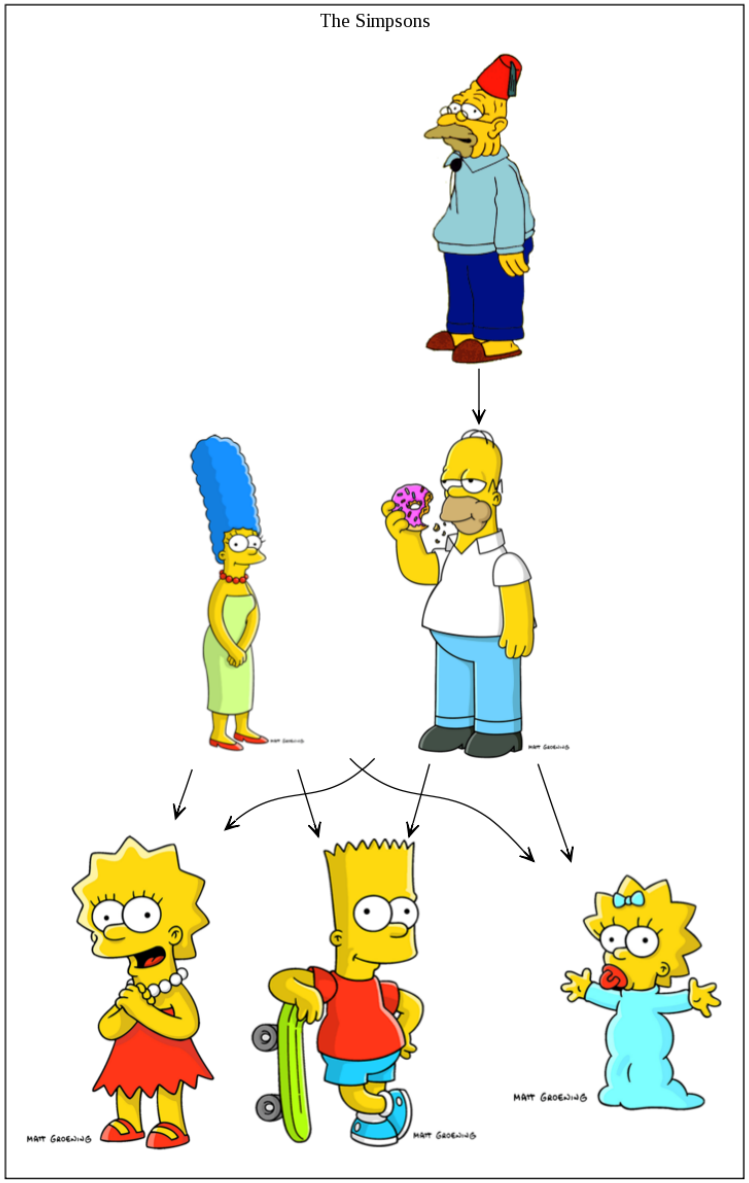
FRANz: Reconstruction of wild pedigrees

Markus Riester

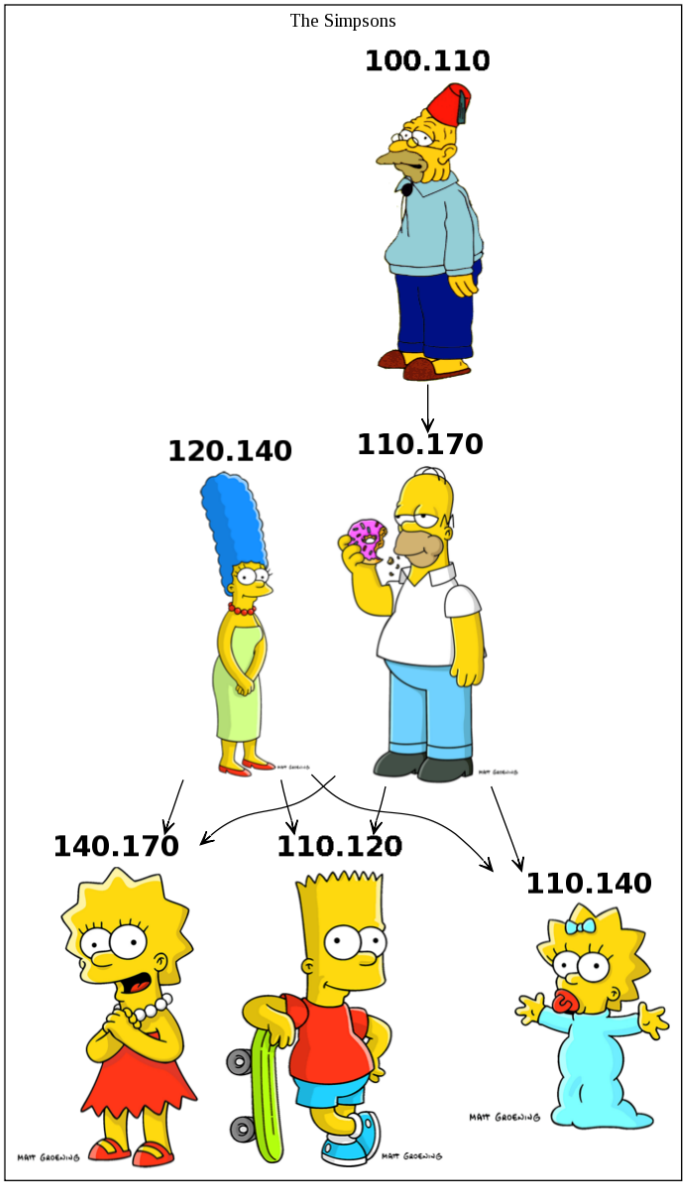
Bioinf Leipzig

February, 2009

A Pedigree...

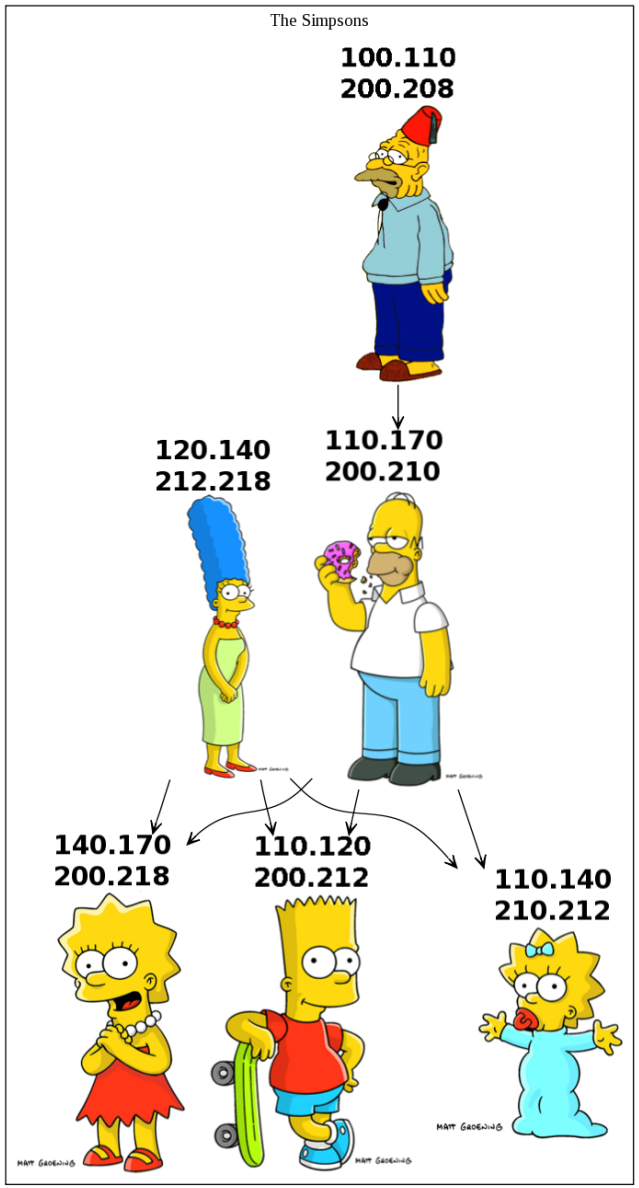


The Pedigree Reconstruction Problem



- Two **chromosomes**
- From **mother** and **father**
- **Locus**
- **Alleles**
- **Genotype** (**homozygous**, **heterozygous**)
- **Microsatellites**
  `CACACACACACACA` (7)

The Pedigree Reconstruction Problem



- Two **chromosomes**
- From **mother** and **father**
- **Locus**
- **Alleles**
- **Genotype** (**homozygous**, **heterozygous**)
- **Microsatellites**
  `CACACACACACACA` (7)

