

Population Genetics of *Populus trichocarpa* for Targeted Breeding as a Biofuel Crop

A Dissertation Presented for the
Doctor of Philosophy
Degree

The University of Tennessee, Knoxville

Cai W. John

August 2022

© by Cai W. John, 2022
All Rights Reserved.

To my parents, friends and mentors who have supported me through my educational career.

Thank you for allowing me the time and space to both learn and grow.

Acknowledgments

I would like to thank the handful of professors and mentors who made time over the years to listen to the incessant questioning of a student. Your patience and guidance have allowed me to become the thinker that I am. Specifically, I would like to thank Dr. Scott J. Emrich and Dr. Wellington Muchero for supporting me through this doctoral journey. The freedom you have granted me has been essential to my growth.

Abstract

Populus trichocarpa (poplar) is a woody species native to the western U.S. and Canada. As a fast-growing crop, it has been under investigation by the Department of Energy as a resource for liquid biofuel production. Having recently expanded the collection of poplar whole-genome sequences so that it spans the entire natural species range, we have the novel opportunity to study adaptive responses across this range. This work starts with an initial proof of concept of methodologies in a well-studied portion of the species range. This region has complete whole-genome sequences and RNA expression data. The completeness of these data allow robust validation of the methods used. Having demonstrated our methods, we expand our study to include the unstudied portion of poplar's range. Using these data, we map nine novel genomic loci as regulators of climactic response. This comes with the implication that poplar's response to abiotic stressors are not a genome-wide phenomenon, but restricted to a small number of key regulators. This novel finding has massive implications on the feasibility of engineering poplar for climate resiliency, and enhanced biomass production. The final study leverages this understanding to investigate the predictive capacity that resides in the genome for these climate and geo-spatial traits. Initial results highlight promising avenues for future investigations using our association mapping targets. We are also able to use this prediction architecture to identify mislabeled genotypes, and subsequently correct their labels.

Table of Contents

1	Introduction	1
1.1	Poplar as a Biofuel Crop	2
1.2	Population Genetics and Poplar	2
1.3	Research Contribution	3
2	Extreme Phenotype Sampling Improves LASSO and Random Forest Marker Selection for Complex Traits	5
2.1	Abstract	6
2.2	Introduction	6
2.3	Methods	8
2.3.1	Multi-omics data	8
2.3.2	Analysis of inherent structure in these data	10
2.3.3	Why Binary Prediction?	10
2.3.4	Classification in poplar	10
2.3.5	Removing confounding factors in malaria expression data	14
2.4	Results	15
2.4.1	Feature Selection of poplar RNA-seq Data	15
2.4.2	Genetic Validation of Expression Results	18
2.5	Applying Techniques to Less Structured Data	18
2.5.1	Feature selection	21
2.5.2	Biological validation	22
2.6	Discussion	25

3	Environmental GWAS Reveals Hot Spots for Climate Adaptability in <i>Populus trichocarpa</i>	29
3.1	Abstract	30
3.2	Introduction	30
3.3	Methods	32
3.3.1	Data Preparation	32
3.3.2	Genome-wide Association Study	32
3.3.3	Allele Frequency Analysis	33
3.3.4	RNA-seq Functional Enrichment	34
3.4	Results	34
3.4.1	Climate Characterization of Species Range	34
3.4.2	Allelic Patterns of Adaptation in Hot Spots	35
3.5	Discussion	35
4	Correcting Low-Confidence Latitude Labels Using Machine Learning in <i>Populus trichocarpa</i>	44
4.1	Abstract	45
4.2	Introduction	45
4.3	Methods	46
4.3.1	Data	46
4.3.2	MinHashing Unaligned Reads	48
4.3.3	Machine Learning Models for Prediction	49
4.4	Results	50
4.4.1	Performance Across Different Sketch Sizes	50
4.4.2	Performance Across Data Representations	51
4.4.3	Predicting Latitude for GW Test Set	51
4.5	Discussion	52
5	Conclusions	57
	Bibliography	59

Appendix	65
Vita	86

List of Tables

2.1	Sample sizes of each data collection. The Mok data and our corrected data have the same sample sizes.	16
2.2	Results of Modelling on Poplar RNA-seq Data. Transcripts are those selected above our significance cut-offs.	16
2.3	Results of Modelling on Malaria RNA-seq Data. Transcripts are those selected above our significance cut-offs.	24
2.4	Top 10 genes found by EPS LASSO compared to 5-split LASSO and EPS RF. Bold entries have been previously linked to potential antimalaria drugs. Genes are listed in order of EPS LASSO rank.	24
2.5	Comparison of rankings from Ford et al. [9] with our LASSO framework on the Zhu data. Genes are listed in order as ranked by [9].	28
3.1	Strongest associations for each climate variable. We see strong colocalization across these major peaks with many climate variables mapping to exactly the same variant.	42
3.2	Annotating genes neighboring the lead variant for each climate variable. . . .	43

List of Figures

2.1	Sample locations of Poplar trees colored by river system.	9
2.2	PCA of Malaria Expression data from Mok et al. [22].	12
2.3	PCA of malaria probes post-regression correction. Substantially less structure driven by life-stage group can be seen in this projection, especially between groups A and B.	13
2.4	Transcript expression for Potri.010G079500 after removing intermediate phenotype samples, demonstrating a clear Columbia vs. non-Columbia difference in expression.	17
2.5	GWAS analysis using Columbia/non-Columbia extremes as the desired phenotype on the poplar data.	17
2.6	PCA of Zhu data [42] using all transcripts. As expected, there is no clear separation of classes.	19
2.7	PCA of Zhu data [42] using LASSO selected transcripts with a slight separation of classes.	20
3.1	A) Sampling locations of previously studied WGS from Slavov et al. (2012) [30] combined with our 207 newly sequenced genotypes. B) Sampling locations of leaf RNA-seq data used for functional enrichment	38
3.2	Collection of GWAS results for climate traits with major peaks from chromosome 1 to 14. Results are ordered by major peak demonstrating the extent of colocalization across the climate traits.	39

3.3	Collection of GWAS results for climate traits with major peaks from chromosome 14 to 19. Results are ordered by the major peak demonstrating the extent of colocalization across the climate traits.	40
3.4	Climate based GWAS implicates strong regulators on Chr13 for seasonality of precipitation and Chr01 for diurnal range. A) Manhattan plot from GWAS based on genotypes shown in panel B, major peak colored in green. B) Climate map of response variable used in GWAS, oriented with North to the right to align with the latitude scale in panel C. Lower quartile genotypes are marked with open triangles and upper quartile genotypes are marked with filled circles. C) Allele frequency shifts computed using all genotypes across the entire species range for variants found in the peak region marked in green in panel A, each line corresponds to the variation of one SNP.	41
4.1	A) Map of the sampling location of all poplar WGS. B) PCA of WGS for the entire poplar dataset. The colorings correspond to our partitioning of the population into broad latitudinal classes that reflect population structure. These colorings illustrate how geographic position in panel A maps to PC coordinates in panel B. C) Showing where the GW-tagged samples map in the PCA.	47
4.2	Comparison of performance across different sketch sizes (s) for sketch-based prediction. Y-axis represents R^2 performance for the regression fit. Each color bar represents a different size of mash sketching, ranging from 500 to 4,000 hash functions. Traits are ordered by performance of the s=4,000 task. . . .	54
4.3	Comparison of performance across different data representations. Y-axis represents R^2 performance for the regression fit. Each color bar represents a different data representation. Traits are ordered by performance on the alignment-based representation.	55

4.4	A) Distribution of the difference between predicted latitude and low-confidence GW latitude labels. Mean value of 0.41 is marked by the vertical black line.	
	B) Examining Willamette river system genotypes within the context of the whole-genome PCA. Genotypes marked in red are Willamette genotypes with high confidence labels that we expect map to the correct position in the PCA. Genotypes marked in blue are GW genotypes that are likely incorrectly labeled as Willamette. Genotypes marked in green are the new neighbors of the GW genotypes based on predicted latitude.	56
1	Isothermality	67
2	Koppen-Geiger Climate Classification	68
3	Max Temperature	69
4	Mean Temperature	70
5	Min Temperature	71
6	Precipitation of the Coldest Quarter	72
7	Precipitation of the Driest Month	73
8	Precipitation of the Driest Quarter	74
9	Precipitation of the Warmest Quarter	75
10	Precipitation of the Wettest Month	76
11	Precipitation of the Wettest Quarter	77
12	Precipitation	78
13	Temp of the Coldest Quarter	79
14	Temp of the Driest Quarter	80
15	Temp Range	81
16	Temp Seasonality	82
17	Temp of the Warmest Quarter	83
18	Temp of the Wettest Quarter	84

19 Chromosome 14 peak region thresholding. This plot is used to select variants in a peak region for visualization in panel C of the climate GWAS figures (i.e., S1-S18). The range threshold is the difference between the maximum and minimum allele frequency of a SNP across the entire species range. We are interested in alleles that show the highest levels of variation. The stepwise pattern observed likely exists due to SNPs in LD with each other. Our goal is to select the LD block containing the SNPs with greatest variation, so in this case a threshold of 0.27 was used and resulted in visualization of 7 SNPs (see S17 panel C). 85

Chapter 1

Introduction

1.1 Poplar as a Biofuel Crop

Populus trichocarpa (poplar) is a woody species native to the Western US and Canada. Its species range extends from southern California all the way to British Columbia encompassing a latitudinal range of 38.88 to 54.25 degrees [30]. It has been under investigation by the Department of Energy (DOE) since the 1970's because of its potential for bioenergy production. More specifically, poplar has remarkably fast growth, allowing it to be harvested every 2-5 years [2], and as a crop can produce energy equivalents of 11 barrels of oil per acre per year [14]. These attributes make it a desirable candidate for biofuel production as these trees produce large amounts of biomass quickly.

Poplar is also well-suited to produce liquid fuel due to its distinct composition of high cellulose and low lignin content. Cellulose provides the carbohydrates needed to produce biofuel, and low lignin content makes this cellulose easier to extract from the biomass relative to other biofuel candidates [pop].

1.2 Population Genetics and Poplar

The advent of high-throughput sequencing technologies has lead to an explosion of whole-genome sequence (WGS) data [8]. This data consists of a long sequence of alleles (either A, T, C, or G) that makes up the chromosomal DNA of the species being sequenced. The position of each one of these alleles is known as a base-pair and a typical whole-genome sequence in humans is made up of about 500 million base-pairs [25].

As populations evolve and reproduce, some of these allele frequencies will change. For a given allele, this is known as a single-nucleotide polymorphism (SNP). These changes (or SNPs) lead to diversity in the population for physical traits or disease outcomes. The goal is to identify regions of the genome (SNPs) that associate with traits of interest, or disease outcomes. Typically, this is done by testing each SNP for association with a trait across the entire genome. This type of study is known as a genome-wide association study (GWAS).

Given their interests in poplar, the DOE has poured resources into the development of a GWAS population for this species. The first poplar genome was sequenced in 2006 [34]

and the population has since been expanded to 1,335 WGS. For the first time since this work began, we now have a population that effectively covers the entire natural species range of poplar. Armed with this, we can now probe the adaptive dynamics that occur between environmental extremes across this species range. Studies that aim to associate abiotic stressors with genomic loci are known as environmental association analysis (EAA).

On top of the WGS data, RNA-seq data has been collected for a subset of this population. To understand RNA-seq data it is important to understand the central dogma of molecular biology which states that DNA makes RNA, and RNA makes proteins, which go on to catalyze reactions in a biological system. RNA-seq data is thus examining the second step of this process. It gives us clues as to how the DNA we study actually gets encoded into proteins. RNA-seq data are represented as counts of the amount of RNA that gets expressed by a portion of DNA. We can use this data to substantiate findings from our GWAS population. It is notable that this data set is much less developed than our GWAS data, consisting of only around 300 different poplar trees.

1.3 Research Contribution

Of major interest to the DOE is the ability to maintain and expand poplar stocks in the face of changing climates. To address this, this work aims to map the genomic hotspots of climate adaptability in poplar. By doing this, it can lead to a mechanistic understanding of how to engineer poplar for optimal growth in any environment. Targeted breeding programs can leverage this knowledge to engineer poplar stocks with adaptive genes and alleles that lead to improved outcomes in the face of changing climates.

This work begins with a proof-of-concept study demonstrating that grouping poplar by its local environment and mapping the genetic differences between these environments produces biologically relevant genes. In this smaller scale study, we utilize only the portion of the species range for which both WGS and RNA-seq data exist. This allows us to validate genomic findings using the transcriptomic data.

Having justified our methods in the prelude, Part 1 expands the analysis to the complete species range which comprises 1,323 WGS. We characterize this species range using 20 different climactic variables and map differences in these environments to the poplar genome.

Finally, in Part 2 we explore how much information regarding the local environment can be recovered from poplar genomes. This study is motivated by the needs of our lab to quickly validate meta data and sampling locations for genotypes before committing to the resource intensive process of short read alignment and variant calling.

Chapter 2

Extreme Phenotype Sampling

Improves LASSO and Random Forest
Marker Selection for Complex Traits

2.1 Abstract

Most attempts to fit a supervised machine learning (ML) model in bioinformatics try to predict the full range of trait or response values. While such prediction tasks effectively capture the entire phenotypic range of the samples, they are cost prohibitive and can be statistically underpowered for detection of rare variants. In a study design known as extreme phenotype sampling (EPS), samples are selected from the two extremes of the phenotypic distribution. This approach is cost-cutting, by reducing genotyping/sequencing costs, as well as capable of increasing statistical power. Although combining EPS with ML algorithms has the potential to enhance association studies by improving their computational efficiency, EPS-ML approaches have seen limited use. In this paper we demonstrate an efficient and effective approach to leverage the EPS study design using LASSO regression and random forests, two commonly used ML algorithms within the broader bioinformatics community. We analyze two distinct data sets: leaf expression values generated from black cottonwood and malaria parasite transcriptome data collected from patients. We demonstrate that focusing only on the phenotypic extremes of these sample sets (by forming binary classes) can select more biologically meaningful features than using the full range. This approach will be useful to investigators when examining complex or novel traits. It is particularly well-suited to RNA-seq data where investigators often want to narrow attention to a small number of candidate transcripts out of a large initial pool. Our approach intentionally leverages existing software with efficient implementations to enable future applications of EPS-ML.

2.2 Introduction

Extreme phenotype sampling (EPS) is a technique originally developed within the context of linkage studies [12] as a means of increasing statistical power while keeping costs low through smaller sample sizes. EPS studies achieve these gains by sampling only individuals from the extremes of the phenotypic distribution. Statistically, the individuals at the extreme are highly informative. Studies [16] have shown that an EPS design can provide equivalent statistical power with smaller sample sizes when compared to traditional GWAS designs.

More recently, Xu et al. [36] have shown EPS can boost association testing power compared to studies that randomly sample the entire phenotypic distribution with comparable sample sizes. This increase in power allows for detection of rare variants not possible using traditional approaches.

Despite these advantages, the algorithms implementing EPS see limited use due to their inability to perform multi-locus tests. EPS has also been almost exclusively applied to linkage studies and SNP marker data. The lack of an established workflow/framework using commonly available tools has been important in limiting use of EPS with ML. Prior work such as Xu et al. [36] have used customized frameworks that are difficult to apply to new applications. Here, we outline an EPS approach for analyzing joint effects for two “off the shelf” machine learning (ML) algorithms: LASSO and random forests (RF). These algorithms were chosen because they have efficient implementations, are widely used, and are demonstrated to work well with a wide-range of data types. They are also suited to multi-locus testing as they simultaneously evaluate all features and select top performers. We demonstrate our approach on two distinct transcriptomic datasets. We illustrate the EPS design’s effectiveness for feature selection relative to the more traditional design of sampling individuals across the entire phenotypic range. We further validate our results by incorporating SNP data, where available, and comparing pathways associated with our markers to those uncovered in previous studies.

We independently hypothesized that an EPS-ML-like approach could enhance discovery of relevant loci in poplar based on our prior work with related bulk segregant analysis [18]. We tested this in the current study.

First, we address an important but outstanding genetic question for future U.S. bio-fuel production. How to best combine pathogen resistance from more northern, low-biomass, poplar trees with the high-biomass, but pathogen susceptibility, of more southern trees? Poplar is an ideal candidate for EPS-ML because its traits strongly correlate with latitudinal degree days and geographically distinct river systems in the Pacific Northwest [21]. Uncovering expressed genes linked to these important traits would enable a more intelligent and targeted breeding program.

Our EPS approach generated a more biologically focused (lower FDR) set of expressed loci for further analysis with a 45% reduction in data size compared to sampling the entire phenotypic range. Second, addressing a serious global parasite, we show that our approach gives comparable feature selection for clinically drug resistant vs. drug susceptible malaria when using 20% less data than full phenotype sampling. In addition to providing genetic features for further study, our results demonstrate the advantages of EPS-ML designs for both cost-effectiveness and increased statistical power. Our approach provides a roadmap to enable and enhance future studies.

2.3 Methods

2.3.1 Multi-omics data

We consider 580 black cottonwood (*Populus trichocarpa*) genotypes from a core region of their latitudinal distribution. Each genotype was drawn from one of four distinct river systems in the Pacific Northwest United States (see Figure 2.1). Of these, leaf RNAseq data were available for 213 individuals comprised of 16,031 differentially expressed genes. Additionally, variant (SNP) data for all 580 genotypes was available to validate expression results. Assembly, genome re-sequencing and RNAseq analysis of this *P. trichocarpa* GWAS population was previously described by Zhang et al. [37].

We also analyze transcriptome data from field-collected malaria infections [22]. These data consist of a total 1,154 plasmodium samples (drawn from distinct patients) with expression reported for 11,005 probes. Note that these expression data have been log normalized relative to a reference lab sample, so some expression values are negative, i.e. lower expression than the reference. Because there are a large number of missing values in these data from [22], we impute missing values as follows:

1. Extract all probes with the same prefix, e.g. ac11rRNA18s_0 and ac11rRNA18s_1
2. If all these probes have a missing value for the same sample, remove all of the corresponding probes from consideration

Source Location of Poplar Genotypes

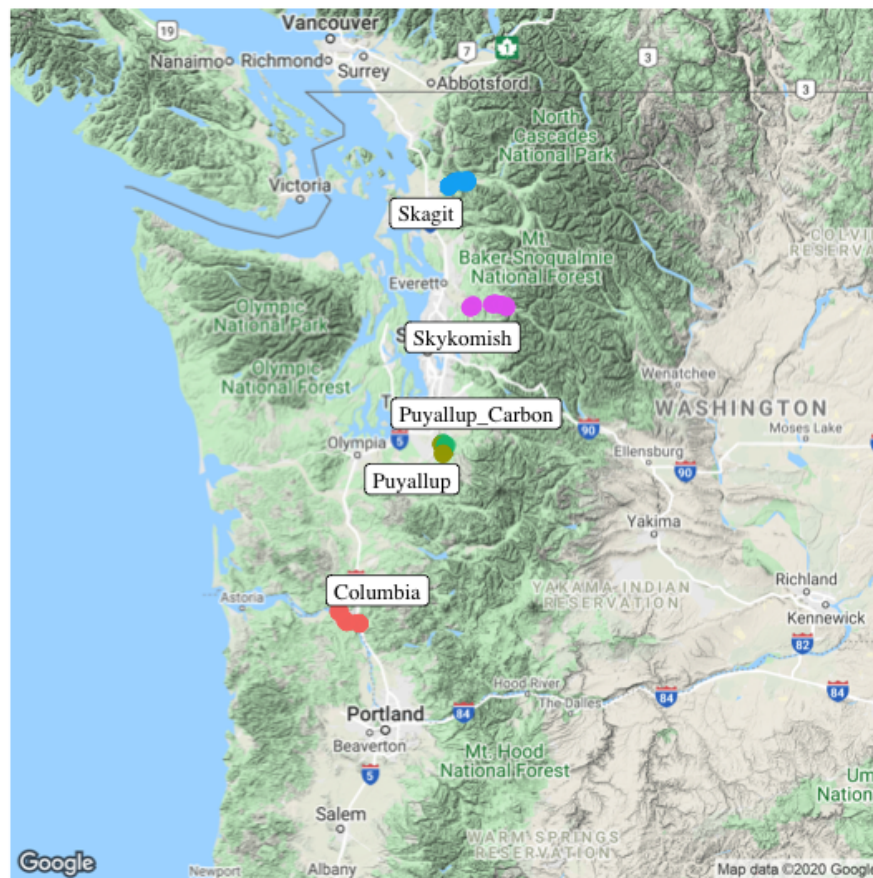


Figure 2.1: Sample locations of Poplar trees colored by river system.

3. If one or more of these probes has a value, replace the missing value(s) with the median of the non-missing probe values

This imputation strategy returns a new, filtered data set with a total of 8,455 probes. Similar to other studies [18, 22], we further filter out highly variable malaria genes that are unrelated to drug resistance with the annotation of “var”, “vsa” or “rif”, which further reduces the data down to a total of 7,894 probes.

2.3.2 Analysis of inherent structure in these data

Principal component analysis (PCA) is performed on both the poplar RNA-seq and variant data. PCAs on expression data are performed using the `prcomp` function in R, and the initial genetic PCA is generated using the `SNPRelate` package, also in R [39]. PCAs of the malaria data are performed using the same software.

2.3.3 Why Binary Prediction?

A common issue with EPS is that the samples at the phenotypic extreme are often vulnerable to measurement errors and phenotypic heterogeneity [16]. Li et al. propose a solution they call “almost-extreme” phenotype sampling where a window excluding the utmost extreme samples is used for sampling, but this demonstrates reduced power relative to true EPS [16]. We mitigate both of these issues by casting the prediction task as a binary class-based one, while accepting this casting from continuous to discrete results in a loss of information [13].

2.3.4 Classification in poplar

LASSO regression [31] is a commonly used approach in bioinformatics to fit a regression while extracting features that are most informative to this fit. We apply this regression to the binary classification task of predicting samples as belonging to one of two classes. This is done for two primary data collections:

1. Using all genotypes in the poplar expression dataset split into classes based on the mean latitudinal value, referenced as “mean-split”.

2. An EPS inspired design using genotypes in the poplar expression dataset drawn from the two extremes (i.e., most northern, most southern), referenced as “EPS”. Data was split on quartiles of the latitudinal distribution to determine northernmost and southernmost, meaning the bottom 25% of data and top 25% of data has been used. This is consistent with Xu et al. [36] where they demonstrate the top and bottom 20% of data are effective for EPS studies. We have expanded it slightly to 25% to account for our smaller data set.

We label these northern and southern classes as non-Columbia and Columbia, respectively, as this represents the river systems that are most proximal to these regions (Figure 1). In performing this EPS split, we have reduced the number of samples in the data by approximately 45% (see Table 2.1). We will consider the efficacy of the EPS results with this in mind, highlighting the EPS design’s increased statistical power at lower sample sizes.

For each data partition, a train and test set are constructed so that we have held out samples to validate our fit. The training set is constructed using 90% of the population from whichever class has less samples: Columbia or non-Columbia. This emphasizes having a balanced training set, although it comes at the cost of training set size. The regression is then fit using all the approximately 16,000 transcripts as predictors for the Columbia/non-Columbia trait. To validate our pipeline is not biased, we shuffle labels and repeat the regression fitting and prediction.

Because of randomness in the cross-validation during LASSO training, feature selection is unstable. To account for this, we bootstrapped the modelling process by repeating it 100 times. On each iteration we track which transcripts are selected. We then obtain a final set of selected transcripts by keeping all transcripts that are selected in greater than 50% of the modelling iterations. We also tried estimating LASSO parameters based on 100 independent runs and assigning feature weights based on the consensus model, but found the former approach is better for these data.

