# The study of 95 identity SNPs for Qatari population using massively parallel sequencing (MPS)

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

[1]Ministry of Interior of Qatar, Doha, Qatar; [2]School of Forensic and Applied Sciences, University of Central Lancashire, Preston, UK

*Corresponding author:* eida-k-al@hotmail.com

## ABSTRACT

For the last three decades, Short Tandem Repeat (STR) markers and capillary electrophoresis-based DNA sequencers have been the gold standard technology for human identification testing in the forensic field. However, Massively Parallel Sequencing (MPS) has the potential to replace the current CE-based technology. All laboratories globally, including the Qatar Forensic Laboratory, have stringent strategies and regulations in improving and processing high numbers of reference and casework samples, using instrumentation and technologies to maximize output using new STR kits consisting of larger number of loci. MPS technology has enabled sequencing of several types of genetic loci in one multiplex including single nucleotide polymorphisms (SNPs) and STRs. One hundred and fifty (150) reference samples were profiled using the ForenSeq$^{TM}$ DNA Signature kit. The average read depth was 20,000 reads for all sequencing runs. The data generated through MPS were analysed using ForenSeq$^{TM}$ Universal software and STRait Razor V3 for the primary, secondary and tertiary analyses. The analyses of the sequence of alleles in STRait Razor software were able to determine novel alleles in the Identntiy SNPs loci. 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 95 identity SNPS allele frequency data for 150 samples were analysed. In conclusion, the results have clearly demonstrated the potential use of MPS methods to study the genetics of Qatari population.

Keywords: Qatari population; MPS; iSNPs; Forensic

## 1.Introduction

### 1.1 Identity SNPs (iSNPs) using Massively Parallel Sequencing (MPS)

Short tandem repeat (STR) analysis using capillary electrophoresis (CE) has been the gold standard method in forensic genetics (4). Advancements in Massively Parallel Sequencing (MPS) led to the development of the ForenSeq™ kit by Illumina (6). The kit enables MPS of 230 genetic markers including STRs and SNP in the same multiplex library preparation run. SNP genotyping methods are amenable to high-throughput sample processing and automated data analysis. Due to their shorter locus size, the PCR amplicons are also shorter, and this assists their ability to be multiplexed and make them suitable target loci in highly degraded sample profiling. In most cases, interpretation of SNP profiles is simple due to the absence of certain artefacts common with STR systems such as the stutters. Recent work has investigated the utility of SNPs as a superior set of SNP markers for use with extremely degraded samples. A number of SNPs were also identified as identity SNPs as these had high heterozygosity within human populations (7). The most significant advantage SNPs can offer is to identity testing with highly degraded remains. The amplicon size of 45-55 bp, for SNPs is smaller than the required length for STR loci. The reduction in amplicon size can provide the maximal number of loci from highly degraded samples. The best SNPs for identity testing are those with the highest heterozygosity and low Fst values, as fewer loci will be needed to reach high levels of discrimination. There are numerous SNP panels which are designed as potential for identification of individuals and they include 52 SNP panel and 95 identity iSNPs which is included in Signature DNA kit (5). This is the first study for the Qatari population using MPS methods and the results would be subsequently useful for casework purposes in the Qatar forensic laboratories.

## 2. Materials

### 2.1 Collection of Qatari Population Samples

A total of 150 saliva samples were collected from unrelated indigenous Qatari males from different origins in Qatar. All samples used in this study were consented for the purpose of the research and were obtained from the Ministry of Interior of Qatar. Ethical approval was granted for PhD project sponsored by Ministry of Interior of Qatar and provided by University of Central Lancashire “Ref STEM454”.

### 2.2 DNA Extraction & Quantification

The samples were extracted using the QIAamp® extraction DNA Mini protocol (Qiagen Ltd, West Sussex, UK) in accordance with Manufacturer’s recommendation. The quantification of the collected samples was carried out using the Quantifiler® Trio DNA Quantification kit (ThermoFisher Scientific) according to the Manufacturer’s instructions.