A Microsatellite Dataset

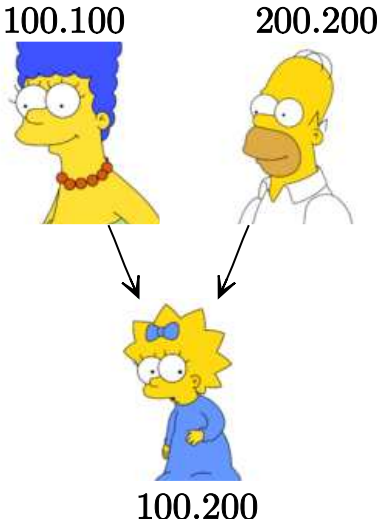
Loci 1-8							
147.157	182.188	210.212	234.240	159.163	168.170	178.180	207.217
151.151	182.188	212.216	234.234	159.159	168.168	178.178	207.237
137.151	182.182	208.214	234.234	159.163	168.168	178.180	207.207
151.151	?.?	216.218	234.240	159.159	168.168	178.180	?.?
147.157	182.182	216.218	234.240	159.159	168.168	178.178	?.?

EDEN (Ecological Diversity and Evolutionary Networks)

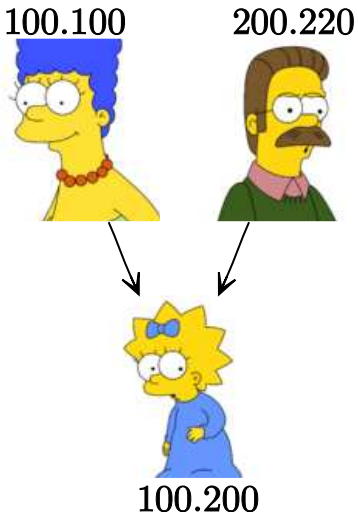
Goals:

- Estimate gene flow, rates of selfing and clonal reproduction, ...
- Investigate pollen flow

