# THE APPLICATION OF NEXT GENERATION SEQUENCING TO STR TYPING AND INVESTIGATIVE GENETICS

Seth A. Faith[1], Dan Bornman[2], Steve Rust[2], Gene Godbold[3], Christine Baker[3], Boyu Yang[4], Curt Barden[1], Scott Nelson[1], Laura Aume[2], Jeremy Craft[2], Jacquelyn Walther[1], Angela Minard-Smith[1], Pearlly Yan[5], Benjamin Rodriguez[5], Ralf Bundschuh[6], Michael Dickens[1], and Brian Young[1]

Battelle Memorial Institute, Divisions of Applied Biology[1], Statistics and Information Analysis[2], International Technology Assessments[3], and Software and Information Engineering[4]. The Ohio State University Medical Center[5] and Department of Physics[6]

## INTRODUCTION

Next-Generation Sequencing (NGS) has the potential to be the ultimate genotyping platform for human identification. NGS is capable of typing the currently important forensic markers, such as short tandem repeats (STRs), mitochondrial and Y-chromosome haplotypes. Using the same DNA sample, NGS is also capable of typing other polymorphisms including single nucleotide polymorphisms (SNPs) that can be investigatively important for DNA phenotyping and determining ancestral origin. While costs for nucleic acid sequencing have decreased dramatically, the current NGS platforms are still too expensive and too slow to be usable for routine forensic analysis. However, if sequencing costs continue on their current downward trajectory, then NGS may soon become cost-competitive with legacy forensic DNA analysis technology. Our prediction of cost-competitiveness for NGS in routine forensic applications will be at least, in part, dependent upon the acceptance of new investigative genetics assays by the forensics community, and the success in developing NGS as a usable platform for a multiplexed suite of assays. That is, if a given DNA sample is to be interrogated for STR profile using a capillary electrophoresis (CE)-based assay, SNP profile by microarray assay and mtDNA by sequencing or mass spectrometry, then a single multiplexed NGS assay may be simpler, faster and cheaper.

However, several challenges need to be overcome to make NGS technology attractive for routine forensics work. The current data analysis paradigm whereby the entire genome of an individual is assembled and aligned to a reference genome is currently too expensive and slow, and for some assays such as STR typing is not usable. Thus, new data analysis pipelines must be developed. In addition, the data interpretations achievable from NGS must be made compatible with needs of the forensic community. One important issue is compatibility with legacy databases. If the legacy STR databases such, as CODIS, consist of allele calls derived from CE mobility measures of allele length, then NGS data must also be formatted as allele calls consistent with the set of alleles in the database. The inherent ability of NGS to uncover many novel alleles that are unobservable to CE-based methods is not necessarily beneficial if it results in an inability to match legacy profiles. Other important issues include the relatively high error rates in current NGS data, development of appropriate sample collection and preparation protocols, and the standardization of methods in the face of rapidly changing technology. In this study, we demonstrate the multi-assay potential of NGS by interrogating saliva-derived DNA samples for both STR and SNP markers as well as for the oral microbiome.

## METHODS

### Sample collection

Human DNA (hDNA) was obtained from anonymous donors using a protocol approved by the Battelle Memorial Institute Internal Review Board. Volunteers completed a survey to report selected phenotypes (hair and eye color, height, weight, gender) and self-reported ancestry using a pedigree chart covering at least three generations. Saliva samples were collected by volunteers using an Oragene®-DNA (DNA Genotek) kit, and DNA was purified using the manufacturer's recommend protocols.

### STR typing by capillary electrophoresis

Human DNA samples were processed with a Promega PowerPlex®16 kit using an Applied Biosystems 3130 Genetic Analyzer, and output data were analyzed with GeneMapper® 4.0 software. All PQVs were evaluated to identify genotype quality. STR profiles generated by this method were used as truth data for comparison to NGS-based STR typing.

