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.

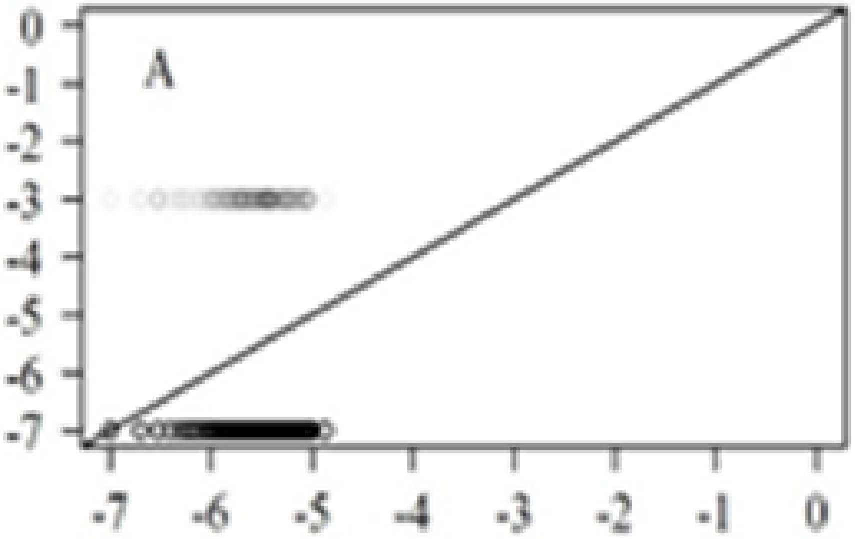
The population is subdivided into 10 subpopulations initially of 100,000 each.

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

Counting

1 in 1,000

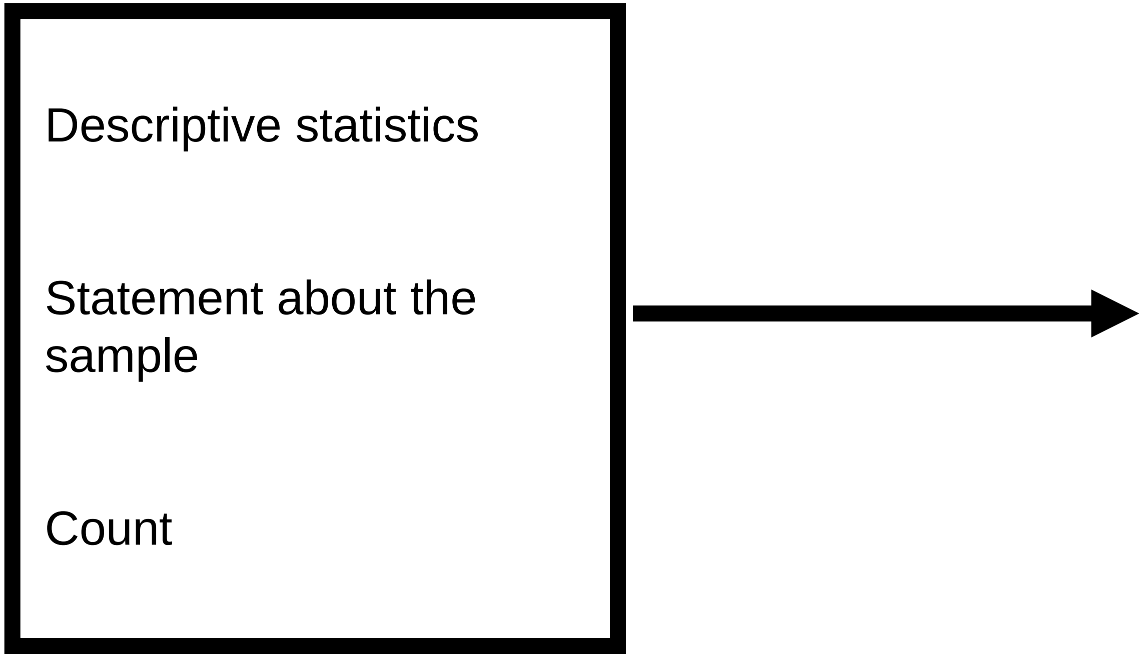
0 in 1,000



Actual frequency in 10,000,000

Factual

Inferential



Descriptive statistics

Statement about the sample

Count

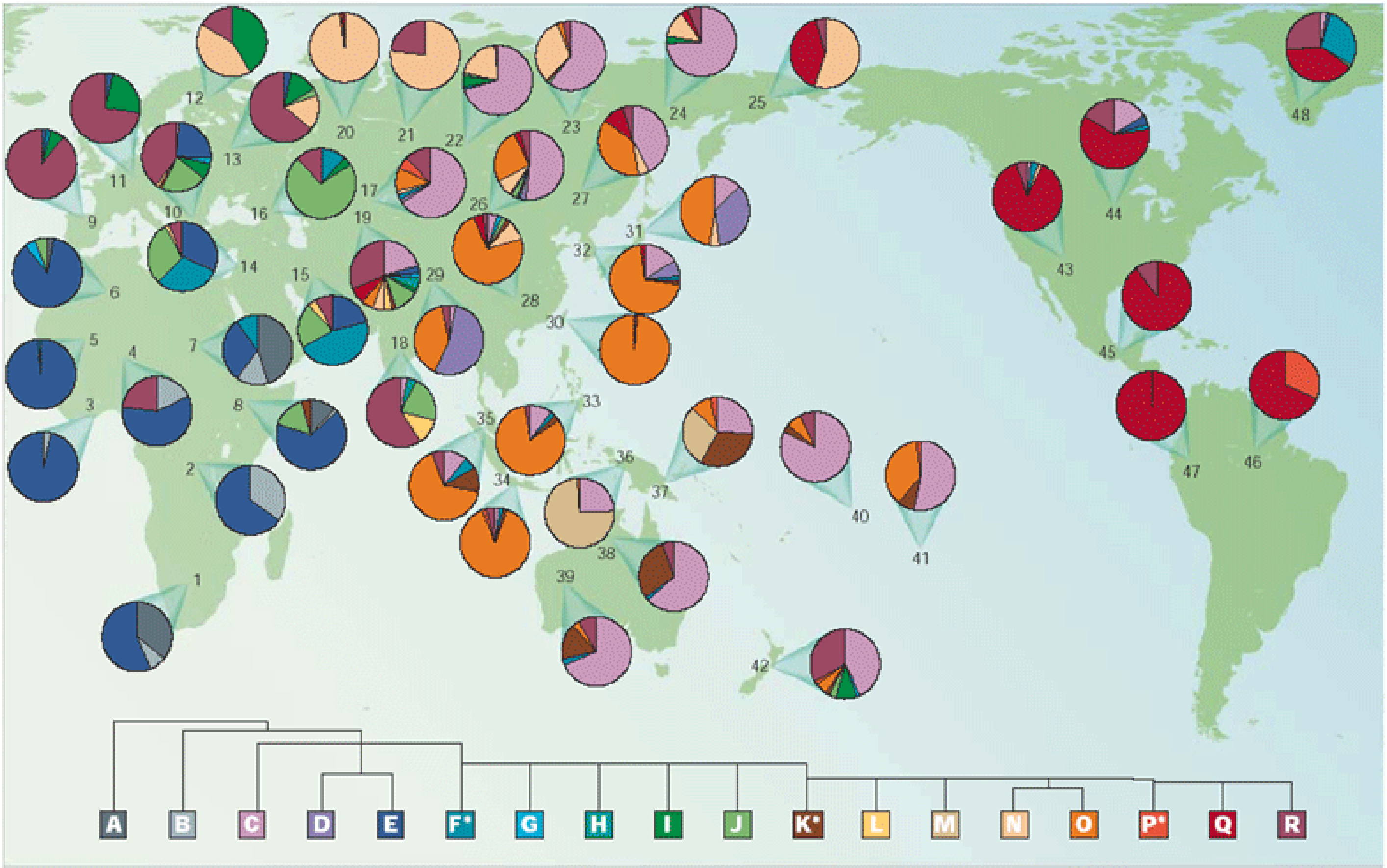
Population

Substructure

Sampling uncertainty

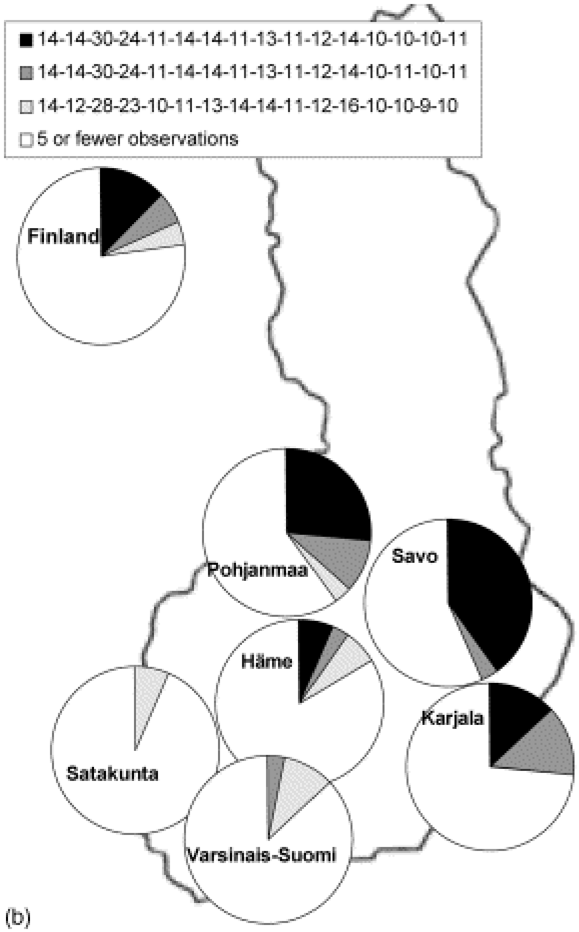
- Do we need a theta correction?
- If so how?
- Does it work?
- How do we combine with autosomal?