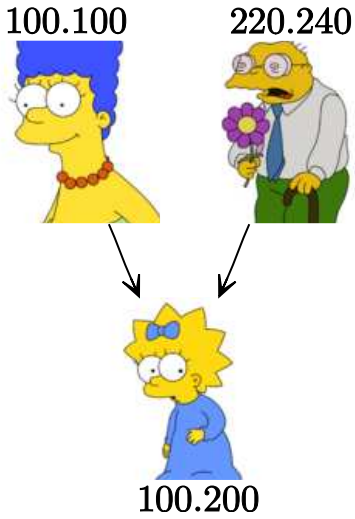
Parentage Probability $T(O|M, F_i)$



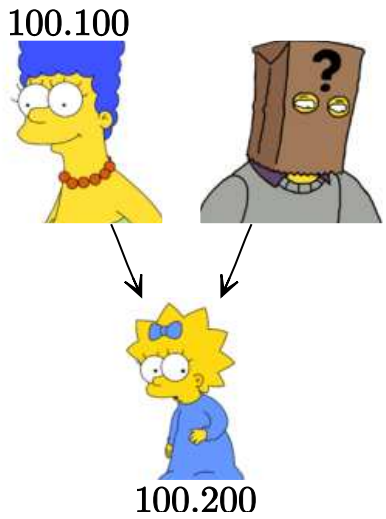
$P = 1 \cdot 1$



$P = 1 \cdot 0.5$



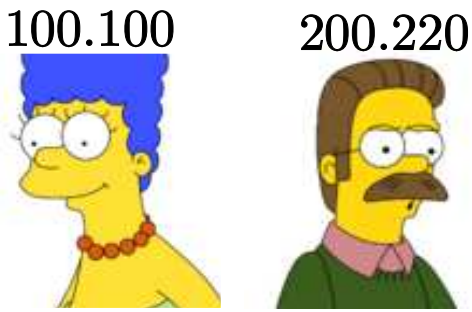
$P = 1 \cdot 0$



$P = 1 \cdot P(200)$

T. Meagher and E.A. Thompson, 1986. Neff et al., 2001. J.D. Hadfield et al., 2006

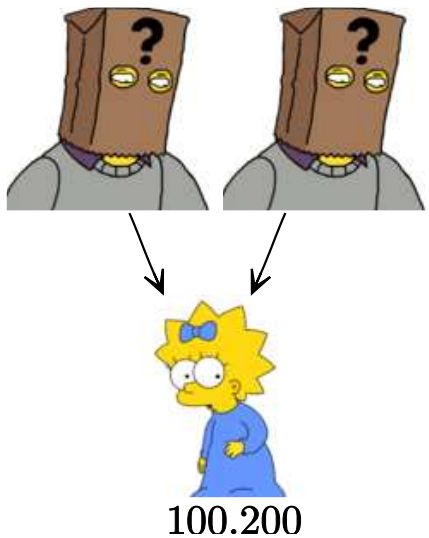
Genotype Probabilities $P(G_A|A)$



$$P(100.100|A) = P(100)^2$$

$$P(200.220|A) = 2 \cdot P(200) \cdot P(220)$$

Genotype Probabilities cont'd



Parentage *Posterior* Probability $P(O|M,F,A,N)$

$$P(\text{Lisa} \mid \text{Marge}, \text{Homer}, A, N) = \frac{\text{Homer}}{\text{Homer} + \text{Ned} + \text{Flowers} + (N-3)\ \text{?}}$$

A = Allele frequencies

N = Number of breeding individuals (sampled and unsampled)

R. Nielsen et al., *Statistical Approaches to Paternity Analyses in Natural Populations. . .*, 2001.

Pedigree Likelihood $L(\mathscr{P}|D, A, N, \vec{\epsilon})$

$$\log P(O_i|M_i, F_i, A, N, \vec{\epsilon}) = \sum_{j}^{N_L} \log T(O_{ij}|M_{ij}, F_{ij}, A_j, N, \epsilon_j)$$

$$\log L(\mathscr{P}|D, A, N, \vec{\epsilon}) = \sum_{i}^{N_I} \log P(O_i|M_i, F_i, A, N, \vec{\epsilon})$$

FRANz: The Pedigree Reconstruction Steps

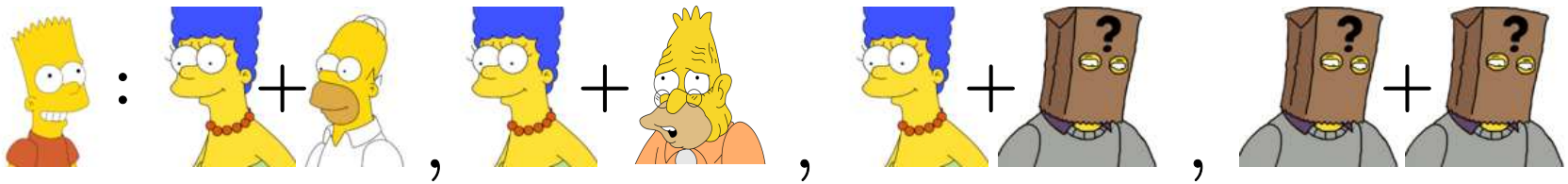
1. Search Space Reduction
2. MCMC Sampling
3. MCMC path analysis

Step 1: Reduce Search Space

- Use prior knowledge: **known relationships** (mother-offspring, **siblings**), **age**, sex, sampling locations to reduce the number of candidates
- More likely than randomly drawn from the population?

$$\frac{T(O|M,F,\vec{\epsilon})}{P(O|A)} > 1$$

- Denote the possible parent combinations of individual i as $\mathscr{H}(i)$



Step 2: MCMC

$$\pi(\mathscr{P}, N|D, A, \vec{\epsilon}) = \frac{f(\mathscr{P}, N)L(\mathscr{P}, N|D, A, \vec{\epsilon})}{\sum_i \sum_j f(\mathscr{P}_i, N_j)L(\mathscr{P}_i, N_j|D, A, \vec{\epsilon})}$$

Step 2: MCMC cont'd

- Change Pedigree $\mathscr{P} \Rightarrow \mathscr{P}'$
- (or $N \Rightarrow N'$)
- Calculate probability of new candidate
- The corresponding acceptance function:

