agriculture

MDPI

Article

Exploring the Effect of Climate Factors on SNPs within FHA Domain Genes in Eurasian *Arabidopsis* Ecotypes

Tamer Aldwairi [1,2,*], David J. Chevalier [3,4] and Andy D. Perkins [1]

[1] Department of Computer Science and Engineering, Mississippi State University, Starkville, MS 39762, USA; perkins@cse.msstate.edu
[2] Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA
[3] Department of Biological Sciences, Mississippi State University, Starkville, MS 39762, USA; dchevalier@ega.edu
[4] Department of Biology, East Georgia State College, Swainsboro, GA 30401, USA
* Correspondence: taldwairi@gmail.com or tamer.aldwairi@temple.edu; Tel.: +1-662-313-8462

Abstract: The rapid developments in high-throughput sequencing technologies have allowed researchers to analyze the full genomic sequence of organisms faster and cheaper than ever before. An important application of such advancements is to identify the impact of single nucleotide polymorphisms (SNPs) on the phenotypes and genotypes of the same species by discovering the factors that affect the occurrence of SNPs. The focus of this study is to determine whether climate factors such as the main climate, the precipitation, and the temperature affecting a certain geographical area might be associated with specific variations in certain ecotypes of the plant *Arabidopsis thaliana*. To test our hypothesis we analyzed 18 genes that encode Forkhead-Associated domain-containing proteins. They were extracted from 80 genomic sequences gathered from within 8 Eurasian regions. We used k-means clustering to separate the plants into distinct groups and evaluated the clusters using an innovative scoring system based upon the Köppen-Geiger climate classification system. The methods we used allow the selection of candidate clusters most likely to contain samples with similar polymorphisms. These clusters show that there is a correlation between genomic variations and the geographic distribution of those ecotypes.

Keywords: single nucleotide polymorphisms; *Arabidopsis thaliana*; Köppen-Geiger climate classification system; Fork-head-associated domain

Citation: Aldwairi, T.; Chevalier, D.J.; Perkins, A.D. Exploring the Effect of Climate Factors on SNPs within FHA Domain Genes in Eurasian *Arabidopsis* Ecotypes. *Agriculture* **2021**, *11*, 166. https://doi.org/10.3390/agriculture11020166

Academic Editor: Ritaban Dutta

Received: 10 January 2021
Accepted: 14 February 2021
Published: 18 February 2021

1. Introduction

The potential use of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae) as a model system for genetic studies was first reported by Titova in 1935 [1]. There are many advantages to using *Arabidopsis* as a model in research studies [2] that aim to understand the genetic, cellular, and molecular biological structure of plants. To analyze the sequence variations within *A. thaliana* the 1001 Genomes Project, available at http://1001genomes.org (accessed on 16 June 2020), was launched with the specific goal to discover the whole sequence variation in at least 1001 strains of the reference plant [3–12].

The goal of many association studies is to find genotype differences between and in some cases within certain species and examine how these changes are reflected in the phenotypic characteristics of those species. In the case of *Arabidopsis*, some studies have concentrated on finding phenotype and genotype associations related to alternative splicing and transposable element effects [13–16]. Here, we seek to relate certain genotypic characteristics like SNPs within FHA domain genes to the distribution of those plant ecotypes collected from within different climate regions. Our clustering-based approach, combined with the climatic scoring, represents a unique approach in *A. thaliana* research studies.

We aim to determine whether SNPs appearing in these 18 gene sequences may be related to climate properties at the locations from which samples were collected. Previous studies have addressed similar issues, such as investigating the association between flowering time and clinical variations with latitude [17–19]. The local adaptation in *A. thaliana* loci was also investigated in [20]. The study found that loci related to certain environments show geographic and climatic patterns of adaptation. In another study [21], the authors were able to predict relative fitness among *A. thaliana* accessions that were gathered from different geographic locations but grown together in a similar environment. Their results provide insights that mutations that increase fitness play an important role in the adaptation of *A. thaliana*. In [22], a Bayesian method to identify certain loci correlations with environmental variables was developed through estimating covariance in allele frequencies between populations and use that as a model to test SNPs. The method distinguishes interesting loci for further investigation. While most of the previous approaches support the notion that different environmental factors can have a certain effect on the genotypic characteristics of *A. thaliana*, they differ in the approaches and methodologies used. For example, the different statistical tests used for quantitative traits, such as the ANOVA test, compared to control/case studies that use the chi-square test [23]. There are also differences in the technologies used by those studies such as Affymetrix versus Illumina. These differences affect the final results established by those studies, and the conclusions inferred based upon those results. Our approach here provides a novel way to weigh the association between SNP variations in *A. thaliana* and environmental factors (climate, precipitation, and temperature) affecting those plants using the Köppen-Geiger classification system to measure those factors. The Köppen-Geiger climate classification [24,25] has been used to associate the mapping of mean climate with the ecosystem conditions in certain geographic areas and more recently in identifying potential changes in vegetation over time and climate variability on various temporal scales [26]. Other studies concentrated on the variation in the phenotypic characteristics based on environmental factors [27,28], altitude [29–33], and longtitude [34].

2. Methods

We conducted our study on 80 *A. thaliana* genome sequences taken from the MPI-Cao2010 project [3], hosted under the 1001 Genomes website [10]. These plants were gathered from 8 Eurasian regions: Spain, North Africa, Swabia in the southwest of Germany, South Tyrol in the North of Italy, Southern Italy, Eastern Europe, Caucasus, Southern Russia, and Central Asia. The sequences were generated using Illumina's sequencing-by-synthesis (SBS) technology. At the time when the study was conducted, whole-genome information was only available for those 80 ecotypes.

