Candidate Reference Family Data: A Tool for Validating Kinship Analysis Software

Poster #35

21th International Symposium on Human Identification
San Antonio, TX, October 11-14, 2010

Email: kristen.oconnor@nist.gov
Phone: 301-975-5205

Kristen L. O'Connor[1], Steven P. Myers[2], Erica L.R. Butts[1], Carolyn R. Hill[1], John M. Butler[1], and Peter M. Vallone[1]

[1] National Institute of Standards and Technology, Biochemical Science Division, Gaithersburg, MD 20899-8312
[2] California Department of Justice, Jan Bashinski DNA Laboratory, Richmond, CA 94804

Copy of poster available:
http://www.cstl.nist.gov/biotech/strbase/pub_pres/Promega2010_OConnor.pdf

Using DNA testing and statistical calculations, kinship analysis evaluates the strength of proposed familial relationships between individuals. Kinship analysis has a variety of applications: criminal and civil paternity cases, mass disaster victim identifications, missing persons identifications, military identifications, and immigration cases. Many software tools are commercially or freely available to aid kinship analysis; however, there is no standard dataset of familial genotypes to help validate calculations made by a software program. Currently, a laboratory must generate pedigrees and genotypes for individuals with known familial relationships. These genotypes are either simulated or taken from previous casework in the laboratory.

The goal of our work is to develop standard reference family data (SRFD) as a tool to aid laboratory validation of kinship analysis software. We are developing an artificial four-generation pedigree as a candidate SRFD based on data collected from six different family groups analyzed with 46 autosomal STRs and 17 Y-STRs. The genotypes of the pedigree reflect observed Mendelian inheritance patterns, including mutations and rare alleles, within real families. The pedigree structure allows for kinship testing of pairwise comparisons (parent-offspring, full siblings, half siblings, first cousins, etc.), paternity trios, and motherless paternity. Due to the size of the pedigree, more complex tests (e.g., incest) can be constructed in the future. The SRFD can be used to verify the functionality of calculations performed by kinship analysis programs including the handling of mutations, rare alleles, and null alleles. Illustrations of how the pedigree data can be used are demonstrated with GeneMarker® HID v. 1.90, a commercially available program from SoftGenetics, and KIn CALc v. 4.0, an Excel®-based freeware program developed at the California Department of Justice. To assist validation for the kinship testing community, the SRFD and pedigree, allele frequency data from major U.S. populations [1-3], and published likelihood ratio formulas will be made available on STRBase (http://www.cstl.nist.gov/strbase/), an online resource for the forensic genetics community [4].

References

[1] Butler, J.M., et al. (2003) Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. J. Forensic Sci. 48(4):908-911.

[2] Hill, C.R., et al. (2008) Characterization of 26 miniSTR loci for improved analysis of degraded DNA samples. J. Forensic Sci. 53(1):73-80.

[3] Hill, C.R., et al. (2010) Concordance and population studies along with stutter and peak height ratio analysis for the PowerPlex® ESX 17 and ESI 17 Systems. Forensic Sci. Int. Genet. (2010), doi:10.1016/j.fsigen.2010.03.014.

[4] Ruitberg, C.M., Reeder, D.J., Butler, J.M. (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res. 29: 320-322.

[5] Hill, C.R., Butler, J.M., Vallone, P.M. (2009) A 26plex autosomal STR assay to aid human identity testing. J. Forensic Sci. 54(5): 1008-1015.

