

Toward a universal DNA database

DNA PROFILES: A PRIVACY PERSPECTIVE

Michael Seringhaus
Yale Law School 2010

Overview

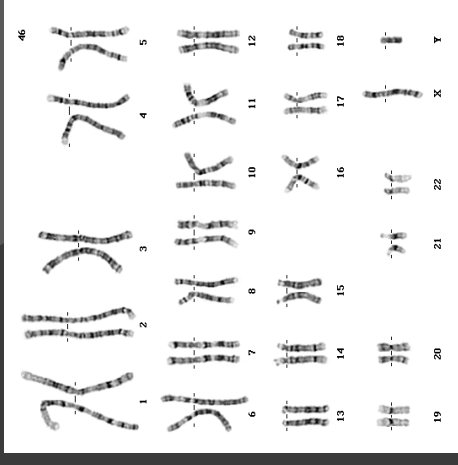
- The genome from a privacy standpoint
- What DNA profiles are (and aren't)
- Legal justification for DNA collection
- The rapidly expanding DNA database
- In favor of a universal DNA database

Overview

- The genome from a privacy standpoint
- What DNA profiles are (and aren't)
- Legal justification for DNA collection
- The rapidly expanding DNA database
- In favor of a universal DNA database

Human Genome Stats

- Human genome is...
 - ~3 billion nucleotides of DNA
 - (Each is A, C, G or T)
 - Divided among 23 chromosomes
 - (You have 2 copies of each, so total = 46 per cell)
 - ~2% coding (~genes)
 - ~98% non-coding (once called “junk DNA”)
 - 99.5% similar between any 2 humans
 - Genome is a privacy nightmare...



Genome is information-rich...

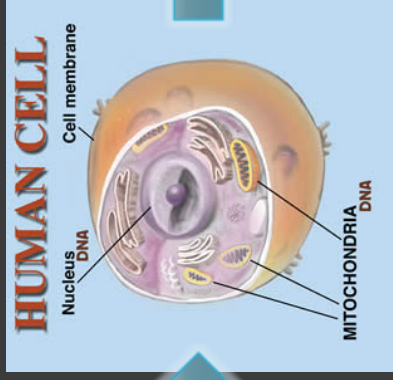
- Your DNA is your biological blueprint.
- Analyzing your genome can reveal...
 - Your phenotypic info (what you look like)
 - Your health information
 - Your disease propensities
 - & much more



... And you leave it everywhere.

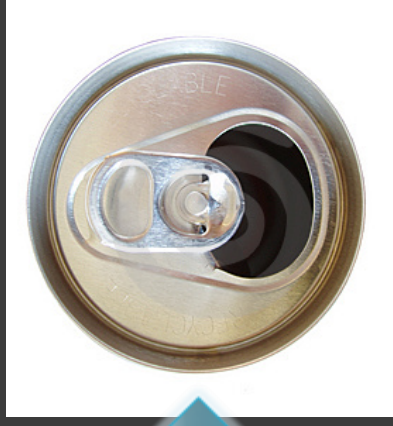


1 person



= 50+ trillion cells

Nearly all carry complete copy of genome



= full copies of genome shed routinely, countless times per day

Saliva, shed skin cells, etc.

Overview

- The genome from a privacy standpoint
- What DNA profiles are (and aren't)
- Legal justification for DNA collection
- The rapidly expanding DNA database
- In favor of a universal DNA database

DNA profiles are not the genome!

- Genome is rich with biologically meaningful information
- But DNA profile discards 99.999996% of this information
- Leaves only biologically meaningless remnants
 - repeated sequences used as identifiers

A DNA profile is...

- A set of 26 numbers
- A means to uniquely identify a given human individual based upon abstract representations of a few specific DNA marker sequences

A DNA profile is not...

- ⦿ Sequence information
 - e.g., ACGTCTATC...
- ⦿ Biologically meaningful
- ⦿ Capable of disclosing any health-related data, propensities, etc.

Building the Profile: STRs

- Genome contains many repeated sequences
 - Located in what was once called “junk DNA”
- One class of repeat is Short Tandem Repeat (STR)
 - Short sequence block (2-10 nucleotides)
 - Repeated anywhere from 5 to 100+ times



- Over 10,000 STRs in human genome
- FBI DNA profile characterizes just 13 of these
 - x 2 copies of each chromosome
 - = 26 loci total

What a DNA profile looks like

STR Locus	Allele 1 (# repeats)	Allele 2 (# repeats)
1	13	14
2	14	17
3	11	11
4	13	16
5	5	8
6	12	12
7	12	14
8	10	9
9	13	14
10	14	17
11	11	11
12	12	16
13	9	14

Summary of DNA profiles

- DNA profile is number of repeats at each of 26 alleles
- No correlation between number of repeats and any known biological characteristic
 - No health data, no genetic predisposition
- No sequence data! Just 26 numbers
- Despite this, they're robust identifiers
 - 26 alleles suffice for precise sample-to-source matching
 - Odds of exact match being wrong = 1 in several billion