As a first step in identifying possible associations between identified SNPs and different climate factors in these ecotypes, we decided to focus on a specific gene family that includes 18 genes. These 18 genes encode a protein with a Fork-head associated domain (FHA) domain. A list of AGI codes for the 18 genes is given in the Supplementary Material section. FHA domain is a phosphothreonine binding domain and is usually part of a multi-domain protein [35]. FHA domain is present in bacteria, animals, humans, and plants. It mediates protein-protein interactions controlling biochemical and cellular function in growth and development such as DNA repair or cell-cycle progression [36]. Within the *A. thaliana* genome there are 18 genes that encode proteins with a FHA domain. So far, the function of some of these genes is known. These functions include DNA repair, signal transduction, control of meiosis, development, hormone synthesis, and microRNA biogenesis [11,37–42]. We chose this family of genes because proteins with an FHA domain have important cellular functions in bacteria, animals, and plants. FHA domain is a small protein domain that recognizes phosphothreonine on proteins regulating their activities. Since phosphorylation is widely used protein modification, FHA domain can interact with a wide range of proteins. In this respect we belive that we can possibly identify relations

between the occurrence of SNPs and the climate factors without the need to examine all the genes in the plant.

To identify any association between the different plant ecotypes, we created a binary matrix. Then we portioned the plants into different groups using K-means clustering. We also used hierarchical clustering, but the results did show a significant correlation between the SNPs and the climate factors. Subsequently, we assess the resulting clusters based upon climate data in which those plants originated from. It is informative to look at the process as a whole, since the process we are proposing is a multi-step process. It is befitting here to summarize the steps involved in the process as shown in Figure 1 below, before we explain each step in much more detail.

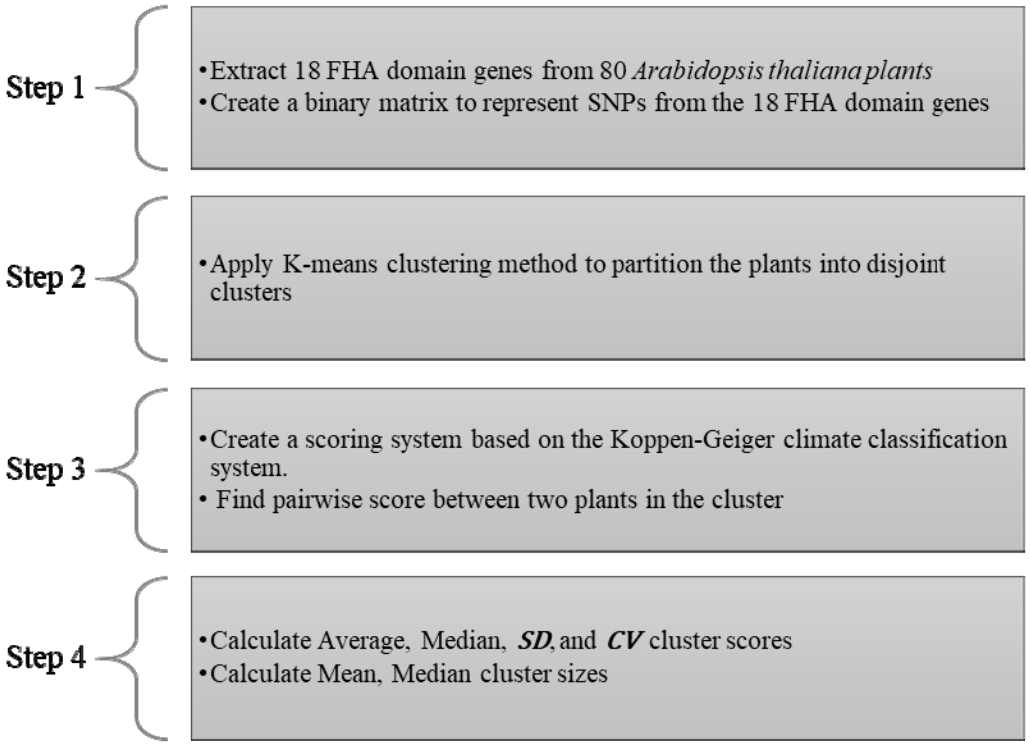

Figure 1. A summary of the steps used to calculate the score and the size of each cluster for the 80 plants at each k-value.

2.1. Analysis at the Tiner Level

We examined the effects of climate on plants at a finer level of granularity by analyzing the temperature and rainfall data for each of the 80 accessions using the closest registered cities data to their locations [43]. We then designed a custom program to calculate the Euclidean distances between any two plants using the temperature and rainfall information. The program calculates the cluster score for temperature and rainfall based on the number of plants within the cluster and the distances between the plants in each cluster for both criteria. The cluster scores are then averaged to obtain the mean temperature and rainfall scores at different k-values. From our analysis, we did not find a clear relation between the distribution of the SNPs within the accessions and the temperature and rainfall at this finer level. This led us to examine relationships between the distribution of SNPs within the plants and the general climate, examining the relationship at a coarser level of granularity.

2.2. Analysis at the Coarser Level

To identify possible climatic associations with the SNPs from the plant ecotypes, we created a binary matrix to represent the SNPs identified in the 18 genes across the 80 accessions. The sequence of each of these accessions was compared to the *A. thaliana*

reference sequence [2], and if the nucleotide in a particular position in the sample did not match the reference, a 1 was placed in the matrix and a 0 otherwise.