### STR typing by NGS

The CODIS core 13 loci were amplified with custom designed primers covering 1600-2000 bp amplicons per locus and Phusion® High Fidelity DNA Pol (NEB). Sequencing libraries were constructed with Illumina DNA TruSeq™ Library kit using index tags. Six multiplexed samples were sequenced in one lane of an Illumina® GAIIx (The Ohio State University) for 150 bp reads (single end). Over five million quality filtered reads were generated for each sample. A custom *in silico* reference sequence was designed for each of the common STR alleles in FASTA format with corresponding annotations in .bed format. Each allele reference contained the full repeat region and 1000 bps of the 5' and 3' flanking sequence. Raw sequence data (.qseq) files were converted to fastq format and then analyzed with reference alignment using Bowtie (Langmeade and Trapnell), SAMtools (Li et. al. 2009), custom software for enumerating sequences, and visualized with IGV (Broad Institute). The Battelle *Galileo* HPC cluster was used to process data.

### Whole genome sequencing

Human DNA was prepared with Epicentre® Nextera™ DNA library kit (Illumina) for sequencing on an Illumina® GAIIx (The Ohio State University) for 10 lanes of 150 bp reads (single end). This data set represented > 4 x $10^8$ quality filtered reads, corresponding to ~12x coverage for entire genome and >250x coverage on the mitochondrial chromosome. Raw data (.qseq) files were converted to fastq format and then analyzed with reference alignment to the human reference ChGr37/hg19 (UCSC, Lander et. al. 2001) using the Burroughs-Wheeler Aligner, SAMtools (Li et. al. 2008, 2009). SNPs were called using GATK (McKenna et al. 2008) and visualized with IGV (Broad Institute). The Battelle *Galileo* HPC cluster was used to process data.

### Phenotype and ancestry predictions

Maximum likelihood estimation model (Frudakis 2008) was implemented using a custom Java 6.0 tool that treated each SNP as an independent measure of probability for the characteristic evaluated. Previously published data for SNP frequencies and correlation to phenotype (Duffy et al. 2007, Kayser et al. 2008) and ancestry (Nassir et al. 2009) were used in the MLE model.

### Microbiome analysis

The unmapped reads remaining from human reference alignment (representing 10-20% of the total reads produced) were analyzed using the BLAST and bacteria database at NCBI. Genetic matches were recorded and unique organisms were identified.

## RESULTS

### STR typing by reference alignment

The NGS protocol resulted in coverage depths ranging from 700 to 20,000 for the STR loci. By aligning these reads to the in-silico reference "genome" consisting of the common alleles for each locus, the method employed was able to correctly call 12 of the 13 CODIS core STR loci for two individual samples, as determined by comparison to capillary electrophoresis (CE) methods (Table 1). However, one complex locus, D21S11, showed ambiguous alignment to the reference. In the case of sample VFD10-001, two CE-determined alleles for that individual were detected. However, the amplicons also aligned significantly to a third allele (30.2). For individual VFD10-018 significant alignment occurred only for one of the two CE-determined alleles. For a 1:1 mixture sample of samples VFD10-001 and VFD10-018, the majority of STR alleles were called correctly with a few exceptions (e.g., D21S11, FGA, and vWA) that contained "drop-out", misalignment or additional calls.

