Article

Assessment of Genetic Diversity and Phylogenetic Relationship of Local Coffee Populations in Southwestern Saudi Arabia using DNA Barcoding

Habib Khemira[1]*, Mosbah Mahdhi[1,2], Muhammad Afzal[3], Mohammed Oteef[4], Taieb Tounekti[5], Zarraq Al-Faifi[6] and Wail Solami[6]

[1]Centre for Environmental Research and Studies, Jazan University, Jazan 45142, P.O. Box 114, Saudi Arabia.
[2]Laboratory of Biodiversity and Valorization of Bioresources in Arid Zones, Faculty of Sciences of Gabes, University of Gabes, Gabes, Tunisia.
[3]Department of Plant Production, College of Food and Agricultural Sciences, King Saud University, Riyadh, Saudi Arabia.
[4]Department of Chemistry, College of Science, Jazan University, Jazan 45142, P.O. Box 114, Saudi Arabia.
[5] Laboratory of Process Engineering & Industrial Systems (LR11ES54), National Engineering School of Gabes, University of Gabes, Gabes 6029, Tunisia.
[6]Department of Biology, College of Science, Jazan University, Jazan 45142, P.O. Box 114, Saudi Arabia.
* Correspondence: author: Hkhemira@jazanu.edu.sa; habibkhemira@yahoo.com.

Abstract: The genetic diversity of local coffee populations is crucial to breed new varieties better adapted to the increasingly stressful environment due to climate change and the evolving consumer preferences. Unfortunately, local coffee germplasm conservation and genetic assessment have not received much attention. Molecular tools offer substantial benefits in identifying and selecting new cultivars or clones suitable for sustainable commercial utilization. New annotation methods, such as chloroplast barcoding, are necessary to produce accurate and high-quality phylogenetic analyses. This study used DNA barcoding techniques to examine the genetic relationships among fifty-six accessions collected from the southern part of Saudi Arabia. PCR amplification and sequence characterization were used to investigate the effectiveness of four barcoding loci, namely atpB-rbcl, TrnL-TrnF, TrnT-TrnL, and TrnL. The maximum nucleotide sites, nucleotide diversity, and an average number of nucleotide differences were recorded for atpB-rbcl, while TrnT-TrnL had the highest variable polymorphic sites, segregating sites, and haploid diversity. Among the four barcode loci, TrnT-TrnL recorded the highest singleton variable sites, while TrnL recorded the highest parsimony information sites. Furthermore, the phylogenetic analysis clustered the *Coffea arabica* genotypes into four different groups, with three genotypes (KSA31, KSA38, and KSA46) found to be the most divergent genotypes standing alone in the cluster and remained apart during the analysis. The study demonstrates the presence of considerable diversity among coffee populations in Saudi Arabia. Furthermore, it also shows that DNA barcoding is an effective technique for identifying local coffee genotypes, with potential applications in coffee conservation and breeding efforts.

Keywords: *Coffea arabica*; chloroplastic DNA; barcode; genetic diversity; population structure; Saudi Arabia

1. Introduction

Coffee is one of the most commercially significant crops, and the second most traded commodity after oil [1]. In addition to its high export value, coffee has also gained in cultural significance over the past few decades. Despite there being more than 125 reported species in the genus *Coffea*, only two species, *Coffea arabica* L. (also known as Arabica coffee) and *C. canephora Pierre ex* A. Froehner (known as Robusta coffee) are grown commercially [2]. The total annual global coffee production in 2020 was over 105.21 million tons, with 63.15 million tons of Arabica coffee accounting for 60.02% of the total, and 42.06 million tons of Robusta coffee comprising 39.97%. Coffee's genetic development is progressing at a sluggish pace despite its enormous economic relevance. The collection, characterization, and wise use of accessible germplasm material for any crop plant species contribute to its genetic development and long-term viability [3]. Therefore, enhancing diversity from both local and foreign sources is critical for the improvement of crops [4]. For historic reasons, the main issue with Arabica coffee has been its narrow genetic base that limits its adaptation to changing environments [2]. To get around this problem, breeders made use of wild coffee diploid species to introduce

new genes into Arabica genotypes [2]. For instance, the leaf rust resistant Arabica cultivar Timor Hybrid got its resistance from its *C. canephora* parent; it was later used as a parent to develop several new rust resistant cultivars such as Catimor and Ruiru 11 [5]. For bean and liquor quality traits, the wild tetraploid Arabica genotypes from the species' center of origin and the little-known ancient varieties from the Arabian Peninsula offer a wide gene pool to explore [6]. Despite the potential importance of coffee heirlooms in the Arabian Peninsula as a source of genetic diversity, there is limited information available on these genotypes. This information is essential for the development of new coffee varieties that can better adapt to changing environmental conditions, such as pest and disease pressure, abiotic stress, and changes in consumer preferences [7]. Furthermore, since over 60% of wild coffee species are in danger of extinction due to accelerated environmental change, gathering complete information, and characterizing this germplasm is of utmost importance [8]

Another issue facing the coffee industry as it struggles to cope with an over-supplied market, is adulteration. It has long been known that coffee is often adulterated with less expensive and readily available plant material [9]. Coffee adulteration has become a more serious issue for the industry in recent years due to the significant expansion in the variety of coffee recipes, stores, and ultimately consumers [10]. Therefore, developing molecular means like genetic barcodes to identify and authenticate the varieties can help mitigate the problem.

In Saudi Arabia and Yemen, *C. arabica* has been cultivated for at least four centuries on the terraced slopes and narrow valleys of the western mountains at different altitudes ranging mostly from 1200 to 2000m above sea level [11,12]. Most of what is grown now in southwestern Saudi Arabia are old cultivars that have been around for hundreds of years [12]. It is likely that these diverse populations are a result of successive introductions of genetic material from Eastern Ethiopia by Arab traders over centuries of uninterrupted exchange across the narrow strait of Bab El-Mandeb [13]. Therefore, it is safe to assume that the southwestern corner of the Arabian Peninsula contains the most genetic diversity of *C. arabica* outside the species' center of origin in the Ethiopian highlands [6]. Unfortunately, these genetic resources have not attracted enough attention from the scientific community, except for the 1989 FAO expedition to southern Yemen led by A. Eskes [14] and two recent studies by Tounekti et al. [12] and Montagnon et al. [6]. These three studies reported the existence of considerable diversity among coffee populations in the Arabian Peninsula. It is worth noticing that the present coffee populations have evolved over hundreds of years in a semi-arid environment [15] marked by recurring droughts, uneven distribution of rainfall, heat stress and high irradiance. Therefore, it is expected that these genotypes could be the source of interesting genes that confer stress tolerance [16].

The molecular phylogenies of coffee species have been established using variations in intergenic spacer sequences [17-19] and introns [20] of plastid DNA, as well as internal transcribed spacer (ITS) sequences of rDNA [21] and a combination of four plastid regions and ITS [22]. Chloroplast DNA (cpDNA) sequence variation is widely used in systematics and for making phylogenetic inferences at different taxonomic levels. Introns and intergenic spacers are known to exhibit high rates of mutation [23]. The trnT-trnL and trnL-trnF intergenic spacers, the trnL intron, and the atpB-rbcL intergenic region are useful for evolutionary studies at low taxonomic levels, and these regions have been extensively used in phylogenetic studies to analyze cytoplasmic polymorphism and to study the demographic history of several species [23,24,25,26,27].

Overall, further research is necessary to fully comprehend the diversity and potential of diploid and tetraploid coffee species and to utilize this information to develop new coffee varieties that can better meet the needs of farmers and consumers in the future. The present study aims to estimate the genetic diversity of local coffee populations in Saudi Arabia and to examine their genetic relatedness using chloroplast intergenic spacer markers.

2. Results

The successful amplification of all four intergenic spacer barcode sequences (atpB-rbcl, TrnT-TrnL, TrnL-trnF, TrnL) was achieved, resulting in a single band of the expected size (see supplementary Figure). The respective sequences for each barcode were submitted to the National Center for Biotechnology Information (NCBI) via Bankit submission, and accessions numbers for each barcode sequence are presented in Table 2. This study utilized the *Coffea arabica* (MN894552.1) chloroplast genome sequence, available on NCBI, as a reference genome sequence to detect polymorphisms among DNA sequences in 56 *Coffea arabica* accessions across all four barcoding primers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 June 2023 doi:10.20944/preprints202306.1304.v1

Table 2. Accession numbers of four barcode primers of 56 *Coffea arabica* genotypes.

