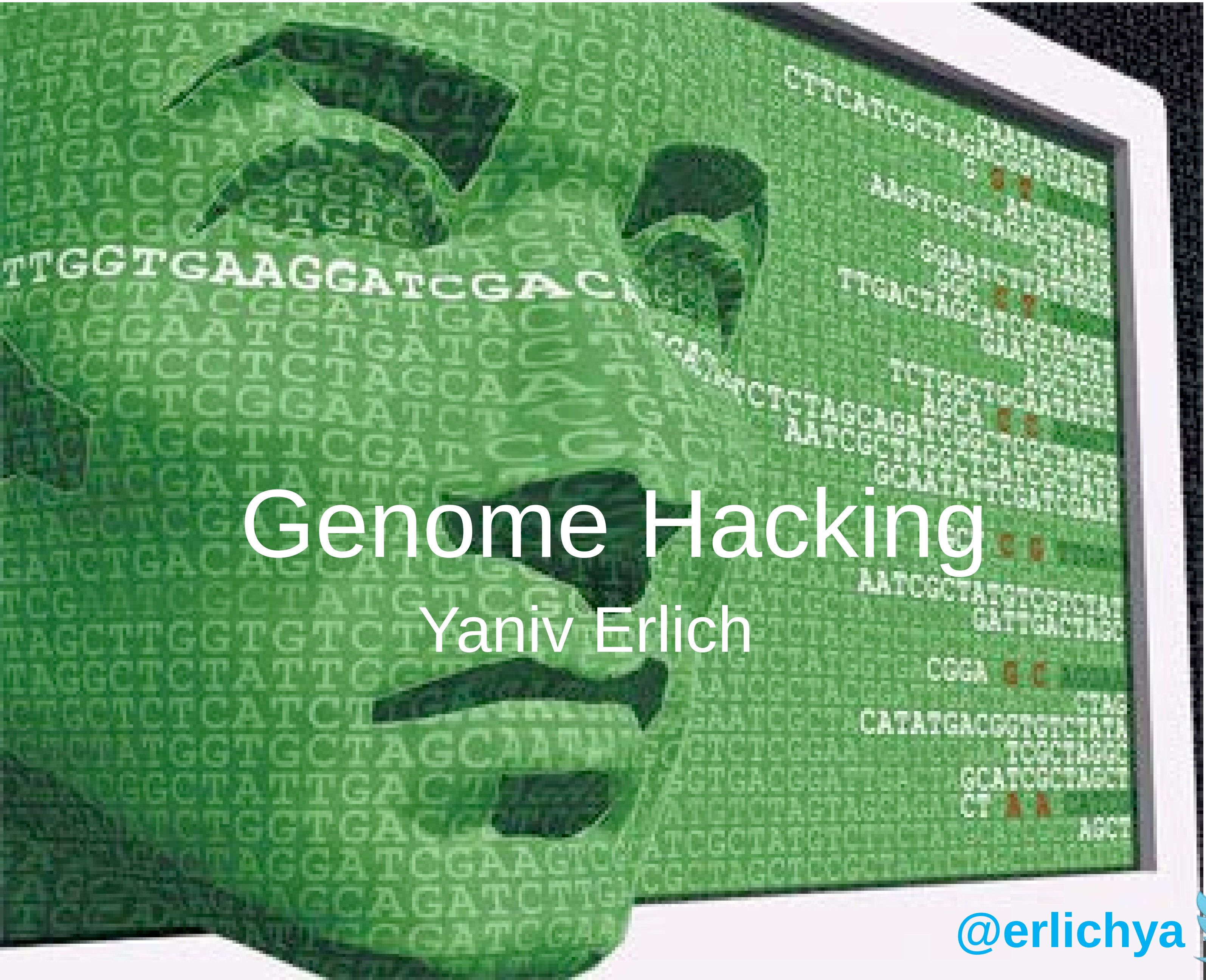


Genome Hacking

Yaniv Erlich

Columbia University

New York Genome Center

@erlichya

@erlichya

We need to share genetic information

Hereditary Spastic Paraparesis
(Erlich et al.)

GENOME RESEARCH

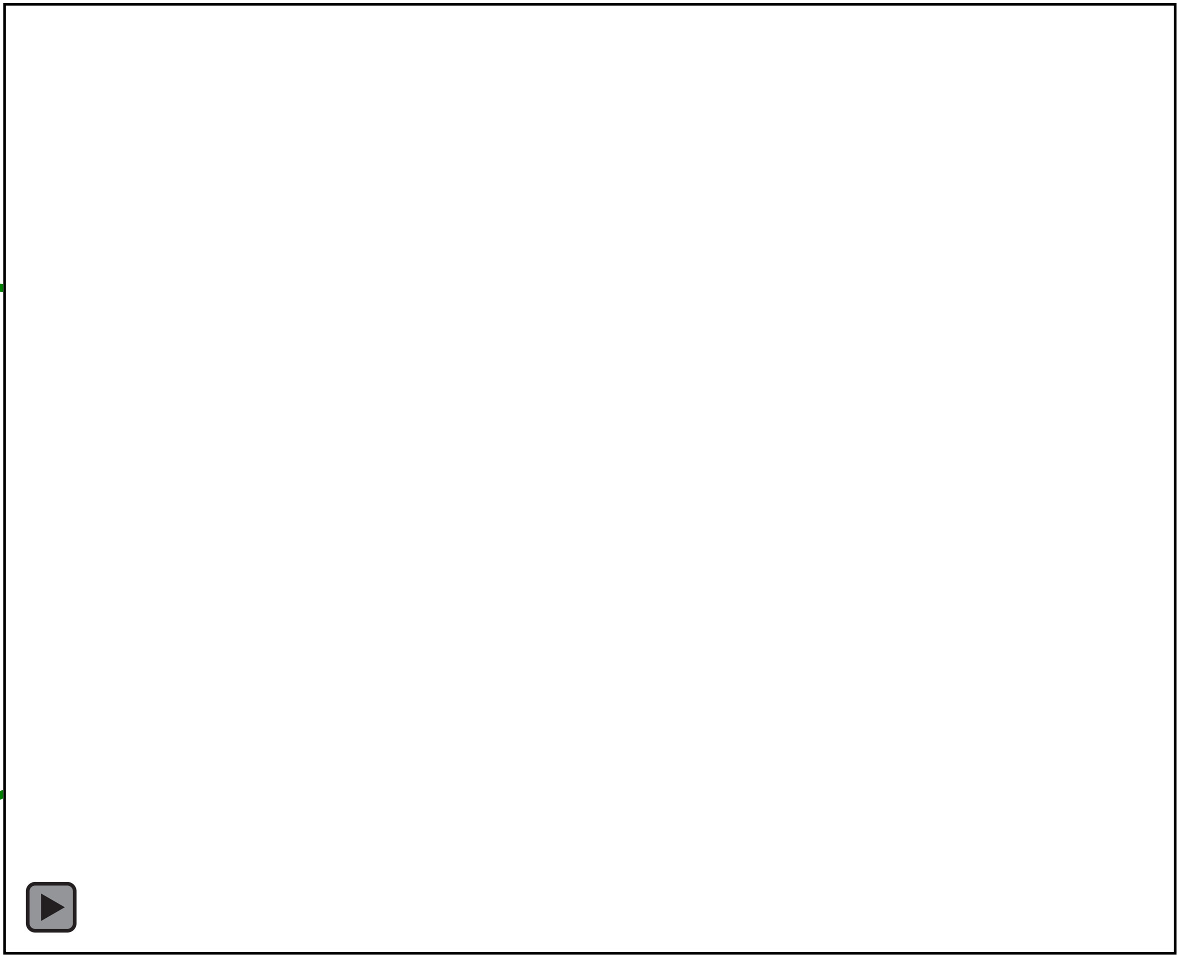
Joubert syndrome
(Endevson et al.)

