UPPSALA
UNIVERSITET

Microsatellite variation in populations from Sudan

Hiba Moh. Ali Babiker

Degree project in biology, Master of Science (2 years), 2010
Examensarbete i biologi 45 hp till masterexamen, 2010
Biology Education Centre and Department of Evolutionary Biology, Uppsala University
Supervisor: Mattias Jakobsson

Abstract

Studies of genetic variation in African populations have gained attention since evidence from both archeology and genetics confirmed that the genetic diversity in African populations is the highest compared to populations in other continents. Sudan is located at the northeastern part of Africa covering major part of the Nile valley which is hypothesized to be a way out of Africa for human migrations. Sudan with its linguistically and ethnically diverse populations has been involved in Y-chromosome STRs and mtDNA studies. In this study I sampled 498 individuals representing different ethnic, linguistic, and geographic regions of Sudan and typed these for the 15 microsatellite loci included in the AmpFℓSTR Identifiler PCR Amplification Kit to determine allele frequencies, genetic variation, allelic diversity, the informativeness of the loci and population structure. From the observed allele frequencies, I calculated the polymorphism information content which was found the highest for the locus D18S51 and the lowest for the locus TPOX. None of the loci deviated from Hardy-Weinberg equilibrium expectations after the Bonferroni correction. The combined power of exclusion was 0.9999981 and the combined match probability was 1 in 7.4 x 10^{17}. All the studied parameters lead us to encourage the use of these loci in human identification for Sudanese population. The allelic richness among Sudanese was the highest in Zagawa and Nuba populations. Besides that the private allelic richness and the uniquely shared alleles were the highest in Zagawa, Nilotics and Nuba populations. I used published data for Ugandans, Egyptians and Somali to study the allelic diversity between these populations and Sudanese populations in this study. The informativeness level for inference of ancestry was low for the 15 loci compared to many other microsatellites. However using the Structure program, individuals from Sudanese populations were assigned to two clusters. The genetic variation among the Sudanese populations revealed that the most genetic variation is within populations and slightly more variance was explained by geography grouping than linguistic grouping. The Principal Component Analysis showed high similarity between the Copts and Nubians with Egyptians and both the PCA and allelic diversity analysis showed the high affinity between Nuba, Zagawa, Nilotics with Ugandans. These findings are similar to results from other studies based on Y-chromosome and mtDNA as well as historical and linguistic evidence.

Contents

1. Introduction

Archeological evidence and genetic evidence based on mtDNA, Y-chromosome and autosomal DNA markers suggest that humans originated in Africa 100,000-200,000 years ago (Grun and Stringer, 1991), and that the favorable environment offered by the Nile Valley explain the permanent settlements in this region which is dated to ~ 18,000 years ago (Phillipson, 1993). Studies on genome-wide data confirmed the previous findings from mtDNA, X-chromosome, and Y-chromosome analyses. For example the recent survey of autosomal microsatellite, insertion/deletion (INDEL) and single nucleotide polymorphism (SNP) markers all agreed that diversity is higher in Africans compared to the rest of the world and that more private alleles are present in Africa (Rosenberg *et al*., 2002; Campbell and Tishkoff, 2008; Jakobsson *et al*., 2008; Tishkoff *et al*., 2009; Campbell and Tishkoff, 2010).

Studies using mtDNA and nuclear DNA markers consistently indicate that Africa is the most genetically diverse region of the world (Tishkoff and Williams, 2002). Studying the levels and patterns of genetic diversity among the ethnically diverse African populations is a key part to answering questions about human evolutionary history, the genetic basis of phenotypic variation and genetic associations with diseases.

Sudan is located to the northeastern part of Africa, with a total 133 living languages listed by Lewis (2009) belonging to three of the major African linguistic groups proposed by Greenberg (1963), namely the Niger-Congo, Nilo-Saharan and Afro-Asiatic. Population samples from Sudan raised attention in studies concerned with migration and genetic diversity, because of its unique location in the Nile valley, which is hypothesized to be part of the traditionally favored model of the migratory route out of Africa (Mellars, 2006).

Although Sudanese populations are important for genetic diversity studies they remained underrepresented in genetic studies because of the need for key research and medical centers. In addition to that, genetic studies in Sudan, which started late compared to other neighboring African countries, have been biased toward mtDNA and Y-chromosome STRs (Krings *et al*.,

1999; Salas *et al*., 2002; Hassan *et al*., 2008). To the best of our knowledge no population data has been published for Sudanese populations regarding the polymorphic autosomal markers included in the AmpFℓSTR Identifiler PCR Amplification Kit.

Scattered throughout the human genome there are many repeated DNA sequences and these repeat sequences are typically located between genes, varying in size from individual to individual without impacting the genetic health of the individual. These repeated DNA sequences are typically designated by the length of the core repeat unit and overall length of the repeat region. Long repeat units may contain several hundred to several thousand bases in the core repeat (Butler, 2009).

Microsatellite DNA, simple sequence repeats (SSRs) or short tandem repeats (STRs) loci are DNA stretches containing core repeat units in the range of 2-7 bp (Butler, 2009). Microsatellites account for 3% of the total human genome. They are among the most variable types of DNA sequence in the genome and their polymorphisms originate mainly from variability in length rather than in the primary sequence. Microsatellites rapidly became the markers of choice in genome mapping, forensics, and subsequently also in population genetic studies and related areas (Ellergen, 2004). Besides that, STR polymorphism has received increasing awareness as an effective genetic marker for analyzing medical and anthropological specimens.

STR typing is the most convenient and valuable technique of DNA fingerprinting technologies. Most forensic studies on STRs favored these genetic markers because of their reasonable rapid processing, their abundance throughout the genome, their high variability within various populations and their small size range, which allow multiplex PCR, where multiple markers are simultaneously amplified in a single reaction (Butler, 2003).

STRs are classified according to the number of nucleotides within a single core repeat unit. Tetranucleotide repeats are the most popular STR markers for human identification. The primary reason for tetranucleotides' popularity is the lower stutter percentage in comparison to stutter percentages for di or trinucleotides. Stutters are DNA amplification artifacts resulting from

slipped-strand mis-pairing in the PCR process. Also they account for the high mutation rate and thus the high level of polymorphism found at these loci (Butler, 2005).

Most STR loci used in population studies or in forensic work were chosen to be on different chromosomes, or at least located far from each other along a chromosome. The few markers that are found on the same chromosome in several studies failed to show any signs of significant linkage between loci as they are separated by millions of bases (Butler, 2006).

I sampled 498 unrelated individuals representing 18 Sudanese populations. In most cases, the individuals were sampled in different geographic locations around Sudan, but for some groups, the sampling took place in Khartoum and its surroundings since populations forming these groups recently migrated from their hometowns.

I use the AmpFℓSTR Identifiler® PCR Amplification Kit, (Applied Biosystems, USA) which is a fluorescent STR kit for human identification. It uses 5-dye chemistry and co-amplifies in one PCR reaction 15 STR loci (CSF1P0, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D21S11, FGA, TH01, TPOX, vWA) and the gender marker Amelogenin.

This study aims at assessing genetic diversity among the Sudanese populations using 15 autosomal markers. I explore the potential for human identification using standard statistical data analysis. I also investigated levels of population structure within Sudan, and, using previously published data from Egyptians (Omran *et al.*, 2009), Ugandans (Gomes *et al.*, 2009), and Somali (Tillmar *et al.*, 2009), I studied population structure among populations from east Africa.

I study the genetic variation in Sudanese sample populations from unrelated individuals, representing different ethnic, linguistic, and geographic groups which include Gaalien, Shaigia, Bataheen, Dinka, Shilluk, Nuer, Mahas, Danagla, Halfawieen, Hadendawa, Beni-Amer, Zagawa,

Hausa, Messiria, Nuba, Copts, Bari and Gemar. The group identification of individuals was based on self-reported ethnicity.

Furthermore, the 18 Sudanese populations will be grouped in larger context according to self-reported ethnicity and linguistic affiliation (Appendix A) and together with data from Egyptians (Omran *et al.*, 2009), Ugandans (Gomes *et al.*, 2009), and Somali (Tillmar *et al.*, 2009), I apply a Bayesian clustering approach using the genotyped STRs to detect the population structure by identifying clusters of genetically similar individuals (Pritchard *et al.*, 2000). I furthermore use statistical methods to address the informativeness of these markers for the inference of ancestry (Rosenberg *et al.*, 2003). In addition, I outline estimates of allelic richness and private allelic richness (Szpiech *et al.*, 2008) for the populations from present study and from previously published data.

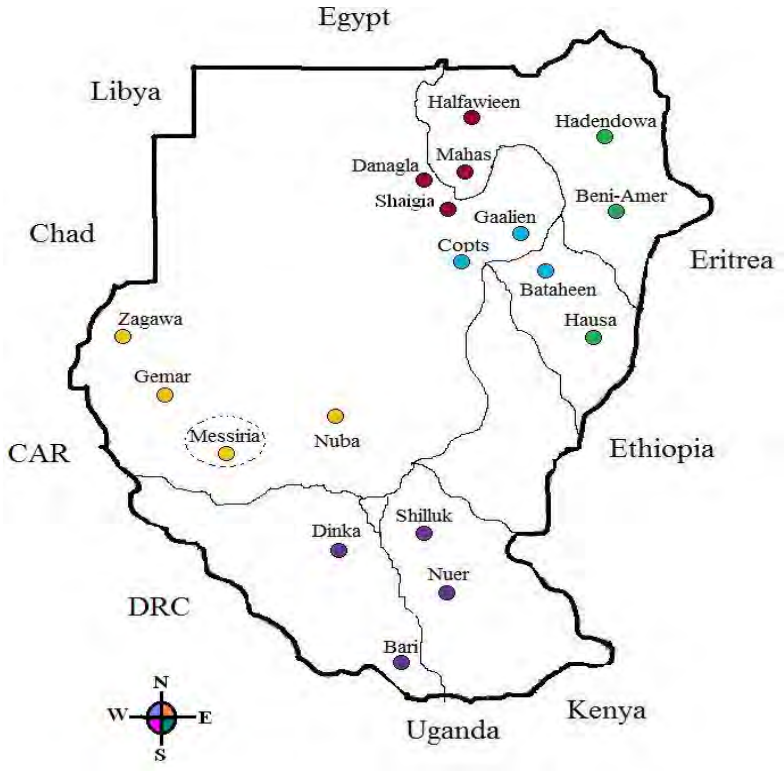


Figure 1

Map of Sudan representing the geographic locations of the 18 studied populations that were genotyped using the AmpFℓSTR Identifiler PCR Amplification Kit. The studied populations from northern, central, eastern, western and southern Sudan are represented by the red, light blue, green, yellow and purple colors respectively. The dotted circle represents the suggested location of Messiria population.

2. Materials and Methods

2.1. Blood Collection

Blood samples were collected from unrelated Sudanese individuals representing different Sudanese populations (geographic locations are shown in Figure 1). Prior to the collection of samples, a sample collection form[1] on self-reported ethnicity, parents, and grandparents was completed and a declaration of an informed consent[2] was obtained from all participants (Appendix B and C). Blood was collected from 498 individuals, 469 representing 18 Sudanese populations and 29 from various other Sudanese populations. Blood samples were taken from the finger using a new sterile lancet for each individual and, as described in Whatman FTA Protocol BD09, the blood drops were applied on the FTA card sample areas (WB120205: FTA Classic Card with 4 sample areas per card) and were air dried for at least 1 hour (Whatman, UK).

