Statistics with Y-STR Haplotypes
The primary challenge is the lack of informativeness of the database.

Consider 23 loci with 6 alleles each.
This is $789,730,223,053,603,000$ potential haplotypes.
Starting population 1,000,000 growing to 10,000,000 over 20 generations. The population is subdivided into 10 subpopulations initially of 100,000 each. 100 samples of size 1,000 are drawn from the whole population.

Counting

1 in 1,000

0 in 1,000

Actual frequency in 10,000,000
Factual

Descriptive statistics
Statement about the sample
Count

Inferential

Population
Substructure
Sampling uncertainty
• Do we need a theta correction?
• If so how?
• Does it work?
• How do we combine with autosomal?
Lineage Markers...Y-SNPs


Slide courtesy of John M. Butler (NIST)
Distribution of the most common 16-loci Y haplotypes in Finnish subpopulations ($n=200$).

M. Hedman, V. Pimenoff, M. Lukka, P. Sistonen and A. Sajantila

Lineage Markers...mtDNA

Geographical distribution of major mtDNA clades

Finnish mitochondrial DNA HVS-I and HVS-II population data


M. Hedman, A. Brandstätter, V. Pimenoff, P. Sistonen, J.U. Palo, W. Parson and A. Sajantila
Fig. 2. Mitochondrial DNA haplogroup distribution among 853 regional United States “Hispanics”. All inter-population pairwise Fst values are significant at the 0.05 level.
Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes

L. Roewer\textsuperscript{a,*}, M. Krawczak\textsuperscript{b}, S. Willuweit\textsuperscript{a}, M. Nagy\textsuperscript{a}, C. Alves\textsuperscript{c}, A. Amorim\textsuperscript{c}, K. Anslinger\textsuperscript{d}, C. Augustin\textsuperscript{e}, A. Betz\textsuperscript{f}, E. Bosch\textsuperscript{g}, A. Caglia\textsuperscript{h}, A. Carracedo\textsuperscript{i}, D. Corach\textsuperscript{j}, A.-F. Dekairelle\textsuperscript{k}, T. Dobosz\textsuperscript{l}, B.M. Dupuy\textsuperscript{m}, S. Füredi\textsuperscript{n}, C. Gehrig\textsuperscript{o}, L. Gusmaõ\textsuperscript{c}, J. Henke\textsuperscript{p}, L. Henke\textsuperscript{p}, M. Hidding\textsuperscript{q}, C. Hohoff\textsuperscript{r}, B. Hoste\textsuperscript{k}, M.A. Jobling\textsuperscript{g}, H.J. Kärgel\textsuperscript{s}, P. de Knijff\textsuperscript{t}, R. Lessig\textsuperscript{u}, E. Liebeherr\textsuperscript{v}, M. Lorente\textsuperscript{w}, B. Martínez-Jarreta\textsuperscript{x}, P. Nievas\textsuperscript{x}, M. Nowak\textsuperscript{y}, W. Parson\textsuperscript{z}, V.L. Pascali\textsuperscript{h}, G. Penacino\textsuperscript{j}, R. Ploski\textsuperscript{y}, B. Rolf\textsuperscript{d}, A. Sala\textsuperscript{j}, U. Schmidt, C. Schmitt\textsuperscript{q}, P.M. Schneider, R. Szibor, J. Teifel-Greding, M. Kayser
Minimal Haplotype Result

DYS19 – 14
DYS389I – 13
DYS389II – 29
DYS390 – 24
DYS391 – 11
DYS392 – 14
DYS393 – 13
DYS385 a/b – 11,15

7 matches in 27,773 individuals from 236 worldwide populations
“The estimated mtDNA haplotype frequencies should be interpreted in the light of the data available concerning the distribution of the mtDNA haplotypes and the possible subpopulation structures within in the relevant population(s)”

clusters of regional groups could be identified in Europe ... indicating Y-STR haplotype-based population substructure [51]. These effects thus need to be considered as well when haplotype frequencies are estimated.

Recommendations on the estimation of Y-STR haplotype frequencies and estimation of the weight of the evidence of Y-STR typing will be presented separately as guidelines for the interpretation of forensic genetic evidence.
Marianne Vaatstra case

Arnoud Kal and Charissa van Kooten, Netherlands Forensic Institute;

Ron Rintjema, Jelle Tjalsma and Cor Reijenga, 3-D team Friesland Police

Peter de Knijff, University of Leiden, The Netherlands
Ronny Decorte, University of Leuven, Belgium

Case: rape and murder of a 16-year old girl in 1999. Sperm, blood and hair of an unknown male were recovered from the crime scene. Haplogroup R1b, probably a local man.

