Kinship testing with with (many) more markers

Maarten Kruijver
m.v.kruijver@vu.nl

~~VU University, Amsterdam, Netherlands~~

September 2, 2016

Outline

Finding a relative among database profiles

POLICYFORUM

HUMAN GENETICS

Finding Criminals Through DNA of Their Relatives

Analyses of the DNA databases maintained by criminal justice systems might enable criminals to be caught by recognizing their kin, but this raises civil liberties issues.

Frederick R. Bieber,[1]* Charles H. Brenner,[2] David Lazer[3]

DNA methods are now widely used for many forensic purposes, including routine investigation of serious crimes and for identification of persons killed in mass disasters or wars (1–4). DNA databases of convicted offenders are maintained by every U.S. state and nearly every industrialized country, allowing comparison of crime scene DNA profiles to one another and to known offenders (5). The policy in the United Kingdom stipulates that almost any collision with law enforcement results in the collection of DNA (6). Following the U.K. lead, the United States has shifted steadily toward inclusion of all felons, and federal and six U.S. state laws now include some provision for those arrested or indicted. At present, there are over 3 million samples in the U.S. offender/arrestee state and federal DNA databases (7). Statutes governing the use of such samples and protection against misuse vary from state to state (8).

Although direct comparisons of DNA profiles of known individuals and unknown biological evidence are most common, indirect genetic kinship analyses, using the DNA of biological relatives, are often necessary for humanitarian mass disaster and missing person identifications (1, 2, 9). Such methods could potentially be applied to searches of the convicted offender/arrestee DNA databases. When crime scene samples do not match anyone in a search of forensic databases, the application of indirect methods could identify individuals in the database who are close relatives of the potential suspects. This raises compelling policy questions about the balance between collective security and individual privacy (10).

To date, searching DNA databases to identify close relatives of potential suspects has been used in only a small number of cases, if some-

[1]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. [2]DNAVIEW, Oakland, CA 94611, and School of Public Health, University of California, Berkeley, CA 94720, USA. [3]John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138, USA.

*Authors are alphabetized to reflect equal contributions. Comments and ideas expressed herein are their own.

*Author for correspondence. E-mail: fbieber@partners.org

times to dramatic effect. For example, the brutal 1988 murder of 16-year-old Lynette White, in Cardiff, Wales, was finally solved in 2003. A search of the U.K. National DNA Database for individuals with a specific single rare allele found in the crime scene evidence that identified a 14-year-old boy with a similar overall DNA profile. This led police to his paternal uncle, Jeffrey Gafoor (11). Investigation of the 1984 murder of Deborah Sykes revealed a close, but not perfect, match to a man in the North Carolina DNA offender database, which led investigators to his brother, Wilford Brown (12). Both Gafoor and Brown matched the DNA from the respective crime scenes, confessed, and were convicted.

Although all individuals have some genetic similarity, close relatives have very similar DNA profiles because of shared ancestry. We demonstrate the potential value of kinship analysis for identifying promising leads in forensic investigations on a much wider scale than has been used to date.

Let us assume that a sample from a crime scene has been obtained that is not an exact match to the profile of anyone in current DNA databases. Using Monte Carlo simulations (13, 14), we investigated the chances of successfully identifying a biological relative of someone whose profile is in the DNA database as a possible source of crime scene evidence (15). Each Monte Carlo trial simulates a database of known offenders, a sample found at a crime scene, and a search. The search compares the crime sample with each catalogued offender in turn by computing likelihood ratios (LRs) that assess the likelihood of parent-child or of sibling relationships (1, 16). We used published data on allele frequencies of the 13 short tandem repeat (STR) loci on which U.S. offender databases are based and basic genetic principles (17–19). A high LR is characteristic of related individuals and is an unusual but possible coincidence for unrelated individuals. The analysis of each simulation therefore assumes that investigators would follow those leads in priority order, starting with those in the offender database with the highest LR for being closely related to the owner of the crime scene DNA sample.

Our simulations demonstrate that kinship analysis would be valuable now for detecting potential suspects who are the parents, children, or siblings of those whose profiles are in forensic databases. For example, assume that the unknown sample is from the biological child of one of the 50,000 offenders in a typical-sized state database. Of the 50,000 LRs comparing the "unknown" sample to each registered offender in the database, the child corresponds to the largest LR about half the time, and has a 99% chance of appearing among the 100 largest LRs (see chart). An analysis of potential sibling relationships produced a similar curve (13).

These results could be refined by additional data—for example, large numbers of single-nucleotide polymorphisms (SNPs). Better and immediately practical, a seven-locus Y-STR haplotype analysis on the crime scene and the list of database leads would eliminate 99% of those not related by male lineage (20). Data-mining (vital records, genealogical and geographical data) for the existence of suitable suspects related to the leads can also help to refine the list.

The potential for improving effectiveness of DNA database searches is large. Consider a hypothetical state in which the "cold-hit" rate—the chance of finding a match between a crime scene sample and someone in the offender database—is 10%. Suppose that

Figure: Bieber et al. (2006) [1]

- Offender's profile is available, but yields no match in the database. Maybe a close relative is in the database?
- Procedure: compute LR in favor of a full sibling or parent/offspring relationship for all database members and further investigate a candidate list of 'large' LRs (say $> 1,000$)
- Exclude false leads by additional (genetic) research (mostly Y-STRs)
- Currently: limited power and many false leads to eliminate

Two strategies for selecting a candidate list

A search strategy strikes a balance between the power to detect a true relative and the workload (number of false leads to be eliminated)

Top-k

- Investigate a candidate list of fixed length
- In California, a list of 168 candidates is further investigated [2]

Fixed threshold

- Investigate LRs exceeding a threshold, e.g. Netherlands Forensic Institute
- A fixed threshold is optimal in the long run [3]

