Technical Note

VEROGEN

SNP Typing in Universal Analysis Software and Kinship Estimation with GEDmatch PRO

Matched toolsets, flexible thresholds, and an integrated workflow promote long-range kinship outcomes.

Highlights

- **Compare data in a dedicated forensic portal**
  Genomic data for investigating violent crimes.
- **Flexible options for kinship estimation**
  Results compatible with segment-based methods.
- **Partner with committed data stewards**
  The only solution for sites concerned about data privacy.

Introduction

Forensic genetic genealogy (FGG) has proven instrumental in solving cold cases, including missing persons identification, exonerations for innocence projects, and sexual assaults and other violent crimes. To date, FGG has aided more than 200 cases. Current methods of generating genomic data for FGG, such as whole-genome sequencing (WGS) and microarray-based genotyping, are constrained by DNA input and quality requirements aligned to research and clinical samples instead of forensic samples, which are typically low-quantity and degraded. The ForenSeq® Kintelligence Kit overcomes this barrier, offering a targeted sequencing solution for low DNA inputs and highly degraded DNA samples. When paired with the ForenSeq Kintelligence Analysis Module in Universal Analysis Software (UAS) and GEDmatch® PRO, the kit provides a workflow that generates new investigative leads to help solve violent crimes and missing persons cases when other options have failed.

UAS enables single nucleotide polymorphism (SNP) typing, guided exploration, rich visualization, and meticulous reviews of allele calls and generates human-readable reports. GEDmatch PRO accepts data from targeted sequencing, WGS, and microarrays to enable long-range kinship estimation. It includes matched toolsets that apply a nonsegment-based method of kinship estimation, the One-to-Many Kinship tool, to support analysis of targeted sequencing data alongside the traditional, segment-based method that supports data from WGS and microarrays. Importantly, segment-based tools in GEDmatch PRO and GEDmatch can analyze the match candidates that One-to-Many Kinship generates. This technical note describes how the nonsegment-based Verogen method provides robust outcomes for forensic DNA samples.

Rapid, confident SNP calls and reporting

The ForenSeq Kintelligence Analysis Module analyzes the targeted sequencing data from libraries prepared with the ForenSeq Kintelligence Kit. When sequencing is complete, the MiSeq FGx® Sequencing System automatically transfers raw base calls to UAS as BCL files. UAS converts the base calls into sequence reads in FASTQ file format. The reads are then demultiplexed and assigned to the appropriate sample based on the index adapter sequences specified in the sample sheet. The use of Unique Dual Indexes (UDIs) improves demultiplexing efficiency and optimizes data recovery. Low-quality reads are trimmed and the FASTQ files are aligned against ForenSeq Kintelligence amplicons to generate alignment files in BAM file format. The alignment step uses both sequencing reads, which improves the sensitivity of SNP allele calls. The BAM files reference the Genome Reference Consortium Human Build 38 (hg38).[1,2]

UAS evaluates SNP allele calls based on the total number of reads and classifies alleles as references or alternatives, providing locus call rate and heterozygosity data with quality control (QC) indicators to facilitate data review. Estimates of contributor status and gender indicate whether each sample is a mixture and the biological sex. When analysis is complete, users can manually inspect each SNP call or use the multi-level filtering and sorting options to focus the review on SNP metadata, calls flagged with QC indicators, and specific SNP categories (Table 1). This combination of filter and sort capabilities minimizes the time required for data review.

To supplement the guided exploration of onscreen results, the software offers four types of reports for packaging SNP data:

- Phenotype and Ancestry—Estimates of hair color, eye color, and biogeographical ancestry.
- GEDmatch PRO—SNP genotype calls configured for secure upload to the database.
- Project—Results for all samples in a project compiled in one report.
- Sample—Allele calls and read depth for each SNP type.

Figure 1 summarizes the complete UAS analysis process, from BCL file conversion through data review and reporting. The *ForenSeq Universal Analysis Software v2.0 Reference Guide (document # VD2019002)* provides further detail on how the software estimates contributor status and gender.

Determining thresholds and guidelines for robust SNP calling

UAS provides a default analysis method, the Verogen Kintelligence Analysis Method, that calls and evaluates SNPs for ForenSeq Kintelligence libraries. The method includes recommended settings that Verogen based on extensive testing (Table 2). Laboratories can implement this method or use it as a template, configuring the settings based on their internal assessments of the ForenSeq Kintelligence Kit. Configurable settings include analytical threshold (AT), interpretation threshold (IT), intralocus balance, and which loci to analyze. The AT provides the lower limit of allele calling by multiplying the AT percentage by the total number of reads at a locus, so only calls at or above the AT are visible. Setting the IT above the AT provides an additional threshold.

