genebank with countries of origins. (A) Result of PCA performed by using all of the available accessions. (B) Result of PCA performed by using 72 geographically diverse accessions and projecting the remaining accessions to PC spaces. The 244 B1K+ accessions are represented by blue boxes with crosses inside. The blue open dots represent IPK accessions originating in Israel and the gray closed triangles represent IPK accessions with unknown origins. The country abbreviations are according to ISO 3166-1 α-3.

A.6 Supplementary Files

File S1 Geographical coordinates and environmental data of 244 B1K+ used in this study. Both raw environmental data and twelve environmental variables used in genome-environment association analysis are included in the file. Grouping information of 58 demes and 10 manually defined regions used in gene flow analysis are also included.

File S2 A demonstration showing the influence of marker number and additional samples (HOH accessions) on the result of population structure analyses.

File S3 Statistics of the genome-wide scan methods for 27,147 SNPs, including X^TX of *BAYPASS*, and the p and q values of the simple RDA, partial RDA and LFMM.

File S4 Annotations of genes in the upstream or downstream 500 bp of the candidate SNPs detected by the genome scan methods (*BAYPASS*, simple RDA, partial RDA and LFMM).

File S5 A table of European Nucleotide Archive (ENA) sample IDs.

Appendix B

GGoutlieR: an R package to identify and visualize unusual geo-genetic patterns of biological samples - Supplementary information

B.1 Methods

To facilitate the investigation of geo-genetic patterns, we introduced a heuristic framework, named **G**eo-**G**enetic **outlieR** (*GGoutlieR*). It quantifies the deviation from isolation-by-distance expectation on an individual basis, providing a data-driven baseline to identify outliers. The analytical framework is available in the R package *GGoutlieR*. Also, the package supports the visualization of unusual geo-genetic patterns. We described the *GGoutlieR* framework in detail below.

B.1.1 Overview

Under the isolation-by-distance (IBD) assumption, the association between geographical distances and genetic distances of individuals enables the prediction of geographical origins based on genetic variation and, reversely, the prediction of genetic components according to geographical origins. In both prediction scenarios, prediction errors of a given sample increase proportionally to the degree of deviation from the IBD expectation. Based on the concept mentioned above, we implemented the K-nearest neighbors (KNN) regression, a non-parametric and non-linear method, to characterize the geo-genetic relationship of each sample with its corresponding nearest neighbors. The prediction errors of KNN regression can be converted into distance-based statistics (hereafter named D statistics), which are assumed to follow a Gamma distribution. Statistical tests are conducted based on the D statistics to spot outlier samples deviating from the IBD expectation. The analytical framework is performed with the R function *ggoutlier* in the *GGoutlieR* package. Three approaches are available by setting the argument *method* in the function *ggoutlier*. We introduced them sequentially below.

B.1.2 Assumptions

To develop the *GGoutlieR* framework, we assume that IBD patterns are pervasive among samples, so the population genetic structure is generally accordant to geographical habitats.

Let geographical distribution and genetic variation of samples be described with two coordinate systems, S_{geo} and $S_{genetic}$. We assume that the coordinates of an individual in $S_{genetic}$ are predictable from its neighbors in S_{geo} and vice versa if the IBD assumption holds.

We further assume that the prediction errors, defined as geographical distances between true and predicted coordinates in S_{geo}, follow a Gamma distribution with unknown parameters, written as $\Gamma_{geo}(\alpha, \beta)$. This assumption is made with the expectation of prediction errors approximating zero under IBD. Similarly, the mean of squared prediction errors of coordinates in $S_{genetic}$ is assumed to follow $\Gamma_{genetic}(\alpha, \beta)$.

B.1.3 Geo-genetic outlier detection with K nearest neighbors

Definition of coordinate system S_{geo} and $S_{genetic}$

The sample coordinates in S_{geo} are defined as geographical coordinates of collection sites with decimal degree format. For $S_{genetic}$, we used ancestry coefficients (Pritchard et al. 2000) to represent samples' coordinates for the empirical applications because ancestry coefficients are more interpretable and easier to visualize on a geographical map than principal component values. We regard a matrix of ancestry coefficients ($Q_{N \times F}$) estimated based on F ancestral populations as the coordinates of N samples distributing in a space $S_{genetic}$ with F dimensions.

