agronomy

MDPI

Article

Genetic Diversity, Linkage Disequilibrium, and Population Structure in a Common Bean Reference Collection

Daniel Ambachew [1], Jorge Mario Londoño [2], Nohra Rodriguez Castillo [3], Asrat Asfaw [4] and Matthew Wohlgemuth Blair [1,*]

[1] Department of Agricultural Sciences, Tennessee State University, Nashville, TN 37209, USA; dambachew@tnstate.edu
[2] Grupo Investigación Biodiversidad y Biotecnología (GIBUQ), Universidad de Quindío, Armenia 630001, Colombia; jmlondono@uniquindio.edu.co
[3] Departamento de Biología, Universidad del Valle, Cali 760032, Colombia; nohra.rodriguez@correounivalle.edu.co
[4] International Institute for Tropical Agriculture (IITA), Kano 700102, Nigeria; a.asfaw@cgiar.org
* Correspondence: mblair@tnstate.edu

Abstract: An in-depth understanding of the extent and pattern of genetic diversity and population structure in crop populations is of paramount importance for any crop improvement program to efficiently promote the translation of genetic diversity into genetic gain. A reference collection of 150 common bean genotypes selected from the International Center for Tropical Agriculture's global core collection was evaluated using single-nucleotide polymorphism (SNP) markers to quantify the amount of genetic diversity, linkage disequilibrium, and population structure. The cultivars and landraces of the collection were diverse and originated from 14 countries, and wild accessions were used as controls for each gene pool. The collection was genotyped using an SNP array, generating a total of 5398 locus calls distributed across the entire bean genome. The SNP data quality was checked, and two datasets were generated. The first dataset (Dataset_1) comprised a set of 5108 SNPs and 150 genotypes after filtering for 10% missing alleles and an MAF < 0.05. The second dataset (Dataset_2) comprised a set of 2300 SNPs that remained after removing any null-allele SNPs and LD pruning for a criterion of $r^2 < 0.2$. Dataset_1 was used for a principal coordinate analysis (PCoA), phylogenetic relationship determination, an analysis of molecular variance (AMOVA), and a discriminant analysis of principal components. Dataset_2 was used for a population structure analysis using STRUCTURE software and is proposed for a genome-wide association study (GWAS). The population structure analysis split the reference collection into two subpopulations according to an Andean or Mesoamerican gene pool. The Mesoamerican populations displayed higher genetic differentiation and tended to split into more groups that were somewhat aligned with common bean races. Andean beans were characterized by a larger average LD but lower LD percentage, a small average genetic distance between members of the population, and a higher major allele frequency, which suggested narrower genetic diversity compared to the Mesoamerican gene pool. In conclusion, the results indicated the presence of high genetic diversity, which is useful for a GWAS. However, the presence of significant linkage disequilibrium requires that genetic distance be considered as a co-factor for any further genetic studies. Overall, the molecular variation observed in the genotypes shows that this reference collection is valuable as a genebank-derived diversity panel which is useful for marker trait association studies.

Keywords: analysis of molecular variance; discriminant analysis of principal components; GWAS population; principal coordinate analysis

check for updates

Citation: Ambachew, D.; Londoño, J.M.; Rodriguez Castillo, N.; Asfaw, A.; Blair, M.W. Genetic Diversity, Linkage Disequilibrium, and Population Structure in a Common Bean Reference Collection. *Agronomy* **2024**, *14*, 985. https://doi.org/10.3390/agronomy14050985

Academic Editor: Shikai Hu

Received: 6 March 2024
Revised: 7 April 2024
Accepted: 25 April 2024
Published: 8 May 2024

1. Introduction

The common bean (*Phaseolus vulgaris* L.) is one of the principal grain legumes in the agri-food system. Common beans originated in a north–south geographical arc from present-day Mexico to Central America and the Andes Mountains [1,2]. Common beans were domesticated in Mesoamerica near the central volcanic axis and in the Andes Mountains of Peru and northwestern Argentina [3]; the Lerma-Santiago basin of Mexico is a possible site from 7000 to 8000 years ago. The domestication of common beans took place in wild, ancestral gene pools that had already diverged and were distributed from Northern Mexico to Colombia (the Mesoamerican gene pool) and from Colombia to Argentina (the Andean gene pool) [4].

Today, the crop is widely cultivated on all major continents except Antarctica, ranging in latitude from 52° N to 32° S and from sea level in the United States and Europe to elevations of more than 3000 m in Andean South America [5]. The common bean is cultivated for its edible mature or immature seeds as well as green pods as a popular vegetable [6]. Its green leaves are sometimes used for human consumption [7] and in some regions of the world, its straw is also used as fodder for animals [8]. As they are protein-rich, common bean seeds play an especially significant role in the human diet. The dry bean crop provides nutrients such as proteins, carbohydrates, mineral elements (iron and zinc), and vitamins (e.g., folate) which are vital in the human diet [9].

As the crop plays an important role in the sociocultural and economic settings of many communities in the world, bean improvement programs are active across the globe [2,10,11]. Reviewing and assessing genetic diversity within a plant species is a pivotal step in bean improvement programs, which range from choosing cross-parents for hybridization blocks and determining a methodology for the selection of multiple segregating generations to determining the best management and maintenance practices for important common bean germplasm collections, whether nationally or internationally [12].

Since the early 1980s, molecular genetic markers have been used to explore the genetic diversity, center of origin, and domestication of common beans. For example, Brown et al. [13] identified three groups of common beans based on their globulin-1 (G1) polypeptide protein subunit composition. In another example, Gepts et al. [1] used variations in Phaseolin seed storage protein to show evidence for different centers of domestication. Other early molecular marker used were allozymes, which also helped us understand the genetic structure of common beans [14]. Since then, many studies have been conducted on common bean genetic diversity, especially since the advent of modern markers for the crop, such as simple sequence repeats (SSRs) and single-nucleotide polymorphisms (SNPs). These have been used to study overall population structure [4,15] and have become the most frequently used markers in the last decade for assessing the population structure and genetic divergence of common beans, as well as providing better analyses of its centers of domestication, diversity, and origin (for example, Cichy et al. [16]).

