UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Use of latent factor mixed model to analyze allele climate association in a panel of ICARDA durum wheat accessions

Melvis Emade Ngeme-Ndie

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

Confidential

SUPERVISOR :
Prof. dr. Christel FAES

SUPERVISOR :
Mr. Zakaria KEHEL

2017
2018

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics

Master's thesis

Use of latent factor mixed model to analyze allele climate association in a panel of ICARDA durum wheat accessions

Melvis Emade Ngeme-Ndie

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

Confidential

SUPERVISOR :
Prof. dr. Christel FAES

SUPERVISOR :
Mr. Zakaria KEHEL

Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank my thesis advisor Prof. dr. Christel FAES of the Interuniversity Institute for Biostatistics and statistical Bioinformatics, University of Hasselt. I'm immensely grateful for her relentless and sacrificial guidance through out the life of the current study. She consistently allowed this study to be my own work, but steered me in the appropriate direction whenever we had discussion meetings. I would also express my sincere gratitude to Dr. Zakaria KEHEL of ICARDA organization in Morocco for encouragement and support through out my internship and life of the current study. He was extremely supportive in all feasible domains during my internship in Morocco.

Abstract

Cluster of loci were identified and selected and functional information such as correlation between them and ecological gradients were retrieved. Techniques for addressing allele-climate association analysis required the aptness to unravel likely environmental variable effects from confounding effects tabled by population ancestry. Recommendations for analysis of genome-wide datasets propounded mechanisms which were computationally fast and efficient on inference. We proffered a routine based on population genetics, ecological modeling, and statistical learning procedures to identify and select genomes which were adaptable locally. This method was latent factor mixed models.

Latent factor mixed model was implemented in the LEA package in R statistical software. This computationally efficient and swift strategy diagnosed associations between ecological variables and clusters of loci on specific chromosomes while simultaneously performing inference on underlying population stratas, visualized by Manhattan graphical tool. We then introduced this model to durum wheat genetic and environmental data sets, identifying multiple loci that exhibited strong correlations with climatic gradients.

KEYWORDS: *Local Adaptation, Genome Scans, Environmental Gradients, Population Structures, Latent Factor Mixed Models.*

1 Introduction

1.1 Background

The International Center for Agricultural Research in the Dry Areas (ICARDA) gene bank holds more than 140,000 accessions of barley, durum wheat, food legumes (lentil,chickpea and faba bean) and feed legumes (vetch and grass pea), composed mainly by landraces and wild species.

ICARDA gene bank is a place where significant quantities of genetic materials are collected, stored, classified and processed. This is to preserve genetic diversity for outsourcing for research and plant breeding so as to sustain food productivity, mitigate poverty and overcome hunger for a rapidly growing world population. Such genetic resources also allow for adaptation to deforestation, desertification and environmental uncertainties. Accessions refer to a distinct, uniquely identifiable sample of seeds, which represent cultivated varieties selected by humans or a population, and stored for conservation and use. The ICARDA gene bank possesses both landraces and wild accessions of cereals, legumes and forages, which are collected from dryland areas around the world. The wild species and landraces are relevant to breeders as they possess resistant genes for diseases and favorable alleles for adaptation to abiotic stress. They also hold alleles for nutritional and quality attributes.

Durum wheat (Triticum turgidum) is the only tetraploid (four sets of chromosomes) and contains the hardest shell of all the different wheat types. It was artificially bred from the Emmer wheat (Triticum dicoccum), which is a member of the annual grass wheat family, because of its very high micro-nutritional contents. It is grown in the spring in the Mediterranean area, Canada and along the Great Northern Plains in the US. Durum wheat is used to make most raised and flat breads, couscous, dried pasta, as well as semolina flour pasta. Durum wheat is sold at 20% to 30% higher prices than other common wheats because it is preferentially used by the food industry to produce high value foods. It is an important staple food, with about 35 % of the human population consuming it.

Durum wheats, which conveniently fit themselves to the prevailing environmental conditions, tend to survive and produce more offsprings. Their abilities to adapt well to their environment is dependent on the differences in genetic traits among the population, for characteristics which are considered to have influence on its fitness (Darwin 1859; Williams 1966). This postulation is regarded as the main process which brings about evolution, conservation, and global change

in biology (Joost et al. 2007; Manel et al. 2010; Barrett and Hoekstra 2011; Jay et al. 2012; Schoville et al. 2012).

Durum wheat contains many traits to adapt to several environmental and commercial uncertainties, such as disease resistance, heat, cold temperatures, drought and salinity tolerances. Plant breeders make use of genetic material of plants in the gene bank to develop superior collection of cultivated varieties. This is achieved through gene mining and new genetic diversity techniques, such as Focused Identification of Germplasm Strategy (FIGS), which improves the resilience and building unit of these cultivated varieties, fortifying them for future environmental and commercial uncertainties. FIGS is essential in identifying adaptive traits during breeding programs.

(a) Emmer wheat (b) Durum Wheat

Figure 1: *Wheats representation from Isterra-seeds.com*

Knowledge of the presence of important traits, alleles and genes are instrumental in plant accession duplications and seed stock maintenance decision making. DartSeq markers (https://www.diversityarrays.com/index.php/technology-and-resources/dartseq/), which are DNA fragments associated with particular locations of the genome assay genome polymorphisms. Information derived from DNA polymorphism are vital. They can be used to classify and explain population differentiation. Population differentiation is brought about by gene diversity and gene flow estimates. Also, the information serves as a guide in transferring superior quality traits such as biotic resilience, which often results in genetic variation. Marker-assisted selection makes use of this information to develop germplasm resources (http://www.isaaa.org/resources/publications/pocketk/19/default.asp).

Some of these identified loci show correlation with their environment. A study carried by Joost et al. (2007) reported that associations between loci and environmental variables can be tested by using regression models. Other studies, based on humans (Hancock et al.,2008; Fumangalli

et al.,2011; Frichot et al.,2013), and on loblolly pines (Eckert et al. ,2010), agreed with the use of regression models in evaluating associations between identified loci and climatic gradients. The study carried out by Jones et al. (2012) reported the use of ecological correlation methods when marine and sticklebacks were compared, to investigate the association between identified loci with the habitat.

In addition, these different suggested methods in testing for associations between detected loci and the environment were adopted by use of models. Some of these models do account for correction of confounding effects due to population structure or isolation-by-distance, whereas others do not. These models were viewed in different categories. A first category, applicable to continuous populations or populations characterized by high gene flows (Joost et al.,2007), was the logistic regression model or simple mantel test. A drawback of using this model was the generation of vast numbers of false positive associations (Schoville et al.,2012, De Mita et al.,2013, Frichot et al., 2013). A second category of models were those statistically dependent on residual terms, which examined effects of ecological variables on allele frequencies after overcoming confounding effects due to geographic structure (demographic history or isolation-by-distance) (Grafen, 1989; Harvey and Pagel, 1991). A third category of methods which evaluated effects of unknown latent factors while appraising the relationship between allele frequencies and environmental gradients are those galvanized by Genome-wide association studies and mixed models (Yu et al., 2006; Frichot et al., 2013; Yoder et al., 2014). In mixed models, fixed effects modeled association between allele frequencies and known environmental variables, whereas unknown hidden factors resulted from population structures due to demographic history or isolation-by-distance. A study presented by Frichot et al. (2013) reported that this mixed model methodology is implemented in the software Latent Factor Mixed Model (LFMM 2.1). Another study carried out by De Villemereuil et al. (2014), nominated Latent Factor Mixed Model as the most reliable method in assaying multiple genome-scan methods.

The current study used Latent Factor Mixed Models (LFMMs) based on regression models to evaluate the association between allele frequencies and candidate environmental variables while extinguishing unknown underlying hidden factors.

1.2 Objective

The purpose of the current study was to appraise desirable associations between allele frequencies and environmental variables for a set of North African durum wheat landraces using latent factor mixed models. Specifically:

- To examine the correlation, covariance and variance structure of the data.
- To assess probable confounding effects which could impact results interpretation.
- To assay the association between allele frequencies and environmental variables after correction of latent factors.

2 Data Description

2.1 Genomic data

Genotypic data for durum wheat landraces were selected from the ICARDA gene bank. These species were collected from five different countries in North Africa, representing the North African durum wheat genetic diversity. A total of 919 species were collected, where 347, 34, 9, 183 and 346 were from Algeria, Egypt, Libya, Morocco and Tunisia respectively. Different accessions were collected from the same sites. Accessions had unique identification referred to as ICARDA_ID. Collection sites also had distinct identification known as SiteCode. These diploid species were recored in a genotypic matrix, as illustrated below.

$$\mathbf{Y} = \begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,8366} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,8366} \\ \vdots & \vdots & & \vdots \\ y_{919,1} & y_{919,2} & \cdots & y_{919,8366} \end{pmatrix}$$

The genomic data for this study contained 919 rows representing individual genotypes and 8366 columns tabling genomic positions or loci. Loci were referred to as single nucleotide polymorphisms (SNPs), where $y_{i,j}$ were i^{th} individual at j^{th} locus.

2.2 Climatic Data

The environmental predictors contained geographic coordinates (Longitude and latitude) of the sites from which different landraces were collected. In total, 55 environmental variables were considered in the current study as predictors illustrated below

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,55} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,55} \\ \vdots & \vdots & & \vdots \\ x_{919,1} & x_{919,2} & \cdots & x_{919,55} \end{pmatrix}$$

Amongst the 55 environmental gradients were 36 monthly climatic variables (precipitation, minimum and maximum temperatures) and 19 bioclimatic variables (Hijmans et al., 2005). The data was derived by interpolating monthly average climatic data from weather stations on a 30 arc-second resolution grid usually known as "1 km^2" resolution (International Journal of Climatology 25: 1965-1978). The 19 bioclimatic variables were generated from monthly temperatures and rainfall values, which aided as a support tool to develop more variables which were biolog-

ically more meaningful. They were adopted in some ecological modeling approaches as well as modeling of distribution of species. The bioclimatic variables demonstrated yearly inclination (e.g., mean annual temperature, annual precipitation), seasonality (e.g., annual range in temperature and precipitation), and extreme or constraint ecological attributes (e.g., temperature of the coldest and warmest month, and precipitation of wettest and driest quarters). A quarter is a three months period (1/4 of an annum). The table below showed a concised description of climatic variables considered in the current study.

Table 1: *Description of the climatic study variables.*

Variable	**Type**	**Description**
ICARDA_IG	Categorical	Identification of accessions held at ICARDA gene bank.
Predictor Variables		
tmin1 -tmin12	Continuous	Minimum temperatures from January to December (^{0}C).
tmax1-tmax12	Continuous	Maximum temperatures from January to December (^{0}C).
prec1-prec12	Continuous	Precipitation from January to December (mm).
bio1	Continuous	Annual Mean Temperature (^{0}C).
bio2	Continuous	Mean Diurnal Range (Mean of monthly (max temp - min temp))(^{0}C).
bio3	Continuous	Isothermality (bio2/bio7) (* 100) (^{0}C).
bio4	Continuous	Temperature Seasonality (standard deviation *100) (^{0}C).
bio5	Continuous	Max Temperature of Warmest Month (^{0}C).
bio6	Continuous	Min Temperature of Coldest Month (^{0}C).
bio7	Continuous	Temperature Annual Range (bio5-bio6) (^{0}C).
bio8	Continuous	Mean Temperature of Wettest Quarter (^{0}C).
bio9	Continuous	Mean Temperature of Driest Quarter (^{0}C).
bio10	Continuous	Mean Temperature of Warmest Quarter (^{0}C).
bio11	Continuous	Mean Temperature of Coldest Quarter (^{0}C).
bio12	Continuous	Annual Precipitation (mm).
bio13	Continuous	Precipitation of Wettest Month (mm).
bio14	Continuous	Precipitation of Driest Month (mm).
bio15	Continuous	Precipitation Seasonality (Coefficient of Variation) (mm).
bio16	Continuous	Precipitation of Wettest Quarter (mm).
bio17	Continuous	Precipitation of Driest Quarter (mm).
bio18	Continuous	Precipitation of Warmest Quarter (mm).
bio19	Continuous	Precipitation of Coldest Quarter (mm).

3 Methodology.

3.1 Exploratory Data Analysis.

3.1.1 Genetic exploratory analysis.

Motivation for performing exploratory analysis on genomic data were:

- to estimate the number of latent factors that will be used for allele-environment analysis.
- to investigate population ancestry.

Suggested number of latent factors.

To be able to have an insight of population ancestry and the number of latent factors to be used for further allele-environment analysis, principal component analysis was performed based on covariance matrix of the genotypic data. Genotypes were represented as $\mathbf{Y}$ = 919 x 8366 matrix, where 919 individuals were the rows and 8366 SNPs were the columns. $\mathbf{Y^T}$was the notation used for the transpose of the matrix $\mathbf{Y}$. Matrix $\mathbf{Y}$ was mean-centered by subtracting the per column (SNPs) average from each column, so that means of each j^{th} column were zero. Principal component analysis was dependent on evaluating eigenvectors of 919 x 919 covariance matrix $\mathbf{\Sigma = YY^T}$.

The decomposition was done using eigen decomposition approach, where the covariance matrix denoted Σ was initially computed. Eigen decomposition was performed such that $\mathbf{\Sigma = U\Lambda U^T}$, where $\mathbf{U}$ was a 919 x k matrix ($\mathbf{U}^T\mathbf{U}$=$\mathbf{I}$), with columns that represented eigenvectors of $\mathbf{YY^T}$, $\mathbf{D}$ a k x k diagonal matrix of singular values (square root of the eigenvalues of $\mathbf{YY^T}$ and $\mathbf{Y^TY}$) and $\mathbf{V}$ was a 8366 x k matrix ($\mathbf{V^TV}$=$\mathbf{I}$) of the eigenvectors $\mathbf{Y^TY}$, where k is the matrix rank, dig(Λ)= $\lambda_1, \ldots, \lambda_k$ = diag(D^2) were the eigenvalues, $\mathbf{U}$ is the matrix of eigenvectors $\mathbf{YY^T}$. Λ and $\mathbf{D}$ were similar (having same eigenvalues, same number of independent eigenvectors but potentially different eigenvectors) but not same (same eigenvalues, same number of independent eigenvectors and probably same eigenvectors). The principal components $\mathbf{P}$ of the genomic data were given by the projection of the data onto the eigenvectors $\mathbf{P}$=$\mathbf{YV}$=$\mathbf{UD}$. PCA for genomic data was computed using pca program in LEA package in R statistical software, which outputted eigenvectors $\mathbf{U}$ as principal components without being weighted by singular values $\mathbf{D}$, which led to principal components with different scales. Since covariance was scaled by a factor $\frac{1}{\mathbf{N}-1}$, then singular values were scaled by a factor of $\frac{1}{\sqrt{\mathbf{N\text{-}1}}}$ for relevance of interpretation of singular

values $\mathbf{D}$ as square-root of eigenvalues of scaled covariance $\frac{1}{\mathbf{N}-1}\mathbf{YY^T}$. Differences in these scales did not affect interpretation of principal components when making inference on population structure for genomic data. The number of principal components retained was based on Cattell's scree criterion which involved visual inspection of scree plot. The benchmark trend in a scree plot was a steep curve, followed by a bend and then by a straight line. Eigenvalues coincided to population structure that lied on the steep curve. Principal components retained were those whose eigenvalues were on the steep curve before the bend and before the curve eventually leveled off to a straight line. The number of principal components retained suggested the number of ancestral populations that explained variation in the entire genomic population. The number of ancestral population recommended the number of latent factors that were adopted for the allele-environment analysis.

Population ancestry for the tabled latent factors.

To be able to trace genetic clusters, make inference about ancestral origin of durum wheat, screen genomes for signatures of natural selection and perform statistical corrections in genome-wide association studies (Pritchard et al. 2000b, Marchini et al, 2004; Price et al. 2006; Frichot et al. 2013), individual admixture coefficients were estimated using classical algorithms based on least-squares regression of alleles frequencies in genomic population (Robert and Hiorns, 1965; Cavalli-Sforza and Bodmer, 1971). Statistical inference of ancestry proportions for durum wheat genotypic data was implemented in R package LEA using sparse nonnegative least-square optimization function *(sNMF)*.

