Evaluation of Massively Parallel Sequencing for Missing Persons Identification

Submitted to the National Institute of Justice

Grant Number 2016-DN-BX-0172

Elisa Wurmbach, Ph.D., City Research Scientist

Ewurmbach@ocme.nyc.gov

Recipient Name and Address:

New York City Office of Chief Medical Examiner

Department of Forensic Biology

421 East 26th Street, New York, NY, 10016

Grant Period: 01/01/2017 to 12/31/2019

Final Summary Overview submitted: February 27, 2020

Elisa Wurmbach Feb 27, 2020

Elisa Wurmbach, Ph.D.

City Research Scientist

SUMMARY OF THE WORK

The number of missing persons and unidentified human remains (decedents' who have not been identified) has been described as a "mass disaster over time". The DNA Missing Persons Group at the NYC OCME processes hundreds of missing persons cases per year. Often only skeletal remains or body parts (e.g. limbs) are available and consequently, offer limited physical characteristics to make a match in missing persons databases. Routinely, DNA is first used in an attempt to amplify genomic markers (STRs) to generate an identity and gender profile to compare to DNA databases. If database comparisons are unsuccessful, or if the genomic DNA is too degraded to yield a profile, then mitochondrial DNA can be analyzed by sequencing the regulatory hyper-variable regions HV1/HV2. In addition, missing persons databases more commonly have ancestry, phenotypic and gender information than DNA profiles.

The basic goal of this study was to evaluate the use of massively parallel sequencing (MPS) in conjunction with ancestry, phenotypic and identity SNPs and STRs, as well as sequencing of the mitochondrial HV1/HV2 regions and complete genome, for an improved identification of missing persons. In this regard the following forensic MPS kits were assessed: For nuclear DNA: (i) ForenSeq™ DNA Signature Prep kit Primer Mix B: Amelogenin, 26 autosomal STRs, 24 Y-STRs, 7 X-STRs, and 94 identity SNPs, as well as 24 phenotypic SNPs to predict eye and hair coloration and 56 ancestry SNPs, of which two overlap, and (ii) Promega PowerSeq™ 46GY System Prototype: Amelogenin, 22 autosomal loci, and 23 Y-STR loci. For mitochondrial

DNA: (i) PowerSeq™ CRM Nested kit (Promega): HV1 (16024-16365), HV2 (73-640) and HV3 (438-574), (ii) Human mtDNA D-loop hypervariable regions (Illumina): HV1 and HV2, and (iii) Human mtDNA Genome (Illumina): entire human mitochondrial genome (16,569bp).

For both kits testing STRs, concordance could be verified, however, three samples revealed sequence differences at D12S391, D16S539, and D10S1248, by using the beta version of the

Promega PowerSeqTM 46GY System. The lowest DNA input that resulted in a full profile was 200pg for the ForenSeq™ DNA Signature Prep kit Primer Mix B and 50pg for the Promega's PowerSeq™ 46GY System.

The evaluation of the PowerSeq™ CRM Nested kit and Human mtDNA D-loop hypervariable regions revealed concordance with data obtained by Sanger sequencing and frequently full variant profiles were obtained from nuclear DNA of 50pg.

Nine experimental runs were performed using the ForenSeq™ DNA Signature Prep kit Primer Mix B; testing 266 samples, in order to evaluate phenotypic and ancestry prediction. The call-rate for eye and hair color prediction was 39.1%. Analysis revealed that there were difficulties in predicting the intermediate (amber/green) eye color. UAS had problems in predicting South Asians, individuals of Indian descent.