However, most recent studies were conducted on germplasm collections from a specific country, region, or continent. For example, the Mesoamerican diversity panel (MDP) was collected from North American bean-breeding programs [17]. Similarly, the Andean bean diversity panel (ADP) was mainly collected from inter-related breeding programs in Africa, the Caribbean, and North America [16]. Furthermore, the MDP was entirely a collection of commercial cultivars, while the ADP included breeding lines, land races, and commercial cultivars.

Our objective was to create a reference collection diversity panel from common beans in an international genebank held by the CGIAR network and duplicated in the Svalbard Vault which will be useful in perpetuity. Our study generates information on the linkage disequilibrium, genetic diversity, and population structure of this collection with a widely used set of 5393 SNPs. In this way, we predict the utility of the collection for future genome-wide genotype × trait association research.

2. Materials and Methods

2.1. Plant Material Used in the Study

A reference collection of 150 bush bean accessions was randomly selected from the core collection of *Phaseolus vulgaris* held by the Consultative Group for International Agricultural

Research (CGIAR), which is multiplied at the Center for Tropical Agriculture (CIAT, Cali, Valle de Cauca, Colombia) and duplicated in various vaults around the world (CIMMYT, Mexico City, Mexico, and Svalbard, Norway). The accessions were diverse and originated from 14 countries, including cultivated and wild common beans. Climbing cultivars were not used due to excess growth and problematic agronomic management or adaptation in other latitudes. The collection included 62 genotypes from Mexico, 42 from Peru, 21 from Guatemala, 5 from Colombia, 5 from Ecuador, 3 from Honduras, 3 from the Czech Republic, 2 from Brazil, 2 from Chile, and 1 each from the Dominican Republic, El Salvador, France, Haiti, and Nicaragua. The genotypes were grown in a greenhouse at the AREC research station of Tennessee State University, and young leaves were taken for DNA extractions and analyses.

2.2. DNA Extraction and SNP Genotyping

Genomic DNA was extracted from the first fresh trifoliate leaves using the 2X protocol for DNA isolation explained in Vega Vela and Chacon Sanchez [18]. In total, 150 mg of chilled leaf sample was ground to a powder in a Retsch mill using stainless steel balls and grinding chambers cooled to below freezing with liquid nitrogen. A threshold value of 1.8 for a 260/280 nm absorbance ratio was considered ideal to select for high DNA purity and was confirmed by electrophoresis on a 1% agarose gel run at 110 V in TAE (1X) buffer. The DNA samples were then diluted to an appropriate concentration for SNP genotyping (1 ng μL^{-1}), which was performed at the Soybean Genomics and Improvement Laboratory, Beltsville Agricultural Research Center–West, USDA, ARS, Beltsville, MD 20705-2350. The reference collection DNA was run using an Illumina iSelect 6K Gene Chip (BARCBEAN6K) SNP array [19], and SNP calling was performed using a locus genotyping module in GenomeStudio software 2.0.5 (Illumina Inc., San Diego, CA, USA [20]). The data were then exported using the PLINK input report plug-in V2.14 [21]. A total of 5398 SNPs distributed across the genome were generated from the SNP chip.

2.3. SNP Data Analysis

The raw SNP data were filtered for 10% missing data and a minor allele frequency (MAF) < 0.05. The remaining SNPs were then imputed using the LD KNNi imputation method [22] plugin of TASSEL software v5 [23], using the default parameters. The imputed data were then exported to appropriate input file formats for downstream analysis. The distribution of the SNPs in the genome was assessed using a per chromosome density plot visualized using R-based CMplot [24] and is presented for Dataset_1 in Supplementary Figure S1.

2.4. Population Structure

The pattern of the population structure was first examined using a PCoA. A distance matrix was calculated on the whole set of SNPs remaining after filtering for missingness and the MAF, and principal coordinates were calculated from the distance matrix using the multi-dimensional scale (MDS) function in TASSEL [23]. A biplot of the PCoA was then visualized using the "ggplot2" [25] R package. Subsequently, the phylogenetic relationships were assessed using the s neighbor-joining method in DARwin v6.0 software [26], using a dissimilarity matrix produced with 1000 bootstraps [26,27]. An evaluation of the population structure was also conducted using a Bayesian clustering analysis in the STRUCTURE software program (Version 2.3.3) [28] (Pritchard et al., 2000). SNPs with more than two alleles and a MAF < 0.05 were filtered out using TASSEL software [23].

For a further analysis of the SNPs that were in linkage disequilibrium (LD), SNP pruning for LD, which used the criterion $r^2 \leq 0.2$, was conducted using PLINK software [21]. After filtering and LD pruning, the remaining SNPs were used for a population structure analysis in STRUCTURE as before but with parameters set to a burn-in period of 10,000 and 10,000 Markov chain Monte Carlo (MCMC) repetitions with 10 iterations for the number of populations (K) from 2 to 10 and admixture included in the model. The number of subpopulations was inferred using ΔK [29], which is the change in the log-likelihood between K values using the program Structure Harvester [30]. Individuals were assigned to populations based on a population membership probability (Q-value) threshold fixed at 0.9. Any individuals below that high threshold were considered admixed.

To better study and explain the genetic structure, we employed a discriminant analysis of principal components (DAPC) using the "adegenet" v1.31 in R package [31]. DAPC is a novel statistical and multivariate approach that does not assume LD between markers and partitions the genetic variance into between-group and within-group components. As a DAPC requires prior population information, we used the inferred population assignment obtained from STRUCTURE, the phylogenetic relationship, and the PCoA analysis. A DAPC was conducted using the "dapc" function, and the number of clusters was determined using a "find.clusters" function in the "adegenet" package [32]. A determination of the optimal number of clusters using a Bayesian information criterion for K = 1–10 by k-means clustering with 100,000 iterations and 10 randomly chosen starting centroids was made. Other R packages, including "poppr" [33], "hierfstat" [34], "phangorn" [35], "ape" [36], "pegas" [37], "ggplot2" [25], and "reshape2" [38], were used to process and analyze the data and visualize the results.