1. Modeling ancestry coefficients.

Genotypic data of durum wheat for 919 individuals at 8366 loci. The data was stored in genotypic matrix **(Y)**, where each entry was the number of derived allele at j^{th} locus for i^{th} individual. Autosomes in diploid durum wheat, derived number of alleles at j^{th} locus was 0, 1 or 2. Three bit of information were used for encoding each 0, 1 or 2 value as indicators of heterozygote or homozygote locus. That is value 0 was encoded as 1 0 0, 1 as 0 1 0 and 2 as 0 0 1. Binary coding usage recommended entries summed up to 8366 for each row of transformed data matrix in the implemented sparse nonnegative matrix factorization procedure. Admixture model expected that genotypic data originated from an admixture of K ancestral populations (from K principal component retained from principal component analysis of genomic data). Given K populations, probability that i^{th} individual carried j^{th} derived alleles at l^{th} locus was

written as

$$p_{il}(j) = \sum_{k=1}^{K} q_{ik} g_{kl}(j), \qquad j = 0, 1, 2 \tag{1}$$

where q_{ik} represented fraction of i^{th} individual genotype that emanated from K ancestral population and $g_{kl}(j)$ represented the homozygote (j = 0, 2) or heterozygote (j=1) frequency at l^{th} locus in population K. Based on the binary coding, equation 1 was written as

$$P = QG \tag{2}$$

where $P = (p_{il})$ was a 919 x 3.8366 matrix, $Q = (q_{ik})$ was 919 X 9 matrix, and $G = g_{kl}(j)$ was a 9 x 3.8366 matrix. Q matrix recorded ancestry proportions for each individual in the sample.

2. Least square estimates of ancestry proportions.

Inference of ancestry coefficients was implemented by least-square (LS) optimization algorithms (Engelhardt and Stephens, 2010). Estimates of **Q** and **G** matrices were derived after minimizing the least squares criterion

$$\mathbf{LS}(\mathbf{Q}, \mathbf{G}) = ||\mathbf{Y} - \mathbf{QG}||_F^2 \tag{3}$$

where $||\mathbf{M}||_F$ was the Frobenium norm of a matrix M (Berry et al. 2007). Least-square solutions were given by singular value decomposition of the matrix **Y**, when no constraints were imposed on **Q** and **G**, thus the resulted matrices **Q** and **G** contained scores and loadings of PCA. Ancestry coefficients were derived from matrices **Q** and **G** constrained with nonnegative entries such that

$$\sum_{k=1}^{K} q_{ik} = 1, \qquad \sum_{j=0}^{2} g_{kl}(j) = 1 \tag{4}$$

Estimation of ancestry coefficients and genotypic frequencies was equivalent to performing nonnegative matrix factorization of the genomic data matrix **Y** based on the constraints on equation 3. In sNMF, estimates of **Q** and **G** were computed using alternating nonnegativity-constrained least squares (ANLS) algorithm (Berry et al. 2007; Kim and Park, 2011). The least-square algorithm was divided into two phases. The 1st phase was characterized by assigning nonnegative values to matrix **Q's** entries, whose iterated cycles continued until convergence. Computation of nonnegative matrix **G** was also a feature of phase 1 which minimized the quantity

$$\mathbf{LS_1}(\mathbf{G}) = ||\mathbf{Y} - \mathbf{QG}||_F^2, \qquad \mathbf{G} \geq 0 \tag{5}$$

Subsequent to suggested solutions for linear regression problems, matrix **G** was derived by equalizing all nonnegative entries to zero. The resulted solution was then normalized so that its entries satisfied equation 4. Second phase of the cycle was marked by computation of a

nonnegative matrix $\mathbf{Q}$ which minimized the quantity

$$\mathbf{LS_2(Q)} = ||\begin{pmatrix} \mathbf{G^T} \\ \sqrt{\alpha} e_1 \times K \end{pmatrix} \mathbf{Q} - \begin{pmatrix} \mathbf{Y^T} \\ 0_1 \times n \end{pmatrix}||_F^2 \tag{6}$$

where $e_1 \times$K was a row vector that contained all entries equal to 1, $0_{1\times n}$ was a vector of length 919 with all entries equal to 0, α was a nonnegative regularization parameter. This minimization question was settled based on the block principal pivoting approach reported by Kim and Park (2011). The established solution, $\mathbf{Q}$, was then normalized which ensured row entries summed up to 1. Termination of iterations was basically dependent on a stationarity criterion recommended from the Karush–Kuhn–Tucker conditions (Kim and Park 2011) and on when the corresponding difference between two consecutive values of the criterion were less than a tolerance threshold of $\varepsilon = 10^{-4}$.

For $\alpha > 0$, the approach resulted in performing sNMF for the data matrix $\mathbf{Y}$. Values of $\alpha > 0$ were tested, because variances of $\mathbf{Q}$ and $\mathbf{G}$ approximated for smaller data sets could be diminished, insignificant approximations could be forced to zero, and numerical behavior of the ANLS minimization procedure could be improved. Time spent to access memory was optimized practically by the programming features used in sNMF.

3. Cross entropy criterion.

To be able to assess the quality (error) of ancestral estimation, a cross validation procedure was implemented which calculated the cross-entropy criterion. The algorithm depended on evaluating cross-entropy criterion which was a value based on forecast for entire genomic data and on disguised genotypes from sparse nonnegative matrix factorization (sNMF) output (Wold 1978; Eastment and Krzanowski 1982). The genomic data was split into two portions, training and test sets by the algorithm. The test set was created by randomly choosing and highlighting certain proportion (5% in the current study) of the entire genomic data and flagging as missing values. The resulted probabilities for the claoked entries were assessed based on the sNMF outputs retrieved from training sets illustrated by the subsequent formula

$$p_{il}^{pred}(j) = \sum_{k=1}^{K} q_{ik} g_{kl}(j), \qquad j = 0, 1, 2 \tag{7}$$

Comparison was made between forecasted values and cloaked values, $\mathbf{Y_{il}}$, by evaluating the mean of $-logp_{il}^{pred}(y_{il})$ on every SNPs in the test data set. Statistically, estimates of cross-entropy criterion were provided by the quantity on equation (8) following,

$$H(p^{sample}, p^{pred}) = \sum_{j=0}^{2} p^{sample}(j) - logp_{il}^{pred}(j), \qquad j = 0, 1, 2 \tag{8}$$

Cross-entropy between p^{sample} and p^{pred} was represented by equation (8). Number of ancestral

populations (K) and regularization parameter (α) were chosen with the goal of minimizing the cross-entropy criterion. The cross-entropy criterion had standard error of order $\frac{1}{\sqrt{n_L}}$, where n_L represented the number of masked genotypes. The yardstick was cross-entropy estimation for entire genomic data being always lower than that for disguised genotypic data. A smaller value of cross-entropy meant a better run in terms of strength of forecast. The criterion was supportive in comparing and highlighting the best run in the cross-entropy estimation for cloaked data from sNMF output.

3.1.2 Climatic explanatory analysis

Objectives of performing explanatory analysis on climatic data were:

- to examine variation pattern in the data.
- to assess the correlation structure of data.
- to evaluate potential confounding factors which could have influence on results interpretation.
- to select variables which will be used for further analysis.
- to perform descriptive statistics on selected climatic variables.
- to rearrange selected climatic variables by country.

Variance structure for climatic data

To have an understanding of summary variability pattern across the data, principal component analysis (PCA) was performed on the set of correlation matrix of predictors (climatic variables). It was a dimension reduction algorithm for climatic variables, by extracting variables of low dimensions in form of components from a high dimensional data set, with goal to capture as much information as feasible. This was implemented using the *princomp function* in R. The output of analysis was proportion of variance explained by each component, shown by use of scree plot graphical tool. Number of principal components retained was deduced by the number of components which were before the point where the curve's slope leveled off clearly (elbow), since these components tend to explain as much variation as the original variables. PCA had advantages in that, visualization became more meaningful with fewer variables. PCA supported in making predictors independent and avoid correlation problem between variables

(multicollinearity problem). It also visaed interpretation of variables using two-dimensional score plot.

Correlation structure for climatic data

To have insight on pattern of correlation for climatic variables, Pearson correlation was appraised between environmental predictors using *cor function* in R statistical software. The results of evaluation was correlation matrix visualized with correlogram graphical tool.

To be able to assess feasible climatic confounding effects, very highly correlated environmental gradients (pairs of variables whose correlation coefficient magnitude were between 0.9 and 1.0), were appraised using *function findcorrelation* from Caret package in R statistical software. This was because, the closer the correlation between two variables were close to 1, the more related their behavior and the more correlated one is with respect to the other. Extremely high correlation between climatic variables was referred to as multicollinearity. This tabled impact on result interpretation, since precision of established coefficients were reduced, which intend weakened statistical power of analysis. The p values might not be trusted in examining statistical significance of climatic variables. Model justification became a problem if many p values were not statistically significant. Estimated coefficients also became extremely sensitive to trivial perturbations in the analysis, since approximated coefficients could vary widely depending on which other predictors were present in the analysis. This suggested difficulties in specifying the correct model. Therefore violation in establishing appropriate relationship between allele-frequencies and each checklisted predictor.

Variable selection for climatic data

Presence of very highly correlated predictors in climatic data resulted in difficulties in appropriate model specification, together with model justification if the majority of predictor p values did not appear to be significant. These portrayed the need to catalog predictors correlated below correlation coefficient magnitude threshold of 0.90. Variables selection for subsequent analysis was based on the high correlation filter approach. If predictors were very highly correlated with one another, any one of these predictors were used as proxies for all others. Pairs of predictors with correlation coefficients magnitude higher than the threshold of 0.90 were reduced to only 1, by comparing their means and retaining the predictor with higher mean. Predictors which were correlated at a correlation coefficient magnitude threshold of 0.90 were evaluated using the *function findcorrelation* and filtered out. Results of the variables correlated below the correlation coefficient magnitude threshold of 0.90 were observed using correlogram graphical tool. These

variables were then used for further analysis. High correlation filter algorithm had advantages that, it did not rely on any machine learning algorithm. It very extremely computationally easy to adopt.

Summary statistics on selected climatic variables

Summary statistics such as minimum values, maximum values, means, medians and standard deviations were appraised for the selected climatic variables.

Cluster analysis for climatic data

To have insight on whether these selected climatic variables were rearranged to form any pattern based on their country of collection, clustering analysis was performed using the K-means clustering approach. This approach assigned observations to 5 clusters randomly and computed centroids for each clusters. Five because the study was interested in organizing the climatic variables into 5 countries namely Algeria, Egypt, Libya, Morocco and Tunisia. The algorithm iterated through reassigning data points to clusters whose centroids were closest and computed new centroids for each clusters. The iterations process continued until the within variance cluster disparity seemed to be null.

3.2 Allele-environment association analysis

3.2.1 Latent Factor Mixed Models

Mixed model algorithm used in gene-environment association study, examined adaptation of SNPs to environmental gradients in five Northern African countries inbred lines of the model plant species durum wheat. Durum wheat was genotyped using DartSeq marker containing 8366 SNPs. Preliminary analyses on exploration of both climatic and genotypic data sets were performed. Analyses on climatic data set extracted 19 climatic variables from an initial total of 55, using high correlation filter procedure. Analyses on genotypic data was reliant on population ancestry using ancestry estimation program sNMF (Frichot et al., 2014), which tabled nine clusters (nine principal components) which described the desirable genetic variation in durum wheat. In succeeding analyses of testing for association between alleles and environmental gradients, latent factor mixed model was adopted. These nine clusters were then used as number of latent factors to perform genome scans for selection.

Latent Factor Mixed Models (LFMMs) were statistical linear regression models with genotypic

matrix as response variables and environmental gradients for a set of individuals as explanatory variables, that struggled to examine unobserved latent factors that modeled confounding effects due to genetic variations such as population ancestry, not explained by adaption to their ecological inhabitat. When these confounding effects were extinguished , desirable association between allele frequencies and environmental variables were elucidated as evidence for selection at particular loci.

For allele frequencies, (Y_{ij}) and a set of environmental variables, (X_i), we had standard tests being based on regression models as below:

$$Y_{ij} = \lambda_j + \beta_j^T X_i + \varepsilon_{ij} \tag{9}$$

where Y_{ij} were allele frequencies for i^{th} individual at j^{th} locus, β_j were environmental effects and ε_{ij} independent residuals, i= 1,..., 919, j=1,...,8366.

Principal component analysis was related to factor analysis through maximum likelihood approximations (Tipping and Bishop, 1999; Engelhardt and Strphens, 2010)

$$Y_{ij} = \lambda_j + C_i^T D_j + \varepsilon_{ij} \tag{10}$$

where Y_{ij} were allele frequencies for i^{th} individual at j^{th}locus, $C_i \sim \mathrm{N}(0, \sigma_{C_i})$ and $D_j \sim \mathrm{N}(0, 1)$ respectively with K (number of retained principal components) dimensions suggested by scores and loadings of principal components, ε_{ij} were uncorrelated residuals suggested by K dimensions ($\mathrm{K} \leq \mathrm{n}$) with i= 1,..., 919, j=1,...,8366.

Latent factor mixed models followed from equations (9) and (10) as

$$Y_{ij} = \lambda_j + \beta_j^T X_i + C_i^T D_j + \varepsilon_{ij} \tag{11}$$

where Y_{ij} were allele frequencies for i^{th} individual at j^{th} locus, β_j were environmental effects, C_i and D_j scores and their loadings respectively, ε_{ij} independent residuals, i= 1,..., 919, j=1,...,8366.

Latent factor mixed model considered in the current study was recommended as:

$$\begin{cases} Y_{ij}|C_i, D_j, \beta_j, \lambda_j, \sigma^2 \sim \mathrm{N}(U_{ij}, \sigma^2)^{I_{i,j}}, \quad where \quad U_{ij} = \lambda_j + \beta_j^T X_i + C_i^T D_j + \varepsilon_{ij}. \quad Likelihood. \\ \lambda_j|\sigma^2_{\lambda_j} \sim \mathrm{N}(0, \sigma^2_{\lambda_j}), \quad \beta_{jl}|\sigma^2_{\beta_l} \sim \mathrm{N}(0, \sigma^2_{\beta_l}), \quad C_i|\sigma^2_{C_i} \sim \mathrm{N}(0, \sigma^2_{C_i} I_K), \quad D_j \sim \mathrm{N}(0, I_K), \quad Prior \end{cases} \tag{12}$$

where:

- Y_{ij} the response variable, was a genotypic matrix of allele frequency entry records of i^{th} subject at j^{th} locus, $i = 1, ..., 919$, $j = 1, ..., 8366$, 919 and 8366 being the total sample size and the total number of loci respectively.

- $\sigma^2_{\lambda_j} \sim \Gamma^{-1}(1+\frac{L}{2}, \frac{1}{2}\sum_j \lambda_j^2 + 1)$, $\sigma^2_{\beta_l} \sim \Gamma^{-1}(1+\frac{L}{2}, \frac{1}{2}\sum_l \beta_{jl}^2 + 1)$, $\sigma^2_{C_i} \sim \Gamma^{-1}(\eta + \frac{nK}{2}, \frac{1}{2}\sum_i C_i^T C_i + \eta)$.
- $I_{i,j}$ an indicator variable equal zero when the genotype data were missing at locus j for individual i, and equal to 1 otherwise.
- λ_j was a locus-specific mean.
- β_j were vectors of regression coefficients.
- X_i were vectors of environmental covariates modeled as the fixed effects.
- C_i and D_j were scores and their loadings respectively. The matrix term $C_i^T D_j$ modeled that part of the genetic variation unaccounted for by ecological adaptation.
- $\varepsilon_{ij} \underset{\sim}{iid} \mathrm{N}(0, \sigma^2)$

In Bayesian system, unknown parameters were evaluated by posterior means. They were examined by use of Markov Chain Monte Carlo (MCMC) algorithm (Gilks et al., 1994) and generated samples from the full conditionals using the Gibbs sampler. Sampled averages were considered as the posterior means of the parameters of interest.

Genome scan for selection used five particular statistical models fitted using LFMMs as implemented in the R package LEA (Frichot at al. 2014). These five models represented five specific runs of the program with distinct initial values. Each model coincided with specific local optima of the likelihood function and were described by particular set of z-scores.