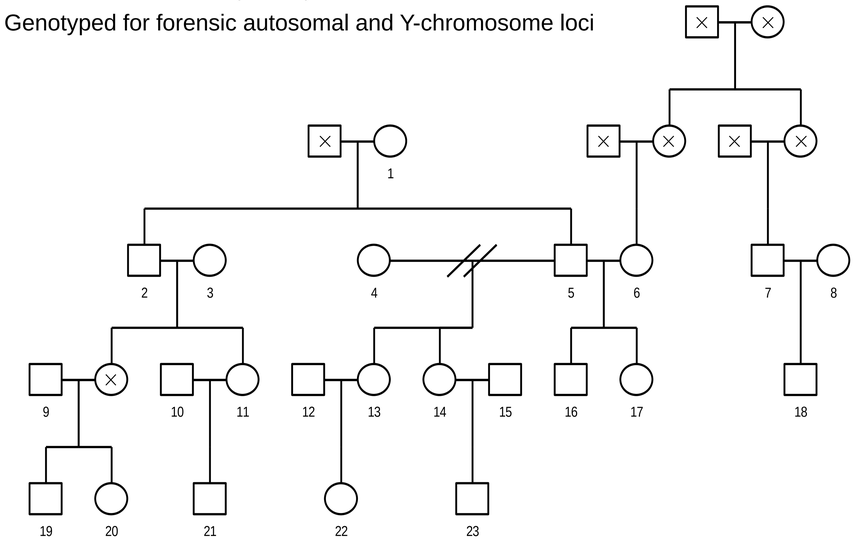
[6] O'Connor, K.L., et al, Linkage disequilibrium analysis of D12S391 and vWA in U.S. population and paternity samples, Forensic Sci. Int. Genet. (accepted)

[7] AABB (2009) Guidance for Standards for Relationship Testing Laboratories, 9th ed. AABB, Bethesda, Maryland, 165 pp.

[8] Slooten, K. Validation of DNA-based identification software by computation of pedigree likelihood ratios. Forensic Sci. Int. Genet. (2010), doi:10.1016/j.fsigen.2010.06.005.

Candidate Reference Family Pedigree

Caucasian individuals (n = 23) with known relatedness
Genotyped for forensic autosomal and Y-chromosome loci



Forensic Markers Genotyped

Autosomal Loci (46 unique loci)
- AmpF/STR® Identifiler® (Applied Biosystems)
- PowerPlex® ESI 17 (Promega)
- NIST-developed assay for 25 STRs [5]

Y-chromosome Loci (17 loci)
- AmpF/STR® Y-filer® (Applied Biosystems)

Common Relationships in Pedigree
(# of possible tests, N)

Pairwise Comparisons	N
Parent-offspring	22
Full siblings	4
Half siblings	4
Uncle/aunt-nephew/niece	9
Grandparent-grandchild	15
First cousins	8
Great grandparent-great grandchild	5
Grand uncle/aunt-grand nephew/niece	5
First cousin once removed	21
Second cousins	8

Paternity Tests	
Paternity trios	9
Reverse paternity trios	9
Motherless paternity	11

Notable Inheritance Patterns

Mutations

D12S391
Individual 9 (paternal mutation of allele 21 → allele 20)

D21S11
Individual 10 (paternal mutation of allele 31 → allele 30)
or
Individual 11 (maternal mutation of allele 29 → allele 30)

Rare Alleles
Observed < 5 times in NIST or FBI Caucasian allele frequency databases

D3S1358
Individuals 4, 7, 13, 14, 23 (allele 13)

vWA
Individual 12 (allele 20)

D16S539
Individuals 7, 18 (allele 6)

D2S1338
Individuals 6,16 (allele 13)

SE33
Individuals 1,2,5,13 (allele 12)
Individual 12 (allele 25)

Application of Standard Reference Family Data (SRFD) to Assist Kinship Analysis

Standard Reference Family Data can be used to verify the functionality of algorithms for kinship analysis.
In addition, the SRFD can be used to evaluate the discriminatory power of adding genetic information to a core set of loci.
Two software programs, GeneMarker® HID v. 1.90 and KIn CALc v. 4.0, were used to demonstrate how the pedigree data can assist validation.

Using SRFD for Validation of Kinship Analysis Methods

1. Determine the parameters to be validated or tested
   - Number and types of familial relationships
   - Loci genotyped
   - Allele frequency database
   - Mutation formula
   - Minimum allele frequency formula
2. Identify the relationships to be tested in pedigree
3. Choose the corresponding genotype data for these individuals
4. Calculate kinship statistics using program or algebraic formulas
   - Likelihood ratio (LR) value or kinship index; posterior probability
5. Compare kinship statistics between program(s) or algebraic formulas
6. Troubleshoot any inconsistent results
7. Validate algorithm between programs(s) or algebraic formulas

Examples of Validation Tests

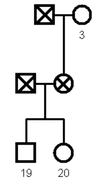
Identifiler® genotypes of individuals from the Candidate Reference Family Data
KIn CALc v. 4.0 for analysis

Complex Kinship Analysis

Family reunification example with individuals 3, 19, and 20.