2.5. Linkage Disequilibrium and Genetic Diversity

After determining the number of clusters and identifying the best iteration using the lowest alpha value, accessions were regrouped based on the inferred population assignment. LD values were calculated for the entire reference collection and for Andean and Mesoamerican groups separately using TASSEL software. Populations were compared based on the percentage of SNP markers in LD, the average LD (r^2) value, and the percentage of significant LD identified using a two-tailed Fisher's exact probability test at a $p < 0.0001$ significance threshold. The same dataset was used for a diversity and genetic distance analysis within the entire population, between subpopulations (Andean, Mesoamerican, and admixtures), and within subpopulations using Molecular Evolutionary Genetic Analysis (MEGA) vX software [39]. A bootstrap method with 500 permutations followed by a composite maximum likelihood model including gamma distribution and transitions and transversions were used to estimate the average genetic distances between and within subpopulations. Major allele frequency, gene diversity, heterozygosity, and polymorphism information content (PIC) were also determined using PowerMarker v3.25 [40]. An analysis of molecular variance (AMOVA) was also conducted to estimate the genetic variance within and among populations based on a PhiPT value of 1000 permutations using the "adegenet package" [32] in R. The two gene pools and the admixtures were considered populations in the AMOVA. All of these analyses were conducted using a set of 5108 SNPs with a MAF > 0.05.

3. Results

3.1. SNP Set Development

After filtering the SNP data, two datasets were generated. The first dataset (Dataset_1) comprised a set of 5108 SNPs and 150 genotypes remaining after filtering for 10% missing alleles and a MAF < 0.05. The second dataset (Dataset_2) comprised a set of 2300 SNPs remaining after removing SNPs with a null allele and LD pruning with a criterion of $r^2 < 0.2$. Dataset_1 was used for all analyses considered here except population structure, which was analyzed using Dataset_2.

The SNPs were evenly distributed across both the telomeric and centromeric regions of the genome (Supplementary Figure S1). The average number of SNPs per chromosome was 462. The number of SNPs per chromosome ranged from 287 on chromosome Pv06 to 564 on chromosome Pv05. Some telomeres did have a greater SNP concentration, notably the long arms of Pv01 and Pv05 and the short arm of Pv04.

3.2. Population Structure and Phylogenetic Relationships

The phylogenetic tree (Figure 1) and the PCoA (Figure 2) showed the reference collection divided into two subpopulations, namely Andean and Mesoamerican gene pools. Four individual races within each gene pool were also identified: the Mesoamerican ones are potentially Durango–Jalisco (DJ) and Mesoamerica (M), while the Andean ones are Nueva Granada (NG) and Peru (P).

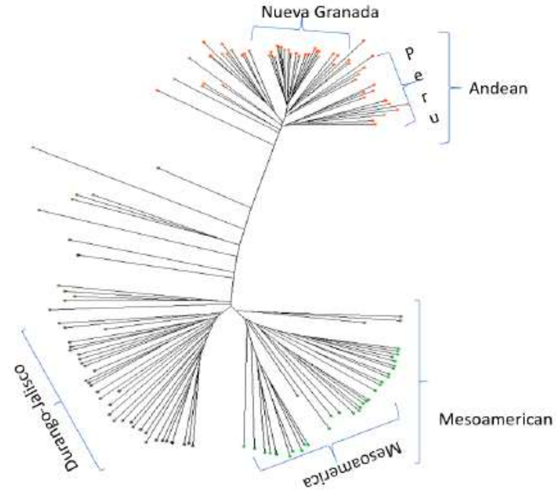

Figure 1. Phylogram visualization of the diversity in the reference collection, showing the phylogenetic relationships between genotypes based on the 5108 SNPs found in this study.

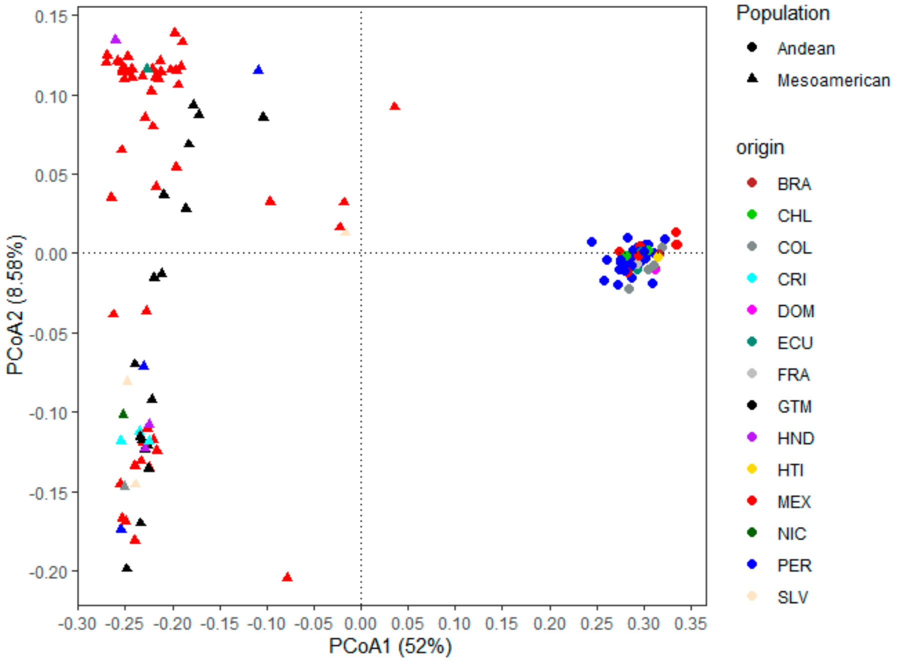

Figure 2. Principal coordinate analysis (PCoA) of the reference collection based on SNP genotyping, with genotypes labeled by a priori population assignment to the Andean and Mesoamerican gene pools, each represented by a circle or triangle, respectively. Points on the graph are color-coded based on countries of origin, which are listed in an abbreviated form.