LR distributions: full siblings and unrelated

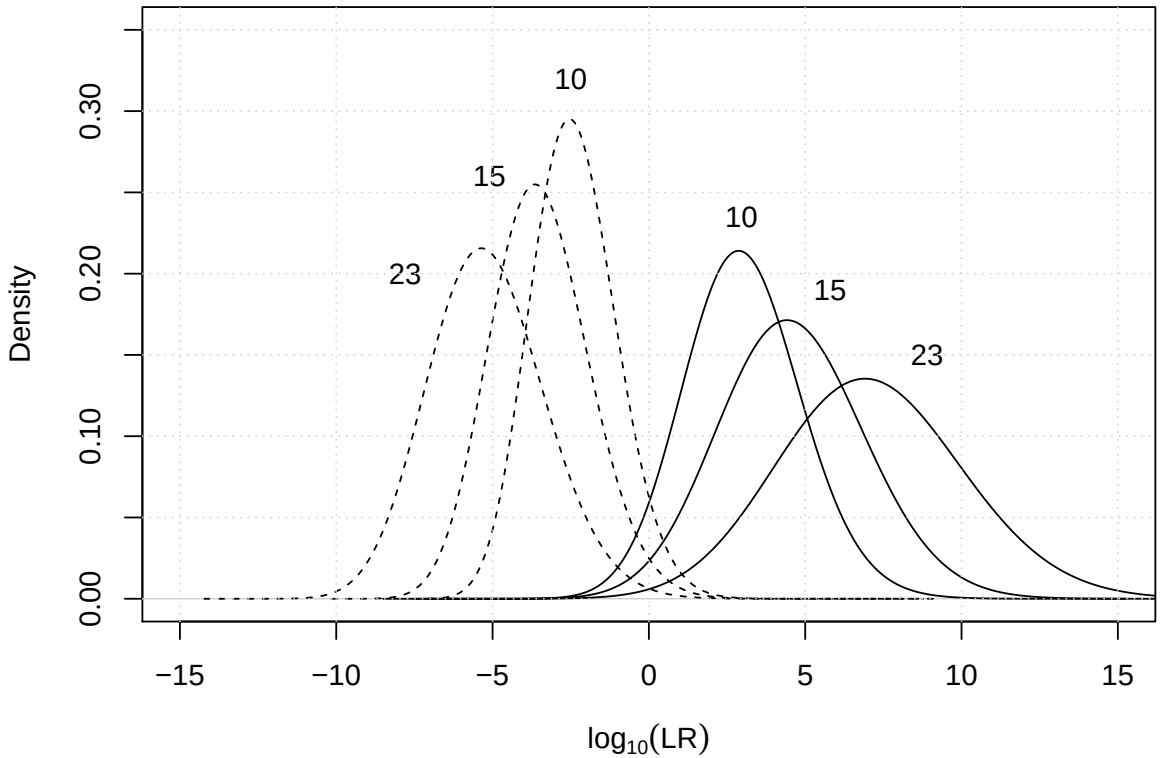


Figure: LR distributions for true full siblings (solid curves) and unrelated persons (dashed curves) for 10, 15 and 23 loci

ROC curve: full siblings and unrelated

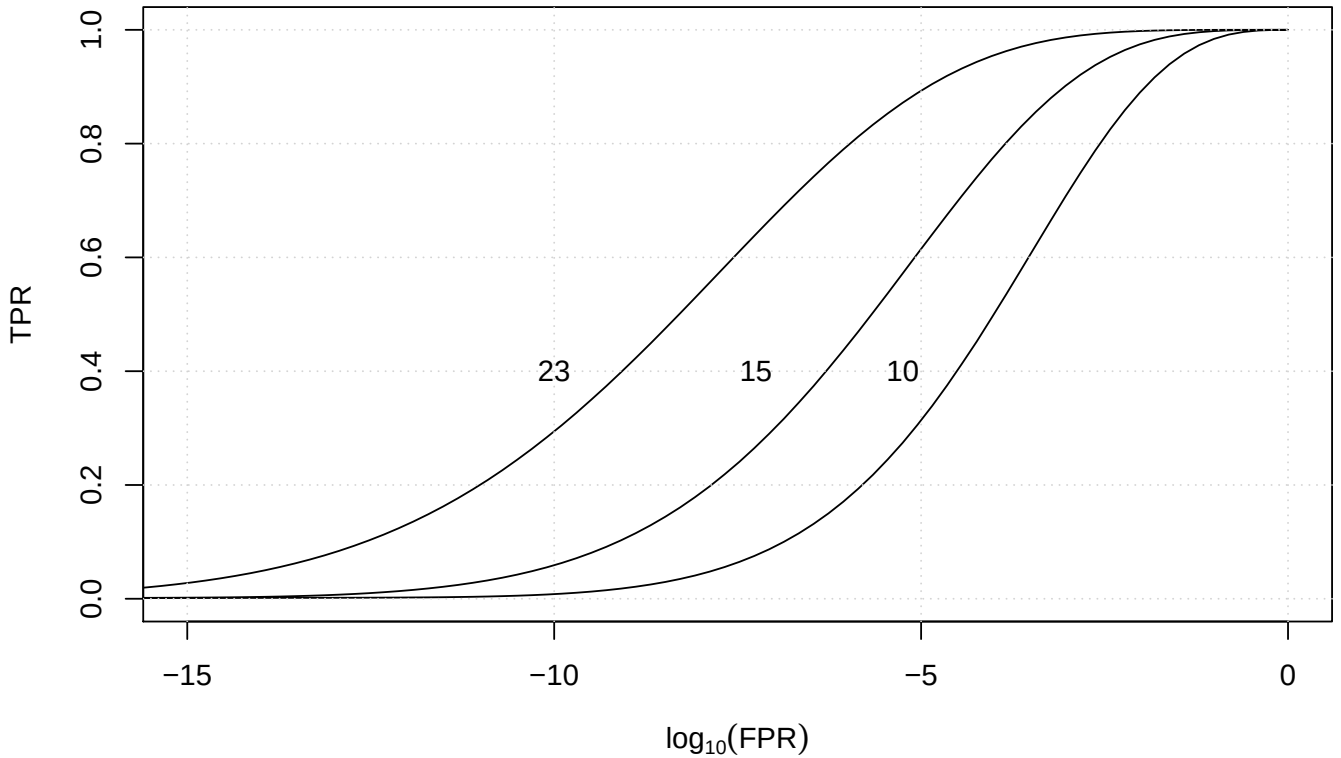


Figure: ROC curve: TPR (exceedance probability for a true relative) versus FPR (exceedance probability for an unrelated profile)