We further substantiate these results by also fitting a random forest (RF) model to the data for the same binary prediction task, in part to show generalizability and to show the EPS-ML approach can be applied independent of LASSO [36]. Train and test sets are constructed in an identical way to the LASSO fitting. The Ranger algorithm from the Keras

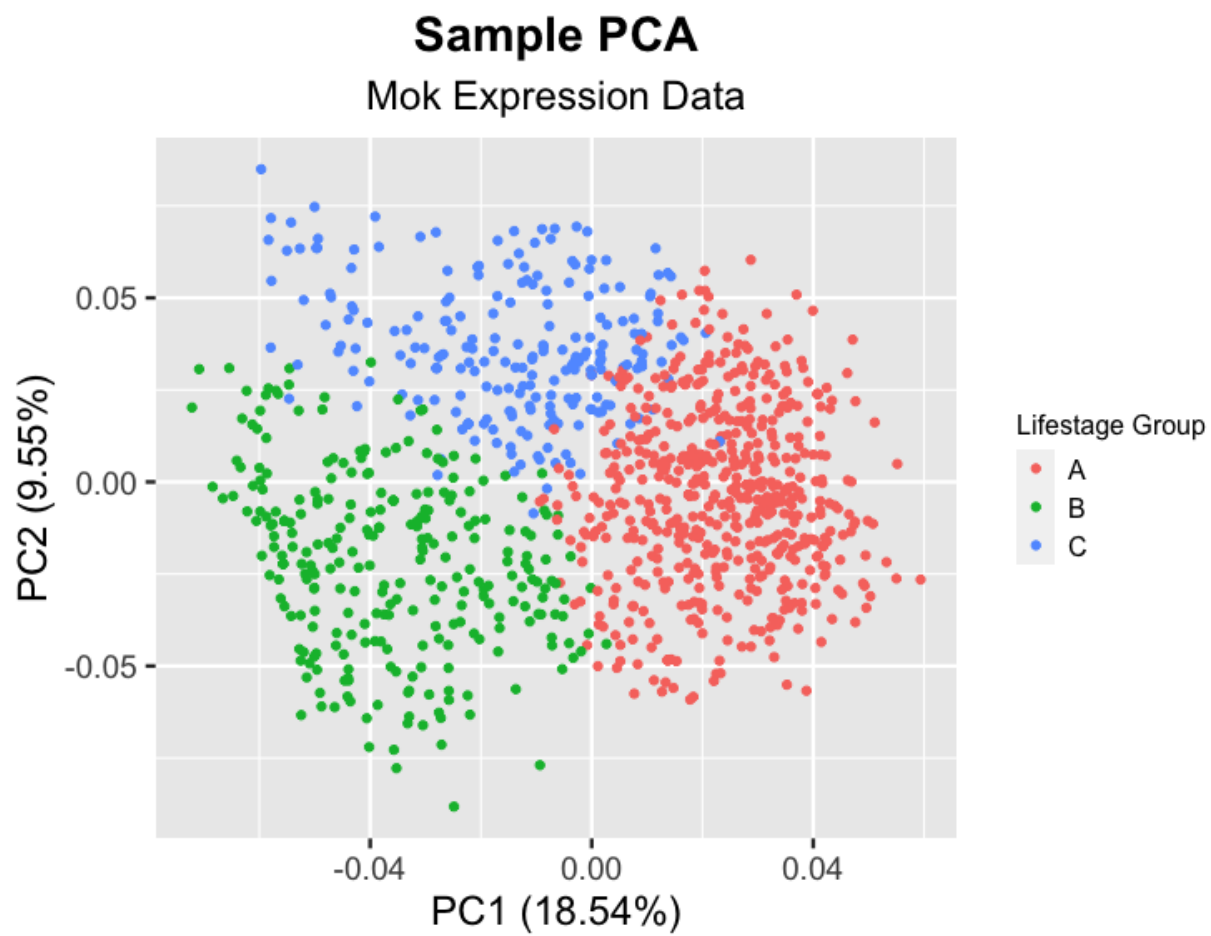


Figure 2.2: PCA of Malaria Expression data from Mok et al. [22].

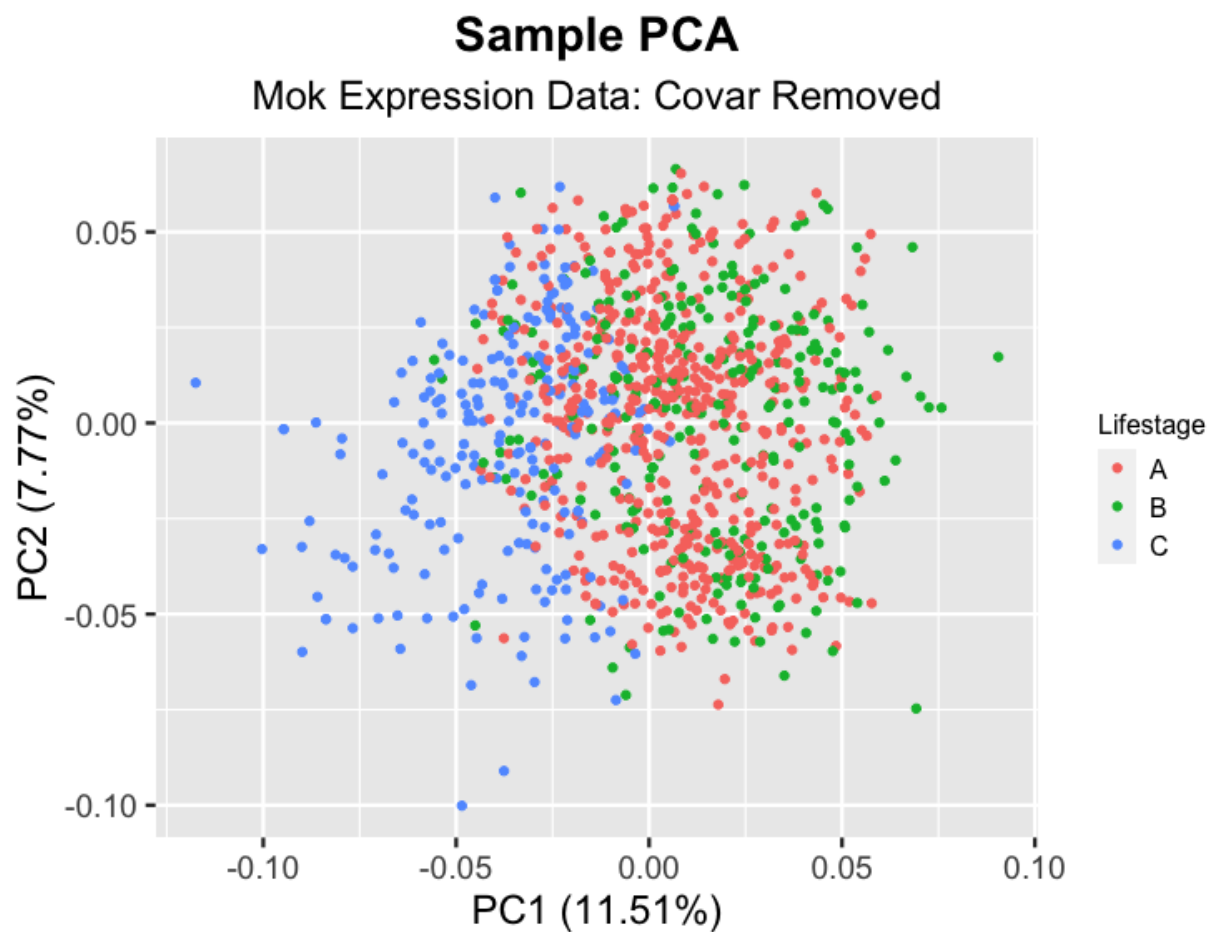


Figure 2.3: PCA of malaria probes post-regression correction. Substantially less structure driven by life-stage group can be seen in this projection, especially between groups A and B.

package in R is used to fit the model [6]. Variable importance is determined using an impurity measure with scaled values between 0 and 100. We select transcripts if they have a variable importance greater than 50.

Finally, to validate these findings, we leverage the genetic sequence data and perform a state-of-the-art genome wide association study (GWAS) using the GEMMA algorithm [41]. This allows us to identify single-nucleotide polymorphisms (SNPs) associated with each class independent of the expression data analyzed with LASSO. GEMMA is implemented using a kinship matrix to account for any inherent population structure.

2.3.5 Removing confounding factors in malaria expression data

As previously described in [22], there is a strong association between these malaria expression data and the life-stage of the parasites when they were sampled. Although there are only tens of known transcription factors in *Plasmodium falciparum*, it is clear that most genes follow a temporal up-down pattern that repeats every 36-48 hours [42]. We can use PCA to visualize this pattern as shown in Figure 2.2 with life-stage labels drawn directly from [22].

At least two approaches can be used to counter the confounding effects of parasite life-stage at sampling. First, a linear regression relative to published metadata can be used to minimize likely confounding effects. For each probe, we run a linear regression using the covariates as predictors. If the regression coefficient for a given covariate has a p-value of < 0.05 then we label that probe as significantly associated with that covariate. We then predict the expected value of the probe based on our fit to the covariate and generate a new dataset by computing the difference between this prediction and the measured value of the probe (see Figure 2.3). Alternatively, Zhu et al. [42] have released their own stage- and location-corrected version of this data set with 773 samples and 5,062 transcripts. This gives us three separate datasets to investigate parasitic resistance: the uncorrected Mok data, our regression corrected Mok data, and the stage- and location-corrected data from Zhu et al. Similar to the poplar analysis, we define two separate extreme classes that correspond to drug resistant and drug susceptible. This is done for each of the above datasets:

1. Using all samples split based on the clinical definition of drug resistance, which is a clearance rate of 5 hours or longer [33], referenced as “5-split”
2. An EPS inspired framework in which samples with intermediate phenotypes (clearance rate between 4 and 6 hours) are omitted prior to analysis, referenced as “EPS”

Again, the EPS-inspired approach reduces our sample size. For the Mok data set we have an approximate 19% reduction in sample size between the 5-split and EPS design. For the Zhu data this reduction is 21%. See Table 2.1 for a breakdown of sample sizes for each dataset.

2.4 Results

2.4.1 Feature Selection of poplar RNA-seq Data

PCA is used as a preliminary investigative step and, as expected, analysis of the RNA-seq data reveals no clear clustering of samples (data not shown). This is unsurprising due to the well known levels of noise present in expression data.

Having defined multiple binary class-like structures for these data (see Methods), we then use them to identify informative transcripts. To first show that our pipeline is not biased, we shuffle labels and perform regression fitting and prediction, and achieve an accuracy of 52.5% on the test set. This is to be expected for an arbitrary classifier on a binary classification task. We next fit the LASSO and RF models and Table 2.2 shows the results. Selected transcripts are those included in over 50% of models and accuracy is measured based on the held-out test set.

Performing a t-test between the two LASSO accuracy distributions concludes they are statistically different with a p-value of 2.97×10^{-05} . The same top-performer is selected by both LASSO models. Its expression is visualized in Figure 2.4. We see that differential expression between the classes at this specific locus is consistent with the perfect classification accuracy achieved using ML. Interestingly, this is the single gene selected by RF on the EPS task (thus the high accuracy) but not the single gene found during the RF mean-split task, supporting our hypothesis that EPS-ML can uncover more biologically relevant features.

Table 2.1: Sample sizes of each data collection. The Mok data and our corrected data have the same sample sizes.

Domain	Source	Partition	Sample Size
Poplar	-	Mean Split	205
Poplar	-	EPS	112
Malaria	Mok	5-split	1045
Malaria	Mok	EPS	851
Malaria	Zhu	5-split	769
Malaria	Zhu	EPS	607

Table 2.2: Results of Modelling on Poplar RNA-seq Data. Transcripts are those selected above our significance cut-offs.

Model	Partition	Transcripts	Accuracy
LASSO	Mean split	47	93%
LASSO	EPS	9	96%
RF	Mean split	1	87%
RF	EPS	1	100%

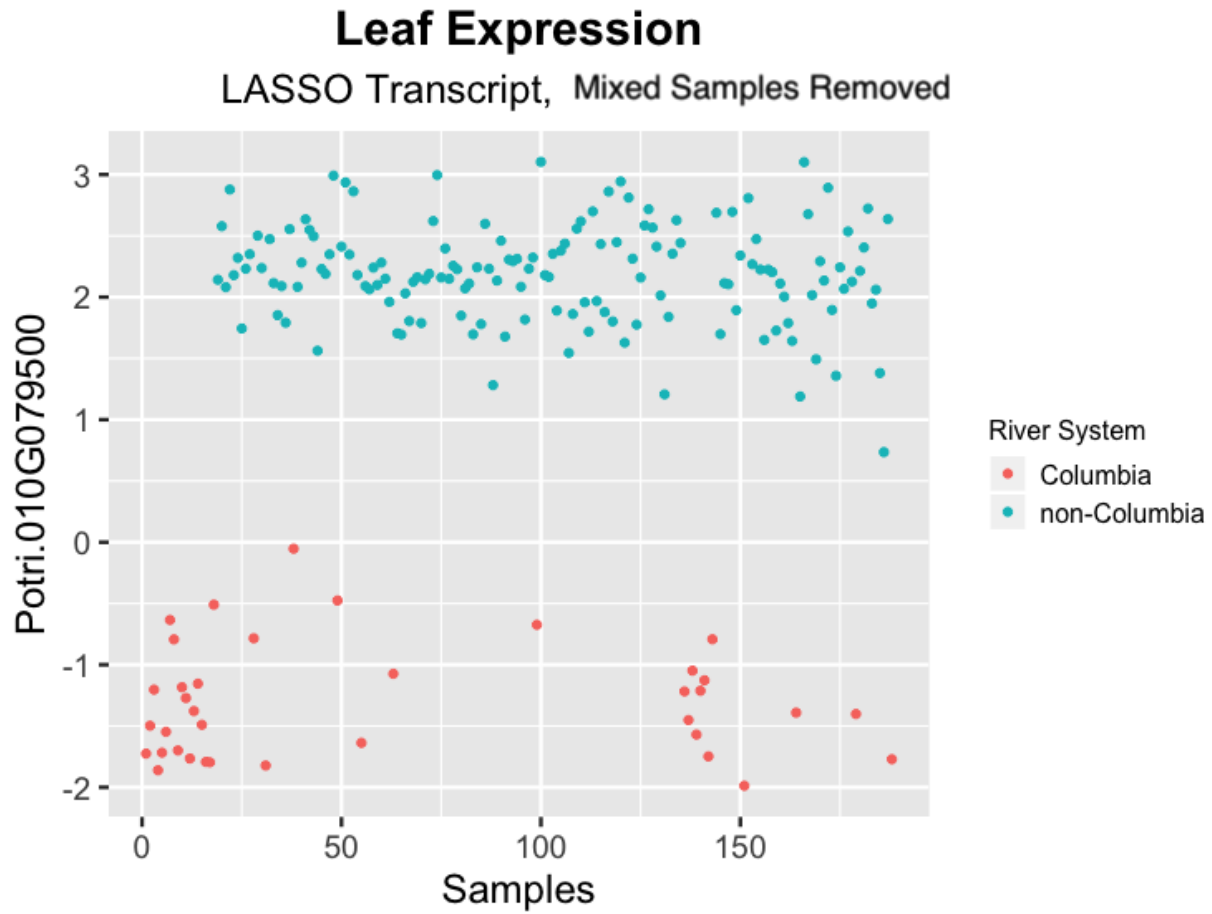


Figure 2.4: Transcript expression for Potri.010G079500 after removing intermediate phenotype samples, demonstrating a clear Columbia vs. non-Columbia difference in expression.

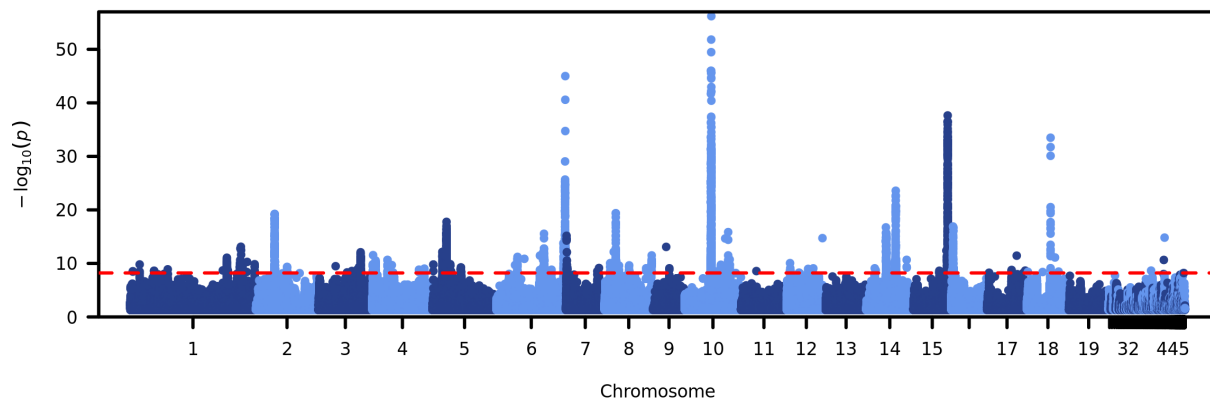


Figure 2.5: GWAS analysis using Columbia/non-Columbia extremes as the desired phenotype on the poplar data.

It is also worth noting that the improvements in feature selection and accuracy seen in Table 2.2 while using the EPS design occur with approximately half as many samples. This demonstrates the boost to statistical power that is possible with EPS and highlights how future studies could cut their costs without sacrificing ability to detect associations.

2.4.2 Genetic Validation of Expression Results

Using the variant data for these genotypes, we perform a traditional genome wide association (GWAS; see Methods) using the binary trait of Columbia (encoded as 1) and non-Columbia (encoded as 0). The results are shown in Figure 2.5. We observe strong signals on chromosome 6, 10, and 15. Importantly, the leaf-expressed transcript identified in our RNA-seq analysis above (see Figure 2.4) co-locates within this GWAS peak on Chr 10. This strengthens our hypothesis that the peak seen on chromosome 10 differentiates the Columbia/non-Columbia river system populations. This is also consistent with prior GWAS work in poplar that showed that this locus, encoding a 3-ketoacyl CoA synthase, is highly associated with multiple adaptive traits [11]. This further validates this transcript’s relevance to the target trait, and reinforces that our EPS-ML results more accurately reflect the underlying biology than sampling the entire phenotypic range.

2.5 Applying Techniques to Less Structured Data

We apply our approach to *in vivo* malaria drug-resistance data from 1,045 samples with expression for up to 7,910 probes. These data are taken from [22] and can be accessed through the Gene Expression Omnibus with Accession number GSE59097. As previously described in [22], there is a strong association between these expression data and the life-stage the parasite was in when it was sampled. To account for this, we use regression and a previously published corrected data set for comparison (see Methods).

Consistent with previously published literature in malaria, and this dataset in particular, both the annotated life-stage group and the East-West covariate (identified by [42]) associate with about 80% and 60% of differentially expressed (DE) probes, respectively. Using the fitted values from this regression (see Methods), we generate a new dataset with the same

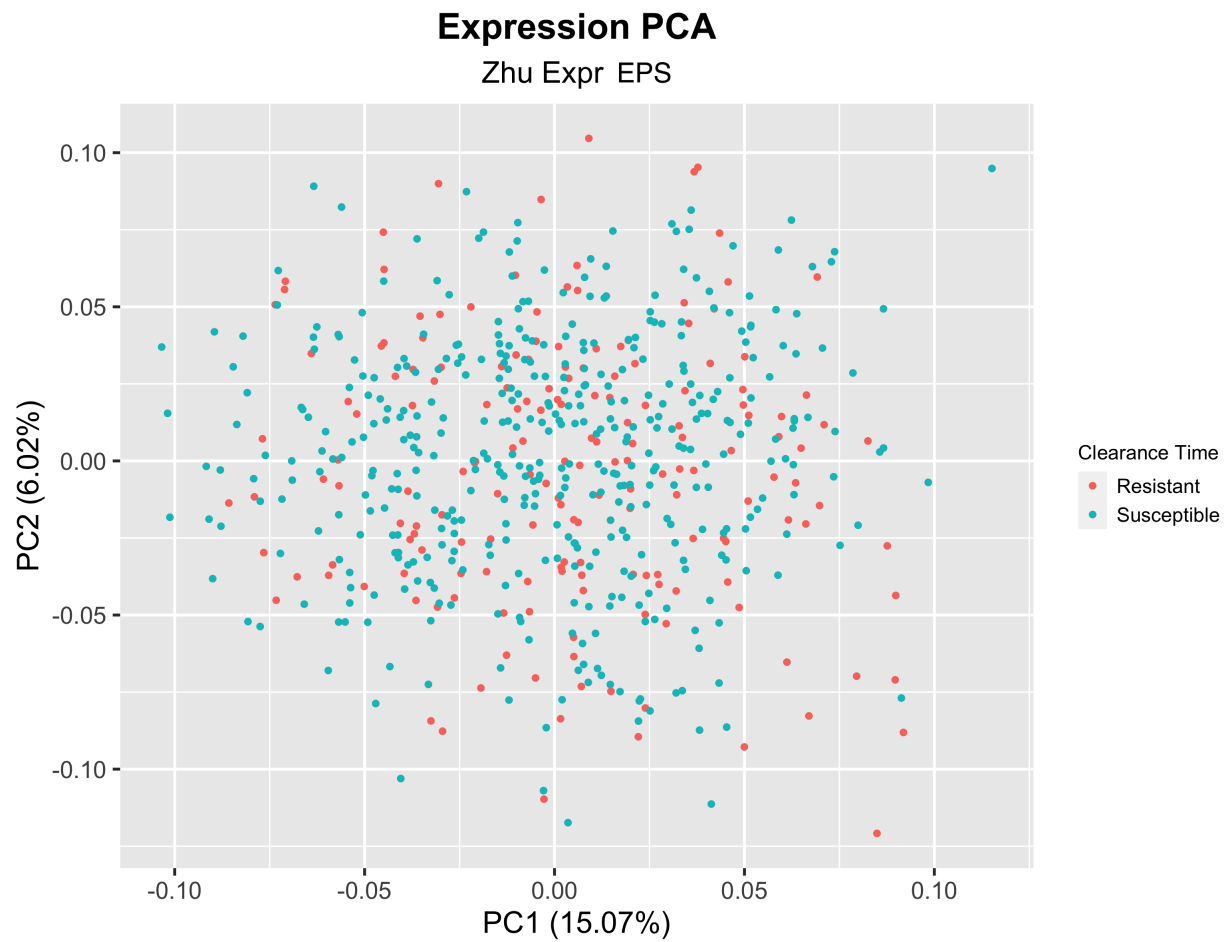


Figure 2.6: PCA of Zhu data [42] using all transcripts. As expected, there is no clear separation of classes.

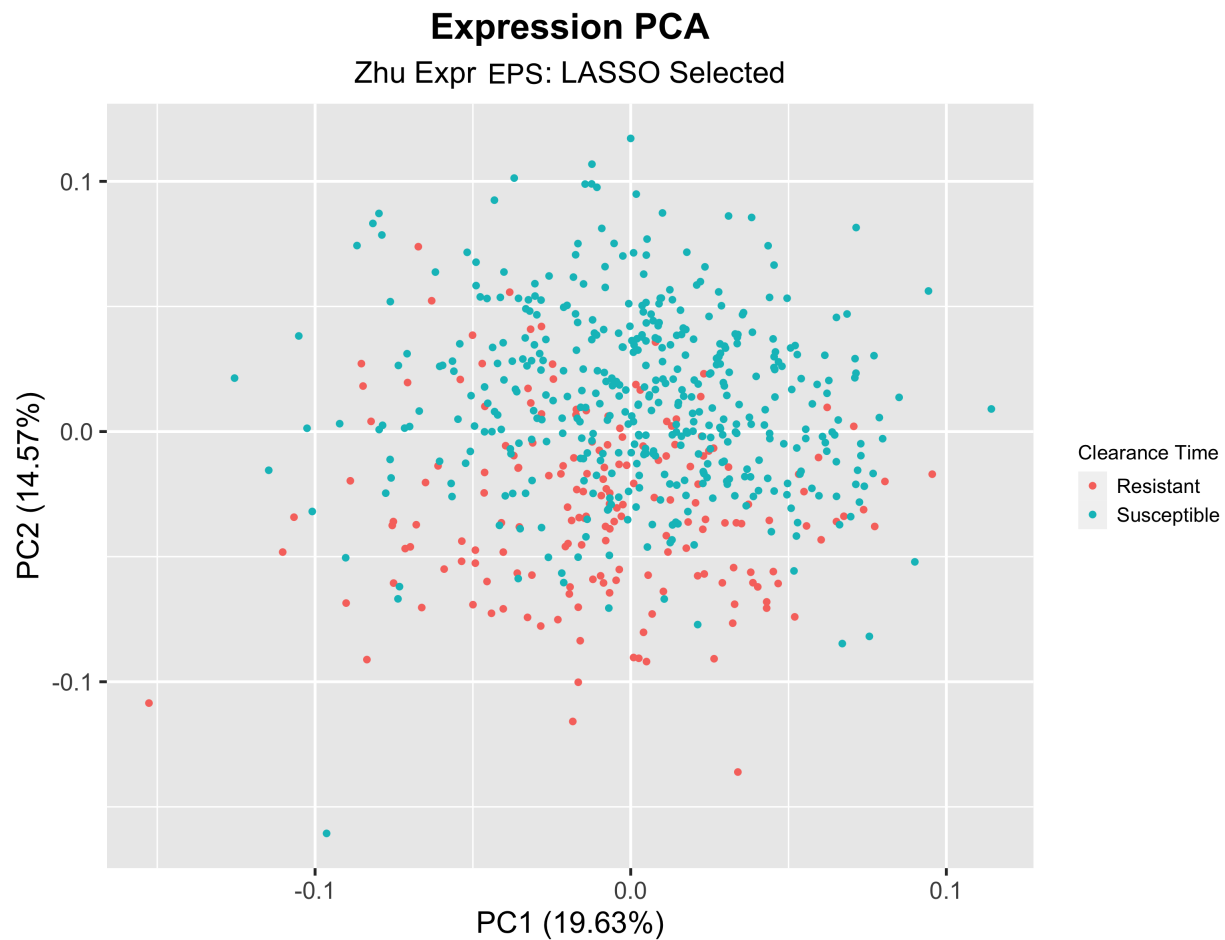


Figure 2.7: PCA of Zhu data [42] using LASSO selected transcripts with a slight separation of classes.

dimensions as the original but with altered expression values. This new dataset, which is visualized in Figure 2.3, has substantially reduced confounding effects, especially between stages A and B.