| | Sample | | | | | |
|---|---|---|---|---|---|---|
| | VFD10-001 | | VFD10-018 | | VFD10-118 (Mixture) | |
| **Loci** | Powerplex16[a] | Illumina® | Powerplex16[a] | Illumina® | Powerplex16 | Illumina® |
| **CSF1PO** | 11,12 | 11,12 | 10,12 | 10,12 | ND | 10,11,12 |
| **D13S317** | 11,13 | 11,13 | 11,12 | 11,12 | ND | 11,12,13 |
| **D16S539** | 12,13 | 12,13 | 11,14 | 11,14 | ND | 11,12,13,14 |
| **D18S51** | 13,17 | 13,15,17 | 13,15 | 13,15 | ND | 13,15,17 |
| **D21S11** | 30,34.2 | 30,30.2,34.2 | 28,30.2 | 30.2, - | ND | 30,30.2,34.2,28? |
| **D3S1358** | 16,16 | 16,16 | 16,16 | 16,16 | ND | 16 |
| **D5S818** | 11,13 | 11,13 | 11,13 | 11,13 | ND | 11,13 |
| **D7S820** | 10,11 | 10,11 | 11,11 | 11,11 | ND | 10,11 |
| **D8S1179** | 13,13 | 13,13 | 13,15 | 13,15 | ND | 13,15 |
| **FGA** | 22,23 | 22,23 | 22,26 | 22,26 | ND | **21,22,23,26** |
| **TH01** | 6,9.3 | 6,9.3 | 6,9 | 6,9 | ND | 6,9,9,3 |
| **TPOX** | 8,8 | 8,8 | 8,10 | 8,10 | ND | 8,10 |
| **vWA** | 16,17 | 16,17 | 14,18 | 14,18 | ND | **11,14,16,17,18** |

**Table 1. Comparison of reference alignment to CE**

Quantifying the number of reads aligning to alleles in the in-silico reference allows one to represent the data in the form of a histogram that is familiar to forensic scientists. The area under the histogram represents depth of coverage and this depth of coverage is amenable to ratio analysis of the sort used when interpreting CE electropherograms. For example, the TPOX locus, sample VFD10-001 presented as homozygous for allele 8, with 1704 total reads, and sample VFD10-018 presented as heterozygous for alleles 8 and 10, with 886 and 758 reads respectively, respectively. The 1:1 mixture sample demonstrated that the ratio of reads mapping to the 10/8 alleles was 0.24 (653/2725 reads), an expected value for the presence of three TPOX_8 alleles and one TPOX_10 allele in a proportionally mixed sample. The quantitation of the mapped reads also showed some stutter as seen with some low amount of -1 repeat alleles. The stutter observed was within the expected range for CE assays, (1-8%), and was higher for alleles D3S317 and D8S1179 (≤15%).