Genotype ID	atpB-rbcl	TrnL-TrnF	TrnT-TrnL	TrnL
KSA1R	OQ718327	OQ914867	OQ914923	OQ953999
KSA2R	--------	OQ914868	OQ914924	OQ954000
KSA3R	OQ844066	OQ914869	OQ914925	OQ954001
KSA4R	OQ914863	OQ914870	OQ914926	OQ954002
KSA5R	OQ914864	OQ914871	OQ914927	OQ954003
KSA6R	OQ914865	OQ914872	OQ914928	OQ954004
KSA7R	OQ914866	OQ914873	OQ914929	OQ954005
KSA8R	OQ850301	OQ914874	OQ914930	OQ954006
KSA9R	OQ850302	OQ914875	OQ914931	OQ954007
KSA10R	OQ850303	OQ914876	OQ914932	OQ954008
KSA11R	OQ850304	OQ914877	OQ914933	OQ954009
KSA12R	OQ850305	OQ914878	OQ914934	OQ954010
KSA13R	OQ850306	OQ914879	OQ914935	OQ954011
KSA15R	OQ851715	OQ914880	OQ914936	OQ954012
KSA16R	OQ851716	OQ914881	OQ914937	OQ954013
KSA17R	OQ851717	OQ914882	OQ914938	OQ954014
KSA18R	OQ851718	OQ914883	OQ914939	OQ954015
KSA19R	OQ851719	OQ914884	OQ914940	OQ954016
KSA20	OQ851720	OQ914885	OQ914941	OQ954017
KSA21	OQ872544	OQ914886	OQ914942	OQ954018
KSA22	OQ872545	OQ914887	OQ914943	OQ954019
KSA23	OQ872546	OQ914888	OQ914944	OQ954020
KSA24	OQ872547	OQ914889	OQ914945	OQ954021
KSA25	OQ872548	OQ914890	OQ914946	OQ954022
KSA26	OQ872549	OQ914891	OQ914947	OQ954023
KSA27	OQ872550	OQ914892	OQ914948	OQ954024
KSA28	OQ872551	OQ914893	OQ914949	OQ954025
KSA29	OQ872552	OQ914894	OQ914950	OQ954026
KSA30	OQ872553	OQ914895	OQ914951	OQ954027
KSA31	OQ872554	OQ914896	OQ914952	OQ954028
KSA32	OQ872555	OQ914897	OQ914953	OQ954029
KSA33	OQ872556	OQ914898	OQ914954	OQ954030
KSA34	OQ872557	OQ914899	OQ914955	OQ954031
KSA35	OQ872558	OQ914900	OQ914956	OQ954032
KSA36	OQ872559	OQ914901	OQ914957	OQ954033
KSA37	OQ872560	OQ914902	OQ914958	OQ954034
KSA38	OQ872561	OQ914903	OQ914959	OQ954035
KSA39	OQ872562	OQ914904	OQ914960	OQ954036

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 June 2023 doi:10.20944/preprints202306.1304.v1

KSA40	OQ872563	OQ914905	OQ914961	OQ954037
KSA41	----------	OQ914906	OQ914962	OQ954038
KSA42	----------	OQ914907	OQ914963	OQ954039
KSA43	----------	OQ914908	OQ914964	OQ954040
KSA44	OQ852764	OQ914909	OQ914965	OQ954041
KSA45R	OQ852765	OQ914910	OQ914966	OQ954042
KSA46	OQ852766	OQ914911	OQ914967	OQ954043
KSA47	OQ852767	OQ914912	OQ914968	OQ954044
KSA48	OQ852768	OQ914913	OQ914969	OQ954045
KSA49	OQ852769	OQ914914	OQ914970	OQ954046
KSA50	OQ852770	OQ914915	OQ914971	OQ954047
KSA51R	OQ852771	OQ914916	OQ914972	OQ954048
KSA52	OQ852772	OQ914917	OQ914973	OQ954049
KSA59	OQ852773	OQ914918	OQ914974	OQ954050
KSA60	OQ852774	OQ914919	OQ914975	OQ954051
KSA61	OQ852775	OQ914920	OQ914976	OQ954052
KSA62	OQ852776	OQ914921	OQ914977	OQ954053
KSA63	OQ852777	OQ914922	OQ914978	OQ954054

Note: KSA2; KSA41; KSA42; KSA43 were not identified in the data base for atpB-rbcl barcode

The percentage of polymorphic sites for each sequence was determined by dividing the number of variable nucleotides by the length of the entire region and multiplying the result by 100. Table 3 presents the number of nucleotide sites (NNS), variable polymorphic sites (VPS), number of segregating sites (NSS), number of haplotypes (NH), nucleotide diversity (ND), and average number of nucleotide differences (ANND) for each barcode primer and the cumulative results for all four primers. The combined sequences showed the highest NSS, followed by the atpB-rbcl primer, while the TrnL primer had the lowest NNS. The TrnT-TrnL primer had the highest VPS (341), followed by atpB-rbcl, while the lowest (154) was recorded for **TrnL-TrnF**. The atpB-rbcl had the highest ND, followed by TrnT-TrnL (0.051), with **TrnL-TrnF** showing the lowest ND. Additionally, the atpB-rbcl had the highest ANND (185.54), while **TrnL-TrnF** exhibited the lowest value (25.23) for ANND.

Table 3. Summary of nucleotide sites, variable polymorphic sites, number of segregating sites, haploid diversity, nucleotide diversity, and average number of nucleotide difference.

Barcode Name	**Individual**	**NNS**	**VPS**	**NSS**	**NH**	**ND**	**ANND**
atpB-rbcl	56	1139	341	341	17	0.54	185.54
TrnL	55	551	237	237	31	0.056	18.93
TrnL-TrnF	56	1055	154	154	17	0.046	25.23
TrnT-TrnL	56	988	421	421	50	0.051	40.50
atpB-rbcl + TrnL+ TrnL-TrnF + TrnT-TrnL	223	4114	651	651	37	0.11	295

 doi:10.20944/preprints202306.1304.v1

NNS=Number of nucleotide sites, VPS= variable polymorphic sites, NSS= number of segregating sites, NH= number of haplotypes, ND= Nucleotide diversity, ANND= average number of nucleotide difference

The nucleotide base composition of each barcode primer was determined and is presented in Table 4. The average nucleotide base composition of atpB-rbcl was recorded as 33.15% T(U), 16.60% C, 34.49% A, and 15.76% G. For TrnC, the composition was 26.7% T(U), 15.9% C, 37.6% A, and 19.8% G. TrnT-TrnL had a composition of 39.18% T(U), 13.88% C, 33.84% A, and 13.10% G (Table 4). The singleton variable sites (STVS) and parsimony information sites (PIS) for each chloroplast barcode are presented in Table 5. The TrnT-TrnL barcode recorded the highest number of STVS (338), followed by **TrnL-TrnF** (133), TrnL (52), and atpB-rbcl (1) had the lowest. Similarly, for PIS, TrnL had the highest number (155), followed by atpB-rbcl (137), while the minimum was recorded for **TrnL-TrnF** (45) across all sequences (Table 5). A phylogenetic analysis was constructed based on concatenated sequences of all four barcodes using the maximum likelihood method and Kimura 2-parameters model. The percentage of trees in which the associated taxa clustered together is shown next to the branches. This analysis involved 56 nucleotide sequences, and the final dataset comprised a total of 4114 positions. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The final phylogenetic tree divided the 56 accessions into five different groups.

 doi:10.20944/preprints202306.1304.v1

Table 4. Nucleotide base substitution matrix of four barcoding markers in Arabica coffee.

