Forensic Science International: Genetics Supplement Series 3 (2011) e447–e448

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/FSIGSS

# Assessing the potential of next generation sequencing technologies for missing persons identification efforts

Jodi Irwin *, Rebecca Just, Melissa Scheible, Odile Loreille

*Armed Forces DNA Identification Laboratory, Armed Forces Medical Examiner System, 1413 Research Boulevard, Rockville, MD, 20850, USA*

ARTICLE INFO

*Article history:*
Received 31 August 2011
Accepted 15 September 2011

*Keywords:*
Next generation sequencing
Missing person
STR
Mitochondrial DNA
Target enrichment

ABSTRACT

To assess the utility of next generation sequencing (NGS) technologies for missing persons applications, we have recently initiated a study of various platforms and target enrichment strategies for sample types regularly encountered in our large-scale identification efforts. Specific laboratory workflows based on target marker enrichment and NGS platform are being considered for different sample types, and the overall effort is being undertaken with a strong emphasis on both raw data and final consensus sequence quality. The current study is first and foremost a general evaluation of these data and technologies from the standpoint of forensic application, yet the strategy we are pursuing is ultimately intended to facilitate NGS integration into standard casework laboratories. We are, therefore, evaluating NGS workflows and data for the typical nuclear and mitochondrial DNA markers used in forensics, while still allowing for future work that may take greater advantage of the strengths of these technologies. Here, we present an overview of our NGS strategy.

Published by Elsevier Ireland Ltd.

## 1. Introduction

Developments in DNA sequencing methods have advanced rapidly over the past few years, and new technologies that produce large volumes of data at low cost (relative to current platforms) are being broadly applied to various questions in medical genetics, evolutionary biology, molecular anthropology, phylogeny, epidemiology and metagenomics. These so-called next generation sequencing (NGS) technologies are permitting the *de novo* characterization of entire genomes, and are also being used to resequence 1000 human genomes [1]. Given these types of applications, it is not difficult to imagine the potential implications of this technology for missing persons and disaster victim identification purposes. Historically, the recovery of large numbers of markers (more than 50 STRs or SNPs, for example) in a single assay has been restricted by technical limitations of current, established capillary based sequencing and genotyping technologies, particularly when considering the typical damaged and degraded forensic specimens regularly encountered in missing persons casework. However, these limitations do not apply to NGS. Instead, the simultaneous recovery of the standard autosomal DNA (STRs and SNPs), mitochondrial DNA, as well as X and Y-chromosomal markers regularly assayed in forensic genetics may well be possible with these new technologies.

Although NGS technologies have the potential to improve the overall cost, throughput, and discriminatory power of applied DNA-based assays, their practical utility for forensic genetics specifically is as yet unproven. The feasibility of implementing even standard NGS approaches in routine casework laboratories has not yet been investigated. And while the benefits of the technology have been demonstrated on some of the most difficult specimens [2–6], in an applied casework laboratory these benefits must be carefully weighed against their cost in terms of time and money. Furthermore, and even more importantly, neither NGS data quality assessments nor data analysis/interpretation procedures have yet been performed according to the strict standards required for forensic genetics [7,8]. Thus, it remains to be seen if NGS technologies are suitable for forensic applications [9].

The goal of our current effort is to comprehensively characterize NGS data from the standpoints of quality and interpretation as required for forensic purposes. Data from various NGS platforms on a variety of relevant sample material are being evaluated, with the objective of detailing the characteristic features, strengths, weaknesses and limitations of the NGS chemistries and platforms, using the genetic markers regularly employed by our laboratory. Generally speaking, the effort is intended to establish a comprehensive understanding of the data so that laboratory assays can be improved to mitigate data artifacts, and appropriately robust and reliable NGS data analysis and interpretation guidelines can be developed.

## 2. Strategy

### 2.1. Samples

The range of sample types and qualities regularly encountered in the Armed Forces DNA Identification Laboratory's (AFDIL) human identification casework are being targeted in these studies. That is, common reference material such as buccal swabs and bloodstains are being used to develop assays and NGS data for

* Corresponding author. Tel.: +1 301 319 0244; fax: +1 301 295 5932.
*E-mail address:* jodi.a.irwin@us.army.mil (J. Irwin).

1875-1768/$ – see front matter. Published by Elsevier Ireland Ltd.
doi:10.1016/j.fsigss.2011.09.085

high-quality specimens, while representative degraded skeletal remains (nonprobative case samples), which vary dramatically in terms of quality, are being used to develop data generation and analysis procedures for moderately and severely degraded specimens. It has been our experience that only by using actual casework material can the utility of any assay, or the value of the resulting data, be adequately assessed for application to degraded specimens. Thus, our NGS efforts are targeting these samples from the start. All samples have been specifically selected based on the availability of Sanger mtDNA sequence data and/or standard STR profiles to serve as reference data.

### 2.2. Target enrichment

The markers being targeted for enrichment are the standard nuclear and mitochondrial loci used in our current identification efforts. Although this strategy of targeting very few markers does not take full advantage of the strengths of NGS in terms of data production, it permits a direct comparison of the NGS data to existing high-quality control data. We anticipate increasing the number of target markers in the future. However, at this stage of the study we are simply evaluating the potential of these new technologies, and the best way to do this is by restricting testing to the markers with which we are most familiar and for which we have standard reference data from both pristine and degraded specimens.