2.2. DNA Extraction

DNA was extracted from all dried blood samples on FTA cards following the manufacture's procedure as described in Whatman FTA Protocol BD01 except that the Whatman FTA purification reagent was modified to half the volume. A 1.2mm diameter disc was punched from each FTA card with a puncher on a cutting mat. The discs were transferred to new eppendorf tubes and washed 3 times in 100µl Whatman FTA purification reagent. Each wash was incubated for 5 minutes at room temperature with moderate manual mixing and the reagent was discarded between washing steps. The discs were then washed twice in 200µl TE buffer (10 mM Tris-HCl, 0.1 mM EDTA, pH 8.0), the buffer was discarded and the discs were left to dry at room temperature for 1 hour.

2.3. DNA Amplification

Extracted DNA was amplified using AmpFℓSTR Identifiler® PCR Amplification Kit (Applied

1, 2 The sample collection and consent forms are available for review by request from Dr. Mattias Jakobsson, Department of Evolutionary Biology, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden, mattias.jakobsson@ebc.uu.se

Biosystems, USA). The PCR mix was prepared according to the manufacturer's protocol by adding 10.5µl AmpFℓSTR Identifiler PCR Reaction Mix, 5.5µl AmpFℓSTR Identifiler Primer Set, and 2.5units AmpliTaq Gold DNA Polymerase to each tube, in addition to 10µl of ddH_2O.

The washed and dried discs were transferred into the PCR tubes in a total volume of 25µl. With each sample set a negative control sample (15µl PCR mixture and 10µl ddH_2O) and a positive control sample (15µl PCR mixture and 10µl (1 ng) AmpFℓSTR control DNA 9947A (0.1 ng/µl)) were included. The tubes were vortexed, centrifuged, and directly placed in a PCR machine (MJ Research Thermo Cycler PTC-225). The cycler was programmed as follows: Initial incubation step at 95 ºC for 11 minutes, 25 cycles of a denaturing step at 94 ºC for 1 minute, an annealing step at 59 ºC for 1 minute, and an extension step at 72 ºC for 1 minute, the final extension step occurred at 60 ºC for 60 minutes and holding was at 12 ºC. Some samples were extracted and amplified twice due to amplification failure at some of the loci during the first run.

2.4. DNA Genotyping and Analysis

The amplified Identifiler PCR products were genotyped with the ABI 3730 XL Genetic Analyzer (Applied Biosystems, USA). Each sample was prepared following the protocol of the EBL (2009). In each reaction tube, 1µL of PCR product was added to previously prepared mixture of 9.8 µL of Hi- Di Formamide (Applied Biosystems, USA), and 0.2 µL of GeneScan-600 LIZ Size Standard (Applied Biosystems, USA). The reaction tubes alongside with a tube including the same mixture in addition to 1µL of AmpFℓSTR Identifiler Allelic Ladder (Applied Biosystems, USA) were then placed in a PCR machine (MJ Research PTC-100 Thermal Cycler) for 3 minutes at 95°C to denature the DNA after which the samples were directly cooled on ice for 5 minutes prior to loading on the Genetic Analyzer. The conditions of the ABI 3730 XL Data Collection Software v.3.0 (Applied Biosystems, USA) were set as follows: module GS STR POP7 (1mL) G5v2, injection time of 15 seconds, injection voltage of 16.0 kV, run voltage of 15.0 kV and run temperature of 63°C. After the completion of run with the Data Collection software v. 3.0, the samples were analyzed with GeneMapper v.4.0 for Windows (Applied Biosystems, USA) applying the Microsatellite STR analysis method, which included the following settings: panel: Identifiler_v1, size standard: GS600LIZ (-20,600), peak detection mode: advanced, analysis range (data point): 3200–6300 and sizing range (bp): 75-400. For

some samples the analysis were repeated on the ABI genetic analyzer more than once to confirm the results.

2.5. Inference of relatedness

The genotyped data obtained from the GeneMapper software was analyzed using Relpair v2.0.1 to infer the relationships of pairs of individuals. The program calculates and compares the multipoint probability of genetic marker data conditional on different pairwise genetic relations, and infers the relationship that makes the data most likely (Boehnke and Cox, 1997; Epstein *et al.*, 2000).

2.6. Calculations of forensic related parameters

The allele frequencies, power of discrimination, power of exclusion, combined power of exclusion, random match probability, combined match probability and polymorphism information content for each genotyped STR locus were calculated using the Promega PowerStats v.12 (Promega, USA). In addition, the observed heterozygosity, expected heterozygosity, and exact test probability values for Hardy-Weinberg equilibrium were calculated using Arlequin v. 3.11 (Excoffier *et al.*, 2005).

2.7. Statistical data analysis

2.7.1. Analysis of allelic diversity

The mean number of distinct alleles (allelic richness), the mean number of private alleles (private allelic richness) and the mean number of shared alleles between populations were analyzed using the ADZE program (Sziepch *et al.*, 2008). The program uses a rarefaction method (Kalinowski, 2004) that correct for the sample size across populations. This method estimates the number of distinct alleles in each set of population and estimates the number of alleles in each set of population but absent in all remaining populations. It also computes the private allelic richness to measure the number of distinct alleles private to a group of populations considering equal-sized sub-samples from each population.

The ADZE program implements these calculations for all 15 loci in my data set and gives the mean, variance and standard error for these loci. The analysis was applied to populations from this study alongside populations from previously published data.

2.7.2. Statistical measurement for the inference of ancestry

The informativeness for assignment (In) provides a natural measure of potential for assignment of an allele to one population compared with the average population. The informativeness for ancestry coefficients (Ia) provides a natural measure of potential for assignment of an allele to a point on the set of all possible ancestry coefficient vectors. The optimal rate of correct assignment (ORCA) gives the probability of correct assignment of an allele using the decision rule with lowest risk (Rosenberg *et al.*, 2003). 1-allele version of ORCA is the optimal rate of correct assignment for a locus if a randomly chosen allele in a pooled collection of populations is assigned to its most likely source population. A 2-allele version of ORCA is the optimal rate of correct assignment for a locus if a randomly chosen diploid genotype is assigned to its most likely source population. The statistics were calculated assuming that each population is equally likely to be the source population to determine the amount of information in the studied loci in an attempt to assess the informativeness of these loci for inference of ancestry. I used the program Infocalc (Rosenberg *et al.*, 2003) to calculate the informativeness for assignment (In), the informativeness for ancestry coefficients (Ia), and optimal rate of correct assignment (ORCA [1-allele]) and (ORCA [2-allele]).

2.7.3. Population structure

The computer program Structure is a multilocus model-based clustering method that assigns individuals to a predefined number of clusters (K) and detects admixed/migrant individuals (Pritchard *et al.*, 2000). Structure was run with the full data of 1167 individuals including 454 Sudanese from this study, 218 Ugandans (Gomes *et al.*, 2009), 265 Egyptians (Omran *et al.*, 2009) and 230 Somali (Tillmar *et al.*, 2009). We ran Structure ten times for K-values from 2 to 10, employing the admixture model for individual ancestry, the F model for allele frequency correlation and a burn-in period of 100,000 followed by 10,000 repeats.

The resulted runs were uploaded into Structure Harvester (Evanno, 2005) to analyze the (k) values and then combined with the program CLUMPP (Jakobsson and Rosenberg, 2007) and the combined data was visualized with DISTRUCT (Rosenberg, 2004).

2.7.4. Analysis of molecular variance (AMOVA)

Analysis of the molecular variance (AMOVA) was implemented to determine statistical differences between all the predefined linguistic and geographic groups (Appendix A and Figure 1) using Arlequin v. 3.11 (Excoffier *et al.*, 2005). The Nuba was excluded from linguistic group analysis as the individuals belonged to both the Nilo-Saharan and Niger-Congo languages and the Messiria was excluded from the geographic group since it is a widely spread nomadic group.

2.7.5. Principal component analysis (PCA)

The genetic data for the populations from this study alongside with the published data for Ugandans, Egyptians, and Somali were used to compute the genetic distances for microsatellite loci described by Goldstein *et al.* (1995) and Slatkin (1995) using Populations v. 1.2.30. The created genetic distance matrix for the populations was used to perform principal component analysis and displayed graphically using the PAST software (available at http://folk.uio.no/ohammer/past/).

3. Results

3.1. Inference of relatedness

Based on Relpair v. 2.0.1 analysis, seven individuals with first-degree relationship as the most likely relationship and with a scaled likelihood ratio of 1.00 were excluded from the dataset and from further population genetic analysis, which resulted in 491 individuals out of 498 genotyped individuals.

3.2. Calculations of forensic related parameters

For the 498 samples, forensic calculations were carried out for each locus with different number of samples due to missing alleles at some loci (see Table 1). The number of observed alleles, the most frequent and the least frequent alleles and their percentages were calculated for each locus and are shown in Table 1. The frequency of the most common allele ranged from 0.397 for allele 7 at the locus TH01 to 0.14 for allele17 at the locus D18S51.

Table 1 Number of observed alleles, frequency of the most frequent alleles, and frequency of the least frequent alleles for the studied 15 loci in the Sudanese population

Locus	Number of observed alleles	Most frequent allele and percentage	Least frequent allele and percentage
CSF1PO	8	10 (31.6%)	7 (0.7%)
D13S317	8	12 (38.4%)	15 (0.3%)
D16S539	8	11 (32%)	15 (0.1%)
D18S51	23	17 (14%)	9, 10.2, 11.2, 14.2, 16.2, 19.2 (0.1% each)
D19S433	18	14 (26.1%)	18, 18.2 (0.1% each)
D21S11	20	29 (28.6%)	25.2 (0.1%)
D2S1338	14	19 (20.9%)	15, 27, 28 (0.1% each)
D3S1358	11	15 (30.5%)	12, 15.2, 16.2 (0.1% each)
D5S818	9	12 (36.6%)	7, 15 (0.1% each)
D7S820	7	10 (34.3%)	7 (0.4%)
D8S1179	13	14 (29.8%)	7, 8, 19 (0.1% each)
FGA	19	22 (18.3%)	18.2, 19.2 (0.1% each)
TH01	7	7 (39.7%)	4 (0.1%)
TPOX	8	8 (39%)	13 (0.1%)
vWA	11	16 (25.9%)	22 (0.1%)

The observed allele frequencies for each of the 15 STR loci in the Sudanese population sample are shown in table 2. However, since there are some alleles which were not sampled sufficiently and that an estimate of an allele frequency is uncertain if the allele is so rare that it is represented only once or a few times in a dataset, it is recommended that each allele be observed at least five times to be used in forensic calculations (Butler, 2009). Therefore, I calculated the minimum allele frequency for each locus where some alleles were not sufficiently sampled (shown in table 3). The minimum allele frequency is 5/ (2n) where n is the number of individuals sampled and 2n is the number of chromosomes (as autosomes are in pairs due to inheritance of one chromosome from each parent).

