

The statistical evaluation of DNA crime stains in R

Miriam Marušíaková

Department of Statistics, Charles University, Prague
Institute of Computer Science, CBI, Academy of Sciences of CR

UseR! 2008, Dortmund

Introduction

Single Crime Scene

- Independence

- Population substructure

- Relatedness

- R package forensic

Mixed stain

- Independence

- Population substructure

- People versus Simpson

- More general situation

Conclusion

Outline

Introduction

Single Crime Scene

- Independence

- Population substructure

- Relatedness

- R package forensic

Mixed stain

- Independence

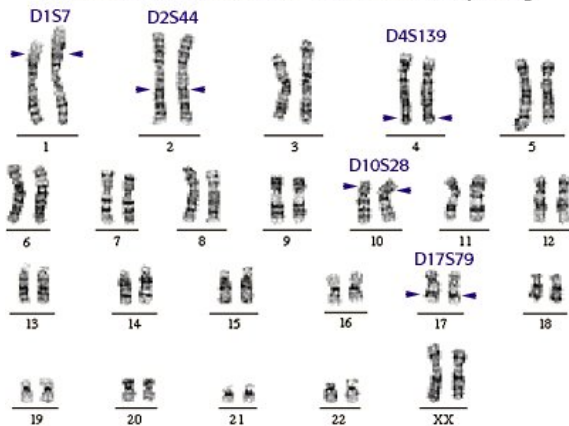
- Population substructure

- People versus Simpson

- More general situation

Conclusion

Chromosomal locations of RFLP markers used in DNA profiling



Human Female Karyotype

Single crime scene stain

- ▶ Blood stain at the crime scene
- ▶ Believed it was left by offender
- ▶ Suspect arrested for reasons unconnected with his DNA profile
- ▶ Crime sample, suspect sample

Hypothesis

- ▶ H_p (prosecution): The suspect left the crime stain.
- ▶ H_d (defense): Some other person left the crime stain.

Notation

- ▶ DNA typing results $E = \{G_C, G_S\}$
- ▶ non-DNA evidence I

Evidence interpretation

- ▶ Prior odds

$$\frac{Pr(H_p|I)}{Pr(H_d|I)}$$

- ▶ Posterior odds

$$\frac{Pr(H_p|E, I)}{Pr(H_d|E, I)}$$

- ▶ Bayes' theorem

$$\frac{Pr(H_p|E, I)}{Pr(H_d|E, I)} = \underbrace{\frac{Pr(E|H_p, I)}{Pr(E|H_d, I)}}_{LR} \times \frac{Pr(H_p|I)}{Pr(H_d|I)}$$

- ▶ Balding and Donnelly (1995), Robertson and Vignaux (1995)

Evidence interpretation

$$\begin{aligned}
 LR &= \frac{Pr(E|H_p, I)}{Pr(E|H_d, I)} = \frac{Pr(G_S, G_C|H_p, I)}{Pr(G_S, G_C|H_d, I)} \\
 &= \frac{Pr(G_C|G_S, H_p, I)}{Pr(G_C|G_S, H_d, I)} \times \frac{Pr(G_S|H_p, I)}{Pr(G_S|H_d, I)} \\
 &= \frac{Pr(G_C|G_S, H_p, I)}{Pr(G_C|G_S, H_d, I)} \times 1 \\
 &= \frac{1}{Pr(G_C|G_S, H_d, I)} \\
 &= \frac{1}{Pr(G_C|H_d, I)} \quad \text{if independence assumed}
 \end{aligned}$$

Errors and fallacies

Statement

The probability of observing this type
if the blood came from someone other than suspect
is 1 in 100.

$$Pr(G_C|H_d, I) = 1/100$$

Common error

The probability that the blood came from someone else is 1 in 100.

$$Pr(H_d|G_C, I) = 1/100$$

There is a 99% probability that it came from the suspect.