Lineage Markers…Y-SNPs



Jobling MA and Tyler-Smith C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics* 4: 598-612

Slide courtesy of
John M. Butler (NIST)



Distribution of the most common 16-loci Y haplotypes in Finnish subpopulations (*n*=200).

Analysis of 16 Y STR loci in the Finnish population reveals a local reduction in the diversity of male lineages

Forensic Science International, Volume 142, Issue 1, 28 May 2004, Pages 37-43

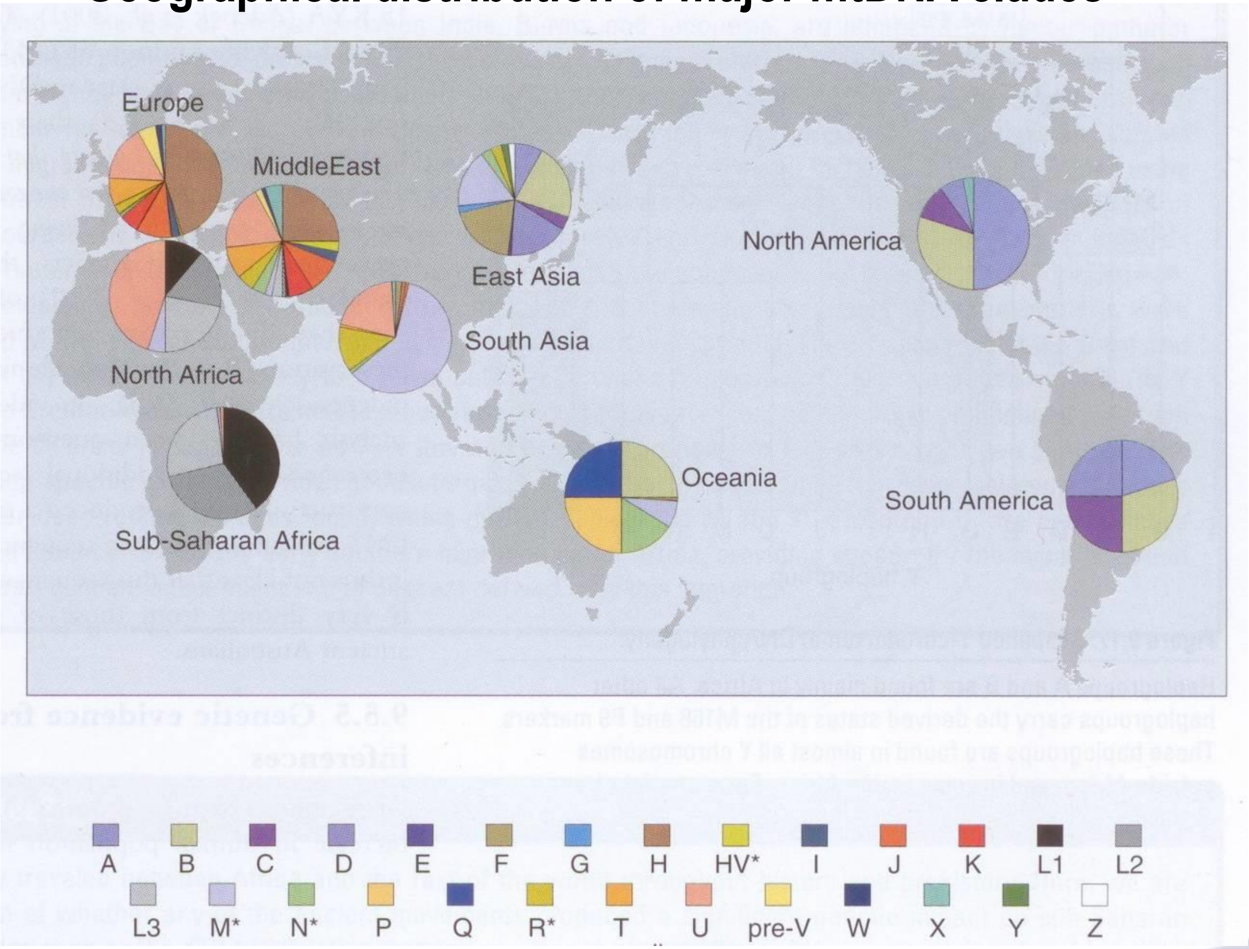
M. Hedman, V. Pimenoff, M. Lukka, P. Sistonen and A. Sajantila

Jobling MA and Tyler-Smith C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics* 4: 598-612

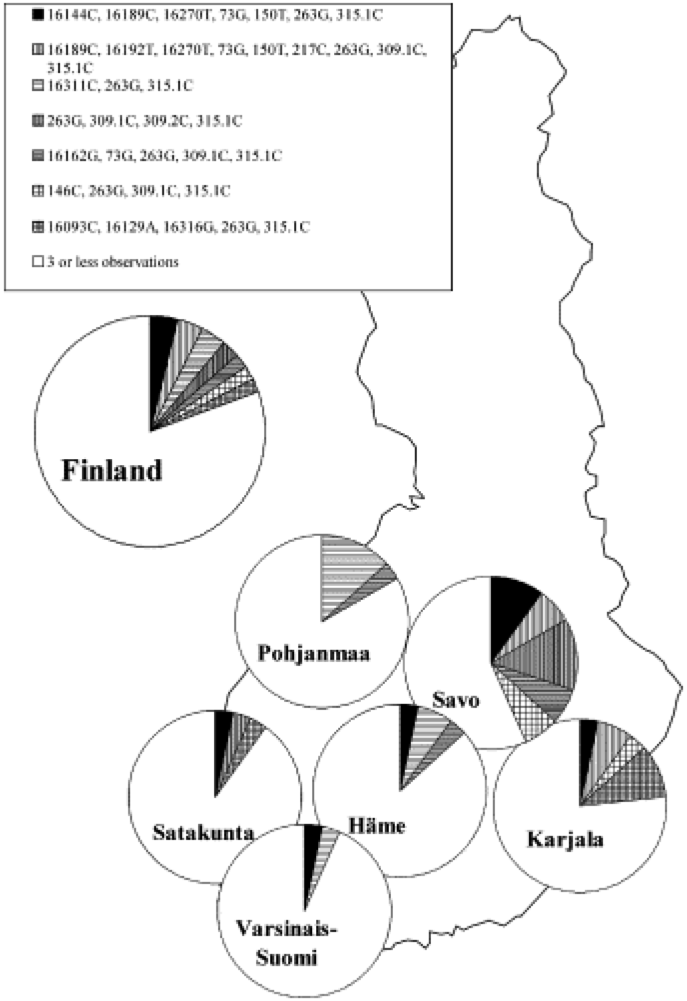
Slide courtesy of John M. Butler (NIST)

Lineage Markers…mtDNA

Geographical distribution of major mtDNA clades



Jobling, M. A., Hurles, M. E. and Tyler-Smith, C. (2004) *Human Evolutionary Genetics*. Garland Science: New York, USA, pp. 291.


■ 16144C, 16189C, 16270T, 73G, 150T, 263G, 315.1C
▥ 16189C, 16192T, 16270T, 73G, 150T, 217C, 263G, 309.1C, 315.1C
▤ 16311C, 263G, 315.1C
▥ 263G, 309.1C, 309.2C, 315.1C
▤ 16162G, 73G, 263G, 309.1C, 315.1C
⊞ 146C, 263G, 309.1C, 315.1C
▦ 16093C, 16129A, 16316G, 263G, 315.1C
□ 3 or less observations
Finland
Pohjanmaa
Savo
Satakunta
Häme
Karjala
Varsinais-Suomi

mtDNA
Finland

Finnish mitochondrial DNA HVS-I and HVS-II population data
Forensic Science International, ***In Press, Corrected Proof***, *Available online 2 March 2007*,
M. Hedman, A. Brandstätter, V. Pimenoff, P. Sistonen, J.U. Palo, W. Parson and A. Sajantila