A Gibbs sampler algorithm was implemented to simultaneously estimate scores (C_i) and loadings (D_j), environmental effects (β_j), and means (λ_j), which were LFMMs parameters. A total of 10 000 Gibbs sampling iterations were run after a burn-in period of 5000 iterations. Model parameters were initialized to zero, that is $C_i = 0_{K,n}$, $D_j = 0_{K,p}$, $\beta_j = 0_{p,d}$, $\lambda_j = 0_{p,1}$. Residual variance, σ^2_{ij} parameter was updated at each iteration using the current residual variance. Other model parameters were updated through Gibbs sampling iterations. At the first iteration when model parameters were initialized, potential missing values in the genotypic matrix Y_{ij} were imputed for t=1, ...,10 000 as follows

$$Y_{ij} \leftarrow C_i^{(t-1)T} D_j^{(t-1)} + \beta_j^{t-1} X_i^{t-1}$$

Variance parameter σ^2_{ij} was then updated after that first iteration as follows

$$\sigma_{ij}^{2(t)} = Var(Y_{ij} - C_i^{(t-1)T} D_j^{(t-1)} - X_i^{(t-1)} \beta_j^{(t-1)})$$

Hyperparamters were then sampled as follows

$\sigma_{C_i}^{2(t)} \sim P(\sigma_{C_i}^2|C_i^{(t-1)},\eta), \qquad \sigma_{\beta_j}^{2(t)} \sim P(\sigma_{\beta_j}^2|{\beta_j}^{(t-1)}), \qquad \sigma_{\lambda_j}^{2(t)} \sim P(\sigma_{\lambda_j}^2|{\lambda_j}^{(t-1)})$

Samples from each locus were:

$\sigma_{\lambda_j}^{(t)} \sim P(\lambda_j|C_i^{(t-1)},D_j^{(t-1)},\beta_j^{(t-1)},\sigma_{\lambda_j}^{2(t)},\sigma_{ij}^{2(t)}), \quad \sigma_{\beta_l}^{(t)} \sim P(\beta_j|C_i^{(t-1)},D_j^{(t-1)},\sigma_{\lambda_j}^{(t)},\sigma_{\beta_1}^{2(t)},\cdots,\sigma_{\beta_l}^{2(t)},\sigma_{ij}^{2(t)})$

Samples for each i^{th} individual were:

$C_i^{(t)} \sim P(C_i|D_j^{(t-1)},\sigma_{\lambda_j}^{(t)},{\beta_j}^{(t)},\sigma_{C_i}^{2(t)},\sigma_{ij}^{2(t)})$

Samples for each j^{th}locus were:

$D_j^{(t)} \sim P(D_j|\sigma_{\lambda_j}^{(t)},\beta_j^{(t)},C_i^{(t)},\sigma_{ij}^{2(t)})$

Parameter estimates were appraised as follows:

$C_i = mean(C_i^{(5001)},\cdots,C_i^{(10000)}), \qquad D_j = mean(D_j^{(5001)},\cdots,D_j^{(10000)}),$

$\beta_i = mean(\beta_i^{(5001)},\cdots,\beta_i^{(10000)}), \qquad \lambda_j = mean(\lambda_j^{(5001)},\cdots,\lambda_j^{(10000)})$

Z-score was evaluated as follows:

$$Z_j = \frac{mean(\beta_j^{(5001)},\cdots,\beta_j^{(10000)})}{(Var(\beta_j^{(5001)},\cdots\beta_j^{(10000)}))^{1/2}}$$

Fitted latent factor mixed model evaluated the extend to which each SNPs was associated with each environmental gradient. While interest was in parameter estimates, primary interest was reliant in determining whether estimates were non-zero. Several algorithms for hypothesis testing are likelihood ratio statistics, wald statistics and score statistics. Z-scores statistics were used in the current study, since it does not require estimation of the parameters being tested.

In order to establish combined p-values adopted in testing association between SNPs and climatic gradients, which could portray collective significant results rather than series of non significant outcome, z-scores at each locus were combined for the five repetitions of latent factor mixed model by Stouffer method (Brown 1975, Whitlock 2005). Stouffer algorithm recommended each test assigned weights proportional to inverse of their squared standard error. Merit of combining z-scores using Stouffer method was the relative ease involved in introducing weights.

Median of z-scores substituted mean of z-scores at each locus. Bias established by the merge of five recognizable z-scores from the five model runs were revised by initiation of an inflation factor, which established the basis of genomic control. Inflation factors were constant values, λ , purposed to recycle test statistics ensuring inflation due to population ancestry and confounding factors were mitigated. The intention for rescaling approach was to customize test statistics which resulted to flat histograms for significant values. Significant values were appraised from chi-squared distribution with one degree of freedom. Their test statistics were nominated as

squared z-scores in the rest of the current study. For chi-squared tests, recycled statistics was $\frac{Z_l^2}{\lambda}$, where Z_l was score evaluated at l^{th} locus, with invariant degree of freedom for test statistics. This approach revised baseline hypothesis of $H_o : Z_l^2 = 1$ and customized it with a new null-hypothesis, $H_o : Z_l^2 = \lambda$, where λ was approximated from the data.

In testing for association between several SNPs with each climatic gradient, multiple testing comparisons were performed. Several tested null hypotheses portrayed significant outcome even if all the hypotheses were false. These were referred to as false discoveries. False discovery rate (FDR) was described as an expected proportion of false positives among a checklist of positive tests. Tests were calibrated at a specific level of significance to suggest largest possible catalog of loci at that level. FDRs were then controlled at an expected level of 10% using false discovery rate control approach examined by Benjamini-Hochberg, 1995. Each individual p value was compared to its Benjamini-Hochberg critical value, $q * (1 : L_j)/L_j$, where q was the chosen expected false discovery rate, $1 : L_j$ were the rankings and L_j was the total number of tests performed.

The cut-off for significance was established at the largest p value that had a lower Benjamini-Hochberg critical value (significant). All other p values smaller than this cut-off point were also tabled significant, even those whose p values were not less than their coincided Benjamini-Hochberg critical value.

Benjamini-Hochberg algorithm for controlling false discovery rates was preferred as it was less sensitive than other algorithms for controlling multiple comparison issues, since if number of tests were increased and the distribution of p values was same in the newly incremented tests as in the fundamental one, same proportion of significant results were recommended by Benjamini-Hochberg which was not the scenario for other methods. Test p values behaved as uniformly distributed random variables with corrected null hypotheses. Graphical techniques of displaying histograms of corrected p values were vital (Balding 2006). Results of the analyses were observed using graphical technique of Manhattan plots.

This model had the advantage that it efficiently estimated random effects due to population history when computing gene-environment correlations, though its limitation was in appropriately shortlisting number of latent factors that was used in running the LFMMs.

3.2.2 Statistical software

All the analyses performed in the current study were performed using R statistical software version 3.4.4, RStudio for windows and latex.

4 Results

4.1 Exploratory data analysis

4.1.1 Genetic exploratory analysis

Suggested number of latent factors

Results of principal component analysis on genotypic data were visualized by use of scree plot graphical tool as presented on Figure 2. Number of principal components retained was based on number of principal components on the steep curve before the bend and before the curve eventually leveled off, since estimates of eigenvalues presented measures of the amount of original total variance explained by each of the new derived variables (principal components), as sum of all eigenvalues equals sum of variances of SNPs. Scree plot illustrated only nine principal components whose eigenvalues were situated within the sharp decent before the bend, since those principal components at and after the bends had smaller comparable sizes. Nine principal components represented nine ancestral populations which explained variation in the entire genomic data and were suggested as the number of latent factor used for allele-environment analysis.

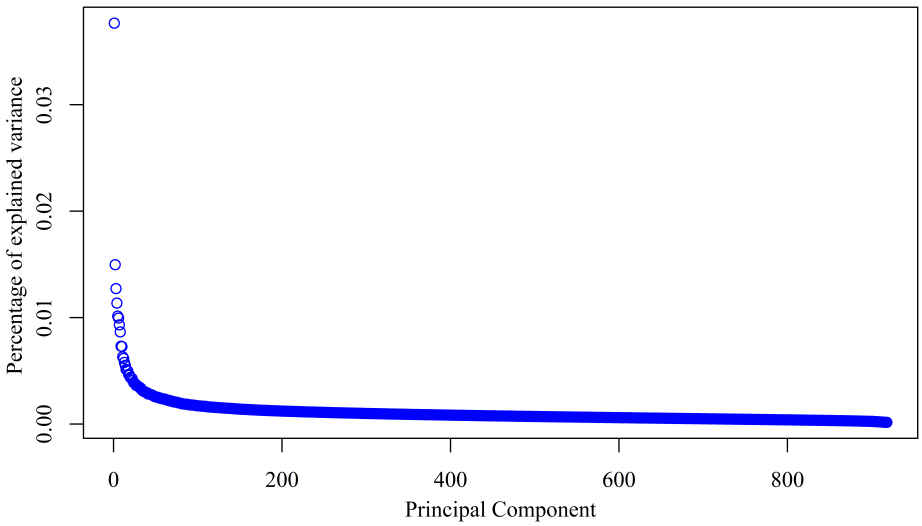

Figure 2: *Scree plot for percentage of variance explained by SNPs.*

q

Scores of principal components 1 and 2 for genomic data were plotted in a scatter plot shown on Figure 3. Such a scatter plot readily showed that there seemed to had been no grouping in genomic data, since individuals did not appear to be distinctly different in the score plot based on countries. There was no observable trend of variation of individuals through out genomic data. This presumed it was not feasible to classify individuals based on countries.

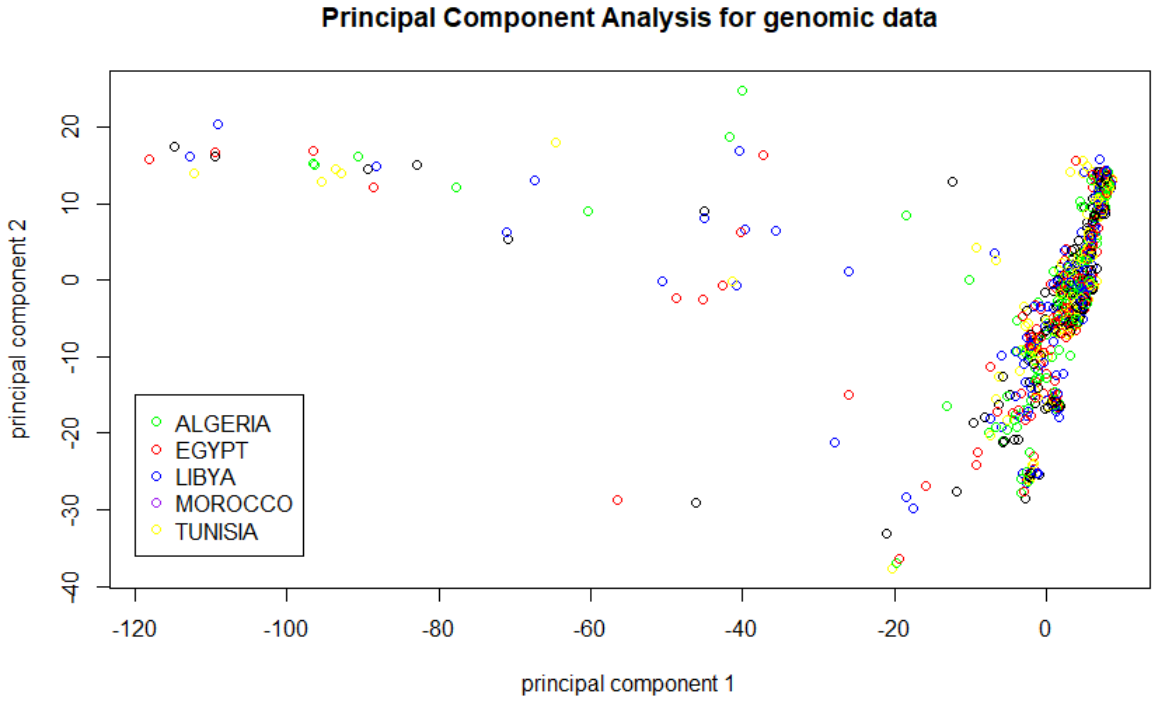

Figure 3: *Scatter plot of principal components scores for genomic data.*

Population ancestry for the tabled latent factors.

1. Modeling ancestry coefficients and least square estimates of ancestry proportions.

Outputs of estimated ancestry proportions based on least-squares procedure were **Q** and **G** matrices. The **Q**-matrix was an ancestry matrix that contained admixture coefficients, precisely recorded ancestry proportions for every individual in the sample. It was a 919×9 matrix, where 919 was the number of individuals in the study and 9 the number of ancestral populations. Graphical illustration of ancestry estimates obtained for 8366 genotypic data set of durum wheat was shown on Figure 4, which showed estimated ancestry coefficient using sNMF with K=9 ancestral population, a regularization parameter $\alpha = 10$, and a cross entropy $= 0.0905$. It was observed that individuals were not separated by their ancestral populations as there did

not seem to be any visible distinct cluster within the populations. Thus, we could ascertain that individuals seemed to have had same ancestor. Table 2 was a representation of how the individuals in the current study were allocated to the different clusters (nine ancestral populations). The **G**-Matrix which was the other output of the least-square algorithm was the genotypic allele frequency matrix used in the allele-environment analysis.

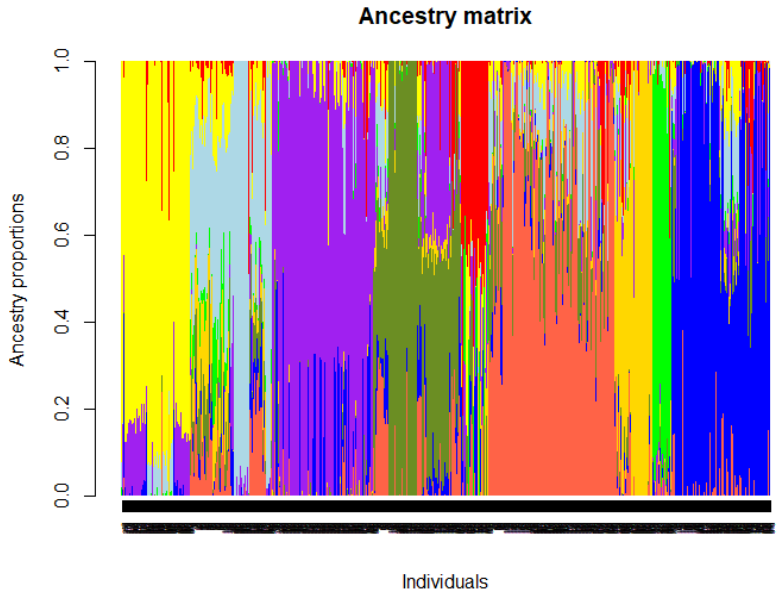

Figure 4: *Bar chart of ancestry coefficients for the best run and K = 9 populations.*

Table 2: *Clustering of individuals in K=9 ancestral population.*

Number of clusters	1	2	3	4	5	6	7	8	9
Number of Individuals	178	139	123	55	145	27	116	98	38

3. Cross entropy criterion.

To be able to select a value for K, sNMF for each value of K from 1 to 20 was projected. Figure 5, illustrated the cross-entropy value obtained for each value of K. K = 9 was the best value based on the cross-entropy criterion because the cross-entropy did not follow the trend of sharp decrease for K greater than 9. There appeared to be consecutive minimal increases and decreases after values of K greater than 9. The analyses were completed by the displayed Q-matrices associated with each run on Figure 5. The best run was shortlisted based on the run with the minimal cross-entropy after comparing the cross-entropy of all the runs. Table 3 showed that run 4 for K=9 ancestral population was the best run, since it had the lowest entropy

of 0.0905. The cross-entropy for the training set was 0.0832 which was lower than that for the tested set of 0.0924, hence met the benchmark of the criterion.

