Improving Forensic SNP-Based Outcomes: Insights from Case Studies, Comparative Sequencing & Bioinformatic Approaches

Alaina Addison, MS[1], Jessica Bouchet, BS[1], Meghan Didier, MS[1], Sheila Diepold, MS[1], Kyla Hackman, MFS[1], Kevin Lord[2], Richard E. Green, PhD[2], Cydne Holt, PhD[1]

[1]*Forensic Services Division, Tetracore®, Inc., Rockville, MD 20850*

[2]*Astrea Forensics™ LLC, Scotts Valley, CA, 95066*

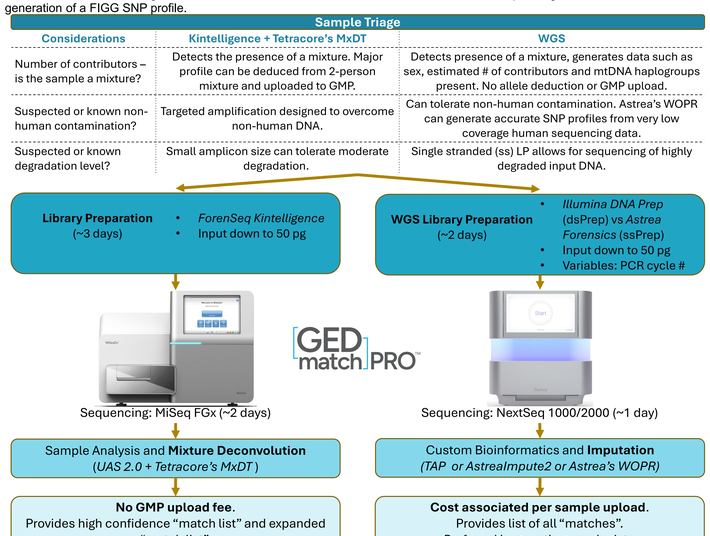
INTRODUCTION

Forensic Investigative Genetic Genealogy (FIGG) is driving innovation in advanced DNA techniques for solving recent problematic cases, cold cases, and unidentified remains investigations. Key areas of focus include assay chemistries, bioinformatics, computational advancements, and mixture resolution. Methodological transparency in SNP sequencing and data handling is essential for ensuring ethical practices and for supporting evidentiary reliability and investigative leads to aid case solvability.

This study evaluated 3 DNA library preparation (LP) kits (one targeted; two whole genome sequencing) with a focus on bioinformatic approaches, mixture resolution, and kinship detection after GEDmatch PRO query. Performance metrics such as SNP yield, quality, mixture deconvolution, kinship estimation and genealogical reach were analyzed. As labs assess assay strengths and limitations, comparison of workflows helps FIGG become more adaptable and reliable, reducing the unsolved case backlog and preventing future offenses. The findings provide real-world insight into using technology and bioinformatics to improve investigative outcomes across varied forensic sample types.

MATERIALS AND METHODS

Figure 1. Laboratory Processing Schematic. Processing decisions are driven by sample details such as quantity, quality and contributor number. Kintelligence and WGS offer benefits under different circumstances. Sample triage is critical to the successful generation of a FIGG SNP profile.



All data were generated from an individual donor DNA that has 22 known family members in the GEDmatch database, opted-in for law enforcement purposes. The degree of relation is known for 12 of those relatives, out to third cousin once removed (3C1R); the degree of relation is not confirmed for the additional 10 putative relatives. This study was conducted under IRB approval.

Sample types included rootless hairs (extracted using the Astrea rootless hair protocol and the extraction protocol described in Begg et al.[1] and quantified using the Qubit 4.0 Fluorometer), a buccal swab extracted using the Prepfiler Express BTA™ Forensic DNA Extraction Kit on the Applied Biosystem Automate Express™ DNA Extraction System and quantified using the Quantifiler Trio Kit on the Applied Biosystems QuantStudio™ 5 Real Time PCR System.

DNA input down to 50pg was analyzed and the extended family served as a testbed for assessing sensitivity and specificity using the 'shared cM' metric in GEDmatch PRO (GMP) across variables such as PCR cycles, NextSeq 1000/2000 read lengths and three FIGG-compatible bioinformatic pipelines for WGS SNP calling, QC and imputation as follows: Tetracore Analysis Pipeline (TAP; a custom in-house tool using GLIMPSE[2]), AstreaImpute2 and Astrea's WOPR (Astrea Forensics). Additionally, mock two-person mixtures (~1:3 & ~1:9 ratios; using the donor and a known female source) were processed through Tetracore's validated Kintelligence mixture deconvolution tool (MxDT).

