Identifying Personal Genomes by Surname Inference

Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y.

Science. 2013 Jan 18;339(6117):321-4. doi: 10.1126/science.1229566.

Journal Club

Kairi Raime
04.02.2013

Introduction

- **Surnames** are paternally inherited -> cosegregation with **Y-chromosome haplotypes**
- **Short tandem repeats** across Y chromosome (**Y-STRs**)
- Multiple **genetic genealogy companies**:

  (surname + Y-STR haplotype records)

  - **Ysearch** website (FamilyTreeDNA)
  - **SMGF** website (Sorenson Molecular Genealogy Foundation)
  - …

Introduction

- **Lunshof et al. (2008):**
  - Genetic genealogy databases: Y-chromosome haplotypes -> surname
  - + demographic information -> speculation that full identification of participants in sequencing projects is possible
- **Gitschier (2009):**
  - empirically approached this hypothesis by testing 30 Y-STR haplotypes of CEU participants in genetic genealogy databases
  - Results: potential surnames can be detected, but surnames could match thousands of individuals

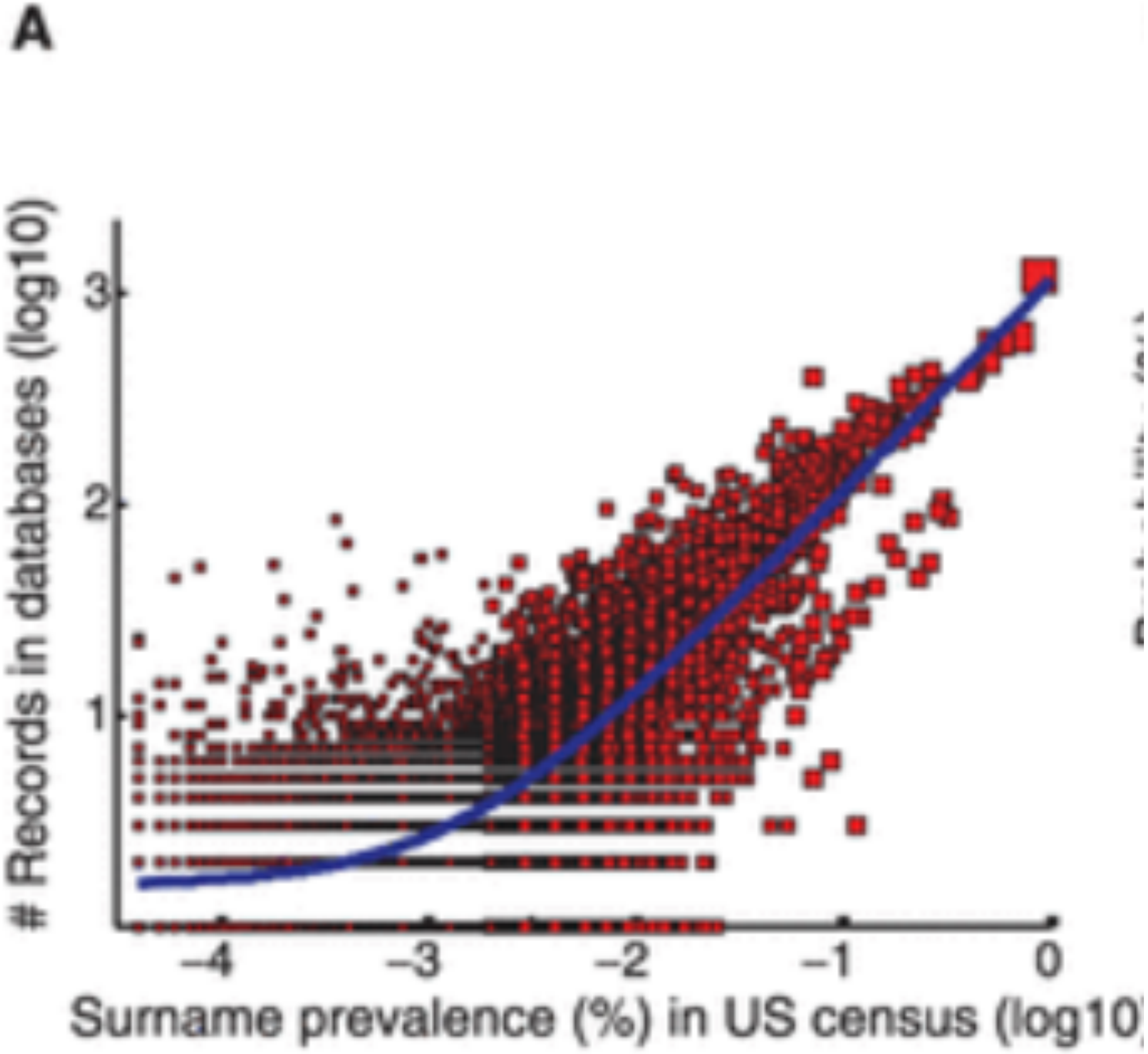
The goal of this research

- To show how readily surname inference might be possible in more general population (quantitative assessment)
- To show that full identities of personal genomes can be exposed via surname inference from recreational genetic genealogy databases followed by internet searches
- To demonstrate end-to-end identification of individuals with only public information

Public genetic genealogy databases:

- **Ysearch** (www.ysearch.org) and **SMGF (www.smgf.org) –** two largest databases
- **Free-of-charge**
- **Built-in search engines**
- **Search:** insert a combination of Y-STR alleles -> matching records on the bases of genetic similarity ->
- **retrieved records:** surnames, information about patrilineal line (geographical locations, potential spelling variants) and pedigrees
- Contain **~39 000** unique surname entries from **~135 000** records
- The distribution of records per surname is significally correlated ($R^2 = 0.78$, $p < 1.20 \times 10^{-6}$) with surname frequences in US

Fig. 1 Quantitative assessment of identification via surname inference.

