k2v: A Containerized Workflow for Creating VCF Files from Kintelligence Targeted Sequencing Data

Stephen D. Turner
Signature Science, LLC
Charlottesville, Virginia

Michelle A. Peck
Signature Science, LLC
Charlottesville, Virginia

November 21, 2022

Abstract

The ForenSeq Kintelligence kit developed by Verogen is a targeted Illumina sequencing assay that genotypes 10,230 single nucleotide polymorphisms designed for forensic genetic genealogy, forensic DNA phenotyping, and ancestry inference. We developed `k2v`, a containerized workflow for creating standard specification-compliant variant call format (VCF) files from the custom output data produced by the Kintelligence Universal Analysis Software. VCF files produced with `k2v` enable the use of many pre-existing, widely used, community-developed tools for manipulating and analyzing genetic data in the standard VCF format. Here we describe the `k2v` implementation, demonstrate its usage, and use the VCF produced by `k2v` to demonstrate downstream analyses that can easily be performed with pre-existing tools using VCF data as input: concordance analysis, ancestry inference, and relationship estimation. `k2v` is distributed as a Docker container available on Docker Hub. Documentation and source code for `k2v` is freely available under the GNU Public License (GPL-3.0) at `https://github.com/signaturescience/k2v`.

Keywords: Forensic genetics · Variant call format · VCF · Forensic genetic genealogy · Kintelligence

1 Introduction

Kinship analysis using single nucleotide polymorphisms (SNPs) is a cornerstone of forensic genetic genealogy (FGG) analysis (Greytak, Moore, and Armentrout 2019; Kling and Tillmar 2019; Glynn 2022). Originally used by hobbyist and professional genealogists, databases populated with direct-to-consumer DNA tests such as GEDmatch and Family Tree DNA are now widely used by law enforcement agencies to generate investigative leads on unknown crime scene samples or unidentified human remains by finding distant relatives of the unknown DNA sample. The technique has gained enough traction in the community to prompt the Federal Bureau of Investigation (FBI) to call for self-regulation and development of guidelines for practicing responsible genetic genealogy (Callaghan 2019).

FGG routinely relied on genome-wide SNP data using whole genome sequencing or microarrays (Russell et al. 2022; de Vries et al. 2022; Davawala et al. 2022; Tillmar et al. 2020). However, targeted sequencing assays have recently been developed that leverage a sparse set of SNPs for kinship analysis, including the community-developed FORCE panel (Tillmar et al. 2021) and the commercially available ForenSeq Kintelligence kit (Verogen, Inc., San Diego CA), which we internally validated (Peck et al. 2022).

The Kintelligence kit provides genotypes on 10,230 SNPs, of which 9,867 are kinship-informative SNPs used for SNP-based kinship analysis across populations, while the remaining 363 are used to inform biogeographical ancestry, identity, hair and eye color, and biological sex (Snedecor et al. 2022). The Kintelligence kit integrates data analysis in a browser-based application, and also produces Excel and text file outputs containing genotype data in a custom format.

The Variant Call Format (VCF) is the *de facto* standard file format for storing genetic variant data that has been used by genetics and bioinformatics community for over a decade (P. Danecek et al. 2011). A vast ecosystem of widely used, community-developed, and community-supported tools exist for managing and analyzing variant data in VCF files including tools for general data manipulation and analysis (Petr Danecek et al. 2021), and tools specifically designed for forensic genetics-focused analysis (Turner et al. 2022; V. P. Nagraj et al. 2022b).

In this paper we describe a new tool, `k2v`, for converting Kintelligence data into a standard specification-compliant VCF, which enables the use of a vast pre-existing infrastructure for genetic data manipulation and analysis. The `k2v` workflow is implemented as a Docker container, and can be run on any system where Docker or Singularity is available. We provide a demonstration of the `k2v` workflow and several example forensic genetics-relevant analyses that can easily be accomplished using standard tools that use VCF data as input.