LR distributions: parents/offspring and unrelated

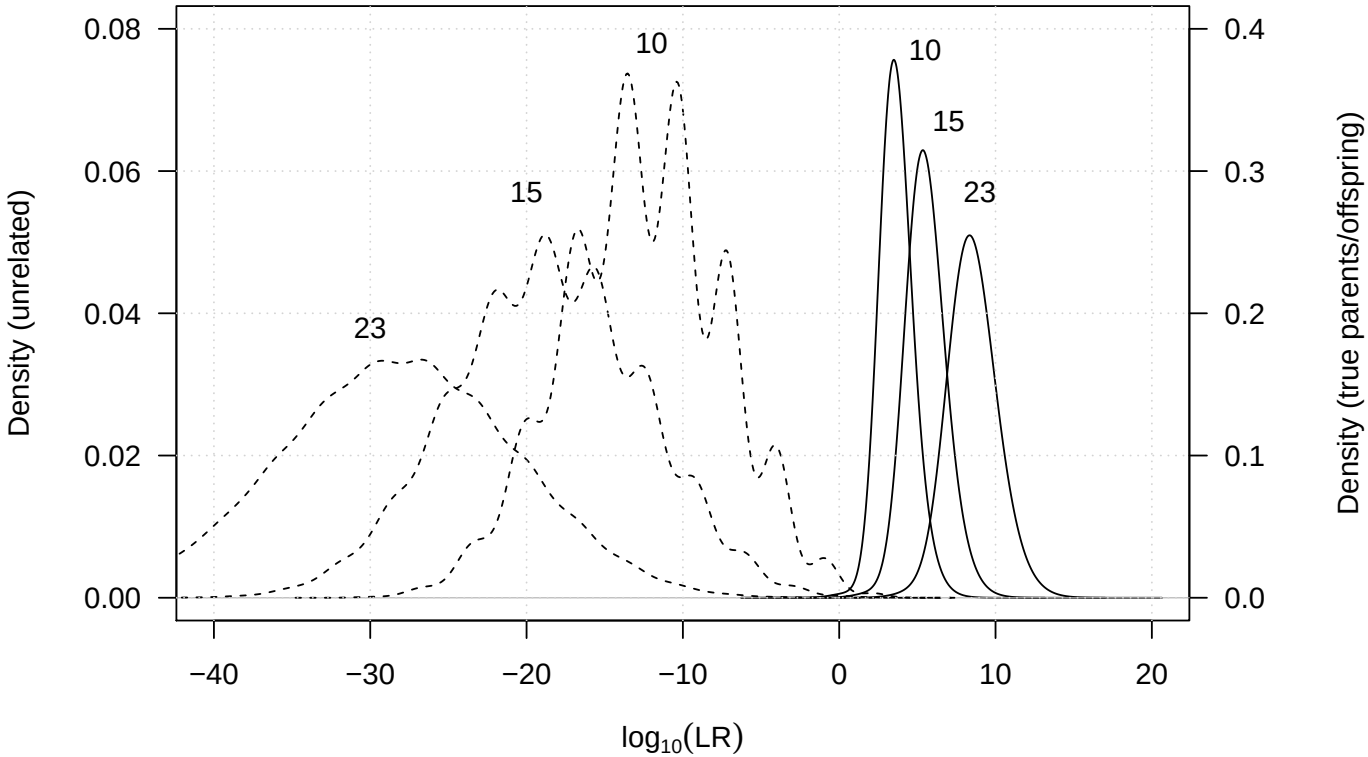


Figure: LR distributions for true parents/offspring (solid curves) and unrelated persons (dashed curves) for 10, 15 and 23 loci

ROC curve: parents/offspring and unrelated

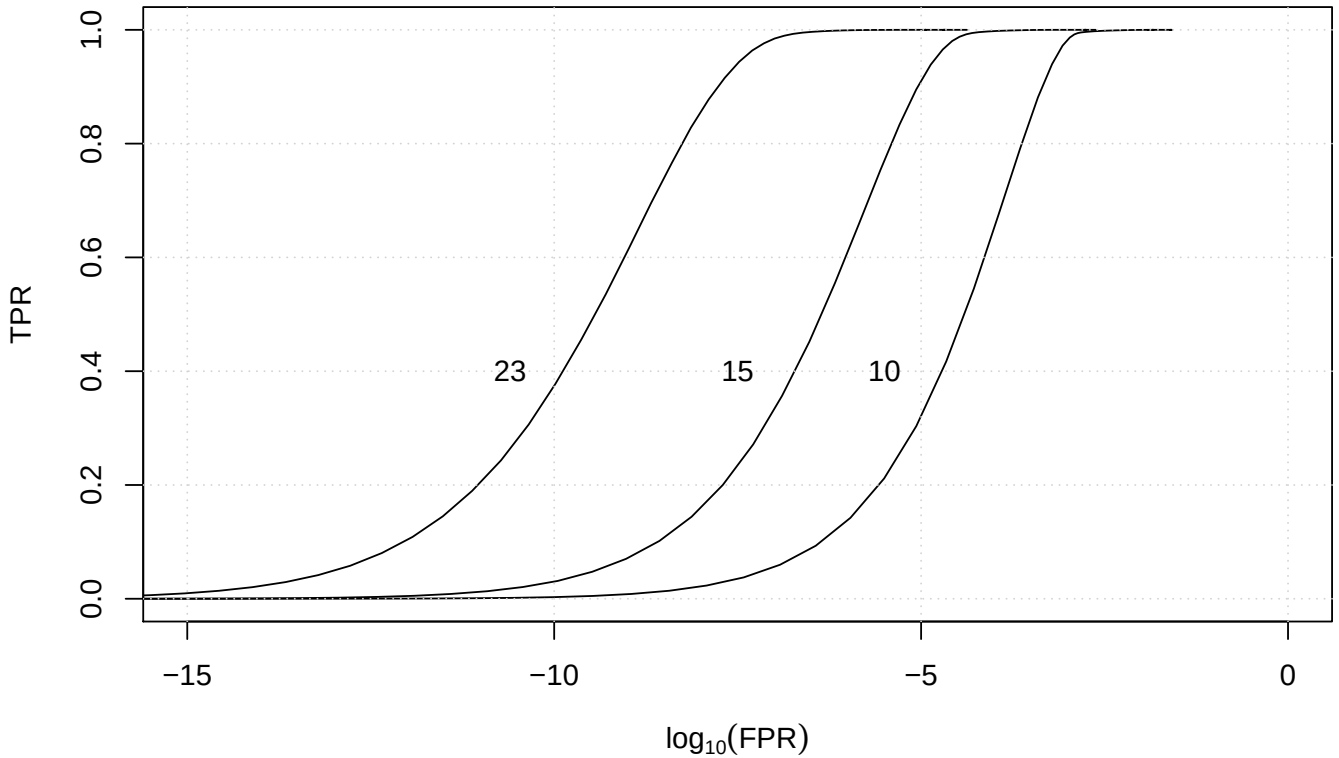


Figure: ROC curve: TPR (exceedance probability for a true relative) versus FPR (exceedance probability for an unrelated profile)