We then applied the k-means clustering method [44] implemented in the R-language for statistical computing [45] to the binary matrix to partition the plants into disjoint clusters. Other clustering methods were also investigated (i.e., hierarchical clustering) before choosing k-means as our method of choice.

The resulting clusters represent a partitioning of the plants based on occurrences of SNPs at a similar set of locations. A custom scoring system based upon the Köppen climate classification [46] properties was designed to determine whether cluster members tended to share climates that are similar or closely related to each other. The Köppen climate classification system divides the land regions based upon the vegetation appearing in various regions, along with different aspects of temperature and precipitation to describe the different climates in the world.

The climate distribution is described by three factors. The first factor is the main climate classification, which may be one of the main five climate groups such as equatorial (A), arid (B), warm temperature (C), snow (D), and polar (E). The second is the seasonal precipitation which takes on values of desert (W), steppe (S), fully humid (f), summer-dry (s), winter dry (w), and monsoonal (m). The third is temperature or level of heat, which can be hot arid (h), cold arid (k), hot summer (a), warm summer (b), cool summer (c), extremely continental (d), polar frost (F), and polar tundra (T). Figure 2 shows the locations of the plants in each country and region.

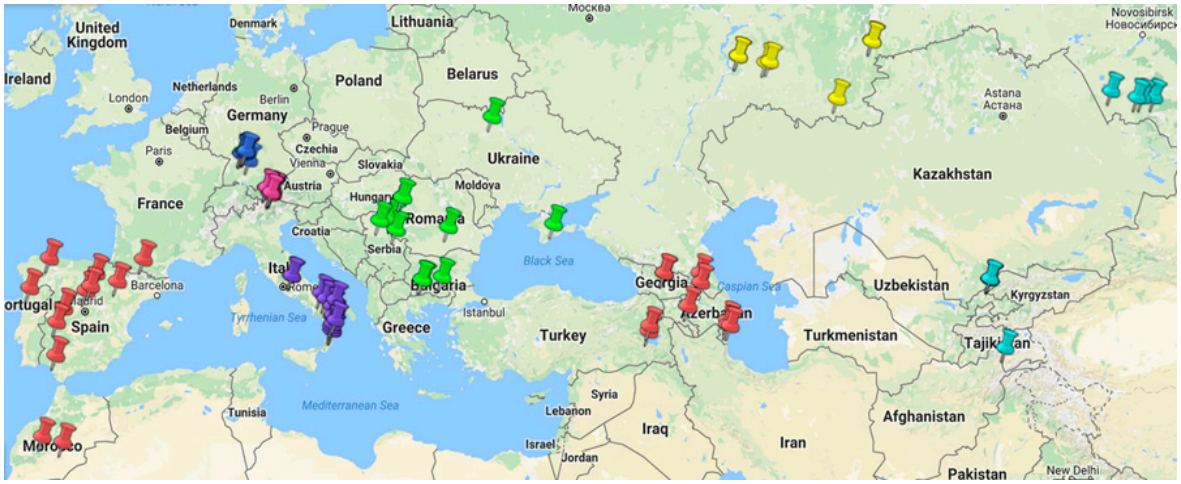

Figure 2. The distribution of the plants across different regions.

Our scoring system was built by assigning a weight point method to every individual letter in the system based on the three factors that make up the Köppen-Geiger climate classification system. A two-point weight difference was assigned for each letter representing one of the main climates. A one-point difference between each precipitation and half a point difference between each temperature classification. Considering the three factors mentioned above, for each category, the closer these factors are to each other, the lesser the score difference there will be between them.

An important thing to note here is that the assignments of different weights to the different factors are not random but are founded upon the extent to which those factors affect the climate and the plants inhabiting each region. The main climate has the highest influence on the plants since even the precipitation and temperature can vary up to a certain degree within each climate type. For such reason, we allocate for the main climate a higher weight, in this case, two points for each change in the main climate. We also applied as shown in Table 1 a various number of weights for precipitation and temperature.

Table 1. Scores assigned for each climate factor.

Main Climate	Precipitation		Temperature	
A = 2	W = 1	m = 6	h = 0.5	d = 3
B = 4	S = 2	_ = 7	k = 1	F = 3.5
C = 6	f = 3		a = 1.5	T = 4
D = 8	s = 4		b = 2	
E = 10	w = 5		c = 2.5	

When we decided how to factor the weights of temperature and precipitation, we considered that for each category in the precipitation category there are several different temperatures. Therefore, we apply a higher weight for the precipitation factor compared to the temperature factor. One can also argue that the value given to each factor in the process does not represent the exact level of significance than the factor has on the overall climate. This is a valid argument, but it cannot stand on a strong basis since it is extremely hard to get an exact measure of the level of contribution each specific category has on the overall process could be difficult or even impossible to measure. The point of the weight mechanism is to give a rough estimate regarding how important each factor is impacting the entire process and not the exact level each factor contributes to the process. Table 1 shows the individual letter score assigned to each of the three factors.

