Genome-Wide Scanning Enabled SNP Discovery, Linkage Disequilibrium Patterns and Population Structure in a Panel of Fonio (Digitaria exilis [Kippist] Stapf) Germplasm

Abdou R Ibrahim Bio Yerima, Kassari Ango Issoufou, Charlotte A Adje, Aissata Mamadou, Happiness Oselebe, Mame Codou Gueye, Claire Billot, Enoch G Achigan-Dako

To cite this version:

HAL Id: hal-03442490
https://hal.inrae.fr/hal-03442490
Submitted on 23 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Genome-Wide Scanning Enabled SNP Discovery, Linkage Disequilibrium Patterns and Population Structure in a Panel of Fonio (Digitaria exilis [Kippist] Stapf) Germplasm

Abdou R. Ibrahim Bio Yerima1,2,*, Kassari Ango Issoufou2, Charlotte A. Adje1, Aissata Mamadou2, Happiness Oselebe3, Mame Codou Gueye4, Claire Billot5 and Enoch G. Achigan-Dako1

1Laboratory of Genetics, Biotechnology and Seed Sciences (G BioS), Faculty of Agronomic Sciences (FSA), University of Abomey-Calavi (UAC), Cotonou, Benin, 2Department of Rainfed Crop Production (DCP), National Institute of Agronomic Research of Niger (INRAN), Niamey, Niger, 3Department of Crop Production & Landscape Management, Faculty of Agriculture & Natural Resources Management, Ebonyi State University, Abakaliki, Nigeria, 4Centre d’Étude Régional pour l’Amélioration de l’Adaptation à la Sécheresse (CERAAS), Institut Sénégalais de Recherches Agricoles (ISRA), Thiès, Senegal, 5UMR AGAP Institut, Univ Montpellier, Cirad, Inrae, Institut Agro, Montpellier, France

White fonio (Digitaria exilis) is a staple food for millions of people in arid and semi-arid areas of West Africa. Knowledge about nutritional and health benefits, insights into morphological diversity, and the recent development of genomic resources call for a better understanding of the genetic structure of the extant germplasm gathered throughout the region in order to set up a robust breeding program. We assessed the genetic diversity and population structure of 259 fonio individuals collected from six countries from West Africa (Nigeria, Benin, Guinea, Mali, Burkina Faso and Niger) in this study using 688 putative out of 21,324 DArTseq-derived SNP markers. Due to the inbreeding and small population size, the results revealed a substantial level of genetic variability. Furthermore, two clusters were found irrespective of the geographic origins of accessions. Moreover, the high level of linkage disequilibrium (LD) between loci observed resulted from the mating system of the crop, which is often associated with a low recombination rate. These findings fill the gaps about the molecular diversity and genetic structure of the white fonio germplasm in West Africa. This was required for the application of genomic tools that can potentially speed up the genetic gain in fonio millet breeding for complex traits such as yield, and other nutrient contents.

Keywords: Digitaria exilis, staple food, genotyping by sequencing, genetic diversity, population structure, linkage disequilibrium patterns
INTRODUCTION

Plant breeders often use the diversity of plant genetic resources to create new and improved cultivars with desirable features such as production potential, nutritional quality, large seed size, and resistance to biotic and abiotic stresses (Govindaraj et al., 2015). Modern crop varieties are typically inappropriate for low-input agriculture in marginal environments, despite the fact that they were created largely for high yielding potential under well-defined conditions (Evenson and Gollin, 2003), compared to traditional varieties or landraces that are well-adapted to various farming systems under different ecological regions, thus contributing to food and nutrition security of subsistent small-scale farmers (Ceccarelli and Grando, 2002). Among traditional cereals, fonio is one of West Africa’s oldest native cereal crops, with social, economic, and nutritional significance (Dansi et al., 2010). Fonio grain is used to make salads, couscous, stews, candies, and alcoholic and non-alcoholic beverages, as well as traditional porridges, fonio Jollof, and flour creams (Abdul and Jideani, 2019). Furthermore, the ecological versatility of fonio, makes the crop an important component to be considered for tackling issues related to food insecurity and malnutrition in the future (Raneri et al., 2019). However, in spite of its inherent advantages in playing a critical role for food and nutrition security, and income generation, no improved varieties have been developed so far. With the current advances in molecular breeding, additional efforts are therefore needed to bridge the gaps and generate knowledge that can contribute significantly to increase precision and accuracy in fonio improvement using molecular data.