The Y-STR profile of the perpetrator did not match any Y-STR profiel in the YHRD and USYSTR databases nor in several genetic genealogy databases (over 200,000 in total).

After several new lines of investigation turned out negative, it was decided to perform a voluntary large scale Y-STR base familial search among 7300 male individuals in the area within a 3-mile radius from the crime scene.
Marianne Vaatstra case

In 7 weeks …. generated 3880 Y-STR profiles.  
23 men matched 17 of 17 Y-filer loci, surnames A, B, C.  
5 men matched 16 of 17 Y-filer loci, surnames A, B, D, E.  
7 men matched 15 of 17 Y-filer loci, surnames F, G, H.  
These Y haplotypes corresponded to 8 different surnames.  
Autosomal DNA indicated no parent-child relationships and no indication of sibling relationship.

38 Y-STRs and 15 RM-Y-STRs indicated the perpetrator could be found within family A. A pedigree was constructed, back to the year 1748. Families A and B turned out to have a common ancestor.
Contrast

\[
\frac{0}{200,000} \quad \frac{23}{3,880}
\]
Factual

Descriptive statistics

Statement about the sample

Count

Inferential

Population

Substructure

Sampling uncertainty
This will be dominant
<table>
<thead>
<tr>
<th>Loci</th>
<th>For African Americans, Asians, Caucasians &amp; Hispanics</th>
<th>For African Americans, Asians, Caucasians, Hispanics &amp; Native Americans</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>5</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>6</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>7</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>8</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>9</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>10</td>
<td>0.0006</td>
<td>0.002</td>
</tr>
<tr>
<td>11</td>
<td>0.0004</td>
<td>0.0009</td>
</tr>
<tr>
<td>12</td>
<td>0.0002</td>
<td>0.0007</td>
</tr>
<tr>
<td>13</td>
<td>0.0002</td>
<td>0.0006</td>
</tr>
<tr>
<td>14</td>
<td>0.0001</td>
<td>0.0005</td>
</tr>
<tr>
<td>15</td>
<td>0.00008</td>
<td>0.0005</td>
</tr>
<tr>
<td>16</td>
<td>0.00006</td>
<td>0.0004</td>
</tr>
<tr>
<td>17</td>
<td>0.00003</td>
<td>0.0004</td>
</tr>
<tr>
<td>18</td>
<td>0.00003</td>
<td>0.0004</td>
</tr>
<tr>
<td>19</td>
<td>0.00003</td>
<td>0.0003</td>
</tr>
<tr>
<td>20</td>
<td>0.00002</td>
<td>0.0003</td>
</tr>
<tr>
<td>21</td>
<td>0.00002</td>
<td>0.0003</td>
</tr>
<tr>
<td>22</td>
<td>0.00002</td>
<td>0.0003</td>
</tr>
</tbody>
</table>
Does it work? How can we test?

Need to test the prediction against the observed value.
Does it work? How can we test?

Need to test the prediction against the observed value.

Population

Predict match probability in a subpopulation

Subpop 1
Observed match probability

Subpop 2
Observed match probability

Subpop 3
Observed match probability

Subpop 4
Observed match probability
Does it work? How can we test?

Need to test the prediction against the observed value.

\[
\sum_{i} \theta \tilde{p}_i + (1 - \theta) \tilde{p}_i^2
\]
<table>
<thead>
<tr>
<th>Region</th>
<th>Subpopulation</th>
<th>1st Half</th>
<th>2nd Half</th>
<th>Observed $M_W$</th>
<th>$\hat{M}_W = \hat{\beta} + (1 - \hat{\beta})\sum u \hat{p}_u^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>45 European meta populations</td>
<td>1st half</td>
<td>0.0085</td>
<td>0.0073</td>
<td>0.0161</td>
<td>0.0157</td>
</tr>
<tr>
<td></td>
<td>2nd half</td>
<td>0.0085</td>
<td>0.0076</td>
<td>0.0156</td>
<td>0.0160</td>
</tr>
<tr>
<td>2 subpopulations, Eastern and Western</td>
<td>1st half</td>
<td>0.0064</td>
<td>0.0101</td>
<td>0.0156</td>
<td>0.0164</td>
</tr>
<tr>
<td></td>
<td>2nd half</td>
<td>0.0056</td>
<td>0.0086</td>
<td>0.0135</td>
<td>0.0142</td>
</tr>
<tr>
<td>Belorussia, Kiev, Ljubljana, Moscow, Novgorod, Poland, Riga, Vilnius, Zagreb</td>
<td>1st half</td>
<td>0.0009</td>
<td>0.0115</td>
<td>0.0110</td>
<td>0.0124</td>
</tr>
<tr>
<td></td>
<td>2nd half</td>
<td>0.0051</td>
<td>0.0113</td>
<td>0.0160</td>
<td>0.0164</td>
</tr>
<tr>
<td>Emilia Romagna, London, Portugal, Pyrenees, South Holland, Southern Ireland, Spain, Strasbourg</td>
<td>1st half</td>
<td>0.0040</td>
<td>0.0174</td>
<td>0.0275</td>
<td>0.0213</td>
</tr>
<tr>
<td></td>
<td>2nd half</td>
<td>0.0029</td>
<td>0.0261</td>
<td>0.0263</td>
<td>0.0289</td>
</tr>
</tbody>
</table>