15 tissue samples of three persons from the Body Farm, with degradation indices (DI) ranging from 0.8-10.7, were tested, as well as DNA from 12 bone samples, with DIs of 1.0-9.6. For the PowerSeq™ CRM Nested System, concordance could be verified with Sanger sequencing. All major variants were detected, however, from one person some minor variants were not observed. The bone samples had insufficient DNA to be sequenced by the Sanger method but were used with the PowerSeq™ CRM Nested System and showed all major variants. The Human mtDNA Genome, sequencing the whole genome, used two PCR reactions, of 9.1kb and 11.2kb, one of which was failing to result in a product. Verogen launched a novel supported kit with an integrated workflow testing the mitochondrial D-Loop, the ForenSeq™ mtDNA Control Region Kit. This kit showed all major variants, but some minor variants within the C-stretch were missing, which could be due to the UAS default settings. This kit is preferred over the Nextera® XT DNA Library Preparation Kit. The experimental run testing degraded samples was performed but failed.

The Beckman Biomek FX liquid handling robot was programmed to perform library preparations in order to have a standardized workflow and to avoid variation from manual handling. The following workflows were programmed: (i) ForenSeq™ DNA Signature Prep kit (ii) Promega's PowerSeq™ CRM Nested kit (iii) Nextera® XT DNA Library Preparation Kit for D-loop and for whole mito genome. All test runs were performed successfully.

1. PURPOSE OF THE PROJECT

Identification of missing persons and human remains is one of the major challenges in forensic genetics. When other approaches such as anthropology, odontology and/or medicolegal investigations are not successful, DNA analysis becomes extremely valuable. However, there are only a few qualified laboratories in the U.S. that have full capabilities to analyze missing person cases, to which the OCME belongs. These capabilities include the full battery of genetic markers: autosomal STRs, Y-STRs, and mtDNA. Current practices to test these markers include PCR amplification followed by capillary electrophoretic separation of autosomal and Y-STRs, and PCR amplification of the regulatory region of the mitochondrial genome followed by Sanger sequencing.

There are some limitations to the current technologies: Capillary Electrophoresis (CE) and Sanger Sequencing. In CE, DNA identification of individuals is performed by separating short but distinct length of amplified DNA (STRs) by CE. Because STR detection by CE is based on amplicon lengths and uses a limited number of fluorescent dyes, it is not possible to detect similar sized alleles, nor is there sufficient chromatographic room to add more alleles. Consequently, forensic improvements requiring additional amplicons for better discrimination are limited in CE. Further, MPS sequencing of STRs reveals that significant numbers of sequence polymorphisms are present in many STRs that make them far more informative than

their size alone. CE cannot detect any polymorphism. Sanger sequencing on the other hand is generally a very robust technique for sequencing, however, each amplicon must be sequenced individually, requiring numerous capillaries and therefore multiple instruments – increasing costs to maintain throughput. MPS sample bar-coding combined with sequencing by synthesis in a flow cell permits sample multiplexing in a single reaction and consequently analyzes multiple genetic polymorphisms simultaneously. MPS technique has improved resolution, throughput, reduced costs, and may aid in mixture interpretations. MPS techniques result in more and detailed information.

The basic goal of the study was to evaluate the use of MPS in combination with ancestry, phenotypic and identity STRs and SNPs, as well as complete mitochondrial genome sequencing, for an improved identification of missing persons. The purpose of this project is to determine how accurate and robust these new markers are with respect to correct identification and their ability to be detected in degraded samples. Our objective is to improve missing person's identification by adding DNA obtainable ancestry and phenotypic characteristics to queries of missing person's databases. Knowing these human characteristics can help to winnow down otherwise impossible long lists to ones with reasonable expectations of success.

The **specific aims** of this application are to **i)** evaluate the sensitivity and repeatability of MPS technology including phenotypic, ancestry and identity markers, **ii)** evaluate Illumina's ancestry and phenotypic SNPs on a diverse population of nearly 300 individuals, **iii)** assess the ability of MPS technology to work with degraded samples typical of missing persons cases, and **iv)** to automate the complex DNA library preparation method in order to improve consistent results.