The dendrogram and PCoA visualizations were supported by an evaluation of the population structure through unsupervised Bayesian genetic clustering allowing for admixture and evaluated with a ΔK test for the best K-value (Figure 3). The separation of the populations at each K-value was enlightening. At K = 2, the reference collection was divided in to Mesoamerican and Andean gene pools; most of the genotypes (58.66%) were grouped into the former. Meanwhile, 34% (n = 51) of the genotypes were in the Andean gene pool, and 7.33% (n = 11) were admixed between the two gene pools.

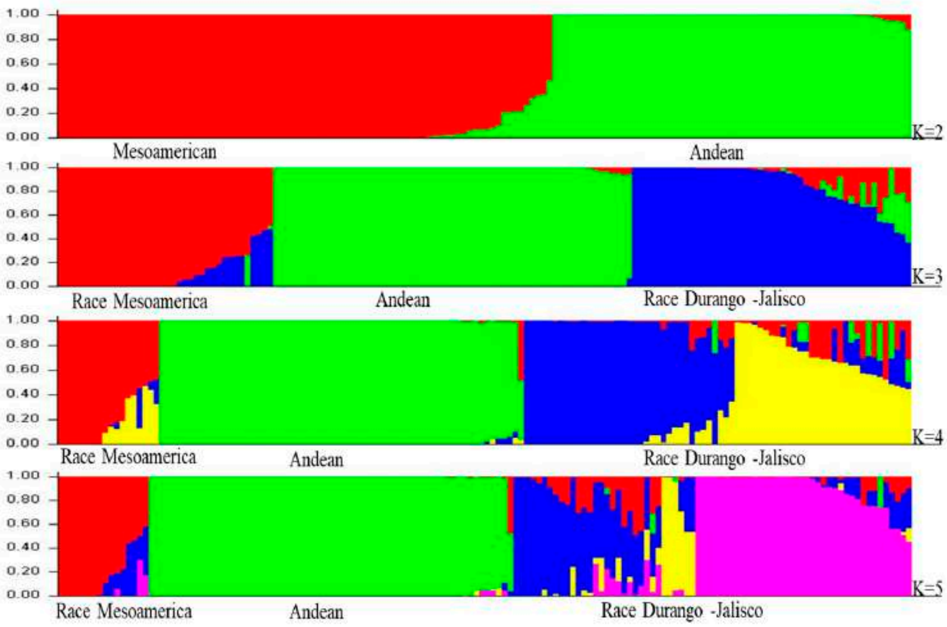

Figure 3. Population structure from K = 2 to K = 5 for bean cultivars in the reference collection, using 2300 SNP markers obtained after filtering for null alleles and linkage disequilibrium (LD) blocks with a criterion of $r^2 < 0.2$. Color differences represent clade groups and possible bean races.

At K = 3, the Mesoamerican gene pool was further divided into two subpopulations corresponding to the Mesoamerica race and the combined Durango–Jalisco race. Increasing the number of clusters from K = 3 to K = 4 and above further divided the Mesoamerican gene pool, especially the Durango–Jalisco race, while the Andean gene pool was not subdivided any further (Figure 3). Though we did not see a significant differentiation in the Andean beans, in the phylogenetic tree, we recognize small clade groups representing the separation between the race Nueva Granada (NG) and the race Peru.

According to the inferred population assignment from the above complementary diversity and population structure analyses (PCoA, STRUCTURE, and the phylogenetic relationship), the majority of the genotypes that clustered in the Andean gene pool were from Peru, Mexico, and Colombia, while those grouped in the Mesoamerican gene pool were dominated by genotypes collected from Mexico and Guatemala. The majority of admixed or inter-gene pool genotypes were from Mexico and Guatemala or Peru. The genotypes of other countries were distributed in all of the subpopulations.

The DAPC results (Figure 4, left) also confirmed the patterns of germplasm divisions from the STRUCTURE analysis of the clusters from K = 2 to K = 6. We observed the same trend: the Mesoamerican gene pool split as the K number increased, while the Andean bean population remained intact. Similar to the PCoA results, the first linear discriminant split the reference collection into Mesoamerican and Andean beans, while the second linear discriminant split the Mesoamerican beans into more groups. The markers with the greater contributions to the first and second principal coordinates were identified using a loading plot. There were close to ten SNP markers were found to have made a greater contribution to the separation of the Andean and Mesoamerican gene pools with the first principal component, while six SNP markers were found to have made a greater contribution to the second principal component, which enabled the further separation of separate the Mesoamerican gene pool as the Delta K increased (Figure 4, right). The cluster with the largest ΔK (K = 2) was used to define the number of gene pools in the reference collection, while K = 6 was allowed for the DAPC of common beans.

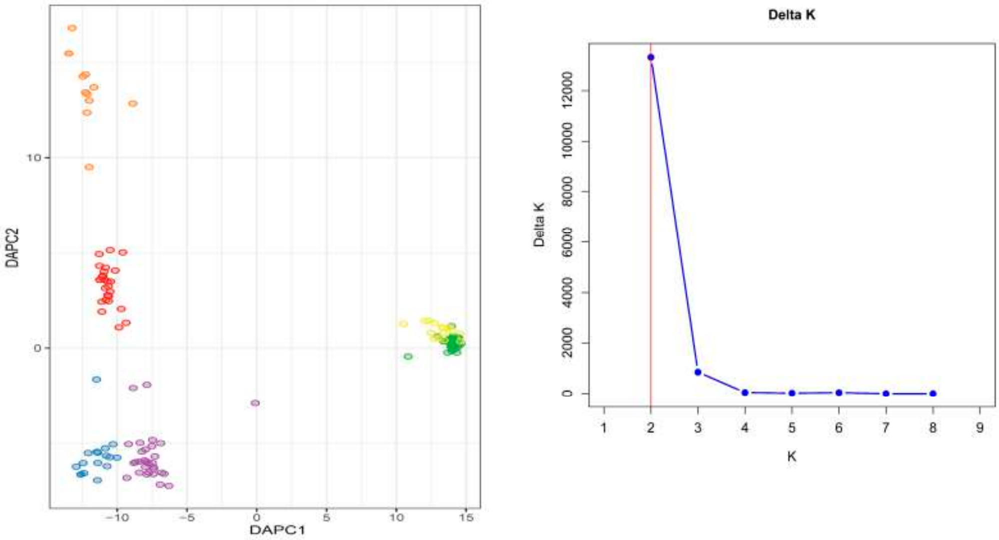

Figure 4. Results of SNP genotyping with Dataset_2, showing (**left**) discriminant analysis of principal components (DAPC) and (**right**) ΔK values based on K as the number of clusters.