2 Implementation

`k2v` is implemented as a Docker container and is distributed to run directly from a command line interface. The `k2v` Docker image has bcftools (Petr Danecek et al. 2021), htslib (Bonfield et al. 2021), and R (R Core Team 2017) installed on a lightweight Alpine Linux base image. Running the `k2v` container instantiates a script inside the container that takes as input the custom `*.SnpResult.txt` file that is produced by Kintelligence, joins that data to a table of reference and alternate alleles obtained by extracting Kintelligence sites from the GnomAD site VCF (Karczewski et al. 2020) to create an intermediate tabular file, which is then converted to VCF using bcftools, and written out to the host filesystem after being compressed with bgzip and indexed with tabix. Documentation and implementation details are available in the project repository at `https://github.com/signaturescience/k2v`.

3 Use cases

3.1 Validation data analysis

With only the custom format text file produced by the Kintelligence platform, custom analysis workflow development would be required for basic data analysis tasks such as call rate, concordance, and heterozygosity analysis that may be used in a validation study. Instead, we can convert the Kintelligence data into VCF then use pre-existing tools for data analysis to streamline analysis.

In this example we convert Kintelligence data generated on the NIST Standard Reference Material sample HG002 (NA24385) to VCF using `k2v`. We first use bcftools (Petr Danecek et al. 2021) to calculate simple statistics on the VCF. We then use the previously published `nrc` tool (V. P. Nagraj et al. 2022b) to get a detailed analysis of how the Kintelligence genotype data differs from the benchmark callset from NIST/Genome-in-a-Bottle. Additional details on `k2v` usage and getting example data can be found on the `k2v` GitHub README. This analyses required less than one second of compute time. Table 1 shows the count of the number of substitutions and variant types in sample HG002. Table 2 shows a subset of results from running the `nrc` tool on the VCF produced by `k2v`, comparing the Kintelligence VCF to the GIAB high-confidence call set.

```
# Convert kintelligence data (HG002.SnpResult.txt) to VCF using the k2v container
docker run -v $(pwd):$(pwd) -w $(pwd) sigsci/k2v HG002.SnpResult.txt
# Simple statistics with bcftools
bcftools stats HG002.vcf.gz
# Use the nrc docker container to calculate concordance stats with GIAB sample data
docker run -v $(pwd):$(pwd) -w $(pwd) sigsci/nrc HG002.giab.vcf.gz HG002.vcf.gz
```

3.2 Ancestry inference

The use of principal components analysis (PCA) for studying genetic variation data was introduced over 40 years ago (Menozzi, Piazza, and Cavalli-Sforza 1978). Since its introduction PCA has been widely used in population and medical genetics for biogeographic ancestry analysis (Patterson, Price, and Reich 2006; Novembre et al. 2008; Li et al. 2008; Wellcome Trust Case Control Consortium 2007; Tian et al. 2008; Price et al. 2008; Shriver and Kittles 2004), and has recently gained in the forensic genomics community for statistical inference of an unknown DNA donor's ancestry (Phillips 2015).

There are several existing tools for conducting PCA with genetic variant data including EIGENSTRAT (Price et al. 2006) and PLINK (Purcell et al. 2007), both widely used and well-supported by the genetic epidemiology community. The Ancestry and Kinship Tools (AKT) developed by Illumina is a permissively licensed open-source implementation of PCA (and other analyses relevant for forensic genomics), which uses the htslib API allowing it to work directly with VCF files.

In this example we use AKT to perform PCA on the HG002 VCF produced by Kintelligence against the 2,504 individuals in the 1000 Genomes Project phase 3 release data. In this example we start with the 1000 Genomes data on the kintelligence sites, merge in the HG002 VCF produced by `k2v`, then run PCA using AKT. We read in the output into R, annotate the 1000 Genomes samples with continental ancestry (Turner 2022). This analysis required less than 10 seconds of compute time. Figure 1 shows a PCA biplot showing the first two principal components, with HG002 highlighted against the background of 1000 Genomes samples.

```
# Merge HG002 with the 1000 Genomes Project reference samples using bcftools
bcftools merge 1000g.kintelligence.vcf.gz HG002.vcf.gz | bcftools sort -Oz -o pca.in.vcf.gz
tabix -f pca.in.vcf.gz
# Conduct a principal components analysis using AKT
akt pca --force pca.in.vcf.gz -Oz -o pca.out.vcf.gz > pca.txt
```

3.3 Relationship estimation

Relationship inference methods using genome-wide SNP data typically fall into two broad classes of methods: measures that directly infer relatedness genome-wide, and identity by descent (IBD) segment approaches (Ramstetter et al. 2017). These approaches to relationship inference have been benchmarked using data generated in ideal conditions (Ramstetter et al. 2017), and more recently assessed using low-quality microarray data on forensic samples (Turner et al. 2022) or with low-coverage whole genome sequencing data followed by imputation (V. P. Nagraj et al. 2022a).