To find out how close every pair of plants that are grouped together in a certain cluster, we took the sum of the absolute values for the difference between each of the three factors (climate, precipitation, temperature) for those two plants. This is noticeably expressed in the numerator part of Equation (1) To show how we calculate the score between two plants let's consider this example. If we take one plant from a region classified in the "BSk" climate and another plant from a nearby region but with a different climate, let's say "Csa". Using the proposed climate scoring method, the difference in the first factor is one level (B = 4 points and C = 6 points) which is equal to two points. For the second factor, there is a difference of two levels (S = 2 points and s = 4 points) which gives us a total of two points. For the third factor, it is a one-level difference (k = 1 points and **a** = 1.5 points) which is worth 0.5 points. If we take the sum of all the scores (2, 2, and 0.5) for each of the factors the total will be equal to 4.5 points.

We repeat this process for every plant grouped in the same cluster which means that every plant will have a certain score that reflects the proximity level of all the plants in each cluster. To further illustrate this point let us assume we have a cluster that contains eight plants. This suggests that every plant will have seven different scores. Each of the scores is a representation of how close those factors we are measuring in each plant is, in comparison to the other seven plants grouped in the same cluster. Then we calculate the mean pairwise climate score (i.e., average cluster score) for that cluster by adding all the pairwise scores for all the plants in the cluster and dividing the total score by the number of plants in the cluster using the following equation:

$$\Sigma_{i=1}^{n} \frac{(|\boldsymbol{C}_{i2} - \boldsymbol{C}_{i1}| + |\boldsymbol{P}_{i2} - \boldsymbol{P}_{i1}| + |\boldsymbol{T}_{i2} - \boldsymbol{T}_{i1}|)}{\boldsymbol{n}} \quad (1)$$

Here ($\boldsymbol{C}$) represents the main climate, ($\boldsymbol{P}$) the precipitation, and ($\boldsymbol{T}$) the temperature. ($\boldsymbol{n}$) Indicates the number of comparisons between the plants within the same cluster and ($i1$) and ($i2$) represents the two plants that are compared. Generally speaking, the lower the mean climate score of a certain cluster, the more similar those ecotype climates are to each other.

Since the k-means algorithm selects random cluster centers at the beginning of the clustering process, clusters produced may be different for subsequent runs at the same k value. Therefore, the clustering process was repeated five times for each k value to avoid bias caused by individual runs.

To find the overall mean score (i.e., K-score) for that run at a certain k value we sum all the average cluster scores for all the plants within each cluster divided by the number of clusters as shown in Equation (2). For example, if we have three plants taken from three different climates BSk, Csa, and Dsb. Referring back to the example we mentioned earlier we found out that the score difference between BSk and Csa is four and half points. Using the same approach, we can calculate the score difference between BSk and Dsb which happen to be 7 points, and for Csa and Dsb 2.5 points. The K-score will be the sum of all the scores divided by three which is approximately equal to 4.67 points:

$$\Sigma_{j=1}^{N} \frac{\Sigma_{i=1}^{n} \frac{(|C_{i2}-C_{i1}|+|P_{i2}-P_{i1}|+|T_{i2}-T_{i1}|)}{n}}{N} \quad (2)$$

To account for outliers in the score, we calculated the median, and to ensure the quality of our results and that our scores are fairly dispersed around the mean, we calculated the standard deviation (SD) and the coefficient of variation (CV). This will assist in finding the degree of disparity in the results across all the average scores.

We notice that the (SD) for all k-mean runs is small. This implies, that the results from the various k-mean runs are not dependent upon the randomness inherent in the k-means algorithm and hence, the results are highly repeatable. Table 2 shows each specific K-score, their means, standard deviations, and medians at different k-values.

Table 2. K-scores at different k-values.

K.	Run 1	Run 2	Run 3	Run 4	Run 5	Mean	Median	SD	CV (%)
5	2.06	2.48	2.72	1.38	2.29	2.18	2.29	0.51	23.35
6	1.45	1.25	2.14	1.25	2.02	1.62	1.45	0.43	26.51
7	2.26	2.32	2.06	1.81	1.31	1.95	2.06	0.41	21.01
8	1.55	1.55	2.27	2.31	2.40	2.02	2.27	0.43	21.33
9	2.12	1.61	1.42	1.81	1.89	1.77	1.81	0.27	15.26
10	1.71	1.94	2.11	1.38	1.79	1.79	1.79	0.27	15.11
11	1.69	2.03	1.62	1.47	1.54	1.67	1.62	0.22	13.16
12	1.75	1.76	1.26	2.10	1.39	1.65	1.75	0.33	20.00
13	1.57	1.99	1.94	1.18	1.24	1.58	1.57	0.38	24.01
14	1.72	1.54	1.29	1.10	1.44	1.42	1.44	0.24	16.96
15	1.11	1.07	1.17	1.75	1.58	1.34	1.17	0.31	23.19
16	1.00	1.64	1.92	1.36	1.19	1.42	1.36	0.37	26.01
17	1.44	1.57	1.09	1.30	0.97	1.27	1.30	0.24	18.84
18	1.27	1.18	1.09	1.14	1.34	1.20	1.18	0.10	8.31
19	1.16	1.51	1.38	1.21	1.06	1.26	1.21	0.18	14.27
20	0.94	1.00	0.98	1.36	1.38	1.13	1.00	0.22	19.47

The table shows the average climate score for each cluster at different values of k for each of the 5 runs of the k-means algorithm.

