Forensic Science International: Genetics Supplement Series 8 (2022) 20–22

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

Whole-genome sequencing of degraded DNA for investigative genetic genealogy

Janet Cady *, Ellen M. Greytak

Parabon NanoLabs, Inc., 11260 Roger Bacon Dr., Ste 406, Reston, VA 20190, USA

ARTICLE INFO

Keywords:
Investigative genetic genealogy
Whole-genome sequencing
Low coverage
Imputation

ABSTRACT

Whole genome sequencing has opened the doors to Investigative genetic genealogy (IGG) analysis of challenging forensic samples that are not suitable for microarray genotyping. These samples still do not typically achieve high enough coverage for direct genotype calling, therefore a pipeline for imputation from low coverage sequencing data was evaluated using data from the 1000 Genomes Project. This pipeline generated results suitable for IGG down to 0.25X coverage. Additionally, forensic samples from a variety of tissue types and input amounts were sequenced and successfully uploaded to genetic genealogy databases after imputation.

1. Introduction

Investigative Genetic Genealogy (IGG) is the attempt to identify an unknown individual by finding genetic relatives using public genetic genealogy databases and combining the results with traditional genealogical research of public records. High-quality SNPs are required in order to obtain accurate database match results, and most solved cases have used microarray genotyping. However, for highly degraded samples, such as bone, microarray genotyping is ineffective; instead, whole-genome sequencing (WGS) must be used. Due to degradation and the presence of non-human DNA in such samples, sequencing coverage is typically too low for direct genotype calling, and imputation is required.

2. Material studied, methods, techniques

2.1. Imputation validation

The WGS alignment file for a subject of British ancestry, HG00119, from the 1000 Genomes Project [1] was downloaded, which has ~5.3× coverage. The alignment was randomly subsampled to 2.5X, 0.5X, 0.25X, and 0.05X. Low-coverage imputation [2] was run on each subsample using a reference panel from the 1000 Genomes Project with subject HG00119 removed. Accuracy was determined by comparing the imputed genotypes for the ~2.2 million target SNPs to their genotypes in the 1000 Genomes phase 3 call set.

IBIS [3] was used to detect shared identical-by-descent segments between samples. The phase 3 call set (actual) and imputed genotypes were used to determine the amount of shared DNA between HG00119 and a known 2nd degree relative that was not included in the reference panel as well as 18 other unrelated subjects with British ancestry from the 1000 Genomes data. For each pair, all shared segments > 7 cM were summed to determine the total shared DNA.

2.2. Forensic samples

Samples were submitted to either HudsonAlpha Discovery or Arbor Biosciences for library preparation and sequencing. For each sample, a small amount of the prepared library was sequenced to determine the quality and human DNA content. A whole human genome enrichment (WGE) step was added for samples with a low percentage of reads aligned to the human genome. All samples were then sequenced to a target of 30× coverage (~100 Gbases per sample) using 2 × 151 bp paired-end reads on an Illumina NovaSeq. FASTQ files were trimmed to remove adapter sequences and poly-G strings introduced in the library prep, then aligned to the human reference genome using bwa-mem. A custom filter was applied to remove alignments with significant soft clipping as these are likely to be from non-human sources. Duplicate reads were marked using the Picard Tools MarkDuplicates command. Coverage was calculated conservatively, only counting the aligned, non-overlapping portions of unique reads.

* Corresponding author.
E-mail address: janet@parabon.com (J. Cady).

https://doi.org/10.1016/j.fsigss.2022.09.008
Received 19 September 2022; Accepted 23 September 2022
Available online 24 September 2022

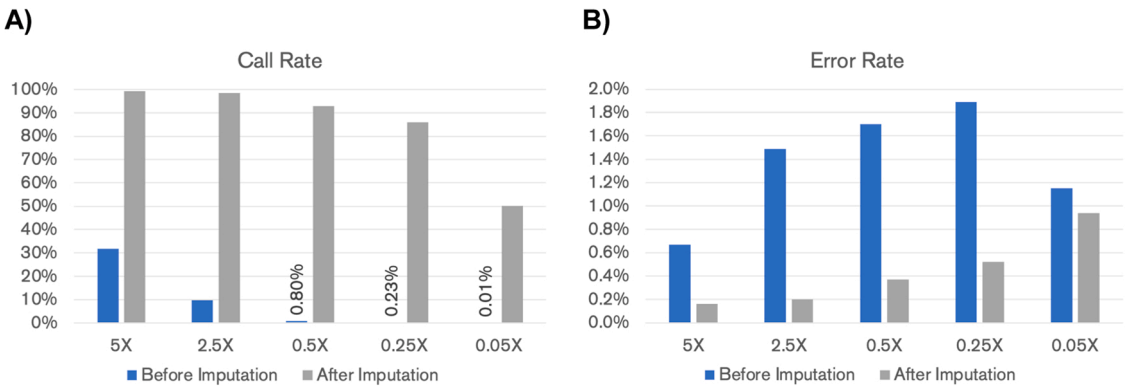

Fig. 1. A) Call rates and B) Error rates before and after low coverage imputation.