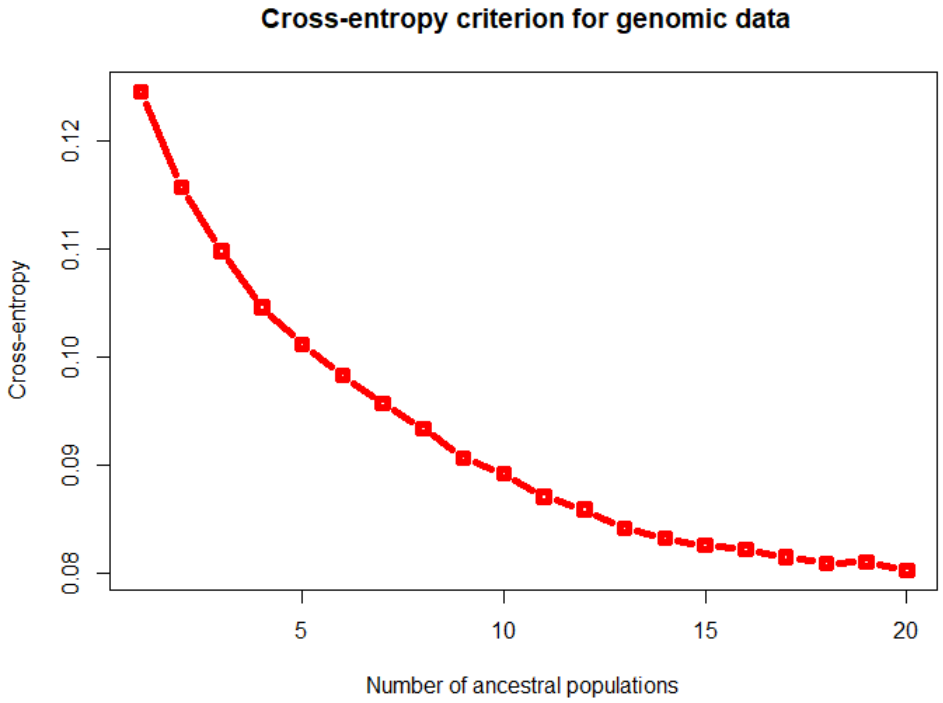

Figure 5: *Value of the cross-entropy criterion for 20 sNMF runs (genomic dataset).*

Table 3: *Cross-entropy evaluated with K=9 ancestral populations for 10 runs*

Number of runs	1	2	3	4	5	6	7	8	9	10
Cross entropy	0.0912	0.0932	0.0909	0.0905	0.0920	0.0918	0.0925	0.0934	0.0916	0.0930

4.1.2 Climatic explanatory analysis

Variance structure

Visual output of Principal component analysis on climatic data was presented on the scree plots on Figure 6. The left paneled scree plot depicted five principal components, since five principal components were visible before the curve's slope clearly leveled off (elbow). This plot illustrated that about 5 components explained nearly 95.6% of variance in the climatic data set. In order words, the 55 predictors were reduced to 5, without compromising on explained variance using principal component analysis.

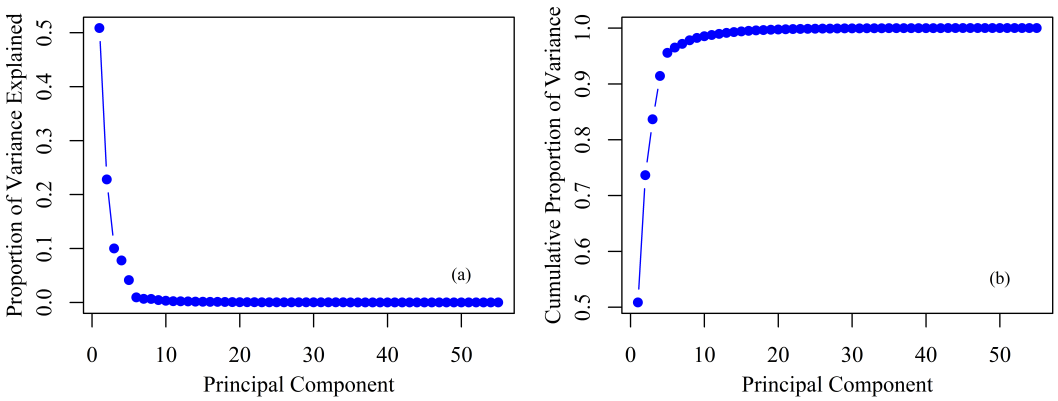

Figure 6: *Scree plots illustrating proportion and cumulative proportion of variance explained by climatic variables (predictors).*

The right paneled scree plot on Figure 6, showed confirmatory check on the number of components which explained the proportion of variance in the climatic data set. This plot showed that 5 components resulted in variance close to about 96%. The number of components selected were therefore 5 as illustrated in the appendix.

Retained principal components were interpreted reliant on the Pearson correlation appraised between the principal component scores and each of the 55 climatic variables by the *cor function* in R statistical software. The computed correlations were tabled on the appendix and interpretations followed subsequently with description of climatic variables on Table 1.

Principal component 1 (PC1)

Principal component 1 (PC1) was very highly negatively correlated with tmin3, tmin4, tmin5, tmax2, bio1 and bio11 with correlation coefficient of magnitudes -0.9016, -0.9204, -0.9288, -0.9041, -0.9676 and -0.9066 respectively. It was depicted that there seemed to have been strongly negatively correlation between PC1 and tmin1, tmin2, tmin6, tmin7, tmin8, tmin9, tmin10, tmin11, tmin12, tmax1, tmax3, tmax4, tmax5, tmax10, tmax11, tmax12, bio6, bio8 and bio10 with correlation coefficient of magnitudes -0.8012, -0.8468, -0.8800, -0.7762, -0.7807, -0.8561, -0.8916, -0.8600, -0.7953, -0.8791, -0.8780, -0.7914, -0.7506, -0.8926, 0.8855,-0.8590, -0.8012, -0.7487 and -0.7785 respectively. It was observable that PC1 was moderately negatively correlated with tmax6, tmax9 and bio9 with correlation coefficient of magnitudes of -0.5122, -0.6856 and -0.6437 respectively. There seemed to had almost been weakly negatively correlation between PC1 and tmax8, bio5, bio15 with correlation coefficient of magnitudes -0.3442, -0.3003 and -0.4951 respectively. It was deduced that there did not seem to be any linear relationship between PC1 and bio3,tmax7 with correlation coefficient of magnitudes of-0.1146 and

-0.2663. There seemed to have existed a strongly positively correlation between prec5, prec6, prec7, bio14, bio17, bio18 and PC1 with correlation coefficient of magnitude of 0.8457, 0.8595, 0.7541, 0.7561, 0.8210 and 0.7401 respectively. There was also moderately positively correlation between prec3, prec4, prec8, prec9, prec10, prec11, bio12 and PC1 with correlation coefficient of magnitudes 0.5771, 0.6783, 0.6756, 0.6457, 0.6026, 0.5236 and 0.6660 respectively. It was illustrated that there was weakly positively correlation between prec1, prec2, prec12, bio7, bio13, bio16, bio19 and PC1 with correlation coefficient of magnitudes 0.4827, 0.4914, 0.4250, 0.3111, 0.4580, 0.4741, and 0.4706 respectively. There did not seem to be any positively linear relationship between bio2, bio4 and PC1 with correlation coefficient of magnitudes 0.1633 and 0.2867 respectively.

Principal component 2 (PC2)

There were observed strongly positively correlations between prec1, prec3, prec12, bio13, bio16, bio19 and PC2 with correlation coefficient of magnitudes 0.7236, 0.5709, 0.7566, 0.7469, 0.7383 and 0.7317 respectively. There were depicted moderately positively correlation between tmin1, tmin12, prec2, prec10, prec11, bio12, bio15 and PC2 with correlation coefficient of magnitudes 0.5042, 0.5248, 0.6813, 0.5058, 0.6894, 0.5053, 0.5752, 0.6873 respectively. Deduced weakly positively correlation seemed to have existed between tmin2, tmin3, tmin4, tmin10, tmin11, prec4, bio11 with PC2 with correlation coefficient of magnitudes 0.4684, 0.3958, 0.3185, 0.3220, 0.4130, 0.4279 and 0.3226 respectively. No positively linear relationship appeared to have existed between tmin5, tmin6, timn7, tmin8, tmin9, tmax12, prec5, bio1 and PC2 with correlation coefficient of magnitudes 0.1406, 0.0693, 0.0667, 0.1631, 0.2447, 0.1443, 0.1878 and 0.0133 respectively. There appeared to have been a strongly negatively correlation between bio2, bio4, bio5, bio7 and PC2 with correlation coefficient of magnitudes -0.7033, -0.6890, -0.7384, -0.8284 respectively. There also seemed to be weakly negatively correlation between prec7, prec8, bio10, bio14, bio17 and PC2 with correlation coefficient of magnitudes -0.4603, -0.3310, -0.4613, -0.4564 and -0.3502 respectively. There appeared to be any depictable negatively linear relationship between prec6, prec9, bio3, bio8, bio9, bio18 and PC2 with correlation coefficient of magnitudes -0.2196, -0.1195, -0.1740, -0.1869, -0.2916 and -0.1726 respectively.

Principal component 3 (PC3)

There seemed to have been any positively linear relationship between tmin1 to tmin 9, bio 6, bio8 and PC3 with correlation coefficient of magnitudes 0.0151, 0.0225, 0.0280, 0.0217, 0.0610, 0.0523, 0.1099, 0.1217, 0.0690, 0.0078 and 0.1082 respectively. There appeared to have been any

negatively linear relationship between tmin10 to tmin12, tmax1 to tmax3, tmax11 to tmax12, prec5 to prec9, bio1, bio4, bio9, bio10, bio11, bio14, bio15, bio17, bio18 and PC3 with correlation coefficient of magnitudes -0.0003, -0.0319, -0.0244, -0.2693, -0.2665, -0.3448, -0.2976, -02932, -0.1714, -0.1254, -0.1000, -0.1076, -0.2234, -0.2163, -0.00921, -0.2464, -0.2643, -0.1469, -0.0995, -0.0939, -0.1194, -0.0969. PC3 appeared to have had weakly negatively correlation with tmax3, tmax4, tmax5, tmax10, prec1, prec4, prec10, prec11, prec12, bio3, bio7, bio12, bio13, bio16 and bio 19. They had correlation coefficient of magnitudes -0.3449, -0.3708, -0.3669, -0.3595, -0.4266, -0.4263, -0.4723, -0.3423, -0.4605, -0.3868, -0.3740, -0.4551, -0.4497, -0.4475, -0.4688 respectively. There were established moderately negatively correlation relationships between tmax6 to tmax9, prec2 to prec3, bio2, bio 5 and PC3 with correlation coefficients of magnitude -0.5020, -0.5389, -0.5528, -0.5480, -0.5047, -0.5004, -0.5154, -0.5415 respectively.

Principal component 4 (PC4)

There appeared to be any negatively linear relationship between tmin1 to tmin4, tmin10 to tmin12, tmax6 to tmax9, prec1, prec2, prec4, prec6 to prec8, prec10 to prec12, bio1, bio5 to bio7, bio12 to bio14, bio16 to bio19 with correlation coefficients of magnitudes -0.1080, -0.1388, -0.1075, -0.1889, -0.2109, -0.0663, -0.0475, -0.0739, -0.2243, -0.2325, -0.0957, -0.1178, -0.0277, -0.0091, -0.2201, -0.0696, -0.1353, -0.2009, -0.1535, -0.0659, -0.0971, -0.2211, -0.1074, -0.0816, -0.1301, -0.0457, -0.0708, -0.0851, -0.1697, -0.2485, -0.0723 respectively. It seemed there were weakly negatively correlation between tmin5, tmin6, tmin9, prec5, prec9, bio9, bio10 and PC4 with correlation coefficient of magnitudes -0.3085, -0.4425, -0.4260, -0.3023, -0.3385, -0.4701 and -0.4475 respectively. There appeared there was moderately negatively correlation between tmin7, tmin8, bio4 and PC4 with correlation coefficient of magnitudes -0.5846, -0.5617, -0.5865. There was any shown positively linear relationship between tmax5, tmax10, tmax11, prec3, bio8, bio11, bio15 with PC4 with correlation coefficient of magnitudes 0.0604, 0.1360, 0.2707, 0.0058, 0.0501, 0.1177 and 0.2890 respectively. There were suggested weakly positively correlation between tmax1 to tmax4, tmax12, bio2 with PC4 with correlation coefficients of magnitudes 0.3150, 0.2875, 0.2675, 0.2429, 0.3337, 0.4534 respectively. There was an implied strongly positively correlation between bio3 and PC4 with correlation coefficient of magnitude 0.8606.

Principal component 5 (PC5)

There were recommended moderately positively correlation between prec9, bio18 and PC5 with correlation coefficients of magnitudes of 0.5631 and 0.5431 respectively. There were tabled

weakly positively correlation between prec7, prec8, bio 14, bio 17 and PC5 with correlation coefficients of magnitudes of 0.3772, 0.4308, 0.3794, 0.3689 respectively. There seemed there were any portrayed positively linear relationship between tmin1 to tmin5, tmin9 to tmin12, tmax1 to tmax3, tmax9 to tmax12, prec6, prec10, bio1, bio3, bio6, bio8, bio11 with PC5 with correlation coefficients of magnitudes of 0.2805, 0.1632, 0.0794, 0.0094, 0.0440, 0.0850, 0.1949, 0.2585, 0.2638, 0.1899, 0.0197, 0.0048, 0.0350, 0.1209. 0.1899, 0.1822, 0.2876, 0.1547, 0.0776, 0.2099, .02781, 0.0413, 0.1761 respectively. There did not seem to be any recommended negatively linear relationship between tmin6 to tmin8, tmax4 to tmax8, prec1 tp prec5, prec11, prec12, bio2, bio4, bio5, bio7, bio10, bio12, bio13, bio15, bio16, bio19 and PC5 with correlation coefficients of magnitudes -0.0118, -0.1296, -0.1224, -0.0044, -0.0126, -0.0148, -0.1159, -0.1283, -0.0391, -0.1706, -0.0882, -0.0358, -0.2094, -0.0536, -0.0776, -0.2804, -0.1122, -0.2555, -0.0801, -0.0910, -0.0045, -0.0700, -0.2487, -0.0811, -0.0444 respectively.

For all the principal components retained, some climatic variables did not seem to suggest any type of linear relationship with the respective principal components. Therefore, aggregately, the principal components were undefined and not used in allele-environment association analysis.

Figure 7, showed a visual output of principal component 1 by principal component 2. Total variance is equal to the sum of the variances of individual variables. The first principal component accounts for the 50.85% of the total variation, whereas the first two components cover almost 73.64% of the total variability. Scatter plot suggest that the individuals were not separated by country, since no pattern of clusters were evident for each group of countries.

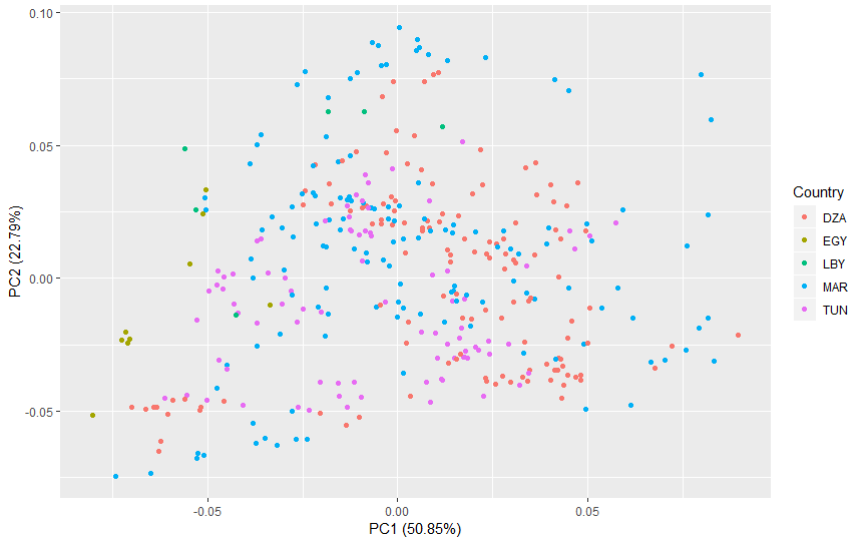

Figure 7: *Scatter plot of principal component 1 by principal component 2, with country effect for climatic data.*

Correlation structure

To have understanding of correlation pattern of climatic data, visualized correlation matrix, computed by Pearson correlation between these climatic variables, were presented by use of correlogram graphical tool in Figure 8. Correlations were illustrated in two ways. The red color showed negative correlation whereas the blue color positive correlation. The size and intensity of color, gave an indication of the correlation. Correlation coefficients whose magnitudes were between 0.90 and 1.00, 0.70 and 0.89, 0.50 and 0.69, 0.30 and 0.49, and between 0.00 and 0.29 were very highly correlated, highly correlated, moderately correlated, weakly correlated and no linear relationship (no relationship) respectively as reported by J. Rumsey (2016).