3. Results

The mean, SD, median, and CV for each individual K-score all assist in finding a stable k-value for the clustering. A stable k-value is likely to hold a low average score for all four criteria, in comparison with the other runs, or at least, for three of the criteria used with the fourth one not exhibiting a very high score value. A low score for a cluster marks evidence of similarity among the plants grouped within the same cluster. The lower the score the closer the plants grouped in the same cluster are to each other. A score of zero means that all the plants grouped in that cluster are identical in all three criteria. This could happen sometimes to one group out of all the k-groups within the cluster. We observed

this happening to two plants that were taken from the ET region (those were the only two plants from that region) where they were grouped together in one cluster during some runs. Calculating the median is necessary to make sure that the mean K-scores are not influenced by the existence of outliers.

The K-score represents how far the cluster scores are apart from each other for each individual run. The lower the score, the more uniform is the grouping of the plants into one cluster. Examining all the k-values from 5 to 20 overall of the 5 runs, we noticed that the maximum score over all the runs was 2.72 points, and the mean climate scores for all the runs are at most 2.18 and the highest median was 2.29 which are low scores. This reflects the similarity between the plants grouped in each cluster.

We calculated the quartiles and inner and outer fences as shown in Table 3 to check for the presence of major or minor outliers. All the scores were in the inner fence's range which denotes the absence of any outliers in the scores. Examining the average (mean) and the median scores in Figure 3 below we noticed that the scores are generally very close and, in some cases, nearly identical. This suggests along with the absence of any outliers that the clustering of the values was more uniform and that the plants grouped in the same cluster are usually close to each other regarding the factors that we are measuring.

Table 3. Quartiles, inner and outer fence ranges.

First Quartile	1.25	
Second Quartile	1.52	
Third Quartile	1.91	
IQR	0.67	
Inner Fence	0.25	2.91
Outer Fence	−0.75	3.91

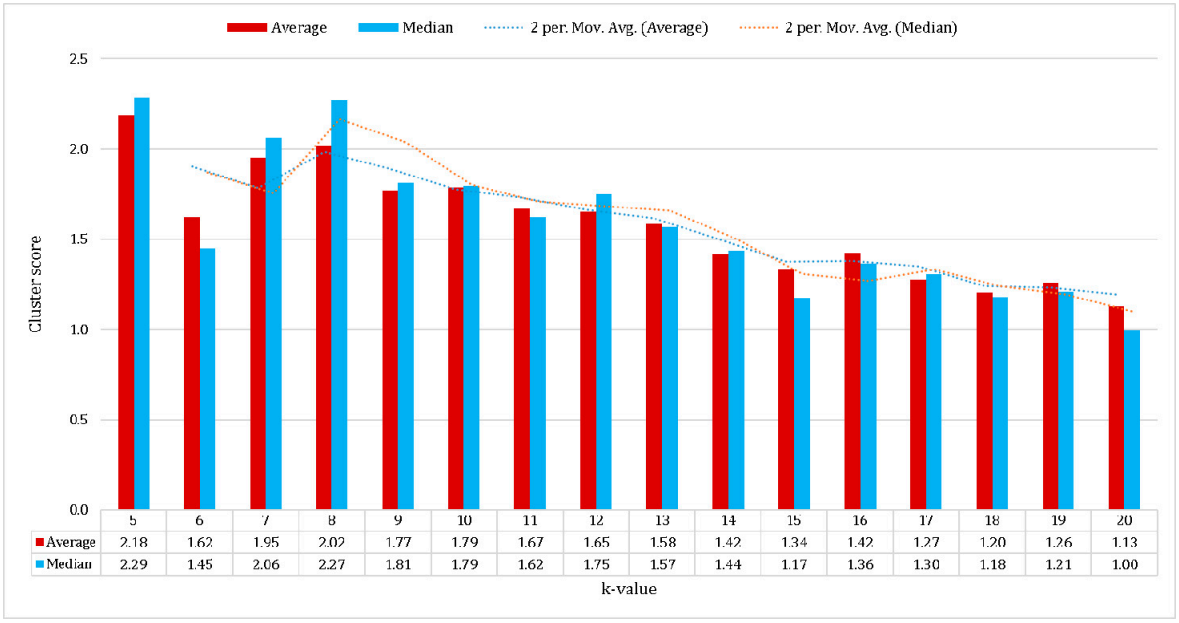

Figure 3. The average and median scores at different k-values.

4. Discussion

To identify which value of k-clustering of the plant ecotypes groups has a stable score we need to examine our results for the (SD) and the (CV) scores at each k-value. A lower score for the SD and the CV is a strong indication that the plants within each cluster group

have scores that are close to each other. We demonstrate our point with a simple example where we choose the k-value to be equal to 10. Let us assume that the plants were equally distributed across all 10 clusters with 8 plants in each cluster. If we take the first two clusters denoted by C1 and C2 for cluster 1 and cluster 2, respectively. If the average pairwise difference between all the plant's scores in C1 is less than that of C2 then this an indication that the eight plants grouped in C1 are much more similar in respect to the factors we are measuring when compared to C2.

If the plants grouped together exhibit a low score then this could be an indication that there is a correlation between the presence of SNPs within their FHA domain genes and the climatic factors that affect those plants in different geographic areas.

This can be observed through finding a stable k-value which can be defined as a value for k with both low SD and CV. Examining the (SD) and (CV)climate scores in Figure 4 we find that the k-value scores for SD and CV generally exhibit low values starting from k = 9 up until k = 20 with the lowest three scores in both categories at k = 18 with SD = 0.1 and CV = 8.31, at k = 11 with SD = 0.22 and CV = 13.16 and at k = 19 with SD = 0.18 and CV = 14.27. Those clusters contain ecotypes gathered from locations that are affected by similar climates which leads us to believe that the plant ecotypes with FHA domain genes might hold key information in deciding if these plants developed in very similar climates. The low values of SD and CV are a strong indication that the clusters at these values are stable and therefore these clusters contain plants that share a similar or a closely related climate and since these clusters were created using the SNPs found in the 18 FHA domain genes of those plants. It follows that the presence of specific variations in those 18 genes contributed significantly to the correlation between genomic variations and the geographic distribution of those ecotypes.

