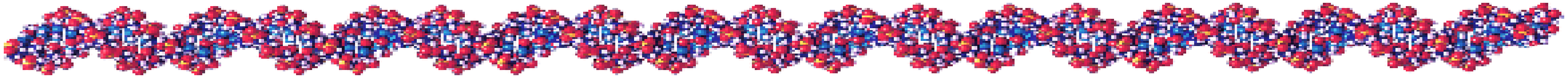
Familial Searching of Forensic DNA Databases



Kristen Lewis O’Connor, Ph.D.

National Institute of Standards and Technology

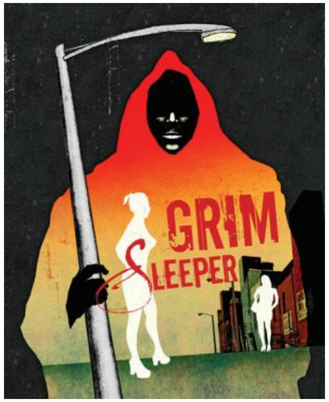
ICFIS, Seattle, WA

July 20, 2011

Outline

- Case aided with familial searching
- Fundamentals of searching for relatives
- Research with New Zealand DNA database
- Ways to increase efficiency of familial searching

Grim Sleeper Case



http://www.laweekly.com/photoGallery/index/316444/0/

Over a 13 year gap in detected crimes, hence the "Sleeper" nickname

- 12 victims murdered in Los Angeles (1985-2007)
- Cases linked through firearms analysis
- DNA evidence recovered and searched against state and national database
- California Dept. of Justice initiated a research program to evaluate the use of familial searching
  - Program was developed and validated using NIST population data from autosomal and Y-STR markers
  - Data are freely available on the STRBase website

www.cstl.nist.gov/strbase/

Myers et al., Searching for first-degree familial relationships in California's offender DNA database. FSI Genetics (in press)
Butler, J.M. (2011) *Advanced Topics in Forensic DNA Typing: Methodology*. Elsevier Academic Press: San Diego. (in press)

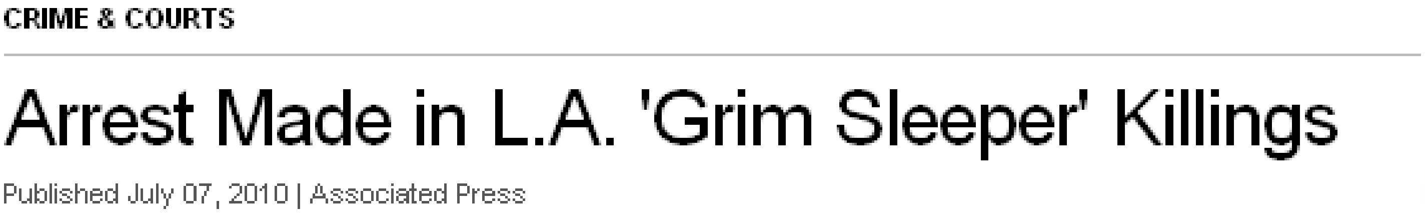
Familial Search for the Grim Sleeper

- October 2008: First familial search of the California database (over 1.1 million profiles) yielded no strong possibilities

- June 30, 2010: Second familial search of the California database (over 1.3 million profiles) yielded one likely relative
  - Database profile from Christopher Franklin (31 years old)
  - Profile added in 2009 after a felony weapons possession charge

- Profiles from Grim Sleeper evidence and C. Franklin shared one allele at all 15 loci

- Both individuals shared the same Y-STR profile



CRIME & COURTS

Arrest Made in L.A. 'Grim Sleeper' Killings

Published July 07, 2010 | Associated Press

CLEARCHANNEL
REWARD
up to $500,000
FOR INFORMATION ON THE MURDER OR ATTEMPTED MURDER OF 12 VICTIMS BETWEEN 1985 AND 2007
3 Ways to provide information to the Los Angeles Police Department
006337

Familial Searching

- Search unknown evidence profile against forensic DNA database to identify possible close relatives of the true offender

- For no suspect cases, cold cases, violent crimes to develop **investigative leads**

- Success in the United Kingdom (2004 – Jan. 2011)
  - 179 cases submitted; 36 successes/81 cases completed (44.4% success rate)
  - **Metadata (age, locality, ethnicity) increase success**

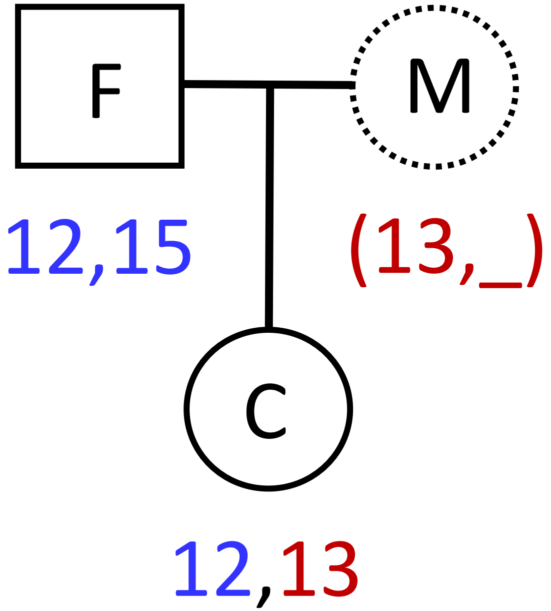
- Familial searching programs in the U.S.
  - Colorado: all forensic unknowns, 10 identifications, 1 conviction (as of June 2011)
  - California: 13 searches, 2 arrests (as of March 2011)
  - Virginia: validation completed (March 2011)
  - Texas