2. PROJECT DESIGN AND METHODS

Specific aim i) Evaluation of sensitivity and repeatability of MPS technology including phenotypic, ancestry and identity markers:

The following kits were assessed with benchmark (following the recommendations of the manufacturer) runs and for sensitivity: (i) ForenSeq™ DNA Signature Prep kit Primer Mix B (800pg, 400pg, 200pg, 100pg, and 50pg DNA input), (ii) Promega PowerSeq™ 46GY System Prototype (500pg, 400pg, 200pg, 100pg and 50pg DNA input), (iii) Promega PowerSeq™ CRM Nested kit (500pg, 400pg, 200pg, 100pg and 50pg DNA input), (iv) Illumina Human mtDNA D-loop hypervariable regions (800pg, 400pg, 200pg, 100pg, and 50pg DNA input), and (v) Illumina Human mtDNA Genome.

Specific aim ii) Evaluation of Illumina's ancestry and phenotypic SNPs on a diverse population of nearly 300 individuals.

In nine experimental runs (32 samples including controls per run, 1ng DNA input) samples of 266 individuals from various populations including African American, East Asian, South Asian, European, and mixed populations were tested.

Specific aim iii) Assessment of the ability of MPS technology to work with degraded samples typical of missing persons cases

Samples obtained from the body farm: 15 tissue samples from three persons as well as 12 DNA samples from bones. DNA was extracted from homogenized tissues utilizing Microcon filters.

Specific aim iv) Automation of the complex DNA library preparation method in order to improve consistent results.

Beckman Biomek FX liquid handling instrument was programmed with the following library preparation workflows: (i) ForenSeq™ DNA Signature Prep kit Primer Mix B, (ii) Promega PowerSeq™ CRM Nested kit, and (iii) Nextera® XT DNA Library Preparation Kit: Illumina Human mtDNA D-loop hypervariable regions and Illumina Human mtDNA Genome.

3. DATA ANALYSIS

The following software/programs were used for the data analysis: (i) ForenSeq™ DNA Signature Prep kit Primer Mix B: ForenSeq Universal Analysis Software (UAS), Illumina, (ii) Promega PowerSeq™ 46GY System Prototype: GeneMarker HTS, Softgenetics, (iii) Promega PowerSeq™ CRM Nested kit: CLC Workbench AQME tool, Qiagen and GeneMarker HTS, Softgenetics, and (iv) Illumina Human mtDNA D-loop hypervariable regions: CLC Workbench with AQME tool from Qiagen and GeneMarker HTS from Softgenetics.

4. PROJECT FINDINGS

Specific aim i:

- Nuclear DNA testing:

ForenSeq™ DNA Signature Prep kit Primer Mix B: Data are concordant with CE technology, full profile could be obtained from 200pg DNA input, outcomes were repeatable, drop-outs were frequently observed, including in positive controls.

Promega PowerSeq™ 46GY System Prototype: Utilizing the beta version, three sequence differences were detected at D12S391, D16S539, and D10S1248. The following artifacts were frequently seen: (-1, +1, -2)-stutter, and sequence errors at Y-STRs. Sensitivity testing revealed that from six samples with 50 pg DNA input no dropouts were observed. However, it should be noted this kit is still not launched yet.

Table 1: Overview of MPS kits testing nuclear DNA

	Runs	Concordance	Artifacts	Sensitivity (full profile)
ForenSeq A/B	12	Yes	+ and – Stutter Sequence Errors	A: 50 pg DNA B: 200 pg DNA
PowerSeq (beta version)	2	3 differences	+ and – Stutter Sequence Errors	50 pg DNA

- Mitochondrial DNA testing:

Promega PowerSeq™ CRM Nested kit: Data are concordant with data obtained from Sanger sequencing; full variant profiles were obtained with 50pg nuclear DNA input, outcomes were repeatable.

Illumina Human mtDNA D-loop hypervariable regions: Data are concordant with data obtained from Sanger sequencing; full variant profiles were frequently obtained with 50pg nuclear DNA input.

Illumina Human mtDNA Genome: PCR reaction for 11.2kb amplicon was not working for none of the 31 reactions, including the positive control. Repeating these also led to no products to proceed. However, the other PCR reaction of 9.1kb was working. Therefore, in order to cover the whole mito genome more robust PCR reactions should be designed.