During the last decades, the efficiency and precision of conventional breeding have been remarkably improved through the application of molecular tools in both basic and applied research (e.g., diversity studies, phylogenetic and evolutionary analyses, linkage QTL mapping, association mapping studies, marker-assisted selection, genomic-assisted selection) (Gross-Germain and Viruel, 2013). Various molecular markers have been reported earlier in fonio millet (Digitaria exilis [Kippist] Stapf) for diversity, phylogenetic and evolutionary studies. These markers included isozyme (Adoukonou-Sagbadja et al., 2010), Random Amplified Polymorphic DNA (RAPD) (Hilu et al., 1997; Kuta et al., 2005), Inter-Simple Sequence Repeat (ISSR) (Anumasun et al., 2018), and Simple Sequence Repeat (SSR) (Barnaud et al., 2012; Ngom et al., 2018; Oloko et al., 2019). Amplified Fragment Length Polymorphism (AFLP) markers were found to be effective in detecting genetic diversity and determining population differentiation, and moderate levels of genetic diversity were detected in Digitaria exilis (Adoukonou-Sagbadja et al., 2007). Furthermore, SSR markers were frequently used for diversity, phylogenetic, and evolutionary investigations (Ellegren, 2002) due to their co-dominance, high polymorphism, high global mutation rates, and repeatability, even if designing primers requires a thorough understanding of DNA sequences (Xu and Crouch, 2008; Mondini et al., 2009). With the recent advances in genotyping by sequencing (GBS) technologies, single nucleotide polymorphisms (SNPs) are becoming the markers of choice in crop genetic studies because of their bi-allelic nature, which allows for accurate variant calling, high reproducibility, and the fact that they provide a rapid, high-throughput, and cost-effective tool for exploring plant genetic diversity on a genome-wide scale. Ching et al. (2002), Chagné et al. (2008), and Studer and Kölliker (2013) without prior knowledge of the species of interest’s genome. Most importantly, these markers enabled dissection of the genetic basis underlying complex quantitative traits through biparental linkage and mapping approaches (Sehgal et al., 2020) as well as comparative genomics (Abrouk et al., 2020; Wang et al., 2021). Unlike linkage mapping, the association mapping approach, also known as linkage disequilibrium (LD) mapping, makes use of the existing genetic diversity in natural germplasm populations and exploits their historic recombination events to map or fine map genes with greater precision (Buckler IV and Thornsberry, 2002). Thus, for a successful association mapping and genomic prediction, a better understanding of LD’s patterns and decay within a population is very critical to figure out the number of markers needed (Porto-Neto et al., 2014). Khatkar et al. (2008) reported that when low LD levels exist in a population, to capture the genetic variation across the genome, a higher marker density is required. To date, the most commonly used measure of LD is the $r^2$ being the squared value of the correlation coefficient between two polymorphic loci’s allelic states (Pritchard and Przeworski, 2001). LD patterns based on SNP markers have been widely studied in different crop species including cereals (Chen et al., 2017; Serba et al., 2019) and legumes (Otyama et al., 2019; Sodedji et al., 2020). However, assessment of genetic diversity, population structure analysis, and Linkage Disequilibrium patterns of fonio millet at genome-wide scale were not performed. Therefore, to better characterize the diversity of the current collection of fonio at genomic level and understand the population structure at regional level, advanced high throughput genotyping platforms is needed. This is the first time, as far as we know that genome-wide diversity is explored on a regional basis in fonio millet including collections from Guinea, Mali, Niger, Benin, Nigeria and Burkina Faso. The paper intends to respond to the following questions: (a) What is the genetic basis of the population structure? (b) how are fonio individuals from various ecological regions related? (c) Does population differentiation link with geographical patterns? and What are the LD patterns? The aims of this research are to investigate the genetic diversity in fonio millet accessions, to determine the structure and level of population differentiation, and to measure LD. We assume that investigating the genetic diversity on a genome wide-scale using genotyping by sequencing analysis will reveal association that is useful in fonio molecular breeding.

MATERIALS AND METHODS

Description of Plant Materials

The plant material included 259 fonio millet accessions collected across six West African countries (Benin, Burkina Faso, Guinea, Mali, Niger, and Nigeria) and kept at the genebank of the Laboratory of Genetics, Biotechnology and Seed Science (GBioS) of the University of Abomey-Calavi (UAC) in Benin. The diversity panel has also been collected...
from a range of ecological regions including arid, semi-arid and sub-humid. The most accessions in the collection were from semi-arid ecology (183), followed by arid (50) and sub-humid (26) ecologies. According to the country of provenance, 42 fonio accessions were collected from Benin, four from Burkina Faso, 29 from Guinea, 28 from Mali, 28 from Niger, and 128 from Nigeria. The distribution map of fonio accessions collected from six West African countries is presented below (Figure 1).

