High-throughput sequencing of forensic STRs and SNPs using the MiSeq benchtop sequencer

C. Xavier[1] and W. Parson[1, 2]
[1]Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria
[2] Forensic Science Program, The Pennsylvania State University, PA, USA
catarina.gomes@i-med.ac.at

Introduction

Next Generation Sequencing (NGS) has been increasing its applications in forensic genetics in the last couple of years. Such methods bring a high multiplexing capacity and a deep sequencing resolution that put data under a new perspective. After proving its capacity with mitochondrial DNA and SNP markers [1-3], new assays comprising length polymorphisms are currently entering the market. Testing and evaluating such applications is then of utmost importance for validation and possible integration into routine casework laboratories.

Therefore, the results of a developmental validation study are presented here. We tested Illumina's ForenSeq DNA Signature Prep Kit performance in several assays as reproducibility, concordance, sensitivity, mixture deconvolution and difficult samples. This kit presents two different primer multiplexes to answer different needs in current forensics casework, both comprising autosomal, Y-chromosome and X-chromosome STRs and different panels of SNPs (**Fig.1**). Such design prevents the usage of different STR kits on the same sample by allowing the analysis of multiple markers in one reaction. The deep sequencing resolution will also contribute to a better description of STR alleles and consequently increase the discrimination power and deepen the mixture deconvolution capacity of these markers.

Materials & Methods

The ForenSeq DNA Signature protocol follows a double PCR design for library preparation (**Fig.1**). An amplification PCR is performed to target specific regions of the genomic DNA and tags the target regions, then an enrichment PCR inserts the specific indices per library. After a purification step, a bead-based normalization protocol provides an equimolar sample pooling. All protocols were followed according to the instructions of the manufacturer [4].

Two Illumina MiSeq runs (1X351, 1X31) were performed to cover all assays. The reproducibility and sensitivity of the method were evaluated by a dilution series starting at 1ng down to 50pg of input 2800M control DNA. 20 buccal swap samples were used for concordance and mixtures (1:1, 1:10 and 1:20 ratios) studies and GEDNAP, DNase degraded and aDNA samples [5] were used as casework type samples. All donors have signed an informed consent declaration.

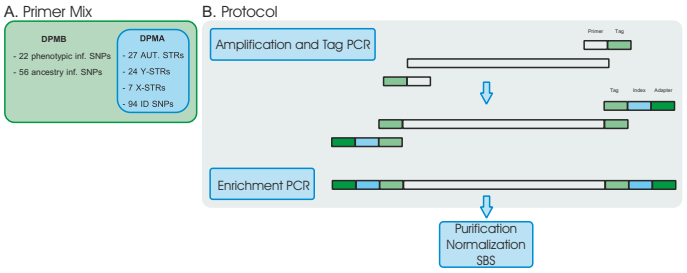

Fig.1A Composition of the two different Illumina designed multiplexes for the ForenSeq Aplication. **1B** ForenSeq protocol overview [4].

Discussion

Quality metrics were calculated per run, namely cluster density, percentage of clusters passing filters, phasing and prephasing. As observed (**Table 1**) in both runs, manufacturer's **quality scores fell within the established values**. Mean sample representation per run was 240393.07 and 111963.36 reads, this difference is probably due to the number of samples loaded in each run(run 1 n=15 and run 2 n=44, **Fig. 1**). Lower values were observed in low input DNA (50pg) and ancient DNA samples (**Fig. 1**).

Mean coverage is not evenly distributed per locus, showing generally higher values on the STR markers than in the SNPs (**Fig.2**). It was also observed that mean coverage distribution within STRs is more heterogeneous than within SNP markers, which could be related to the design of the assay.

The majority of STR markers (**Fig.2**) were called down to 50pg of input DNA (mean of 93.2% of called loci). The highest percentage of dropout was 3.3% at 50pg , ratifying the high sensitivity of NGS methods. Considering the SNP markers (**Fig.2**), a lower percentage of called loci was observed at 50pg (mean of 85.5%) and a higher percentage of dropouts (5.3%).

Concordance has been found in all buccal swap samples for **1014 alleles** compared, excepting from one (DXS10148), in which an off ladder allele had been described in CE results. This marker was posteriorly removed from the panel. Other issues have been found on SE33 (also removed) and Y GATA H4 (corrected). Concordance analysis of the GEDNAP samples was also very satisfactory, showing identical profiles for all 3 single source samples. GEDNAP 2 mixture samples showed five loci in which the ForenSeq contributed for a better deconvolution of the different alleles (**Fig.3**).

In regard to the other mixture assays, the identification of the minor donor was still visible at **1:20 ratio**. Another interesting observation was the deconvolution of stutter produced from the major donor from the minor allele. It shows that NGS sequencing will help further on mixture analysis and separation of different profiles not accessible with CE technologies.

We obtained profiles above **80%** in the gednap samples and above **50%** in the ancient DNA samples (**Fig.1**). The lowest pofile has 58% of typed loci, which corresponds to 89 loci, from which 16 are autosomal STRs (**Fig.1**).

Next generation sequencing brings a new multiplexing dimension and a sharp resolution to the field that will boost casework processing and databasing. However, these technologies will also bring the need to reconstruct database systems as well as improve the nomenclature system to allow the integration of new markers.