Approach 1: outlier identification with geographical KNNs

The first approach in the *GGoutlieR* framework aims at identifying outliers that are genetically differentiated from their geographical KNNs.

Step 1. Compute pairwise geographical distance matrix.

Step 2. Find KNNs for each individual according to pairwise geographical distances with a given K. To avoid a divisor of zero in the equation B.1 of **Step 3**, *GGoutlieR* will ignore neighbors within 100 meters by the default (controlled by the *min_nn_dist* argument).

Otherwise, one unit of distance is added to the off-diagonal values of geographical distance matrix before searching KNNs if any pairwise distance is zero and *min_nn_dist* is set to zero.

Step 3. Predict $\hat{x}_{genetic,i,j}$ using a weighted KNN approach. $\hat{x}_{genetic,i,j}$ is the predicted coordinate of an individual i in the dimension j of $S_{genetic}$, where $i = \{1, 2, ..., N\}$ and $j = \{1, 2, ..., F\}$. N is the number of individuals and F is the number of dimensions in $S_{genetic}$, i.e. the number of ancestral populations. The weight of the k th nearest neighbor of an individual i is computed as

$$w_{i,k} = \frac{\frac{1}{d_{i,k}}}{\sum_{k=1}^{K} \frac{1}{d_{i,k}}} \tag{B.1}$$

where $d_{i,k}$ is geographical distance between the individual i and its k th nearest neighbor. $\hat{x}_{genetic,i,j}$ is calculated as

$$\hat{x}_{genetic,i,j} = \frac{1}{K} \sum_{k=1}^{K} w_{i,k} \; x_{genetic,i,j,k} \tag{B.2}$$

where K is a given number of nearest neighbors. The default of *GGoutlieR* searches the optimal K with a range of values (see **Step 5.1**). $x_{genetic,i,j,k}$ is the coordinate of k th neighbor of individual i in the dimension j of $S_{genetic}$.

Step 4. Compute mean of squared prediction errors as

$$D_{genetic,i} = \frac{1}{F} \sum_{j=1}^{f} \hat{\varepsilon}_{i,j}^2 = \frac{1}{F} \sum_{j=1}^{f} (x_{genetic,i,j} - \hat{x}_{genetic,i,j})^2 \tag{B.3}$$

where $x_{genetic,i,j}$ and $\hat{x}_{genetic,i,j}$ are the true and predicted coordinates of the individual i in the dimension j of $S_{genetic}$, respectively.

Step 5.1. Search optimal number of nearest neighbors (K) by minimizing $\sum_{i=1}^{n} D_{genetic,i}$. The **Step 1-4** are repeated with a range of K values (the default is from 3 to 50). As $D_{genetic}$ represents the size of prediction errors, we define optimal K as the K value resulting in the lowest $\sum_{i=1}^{n} D_{genetic,i}$.

Step 5.2. Repeat **Step 1-4** with the optimal K.

Step 6. Obtain an empirical null distribution $\Gamma_{genetic}(\alpha, \beta)$. α and β are evaluated by maximum likelihood estimation.

Step 7. Test individuals with the empirical null distribution $\Gamma_{genetic}(\alpha, \beta)$ from **Step 6**. The null hypothesis is that a focal individual follows the IBD expectation, whereas the alternative hypothesis is that a focal individual is genetically differentiated from its K geographically nearest neighbors. Considering that a true outlier may induce the significance of its neighbors, we perform the test in a multi-stage manner. In each iteration, we drop the most significant individual and repeat the **Step 2-4** to exclude the influence from the most significant outlier. This procedure is repeated until no outlier is identified with a given significant level.

To use the genetic KNN approach, users have to set the argument *method = "geoKNN"* for the *ggoutlier* function.