Relative identification (23 loci) versus direct matching

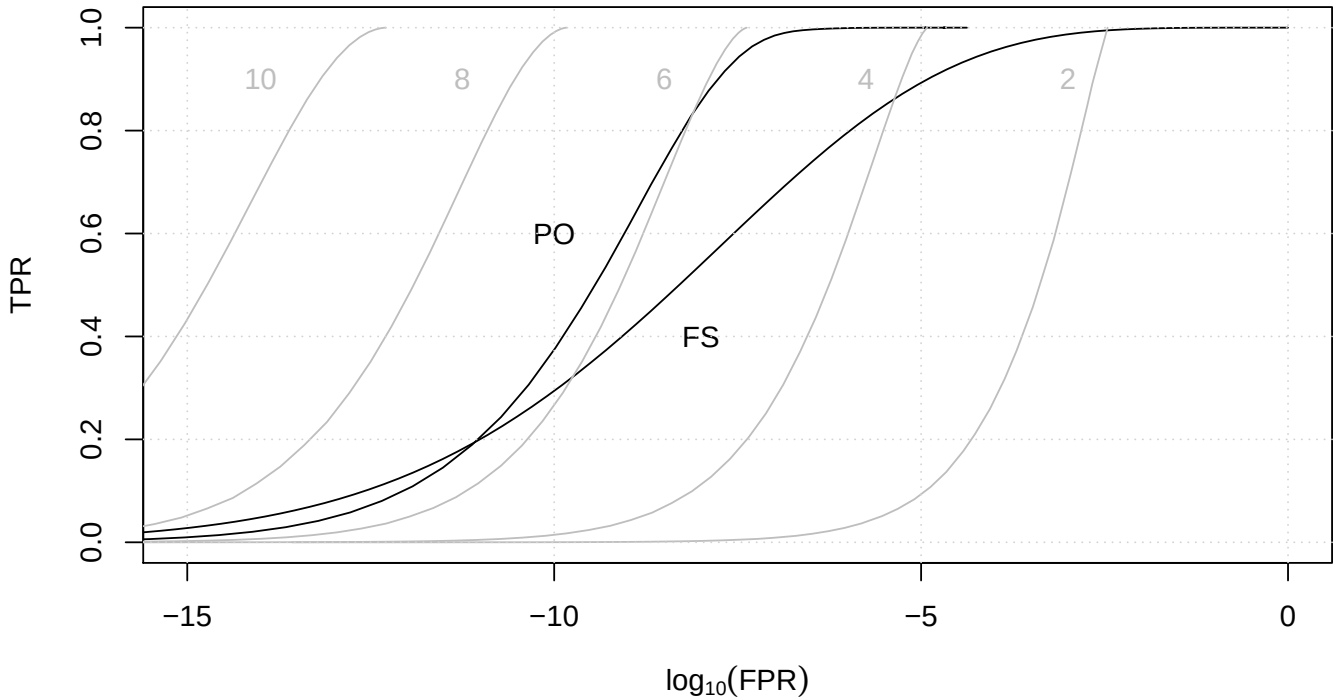


Figure: ROC curves for discrimination between first degree relatives and unrelated pairs using 23 loci (black curves) compared to those for direct identification using 2, 4, 6, 8, 10 ind. copies of D8S1179 (gray curves)

Top-k probabilities for full sibs in a 1M database

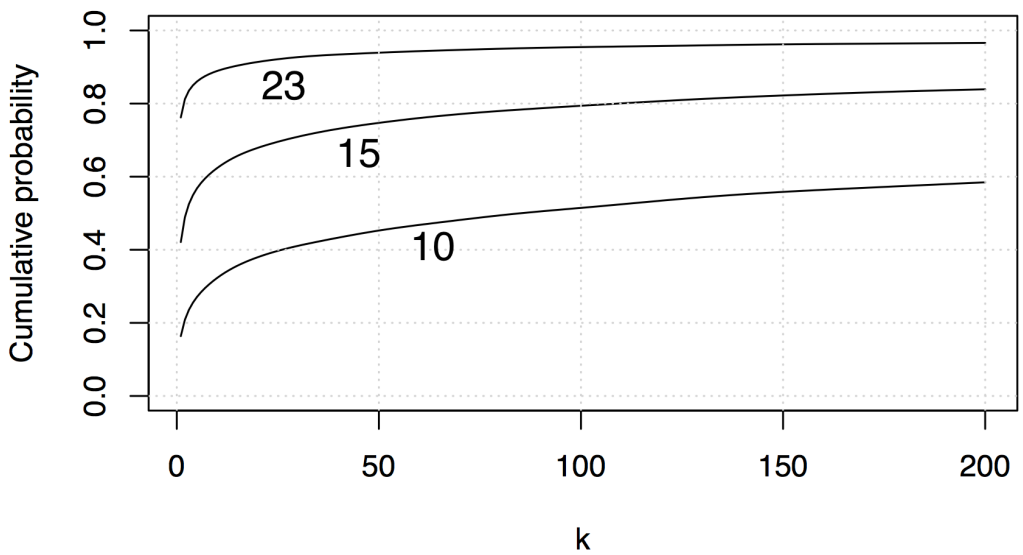


Figure: Probability for a true full sibling to appear in the top-k for a database of one million unrelated profiles for the three multiplexes