To replicate the two extremes (north/south) for these malaria data, we used the provided clearance time metadata to define a highly drug susceptible group (low clearance time) and a likely drug resistant group (high clearance time). We construct these classes in two different ways. The first approach “5-split” defines classes split on a 5 hour clearance time, which is the clinical definition of drug resistance [33]. The second approach leverages EPS in a manner similar to our work in poplar: we omit samples around the phenotypic mean. In practice, this is realized by labeling samples with clearance times less than 4 as susceptible and samples with clearance times greater than 6 as resistant, omitting all samples in-between. Results obtained from this task are referenced as “EPS”.

In creating these two splits we end up with different numbers of samples for each data collection. The Mok data and our corrected set share sample sizes and the Zhu data has slightly less (see Table 2.1). In this case the EPS sets represent approximately 20% reductions in sample size.

2.5.1 Feature selection

As with poplar, we first use LASSO regression to predict between the binary classes. LASSO selects the features of most importance to the task, which corresponds to the transcripts that are most associated with malarial drug resistance. We can use this approach to validate previously identified loci as well as to explore new loci that were previously undiscovered.

LASSO is implemented using the GLMNET software in R [10]. We utilize 90% of our samples to create a balanced training set and the remaining 10% as a test set. Accuracy metrics on this test set can be used to compare between model performances on the different datasets as well as to clarify how confident we should be in the selected features. Checking our pipeline for bias, we run it with class labels shuffled. Bootstrapping this process in the same way we do for the unshuffled LASSO runs, we obtain an average classification accuracy of 50%, exactly as expected for an arbitrary predictor on a binary classification task.

As was demonstrated in poplar, if transcripts are identified that show strong differential expression between the classes, we will see clear separation in the post-LASSO PCA. Figure 2.6 and 2.7 show our results on the data from Zhu [42] before and after LASSO, respectively. We can see clearer separation between the classes when we use only the LASSO selected transcripts but it is not perfect, which is indicative of the difficulty of this task and shows that the asynchronous parasites sampled from patients have much stronger confounding factors than the previously analyzed poplar data.

In the malaria data we use a threshold of inclusion in over 75% of bootstrapped LASSO models, or a scaled variable importance of greater than 75 for the RF models. This is appropriate due to the data being noisier. This more stringent cut-off works to counter this effect. Our results for both model types on all data sets are presented in Table 2.3, where the number of transcripts are selected in this way.

2.5.2 Biological validation

Based on Table 2.3 we note that in general LASSO selects more transcripts than RF and it also has slightly lower classification accuracy. Based on maximizing classification accuracy, we would be tempted to conclude that RF is the preferred algorithm. However, the features selected by LASSO are consistently more strongly linked to previously studied drug resistance loci. In fact, although only 1-3 transcripts pass our selection thresholding with the RF, these transcripts are mostly associated with life-stage effects, the major confounding factor in these data.

Comparing the RF performance on the Zhu data to the Mok and our corrected datasets, in general we see that models trained on the Zhu data achieve lower accuracy. This seems to validate our hypothesis and suggests that the high classification performance seen when using RF on the Mok and our corrected datasets is largely a result of the model fitting to existing life stage effects between the resistance groups, and does not reflect drug resistance.

We conclude that the LASSO framework is more sensitive to the resistance/susceptible signal in these data compared to the RF framework. We can further validate this conclusion by examining Table 2.4 which shows the top 10 ranked transcripts from the EPS LASSO run on the Zhu data. The top four genes from EPS LASSO are also found in the 5-split

LASSO with a slight re-ordering, but for EPS RF only one of these genes falls within the top five. Of these top four, three of them have been previously associated with malarial drug resistance. This gives us confidence that the high ranking is appropriate and that LASSO is detecting signal of interest. Significantly, one of these genes (AP-2 Mu) was only very recently uncovered experimentally [3].

Table 2.4 also demonstrates that EPS LASSO and 5-split LASSO identify similar top candidates. However, EPS LASSO also consistently further downranks false positives and achieves both of these accomplishments while using only 80% of the data that 5-split LASSO was trained on. These results highlight that EPS studies could be effective option in situations where there are small sample sizes.

Another group, Ford et al., have published their machine learning results for predicting malaria clearance time on these same data [9]. Their study provides a non-EPS/uncorrected data contrast to the results presented here, as they applied multiple ML approaches using standard validation techniques. From their study, only five of their top ten candidates are present in our top 50 from the EPS LASSO rankings on the Zhu data (see Table 2.5). The strongest known gene associated with reduced clearance is multidrug resistance 1 (PF3D7-052300), which is ranked 1st with our EPS LASSO model but is only ranked 6th by Ford et al. [9]. Ford et al. also identify four likely stage-specific genes in their top 10 that are probable false positives. These are ranked by our EPS LASSO model as the 143rd (2nd in Ford), 526th (7th in Ford), 281st (8th in Ford), and 392nd (9th in Ford) demonstrating that the LASSO regression has effectively down-ranked the life-stage signal when using well corrected data. We note that our RF approach also selects a stage specific gene as being most informative, and therefore is similarly confounded by life stage effects as Ford et al.

We performed GO term enrichment using PlasmoDB (www.plasmodb.org) on the set of selected transcripts as presented in Table 2.3. Given the small number of selected features, none of the available GO terms show statistically significant enrichment. However, the functional pathways associated with the transcripts from our EPS LASSO framework on the Zhu data show correspondence to experimentally determined proteins involved in drug response as reported by [3], for example, response to drug (GO:0042493) and response to chemical (GO:0042221). Comparing between the EPS

Table 2.3: Results of Modelling on Malaria RNA-seq Data. Transcripts are those selected above our significance cut-offs.

Model	Data	Partition	Transcripts	Accuracy
LASSO	Mok	5-split	15	77%
LASSO	Mok	EPS	10	81%
LASSO	Corrected	5-split	9	76%
LASSO	Corrected	EPS	8	81%
LASSO	Zhu	5-split	10	75%
LASSO	Zhu	EPS	7	81%
RF	Mok	5-Split	1	79%
RF	Mok	EPS	2	88%
RF	Corrected	5-Split	1	82%
RF	Corrected	EPS	2	88%
RF	Zhu	5-Split	3	76%
RF	Zhu	EPS	1	83%

Table 2.4: Top 10 genes found by EPS LASSO compared to 5-split LASSO and EPS RF. Bold entries have been previously linked to potential antimalaria drugs. Genes are listed in order of EPS LASSO rank.

Gene name	5-sp. Rank	RF Rank	Annotation
0523000	2	3	multidrug resistance 1
1349200	3	9	glutamate-tRNA ligase
1038700	4	8	<i>Plasmodium</i> exported protein
1218300	1	17	AP-2 complex subunit mu
1328400	7	2	conserved protein, unknown function
0404900	16	30	6-cysteine protein P41
1255200	23	62	PfEMP1 membrane protein
0202000	73	16	knob-associated histidine-rich protein
1242600	9	7	farnesyltransferase subunit alpha
1212900	39	24	bromodomain protein 2

and 5-split runs on the Zhu data, relevant GO terms have smaller p-values using EPS. We see a similar signal for EPS LASSO applied to our corrected data, but, as expected, there are much more stage effects with cell-cell adhesion (GO:GO:0098609) being the most significant enrichment found. We see no such enrichment using either RF or the 5-split approach on our corrected data.

We can conclude from this analysis that our LASSO framework applied to the Zhu data is selecting markers biologically associated with drug resistance/susceptibility. Importantly, we observe that performance between the LASSO EPS and LASSO 5-split framework is comparable (see Table 2.4). This occurs in spite of the 20% reduction in number of samples used for the EPS framework, highlighting EPS’s boosted statistical power to detect associations. Unlike poplar, where we saw improvements in feature selection with 45% reductions in data size, the malaria data confers comparable feature selection under EPS. This should encourage investigators as we see that in both examples cost-savings can be obtained by reducing genotyping costs without sacrificing association power. We hypothesize that the difference in feature selection performance between the poplar and malaria data is due to stronger confounding effects in the malaria data.

2.6 Discussion

In this study we have focused on the ability of EPS-ML approaches to increase associative power in multi-locus transcriptomic analysis. In doing so, we focus on the ability of the model frameworks to select meaningful biological features, as opposed to the more traditional ML focus of model performance. Identifying feature relevance is a more convoluted task than evaluating model performance, but one that is more impactful to biologists and we leveraged multiple strategies to achieve this goal.

In poplar, we were able to validate our transcriptomic findings by using SNP data. This allowed us to confirm that the top-performer from EPS LASSO and the only selected feature from EPS RF co-locate with the major peak from our GWAS study (see Figures 2.4 and 2.5). Previous work has also shown that this locus, encoding a 3-ketoacyl CoA synthase, is highly associated with multiple adaptive traits [11]. Significantly, EPS-ML required 45%

less samples than non-EPS-ML approaches and produced more actionable results. We will be pursuing this further for our work in U.S. bio-fuel production, highlighting the associative power of EPS to detect loci of fundamental significance to biological questions.

In malaria, we use functional annotations to determine if our top candidates are relevant to clinically observed drug resistance or false positives that reflect the impact of confounding effects within these data. Contrary to our poplar example where we had well-structured, low-noise data, the malaria data is beset by overwhelming life-stage effects. We have overcome these effects through our own means of correction as well as by studying a version of the data from Zhu et al. that was previously corrected using a more complicated, biologically-informed approach [42]. Our best results are obtained from this previously corrected Zhu data, illustrating the importance of leveraging a biological understanding of the data prior to model implementation.

With the Zhu data, we use EPS-ML to extract new and important underlying mechanisms. For example, PF3D7-1328400 is a conserved protein of unknown function that has been selected as a top-10 candidate in the EPS LASSO, 5-split LASSO and both RF models. Another candidate gene selected by our approach (AP-2 Mu) was only very recently uncovered experimentally [3]. Thus EPS-ML is powerful for elucidating interesting, biologically relevant candidates for further investigation.

Our studies show that EPS’s associative power seems to suffer in the face of strong confounding effects. In this context, we found the LASSO algorithm performed better than the RF, although the RF provided a more focused candidate set in poplar. Panarella et al. [24] also observe that in the face of strong confounding effects EPS suffers from an inflated false positive rate. This can be controlled, however, by including the top principal components in a logistic regression model. This leads us to suggest that investigators take steps to counter such effects before implementing EPS. Despite this, EPS still provided similar performance to the non-EPS framework while only utilizing 80% of the data.

Areas for future work include investigations into the trade-off between using a binary class structure for EPS or preserving the continuous phenotype. Multiple studies have highlighted that the binary class structure results in information loss [16, 36] relative to the continuous phenotype. In support of the binary structure, the previous work by Panarella

et al. demonstrated that such a structure can help counter confounding effects [24] and it certainly mitigates concerns about phenotypic heterogeneity at the extremes as raised by Li et al. [16]. A systematic investigation into these trade-offs would greatly benefit future work in EPS-ML.

As a final point we highlight our use of bootstrapping in the EPS-LASSO framework. This bootstrapping counters the instability of feature selection due to the random partitioning of data during cross-validation. We overcome this issue using a simple looping structure within a straightforward implementation. This approach avoids the need for specialized likelihood estimators that led previous studies to rely on niche software [36]. The downside of bootstrapping is the reliance on heuristic cut-offs of significance based on model inclusion. For this reason, the RF framework may be preferred over LASSO in the absence of confounding effects.

We have demonstrated that EPS can boost the statistical power of association studies when using less samples compared to designs that sample the entire phenotypic distribution. Our results in poplar demonstrate more focused and biologically relevant feature selection than the non-EPS design, even though the EPS study was performed with 45% less data. In malaria, we show comparable selection of features between the EPS and non-EPS designs despite the EPS design using 20% less data. These results corroborate what has been observed with EPS designs applied to linkage studies and GWAS, but extend the findings to transcriptomics and ML. Our study demonstrates a straightforward approach to EPS-ML that will facilitate future studies to seek the associative boost and cost reductive power of EPS.

Table 2.5: Comparison of rankings from Ford et al. [9] with our LASSO framework on the Zhu data. Genes are listed in order as ranked by [9].

Gene name	EPS Rank	5-split Rank	Annotation
1245300	46	13	Drug linked
1107700	143	72	Stage
1328400	5	7	Unknown
1372000	30	25	Likely stage
1115600	26	26	Drug linked
0523000	1	2	Resistance
1129100	526	299	Stage
0935400	281	58	Stage
1246800	392	NA	Stage
0805000	93	371	Unknown

Chapter 3

Environmental GWAS Reveals Hot Spots for Climate Adaptability in *Populus trichocarpa*

3.1 Abstract

Populus trichocarpa (poplar) is a woody species native to the Western US and Canada. It is an important biofuel crop for national security resource diversification and therefore is a major target of Department of Energy (DOE) breeding programs. Climate change is a top national security threat demanding the attention of researchers across agriculture and energy production. In this study, we showcase a new data resource of whole-genome sequences across the complete species range of poplar and use it to identify loci for climate change resiliency. We further demonstrate this resource as an important tool for scientists to probe the evolutionary dynamics of local adaptation across a diverse landscape of environments. Performing genome-wide association studies across 20 different climate variables, we identify nine novel genetic markers linked to climate variability. We suspect these loci to be responsible for local adaptation to large fluctuations in temperature and rainfall and thus important for climate change resiliency which aligns with climate change predictions of increasing extreme weather patterns and volatility. These findings illustrate that response to environmental stressors in poplar may be regulated by only a handful of key loci. This novel result has large implications for the ease of engineering a poplar stock well-suited to harsh environments. Molecular validation is underway to substantiate the effect of these loci and the role they play in developing a poplar variety more resilient to abiotic stressors.

3.2 Introduction

The *P. trichocarpa* species range extends from southern California all the way to British Columbia encompassing a latitudinal range of 38.88 to 54.25 degrees [30]. This range includes a diversity of macro and micro-environments that have likely shaped subpopulations of this species. Such local adaptation may aid in developing poplar biofuel feedstocks with greater biomass yield and enhanced resilience, especially with a predicted changing climate. An ideal research outcome is to understand how this environmental stratification drives evolution of genetic mechanisms that regulate bioenergy traits such that they could be modified in vivo to address future climate variability.

Relevant data, however, have not previously been available for the southern portion of the species range. With the recent sequencing of 207 new poplar genomes from the southern extent of poplar’s range [37], we are now positioned to probe the natural variation across poplar’s entire species range (see Figure 3.1). Unfortunately, little phenotypic data (e.g., drought or disease tolerance) are available for these recently sequenced southern individuals, preventing study of gene-by-phenotype effects directly.

In this study, we use 19 bioclimatic variables from the WorldClim database [7] to perform genome-wide association studies (GWAS). For each climate variable, we sample values only from the top and bottom 25% of its distribution leading to disparate climate groups (i.e. high and low precipitation), which amplifies association with meaningful loci as previously reported [15]. In addition to the WorldClim variables, we partition the species range using the Koppen-Geiger climate classification system [28]. Across the range of possible classifications, poplar falls within two main classes: a temperate oceanic climate (Cfb) in the north, and a hot-summer Mediterranean climate (Csa) in the south. In general, the northern environmental group experiences more stable temperature and rainfall throughout the year, with higher levels of precipitation. While the southern population is particularly amenable to climate change, demonstrating adaptation to large fluctuations in temperature and rainfall across the year. We hypothesize that genetic responses to these contrasting environments can be uncovered using genome-wide association analysis.

Our results reveal that a small number of genomic loci on multiple chromosomes could be responsible for adaptation to this wide range of climatic pressures. This suggests that genetic modification may be necessary at only a handful of loci to engineer a climate resilient poplar. Further, this new data collection of poplar genomes presents a powerful resource for investigators interested in studying gene-by-environment effects and their role in local adaptation. We show that grouping trees by their environmental context can be a powerful study design for GWAS in the absence of complete phenotypic data across the whole range. We leave it up to researchers to group and probe the data in ways that are most relevant to their own questions.

3.3 Methods

3.3.1 Data Preparation

Whole genome resequencing was previously described by Zhang et al. [37] for 1,023 individuals of this poplar GWAS population. In this study we use the same protocols to incorporate an additional 207 new accessions from the southern portion of the species range. BWA-mem [17] is used for alignment to the *Populus trichocarpa* reference genome v3.0 and gatk HaplotypeCaller [35] is used for variant calling. RNA-seq data collection and processing is previously described in Zhang et al [37].

3.3.2 Genome-wide Association Study

Our GWAS analyses are performed using a GEMMA [40] implementation that utilizes a kinship matrix to correct for population structure. Climate variables are studied using an extreme phenotype sampling design. Each variable has the middle portion of its range omitted from the study, using only genotypes sampled from the upper and lower quartiles. Results are interpreted using the stringent Bonferroni correction to identify all major and minor peaks. Major peaks are interpreted as the single largest peak against the background signal.

Common concerns with association studies involving environment are that the neutral genetic structure of genotypes will lead to false positive associations. It is therefore important to correct for this in association tests. Previous work has shown that computing spatial auto-correlation and correcting for it along with genetic relatedness is overly conservative [27]. No previous work has been done to demonstrate that relatedness alone appropriately accounts for neutral genetic structure. Here, we argue that using the kinship matrix alone (to account for relatedness) seems to give appropriate results. We map 20 different variables using similar spatial distributions of genotypes for each, yet Figure 3.2 and 3.3 demonstrate broadly different patterns of association on the genome-wide level. We would expect similar patterns across all variables if we were mapping primarily neutral genetic structure.

3.3.3 Allele Frequency Analysis

A pipeline to analyze allele frequency shifts across the entire latitudinal range (not restricted to the extremes) was constructed. This allows finer granularity of target identification within GWAS peaks. It also demonstrates patterns of selective sweeps exerted by the environmental pressures across the landscape. This analysis also allows us to roughly assess the size of LD blocks in these regions, providing insight into the extent of recombination here.

The pipeline was built using VCFtools and bash. The algorithm is:

1. Genotype data is filtered to chromosomal regions of interest based on GWAS peaks
2. A list of samples sorted by latitude is created
3. Starting with the most southern sample, a window of 100 samples is selected
4. For these 100 samples, the allele frequency of each locus in the chromosomal regions is computed
5. These frequencies are labeled with the average latitude of the 100 samples selected
6. The procedure is repeated starting with the second most southernly sample and iterated until the northernmost sample has been reached

This framework produced an allele frequency (AF) matrix, consisting of a column for each SNP in the chromosomal regions, and a row for each set of 100 genotypes that were labeled by their average latitude.

For each chromosomal region, we can select variants by the amount of variation in allele frequency they exhibit across the latitudinal range. The idea here is that variants in strong LD to each other will show similar levels of variation. With this in mind, our goal is to filter out all variants except for those in the LD block that exhibit the highest levels of variation. We achieve this by plotting the number of selected SNPs against the variance threshold cut-off (see Supplementary Figure 19) and use a threshold that chooses only the last set of SNPs.

3.3.4 RNA-seq Functional Enrichment

Interested in the biological pathways involved in response to these climate variables we perform GO term functional enrichment using topGO [26]. This is done for each climate variable individually. For a given climate variable, we split samples into a high and low group based on the quartiles of the distribution. We then identify differentially expressed (DE) transcripts across these groups in the leaf and xylem RNA-seq data. Differentially expressed transcripts are subsequently annotated for function.

3.4 Results

3.4.1 Climate Characterization of Species Range

We consider 1,273 genotypes of *Populus trichocarpa* from a representative sampling of the latitudinal distribution of its species range (see Figure 3.1). Assembly and genome re-sequencing of this population was previously described by Zhang et al. [37]. Climate data is downloaded from WorldClim [7] and consists of 19 different bioclimatic variables.

Beyond explicit climate variables, we also classify the species range using the Koppen-Geiger climate classification system. We see our tree’s span from a temperate oceanic climate (Cfb) in the north, to a hot-summer Mediterranean climate (Csa) in the south [28]. Climate modelling from Rubel et al. (2010) [28] predicts that over the next 75 years this Csa climate will spread substantially across the western US. Identifying genotypes and their loci that are well adapted to the Csa climate provides an important resource for marker-assisted breeding programs, especially in the face of climate change. We probe this question using a GWAS across all the Koppen-Geiger classes that appear in our dataset.

Using these 20 variables, we perform genome wide association studies (GWAS) to probe the genetics of adaptation. Results for all association tests are shown in Figures 3.2 and 3.3, where they are listed by major chromosome of association.

Looking closely at seasonality of precipitation and diurnal range in Figure 3.4, we note strong genetic responses to climatic pressures on chromosome 13 (Chr13) and chromosome 1 (Chr01) respectively. Looking explicitly at shifts in allele frequency (see Methods; Figure 3.4

panel C) for variants within these peaks reveals patterns characteristic of adaptive response to environmental clines [19] and alleles that have been subject to a strong selective sweep [27]. Similar visualizations are available for all 20 variables in the Appendix, showing similarly strong patterns of selective sweeps.

Across the 20 climate variables studied we identify only 9 major associated genomic loci, including the major peaks on Chr01 and Chr13 shown in Figure 3.4. An interesting feature of this result is that many climate variables share the same major peak. We refer to separate climate variables sharing a major peak as colocalization. The extent of this across the genome is visible in Figures 3.2 and 3.3, one example being Diurnal Range and Max Temp sharing a major peak on Chr01.

3.4.2 Allelic Patterns of Adaptation in Hot Spots

Beyond GWAS peaks, we investigate allele level variation across the entire population at the genomic regions associated with climate variables. Examples of these results are shown in panel C of Figure 3.4 (see Methods). The revealed patterns seem indicative of response to a strong selective sweep [27]. We observe groups of SNPs behaving in unison (LD blocks) with smooth response curves to a latitudinal cline that is known to correlate with environmental variables [20]. We can make inferences about the size of these LD blocks by looking at the number of SNPs plotted. Figure 3.4 panel C shows that Chr13 has a very small LD block associating with seasonality of precipitation. This implies recombination is high in this region, having whittled away hitchhiking variants from the causal variant within this block. This is validated by an examination of LD across this window where we observe a region of decreasing LD as you approach this locus. The Chr01 association with diurnal range on the other hand seems to show two contrasting LD blocks of approximately equivalent size.

3.5 Discussion

We observe wide-spread colocalization of GWAS peaks across 20 different climate variables. The ramifications of this are that flexible and robust modulation of climate adaptability may be possible using only a handful of loci. The Chr16 peak shows up as the major driver

of multiple warm-weather precipitation related variables. The Chr13 peak that associates with seasonality of precipitation maps to a putative chloroplast integration event in the nuclear genome (Chr13:14,730,412-14,896,306) [38] and thus we may be mapping variation in the chloroplast. This chloroplast event is further supported by functional enrichment for seasonality of precipitation which shows phosphatidylinositol phosphorylation as the top GO term. Recent studies [29], have linked this process to regulation of chloroplast processes. Peaks on Chr10 and Chr01 consistently co-occur, alternating between which is dominant. This suggests these loci are co-inherited.