AJHG The American Journal of Human Genetics

Hemifacial Microsomia
(Zielinski,.., & Erlich)

PLoS One

Vulnerability research



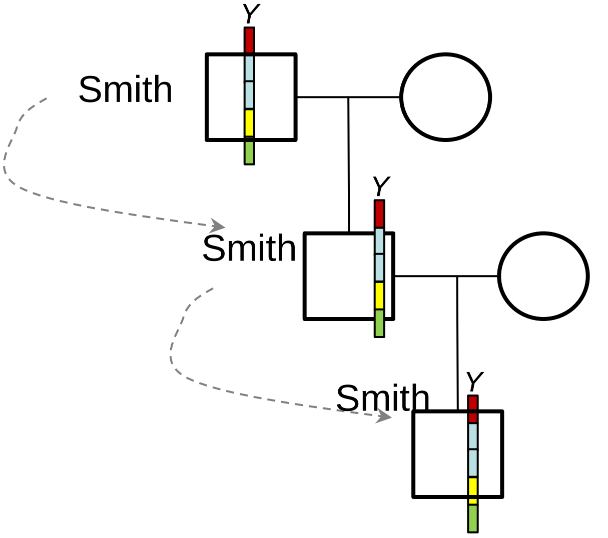
Intercom

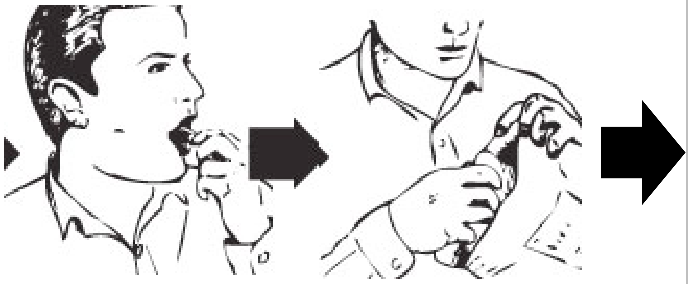
Me

IT department of a major bank

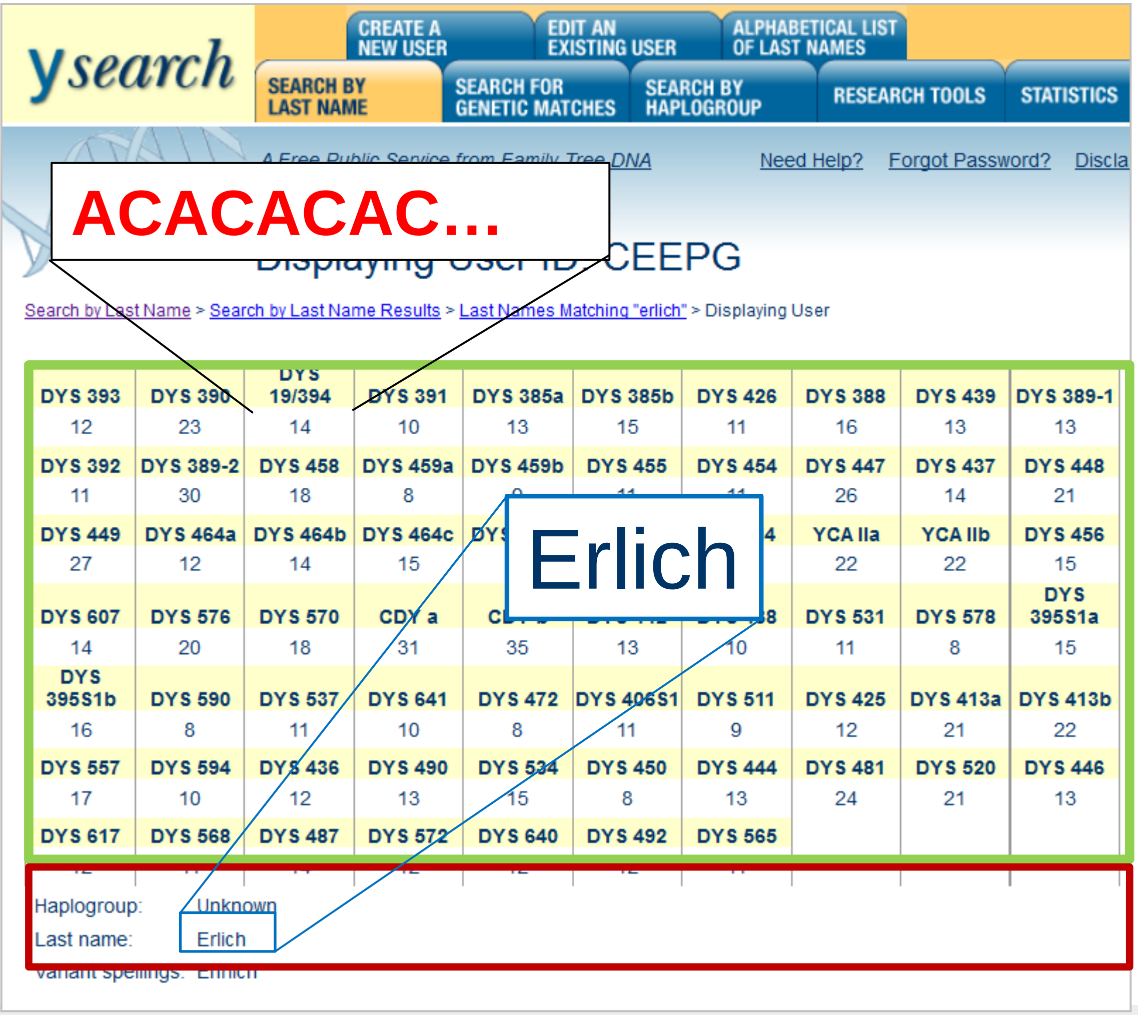
Fingerprint reader

Correlation between Y-chr and surnames





www.ysearch.org:



Displaying User ID: CEEPG

Search by Last Name > Search by Last Name Results > Last Names Matching "erlich" > Displaying User

DYS 393	DYS 390	DYS 19/394	DYS 391	DYS 385a	DYS 385b	DYS 426	DYS 388	DYS 439	DYS 389-1
12	23	14	10	13	15	11	16	13	13
DYS 392	DYS 389-2	DYS 458	DYS 459a	DYS 459b	DYS 455	DYS 454	DYS 447	DYS 437	DYS 448
11	30	18	8	9	11	11	26	14	21
DYS 449	DYS 464a	DYS 464b	DYS 464c	DYS [illegible]	[illegible]	[illegible]4	YCA IIa	YCA IIb	DYS 456
27	12	14	15	[illegible]	[illegible]	[illegible]	22	22	15
DYS 607	DYS 576	DYS 570	CDY a	CDY b	DYS 442	DYS 438	DYS 531	DYS 578	DYS 395S1a
14	20	18	31	35	13	10	11	8	15
DYS 395S1b	DYS 590	DYS 537	DYS 641	DYS 472	DYS 406S1	DYS 511	DYS 425	DYS 413a	DYS 413b
16	8	11	10	8	11	9	12	21	22
DYS 557	DYS 594	DYS 436	DYS 490	DYS 534	DYS 450	DYS 444	DYS 481	DYS 520	DYS 446
17	10	12	13	15	8	13	24	21	13
DYS 617	DYS 568	DYS 487	DYS 572	DYS 640	DYS 492	DYS 565			
12	11	14	12	12	12	11			