Results

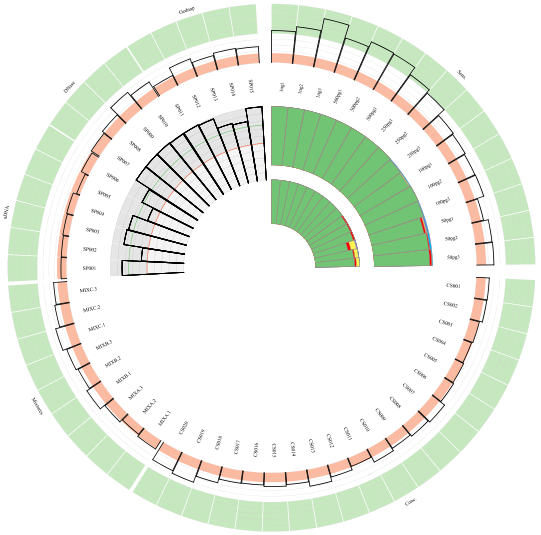

Fig.1 Overview of the performance of each sample. Outer circular histogram shows sample representation (number of reads passing quality filters), red background represents Illumina's threshold of 80000 reads and green background represents values above 250000 reads. Right-side inner graph describes the sensitivity study per sample both considering STRs (outer) and SNPs (inner), color key is established as follows: green - % correct profile, red - % dropout, yellow - % dropin, purple - % discordancies and blue - not covered. Left-side graph shows the percentage of profile obtained in casework-type samples, red line represents 50% and green line 70% respectively. The two runs are separated by large breaks in the circle design, smaller breaks divide different assays performed in the second run. Plot designed using Circos platform [6].

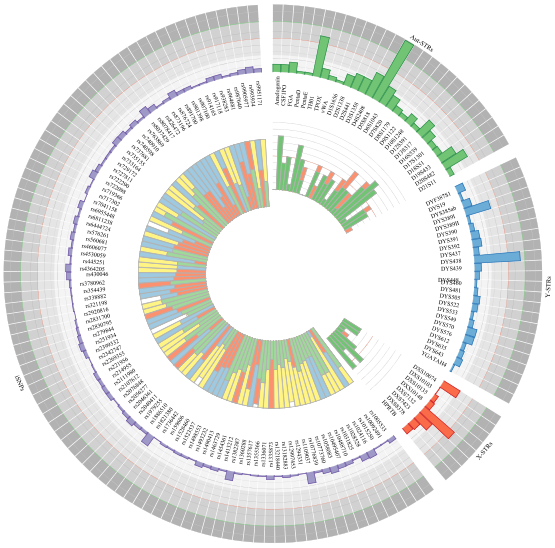

Fig. 2 Outer circular histogram depicts mean coverage per loci, different colors represent different marker types (Autosomal STRs, Y-STRs, X-STRs and iSNPs). Y axis values vary from 0 to 7000, red line is set at 3500 and green line at 5250. Right-side inner histogram shows the number of isometric alleles (green) and the number of isometric heterozygotes (light red). Left-side inner histogram represents allelic frequence in the different SNPs, color key as follows: green - A, red - T, blue - C, yellow - G and white - not covered. Plot designed using Circos platform [6].

	Mean SP (above 80000)	Cluster Density (400-1650 k/mm2)	Cluster PF (above 80%)	Phasing (below 0.25%)	Pre-phasing (below 0.15%)
Run 1	240393.07	607	96.34	0.13	0.04
Run 2	111963.36	687	94.58	0.12	0.05

Table 2. Quality metrics summary per run. Mean SP stands for mean sample representation and is the mean read number per sample achieved in each run. Cluster density is the number of clusters per mm2. Cluster PF is the number of clusters passing filters. Phasing is the number of strands out of phase, falling behind and prephasing is the number of strands jumping a base ahead of phase. Manufacturer's expected values can be found within brackets.

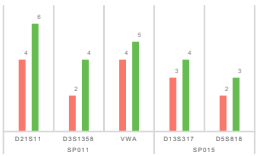

Fig.3 Graph depicting the number of alleles found in the shown markers for two mixture samples with CE (red) and ForenSeq (green) .

References & Acknowledgements

[1]Parson W et al. (2013) *Forensic Sci Int Genet*; 7: 543-549
[2]Parson W et al. (2015) *Forensic Sci Int Genet*; 15: 8-15
[3]EduardoffM et al. (2015) *Forensic Sci Int Genet*; 17: 110-21
[4]Illumina ForenSeq DNA Signature Library Prep Guide (http://support.illumina.com/downloads/forenseq-dna-signature-prep-guide-15049528.html)
[5]Bauer CM et al. (2013) *Forensic Sci Int Genet*; 7:581-586
[6]Krzywinski, M et al. (2009) *Genome Res*; 19: 1639-1645

C. Xavier has a PhD grant (SFRH/BD/90873/2012) from Fundação para a Ciência e Tecnologia (FCT). The authors would like to thank D. Niederwieser, S. Nagl, G. Huber, P. Kralj and H. Niederstaetter for technical support and M. Eduardoff for helpful discussions. The authors would like to thank Angela Kalbande, Jennifer Neumann, Nicola Oldroyd and Illumina for support.