Table 2 Allele Frequencies for each of the genotyped 15 STR loci in this study

Allele	CSF1PO	D13S317	D16S539	D18S51	D19S433	D21S11	D2S1338	D3S1358	D5S818	D7S820	D8S1179	FGA	TH01	TPOX	vWA
	N: 485	N: 489	N: 491	N: 485	N: 491	N: 491	N: 489	N: 490	N: 490	N: 485	N: 491	N: 486	N: 489	N: 491	N: 491
4													0.001*		
6													0.228	0.006	
7	0.007								0.001*	0.004*	0.001*		0.397	0.005	
8	0.048	0.069	0.039	0.002*					0.084	0.228	0.001*		0.091	0.390	
9	0.037	0.048	0.189	0.001*	0.002*				0.037	0.106	0.003*		0.167	0.292	
9.3													0.083		
10	0.316	0.027	0.069	0.004*	0.014				0.098	0.343	0.027		0.034	0.100	
10.2				0.001*											
11	0.242	0.311	0.320	0.010	0.019				0.202	0.212	0.053			0.189	
11.2				0.001*	0.004*										
12	0.292	0.384	0.205	0.101	0.075			0.001*	0.366	0.091	0.113			0.017	0.002*
12.2					0.010										
13	0.048	0.128	0.147	0.087	0.225			0.003*	0.203	0.015	0.241			0.001*	0.003*
13.2				0.002*	0.055										
14	0.008	0.031	0.031	0.096	0.261			0.065	0.008		0.298				0.077
14.2				0.001*	0.078										
15		0.003*	0.001*	0.111	0.092		0.001*	0.305	0.001*		0.188				0.156
15.2					0.082			0.001*							
16				0.136	0.033		0.051	0.263			0.060				0.259
16.2				0.001*	0.033			0.001*							
17				0.140	0.008		0.139	0.271			0.010	0.008			0.245
17.2				0.002*	0.007										
18				0.114	0.001*		0.091	0.082			0.003*	0.009			0.155
18.2					0.001*							0.001*			
19				0.078			0.209	0.005			0.001*	0.047			0.075
19.2				0.001*								0.001*			
20				0.059			0.128	0.002*				0.065			0.024
21				0.034			0.060					0.108			0.003*
21.2												0.005*			
22				0.013			0.128					0.183			0.001*
22.2												0.002*			
23				0.003*			0.091					0.166			
24							0.056					0.143			
24.2						0.003*						0.002*			
25							0.035					0.091			
25.2						0.001*									
26						0.002*	0.009					0.039			
27						0.030	0.001*					0.026			
28						0.123	0.001*					0.063			
29						0.286						0.035			
30						0.226									
30.2						0.005						0.006			
31						0.082									
31.2						0.039									
32						0.017									
32.2						0.080									
33						0.007									
33.2						0.034									
34						0.009									
34.2						0.005									
35						0.026									
36						0.017									
37						0.005									
38						0.002*									

* = Alleles for which less than 5 copies were sampled from the population; the minimum allele frequency for this dataset is (5/2n) and should be used in forensic calculations (Butler, 2009).

Table 3 shows the results of additional calculated forensic parameters for the 15 loci. The power of discrimination is the potential power of a genetic marker to differentiate between any two

individuals chosen at random; PD = $1 - \sum_j P(i,j)^2$ where p(i, j) is the frequency of genotype j for locus i and it ranged from 0.979 (D18S51) to 0.874 (TPOX), the observed heterozygosity from 0.895 (D18S51) to 0.684 (TPOX), and the polymorphism information content from 0.889 (D18S51) to 0.67 (TPOX), which all indicate that the (D18S51) is the most polymorphic and discriminating locus.

The power of exclusion is the probability to exclude for example a non-true father in a paternity test and it can be calculated for each locus as PE = h^2 (1-2*h*H^2) where h = observed heterozygosity and H = observed homozygosity. The power of exclusion ranged from 0.785 (D18S51) to 0.406 (TPOX), presumably the higher the number the more useful the marker is. Given that power of exclusion and power of inclusion should add up to 1 or 100% and so power of exclusion = (1 - Power of inclusion) I used the calculated power of exclusion for each locus to calculate the power of inclusion for each locus and then use it to calculate the combined power of exclusion for the 15 loci from 1-((1-PE_1) (1-PE_2)...(1 PE_n), where PE_n is the power of exclusion for marker n. The combined power of exclusion for all the 15 loci was calculated to be 99.99981% (Brenner *et al.*, 1990; Butler, 2009).

The match probability is the probability for a random match between two unrelated individuals drawn from the same population. Simply, it is the estimated frequency at which a particular STR profile would be expected to occur in a population. It is the sum of the frequency squared of each genotype e.g. PM_i = $\sum_j P(i,j)^2$ where p(i, j) is the frequency of genotype j for locus i (Butler, 2009). The total or combined match probability (MP) for the 15 loci is calculated as PM_tot = PM_1*PM_2*...*PM_n which was 1.35 x 10^{-18} or 1 in 7.4 x 10^{17}.

Hardy-Weinberg Equilibrium expectations were met for all loci except for D13S317, D2S1338, D7S820, D8S1179, vWA (p-values: 0.021, 0.022, 0.015, 0.005 and 0.018, respectively) which are < 0.05 (significance level). However, after applying the Bonferroni's correction which assumes that a 0.05 significance level obtained for 15 tests (one per locus) gives an actual significance level of $p < 0.0033$. Therefore the values for these loci were not significant.

Table 3 Forensic statistical data and the exact test probability values for Hardy-Weinberg equilibrium

Locus	CSF1PO	D13S317	D16S539	D18S51	D19S433	D21S11	D2S1338	D3S1358	D5S818	D7S820	D8S1179	FGA	TH01	TPOX	vWA
	N: 485	N: 489	N: 491	N: 485	N: 491	N: 491	N: 489	N: 490	N: 490	N: 485	N: 491	N: 486	N: 490	N: 491	N: 491
MP	0.103	0.114	0.073	0.021	0.043	0.047	0.028	0.106	0.090	0.093	0.069	0.026	0.100	0.126	0.066
PD	0.897	0.886	0.927	0.979	0.957	0.953	0.972	0.894	0.91	0.907	0.931	0.974	0.9	0.874	0.934
PE	0.504	0.409	0.552	0.785	0.665	0.611	0.68	0.533	0.519	0.469	0.503	0.748	0.431	0.406	0.681
PIC	0.708	0.689	0.762	0.889	0.833	0.816	0.865	0.711	0.732	0.729	0.769	0.874	0.711	0.67	0.79
H_{obs}	0.746	0.687	0.774	0.895	0.833	0.804	0.843	0.763	0.755	0.726	0.745	0.877	0.702	0.684	0.841
H_{exp}	0.751	0.731	0.792	0.899	0.850	0.834	0.878	0.754	0.767	0.766	0.798	0.886	0.747	0.718	0.813
P	0.955	0.021	0.335	0.357	0.195	0.211	0.022	0.335	0.258	0.015	0.005	0.749	0.311	0.400	0.018
MAF	0.0052	0.0051	0.005	0.0052	0.005	0.005	0.0051	0.0051	0.0051	0.0052	0.005	0.0052	0.0051	0.005	0.005

(HWE exact test with 100,000 steps in markov chain and 1000 of dememorization steps), MP (Matching Probability), PD (Power of discrimination), PE (Power of exclusion), PIC (Polymorphism information content), H_{obs} (Observed Heterozyosity), H_{exp} (Expected Heterozygosity), P (Hardy-Wienberg equilibrium, exact test P value based), MAF (Minimum allele frequency (5/2n))

3.3. Population data analysis

For the 498 genotyped individuals, a total of 454 individuals representing 18 ethnic groups were included in the population data analysis as I excluded 44 individuals, 29 individuals from other population groups, seven individuals with first-degree relationship and eight individuals due to missing alleles at some loci.

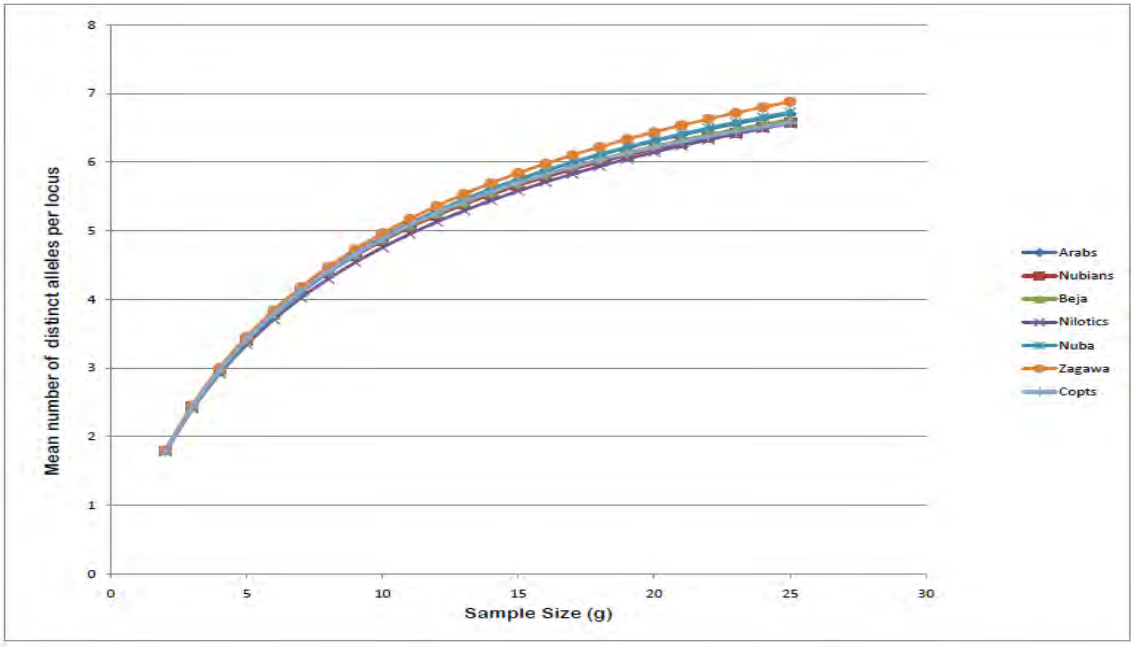
3.3.1. Analysis of allelic diversity

ADZE (Sziepch *et al.*, 2008) was used to study the allelic diversity in the examined populations by mean number of distinct alleles, mean number of private alleles and mean number of shared alleles between combinations of two populations as a function of a standardized sample size. Even though the program corrects for the sample size, I excluded the Messeria, Hausa, Bari, Gemar populations from this analysis because of their small sample size to raise the minimum sample size from each population.