For target enrichment, three different approaches are currently being pursued: a standard PCR-based approach, hybridization capture [10] and primer extension capture [2]. Different target enrichment strategies are being employed for different sample types (in terms of quality) and target markers. The ease of any given enrichment method and its potential for near-term implementation in a production setting are also considerations. However, the quality of the NGS data produced with a given enrichment method on a given sample type will ultimately dictate the strategies we elect to further optimize.

### 2.3. Platforms

NGS data from both pristine and degraded samples are being generated at this time on two different platforms: the 454 GS Junior system (Roche Diagnostics Corporation), and the Illumina Genome Analyzer (Illumina, Inc). The Roche 454 Jr is being used to sequence STR amplicons since the short read lengths of the current Illumina sequencing chemistry (at present, the maximum read length is 150 base pairs for single-end sequencing) do not permit complete coverage of most repeat regions. The Illumina Genome Analyzer is being used primarily for mtDNA sequencing efforts, as this is the most cost-effective approach for generating deep read coverage. High coverage depth is desirable for mtDNA applications because it (1) permits better detection of low level heteroplasmy and (2) facilitates the interpretation of highly damaged DNA templates.

### 2.4. NGS data review

NGS data are being directly compared both to fragment analysis profiles generated with standard, commercially available STR kits, as well as Sanger-based sequence data that have been generated from the same samples. The control data are serving as the benchmark by which to assess the overall signal to noise ratio in the NGS data as measured by, among other things: the overall frequency of NGS errors; mtDNA heteroplasmy detection; the ease of distinguishing heteroplasmy from sequencing error; and the accuracy of both the individual reads and the consensus sequences across homopolymer regions, short tandem repeats and other complicated motifs. In all cases, the NGS data are being tied back to the sample preparation/target enrichment steps, so that the effect of different library construction protocols on data recovery, data quality, and data consistency with Sanger profiles can be determined.

In addition, since the different NGS platforms under investigation are known to exhibit well-described differences in the resulting data, our data are being reviewed with these chemistry-specific characteristics in mind. 454 data, for instance, are known to be problematic in homopolymer regions due to the specific chemistry/signal detection of this platform. As a result, the 454 sequences are being evaluated under the assumption that these chemistry-specific artifacts are likely to manifest in the sequence reads in other ways as well.

## 3. Discussion

By understanding and characterizing the basic features and footprints of the various NGS chemistries and then tying those assessments back to the laboratory assays, NGS raw data quality can be improved and data analysis/interpretation strategies that accommodate these features and are appropriate for forensic applications can be devised. Our preliminary results from both the Illumina and 454 platforms with pristine and degraded samples already suggest that this should be feasible. Though the fundamental differences between the various NGS technologies may require the development of platform-dependent data quality, analysis and interpretation guidelines, a comprehensive understanding of the data will help mitigate particular issues at both the laboratory assay and data analysis stages. In the end, we hope to generate the same high-quality raw and consensus data using NGS technologies that are routinely developed via Sanger sequencing. This will, in turn, lead to the same level of comfort with and confidence in the use of NGS data for forensics as we have with Sanger-based data.

## Conflict of interest

None.

## Acknowledgements

The authors would like to thank LTC Louis Finelli, James Ross, Lanelle Chisolm, Shairose Lalani, James Canik, Brion Smith, Marjorie Bland, Suzanne Barritt and the American Registry of Pathology for their help in making this work possible. We are also grateful to Walther Parson, Harald Niederstätter, Michael Hofreiter, Adrian Briggs and Mark Whitten for valuable discussion. Funding for this research was provided by the Armed Forces DNA Identification Laboratory. The opinions and assertions contained herein are solely those of the authors and are not to be construed as official or as views of the US Department of Defense or the US Department of the Army.

## References

[1] 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, Nature 467 (2010) 1061–1073.
[2] A.W. Briggs, J.M. Good, R.E. Green, et al., Targeted retrieval and analysis of five Neandertal mtDNA genomes, Science 325 (2009) 318–321.
[3] R.E. Green, J. Krause, A.W. Briggs, et al., A draft sequence of the Neandertal genome, Science 328 (2010) 710–722.
[4] J. Krause, Q. Fu, J.M. Good, et al., The complete mitochondrial DNA genome of an unknown hominin from southern Siberia, Nature 464 (2010) 894–897.
[5] M. Rasmussen, Y. Li, S. Lindgreen, et al., Ancient human genome sequence of an extinct Palaeo-Eskimo, Nature 463 (2010) 757–762.
[6] D. Reich, R.E. Green, M. Kircher, et al., Genetic history of an archaic hominin group from Denisova Cave in Siberia, Nature 468 (2010) 1053–1060.
[7] O. Harismendy, P.C. Ng, R.L. Strausberg, et al., Evaluation of next generation sequencing platforms for population targeted sequencing studies, Genome Biol. 10 (2009) R32.
[8] M.V. Zaragoza, J. Fass, M. Diegoli, et al., Mitochondrial DNA variant discovery and evaluation in human Cardiomyopathies through next-generation sequencing, PLoS ONE 5 (2010) e12295.
[9] H.J. Bandelt, A. Salas, Current next generation sequencing technology may not meet forensic standards, Forensic Sci. Int. Genet. (2011).
[10] T. Maricic, M. Whitten, S. Paabo, Multiplexed DNA sequence capture of mitochondrial genomes using PCR products, PLoS ONE 5 (2010) e14004.