Table 1: Filters and sort options

Feature	Options
Metadata filters	Chromosome
	Typed or Untyped
	Homozygous or Heterozygous
QC indicator filters	Allele Count
	Analytical Threshold
	Imbalanced
	Interpretation Threshold
	No QC Indicators
	Not Detected
	Unexpected Allele
	User Modified
SNP type filters	Ancestry SNP
	Identity SNP
	Kinship SNP
	Phenotype SNP
	X-SNP
	Y-SNP
Sorting	Allele Count
	Amplicon Size
	Chromosome
	ILB
	Intensity
	SNP Name
	SNP Type

UAS also includes guidance for the sample read count, which is a metric users can access from the sample representation bar chart presented on the user interface. For each sample in a run, the bar chart displays the number of reads (intensity). Additionally, users can create multiple analysis methods and reanalyze samples using the different methods. UAS preserves the data from all analysis methods so users can easily switch between methods and review results with a variety of settings applied.

VEROGEN

Analytical and interpretation thresholds

To determine the default setting of 3% for both AT and IT, 10 operators performed 25 runs on the MiSeq FGx System using the MiSeq FGx Reagent Kit. Each run sequenced three libraries, which is the recommended plexity, prepared with the ForenSeq Kintelligence Kit. Data from 31 1 ng positive amplification controls and 21 negative amplification controls were analyzed using various AT and IT settings with the aim of identifying the combination that maximizes concordant allele calls and minimizes discordant allele calls. The lowest combination Verogen evaluated was 1.5% AT and 1.5% IT, which represents a minimum read count of 10. The increase to 3% raises the minimum read count to 20 and results in an overall positive effect on concordant call rates and a negligible effect on the detection of true allele calls (Table 3). Given these results, Verogen recommends settings that call fewer SNPs rather than introduce potential errors into kinship estimation.

Sample read count guideline

To determine the sample read count guideline of 15 million reads per sample, Verogen evaluated the total number of reads obtained from 29 samples at 1 ng input. Libraries were prepared with the ForenSeq Kintelligence Kit and sequenced on the MiSeq FGx System using the MiSeq FGx Reagent Kit. Each run sequenced three libraries. Samples that reached 15 million reads demonstrated a median call rate of ≥ 99.8% and a minimum call rate of 98.6%. However, samples that do not reach 15 million reads might still generate sufficient data for analysis.