Statistics from: Chris Maguire (UK), Mitch Morrissey (CO), California DOJ, and http://www.fbi.gov/about-us/lab/codis/familial-searching

Fundamentals of Searching for Relatives

Allele Sharing: **Parent-Offspring**

Single locus example



1 allele shared between any parent and child

Probability of sharing alleles from a common ancestor (per locus)

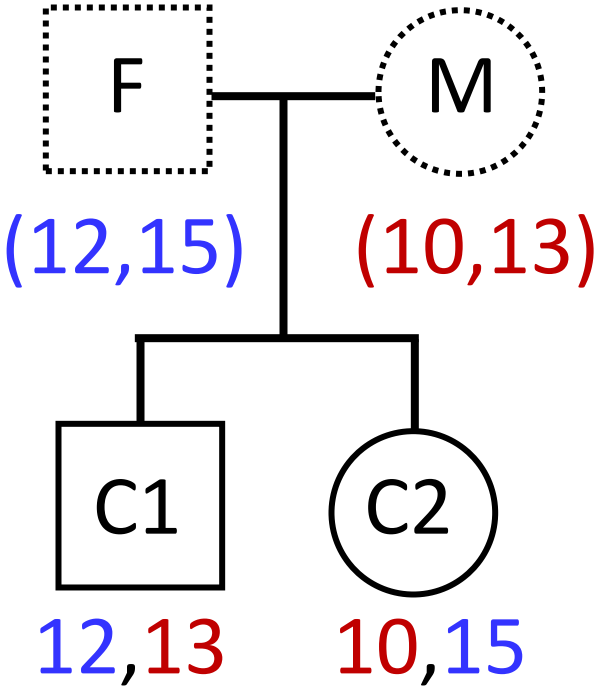
Pr(0 alleles) = 0

Pr(1 allele) = 1

Pr(2 alleles) = 0

Allele Sharing: **Full Siblings**

Single locus example



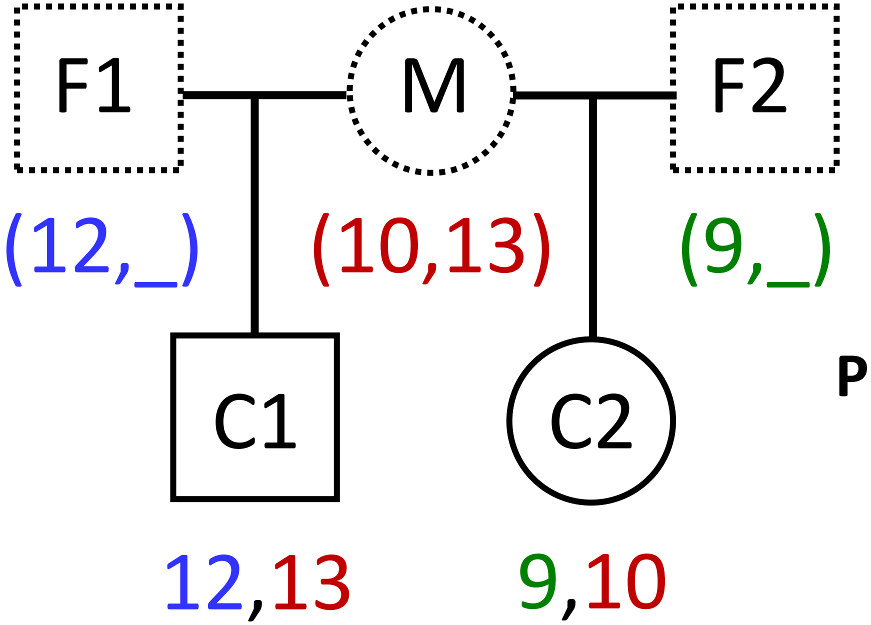
0 alleles shared between
these full siblings

Probability of sharing alleles from a common ancestor (per locus)

Pr(0 alleles) = 1/4
Pr(1 allele) = 1/2
Pr(2 alleles) = 1/4

Allele Sharing: **Half Siblings***

Single locus example



0 alleles shared between
these half siblings

Probability of sharing alleles from a common ancestor (per locus)

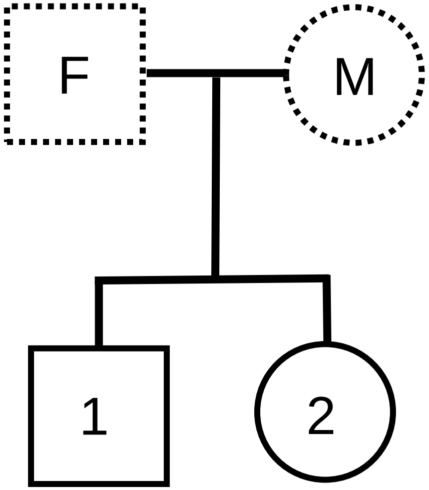
Pr(0 alleles) = 1/2
Pr(1 allele) = 1/2
Pr(2 alleles) = 0

* Allele sharing equivalent for uncle/nephew and grand-parent/grand-child

How is kinship assessed?