DNA Extraction and Genotyping
Fonio plants were grown in plastic bags in the GBioS laboratory at the University of Abomey-Calavi (Benin) and three weeks old leaves were sampled from a single plant and collected into 96 deep well samples collection plates and sent to the Integrated Genotyping Service and Support (IGSS) platform (https://ordering.igssafrica.org/cgi-bin/order/login.pl) now SEQART AFRICA located at Biosciences Eastern and Central Africa (BecA-ILRI) Hub in Nairobi for Genotyping. The Nucleomag Plant Genomic DNA extracting kit was used to extract genomic DNA from 50–100 ng/ul, with the quality and quantity of the DNA extracted being checked on 0.8% agarose. DAoT-Seq has been optimized for *D. exilis* by choosing the best method to reduce complexity (*PstI*-*MseI* restriction enzymes). The DNA samples were processed using Kilian et al. (2012) digestion and ligation reactions method, but they replace a single *PstI*-compatible adapter with two different adapters that match two different restriction enzymes. The Illumina flow cell attachment sequence, sequencing primer sequence, and staggered, varying length barcode have been built into *PstI*-compatible adapter while the reverse adapter contained the flow cell fitting region and the overhang sequence that was *MseI*-compatible. PCR was used to amplify only mixed fragments (*PstI*-*MseI*) following these reactions conditions: initial denaturation stage (1 min at 94°C); denaturation stage (30 rounds each consisting of 20 s at 94°C); annealing stage (30 s at 58°C); extension stage (45 s at 72°C); and finally, an extension step of 7 min at 72°C. Following PCR, equimolar amounts of amplification products were bulked and applied to c-Bot bridge PCR for each sample of the 96-well microtiter plate and sequenced on Illumina Hiseq2500. Single read sequencing took place over 77 cycles. Sequences produced by every lane were processed using proprietary DAoT analytical pipelines. In order to filter poor-quality sequences, the FASTQ files were first processed in the primary pipeline with more stringent selection criteria compared to the remaining sequence in the barcode region. The sequences of specific samples contained in the barcode split-step were therefore assigned more consistently. The barcode/sample sequences were used in the marker calling with ~2,500,000 (±7%). At the end, the same sequences were collapsed into “fastcall files.” and these files were used in the secondary pipeline for DAoT P/Lxs proprietary SNP and SilicoDAoT (present = 1 and absent = 0) calling algorithms (DAoTsoft14). The sequence data were processed using analytical pipeline. SNP markers were aligned to the reference genome of Fonio CM05836 (*D. exilis*) (Abrouk et al., 2020) to identify chromosome positions.

Data Cleaning, Imputation and Filtering
TASSEL (Trait Analysis by aSsociation, Evolution and linkage) software version 5.2.56 (Bradbury et al., 2007) was used for data cleaning or quality control. Prior to the analysis of genetic parameters, tree filters were generated before missing data imputation, whereby loci with minor allele frequency (MAF) <0.05 along with loci departing from Hardy–Weinberg equilibrium at 5% and loci with a call rate <70 % were prune using the PCCA algorithm in KDCompute (https://kdcompute. igss-africa.org/kdcompute/login). The imputed and filtered data were used for structure and genetic relatedness analyses, the analysis of molecular variance (AMOVA), and as well as the discriminant analysis of principal components (DAPC).

Genetic Diversity Analysis
Parameters of genetic diversity such as observed and expected heterozygositities and inbreeding coefficient were estimated at different levels including the entire panel, ecological regions and as well as country of provenance using the package “adegenet” (Jombart, 2008) implemented in R software version 3.6.1 (Core Team, R, 2019). The analysis of molecular variance (AMOVA) was estimated using the package “poppr” (Kamvar et al., 2014) implemented also in R software.

Population Structure Analysis
In order to cluster fonio accessions onto different groups, population structure was investigated using DAoTseq-derived SNP markers distributed across the fonio genome using the Bayesian clustering approach in STRUCTURE version 2.3.4 (Pritchard et al., 2000). The analysis included three ecological regions (e.g., arid, semi-arid and sub-humid zones) as putative phytogeographic origins of accessions. A burn-in period of 10,000 Markov Chain Monte Carlo (MCMC) iterations and 10,000-run length, with an admixture model following Hardy-Weinberg equilibrium and correlated allele frequencies were considered in the structure analysis (Evanno et al., 2005). For each value of K (number of subpopulations), 10 independent runs were performed ranging from 1 to 7. The Structure Harvester (Earl, 2012) was used to analyze the results from the Structure, which enabled the identification of the best K-value as the distinct peak in the change of likelihood (ΔK) (Evanno et al., 2005). The extent of genetic differentiation (Fst) between the three ecological regions was estimated with Structure harvester simulation results (Supplementary File 1). DAoTseq SNP data were numerically coded as follows: A = 1, T = 2, C = 3, G = 4 as suggested in Structure V2.3.4 user manual.

Genetic Relatedness and Population Differentiation
A Neighbor-Joining (NJ) dendrogram was drawn using Darwin software V6.0.21 (Perrier and Jacquemoud-Collet, 2016), in order to investigate the genetic relatedness among accessions. Prior to the analysis, the 259 accessions were grouped according to their country of provenance to describe the distribution in the identified clusters. Furthermore, discriminant analysis of principal components (DAPC) was performed to infer and describe clusters in population of genetically related individuals.
using \texttt{adegenet} (Jombart, 2008) and \texttt{hierfstat} (Goudet, 2005) packages implemented in R V.3.6.1 (Core Team, R, 2019). In DAPC, data were set first for principal components analysis (PCA) and subsequently clustered using discriminant analysis (DA) with respective eigenvalues. Moreover, in order to estimate genetic differentiation among fonio clusters, Weir and Cockerham’s Fst-values (Weir and Cockerham, 1984) were calculated using R software.