***Approach 2*: outlier identification with genetic KNNs**

The second approach of the *GGoutlieR* framework aims at identifying outliers that are geographically remote from genetically similar individuals, i.e. their corresponding KNNs in $S_{genetic}$. The rationale is similar to the first approach as described in the previous section.

Step 1. Compute pairwise Euclidean distances according to a given matrix of genetic components, i.e. ancestry coefficients. If any pairwise distance is zero, 10^{-6} is added to the off-diagonal values of the genetic distance matrix to avoid a divisor of zero in the equation B.4 of **Step 3**. As an alternative option, *GGoutlieR* accepts a distance matrix given by users in this step if users prefer a customized calculation of individual-based genetic distances.

Step 2. Find KNNs for each individual according to pairwise genetic distances with a given K.

Step 3. Predict $\hat{x}_{geo,i,j}$ using a weighted KNN approach. $\hat{x}_{geo,i,j}$ is the predicted coordinate of an individual i in the dimension j of S_{geo}, where $i = \{1, 2, ..., N\}$ and $j = \{1, 2\}$. N is the number of individuals and j corresponds to longitude and latitude. The weight of the k th nearest neighbor of an individual i is computed as

$$w_{i,k} = \frac{\frac{1}{d_{i,k}^2}}{\sum_{k=1}^{K} \frac{1}{d_{i,k}^2}} \tag{B.4}$$

where $d_{i,k}$ is genetic distance between the individual i and its k th nearest neighbor computed in the **Step 1**. $\hat{x}_{geo,i,j}$ is calculated as

$$\hat{x}_{geo,i,j} = \frac{1}{K} \sum_{k=1}^{K} w_{i,k}\ x_{geo,i,j,k} \tag{B.5}$$

where K is a given number of nearest neighbors. The default of *GGoutlieR* searches the optimal K with a range of values (see **Step 5.1**). $x_{geo,i,j,k}$ is the coordinate of k th neighbor of individual i in the dimension j of S_{geo}.

Step 4. Compute prediction errors as

$$D_{geo,i} = GeoDist(x_{geo,i}, \hat{x}_{geo,i}) \tag{B.6}$$

where $GeoDist(x_{geo,i}, \hat{x}_{geo,i})$ is the geographical distance between the true and predicted locations of the individual i, which is calculated with the *geosphere* package (Hijmans 2019).

Step 5.1 Search optimal number of nearest neighbors (K) by minimizing $\sum_{i=1}^{n} D_{geo,i}$. The **Step 1 - 4** are repeated with a range of K values. The K value resulting in the lowest $\sum_{i=1}^{n} D_{geo,i}$ is considered as the optimal K for the given data set. The default of *GGoutlieR* tests a range K from 3 to 50.

Step 5.2 Repeat **Step 1-4** with the optimal K.

Step 6 Obtain an empirical null distribution $\Gamma_{geo}(\alpha, \beta)$. α and β are identified by maximum likelihood estimation.

Step 7 Test individuals with the empirical null distribution $\Gamma_{geo}(\alpha, \beta)$. The null hypothesis is that a focal individual follows the IBD expectation. The alternative hypothesis is that a focal individual is geographically remote from K individuals that are genetically most similar to a focal individual. The test is carried out in a multi-stage manner as described in the **Step 7** of

Approach 1.

To use the geographical KNN approach, users have to set the argument *method = "geneticKNN"* for the *ggoutlier* function.

***Approach 3*: composite approach**

The geographical KNN and genetic KNN approach above attempt to identify geo-genetic outliers from different perspectives. To leverage both approaches, a composite method first carries out the **Step 1 - 6** of geographical KNN and genetic KNN approaches. Next, instead of doing multi-stage tests (**Step 7**) separately, the composite approach sequentially removes the most significant outlier among the results of two KNN approaches and then repeats the KNN searching and p-value computation to identify outliers with two KNN approaches. This iterative procedure continues until no new outlier raises with the given significant threshold.

To use the composite KNN approach, users have to set the argument *method = "composite"* for the *ggoutlier* function.