Likelihood Ratio (LR)

Evaluate genotypes to give weight (strength) to compared relationships



LR = $\frac{\text{Probability of genotypes if 1,2 are full siblings}}{\text{Probability of genotypes if 1,2 are unrelated}}$

By the definition of a LR:
LR > 1 supports the numerator (alleged relationship)
LR < 1 supports the denominator (unrelated)

Larger LR values provide more support for the alleged relationship

Research with New Zealand DNA Database

Ph.D. dissertation with
Bruce Weir and Mary-Claire King

UW Genome Sciences

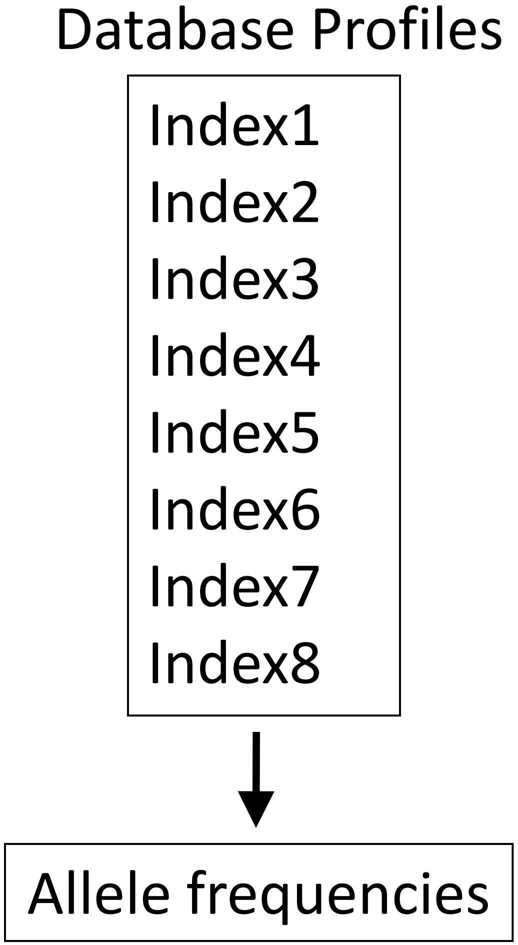
Statistical Modeling

- Assessed the effectiveness of searching for parent-offspring, full sibling, and half sibling (or equivalent) relationships

- Used the New Zealand DNA Database
  - 80,000 subjects
  - 10-locus profiles

- Performed 1,000 simulations by generating one true relative pair per search

Statistical Modeling

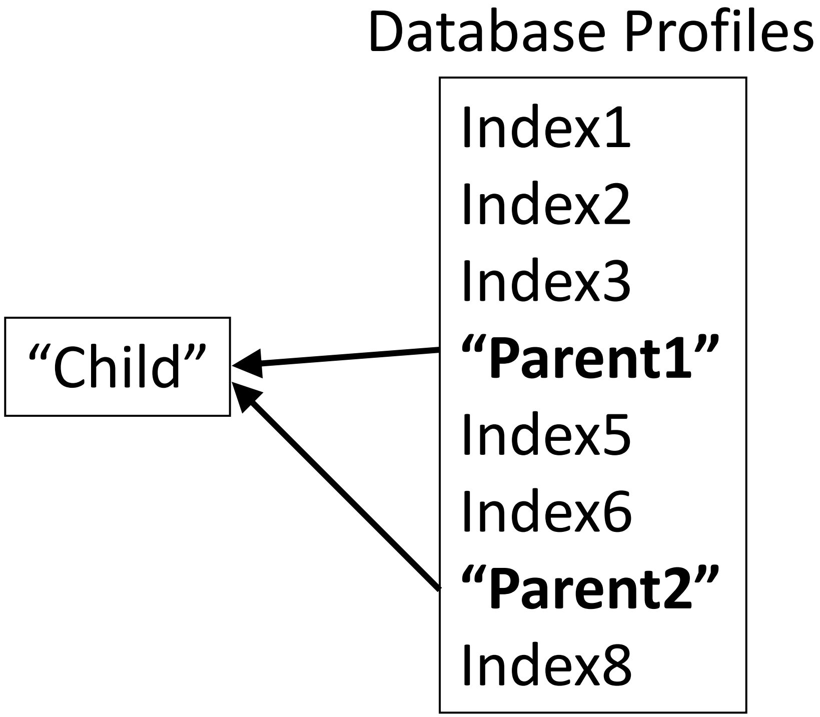


Database Profiles

Index1
Index2
Index3
Index4
Index5
Index6
Index7
Index8

Allele frequencies

Statistical Modeling



Database Profiles

Index1

Index2

Index3

“Parent1”