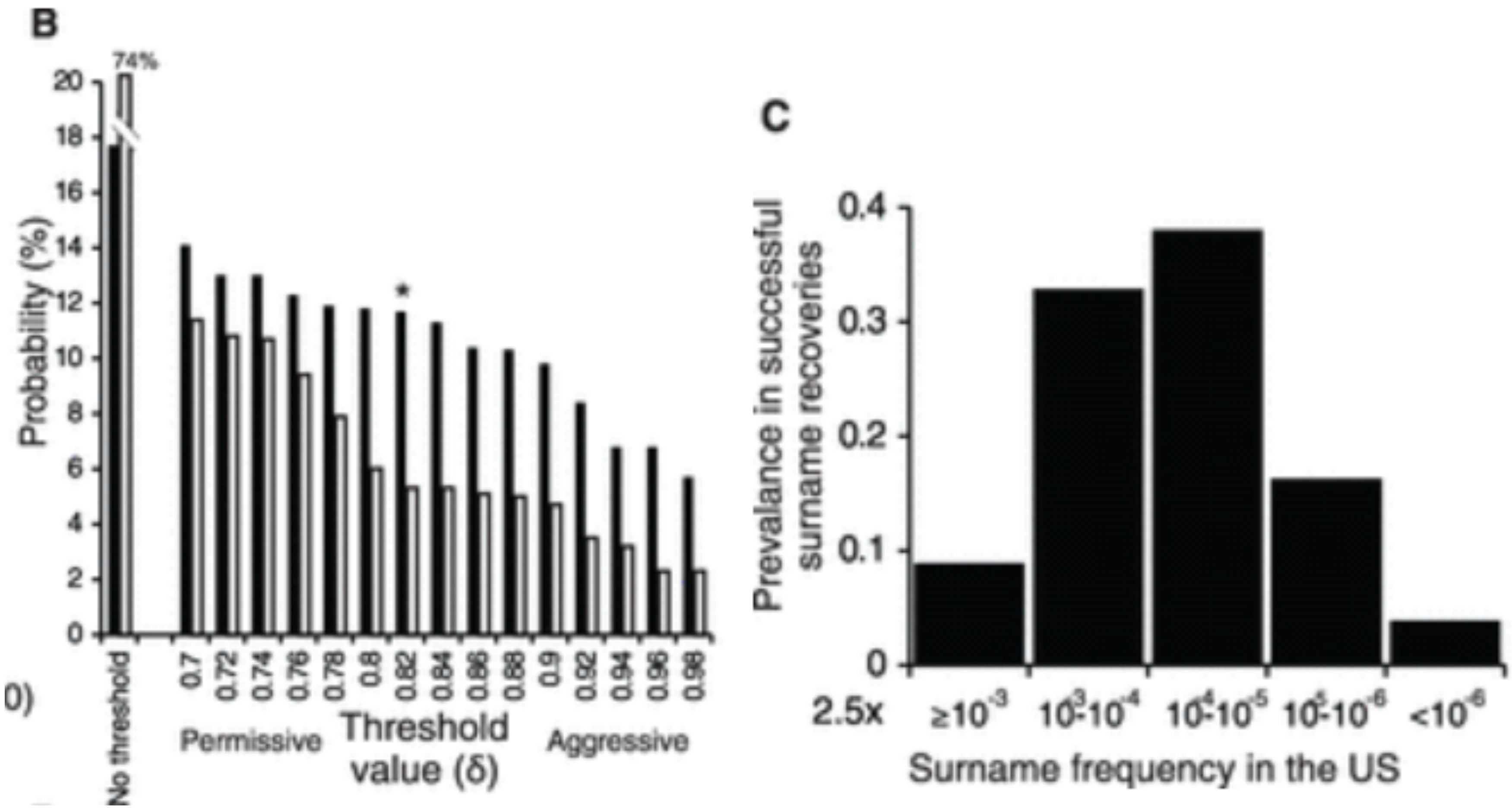


M Gymrek et al. Science 2013;339:321-324

Testing probability of surname inference 1.1

- **Challenging the two databases** with cohort of **Y-STR haplotypes** (consisting of 34 markers) from 911 individuals, Caucasian ancestry, with 521 surnames (known) – compiled from **YBase database**
- **Haplotype query** -> algorithm -> retrieving the database record with the shortest TMRCA (the time to most recent common ancestor)
- Calculating a **confidence score,**
  - If the score passed a user-defined theshold -> surname to the input haplotype, otherwise -> „unknown“
- **Testing the algorithm with range of confidence thresholds**: trade-off between successful vs wrong recovery of surnames.
- **Weighted the results** using a stratified sampling approach -> the expected distribution of recovered surnames as a function of their prevalence

Fig. 1 Quantitative assessment of identification via surname inference.



M Gymrek et al. Science 2013;339:321-324

Figure S2:

