applied biosystems

E10: Development Process Validation for Kinship Analysis Algorithm

Carolina Dallett MS; Sharada Vijaychander, MS; Aniketan Swami, MS; Gloria Lam, MS; Narasimhan Rajagopalan, MS

Human Identification, Thermo Fisher Scientific, 180 Oyster Pt Blvd, South San Francisco, CA 94080, USA

ABSTRACT

Computing LR by a kinship analysis algorithm for autosomal markers is straightforward and well defined. Such calculation provides a value for evidence given the prosecution versus the defense proposition. It is recommended and widely used in forensics, missing person and paternity. The forensics community has validated stand-alone software for calculating LR using trios and many biologically related family members. Software such as Familias[1], [2] and MPKin [3] are used regularly for such calculations. Because these implementations are standalone, transcription errors can occur on transferring data from data collection, table input and result storage, in addition it can also be time consuming.

Life Technologies (LT) [5] has incorporated its version of a kinship algorithm, based on ES algorithm[2, 3, 4], to data collection and storage for ease of use and reliability of results, therefore avoiding human transcription errors. This presentation will encompass the steps taken by our team to validate our kinship analysis algorithm given the available methods, data and external collaborators. Building on previous literature, we have used NIST, CEPH and real data from collaborators to compare results of the kinship algorithm to those currently used in the paternity and forensics laboratories. We show that the standard calculations, including complex pedigree trees, mutations and rare alleles concur with currently used methods.

Through this work we have established the LT-kinship algorithm, a more flexible implementation with state-of-the-art models, as accurate. We have further tested the algorithm with SNP data, showing that for a small number of SNPs, the algorithm produces Likelihood ratio (LR) values, which may be an option once expert data and tables become available.

INTRODUCTION

The LT kinship algorithm is similar to established kinship algorithms. It provides statistical basis for family inference of a specified biological relationship by providing a Likelihood ratio (LR). Building on extensive literature and work by the forensics community we have chosen to use Familias, MPKin and Lisa (a commercial software - Future Technology Inc.) for verification and validation of the kinship algorithm with STR loci.

We have now shown that the algorithm also works for simple and complex Kinship relationships. We have obtained STR and SNP genotypes from 4 related people for Likelihood Ratio (LR) calculations. Our aim was to compare the LRs for higher discriminating power.

METHODS

Concordance to algebraic formulas and to established software were performed after extensive literature review and interviews with knowledge and business leaders (Bruce Budowle, Jianye Ge, Arthur Eisenberg, Peter Schneider, Orchid Inc) . Requirements for a flexible kinship algorithm were developed. We obtained guidance on relevant mutation models and rare allele models to implement, in addition to probability tables commonly used on the field.

We further collaborated with Bruce Budowle and Jianye Ge of University of North Texas (UNT) verifying the calculations against their kinship algorithm MPKin (used extensively by the FBI). UNT also provided us with the statistical dataset (Fig 2) which was used as input of validation studies to the Lisa software. For internal verification we used algebraic formulas and Familias. Multiple rounds of testing were performed (Fig1).

For Performance we executed simple to complex scenarios with 1 and 5 loci in a laptop, with the algorithm reading data from a local PostgreSQL database. Time was measured with SAHI (free version).

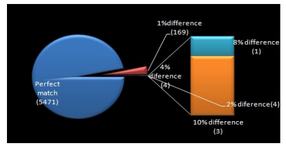
For the SNP/STR data comparison, samples were genotyped using the ION PGM™ System and the following data was used: 37 Identity SNPs from the HID-Ion AmpliSeq™ Identity Panel and 9 CODIS loci. The Frequency table for the SNPs were created using the 1000 genomes using the Yoruba population as reference.