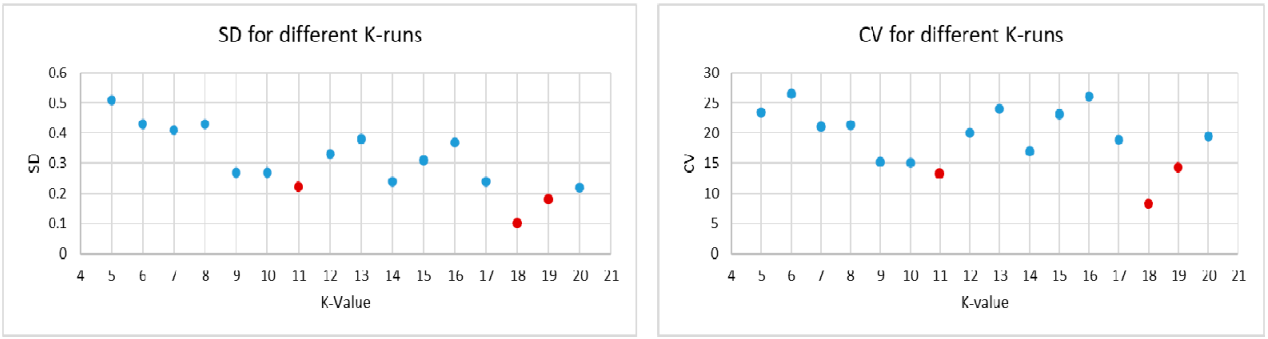

Figure 4. The SD and CV values for difference K-runs the red dots indicate k-values that exhibit both low SD and CV values.

5. Conclusions

In this paper, we investigated possible associations between the distribution of plant ecotypes in various locations and SNPs appearing in these ecotypes within 18 FHA domain genes. SNPs data were clustered using the k-means clustering method and assigned a score using an array of climate factors for a range of k values. Based upon our analysis of these values, at different runs, and examining the K-run mean, median, (SD), and (CV) for the climate scores along with, the average and average median cluster sizes of those clusters. We established that using k-means clustering to find an association relationship between plant ecotypes at different locations and SNPs appearing in those ecotypes is generally stable for most k-values starting at k = 9 with k = 18, 11, and 19 possessing the most stable SD and CV clustering values. Our analysis shows that there is a correlation between the presence of SNPs in FHA domain genes in *A. thaliana* ecotypes and their geographic distribution. Future work includes utilizing machine learning to build models for predicting the association between SNPs and the distribution of plant ecotypes in various locations. In addition, we will continue to investigate the probability that the clusters at these k-values can effectively group plants sharing the same climate characteristics.

Supplementary Materials: The following are available online at https://www.mdpi.com/2077-0472/11/2/166/s1, Refer to supplementary File S1 for a list of AGI codes of the 18 FHA Domain Genes.

Author Contributions: Conceptualization, T.A., D.J.C. and A.D.P.; methodology, T.A., D.J.C. and A.D.P.; software, T.A.; investigation, T.A.; writing—original draft preparation, T.A.; writing—review and editing, T.A., D.J.C. and A.D.P.; visualization, T.A.; supervision, A.D.P.; project administration, A.D.P.; funding acquisition, A.D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation under grants EPS-0903787 and EPS-1006883.