**Linkage Disequilibrium (LD)**

Genome-wide LD patterns of known position of SNPs on the chromosomes (Huang et al., 2012) out of the complete set of 21,324 polymorphic markers were estimated using TASSEL (Trait Analysis by aSSociation, Evolution and Linkage) V5.3.1 software (Bradbury et al., 2007). LD patterns between markers was estimated using $r^2$, which is the correlation between alleles at two loci, as well as an indicator for evaluating the resolution of association approaches. Farnir et al. (2000) stated that for all possible combination of alleles and then weighting them according to the allele’s frequency, a weighted mean of $r^2$ was calculated between the two loci. In order to test the LD’s significance, we also obtained $P$-values determined by permutational test to calculate the ratio of permuted gamete distributions that was less likely than the gamete distribution observed under the null independence hypothesis (Weir, 1996). Furthermore, the rate of LD decay was drawn as the variation $r^2$ and the physical genetic distance (Mb) between pairs of SNP markers using “dartR” (Gruber et al., 2018) and “Sommer” (Covarrubias-Pazaran, 2016) packages implemented in R.

**RESULTS**

**Genetic Diversity**

A total of 21,324 codominant single nucleotide polymorphisms (SNPs) were obtained from the DArTseq analysis with 46.12% of missing data of the overall fonio accessions. The call rate ranged from 37.83 to 100% with a mean of 53.87% ± 0.080. These markers showed high reproducibility varying from 0.90 to 1.00 with an average of 0.99 ± 0.011. The minor allele frequency (MAF) ranged from 0.002 to 0.5 with a mean of 0.153 ± 0.08 while the major allele frequency (MaF) varied from 0.5 to 0.99 with an average of 0.84 ± 0.08 (Table 1). Furthermore, 17,821 markers (83.57%) had minor allele frequency >0.05 and...
the majority (83.83%) had a polymorphic information content (PIC) value >0.1 with a mean value of 0.25 ± 0.11. About 80% (17,045) of SNPs were aligned to the reference genome of fonio CM05836. These markers were almost uniformly distributed across the genome with the most SNPs observed on chromosome 09B (2,004 markers) followed by chromosome 05B (1,590 markers). The lowest numbers of SNP markers were observed on chromosomes 06A and 08A with 406 and 386 markers, respectively (Supplementary Figure 1).

The discovered SNPs were also categorized according to nucleotide substitutions as either transitions (A→G or C→T) or transversions (A→C, C→G, A→T, G→T). Our results of nucleotide changes showed that guanine (G) and cytosine (C) nucleotides mutate at 1.2-fold higher rates than adenine (A) and thymine (T) nucleotides (Table 2). The results also indicated that at a higher proportion of missing data, both minor allele frequency (MAF) and proportion of heterozygous reached lower values (Figure 2). Six hundred and eighty-eight (688) putative out of 21,324 SNP markers obtained after filtering and imputation were used for downstream analyses. The gene diversity (GD) analysis revealed that the entire panel varied from 0.09 to 0.50 with an average of 0.30 ± 0.12. Moreover, among the three ecological regions, the highest level of genetic diversity was observed in accessions collected from sub-humid ecology (0.201), and the lowest in the arid ecology (0.004). Also, based on country of provenance, accessions from Guinea and Nigeria showed the highest level of genetic diversity with respectively 0.209 and 0.200 while the least level was observed in accessions from Niger (0.003) (Table 3). The AMOVA results indicated that a large proportion of total genetic variation (67%) segregates among fonio accessions rather than between ecological zones, as well as the country of provenance (Table 4).

### Population Structure and Genetic Relatedness

The outputs of the structure analysis indicated that the likelihood of Delta K (ΔK) peaked at K = 2 (Figure 3A) revealing that two sub-populations (population I and population II) showed the highest probability for population clustering (Figure 3B). The distribution of fonio accessions among the two sub-populations revealed that population II had more fonio accessions (61.39%) than population I (38.61%). The accessions from population I were mainly composed of accessions from arid and semi-arid ecologies (Nigeria, Mali and Niger) whereas most accessions from population II were from semi-arid and sub-humid ecologies (Benin, Guinea, Burkina Faso). In addition, both of the two sub-populations diverged significantly as explained by the mean value of fixation index (Fst) for each of the sub-populations with 0.355 and 0.935 for population I and population II respectively (Table 5).

Furthermore, the Neighbor-Joining dendrogram results based on dissimilarity among 259 fonio accessions revealed also two heterogenetic clusters confirming the results of the Structure analysis. Cluster I (45.18%) was mainly composed of accessions originating from Benin, Niger and Mali while Cluster II (54.82%) was essentially composed of accessions from Nigeria, Guinea, Mali, and Burkina Faso (Figure 4).

### Population Differentiation

The degree of differentiation among fonio genotypes from different ecological regions and countries of provenance are shown in Table 6. Significant genetic differentiation was found among fonio accessions from different ecologies as indicated by the overall Weir and Cockerham’s Fst-value of 0.24. Pairwise Fst-values varies from 0.19 to 0.71. The lowest Fst-value (0.19) was observed between arid and sub humid ecologies while the highest Fst-value (0.71) was observed between arid and sub-humid ecologies. However, the overall Weir and Cockerman’s Fst-value based on country of provenance of fonio accessions was merely higher (0.26) compared to that of different ecological regions (0.24). Furthermore, pairwise Fst-values varied from 0.10 to 0.59. The highest Fst-value was observed between Nigeria and Benin (0.59) while the lowest value was observed between Burkina Faso and Guinea (0.10) (Table 6).