Index5

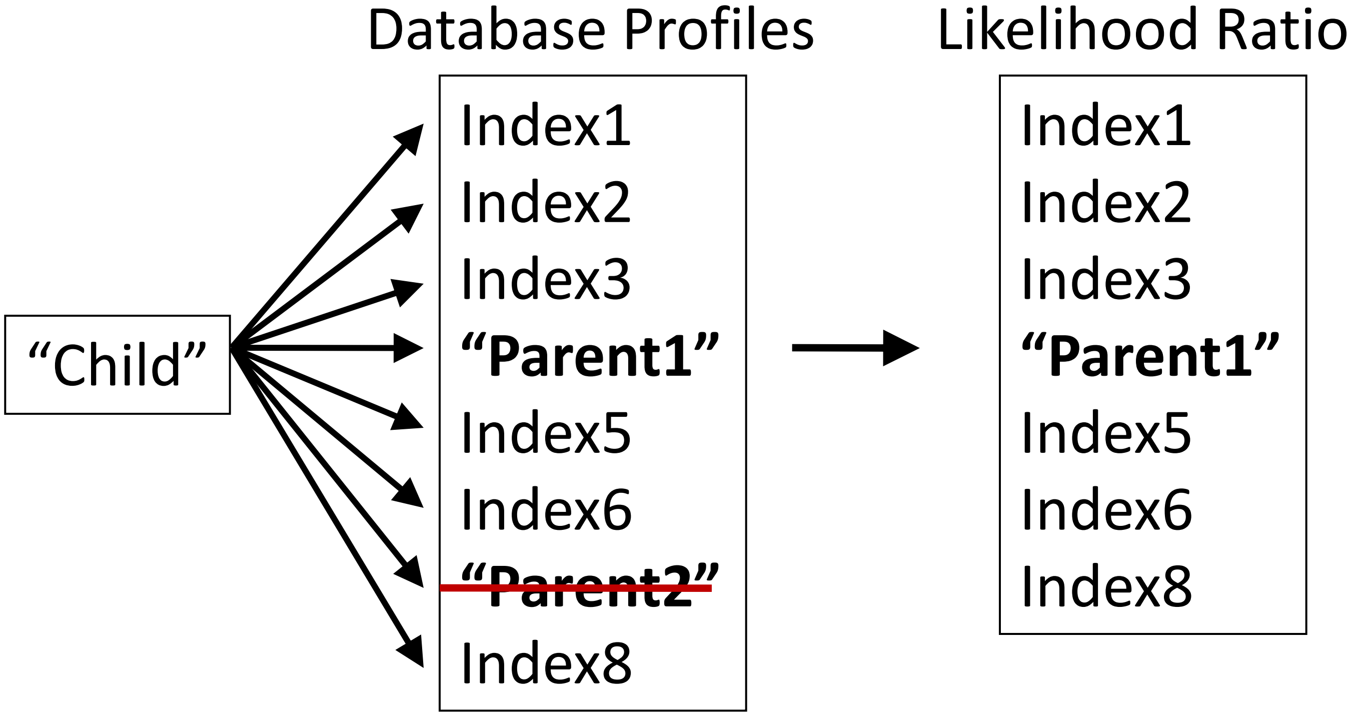
Index6

“Parent2”

Index8

“Child”

Statistical Modeling



Database Profiles

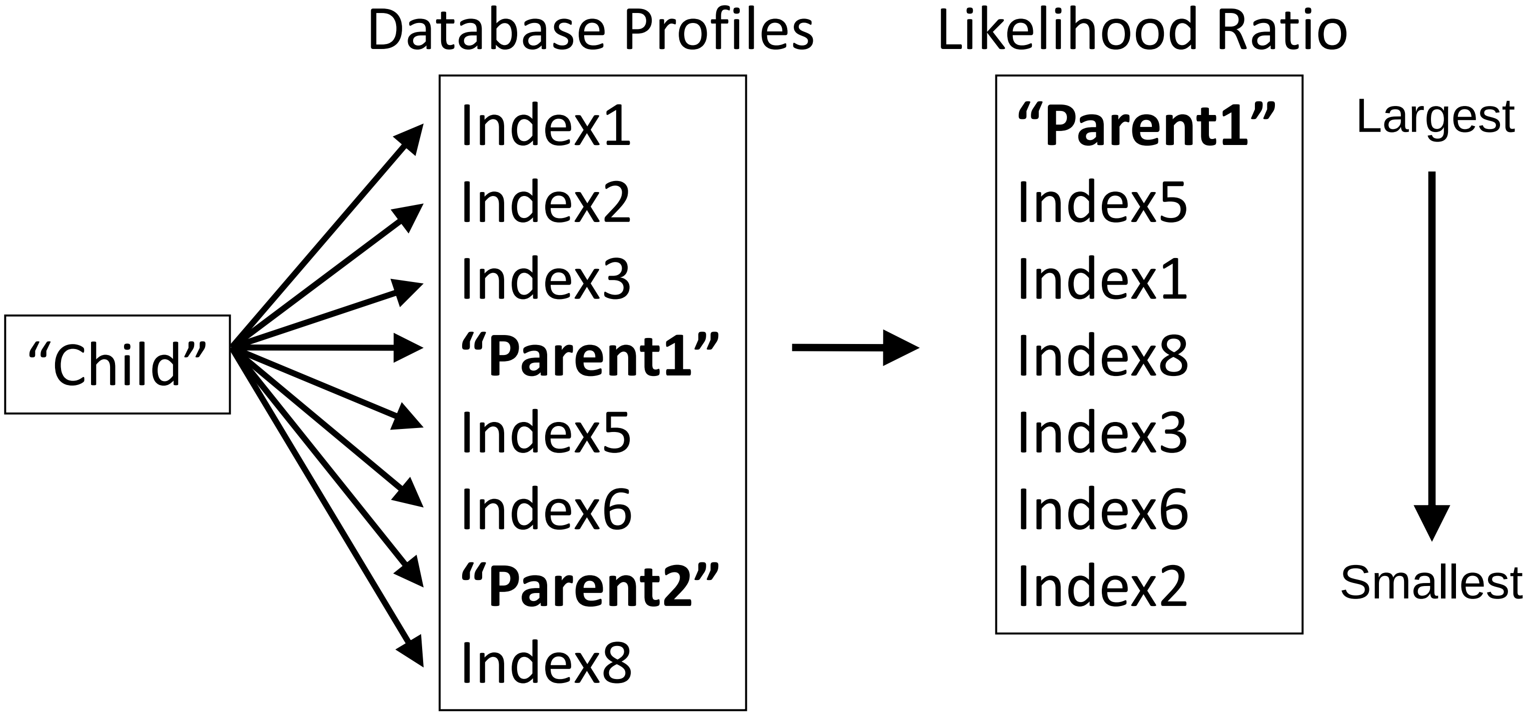
Index1
Index2
Index3
“Parent1”
Index5
Index6
~~**“Parent2”**~~
Index8