Across all climate variables we see consistent enrichment (see Methods) for metabolic and microtubule-based processes. This is also validated by some of the minor GWAS peaks isolated, with a notable candidate on Chr07 mapping to a hexokinase, further validating the role of energy production and metabolic processes in response to climate stress. We currently have experiments underway to validate whether known energy pathway interventions can boost stress response in poplar.

Looking at the directionality of trait regulation where relevant data exists (see Methods), we see patterns previously identified in poplar with tree height and bud set showing an inverse relationship.

The strong patterns of differentiation seen between the Koppen-Geiger climate groups indicates that these differing environments are responsible for large shifts in adaptation. Considering the implications of climate change across this region of the country, being able to map the drivers of adaptive advantage will likely determine whether poplar’s arable land space increases or decreases over the coming century. As stated, predictions by climatologists suggest that the southern climate context known as hot-summer Mediterranean (Csa) will expand massively over a 100-year period. Proactively breeding poplar stocks well-suited to this environment will allow it to remain viable as a source of liquid biofuels. This study presents two novel loci on chromosome 10 and chromosome 18 (see Supplementary Figure 2) as top targets for adaptation to this context.

The recent sequencing of 207 new genotypes from the southern portion of poplar’s range opens the opportunity to explore species-wide adaptive dynamics. In this study, we leveraged this opportunity by constructing macro-environment groups across 20 different

climate variables. This approach revealed specific targets for experimental follow-up as well as broader patterns of climate response. We believe this is only an initial step towards what is possible with this data, and we hope other investigators will view this study as a showcase for this new data resource.

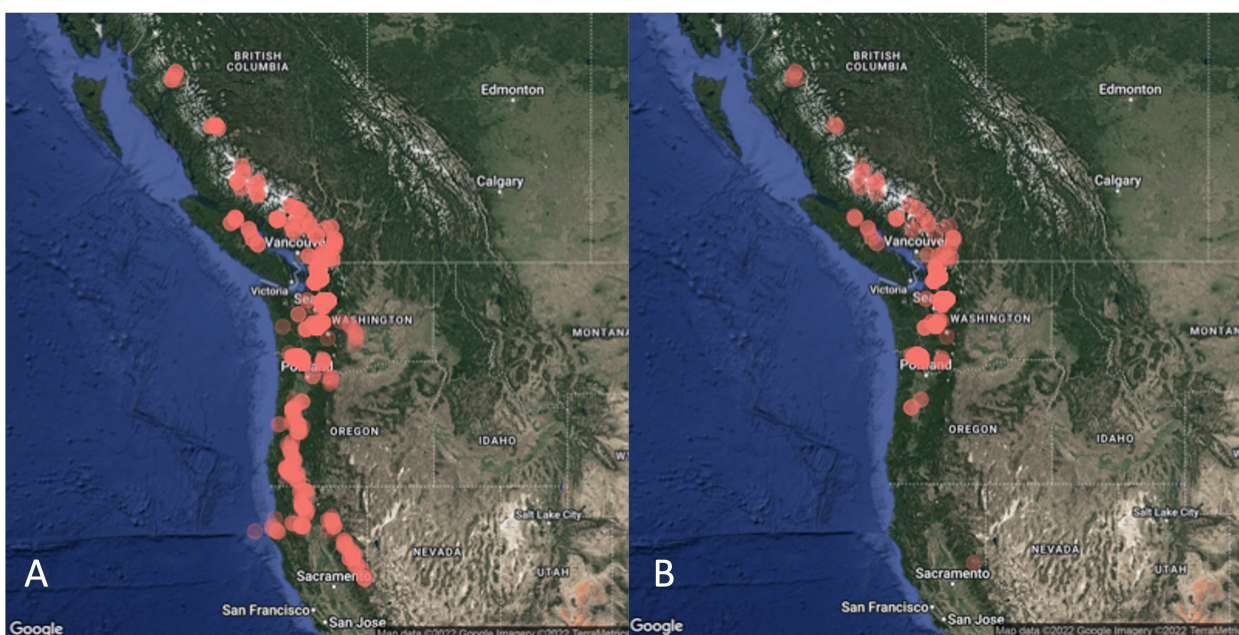


Figure 3.1: A) Sampling locations of previously studied WGS from Slavov et al. (2012) [30] combined with our 207 newly sequenced genotypes. B) Sampling locations of leaf RNA-seq data used for functional enrichment

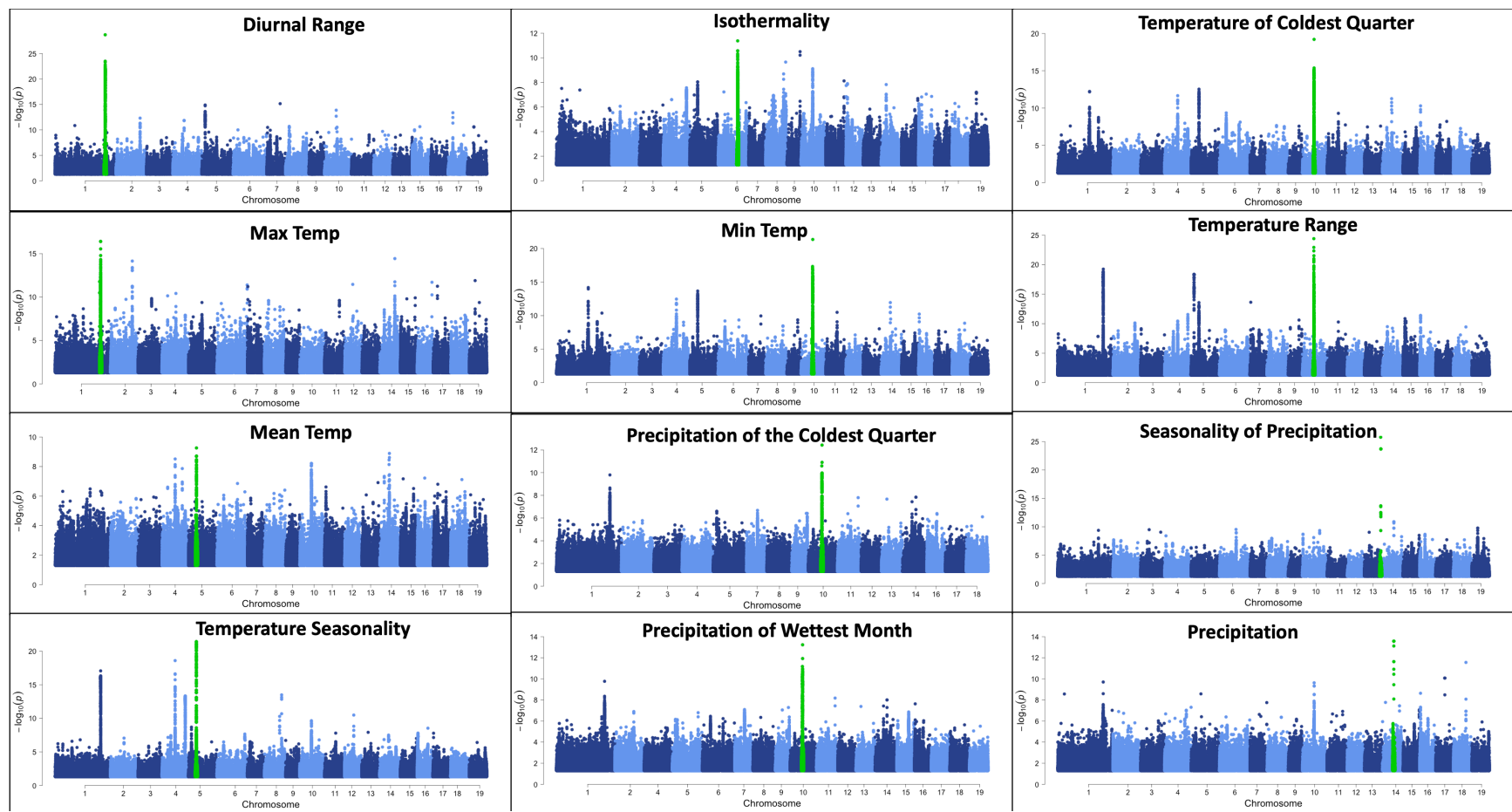


Figure 3.2: Collection of GWAS results for climate traits with major peaks from chromosome 1 to 14. Results are ordered by major peak demonstrating the extent of colocalization across the climate traits.

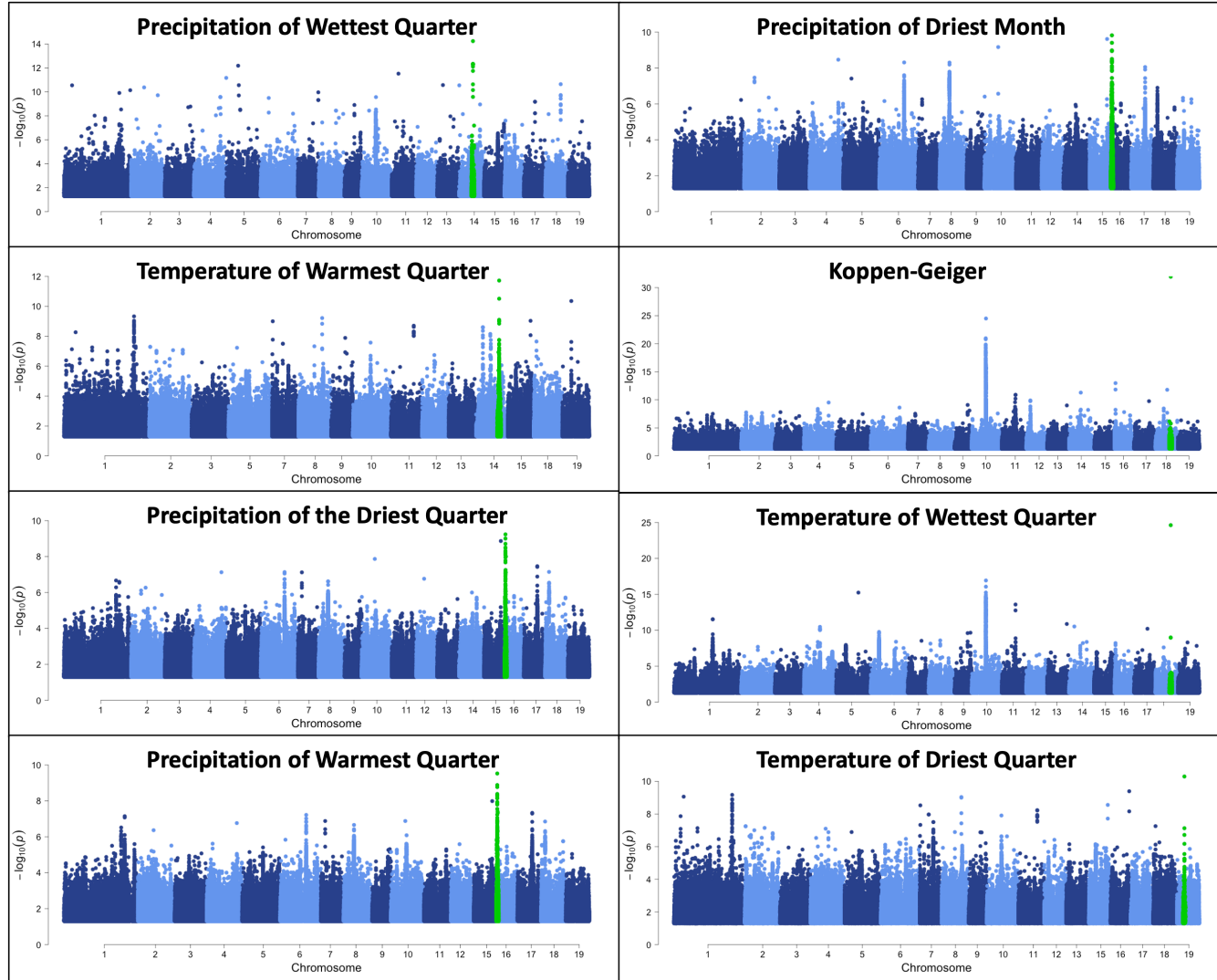


Figure 3.3: Collection of GWAS results for climate traits with major peaks from chromosome 14 to 19. Results are ordered by the major peak demonstrating the extent of colocalization across the climate traits.

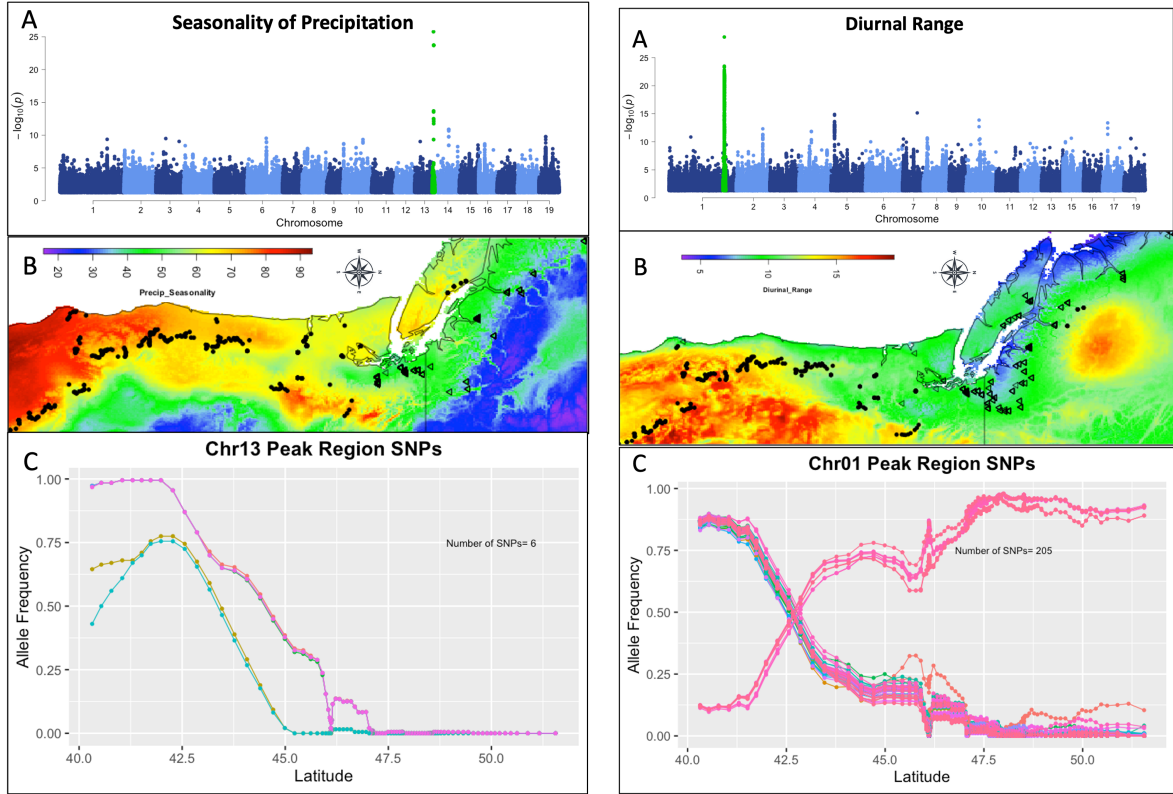


Figure 3.4: Climate based GWAS implicates strong regulators on Chr13 for seasonality of precipitation and Chr01 for diurnal range. A) Manhattan plot from GWAS based on genotypes shown in panel B, major peak colored in green. B) Climate map of response variable used in GWAS, oriented with North to the right to align with the latitude scale in panel C. Lower quartile genotypes are marked with open triangles and upper quartile genotypes are marked with filled circles. C) Allele frequency shifts computed using all genotypes across the entire species range for variants found in the peak region marked in green in panel A, each line corresponds to the variation of one SNP.

Table 3.1: Strongest associations for each climate variable. We see strong colocalization across these major peaks with many climate variables mapping to exactly the same variant.

	Climate	Chr	Pos	-Log10(P)	Gene Region
1	Diurnal Range	Chr01	41471877	28.69	Potri.001G395200__300
2	Max Temp	Chr01	41465261	16.41	Potri.001G395200
3	Mean Temp	Chr05	7166930	9.26	Potri.005G095400__500
4	Temp Seasonality	Chr05	7170385	21.45	Potri.005G095500
5	Isothermality	Chr06	17835645	11.38	Potri.006G170700__800
6	Min Temp	Chr10	10639623	21.35	Potri.010G080300__400
7	Precip Coldest Quarter	Chr10	10639623	12.43	Potri.010G080300__400
8	Precip Wettest Month	Chr10	10639623	13.23	Potri.010G080300__400
9	Temp Coldest Quarter	Chr10	10639623	19.21	Potri.010G080300__400
10	Temp Range	Chr10	10457195	24.40	Potri.010G079400__500
11	Koppen Geiger	Chr13	14670772	72.66	Potri.013G137000__100
12	Precip Seasonality	Chr13	14670772	25.77	Potri.013G137000__100
13	Precip	Chr14	10439491	13.57	Potri.014G136400__500
14	Precip Wettest Quart	Chr14	10439559	14.25	Potri.014G136400__500
15	Temp Warmest Quarter	Chr14	13653701	11.72	Potri.014G170300__400
16	Precip Driest Month	Chr16	618375	9.82	Potri.016G011500__600
17	Precip Driest Quart	Chr16	618720	9.24	Potri.016G011500__600
18	Precip Warmest Quart	Chr16	618720	9.52	Potri.016G011500__600
19	Temp Wettest Quarter	Chr18	11505780	24.63	Potri.018G088000__100
20	Temp Driest Quart	Chr19	5323667	10.29	Potri.019G045000__100

Table 3.2: Annotating genes neighboring the lead variant for each climate variable.

	Gene	Arabidopsis Ortholog	Annotation
1	Potri.001G395200	AT2G20562	NA
2	Potri.001G395300	AT3G29635	Isoflavone-7-O-beta-glucoside
3	Potri.005G095400	NA	NA
4	Potri.005G095500	AT5G23960	Germacradienol synthase
5	Potri.006G170700	AT5G57610	BEM1P Kinase
6	Potri.006G170800	AT5G57620	MYB-LIKE Protein
7	Potri.010G080300	NA	NA
8	Potri.010G080400	AT2G26640	Very-long-chain 3-ketoacyl-CoA
9	Potri.010G079400	AT2G26640	Very-long-chain 3-ketoacyl-CoA
10	Potri.010G079500	AT2G26640	Very-long-chain 3-ketoacyl-CoA
11	Potri.013G137000	ATCG00700	PsbN protein (psbN)
12	Potri.013G137100	ATCG00680	chlorophyll apoprotein
13	Potri.014G136400	AT2G48010	LRR Receptor
14	Potri.014G136500	AT2G48020	Sugar Transporter ERD6-Like
15	Potri.014G170300	AT2G04270	glycoside hydrolase
16	Potri.014G170400	AT2G04240	RING zinc finger protein
17	Potri.016G011500	NA	H(+)-transporting two-sector
18	Potri.016G011600	NA	NA
19	Potri.018G088000	AT4G29120	2-hydroxy-3-oxopropionate
20	Potri.018G088100	AT5G25930	Leucine-rich repeat
21	Potri.019G045000	AT5G25930	Polyphosphate kinase
22	Potri.019G045100	AT5G23960	Casbene synthase

Chapter 4

Correcting Low-Confidence Latitude Labels Using Machine Learning in *Populus trichocarpa*

4.1 Abstract

This study leverages machine learning algorithms to correct low-confidence latitude labels for whole-genome sequences. Our protocol reduces the dimensionality of these sequences through a principal component analysis to reduce their sparsity prior to input to the predictive algorithms. We demonstrate the effectiveness of this approach by showing our results are coherent with current labels where high-confidence labels exist. We validate the predictions on low-confidence labels by looking at neighboring clusters in the principal component space. On top of this, we make strides to achieve predictive capacity in this domain using unaligned sequence fragments. Despite not achieving predictive levels to correct labels from these unaligned fragments directly, we develop a novel utilization of minHashing whole genomes that shows promise for future development.

4.2 Introduction

Ensuring meta-data labels correctly match genotyped samples is a problem inherent to data collection and distribution. In this study, we test the efficacy of a tool for predicting local environment traits from aligned and unaligned sequence data. The motivation for such a tool is to provide researchers with a quick diagnostic to assess which meta-data collection should be assigned to a genotype, when its suspected the genotype-meta data pair has been incorrectly assigned. We tackle this problem in the context of *P. trichocarpa* (poplar) as it is the specific use case motivating the tool, however, similar applications are possible across the range of sequenced plant species.

Poplar’s species range extends from southern California all the way to British Columbia encompassing a latitudinal range of 38.88 to 54.25 degrees [30]. This range includes a diversity of macro and micro-environments that have likely shaped subpopulations of this species. These subpopulations can be characterized by both population structure, as previously identified in Slavov et al. (2012) [30], as well as climactic variables. Our goal is to predict the latitudinal coordinates of these genotypes from their sequence data. We extend

this task by also attempting prediction on climactic variables from unaligned reads of these genotypes.

We make predictions from both aligned, variant-called data as well as unaligned sequence fragments. As anticipated, we find aligned sequences provide stronger signals of genetic architecture, leading to higher predictive capacity. Only low-levels of success are attained using the unaligned reads, but this null result is motivated by a desire to pioneer a tool that could be used to test the integrity of meta-labels prior to variant-calling and alignment. It is also of interest to researchers to understand what level of post-alignment signal is recoverable from pre-alignment fragments.

4.3 Methods

4.3.1 Data

We consider 1,273 poplar genotypes from a representative sampling of the latitudinal distribution of its species range. Assembly and genome re-sequencing of this population was previously described by Zhang et al. [37]. BWA-mem [17] is used for alignment to the *Populus trichocarpa* reference genome v3.0 and gatk HaplotypeCaller [35] is used for variant calling. Climate data is downloaded from WorldClim [7] and consists of 19 different bioclimactic variables.

For the 1,273 whole-genome sequence (WGS) SNP-sets we use a principal component analysis (PCA) to examine the genetic structure of this population (see Figure 4.1). Similar to previously published work examining population structure [30], we find strong coherence with latitude. We use the structure seen in the PCA to identify likely mis-labeled genotypes. Based on Figure 4.1, we assert that genotypes drawn from similar latitudinal positions should cluster together, therefore if a genotype has an incorrect latitudinal label we should see it clustering near genotypes with different labels. We can use river system labels to test this, as they are finer grained than our broad latitudinal classes shown in Figure 4.1, yet they provide a discrete label to use for comparing neighbors in the PCA.

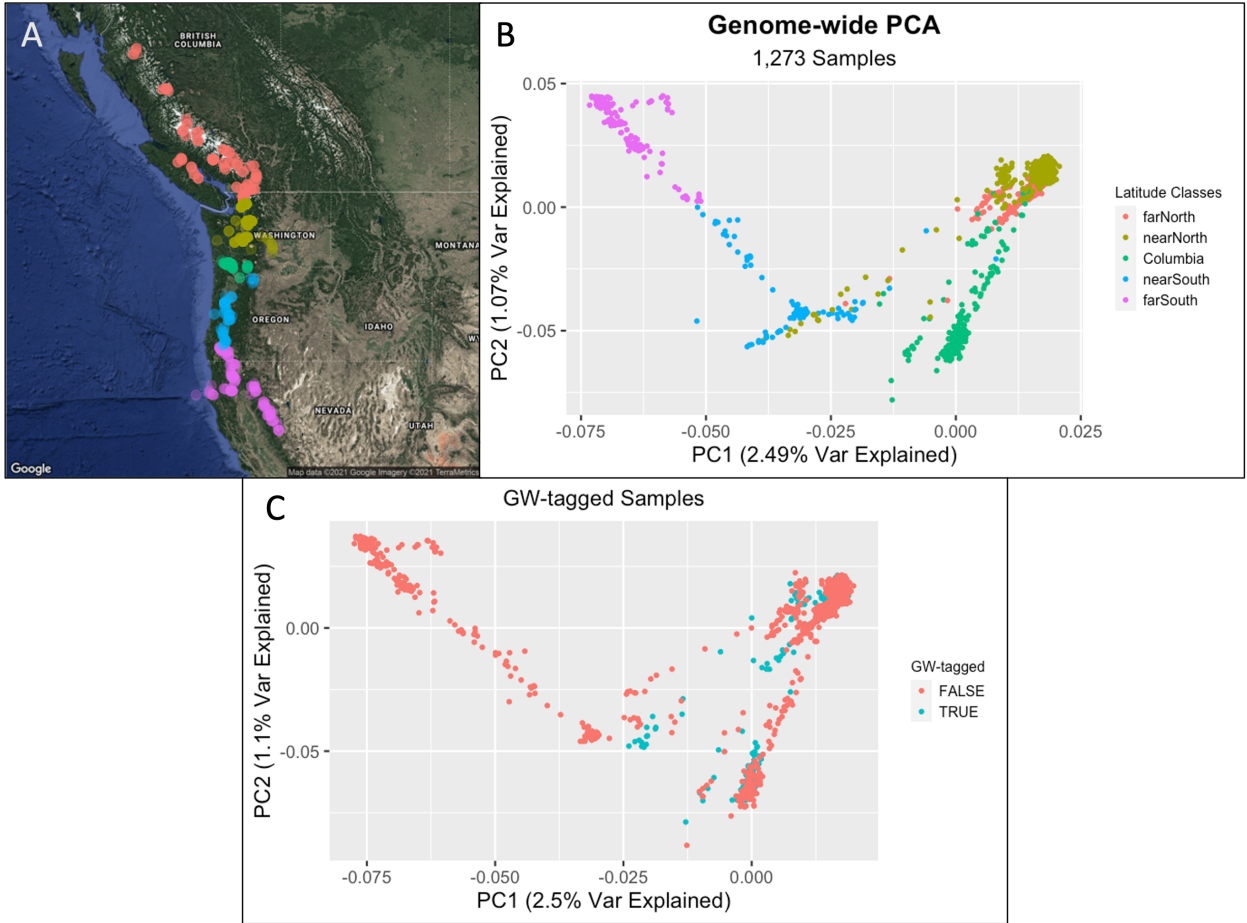


Figure 4.1: A) Map of the sampling location of all poplar WGS. B) PCA of WGS for the entire poplar dataset. The colorings correspond to our partitioning of the population into broad latitudinal classes that reflect population structure. These colorings illustrate how geographic position in panel A maps to PC coordinates in panel B. C) Showing where the GW-tagged samples map in the PCA.