Figure 1. Settings available

1) Mutation models
   - 2-Phase [3]
   - 1-Step [5]
   - No mutation
2) Rare allele models
   - 1/(N+1) [6]
   - 5/(2N) [4]
3) Micro variance
   - Micro variance mutation
4) Import customized probability table
5) Import of profiles In GeneMapper-IDX export format
6) Control over prior probability value,
7) In the event of unknown ethnicity, algorithm can calculate LR on all available ethnicities and will provide Mix and Average LR values
8) Algorithm can be used for SNP genotype

RESULTS

Figure 2. Statistical testing of LT kinship algorithm – Percent Error, in parenthesis number of tests

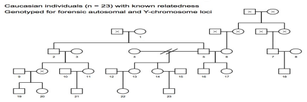


Statistical Test:
Percent error calculated as 100 x (True value – LT-algorithm value) / True value (at the fourth digit)

Paternity Pedigrees: Classical trios, Motherless, siblings, half siblings were input into software. Results were compared to values provided by collaborator, who used Lisa (a commercial software), DNA-view and Cal-DOJ algorithm.

Dataset: The dataset consists of 5,652 permutation of real cases. For example a family with 2 parents and 3 siblings would be permutated into 2 parents and 1 sibling, 2 parents and 2 siblings, 1 parent and 1 sibling, etc… Genotype only data obtained in collaboration with UNT .

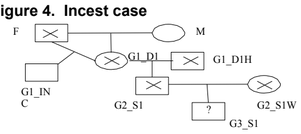
Figure 3. Corner case testing. NIST dataset



Kinship Pedigrees. Complicated cases, cousins, half siblings and avuncular (uncle/ aunt) relationships were entered in LT Kinship algorithm. Results were compared to Familias.

Mutations. The NIST dataset also has members with documented mutations. Trio with mutation were entered in the LT algorithm. Results were compared to MPKin to validate the 2- phase model. 1-step model was not tested based on availability.

Figure 4. Incest case



Incest case - Pedigree used by Ge et. al. for validation of MPKin [3]. Pedigree was entered in the LT algorithm. Results were compared to MPKin .

Table 1. Incest result

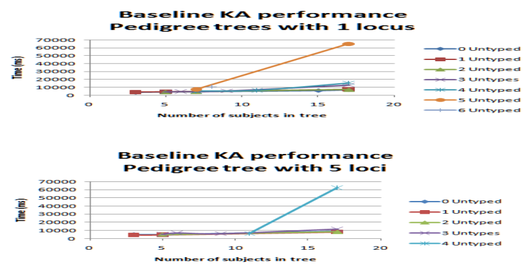
	MPKin	LT Kinship
Marker	LR	LR
CSF1PO	0.895917	0.897718458
D16S539	1.1395515	1.141900818
D7S820	6.9244558	6.933369866
D13S317	0.9936928	0.99000392
D5S818	2.3529993	2.359339727
D3S1358	1.3640239	1.365845639
D8S1179	0.5679763	0.568661701
D18S51	1.4404304	1.442335316
D21S11	0.7852543	0.786719312
FGA	1.3164179	1.31987114
VWA	0.8120352	0.812606708
TPOX	1.7515461	1.75263184
TH01	0.9973072	0.99753224
PENTAD	0.5002333	0.50000000
PENTAE	1.1168459	1.119312696
Combined LR	**15.112767**	**15.35440384**

LR for individual markers and combined LR calculated independently between MPKin and LT-Kinship algorithm – MPKin data provided by Dr. Ge. Collaborators educated us that minor difference observed (~1.5% difference) is expected due to algorithm implementation differences.

Table 2. Improvement over other market products

LT Kinship algorithm	Other Market Products
Implemented with 2-Phase and 1-step mutation models.	Other mutation models have been deemed not adequate for STR mutation [3]
Algorithm allows for 2 different rare allele models (1/N+1 and 5/2N) [4,5]	User must hard code every rare allele value.
Input: GeneMapper IDX output (Profile) Probability table	Input: Specific for application, not intuitive to create. Prone to transcription errors
Drag and drop pedigree interface with genotype and metadata review for cross-check	Progressive choosing of parent for each node from a drop down of names
Algorithm can be attached to a DB for profile storage	Profiles need to be stored in separate text files

Figure 5. Performance of Kinship algorithm for CE- STR data



Performance of the LT Kinship algorithm was obtained by input of progressively more complex pedigree trees. Trees ranged from basic trios, grandparents, 4 generations to 5 generations . Within these pedigree trees, we varied the number of Untyped nodes (ex. A trio without a typed mother) . We observed that the algorithm is generally fast (~5s for a 7 people pedigree tree), the number of Untyped nodes, causes it to slow down (~10s for a 7 people pedigree with 6 Untyped nodes). Experiments were carried out in a local laptop.