Likelihood Ratio

Index1
Index2
Index3
“Parent1”
Index5
Index6
Index8

“Child”

Statistical Modeling



“Child”

Database Profiles

Index1

Index2

Index3

“Parent1”

Index5

Index6

“Parent2”

Index8

Likelihood Ratio

“Parent1”

Index5

Index1

Index8

Index3

Index6

Index2

Largest

Smallest

Challenge of Identifying True Relatives in a Database

Rank	Index #	Likelihood Ratio (LR)	
1	Index5243	7048	— False positive
2	Index1438	5503	— False positive
3	**Parent1**	45	— True positive
4	Index45677	3	— False positive
5	Index39732	0	
6	Index134	0	
7	Index7701	0	
.	.	.	
.	.	.	
412	Index22093	0	
413	Index208	0	

- Unrelated individuals may have higher LRs due to chance allele sharing
  - Included in subsequent investigation → "False positive"
- True relatives will not always have the highest LR
  - Potentially not included in subsequent investigation →"False negative"

Evaluate how often a true relative will be found in a database search

1. Tracked the ordered rank of the true relative for each of the 1000 simulations
2. Calculated the cumulative frequency of true relatives (counts per rank/1000)
3. Think of frequencies as the empirical probability of finding true relative after investigating a certain proportion of individuals in the database

Ordered Rank	Proportion of Database = rank/database size	Count/1000 = Frequency	
1	0.002	108	0.180
2	0.004	88	0.088
3	0.006	83	0.083
4	0.008	46	0.046
5	0.010	47	0.047
6	0.012	34	0.034
7	0.014	27	0.027
.	.	.	.
.	.	.	.
412	0.824	0	0
413	0.826	1	0.001