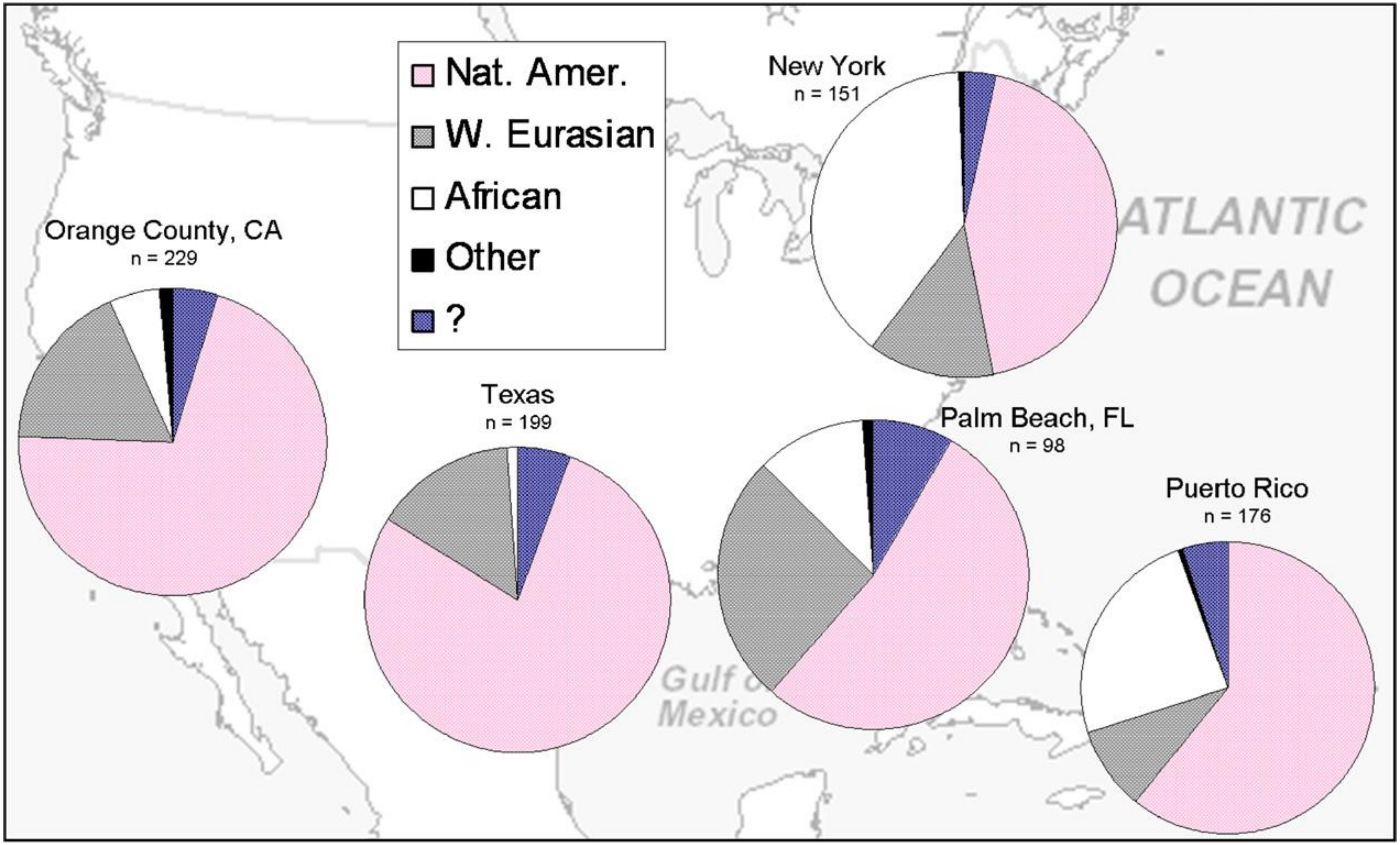
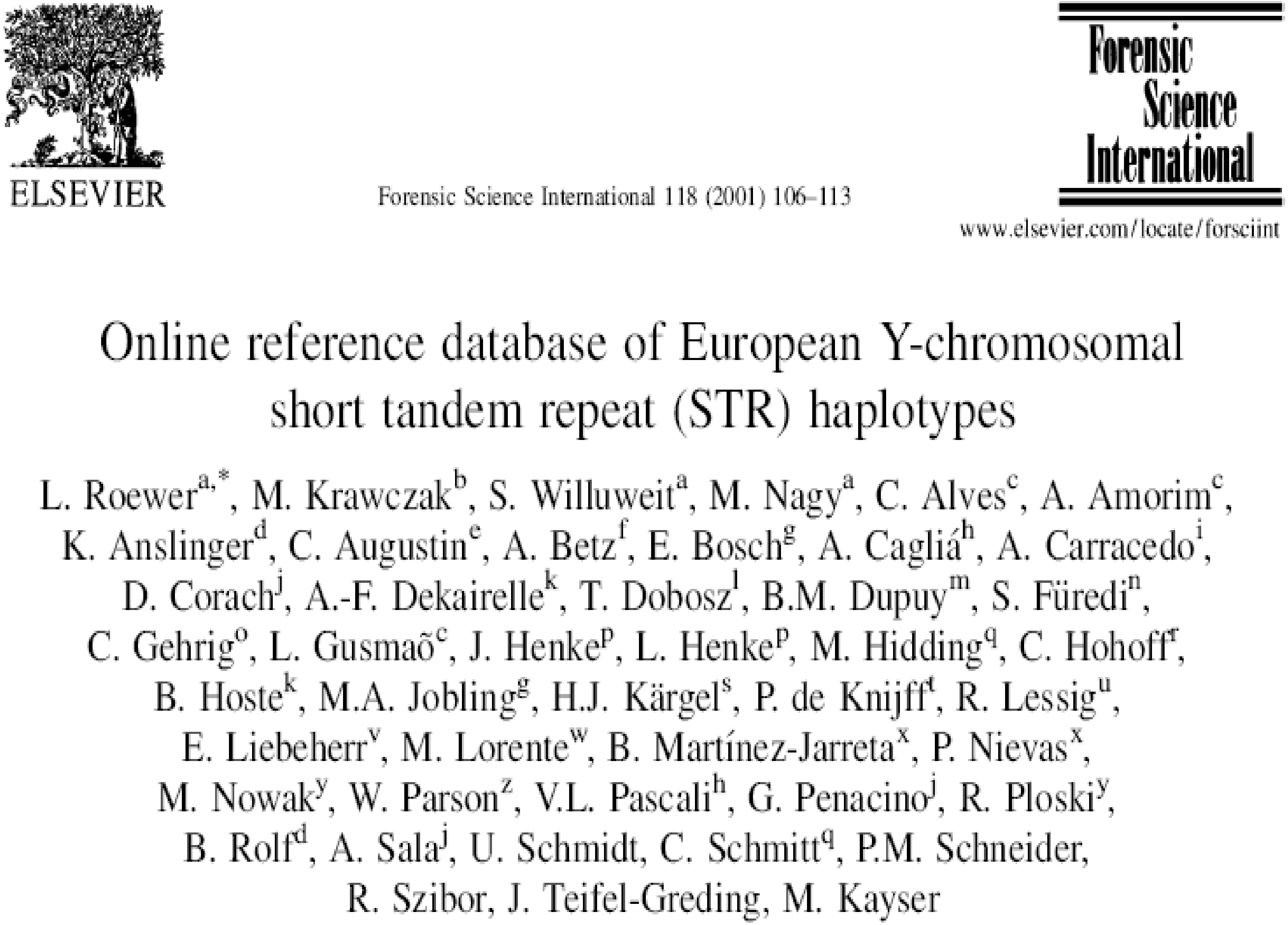


Fig. 2. Mitochondrial DNA haplogroup distribution among 853 regional United States ‘‘Hispanics’’. All inter-population pairwise Fst values are significant at the 0.05 level.



