Forensic Science International: Genetics Supplement Series 8 (2022) 112–114

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

# Evaluating probabilistic genotyping for low-pass DNA sequencing

Sammed N. Mandape [a,*], Kapema Bupe Kapema [a], Tiffany Duque [b], Amy Smuts [a], Jonathan L. King [a], Benjamin Crysup [a], Jianye Ge [a,b], Bruce Budowle [c], August E. Woerner [a,b]

[a] Center for Human Identification, University of North Texas Health Science Center, Fort Worth, TX, USA
[b] Department of Microbiology, Immunology and Genetics, University of North Texas Health Science Center, Fort Worth, TX, USA
[c] Department of Forensic Medicine, University of Helsinki, FI

ARTICLE INFO

*Keywords:*
Low-pass DNA
Genetic genealogy
Probabilistic genotyping
Genotype likelihood

ABSTRACT

Most genomic methods consider the sample genotype. Data are evaluated at some location, and if the signal strength is sufficient, a genotype call is made. Conversely, sites that lack sufficient signal are treated as missing data. Such methods for genotype calling are binary, and this dichotomy limits genomic analyses to relatively high-coverage (and high-cost) massively parallel sequencing (MPS) data. It follows that bioinformatic methods that rely on genotypes may not be ideal for trace DNA samples, such as those sometimes encountered in forensic investigations, but even when applicable such analyses can be expensive. However, there are some genomic analyses where having many uncertain genotypes (with measured uncertainty) assayed over the entirety of the genome may be more powerful than current multi-locus approaches that consider a limited number of well-characterized markers. Methods for such problems may rely on genotype likelihood, which expresses the likelihood of alternative genotype calls in addition to the most likely call. One application that can benefit from genotype likelihoods is kinship analysis. NgsRelate is a bioinformatic tool that infers pairwise relatedness using a probabilistic genotyping framework, which accommodates the uncertainty associated with genotype calls for low-pass MPS data. Here, NgsRelate was used to infer kinship coefficients from low-pass whole genome sequencing data from a known pedigree. Multiple samples in a titration series (ranging from 50 ng to 0.5 ng) on a single MPS S4 flow cell were assessed. A reproducible scientific bioinformatic workflow was developed to evaluate kinship coefficients considering up to 3rd degree relatives. NgsRelate was found to provide robust assessments of kinship. Further, the use of low-pass MPS data provides a more cost-effective way to conduct forensic investigations.

## 1. Introduction

The use of single nucleotide polymorphisms (SNPs) has emerged as a powerful approach in kinship analysis [1]. SNP profiles can be assayed using different technologies, commonly by whole-genome sequencing (WGS), targeted sequencing or microarray-based genotyping. Microarray-based genotyping is accurate, easy to genotype (with less demanding bioinformatic analysis), and inexpensive [2]. However, accurately determining the SNP genotype requires ample DNA, an infeasible requirement in many forensic investigations [1,2]. Targeted sequencing (for example, Verogen's ForenSeq Kintelligence kit) focuses on a smaller subset of SNPs and technique such as hybridization capture method lacks a large SNP panel that will be applicable in forensic genetics [2]. WGS may seem expensive, although, it offers much more refined and advanced workflow than other technologies [2]. Additionally, WGS potentially increases the detection power for distant relatives [3]. The cost of WGS analyses (a limiting factor in most cases) can be lowered, as shown in current study, by sequencing more than the recommended number of samples in a single run. However, low-pass sequencing data can lead to inaccurate genotype calls [4]. An alternative approach, inferring pairwise relatedness, employs genotype likelihoods rather than genotype calls (see [5,6]). In this study, the genotype likelihood based KING-robust estimator [7], implemented in NgsRelate v2 [5], was used to infer kinship coefficients from a low-pass WGS data from a known pedigree. NgsRelate is a bioinformatic tool that can infer pairwise relatedness using a probabilistic genotyping framework, thus, it is well-suited to low-pass massively parallel sequencing (MPS) data.