	Atpb-rbcl					TrnL					TrnC					TrnT-TrnL				
Genotypes	T(U)	C	A	G	Total	T(U)	C	A	G	Total	T(U)	C	A	G	Total	T(U)	C	A	G	Total
KSA1R	36.36	16.01	30.48	17.15	968	32.94	20.38	32.46	14.22	422	26.9	15.7	37.9	19.6	562	39.48	13.49	33.79	13.24	808
KSA2R						33.25	19.21	33.50	14.04	406	28.5	15.5	35.8	20.2	1017	39.58	14.02	33.75	12.66	806
KSA3R	35.60	15.89	31.48	17.03	969	33.25	19.76	32.53	14.46	415	27.3	15.5	38.0	19.3	561	39.14	13.83	34.20	12.84	810
KSA4R	35.62	15.86	31.12	17.40	977	33.74	19.56	32.03	14.67	409	26.9	15.6	38.0	19.5	558	36.78	14.23	33.72	15.28	949
KSA5R	35.34	15.89	31.57	17.21	982	33.58	20.34	31.86	14.22	408	27.6	15.7	37.7	19.0	562	37.93	14.81	33.60	13.67	878
KSA6R	34.38	15.57	31.90	18.15	1047	33.66	19.85	32.20	14.29	413	27.9	16.1	36.9	19.1	559	39.63	13.46	34.69	12.22	810
KSA7R	34.34	16.49	31.90	17.27	1025	32.94	19.81	32.94	14.32	419	26.4	16.3	38.0	19.3	569	39.15	13.66	34.88	12.32	820
KSA8R	35.20	16.02	30.92	17.86	980	33.33	20.29	31.88	14.49	414	26.4	15.5	38.4	19.7	549	38.00	13.65	34.47	13.88	850
KSA9R	35.79	15.79	30.97	17.44	975	27.97	17.72	34.27	20.05	429	26.4	15.6	38.3	19.6	556	38.63	13.92	33.18	14.27	862
KSA10R	35.35	16.14	31.14	17.37	973	30.37	22.73	32.64	14.26	484	26.3	15.6	38.1	20.0	551	38.31	14.24	32.54	14.92	885
KSA11R	31.89	17.34	35.36	15.41	1038	32.39	22.98	29.76	14.88	457	25.9	16.5	37.3	20.3	557	36.42	13.22	31.97	18.39	832
KSA12R	32.69	16.04	36.33	14.93	991	34.00	18.60	32.00	15.40	500	27.0	15.7	38.0	19.3	548	39.53	13.84	33.54	13.09	802
KSA13R	33.13	16.36	35.60	14.92	972	33.25	20.15	31.80	14.81	412	26.7	16.1	37.5	19.7	554	40.08	13.94	34.17	11.81	796
KSA15R	32.80	16.10	36.22	14.89	994	33.66	20.34	31.72	14.29	413	27.0	15.4	37.9	19.7	544	39.85	14.25	33.92	11.98	793
KSA16R	33.13	16.26	35.38	15.24	978	32.89	20.00	32.00	15.11	450	26.7	16.4	37.5	19.4	566	39.43	14.18	33.58	12.81	804
KSA17R	32.66	16.13	36.19	15.02	992	33.98	19.28	32.05	14.70	415	26.8	15.5	37.6	20.1	548	39.87	13.96	33.71	12.45	795
KSA18R	32.49	16.05	35.81	15.66	1022	33.09	20.19	31.63	15.09	411	26.9	15.8	37.7	19.6	551	40.30	14.23	33.88	11.59	794
KSA19R	32.79	15.94	36.14	15.13	985	31.80	19.80	32.00	16.40	500	26.8	15.6	37.9	19.7	557	39.53	14.00	34.20	12.27	807
KSA20	32.40	15.90	36.40	15.30	1000	33.41	19.95	31.73	14.90	416	26.1	15.4	38.3	20.1	566	38.99	13.74	33.79	13.49	808
KSA21	33.16	15.76	36.23	14.84	977	30.93	19.77	34.19	15.12	430	26.1	15.3	38.2	20.4	555	39.88	14.00	33.50	12.63	800
KSA22	32.99	16.55	35.65	14.81	979	32.49	19.22	33.18	15.10	437	26.6	15.0	38.4	20.0	515	39.88	14.00	33.75	12.38	800
KSA23	33.07	15.95	35.80	15.18	1028	32.85	20.05	31.64	15.46	414	27.1	15.4	37.9	19.6	565	39.88	14.00	33.38	12.75	800
KSA24	32.40	16.15	35.95	15.50	929	33.50	20.39	31.55	14.56	412	26.4	14.9	38.2	20.5	523	39.48	13.82	33.62	13.08	803
KSA25	32.97	15.55	36.22	15.26	1016	28.67	14.00	34.89	22.44	450	26.5	16.6	37.0	19.9	548	39.88	14.00	33.63	12.50	800

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 June 2023 doi:10.20944/preprints202306.1304.v1

KSA26	32.90	15.33	36.45	15.33	1070	34.12	19.43	31.75	14.69	422	26.6	15.8	37.9	19.7	549	40.03	13.80	33.75	12.42	797
KSA27	31.87	15.99	36.05	16.09	957	33.65	19.47	32.21	14.66	416	27.2	15.7	37.4	19.7	548	39.50	14.13	33.75	12.63	800
KSA28	32.47	16.80	35.57	15.15	970	32.44	19.11	32.00	16.44	450	26.2	16.3	36.4	21.1	583	39.63	14.00	33.75	12.63	800
KSA29	35.42	25.33	21.87	17.38	1070	33.57	19.08	31.40	15.94	414	25.7	16.8	36.0	21.5	600	39.15	14.09	33.29	13.47	802
KSA30	32.48	16.45	36.16	14.91	979	32.77	19.76	32.77	14.70	415	26.0	15.6	38.3	20.0	569	39.60	13.70	33.62	13.08	803
KSA31	33.23	16.16	35.69	14.93	978	32.53	20.24	32.53	14.70	415	25.1	16.0	39.2	19.7	589	33.37	11.21	40.97	14.45	803
KSA32	32.72	16.62	35.49	15.18	975	33.41	20.58	32.20	13.80	413	27.1	15.9	37.7	19.3	554	39.88	13.88	33.63	12.63	800
KSA33	32.93	16.21	36.05	14.80	993	33.57	20.05	31.88	14.49	414	26.5	16.2	37.5	19.8	550	38.83	14.02	34.12	13.03	806
KSA34	32.28	16.65	36.06	15.02	979	33.50	19.90	31.80	14.81	412	26.0	16.2	38.4	19.4	573	38.64	13.73	34.26	13.37	823
KSA35	32.67	16.04	36.16	15.14	1004	32.60	19.95	32.85	14.60	411	26.7	15.7	38.1	19.5	554	39.29	13.92	33.62	13.18	812
KSA36	33.54	15.90	35.28	15.28	975	31.66	19.36	33.94	15.03	439	26.5	16.0	37.9	19.6	551	39.41	13.63	34.32	12.64	807
KSA37	32.99	15.73	35.96	15.32	979	32.50	19.09	32.73	15.68	440	27.1	15.5	38.7	18.8	595	39.70	13.52	34.37	12.41	806
KSA38	32.96	15.61	36.14	15.30	974	33.49	20.48	31.33	14.70	415	27.1	15.9	37.7	19.3	584	40.08	14.03	34.13	11.76	791
KSA39	33.03	15.64	36.40	14.93	978	32.94	20.56	31.31	15.19	428	27.3	15.7	37.8	19.3	535	39.41	13.88	33.46	13.26	807
KSA40	32.39	16.47	35.21	15.93	923	33.01	19.86	32.54	14.59	418	26.4	15.7	38.1	19.9	554	39.18	13.84	33.37	13.60	809
KSA41						33.49	19.95	32.80	13.76	436	26.6	14.9	38.2	20.3	523	39.67	14.11	33.75	12.47	794
KSA42						34.49	20.14	30.79	14.58	432	26.4	15.5	37.3	20.8	576	40.05	14.30	33.58	12.06	804
KSA43						33.82	20.19	31.63	14.36	411	27.0	14.7	37.9	20.4	530	39.88	13.75	33.75	12.63	800
KSA44	31.89	16.76	35.57	15.78	925	31.65	19.50	31.19	17.66	436	26.8	15.9	37.7	19.6	560	39.30	13.81	33.21	13.68	804
KSA45R	32.86	15.84	35.68	15.62	922	32.06	19.38	33.01	15.55	418	27.2	15.7	37.2	20.0	541	39.60	13.82	33.50	13.08	803
KSA46	32.86	16.43	35.61	15.10	980	32.64	19.44	32.18	15.74	432	26.7	15.8	37.3	20.2	544	39.35	14.16	33.58	12.91	798
KSA47	33.10	15.68	35.95	15.27	982	33.17	19.61	32.93	14.29	413	26.1	16.0	38.2	19.7	563	39.63	14.13	33.50	12.75	800
KSA48	32.39	16.30	35.65	15.65	920	33.63	19.41	32.05	14.90	443	28.6	21.6	32.4	17.4	574	39.63	14.16	33.66	12.55	805
KSA49	32.65	16.05	35.79	15.51	922	32.61	20.86	31.89	14.63	417	26.9	17.7	35.5	19.9	583	39.30	13.93	33.58	13.18	804
KSA50	32.68	16.02	35.82	15.48	924	32.27	21.59	30.00	16.14	440	26.9	15.4	37.9	19.8	551	39.78	13.97	33.29	12.97	802
KSA51R	32.39	15.71	36.51	15.38	923	33.58	19.95	32.12	14.36	411	26.5	16.5	37.9	19.1	570	39.46	13.93	33.91	12.70	811
KSA52	32.68	16.07	35.40	15.85	921	32.85	20.68	31.39	15.09	411	26.7	15.7	37.1	20.5	536	38.75	14.02	33.58	13.65	813