ELSEVIER

Forensic Science International 118 (2001) 106–113

Forensic Science International

www.elsevier.com/locate/forsciint

Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes

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

Y-Chromosome Haplotype Reference Database

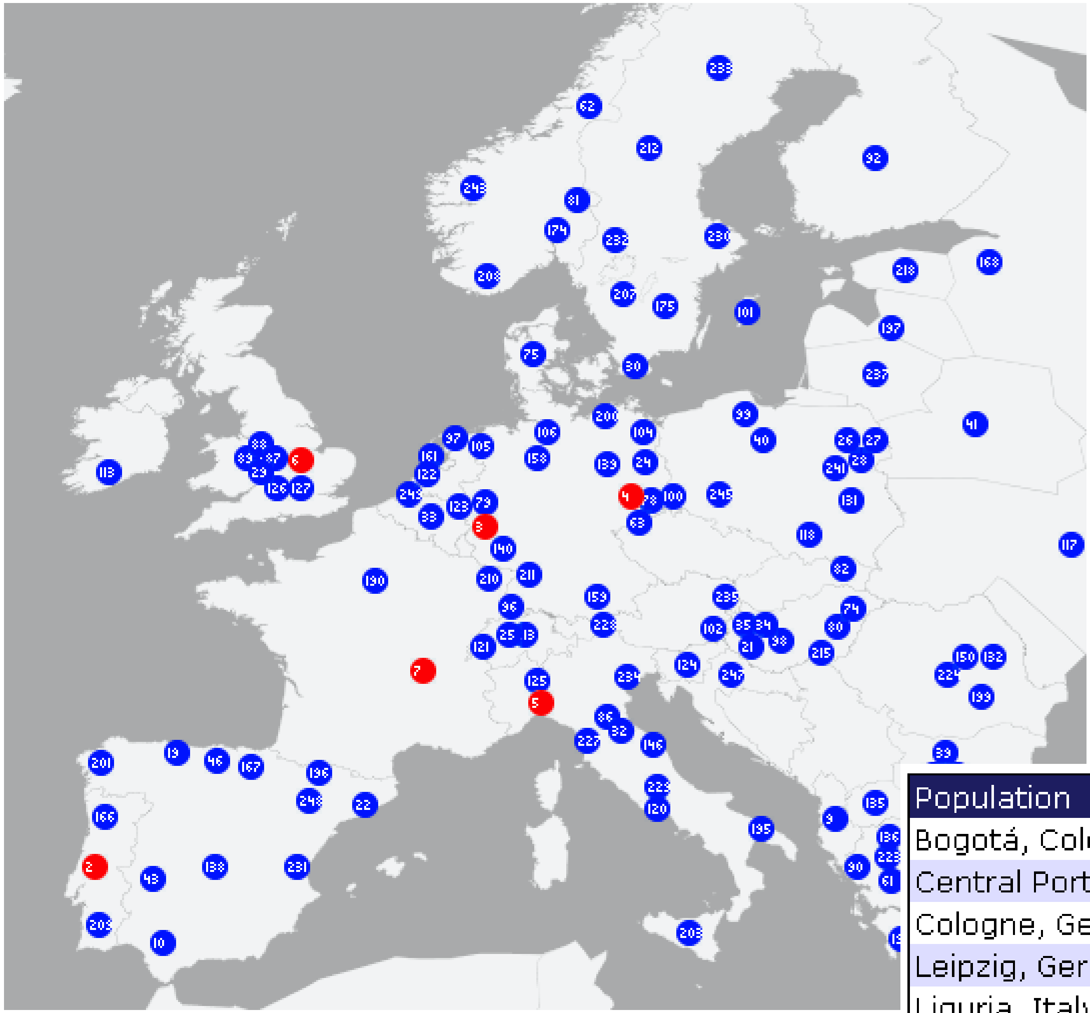
www.YHRD.org

Release "15" from 2004-12-17 16:11:24

7 matches in 27,773 individuals from 236 worldwide populations

Minimal Haplotype Result

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

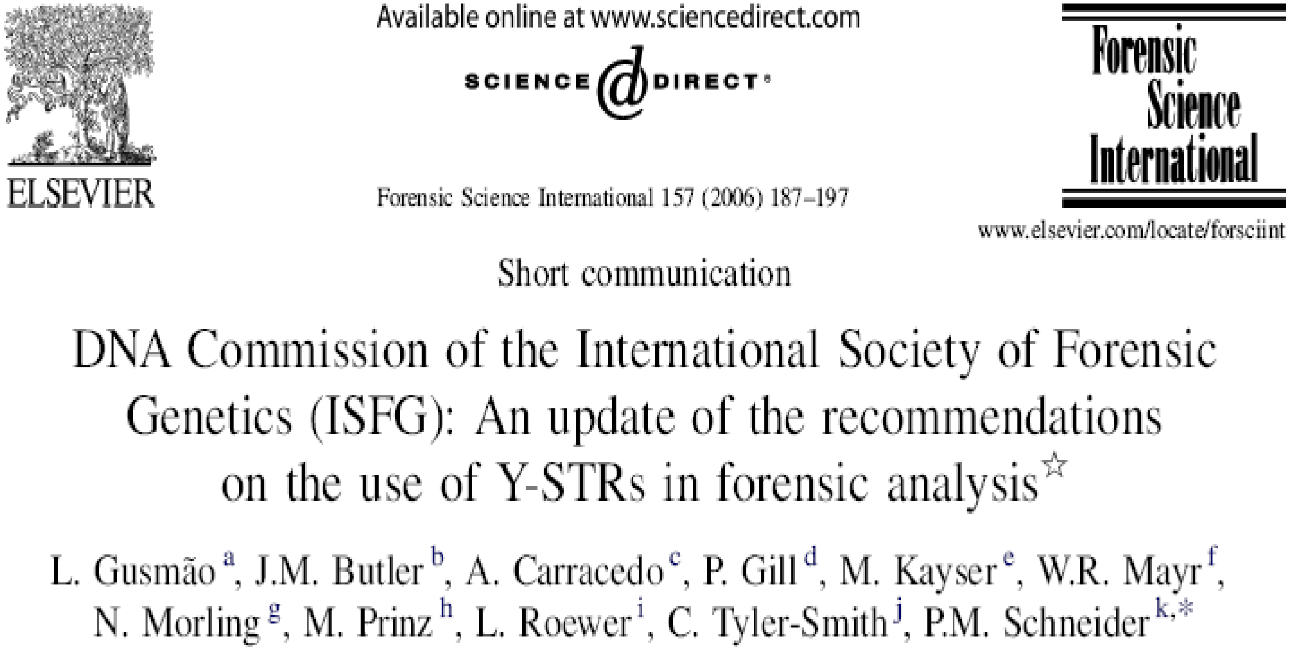


Population	#	Metapopulation
Bogotá, Colombia [European]	1 / 147	Eurasian MP / European MP
Central Portugal	1 / 230	Eurasian MP / European MP
Cologne, Germany	1 / 135	Eurasian MP / European MP
Leipzig, Germany	1 / 661	Eurasian MP / European MP
Liguria, Italy	1 / 81	Eurasian MP / European MP
London, UK	1 / 285	Eurasian MP / European MP
Lyon, France	1 / 125	Eurasian MP / European MP

Slide courtesy of
John M. Butler (NIST)

- ***“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)”***

Carracedo, A, Bär, W, Lincoln, P, Mayr, W, Morling, N, Olaisen, B, et al. DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Forensic Science International. 110(2000);(2):79-85



Available online at www.sciencedirect.com

SCIENCE DIRECT

Forensic Science International 157 (2006) 187–197

Forensic Science International

www.elsevier.com/locate/forsciint

Short communication

DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis☆

L. Gusmão[a], J.M. Butler[b], A. Carracedo[c], P. Gill[d], M. Kayser[e], W.R. Mayr[f], N. Morling[g], M. Prinz[h], L. Roewer[i], C. Tyler-Smith[j], P.M. Schneider[k,*]

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;

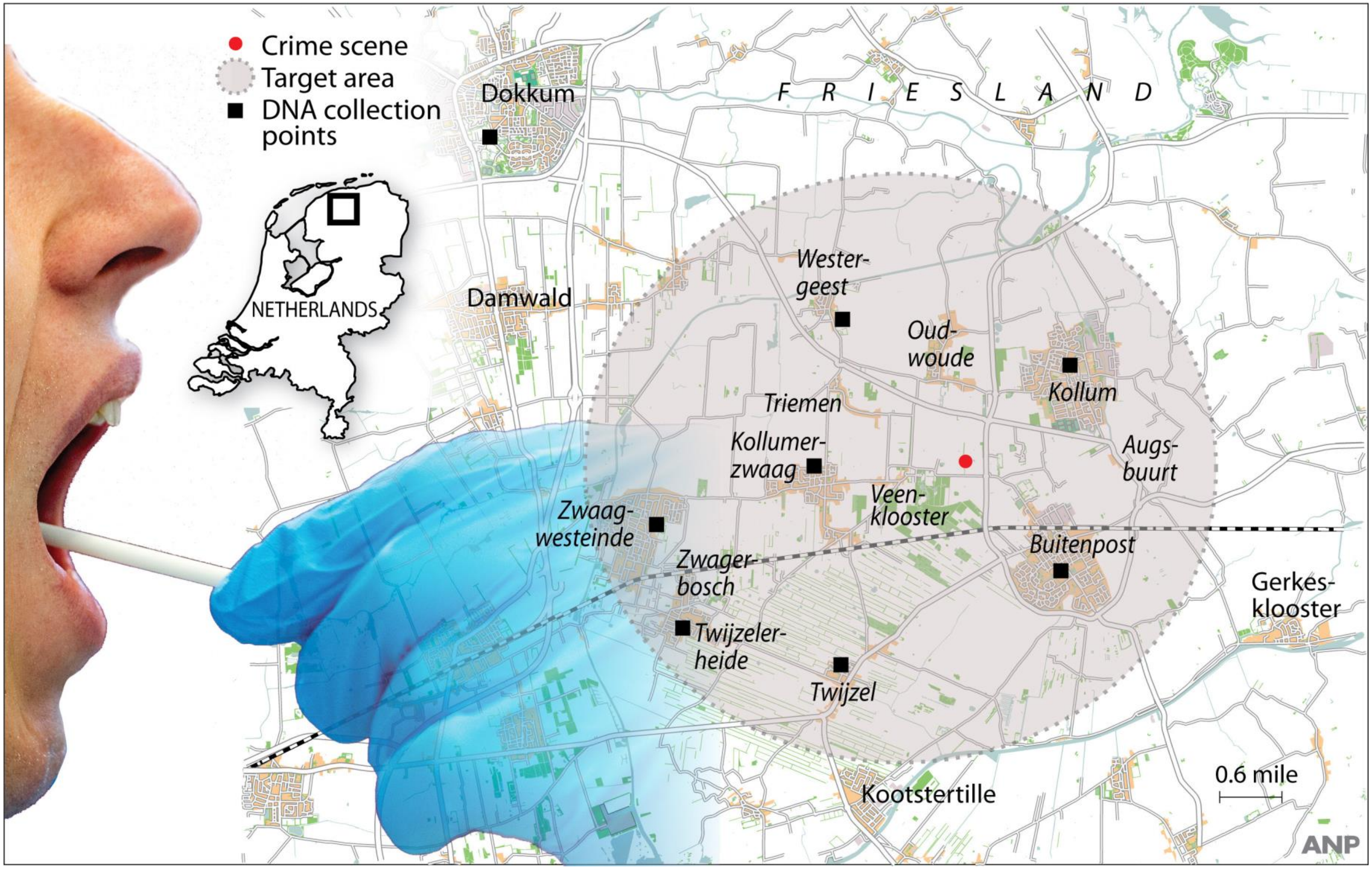
Ron Rintjema, Jelle Tjalsma and Cor Reijenga, 3-D team Friesland Police

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.