In this example, we use the KING-robust estimator (Manichaikul et al. 2010) implemented in the PLINK 2 software (Chang et al. 2015) to estimate relatedness between two Ashkenazi Jewish individuals from the Personal Genome Project: HG002 (NA24385; the son), and HG004 (NA24143, the mother), using VCFs produced by `k2v` from Kintelligence data. This analysis required less than 50 milliseconds of compute time.

```
# Convert each Kintelligence text file to VCF
docker run -v $(pwd):$(pwd) -w $(pwd) sigsci/k2v HG002.SnpResult.txt
docker run -v $(pwd):$(pwd) -w $(pwd) sigsci/k2v HG004.SnpResult.txt
# Merge with bcftools
bcftools merge HG002.vcf.gz HG004.vcf.gz > merged.vcf.gz
# Estimate relatedness
plink2 --make-king-table --vcf merged.vcf.gz
```

The kinship coefficient between these two samples as inferred using the KING-robust estimator is $\phi = 0.2384$, which clearly indicates that this sample pair is a first-degree relationship. The proportion of loci where zero alleles are shared identical by state is $\kappa_0 = 0.0039$. Because parent-child offspring are expected to share $\kappa_{0,1,2} = (0, 1, 0)$ while full siblings are expected to share $\kappa_{0,1,2} = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$, we correctly infer that HG002 and HG004 are a parent-offspring pair.

4 Conclusion

To use data from a sequencing or genotype assay that produces variant calls in a custom format requires either (a) custom development of new analysis tools for each downstream analysis task, or (b) development of a single tool to convert data to a common format for which tools already exist. `k2v` takes the latter approach, enabling the rapid conversion of Kintelligence variant data in a custom format to a standard specification-compliant VCF file. Once data exists in a VCF, nearly any downstream analyses and manipulation tasks can readily be achieved with existing infrastructure. `k2v` is currently limited to SNPs only, and cannot capture the one insertion that Kintelligence targets (rs796296176).

`k2v` is freely available under the GNU Public License (GPL) v3 at `https://github.com/signaturescience/k2v` with further documentation on the repository README. A pre-built Docker image is available on Docker hub and can be installed via `docker pull sigsci/k2v`.

Table 1: Count of the number of substitutions and variant types in sample HG002.

Substitution / variant type	Count
A>C	288
A>G	1,439
A>T	4
C>A	289
C>G	4
C>T	1,614
G>A	1,515
G>C	2
G>T	302
T>A	1
T>C	1,424
T>G	307
Homozygous REF	2,897
Homozygous ALT	2,590
Heterozygous	4,599
Transitions	5,992
Transversions	1,197
Missing	55

Table 2: Subset of output from the `nrc` tool comparing the Kintelligence data converted to VCF using `k2v` against the high-confidence GIAB callset from HG002. The table shows the total number of biallelic SNPs compared between the two samples, discordance and concordance statistics, non-reference discordance (NRD), non-reference concordance (NRC), and the full array of each type of genotype match and mismatch. The vector of match/mismatch counts can be interpreted as follows: RRRR = number of genotype matches where the GIAB and Kintelligence calls are both homozygous for the reference allele; RRRA = number of genotype mismatches where the GIAB call is homozygous reference (Ref/Ref), but the Kintelligence VCF is heterozygous (Ref/Alt); RARA = number of genotype matches where both the GIAB callset and the Kintelligence VCF are both heterozygous (Ref/Alt); etc. See (V. P. Nagraj et al. 2022b) for definitions of NRD and NRC. Note that the total number of SNPs compared is fewer than the number of SNPs assayed by Kintelligence – this is because the GIAB call set only includes autosomes, and the high-confidence regions do not cover all the Kintelligence-assayed sites.