In the DAPC from the figure above, the first principal component separated the Andean and Mesoamerican gene pools, while the second principal component separated the Mesoamerican gene pool into four distinct subpopulations corresponding to color-coded races for each gene pool, with green and yellow circles representing Andean accessions and then orange, red, blue, or purple circles representing Mesoamericans accessions, agreeing with the Evanno test for population structure.

3.3. Linkage Disequilibrium and Genetic Diversity

The distribution of SNPs used in the initial part of this study was confirmed to be genome-wide and even across linkage groups (Supplemental Figure S1). Given this, LD values between pairwise SNPs were explored across the whole genome for the entire collection and for the Mesoamerican and Andean beans only. For the entire 150 genotypes in the reference collection, only 15% of the pairwise SNPs across the 11 common bean chromosomes were in linkage disequilibrium (LD), 100% of which were in significant LD ($r^2 < 0.5$ and $p < 0.0001$). The LD levels in the Mesoamerican and Andean beans were 3.9 and 2.2%, respectively. Though the percentages of pairwise SNPs in LD were low, the percentages of pairwise SNPs in significant LD were high: 99.25% in the Mesoamerican beans and 89.22% in the Andean beans (Table 1).

Table 1. Average and percent of pairwise SNP markers in linkage disequilibrium (LD) at $r^2 \geq 0.5$, calculated for the complete reference collection of common beans presented in this study and then separately for each gene pool (Andean and Mesoamerican beans within the reference collection).

Population	Average LD (r^2)	% Loci in LD	% of LD Loci Highly Significant [1]
Andean	0.27	2.20	89.22
Mesoamerican	0.15	3.94	99.25
Complete	0.26	15.00	100.00

[1] significant at $p < 0.0001$.

The admixture (inter-gene pool) population showed no evidence of pairwise LD or the significance of LD. The genetic diversity of the reference collection was also assessed based on the average genetic distance between members of different groups (Mesoamerican, Andean, and admixed populations), within members of same group, and between group members in the case of the complete collection (Table 2, left half).

Table 2. Measures of average genetic distance within and among subgroups of the common bean reference collection, showing Andean and Mesoamerican gene pools and the admixture subgroup. Average genetic distances within subgroups are above the diagonal, while between subgroups, they are below the diagonal. The average major allele frequency (MAF), gene diversity (GD), expected heterozygosity (H_e), and polymorphism information content (PIC) of all loci for each subgroup are given. Bolding of numbers indicate the diagonal in average distances.

Population	Average Distance			MAF	GD	H_e	PIC
	Andean	**Meso Am**	**Admix**				
Andean	**0.12**	1.93	0.69	0.757	0.324	0.068	0.264
Mesoamerican	2.19	**0.39**	0.28	0.753	0.306	0.133	0.265
Admixture	1.17	0.89	**0.84**	0.705	0.377	0.073	0.299
Complete	1.14			0.907	0.125	0.067	0.105

Subsequently, the net genetic distance between members of different groups and the overall average genetic distance were calculated for the entire reference collection (Table 2, right half). The average distance for the entire collection was 1.14. The admixed population had the largest within-group average genetic distance (0.84). The genetic distance between the Andean and Mesoamerican beans was large, and it was almost twice as much as the overall average genetic distance. The genetic diversity was also assessed using major allele frequency, gene diversity, heterozygosity, and polymorphic PIC values. The complete collection was characterized by a higher MAF (0.907) and lower genetic diversity (0.125), heterozygosity (0.067), and PIC (0.105) values than the gene pools individually. The Mesoamerican population had the highest heterozygosity (0.133), whereas for the MAF and PIC, it had values between the Andean and admixed populations. Finally, the admixed and the Andean populations displayed the highest gene diversity (0.377 and 0.324, respectively) compared to Mesoamerican beans (0.306).

The analysis of molecular variance (Table 3) confirmed the presence of significant variation between populations, between genotypes within a population, and between genotypes in the entire population. The highest proportion of genetic variation (45%) was contributed by the difference between populations, followed by differences between samples within populations.

Table 3. Analysis of molecular variance between Andean, Mesoamerican, and admixed common beans subpopulations of the 150 genotypes in the reference collection based on an inferred population assignment obtained from a STRUCTURE analysis using the 5108 SNPs found in this study.

Source of Variation	df [1]	MS	Variance	% Variation	*p*-Value
Between Populations	2	91,819	1116	45	0.001
Between Samples within population	147	2326	977	40	0.001
Within Samples	149	371	371	15	0.001

[1] Abbreviations: df = degrees of freedom; MS = mean square; *p*-value = probability.

4. Discussion

An in-depth understanding of the extent of genetic diversity and population structure in a crop population is of paramount importance for any crop improvement program to design and frame the breeding objectives and methods that can efficiently promote the translation of genetic diversity into genetic gain.

Current knowledge on the pattern of genetic diversity, population structure, and domestication events in common beans is based on various studies conducted using different morphological, biochemical, and molecular genetic markers. In mid-1980s, seed protein markers were used to study genetic diversity in common beans from Latin America [1].

Later, with the advancement of molecular marker technology, DNA markers became prevalent. For example, since the development of the first SSR markers for common beans [6], they have been widely used for genetic diversity and population structure studies in the crop. Among SSR studies, examples of Blair et al. [41] and Díaz and Blair [42] are some from our laboratory finding gene pool division and admixture events. Most SSRs were evaluated by precise fluorescent labeling.