ForenSeq™ mtDNA Control Region Kit is a novel supported kit with an integrated workflow, which was launched by Verogen in fall 2019, testing the mitochondrial D-Loop. This kit was tested in addition to the proposed specific aims. The benchmark experiment showed all major variants, but some minor variants in the C-stretch were missing, which could be due to the UAS default settings. This kit is preferred over the Nextera® XT DNA Library Preparation Kit. The experimental run testing degraded samples was performed but failed, because of miscommunication between server and MiSeq. It is planned to repeat this run, after completion of the grant.

Table 2: Overview of MPS testing mitochondrial DNA:

	Runs	Failed	Concordance to Sanger	Sensitivity
PowerSeq D-Loop	5		Yes	50 pg DNA
Nextera D-Loop	2		Yes	50 pg DNA
Nextera mt genome	1	1 (library prep)		
ForenSeq D-Loop	2	1 (sequencing)	Yes	

Specific aim ii:

Nine experimental runs were performed using the ForenSeq™ DNA Signature Prep kit Primer Mix B; testing 266 samples, in order to evaluate phenotypic and ancestry prediction. The call-rate for eye and hair color prediction was 39.1%, because of dropouts. Analysis revealed that there were difficulties in predicting the intermediate (amber/green) eye color. Sometimes the hair color of opposite shade was predicted. UAS had problems in predicting South Asians, individuals of Indian descent. For more details, please see Sharma et al. 2019 *Electrophoresis*.

Specific aim iii:

DNA from 12 bone samples were obtained from Amy Mundorff. Their degradation indices ranged from 1.0-9.6. These samples were tested with the ForenSeq™ DNA Signature Prep kit as well as with the PowerSeq™ CRM Nested System. Table 3 shows the outcomes. Concordance could not be verified, as there was too little DNA to be sequenced by the Sanger method.

Table 3: STR and mito profiles from DNA derived of bones

Bone DNA	DNA [ng/µl]	D.I.	ForenSeq: STR	Mito PCR	PowerSeq
201	0.453	1.0	Profile: DO	product	Profile
211	0.246	1.1	Profile	product	Profile
221	0.017	1.1	Profile	product	Profile
301	0.027	3.3	Profile: DO	product	Profile
311	0.097	1.4	Profile	product	Profile
321	0.045	1.1	Profile	product	Profile
401	0.332	2.1	Profile	product	Profile
411	0.084	1.6	Profile	product	Profile
421	0.048	5.1	Profile	product	Profile
501	0.014	4.5	Profile: DO	product	Profile
511	0.015	3.6	Profile: DO	product	Profile
521	0.059	9.6	Profile: DO	product	Profile

DO: dropouts observed

15 tissue samples of three persons from the Body Farm (Amy Mundorff, University of Tennessee) were selected for DNA extraction. Degradation indices ranged from 0.8-10.3. The nuclear DNA input ranged per sample from 500-50pg. Table 4 shows the outcomes. Concordance for the 3 persons from the body farm could be verified with Sanger sequencing. All major variants were detected. Based on these data, MPS technology is more sensitive compared to Sanger sequencing. It is planned to further analyze the data and publish results and conclusions in a peer reviewed journal.

Table 4: STR and mito profiles from tissues collected from the body farm:

Tissue	Day of collection	DNA [ng/µl]	D.I.	ForenSeq: STR	Mito PCR	PowerSeq:mito
H1	0	10	1.3	Profile	product	Profile
H11/12	6	0.38	4.8	INC	product	Profile
H20	12	0.002	3.6	No DNA	product	Profile
H35	35	0.002	Ø	INC	product	Profile
H38	48	0.02	Ø	INC	product	Profile
I2	0	12.48	0.8	Profile	product	Profile
I11	6	0.21	3.4	Profile: DO	product	Profile
I20	12	0.23	8.9	Profile: DO	product	Profile
I35	30	0.001	3.5	No DNA	product	Profile
I38	45	0.03	Ø	INC	product	Profile
J2	0	10.67	1.1	Profile	product	Profile
J4	10	9.91	1.1	Profile	product	Profile
J7	20	0.013	10.3	INC	product	Profile
J11	30	0.03	2.2	Profile	product	Profile
J15	40	0.06	1.7	Profile: DO	product	Profile