PRELIMINARY RESULTS COMPARISON OF LIKELIHOOD RATIO BETWEEN SNP AND STR DATA

Figure 6. Kinship relationship depicted as Hypothesis H0 and Hypothesis H1, the ratio of H0/H1 gives the Likelihood ratio. The same relationship was tested for SNPs and STRs. H1 here depicts all members not related.

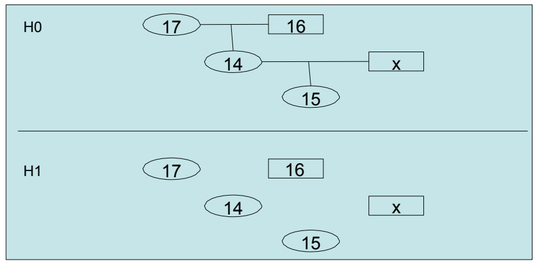


Table 3. Kinship Input: Subset of HID-Ion AmpliSeq™ Identity Panel – 43 Ken Kidd SNPs, Subset of CODIS STRs

	SNPs			STRs		
Sample	rs7520386	rs560681	rs1109037	CSF1PO	D16S539	D3S1358
15	A,G	A,G	A,G	12,12	9,13	16,17
16	A,G	A,A	G,G	11,13	9,9	15,17
17	A,G	A,G	G,G	1,1	11,13	9,9
14	A,G	A,A	A,G	12,13	9,12	16,17

Table 4. Input: Subset SNP Probability table

Population:	Hapmap African				
Sample Size:	1000	1000	1000	1000	1000
Pater_Mute:	0	0	0	0	0
Mater_Mute:	0	0	0	0	0
Allele	rs10092491	rs10488710	rs1058083	rs10773760	rs10776839
A	-	-	0.39	0.703	-
G	-	0.322	0.61	0.297	0.525
C	0.551	0.678	-	-	-
T	0.449	-	-	-	0.475

In comparing STR genotypes to SNP genotypes, we postulated the same algorithm used for STR kinship calculation could be used for SNPs. We further postulated there would be a higher discriminating power when calculating kinship with the SNP dataset.

We had access to the genotype of a complex trio (Figure 6), genotyped in-house, of a grandmother, grandfather, mother and child. The genotypes were generated using the ION PGM™ System and the following data was used: 37 Identity SNPs from the HID-Ion AmpliSeq™ Identity Panel and 9 CODIS loci (generated using an in-house STR panel).

We further generated a frequency table from the 1000 genome site (partially depicted in Table 4) for the SNPs for the Yoruba population. For STR's we used the standard FBI table for African-American population generated by Budowle et al. Using the LT kinship algorithm, we calculated the Likelihood ratio for the same relationship separately for STRs and for SNPs (Table 4).

The LR for STRs was quite strong (3.00E8), This gave us confidence we had the correct relationship, as we tried different combinations and got much lower results (e.g. 15 being the mother of 14 with LR=168.73). When we used the 43 Ken Kidd SNPs [6] of these individuals we obtained a LR=3.35E46. Several orders of magnitude stronger.

Table 5. Partial results from LT-Kinship

STR				SNP			
Combined Likelihood Ratio = 3.00E8				Combined Likelihood Ratio = 3.34E46			
Combined H0 Likelihood = 1.64E-33				Combined H0 Likelihood = 8.80E-121			
Combined H1 Likelihood = 5.45E-42				Combined H1 Likelihood = 2.63E-167			
Locus	HP	HD	LR	Locus	HP	HD	LR
CSF1PO	1.37E-04	1.08E-05	1.27E+01	RS10092491	2.75E-02	2.44E-02	1.13E+00
D16S539	1.79E-04	2.10E-05	8.50E+00	RS10488710	1.53E-02	2.05E-03	7.49E+00
D3S1358	1.31E-03	3.13E-04	4.20E+00	RS1058083	3.45E-02	3.13E-02	1.10E+00
D5S818	1.97E-03	7.09E-04	2.78E+00	RS10773760	3.06E-02	3.60E-02	8.52E-01
D7S820	1.85E-04	3.88E-06	4.76E+01	RS10776839	1.48E-02	7.00E-03	2.11E+00
D8S1179	2.97E-05	1.30E-06	2.28E+01	RS1109037	4.43E-02	4.01E-02	1.10E+00
TH01	1.21E-04	8.61E-06	1.40E+01	RS12997453	1.46E-02	1.80E-03	8.08E+00