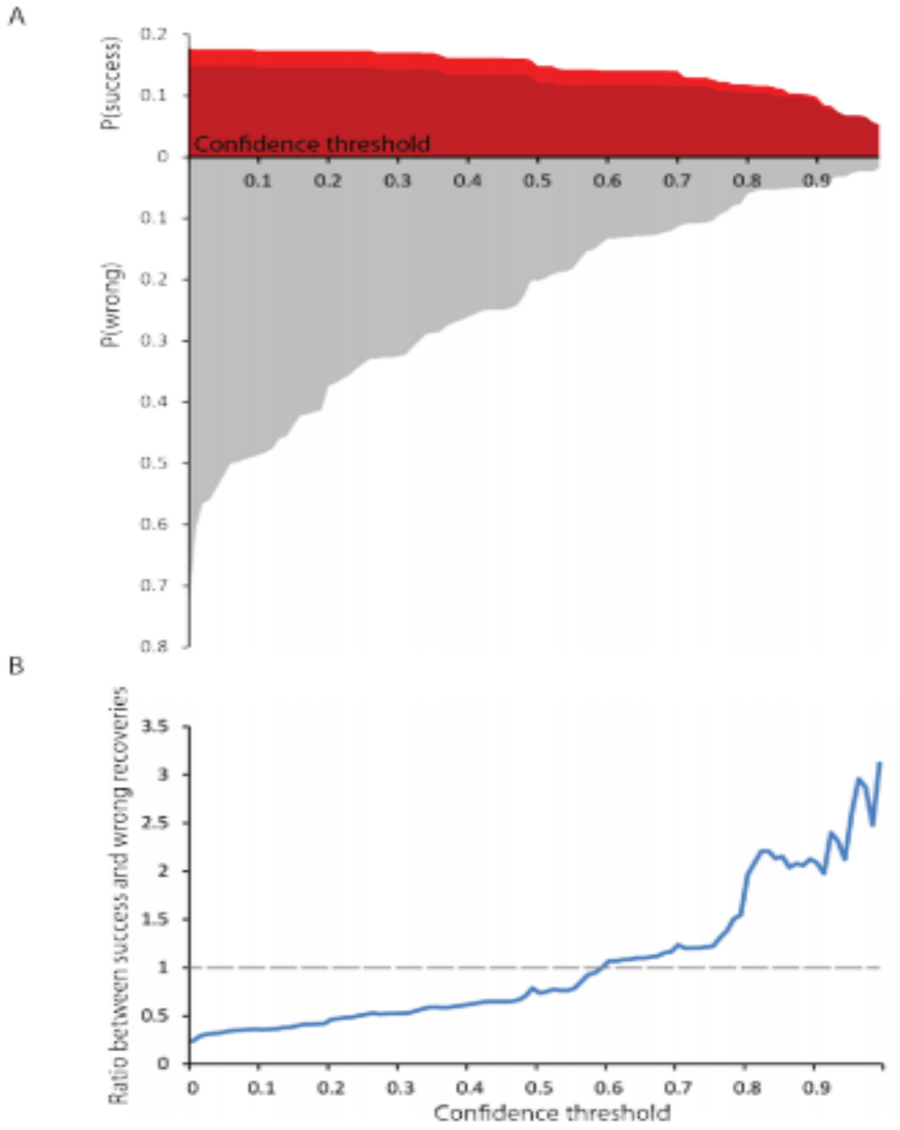


Figure S2: Performance of surname recovery at different confidence thresholds. (**A**) The rate of successful recovery with exact matches (dark red) and spelling variants (light red) versus the wrong recovery rate (gray) as a function of confidence threshold level. (**B**) The ratio between successful recoveries to wrong recoveries.