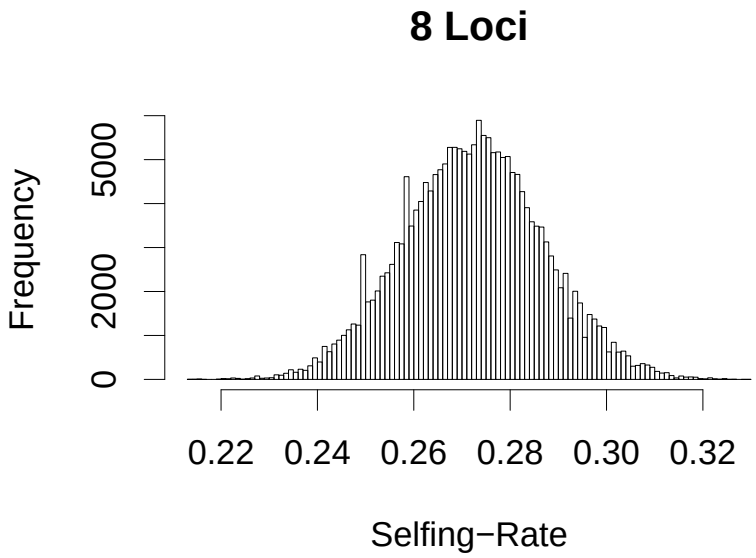
$$\alpha(\mathscr{P}', N' | \mathscr{P}, N) = \begin{cases} 0 & \text{if change introduced a directed cycle in } \mathscr{P} \\ \min & \left[1, e^{[LL(\mathscr{P}', N') - LL(\mathscr{P}, N)]/T}\right] \end{cases}$$

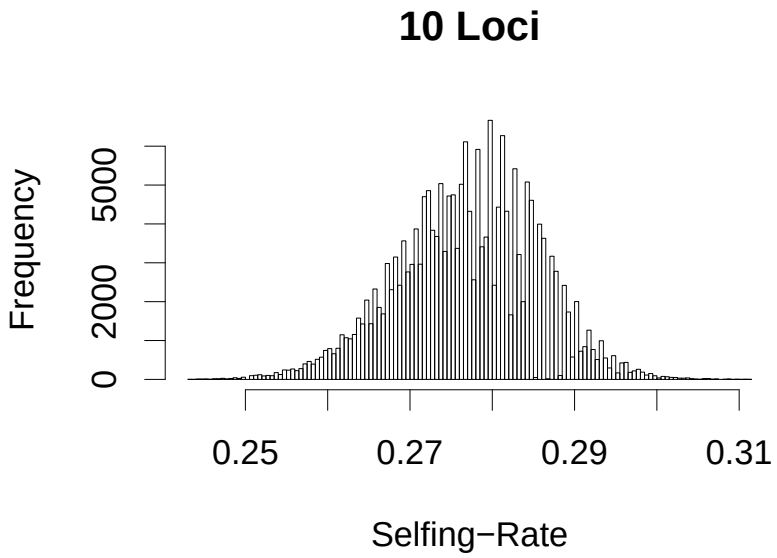
Step 3: Analyze MCMC path

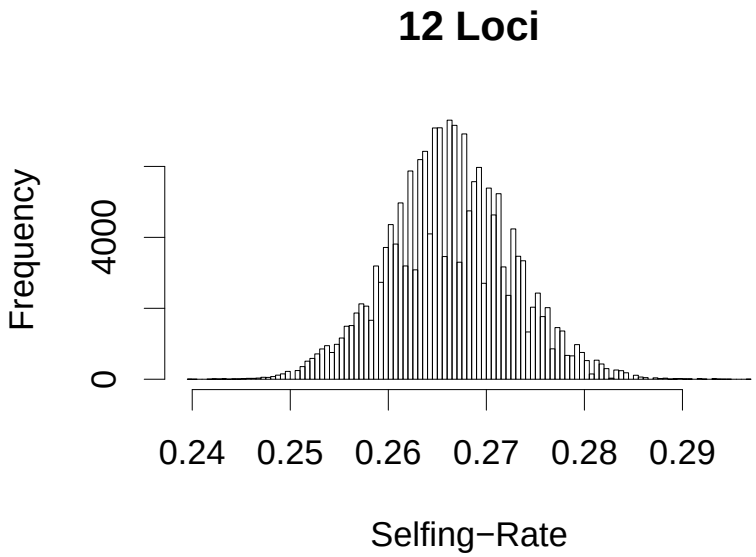
Example: Estimation of the selfing rate of a simulated population

