Analysis of 55 Kidd Ancestry SNP of Qatari population using Structure & ForenSeq Universal Software

Almohammed. E [1*,2 *];Hadi. S[1]

[1]*University of Central Lancashire, School of Forensics and Applied Sciences, Preston, UK*

[2]*Ministry of Interior of Qatar, Doha, Qatar;*

**Corresponding author:*

eida-k-al@hotmail.com

Abstract

SNPs are good predictors of ethnicity and serval panels have been published (1). The ForenSeq Signature kit (Illumina) offers coverage of 230 different markers including 55 ancestry SNPs (AISNPS). The ForenSeq Universal Analysis Software (UAS) brings the capability to analyse the sequencing data, represent results and accomplish statistical estimates of biogeographic ancestry. The ancestry prediction results in UAS are based on Principal Component Analysis (PCA) collected on several reference populations comprised in the 1000 Genomes project. This set does not include Qatari population. The data was analysed using STRUCTURE software. These data serves as an addition to the existing Middle Eastern population data for the 55 AISNPS. The MPS kit included 55 Ancestry Informative Marker SNPs (AISNPs). The Qatari population has been a melting pot of various populations and this forensic study was the first of its kind to generate new data on the genetics of Qatari population. The 55-ancestry marker allele frequency data for 150 samples were analysed and compared to 139 world populations using STRUCTURE software. The Qatari population data for the Kidd AISNPs were found to be similar in patterns when compared to other Middle Eastern populations. FROG-kb analysis led to all Qatari samples being correctly assigned to Middle Eastern populations showing that the Illumina software needs to be enhanced by adding more databases from the Middle East. In conclusion, the results have clearly demonstrated the potential use of MPS methods to study the genetics of Qatari population.

Keywords: Ancestry SNPS; ForenSeq™ kit; MPS; Qatar.

1. Introduction

2. Qatar State

The Qatari population comprises of than 2.5 million populations, included of ~300,000 nationals with different origins (Qatar Statistics Authority 2018, www.qsa.gov.qa). Based on family names and oral history, it is believed that the majority of the Qatari population originated from the Arabian Peninsula (2), (3). Rodriguez-Flores et al. (2016), also described that the Qatari population can be divided into at least three genetic subpopulations that can reveal the historical migration patterns in the region: Bedouin, Persian-South Asian and African (4), (2). A number of 48 SNPs were genotyped previously by for classification into one of the three subpopulations completed on >70% ancestry in one cluster in a STRUCTURE analysis for identify individuals (2). It has shown primary attention was the for the Bedouin genetic subpopulation because of its deepest ancestry in Arabia. The deepest ancestral roots in the Peninsula are the Bedouin subpopulation, who still practises tribal marriages based on the Arabic traditions, which in turn might affect the level of homozygosity (3).

The HGDP set of 52 populations is the commonly used group of reference samples totalling 1052 individuals. This shows that (3%) of the 1397 SNPs can happen in different panels (7). Soundararajan et al. (2016) noted that the 128 AISNPs from the Seldin and Kidd set of 55 AISNPs at present have the largest database of reference populations with a major world regions. In this study, 140 population data of Pakstis et al. (2017) were re-analysed after inclusion of the Qatari population data for 55 Kidd panel AISNPs generated from ForenSeq™ Signature kit. The aim of this study was to analyse the 55 Kidd AISNPs in Qatari population in order to understand the ancestry of this population in a better way.

3. Materials and Methods

In this study, a total number of 150 Qatari population data for 55 Kidd AISNPs generated from ForenSeq™ Signature kit and used in this study. The STRUCTURE 2.3.4 software provided unique method of evaluating how good a set of loci tested on several individual could infer ancestry. STRUCTURE software is a freely accessible programme for population analysis established by Pritchard et al. (2000). It used a systematic Bayesian method applying Markov Chain Monte Carlo (MCMC) evaluation could begin by randomly assigning samples to a pre-determined number of groups. The K value can indicate likely number of clusters as estimated as log probability of data Pr (X|K) mentioned in Pritchard et al. (2000). In order to assume correlated allele frequencies, the admixture model was more appropriate for this study and this was applied using K values that were altered from 6 to 11 and the 20 runs.