* Corresponding author.
*E-mail address:* sammed.mandape@unthsc.edu (S.N. Mandape).

https://doi.org/10.1016/j.fsigss.2022.10.001
Received 19 September 2022; Accepted 5 October 2022
Available online 7 October 2022

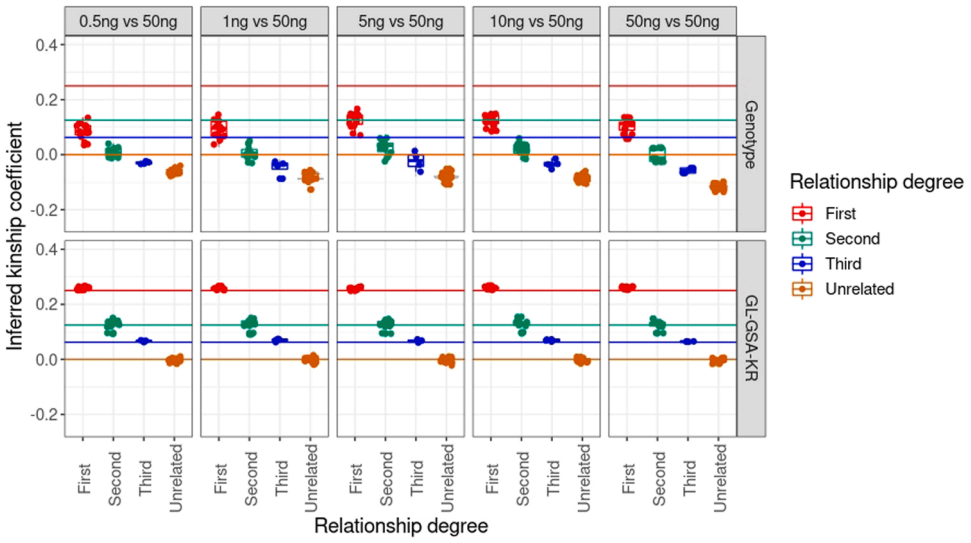


**Fig. 1.** Inferred kinship coefficients (y-axis and colors) are contrasted against the relationship degree (x-axis and colored horizontal lines). Kinships coefficients were estimated using two approaches (outer rows): genotype calls from WGS and from SNPs pre-selected to be variable in the population and considering the genotype likelihood (GL-GSA-KR). Kinship coefficients were inferred for serially diluted DNA input from 0.5 ng to 50 ng (outer columns) over a range of expected relationships.

## 2. Methods

In this study, WGS was performed on forty samples selected from a known three-generation pedigree. The DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. Eight samples were serially diluted from 50 ng to 0.5 ng (50 ng, 10 ng, 5 ng, 1 ng, 0.5 ng in NA13050, NA13047, NA07013, NA07035, NA07028, NA07437, NA07046, NA13053). Illumina DNA Prep-Kit was used to prepare DNA libraries according to the manufacturer's protocol. Briefly, DNA samples were tagmented and amplified using unique dual indexed adapters from the IDT®. The amplified forty libraries were purified, quantified, normalized, pooled (2 nM), and sequenced on a NovaSeq 6000 (S4 Reagent Kit v1.5, 300 cycles).

FASTQ files from the NovaSeq were mapped to the GRCh38 reference genome and sorted/indexed BAM files were created [8,9]. Optical and PCR duplicates were marked [10], indels were realigned [10], and base-quality scores were recalibrated [10]. Variant calling was performed with BCFtools [9] on pairs of BAM files, generating VCF files with both genotypes and genotype likelihoods. Problematic regions exist in the genome, for example repetitive regions may be under-represented relative to the actual sequence present [11]. These regions lead to likely artifacts in genome assemblies and inaccurate interpretation [11]. Such regions were identified and removed. VCF files were annotated with non-Finnish European population allele frequencies from gnomAD [12] and optionally down-sampled to only include well-described sites. For the latter, the union of sites in Illumina's Infinium Global Screening Array (GSA), a sparse set of 'reliable' SNPs useful for kinship analysis [13] and SNPs from Verogen's ForenSeq Kintelligence system were taken. Collectively these sites are termed GSA-KR. Kinship coefficients were estimated using either the genotype [14] or the genotype likelihood based estimator of the KING-robust algorithm [7] as implemented in NgsRelate (v2). The root-mean squared error (RMSE) was computed by comparing the observed (inferred) versus the expected kinship coefficient.