● Crime scene
Target area
■ DNA collection points
NETHERLANDS
FRIESLAND
Dokkum
Damwald
Westergeest
Oudwoude
Kollum
Triemen
Kollumerzwaag
Augsbuurt
Veenklooster
Zwaagwesteinde
Buitenpost
Zwagerbosch
Gerkesklooster
Twijzelerheide
Twijzel
Kootstertille
0.6 mile
ANP

Marianne Vaatstra case

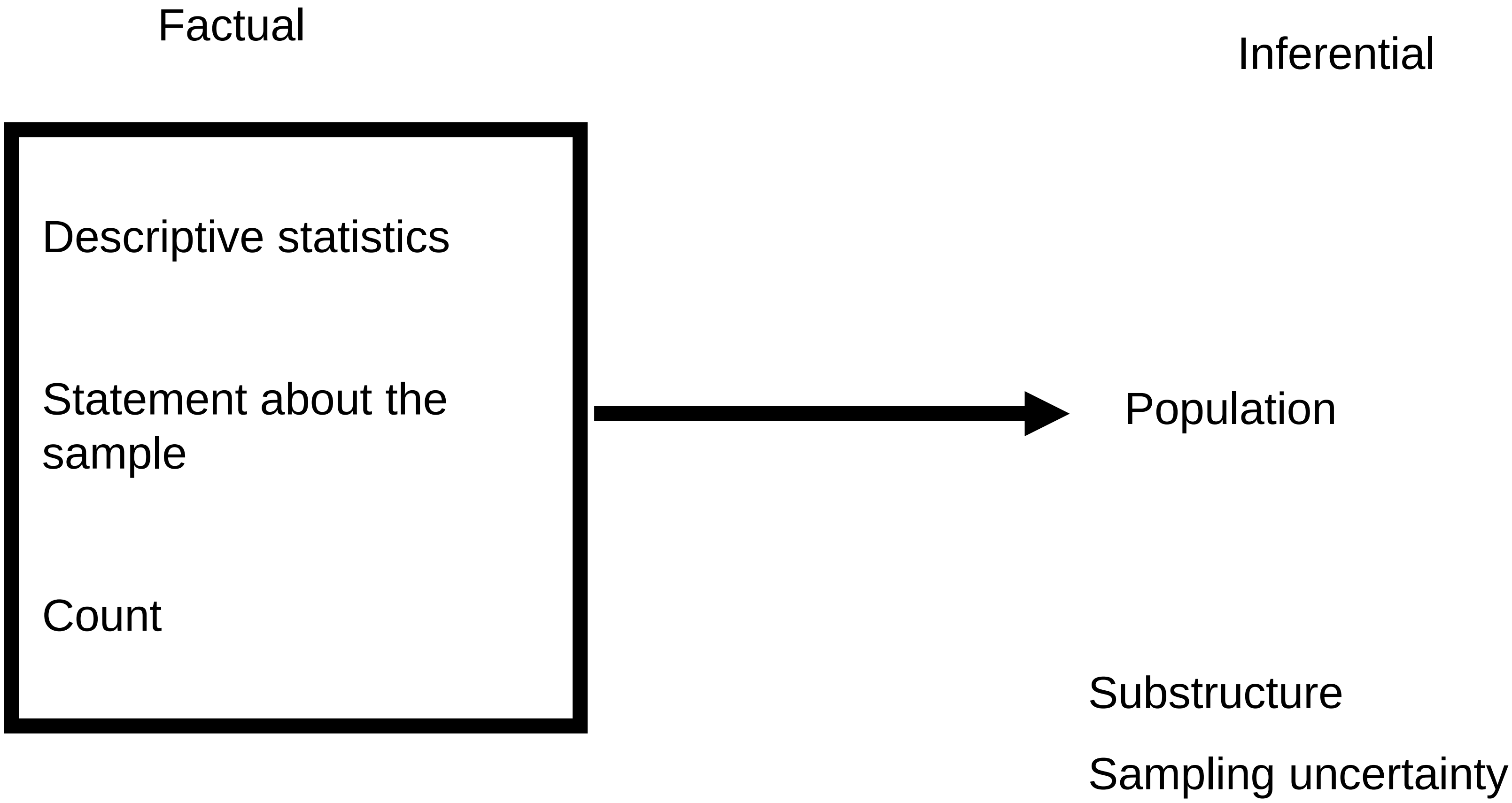
In 7 weeks …. generated 3880 Y-STR profiles.
23 men matched 17 of 17 Y-filer loci, surnames A, B, C.
5 men matched 16 of 17 Y-filer loci, surnames A, B, D, E.
7 men matched 15 of 17 Y-filer loci, surnames F, G, H.
These Y haplotypes corresponded to 8 different surnames.
Autosomal DNA indicated no parent-child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

$$\frac{0}{200,000}$$

$$\frac{23}{3,880}$$



Factual

Inferential

Descriptive statistics

Statement about the sample

Count

Population

Substructure

Sampling uncertainty

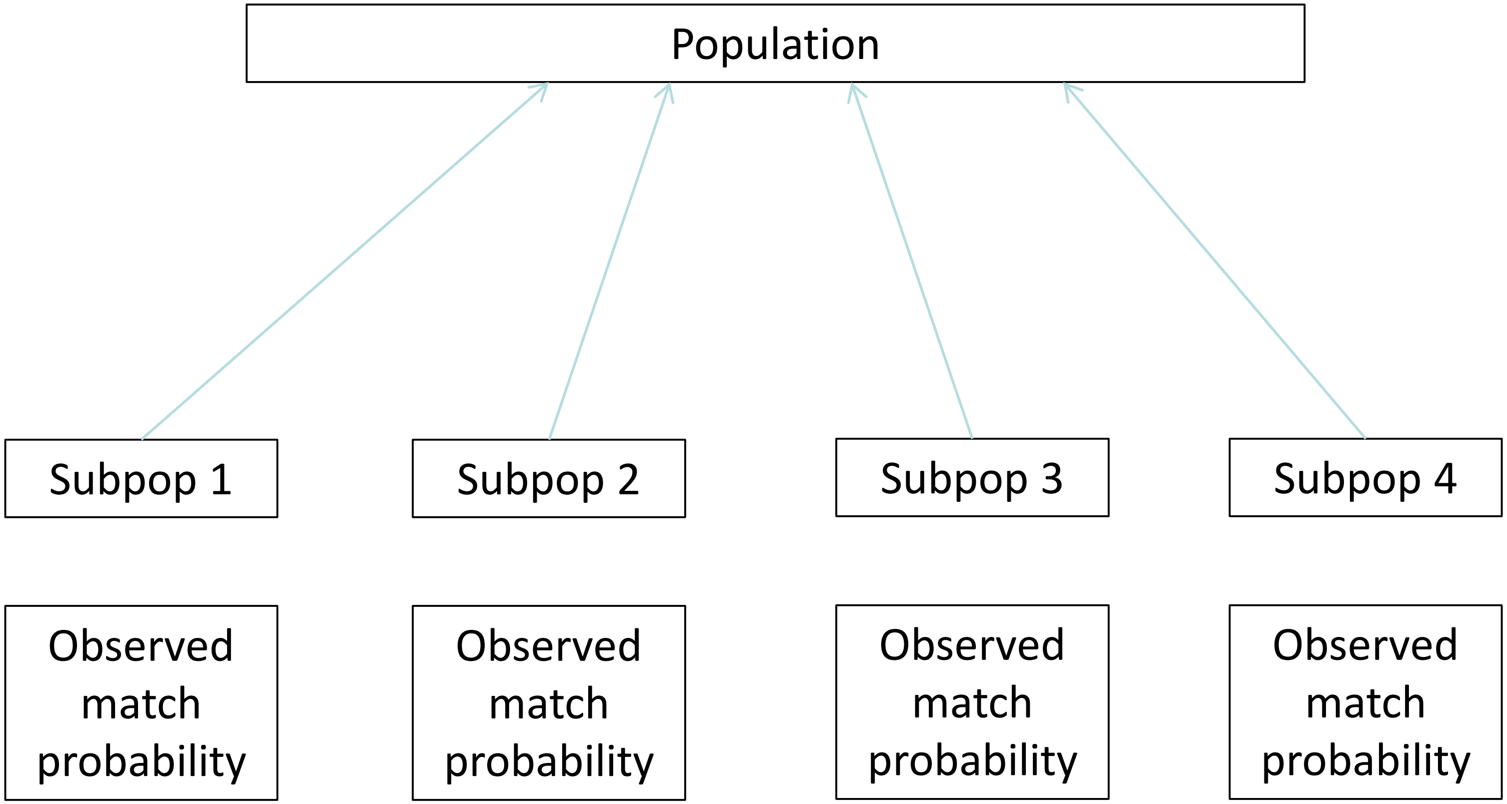
$$f' \approx \underset{\uparrow}{\theta} + (1-\theta) f$$

This will be dominant

Loci[3]	For African Americans, Asians, Caucasians & Hispanics	For African Americans, Asians, Caucasians, Hispanics & Native Americans
1	0.06	0.06
2	0.04	0.04
3	0.03	0.03
4	0.02	0.02
5	0.008	0.008
6	0.005	0.005
7	0.003	0.003
8	0.002	0.002
9	0.001	0.002
10	0.0006	0.002
11	0.0004	0.0009
12	0.0002	0.0007
13	0.0002	0.0006
14	0.0001	0.0005
15	0.00008	0.0005
16	0.00006	0.0004
17	0.00003	0.0004
18	0.00003	0.0004
19	0.00003	0.0003
20	0.00002	0.0003
21	0.00002	0.0003
22	0.00002	0.0003

Does it work? How can we test?