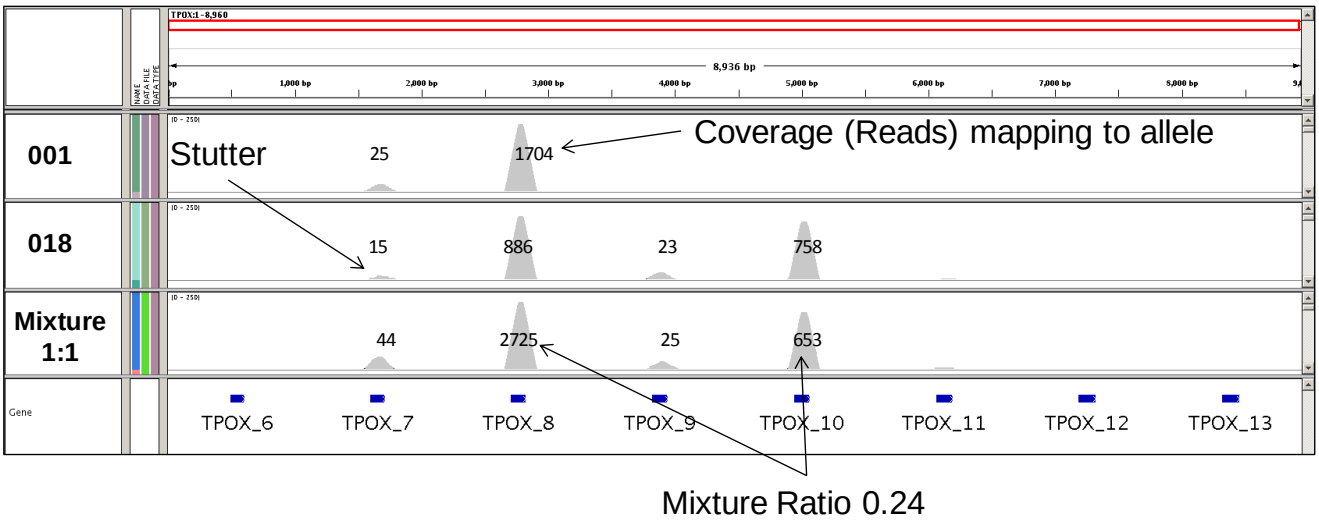


**Figure 1. Reference alignment results for TPOX locus for two individual samples and a 1:1 mixture**

The sequencing method also allowed for identification of SNPs within STRs. In Figure 2, the locus D3S1358, is shown for two individual samples and a 1:1 mixture. Both individuals presented with homozygous alleles containing 16 repeats both by PowerPlex16 analysis and the NGS reference alignment method. However, three separate alleles were present between the two individuals when SNPs were examined. Sample VFD10-001 had the prototypical 16 repeat allele TCTA$[TCTG]_3[TCTA]_{12}$ and an allele with a transition (G→A) SNP, TCTA$[TCTG]_2[TCTA]_{13}$. Sample VFD10-018 had the same transition (G→A) SNP allele, TCTA$[TCTG]_2[TCTA]_{13}$, plus a double transition 2(G→A) SNP allele TCTA[TCTG][TCTA]$_{14}$. The power of obtaining sequence information is further demonstrated by analysis of the ratios for the 1:1 mixture sample, which showed appropriate proportions for the three alleles.