Top-k probabilities for parents/offspring in a 1M database

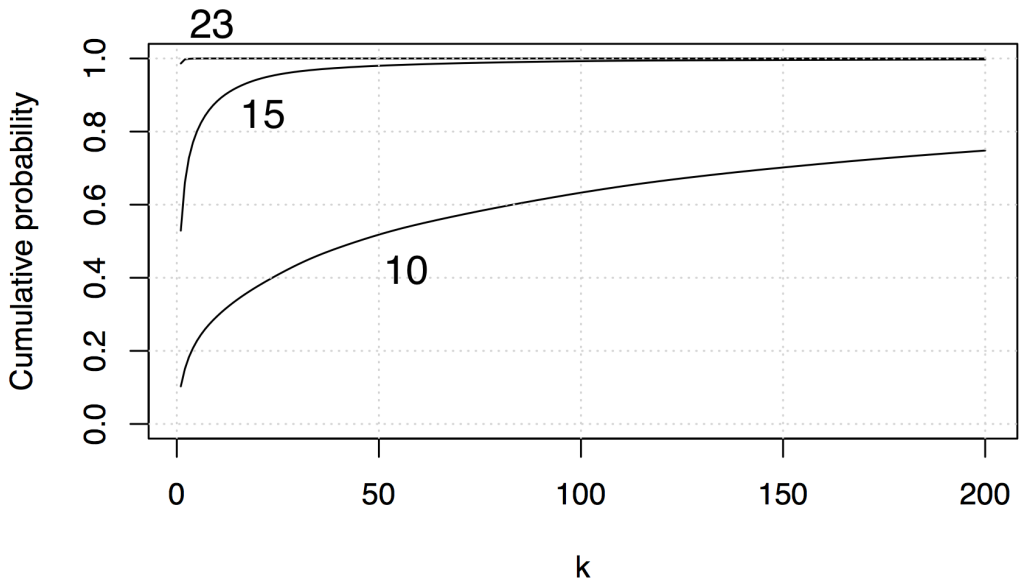


Figure: Probability for a true parent/offspring to appear in the top-k for a database of one million unrelated profiles for the three multiplexes

Top-k probabilities for full siblings

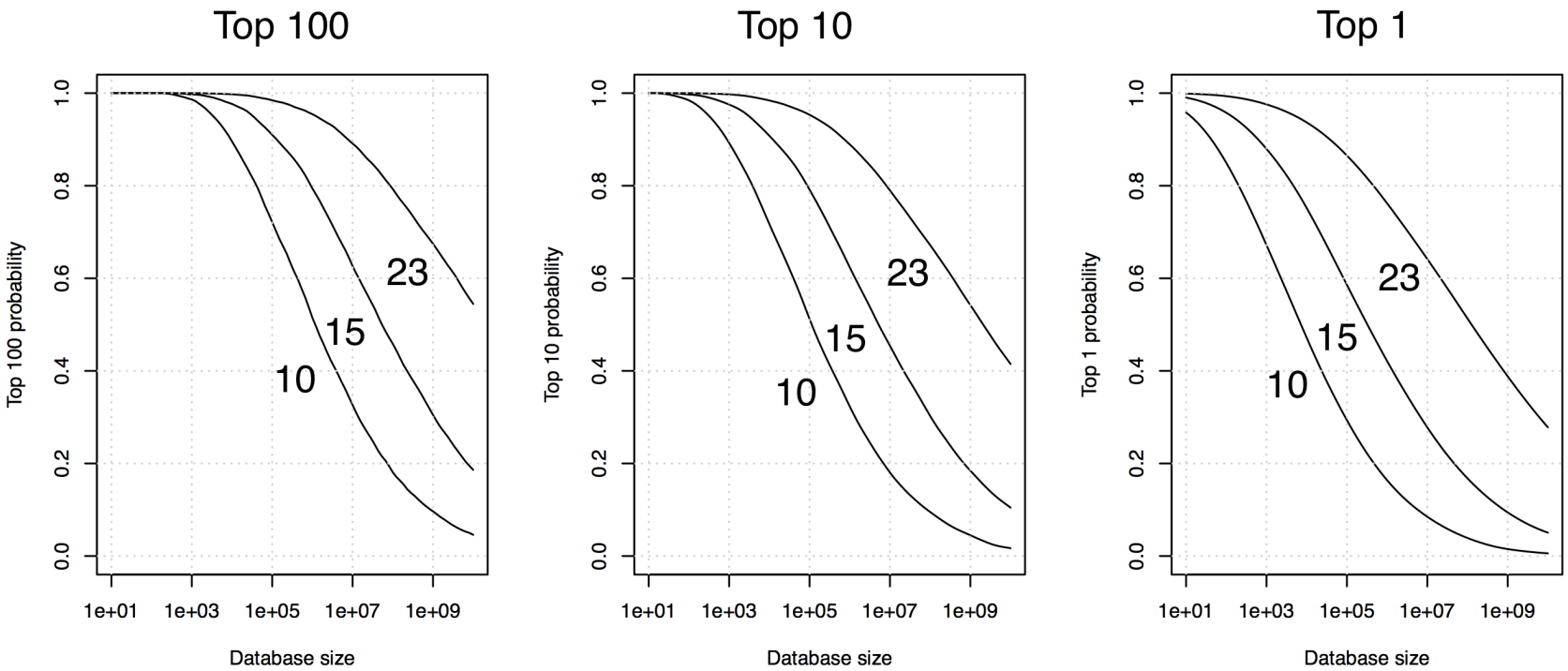


Figure: Probability to find a true relative in the top-k for databases of different sizes