Errors and fallacies (cont.'d)

Statement

The evidence is 100 times more probable **if**
the suspect left the crime stain **than if** some unknown person left it.

$$\frac{1}{Pr(G_C|H_d, I)} = 100$$

Common error

It is 100 times more probable that
the suspect left the crime stain than some unknown person.

$$\frac{Pr(H_p|G_C, I)}{Pr(H_d|G_C, I)} = 100$$

Outline

Introduction

Single Crime Scene

- Independence

- Population substructure

- Relatedness

- R package forensic

Mixed stain

- Independence

- Population substructure

- People versus Simpson

- More general situation

Conclusion

Outline

Introduction

Single Crime Scene

Independence

Population substructure

Relatedness

R package forensic

Mixed stain

Independence

Population substructure

People versus Simpson

More general situation

Conclusion

Product rule

Model of an ideal population

- ▶ Infinite size
- ▶ Random mating
- ▶ Model reliable for most real-world problems

Hardy-Weinberg equilibrium

$$Pr(G = A_i A_i) = p_i^2$$

$$Pr(G = A_i A_j) = 2p_i p_j, \quad i \neq j$$

Likelihood ratio

$$LR = \frac{1}{Pr(G_C | G_S, H_d, I)} = \frac{1}{Pr(G_C | H_d, I)} = \frac{1}{p_i^2} \left(\frac{1}{2p_i p_j} \right)$$

Outline

Introduction

Single Crime Scene

Independence

Population substructure

Relatedness

R package forensic

Mixed stain

Independence

Population substructure

People versus Simpson

More general situation

Conclusion

Population substructure

- ▶ F - measure of uncertainty about allele proportions in the population of suspects
- ▶ Genetic interpretation of F (Wright, 1951)
- ▶ How to estimate F (Weir and Cockerham, 1984)
- ▶ Recommendations (National Research Council, 1996)
 - $F = 0.01$ large subpopulations (USA)
 - $F = 0.03$ small isolated subpopulations
- ▶ Match probabilities (Balding and Nichols, 1994)

$$P(G_C = A_i A_i | G_S = A_i A_i) = \frac{[2F + (1 - F)p_i] [3F + (1 - F)p_i]}{(1 + F)(1 + 2F)}$$

$$P(G_C = A_i A_j | G_S = A_i A_j) = \frac{2 [F + (1 - F)p_i] [F + (1 - F)p_j]}{(1 + F)(1 + 2F)}$$

Effects of F corrections

Likelihood ratio - some numerical values

- ▶ Heterozygotes A_iA_j , $p_i = p_j = p$

	$F = 0$	$F = 0.001$	$F = 0.01$	$F = 0.03$
$p = 0.01$	5 000	4 152	1 295	346
$p = 0.05$	200	193	145	89
$p = 0.10$	50	49	43	34

- ▶ Homozygotes A_iA_i , $p_i = p$

	$F = 0$	$F = 0.001$	$F = 0.01$	$F = 0.03$
$p = 0.01$	10 000	6 439	863	157
$p = 0.05$	400	364	186	73
$p = 0.10$	100	96	67	37

Outline

Introduction

Single Crime Scene

Independence

Population substructure

Relatedness

R package forensic

Mixed stain

Independence

Population substructure

People versus Simpson

More general situation

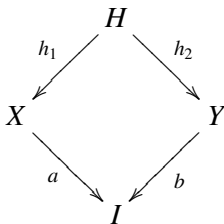
Conclusion

Inbreeding

- ▶ Individuals with common ancestors - **related**
- ▶ Their children - **inbred**
- ▶ Alleles are **ibd** (identical by descent) - copies of the same allele

Example

Alleles h_1, h_2 transmitted from parent H to X, Y who transmit a, b to I



$$Pr(h_1 \text{ is ibd to } h_2) = Pr(h_1 \equiv h_2) = 0.5$$

$$Pr(a \equiv b) = Pr(a \equiv h_1, b \equiv h_2 | h_1 \equiv h_2) Pr(h_1 \equiv h_2) = 0.5^3 = 0.125$$

Match probabilities for close relatives

Balding and Nichols (1994)

$$Pr(G_C | G_S, H_d, I) = \left\{ \begin{array}{l} (k_0 p_i^4 + k_1 p_i^3 + k_2 p_i^2) / p_i^2 \\ (4k_0 p_i^2 p_j^2 + k_1 p_i p_j (p_i + p_j) + 2k_2 p_i p_j) / 2 p_i p_j \end{array} \right.$$