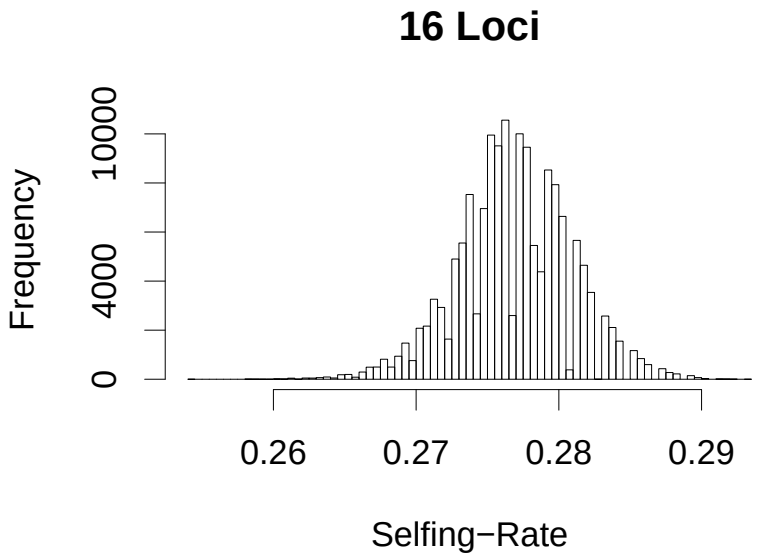
- 2000 of 10000 individuals
- Typing error of 1%
- Selfing-rate of population 0.260, of sample 0.277
- Prior knowledge:
  - Upper bound number of unsampled genotypes $N_{max} = 20000$
  - Typing error rate 1%.

Step 3: Analyze MCMC path cont'd







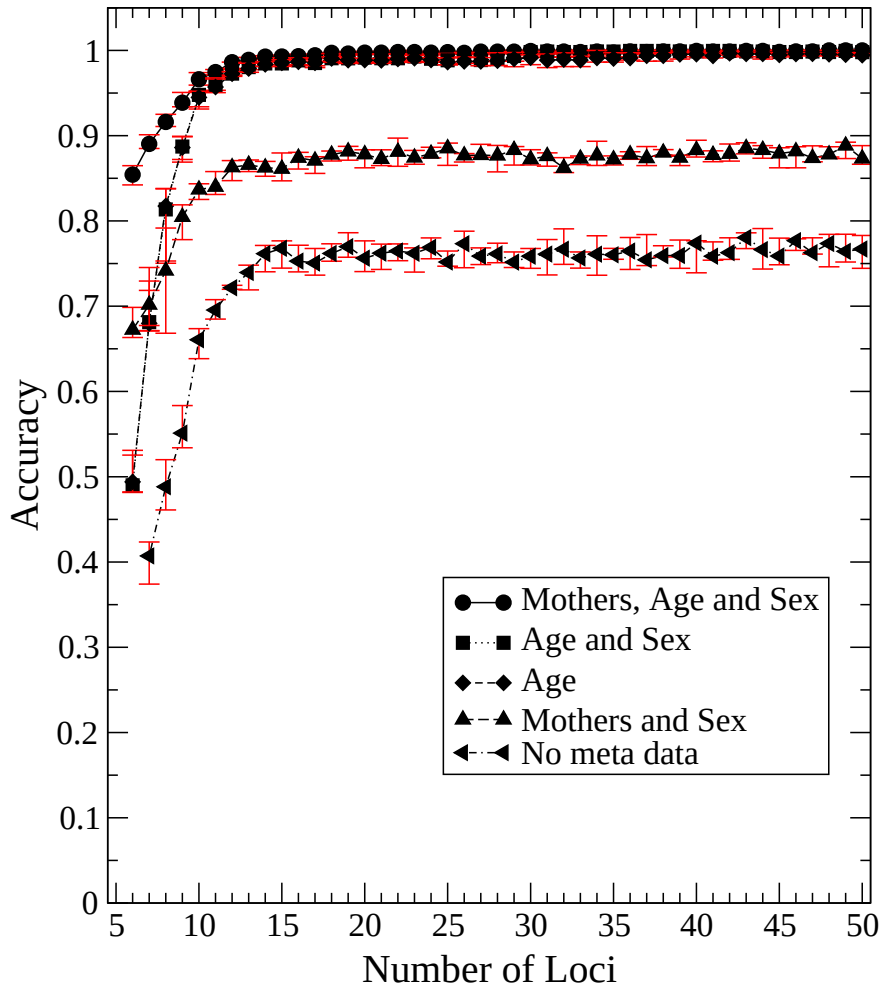


Prior Knowledge

- Simulated populations based on statistics of the **German population**
- 1000 of 2000 Individuals
- **Deep Pedigrees: 5 to 9 generations**
- Prior knowledge:
  - Upper bound number of unsampled genotypes $N_{max} = 10000$
  - Typing error rate 1%.
- Influence of other **prior knowledge**?
  - Age
  - Known Mothers
  - Sex