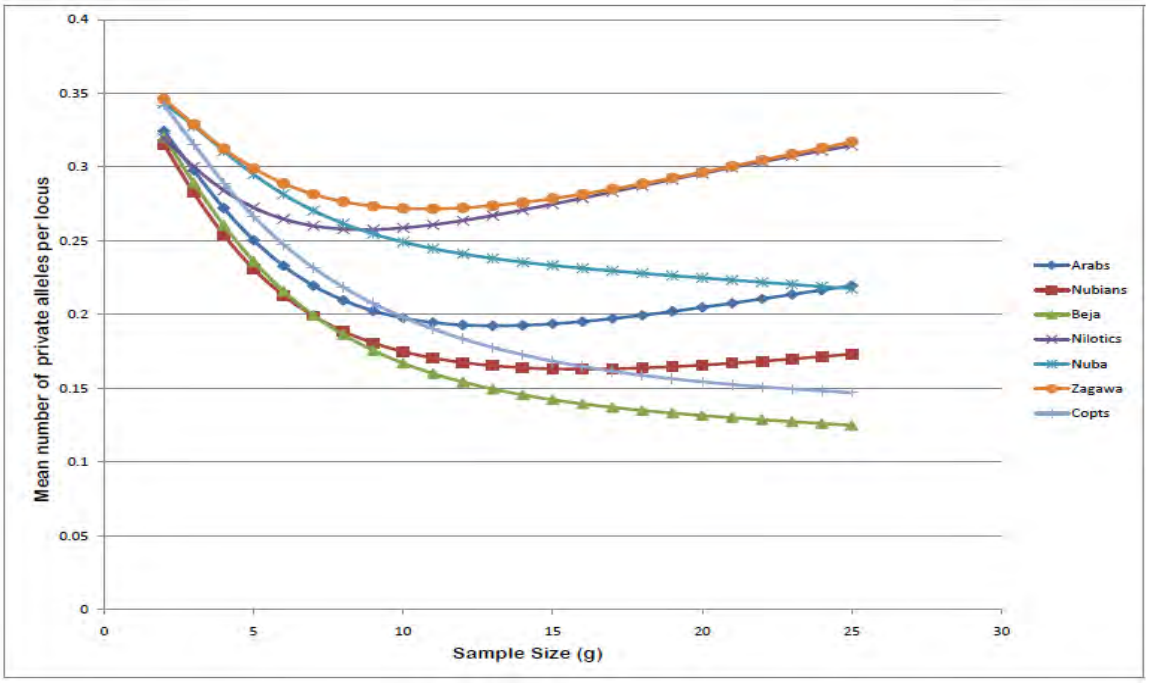
In this context the allelic richness, the private allelic richness and the uniquely shared alleles were calculated for the studied populations shown in figures 2A, 2B and 2C respectively, considering equal-sized sub-samples from each population. These computations demonstrated that the Zagawa, Nuba and Arabs are the richest in distinct alleles, whereas the Nubians, Beja, Copts and Nilotics have fewer distinct alleles. On the other hand, the mean number of the

private alleles per locus was found to be the richest in Zagawa, Nilotics and Nuba compared to the Arabs, Nubians, Copts and Beja. Moreover, the shared alleles between combinations of Sudanese populations (Figure 2C) resulted in a higher mean number of shared alleles between the Nuba and Zagawa followed by the Nuba and Nilotics. Thus, the Zagawa, Nuba and Nilotics stand out as being more diverse populations, and that are most distinct from other Sudanese populations.

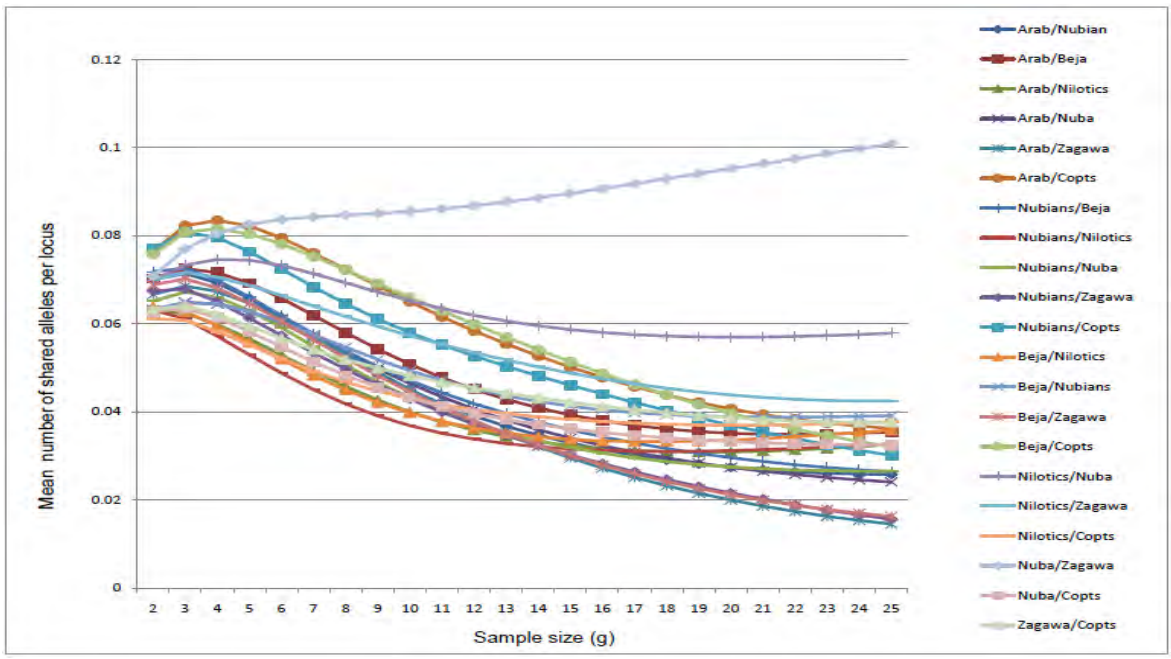
In addition, and as illustrated in figure 3A, 3B and 3C, the allelic richness, the private allelic richness and uniquely shared alleles were calculated for the populations from this study with one Ugandan, one Egyptian and one Somali population, altogether considering equal-sized sub-samples from each population. The mean number of distinct alleles per locus was the highest in the Zagawa, Ugandans, Nuba and Arabs while in Egyptians, Beja, Nilotics, Somali Nubians and Copts was the lowest. The mean number of private alleles was high in the Somali, Nilotics, Zagawa, and Ugandans and lower in Arabs, Egyptians, Nuba and Nubians and the lowest in the Copts and the Beja. In addition, the uniquely shared alleles between Sudanese, Ugandans, Egyptians and Somali resulted in a higher mean number for Zagawa and Ugandans followed by Nuba and Ugandans, Nuba and Zagawa and Nilotics and Ugandans (Figure 3C).



A Mean number of distinct alleles per locus.

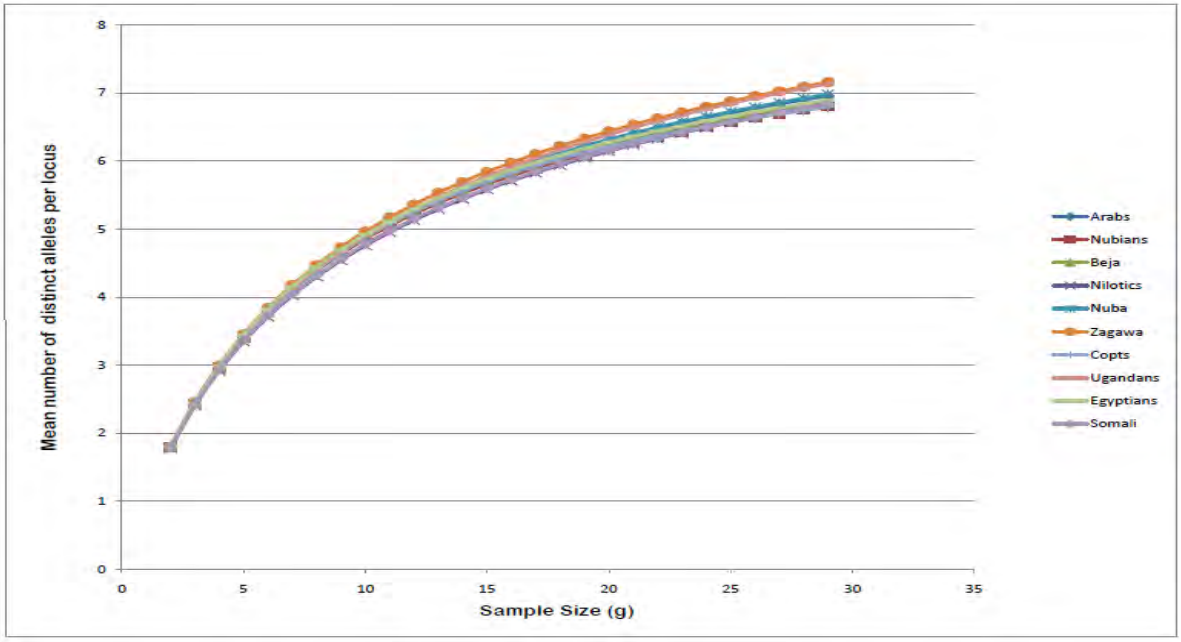


B Mean number of Private alleles per locus.

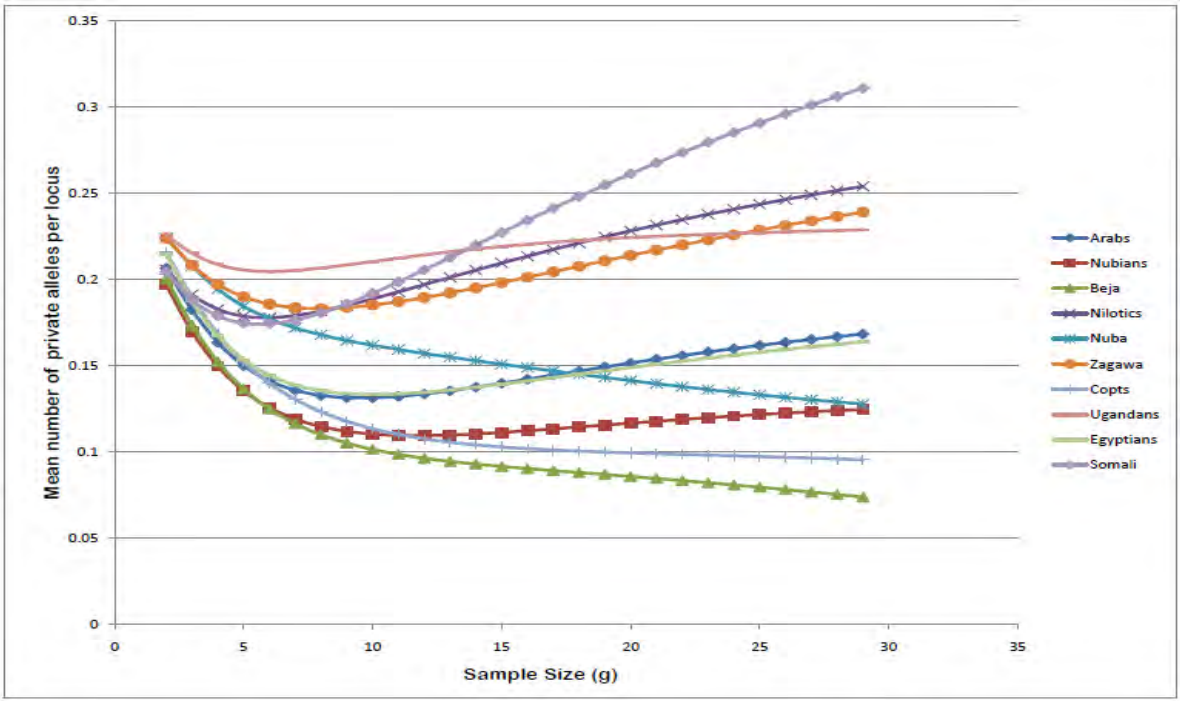


C Mean number of uniquely shared alleles per locus.