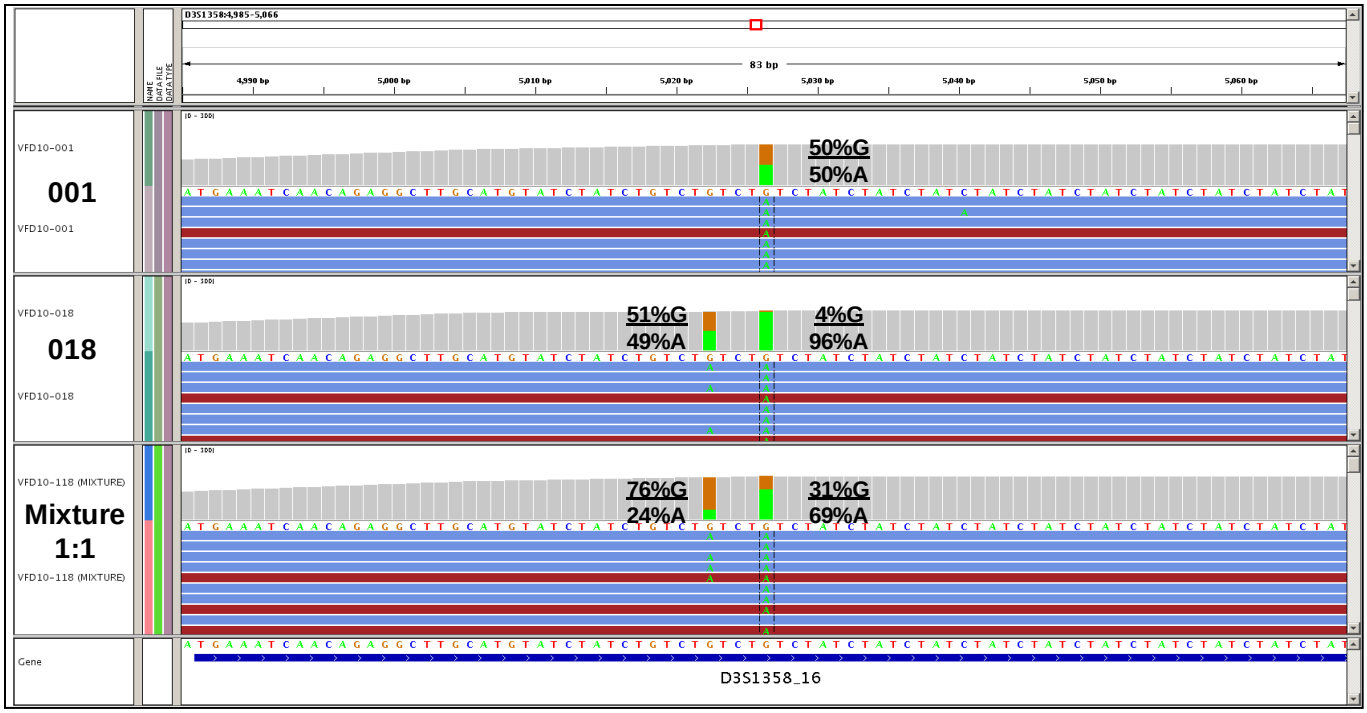


**Figure 2. Sequence information and SNP identification for locus D3S1358**

Overall, the reference alignment method was effective for calling STR alleles, finding SNPs and sorting out mixed samples from NGS data produced on the Illumina GAIIx. From the samples examined, very few mis-alignments, “drop-outs” or miscalls were observed. Like capillary electrophoresis, some stutter was observed (Figure 1). Upon initial inspection stutter was in range of observed CE assays (1-8%, ≤15% D3 and D8). We note that the PCR and library preparation extension temperatures were performed at 72°C and stutter may be reduced by optimizing PCR conditions.

Future challenges for this approach include: detect and report SNPs by automated processes, eliminate/reduce misalignment, “drop-out” or mis-calling, detect and report off-ladder alleles (indels & rare alleles), call alleles for complex and large STRs (i.e., D21), develop validation methods and PQVs, increase multiplexing potential and decrease turnaround time. From our studies it appears that most of these improvements may be attained through improved bioinformatics and custom pipeline analyses.

## SNP genotyping for ancestry and phenotype

For this study, the entire genome of sample VFD10-018 was sequenced, and SNPs were called for the purpose of predicting the individual’s ancestry, and physical phenotype. Relying on a panel of previously evaluated ancestry informative markers (AIMs) from Nassir et al. 2009, we employed a maximum likelihood estimation (MLE) model, as described by Frudakis (2008), to evaluate the ancestry of the individual contributing sample VFD10-018. The analysis showed that the individual had the highest association with a group defined as Utah residents with Northern and Western European ancestry, 0.65 probability (Table 2). The individual had self-reported as having German ancestry on both the paternal and maternal lineages for at least three previous generations. Therefore, the prediction matched the self-reported data.

| Group Identity Prediction [a] Sample 018 | | Self-Reported Ancestry |
|---|---|---|
| 0.65 | Utah residents with Northern and Western European ancestry | German, at least 3 generations |
| 0.64 | Tuscan in Italy | |
| 0.55 | Gujarati Indians in Houston, Texas | |
| 0.44 | Japanese in Tokyo, Japan | |
| 0.43 | Mexican ancestry in Los Angeles, California | |
| 0.42 | Han Chinese in Beijing, China | |
| 0.40 | Maasai in Kinyawa, Kenya | |
| 0.38 | Yoruban in Ibadan, Nigeria | |
| 0.37 | African ancestry in Southwest USA | |
| 0.36 | Chinese in Metropolitan Denver, Colorado | |
| 0.27 | Luhya in Webuye, Kenya | |

[a] Diversity Panel data obtained from Nassir R et al., BMC Genetics, 10:39 (2009)

**Table 2. Prediction of ancestry from AIM panel**

Phenotypes were evaluated using a similar MLE method and SNPs previously characterized. Table 3 shows the results that demonstrated the prediction of eye color to be brown (95.5% chance), which matched the self-reported data. Hair color was predicted to be dark brown or black, which also matched the self-reported data. In addition, the same sample showed genotype for two SNPs (rs4988235 and rs6113491) that would suggest increased odds of having male-pattern baldness. This fact and the possibility of graying hair must be taken into account when evaluating the hair color predictions made. The survey used for this study did not inquire on baldness or graying.