Discriminant analysis of principal components (DAPC) was estimated using the detected number of clusters. The 10 first PCs (60% of variance) of PCA and two discriminant eigenvalues were retained (Figure 5). The results showed that fonio accessions can be clustered in three groups and discriminant function 1 clearly differentiated cluster 2 from cluster 3. Cluster 1 consisted of 66.02% (171 accessions), while cluster 3 and cluster 1 had 20.46% (53 accessions) and 13.52% (35 accessions), respectively. Accessions from Benin, Burkina Faso, Mali, and Niger were more dominant in cluster 1 whereas the majority of accessions from clusters 2 and 3 were originated from Nigeria and Guinea. Results of pairwise Fst revealed high genetic differentiation among clusters with overall Weir and Cockerham’s Fst-value of 0.68. Pairwise Fst-values ranged from 0.31 to 0.72 (Supplementary Table 1). The lowest Fst-value (0.31) was found

---

**Table 1**: Genetic properties of markers.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>CR</th>
<th>MAF</th>
<th>MaF</th>
<th>GD</th>
<th>PIC</th>
<th>RP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>0.3783</td>
<td>0.002</td>
<td>0.5</td>
<td>0.09</td>
<td>0.00</td>
<td>0.90</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.00</td>
<td>0.5</td>
<td>0.99</td>
<td>0.50</td>
<td>0.5</td>
<td>1.00</td>
</tr>
<tr>
<td>Mean</td>
<td>0.5387</td>
<td>0.153</td>
<td>0.84</td>
<td>0.30</td>
<td>0.25</td>
<td>0.99</td>
</tr>
<tr>
<td>SD</td>
<td>0.080</td>
<td>0.08</td>
<td>0.08</td>
<td>0.12</td>
<td>0.11</td>
<td>0.011</td>
</tr>
</tbody>
</table>

CR, call rate; MAF, minor allele frequency; MaF, major allele frequency; GD, total genetic diversity; PIC, polymorphic information content; RP, reproductibility; SD, standard deviation.
between cluster 2 and cluster 3. An intermediate Fst-value (0.69) was observed between cluster 1 and cluster 2. Moreover, the highest Fst-value (0.72) was detected between cluster 1 and cluster 3.

Linkage Disequilibrium (LD)
LD was analyzed across the fonio genome. Of the 21,324 SNPs, 17,045 (80%) covering the entire genome were in true LD based on the Comparison between pairs of markers. The LD patterns analysis revealed that the large red blocks of haplotypes along the diagonal of the triangle plot indicate the high level of LD between the loci in the blocks (Figure 6).

The LD decay plot showed that few LD values were found over a short physical genetic distance which decreased rapidly, with very low decline at longer distances between markers across the genome. In the present study, LD decayed rapidly below $r^2 = 0.64$, reaching 0.56 at a distance of 5 Mb (Figure 7).

FIGURE 2 | Histograms showing basic SNP characteristics: (A) proportion of missing data, (B) minor allele frequency, (C) proportion of heterozygosity.

DISCUSSION
SNP Markers Discovery and Genetic Diversity
Genotype-by-sequencing (GBS) provides new technologies for breeders with cost-effective, high-depth coverage, and high throughput sequencing platforms (Elshire et al., 2011; Fu and...
TABLE 3 | Genetic diversity estimations of 259 fonio accessions based on ecological regions and country of provenance.

<table>
<thead>
<tr>
<th>Population</th>
<th>Number of individuals</th>
<th>Observed heterozygosity</th>
<th>Expected heterozygosity</th>
<th>Inbreeding coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire panel</td>
<td>259</td>
<td>0.026</td>
<td>0.274</td>
<td>0.87</td>
</tr>
<tr>
<td>Ecology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arid</td>
<td>50</td>
<td>0.004</td>
<td>0.005</td>
<td>0.22</td>
</tr>
<tr>
<td>Semi-arid</td>
<td>183</td>
<td>0.018</td>
<td>0.164</td>
<td>0.86</td>
</tr>
<tr>
<td>Sub-humid</td>
<td>26</td>
<td>0.042</td>
<td>0.201</td>
<td>0.79</td>
</tr>
<tr>
<td>Country of provenance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benin</td>
<td>42</td>
<td>0.004</td>
<td>0.057</td>
<td>0.92</td>
</tr>
<tr>
<td>Burkina Faso</td>
<td>4</td>
<td>0.005</td>
<td>0.006</td>
<td>0.19</td>
</tr>
<tr>
<td>Guinea</td>
<td>29</td>
<td>0.036</td>
<td>0.209</td>
<td>0.83</td>
</tr>
<tr>
<td>Mali</td>
<td>28</td>
<td>0.006</td>
<td>0.005</td>
<td>0.17</td>
</tr>
<tr>
<td>Niger</td>
<td>28</td>
<td>0.004</td>
<td>0.003</td>
<td>0.32</td>
</tr>
<tr>
<td>Nigeria</td>
<td>128</td>
<td>0.028</td>
<td>0.200</td>
<td>0.88</td>
</tr>
</tbody>
</table>

TABLE 4 | Analysis of molecular variance (AMOVA) results.