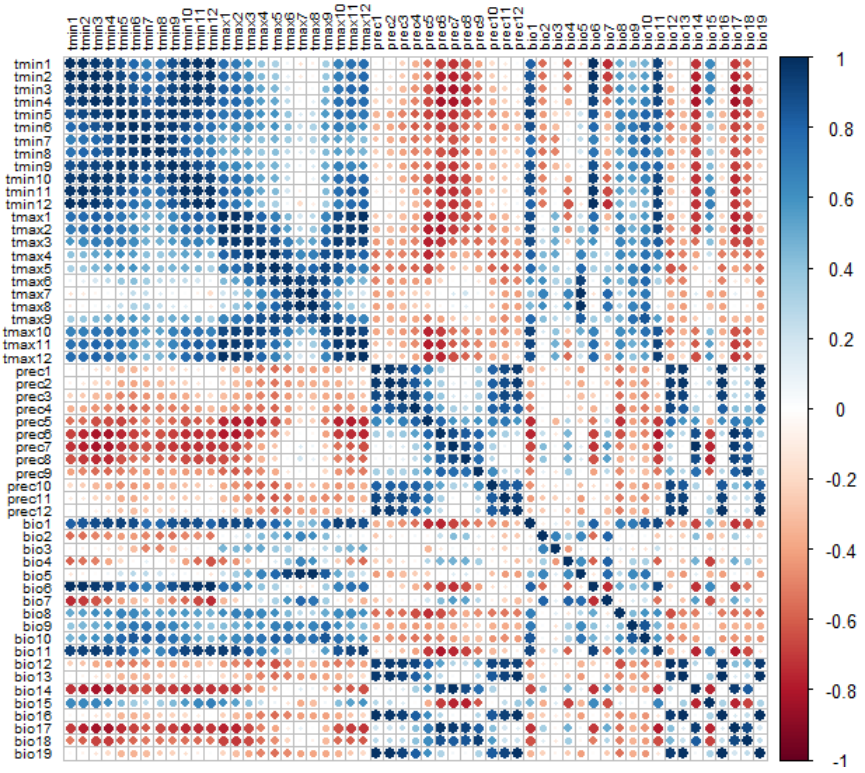

Figure 8: *Correlogram illustrating correlation matrix for climatic variables.*

Outputs of very highly correlated (correlation coefficient whose magnitude is between 0.9 and 1), were presented on Table 6 in the appendix. Correlations between pairs of predictors were computed for a correlation coefficient magnitude threshold of 0.90. Means of each of these pairs were also computed and compared. Predictors with lower means were discarded, retaining ones with higher means. Below were listed independent variables responsible for probable climatic confounding effects. bio1, tmin5, tmax2, tmax10, tmax3, tmin4, tmax11, tmin6, bio11, tmin3, tmax1, tmin10, tmax4,prec6, tmin9, tmin11, tmax5, tmin2, bio17, tmin7, bio6, tmin12, tmax9,

prec7, bio12, prec3, prec9, tmax6, prec11, prec1, prec2, bio16, bio19, prec12, tmax8, tmax7. Description of these variables were demonstrated on Table 1.

Variable selection for climatic data

Figure 9, is a correlogram which showed filtered out variables which were not correlated more than a correlation coefficient magnitude threshold of 0.9 using *function findcorrelation* from Caret package in R statistical software. These filtered out climatic variables were the variables that were used in the allele-environment association analysis. These variables were tmin1, tmin8, tmax12, prec4, prec5, prec8, prec10, bio2, bio3, bio4, bio5, bio7, bio8, bio9, bio10, bio13, bio14, bio15, bio18. Description of these variables were presented on Table 1, above.

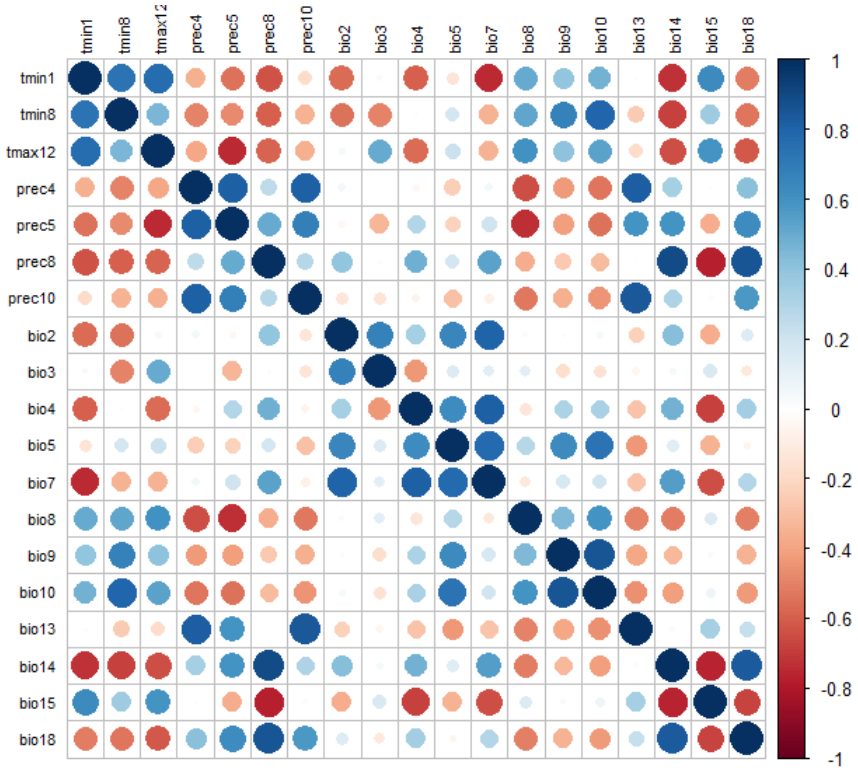

Figure 9: *Correlogram illustrating correlation matrix for climatic variable at correlation below threshold 0.9.*

Summary statistics of selected climatic variables

Results of evaluation of summary statistics such as minimum values, means, medians, maximum values and standard deviations, for these correlated environmental variables at a correlation coefficient below correlation coefficient magnitude threshold of 0.9, were presented on Table 4. Description of the variables were on Table 1.

Table 4: *Summary statistics for environmental variables correlated below threshold of 0.9.*

Variable	Minimum	Mean	Median	Maximum	Std Dev
tmin1	-60.00	37.43	37.00	97.00	26.47836
tmin8	114.0	194.1	197.0	266.0	28.0506
tmax12	72.0	142.3	140.0	237.0	28.67522
prec4	0.00	35.92	37.00	99.00	17.87187
prec5	0.00	28.96	33.00	62.00	14.31776
prec8	0.000	7.317	6.000	29.000	5.517535
prec10	0.000	36.63	37.00	89.00	14.07705
bio2	66.0	115.1	119.0	173.0	22.88504
bio3	27.00	38.17	39.0	52.00	4.957282
bio4	2415	6269	6443	8110	871.935
bio5	230.0	334.7	329.0	446.0	28.29171
bio7	137.0	297.4	297.0	410.0	41.51772
bio8	32.0	115.6	107.0	232.0	39.75605
bio9	120.0	249.1	251.0	325.0	22.47946
bio10	185.0	252.2	253.0	331.0	22.0311
bio13	1.0	57.9	57.0	174.0	30.54938
bio14	0.00	3.83	3.00	14.00	3.310628
bio15	26.00	53.31	55.00	109.00	15.97353
bio18	0.00	30.83	30.00	64.00	15.75293

Cluster analysis on climatic data

Table 5 illustrated comparison of clusters to countries. K-means clustering on climatic data with 5 clusters of sizes 66, 108, 337, 158, 250. The table showed data that belonged to Algeria got grouped mainly into cluster 2, 3 and 4, Egypt mostly into cluster 5, Libya substantially into cluster 1, Morocco largely into clusters 1, 4 and 5, and Tunisia basically into clusters 3 and 5. The algorithm wrongly classified 2 data points into cluster 2 and 27 data points into cluster 5, that belonged to Algeria. 6 data points that belonged to Egypt were wrongly classified into cluster 1.

A data point that belonged to Libya was wrongly classified into cluster 5. 12 and 16 data points were wrongly classified into cluster 2 and 3 respectively, that belonged to Morocco, and 25 and 26 data points that belonged to Tunisia, were wrongly classified to clusters 2 and 4 respectively.

Table 5: *Comparing clusters of climatic variables with country.*

Cluster	DZA	EGY	LBY	MAR	TUN
1	2	6	8	50	0
2	71	0	0	12	25
3	165	0	0	16	156
4	82	0	0	49	26
5	27	28	1	56	139
Total	347	34	9	183	346

4.2 Allele-environment association analysis

4.2.1 Latent factor mixed models

Results of latent factor mixed models were visualized using histogram of p values. The principal goal of latent factor mixed models analysis was to arrive at sets of corrected p values, confirming the fact that potential confounding effects that resulted from population ancestry or other sources not accounted for by adaptation to their inhabitats were overpowered. Histograms of corrected (adjusted) p values, appeared like plateaus with peaks close to zero. Manhattan plots tabled at the appendix were presented to demonstrate possible SNPs which were identified associated with the respective ecological variables. SNPs with highest peaks above significant genome-wide line (red horizontal line of Manhattan plots visualized in the appendix) were portrayed as strongly associated meanwhile those above the suggestive genome-wide line (blue horizontal line of Manhattan plots observed in the appendix) were recommended as associated with the respective climatic variables. Manhattan plots of the various climatic variables considered in the current study were observed at the appendix.

For all the 19 environmental gradients adopted in the LFMMs analysis, it was observed that, all the adjusted p values histograms seemed to have flat peaks close to zero. These appeared to demonstrate uniform distributions of the adjusted p values. Thus, it was tabled that probable confounding effects in the analyses were extinguished.

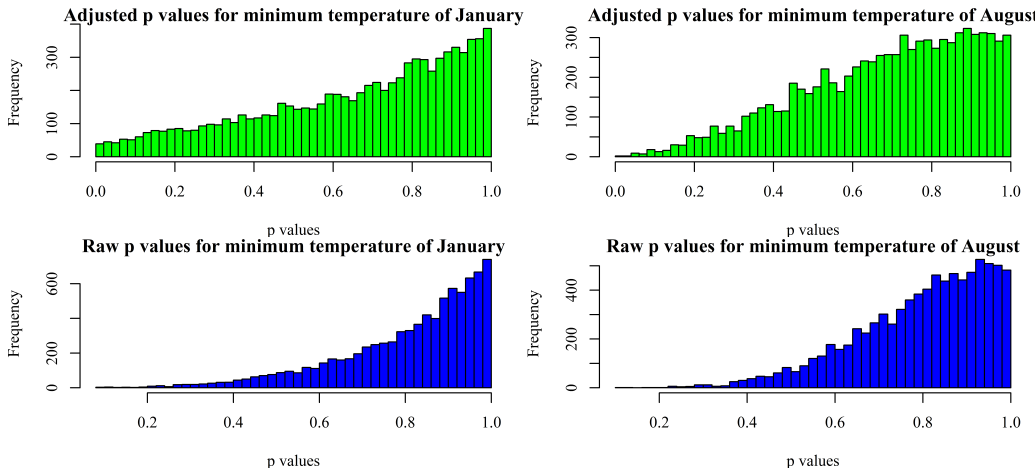

Figure 10: *Histograms of p values for minimum temperature of January and August.*

Left and right panels of Figure 10, suggested heavily left skewed distributions with peak closest to zero for unadjusted p values. These suggested, individuals of durum wheat were sampled from dissimilar distributions. Originally, any SNPs were shortlisted to be associated by any of panels, since none was observed to have had raw p values<0.05 (significant unadjusted p values). Roughly 50 SNPs were observed to obtain adjusted p values<0.05, but were not suggested to be highlighted as associated with the climatic variables, since they did not appear to have adjusted p values < Benjamini-Hochberg critical values.

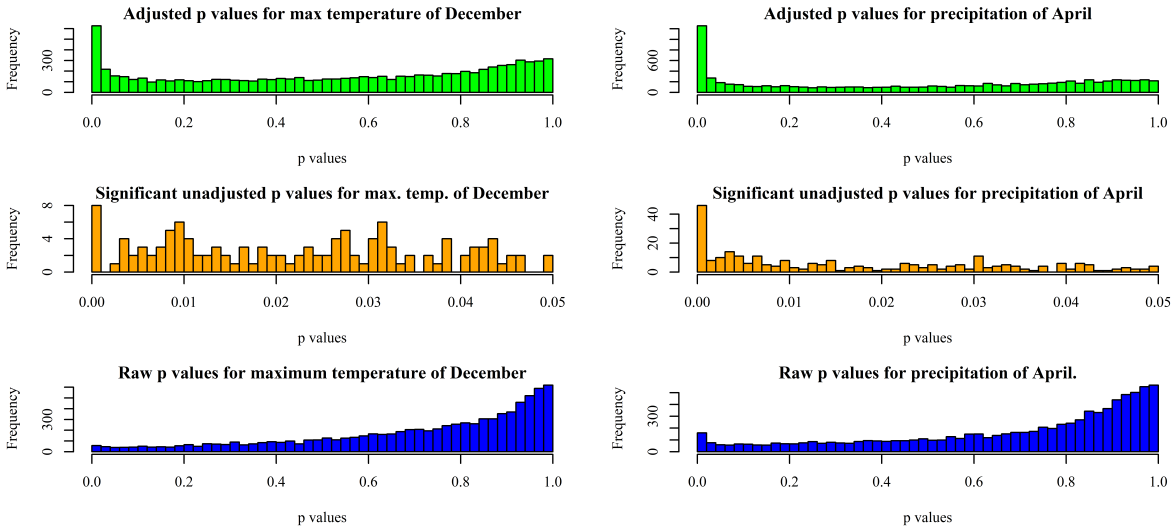

Figure 11: *Histograms of p values for maximum temperature of December and precipitation of April.*

Figure 11 tabled left skewed distributions for both right and left panels. Virtually 123 and 263 SNPs and almost 371 and 1158 SNP were suggested as associated with minimum temperature of

december and precipitation of april respectively, since these 123 and 236 SNPs obtained raw p values<0.05 and 371 and 1158 SNPs appraised adjusted p values < agreed Benjamini-Hochberg critical values.

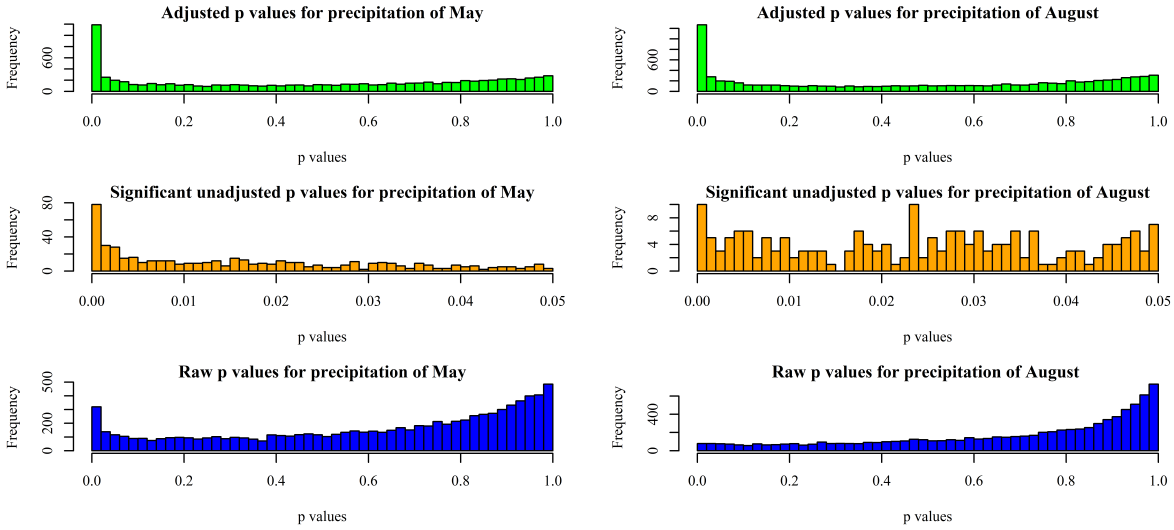

Figure 12: *Histograms of p values for precipitation of May and August.*

Raw p values of left panel of of Figure 12 recommended almost uniform distribution whereas the right panel of same figure portrayed left skewed distribution with peak close to zero. Approximately 221 and 906 SNPs were depicted as associated with precipitation of May, since 221 SNPs obtained raw p values < 0.05 and 906 SNPs examined adjusted p values < concurred Benjamini-Hochberg critical values. There were any recommended SNPs associated to precipitation of August, as there were any SNPs that established either raw p values < 0.05 or evaluated adjusted p values < coincided Benjamini-Hochberg critical values.