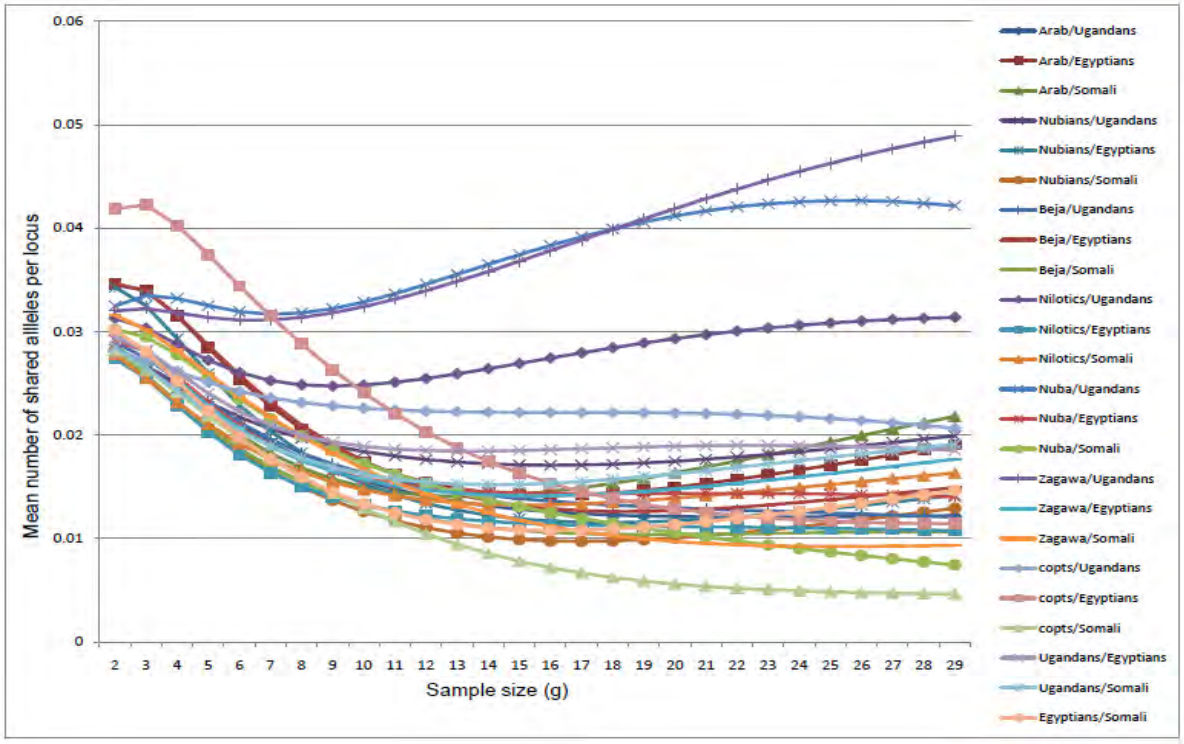
Figure 2 Inference of the mean number of distinct alleles, mean number of private alleles and mean number of uniquely shared alleles. A) Mean number of distinct alleles within Sudanese populations. B) Mean number of private alleles within Sudanese populations. C) Mean number of uniquely shared alleles between Sudanese populations.



A Mean number of distinct alleles per locus.



B Mean number of private alleles per locus.



C Mean number of uniquely shared alleles per locus.

Figure 3 Inference of the mean number of distinct alleles, mean number of private alleles and mean number of uniquely shared alleles. A) Mean number of distinct alleles between Sudanese, Egyptian, Ugandan and Somali populations. B) Mean number of private alleles between Sudanese, Egyptian, Ugandan and Somali populations. C) Mean number of uniquely shared alleles between Sudanese, Egyptian, Ugandan and Somali populations.

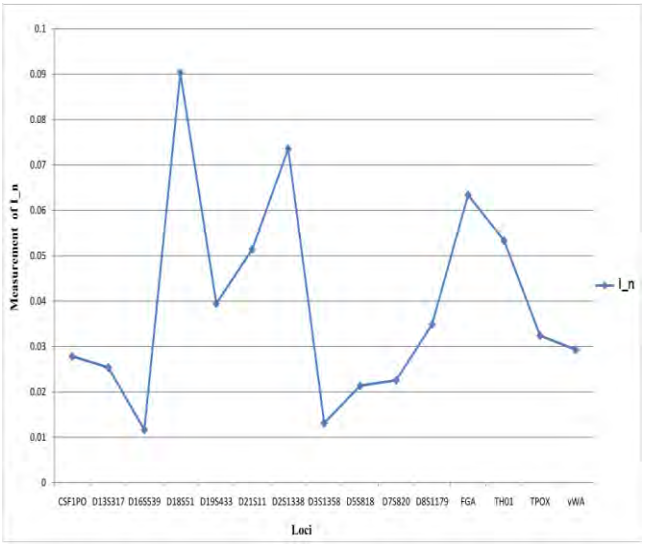
3.3.2. Statistical measurement for the inference of ancestry

I determined the informativeness for assignment measurements for each of the studied 15 loci and they are summarized in Table 4. It shows the values of the statistical measurement: the informativeness for assignment (In), the informativeness for ancestry coefficients (Ia), and optimal rate of correct assignment (ORCA). We focused on the computed In as Rosenberg *et al.* (2003) stated that the statistics of In, Ia, and ORCA produce similar estimates and that the In is most robust.

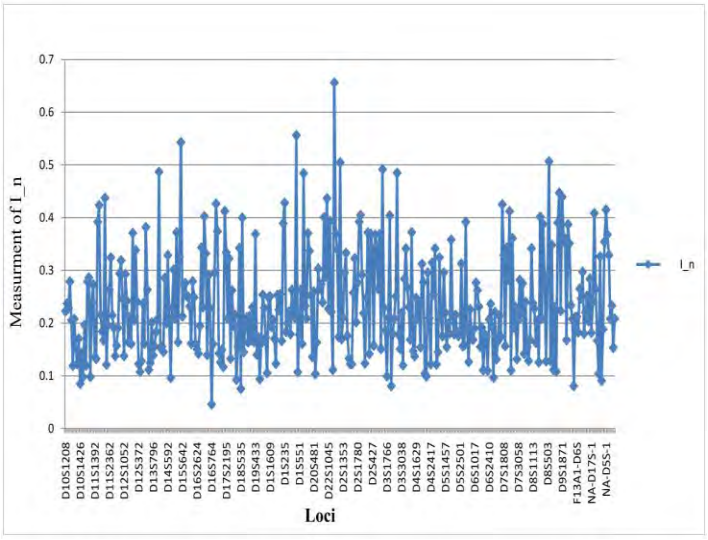
Table 4 Computed informativeness statistics

Locus	In	Ia	ORCA[1-allele]	ORCA[2-allele]
CSF1PO	0.027825	0.005543	0.304099	0.338277
D13S317	0.025381	0.005047	0.303437	0.339592
D16S539	0.011621	0.002336	0.300239	0.317372
D18S51	0.090298	0.017898	0.363192	0.425311
D19S433	0.039461	0.007906	0.325635	0.366693
D21S11	0.051414	0.009988	0.331272	0.379159
D2S1338	0.073571	0.014866	0.361602	0.419577
D3S1358	0.013129	0.002604	0.288327	0.305432
D5S818	0.021359	0.004273	0.306461	0.339397
D7S820	0.022552	0.004653	0.307257	0.328783
D8S1179	0.034887	0.007111	0.323764	0.366753
FGA	0.0634	0.012136	0.324558	0.37189
TH01	0.053337	0.010916	0.343666	0.389454
TPOX	0.032397	0.00647	0.325244	0.349902
vWA	0.029262	0.00557	0.311902	0.344487

In gives the amount of information gained about population assignment from observation of a single randomly chosen allele at a locus. In measured from this study is shown in Figure 3A for the 15 included loci and compared to In values for 377 loci in 6 African populations studied by Rosenberg *et al.* (2003), which was obtained from online supplement of Rosenberg *et al.* (2003) (shown in figure 3B). The difference in informativeness level for inference of ancestry is clearly evident from the comparison of In produced by the loci in my study compared to those studied by Rosenberg *et al.* (2003).



A



B

Figure 3 A) Informativeness for assignment for the studied 15 loci in populations of my recent study and the three populations from the published data. **B)** Informativeness for assignment for the 377 loci in 6 African populations studied by Rosenberg *et al.* (2003)

3.3.3. Population structure

I estimated population structure for the Sudanese populations together with Somali, Egyptian and Ugandan populations. The *L(K)* estimate for the number of groups given by structure often does not correspond to the real number and the most likely K observed at k = 2 from the highest *lnP (X/K)* resulted from Structure Harvester, however as stated by Evanno *et al.* (2005) that the real number of groups is best detected by the modal value of ΔK, a quantity based on the second order rate of change with respect to K of the likelihood function. Although most populations showed substantial contributions from all K clusters, three main cluster patterns could be distinguished (illustrated in Figure 5). The first cluster pattern includes the Nilo-Saharan speaking groups from Sudan (Zagawa, Nuba, Dinka, Nuer, Shilluk, Bari, and Gemar) and the Afro-Asiatic Hausa with the Nilo-Saharan speaking group from Uganda (Karamoja). The second cluster pattern includes the Nilo-Saharan speaking group Danagla, Mahas and Halfawieen and the Afro-Asiatic speaking groups Bataheen, Gaalien, Shaigia, Messiria, Hadendowa, Beni-Amer and Copts with the Afro-Asiatic speaking group the Egyptians. The third cluster pattern represents the Somali population.

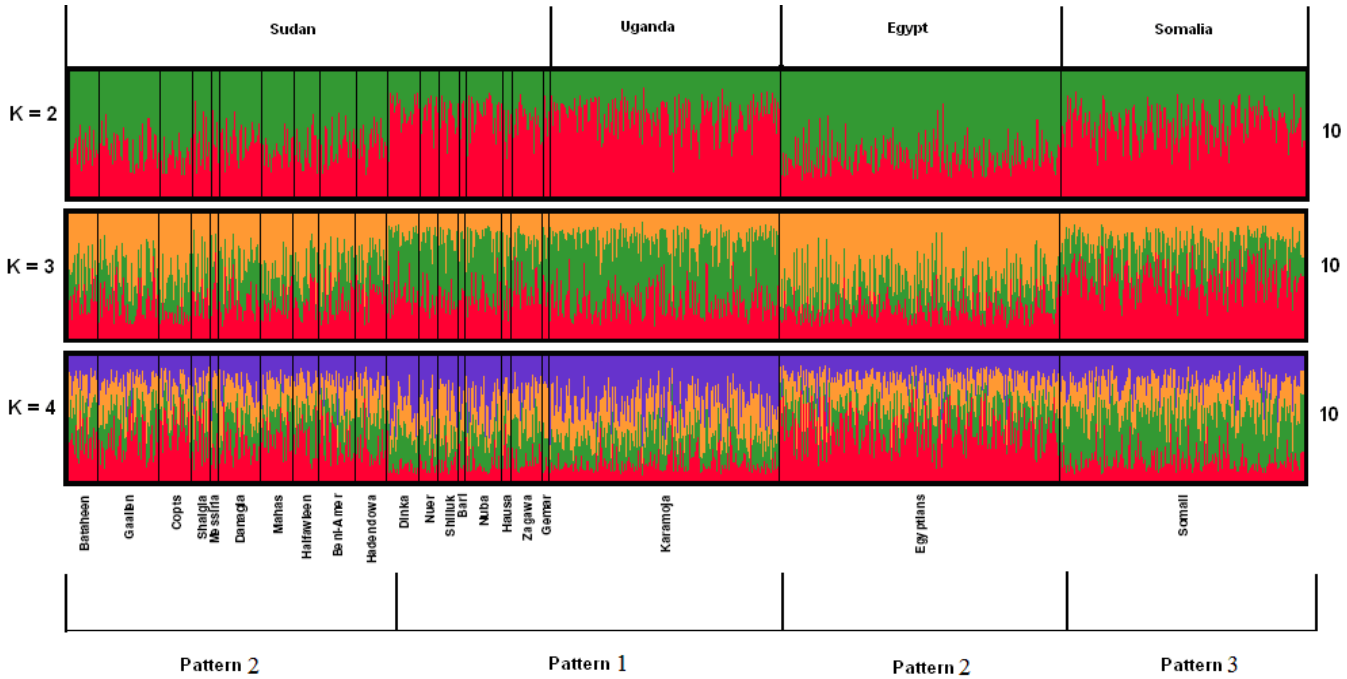


Figure 5

Results of Structure analysis for the studied Sudanese populations, Ugandans, Egyptians and Somali at K = 2-4. Each population is represented by a column divided into K colors with each color representing a cluster and each single line representing an individual. Different populations are separated by a black line and are labeled below the figure by self reported ethnicity and above the figure by geographic region.