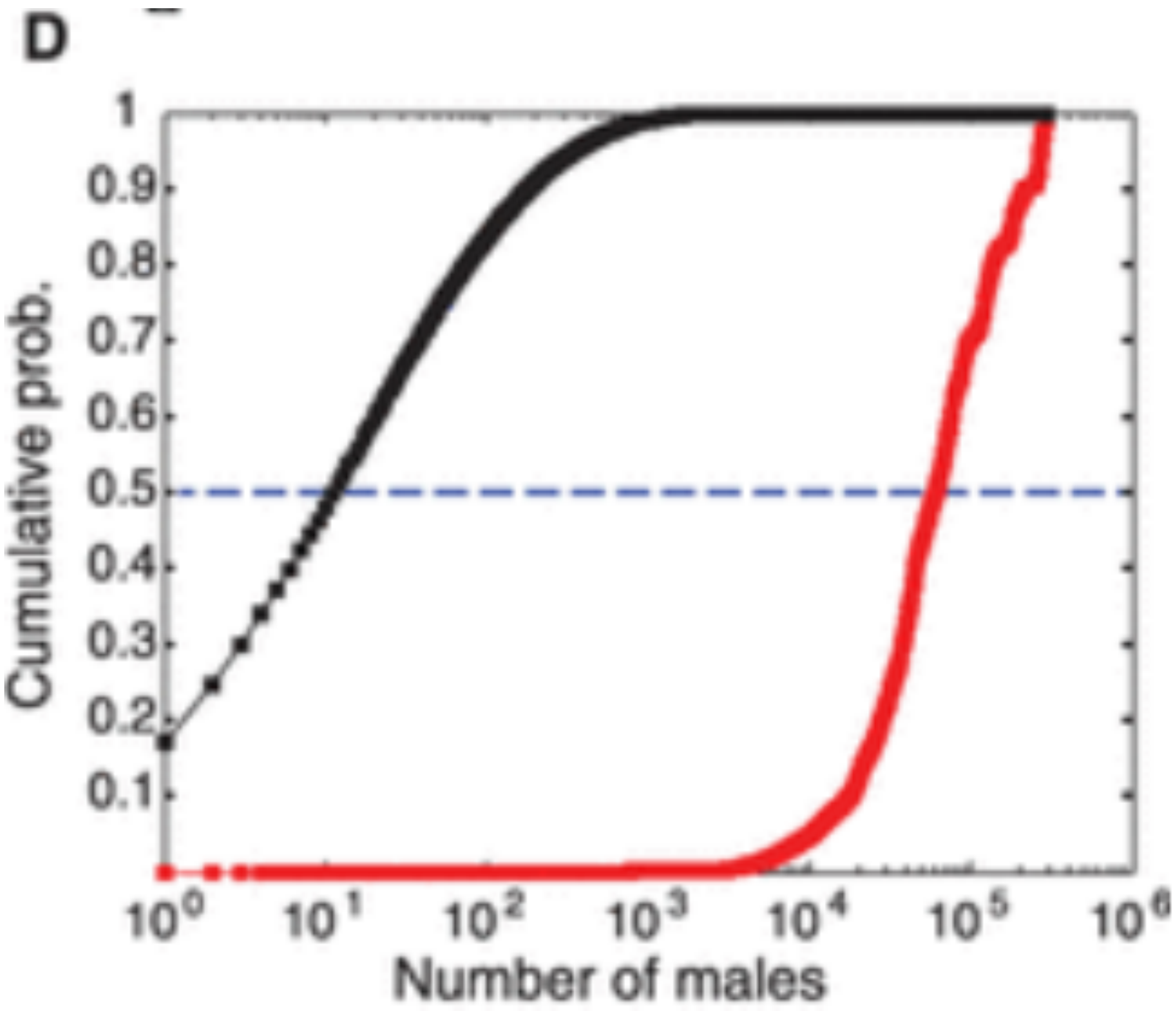
M Gymrek et al. Science 2013;339:321-324

Testing probability of surname inference 2.2

- Combining the recovered **surname with demographic data**
- Various online public record search engines: PeopleFinders.com, USA-people-search.com,…
  - – Search individuals by year of birth, state, surname combinations
- **Results:** (Using U.S. Census data)

  year of **birth + state alone** are weak identifies, but combination of **age+ state+ surname** narrowed significantly the list of matched individuals

Fig. 1 Quantitative assessment of identification via surname inference.