Top-k probabilities for parents/offspring

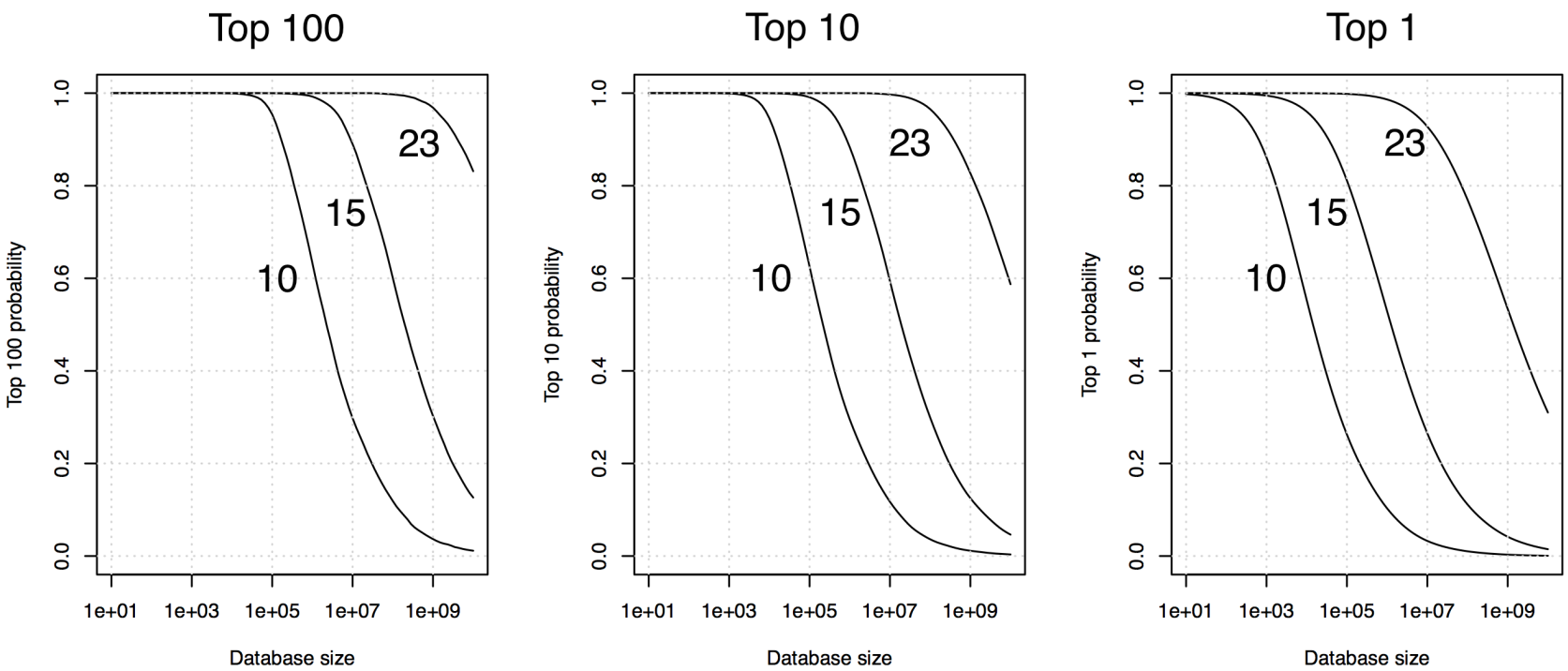


Figure: Probability to find a true relative in the top-k for databases of different sizes

LR $> 1,000$

		True relationship			
LR	# loci	PO	FS	HS	UN
PO vs UN	10	**7.48e-01**	2.57e-01	3.57e-02	1.73e-04
PO vs UN	15	**9.85e-01**	2.88e-01	2.50e-02	3.10e-05
PO vs UN	23	**9.99e-01**	3.40e-01	1.71e-02	1.11e-06
FS vs UN	10	3.37e-01	**4.97e-01**	2.97e-02	8.95e-05
FS vs UN	15	7.22e-01	**7.56e-01**	6.40e-02	6.71e-05
FS vs UN	23	9.69e-01	**9.29e-01**	1.28e-01	2.74e-05

Table: Exceedance probabilities for LR $> 1,000$ for different choices of H_p, different true relationships and different number of loci

LR $> 1,000,000$

		True relationship		
LR	# loci	PO	FS	UN
PO vs UN	10	**2.03e-02**	1.37e-02	8.47e-09
PO vs UN	15	**3.29e-01**	9.35e-02	9.81e-08
PO vs UN	23	**9.59e-01**	1.65e-01	4.44e-08
FS vs UN	10	8.06e-03	**5.71e-02**	1.79e-08
FS vs UN	15	8.85e-02	**2.71e-01**	5.22e-08
FS vs UN	23	5.02e-01	**6.51e-01**	5.73e-08

Table: Exceedance probabilities for LR $> 1,000,000$ for different choices of H_p, different true relationships and different number of loci

How many independent loci would be needed?

- A familial search to be feasible if there is a reasonable probability (say 50%) to detect a true relative (if present) at a reasonably small FPR (say 10^{-4})
- We investigate how many independent loci are needed for a familial search to be feasible for common pairwise relationships (FS, PO, HS, FC, SC)
- Procedure: find for each number of loci the LR threshold such that FPR $= 10^{-4}$ and compute the corresponding TPR

How many independent loci would be needed?