Need to test the prediction against the observed value.



Population

Subpop 1

Subpop 2

Subpop 3

Subpop 4

Observed match probability

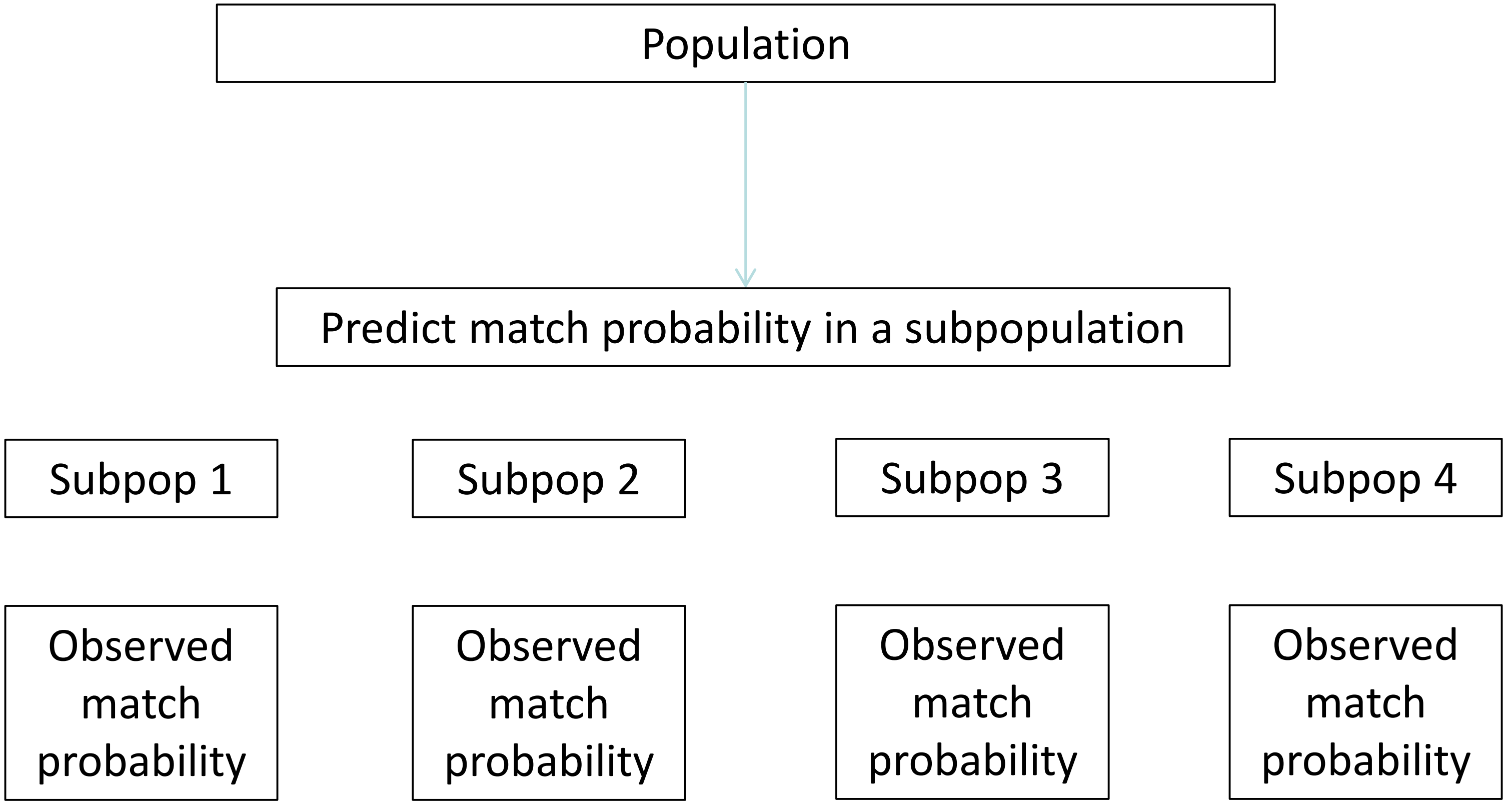
Observed match probability

Observed match probability

Observed match probability

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

Subpop 2

Subpop 3

Subpop 4

Observed match probability

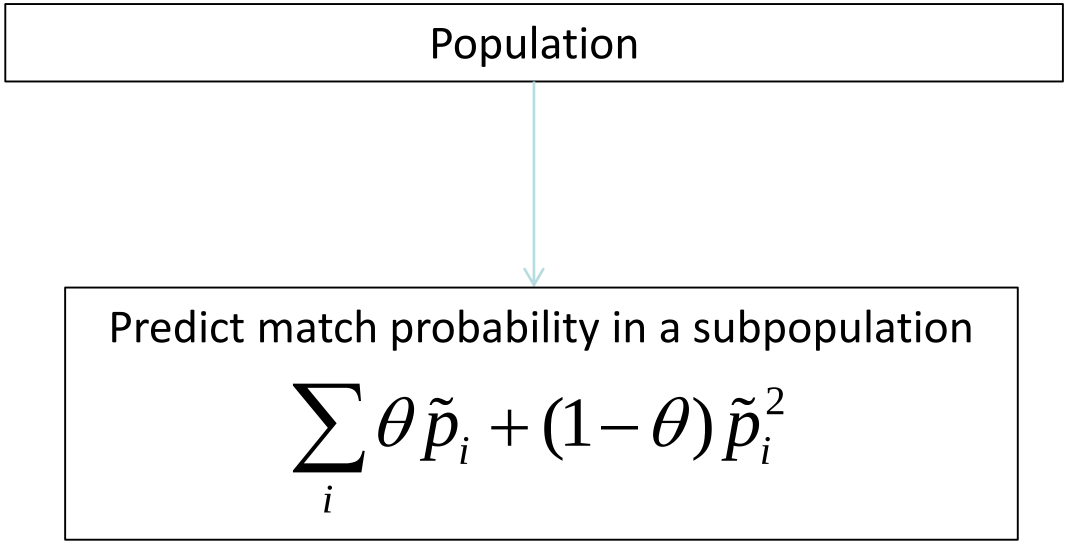
Observed match probability

Observed match probability

Observed match probability

Does it work? How can we test?

Need to test the prediction against the observed value.



Population

Predict match probability in a subpopulation

$$\sum_i \theta \tilde{p}_i + (1-\theta)\, \tilde{p}_i^2$$

This is the most important slide for this section

		$\hat{\beta}_W = \frac{\frac{r-1}{r}\frac{\tilde{M}_W - \tilde{M}_B}{1-\tilde{M}_B}}{1-\frac{1}{r}\frac{\tilde{M}_W - \tilde{M}_B}{1-\tilde{M}_B}}$	$\sum_u \tilde{p}_u^2$	**Observed M_W**	$\hat{M}_W = \hat{\beta} + (1-\hat{\beta})\sum_u \tilde{p}_u^2$
45 European meta populations	1st half	0.0085	0.0073	0.0161	0.0157
	2nd half	0.0085	0.0076	0.0156	0.0160
2 subpopulations, Eastern and Western	1st half	0.0064	0.0101	0.0156	0.0164
	2nd half	0.0056	0.0086	0.0135	0.0142
Belorussia, Kiev, Ljubljana, Moscow, Novgorod, Poland, Riga, Vilnius, Zagreb	1st half	0.0009	0.0115	0.0110	0.0124
	2nd half	0.0051	0.0113	0.0160	0.0164
Emilia Romagna, London, Portugal, Pyrenees, South Holland, Southern Ireland, Spain, Strasbourg	1st half	0.0040	0.0174	0.0275	0.0213
	2nd half	0.0029	0.0261	0.0263	0.0289

Credit Myers, Roewer, Weir, Willeuit, and Buckleton

UNITED STATES V. KOOTSWATEWA

- 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

UNITED STATES V. KOOTSWATEWA

- 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 …

https://casetext.com/case/united-states-v-kootswatewa-1

UNITED STATES V. KOOTSWATEWA

- 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

$$\hat{p}_x = C / D$$

- 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

$$\hat{p}_x = C / D$$

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

- Clopper CJ, Pearson ES. The use of confidence or fiducial intervals illustrated in the case of the binomial. Biometrika. 1934;26:404-13.