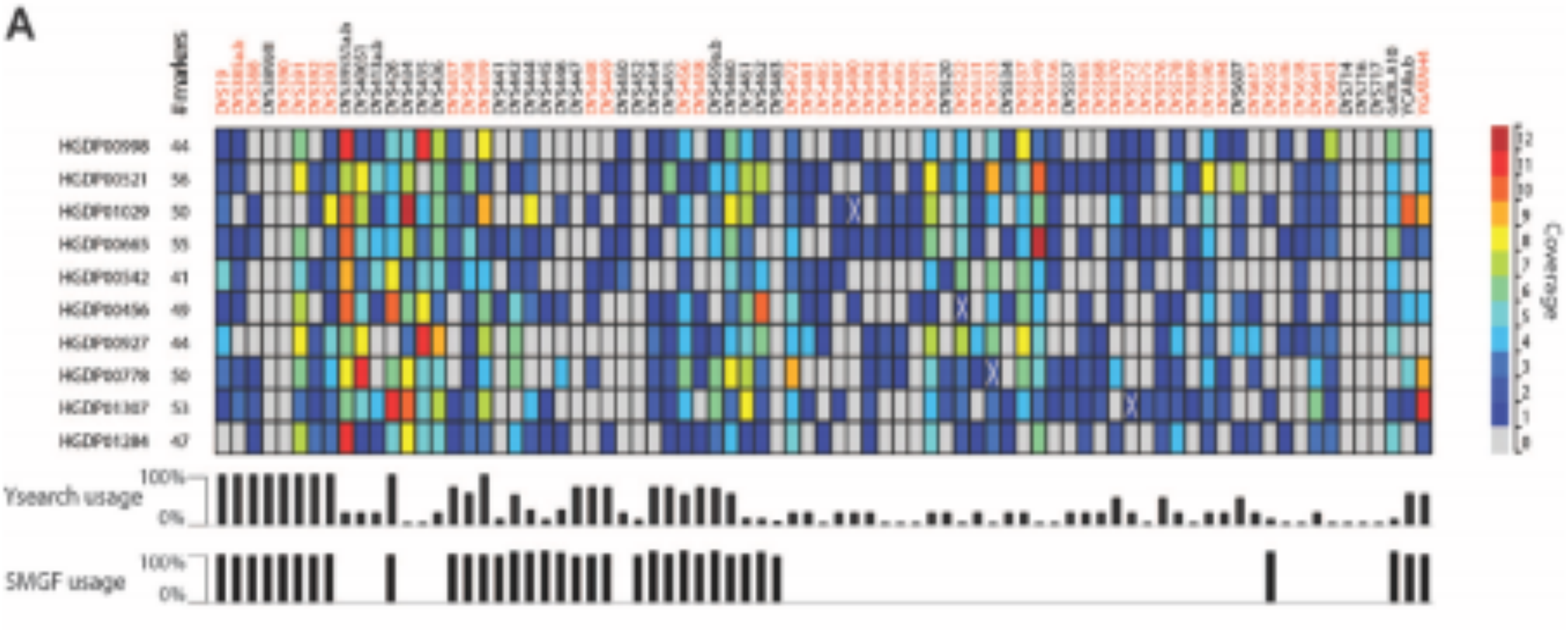
M Gymrek et al. Science 2013;339:321-324

Feasibility of Illumina sequencing to produce accurate Y-STR haplotypes

- **lobSTR** v2.0.0 – algorithm for STR profiling from raw sequencing reads
- **10 high-coverage male genomes** from the HGDP (Human Genome Diversity Panel) (downloaded from NCBI Short Read Archive)
- -> **Y-STR haplotypes** (with an average number of 53 out of the possible 79 genealogical markers)
- Comparing these haplotypes (47 markers) to capillary electrophoresis results -> 99% accuracy
- Even at lower sequencing coverage of 10x, lobSTR can give informative haplotypes

Fig S4. lobSTR calling performance on Y-STR haplotypes from ten male genomes

Figure S4:



M Gymrek et al. Science 2013;339:321-324

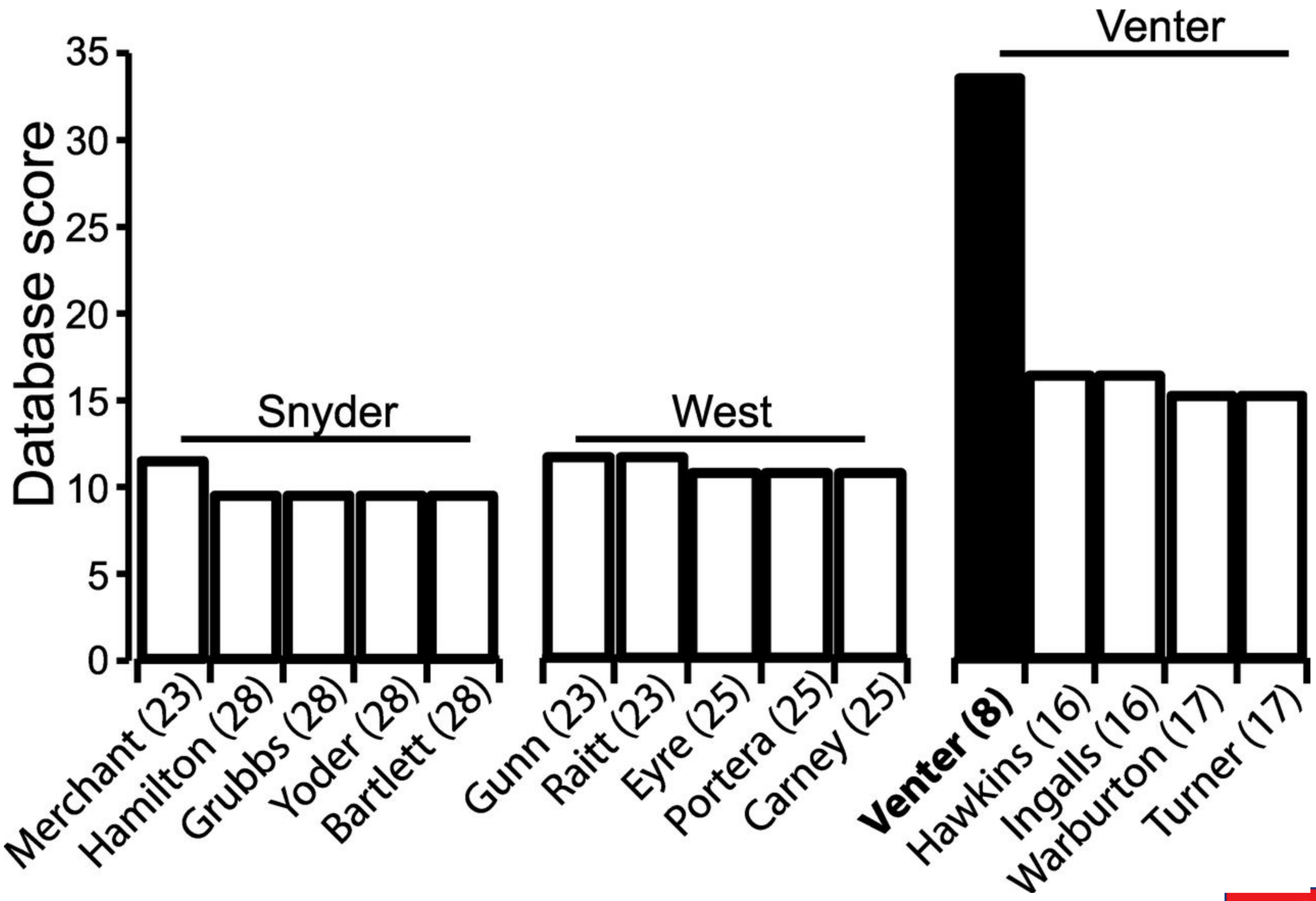
The ability to retrieve genetic genealogy records with the Illumina haplotypes

- **Genome of US Caucasian male** (Illumina 100-pb reads, coverage of 13x) -> **profiled STRs**
- -> **Ysearch** database
- **Results**: a search with the Illumina haplotype returned his Ysearch entry as a top record

Identifing John West, Michael Snyder and Craig Venter

- **Genomes from identified individuals** in NCBI archives – good test cases for identification via surname inference
- **Genomic data -> Y-STR haplotypes** (using **lobSTR**)
- **Ysearch** and **SMGF** databases
- **Results:**
  West and Snyder did not return their surnames, Venter`s haplotype returned a clear match to a „Venter“ record (33 comparable markers and TMRCA less than 8 generat.)
- **Combining the inferred surname with demographic profiling** (age+ state+ surname) -> 2 matching records of male

Fig. 2 The top five records retrieved after searching Ysearch with the Y-STR haplotypes of Michael Snyder, John West, and Craig Venter.