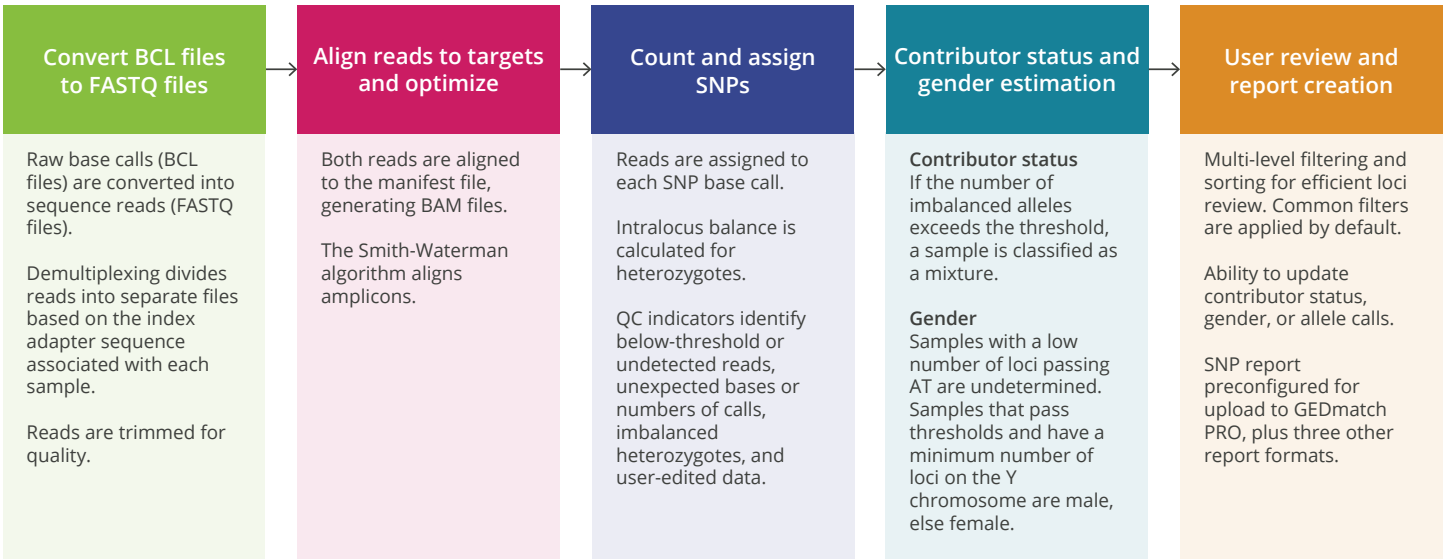

Figure 1: UAS automatically performs SNP calling, completing analysis in less than one hour. Single-click reporting produces a GEDmatch PRO-ready report.

Table 2: Thresholds and guidelines in the Verogen Kintelligence Analysis Method

Threshold or guideline	Description	Default Value	Configurable
Analytical threshold	The value that a read count must reach for the software to type an allele.	3%	Yes
Interpretation threshold	The value that an allele must reach to contribute to a call.	3%	Yes
Intralocus balance	The balance of read counts between typed alleles at a heterozygous locus.	50%	Yes
Sample representation	The number of reads per sample for a run, providing quantitative sample and run information.	15 million*	No

* The sample representation value is provided for guidance only, and is not a setting in the software.

Table 3: Results for different threshold settings

Metric	1.5% AT and IT	3% AT and IT
Average number of SNPs typed in the negative controls out of 10,230	33	16
Average number of discordant allele calls in the positive control*	13	0.4
Maximum number of discordant allele calls in the positive control	43	4
Average concordant allele call rate for kinship SNPs	99.9%	99.8%
Average concordant allele call rate for nonkinship SNPs	≥ 99.2%	≥ 99.2%
Range of concordant call rates	98.6–100%	98.6–100%

* PCR or sequence-based noise and drop-ins can cause discordant calls in positive controls.

Kinship estimation with dense SNP data

GEDmatch PRO is a dedicated forensic portal for kinship estimation in cases of unidentified human remains, missing persons, and violent crimes. It supports upload of SNP reports (kits) that WGS, microarrays, and targeted sequencing generate, enabling comparisons against profiles voluntarily shared in GEDmatch. Output from both segment-based and nonsegment-based comparisons is a list of candidate matches. Candidate matches from the nonsegment-based tool can be used as input for tree-building and other segment-based genealogy tools. Users can also leverage the segment-based tool to properly place associations on a family tree.

Segment-based comparisons

Traditional DNA-based genealogy methods generate a sizable number of SNP calls, typically three billion for WGS and 650,000–2,500,000 for microarrays, and employ a segment-based approach to determine genetic relatives. The DNA between SNP locations on a chromosome is called a segment. If two kits have similar SNP alleles between contiguous segments, they are considered a candidate. A centimorgan (cM) is a measure of genetic distance that denotes the size of matching DNA segments in autosomal DNA tests. Segments that have many centimorgans in common are more likely to be significant and to indicate a common ancestor within a genealogical timeframe, indicating biological inheritance.

Users can leverage the Shared cM Project or similar tools to evaluate the number of shared centimorgans, average segment length, and the longest segment across matches to understand the most likely relationships that the total number of shared centimorgans across chromosomes indicates. Different genealogy resources, such as census, family, birth, and death records, further refine these candidates. Evaluating the top match candidates, building and merging family trees, identifying the most recent common ancestor, and admixture analysis also help with genealogical assessments.[3,4]

Nonsegment-based comparisons

In contrast to traditional, segment-based methods, the Verogen method does not rely on overlap between long, contiguous segments to compare kits and generate matches. Instead, the One-to-Many Kinship tool leverages a variation of principal component analysis (PCA) to analyze the dense set of 10,230 forensically curated SNPs included in the ForenSeq Kintelligence Kit. Typically used in genome-wide association studies with genotype SNP data to infer and correct for population structure (i.e., ancestry interference), PCA takes SNPs as input and performs dimension reduction to calculate principal components (PCs) that capture variability in the data. The top PCs generally reflect population structure among the samples. However, when a population contains known or unknown relatives, PC is confounded by the family structure and presents clusters of close relatives instead of indicating variations across populations.[5,6]