Evaluate probabilities of grandmother-grandchildren kinship vs. unrelated.

LR = Pr(genotypes|3 is the grandmother of 19 and 20) = 5,245
Pr(genotypes|3 is unrelated to 19 and 20)

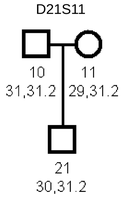


Mutations

Paternity analysis with individuals 10 (alleged father),11 (mother), and 21 (child). Either a paternal or maternal mutation event occurred at D21S11 to result in allele 30 in the child.

Compare likelihood ratio (LR) values when different mutation algorithms are used.

	LR Values for Different Mutation Algorithms		
	No Mutation Allowed	Mutation Always Allowed and Considered	Mutation Considered Only if Required
D21S11	0	0.0033	0.0024
Other 14 loci	760,771	756,069	760,771
Profile	0	2,471	1,836

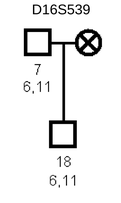


Minimum Allele Frequency for Rare Alleles

Motherless paternity with individuals 7 and 18. Allele 6 at D16S539 is rare in both individuals.

Compare LR values when different algorithms are used.

	LR Values for Different Minimum Allele Frequency Algorithms		
	No Minimum Allowed	5/2N	1-0.05(0.5N)
D16S539	0/0	21	35
Profile	0/0	935,493	1,539,708



Effect of Additional Loci on Kinship Analysis

GeneMarker® HID v. 1.90 for analysis

Known genetic pedigree data are helpful to illustrate the discriminatory power gained by adding loci to a common U.S. forensic panel. The candidate SRFD were used for this analysis.

Pairwise comparisons of candidate SRFD provide a good model to evaluate the effect of additional loci on likelihood ratio (LR) values.

- Five loci, recently adopted in Europe, were added to the 15 Identifiler® loci.
- SE33 was added to further expand the set of loci.
- Locus vWA was removed from analyses with D12S391 due to linkage disequilibrium between the loci [6].

Known Relationship Tested vs Unrelated	Likelihood Ratio Values		
	15 STRs (Identifiler)	19 STRs (Identifiler-vWA + 5 Euro)	20 STRs (Identifiler-vWA + 5 Euro + SE33)
Parent-offspring (3 vs 11)	37,901	1,088,585	4,392,287
Full siblings (2 vs 5)	3.0	291	161,550
Half siblings (13 vs 16)	0.4	0.3	0.2
Uncle-niece (2 vs 14)	0.3	0.7	2.4
Grandmother-grandchild (1 vs 17)	4.0	2.1	1.0
Cousins (22 vs 23)	1.5	2.7	2.0

- Five additional STR loci increase discriminatory power for identification of close relatives (parent-offspring, full siblings) by 2-3 orders of magnitude on average (data not shown).
  - However, by typing more loci, there is a greater chance of mutation events.
- A single highly polymorphic locus (e.g., SE33) can be powerful for identifying first degree relatives if an allele is shared (e.g., in the full sibling example above, locus LR at SE33 = 555).
  - However, a large amount of allelic variation exists due to a high mutation rate
- More distant relationships remain difficult to identify with 20 STR loci.
  - Lineage markers, more putative family members, or non-genetic information (with Bayesian statistics) can increase confidence in a kinship test

SRFD on STRBase http://www.cstl.nist.gov/strbase/kinship.htm

To assist kinship validation activities, the following data will be made available on STRBase, an online resource for the forensic genetics community [4]:
- Standard Reference Family Data genotypes and pedigree
- Notations of mutations, rare alleles, and null alleles in pedigree
- Allele frequency data from major U.S. populations [1-3]
- Published likelihood ratio formulas [e.g., 7,8]

We welcome your feedback and ideas about how NIST may assist kinship analysis through standardized datasets and validation support.
Contact kristen.oconnor@nist.gov

Acknowledgments and Disclaimer

We thank Teresa Snyder-Leiby from SoftGenetics for software demonstration and support. This project was funded by the FBI Criminal Justice Information Services Division and by the National Institute of Justice through interagency agreement 2008-DN-R-121 to the NIST Office of Law Enforcement Standards. Points of view are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice. Certain commercial equipment, instruments and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.