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 June 2023 doi:10.20944/preprints202306.1304.v1

KSA59	32.36	16.13	35.71	15.80	924	33.50	19.75	31.50	15.25	400	26.3	15.6	38.5	19.6	556	38.27	13.49	33.54	14.70	823
KSA60	32.50	16.36	35.75	15.38	923	33.64	20.23	30.45	15.68	440	26.7	16.2	37.6	19.6	551	38.56	14.20	33.78	13.46	817
KSA61	32.29	15.87	35.64	16.20	926	32.70	20.14	32.46	14.69	422	25.9	16.6	37.0	20.5	595	39.23	14.20	33.50	13.08	803
KSA62	32.86	16.02	36.22	14.90	980	32.00	18.00	35.50	14.50	400	27.4	15.6	36.9	20.0	544	39.70	13.86	33.58	12.86	801
KSA63	33.09	15.93	35.96	15.02	979	32.81	19.82	32.21	15.16	426	26.4	15.9	37.7	19.9	552	39.57	13.43	34.05	12.95	834
Avg.	33.15	16.60	34.49	15.76	963.8						26.7	15.9	37.6	19.8	566	39.18	13.88	33.84	13.10	812

Table 5. Singleton variable sites and parsimony information sites of the four chloroplasts barcode.

Gene	Parameter	Sites	Position	Total	Grand Total
atpB-rbcl.	STVS	Two variants' sites	448	1	1
		Three variants' sites		0	
	PIC	Two variants' sites	338 362 408 442 443 444 447 457 469 878 913	11	137
		Three variants' sites	335 339 340 350 356 360 361 363 364 366 369 370 383 387 388 394 395 396 398 399 400 403 404 405 410 414 416 417 430 432 433 434 435 437 438 439 441 445 446 451 452 453 454 455 459 460 464 465 466 467 470 471 488 489 490 494 495 496 497 498 499 500 505 507 510 512 513 515 520 522 523 524 530 531 534 536 539 540 543 545 546 549 552 553 554 555 562 563 564 567 568 569 570 595 602 766 768 771 776 779 780 784 789 794 795 826 828 829 834 844 845 855 866 867 869 874 882 896 897 907 908 911 914 915 921 924	126	
TrnL	STVS	Two variants' sites	40 60 103 134 149 155 165 173 178 182 186 192 215 216 218 219 237 242 246 335 343 344 369 378 386 390 391 397 398 400 402 405 407 408 411 413 417 418 419 427 430 431	42	52
		Three variants' sites	124 157 217 379 406 410 414 428 429 439	10	
	PIC	Two variants' sites	37 42 49 50 58 63 64 71 72 73 74 76 77 79 84 85 86 87 88 92 93 99 100 102 105 106 108 109 112 113 114 116 117 123 126 129 130 131 132 140 141 142 146 148	155	182

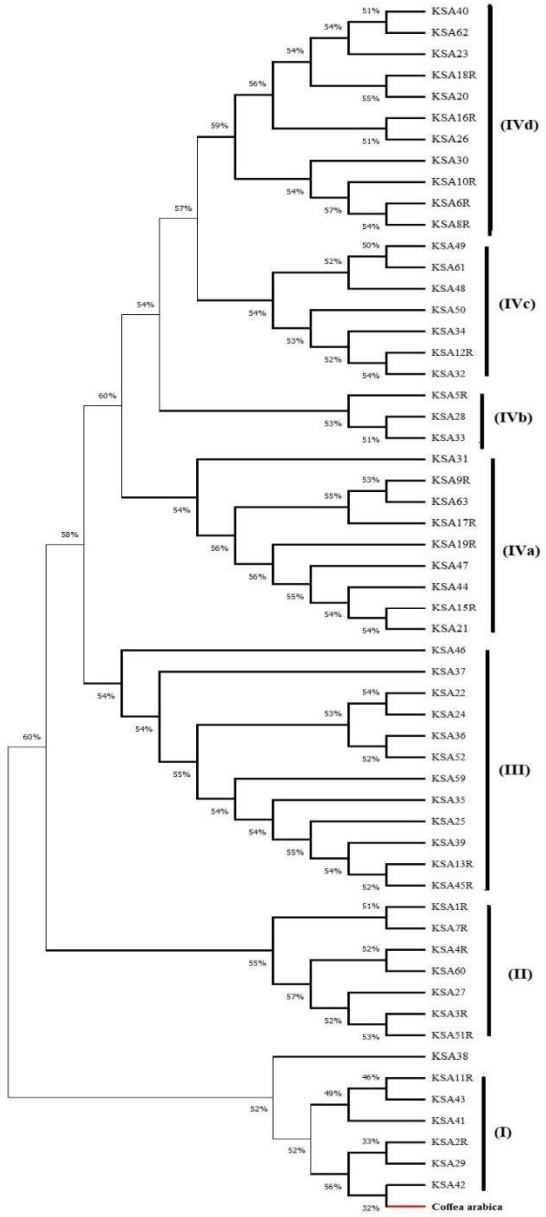
 doi:10.20944/preprints202306.1304.v1

			154 156 161 162 163 167 168 170 172 180 184 190 191 195 196 198 199 200 201 203 204 221 223 226 227 228 232 233 239 244 245 248 250 251 252 253 255 256 258 260 261 263 266 267 270 273 274 275 276 278 280 281 282 284 288 289 294 295 297 302 303 304 305 309 312 313 314 315 318 320 321 322 323 324 327 328 329 330 332 333 334 337 341 342 346 347 348 349 351 352 353 354 356 357 358 360 361 362 363 365 367 373 375 382 387 392 393 401 426 437 440		
		Three variants' sites	21 30 43 44 45 53 78 89 90 118 181 185 291 345 350 355 366 368 372 374 376 377 381 383 384 403 432	27	
TrnL-TrnF	STVS	Two variants' sites	32 33 50 51 64 91 224 256 268 271 273 281 282 284 285 286 287 290 291 293 297 300 301 307 308 310 312 314 315 318 320 321 324 325 327 328 330 332 336 337 338 340 346 347 348 349 350 351 353 354 356 357 362 363 364 365 366 367 369 370 371 373 374 375 376 385 388 389 392 393 396 403 406 407 410 411 416 419 426 427 429 430 432 443 444 445 447 448 452 455 456 457 459 460 467 470 471 472 475 476 484 487 490 494 495 497 501 504 505 506 508 510 520 521 535 543 548 549 551	119	133
		Three variants' sites	37 368 394 399 404 420 446 454 464 482 488 489 530 552	14	
	PIC	Two variants' sites	31 46 47 86 88 382 386 400 414 415 422 423 431 442 449 450 453 458 463 473 474 477 478 479 485 486 492 493 502 518 525 533 538 541 547	35	45
		Three variants' sites	23 441 499 500 522 531 532 539 545 546	10	