Appendix C

Predicting the geographical origins of barley genebank accessions using deep learning: Can large sample sizes improve genome-environment association studies? - Supplementary information

C.1 SLiM simulation

C.1.1 Mutational effect of QTLs

In our simulation, we assumed that the environmental variables of all sites are located in a 95% interval of the expected genetic variation. The expected genetic variation is calculated as $\sigma_g^2 = 2N\sigma_{qtl}^2$, where N is the number of QTLs ($N = 100$) and 2 is for diploidy. Let the boundary of the 95% interval of the expected genetic variation be $\pm c\sigma_g$, where c is a constant, and let the most extreme environmental variable after centering be x. With our assumption, $|x|$ should be equal to or less than $c\sigma_g$. Therefore, we have $c\sigma_g \geq |x|$. We can rewrite it as $c\sqrt{2N\sigma_{qtl}^2} \geq |x|$, and it gives $\sigma_{qtl} \geq \sqrt{\frac{|x|^2}{2Nc^2}}$. With the equation above, we set $\sigma_{qtl} = 0.45$.

C.1.2 Plasticity of selection

We determined the plasticity, standard deviation (SD) of fitness bell curve, based on the environmental contrasts between connected sites. This was done by calculated the absolute difference between the environmental variables of the two connected sites, denoted as $C_{env} = |Env_i - Env_j|$, where site i is a connected neighbor of the site j with gene flow.

To simulate a sufficiently strong isolation by environment, we assumed that 90% of C_{env} values fall within a 95% interval of the fitness bell curve. Among the selected 312 sites, we found that 90% of C_{env} values were less than 5.7. To set the plasticity parameter ($\sigma_{plasticity}$) for the SLiM simulation, we chose a value of 2.85, such that $2\sigma_{plasticity} = 5.7$, approximately covering the 95% interval under a normal distribution.

C.2 Supplementary Figures

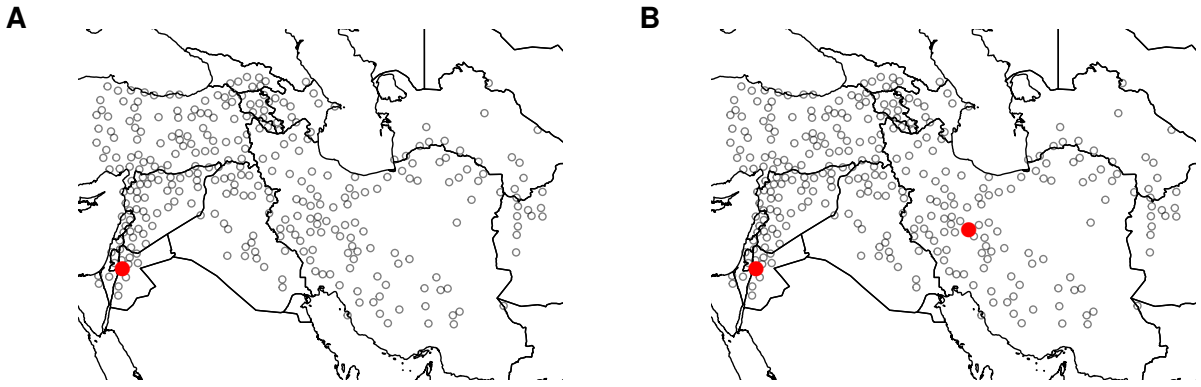

Figure C.1 Geographical distribution of sub-populations in two demographic scenarios of SLiM simulation. **A**. Population expansion from one refugium (1R). **B**. Population expansion from two refugium (2R). Red dots indicate the starting points of population expansion.