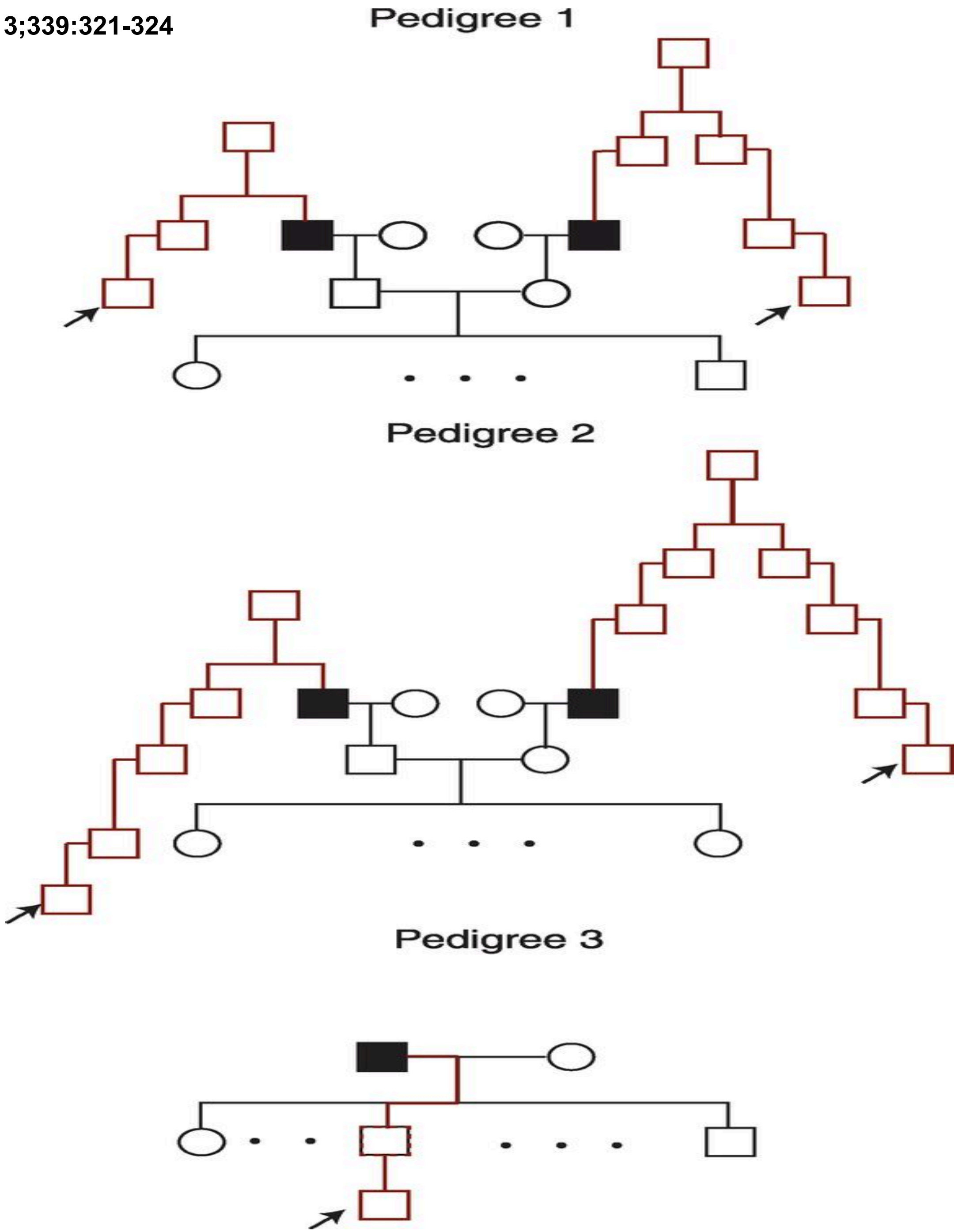
M Gymrek et al. Science 2013;339:321-324

Published by AAAS

Conclusions

- Each complete pedigree re-identification took **3 to 7 hours** by single person
- Data release, even of a few markers, from one person can spread through deep genealogical ties and lead to the **identification of another person** who might have no acquaintance with the person who released his genetic data
- This identification technique entirely relies on **free, publicly available resources**
- Genetic genealogy enthusiasts add thousands of records to these databases every month
- **Third-generation sequencing platform**, longer reads -> higher coverage of Y-STR markers -> further strengthening the ability to link haplotypes and surnames

M Gymrek et al. Science 2013;339:321-324



Published by AAAS

Conclusions - solutions

- **Masking Y-STR markers** – not sustainable
- **Restricting genetic genealogy information** – not practical
- **Controlled-access databases with data use agreements** – may mediate the exposure of genomic information to surname inference
- **Clear policies for data sharing, educating participants** about benefits and risks of genetic studies, and the **legislation of proper usage of genetic information**

Thanks for listening!