 doi:10.20944/preprints202306.1304.v1

TrnT-TrnL	STVS	Two variants' sites	33 35 36 37 39 40 41 42 43 45 46 49 51 52 53 60 62 63 65 69 73 76 77 78 80 81 82 84 86 87 88 91 92 96 97 98 99 101 106 107 108 109 114 115 122 124 125 126 127 129 130 131 132 136 137 138 141 142 143 150 151 152 154 155 161 162 163 166 167 169 174 175 177 178 180 184 189 190 191 192 195 196 197 202 204 211 213 217 221 222 223 224 227 229 230 233 236 237 249 253 254 256 257 258 259 260 261 262 265 267 268 270 271 273 275 276 283 284 287 288 294 299 301 303 304 305 306 310 311 312 313 321 322 326 327 328 330 332 335 339 346 347 351 352 354 356 357 358 359 366 368 371 375 377 379 381 382 385 386 391 394 398 400 403 405 411 414 415 416 417 420 422 428 432 435 437 445 446 447 449 451 456 457 464 468 471 473 475 476 477 481 482 490 491 492 493 497 498 499 500 502 504 509 513 514 515 516 519 520 527 528 529 532 533 534 535 538 542 543 544 545 546 548 551 554 565 568 576 579 580 582 595 598 599 600 601 604 605 618 626 630 631 633 638 639 640 642 643 645 646 647 649 652 654 655 656 661 664 668 670 674 676 678 679 680 682 684 685 688 691 695 696 697 698 701 703 706 709 711 714 715 716 718 719 722 723 725 727 749 752 756 760 762 766 768 771 774 777 778 782 791 800 814	303	338
		Three variants' sites	61 72 550 570 581 606 607 613 622 625 627 628 653 659 675 707 710 713 721 724 728 732 734 741 754 763 769 770 775 776 780 792 795 796 797	35	
	PIC	Two variants' sites	28 240 393 421 444 452 470 530 541 547 549 573 574 586 590 592 594 608 614 619 620 621 624 636 650 660 665 666 671 677 683 705 708 712 731 748 751 753 755 798 799 802 805 812 819 820 821	47	74
		Four variants' sites	22 24 27 48 635 641 648 672 673 689 720 726 739 740 764 765 767 779 783 794 801 806 807 809 810 816 822	27	

 doi:10.20944/preprints202306.1304.v1



 doi:10.20944/preprints202306.1304.v1

Figure 1. Evolutionary analysis by Maximum Likelihood method using four barcodes with 1000 bootstraps constructed in MEGA 10.0 using the concatenated sequence of atpB-rbcl, TrnLc, TrnT-TrnL and TrnL-TrnF.

The phylogenetic analysis grouped the coffee accessions into four different groups. The first group contained six accessions (KSA42, KSA29, KSA2R, KSA41, KSA43, KSA11R, and KSA38) that were mostly from the Rayda district of Assir region. The second group contained seven accessions (KSA51R, KSA3R, KSA27, KSA60, KSA4R, KSA7R, and KSA1R), all from Jazan Region except KSA60 was from Assir. The third group contained 12 accessions (KSA45R, KSA13R, KSA39, KSA25, KSA35, KSA59, KSA52, KSA36, KSA24, KSA22, KSA37 and KSA46), all collected from Jazan Region except KSA59 from the north of Assir Region. The fourth and largest group contained 43 accessions that can be further subdivided into four subgroups. The first subgroup (IVa) was a diverse one and contained 12 accessions originating from the three regions. Subgroup IVb contained three accessions (KSA33, KSA28 and KSA5R) all from Jazan Region. Subgroup IVc contained seven accessions, six from Jazan and one from Al-Baha. Subgroup IVd contained 11 accessions, eight from Jazan region, two from Assir and one from Al-Baha.

3. Discussion:

The genetic diversity present in crop wild or primitive relatives (CWR) plays a crucial role in the effectiveness of crop improvement programs. These wild or unknown genotypes exit in diverse habitats, many of which are currently facing significant threats due to habitat degradation and climate change [8,14]. It is estimated that approximately 60% of wild coffee species are at risk of extinction worldwide; similarly, underutilized old varieties are disappearing from the orchards. This emphasizes the urgent need to conserve these species through both in situ and ex situ measures, which can help preserve their genetic diversity for future utilization.

While morphological descriptors are commonly used to characterize different coffee species, molecular markers are considered more efficient in distinguishing closely related species and cultivars [23], and more precise and reliable than morphological and biochemical markers [24]. Furthermore, several studies have demonstrated that specific regions of the chloroplast genome can serve as DNA barcodes for a wide variety of plant species [25,26]. Selection of suitable plastid genomes offer sufficient genetic information for distinguishing between genotypes. Additionally, when choosing suitable DNA barcoding loci, the variable regions should be given a primary consideration [27]. Therefore, the objective of this study was to evaluate the evolutionary and phylogenetic relationships among fifty-six Arabica coffee species by utilizing four DNA barcoding markers (atpB-rbcl, **TrnL-TrnF**, TrnL, and TrnT-TrnL). This research aimed to investigate the potential of four DNA barcode loci (specifically, atpB-rbcL, TrnL, TrnL-TrnF, and trnL-trnL from the chloroplast region) for the identification and provision of phylogenetic information on local Arabica coffee species. All four regions were successfully amplified using universal primers, yielding clear and reliable results. However, earlier studies have indicated that there were cases of partial amplification from the respective barcode loci's universal primers [28,29].

Similarly, other studies [30,31] have indicated that additional barcode primers, including matK, rbcL, and trnL-trnF, have demonstrated successful amplification within coffee species, which aligns with the findings of the present study. Despite the abundance of available data on DNA barcoding of angiosperms, there is currently limited information regarding specific barcodes that can guarantee an accurate species identification in all cases [32]. Often, a barcode that performs effectively for one group of plants may prove inadequate for another group, especially in the case of recently diverged species [33]. The current study successfully identified all fifty-six accessions as *Coffea arabica*, showcasing the effectiveness of the universal DNA barcode primers. Likewise, multiple studies have extensively documented the reliability of matK and rbcL, either individually or in combination, as DNA barcodes that can be used with confidence across various plant species [34]. Additional reports have recommended the utilization of rbcL as a valuable DNA barcode locus, primarily due to its relatively compact length of 500 bp, robust universal primer, high success rate in PCR amplification,

 doi:10.20944/preprints202306.1304.v1

and excellent sequencing quality [35]. However, other DNA barcodes, such as TrnL-TrnF and the TrnL spacer, have also been suggested as reliable alternative barcodes for identification of species [36]. The extent of sequence variation among the species or terminals under analysis is a crucial factor in determining the effectiveness of any barcoding locus [34]. Other reports have recommended the utilization of rbcL as a valuable DNA barcode locus due to its relatively compact length of 500 bp, robust universal primer, high success rate in PCR amplification, and excellent sequencing quality [25], along with other DNA barcodes such as TrnL-TrnF and the TrnL spacer [36]. The degree of sequence variation present among the species or terminals being analyzed is a determining factor in the efficacy of any barcoding loci [34].

The number of singleton variable sites was found to be higher in TrnLc, TrnL, and TrnL compared to atpB-rbcl. Similarly, TrnL and atpB-rbcl had more parsimony information sites than the rbcL barcode spacer region. These findings are consistent with a previous study conducted by [37], which reported that TrnL-trnF and matK barcodes exhibited greater variability than rbcl in Indian coffee arabica genotypes. The present study also found similar results for PIC among the four barcodes analyzed. Similarly, previous research has indicated that TrnL-TrnF and matK loci exhibit greater sequence polymorphism than rbcL, as suggested by [38,39], and [40]. The current study's results support these findings. Hence, the present study found that all four barcode sequences, which were evaluated as candidate barcode markers, met the DNA barcoding criteria outlined by [33]. Specifically, these markers exhibited sufficient sequence variability to enable effective discrimination among Saudi coffee genotypes. Similarly, [41] previously suggested that indels involving multiple residues may serve as useful diagnostic markers for species discrimination and phylogenetic analyses [42].

The phylogenetic analysis grouped the Saudi *Coffea arabica* genotypes into four groups with a clear influence of geographic origin suggesting the genotypes of each region share one or more common ancestor (Figure 2). For instance, accessions KSA11R, KSA41, KSA42 and KSA43 from the isolated Rayda district of Assir region were grouped in clusters I and II. The accessions representing very old trees (KSA36, KSA44, KSA46, KSA47) segregated in the middle of the phylogenic tree in groups III and IVa. Similar results were reported by [37] where the grouping using single and multilocus barcode primers was strongly influenced by the geographic origin of the genotypes. A molecular analysis of *coffee arabica* genotypes from Saudi Arabia using SRAP markers grouped them into five distinct groups based mostly on their origin [23]. The accessions collected from Jazan region primarily clustered in groups II and IV, whereas those from Al-Baha and Assir regions constituted a different group. Similar surveys of genetic diversity among coffee populations in northern Yemen [6] and southern Yemen [14] found that each district (valley) have its own cultivars. Another study using genotyping by sequencing (GBS) showed that genetic closeness correlated with geographic proximity [43].The current study provides further evidence to support this finding. It was also suggested chloroplast sequences provide high level effective and accurate information for plant evolution and high level preservation [44]. For future studies on this economically significant crop, we recommend using resequencing and genome-wide association studies (GWAS) to discover additional polymorphic markers associated with important agro-morphological traits. These markers would be beneficial for a range of investigations in *Coffea* crop. Ultimately, the polymorphic markers established and confirmed in this research hold potential as a valuable genomic asset for molecular breeding, identification, and biogeography studies of Arabica coffee.