Skin tone was also examined and the MLE model predicted the individual to have olive/dark (47.40%) to medium (42.1%) skin tone (Table 3). This study did not ask volunteers to report skin tone. Lastly, SNPs

rs2153271, rs4778138, and rs1805007 were examined to determine likelihood of freckling. The genotype of the individual suggested only slightly higher odds of freckling skin.

| Phenotype Examined | Genoset Examined | Probability % Sample 018 | Self-Reported |
|---|---|---|---|
| Eye Color[a] | SNP:<br>rs7495174<br>rs4778138<br>rs4778241<br>rs916977<br>rs12913832 | **Brown 95.50**<br>Blue/Gray 4.50<br>Green/Hazel 0.00 | Brown |
| Hair Color[b] | SNP:<br>rs7495174<br>rs4778138<br>rs4778241 | **Black 45.50**<br>**Dark/Brown 45.50**<br>Light/Brown 4.50<br>Fair/Blond 4.50<br>Red/Auburn 0.00 | Brown |
| Skin Color[b] | SNP:<br>rs7495174<br>rs4778138<br>rs4778241 | Olive/Dark **47.40**<br>Medium 42.10<br>Fair/Pale 10.50 | Not reported[c] |

[a] Frequency data obtained from Kayser et .al., The Am J of Human Gen 82, 411-423, (2008).
[b] Frequency data obtained from Duffy et .al., The Am J of Human Gen, 80, 241-253 (2007).
[c] Skin pigmentation data was not self- reported. .

**Table 3. Prediction of phenotype using MLE**

## Oral Microbiome

Oral microbiome analysis was also performed with the genetic data from the saliva samples. This analysis revealed a microbiome profile in terms of species level identification and relative abundance of microorganisms (Figure 3). Such a profile may be unique to the individual and thereby useful for investigative genetics purposes.

In addition to possible uniqueness of the microbiome, the microbiome can also reveal investigatively important information about an individual's recent activities. For example, the microbiome of individual VFD10-018 contained genetic signatures of the bacterium *Streptococcus thermophilus*, which is used in industrial processes to make cheeses and yogurt. This is consistent with recent ingestion of cheese or yogurt.

The microbiome can also provide information on an individual's geographic location. For example, the individual contributing sample VFD10-018 contained a GG genotype at SNP rs4988235, resulting in little or no production of the lactase enzyme. Lactose intolerance has been associated with specific ancestries and could serve as an additional clue of the individual's biogeographic origin.

The oral microbiome of the individual contributing sample VFD10-018 also contained genetic signatures of the fungus *Histoplasma capsulatum*. This fungus has a unique geographic restriction, primarily the North American mid-west region with a high concentration in the Mississippi river valley area, the so-called Histoplasmosis belt. Sample VFD10-018 was collected in Columbus, OH, which is within the geographic distribution range of *H. capsultum*.