Metric	Value
Total	9,802
Discordance	0.00867
Concordance	0.991
NRD	0.0121
NRC	0.988
RRRR	2,779
RRRA	2
RRAA	0
RARR	40
RARA	4,511
RAAA	42
AARR	0
AARA	1
AAAA	2,427

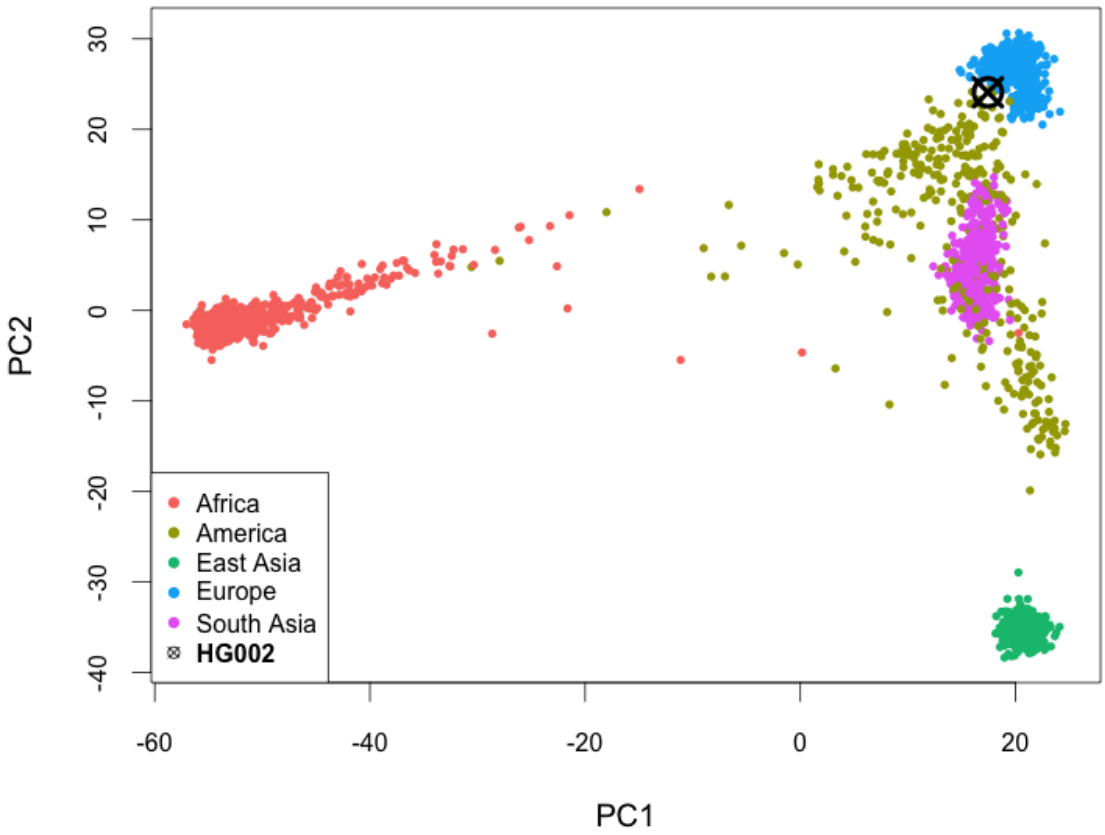

Figure 1: Principal components analysis (PCA) biplot, showing the first two principal components highlighting the HG002 VCF produced by `k2v` against the 2,504 samples in the 1000 Genomes Project phase 3 release data.

Author contributions

Stephen Turner: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration.

Michelle Peck: Resources, Data Curation, Writing – Review & Editing.

Author disclosure statement

The authors have no competing financial interests to disclose. Any mention of commercial products was done for scientific transparency and should not be viewed as an endorsement of the product or manufacturer.

Acknowledgements

The authors thank Erin Gorden, Christina Neal, Dr. Carmen Reedy, and Dr. Alex Koeppel for helpful discussions on the utility of this tool.

Funding

This research received no external funding.

References

Bonfield, James K, John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, and Robert M Davies. 2021. "HTSlib: C Library for Reading/Writing High-Throughput Sequencing Data." *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab007.