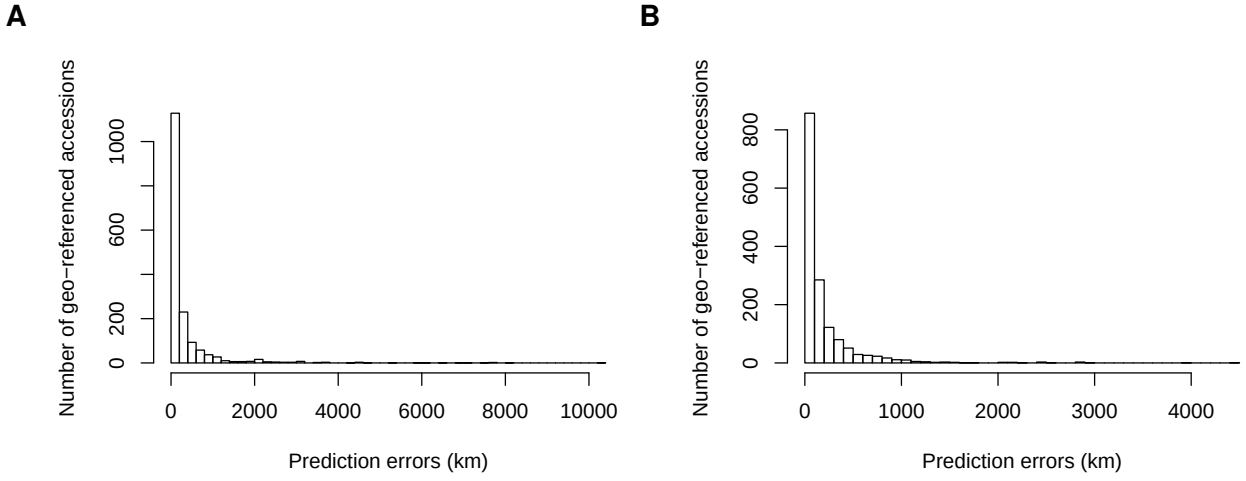

Figure C.2 Mean prediction errors in kilo-meters estimated from cross-validation of IPK barley landrace collection. **A**. Prediction errors of original samples. **B**. Prediction errors of samples excluding outliers with unusual geo-genetic patterns.