In recent years, there has been a cost reduction in genome sequencing and allele-calling technologies. As a result, SNP markers have become the markers of choice. Some SNP studies have used KASP technology [3,15] or enzyme digestions [43,44]; however, the majority have used Illumina chips, as represented by the BARCBean6K_3 SNP array we used as well. This array was first tested for GWAS analyses after a genetic diversity assessment in both Andean and Mesoamerican common bean breeding material [16,17]. Our current study explored and quantified the level of genetic diversity and interrelationships in the reference collection of 150 common beans using the same BARCBean6K_3 Illumina SNP array but with more diverse genebank materials.

From previous studies and their passport data, we identified that the Mesoamerican gene pool was divided into the races Mesoamerica (M) and Durango–Jalisco (DJ). Since the genotypes for our reference collection were directly from the centers of the origin and domestication of beans in North and South America, the population structure analysis split the reference collection cleanly into two subpopulations, the Andean and Mesoamerican gene pools.

Furthermore, in our study, the Mesoamerican populations displayed higher genetic differentiation and tended to split into more groups based on the subraces of DJ and M1 and M2. Even at K = 4, the Mesoamerican gene pool was grouped into three groups (DJ and M). In other words, at K= 4, while the race Mesoamerica remained intact, the race Durango–Jalisco was further divided in two.

This result is consistent with the bean population structure established in previous studies in which two gene pools and multiple races were present. For example, in Cuba [45], two distinct populations were found using SSR markers corresponding to each gene pool there. Common bean genotypes collected from Ethiopian and Kenyan production ecologies were found to have considerable diversity that corresponded well to the Andean and Mesoamerican gene pools [12]. In our previous study of the international core collection using SSR markers, we found an excellent separation of the two gene pools using land races but also gene pool admixtures and subdivisions of races within gene pools [4].

With all these studies agreeing on the two gene pool structures, it is important to know the degree of separation between them. In our PCoA, the first principal component of the PCoA accounted for 52% of the variance and separated the Mesoamerican and Andean gene pools into the same groups as those that were found at K = 2 using STRUCTURE. The second principal component only accounted for 8.58% of the variance but separated the Mesoamerican gene pool into two subgroups, possible races, as was found in the phylogenetic tree (see Figure 1, the groups labeled Mesoamerica and Durango–Jalisco). These races would align with the subpopulation structure found for K = 3.

Other important results of our study included discovering the PIC values of individual SNPs as one parameter for measuring genetic diversity in each subpopulation. The mean PIC values of all SNPs for the admixture group (0.299) was higher than the PIC values for either the Andean group or Mesoamerican group individually or compared to the complete collection (PIC = 0.105). This showed the reference collection to be highly diverse, with a significant contribution of the Mesoamerican gene pool and admixture events occurring between the two main common bean gene pools.

Exploring the extent and patterns of linkage disequilibrium (LD) in a population is a crucial factor in any marker-trait association study and marker-assisted selection. Association mapping exploits the LD present among individuals from natural populations or germplasm collections to dissect the genetic basis of complex trait variation [46].

Linkage disequilibrium is extremely population-specific and can determine the utility of a panel for marker trait associations. This study examined the reference collection and compared the subpopulations based on the proportion of paired SNP markers that are in significant LD at $p < 0.0001$ and an LD cutoff value of $r^2 > 0.5$.

The average LD for the Andean gene pool was equal to the average LD in the entire population but had the lowest percentage of significant LD when compared to the 100% significant LD for Mesoamerica and the entire population. From the total pairwise comparisons using 5108 SNPs, only 15% of the pairwise SNPs were significant in the entire population while a smaller LD percentage was found in the Mesoamerican and Andean beans. Similar results were reported by Campa et al. [47], who found that 10% of pairwise SNP markers were in LD using a national panel of 308 Spanish common beans evaluated with 3099 SNPs from genotyping by sequencing (GBS), a method that is less precise for genotyping than our use of the Illumina chip. Unlike that previous study, we emphasized bush beans that are easy to grow in multiple environments and de-emphasized climbing beans as they are often difficult to grow in the greenhouse or field due to the large biomass that needs to be staked.

In general, the Andean beans were characterized by a larger average LD and lower LD percentage, a small average genetic distance between members of the population, a higher major allele frequency, and lower gene diversity and PIC values, which suggest narrower genetic diversity compared to the Mesoamerican genepool. This result is consistent with previous reports. Campa et al. [47] and Bitocchi et al. [48] reported narrower genetic variation in the Andean beans compared to their Mesoamerican counterparts, with the latter authors suggesting a domestication bottleneck that is especially evident in gene sequences to explain why Andean beans could be less diverse in SNP polymorphisms than Mesoamerican beans. Blair et al. [49] found similar level of simple-sequence-repeat-based diversity in the two gene pools, making GWASs appropriate across groups.

5. Conclusions

The results of this study indicated the presence of high genetic diversity in the reference collection but also the necessity of considering linkage disequilibrium, genetic distance, and the proportion of genetic variation when describing or utilizing the collection. Given the observed molecular variation, the reference collection could be used as a diversity panel for GWASs and could be interesting for marker trait association results. While other reports have included more genotypes in their panels, they can be unwieldy due to narrow adaptation or the later flowering time of certain beans, such as climbers, which we considered undesirable for a reference collection of mainly bush beans. This is important for agronomic management as heavily vined, climbing beans are much more complicated to manage agronomically than free-standing bush beans. Finally, the reference collection we developed was intended for global use rather than as a specific breeding program collection for one region or one country. Therefore, we hope to analyze the utility of this international reference collection diversity panel for GWAS evaluation of traits important to common bean in Africa, Asia, Europe, and North or South America.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/agronomy14050985/s1, Figure S1: Distribution and density of Dataset_1 common bean SNPs.

Author Contributions: Conceptualization, M.W.B., N.R.C., A.A. and D.A.; funding acquisition, M.W.B.; investigation, D.A., J.M.L. and N.R.C.; methodology, D.A., J.M.L., N.R.C. and M.W.B.; project administration, M.W.B.; resources, A.A.; software, J.M.L.; supervision, N.R.C., A.A. and M.W.B.; visualization, D.A. and M.W.B.; writing—original draft, D.A. and M.W.B.; writing—review and editing, M.W.B. All authors have read and agreed to the published version of the manuscript.