Data Availability Statement: Data analyzed was retrieved from the publicly-available data repository at http://www.1001genomes.org/ (accessed on 16 June 2020).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Titova, N.N. In search for plant Drosophila. *Sov. Bot.* **1935**, *2*, 61–67. (In Russian)
2. Initiative, T.A.G. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nat. Cell Biol.* **2000**, *408*, 796–815. [CrossRef]
3. Cao, J.; Schneeberger, K.; Ossowski, S.; Günther, T.; Bender, S.; Fitz, J.; Koenig, D.; Lanz, C.; Stegle, O.; Lippert, C.; et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **2011**, *43*, 956–963. [CrossRef]
4. Long, Q.; A Rabanal, F.; Meng, D.; Huber, C.D.; Farlow, A.; Platzer, A.; Zhang, Q.; Vilhjálmsson, B.J.; Korte, A.; Nizhynska, V.; et al. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat. Genet.* **2013**, *45*, 884–890. [CrossRef]
5. Schmitz, R.J.; Schultz, M.D.; Urich, M.A.; Nery, J.R.; Pelizzola, M.; Libiger, O.; Alix, A.; McCosh, R.B.; Chen, H.; Schork, N.J.; et al. Patterns of population epigenomic diversity. *Nat. Cell Biol.* **2013**, *495*, 193–198. [CrossRef] [PubMed]
6. Schneeberger, K.; Ossowski, S.; Ott, F.; Klein, J.D.; Wang, X.; Lanz, C.; Smith, L.M.; Cao, J.; Fitz, J.; Warthmann, N.; et al. Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 10249–10254. [CrossRef] [PubMed]
7. Gan, X.; Stegle, O.; Behr, J.; Steffen, J.G.; Drewe, P.; Hildebrand, K.L.; Lyngsoe, R.; Schultheiss, S.J.; Osborne, E.J.; Sreedharan, V.T.; et al. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nat. Cell Biol.* **2011**, *477*, 419–423. [CrossRef]
8. Alonso-Blanco, C.; Andrade, J.; Becker, C.; Bemm, F.; Bergelson, J.; Borgwardt, K.M.; Cao, J.; Chae, E.; Dezwaan, T.M.; Ding, W.; et al. 1135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **2016**, *166*, 481–491. [CrossRef] [PubMed]
9. Ossowski, S.; Schneeberger, K.; Clark, R.M.; Lanz, C.; Warthmann, N.; Weigel, D. Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.* **2008**, *18*, 2024–2033. [CrossRef] [PubMed]
10. 1001 Genomes. Available online: https://1001genomes.org/ (accessed on 16 June 2020).
11. Morris, E.R.; Chevalier, D.; Walker, J.C. DAWDLE, a Forkhead-Associated Domain Gene, Regulates Multiple Aspects of Plant Development. *Plant Physiol.* **2006**, *141*, 932–941. [CrossRef]
12. Kawakatsu, T.; Huang, S.-s.C.; Jupe, F.; Sasaki, E.; Schmitz, R.J.; Urich, M.A.; Castanon, R.; Nery, J.R.; Barragan, C.; He, Y.; et al. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell* **2016**, *166*, 492–505. [CrossRef]
13. Zhang, X.; Wessler, S.R. Genome-wide comparative analysis of the transposable elements in the related species Arabidopsis thaliana and Brassica oleracea. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5589–5594. [CrossRef]
14. Filichkin, S.A.; Priest, H.D.; Givan, S.A.; Shen, R.; Bryant, D.W.; Fox, S.E.; Wong, W.-K.; Mockler, T.C. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res.* **2009**, *20*, 45–58. [CrossRef]
15. Filiault, D.L.; Maloof, J.N. A Genome-Wide Association Study Identifies Variants Underlying the Arabidopsis thaliana Shade Avoidance Response. *PLoS Genet.* **2012**, *8*, e1002589. [CrossRef] [PubMed]
16. Atwell, S.; Huang, Y.S.; Vilhjálmsson, B.J.; Willems, G.; Horton, M.W.; Li, Y.; Meng, D.; Platt, A.; Tarone, A.M.; Hu, T.T.; et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nat. Cell Biol.* **2010**, *465*, 627–631. [CrossRef] [PubMed]
17. Kurbidaeva, A.S.; Zaretskaya, M.V.; Soltabaeva, A.D.; Novokreshchenova, M.G.; Kupriyanova, E.V.; Fedorenko, O.M.; Ezhova, T.A. Genetic Base of Arabidopsis thaliana (L.) Heynh: Fitness of Plants for Extreme Conditions in Northern Margins of Species Range. Генетика **2013**, *49*, 943–952. [CrossRef]
18. Stinchcombe, J.R.; Weinig, C.; Ungerer, M.; Olsen, K.M.; Mays, C.; Halldorsdottir, S.S.; Purugganan, M.D.; Schmitt, J. A latitudinal cline in flowering time in Arabidopsis thaliana modulated by the flowering time gene FRIGIDA. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4712–4717. [CrossRef]
19. Lewandowska-Sabat, A.M.; Fjellheim, S.; Rognli, O.A. The continental-oceanic climatic gradient impose clinal variation in vernalization response in Arabidopsis thaliana. *Environ. Exp. Bot.* **2012**, *78*, 109–116. [CrossRef]