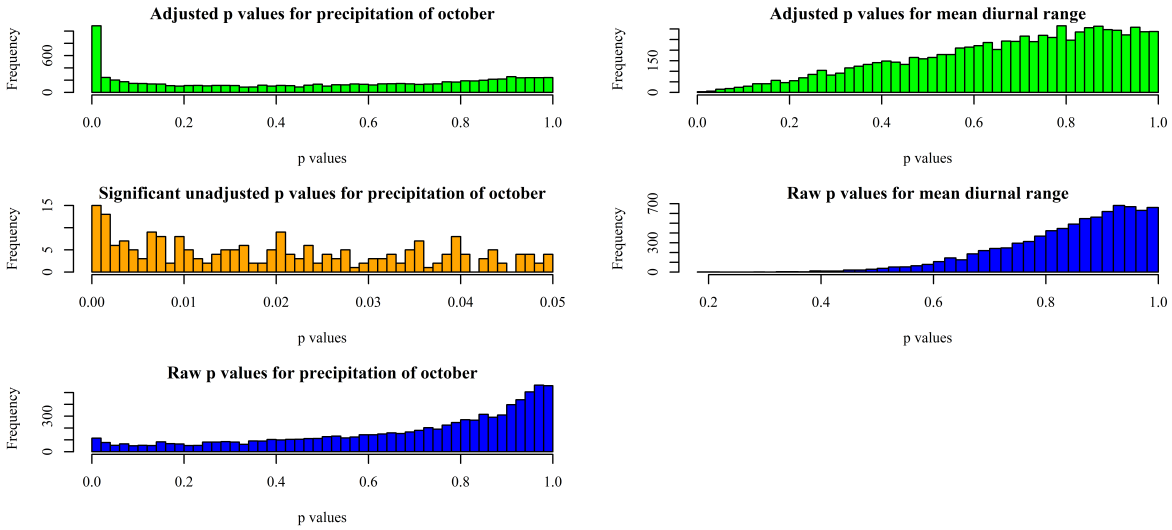

Figure 13: *Histograms of p values for precipitation of October and mean diurnal range.*

There seemed to be an observed almost uniform distribution of raw p values at the left panel of Figure 13 with a heavily left skewed distribution of SNPs illustrated by raw p values at the right panel of same figure. Nearly 221 and 906 SNPs were tabled as associated with precipitation of October, since 221 SNPs obtained raw p values < 0.5 and 906 appraised adjusted p values $<$ coincided Benjamini-Hochberg critical. There did not seem to be any SNPS recommended as associated with mean diurnal range, since any SNPs did not obtain either raw p values <0.05 or established adjusted p values $<$ accorded Benjanmini-Hochberg critical values.

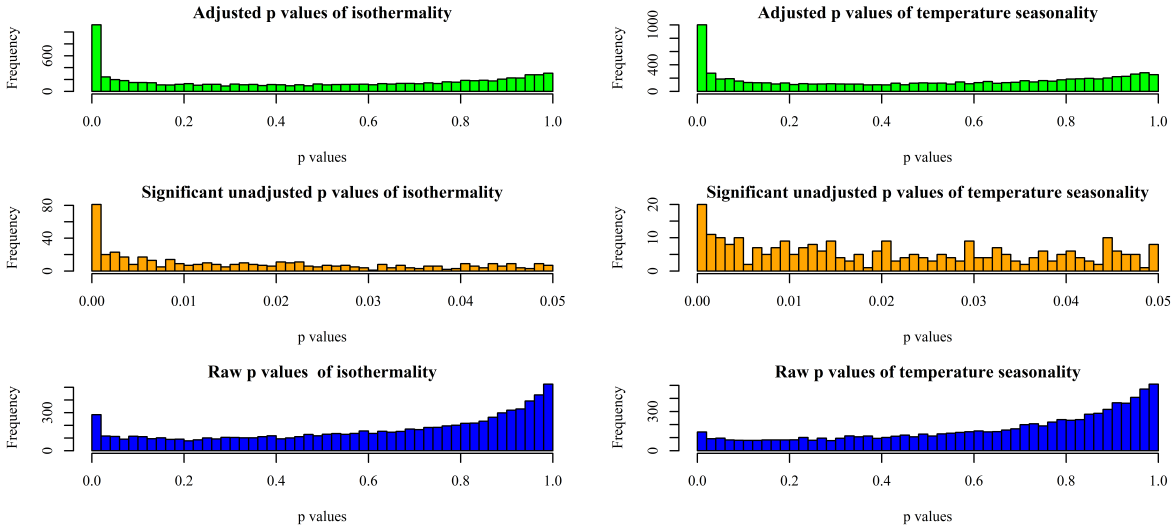

Figure 14: *Histograms of p values for Isothermality and temperature seasonality.*

Figure 14 illustrated virtually uniform distributions of raw p values for both climatic variables. There were recommended associations between SNPs and both climatic variables, since 446 and

285 SNPs evaluated raw p values < 0.05 and 995 and 796 SNPs approximated adjusted p values $<$ corresponded Benjamini-Hochberg critical values for isothermality and temperature seasonality respectively. Thus 995 and 796 SNPs were the demonstrated candidate loci associated with isothermality and temperature seasonality respectively.

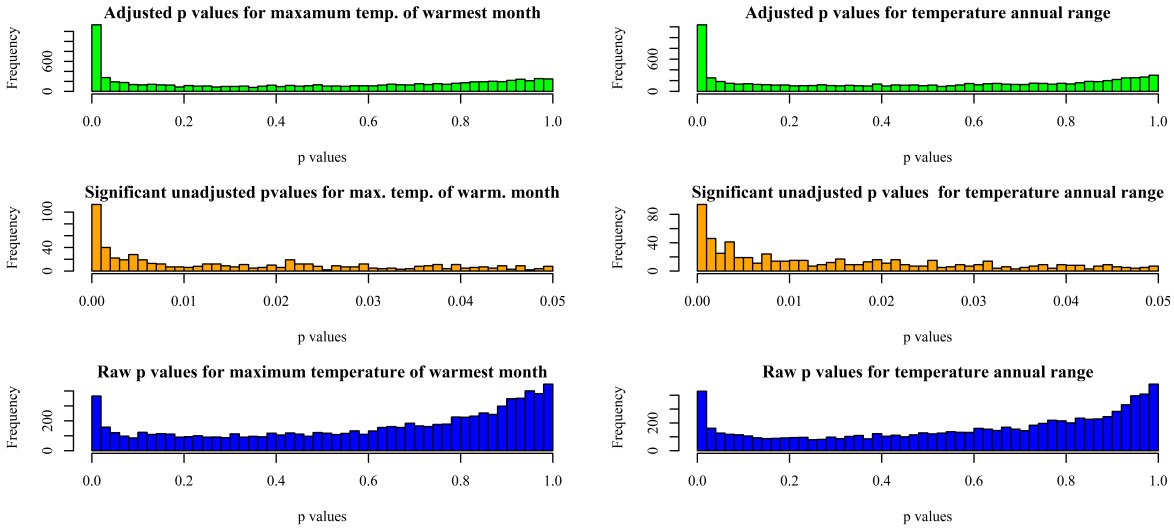

Figure 15: *Histograms of p values for maximum temperature of warmest month and temperature annual range.*

It was observed from Figure 15 that, raw p values of SNPs showed almost uniform distributions for both climatic variables. Approximately 582 and 653 SNPs were tabled associated with maximum temperature of warmest month and temperature annual range respectively, since they assessed raw p values < 0.05. Nearly 1233 and 1135 SNPs were recommended candidate loci with maximum temperature of warmest month and temperature annual range respectively, since they established adjusted p values $<$ coincided Benjamini-Hochberg critical values.

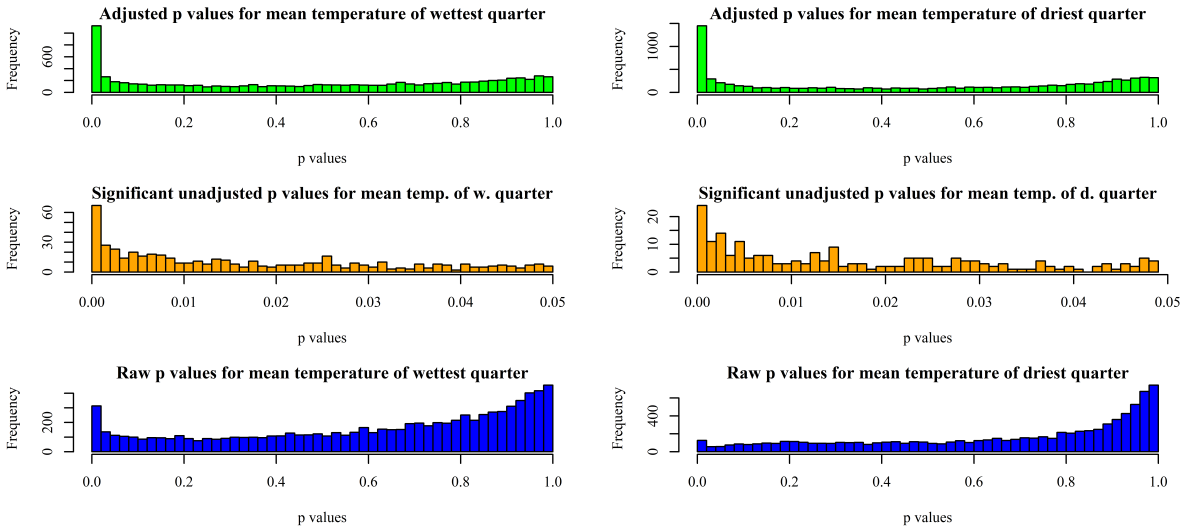

Figure 16: *Histograms of p values for mean temperature of wettest and driest quarters.*

Almost observable uniform distribution was recommended by raw p value of SNPs at left panel of Figure 16 with a seemly left skewed distribution of SNPs with peaks close to zero at the right panel. Roughly 511 and 204 SNPs were tabled associated with mean temperature of wettest quarter and mean temperature of driest quarter respectively, since they both established raw p values < 0.05. About 963 and 1377 candidate loci were observed in both panels of Figure 16, since they approximated adjusted p values $<$ corresponded Benjamini-Hochberg critical values.

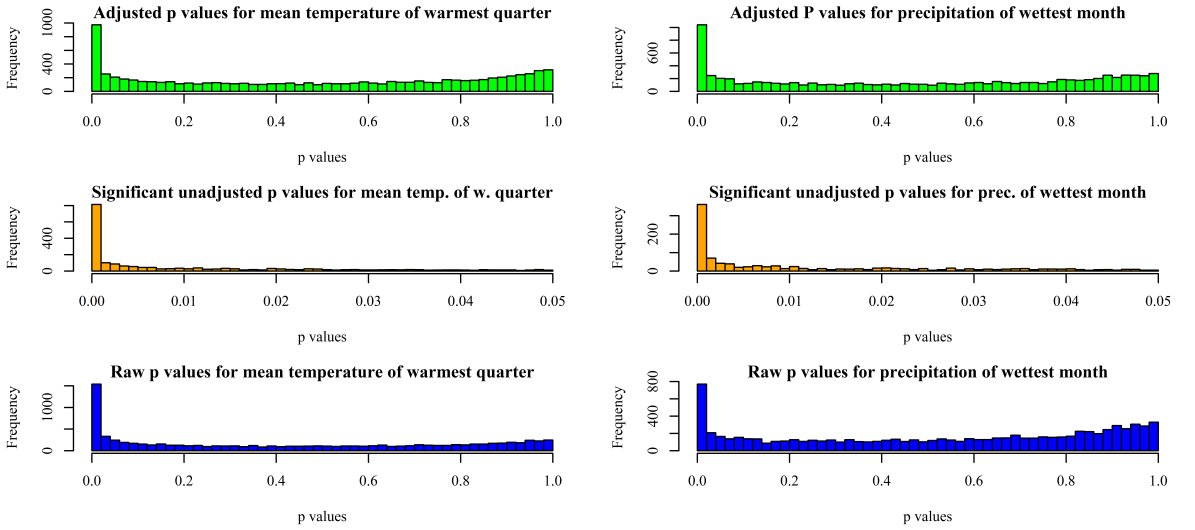

Figure 17: *Histograms of p values for mean temperature of warmest quarter and precipitation of wettest month.*

Raw p values of SNPs in left panel of Figure 17 demonstrated virtually right skewed distribution with peak close to zero and almost uniform distributions was illustrated at the right panel.

Nearly 1992 and 1050 SNPs assayed raw p values < 0.05, thus tabled associated with mean temperature of warmest quarter and precipitation of wettest month respectively. Moreover, 782 and 878 SNPs were presented as candidate loci for mean temperature of warmest quarter and precipitation of wettest month respectively, since they both appraised corrected p values < established Benjamini-Hochberg critical values. It was depicted that 1210 and 172 SNPs preliminarily recommended associated were not portrayed as candidate loci for mean temperature of warmest quarter and precipitation of wettest month respectively. Thus 1210 and 172 SNPs were the suggested false discoveries for mean temperature of warmest quarter and precipitation of wettest month respectively.

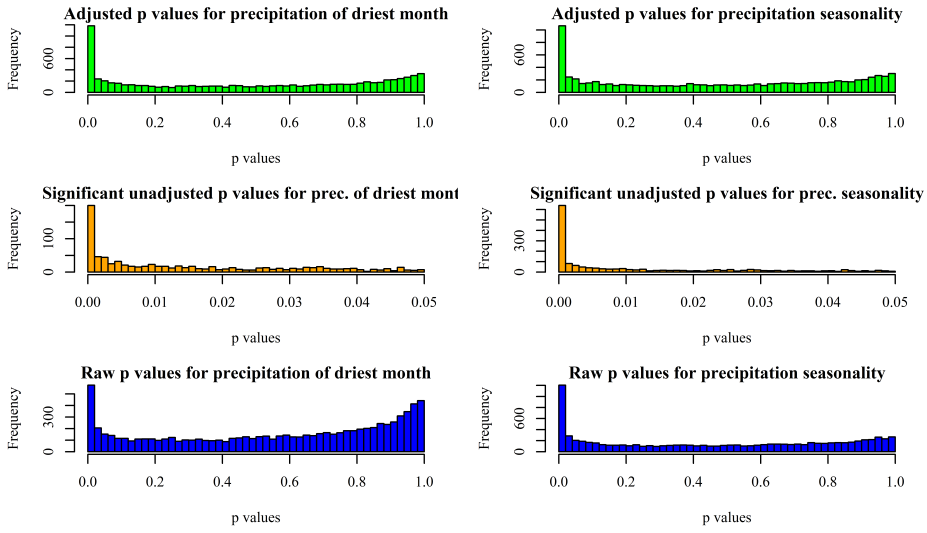

Figure 18: *Histograms of p values for precipitation of driest month and precipitation seasonality.*

Raw p values of SNPs at left panel of Figure 18 suggested almost uniform distribution with a right skewed distribution presented by raw p values at the right panel of same figure. Roughly 851 and 1603 SNPs evaluated raw p values < 0.05, thus tabled associated with precipitation of driest month and precipitation seasonality respectively. Nearly 1050 and 901 SNPs established adjusted p values < coincided Benjamini-Hochberg critical values. Therefore, about 1050 SNPs were flagged candidate loci for mean temperature of warmest quarter and almost 901 suggested candidate loci precipitation of wettest month. It was deduced that virtually 702 SNPs were check listed as false discoveries, since an excess of practically 702 SNPs that adopted raw p values < 0.05 did not table appraised adjusted p values < corresponded Benjamini-Hochberg critical values.