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum of Squared</th>
<th>Mean Squared</th>
<th>Variance</th>
<th>% Est.variance</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Population (ecological regions)</td>
<td>2</td>
<td>4512.53</td>
<td>2256.26</td>
<td>29.91</td>
<td>28.66</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Between Subpopulation (country of provenance)</td>
<td>5</td>
<td>820.04</td>
<td>164.006</td>
<td>4.72</td>
<td>4.52</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Within samples (accessions)</td>
<td>251</td>
<td>17503.34</td>
<td>69.73</td>
<td>69.73</td>
<td>66.82</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Total</td>
<td>258</td>
<td>22835.91</td>
<td>88.51</td>
<td>104.36</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 3 | The STRUCTURE results of 259 fonio accessions using 688 SNP markers: (A) the optimal value of K (K = 2); (B) Bar plot of accessions showing two subpopulations (Population I = red color and Population II = green color).

Peterson, 2011; Egan et al., 2012; Poland and Rife, 2012; Mir et al., 2013; Kim et al., 2016; Voss-Fels and Snowdon, 2016. The high throughput multiplexing genotyping platforms used in this study for assessing genetic diversity and population structure in a panel of 259 fonio accessions yielded 21,324 quality SNP markers. The relatively high proportion of missing data observed in this study might be due to the low read coverage in this dataset during genotyping because SNPs with low read coverage are more prone to errors than SNPs with high read coverage (Furuta et al., 2017). This shows the importance of sequencing to higher read depths to minimize missing data. On the other hand, this could also be explained by DNA contamination during leaf sampling which affects the quality of DNA. This result is in accordance with that of Taberlet et al. (1996) who reported that both DNA quality and or low quantity cause genotyping errors by favoring allelic dropouts and false alleles. Furthermore, in this study, the observed nucleotide changes may be due to inherent characteristics of DNA (e.g., cytosine deamination) that often cause a mutational bias in plant genomes (Gaut et al., 2011). However, the mutation rate obtained in this study was
lower compared to that reported earlier in maize (Morton et al., 2006).

Previously, genetic diversity of fonio millet was investigated with a different type of markers and based on small data sets limiting the chance of getting high polymorphic and informative markers (Hilu et al., 1997; Adoukonou-Sagbadja et al., 2007; Animasaun et al., 2018). The results revealed that the markers developed in this study were highly reproducible, with a high call rate and high polymorphism information content suggesting that these markers are very important for future fonio genetic improvement. This shows the extreme importance of genotyping by sequencing analysis for SNPs discovery, harnessing genetic diversity, and genotyping in most orphan small cereals where reference genome is not yet available. The high polymorphic markers obtained in this study are within the range of values reported in earlier SNP-based genetic diversity studies in other small millets (Kumar et al., 2016; Johnson et al., 2019). However, since genetic diversity is a true indicator of the degree of genetic variance within individuals in a population, in this study a moderate level of genetic diversity was observed among fonio accessions. This was further supported by the result of AMOVA, with a large proportion of total genetic variance found within fonio accessions reflecting that to some extent our collection has captured the diversity in the crop. This result is also consistent with that reported earlier by Adoukonou-Sagbadja et al. (2007) who assessed genetic diversity of 122 fonio accessions using AFLP markers. Furthermore, compared to these results a lower level of genetic diversity was reported in foxtail millet (Wang et al., 2010). This could be attributed to factors such as inbreeding due to the self-pollinated nature and in other hand the small population size (Barnaud et al., 2012; Abrouk et al., 2020). These factors have been reported to contribute to a reduction in genetic diversity in crop plants (Bhandari et al., 2017). The results revealed also that the genetic diversity in D. exilis is concentrated in sub humid and semi-arid ecologies of West Africa comprising the Upper Niger basin (Guinea and Mali), Northwestern Benin Republic,
TABLE 6 | Pairwise Fst analysis results among ecological regions and countries of provenance of 259 fonio millet accessions.

<table>
<thead>
<tr>
<th>Ecology</th>
<th>Arid</th>
<th>Semi-arid</th>
<th>Sub-humid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arid</td>
<td>-</td>
<td>0.19*</td>
<td></td>
</tr>
<tr>
<td>Semi-arid</td>
<td>0.19*</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Sub-humid</td>
<td>0.71**</td>
<td>0.19*</td>
<td>-</td>
</tr>
</tbody>
</table>

Overall Weir and Cockerham’s Fst-value = 0.24*

Country of provenance

<table>
<thead>
<tr>
<th>Country of provenance</th>
<th>Mali</th>
<th>Niger</th>
<th>Benin</th>
<th>Guinea</th>
<th>Burkina Faso</th>
<th>Nigeria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mali</td>
<td>-</td>
<td>0.29*</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Niger</td>
<td>0.29*</td>
<td>-</td>
<td>0.28*</td>
<td>0.32**</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Benin</td>
<td>0.28*</td>
<td>0.52**</td>
<td>-</td>
<td>0.28*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Guinea</td>
<td>0.30*</td>
<td>0.27*</td>
<td>0.13*</td>
<td>0.10*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Burkina Faso</td>
<td>0.07*</td>
<td>0.27*</td>
<td>0.13*</td>
<td>0.10*</td>
<td>-</td>
<td>0.17*</td>
</tr>
<tr>
<td>Nigeria</td>
<td>0.29**</td>
<td>0.53***</td>
<td>0.59***</td>
<td>0.27**</td>
<td>0.17*</td>
<td>-</td>
</tr>
</tbody>
</table>

Overall Weir and Cockerham’s Fst-value = 0.26**

* = significant at 5%, ** = significant at 1%, *** = significant at 0.1%.