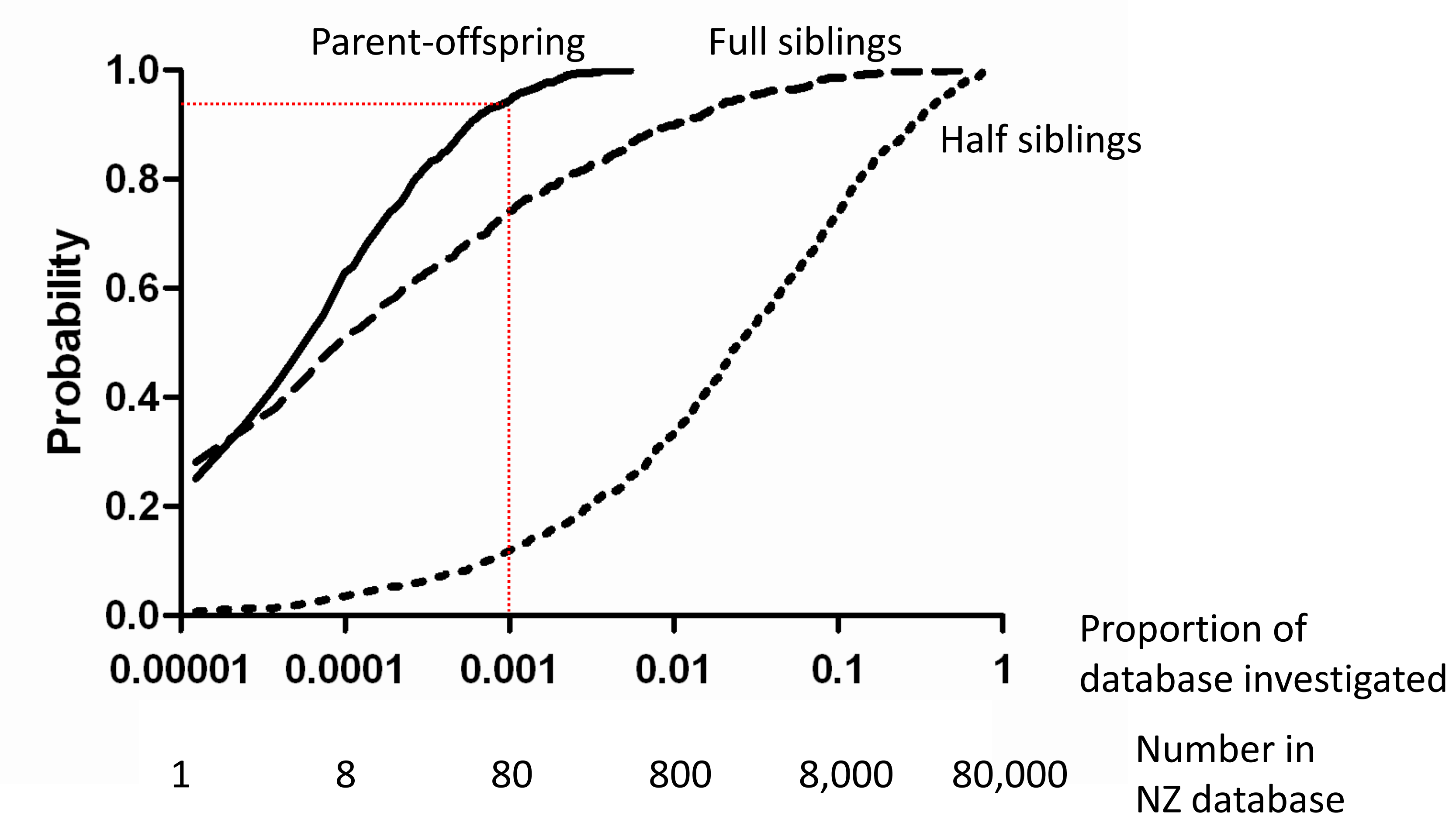
Top 1%

0.444 cumulative frequency

"In an example database of 500 profiles, the probability of finding the true parent-offspring is 0.444 if the top 1% of LR values are investigated after familial searching."

Probability of finding a true relative given the proportion of the NZ database investigated (10 STR loci, n = 80,000)

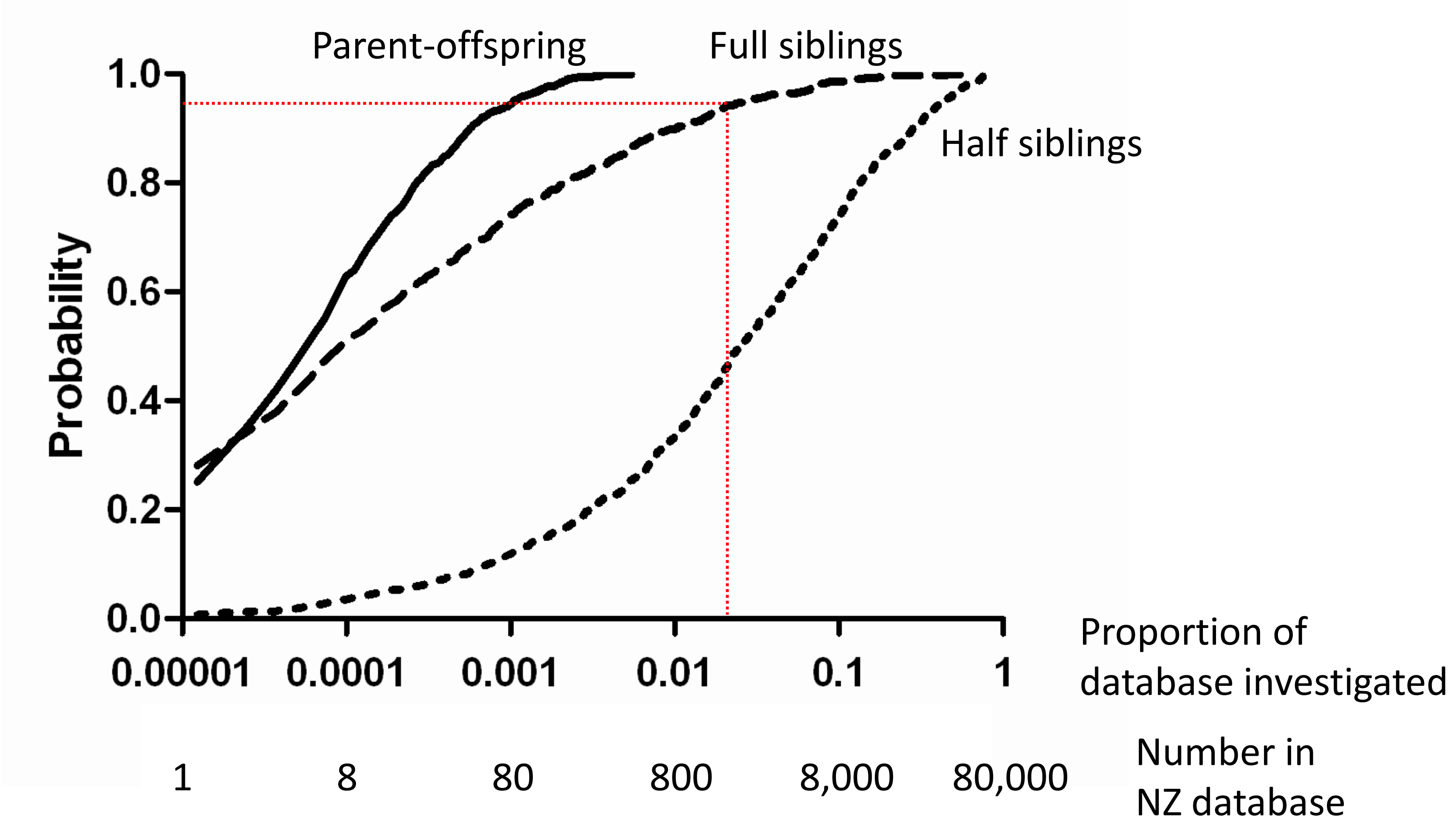
"Simulations indicate that the probability of finding the true parent-offspring is approximately 0.95 if the top 0.1% of LR values are investigated."