Callaghan, Thomas F. 2019. "Responsible Genetic Genealogy." *Science* 366 (6462): 155–55. https://doi.org/10.1126/science.aaz6578.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (1). https://doi.org/10.1186/s13742-015-0047-8.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. "The Variant Call Format and VCFtools." *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab008.

Davawala, A., A. Stock, M. Spiden, R. Daniel, J. McBain, and D. Hartman. 2022. "Forensic Genetic Genealogy Using Microarrays for the Identification of Human Remains: The Need for Good Quality Samples – A Pilot Study." *Forensic Science International* 334 (May): 111242. https://doi.org/10.1016/j.forsciint.2022.111242.

de Vries, Jard H., Daniel Kling, Athina Vidaki, Pascal Arp, Vivian Kalamara, Michael M. P. J. Verbiest, Danuta Piniewska-Róg, Thomas J. Parsons, André G. Uitterlinden, and Manfred Kayser. 2022. "Impact of SNP Microarray Analysis of Compromised DNA on Kinship Classification Success in the Context of Investigative Genetic Genealogy." *Forensic Science International: Genetics* 56 (January): 102625. https://doi.org/10.1016/j.fsigen.2021.102625.

Glynn, Claire L. 2022. "Bridging Disciplines to Form a New One: The Emergence of Forensic Genetic Genealogy." *Genes* 13 (8): 1381. https://doi.org/10.3390/genes13081381.

Greytak, Ellen M., CeCe Moore, and Steven L. Armentrout. 2019. "Genetic Genealogy for Cold Case and Active Investigations." *Forensic Science International* 299 (June): 103–13. https://doi.org/10.1016/j.forsciint.2019.03.039.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." *Nature* 581 (7809): 434–43. https://doi.org/10.1038/s41586-020-2308-7.

Kling, Daniel, and Andreas Tillmar. 2019. "Forensic Genealogy—A Comparison of Methods to Infer Distant Relationships Based on Dense SNP Data." *Forensic Science International: Genetics* 42 (September): 113–24. https://doi.org/10.1016/j.fsigen.2019.06.019.

Li, Jun Z., Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, et al. 2008. "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation." *Science* 319 (5866): 1100–1104. https://doi.org/10.1126/science.1153717.

Manichaikul, Ani, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. 2010. "Robust Relationship Inference in Genome-Wide Association Studies." *Bioinformatics (Oxford, England)* 26 (22): 2867–73. https://doi.org/10.1093/bioinformatics/btq559.

Menozzi, P., A. Piazza, and L. Cavalli-Sforza. 1978. "Synthetic Maps of Human Gene Frequencies in Europeans." *Science* 201 (4358): 786–92. https://doi.org/10.1126/science.356262.

Nagraj, V P, Matthew Scholz, Shakeel Jessa, Jianye Ge, Meng Huang, August E Woerner, Dixie Peters, Bruce Budowle, Michael D Coble, and Stephen D Turner. 2022a. "Relationship Inference with Low-Coverage Whole Genome Sequencing on Forensic Samples." *Forensic Genomics* 2 (3): 81–91. https://doi.org/10.1089/forensic.2022.0009.

Nagraj, V. P., Matthew Scholz, Shakeel Jessa, Jianye Ge, August E. Woerner, Bruce Budowle, Meng Huang, and Stephen D. Turner. 2022b. "Vcferr: Development, Validation, and Application of a SNP Genotyping Error Simulation Framework." Preprint. Bioinformatics. https://doi.org/10.1101/2022.03.28.485853.

Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, et al. 2008. "Genes Mirror Geography Within Europe." *Nature* 456 (7218): 98–101. https://doi.org/10.1038/nature07331.

Patterson, Nick, Alkes L. Price, and David Reich. 2006. "Population Structure and Eigenanalysis." *PLoS Genetics* 2 (12): e190. https://doi.org/10.1371/journal.pgen.0020190.

Peck, Michelle A, Alexander F Koeppel, Erin M Gorden, Jessica Bouchet, Mary C Heaton, David A Russell, Carmen R Reedy, Christina M Neal, and Stephen D Turner. 2022. "Internal Validation of the ForenSeq Kintelligence Kit for Application to Forensic Genetic Genealogy." *bioRxiv*, January, 2022.10.28.514056. https://doi.org/10.1101/2022.10.28.514056.