4. Materials and Methods

4.1. Plant Material

A survey was carried out at several sites in the Sarawat mountain range, running parallel to the Red Sea from the southeast to the northwest through the three administrative regions of Jazan, Assir, and Al-Baha. The survey covered a strip of terraced mountains located between latitudes 17˚N and 20˚N, the most northern location where coffee is commercially grown in the world. The coffee gardens included in the survey were found at altitudes ranging from 900 to 2000 m a.s.l. In total, we

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 June 2023 doi:10.20944/preprints202306.1304.v1

collected young leaves from 56 accessions, from Jebel Fayfa (Fayfa district), Eddayer, Maadi (Haroub district), Jebel Al-Gahr (Al-Rayth district), Rayda valley (Assouda district in Assir region), Mahayel Assir district, Al-Majarda district and Jebel Shada (Al-Mekhwah district of Al-Baha region) (Table 1). We tagged and sampled 1-3 trees representing each tree population. Each accession was given a code starting with the acronym "KSA" (e.g., KSA-1), but, for the sake of simplicity, we dropped the acronym in the figures. The letter "R" was added to the code of accessions 1-19, 45, and 51 to indicate that they were sourced from a small, local coffee germplasm collection established in the Fayfa district.

Table 1. Altitude and latitude of the sites where the coffee accessions were sourced. The sites are located between longitudes 42°22' and 43°07'E.

#	Accession no.	Region	District	Altitude (m a.s.l.)	Latitude
1	KSA1R	Jazan	Khacher/Al-Zoughli	1254	17°18'03"N
2	KSA2R	Jazan	Khacher/Al-Guatil	1484	17°19'01"N
3	KSA3R	Jazan	Khacher/Al-Guatil	1484	17°19'01"N
4	KSA4R	Jazan	Jebel Fayfa	1541	17°15'21"N
5	KSA5R	Jazan	Wadi Dafa	1254	17°25'41"N
6	KSA6R	Jazan	Tallan	1672	17°23'12"N
7	KSA7R	Jazan	Tallan	1672	17°23'12"N
8	KSA8R	Jazan	Tallan	1546	17°23'01"N
9	KSA9R	Jazan	Tallan	1672	17°23'12"N
10	KSA10R	Jazan	Khacher/Al-Zoughli	1254	17°18'03"N
11	KSA11R	Assir	Rayda	1594	18º11'37"N
12	KSA12R	Jazan	Maaddi	1287	17º29'29"N
13	KSA13R	Jazan	Maaddi	1344	17º29'29"N
14	KSA15R	Al-Baha	Shada Al-ala	1548	19º50'54"N9
15	KSA16R	Assir	Rayda	1594	18º11'37"N
16	KSA17R	Assir	Rayda	1519	18º11'37"N
17	KSA18R	Assir	Al-Majarda	1329	19º09'35"N
18	KSA19R	Assir	Al-Majarda	1300	19º09'35"N
19	KSA20	Jazan	Jebel Fayfa	1260	17º15'20"N
20	KSA21	Jazan	Jebel Fayfa	1260	17º15'20"N
21	KSA22	Jazan	Jebel Fayfa	1260	17º15'20"N
22	KSA23	Jazan	Jebel Fayfa	1260	17º15'20"N
23	KSA24	Jazan	Jebel Fayfa	1260	17º15'20"N
24	KSA25	Jazan	Jebel Fayfa	1260	17º15'20"N
25	KSA26	Jazan	Jebel Fayfa	1550	17º15'24"N
26	KSA27	Jazan	Jebel Fayfa	1550	17º15'24"N
27	KSA28	Jazan	Jebel Fayfa	1550	17º15'24"N
28	KSA29	Jazan	Al-Gahr	1846	17º38'08"N
29	KSA30	Jazan	Al-Gahr	1846	17º38'08"N
30	KSA31	Jazan	Al-Gahr	1846	17º38'08"N
31	KSA32	Jazan	Al-Gahr	1846	17º38'08"N

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 June 2023 doi:10.20944/preprints202306.1304.v1

32	KSA33	Jazan	Al-Gahr	1846	17º38'08"N
33	KSA34	Jazan	Jebel Fayfa	1660	17º15'55"N
34	KSA35	Jazan	Jebel Fayfa	1660	17º15'55"N
35	KSA36	Jazan	Jebel Fayfa	1450	17º15'59"N
36	KSA37	Jazan	Eddayer	1100	17º22'10"N
37	KSA38	Jazan	Eddayer	1228	17º22'10"N
38	KSA39	Jazan	Eddayer	1228	17º22'10"N
39	KSA40	Jazan	Haroub	1100	17º29'29"N
40	KSA41	Assir	Rayda	1450	18º11'37"N
41	KSA42	Assir	Rayda	1450	18º11'37"N
42	KSA43	Assir	Rayda	1400	18º11'37"N
43	KSA44	Jazan	Jebel Fayfa	1524	17º15'48"N
44	KSA45R	Jazan	Jebel Fayfa	1524	17º15'48"N
45	KSA46	Jazan	Al-Gahr	1750	17º39'01"N
46	KSA47	Jazan	Al-Gahr	1750	17º39'01"N
47	KSA48	Jazan	Jebel Fayfa	1260	17º15'20"N
48	KSA49	Jazan	Jebel Fayfa	1260	17º15'20"N
49	KSA50	Jazan	Jebel Fayfa	1260	17º15'20"N
50	KSA51R	Jazan	Jebel Fayfa	1524	17º17'13"N
51	KSA52	Jazan	Jebel Fayfa	1550	17º15'24"N
52	KSA59	Assir	Al-Majarda	1329	19º09'35"N
53	KSA60	Assir	Al-Majarda	1300	19º09'35"N
54	KSA61	Al-Baha	Shada Al-ala	1548	19º50'54"N
55	KSA62	Al-Baha	Shada Al-ala	1548	19º50'54"N
56	KSA63	Al-Baha	Shada Al-ala	1548	19º50'54"N

4.2. DNA extraction

Plant material, consisting of young leaves from various *C. arabica* accessions, was collected from representative trees in each population and transported to the lab in a cooler. The leaves were sanitized by immersing them in a 5% sodium hypochlorite solution for 1-2 minutes and then rinsing them with sterile distilled water. The material was then ground in liquid nitrogen and stored in an -80ºC freezer. DNA was extracted from 100 mg of mixed powder using an innuPREP Plant DNA Kit (Analytik Jena), following the manufacturer's protocol. DNA quality and concentration were determined using a Nanodrop ND-1000 spectrophotometer (Saveen Werner, Sweden).

4.3. Chloroplastic DNA amplification and sequencing

Four chloroplast DNA regions were considered (Table 2). PCR was performed in a 25 µl volume containing 2 µl of template DNA, 10 µl of 1X innuMix Standard PCR, and 1 µM of each primer (Table 2). The Gene Amp PCR System 9700 was used with the following program: initial denaturation at 94°C for 5 min, 35 cycles of denaturation at 94°C for 1 min, annealing at 49-52°C for 60-75s, and elongation at 72°C for 60-75s, followed by a final polymerization at 72°C for 10 min (Table 2). To check the effectiveness of PCR, positive control using sterile water was included in all amplifications. The PCR products were checked by electrophoresis on 1% agarose gel in TAE buffer, and DNA was visualized under UV light after staining with ethidium bromide.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 June 2023 doi:10.20944/preprints202306.1304.v1

Table 2. General information about the PCR primers used in this study.