### 2.3 ForenSeq™ DNA Signature Prep Kit (Beta version)

The beta version of the ForenSeq™ kit provided PCR primer mixes for the targeted amplification of STRs (27 autosomal STRs, 24-Y-STRs, and 9 X-STRs) and 95 identity informative SNPs (iSNPs). Library Generation process allowed simultaneous preparation of 96 samples to generate libraries of PCR products within a single plate. Each library was a collection of amplified DNA fragments from a single sample. The Human Sequencing Control was used to determine running completion and to highlight possible sequencing issues. Massively parallel sequencing and data analysis were performed on the MiSeq FGx™ desktop sequencer (Illumina) with a ForenSeq™ sequencing kit (351-31bp) (Illumina) according to the Manufacturer’s has recommended protocol. Data analysis was performed using the ForenSeq™ Universal Analysis Software (UAS) and STRait Razor v3.0 (5). DoC, also known as read depth, and ACRs, i.e., heterozygote balance, were calculated for each identity informative SNP locus. ACRs were calculated by dividing the lower coverage allele by the higher coverage allele at that locus. SCRs used the bp sequence of the STR alleles versus the repeat structure of the alleles. The ForenSeq Universal Analysis Software determined the length of the STR sequence and the number of repeats within that sequence and aligned to the hg19 human reference sequences that corresponded to the loci. Read numbers below 10 were not displayed by UAS software using default settings. ISNPs were typed above a threshold of 30 reads. STRait Razor v3.0 has been used to identify the STR allele sequences and has led to the identification of novel alleles (9,5). The nomenclature used herein was based on practices in the literature but used standards for variant allele naming likely to be established by the International Society of Forensic Genetics (ISFG), especially when novel variants were observed where permutations might exist for allele nomenclature (7).

## 3. Results

### 3.1 Primary and Secondary Analysis Using Universal Analysis Software (UAS)

The sample representation and the quality metrics were developed using the UAS for all the three runs. The cluster pass filter (PFs) represented the filters the clusters that did not attain the standard quality scores to accurately quantitate the fluorescence signal by the lasers. The cluster PFS were: run 1 (≥87.13%), run 2 (≥89.67%) and run 3 (≥89.67%). Phasing and pre-phasing values indicate the percentage of molecules that fall one base behind (phasing) (≤0.166%) or ahead (pre-phasing) (≤0.089%) in each cycle. The majority of loci amplified were <150 bp in fragment size. Of all the markers amplified the highest proportion present were 100% for pSNPs, 92.5 % for auSTRs, 85.7 % X STRs, 75% iSNPs and 94.6% AISNPs when the data for all runs was combined. Qatari samples showed a high allele recovery for all STRs and SNPs loci. In this study, STRait Razor v3.0 was used in the analysis to identify both the alleles and the novel alleles (5,6). It calculates the depth of coverage by summing all effective reads within the loci. It is designed to detect forensically relevant STR alleles in FASTQ sequence data, based on allelic length. The software also calculated the Allele Coverage ratio (ACRs) for auSTRs, which is a ratio of the coverage of two alleles in heterozygote individual.

### 3.1.1 Tertiary analysis using ForenSeq™ Universal Analysis Software and STRait Razor Software

The analysis produced an average DoC values were ranging from 1209 - 9839 for SNPs. For auSTRs in the Qatari population, DoC was above 400,000 reads. The highest DoC for iSNP loci was for six loci (rs1109037 showing rs140313; rs192655; 14031320, rs4364205 and rs8037429) with a 140, 220 coverage depth. The rest of SNPs showed a DoC below were below 40,000. The Allele coverage ratio (ACR) was calculated by UAS as a ratio of the number of reads for two alleles in heterozygous loci (Lower coverage/Higher Coverage). For SNP loci, ACR is equivalent to the ratio of reference allele coverage to reference and alternative allele coverage. The average ACR values obtained were 0.83 ± 0.069 for SNPs. Overall, the ACRs were quite even across the loci. Across, all samples, between 90 and 100% of the heterozygous loci, showed allelic balance ratios of ~ 0.5. This study identified several novel identity SNPS were identified. The novel variants were compared to the human genome reference sequence (GRCh38) (Supplementary Materials Table 1) (8, 10). The repeat pattern variants were compared to the existing database published literature and the database STRBase (strbase.nist.gov) [Accessed Date: 13-09-2019]. Allele frequencies for 94 iSNPs were calculated (Fig 1). The data showed that most frequencies were above 0.5 indicating the forensic value of the markers. AuSTR and iSNP RMPs for both length match probability and sequence match probabilities were calculated using STRait Razor software (Fig 1A). For auSTRs, these values were $2.21\times10^{-31}$ and $1.29\times10^{-35}$. For iSNPs, these were $9.5\times10^{-38}$ and $8.7\times10^{-41}$. Loci rs2056277, rs722098, rs97118 and rs938283 showed the highest increases in RMP due to the inclusion of sequence-based (SB) alleles in the Qatari populations. Such data have not been reported earlier. The locus rs2056277 was previously reported by Poulsen et al. (2011) to be relatively uninformative for identity testing purposes. The iSNP full profile RMP was calculated to be $1.08 \times 10^{-32}$. STRait Razor was used to generate the combined MPs for auSTRs and iSNPs and the combined match probabilities for Length-Based (LB) and Sequence-Based (SB) alleles were $2.9\times10^{-68}$ and $1.12\times10^{-75}$.