Southwestern Burkina Faso and Northern Nigeria. This result confirms the Upper Niger basin as center of diversity of *D. exilis* as reported earlier by Portères and Harlan (1976). However, the level of genetic diversity observed in Nigeria suggests the proposition of Northern Nigeria as the second center of diversity of white fonio. The low genetic diversity found in the arid ecology of Niger and part of Mali may therefore be related to the founder effect that occurs when a small group of individuals that is not genetically representative of the population from which they came establish in a new area. Similar results were reported by Adoukonou-Sagbadja et al. (2007) who stated that founder effect was the main cause of low genetic variation observed in white fonio from the Atacora mountain zone (Benin and Togo areas).

**Clustering and Genetic Relatedness of Fonio Accessions**

Structure analyses revealed two divergent subpopulations while the Neighbor-Joining dendrogram clustered fonio genotypes in two distinct groups based on ecological regions and country of provenance, respectively. On the contrary, previous studies revealed that fonio accessions clustered in three groups and six subpopulations based on principal component analysis (PCA) and structure analysis, respectively (Abrouk et al., 2020). This difference in clustering fonio accessions could be attributed to the sample size and geographical sources where fonio accessions were collected, as well as the number of SNP markers used. For instance, in our study, we used 21,324 SNP markers and
FIGURE 6 | LD patterns plot based on 17,045 SNP markers. The relationship between two markers is represented by each cell with the color codes for the presence of significant LD. In both diagonals, colorful bar codes are shown for the significant threshold levels.

259 fonio accessions collected from six West African countries while in previous studies, Abrouk et al. (2020) used 11,046,501 SNP markers and 166 *D. exilis* accessions collected from seven West African countries (Guinea, Mali, Benin, Togo, Burkina Faso, Ghana, and Niger). Nonetheless, Wang et al. (2021) used 13,546 SNP markers and identified three clusters while assessing the genetic diversity of 130 fonio accessions collected from Mali and Niger based on principal coordinates and dendrogram relationship analyses. This confirms our results of discriminant analysis of principal components, although the numbers of countries where fonio germplasm was collected, sample size, and SNP markers were lower compared to in our study. This shows the importance of diversifying sources of germplasm collection and using dense markers in order to capture the maximum diversity. Furthermore, the discrepancies between the results of our study and those of Adoukonou-Sagbadja et al. (2007) may be due to different type of markers (SNP vs. AFLPs markers) and the small data sets used, although fonio accessions were collected from nearly the same countries of provenance.

Varshney et al. (2007) reported that AFLP and SSR markers were more appropriate for fingerprinting and diversity analysis while SNP markers were considered as the best class for characterizing and conserving the genebank materials. Importantly, SNP markers can be digitalized or represented as graphic formats (Garafutdinov et al., 2020) across different genotypes making them the markers of choice for genetic diversity studies with recent advances in next generation sequencing technologies (NGS). The results of clustering fonio accessions in different groups demonstrate the influence of environmental factors and human activities on genetic variability as reported earlier (Abrouk et al., 2020). According to Wondimu et al. (2021), populations were structured based on geographical
patterns and botanical races in sorghum. This is supported by the finding of Deu et al. (2010) who stipulated that seed exchange and food preferences were other factors that affected population structure in sorghum. However, in this study fonio landraces clustered irrespective of their ecological regions and country of provenance backgrounds implying that alleles from different sources co-occur within individuals. This shows the importance of admixture, a phenomenon that potentially increases genetic variation by decreasing inbreeding depression (Shi et al., 2018). This phenomenon could likely be due to the existence of gene flow through germplasm exchanges that frequently occur between different regions as reported earlier (Sidibé et al., 2020). This result is consistent with that of Kabbaj et al. (2017) who reported among wheat genotypes including landraces, varieties, and elite lines collected from diverse origins higher admixtures. The difference observed between the two clustering approaches Structure model and DAPC analysis could be attributed to the sensitivity of the Bayesian clustering algorithm to sample size, number of populations, number of loci scored and the type of markers (Evanno et al., 2005). On the other hand, with DAPC a considerable inaccuracy in terms of both the suggested number of clusters and the placement of individuals in those clusters was reported when the migration rates were high (Miller et al., 2020).