Our task comes with a natural test set of 282 genotypes. These were collected from the field prior to GPS technology so geo-spatial labels were only assigned recently, using the best available information from the sampling records. Because of this protocol, we have low-confidence in their geo-spatial labels relative to the rest of the population. These genotypes are marked with a ‘GW’ tag, and we will refer to them as the GW test set (shown in Panel C of Figure 4.1).

Our desired use-case is to predict the latitudinal location of these samples. Without ground-truth labels, we will be unable to quantify our performance on them explicitly so we will use evidence based on where they cluster in the aligned genotype PCA to determine whether our new class predictions are plausible.

4.3.2 MinHashing Unaligned Reads

To predict from unaligned sequences we use the Mash software [23] to minHash our reads. In summary, this process creates a reduced representation of the genome that has been shown to accurately reflect genome-wide structure [23]. It does this by randomly sub-sampling kmer’s from the read fragments using a minHashing approach. When using Mash, the user must define the kmer length to use, as well as the number of hash functions to store determining the sketch size (s). For our study we chose a kmer length of 21. This is the default in MASH and their studies demonstrate this kmer length robustly maps to Average Nucleotide Identity (an alignment based measure of mutation distance) across different sketch sizes.

Mash states that “Increasing sketch size improves the accuracy of Mash estimates, especially for divergent genomes”. Given that our study takes place within the context of a single species we anticipate that sketch size will only have a small effect on our predictive capacity. To test this, we run MASH at three different sketch sizes: $s=500$, 2000, and 4000. We train and test our prediction algorithms across all three sketch sizes and compare performance (see Figure 4.2).

Once sketched, we implement a novel application of the Mash output. Typically, these sketches are used for a simple pairwise comparison of genomes to estimate genetic distance. For a pair of genotypes, this is done by comparing the sets of hash functions in each genome sketch. Here, instead of only looking at pairwise comparisons, we look at set membership

across the entire population. This is achieved by building a presence-absence matrix for the hash functions in each sketch across the population. Assuming no hash collisions, this means our matrix represents a random sampling of kmers, with a 1 indicating that kmer as present in a genotype and 0 indicating its absence. As a proof of concept, we were curious what percentage of aligned predictive capacity could be recovered using this mapping in an unaligned context. We thus proceed with two different input types for unaligned prediction: mash-based pairwise distance estimates, and our novel presence-absence matrix built off mash sketches. We will refer to these inputs as the distance-based and sketch-based inputs respectively.

4.3.3 Machine Learning Models for Prediction

Having built sketches and used them to encode a presence-absence matrix, we now pass both these inputs to machine learning (ML) models for trait prediction. With sketch-based input, we denote a matrix X with size (N, F) with each N row corresponding to samples and each F column corresponding to hashes. Before prediction, we perform a principal component analysis (PCA) to reduce the dataset to 50 features, which are low-dimensional projected versions of the hashes in the original matrix. This dimensionality reduction eases the sparsity burden of the original input and we pass these 50 principal components to our prediction methods.

With distance-based input, denote a distance matrix D as an (N, N) matrix with N corresponding to the number of samples in the dataset. Each entry $D_{i,j}$ is the Mash distance between samples i and j . This matrix is passed directly to the prediction algorithms.

Similar to the sketch-based representation, the alignment-based representation consists of a matrix A of size (N, F) , with N samples and F variants from the whole genome VCF file. A PCA is performed on this VCF file using the SNPRelate package [39] and we use all the principal components ($n=32$) from this projected version of the aligned whole genome as input for prediction.

We use four different machine learning regression models for the sketch-based and alignment-based learning: ElasticNet [43], Random Forest [4], Lasso [32], and XGBoost [5]. We chose these algorithms to represent a diversity of machine learning techniques,

including non-linear and linear methods. ElasticNet and Lasso are both linear models, and Random Forest and XGBoost are both tree-based nonlinear models. For the distance-based representations, we use k-Nearest Neighbors algorithm because it predicts based on distances rather than vector inputs.

For each data representation, all models are fit via 5-fold cross validation, and R^2 is used to score the model’s performance. After obtaining R^2 scores on each fold, the score is averaged across the 5 folds to obtain the final score. There is some variability in performance across the different models but overall XGBoost gives the most promising performance for each task. Thus, we will continue using XGBoost as the main predictive model in future analyses.

4.4 Results

4.4.1 Performance Across Different Sketch Sizes

Figure 4.2 shows prediction performance across different sketch sizes. Most notably, for seasonality of precipitation, precipitation of the driest quarter, and precipitation of the warmest quarter sketch size seems to have a large impact on prediction performance. This goes against our hypothesis that prediction results would be largely unchanged by shifting sketch size due to a lack of strong divergence in our population of whole-genome sequences. In general, we observe a decreasing trend in prediction performance with decreasing sketch size.

Interestingly, in association mapping results for these traits (see Chapter 3) we observe that three out of four of our top performing traits have a strong regulator at the same loci on chromosome 16. This leads us to hypothesize that the boosted performance seen for these traits could be due to the chance sampling of kmer’s from this region while mash sketching. This would imply that for traits showing very poor performance, we could boost this with sufficient sampling of kmer’s from regulatory regions.

In general, we see poor predictive capacity from unaligned reads. This is especially true for the geo-spatial trait of greatest interest in this study: latitude. Figure 4.3 demonstrates

we achieve much better performance using the alignment-based representations and we will thus be using a model built off this data for correcting the geo-spatial labels in the GW test set.

4.4.2 Performance Across Data Representations

Figure 4.3 shows results across all three data representations. We observe that the alignment-based protocol substantially outperforms the others. The assumption is that correctly aligning reads builds whole-genome sequences into a larger scale structure with more meaningful signal than the short fragments. The PCA should capture this longer range signal, thereby taking advantage of the benefits of aligned genomes. With this in mind, we can use the alignment-based pipeline to define a predictive ceiling. Future attempts to boost the performance of prediction from unaligned modalities can quantify performance by the percentage of this upper limit that is recovered.

Notably for our use-case, performance is highest for alignment-based prediction of latitude. Our next step is to assess whether this high performance can accurately predict the latitude of genotypes in our GW test set.

4.4.3 Predicting Latitude for GW Test Set

In order to assess whether predicted latitude labels are plausible for the GW test set we can first examine the performance on genotypes with high-confidence labels. To achieve this, we remove the GW test set from the training data and create a new test set of 200 genotypes. We examine the distribution of latitude differences between these high confidence labels and our prediction. This difference is normally distributed with a mean of 0.09. The skew to the right is due to outliers in the right tail. This implies we have a strong coherence between our high-confidence latitude labels and our predictions, which is expected given the high R^2 value of 0.97 from the prediction results in Figure 4.3.

We can use these results to investigate the accuracy of the labels in the GW test set. If the GW labels are accurate, we expect to see a similar distribution for the difference between metadata and predictions. Figure 4.4 panel A shows this distribution with a mean value of

0.41 marked with the black vertical line. Although highly populated near 0, there is a long tail extending to the right. This tail is much longer than that seen for the high confidence labels. This suggests that many of the GW samples are being predicted as further north than their meta data claims.

If we look at the genotypes that reside in these tails (those with $abs(diff) > 2$), we find an enrichment of samples from the Willamette river system. Pursuing this, we can inspect both GW and non-GW samples from Willamette and compare their PCA mappings by partitioning them into two groups: high confidence labels (non-GW) and low confidence labels (GW). If incorrectly labeled we would expect to see that these groups don't cluster together. Panel B in Figure 4.4 shows separation between these two clusters. Using predicted latitude, we can define new neighbors for these mis-labeled genotypes and look at the PCA to see if they map more closely to the GW samples. Figure 4.4 panel B shows these new neighbors colored in green and we do indeed see greater correspondence between them and the previously assigned GW Willamette samples. This all builds to suggest that these GW-tagged samples have been incorrectly assigned to the Willamette river system. Our analysis reveals that the genetic architecture of these genotypes imply they were sampled from latitudes further North than their labels claim. We can thus re-label these genotypes with their predicted latitude and have confidence this more accurately reflects their true sampling location.

4.5 Discussion

In this study we have leveraged machine learning algorithms applied to whole-genome sequences to predict the latitudinal coordinates of mislabeled samples. This approach relies solely on the genetic fingerprint of each genotype to identify the location it was sampled from. This allows us to both validate the sampling location of genotypes that were previously assigned high confidence labels, as well as correct the sampling location of genotypes with low confidence labels.

Future studies will attempt to further develop our predictive capacity using unaligned sequence fragments. This is an attractive objective as it increases the utility of our tool,

allowing researchers to assign genotype labels a confidence score prior to the time-consuming protocols of variant-calling and alignment. The initial studies undertaken in this paper outline two paths to improvement. The first is to define a target set of kmer's to screen for. Previous association mapping studies have identified regulatory hotspots for these traits of interest. We hypothesize that if we could sample kmer's directly from these hotspots we would significantly boost performance. Our assumption is that this happened coincidentally for the traits that show superior performance in Figure 4.2.

A second approach would be to increase the sketch size of the mash sketching procedure. The computational time of this sketching procedure is largely insensitive to sketch size. We hypothesize that increasing the number of hash functions that are retained may result in significant boosts in performance if the genome could be sufficiently sampled. This aligns with the results in Figure 4.2 where we see increasing performance with increasing sketch size.

In the age of large data sets collected by humans, errors in data collection are common. Being able to identify and correct such mistakes without introducing further biases provides a means of cost-recovery, especially in the context of whole-genome sequences where data production is expensive. This study serves as an initial example of tools that can be built for this purpose, as well as methods to validate their correctness. All code for the project is freely distributed on GitHub (<https://github.com/owencqueen/MashPredict>) and we hope that it might serve as a starting point for future investigations.

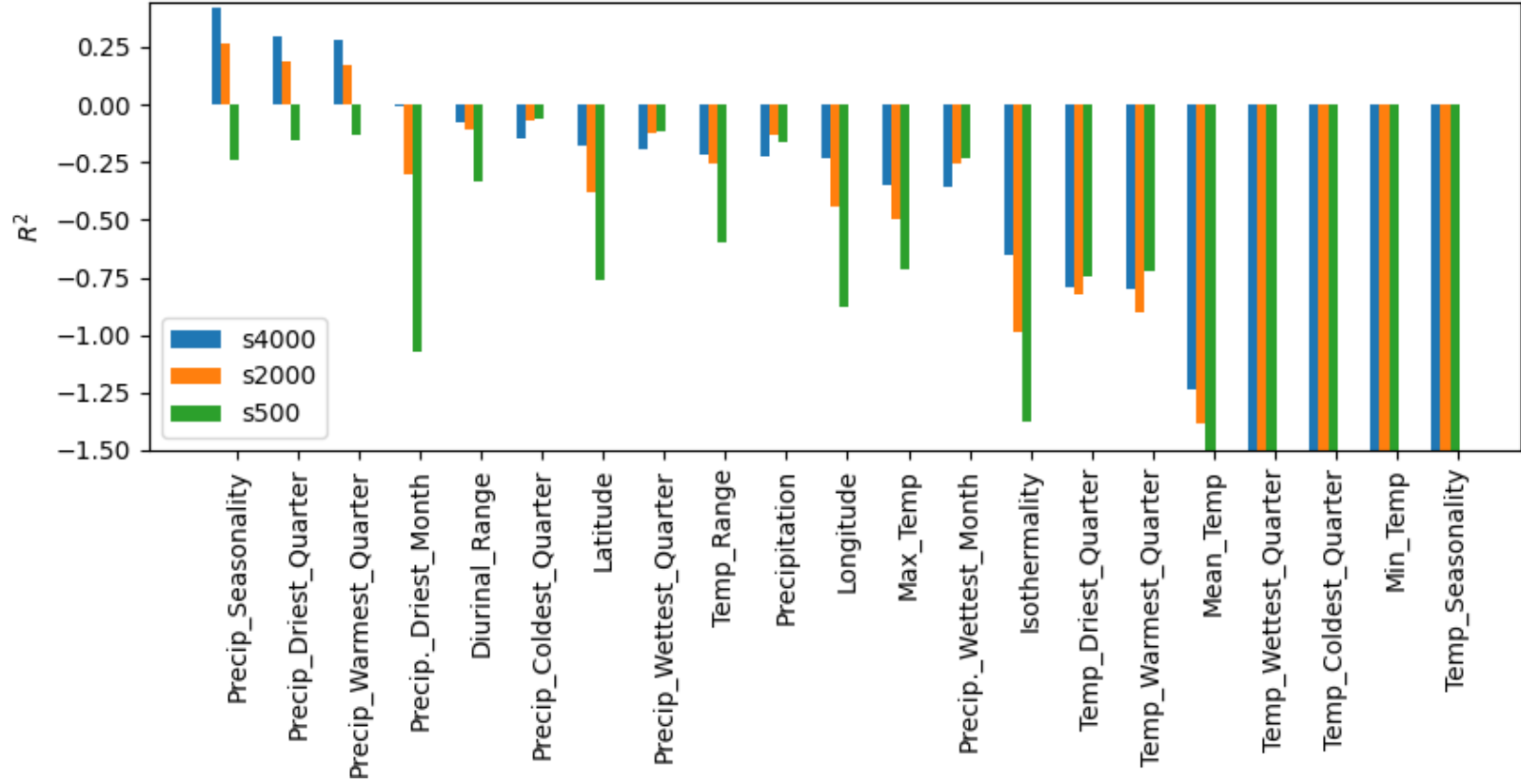


Figure 4.2: Comparison of performance across different sketch sizes (s) for sketch-based prediction. Y-axis represents R^2 performance for the regression fit. Each color bar represents a different size of mash sketching, ranging from 500 to 4,000 hash functions. Traits are ordered by performance of the $s=4,000$ task.

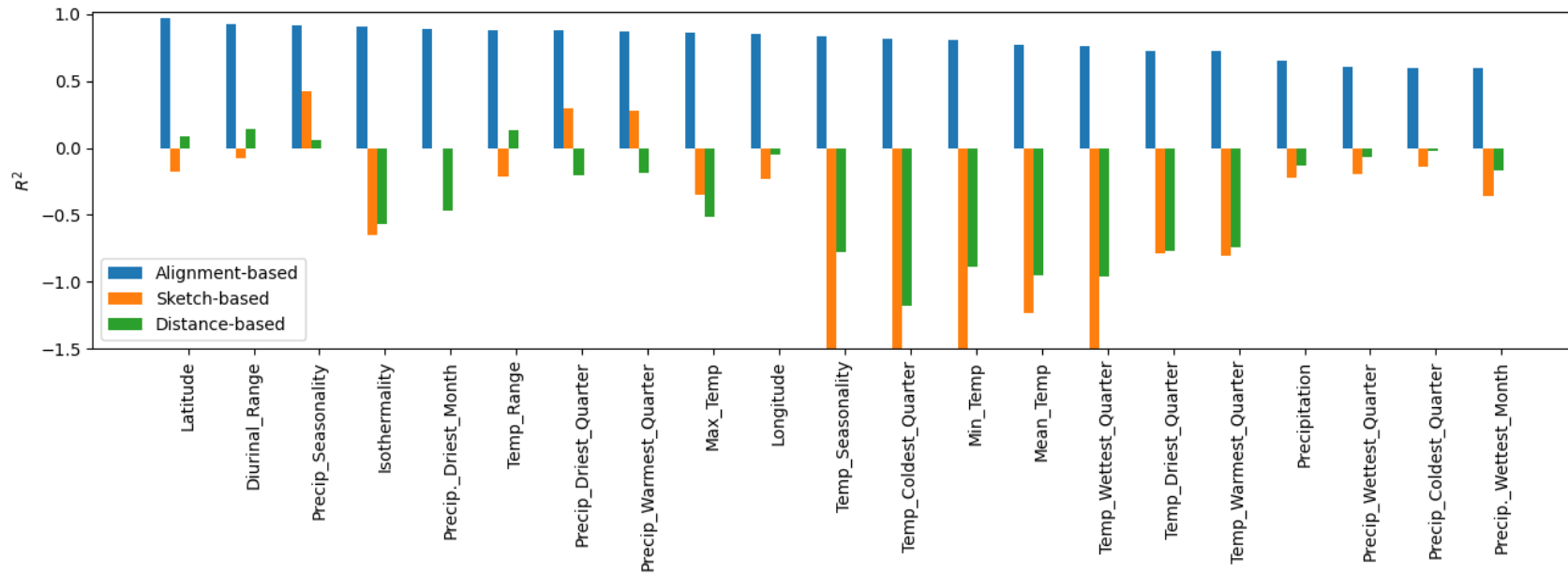


Figure 4.3: Comparison of performance across different data representations. Y-axis represents R^2 performance for the regression fit. Each color bar represents a different data representation. Traits are ordered by performance on the alignment-based representation.

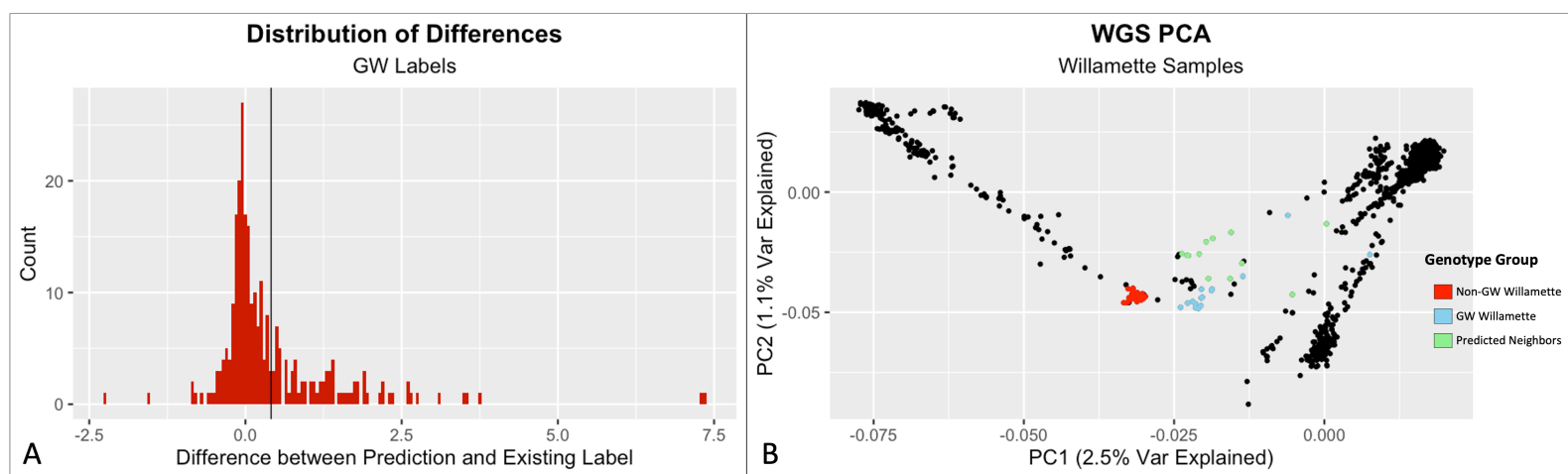


Figure 4.4: A) Distribution of the difference between predicted latitude and low-confidence GW latitude labels. Mean value of 0.41 is marked by the vertical black line. B) Examining Willamette river system genotypes within the context of the whole-genome PCA. Genotypes marked in red are Willamette genotypes with high confidence labels that we expect map to the correct position in the PCA. Genotypes marked in blue are GW genotypes that are likely incorrectly labeled as Willamette. Genotypes marked in green are the new neighbors of the GW genotypes based on predicted latitude.

Chapter 5

Conclusions

Whole-genome sequences are a powerful tool for engineering biofuel stocks to address national security concerns. The recent resequencing of 207 new poplar genomes completes our data collection so that it spans the entire natural species range. For the first time, this allows investigation of species wide dynamics in response to climactic pressures. A mechanistic understanding of these responses would mean that poplar could be bio-engineered for success within any environment that falls within its range of natural variation.

This study has leveraged this data to make strides towards this goal. Chapter 2 utilized a proof of concept analysis restricted to a well studied portion of the species range. This demonstrated that the techniques to be applied species wide were well founded and produced results that are biologically plausible.

Chapter 3 extended this work to identify regulatory hot spots as responses to a wide range of climactic variables. This provided a long list of high probability targets for genome editing. Results of downstream experimental set-ups will validate the effectiveness of modulation of these loci for climate response, providing mechanistic understanding of how poplar adapts.

Chapter 4 took a data centric approach to validating and correcting meta-data labels for this new data collection. Mislabeling is common in data collection and using the underlying genomic signal to validate whether low-confidence labels are accurate provides a useful diagnostic tool that can be used in the lab moving forward.