Kinship coefficients

Relationship	k_0	k_1	k_2
Parent - child	0	1	0
Siblings	1/4	1/2	1/4
Grandparent - grandchild	1/2	1/2	0
Uncle - nephew	1/2	1/2	0
Cousins	3/4	1/4	0
Unrelated	1	0	0

k_i - probability that two persons will share i alleles ibd

$i = 0, 1, 2$

Outline

Introduction

Single Crime Scene

Independence

Population substructure

Relatedness

R package forensic

Mixed stain

Independence

Population substructure

People versus Simpson

More general situation

Conclusion

R function Pmatch

Usage

```
Pmatch(prob, k = c(1, 0, 0), theta = 0)
```

Example

```
> p <- c(0.057, 0.160, 0.182, 0, 0.024, 0.122)
> Pmatch(p)
```

```
$prob
```

	locus 1	locus 2	locus 3
allele 1	0.057	0.182	0.024
allele 2	0.160	0.000	0.122

```
$match
```

	locus 1	locus 2	locus 3
[1,]	0.01824	0.033124	0.005856

```
$total_match
```

```
[1] 3.538088e-06
```

Outline

Introduction

Single Crime Scene

- Independence

- Population substructure

- Relatedness

- R package forensic

Mixed stain

- Independence

- Population substructure

- People versus Simpson

- More general situation

Conclusion

Mixtures

- ▶ Prosecution and defense hypothesis

H_p : Contributors were the victim and the suspect.

H_d : Contributors were the victim and an unknown person.

- ▶ Likelihood ratio for the mixture

$$\begin{aligned}
 LR &= \frac{\Pr(E_C, G_V, G_S | H_p, I)}{\Pr(E_C, G_V, G_S | H_d, I)} \\
 &= \frac{\Pr(E_C | G_V, G_S, H_p, I)}{\Pr(E_C | G_V, G_S, H_d, I)} \times \frac{\Pr(G_V, G_S | H_p, I)}{\Pr(G_V, G_S | H_d, I)} \\
 &= \frac{\Pr(E_C | G_V, G_S, H_p, I)}{\Pr(E_C | G_V, G_S, H_d, I)} = \frac{1}{\Pr(E_C | G_V, G_S, H_d, I)} \\
 &= \frac{1}{\Pr(E_C | G_V, H_d, I)} \quad (\text{if independence assumed})
 \end{aligned}$$

Outline

Introduction

Single Crime Scene

- Independence

- Population substructure

- Relatedness

- R package forensic

Mixed stain

- Independence**

- Population substructure

- People versus Simpson

- More general situation

Conclusion

Match probability (independence assumption)

Weir et al (1997)

$$P_x(U, C) = (T_0)^{2x} - \sum_{i \in U} (T_{1i})^{2x} + \sum_{i, j \in U: i < j} (T_{2ij})^{2x} - \dots$$

$$T_0 = \sum_{l \in C} p_l$$

$$T_{1i} = \sum_{l \in C \setminus \{i\}} p_l, \quad i \in U$$

$$T_{2ij} = \sum_{l \in C \setminus \{i, j\}} p_l, \quad i, j \in U, \quad i < j$$

U - set of alleles from the crime sample C not carried by known contributors

x - number of unknown contributors

R

- ▶ `Pevind(alleles, prob, x, u = NULL)`
- ▶ `LR.ind(alleles, prob, x1, x2, u1 = NULL, u2 = NULL)`

Outline

Introduction

Single Crime Scene

Independence

Population substructure

Relatedness

R package forensic

Mixed stain

Independence

Population substructure

People versus Simpson

More general situation

Conclusion

Match probability for structured population

Assumption

All the involved people come from the same subpopulation with parameter F

Fung and Hu (2000), Zoubková and Zvárová (2004)

$$MP = \sum_{r_1=0}^r \sum_{r_2=0}^{r-r_1} \dots \sum_{r_{c-1}=0}^{r-r_1-\dots-r_{c-2}} \frac{(2n_U)! \prod_{i=1}^c \prod_{j=t_i+v_i}^{t_i+u_i+v_i-1} [(1-F)p_i + jF]}{\prod_{i=1}^c u_i! \prod_{j=2n_T+2n_V}^{2n_T+2n_U+2n_V-1} [(1-F) + jF]}$$