Adapted from K. Lewis, Ph.D. Dissertation, University of Washington, 2009

Probability of finding a true relative given the proportion of the NZ database investigated (10 STR loci, n = 80,000)

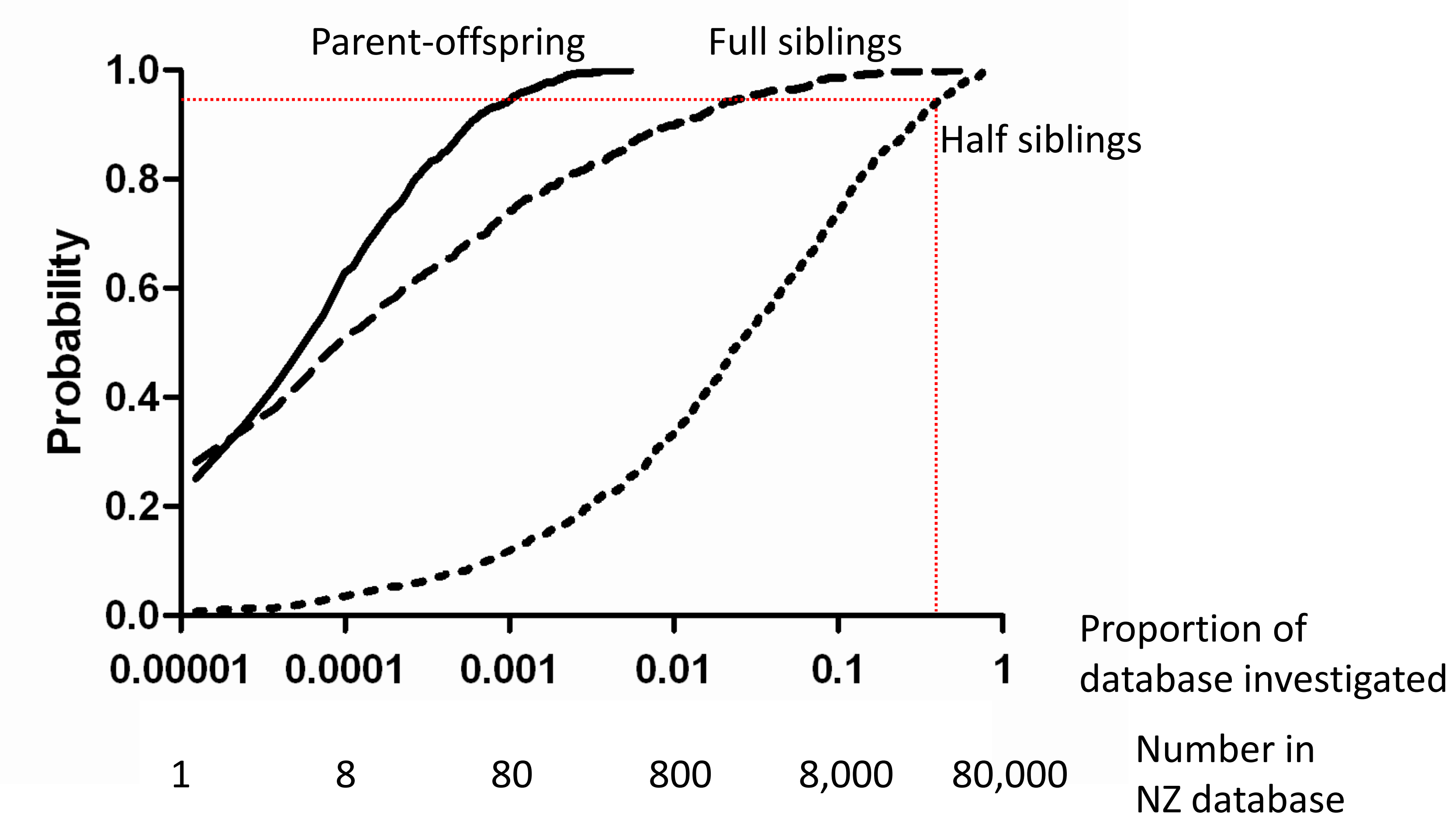
"Simulations indicate that the probability of finding the true full sibling is approximately 0.95 if the top 3% of LR values are investigated."



Adapted from K. Lewis, Ph.D. Dissertation, University of Washington, 2009

Probability of finding a true relative given the proportion of the NZ database investigated (10 STR loci, n = 80,000)

"Simulations indicate that the probability of finding the true half sibling is approximately 0.95 if the top 44% of LR values are investigated."