Bibliography

- [pop] Poplar (populus spp.) trees for biofuel production. <https://farm-energy.extension.org/poplar-populus-spp-trees-for-biofuel-production/>. Accessed: 2021-11-29. 2
- [2] Balatinecz, J. J. and Kretschmann, D. E. (2001). Properties and utilization of poplar wood. poplar culture in north america. *NRC Research Press*, pages 277–291. 2
- [3] Birnbaum, J., Scharf, S., Schmidt, S., Jonscher, E., Hoeijmakers, W. A. M., Flemming, S., Toenhake, C. G., Schmitt, M., Sabitzki, R., Bergmann, B., et al. (2020). A kelch13-defined endocytosis pathway mediates artemisinin resistance in malaria parasites. *Science*, 367(6473):51–59. 23, 26
- [4] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. 49
- [5] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. 49
- [6] Chollet, F. (2015). keras. <https://github.com/fchollet/keras>. 14
- [7] Fick, S. and Hijmans, R. (2017). Worldclim 2: new 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37:4302–4315. 31, 34, 46
- [8] Flint, J., Valdar, W., Shifman, S., and Mott, R. (2005). STRATEGIES FOR MAPPING AND CLONING QUANTITATIVE TRAIT GENES IN RODENTS. *Nature Reviews Genetics*, 6(April). 2
- [9] Ford, C. and Janies, D. (2020). Ensemble machine learning modeling for the prediction of artemisinin resistance in malaria. *F1000Research*, 9(62). ix, 23, 28
- [10] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22. 21
- [11] Gonzales-Vigil, E., Hefer, C. A., von Loessl, M. E., La Mantia, J., and Mansfield, S. D. (2017). Exploiting natural variation to uncover an alkene biosynthetic enzyme in poplar. *The Plant Cell*, 29(8):2000–2015. 18, 25

- [12] Gu, C., Todorov, A., and Rao, D. (1997). Genome screening using extremely discordant and extremely concordant sib pairs. *Genetic epidemiology*, 14(6):791–796. [6](#)
- [13] Huang, B. and Lin, D. (2007). Efficient association mapping of quantitative trait loci with selective genotyping. *The American Journal of Human Genetics*, 80(3):567–576. [10](#)
- [14] Isebrands, J. G. et al. (2014). *Poplars and Willows: Trees for Society and the Environment*. The Food and Agriculture Organization of the United Nations. [2](#)
- [15] John, C., Muchero, W., and Emrich, S. (2020). Extreme phenotype sampling improves lasso and random forest marker selection for complex traits. *International Conference on Bioinformatics and Biomedicine*, pages 791–796. [31](#)
- [16] Li, D., Lewinger, J. P., Gauderman, W. J., Murcray, C. E., and Conti, D. (2011). Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic epidemiology*, 35(8):790–799. [6](#), [10](#), [26](#), [27](#)
- [17] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *BioarXiv*. [32](#), [46](#)
- [18] Li, X., Kumar, S., McDew-White, M., Haile, M., Cheeseman, I. H., Emrich, S., Button-Simons, K., Nosten, F., Kappe, S. H. I., Ferdig, M. T., Anderson, T. J. C., and Vaughan, A. M. (2019). Genetic mapping of fitness determinants across the malaria parasite *Plasmodium falciparum* life cycle. *PLOS Genetics*, 15(10):1–25. [7](#), [10](#)
- [19] McGoey, B., Hodgins, K., and Stinchcombe, J. (2020). Parallel flowering time clines in native and introduced ragweed populations are likely due to adaptation. *Ecology and Evolution*. [35](#)
- [20] McKown, A. D. et al. (2014a). Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *populus trichocarpa*. *New Phytologist*. [35](#)
- [21] McKown, A. D., Guy, R. D., Klápště, J., Geraldes, A., Friedmann, M., Cronk, Q. C., El-Kassaby, Y. A., Mansfield, S. D., and Douglas, C. J. (2014b). Geographical and

- environmental gradients shape phenotypic trait variation and genetic structure in *populus trichocarpa*. *New Phytologist*, 201(4):1263–1276. [7](#)
- [22] Mok, S. and Ashley, E. A. e. a. (2015). Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. *Science*, 347(6220):431–435. [x](#), [8](#), [10](#), [12](#), [14](#), [18](#)
- [23] Ondov, B., Treangen, T., Melsted, P., et al. (2016). Mash: fast genome and metagenome distance estimation using minhash. *Genome Biology*. [48](#)
- [24] Panarella, M. and Burkett, K. M. (2019). A cautionary note on the effects of population stratification under an extreme phenotype sampling design. *Frontiers in genetics*, 10:398. [26](#), [27](#)
- [25] Piovesan, A., Pelleri, M., Antonaros, F., et al. (2019). On the length, weight and gc content of the human genome. *BMC Res Notes*. [2](#)
- [26] Rahnenfuhrer, A. A. (2022). Enrichment analysis for gene ontology. [34](#)
- [27] Rellstab, C., Gugerli, F., et al. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*. [32](#), [35](#)
- [28] Rubel, F. and Kottek, M. (2010). Observed and projected climate shifts 1901-2100 depicted by world maps of the köppen-geiger climate classification. *Meteorol. Z.* [31](#), [34](#)
- [29] Schroda, M. (2020). Phosphoinositides regulate chloroplast processes. *Proceedings of the National Academy of Science*. [36](#)
- [30] Slavov, G. T., DiFazio, S. P., et al. (2012). Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *populus trichocarpa*. *New Phytologist*, 196:713–725. [x](#), [2](#), [30](#), [38](#), [45](#), [46](#)
- [31] Tibshirani, R. (1996a). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. [10](#)
- [32] Tibshirani, R. (1996b). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. [49](#)

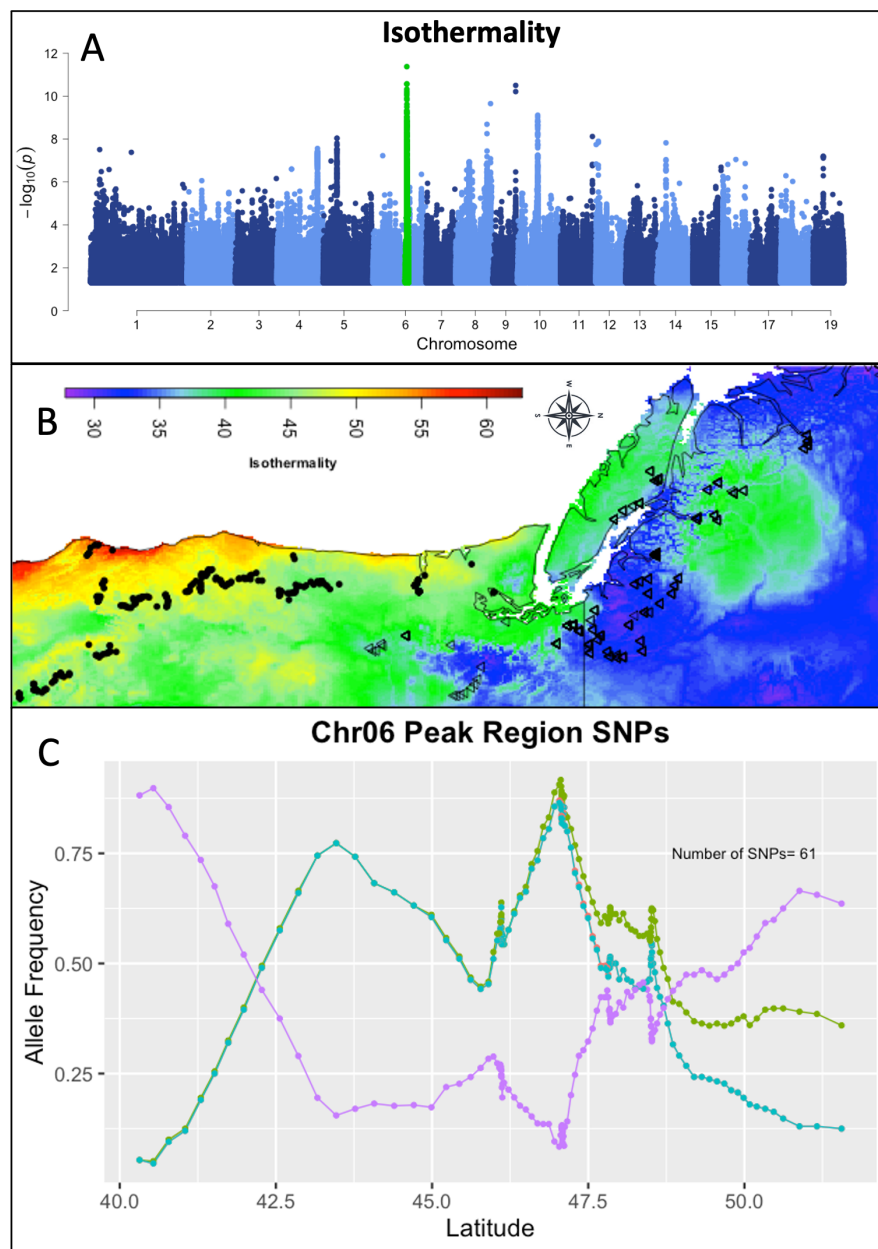
- [33] Tilley, L., Straimer, J., Gnädig, N. F., Ralph, S. A., and Fidock, D. A. (2016). Artemisinin action and resistance in *Plasmodium falciparum*. *Trends in parasitology*, 32(9):682–696. [15](#), [21](#)
- [34] Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., dePamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., de Peer, Y. V., and Rokhsar, D. (2006). The genome of black cottonwood, *populus trichocarpa* (torr. & gray). *Science*, 313(5793):1596–1604. [2](#)
- [35] Van der Auwera, G. and O’Connor, B. (2020). Genomics in the cloud: Using docker, gatk, and wdl in terra. *O’Reilly Media*. [32](#), [46](#)
- [36] Xu, C., Fang, J., Shen, H., Wang, Y.-P., and Deng, H.-W. (2018). Eps-lasso: test for high-dimensional regression under extreme phenotype sampling of continuous traits. *Bioinformatics*, 34(12):1996–2003. [7](#), [11](#), [26](#), [27](#)
- [37] Zhang, J., Yang, Y., et al. (2018). Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of *hct2* in caffeoylquinic acid biosynthesis

- and its regulation by defense-responsive transcription factors in populus. *New Phytologist*, 220(2):502–516. [8](#), [31](#), [32](#), [34](#), [46](#)
- [38] Zhang, L., Wang, Z., Wang, Y., and Huang, B. (2017). Transcriptomic profile of lignocellulose degradation from *trametes versicolor* on poplar wood. *BioRes.* [36](#)
- [39] 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. [10](#), [49](#)
- [40] Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44:821–824. [32](#)
- [41] Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407. [14](#)
- [42] Zhu, L., Tripathi, J., Rocamora, F. M., Miotto, O., van der Pluijm, R., Voss, T. S., Mok, S., Kwiatkowski, D. P., Nosten, F., Day, N. P., et al. (2018). The origins of malaria artemisinin resistance defined by a genetic and transcriptomic background. *Nature Communications*, 9(1):5158. [x](#), [14](#), [18](#), [19](#), [20](#), [22](#), [26](#)
- [43] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: series B (Statistical Methodology)*, 67(2):301–320. [49](#)

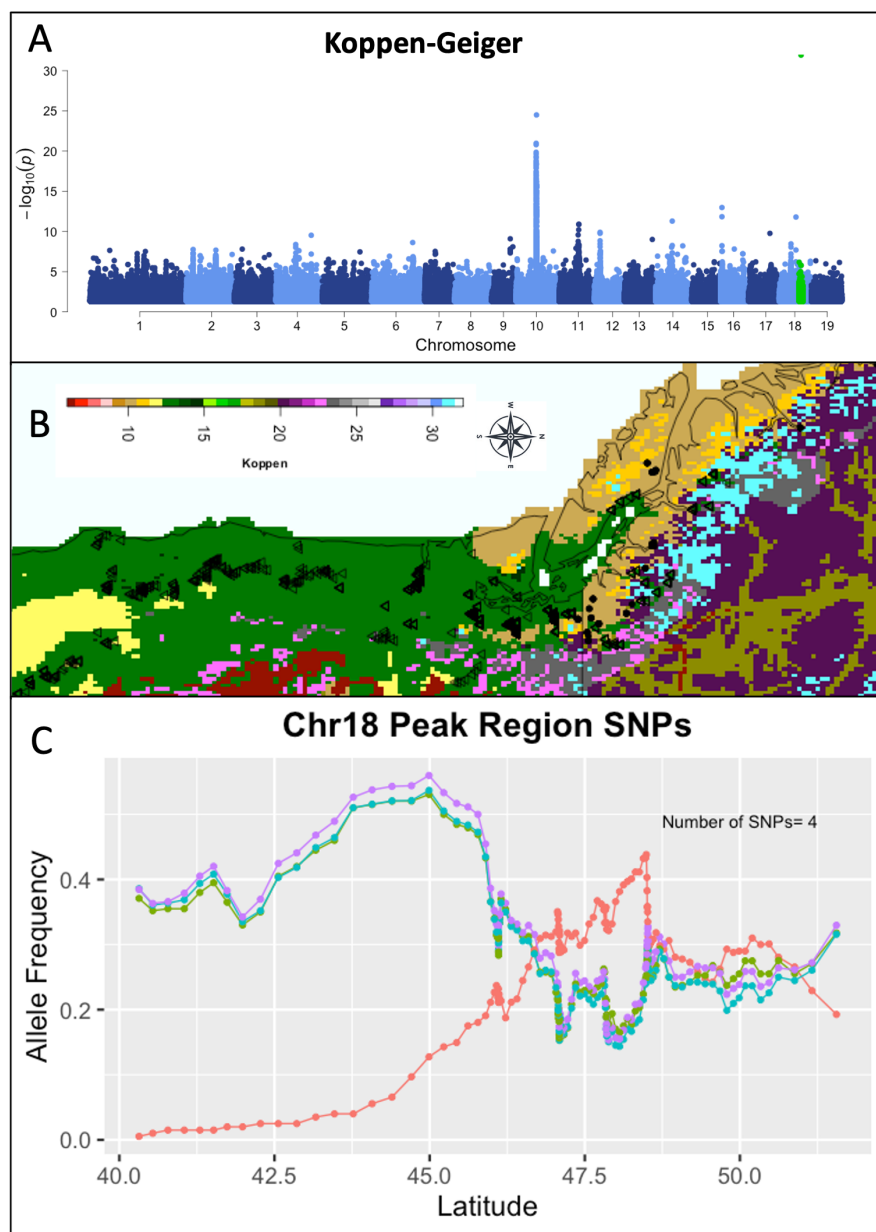
Appendix

Supplementary Figures

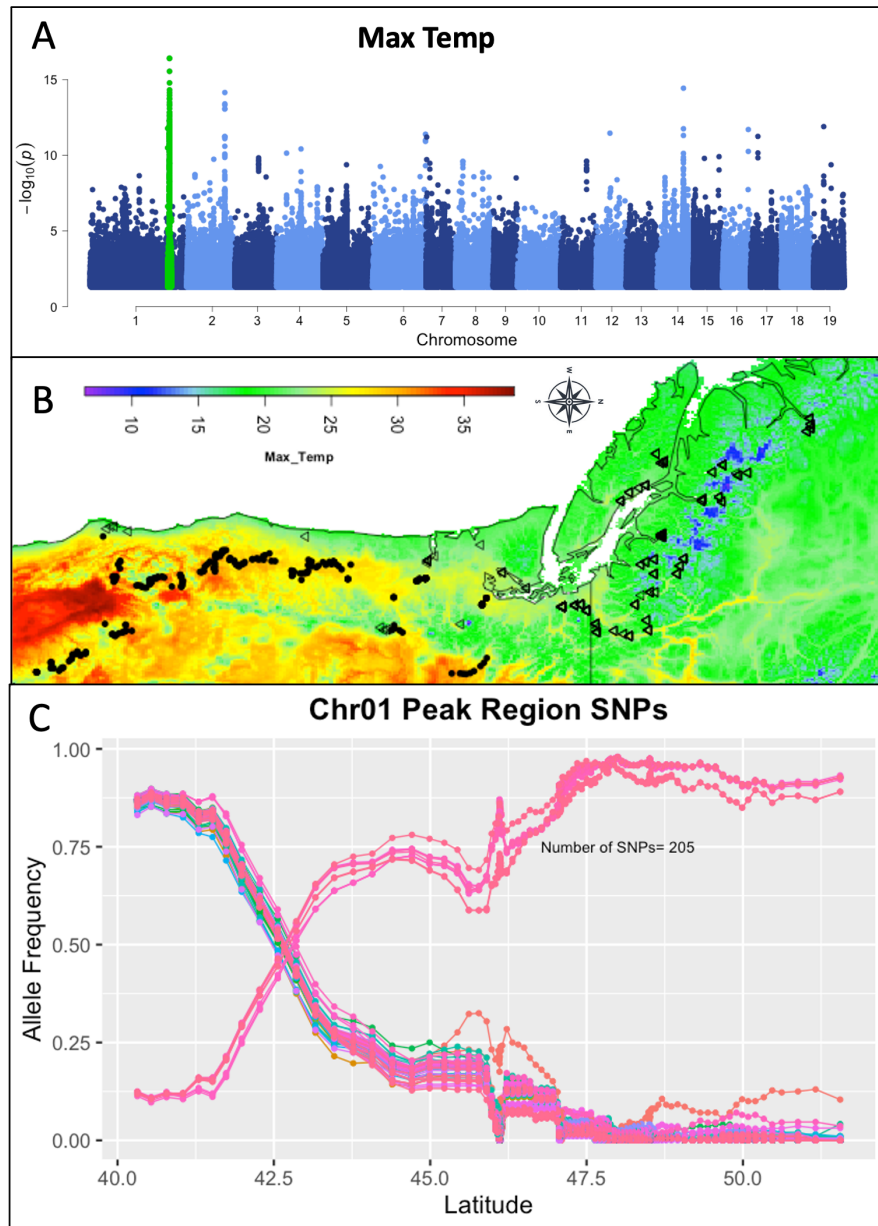
Supplementary Figures 1-18 (S1-S18) below show GWAS results for each climate variable. They have the same presentation as Figure 3.4 above. Panel A shows GWAS p-values for the genotypes sampled from the map in panel B. Panel C shows the results of the allele frequency analysis (see Methods), demonstrating allelic variation and LD block size of correlated variants.



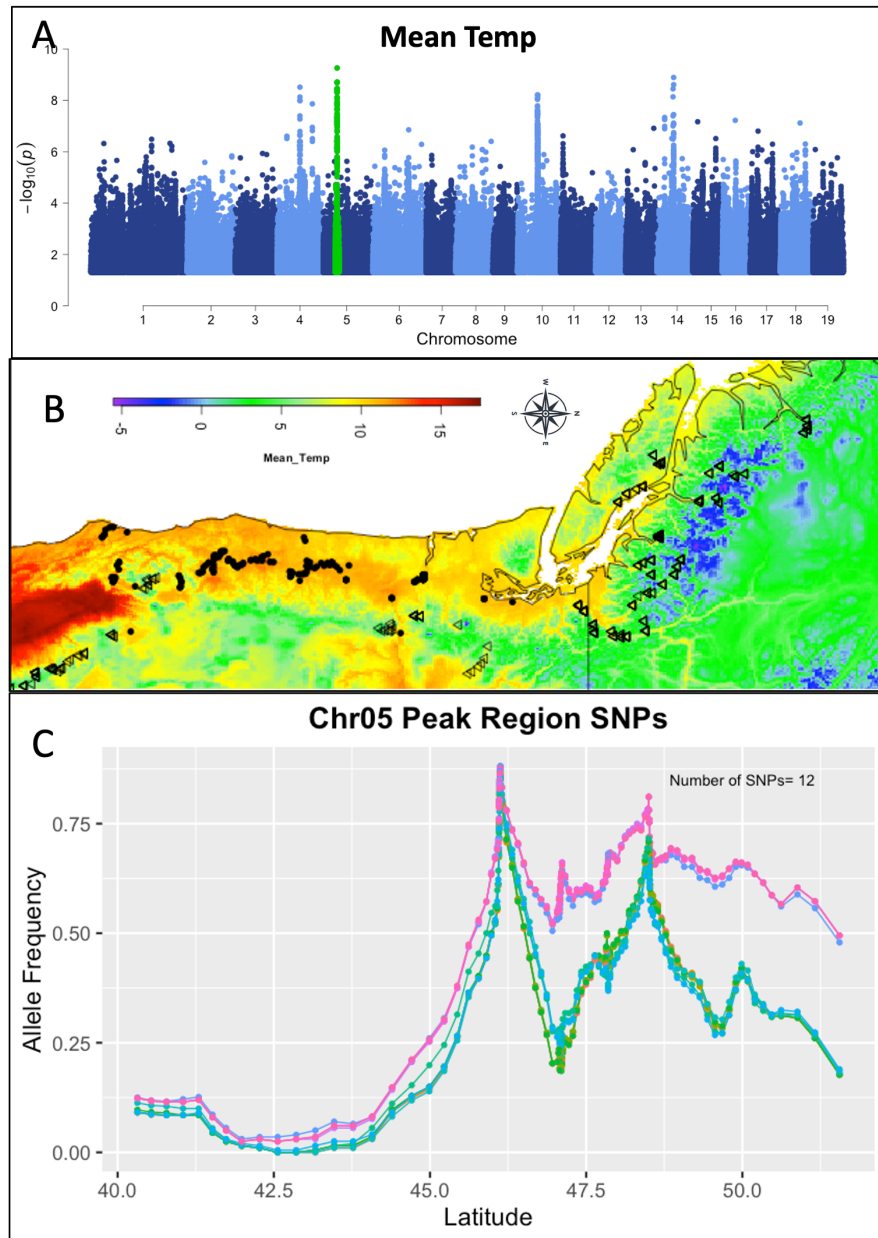
Supplementary Figure 1: Isothermality



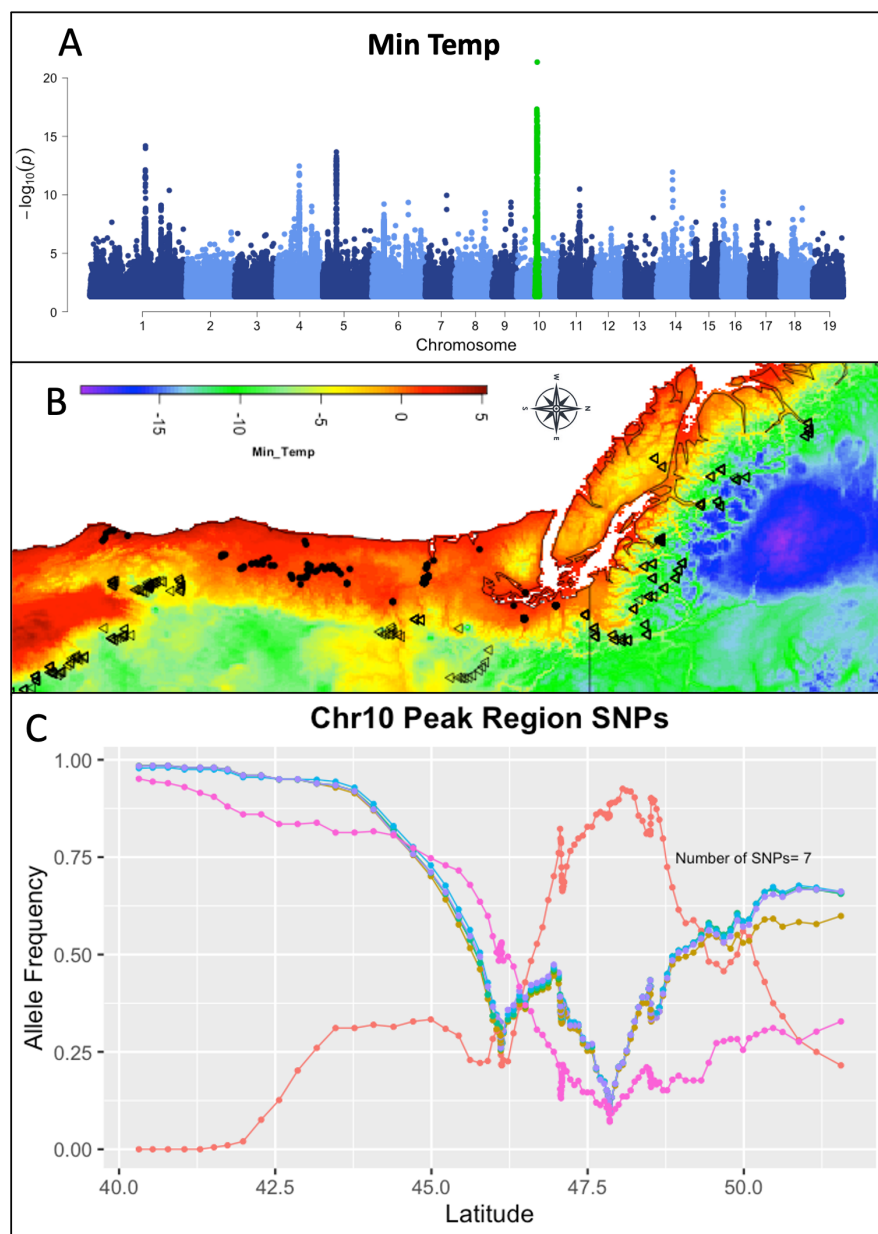
Supplementary Figure 2: Koppen-Geiger Climate Classification