Funding: Funding for this project was provided by the Evans Allen Fund of the National Institute of Food and Agriculture (NIFA) as part of the United States Department of Agriculture (USDA) grant TENX-07 to Tennessee State University. Start-up funding was also provided by Evolutionary Genomics (EG) LLC through a project with the Bill and Melinda Gates Foundation (BMGF).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Acknowledgments: We thank program leader, and technicians at the USDA-Beltsville, Soybean for performing SNP genotyping.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Gepts, P.; Bliss, F.A. Phaseolin Variability among Wild and Cultivated Common Beans (*Phaseolus vulgaris*) from Colombia. *Econ. Bot.* **1986**, *40*, 469–478. [CrossRef]
2. Blair, M.W.; Soler, A.; Cortés, A.J. Diversification and Population Structure in Common Beans (*Phaseolus vulgaris* L.). *PLoS ONE* **2012**, *7*, e49488. [CrossRef] [PubMed]
3. Gaut, B. The complex domestication history of the common bean. *Nat. Genet.* **2014**, *46*, 663–664. [CrossRef] [PubMed]
4. Bitocchi, E.; Bellucci, E.; Giardini, A.; Rau, D.; Rodriguez, M.; Biagetti, E.; Santilocchi, R.; Spagnoletti Zeuli, P.; Gioia, T.; Logozzo, G.; et al. Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. *New Phytol.* **2013**, *197*, 300–313. [CrossRef] [PubMed]
5. Bitocchi, E.; Rau, D.; Bellucci, E.; Rodriguez, M.; Murgia, M.L.; Gioia, T.; Santo, D.; Nanni, L.; Attene, G.; Papa, R. Beans (*Phaseolus* ssp) as a model for understanding crop evolution. *Front. Plant Sci.* **2017**, *8*, 722. [CrossRef] [PubMed]
6. Arkwazee, H.A.; Wallace, L.T.; Hart, J.P.; Griffiths, P.D.; Myers, J.R. Genome-Wide Association Study (GWAS) of White Mold Resistance in Snap Bean. *Genes* **2022**, *13*, 2297. [CrossRef] [PubMed]
7. Rodríguez De Luque, J.J.; Creamer, B. Major constraints and trends for common bean production and commercialization; establishing priorities for future research. *Agron. Colomb.* **2014**, *32*, 423–431. [CrossRef]
8. Ambachew, D.; Mekbib, F.; Asfaw, A.; Blair, M.W. Trait associations in common bean genotypes grown under drought stress and field infestation by BSM bean fly. *Crop J.* **2015**, *3*, 305–316. [CrossRef]
9. Cichy, K.; Chiu, C.; Isaacs, K.; Glahn, R. Dry Bean Biofortification with Iron and Zinc. In *Biofortification of Staple Crops*; Kumar, S., Dikshit, H.K., Mishra, G.P., Singh, A., Eds.; Springer: Singapore, 2022. [CrossRef]
10. McClean, P.E.; Terpstra, J.; McConnell, M.; White, C.; Lee, R.; Mamidi, S. Population structure and genetic differentiation among the USDA common bean (*Phaseolus vulgaris* L.) core collection. *Genet. Resour. Crop Evol.* **2012**, *59*, 499–515. [CrossRef]
11. Nadeem, M.A.; Yeken, M.Z.; Shahid, M.Q.; Habyarimana, E.; Yılmaz, H.; Alsaleh, A.; Hatipoğlu, R.; Çilesiz, Y.; Khawar, K.M.; Ludidi, N.; et al. Common bean as a potential crop for future food security: An overview of past, current and future contributions in genomics, transcriptomics, transgenics and proteomics. *Biotechnol. Biotechnol. Equip.* **2021**, *35*, 759–787. [CrossRef]
12. Swarup, S.; Cargill, E.J.; Crosby, K.; Flagel, L.; Kniskern, J.; Glenn, K.C. Genetic diversity is indispensable for plant breeding to improve crops. *Crop Sci.* **2021**, *61*, 839–852. [CrossRef]
13. Brown, J.W.S.; Ma, Y.; Bliss, F.A.; Hall, T.C. *Genetic Variation in the Subunits of Globulin-1 Storage Protein of French Bean*; Springer: Berlin/Heidelberg, Germany, 1981; Volume 59.
14. Singh, S.; Gepts, P.; Debouck, D. Races of Common Bean (*Phaseolus vulgaris*, Fabaceae). *Econ. Bot.* **1991**, *45*, 379–396. [CrossRef]
15. Cortés, A.J.; Chavarro, M.C.; Blair, M.W. SNP marker diversity in common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* **2011**, *123*, 827–845. [CrossRef] [PubMed]
16. Cichy, K.A.; Wiesinger, J.A.; Mendoza, F.A. Genetic diversity and genome-wide association analysis of cooking time in dry bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* **2015**, *128*, 1555–1567. [CrossRef]
17. Moghaddam, S.M.; Mamidi, S.; Osorno, J.M.; Lee, R.; Brick, M.; Kelly, J.; Miklas, P.; Urrea, C.; Song, Q.; Cregan, P.; et al. Genome-Wide Association Study Identifies Candidate Loci Underlying Agronomic Traits in a Middle American Diversity Panel of Common Bean. *Plant Genome* **2016**, *9*, 1–21. [CrossRef] [PubMed]
18. Vega-Vela, N.E.; Chacón-Sánchez, M.I. Isolation of high-quality DNA in 16 aromatic and medicinal Colombian species using silica-based extraction columns. *Agron. Colomb.* **2011**, *29*, 349–357.
19. Song, Q.; Jia, G.; Hyten, D.L.; Jenkins, J.; Hwang, E.-Y.; Schroeder, S.G.; Osorno, J.M.; Schmutz, J.; Jackson, S.A.; McClean, P.E.; et al. SNP assay development for linkage map construction, anchoring whole-genome sequence, and other genetic and genomic applications in common bean. *G3 Genes Genomes Genet.* **2015**, *5*, 2285–2290. [CrossRef]
20. Zhao, S.; Jing, W.; Samuels, D.C.; Sheng, Q.; Shyr, Y.; Guo, Y. Strategies for processing and quality control of Illumina genotyping arrays. *Brief. Bioinform.* **2018**, *19*, 765–775. [CrossRef]
21. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef]
22. Money, D.; Gardner, K.; Migicovsky, Z.; Schwaninger, H.; Zhong, G.Y.; Myles, S. LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. *G3 Genes Genomes Genet.* **2015**, *5*, 2383–2390. [CrossRef]
23. Bradbury, P.J.; Zhang, Z.; Kroon, D.E.; Casstevens, T.M.; Ramdoss, Y.; Buckler, E.S. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **2007**, *23*, 2633–2635. [CrossRef] [PubMed]