## 3. Results and discussion

### 3.1. WGS throughput

A total of 15.3 billion reads was generated. Of the total reads, 12% were optical and/or PCR duplicates. Coverage was estimated using 10,000 random positions from non-duplicate reads. Coverage ranged from 3.3 × to 8 × across DNA inputs (0.5 ng – 50 ng). Coverage was variable both within and between subjects. For instance, NA07028 had 3.3 × and 6 × coverage for 1 ng and 50 ng of DNA, respectively, while NA13053 showed roughly the opposite (6.5 × and 3.5 × for the same DNA inputs). No clear association between the amount of DNA used and read-depth was apparent, as may be expected given library normalization [15].

### 3.2. Kinship estimation

Kinship coefficients were estimated comparing all 50 ng samples to all samples (0.5 ng to 50 ng) so as to emulate the forensic use-case. Kinship inferred using genotype likelihood down-sampled to only GSA-KR (herein, GL-GSA-KR) sites were more accurate (RMSE 0.011) than from genotypes assessed across the whole genome (herein, genotypes, RMSE 0.116). Using genotypes also led to systematically lower kinship coefficient estimates for all the degrees of relatedness, including the unrelated class (Fig. 1). Only considering genotypes (and not genotype likelihoods) with GSA-KR SNPs showed comparable levels of error to the whole genome SNP assessment (RMSE 0.115), suggesting a benefit to using genotype likelihoods in conjunction with SNPs that are well-characterized.

## 4. Conclusion

Accurately and cost-effectively estimating kinship coefficients is essential in forensic investigations. With advances in MPS technologies, WGS is a promising approach. However, costs associated with WGS and the resulting uncertain genotypes from low-pass sequencing data remain a challenge. This study demonstrates multiple samples sequenced together can be a cost-effective approach to WGS. Additionally, selecting sites known to vary in the population (e.g., GSA-RK SNPs) can be used to improve estimates of kinship from low-pass sequencing data.

## Conflict of interest statement

None.

## Acknowledgments

This research was supported in part by award 2019-DU-BX-0046 awarded by the National Institute of Justice, Office of Justice

Programs, U.S. Department of Justice and by internal funds from the Center for Human Identification.

## References

[1] K. Yagasaki, et al., Practical forensic use of kinship determination using high-density SNP profiling based on a microarray platform, focusing on low-quantity DNA, Forensic Sci. Int. Genet. (2022), 102752.
[2] D. Kling, et al., Investigative genetic genealogy: current methods, knowledge and practice, Forensic Sci. Int. Genet. 52 (2021), 102474.
[3] H. Li, et al., Relationship estimation from whole-genome sequence data, PLoS Genet. 10 (2014), e1004144.
[4] R. Nielsen, et al., SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data, PLoS One 7 (2012), e37558.
[5] T.S. Korneliussen, et al., NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data, Bioinformatics (2015) btv509.
[6] A.K. Nøhr, et al., NGSremix: a software tool for estimating pairwise relatedness between admixed individuals from next-generation sequencing data, G3 Genes Genomes|Genet. (11) (2021) jkab174.
[7] R.K. Waples, et al., Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data, Mol. Ecol. 28 (2019) 35–48.
[8] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013), https://doi.org/10.48550/ARXIV.1303.3997.
[9] P. Danecek, et al., Twelve years of SAMtools and BCFtools, GigaScience 10 (2021) giab008, https://doi.org/10.1093/gigascience/giab008.
[10] A. McKenna, et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303.
[11] H.M. Amemiya, et al., The ENCODE Blacklist: identification of problematic regions of the genome, Sci. Rep. 9 (2019) 9354.
[12] K.J. Karczewski, et al., The mutational constraint spectrum quantified from variation in 141,456 humans, Nature 581 (2020) 434–443.
[13] R. Arthur, et al., AKT: ancestry and kinship toolkit, Bioinformatics 33 (2017) 142–144.
[14] A. Manichaikul, et al., Robust relationship inference in genome-wide association studies, Bioinformatics 26 (2010) 2867–2873.
[15] S. Riman, et al., Understanding the characteristics of sequence-based single-source DNA profiles, Forensic Sci. Int. Genet. 44 (2020), 102192.