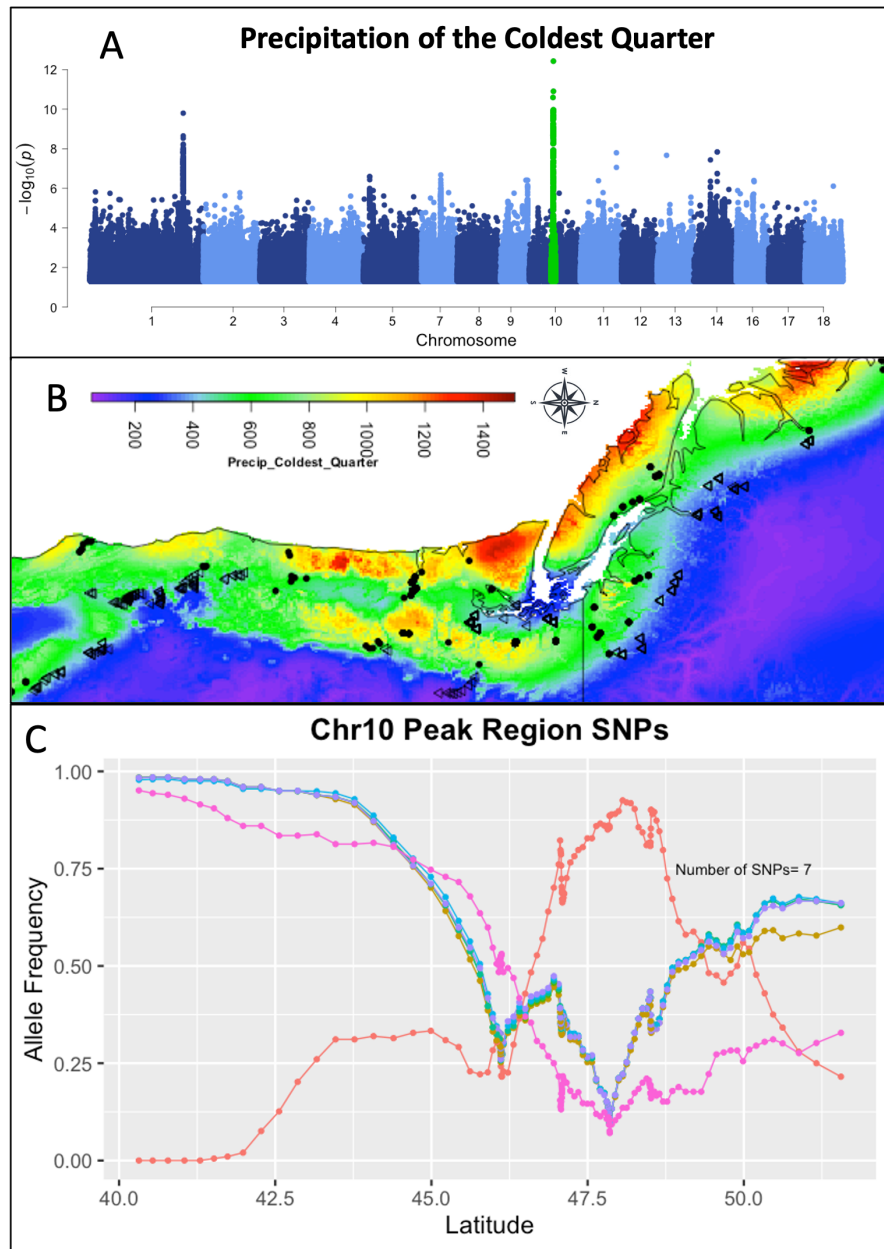
Supplementary Figure 3: Max Temperature



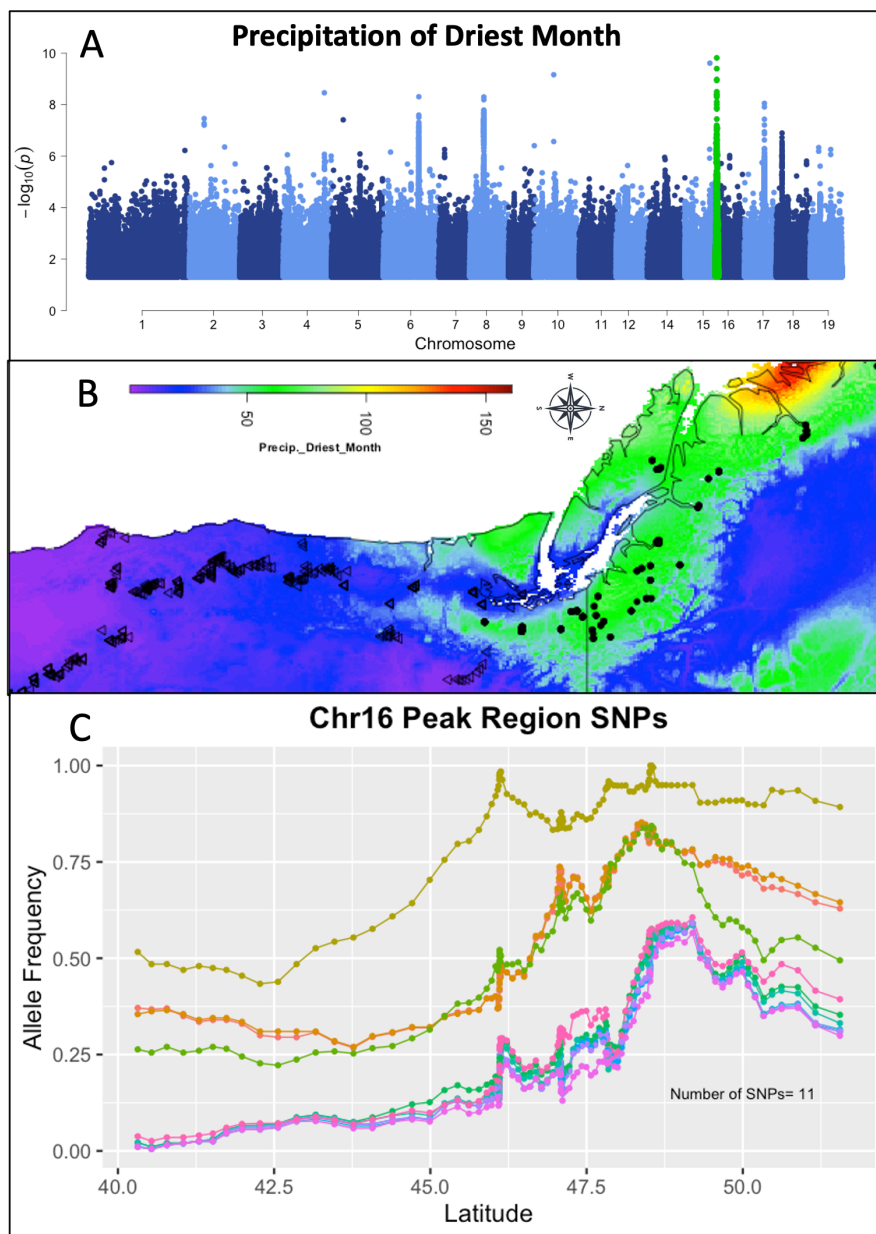
Supplementary Figure 4: Mean Temperature



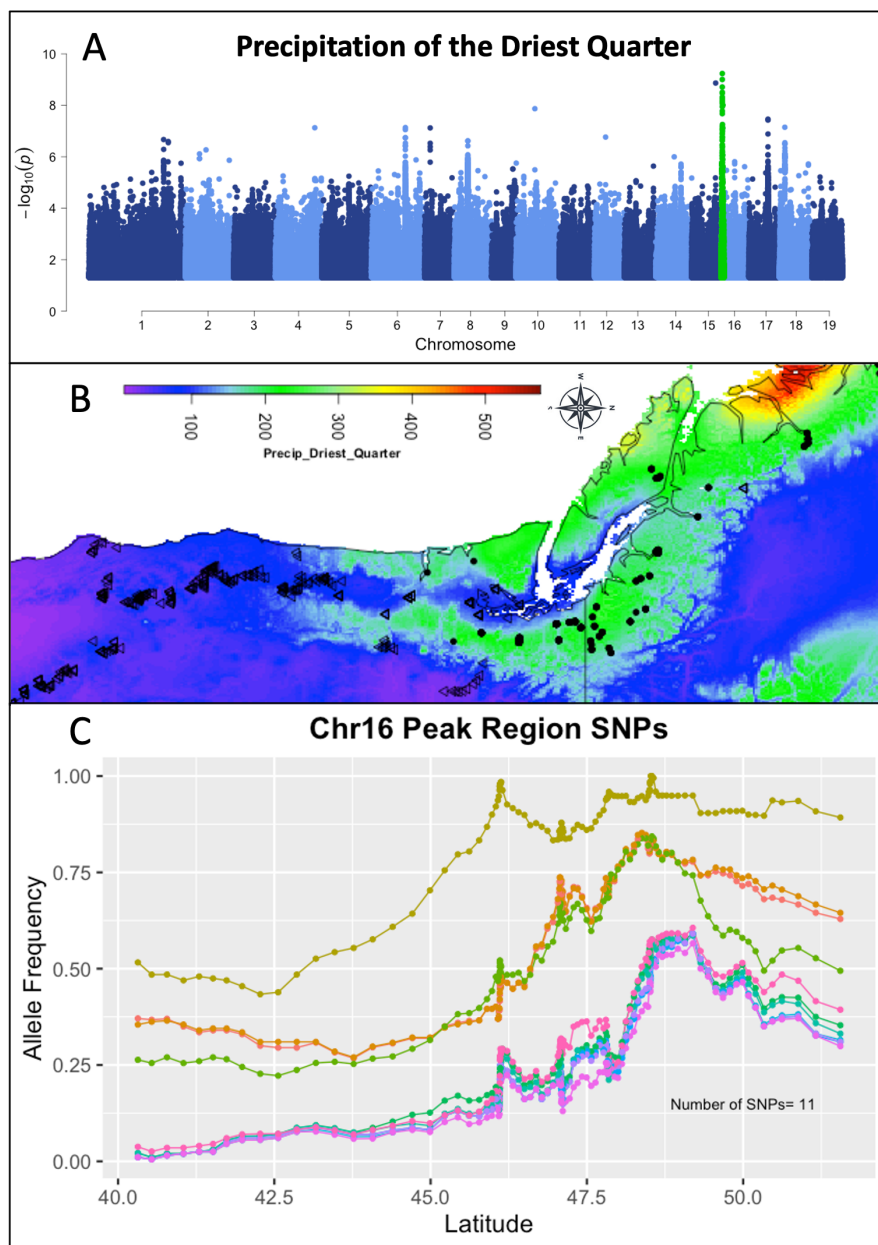
Supplementary Figure 5: Min Temperature



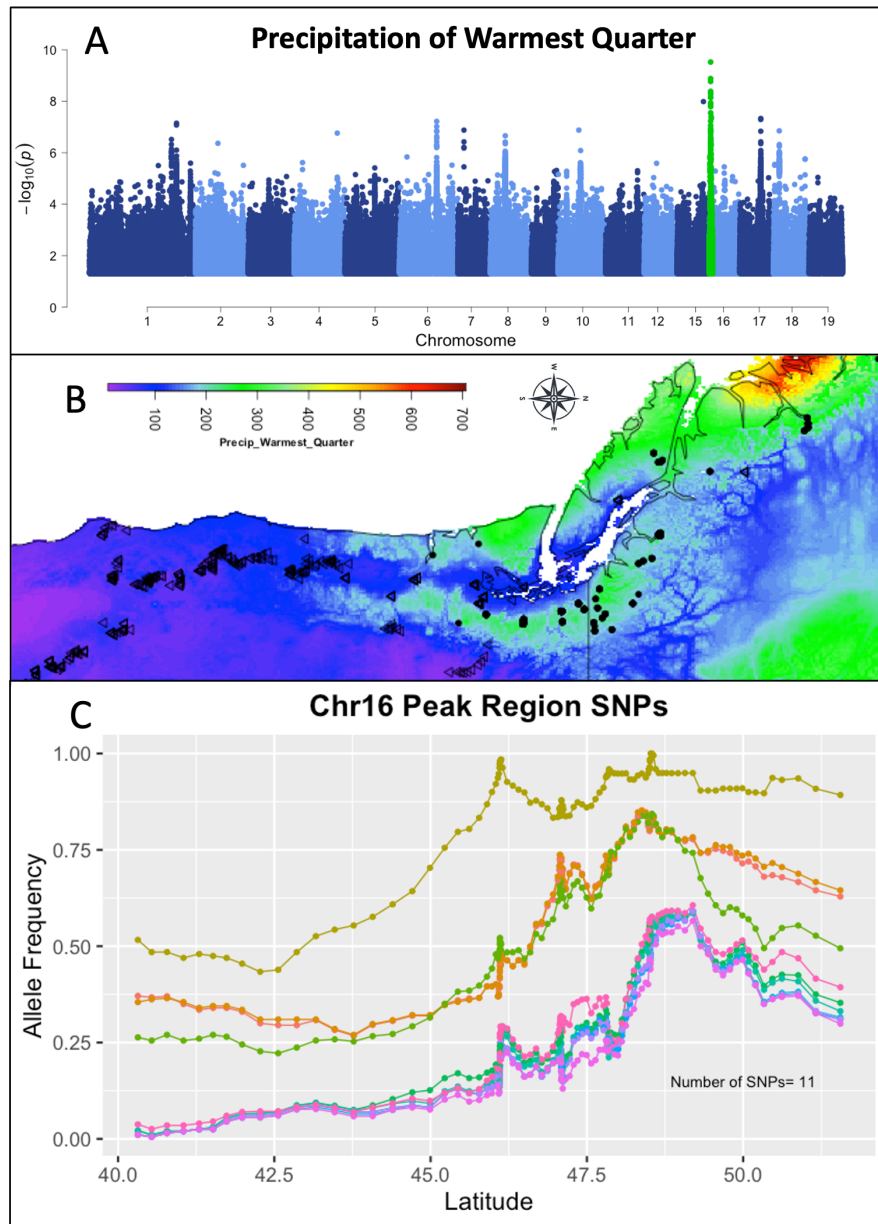
Supplementary Figure 6: Precipitation of the Coldest Quarter



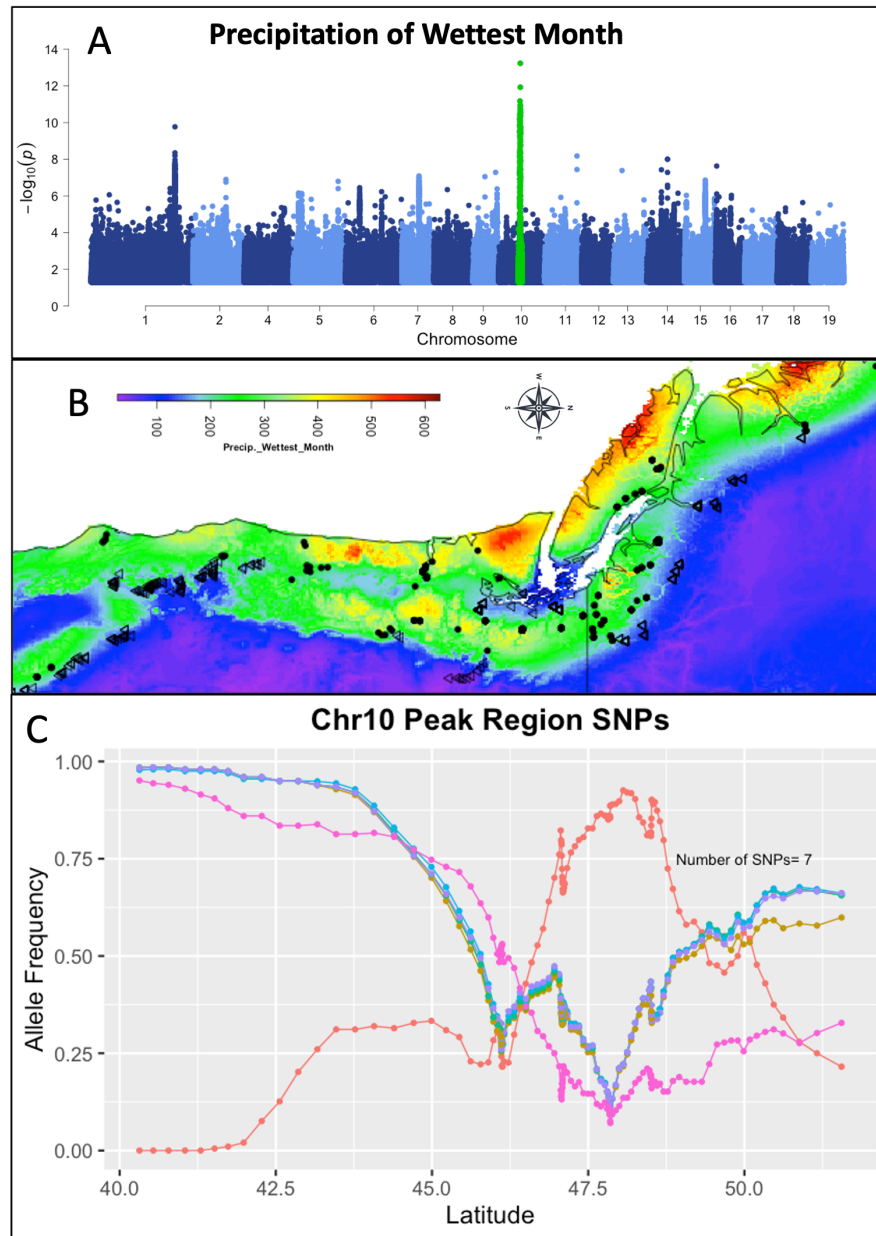
Supplementary Figure 7: Precipitation of the Driest Month



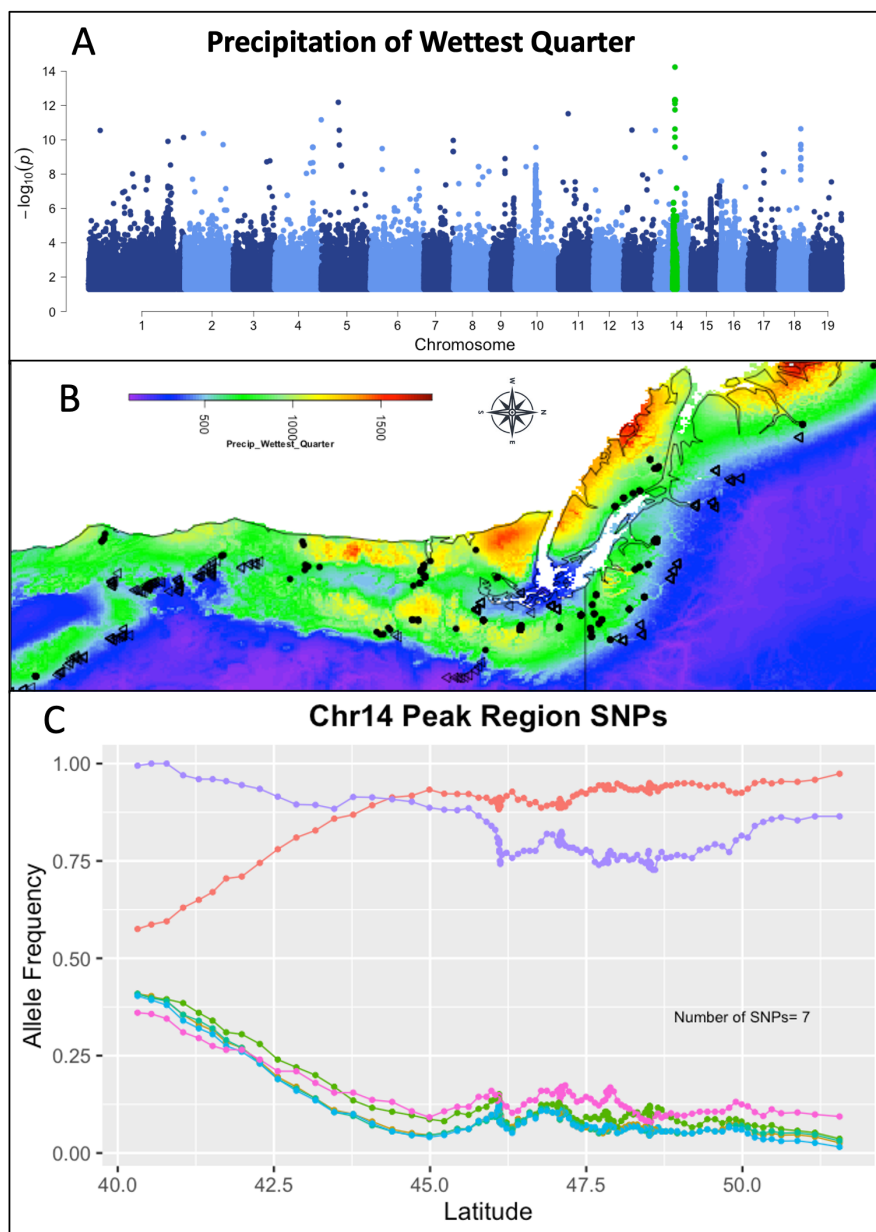
Supplementary Figure 8: Precipitation of the Driest Quarter



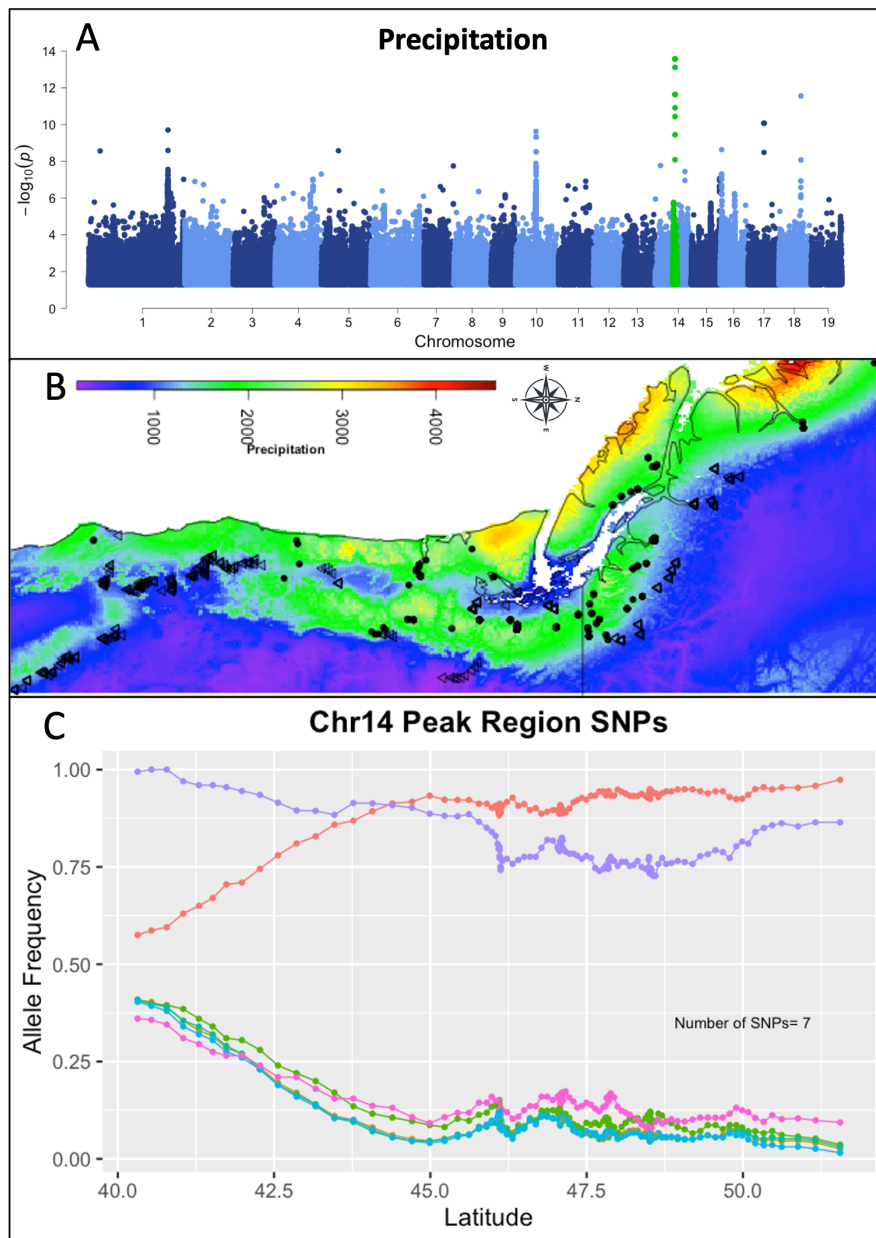
Supplementary Figure 9: Precipitation of the Warmest Quarter