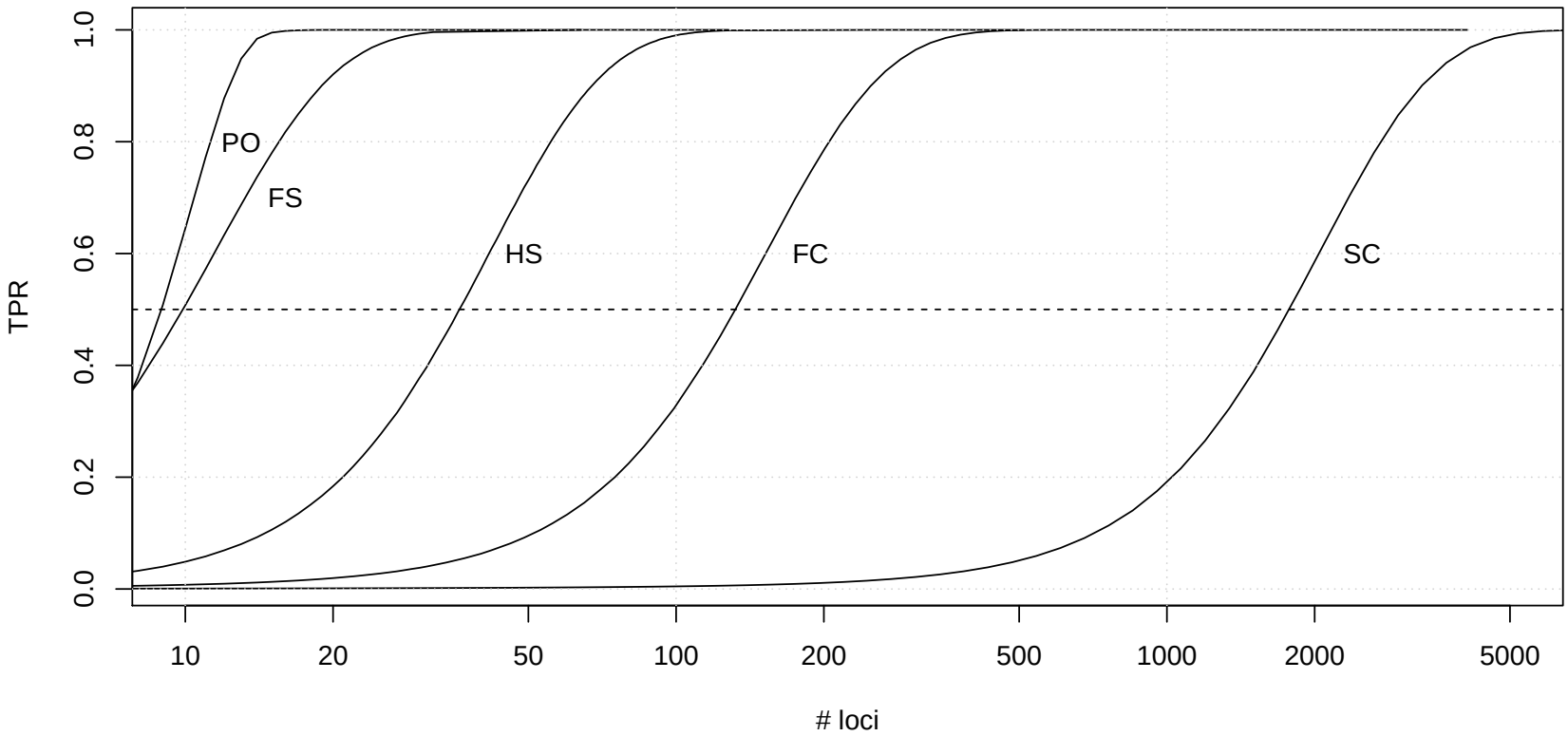
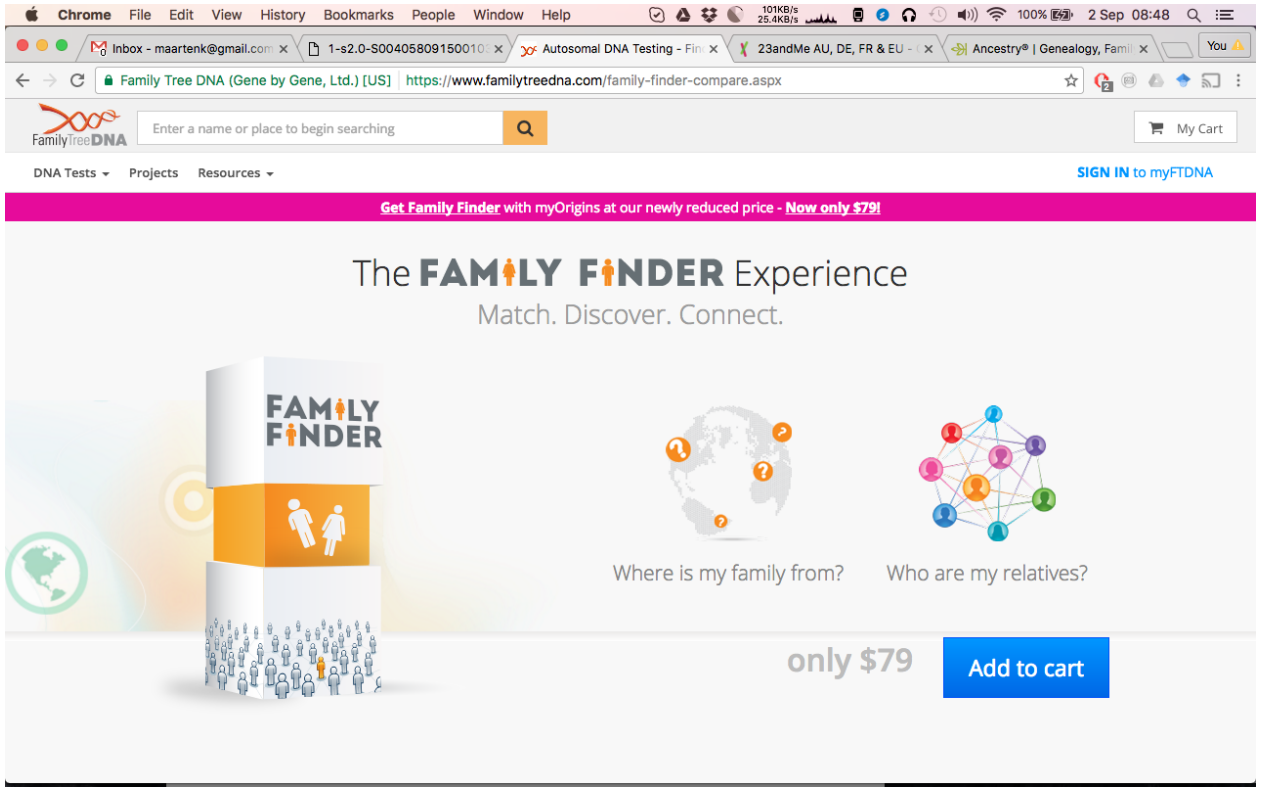