Credit Myers, Roewer, Weir, Willeit, and Buckleton
The Government alleges that Defendant Theodore Kootswatewa, a Hopi adult, sexually assaulted a Hopi girl inside an abandoned trailer owned by a Hopi woman on the Hopi reservation.

- Yfiler 17 STR loci
- Applied Biosystems database and determined that the profile has not been observed $N = 105$ Native Americans
- 1 in 35

https://casetext.com/case/united-states-v-kootswatewa-1
Charles H. Brenner... it is questionable whether there would be any genetic common ancestry among Native Americans today because of the isolation of specific tribes and the natural mutation process.

...pooling Native Americans into a single genetic classification could manufacture diversity, thereby inflating random match probabilities ...
Native American pooled data when the suspect is a Hopi charged with an offense on the Hopi reservation likely results in "a hugely exaggerated statistic,

" and that by using the pooled data "[y]ou'll be framing the suspect."

Dr. Brenner opined that the "1 in 35 Native Americans" statistic generated by Ms. Daniel's analysis is not reliable because it cannot be known whether the Applied Biosystems database includes an appropriately representative population of any particular Native American tribe.
UNITED STATES V. KOOTSWATEWA

• the Court finds that Ms. Daniel's testimony about the probability of a random match of the Y-STR partial DNA profile identified on the victim is not reliable under Rule 702
• The Counting method,
• Augmented counting method
• The Clopper and Pearson 95% confidence interval
• The application of a subpopulation correction
• The Kappa method
• The Discrete Laplace method
• The Generalised Good method
• The coalescent method

Credit Duncan Taylor, James Curran and John Buckleton
• The Counting method
• estimate of the population proportion

\[ \hat{p}_x = \frac{C}{D} \]

• C is the count in a database of size D
• This is the traditional and incumbent method
• C is often 0
• Augmented counting
• Add the observation to the database

\[ \hat{p}_x = \left( C + 1 \right) / \left( D + 1 \right) \]
• The Clopper and Pearson 95% confidence interval

\[ \hat{p}_x = \frac{C}{D} \]

• Adds an exact confidence interval to either the counting or augmented counting method

• Subpopulation correction – BS Weir

\[
\hat{p}_x = \hat{\beta}_w + (1 - \hat{\beta}_w) \left( \frac{C}{D} \right)
\]
• The Kappa method – Charles Brenner

\[ \hat{p}_x = \frac{(C + 1)(1 - \kappa)}{D + 1} \]

\( \kappa \) denotes the fraction of hapltypes that have been observed only once, i.e. singletons, in the database augmented by \( x \)
• The Discrete Laplace method
• The Discrete Laplace (hereafter Laplace) method gives a profile probability. It uses the following genetic assumptions to model a probability distribution:
  • A population of haplotypes is composed of clades of haplotypes,
  • Each clade has arisen from one ancestral haplotype by stepwise mutation, and
  • Mutations occur independently of each other.
• The Generalised Good
• This method calculates a likelihood ratio rather than a haplotype probability or a match probability, however we will display the inverse of the $LR$ in order to allow it to be compared

$$LR = \frac{(D - C - 1) D_{C+1}}{(C + 2) D_{C+2}} \approx \frac{DD_{C+1}}{(C + 2) D_{C+2}}$$

$D_{C+1}$ for example, using $C = 1$, the $D_2$ is the number of matching pairs in the database
THE COALESEENCE METHOD

Assumes some ancient state of a population where a single haplotype existed and that all current haplotype diversity is from mutations of that ancient state haplotype.

These haplotypes, and the haplotype of the suspect, $h_s$, are ordered into a large number of coalescent trees.

The donor of the trace, $x$, with haplotype $h_x$, is trialled in different positions in the trees.

Starting population 1,000,000 growing to 10,000,000 over 20 generations.

The population is subdivided into 10 subpopulations initially of 100,000 each.

100 samples of size 1,000 are drawn from the whole population.

The various estimation procedures applied to 1,000 haplotypes drawn from the whole population using the sample.