3.3.4. AMOVA the molecular variance

I performed an AMOVA analysis in order to determine the amount of genetic variation within and among populations. The 18 Sudanese populations showed that almost all the genetic variation was found to be within populations. The results are shown in table 5.

AMOVA was analyzed for the studied populations in five groups according to geographic locations (Central, Northern, Eastern, Western and Southern) and two groups according to linguistic affiliation (Afro-Asiatic and Nilo-Saharan). The Nuba population was excluded from linguistic grouping as the sample group contained both the Nilo-Saharan and Niger-Congo speaking groups and the Messiria was excluded from geographic grouping because it is a widely spread nomadic group.

When populations were grouped according to geographic locations and linguistic affiliation, the highest genetic variance was found within populations (98.99%) and (99.07%) respectively. In linguistic groups, the variance among populations within groups was (0.73%) and among the groups was (0.21%), while in geographic groups the variance among populations within groups (0.49%) and among the groups (0.52%). Therefore, geography appears to be slightly better at accounting for genetic variation than the linguistic classification (0.52% vs. 0.21%).

Table 5 Analysis of molecular variance (AMOVA)

		Within populations		Among populations within groups		Among groups	
Group	No. of groups	Variance (%)	F_{ST}	Variance (%)	F_{SC}	Variance (%)	F_{CT}
Linguistic groups[a]	2	99.07	0.00934	0.73	0.00729	0.21	0.00207
Geographic groups[b]	5	98.99	0.01011	0.49	0.00497	0.52	0.00516

For all F-statistics, *P*-values are less than 0.01

a Afro-Asiatic and Nilo-Saharan (see Appendix A).

b Central, Northern, Eastern, Western and Southern (see Figure 1).

3.3.5. Principal Component Analysis

The genotype data for the 15 microsatellite loci was used to create genetic distance matrix for the populations. The microsatllite distances described by Goldstein (1995) and Slatkin (1995) were computed and displayed graphically in a form of principal component plot (Figure 6). The Sudanese populations were grouped into larger context (Appendix A) according to self-reported ethnicity.

Principal component 1 explained (24.85%) of the genetic variation distinguishes Egyptians, Somali, Nubians and Copts from Ugandans and Arabs, Hausa, Beja, Nuba, Nilotics, Zagawa and Gemar from Sudan (Figure 6).

The second PC (12.78%) distinguishes Egyptians and Copts from the other groups. The Somali and Nubians slightly clustered towards Egyptians (Figure 6). In addition, the third PC (11.05%) distinguishes Somali population from all other populations (Figure 7).

PCA results indicate genetic similarity for the Copts and Nubians with Egyptians. Furthermore the Nuba, Zagawa, Nilotics and Gemar associate with Ugandans as also suggested by the Structure analysis. It is worth noting that the clustering of the populations could be affected by uneven sampling (McVean, 2009) even though I grouped the major samples in larger context to level sample sizes.

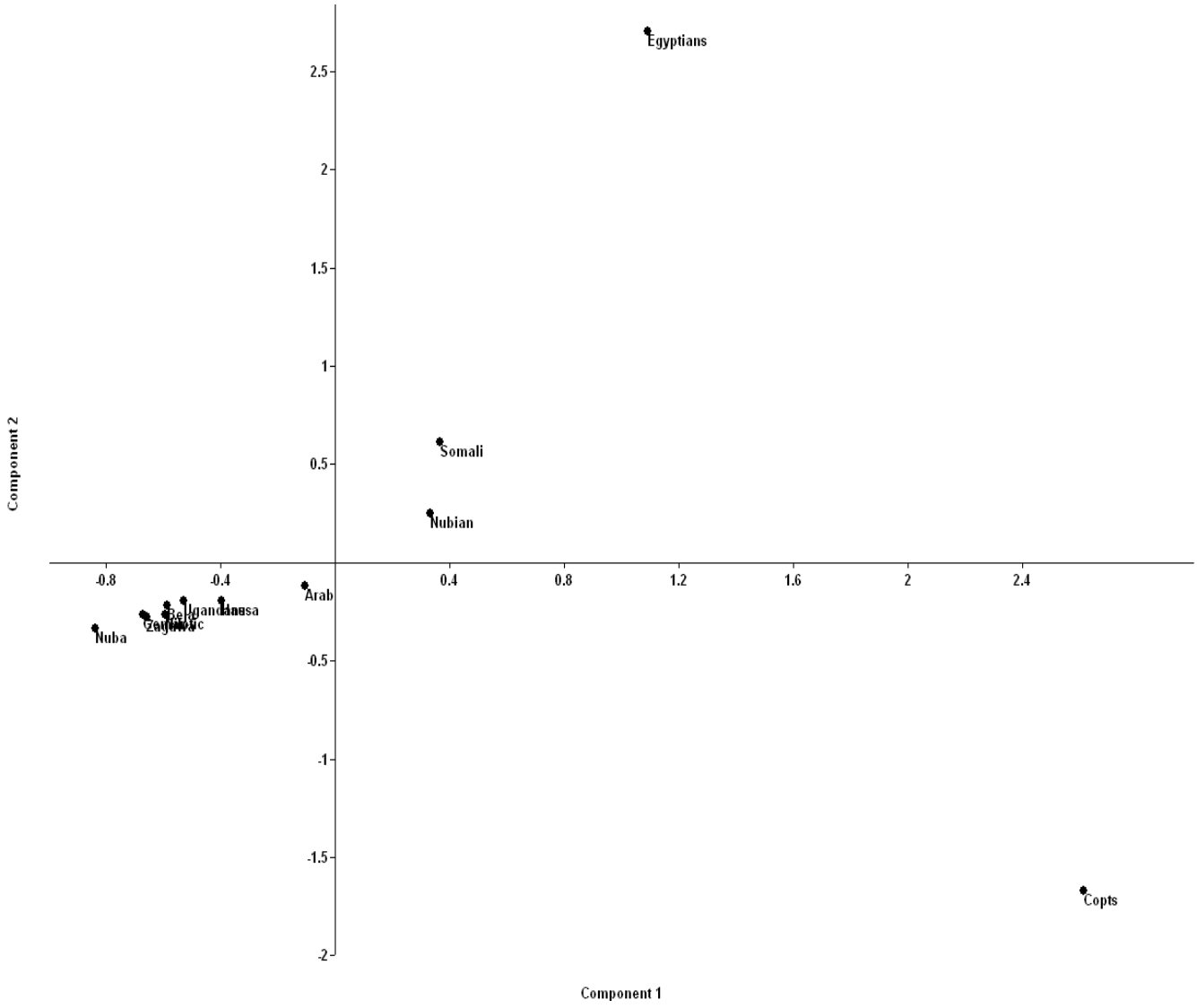


Figure 6

Plot showing the first and the second principal components of the genotypic data for Sudanese populations from the present study compared with Ugandans, Egyptians and Somali.

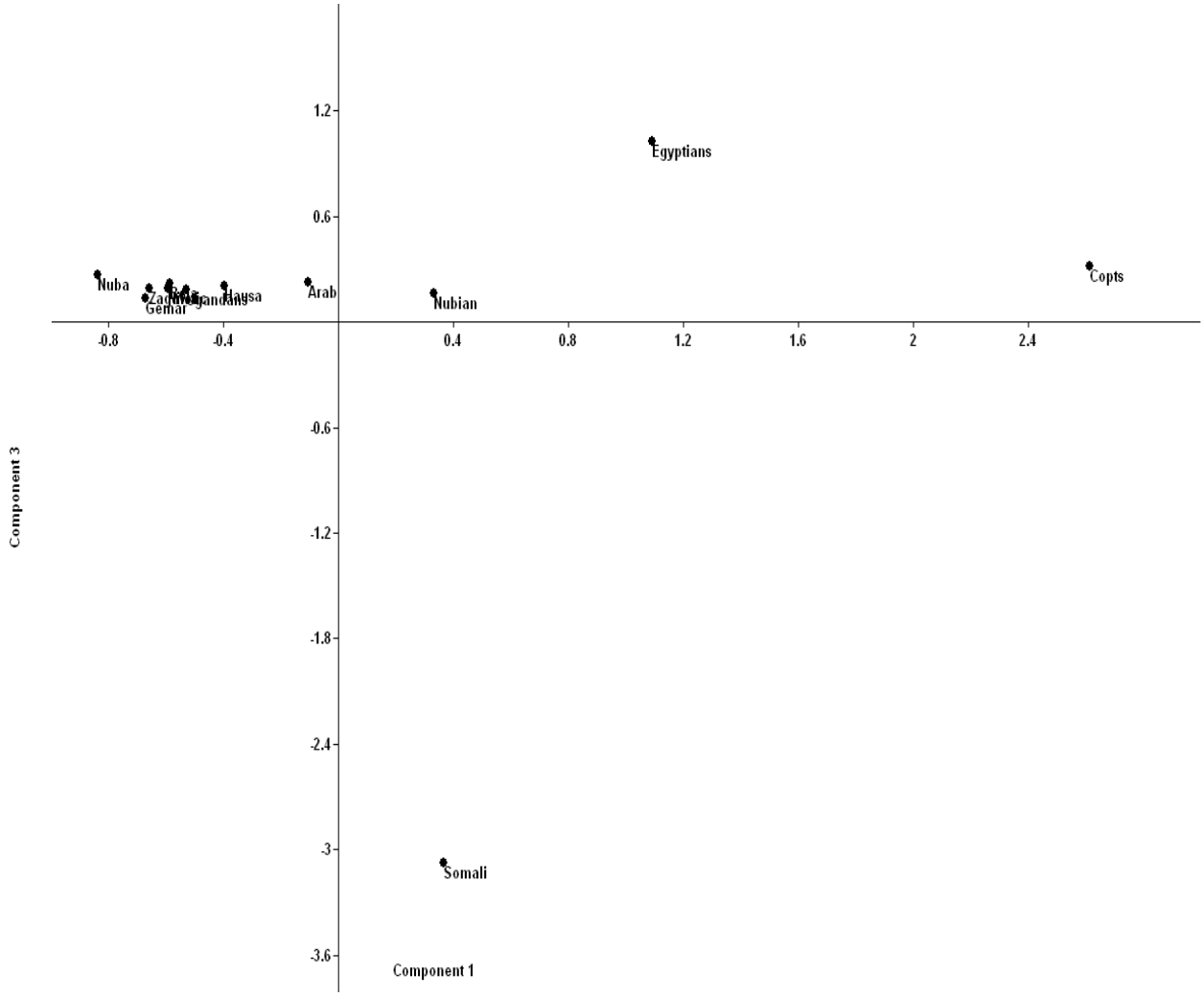


Figure 7

Plot showing the first and the third principal components of the genotypic data for Sudanese populations from the present study compared with Ugandans, Egyptians and Somali.