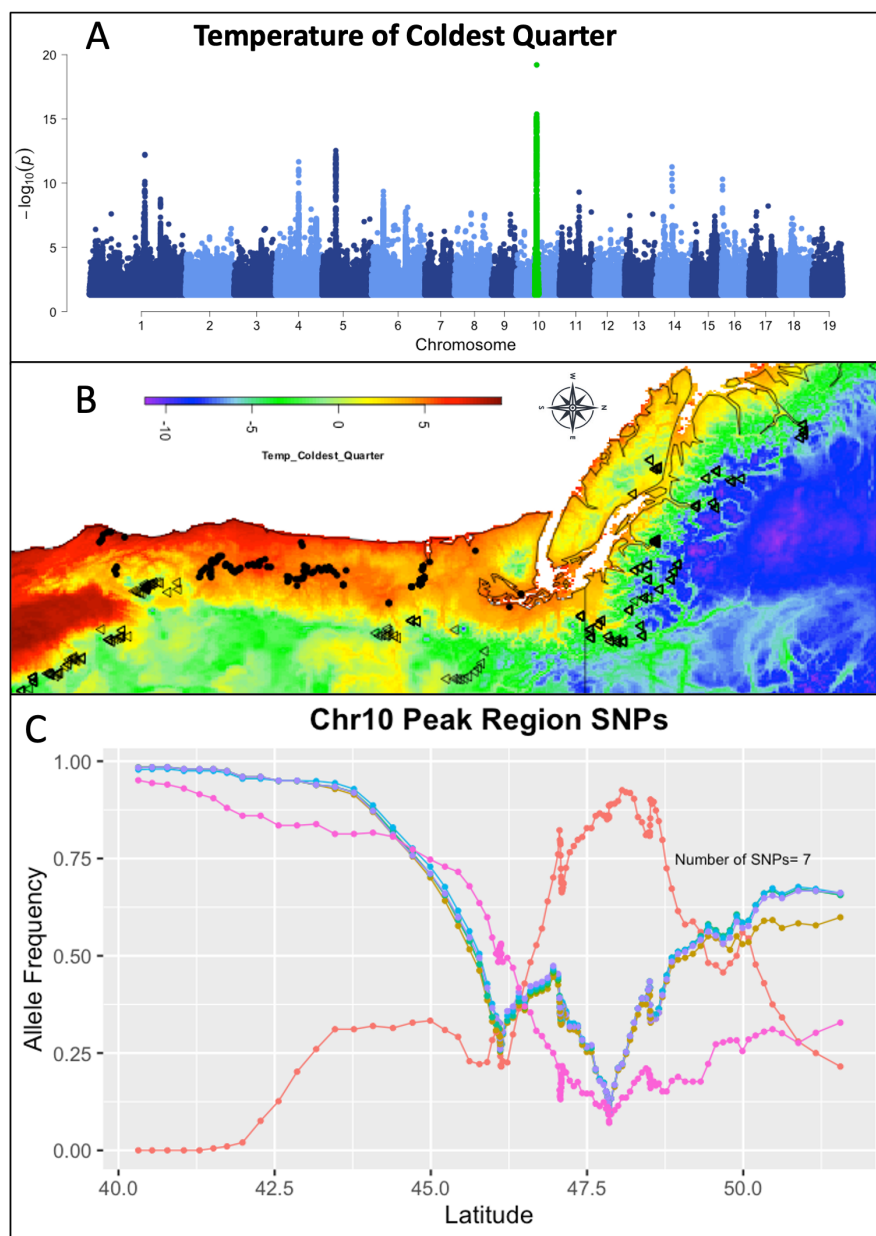
Supplementary Figure 10: Precipitation of the Wettest Month



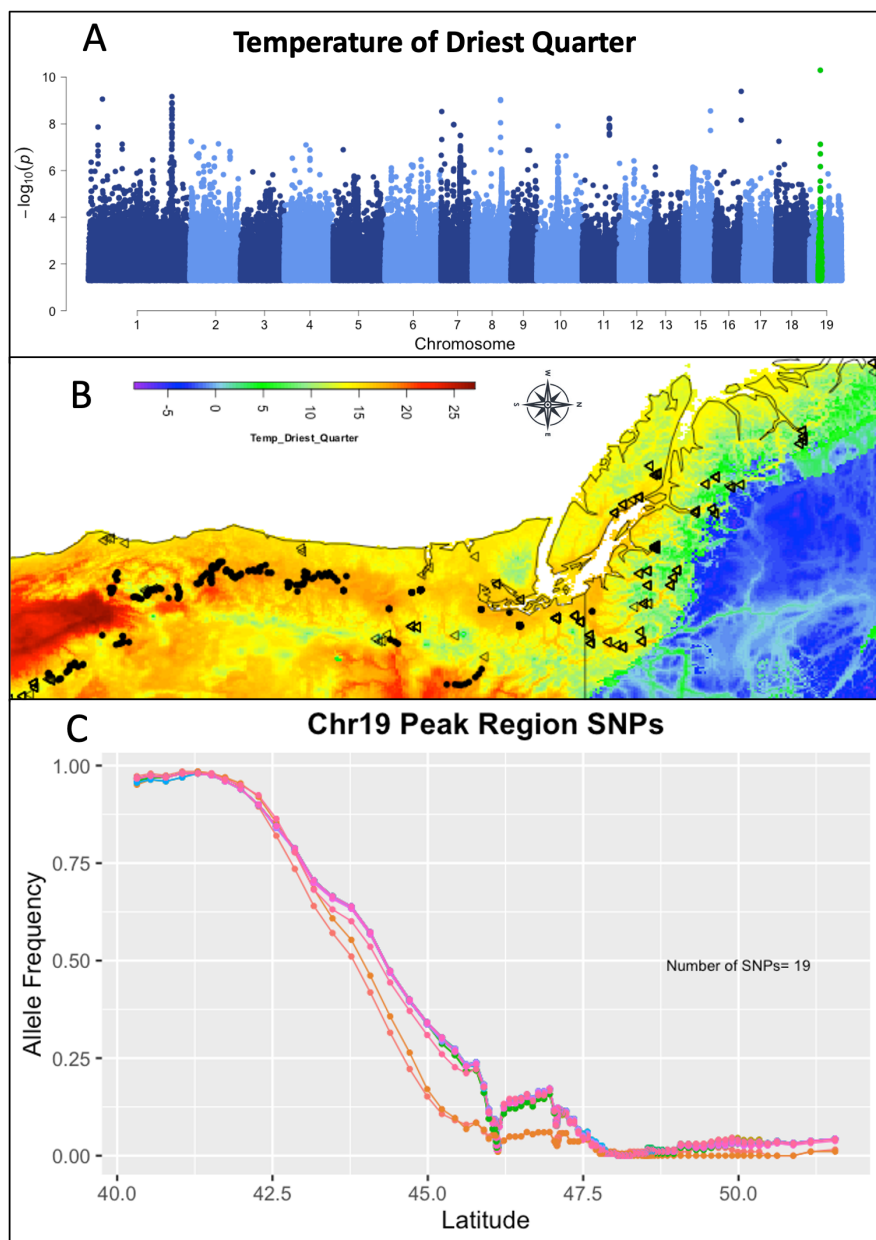
Supplementary Figure 11: Precipitation of the Wettest Quarter



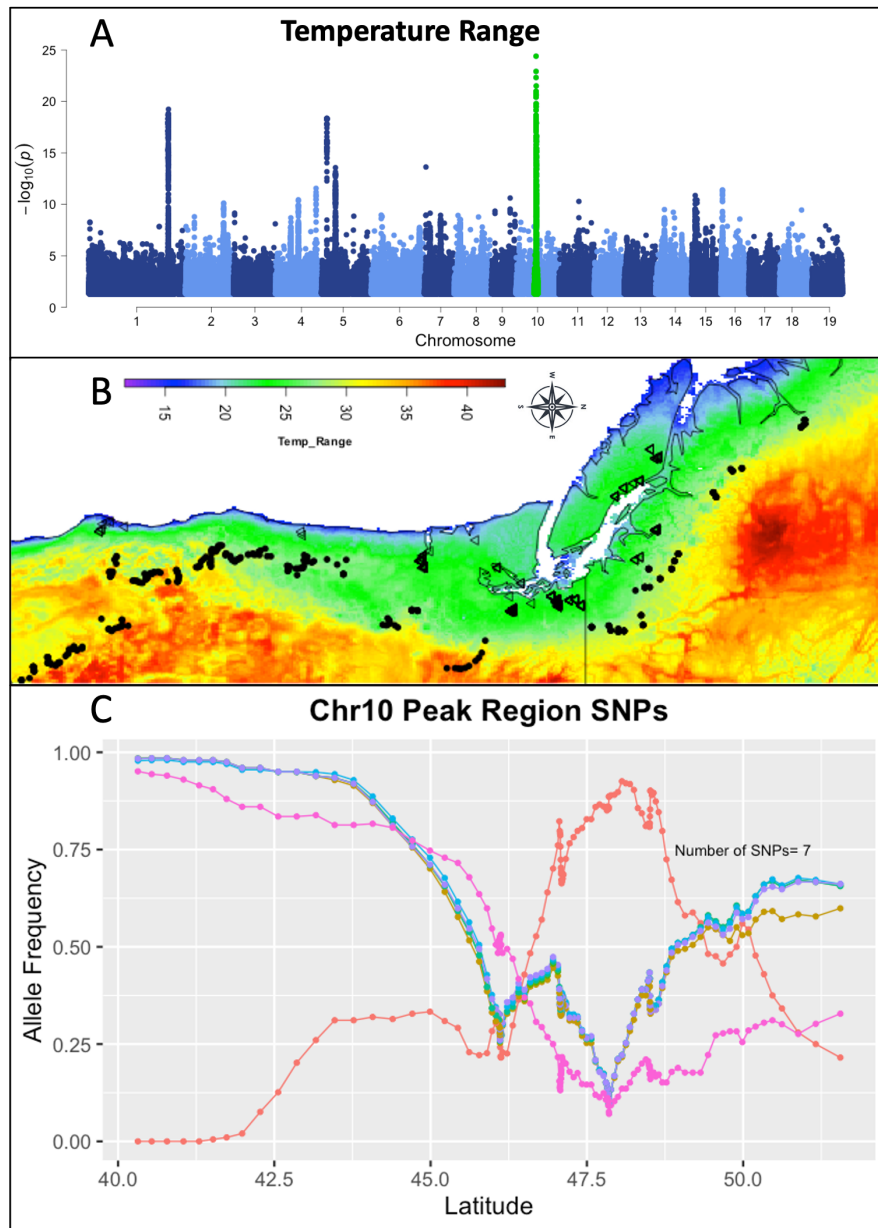
Supplementary Figure 12: Precipitation



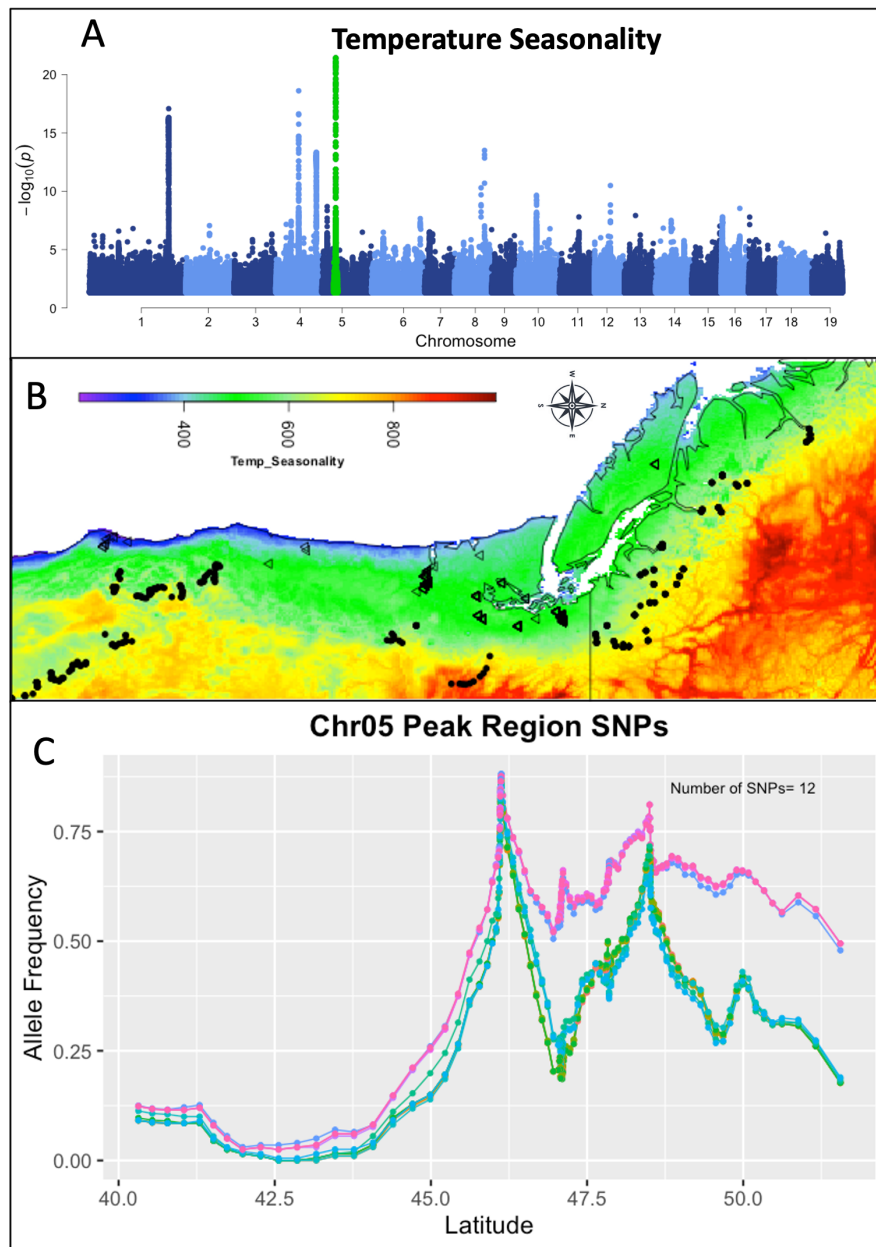
Supplementary Figure 13: Temp of the Coldest Quarter



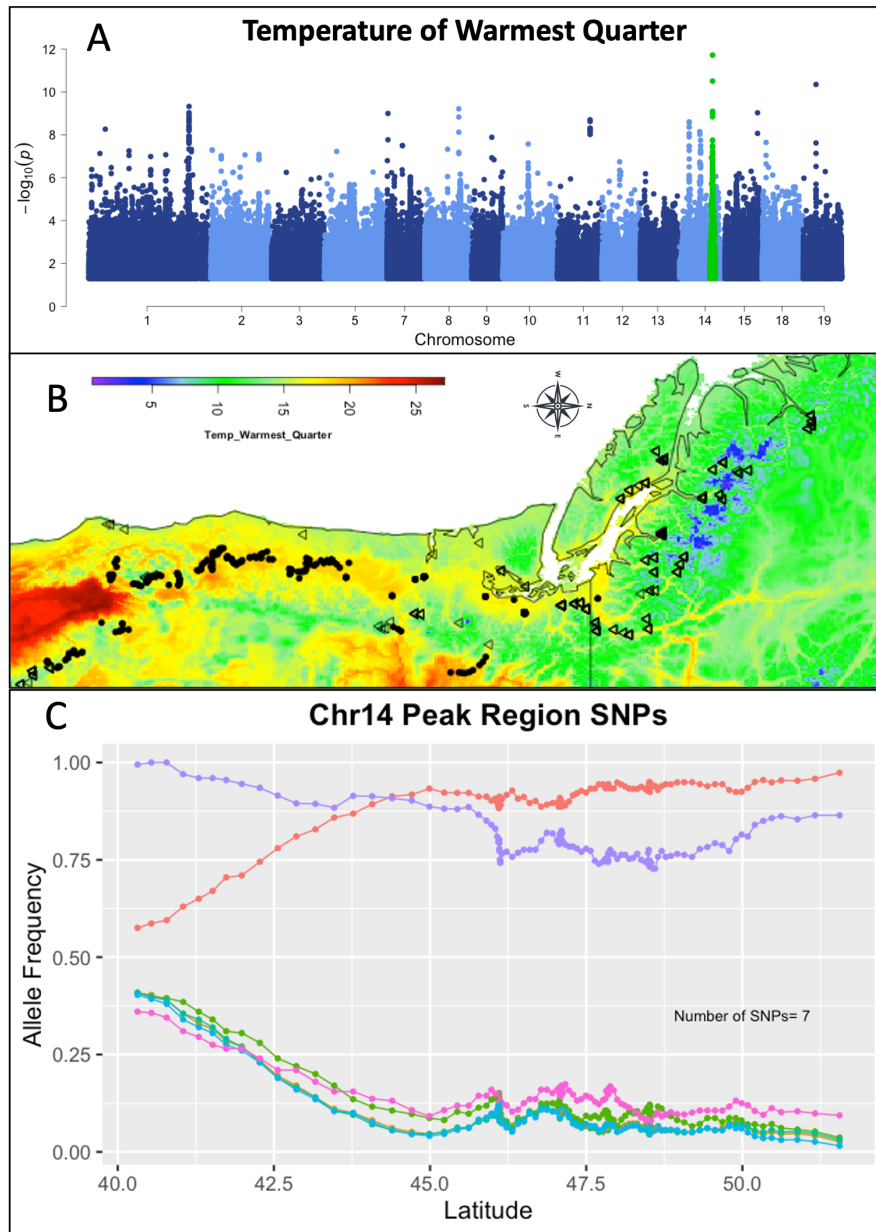
Supplementary Figure 14: Temp of the Driest Quarter



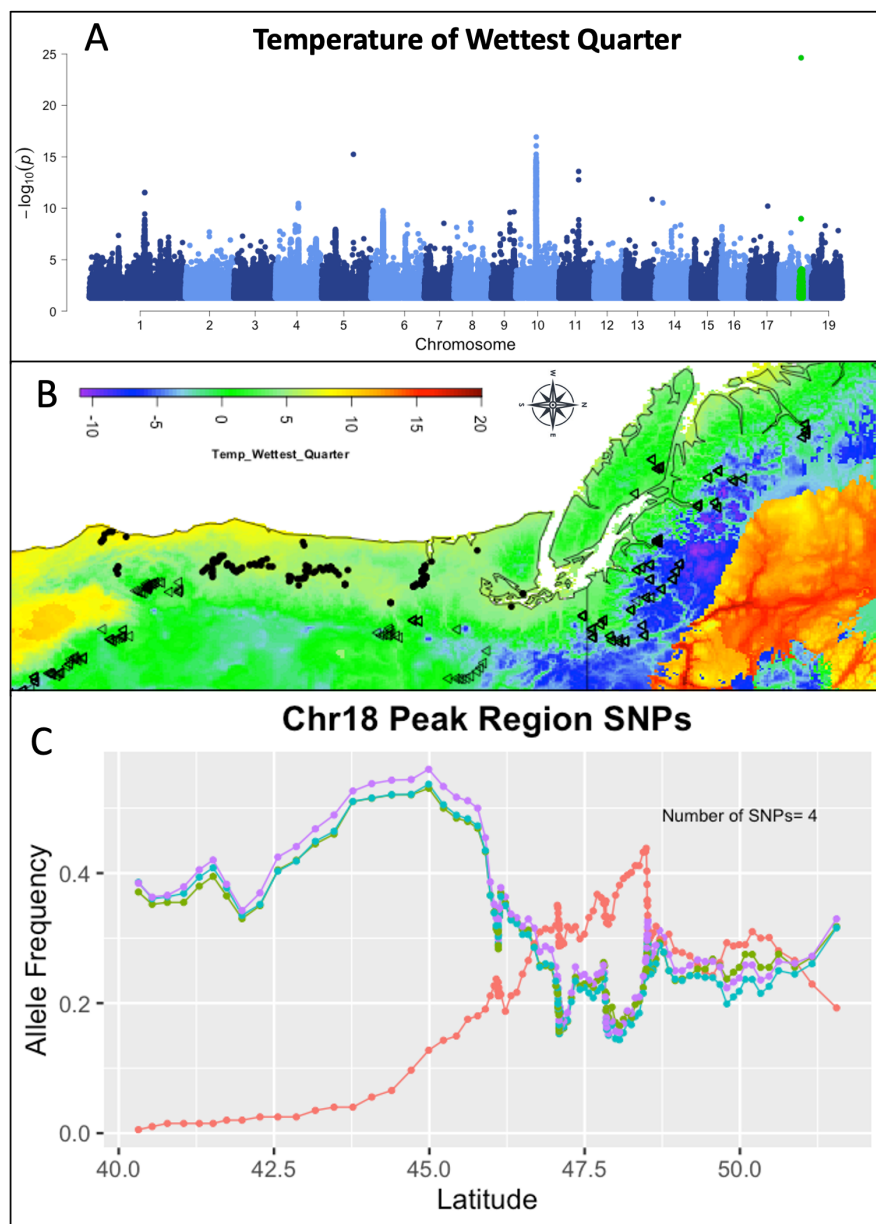
Supplementary Figure 15: Temp Range



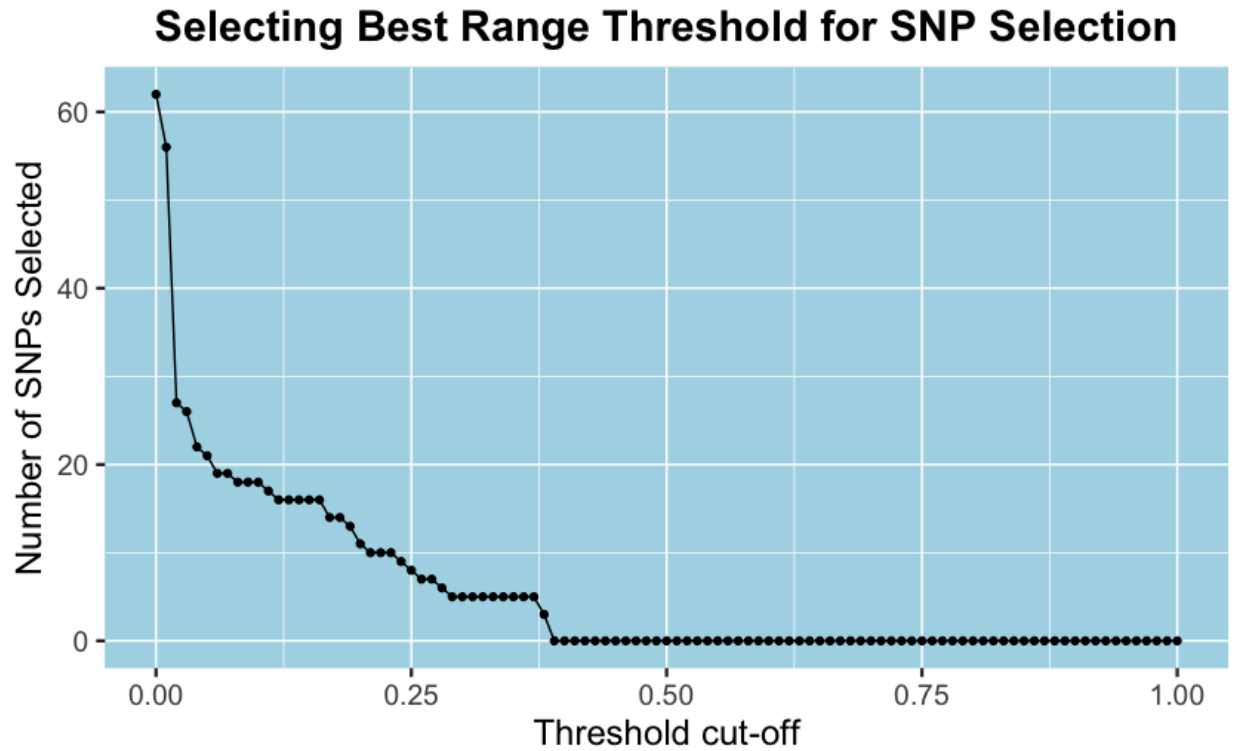
Supplementary Figure 16: Temp Seasonality



Supplementary Figure 17: Temp of the Warmest Quarter



Supplementary Figure 18: Temp of the Wettest Quarter



Supplementary Figure 19: Chromosome 14 peak region thresholding. This plot is used to select variants in a peak region for visualization in panel C of the climate GWAS figures (i.e., S1-S18). The range threshold is the difference between the maximum and minimum allele frequency of a SNP across the entire species range. We are interested in alleles that show the highest levels of variation. The stepwise pattern observed likely exists due to SNPs in LD with each other. Our goal is to select the LD block containing the SNPs with greatest variation, so in this case a threshold of 0.27 was used and resulted in visualization of 7 SNPs (see S17 panel C).

Vita

Cai W. John is a statistical geneticist currently graduating with a Ph.D. in Data Science and Engineering from the University of Tennessee Knoxville. Prior to his doctoral work, he worked in human genetics striving to characterize sub-types of Alzheimer's disease using brain image data and RNA expression.

With a background in physics, Cai's combination of mathematical literacy and computational confidence allow him to rapidly integrate into new problem contexts across biology. His current interests include mapping casual variants to druggable targets in human systems, increased stress tolerance by supplementation of metabolic precursors in mammalian contexts, and cancer genomics.