Haplogroup: Unknown
Last name: Erlich
Variant spellings: Erlich

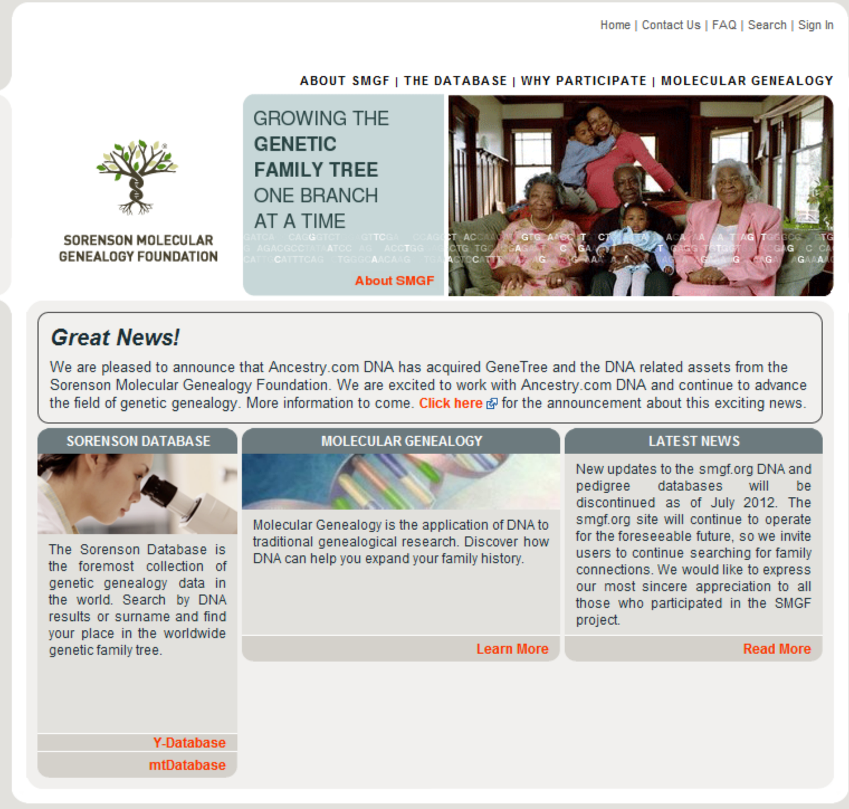
The main idea

A systematic study: can we recover the identity of anonymous genomic datasets?

Databases of interest

140,000 **publicly** accessible surname-Ychr records

www.smgf.org


Home | Contact Us | FAQ | Search | Sign In
ABOUT SMGF | THE DATABASE | WHY PARTICIPATE | MOLECULAR GENEALOGY
SORENSON MOLECULAR GENEALOGY FOUNDATION
GROWING THE GENETIC FAMILY TREE ONE BRANCH AT A TIME
About SMGF

Great News!

We are pleased to announce that Ancestry.com DNA has acquired GeneTree and the DNA related assets from the Sorenson Molecular Genealogy Foundation. We are excited to work with Ancestry.com DNA and continue to advance the field of genetic genealogy. More information to come. Click here for the announcement about this exciting news.

SORENSON DATABASE

The Sorenson Database is the foremost collection of genetic genealogy data in the world. Search by DNA results or surname and find your place in the worldwide genetic family tree.

Y-Database

mtDatabase

MOLECULAR GENEALOGY

Molecular Genealogy is the application of DNA to traditional genealogical research. Discover how DNA can help you expand your family history.

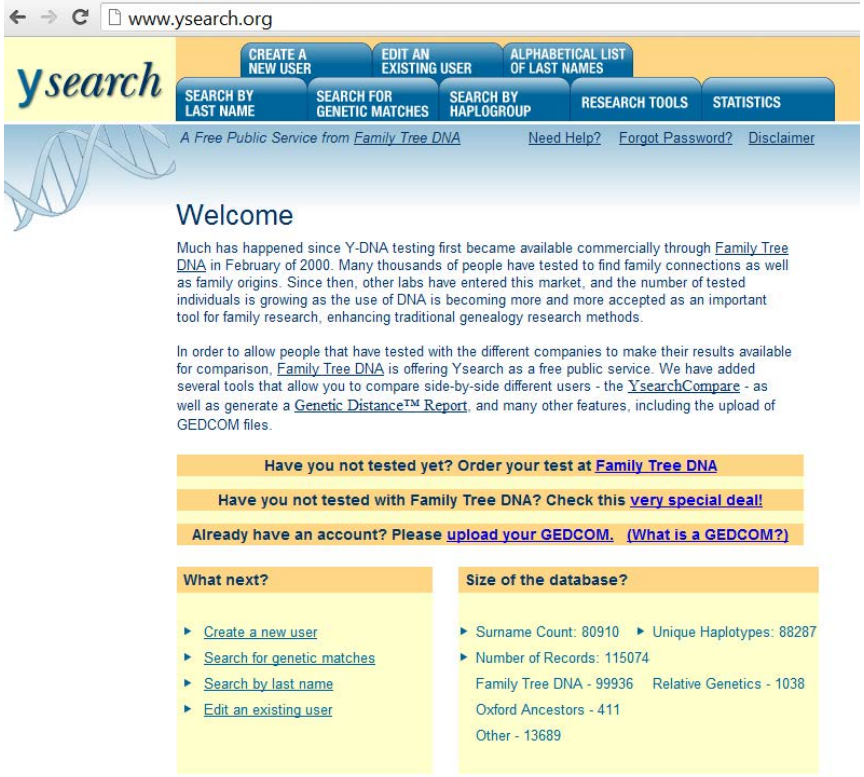
Learn More

LATEST NEWS

New updates to the smgf.org DNA and pedigree databases will be discontinued as of July 2012. The smgf.org site will continue to operate for the foreseeable future, so we invite users to continue searching for family connections. We would like to express our most sincere appreciation to all those who participated in the SMGF project.

Read More

www.ysearch.org



Welcome

Much has happened since Y-DNA testing first became available commercially through Family Tree DNA in February of 2000. Many thousands of people have tested to find family connections as well as family origins. Since then, other labs have entered this market, and the number of tested individuals is growing as the use of DNA is becoming more and more accepted as an important tool for family research, enhancing traditional genealogy research methods.

In order to allow people that have tested with the different companies to make their results available for comparison, Family Tree DNA is offering Ysearch as a free public service. We have added several tools that allow you to compare side-by-side different users - the YsearchCompare - as well as generate a Genetic Distance™ Report, and many other features, including the upload of GEDCOM files.

Have you not tested yet? Order your test at Family Tree DNA

Have you not tested with Family Tree DNA? Check this very special deal!

Already have an account? Please upload your GEDCOM. (What is a GEDCOM?)

What next?