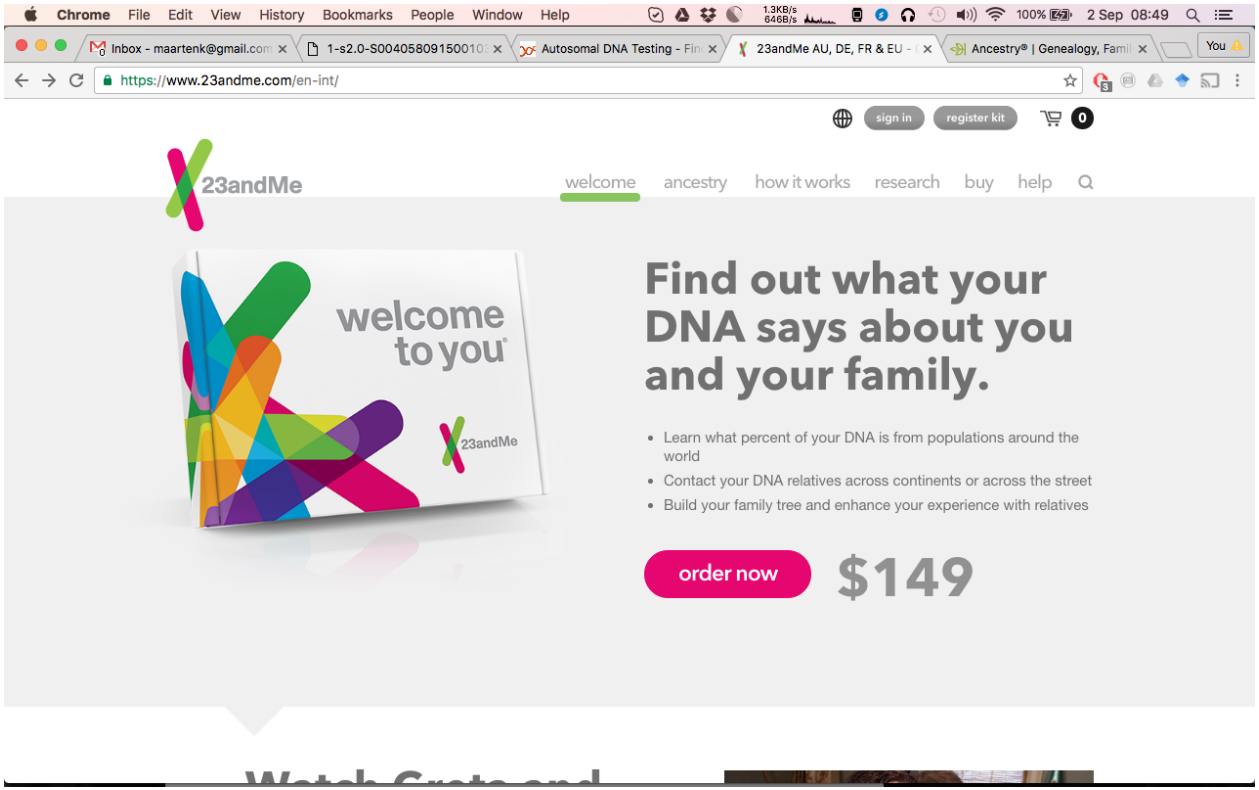
Figure: TPR for threshold such that FPR $= 10^{-4}$

Who's on top?



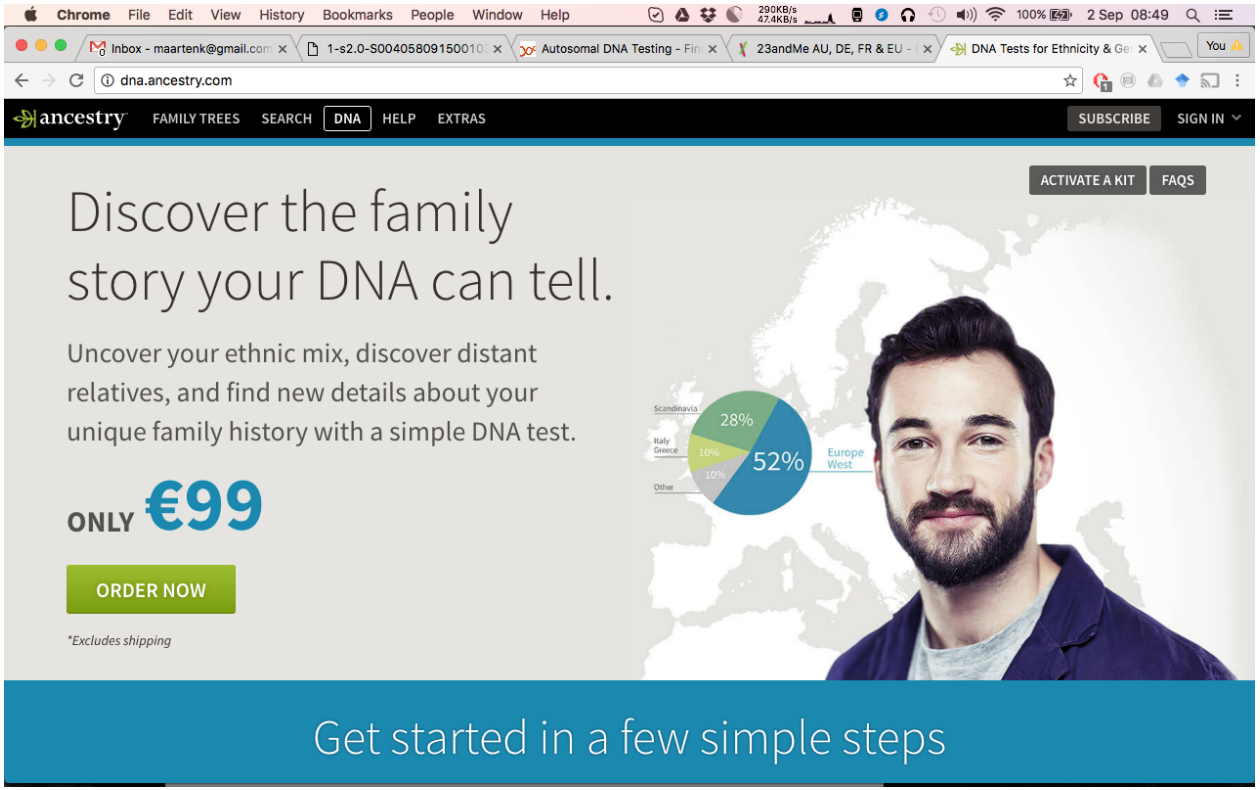
Figure

Who's on top?



Figure

Who's on top?



Figure

Conclusions

- Since DNA databases profiles comprise just 10-15 loci, familial searches have limited power, are only suited for finding first degree relatives and it requires significant effort to eliminate false leads
- With 23 loci, finding first degree relatives is possible without the need to eliminate many false leads
- Familial searches for full siblings will become feasible in larger databases (e.g. United States NDIS)
- A parent/offspring search will be practically match/no match
- Searching second degree relatives, on the other hand, remains problematic, even if many more markers would be included

References I


F. Bieber, C. Brenner, D. Lazer, Finding Criminals Through DNA of Their Relatives, Science 312 (5778) (2006) 1315. doi:10.1126/science.1122655.


S. Myers, M. Timken, M. Piucci, G. Sims, M. Greenwald, J. Weigand, K. Konzak, M. Buoncristiani, Searching for first-degree familial relationships in California's offender DNA database: Validation of a likelihood ratio-based approach, Forensic Science International: Genetics 5 (5) (2011) 493–500.

doi:10.1016/j.fsigen.2010.10.010.

References II

M. Kruijver, R. Meester, K. Slooten, Optimal strategies for familial searching, Forensic Science International: Genetics 13 (2014) 90–103.
doi:10.1016/j.fsigen.2014.06.010.