Ø: not detected; DO: dropouts observed; INC: most STRs dropped out

Specific aim iv:

The Beckman Biomek FX liquid handling instrument was programmed to perform the library preparations in order to have a standardized workflow and to avoid variation from manually handling. This instrument was used for the following library preparations: (i) ForenSeq™ DNA Signature Prep kit (ii) Promega's PowerSeq™ CRM Nested kit (iii) Nextera® XT DNA Library

Preparation Kit for D-loop and for whole mito genome. All test runs were performed successfully.

5. RESULTS OF THE FUNDED PROJECT

Peer reviewed publications:

Sharma V, Jani K, Khosla P, Butler E, Siegel D, Wurmbach E. (2019) Evaluation of ForenSeq™ Signature Prep Kit B on predicting eye and hair coloration as well as biogeographical ancestry by using Universal Analysis Software (UAS) and available web-tools. *Electrophoresis* 2019 Feb 15. doi: 10.1002/elps.201800344.

A second publication is planned, describing the findings from the degraded samples using MPS technology, with the tentative title: “Examination of degraded tissue samples collected from the body farm for human identification using MPS”.

Oral presentations of this funded project:

April 27, 2018: Seminar for DNA specialists at the NYC OCME

April 23, 2019: Promega Tech Tour, part of the Bode Meeting in Phoenix, AZ

Internal validation:

Based on the outcomes of this grant, the decision was made for the New York City Office of Chief Medical Examiner (NYC OCME), to bring massively parallel sequencing online for casework. For mitochondrial DNA testing, the OCME is currently validating Promega’s PowerSeq™ CRM Nested Custom Assay, testing the D-Loop, with the CLC Genomics

Workbench and the AQME plug-in tool. Library preparation will be performed on the Beckman Biomek FGx liquid handling instrument.

6. IMPLICATIONS FOR CRIMINAL JUSTICE POLICY AND PRACTICE IN THE UNITED STATES

All experiments, including forensic casework require positive and negative controls to ensure correct working of a kit. However, all positive controls (2800M) using the ForenSeq™ DNA Signature Prep kit mix B, showed variable losses of typed loci. Issues with the positive control indicate limited validity in the run(s) and/or kit as case to case comparisons cannot be performed. In case of dropouts, within the positive control, rules should be established for data analysis in casework: (i) repeat the run, (ii) disregard the group that showed dropouts (e.g. iSNP when there is a high rate of dropouts), or (iii) just disregard the affected loci.

For the phenotypic predictions a dropout rate of around 39% was observed, as these SNPs were used for making the eye and hair color predictions. In order to obtain this data, the run should be repeated.

The ForenSeq™ DNA Signature Prep kit Primer Mix B failed to predict green eye color in 100% of the samples tested. It also predicted the opposite shade of hair color to the true color for a low percentage of cases. For these situations, this kit would not be very useful in describing a missing person's eye and/or hair coloration.

The Illumina whole mitochondrial DNA kit did not produce successful results, as the longer of the two PCR reactions (11.2kb and 9.1kb) covering the whole genome was not resulting in an amplicon, in none of the 31 reactions, including the positive control, despite the use of a special DNA polymerase (Takara, Bio USA, Inc. Mountain View, CA). We suggest a re-design of amplicons and primers to result in better and more robust PCR reactions.

PowerSeq™ CRM Nested Custom Assay, testing the D-Loop, with the CLC Genomics Workbench and the AQME plug-in tool for data analysis to test mitochondrial DNA can be used for forensic casework.