Adapted from K. Lewis, Ph.D. Dissertation, University of Washington, 2009

Trade-off between False Positives and False Negatives

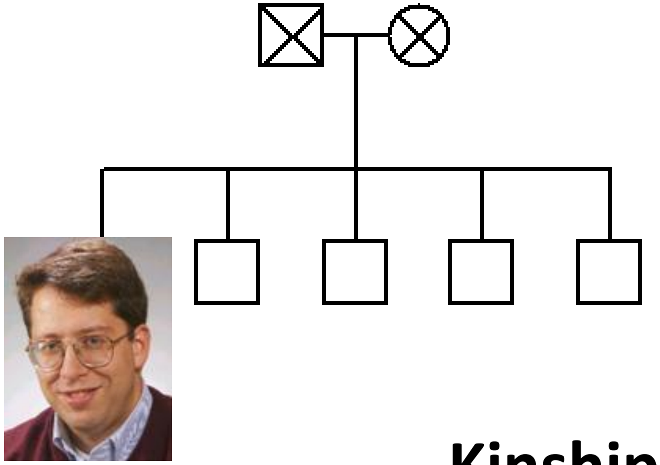
1. Set LR threshold to filter ranked list of potential relatives.
2. What is the probability that a true relative is in this filtered list (PT)?
3. How many false positives will be included in filtered list (FP)?

LR Threshold	**Parent-Offspring**		**Full Siblings**		**Half Siblings**	
	PT	FP	PT	FP	PT	FP
100	0.95	18	0.70	5	0.03	2
10	1.00	45	0.86	60	0.27	184
1	1.00	79	0.95	440	0.73	4,570

Increasing the LR threshold makes familial searching less efficient but reduces the number of false positive leads

Assuming a database (n = 100,000) extrapolated from the NIST U.S. population data (n = 572) with 13 CODIS loci

The Reality of Full Sibling Searches



Range of likelihood ratios for true brothers illustrates chance for false leads, even with additional loci

Kinship Statistics for True Full Siblings

Comparison	Likelihood Ratio	
	13 STRs	19 STRs
Brother 1	571	21,239
Brother 2	2703	1360
Brother 3	1	19,991
Brother 4	2	2

13 STRs: CODIS core

19 STRs: Recommended for expanded CODIS autosomal STRs (D. Hares, FSI Genetics (2011) in press)

Ways to Increase the Efficiency of Familial Searching

More Success for Within-State Searches

Number of profiles that would have to be investigated for 90% chance of finding true relative*

Relationship	**Median state database**	**National database**
	n=100,000	n=10,000,000
Parent-offspring	37	3,700
Full Siblings	134	13,400
Half Siblings	17,441	1,744,100

87% of CODIS hits are within state

* Extrapolated results from searching the New Zealand database with 13 CODIS loci

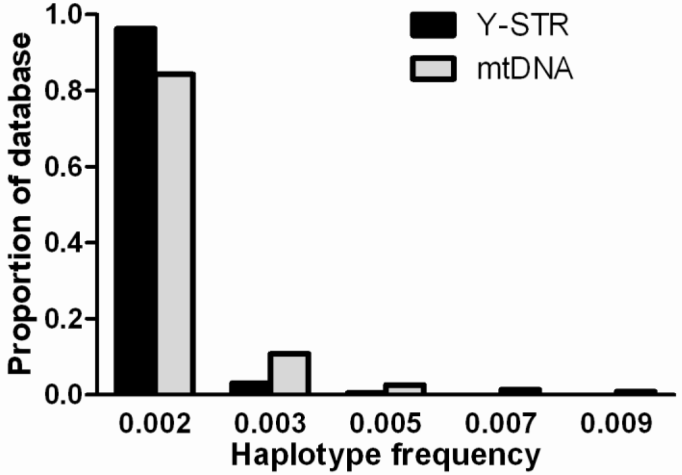
Filter on Y-chromosome

- Missouri has 12 Y-STRs typed on 45,000 (20%) database samples
- Females are “noise” – contribute to false positives
  - Up to 20% of database profiles are female
  - Not investigated if additional testing (Y-STR) is required
  - Remove female profiles prior to familial search or follow-up with non-genetic information
- What about mitochondrial DNA?
  - Sequencing costs are prohibitive
  - Linear arrays have low resolution
- Incorporate LR of Y-STR match into search statistic

$$\boxed{\text{``Odds''} = LR_{\text{autosomal STR}} * LR_{\text{Y-STR}} * 1/N}$$

Myers et al., Searching for first-degree familial relationships in California's offender DNA database. FSI Genetics (in press)

Frequency of 17 Y-STR and HVI/HVII mtDNA haplotypes in the NIST database (n = 572)



Database Longevity Leads to Parent-Offspring Searches

Demographic of male inmates held in custody in U.S. state or federal prison or in local jails, by age, as of June 30, 2009

Age	Proportion of Male Profiles
18-19	0.07
20-24	0.14
25-29	0.15
30-34	0.16
35-39	0.14
40-44	0.12
45-49	0.09
50-54	0.06
55-59	0.04
60-64	0.02
65 or older	0.01

Assume 20-year age gap between father/son

Adapted from http://bjs.ojp.usdoj.gov/content/pub/pdf/pim09st.pdf (Table 19 data).

Conclusions

- Using science, a cost-benefit analysis is necessary to balance effort to find relatives against spending limited resources on false leads

- National database searches are not efficient due to large number of false positives

- Not yet a way to effectively follow up on female profiles

- Database age will increase the utility of parent-offspring searches and the efficiency of familial searching

Acknowledgments

NIST

John Butler
Peter Vallone

North Umbria University

Chris Maguire

FBI

Doug Hares

University of Washington

Bruce Weir
Mary-Claire King

ESR, New Zealand

John Buckleton

Funding

NRC Postdoctoral Fellowship
NIJ

Presentation will be available on STRBase

http://www.cstl.nist.gov/strbase/

kristen.oconnor@nist.gov