24. Yin, L.; Zhang, H.; Tang, Z.; Xu, J.; Yin, D. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study. *Genom. Proteom. Bioinform.* **2020**, *19*, 619–628. [CrossRef]
25. Wickham, H.; Chang, W.; Henry, L.; Pedersen, T. Package 'ggplot2'. 2024, pp. 1–318. Available online: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf (accessed on 2 March 2024).
26. Perrier, X.; Jacquemoud-Collet, J.P. DocDarwin6. 2010, Volume 1, pp. 1–116. Available online: https://darwin.cirad.fr/ (accessed on 2 March 2024).
27. Perrier, X.; Flori, A.; Bonnot, F. *Genetic Diversity of Cultivated Tropical Plants*; Hamon, J., Seguin, M., Perrier, X., Eds.; Science Publishers: Enfield, NH, USA, 2003; pp. 43–76.
28. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of Population Structure Using Multilocs Genopyte Data. *Genet. Soc. Am.* **2000**, *155*, 945–959. [CrossRef]
29. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* **2005**, *14*, 2611–2620. [CrossRef]
30. Earl, D.A.; vonHoldt, B.M. Structure harvester: A website and program for visualizing Structure output and implementing the Evanno method. *Conserv. Genet. Resour.* **2012**, *4*, 359–361. [CrossRef]
31. Jombart, T.; Devillard, S.; Balloux, F. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.* **2010**, *11*, 94. [CrossRef]
32. Jombart, T.; Ahmed, I. Adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* **2011**, *27*, 3070–3071. [CrossRef] [PubMed]
33. Kamvar, Z.N.; Tabima, J.F.; Grünwald, N.J. Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2014**, *2014*, e281. [CrossRef]
34. Goudet, J. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **2005**, *5*, 184–186. [CrossRef]
35. Schliep, K.; Potts, A.J.; Morrison, D.A.; Grimm, G.W. Intertwining phylogenetic trees and networks. *Methods Ecol. Evol.* **2017**, *8*, 1212–1220. [CrossRef]
36. Paradis, E.; Schliep, K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **2019**, *35*, 526–528. [CrossRef]
37. Paradis, E. Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics* **2010**, *26*, 419–420. [CrossRef]
38. Wickham, H. Reshaping Data with the reshape Package. *J. Stat. Softw.* **2007**, *21*, 1–20. [CrossRef]
39. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. [CrossRef]
40. Liu, K.; Muse, S.V. PowerMaker: An integrated analysis environment for genetic marker analysis. *Bioinformatics* **2005**, *21*, 2128–2129. [CrossRef]
41. Blair, M.W.; Giraldo, M.C.; Buendía, H.F.; Tovar, E.; Duque, M.C. Microsatellite marker diversity in common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* **2006**, *113*, 100–109. [CrossRef]
42. Díaz, L.M.; Blair, M.W. Race structure within the Mesoamerican gene pool of common bean (*Phaseolus vulgaris* L.) as determined by microsatellite markers. *Theor. Appl. Genet.* **2006**, *114*, 143–154. [CrossRef]
43. Mamidi, S.; Rossi, M.; Annam, D.; Moghaddam, S.; Lee, R.; Papa, R.; McClean, P. Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. *Funct. Plant Biol.* **2011**, *38*, 953–967. [CrossRef]
44. Valdisser, P.A.M.R.; Pappas, G.J.; de Menezes, I.P.P.; Müller, B.S.F.; Pereira, W.J.; Narciso, M.G.; Brondani, C.; Souza, T.L.P.O.; Borba, T.C.O.; Vianello, R.P. SNP discovery in common bean by restriction-associated DNA (RAD) sequencing for genetic diversity and population structure analysis. *Mol. Genet. Genom.* **2016**, *291*, 1277–1291. [CrossRef]
45. Blair, M.W.; Lorigados, S.M. Diversity of common bean landraces, breeding lines, and varieties from Cuba. *Crop Sci.* **2016**, *56*, 322–330. [CrossRef]
46. Myles, S.; Peiffer, J.; Brown, P.J.; Ersoz, E.S.; Zhang, Z.; Costich, D.E.; Buckler, E.S. Association mapping: Critical considerations shift from genotyping to experimental design. *Plant Cell* **2009**, *21*, 2194–2202. [CrossRef]
47. Campa, A.; Murube, E.; Ferreira, J.J. Genetic diversity, population structure, and linkage disequilibrium in a Spanish common bean diversity panel revealed through genotyping-by-sequencing. *Genes* **2018**, *9*, 518. [CrossRef]
48. Bitocchi, E.; Nanni, L.; Bellucci, E.; Rossi, M.; Giardini, A.; Zeuli, P.S.; Logozzo, G.; Stougaard, J.; McClean, P.; Attene, G.; et al. Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E788–E796. [CrossRef]
49. Blair, M.W.; Cortés, A.J.; Penmetsa, R.V.; Farmer, A.; Carrasquilla-Garcia, N.; Cook, D.R. A high-throughput SNP marker system for parental polymorphism screening, and diversity analysis in common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* **2013**, *126*, 535–548. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.