4. Results

The results of this study revealed that was no significant (P>0.05) deviations from (HW) Hardy-Weinberg observed on those expected SNP that were described in the ALFRED database. Allele frequencies for the whole 55 SNPs in all 140-population samples were mentioned in ALFRED. Added populations have been studied in scientific publications for some of the SNPs. Therefore, some SNPs have data on more than 139 populations in ALFRED. The Qatari population was uploaded to ALFRED recently. The set of 55 AISNPs in FROG-kb, has an extensive allele frequency data for all 140-reference population samples. The completeness of the data permits the likelihoods ratios to be calculated for all of these 140 populations for DNA profile for the 55 AISNPs. The STRUCTURE results were reported for assessed cluster membership values at K = 6, 7, 8, 9, 10 and 11 in 140 reference populations. The results for the STRUCTURE analysis are displayed in Fig 1. The data showed the maximum likelihood run of the most frequently occurring (for 10 of 20 runs) cluster pattern at K = 9. This study comprises 140 reference populations with 8,184 individuals including Qatari new population with the rest of the population reported previously by Pakstis et al. (2017). Of the 8,184 individuals analysed, 72.3% had all 55 AISNP genotypes currently; 91.0% of individuals had no more than three missing profiling and 1.9% of the possible genotypes were missing. Among the 129 individuals included, there were 3.2% of 7095 missing. The new Qatari population data generally showed a very strong similarity to previously analysed populations of the Middle Eastern population who were well known to be closely related in the geographical location (2). The overall pattern of the results showed that Qatari population was similar to surrounding groups like the Kuwaiti, U.A.E, and Saudi as shown in Fig 1(A). A more focused STRUCTURE analysis was also done for reduced number of population data using sixty-nine data set by omitting the 25 populations from East Asia, Pacific and Americas. With the additional analysis, using less number of population, the STRUCTURE software could allow for further investigations if any new cluster patterns could be recognized within North Africa. The nearby regions remaining in the analysis did not emerge in the previous 139-population analysis because of the distraction provided by the strong clusters in East Asia, Pacific and Americas (6). The individual bar plots for the most frequently occurring cluster at each K level from was from 6 to 7 (Fig 1B). The results show that the most common cluster pattern showing the highest likelihood run in this more focused STRUCTURE analysis. The cluster pattern observed for North African, Southwest Asian populations and sub-Saharan at K = 10, k =11 in this analysis focused on 140 populations. This was similar to what was reported in earlier publication of Paskstis et al. (2017).

A total number of three Qatari individual 55 AISNP genotypes generated from ForenSeq™ Signature kit, were analysed in FROG-kb for the 55 AISNP panel. The Qatari population samples from which the individuals were compared with the population data previously were used for FROG-kb calculations. The likelihood ratio was estimated as the probability of the greatest population divided by the probability of the specified population. The highest ranked likelihood results for the top 15 populations were within one order of magnitude of each other. A total number of 45 populations was listed by the probability calculated by FROG-kb for the three Qatari individuals. Using the rule of likelihood, the population with the utmost probability of this genotype becomes the most likely population of the closest origin and

likelihood ratios indicated how much more likely the best population was compared to other population. Table (1C) showed that the ratio of 100 or more in the populations is significantly less likely to be the ancestry of the sample.

For the Qatari individuals, the populations within one order of magnitude were mostly from the Mediterranean area. The high-ranking populations for the Qatari population individuals were from the Middle Eastern, Africa and South Asian regions. Tunisian individuals clustered together or with populations from the nearby Middle East (1). Thus, populations with non-significant likelihood ratios for ancestry assignment were more specific than using only the single most likely population. The Qatari populations have similar geographic relatedness with the origin of Middle Eastern population and much more similar to the groups of UAE, Kuwaiti and Saudi Arabian populations. FROG-kb for individual ancestry assignment at the population levels was found to be quite precise in assignment the samples to relevant population groups. The results from FROG-kb provided relative likelihoods of ancestry from different populations for user-specified genotypes. Such data are probably in principle and these results should not be taken as absolute values. The Qatari population assignment data on the STRUCTURE results revealed that the population of origin was most likely to be very close to the Middle Eastern related populations among the reference populations when compared on FROG-kb. The ForenSeq™ Universal Software did not include any Middle Eastern populations so the ancestry of these samples showed that the Qatari individuals clustered with the American Admixed origin, which was not the correct result. The software needs to be updated with a more comprehensive reference database with bigger number of populations/samples. For forensic applications, it is necessary to have an accurate biogeographic ancestry assignment and for this sufficient number of AISNPs need to be used and a reasonable reference data on the appropriate and suitable populations is required.