- Create a new user
- Search for genetic matches
- Search by last name
- Edit an existing user

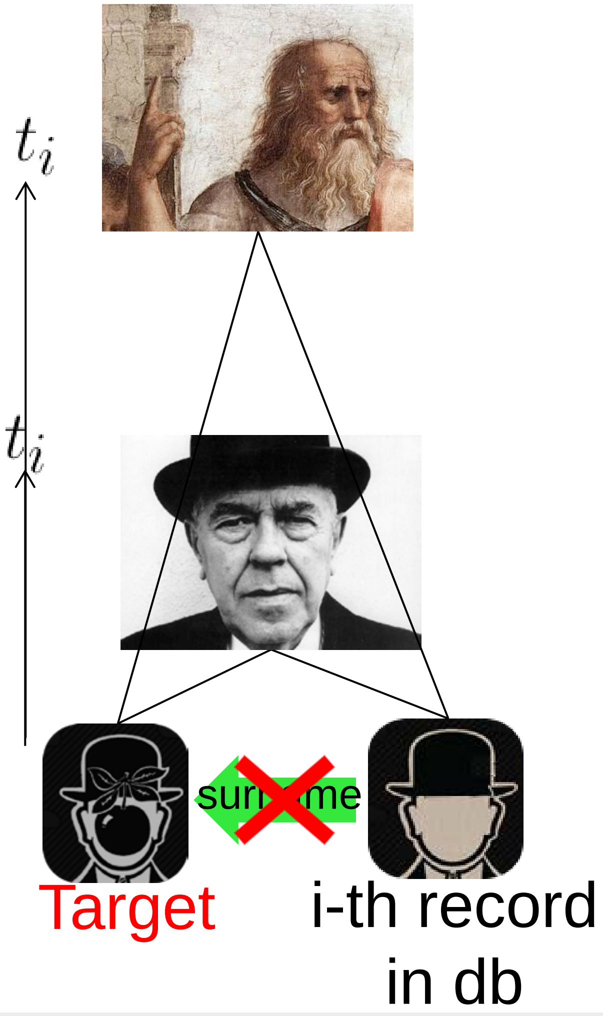
Size of the database?

- Surname Count: 80910
- Unique Haplotypes: 88287
- Number of Records: 115074

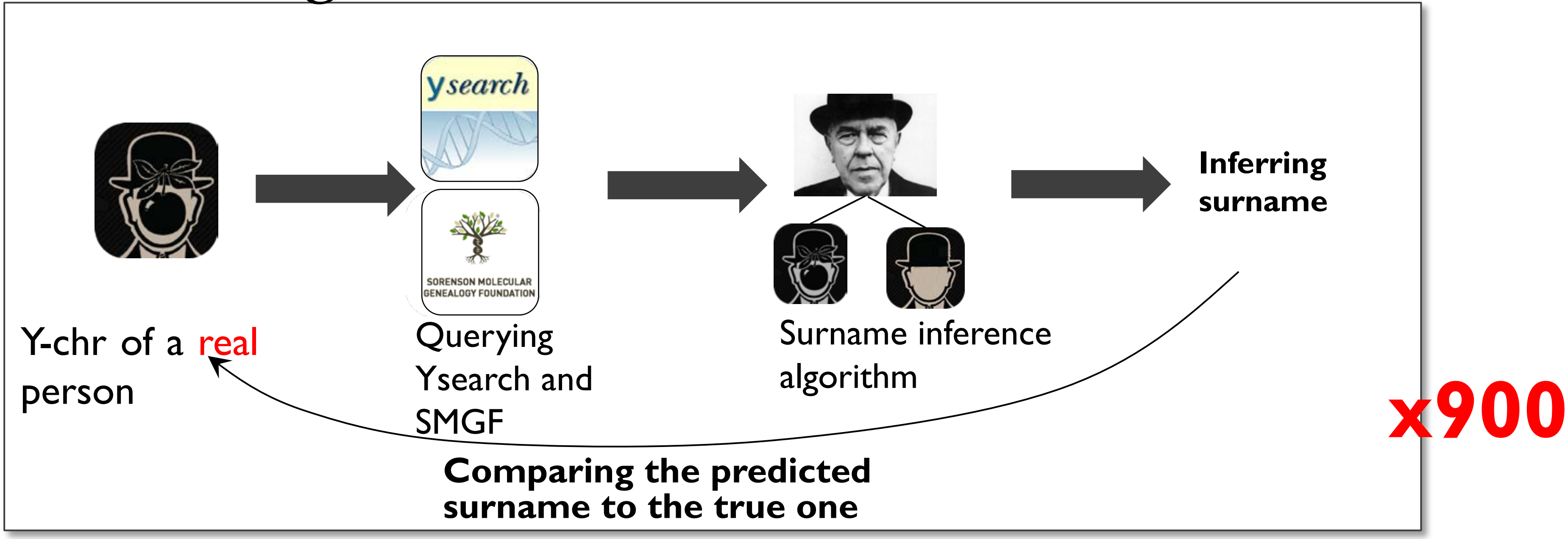
Family Tree DNA - 99936
Relative Genetics - 1038
Oxford Ancestors - 411
Other - 13689

How to find surnames?

Estimating the **time** to most recent common ancestor



Empirical test to determine the probability of recovering a US surname



Y-chr of a real person

Querying Ysearch and SMGF

Surname inference algorithm

Inferring surname

Comparing the predicted surname to the true one

x900

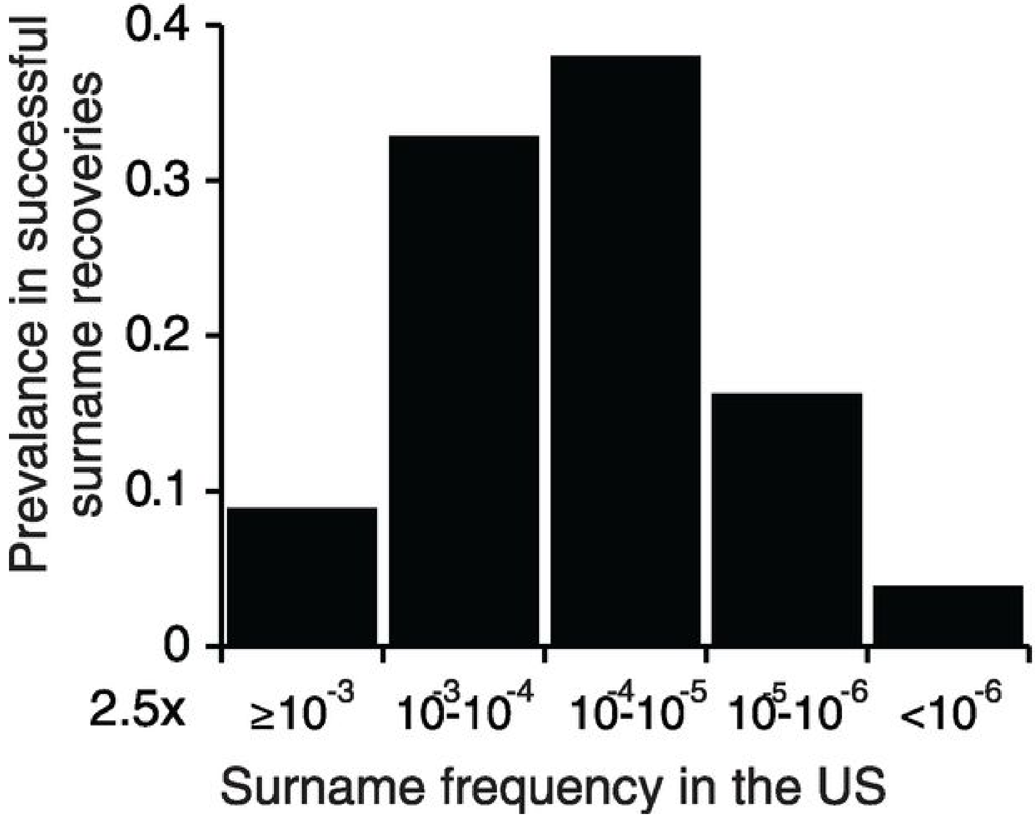
For US Caucasian males:

12% Successful recoveries

5% Wrong recoveries

83% Unknown

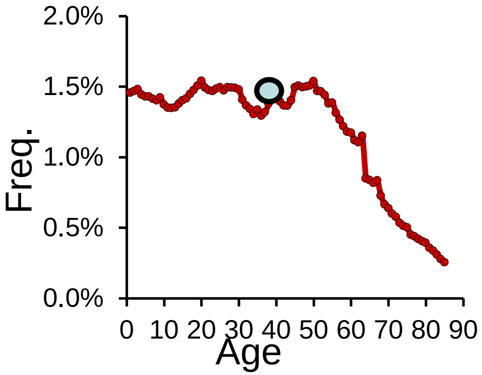
Distribution of inferred surnames



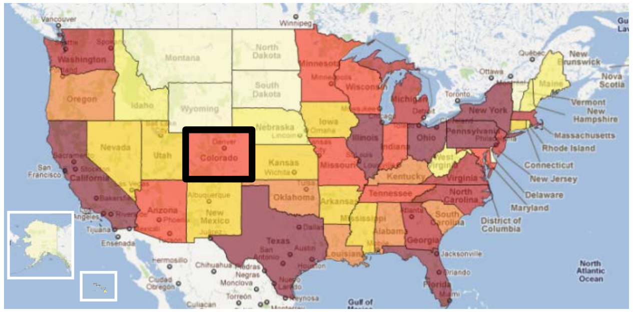
Most of the inferred surnames are relatively rare

Triangulate individuals with metadata

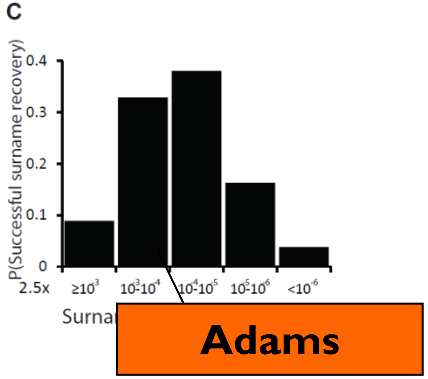
Age



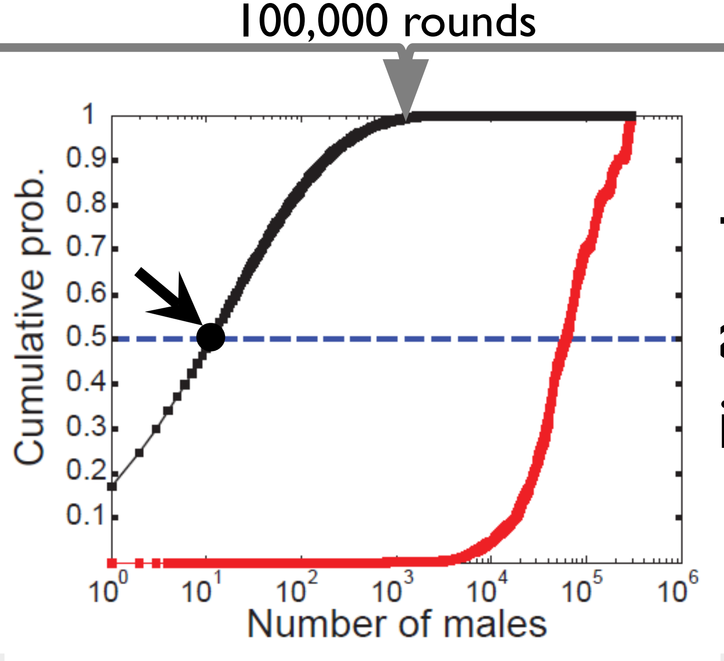
State



Surname



100,000 rounds

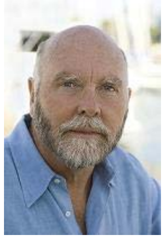


■ Age+state+surname

🟥 Only age+state

The median of age+state+surname is12 males.

Putting it all together: the Venter case



lobSTR

We got a surname from whole genome sequencing data

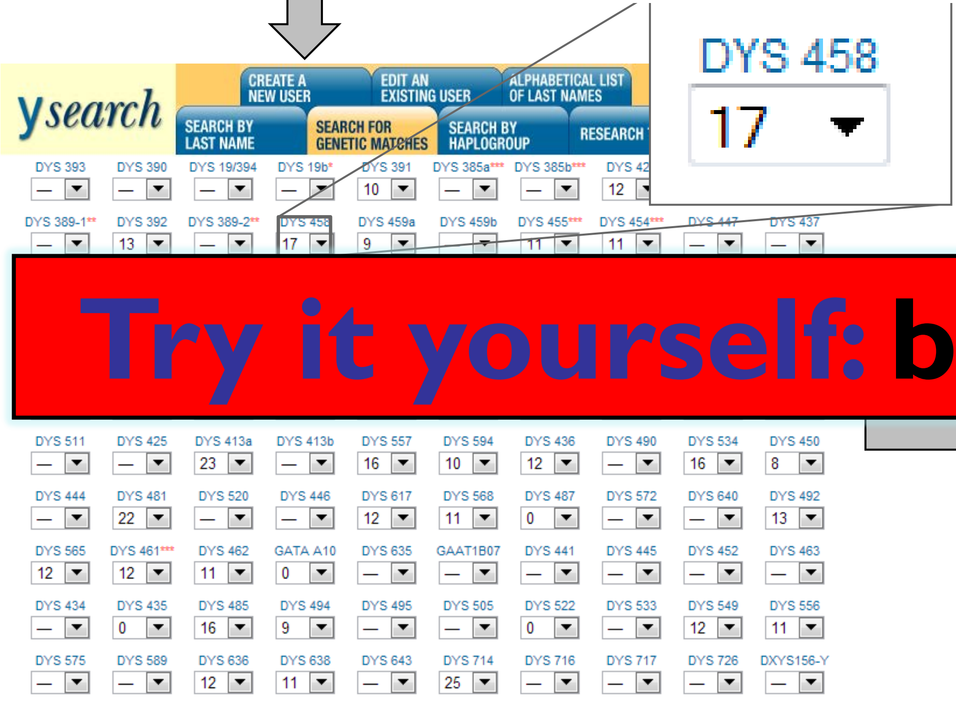
Method

GENOME RESEARCH

lobSTR: A short tandem repeat profiler for personal genomes

Melissa Gymrek,[1,2] David Golan,[2,3] Saharon Rosset,[3] and Yaniv Erlich[2,4]

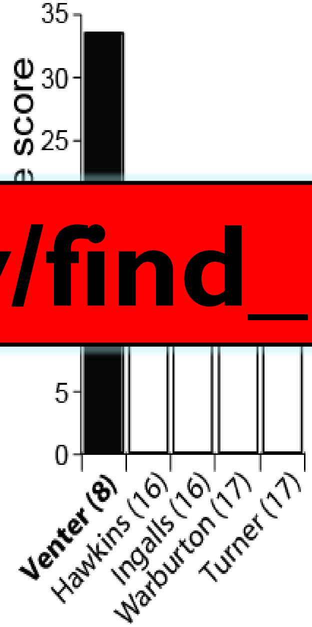
[1] Harvard–MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [2] Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA; [3] Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel



DYS 458

17

DYS 393	DYS 390	DYS 19/394	DYS 19b*	DYS 391	DYS 385a***	DYS 385b***	DYS 42		
—	—	—	—	10	—	—	12		
DYS 389-1**	DYS 392	DYS 389-2**	DYS 458	DYS 459a	DYS 459b	DYS 455***	DYS 454***	DYS 447	DYS 437
—	13	—	17	9		11	11	—	—
DYS 511	DYS 425	DYS 413a	DYS 413b	DYS 557	DYS 594	DYS 436	DYS 490	DYS 534	DYS 450
—	—	23	—	16	10	12	—	16	8
DYS 444	DYS 481	DYS 520	DYS 446	DYS 617	DYS 568	DYS 487	DYS 572	DYS 640	DYS 492
—	22	—	—	12	11	0	—	—	13
DYS 565	DYS 461***	DYS 462	GATA A10	DYS 635	GAAT1B07	DYS 441	DYS 445	DYS 452	DYS 463
12	12	11	0	—	—	—	—	—	—
DYS 434	DYS 435	DYS 485	DYS 494	DYS 495	DYS 505	DYS 522	DYS 533	DYS 549	DYS 556
—	0	16	9	—	—	0	—	12	11
DYS 575	DYS 589	DYS 636	DYS 638	DYS 643	DYS 714	DYS 716	DYS 717	DYS 726	DXYS156-Y
—	—	12	11	—	25	—	—	—	—



Try it yourself: bit.ly/find_craig

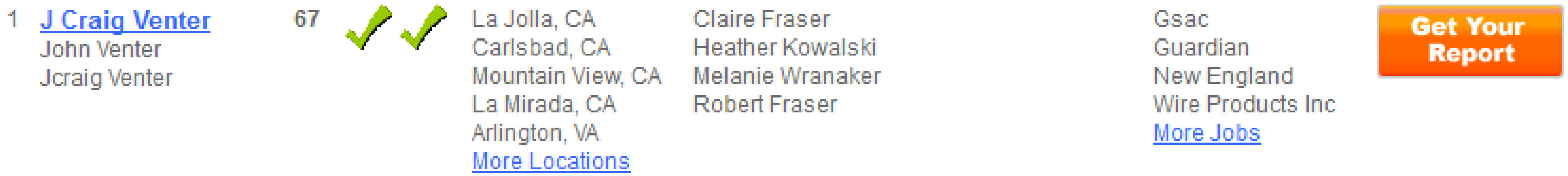
Getting to Craig Venter

Searching for:

1. Venter
2. California
3. Born in 1946
4. Male

In USSearch.com

Two matches, including:



1	J Craig Venter John Venter Jcraig Venter	67	✓ ✓	La Jolla, CA Carlsbad, CA Mountain View, CA La Mirada, CA Arlington, VA More Locations	Claire Fraser Heather Kowalski Melanie Wranaker Robert Fraser	Gsac Guardian New England Wire Products Inc More Jobs	Get Your Report

Can we identify anonymous genomes?

1000 Genomes cases

1000 Genomes A Deep Catalog of Human Genetic Variation

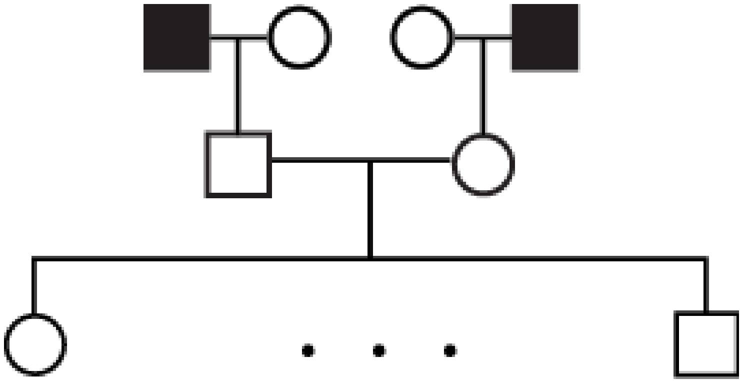
EMBL-EBI

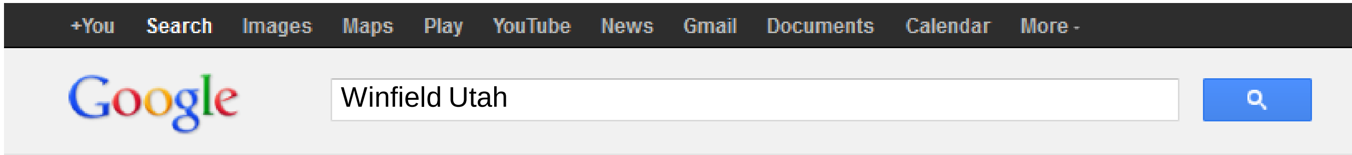
10 CEU (Utah) genomes

SORENSON MOLECULAR GENEALOGY FOUNDATION

ysearch

surname predictions





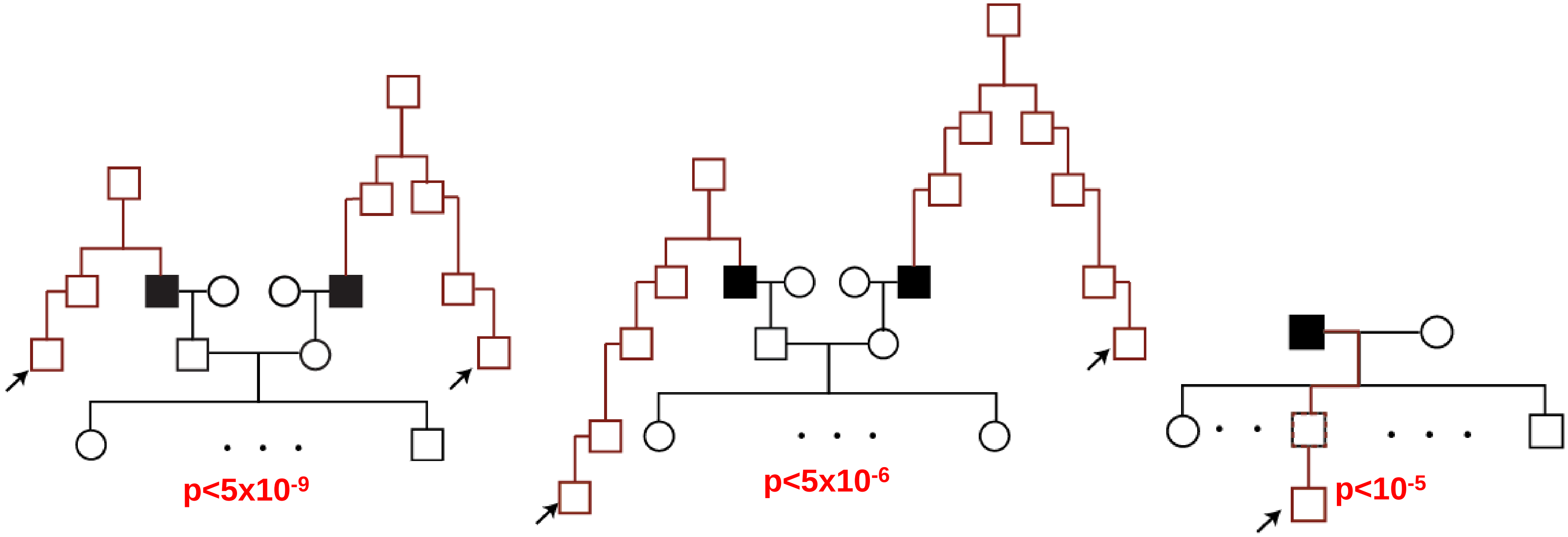
Winfield Utah

Found an obituary that has the exact description of the pedigree

Probability of a random match < 5×10^{-9}

*Some of the details in this slide were modified to respect the identity of the family

Beginner’s luck?