To bypass this issue, Verogen applied a variation of PCA to kits opted in for law enforcement use. One-to-Many Kinship builds on published, peer-reviewed methods such as PC-Relate and PC-AiR and uses targeted sequencing data to infer population structure and make genetic correlations without the need for ancestry or reference population information. PC-Air accounts for relatedness in the population to provide ancestry estimations that are not confounded by family structure. Accordingly, the PCs that PC-AiR generates are robust to known or cryptic relatedness. When implemented on kits in GEDmatch, PC-AiR identifies mutually unrelated kits that are maximally ancestrally diverse.[7,8]

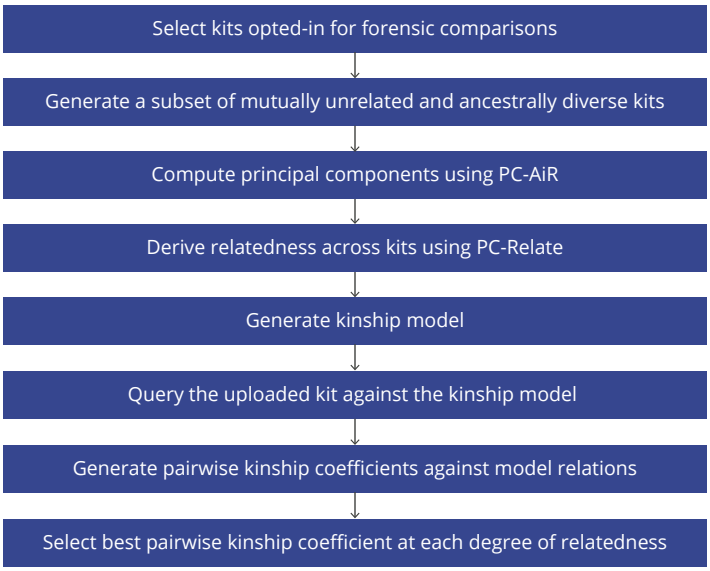

Figure 2: The Verogen method of kinship estimation yields a simple measure of relatedness.

Pairwise kinship coefficients, pairwise identical by descent (IBD) sharing probabilities, and individual inbreeding coefficients are among the current methods for estimating recent genetic relatedness. In the presence of population structure and ancestry admixture, however, these methods have limitations or require the appropriate reference population allele frequency. When population allele frequencies are absent, kinship estimates might be biased. PC-Relate is used to estimate measures of recent genetic relatedness in samples with an unknown or unspecified population structure without reference population allele frequencies, even when endogamy or consanguinity are present. PC-Relate identifies ancestry-representative PCs that adjust for family structure and generate relatedness estimates as kinship coefficients in the presence of population structure, admixture, and departures from the Hardy-Weinberg equilibrium. ForenSeq Kintelligence uploads are compared to all other kits in GEDmatch PRO to generate the list of candidate matches (Figure 2).[8,9]

Table 4: Expected kinship coefficients and associated degrees of relatedness

Degree	Relationship	Average Kinship Coefficient	Kinship Coefficient Range	Expected cM*	Expected cM Range
0	Self	0.5	0.484–0.514	3560	3560
1	Parent, child, sibling	0.251	0.196–0.31	3560	2787–3560
2	Grandparent, grandchild, aunt or uncle, niece or nephew, half-sibling	0.126	0.76–0.17	1799	1083–2471
3	Great-grandparent, great-grandchild, great-aunt or uncle, great-niece or nephew, half-aunt or uncle, half-niece or nephew, cousin	0.064	0.023–0.11	917	326–1566
4	Great-great-grandparent, great-great-grandchild, great-great aunt or uncle, great-great-niece or nephew, half great-aunt or uncle, half-great-niece or nephew, half-cousin, cousin once removed, cousin twice removed	0.034	0.064–0.076	480	91–1091
5	Great-great-great grandparent, great-great-great-grandchild, great-great-great aunt or uncle, great-great-great-niece or nephew, half great-great-aunt or uncle, half great-great-niece or nephew, second cousin, half-cousin once removed	0.019	-0.0009–0.059	265	0–840
6	Great-great-great-great aunt or uncle, cousin three times removed, half-cousin twice removed, half second cousin, second cousin once removed	0.011	-0.006–0.031	152	0–445
7	Third cousin, second cousin twice removed, half-cousin three times removed, half second cousin once removed	0.007	-0.009–0.025	152	0–445