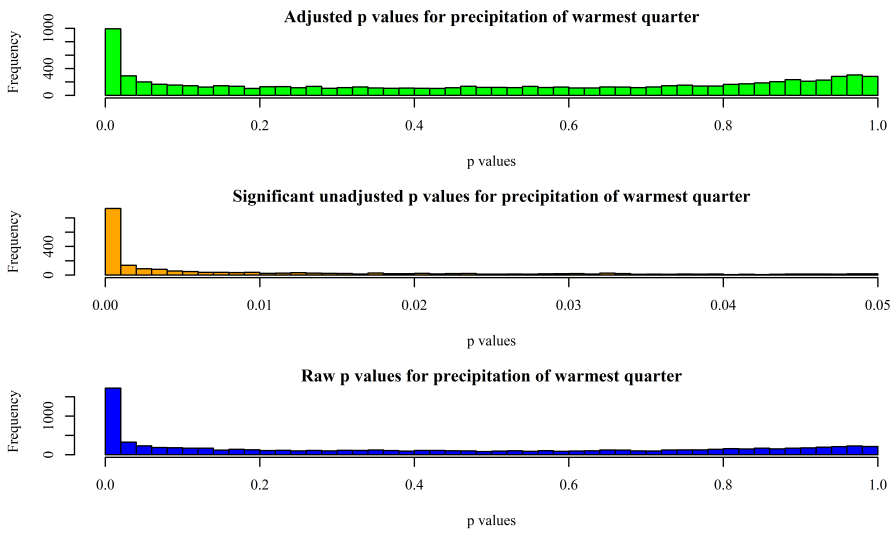

Figure 19: *Histograms of p values for precipitation of warmest quarter.*

Figure 19 suggested right skewed distribution of raw p values of SNPs with precipitation of warmest quarter. Almost 2171 SNPs examined raw p values < 0.05, thus portrayed associated with precipitation of warmest quarter. Among the 2171 SNPs flagged associated with precipitation of warmest quarter, virtually 779 SNPs were recommended candidate loci for precipitation of warmest quarter. It was deduced that 1392 SNPs which were initially suggested as associated were not flagged candidates, thus tabled false discoveries.

5 Discussion and Conclusion

5.1 Discussion

The goal of the current study was to examine association between alleles and environmental gradients. Alleles was the outcome variable of 919 $\times$ 8366 matrix. Environmental gradients coincided to the predictors. Latent factor mixed model was adopted in allele-environment analysis.

5.1.1 Genotypic data

Preliminary exploratory analysis on genetic data, based on principal component analysis, revealed nine principal components which explained nine clusters within the genetic population (figure 2). Score plot of the first two principal components inclusive of country effects, suggested genes were not separated by their countries of origin, since there did not seem to be any visible clusters of genes from same countries (figure 3). It was of interest to examine the ancestry of these nine clusters. Appraisal was feasible by use of ancestry matrix graphical technique which was a plot that described ancestry proportions of each individual in the population under study (figure 4). It was depicted that nine clusters appeared to have had similar ancestors as the ancestry map did not table any visible isolation of distinct clusters. A cross-entropy validation algorithm was implemented to ascertain the number of sub genetic populations that described the entire genetic population. The approach was employed using sNMF program in LEA package in R statistical software (Frichot et al., 2015). An aggregate of 20 values of K were projected, with 20 coincided cross-entropy values. It was shown that there was an initial sharp descent pattern for the first nine values of K, with succeeding alteration of increments and decrements of the remaining values of K (figure 5). It was deduced that nine factors were suggested by the sNMF program. Run 4 was observed as the ideal run, since it recommended the least cross-entropy (table 3). The procedure seconded the tabled nomination of nine genetic clusters by principal component analysis for genetic data. The evaluated nine cluster of genetic data were employed as portrayed number of latent factors in allele-environment analysis using latent factor mixed models.

5.1.2 Climatic gradients

A total of 55 ecological variables were present in climatic data. Employing these entire climatic predictors in alleles-environment analysis posted issues. These problems included high

computational price to perform learning and inference, data visualization and interpretation and feasible over fitting. In resolving these bottlenecks, preliminary analysis was done based on principal component analysis on correlation matrix. It was depicted that five principal components explained almost 96% of variability in the climatic data (figure 6). It was noticed from the score plot of the first two principal climatic component that, there did not appear to be in any trend of spread of individuals in the different countries. Individuals from different countries did not form any observable cluster (figure 8). However, since interpretation was required between alleles and each ecological gradient, using the first principal components that preserved the majority of the relevant information in the climatic data, will not satisfy the study motivation. High correlation filter approach was implemented on climatic variables. A correlogram was the graphical tool used to visualize possible correlation that existed between environmental variables (figure 9). Pairs of climatic variables with correlation coefficient magnitude threshold of 0.9. were discarded with one with higher mean retained. Climatic variables with correlation coefficient magnitude of threshold below 0.9 were retained (figure 10). An aggregate of 19 environmental variables were retained as the predictors used in alleles-environment analysis (figure 10). Summary statistics of these climatic variables were computed and presented on table 6. It was also of interest to examine how individuals were grouped in the different countries. K-means clustering procedure was used to depict the clustering pattern (table 7). It was observed that the majority of the individuals were correctly group in their respective countries with some minimal wrong groupings.

5.1.3 Latent factor mixed models

Fundamental motivation of LFMMs analyses were to obtain histograms of adjusted p values which seemed to describe uniform distribution. This trend of uniform distribution seemed to be visualized from figures (11) to (20). These suggested that potential confounding effects were extinguished. Various distributions of raw p values were illustrated from figures (11) to (20). Probable associated SNPs with respective climatic variables adopted in the current analysis were also visualized from figures (11) to (20). Feasible false discoveries were highlighted on figures (19) and (20).

5.2 Conclusion

SNPs flagged associated with the various climatic variables adopted in LFMMs analysis (candidate loci), were those which recommended achievable association with the 19 ecological gradients

after confounding effects were overpowered. Among the 8366 SNPs considered in the current study, none were associated with minimum temperature for January, August and mean diurnal range . Nearly 371 SNPs were selected as candidate loci for maximum temperature of December. Roughly 1158 , 1024, 1173 and 906 SNPs were archived candidate loci for precipitation for April, May, August and October respectively. Virtually 955, 796, 1233 SNPs were balloted candidate loci for isothermality, temperature seasonality and maximum temperature of warmest month respectively. Approximately 1135, 963 and 1377 SNPs were cataloged candidate loci for temperature annual range, mean temperature of wettest quarter and mean temperature of driest quarter respectively. Practically 782, 878 and 1050 SNPs were checklisted candidate loci for mean temperature of warmest quarter, precipitation of wettest month and precipitation of driest month respectively. Finally, approximately 901 and 779 SNPs were flagged candidate loci for precipitation seasonality and precipitation of warmest quarter respectively. These associations were observed on the Manhattan plots in the appendix.

References

[1] Alexander,D.H. & Lange, K.(2011), Enhancements to the DAMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics,* **12**,246.

[2] Atwell,S., Huang Y.S., Vilhjalmsson, B.J., Willems, G., Horton, M.,Li, Y., et al. (2010) Genome-Wide Association study of 107 phenotypes in Arabidopsis thaliana inbred lines.Nature,**465**,627-631.

[3] Benjamini, Y. & Hochberg, Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing.*Journal of the Royal Statistical Society.Series* **B** *(Methodological)* **57**, 289-300.

[4] Carbon, S. Ireland,A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., et al(2009)AmiGO:online access to ontology and annotation data. Bioinformatics, **25**, 288-289.

[5] Frichot, E., Schoville, S.D., de Villemereuil, P.,Gaggiotti, O.E. & Francois,O.(2015)Detecting adaptive evolution based on association with ecological gradients: Orientation matters! *Heredity*, **115**, 22-28.

[6] Frichot, E., Mathieu, F., Trouillon,T., Bouchard, G. & Francois,O.(2014) Fast and efficient estimation of individual ancestry coefficients.*Genetics*,**196** ,973-983.

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

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

[9] Pritchard,J.K., Stephens, M.& Donnelly, P.(2000).Inference of population structure using multilocus genotype data *Genetics*,**155** ,945-959.

[10] Patterson,N., Price, A.L& Reich, D.(2006) Population structure and eigenanalysis *PLo Genetics*,**2** ,20.

[11] Currat,M., Ray, N.& Excoffier, L.(2004) SPLATCHE:a program to simulate genetic diversity taking into account environmental heterogeneity.*Molecular Ecology Notes*,**4** , 139-142.

[12] Darwin,C.(1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* John Murray, London.

[13] Devlin,B.& Roeder, K.(1999) Genomic control for association studies *Biometrics .* **55** , 997-1004. bibitemFrancois Francois,O. & Durand,E.(2010) Spatially explicit bayesian clustering models in population genetics.*Molecular Ecology Resources*, **10**, 773-784.

[14] Williams,G.C.(1966) *Adaptation and Natural selection, volume 1996.* Princeton University Press, Princeton, New Jersey.

[15] Schoville,S.D., Bonin, A.,Francois,O., Lobreaux,S., Melodelima,C.& Manel,S.(2012) Adaptive genetic variation on landscape: Methods and cases.*Annual Review of Ecology and Systematics*, **43**, 23-43.

[16] Fumagalli,M., Sironi, M.,Pozzoli, U.,Ferrer-Admettla,A.,Pattini, L.& Nielsen, R.(2011) Signatures for environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution *PLo Genetics*,7,e1002355.

[17] Liptak,T.(1959) On the combination of independent tests.*Publications of the Mathematical Institute of the Hungarian Academy of Science.***3**, 171-197.

[18] Haldane,J.B.S.(1948) The Theory of cline. *The Journal of Genetics*,**48**, 277-284.

[19] Hancock, A.M., Witonsky, D.B.,Gordon, A.S., Eshel, G., Pritchard,J.K., Coop, G. & Di Rienzo,A. (2008) Adaptations to climate in candidate genes for common metabolic disorders *PLoS Genetics*, **4**,13.

[20] Manel, S., Joost, S., Epperson, B.K., Holderegger, R., Storfer, A., Rosenberg, M.S., Scribner, K.T., Bonnie, A. & Fortin,M.J. (2010) Perspectives on the use of landscape genetics to detect adaptive variations in the field. *Molecular Ecology*, **19**,3760-3772.

[21] Jay, F., Manel, S., Alvarez, N., Durand, E.Y., Thuiller, W., G., Holderegger, R., Taberlet, P. & Francois,O. (2012) Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, **21**,2354-2368.

[22] Frichot, E., & Francois,O.(2015). LEA: An R Package for Landscape and Ecological Association Studies.

[23] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44:821–824. [PMC free article] [PubMed].

[24] Yu J., Pressoir G., Briggs W.H., Vroh Bi I., Yamasaki M., Doebley JF., McMullen MD., Gaut BS., Nielsen DM., Holland JB., Kresovich S., Buckler ES., A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38:203–208. [PubMed].

[25] Young JH, Chang YC, Kim JD, Chretien JP, Klag MJ, Levine MA, Ruff CB, Wang NY, Chakravarti A. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. PLoS Genet. 2005;1:e82. [PMC free article] [PubMed]].

[26] West M. Bayesian factor regression models in the "large p, small n" paradigm. Bayesian Stat. 2003;7:723–732.

[27] Tipping ME, Bishop CM. Probabilistic principal component analysis. J Roy Stat Soc B. 1999;61:611–622.

[28] Thibert-Plante X, Hendry AP. When can ecological speciation be detected with neutral loci? Mol Ecol. 2010;19:2301-2314. [PubMed].

[29] Storz JF, Wheat CW. Integrating evolutionary and functional approaches to infer adaptation at specific loci. Evolution. 2010;64:2489-2509. .

[30] Storz JF. Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol Ecol. 2005;14:671-688.

[31] Smouse PE, Long JC, Sokal RR. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst Biol. 1986;35:627-632.

[32] Salakhutdinov R, Mnih A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. ICML. 2008;25:880-887.

[33] Saccone SF, Quan J, Mehta G, Bolze R, Thomas P, Deelman E, Tischfield JA, Rice JP. New tools and methods for direct programmatic access to the dbSNP relational database. Nucleic Acids Res. 2011;39:D901-D907.

[34] R Development Core Team. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing; 2012.

[35] Prugnolle F, Manica A, Charpentier M, Gugan JF, Guernier V, Balloux F. Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol. 2005;15:1022-1027.

[36] Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol. 2010;20:R208-R215.

[37] Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 2009;5:e1000519.

[38] Poncet BN, Herrman D, Gugerli F, Taberlet P, Holderegger R, Gielly L, Rioux D, Thuiller W, Aubert S, Manel S. Tracking genes of ecological relevance using a genome scan in two independent regional population samples of Arabis alpina. Mol Ecol. 2010;19:2896-2907.

[39] Pavlidis P, Jensen JD, Stephan W, Stamatakis A. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. Mol Biol Evol. 2012;29:3237-3248.

[40] Novembre J, Di Rienzo A. Spatial patterns of variation due to natural selection in humans. Nat Rev Genet. 2009;10:745-755.

[41] Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005;39:197-218.

[42] Nei M. Genetic distance between populations. Am Nat. 1972;106:283-292.

[43] Meirmans PG. The trouble with isolation by distance. Mol Ecol. 2012;21:2839-2846.

[44] Li JZ, Absher DM, Tang H, et al. (11 co-authors) Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008;319:1100–1104.

[45] Lenormand T. Gene flow and the limits to natural selection. Trends Ecol Evol. 2002;17:183-189.

[46] Legendre P, Legendre L. Numerical ecology. 3rd English ed. Amsterdam (Netherlands): Elsevier; 2012.

[47] Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8:e1002453.

[48] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer. 2009;8:30-37.

[49] Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM. Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. 2006;16:980-989.

[50] Jolliffe IT. Principal component analysis. New York: Springer Verlag; 1986.

[51] Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002;18:337-338.

[52] Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. 2012;40(Database issue):D261-D270.

[53] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106:9362-9367.

[54] Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25:1965-1978.

[55] Harmon LJ, Glor RE. Poor statistical performance of the Mantel test in phylogenetic comparative analyses. Evolution. 2010;64:2173-2178.

[56] Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A. Adaptations to climate-mediated selective pressures in humans. PLoS Genet. 2011;7:e1001375.

[57] Frichot E, Schoville SD, Bouchard G, François O. Correcting principal component maps for effects of spatial autocorrelation in population genetic data. Front Genet. 2012;3:254.

[58] Engelhardt BE, Stephens M. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genet. 2010;6:e1001117.

[59] Endler JA. Geographic variation, speciation, and clines. Princeton (NJ): Princeton University Press; 1977.

[60] Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.

[61] Eckert AJ, Bower AD, Gonzàlez-Martìnez SC, Wegrzyn JL, Coop G, Neale DB. Back to nature: ecological genomics of loblolly pine (Pinus taeda, Pinaceae) Mol Ecol. 2010;19:3789-3805.

[62] Eckart C, Young G. The approximation of one matrix by another of lower rank. Psychometrika. 1936;1:211-218.

[63] Durand E, Jay F, Gaggiotti OE, François O. Spatial inference of admixture proportions and secondary contact zones. Mol Biol Evol. 2009;26:1963-1973.

[64] Chen C, Durand E, Forbes F, François O. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. Mol Ecol Notes. 2007;7:747-756.

[65] Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. High-dimensional sparse factor modeling: applications in gene expression genomics. J Am Stat Assoc. 2008;103:1438-1456.

[66] Berry A, Kreitman M. Molecular analysis of an allozyme cline: alcohol dehydrogenase in Drosophila melanogaster on the East Coast of North America. Genetics. 1993;134:869–893.

[67] Beaumont MA, Nichols RA. Evaluating loci for use in the genetic analysis of population structure. Proc R Soc B Biol Sci. 1996;263:1619–1626.

[68] Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol. 2004;13:969–980.

[69] Kevin Caye and Olivier Francois. LFMM 2.0: Latent factor models for confounder adjustment in genome and epigenome-wide association studies.

[70] Barrett RD, Hoekstra HE. Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet. 2011;12:767–780.

[71] Akey JM. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res. 2009;19:711–722.

[72] Flori,L., Fritz, S., Jaffrezic, F., Boussaha, M., Gut, I., Health, S., Foulley, J., Gautier, M., (2009). The genome response to artificial selection: a case study in dairy cattle.PLoS One 4(8), e6595.

[73] Davey, J., Hohenlohe, P., Etter, P., Boone, J., Catchen, J., Blaxter, M.,(2011). Genome-Wide genetic marker discovery and genotyping using next-generation sequencing. Nature Review Genetics. **12**, 499-510.

[74] Yamasaki, M., Tenaillon, M.I., Vroh Bi, I., Schroeder, S., Sanchez-Villeda, H., Doebley, J., Gaut, B., McMullen, M.,(2005). A large scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement plant cell. **17(11)**, 2859-2872.