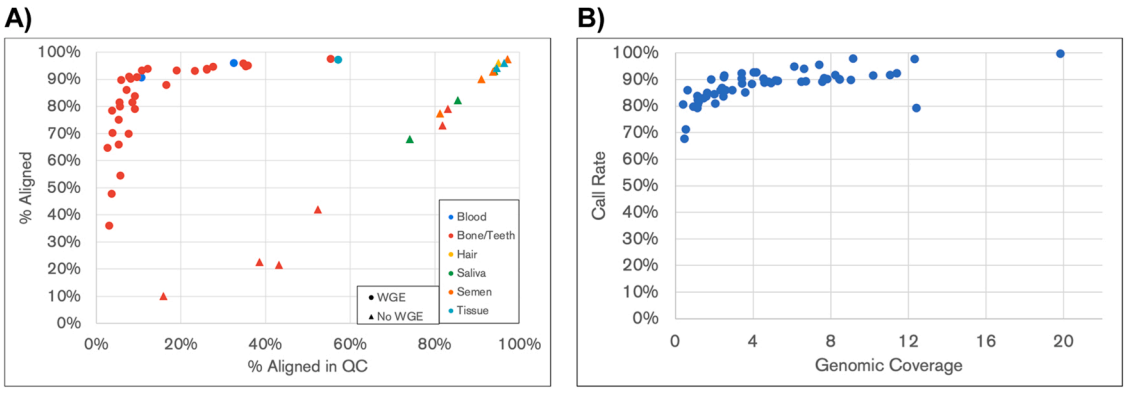

Fig. 2. Results of sequencing forensic samples showing A) alignment rates to the human genome in QC and final sequencing and B) genomic coverage and call rates.

3. Results

3.1. Simulated low coverage sequencing

The low coverage imputation pipeline significantly improved call rates and error rates over direct genotype calling alone (Fig. 1). While error rates prior to imputation are under 2 % at all coverage levels, the call rates were very low (< 30 %).

The results from the imputation pipeline at each coverage level were compared to other subjects from the 1000 Genomes Project using IBIS. Imputed genotypes from alignments down to 0.25× coverage resulted in negligible differences from the expected amounts of shared DNA. While the 0.05× coverage data detected the correct amount of shared DNA with truly related samples, it showed a large inflation in the amount of DNA shared with unrelated subjects.

3.2. Forensic samples

Over 70 forensic cases have been successfully uploaded to a genealogy database after low coverage imputation with as little as 0.114 ng of DNA submitted. The majority of extracts were from bone. Most samples showed high levels of degradation measured by the Quality Index. High quality DNA is expected to have a Quality Index of 1 with lower scores corresponding to higher levels of degradation. Only one forensic sample had a quality index of 1 or higher. Shallow sequencing of the prepared libraries showed that extracts from bone in particular had very high amounts of non-human DNA present and the WGE significantly improved the proportion of human reads generated (Fig. 2A). Despite being sequenced to a target of 30X coverage, roughly 90% of cases resulted in 10X coverage or less, with 61% under 5X coverage (Fig. 2B). This analysis pipeline was successful on samples down to 0.39X coverage.

4. Discussion

The low coverage imputation pipeline was highly effective at achieving high call rates and low error rates from very low coverage data. The results from this pipeline were found to be suitable for use in IGG down to 0.25× coverage. At 0.05× coverage, there was a significant inflation in the amount of DNA shared between unrelated subjects, therefore care should be taken when performing IGG with very low quality to data as there may be false matches. While low coverage was simulated in this evaluation, other challenges seen in forensic samples such as contamination with non-human DNA and uneven distribution of coverage are not captured in this analysis. The former issue is significantly reduced by human genome enrichment prior to sequencing while the latter results in lower call rates than those seen in the validation data and variation in post-imputation call rates for samples with equal overall coverage. Despite these difficulties, over 70 forensic samples have been successfully imputed and uploaded to genetic genealogy databases.

5. Conclusion

Forensic samples with low amounts of DNA, high levels of degradation, or high amounts of non-human DNA aren't suitable for microarray genotyping, however WGS can often be used instead. High call rates can be achieved from these samples using an accurate low coverage imputation pipeline. Combining lab techniques to select for human DNA with low-coverage imputation greatly expands the pool of cases that are candidates for IGG.

Conflict of interest

J.C. and E.G. are employees of Parabon NanoLabs, Inc., which provides IGG services to law enforcement.

References

[1] The 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, Nature 467 (7319) (2010) 1061–1073.

[2] R. Hui, et al., Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes, Sci. Rep. 10 (1) (2020), 18542.

[3] D.N. Seidman, et al., Rapid, phase-free detection of long identity-by-descent segments enables effective relationship classification, Am. J. Hum. Genet. 106 (4) (2020) 453–466.