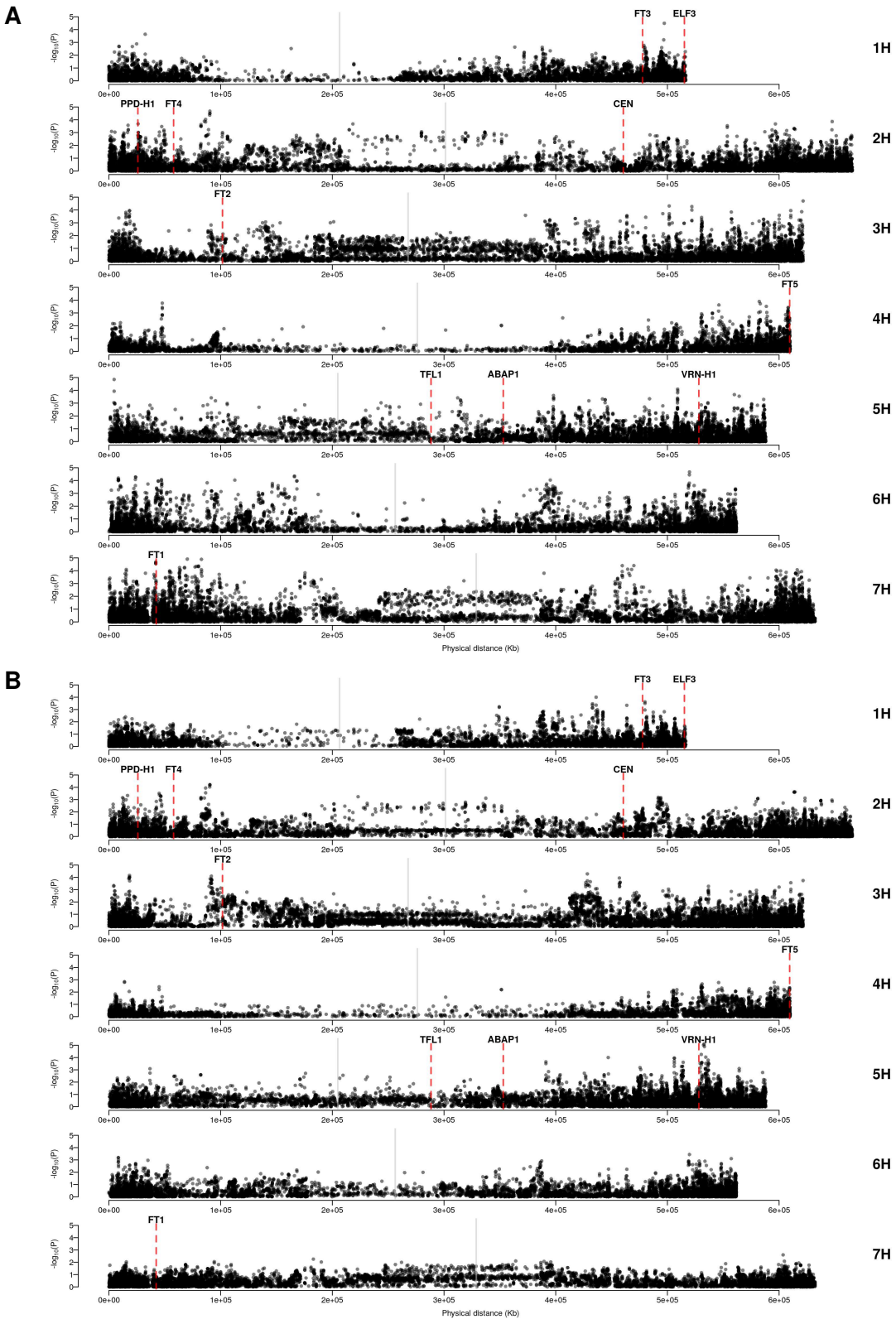

Figure C.3 Regular GEA of IPK landraces with the environmental principal component (PC). **A**. Regular GEA with environmental PC2. **B**. Regular GEA with environmental PC3. Blue and red horizontal lines are the significant levels of FDR = 0.05 and FDR = 0.01. Grey vertical lines indicate the positions of centromeres. Red dashed lines indicate the positions of flowering time genes.