20. Fournier-Level, A.; Korte, A.; Cooper, M.D.; Nordborg, M.; Schmitt, J.; Wilczek, A.M. A Map of Local Adaptation in Arabidopsis thaliana. *Science* **2011**, *334*, 86–89. [CrossRef] [PubMed]
21. Hancock, A.M.; Brachi, B.; Faure, N.; Horton, M.W.; Jarymowycz, L.B.; Sperone, F.G.; Toomajian, C.; Roux, F.; Bergelson, J. Adaptation to Climate Across the Arabidopsis thaliana Genome. *Science* **2011**, *334*, 83–86. [CrossRef]
22. Coop, G.; Witonsky, D.; Di Rienzo, A.; Pritchard, J.K. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genet.* **2010**, *185*, 1411–1423. [CrossRef]
23. Bush, W.S.; Moore, J.H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* **2012**, *8*, e1002822. [CrossRef]
24. Beck, H.E.; Zimmermann, N.E.; McVicar, T.R.; Vergopolan, N.; Berg, A.; Wood, E.F. Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Sci. Data* **2018**, *5*, 180214. [CrossRef]
25. Naranjo, L.; Glantz, M.H.; Temirbekov, S.; Ramírez, I.J. El Niño and the Köppen–Geiger Classification: A Prototype Concept and Methodology for Mapping Impacts in Central America and the Circum-Caribbean. *Int. J. Disaster Risk Sci.* **2018**, *9*, 224–236. [CrossRef]
26. Chen, D.; Chen, H.W. Using the Köppen classification to quantify climate variation and change: An example for 1901–2010. *Environ. Dev.* **2013**, *6*, 69–79. [CrossRef]
27. El-Soda, M.; Malosetti, M.; Zwaan, B.J.; Koornneef, M.; Aarts, M.G. Genotype × environment interaction QTL mapping in plants: Lessons from Arabidopsis. *Trends Plant Sci.* **2014**, *19*, 390–398. [CrossRef] [PubMed]
28. Provart, N.J.; Alonso, J.; Assmann, S.M.; Bergmann, D.C.; Brady, S.M.; Brkljacic, J.; Browse, J.; Chapple, C.; Colot, V.; Cutler, S.R.; et al. 50 years of Arabidopsis research: Highlights and future directions. *New Phytol.* **2016**, *209*, 921–944. [CrossRef]
29. Wolfe, M.D.; Tonsor, S.J. Adaptation to spring heat and drought in northeastern SpanishArabidopsis thaliana. *New Phytol.* **2014**, *201*, 323–334. [CrossRef]
30. Vidigal, D.S.; Marques, A.C.S.S.; Willems, L.A.J.; Buijs, G.; Méndez-Vigo, B.; Hilhorst, H.W.M.; Bentsink, L.; Picó, F.X.; Alonso-Blanco, C. Altitudinal and climatic associations of seed dormancy and flowering traits evidence adaptation of annual life cycle timing inArabidopsis thaliana. *Plant Cell Environ.* **2016**, *39*, 1737–1748. [CrossRef] [PubMed]
31. Singh, A.; Tyagi, A.; Tripathi, A.M.; Gokhale, S.M.; Singh, N.; Roy, S. Morphological trait variations in the west Himalayan (India) populations of Arabidopsis thaliana along altitudinal gradients. *Curr. Sci.* **2015**, *108*, 2213–2222.
32. Botto, J.F. Plasticity to simulated shade is associated with altitude in structured populations ofArabidopsis thaliana. *Plant Cell Environ.* **2015**, *38*, 1321–1332. [CrossRef]
33. Luo, Y.; Widmer, A.; Karrenberg, S. The roles of genetic drift and natural selection in quantitative trait divergence along an altitudinal gradient in Arabidopsis thaliana. *Heredity* **2014**, *114*, 220–228. [CrossRef]
34. Brachi, B.; Meyer, C.G.; Villoutreix, R.; Platt, A.; Morton, T.C.; Roux, F.; Bergelson, J.; Zhang, X.; Gui, L.; Zhang, X.; et al. Coselected genes determine adaptive variation in herbivore resistance throughout the native range of Arabidopsis thaliana. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 4032–4037. [CrossRef]
35. Durocher, D.; Henckel, J.; Fersht, A.R.; Jackson, S.P. The FHA Domain Is a Modular Phosphopeptide Recognition Motif. *Mol. Cell* **1999**, *4*, 387–394. [CrossRef]
36. Scolnick, D.M.; Halazonetis, T.D. Chfr defines a mitotic stress checkpoint that delays entry into metaphase. *Nat. Cell Biol.* **2000**, *406*, 430–435. [CrossRef]
37. D'Erfurth, I.; Jolivet, S.; Froger, N.; Catrice, O.; Novatchkova, M.; Simon, M.; Jenczewski, E.; Mercier, R. Mutations in AtPS1 (Arabidopsis thaliana Parallel Spindle 1) Lead to the Production of Diploid Pollen Grains. *PLoS Genet.* **2008**, *4*, e1000274. [CrossRef] [PubMed]
38. Akutsu, N.; Iijima, K.; Hinata, T.; Tauchi, H. Characterization of the plant homolog of Nijmegen breakage syndrome 1: Involvement in DNA repair and recombination. *Biochem. Biophys. Res. Commun.* **2007**, *353*, 394–398. [CrossRef]
39. Lee, S.-Y.; Kim, H.; Hwang, H.-J.; Jeong, Y.-M.; Na, S.H.; Woo, J.-C.; Kim, S.-G. Identification of Tyrosyl-DNA Phosphodiesterase as a Novel DNA Damage Repair Enzyme in Arabidopsis. *Plant Physiol.* **2010**, *154*, 1460–1469. [CrossRef] [PubMed]
40. Stone, J.M.; A Collinge, M.; Smith, R.D.; A Horn, M.; Walker, J.C. Interaction of a protein phosphatase with an Arabidopsis serine-threonine receptor kinase. *Science* **1994**, *266*, 793–795. [CrossRef]
41. Xiong, L.; Lee, H.; Ishitani, M.; Zhu, J.-K. Regulation of Osmotic Stress-responsive Gene Expression by theLOS6/ABA1 Locus inArabidopsis. *J. Biol. Chem.* **2002**, *277*, 8588–8596. [CrossRef] [PubMed]
42. Yu, B.; Bi, L.; Zheng, B.; Ji, L.; Chevalier, D.; Agarwal, M.; Ramachandran, V.; Li, W.; Lagrange, T.; Walker, J.C.; et al. The FHA domain proteins DAWDLE in Arabidopsis and SNIP1 in humans act in small RNA biogenesis. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 10073–10078. [CrossRef] [PubMed]
43. Home | World Weather Information Service. Available online: http://worldweather.wmo.int/en/home.html (accessed on 16 June 2020).
44. Macqueen, J. Some methods for classification and analysis of multivariate observations, 5-TH BERKELEY Symp. *Math. Stat. Probab.* **1967**, *1*, 281–297. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.308.8619 (accessed on 30 June 2019).
45. Venables, W.N.; Ripley, B.D. Package MASS. Available online: http://www.r-project.org (accessed on 16 June 2020).
46. Peel, M.C.; Finlayson, B.L.; McMahon, T.A. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* **2007**, *11*, 1633–1644. [CrossRef]