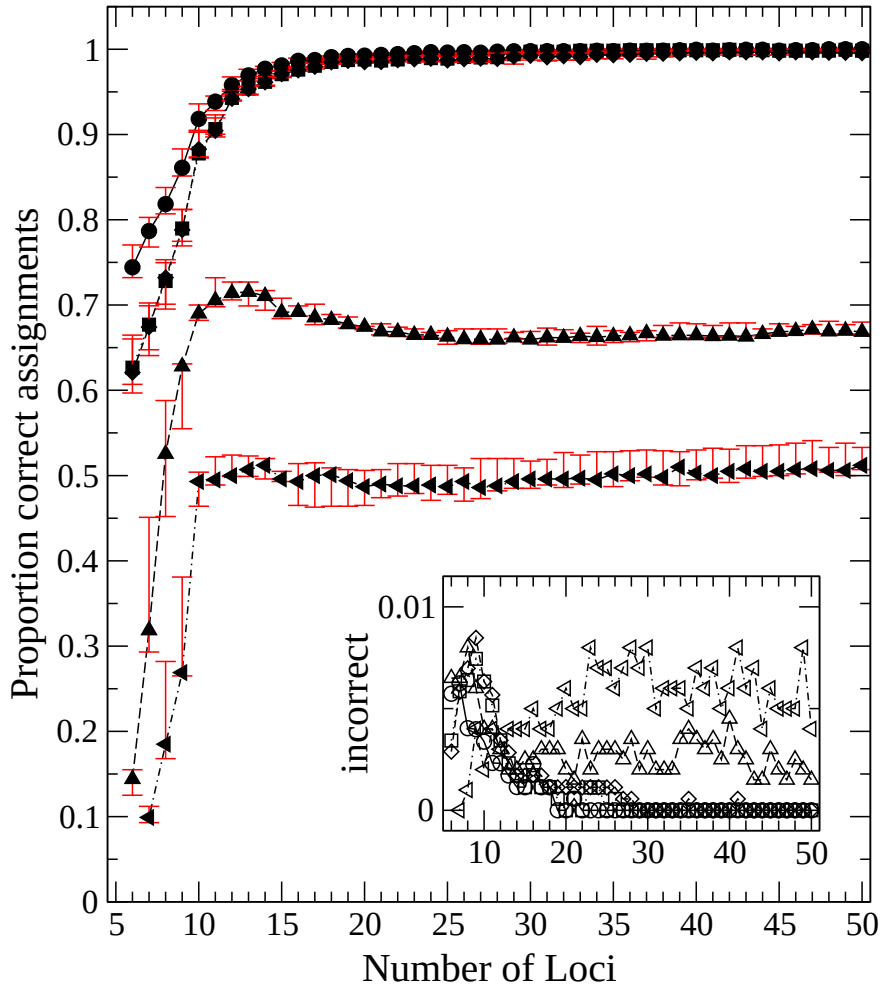
Simulated Data

Accuracy of the Maximum Likelihood Pedigree (Simulated Annealing)



Simulated Data cont'd

If we look at parentages with a posterior probability > 0.95 (MCMC), are they correct?



FRANz

`http://www.bioinf.uni-leipzig.de/Software/FRANz`

- **Flexible:** Can be used like classical parentage inference programs
- Very comprehensive marker suite analyses
- **Simulated Annealing** for finding the **Maximum Likelihood Pedigree**
- Written in ANSI C, completely multi-threaded
- GPL
- FRANz: Reconstruction of wild multi-generation pedigrees. Markus Riester, Peter F. Stadler and Konstantin Klemm. Bioinformatics 2009 (in press).

Acknowledgments

`http://www.bioinf.uni-leipzig.de/Software/FRANz`

Thank you!

Universität Leipzig

- Konstantin Klemm
- Peter F. Stadler
- Camille Stephan-Otto Attolini

IFISC/IMEDEA-UIB, Palma de Mallorca

- Emilio Hernández-García

CCMAR Faro

- Sophie Arnaud-Haond
- Gareth Pearson
- Ester Serrão