Sr#	**Sequence 5′-3′**	**Target**	**PCR condition**	**Source**
1	CATTACAAATGCGATGCTCT	**trnT-trnL**	Hybridation : 50°C/1 min	[45]
	TCTACCGATTTCGCCATATC		Elongation : 72°C/1min	
2	CGAAATCGGTAGACGCTACG	**TrnL**	Hybridation : 49°C/1.15 min	[45]
	GGGGATAGAGGGACTTGAAC		Elongation : 72°C/1.15 min	
3	GGTTCAAGTCCCTCTATCCC	**TrnL-trnF**	Hybridation :52°C/1 min	[45]
	ATTTGAACTGGTGACACGAG		Elongation : 72°C/1min	
4	GAAGTAGTAGGATTGATTCTC	**atpB-rbcL**	Hybridation : 50°C/1 min	[46]
	TACAGTTGTCCATGTACCAG		Elongation : 72°C/1min	

The amplified products were purified using the GFX PCR kit (GE Healthcare). Sequencing reactions were carried out by Congenic using Sanger technology, separately for each strand to obtain independent forward and reverse sequences. The forward and reverse fragments were aligned, and additional reactions were conducted in case of any discrepancies.

4.4. Sequence Analysis:

The scanner software-2 was utilized to determine the quality of the sequences. The four barcode samples of each coffee arabica genotype were manually curated and aligned using the contig assembly program in Bio Edit 7.0 software to ensure high-quality sequences. Nucleotide sequences obtained from the 57 accessions were initially aligned using CLUSTAL W [47] and analyzed with MEGA program version X. The number of individuals, number of nucleotide sites, variable polymorphic sites, number of segregating sites, number of haplotypes, nucleotide diversity, and average number of nucleotide differences of each barcode marker and consensus sequence were measured using DNAsp (v6) [48]. The quantification of insertion events in the sequence was determined by the number of variable sites where the addition of one or more nucleotides signals polymorphism. Likewise, the number of deletions was determined by the variable sites where polymorphism arises due to the removal of one or more nucleotides. The identification of the number of transitions in the sequences was based on the number of variable sites where polymorphism occurred due to the exchange between two purines (A and G) or two pyrimidines (C and T). On the other hand, the number of transversions was determined by the variable sites where polymorphism resulted from the replacement of a purine with a pyrimidine (Table 4). To determine the number of mutation events that have occurred in a sequence, the sum of variable sites and the number of distinct mutations observed at the same nucleotide site across different samples are combined. This quantification considers both different types of polymorphisms and multiple occurrences of mutations within the sequence. Various parameters were estimated for each sequence region to differentiate them, based on the number of monomorphic or polymorphic sites, the number of parsimony informative sites (PIC), nucleotide diversity (π), haplotype diversity (Hd), and the total number of mutations [49], singleton variable site (STVC) [50] (Table 5).

4.5. Evolutionary analysis by Maximum Likelihood method:

The evolutionary history was inferred by using the Maximum Likelihood method and Kimura 2-parameter model [51]. The tree with the highest log likelihood (-22360.57) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The tree is drawn to

 doi:10.20944/preprints202306.1304.v1

scale, with branch lengths measured in the number of substitutions per site. The proportion of sites where at least 1 unambiguous base is present in at least 1 sequence for each descendent clade is shown next to each internal node in the tree. This analysis involved 57 nucleotide sequences. There was a total of 5381 positions in the final dataset. Evolutionary analyses were conducted in MEGA X [52].

5. **Conclusion:**

To summarize, this study utilized a DNA barcoding approach to investigate the molecular relationships among fifty-six Arabica coffee samples collected from the southern region of Saudi Arabia. The three-barcode region, namely TrnT-TrnL, **TrnL-TrnF**, and TrnL, exhibited higher sequence variability compared to the atpB-rbcl barcode region and effectively differentiated the local coffee genotypes by the presence of unique variable sites (singletons and parsimony). Moreover, the combination of DNA sequences from these barcode loci, analyzed through phylogenetic analysis using the maximum likelihood method, grouped similar coffee genotypes, providing improved resolution, and understanding of the studied genotypes. These findings will contribute to future research programs in the identification and conservation of Arabica coffee using DNA barcoding markers.

Acknowledgments: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number 1.092.

References

1. Mussatto, S.I.; Machado, E.M.; Martins, S.; Teixeira, J.A. Production, composition, and application of coffee and its industrial residues. *Food and Bioprocess Technology* **2011**, *4*, 661-672.
2. Mishra, M.K. Genetic resources and breeding of coffee (*Coffea* spp.). *Advances in Plant Breeding Strategies: Nut and Beverage Crops*, **2019**, *4*, 475-515.
3. Nguyen, G.N.; Norton, S.L. Genebank phenomics: A strategic approach to enhance value and utilization of crop germplasm. *Plants* **2020**, *9*, 817.
4. Migicovsky, Z.; Warschefsky, E.; Klein, L.L.; Miller, A.J. Using living germplasm collections to characterize, improve, and conserve woody perennials. *Crop Science* **2019**, *59*, 2365-2380.
5. Word Coffee Research. High-yield-ing, dwarf/compact hybrid tolerant to coffee leaf rust and resistant to coffee berry disease (CBD) released in Kenya. *https://varieties.worldcoffeeresearch.org/varieties/ruiru-11*, **2023**, (accessed on 6 June 2023).
6. Montagnon, C.; Mahyoub, A.; Solano, W.; Sheibani, F. Unveiling a unique genetic diversity of cultivated Coffea arabica L. in its main domestication center: Yemen. *Genetic Resources and Crop Evolution* **2021**, *68*, 2411-2422.
7. Herrera, J.C.; Lambot, C. The coffee tree—Genetic diversity and origin. In *The craft and science of coffee*; Elsevier: **2017**, pp. 1-16.
8. Davis, A.P.; Chadburn, H.; Moat, J.; O'Sullivan, R.; Hargreaves, S.; Nic Lughadha, E. High extinction risk for wild coffee species and implications for coffee sector sustainability. *Science advances* **2019**, 5, eaav3473.
9. Oliveira, L.S.; Franca, A.S. An overview of the potential uses for coffee husks. *Coffee in health and disease prevention* **2015**, 283-291.
10. Choudhary, A.; Gupta, N.; Hameed, F.; Choton, S. An overview of food adulteration: Concept, sources, impact, challenges and detection. *International Journal of Chemical Studies* **2020**, *8*, 2564-2573.
11. Al-Asmari, K.M.; Zeid, I.M.A.; Al-Attar, A.M. Coffee Arabica in Saudi Arabia: An Overview. International Journal of Pharmaceutical and Phytopharmacological Research (eIJPPR) **2020**, 10, 71-78.
12. Tounekti, T.; Mahdhi, M.; Al-Turki, T.A.; Khemira, H. Genetic diversity analysis of coffee (*Coffee arabica* L.) Germplasm accessions growing in the Southwestern Saudi Arabia using quantitative traits. *Natural Resources* **2017**, *8*, 321-336.
13. Montagnon, C.; Sheibani, F.; Benti, T.; Daniel, D.; Bote, A.D. Deciphering Early Movements and Domestication of Coffea arabica through a Comprehensive Genetic Diversity Study Covering Ethiopia and Yemen. *Agronomy* **2022**, *12*, 3203.