Phillips, Chris. 2015. "Forensic Genetic Analysis of Bio-Geographical Ancestry." *Forensic Science International: Genetics* 18 (September): 49–65. https://doi.org/10.1016/j.fsigen.2015.05.012.

Price, Alkes L, Johannah Butler, Nick Patterson, Cristian Capelli, Vincenzo L Pascali, Francesca Scarnicci, Andres Ruiz-Linares, et al. 2008. "Discerning the Ancestry of European Americans in Genetic Association Studies." Edited by Jonathan K Pritchard. *PLoS Genetics* 4 (1): e236. https://doi.org/10.1371/journal.pgen.0030236.

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9. https://doi.org/10.1038/ng1847.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3): 559–75. https://doi.org/10.1086/519795.

R Core Team. 2017. "R: A Language and Environment for Statistical Computing."

Ramstetter, Monica D, Thomas D Dyer, Donna M Lehman, Joanne E Curran, Ravindranath Duggirala, John Blangero, Jason G Mezey, and Amy L Williams. 2017. "Benchmarking Relatedness Inference Methods with Genome-Wide Data from Thousands of Relatives." *Genetics* 207 (1): 75–82. https://doi.org/10.1534/genetics.117.1122.

Russell, David A, Erin M Gorden, Michelle A Peck, Christina M Neal, Mary C Heaton, Jessica Bouchet, Alexander F Koeppel, Elayna Ciuzio, Stephen D Turner, and Carmen R Reedy. 2022. "Developmental Validation of the Illumina Infinium Assay Using the Global Screening Array (GSA) on the iScan System for Use in Forensic Laboratories." *bioRxiv*, January, 2022.10.10.511614. https://doi.org/10.1101/2022.10.10.511614.

Shriver, Mark D., and Rick A. Kittles. 2004. "Genetic Ancestry and the Search for Personalized Genetic Histories." *Nature Reviews Genetics* 5 (8): 611–18. https://doi.org/10.1038/nrg1405.

Snedecor, June, Tim Fennell, Seth Stadick, Nils Homer, Joana Antunes, Kathryn Stephens, and Cydne Holt. 2022. "Fast and Accurate Kinship Estimation Using Sparse SNPs in Relatively Large Database Searches." Preprint. Genomics. https://doi.org/10.1101/2022.08.22.504804.

Tian, Chao, Robert M Plenge, Michael Ransom, Annette Lee, Pablo Villoslada, Carlo Selmi, Lars Klareskog, et al. 2008. "Analysis and Application of European Genetic Substructure Using 300 K SNP Information." Edited by Jonathan K Pritchard. *PLoS Genetics* 4 (1): e4. https://doi.org/10.1371/journal.pgen.0040004.

Tillmar, Andreas, Peter Sjölund, Bo Lundqvist, Therese Klippmark, Cajsa Älgenäs, and Henrik Green. 2020. "Whole-Genome Sequencing of Human Remains to Enable Genealogy DNA Database Searches – A Case

Report." *Forensic Science International: Genetics* 46 (May): 102233. https://doi.org/10.1016/j.fsigen.2020.102233.

Tillmar, Andreas, Kimberly Sturk-Andreaggi, Jennifer Daniels-Higginbotham, Jacqueline Tyler Thomas, and Charla Marshall. 2021. "The FORCE Panel: An All-in-One SNP Marker Set for Confirming Investigative Genetic Genealogy Leads and for General Forensic Applications." *Genes* 12 (12): 1968. https://doi.org/10.3390/genes12121968.

Turner, Stephen D. 2022. "KGP: An R Package with Metadata from the 1000 Genomes Project." *arXiv* 2210.00539 (October). https://doi.org/10.48550/arXiv.2210.00539.

Turner, Stephen D., V. P. Nagraj, Matthew Scholz, Shakeel Jessa, Carlos Acevedo, Jianye Ge, August E. Woerner, and Bruce Budowle. 2022. "Skater: An R Package for SNP-based Kinship Analysis, Testing, and Evaluation." *F1000Research* 11: 18. https://doi.org/10.12688/f1000research.76004.1.

Wellcome Trust Case Control Consortium. 2007. "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls." *Nature* 447 (7145): 661–78. https://doi.org/10.1038/nature05911.