- Subpopulation correction – BS Weir

$$\hat{p}_x = \hat{\beta}_W + (1 - \hat{\beta}_W)\left(\frac{C}{D}\right)$$

- The Kappa method – Charles Brenner

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

κ 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.
- Andersen MM, Caliebe A, Jochens A, Willuweit S, Krawczak M. Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. Forensic Science International: Genetics. 2013;7:264-71.

- 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

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

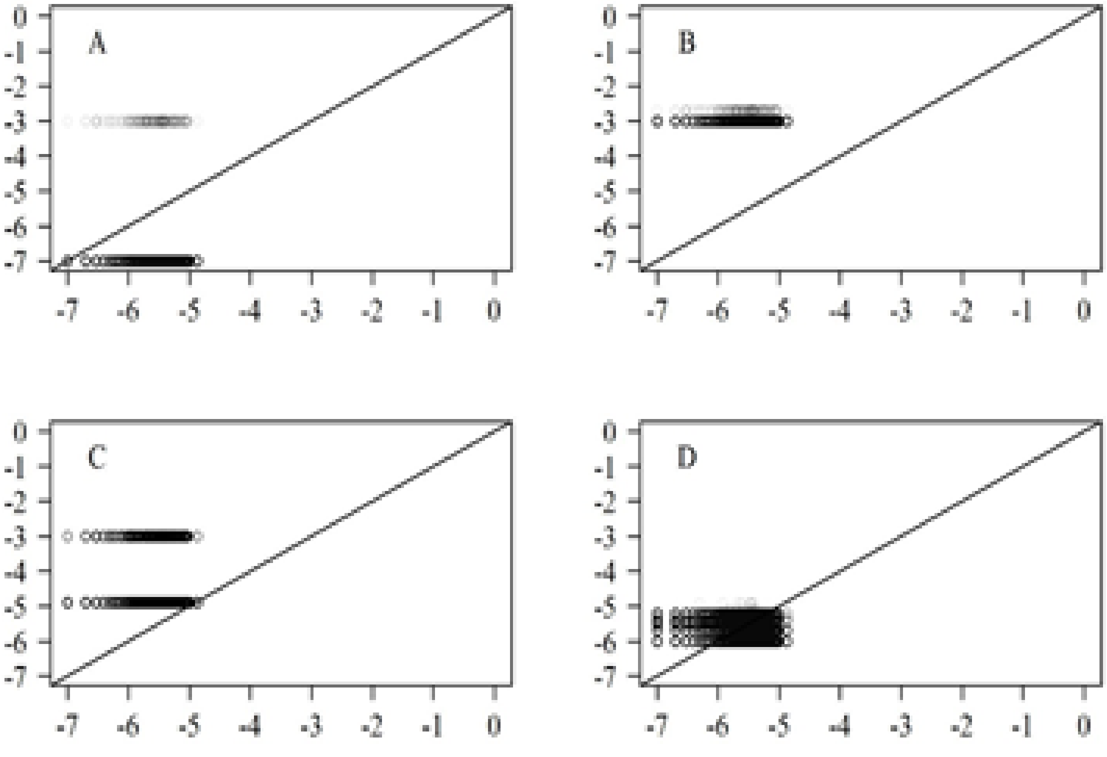
D_{C+1} for example, using $C = 1$, the D_2 is the number of matching pairs in the database

- THE COALESCENCE 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.

Wilson IJ, Weale ME, Balding DJ. Inferences from DNA Data: Population Histories, Evolutionary Processes and Forensic Match Probabilities. Journal of Royal Statistical Society Series A. 2003;166:155-201.

Counting Conservative

Augmented Counting



Subpopulation

Kappa

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.

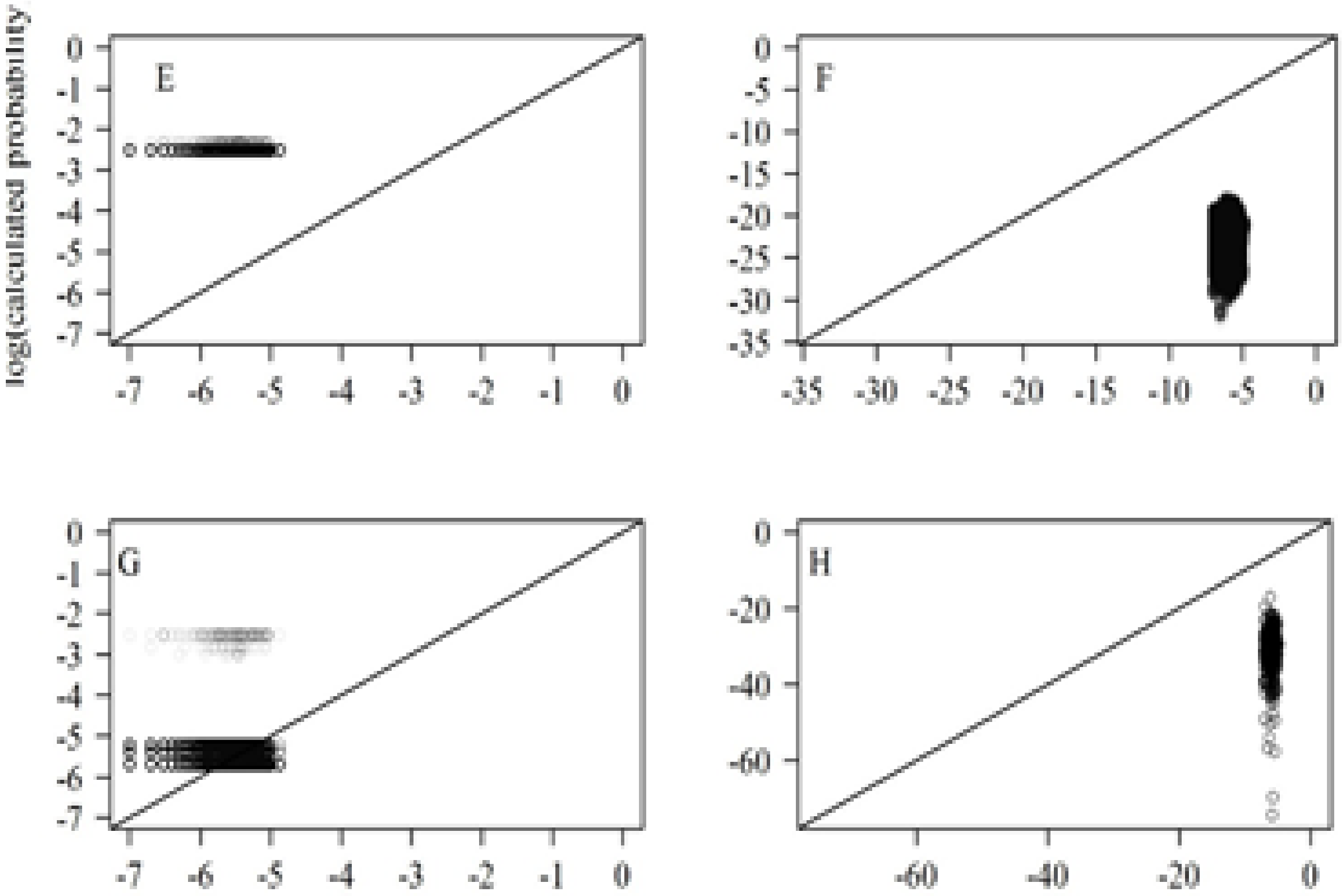
100 samples of size 1,000 are drawn from the whole population.

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:

Donor 1	Donor 2
a	b
c	d

This could be explained as:

ac:bd → do exist

ad:bc→ may or may not exist in the database or references

Please mentally extend to 23 or 27 loci

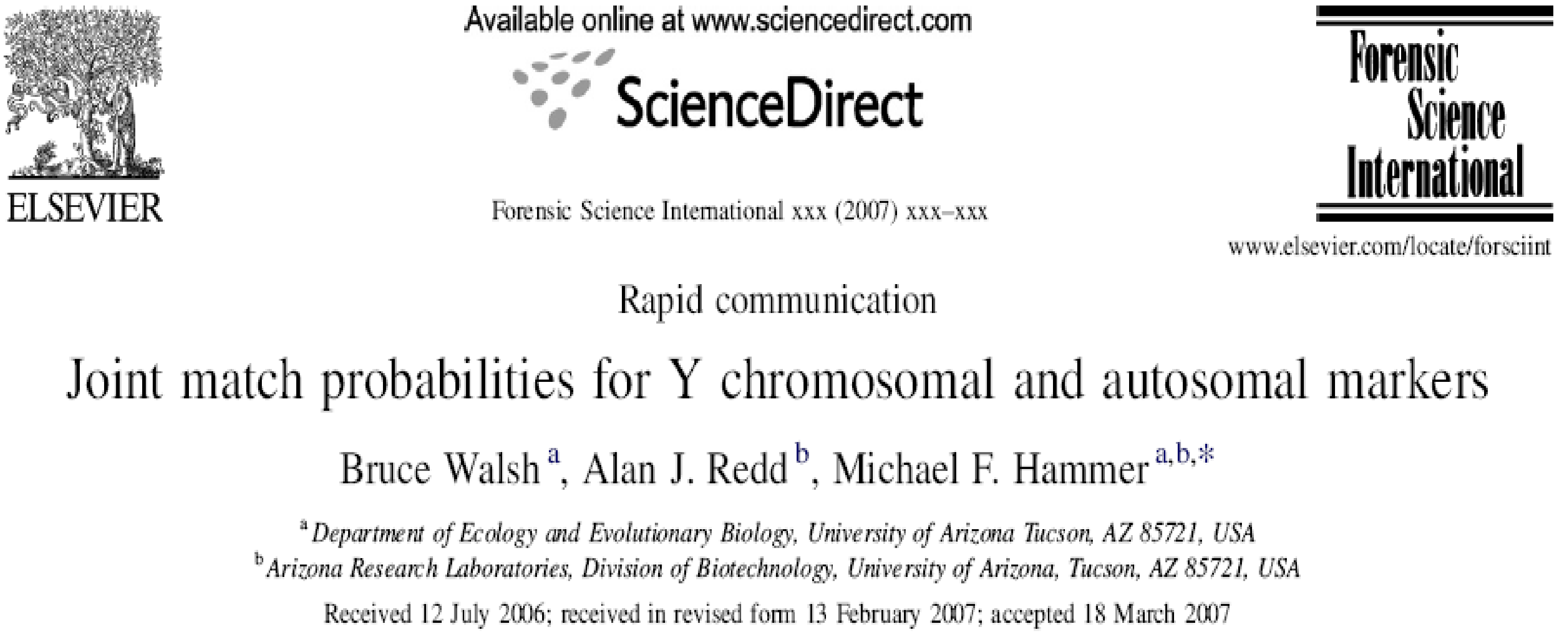
There are about 4 – 67 million haplotype combinations for this simple mix