CONCLUSIONS

By testing simple trios, deficient trios (e.g. motherless) and more complex kinship relationships (3-5 generations, incest, half-siblings, avuncular) we have shown the Life Technologies kinship algorithm is concordant to algebraic formulas (work not shown) and to standard software currently used for kinship analysis calculations. We have further established the kinship algorithm can process SNPs. Flexible and robust, we present here a user-friendly tool to anybody with minimal training, accommodating SOPs.

During the preparation phase of this project we have pre-empted much of the concern the community has about limitation of software needs such as implementation of different rare allele models and mutation models. The knowledge leaders in the field were paramount to defining these requirements.

It is interesting that the implementation of algorithms sometimes yield very small differences between results, which led us to verify using percent Error. In communication with the experts in the field, they attest these small differences of decisions such as using a mutation model even if loci are compatible (which our algorithm does not do). We have verified that these small differences do not yield changes in probability of paternity. Labs should keep issues such as this in mind when validating an algorithm.

The LT kinship algorithm is meant to be used in the context of a profile database where users can store profiles and probability tables and change parameters from a drop down menu. This would allow for a more robust chain verification of calculations.

Finally, we are excited to share the preliminary results using SNPs as genotype input. The experts we spoke with are aware that other algorithms can be used for a set of 2 members with a large number of SNPs. However they also believe for a smaller set of SNPs such as the Ken Kidd HID SNPs [6] we can perform kinship analysis using this algorithm and get meaningful and statistically significant results with higher discriminating power .

We hope the forensics community will benchmark a set number of SNPs for kinship and provide probability tables. This could lead to the future work of verifying the LR obtained from this algorithm is robust.

REFERENCES

[1] Fung, W. K. User-friendly programs for easy calculations in paternity testing and kinship determinations. **136,** 22–34 (2003).
[2] Azevedo, D. A., Souza, G. R. B., Silva, I. H. E. F. & Silva, L. A. F. Genetic kinship analysis: A concordance study between calculations performed with the software Familias and algebraic formulas of the American Association of Blood Banks. *Forensic Sci. Int. Genet. Suppl. Ser.* **3,** e186–e187 (2011).
[3] Ge, J., Budowle, B. & Chakraborty, R. DNA identification by pedigree likelihood ratio accommodating population substructure and mutations. *Investig. Genet.* **1,** 8 (2010).
[4] Budowle, B., Monson, K. L. & Chakraborty, R. Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci. *Int. J. Legal Med.* **108,** 173–6 (1996).
[5] Brenner, C. H. Symbolic kinship program. *Genetics* **145,** 535–542 (1997).
[6] Pakstis, A. J. *et al.* SNPs for a universal individual identification panel. *Hum. Genet.* **127,** 315–24 (2010).
[7] Egeland, T, Mostad, P.F, Mevåg, B, and Stenersen,M, "Beyond traditional paternity and identification cases. Selecting the most probable pedigree.," *Forensic Sci. Int.*, vol. 110, no. 1, pp. 47–59, May 2000
[8] Gjertson, D.W, Brenner, C.H, Baur, M.P, Carracedo, A, Guidet, F, Luque, J.A, Lessig, R, Mayr, W.R, Pascali, V.L, Prinz, M, Schneider, P.M, and Morling, N, "ISFG: Recommendations on biostatistics in paternity testing.," *Forensic Sci. Int. Genet.*, vol. 1, no. 3–4, pp. 223–31, Dec. 2007.

DATASETS:

- CEPH: http://www.cephb.fr/en/cephdb/
- NIST: K. L. O. Connor *et al* "Candidate Reference Family Data : A Tool for Validating Kinship Analysis Software," p. 5205, 2010.
- HAPMAP for SNP probability (s223624570-pilot_1_YRI_low_coverage_panel) : http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=rs10092491

Author Contact Information
Sharada.Vijaychander@thermofisher.com
Ph: +001 650-504-6519

ThermoFisher SCIENTIFIC