**Regional Differentiation**

A large genetic differentiation was found among fonio accessions based on country of provenance compared to that of ecological regions suggesting that the regional differentiation of fonio accessions is a country based rather than ecological regions. This may be likely due to farmers’ management practices that are specific to ethnolinguistic groups because fonio accessions used in this study were provided by different ethnic groups of farmers from different countries. This is supported by the findings of Abrouk et al. (2020) who cited farmers’ ethnolinguistic groups as among the factors that shaped genetic diversity in fonio. A possible explanation can be related to the self-pollination nature of the crop (Barnaud et al., 2012) that limits gene flow among fonio genotypes. This result corroborates the findings of Adoukonou-Sagbadja et al. (2007) who reported a high degree of differentiation within fonio accessions. Conversely, a low level of regional differentiation was reported by Wondimu et al. (2021) for 338 cultivated sorghum in Ethiopia. However, the low degree of differentiation observed between semi-arid and sub-humid zones can be explained by seed exchanges among fonio farmers due to the proximity of these zones. This finding is in agreement with that reported previously by Akohoue et al. (2020) in Kersting's groundnut and confirms the high genetic distance observed between arid and sub-humid ecologies probably due to isolation by distance (Wright, 1946). Furthermore, the great divergence observed between fonio accessions from Nigeria and the remaining countries could be also explained by geographic isolation leading to speciation which is the formation of distinct species in the course of evolution. Surprisingly, fonio accessions from Nigeria and Benin Republic for instance were collected from the same ecology (semi-arid region), but the highest divergence among fonio was observed between these two countries confirming the hypothesis that regional differentiation was country-based. Since the two groups of population (Nigeria and Benin) are thriving in the same ecology, the divergence may likely arise from management practices firstly because farmers are from different ethnolinguistic groups (e.g., Hausa ethnic group from Northern Nigeria and Ditamari from the Northwestern Benin Republic). This is further supported by Abrouk et al. (2020) who highlighted the strong impact of anthropogenic
factors on shaping the genetic diversity of fonio. Thus, future research should investigate drivers that influence management practices of fonio millet (*D. exilis*) across different ethnolinguistic groups of West Africa. Secondly, the two populations might have experienced different pressures of biotic and abiotic stresses and thirdly, the informal seed systems that limit the seed exchange of improved fonio varieties (Adoukonou-Sagbadja et al., 2006; Sekloka et al., 2015) may probably cause this genetic divergence.

**Linkage Disequilibrium (LD) Patterns**

Even though any quantitative traits loci (QTLs) and/or markers associated with different traits on the chromosomes have been reported, yet in fonio the results show potential markers distributed on different chromosomes in the genome that were in true LD, paving ways for genome-wide association studies and genomic selection. Hence, further investigation should be made in order to map genomic regions associated with important well-preferred traits for both farmers and consumers. Furthermore, the results indicated a high level of non-random association between the loci in the blocks suggesting that there has been a low or no recombination rate since LD block formations. Moreover, our finding revealed that on average fonio germplasm was characterized by a high linkage disequilibrium with slow LD decay over long physical distances between markers. This could be attributed to the mating system of the crop because selfing limits heterozygosity, while increasing homozygosity with the consequence of reducing the effectiveness of recombination rate. As a result LD is expected to be maintained over long physical distances (Gaut and Long, 2003). Similar results were reported in wild barley (Morrell et al., 2005). The application of artificial hybridization techniques between distant genetic parents, mutagenesis and transgenesis could potentially increase recombination rate, thereby increase genetic diversity among fonio germplasm. In addition, the results of LD patterns provide relevant information on the recombination history under natural selection since no breeding efforts through artificial hybridization have been reported yet. Therefore, future LD patterns analysis should be investigated in-depth in order to shed light on other factors that alter certain genomic regions leading to the differentiation in the LD patterns among fonio accessions.

**CONCLUSION**

The present study demonstrated the potential of genotyping by sequencing (GBS) to discover SNPs, to assess genetic diversity and population structure of fonio millet accessions. The results indicated also that grouping of fonio accessions was irrespective of geographical origin backgrounds while the regional differentiation was based on the country of provenance rather than the ecological regions. Furthermore, the high genetic divergence observed between accessions from Nigeria to the other countries particularly the Benin Republic suggests geographic isolation. Moreover, the relatively high non-random association between alleles at two loci observed in this study could be attributed to the selfing nature of the crop, which reduces the recombination rate. Finally, our findings highlighted the importance of population structure, and Linkage Disequilibrium patterns in genetic diversity study and provide relevant information for association mapping and genomic selection in fonio millet.

**DATA AVAILABILITY STATEMENT**

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: Dryad and https://doi.org/10.5061/dryad.vmcvdncsr.

**AUTHOR CONTRIBUTIONS**

AI and EA-D conceptualized the study, wrote the manuscript, and edited the manuscript. AI, KI, and AM carried out data analysis. EA-D, CA, KI, AM, HO, MG, and CB critically revised the draft and updated the manuscript for publication. All authors contributed to the article and approved the submitted version.

**FUNDING**

AI, from the National Institute of Agronomic Research of Niger (INRAN), is a scholar of the Intra-Africa Academic Mobility Scheme under the project grant number 2016–2988 on Enhancing training and research mobility for novel crops breeding in Africa (MoBread) funded by the Education, Audiovisual and Culture Executive Agency (EACEA) of the European Commission. The project provided a scholarship for academic training and research mobility and a research grant to the AI to complete a PhD degree at the University of Abomey-Calavi (Republic of Benin).

**ACKNOWLEDGMENTS**

The authors wish to thank MoBread program for the financial support. We are thankful to Prof Simon Krattinger and Michael Abrouk for providing data for the annotation of fonio CM05836 genome. We also thank Félicien Akohoue and Florence Kombieni from GBioS Lab for their help in leaf DNA sampling.

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsufs.2021.699549/full#supplementary-material

**REFERENCES**


Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.