Overview

- The genome from a privacy standpoint
- What DNA profiles are (and aren't)
- Legal justification for DNA collection
- The rapidly expanding DNA database
- In favor of a universal DNA database

FBI CODIS Database

- Combined DNA Index System (CODIS)
- Founded 1998, expanded rapidly since
- Now over 7 million profiles (world's largest)



Fourth Amendment Search

- Compulsory DNA collection constitutes a search under the Fourth Amendment
- Thus, in order to be constitutionally permissible, this search must be reasonable



Legality of DNA collection

- Circuit courts have upheld compulsory DNA collection in the face of Fourth Amendment challenges
- Totality of circumstances balancing test:
 - Legitimate government interest
 - Preventing crime!
 - Convict's diminished expectation of privacy

Overview

- The genome from a privacy standpoint
- What DNA profiles are (and aren't)
- Legal justification for DNA collection
- The rapidly expanding DNA database
- In favor of a universal DNA database

Who's in CODIS?

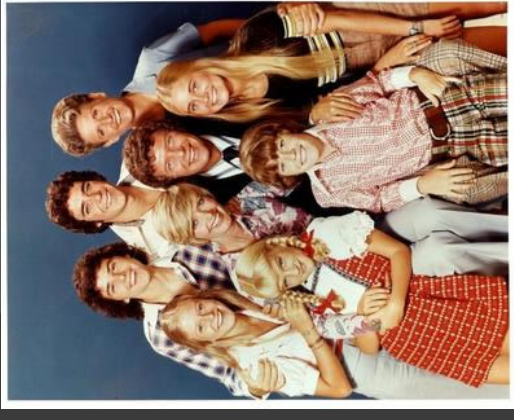
- 1998: Convicted violent felons (certain crimes)
- 2000s: Added more 'qualifying' offenses
- 2009: Arrestees
 - FBI and at least 15 states (including California) now profile arrestees

Arrestee DNA: Courts

- State courts have been mixed
 - Minnesota: Not OK to include
 - Virginia: OK to include
- 2009 case of first impression for federal courts
 - *US v Pool*, E.D.Cal. May 27, 2009
 - Holds: DNA collection from arrestees is OK
 - Court cites finding of judicial probable cause necessary to effect arrest in the first place
 - Again invokes diminished expectation of privacy

Familial DNA Search

- You commit a crime → forensic sample
- CODIS search using your sample returns no exact hits
 - You're not in CODIS
- What if your brother is profiled in CODIS?
- Partial search is much more likely to match a close relative of yours than some random person
 - Relatives share some DNA!
- Cops know your brother didn't do it – it's not a match – but they also know it's likely that one of his relatives did
- Cops use him to get to you
- You're effectively reachable through CODIS because your brother is profiled



Familial DNA: Legal status

- De facto expansion of CODIS database
- FBI doesn't do familial search
 - But now allows states to do so
 - California & Colorado do it
 - Only Maryland has banned it
 - Used regularly in the UK
- Courts have yet to hear the issue
- Hard to justify?
 - Surely, relatives of criminals do not have a diminished expectation of privacy!
 - But have they even been searched?
 - Search = actual cheek swab DNA sample?
 - Search = (effective) inclusion in the database?

Familial DNA: Technical Issues

- Technique doesn't work very well
- 26 alleles are **NOT enough** to effectively match relatives!
- MANY false positives
 - Odds of matching 13/26 alleles in general population is ~3%
 - i.e., about 200,000 partial matches from 7m records in CODIS
 - Partial matches may or may not be relatives
 - If not, then waste of time to pursue
 - Why should some random person know anything about the true perpetrator?
- CODIS software is not optimized to find relatives
 - Relatives show certain inheritance patterns
 - Software isn't configured to spot these

Aside: Racial bias in CODIS

- Certain races are overrepresented in CODIS database
 - Reflects reality of national crime statistics
 - Still a source of complaint about the DB
- BUT
 - If we start searching relatives, we greatly amplify this bias
 - We approach universal coverage for certain races only
 - Is this OK legally? Politically?

Overview

- The genome from a privacy standpoint
- What DNA profiles are (and aren't)
- Legal justification for DNA collection
- The rapidly expanding DNA database
- In favor of a universal DNA database

Forensic DNA vs. Genomic DNA

- DNA profile is a very limited use of genomic information
 - Contains NO sequence data
 - Contains NO biologically meaningful information at all
 - Resembles a (biologically-derived) social security number
- As a scientist, I was shocked by how much genomic information forensic labs DO NOT use
 - Police could theoretically work up a full phenotypic profile of a suspect based on crime scene DNA sample
 - “You’re looking for a white male, 6’1, red hair, brown eyes...”