* Expected cM = min(3560, 4 × max(0, kinship_coeff) × 3560

Leveraging a targeted SNP set and building on established methods with a history of use in forensic paternity cases and mass disaster victim identification (DVI), One-to-Many Kinship is less sensitive to typing errors or partial data compared to segment-based methods because it does not rely on continuous stretches of SNPs that are physically adjacent on a chromosome. This advantage makes One-to-Many Kinship more useful for processing crime-scene samples and unidentified human remains. By using the kits and pedigrees in GEDmatch as inputs to train and test the model, the Verogen method is optimized to identify informative relationships in GEDmatch PRO.

Flexibility for different methods of data generation

GEDmatch PRO is easy to use, displaying kinship coefficients with the equivalent centimorgan values for context and allowing users to switch between two established methods of assessing relatedness (Table 4). Because the segment-based approach uses centimorgans interpreted with the Shared cM Project, GEDmatch PRO displays a schematic adapted from this project. The schematic summarizes the likelihood of various relationships with associated kinship coefficients, allowing it to function as a conversion tool between centimorgan values and kinship coefficients. In general, the more distant a putative relative, the more likely a one-to-many query, segment-based or not, returns a centimorgan value or kinship coefficient that overlaps at least two orders of relatedness. A simple measure of relatedness, kinship coefficient is the probability that a set of randomly sampled alleles were inherited from the same ancestor.

Conclusion

Reserved for challenging cases, FGG provides an opportunity for resolution after traditional methods have failed. Granting law enforcement access to opted-in kits though GEDmatch PRO helps improve outcomes for homicides and sexual assaults. Investigators interact with GEDmatch PRO as an impartial tool to accelerate the generation of genealogically significant leads for forensic use. GEDmatch PRO minimizes intrusion by assessing fewer SNPs than GEDmatch and surfacing the results of consenting users only. Capabilities such as user provisioning and access control reduce the possibility of unintentionally sharing kits, allowing secure collaboration. Coupled with the power to process degraded, low-input samples and quickly and easily generate high-confidence SNP calls, One-to-Many Kinship caps an integrated sequencing solution that delivers on the potential of FGG without compromising quality or privacy.

References

1. Schneider, Valerie A., Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, et al., "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly," *Genome Research* 27, no. 5 (May 2017): 849–864.
2. Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al., "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature* 581 (May 2020): 434–443, https://doi.org/10.1038/s41586-020-2308-7.
3. Blaine T. Bettinger, "The Shared cM Project 4.0 tool v4," DNA Painter, updated March 26, 2020, https://dnapainter.com/tools/sharedcmv4.
4. Jonny Perl, "Introducing the updated shared cM tool," *DNA Painter Blog*. DNA Painter, March 27, 2020. https://dnapainter.com/blog/introducing-the-updated-shared-cm-tool/.
5. Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics* 38 (July 2006): 904–909, https://doi.org/10.1038/ng1847.
6. Heath, Simon C., Ivo G. Gut, Paul Brennan, James D. McKay, Vladimir Bencko, Eleonora Fabianova, Lenka Foretova, et al., "Investigation of the fine structure of European populations with applications to disease association studies," *European Journal of Human Genetics* 16, (November 2008): 1413–1429, https://doi.org/10.1038/ejhg.2008.210.
7. Conomos, Matthew P., Alexander P. Reiner, Bruce S. Weir, and Timothy A. Thornton "Model-free Estimation of Recent Genetic Relatedness," *American Journal of Human Genetics* 98, no. 1 (January 2016): 127–148.
8. Conomos, Matthew P., Michael B. Miller, and Timothy A. Thornton, "Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness," *Genetic Epidemiology* 39, no. 4 (May 2015): 276–293.
9. Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al., "The mutational constraint spectrum quantified from variation in 141,456 humans," *Nature* 581 (May 2020): 434–443, https://doi.org/10.1038/s41586-020-2308-7.

Verogen • 1.833.837.6436 toll-free (North America) • +1.858.285.4101 tel • +44 (0) 208 054 8706 (United Kingdom) info@verogen.com • www.verogen.com
 | Document # VD2020058 Rev. A | Current as of March 2021