■ Successful surname recovery (targeted individual)

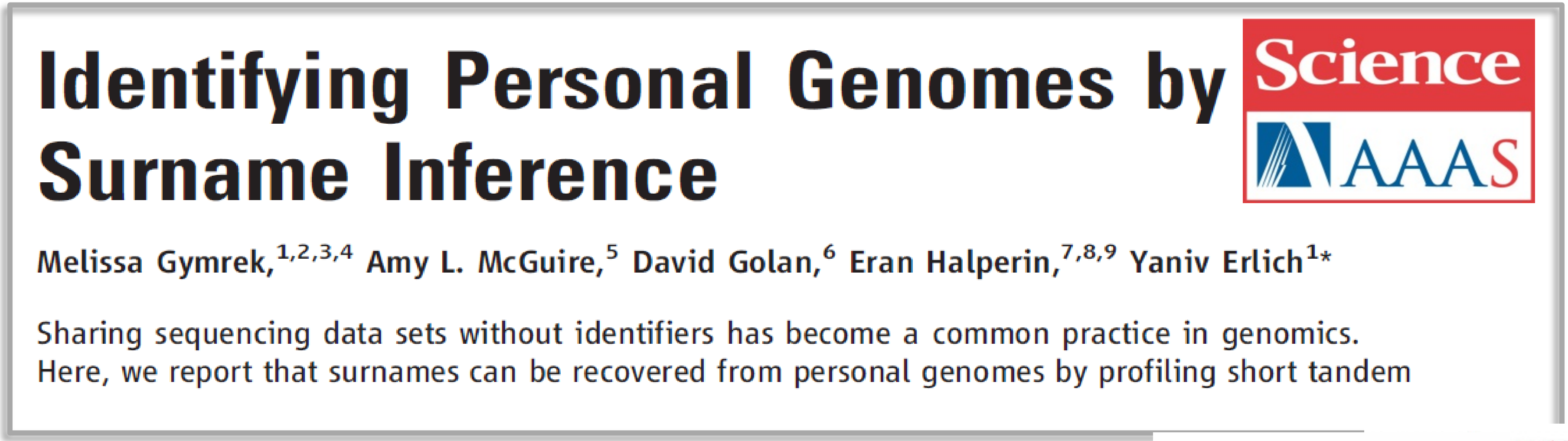
↗ Person tested by genetic genealogy service (source)

— Patrilineal line from source to target

Breaching the privacy of close to **50** CEU samples.

Aftermath

Our study



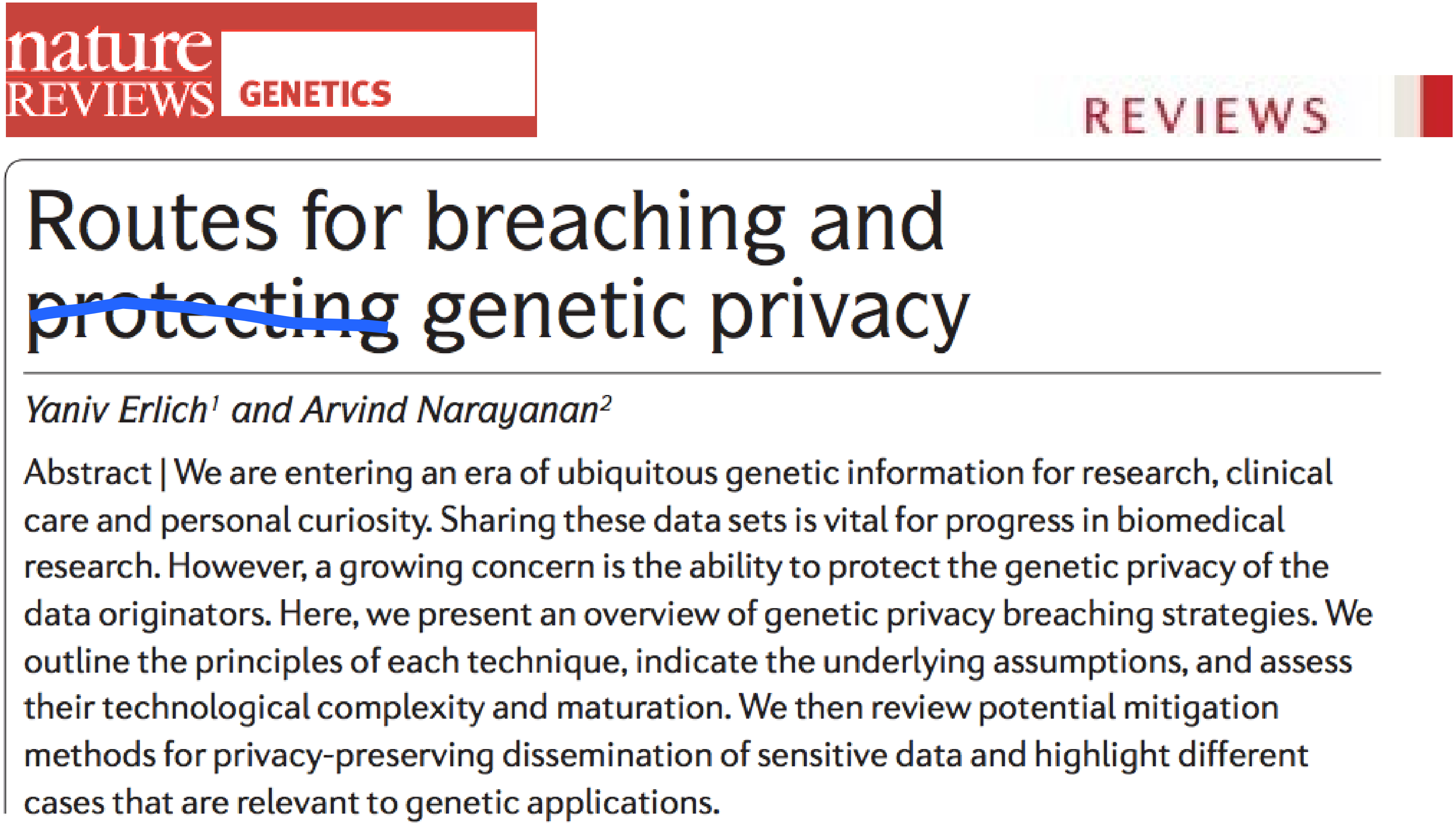
Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem

"..And Erlich respon
findin
resea
an important precede
should
emplary way
d other genetic
ts
aling with
t response
other researchers
npr
The New York Times
Technology Review
Los Angeles Times
THE WALL STREET JOURNAL
EL PAÍS
HAARETZ
USA TODAY
Bloomberg NEWS
SRF Schweizer Radio und Fernsehen
NBC NEWS
BBC NEWS
The Boston Globe
NewScientist
גלצ
La Recherche
TheScientist

The hitchhiker guide to genome hacking



nature REVIEWS GENETICS

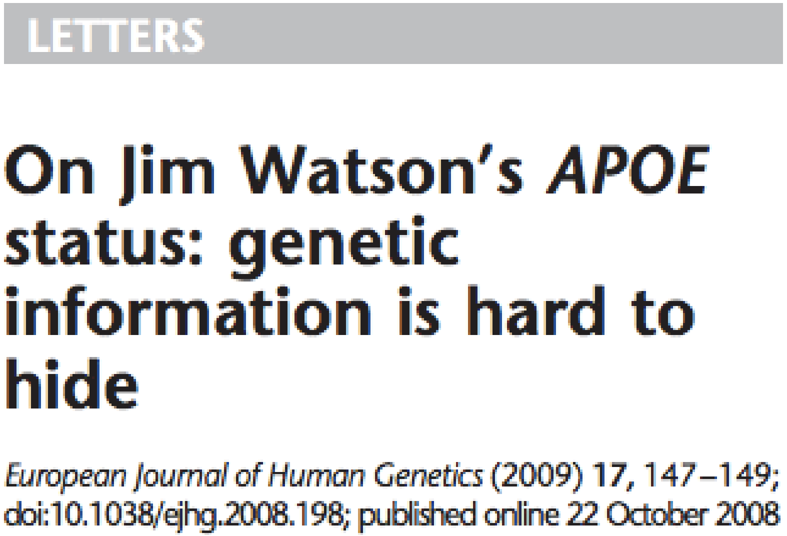
REVIEWS

Routes for breaching and ~~protecting~~ genetic privacy

Yaniv Erlich[1] and Arvind Narayanan[2]

Abstract | We are entering an era of ubiquitous genetic information for research, clinical care and personal curiosity. Sharing these data sets is vital for progress in biomedical research. However, a growing concern is the ability to protect the genetic privacy of the data originators. Here, we present an overview of genetic privacy breaching strategies. We outline the principles of each technique, indicate the underlying assumptions, and assess their technological complexity and maturation. We then review potential mitigation methods for privacy-preserving dissemination of sensitive data and highlight different cases that are relevant to genetic applications.