RESULTS & DISCUSSION

A total of ten (10) Kintelligence and nine (9) whole genome sequencing events were completed using DNA from the known donor. The resulting FIGG SNP profiles were uploaded to GMP and queries performed using the one-to-many tool. Details are shown in the following table. AT = analysis threshold for allele calling.

Workflow	Details	Allele Call Rate / SNP Coverage	FIGG Results
Kintelligence (n=10)	1.5 % AT Input ranged from 4 ng to 50 pg	Call rates ranged from 92.4 % (50 pg input) to 97.9 % (4 ng input), 9,453 & 10,015 total SNPs, respectively, typed out of the 10,231 total SNPs targeted.	All FIGG SNP profiles, down to 50 pg input, detected eight relatives out to 1C1R, in the expected order and within the expected shared cM ranges.
WGS (n=9)	Input ranged from 1 ng to 50 pg PCR cycles assessed: 9-13	Depth of coverage across human genome ranged from 2.83 – 21.1X. Rootless hair samples from the donor generated depth of coverage ranging from 0.58 – 2.32X.	All FIGG SNP profiles detected eight relatives out to 1C1R. Beyond 1C1R, more distant relatives are identified, but with increased variability in rank order and in shared cM, as compared to Kintelligence outcomes.

Bioinformatic Pipeline Comparison:

- All three rootless hair datasets were analyzed with Astrea's WOPR pipeline. WOPR-generated FIGG SNP profiles detected the 8 known relatives out to 1C1R, in expected order.
- The additional 14 relatives were consistently detected for rootless hairs, but in variable order.
- WOPR-generated data produced an increased shared cM as compared to all other data analyzed in this study (Figure 4), including those from higher quality input DNA as compared to the rootless hair.
- One rootless hair dataset was analyzed with AstreaImpute2. AI2 detected the 8 known relatives in expected order & 12 of the 14 additional relatives in variable order (Figure 4).
- Prior analysis of the lowest coverage rootless hair (0.58X) did not produce a FIGG SNP profile. WOPR demonstrated (1) ability to produce a successful GMP upload from previously unusable data and (2) the most accurate FIGG SNP profiles, even from the most challenging sample type and from "very low" depth sequencing data.

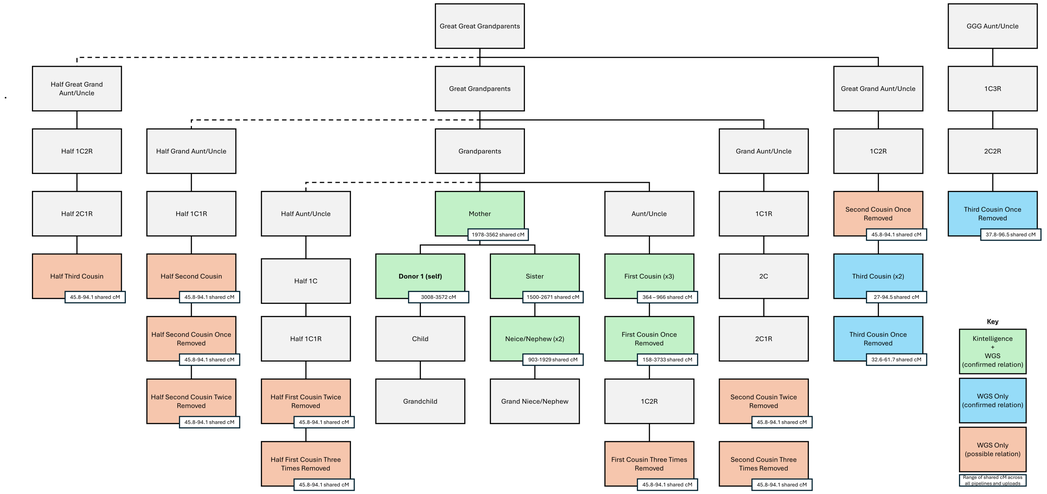


Figure 2. Comparison of GMP outcomes between Kintelligence and Whole Genome Sequencing methods. The relatives highlighted in green and blue (and self) represent known relatives with array-based data, opted in for law enforcement query in GMP. An additional 10 distant, known relatives with array-based data were also available for GMP query. For these 10 relatives, the exact degree of relation is unknown, but likely extends out beyond 7th degree. Relatives highlighted in orange represent estimated relationships to these more distant relatives as returned in the WGS GMP queries. The range of shared cM below the boxes represents the WGS GMP query results across nine sequencing events.