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

ELSEVIER

Forensic Science International xxx (2007) xxx–xxx

Forensic Science International

www.elsevier.com/locate/forsciint

Rapid communication

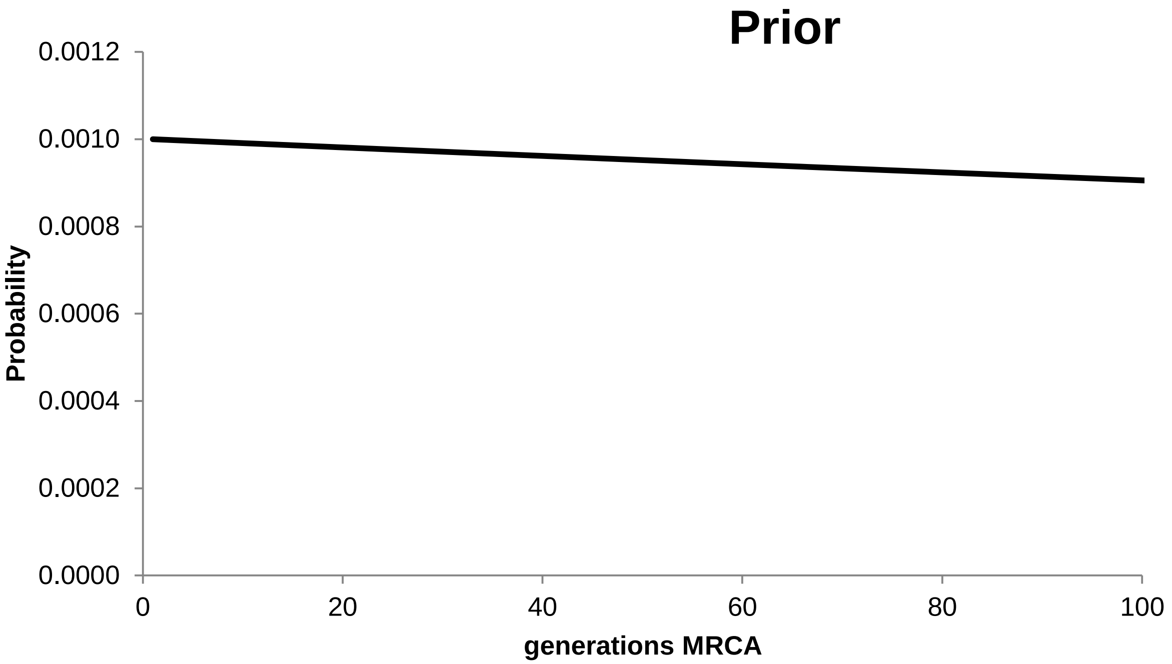
Joint match probabilities for Y chromosomal and autosomal markers

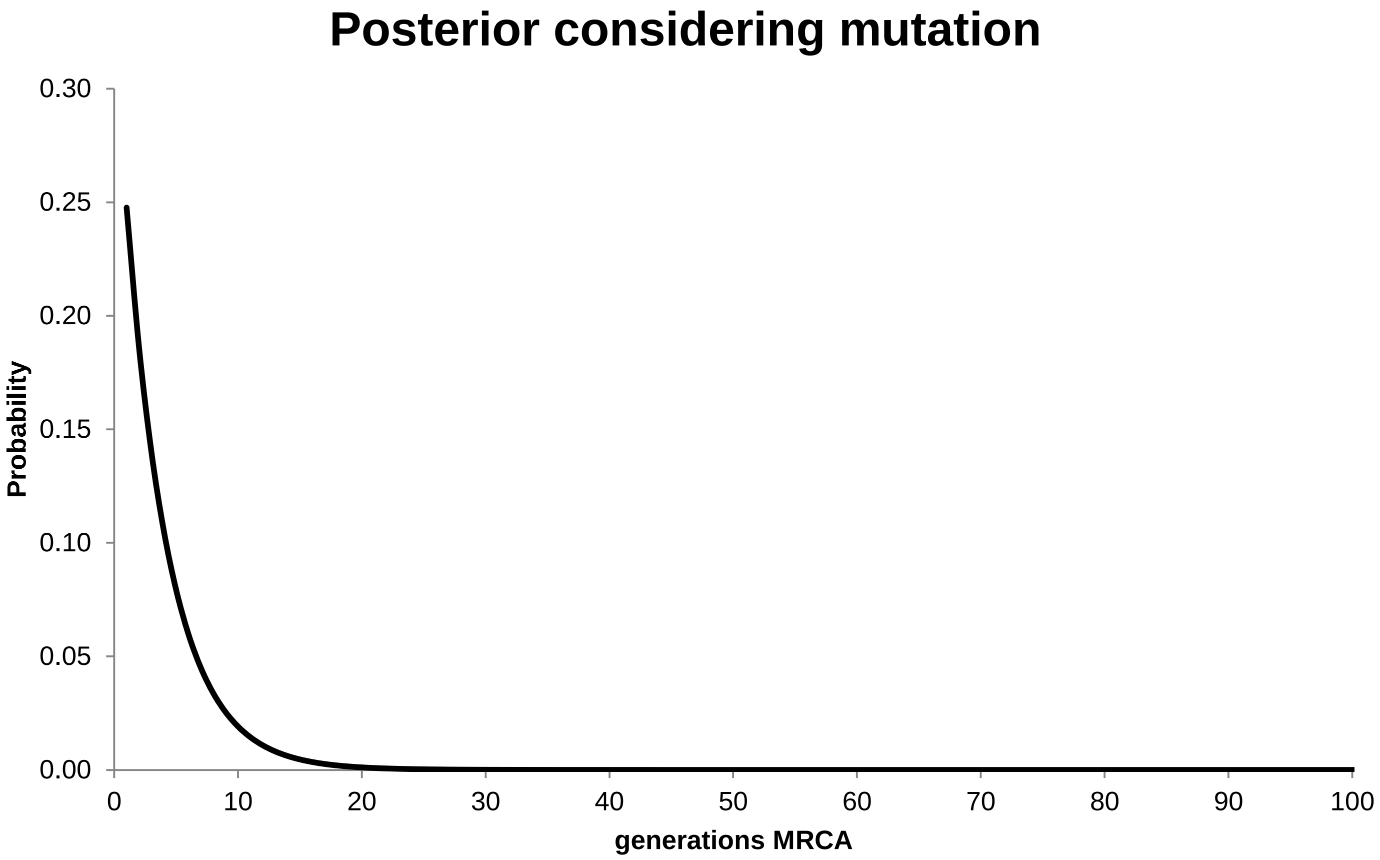
Bruce Walsh[a], Alan J. Redd[b], Michael F. Hammer[a,b,*]

[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
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Probability
0
20
40
60
80
100
generations MRCA


Posterior considering mutation
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Probability
0
10
20
30
40
50
60
70
80
90
100
generations MRCA

Use TMRCA posterior to compute theta for autosomal for “unrelated = not sibs or cousins”

Multiplex		PPY	Yfiler	PowerPlex Y23	Yfiler Plus
Loci (l)		11	16	22	25
$\mu_{ave.}$		0.0021	0.0026	0.0035	0.0057
	N_Y				
0.001	100	0.002	0.003	0.004	0.006
	1,000	0.002	0.002	0.004	0.005
	10,000	0.002	0.002	0.004	0.005
	100,000	0.002	0.002	0.004	0.005
0.01	100	0.011	0.012	0.013	0.014
	1,000	0.011	0.011	0.013	0.014
	10,000	0.011	0.011	0.013	0.014
	100,000	0.011	0.011	0.013	0.014
0.03	100	0.031	0.032	0.033	0.034
	1,000	0.031	0.031	0.033	0.034
	10,000	0.031	0.031	0.033	0.034
	100,000	0.031	0.031	0.033	0.034

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.