5. Discussion

The current study has shown that the Qatari populations have similar cluster pattern equally to the geographic relatedness with the origin of Middle Eastern population. These are rather more similar to the groups of UAE, Kuwaiti and Saudi Arabia. The Qatari population assignment data on STRUCTURE and Frog-kb revealed that the population of origin was most closely linked to Middle Eastern region. FROG-kb for sample origin assignment at the population levels was more accurate than ForenSeq™ software using the same loci. The results from FROG-kb provided likelihoods of ancestry from different populations for user-specified genotypes. In the current study, the available 140 populations, including the Qatari population show similar results when compared to previously published STRUCTURE result and similar to the latest publication by (6). Regarding the literatures, the history of Qatar suggested the Qatari populations could be divided into many groups (3), (4). These include the group of Arab origin descendants of the Bedouin, a second that has strong affinity with Persian and Central Asia, and a third that has strong attraction with African ancestry. This proposes that, like the Qatari population, the Saudi, UAE and Kuwaiti population showed a various array of genetic contributions following centuries of active trade and this is not simply a relic of the ancient out-of-Africa migration. A total number of 55 AISNPs results for the Qatari population demonstrated that this set has enough power to designate the samples to the correct population though in a few cases, the top assigned populations were not currently Middle Eastern populations. The obvious reason is population migrations and variable ancestral origin. Historical patterns of migration into the Qatar Peninsula from the subcontinent, especially from the coastal regions affected the Qatari ancestry estimations. With increase in number of populations studied for the 55 AISNPs (5), it has clarified that the database needed to be enhanced and the current study contributed to a new dataset which would add more reliability to the analytical results for Middle Eastern populations. The more reference populations used and added, especially from numerous relatively disregarded geographical regions and smaller ethnic groups in better-studied areas, needs to be considered. There are chances to develop and fine-tune the best AISNP panels for specific geographical

regions that can be developed further in future. The existing panels for ancestry estimations can be improved through such studies.

5. Conclusion

In conclusion, these results, in comparison to the previous findings, reveal the value of more populations studied for a slight number of informative SNPs. Further AISNPs can be recognised in the upcoming, particularly those with a high frequency variant of restricted geographic extent. Such SNPs may be inefficient for a global analysis of ancestry nevertheless very valuable for refined analysis of ancestry within a biogeographic region of the world such as Middle East.

9. References

1. CHERNI, L., PAKSTIS, A. J., BOUSSETTA, S., ELKAMEL, S., FRIGI, S., KHODJET-EL-KHIL, H., BARTON, A., HAIGH, E., SPEED, W. C., BEN AMMAR ELGAAIED, A., KIDD, J. R. & KIDD, K. K. 2016. Genetic variation in Tunisia in the context of human diversity worldwide. *American Journal of Physical Anthropology,* 161**,** 62-71.
2. Rodriguez-Flores, J.L., Fakhro, K., Agosto-Perez, F., Ramstetter, M.D., Arbiza, L., Vincent, T.L., Robay, A., Malek, J.A., Suhre, K., Chouchane, L. and Badii, R., 2016. Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations. Genome research, 26(2), pp.151-162.
3. HUNTER-ZINCK, H., MUSHAROFF, S., SALIT, J., AL-ALI, K. A., CHOUCHANE, L., GOHAR, A., MATTHEWS, R., BUTLER, M. W., FULLER, J., HACKETT, N. R., CRYSTAL, R. G. & CLARK, A. G. 2010. Population Genetic Structure of the People of Qatar. *The American Journal of Human Genetics,* 87**,** 17-25.
4. OMBERG, L., SALIT, J., HACKETT, N., FULLER, J., MATTHEW, R., CHOUCHANE, L., RODRIGUEZ-FLORES, J. L., BUSTAMANTE, C., CRYSTAL, R. G. & MEZEY, J. G. 2012. Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. *BMC genetics,* 13**,** 1.
5. PAKSTIS, A. J., HAIGH, E., CHERNI, L., ELGAAIED, A. B. A., BARTON, A., EVSANAA, B., TOGTOKH, A., BRISSENDEN, J., ROSCOE, J., BULBUL, O., FILOGLU, G., GURKAN, C., MEIKLEJOHN, K. A., ROBERTSON, J. M., LI, C.-X., WEI, Y.-L., LI, H., SOUNDARARAJAN, U., RAJEEVAN, H., KIDD, J. R. & KIDD, K. K. 2015. 52 additional reference population samples for the 55 AISNP panel. *Forensic Science International: Genetics,* 19**,** 269-271.
6. PAKSTIS, A. J., KANG, L., LIU, L., ZHANG, Z., JIN, T., GRIGORENKO, E. L., WENDT, F. R., BUDOWLE, B., HADI, S., AL QAHTANI, M. S., MORLING, N., MOGENSEN, H. S., THEMUDO, G. E., SOUNDARARAJAN, U., RAJEEVAN, H., KIDD, J. R. & KIDD, K. K. 2017. Increasing the reference populations for the 55 AISNP panel: the need and benefits. *International Journal of Legal Medicine,* 131**,** 913-917.
7. SOUNDARARAJAN, U., YUN, L., SHI, M. & KIDD, K. K. 2016. Minimal SNP overlap among multiple panels of ancestry informative markers argues for more international collaboration. *Forensic Science International: Genetics,* 23**,** 25-32.