Controlling Access vs. Use

- Controlling access to genomic DNA sample is difficult / impossible
 - Remember, we shed cells all the time
- Better to control use of data through regulation?
 - Current FBI DNA profiles are a surprisingly non-invasive use
 - They distill complex biological information down to a 26-field numeric identifier

Problems facing CODIS database

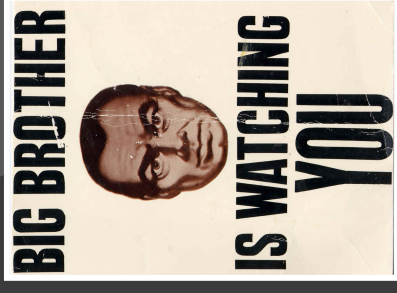
- Current expansion threatens to overstep legal grounds for profile inclusion
 - All based on diminished expectation of privacy
- Racial bias in database
- Government retains biological DNA sample
 - Blood, saliva, tissue, etc.
 - Risk:
 - Government could revisit this sample later and do new and different analyses
 - Genome sequence, health info / disease predispositions
- But this is a regulatory problem (what can they do with DNA?) more than a sample access problem

Possible solution: Universal DB

- ⦿ Restrict DB use to serious crimes (felonies?)
 - ⦿ No dragnet / DNA from litter!
- ⦿ Make collection 'quasi-mandatory':
 - Could require DNA profiling in order to get SSN, driver's license, etc.
 - Can envision legitimate governmental interests here
- ⦿ No sample retention
 - Or at least, robust regulations controlling what the government can and cannot do with your biological DNA sample
 - We'll need these rules even without a universal database!
 - Stiff penalties for misuse

Arguments against Universal DB

- ⦿ “Genetic surveillance”
 - But: Use database only for serious investigations!
- ⦿ Privacy of sensitive genetic data
 - I agree – but that’s not what DNA profile contains
- ⦿ Can’t justify 4th Amend. search of ordinary citizens
 - But: Justify under special needs / suspicionless search?
 - Citizens are photographed & ID’d routinely



Arguments for Universal DB

- No more (legally dubious) familial DNA search
 - If everyone is in the database, then no need to search relatives
 - Stick with powerful exact matches rather than error-prone partial match
- Cure racial bias
- Catch more criminals, sooner
 - including first-timers
 - catch criminals who offend multiple times before first arrest
- Prevent wrongful convictions
 - & exonerate wrongfully convicted
- Good deterrent!
- Low privacy risk if implemented properly

Conclusions

- Genome itself is a privacy nightmare
 - We need to think (fast) on how to protect this data
 - GINA does not go far enough, etc.
- BUT: DNA profiles contain no meaningful biological information
 - They're just identifier tags
- When used only to create such profiles, I don't think DNA collection needs to be justified using the 'diminished expectation' standard
 - And lots of good will result from having a comprehensive forensic database

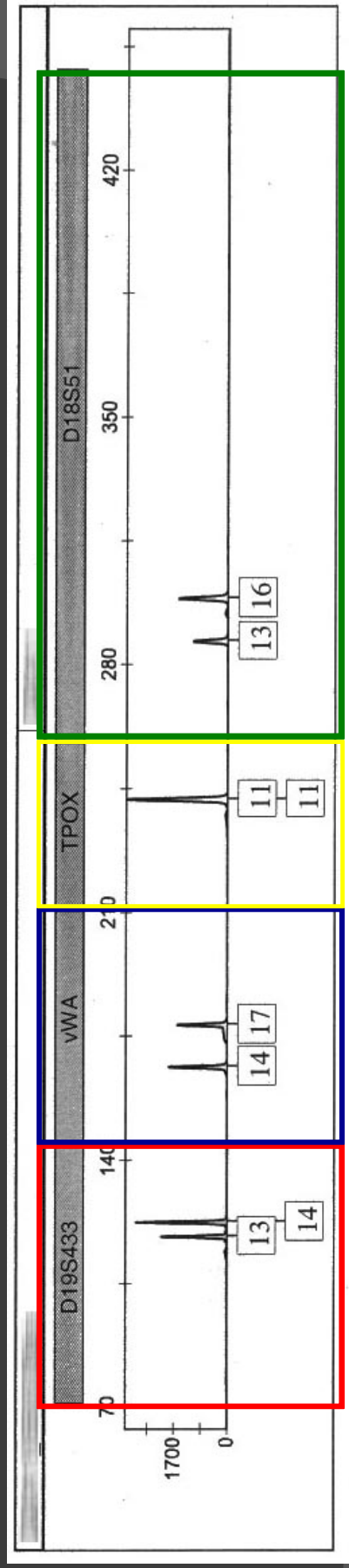
● Thanks!





What a DNA profile looks like

- Amplify each STR and determine total length of repeated sequence
 - This tells you number of repeats
- Recall: You have 2 copies of each chromosome
 - (One maternal copy, one paternal copy)
- So your DNA profile is 13 STRs x 2 copies of each
 - = 26 numbers



Familial DNA Search

- Identical 23/23 match = same person
- By searching at lower stringency, system can return partial profile matches
- Relatives share some DNA!
 - Siblings and parents/children share, on average, at least half their DNA sequence
 - Thus, at least half their 26 STR alleles
- By searching for partial matches, can match a crime scene sample to a relative of the perpetrator, if the relative is in CODIS
- De facto expansion of CODIS database!