R

- ▶ `Pevid.gen(alleles, prob, x, T = NULL, V = NULL, theta = 0)`

T, V - genotypes of known contributors, known non-contributors

Outline

Introduction

Single Crime Scene

- Independence

- Population substructure

- Relatedness

- R package forensic

Mixed stain

- Independence

- Population substructure

- People versus Simpson**

- More general situation

Conclusion

People versus Simpson (Los Angeles County Case)

DNA evidence

- ▶ Material at the crime scene - alleles a, b, c (locus D2S44)
- ▶ Suspect's genotype ab
- ▶ Victim's genotype ac

Hypotheses (Weir et al, 1997, Fung and Hu, 2000)

H_p : Contributors were the victim, suspect and m unknowns.

H_d : Contributors were n unknowns.

R

- ▶ $a = c('a', 'b', 'c')$
- ▶ $p = c(0.0316, 0.0842, 0.0926)$
- ▶ `suspect <- 'a/b', victim <- 'a/c'`

Likelihood ratios for the Simpson case

Defense Prosecution	$n = 2$			$n = 3$		
	$F = 0$	0.01	0.03	$F = 0$	0.01	0.03
$m = 0$	1623	739	276	21606	5853	1150
$m = 1$	70	44	26	938	345	107

Prosecution proposition ($F = 0$)

- ▶ `Pevind.ind(alleles = a, prob = p, x = m)`
- ▶ `Pevind.gen(alleles = a, prob = p, x = m,
T = c(victim, suspect))`

Defense proposition ($F = 0$)

- ▶ `Pevind.ind(alleles = a, prob = p, x = n,
u = c('a', 'b', 'c'))`
- ▶ `Pevind.gen(alleles = a, prob = p, x = n)`

Outline

Introduction

Single Crime Scene

Independence

Population substructure

Relatedness

R package forensic

Mixed stain

Independence

Population substructure

People versus Simpson

More general situation

Conclusion

More general situation

Contributors from different ethnic groups

Fung and Hu (2001)

$P_{\text{evid.ind}}$, LR_{ind}

(independence within and between ethnic groups)

Presence of related people

Hu and Fung (2003)

Example

Suspect not typed, his relative is tested

Two related people among unknown contributors

$P_{\text{evid.rel}}$

Outline

Introduction

Single Crime Scene

- Independence

- Population substructure

- Relatedness

- R package forensic

Mixed stain

- Independence

- Population substructure

- People versus Simpson

- More general situation

Conclusion

Conclusion

Single crime stain

- ▶ Suspect and offender are unrelated
- ▶ are members of the same subpopulation
- ▶ are close relatives

P_{match}

Mixed crime stain

- ▶ Contributors unrelated
- ▶ members of the same subpopulation
- ▶ may be related
- ▶ from different ethnic groups

$P_{\text{evid.ind}}, LR_{\text{ind}}$

$P_{\text{evid.gen}}$

$P_{\text{evid.rel}}$

Thank you for your attention!

Acknowledgement

The work was supported by GAČR 201/05/H007 and the project 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic.

Important publications



BALDING AND NICHOLS(1994)

DNA profile match probability calculation

Forensic Science International 64, p. 125-140



WEIR ET AL(1997)

Interpreting DNA mixtures

Journal of Forensic Sciences 42, p. 213-222



FUNG AND HU(2000)

Interpreting forensic DNA mixtures: allowing for uncertainty in population substructure and dependence

Journal of Royal Statistical Society A 163, p. 241-254



ZOUBKOVÁ, SUPERVISOR ZVÁROVÁ (2004)

Master thesis (in Czech)

Charles University, Prague

Software



R DEVELOPMENT CORE TEAM (2006)

R: A language and environment for statistical computing.

R Foundation for Statistical Computing, Vienna, Austria

ISBN 3-900051-07-0, URL <http://www.R-project.org>



MARUŠIAKOVÁ, M (2007)

forensic: Statistical methods in forensic genetics

R package version 0.2