4. Discussion

In this study I determined the frequencies of the 15 Identifiler STR loci in populations from Sudan. The calculated combined power of exclusion for the 15 loci in the Sudan sample was 99.99981%, demonstrating that these markers can be used for paternity determinations for the Sudanese. Moreover, the calculated combined match probability, which was 1 in 7.4 x 10^{17}, will have an important role in the analysis of mixed DNA profiles such as in sexual assault cases. The allele frequency population dataset presented in this study should be used when calculating random match probabilities and paternity indices among individuals of Sudan. In addition, the calculated minimum allele frequency for each of the genotyped loci will be useful in forensic cases, in particular when an allele is found which has an estimated frequency below the threshold. In such cases, the minimum allele frequency should be used instead of an observed allele's frequency, which has been sampled insufficiently to get reliable estimates for the studied case profile.

The studied 15 STR markers have become important tools in modern society and are the most popular markers for human identification both in forensic casework and paternity testing and will continue to be widely used for many years because of their high degree of variability and the relatively high rate of mutation which are advantageous properties for individual identification rather than for estimating ethnic origin (Butler, 2006). Moreover, the 15 Identifiler STRs are mostly preferred in the field of forensic because of their lower stutter ratio, which is beneficial in the case of identifying mixed samples that are very common in biological evidences (Butler, 2005). In addition, low amounts of DNA, even in a degraded form, can be successfully typed with these STRs compared to previously used technologies (Ruitberg *et al.*, 2001).

The microsatellites studied by Rosenberg *et al.* (2003) for the worldwide dataset (including African populations) reported tetranucleotides as the least informative class of microsatellites for population structure inference compared to both dinucleotides and trinucleotides. From the comparison of the computed In values for the 15 Identifiler STRs from this work to those from Rosenberg *et al.* (2003), I found that the 15 Identifiler STRs (tetranucleotide microsatellites) have very low informativeness for inference of population structure. This observation can be

explained by the relatively high rate of mutation with these STRs, which makes it challenging to separate alleles that are identical by state from those identical by descent (Butler, 2006).

A recent study by Philips *et al.* (2010) investigated global variability in the 15 Identifiler STR loci and 5 new European Standard Set (ESS) STRs, which were applied to the CEPH human genome diversity panel. The Identifilier STR loci showed a higher error rate compared to ESS loci, even though Philips *et al.* (2010) demonstrated that Africans are better differentiated with the Identifiler loci than by ESS loci.

The studied informativeness of the Identifiler STRs from my study and from Philips *et al.* (2010) in addition to that from Rosenberg *et al.* (2003) demonstrate that combining loci will result in improved inference of ancestry.

Within Sudan, the most private alleles and the most uniquely shared alleles observed from allelic diversity analysis were found in Zagawa, Nilotics and Nuba. This reflects a higher level of diversity in these groups compared with other Sudanese populations in this study. This observation is similar to results from the Y-Chromosome (Hassan *et al.*, 2008) where the Nilo-Saharan populations except the Nubians show a little evidence of gene flow and low migration rate with other Sudanese populations. On the other hand the observed low diversity seen in the Beja and the Nubians could result from the bidirectional migration due to the geographic locations of these populations at entering ports to Sudan.

The population structure analysis revealed that the Copts, Nubians, Arabs and Beja from Sudan cluster with the Egyptians. The Nuba, Nilotics, Zagawa, Gemar and Hausa cluster with Ugandans population. Krings *et al.* (1999) found that people from Egypt and Nubia (northern Sudan) have similar mtDNA types. These findings are also consistent with the historical evidence for long-term interactions between Egypt and Nubia which probably resulted in gene flow between these two regions. My results, in addition to mtDNA (Krings *et al.*, 1999) and Y-chromosome data (Lucotte and Mercier, 2003; Keita, 2005; Hassan *et al.*, 2008) indicate that migration, potentially bidirectional, occur along the Nile.

Even though most of the genetic variation was observed within populations and the genetic difference between linguistic and geographic groups was not great, it seems that, in Sudan,

geography plays an important role in determining differences between the groups, whereas language plays a lesser role. This result is similar to the findings of Hassan *et al.* (2008) based on a Y-chromosome study where the genetic variation among geographic groups was also higher than among linguistic groups.

The number of key migration events that took place in Africa which were described by Campbell and Tishkoff (2010), reflect the population structure results that I obtained. The studied populations from Sudan representing the Nilo-Saharan speakers (Nilotics, Zagawa, Gemar) and the Nilo-Saharan and Niger-Congo speakers (Nuba) clustered with the Karamoja population from Uganda. The Zagawa, Nuba and Nilotics showed the highest mean number of uniquely shared alleles with Ugandans, which demonstrate some level of correlation between linguistic, geographic and genetic diversity (Cavalli-Sforza, 1997). Typically these findings are in agreement with the historical findings that suggested the southeast corner of the Sudan as the homeland location for proto-Nilotic where both the Nilotes from Sudan and Uganda originated (David, 1982).

My results from the Structure analysis (for Beja, Nuba and Nilotic groups) revealed a similar pattern as that previously reported by Tishkoff *et al.* (2009) from nuclear microsatellites and insertion/deletion markers. Tishkoff *et al.* (2009) argued that clustering of these populations is primarily associated with language families, including Afro-Asiatic, Niger-Kordofanian and Nilo-Saharan.

In all my analysis, the Somali population stands out from other populations. This observation could be related to linguistic and geographic distance between the Somali and the other groups. A recent Y-chromosome study suggested that the Somali population is both of Eurasian and sub-Saharan origin (Sanchez *et al.*, 2005), potentially explaining the differentiation from African groups.

I conclude that, even though the 15 microsatellite loci from this study showed a low level of informativeness for inference of ancestry, these markers are informative enough to differentiate (to some degree) between distinct geographic and linguistic groups. Graydon *et al.* (2009) stated that "STR profiles could be used for inference of ancestry, but with varying degrees of certainty depending on the ethnicities compared". I found that the Identifiler microsatellites can be used

to detect clearly structured groups, but for fine level population structure inference, they have little power.

It is worth noting that some populations were represented by a small number of individuals and some of the populations are likely to share recent common ancestry, therefore these results do not reflect the full extent of population structure in Sudan. My future project with high density genome wide markers has the potential to detect more extensive population substructure within Sudan.

Most importantly, my results show that the combination of the 15 STR loci (CSF1P0, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D21S11, FGA, TH01, TPOX, and vWA) is a useful and powerful tool for personal identification and parentage analysis in the Sudanese population.

Finally, genetic studies on Sudanese populations have been underrepresented and mostly biased toward maternal and paternal line genetic studies. Further research should include groups not considered in this study and should incorporate more intensive genetic studies including genome-wide analysis to lead to the identification of novel variability. Moreover genetic studies should also include highly informative loci, both microsatellites (dinucleotides and trinucleotides) and SNPs, to further shed light on Sudanese population structure and to lower the genotyping efforts. I would also like to encourage improving the coverage and sample sizes of the allele frequencies listed in this study to improve their use in real case studies.

Acknowledgements

I would first like to express my sincere gratitude to Dr. Mattias Jakobsson of the Department of Evolutionary biology at Uppsala University, for giving me the opportunity to work on this project. I am grateful to you for funding this project, for your enthusiasm, your encouragement and for patiently guiding me through the various analysis stages, and for providing helpful criticism and feedback throughout the writing process. Furthermore, I would like to thank Dr. Hisham Yousef from the University of Khartoum for all the hours he spent with me discussing the sample collection plan and for helping and guiding me through the project. I would also like to thank Eva Daskalaki from the Department of Evolutionary Biology for her guidance and help in lab preparation. In addition, I would like to thank our group members, in particular Carina Schlebusch for her time in explaining different software programs and discussing and reviewing my thesis and Pontus Skoglund for his help in managing data files format. Their comments and perspectives greatly helped to make it a success. My deepest thanks should go to Mr. Sideeg Musa Elshareef, Mr. Ahmed Meirghani Suliman, Sami Sayed, Zaher Khalaf Allah and Abdullah Alhag for their great help during sample collection. I would also like to thank everyone through the blood samples collection and all the participants from Shendi, Algelaiaa, Kerma, Alburgaig Hospital, Dongola, Artigasha, Port Sudan, Khartoum, Bahri and Omdurman, who helped or donated their blood to make this work see the light. Last, but most important, I would like to send my respect and appreciation to my family and my husband, for their support and tolerance during my travelling.

References

Boehnke M, and Cox N, (1997). Accurate inference of relationships in sib-pair linkage studies, American Journal of Human Genetics, 61:423-429. (programme available at http://csg.sph.umich.edu/boehnke/relpair.php).

Brenner C, and Morris J, (1990). Paternity index calculations in single locus hypervariable DNA probes: validation and other studies, p. 21-53. In Proceedings for the International Symposium on Human Identification 1989. Promega Corporation, Madison, WI.

Butler J, (2003). For Methods in Molecular Biology: Forensic DNA Typing Protocols, Constructing STR Multiplex Assays, Humana Press, MD 20899-8311.

Butler J, (2005). Forensic DNA typing: Biology, Technology, and Genetics of STR Markers, 2nd ed., Elsevier Academic Press, New York.

Butler J, (2006). Genetics and genomics of core STR loci used in human identity testing, Journal of Forensic Science 51, 2: 253-265.

Butler J, (2009). Fundamentals of Forensic DNA typing, Burlington, Elsevier Academic Press, San Diego.

Campbell M, and Tishkoff S, (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annual Review of Genomics and Human Genetics, 9: 403-33.

Campbell M, and Tishkoff S, (2010). The evolution of human genetic and phenotypic variation in Africa, Current Biology, 20: 166-173.

Cavalli-Sforza L, (1997). Genes, peoples, and languages, Proceedings of the National Academy of Sciences USA, 94: 7719-7724.

David N, (1982). The BIEA Southern Sudan Expedition of 1979: Interpretation of the archeological data, British Institute in East Africa, 8: 51-53.

EBL 2009. Evolutionary Biology Laboratory, Genotyping preparation protocol. Available at http://www.egs.uu.se/evbiol/abi3730/prep_plates.html. Access date 16 October 2009

Ellegren H, (2004). Microsatellites: Simple Sequence with Complex Evolution, Nature, 5: 435-445.

Epstein M, Duren W, and Boehnke M, (2000). Improved inference of relationship for pairs of individuals, American Journal of Human Genetics, 67: 1219-1231.