[75] Hansen, M.M., Olivieri, I., Waller, D.M., Nielsen, E.E., GEM Working Group(2012). Monitoring adaptive genetic responses to environmental change. *Molecular Ecology***21(6)**, 1311-1329.

[76] Guillot,G., et al., Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies.Spatial statistics(2013),http//dx.doi.org/10.1016/j.spasta.2013.08.001.

[77] De Mita, S., Thuillet, A., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., Vigouroux, Y.,(2013). Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Molecular Ecology. **22(5)**,1383-1399.

[78] Conord, C., Lemperiere, G., Taberlet, P., Despres, L.,(2006). Genetic Structure of the forest pest *Hylobius abietis* on conifer plantations at different spatial scales in Europe. Heredity, 97, 46-55.

[79] Chiles, J., Delfiner, P.,(1999).Geostatistics: Modeling Spatial Uncertainty. Wiley, Hoboken, NJ, USA.

[80] Rebaudo, F., Le Rouzic, A., Dupas, S., Silvain, J., Harry, M., Dangles, O.,(2013). SimAdapt: An individual-based genetic model for simulating landscape management impacts on populations. Methods in Ecology and Evolution **4(6)**, 595-600.

[81] Nick J. Patterson., Priya Moorjani, Yonyao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, & David Reich. Ancient admixture in human history. *Genetics, doi:10.1534/genetics.112.145037*,2012.

[82] Francois O., & J., Durand E.,(2010). Spatially explicit bayesian clustering models in population genetics.*Molecular Ecology Resource* 10:773-784.

[83] Pluess A.R., Frank A., Heiri C., Lalagüe H., Vendramin G.G., Oddou-Muratorio S,. Genome-environment association study suggests local adaptation to climate at the regional scale in Fagus sylvatica.

[84] Smith O.L., The influence of enviornmental gradients on ecosystem stability.

[85] http://www.worldclim.org/bioclim.

[86] Xiaoyi Gao^a Eden R. $Martin^b$,Using Allele Sharing Distance for Detecting Human Population Stratification.

[87] Felicity C. Jones, Manfred G. Grabherr[...]David M. Kingsley, The genomic basis of adaptive evolution in threespine sticklebacks

[88] Eric Frichot, François Mathieu, Théo Trouillon, Guillaume Bouchard, and Olivier François. 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics 196:973-983.

[89] Geoff Gordon & Ryan Tibshirani, Karush-Kuhn-Tucker conditions. Optimization 10-725 36-725.

[90] Frichot et al., 2014. Least-squares estimates of ancestry proportions.

[91] Alexander DH and Lange K. 2011. Enhancement to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12:246.

[92] Wold 1978. Cross-validatory estimation of the number of components in factor and principal component models.Technometrics 20:397-405.

[93] Eastment and Krzanowski; 1982. Cross-validatory choice of the number of components from a principal component analysis.Technometrics 24:73-77.

[94] Ziqiang Shi, Rujie Liu, (2016). arXiv:1608.04826 [math.OC].

[95] Jun Li, Daniela M. Witten, Iain M. Johnstone, Robert Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data.

[96] David J. Balding, 2006. A tutorial on statistical methods for population association studies.

6 Appendix

Retained principal components expressed as a linear combination of climatic variables in the current study.

$$\textbf{PC1} = -0.1515 \times tmin1 - 0.1601 \times tmin2 - 0.1705 \times tmin3 - 0.1740 \times tmin4 - 0.1756 \times tmin5 - 0.1664 \times tmin6 - 0.1468 \times tmin7 - 0.1476 \times tmin8 - 0.1619 \times tmin9 - 0.1686 \times tmin10 - 0.1626 \times tmin11 - 0.1504 \times tmin12 - 0.1662 \times tmax1 - 0.1709 \times tmax2 - 0.1660 \times tmax3 - 0.1497 \times tmax4 - 0.1419 \times tmax5 - 0.0969 \times tmax6 - 0.0504 \times tmax7 - 0.0651 \times tmax8 - 0.1296 \times tmax9 - 0.1688 \times tmax10 - 0.1674 \times tmax11 - 0.1624 \times tmax12 + 0.0913 \times prec1 + 0.0929 \times prec2 + 0.1091 \times prec3 + 0.1283 \times prec4 + 0.1599 \times prec5 + 0.16253 \times prec6 + 0.1426 \times prec7 + 0.1277 \times prec8 + 0.1221 \times prec9 + 0.1139 \times prec10 + 0.0990 \times prec11 + 0.0804 \times prec12 - 0.1830 \times bio1 + 0.0309 \times bio2 - 0.0217 \times bio3 + 0.0542 \times bio4 - 0.0568 \times bio5 - 0.1515 \times bio6 + 0.0588 \times bio7 - 0.1416 \times bio8 - 0.1217 \times bio9 - 0.1472 \times bio10 - 0.1714 \times bio11 + 0.1259 \times bio12 + 0.0866 \times bio13 + 0.1430 \times bio14 - 0.0936 \times bio15 + 0.0897 \times bio16 + 0.1552 \times bio17 + 0.1399 \times bio18 + 0.0890 \times bio19.$$

$$\textbf{PC2} = 0.1424 \times tmin1 + 0.1323 \times tmin2 + 0.1118 \times tmin3 + 0.0900 \times tmin4 + 0.0397 \times tmin5 + 0.0196 \times tmin6 + 0.0188 \times tmin7 + 0.0461 \times tmin8 - 0.0691 \times tmin9 + 0.0909 \times tmin10 + 0.1166 \times tmin11 + 0.1482 \times tmin12 + 0.0407 \times tmax1 + 0.0092 \times tmax2 - 0.0448 \times tmax3 - 0.1105 \times tmax4 - 0.1462 \times tmax5 - 0.1891 \times tmax6 - 0.2121 \times tmax7 - 0.1976 \times tmax8 - 0.1218 \times tmax9 - 0.0349 \times tmax10 + 0.0151 \times tmax11 + 0.0408 \times tmax12 + 0.2044 \times prec1 + 0.1924 \times prec2 + 0.1613 \times prec3 + 0.1208 \times prec4 + 0.0531 \times prec5 - 0.0620 \times prec6 - 0.1300 \times prec7 - 0.1303 \times prec8 - 0.0337 \times prec9 + 0.1429 \times prec10 + 0.1947 \times prec11 + 0.2137 \times prec12 + 0.0038 \times bio1 - 0.1986 \times bio2 - 0.0491 \times bio3 - 0.1946 \times bio4 - 0.2085 \times bio5 + 0.1427 \times bio6 - 0.2340 \times bio7 - 0.0528 \times bio8 - 0.0823 \times bio9 - 0.0935 \times bio10 - 0.0934 \times bio11 + 0.1625 \times bio12 + 0.2109 \times bio13 - 0.1289 \times bio14 + 0.1941 \times bio15 + 0.2085 \times bio16 - 0.0989 \times bio17 - 0.0487 \times bio18 + 0.2067 \times bio19.$$

$$\textbf{PC3} = 0.0064 \times tmin1 + 0.0096 \times tmin2 + 0.0119 \times tmin3 + 0.0093 \times tmin4 + 0.0260 \times tmin5 + 0.0223 \times tmin6 + 0.0468 \times tmin7 + 0.0461 \times tmin8 + 0.0294 \times tmin9 - 0.0001 \times tmin10 - 0.0136 \times tmin11 - 0.0104 \times tmin12 - 0.1148 \times tmax1 - 0.1136 \times tmax2 - 0.1470 \times tmax3 - 0.1580 \times tmax4 - 0.1564 \times tmax5 - 0.2139 \times tmax6 - 0.2297 \times tmax7 - 0.2356 \times tmax8 - 0.2335 \times tmax9 - 0.1532 \times tmax10 - 0.1268 \times tmax11 - 0.1250 \times tmax12 - 0.1818 \times prec1 - 0.2151 \times prec2 - 0.2132 \times prec3 - 0.2013 \times prec4 - 0.0730 \times prec5 - 0.0534 \times prec6 - 0.0426 \times prec7 - 0.0459 \times prec8 - 0.0952 \times prec9 - 0.1817 \times prec10 - 0.1459 \times prec11 - 0.1963 \times prec12 - 0.0922 \times bio1 - 0.2197 \times bio2 - 0.1648 \times bio3 - 0.0393 \times bio4 - 0.2308 \times bio5 + 0.0033 \times bio6 - 0.1594 \times bio7 + 0.0461 \times bio8 - 0.1050 \times bio9 - 0.1126 \times bio10 - 0.0626 \times bio11 - 0.1940 \times bio12 - 0.1917 \times bio13 -$$

$0.0424 \times bio14 - 0.0400 \times bio15 - 0.1907 \times bio16 - 0.0509 \times bio17 - 0.0413 \times bio18 - 0.1998 \times bio19$

$\mathbf{PC4} = -0.0522 \times tmin1 - 0.0671 \times tmin2 - 0.0520 \times tmin3 - 0.0914 \times tmin4 - 0.1493 \times tmin5 - 0.2141 \times tmin6 - 0.2828 \times tmin7 - 0.2717 \times tmin8 - 0.2061 \times tmin9 - 0.1020 \times tmin10 - 0.0321 \times tmin11 - 0.0230 \times tmin12 + 0.1524 \times tmax1 + 0.1391 \times tmax2 + 0.1294 \times tmax3 + 0.1175 \times tmax4 + 0.0292 \times tmax5 - 0.0358 \times tmax6 - 0.1085 \times tmax7 - 0.1125 \times tmax8 - 0.0463 \times tmax9 + 0.0658 \times tmax10 + 0.1310 \times tmax11 + 0.1614 \times tmax12 - 0.0570 \times prec1 - 0.0134 \times prec2 + 0.0028 \times prec3 - 0.0044 \times prec4 - 0.1462 \times prec5 - 0.1065 \times prec6 - 0.0337 \times prec7 - 0.0655 \times prec8 - 0.1637 \times prec9 - 0.0972 \times prec10 - 0.0742 \times prec11 - 0.0319 \times prec12 - 0.0470 \times bio1 + 0.2193 \times bio2 + 0.4163 \times bio3 - 0.2837 \times bio4 - 0.1070 \times bio5 - 0.0519 \times bio6 - 0.0395 \times bio7 + 0.0242 \times bio8 - 0.2274 \times bio9 - 0.2165 \times bio10 + 0.0570 \times bio11 - 0.0630 \times bio12 - 0.0221 \times bio13 - 0.0342 \times bio14 - 0.1399 \times bio15 - 0.0412 \times bio16 - 0.0821 \times bio17 - 0.1202 \times bio18 - 0.0350 \times bio19$

$\mathbf{PC5} = 0.1859 \times tmin1 + 0.1082 \times tmin2 + 0.0526 \times tmin3 + 0.0062 \times tmin4 + 0.0292 \times tmin5 - 0.0078 \times tmin6 - 0.0859 \times tmin7 - 0.0811 \times tmin8 + 0.0563 \times tmin9 + 0.1291 \times tmin10 + 0.1713 \times tmin11 + 0.1748 \times tmin12 + 0.0555 \times tmax1 + 0.0131 \times tmax2 + 0.0032 \times tmax3 - 0.0029 \times tmax4 - 0.0083 \times tmax5 - 0.0098 \times tmax6 - 0.0768 \times tmax7 - 0.0851 \times tmax8 + 0.0232 \times tmax9 + 0.0801 \times tmax10 + 0.1259 \times tmax11 + 0.1208 \times tmax12 - 0.0259 \times prec1 - 0.0250 \times prec2 - 0.1131 \times prec3 - 0.0584 \times prec4 - 0.0237 \times prec50.1906 \times prec6 + 0.2499 \times prec7 + 0.2855 \times prec8 + 0.3732 \times prec9 + 0.1024 \times prec10 - 0.1388 \times prec11 - 0.0355 \times prec12 + 0.0514 \times bio1 - 0.0501 \times bio20.1391 \times bio3 - 0.1858 \times bio4 - 0.0744 \times bio5 + 0.1843 \times bio6 - 0.1693 \times bio7 + 0.0274 \times bio8 - 0.0531 \times bio9 - 0.0603 \times bio10 + 0.1167 \times bio11 - 0.0030 \times bio12 - 0.0464 \times bio13 + 0.2514 \times bio14 - 0.1648 \times bio15 - 0.0537 \times bio16 + 0.2445 \times bio170.3599 \times bio18 - 0.0294 \times bio19.$

Table 6: *Correlated predictors at correlation coefficient magnitude threshold of 0.90*

Compare row 37 and column 5 with corr 0.923	
Means	0.645 vs 0.491 so flagging column 37
Compare row 5 and column 4 with corr 0.969	
Means	0.615 vs 0.485 so flagging column 5
Compare row 14 and column 22 with corr 0.931	
Means	0.6 vs 0.48 so flagging column 14
Compare row 22 and column 15 with corr 0.959	
Means	0.592 vs 0.476 so flagging column 22
Compare row 15 and column 23 with corr 0.94	
Means	0.582 vs 0.471 so flagging column 15
Compare row 4 and column 6 with corr 0.922	
Means	0.581 vs 0.467 so flagging column 4
Compare row 23 and column 47 with corr 0.924	
Means	0.557 vs 0.462 so flagging column 23
Compare row 6 and column 9 with corr 0.959	
Means	0.566 vs 0.459 so flagging column 6
Compare row 47 and column 3 with corr 0.946	
Means	0.543 vs 0.454 so flagging column 47
Compare row 3 and column 10 with corr 0.964	
Means	0.536 vs 0.45 so flagging column 3
Compare row 13 and column 24 with corr 0.987	
Means	0.521 vs 0.447 so flagging column 13
Compare row 10 and column 9 with corr 0.955	
Means	0.514 vs 0.443 so flagging column 10
Compare row 16 and column 17 with corr 0.963	
Means	0.526 vs 0.439 so flagging column 16
Compare row 30 and column 53 with corr 0.972	
Means	0.512 vs 0.436 so flagging column 30
Compare row 9 and column 2 with corr 0.91	
Means	0.498 vs 0.432 so flagging column 9
Compare row 11 and column 2 with corr 0.974	
Means	0.456 vs 0.429 so flagging column 11
Compare row 17 and column 21 with corr 0.92	
Means	0.519 vs 0.428 so flagging column 17
Compare row 2 and column 1 with corr 0.981	
Means	0.444 vs 0.424 so flagging column 2
Compare row 53 and column 50 with corr 0.958	
Means	0.453 vs 0.423 so flagging column 53
Compare row 7 and column 8 with corr 0.988	
Means	0.456 vs 0.42 so flagging column 7
Compare row 1 and column 42 with corr 0.999	
Means	0.386 vs 0.42 so flagging column 42
Compare row 1 and column 12 with corr 0.992	
Means	0.367 vs 0.423 so flagging column 12
Compare row 21 and column 18 with corr 0.91	
Means	0.451 vs 0.423 so flagging column 21
Compare row 50 and column 31 with corr 0.999	
Means	0.372 vs 0.424 so flagging column 31
Compare row 48 and column 28 with corr 0.919	
Means	0.545 vs 0.424 so flagging column 48
Compare row 54 and column 33 with corr 0.942	
Means	0.385 vs 0.42 so flagging column 33
Compare row 27 and column 35 with corr 0.9	
Means	0.482 vs 0.419 so flagging column 27
Compare row 18 and column 20 with corr 0.93	
Means	0.466 vs 0.417 so flagging column 18
Compare row 35 and column 25 with corr 0.905	
Means	0.483 vs 0.41 so flagging column 35
Compare row 25 and column 26 with corr 0.966	
Means	0.473 vs 0.406 so flagging column 25
Compare row 26 and column 52 with corr 0.979	
Means	0.441 vs 0.4 so flagging column 26
Compare row 52 and column 55 with corr 0.995	
Means	0.428 vs 0.398 so flagging column 52
Compare row 55 and column 49 with corr 0.988	
Means	0.4 vs 0.396 so flagging column 55
Compare row 49 and column 36 with corr 0.987	
Means	0.371 vs 0.398 so flagging column 36
Compare row 20 and column 41 with corr 0.994	
Means	0.409 vs 0.398 so flagging column 20
Compare row 41 and column 19 with corr 0.998	
Means	0.385 vs 0.399 so flagging column 19

Candidate loci demonstrated in Manhattan plots.