14. Eskes, A.B. Identification, description and collection of coffee types in P. D. R. Yemen. IPGRI, Rome, **1989**. pp. 70.
15. De Pauw, E. An agroecological exploration of the Arabian Peninsula. ICARDA, Aleppo, Syria, **2022**, pp. 77.
16. Tounekti T., M. Mahdhi, T. A. Al-Turki, H. Khemira. Water relations and photo-protection mechanisms during drought stress in four coffee (*Coffea arabica*) cultivars from southwestern Saudi Arabia. *South African Journal of Botany*, **2018**, *117*:17-25.
17. Lashermes, P.; Trouslot, P.; Anthony, F.; Combes, M.-C.; Charrier, A. Genetic diversity for RAPD markers between cultivated and wild accessions of *Coffea arabica*. *Euphytica* **1996**, *87*, 59-64.
18. Cros, J.; Combes, M.-C.; Trouslot, P.; Anthony, F.; Hamon, S.; Charrier, A.; Lashermes, P. Phylogenetic analysis of chloroplast DNA variation in *Coffea* L. *Molecular phylogenetics and evolution* **1998**, *9*, 109-117.
19. Tesfaye, K.; Govers, K.; Bekele, E.; Borsch, T. ISSR fingerprinting of Coffea arabica throughout Ethiopia reveals high variability in wild populations and distinguishes them from landraces. *Plant systematics and evolution* **2014**, *300*, 881-897.
20. Tesfaye, K.; Borsch, T.; Govers, K.; Bekele, E. Characterization of Coffea chloroplast microsatellites and evidence for the recent divergence of C. arabica and C. eugenioides chloroplast genomes. *Genome* **2007**, *50*, 1112-1129.
21. Lashermes, P.; Combes, M.-C.; Trouslot, P.; Charrier, A. Phylogenetic relationships of coffee-tree species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA. *Theoretical and Applied Genetics* **1997**, *94*, 947-955.
22. Maurin, O.; Davis, A.P.; Chester, M.; Mvungi, E.F.; Jaufeerally-Fakim, Y.; Fay, M.F. Towards a phylogeny for Coffea (Rubiaceae): identifying well-supported lineages based on nuclear and plastid DNA sequences. *Annals of Botany* **2007**, *100*, 1565-1583.
23. Al-Ghamedi, K.; Alaraidh, I.; Afzal, M.; Mahdhi, M.; Al-Faifi, Z.; Oteef, M.D.; Tounekti, T.; Alghamdi, S.S.; Khemira, H. Assessment of genetic diversity of local coffee populations in Southwestern Saudi Arabia using SRAP markers. *Agronomy* **2023**, *13*, 302.
24. Jingade, P.; Huded, A.K.; Kosaraju, B.; Mishra, M.K. Diversity genotyping of Indian coffee (*Coffea arabica* L.) germplasm accessions by using SRAP markers. *Journal of Crop Improvement* **2019**, *33*, 327-345.
25. Mashaly, A.; Alajmi, R.; Mustafa, A.E.-Z.; Rady, A.; Alkhedir, H. Species abundance and identification of forensically important flies of Saudi Arabia by DNA barcoding. *Journal of medical entomology* **2017**, *54*, 837-843..
26. Al-Dakhil, M.; Alghamdi, S.; Migdadi, H.; Afzal, M.; Ali, A.A. Morphological characterization and DNA barcoding of Duckweed species in Saudi Arabia. *Plants* **2021**, *10*, 2438.
27. Mekbib, Y.; Saina, J.K.; Tesfaye, K.; Eshetu, G.; Hu, G.; Chen, J.; Li, Z. Chloroplast genome sequence variations and development of polymorphic markers in Coffea arabica. *Plant Molecular Biology Reporter* **2020**, *38*, 491-502.
28. Cuénoud, P.; Savolainen, V.; Chatrou, L.W.; Powell, M.; Grayer, R.J.; Chase, M.W. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences. *American Journal of Botany* **2002**, *89*, 132-144.
29. Yan, L.J.; Liu, J.; Möller, M.; Zhang, L.; Zhang, X.M.; Li, D.Z.; Gao, L.M. DNA barcoding of Rhododendron (Ericaceae), the largest Chinese plant genus in biodiversity hotspots of the Himalaya–Hengduan Mountains. *Molecular ecology resources* **2015**, *15*, 932-944.
30. Wu, F.; Li, M.; Liao, B.; Shi, X.; Xu, Y. DNA barcoding analysis and phylogenetic relation of mangroves in Guangdong Province, China. *Forests* **2019**, *10*, 56.
31. Vickers, N.J. Animal communication: when i'm calling you, will you answer too? *Current biology* **2017**, *27*, R713-R715.
32. Weigand, H.; Beermann, A.J.; Čiampor, F.; Costa, F.O.; Csabai, Z.; Duarte, S.; Geiger, M.F.; Grabowski, M.; Rimet, F.; Rulik, B. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of the Total Environment* **2019**, *678*, 499-524.
33. Li, X.; Yang, Y.; Henry, R.J.; Rossetto, M.; Wang, Y.; Chen, S. Plant DNA barcoding: from gene to genome. *Biological Reviews* **2015**, *90*, 157-166.
34. Carneiro de Melo Moura, C.; Brambach, F.; Jair Hernandez Bado, K.; Krutovsky, K.V.; Kreft, H.; Tjitrosoedirdjo, S.S.; Siregar, I.Z.; Gailing, O. Integrating DNA barcoding and traditional taxonomy for the identification of dipterocarps in remnant lowland forests of Sumatra. *Plants* **2019**, *8*, 461.

35. Hong, Z.; He, W.; Liu, X.; Tembrock, L.R.; Wu, Z.; Xu, D.; Liao, X. Comparative Analyses of 35 Complete Chloroplast Genomes from the Genus *Dalbergia* (Fabaceae) and the Identification of DNA Barcodes for Tracking Illegal Logging and Counterfeit Rosewood. *Forests* **2022**, *13*, 626.
36. Kang, Y. Molecular identification of *Aquilaria* species with distribution records in China using DNA barcode technology. *Mitochondrial DNA Part B* **2021**, *6*, 1525-1535.
37. Mishra, M.; Jingade, P.; Huded, A.K.C. DNA barcoding analysis and phylogenetic relationships of Indian wild coffee species. *Turkish Journal of Botany* **2022**, *46*, 109-122.
38. Hao, D.C.; Huang, B.L.; Chen, S.L.; Mu, J. Evolution of the chloroplast trnL-trnF region in the gymnosperm lineages Taxaceae and Cephalotaxaceae. *Biochemical Genetics* **2009**, *47*, 351-369.
39. Skuza, L.; SzuĆKo, I.; Filip, E.; Adamczyk, A. DNA barcoding in selected species and subspecies of Rye (Secale) using three chloroplast loci (matK, rbcL, trnH-psbA). *Notulae Botanicae Horti Agrobotanici Cluj-Napoca* **2019**, *47*, 54-62.
40. Meena, R.K.; Negi, N.; Uniyal, N.; Shamoon, A.; Bhandari, M.S.; Pandey, S.; Negi, R.K.; Sharma, R.; Ginwal, H.S. Chloroplast-based DNA barcode analysis indicates high discriminatory potential of matK locus in Himalayan temperate bamboos. *3 Biotech* **2020**, *10*, 1-13.
41. Mahadani, P.; Ghosh, S.K. Utility of indels for species-level identification of a biologically complex plant group: a study with intergenic spacer in Citrus. *Molecular biology reports* **2014**, *41*, 7217-7222.
42. Wu, L.; Wu, M.; Cui, N.; Xiang, L.; Li, Y.; Li, X.; Chen, S. Plant super-barcode: a case study on genome-based identification for closely related species of Fritillaria. *Chinese medicine* **2021**, *16*, 1-11.
43. Hamon, P.; Grover, C.E.; Davis, A.P.; Rakotomalala, J.-J.; Raharimalala, N.E.; Albert, V.A.; Sreenath, H.L.; Stoffelen, P.; Mitchell, S.E.; Couturon, E. Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (*Coffea*) and insights into the evolution of caffeine content in its species: GBS coffee phylogeny and the evolution of caffeine content. *Molecular phylogenetics and evolution* **2017**, *109*, 351-361.
44. Guyeux, C.; Charr, J.-C.; Tran, H.T.; Furtado, A.; Henry, R.J.; Crouzillat, D.; Guyot, R.; Hamon, P. Evaluation of chloroplast genome annotation tools and application to analysis of the evolution of coffee species. *PLoS One* **2019**, *14*, e0216347.
45. Taberlet, P.; Gielly, L.; Pautou, G.; Bouvet, J. Universal primers for amplification of three non-coding regions of. *Plant molecular biology* **1991**, *17*, 1105-1109.
46. Savolainen, V.; Corbaz, R.; Moncousin, C.; Spichiger, R.; Manen, J.-F. Chloroplast DNA variation and parentage analysis in 55 apples. *Theoretical and Applied Genetics* **1995**, *90*, 1138-1141.
47. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **1994**, *22*, 4673-4680.
48. Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.; Sánchez-Gracia, A. DnaSP version 6 for 32-bit and 64-bit environments (ZIP file of 11.7 Mb).
49. Hosein, F.N.; Austin, N.; Maharaj, S.; Johnson, W.; Rostant, L.; Ramdass, A.C.; Rampersad, S.N. Utility of DNA barcoding to identify rare endemic vascular plant species in Trinidad. *Ecology and Evolution* **2017**, *7*, 7311-7333.
50. Pettengill, J.B.; Neel, M.C. An evaluation of candidate plant DNA barcodes and assignment methods in diagnosing 29 species in the genus *Agalinis* (Orobanchaceae). *American journal of botany* **2010**, *97*, 1391-1406.
51. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* **1980**, *16*, 111-120.
52. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution* **2018**, *35*, 1547.