Evanno G, Regnaut S, and Goudet J, (2005). Detecting the number of clusters of individuals using the software structure: a simulation study, Molecular Ecology, 14, 8: 2611-2620. (programme available at http://taylor0.biology.ucla.edu/struct_harvest/). Access date 6 April 2010.

Excoffier L, Laval G, and Schneideret S, (2005). Arlequin ver.3.0: An Integrated Software Package for Population Genetics Data Analysis, Evolutionary Bioinformatics Online, 1: 47–50.

Goldstein D, Linares A, Cavalli-Sforza L, and Feldman M, (1995). An Evaluation of Genetic Distances for Use With Microsatellite Loci, Genetics, 139:463-471. (programme available at http://bioinformatics.org/~tryphon/populations/).

Gomes V, Sanchez-Diz P, Alvesa C, Gomes I, Amorim A, Carracedo A, and Gusmao L, (2009). Population data defined by 15 autosomal STR loci in Karamoja population (Uganda) using AmpFℓSTR Identifiler® kit, Forensic Science International: Genetics, 3: 55–58.

Graydon M, Cholette F, and Ng L K, (2009). Infering ethnicity using 15 autosomal STR loci-Comparisons among populations of similar and distinctly different physical traits, Forensic Science International: Genetics, 3: 251–254.

Greenberg J, (1963). The languages of Africa, Bloomington: Indiana University Press.

Grun R, and Stringer C, (1991). Electron spin resonance dating and the evolution of modern humans. Archaeometry, 33:153–199.

Hassan H, Underhill P, Cavalli-Sforza L, and Ibrahim M, (2008). Y-chromosome variation among Sudanese: restricted gene flow, concordance with language, geography, and history, American Journal of Physical Anthropology, 137: 316–323.

Jakobsson M, and Rosenberg N, (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, Bioinformatics, 14: 1801–1806. (programme available at http://rosenberglab.bioinformatics.med.umich.edu/clumpp.html)

Jakobsson M, Scholz S, Scheet P, Raphael Gibbs J, VanLiere J, Fung HC, Szpiech Z, Degnan J, Wang K, Guerreiro R, Bras J, Schymick J, Hernandez D, Traynor B, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann H, Hardy J, Rosenberg N, and Singleton A, (2008). Genotype, haplotype and copy-number variation in worldwide human populations, Nature, 451: 998-1003.

Kalinowski S, (2004). Counting alleles with rarefaction: private alleles and hierarchical sampling designs, Conservation Genetics, 5: 539-543.

Keita S.O.Y, (2005). History in the interpretation of the pattern of p49a, f TaqI RFLP Y-chromosome variation in Egypt: a consideration of multiple lines of evidence, American Journal of Human Biology, 17: 559–567.

Krings M, Salem A, Bauer K, Geisert H, Malek A, Chaix L, Simon C, Welsby D, Rienzo A, Utermann G, Sajantila A, Paabo S, and Stoneking M, (1999). mtDNA Analysis of Nile River Valley Populations: A Genetic Corridor or a Barrier to Migration? American Journal of Human Genetics, 64: 1166-1176.

Lewis M, (2009). Ethnologue: Languages of the World, Sixteenth edition, Dallas, Tex.: SIL International. Online version available at http://www.ethnologue.com/. Access date 24 October 2009.

Lucotte G, and Mercier G, (2003). Brief Communication: Y-Chromosome Haplotypes in Egypt, American Journal of Physical Anthropology, 121:63–66.

McVean G, (2009). A Genealogical Interpretation of Principal Components Analysis. PLoS Genetics 5(10): e1000686. doi:10.1371/journal.pgen.1000686.

Mellars P, (2006). Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model, Proceedings of the National Academy of Sciences USA, 103: 9381-9386.

Omran G, Rutty G, and Jobling M, (2009). Genetic variation of 15 autosomal STR loci in Upper (Southern) Egyptians, Forensic Science International: Genetics, 3: 39- 44.

Philips C, Fernandez-Formoso L, Garcia-Magarinos M, Porrasa L, Tvedebrink T, Amigo J, Fondevila M, Gomez-Tato A, Alvarez-Diosd J, Freire-Aradas A, Gomez-Carballa A,. Mosquera-Miguel A, Carracedo A, and Lareua M. V, (2010) Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel, Forensic Science international doi:10.1016/j.fsigen.2010.02.003.

Phillipson D, (1993). African Archeology, Cambridge, Cambridge University Press.

Pritchard J, Stephens M, and Donnelly P, (2000). Inference of population structure using multilocus genotype data, Genetics, 155: 945-959. (programme available at http://www.stats.OX.ac.uk/~pritch/home.html).

Rosenberg N, (2004). DISTRUCT: a program for the graphical display of population structure, Molecular Ecology Notes, 4: 137–138. (programme available at http://rosenberglab.bioinformatics.med.umich.edu/distruct.html)

Rosenberg N, Li L, Ward R, and Pritchard J, (2003). Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics, 73:1402-1422.

Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L, and Feldman M, (2002). Genetic structure of human populations, Science, 298: 2381-2385.

Ruitberg C, Reeder D, and Butler J, (2001). STRBase: a short tandem repeat DNA database for the human identity testing community, Nucleic Acids Research, 29 1: 320-322

Salas A, Richards M, De la Fe T, Lareu M V, Sobrino B, Sanchez-Diz P, Macaulay V, and Carracedo A, (2002). The making of the African mtDNA landscape, American Journal of Human Genetics, 71:1082–1111.

Sanchez J, Hallenberg C, Børsting C, Hernandez A, and Morling N, (2005). High frequencies of Y chromosome lineages characterized by E3b1, DYS19-11, DYS392-12 in Somali males, European Journal of Human Genetics, 13: 856–866

Slatkin M, (1995). A Measure of Population Subdivision on Microsatellite Allele Frequencies. Genetics, 139:457-462.

Szpiech Z, Jakobsson M, and Rosenberg N, (2008). ADZE: a rarefaction approach for counting alleles private to combinations of populations, Bioinformatics, 24: 2498–2504. (programme available at http://rosenberglab.bioinformatics.med.umich.edu/adze.html)

Tillmar A, Backstrom G, and Montelius K, (2009). Genetic variation of 15 autosomal STR loci in a Somali population, Forensic Science International: Genetics, 4: 19-20.

Tishkoff S, and Williams S, (2002). Genetic Analysis of African Populations: Human Evolution and Complex Disease, Nature Review Genetics, 3: 611-621.

Tishkoff S, Reed F, Friedlaender F, Ehret C, Ranciaro A, Froment A, Hirbo J, Awomoyi A, Bodo J, Doumbo O, Ibrahim M, Juma A, Kotze A, Lema G, Moore J, Mortensen H, Nyambo T, Omar S, Powell K, Pretorius G, Smith M, Thera M, Wambebe C, Weber J, and Williams S, (2009). The genetic structure and history of Africans and African Americans, Science, 324: 1035 – 1044.

Appendix A

Sample Size and linguistic affiliation of the populations in the study

Ethnic Groups			Linguistic Affiliation	
Group	Subgroup	N	Family	Level
Arabs				
	Bataheen	29	Afro-Asiatic	Semetic
	Gaalien	57	Afro-Asiatic	Semetic
	Shaigia	17	Afro-Asiatic	Semetic
	Messiria	8	Afro-Asiatic	Semetic
Copts		31	Afro-Asiatic	Ancient Egyptian
Hausa		10	Afro-Asiatic	Chadic
Beja				
	Beni-Amer	35	Afro-Asiatic	Cushitic
	Hadendowa	29	Afro-Asiatic	Cushitic
Nubians				
	Danagla	40	Nilo-Saharan	Eastern Sudanic
	Mahas	31	Nilo-Saharan	Eastern Sudanic
	Halfawieen	24	Nilo-Saharan	Eastern Sudanic
Nilotics				
	Dinka	30	Nilo-Saharan	Eastern Sudanic
	Nuer	19	Nilo-Saharan	Eastern Sudanic
	Shilluk	19	Nilo-Saharan	Eastern Sudanic
	Bari	6	Nilo-Saharan	Eastern Sudanic
Nuba		34	Nilo-Saharan + Niger-Congo	Eastern Sudanic + Kordofanian
Zagawa		29	Nilo-Saharan	Saharan
Gemar		6	Nilo-Saharan	Eastern Sudanic

Appendix B

Evolutionary Biology Center

Uppsala University

Sudanese Genetic Diversity Project 2009-2010

Sample Collection Form

Name: …….

Age: ……

Sex: …….....

ID

Ethnic Origin:-

Father: ……. G. mother: ……..

Mother: …… G. mother: ……..

Locality/City: …….....

Type of sample(s): ……..

Name of researcher: Hiba Babiker

Date of collection: / /

Appendix C

Consent Form for Research Study

استمارة موافقة للمشاركة في دراسة بحثية

Genetic Patterns of Autosomal Short Tandem Repeats among Sudanese population

دراسة الأنماط الجينية للتكرارات الترادفية القصيرة بين السودانيين

By signing this form, I agree that:

The study has been explained to me. Possible harm and discomforts and possible benefits (if any) of this study have been explained to me. I understand that I have the right not to participate and the right to stop at any time. I understand that I may refuse to participate. The only very slight risk to me is that there might be the possibility of discomfort at the site where blood is drawn. These discomforts are brief and transient. I understand that I have a choice of not answering any specific questions. I have been told that my personal information will be kept confidential and I understand that no information that would identify me will be released or printed.

من خلال التوقيع على هذه الاستمارة، أوافق على ما يلي :

تم توضيح وشرح أهداف هذه الدراسة لي، وتمت الإجابة على جميع تساؤلاتي. كما تم توضيح الضرر والفوائد الممكنة (إن وجدت) من هذه الدراسة. وأن لي الحق في عدم المشاركة، والحق في التوقف في أي وقت يتم إخطاري به، مع إنه ربما تواجهني بعض الآلام الطفيفة في موضع أخذ العينة. وإن لدي الخيار في الإجابة كما أن جميع المعلومات الشخصية المتعلقة بي سيتم التعامل معها بسرية ولن تتم طباعتها أو نشرها.

If you agree to be in this study, you will give a blood sample 500 µl. The blood will be taken from the finger using a lancet. This will take about 3 minutes.

اذا كنت توافق على أن تكون ضمن هذه الدراسة فسوف تتبرع بعينة دم (500 ميكرولتر) ستؤخذ من الأصبع بواسطة إبرة صغيرة. وستستغرق عملية أخذ العينة حوالي ثلاث دقائق.

I hereby consent to participate.

فأنا أوافق على المشاركة.

____

Signature   Date

التوقيع   التاريخ

Name and Age of the Participant: ___________________________________

اسم وعمر المشارك

Name of person who supervised the filling of the form: ___________________________

اسم الشخص المشرف على ملأ الاستمارة

____

Signature   Date

التوقيع   التاريخ

To whom correspondence should be sent:

Dr. Mattias Jakobsson
Dept. Evolutionary Biology, Uppsala University
Norbyv. 18D
SE-752 36 Uppsala
Sweden

للاستفسار والمراسلة :

هبة محمد علي بابكر
وزارة الداخلية الأدلة الجنائية
إدارة المختبرات الجنائية
ص. ب. 981 الخرطوم   ت: 0912975498
السودان