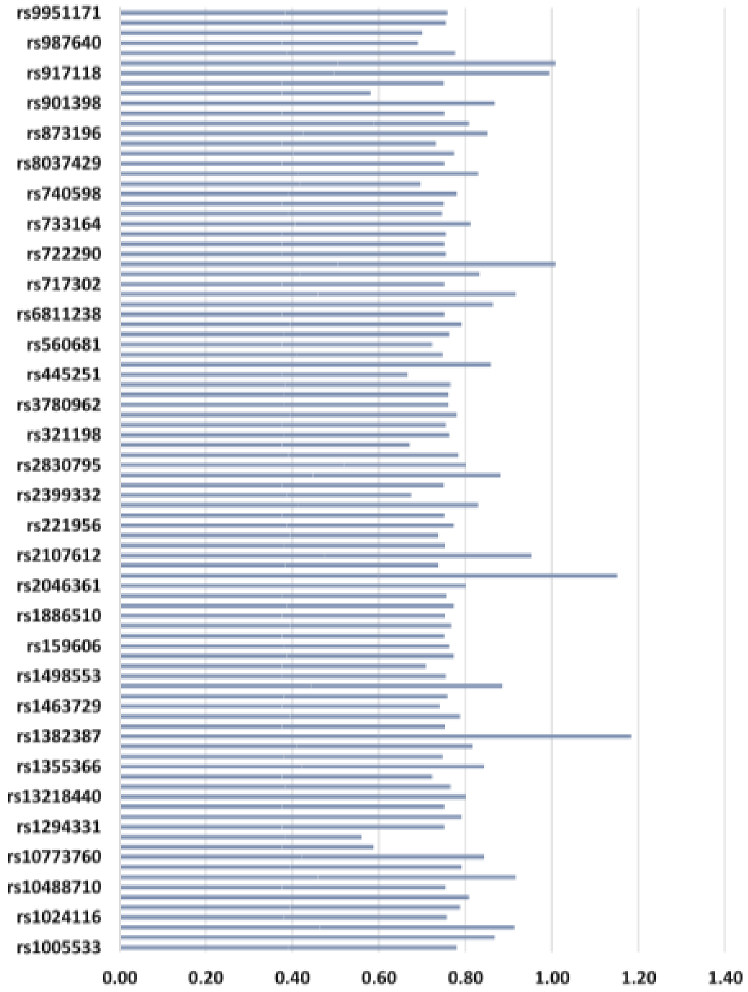

Figure 1. (A) Bar chart showing the Random Match Probability (RMP) increase due to Sequence-Based Alleles. These were detected in Massively Parallel Sequencing of identity SNPs (The Combined Match Probability = 2.9×10$^{-68}$ and 1.12×10$^{-75}$).

## 4. Discussion

MPS technology is either standalone or additional technology which has the potential to overcome the limitations of CE-based STR typing. Furthermore, MPS enables the simultaneous analysis of a large number of various types of markers. The Qatari population had not been studied before using the MPS technology and the ForenSeq™ DNA Signature kit allowing for the collection of population data for the markers which are included in the kit besides the determination of several new sequence-based alleles. Massively parallel sequencing (MPS) technology has a great potential to become an alternate or additional method to overcome the limitations of CE-based STR typing as MPS does not require size separation between amplicons and therefore enables the simultaneous analysis of a large number of markers. In addition, the application of MPS technology to STR analysis can provide data on sequence variation. There allowing for larger number of alleles due to sequenced differences within same sized alleles and consequently remarkably higher powers of discrimination. This study has shown similar trends for STRs and SNP markers in the Qatari population. The interpretation threshold for most loci was >4.5% and the analytical threshold, which distinguishes signal from noise, was kept at the default level of >1.5%. In an early study on the use of MPS for STR sequencing by Bornman et al. (2012), it was demonstrated that a >99% confidence in allele designation can be estimated for DoC of 18,500 reads using the limited 150 base single read sequences on the Illumina MiSeq FGx platform