Mixed Sample Analysis:

Several mock mixed samples were prepared using DNA from the donor and processed through WGS and Kintelligence methods. Both workflows detected the presence of a mixed sample and informed downstream analysis. FIGG SNP profiles were not generated from WGS data determined to contain more than one contributor. Tetracore developed and validated a Kintelligence Mixture Deconvolution Tool (MxDT) to deduce major alleles from two-person mixed samples that contain a major contribution of ≥ ~60%. The major profile was deduced with and without consideration of the known minor contributor profile. The MxDT-generated major alleles were queried on GMP and successfully detected all 8 expected/known relatives out to 1C1R, in the expected order and within the expected shared cM range (Figure 3), consistent with single source results.

1:3 F:M Mixture - Deduced Major Contributor Expected & Observed Centimorgans

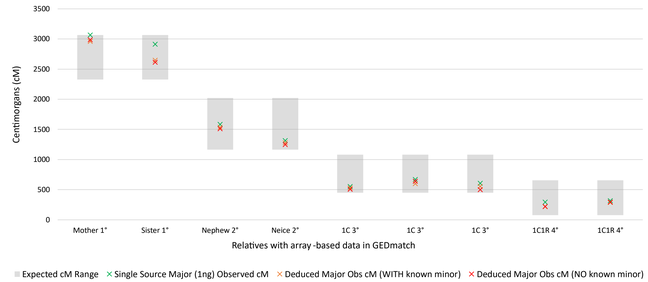


Figure 3. GMP outcomes for a Kintelligence-deduced major profile generated with & without consideration of the known minor profile. The 8 known relatives were detected within the expected shared cM ranges for the degree of relation. The gray bars represent the expected range of shared cM for the known degree of relation.

GMP Outcomes

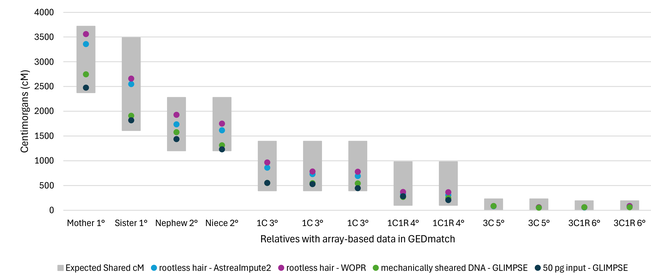


Figure 4. GMP outcomes for WGS data. Across the two WGS library preparation methods and three pipelines, WOPR produced the highest shared cM values, even from rootless hairs, one of the most challenging sample types. In some instances, the observed increased shared cM metric from Astrea's WOPR analyses placed the shared cM value outside the overlap range with the next degree of relation, which can provide less ambiguous estimations for the IGG research phase of casework. The gray bars represent the expected range of shared cM for the known degree of relation.

CONCLUSIONS

These studies underscore that thoughtful workflow selection during sample triage and flexible bioinformatic strategies, in combination, are critical for the success of challenging samples.

1. Throughout this study, WGS library prep and analysis pipelines achieved greater genealogical reach, compared to Kintelligence targeted sequencing, with Kintelligence reliably detecting 4th degree relatives and WGS reliably detecting at least 7th degree.
2. The Kintelligence MxDT successfully resolved major contributor alleles and produced GMP candidate relative lists for mock mixtures and sexual assault casework evidence. Further MxDT development is anticipated to support analysis of increasingly challenging mixture ratios, DNA inputs & minor contributors, as well as enabling WGS-generated SNP mixture deconvolution.
3. Astrea's WOPR analysis tool provides highly accurate FIGG SNP profiles from very low coverage data, providing genetic genealogists with refined data for IGG research. From these studies, Tetracore developed and validated a menu of sample processing techniques and analysis tools that allow for a fully customized workflow on a sample-by-sample basis. The benefit of flexibility and a well-established pool of strategies allows for refined attention to specific details of each unique forensic sample. These approaches have been used to successfully process ancient DNA samples such as highly degraded and bacterially-contaminated bones that previously were unsuccessful in both targeted NGS and WGS.

ACKNOWLEDGEMENTS

This work was supported by Tetracore, Inc. and Astrea Forensics. The authors thank the donor and their family members, as well as the GMP team at Qiagen, without whom this research would not be possible.

REFERENCES

1. Begg, *et al*., Genomic analyses of hair from Ludwig van Beethoven, *Current Biology* (2023), https://doi.org/10.1016/j.cub.2023.02.041
2. Rubinacci, S. *et al*., Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021), https://doi.org/10.1038/s41588-020-00756-0 Available on github: https://odelaneau.github.io/GLIMPSE/glimpse1/