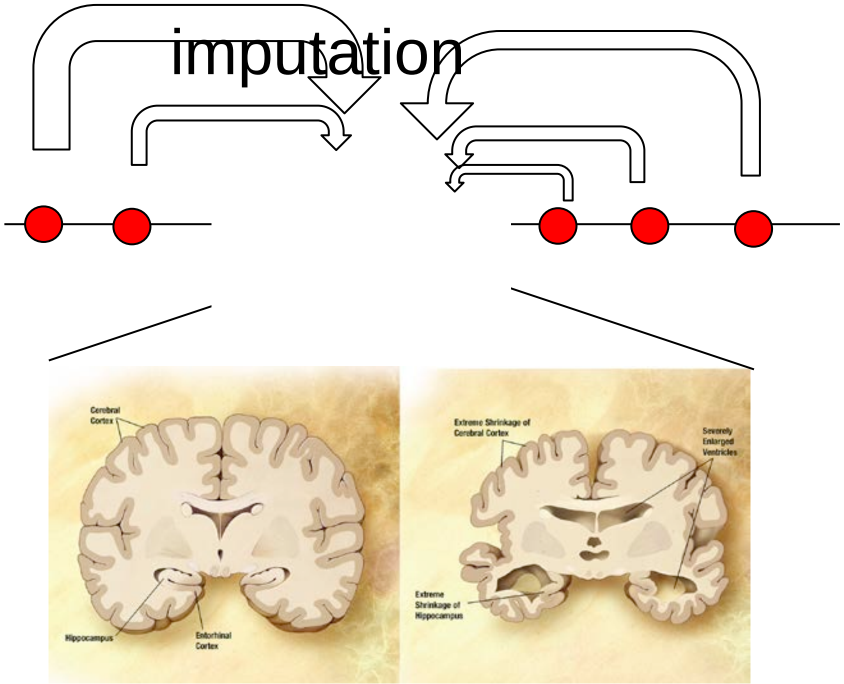
The hitchhiker guide to genome hacking



LETTERS

On Jim Watson's *APOE* status: genetic information is hard to hide

European Journal of Human Genetics (2009) 17, 147–149; doi:10.1038/ejhg.2008.198; published online 22 October 2008

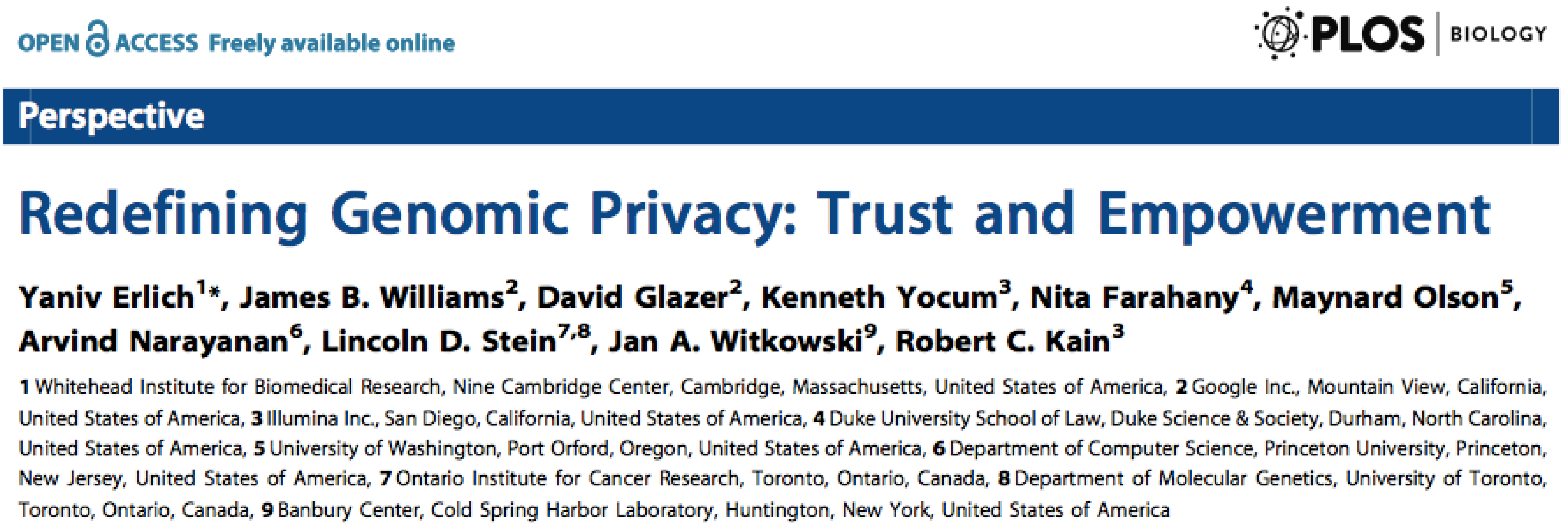

Cerebral Cortex
Hippocampus
Entorhinal Cortex
Extreme Shrinkage of Cerebral Cortex
Severely Enlarged Ventricles
Extreme Shrinkage of Hippocampus

Genetic imputation

Alzheimer's disease

Barack Obama is the President

The path forward



OPEN ACCESS Freely available online

PLOS | BIOLOGY

Perspective

Redefining Genomic Privacy: Trust and Empowerment

Yaniv Erlich[1]*, James B. Williams[2], David Glazer[2], Kenneth Yocum[3], Nita Farahany[4], Maynard Olson[5], Arvind Narayanan[6], Lincoln D. Stein[7,8], Jan A. Witkowski[9], Robert C. Kain[3]

1 Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts, United States of America, **2** Google Inc., Mountain View, California, United States of America, **3** Illumina Inc., San Diego, California, United States of America, **4** Duke University School of Law, Duke Science & Society, Durham, North Carolina, United States of America, **5** University of Washington, Port Orford, Oregon, United States of America, **6** Department of Computer Science, Princeton University, Princeton, New Jersey, United States of America, **7** Ontario Institute for Cancer Research, Toronto, Ontario, Canada, **8** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, **9** Banbury Center, Cold Spring Harbor Laboratory, Huntington, New York, United States of America

Key points: transparency, reputation system, compensation

NEW YORK GENOME CENTER

Yaniv Erlich

@erlichya

COLUMBIA UNIVERSITY

Acknowledgements

Team Genetic Privacy

Melissa Gymrek (HST – Harvard/MIT)
Amy McGuire (Baylor)
David Golan (Tel-Aviv University)
Eran Halperin (Tel-Aviv University)

Funding:

BURROUGHS WELLCOME FUND

National Human Genome Research Institute

BROAD INSTITUTE

Andria and Paul Heafy