(1). The obtained results of this study showed that AutSTRs and SNPs used in the ForenSeq™ Signature kit were very informative for forensic purposes in Qatari population. The length-based combined match probability of $2.9\times10^{-68}$ and sequence combined match probability of $1.12\times10^{-75}$ was obtained for auSTR and ID SNPs for the Qatari population. It has been demonstrated that MPS can reveal micro-heterogeneity of STR alleles (2, 4). The possibility to detect various STRs and SNPs in a single sequencing run allowed for immense increase in discrimination power from DNA recovered from crime scenes. These findings also demonstrated the value of employing MPS for forensic casework. Observed and expected heterozygosity were determined for the Identity SNPs. All loci were highly polymorphic; as expected, the heterozygosity increased at those loci. Tests for Hardy-Weinberg equilibrium were performed separately for LB alleles for 95 Identity SNPs in Qatari population.

## 5. Conclusion

This study serves as the first investigation into Qatari population genetics with respect to forensically-relevant loci as well as the first set of population data reported using the ForenSeq™ DNA Signature Prep kit. Identity SNP allele frequencies have been reported for 150 Qatari samples. These findings demonstrate the value of increased heterozygosity and discrimination potential due. This new platform will provide additional discrimination power to STRs routinely used in forensic genetics considering its ability to reveal intra-repeat sequence variants. The possibility to detect 59 autosomal STRs together with 94 Identity SNPs in a single sequencing run, allows for as better DNA recovery from a crime scene. This current study serves as the first investigation into the Qatari population with respect to forensically-relevant loci as well as the first set of population data reported using the ForenSeq™ DNA Signature Prep kit. In this study, STR and Identity SNP allele frequencies have been reported for 150 Qatari samples. The data presented in this study produce combined STR and SNP RMPs of the combined match probability of $2.9\times10^{-68}$ and $1.12\times10^{-75}$ for length-based and sequence-based autosomal STR alleles respectively. The magnitude of these RMP values highlights the power of a combined STR and SNP approach towards source attribution in forensic DNA typing.

## 6. Acknowledgements

I would like to thank the Qatar Ministry of Interior who supported my PhD project. A special thanks to Qatar Embassy (Military and Cultural Attaché office) for their help and support during my research project

## 7. References

1. Bornman, D.M., Hester, M.E., Schuetter, J.M., Kasoji, M.D., Minard-Smith, A., Barden, C.A., Nelson, S.C., Godbold, G.D., Baker, C.H., Yang, B. and Walther, J.E., 2012. Short-read, high-throughput sequencing technology for STR genotyping. *BioTechniques. Rapid dispatches, 2012,* p.1.

2. BØRSTING, C. & MORLING, N. 2015. Next generation sequencing and its applications in forensic genetics. *Forensic Science International: Genetics,* 18**,** 78-89.

3. BUERMANS, H. P. J. & DEN DUNNEN, J. T. 2014. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease,* 1842**,** 1932-1941.

4. CARATTI, S., TURRINA, S., FERRIAN, M., COSENTINO, E. & DE LEO, D. 2015. MiSeq FGx sequencing system: A new platform for forensic genetics. *Forensic Science International: Genetics Supplement Series,* 5**,** e98-e100.

5. KING, J. L., WENDT, F. R., SUN, J. & BUDOWLE, B. 2017. STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems. *Forensic Science International: Genetics,* 29**,** 21-28.

6. MISEQFGX™MANUAL 2015. *MiSeq FGx™ Instrument Reference Guide.*www.illumina.com/systems/miseq-fgx.html.

7. PARSON, W., BALLARD, D., BUDOWLE, B., BUTLER, J. M., GETTINGS, K. B., GILL, P., GUSMÃO, L., HARES, D. R., IRWIN, J. A., KING, J. L., KNIJFF, P. D., MORLING, N., PRINZ, M., SCHNEIDER, P. M., NESTE, C. V., WILLUWEIT, S. & PHILLIPS, C. 2016. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Science International: Genetics,* 22**,** 54-63.

8. Phillips, C., Gettings, K.B., King, J.L., Ballard, D., Bodner, M., Borsuk, L. and Parson, W., 2018. “The devil’s in the detail”: Release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide. *Forensic Science International: Genetics*, *34*, pp.162-169.

9. WARSHAUER, D. H., KING, J. L. & BUDOWLE, B. 2015. STRait Razor v2.0: The improved STR Allele Identification Tool – Razor. *Forensic Science International: Genetics,* 14**,** 182-186.

10. Young, B., Faris, T. and Armogida, L., 2019. A Nomenclature for Sequence-Based Forensic DNA Analysis. *Forensic Science International: Genetics*.