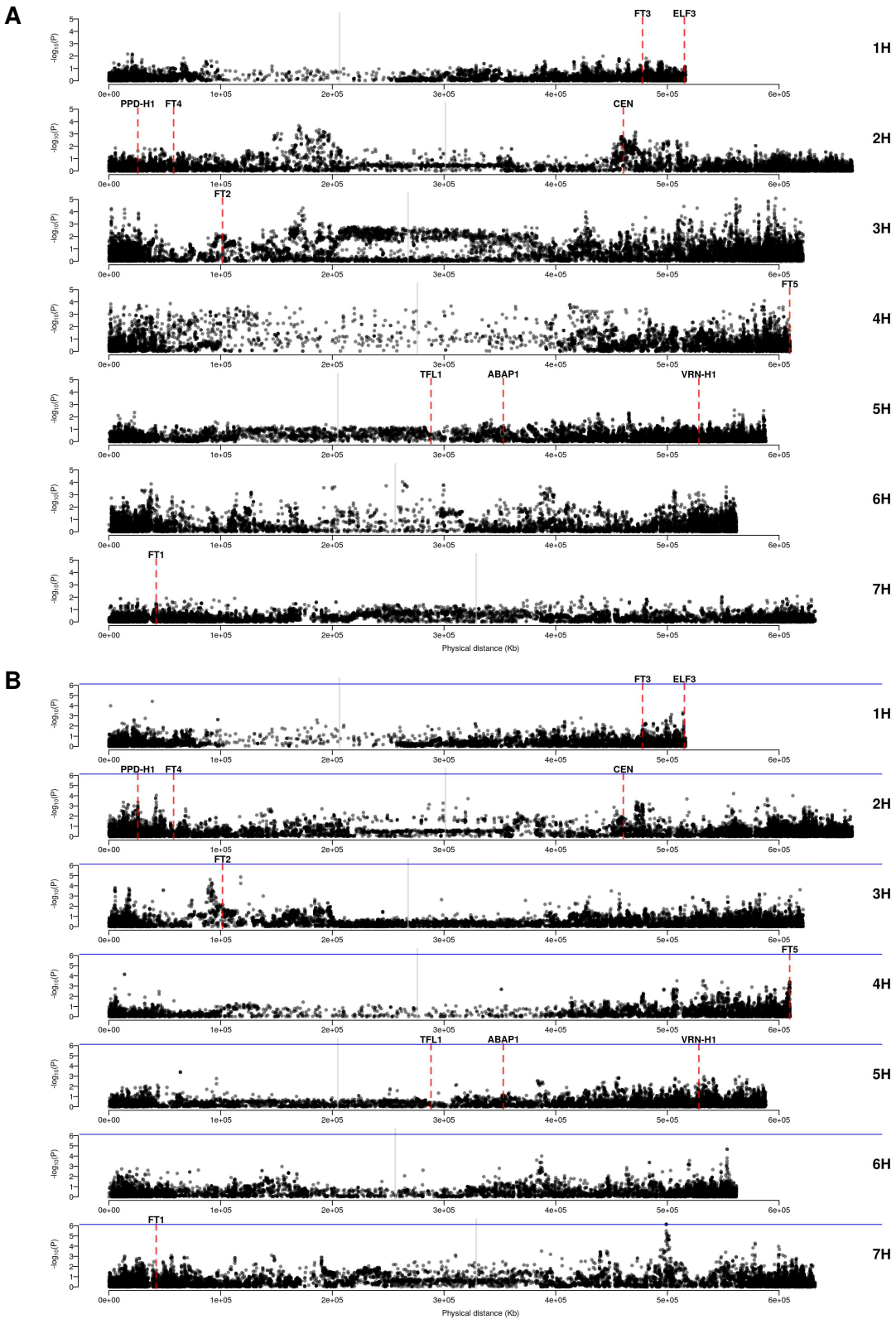

Figure C.4 *GEAplus* of IPK landraces with the environmental principal component (PC). **A**. *GEAplus* with environmental PC1. **B**. *GEAplus* with environmental PC3. Blue and red horizontal lines are the significant levels of FDR = 0.05 and FDR = 0.01. Grey vertical lines indicate the positions of centromeres. Red dashed lines indicate the positions of flowering time genes.

Appendix D

Curriculum Vitae

Che-Wei Chang
Date and Place of Birth: 17 December 1992, Taipei, Taiwan

Work Experience

July 2023 – present	Research Associate Quantitative Genetics	Breeding Technology Centre (BTC), Dümmen Orange B.V., the Netherlands
August 2018 – September 2018	Research Assistant	Department of Agronomy, National Taiwan University, Taiwan
July 2017 – July 2018	Mandatory Military Service	Taiwan

Education

December 2018 – present	Doctoral student in Crop Biodiversity and Breeding Informatics	University of Hohenheim, Germany
July 2015 – June 2017	Master of Science in Genetics and Molecular Breeding	National Taiwan University, Taiwan
July 2012 – June 2015	Bachelor of Science in Agronomy (transferred from National Chung Hsing University in July 2012)	National Taiwan University, Taiwan
July 2011 – June 2012	Bachelor of Science in Agronomy	National Chung Hsing University, Taiwan

Delft, the Netherlands, 26.01.2025

Annex 3

Declaration in lieu of an oath on independent work

according to Sec. 18(3) sentence 5 of the University of Hohenheim's Doctoral Regulations for the Faculties of Agricultural Sciences, Natural Sciences, and Business, Economics and Social Sciences

1. The dissertation submitted on the topic

Exploring adaptive genetic variation in exotic barley germplasm with landscape genomics

is work done independently by me.

2. I only used the sources and aids listed and did not make use of any impermissible assistance from third parties. In particular, I marked all content taken word-for-word or paraphrased from other works.

3. I did not use the assistance of a commercial doctoral placement or advising agency.

4. I am aware of the importance of the declaration in lieu of oath and the criminal consequences of false or incomplete declarations in lieu of oath.

I confirm that the declaration above is correct. I declare in lieu of oath that I have declared only the truth to the best of my knowledge and have not omitted anything.

Delft, the Netherlands, 26.01.2025 [redacted]

Place, Date Signature