The observed match probability is calculated by the frequency of the haplotype in the whole population.
Conservative  Clopper and Pearson  Laplace

Generalised Good  Coalascent
Y STR mixtures: The challenge

Consider a profile that has ground truth a 1:1 mix of:

<table>
<thead>
<tr>
<th>Donor 1</th>
<th>Donor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

This could be explained as:
ac:bd $\rightarrow$ do exist
ad:bc $\rightarrow$ may or may not exist in the database or references

Please mentally extend to 23 or 27 loci

There are about 4 – 67 million haplotype combinations for this simple mix
Most of these exist neither in the database nor references
• This is a novel problem
• We have never previously needed the probability of a profile neither in the database nor references
• The type of summations in LRs for Y mixtures will involve millions of these.
• Laplace does do this but has a worrying non-conservativeness
• I have not yet worked out whether being conservative in a haplotype probability always leads to conservative LRs.
• I think the answer probably is nearly always.
Statistics

- Combining Y and mito with autosomal
- Bruce Walsh

Available online at www.sciencedirect.com

ScienceDirect


Rapid communication

Joint match probabilities for Y chromosomal and autosomal markers

Bruce Walsh\(^a\), Alan J. Redd\(^b\), Michael F. Hammer\(^{a,b,*}\)

\(^a\)Department of Ecology and Evolutionary Biology, University of Arizona Tucson, AZ 85721, USA
\(^b\)Arizona Research Laboratories, Division of Biotechnology, University of Arizona, Tucson, AZ 85721, USA

Received 12 July 2006; received in revised form 13 February 2007; accepted 18 March 2007
Suggests a geometric distribution of TMRCA
Prior

Probability vs. generations MRCA
Posterior considering mutation
Use TMRCA posterior to compute theta for autosomal for “unrelated = not sibs or cousins”

<table>
<thead>
<tr>
<th>Multiplex</th>
<th>PPY</th>
<th>Yfiler</th>
<th>PowerPlex</th>
<th>Yfiler Plus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loci (l)</td>
<td>11</td>
<td>16</td>
<td>22</td>
<td>25</td>
</tr>
<tr>
<td>$\mu_{\text{ave.}}$</td>
<td>0.0021</td>
<td>0.0026</td>
<td>0.0035</td>
<td>0.0057</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N_Y$</th>
<th>PPY</th>
<th>Yfiler</th>
<th>PowerPlex Y23</th>
<th>Yfiler Plus</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.002</td>
<td>0.003</td>
<td>0.004</td>
<td>0.006</td>
</tr>
<tr>
<td>1,000</td>
<td>0.002</td>
<td>0.002</td>
<td>0.004</td>
<td>0.005</td>
</tr>
<tr>
<td>10,000</td>
<td>0.002</td>
<td>0.002</td>
<td>0.004</td>
<td>0.005</td>
</tr>
<tr>
<td>100,000</td>
<td>0.002</td>
<td>0.002</td>
<td>0.004</td>
<td>0.005</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N_Y$</th>
<th>PPY</th>
<th>Yfiler</th>
<th>PowerPlex Y23</th>
<th>Yfiler Plus</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.011</td>
<td>0.012</td>
<td>0.013</td>
<td>0.014</td>
</tr>
<tr>
<td>1,000</td>
<td>0.011</td>
<td>0.011</td>
<td>0.013</td>
<td>0.014</td>
</tr>
<tr>
<td>10,000</td>
<td>0.011</td>
<td>0.011</td>
<td>0.013</td>
<td>0.014</td>
</tr>
<tr>
<td>100,000</td>
<td>0.011</td>
<td>0.011</td>
<td>0.013</td>
<td>0.014</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N_Y$</th>
<th>PPY</th>
<th>Yfiler</th>
<th>PowerPlex Y23</th>
<th>Yfiler Plus</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.031</td>
<td>0.032</td>
<td>0.033</td>
<td>0.034</td>
</tr>
<tr>
<td>1,000</td>
<td>0.031</td>
<td>0.031</td>
<td>0.033</td>
<td>0.034</td>
</tr>
<tr>
<td>10,000</td>
<td>0.031</td>
<td>0.031</td>
<td>0.033</td>
<td>0.034</td>
</tr>
<tr>
<td>100,000</td>
<td>0.031</td>
<td>0.031</td>
<td>0.033</td>
<td>0.034</td>
</tr>
</tbody>
</table>

Credit John Buckleton and Steven Myers
But derived from the Walsh et al insight
• Do we need a theta correction? I think so.
• If so how? Bruce Weir’s method.
• Does it work? Yes but we’d love more data.
• How do we combine with autosomal? Walsh method? Decide what we mean by unrelated.