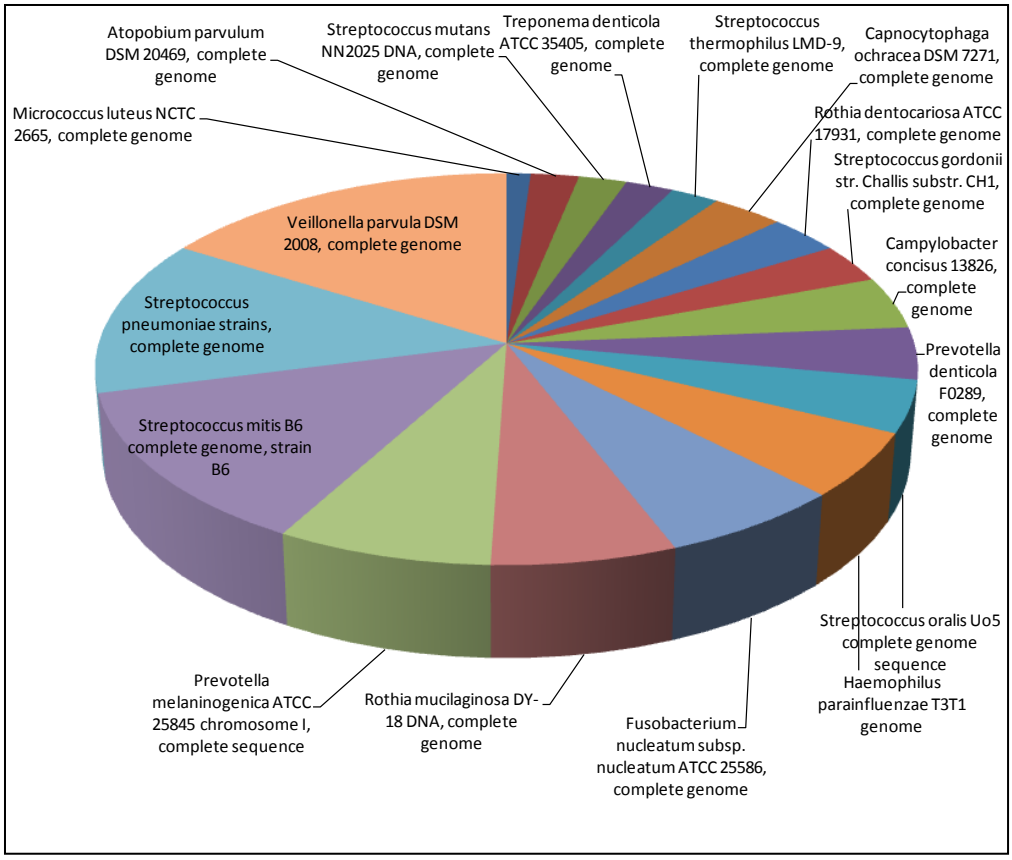


**Figure 3. Diversity and relative abundance of oral microbiome in sample VFD10-018**

In summary, our case study using NGS on a saliva sample yielded forensic data of the individual's STR profile. We also determined gender (Male), mitochondrial haplotype, and Y-CHM haplotype by SNP or STR (data not shown), which is highly valuable for forensics profiling and kinship analysis. In addition, investigative genetics were used on the NGS data to further show that the individual was of European ancestry, had brown/dark eyes and hair, olive to light/medium skin tone, may have had freckles, possibly was balding or bald, may have ingested yogurt/cheese, and may have lived or have traveled in a region where they were exposed to *H. capsulatum* (Midwest, MS river valley area). Together, this set of data demonstrated the potential of NGS to be used as a multipurpose genotyping platform. A vibrant community of developers in NGS technology could expand upon this work to help this technology move into the forensics laboratories for routine use in human genotyping.

## ACKNOWLEDGEMENTS

This work was supported by Battelle internal research and development funds project 100000279. Sequencing on the Illumina GAIIx was performed at the sequencing core at The Ohio State University Medical Center.

## REFERENCES

Duffy, D. L., G. W. Montgomery, et al. (2007). "A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation." Am J Hum Genet **80**(2): 241-52.

Frudakis, T. (2008). Molecular Photofitting: Predicting Ancestry and Phenotype Using DNA, Academic Press.

Kayser, M., F. Liu, et al. (2008). "Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene." Am J Hum Genet **82**(2): 411-23.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics **25**(14): 1754-60.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-9.

Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Res **18**(11): 1851-8.

McKenna, A., M. Hanna, et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome Res **20**(9): 1297-303.

Nassir, R., R. Kosoy, et al. (2009). "An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels." BMC Genet **10**: 39.